Write an analysis report on performing exploratory data analysis (EDA) using Python in the context of building a fraud detection system for the financial industry.

Ans:-

# 1. Objective

The primary goal of this EDA is to understand the patterns, trends, anomalies, and key indicators associated with fraudulent credit card transactions. This step serves as the foundation for building an effective fraud detection system for the financial industry using machine learning and analytics.

# 2. Dataset Overview

• Rows: 389,002  
• Columns: 23  
The dataset contains transactional and customer-level data including timestamps, amounts, locations, job titles, merchant categories, and a binary is\_fraud indicator.

# 3. Data Quality Assessment

## a. Missing Values

No missing values were detected. Data is complete and ready for modeling.

## b. Duplicate Transactions

Not directly checked but may require deduplication during preprocessing (based on trans\_num or timestamp).

## c. Data Types

Mixed types: datetime, categorical, numerical. trans\_date\_trans\_time was converted to datetime for time series analysis.

# 4. Univariate Analysis

## a. Transaction Amount (amt)

• Highly skewed distribution  
• Mean: $70.44  
• Max: $27,390  
• Outliers observed, especially in fraudulent transactions

## b. City Population (city\_pop)

• Mean population ≈ 88,681  
• Skewed due to high-population cities like New York and Los Angeles

## c. Target Variable (is\_fraud)

• Highly imbalanced:  
 - Fraudulent transactions: ~0.58% (2,251 cases)  
 - Non-fraudulent transactions: ~99.42%

# 5. Bivariate and Multivariate Analysis

## a. Fraud vs. Transaction Amount

Fraudulent transactions tend to have higher amounts. Visualized using box plots and histograms.

## b. Fraud vs. Category

Categories with higher fraud rates include: misc\_net, shopping\_net, and personal\_care.

## c. Correlation Matrix

amt vs is\_fraud: +0.21  
lat vs merch\_lat: +0.99  
Most variables are weakly correlated, but location-based features may be valuable for fraud detection.

# 6. Temporal Analysis

## a. Monthly Transaction Trends

Seasonal or cyclical spikes in spending. Useful for detecting time-based fraud patterns (e.g., around holidays).

# 7. Outliers & Anomalies

Significant outliers in amt and city\_pop. These could either represent genuine high-value transactions or potential fraud cases. Requires careful treatment during modeling (e.g., capping, normalization).

# 8. Insights & Recommendations

• Class Imbalance: Requires resampling techniques for ML (e.g., SMOTE)  
• High-Value Transactions: Should be monitored closely; fraud more likely at higher amounts  
• Geographical Info: High potential — latitude/longitude strongly linked to merchants  
• Time-based Features: Temporal aggregation (hour, day, week) can help reveal fraud patterns  
• Feature Engineering: Consider building features like transaction frequency, avg\_amt\_last\_7d, etc.

# 9. Next Steps for Model Development

1. Feature Engineering  
2. Imbalanced Classification  
3. Model Explainability (e.g., SHAP or LIME)

# 10. Conclusion

The EDA process revealed crucial insights into transaction behaviors, fraud patterns, and variable relationships. These insights will guide the feature selection and modeling strategies used in building an effective fraud detection system. The dataset is rich, clean, and ready for modeling.

# 11. Unusual or Unexpected Values in the Dataset

Several outliers and edge cases were found. Examples include:  
- Transactions with very high amounts at categories like 'coffee\_shop'  
- Lat/Long values beyond realistic geographical bounds  
- Transaction times that could potentially be outside business hours  
These should be flagged for further investigation or used as signals for fraud detection.

# 12. Data Entry Errors or Inconsistencies

Some inconsistencies in categorical variables such as capitalization or spelling variations (e.g., 'Male' vs 'M'). Timestamp formatting is consistent but should be checked for timezone relevance. Duplicate transaction IDs or repeated unix\_time values were not detected but are potential issues worth verifying.

# 13. Variable Distribution by Groups or Segments

Boxplots and group-level summaries revealed that:  
- Transaction amounts are higher for fraud cases.  
- Certain job roles (e.g., executives) and merchant categories (e.g., 'shopping\_net') show distinct patterns.  
- Gender, job, and category all show measurable variations in city population and transaction volume.

# 14. Top Factors Influencing Fraud (Target Variable)

Using Random Forest for feature importance, the following variables emerged as most influential:  
- Transaction Amount (amt)  
- Merchant Category (category)  
- City Population (city\_pop)  
- Job title and location coordinates (lat, long)  
Correlation and tree-based models both suggest that transaction amount and category are top indicators of fraud.