1 Introduction

1.1 Aim and Objectives

The primary aim of this study is to analyse a credit card fraud dataset to the identify patterns and gain a comprehensive insight that can aid in the detection and prevention of fraudulent activities. This will assist the banking industry in making critical step by which their customers use their credit card for transactions.
The transaction data set that is being explored will also assist the company in exploring their customer spending habits, age groups that transact the most, time period where transactions occur, helping the company plan resource allocation there by optimising response time from their customer care departments (Author, 2024).

The objectives that will be carried out in this study are as follows:

1. Perform exploratory data analysis on the fraud dataset to understand the nature and characteristics of the data.
2. Identify key factors and variables that contribute to the likelihood of fraudulent transactions.
3. Develop strategies and techniques for detecting and preventing fraud based on the insights gained from the analysis.
4. Compare the effectiveness and efficiency of SAS and Tableau in performing fraud analytics and data visualization.

1.2 Background Information

Fraud is a pervasive issue that affects businesses across various industries, leading to significant financial losses and reputational damage. With the advent of digital transactions and e-commerce, the risk of fraud has escalated, necessitating robust fraud detection and prevention mechanisms. This study aims to leverage data analytics and historical transaction data to uncover insights and develop robust fraud prevention strategies and anomalies that can indicate fraudulent activities, thereby enabling organizations to take proactive measures to mitigate fraud risks (Author, 2024)..

1.3 Description of the Fraud Dataset

The fraud dataset used in this study is a comprehensive collection of 1,852,394 transactions, which was obtained from the Kaggle website. The data set comprises of 23 variables representing various attributes of financial transactions including both legitimate and fraudulent transactions. The data set also captures transaction amount, customer demographics, geographic location, transaction time, and customer details like, date of birth, occupation, and gender. Table 1 below highlights the important variable that will aid the analysis and detection of fraud.

Variable Namae	Description
Transaction Date Time	This shows the date and the time when a transaction was executed
Category	This describes the category of goods being purchased. Customer spending habit can be computed using the columns. Targeted marketing can also be achieved using this field. As well as pinpointing the category which fraudsters target to defraud customers.
Amount	This is the amount for the individual transaction
Gender	The describes the customer's gender. This variable can highlight if a certain gender is often targeted for fraud.
City	The city in which the transaction occurred. This variable can highlight how customer move across the state.
State	The state in which the transaction occurred. Segmentation of
Job	The occupation which the customer holds. This is a category variable that describes the job of the customer
Date of Birth	The date the customer was born. Using this data we can segment the customer into age group, thereby giving us insight to the spend habit of different age group.
Is_fraud	Discrete numerical variable indicating fraud or valid transaction

Table 1 – key variables and description in the *credit card fraud data set.*

2. Analytics Design

The analytics design employed in this study follows a structured approach, incorporating various data mining techniques and statistical methodologies. The process begins with data preprocessing, which involves cleaning and transforming the raw data to ensure its quality and suitability for analysis. Subsequently, exploratory data analysis (EDA) is conducted to gain a deeper understanding of the dataset's characteristics and identify potential relationships between variables.

To illustrate the analytics design process, a Unified Modelling Language (UML) activity diagram is presented below:

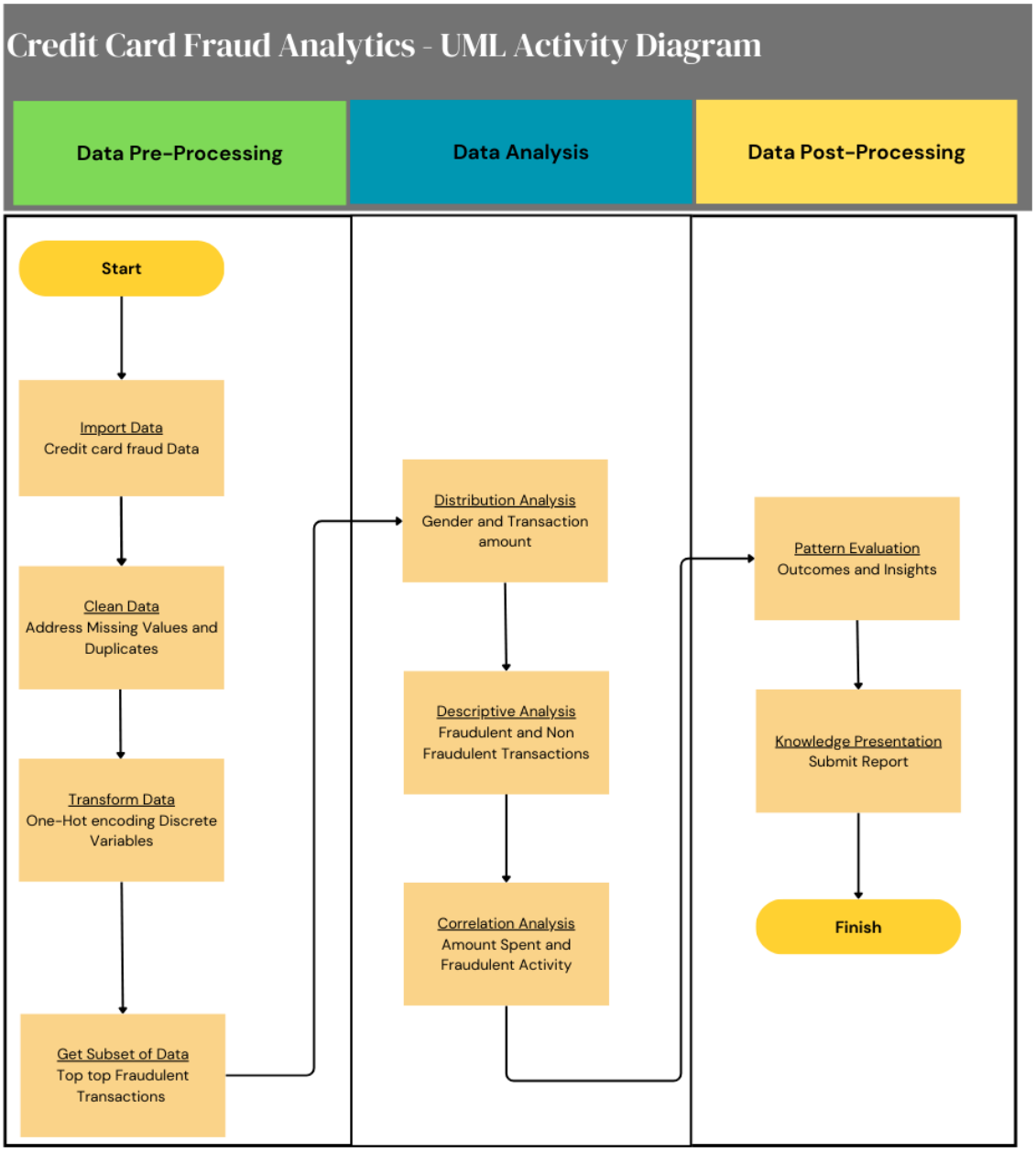


Fig 2.1 Unified Modelling Language (UML) activity diagram

3. Data Analytics

Tableau Desktop and SAS academy were the analytic used to analyse and carry out the prediction of fraud in this section.

3.1 Data Pre-Processing

Before proceeding with the analysis, several data pre-processing steps were performed to ensure the quality and consistency of the data.

3.1.1 Data Cleaning

The dataset was thoroughly examined for missing values, duplicates, and outliers. Missing values were imputed using appropriate techniques and removed when they constituted a significant portion of the data. Duplicate transactions were identified and removed to avoid redundancy and skewed results.

3.1.2 Data Transformation

Categorical variables were encoded using techniques such as one-hot encoding or label encoding to make them suitable for analysis. The encoded numerical variable is_fraud was replaced by a new binary variable that is for the unique integer value.

3.1.3 Data Subsetting

The data set is about 1.8 million observations and this quite large and will cause performance issues when running the analysis. To resolve this issue, a subset of the data was taken. This was done by looking at the top ten highest transaction by state and the top ten highest fraudulent transaction. New York was the state with the highest fraudulent transaction and had the second highest transaction. Our focus moving forward will be focused on the state of New York.

Top 10 transaction by state

	Transaction_Type	Transaction_Type
	Non Fraudulent Transaction	Non Fraudulent Transaction
	TotalTransactionAmount	NumberOfTransaction
	Sum	Sum
StateOfTransaction		
AL	$3,661,568	58243.00
CA	$5,700,931	80093.00
FL	$4,274,728	60441.00
IL	$4,157,768	61888.00
MI	$4,523,936	65526.00
MO	$3,606,432	54642.00
NY	$8,140,945	118689.00
OH	$4,631,536	66267.00
PA	$7,898,197	113601.00
TX	$9,239,298	134677.00

Fig 3.1 Table showing the top 10 transactions by state.

Top 10 fraudulent transaction by state

	Transaction_Type	Transaction_Type
	Fraudulent Transaction	Fraudulent Transaction
	TotalTransactionAmount	NumberOfTransaction
	Sum	Sum
StateOfTransaction		
AL	$149,072	278.00
CA	$205,939	402.00
FL	$173,648	334.00
IL	$164,099	324.00
MI	$158,766	299.00
MN	$148,535	280.00
NY	$395,295	730.00
OH	$192,000	360.00
PA	$307,432	572.00
TX	$318,117	592.00

Fig 3.2 Table showing the top 10 fraudulent transactions by state.

3.2 Exploratory Data Analysis

To gain insights into the characteristics of the credit card fraud dataset and identify potential patterns and relationships an exploratory data analysis (EDA) techniques were employed.

As shown in the previous fig

- Non-Fraudulent Transactions: Texas (TX) has the highest number of non-fraudulent transactions (134,677) and total transaction amount ($9,239,298), followed by New York (NY) and California (CA).
- Fraudulent Transactions: New York (NY) leads in fraudulent transactions both in count (730) and total amount ($395,295), followed by Texas (TX) and Pennsylvania (PA).

3.2.1 Distribution Analysis

EDA was carried out to gain a broad understanding of the dataset's features and attributes and uncover potential relationships between variables. Summary statistics, such as mean, median, and standard deviation, were calculated for numerical variables, while frequency distributions were examined for categorical variables.

summary statistics of customer gender

gender	N Obs	Variable	Mean	Median	Sum	Std Dev	Minimum	Maximum	N	Lower Quartile	Upper Quartile
F	68832	amt	71.3256138	44.1000000	4909484.65	147.9331102	1.0000000	10926.44	68832	9.2100000	85.0400000
		city_pop	199498.05	3487.00	13731849938	600665.96	69.0000000	2504700.00	68832	824.0000000	8830.00
		age	43.7014906	43.0000000	3008061.00	16.1110143	20.0000000	94.0000000	68832	29.0000000	58.0000000
		hourOfTransaction	12.9254998	14.0000000	889688.00	6.8091771	0	23.0000000	68832	7.0000000	19.0000000
M	50587	amt	71.6934396	50.6300000	3626756.03	200.5520735	1.0000000	27390.12	50587	10.9000000	82.7300000
		city_pop	85322.80	2258.00	4316224421	320755.43	176.0000000	1382480.00	50587	1666.00	10717.00
		age	52.0617945	51.0000000	2633650.00	19.2995604	19.0000000	93.0000000	50587	41.0000000	64.0000000
		hourOfTransaction	12.3112460	13.0000000	622789.00	6.9017918	0	23.0000000	50587	6.0000000	18.0000000

Fig 3.3 Summary Statistics of Gender with regards to transaction population and age.

Total number of transaction and average amount spent by gender

Transaction Amount		Transaction Amount			
		Number of Transaction		Average Amount Spent	
		Customer Gender		Customer Gender	
Number of Transaction	Average Amount Spent	F	M	F	M
119419	$71	68832	50587	$71	$72

Fig 3.4 Average spend and number of transactions by gender.

The above table shows the average amount spent by customer gender. On average Male customer spend more money per transaction when compared to female customer and the overall average. However, female customers carry out more transactions and in essence spend more money overall.

To understand the age distribution of the customers using credit card a box plot is shown in the below figure.

