Exploratory Data Analysis Of Insurance Claims Dataset

Sumit Trivedi – k00302092

2024-10-12

## Problem Statement

In today’s competitive Car Insurance Industry landscape, effectively managing and analyzing insurance claims data is crucial for minimizing fraud, and enhancing customer satisfaction.The Insurance Company Name has commissioned an exploratory data analysis (EDA) of its car insurance claims Dataset to uncover valuable insights and identify potential areas for improvement.

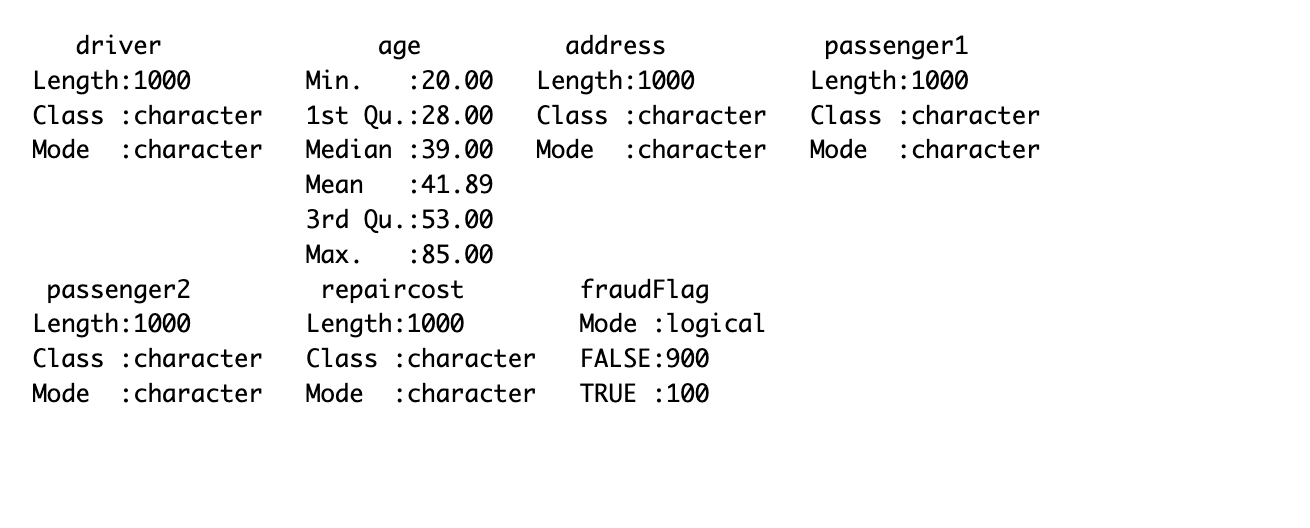
This Report aims to provide comprehensive insights into the company’s car insurance claims data by conducting a thorough exploratory data analysis. By leveraging feature engineering techniques, we will enhance the dataset to uncover hidden patterns, trends, and anomalies.

## Solution Summary

The key outcomes and solutions derived from this car insurance claims Dataset analysis are summarized as follows:

1. **Data Quality Improvement**: Conducted a thorough evaluation of the dataset to identify and address missing, inconsistent, or erroneous data. Enhancement: Cleaned the data to improve its accuracy and reliability, ensuring it is suitable for in-depth analysis
2. **Descriptive Analysis**: Provided a detailed summary of the dataset, including mean, median, mode, standard deviation, and distribution of key variables.
3. **Visualization**: Created visual representations (e.g., histograms, box plots, scatter plots) to understand the data distribution and relationships better.
4. **Fraud Detection Insight**: Identified key indicators and patterns associated with potentially fraudulent claims.
5. Created new, meaningful **features** that enhance the dataset’s richness and improve the predictive power of future models.
6. **Conclusions** highlighting the insights you have gained

## Descriptive Statistics:-



Summary statistics - overview of our dataset

The dataset consists of 1,000 observations and includes the following variables: driver, age, address, passenger1, passenger2, repaircost, and fraudFlag. Below is a detailed summary of each variable:

1. **Driver**:

* **Length**: 1,000
* **Class**: Character
* **Description**: This variable represents the name of the driver involved in the claim.

