PROJECT REPORT

CS595 : TERMINATION PROJECT SPRING 2019

TOPIC :- Data Visualization

Manish Kumar, B00715673

Email: mkumar7@binghamton.edu

PROJECT OVERVIEW :-

This project has two parts that demonstrate the importance and value of data visualization techniques in the data analysis process. In the first part, you will use Python visualization libraries to systematically explore a selected dataset, starting from plots of single variables and building up to plots of multiple variables. In the second part, you will produce a short presentation that illustrates interesting properties, trends, and relationships that you discovered in your selected dataset. The primary method of conveying your findings will be through transforming your exploratory visualizations from the first part into polished, explanatory visualizations.

What do I need to install?

This project uses Python 3 and is designed to be completed through the Jupyter Notebooks IDE. It is highly recommended that you use the Anaconda distribution to install Python, since the distribution includes all necessary Python libraries as well as Jupyter Notebooks. The following libraries are expected to be used in this project:

- NumPy
- pandas
- Matplotlib
- Seaborn

Why this project?

Data visualization is an important skill that is used in many parts of the data analysis process. **Exploratory** data visualization generally occurs during and after the data wrangling process, and is the main method that you will use to understand the patterns and relationships present in your data. This understanding will help you approach any statistical analyses and will help you build conclusions and findings. This process might also illuminate additional data cleaning tasks to be performed. **Explanatory** data visualization techniques are used after generating your findings, and are used to help communicate your results to others. Understanding design considerations will make sure that your message is clear and effective. In addition to being a good producer of visualizations, going through this project will also help you be a good consumer of visualizations that are presented to you by others.

What will I learn?

After completing this project, you will be able to:

- Supplement statistics with visualizations to build understanding of data.
- Choose appropriate plots, limits, transformations, and aesthetics to explore a dataset, allowing you to understand distributions of variables and relationships between features.
- Use design principles to create effective visualizations for communicating findings to an audience.

PROJECT OVERVIEW :-

This project is divided into two major parts. In the first part, you will conduct an **exploratory** data analysis on a dataset of your choosing. You will use Python data science and data visualization libraries to explore the dataset's variables and understand the data's structure, oddities, patterns and relationships. The analysis in this part should be structured, going from simple univariate relationships up through multivariate relationships, but it does not need to be clean or perfect. There is no one single answer that needs to come out of a given dataset. This part of the project is your opportunity to ask questions of the data and make your own discoveries. It's important to keep in mind that sometimes exploration can lead to dead ends, and that it can take multiple steps to dig down to what you're truly looking for. Be patient with your steps, document your work carefully, and be thorough in the perspective that you choose to take with your dataset.

In the second part, you will take your main findings from your exploration and convey them to others through an **explanatory** analysis. To this end, you will create a slide deck that leverages polished, explanatory visualizations to communicate your results. This part of the project should make heavy use of the first part of the project. Select one or two major paths in your exploration, choose relevant visualizations along that path, and then polish them to construct a story for your readers to understand what you found.

Step 1.1: Choose your Dataset

First, you will choose a dataset from the Dataset Options.

Download the Dataset Options file for full details & descriptions from the Resources Tab.

1. Click on Resources in the leftmost panel of your classroom

2. Click the File Name to start download
   Quick List Below:
   Dataset Options

   Ford GoBike System Data Flights

   Or select your own dataset! See guidelines in the Dataset Options download in the Resources tab on whether or not a dataset will be appropriate for use in this project Remember that finding and cleaning your own data set could take significant time and effort!

   A Google Doc download option with identical info is available below as well, if you prefer it. This is not accessible on all networks. Google Doc Download

Step 1.3: Explore Your Data

It's time to get to the interesting bits. Explore your data and document your findings in a report. The report should briefly introduce the dataset, then systematically walk through the points of exploration that you conducted. You should have headers and text that organize your thoughts and findings. Visualizations in this part of the project need not be completely polished: this is just your own exploration at this point. However, you should still make sure that you adhere to principles of using appropriate plot types and encodings so that accurate conclusions can be drawn, and that you have enough comments and labeling so that when you return to your work, you can quickly grasp your analysis steps.

If you use a Jupyter Notebook for this step of the project, don't forget to export the notebook as an html file for the project submission.

Step 2.1: Document your Story

At the end of your exploration, you probably have a bunch of things that you've discovered. Now it's time to organize your findings and select a story that you will convey to others. In your readme document, you should summarize your main findings and reflect on the steps you took in your data exploration. You should also lay out the key insights that you want to convey in your explanatory report as well as any changes to visualizations, or note new visualizations that will be created to bridge between your insights.

Step 2.2: Create your Slide Deck

Follow the plans you laid out in the previous step and create a slide deck with explanatory data visualizations to tell a story about the data you explored. You can start with code that you used in your exploration, but you should make sure that the code is revised so that your plots are polished. Make sure that you also pay attention to aspects of design integrity in your revisions.

Step 2.3: (Optional) Get Feedback

Though not required, it is highly recommended that you try to get feedback from at least one person before you submit your project. By sharing your work with others, you can get input from a different perspective that catches things that you may have originally missed. Share your slide deck with someone in person and have them provide live feedback on what they get from your slide deck. Alternatively, you can also share your work with your fellow students. Post a message in a student community channel for this project with a link to your project and ask for feedback. Be sure to keep an eye out for others who are also seeking feedback and return the favor!

You might need to ask specific questions to prompt your reader. The following questions might be good starters; be sure to follow up or come up with your own questions:

- What do you notice about each visualization?
- What questions do you have about the data?
- What relationships do you notice?
- What do you think is the main takeaway from the slide deck?
- Is there anything that you don’t understand from the plots?

If you get feedback from others, then add their feedback to your readme document. Note what changes you make to your slide deck and designs based on that feedback. You can also include feedback from your reviewer as part of this revision process.