TASK – L2T18

Exploratory Data Analysis on the “Credit Card Fraud Detection" dataset

INTRODUCTION

Credit card fraud is a widespread problem that affects both credit card issuers and cardholders. It involves the use of stolen or counterfeit credit card information to make unauthorized purchases or cash withdrawals. In this EDA, I will be analyzing a dataset from Kaggle that contains credit card transactions made by European cardholders to explore the relationships between the variables and identify patterns that can help detect fraudulent transactions.
The aim is to extract stories and assumptions based on the visualizations of the data.

Summary of the data set:

The dataset contains 31 columns for each instance, 28 of which have been anonymized (To preserve private, confidential & sensitive information) and are labeled V1 through V28, while the remaining three columns contain time and amount features.

Anonymized features generally include: Names | Addresses | Phone Numbers | IP Address | Email | Geolocation | etc.

DATA CLEANING & PROCESSING:

Data cleaning:

The EDA starts with data collection and loading, followed by a summary of the dataset using the head() and describe() functions to get an idea of what the data looks like.

During data cleaning, I checked for missing data.
It is important to handle the missing values appropriately.
Many machine learning algorithms fail if the dataset contains missing values.
You may end up building a biased machine learning model, leading to incorrect results if the missing values are not handled properly.
Missing data can lead to a lack of precision in the statistical analysis.

I also conducted some visualizations to better understand the data.

Summary of the methods done during data cleaning:

During data cleaning, I used the isnull() and sum() methods.

MISSING DATA:

Missing data is defined as *the values or data that is not stored or not present for some variable/s in the given dataset.*

Missing Data can be of three types:

1] Missing Completely At Random (MCAR)

2] Missing At Random (MAR)

3] Missing Not At Random (MNAR).

In this specific dataset there happened to be no missing data, hence no further action was required. Though normally, Missing data can be handled in various ways;

- Dropping / Deleting the Missing Value(s):
  - Please note that this approach is not recommended especially for (MNAR) type.
  - If the missing value is of type Missing At Random (MAR) or Missing Completely At Random (MCAR) then it can be deleted.

- Filling the missing values with appropriate data based on the data type (Imputation).
  - Replace with: Arbitrary value, mean, mode, median, forward fill (previous value), backward fill (next value) or Interpolation methods {Polynomial, Linear [Default], Quadratic}.

CODE FOR THE VISUALIZATIONS:

To visualize the data, I used Jupyter Notebook and I imported the necessary libraries such as:

Numpy | pandas | seaborn | matplotlib.pyplot.

I then loaded the data set into the pandas dataframe using pd.read_csv.

To gain insights into the data, various visualizations were conducted using;

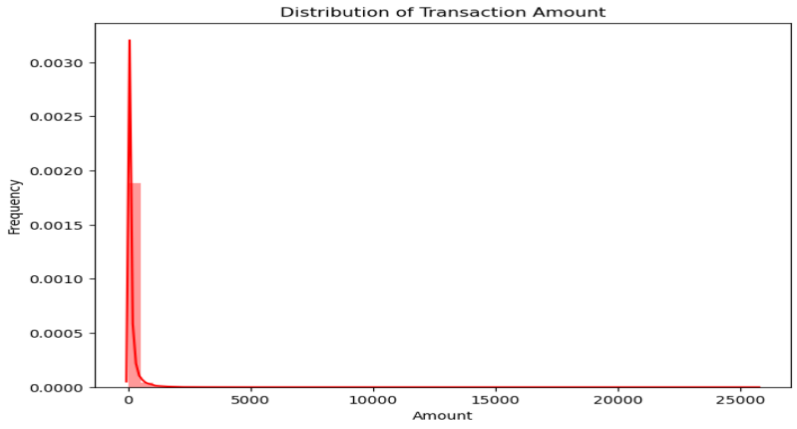
sns.displot (Histogram), sns.scatterplot, sns.heatmap, sns.boxplot,
sns.kdeplot (kernel density estimate) and plt.show.

