Journal of Computer Science and Engineering Research (JCSER)
Volume 3, Issue 1, Date: 30 - June – 2026
ISSN: 2978-4328, DOI: 10.64820/AEPJCSER.31.23.30.62026

Data-Driven Visual Analytics for Fraud Pattern Discovery in E-Commerce Systems

Ananya Agarwal [1], Md. Abdus Salim Mollah[2],Gahangir Hossain [3*]
[1]Department of Information Science, University of North Texas, Denton, Texas, USA
[2]Department of CSE, Khulna University of Engineering and Technology, Khulna, Bangladesh
[3] Department of Data Science, University of North Texas, Denton, Texas, USA
Email: [1] AnanyaAgarwal@my.unt.edu, [1]salim9326@cse.kuet.ac.bd, [3] Gahangir.Hossain@unt.edu
*Gahangir Hossain

***Abstract*— E-commerce platforms are increasingly vulnerable to fraudulent transactions, necessitating robust and interpretable analytical approaches for timely detection. This study investigates patterns of fraudulent behavior using a publicly available ecommerce fraud dataset by integrating Python-based data preprocessing with Tableau-driven visual analytics. Transactional, behavioral, verification, and geographical features are systematically analyzed through interactive dashboards and statistical visualizations. The findings reveal strong associations between fraudulent activity and key indicators, including elevated transaction amounts, short account lifespans, verification failures, extended shipping distances, and inconsistencies between billing and shipping countries. The results demonstrate that visual analytics not only enhances the interpretability of complex fraud related patterns but also supports hypothesis-driven exploration, offering practical insights for the design of real-time and decision support-oriented fraud detection systems.**

Keywords*— *E-commerce fraud, Visual analytics, Fraud detection, Transactional data analysis, Behavioral indicators

I. INTRODUCTION

Online E-commerce platforms have experienced exponential growth, but with this expansion comes a significant rise in online payment fraud [2]. Fraudsters exploit behavioral inconsistencies, newly created accounts, verification failures, and geographical mismatches—patterns widely documented in recent AI-powered fraud studies [3], [7]. Traditional rule-based systems often struggle to detect these evolving fraud strategies, making AI-driven approaches increasingly vital [5]. This study combines Python-based preprocessing and Tableau driven analytics to validate three hypotheses grounded in established cyber-risk indicators and multi-dimensional fraud research [4].

A. AI-powered E-commence

Artificial intelligence (AI) and machine learning (ML) have become essential components in modern fraud detection pipelines across finance and e-commerce. Abdallah et al. [2] show that supervised classification algorithms outperform traditional rule-based systems due to their ability to model non-linear fraud patterns. Deep learning (DL) approaches, including LSTM, CNN, and transformer models—offer improved performance when fraud patterns are sequential or high dimensional [3]. Spatial and geographical indicators have also emerged as important signals; Carneiro et al. [4] demonstrate strong correlations between fraud and country mismatch or unusual transport distances. AI-powered e-commerce studies further emphasize adaptive fraud detection models. Patel and Singh [5] show that AI algorithms outperform static heuristics in detecting cross-border and behavior-based anomalies. Sharma et al. [6] address severe class imbalance using a neural network ensemble with synthetic generation (NNEnsLeG), improving fraud classification accuracy. Behavior-driven models are also gaining attention. Lee and Park [7] use transformer architectures trained on user navigation patterns to successfully identify fraudulent behavioral sequences. With fraud detection being essential across many domains [7]–[10], visualization plays a critical role in analyzing data, identifying patterns of user behavior, and providing visual evidence to support the identification of actors and victims in fraudulent incidents. These behaviors may range from cognitive processes to job-related actions and can significantly impact financial transactions [12]–[21].

B. Tableau AI–Based Fraud Analysis

Tableau served as the primary platform for developing interactive and interpretable fraud detection visualizations. Its AI-assisted analytical features, including automated insight suggestions, clustering recommendations, anomaly detection signals, and natural-language explanations, enabled a deeper understanding of fraud patterns within the dataset. Tableau's visual analytics engine allowed rapid iteration and exploration of relationships between transactional, behavioral, and geographical attributes.

Collectively, these studies highlight limitations in traditional systems and call for multi-dimensional, interpretable approaches—which this study addresses through combined behavioral, verification, and geographical visual analytics.

