School of Computer Science and Engineering
(Computer Science & Engineering- Artificial Intelligence & Data Engineering)
Faculty of Engineering & Technology
Jain Global Campus, Kanakapura Taluk - 562112
Ramanagara District, Karnataka, India

2023-2024
(IV Semester)

A Project Report on

"SYNTHETIC FINANCIAL DATA ANALYSIS"

Submitted in partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING-ARTIFICIAL INTELLIGENCE & DATA ENGINEERING

Submitted by

AKANSHA SHETTY, CHIMIRALA KOWSTUBHA, KAPAROTU VENKATA SURYA THARANI
22BTRAD002, 22BTRAD012, 22BTRAD018

Under the guidance of
Mr. Arnab Roy
Project Practice Head and Mentor
Futurense Technologies

FACULTY OF ENGINEERING AND TECHNOLOGY

Department of Computer Science and Engineering- Artificial Intelligence & Data Engineering

School of Computer Science & Engineering

Faculty of Engineering & Technology

Jain Global Campus, Kanakapura Taluk - 562112

Ramanagara District, Karnataka, India

CERTIFICATE

This is to certify that the project work titled **"SYNTHETIC FINANCIAL DATA ANALYSIS"** is carried out by **Akansha Shetty (22BTRAD002), Chimirala Kowstubha (22BTRAD012), Kaparotu Venkata Surya Tharani (22BTRAD018),** a bonafide student(s) of Bachelor of Technology at the School of Engineering & Technology, Faculty of Engineering & Technology, JAIN (Deemed-to-be University), Bangalore in partial fulfillment for the award of degree in Bachelor of Technology in Computer Science and Engineering- Artificial Intelligence & Data Engineering, during the year **2023-2024**.

Mr. Arnab Roy
Project Practice Head and Mentor

Date:

Dr. Aditya Pai H ,
Program Head,
Computer Science and Engineering- Artificial Intelligence & Data Engineering,
School of Computer Science & Engineering
Faculty of Engineering & Technology
JAIN (Deemed to-be University)
Date:

Dr. Geetha G
Director,
School of Computer Science & Engineering
Faculty of Engineering & Technology
JAIN (Deemed to-be University)
Date:

Name of the Examiner

Signature of Examiner

1.

2.

DECLARATION

We , **Akansha Shetty (22BTRAD002), Chimirala Kowstubha (22BTRAD012), Kaparotu Venkata Surya Tharani (22BTRAD018)** students of IVth semester B.Tech in **Computer Science and Engineering- Artificial Intelligence & Data Engineering**, at School of Engineering & Technology, Faculty of Engineering & Technology, **JAIN (Deemed to-be University)**, hereby declare that the internship work titled **"Synthetic Financial Data Analysis"** has been carried out by us and submitted in partial fulfilment for the award of degree in **Bachelor of Technology in Computer Science and Engineering- Artificial Intelligence & Data Enginnering** during the academic year **2023-2024**. Further, the matter presented in the work has not been submitted previously by anybody for the award of any degree or any diploma to any other University, to the best of our knowledge and faith.

Name1: Akansha Shetty Signature
USN : 22BTRAD002
Name2: Chimirala Kowstubha Signature
USN : 22BTRAD012
Name3: Kaparotu Venkata Surya Tharani Signature
USN : 22BTRAD018

Place : Bangalore
Date :

ACKNOWLEDGEMENT

It is a great pleasure for us to acknowledge the assistance and support of a large number of individuals who have been responsible for the successful completion of this project work.

First, we take this opportunity to express our sincere gratitude to Faculty of Engineering & Technology, JAIN (Deemed to-be University) for providing us with a great opportunity to pursue my Bachelors Degree in this institution.

We are deeply thankful to several individuals whose invaluable contributions have made this project a reality. We wish to extend our heartfelt gratitude to ***Dr. Chandraj Roy Chand, Chancellor****, for his tireless commitment to fostering excellence in teaching and research at Jain (Deemed-to-be-University). We are also profoundly grateful to the honorable* ***Vice Chancellor, Dr. Raj Singh, and Dr. Dinesh Nilkant, Pro Vice Chancellor****, for their unwavering support. Furthermore, we would like to express our sincere thanks to* ***Dr. Jitendra Kumar Mishra, Registrar****, whose guidance has imparted invaluable qualities and skills that will serve us well in our future endeavors.*

We extend our sincere gratitude to ***Dr. Hariprasad S A, Director*** *of the Faculty of Engineering & Technology,* ***and Dr. Geetha G, Director*** *of the School of Computer Science & Engineering within the Faculty of Engineering & Technology, for their constant encouragement and expert advice. Additionally, We would like to express our appreciation to* ***Dr. Krishnan Batri, Deputy Director (Course and Delivery), and Dr. V. Vivek, Deputy Director (Students & Industry Relations),*** *for their invaluable contributions and support throughout this project.*

It is a matter of immense pleasure to express our sincere thanks to ***Dr. Aditya Pai H,*** *program head,* ***Computer Science and Engineering- Artificial Intelligence & Data Engineering****, School of Computer Science & Engineering Faculty of Engineering & Technology for providing right academic guidance that made our task possible.*

We would like to thank our guide ***Mr. Arnab Roy Project Practice Head and Mentor****,* ***Futurense Technologies****, for sparing his valuable time to extend help in every step of our work, which paved the way for smooth progress and fruitful culmination of the project.*

We would like to thank our Project Coordinator ***Mr. Akash Das AVP and Project Manager, Futurense Technologies****, and all the staff members of Computer Science and Engineering for their support.*

We are also grateful to our family and friends who provided us with every requirement throughout the course.

We would like to thank one and all who directly or indirectly helped us in completing the work successfully.

Signature of Student(s)

ABSTRACT

This capstone project presents a comprehensive analysis of synthetic financial data conducted during an internship program. The dataset comprises transaction records capturing various attributes such as transaction amount, customer demographics, merchant details, and transaction timestamps. Leveraging this dataset, the analysis aims to uncover valuable insights into transaction patterns, customer behavior, and fraud detection within the financial ecosystem.

The analysis begins with an exploration of temporal trends in transaction volumes, identifying peak periods of activity and potential patterns in transaction behavior. Subsequently, customer segmentation analysis is performed to delineate distinct customer segments based on demographic factors, transaction frequency, and preferred purchase categories. This segmentation enables targeted marketing strategies and personalized recommendations, enhancing customer engagement and satisfaction.

Additionally, the analysis explores transaction data to identify potential anomalies or irregularities that may indicate fraudulent activities. By scrutinizing transaction patterns and anomalies, this analysis aims to enhance awareness of potential risks associated with fraudulent transactions within the financial ecosystem, providing valuable insights for risk management strategies.

The findings from this analysis provide actionable insights that can empower financial institutions to optimize operational efficiency, enhance security measures, and deliver personalized experiences to their customers. This project underscores the significance of data-driven decision-making in driving strategic initiatives and fostering innovation within the financial sector.

Keywords: Financial data analysis, Transaction patterns, Customer segmentation, Fraud detection

TABLE OF CONTENTS

LIST OF FIGURES

NOMENCLATURE USED

EDA	Exploratory Data Analysis

Chapter 1

1. INTRODUCTION

1.1. Background & Motivation

Financial datasets play a pivotal role in various domains, ranging from banking and investment to risk assessment and fraud detection. The analysis of financial data provides valuable insights for decision-making processes, risk mitigation, and the formulation of effective strategies.

In recent years, the advent of synthetic datasets has gained prominence due to their ability to replicate the statistical characteristics of real-world data while preserving privacy and confidentiality. Synthetic financial datasets, in particular, are crafted to mimic the complexities and patterns inherent in genuine financial data, enabling researchers and practitioners to perform analyses, develop models, and validate algorithms without compromising sensitive information.

The motivation behind utilizing synthetic financial datasets stems from several key factors:

1.1.1. Data Privacy and Confidentiality:

Financial data often contains sensitive information about individuals, businesses, and institutions. Maintaining data privacy is paramount to comply with regulations such as GDPR and safeguard against potential breaches. Synthetic datasets offer a viable solution by generating realistic data while ensuring anonymity and confidentiality.

1.1.2. Accessibility and Availability:

Real-world financial datasets may be limited in availability due to restrictions imposed by regulatory bodies or proprietary concerns. Synthetic datasets alleviate these constraints by providing readily accessible data for research, experimentation, and educational purposes.

1.1.3. Model Validation and Benchmarking:

Evaluating the performance of financial models and algorithms requires extensive testing on diverse datasets. Synthetic financial datasets serve as valuable tools for validating models, benchmarking algorithms, and assessing their robustness under various scenarios.

1.1.4. Risk-Free Experimentation:

Experimenting with financial data entails inherent risks, especially when dealing with sensitive information or market-sensitive data. Synthetic datasets offer a risk-free environment for experimentation, allowing researchers and practitioners to explore new methodologies, techniques, and hypotheses without exposure to potential liabilities.

1.1.5. Algorithm Development and Training:

Machine learning algorithms, such as those used in fraud detection, credit scoring, and portfolio management, heavily rely on the availability of high-quality training data. Synthetic financial datasets serve as reliable sources for training algorithms, enabling developers to enhance model accuracy, generalize performance, and adapt to dynamic market conditions.

By leveraging synthetic financial datasets, researchers, data scientists, and industry professionals can overcome challenges associated with data privacy, accessibility, and risk, thereby accelerating innovation, fostering collaboration, and driving advancements in financial analytics and decision support systems.

In this report, we delve into the analysis of a synthetic financial dataset, exploring its characteristics, patterns, and implications for decision-making. Through rigorous examination and interpretation of the data, we aim to extract actionable insights, uncover hidden trends, and demonstrate the utility of synthetic datasets in driving informed decision-making processes within the financial domain.

1.2. Objective

The project aims to analyze synthetic financial data, with a specific focus on transaction details such as amount, timestamp, and fraud indicators to gain valuable insights into fraud detection.

Exploratory Data Analysis (EDA): The primary methodology employed involves conducting an EDA on the dataset.

This process entails:

- **Characterization:** Understanding the fundamental characteristics of the financial data.
- **Informative Features Identification:** Analyzing the data to identify key characteristics and patterns that differentiate fraudulent transactions from legitimate ones. This will involve exploring features such as transaction amount, card type, location, purchase category, and potentially transaction description
- **Anomaly Detection:** Recognizing any irregularities or outliers that may indicate potential fraudulent activities.
- **Further Investigation:** Highlighting areas within the data that require deeper investigation or analysis.
- **Insight Generation:** Through the EDA process, the project aims to generate actionable insights essential for:
    - **Informed Decision-Making:** Providing stakeholders with valuable information for making strategic decisions.
    - **Risk Management:** Facilitating the identification and mitigation of financial risks associated with fraudulent activities.
    - **Business Optimization:** Identifying opportunities for process optimization or improvement within financial operations.

1.3. Delimitation of research

1.3.1. Data Quality and Generalizability

1.3.1.1. Synthetic vs. Real Data

Similar to the previous scenario, if the analysis relies solely on synthetic data, it might not perfectly reflect real-world patterns. Real anonymized transaction data (including transaction_amount, transaction_time, card_type, location, purchase_category, and customer_age) can improve generalizability.

1.3.1.2. Limited Scope

The data might not capture the full range of fraud scenarios. Including additional columns like merchant_id and transaction_description could provide richer context for fraud detection. For instance, analyzing purchase descriptions (e.g., "TV" vs "Gift Card") might reveal suspicious spending patterns.

1.3.2. Analysis Methods:

1.3.2.1. Limited Techniques:

Basic techniques like analyzing transaction_amount by customer_age or location are helpful, but leveraging machine learning on the entire dataset (including customer_id, merchant_id, card_type, etc.) could uncover hidden patterns and improve fraud detection accuracy.

1.3.2.2. Python Libraries:

The choice of libraries like pandas, scikit-learn, or TensorFlow can impact the types of insights extracted. Exploring a wider range of libraries might reveal more comprehensive information.

1.3.3. Interpretation and Actionability:

1.3.3.1. Overfitting:

If the model overfits the data, it might not perform well on unseen real-world transactions. Techniques like cross-validation can help mitigate this.

1.3.3.2. Actionable Insights:

The analysis should translate findings into clear recommendations. For example, identifying locations (e.g., City-30 with high gas station fraud) or customer age groups (e.g., young adults with high online shopping fraud) susceptible to fraud can help businesses focus their resources.

1.3.4. Additional Considerations:

1.3.4.1. Data Anomalies:

Investigate outliers or inconsistencies. For instance, a negative value in transaction_amount would require correction.

1.3.4.2. Missing Information:

The write-up should mention the Python libraries used, the complexity of the analysis, and how the findings were validated.

1.3.5. Mitigating Limitations:

1.3.5.1. Advanced Techniques:

Utilize machine learning algorithms to identify suspicious patterns in various data combinations (e.g., high transaction amounts for a specific customer_age and location).

1.3.5.2. Actionable Recommendations:

Translate findings into clear recommendations for businesses. This could involve suggesting stricter fraud checks for specific card_type or merchant_id combinations.

1.3.5.3. Data Validation:

Ensure proper data cleaning and address anomalies before drawing conclusions.

1.3.5.4. Document the Process:

Clearly document the Python libraries used, the steps involved in the analysis, and the validation methods employed.

1.3.6. Technical Skills Development:

1.3.6.1. Python Programming:

Hands-on experience with Python libraries like pandas, scikit-learn, or TensorFlow strengthens their programming skills. This is highly valuable in today's data-driven world.

1.3.6.2. Machine Learning Fundamentals:

The project exposes us to machine learning concepts like data analysis, model building, and evaluation. Understanding these fundamentals is crucial for various fields beyond finance.

1.3.6.3. Data Analysis Techniques:

We can learn data cleaning, manipulation, and visualization techniques using tools like pandas and Matplotlib. These skills are essential for any data-related career path.

1.3.7. Problem-Solving and Critical Thinking:

1.3.7.1. Identifying Patterns:

The analysis process requires identifying patterns and anomalies within the data. This develops their critical thinking and problem-solving abilities.

1.3.7.2. Formulating Hypotheses:

We will learn to formulate hypotheses about fraud based on data analysis and test them using their models. This strengthens our ability to approach complex problems systematically.

1.3.8. Understanding Financial Fraud:

1.3.8.1. Fraud Detection Methods:

The project provides insights into real-world fraud detection methods and the importance of data analysis in combating financial crime.

1.3.8.2. Impact of Fraud:

We have gained a deeper understanding of the financial and reputational risks associated with fraud for businesses and consumers.

1.4. Benefits of research

1.4.1. Improved Fraud Detection:

1.4.1.1. Identify Suspicious Patterns:

By analyzing vast amounts of data (including transaction_amount, transaction_time, card_type, location, purchase_category, and customer_age), machine learning algorithms can identify complex and hidden patterns that might be missed by traditional methods. This can lead to earlier detection of fraudulent activity and potential losses.

1.4.1.2. Focus Resources:

The analysis can pinpoint areas with higher fraud risk (**e.g., City-30 with high gas station fraud or young adults with high online shopping fraud**). This allows businesses to focus their resources on high-risk areas and implement stricter security measures for specific card_type or merchant_id combinations.

