Data Structures and Algorithms in Python

Project 1 – Tell a Data Science Story

For this project you should turn in the following:

- Typed answers to the Data, Data, Data reflection questions and sketches/notes from your planning phase.
- .ipynb file with your data analysis
- PowerPoint presentation with figures you generated using Python

Please review the assignment rubric for the specific requirements. A presentation template is attached - you do not have to use the template, but you should have all the sections indicated.

Please place the three files in a .zip folder and upload one per group to this assignment. Many options for data sources are available on the CREEKnet topics page. Below are some helpful tips for the planning & preparation phase of your project:

Hypothesis generation

- Refine the question or hypothesis you want to explore in your project
- Make a plan for what steps you need to take to answer the question
- Sketch out potential plots including x and y axes (do this on paper with your group)

Data cleaning

- If necessary, clean your data programatically. Add commands to your .ipynb to do this.
- You should be using pandas, check out documentation
- To help with data frame manipulation in pandas check out this Jupyter Notebook
- What variables do you need? What outliers should you remove? What variable has too much missing data to be reliable?

Assignment Rubric

Reflection: Dataset choice

Excellent	Proficient	Developing	Missing
20 points	*18 points*	*16 points*	*0 points*
The written reflection questions are all answered fully. You have stated a clear hypothesis or question and developed a plan to address it.			

Programming: Importing a meaningful dataset

Yes	No
10 points	*0 points*
You have imported a dataset in your program (examples used in previous assignments do not count). You included a comment or markdown cell in your .ipynb file indicating the source of the data.	

Programming: Use Pandas for Data manipulation

Excellent	Proficient	Developing	Missing
20 points	*18 points*	*16 points*	*0 points*
Pandas is used to filter and/or subset data multiple times in your code. Pandas and/or numpy is used to generate summary statistics such as an average, correlation, standard deviation, etc.			

Programming: Use Seaborn or Matplotlib for Data visualization

Excellent	Proficient	Developing	Missing
20 points	*18 points*	*16 points*	*0 points*
At least four plots are generated in your code. These plots should have correctly labeled axes, and should be relevant to the questions/ hypotheses in your project.			

Presentation

Excellent	Proficient	Developing	Missing
30 points	*26 points*	*22 points*	*0 points*
Presentation is visually appealing and states a concrete question/hypothesis. All required sections from the template are present and fully developed. Presenters maintain eye contact and answer audience questions confidently.			