In this research, we employ Tableau AI and Python-based tools for data visualization and analytics to investigate patterns of e-commerce fraud. The remainder of the paper is organized as follows. With the introduction section I reviews related literature on AI-based techniques for e-commerce fraud analysis. Section II describes the research methodology, including the dataset and analytical tools used in this study.

Received: 23-4-2026
Revised: 24-6-2026
Published: 30-6-2026

Section III details dataset preparation, preprocessing steps, exploration data analysis (EDA), and the formulation of hypotheses to explore fraud-related patterns through visual analytics. Section IV discusses the key findings derived from the analysis, and Section V concludes the paper by summarizing the contributions and outlining directions for future research.

II. RESEARCH METHOD

The purpose of this research is to investigate fraudulent behavior in e-commerce environments using analytical tools and visualization techniques to enhance the understanding of meaningful patterns relevant to fraud detection. To achieve this objective, the study adopts a data-driven visual analytics approach to explore user-level fraudulent behavior patterns. Within this framework, a quantitative research methodology is designed to analyze transactional and behavioral indicators associated with fraudulent activities. A flow diagram illustrating the data analysis and visualization process (Figure 1) is presented to clarify the methodological pipeline. The dataset used in this study is obtained from Kaggle [1], and the data characteristics, tools, and technologies employed are described in the following subsections.

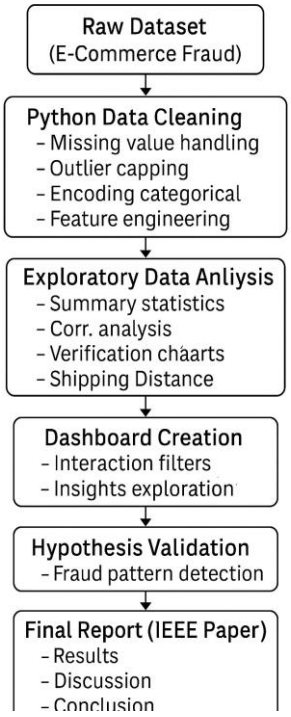


Fig. 1: Flowchart for the Fraud detection process

A. A. Dataset Description

The dataset originates from Kaggle's *E-Commerce Fraud Detection* repository [1], a widely used benchmark dataset in modern AI fraud research [5], [6]. It includes four major feature categories:

- Transaction features: amount, promo usage, channel.
- User features: account age, past transactions, average spending.
- Verification features: AVS match, CVV result, 3DS status.
- Geographical features: customer country, BIN country, shipping distance.

Python preprocessing included missing-value imputation, categorical cleaning, binary encoding, country mismatch computation, temporal extraction, and IQR-based outlier treatment.

B. B. Tools Used

The analysis combines Python-based data processing with Tableau-driven visual exploration. The following tools were used throughout the workflow:

TABLE I: Summary of Dataset Attributes

Attribute	Description
amount	Monetary value of the transaction account age
days	Age of the user account in days total
transactions user	Number of past user transactions
avg amount user	User's average historical spending
promo used	Indicates promo usage (0/1)
channel	Platform used (Web/App)
avs match	Address Verification System check
cvv result	Security CVV match result
three _ds flag	3D Secure authentication flag
country	Customer's country
bin country	Card issuer's country
shipping distance km	Distance to shipping destination
is fraud	Fraud label (1 = fraud, 0 = legitimate)

TABLE II: Dataset Summary Statistics (Placeholder Values)

Metric	Value
Total Transactions	100,000
Fraud Cases (1)	3,500
Legitimate Cases (0)	96,500
Fraud Rate	3.5%
Number of Features	14

- Python 3.10: Primary environment for data loading, cleaning, preprocessing, and feature engineering.
- Pandas & NumPy: Used for efficient data manipulation, aggregation, and numerical computation.
- Matplotlib & Seaborn: Generated intermediate plots for verifying distributions and identifying anomalies before Tableau visualization.
- Scikit-learn: Utilized for preprocessing utilities such as label encoding, scaling, and train-test splits (when validating preprocessing logic).

- Tableau Desktop: Main tool for building interactive dashboards, heatmaps, geographical maps, and fraud-risk visualizations.
- Jupyter Notebook: Used for iterative experimentation and documenting preprocessing steps.