1. **Age**:

* **Minimum**: 20.00
* **1st Quartile**: 28.00
* **Median**: 39.00
* **Mean**: 41.89
* **3rd Quartile**: 53.00
* **Maximum**: 85.00
* **Description**: The age of the drivers ranges from 20 to 85 years, with a median age of 39 and a mean age of 41.89.

1. **Address**:

* **Length**: 1,000
* **Class**: Character
* **Description**: This variable contains the addresses of the drivers.

1. **Passenger1**:

* **Length**: 1,000
* **Class**: Character
* **Description**: This variable represents the name of the first passenger in the vehicle, if present.

1. **Passenger2**:

* **Length**: 1,000
* **Class**: Character
* **Description**: This variable represents the name of the second passenger in the vehicle, if present.

1. **Repair Cost**:

* **Length**: 1,000
* **Class**: Character
* **Description**: This variable indicates the estimated cost of repairs for the vehicle, with various formats that have been standardized during the data cleaning process.

1. **Fraud Flag**:

* **Mode**: Logical
* **False**: 900
* **True**: 100
* **Description**: This binary variable indicates whether a claim is suspected of being fraudulent (TRUE) or not (FALSE). Out of 1,000 claims, 100 are flagged as fraudulent.

With **skim()** we can find a more detailed statistical analysis of each column with the count of missing data as shown in [Figure 1](#fig-Skim-functions-stats) which is very useful moving further.

|  |
| --- |
| Figure 1: Skim() function Stats |

## Data Cleaning:

Data Cleaning in R involves the process of identifying and correcting errors, inconsistencies and missing values in the given dataset to improve its quality and accuracy.

Data cleaning for the claim\_fraudflag\_data dataset included several steps to ensure the data is accurate, consistent, and ready for analysis. Specific details of what was done during the data cleaning phase are outlined below:

1.**Repaircost column Data Cleaning**

**str\_replace\_all()**: This function replaces some characters and patterns in the repaircost column with a standard value.

* Replaced the characters [k~] with 000, converting shorthand representations of thousands, such as “approx 2k”, to their full numeric form, such as “2000”.
* Replaced the characters [!,\*] with 0 to account for anomalies and mistakes within the data.
* Replaced the character [$] with 5, which was a manual correction in the dataset.

**str\_trim()**: After the replacement, str\_trim() was applied to **remove leading or trailing spaces** created by these replacements.

This step ensured the values in the repaircost column were clean and consistent.

**as.numeric()** : Converting the column values to numerical values to understand the data in a better way.

[Figure 2](#fig-Distinct-Repaircost-values) shows the distinct values before cleaning and can clearly be observed that there are inconsistencies and then the [Figure 3](#fig-Distinct-Cleaned-Repaircost-Values) values after cleaning which are more readable and easy to use.

|  |
| --- |
| Figure 2: Distinct values in repair cost column before Cleaning |

|  |
| --- |
| Figure 3: Distinct values in repair cost column after cleaning. |

1. **Passenger column Data Cleaning**:

**mutate()** and **case\_when()**: To manage missing data and create meaningful columns. passenger 1 and passenger 2 columns were converted to binary. These columns indicate the presence of passengers based on whether the respective values (passenger1, passenger2) are NA. This was an important step to handle missing values in a structured way. 1 indicates the passenger present on the other hand 0 indicate no passenger.