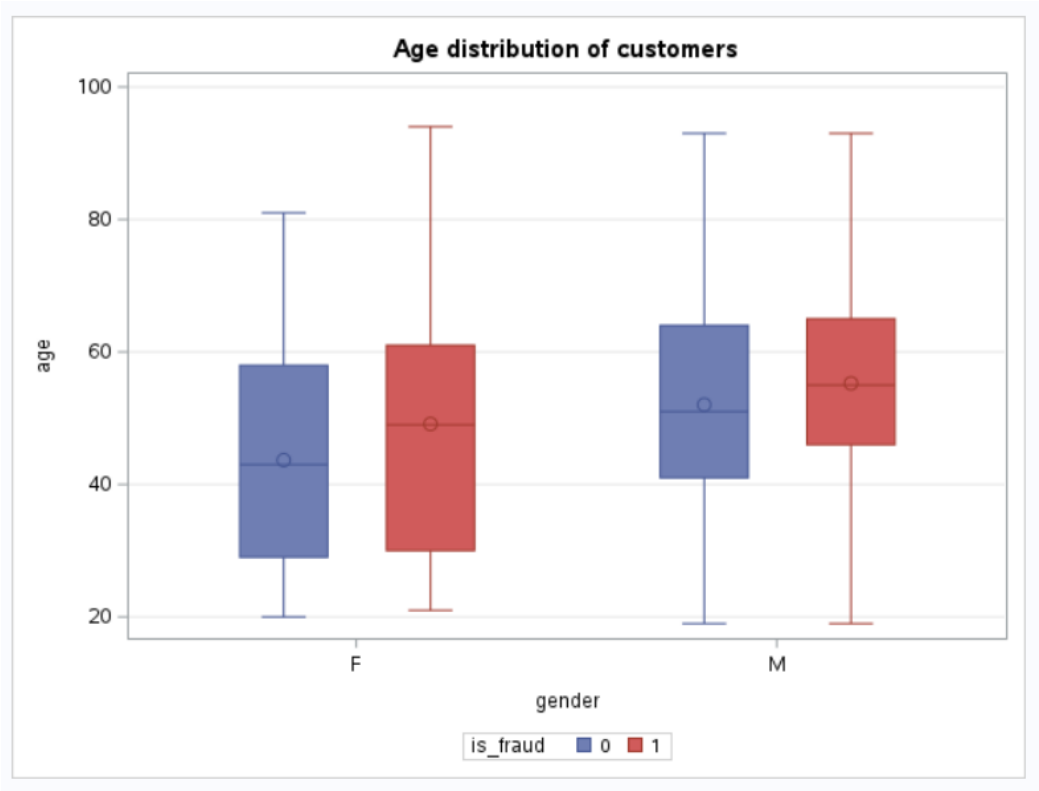


Fig 3.5 Box plot of age of customer by gender

In the above figure, it can be observed that on average the male customers are older than the female customer. With the female customer on average being in their forties and the male customers are in their fifties. Another trend that can be observed from the figure is that fraudulent transactions tend to happen in both genders. This can be alluded to the fact that older individuals are more susceptible to falling for scam as they are not technologically as

savvy as their younger counterparts. This insight os quite significant as stakeholders in the customer experience team can use this detial and reach out to older customers especially, giving them training on how to spot fraudulent transactions there by curbing the risk of them falling victim to scam.

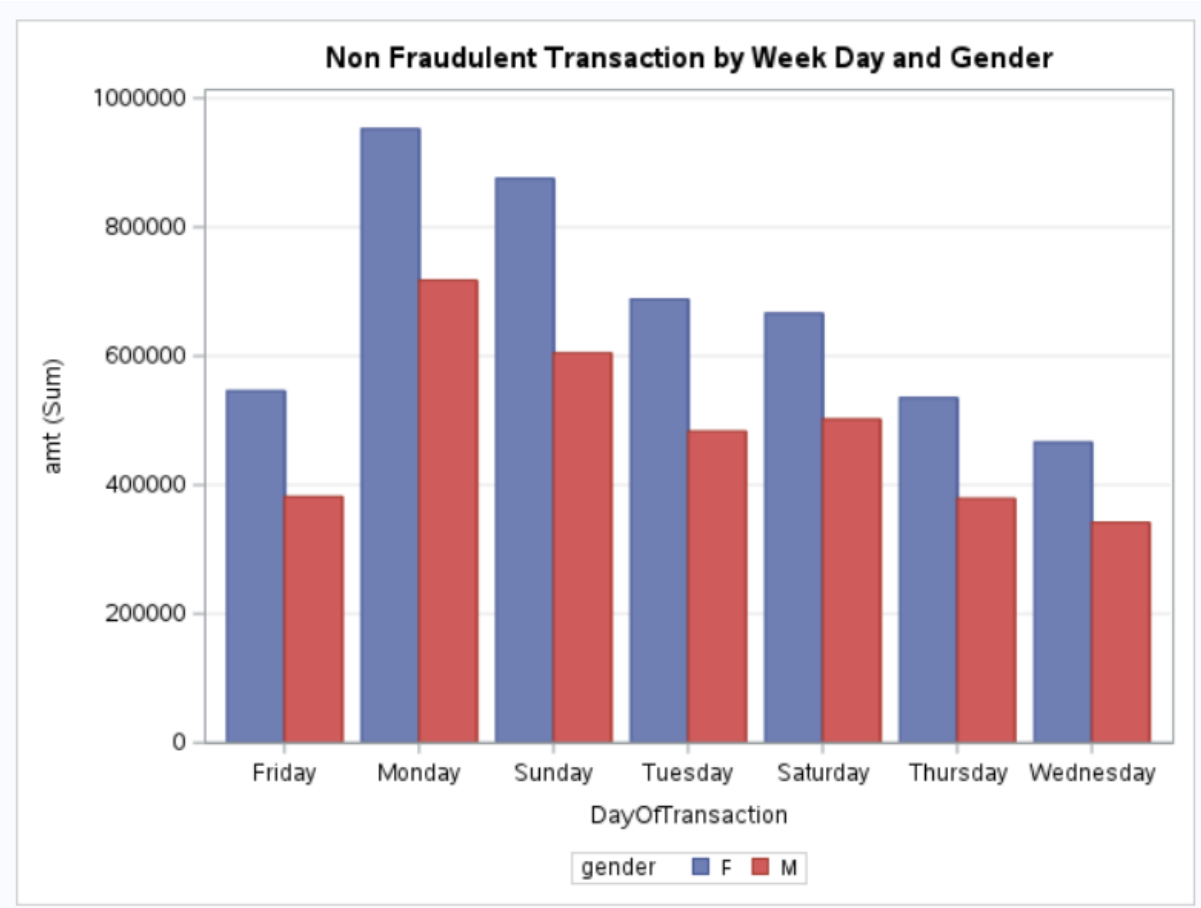


Fig 3.6 Frequency of legitimate transaction on week days

The above figure shows the frequency of legtimate transaction and the week day in which they occur the most. Monday and Sunday are the days where most transaction takes place and this is regardless of the gender making the transaction. This insight will enable stakeholders allocate resources to begining of the week and ensure there are no failures.

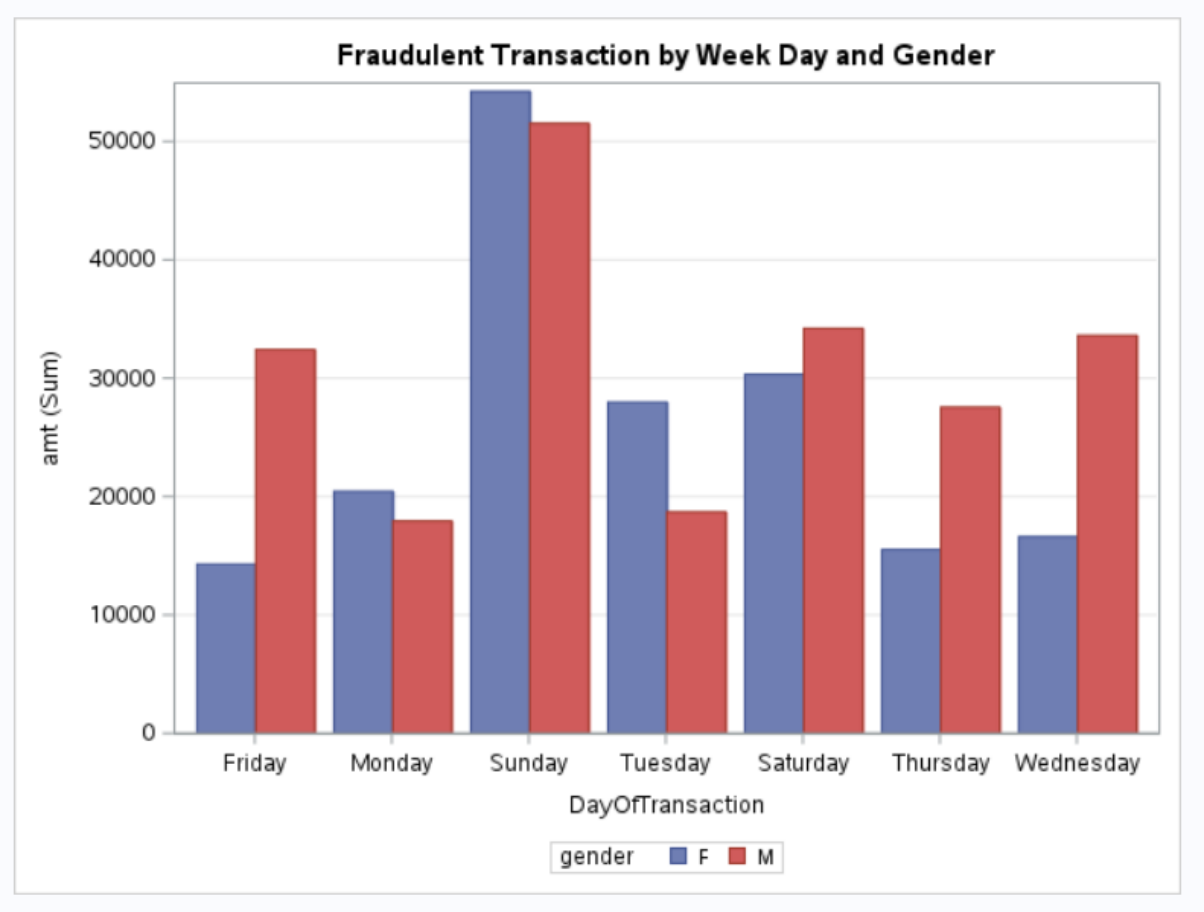


Fig 3.7 Frequency of legitimate transaction on week days

In contrast to fig 3.6 the above figure shows that fraudulent transactions for male customers exceeds that of female customers from Wednesday to Saturday. However, fraudulent transaction mostly take place from Sunday to Tuesday for female customer. The spike days for fraudulent transactions are mostly Sundays and affect female customers the most. This insight can be utilized to educate customers with what to look out for with regards to fraudulent activities on their credit card.

Transaction Category by Customer Gender

	Customer Transaction Category													
	entertainment	food_dining	gas_transport	grocery_net	grocery_pos	health_fitness	home	kids_pets	misc_net	misc_pos	personal_care	shopping_net	shopping_pos	travel
	Transaction Amount	Transaction Amount	Transaction Amount	Transaction Amount	Transaction Amount	Transaction Amount	Transaction Amount	Transaction Amount	Transaction Amount	Transaction Amount	Transaction Amount	Transaction Amount	Transaction Amount	Transaction Amount
	Sum	Sum	Sum	Sum	Sum	Sum	Sum	Sum	Sum	Sum	Sum	Sum	Sum	Sum
Customer Gender														
Female	$302,172	$250,838	$423,965	$144,053	$898,490	$224,253	$357,778	$361,854	$281,584	$288,161	$258,997	$475,415	$512,303	$129,621
Male	$235,216	$189,416	$334,140	$74,649	$511,815	$206,776	$303,332	$255,458	$188,559	$183,374	$156,224	$341,314	$344,653	$301,830

Fig 3.8 Transaction category by gender

Fig 3.8 gives us insights into customer spending habits. We can see from the table that female customer out spends the male customers in every category except Travel. The highest category where customers spend their money is in Grocery point of sale i.e. where they physically swipe their card at the counter. The least amount spent in a category was also in grocery, but the

payment method is online. It will make sense to market travel insurance to male customer and market coupons to female customers who physically shop for Grocery items.

