# Exploratory Data Analysis

The document provides a thorough exploratory data analysis (EDA) of a weekly synthetic dataset containing 10,000 financial transactions that happened every second, each having geographical, temporal, and behavioral variables. The primary goal of this EDA is to evaluate the dataset's structure and completeness, identify hidden patterns in transactional activity, and acquire vital insights on the nature and prevalence of fraudulent behavior. To further understand data variability and outliers, the study employs a variety of analytical tools, including missing values diagnostics, class imbalances in fraud labels, and transaction amount distributions (both raw and log-transformed). Temporal trends are investigated using hourly and daily fraud patterns, while spatial dynamics are recorded using city-level mappings, coordinate visualizations, and DBSCAN-based clustering to detect possible fraud hotspots.

Furthermore, the significance of individual variables is assessed using a Random Forest model, which provides a data-driven basis for future predictive modeling. Collectively, this EDA not only confirms the dataset's preparedness for machine learning applications but also identifies significant locations where fraudulent behavior is most prevalent, both in time and geography.

## Data Preparation

## Baseline Exploratory Data Analysis

The first data preparation began with the import of major Python libraries such as Pandas, Numpy, Matplotlib, Seaborn, and scikit-learn modules for machine learning and assessment. The synthetic financial transaction dataset (synthetic_financial_dataset.csv) was imported into a pandas dataframe for structured analysis.

To increase geographical interpretability, anonymized city codes (e.g. City-1 and City-2) were assigned to 15 major US cities. This mapping was added to the new city_name column. Each city was then allocated latitude and longitude values via a customized dictionary, allowing geographic charting and spatial fraud trend analysis.

A check for missing values revealed that the dataset included no blank entries, providing a clean input for subsequent processing. The goal variable is_fraudulent and the temporal parameter transcation_time were separated, and the alter was deleted from

the modeling feature set. Categorical variable were encoded with encoding pd.get_dummies(), resulting in data appropriate for machine learning models.

The initial investigation comprised descriptive statistics and histogram for the quantity variable, as well as class balance checks with count plots. These processes guaranteed that the system was ready for downstream modeling and visual fraud pattern identification.

## Population Mapped Exploratory Data Analysis

The second round of data preparation brought a population-aware upgrade by including actual demographic data. The top 15 most populous US cities were assigned synthetic city codes based external population figures from US_Cities.xlsx file, which includes fields like city, state_name, and population.

To synchronize both datasets, city names were standardised, and a mapping was built to replace synthetic city codes with realistic ones. This added demographic context to the information, allowing for the calculation of population-adjusted fraud measures (such as fraud rates per 100,000 people) and providing a more relevant foundation for geographical insights.

The improved dataset allowed for more complex visualizations, such as scatter plots and bubble charts that showed fraud rates vs population. These findings highlighted fraud exposure trends at the metropolitan scale, providing a more grounded insight of where fraud is concentrated and how it scales with city sizes.

This version of EDA emphasized realism and policy relevance, with implications for both pattern detection and city-level fraud reduction methods. The combined information provided the platform for visual storytelling about fraud distribution and demographic correlations.

## Key Visualizations and Insights

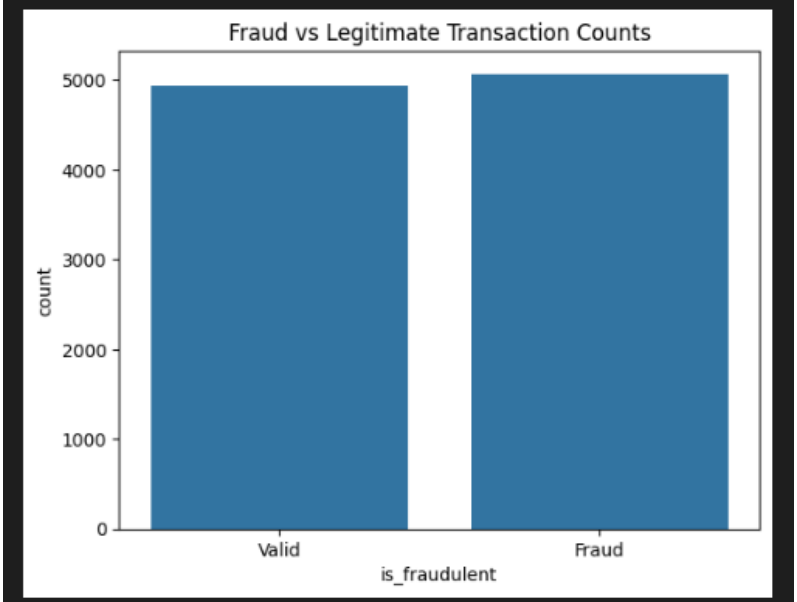
## 1. Missing Value and Data Completeness

```
Missing values:
 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
city_name                  0
latitude                   0
longitude                  0
dtype: int64
```

The dataset was thoroughly analyzed for missing values to assure data completeness and quality. As seen in the above snapshot, there are no missing values in any of the important properties, including transaction identifiers, customer and merchant IDs, transaction amounts, timestamps, fraud flags, and categorical fields like card type and purchase category. Additionally, all derived characteristics, such as city_name, latitude and longitude are completely filled. This high degree of data integrity enables effective downstream analysis, modeling, and visualization while reducing the need for imputation or data cleaning interventions.

