TUNIS BUSINESS SCHOOL
UNIVERSITY OF TUNIS

BI AND DBMS PROJECT REPORT

Car Insurance Claims Fraud & Cost Analysis

Prepared by:
Noursine Amira
Maycem Azaza
Rahma Haddouchi
Lina Brahmi

Professor: Ameni Azzouz

Academic Year:
2025-2026

Contents

Chapter 1

General Introduction

1.1 Business Needs

The project operates within the **Property & Casualty (P&C) Insurance** sector. Profitability in this sector relies heavily on accurate risk assessment and efficient claims management. Currently, the business faces significant challenges:

- **Undetected Fraud:** Insurance fraud (staged accidents, fake injuries) drains profits and raises premiums for honest customers.
- **Operational Inefficiency:** Claims adjusters waste time investigating low-risk claims manually.
- **Lack of Visibility:** There is no centralized view connecting driver demographics and incident details to financial losses.

1.2 Project Goals

The primary objective is to build an end-to-end Business Intelligence solution that:

1. Identifies fraud patterns and high-risk segments.
2. Monitors Key Performance Indicators (KPIs) such as *Fraud Rate* and *Loss Ratio*.
3. Optimizes the claims approval process by reducing the average settlement time.

1.3 Key Questions to Explore

To achieve these goals, our analysis addresses the following questions:

1. What is the overall fraud rate across the portfolio?

2. How does the age of the driver correlate with fraud probability?
3. Are specific vehicle categories (e.g., Luxury, Utility) more prone to fraud?
4. Does the presence of witnesses or police reports reduce the likelihood of fraud?
5. Do past offenders (people with prior claims) tend to re-offend?

1.4 Deliverables

- **Cleaned Dataset:** `insurance_fraud_data_cleaned.csv`
- **ETL Pipeline:** Python Notebook for data transformation.
- **Data Model:** Star Schema designed in Power BI.
- **Interactive Dashboards:** A comprehensive report containing Executive, Deep Dive, and Operational views.

Chapter 2

Technical Implementation

2.0.1 Tools and Technologies

The following tools were utilized to process the data and build the visualization suite:

- **Python**: The core programming language used for backend data preprocessing.
- **Pandas**: Utilized for data cleaning, transformation, and standardization (e.g., standardizing `gender` and `fraud_reported` columns).
- **NumPy**: Employed for numerical operations and handling arrays.
- **Microsoft Power BI**: The central platform used to design and deploy the interactive dashboards, including the *Executive Summary*, *Deep Dive*, and *Claim Details* views.
- **Microsoft Azure Maps**: Integrated within the Power BI environment to visualize the geospatial distribution of international claims.

To facilitate data manipulation, we utilized standard Python data science libraries:

```
import pandas as pd
import numpy as np

# Pandas is used for data frame manipulation
# NumPy is used for numerical operations
```

Listing 2.1: Importing Required Libraries

2.1 Project Phases

2.1.1 Data Gathering

The raw dataset (`insurance_fraud_data.csv`) contains approximately 12,000 records. It includes policy details, claim amounts, driver demographics, and incident descriptions.

2.1.2 Data Preparation (ETL Showcase)

We implemented a robust ETL pipeline using Python. Below are the key steps performed in the `Data Preparation.ipynb` notebook:

1. Handling Missing Values
We imputed missing numerical values (e.g., `age_of_driver`) with the median and categorical values with the mode.

```
# Replace '?' with NaN and impute
df.replace('?', np.nan, inplace=True)
df['age_of_driver'].fillna(df['age_of_driver'].median(), inplace=True)
df['gender'].fillna(df['gender'].mode()[0], inplace=True)
```

Listing 2.2: Imputing Missing Values

2. Data Standardization
To ensure clean filtering in the dashboard, we standardized the `gender` column and converted the target variable `fraud_reported` into a binary format.

```
# Standardizing Gender text
df['gender'] = df['gender'].replace({'M': 'Male', 'F': 'Female'})

# Converting Fraud Reported to Binary (1/0)
df['fraud_reported'] = df['fraud_reported'].apply(lambda x: 1 if x == '
    Y' else 0)
```

