I

By Palakpreet Kaur

Insurance Claims Analytics: SQL Reporting & Insights



# Project Overview

This project focuses on analyzing an insurance claims dataset to uncover fraud patterns, quantify financial exposure, and provide actionable recommendations to reduce fraud risk. Fraudulent claims can significantly impact an insurer’s profitability by inflating loss ratios, increasing reserves, and driving up costs for honest customers.

The dataset contains 1,000 claims, including attributes such as claim amount, incident type, state, deductible, customer demographics, and documentation details. Using SQL-based reporting and data analysis, the project aims to:

* Measure overall fraud rate and financial exposure
* Identify high-risk segments based on geographic, pricing, and behavioral patterns
* Detect operational weaknesses (e.g., missing documentation) correlated with fraud
* Generate insights for underwriting, pricing, and fraud detection strategies

**The analysis process included:**

* Data Cleaning: Handling missing and inconsistent values, standardizing categorical fields.
* SQL Query Development: Creating reusable views and aggregations to explore fraud across multiple dimensions.
* Risk Metric Calculation: Computing fraud rates, claim severity, and fraud lift for comparative risk assessment.
* Business Insight Generation: Translating raw findings into actionable recommendations for fraud mitigation.

# Executive Summary

# Dataset Description

The dataset comprises **1,000 insurance claims** with both fraudulent and non-fraudulent cases. It includes **demographic, policy, and claim-related attributes** relevant for fraud detection. Key details:

* **Number of Records:** 1,000
* **Key Columns:**
  + fraud\_reported – Indicates whether the claim was flagged as fraudulent (Y/N)
  + total\_claim\_amount – Monetary value of the claim
  + incident\_type – Type of loss (e.g., Single Vehicle Collision, Multi-Vehicle Collision, Vehicle Theft, Parked Car)
  + incident\_state and incident\_city – Geographic location of the claim
  + policy\_deductible – Deductible amount on the policy
  + insured\_occupation and insured\_education\_level – Demographic attributes
  + property\_damage, police\_report\_available – Documentation indicators
  + Other attributes: hobbies, age category, employment type

1. **Data Preparation Steps**

* Replaced placeholder values (e.g., ?, UNKNOWN) with standardized terms
* Checked for logical inconsistencies and validated relationships between columns
* Created additional derived metrics for analysis (e.g., Fraud Rate, Fraud Lift)

# Report R1: Geographic Distribution of Fraud Cases

**Objective**

To identify the **locations (states and cities)** with the highest number of **fraud-reported insurance claims**. This insight helps prioritize fraud investigations geographically and allocate resources efficiently.

SQL View Created:

* 1. **View: Fraud\_cases\_by\_state**

A screenshot of a computer

AI-generated content may be incorrect.

This view returns the top 10 states (only shown 5) with the most fraud cases.

**Observation:**

1. South Carolina has the most fraud cases (73), consisting 30% of total fraud cases. For context, the total number of fraud cases are 247.
2. Followed, New York and West Virginia consist approximately 23% and 15% respectively.

Let’s dig deeper in South Carolina through a view of States and their cities where fraud is reported.

* + 1. **View: Fraud\_cases\_by\_state\_city**

A screenshot of a computer

AI-generated content may be incorrect.

This view groups fraud cases by both city and state, allowing us to drill into high-risk metro areas.

**Observation:**

1. Columbus scores the highest contributor to fraud cases, consisting 24% of South Carolina fraud cases (7% of total fraud cases).

# Report R2: Fraud Rate by Vehicle Make & Model

**Objective:**

To identify vehicle models with a disproportionately high rate of fraud-reported insurance claims, helping the business assess risk exposure linked to specific vehicles.

* 1. **View: fraud\_cases\_by\_vehicle**

A screenshot of a data

AI-generated content may be incorrect.

This view computes fraud rates by vehicle make and model.

