



A Resource for Students Conducting a DataJam Project

A Step-by-Step Guide Through a DataJam
Project

By: Tony Robol

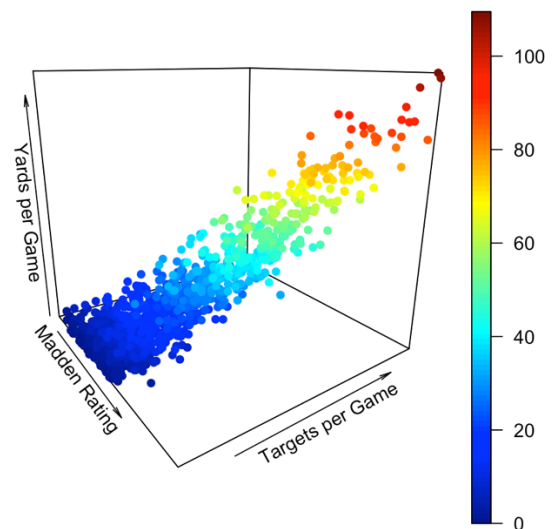


TABLE OF CONTENTS

| | |
|---|-----------|
| <i>PART ONE: A Statistical Approach to Analyzing a DataJam Project</i> | 4 |
| <i>Variables</i> | 5 |
| <i>Descriptive Statistics</i> | 6 |
| MEASURES OF CENTER | 6 |
| MEASURES OF SPREAD | 7 |
| MEASURES OF SHAPE | 7 |
| <i>Types of Distributions</i> | 8 |
| THE NORMAL DISTRIBUTION | 8 |
| SKEWED DISTRIBUTIONS | 8 |
| <i>Observations and Outliers</i> | 9 |
| WHAT ARE OBSERVATIONS? | 9 |
| WHAT ARE OUTLIERS? | 9 |
| <i>Control and Confounding Variables</i> | 10 |
| WHAT ARE CONTROL VARIABLES? | 10 |
| WHAT ARE CONFOUNDING VARIABLES? | 10 |
| <i>Sampling and Bias**</i> | 11 |
| SAMPLING AND TYPES OF SAMPLING TECHNIQUES | 11 |
| <i>Types of Visualizations for Data</i> | 12 |
| ONE CATEGORICAL VARIABLE | 13 |
| ONE QUANTITATIVE VARIABLE | 14 |
| ONE CATEGORICAL, ONE QUANTITATIVE VARIABLE | 15 |
| TWO CATEGORICAL VARIABLES | 16 |
| TWO QUANTITATIVE VARIABLES | 17 |
| <i>Regression</i> | 18 |
| SIMPLE LINEAR REGRESSION AND THE BEST-FIT LINE | 18 |
| The Best-Fit Line | 18 |
| The Correlation Coefficient (r) and R^2 values | 19 |
| MULTIPLE REGRESSION | 20 |
| <i>Regression Diagnostics and Conditions</i> | 22 |
| CONDITIONS FOR A LINEAR MODEL | 22 |
| UNNECESSARY PREDICTORS IN A MODEL | 23 |
| <i>Inferential Statistics</i> | 24 |

| | |
|---|-----------|
| CONFIDENCE INTERVALS | 24 |
| SIGNIFICANCE TESTS | 24 |
| <i>Glossary</i> | 26 |
| <i>PART TWO: How to Conduct a DataJam Project from Start to Finish</i> | 29 |
| <i>How to Conduct a DataJam Project: A SUMMARY</i> | 30 |
| <i>Exploring Topics</i> | 32 |
| AFTER CREATING YOUR LIST OF IDEAS | 32 |
| <i>Searching for Data</i> | 33 |
| <i>Writing Your Research Questions</i> | 34 |
| <i>Collecting and Preparing Data</i> | 35 |
| PREPARING DATA USING SOFTWARE | 35 |
| VERIFYING ENTRY ACCURACY | 35 |
| <i>Filling in Data Gaps</i> | 36 |
| <i>Finalizing Your Data</i> | 37 |
| DECIDING WHICH VARIABLES TO USE | 37 |
| FINALIZING YOUR VARIABLES | 37 |
| FINALIZING YOUR SPREADSHEET | 38 |
| <i>Analyzing Your Data</i> | 39 |
| EXCEL/GOOGLE SHEETS | 39 |
| MINITAB | 40 |
| TABLEAU PUBLIC | 42 |
| <i>Interpreting Data and Writing Conclusions</i> | 44 |
| <i>Limitations and Suggestions for Future Research</i> | 46 |

DISCLAIMER: This manual was designed in short, 1-2 page sections so that you may access the desired section quickly and not have to read through the entire manual.

**PART ONE: A Statistical
Approach to Analyzing a
DataJam Project**

Variables

The field of statistics revolves around variables because we use them to classify observations in a study/experiment and draw connections between these different classes of observations. A **variable** is defined as any measurable quantity or characteristic that can also be counted. Variables are the way we classify each category of data. There are several important classifications of variables that you must be aware of when conducting a DataJam project.

| Speed (mph) | Views |
|-------------|-------|
| 34.22 | 1217 |
| 23.83 | 1753 |
| 25.01 | 2768 |
| 71.33 | 4952 |
| 43.86 | 516 |
| 33.97 | 2833 |
| 37.89 | 1987 |
| 22.12 | 5657 |
| 45.75 | 4362 |
| 63.42 | 923 |

First and foremost is the quantitative variable. A **quantitative variable** is simply a variable that expresses a measurable quantity such as a number. Examples of quantitative variables include height, age, speed, number of views a YouTuber’s video gets, or mileage on all the cars in your school parking lot. Quantitative variables can further be classified into two categories, discrete and continuous variables. The difference between continuous and discrete variables is that **continuous variables** can be measured as fractional values of a whole, while **discrete variables** can only occupy sets of distinct whole values. For example, number of views a YouTuber’s video gets would be considered a discrete variable because views are not measured in fractional amounts; one click on the video equals one view. On the other hand, an example of a continuous variable would be speed, because you can travel at a fraction of a mile per hour (mph) or meter per second (m/s). Shown to the left is an example of two quantitative variables, speed and views on a theoretical up and coming YouTuber’s videos, and observations under those variables.

The other type of major variable classification you should be aware of is the categorical variable. **Categorical variables** are typically not represented by numbers (although they can be), instead usually having values that describe *qualities* rather than *quantities*. Examples of categorical variables include color, gender, letter grade, and month of the year. Shown to the right are examples of categorical variables, color and letter grade, along with several observations under those variables.

| Color | Letter Grade |
|--------|--------------|
| Red | C |
| Yellow | A |
| Blue | B |
| Blue | F |
| Red | B |
| Green | A |
| Orange | C |
| Orange | C |
| Yellow | D |
| Blue | A |

Now, the question you may be asking is: *How do I distinguish from categorical and quantitative variables?* There is actually a pretty simple answer to this question. First, ask yourself “Can I take the mean or median of this data set?”. If the answer is no, you have yourself a categorical variable. If the answer is yes, you have yourself a quantitative variable. There is no such thing as an average-colored car, but there is such thing as an average or median family income across the United States. This is because income is a quantitative variable, while car color is a categorical variable.

Take-Home Message: It is important to identify which of your variables are categorical and which of your variables are quantitative in a DataJam project!

Descriptive Statistics

Descriptive statistics do exactly what the name implies: describe the characteristics of your variables given the variables' observations. Using descriptive statistics can help you gain an immediate summary of what your dataset looks like as well as provide an idea of what the shape, center, spread, and other characteristics of your dataset are. Examples of commonly used descriptive statistics are as follows:

The **number of observations** is a commonly used descriptive statistic used to denote how many observations there are under a certain variable. This is usually denoted by n or *Total Count* in statistics. As seen to the right, for all variables in this dataset, $n = 60$.

Statistics

| Variable | Total Count | N | Mean |
|--------------------------------|-------------|----|--------|
| US Suicide Rate (15+) per mil | 60 | 60 | 9.7584 |
| US Homicide Rate per million | 60 | 60 | 4.8201 |
| Unemployment Rate (%) | 60 | 60 | 6.783 |
| Alcohol Poisonings per million | 60 | 60 | 0.4617 |
| Real Alcohol Sales (millions) | 60 | 60 | 8304 |
| US H1N1 Thousand Deaths | 60 | 60 | 0.212 |
| US H1N1 Hospitalizations | 60 | 60 | 4661 |
| Drug Overdoses per Million | 60 | 60 | 7.7300 |

MEASURES OF CENTER

Statistics

| Variable | Total Count | N | Mean |
|--------------------------------|-------------|----|--------|
| US Suicide Rate (15+) per mil | 60 | 60 | 9.7584 |
| US Homicide Rate per million | 60 | 60 | 4.8201 |
| Unemployment Rate (%) | 60 | 60 | 6.783 |
| Alcohol Poisonings per million | 60 | 60 | 0.4617 |
| Real Alcohol Sales (millions) | 60 | 60 | 8304 |
| US H1N1 Thousand Deaths | 60 | 60 | 0.212 |
| US H1N1 Hospitalizations | 60 | 60 | 4661 |
| Drug Overdoses per Million | 60 | 60 | 7.7300 |

The **mean** is a very commonly used measure to indicate the center of a *quantitative* dataset. Put simply, the mean is the average value of all observations within a variable. It is found by adding together the values of all observations and dividing by the total number of observations. As shown in the figure to the left, the mean US suicide rate per million people (ages 15 and up) is higher than the US homicide rate per million. Therefore, we can conclude that *on average*, during the years 2006-2010, more people committed suicide each month than homicide of another individual. It is worth noting that the mean should not be used to measure the center of a

dataset when there are outliers skewing the distribution of the variable.

The **median** is the other commonly used measure to indicate the center of a quantitative dataset. The median is the middle value of a numeric dataset and can be found by hand by following the note on the bottom right, assuming a relatively small dataset. In the event that there are two middle values (as is the case with all datasets with an even-numbered n), then the median is found by taking the mean of those two middle values. For larger datasets, it is best to use a software package or calculator to find the median. As shown to the right, the median suicide rate is yet again higher than the median homicide rate, so we reach the same conclusion as when we utilized the mean in this case.

| Variable | Q1 | Median |
|--------------------------------|--------|--------|
| US Suicide Rate (15+) per mil | 9.1980 | 9.8238 |
| US Homicide Rate per million | 4.4715 | 4.8547 |
| Unemployment Rate (%) | 4.700 | 5.700 |
| Alcohol Poisonings per million | 0.3600 | 0.5219 |
| Real Alcohol Sales (millions) | 7790 | 8495 |
| US H1N1 Thousand Deaths | 0.000 | 0.000 |
| US H1N1 Hospitalizations | 0 | 0 |
| Drug Overdoses per Million | 7.4567 | 7.7591 |

Finding the median

| | | | | | | | | |
|--|---------------|---------------|---------------|------|---------------|---------------|---------------|---------------|
| 10 | 37 | 41 | 26 | 17 | 54 | 33 | 27 | 8 |
| ↓ Arrange from least to greatest | | | | | | | | |
| 8 | 10 | 17 | 26 | 27 | 33 | 37 | 41 | 54 |
| ↓ Alternate crossing out numbers from each end until the middle is reached | | | | | | | | |
| 8 | 10 | 17 | 26 | (27) | 33 | 37 | 41 | 54 |
| Median = 27 | | | | | | | | |

For a normal distribution,
IQR > Standard Deviation

MEASURES OF SPREAD

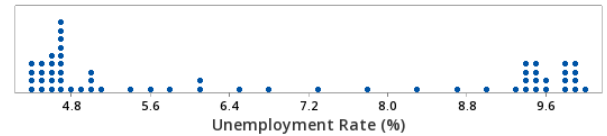
The **standard deviation** is the most common way to measure the spread of a quantitative dataset, or the way in which observations are arranged in a distribution. The standard deviation indicates how closely observations are situated to the mean; a larger value means that observations are located further away from the mean. This measure is typically difficult to calculate by hand, especially with variables containing a large n , so calculating this measure is usually best left to a calculator or software package such as Minitab or Excel. When dealing with outliers in a dataset, perhaps one of the best ways to measure spread is the **interquartile range**. The interquartile range is the difference between the observation with a value at the 75th percentile within a dataset and the observation with a value at the 25th percentile in a dataset. The interquartile range is usually best found through the use of a software package or calculator, especially with large datasets.

Statistics

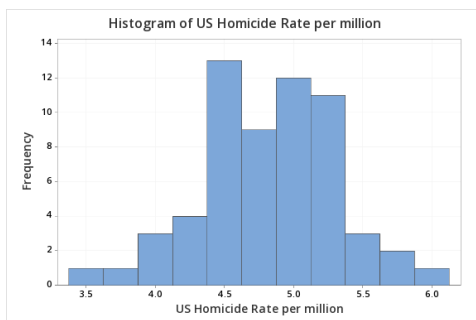
| Variable | StDev | IQR |
|--------------------------------|--------|--------|
| US Suicide Rate (15+) per mil | 0.6598 | 1.0337 |
| US Homicide Rate per million | 0.5030 | 0.7079 |
| Unemployment Rate (%) | 2.287 | 4.775 |
| Alcohol Poisonings per million | 0.2094 | 0.2220 |
| Real Alcohol Sales (millions) | 981 | 1189 |
| US H1N1 Thousand Deaths | 0.812 | 0.000 |
| US H1N1 Hospitalizations | 17362 | 0 |
| Drug Overdoses per Million | 0.3832 | 0.5602 |

MEASURES OF SHAPE

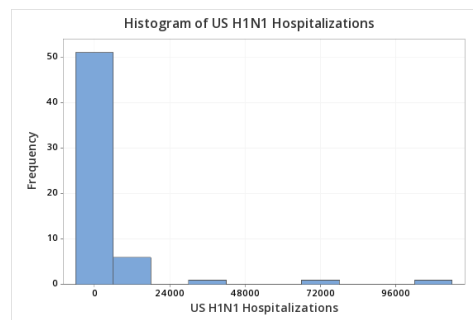
The **mode** is an indicator of the most frequent observation within a variable and/or dataset. It is found by simply identifying the value of the observation that appears most frequently in a dataset. The mode is usually a good indicator of where the peak(s) is within a distribution of a variable. As shown to the right, 4.7% is the mode unemployment rate, as it appears 9 times throughout the years 2006-2010, more frequently than any other unemployment rate.



Finally, two measures to check whether a distribution is (approximately) normal are the **skewness** and the **kurtosis**. The skewness is, as the name implies, the degree of skew away from a normal distribution, while the kurtosis is a measure of the thickness of the tails in a distribution. These measures are most easily calculated by a software package such as Minitab or Microsoft Excel. In order to safely assume a distribution as approximately normal, the skewness must fall within the range of -0.5 to 0.5 and the kurtosis within -2 to 2. (Usually if a distribution has high skewness, it also has high kurtosis and vice versa)



Low skewness, Low kurtosis



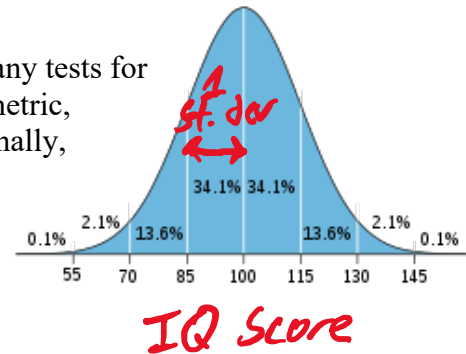
High skewness, High kurtosis

Types of Distributions

In statistics, it is important to know what kind of distributions you have across your variables to check the conditions for statistical significance. In most cases, it is important that your variables follow a normal distribution, as it is the distribution upon which many tests for statistical significance are based. Ensuring your distribution(s) is normal is extremely important for DataJam projects, else you will need to use analyses not requiring normal distributions. The two other major types of distributions are skewed left and skewed right; these distributions usually contain outliers and are asymmetrical.

THE NORMAL DISTRIBUTION

The **normal distribution** is the most important distribution in statistics, as many tests for statistical significance are based around it. As seen to the right [2], it is a symmetric, unimodal (one peak) distribution, where the mean equals the median. Additionally, close to all observations fall within three standard deviations of the mean (or median). Good examples of (approximately) normal distributions include IQ (seen to the right) and human height.

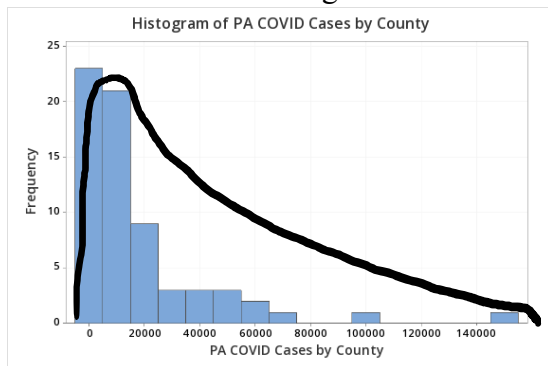


SKEWED DISTRIBUTIONS

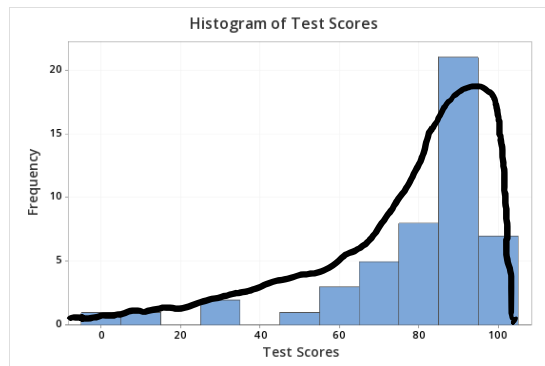
Skewed left and **skewed right** distributions often include outliers that drag the mean away from the median, or peak, of the distribution. Likewise, in both of these cases, the median does not equal the mean. More specifically, in a left-skewed distribution, the mean is less than the median, and in a right-skewed distribution, the mean is greater than the median. A trick to remember this for skewed distributions is that “the mean follows the tail”. A good example of a skewed left distribution is test scores (shown below), while good examples of right-skewed distributions are income and PA COVID-19 cases by county [3] (shown below).

Now, we know that mean and median both describe the center of a distribution. But, for a skewed distribution, they are different! So the question becomes, which one do we use? The answer is the median, as its value is *resistant* to the presence of outliers. Thus, its value does not change much from the true center with the addition of an outlier, unlike the mean, whose value changes drastically.

Skewed Right



Skewed Left



Observations and Outliers

WHAT ARE OBSERVATIONS?