Listing 2.3: Standardizing Categorical Data

3. Time Series Preparation
We parsed the claim dates to enable Month-over-Month (MoM) and Year-over-Year (YoY) analysis.

```
# Extracting Year and Month
df['claim_date'] = pd.to_datetime(df['claim_date'])
df['claim_year'] = df['claim_date'].dt.year
df['claim_month'] = df['claim_date'].dt.month_name()
```

Listing 2.4: Date Parsing

2.1.3 Data Warehouse Creation & Modeling

We designed a **Star Schema** to optimize the data for Power BI performance.

- **Fact Table:** `Fact_Claims` (Quantitative Data).
- **Dimension Tables:** `Dim_Driver`, `Dim_Vehicle`, `Dim_Date`, `Dim_Incident`.

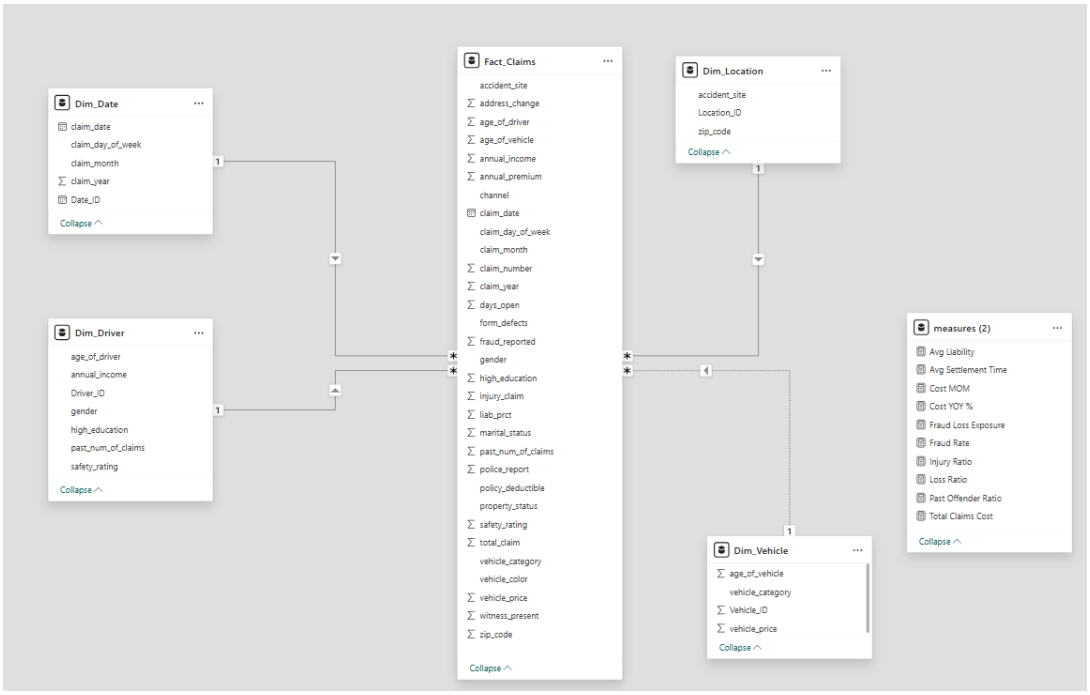

Figure 2.1: Star Schema Relationship Diagram

2.1.4 Data Analysis Phase (Dashboards)

The visualization layer consists of three key dashboards designed for different stakeholders.

1. Executive Summary

This dashboard provides high-level visibility into the strategic health of the insurance portfolio. It features critical KPIs including **Total Claims Cost**, **Fraud Rate**, **Avg Settlement Time**, and **Fraud Loss Exposure**. Additionally, a dynamic slicer has been integrated at the top-left, enabling users to filter and analyze these performance metrics by **Quarter**.