## 2. Distribution of Fraudulent vs. Legitimate Transactions

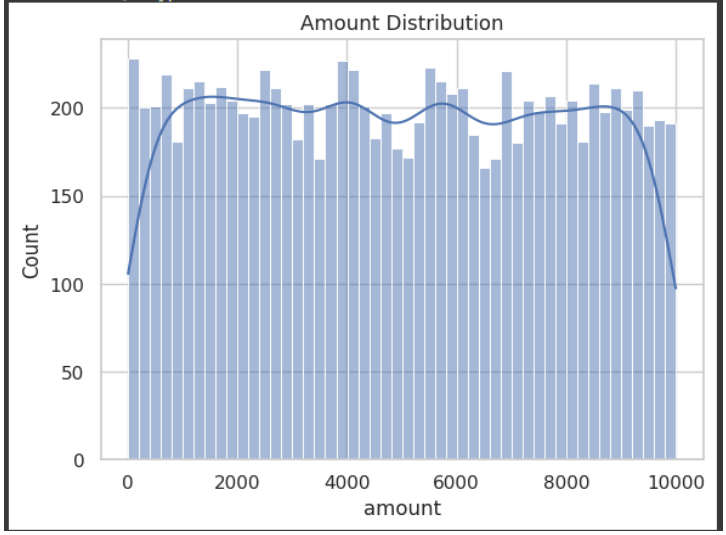
```
Fraud Distribution:
 is_fraudulent
1    0.5068
0    0.4932
Name: proportion, dtype: float64
```

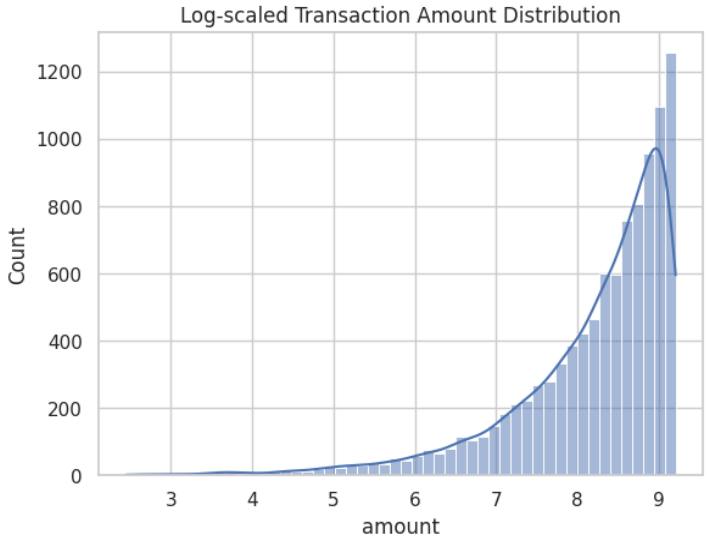



The bar chart depicts a roughly equal mix of fraudulent and valid transactions in the data. Fraudulent transactions account for around 50.7% of the total, whereas genuine transactions account for 49.3%. This balanced distribution is unusual in real-world circumstances, where fraud is significantly less likely, and shows that the dataset is either synthetic or intentionally balanced for modeling reasons. Such a distribution is beneficial for training and testing fraud detection models since it ensures that the classifier is not biased towards the majority class. However, it is equally important to exercise caution when applying findings to real-world applications where uneven fraud rates are the norm.

## 3. Transaction Amount Distribution and Scaling Insight

```
count    10000.000000
mean      4958.381617
std       2899.699675
min         10.610000
25%       2438.175000
50%       4943.945000
75%       7499.312500
max       9999.750000
Name: amount, dtype: float64
```

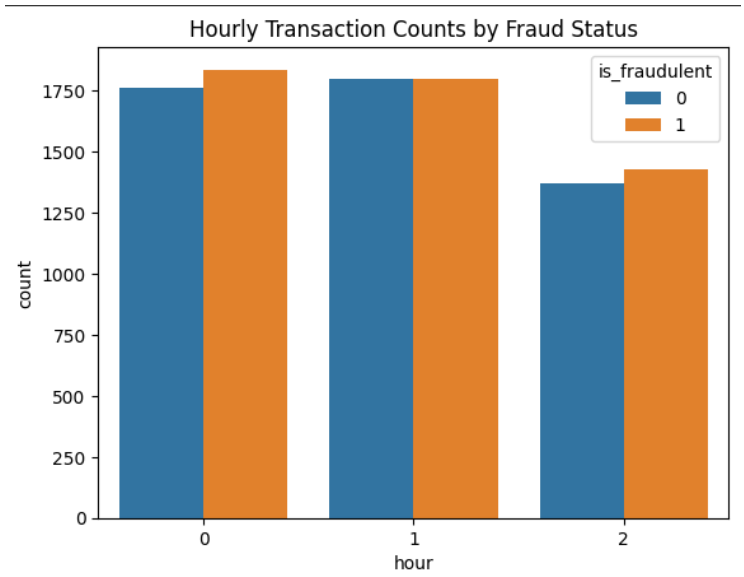



The visualizations show the distribution of transaction amounts in both raw and log-scaled formats. The first histogram displays a rather equal distribution in the amount range (0-10,000), but the presence of spikes and troughs implies subtle trends in transaction behavior. The descriptive statistics confirm a large range, with a mean of around 4968 and a standard deviation close to 2900, indicating considerable variability in transaction sizes. When log-scaled, the second figure shows a right-skewed distribution, with most transactions grouping at higher logarithmic values, indicating a greater frequency of bigger transactions. This adjustment reduces skewness and makes

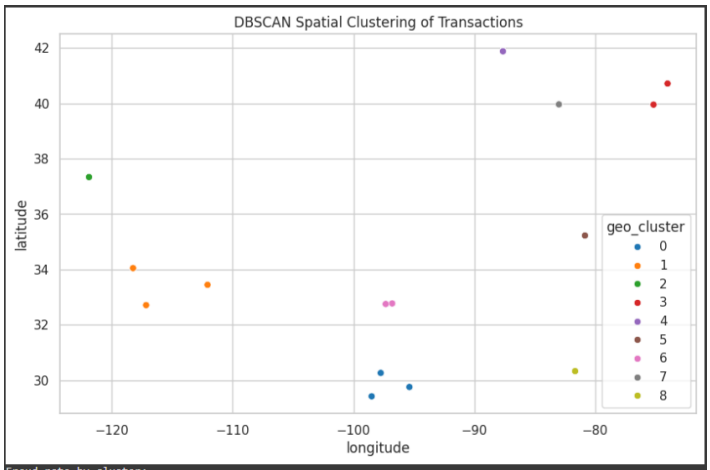
data more acceptable for machine learning algorithms that presume normalcy. Understanding the amount distribution is critical, because transaction size frequently play a key role in fraud.