1.4.1.3. Adapt to Evolving Fraud Techniques:

Fraudsters constantly develop new methods. **Machine learning models** can be continuously trained on new data to stay ahead of evolving fraud tactics.

1.4.2. Enhanced Customer Experience:

1.4.2.1. Reduced False Positives:

By fine-tuning the analysis, the system can differentiate between legitimate and fraudulent transactions more accurately. This reduces the number of false positives where legitimate transactions are flagged for verification, improving customer experience.

1.4.2.2. Personalized Security Measures:

Based on factors like customer_age, location, and spending habits, the system can implement personalized security measures. This could involve requiring additional verification for high-risk transactions (**e.g., large purchases at a new merchant_id**) while streamlining the process for low-risk transactions.

1.4.3. Informed Business Decisions:

1.4.3.1. Customer Segmentation:

The analysis can reveal spending patterns by demographics (e.g., customer_age) and location. This allows businesses to segment their customer base and tailor marketing strategies and product offerings more effectively.

1.4.3.2. Optimize Risk Management:

By understanding fraud patterns, businesses can optimize their risk management strategies. This could involve setting spending limits based on customer_age and purchase_category or partnering with specific card_type providers known for strong security measures.

1.4.4. Additional Benefits:

1.4.4.1. Scalability:

Python analysis can handle large datasets efficiently, making it suitable for large financial institutions with vast amounts of transaction data.

1.4.4.2. Customization:

Python offers a wide range of libraries and tools, allowing researchers to customize the analysis to fit specific needs. This can include incorporating additional data sources like customer demographics or social media activity for a more holistic view.

Chapter 2

2. LITERATURE SURVEY

2.1. Literature Review

[1]. Li, H., Yang, W., & Li, W. (2020). A survey on deep learning-based fraud detection in banking and finance. Information Fusion, 63, 144-157.

[2]. Wang, S., & Wang, Q. (2021). Customer segmentation in financial services: A systematic literature review. Journal of Retailing and Consumer Services, 59, 102429.

[3]. Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, Inc.

[4]. Zhang, H., Wang, L., & Du, Y. (2019). Deep learning for credit card fraud detection in financial transaction networks. Computers & Security, 81, 101-113.

[5]. Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2011). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research, 218(1), 211-229.

2.2. Inferences Drawn from Literature Review

The inferences drawn from the literature review provide valuable insights into the existing body of knowledge on the research topic. Here are some key inferences:

2.2.1. Understanding Customer Behavior

Existing literature emphasizes the importance of understanding customer behavior in the financial domain, including spending patterns, preferences, and transaction habits.

Studies suggest that customer demographics, such as age, income level, and geographic location, significantly influence spending behavior and fraud susceptibility.

2.2.2. Emerging Trends in Fraud Detection

Recent research highlights the growing adoption of advanced technologies, such as machine learning and deep learning, for fraud detection in financial transactions.

There is a shift towards proactive fraud prevention strategies that leverage predictive analytics and anomaly detection algorithms to identify suspicious activities in real-time.

2.2.3. Implications for Current Research

The literature review underscores the need for a holistic approach to fraud detection, integrating both traditional rule-based methods and modern data-driven techniques.

It provides valuable insights into the limitations and challenges of existing fraud detection systems, guiding the development of more robust and adaptive solutions.

Chapter 3

3. PROBLEM FORMULATION AND PROPOSED WORK

3.1. Introduction

Financial institutions handle vast amounts of transaction data, making fraud detection a critical and ongoing challenge. Conventional approaches frequently depend on real-world data, which has limitations due to privacy issues and the unbalanced nature of fraudulent transactions (which are uncommon in comparison to legal ones).

This project leverages a synthetic financial dataset. The use of synthetic data has various advantages:

Privacy Protection: It allays worries about private client data being revealed. Controlled Environment: Enables more in-depth analysis of particular areas of fraud detection by allowing customisation of data generating parameters, including fraud scenarios.
Data Imbalance Control: Facilitates more thorough analysis by distributing the quantity of authentic and fraudulent transactions.

This report details the analysis of a synthetic financial dataset aimed at identifying informative features and evaluating techniques for accurate fraud detection. In order to identify patterns and trends linking fraudulent transactions to legitimate ones, we examine the data. The goal is to gain insights that can contribute to developing more effective fraud detection strategies for financial institutions.

3.2. Problem Statement

What insights can a synthetic financial dataset provide to identify informative features for fraud detection, and how can we use it to evaluate methods for achieving high accuracy?

Financial institutions encounter a formidable challenge in effectively detecting fraudulent transactions within vast quantities of transactional data. Conventional approaches heavily rely on real-world data, which presents inherent limitations such as privacy concerns and the disproportionate occurrence of fraudulent transactions relative to legitimate ones.

To address these limitations, this project utilizes a synthetic financial dataset. Synthetic data offers distinct advantages, including heightened privacy protection, a meticulously controlled environment for tailored analysis, and equitable representation of fraudulent and legitimate transactions.

The primary objective of this report is to conduct a comprehensive analysis of the synthetic financial dataset, with the specific aim of identifying discerning features and evaluating methodologies for precise fraud detection. By meticulously scrutinizing patterns and trends inherent within the dataset, the overarching goal is to derive actionable insights that can significantly augment the development of robust fraud detection strategies tailored for financial institutions.

3.3. System Architecture /Model

3.3.1. Data Preprocessing

3.3.1.1. Data Exploration

We'll conduct an in-depth exploration of the dataset to understand its characteristics, such as data types, missing values, and outliers. Descriptive statistics will be used to summarize key features.

3.3.1.2. Data Cleaning

Any identified missing values or outliers will be addressed using appropriate techniques. As our synthetic financial dataset contains no null values, outliers, or duplicates, we can proceed with analysis.

3.3.1.3. Feature Engineering

Depending on analysis needs, we'll employ feature engineering techniques to create new features from existing ones. For example, we'll create a column for customer age groups to segment customers according to age ranges.

3.3.2. Exploratory Data Analysis (EDA)

3.3.2.1. Visualization

Data visualization techniques like histograms, boxplots, and scatterplots will be used to uncover patterns, trends, and relationships between features. Special focus will be given to how fraudulent transactions differ from legitimate ones across various features.

3.3.2.2. Statistical Analysis

Basic statistical analysis will compare feature distributions of authentic and fraudulent transactions.

3.3.3. Analysis and Interpretation

We will conduct a thorough examination of the EDA findings, emphasizing pivotal features essential for discerning between fraudulent and legitimate transactions.

3.3.4. Tools and Techniques

We'll utilise a variety of Python libraries for this analysis, including:

- Pandas for creating dataframes
- Numpy for scientific computing
- Matplotlib and Seaborn for data visualization
- Plotly Express and Plotly Graph Objects for interactive visualizations
- Scikit-learn for machine learning algorithms such as KMeans clustering
- Networkx for network analysis

3.3.5. Importance

This detailed analysis is vital for improving understanding, identifying patterns, and uncovering potential areas of concern within the **synthetic financial data**. It lays the foundation for ensuring data integrity and reliability for subsequent analyses and decision-making processes.

3.4. Proposed Algorithms

Data Exploration through Visualization: The following techniques will be used to visually analyze the additional data points and identify potential patterns associated with fraudulent transactions:

3.4.1. Transaction Frequency (transaction_frequency):

Line plots will be used to analyze trends in transaction frequency over time for both fraudulent and legitimate transactions.
These plots can reveal spikes or dips in activity that may indicate suspicious behavior, such as a sudden burst of transactions followed by a period of inactivity.

3.4.2. Transaction Duration (transaction_duration):

Scatterplots will be used to explore the relationship between transaction duration and transaction amount.
If fraudulent transactions tend to have shorter durations or unusual patterns (e.g., very long durations for small purchases), this could be valuable for detecting anomalies.

3.4.3. Merchant Reputation (merchant_reputation):

Bar charts will compare the reputation scores of merchants involved in fraudulent and legitimate transactions.
Significant differences in reputation scores may signal the importance of incorporating merchant reputation into fraud detection algorithms.
Businesses with consistently low reputation scores could be flagged for further scrutiny.

3.4.4. Customer Behavior (customer_behavior):

Heatmaps will be used to visualize the distribution of customer behavior metrics such as transaction frequency or average transaction amount.
Identifying unusual behavior patterns, such as a sudden increase in average transaction amount or a significant change in purchase categories, can aid in detecting fraudulent activities.

3.4.5. Transaction Device (transaction_device)

Pie charts will compare the distribution of transaction devices used for fraudulent and legitimate transactions.
Variations in device usage patterns may indicate potential fraud vectors or security vulnerabilities.
For instance, a high concentration of fraudulent transactions originating from mobile devices in a region known for mobile device hacking could be a red flag.

3.4.6. Transaction Amount Distribution

Density plots will visualize the distribution of transaction amounts for both fraudulent and legitimate transactions.
This can reveal any skewness or abnormalities in the distribution that may signal fraudulent behavior, such as a higher concentration of transactions just below a specific threshold amount that might trigger additional verification steps.

3.4.7. Transaction Velocity (transaction_velocity)

Boxplots will be used to compare the velocity of transactions (e.g., transactions per minute, hour) between fraudulent and legitimate transactions.

Deviations from normal velocity patterns, such as a sudden surge in transactions within a short time frame, may indicate fraudulent activity.

3.4.8. Transaction Recurrence (transaction_recurrence)

Line plots will be used to analyze the recurrence of transactions over time, particularly for recurring payments or subscriptions.
Sudden changes or disruptions in recurrence patterns, such as a missed payment on a recurring bill historically paid on time, can be indicative of fraudulent behavior.

3.5. Proposed Work

By leveraging a comprehensive array of analytical techniques and methodologies, including Exploratory Data Analysis (EDA) and various plots, the proposed work aims to extract actionable insights, enhance fraud detection capabilities, and mitigate risks associated with fraudulent financial transactions within the synthetic financial dataset.

3.5.1. Enhanced Descriptive Analysis

Conduct comprehensive Exploratory Data Analysis (EDA) to succinctly summarise and vividly portray the intrinsic attributes of the synthetic financial dataset.

Utilise advanced visualisation techniques, including histograms, box plots, and scatter plots, to depict spending behaviours across various demographic segments, emphasising key categories such as Groceries and Retail as predominant expenditure areas.

Employ statistical methods such as mean, median, and standard deviation to identify trends in spending patterns, including the relationship between customer age groups and expenditure preferences.

3.5.2. Advanced Multivariate Analysis

Utilize multivariate statistical techniques such as Principal Component Analysis (PCA) and Factor Analysis to delve into numerous variables concurrently, unveiling intricate patterns and interconnections within the synthetic financial dataset.

Investigate complex relationships between customer demographics (age group), transaction characteristics, and spending behaviors using correlation matrices and heatmap visualizations to derive nuanced insights.

Utilize machine learning algorithms such as clustering and classification models to identify subtle correlations and dependencies, enabling a deeper understanding of customer behavior and transaction dynamics.

3.5.3. Precision Anomaly Detection

Deploy advanced anomaly detection algorithms such as Isolation Forest and Local Outlier Factor to precisely identify aberrant patterns or outliers indicative of potential fraudulent activities within the synthetic financial data.

Leverage anomaly detection techniques to scrutinize transaction amounts, frequencies, and locations, flagging deviations from expected patterns for further investigation using interactive plots and dashboards.

Utilize anomaly detection models to highlight irregularities in fraud rates across different card types and customer age groups within the synthetic financial dataset, facilitating targeted fraud prevention strategies.

3.5.4. Bivariate Analysis

Conduct thorough bivariate analysis to examine the relationship between pairs of variables within the synthetic financial data, uncovering correlations or patterns suggestive of fraudulent behavior.

Investigate the association between card types and spending behaviors across diverse age groups, discerning potential indicators of fraudulent transactions using scatter plots and regression analysis.

Analyze correlations between transaction amounts, customer demographics, and fraud incidence rates within the synthetic financial data to identify potential risk factors and mitigation strategies using correlation plots and pair plots.

3.5.5. Network Analysis

Employ network analysis techniques such as graph theory and network visualization to scrutinize the intricate network structure within the synthetic financial dataset, focusing on relationships and connections between entities such as customers, merchants, and transactions..2

Identify clusters of transactions or suspicious patterns indicative of coordinated fraudulent activities within the synthetic financial data, enabling proactive fraud detection measures using network plots and centrality measures.

Utilize graph-based algorithms such as community detection and anomaly detection to uncover hidden patterns and anomalies within the transaction network of the synthetic financial dataset, enhancing fraud detection capabilities.

3.5.6. Data Preprocessing

Conduct thorough data preprocessing to clean and prepare the synthetic financial transaction dataset, addressing issues such as missing values, categorical variable encoding, and feature scaling using preprocessing techniques such as imputation and normalization.

Implement robust data transformation techniques such as feature engineering and dimensionality reduction to ensure data quality, consistency, and compatibility with advanced analytical models and machine learning algorithms.

3.5.7. Univariate Analysis

Perform detailed univariate analysis to examine individual variables within the synthetic financial dataset, elucidating their distributions, characteristics, and potential outliers using descriptive statistics and distribution plots.

Explore the distribution of transaction amounts, fraud rates, and customer demographics within the synthetic financial data, gaining insights into key variables influencing fraudulent activities using histograms and density plots.

3.5.8. Temporal Analysis

Conduct comprehensive temporal analysis to identify trends and patterns in transaction data over time intervals, including hourly, daily, and monthly trends within the synthetic financial data using time series analysis and seasonal decomposition.

Analyze transaction volume and spending behavior by time of day, day of the week, and month within the synthetic financial data, discerning temporal variations and recurring patterns indicative of fraudulent activities using time series plots and heatmaps.

3.5.9. Collaboration

Foster collaboration with domain experts to interpret findings, derive actionable insights, and formulate effective fraud detection and prevention strategies.

Collaborate with stakeholders to disseminate findings and recommendations, fostering a holistic approach to combating fraudulent activities within the financial ecosystem.

CHAPTER 4

4. IMPLEMENTATION

4.1 Software Implementation

The analysis of the synthetic financial dataset is conducted using a combination of programming languages, libraries, and tools to facilitate data manipulation, visualization, statistical analysis, and machine learning modeling. The following software stack is utilized for the analysis:

4.1.1. Programming Language

Python, a versatile programming language renowned for its simplicity, readability, and extensive ecosystem of libraries, is chosen as the primary language for data analysis and modeling.

4.1.2. Libraries and Frameworks

4.1.2.1. Pandas:

Pandas is employed for data manipulation, exploration, and transformation tasks. It provides powerful data structures and functions for handling structured data, facilitating seamless data preprocessing and cleaning.

4.1.2.2. NumPy:

NumPy is utilized for numerical computing operations such as array manipulation, mathematical functions, and linear algebra operations. It serves as a foundation for many scientific computing tasks within Python.

4.1.2.3. Matplotlib and Seaborn:

Matplotlib and Seaborn are employed for data visualization, offering a wide range of plotting functions and customization options to create insightful visualizations of the dataset's characteristics, patterns, and relationships.

4.1.2.4. Scikit-learn:

Scikit-learn is utilized for machine learning modeling, encompassing various algorithms for classification, regression, clustering, and model evaluation. It provides user-friendly interfaces and robust implementations for building and evaluating predictive models.