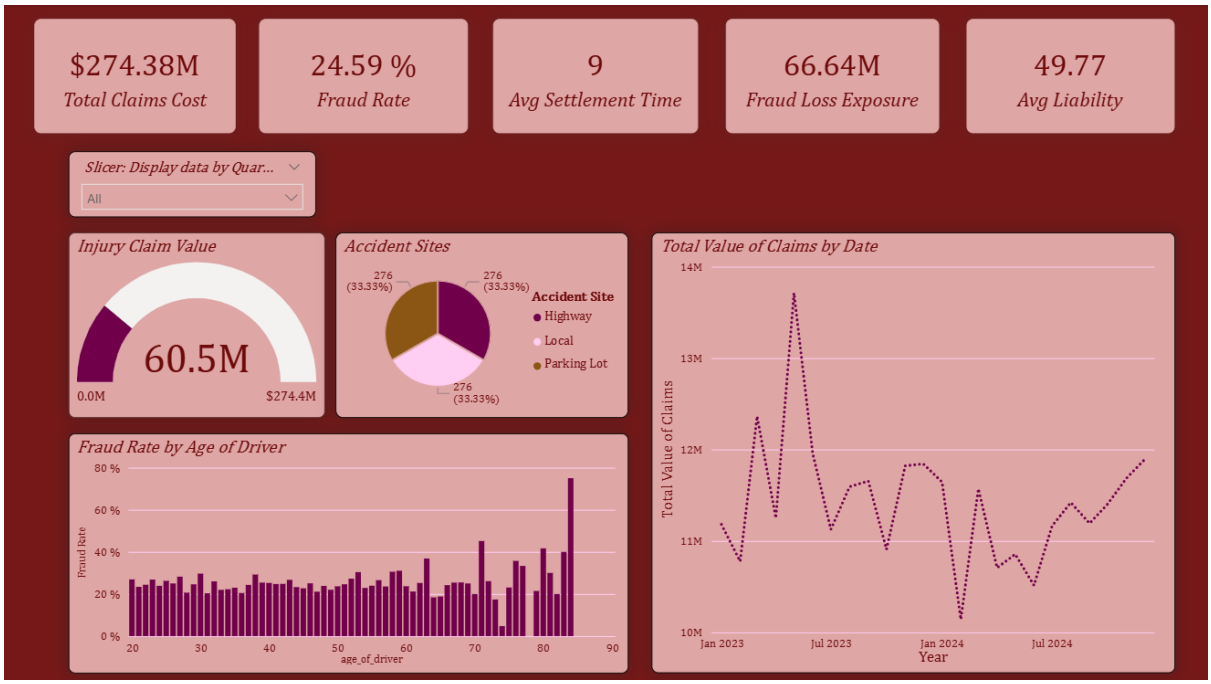

Figure 2.2: Executive Summary Dashboard

2. Deep Dive Analysis

The Deep Dive dashboard facilitates a granular investigation into the underlying drivers of risk and liability. Unlike the executive summary, this view allows stakeholders to isolate specific behavioral and mechanical correlations using advanced interaction features.

Visual Breakdown:

- **Recidivism & Liability (Top Left):** A combo chart juxtaposes the *Total Claim Value* against *Average Liability Percentage*. The secondary axis reveals a critical trend: while first-time claimants represent the highest total cost volume, the liability percentage fluctuates significantly based on the number of past claims, highlighting recidivism risks.

- **Safety vs. Police Interaction (Top Right):** This dual-axis chart plots the density of *Past Claims* and *Police Reports* against *Safety Ratings* (0-100). It allows analysts to determine if drivers with lower safety scores are more likely to be involved in incidents requiring police intervention.
- **Vehicle Category Analysis (Bottom Left):** Using "small multiples," this section compares *Safety Ratings* and *Liability* across three distinct vehicle classes: **Compact, Large, and Medium**. This faceted view helps isolate whether specific vehicle sizes are more prone to defects or liability spikes.
- **Fraud Status Verification (Bottom Right):** A stacked bar chart contrasts the volume of *Police Reports* for legitimate claims ("No Fraud") versus fraudulent ones, serving as a quick validation metric for investigators.