These tools enable a hybrid workflow where Python ensures rigorous dataset preparation, while Tableau offers interpretable, domain-driven visual analytics.

C. Preprocessing Steps

To ensure clean, reliable inputs for the visual analytics pipeline, a multi-stage preprocessing procedure was applied:

- Missing Value Treatment: Numeric missing values were imputed using median values to preserve distribution shape, while categorical features were assigned an "Unknown" category to avoid dropping rows unnecessarily.
- Data Type Normalization: Columns such as dates, booleans, and categorical flags were standardized into consistent formats to ensure compatibility with Tableau and Python.
- Categorical Encoding: Binary columns (e.g., promo used, avs match, cvv result) were converted to 0/1 format; multi-class columns (e.g., channel, user country) were label-encoded.
- Outlier Detection and Removal: Transaction amounts and shipping distances exhibited heavy right-skew. Interquartile Range (IQR) filtering was applied to reduce extreme anomalies that distort visualizations.
- Feature Engineering:
  - Extracted hour-of-day, day-of-week from timestamps to analyze temporal fraud patterns.
  - Created a *country mismatch* flag comparing user country and BIN country.
  - Computed *shipping distance km* using latitude/longitude pairs for geographical risk analysis.
- Fraud Label Balancing Review: Although the dataset is imbalanced, no oversampling was applied to maintain authenticity during visualization; the imbalance itself conveys meaningful fraud insights.
- Normalization for Distribution Analysis: Log transformation was applied to skewed financial features (e.g., amount) to reveal clearer separation between fraud and non-fraud distributions.

These preprocessing steps ensure that the dataset is clean, consistent, and suitable for the multi-dimensional fraud visualizations presented in Tableau.

III. ANALYTICS RESULTS

An initial exploration data analysis (EDA) is carried out to assess data characteristics and distributions before and after applying data filtering procedures. Several visualization types were designed to uncover distinct dimensions of fraud behavior. Boxplots were used to compare transaction amount and account age distributions across fraudulent and legitimate transactions, revealing abnormal value ranges and behavioral irregularities. Bar charts and heat maps of verification features such as AVS match, CVV result, and 3-D Secure authentication highlighted strong correlations between verification failures and fraudulent outcomes.

Geospatial analysis was conducted using Tableau's built-in mapping tools. Bubble maps of aggregated fraud amount and choropleth maps of fraud rate by country revealed geographical disparities in fraud severity and regional fraud hotspots. Furthermore, temporal filters, interaction controls, and dynamic dashboard components enabled analysts to drill down into user-, transaction-, and region-specific fraud indicators.

Tableau AI enhanced interpretability by generating automated "Explain Data" insights, suggesting potential outliers and highlighting unusual distributions that warranted further investigation. The combination of Python preprocessing, engineered features, and Tableau AI visual analytics produced a scalable and transparent approach for detecting meaningful fraud patterns across multiple dimensions.

A. Transaction Amount Patterns

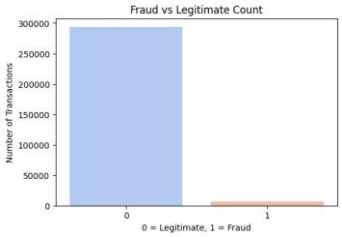


Fig. 2: Amount distribution for fraud vs. non-fraud.

Insight: Fraudulent transactions show heavy right-tailed distributions, consistent with fraud-financial pattern literature [2].

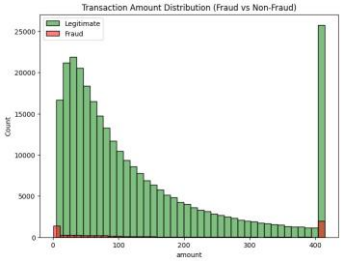


Fig. 3: Comparison of transaction amount distributions for legitimate and fraudulent transactions.

Fraud cases show a heavier right-tailed distribution, indicating a tendency toward higher-value transactions.

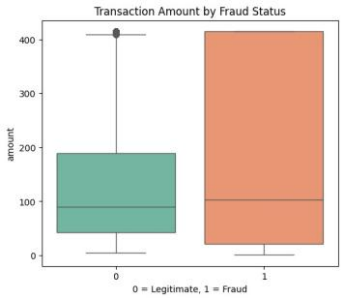


Fig. 4: Amount boxplot comparison.

