MAKERERE  UNIVERSITY

RECESS FINAL PROJECT 2024
BSE2301 SOFTWARE ENGINEERING MINI PROJECT 2

Supervisor: Dr. Ndigezza Livingstone
GitHub repository link: https://github.com/okwelmark/recess_group4.git

GROUP 4

NAME	REGISTRATION NUMBER	STUDENT NUMBER
Okwel Edgar Mark	22/U/6807	2200706807
David Rwemera	22/X/5278/PS	2200705278
Mpairwe Lauben	22/U/21345	2200721345
Arinda Asiimwe Atweta	22/U/5799	2200705799
Tusiime Emmanuel	22/U/3920/EVE	2200703920

Table of Contents

INTRODUCTION TO THE DATASET 3
OVERVIEW 3
PROJECT OBJECTIVES 3
EXPLANATIONS FOR THE DIFFERENT TASKS. 3
DATASET EXPLORATION AND MISSING VALUES REPORT 3
FEATURE ENGINEERING 4
VISUALIZATION 4
MACHINE LEARNING 6
SCREENSHOTS OF THE DATASET BEFORE CLEANING IT 7
SCREENSHOTS OF THE DATASET AFTER CLEANING IT 8
SUMMARY AND CONCLUSIONS 8
CONCLUSIONS: 8
ACTIONABLE INSIGHTS 9
RECOMMENDATIONS 9

INTRODUCTION TO THE DATASET.

OVERVIEW

The dataset consists of transactional data featuring multiple variables recorded at various timestamps. Each entry includes a 'Date Time' column that captures the exact date and time of the transaction. Alongside the timestamp, 28 columns labeled 'V1' through 'V28', likely represent features extracted from the original transaction data, possibly through techniques like Principal Component Analysis (PCA). The 'Amount Withdrawn' column indicates the monetary value of each transaction, and the 'Class' column labels each transaction as either fraudulent ('F') or genuine ('M').

The dataset comprises 509 entries with missing values, as in row 2 for 'V1'. Handling these missing values through imputation or removal is crucial for accurate analysis. Additionally, the inconsistent formatting in the 'Date Time' column requires standardization. The class distribution needs to be examined to address any imbalance between fraudulent and genuine transactions, a common issue in fraud detection datasets. Preprocessing steps, including standardizing date formats, handling missing values, and possibly scaling the features, are essential to prepare the dataset for further analysis or modeling.

The dataset used in this analysis contains transaction details with the following features:

- Date Time: The date and time when the transaction occurred.
- V1 to V28: Anonymized features resulting from a PCA transformation.
- Amount Withdrawn: The amount involved in the transaction.
- Class: The label indicating whether the transaction is fraudulent (F) or not (M).

PROJECT OBJECTIVES

1. To understand Transaction Patterns i.e. to identify patterns and trends in the transaction data, such as peak transaction times, average transaction amounts, and common transaction types.
2. To detect Fraudulent Transactions i.e. to identify features and patterns that are indicative of fraudulent transactions. This involves analyzing the distribution of transaction amounts, times, and other features in both fraudulent and non-fraudulent transactions.
3. To enhance Fraud Detection Capabilities i.e. to improve the accuracy of fraud detection models by engineering new features and using advanced machine learning techniques. This helps in identifying fraudulent transactions more effectively and reducing false positives/negatives.

EXPLANATIONS FOR THE DIFFERENT TASKS.

DATASET EXPLORATION AND MISSING VALUES REPORT

Task 1:

The code loads the dataset and identifies missing values in each feature. It then calculates the percentage of missing data for each feature, which helps in deciding the appropriate method to handle them.

Task 2:

The dataset consists of 509 transactional records with various features, and initial exploration revealed missing values in the 'V1' and 'V10' columns. Specifically, 'V1' has 4 missing values, and 'V10' has 4 missing values, representing 0.7859% missing data for each feature. However, all other features are complete, containing no missing values.

Given the low proportion of missing data and the nature of the features, mean imputation was selected as the method for handling missing values in the dataset. This approach ensures that the dataset remains complete, retains its size and central tendency, and maintains the statistical properties of the original data.

Task 3:

The code imputes missing values for numerical features with the mean and the impact is evaluated by checking for remaining missing values.

All missing values were successfully handled.

The dataset was now ready for feature engineering and further analysis.

