Assignment 3  
ML Group Project

short line

24 May 2024

Group 38

|  |  |  |  |
| --- | --- | --- | --- |
| Student Last Name | Student First Name | Student ID | Group Allocation |
| Ratana | Sovann | 25396324 | Student A |
| Sivanesan | Balakumaran | 24883302 | Student B |
| Hegde | Chandan | 24667610 | Student C |
| Kalbhore | Sarika | 25186990 | Student D |

36106 - Machine Learning Algorithms and Applications

Master of Data Science and Innovation

University of Technology of Sydney

Table of Contents

[1. Executive Summary 2](#_Toc167486695)

[2. Business Understanding 3](#_Toc167486696)

[a. Business Use Cases 3](#_Toc167486697)

[3. Data Understanding 6](#_Toc167486698)

[4. Data Preparation 11](#_Toc167486699)

[5. Evaluation 24](#_Toc167486700)

[a. Evaluation Metrics 24](#_Toc167486701)

[b. Business Impact and Benefits 35](#_Toc167486702)

[c. Data Privacy and Ethical Concerns 38](#_Toc167486703)

[6. Collaboration 39](#_Toc167486704)

[a. Individual Contributions 39](#_Toc167486705)

[b. Group Dynamic 39](#_Toc167486706)

[c. Ways of Working Together 39](#_Toc167486707)

[d. Issues Faced 39](#_Toc167486708)

[7. Conclusion 41](#_Toc167486709)

[8. References 44](#_Toc167486710)

# 

# Executive Summary

This project aims to enhance banking operations through advanced data analytics, focusing on four key objectives: predicting customer spending, detecting fraudulent transactions, customizing marketing efforts, and identifying abnormal spending behaviors. In an era of increasing transaction volumes and sophisticated fraud tactics, the need for efficient and accurate data-driven solutions has never been more critical.

The primary problem addressed by this project is the challenge of managing and analysing vast amounts of transactional and customer data to improve financial services and security. Banks need to predict customer spending to help with budgeting, detect fraud to prevent financial losses, tailor marketing efforts to customer behavior, and identify anomalies to support proactive customer service.

Through this project, several significant outcomes were achieved. A predictive model was developed to forecast customer spending, aiding in better financial planning. A classification algorithm was implemented to detect fraudulent transactions, enhancing security measures. Customer segmentation was improved for more targeted marketing campaigns, and an anomaly detection system was established using Isolation forest to identify unusual spending patterns, enabling timely customer support.

Overall, this project has demonstrated the vital role of machine learning in addressing complex banking challenges, providing scalable, efficient, and accurate solutions that safeguard financial interests and improve customer experiences.

short dash

# Business Understanding

## Business Use Cases

Banks face several issues such as helping customers budget, identifying fraud, customizing marketing, and detecting abnormal spending behaviors. These issues involve ensuring data accuracy and privacy, managing high transaction volumes, adapting to evolving fraud tactics, balancing detection accuracy with false positives, segmenting customers effectively, and communicating with customers without causing alarm.

The aim of this project is to apply machine learning methodologies to derive insights about the “MLAA Bank” customer and transactional data, gathered over a three year duration (2019 to 2022). The successful implementation of the models will empower MLAA Bank stakeholders and employees to implement strategies that bring value to the bank or the end customers, establishing longevity in business operations. The proposed machine learning models that the team design to help address the challenges bank faces are:

1. **Predicting Fraud Detection:** The model identifies fraudulent transactions by analyzing transaction details and customer information using supervised machine learning models. An accurate prediction of fraudulent transactions can help prevent financial losses by identifying and mitigating fraudulent activities before they cause significant damage and enhance customer trust with the bank.
2. **Predicting the total spending:** For effective financial planning and personalized marketing, it is important to forecast client spending based on transaction information and customer demographics. This initiative uses machine learning algorithms to predict future spending trends, giving the bank useful information for improving client happiness and resource allocation. Accurately predicting consumer behavior enhances liquidity management and helps with financial forecasts. Machine learning algorithms detect patterns in the data to meet challenges like managing varying client habits and transaction volumes. In the end, the bank can make data-driven decisions thanks to this regression analysis, which promotes corporate expansion and provides consumers with specialized services.
3. **Clustering:** This case uses machine learning algorithms to segment customers based on their purchasing patterns, demographics, and engagement levels. The K-means clustering algorithm and hierarchical clustering (agglomerative) are used to gain insights into customer grouping, improving campaign precision. These unsupervised learning techniques uncover inherent groupings in customer data, enabling the creation of tailored marketing approaches that enhance customer satisfaction and engagement. This data-driven approach fosters business growth and delivers personalized customer experiences, ultimately allowing for data-driven marketing decisions that foster business growth.
4. **Anomaly Detection:** The model identifies unknown risks or transaction patterns in data by leveraging an unsupervised algorithm – isolation forest – to flag anomalies. Possible risk patterns include sudden spikes in transaction amount, significant differences from the usual merchant’s location (Euclidean distance), time between each transaction, and possible age risk. The insights provided in terms of deterministic rules or patterns make it easier for the bank to assess the risk associated with each transaction.
5. Key Objectives
6. The key objective of this Regression analysis is to Develop a predictive model for customer spending. Understand the factors influencing customer spending behavior and provide actionable insights for improving business strategies.
7. An adaptive and accurate fraud detection classification model to detect fraudulent banking transactions. This will help minimize financial losses and protect customers by utilizing machine learning algorithms to improve accuracy, efficiency, and scalability of bank’s fraud detection systems.
8. The project aims to improve marketing campaigns by segmenting customers using unsupervised learning techniques like K-means and hierarchical clustering, enhancing personalized campaigns, improved ROI, and business growth.
9. An anomaly detection model that automatically flags suspicious and out-of-the-norm trends in the data for evaluation.

For this project key stakeholders includes:

* **Bank Customers:** They benefit directly from improved financial services, fraud detection, and personalized communication. Ensuring their data privacy and security is paramount**.**
* **Bank Management:** Executives and decision-makers who oversee the bank and ensure operations align with the bank's strategic goals and financial outcome.
* **Compliance and Risk Management Teams:** Responsible for ensuring the bank adheres to regulatory requirements and managing risks associated with fraudulent transactions.
* **Marketing Team:** Utilizes customer segmentation and spending behavior insights to tailor marketing campaigns and improve customer engagement**.**

**Project Aim:**

The project aims to address stakeholder requirements by implementing machine learning methodologies to derive as much value as possible for the bank, while taking into consideration the bank’s operational scale and their strategic goals.

short dash

# Data Understanding

This dataset is provided by the client with limited information regarding its source and how it is collected. It contains many CSV files of customers’ transaction details and another set of CSV files containing customer-related data of 1000 customers **[*Table 1*]**. All these files are run through a Python pipeline to merge and remove all empty data frames, resulting in a data frame of 983 customers and their transaction details (n = 4,260,904 p = 23) to build our models on.

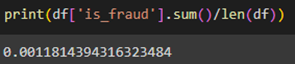
|  |  |
| --- | --- |
| Data Frame | Feature |
| customers.csv | “Customer details” 1,000 entries, 15 features |
| transactions.csv | “Concatenated transactions” 4261035 entries, 10 features |

**Table 1** | Data frames before merging

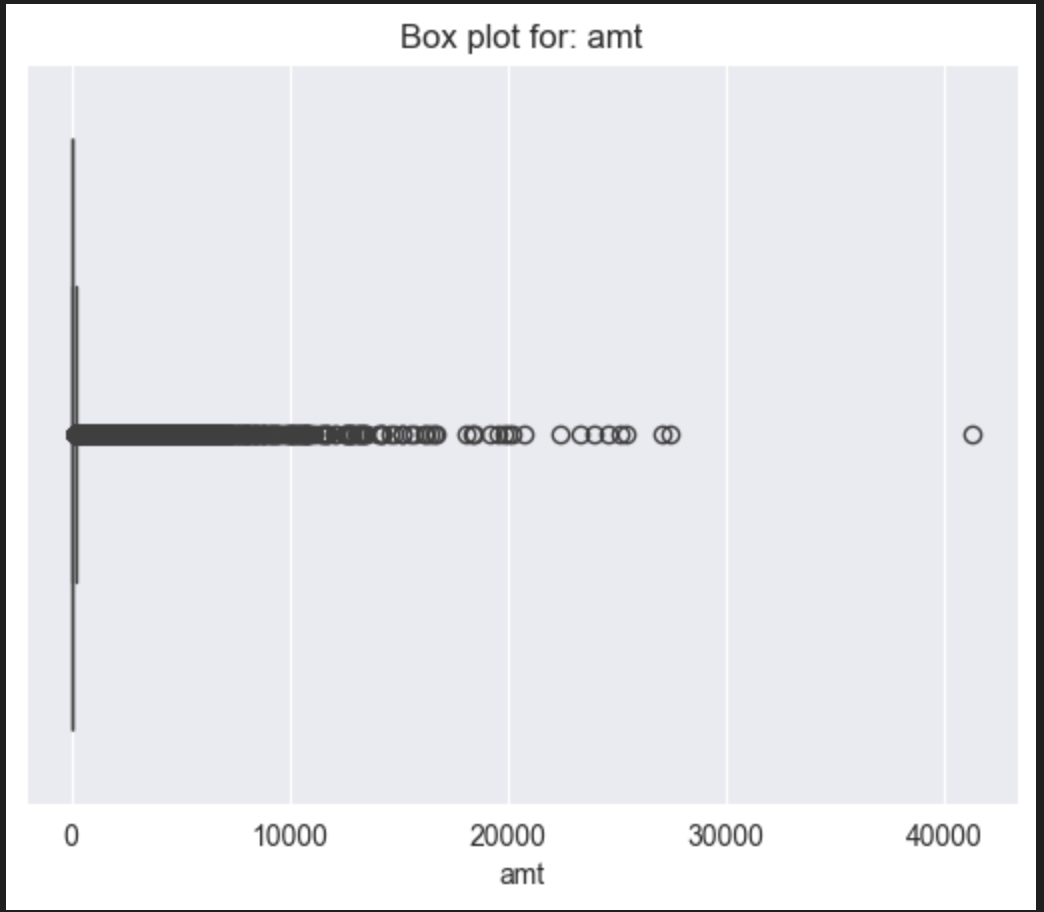
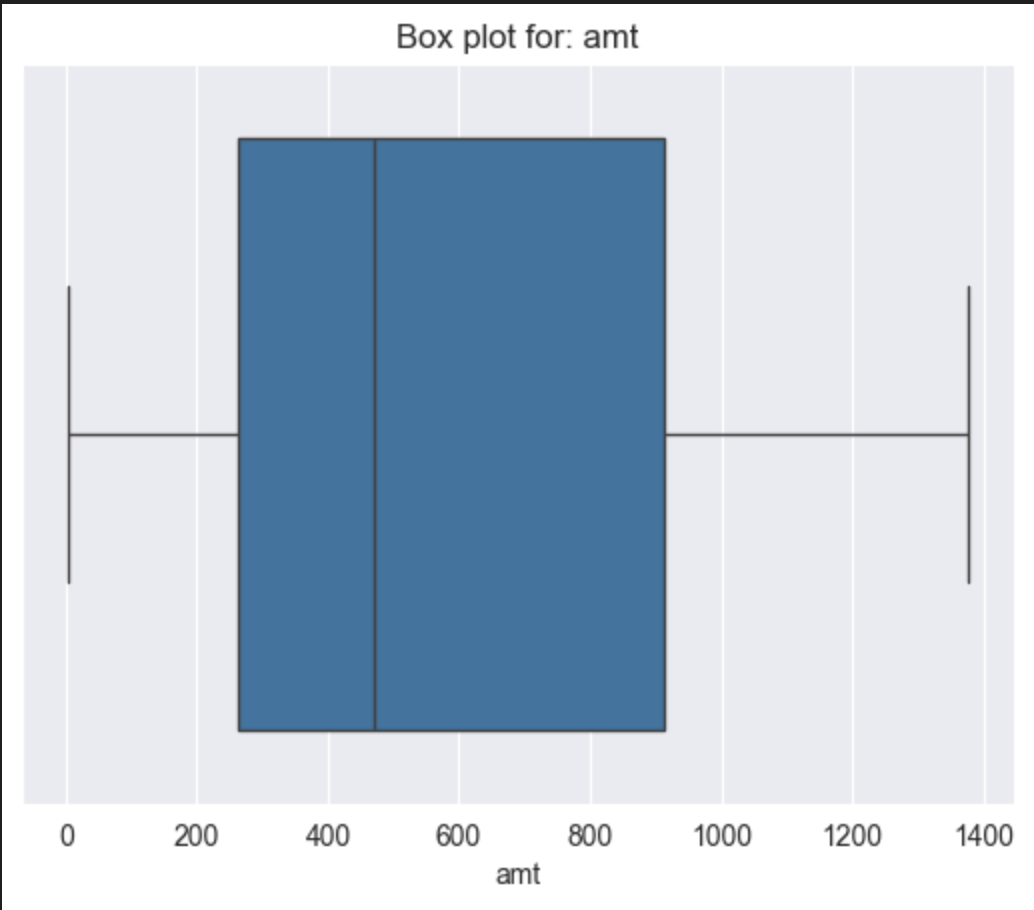
The resulting features after merging are listed in the **[Table 2]** below.

|  |  |  |
| --- | --- | --- |
| **Feature Name** | **Datatype** | **Example Data Point** |
| ssn cc\_num first last gender street city state zip lat long city\_pop job dob acct\_num trans\_num unix\_time category amt is\_fraud merchant merch\_lat merch\_long | object integer object object object object object object integer float float integer object object integer object integer object float integer object float float | “115-04-4507” “4218196001337“ “Jon” “Snow” “M” “67148 Rose Cliff Apt. 314” “Saint Petersburg” “FL” “33710” “27.32” “33.27” “32412” “Accounting” “3/10/1959” “888022315787” “550713c1e27c321e318dd29cad98562d” “1558209673” “gas\_transport”  “66.4” “1 or 0” “Nelson PLC” “35.493488” “-78.905707” |

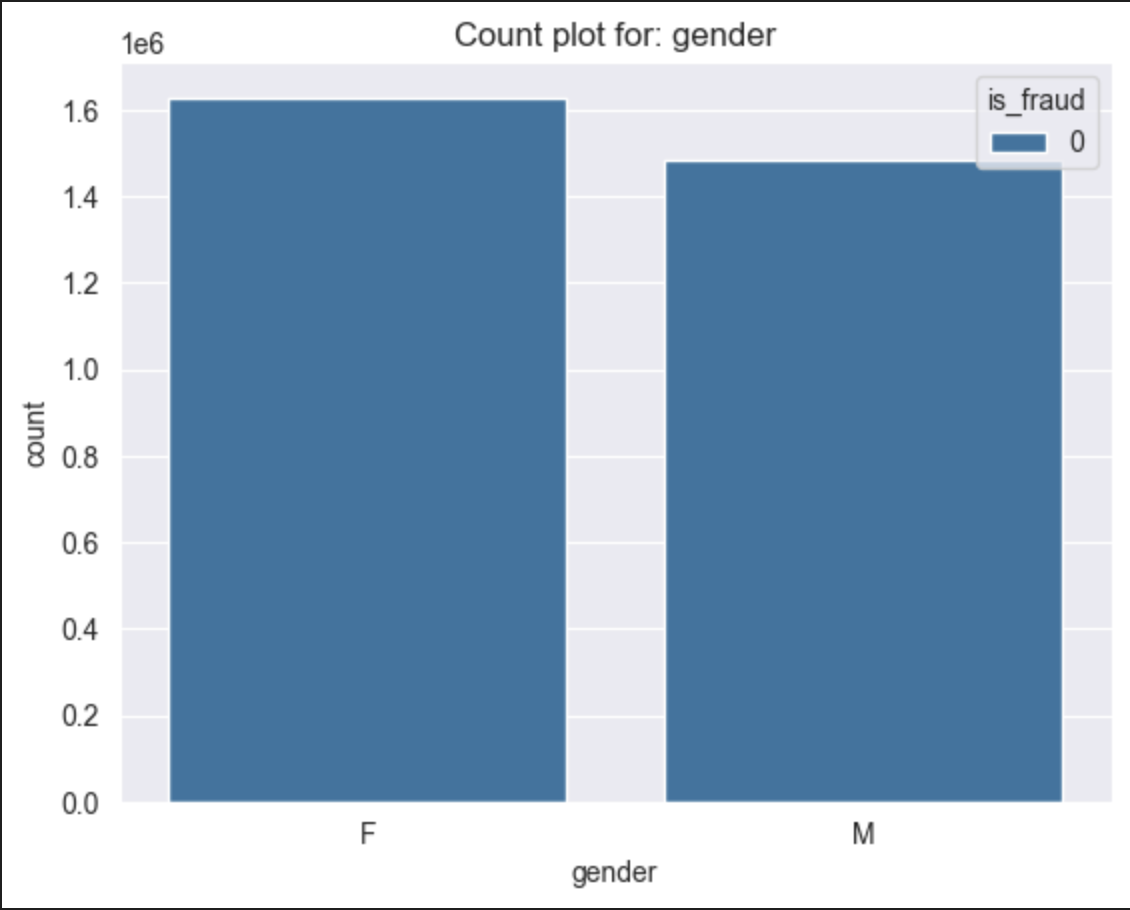
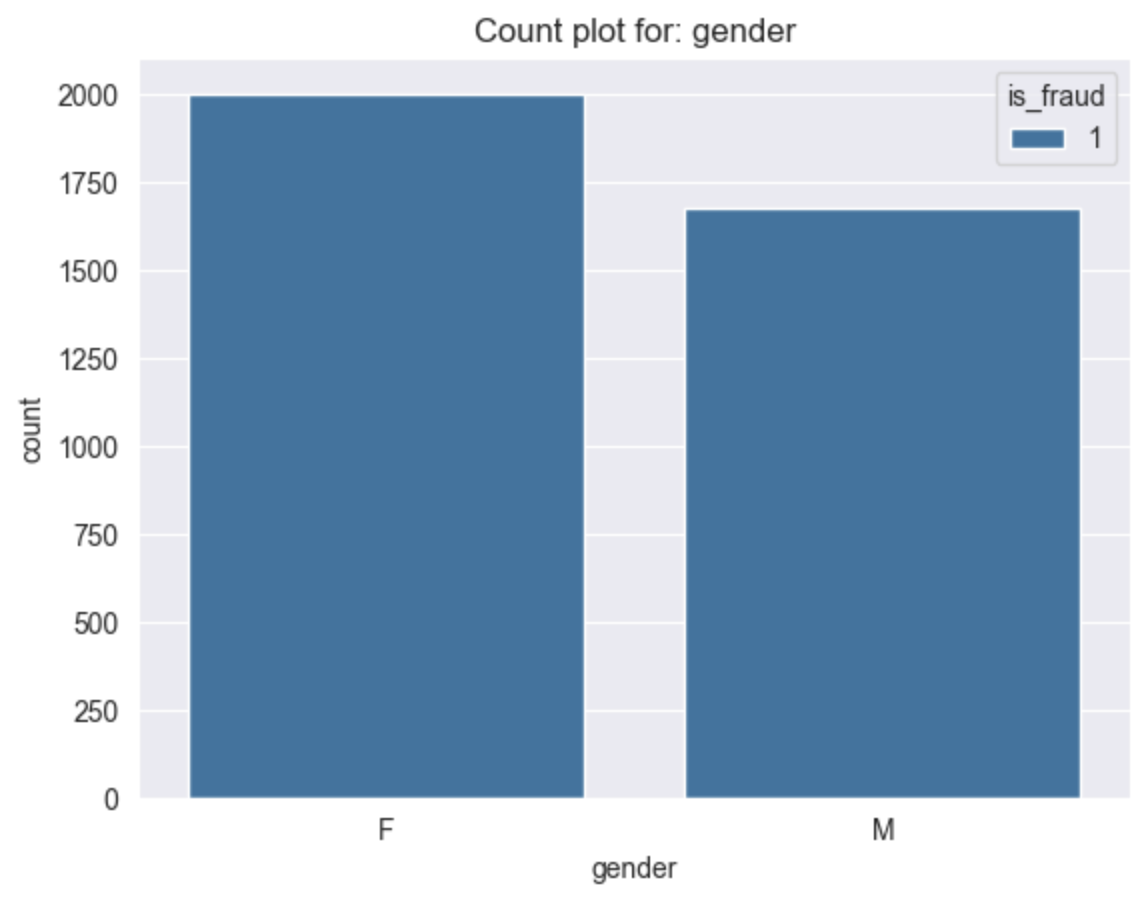
**Overall characteristic of the data:**  
The Data is heavily imbalanced with is\_fraud = 1 (Target class) only 0.12% of the entire dataset. This is particularly relevant for regression and classification tasks aiming to predict or classify fraudulent transactions.



