Exploratory Data Analysis: Fraud Detection

Tommy O'Gorman

Student ID: K00301970

Lecturer: Dr Matthew Horrigan

Date: April 19, 2025

Table of contents

1 Introduction

1.1 Problem Statement

Our company is tasked with uncovering insights from motor insurance claims data to support the early detection of fraudulent activity. Fraudulent claims lead to substantial financial losses, and existing detection methods are often inefficient or reactive. Exploratory Data Analysis (EDA) is a vital first step in identifying trends, behaviours, and risk factors that distinguish fraudulent from legitimate claims.

This report presents an EDA of motor insurance claim records, focusing on behavioural, demographic, and cost-related variables. The analysis aims to uncover fraud-related patterns that can inform future model development and strengthen fraud prevention strategies.

1.2 Solution Statement

This EDA followed a structured and reproducible process to extract insights from raw insurance claims data. The approach involved:

1. **Data Cleaning**: Resolved inconsistencies, applied consistent formatting, and ensured data quality to enable accurate and trustworthy analysis.
2. **Feature Engineering**: Created new variables to capture behavioural, demographic, and financial signals relevant to fraud detection. These additions helped generate new insights from the original data and supported deeper EDA.
3. **Exploratory Visualisation**: Used targeted plots to examine relationships between the fraud flag indicator and key factors such as repair cost, geographic region, driver age, and car occupancy (e.g. number of people in the car).
4. **Pattern Detection**: Identified repeated individuals, regional fraud hotspots, and cost-related anomalies to highlight subtle indicators of suspicious behaviour.
5. **Insight Generation**: Delivered findings to support operational fraud prevention strategies and guide potential future predictive modelling. These insights were grounded in EDA and aimed at real-world application.

Each step enhanced the dataset's interpretability and provided a solid foundation for both practical use and future analytical work.

2 Overview of Dataset

The dataset contains **1,000 motor insurance claim records**, each providing granular information on the individuals involved, the nature of the incident, and the claim outcome. Specifically, it includes:

- **Identifiable names** of drivers and passengers
- **Driver age**, included as a numeric variable
- **Residential addresses**, including street names and house numbers
- **Context of the incident**, including any passengers present
- **Claim details**, including repair costs and a fraud flag indicating whether the claim was identified as fraudulent or not fraudulent

The presence of **real names and addresses** suggests that this dataset has not been anonymised, and therefore requires **ethical and responsible handling** throughout the analysis.

driver	age	address	passenger1	passenger2	repaircost	fraudFlag
JOSEPH MCGRATH	24	3 CO$RIB VIEW	JOSEPH GRIFFIN	NA	approx 3k	FALSE
MARY BRENNAN	53	21 BLACKWATER VIEW	NA	NA	approx 2k	FALSE
JOSEPH COLLINS	48	7 SLANEY LODGE	NA	NA	approx 2k	FALSE
ROBERT WALSH	40	12 LIFFEY GROVE	NA	NA	approx 500	FALSE
KEVIN OCONNELL	27	21 BLACKWATER GLADE	NA	NA	approx 500	FALSE

Figure 1. Sample of Motor Insurance Claims Data

3 Data Cleaning Overview

The dataset required structured cleaning to ensure consistency, accuracy, and readiness for analysis. These steps reduced noise, removed inconsistencies, and prepared the data for downstream feature engineering and modelling.

- **Renamed columns using `snake_case`**: Improves readability and coding consistency, making variable names easier to reference and manage throughout the project.
- **Standardised text fields to uppercase**: Although these fields were already capitalised in the original dataset, they were converted to uppercase programmatically to ensure consistency and reproducibility in case future datasets (e.g., Excel uploads) are not already formatted this way.
- **Cleaned the `address` column**: Removed numeric prefixes, punctuation, and extra whitespace to standardise address strings and support the creation of routing-based geographic variables.
- **Extracted `street_number`**: Separating the numeric part of addresses allows for potential spatial or sequential analysis (e.g., fraud clusters by street range).
- **Created a `routing_area` column**: Mapped address patterns — including common misspellings (e.g., "BLACATER" for "BLACKWATER") — to their correct routing zones. This supports geographic trend analysis, which is key in fraud detection.
- **Cleaned the `repair_cost` field and converted to numeric (`repair_cost_eur`)**: Raw cost entries included symbols and inconsistent formatting (e.g. "~1500" or "€2k"). Standardising to numeric euros allowed reliable aggregation and visualisation.
- **Enforced consistent data types**: Applied appropriate formats (e.g. factors, characters, integers) to ensure the dataset behaves predictably in analysis, visualisation, and modelling.

This structured approach ensured the data was clean, consistent, and ready for analysis — helping to avoid misleading insights later in the process.

A preview of the cleaned dataset is shown below.

driver	age	street_number	address	routing_area	passenger_1	passenger_2	repair_cost	repair_cost_eur	fraud_flag	fraud_flag_num
JOSEPH MCGRATH	24	3	CORIB VIEW	CORRIB	JOSEPH GRIFFIN	NA	approx 3k	3000	FALSE	0
MARY BRENNAN	53	21	BLACKWATER VIEW	BLACKWATER	NA	NA	approx 2k	2000	FALSE	0
JOSEPH COLLINS	48	7	SLANEY LODGE	SLANEY	NA	NA	approx 2k	2000	FALSE	0
ROBERT WALSH	40	12	LIFFEY GROVE	LIFFEY	NA	NA	approx 500	500	FALSE	0
KEVIN OCONNELL	27	21	BLACKWATER GLADE	BLACKWATER	NA	NA	approx 500	500	FALSE	0

Figure 2. Preview of Cleaned Dataset

3.1 Feature Engineering

Feature engineering was used to create structured, meaningful variables that support deeper insight during the EDA process. While predictive modelling is not the focus here, these features may also prove useful in later modelling stages.

- **Binned age into `age_group`**: Created logical age bands (`under_25`, `25_to_65`, `over_65`) aligned with categories commonly used in insurance-related analysis. This transformation allows for clearer comparisons across age segments and supports non-linear analysis.
- **Created `num_passengers` and `total_in_car`**: These variables quantify vehicle occupancy, helping to assess whether the number of individuals in the car at the time of the claim is associated with fraudulent behaviour.
- **Created `carpool` flag**: A binary variable indicating whether more than one person was in the car. This simplified analysis of shared travel behaviour and its potential relationship with fraudulent claims.
- **Flagged `high_cost` claims**: Created a binary flag for claims exceeding €1,500 — a common industry threshold. This complemented the analysis of `repair_cost_eur` as a continuous variable, allowing comparisons both across cost bands and detailed distributions.
- **Converted `fraud_flag` to numeric (`fraud_flag_num`)**: A new column was created to convert the original TRUE/FALSE fraud flag into a numeric format (1/0). This allowed for inclusion in correlation analysis and visual comparisons (e.g. bar charts, matrices), while keeping the original flag intact for reference.
- **Applied one-hot encoding to `routing_area`**: Converted each routing area into a separate numeric column, making it suitable for correlation analysis and easier to include in visual comparisons of fraud patterns across geographic regions.

All newly engineered variables were logically integrated into the dataset with appropriate data types for analysis. A preview of the feature-engineered dataset is shown below.

driver	age	age_group	street_number	address	routing_area	passenger_1	passenger_2	num_passengers	total_in_car	carpool	repair_cost	repair_cost_eur	high_cost	fraud_flag	fraud_flag_num
JOSEPH MCGRATH	24	under_25	3	CORIB VIEW	CORRIB	JOSEPH GRIFFIN	NA	1	2	1	approx 3k	3000	1	FALSE	0
MARY BRENNAN	53	25_to_65	21	BLACKWATER VIEW	BLACKWATER	NA	NA	0	1	0	approx 2k	2000	1	FALSE	0
JOSEPH COLLINS	48	25_to_65	7	SLANEY LODGE	SLANEY	NA	NA	0	1	0	approx 2k	2000	1	FALSE	0
ROBERT WALSH	40	25_to_65	12	LIFFEY GROVE	LIFFEY	NA	NA	0	1	0	approx 500	500	0	FALSE	0
KEVIN OCONNELL	27	25_to_65	21	BLACKWATER GLADE	BLACKWATER	NA	NA	0	1	0	approx 500	500	0	FALSE	0

Figure 3. Preview of Feature Engineered Dataset

Exploratory Data Analysis

With the cleaned and feature-enhanced dataset now in place, we conducted an Exploratory Data Analysis (EDA) to uncover meaningful patterns related to fraudulent motor insurance claims. The goal of this section is to identify potential behavioural, demographic, and geographic indicators of fraud that can inform future predictive modelling efforts.

Each visual in this section is accompanied by a brief interpretation, highlighting insights that may help the company reduce financial losses and improve the early detection of suspicious claims.

4 Financial Overview of Fraud

4.1 Overall Financial Impact of Fraud

Understanding the overall financial burden of fraudulent claims helps to prioritise early detection efforts. While fraudulent claims are less frequent, they may account for a disproportionate share of total costs, making them a critical focus for insurers.

The bar chart below compares the total repair costs of non-fraudulent versus fraudulent claims.