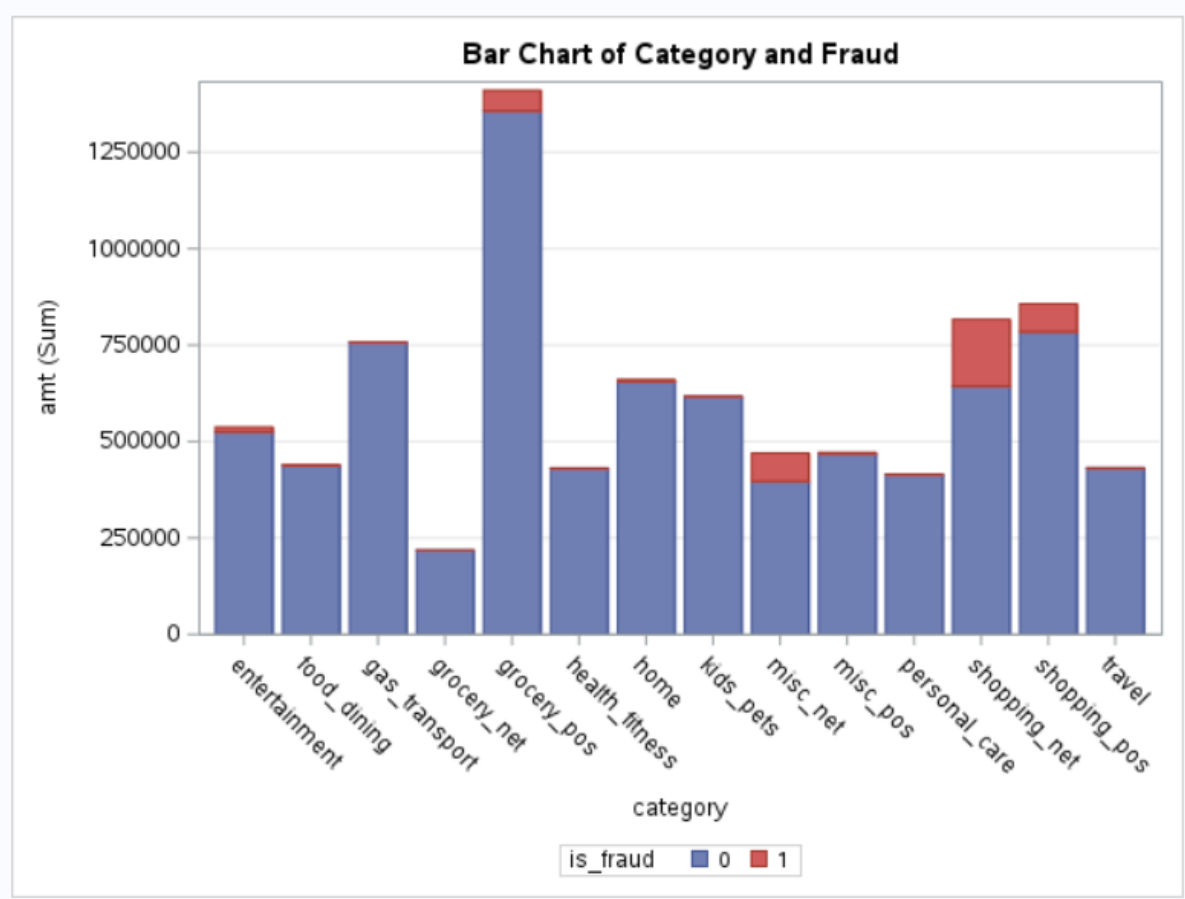


Fig 3.9 Category of fraudulent and non-fraudulent transaction.

Fig 3.9 is a bar graph that splits the total transaction amount into the different category which customers carried out transaction. It can be seen that the bulk of the transaction was done in the grocery category. While shopping online had the most fraudulent transactions followed by miscellaneous shopping online. This is a good insight as stake holders can flag easily a transaction to be fraudulent when it falls into these categories and a warning message alerting the customer can be sent.

Total Transaction by Category and Average Spend by Customer

	Amount per Transaction		
	Total Amount Spent	Average Amount Spent	Median Amount Spent
Customer Transaction Category			
entertainment	$537,388	$64	$48
food_dining	$440,254	$53	$45
gas_transport	$758,104	$63	$62
grocery_net	$218,702	$56	$52
grocery_pos	$1410305	$122	$109
health_fitness	$431,029	$54	$42
home	$661,110	$58	$48
kids_pets	$617,312	$58	$47
misc_net	$470,143	$81	$10
misc_pos	$471,535	$65	$12
personal_care	$415,221	$49	$33
shopping_net	$816,729	$90	$9
shopping_pos	$856,956	$79	$8
travel	$431,451	$112	$6

Fig 3.10 Average amount spend by customers.

To understand the spending habit of customers the table in fig 3.10 was created, it shows that personal care and online grocery shopping are the two least categories where customers spend their money. On average, the least amount spend by a customer is on personal care but the most amount spent on average falls within the travel and physical grocery category. When compared to the median spent by customers the story slightly changes, miscellaneous and shopping but in physical stores and online shopping categories were the least amount were customers spend their money.

3.2.2 Time Series Analysis

To understand the difference in characteristics of fraudulent transactions and non-fraudulent transactions, a time series analysis of both kinds of transactions was carried out. This will create a clear differentiation between what is considered to be fraudulent transactions and what is not.

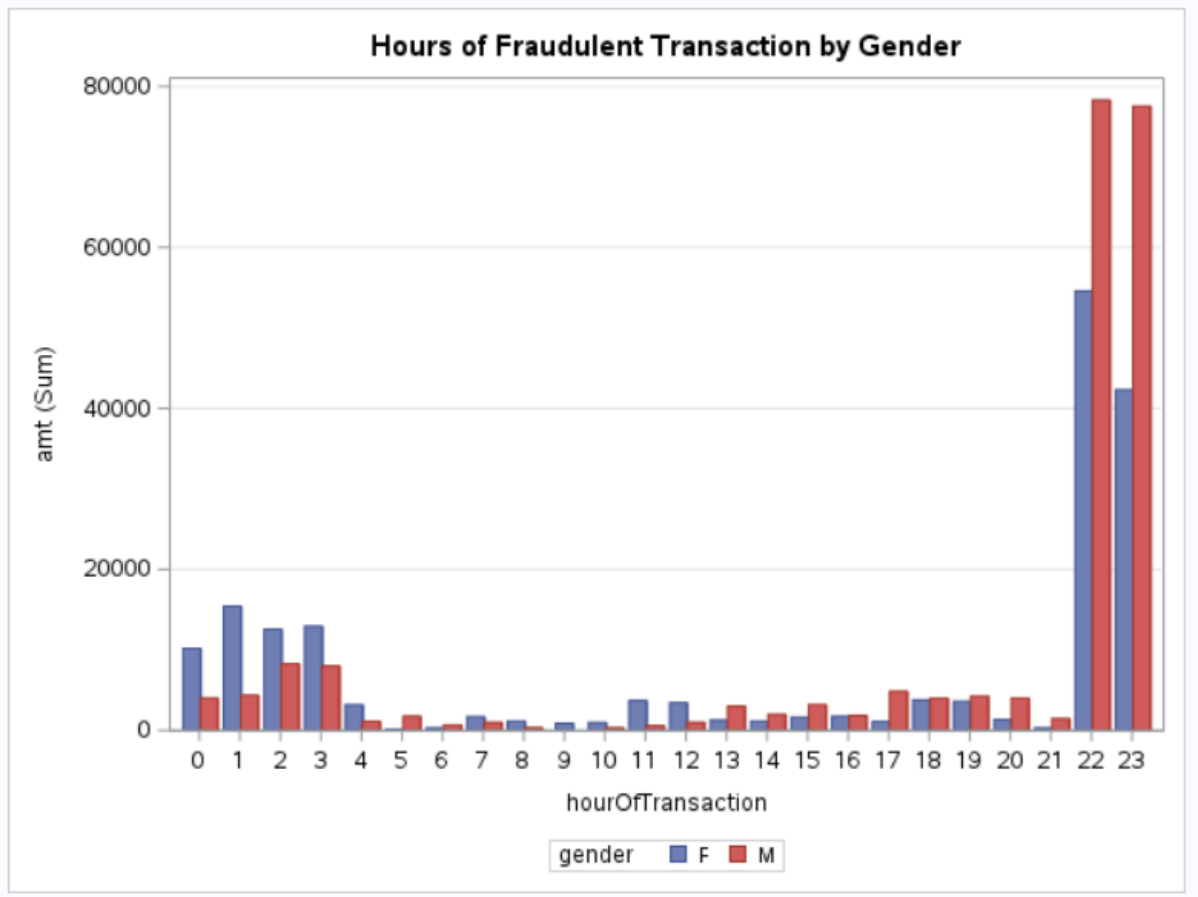


Fig 3.11 Hours of the day plotted against fraudulent transactions.

The figure above shows fraudulent transactions plotted by the hour of the day in which the transaction occurred. Instantly a theme can be seen, the scammers that carry out fraudulent transactions do so in the late hours of the day into the early hours of the day. This kind of information already gives a good red flag of when transactions can be halted if suspected of fraudulent activities.

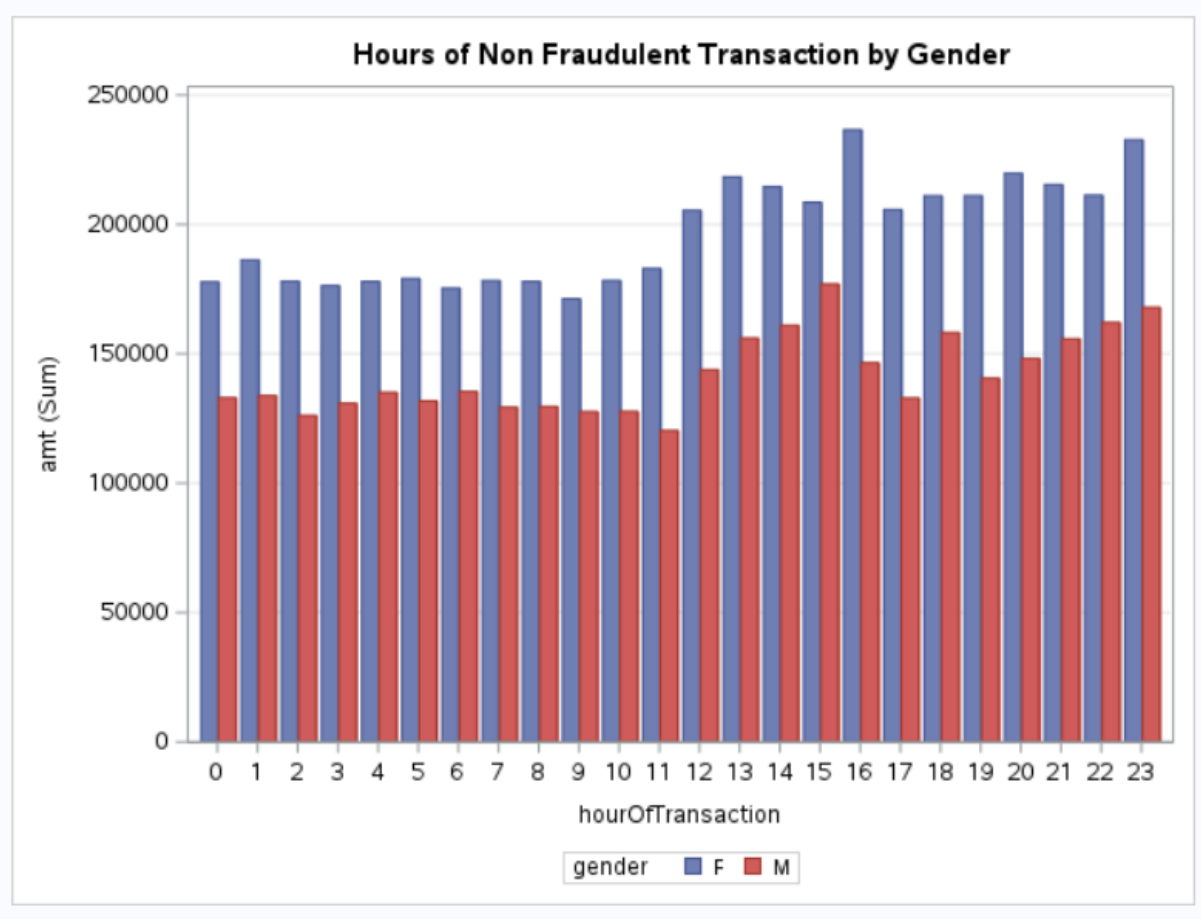


Fig 3.12 Hours of the day plotted against non-fraudulent transactions.

Fig 3.12 shows the hours in which customers carried out transactions by both male and female. There is an even spread of the data, with peak periods starting at mid-day and least transactions happening at midnight. When fig 3.12 is compared to fig 3.11, it is clear how fraudulent transactions can be profiled and as such detected using the hour of the day as a feature in a model that will predict fraudulent transactions.

3.2.3 Correlation Analysis

To effectively predict fraud, a correlation analysis was carried out using the amount spent by a customer and the likelihood of fraud to be committed. To achieve this, the amount being spent by customers were broken down into bins of eight. Each been representing a fraction of the total amount spent. The range of the bins are $20, $40, $100, $200, $500, $1,000, and above $1,000. The table shows that as the amount of money spent increases so does the likelihood of the transaction to be fraudulent.