**Observation:**

1. Several high-end or large utility vehicles—**BMW X6 (43.75%)**, **Chevrolet Silverado (40.91%)**, **Mercedes ML350 (40.00%)**, **Mercedes C300 (38.89%)**, and **Chevrolet Tahoe (37.50%)**—show significantly elevated fraud rates relative to their total number of claims.
2. Although these models are not the most frequently insured, their **disproportionately high fraud percentages** suggest targeted fraud risks, potentially due to the **higher value or resale potential** of these vehicles.

# Report R3: Average Claim amount by fraud status

**Objective:**

To identify claim amount associated with the fraud or non-fraud cases, helping the business to identify risk with claim values.

* 1. **View: fraud\_non-fraud\_avg\_claim**

A screenshot of a computer

AI-generated content may be incorrect.

This view shows the avg\_claim\_amount for both the fraud scenarios.

Observation:

1. This indicates that **fraudulent claims are approximately 20% higher** in dollar value than legitimate claims.
2. **High-value claims**, especially those submitted in high-fraud regions or for high-risk vehicle models, should be subject to **additional verification**.

*I have something to report here – to show you min and max claim amounts for fraud scenarios:*

* + 1. **View: min\_claim\_amount\_by\_fraud**

A screenshot of a computer

AI-generated content may be incorrect.

* + 1. **View: max\_claim\_amount\_by\_fraud**

A screenshot of a computer

AI-generated content may be incorrect.

These views show the minimum and maximum claim amounts for fraudulent activities.

**Observation:**

1. While the **maximum claim amounts** are similar for both fraud and non-fraud cases, the **minimum amount for fraudulent claims is significantly higher**. This suggests that fraudulent claims tend to occur in **moderate to high-value ranges**, possibly to avoid suspicion that comes with extreme values or small claims.

# Report R4: Customer segmentation for fraud

To segment customers that are more likely to cause fraud based on their insured featues like education, occupation, hobbies, relationship

* 1. **View: Fraud cases by occupation**

**A screenshot of a computer

AI-generated content may be incorrect.**

This view shows the top 5 occupation that are more likely to be involved in a fraud. “Exec-managerial” is at the top.

**Observation:**

1. Fraud rates vary significantly by occupation, with **executive/managerial roles exhibiting the highest fraud rate at 36.84%**, followed by **farming-fishing (30.19%)** and **craft-repair (29.73%)**.
2. This finding is counterintuitive and suggests that fraud is not limited to lower-income or high-risk job categories. It indicates that **occupation is a potentially strong predictor** for fraud risk and should be considered in fraud detection strategies.
   1. **View: Fraud cases by hobbies**

**A screenshot of a computer

AI-generated content may be incorrect.**

This view includes the fraud cases and percentage based on the person hobbies.

**Observation:**

1. Hobbies appear to be a strong differentiator in fraud likelihood. Claims associated with chess (82.61%) and Cross-fit (74.29%) show disproportionately high fraud rates compared to other hobbies such as Polo (27.66%) or Board Games (29.17%).
2. These patterns could be linked to **fraud clusters, identity fraud, or behavioral trends** that merit further investigation. Given the non-trivial sample size (nearly 23% of the dataset across the top 5 hobbies), this variable is a **strong candidate** for inclusion in fraud detection models or business rules.
   1. **View: Customer segmentation by relationship and hobbies**

A screenshot of a computer

AI-generated content may be incorrect.

This view shows the fraud percentage when certain hobbies and occupation is combined.

**Observation:**

1. **Chess** as a hobby across multiple job categories (e.g., Private house service, Sales, Machine operators) showed **100% fraud rates** with 4–6 claims each.
2. Cross-fit and Yachting combined with Executive or Transport roles also produced 80% fraud rates. Farming-fishing + Skydiving had a notable 57% fraud rate.
3. These combinations may indicate **behavioral fraud patterns** or **identity misrepresentation** and are strong candidates for **rule-based fraud alerts or advanced feature interactions in predictive modeling**.
   1. **View: Customer segmentation based on occupation and age**

A screenshot of a data table

AI-generated content may be incorrect.