DATA STORIES AND VISUALISATIONS:

(A) Distplot:

The univariate analysis begins with a visualization of the distribution of transaction amount using a histogram created using the distplot function from the seaborn library.
The graph shows the frequency of transaction amounts, with the x-axis representing the transaction amounts and the y-axis representing the frequency of transactions.

The second histogram shows that most transactions are for small amounts, with very few transactions for amounts greater than $2,000. Further analysis is done by calculating and displaying additional statistics such as mean, median, and standard deviation. The mean transaction amount is around $88, while the median is $22, indicating that the distribution is skewed to the right.



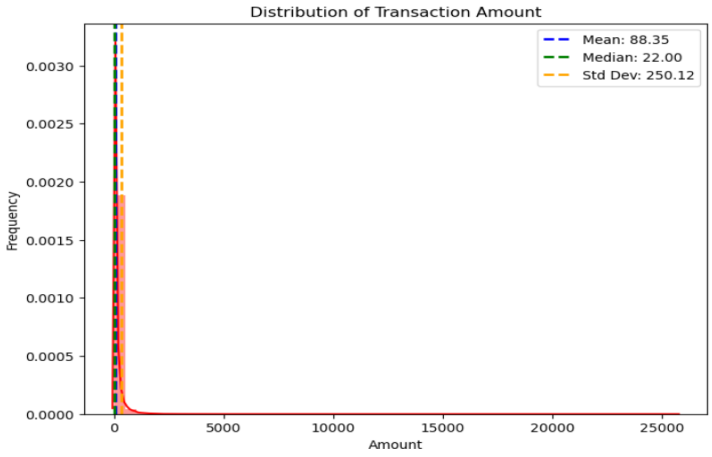
From the above visual, the distribution is highly skewed with a long tail to the one side.

This suggests that most transactions are relatively small, with the frequency of transactions decreasing as the transaction amount increases.

Added Stats:

The below visualization shows the distribution of transaction amounts along with the mean, median, and standard deviation. The blue dashed line represents the mean amount. The green dashed line represents the median amount. The orange dashed line represents one standard deviation away from the mean.

By adding these statistics to the visualization, we can better understand the central tendency and spread of the transaction amounts.



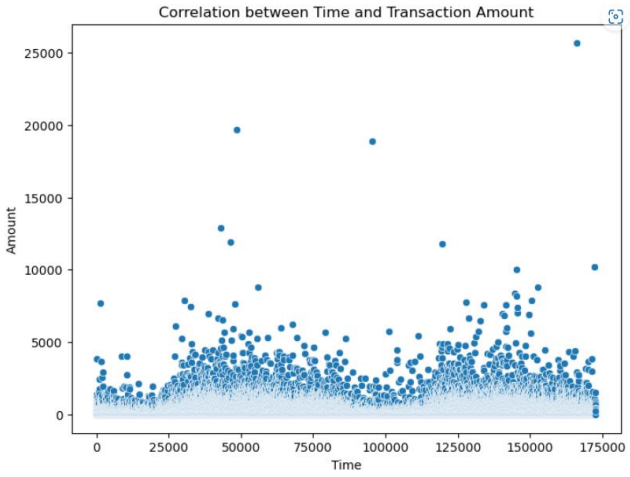
Understanding the central tendency (such as mean or median) and spread (such as standard deviation or range) of transaction amounts is important because it provides a summary of the distribution of the data. This information helps to identify the typical or average transaction amount, as well as the variability or dispersion around that central value.

By understanding the central tendency and spread of transaction amounts, one can detect outliers or extreme values that may indicate fraudulent activity or errors in the data. It can also help to identify potential trends or patterns in the data over time or across different subsets of the data.

Central tendency and spread of transaction amounts is important for making informed decisions in areas such as risk management, fraud detection, and financial forecasting.