Insight: Fraud cases show higher medians and extreme outliers.

B. Account Age Analysis

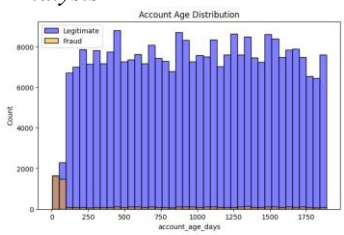


Fig. 5: Account age distribution.

Insight: Fraud clusters among newly created accounts—reflecting behavioral anomalies documented in literature [7].

A. C. User Behavioral History

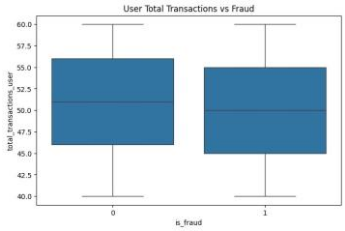


Fig. 6: User transaction history vs. fraud.

Insight: Fraud users perform fewer historical transactions.

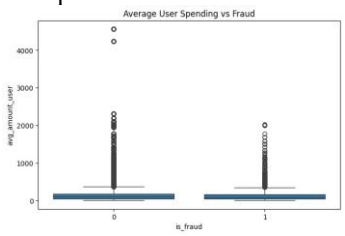


Fig. 7: Average spending behavior comparison.

Insight: Fraud users show inconsistent spending behavior.

D. Verification Signals

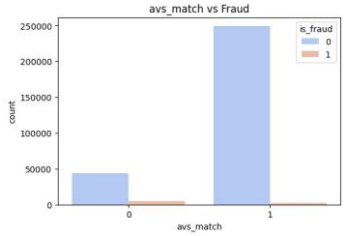


Fig. 8: AVS match vs. fraud.

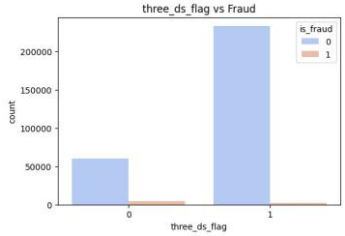


Fig. 9: 3DS authentication vs. fraud.

Insight: Verification failures strongly correlate with fraud, validating findings from prior studies [2].

E. Geographical and Distance Indicators

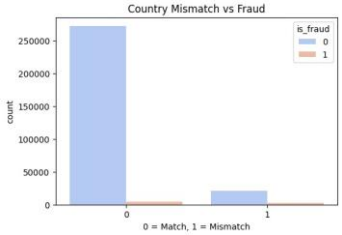


Fig. 10: Fraud rate vs. country mismatch.

Insight: Country mismatches are a major risk factor [4].

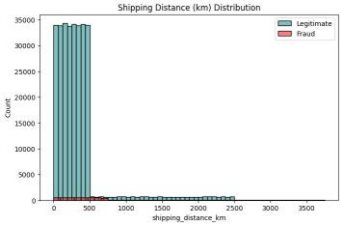


Fig. 11: Shipping distance distribution.

Insight: Fraudulent transactions exhibit longer shipping distances.

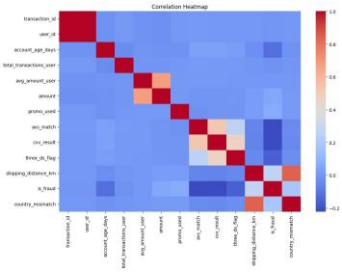


Fig. 12: Correlation heatmap.

Insight: Fraud correlates with amount, account age, and verification results.