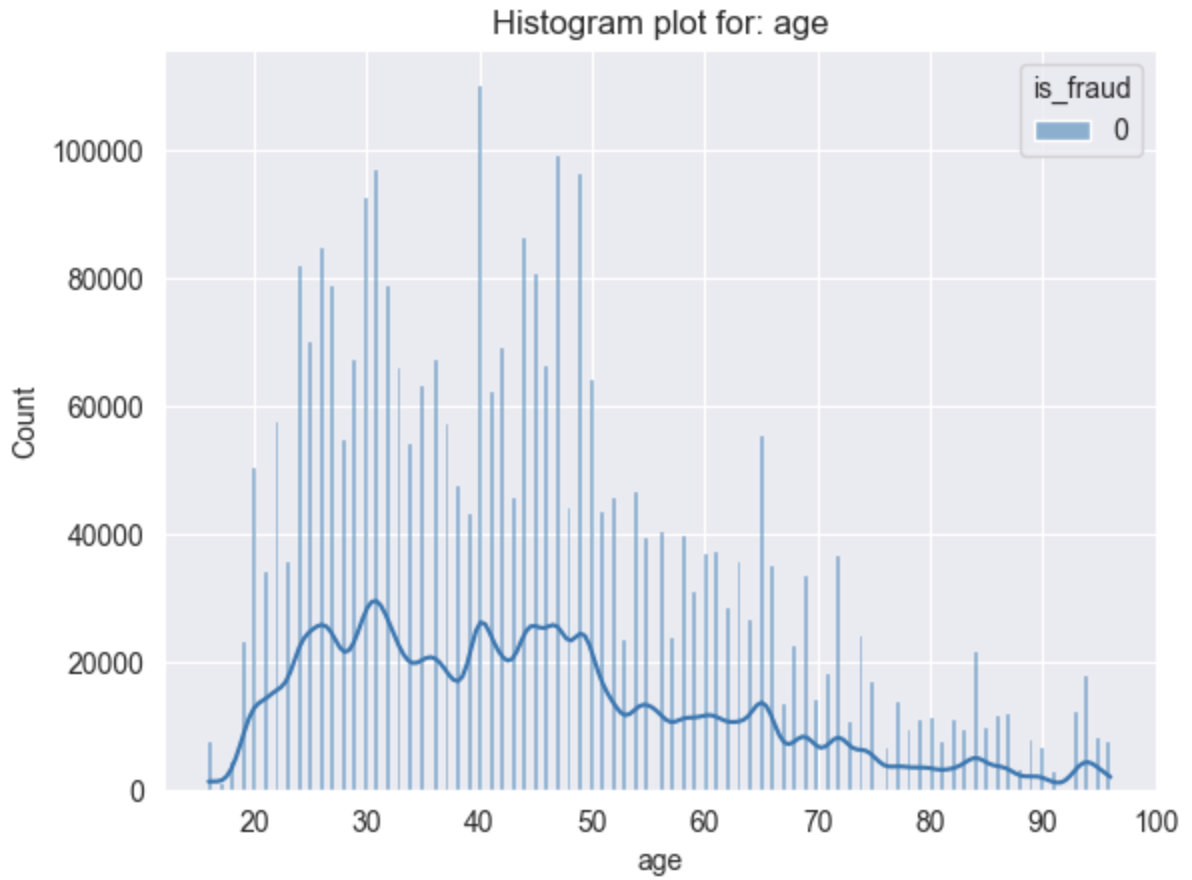
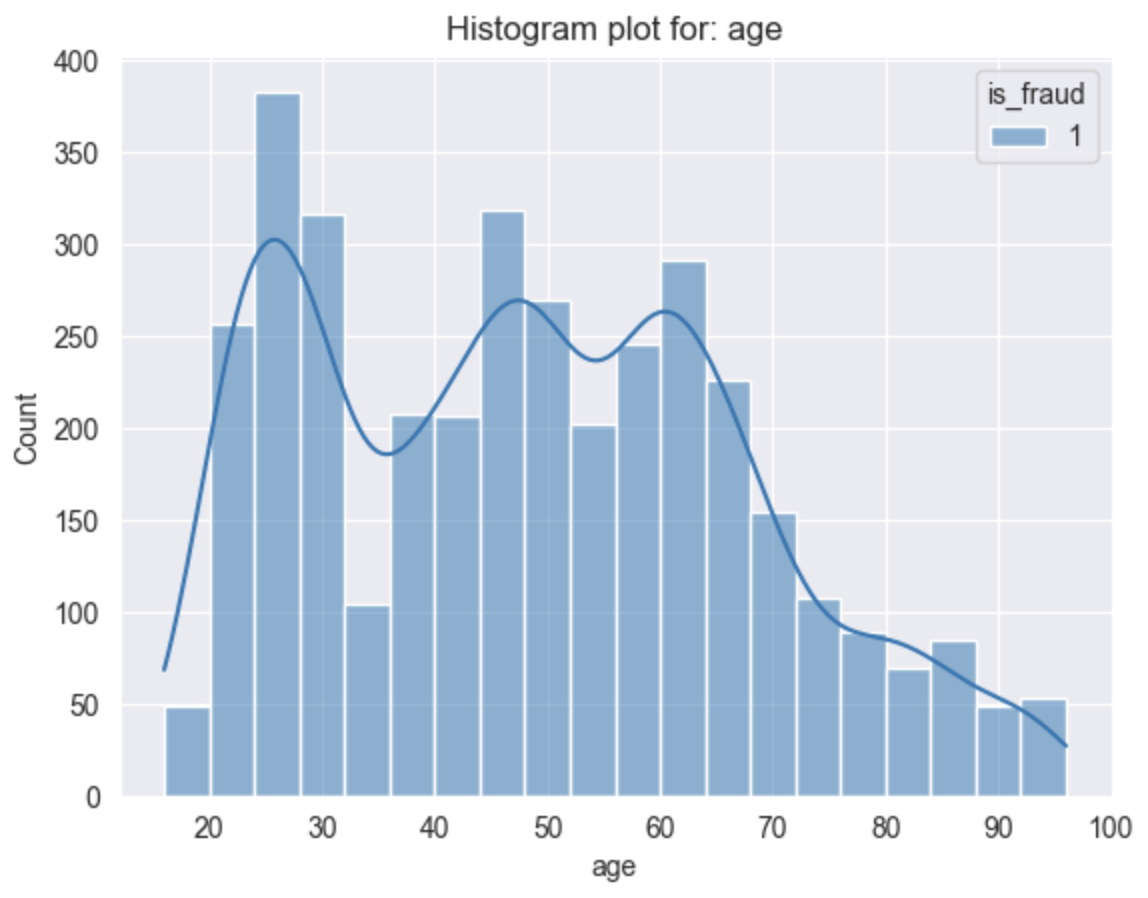
The following feature were used to analyze potential fraudulent transaction:

    
**Figure** **3.1.1** & **3.1.2**

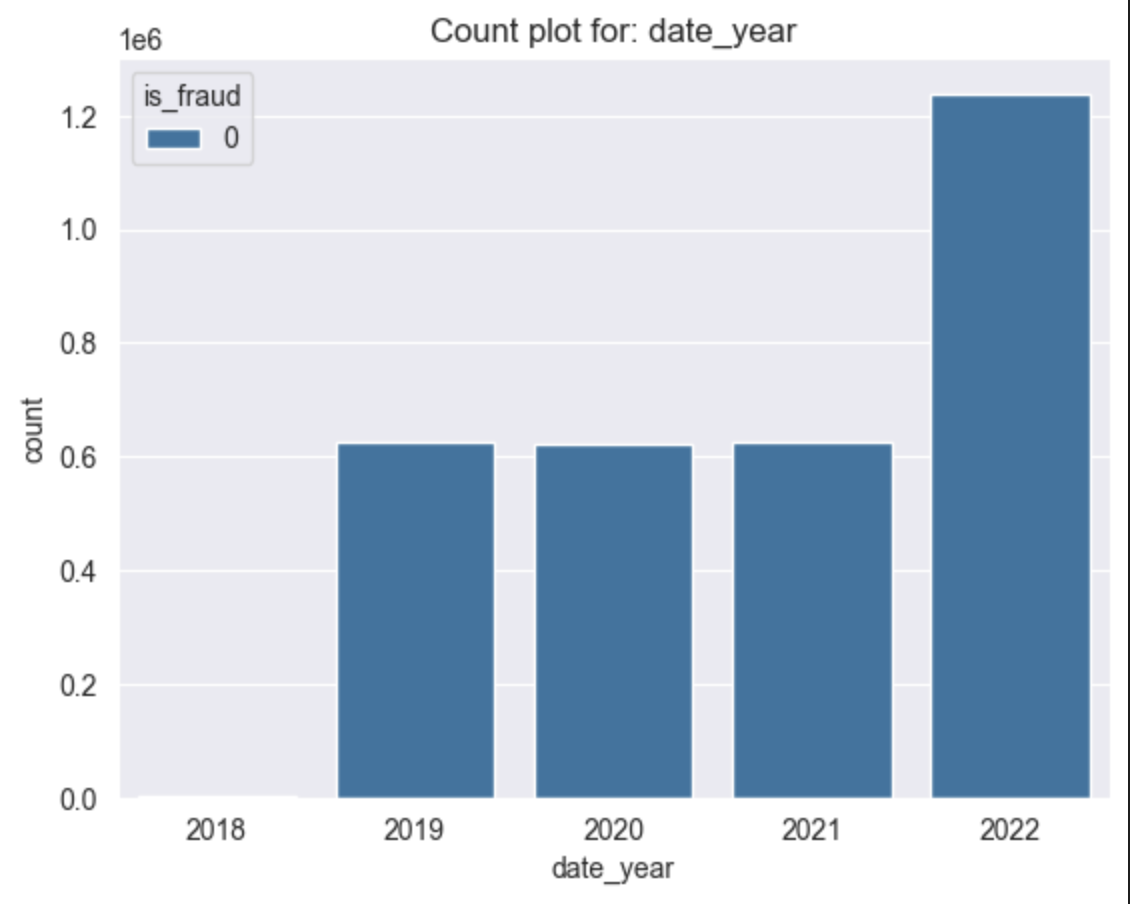
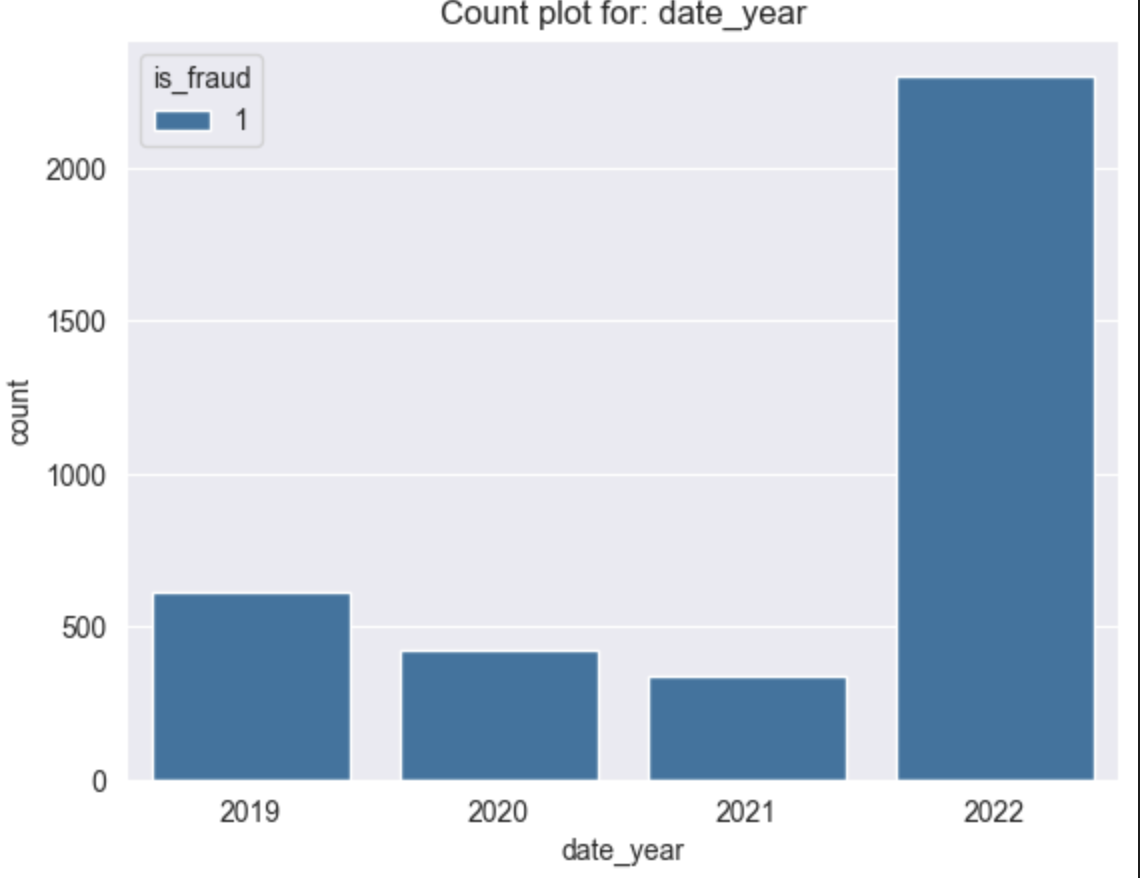
Fraudulent activities primarily occur below $1400, with most cases falling between $200 and $950. To detect more fraud cases, we should focus more closely on transactions within this range.

   
**Figure** **3.1.3** & **3.1.4**

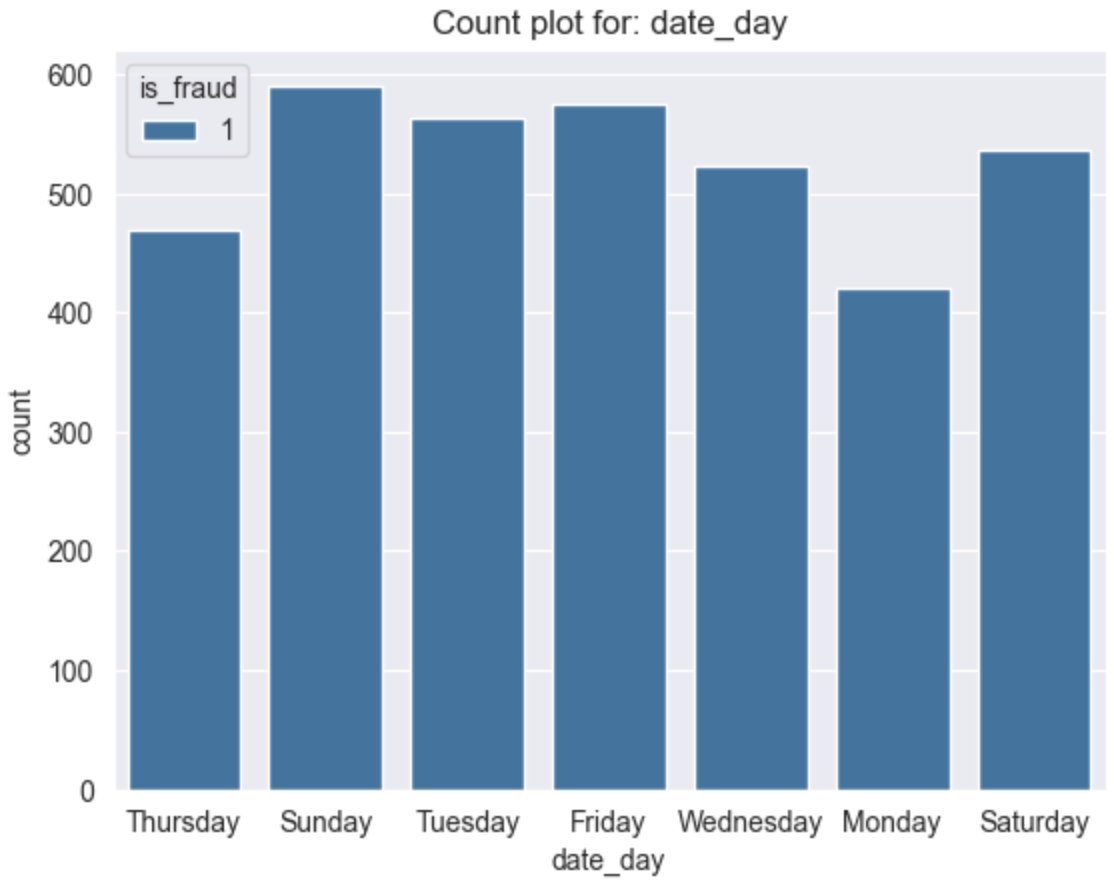
The distribution of fraudulent transactions seemed to be equal across Gender.