The Below Figure [Figure 4](#fig-passenger-before-cleaning) shows missing values in both passenger columns whereas [Figure 5](#fig-passenger-after-Cleaning) shows the passenger columns after cleaning which makes them ready for analysis.

|  |
| --- |
| Figure 4: Values in passenger Column before Cleaning |

|  |
| --- |
| Figure 5: passenger column after Cleaning |

3.**Address column Data Cleaning**

**str\_replace\_all()**: On the address column, multiple patterns that should be replaced by standardized counterparts, such as special character regex [R/], [!][!], [K%], etc., standardized as RR, LIF, VER, etc. It has been an important step as it had to deal with addresses having irregular characters or formatting on many occasions.

Several steps of str\_replace\_all() were run consecutively to deal with a number of different potential patterns and anomalies to regularize the address values for consistency and accuracy in the dataset.

4.**Driver column Data Cleaning**

Similary on driver column some patterns were recognized that needed to be cleaned. Used str\_replace\_all function to do the cleaning for driver names.

**Key Functions Used**:

**str\_replace\_all()**: This was a core function used for cleaning, replacing multiple patterns or characters within string columns for standardization of data and dealing with inconsistencies.

**mutate()**: Used mainly for creating new variables or even changing the values of old ones. It was crucial to be used for creating binary flags indicating passenger presence.

**case\_when()**: A conditional function where values are assigned depending on logical conditions; this is utilized to create flags based on passenger presence. **str\_trim()**: This function was applied after string replacements to remove unwanted spaces and ensure that the repaircost column was free from extraneous whitespace.

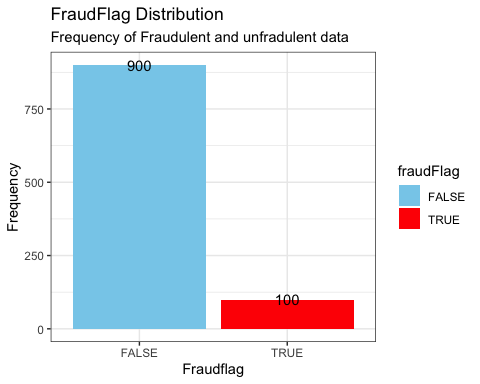
**as\_numeric()** : A column was converted to numerical datatype for better understanding our dataset.

Cleaning the data was a crucial step in transforming raw data into a structured and analyzable form. This was done to standardize the dataset, handle missing values, and correct errors within the data. This means the dataset is now ready for detailed analysis and modeling with consistent and accurate values in the key columns.

## Engineered Features:

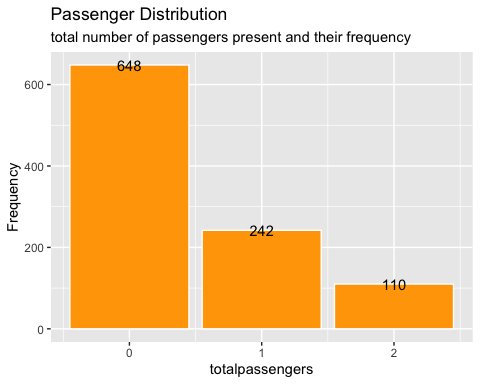
**total\_passengers**: Both the pre-existing passenger1 and passenger2 column indicating the presence of passengers with 1’s and 0’s have been converted to numerical type and after the conversions a new column was generated as a part of feature engineering to make our analysis more insightful with the name total\_passengers by doing arithimatic sum of passenger1 and passenger2 columns indication the total number of passengers presents.

## Exploratory Visualisations



Frequency of Fraudulent & Unfradulent Data

It can be observed from the above bar chart of **FraudFlag Distribution** that out of 1000 rows of claims data in our dataset their are a total of **100 Claims which were flagged as frauds**. The red color bar represents the fraudulent data and its count.



