Why do we do EDA?

link= "https://towardsdatascience.com/exploratory-data-analysis-topic-that-is-neglected-in-data-science-projects-9962ae078a56#:~:text=By%20completing%20the%20EDA%20you,from%20exploring%20your%20(https://towardsdatascience.com/exploratory-data-analysis-topic-that-is-neglected-in-data-science-projects-9962ae078a56#:~:text=By%20completing%20the%20EDA%20you,from%20exploring%20your%20
Exploratory Data Analysis is a crucial step before you jump to machine learning or modeling your data. By doing this you can get to know whether the selected features are good enough to model, are all the features required, are there any correlations based on which we can either go back to the Data Preprocessing step or move on to modeling.

Once EDA is complete and insights are drawn, its feature can be used for supervised and unsupervised machine learning modeling.

In every machine learning workflow, the last step is Reporting or Providing the insights to the Stake Holders and as a Data Scientist you can explain every bit of code but you need to keep in mind the audience. By completing the EDA you will have many plots,heat-maps, frequency distribution, graphs, correlation matrix along with the hypothesis by which any individual can understand what your data is all about and what insights you got from exploring your data set.

We have a saying “A picture is worth a thousand words”.

I want to modify it for data scientist as “A Plot is worth a thousand rows”

What are the steps in EDA?

There are many steps for conducting Exploratory data analysis. I want to discuss regarding the below few steps

- Description of data
- Handling missing data
- Handling outliers
- Understanding relationships and new insights through plots

In [ ]:
```
1
```

a) Description of data:

We need to know the different kinds of data and other statistics of our data before we can move on to the other steps. A good one is to start with the describe() function in python. In Pandas, we can apply function describe on a data frame which helps in generating descriptive statistics that summarize the central tendency, dispersion, and shape of a dataset’s distribution, excluding “NaN” values.

For numeric data, the result's index will include count, mean, std, min, max as well as lower, 50 and upper percentiles. By default, the lower percentile is 25 and the upper percentile is 75. The 50 percentile is the same as the median.

For object data (e.g. strings or timestamps), the result's index will include count, unique, top, and freq. The top is the most common value. Freq is the most common value frequency. Timestamps also include the first and last items.

In [ ]:
```
1
```

b) Handling missing data:

Data in the real world are rarely clean and homogeneous. Data can either be missing during data extraction or collection due to several reasons. Missing values need to be handled carefully because they reduce the quality of any of our performance metrics. It can also lead to wrong prediction or classification and can also cause a high bias for any given model being used. There are several options for handling missing values. However, the choice of what should be done is largely dependent on the nature of our data and the missing values. Below are some of the techniques:

- Drop NULL or missing values
- Fill Missing Values
- Predict Missing values with an ML Algorithm

(i) Drop NULL or missing values:

This is the fastest and easiest step to handle missing values. However, it is not generally advised. This method reduces the quality of our model as it reduces sample size because it works by deleting all other observations where any of the variables are missing.

Python code :

dataset.dropna()

(ii) Fill Missing Values:

This is the most common method of handling missing values. This is a process whereby missing values are replaced with a test statistic like mean, median or mode of the particular feature the missing value belongs to.

Python code :

dataset['Column_name']=dataset['Column_name'].fillna(mean_value).

(iii) Predict Missing values with an ML Algorithm:

This is by far one of the best and most efficient methods for handling missing data. Depending on the class of data that is missing, one can either use a regression or classification model to predict missing data.

In [ ]:
```
1
```

c) Handling outliers:

An outlier is something separate or different from the crowd. Outliers can be a result of a mistake during data collection or it can be just an indication of variance in your data. Some of the methods for detecting and handling outliers:

- Box Plot
- Scatter plot
- Z-score
- IQR(Inter-Quartile Range)

(i) Box Plot:

A box plot is a method for graphically depicting groups of numerical data through their quartiles. The box extends from the Q1 to Q3 quartile values of the data, with a line at the median (Q2). The whiskers extend from the edges of the box to show the range of the data. Outlier points are those past the end of the whiskers. Box plots show robust measures of location and spread as well as providing information about symmetry and outliers.

(ii) Scatter plot:

A scatter plot is a mathematical diagram using Cartesian coordinates to display values for two variables for a set of data. The data are displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis. The points that are far from the population can be termed as an outlier.

(iii) Z-score:

The Z-score is the signed number of standard deviations by which the value of an observation or data point is above the mean value of what is being observed or measured. While calculating the Z-score we re-scale and center the data and look for data points that are too far from zero. These data points which are way too far from zero will be treated as the outliers. In most of the cases, a threshold of 3 or -3 is used i.e if the Z-score value is greater than or less than 3 or -3 respectively, that data point will be identified as outliers.

Python Code:

z = np.abs(stats.zscore(dataset)) Once we get the z-score we can fit our datset base on that.

Python Code:

dataset = dataset[(z < 3).all(axis=1)]

(iv) IQR:

The interquartile range (IQR) is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles.

IQR = Q3 − Q1.

Python Code:

Q1 = dataset.quantile(0.25) Q3 = dataset.quantile(0.75) IQR = Q3 — Q1

Once we have IQR scores below code will give an output with some true and false values. The data point where we have False means values are valid and True indicates the presence of an outlier.

Python Code:

print(boston_df_o1 < (Q1–1.5 * IQR)) |(boston_df_o1 > (Q3 + 1.5 * IQR))

In [ ]:
```
1
```

d) Understanding relationships and new insights through plots :

We can get many relations in our data by visualizing our data set.Let's go through some techniques in-order to see the insights.

- Histogram
- Heat Maps

(i) Histogram:

A histogram is a great tool for quickly assessing a probability distribution that is easily understood by almost any audience. Python offers a handful of different options for building and plotting histograms.

(ii) Heat Maps:

The Heat Map procedure shows the distribution of a quantitative variable over all combinations of 2 categorical factors. If one of the 2 factors represents time, then the evolution of the variable can be easily viewed using the map. A gradient color scale is used to represent the values of the quantitative variable. The correlation between two random variables is a number that runs from -1 through 0 to +1 and indicates a strong inverse relationship, no relationship, and a strong direct relationship, respectively.

In [ ]:
```
1
```

What are the tools used for EDA?

There are plenty of open-source tools exist which automate the steps of predictive modeling like data cleaning, data visualization. Some of them are also quite popular like Excel, Tableau, Qlikview, Weka and many more apart from the programming.

In programming, we can accomplish EDA using Python, R, SAS. Some of the important packages in Python are:

- Pandas
- Numpy
- Matplotlib
- Seaborn
- Bokeh

In [ ]:
```
1
```

What happens if we don’t do EDA?

Many Data Scientists will be in a hurry to get to the machine learning stage, some either entirely skip exploratory process or do a very minimal job. This is a mistake with many implications, including generating inaccurate models, generating accurate models but on the wrong data, not creating the right types of variables in data preparation, and using resources inefficiently because of realizing only after generating models that perhaps the data is skewed, or has outliers, or has too many missing values, or finding that some values are inconsistent.

In our Trip example, without any prior exploration of the place you will be facing many problems like directions, cost, travel in the trip which can be reduced by EDA the same applies to the machine learning problem.

In [ ]:
```
1
```