(B) Scatterplot:

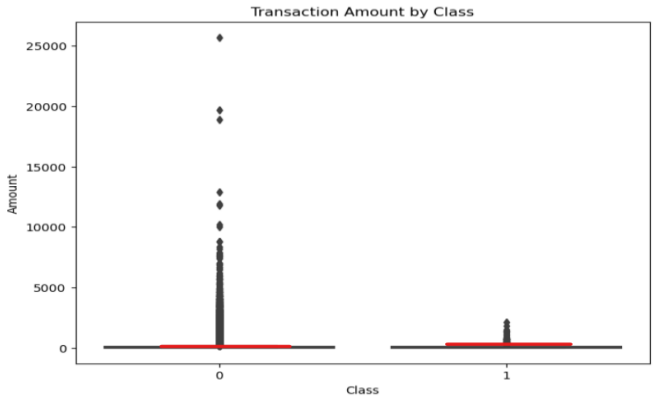
The bivariate analysis starts with a scatter plot of transaction amount versus time, aimed at visualizing the correlation between the two variables. The plot shows no clear correlation between the two variables.



From the above plot, there doesn't seem to be a clear linear relationship between time and transaction amount. While there are a few high transaction amounts that occur at various points in time, there is no clear trend in the data. Though, this plot does provide us with a good starting point to identify trends / patterns over time.

(C) BOXPLOT:

- A boxplot is a type of visualization that shows the distribution of a numerical variable (in this case, transaction time) by grouping the data based on a categorical variable (in this case, fraudulent status).
- The box in the plot represents the interquartile range (IQR), which is the range between the 25th and 75th percentiles of the data. The whiskers extending from the box represent the range of the data within 1.5 times the IQR. Any data points beyond the whiskers are considered outliers and are plotted as individual points.
- In fraud detection, typically you will find class 0 represents non-fraudulent transactions, and class 1 represents fraudulent transactions.



The above Boxplot was created to show the distribution of transaction amounts by class (fraudulent vs. non-fraudulent) to compare the distribution of amounts between the two classes. This can help identify if there are any significant differences in the transaction amounts for fraudulent vs. non-fraudulent transactions. The x-axis of the boxplot represents the two classes of transactions: non-fraudulent (Class 0) and fraudulent (Class 1). The y-axis represents the transaction amount.

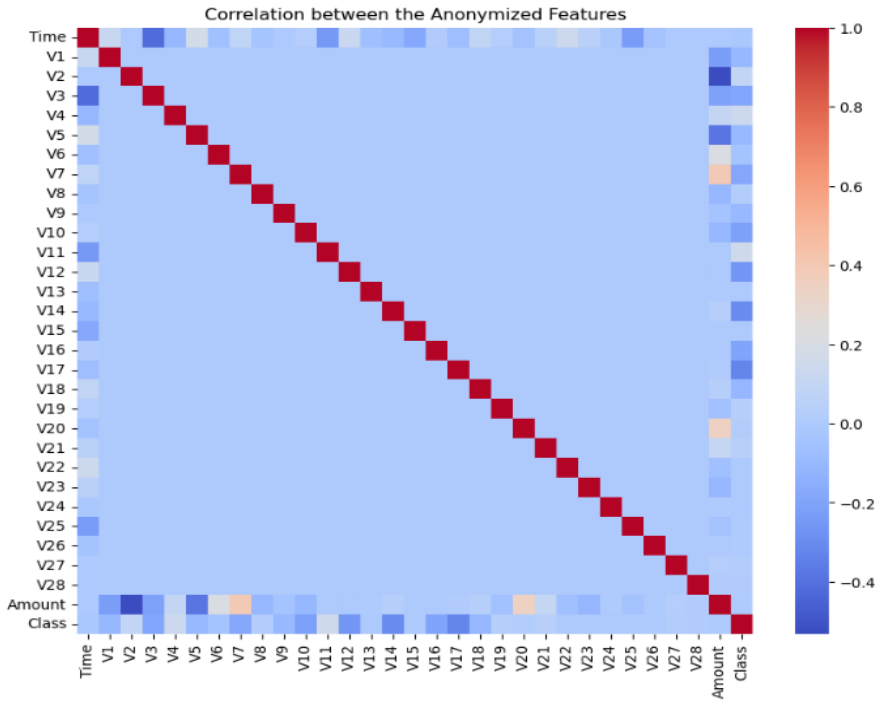
The box itself is divided into three parts: the bottom of the box represents the 25th percentile of the data, the middle line represents the median (50th percentile), and the top of the box represents the 75th percentile.