From the two Visualization of *Passenger distribution and FraudFlag by passenger distribution* we can see that In the dataset given there are

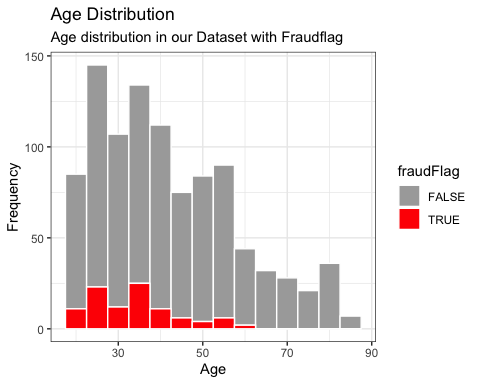
*648 claims which had no passengers* sitting with the driver out of which **7 were flagged as fraud** ,

*242 claim applications said to have 1 passenger* with the driver from which **26 were flagged as fraud** ,

*11% i.e. 110 of the claim applications had 2 passengers* with the driver and **67 were flagged as Fraud**.

can be interpreted from above & below bar graphs.



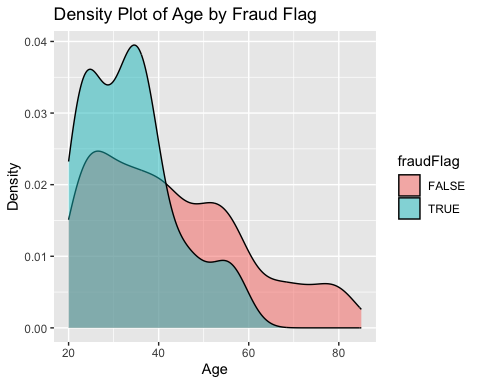


The **age distribution is approximately right skewed** with more individuals in age groups 22-40.

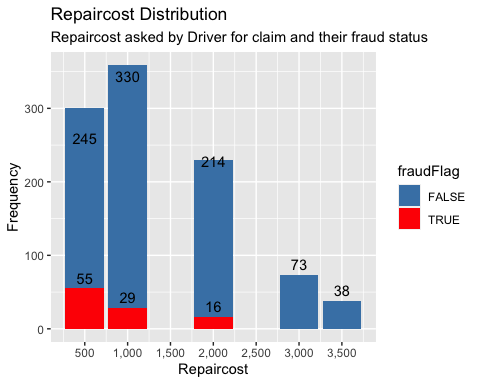
Across all age groups the red bars(frauds) are less as compared to the grey bars.

It can also be observed that **younger age groups show a higher count of frauds**(red bars).

Fraud cases seems to diminish as age increases particularly after 55 where fraud count dissapears.



The Above shown density plot also represents that as the **Age increases the fraud cases decreases and diminishes** at around a certain age.Fraud cases are almost negligible for individuals over 60.Younger individuals are more susceptible to fraud, as suggested by the higher density of fraud cases in these age groups.

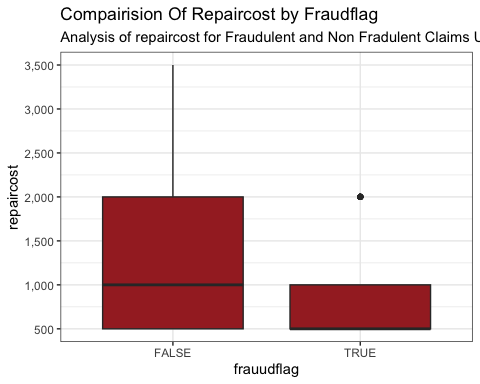


The highest number of claims fall in the range of 1000 repaircost, with a total of 330 claims. This is followed by 500 and 2000 with a total of 245 and 214 claims, respectively. This would mean that repaircost of 1000 might be a typical claim for repairs.