4.1.2.5. TensorFlow or PyTorch (Optional):

For advanced machine learning tasks such as deep learning, TensorFlow or PyTorch can be incorporated to develop and train neural network models for fraud detection, anomaly detection, or predictive analytics.

4.1.3. Integrated Development Environment (IDE):

Jupyter Notebook or JupyterLab: Jupyter Notebook or JupyterLab is chosen as the development environment for its interactive computing capabilities, support for inline visualization, and documentation features. It facilitates an iterative and exploratory approach to data analysis, enabling seamless integration of code, visualizations, and narrative explanations within a single document.

4.1.4. Version Control:

Git: Git is utilized for version control, enabling collaborative development, tracking changes, and managing project iterations effectively. Git repositories are used to store code, documentation, and experimental notebooks, facilitating reproducibility and collaboration among team members.

4.1.5. Dependency Management:

pip: pip is employed for package management, allowing for the installation, upgrading, and removal of Python packages and dependencies. Virtual environments can be utilized to isolate project dependencies and ensure reproducibility across different environments.

The software implementation follows best practices for data analysis and machine learning, including modular code design, documentation, version control, and reproducibility. By leveraging the aforementioned software stack, we aim to conduct a comprehensive analysis of the synthetic financial dataset, develop predictive models, and derive actionable insights to inform decision-making processes within the financial domain.

4.2. Software Alogrithm

The analysis of the synthetic financial dataset involves the implementation of various algorithms aimed at extracting insights, detecting patterns, and making predictions. The following algorithms are utilized in the analysis:

4.2.1. Exploratory Data Analysis:

EDA is performed to gain a comprehensive understanding of the dataset's characteristics, distribution, and relationships among variables. Descriptive statistics, data visualization techniques, and summary statistics are employed to identify trends, anomalies, and potential areas of interest within the dataset.

4.2.2. Data Preprocessing:

Data preprocessing techniques are applied to clean, transform, and prepare the dataset for further analysis and modeling. This includes handling missing values, encoding categorical variables, scaling numerical features, and splitting the dataset into training and testing sets for model evaluation.

4.2.3. Fraud Detection Alogrithms:

Various fraud detection algorithms are employed to identify fraudulent transactions within the dataset. These algorithms include:

4.2.3.1. Supervised Learning Algorithms:

Classification algorithms such as Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machines (SVM) are trained on labeled data to distinguish between fraudulent and legitimate transactions based on features such as transaction amount, location, time, and card type.

4.2.3.2. Unsupervised Learning Algorithms:

Anomaly detection techniques such as Isolation Forest, Local Outlier Factor (LOF), and One-Class SVM are utilized to identify unusual or suspicious patterns in transactional behavior that deviate from the norm.

4.2.3.3. Hybrid Approaches:

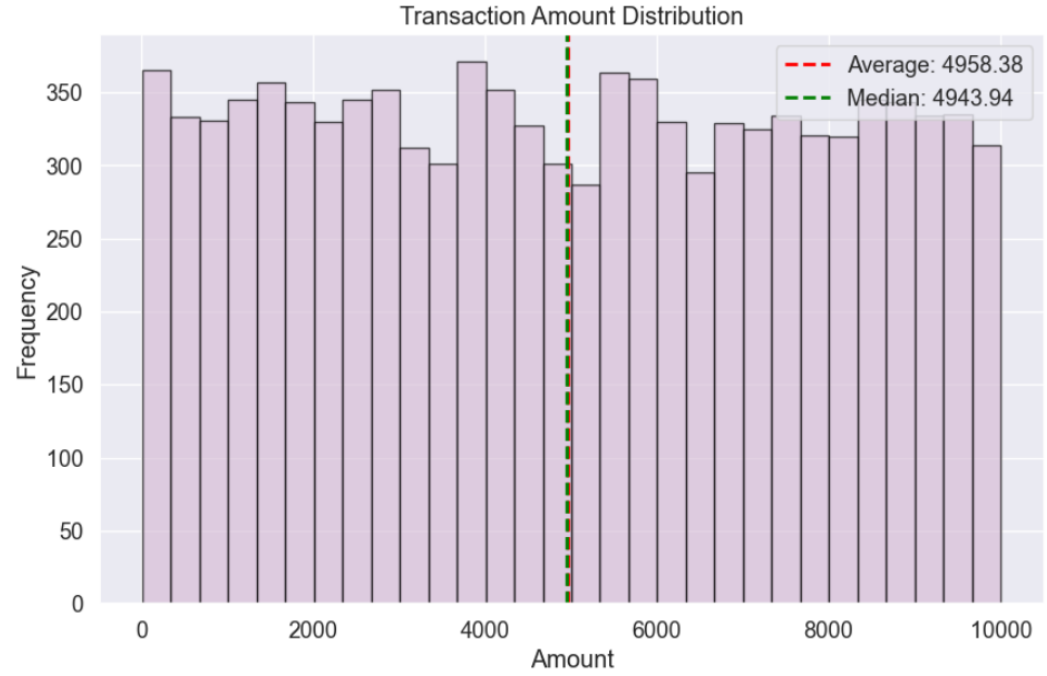
Hybrid approaches combining supervised and unsupervised techniques, ensemble methods, or deep learning architectures are explored to enhance fraud detection performance and adapt to evolving fraud patterns.

Chapter 5

5. RESULTS AND DISCUSSIONS

5.1. Distribution of features

5.1.1. Distribution of Amount




```
Average transaction amount: 4958.38
Median transaction amount: 4943.94
Maximum transaction amount: 9999.75
Minimum transaction amount: 10.61
```

Fig. 5.1.1 Transaction Amount Distribution

5.1.2. Distribution of Fraudulent Transactions:

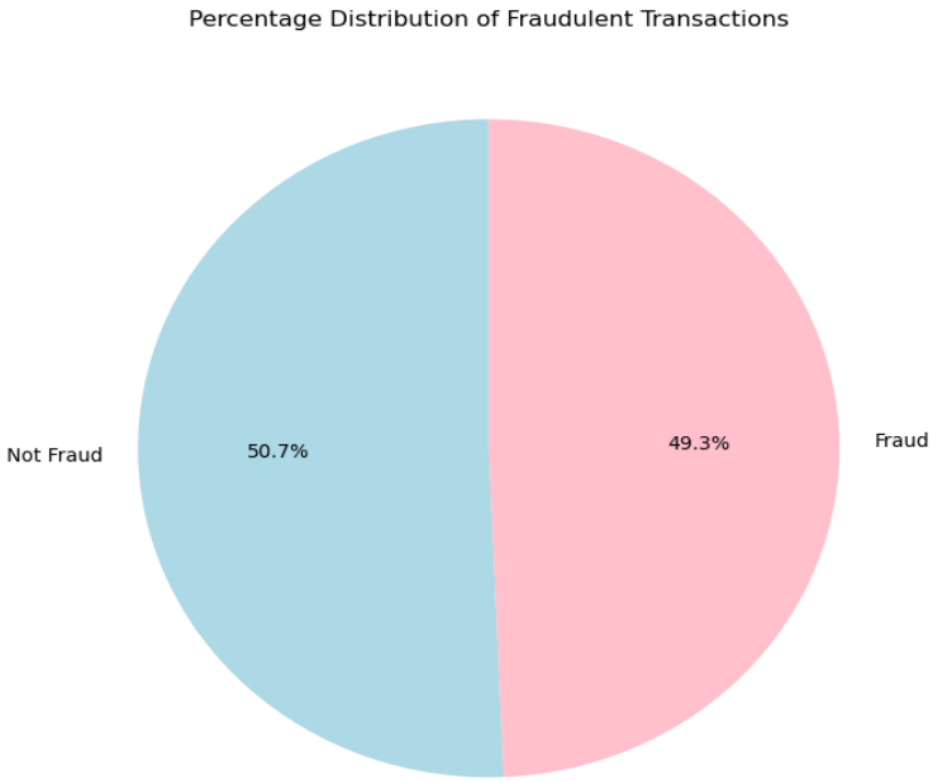


Fig. 5.1.2 Percentage distribution of fraudulent transactions

5.1.3. Distribution of Purchase Category:

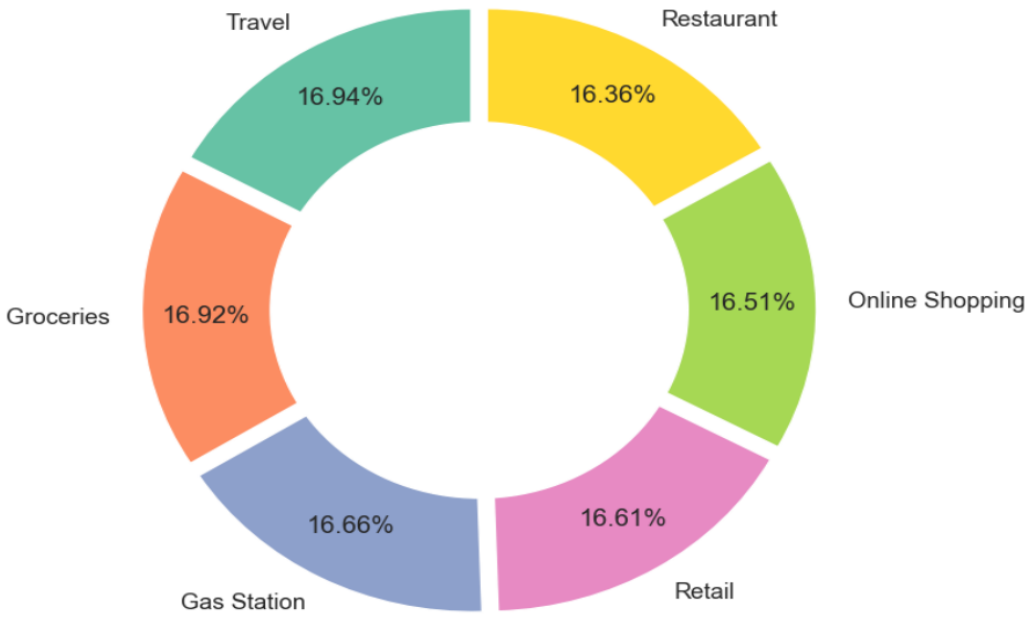


Fig. 5.1.3 Distribution of purchase category

5.1.4. Distribution of Card type:

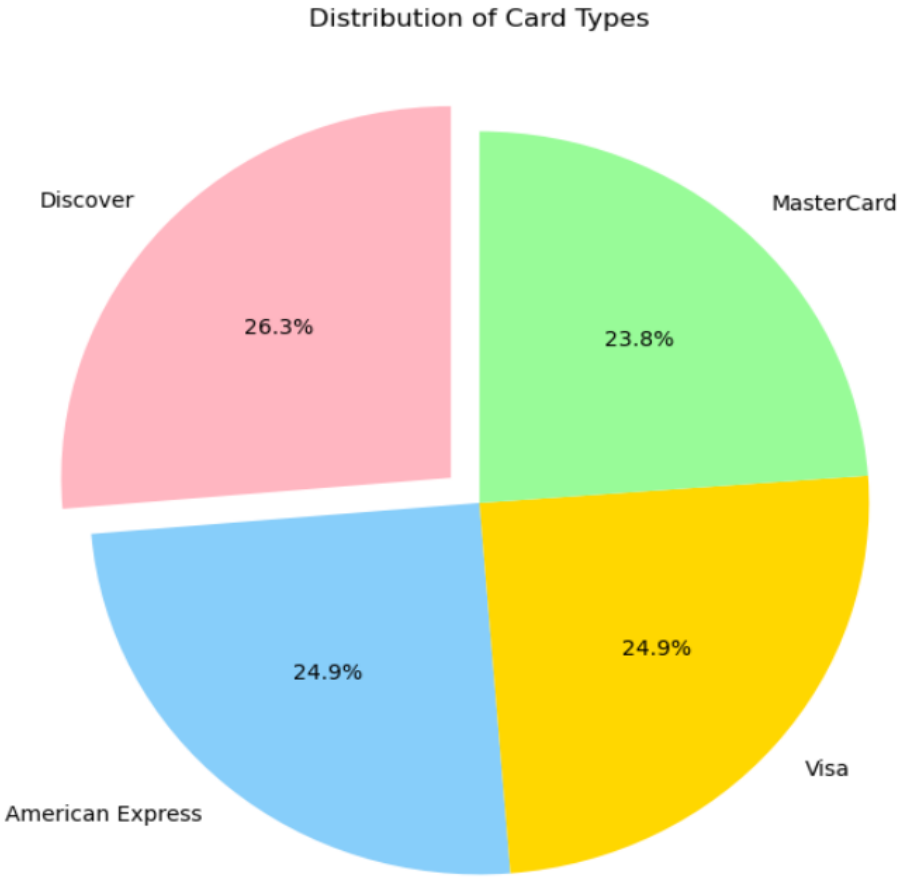


Fig. 5.1.4 Distribution of card types

5.2. Transaction Time Analysis:

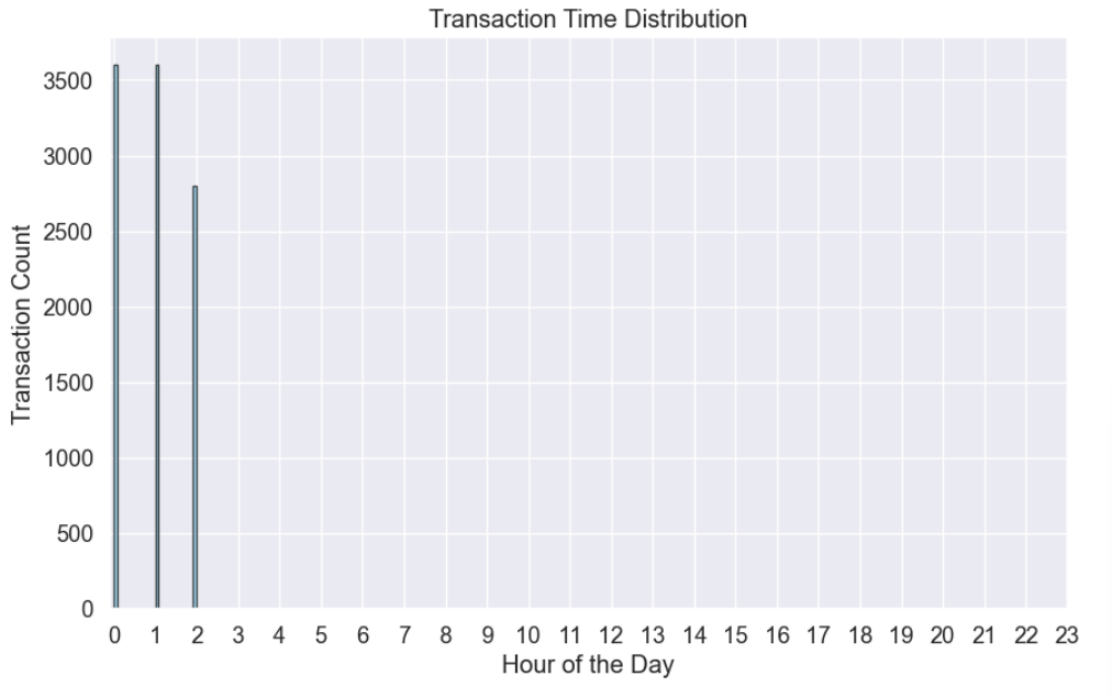


Fig. 5.2. Transaction Time Distribution

From figure 5.2, we can observe that in the dataset we are working, we have only transactions history occurred from 0 to 2 hours that is from midnight 12:00 AM to 2:00 AM.

