*Name: Sharanya R (Section A)*

Report for data analysis of US crime data

Description of the dataset:

It is a CSV file containing information about the different crimes of different US states. It has numerical and categorical columns describing various crime data such as crime type, police response time, arrest status and many more. Such datasets are crucial for law enforcements, lawmakers and researchers to analyze and understand patterns, trends and characteristics of criminal behavior. The dataset was taken from kaggle. The size of the dataset is 949kb. It has 6050 rows and 20 columns. Analyzing such datasets can reveal patterns and insights that contribute to the development of effective crime prevention policies and public safety initiatives.

Problem Statement:

Which US State has the best police response time for homicide?

Data Preprocessing:

The total null values in the dataset is 4221. All the columns have null values. The unnamed: 0 column was removed. The numerical columns were filled using their median values. Median values were preferred over mean values as they were more accurate and had less number of decimal points. The categorical columns were filled using their mode values.

There were 1022 duplicate values. They were dropped using the drop.duplicates() function. After handling the null values and duplicates, the dataset has 5028 rows and 19 columns.

Statistical Analysis:

The statistical analysis was done by get\_group, groupby and agg functions. The statistical analysis was performed for police response time in the state of Massachusetts. When finding the mean of police response time in each state, the state of Wisconsin had a better police response time of 0.94 compared to other states, so the state of Wisconsin as chosen for statistical analysis. The values are calculated in minutes.

Measure of Central Tendencies:

The mean of the police response time for homicide in the state of Wisconsin is 0.74 minutes.

The median of the police response time for homicide in the state of Wisconsin is 0.665 minutes.

Measure of the spread:

The standard deviation of the police response time for homicide in the state of

Wisconsin is 0.4102695590841601 minutes.

The variance of the police response time for homicide in the state of Wisconsin is 0.16832111111111112 minutes.

The maximum police response time for homicide in the state of Wisconsin is 1.94 minutes.

The minimum police response time for homicide in the state of Wisconsin is 0.11 minutes.

The range is 1.47 minutes.

Quantiles and Percentages:

The Quantile was measured for the police response time

The 25th percentile for the police response time is 0.64 minutes.

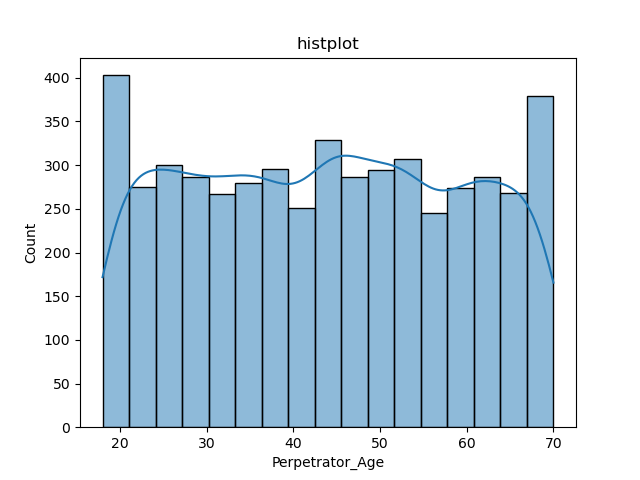
The 35th percentile for the police response time is 0.84 minutes.

The 50th percentile for the police response time is 0.96 minutes.

The 75th percentile for the police response time is 1.47 minutes.

The 95th percentile for the police response time is 1.88 minutes.

Shape of Distribution:



The value of skew for the perpetrator age is 0.013900634798278463. As the value lies

between -0.5 and 0.5, the distribution is approximately symmetric. It also shows zero

skewness. It is also known as a “symmetric distribution or normal distribution”. It

