MODULE #4: Finding Correlations in Datasets

Notes & Considerations:

- Feel free to adapt the timeline and materials to your individual course and student needs.
- You may edit all of the instruction sheets, activity sheets and project rubrics for your individual course and student needs.
- Ideas for extensions are listed under applicable activities.
- Email Lee Cristofano cristofano.lee@bphawks.org or Emily Smoller smoller.emily@bphawks.org with any questions.

Need to Know Information for Teachers (in order to complete the module):

- How to create a scatter plot
- How to add in the trendline and r-squared value
- How to determine whether or not there is a correlation between two variables (using r-squared)

Enduring Understanding: Variables can show positive or negative correlations to each other. What does this imply and what information can be gained from comparing two variables against each other?

Big Question: Are there relationships between variables, and how can we show them? How "strong" are these relationships to each other?

Why? There are times when we notice that a change in one factor (variable) might have some direct effect, or cause some kind of change, on another factor (variable). What does this suggest about the relationship among variables?

Course Objectives:

- To allow students to explore the relationship between two (or more) variables to see if a change in one variable has a measurable effect on another.
- To utilize basic spreadsheet functions and built-in statistical measures to quantify these relationships.

Student Objective:

- To be able to generate data, obtain data, and create a basic scatter plot of the data to conduct basic curve-fitting analysis and measure a statistical quantity called R^2 ("R-squared", the "correlation coefficient", or "coefficient of determination").

  Note: Our purpose here is not to teach statistics as a math course might cover, but rather to apply and interpret the meaning of the R^2 value. It might be appropriate to collaborate with a statistics teacher for these activities.

PA Standards:

PA Standard Area: 2.4 Measurement, Data & Probability

- CC.2.4.HS.B.1 - Summarize, represent, and interpret data on a single count or measurement variable.
- CC.2.4.HS.B.2 - Summarize, represent, and interpret data on two categorical and quantitative variables.

- CC.2.4.HS.B.3 - Analyze linear models to make interpretations based on the data.
- CC.2.5.HS.B.4 - Recognize and evaluate random processes underlying statistical experiments.
- CC.2.4.HS.B.5 - Make inferences and justify conclusions based on sample surveys, experiments and observational studies.
- CC.2.4.HS.B.6 - Use the concepts of independence and conditional probability to interpret data.

PA Standard Area: Computer Science - 3A.DA Data Analysis

- 3A.DA.11 - Create interactive data visualizations using software tools to help others better understand real-world phenomena.

PA Standard Area: Computer Science - 3B.DA Data Analysis

- 3B.DA.05 - Use data analysis tools and techniques to identify patterns in data representing complex systems.
- 3B.DA.06 - Select data collection tools and techniques to generate data sets that support a claim or communicate information.
- 3B.DA.07 - Evaluate the ability of models and simulations to test and support the refinement of hypotheses.

PA Standard Area: 15.3 Communication

- 15.3.12.C - Create a research project based upon defined parameters.
- 15.3.12.G - Employ appropriate presentation skills to lead discussions and team activities.
- 15.3.12.W - Collaborate via electronic communication with peers, educators, and/or professionals to meet organizational goals.

PA Standard Area: English Language Arts

- CC.1.4.11-112.A - Write informative/ explanatory texts to examine and convey complex ideas, concepts, and information clearly and accurately.
- CC.1.4.11-12.D - Organize ideas, concepts, and information to make important connections and distinctions; use appropriate and varied transitions to link the major sections of the text; include formatting when useful to aiding comprehension; provide a concluding statement or section.
- CC.1.4.11-12.F - Demonstrate a grade-appropriate command of the conventions of standard English grammar, usage, capitalization, punctuation, and spelling.
- CC.1.4.11-12.G - Write arguments to support claims in an analysis of substantive topics.
- CC.1.4.11-12.J - Create organization that establishes clear relationships among claim(s), counterclaims, reasons, and evidence; Use words, phrases, and clauses to link the major sections of the text, create cohesion, and clarify the relationships between claim(s) and reasons, between reasons and evidence, and between claim(s) and counterclaims; provide a concluding statement or section that follows from and supports the argument presented.
- CC.1.4.11-12.U - Use technology, including the Internet, to produce, publish, and update individual or shared writing products, taking advantage of technology's capacity to link to other information and to display information flexibly and dynamically.
- CC.1.4.11-12.V - Conduct short as well as more sustained research projects to answer a question (including a self-generated question) or solve a problem; narrow or broaden the inquiry when appropriate; synthesize multiple sources on the subject, demonstrating understanding of the subject under investigation.
- CC.1.4.11-12.W - Gather relevant information from multiple authoritative print and digital sources, using advanced searches effectively; assess the usefulness of each source in answering the research question; integrate information into the text selectively to maintain the flow of ideas, avoiding plagiarism and following a standard format for citation.
- CC.1.5.11-12.A - Initiate and participate effectively in a range of collaborative discussions on grades level topics, texts, and issues, building on others' ideas and expressing their own clearly and persuasively.