This view shows how fraud cases are distributed over occupation and age distribution.

**Observation:**

1. **Craft-repair professionals aged 55–64** had the **highest fraud rate (60%)** among all significant groups. **Executive/managerial professionals aged 25–34** followed with a fraud rate of **53.85%**, and those aged 35–44 had a rate of **36.67**%.
2. These insights suggest that both **age and occupation together** can explain fraud behavior better than either variable alone. This segmentation should be considered for **risk scoring models**, **manual review prioritization**, and **targeted education or fraud prevention programs**.

# Report R5: Policy Evaluation

* 1. **View: Policy deductible associated with larger claims**

A screenshot of a data

AI-generated content may be incorrect.

This view shows the states with their policy deductibles with their claims and values.

**Observation:**

1. Ohio shows the highest fraud rate for $500 deductibles (62.5%) and elevated fraud for $2000 as well.
2. South Carolina, while having more claims overall, exhibits a 34.5% fraud rate for $500 deductibles, and ~26–27% for higher deductibles.
3. North Carolina has moderate to high fraud rates across both $1000 and $2000 levels (~32–37%).
4. These trends highlight how deductible structures interact with location-specific behavior, suggesting the need to incorporate state-deductible pairings into fraud risk models and rule-based detection systems.

# Report R6: Incident types

* 1. **View: Incident type to highest total claims**

A screenshot of a computer

AI-generated content may be incorrect.This shows the total claims based on incident types with their claim amount and fraud rate.

Observation:

1. Incident type analysis reveals that **fraud is most prevalent in collision-related claims**, with **Single Vehicle Collisions at 29.03% fraud rate** and **Multi-Vehicle Collisions at 27.21%**. These incidents also involve substantially higher claim values, averaging **$60,000+ per claim**, making them a critical area for fraud prevention.
2. Conversely, **Vehicle Theft (8.51%)** and **Parked Car incidents (9.52%)** show significantly lower fraud rates and lower claim amounts (~$5,000), suggesting that **fraud risk in these categories is relatively low**. Prioritizing resources for **collision-related claims** could yield the highest return in fraud mitigation efforts.

# Report R7: Operational Gaps

* 1. **View: How 'unknown values' impact the fraud**

A screenshot of a computer

AI-generated content may be incorrect.

This view shows how the ‘unknown values’ impact the fraud rate, enforcing the need for stricter reporting requirements, or apply extra fraud checks when fields are unknown.

Observation:

1. Analysis of fraud rates by property damage and police report availability shows that **missing or unclear documentation significantly increases fraud risk**.
2. For instance:

**Property Damage = NO and Police Report = UNKNOWN** → **31.53% fraud rate**

**Both fields UNKNOWN** → **29.03% fraud rate**

1. Even fully documented claims (**YES/YES**) show a fraud rate of **24.44%**, but this is lower than incomplete or inconsistent cases.
2. **Recommendation:** Prioritize stricter enforcement of documentation and apply additional fraud checks when these fields are missing or marked as UNKNOWN.

# Report R8: Fraud Lift

* 1. **View: How 'unknown values' impact the fraud**

A screenshot of a data

AI-generated content may be incorrect.

This view illustrates the relative risk in the report.

Observation:

1. Fraud risk analysis by **State × Deductible** shows that certain combinations significantly exceed the portfolio average.
2. For instance: **Ohio ($500 deductible)** exhibits a **62.5% fraud rate**—**2.5× higher than baseline**.
3. **Pennsylvania ($1000 deductible)** and **Ohio ($2000 deductible)** also show elevated risk (1.6× and 1.8× lift, respectively).
4. **South Carolina** presents a major exposure scenario: while its fraud lift is lower (1.4×), its **claim volume is 87**, and average claim size exceeds $55K, making this a critical focus for fraud detection and resource allocation.