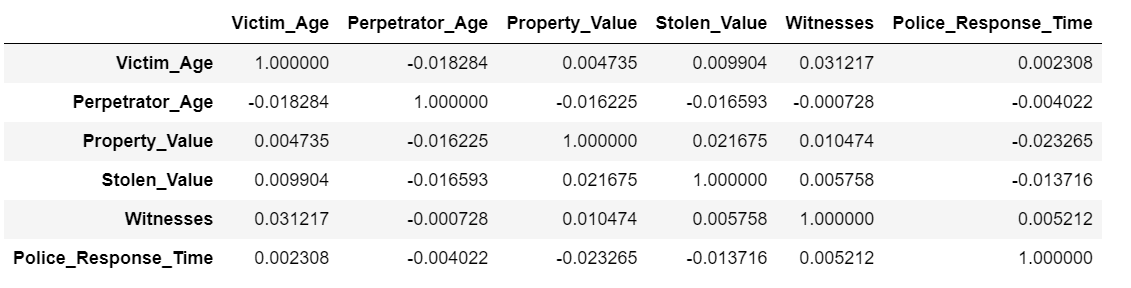
signifies that distribution of data is evenly distributed around the mean, with no long

tails on either end of the distribution. (Mean>Median>Mode). The value of kurtosis is -1.1858589590062802. It shows a slightly platykurtic type of kurtosis. Platykurtic

having a thin tail and stretched around the centre means most data points are present

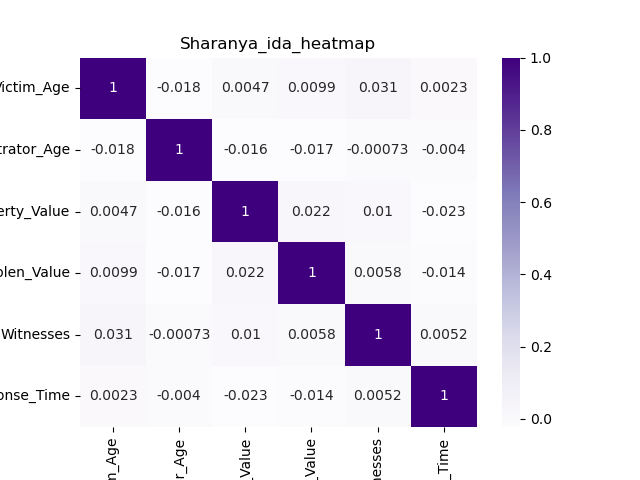
in high proximity to the mean. This is negative kurtosis.

Correlation Matrix and Heatmap:



The correlation matrix, presented in the table, illustrates the pairwise correlations between the variables under consideration: Victim age, Perpetrator age, Property Value, Stolen Value, Witnesses and Police Response Time.

The correlation coefficient between most of the columns is very weak and is statistically irrelevant. The weak correlation suggests that there is no clear linear relationship between the columns. The data is significantly not related to each other.



The heatmap provides a graphical representation of the correlation matrix, illustrating the strength and direction of relationships between the columns. Upon careful examination of the heatmap, it is evident that there are no prominent patterns or discernible relationships between the columns. The colours on the heatmap, indicating correlation coefficients, do not exhibit a consistent trend or clustering.

Many cells in the correlation matrix appear neutral, indicating weak or no associations between the columns. The absence of strong colours suggests that changes in one variable are not consistently associated with changes in another. The heatmap analysis reveals a lack of discernible relationships between the variables in the dataset. This suggests that, based on the data examined, the columns do not exhibit a consistent linear correlation pattern.

Data Encoding:

The purpose of data encoding is to prepare numerical features for machine learning models. By scaling and standardizing these features, we ensure that they contribute equally to the learning process, preventing any single feature from dominating the model's training.

The purpose of data encoding is to prepare numerical features for machine learning models. By scaling and standardizing these features, we ensure that they contribute equally to the learning process, preventing any single feature from dominating the model's training.

Normalization:

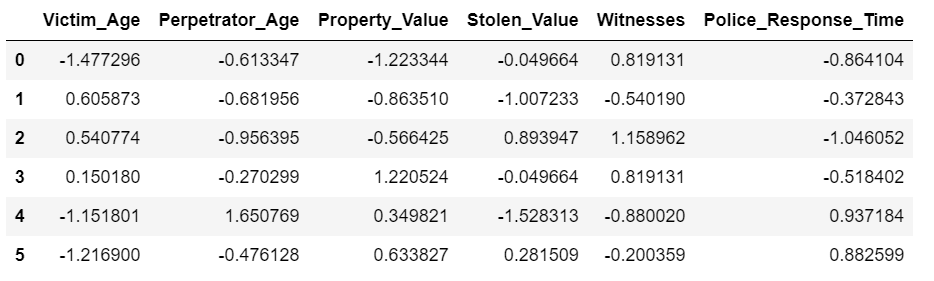
For normalization, we utilized the Min-Max scaling technique. This involves transforming the values of each feature to a fixed range, typically between 0 and 1.

