Pittsburgh Data Jam 2016-2017 - Data analysis Workshop Exercises

Our Question: What are the factors that influence reading and mathematics proficiency scores for students in Allegheny County?

Note: There are no right or wrong answers in these exercises. The goal is to learn how use Excel to understand and model the data. Focus on the process and not the end goal. Try using many Excel functions.

Data Sets: You will be using two data sets that contain different information.

1. **2012 math and reading aggregated allegheny.xlsx**
    - Contains information about the 2012 PSSA Mathematics and Reading District-wide Level Proficiency Results
    - Broken out by school districts
2. **teacher demographics.xls**
    - Contains salary, teaching experience and education level of all students teaching in Allegheny County School Districts.
    - Each row represents an individual teacher. Names of the teacher are not included.

Excel help:

- Using excel.doc – Located in the Other Documents directory on USB
- Use google

Data Exploration Exercise

Objective: To understand the data that you have to work with and get a better understanding of the variables. After this step you should have a good understanding of which data variables you are going to keep working with. Make sure you look at both data sets.

Excel Skills: In this exercise you should use:

- Sort
- Filter
- Summarize
- Create Crosstabs (Pivot Table)
- Charting

Questions to answer:

- Do I need all of the variables?
- Which are likely the most helpful in answering our question?
- Do the variables offer enough information? i.e. Is there enough differentiation in the data points?
- Are means and medians similar?
- Are there outliers? What will you do with them?
- Are there missing values? How will you fix them?
- What do graphs tell us about the data?
- Should I create new variables? What would they be?

Steps to take:

Data Set 1:

- Open 2012 math and reading aggregated Allegheny spreadsheet
- Do you need all the rows? Can you delete some?
- Investigate the variable (columns) names. Do you understand what each variable represents?
- How Many rows of data do you have? Keep track of this to ensure no data is lost.
- Do you need all columns? Delete if unnecessary.
- Calculate Minimums, Maximums, Mean, and Mode for your variables.
    - Sum?
- Sort the School Districts based on different variable.
    - Do you see any outliers?
    - Is all the data present?

Data Set 2:

- Open teacher demographics spreadsheet
- Do you need all the rows? Can you delete some?
- Investigate the variable (columns) names. Do you understand what each variable represents?
- How Many rows of data do you have? Keep track of this to ensure no data is lost.
- Do you need all columns? Delete if unnecessary.
- Calculate Minimums, Maximums, Mean, and Mode for your variables.
    - Sum?
- Copy these values to another sheet.
- Sort the rows based on different variable.
    - Do you see any outliers?
    - Is all the data present?
- Add Filter – Select all rows where AnnualSalary, YearsInEd, YearsInSD **OR** School_District has no value.
- Add a new variable (Column) called YearsNotinSD
    - Fill in formula subtracting YeradInSD from YearsInEd
- Are there other variables you could create?
- Create a PivotTable using all columns.
    - Experiment with different combinations of values in rows, columns and values.
- Create a final pivot table with rows = School District and values whatever you like to combine with the Proficiency spreadsheet.
- Create a bar chart
- Create a line chart
- Try some other charts
- What insights do you get from the charts?

Data Preparation Exercise

Objective: In this step you will fix any outliers or missing values you encounter. You may decide to not make any changes, but make sure you understand why.

Questions to answer:

- Did I address all Outliers for the variables I am interested in using?
- Did I fix all missing values and bad data elements?
- Did I create new variables?
- Did I bring all my data together?

Excel Skills: In this exercise you should use:

- Copy/Paste
- Sort
- Filter
- Cell Functions

Steps to take:

- Fix any outliers from prior exercise
  - Use sorting to help find rows
- Fix any missing values from prior values
  - Use Filters to find missing values
- Merge the two data sets.
  - Make sure to copy aggregated data from teacher demographic data set to proficiency data set. Use a cross tab in the proficiency data set
  - Ensure both data sets are sorted the same and have the same number of rows
  - Delete the extra column after data is copied.
- Create a new variable based on two other variables

Data Modeling Exercise

Objective: You should create at least one trend model that will show the correlation of two variables.

Questions to answer:

- What are the model formulas for your analysis?
- What are the best variables to use in predicting your dependent variable?
- Which model has the highest R^2 score? What does that tell you about your model?
- Is it a good predictor?
- What conclusions can you make from your analysis?
- What actions would you recommend?
- What additional data might be useful?
- What new questions would you be able to ask?

Excel Skills: In this exercise you should use:

- Scatter Plots

Steps to take:

- Create scatter plots for a pair of variables.
- Create a trend line
- Show trend line formula
- Show coefficient of determination (R^2)
- Repeat for a few combinations.
- Which one has the best correlation?