5.3. Demographic Analysis:

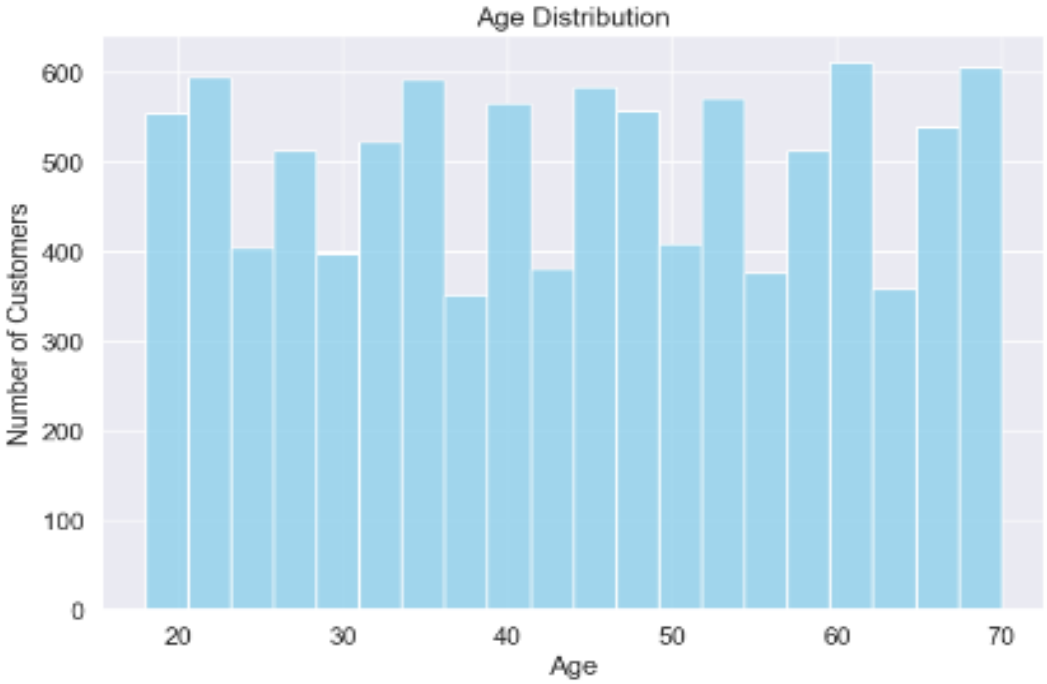


Fig. 5.3. Age Distribution

From the figure 5.3, we can observe that maximum number of customers are of age above 60.

5.4. Fraud Incidence Rate Analysis:

Fig. 5.4. Fraud incidence rates by segment

This graph answers the question Can customer segmentation help identify potential fraudsters or anomalous behavior? Are there specific segments with higher fraud incidence rates? Analyzing transaction patterns within segments can aid in fraud detection and prevention efforts.

In figure 3.4, Segment 0 indicates younger customers with lower transaction amounts. Segment 1 indicates customers with moderate transaction amounts across diverse age groups. Segment 2 indicates older, affluent customers with high transaction amounts. Segment 3 indicates younger customers with lower to moderate transaction amounts. Segment 4 indicates middle-aged to older customers with very high transaction amounts.

5.5. Analysis On Spending Patterns:

5.5.1. Spending patterns based on card type:

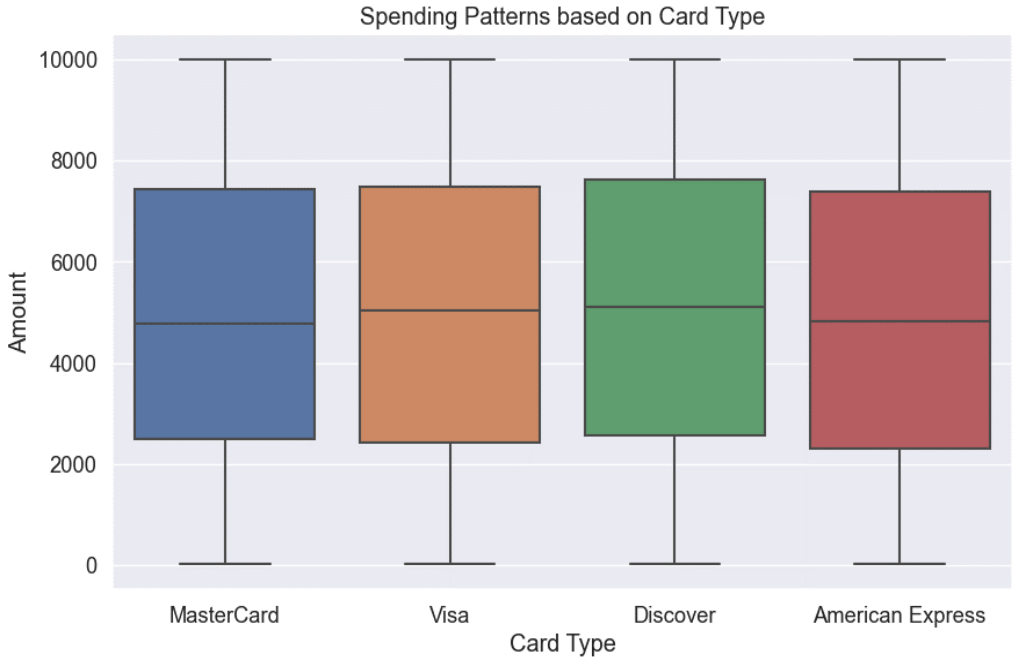


Fig. 5.5.1 Spending patterns based on card type

5.5.2. Spending patterns based on purchase category:

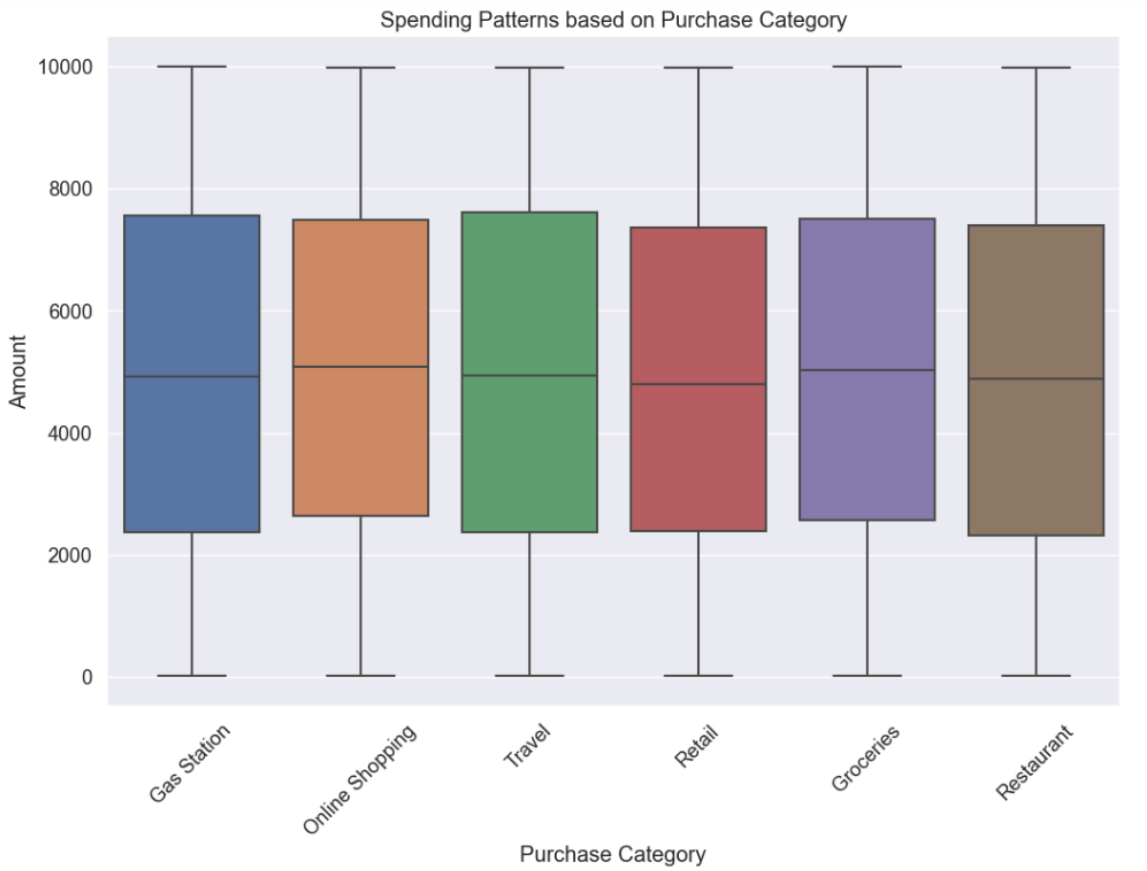


Fig. 5.5.2 Spending patterns based on purchase category

5.5.3. Spending behaviour comparison:

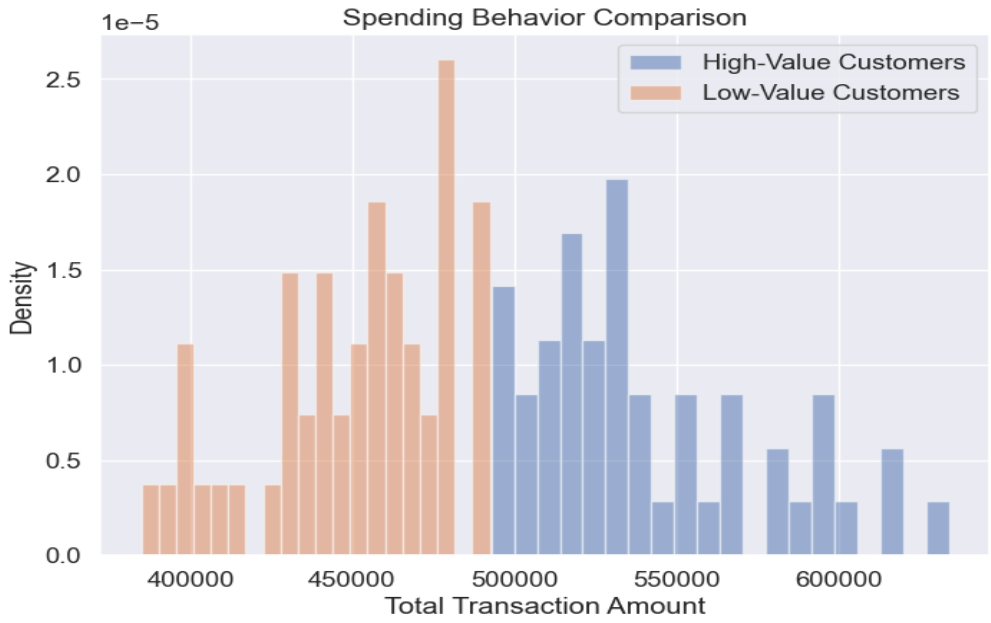


Fig. 5.5.3. Spending behaviour comparision

In the above graph, density represents the relative likelihood of observing transaction amounts within each group of customers, allowing for comparisons of spending behavior between high-value and low-value customer segments.

The total transaction amount for each customer is obtained by summing the transaction amounts associated with their respective customer IDs after grouping the dataset by 'customer_id'. This allows us to identify the customers who contribute the most to the total revenue.

5.5.4. Average Spending on Purchase Categories by Age Group:

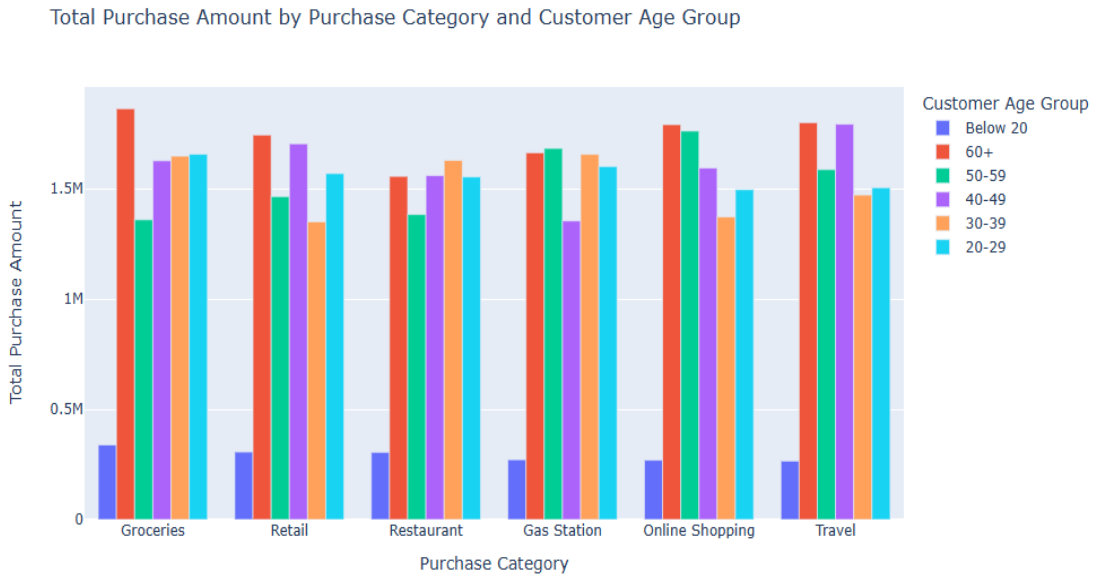


Fig. 5.5.4 Average Spending on Purchase Categories by Age Group

From figure 5.5.4, we can observe that:

Groceries and Retail Lead Spending: Across all age groups, groceries and retail seem to be the top spending categories.

Spending Increases with Age: The total purchase amount appears to increase as the customer age group gets older (except possibly 60+). This could be due to factors like higher income or larger household sizes in older demographics.

Younger Age Groups Spend More on Online Shopping: The 20-29 age group seems to have a higher proportion of spending on online shopping compared to older groups.

