MODULE #2: An Introduction to Spreadsheet Analysis

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):
- Ability to find sources of open data
- Ability to download a dataset as a CSV, Google Sheets and Excel Spreadsheet file
- Ability to sort data
- Ability to sum, count and average data
- Ability to create a Pivot Table in Google Sheets and/or Excel
- Ability to manipulate a Pivot Table to analyze data and answer questions
- Ability to create a bar or pie chart in Google Sheets and/or Excel
- Ability to answer questions with data analysis
- Ability to uncover a research problem and formulate a research question
- Ability to recommend potential policy to a party in order to help solve your research problem
- Ability to tell a story with data

Enduring Understanding: We can tell stories with data.

Module Objective: Begin spreadsheet training to give students the skills needed to locate and download useful data, sort data, construct Pivot Tables to segment data, visualize data to answer questions and tell stories.

Student Objectives:
- To think like a data scientist...
    - Is there data available to answer this question?
    - How do we get the data?
    - What do we do with the data to answer questions?

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.6 - Use the concepts of independence and conditional probability to interpret data.

PA Standard Area: Computer Science - 3A.DA Data Analysis
- 3A.DA.10 - Evaluate the tradeoffs in how data elements are organized and where data is stored.
- 3A.DA.11 - Create interactive data visualizations using software tools to help others better understand real-world phenomena.

- 3A.DA.12 - Create computational models that represent the relationships among different elements of data collected from a phenomenon or process.

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.

PA Standard Area: 15.4 Computer & Information Technologies
- 15.4.12.A - Apply the creative and productive use of emerging technologies for educational and personal success.

PA Standard Area: English Language Arts
- 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.

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.

Interpret linear models

CCSS.MATH.CONTENT.HSS.ID.C.9
Distinguish between correlation and causation.

Make inferences and justify conclusions from sample surveys, experiments, and observational studies
CCSS.MATH.CONTENT.HSS.IC.B.6
Evaluate reports based on data.

Materials:
- Baseball Data Activity Resources
    - Spreadsheet: Module 2: Baseball Sample Data
    - Instruction Sheet
    - Video
- Dog License Data Activity Resources
    - Video Instructions (using Google Sheets): Dog License Activity Instructions

- ○ Video Instructions (using Excel):
  - ■ [Downloading the data set from WPDC](#) (also counting & sorting)
  - ■ [Creating Pivot Tables](#)
- ● General Spreadsheet Video Resources:
  - ○ [Using spreadsheets, arranging and working with data](#)
  - ○ [Creating Pivot Tables](#)
  - ○ [Creating graphs](#) in Excel
- ● [Western PA Data Center Website](#)
- ● Google Sheets or Microsoft Excel
- ● Post-It Notes

Prerequisites: The teacher will need to be able to...
- ● Search/download the appropriate dataset
- ● Use spreadsheet functions including:
  - ○ Sorting data
  - ○ Inserting columns/rows
  - ○ Calculating sum/count
  - ○ Creating Pivot Tables with one and multiple variables
  - ○ Creating charts and graphs

Activity 1: Baseball Data Activity
- ● Have students open [Module 2: Baseball Sample Data](#)
- ● Discussion:
  - ○ This data set is very small (only 21 rows of data).
  - ○ Other data sets they will be using will have 10s of thousands of rows of data.
  - ○ This sample data set was created to teach students how to analyze data using Pivot Tables before getting into the larger data sets.

- ● Spreadsheet Analysis:
  - ○ **Resources**: [Instruction Sheet](#) and [Instructional Video](#)
  - ○ Identify the column headers (team, batting average, hits, home runs, etc.). These are the variables which you may analyze. All data sets will have these.
  - ○ **Counting** the number of rows of data.
  - ○ **Sorting** the data by a desired field (ex: team, age, year in league, etc.)
  - ○ Quick counting subsets of the data by selecting multiple rows.

- ● Creating & Utilizing Pivot-Tables:
  - ○ Questions to answer with your PivotTable:
    - ■ *How many 12 year olds are in the league?*

- *How many participants are batting .500 or higher?*
- *How many participants have not hit a home run?*
- *How many participants are on each team?*
- *What team has the highest batting average?*
- *Do older participants have a higher batting average?*
- *Do participants with more experience have a higher batting average?*
- *Which team has the most home runs?*
- Any other questions you wish to investigate!

Evaluation - award points for successfully:
- Sorting the data
- Performing basic Counts and Sums
- Creating a Pivot Table
- Answering the questions listed above on a sheet of paper or Google Doc

Activity 2: Dog License Data Activity
- This activity will have students use the skills they learned in Activity 1 to analyze a larger data set in order to answer research questions.
- Question - *What are the most common dog names in Allegheny County?*
- Post-it notes on board: your best guesses (5 mins or fewer).
- Do some Affinity Clustering - are there trends among student guesses?
- Class discussion:
  - Why did students choose these names?
  - How would we decisively answer such a question? (*Conduct surveys, random phone calls, knock on doors, veterinarians' records, etc.*)
  - What are disadvantages to these techniques? (*Cost, time consuming, cooperation, etc.*)

Activity 3:
- Explore the WPRDC.org website and see all the Open Data sets available to anyone…
- Students explore the Dog License Database
- Class discussion:
  - What information is available? What are the column headers (variables)?
  - Notice the database is quite large, and would take a long time to look through and tally up all the names by hand…

Activity 4:
- Students download the dataset to their device and open it with a spreadsheet application (Google Sheets or Microsoft Excel).
- Discuss with students how data comes in different formats (XLSX,CSV, Google Sheet, etc.) and which types will be appropriate for our activity. Resources
- Teacher leads the process of downloading the dataset in CSV form and opening it with the appropriate tool (Google Sheets, Excel, etc.)