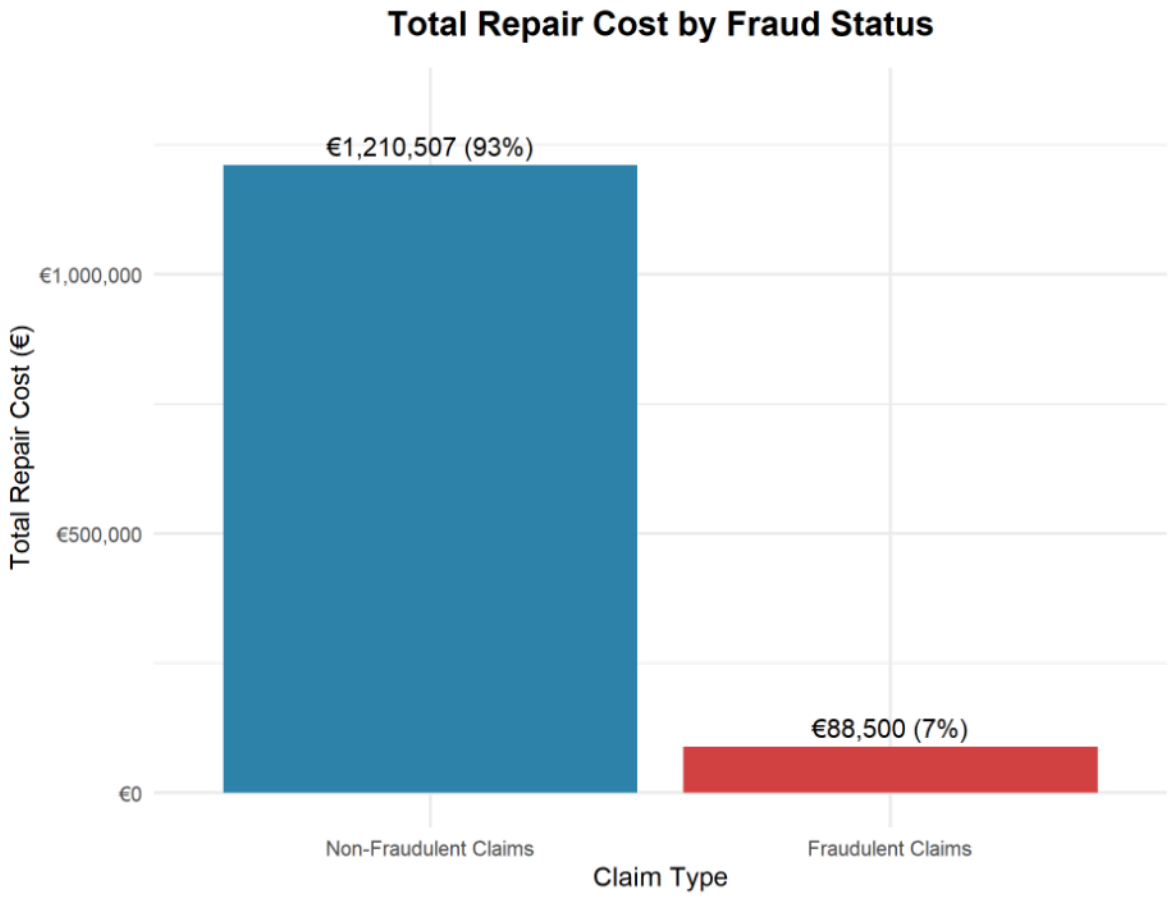

Figure 4. Total repair cost associated with non-fraudulent and fraudulent claims. Fraudulent claims account for just 7% of claim volume but contribute significantly to total repair costs.

Despite representing a smaller proportion of all claims, fraudulent claims account for €88,500 — or 7% — of total repair costs. In contrast, non-fraudulent claims account for the remaining 93%. Although fraud is relatively infrequent, its financial impact is substantial, and even a small number of high-cost claims can result in significant losses.

5 Correlation with Fraud

Understanding how features relate to fraudulent claims helps identify strong predictors for future modelling. This section presents two distinct visualisations: one showing **correlation with the fraud flag** specifically, and another showing **correlation between all numeric features** to inform feature selection.

5.1 Correlation Strength with Fraud Flag

The first chart focuses on identifying which features are most directly associated with fraudulent claims. It displays the correlation of each numeric feature with `fraud_flag_num`, where a higher value suggests a stronger linear relationship with fraud likelihood.

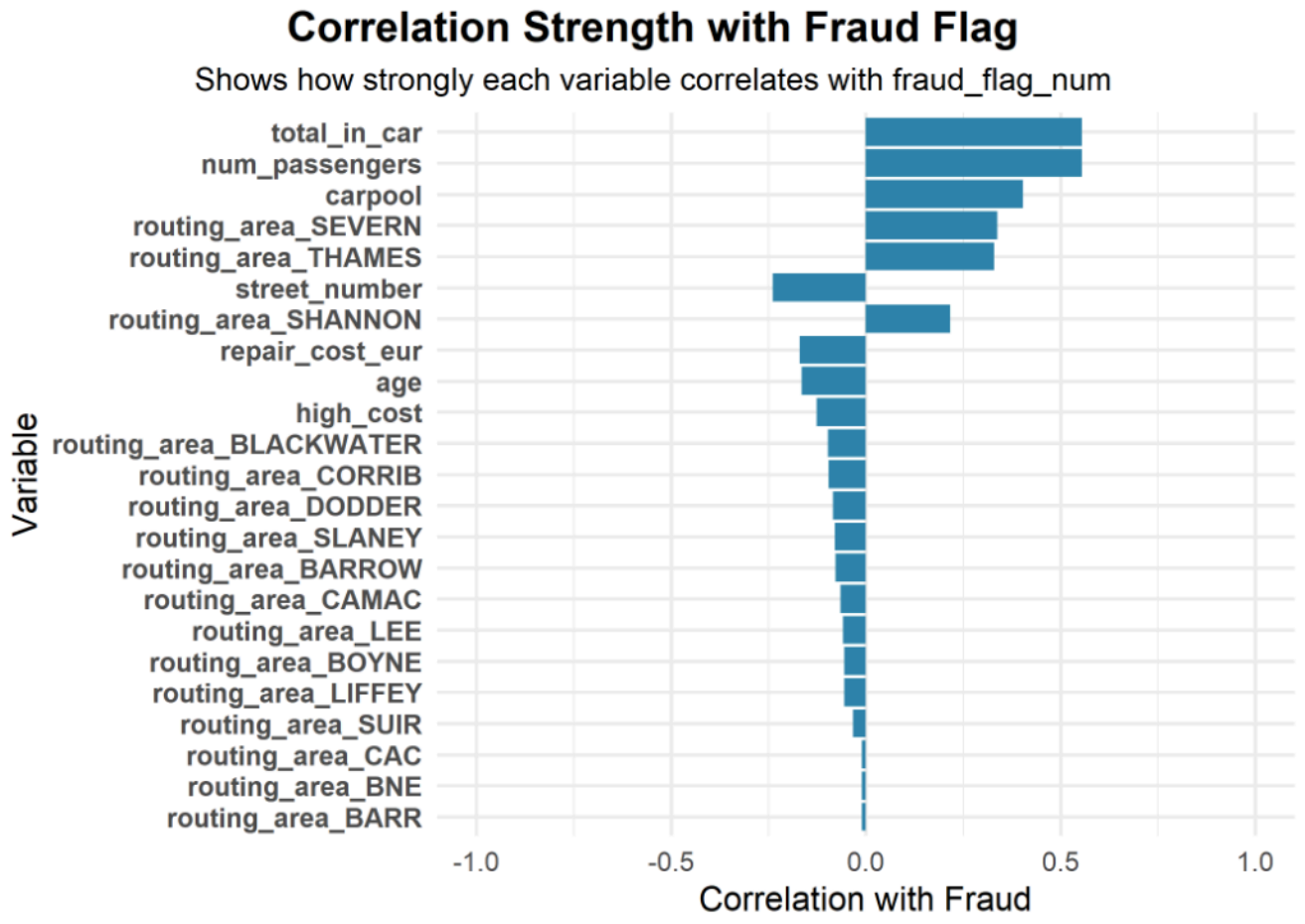

Figure 5. Correlation strength between selected numeric features and the fraud flag. Stronger correlations suggest higher predictive value.

The plot highlights `total_in_car`, `num_passengers`, and `carpool` as the three variables most strongly correlated with `fraud_flag_num`, all showing weak to moderate positive correlations. This suggests that as the number of occupants in a vehicle increases, so too does the likelihood of a fraudulent claim — potentially pointing to inflated or coordinated claims involving passengers.

Other positively associated variables include `repair_cost_eur` and `high_cost`, though their correlations with fraud are weaker. These results are consistent with the general expectation that more expensive repairs may be more closely scrutinised for fraud. However, as explored later in the report, the actual distribution of fraudulent claim costs reveals a more nuanced pattern — with many falling into a modest, mid-range cost band rather than the highest-cost bracket. A small number of `routing_area_*` features — such as `routing_area_SEVERN` and `routing_area_THAMES` — also show weak positive correlations, hinting at potential spatial patterns worth further investigation.