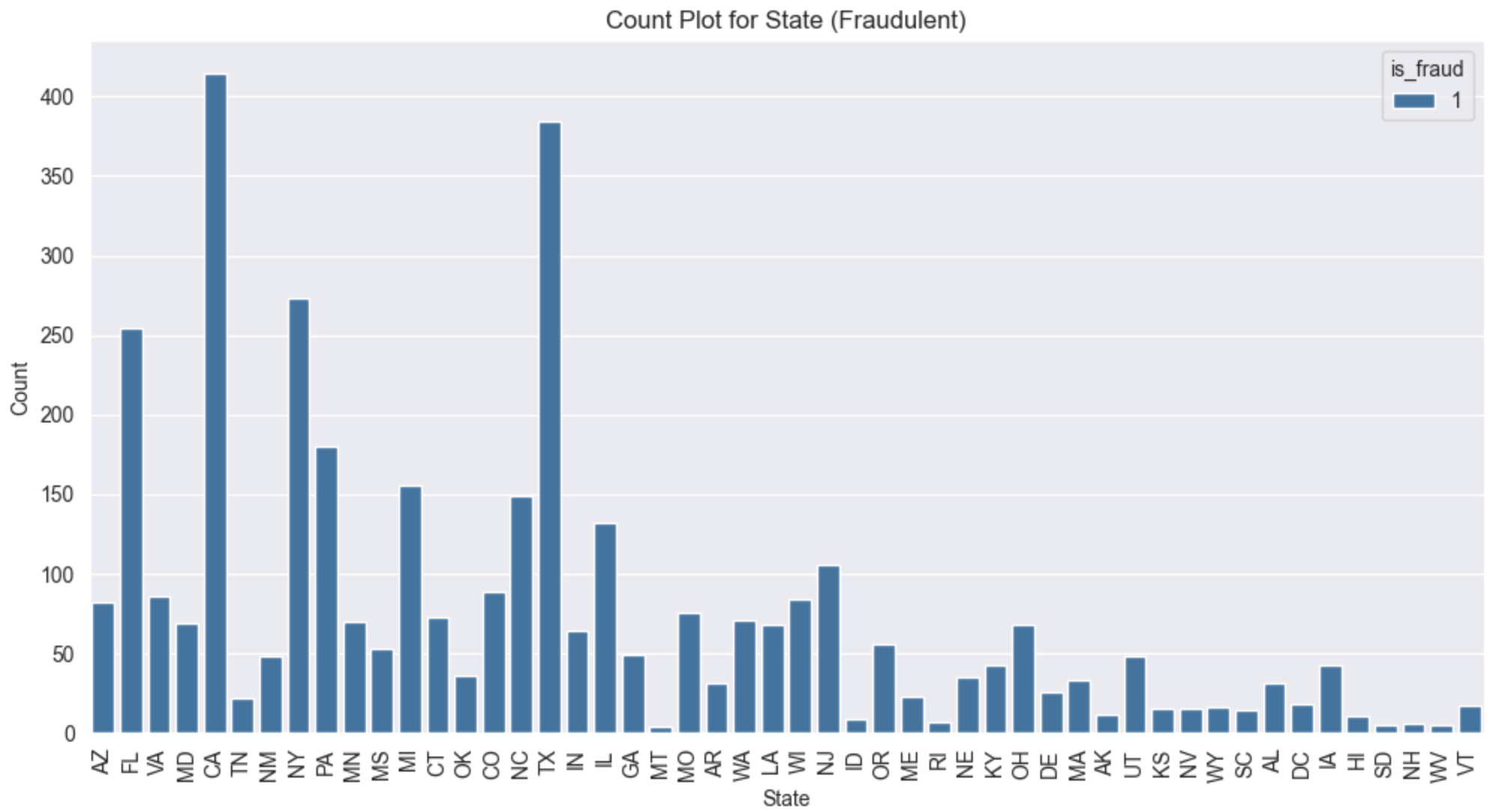
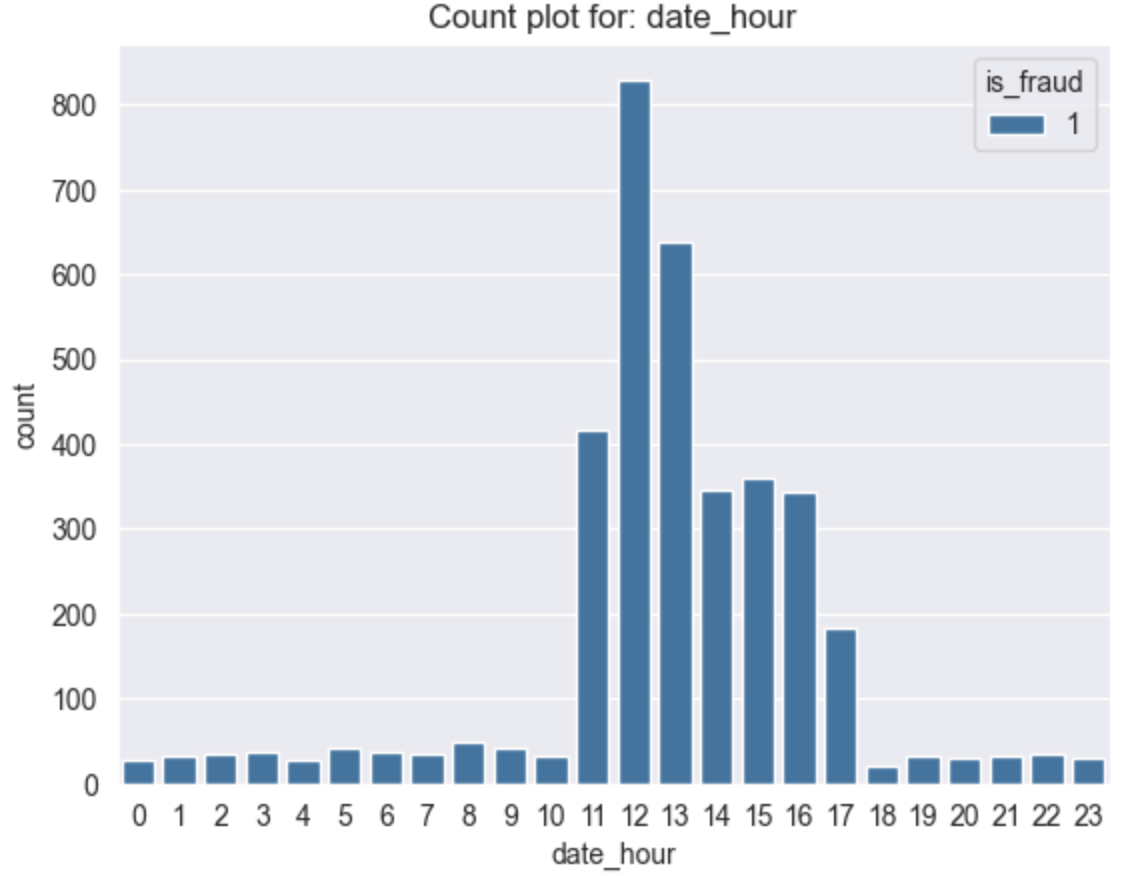
   
**Figure** **3.1.5** & **3.1.6**

A significant number of fraudulent activities are observed by individuals between the ages of 25 and 60.

   
**Figure** **3.1.7** & **3.1.8**

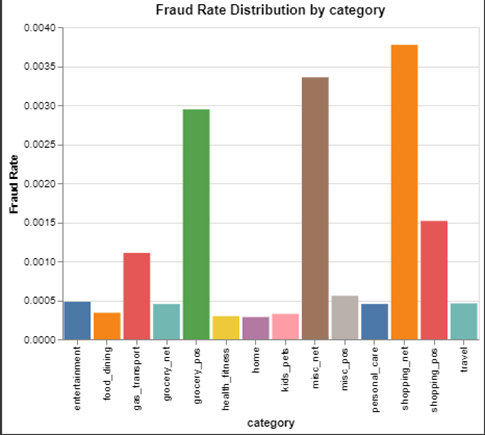
Although 2019-2021 had similar transactions, the fraudulent activities experienced a decreasing trend from 2019 to 2021 but suddenly spiked in 2022 which is alarming for the bank.

   
**Figure** **3.1.9** & **3.1.10**  
Days of the week does not seem to be indicative of fraudulent transactions.

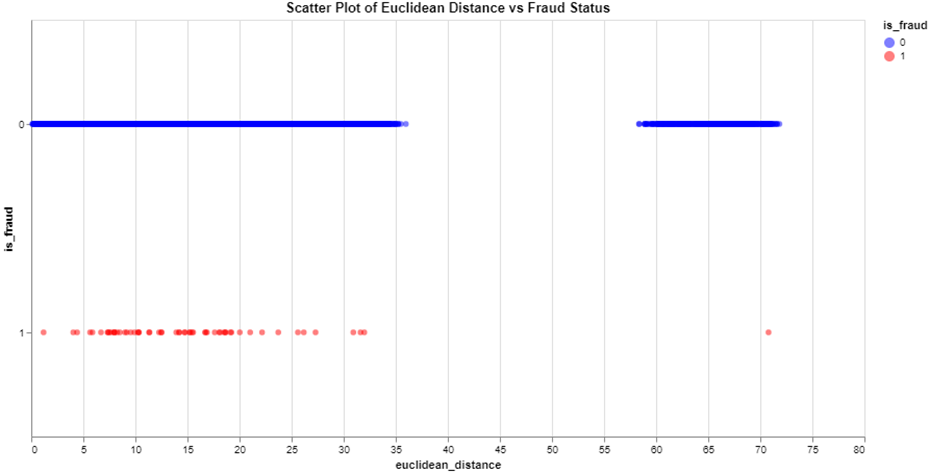
  
**Figure** **3.1.11** & **3.1.12**

The majority of transactions exhibiting significant fraudulent behavior occur between 11 A.M. and 5 P.M., with the highest concentration of fraudulent activities happening around noon, at 12 P.M **[Figure 3.1.11]** .

The state features **[Figure 3.1.12]** may be indicative of some pattern that helps detect fraud. However, it will not be included in the fraud risk related model due to ethical concerns discussed in 6.c

  
**Figure** **3.1.13**

There seems to be some pattern of risk with some categories although it is spread out. Most of the fraudulent coincide with the grocery\_pop and Shopping\_pops category

  
**Figure** **3.1.14**

Euclidean distance from the average merchant location seems to not be indicative of fraudulent risk due to the high overlap between the two.

short dash

# Data Preparation

Prior to modeling, data are prepared by applying a range of industry standard cleaning and manipulation techniques. Moreover, the data goes through different preparation steps catering towards specific use cases. These are illustrated on the following page.

**4.1 Data preparation for Regression:**

Cleaning the datasets to guarantee quality and consistency was the first stage in getting the data ready for regression analysis. This involved managing missing values and looking for and eliminating duplicate entries. To prevent redundancy, for example, duplicate rows in the transaction and customer data were found and eliminated.

Subsequently, the transaction and customer data were combined using the shared identifier 'acct\_num' to incorporate pertinent details for every transaction. In order to correlate transactional data with client demographics, this step was essential.

There were both category and numerical characteristics in the dataset, and each required a distinct preprocessing method. StandardScalar was used to scale the numerical features, while OneHotEncoder was used to encoder the categorical data.

Finally, the data was split into training and testing sets to evaluate the model's performance. This split ensures that the model can generalize well to unseen data, a critical aspect of building robust predictive models.

**4.2 Data preparation for Classification:**

In preprocessing the training dataset, one-hot encoding is used to convert specific categorical features into binary representations. Employing the OneHotEncoder module from sklearn, columns such as 'category', 'age\_cat', 'date\_year', and 'date\_day' undergo this transformation. These binary features are integrated with the original dataset, replacing the categorical columns. Subsequently, to mitigate multicollinearity, the variables 'age' and 'date' are removed. Additionally, other categorical variables like "Gender" and "State" are similarly encoded using one-hot encoding or similar methods, enabling their conversion into numerical formats compatible with machine learning models.

***Scaling numerical features:*** Numerical features like "Age" and "Amount" are scaled to ensure that they have a similar range of values. This is often done to prevent features with larger magnitudes from dominating the model training process.

***Selecting relevant features:*** The features selected for the final model involve considering domain knowledge and using selection techniques such as correlation analysis or feature importance scores from tree-based models.

**4.3 Data preparation for Clustering:**

Certain features (amt, city\_pop, lat, long, category, city, state, gender, and job) were chosen and categorised into numerical and categorical categories in order to prepare the data for a K-means clustering algorithm.

StandardScaler was used to standardise numerical characteristics in order to guarantee that they were on the same scale, which is necessary for K-means.

OneHotEncoder was used to transform categorical characteristics into a numerical representation. These transformations were performed to the corresponding features using a ColumnTransformer, and the fit\_transform function produced a preprocessed dataset that was prepared for clustering. In order to verify that the preprocessing was effective, the altered data's shape was lastly examined. By taking these actions, you can be confident the data is K-means algorithm ideal.

**4.4 Data preparation for Anomaly detection:**

**Known features:** amt, category

**Feature Engineering needed:** Time between transactions, number of transactions made to a particular merchant, day of the week’s transaction was made and age, euclidean distance.

* The time between transactions was calculated by sorting transactions by acct\_num and time\_stamp, then calculating the time difference between consecutive transactions within each group: *time\_diff*
* The number of transactions made to a particular merchant is derived from the cumulative sum of each transaction made to that merchant: *num\_transactions*
* Age is derived from the difference between DOB and transaction date.

Other features are not selected for the following reasons:

* Transaction ID, Account Number, and SSN are just unique identifiers for transactions and accounts.
* Customer latitude and longitude did not reveal any useful risk pattern from other models.
* City, State, Zip Code, Street, Jobs could raise ethical concerns (discuss in 6.c)

**5.1: Regression Analysis for Prediction:**

Machine Learning Algorithms Used:

For this project, I utilized two primary machine learning algorithms: Dummy Regressor and RandomForest Regressor.

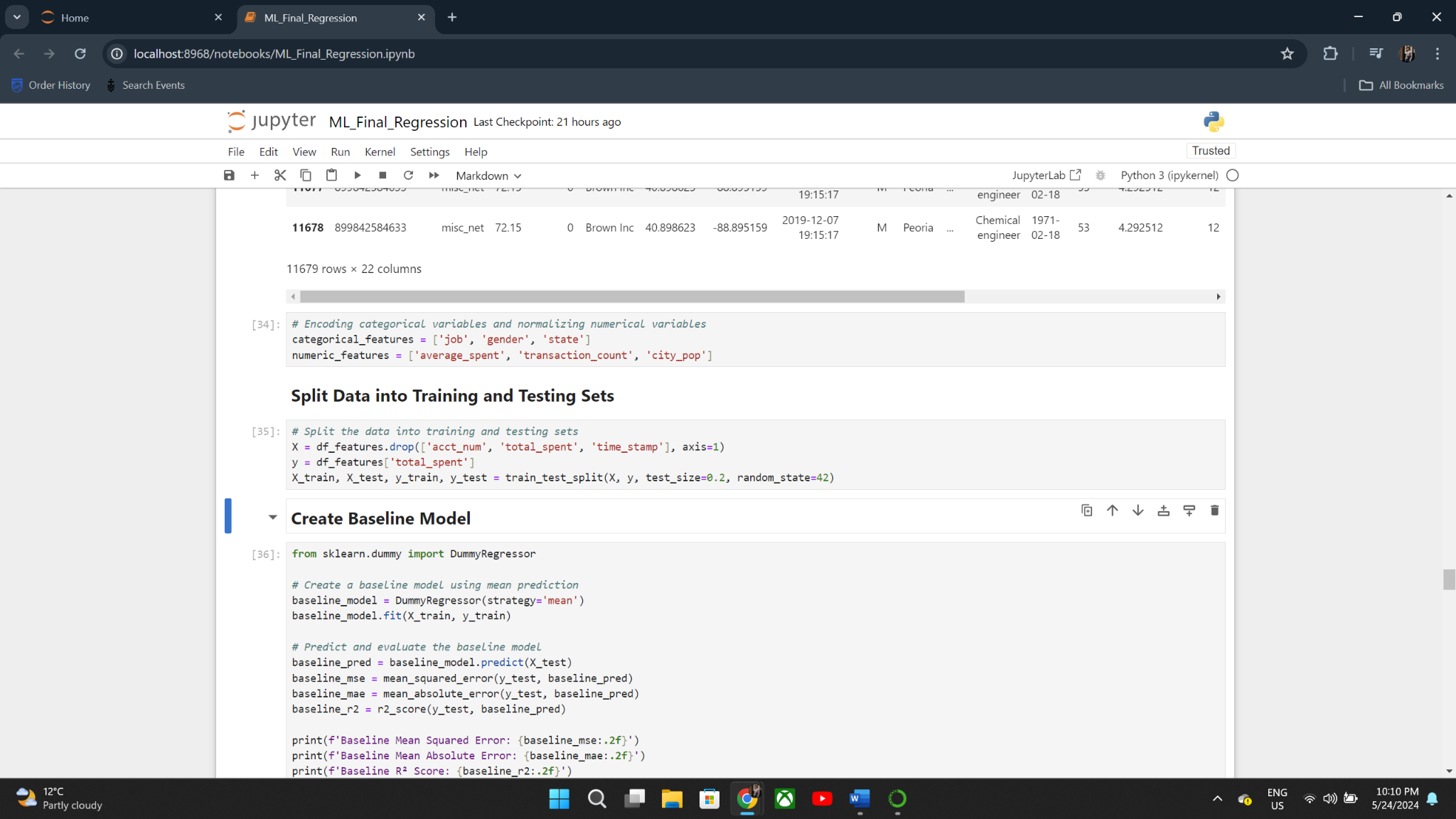
* Dummy Regressor was used as a baseline model. It predicts the mean of the target variable and serves as a performance benchmark to evaluate the effectiveness of more sophisticated models.
* RandomForest Regressor was selected for its robustness and ability to handle large datasets with high dimensionality. It aggregates the predictions of multiple decision trees, reducing overfitting and providing high accuracy.

Rationale for Algorithm Selection:

* RandomForest Regressor: This algorithm was chosen due to its ability to handle both numerical and categorical features, its robustness against overfitting, and its high interpretability through feature importance scores. It is well-suited for predicting customer spending, where the relationship between features and the target variable can be complex and nonlinear.
* Dummy Regressor: The Dummy Regressor was used to set a baseline performance. By comparing the RandomForest model's performance against this simple baseline, we can quantify the improvement and justify the use of more complex algorithms.

Parameter Tuning and Model Selection:

For this project, we did not employ GridSearchCV for hyperparameter tuning but instead used the default settings and manually adjusted parameters based on standard practices and project needs.