Interactivity and Scenario Testing:
The dashboard features dynamic slicers that allow for hypothesis testing. As shown in the figures below, we performed a specific segmentation test:

1. **General View (Figure 2.3):** Shows the dataset in its entirety.
2. **Filtered Scenario (Figure 2.4):** We isolated drivers who **Own** their vehicles and drive **Red** cars. This interaction instantly updates all metrics, revealing how specific demographics contribute to the overall risk profile.

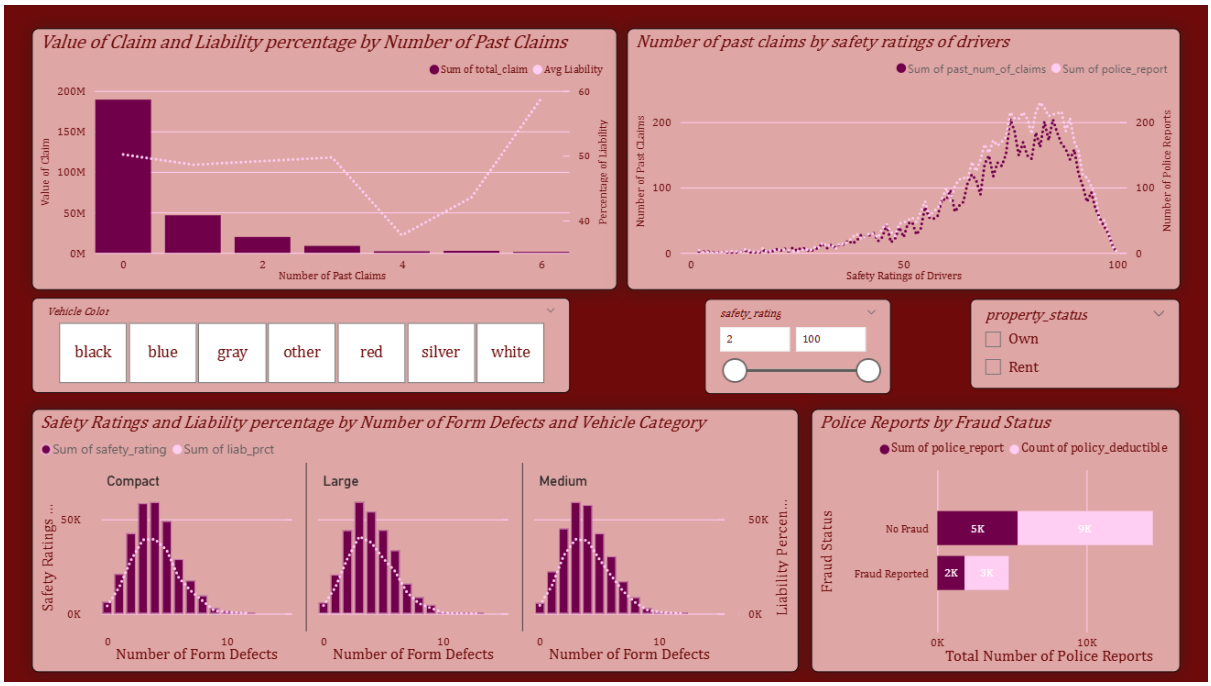

Figure 2.3: Deep Dive Analysis: General Overview (Unfiltered)