Correllation between amount spent and likelihood of fraud

	Activity		Activity	
	Non Fraudulent	Fraudulent	Non Fraudulent	Fraudulent
	total_trans	total_trans	count_trans	count_trans
	Sum	Sum	Sum	Sum
amount_split				
$0 - $20	$298,639	$1,593	39871.00	123.00
$21 - $40	$459,650	$477	14781.00	21.00
$41 - $70	$1,402,771	$302	25069.00	6.00
$71 - $100	$1,480,545	$381	17696.00	4.00
$101 - $200	$2,172,526	$1,622	15923.00	13.00
$201 - $500	$1,180,495	$66,045	4204.00	207.00
$501 - $1000	$549,788	$226,118	843.00	264.00
>= $1000	$596,532	$98,757	302.00	92.00

Fig 3.13 Table showing increase in amount spent and the likelihood of fraud.

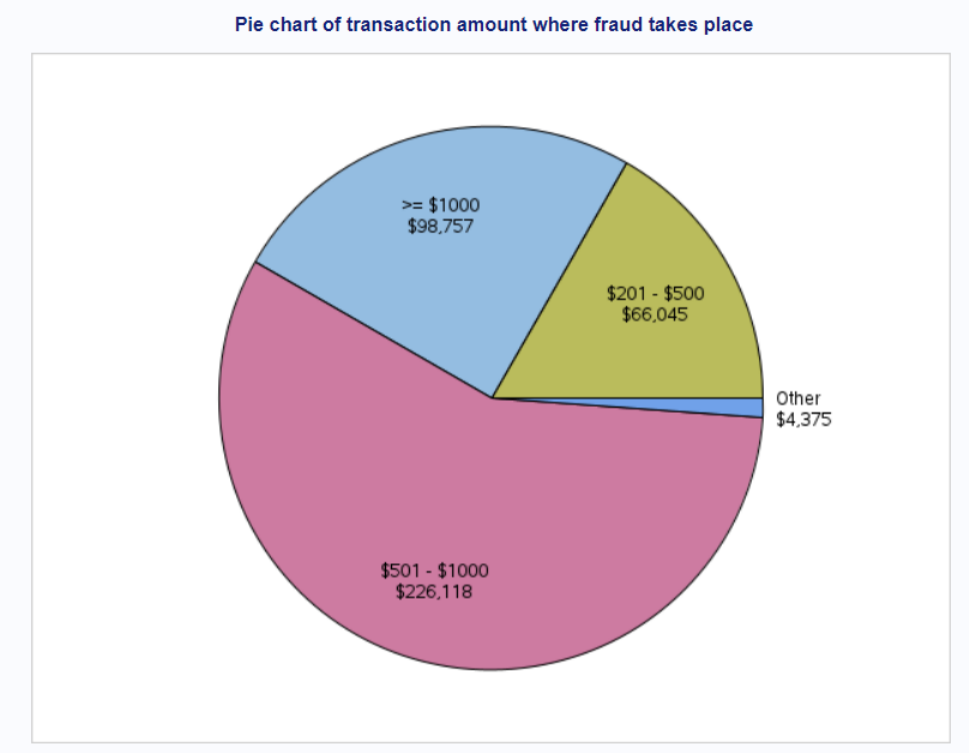


Fig 3.14 Pie chart showing the spend range and total fraudulent amount.

Our earlier analysis showed that the average amount spent on any transaction by a customer, either female or male was $71, however when you look at fig 3.14, which is a characteristic of fraudulent transactions, the observation is that fraudulent transactions tend to be three times or higher the average spend of a customer. This makes sense as scammers would want to cash in big in order to get away with as much more as they can.

3.2.3 Customer Age Analysis

At the beginning of our analysis, we did a distribution of age within the customer base using the credit card. This showed that older customers tend to get defrauded. The below table is a comprehensive analysis of age of customers and the likelihood of getting defrauded. Just like we did with the correlation analysis, we will group the customer base into age range. Teenagers, customers below the age of 20 years old, Young Adults, customers between the age of 20 – 35 years old, middle age between 35 and 60 years old, and then lastly Elderly, customers older than 60 years. This grouping will also be further grouped into gender, so we can have a clear view of how the age structure is between genders as well. The below figure shows in clear details this analysis.

Age Distribution and Faudulent Activity

		Activity		Activity	
		Fraudulent	Non Fraudulent	Fraudulent	Non Fraudulent
		TotalTrans	TotalTrans	NumOfTrans	NumOfTrans
		Sum	Sum	Sum	Sum
Gender	age_split				
Male	Teenager	$8,335	$19,467	9.00	306.00
	Young Adult	$27,579	$617,128	49.00	10252.00
	Middle Aged Adult	$114,534	$1,902,342	187.00	25840.00
	Elderly	$65,476	$871,896	120.00	13824.00
Female	Young Adult	$50,745	$1,885,196	125.00	28404.00
	Middle Aged Adult	$56,509	$1,987,452	121.00	26232.00
	Elderly	$72,117	$857,465	119.00	13831.00

Fig 3.16 Age distribution, gender, and fraudulent activity

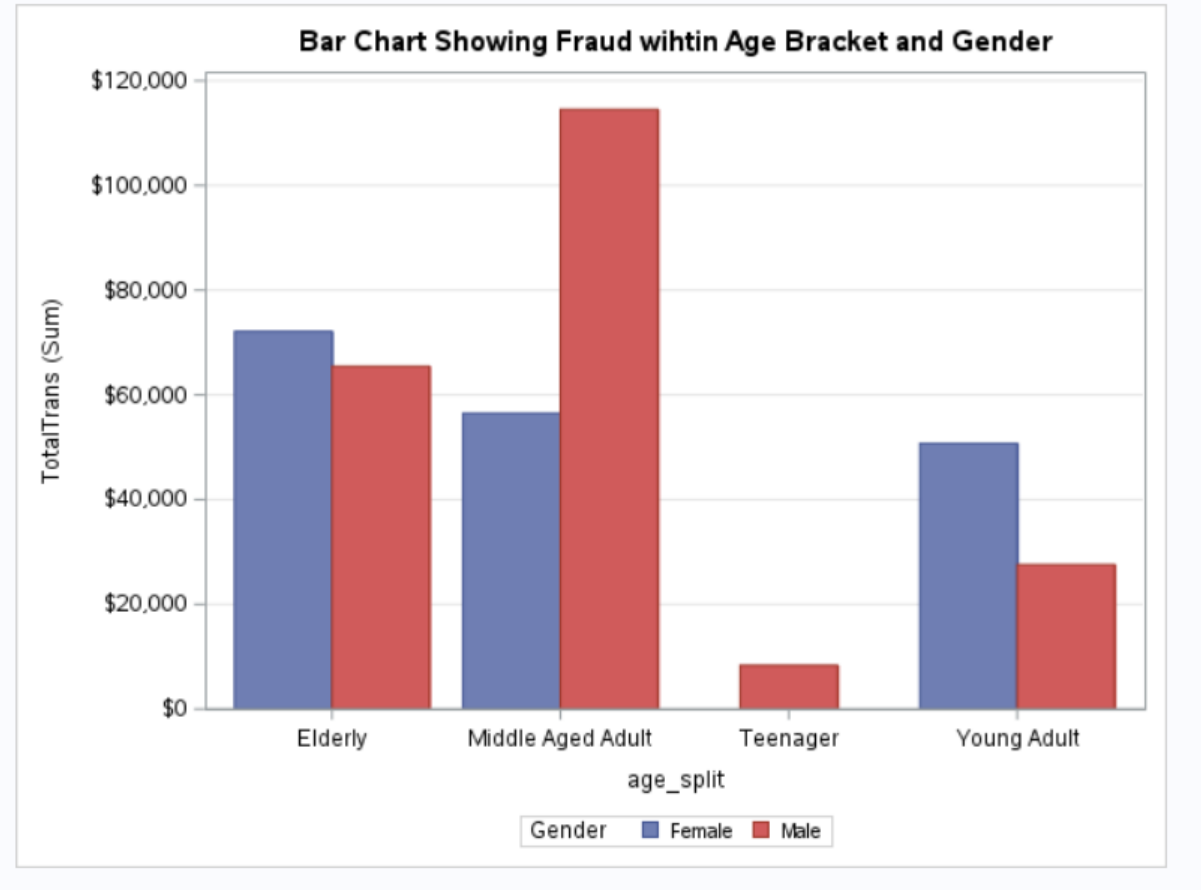


Fig 3.17 Bar chart of age bracket targeted by fraudsters

Imploring a bar graph to assist in buttressing the insight of our analysis, we can see that middled aged male customer are the most defrauded within our customer base followed by elderly women and young women. Teenage male customers also tend to fall victim as well.

4. Critical Comparison of SAS and Tableau

4.1 Introduction

In this study, both SAS and Tableau were utilized for data analysis and visualization tasks related to fraud detection. The study includes **a critical comparison of** these two **tools to** highlight their strengths, weaknesses, and suitability for different aspects of the fraud analytics process.

4.2 SAS Overview

SAS (Statistical Analysis System) is a powerful software suite developed by the SAS Institute for advanced analytics, business intelligence. and data management. It provides a comprehensive

set of tools for data manipulation, statistical analysis, predictive modelling, and reporting (SAS, 2023; Dietz, **2020).**

4.3 Tableau Overview

Tableau is a leading data visualization and business intelligence platform renowned for its user-friendly drag-and-drop interface and expansive visualisation capabilities. It allows users to create interactive dashboards, charts, and reports by connecting to various data sources (Tableau, 2024).

4.4 Critical Evaluation of SAS and Tableau

4.4.1 Data Pre-Processing and Manipulation

SAS offers a robust programming language and extensive data manipulation abilities, which makes it well-suited for compound data pre-processing tasks, such as data cleaning, transformation, and integration. Its SQL-like syntax and powerful procedures (e.g., PROC SQL, PROC SORT, PROC TRANSPOSE) provide a high degree of control and flexibility in handling large datasets (SAS, 2014).

Tableau, while primarily focused on data visualization, also offers data preparation capabilities through its built-in data interpreter and data blending features. However, for more advanced data manipulation tasks, it may require additional scripting or integration with other tools (DataCamp, 2024).

4.4.2 Exploratory Data Analysis and Visualization

Both SAS and Tableau excel in exploratory data analysis and visualization, albeit with different strengths. SAS offers a comprehensive variety of statistical procedures (e.g., PROC UNIVARIATE, PROC MEANS, PROC CORR) and powerful graphing capabilities through PROC SGPLOT and SAS/GRAPH (Smith, 2006). It allows users to generate various plots, charts, and statistical summaries for in-depth data exploration.

Tableau, on the other hand, is widely acknowledged for its intuitive and interactive data visualization abilities. Its drag-and-drop interface and extensive variety of visualization types (e.g., line charts, bar charts, scatter plots, maps, and dashboards) make it easy to create visually appealing and insightful representations of data (Taylor, 2019).

4.4.3 Model Building and Evaluation

SAS excels in model building and evaluation for fraud detection. It delivers a comprehensive collection of procedures and tools for building and evaluating various machine learning models, such as logistic regression (PROC LOGISTIC), decision trees (PROC ARBORETUM), and neural networks (PROC NEURAL). Additionally, SAS offers advanced techniques like ensemble modelling and model validation through procedures like PROC HPSPLIT and PROC HPSAMPLE.

While Tableau does not have built-in model-building capabilities, it can integrate with other tools like R or Python for advanced analytics and machine learning tasks. However, this integration may require additional setup and coding efforts.

4.4.4 Scalability and Performance

SAS and Tableau are both designed to handle large datasets, but they differ in their approaches and capabilities. SAS is known for its ability to process and analyze massive datasets efficiently, thanks to its high-performance computing capabilities and support for distributed computing environments. It can take advantage of parallel processing and in-memory computing to accelerate data-intensive operations, making it suitable for handling large-scale fraud detection workloads (SAS, 2024).

Tableau, while having the capabilities to handle large datasets, may experience performance bottlenecks when the datasets are extremely large or complex. However, it offers various optimization techniques, such as data extracts, live data connections, and data engine caching, to improve performance and responsiveness.

4.4.5 Ease of Use and User Experience

SAS has a higher initial learning difficulty, especially for users without prior programming experience. Its syntax-based programming language requires a certain level of proficiency to write and execute code effectively. However, once mastered, SAS offers a high degree of control and flexibility for advanced analytics tasks.