`street_number`, while appearing mid-ranked in the visual, shows no meaningful correlation with `fraud_flag_num` and is likely unrelated to fraud. Its position may reflect noise rather than signal — a reminder that not all patterns are meaningful, and care must be taken to avoid being misled by randomness. Similarly, `age`, `repair_cost_eur`, and most `routing_area_*` variables show near-zero correlation and are unlikely to be strong predictors on their own.

It's worth noting that **`age` as a continuous variable shows weak linear correlation with fraud**, but this doesn't rule out its importance. Later in the analysis, we explore age through **nonlinear groupings**, which reveal clearer fraud patterns and support the use of **banded age categories** — a common practice in insurance analytics.

5.2 Correlation Between All Numeric Features

While the first chart focuses on how each variable relates to fraud specifically, the matrix below provides a broader overview of how all numeric variables interact with one another. This helps spot redundant variables and uncover hidden relationships — important steps in understanding the structure of the data before any future modelling.

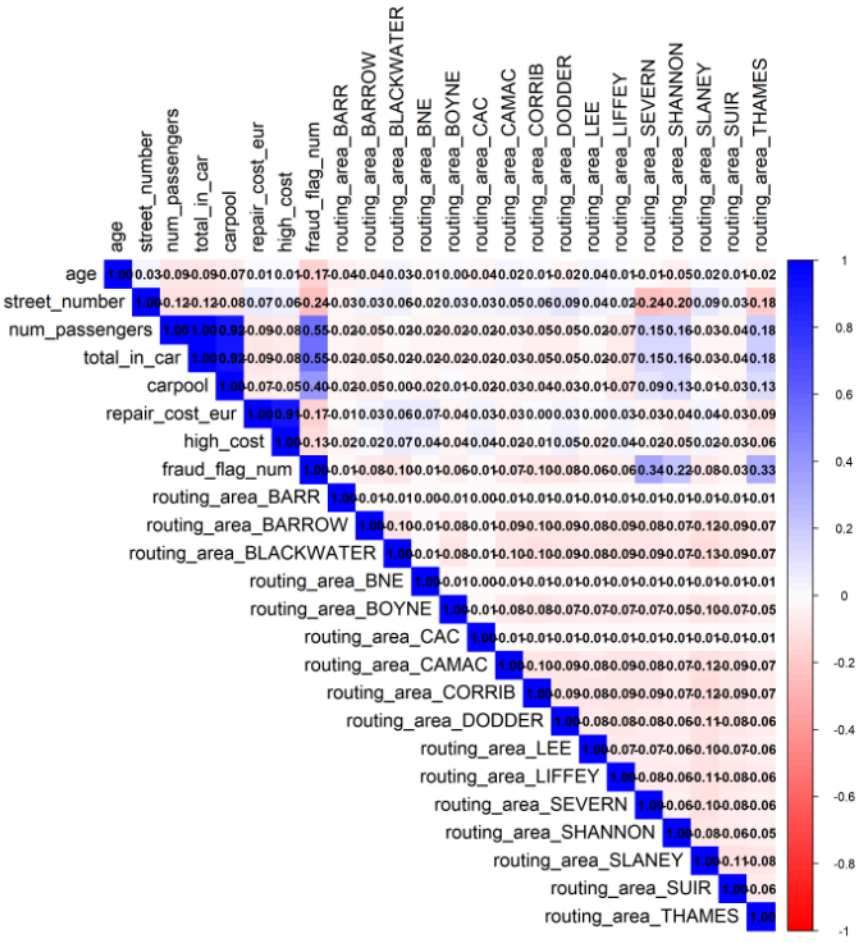

Figure 6. Correlation matrix showing relationships between all numeric variables. Darker colours and larger values indicate stronger correlations.

The correlation matrix highlights several features that are highly collinear by design (capturing similar information in different ways). For example, `total_in_car` and `num_passengers` have a correlation of 0.92, and both also strongly correlate with `carpool` (r = 0.91). These features all capture car occupancy in different forms. While this redundancy wouldn't affect EDA, it's worth noting for any future modelling. `carpool` may offer clearer interpretation, while `total_in_car` retains more detail.

Similarly, `repair_cost_eur` and `high_cost` are closely related (r = 0.91), with the latter acting as a simplified thresholded version of the former. The choice between them depends on whether a continuous or binary view of cost is more relevant for the analysis.

Most other variables — including `age`, `street_number`, and the `routing_area_*` features — show weak correlations with each other and with `fraud_flag_num`. This independence is useful in EDA, as it suggests these variables may contain distinct signals, particularly when analysed with grouping or non-linear

techniques.

5.3 Summary of Correlation Insights

- Car occupancy (`total_in_car`, `num_passengers`, `carpool`) and high-cost claims show the strongest associations with fraud.
- Strong correlations between some variables (e.g. `carpool` and `total_in_car`) highlight the need for **careful variable selection** in future modelling stages.
- Subtle but consistent spatial signals justify further exploration of **geographic patterns** in fraud.
- These insights guide **feature selection** and help frame the next steps in the analytics process.

6 People in Car vs Fraud

Understanding how the number of people in a vehicle impacts fraud likelihood can reveal behavioural patterns and inform risk profiling. This section explores the relationship between car occupancy — both as a count (`total_in_car`) and as a behaviour (`carpool`) — and the observed fraud rate.

6.1 Average Fraud Rate by Number of Occupants

The bar chart in **Figure 7** illustrates the fraud rate by total number of people in the vehicle (including the driver). A clear upward trend is observed: claims involving just the driver show a fraud rate of **1%**, which rises to **11%** with two people and spikes dramatically to **61%** with three occupants.

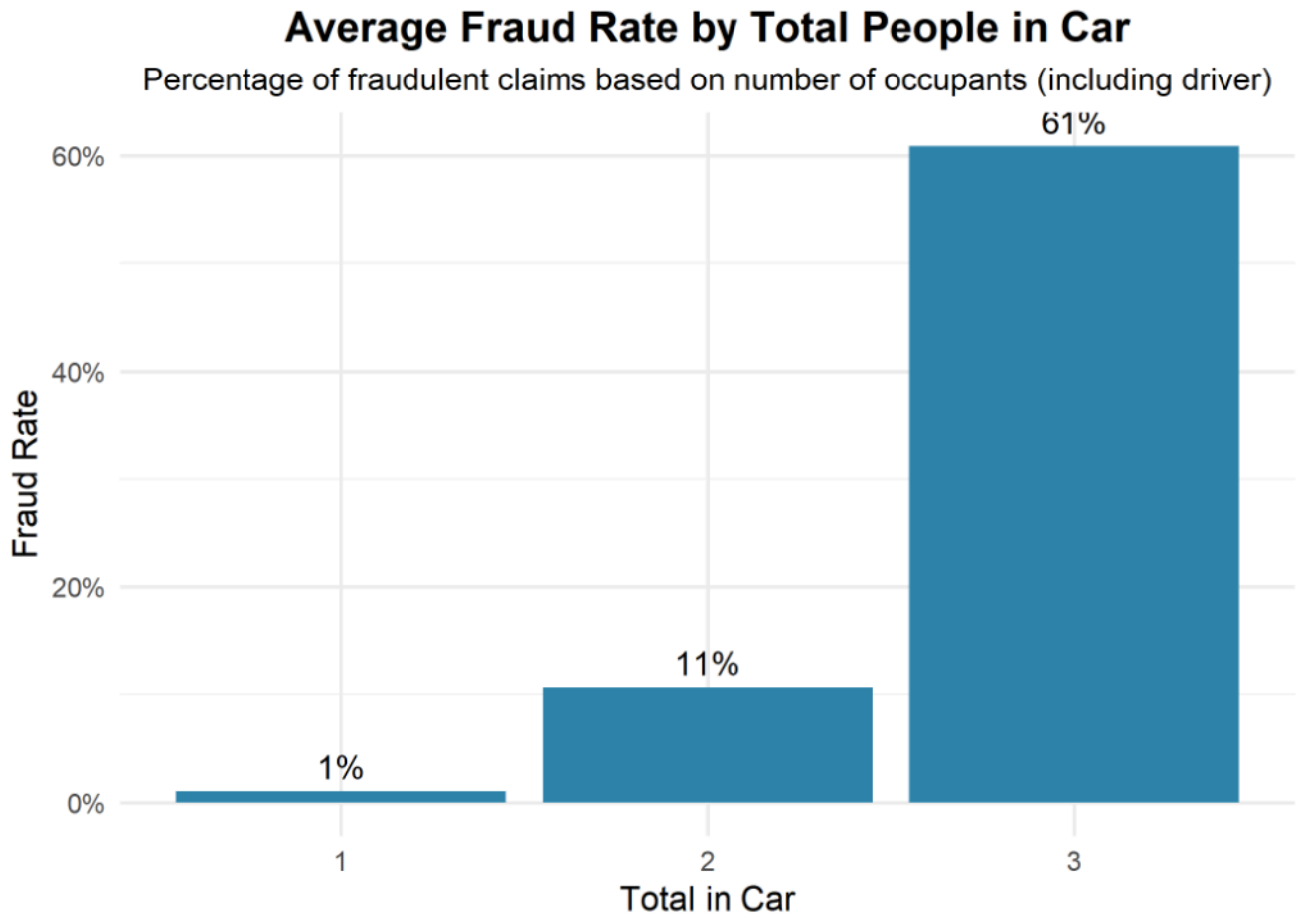

Figure 7. Average fraud rate based on total number of people in the car at time of claim.

