Name: Adwait Hegde
Roll No: 2019130019
TE Comp (Batch-A)

EXPERIMENT 1

Aim:

To perform Exploratory Data Analysis such as number of data samples, number of features, number of classes, number of data samples per class, removing missing values, conversion to numbers, using seaborn library to plot different graphs.

Theory:

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

It is a good practice to understand the data first and try to gather as many insights from it. EDA is all about making sense of data in hand, before getting them dirty with it. The EDA approach is precisely that--an approach--not a set of techniques, but an attitude/philosophy about how a data analysis should be carried out.

Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to

1. maximize insight into a data set;
2. uncover underlying structure;
3. extract important variables;
4. detect outliers and anomalies;
5. test underlying assumptions;
6. develop parsimonious models; and
7. determine optimal factor settings.

EDA is not identical to statistical graphics although the two terms are used almost interchangeably. Statistical graphics is a collection of techniques--all graphically based and all focusing on one data characterization aspect. EDA encompasses a larger venue; EDA is an approach to data analysis that postpones the usual assumptions about what kind of model the data follow with the more direct approach of allowing the data itself to reveal its underlying structure and model. EDA is not a mere collection of techniques; EDA is a philosophy as to how we dissect a data set; what we look for; how we look; and how we interpret. It is true that EDA heavily uses the collection of techniques that we call "statistical graphics", but it is not identical to statistical graphics per se.

Implementation:
Dataset was from Boston Police Department about Crime Incident Report of 2020

Colab Link:
https://colab.research.google.com/drive/1UShCSXWa2rgW2iU9hTsqg4a9JWN50nsw?usp=sharing

GitHub Link:
https://github.com/adwait-hegde/DataAnalytics-Lab/tree/main/Exp1

Dataset:
https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system/resource/be047094-85fe-4104-a480-4fa3d03f9623

Conclusion:

Dataset initially had 70894 data samples and 17 features out of which 4 were not useful for the EDA (were NaN or all unique) and thus were dropped.
Analysis showed that

- Crimes were evenly spread across days of the week and comparitively least were reported on Sundays.
- 0th hour had the greatest number of reported crime and then the data falls drastically which indicates the chances of unknown data being set to 0th hour. Daytime crimes were more as compared to the night time.
- One can find about 4500-7000 cases each month and highest in the month of October and lowest in April
- Shooting cases were very less (1.6%) as compared to one that did not involve shooting
- Then analysis was done on the cases that involved shooting that showed:
    - Cases were maximum during the mid year.
    - Shooting cases showed a reverse trend and were more during the night as compared to daytime and all the months showed similar trend
- Scatter plot was done using the latitude and the longitude data, but there were some data about 1.8% with anomaly (the data was initially 0, 0) which were set to the mean of all the data
- Correlation matrix showed that there was no corelation between the features