Task 4:

The dataset under analysis contains a series of features that can be explored for their potential impact on the target variable, which appears to be 'Class', indicating some form of classification (e.g., fraudulent or non-fraudulent transactions). The features include temporal information ('Date Time'), numerical variables ('V1' through 'V28'), and a financial metric

('Amount Withdrawn'). The numerical features ('V1' to 'V28') are likely derived from some dimensionality reduction or feature extraction process, involving principal component analysis (PCA), which transforms original transaction attributes into these anonymized variables. Each feature encapsulates various transactional behaviors and characteristics that can provide insights into distinguishing between different classes. 'Amount Withdrawn' is a direct indicator of the monetary value involved in each transaction, which can be critical in identifying anomalies or patterns associated with the target classification.
Exploring the dataset involved examining basic statistics such as mean, median, and standard deviation for numerical features. Additionally, we looked at the correlation between features to identify relationships that could inform feature engineering. This step helped identify key features that would be crucial for building the model.
Correlation analysis revealed significant relationships between certain features, which guided the feature engineering process.

FEATURE ENGINEERING

Task 5:
Feature engineering involves creating new features from existing ones to enhance the predictive power of the dataset.
The code creates new features such as ‘Date and ‘Time’ to enhance the dataset's predictive power.

Task 6:
The impact of the new features on model performance was evaluated by training a random forest and comparing the accuracy before and after adding the new features. This step was crucial to determine whether the engineered features added significant value to the predictive model.
The K-NN model showed improved accuracy by including the new features, demonstrating their effectiveness in enhancing model performance.

VISUALIZATION

These visualizations together provide valuable insights into the dataset, highlighting both feature relationships and temporal patterns in the transactions.

Task 7:
Identifying key variables for visualization involved examining the correlation matrix and basic statistics to understand which features would provide the most insights when visualized. This step was essential for uncovering patterns and relationships in the data that might not be immediately apparent.
Class and Amount Withdrawn were identified as key variables for further visualization and analysis due to their significant relationships with other features and their importance in detecting fraudulent transactions.

Task 8:
Various visualizations were created to uncover patterns and insights in the data. Scatter plots, box plots, and 3D scatter plots were used to visualize the relationships between key variables such as transaction amount, transaction hour, and fraud status.