This sharp increase suggests that multi-occupant claims — especially those with three individuals — are significantly more likely to be fraudulent. This insight supports earlier correlation findings and may reflect attempts to exaggerate injury claims or fabricate witness support. `total_in_car` stands out as a particularly insightful feature for understanding fraud patterns and may hold strong potential for future modelling.

6.2 Carpooling and Fraud Likelihood

Figure 8 explores a related but simplified version of the previous analysis, grouping drivers into two categories: those **not carpooling** (i.e., solo) and those **carpooling** (i.e., with at least one passenger). The fraud rate increases from **1%** for solo drivers to **26%** for carpooling cases.

While similar to Figure 7 on `total_in_car`, this chart presents a **binary view of car occupancy** rather than exact counts. Both visuals reinforce the same pattern — that shared trips are more frequently linked with fraud — but do so at different levels of detail.

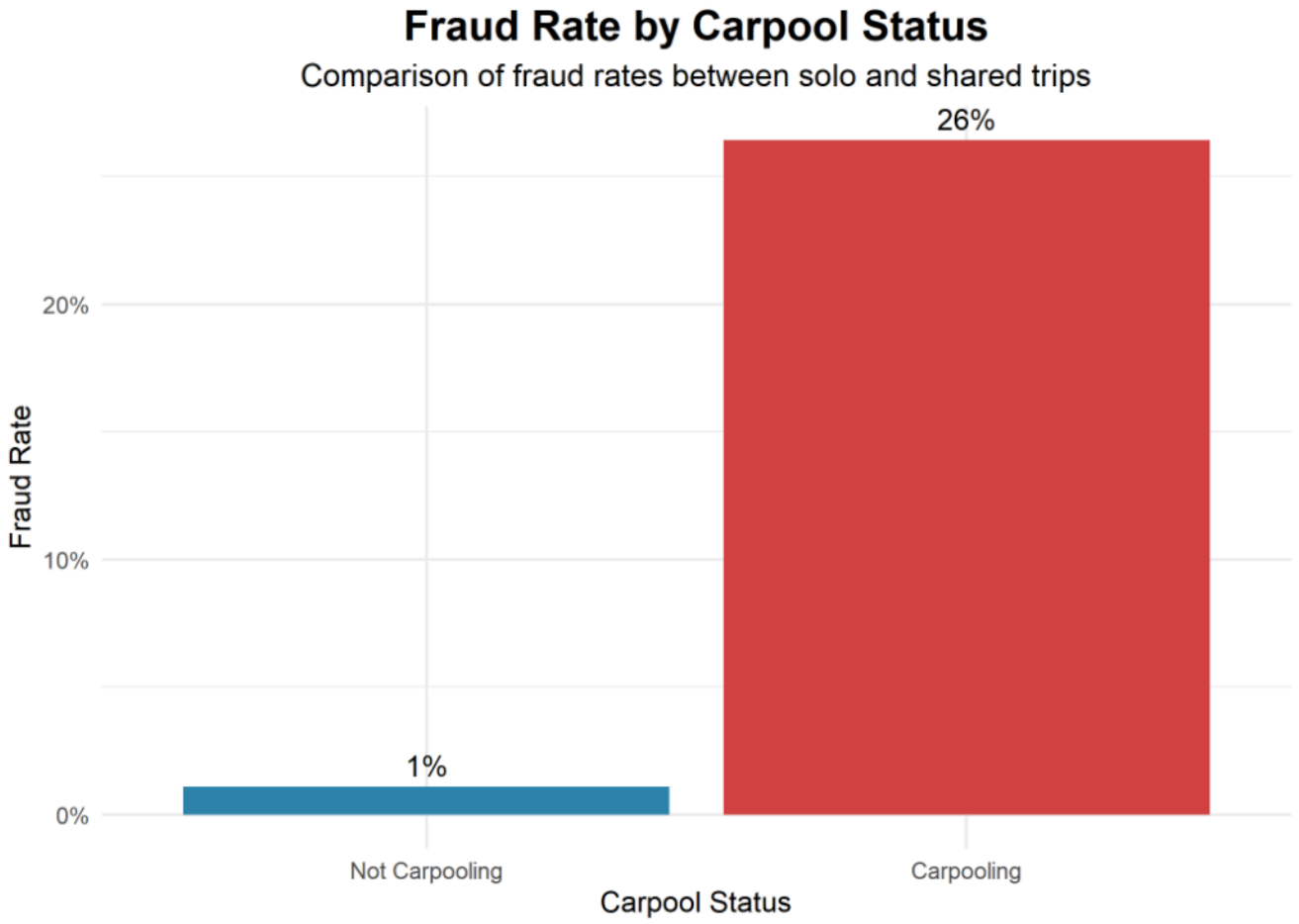

Figure 8. Comparison of fraud rates between solo drivers and those travelling with passengers.

Although this binary breakdown lacks the nuance of `total_in_car`, it offers greater interpretability and practical value. The elevated fraud risk in carpooling cases may stem from reduced personal accountability, coordinated exaggeration of injuries, or inclusion of fictitious passengers.

6.3 Key Insights from People in Car vs Fraud

- **Sharp risk gradient**: Fraud likelihood increases steeply with each additional person — especially from 2 to 3 people — suggesting a tipping point for potential collusion.

- **Carpool flag effectiveness**: The binary `carpool` feature retains strong explanatory value and offers a clean input for segmentation or rule-based systems.
- **Model design implication**: Both `carpool` and `total_in_car` are valuable. While collinear, they offer different advantages — `total_in_car` for richer insights, `carpool` for simplicity and speed. As seen in the correlation matrix (Figure 6), these features share a correlation of **0.91** — above the common **0.9 threshold** — suggesting that only one should be selected for future modelling to avoid confusing the model or introducing redundancy.

7 Geographic Patterns in Fraudulent Claims

Understanding how fraud rates vary across different routing areas can help identify regional risk factors. Certain locations may experience elevated fraud due to factors such as local claim culture, socioeconomic conditions, or inconsistencies in claim verification processes.

7.1 Fraud Rate by Routing Area

The bar chart in **Figure 9** shows the average fraud rate for each routing area. Fraud appears to be highly concentrated in a few locations. `routing_area_THAMES` has the highest fraud rate at **56%**, followed by `routing_area_SHANNON` at **40%** and `routing_area_SUIR` at **7%**. In contrast, the majority of other routing areas show very low fraud rates, often at or near **0%**.