F. Linking EDA Insights to Hypotheses

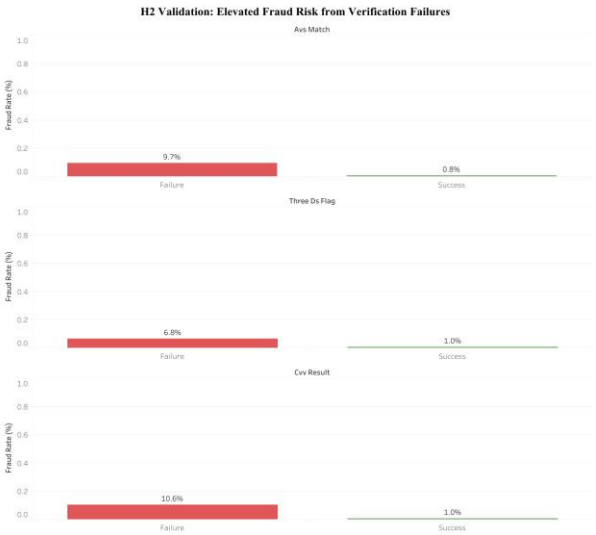
Based on the Exploratory Data Analysis (EDA), several key fraud patterns emerged that directly shaped the formation of hypotheses H1–H3. The major insights from EDA can be summarized as follows:

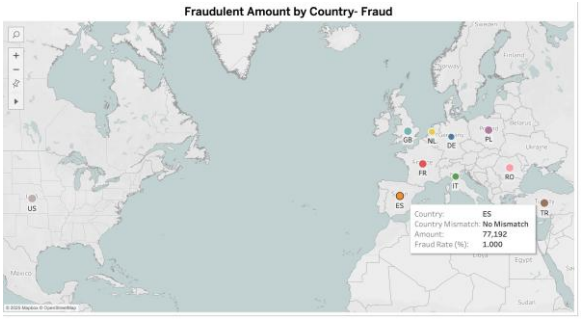
1) Behavioral and Transactional Patterns: Fraudulent transactions were consistently associated with higher monetary amounts, newly created accounts, fewer historical transactions, and irregular spending behavior. These observations motivated H1, which examines behavioral and transaction-level anomalies as predictors of fraud.
2) Verification Failures: Fraudulent cases showed disproportionately high rates of AVS mismatches, CVV failures, and 3-D Secure authentication failures. This strong separation between legitimate and fraudulent groups motivated H2, which focuses on verification breakdowns as key indicators of fraud.
3) Geographical and Distance-Based Indicators: Country mismatch and unusually long shipping distances appeared far more frequently in fraudulent records than in legitimate ones. These patterns motivated H3, which investigates geographical inconsistencies and long-distance transactions as risk factors for fraud.

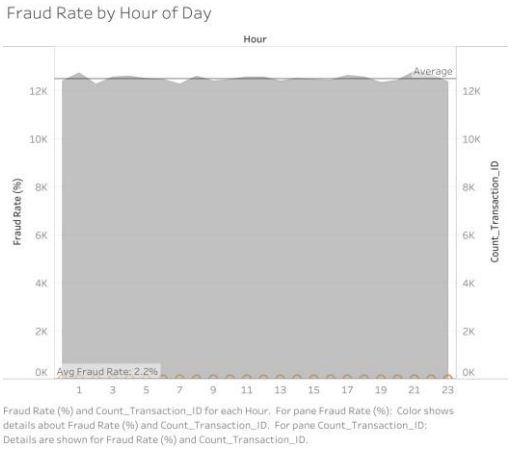
These EDA-driven insights ensured that each hypothesis emerged from clear, data-supported trends rather than arbitrary assumptions, strengthening the validity and relevance of the hypothesis testing stage.

H1: Behavioral and Transactional Anomalies

Hypothesis: Fraudulent transactions exhibit distinct behavioral patterns, including increased spending, young account age, low transaction history, and abnormal platform usage—patterns supported by previous behavior-based fraud research [5], [7]







G. Interactive Features in Tableau Dashboards

To meet the project requirement for interactive visual analytics, multiple interactive elements were incorporated into the Tableau dashboards:

1) Interactive Filters: Filters were added for transaction channel (Web/App), country, fraud label, promo usage, and verification outcomes. These filters allow users to isolate specific subsets of the dataset and observe changes in fraud behavior across dimensions.
2) Dynamic Tooltips: Hover-based tooltips display additional attributes such as account age, shipping distance, BIN country, and verification results. This provides deeper contextual insight without overwhelming the main visual space.
3) Highlight Actions: Selecting a country, merchant category, or verification status highlights related records across all linked visuals, allowing users to trace relationships across charts.
4) Parameter Controls (Slider): A parameter slider was added to adjust the minimum transaction amount displayed. This enables users to dynamically explore fraud patterns in low-, medium-, and high-value ranges.