Normalization was implemented using the MinMaxScaler from the sklearn. preprocessing library in Python. All the numerical columns were independently normalized to a range between 0 and 1.



For standardization, we employed the Z-score standardization technique. This involves transforming the values of each feature to have a mean of 0 and a standard deviation of 1.

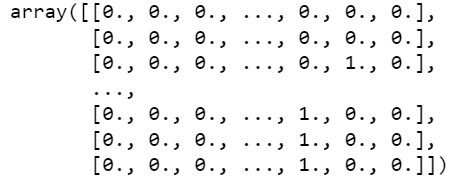
Standardization was implemented using the StandardScaler from the sklearn. preprocessing library in Python. All the numeric columns were independently standardized.

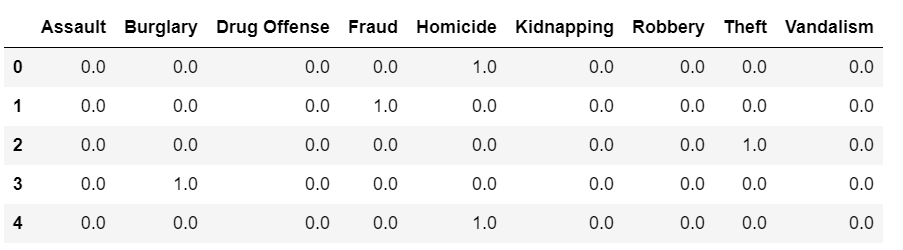


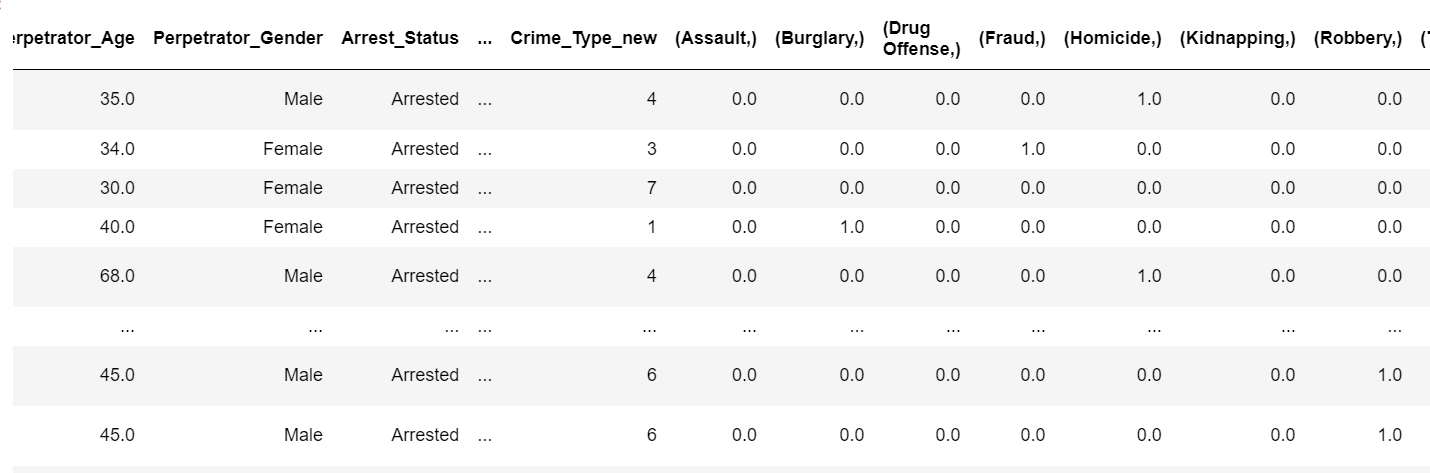
Standardization centres the data around a mean of 0 and scales it by the standard deviation. As the dataset normally distributed, standardization can be a suitable choice.