- CC.1.5.11-12.D - Present information, findings, and supporting evidence clearly, concisely, and logically such that listeners can follow the line of reasoning; ensure that the presentation is appropriate to purpose, audience, and task.
- CC.1.5.11-12.F - Make strategic use of digital media in presentations to add interest and enhance understanding of findings, reasoning, and evidence.
- CC.1.5.11-2.G - Demonstrate command of the conventions of standard English when speaking based on grade 9-10 level and content.

Common Core Standards:

Summarize, represent, and interpret data on a single count or measurement variable

CCSS.MATH.CONTENT.HSS.ID.A.1

Represent data with plots on the real number line (dot plots, histograms, and box plots).

CCSS.MATH.CONTENT.HSS.ID.A.2

Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, standard deviation) of two or more different data sets.

CCSS.MATH.CONTENT.HSS.ID.A.4

Use the mean and standard deviation of a data set to fit it to a normal distribution and to estimate population percentages. Recognize that there are data sets for which such a procedure is not appropriate. Use calculators, spreadsheets, and tables to estimate areas under the normal curve.

Summarize, represent, and interpret data on two categorical and quantitative variables

CCSS.MATH.CONTENT.HSS.ID.B.6

Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.

CCSS.MATH.CONTENT.HSS.ID.B.6.C

Fit a linear function for a scatter plot that suggests a linear association.

Interpret linear models

CCSS.MATH.CONTENT.HSS.ID.C.8

Compute (using technology) and interpret the correlation coefficient of a linear fit.

CCSS.MATH.CONTENT.HSS.ID.C.9

Distinguish between correlation and causation.

Understand and evaluate random processes underlying statistical experiments

CCSS.MATH.CONTENT.HSS.IC.A.1

Understand statistics as a process for making inferences about population parameters based on a random sample from that population.

CCSS.MATH.CONTENT.HSS.IC.A.2

Decide if a specified model is consistent with results from a given data-generating process, e.g., using simulation. For example, a model says a spinning coin falls heads up with probability 0.5. Would a result of 5 tails in a row cause you to question the model?

Make inferences and justify conclusions from sample surveys, experiments, and observational studies

CCSS.MATH.CONTENT.HSS.IC.B.3

Recognize the purposes of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each.

CCSS.MATH.CONTENT.HSS.IC.B.4

Use data from a sample survey to estimate a population mean or proportion; develop a margin of error through the use of simulation models for random sampling.

CCSS.MATH.CONTENT.HSS.IC.B.5
Use data from a randomized experiment to compare two treatments; use simulations to decide if differences between parameters are significant.
CCSS.MATH.CONTENT.HSS.IC.B.6
Evaluate reports based on data.

Materials:

- Data Sets:
  - Western PA Regional Data Center
  - PA Department of Education Data & Reporting
  - Allegheny County Data
  - US Census Data
  - Data USA
- R-Squared and Correlation Resources:
  - Creating Scatter Plots, Adding the Trendline and R-squared in Google Sheets
  - Pitt Student Video: Pitt Data Diaries #2 - Why Correlation Is NOT Causation
  - Pitt Student Video: Pitt Data Diaries #3 - Analyzing Data! Correlations in Excel
  - Pitt Student Video: Correlation Matrices
  - Article & Video: Investopedia
  - Article: How to Interpret R-Squared and Goodness of Fit
  - Article & Video: How to R-Squared in Regression Analysis
- Module #4: Circumference vs. Diameter Record Sheet
- Google Sheets or Microsoft Excel
- Google Slides, PowerPoint, Canva, Prezi or any other slide design software
- Correlation Mini Project Rubric:
  - Correlation Mini-Project (with points)
  - Correlation Mini Project (without points)
- Google Slides template

Teacher Prerequisites:

- Create a scatter plot
- Add the trend line (using Google Sheets or Excel)
- Calculate R-squared (using Google Sheets or Excel)
- At what point is there a strong correlation between the variables

---

Activity 1: Engage students to generate some numerical data

- Draw some circles on paper, whiteboards, sidewalk chalk, etc.

- Find some round objects, for example coffee cups, garbage cans, tree stumps, etc.
- Collect two measurements of these circles: diameter and circumference.
- Students should record all measurements on a sheet of paper. Module #4: Circumference vs. Diameter Record Sheet
- Students can create a class spreadsheet (Google Sheets or Excel) of these two measurements (in consistent units, for instance all measurements in centimeters).
- Enter all of the circumference measurements in one column and the respective diameter measurements in the column to the right
- Now, create a scatter plot of this data, plotting Circumference vs. Diameter
- Add the trendline and R-squared value to the graph
- **Video Resources**:
  - Creating Scatter Plots, Adding the Trendline and R-squared in Google Sheets
  - Pitt Data Diaries #3 - Analyzing Data! Correlations in Excel

