Project: Exploratory Data Analysis on Insurance Claim Fraud Detection

Dataset Source: This analysis utilizes the "EDA on Insurance Claim Fraud Detection" dataset, sourced from Kaggle.

Objective: The goal of this exploratory data analysis is to identify patterns, trends, and irregularities in insurance claims data. The analysis focuses on detecting key indicators of fraudulent claims, supporting the development of effective detection strategies, and delivering actionable business insights.

Tools Used: Microsoft Excel was used for the entire data analysis process, including:

- Data cleaning and preparation
- Creation of pivot tables for summarization
- Generation of charts and visualizations
- Design of an interactive dashboard

Data Cleaning & Preparation

Objective:
Transform raw, unstructured insurance claims data into a clean, consistent, and analysis-ready dataset.

Process:
All missing and blank values were identified and addressed. Numerical blanks were replaced with 0 or the column median, while categorical blanks were labeled as **“Unknown”** or **“Not Provided”** to preserve data integrity. Text fields were standardized by correcting spelling errors, removing extra spaces using Excel’s **TRIM()** function, and ensuring consistent categorical values.

Column headers were formatted in proper case and bolded for readability. Dates were standardized into a consistent format (e.g., DD-MM-YYYY or MM/DD/YYYY), and monetary fields were formatted in currency style ($#,##0.00) for clear financial representation.

Duplicate records were removed using a composite key to ensure each claim was unique, and irrelevant identifiers such as Insured Zip and Policy Number were dropped to focus on variables relevant for fraud detection.

Pivot Tables & Analysis

Overall Fraud Prevalence

Approximately 24.7% of the total claim value ($546,238,648) was flagged as fraudulent ($131,658,461), highlighting a significant area of financial loss for the company.

State-wise Fraud Analysis

Fraud rates varied across states, with Ohio having the highest rate at 43% and West Virginia the lowest at 18%, indicating potential regional patterns and the need for targeted fraud detection strategies.

Average Claim Amount

The average fraudulent claim ($60,302) was notably higher than the average non-fraudulent claim ($50,289), suggesting that larger claims are more likely to be fraudulent and should be prioritized for review.

Incident Type Analysis

Single vehicle collision claims had a high fraud rate of 29%, while vehicle theft and parked car incidents had lower rates of 8.5% and 9.5%, indicating that certain incident types are more prone to fraudulent activity.

Vehicle Make Correlation

Luxury brands such as BMW (32.5%) and Mercedes (31.8%) showed higher fraud rates compared to mainstream brands like Toyota (17.0%) and Honda (22.2%), suggesting that high-value vehicles are more frequently targeted.

Impact of Police Reports

Claims without a police report had a higher fraud rate (29.7%) than those with a report (25.9%), emphasizing the importance of official documentation in deterring fraud.

Time of Day Analysis

Fraud occurred consistently throughout the day, with slightly lower rates at night (21.9%), indicating that fraudulent activity is largely independent of time.

Dashboard & Visualizations

Objective

The dashboard was designed to transform analytical insights into an intuitive, interactive, and visually compelling tool for rapid executive decision-making.

Design & Execution

The main dashboard presents four key visuals: an Overall Fraud Rate KPI (24.7%), Fraud by Claim Type (bar chart), Police Report Analysis (comparative bar chart), and Top States by Fraud (horizontal bar chart). Interactive navigation is provided through hyperlinked buttons, allowing users to access detailed analysis tabs such as Vehicle Analysis, Fraud Patterns, and Geographic Details. Consistent color schemes, clear labels, and strategic chart selection ensure that insights can be interpreted quickly and accurately.

Outcome

The dashboard provides an at-a-glance overview of fraud patterns, enabling users to quickly identify high-risk areas and drill down into details with a single click. This enhances the efficiency of fraud detection and supports informed decision-making across the organization.

Findings

Analysis of the cleaned dataset revealed that 24.7% of the total claim value ($546M) was fraudulent, indicating a significant financial impact. Fraud rates varied by state, with Ohio highest at 43% and West Virginia lowest at 18%, suggesting regional patterns. Fraudulent claims were higher in value (average $60,302) compared to non-fraudulent claims ($50,289), and single vehicle collisions were more prone to fraud than theft or parked car incidents. Luxury vehicles, such as BMW and Mercedes, had higher fraud rates than mainstream brands, and claims without a police report were more likely to be fraudulent (29.7% vs. 25.9%). Fraud occurred consistently throughout the day, with slightly lower rates at night.

Recommendations

Implement stricter verification for high-value and single vehicle collision claims, require police reports for all claims, and prioritize monitoring in states with higher fraud rates. Additionally, focus on claims involving luxury vehicles and continue using dashboards for real-time fraud detection to improve efficiency and reduce financial losses.