In statistics, **observations** are simply the data points you analyze through scatterplots, histograms, boxplots, bar graphs, etc. Observations can contain values for one variable, such as color or mileage, or they can contain multiple corresponding values, such as characteristics of a car: mileage, weight, fuel efficiency, color, number of seats, etc. Usually, good DataJam projects take into account multiple variables when considering their research questions, so your observations will almost always contain multiple corresponding values. In my DataJam project considering the effects of the 2007-2010 H1N1 pandemic, my observations were months from 2006-2010, but the corresponding values for the months included the homicide rate, suicide rate, drug overdoses, current unemployment rate, among other factors shown below:

| Month/Year | US Suicide Rate (15+) per million | US Homicide Rate per million | Unemployment Rate (%) | Alcohol Poisonings per million | Real Alcohol Sales (millions of \$) | US H1N1 Thousand Deaths | Drug Overdoses per Million | H1N1 Thousand Deaths*2 | Seasons | Pandemic? |
|--------------|-----------------------------------|------------------------------|-----------------------|--------------------------------|-------------------------------------|-------------------------|----------------------------|------------------------|---------|-----------|
| January-06 | 9.10476 | 5.23591 | 4.7 | 0.12122 | 38.77337719 | 0 | 6.84877 | 0.00000 | Winter | 0 |
| February-06 | 8.07149 | 4.11481 | 4.8 | 0.11777 | 44.7613719 | 0 | 6.67857 | 0.00000 | Winter | 0 |
| March-06 | 9.20238 | 4.55916 | 4.7 | 0.11095 | 42.36025044 | 0 | 7.39015 | 0.00000 | Winter | 0 |
| April-06 | 9.48430 | 5.12683 | 4.7 | 0.12095 | 38.85981711 | 0 | 7.30389 | 0.00000 | Spring | 0 |
| May-06 | 9.83317 | 5.38155 | 4.6 | 0.08057 | 37.09163615 | 0 | 7.67786 | 0.00000 | Spring | 0 |
| June-06 | 9.59687 | 5.39048 | 4.6 | 0.08721 | 52.74397897 | 0 | 7.82240 | 0.00000 | Spring | 0 |
| July-06 | 9.96046 | 6.02252 | 4.7 | 0.12400 | 37.90231315 | 0 | 8.07695 | 0.00000 | Summer | 0 |
| August-06 | 9.37948 | 5.31426 | 4.7 | 0.11720 | 40.10182061 | 0 | 7.65829 | 0.00000 | Summer | 0 |
| September-06 | 9.06632 | 5.32939 | 4.5 | 0.10037 | 44.831324 | 0 | 7.38353 | 0.00000 | Summer | 0 |
| October-06 | 9.51665 | 5.34831 | 4.4 | 0.09697 | 41.13419799 | 0 | 7.30713 | 0.00000 | Fall | 0 |
| November-06 | 8.93800 | 5.10361 | 4.5 | 0.10620 | 44.00000000 | 0 | 7.18113 | 0.00000 | Fall | 0 |
| December-06 | 8.72327 | 5.33274 | 4.4 | 0.15685 | 28.05310752 | 0 | 7.17484 | 0.00000 | Fall | 0 |

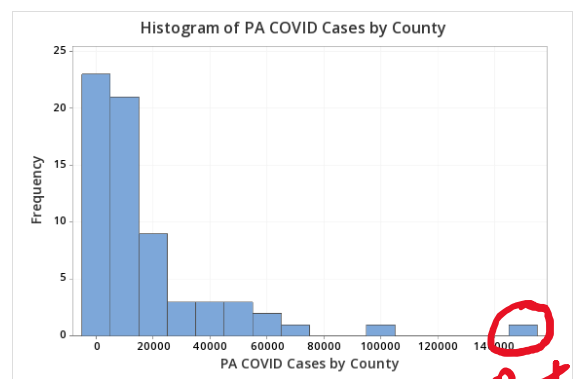
One Observation!

Across 10 variables

WHAT ARE OUTLIERS?

Outliers are observations that are significantly removed from the rest of the observations, having either much higher or lower values than the rest of the observations (example shown to the right). Outliers usually skew the data away from a proper normal distribution, which can be problematic for tests and checks of data requiring there to be a normal distribution. Additionally, outliers make it tough to identify the center and spread of a dataset.

Luckily, there is a solution to this problem: **resistant** measures. Resistant measures are either not or very slightly affected by the presence of outliers in a dataset. Examples of resistant measures are the median for center and the interquartile range for spread (discussed further in module ____). Measures that are not resistant include the mean and standard deviation, as adding a very large or very small value to the dataset can heavily influence their calculations.



Outlier

Take-Home Message: Use *resistant* measures when forced to deal with outliers!

Control and Confounding Variables

In a study or experiment, it is important to identify which factors are held constant throughout, or the **control variables**, and what factors could potentially deviate the observed results from the actual results in the real world. These potentially deviating factors are known as **confounding variables**. In a DataJam project, it is very important you consider the factors that may confound the results, and then control for them through your variables.

WHAT ARE CONTROL VARIABLES?

As stated above, it is imperative in a sophisticated DataJam project that you understand which factors need to be held constant to best measure the desired relationship. For example, say you want to observe the growth of plants in various degrees of sunlight. So, you place the plants in various degrees of shade. However, you water the plants in the sunlight more often than the ones in the shade. Naturally, you observe that the plants in the sunlight grow more than the ones in the shade, but this could either be due to the exposure to sunlight *or* the amount of water given – we don't know! Thus, we can't answer our research question, which is why it is very important to *control*, or hold constant, the amount of water you give each plant to truly measure the relationship between growth and amount of sunlight.

WHAT ARE CONFOUNDING VARIABLES?

Confounding variables are variables that can complicate your results away from what they truly are in the real world. In the previous plant example, the varying amounts of water would be an example of a confounding variable because it complicates, or *confounds*, the relationship between amount of sunlight and plant growth.

In my mental health study, I wanted to measure alcohol consumption as a predictor, but I could only find data on alcohol sales*. I could have just used this as my variable to measure alcohol consumption, but I realized that the monetary value of alcohol sales may have gone up over time merely due to *inflation*. In this case, *inflation* was the confounding variable, and I had to include an inflation adjustment calculation to *control* for inflation and measure the desired quantity: alcohol consumption (measured by *real* alcohol sales).

| Alcohol Sales (millions of \$) | Inflation Adjustment | Real Alcohol Sales (millions of \$) |
|--------------------------------|----------------------|-------------------------------------|
| 6066 | 1 | 6066 |
| 6590 | 1.00202 | 6576.715036 |
| 7923 | 1.00756 | 7863.55155 |
| 7335 | 1.01614 | 7218.493515 |
| 8843 | 1.02118 | 8659.589886 |
| 9327 | 1.02320 | 9115.519937 |
| 7792 | 1.02622 | 7592.9138 |
| 9156 | 1.02824 | 8904.535906 |
| 8037 | 1.02320 | 7854.769351 |
| 8640 | 1.01765 | 8490.148872 |

Handwritten notes in green:
 Divide This... (next to Alcohol Sales)
 By This... (next to Inflation Adjustment)
 To get This! (next to Real Alcohol Sales)

*Sometimes you may not find the exact variable you are looking for, so this is where you will need to get creative with the data you are given!

Sampling and Bias**

If you ever happen to collect data during a DataJam project (unlikely), it is important to know how to appropriately generalize the results of your study and avoid bias so you can have as accurate a conclusion as possible. **Bias** is the deviation in the results of your study from the true value of the population parameter. (Note: a **parameter** is simply a statistic that pertains to an entire population). As such, we want to avoid bias as much as possible in DataJam studies so our results come as close as possible to the actual value of the population parameter. This can be done using sampling techniques, as described below.

SAMPLING AND TYPES OF SAMPLING TECHNIQUES

In statistics, **sampling** is the process by which a subset of the population is taken and observed for a characteristic of interest. Then, once the researcher collects data on the subset, they then attempt to generalize the results to the entire population. However, there is usually bias associated with this generalization, as any sample is usually not truly representative of the entire population. Now, you might be thinking, *why not just collect the data directly from the whole population?* Such a technique might be plausible for very small populations, but for very large populations, such as the United States, this is nearly impossible. This is why we use samples in statistics.

A **simple random sample** is the best type of sample for a statistician to conduct and often yields the best results with the least amount of bias. However, it is typically very difficult to conduct, as all members of the sample are randomly generated, and thus need to be both sought out and willing to respond to the study/experiment. When this is not possible, easier, yet more biased methods explained below must be used.

A **cluster sample** is sort of like a simple random sample, but easier to conduct as observations are usually not located all over the place. First, you need to divide the population in smaller “clusters”. Then, use a randomizer to select certain clusters, and then interview all observations in those clusters.

A **systematic sample** is when the researcher collects observations by interviewing every k^{th} person they, for example, pass by. For instance, if I wanted to use a systematic sample to find out what the average weight of a student’s backpack is in high school, and I need 100 data points out of a high school with 900 students, I would interview every 9th student that walks in the building.

A **convenience sample** is the easiest type of sample to conduct, but often yields the most amount of bias due to it usually not being representative of the true population. It involves simply interviewing people that are closest to you and/or willing to respond to the study/experiment. Try this technique as a last resort if you cannot conduct any of the above samples.

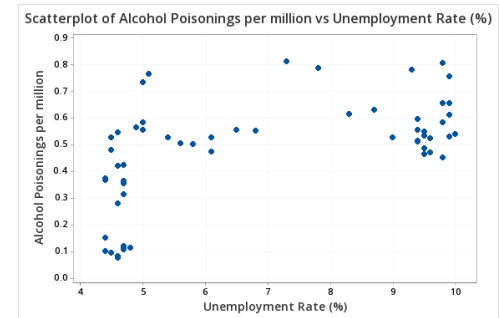
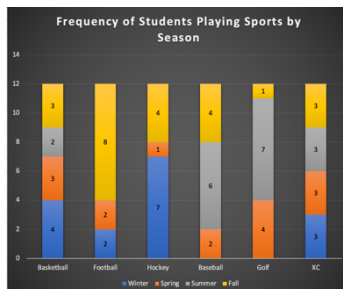
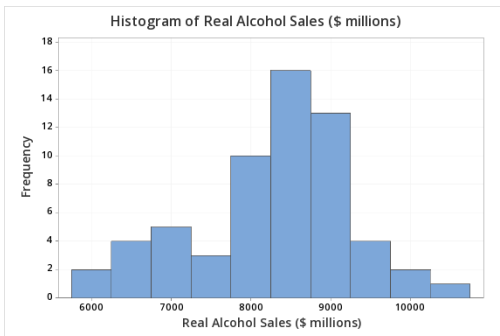
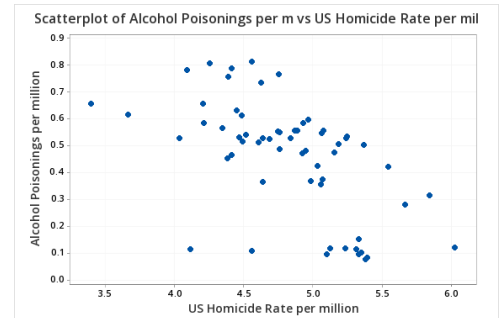
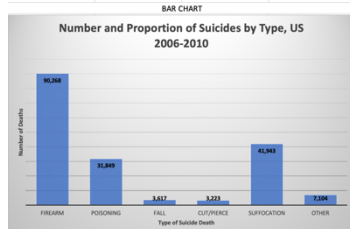
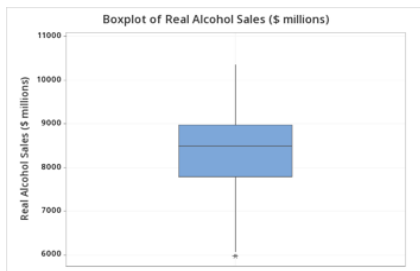
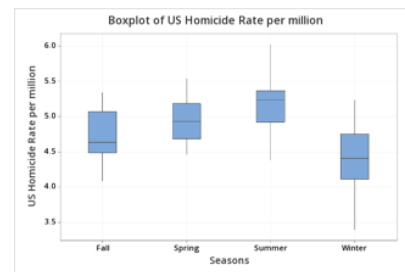
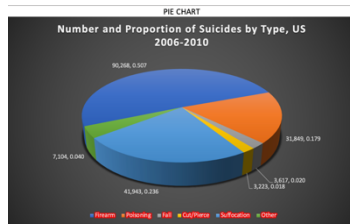
Take home message: In a DataJam project, it is best to minimize bias!

**Note: This section is only relevant to you if you must collect your own data, which is very rare and not required for you to have to do in a DataJam project. Instead, just ensure your data is reliable.

Types of Visualizations for Data

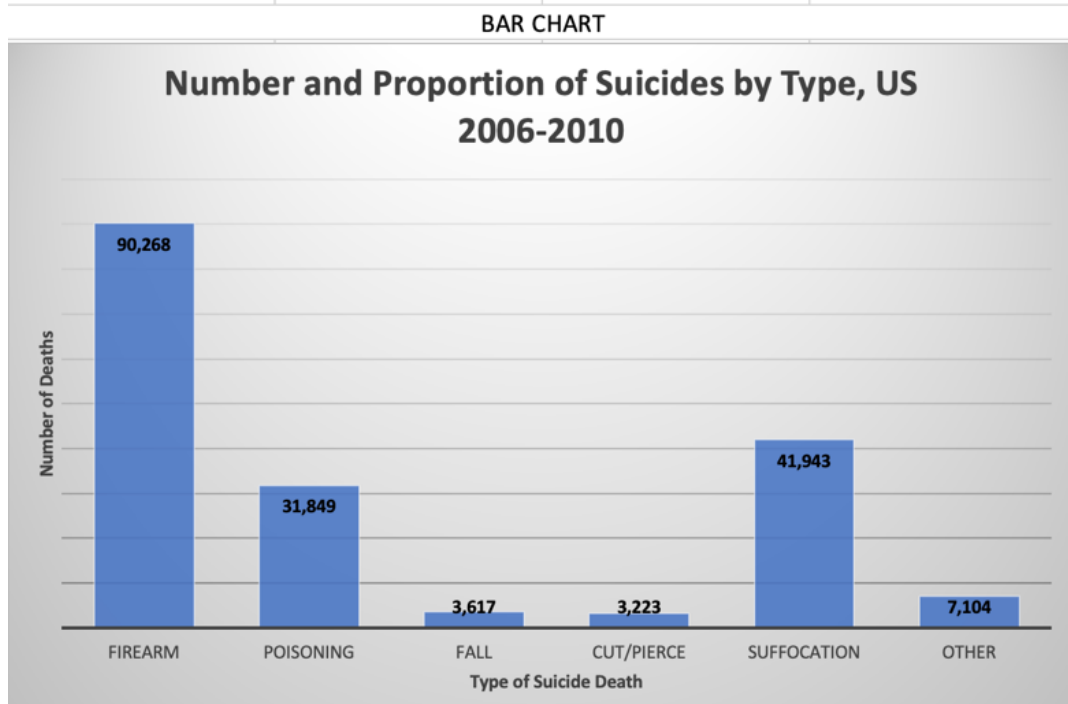
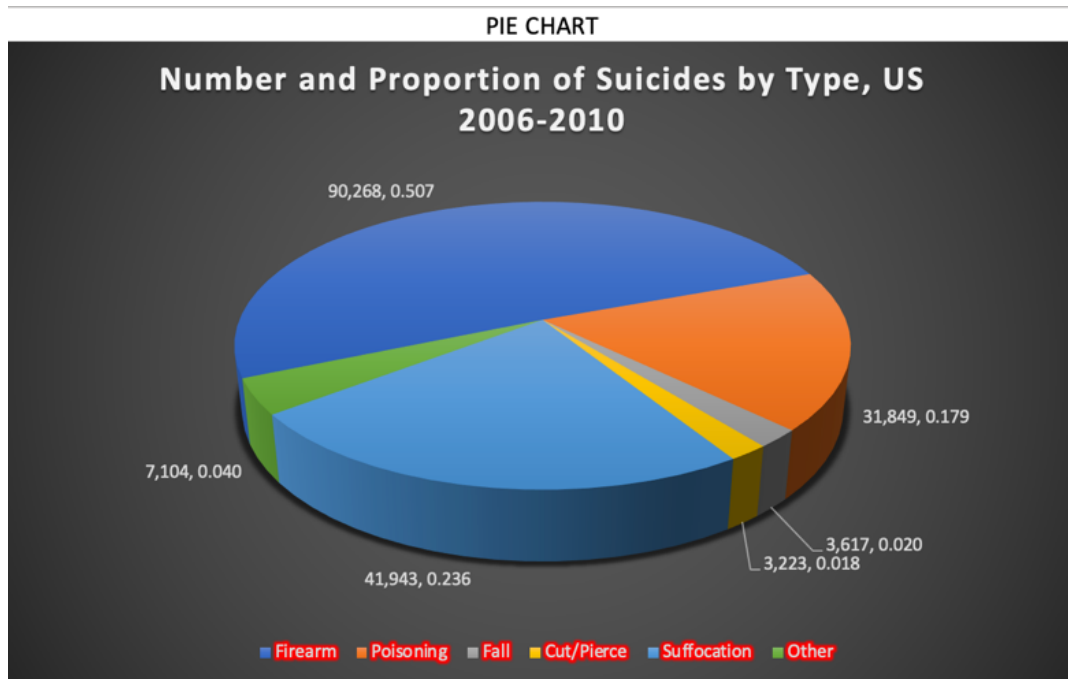
One of the most important aspects of a DataJam project is visualizing your results with a graphic for a diverse audience to understand. This is often the most exciting part of the project as well: finally getting to see for yourself what your results look like. However, to do this, you must know what type of graphic to use so it works with the type of variables you are working with (if you are having trouble classifying the variables you are working with, see the [Variables](#) section). A preview of all the graphics that will be discussed on the following pages is shown below.

Note: When using graphical displays, it is imperative that you always LABEL your graphics so others can better understand them!! This includes labeling every category, axis, etc. and giving your graph a title!



