Bailey Korfhage

DS-160-01

Exam 1

Exploratory data analysis (EDA) is an objective way for analysts to gather data and conclusions. To the human eye, small amounts of data can be transcribed into useful conclusions. However, when data sets become too large to analyze with simple logic, analysts turn to EDA. SQL, pandas, python, seaborn, tableau, and matplotlib are all programs used in EDA. For the purposes of this essay, details on EDA will be provided with python applications.

Python applications require that you install pandas, seaborn, matplotlib, and numpy using “!pip install”. Once the applications have installed on the computer, they must be imported. Usually, as the applications are imported, they are shorted to pd, sns, plt, and np, respectively. To start EDA, there must either be a file to analyze (usually a csv) or the analyst must create a data frame using pandas. If it is a file, pandas can read the csv and save it to a variable. To create a data frame, the original data must be stored in a dictionary or list and converted properly. Once imported or created, the data frame can be given basic diagnostics such as “.head()”, “.tail()”, “.sample()”, “.info()”, “.describe()”, and “.columns”. The “head”, “tail”, and “sample” function pulls the first five rows, last five rows, and a sample of five rows, respectively. “Info” gives the column name, number of null values in each column, and the type of values in each column. “Describe” gives basic statistics of all columns that have numeric values. The “columns” function lists the columns of the data set for quick reference.

Before analysis can occur, an analyst must prepare the data set. For example, there could be values missing that must be filled before proper EDA can be done. With a column containing a numerical value, mean or median is usually best to fill the null values. However, if the column contains a categorical value, mode is the best way to account for the missing value. The analyst might also want to clean up the data to be more specific. Pandas allows the data to be sliced into certain sections to give the analyst a chance to home in on specific portions that would be useful in the conclusions. Also, if there is a column that isn’t necessary, it can be removed using the “. drop()” function.

Once the analyst has properly assessed the data for adjustments, simple plots can be created. To assure correct analysis is done, the values of each column must be the correct data type. Dates/times must be converted to “datetime” and numbers must be converted to integers using “int64”. Histograms, line plots, and box plots are all types of descriptive charts that give a visual representation of the data. Histograms are helpful for the dispersion of the data. Line plots are helpful for seeing how the data is over time. Box plots give the five number summary (min, lower quartile, median, upper quartile, max) and show any outliers.

Plotting the data is a great way to turn data from a jumble of numbers into a more universal recognition tool. However, plotting data doesn’t do much more than put the numbers and words into a different model. The best way to perform EDA is through correlation and statistics. The correlation between two variables is a numeric representation of their relationship. It ranges from -1 to +1 with a positive correlation indicating a direct relationship (as one variable increases, so does the other) and a negative relationship indicating the opposite. The weaker the relationship, the lower the correlation number. In statistical analysis, if the number is above .8, (or below -8), the relationship is so strong that one variable predicts the other in many instances. Pandas, matplotlib and seaborn can all calculate and plot this number.

Numpy, pandas, matplotlib, and seaborn all have their own unique properties for EDA. Numpy is very helpful for creating a data frame. It can create an array out of a list to then turn into a data frame. Pandas pretty much takes point on the organization of the data. It imports the data into the system and can manipulate the set into the specific information that the analyst requires. Pandas also is great for simple plots with the data extracted. Matplotlib is like pandas but has many more features for EDA. In addition to histograms, box plots, and line plots, matplotlib can make graphs that allow for more comparison. It can do multiple plots in one figure, twin axes, and stack plots. These graphs allow a visual comparison of data columns. Seaborn takes it even further. The program adds a more statistical approach to the graphs. It automatically adds a confidence interval, title, and axis labels. It is a more pristine program for adding three columns to a set. For example, a bar graph does not only have to be graphed by the x and y axis, it can be colored, divided, and styled by a different data set column’s parameters. Overall, seaborn is the most sophisticated of the three plotting programs, but EDA can be efficiently performed on any python applications.