**Jalin Roberts – Exploratory Data Analysis Essay**

Exploratory Data Analysis Is how we describe the process of going from having a lot of raw data, and performing tasks on this data to better understand. Specifically, EDA is the process of taking that raw data, performing tasks on it using tools of programming languages such as Python, R, and SQL, and then manipulating it to better understand. As these tools are used, their usefulness, application, and enrichment helps us visualize our information. Tools such as Matplotlib, Pandas, Seaborn, and NumPy are what we use to perform our Exploratory Data Analysis. Personally, this exploratory process is making sense of information that does not completely make sense. Though we may have certain sets of information, both given and enriched, have the ability to visualize our data makes a major impact. This part of the process is what truly allows both programmers, stockholders, and non-computer science personnel all able to understand what has happened along this process.

Data Analysis begins with the most important step along the process…. Inputting our raw data! We cannot do any single part of the process without our raw data, so before we can begin, we must have data to analyze. Raw data is data that we have not processed, coded, or done anything with. How raw data is prepared for analysis is another story, but our process simply begins by loading our raw data. From there we begin with the first library that we have covered this semester…. NumPy!

NumPy is used for creating arrays, and we can create almost anything array-wise with this library. We can also create things like multi-dimensional arrays and matrices, along with more functions that we can use to operate on these arrays. This library is used heavily for things like statistics, data analysis, and research. This library also lets use perform element-wise operations very easily without having to use any loops. With NumPy, you are kind of limited as to the kind of data you can work with. This library works best with data that is uniform. So while its usefulness with arrays is quite helpful, this library does have some shortcomings. This is where Pandas comes into the process.

Pandas is a Python library that stands for Panel Data. This library works best with data frames, and luckily it can handle different kinds of data. This library is probably the most used library for exploratory data analysis. By importing our raw data, we can now begin to visualize, interpret, and brainstorm ideas for what in our data that we would like to change, or know more about. This library will help us create Python objects with rows and columns (dataframes) that look very similar to tables. What this does is allows us to have another layer of visualization to our data. Speaking of having added data visualization, what we do in Pandas leads us to another step in the analysis process, visualization using MatPlotLib.

This library has a lot to it; it is a library that we use for visualization in Python. More specifically, a lot of what we can do revolves around things like plotting. If we wanted to quickly visualize a dataset in a pie chart, we can easily use MatPlotLib to make this process even easier. We can also customize our styles of plots and layouts. Typically, we can use the main types of graphs that we think about; histograms, pie charts, box plots, etc. However, where MatPlotLib comes into play being useful for plotting, we can also use another library that takes this library to the next level. That library is called Seaborn.

Seaborn is a Python library that we use that is built on top of matplotlib. More specifically, we can only use matplotlib to plot graphs using Pandas and NumPy. With Seaborn, we can extend our plotting capabilities to utilize MatPlotLib, NumPy, and Pandas to plot our graphs. When we only want to plot basic graphs for visualization, MatPlotLib will be the way to go. However, Seaborn’s built-in default themes afford us to work with more interesting visualizations. We can also get even more creative in our plots; heat maps, lmplots, etc. In a nutshell, Seaborn is just a more visually appealing version of MatPlotLib, with extra visualizations.

The overall process of exploratory data analysis is to take our raw data, perform some manipulations on it, and return our new information. Along we way we will answer questions, create new ones, and eventually be able to properly visualize our answers. Through the use of these Python libraries, we can make this process far easier and faster. As we answer the questions we are looking to answer, we can properly plot, visualize, and represent our findings. The added effort we put into our visualizations helps use adequately explain our findings to anyone in Data Science, or outside of this field.