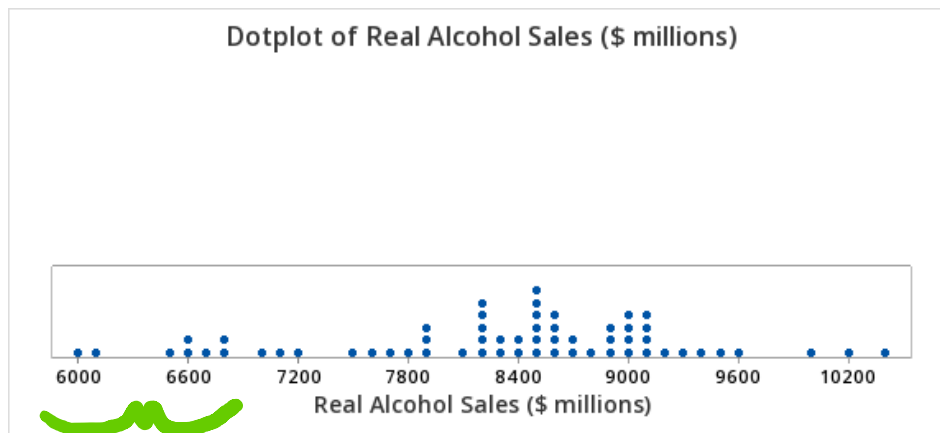
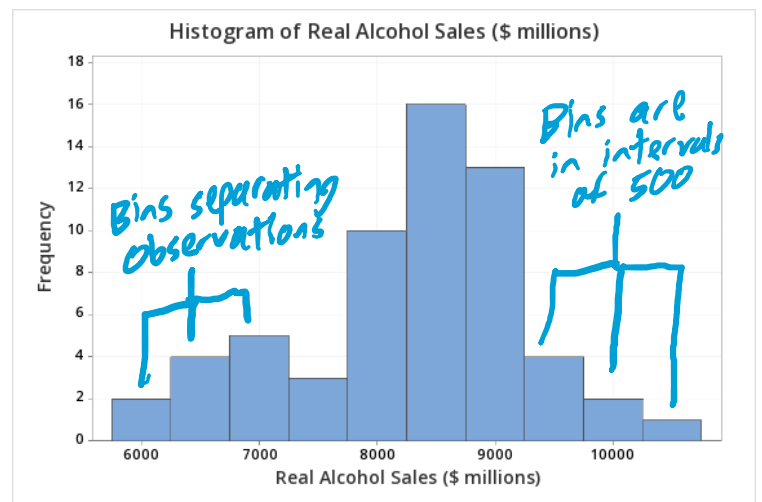
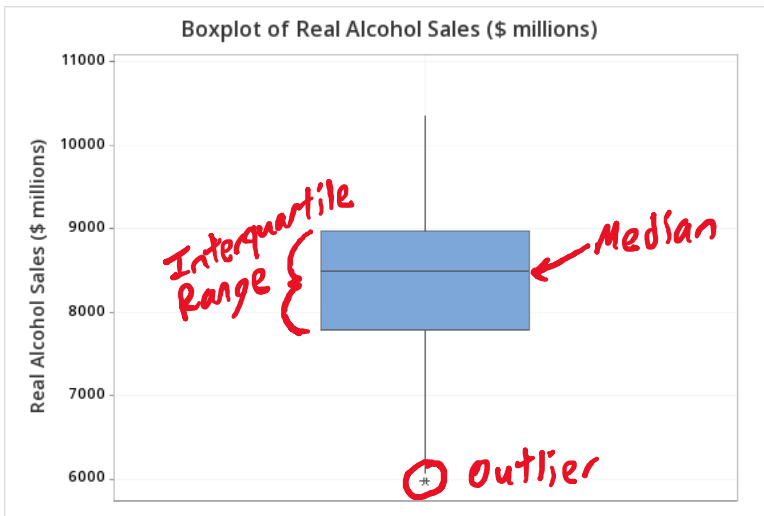
ONE CATEGORICAL VARIABLE

For displays involving just one categorical variable, you will want to use either a **pie chart** or **bar graph** showing the proportions or percentage of each category with respect to the whole. In my case, if I wanted to investigate *how* individuals committed suicide rather than just the number each month, I could include a bar graph or pie chart showing the proportion of people that committed suicide via firearm, intentional overdose, etc. An example of what this would look like is shown below:



ONE QUANTITATIVE VARIABLE

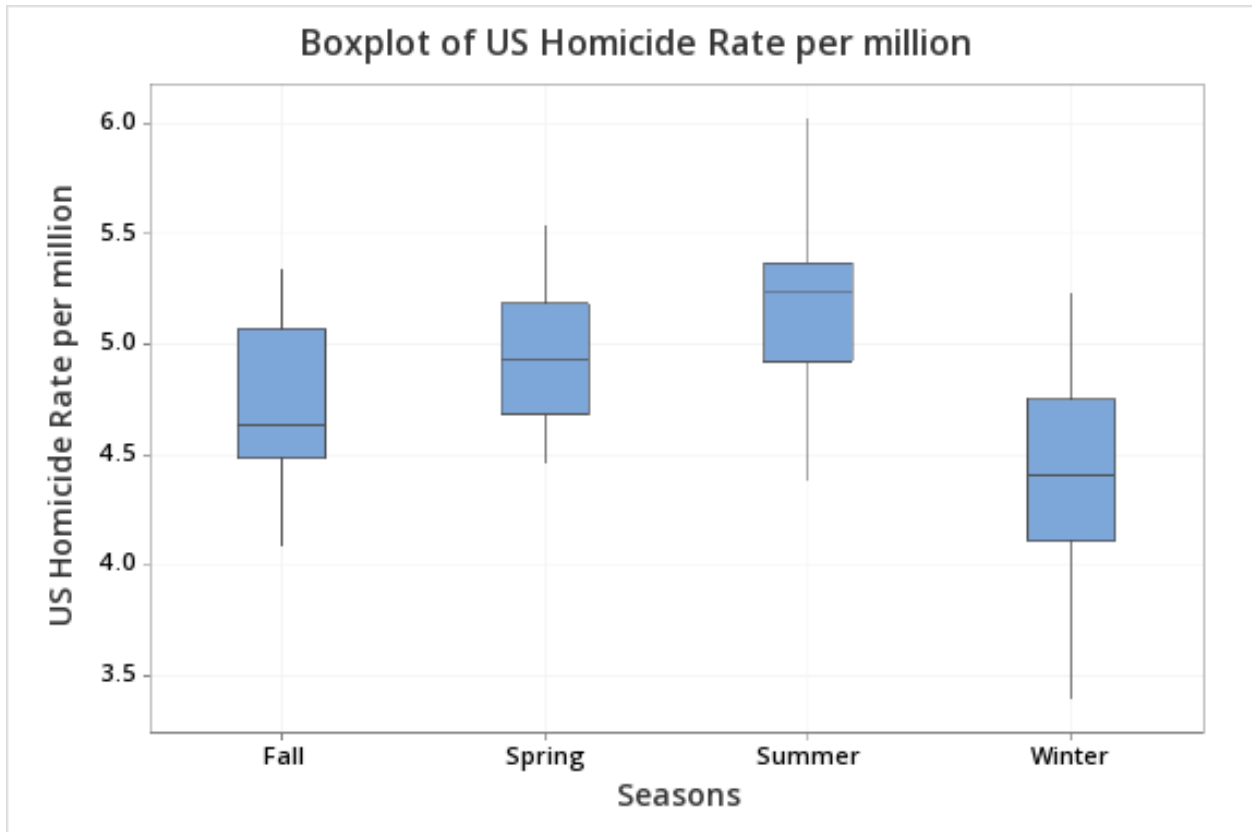
For displays involving one quantitative variable, a histogram, boxplot (AKA box-and-whisker plot), or dot plot are the best options. However, each one has its own strengths in describing shape, center, and spread, as discussed above in the descriptive statistics section. A **boxplot** is best for showing the center and interquartile range (spread), as denoted by the center line in the box and two ends of the box, respectively. Boxplots can also show outliers, as denoted by asterisks (*). On the other hand, while a boxplot can also show a distribution's shape, a **histogram** or **dot plot** is best for that. Histograms and dot plots are fundamentally very similar, except a histogram shows the number of observations in a certain bin of the histogram on the y-axis, while a dot plot depicts the number of observations by the number of dots above that bin. An example of each of these is shown below:



Smaller bins,
of dots corresponds
to # of observations
in that bin

ONE CATEGORICAL, ONE QUANTITATIVE VARIABLE

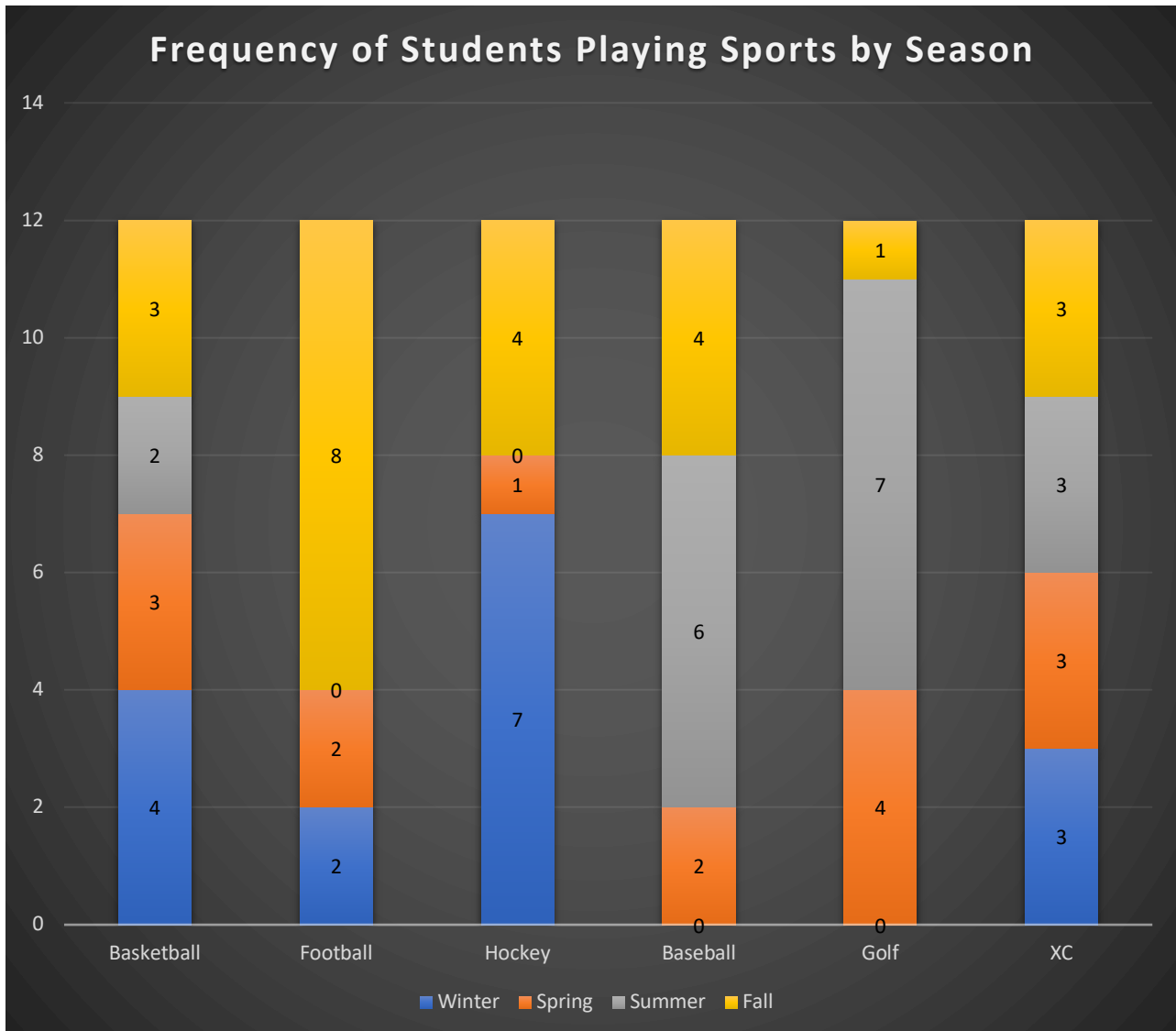
When dealing with one categorical and one quantitative variable together, it is best to use a side-by-side boxplot. A **side-by-side boxplot** shows differences in the distributions of the quantitative variable, as denoted by the boxplots, across the different categories of the categorical variable. These side-by-side boxplots are useful in determining if there is a category that is significantly different than the others and needs to be further examined. In my case, I needed to examine if there was a certain season in which homicide rate was significantly higher or lower than the others, and thus needed to have further examination into. The graphic of the side-by-side boxplot I examined can be found below.



As shown above, there is no one boxplot that is *significantly* higher or lower than the others, so I could conclude that there was no further investigation I needed to do between homicide rate and season.

TWO CATEGORICAL VARIABLES

When your data consists of two categorical variables, the best graphic to use is a stacked bar chart. A **stacked bar graph** shows the proportions of the categories across the observations of the first categorical variable across the categories of the second categorical variable. A stacked bar graph may be useful if you would like to compare these proportions across the categories of another categorical variable. An example can be found below.



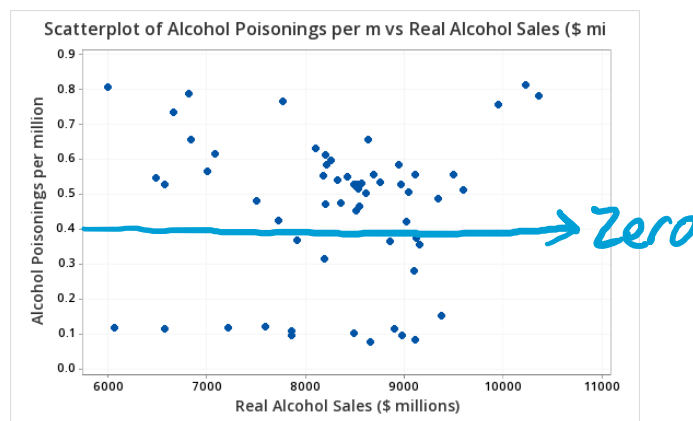
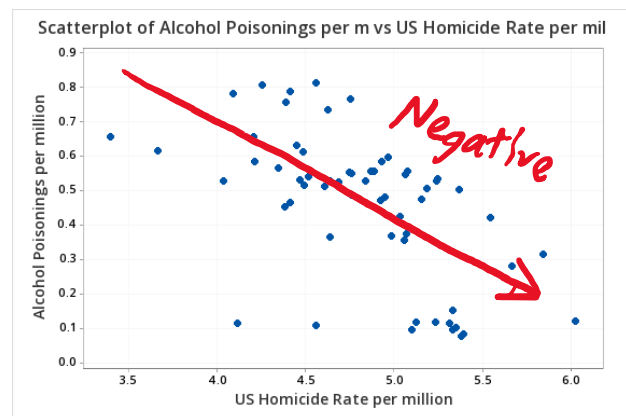
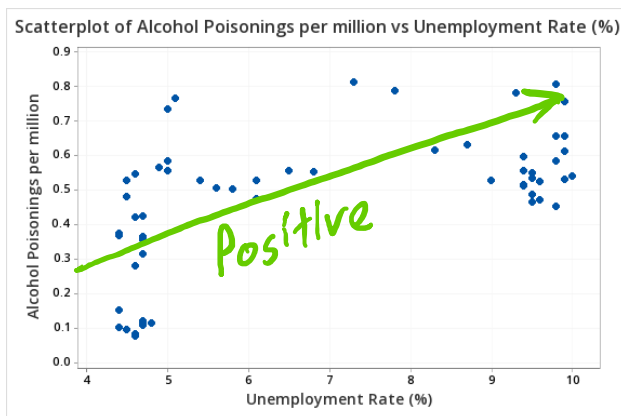
As depicted by the graph, sports such as football and golf are played predominantly in one season: fall and summer, respectively. However, a sport like cross country can be conducted in any season, so it is more spread out across the four seasons. A stacked bar chart can help to show the distributions of these categories.

TWO QUANTITATIVE VARIABLES

When working with two quantitative variables, the type of graphical display to use, and perhaps the most popular display among all DataJam projects, is the scatterplot. A **scatterplot** uses an xy-coordinate system to plot observations between two quantitative variables, one of which is placed on the y-axis and one of which is placed on the x-axis. When working with a scatterplot, it is important to note three things about the distribution of observations:

- The *strength* – how closely the data points fall to a best fit line, can be either strong, moderate, or weak
- The *direction* – Is the relationship positive and upward-sloping, negative and downward-sloping, or is there seemingly no relationship?
- Whether the relationship is *linear* or *nonlinear* – Can the points be modeled by a straight line, or is there a different type of relationship to describe the data points?

There is a lot you can do with a scatterplot, and what we discuss in this module just scratches the surface. However, these are the basics you will need to understand what will be covered in the following modules – all of which deal with scatterplots and regression. Below are several examples of scatterplots depicting different strengths, directions, and fits.



There is a lot more you can do with scatterplots besides examine the directions as shown above. These other things are discussed in the next section.

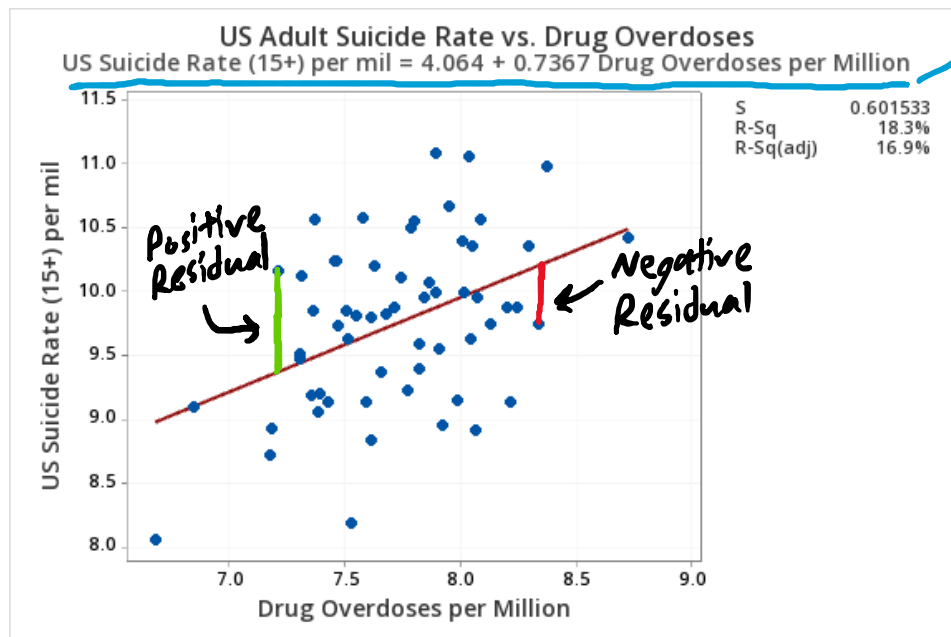
Take-Home Message: Know what type of variables you have before making visualizations!

Regression

Regression is probably the most commonly used technique within a DataJam project, and for good reason – there is truly a LOT you can do with it. Likewise, it is imperative that you understand the basics of it so that you are comfortable using it in your project. In these next few sections, we will start simple, with basic linear regression, and build up to the more complex multiple regression.

SIMPLE LINEAR REGRESSION AND THE BEST-FIT LINE

Simple linear regression involves fitting a best-fit line to data points involving the Y-variable (also known as the *response* variable) and just one X-variable (also known as the *predictor* or *explanatory* variable). This best-fit line is calculated by minimizing what are called residuals. The **residual** of a data point is the difference between its observed (or actual) value and the value predicted by the line of best fit (see below). To calculate the best-fit line, these residuals are squared (so all of them are positive) and then summed; the equation of the line generating the lowest sum of the squared residuals is the best-fit line*.



Regression Equation

The Best-Fit Line

The best-fit line has several very important properties that can be analyzed and interpreted for a conclusion. The first and foremost of these properties is the regression equation. The **regression equation** is simply the equation of the best-fit line and can be used to make predictions about what the Y-variable value would be at a certain X-variable value and vice versa. These predictions can fall into one of two categories: interpolation and extrapolation.

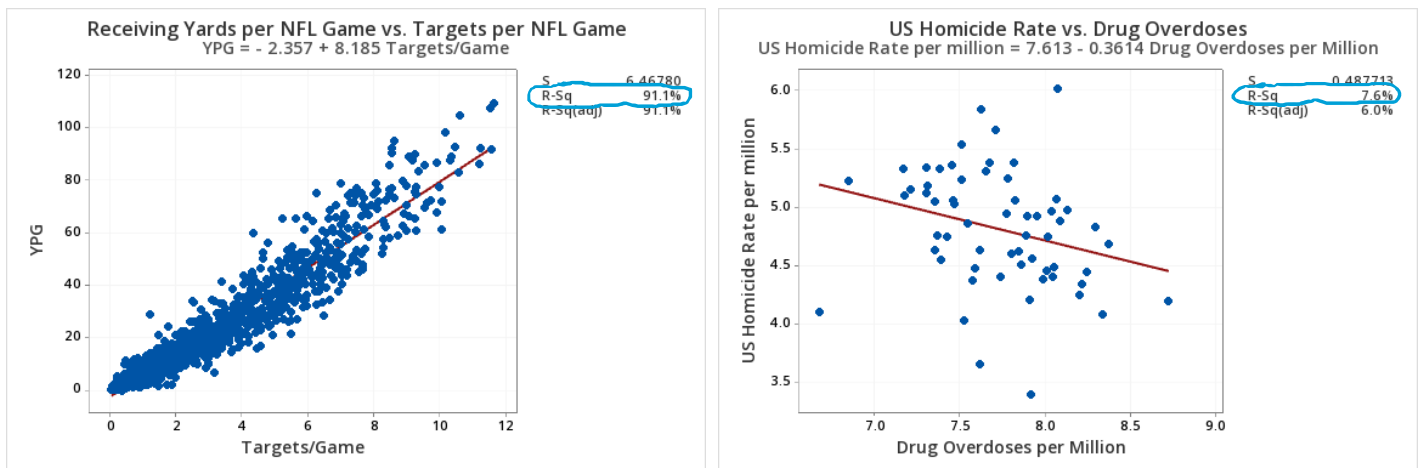
Interpolation involves using the best fit line and a variable input to predict the value of the other variable *within the range of the plotted observations*. For example, in the above graph, new predictions would be considered interpolation if the Drug Overdoses input falls between 7.0 and 8.5 or the Suicide Rate input falls within 8.5 and 11.0.