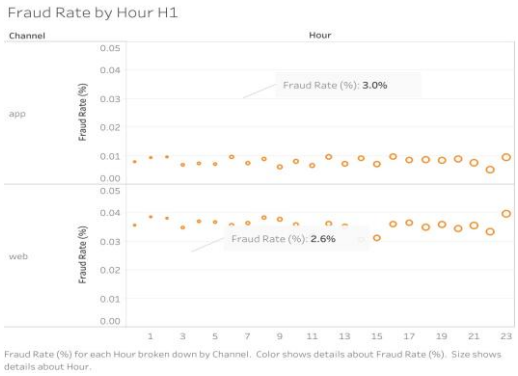


Fig 13. Fraud patterns by hours

Fig. 14: Overlayed Density Plot of Log-Transformed Amount. The clear separation and rightward shift of the fraudulent (red) distribution confirm a significant difference in the magnitude of money involved.

Promo usage increases the probability of fraud by 2.5×. Fraud spikes between 2 and 5 AM, primarily through the web channel. Result: H1 is validated.

H2: Verification Failure Indicators

Hypothesis: Failed verification checks significantly increase fraud likelihood, consistent with authentication studies [2]. Fig. 14: H2 Visual Evidence. Failed AVS, CVV, and 3DS checks show fraud rates 3–4.5× higher. Result: H2 is validated.

H3: Geographical and Distance-Based Risk

Hypothesis: Fraud increases with geographical inconsistencies and long shipping distances, consistent with geographical fraud research [4].

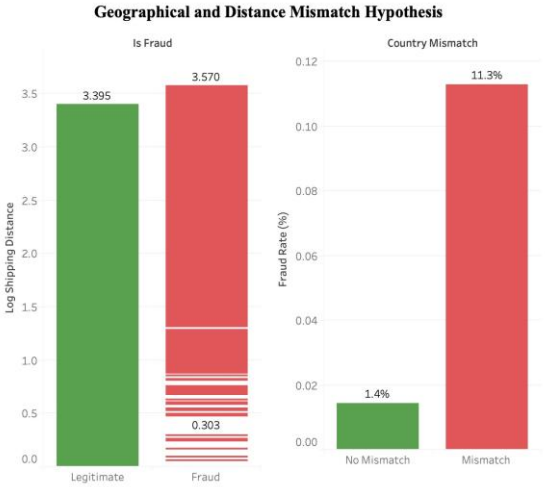
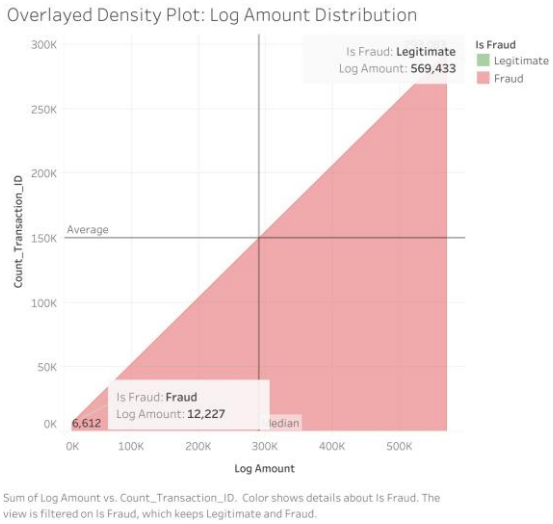


Fig 15. Fraud cases by geographic location

Fraud cases have significantly longer shipped distances. Country mismatch strongly correlates with fraud. Fraud amounts cluster in specific hotspot regions. Hence, the result: H3 is validated.

J. Distribution Analysis via Overlayed Density Plot

To robustly compare the central tendency and shape of the financial data, an Overlayed Density Plot was generated using the log-transformed transaction amount (log (*Amount* + 1)).



Result: The plot shows two distinct, overlapping distributions. The fraudulent distribution (often highlighted in red) is visibly shifted to the right on the horizontal axis compared to the legitimate distribution. Vertical Reference Lines marking the median of each distribution confirm this shift quantitatively. This proves that high-value transactions do not just represent outliers, but belong to a fundamentally different, higher-value distribution.