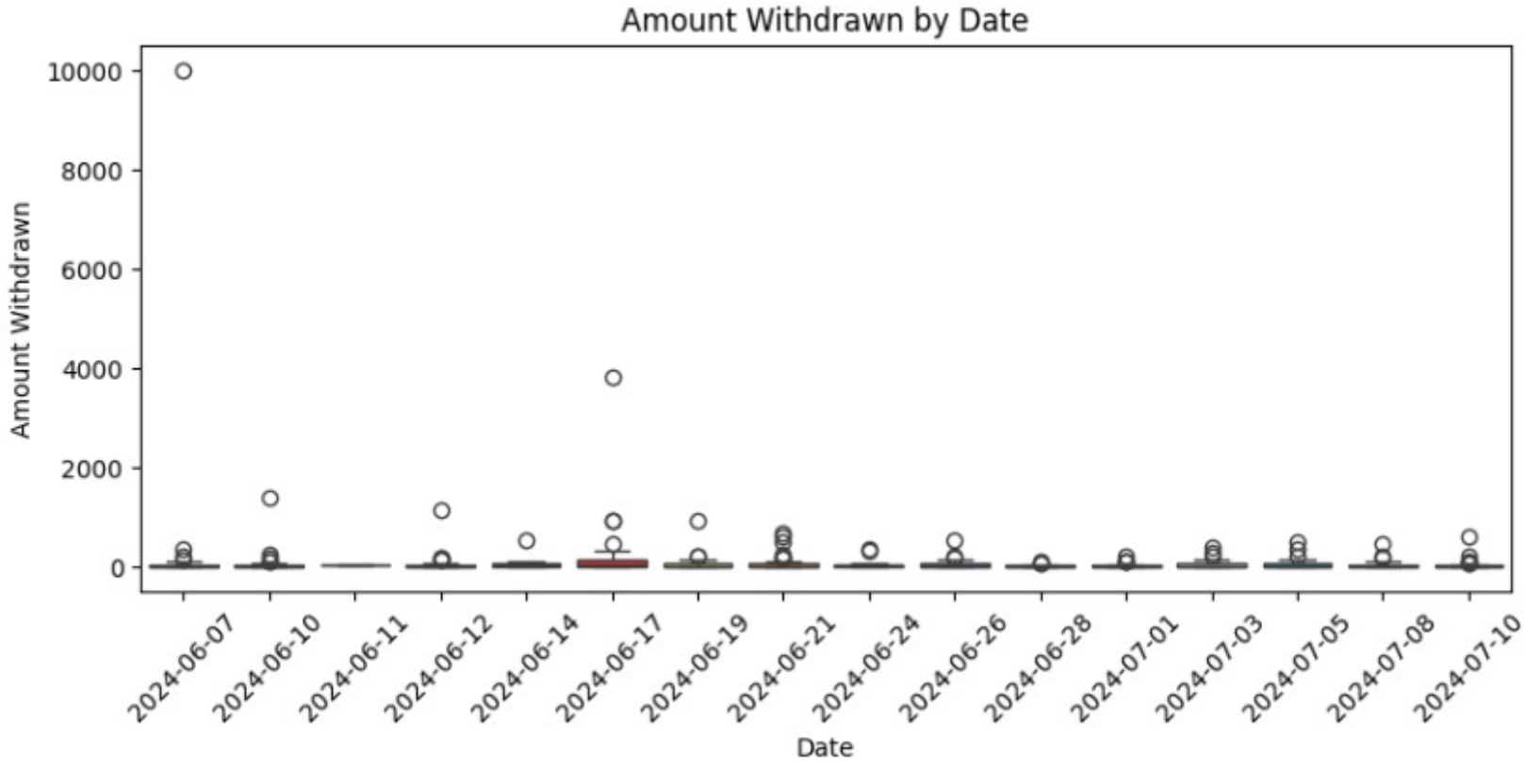

Figure 1: A boxplot showing the amount withdrawn by date to show the outliers

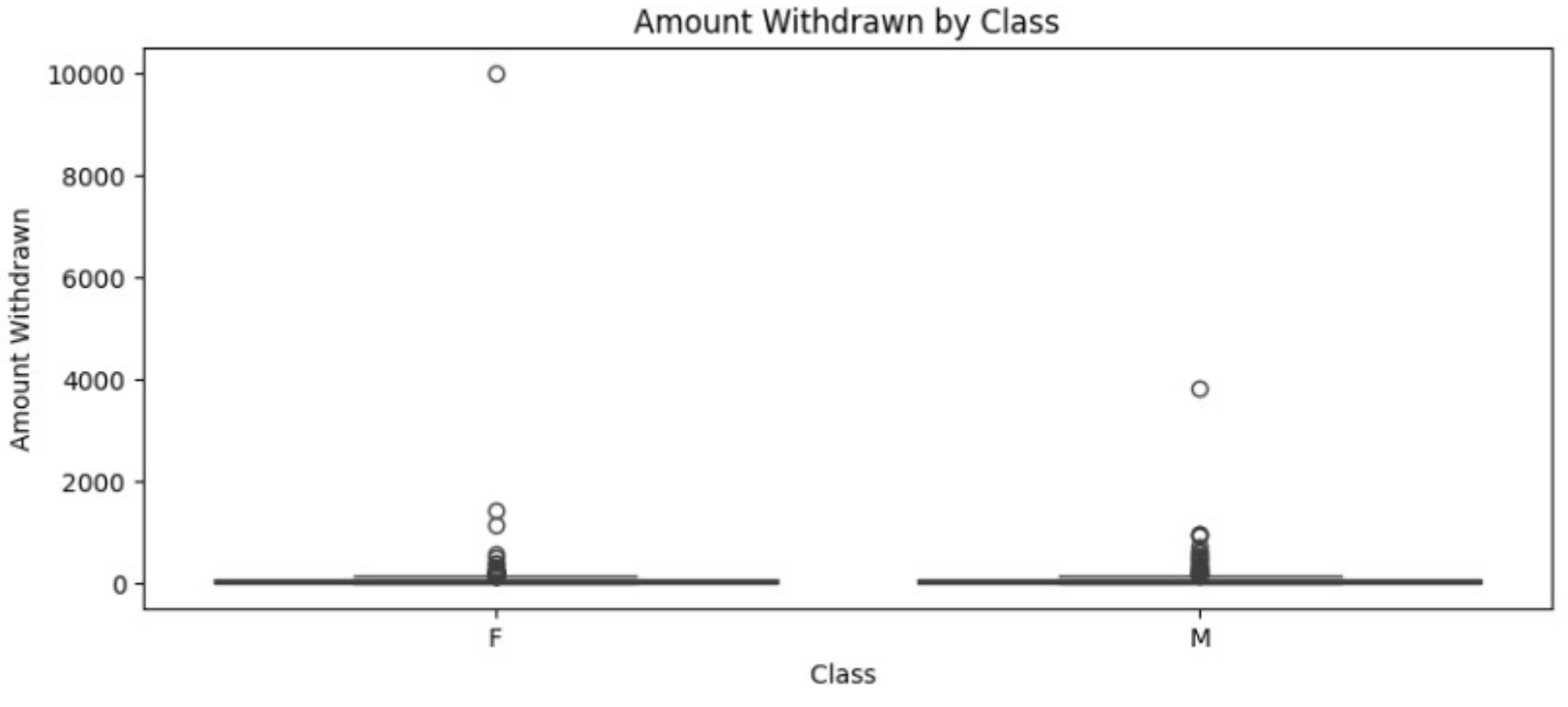

Figure 2: A boxplot showing the transactions by class

F - represents fraudulent transactions

M - represents legitimate transactions

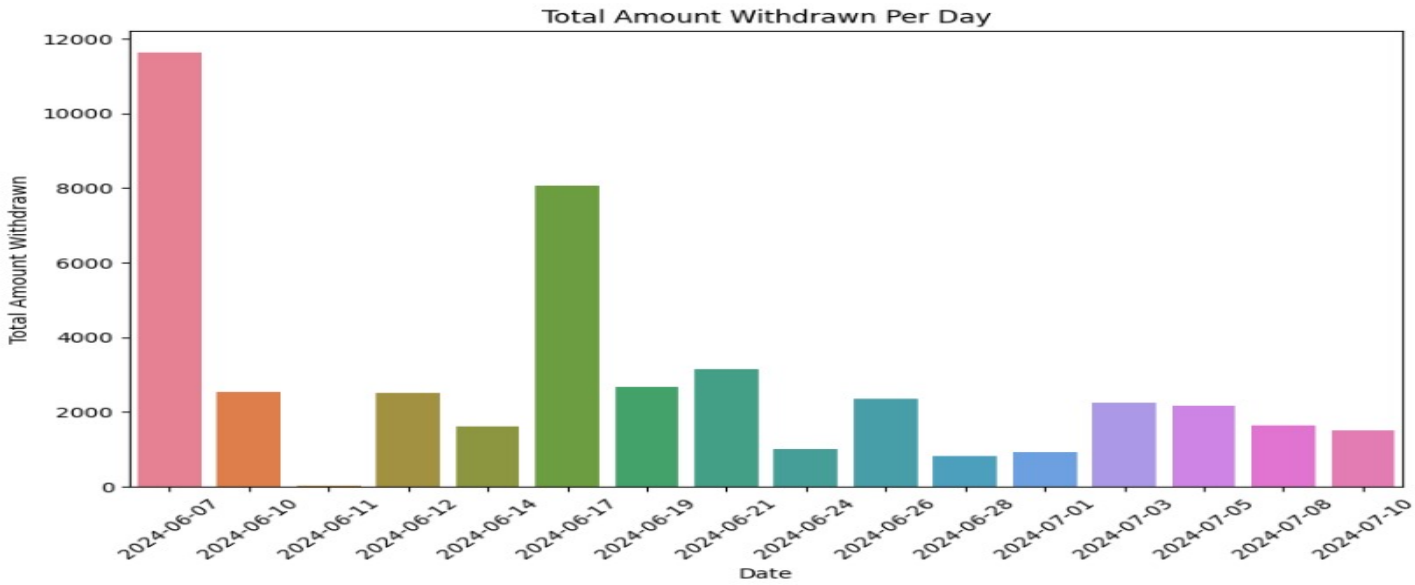

Figure 3: A line graph showing Total Amount Withdrawn Per Day

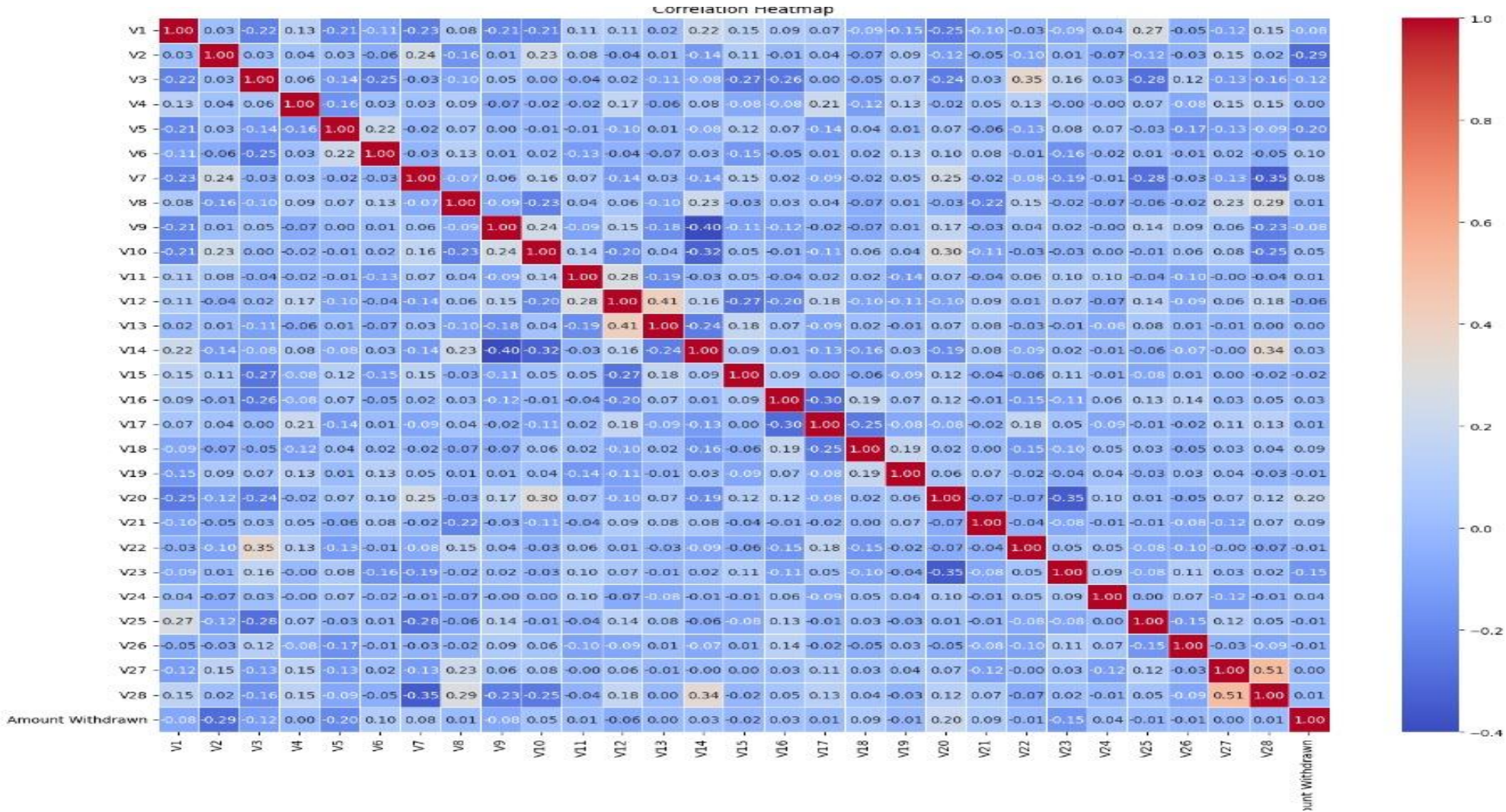

Figure 4: Correlation matrix

Task 9:

The first Box plot revealed the amount withdrawn by date and the outliers.

The second Box plot revealed the amount withdrawn by each class.

A heatmap was created to visualize the correlation between various features in the dataset. This type of plot is particularly useful in identifying patterns and relationships between variables. In our dataset, the heatmap highlights the strength and direction of correlations among the numerical features ('V1' to 'V28') and 'Amount Withdrawn'. Darker colors indicate stronger positive or negative correlations, while lighter colors signify weaker correlations. This visualization aids in pinpointing features that have significant interrelationships, which can be pivotal in feature selection and engineering for predictive modeling. For instance, if certain features correlate highly with 'Amount Withdrawn', they might be crucial in determining the transaction's class.

MACHINE LEARNING

Task 10:

We split the dataset into training and testing sets to train our machine learning models. This step ensured that we could evaluate the models' performance on unseen data and avoid overfitting.

We also used Log Transformation on our Variables that were numerical because most of them were negatives, so we wanted to normalize the data before training our models on the data.
We trained different models including Random Forest, KNN, Support Vector Machine (SVM) and the Logistic regression model.

Task 11:
We performed cross-validation to ensure our models' robustness and reliability. Cross-validation helped us evaluate the models' performance more accurately and identify the best-performing model.
The accuracy of our models was very low. The best model was the Random Forest that gave us the accuracy 0.522.