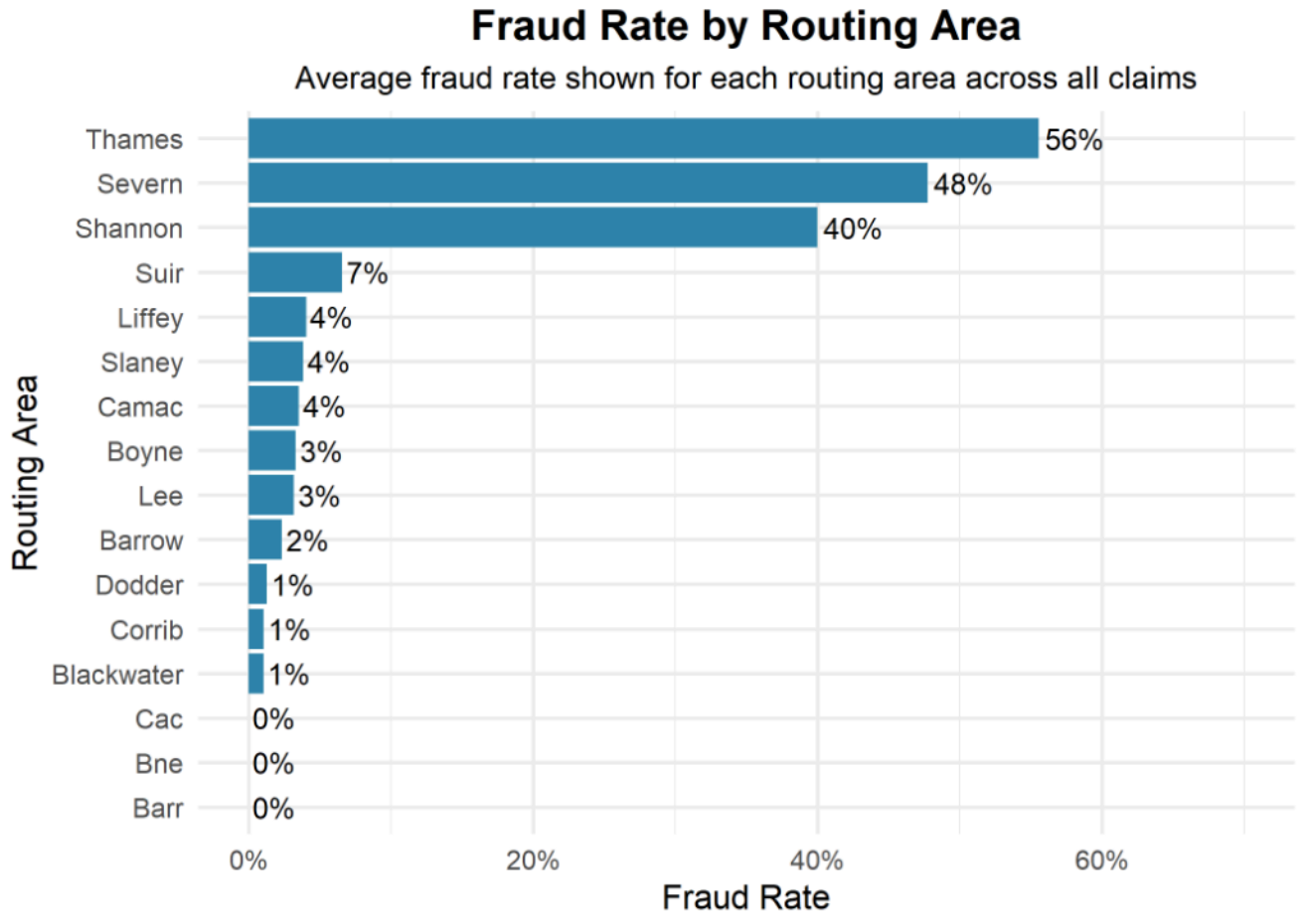

Figure 9. Average fraud rate by routing area across all claims.

These findings suggest that a small number of geographic areas account for a disproportionate share of fraud cases. This insight may support the development of targeted fraud-prevention efforts or additional scrutiny in high-risk regions.

7.2 Fraud Rate by Routing Area and Carpooling Status

To further explore regional dynamics, the chart in **Figure 10** breaks down fraud rates by routing area and `carpool` status. This split allows us to examine whether the elevated risk in some areas is amplified or explained by shared travel behaviour.

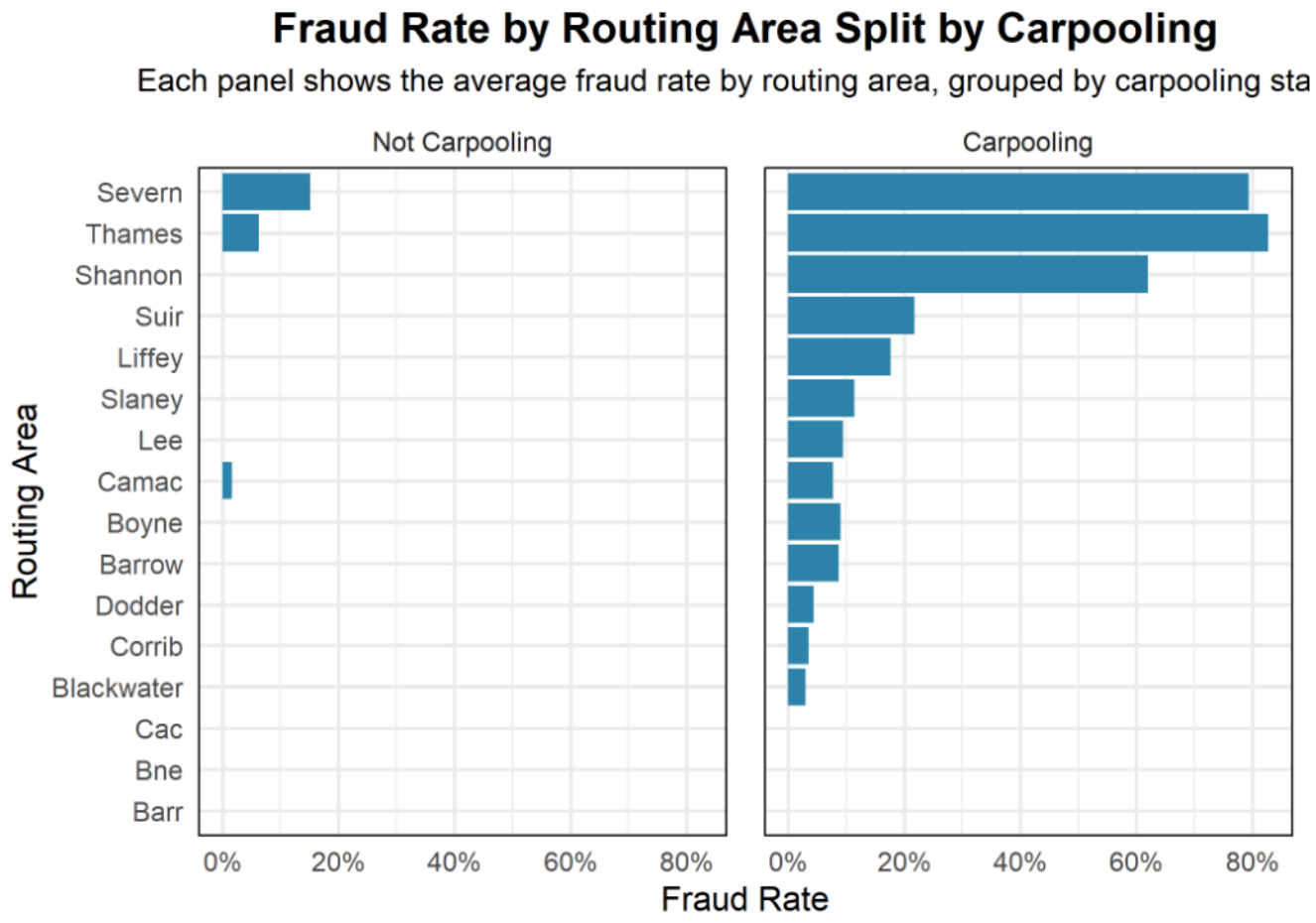

Figure 10. Fraud rate by routing area, split by carpooling status.

The right-hand panel (Carpooling) reveals that `routing_area_THAMES`, `SEVERN`, and `SHANNON` have significantly higher fraud rates when trips involve multiple occupants. For instance, `THAMES` shows a fraud rate of over **75%** for carpooling claims — far higher than the non-carpooling equivalent. In contrast, most routing areas with lower overall fraud see little to no difference by carpooling behaviour.

This suggests that in high-risk regions, **carpooling may amplify fraud exposure**, potentially indicating coordinated or opportunistic claims involving multiple passengers. These nuanced interactions between geography and behaviour may prove valuable in building risk-aware, region-specific fraud models.

7.3 Key Insights from Geographic Patterns in Fraudulent Claims

- **Fraud is geographically concentrated**: Just a few routing areas (e.g. `THAMES`, `SHANNON`) drive most of the fraudulent activity.
- **Shared trips intensify regional risk**: Carpooling increases fraud likelihood in high-risk areas, reinforcing the need to consider behavioural context in fraud analysis.
- **Operational value**: These findings could inform resource allocation, claim verification priorities, or the deployment of region-specific fraud detection thresholds.

8 Age-Related Insights

Age can be a critical factor in understanding behavioural patterns in fraudulent claims. This section explores how age relates to fraud using a combination of histograms, fraud rate comparisons, proportional breakdowns, and cost analysis.

8.1 Distribution of Fraudulent Drivers by Age

The histogram in **Figure 11** shows the age distribution of drivers involved in fraudulent claims. Most cases involve individuals aged between **25 and 40**, with visible peaks in the **25–30** and **35–40** age bands. Fewer fraud cases occur at the extremes of the age spectrum, particularly among older drivers.