5.5.5. Average Spending on Purchase Categories by card type:

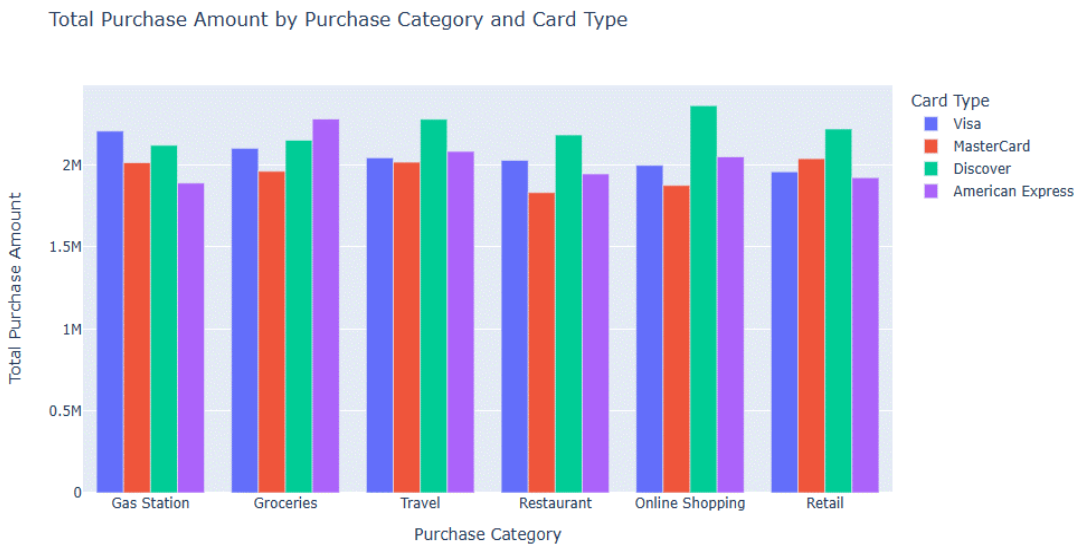


Fig. 5.5.5 Average Spending on Purchase Categories by Card Type

From figure 5.5.5, we can observe that Visa and Mastercard are the dominant card types in terms of total spending across most categories.

5.6. Fraud Rate Analysis:

5.6.1. Distribution of Fradulent Transactions by Purchase Categ

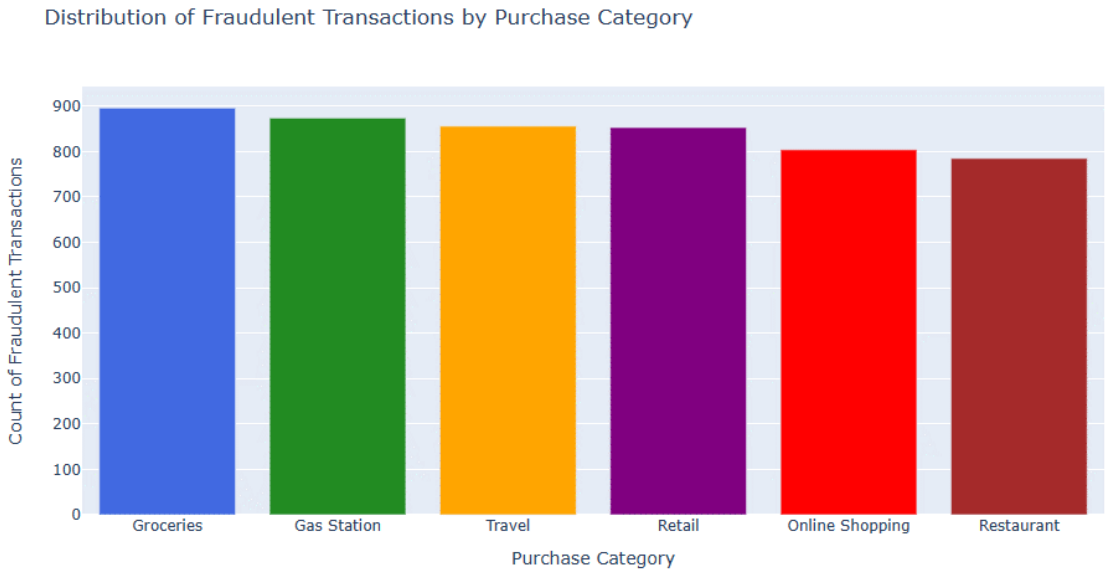


Fig. 5.6.1. Distribution of Fraudulent Transactions by Purchase Category

From the above figure, we can observe that Grocery stores appear to have the highest percentage of fraudulent transactions, while restaurants have the lowest.

5.6.2. Distribution of Fraudulent Transactions according to City:

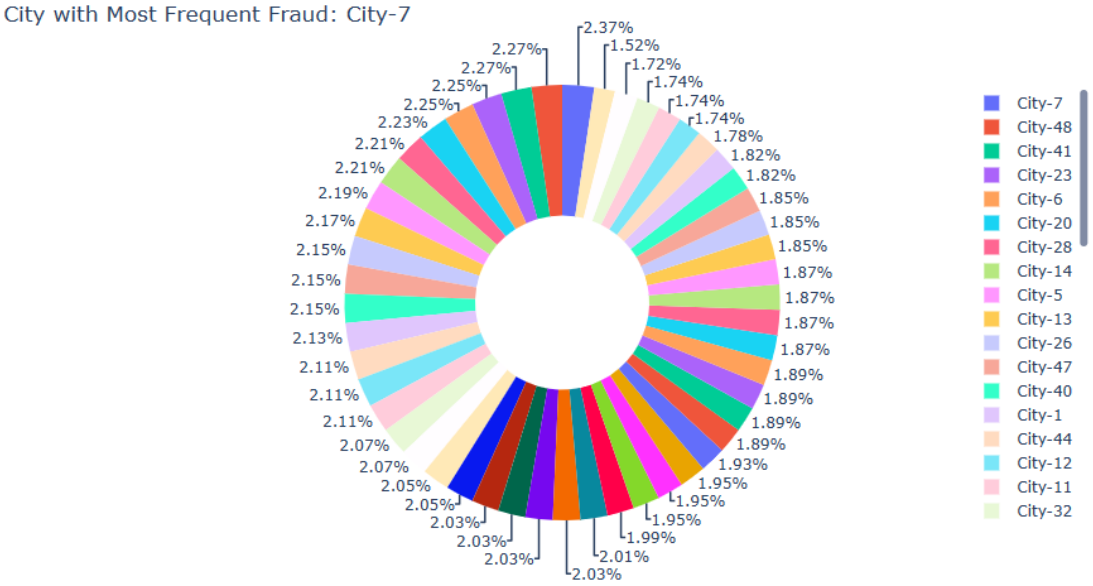


Fig. 5.6.2. Distribution of Fraudulent Transactions by city

5.6.3. Distribution of Fraudulent Transactions by Card Type and Age Group:

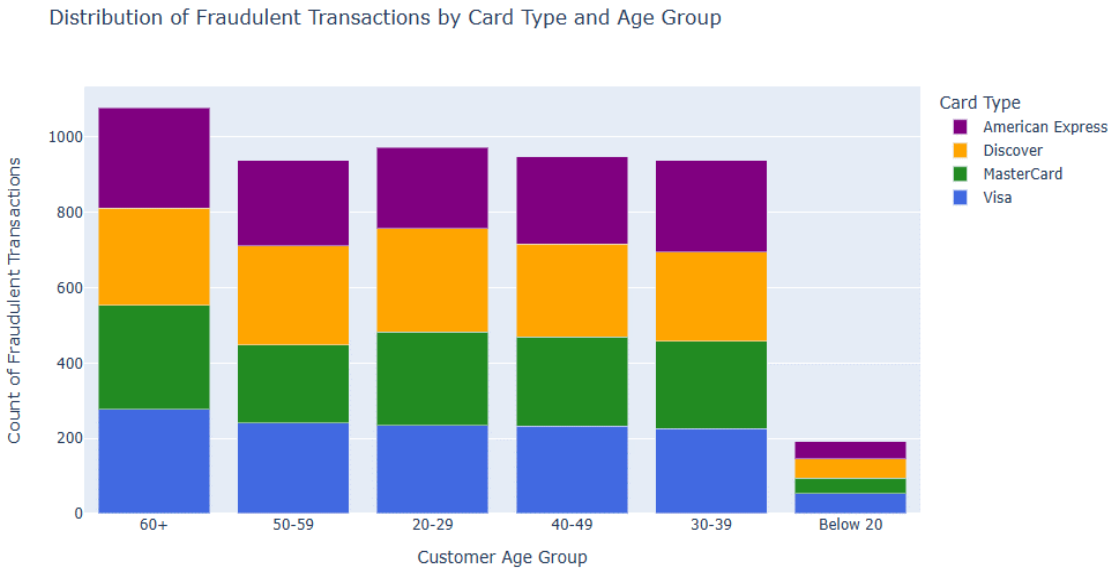


Fig. 5.6.3. Distribution of Fraudulent Transactions by Card type and Age group

From the above graph we can observe that, younger age groups (20-29) seem to have the lowest overall fraud rates across most card types.

5.7. Distrubution of Card Usage by Card Type and Age Group

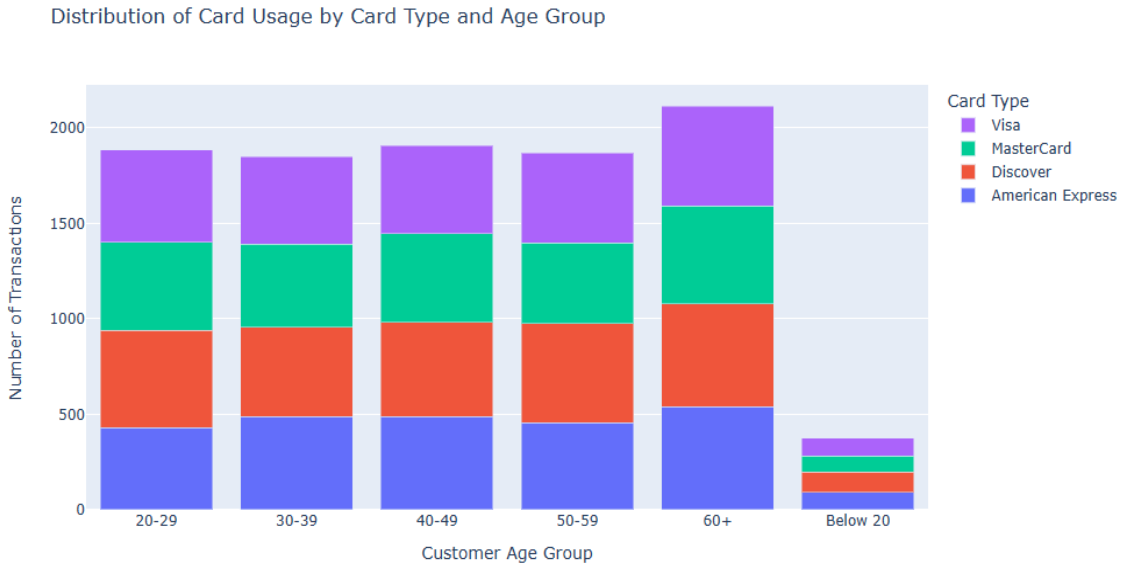


Fig. 5.7. Distribution of Card Usage by Card Type and Age Group

From the above figure, we can observe that customers of age above 60 are doing more transactions and visa card usage is more.

5.8. Transaction Count by Location and Age Group:

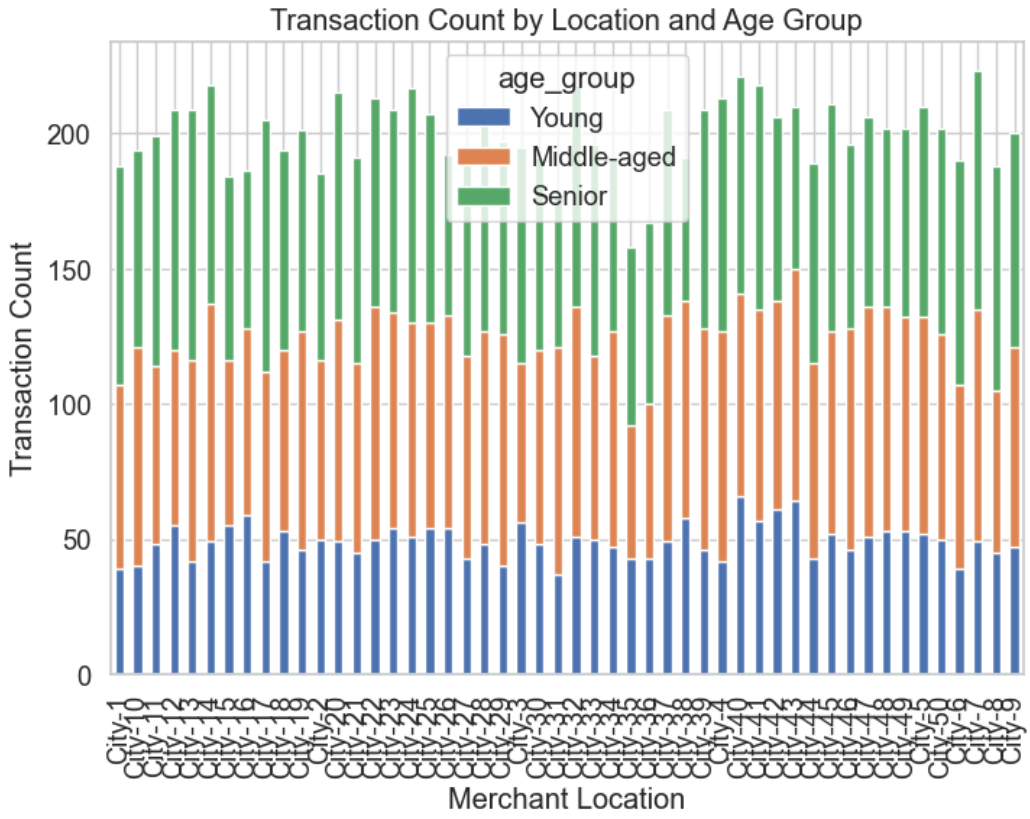


Fig. 5.8. Transaction Count by Location and Age Group

This graph displays transaction counts by location and age group, with color intensity denoting count.

Visibility issues: X-axis (age group) labels are absent, and Y-axis (location) labels are condensed.

Noteworthy: Certain locations like City-37 exhibit higher transaction counts.

Transaction counts vary across age groups within locations; e.g., City-17 favors "young" age group, City-25 favors "middle-aged".

While graph suggests variation in transaction count by location and age group, better axis labeling would enhance analysis.

5.9. Transaction Count by Purchase Category and Age Group:

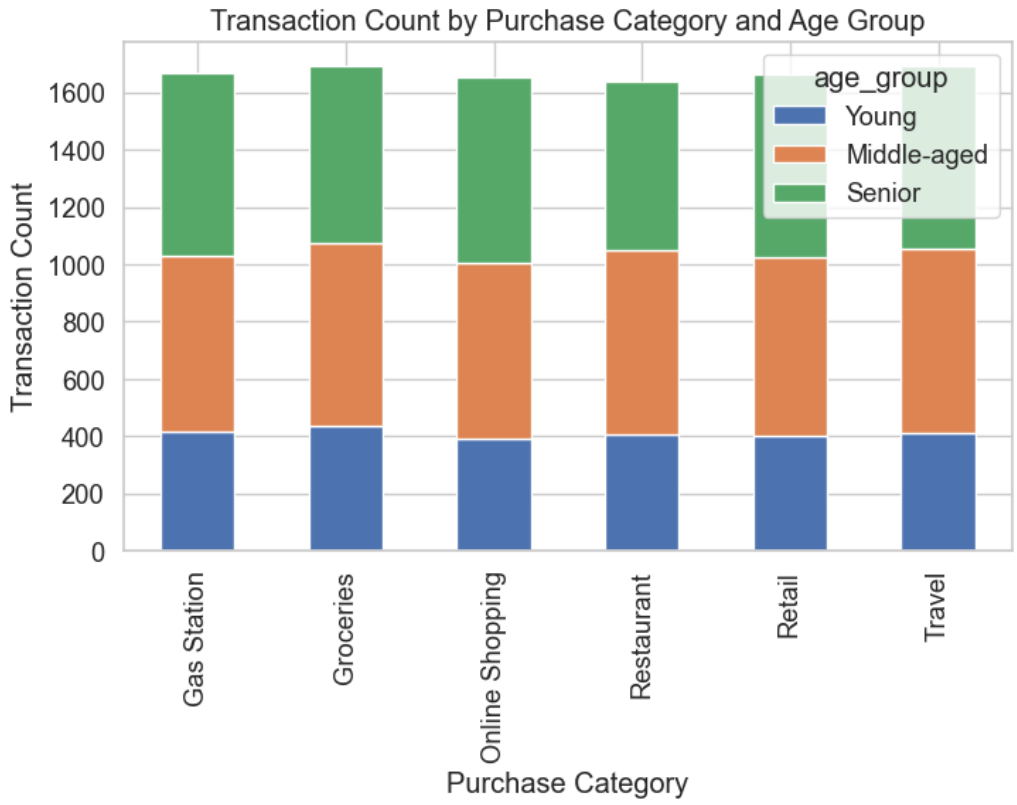


Fig. 5.9. Transaction Count by Purchase Category and Age Group

This graph displays average transaction count by purchase category and age group.