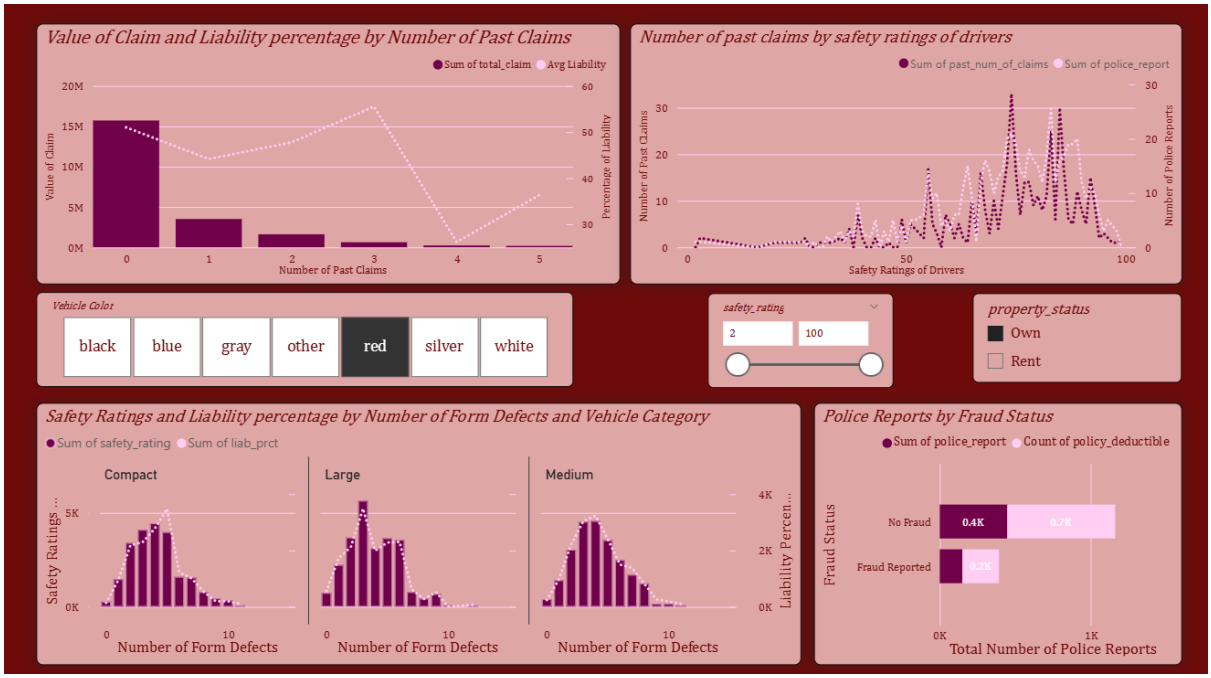

Figure 2.4: Deep Dive Analysis: Filtered Scenario (Red Vehicles, Property Owners)

3. Operational & Claim Details (Geospatial Analysis)

This view shifts focus to the geospatial distribution of risk, utilizing **Microsoft Azure Maps** to visualize the *Total Value of Claims* by location. Unlike the high-level summary, this dashboard provides a multi-tiered geographic analysis that enables stakeholders to pinpoint liability hotspots.

- **Global Footprint**: The macroscopic view reveals an international operational scope. Beyond the primary North American market, the organization manages active claim clusters in **South America** (specifically Brazil), **Europe** (Germany and Italy), and **East Asia**.
- **Regional Density (US & Europe)**:
  - The **United States** view highlights the highest volume of claims, densely clustered across the **Midwest** (Ohio, Indiana) and the **Northeast**.
  - The **European** view isolates specific operational hubs in Central Europe, showing distinct activity in Munich and Northern Italy.
- **Fraud Exposure Visualization**: In all views, map markers function as dynamic pie charts. The color coding allows for immediate risk assessment: **Dark Purple** segments indicate *Fraud Reported*, while **Light Pink** segments represent *No Fraud* claims. Larger markers denote locations with a higher total claim value.