Class discussion: Look at this graph. A teacher-led discussion could include points such as:

- *Does this plot of the data suggest there is a relationship between variables?* (yes - not random)
- *What is the shape of the graph* (linear)
- *What is the slope of the graph?* (positive) (you might even measure the slope if appropriate!)
- *What happens as d increases/decreases?* (C increases/decreases as well)
- *What happens as C increases/decreases?* (d increases/decreases as well)
- *Does this suggest a correlation between the variables?* (yes - discuss whether this is a positive/negative correlation)
- *If so, how strong is the relationship?* (check the R^2 value and discuss its meaning - a value of $R^2 = 1$ shows strong positive correlation and $R^2 = 0$ shows no correlation between the variables. You would typically look for an R-squared value of at least 0.700 to show correlation. Discuss the significance of R^2 values in between these extremes.

Evaluation - award points for successfully:

- Recording the circumference and diameters
- Creating a scatter plot
- Adding in trend line and R-squared value to the scatter plot
- Determining if there is a strong correlation
- Participating in the class discussion

Activity 2: Teacher led discussion:

- Correlation versus causation

    - Video Resource: Pitt Data Diaries #2 - Why Correlation Is NOT Causation

- *What other examples can we think of that might show variables that show correlation?*
- *What other data might we generate/collect to explore this idea further.*

This activity is largely teacher-driven depending on time and interest. You may complete one or both of the following options.

A. **Spreadsheet Analysis:** Take a look at a spreadsheet of interest with multiple columns (variables) of data that would allow for multiple graphs and correlations to be investigated.

   Evaluation - award points for successfully:

   - Creating scatter plots
   - Adding in trend line and R-squared value
   - Determining if there is a strong correlation

B. **Create a Survey:** A survey (Google Form) might be used to collect some data about your students, their families and faculty: Age, shoe size, height, resting heart rate, arm span, forearm length, etc. (This may be anonymously collected, to determine if there are any correlations.) Feel free to create your own study with different variables to collect and analyze!

   Evaluation - award points for successfully:

   - Creating a Google Form survey
   - Collecting data
   - Creating scatter plots
   - Adding in trend line and R-squared value
   - Determining if there is a strong correlation

Activity 3: Student Choice for Finding Correlations

- This activity is largely dependent on the teacher's and student's interest
- Explore the WPRDC.org website and peruse the Open Datasets available
- Have students find some datasets that would be interesting to explore correlations
- Once they find a data set, create scatter plots with a trend line and r-squared to attempt to find correlations
- For instance:
    - Is Air Quality in Allegheny County correlated to temperature?
    - Are Firearm Seizures correlated to crime rates in Pittsburgh neighborhoods?

Evaluation - award points for successfully:

- Finding a data set of interest
- Creating scatter plots

- Adding in trend lines and R-squared values
- Determining if there is a strong correlation
- Creating a few slides to share their findings with the class

Activity 4: Mini project Explore a Data Set as Class

- We propose the following question for students to investigate: ***What drives student achievement?***
    - Is there a correlation between where a student lives and his/her success?
    - Do students who come from wealthier communities have an advantage over others?
    - Is there a correlation between the number of AP and Honors courses a school offers?
    - Is the number of extracurricular activities a school offers a factor?
    - Think of any other variables that you believe may have an impact, and research.
- Here, we can define "success" using, e.g., such metrics as PSSA, Keystone, SAT, and ACT scores which are readily available at the school district level. We can define community "wealth" with metrics such as, e.g., median household income and median home value.
- Students will search for multiple datasets and create their own worksheet with relevant data columns. Here are some websites that should get you started:
    - PA Department of Education Data & Reporting
    - Allegheny County Data
    - US Census Data
    - Data USA
- Once they have created their spreadsheet, students are then free to explore any relevant correlations they may find.
- Students should create multiple analyses and see what correlations exist, if any.
- The teacher can decide how these correlations, if any, are presented (e.g. poster session, Google Slides, PowerPoint, etc.)
- A robust class discussion will surely take place!

  Notes:

    1. This is only a suggested activity that was created from an overarching question - do wealthy kids have an advantage?
    2. Perhaps the teacher, or better yet the students, might have such questions they would like to explore! (Baseball/Football/Soccer payroll vs wins, etc)

Evaluation - award points for successfully:

- Locating data
- Creating a spreadsheet
- Creating scatter plots
- Adding in trend line and R-squared value

- Determining if there is a strong correlation

Activity 5: Tell the Mini project story

- How will students present their findings in a compelling manner? What tools are available? Encourage student creativity!
- Create slides containing your data - Google Slides template
- Share with the class

Evaluation - see rubrics below

- Correlation Mini-Project (with points)
- Correlation Mini Project (without points)

Possible Timeline:
Day 1-2: Activity 1
Day 3-5: Activity 2
Day 6-7: Activity 3
Day 8-10: Activity 4
Day 11-12: Activity 5