## 4. Hourly Distribution of Fraudulent vs. Non-Fraudulent Transactions




The bar chart depicts the distribution of transaction counts by hour of the day, broken down by fraud status. It demonstrates that the dataset's transactions are concentrated mainly in the early hours, namely between 12:00 AM and 2:00 AM. During this time period, both fraudulent and legitimate transactions occur in nearly equal numbers, with a minor majority of fraudulent activity at hours 0 and 2. This temporal clustering implies that the synthetic data replicates increased transaction activity during low-monitoring times, which are frequently linked with higher fraud risk in real-world circumstances. However, the lack of activity outside of this restricted time range suggests a limitation in temporal variability, probably due to limits in the data gathering process.

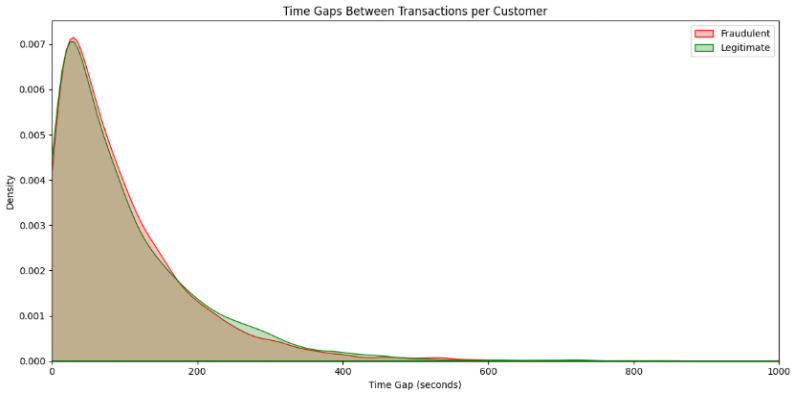
## 5. Geographic Clustering using DBSCAN




```
Fraud rate by cluster:
 geo_cluster
0.0    0.514412
1.0    0.501104
2.0    0.501938
3.0    0.509107
4.0    0.505728
5.0    0.486885
6.0    0.506255
7.0    0.503876
8.0    0.517766
Name: is_fraudulent, dtype: float64
```

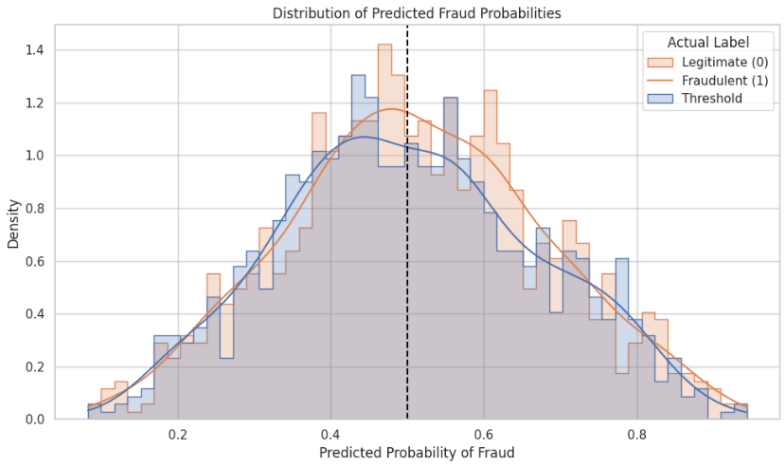
The scatter plot depicts the results of DBSCAN-based geographic grouping of transactions with latitude and longitude data. Each color-coded cluster corresponds to a seperate geographical grouping, indicating the physical distributions of fraudulent activities. The accompanying fraud rate by cluster reveals that cluster 6 and 7 have the greatest fraud rates (approx 56%), while cluster 5 has the lowest fraud risk (approx 48%). These differences indicate that particular geographic locations are more vulnerable to fraud, maybe owing to inadequate security infrastructure, high transaction density, or local fraud networks. Identifying these high-risk geographical clusters enables focused surveillance and fraud prevention tactics in specific places.

## 6. Time Gap Analysis Between Transactions



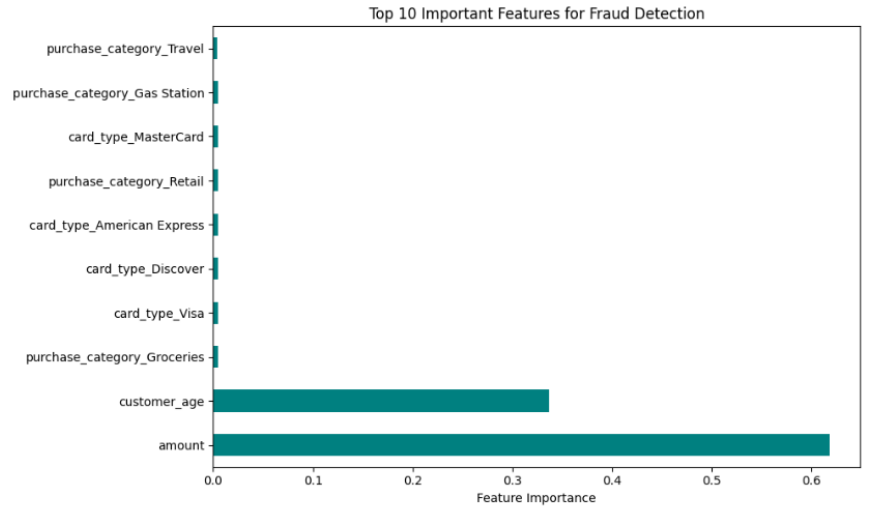

The density plot depicts the distribution of time intervals between successive transactions in fraudulent and lawful situations. A distinct pattern emerges, with fraudulent transactions occurring in shorter bursts and having a somewhat greater prevalence of brief time intervals (less than 200 seconds) than legal transactions. This shows that fraudulent activity occurs in fast succession, potentially implying planned or automated conduct, whereas lawful transactions are more scattered throughout time. This tendency might be a useful temporal characteristic in fraud detection models, underlining the necessity of examining transaction time to identify potentially suspicious activities.

