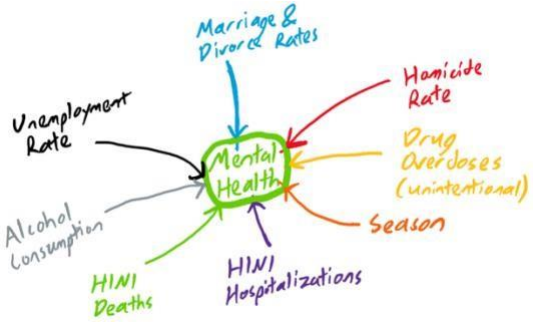
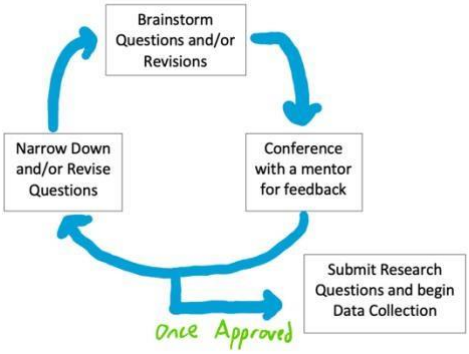
How to Conduct a DataJam Project

This **brief overview** will provide a summary of each of the six steps in the **process of doing a DataJam Project.**



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 Question & Hypothesis** – This is often the section students struggle the most with, as it can be difficult to word your questions in a way that it will allow you to fully explore the topic you are interested in. Writing up your research question and formulating your hypothesis is truly a process that takes time. We recommend first brainstorming questions, then conferencing with a mentor for revision. Then, rinse and repeat until these are worded in a way that they will allow you to do the analyses you would like to do and you get the OK from 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! we 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

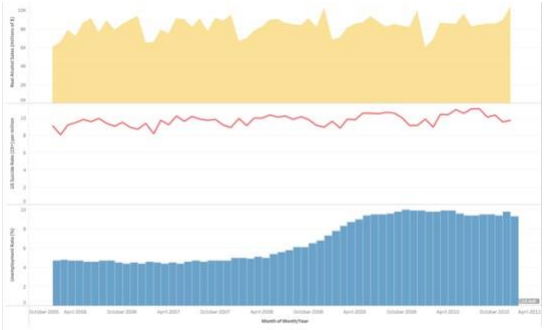
compilation of data, as these two applications make it much easier to organize and clean your data when the time comes. One thing to note about this step: **it is often the most time-consuming, so plan accordingly!** Data collection usually is not pretty, and there are data gaps that usually need to be filled. Think hard and consult a mentor to figure out if there is a way to fill in missing data.

4. **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 and compile 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 from which analyses can be conducted.

5. **Analyzing your Data** – It is finally time to obtain the results of all that hard work preparing and finalizing your data! For this part, you might 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! Analyzing data is often an iterative process where the results of one analysis lead you to realize it would be good to do another analysis.

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%				



6. **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.

For lots more information on how to undertake a DataJam project, look at the DataJam manual prepared by Tony Robol, a DataJam mentor.
It is located on the Resources page.