Categorical Encoding:

Encoding of categorical column in the dataset, specifically addressing the "Crime Type" feature is done. The objective is to prepare the categorical data for machine learning models. The crime dataset comprises various features, including the categorical feature "Crime\_Type," which enumerates different types of crimes, such as "Robbery," "Burglary," and "Assault."



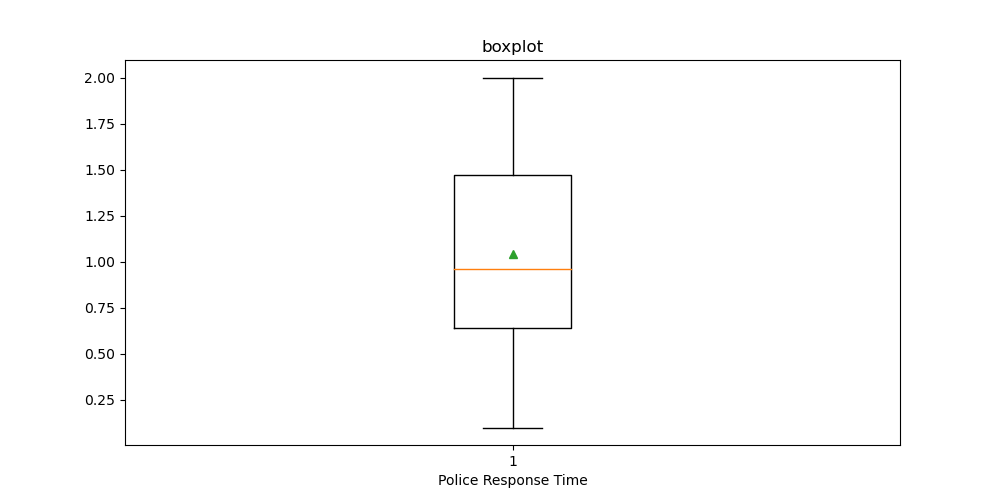




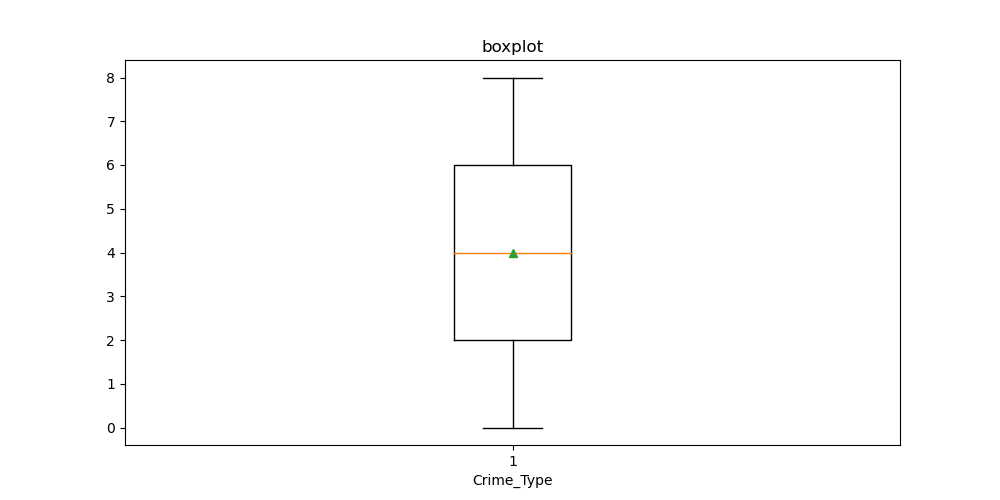
The OneHotEncoder from the sklearn.preprocessing library was utilized. Each category in the "Crime\_Type" feature was assigned a unique binary column. The resulting one-hot encoded data was converted to a NumPy array using the toarray() method. And was joined to the data frame. One-hot encoding eliminates the ordinal relationship between crime types and ensures that the machine learning model does not interpret any inherent order. It is particularly useful when there is no ordinal relationship among categories, in this case. The "Crime\_Type" feature was encoded using cat.codes, resulting in a new column of integer labels.