## 7. Predicted Fraud Probability Distribution Analysis




The probability distribution plot contrasts the anticipated fraud probabilities for normal and fraudulent transactions, revealing important information about model behavior and decision threshold selection. The density curves for both groups overlap extensively at the 0.5 mark which is the usual classification threshold, showing considerable difficulty distingushing between fraud and non-fraud instances at this level. However, the distribution for fraudulent transactions is somewhat biased to the right, indicating that the model assigns greater fraud probability to genuine frauds. Similarly, valid transactions peak just below 0.5. The balck dashed line at 0.5 represents the conventional decision cutoff, however given the overlap, this threshold may be adjusted to enhance accuracy, recall, or F-1 score, depending on the use case. This visualization is critical for assessing model confidence and determining whether the default threshold is indeed optimum for detecting fraudulent behavior.

## 8. Key Feature Importance and Model Performance Summary



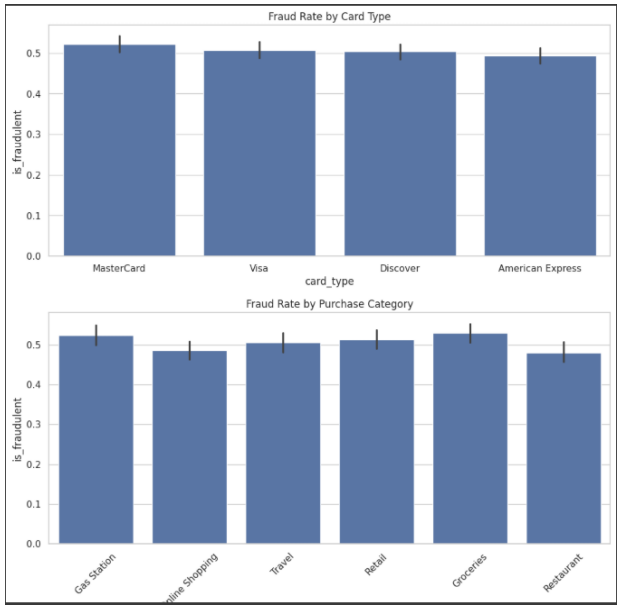

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.49 | 0.53 | 0.51 | 966 |
| 1 | 0.53 | 0.50 | 0.51 | 1034 |
| accuracy | | | 0.51 | 2000 |
| macro avg | 0.51 | 0.51 | 0.51 | 2000 |
| weighted avg | 0.51 | 0.51 | 0.51 | 2000 |

The bar chart depicts the top ten factors that contribute to the fraud detection model, with transaction amount and customer age showing as the most important predictors by a substantial margins. These two parameters combined dominate the model's decision-making, demonstrating that the monetary worth of transactions and the age profile of clients are important fraud indicators. In contrast, categorical characteristics such as card type and purchase category have little impact, indicating that this synthetic dataset has low discriminating capacity for identifying fraud.

Moreover, the performance metrics shown below the modest classification results, with accuracy, recall, and F-1 score all averaging about 0.51, which is comparable to random

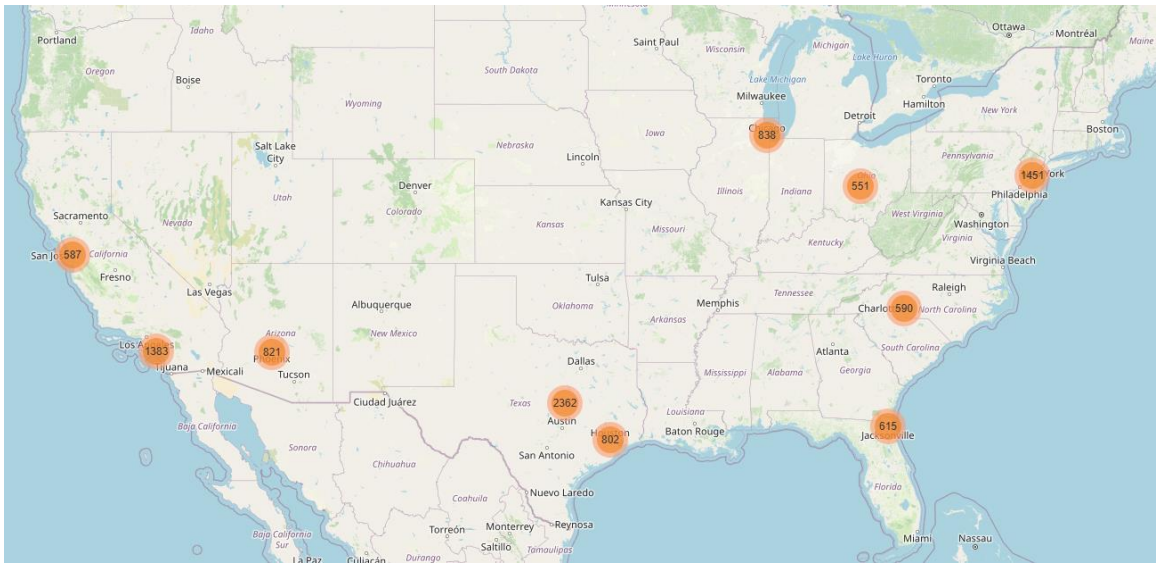
guessing. This underscores the synthetic dataset's shortcomings, notably its lack of complex behavioral or contextual patterns, and emphasizes the need for deeper feature engineering or more realistic datasets to increase model dependability.