We can also observe that there is no data present for claims for repaircost of 1500 and 2000. which is a very unusual behaviour

Taking about fraudulent and non-fraudulent claims as the **repair cost increases the fraudulent cases decreases and there are no frauds recorded for repaircost above 2000 i.r for 3000 and 3500 range repaircost.**

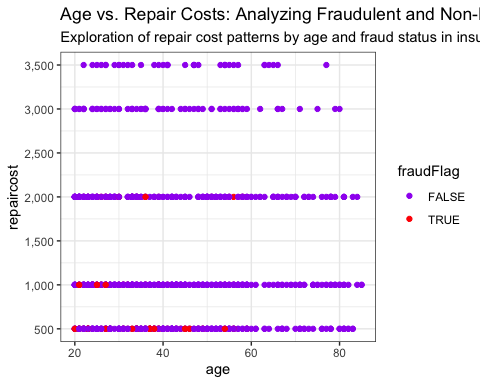
We can conclude that most of the **fraudsters are asking for low repaircost** which is a interesting and more suitable matrix.



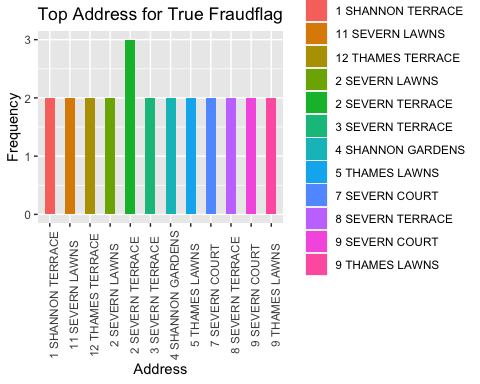
The Repaircosts are higher on average for cases flagged as false(non-fraudulent).

Their is a noticeable **outlier in the fraudulent cases(True)** indicating one or two fraudulent cases have incurred for much higher repaircost.

The median for non-Fraudlent Cases(false) is around 1000 which is highers then that of fraudulent cases(flagged as true).



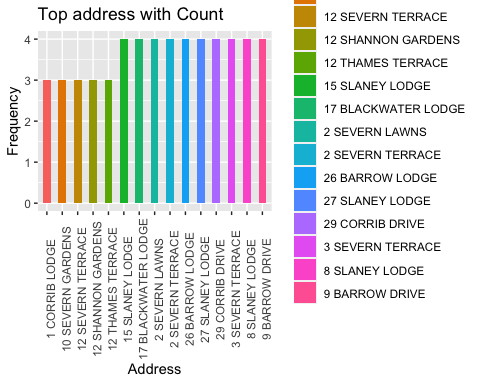
From the above **Age vs Repaircost** scatter plot it can be concluded that there is no clear trend or correlation between repaircost and age. We can only observe that there are no fraudulent claims in repair cost greater than 2000 as there are no red dots presents above this point and can also be concluded that most of the fraudulent claims are asking for small repaircost amount and also after the age of 55 there is no fraudulent cases found.



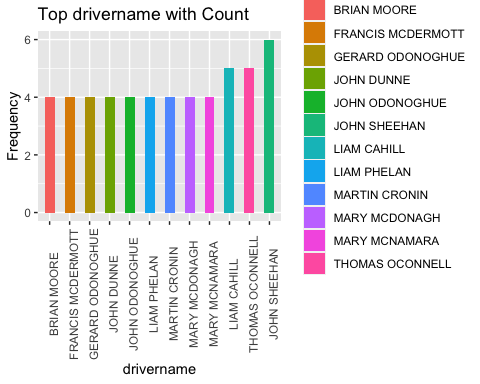
The Above graphs shows us the top 12 address column in our dataset which are flagged as true with how many times they were flagged as true(fraudulent) in our dataset.

So we can see that “**2 SEVERN TERRACE”** has the highest frequency indicating that this address was repeatedly flagged as fraud for 3 times. and the rest addresses such as”**1 Shannon Terrace” , “11 Severen Lawns”** were flagged as true two times as indicated by our dataset.