Visual Representation of Data:

Boxplot:

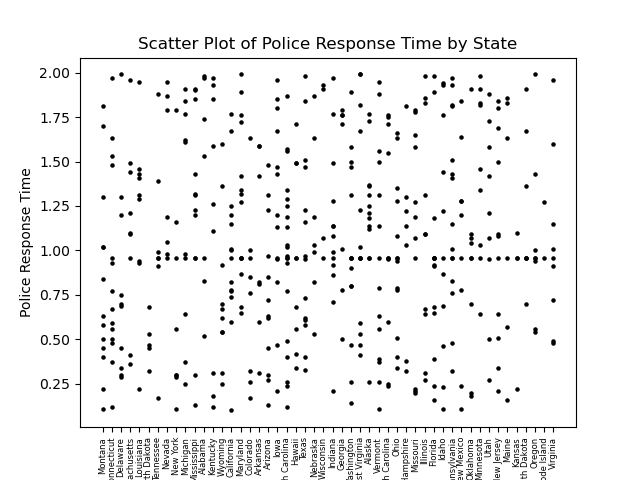


The above boxplot is plotted for Police Response Time. The above boxplot has no outliers and highlights the median. The police response time is not extreme as we have no outliers in this boxplot.



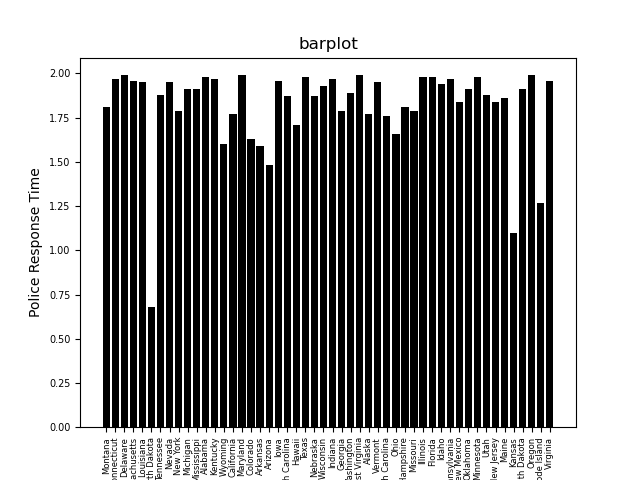
The above boxplot shows the distribution for the crime type data after scaling and categorical encoding. There are no outliers and the median is highlighted. The Crime Type data is not extreme as we have no outliers in this boxplot.

Scatter Plot:

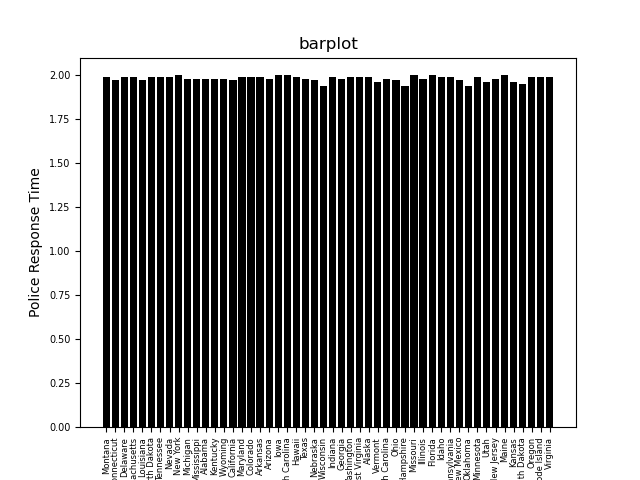


The above scatter plot is plotted for State vs Police Response Time. State is in the x axis while Police Response Time is taken in the y axis. Only the first 500 rows were taken into consideration when plotting this scatter plot. We can see that the distribution is scattered most around 1.00, indicating most states have a police response time around that range.

Barplot:



The barplot represents police response time for different states. This barplot was plotted only for the first 500 rows. From the barplot we can see that the state of Dakota has the best police response time, in the first 500 columns of the data.



The barplot represents police response time for different states. This barplot was plotted for all the 5028 rows. From the barplot we can see that the state of Wisconsin has the best police response among all the US states.

Conclusion:

* From the analysis we can see that the state of Wisconsin has the best police response time for homicide
* If we consider the first 500 rows, the state of Dakota has the best police response time