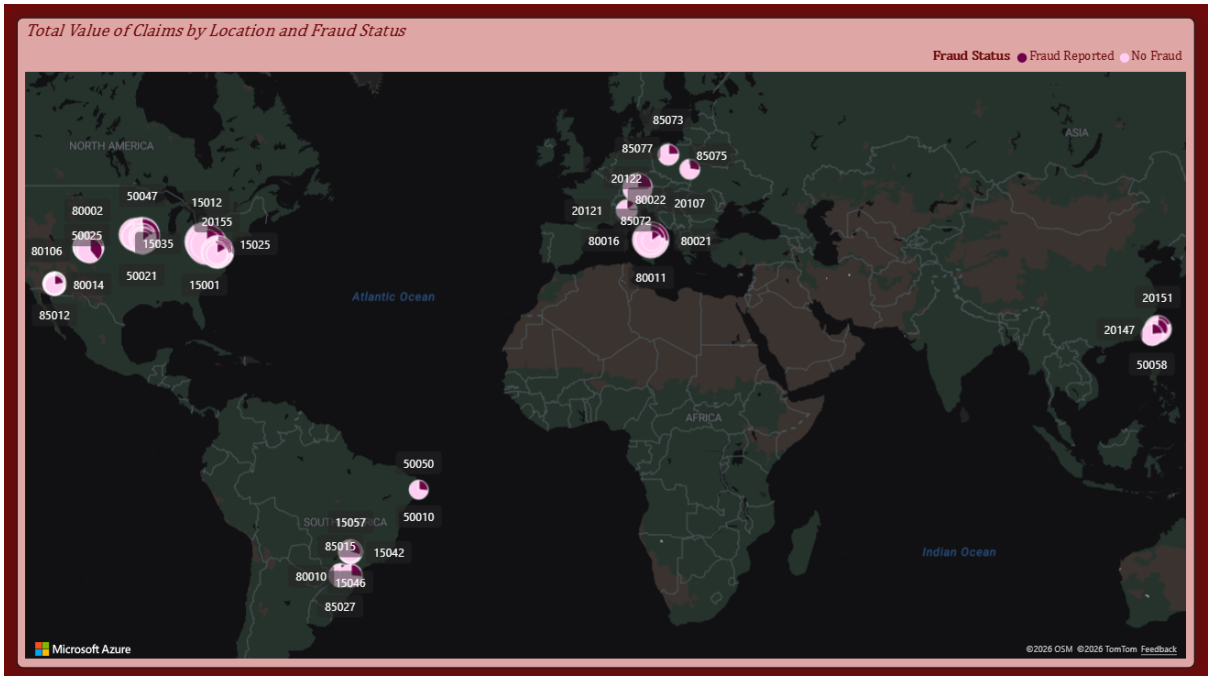

Figure 2.5: Global Operational Scope (Americas, Europe, Asia)