Y-axis represents average transaction count, while X-axis shows purchase categories.

Bars are color-coded for different age groups: young, middle-aged, and senior.

Observation: Senior citizens generally have the fewest transactions across all purchase categories. Young individuals show highest transaction counts at gas stations, followed by restaurants and online shopping. Middle-aged individuals tend to make the most transactions at grocery stores, followed by gas stations and restaurants.

Trend: Transaction amounts at gas stations decrease with age, possibly due to reduced driving or increased fuel efficiency.

Notable: Young individuals exhibit higher online transaction rates compared to other age groups, possibly due to greater comfort with online shopping or distinct spending habits.

5.10. Merchant Performance Analysis:

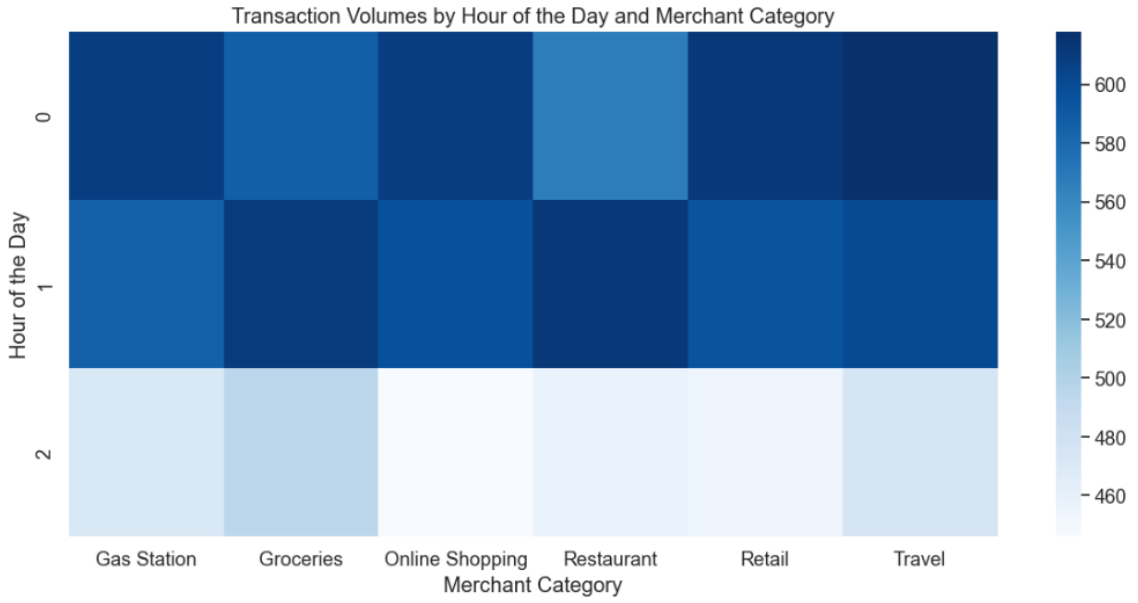


Fig. 5.10. Transaction Volumes by Hour of the Day and Merchant Category

Overall, transaction volume is higher during daytime hours (8am to 8pm) and lower at night (10pm to 6am), indicating increased purchasing activity during waking hours.

Grocery stores exhibit steady transaction volume throughout the day compared to other categories, possibly due to continuous grocery shopping needs.

Gas stations experience transaction volume spikes in the morning (8am to 10am) and afternoon (4pm to 6pm), suggesting increased fuel purchases during commute times.

Online shopping demonstrates transaction volume spikes in the evening (8pm to 10pm), indicating heightened online shopping activity after work hours or during weekends.

5.11. Comparison of Transaction Amounts: Fraudulent vs. Legitimate:

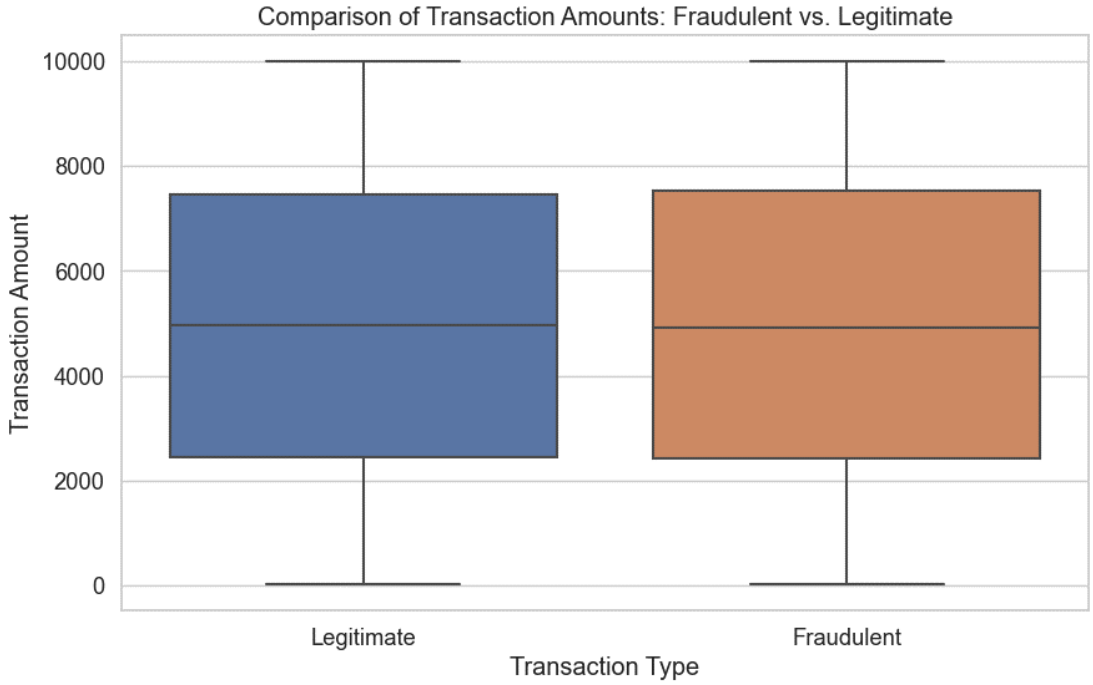


Fig. 5.11. Comparison of Transaction Amounts: Fraudulent vs. Legitimate

Fraudulent transactions involve smaller amounts: The median transaction amount is lower for fraudulent purchases.

Wider range for fraudulent amounts: Fraudulent transactions show a wider spread of values compared to legitimate ones.

More outliers in fraudulent transactions: There are more outliers in the fraudulent transaction amounts, suggesting some suspicious high or low value purchases.

5.12. CorrelationAnalysis:

5.12.1. Relationship between card type and location based on purchase category:

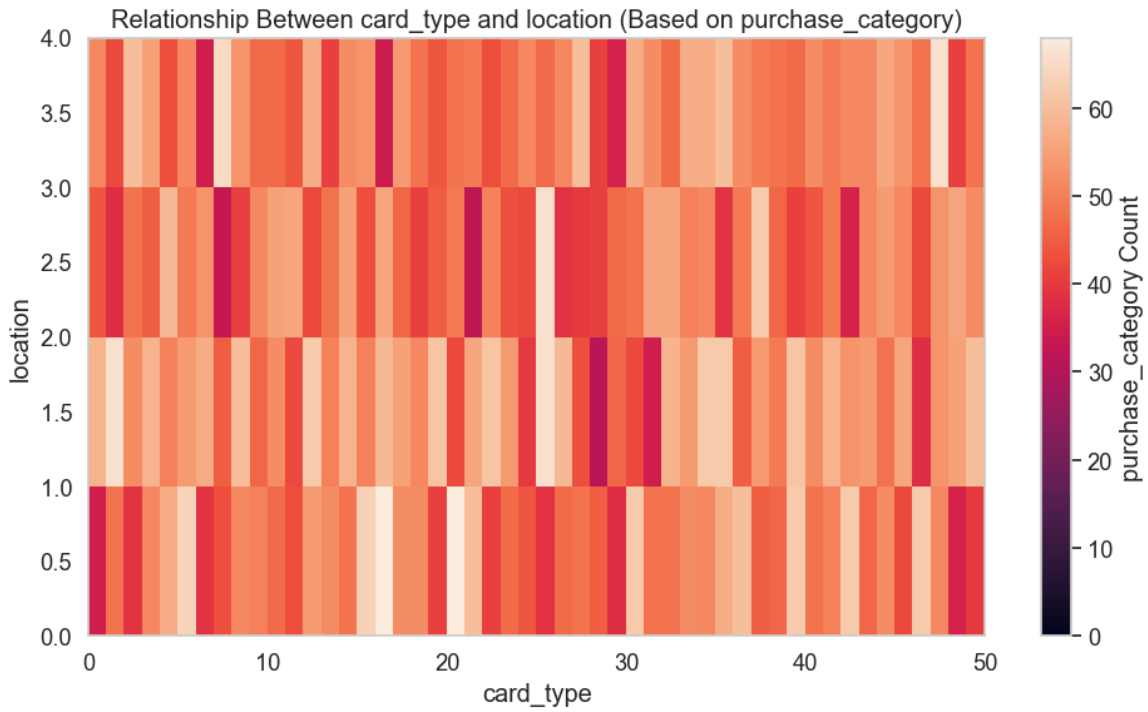


Fig. 5.12.1. Relationship between card type and location based on purchase category

From the correlation analysis, we can discern the predominant card type usage across different purchase categories and infer the favored card types in specific locations.

5.12.2. Relationship between amount and customer age:

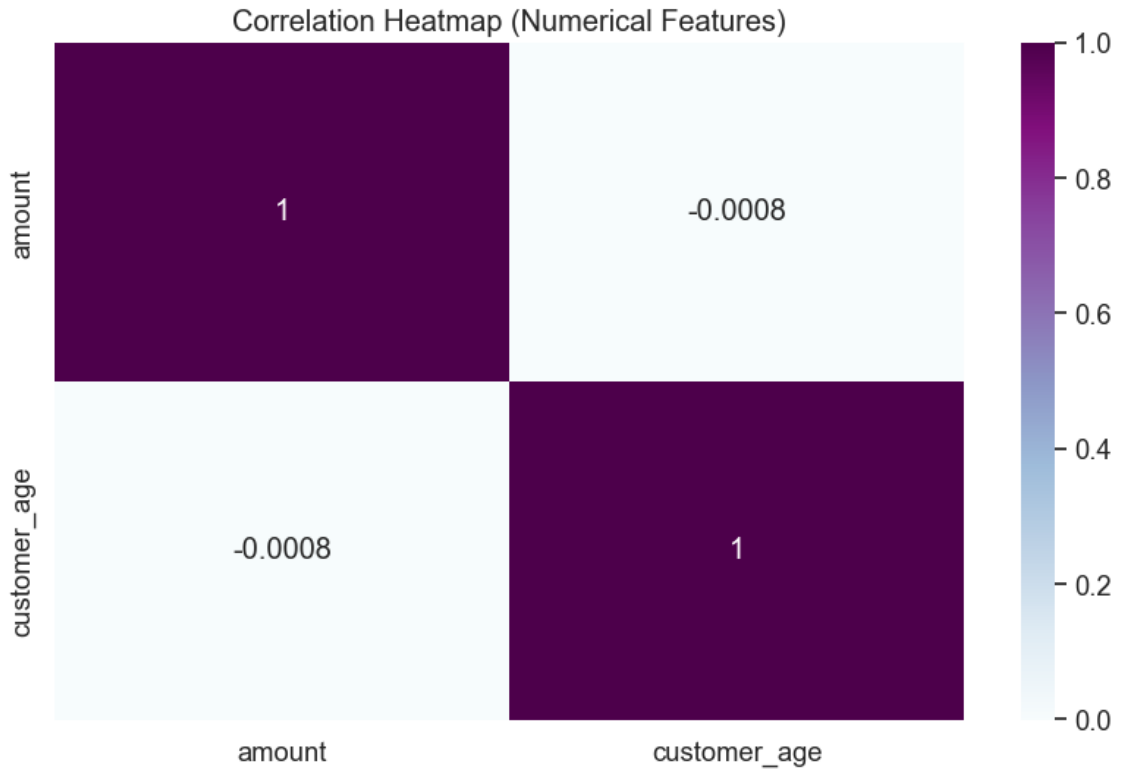


Fig. 5.12.2 Relationship between amount and customer age

Dark colour indicates a positive correlation (values tend to move together), light blue colour indicates a negative correlation (values tend to move in opposite directions).

The closer the colour is to dark blue or light blue, the stronger the correlation. The closer the colour is to white, the weaker the correlation.