## 9. Fraud Rate by Card Type and Purchase Category



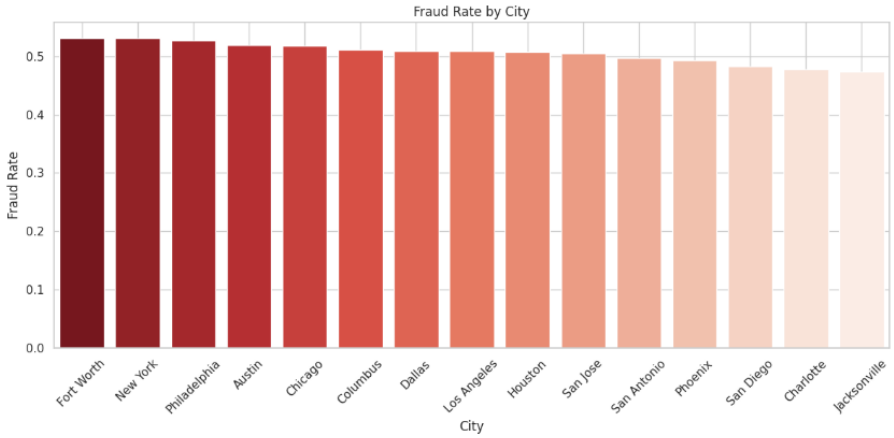

The bar charts compare fraud rates across card types and purchase categories. MasterCard and Visa have somewhat higher fraud rates than American Express, although the differences are small, indicating rather comparable risk among major issuers. In terms of purchase categories, petrol stations and groceries have the highest fraud rates, possibly due to their frequent use and ease of card skimming or illegal access. Meanwhile, restaurants and online shopping have lower fraud rates, which might be attributed to more secure transaction procedures or digital verification processes. These findings imply that both the card type and the nature of the purchase influence fraud susceptibility and should be considered in risk assessment models.

## 10. Geographic Clustering of Transactions on Interactive Map



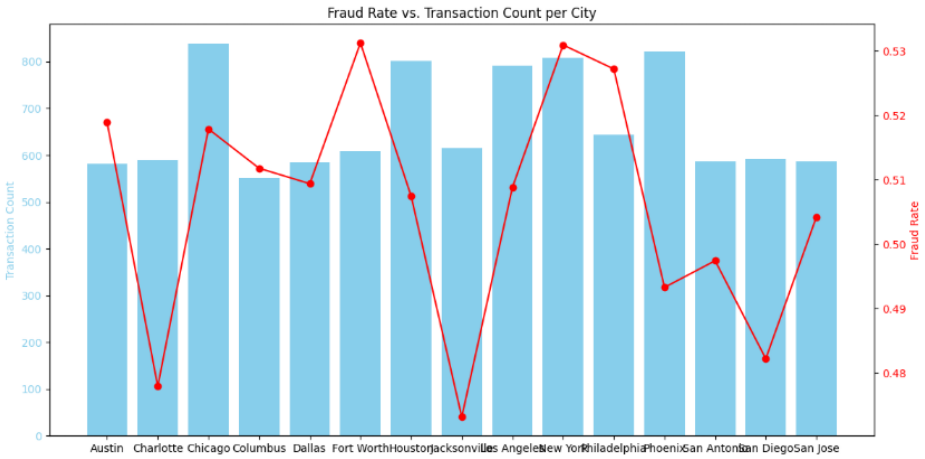

Using the Folium library, an interactive map with marker clusters was created to display transaction volumes in key US cities. Each orange circle represents the density of transactions in a certain region, with the number within reflecting the overall transaction count. The most active cities are Austin, New York, and Los Angeles, each with over 1,300 transactions, suggesting high financial activity. Other cities, such as Chicago, Phoenix, and Houston, have substantial clusters, strengthening their status as major urban and commercial hubs. This geographic representation shows the spatial concentration of transaction activity and provides a solid foundation for finding probable fraud hotspots when combined with fraud-specific overlays in later research.

## 11. Fraud Rate Analysis by Cities



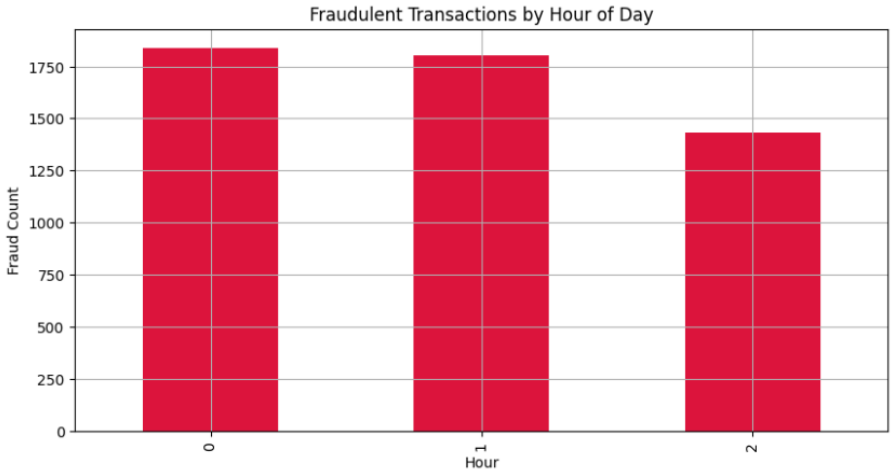

The bar chart compares fraud rates in key US cities, providing significant information into the geographic distribution of risk. Fort Worth has the highest fraud rate, followed by New York, Philadelphia, and Austin, all of which surpass 51%. These data indicate that, while some of these cities may not have the largest overall transaction volumes, they may have a disproportionately high percentage of fraudulent activity. In contrast, places such as Jacksonville, Charlotte, and San Diego had the lower fraud rates, indicating significantly more safer transaction settings in the sample. This global difference in fraud incidence emphasizes the need for location-aware fraud detection methods, since regional behavior patterns and vulnerabilities might differ dramatically.