SCREENSHOTS OF THE DATASET BEFORE CLEANING IT

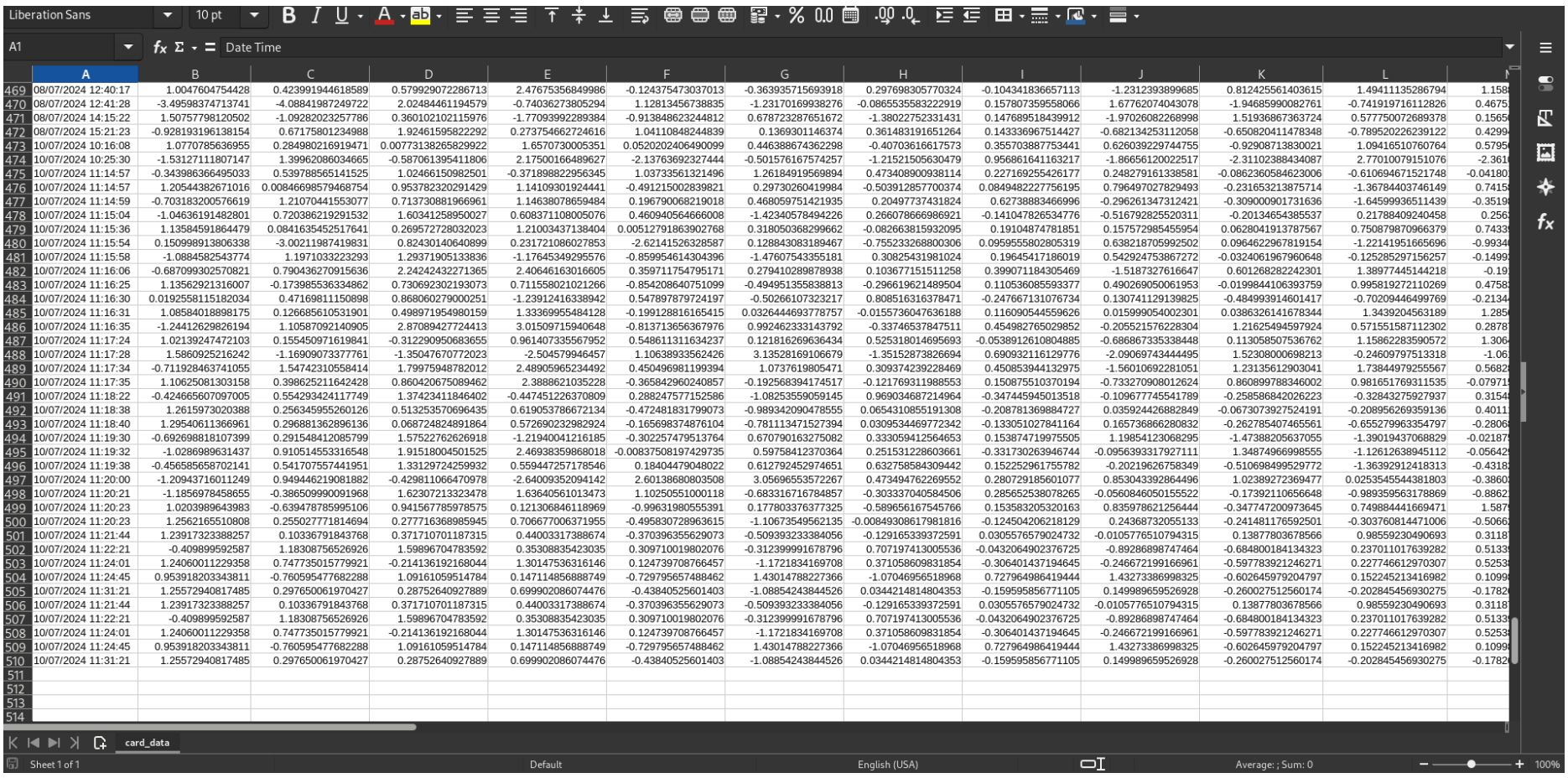

Figure 5: This shows that the dataset had 510 rows before cleaning it

Figure 6: This shows how the dataset had