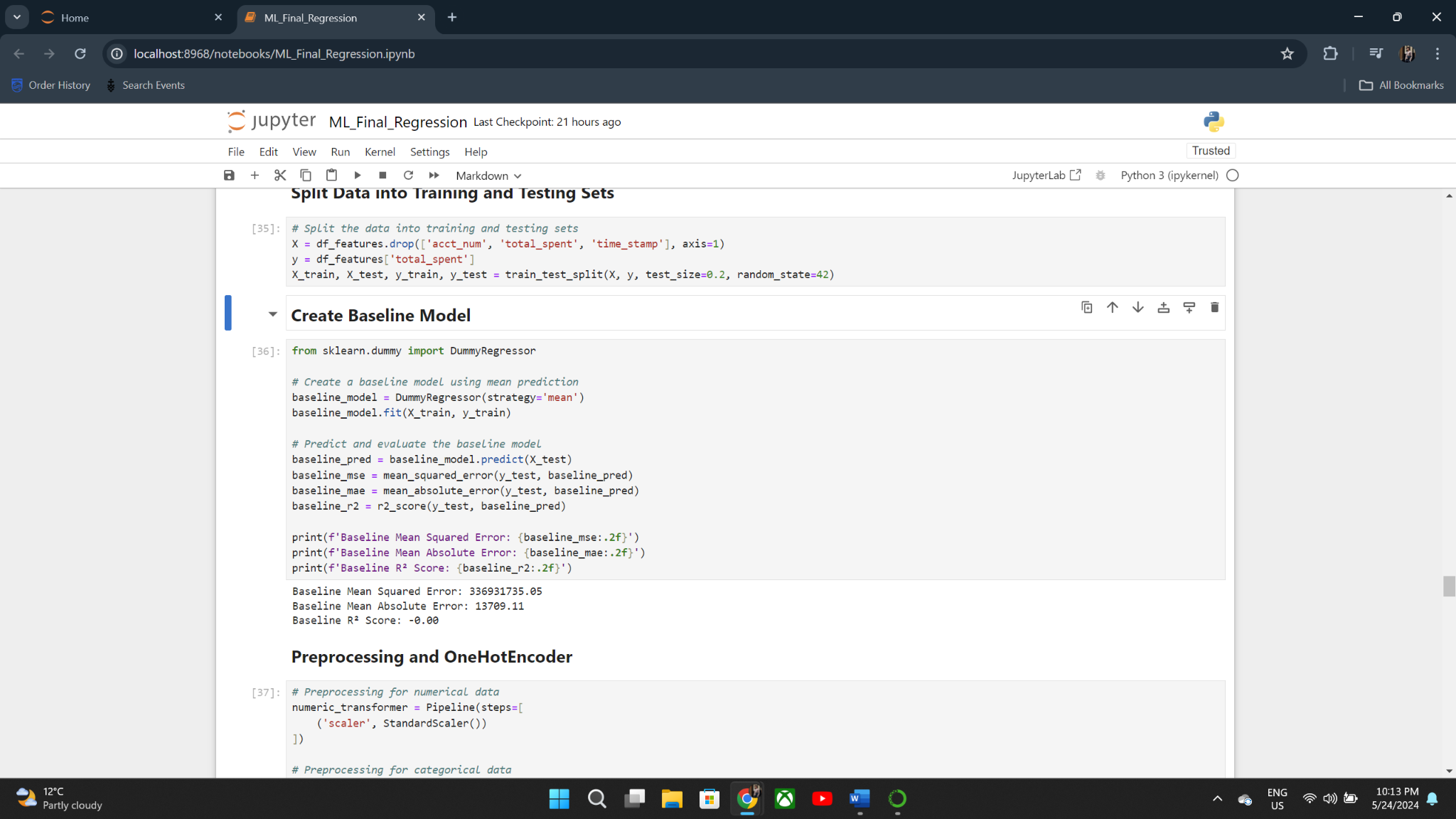


**Approach 1: Dummy Regressor**

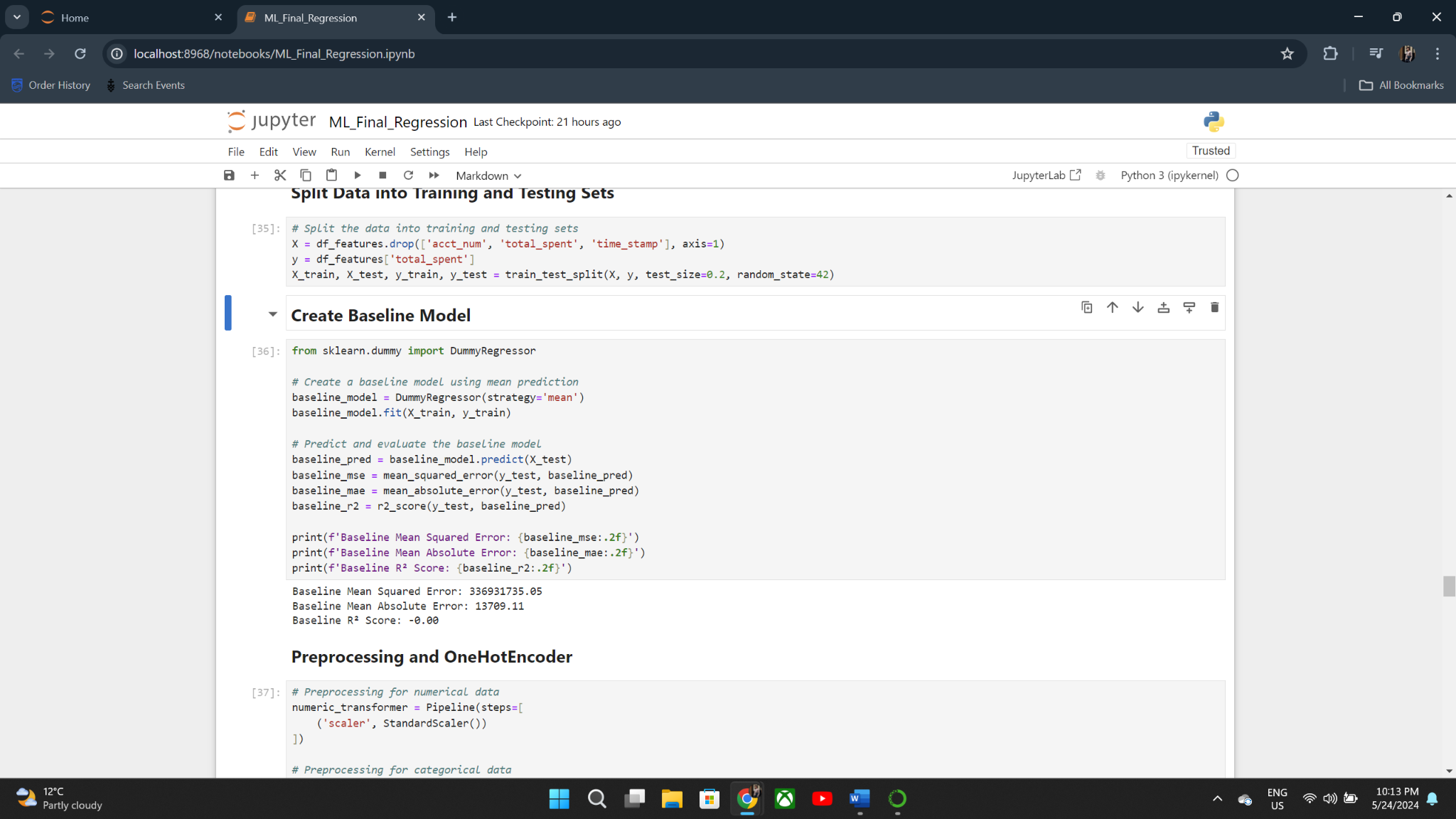
Algorithm: **Dummy Regressor**

Purpose: To serve as a baseline model by predicting the mean of the target variable.

The Dummy Regressor was trained on the same training data to provide a baseline performance for comparison.



Performance:

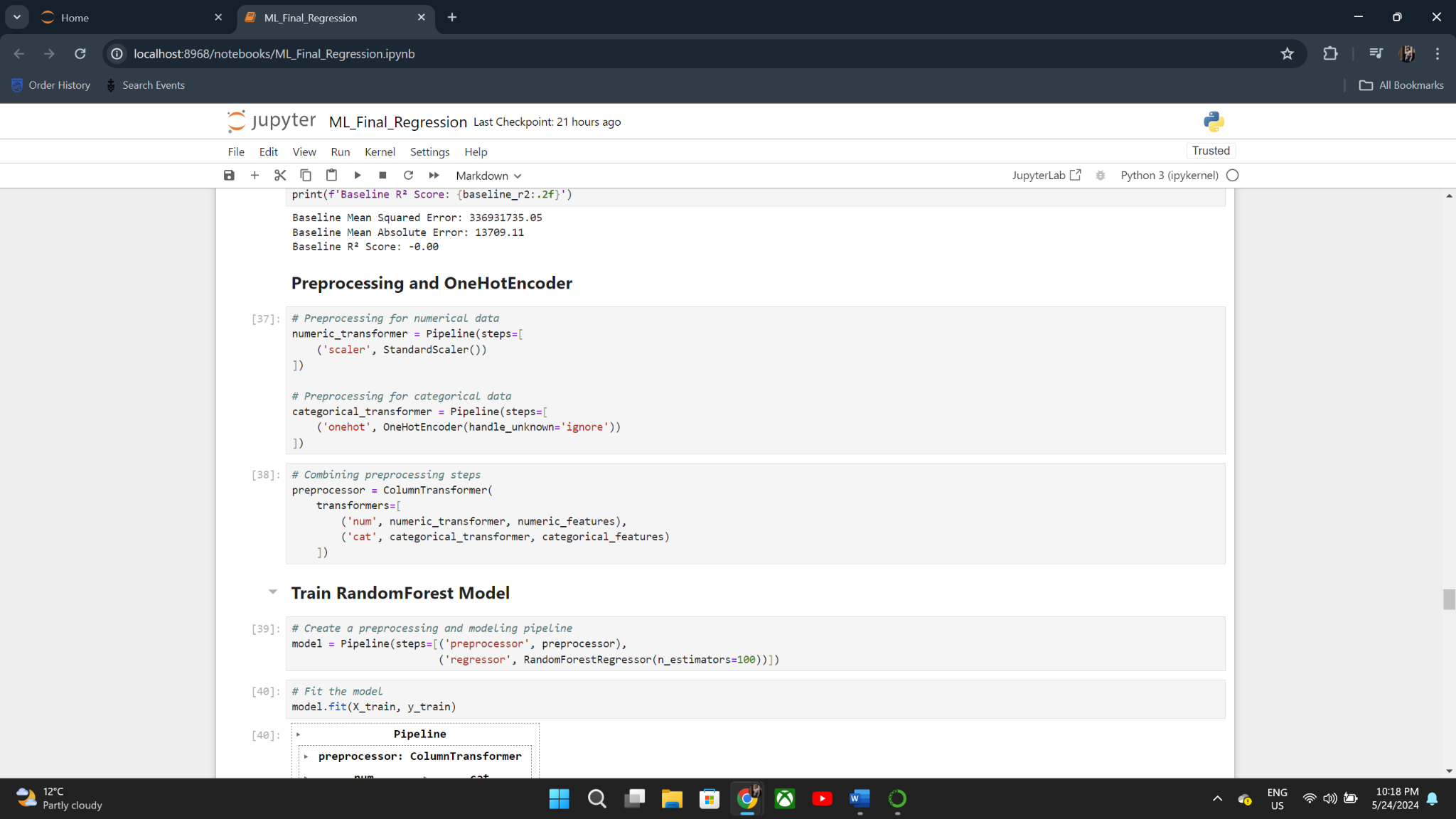


**Approach 2: RandomForest Regressor**

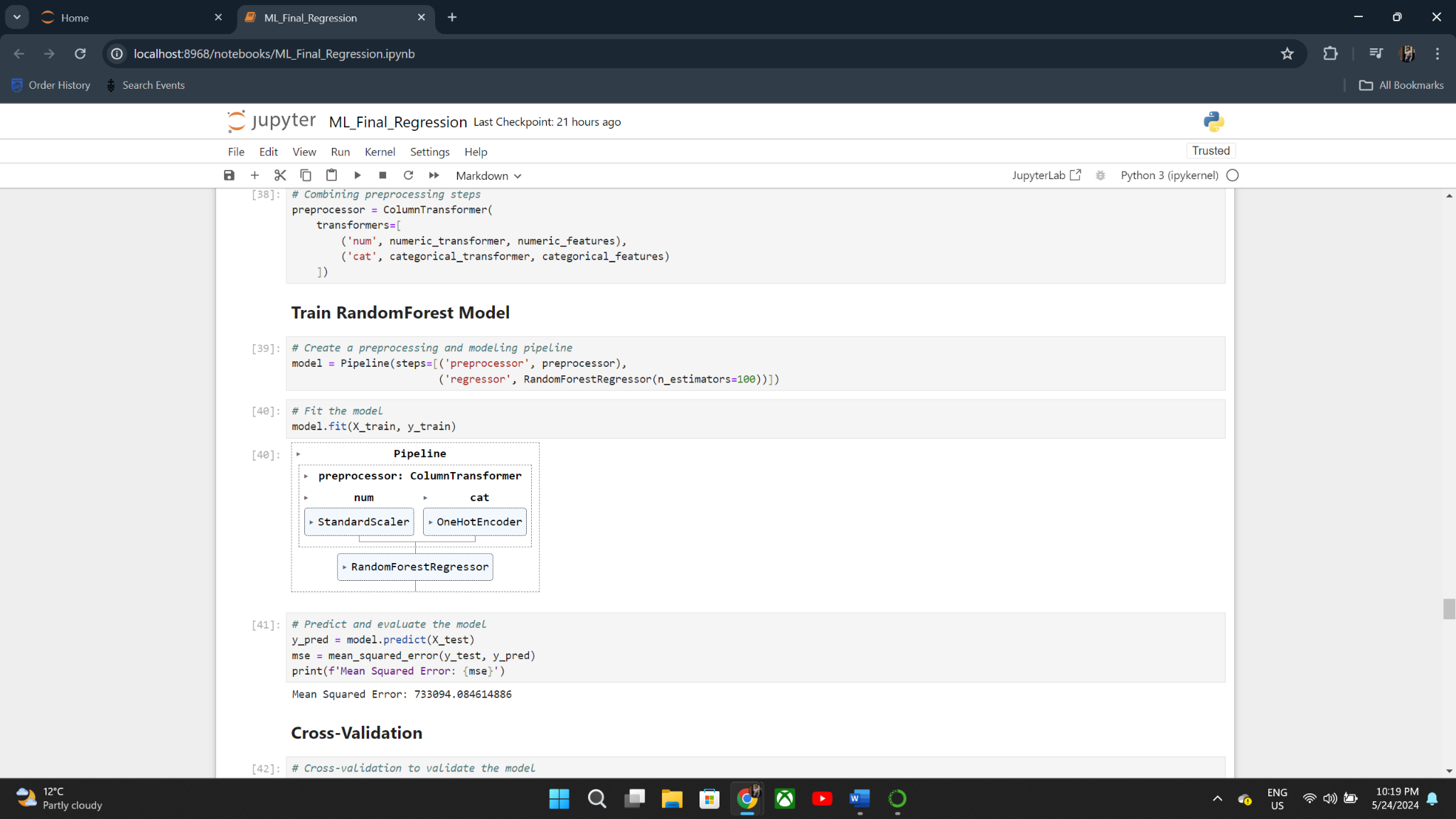
**Algorithm:** RandomForest Regressor

**Key Hyperparameters:** n\_estimators: 100

**Preprocessing and Feature Engineering:** Numerical features were scaled using StandardScaler. Categorical features were encoded using OneHotEncoder.



The training process involved fitting the RandomForest model to the preprocessed training data. Cross-validation was used to validate the model’s performance and ensure its generalizability.



**Performance Metrics:**

**Mean Absolute Error**: 220.88

**Mean Squared Error:** 733094.08

**Root Mean Squared Error**: 848.22

**R² Score**: 1.00

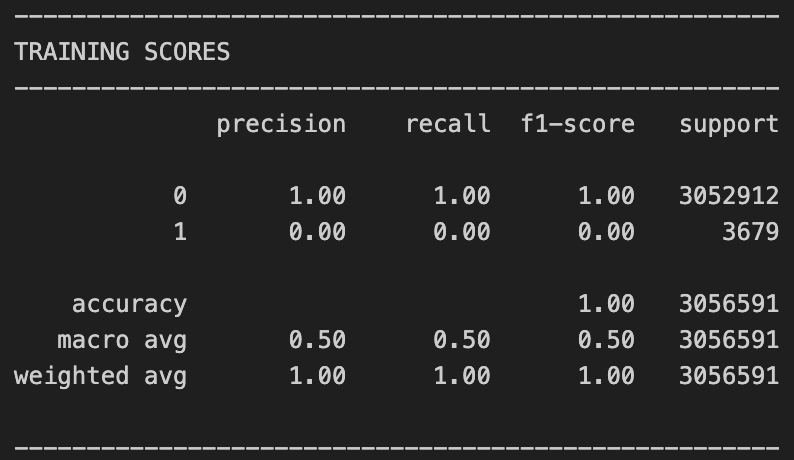
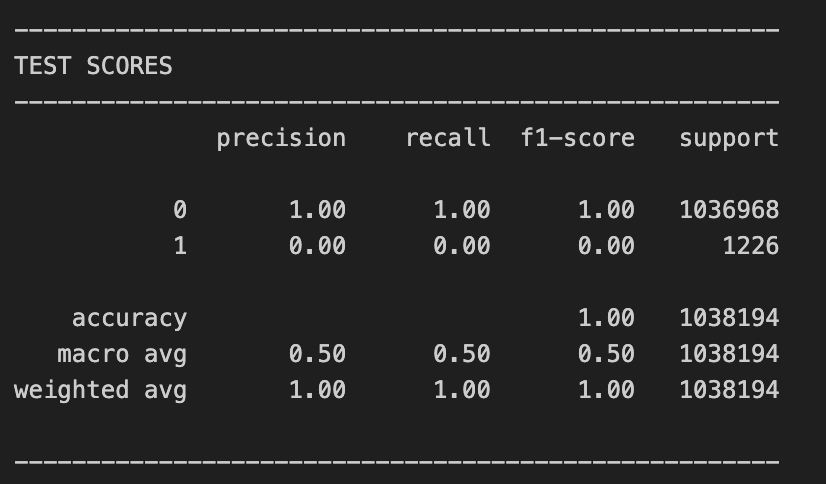
**Adjusted R² Score**: 1.00

The evaluation of the regression models revealed significant improvements in prediction accuracy when contrasting the RandomForest Regressor with the conventional Dummy Regressor. The RandomForest model has excellent performance metrics, such as an R2 Score of 1.00, indicating that it described all of the variance in the target variable. The display of the data further confirmed the algorithm's accuracy in predicting customer expenditure. These results highlight the advantages of using sophisticated machine learning techniques to financial regression analysis.