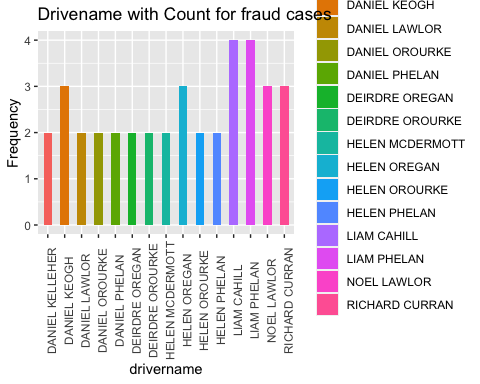
This Suggest there is a Potential repeat Frauds at the same address but no extraordinary concentration at any one location



We can say that there are 10 address in our dataset which are repeated 4 times and compairing the above two bar charts 1 of which has been flagged as True for fraud “**2 SEVERN TERRACE”** 3 times. Similary with the once repeated 3 times in our dataset 1 such has been flagged as true 2 times”**12 Thames Terrace”.** *This was one such pattern that could be seen. there can be many such hiddden patterns in address column.*

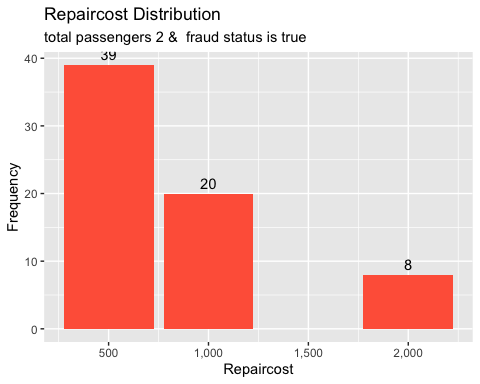


From the above bar graph we can observe that there are repetition in driver names as well with the **highest frequency of 6 by the name** **JOHN SHEEHAN,** followed by **Thomas Oconnnell & Liam Cahhill** with a **frequency of 5.** and many more indicating the repetitiveness of names in our dataset.



This chart identifies drivers associated with fraudulent claims. Each bar represents a driver name, and the bar height indicates the frequency of fraud cases associated with that driver name. Indicating **Liam Cahill** and **Liam Phelena** are the names of driver that have been repeated and been **flagged as fraud the highest times that is 4 times**. Similarly

## Going Deep in the Analysis

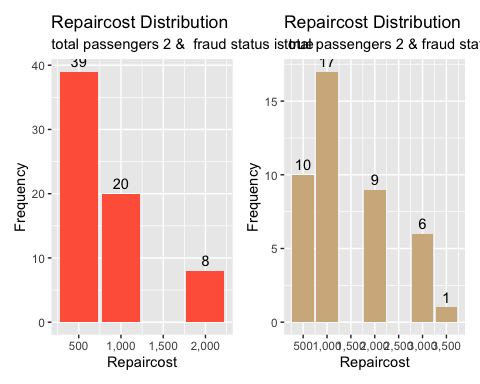


After filtering the data for true as fraud flag and total passengers == 2 we get the graph that shows the distribution of repair costs for claims involving two passengers where the fraud status is true.

The majority of fraudulent claims with two passengers have a repair cost of 500, with a frequency of 39. The second most common repair cost is 1000, with a frequency of 20.The least frequent repair cost is 2000, with a frequency of 8.

Claims with a repair cost of 500 dominate the distribution. This suggests that **fraudulent claims are more common at lower repair costs indicating a potential common threshold for fraudulent claims**.

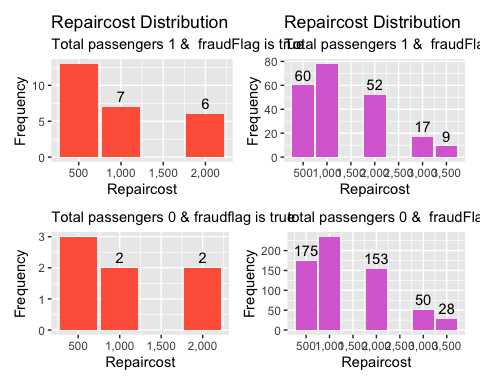
There is a significant drop in the frequency of fraudulent claims as the repair cost increases from 500 to 1000 and further to 2000.