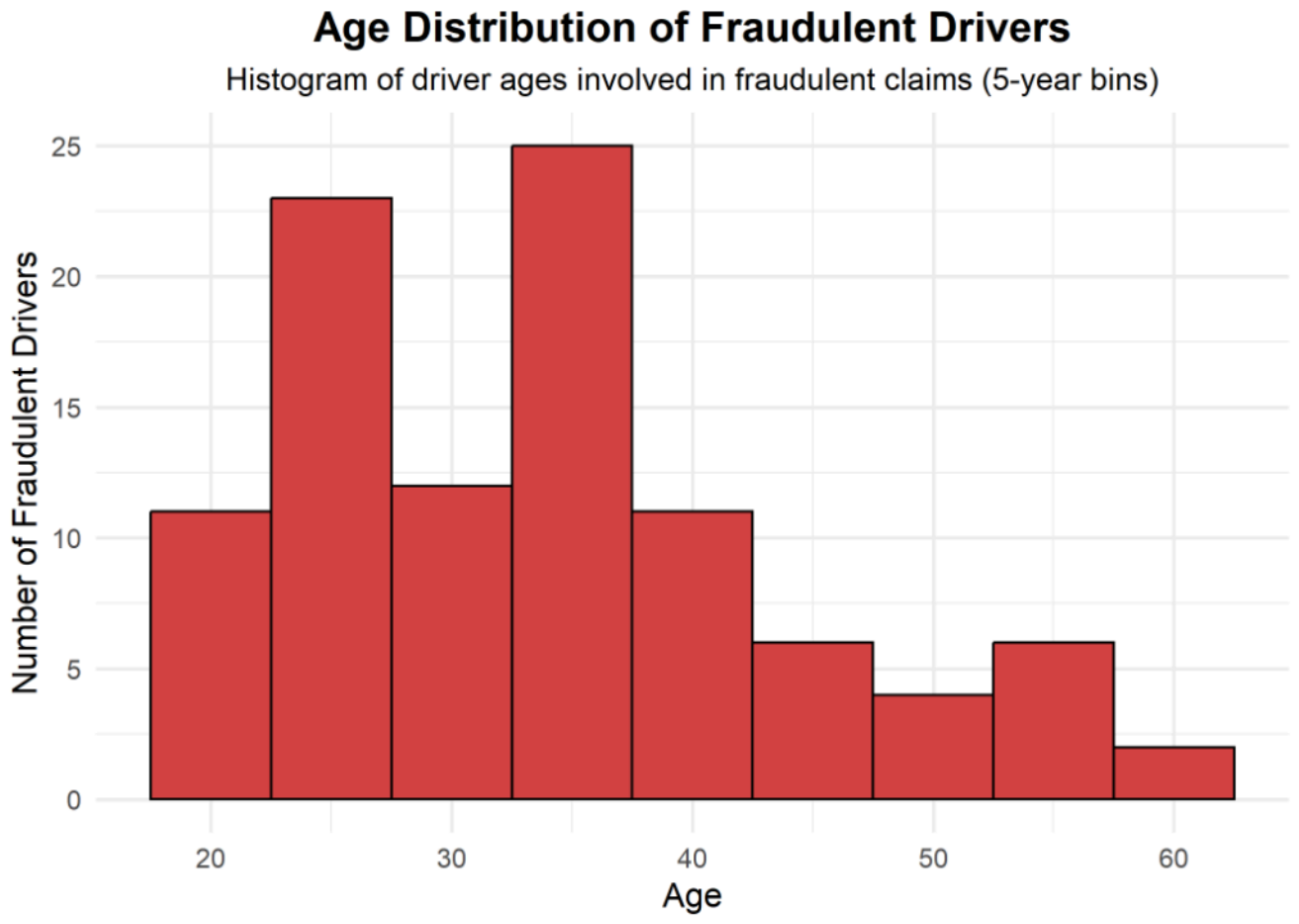

Figure 11. Histogram showing number of fraudulent drivers across 5-year age bins.

This distribution suggests that fraud is most common among drivers in their **working-age years**, possibly linked to financial pressure or higher driving frequency. It highlights the need to explore not just raw counts but fraud rates relative to age group size.

8.2 Fraud Rate Within Each Age Group

Figure 12 compares the **fraud rate** across three age categories: `under_25`, `25_to_65`, and `over_65` (note that these are different bins than the ones used in the previous histograms). While the `under_25` group shows the highest fraud rate (**14.7%**) *within its age bracket*, the majority of fraud costs overall are concentrated in the `25_to_65` and `over_65` groups.

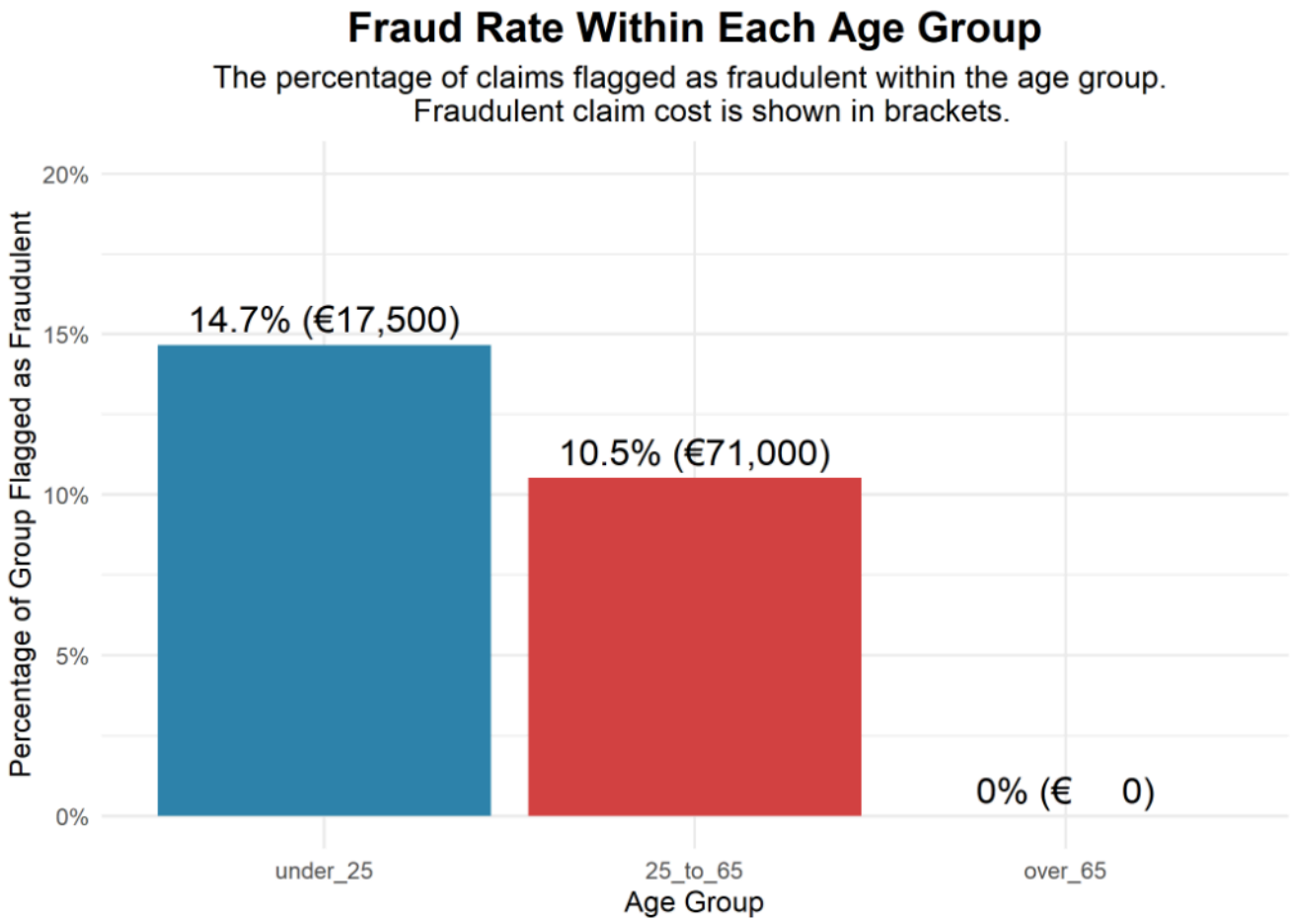

Figure 12. Percentage of claims flagged as fraudulent within each age group, along with associated claim cost.

This confirms that while younger drivers are **more likely to commit fraud**, the overall **cost of fraud** is driven by the larger middle-aged group.

8.3 Contribution to Total Fraud by Age Group

To understand the broader impact, Figure 13 shows each group's contribution to total fraud cases. Although younger drivers have a higher individual fraud rate, the `25_to_65` group accounts for 78% of all fraud cases, contributing over €71,000 in flagged claims. This indicates that while the likelihood of fraud

is highest within the `under_25` group (as shown in Figure 12), the majority of fraud cases and associated costs come from the `25_to_65` group.