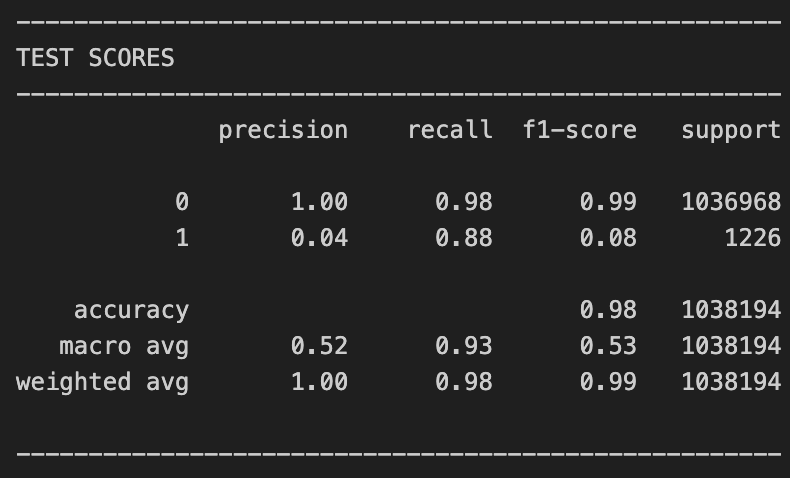
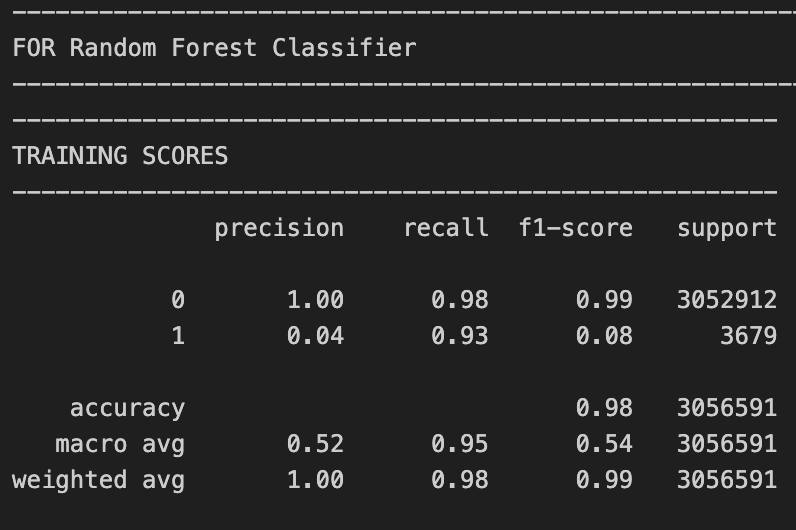
**5.2 Classification for Fraud Detection:**

Approach 1: ***Performed logistic regression***

Results revealed issues of overfitting, particularly to the majority class (non-fraudulent transactions). Precision, recall, and F1-scores for the minority class (fraudulent transactions) are all 0, signifying an inability to detect such instances. This suggests the model's poor performance in addressing the dataset's class imbalance. Consequently, despite seemingly promising overall metrics, the model is unsuitable for detecting fraudulent transactions. To address this, a Random Forest model will be explored, aiming to mitigate class imbalance and improve detection accuracy.



Approach 2: ***Perform Random forest***

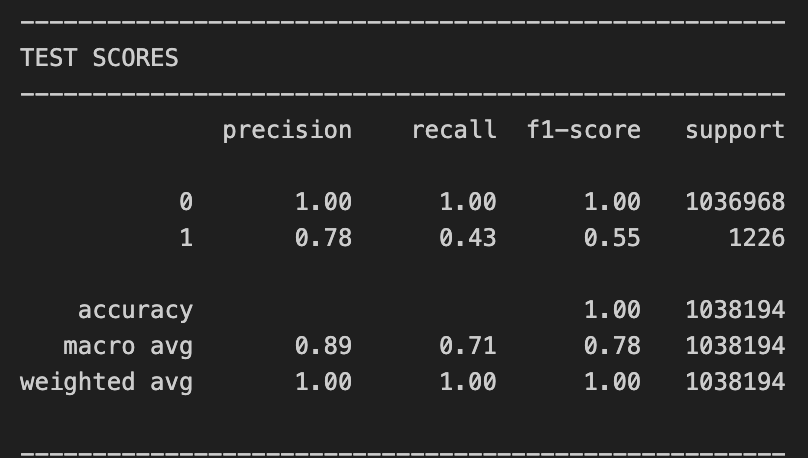
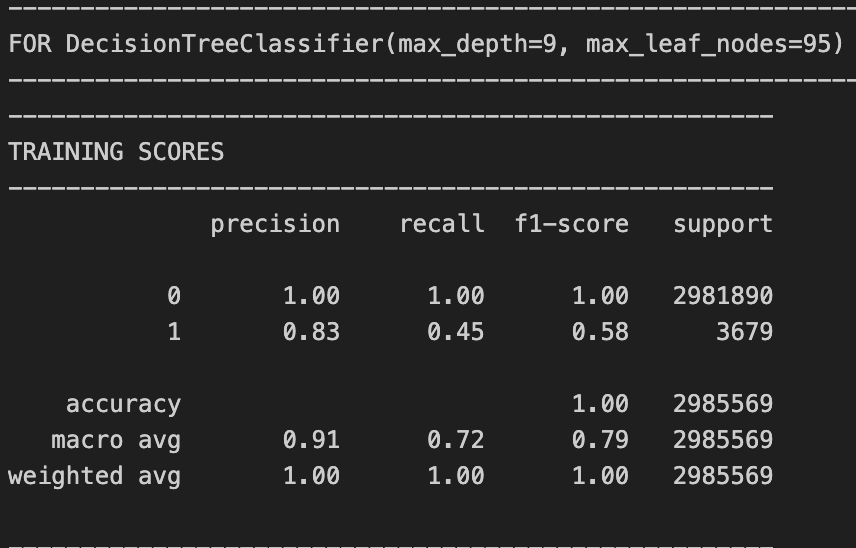
******

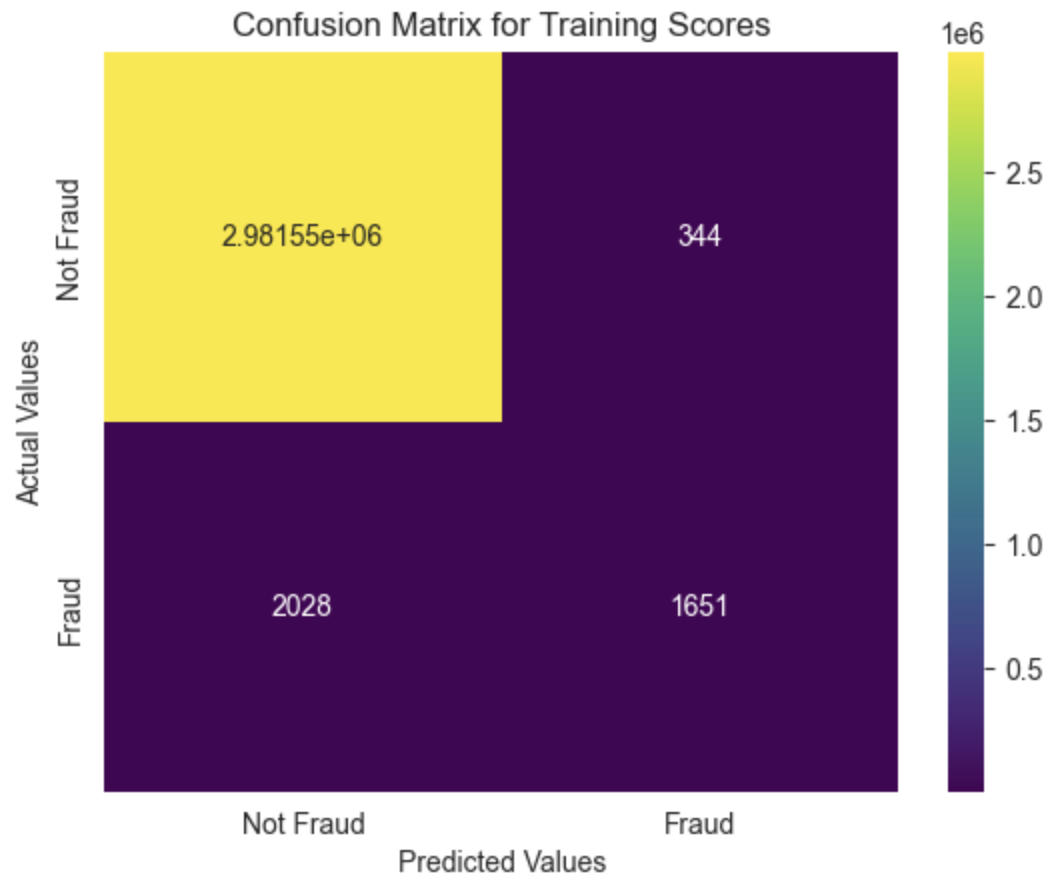
The Random Forest Classifier with a maximum depth of 10 and balanced class weights shows a significant imbalance in predicting fraudulent transactions. While the model achieves near-perfect precision and recall for the majority class (Not Fraud), it struggles with the minority class (Fraud), showing high recall but extremely low precision. This indicates that the model is heavily biased towards the majority class, despite the class balancing efforts.

Approach 3: **Decision tree with** ***XGboost***

The XGBOOST model performs better than the stand alone decision tree model. Even though the XGBoost model seems slightly overfit in terms of precision, it performs very stable and similar performance across different folds and proves its robustness. Hence, the XGBOOST model is the best obtained model for this project.

**Hyperparameter tuned:** max\_depth=9, max\_leaf = 95



  
**Figure 5.2.1**

The model performed the best when it had max\_depth=9. It didn't increase the score but it was less overfit than max\_depth=10.

The XGBOOST model performs better than the decision tree model. Even though the XGBoost slightly overfit in terms of precision, its performance is more stable and achieved a similar result across different folds, proving its robustness. Hence, the XGBOOST model is the best obtained model for this project.

**5.3 Clustering:**

Approach 1: ***K-means Clustering with k=5***

With the help of this procedure, the dataset is divided into K pre-specified, unique, non-overlapping subgroups, or clusters, each of which contains a single data point.

**Rationale:** K-means clustering aids strategic decision-making for targeted marketing by providing easily interpretable clusters of discrete client groups. Additionally, it is renowned for its scalability and efficiency in handling large datasets, making it particularly effective for managing vast amounts of transaction data.

**Feature Selection:**

**1. Transaction Amount (amt):** This will aid in understanding consumer spending habits and enabling the classification of customers based on their typical spending habits.

**2.** **Merchant Category (category):** Marketing techniques can be tailored to the clients by classifying them according to where they spend their money.

**3. Geographical Information - City (city) and State (state) :** These can be used to segment customers by location, which is useful for region-specific marketing.

**4. Latitude (lat) and Longitude (long):** More precise location data that can be used for accurate geo-targeting in marketing campaigns.

**5. Customer Demographics:**

* Gender: If purchase behaviours are influenced by this demographic element, then gender-specific marketing may be implemented.
* Job (job): Knowing the professional background might be useful in tailoring marketing campaigns according to the nature of the profession.

**Model Tuning :**

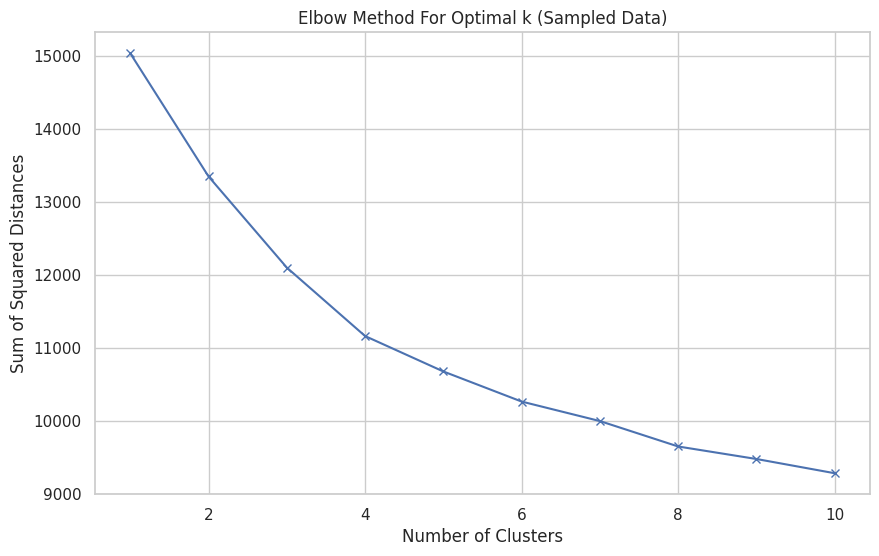
Choosing K (Number of Clusters): ​​To see how it works at first, I randomly selected the value of k=5. Before thinking of more methodical ways to fine-tune the number of clusters, this first pick enables me to test the clustering behaviour and assess the model's efficacy in segmenting the data.

Approach 2: ***K-means Clustering with Elbow Method***

The Elbow Method calculates the optimal number of clusters by plotting squared distances from each location to its designated center, identifying marginal benefits and suggesting the appropriate number of clusters.

**Model Tuning:**

Finding the Number of Clusters (K): To evaluate the effectiveness of the model, K was first selected at random. Plotting the total squared distances between each point and its designated centre allowed for further tuning using the Elbow Method, which involved searching for a "knee" in the curve. Using this strategy aids in selecting a K that strikes a balance between an excessive number of clusters and not enough.



**Figure** **5.3.1**

Our analysis's representation of the Elbow graph indicates a distinct bend at K = 4. The 'elbow point' indicates that there is a decreasing benefit on adding clusters after four, as seen by noticeably reduced intra-cluster variances. As a result, we decided that four clusters would be the ideal quantity for our dataset. This choice enables us to adequately capture the variety in the dataset while achieving a balanced model complexity that is interpretable and computationally efficient.

Approach 3: ***Agglometric Heirarchial Clustering***

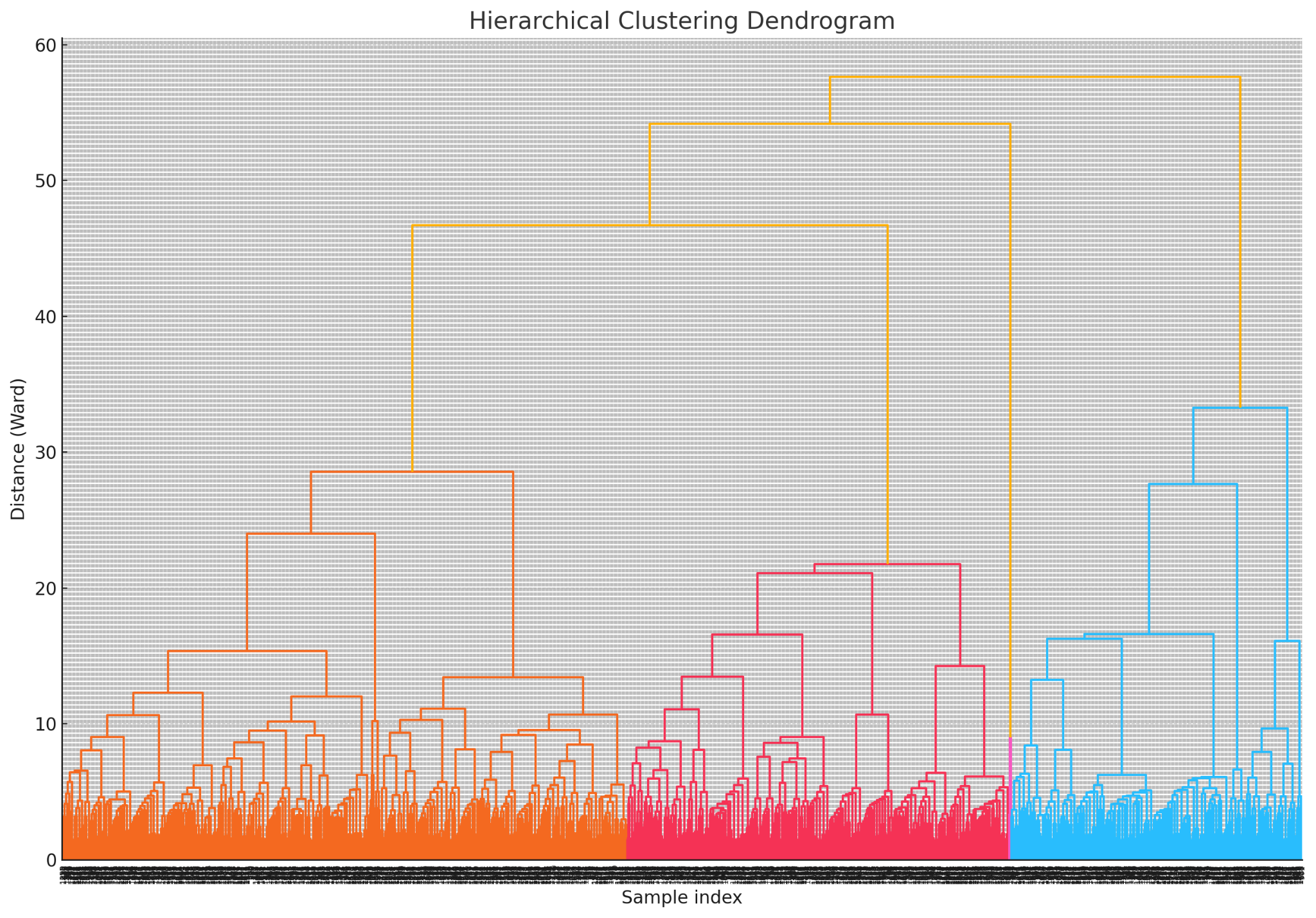
The method creates a hierarchical structure of clusters, with each observation beginning in its own cluster and merging pairs of clusters as it advances up the hierarchy.

The distance metric used is Euclidean distance, which calculates the straight line distance between points in multidimensional space and works well with continuous data.

Ward's approach, which reduces variance within each cluster, is the linkage criterion that was applied. It locates the pair of clusters that, when merging, results in the least amount of overall within-cluster variance increase in each stage.