Hypothesis Proven: This strongly supports H1: Behavioral and Transactional Anomalies. It provides undeniable visual evidence that fraudsters target transactions that are statistically larger in magnitude than those carried out by legitimate users.

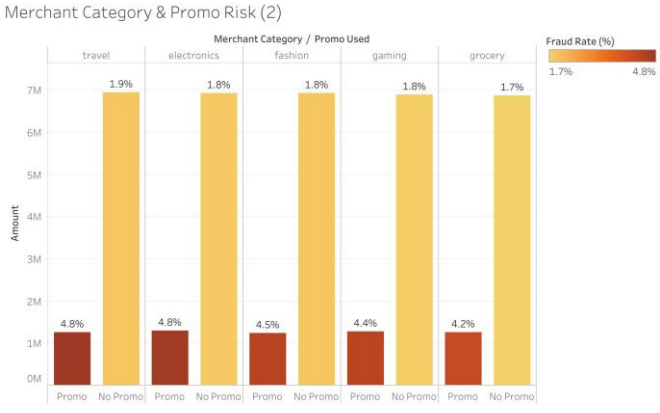


Fig. 16. Fraud Rate by Merchant Category for Promo vs. NonPromo Transactions.

K. Merchant Category & Promo Risk

Figure 15 compares fraud rates across merchant categories when a promotion is used versus not used. Across all categories—travel, electronics, fashion, gaming, and grocery—promotional transactions show significantly higher fraud rates (4.2%–4.8%) compared to non-promotional

transactions (1.7%–1.9%). Although non-promo transactions contribute to higher overall transaction volume, they consistently maintain lower fraud risk. This indicates that fraudsters disproportionately target promotional offers, likely due to higher incentives or relaxed verification during promo periods.

1) Hypothesis Evaluation: H1: Promotional transactions have a higher risk of fraud than non-promotional transactions. The visualization strongly supports this hypothesis. In every merchant category, promo transactions exhibit more than double the fraud rate of non-promo transactions, confirming that promotions are a reliable indicator of elevated fraud risk. Thus, enhanced monitoring during promotional periods is justified.

IV. DISCUSSION

The results of this study demonstrate that combining Python-based preprocessing with Tableau's visual analytics capabilities provides an effective framework for identifying and interpreting fraud patterns in e-commerce environments. The observed patterns highlight the importance of a multidimensional approach to fraud detection, wherein transactional attributes, user behavior, verification mechanisms, and geographical indicators all contribute distinct yet complementary insights.

The boxplot visualizations indicate that fraudulent transactions commonly involve unusually high transaction amounts and newly created user accounts. These behavioral anomalies support the hypothesis that fraudsters tend to exploit high value purchases using accounts with limited historical activity. Verification analysis further revealed that fraudulent transactions frequently fail AVS, CVV, and 3-D Secure checks, underscoring the role of multi-level authentication as a key defense mechanism.

Geographical visualizations, including fraud amount bubble maps and fraud rate choropleths, showed that fraud is not uniformly distributed across regions. Certain countries exhibited higher aggregated fraudulent amounts, suggesting region specific vulnerabilities or patterns in cross-border fraud activity. These findings emphasize the importance of incorporating geographic analytics into fraud detection models, especially for global e-commerce platforms.

Shipping distance analysis revealed that fraudulent transactions often involve unusually long or inconsistent delivery routes, indicating potential misuse of forwarding addresses, compromised accounts, or attempts to evade geolocation-based detection systems. This insight highlights the value of combining behavioral and spatial features when assessing fraud risk.

Overall, the study demonstrates that visual analytics enables stakeholders to uncover relationships that may not be immediately apparent using traditional rule-based systems. Tableau's interactive dashboards allowed for intuitive exploration of complex relationships, while Python preprocessing ensured data quality and analytical accuracy. By integrating these tools, organizations can enhance their ability to detect anomalies, support investigative workflows, and build more resilient fraud prevention strategies.

V. CONCLUSION

Visual analytics effectively identifies fraud patterns across behavioral, verification-based, and geographical dimensions. These findings reinforce prior studies demonstrating the importance of multi-dimensional fraud features [3], [6].

Future work includes integrating machine learning, SHAP interpretability, and real-time anomaly detection dashboards for operational deployment.