5.12.3. Relationship between customer age,amount and is fraudulent:

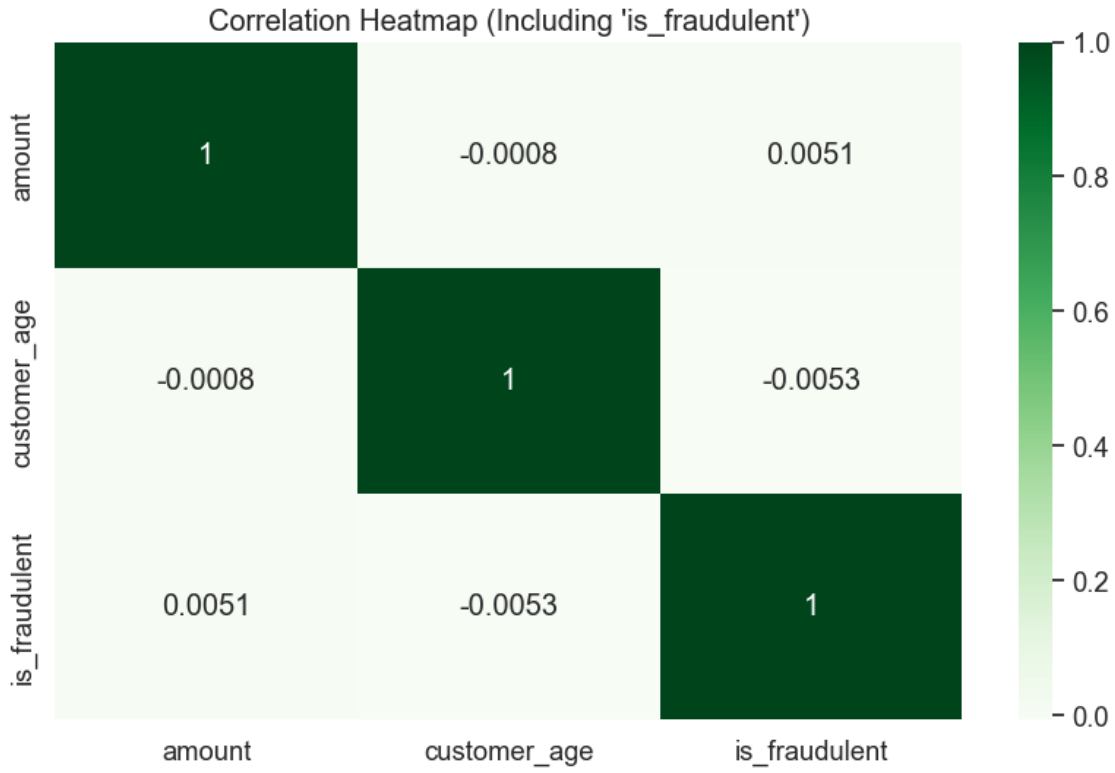


Fig. 5.12.3 Relationship between customer age,amount and is fraudulent

This map indicates that amount and is_fraudulent categories have positive correlation that means they will move in same direction where as customer age, amount and is_fraudulent have negative correlation that means they will move in opposite direction.

CONCLUSION AND FUTURE SCOPE

CONCLUSION

In conclusion, our exhaustive analysis of the financial dataset has yielded pivotal insights into consumer spending behaviors, fraud detection mechanisms, and transactional dynamics. We've discerned that groceries and retail are consistently the top spending categories across all age groups. Moreover, as customers age, their total purchase amount tends to increase, except possibly for those aged 60 and above, which may be attributed to factors like higher income or larger household sizes.

Additionally, younger age groups, particularly the 20-29 demographic, demonstrate a higher proportion of spending on online shopping compared to older groups, indicating evolving consumer behaviour towards e-commerce.

Furthermore, our scrutiny of transactional volumes across demographics and locales has unearthed discernible divergences, underscoring the indispensable role of retail analytics services in refining operational strategies and augmenting customer experiences.

Moreover, our visualization of fraud rates across varying card types and age groups has furnished invaluable insights, facilitating the bespoke customization of fraud prevention solutions catering to specific demographic segments.

By assimilating these findings into our analytics framework, we stand poised to fortify fraud prevention strategies and elevate customer engagement endeavors, thereby fostering sustainable progress within the financial landscape.

FUTURE SCOPE

The future scope lies in leveraging insights from consumer spending behaviors and transaction patterns. Groceries and retail continue to dominate expenditure across all age groups, indicating sustained demand. As customers age, their purchasing power increases, highlighting opportunities for targeted marketing and tailored product offerings. Moreover, the rising trend of online shopping among younger demographics underscores

the need for innovative e-commerce platforms catering to evolving consumer preferences.

In addition, the prevalence of Visa and Mastercard for transactions suggests the importance of optimizing payment systems to enhance convenience and security. Visualizing fraud rates enables proactive fraud prevention measures, particularly for vulnerable demographics such as younger customers. Transaction counts by location and age group offer valuable insights into consumer behaviour, facilitating targeted marketing strategies and operational optimizations.

Overall, the future lies in harnessing data-driven approaches to adapt to changing consumer trends, mitigate risks, and capitalize on emerging opportunities in the retail and financial sectors.

REFERENCES

[1] The 4 Types of Data Analysis [Ultimate Guide] (careerfoundry.com)

[2] Exploratory Data Analysis+ML_Fraud_Detection (kaggle.com)

APPENDIX - I

SOURCE CODE

Data Cleaning and Preprocessing:

Check for null values:

```
df.isnull().any()
```

Check for duplicate values:

```
duplicate_values=df.duplicated().sum()
print(duplicate_values)
```

Outlier Detection:

```
# Convert numeric columns to numeric type
numeric_cols = df.select_dtypes(include=['float64', 'int64']).columns
df[numeric_cols] = df[numeric_cols].apply(pd.to_numeric, errors='coerce')
# Identify non-numeric columns
non_numeric_cols = df.columns.difference(numeric_cols)
# Drop non-numeric columns for outlier detection
numeric_df = df.drop(columns=non_numeric_cols)
# Outlier detection and treatment using IQR method
Q1 = numeric_df.quantile(0.25)
Q3 = numeric_df.quantile(0.75)
IQR = Q3 - Q1
outliers = ((numeric_df < (Q1 - 1.5 * IQR)) | (numeric_df > (Q3 + 1.5 * IQR))).any(axis=1)
print("Number of Outliers:", outliers.sum())
# Removing outliers
```

 # df = df[~outliers]

Univariate Analysis:

```
# Set seaborn style
sns.set_style("whitegrid")
# Filter the dataset to include only relevant columns
relevant_columns = ['customer_id', 'amount', 'customer_age', 'purchase_category']
filtered_df = df[relevant_columns]

# Segment customers based on age groups (e.g., bins of 10 years)
age_bins = [0, 20, 30, 40, 50, 60, 70, 80, 90, 100]  # Define age bins
age_labels = ['0-20', '21-30', '31-40', '41-50', '51-60', '61-70', '71-80', '81-90', '91-100']  #
Define age group labels
filtered_df['age_group'] = pd.cut(filtered_df['customer_age'], bins=age_bins,
labels=age_labels, right=False)
# Create a figure and axis objects
fig, axes = plt.subplots(nrows=2, ncols=1, figsize=(12, 14))

# Analyze spending behavior between different age groups
ax1 = axes[0]
sns.barplot(data=filtered_df, x='age_group', y='amount', estimator='mean', ci=None, ax=ax1,
palette='coolwarm')
ax1.set_title('Average Spending by Age Group', fontsize=16)
ax1.set_xlabel('Age Group', fontsize=14)
ax1.set_ylabel('Average Amount', fontsize=14)
ax1.tick_params(axis='x', labelrotation=45)

# Investigate spending on specific purchase categories by age groups
ax2 = axes[1]
sns.barplot(data=filtered_df, x='age_group', y='amount', hue='purchase_category',
estimator='mean', ci=None, ax=ax2, palette='muted')
ax2.set_title('Average Spending on Purchase Categories by Age Group', fontsize=16)
ax2.set_xlabel('Age Group', fontsize=14)
ax2.set_ylabel('Average Amount', fontsize=14)
ax2.legend(title='Purchase Category', title_fontsize='13', fontsize='12',
bbox_to_anchor=(1.05, 1), loc='upper left')
ax2.tick_params(axis='x', labelrotation=45)

# Adjust layout
plt.tight_layout()

# Show plot
```

```
plt.show()
```

Multivariate Analysis:

Customer Behavior Segmentation:

How do different customer segments interact with merchant attributes (location, category), and how does this interaction influence their transaction patterns? :

```
import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt
```

 # Load the synthetic financial data

```
synthetic_financial_data = pd.read_csv("synthetic_financial_data.csv")
```

 # Group the data by customer segments (e.g., age groups)

 # For this example, let's group customers into three age groups: young, middle-aged, and senior

```
synthetic_financial_data['age_group'] = pd.cut(synthetic_financial_data['customer_age'], bins=[0, 30, 50, float('inf')], labels=['Young', 'Middle-aged', 'Senior'])
```

 # Group the data by location and age group, and count transactions

```
location_age_group_count = synthetic_financial_data.groupby(['location', 'age_group']).size().unstack(fill_value=0)
```

 # Group the data by purchase category and age group, and count transactions

```
category_age_group_count = synthetic_financial_data.groupby(['purchase_category', 'age_group']).size().unstack(fill_value=0)
```

 # Plotting

```
plt.figure(figsize=(14, 6))
```

 # Plot for transaction counts by location and age group

```
plt.subplot(1, 2, 1)

location_age_group_count.plot(kind='bar', stacked=True, ax=plt.gca())

plt.title('Transaction Count by Location and Age Group')
plt.xlabel('Merchant Location')
plt.ylabel('Transaction Count')

# Plot for transaction counts by purchase category and age group
plt.subplot(1, 2, 2)
category_age_group_count.plot(kind='bar', stacked=True, ax=plt.gca())
plt.title('Transaction Count by Purchase Category and Age Group')
plt.xlabel('Purchase Category')
plt.ylabel('Transaction Count')

# Adjust layout
plt.tight_layout()
plt.show()
```

Descriptive Analysis:

Customer Spending Patterns:

Do fraudulent transactions differ significantly in transaction amounts compared to legitimate ones?

```
import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

# Load the synthetic financial data

synthetic_financial_data = pd.read_csv("synthetic_financial_data.csv")

# Plotting

plt.figure(figsize=(10, 6))

# Create box plot to compare transaction amounts between fraudulent and legitimate transactions

sns.boxplot(x='is_fraudulent', y='amount', data=synthetic_financial_data)

plt.title('Comparison of Transaction Amounts: Fraudulent vs. Legitimate')

plt.xlabel('Transaction Type')

plt.ylabel('Transaction Amount')

# Customize x-axis labels

plt.xticks(ticks=[0, 1], labels=['Legitimate', 'Fraudulent'])

plt.show()
```

Temporal Analysis:

```
import pandas as pd
import matplotlib.pyplot as plt

# Load the dataset
df = pd.read_csv('synthetic_financial_data.csv')
# Convert 'transaction_time' column to datetime format
df['transaction_time'] = pd.to_datetime(df['transaction_time'])
# Extract hour information from 'transaction_time'
df['hour'] = df['transaction_time'].dt.hour

# Plot the distribution of transaction times
plt.figure(figsize=(10, 6))
plt.hist(df['hour'], bins=24, color='skyblue', edgecolor='black', alpha=0.7)
plt.title('Transaction Time Distribution')
plt.xlabel('Hour of the Day')
plt.ylabel('Transaction Count')
plt.xticks(range(24))
plt.grid(True)
plt.show()
```

Visualization:

 # Plotting

```
plt.figure(figsize=(24, 6))  # Increase the width of the figure
```

 # Age Distribution

```
plt.subplot(1, 3, 1)
sns.histplot(data=df, x='customer_age', color='skyblue', bins=20)
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Number of Customers')
```

 # Location Analysis

```
plt.subplot(1, 3, 2)
sns.countplot(data=df, y='location', palette='pastel')
plt.title('Location Analysis')
plt.xlabel('Number of Customers')
plt.ylabel('Location')
```

 # Rotate x-axis labels for better readability

```
plt.xticks(rotation=45)
```

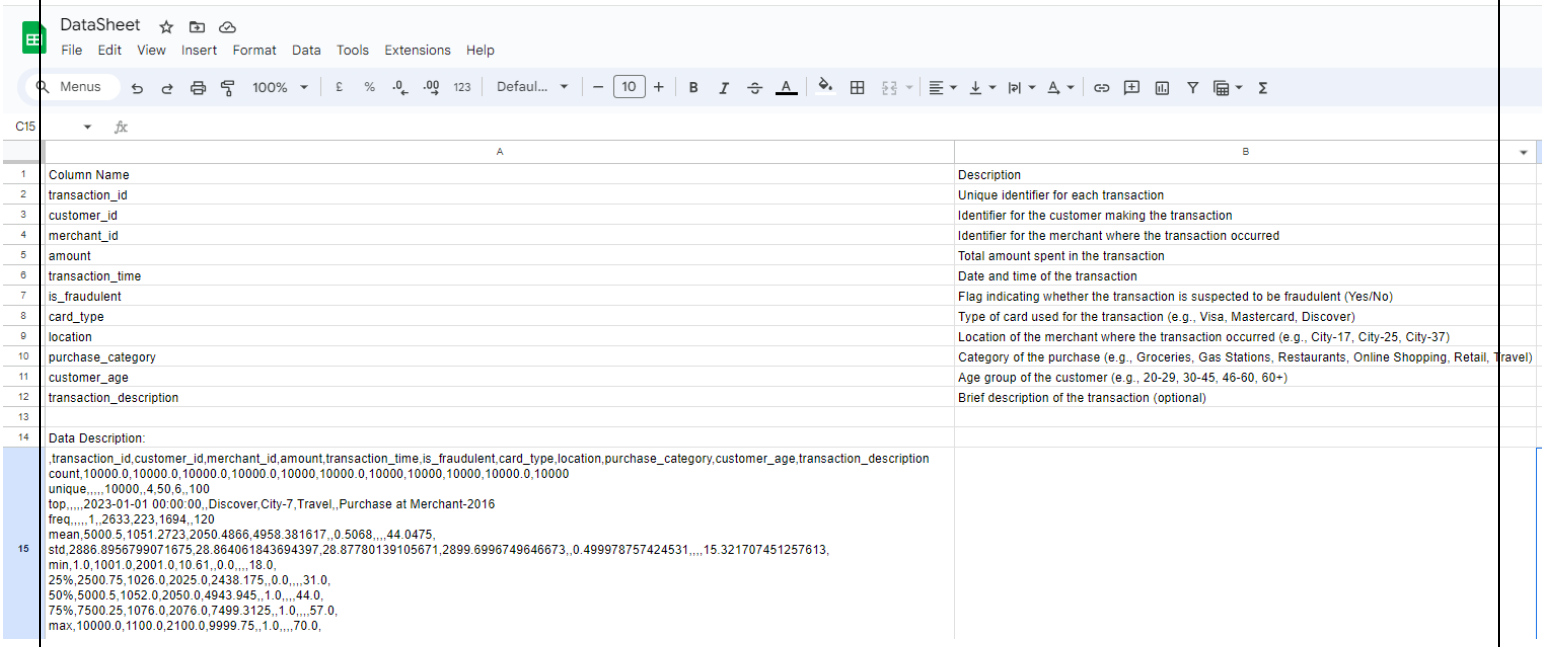
 # Card Type Distribution

```
plt.subplot(1, 3, 3)
card_type_counts = df['card_type'].value_counts()
plt.pie(card_type_counts, labels=card_type_counts.index, autopct='%1.1f%%', startangle=90, colors=['pink', 'saddlebrown','sandybrown','rosybrown'])
plt.title('Card Type Distribution')

plt.tight_layout()
plt.show()
```