Tableau, on the other hand, is renowned for its user-friendly interface and drag-and-drop functionality. It provides a more intuitive and visually oriented experience, making it accessible to a wider range of users, including non-technical stakeholders. Tableau's ease of use facilitates rapid data exploration and visualization, enabling faster insights and decision-making (Tableau, 2024).

4.4.6 Integration and Extensibility

Both SAS and Tableau offer integration capabilities with other tools and platforms, enabling users to leverage their strengths in combination with other technologies.

SAS provides a wide range of integration options, including interfaces for programming languages like Python, R, and Java, as well as support for various data sources and formats. It also offers APIs and SDKs for custom application development and integration with other systems (SAS, 2024; Tableau, 2024).

Tableau also supports integration with a wide range of data sources, which includes databases, cloud platforms, and file formats. It offers APIs and web data connectors for integrating with external applications and services. Additionally, Tableau's extensions gallery allows users to enhance their analysis and visualization capabilities by installing third-party extensions.

4.4.7 Cost and Licensing

SAS and Tableau have different licensing models and cost structures, which can be a consideration for organizations based on their budget and requirements.

SAS typically follows a perpetual licensing model, with upfront costs for the software licenses and annual maintenance fees. It offers different pricing tiers based on the number of users, modules, and computing resources required. SAS can be a significant investment, particularly for small to medium-sized organizations.

Tableau, on the other hand, offers more flexible licensing options, including subscription-based and perpetual licenses. It provides different pricing plans based on the number of users and deployment options (on-premises or cloud-based). Tableau's pricing structure may be more suitable for organizations with varying budgets and scalability requirements (Tableau, 2024).

5 Conclusion

This study aimed to analyse a credit card fraud dataset to gain insights into patterns and characteristics that can aid in the detection and prevention of fraudulent activities. Through exploratory data analysis using SAS and Tableau, several key findings were uncovered.

The analysis revealed that certain factors, such as transaction amount, customer age, gender, and time of day, can be indicators of potential fraud. Fraudulent transactions were found to be more prevalent among older customers, suggesting a need for targeted education and awareness campaigns. Additionally, fraudulent activities were more likely to occur during late-night and early-morning hours, providing a temporal pattern that can be leveraged for fraud detection models.

The critical comparison between SAS and Tableau highlighted the strengths and weaknesses of each tool concerning different aspects of fraud analytics. While SAS excelled in data

manipulation, statistical analysis, and model building, Tableau demonstrated its prowess in interactive data visualization and user-friendly exploration.

The insights gained from this study can be valuable for financial institutions and other organizations in developing robust fraud prevention strategies, optimizing resource allocation, and enhancing customer education and support.

5.1 Future Work

Based on the findings and limitations of this study, several directions for future work can be explored:

1. Model Development and Deployment: Building upon the insights from the exploratory data analysis, machine learning models can be developed and evaluated for real-time fraud detection. These models can leverage the identified patterns and features to classify transactions as fraudulent or legitimate. Subsequently, these models can be integrated into existing fraud detection systems for deployment and continuous monitoring.
2. Real-time Fraud Monitoring: Implementing a real-time fraud monitoring system that incorporates the developed models and leverages streaming data capabilities. This system would analyse transactions as they occur, providing immediate alerts and enabling timely intervention to prevent fraud and minimize losses.
3. Ensemble Techniques and Advanced Algorithms: Exploring ensemble techniques and advanced machine learning algorithms, such as deep learning and neural networks, to improve the accuracy and robustness of fraud detection models. These techniques can capture complex patterns and relationships within the data, potentially enhancing the detection capabilities.
4. Incorporation of Additional Data Sources: Integrating additional data sources, such as customer behavioural data, social media data, and external data sources, into the fraud detection models. These additional data sources can provide valuable context and insights, further enhancing the accuracy of fraud detection.
5. Collaborative Fraud Detection: Investigating collaborative fraud detection approaches where financial institutions and organizations share anonymized transaction data and fraud patterns. This collaborative approach can help identify emerging fraud schemes and patterns that may be difficult to detect within a single organization's data.
6. Interpretability and Explainability: Focusing on developing interpretable and explainable models for fraud detection. This would enable stakeholders to understand the rationale behind the model's predictions, facilitating trust and acceptance, as well as enabling identification of potential biases or limitations.
7. User Experience and Visualization: Enhancing the user experience and visualization capabilities of fraud detection systems. This could involve developing intuitive dashboards and reporting tools that provide real-time insights and alerts to fraud analysts and decision-makers, enabling them to take timely and informed actions.

By pursuing these future directions, organizations can further strengthen their fraud detection and prevention capabilities, staying ahead of evolving fraud tactics and minimizing financial losses while enhancing customer trust and satisfaction.

6 References

1. Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. Decision Support Systems, 50(3), 602-613. https://psycnet.apa.org/record/2011-00451-008
2. Continuous monitoring and updating of the model as new data becomes available. https://neptune.ai/blog/retraining-model-during-deployment-continuous-training-continuous-testing
3. Exploratory Data Analysis:
4. Kaggle Dataset: "Credit Card Fraud Detection" (https://www.kaggle.com/mlg-ulb/creditcardfraud)
5. Model Evaluation:
    i. Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems, 50(3), 559-569. https://www.concur.co.uk/resource-centre/reports/hidden-cost-expense-fraud-and-non-compliance?gclid=CjwKCAjw9cCyBhBzEiwAJTUWNepkncfnT4pJ3E9uo96s3gOHyEGfACstiKwUNjGW_DByKLb6sSy2cBoCx8AQAvD_BwE&pid=ppc&cid=uk_goo_web_br_expense_fraud_prevention&ef_id=CjwKCAjw9cCyBhBzEiwAJTUWNepkncfnT4pJ3E9uo96s3gOHyEGfACstiKwUNjGW_DByKLb6sSy2cBoCx8AQAvD_BwE:G:s&s_kwcid=AL!5231!3!561849526939!b!!g!!expense%20fraud%20prevention!108082403!4637590883&gad_source=1
    b. Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. arXiv preprint arXiv:1009.6119. https://arxiv.org/abs/1009.6119
        i. Sharda, R., Delen, D., & Turban, E. (2018). Analytics, data science, & artificial intelligence: Systems for decision support. Pearson.
    c. Tukey, J. W. (1977). Exploratory data analysis. Addison-Wesley.
    d. Wickham, H., & Grolemund, G. (2017). R for data science: Import, tidy, transform, visualize, and model data. O'Reilly Media, Inc. https://link.springer.com/referenceworkentry/10.1007/978-0-387-32833-1_136
6. Applying machine learning algorithms (e.g., logistic regression, decision trees, random forests, neural networks) to build predictive models for fraud detection. https://www.researchgate.net/publication/378680090_Fraud_detection_using_supervised_learning_Algorithms
    a. Assessing the performance of the developed models using appropriate evaluation metrics (e.g., accuracy, precision, recall, F1-score, area under the ROC curve). https://blogs.ainomic.in/model-evaluation-and-metrics-34a6f7f678c6
7. Comparing the performance of different models to select the best-performing one. https://www.sciencedirect.com/science/article/pii/S2665917422000666
8. Comparison of SAS and Tableau:

9. DataCamp (2024) *Tableau Prep Builder: A Comprehensive Guide to Data Preparation*. Available at: https://www.datacamp.com/tutorial/tableau-prep-builder (Accessed: 20 May 2024).
10. Dietz, S. (2020) 'Statistical Data Analysis using SAS', *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(4), pp. 1829. doi: 10.1111/rssa.12606.
11. Feature Selection: Identifying the most relevant features (variables) that contribute to the prediction of fraudulent transactions. https://content.darwinium.com/fraud-prevention?utm_term=fraud%20detection&utm_campaign=UK%7CSearch%7CProduct%7CFraud&utm_source=google&utm_medium=cpc&hsa_acc=8600739592&hsa_cam=21262682817&hsa_grp=167315109812&hsa_ad=698932873136&hsa_src=g&hsa_tgt=kwd-11240143&hsa_kw=fraud%20detection&hsa_mt=b&hsa_net=adwords&hsa_ver=3&gad_source=1&gclid=CjwKCAjw9cCyBhBzEiwAJTUWNWCyGfKgWOF3DQN1F0GqdQU4K923eGnFFQfFgOVR6KK3jS489O2bmBoC8G4QAvD_BwE
    a. Fraud Detection and Prevention Strategies:
12. Key Factors and Variables: Abdallah, A., Maarof, M. A., & Zainal, A. (2016). Fraud detection system: A survey. Journal of Network and Computer Applications, 68, 90-113. https://translateyar.ir/wp-content/uploads/2019/09/Fraud-detection-survey-1.pdf
13. Model Building: Splitting the data into training and testing sets. https://encord.com/blog/train-val-test-split/
14. Model Deployment: Implementing the selected model in a production environment for real-time fraud detection and prevention. https://www.linkedin.com/pulse/deploying-machine-learning-models-production-analysis-omer-ali-phd
15. SAS Institute Inc. (2014) DS2 Language Reference for SAS 9.4. [PDF] Available at: SAS DS2 https://documentation.sas.com/api/docsets/ds2pg/9.4/content/ds2pg.pdf?locale=en (Accessed: 20 May 2024).
16. SAS Institute Inc. (2024) *SAS Documentation*. Available at: https://support.sas.com/en/documentation.html (Accessed: 20 May 2024).
17. Tableau Software (2024) *Performance Tips*. Available at: https://www.scribbr.co.uk/referencing/harvard-website-reference/ (Accessed: 21 May 2024).
18. Tableau Software (2024) *Pricing for Teams & Organizations*. Available at: https://www.tableau.com/pricing/teams-orgs (Accessed: 20 May 2024).
19. Tableau Software (2024) *Reference Materials*. Available at: https://www.scribbr.co.uk/referencing/harvard-website-reference/ (Accessed: 20 May 2024).
20. Taylor, D. (2019) *What is Tableau? Uses and Applications*. [online] Available at: Guru99.com (Accessed: 20 May 2024).
21. Techniques like correlation analysis, recursive feature elimination, or information gain can be employed for feature selection. https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/
22. Techniques like cross-validation can be used to evaluate and optimize the model's performance. https://deepchecks.com/evaluating-model-performance-using-validation-dataset-splits-and-cross-validation-techniques/
    i. Truscott, W. G. (2003). Data mining techniques for the detection of suspicious financial transactions. In Proceedings of the 2003 IEEE/IAFE Computational Intelligence for Financial Engineering

(CIFEr'03) (pp. 4-9). IEEE. https://technologyadvice.com/business-intelligence/resources/tableau-vs-sas/

23. □

7 Appendix 1 – Coding SAS Code

7.1 Data Pre-processing

```
/*************** DATA PREPROCESSING STEPS ***********************/


/******************** Import Data into Library ***********************/
PROC IMPORT
    DATAFILE= "/home/u63776952/coursework/fraudTotal.csv"
    OUT= mylib.fraudFull
    DBMS=csv
    REPLACE;
    GETNAMES=YES;

RUN;

/******************** List Table attribute of data set ***********************/
ods noproctitle;
ods select attributes variables;

proc datasets;
    contents data=MYLIB.FRAUDFULL order=collate;
quit;



/******************** Show transaction by state ***********************/

/*select top 10 transaction by state in order to slect a case study */

proc sql outobs=10;
    create table Tfraud_temp1 as
    select
            state as StateOfTransaction,
            case when is_fraud=1 then "Fraudulent Transaction"
            else "Non Fraudulent Transaction" end as Transaction_Type,
            count(amt) as NumberOfTransaction,
            sum(amt) FORMAT=dollar10. as TotalTransactionAmount

    from mylib.fraudfull
    where is_fraud=1
    group by StateOfTransaction,Transaction_Type
    order by TotalTransactionAmount desc;
    quit;
run;


proc sql outobs=10;
    create table Tfraud_temp0 as
    select
            state as StateOfTransaction,
            case when is_fraud=1 then "Fraudulent Transaction"
            else "Non Fraudulent Transaction" end as Transaction_Type,
            count(amt) as NumberOfTransaction,
            sum(amt) FORMAT=dollar10. as TotalTransactionAmount

    from mylib.fraudfull
    where is_fraud=0
    group by StateOfTransaction,Transaction_Type
    order by TotalTransactionAmount desc;
    quit;
run;


/* format the total amount and number of fraudulent transaction to highlight pain point */
proc format;
    value bcnfr low - 4000000 = "red"
            4000001 - 8000000  = "yellow"
            8000001 - high = "green";
run;

proc format;
    value bcfr low - 400 = "green"
            401 - 700  = "yellow"
            701 - high = "red";
run;


/* Table to display non fraudulent and fraudulent transaction by state */

title "Top 10 transaction by state";
proc tabulate data=work.Tfraud_temp0 ;
    class StateOfTransaction Transaction_Type;
    var TotalTransactionAmount NumberOfTransaction;
```