Conversely, **extrapolation** involves using the best-fit line and a variable input to predict the value of the other variable *beyond* the range of the plotted observations. For example, if the Drug Overdoses input is over 9.0 or the Suicide Rate input is above 11.5, this would be considered extrapolation. Generally, extrapolation is not advisable due to the unknown nature of the best-fit line beyond the range of the plotted points, so only use extrapolation as a predictive method if there are no other options.

The Correlation Coefficient (r) and R^2 values

Another important feature of the best-fit line is that it comes with a calculation of something called the correlation coefficient, or r value. The **correlation coefficient (r)** measures both the strength and direction (discussed in the scatterplot module briefly) of the correlation between the X and Y variables. r ranges from -1 to 1, with values close to 1 and -1 being considered a strong correlation and values close to 0 being considered little to no correlation. The direction is simply indicated by the sign of the r -value; a positive r means a positive correlation and a negative r indicates a negative correlation.

Perhaps even more important than the r value is the R^2 value, which is simply the square of the r value. The **R^2 value** measures the amount, or percent of variation in the Y-variable that is explained by the X-variable. The remaining percent (out of 100%) is considered error. Likewise, a higher R^2 value indicates greater predictive power of the best-fit line due to less percent being allocated to predictive error, so this quantity is extremely valuable in drawing predictions and conclusions from the data. Shown below are two examples of simple linear regression: one with a weak correlation and one with a strong correlation.



On the left is an earlier analysis I conducted using NFL receiving data from 2017-2019. The R^2 value is a whopping 91.1%, meaning that only 8.9% is allocated to predictive error. Likewise, any predictions of new observations using this line are likely to be very accurate, with a relatively small residual. On the other hand, the right graph only has an R^2 value of 7.6%, meaning that 92.4% is allocated to error. Unfortunately, this means that this line does not have very good predictive power, and that the actual value of new observations may be widely different than the value predicted by the line. Keep in mind that residuals and regression error are closely connected – we want residuals to be small!

Take-Home Message: When performing simple linear regression, it is best to take note of the regression equation and R^2 value, as there is a lot you can infer from both pieces!

MULTIPLE REGRESSION

While it is important to know the principles of simple linear regression, performing a mere simple linear regression with one X and one Y variable for your DataJam project is often not complex enough. A winning DataJam project considers several factors around their main issue rather than just one. However, many students ask: *how do I go beyond a scatterplot to analyze these additional factors?* This is where multiple regression comes into play. **Multiple regression** involves the use of multiple X-variables to explain additional variation in a single Y-variable. In the case of a DataJam project, your single Y-variable will be the main variable you wish to study, with the X-variables being the factors that may explain the values of the Y-variable. There are several key differences between multiple regression and simple linear regression, as highlighted below.

More Variables!

The first and foremost thing to note about multiple regression is that, well, there are more X-variables to deal with (obviously)! Multiple regression allows you to consider these different factors as it relates to the main factor you are trying to study (your single Y-variable). As far as your regression equation goes, adding more variables simply adds more terms to your equation, as shown below:

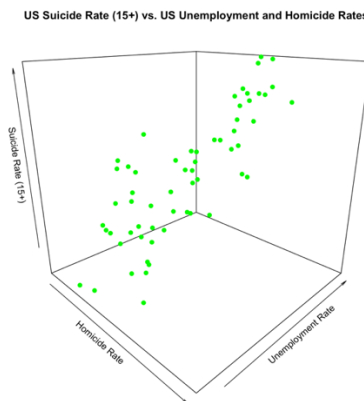
Regression Equation

$$\begin{aligned} \text{US Suicide Rate (15+) per mil} &= 0.28 + 0.901 \text{ US Homicide Rate per million} \\ &+ 0.2080 \text{ Unemployment Rate (\%)} \\ &+ 0.481 \text{ Drug Overdoses per Million} \end{aligned}$$

Simply put, there is an additional term for each factor added. Likewise, in order to predict a Y-value, you need an input for each of the X-variables.

No Scatterplot or Other Form of Visualization...

As nice as multiple regression is for examining your response variable in multiple dimensions, it does come with a major drawback. Unfortunately, it is very difficult to visualize your results of multiple regression other than with tables and equations, as your theoretical scatterplot would most likely exceed three dimensions. If you're curious, it is possible to make a 3D scatterplot if your project contains *only two* X-variables; this can be done using a statistical program more sophisticated than most DataJam teams use, called *R*.



As cool as this may be, it is still NOT advisable to do this as part of a DataJam project, regardless of whether you are a coding wizard. It is hard to truly visualize a 3D plot on a 2D surface or

screen, so do not try to pull this off on a DataJam project. Thus, you will need to stick with the equations and tables from the computer output as your results.

Different Statistics

The last major difference between simple linear regression and multiple regression is that there are different summary statistics for multiple regression. First of all, instead of simply r and R^2 , there are the *multiple-r* and *multiple- R^2* . The **multiple-r** and **multiple- R^2** simply measure the correlation between all of the predictors *collectively* and the response. Since there are more predictors, the multiple-r and multiple- R^2 will be higher than it would be with any individual predictor, as adding more predictors to the model explains more variation in the response and thus helps remove some of the error percentage in R^2 . An example of multiple-r and multiple- R^2 is shown below:

Model Summary

| S | R-sq | R-sq(adj) | R-sq(pred) |
|----------|--------|-----------|------------|
| 0.435189 | 58.71% | 56.50% | 53.18% |

$$\text{Multiple-}R^2 = 58.71\% = 0.5871$$

To find multiple-r, simply take the square root

$$\text{Multiple-r} = \sqrt{R^2} = \sqrt{0.5871}$$

$$\text{Multiple-r} = 0.7662$$

CONCLUDING REMARKS ON SIMPLE AND MULTIPLE REGRESSION

As you have seen, regression and correlation are extremely powerful tools to use within a DataJam project. However, there is one thing you will need to keep in mind as you use these techniques.

It is absolutely imperative you remember that **correlation does not imply**

causation! Because this is an observational study and not a controlled experiment, it is very likely there are confounding variables preventing you from concluding causation. This is one of the most common mistakes made by DataJam students; do not fall into this trap! For more elaboration on this principle, please check out this video from a fellow DataJam mentor, Jackson Filosa:

https://www.youtube.com/watch?v=2W_JNpXD2Pw

Regression Diagnostics and Conditions

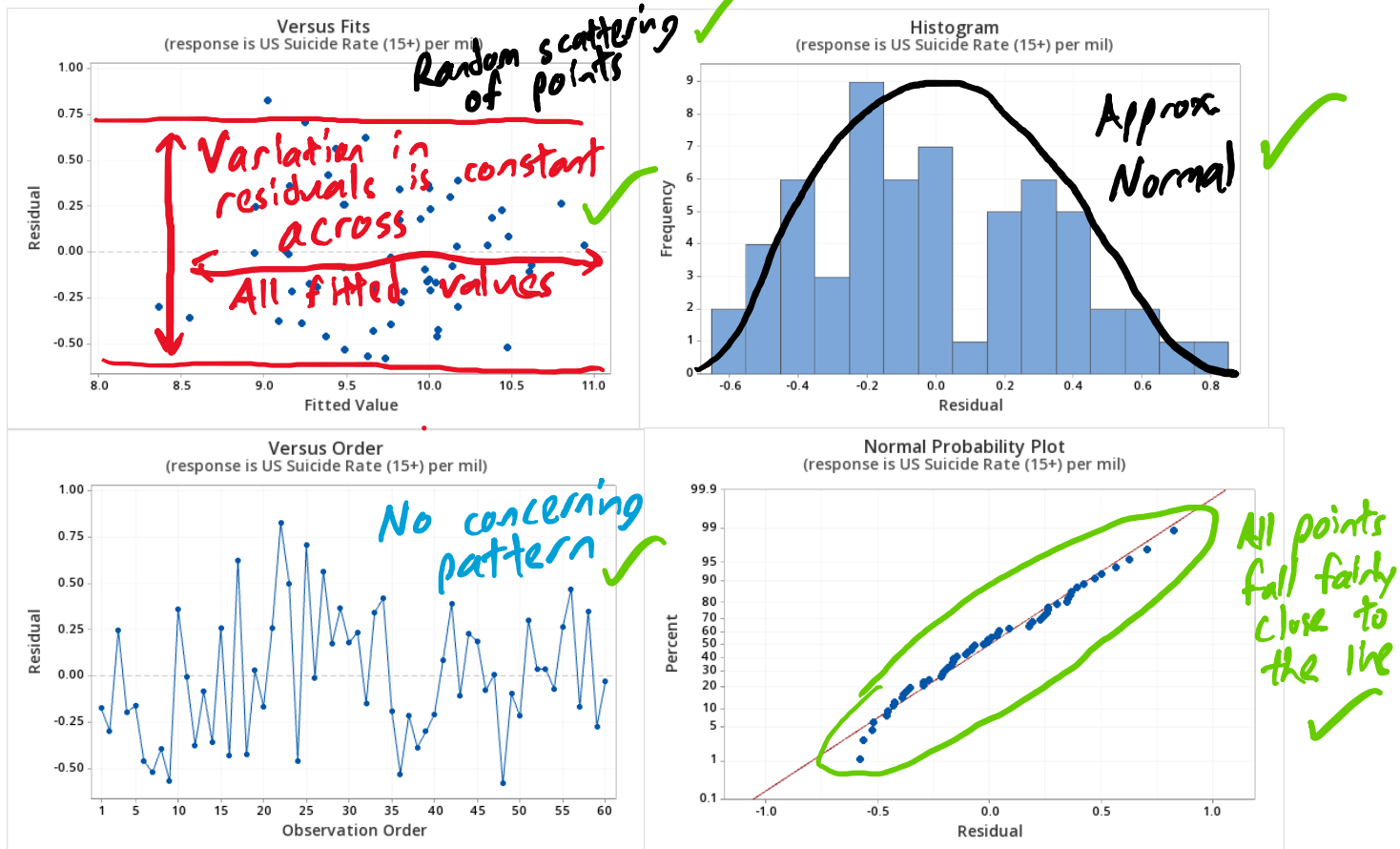
When performing regression in a DataJam project, there are certain conditions that should be met to validate your model. If these conditions are not met, you still should run and display the results of your regression but mention that your model *may not be valid* due to the conditions for your model not being met. The different conditions for regression are listed below.

CONDITIONS FOR A LINEAR MODEL

When running a simple or multiple regression, you almost always assume a linear model. As such, you need to be sure that your model satisfies the underlying conditions for a linear model so that you can safely make this assumption. There are four main conditions that need to be checked and can be done so through examining plots in Minitab or Excel/Google Sheets that correspond to the multiple regression.

- Independence – usually not a problem unless time is a factor in the study; can be a problem if the time series plot shows a sinusoidal pattern or alternates frequently between positive and negative
- Linearity – check the residual plot for a random scattering of points, not a pattern
- Homoscedasticity – residual plot must have a relatively constant spread across all fitted values
- Normality – either the histogram of the residuals is approximately normal or the points on the normal probability plot fall close to the plotted line

Shown below are the results of my study in terms of the conditions to satisfy a linear model.



All the conditions in my project appeared to be satisfied, so I could appropriately apply a linear model.

UNNECESSARY PREDICTORS IN A MODEL

After you've collected your data on numerous factors, you may want to just throw all of them into a multiple regression model and call it a day. While this is a good procedure to begin your analysis, you may have certain predictors that just do not fit well into the model. As such, these predictors need to be removed from the model. There are two main things you need to check for when removing predictors from your full model:

- **Collinearity** – two or more predictors are redundant within the model
- **Insignificance** – a predictor does not affect the response enough to be kept in the model

With collinearity, it is best to look at the VIF (variance inflation factor) within the computer output or a correlation matrix (if VIF is not available). If the VIF is close to or over 10 for a predictor, then that predictor should be removed from the model. Alternatively, a correlation matrix can be examined; two predictors that are highly correlated with each other should have one of them removed to reduce model redundancy.

When dealing with insignificant predictors, it is best to use a model selection technique to determine the best model. If you have seven or less predictors, use *best subsets regression*. This technique allows you to pick the best possible model out of all. Usually, with best subsets regression, you look for the model with the highest **adjusted R²**. The adjusted R² is similar to the regular R², but it applies a penalty to the value for every unnecessary predictor added. Therefore, it is important to look for the model with the highest *adjusted R²* rather than the highest default R². If best subsets regression does not eliminate all significant predictors, then use *forward selection*. This will ensure only significant predictors enter the model. Best subsets regression can be accessed in Minitab by going to Stat → Regression → Regression → Best Subsets. Forward selection can be accessed by going to Stat → Regression → Regression → Fit Regression Model → Stepwise.

Coefficients

| Term | Coef | SE Coef | T-Value | P-Value | VIF |
|--------------------------------|----------|----------|---------|---------|--------|
| Constant | 1.20 | 1.43 | 0.84 | 0.405 | |
| US Homicide Rate per million | 0.824 | 0.183 | 4.51 | 0.000 | 3.70 |
| Unemployment Rate (%) | 0.2292 | 0.0450 | 5.09 | 0.000 | 4.64 |
| Alcohol Poisonings per million | 0.493 | 0.348 | 1.41 | 0.164 | 2.33 |
| Real Alcohol Sales (millions) | 0.000043 | 0.000083 | 0.51 | 0.610 | 2.89 |
| US H1N1 Thousand Deaths | -3.47 | 1.42 | -2.45 | 0.018 | 577.85 |
| US H1N1 Hospitalizations | 0.000164 | 0.000067 | 2.45 | 0.018 | 590.04 |
| Drug Overdoses per Million | 0.276 | 0.173 | 1.60 | 0.117 | 1.92 |
| Seasons | | | | | |
| Spring | 0.647 | 0.154 | 4.19 | 0.000 | 1.99 |
| Summer | 0.465 | 0.173 | 2.69 | 0.010 | 2.49 |
| Winter | 0.404 | 0.242 | 1.67 | 0.101 | 4.88 |
| Pandemic? | | | | | |
| Yes | -0.438 | 0.189 | -2.32 | 0.025 | 2.69 |

Insignificant p-value,
both removed

Coefficients

| Term | Coef | SE Coef | 95% CI | T-Value | P-Value | VIF |
|------------------------------|---------|---------|--------------------|---------|---------|-------|
| Constant | 1.16 | 1.36 | (-1.56, 3.88) | 0.86 | 0.395 | |
| US Homicide Rate per million | 0.779 | 0.181 | (0.415, 1.143) | 4.30 | 0.000 | 3.53 |
| Unemployment Rate (%) | 0.2466 | 0.0449 | (0.1563, 0.3368) | 5.49 | 0.000 | 4.49 |
| US H1N1 Thousand Deaths | 0.554 | 0.291 | (-0.030, 1.137) | 1.91 | 0.062 | 23.68 |
| Drug Overdoses per Million | 0.375 | 0.153 | (0.068, 0.683) | 2.45 | 0.018 | 1.46 |
| H1N1 Thousand Deaths^2 | -0.1187 | 0.0580 | (-0.2352, -0.0022) | -2.05 | 0.046 | 20.02 |
| Seasons | | | | | | |
| Spring | 0.622 | 0.149 | (0.322, 0.921) | 4.17 | 0.000 | 1.80 |
| Summer | 0.435 | 0.160 | (0.113, 0.756) | 2.72 | 0.009 | 2.07 |
| Winter | 0.307 | 0.167 | (-0.028, 0.643) | 1.84 | 0.072 | 2.26 |
| Pandemic? | | | | | | |
| Yes | -0.492 | 0.204 | (-0.902, -0.083) | -2.41 | 0.020 | 3.06 |

Kept via
model
selection

Full Model

Huge VIF,
Hospitalizations
Removed

Final Model

After removing and modifying some of the predictors, I was able to come up with the true best model for the data. H1N1 Deaths has a VIF over 10 only because of the squared term I had to add – consult a mentor if you think you may have a nonlinear term! Additionally, only one of the three “Seasons” terms and one of the H1N1 terms needed to be significant, so this model validated all terms.

Inferential Statistics

Using Inferential Statistics is vital to a DataJam project – it is what you look at to make conclusions of significance! However, depending on your project, there may be different types of inferential statistics you look at. There are two major measures of inference: confidence intervals and significance tests.

CONFIDENCE INTERVALS

A **confidence interval** is an interval estimate in which the true value of the population parameter most likely lies. The interval is constructed with a predetermined confidence level (usually 90%, 95%, or 99%). The most common is a 95% confidence interval, which states that the observer is 95% confident that the true value of the parameter lies within the bounds of the interval. Confidence intervals are not commonly used in a DataJam project, but they can accompany significance tests, the more commonly used measure of inference. Examples of what confidence intervals look like are shown below.

Coefficients

| Term | Coef | SE Coef | 95% CI | T-Value | P-Value | VIF |
|------------------------------|---------|---------|--------------------|---------|---------|-------|
| Constant | 1.16 | 1.36 | (-1.56, 3.88) | 0.86 | 0.395 | |
| US Homicide Rate per million | 0.779 | 0.181 | (0.415, 1.143) | 4.30 | 0.000 | 3.53 |
| Unemployment Rate (%) | 0.2466 | 0.0449 | (0.1563, 0.3368) | 5.49 | 0.000 | 4.49 |
| US H1N1 Thousand Deaths | 0.554 | 0.291 | (-0.030, 1.137) | 1.91 | 0.062 | 23.68 |
| Drug Overdoses per Million | 0.375 | 0.153 | (0.068, 0.683) | 2.45 | 0.018 | 1.46 |
| H1N1 Thousand Deaths^2 | -0.1187 | 0.0580 | (-0.2352, -0.0022) | -2.05 | 0.046 | 20.02 |
| Seasons | | | | | | |
| Spring | 0.622 | 0.149 | (0.322, 0.921) | 4.17 | 0.000 | 1.80 |
| Summer | 0.435 | 0.160 | (0.113, 0.756) | 2.72 | 0.009 | 2.07 |
| Winter | 0.307 | 0.167 | (-0.028, 0.643) | 1.84 | 0.072 | 2.26 |
| Pandemic? | | | | | | |
| Yes | -0.492 | 0.204 | (-0.902, -0.083) | -2.41 | 0.020 | 3.06 |

0 not in the interval, significant

0 in the interval, not significant

These confidence intervals measure the true value of the *slope* of each of these variables and the value of the y-intercept (Coef). Confidence intervals can provide the same information as a significance test; in this case, the slope is significant, or not equal to zero, if zero is not contained within the confidence interval. While this can be inferred from the interval, it is more easily inferred from a significance test.

SIGNIFICANCE TESTS

A **significance test** demonstrates if an observed value differs significantly from a hypothesized value. While the computer does most of the work in Minitab or Excel, it is important to understand the logic these programs are using. The first step of a test for statistical significance is to state the null hypothesis and the alternative hypothesis. The **null hypothesis** usually states that the quantity in question (proportion, mean, slope, etc.) has no relationship/significance in the population. On the other hand, the **alternative hypothesis** states that there is a significant relationship of the statistic with respect to the null hypothesis. A conclusion is made by either rejecting or failing to reject the null hypothesis. In order to make this conclusion, a p-value must be calculated. The **p-value** is the probability, *assuming the*

null hypothesis is true, of obtaining a test statistic as extreme or more extreme as the one calculated in the significance test. As a general rule of thumb, if your p-value is less than 0.05, you may reject the null and conclude statistical significance. Otherwise, fail to reject the null. There are several different types of significance tests, as described below.