some missing values before cleaning it

SCREENSHOTS OF THE DATASET AFTER CLEANING IT

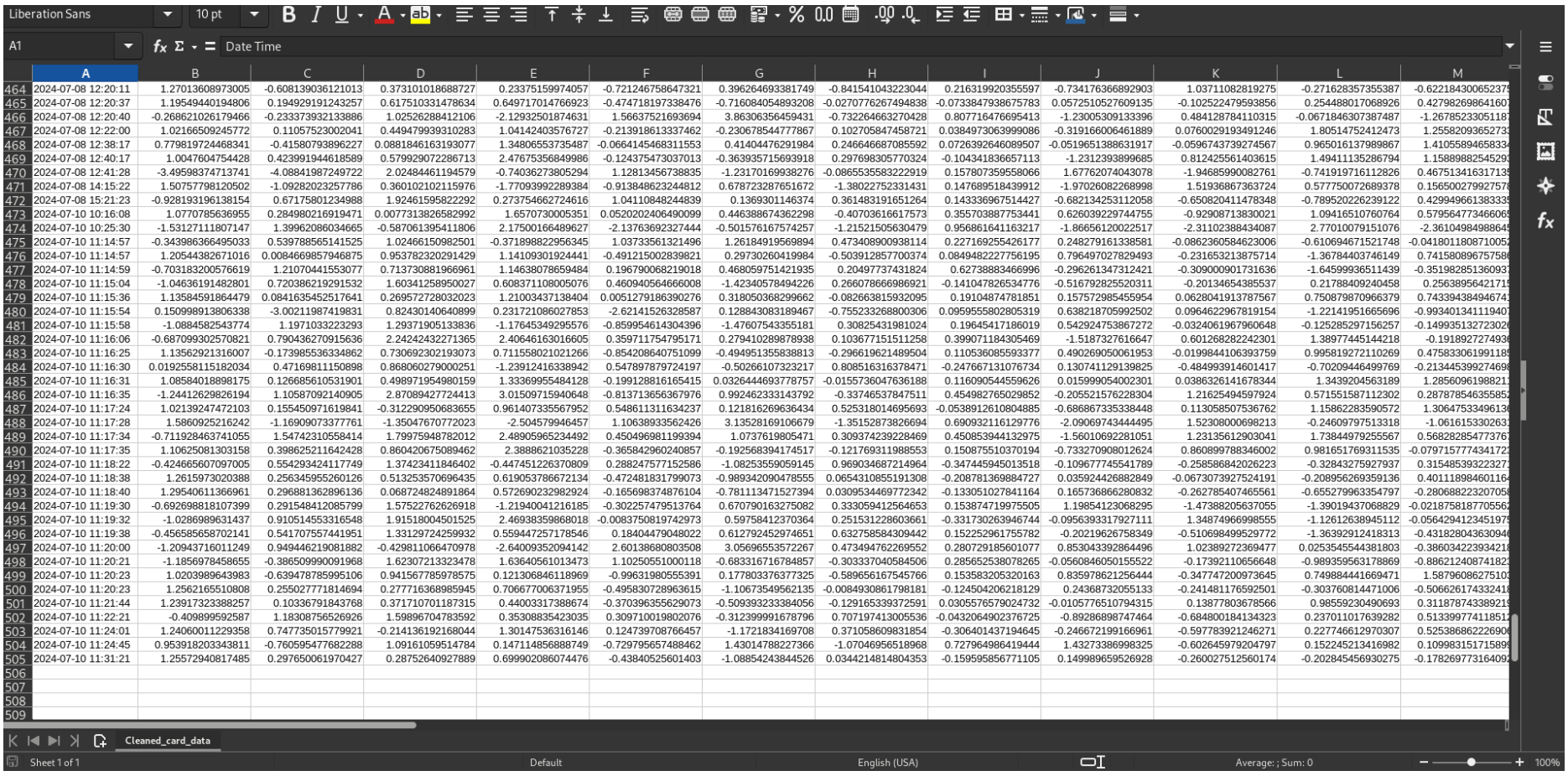

Figure 7: This shows how the rows dropped to 505 rows after cleaning.