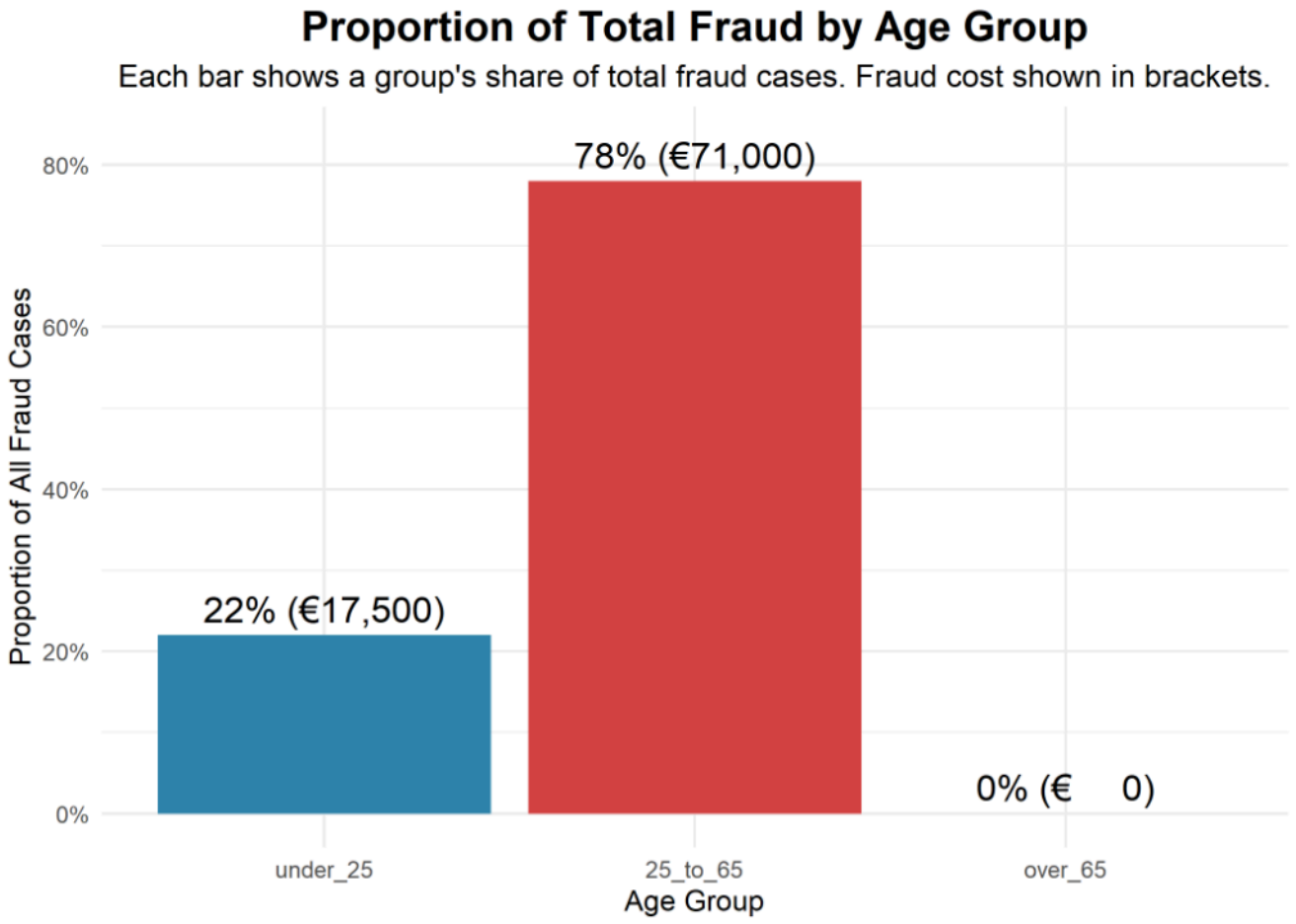

Figure 13. Each age group's contribution to total fraudulent claims, with total cost in brackets.

This reinforces the importance of targeting interventions or fraud checks within the dominant `25_to_65` group, despite the risk signal in younger drivers.

8.4 Repair Costs by Age Group and Fraud Status

The boxplot in **Figure 14** compares **repair costs** across age groups for both fraudulent and non-fraudulent claims. In non-fraud cases, the median and spread of repair costs appear similar across age groups. However, in fraudulent claims, the distributions are **more uniform**, with fewer outliers.

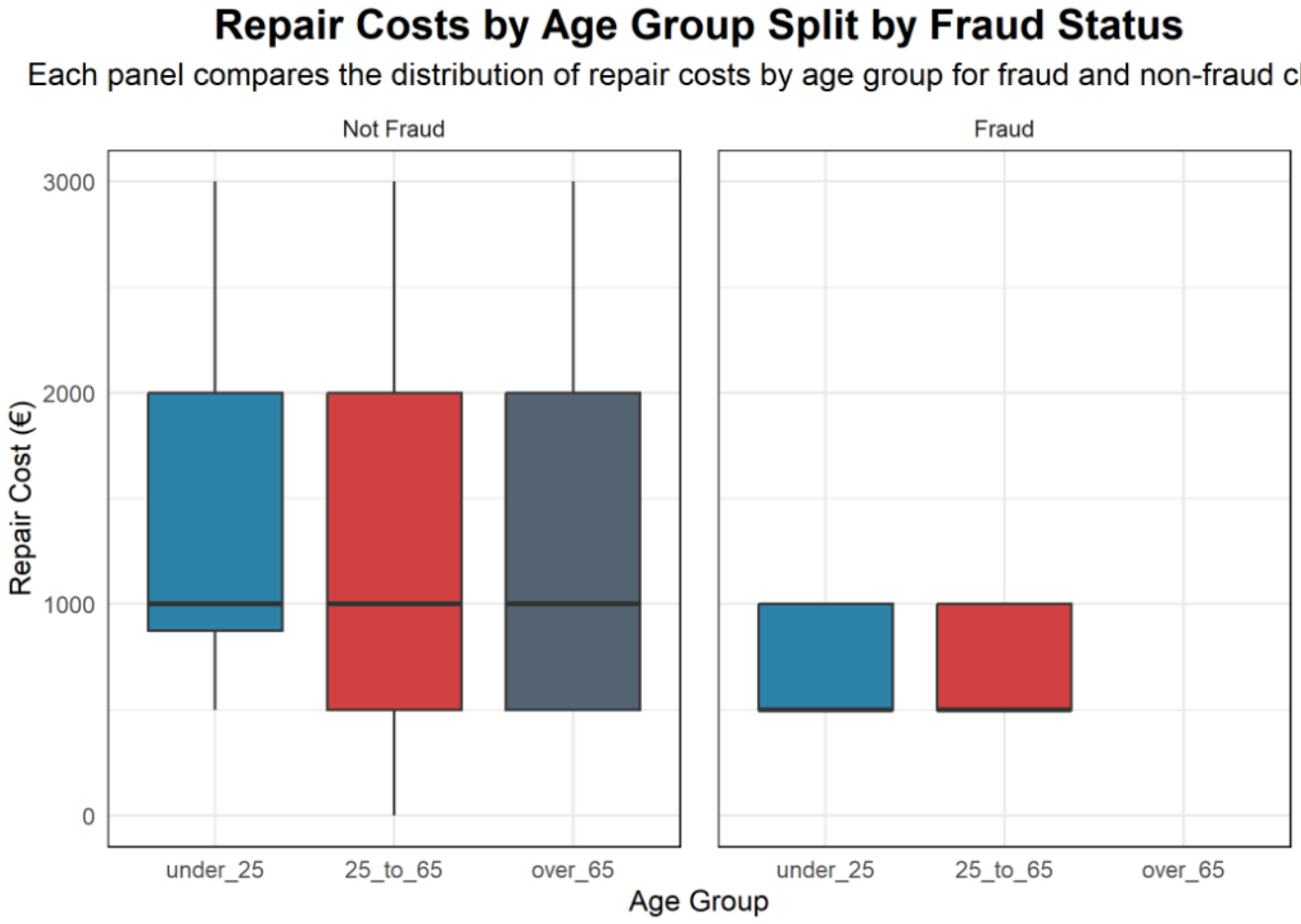

Figure 14. Distribution of repair costs across age groups, split by fraud flag.

Earlier, in **Figure 5**, we observed that higher repair costs were weakly correlated with increased fraud, suggesting that more expensive repairs might warrant closer scrutiny. However, in **Figure 14**, we see a different trend: while legitimate claims show significant variability in repair costs across age groups, fraudulent claims appear **more standardised**, particularly in the `under_25` and `25_to_65` groups. Interestingly, **no fraudulent claims were observed in the over_65 group**, reinforcing earlier findings that fraud is not common in this age bracket.

8.5 Key Insights from Age-Related Insights

- **Younger drivers (<25)** are more likely to commit fraud proportionally, but represent a smaller share of total fraud cost.
- **Drivers aged 25–65** pose the **greatest operational risk**, accounting for the vast majority of fraud cases and financial impact.
- **Older drivers (65+)** appear to present **minimal fraud risk**, with no recorded fraudulent claims in the dataset.
- **Fraudulent repair costs** are more tightly clustered than non-fraudulent ones, which may indicate efforts to standardise or limit claims amounts.

These findings offer valuable direction for segmentation strategies in fraud modelling, helping to balance proportional risk and real-world impact.

9 Repair Cost Patterns

9.1 Repair Cost Distribution by Fraud Status

The boxplot in **Figure 15** compares repair costs between fraudulent and non-fraudulent claims. Interestingly, the **median repair cost for fraudulent claims is lower**, and the overall spread is narrower than in legitimate claims. This suggests that fraudulent claims are intentionally kept modest to avoid drawing attention or triggering audit thresholds. This pattern aligns with earlier findings (**Figure 12** and **Figure 13**) where fraudulent claims were found to be more prevalent among younger drivers (`under_25`), but when focusing on costs, fraud is often concealed by keeping repair expenses lower. These age-related patterns highlight the need to integrate both cost and behavior-based insights in fraud detection.