- **T-test** – measures the statistical significance of a singular mean of some measured quantity within the population to make conclusions about the hypothesized mean of the population
- **1-proportion Z-test** – measures the statistical significance of a singular proportion of some measured quantity within the population to make conclusions about the hypothesized population proportion
- **2-proportion Z-test** – compares two proportions from two different sample groups to determine if those proportions are significantly different
- **2-sample T-test** – compares two means from two different sample groups to determine if those means are significantly different
- **Linear Regression T-test** – commonly used in regression, determines if the slope of the best-fit line is significant, therefore concluding a relationship between the X and Y variable
- **F-test** – also commonly used in regression, similar to the Linear Regression T-test except uses an F-distribution, which is strictly positive and can be used for multiple regression
- **ANOVA** – stands for **A**nalysis **O**f **V**ariance, used as a table in F-tests to display the significance of the regression

You likely will not need to know how to conduct these tests by hand, as software packages such as Minitab or Excel/Google Sheets do this for you. However, it is important to understand what you are looking at in the software output, so that is why we are going over this. An example of an output you might see is as follows:

| ANOVA | | | | | | | | |
|-----------------|---------------------|-----------------------|---------------|----------------|-----------------------|------------------|--------------------|--------------------|
| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>Significance F</i> | | | |
| Regression | 7 | 16.7861652 | 2.3980236 | 14.0088559 | 4.773E-10 | | | |
| Residual | 52 | 8.901314149 | 0.171179118 | | | | | |
| Total | 59 | 25.68747935 | | | | | | |
| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> | <i>Lower 95%</i> | <i>Upper 95%</i> | <i>Lower 95.0%</i> | <i>Upper 95.0%</i> |
| Intercept | 1.36427424 | 1.438363857 | 0.948490348 | 0.34726724 | -1.522014 | 4.25056247 | -1.522014 | 4.25056247 |
| Homicide | 0.88047131 | 0.146990832 | 5.989974332 | 1.9969E-07 | 0.58551263 | 1.17543 | 0.58551263 | 1.17543 |
| Unemployment | 0.24419765 | 0.038961967 | 6.267590231 | 7.2586E-08 | 0.16601474 | 0.32238055 | 0.16601474 | 0.32238055 |
| Alc. Poisonings | -1.0436709 | 0.777063405 | -1.343096151 | 0.18507376 | -2.6029627 | 0.51562093 | -2.6029627 | 0.51562093 |
| Alc. Sales | -0.0198747 | 0.009631097 | -2.063599583 | 0.0440632 | -0.0392009 | -0.0005485 | -0.0392009 | -0.0005485 |
| H1N1 Deaths | 0.15301734 | 0.271668111 | 0.56325102 | 0.57568479 | -0.3921246 | 0.69815929 | -0.3921246 | 0.69815929 |
| Drug Overdoses | 0.43615031 | 0.187521329 | 2.325870414 | 0.02395957 | 0.05986124 | 0.81243939 | 0.05986124 | 0.81243939 |
| H1N1 Deaths^2 | -0.0632171 | 0.056902894 | -1.110964005 | 0.27169447 | -0.1774011 | 0.05096694 | -0.1774011 | 0.05096694 |

Take-Home Message: When working with significance tests, the most important thing to check is to see if the p-value is under 0.05! If so, your test is statistically significant.

Glossary

- Adjusted-R²** – similar to the multiple-R², but applies a penalty to the R² value if an unnecessary X-variable is added to the model
- Alternative hypothesis** – states that there is a significant relationship of the statistic with respect to the null hypothesis
- Bar Chart** – a graphical display for one categorical variable that depicts the frequency of observations in each category by higher/lower bars
- Bias** – the deviation in the results of your study from the true value of the population parameter
- Boxplot** – a graphical display for one quantitative variable that is best for depicting the center (median) and interquartile range of a distribution
- Categorical Variable** – a type of variable whose values describe a quality or characteristic of observations in the data set
- Cluster Sample** – conducted by first dividing the population into clusters, using a randomizer next to select certain clusters, then interviewing everyone in that cluster
- Collinearity** – Occurs when there is a redundant predictor(s) in the model; indicated by a high VIF (>10); can be fixed by removing the redundant predictor
- Confidence interval** – an interval estimate in which the true value of the population parameter likely lies; can become less confident by narrowing the interval or more confident by widening the interval
- Confounding Variables** – variables that can complicate your results away from what they truly are in the real world
- Continuous Variables** – a type of quantitative variable whose values can exist as fractional numbers of a whole, such as gas mileage
- Control Variables** – Variables that are intentionally held constant throughout the course of a project so that the quantity (or quantities) of interest can be most accurately measured
- Convenience Sample** – conducted by simply interviewing the closest or most willing participants to the researcher; easy to conduct but usually leads to biased results
- Correlation Coefficient (*r*)** – measures both the strength and direction of the correlation between the X and Y variables; ranges from -1 to 1, with values close to 1 and -1 being considered a very strong correlation and values close to 0 being considered little to no correlation
- Descriptive Statistics** – describe the characteristics of your variables given the variables' observations
- Discrete Variable** – a type of quantitative variable whose values can only be strictly whole
- Dotplot** – a graphical display for one quantitative variable that is best for displaying the shape and mode of a distribution; it shows the number of observations in a certain bin by the number of vertically stacked dots
- Extrapolation** – involves using the best-fit line and a variable input to predict the value of the other variable *beyond* the range of the plotted observations
- Histogram** – a graphical display for one quantitative variable that is best for displaying the shape of a distribution; it classifies observations into bins based on values
- Interpolation** – involves using the best fit line and a variable input to predict the value of the other variable *within* the range of the plotted observations
- Interquartile range (IQR)** – the difference between the observation with a value at the 75th percentile within a dataset and the observation with a value at the 25th percentile in a dataset; it is usually best found through the use of a software package or calculator, especially with large datasets.

Kurtosis – a measure of the thickness of the tails in a distribution; values indicative of a normal distribution usually fall within -2 to 2

Mean – a very commonly used measure to indicate the center of a quantitative dataset; it is found by taking the sum of the values of all observations then dividing by the total number of observations

Median – the middle value of a numeric dataset, which can be found by arranging the values of a dataset from least to greatest and then crossing out values alternatively from each end until the middle value is reached or by using technology

Mode – an indicator of the most frequent observation within a variable and/or dataset. It is found by simply identifying the value of the observation that appears most frequently in a dataset.

Multiple regression – involves the use of multiple X-variables to collectively explain variation in a single Y-variable

Multiple- r – measures the correlation between all X-variables collectively and the response

Multiple- R^2 – measures the variation in Y that can be explained by all of the X-variables collectively

Normal Distribution – the most important distribution in statistics, as many tests for statistical significance are based around it; it is a symmetric, unimodal (one peak) distribution, where the mean equals the median

Null hypothesis – usually states that the quantity in question (proportion, mean, slope, etc.) has no relationship/significance in the population

Number of Observations – a commonly used descriptive statistic used to denote how many observations there are under a certain variable

Observations – the data points used to conduct analysis through scatterplots, histograms, boxplots, bar graphs, etc; can contain values for one variable or multiple corresponding variables

Outliers – observations that are significantly removed from the rest of the observations, having either much higher or lower values than the rest of the observations

p -value – the probability, assuming the null hypothesis is true, of obtaining a test statistic as extreme or more extreme as the one calculated in the significance test

Parameter - a statistic that pertains to an entire population

Pie Chart – a graphical display for one categorical variable that displays the proportion of the total observations that each category represents

Quantitative Variable – a variable that describes a measurable quantity as a number

R^2 value – measures the amount, or percent of variation in the Y-variable that is explained by the X-variable; the remaining percent (out of 100%) is considered error; is simply the square of the r -value

Regression Equation – the equation of the best-fit line, and can be used to make predictions about what the Y-variable value would be at a certain X-variable value and vice versa

Residual – the difference between its observed (or actual) value and the value predicted by the line of best fit; this is a measure of error in the accuracy of the best-fit line

Resistant Measures – either are not or very slightly affected by the presence of outliers in a dataset; good examples are median and IQR

Sampling – the process by which a subset of the population is taken and observed for a characteristic of interest.

Scatterplot – uses an xy-coordinate system to plot observations between two quantitative variables, one of which is placed on the y-axis and one of which is placed on the x-axis.

Side-by-side boxplot – shows differences in the distributions of the quantitative variable, as denoted by the boxplots, across the different categories of the categorical variable

Significance test – demonstrates if an experimental or observed value differs significantly from a hypothesized value

Simple Linear Regression – involves fitting a best-fit line to data points involving the Y-variable (also known as the *response* variable) and just one X-variable (also known as the *predictor* or *explanatory* variable)

Simple Random Sample – a sample in which subjects are randomly selected from the entire population and then interviewed; usually yields results with the least amount of bias but is very difficult to execute

Skewed Left Distribution – Mean is less than the median, it is identified as a distribution in which there are a few data points far to the left of the rest of the data; a good example is test scores

Skewed Right Distribution – Mean is greater than the median; it is identified as a distribution in which there are a few data points far to the right of the rest of the data; a good example is income

Skewness – the degree of skew away from a normal distribution; values indicative of a normal distribution usually fall within -0.5 to 0.5

Stacked bar graph – shows the proportions of the categories across the observations of the first categorical variable across the categories of the second categorical variable

Standard Deviation – indicates how closely observations are situated to the mean; a larger value means that observations are located further away from the mean; it is best calculated using technology

Systematic Sample – conducted by the researcher interviewing every k^{th} person they encounter; usually more biased than a simple random sample but less biased than a convenience sample

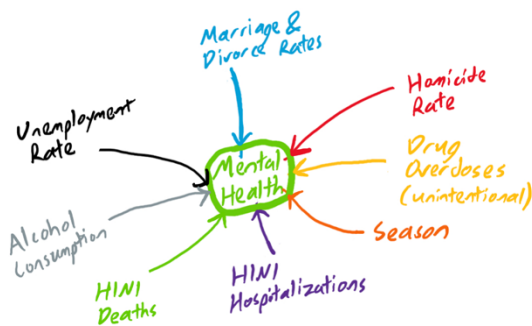
Variable – any characteristic, number, or quantity that can be measured and/or counted

PART TWO: How to Conduct a DataJam Project from Start to Finish

By: Tony Robol

How to Conduct a DataJam Project: A SUMMARY

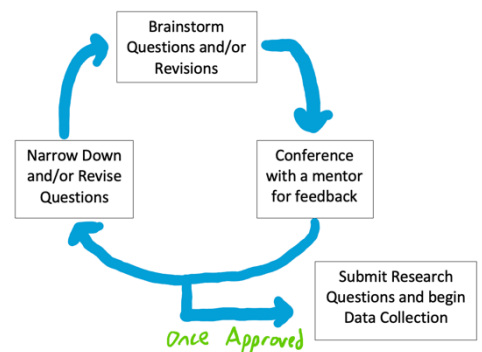
Welcome to the Pittsburgh DataJam competition! In this second part of the resource, you will discover how to accomplish each section of the project from brainstorming ideas and exploring topics to writing your conclusions. Each section will be highlighted by examples from my own project that I conducted over the summer of 2021 with assistance from members from the advisory board itself, Dr. Judy Cameron and Brian Macdonald, as well as fellow mentors Jackson Filosa and Lucas Troy. This section will provide a summary of each of the six steps in the process, **so if you were to read just one section of this entire resource, read this section.**



1. **Exploring Topics** – This section is where you begin to think about what subject you would like to explore as a focus of your project. Begin with many ideas, then narrow it down to a few ideas that you can rank from most exciting to less exciting. One of the most common problems teams face is finding datasets that allow them to answer their questions, so being sure to think of multiple ideas is important in case your top idea is not compatible with available data. Begin

searching for data to see if your top idea is a plausible one to pursue. If not, move down your rankings to see if your other ideas are until you find one that is.

2. **Writing your Research Questions** – This is often the section students struggle the most with, as it is difficult to word your questions in a way that is complex and intriguing enough so that it can fully answered with data on multiple fronts! Writing up research questions is truly a process that takes time, *not* one that can be thought of on a whim and turned in within the span of five minutes. I recommend first brainstorming questions, then conferencing with a mentor for revision. Then, rinse and repeat until you get the OK from both a mentor and the advisory board when you turn your proposal in.



3. **Collecting and Preparing Data** – This section is where you finally start diving into the available data and compiling it! I recommend using either Excel or Google Sheets for

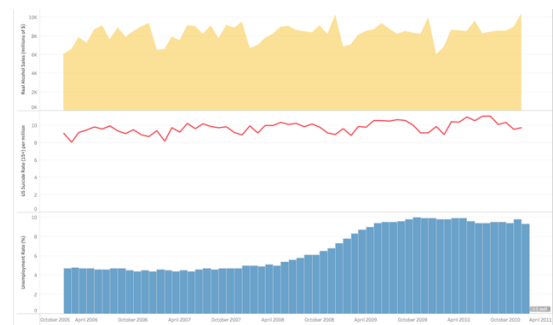
| | D | E | F | G | H | I | J |
|----|-----------------------|-----------------------------------|--------------|------------------------------|-----------------------|--------------------|--------------------------------|
| 1 | US Suicides (Age 15+) | US Suicide Rate (15+) per million | US Homicides | US Homicide Rate per million | Unemployment Rate (%) | Alcohol Poisonings | Alcohol Poisonings per million |
| 2 | 2704 | 9.10476 | 1555 | 5.23591 | 4.7 | 36 | 0.12122 |
| 3 | 2399 | 8.07149 | 1223 | 4.11481 | 4.8 | 35 | 0.11776 |
| 4 | 2737 | 9.20238 | 1356 | 4.55916 | 4.7 | 33 | 0.11095 |
| 5 | 2823 | 9.4843 | 1526 | 5.12683 | 4.7 | 36 | 0.12095 |
| 6 | 2929 | 9.83317 | 1603 | 5.38155 | 4.6 | 24 | 0.08057 |
| 7 | 2861 | 9.59687 | 1607 | 5.39048 | 4.6 | 26 | 0.08721 |
| 8 | 2972 | 9.96046 | 1797 | 6.02252 | 4.7 | 37 | 0.124 |
| 9 | 2801 | 9.37948 | 1587 | 5.31426 | 4.7 | 35 | 0.1172 |
| 10 | 2710 | 9.06632 | 1593 | 5.32939 | 4.5 | 30 | 0.10037 |
| 11 | 2847 | 9.51665 | 1600 | 5.34831 | 4.4 | 32 | 0.10697 |
| 12 | 2676 | 8.938 | 1528 | 5.10361 | 4.5 | 30 | 0.1002 |
| 13 | 2614 | 8.72327 | 1598 | 5.33274 | 4.4 | 47 | 0.15685 |
| 14 | 2820 | 9.40343 | 1518 | 5.06185 | 4.6 | 164 | 0.54687 |
| 15 | 2459 | 8.19406 | 1212 | 4.03871 | 4.5 | 159 | 0.52983 |
| 16 | 2928 | 9.75063 | 1497 | 4.98521 | 4.4 | 111 | 0.36964 |
| 17 | 2776 | 9.23758 | 1487 | 4.94823 | 4.5 | 145 | 0.48251 |
| 18 | 3080 | 10.24181 | 1524 | 5.0677 | 4.4 | 113 | 0.37575 |
| 19 | 2900 | 9.63557 | 1669 | 5.54544 | 4.6 | 127 | 0.42197 |
| 20 | 3073 | 10.20147 | 1760 | 5.84269 | 4.7 | 96 | 0.31869 |
| 21 | 2979 | 9.88058 | 1709 | 5.66832 | 4.6 | 85 | 0.28192 |
| 22 | 2940 | 9.74245 | 1520 | 5.03691 | 4.7 | 129 | 0.42747 |
| 23 | 2975 | 9.85014 | 1528 | 5.05916 | 4.7 | 108 | 0.35758 |
| 24 | 2780 | 9.19655 | 1402 | 4.63797 | 4.7 | 111 | 0.3672 |
| 25 | 2698 | 8.91802 | 1535 | 5.07382 | 5 | 169 | 0.55862 |
| 26 | 3014 | 9.95527 | 1400 | 4.62421 | 5 | 223 | 0.73657 |

compilation of data, as these two applications make it much easier to organize and clean your data when the time comes. One thing to note about this step: **it is often the most time-consuming, so plan accordingly!** Data collection usually is not pretty, and there are data gaps that usually need filled. Use a creative method to complete data filling but be sure it is reasonable and lines up with available data!

- Finalizing your Data** – This is the step where you select the final data you will be using in your analysis, as sometimes data you collect is not useful in analysis within the context of your research questions! Organization and cleaning of the data is a part of this step, but so is working with the data to eliminate confounding variables that would get in the way of proper analysis. The final product of this step should be a spreadsheet in which analysis can be conducted straight from.
- Analyzing your Data** – It is finally time to obtain the results of all that hard work preparing and finalizing your data! For this part, continue to use Google Sheets/Excel, Minitab, and/or Tableau Public to perform statistical analysis on your variables and obtain informative visualizations. If you are having difficulty interpreting these results or using the software, consult a mentor about it!

Analysis of Variance

| Source | DF | Seq SS | Contribution | Adj SS | Adj MS | F-Value | P-Value |
|------------------------------|----|---------|--------------|---------|--------|---------|---------|
| Regression | 9 | 18.7531 | 73.01% | 18.7531 | 2.0837 | 15.02 | 0.00000 |
| US Homicide Rate per million | 1 | 1.0135 | 3.95% | 2.5617 | 2.5617 | 18.47 | 0.00008 |
| Unemployment Rate (%) | 1 | 12.5001 | 48.66% | 4.1742 | 4.1742 | 30.10 | 0.00000 |
| US H1N1 Thousand Deaths | 1 | 1.0138 | 3.95% | 0.5036 | 0.5036 | 3.63 | 0.06245 |
| Drug Overdoses per Million | 1 | 1.1708 | 4.56% | 0.8345 | 0.8345 | 6.02 | 0.01770 |
| H1N1 Thousand Deaths^2 | 1 | 0.2268 | 0.88% | 0.5810 | 0.5810 | 4.19 | 0.04596 |
| Seasons | 3 | 2.0205 | 7.87% | 2.4775 | 0.8258 | 5.95 | 0.00149 |
| Pandemic? | 1 | 0.8077 | 3.14% | 0.8077 | 0.8077 | 5.82 | 0.01951 |
| Error | 50 | 6.9343 | 26.99% | 6.9343 | 0.1387 | | |
| Total | 59 | 25.6875 | 100.00% | | | | |



- Interpreting Data and Writing Conclusions** – This is where you take the results depicted in visualizations or in tables and equations and put it into words that a *general audience* can understand. It is very important that you can convey your results to a wide audience, as this is the basis for data-driven decision-making, which has become a staple in the working world today. Limitations of your project and suggestions for future research are also good things to include in this step.

Exploring Topics

After forming a team for your DataJam project, the first step in your project is to start exploring potential topics for your project. For this part, I only have two suggestions. The first suggestion I have is to brainstorm topics you are passionate about! Passion and motivation go hand in hand. If you choose an uninteresting topic, you likely will not be motivated to work on and hand in a sophisticated project. Second, be creative! This part of the project is where you and your team really get to think outside of the box and home in (hopefully) on something unique. To brainstorm more effectively, here are a few techniques you could use:

- Draw a diagram or web building off of different key ideas
- Use post-it notes to make a “wall of ideas”
- Make individual lists of ideas and compare within the group
- Simply make a list of ideas as a whole group

AFTER CREATING YOUR LIST OF IDEAS

So, you’ve completed brainstorming and compiled your large list of ideas. Now, it is time to get together as a group and narrow things down. However, it is important that **you do NOT simply choose one topic and throw the rest of them out!** This is a common pitfall for students beginning their projects – even I made this mistake! Instead, what you want to do is rank your top 3 or 4 ideas for a project as a group. The main reason for this is *data availability* is a major limiting factor in DataJam projects! Your first idea may be spectacular and extremely creative, but if there is no data available for you to examine, you unfortunately cannot go anywhere with it.