As we compare for non-fraudulent claims with the same amount of passenger as 2 but fraud flag = false we can observe that the highest frequency of claims with a fraud status of false is for a repair cost of 1000, with 17 claims.The next most frequent repair costs are 500 (10 claims), 2000 (9 claims), and 3000 (6 claims).

Non-fraudulent claims show a more varied distribution, with notable frequencies at 1000, 500, 2000, and 3000.

The presence of only three distinct repair costs for fraudulent claims further supports the idea of a pattern or strategy being employed.



Doing the comparison for total\_passengers as 1 and 0 across fraudflag true and false we find that

For cases with total passengers 1 and fraud status true, the repair costs are generally lower. The highest frequency of repair costs is observed around 500. In contrast, for non-fraudulent cases with one passenger, the repair costs are higher, with the most frequent repair cost around 1000.

For both fraud and non-fraud cases with zero passengers, the repair cost distribution is similar, with the highest frequencies for lower repair costs (500 and 1000). The number of passengers seems to have a notable impact on repair cost patterns, especially in fraudulent cases.

|  |
| --- |
| Figure 6: list of addresses with 2 passengers and frequency as 2 |

This is the list of addresses which had total passengers as 2 and the frequency of how many times did they appear in the dataset when filtered for true as fraudflag.

## Engineered Features to improve the model :

**Feature engineering** is a crucial step in the data pre-processing stage because it enhances the predictive power of machine learning models by creating new features or modifying existing ones to better capture the underlying patterns in the data. In this process, I am first removing zero variance predictors that do not contribute any meaningful information to the model. Further, discretizing this continuous variable, ‘age’ into categorical bins. Furthermore, the interaction between ‘age’ and ‘repaircost’ was done to know the combined influence of those two features concerning the target feature, fraudFlag as well as the interaction between total\_passengers:repaircost. These steps finalize the preparation and application of the recipe to the data to ensure that the dataset is enriched with relevant, well-structured features that will result in better model performance.

## Conclusions from Exploratory Analysis

This analysis of the car insurance claims dataset uncovers several critical insights. After thorough cleaning of the data, performing feature engineering and doing Exploratory Data Analysis the dataset has been transformed into a reliable resource of identifying fraud claim trends and patterns. Some key findings included are as follows.

**Fraudulent Claims** are concentrated over smaller repaircost price especially around 500 and as repaircost increases this trend diminishes. This indicates that claims for higher repaircost may go through with a strict validation process in the insurance company.

On the other Hand non-Fraudulent claims tend to have higher repair cost in general.

**Fraud rate decreases with increasing drivers age** and becomes negligible after the age of 55.Which can be understood as Younger drivers might be less experienced and more prone to risky behaviour increasing both legitimate and Fraudulent claims.

On the other Hand older driver may have many things attached with their behavior like their reputation and many more hence less fraudulent activities are observed.

We can also say that Fraudsters often target less amount to grab less attention on themselves hence are clustered in low repaircost prices.

Fraudulent Claims with two passengers are concentrated at lower price indicating it could be a staged crime and both parties of collusion sticking on small amount for better compensation.(common threshold we can say)

Certain driver-names and Addresses are associated with multiple fraudulent claims and are repeating many times from this the address indicates maybe organized crime hot-spots

Hence the insurance company should

* Monitor the younger claims application report keenly (below 30) as these group is more likely to be attracted towards Fraudulent tendencies.
* Flag repeat offenders by tracking addresses and driver names.
* Introduce some threshold for smaller claims repaircost & claims falling under suspision should be given more attention
* can also introduce some enhanced varification process for low claim amounts.
* Should keep an eye on number of passengers as staged accidents could play a vital role in fraudulent claims.

By adopting all these Strategies and keeping all the insights in mind the Insurance company can build a strong and robust fraud detection framework which would minimize the losses and ensure fair treatment to genuine Fraudulent Claims.