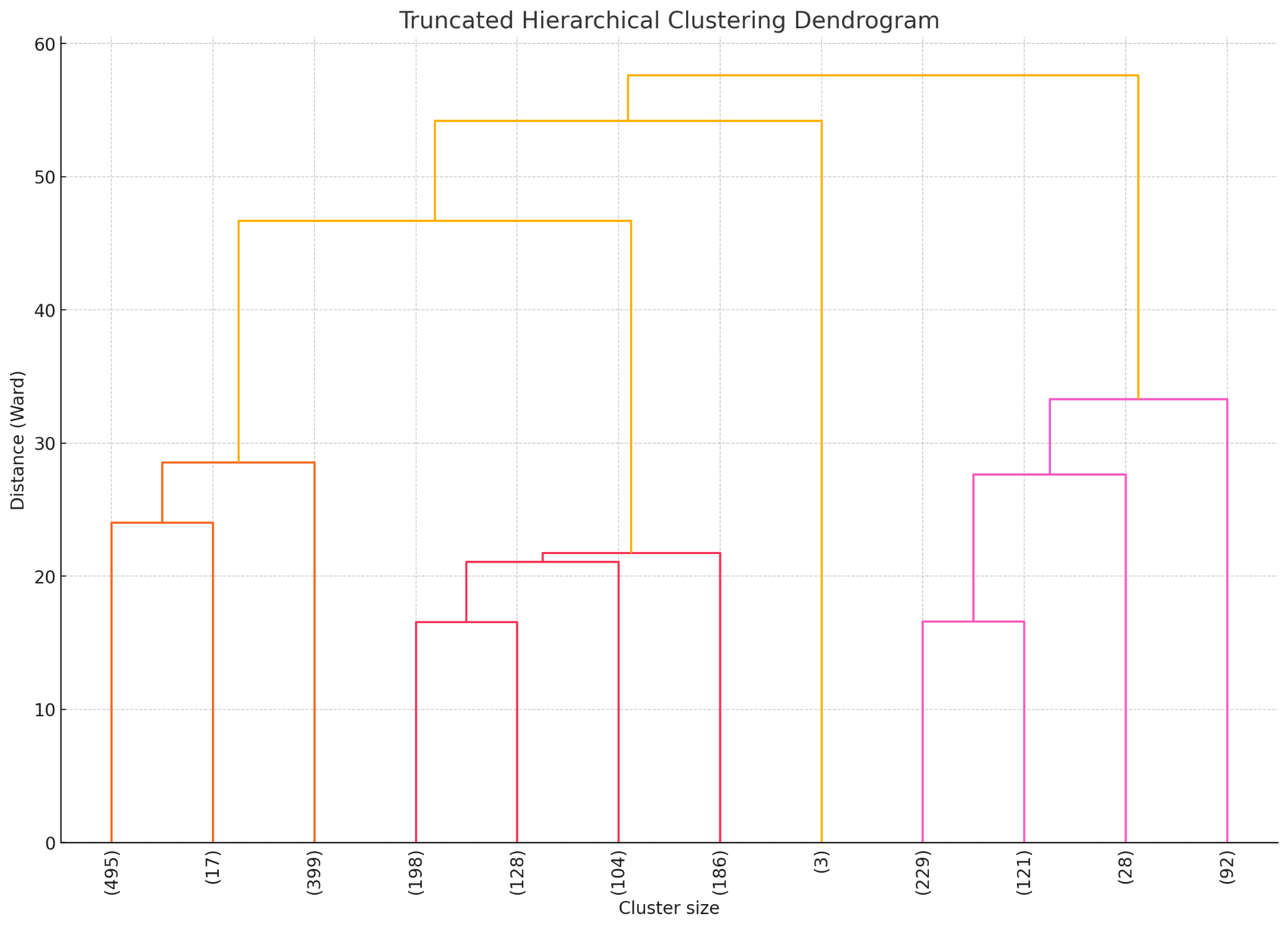
Data Structure Intuition: It offers a dendrogram, a visual depiction of the data structure that might aid in comprehending the relationships between items.

**Figure** **5.3.2**



A single data point (a client) is represented by each line at the bottom.

The distance at which clusters combine is represented by the vertical lines without horizontal connections. Greater dissimilarity between merged clusters is indicated by larger vertical distances.



**Figure** **5.3.3**

**5.4 Anomaly Detection:**

Isolation Forest offers several benefits for anomaly detection in banking, including handling high-dimensional data, efficiency, robustness to noise, and minimal assumptions about data distribution. These advantages make it well-suited for flagging potential fraudulent transactions and other anomalies in complex, large-scale banking datasets.

To capture unusual patterns in transactions’ behavior, the initial features used for the modeling are:

* **Transaction Amount (amt)**: Isolation Forests can effectively identify unusually high or low transaction amounts compared to typical transactions.
* **Transaction Category (category)**: By analyzing transaction categories, the model can detect atypical spending patterns or categories.
* **Time between Transactions (time\_diff)**: Isolation Forests can identify unusual temporal gaps between transactions, such as very rapid or unusually delayed transactions.
* **Number of Transactions to a Merchant (num\_transactions)**: Frequent or infrequent transactions with a particular merchant can be flagged as anomalies.
* **Day of the Week (day\_of\_week)**: Unusual transaction patterns on specific days (e.g., high spending on weekdays vs. weekends) can be detected.
* **Customer Age (age):** Age-based anomalies can be identified, such as spending patterns that are atypical for a particular age group.

However after visualizing each PCA component and each feature in a scatter plot it is revealed that Age, Jobs, category of spending and day of transaction does not make good anomaly detection features. These features may not exhibit significant variation across the dataset or may not have clear boundaries that distinguish normal behavior from anomalies. Following this, some features were removed and new features are generated for the final model:

* **euclidean\_dist:** the euclidean distance from the average merchant location. Significantly different from the average location of transaction for a given merchant could indicate higher risk.
* **normalized\_amt**: the amount normalized against typical spending patterns (unique to each customer)
* **amt**
* **age**
* **num\_transactions**
* **time\_diff**

# Evaluation

## Evaluation Metrics

**5.1.a) Evaluation and results of regression (Prediction for Next Month By Total Spending)**

To assess the performance of the regression models, we used the following evaluation metrics:

* Mean Absolute Error (MAE): Measures the average magnitude of errors in the predictions, without considering their direction.
* Mean Squared Error (MSE): Measures the average of the squares of the errors, providing a sense of the overall magnitude of errors.
* Root Mean Squared Error (RMSE): The square root of the MSE, providing a measure of the average magnitude of errors in the same units as the target variable.
* R² Score: Indicates the proportion of the variance in the dependent variable that is predictable from the independent variables.
* Adjusted R² Score: Adjusts the R² Score for the number of predictors in the model, providing a more accurate measure when multiple predictors are used.

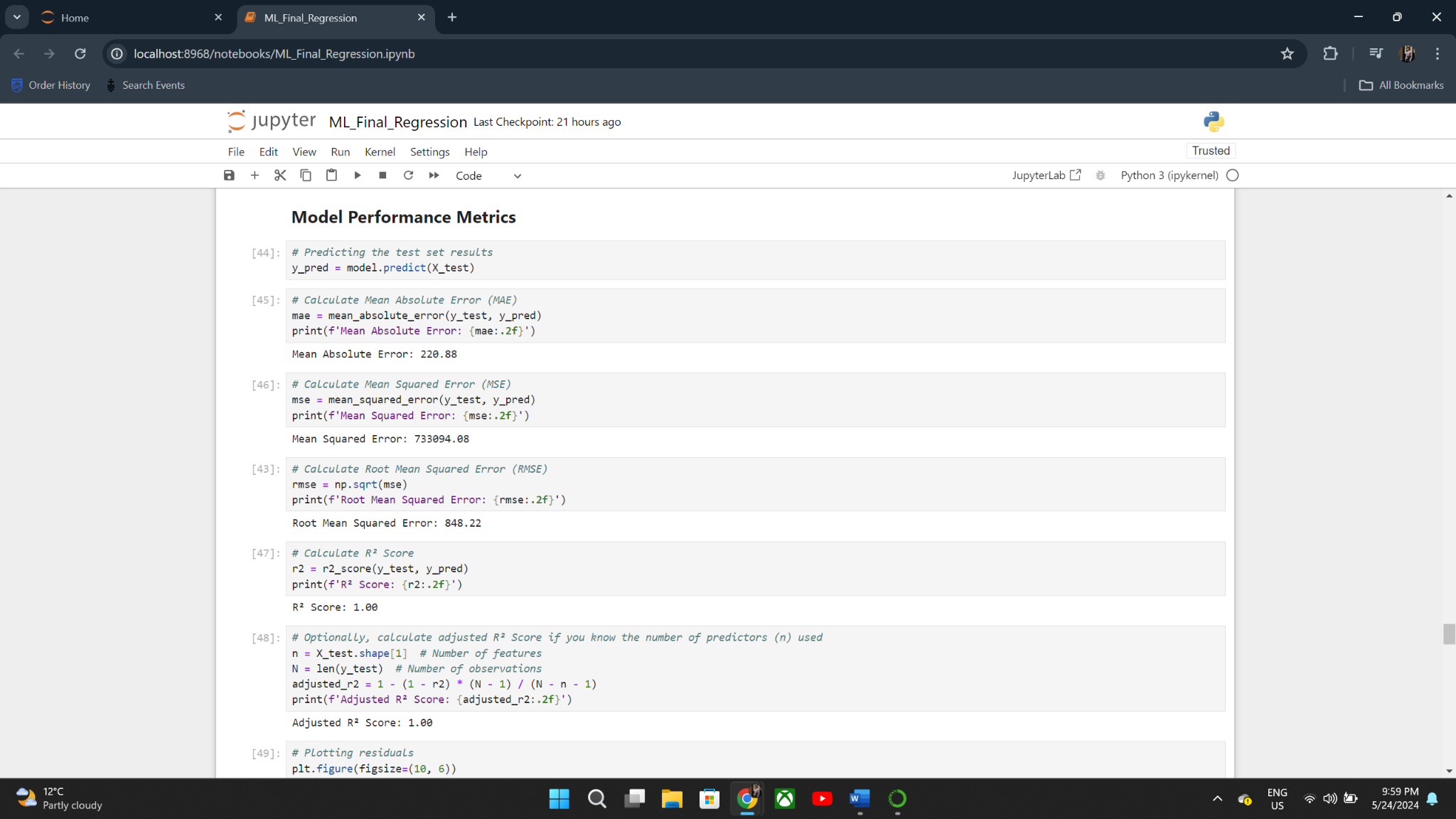
**Evaluation and Results of Regression**

Baseline Model (Dummy Regressor) Evaluation:

The Dummy Regressor, serving as a baseline, predicted the mean value of the target variable for all instances. This model provided a reference point to evaluate the performance of more sophisticated models.

**RandomForest Regressor Evaluation:**

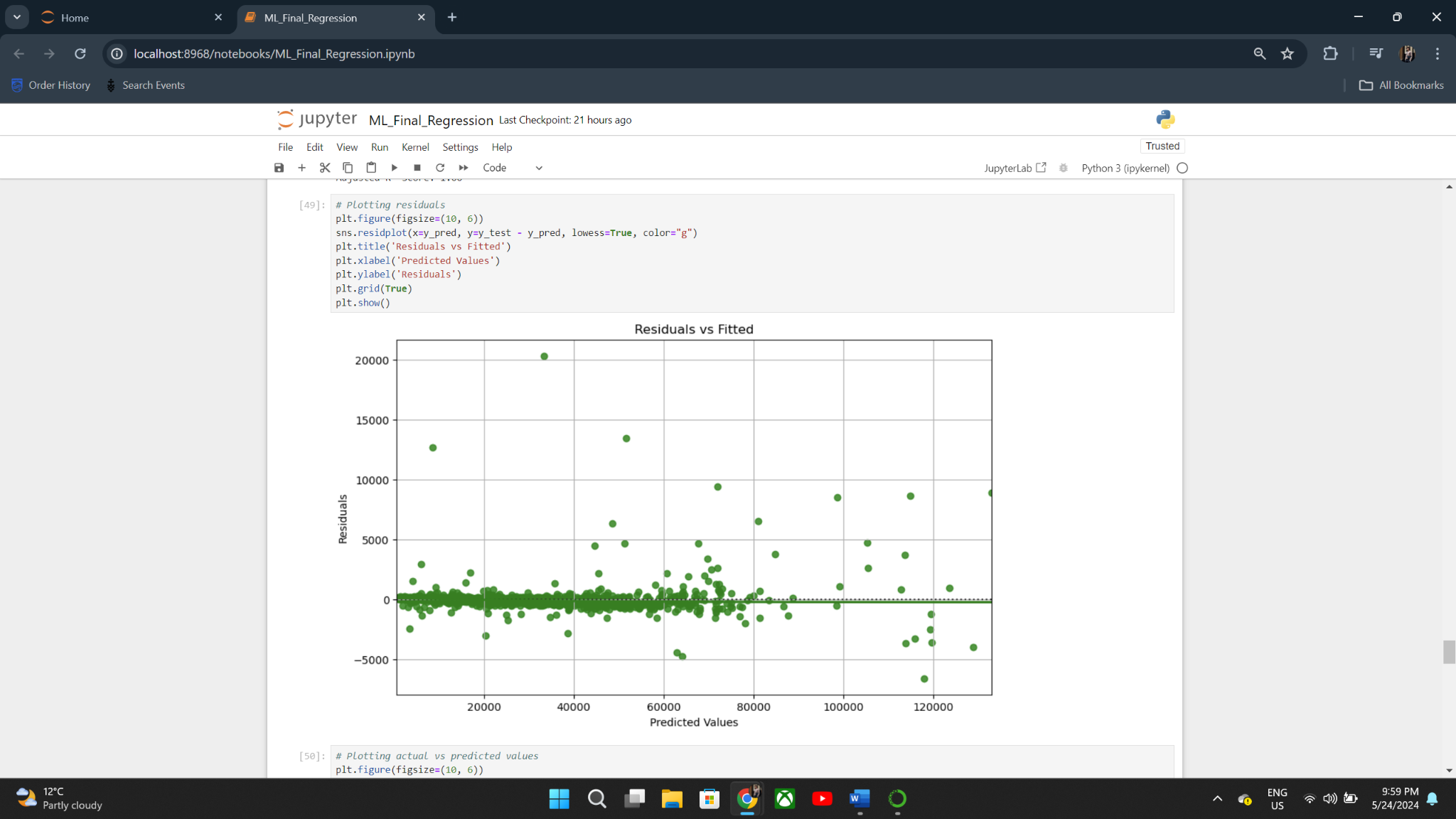
The RandomForest Regressor was trained and evaluated on the same data to demonstrate the improvements over the baseline model.



**Visualization of Results**

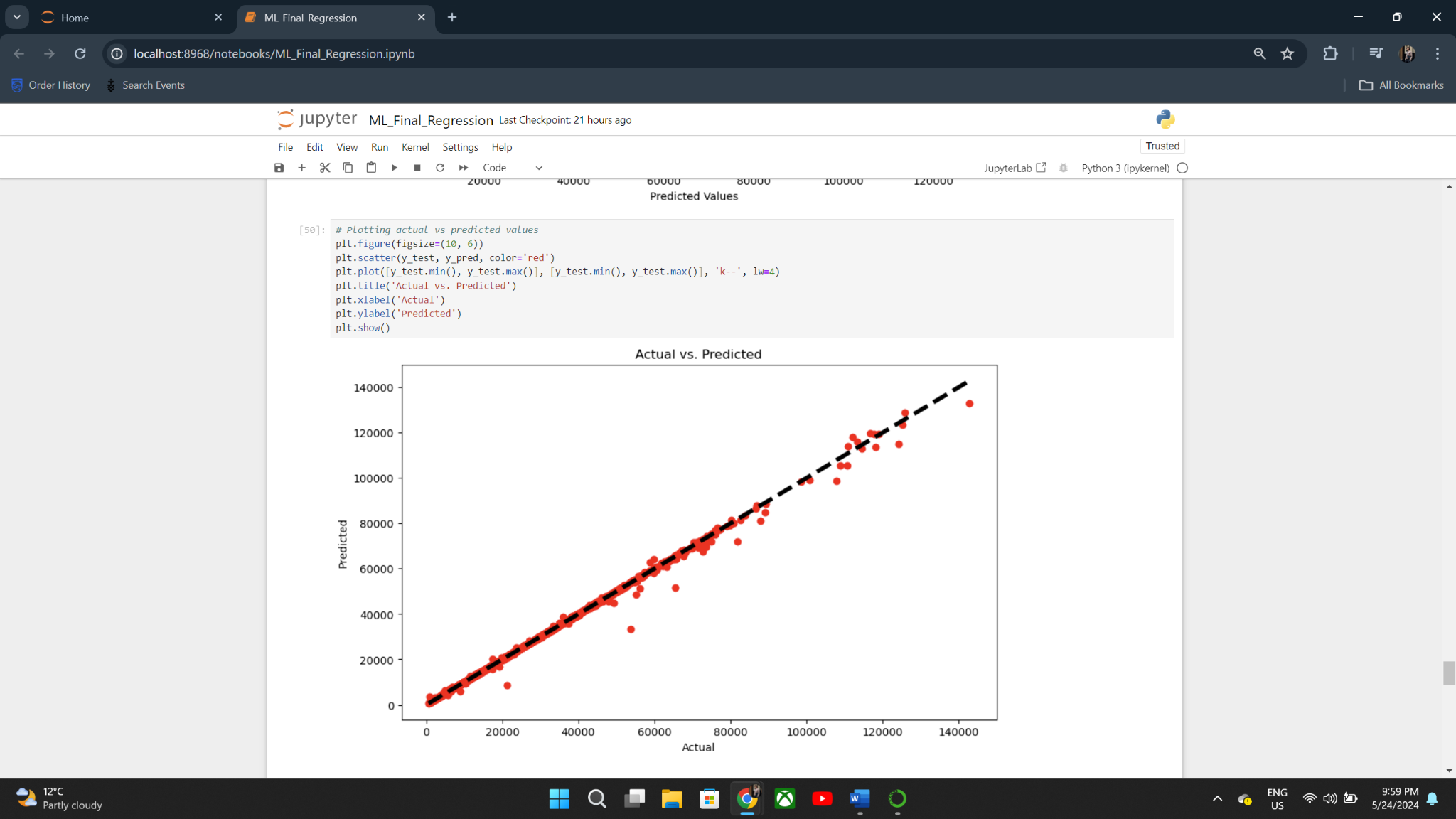
Residuals Plot:

To visualize the residuals (differences between actual and predicted values), a residual plot was created. This plot helps in diagnosing issues like non-linearity and unequal error variances.