The whiskers extending from the box represent the range of the data within a specified limit. The limits

of the whiskers are usually set to 1.5 times the interquartile range (IQR), which is the distance between the 25th and 75th percentile. Any points beyond the whiskers are considered outliers and plotted individually as small circles.

In this plot, we can see that the median transaction amount for fraudulent transactions (Class 1) is slightly higher than for non-fraudulent transactions (Class 0). The distribution of transaction amounts for fraudulent transactions has a higher spread than the distribution for non-fraudulent transactions.
The plot shows that the median transaction amount for fraudulent transactions is higher than that of non-fraudulent transactions, indicating that fraudulent transactions tend to be for larger amounts. This suggests that “transaction amount” may be a useful feature for detecting fraud.

(D) HEATMAP:



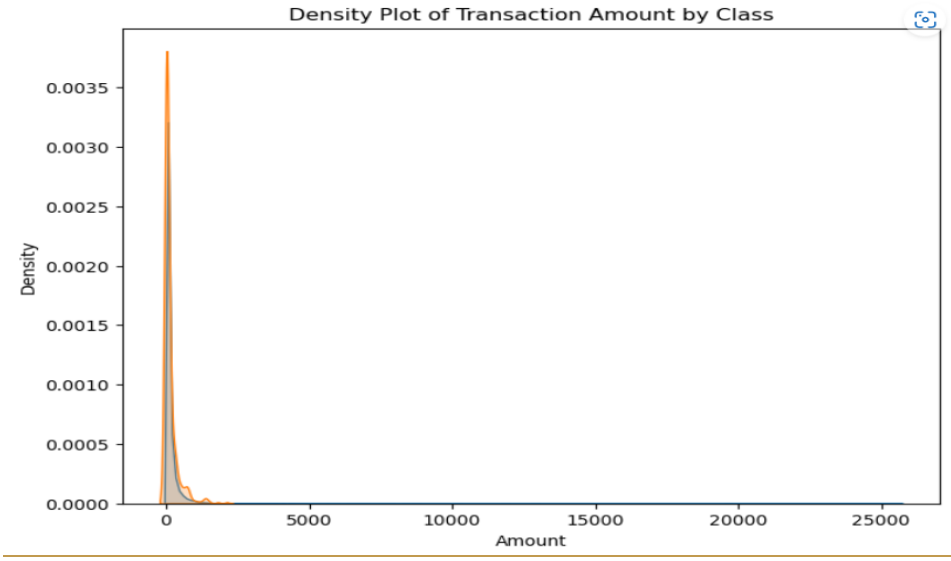
The multivariate analysis includes a heatmap of the correlation matrix to visualize the pairwise correlation between all variables in the dataset.
This can help identify any strong correlations between variables, which can be useful for feature engineering or variable selection.

The heatmap above shows that there are no strong correlations between the anonymized features. This suggests that they are all independent, and each one may provide useful information for detecting fraudulent transactions.

Commented [U1]:

(E) DENSITY PLOTS (KDE – Kernel Density Estimate)

Density plots are created for the transaction amounts broken down by class for comparison of the distribution of amounts between the two classes.
This can help identify if there are any significant differences in the shape of the distribution for fraudulent vs. non-fraudulent transactions. The plots show that the distribution of transaction amounts for fraudulent transactions is more spread out and has a longer tail than the distribution for non-fraudulent transactions. The x-axis represents the transaction amount, and the y-axis represents the density or frequency of that amount.