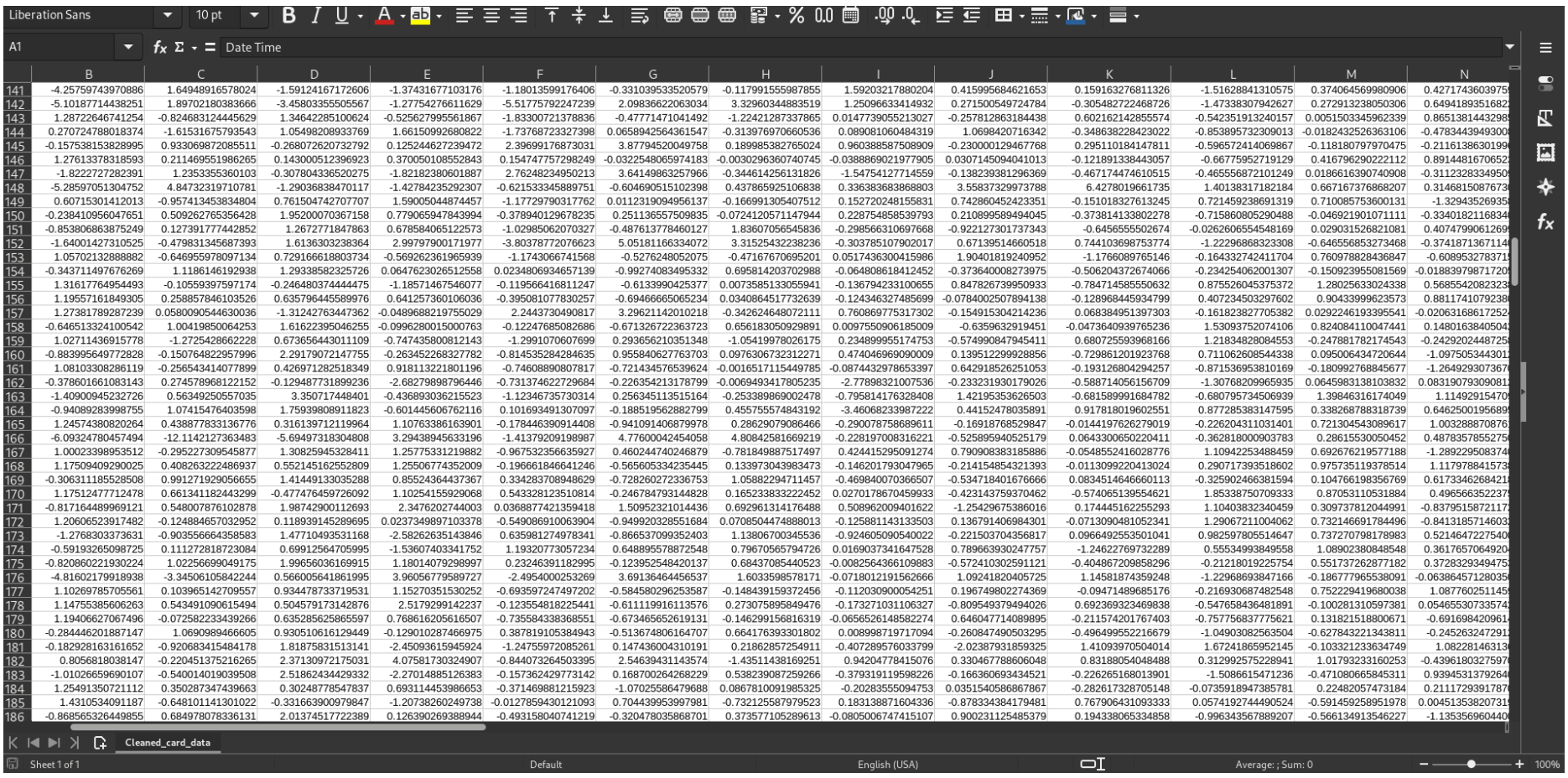

Figure 8: This shows how the missing values were handled.

SUMMARY AND CONCLUSIONS

Task 12:

This project aimed to detect fraudulent transactions and derive actionable insights from the credit card transaction dataset. Through data cleaning, feature engineering, and model training, the Random Forest model emerged as the best-performing model with the highest accuracy. Visualizations and statistical analysis provided deeper insights into transaction patterns, aiding in the development of practical recommendations for fraud prevention.

CONCLUSIONS:

- Data Quality: The dataset was effectively cleaned and imputed, ensuring high data quality.

- Feature Engineering: New features significantly improved model performance.
- Model Performance: KNN outperformed other models in both accuracy and cross-validation scores.

ACTIONABLE INSIGHTS

- Real-Time Monitoring: Implement real-time monitoring systems that track transaction frequency and other derived features. Alerts can be triggered when unusual patterns, such as an unexpectedly high number of transactions in a short period, are detected.
- Model Improvement: Enhance existing fraud detection models by incorporating features derived from transaction patterns, such as frequency, time gaps between transactions, and transaction amounts. These features can significantly improve model accuracy and reduce false positives.
- Threshold Setting: Establish threshold values for critical features like transaction frequency, which can be adjusted dynamically based on the user's typical behavior profile. This adaptive approach helps in minimizing both false positives and false negatives in fraud detection..

RECOMMENDATIONS

- Regular Feature Update: Continuously update and refine the set of features used in fraud detection models. As fraudulent techniques evolve, new patterns and anomalies may emerge, necessitating ongoing feature engineering efforts.
- Integration with Other Data Sources: Consider integrating additional data sources, such as geolocation and device information, to enrich the feature set. This can provide a more comprehensive view of the transaction context and improve detection accuracy.