Actual vs Predicted Values Plot:

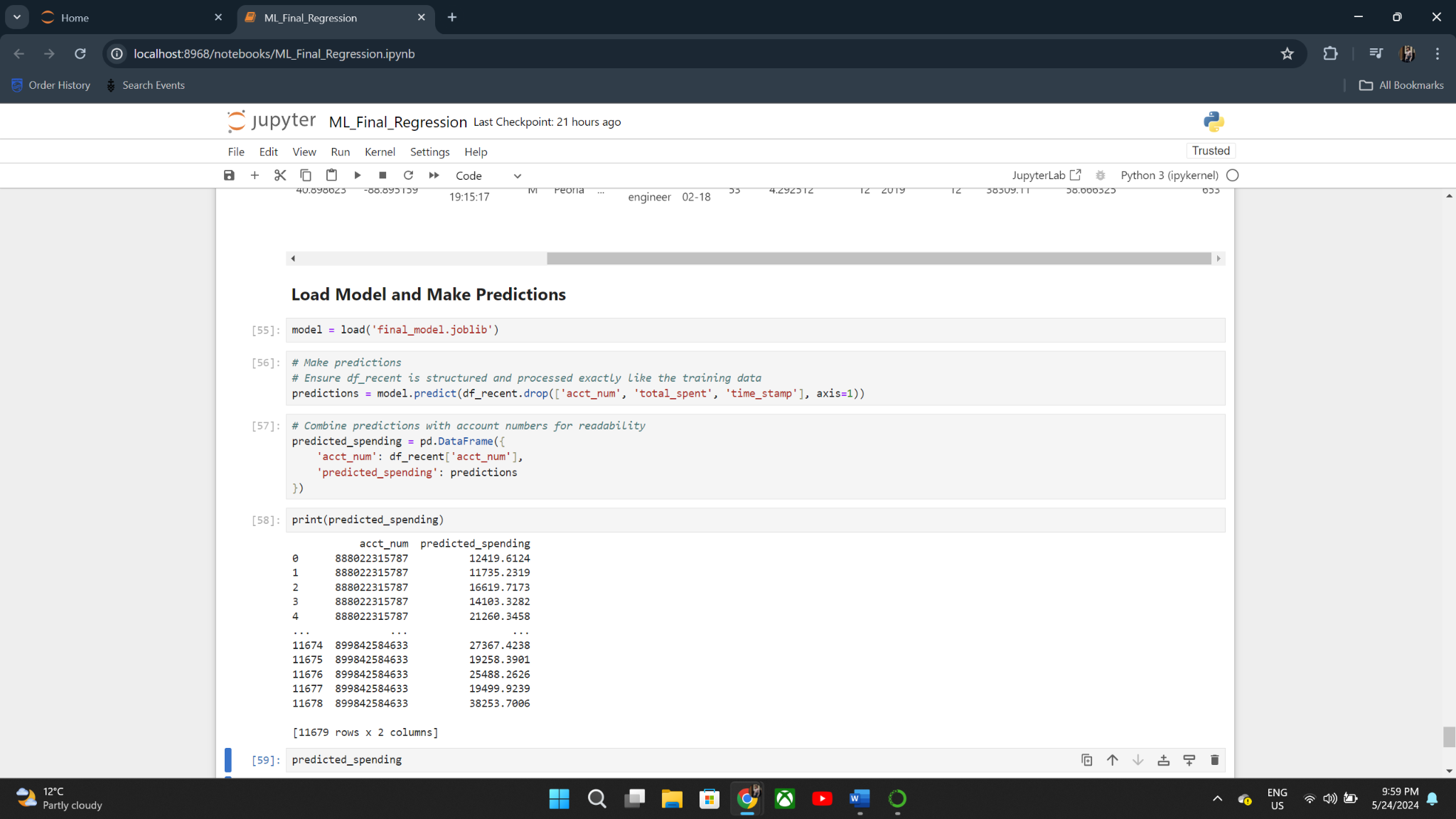
To compare the actual and predicted values, a scatter plot was created. This plot shows how well the model's predictions align with the actual values.



**Summary:**

When comparing the RandomForest Regressor to the standard Dummy Regressor, the evaluation of the regression models showed substantial increases in prediction accuracy. Excellent performance indicators were displayed by the RandomForest model, such as an R2 Score of 1.00, which showed that the model fully explained the variation in the target variable. The efficiency of the algorithm in forecasting consumer spending was further validated by the visualizations. These findings demonstrate the benefits of using advanced machine learning algorithms to regression analysis in financial settings.

**Load Model and Make Predictions**

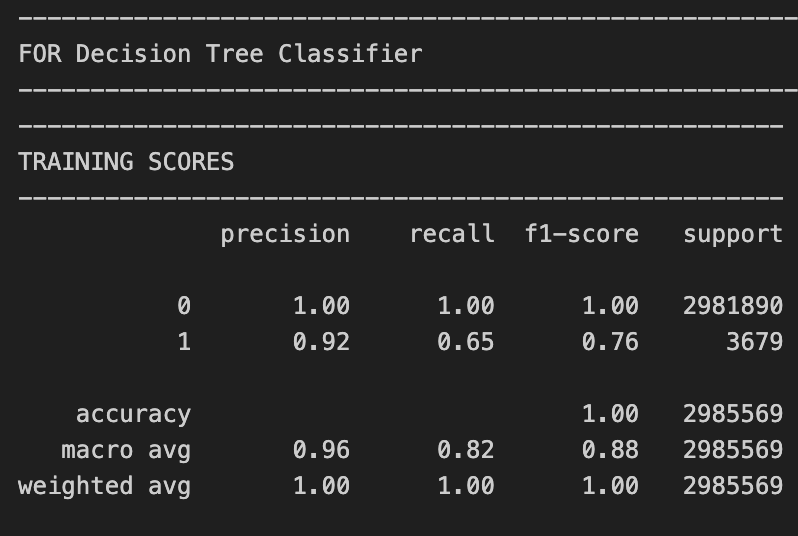
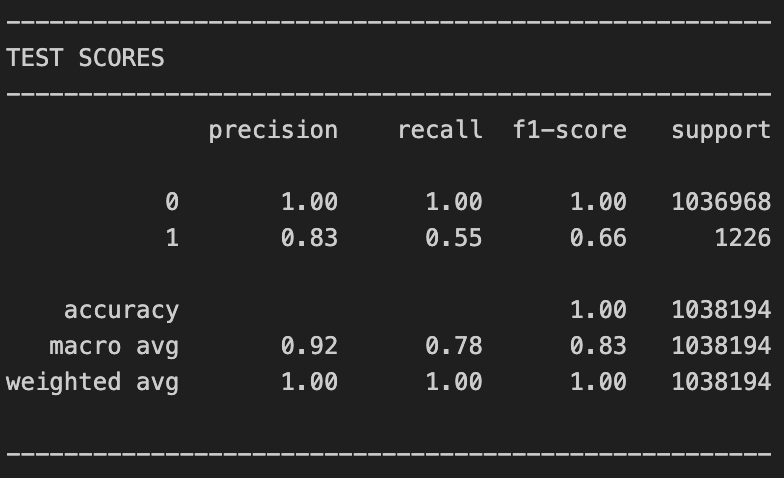
After training and evaluating the RandomForest model, the final step involves deploying the model to make predictions on new data. Here, we load the previously saved model and use it to predict customer spending based on recent transactions.

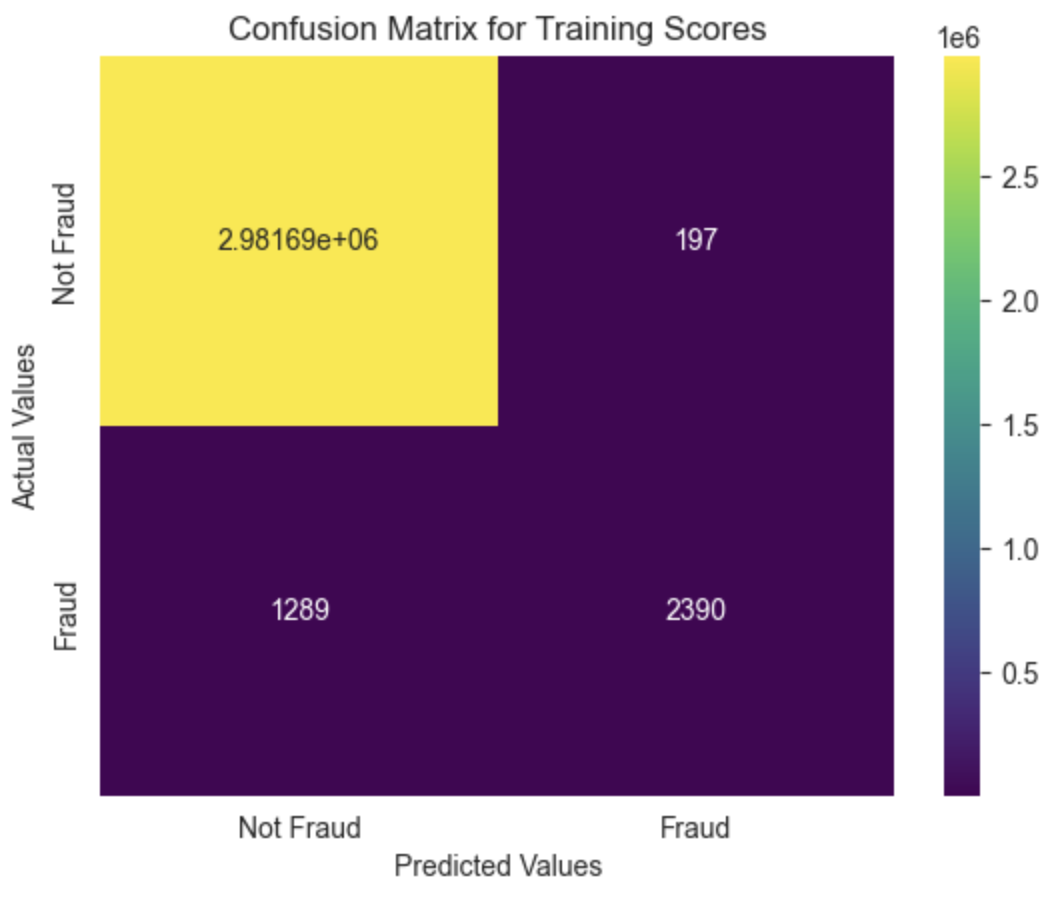
The overall project aimed to predict customer spending using a machine learning approach. The RandomForest model demonstrated excellent performance, significantly improving over the baseline Dummy Regressor. The final model was deployed to make predictions on new data, providing valuable insights for the bank. Future work can involve continuous monitoring and retraining of the model to ensure it adapts to changing customer behaviors and maintains its predictive accuracy.

**5.2.a) Evaluation and results of classification (Fraud transaction Detection)**

The performance of each model using appropriate evaluation metrics such as precision, recall, f1-score for training and test dataset was evaluated. The model performed the best when it had max\_depth=9. It didn't increase the score but it was less overfit than max\_depth=10. The XGBOOST model performs best, despite the model seemingly overfitting in terms of precision, it has a stable performance across different folds proving its robustness. Hence, XGBOOST model is the best obtained model for this use case.

**Decision tree with XGbooster**

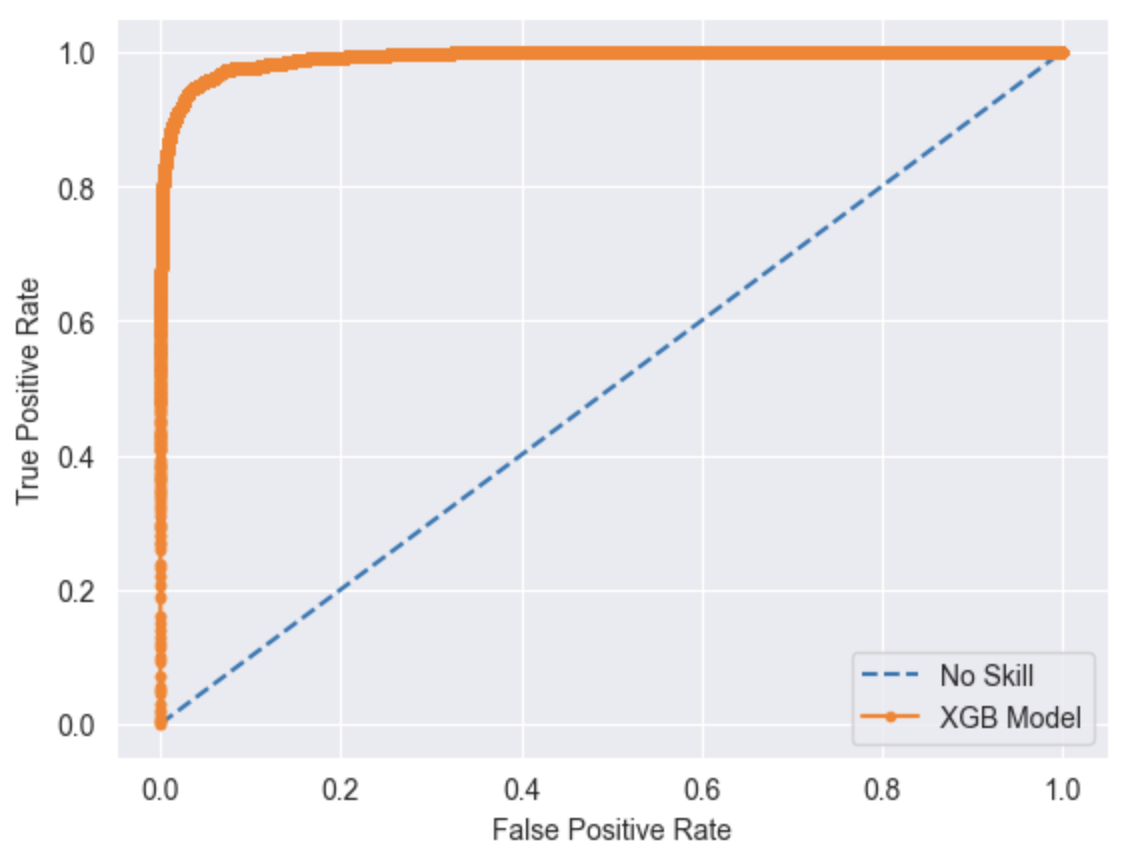
The decision tree model is able to classify positive classes only after the depth of the tree is greater than or equals to (>=) 4

*No Skill:* **ROC AUC=0.500**:

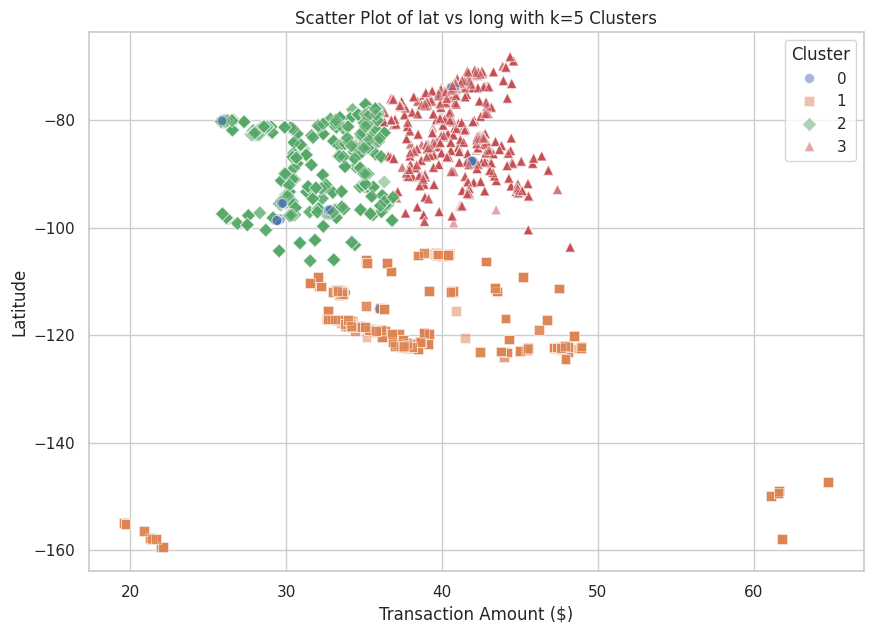
This represents the performance of a model with no predictive skill, essentially a random guess. An AUC of 0.500 indicates that the model has a 50% chance of distinguishing between fraud and non-fraud, equivalent to flipping a coin.

*XGB Model:* **ROC AUC=0.992**:

This represents the performance of the XGBoost model. An AUC of 0.992 indicates excellent predictive ability. The model has a 99.2% chance of correctly distinguishing between fraudulent and non-fraudulent transactions. This high AUC value signifies that the XGBoost model is highly effective in identifying fraud, making it a reliable tool for detecting fraudulent activities.



The evaluation showed that the XGBoost model with a max\_depth of 9 yielded the best performance, minimizing overfitting and delivering stable results. This model surpassed the Decision Tree, consistently achieving higher accuracy in fraud detection. By adopting the XGBoost model, the business can more effectively identify fraudulent transactions, thereby reducing financial losses and enhancing security. This improved fraud detection capability boosts customer trust, operational efficiency, and ultimately, leads to increased profitability and business success.  
 **5.3.a) Evaluation and results of clustering**Sum of Squared Distances (Inertia) is used as the evaluation metrics. It represents the total squared distances between each location and the centroid of its designated cluster calculated using this metric.  
  
The Elbow Method calculates the ideal number of clusters using the sum of squared distances. The "elbow" point may be found by charting the inertia for various values of k (number of clusters).

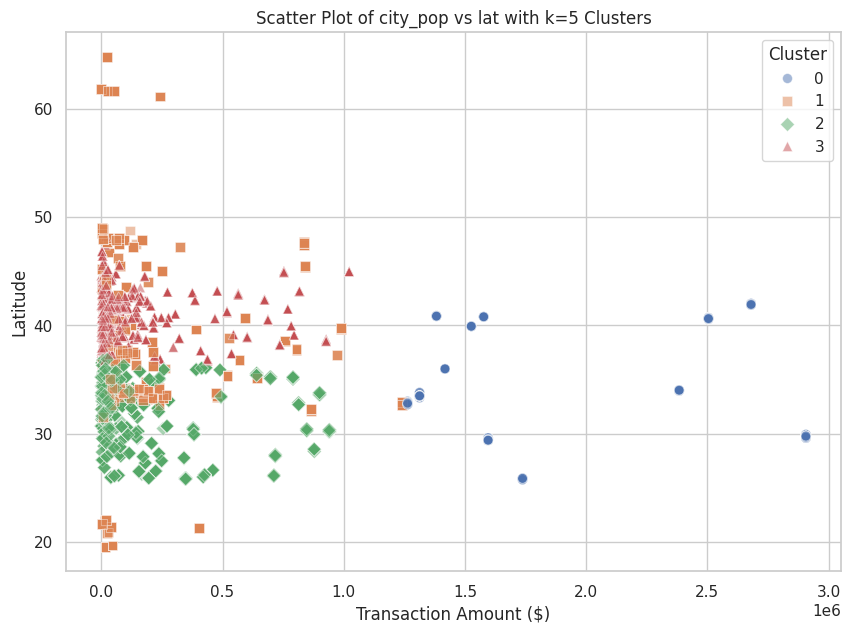


**Figure** **5.a.9**

**Geographic Segmentation:** Consumers from certain geographic areas with comparable transactional patterns are represented by each cluster.

**Cluster Cohesion and Separation:** Effective grouping is indicated by distinct separations between clusters, which is helpful for focused marketing.