My first idea was to examine COVID-19 data to determine the effect of the pandemic on adult mental health, but this was not possible because sufficient data for COVID-19 had not been released. As a result, I had to go all the way back to the drawing board because I had not thought of an alternative topic. If your foremost topic does not have sufficient data available, **do not get discouraged; this is a very common occurrence!** This problem can also be lightened through ranking your top few ideas, as data will more than likely be available for at least one of them.

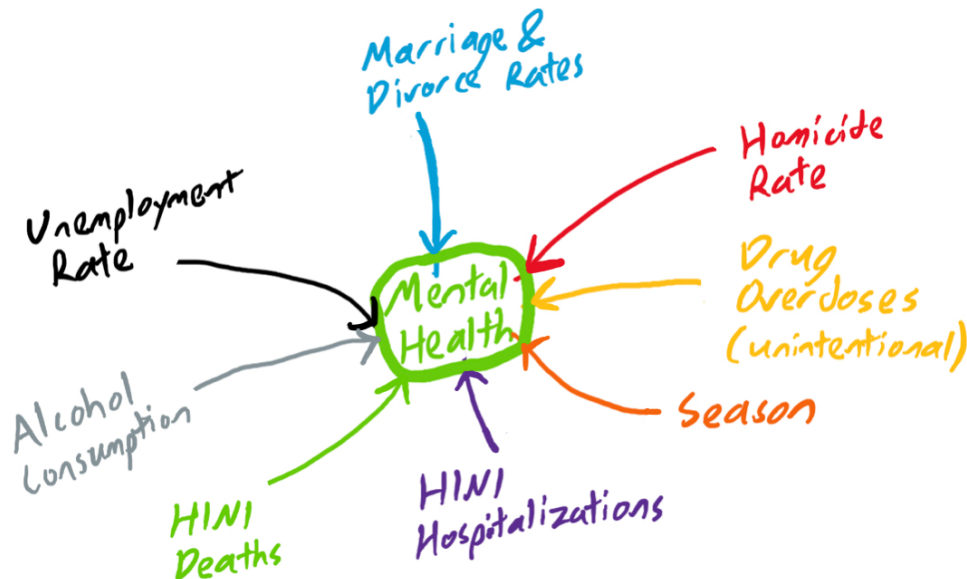
As I had to go back to the drawing board due to not having a second idea, I consulted with a mentor to help think of an alternative topic (Don’t be afraid to use this approach as well!). As a result, I was able to come up with a second project idea: Examining the era of the H1N1 pandemic and Great Recession for increases in adult mental illness, for the purposes of comparing mental illness increases with the COVID-19 pandemic. Luckily, there was data available for this topic, so I was able to go through with it! However, the question remains: *how* did I find this data, and how did I know what to look for? In the next section, we will discuss this.

Take-Home Message: When beginning to brainstorm, do not get fixated on just one idea!

Searching for Data

We have discussed the importance of ensuring there is sufficient data around your idea so you can go through with your project, but the question remains: how do we search for this data, and how do we know what to search for? When brainstorming ideas, it is imperative you search around for data on your ideas to ensure you have a viable project idea (Note: you do not need to formally *collect* all of the data, nor should you in this preliminary phase; just ensure that there is data out there). Consider following this procedure when searching for data to validate your project:

1. Consider your primary (or response) variable you wish to study. Begin by searching for relevant data surrounding this topic from *reliable* (.gov, .edu, or first-party) sources. Here are some of my favorite sources:
 - a. cdc.gov – great for health-related data
 - b. bls.gov – great for economic data
 - c. data.census.gov – great for population and economy data
 - d. nih.gov – another great resource for health and disease data
2. If data is available for your primary variable (in my case, adult mental health* from 2006-2010), begin to think about factors that could influence that variable's values. In my case, I had to think about factors that would influence mental health.

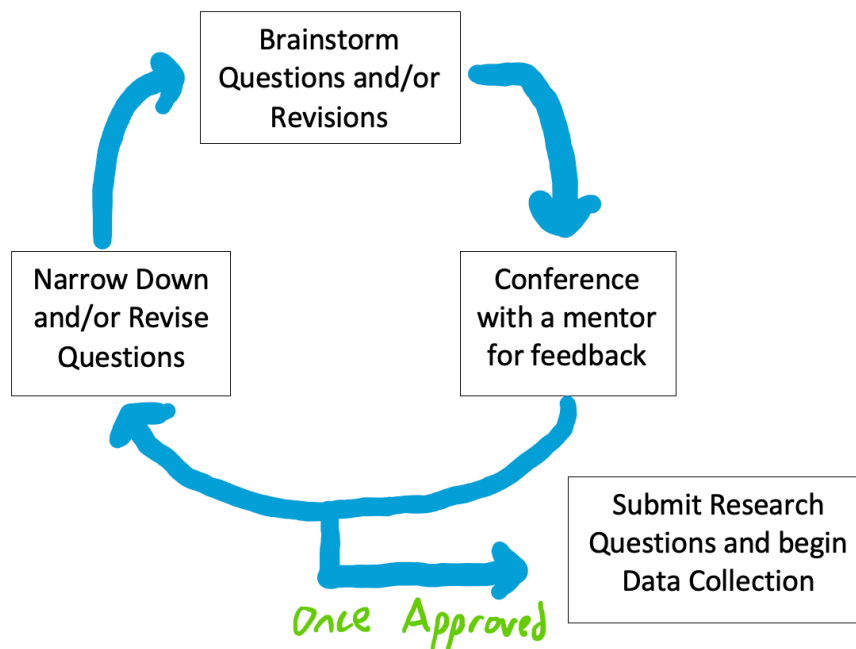


3. After brainstorming a list of factors (the more, the better!) begin to look for data surrounding these factors. You do NOT need to find data on every single factor for your project to be a success! If you can find data on 4-6 of them, that should be plenty.
4. If you have successfully scanned and found data on your primary variable and numerous influencing factors (if applicable), congratulations! You have found a viable topic for your project. Now, it is time to begin thinking about how you will write your research questions.

Take-Home Message: Always search for data from *reliable* sources before deciding on a project!

Writing Your Research Questions

One of the greatest trouble areas in a DataJam project lies in writing specific and thoughtful research questions! As such, it is very important that you understand the proper way to think of, write, and revise your research questions so that you take your project in the right direction all the way through. The best way to design your research questions is to use an *iterative* process, or one that involves repetition of a procedure to obtain more optimal results. In this case, the iterative process that you should look to follow when writing research questions looks like this:



In my case, I had to design research questions for my H1N1 and mental health DataJam project. I first brainstormed, then met with Dr. Judy Cameron and Brian MacDonald, my two mentors, for suggestions, and then revised. After multiple rounds of this process, I came up with the following research questions:

- What factors are correlated most with mental health decline during a pandemic?
- Did the H1N1 pandemic (2009-2010) have a negative effect on mental health?

Once you have submitted your research questions to the DataJam Advisory Board and have had them approved, you may think that you are done. This, however, is NOT the case! Great DataJam projects often have results in either their data collection or analysis that warrant further investigation, and thus new research questions. One of the best qualities in a data scientist is curiosity; never hesitate to dig deeper within your project!

Take-Home Message: When defining research questions, be sure to seek feedback early and often to ensure your research questions end up as insightful as possible! Also, make sure that your research questions are something you really want to know the answer to!

Collecting and Preparing Data

Now that you've finally defined what you will be researching, it is time to start diving into the data! But first, I want to emphasize something very important about this step: **Data collection will more than likely be the step of the process that is the most time-consuming!** You will more than likely run into roadblocks while finding and collecting data, which is why this step is challenging. However, continue to work at it, don't be afraid to seek help, and you will pull through!

PREPARING DATA USING SOFTWARE

When compiling your data, I recommend Microsoft Excel or Google Sheets. Both applications are great for collecting a list of both numeric and nonnumeric data, as well as organizing it using the various built-in tools. When entering the data, it is important that you put your variables in the columns and each data point in the rows. Below is an example of me doing this:

| | D | E | F | G | H | I | J |
|----|-----------------------|-----------------------------------|--------------|------------------------------|-----------------------|--------------------|--------------------------------|
| 1 | US Suicides (Age 15+) | US Suicide Rate (15+) per million | US Homicides | US Homicide Rate per million | Unemployment Rate (%) | Alcohol Poisonings | Alcohol Poisonings per million |
| 2 | 2704 | 9.10476 | 1555 | 5.23591 | 4.7 | 36 | 0.12122 |
| 3 | 2399 | 8.07149 | 1223 | 4.11481 | 4.8 | 35 | 0.11776 |
| 4 | 2737 | 9.20238 | 1356 | 4.55916 | 4.7 | 33 | 0.11095 |
| 5 | 2823 | 9.4843 | 1526 | 5.12683 | 4.7 | 36 | 0.12095 |
| 6 | 2929 | 9.83317 | 1603 | 5.38155 | 4.6 | 24 | 0.08057 |
| 7 | 2861 | 9.59687 | 1607 | 5.39048 | 4.6 | 26 | 0.08721 |
| 8 | 2972 | 9.96046 | 1797 | 6.02252 | 4.7 | 37 | 0.124 |
| 9 | 2801 | 9.37948 | 1587 | 5.31426 | 4.7 | 35 | 0.1172 |
| 10 | 2710 | 9.06632 | 1593 | 5.32939 | 4.5 | 30 | 0.10037 |
| 11 | 2847 | 9.51665 | 1600 | 5.34831 | 4.4 | 32 | 0.10697 |
| 12 | 2676 | 8.938 | 1528 | 5.10361 | 4.5 | 30 | 0.1002 |
| 13 | 2614 | 8.72327 | 1598 | 5.33274 | 4.4 | 47 | 0.15685 |
| 14 | 2820 | 9.40343 | 1518 | 5.06185 | 4.6 | 164 | 0.54687 |
| 15 | 2459 | 8.19406 | 1212 | 4.03871 | 4.5 | 159 | 0.52983 |
| 16 | 2928 | 9.75063 | 1497 | 4.98521 | 4.4 | 111 | 0.36964 |
| 17 | 2776 | 9.23758 | 1487 | 4.94823 | 4.5 | 145 | 0.48251 |
| 18 | 3080 | 10.24181 | 1524 | 5.0677 | 4.4 | 113 | 0.37575 |
| 19 | 2900 | 9.63557 | 1669 | 5.54544 | 4.6 | 127 | 0.42197 |
| 20 | 3073 | 10.20147 | 1760 | 5.84269 | 4.7 | 96 | 0.31869 |
| 21 | 2979 | 9.88058 | 1709 | 5.66832 | 4.6 | 85 | 0.28192 |
| 22 | 2940 | 9.74245 | 1520 | 5.03691 | 4.7 | 129 | 0.42747 |
| 23 | 2975 | 9.85014 | 1528 | 5.05916 | 4.7 | 108 | 0.35758 |
| 24 | 2780 | 9.19655 | 1402 | 4.63797 | 4.7 | 111 | 0.3672 |
| 25 | 2698 | 8.91802 | 1535 | 5.07382 | 5 | 169 | 0.55862 |
| 26 | 3014 | 9.95527 | 1400 | 4.62421 | 5 | 223 | 0.73657 |

VERIFYING ENTRY ACCURACY

Especially when manually typing in data, it is very important that you ensure data was entered accurately, or else your results will be invalid!! Even when copying and pasting data from another source, it is important to check that this procedure was done correctly. To do this, follow this procedure:

1. Have two people collect all the data using Excel/Google Sheets/Other method
2. Cross-verify the compiled data for any differences
3. If there are any differences, have a third person recheck the source to verify which person is correct. If your team only has two members, have both members recheck the source to verify

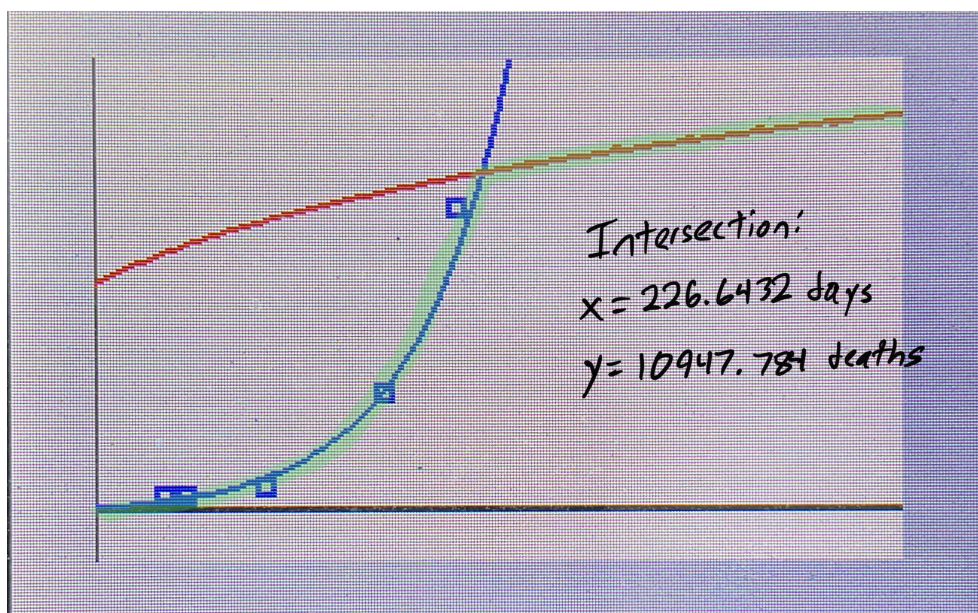
Take-Home Message: Ensuring Accuracy of Entry is imperative to ensure valid results!

Filling in Data Gaps

Filling in missing data within variables is one of the major roadblocks students encounter in data collection. Data may not be available due to legal concerns, differing frequencies of collection (such as data only being collected yearly when a project focuses on monthly), lack of collection to begin with, etc. Nonetheless, if you really want to use this variable in your project, you will need to find a way to fill in the data gaps. This is where it is time to get creative!

In my project, it was tough to find data from 2006-2010 because data collection was not nearly as good as it is today. As such, I was left with numerous data gaps that I had to get creative to fill. The first of which was with the H1N1 pandemic data. Studying this variable was essential to answering my research questions, so I had to find a way to collect monthly data on it. However, not only was data not available every month of the pandemic, but data updates were given in the *middle* of each available month, meaning any new deaths and hospitalizations were split between two months.

The way I corrected for this was by analyzing the pandemic on a *daily* scale so I could frame the data by each calendar month like the rest of my variables. It was tough, but I did this by plotting “Days Since the Start of the H1N1 Pandemic” on the x-axis and “Cumulative Deaths” on the y-axis. I then fit a piecewise curve to get a model I could predict deaths from in the correct monthly frame. Finally, I followed this same procedure for H1N1 hospitalizations.



Yet again, this demonstrates that creativity is one of the best qualities you can have as a young data scientist! You do need to be careful with going too far over the edge, however. While being creative is great, you still need to have viable reasoning behind your method for filling in the gaps. In my case, all my “predicted” points by the curve checked out with the actual data, so I was able to continue with it.

Take-Home Message: When filling in data, be as creative as possible, but make sure your reasoning behind your method for filling is viable!

Finalizing Your Data

Now that you have gone through the long and arduous process of collecting your data, it is time to organize it and prepare it for statistical analysis. This step requires some thinking and the use of the built-in functions in Excel/Google Sheets.

DECIDING WHICH VARIABLES TO USE

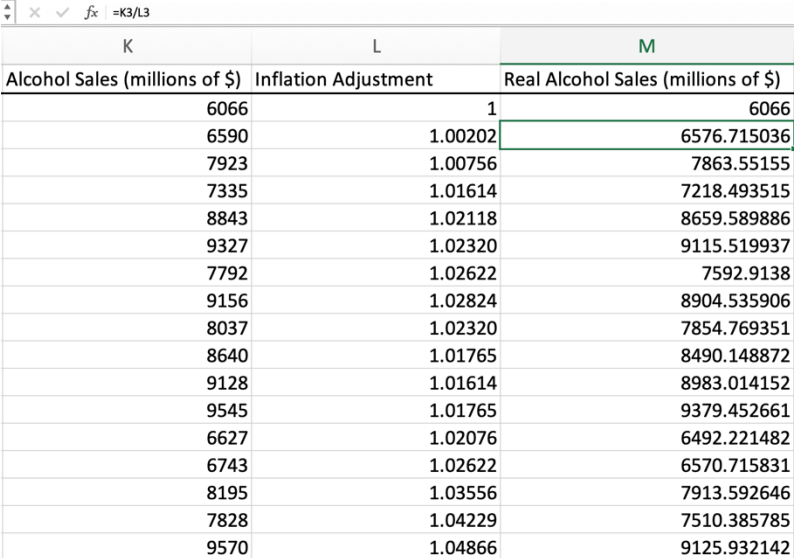
A key step in the organization process is deciding which variables to put through to your analysis. This step involves conferencing with your team and possibly a mentor about which variables will be practically useful in your analysis. Taking time to complete this step is important, as you do not want to put variables that you might have collected data on through to your analysis that might not be viable predictors of your response (Y) variable. Examples of such variables may be those that you were not able to collect and fill sufficient data on or variables whose data simply do not line up with the constraints of the project.

In my project, I had to cut a few variables from my project before beginning analysis. First, I eliminated a variable in which there simply was not sufficient data available: *Antidepressant Prescription Rates*. Then, after consulting with my mentor, I was informed that one of my other variables I wished to examine adult mental illness with, *Marriage and Divorce Rates*, also did not work; this was due to the era in which data was available from (2019) most likely differing from the era in which I was studying (2006-2010).

Overall, I would first cut variables that you simply do not have sufficient data on, then move onto thinking about which of the remaining variables (hopefully all of them) are feasible for analysis within the context of the Y-variable.