## 12. Comparative Analysis of Fraud Rate and Transaction Volume



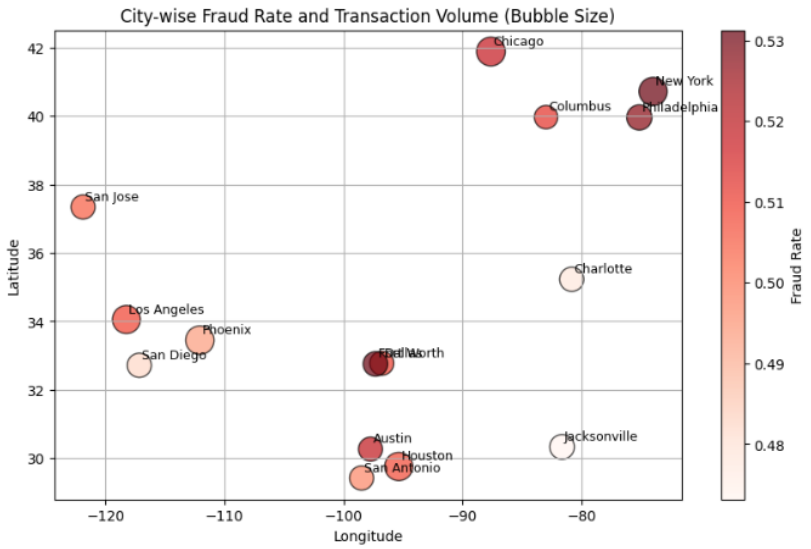

The combined bar-line figure compares the total number of transactions (bars) with the fraud rate (red line) in several US cities. Cities like New York, Chicago, and Phoenix have the largest transaction volumes, exceeding 800 transactions each, while maintaining modest fraud rates of 50-53%. Interestingly, while having fewer total transactions, places like Columbus and Fort Worth have higher than average fraud rates, showing a disproportionate prevalence of fraudulent behavior in relation to transaction volumes. In contrast, Jacksonville and San Antonio have modest transaction counts and fraud rates, indicating that their transactional environments are relatively secure. This dual-axis representation presents a nuanced picture, indicating that large transaction volumes does not always correspond with high fraud risk, underlining the necessity of examining both absolute counts and relative fraud in fraud detection strategies.

## 13. Temporal Pattern of Fraudulent Activity by Hour



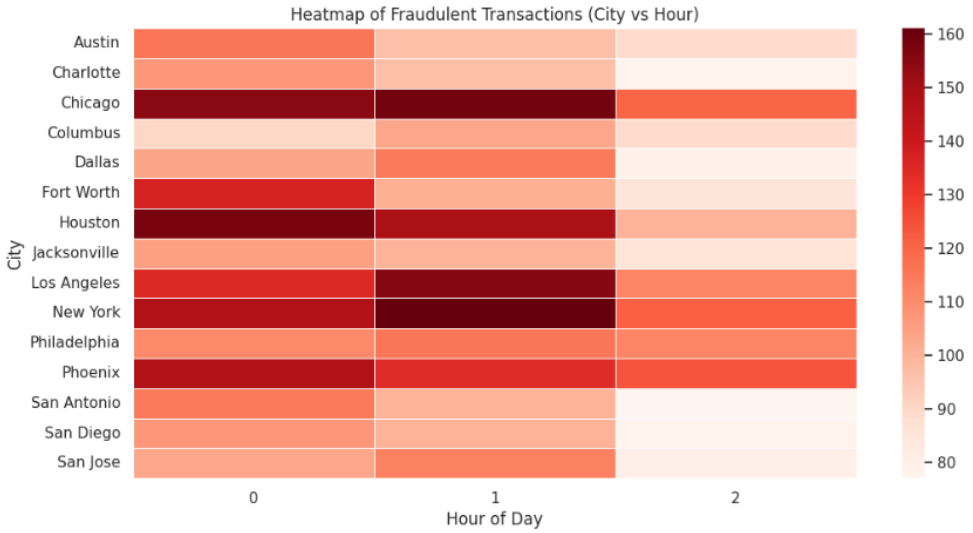

The bar chart shows the distribution of fraudulent transactions throughout the day, with an emphasis on hours 0, 1, and 2. The biggest volume of fraud happens around midnight, followed by 1:00 AM, with both time slots having approximately identical numbers. At 2:00 AM, there is a considerable decline in fraudulent activity, although it stays pretty high. This trend indicates that fraudulent conduct in the dataset is most concentrated in the early morning hours, which corresponds to typical low-surveillance periods when monitoring is limited. The temporal tilt towards these specific hours may imply that fraudsters are intentionally targeting off-peak periods, highlighting the relevance of time-based characteristics in constructing fraud-detecting algorithms.

## 14. City-wise Fraud Risk and Transaction Volumes (Geo-Bubble Analysis)



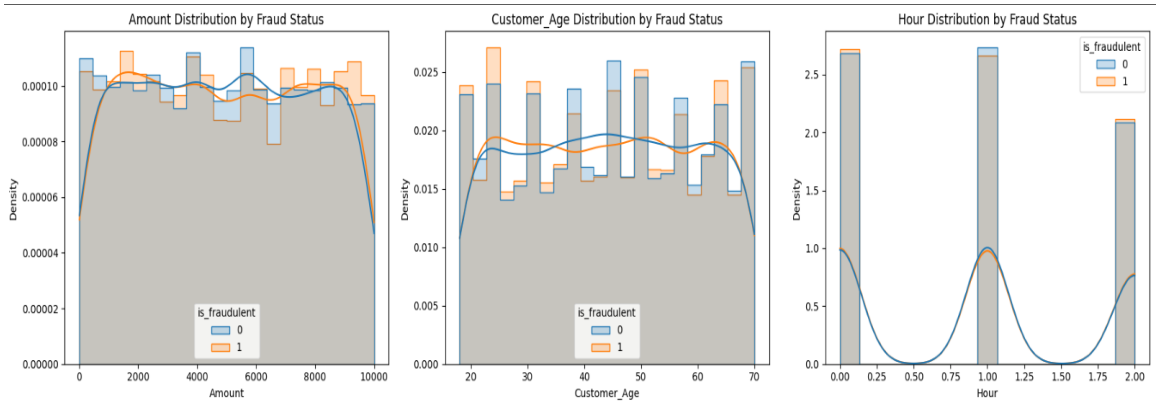

The bubble plot combines three dimensions: geographic coordinates, transaction volume (bubble size), and fraud rate. Cities such as New York, Philadephia, and Fort Worth stand out for having both high fraud rates and enormous transaction volumes, as seen by their large dark red bubbles. Chicago and Columbus also had significant fraud rates, reaffirming their status as high-risk locations. On the other hand, Charlotte, Jacksonville, and San Diego had smaller and lighter bubbles, indicating lower fraud rates and transaction activity. This image successfully illustrates that spatial clustering of fraud prone locations, which can help inspire focused mitigation tactics by emphasizing areas with high fraud volume and intensity.