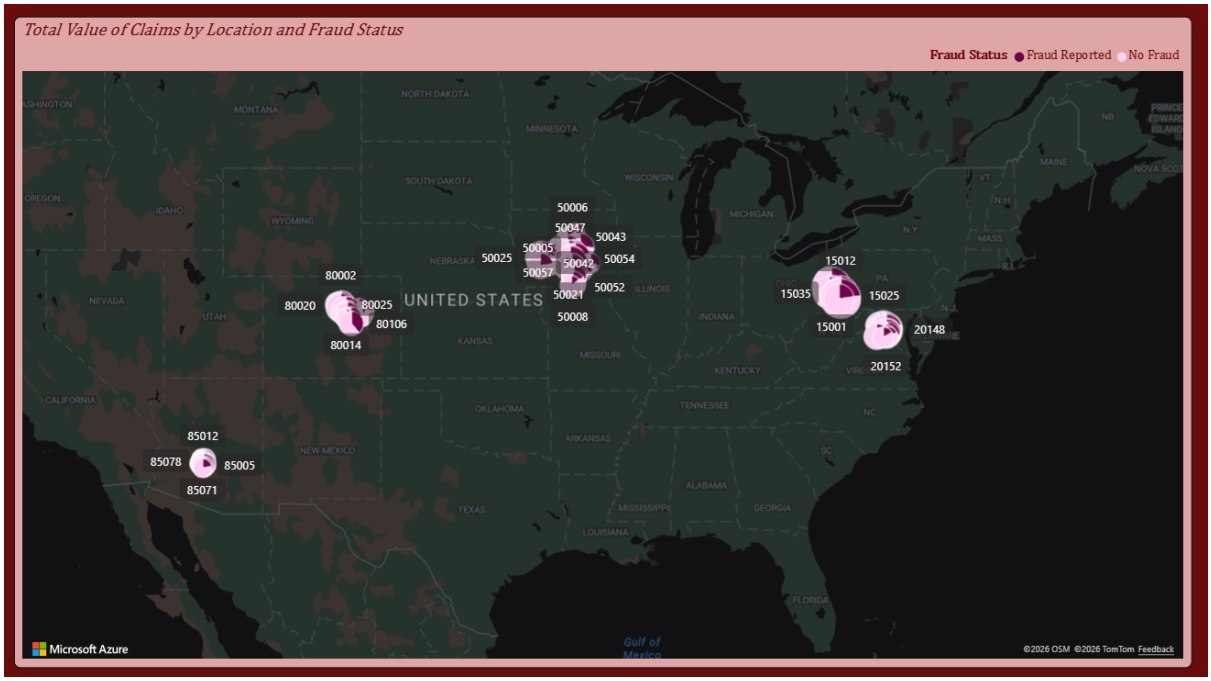

Figure 2.6: High-Density Cluster Analysis: United States Market

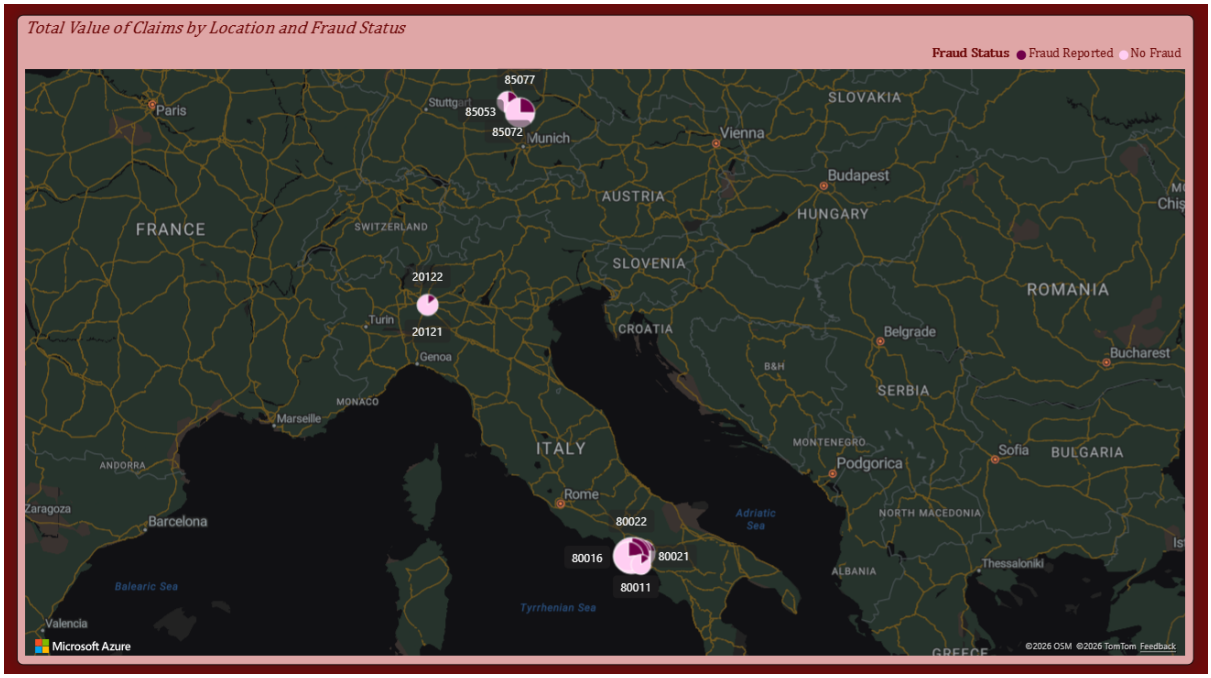

Figure 2.7: Regional Focus: European Market (Germany & Italy)

Chapter 3

Conclusion

3.0.1 Key Findings

The comprehensive analysis of the insurance portfolio through Power BI has revealed critical insights into financial exposure and operational risk:

- **High Financial Exposure:** The organization faces a total **Fraud Loss Exposure of $66.64M**, representing a significant portion of the total claim value. This confirms that nearly one-quarter of the portfolio's value is under threat from fraudulent activity.
- **Global & Regional Risk:** Risk is not evenly distributed. The geospatial analysis identifies the **United States** (specifically the Midwest and Northeast) as the primary hub for claim volume, while distinct operational clusters were identified in **Germany** and **Italy**, requiring cross-border monitoring.
- **Operational Lag:** The **Average Settlement Time** stands at **9 days**. This prolonged lifecycle suggests operational inefficiencies that may be contributing to higher costs.
- **Cost Drivers: Injury Claims** are a massive financial burden, totaling **$60.5M**. This specific claim type represents a significant portion of the $274.38M Total Claims Cost.