FINALIZING YOUR VARIABLES

In many cases, the data that is available is not able to measure the exact thing you are looking for. When this happens, don't get discouraged! Yet again, this is where it is time to get creative! In my case, I wanted to measure alcohol *consumption*, but I could only find data on alcohol *sales*. While I could have just used *sales* as *consumption* data, I would not be taking into account inflation. Inflation in this case is a *confounding variable*, separating the data I have from the quantity I want to measure. Thus, it is best to control for that, and I did so by collecting data on the worth of a dollar in different months, using January as a base month. Finally, I created a new variable, *Real Alcohol Sales*, that adjusts the sales data I had for inflation, giving me consumption data.



| K | L | M |
|--------------------------------|----------------------|-------------------------------------|
| Alcohol Sales (millions of \$) | Inflation Adjustment | Real Alcohol Sales (millions of \$) |
| 6066 | 1 | 6066 |
| 6590 | 1.00202 | 6576.715036 |
| 7923 | 1.00756 | 7863.55155 |
| 7335 | 1.01614 | 7218.493515 |
| 8843 | 1.02118 | 8659.589886 |
| 9327 | 1.02320 | 9115.519937 |
| 7792 | 1.02622 | 7592.9138 |
| 9156 | 1.02824 | 8904.535906 |
| 8037 | 1.02320 | 7854.769351 |
| 8640 | 1.01765 | 8490.148872 |
| 9128 | 1.01614 | 8983.014152 |
| 9545 | 1.01765 | 9379.452661 |
| 6627 | 1.02076 | 6492.221482 |
| 6743 | 1.02622 | 6570.715831 |
| 8195 | 1.03556 | 7913.592646 |
| 7828 | 1.04229 | 7510.385785 |
| 9570 | 1.04866 | 9125.932142 |

FINALIZING YOUR SPREADSHEET

The other thing I would recommend doing before moving into analysis is gathering all your variables that you *will* be using in your analysis and putting in all in one block. This can be on the same or on a separate sheet but be sure to isolate this data from all the unusable data. In my case, I moved all of my data to a new spreadsheet. This process should look like this:

| Month/Year | Month Number | US Population | US Suicide Rate (15+) | US Homeless (Per 100) | US Unemployment Rate (%) | Alcohol Poisonings per million | Real Alcohol Sales (millions of \$) | US H1N1 Thousand Deaths | Drugs Overdoses per Million | H1N1 Thousand Deaths | Seasons | Pandemic? |
|------------|--------------|---------------|-----------------------|-----------------------|--------------------------|--------------------------------|-------------------------------------|-------------------------|-----------------------------|----------------------|---------|-----------|
| 1 | 1 | 286,387,570 | 9.10476 | 5.23591 | 4.7 | 0.12122 | 38.7733719 | 0 | 6.84877 | 0.00000 | Winter | 0 |
| 2 | 2 | 287,243,847 | 8.07149 | 4.11481 | 4.8 | 0.11776 | 40.76148302 | 0 | 6.67857 | 0.00000 | Winter | 0 |
| 3 | 3 | 287,432,085 | 9.20238 | 4.55916 | 4.7 | 0.11095 | 42.36025044 | 0 | 7.39015 | 0.00000 | Winter | 0 |
| 4 | 4 | 287,488,603 | 9.48430 | 5.12683 | 4.7 | 0.12095 | 38.85981711 | 0 | 7.30389 | 0.00000 | Spring | 0 |
| 5 | 5 | 287,488,406 | 9.83317 | 5.38155 | 4.6 | 0.08057 | 37.09163615 | 0 | 6.77676 | 0.00000 | Spring | 0 |
| 6 | 6 | 288,124,821 | 9.59687 | 5.39048 | 4.6 | 0.08721 | 52.74397897 | 0 | 8.22440 | 0.00000 | Spring | 0 |
| 7 | 7 | 288,179,812 | 9.96046 | 5.39048 | 4.7 | 0.12400 | 37.90231315 | 0 | 8.07695 | 0.00000 | Summer | 0 |
| 8 | 8 | 288,179,812 | 9.37948 | 5.31426 | 4.7 | 0.11720 | 40.10182061 | 0 | 7.65829 | 0.00000 | Summer | 0 |
| 9 | 9 | 288,308,360 | 9.06632 | 5.32939 | 4.5 | 0.10037 | 44.836254 | 0 | 7.38353 | 0.00000 | Summer | 0 |
| 10 | 10 | 289,129,196 | 9.51665 | 5.34831 | 4.4 | 0.10697 | 41.13449739 | 0 | 7.30713 | 0.00000 | Fall | 0 |
| 11 | 11 | 289,389,746 | 8.93800 | 5.10361 | 4.5 | 0.10020 | 44.9039619 | 0 | 7.18113 | 0.00000 | Fall | 0 |
| 12 | 12 | 289,389,746 | 8.72327 | 5.33274 | 4.4 | 0.15685 | 28.05310752 | 0 | 7.17484 | 0.00000 | Fall | 0 |
| 13 | 13 | 289,890,446 | 9.40343 | 5.06185 | 4.6 | 0.154687 | 48.1156129 | 0 | 7.82619 | 0.00000 | Winter | 0 |
| 14 | 14 | 289,890,446 | 8.19406 | 4.03871 | 4.5 | 0.152983 | 48.493270509 | 0 | 5.27600 | 0.00000 | Winter | 0 |
| 15 | 15 | 289,890,446 | 9.75063 | 4.98521 | 4.4 | 0.36964 | 11.90323231 | 0 | 8.13218 | 0.00000 | Winter | 0 |
| 16 | 16 | 289,890,446 | 9.23758 | 4.94823 | 4.5 | 0.48251 | 9.326219928 | 0 | 7.77674 | 0.00000 | Spring | 0 |
| 17 | 17 | 289,890,446 | 10.24181 | 5.06770 | 4.4 | 0.37575 | 11.70976968 | 0 | 7.46189 | 0.00000 | Spring | 0 |
| 18 | 18 | 289,890,446 | 9.63557 | 5.45444 | 4.6 | 0.42197 | 10.90121141 | 0 | 7.51574 | 0.00000 | Spring | 0 |
| 19 | 19 | 289,890,446 | 10.20147 | 5.84269 | 4.7 | 0.31869 | 14.7477784 | 0 | 7.62869 | 0.00000 | Summer | 0 |
| 20 | 20 | 289,890,446 | 9.88058 | 5.66832 | 4.6 | 0.28192 | 16.31649099 | 0 | 7.71475 | 0.00000 | Summer | 0 |
| 21 | 21 | 289,890,446 | 9.74245 | 5.03691 | 4.7 | 0.42747 | 10.99480196 | 0 | 7.47252 | 0.00000 | Summer | 0 |
| 22 | 22 | 289,890,446 | 9.85014 | 5.05916 | 4.7 | 0.35758 | 13.14373587 | 0 | 7.36029 | 0.00000 | Fall | 0 |
| 23 | 23 | 289,890,446 | 9.19655 | 4.63797 | 4.7 | 0.36720 | 12.7995443 | 0 | 7.35724 | 0.00000 | Fall | 0 |
| 24 | 24 | 289,890,446 | 8.91802 | 4.57382 | 5.0 | 0.55862 | 8.950691065 | 0 | 8.06853 | 0.00000 | Fall | 0 |
| 25 | 25 | 289,890,446 | 9.95527 | 4.62421 | 5.0 | 0.73657 | 6.788214036 | 0 | 7.84464 | 0.00000 | Winter | 0 |
| 26 | 26 | 289,890,446 | 9.14293 | 4.34702 | 4.9 | 0.56772 | 8.631015547 | 0 | 8.21874 | 0.00000 | Winter | 0 |
| 27 | 27 | 289,890,446 | 10.00379 | 4.75287 | 5.1 | 0.76521 | 6.664844299 | 0 | 8.01820 | 0.00000 | Winter | 0 |
| 28 | 28 | 289,890,446 | 10.00660 | 4.93108 | 5.0 | 0.58672 | 11.04898185 | 0 | 7.89764 | 0.00000 | Spring | 0 |
| 29 | 29 | 289,890,446 | 10.36503 | 4.83833 | 5.4 | 0.53027 | 10.18342579 | 0 | 8.29933 | 0.00000 | Spring | 0 |
| 30 | 30 | 289,890,446 | 10.13010 | 5.18682 | 5.6 | 0.50683 | 11.04898185 | 0 | 7.31289 | 0.00000 | Spring | 0 |
| 31 | 31 | 289,890,446 | 10.24683 | 5.37005 | 5.8 | 0.50313 | 11.52774512 | 0 | 7.45493 | 0.00000 | Summer | 0 |
| 32 | 32 | 289,890,446 | 9.85682 | 5.24383 | 6.1 | 0.52898 | 11.5315637 | 0 | 7.51090 | 0.00000 | Summer | 0 |
| 33 | 33 | 289,890,446 | 10.17024 | 5.15734 | 6.1 | 0.47601 | 12.8148047 | 0 | 7.21240 | 0.00000 | Summer | 0 |
| 34 | 34 | 289,890,446 | 9.81439 | 4.87111 | 6.5 | 0.55764 | 11.65635831 | 0 | 7.55104 | 0.00000 | Fall | 0 |
| 35 | 35 | 289,890,446 | 9.14145 | 4.74936 | 6.8 | 0.55393 | 12.7596484 | 0 | 7.42722 | 0.00000 | Fall | 0 |
| 36 | 36 | 289,890,446 | 8.96499 | 4.56274 | 7.3 | 0.81232 | 8.986617962 | 0 | 7.92339 | 0.00000 | Fall | 0 |

By doing this, it makes it easier to transfer your data directly to a software package such as Minitab or select data to analyze directly in Excel. Making your data easier to read in this way not only makes your life easier in the analysis step but makes it easier for others to understand the data you intend to use.

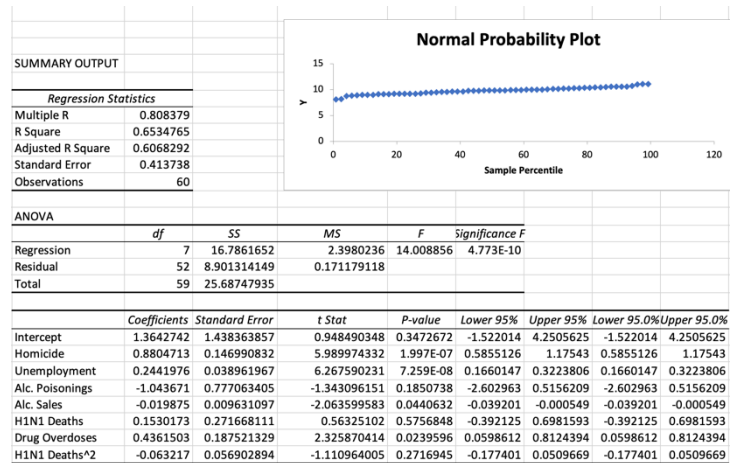
Take-Home Message: Finalizing your variables and organizing them into one place will make the analysis step much easier!

Analyzing Your Data

It is finally time to make graphs, equations, and tables to help you answer your research questions. To do this, there are several different software packages that can be used. These software packages are Minitab, Tableau Public, and Google Sheets/Excel, and all are free to access through the DataJam and online. Below I will discuss the strengths of each of the packages: Excel/Google Sheets, Minitab, and Tableau Public.

EXCEL/GOOGLE SHEETS

Excel/Google Sheets should be used for collecting and organizing your data, but it *can* be used to perform analysis and visualization if Minitab and Tableau are not available. By going to the *Insert* section, you can find a variety of different visualizations for your data; select the appropriate graph based on the type of variable(s) you are working with. Additionally, Excel can perform statistical analysis by going to *Data* → *Data Analysis*. From there, you can see descriptive statistics, perform simple or multiple regression, see a correlation matrix, perform tests of significance, etc. A few examples of these functions are shown below.



| Descriptive Statistics | | | | | | | | |
|------------------------|-----------|--------------------|--------------|--------------------|--------------|--------------------|-----------------|--|
| | Suicide | | Homicide | | Unemployment | | Alc. Poisonings | |
| Mean | 9.7583976 | Mean | 4.82006194 | Mean | 6.7833333 | Mean | 0.461674862 | |
| Standard Error | 0.0851842 | Standard Error | 0.064936482 | Standard Error | 0.2952177 | Standard Error | 0.02703627 | |
| Median | 9.8237772 | Median | 4.854721972 | Median | 5.7 | Median | 0.521875565 | |
| Mode | None | Mode | None | Mode | 4.7 | Mode | None | |
| Standard Deviation | 0.6598341 | Standard Deviation | 0.502995827 | Standard Deviation | 2.2867464 | Standard Deviation | 0.209422044 | |
| Sample Variance | 0.435381 | Sample Variance | 0.253004802 | Sample Variance | 5.229209 | Sample Variance | 0.043857592 | |
| Kurtosis | -0.107422 | Kurtosis | 0.37704251 | Kurtosis | -1.768839 | Kurtosis | -0.581508525 | |
| Skewness | -0.216304 | Skewness | -0.218465029 | Skewness | 0.3254036 | Skewness | -0.49489191 | |
| Range | 3.0096186 | Range | 2.620827682 | Range | 5.6 | Range | 0.73174672 | |
| Minimum | 8.0714854 | Minimum | 3.401695711 | Minimum | 4.4 | Minimum | 0.080572222 | |
| Maximum | 11.081104 | Maximum | 6.022523393 | Maximum | 10 | Maximum | 0.812318943 | |
| Sum | 585.50386 | Sum | 289.2037164 | Sum | 407 | Sum | 27.7004917 | |
| Count | 60 | Count | 60 | Count | 60 | Count | 60 | |

| Correlation Matrix | | | | | | | | |
|--------------------|-----------|--------------|--------------|-----------------|------------|-------------|----------------|---------------|
| | Suicide | Homicide | Unemployment | Alc. Poisonings | Alc. Sales | H1N1 Deaths | Drug Overdoses | H1N1 Deaths^2 |
| Suicide | 1 | | | | | | | |
| Homicide | 0.1986292 | 1 | | | | | | |
| Unemployment | 0.459247 | -0.571024932 | 1 | | | | | |
| Alc. Poisonings | 0.2366153 | -0.573507115 | 0.591489301 | 1 | | | | |
| Alc. Sales | -0.202367 | 0.318143854 | -0.261701657 | -0.834321021 | 1 | | | |
| H1N1 Deaths | -0.006232 | -0.184244344 | 0.359065858 | 0.146699725 | -0.040899 | 1 | | |
| Drug Overdoses | 0.4277791 | -0.275277102 | 0.468024722 | 0.593150388 | -0.390071 | 0.012611035 | 1 | |
| H1N1 Deaths^2 | -0.074052 | -0.125294976 | 0.258205985 | 0.110761902 | -0.033831 | 0.962159741 | -0.029196156 | 1 |

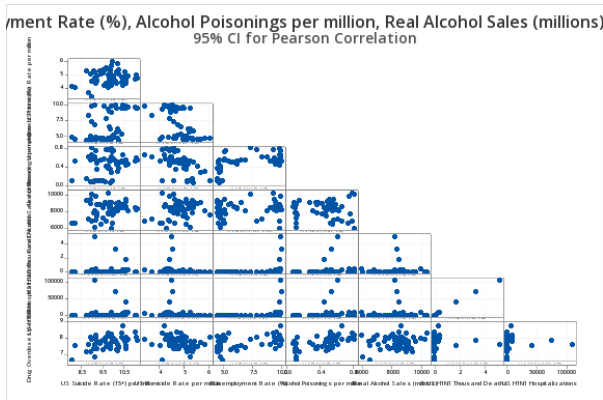
Realistically, you could use Excel or Google Sheets for your entire project. Both applications have all the tools to prepare, organize, and analyze your data. However, I will discuss two additional options in the following sections that give you more advanced options for statistical analysis and visualization respectively: Minitab and Tableau Public.

Take-Home Message: Excel and Google Sheets have tools for analyzing your data as well!

MINITAB

Minitab is a great software that can be used in a DataJam project beyond Excel/Google Sheets, as it truly gives you a lot of options. To name a few things, you can make a ton of different plots, perform multiple regression, test for significance, construct confidence intervals, and perform model selection. Additionally, Minitab includes a spreadsheet where you can copy and paste your data from Google Sheets/Excel. (It is not recommended to directly enter data into Minitab because it is much harder to work with in terms of data entry). Rather than continue to ramble about its benefits, I'd rather show you with pictures.

Correlation Matrix



Correlations

| | US Suicide Rate (15+) per mil | US Homicide Rate per million | Unemployment Rate (%) | Alcohol Poisonings per million | Real Alcohol Sales (millions) |
|--------------------------------|-------------------------------|------------------------------|-----------------------|--------------------------------|-------------------------------|
| US Homicide Rate per million | 0.199 | | | | |
| Unemployment Rate (%) | 0.459 | -0.571 | | | |
| Alcohol Poisonings per million | 0.237 | -0.574 | 0.591 | | |
| Real Alcohol Sales (millions) | 0.271 | 0.216 | 0.167 | 0.025 | |
| US H1N1 Thousand Deaths | -0.006 | -0.184 | 0.359 | 0.147 | -0.006 |
| US H1N1 Hospitalizations | 0.007 | -0.187 | 0.369 | 0.147 | -0.003 |
| Drug Overdoses per Million | 0.428 | -0.275 | 0.468 | 0.593 | 0.114 |

| | US H1N1 Thousand Deaths | US H1N1 Hospitalizations |
|--------------------------------|-------------------------|--------------------------|
| US Homicide Rate per million | | |
| Unemployment Rate (%) | | |
| Alcohol Poisonings per million | | |
| Real Alcohol Sales (millions) | | |
| US H1N1 Thousand Deaths | 0.999 | |
| US H1N1 Hospitalizations | 0.013 | 0.014 |
| Drug Overdoses per Million | | |

Multiple Regression

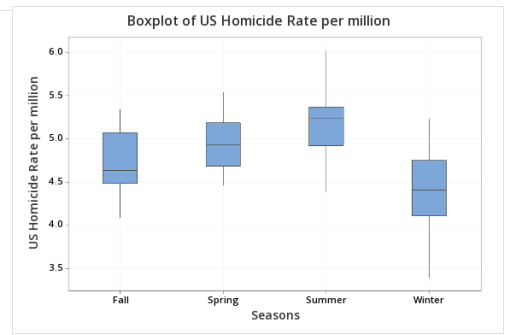
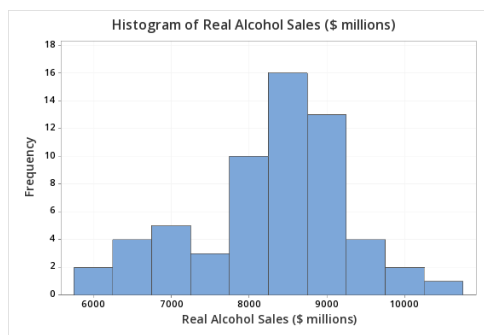
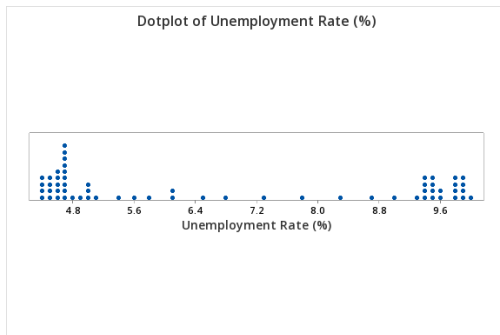
Coefficients

| Term | Coef | SE Coef | 95% CI | T-Value | P-Value | VIF |
|------------------------------|---------|---------|--------------------|---------|---------|-------|
| Constant | 1.16 | 1.36 | (-1.56, 3.88) | 0.86 | 0.395 | |
| US Homicide Rate per million | 0.779 | 0.181 | (0.415, 1.143) | 4.30 | 0.000 | 3.53 |
| Unemployment Rate (%) | 0.2466 | 0.0449 | (0.1563, 0.3368) | 5.49 | 0.000 | 4.49 |
| US H1N1 Thousand Deaths | 0.554 | 0.291 | (-0.030, 1.137) | 1.91 | 0.062 | 23.68 |
| Drug Overdoses per Million | 0.375 | 0.153 | (0.068, 0.683) | 2.45 | 0.018 | 1.46 |
| H1N1 Thousand Deaths^2 | -0.1187 | 0.0580 | (-0.2352, -0.0022) | -2.05 | 0.046 | 20.02 |
| Seasons | | | | | | |
| Spring | 0.622 | 0.149 | (0.322, 0.921) | 4.17 | 0.000 | 1.80 |
| Summer | 0.435 | 0.160 | (0.113, 0.756) | 2.72 | 0.009 | 2.07 |
| Winter | 0.307 | 0.167 | (-0.028, 0.643) | 1.84 | 0.072 | 2.26 |
| Pandemic? | | | | | | |
| Yes | -0.492 | 0.204 | (-0.902, -0.083) | -2.41 | 0.020 | 3.06 |

Analysis of Variance

| Source | DF | Seq SS | Contribution | Adj SS | Adj MS | F-Value | P-Value |
|------------------------------|----|---------|--------------|---------|--------|---------|---------|
| Regression | 9 | 18.7531 | 73.01% | 18.7531 | 2.0837 | 15.02 | 0.00000 |
| US Homicide Rate per million | 1 | 1.0135 | 3.95% | 2.5617 | 2.5617 | 18.47 | 0.00008 |
| Unemployment Rate (%) | 1 | 12.5001 | 48.66% | 4.1742 | 4.1742 | 30.10 | 0.00000 |
| US H1N1 Thousand Deaths | 1 | 1.0138 | 3.95% | 0.5036 | 0.5036 | 3.63 | 0.06245 |
| Drug Overdoses per Million | 1 | 1.1708 | 4.56% | 0.8345 | 0.8345 | 6.02 | 0.01770 |
| H1N1 Thousand Deaths^2 | 1 | 0.2268 | 0.88% | 0.5810 | 0.5810 | 4.19 | 0.04596 |
| Seasons | 3 | 2.0205 | 7.87% | 2.4775 | 0.8258 | 5.95 | 0.00149 |
| Pandemic? | 1 | 0.8077 | 3.14% | 0.8077 | 0.8077 | 5.82 | 0.01951 |
| Error | 50 | 6.9343 | 26.99% | 6.9343 | 0.1387 | | |
| Total | 59 | 25.6875 | 100.00% | | | | |

Simple Plots



For those performing regression: If you begin to form models and are unsure of the best model, Minitab can perform *model selection* to assist you in selecting the best model. To do this, perform *all subsets regression*, detailed in the Part 1 subsection “Unnecessary Predictors in a Model”. The model with the highest adjusted- R^2 is usually the best. Oh also, correlation does not imply causation.