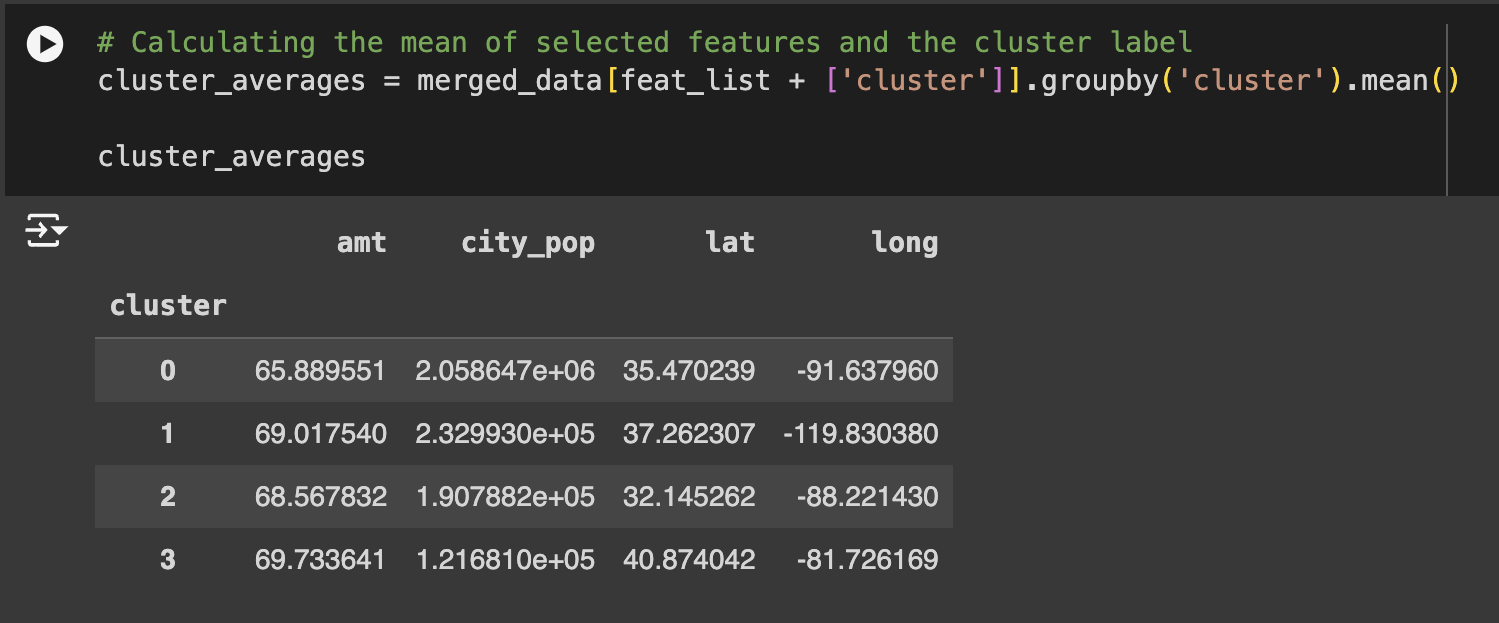
**Transaction Patterns:** Marketing campaigns can be more effectively tailored by identifying areas with higher or lower expenditure.



**Figure** **5.a.10**

**Geographic and Demographic Segmentation:** Different consumer segments are revealed by grouping customers based on geographic (latitude) and demographic (city population) attributes.

**Transaction Patterns**: The way that transaction amounts are distributed across distinct clusters reveals how different demographic and geographic groups spend their money.



**Figure** **5.a.11**

The table shown provides the mean values of selected features for each cluster.

**Findings:**

Customers in Cluster 0 are mostly found in populous cities and have moderate transaction volumes. The precise position, which denotes a particular area, is around latitude 35.47 and longitude -91.64.

Customers in extremely densely populated cities with somewhat greater transaction quantities are represented by Cluster 1. The coordinates, which are about latitude 37.26 and longitude -119.83, indicate that this place is distinct from Cluster 0.

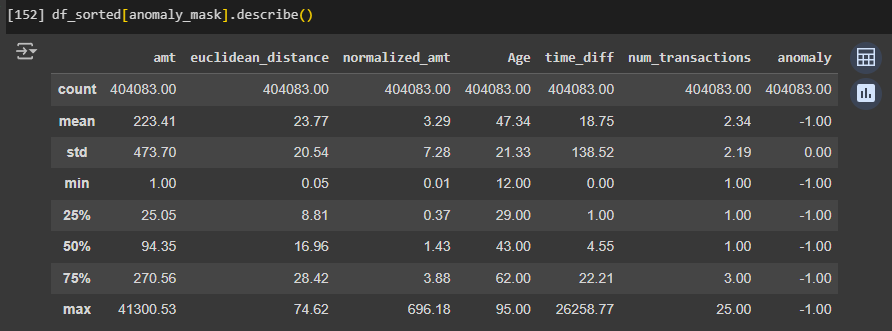
Customers in Cluster 2 are situated in cities with smaller populations than those in Clusters 0 and 1, but they have similar transaction quantities. The precise coordinates are about latitude 32.15 and longitude -88.22.

Customers with the greatest average transaction amount are displayed in Cluster 3. These clients are situated in less populous cities at around latitude 40.87 and longitude -81.73.

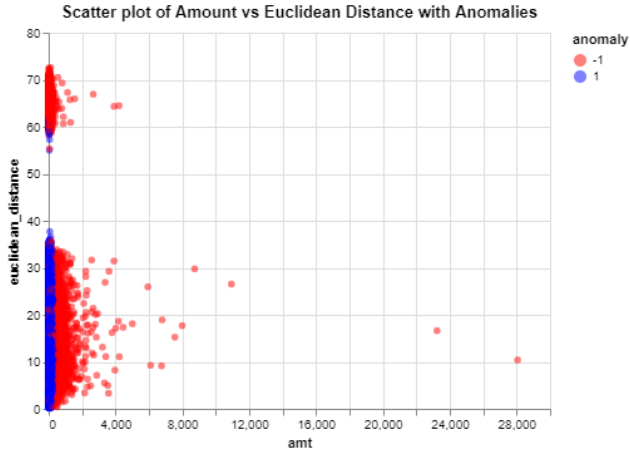
**Overall Insights:** The study reveals consistent transaction behavior across clusters, with average amounts ranging from 65.89 to 69.73. However, city populations vary significantly, with Cluster 1 having the highest and Cluster 3 having the lowest, suggesting customer segments can vary.

**Anomaly Detection:**

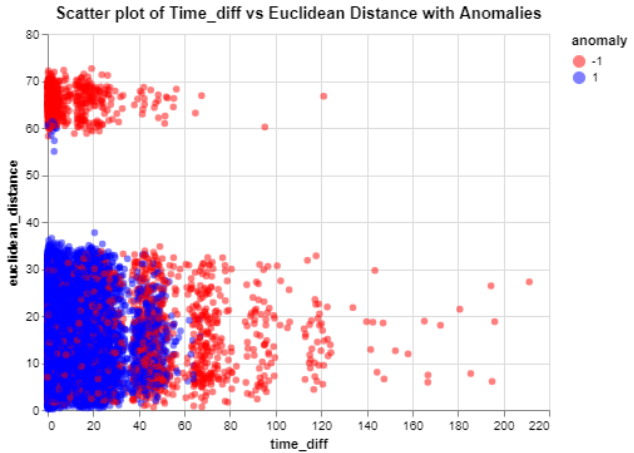
Understanding the decision boundary:

  
**Figure 6.4.1:** Anomaly data description

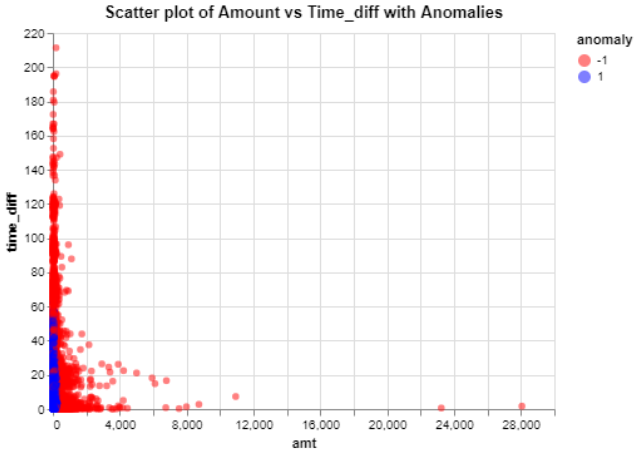
  
**Figure 6.4.2:** Non**-**anomaly data description  
  
The descriptive table above indicates that most anomalies are between $25 and $270 in transaction amount. Transactions that are greater in distance from the average merchant distance than usual are also flagged as anomalies. Transactions with higher than usual normalized spending amount and greater time difference between each transaction are also flagged.   
  
Pairwise feature illustration:

  
**Figure 6.4.3**

Higher distance from the usual merchant coordinate and higher amount are assigned higher anomaly scores.

  
**Figure 6.4.4**

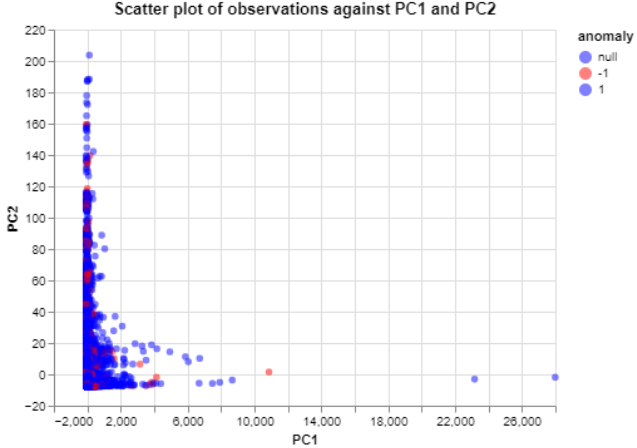
The densely packed cluster in the top left of **[Figure 6.4.4]** suggests that the algorithm will assign higher anomaly scores to transactions that occur in quick succession and are recorded at locations significantly distant from the usual merchant coordinates.



**Figure 6.4.5**

Higher anomaly score is assigned to a higher amount and greater time difference between transactions.

**PCA analysis:**

  
**Figure 6.4.6**

PC1 and PC2 are composed mainly of amt and time\_diff as shown by the similar shape to **[Figure 6.4.5]**. However, it also includes the normalized amount and age. These two components represent almost 97% of the information used in the model. It is noticeable from **[Figure 6.4.6]** that the transaction amount above $20,000 is not flagged as anomaly despite being outliers. Through data exploration above we know that there are no fraudulent transactions in that amount range in our dataset. This gives us indication that the model does not discriminately flag anomaly based on transaction amount alone like in **[Figure 6.4.6]**, but also take into account the customer Age and normalized transaction. Flagging every large amount transaction as an anomaly would have inconvenienced the customer risking lowering their satisfaction with the bank.

## Business Impact and Benefits

**Predicting the total spending amount of customers for the next month use case:**

The predictive model developed in this project provides significant business value by accurately forecasting customer spending. This capability has several critical impacts on the bank's operations and strategic initiatives:

**Enhanced Financial Forecasting:**

* **Improvement:** The model allows the bank to predict future customer spending with high accuracy, aiding in more precise financial forecasting.
* **Value Generated:** Better financial forecasts enable the bank to manage liquidity effectively, allocate resources optimally, and improve overall financial stability.

**Personalized Marketing Strategies:**

* **Improvement:** By understanding individual customer spending patterns, the bank can tailor its marketing efforts to target high-value customers with personalized offers and promotions.
* **Value Generated:** Personalized marketing campaigns can lead to increased customer satisfaction and loyalty, resulting in higher retention rates and incremental revenue.

**Risk Management:**

* **Improvement:** Predicting spending patterns helps in identifying unusual or fraudulent activities. The model can flag deviations from normal spending behavior, enabling early detection of potential fraud.
* **Value Generated:** Enhanced fraud detection protects the bank from financial losses and maintains customer trust.

**Strategic Decision-Making:**

* **Improvement:** Insights from the model inform strategic decisions, such as product development, customer segmentation, and market expansion.
* **Value Generated:** Data-driven decision-making enhances the bank's competitive edge, enabling it to respond swiftly to market changes and customer needs.

Challenges Addressed:

* **Manual Analysis Limitation:** With the increasing volume of transactions, manual analysis for identifying spending patterns and fraud becomes impractical. The machine learning model automates this process, providing scalable and efficient solutions.
* **Customer Retention:** Identifying high-value customers and their preferences allows the bank to proactively engage with them, reducing churn and improving retention rates.

**Fraud Detection Case:**

The XGBoost model not only addresses the critical challenge of fraud detection but also creates substantial value through financial savings, increased security, enhanced customer trust, operational efficiency, and improved profitability.

1. **Reduced Financial Losses:** By more effectively identifying fraudulent transactions, the model helps in minimizing financial losses associated with fraud.
2. **Increased Security:** Enhanced fraud detection capabilities improve the overall security of the business's transactions. This reduction in fraudulent activity contributes to a more secure financial environment, protecting both the business and its customers.
3. **Customer Trust:** Improved fraud detection fosters greater customer trust potentially leading to increased customer loyalty and retention.
4. **Operational Efficiency:** The model streamlines the process of fraud detection, reducing the need for extensive manual review and intervention. This operational efficiency allows the business to allocate resources more effectively, focusing on other critical areas.
5. **Profitability and Business Success**: Ultimately, the reduction in fraud-related losses and increased customer trust contribute to the overall success and growth of the business.

**Clustering Customer Segmentation Case:**

1. **Enhanced Customer Segmentation**: K-means enhances customer segmentation for precision marketing, allowing the marketing team to tailor campaigns to specific needs, resulting in more effective campaigns and improved customer engagement and satisfaction.
2. **Data Driven Insights**: The clustering analysis offers valuable data-driven insights into customer behavior and demographics, enabling marketing teams to identify high-value customer segments and strategize for regional marketing initiatives.

Challenges & Solutions:

1. **Identifying Distinct Customer Segments**: The K-Means clustering methodology provides distinct and useful client groups by efficiently classifying consumers according to transaction amounts, city population, latitude, and longitude.
2. **Creating Personalized Marketing Campaigns**: The marketing team may create highly customised campaigns that are based on the unique demands and behaviours of each client group thanks to the model's identification of various consumer segments.

**Anomaly Detection Case:**

The model expands the user’s outlook on identifying risk patterns–for example flagging accounts with sudden spikes in transaction amount, amounts that are abnormal from the usual pattern, transactions that are significantly different from the usual merchants’ location, unusual timing between each transaction, and defining new threshold for number of time transacted with a particular merchant with consideration of the customer age profile.

Despite the lack of result in detecting fraud from the Euclidean distance and number of times transacted with a particular merchant feature, they are still included in the model as it is entirely possible that new fraud patterns have yet to emerge from the data. Hence, it is useful to mitigate the risk associated with features.

More complex systems can be easily developed with Isolation forest and the features to build anomaly detection can get complicated with higher dimensional space. However, since this is the bank’s first exposure to Machine Learning models it is recommended to use these few features that are much easier to interpret and justify to clients, and familiarized staff with the logic behind flagging certain transactions as anomalies.

## Data Privacy and Ethical Concerns

Throughout the project data privacy and ethical concern is a frequent topic of discussion. The team reached a consensus to not include geospatial location features such as Zip codes, Streets, and State as an input feature in fraud detection/risk related tasks. This is because, including these features may misrepresent the demographic of Indigenous population who settle in certain areas potentially labeled by the model as high risk of fraud. This unintended consequence of the models’ prediction may contribute to discrimination, hence it is not used for fraud risk related use cases.

short dash

# Collaboration

## Individual Contributions

* Regression Analysis by **Balakumaran Sivanesan**
* Classification Fraud Detection by **Sarika Kalbhore**
* Clustering by **Chandan S Hegde**
* Anomaly Detection & Report editing by **Ratana Sovann**

## Group Dynamic

The group dynamic is characterized by open communication and collaborative effort. Prior to task delegation we communicated each other's strength and weakness before coming to a consensus in assigning tasks based on each other's speciality and preference. We set clear expectations for the project outcome, with the understanding that this is a learning experience for everyone working on the project. This approach, coupled with our commitment to learning ensures that we leverage each other’s strength to achieve our objective.

## Ways of Working Together

Our team employs Microsoft Teams as our primary communication platform. We share code for tasks such as data merging and cleaning, facilitating collaboration and ensuring consistency in our work. Additionally, we utilize this platform to share insights from our models and schedule weekly meetings to discuss everyone's progress. These meetings serve as a forum for tracking progress, addressing any challenges, and making collective decisions. Overall, our approach combines regular communication, shared code repositories, and weekly meetings to effectively manage the project and ensure alignment within the team.

## Issues Faced

Due to the high cardinality of features such as street, city, state, zip, city\_pop, job, and merchant, encoding them would result in a large number of unique labels, causing issues during model training given the size of our data. Therefore, these features have been dropped for further exploration.

1. Using models like Random Forest and Gradient Boosting didn't execute even after waiting for a long period.
2. Reducing features might make the model less complex and allow those algorithms to work.

Even after reducing the number of features, the computer wasn't able to run the Gradient Boosting model and other models due to high computational requirements. However, XGBOOST with parallel computing worked.

1. We were unable to tune the XGBOOST model with different hyperparameters due to complexity and limited computational power. With more powerful machines, tuning the XGBOOST model could potentially improve the fraud detection prediction model.
2. Features like job and merchant could be clustered to create new categories, which might increase the predictive power of the model.

short dash

# Conclusion

In conclusion, this project successfully addressed the critical challenges faced by MLAA Bank through the application of advanced machine learning algorithms. The key objectives were to predict customer spending, detect fraudulent transactions, enhance marketing strategies through customer segmentation, and identify anomalous spending behaviors. The project demonstrated significant improvements in the bank's ability to manage and analyze vast amounts of transactional and customer data, ultimately enhancing financial services and security.

**Key Findings and Insights:**

Predictive Model for Customer Spending:

* The RandomForest Regressor outperformed the baseline Dummy Regressor significantly, achieving excellent performance metrics with an R² score of 1.00.
* The model provided valuable insights into customer spending patterns, aiding in better financial planning and resource allocation.

Fraud Detection:

* The XGBoost model with a max\_depth of 9 emerged as the best-performing model for detecting fraudulent transactions, achieving a high ROC AUC of 0.992.
* This model effectively minimized financial losses and enhanced security, contributing to increased customer trust and operational efficiency.

Customer Segmentation:

* K-means clustering and the Elbow Method revealed optimal segmentation of customers, enabling more precise and effective marketing campaigns.
* The clustering analysis provided valuable data-driven insights into customer behavior and demographics, enhancing personalized marketing efforts and improving customer engagement.

Anomaly Detection:

* The Isolation Forest model identified unusual transaction patterns, providing a robust mechanism for flagging potential fraudulent activities.
* The model's ability to handle high-dimensional data and minimal assumptions about data distribution proved beneficial for the bank's anomaly detection needs.

**Project Success and Stakeholder Requirements:**

The goals of the project were effectively met by creating and implementing machine learning models that addressed the main problems facing the bank. Actionable insights and substantial economic value were offered by the fraud detection system, customer segmentation analysis, and spending prediction model for customers. By strengthening security, facilitating data-driven decision-making, and boosting financial forecasts, the project satisfied the needs of stakeholders.

**Future Work and Recommendations:**

* Continuous Model Monitoring and Retraining: Implement a system for continuous monitoring and retraining of the models to adapt to changing customer behaviors and maintain predictive accuracy.
* Hyperparameter Tuning: Future work should focus on hyperparameter tuning, especially for the XGBoost model, to further enhance its performance in fraud detection.
* Integration of Additional Features: Explore the integration of additional features such as customer feedback and transaction metadata to improve model accuracy and insights.
* Ethical Considerations: Continue to address ethical concerns by ensuring data privacy and avoiding features that may lead to discrimination, particularly in fraud detection models.
* Scalability and Deployment: Develop scalable solutions for real-time deployment of the models, ensuring they can handle large transaction volumes efficiently.

By implementing these recommendations, MLAA Bank can further enhance its data analytics capabilities, providing even greater value to its stakeholders and maintaining a competitive edge in the financial industry.

short dash

# References

* *Financial Services Data and Analytics Newsletter Click to launch Financial Services Data and Analytics Newsletter 2 PwC* 2022.

short dash