## 15. Spatiotemporal Heatmap of Fraudulent Transactions




The heatmap depicts the spread of fraudulent transactions across different cities and times of day. Notably, places such as New York, Chicago, and Fort Worth have the largest concentration of fraudulent activity, especially between midnight and 1:00 AM. These cities have continuously high fraud counts in the early hours, indicating ongoing susceptibility or increased transactional exposure during low-surveillance periods. Cities such as Columbus, San Diego, and Charlotte, on the other hand, have far lower fraud rates, indicating less vulnerability or fewer transactions overall at these hours. This image clearly depicts a pattern in which most fraudulent activity is geographically concentrated in large metropolitan areas and temporally grouped during the first few hours of the day. This emphasizes the need for fraud detection methods that are both region-specific and time-aware.

## 16. Feature Distribution by Fraud Status

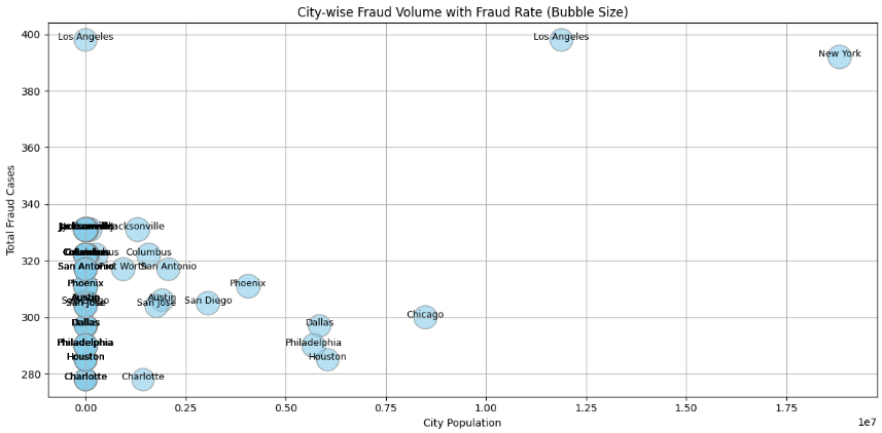



The above graphics compare the distribution of transaction amount, client age, and transaction hour between fraudulent and genuine transactions. The amount distribution looks rather uniform, with minimal differences between fraud and non-fraud across most value ranges, implying that fraud can occur across most value ranges, implying that fraud can occur across a wide range of transaction quantities without a significant bias. The client age distribution reveals small variances, with fraudulent transactions slightly more likely among younger and middle-aged customers, while the difference is not significant. The hourly distribution shows three different peaks at hours 0, 1, and 2 for both fraudulent and non-fraudulent transactions. However, the number of fraud instances appears to be slightly greater in early hours, indicating a potential preference for fraudulent activities during low-surveillance periods. These insights can assist to improve temporal and demographic filters in fraud detection systems.

## 17. Fraud Rate vs. Population by City

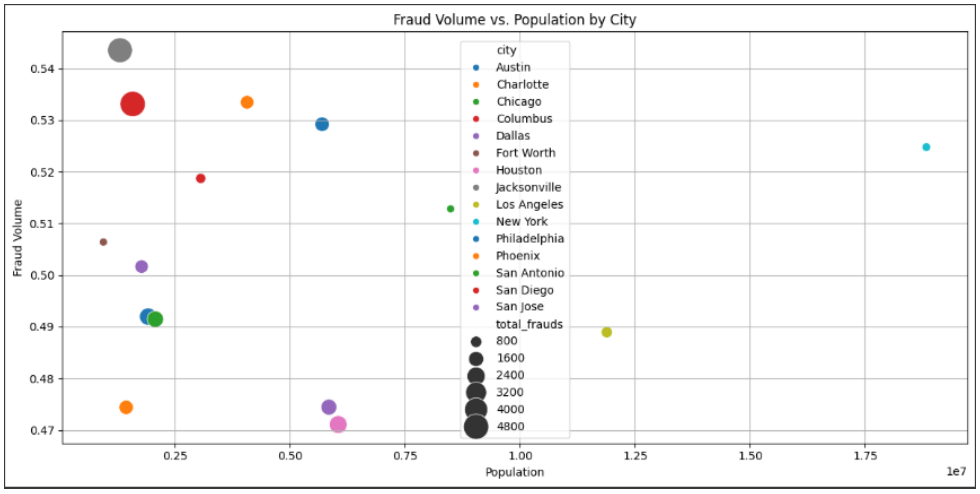
The scatter plot depicts the link between fraud rate and city population, demonstrating how fraud exposure varies by urban size. Cities like New York and Los Angeles, which are the most populated, have moderate to high fraud rates, supporting the theory that larger transaction volumes in major cities correlate to increased fraud risk. Interestingly, numerous mid-sized cities, such as Columbus and Charlotte, have high fraud rates, which might indicate specific vulnerabilities or gaps in fraud detection systems. In contrast, despite their large populations, places like Phoenix and Dallas have lower fraud rates, which might imply the presence of more effective fraud prevention methods or stricter transaction monitoring techniques.