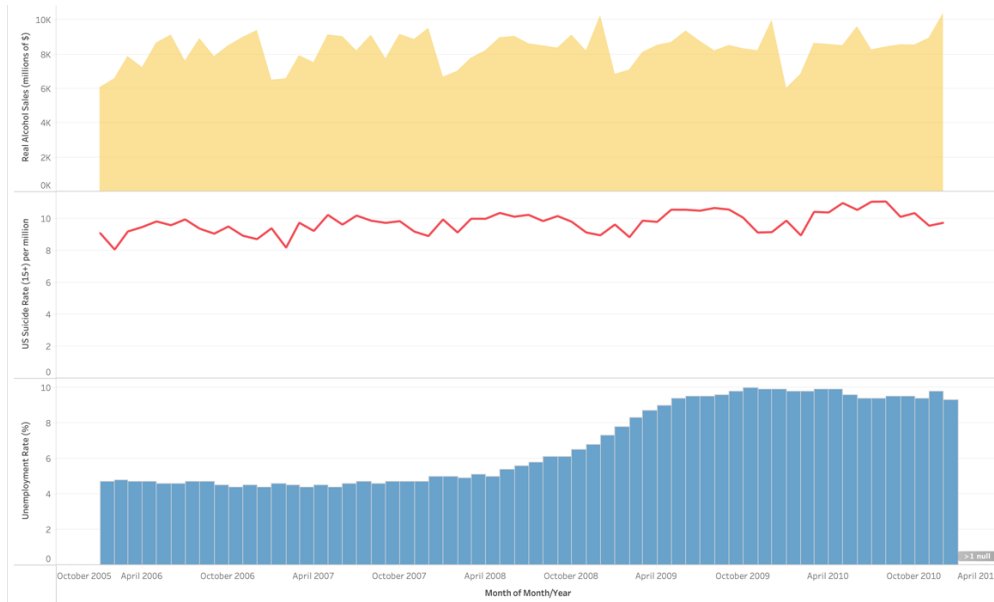
Overall, you can never go wrong with Minitab when performing a statistical analysis, as it has all the tools and then some that you will need during a DataJam project. Most DataJam mentors have experience with Minitab and are available for consultation when it comes to this software package. However, if you do not think you will need Minitab, then Google Sheets or Excel will work just fine for analysis.

Take-Home Message: Minitab gives you a ton of options for statistical analysis in case you will need to undergo an analysis beyond the scope of what Google Sheets or Excel can do!

TABLEAU PUBLIC

One of the best and most popular applications for visualizations in a DataJam project is Tableau (Public). It works similarly to a software offering basic visualizations such as Minitab or Excel/Google Sheets but offers more eye-popping results. It can be a little hard to get used to however, so see a mentor if you get stuck working with the different options! Below are just a few examples of what you can do with this visualization software.

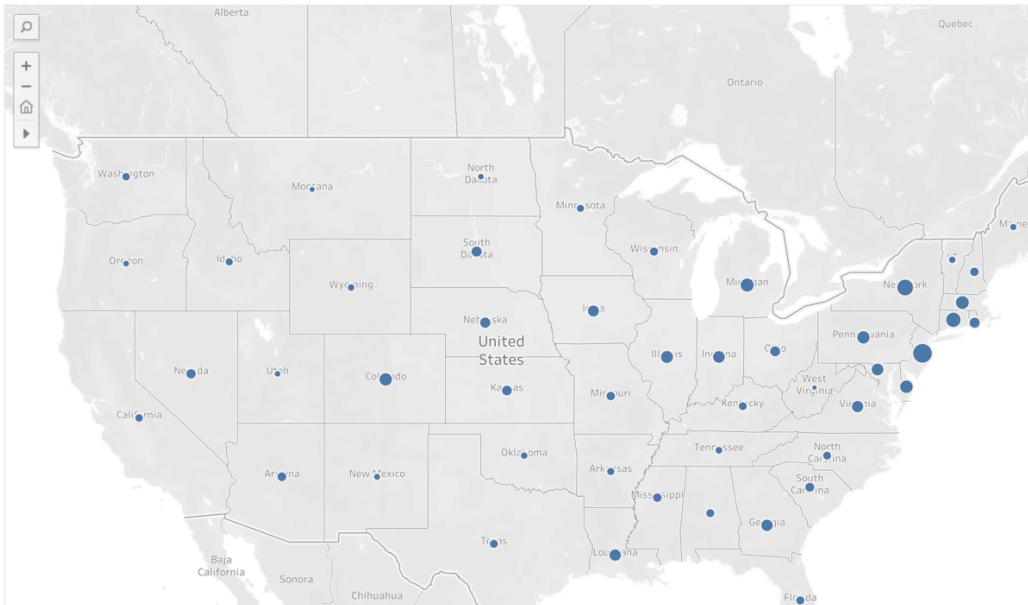
Trends over Time



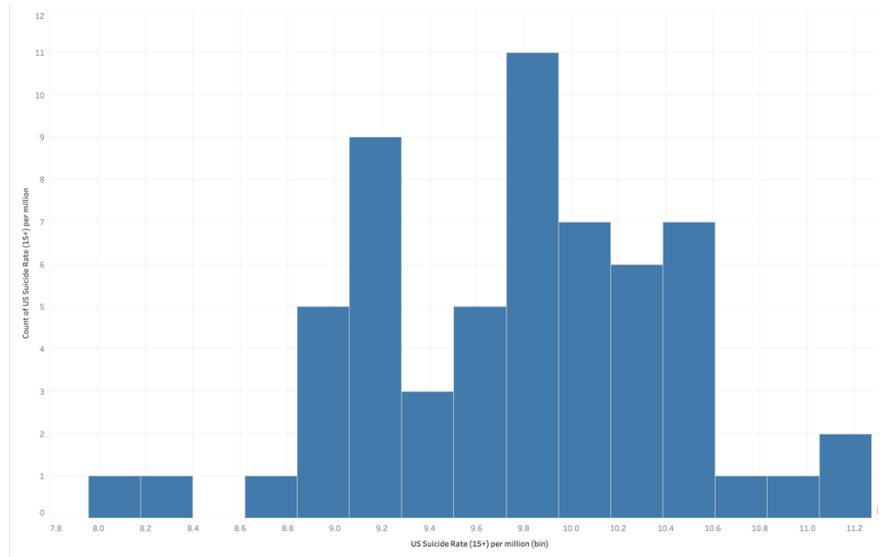
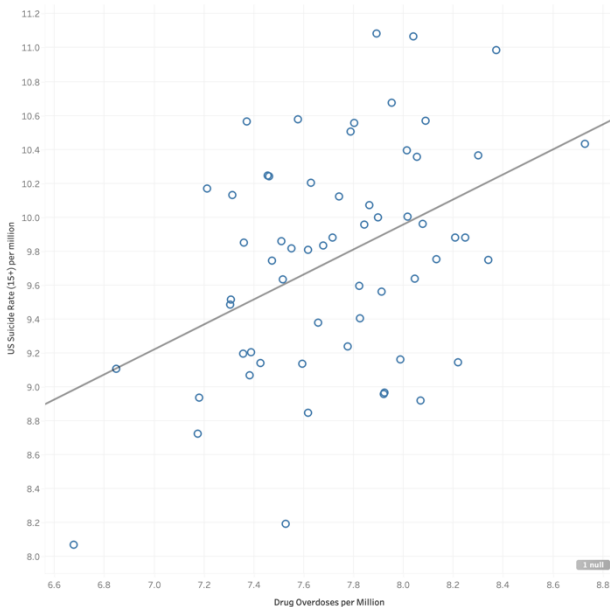
Maps

Proportion of Positive COVID-19 Tests by State by [Tony R.](#)

Sheet 1



Simple Plots



As you can see, Tableau gives you a ton of options when it comes to visualizing your data, so I would recommend Tableau if you would like your graphic to have a bit more flair than it would in Google Sheets/Excel or Minitab. Additionally, they say a picture is worth 1000 words, so having an informative visualization here can go a long way in portraying your results in your final project.

One thing I would like to mention is that when using Tableau Public, it often asks for “measures” when setting up a chart. The default option “SUM” will work just fine, but be sure to place your variables in the “Marks” section as well to ensure every point is being plotted and not just the point that is the summation of all the values. If that does not work, keep at it with a trial-and-error process or ask a mentor!

Take-Home Message: Tableau Public is best used for visualizations, as there are many eye-popping options to choose from!

Interpreting Data and Writing Conclusions

After conducting your analysis and making your visualizations using software, it is time to interpret what the results are saying. In other words, you are taking what the results are saying and putting them into words, allowing anyone to understand the research you have conducted. This is an especially important step, as you want to be sure anyone can understand the significance of your results, not just other data scientists.

Being able to interpret and explain your results to a general audience leads to **data-driven decision-making**, which is how most decisions are starting to be made in the working world today. Data-driven decision-making is becoming so prevalent in today's world for two main reasons, as listed below:

1. Data collection technology is becoming more advanced, making it much easier to collect data from a wide range of sources
2. Numbers don't lie – having relevant data on a product, idea, app, etc. is an ever-reliable way to demonstrate or predict the effectiveness of it or an innovation of it in the future

When interpreting data, it is imperative that you look at the data and think about how it answers the research question(s) you set out to answer. It is in this light that you should frame your interpretations and conclusions. Common things to explain *in light of the research questions* are:

- What's shown in your graphics
- Any outliers or unusual results
 - If applicable, also include an explanation of what could be causing that outlier
- Results of any significance test (proportion z-tests, t-tests, F-tests, etc.) as it relates to your research question(s)
- Regression results
 - Equations and slopes of your variables
 - The model as a whole
 - R and R² values
 - Significance of your predictor variables and the overall regression
 - Technical conditions such as the conditions for a linear fit or regression diagnostics

In my project, I used a multiple regression, so most of my concluding remarks came from regression. Nevertheless, there are a lot of inferences that can be made from other types of analysis such as histograms, bar charts, etcetera, so don't get discouraged if your project does not use a regression! My concluding statement can be found below.

Analysis of Variance

| Source | DF | Seq SS | Contribution | Adj SS | Adj MS | F-Value | P-Value |
|------------------------------|----|---------|--------------|---------|--------|---------|---------|
| Regression | 9 | 18.7531 | 73.01% | 18.7531 | 2.0837 | 15.02 | 0.00000 |
| US Homicide Rate per million | 1 | 1.0135 | 3.95% | 2.5617 | 2.5617 | 18.47 | 0.00008 |
| Unemployment Rate (%) | 1 | 12.5001 | 48.66% | 4.1742 | 4.1742 | 30.10 | 0.00000 |
| US H1N1 Thousand Deaths | 1 | 1.0138 | 3.95% | 0.5036 | 0.5036 | 3.63 | 0.06245 |
| Drug Overdoses per Million | 1 | 1.1708 | 4.56% | 0.8345 | 0.8345 | 6.02 | 0.01770 |
| H1N1 Thousand Deaths^2 | 1 | 0.2268 | 0.88% | 0.5810 | 0.5810 | 4.19 | 0.04596 |
| Seasons | 3 | 2.0205 | 7.87% | 2.4775 | 0.8258 | 5.95 | 0.00149 |
| Pandemic? | 1 | 0.8077 | 3.14% | 0.8077 | 0.8077 | 5.82 | 0.01951 |
| Error | 50 | 6.9343 | 26.99% | 6.9343 | 0.1387 | | |
| Total | 59 | 25.6875 | 100.00% | | | | |

As a result of my multiple regression, several inferences were drawn. First and foremost, the model seemed to do a great job explaining much of the variation in adult mental illness over the years 2006-2010, as the total R² value was 73.01%. This means that 73.01% of the variation in adult mental illness from 2006-2010 can be explained by all seven of the predictors in the model collectively. Based on the R² values obtained, Unemployment Rate (48.66%), Season of the Year (7.87%), and Drug Overdoses (4.56%) were the factors that were most

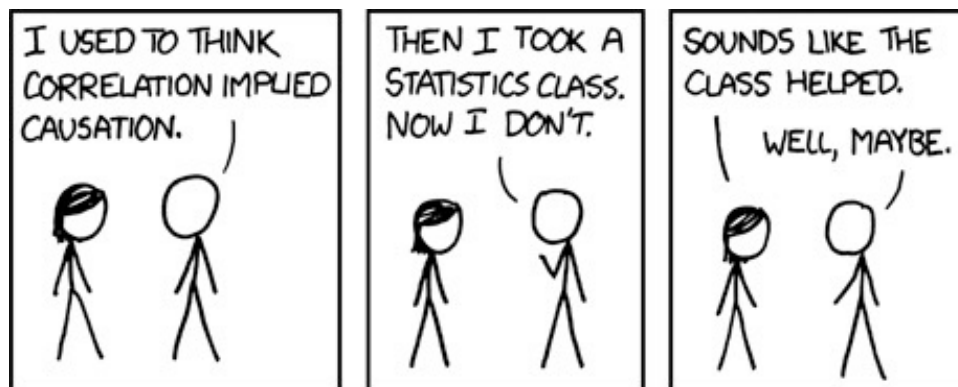
correlated with adult mental illness, measured by suicide rate from 2006-2010. Unemployment rate appears to be especially correlated with adult mental illness, with a partial R^2 of 48.66%, which is much higher than any of the other predictor, or X variables. Therefore, it is by far the variable that is most correlated with adult mental illness.

Oddly enough, the model seemed to show that the effect of a worldwide H1N1 pandemic seemed to decrease adult mental illness. Under the “Coef” column, you can see that the slope coefficients for $H1N1\ Deaths^2$ and $Pandemic?$ were both negative. Additionally, as shown by the graphic to the right, the combination of $H1N1\ Deaths$, $H1N1\ Deaths^2$, as well as the indicator variable $Pandemic?$ were all significant at the 0.05 level. (Note: only one of $H1N1\ Deaths$ or $H1N1\ Deaths^2$ needed to be significant at the 0.05 level for them both to be); this indicates that not only were the slopes negative, but they were also significantly negative.

Coefficients

| Term | Coef | SE Coef | 95% CI | T-Value | P-Value | VIF |
|------------------------------|---------|---------|--------------------|---------|---------|-------|
| Constant | 1.16 | 1.36 | (-1.56, 3.88) | 0.86 | 0.395 | |
| US Homicide Rate per million | 0.779 | 0.181 | (0.415, 1.143) | 4.30 | 0.000 | 3.53 |
| Unemployment Rate (%) | 0.2466 | 0.0449 | (0.1563, 0.3368) | 5.49 | 0.000 | 4.49 |
| US H1N1 Thousand Deaths | 0.554 | 0.291 | (-0.030, 1.137) | 1.91 | 0.062 | 23.68 |
| Drug Overdoses per Million | 0.375 | 0.153 | (0.068, 0.683) | 2.45 | 0.018 | 1.46 |
| H1N1 Thousand Deaths^2 | -0.1187 | 0.0580 | (-0.2352, -0.0022) | -2.05 | 0.046 | 20.02 |
| Seasons | | | | | | |
| Spring | 0.622 | 0.149 | (0.322, 0.921) | 4.17 | 0.000 | 1.80 |
| Summer | 0.435 | 0.160 | (0.113, 0.756) | 2.72 | 0.009 | 2.07 |
| Winter | 0.307 | 0.167 | (-0.028, 0.643) | 1.84 | 0.072 | 2.26 |
| Pandemic? | | | | | | |
| Yes | -0.492 | 0.204 | (-0.902, -0.083) | -2.41 | 0.020 | 3.06 |

As you can see above, when writing my conclusions, I made sure to base all of my answers around the two research questions that I wrote at the beginning of the project. After all, what is the point of writing research questions if you don't intend to answer them! Finally, I want to address one last pitfall that is so common for students to fall into when writing their conclusions and it is that **correlation does not imply causation!** Just because a variable or group of variables might be highly correlated with the response does not mean that those variables *cause* the effect seen in the response data! This is due to potential confounding variables that a DataJam project cannot possibly account for. Additionally, a correlation is not directional, so it can tell you one variable is leading to another when it could easily be the other way around. Instead, simply say that these variables are *highly correlated* with each other.



Take-Home Messages:

1. Be sure to frame your interpretations so that they fully answer your research questions!
2. Correlation does not imply causation!

Limitations and Suggestions for Future Research

The last things you should discuss in your concluding remarks are limitations of your project, as well as suggestions for future research that can build off your project. To discuss the former of the two, every project has its limitations. Yes, even those projects conducted by specialized researchers, in a formal lab, using million-dollar equipment. Likewise, don't be afraid to acknowledge these limitations – it is an important part of the project! These limitations likewise bridge into potential suggestions for future research, which are intended to cover those limitations, build off your project, or both. Successfully completing this step is a hallmark of a great DataJam project – it really shows you understand your project in and out and where it can go in the future.

A few common limiting factors within a DataJam project are time and unavailability of data. While these two principles are good to state as limiting factors, keep in mind that almost every other group who completes this section will also state these two factors. This can get boring for the judges, so this is where it is time to get – yet again – creative! After doing some thinking, some of the other limiting factors I could think of for my project besides time and data availability were as follows:

- Conflicts in team schedules limiting the possible number of meetings
- Possible errors in source data collection/having to rely on third-party sources for data

It is very important to know that most research usually builds off previous research, even in the professional field. As such, it is important to provide suggestions for future research so that fellow researchers have several ideas already in hand should they choose to build off your project.

In my project, I wished to study the effects of the global H1N1 pandemic on mental health, seeing that the COVID-19 pandemic seemingly had a severe increase in adult mental illness. However, since I was unable to find complete data on the COVID-19 pandemic, a suggestion for future research would be for another group to use complete COVID-19 data once the pandemic is through to analyze this effect (voila, a limitation that can be covered through future research). Since the H1N1 pandemic had seemingly no increase in mental illness, it could be that it was either not severe enough or the effects of quarantine were what really led to this increase. As you can see, it is good to think about the results of your own research when developing suggestions for future research.

Even this resource is an example of research, and as such I developed suggestions for its future research. Below you will find these ideas:



SUGGESTIONS FOR FUTURE RESEARCH

- Get feedback from high schoolers in the fall
- Write the manual in different languages?
- Collaborate with Jackson Filosa, a fellow CRF recipient, to create video resource about my project

The graphic also includes a diagram with a central cloud labeled 'FEEDBACK' and arrows pointing to icons for 'IDEA', 'RESPONSE', 'OPINION', 'SURVEY', 'ADVICE', and 'COMMENT'. The top right of the graphic shows a camera lens focusing on a globe.

Take-Home Message: Providing Limitations and Suggestions for Future Research are hallmarks of a sophisticated DataJam project, as it shows you understand your project very well!