- **Notes**:
  - CSV files cannot be saved. Students must download the file in Google Sheets or save in Excel for the data to appear in the next session.
  - You may wish to provide students with a portion of this data set (possibly sort by zip code and provide certain zip codes) as it is very large and may take a long time to download and analyze on Chromebooks or other student devices.
- Video resources:
  - Instructions specific to Dog License Data (in Google Sheets): Dog License Activity Instructions
  - Instructions specific to Dog License Data (in Excel): Downloading the data set from WPDC
  - General spreadsheet information: Using spreadsheets, arranging and working with data

Activity 5:
- As students scroll through the spreadsheet, teachers can prompt discussion questions,
  - How many rows of data are there?
  - What data did the county choose to share with the public?
  - Why was this data selected to be shared?
  - What data has been redacted or hidden from our view and why might it have been withheld?
  - Students may have their own questions to add as well…

Activity 6: You may wish to review, and practice the skills introduced in the Baseball Data Simulation in Activity 1, as a class.
- Counting the number of rows of data
- Sorting the data by a desired field (e.g. Dog name, breed, ZIP code, etc.)
  - How does sorting the data reveal patterns in the data?
  - Note: Be sure to sort the whole data set and not just the column
- Quick counting subsets of the data by selecting multiple rows.
  - Count and Sum in the footer of the window
- Is your dog in the spreadsheet? Why or why not?
- Any additional teacher-led activities with the data that you may wish to add!

Evaluation - award points for successfully:
- Downloading the spreadsheet
- Sorting the data
- Performing basic Counts and Sums

Activity 7: Creating & Utilizing PivotTables
- Video resources:

- - Instructions specific to Dog License Data (in Google Sheets): Dog License Activity Instructions
    - Instructions specific to Dog License Data (in Excel): Creating Pivot Tables
    - General information on creating Pivot Tables: Creating Pivot Tables
- Teacher leads a session to review the process to build a Pivot Tables, which can:
    - Count the number of instances of a given property
    - Segment the dataset
- Answer our original question - *What is the top dog name in Allegheny County?*
- Ask additional questions, such as:
    - *What zip code has the most dogs?*
    - *What is the most popular dog name in your zip code?* (Requires segmentation)
- Students can write their findings on a sheet of paper or type in a Google or Word Doc

Evaluation - award points for successfully:
- Creating a Pivot Table
- Recording the following information on a sheet of paper or Google Doc.
    - Identifying the top 10 dog names in Allegheny County
    - Identifying the zip code with the most dogs
    - Identifying the top ten dog names in their zip code

Activity 8: Visualize the Data
- Video resource: Creating graphs in Excel
- How can the results be visualized?
- Teachers can lead a discussion of various ways data is communicated effectively.
- What tools are available? (e.g. bar chart, pie chart, hand-drawn graph, word cloud, etc)
- Walk students through the steps to create bar chart for the top 10 dog names in Allegheny County
- Good place to encourage student creativity in reporting results. Is there such a thing as a "bad data visualization?"

Evaluation - award points for successfully:
- Creating a bar chart of the top 10 dog names in Allegheny County

Activity 9: Tell the Story
- How will students present their findings in a compelling manner?
- What tools are available? Encourage student creativity!
- Have students create their own visualization displaying this data

Evaluation - award points for successfully:
- Creating their own visualization displaying this data

- Presenting their visualization to the class

Activity 10: Provide Actionable Recommendations

Now that students are familiar with PivotTables, many new questions can be quickly proposed and analyzed.

- Have students brainstorm their own questions of interest to answer (e.g. most popular breed, most popular color, percentage spayed/neutered, number of dogs that live in my ZIP code, etc.) OR
- Teacher might assign two or three questions to be answered as an assessment.
- Have students uncover a question/problem that can be solved with this data. For example...*Where should Petco open a new store? Where should Allegheny County or the City of Pittsburgh create a dog park? Where does Giant Eagle need to stock more dog food?*
    - What entity, organization, business can use this data?
    - How is it actionable?
    - Answer the question with the dataset
    - Visual the data

Evaluation - award points for successfully:

- Uncovering a question/problem
- Creating a visualization
- Determining how the data can be used by an entity/organization/business
- Proposing a recommendation

Extensions:
- Has the top ten list of dog names changed over the years, and how might we answer that question?
- Top 10 baby names over the last decade? How have trends in baby names changed over the years?

Possible Timeline (open to teacher discretion):
Day 1: Activity 1
Day 2: Activities 2-4
Day 3: Activities 5-6
Day 4-5: Activity 7
Day 6: Activity 8
Day 7: Activity 9-10