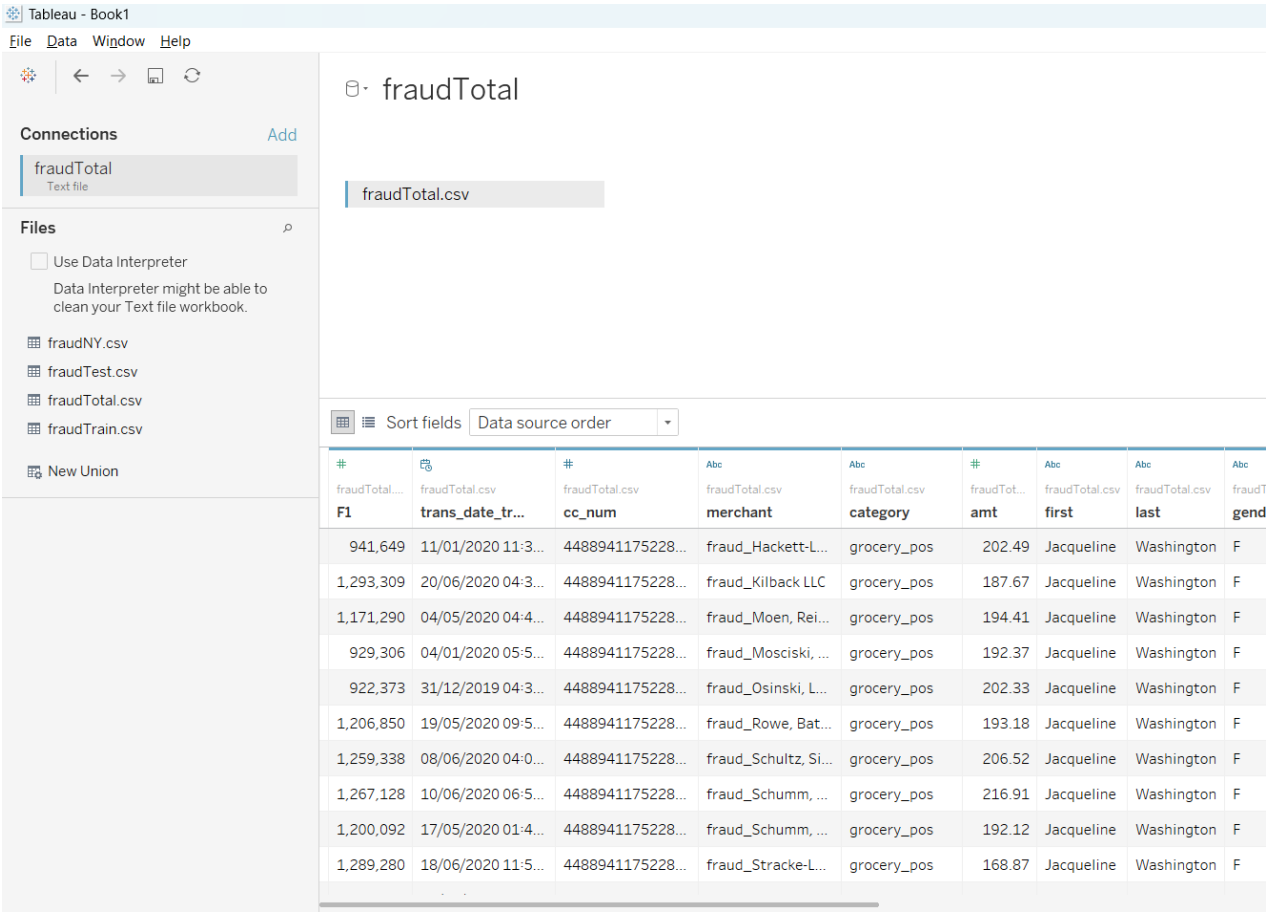
```
    table StateOfTransaction,Transaction_Type*TotalTransactionAmount*{style = {background = bcnfr.}}*(sum *format=dollar1
run;
title;

title "Top 10 fraudulent transaction by state";
proc tabulate data=work.Tfraud_temp1 ;
    class StateOfTransaction Transaction_Type;
    var TotalTransactionAmount NumberOfTransaction;
    table StateOfTransaction,Transaction_Type*TotalTransactionAmount*(sum *format=dollar10.) Transaction_Type*NumberOfTra
run;
title;


/******************** Select the state of New York to work with in determining Fraud **********************/
proc sql noprint;
    create table MYLIB.NYFRAUD as select * from MYLIB.FRAUDFULL where(state EQ
        'NY');
quit;


/* Add new calculated columns*/

Data MYLIB.NYFRAUD ;
 set MYLIB.NYFRAUD;
 age = intck('YEAR', dob, datepart(trans_date_trans_time));
 hourOfTransaction = hour(trans_date_trans_time) ;
 DayOfTransaction = put(datepart(trans_date_trans_time),dowName.) ;
 monthOfTransaction = put(datepart(trans_date_trans_time),monname.) ;
run;

/******************** END OF DATA PREPROCESSING STEPS **********************/




/******************** SUMMARY STATISTICS **********************/
title "summary statistics of customer gender";
proc means data=MYLIB.NYFRAUD chartype mean median sum std min max n q1 q3 vardef=df;
    var amt city_pop age hourOfTransaction;
    class gender;
run;
title;

/******************** Box plot of customer gender and age distribution **********************/

ods graphics / reset width=6.4in height=4.8in imagemap;
title "Age distribution of customers";
proc sgplot data=MYLIB.NYFRAUD;
    vbox age / category=gender group=is_fraud;
    yaxis grid;
run;
title;
ods graphics / reset;


/******************** END OF SUMMARY STATISTICS **********************/







/******************** GRAPHICAL EXPLORATORY DATA ANALYSIS **********************/

/******************** bar chart of day of transaction grouped by gender **********************/

ods graphics / reset width=6.4in height=4.8in imagemap;
title "Non Fraudulent Transaction by Week Day and Gender";
proc sgplot data=MYLIB.NYFRAUD;
    where is_fraud = 0;
    vbar DayOfTransaction / response=amt group=gender groupdisplay=cluster;
    yaxis grid;
run;
title;
ods graphics / reset;
```

```
ods graphics / reset width=6.4in height=4.8in imagemap;
title "Fraudulent Transaction by Week Day and Gender";
proc sgplot data=MYLIB.NYFRAUD;
    where is_fraud = 1;
    vbar DayOfTransaction / response=amt group=gender groupdisplay=cluster;
    yaxis grid;
run;
title;
ods graphics / reset;


/******************** END OF GRAPHICAL EXPLORATORY DATA ANALYSIS **********************/




/******************** TABULAR DATA ANALYSIS **********************/

/******************** Table showing the number of transactions carried out by gender **********************/
title "Total number of transaction and average amount spent by gender";
proc tabulate data=mylib.nyfraud ;
    label gender = "Customer Gender"
        amt = "Transaction Amount";
    keylabel N="Number of Transaction"
            mean="Average Amount Spent";
    class gender;
    var amt;
    table amt * (N mean*format=dollar8.)
            amt *(N mean*format=dollar8.) * gender;
Run;

title;

/******************** Table showing the number of transactions carried out customer category and formatted with color ****

proc format;
    value bckg low - 420000 = "red"
            420001 - 540000 = "yellow"
            540001 - 2000000 = "green";
run;


proc format;
    value bckgg low - 20 = "red"
            21 - 50 = "yellow"
            51 - 2000 = "green";
run;

title "Total Transaction by Category and Average Spend by Customer";

proc tabulate data=mylib.nyfraud format=dollar8.;
    keylabel sum="Total Amount Spent"
            mean="Average Amount Spent"
            median="Median Amount Spent";
    label category = "Customer Transaction Category"
            amt = "Amount per Transaction";
    class category;
    var amt ;
    table category,amt*(sum*{style = {background = bckg.}} (mean median) *{style = {background = bckgg.}});
run;

title;


/******************** Bar Chart of Category and Faud**********************/

ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgplot data=MYLIB.NYFRAUD;
    vbar category / response=amt group=is_fraud groupdisplay=stack;
    yaxis grid;
run;

ods graphics / reset;
```

```
/******************** Table showing the number of transactions carried out customer category and gender ******************
proc format;
    value $ genfmt 'M' = 'Male'
                   'F' = 'Female';
    value $ frfmt '1' = 'Fraudulent'
                  '0' = 'Non Fraudulent';
run;

title "Transaction Category by Customer Gender ";
proc tabulate data=mylib.nyfraud format=dollar10.;
    label gender = "Customer Gender"
        amt = "Transaction Amount"
        category = "Customer Transaction Category";
    format gender $genfmt.;
    class category gender;
    var amt;
    table gender ,category*amt*sum;
run;
title;



/******************** Table showing the number of transactions carried out customer category and gender ******************

title "Customer employment detail by Gender ";
proc tabulate data=mylib.nyfraud ;
    label gender = "Customer Gender"
        amt = "Transaction Amount"
        job = "Customer Job Description";
    keylabel N="NumberofTransaction"
             sum="AmountTransacted";
    format gender $genfmt.;
    class job gender;
    var amt;
    table job,gender*amt*(N sum*format=dollar10.);
run;
title;

/******************** END OF TABULAR DATA ANALYSIS **********************/

/* create transaction chunks to segment customer spend */
proc sql;
    create table fraud_temp as
    select
            case when amt< 21 then "$0 - $20"
                when amt < 41 then "$21 - $40"
                when amt < 71 then "$41 - $70"
                when amt <101 then "$71 - $100"
                when amt < 201 then "$101 - $200"
                when amt < 501 then "$201 - $500"
                when amt < 1001 then "$501 - $1000"
                else ">= $1000"
            end as amount_split,
            is_fraud,
            count(amt) as count_trans,
            sum(amt) FORMAT=dollar10. as total_trans

    from mylib.nyfraud
    group by amount_split,is_fraud
    order by total_trans;
    quit;
run;


proc format;

    value  frfmtt 1 = 'Fraudulent'
                  0 = 'Non Fraudulent';
    value $bcfrr  (notsorted)
            "$0 - $20" = "$0 - $20"
            "$21 - $40" = "$21 - $40"
            "$41 - $70" = "$41 - $70"
             "$71 - $100" = "$71 - $100"
             "$101 - $200" = "$101 - $200"
             "$201 - $500" = "$201 - $500"
             "$501 - $1000" = "$501 - $1000"
             ">= $1000" = ">= $1000";
    value ffamt low - 2000 = 'green'
                2000 - 250000 = 'red'
                250001 - high = 'white';
```

8 Appendix 2 Tableau Visualisation

8.1 Data Pre-Processing



fraudTotal.... F1	fraudTotal.csv trans_date_tr...	fraudTotal.csv cc_num	fraudTotal.csv merchant	fraudTotal.csv category	fraudTot... amt	fraudTotal.csv first	fraudTotal.csv last	fraudT gend
941,649	11/01/2020 11:3...	4488941175228...	fraud_Hackett-L...	grocery_pos	202.49	Jacqueline	Washington	F
1,293,309	20/06/2020 04:3...	4488941175228...	fraud_Kilback LLC	grocery_pos	187.67	Jacqueline	Washington	F
1,171,290	04/05/2020 04:4...	4488941175228...	fraud_Moen, Rei...	grocery_pos	194.41	Jacqueline	Washington	F
929,306	04/01/2020 05:5...	4488941175228...	fraud_Mosciski, ...	grocery_pos	192.37	Jacqueline	Washington	F
922,373	31/12/2019 04:3...	4488941175228...	fraud_Osinski, L...	grocery_pos	202.33	Jacqueline	Washington	F
1,206,850	19/05/2020 09:5...	4488941175228...	fraud_Rowe, Bat...	grocery_pos	193.18	Jacqueline	Washington	F
1,259,338	08/06/2020 04:0...	4488941175228...	fraud_Schultz, Si...	grocery_pos	206.52	Jacqueline	Washington	F
1,267,128	10/06/2020 06:5...	4488941175228...	fraud_Schumm, ...	grocery_pos	216.91	Jacqueline	Washington	F
1,200,092	17/05/2020 01:4...	4488941175228...	fraud_Schumm, ...	grocery_pos	192.12	Jacqueline	Washington	F
1,289,280	18/06/2020 11:5...	4488941175228...	fraud_Stracke-L...	grocery_pos	168.87	Jacqueline	Washington	F