3.0.2 Answers to Analytical Questions

The BI solution provided definitive, data-driven answers to the five key research questions proposed at the project's inception:

- **Q1: What is the overall fraud rate across the portfolio?**
  Answer: The overall fraud rate is **24.59%**. This metric serves as the baseline KPI for measuring the effectiveness of future anti-fraud initiatives.

- **Q2: How does the age of the driver correlate with fraud probability?**
  Answer: The *Fraud Rate by Age of Driver* visualization reveals a non-linear relationship. While rates fluctuate across middle age groups, there is a concerning spike in fraud frequency among drivers in the **80+ age bracket**, where rates approach 80% in specific instances. This suggests targeted auditing is required for senior demographics.

- **Q3: Are specific vehicle categories more prone to fraud?**
  Answer: The analysis segmented vehicles into **Compact, Large, and Medium** categories. The "Small Multiples" analysis reveals that **Large** vehicles exhibit a distinct liability density curve relative to defect counts, suggesting a higher mechanical or reporting correlation compared to Compact cars.

- **Q4: Does the presence of witnesses or police reports reduce the likelihood of fraud?**
  Answer: Yes. The *Police Reports by Fraud Status* chart demonstrates a strong inverse correlation. Legitimate ("No Fraud") claims are supported by a significantly higher volume of police reports (approximately 14,000 total verified reports) compared to fraudulent cases (approximately 5,000 reports). Missing police documentation is a strong leading indicator of fraud.

- **Q5: Do past offenders (people with prior claims) tend to re-offend?**
  Answer: Yes. The "Deep Dive" combo chart shows that while first-time claimants (0 past claims) account for the highest total value, the **Average Liability Percentage** trends upward with history. Specifically, drivers with **6 past claims** show a spike in average liability (reaching nearly 60%), validating recidivism as a critical risk factor.

3.0.3 Strategic Recommendations

Based on these findings, we propose the following business actions:

1. **Geographic Resource Allocation:** Shift the majority of fraud prevention resources to the **US Midwest and Northeast regions**, as the map analysis proves this is where the highest volume of financial liability sits.
2. **Automated Risk Flagging:** Implement a rule in the claims management system to auto-flag any filing where the claimant has **Past Claims $\geq$ 6** or where a **Police Report** is missing, as these are statistically proven high-risk indicators.
3. **Settlement Optimization:** Initiate a process review to reduce the **9-month settlement time**. Accelerating the closure of low-risk claims (verified by police reports) will free up investigators to focus on the **$66.64M** fraud exposure.

3.0.4 Limitations & Future Work

- **Data Standardization:** Integrating global zip codes required careful handling to distinguish between European and US locations. Future iterations will enforce a

strict `Country` field to eliminate mapping ambiguity.

- **Predictive Analytics:** While the current solution describes *what* happened, the next phase should employ Machine Learning (e.g., Random Forest) to predict *future* fraud probability in real-time based on Age, Vehicle Category, and History.

3.0.5 Final Conclusion

This project has successfully transformed raw insurance data into a strategic asset. By visualizing the **$274.38M Total Claims Cost** and isolating the **$66.64M Fraud Loss Exposure**, the organization can now move from reactive observation to proactive mitigation. The path forward is clear: target the high-risk US regions, monitor the senior driver demographic, and streamline the settlement lifecycle to secure profitability.

3.1 GitHub Repository

The complete code, dataset, and Power BI files are available at:
https://github.com/rahmahaddouchi/Car-Insurance-Claims-Fraud-and-Cost-Analysis