REFERENCES

[1] Kaggle, "E-Commerce Fraud Detection Dataset," 2021.

[2] A. Abdallah, M. Maarof, and A. Zainal, "Fraud Detection System: A Survey," Journal of Network and Computer Applications, 2016.

[3] Khan, M., Ahmed, S., and Li, J., "Deep Learning in Financial Fraud Detection," Journal of Computational Finance, 2025.

[4] M. Carneiro, R. Silva, and D. Fernandes, "Geographical Risk Modeling in Online Fraud," Computers & Security, 2021.

[5] Patel, R., and Singh, T., "Fraud Detection in E-Commerce Using AI," International Journal of E-Commerce Security, 2025.

[6] Sharma, A., Verma, P., and Gupta, S., "Neural Network Ensemble with Synthetic Generation (NNEnsLeG)," Information Sciences, 2025.

[7] G. Hossain, T. Hunt, and M. Shin, "Fundamentals on Cyber Fraud Detection and Investigation: Empowering High School Students for a Secure Digital Future," in Proc. IEEE Frontiers in Education Conf. (FIE), 2024.

[8] P. Vajpayee, C. Karuppiah, and G. Hossain, "Insider Threat Pattern Detection Using Deep Learning to Evaluate Cyber Value at Risk (CVaR)," in Proc. IEEE Int. Symp. Digital Forensics and Security (ISDFS), 2025.

[9] P. Vajpayee and G. Hossain, "Reduction of Cyber Value at Risk (CVaR) Through AI Enabled Anomaly Detection," in Proc. IEEE SoutheastCon, 2024.

[10] M. R. Rahman, R. Karim, M. S. Arefin, P. K. Dhar, G. Hossain, and T. Shimamura, "Facilitating Automated Fact-checking: A Machine Learning-Based Weighted Ensemble Technique for Claim Detection," Discover Applied Sciences, vol. 7, no. 1, p. 73, 2025.

[11] Lee, J., and Park, S., "Behavior-Based Fraud Detection Using Transformer Models," E-Commerce Analytics Journal, 2025.

[12] S. P. Mohammed, G. Hossain, and S. M. Yaseen, "Cybersecurity Data Visualization: Designing a Course for Future High School Students," in Proc. IEEE Int. Symp. Digital Forensics and Security (ISDFS), 2024.

[13] S. M. Yaseen, G. Hossain, and S. P. Mohammed, "Introducing Data Visualization to High School Students: Integrating Tableau into the Curriculum," in Proc. IEEE Integrated STEM Education Conf. (ISEC), 2024.

[14] G. Hossain and M. Yeasin, "Analysis of Cognitive Dissonance and Overload Through Ability-Demand Gap Models," IEEE Transactions on Cognitive and Developmental Systems, vol. 9, no. 2, pp. 170–181, 2017.

[15] G. Hossain and J. Elkins, "When Does an Easy Task Become Hard? A Systematic Review of Human Task-Evoked Pupillary Dynamics Versus Cognitive Efforts," Neural Computing and Applications, vol. 30, no. 1, pp. 29–43, 2018.

[16] G. Hossain and J. D. Elkins, "Cognitive Effort Assessment Through Pupillary Responses: Insights from Multinomial Processing Tree Modeling and Neural Interconnections," Online Journal of Communication and Media Technologies, vol. 14, no. 1, 2024.

[17] G. Hossain and M. Yeasin, "Understanding Effects of Cognitive Load from Pupillary Responses Using Hilbert Analytic Phase," in Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), 2014.

[18] J. D. Elkins and G. Hossain, "Multinomial Processing Models in Visual Cognitive Effort Diagnostics," in Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), 2015.

[19] M. H. Myers and G. Hossain, "Dual EEG Alignment Between Participants During Shared Intentionality Experiments," Brain Research, vol. 1790, 2022.

[20] R. R. Ghahari, G. Hossain, et al., "Semi-Aural Interfaces: Investigating Voice-Controlled Aural Flows," Journal of Interacting with Computers, vol. 28, no. 5, pp. 345–356, 2016.

[21] G. Hossain, "Rethinking Self-Reported Measure in the Subjective Evaluation of Assistive Technology," Human-Centric Computing and Information Sciences, vol. 7, no. 23, 2017.