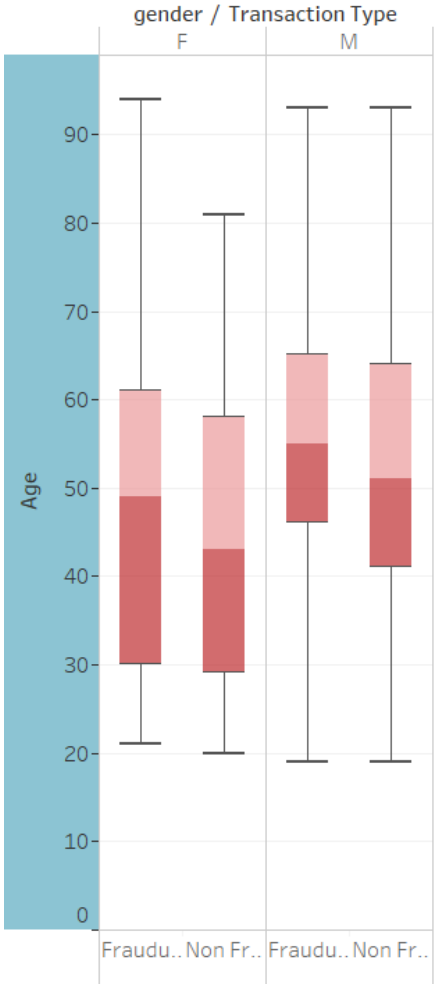
8.2 Data Mining

Age Distribution and Fraudulent Activity

		Transaction Type			
		Number of Records		amt	
Gender L..	Age Split	Fraudule..	Non Fra..	Fraudule..	Non Fra..
Female	Elderly	119	13,831	72,117	857,465
	Middle Aged Adult	121	26,232	56,509	1,987,452
	Young Adult	125	28,404	50,745	1,885,196
Male	Elderly	120	13,824	65,476	871,896
	Middle Aged Adult	187	25,840	114,534	1,902,342
	Teenager	9	306	8,335	19,467
	Young Adult	49	10,252	27,579	617,128

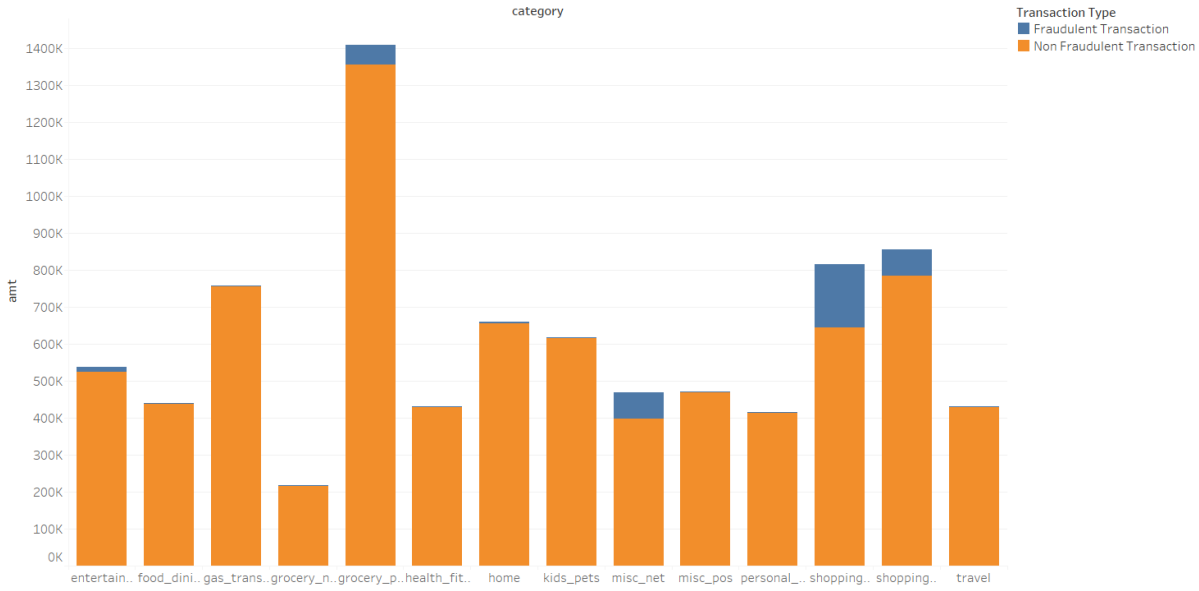
Number of Records and amt broken down by Transaction Type vs. Gender Long and Age Split. The data is filtered on state, which keeps NY.

Age Distribution of Customer



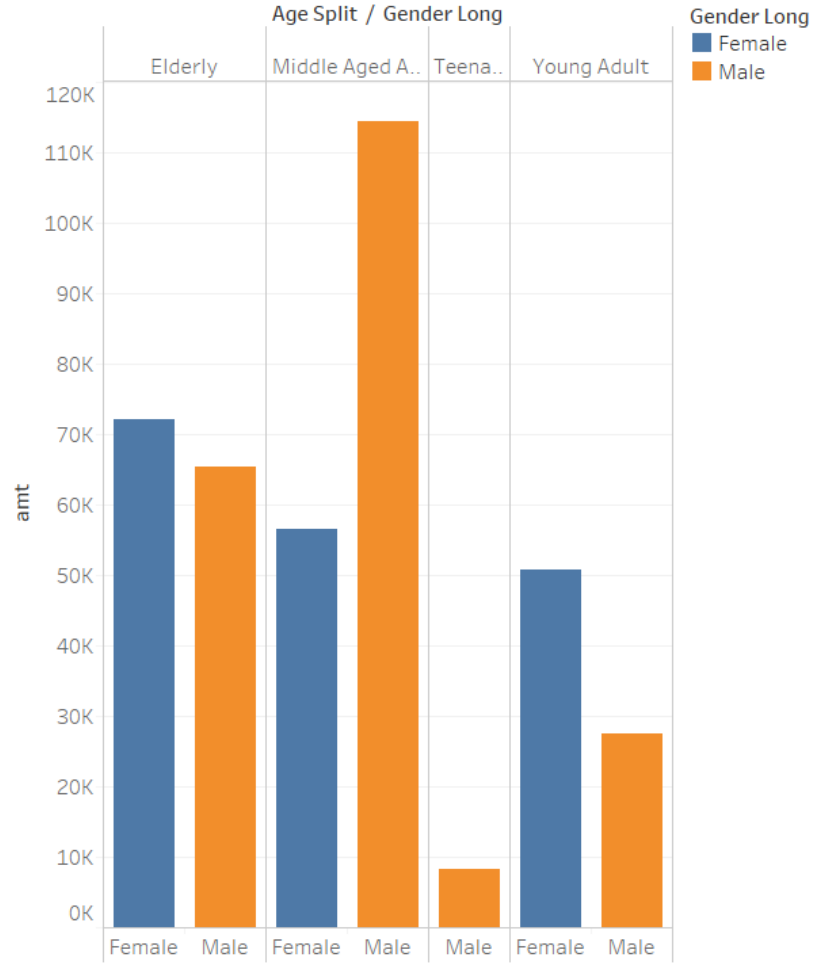
Age for each Transaction Type broken down by gender. The data is filtered on state, which keeps NY.

Bar Chart of Chategory and Fraud



Sum of amt for each category. Color shows details about Transaction Type. The data is filtered on state, which keeps NY.

Bar Chart Showing fraud within age bracket



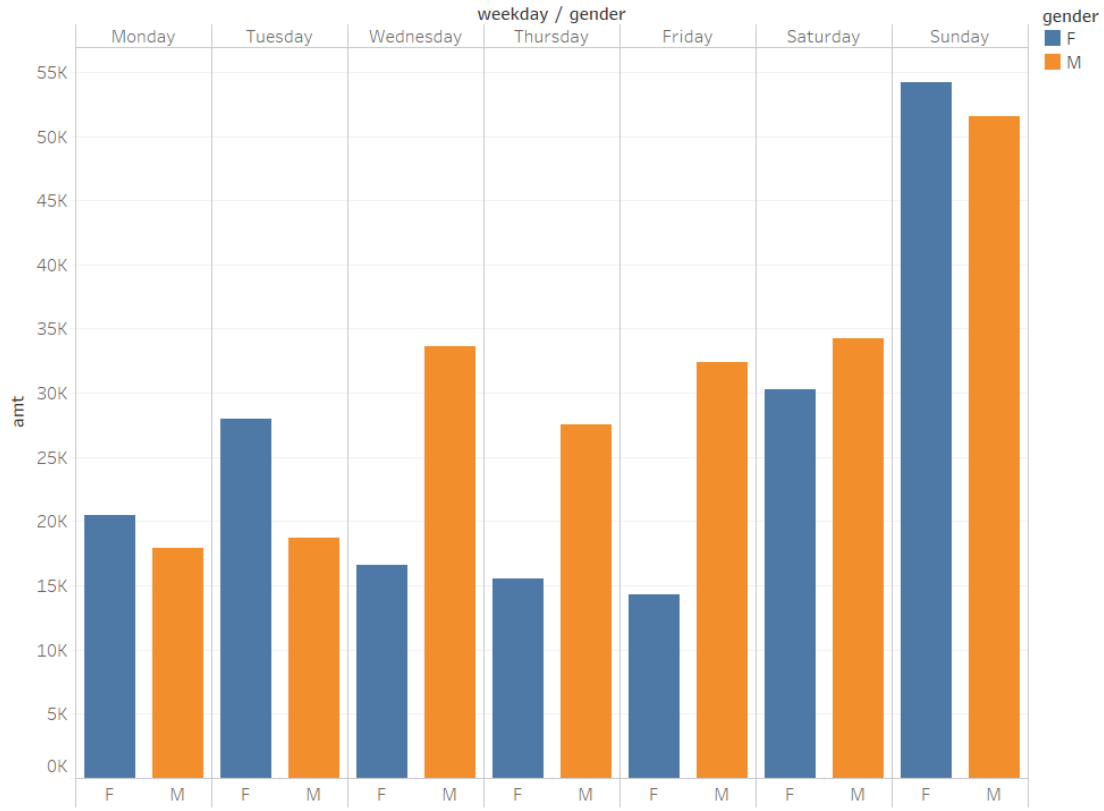
Sum of amt for each Gender Long broken down by Age Split. Color shows details about Gender Long. The data is filtered on state and Transaction Type. The state filter keeps NY. The Transaction Type filter keeps Fraudulent Transaction.

Correlation between amount soent and likelihood of fraud

	Transaction Type			
	Number of Records		amt	
Spend Split	Fraudule..	Non Fra..	Fraudule..	Non Fra..
>= $1000	92	302	98,757	596,532
$0 - $20	123	39,871	1,593	298,639
$21 - $40	21	14,781	477	459,650
$41 - $70	6	25,069	302	1,402,771
$71 - $100	4	17,696	381	1,480,545
$101 - $200	13	15,923	1,622	2,172,526
$201 - $500	207	4,204	66,045	1,180,495
$501 - $1000	264	843	226,118	549,788

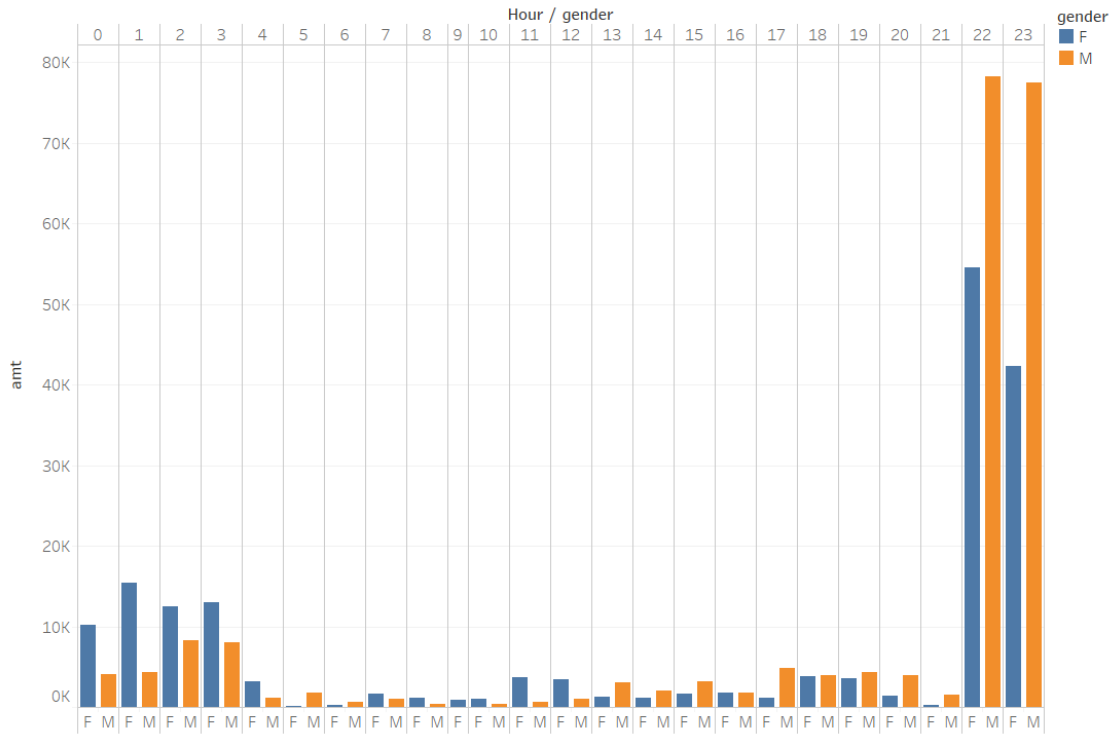
Number of Records and amt broken down by Transaction Type vs. Spend Split. The data is filtered on state, which keeps NY.