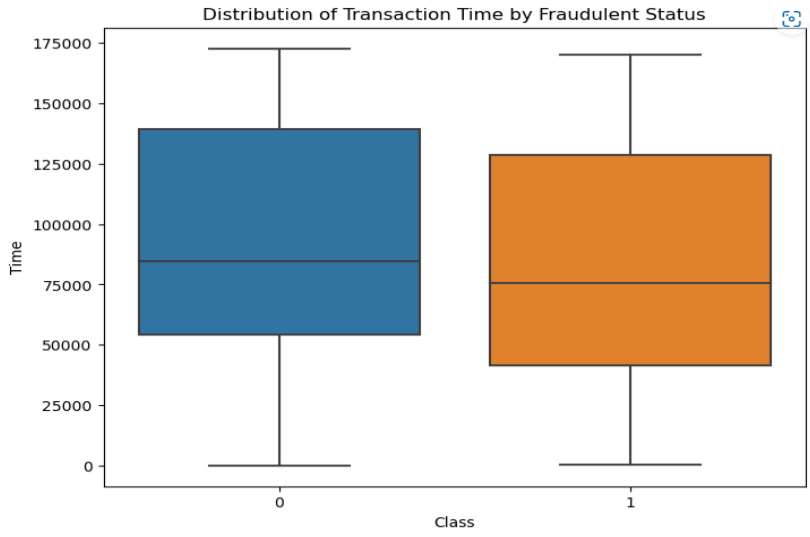


The above graph shows that the density plot for fraudulent transactions has a much larger peak and a longer tail than the density plot for non-fraudulent transactions. This suggests that fraudulent transactions tend to have higher transaction amounts compared to non-fraudulent transactions.

Also, the graph also shows that fraudulent transactions have a larger variance in transaction amount compared to non-fraudulent transactions. Therefore, this visualization highlights the usefulness of exploring the data by class, as it helps identify differences between fraudulent and non-fraudulent transactions.

c) Boxplot:

To Visualize the distribution of time for fraudulent and non-fraudulent transactions.



In this particular graph, the x-axis represents the two possible classes for the fraudulent status of transactions (0 for non-fraudulent and 1 for fraudulent), while the y-axis represents the time of each

transaction. The box for each class shows the IQR, while the whiskers extend to show the range of the data, and any individual points beyond the whiskers (If any) are plotted (individually) as outliers.

There appears to be some difference in the distribution of transaction times between fraudulent and non-fraudulent transactions in the boxplot when looking at the median transaction time and could potentially indicate fraudulent activity occurring during unusual times or intervals as well. Therefore, this suggests that "transaction time" may be a useful feature for detecting fraud.

CONCLUSION:

The exploratory data analysis (EDA) is focused on understanding the distribution and relationship between variables in the creditcard.csv dataset. The dataset contains information about credit card transactions, including transaction amount, time, and whether the transaction is fraudulent or not. The dataset contains 284,807 transactions, out of which 492 (0.17%) are fraud transactions.

The analysis of the dataset shows that most transactions are for small amounts, with very few transactions for amounts greater than $2,000. Fraudulent transactions tend to be for larger amounts than non-fraudulent transactions. There is no clear correlation between transaction amount and time. The heatmap of the correlation matrix shows that most variables are weakly correlated with each other, with the exception of some variables that are moderately correlated with time. The distribution of transaction amounts for fraudulent transactions is more spread out and has a longer tail than the distribution for non-fraudulent transactions.

The EDA of the credit card fraud dataset has helped us to identify some patterns that can help detect fraudulent transactions. The analysis shows that fraudulent transactions tend to be for larger amounts than non-fraudulent transactions, and the distribution of transaction amounts for fraudulent transactions is more spread out and has a longer tail than the distribution for non-fraudulent transactions. There is no clear correlation between transaction amount and time, and there are no strong correlations between the anonymized features.

RECOMMENDATIONS:

Based on the analysis, I have identified a few features that may be useful for detecting fraudulent transactions, I recommend the following:

- Credit card issuers should monitor transactions for larger amounts more closely, as these are more likely to be fraudulent.
- Fraud detection algorithms should take into account the shape of the distribution of transaction amounts, as fraudulent transactions tend to have a longer tail.
- Feature engineering techniques should be considered to better capture potential patterns and correlations between the variables.
- Using a combination of different algorithms such as decision trees, random forests, and neural networks to build a robust fraud detection system.
- Central tendency and spread of transaction amounts are important for making informed decisions in areas such as risk management, fraud detection, and financial forecasting.
- (1) Transaction Amount, (2) Anonymized Features, (3) Transaction Time, (4) Class Exploration

  may provide useful information in detecting fraudulent transactions when developing a fraud detection model.

THIS REPORT WAS WRITTEN BY:
Sameera Mohamed