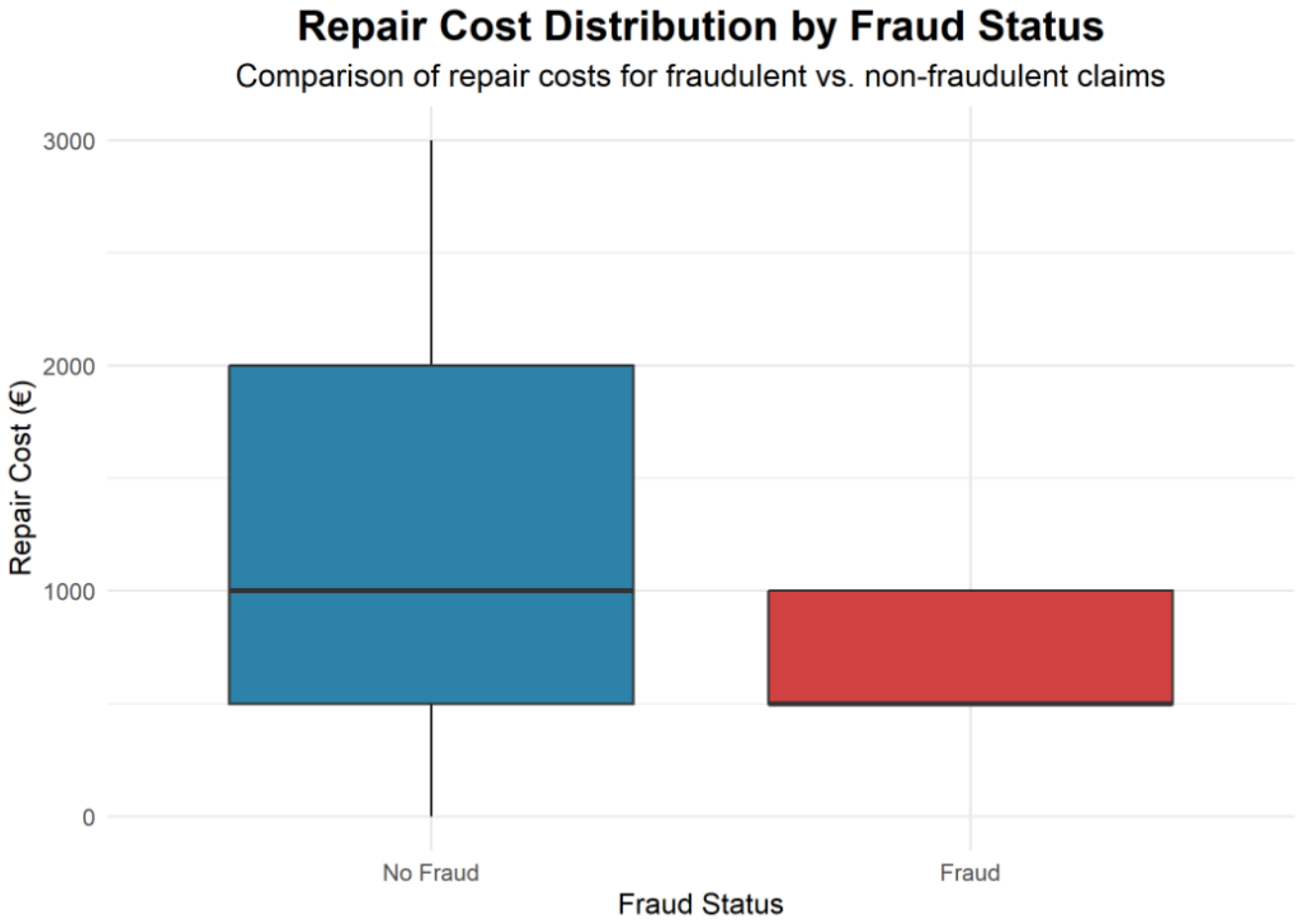

Figure 15. Boxplot comparing repair costs for fraud vs. non-fraud claims.

This finding suggests that fraudulent claims may often be **strategically kept below high-cost thresholds** to reduce scrutiny or bypass manual review. It complements the earlier observation that while higher repair costs can be linked to fraud, **lower-cost claims may be used to disguise intent**, underscoring the importance of combining pattern-based detection with cost considerations.

10 Repeated Names Check

10.1 Repeated Individuals Involved in Claims

The bar chart in **Figure 16** highlights individuals who appear **more than three times** across claims — either as a driver or passenger. A dashed threshold at 3 appearances marks the point beyond which repetition may be considered suspicious.

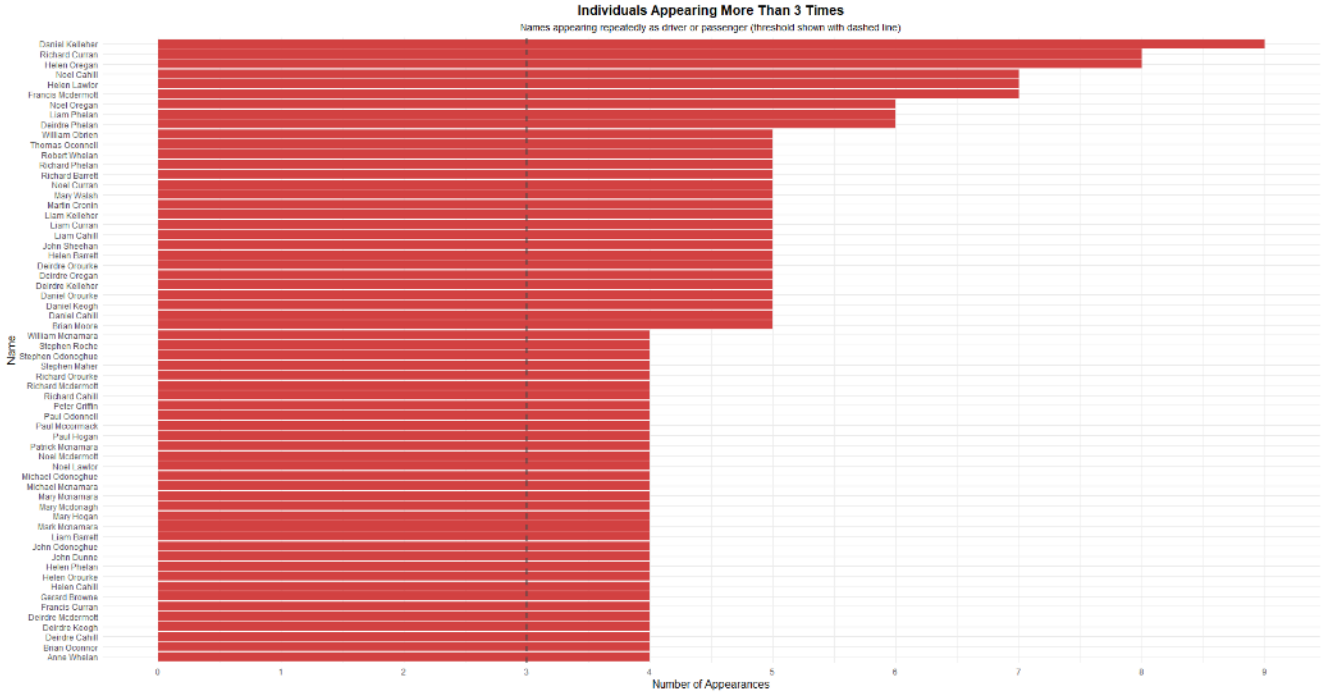

Figure 16. Individuals involved in more than three claims, either as drivers or passengers.

The analysis reveals multiple individuals with **four or more appearances**, including a small number with **six to nine claims**. These repeated appearances may indicate:

- **Legitimate repeat claimants**, such as fleet drivers or frequent passengers, or
- **Potential organised fraud**, where the same individuals are repeatedly involved in suspicious claims.

While further investigation would be needed to confirm intent, this pattern-based flag serves as a **strong operational trigger** for deeper review. Integrating name frequency checks into future fraud models or alert systems could support proactive fraud prevention.

This pattern of repeated names adds to earlier findings about **carpooling** and **fraud hotspots**, showing that looking at how often someone appears in claims can help spot fraud — especially when combined with where the claims happen and how many people are involved.

11 Conclusion and Next Steps

This exploratory data analysis uncovered several meaningful patterns that improve our understanding of fraudulent motor insurance claims and inform the design of future predictive models.

11.1 Key Findings

1. Car occupancy is a major fraud signal
Features such as `total_in_car`, `num_passengers`, and `carpool` consistently showed the strongest associations with fraud. Shared trips, especially those involving three or more individuals, were linked to significantly higher fraud rates. This suggests possible collusion or fabricated injury claims.

2. Fraud isn't always expensive
While high-cost claims were moderately correlated with fraud, the boxplot analysis revealed that many fraudulent claims cluster in lower-cost ranges, with a narrower spread. This may reflect deliberate underreporting to avoid triggering audit thresholds — something cost-based detection alone might miss.

3. Fraud is geographically concentrated
Areas like `THAMES`, `SHANNON`, and `SEVERN` show disproportionately high fraud rates. When combined with carpooling, these hotspots become even more pronounced — suggesting that shared-travel dynamics and regional factors together may drive coordinated fraud.

4. Age patterns matter
Although drivers under 25 had the highest fraud rate within their group, the 25–65 group was responsible for the majority of fraud cases and monetary impact. Notably, no fraud was recorded among over-65s. These insights support age-informed segmentation strategies.

5. Repeated names may signal organised fraud
Several individuals appeared in multiple claims — some up to six or more times. This repetition may point to organised fraud networks. While more investigation is needed, repetition-based checks could be valuable operational triggers or model inputs.

11.2 Implications and Next Steps

Together, these insights demonstrate the importance of combining **behavioural**, **cost-based**, **geographic**, and **demographic** patterns in fraud analysis. Engineered features like `carpool`, `high_cost`, and `age_group` added significant value and should be included in future model pipelines.

The next phase of this project involves developing and testing classification models. Emphasis will be placed on balancing model accuracy and interpretability, evaluating the predictive power of these engineered features, and ensuring any solution works in real-world claims handling to support proactive fraud detection.