Fraudulent Transaction By Weekday



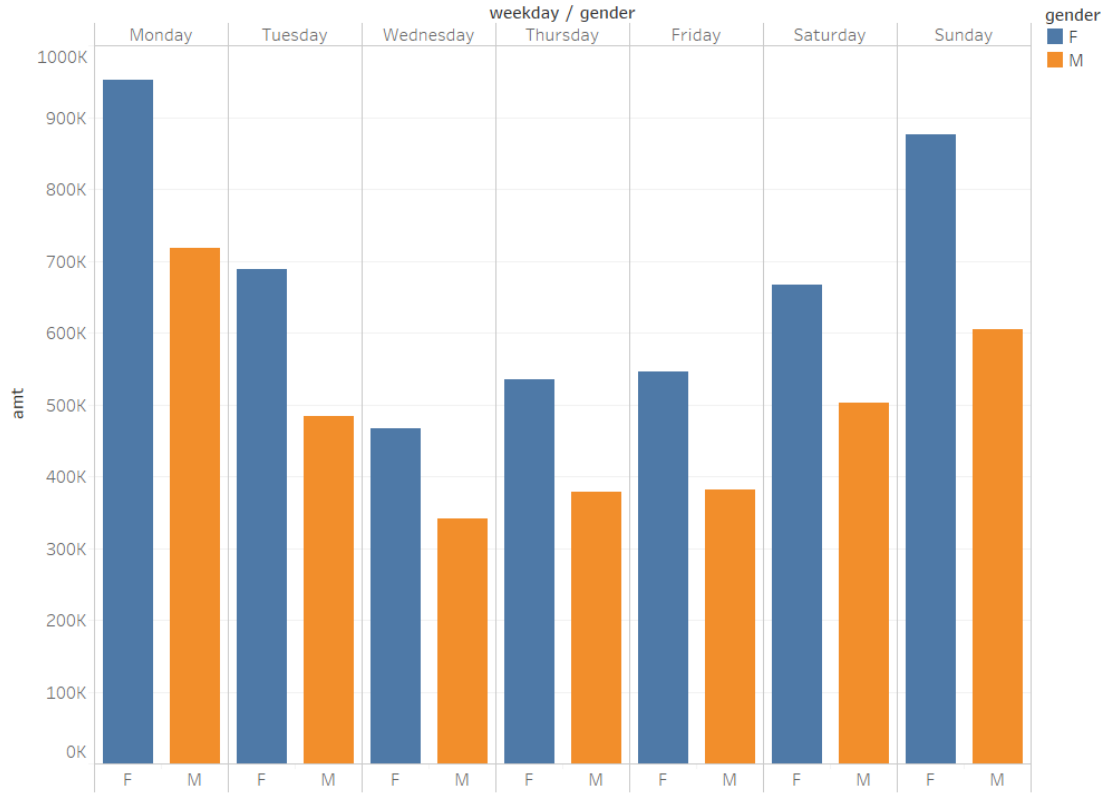
Sum of amt for each gender broken down by weekday. Color shows details about gender. The data is filtered on state and Transaction Type. The state filter keeps NY. The Transaction Type filter keeps Fraudulent Transaction.

Fraudulent Transaction



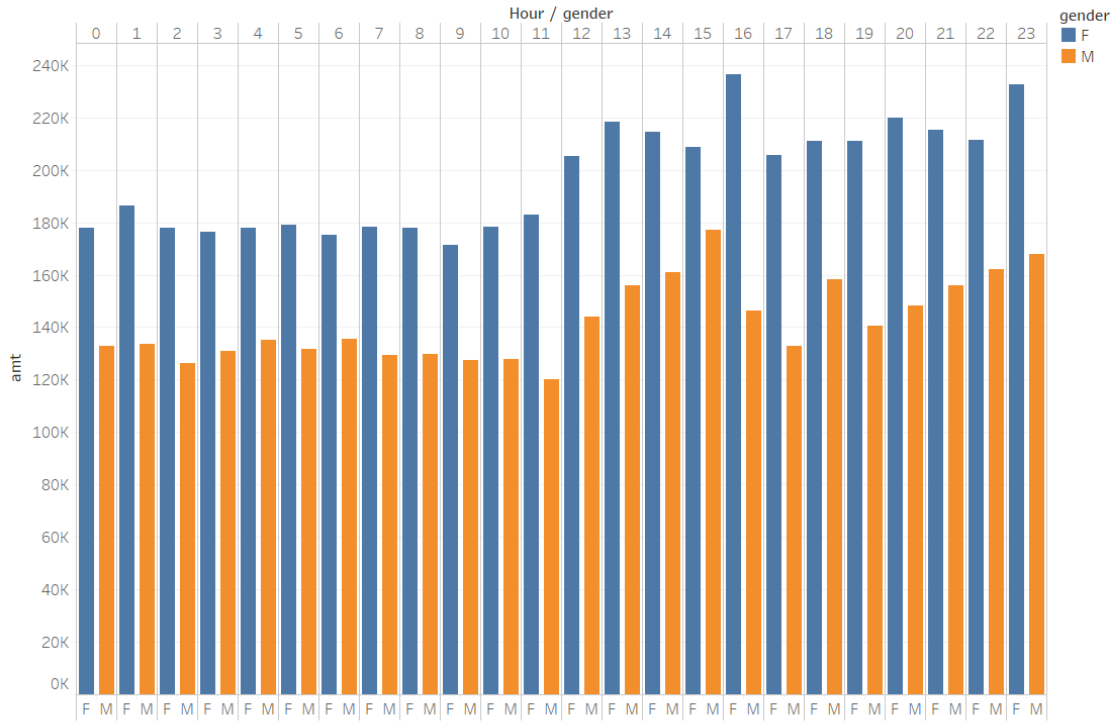
Sum of amt for each gender broken down by Hour. Color shows details about gender. The data is filtered on state and Transaction Type. The state filter keeps NY. The Transaction Type filter keeps Fraudulent Transaction.

Non Fraudulent Transaction By Weekday



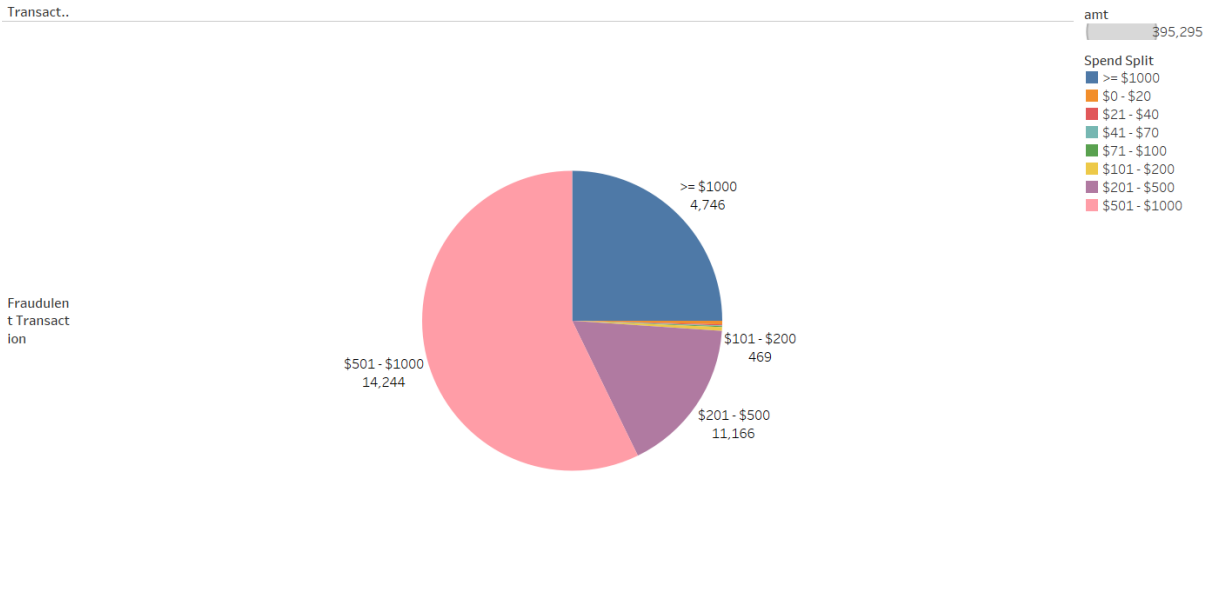
Sum of amt for each gender broken down by weekday. Color shows details about gender. The data is filtered on state and Transaction Type. The state filter keeps NY. The Transaction Type filter keeps Non Fraudulent Transaction.

Non Fraudulent Transaction



Sum of amt for each gender broken down by Hour. Color shows details about gender. The data is filtered on state and Transaction Type. The state filter keeps NY. The Transaction Type filter keeps Non Fraudulent Transaction.

Pie chart of transaction amount where fraud takes place



Spend Split and sum of Age broken down by Transaction Type. Color shows details about Spend Split. Size shows sum of amt. The marks are labeled by Spend Split and sum of Age. The data is filtered on state, which keeps NY. The view is filtered on Transaction Type, which keeps Fraudulent Transaction.

Top 10 transaction by state

	Transaction Type	
	Non Fraudulent Tran..	
state	amt	Number o..
AL	3,661,568	58,243
CA	5,700,931	80,093
FL	4,274,728	60,441
IL	4,157,768	61,888
MI	4,523,936	65,526
MO	3,606,432	54,642
NY	8,140,945	118,689
OH	4,631,536	66,267
PA	7,898,197	113,601
TX	9,239,298	134,677

1

KPI
- GREEN
- RED
- YELLOW

Amt and Number of Records broken down by Transaction Type vs. state. Color shows details about KPI. Size shows minimum of Number of Records. The marks are labeled by amt and Number of Records. The data is filtered on Index, which ranges from 1 to 10. The view is filtered on Transaction Type, which keeps Non Fraudulent Transaction.

Peer Evaluation Form for Group Work

Your Group ID: ____Group 6_____

Write the **Student ID** of each of your group members in a separate column. For each group member, indicate the extent to which you agree with the statement on the left, using a scale of 1-5:

1 = lowest score, 5 = highest score

1: Very poor: unacceptable performance
2: poor: less than acceptable performance
3: Average performance
4: Good performance
5: Excellent performance

Evaluation Criteria	**Group member 1: ID**	**Group member 2: ID**	**Group member 3: ID**	**Group member 4: ID**	**Comment**
	100682459	**100691035**	**100695495**	**100629371**	
	Grade: (1 - 5)	**Grade: (1 - 5)**	**Grade: (1 - 5)**	**Grade: (1 - 5)**	
•Did the individual contribute his/her fair share? •Contributes meaningfully to group discussions. •Contributes significantly to the success of the project. •Demonstrates a cooperative and supportive attitude. •Overall, how would you rank this person's contributions to the group?	**5**	**5**	**5**	**5**	
•Did the individual complete all work in a timely manner?	**5**	**5**	**5**	**5**	
•How would you rate the quality of individuals' work?	**5**	**5**	**5**	**5**	

•Did the individual maintain a positive, respectful attitude? •Attends group meetings regularly and arrives on time. •Would you want to work with this person again?	**5**	**5**	**5**	**5**	
TOTALS	**20**	**20**	**20**	**20**	
PEMark = (TOTALS / 20)	**20**	**20**	**20**	**20**	

PEMark: peer evaluation mark

** If a student **never** joined the group, put "*no participation*" in the corresponding column.

Overall mark for a student (CW) = CW mark * peer evaluation mark