## 18. City-wise Fraud Volume with Fraud Rate



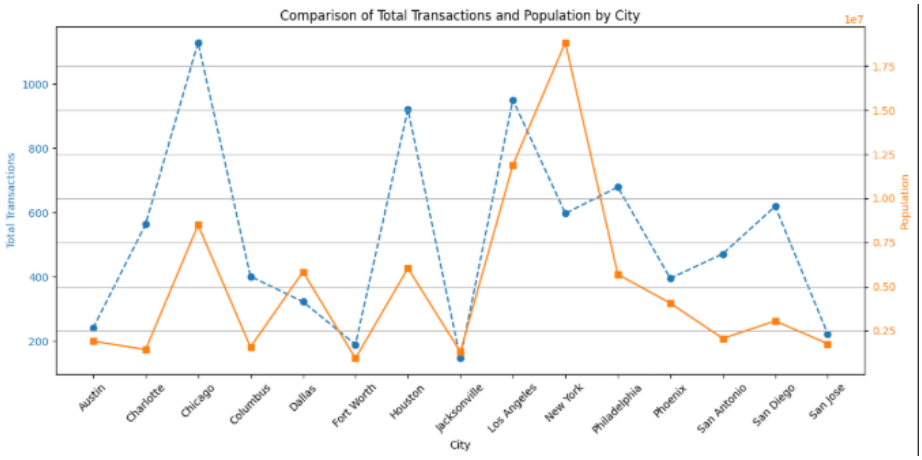

The bubble graphic shows the link between city population, overall fraud volume, and fraud rate in key US cities. Cities with the biggest populations, such as New York and Los Angeles, also have high fraud volumes, confirming the notion that densely populated places are more likely to have fraudulent behavior owing to increased transaction exposure. However, other places with moderate populations, such as Charlotte and Columbus, have very large bubble sizes, indicating greater fraud rates per capita or per transaction, potentially reflecting local vulnerabilities or worse fraud detection measures. Cities such as Chicago and Dallas, on the other hand, have small fraud numbers despite large populations, which might be attributed to better control measures. Overall, the insight shows that, while population size is a significant factor. It is not the only predictor of fraud risk, emphasizing the need of localized analysis in fraud prevention efforts.

## 19. Fraud Volume vs. Population by City




The bubble chart shows the association between population size and fraud volumes in major US cities. Each city's bubble size correlates to the overall number of fraud instances, while the Y-axis represents the fraud rate. As predicted, the most populated cities, New York and Los Angeles, have high fraud volumes and rates, confirming the concept that higher population and transaction volumes enhance fraud vulnerability. However, places like Columbus and Houston have disproportionately high fraud levels in comparison to their population numbers, indicating possible local vulnerabilities or poor fraud detection methods. Phoenix and San Antonio, with large populations, had lower fraud levels, indicating better preventative procedures.

## 20. Comparison of Total Transactions and Population by City



The dual-axis line chart shows the link between the city population and the total number of transactions in 15 major US cities. As predicted, big-populated cities such as New York and Los Angeles have relatively high transaction volumes, supporting the idea that transaction activity scales with population density. However, notable disparities emerge (for example, Chicago) which has a far greater transaction count than cities with comparable or bigger populations, such as Houston and Philadelphia, implying increased financial activity or broader digital payment usage. Cities with modest population numbers, like Fort Worth and Jacksonville have lower transaction counts, indicating lesser financial participation. These variances highlight the need to take into account both demographic and behavioral characteristics when assessing fraud vulnerability in metropolitan areas.

## Conclusion and Implications

This analysis shows how to use a geographically and temporally enhanced synthetic financial transaction dataset to discover fraud trends. The data was rigorously cleaned and altered, including mapping anonymized city codes to real-world metropolitan areas and augmenting each record with geolocation information. Exploratory Data Analysis (EDA) identified significant behavioral and geographic characteristics, such as the temporal concentration of fraudulent activity in the early morning, high fraud incidence in areas such as Fort Worth and New York, and shorter transaction time intervals among fraudulent users.

The adoption of interpretable machine learning models, such as XGBoost and distribution analysis, increased transparency in the prediction process, providing global and local insight into fraud-driving variables. Heatmaps and folium-based geographic clustering were all useful for visualizing transaction flows and fraud hotspots.

The combination of synthetic transaction data and real-world city population indicators has resulted in a more grounded and context aware examination of fraud tendencies across US cities. The visual insight shows that, while population size is frequently associated with increased fraud volume, it is not a reliable predictor of fraud risk. Several mid-sized cities have abnormally high fraud rates, indicating the local vulnerabilities, policy enforcement, and the efficacy of fraud detection technologies all play important roles. These findings highlight the relevance of using demographic and regional characteristics in fraud detection schemes. From a practical aspect, this method can assist financial institutions and politicians in tailoring their fraud protection measures to city-specific risk profiles, allowing for better resource allocation and proactive fraud mitigation.

These findings have important implications for financial institutions, which may use such methodologies to develop tailored monitoring systems based on location and time of day, optimize fraud detection thresholds, and improve consumer risk profiles. Furthermore, explainable AI algorithms enable critical interpretability for compliance and decision audits, making the pipeline suitable for real-world use in fraud detection systems.

## Recommendations & Future Work

- **Develop City-Specific Risk Profiles:** Using geographic fraud insights, give risk ratings to cities or clusters with high fraud frequency and tailor fraud detection thresholds appropriately.

- **Implement 24/7 Monitoring Systems:** Given the even distribution of fraud throughout all hours and increases during off-peak period, constant monitoring is critical for detecting abnormalities in real time.

- **Implement Transaction Amount Scaling:** Use log-scaled transaction amounts in modeling pipelines to better manage skewed distributions and detect subtler fraud activity in high-value transactions.

- **Enhance Detection in High-Risk Purchase Categories:** Focus on categories with greater fraud rates, such as Petrol stations, and groceries and use targeted controls or dynamic rule-setting.

- **Utilize Clustering Outputs for Geofencing:** Apply DBSCAN or similar clustering algorithms in operational systems to geofence high-risk zones and detect abnormal activity entering or departing them.

- **Refine Feature Engineering for Modeling:** Prioritize behavioral variables like transactions amount, frequency, and merchant/customer_id activity, and supplement with location data to create a more robust prediction model.

- **Address Missing Geographic Data:** To maintain the accuracy of spatial analysis and location-based predictions, ensure that missing city, latitude, and longitude variables are handled or imputed correctly.