APPENDIX-II

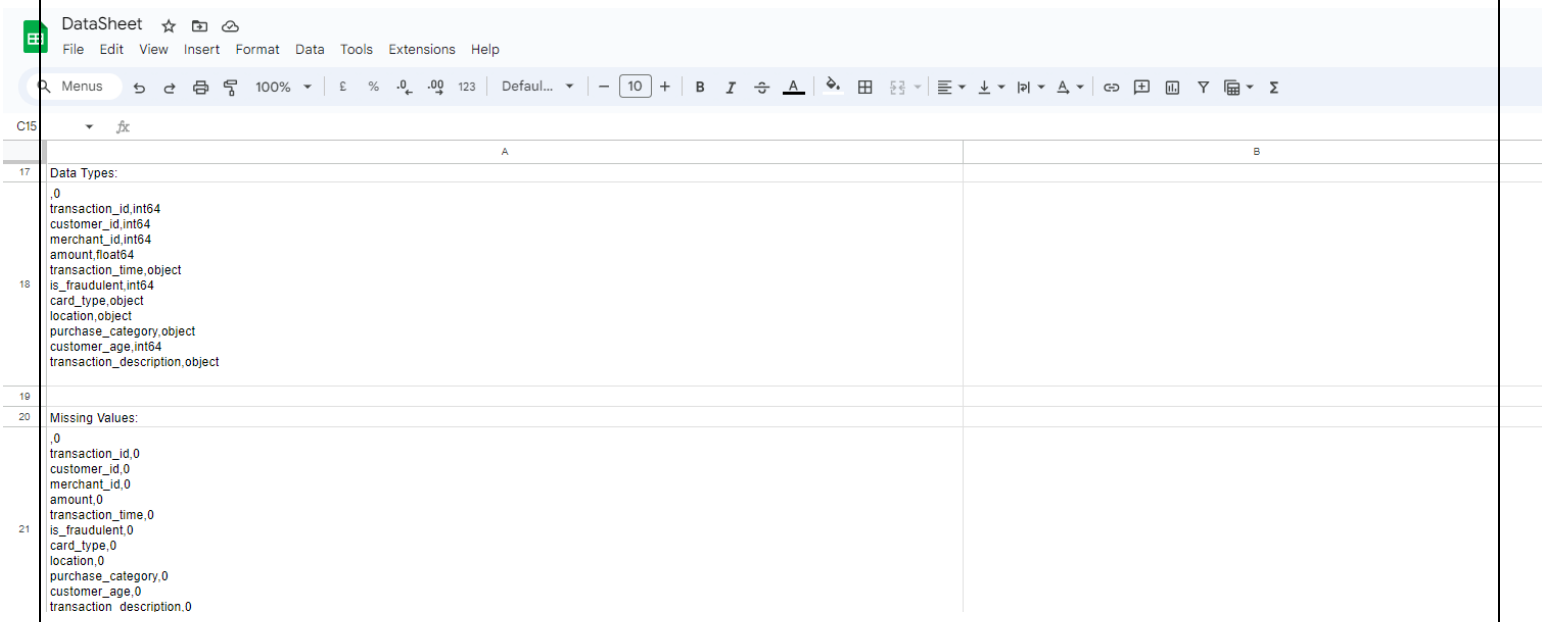
DATASHEETS



DataSheet

File Edit View Insert Format Data Tools Extensions Help

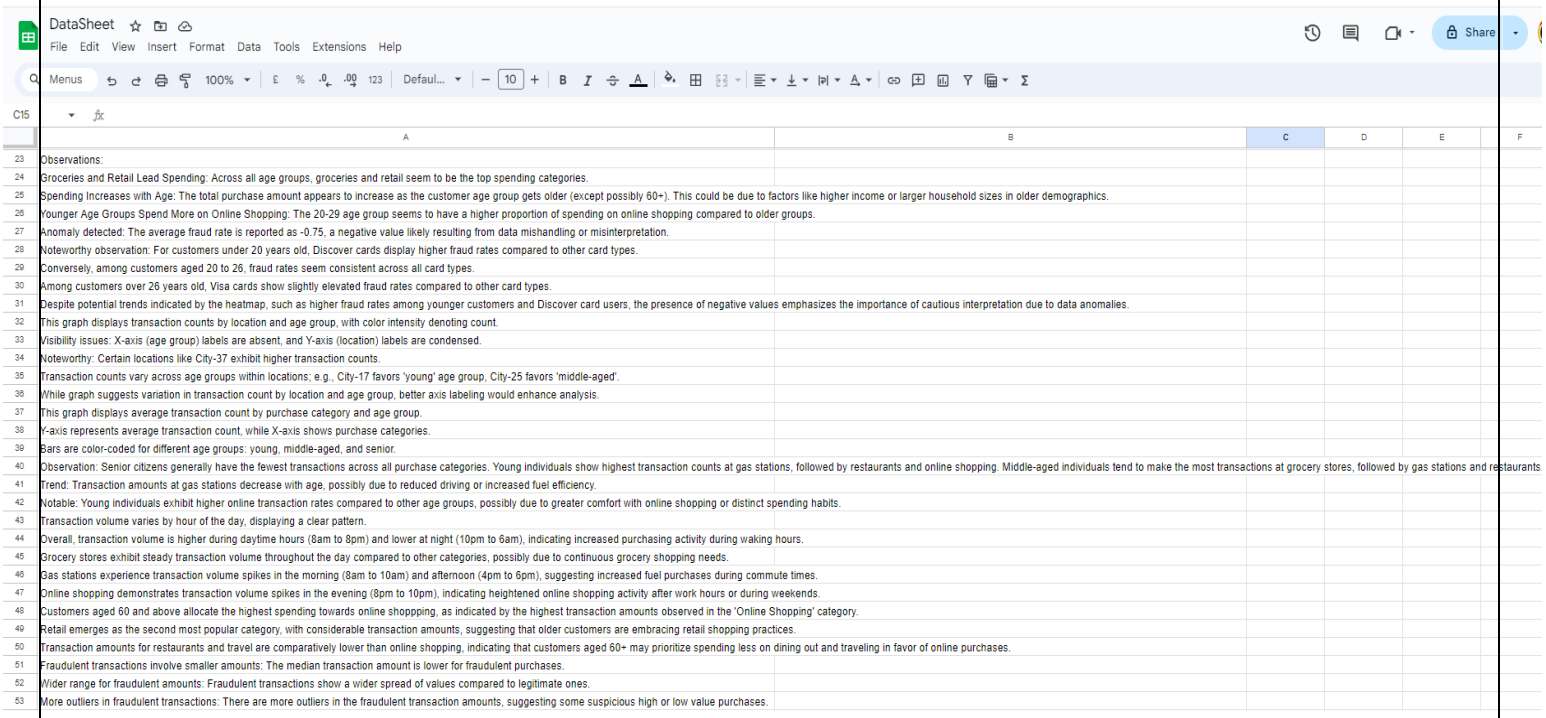
	A	B
1	Column Name	Description
2	transaction_id	Unique identifier for each transaction
3	customer_id	Identifier for the customer making the transaction
4	merchant_id	Identifier for the merchant where the transaction occurred
5	amount	Total amount spent in the transaction
6	transaction_time	Date and time of the transaction
7	is_fraudulent	Flag indicating whether the transaction is suspected to be fraudulent (Yes/No)
8	card_type	Type of card used for the transaction (e.g., Visa, Mastercard, Discover)
9	location	Location of the merchant where the transaction occurred (e.g., City-17, City-25, City-37)
10	purchase_category	Category of the purchase (e.g., Groceries, Gas Stations, Restaurants, Online Shopping, Retail, Travel)
11	customer_age	Age group of the customer (e.g., 20-29, 30-45, 46-60, 60+)
12	transaction_description	Brief description of the transaction (optional)
13		
14	Data Description:	
15	,transaction_id,customer_id,merchant_id,amount,transaction_time,is_fraudulent,card_type,location,purchase_category,customer_age,transaction_description count,10000.0,10000.0,10000.0,10000.0,10000,10000.0,10000,10000,10000,10000.0,10000 unique,,,,,10000,,4,50,6,,100 top,,,,,2023-01-01 00:00:00,,Discover,City-7,Travel,,Purchase at Merchant-2016 freq,,,,,1,,2633,223,1694,,120 mean,5000.5,1051.2723,2050.4866,4958.381617,,0.5068,,,,44.0475, std,2886.895679907167 5,28.864061843694397,28.87780139105671,2899.6996749646673,,0.499978757424531,,,,15.321707451257613, min,1.0,1001.0,2001.0,10.61,,0.0,,,,18.0, 25%,2500.75,1026.0,2025.0,2438.175,,0.0,,,,31.0, 50%,5000.5,1052.0,2050.0,4943.945,,1.0,,,,44.0, 75%,7500.25,1076.0,2076.0,7499.3125,,1.0,,,,57.0, max,10000.0,1100.0,2100.0,9999.75,,1.0,,,,70.0,	

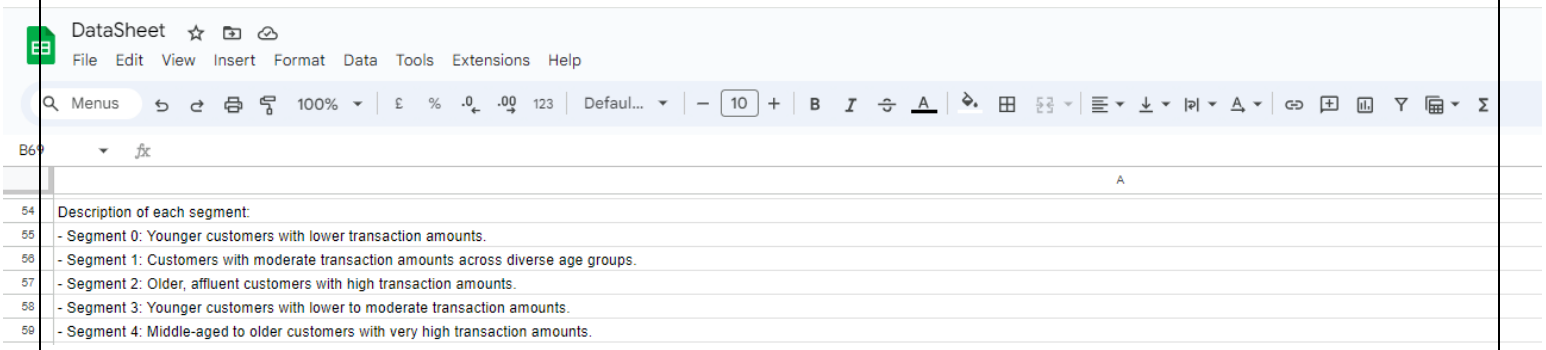


DataSheet

File Edit View Insert Format Data Tools Extensions Help

	A	B
17	Data Types:	
18	,0 transaction_id,int64 customer_id,int64 merchant_id,int64 amount,float64 transaction_time,object is_fraudulent,int64 card_type,object location,object purchase_category,object customer_age,int64 transaction_description,object	
19		
20	Missing Values:	
21	,0 transaction_id,0 customer_id,0 merchant_id,0 amount,0 transaction_time,0 is_fraudulent,0 card_type,0 location,0 purchase_category,0 customer_age,0 transaction_description,0	


DataSheet
File Edit View Insert Format Data Tools Extensions Help
Share
Menus
100%
Defaul...
10
C15
A
B
C
D
E
F
23 Observations:
24 Groceries and Retail Lead Spending: Across all age groups, groceries and retail seem to be the top spending categories.
25 Spending Increases with Age: The total purchase amount appears to increase as the customer age group gets older (except possibly 60+). This could be due to factors like higher income or larger household sizes in older demographics.
26 Younger Age Groups Spend More on Online Shopping: The 20-29 age group seems to have a higher proportion of spending on online shopping compared to older groups.
27 Anomaly detected: The average fraud rate is reported as -0.75, a negative value likely resulting from data mishandling or misinterpretation.
28 Noteworthy observation: For customers under 20 years old, Discover cards display higher fraud rates compared to other card types.
29 Conversely, among customers aged 20 to 26, fraud rates seem consistent across all card types.
30 Among customers over 26 years old, Visa cards show slightly elevated fraud rates compared to other card types.
31 Despite potential trends indicated by the heatmap, such as higher fraud rates among younger customers and Discover card users, the presence of negative values emphasizes the importance of cautious interpretation due to data anomalies.
32 This graph displays transaction counts by location and age group, with color intensity denoting count.
33 Visibility issues: X-axis (age group) labels are absent, and Y-axis (location) labels are condensed.
34 Noteworthy: Certain locations like City-37 exhibit higher transaction counts.
35 Transaction counts vary across age groups within locations; e.g., City-17 favors 'young' age group, City-25 favors 'middle-aged'.
36 While graph suggests variation in transaction count by location and age group, better axis labeling would enhance analysis.
37 This graph displays average transaction count by purchase category and age group.
38 Y-axis represents average transaction count, while X-axis shows purchase categories.
39 Bars are color-coded for different age groups: young, middle-aged, and senior.
40 Observation: Senior citizens generally have the fewest transactions across all purchase categories. Young individuals show highest transaction counts at gas stations, followed by restaurants and online shopping. Middle-aged individuals tend to make the most transactions at grocery stores, followed by gas stations and restaurants.
41 Trend: Transaction amounts at gas stations decrease with age, possibly due to reduced driving or increased fuel efficiency.
42 Notable: Young individuals exhibit higher online transaction rates compared to other age groups, possibly due to greater comfort with online shopping or distinct spending habits.
43 Transaction volume varies by hour of the day, displaying a clear pattern.
44 Overall, transaction volume is higher during daytime hours (8am to 8pm) and lower at night (10pm to 6am), indicating increased purchasing activity during waking hours.
45 Grocery stores exhibit steady transaction volume throughout the day compared to other categories, possibly due to continuous grocery shopping needs.
46 Gas stations experience transaction volume spikes in the morning (8am to 10am) and afternoon (4pm to 6pm), suggesting increased fuel purchases during commute times.
47 Online shopping demonstrates transaction volume spikes in the evening (8pm to 10pm), indicating heightened online shopping activity after work hours or during weekends.
48 Customers aged 60 and above allocate the highest spending towards online shoppping, as indicated by the highest transaction amounts observed in the 'Online Shopping' category.
49 Retail emerges as the second most popular category, with considerable transaction amounts, suggesting that older customers are embracing retail shopping practices.
50 Transaction amounts for restaurants and travel are comparatively lower than online shopping, indicating that customers aged 60+ may prioritize spending less on dining out and traveling in favor of online purchases.
51 Fraudulent transactions involve smaller amounts: The median transaction amount is lower for fraudulent purchases.
52 Wider range for fraudulent amounts: Fraudulent transactions show a wider spread of values compared to legitimate ones.
53 More outliers in fraudulent transactions: There are more outliers in the fraudulent transaction amounts, suggesting some suspicious high or low value purchases.


DataSheet
File Edit View Insert Format Data Tools Extensions Help
Menus
100%
Defaul...
10
A
54 Description of each segment:
55 - Segment 0: Younger customers with lower transaction amounts.
56 - Segment 1: Customers with moderate transaction amounts across diverse age groups.
57 - Segment 2: Older, affluent customers with high transaction amounts.
58 - Segment 3: Younger customers with lower to moderate transaction amounts.
59 - Segment 4: Middle-aged to older customers with very high transaction amounts.

INFORMATION REGARDING STUDENT(S)

STUDENT NAME	EMAIL ID	PHONE NUMBER
AKANSHA SHETTY	AKANSHASHETTY07@GMAIL.COM	7259982774
CHIMIRALA KOWSTUBHA	KOWSTUBHACHIMIRALA@GMAIL.COM	7995110124
KAPAROTU VENKATA SURYA THARANI	MORESPACEE123@GMAIL.COM	8817683282