Fraud trends in vehicle insurance data

Helmi Kaittikkattil Abraham (K00302088)

# Fraud trends in vehicle insurance data”

## Problem Statement

The organization operates in the vehicle insurance domain and aims to identify patterns and factors that influence the fraudulent claims in vehicle insurance data. The data set includes information such as driver demographics, repair costs, and passenger details.By analyzing these,the ultimate objective is to provide insights that can help decision-making for fraud prevention and claim-handling processes.

### Key questions include:

* Do specific age group categories have higher fraud rates?
* How does the presence of passengers influence the likelihood of fraud?
* Identifying trends in fraudulent claims by analyzing repair costs.

## Solution Summary

To explore the dataset and derive insights, the following steps will be undertaken:

* Cleaning and organizing the data to handle missing values and inconsistencies.
* Engineering new features to enhance the analysis.
* Conducting exploratory data analysis (EDA) to visualize trends and correlations.

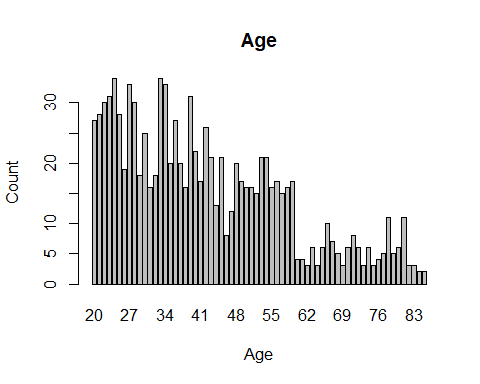
## Exploratory Data Analysis

After loading the data set and examining the summary, decided to explore the attributes in detail, in order to find which pre processing steps are required.

### Exploring each attribute

* driver : Indicates the name of driver.
* Age : Age of driver

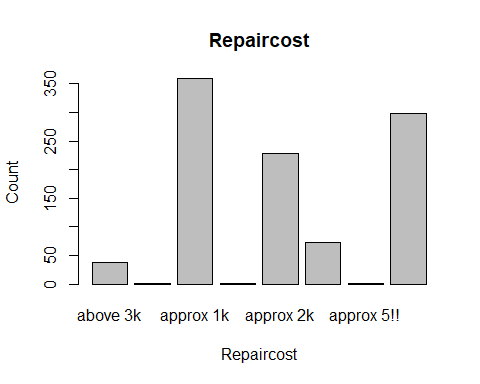
[1] 24 53 48 40 27 38 34 33 21 23 55 74 22 32 44 37 82 56 42 52 70 72 80 71 30  
[26] 31 41 20 58 25 54 39 36 35 46 68 50 61 28 78 43 51 65 66 63 45 59 67 26 49  
[51] 62 64 57 77 75 29 84 81 79 69 73 60 83 47 76 85



It will be better to find fraud flag in an age group rather than each given age. So a new feature with different age group will perform better.

* address : It is the address of driver.
* passenger1 & passenger2 : This indicates whether a passenger was present in the car or not during accident. Sometimes more than one passenger may be present and sometimes there is no one in the car. To verify whether this presence of passenger have any relation to fraudness, we need one column which describes this passenger presence. So feature engineering will help here.
* Repair cost : Indicates the claimed cost of repair

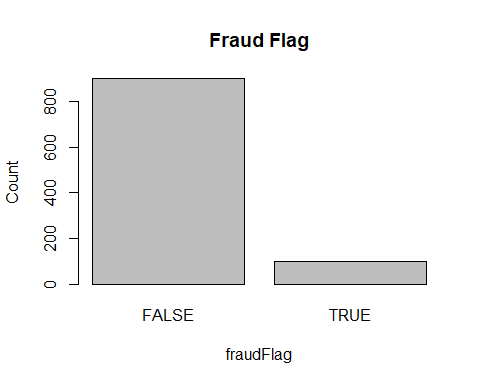
[1] "approx 3k" "approx 2k" "approx 500" "approx 1k" "approx 5!!"  
[6] "above 3k" "approx $\*0" "approx 2~"



From the bar graph, its clear that some unique values in this column contains special characters which we need to clear.

* Fraud flag : This is the output column which contain Boolean values *TRUE* and *FALSE*, where *TRUE* indicates its a fraud claim and *FALSE* indicate it is not

[1] FALSE TRUE



Here is the detailed version of processes undertaken in data cleaning, feature engineering, and the visualizations generated to uncover patterns and insights.

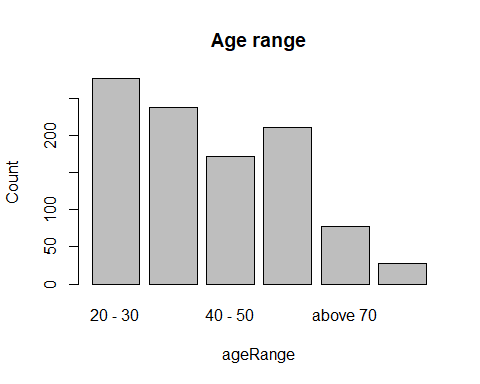
### 1. Data Cleaning

* special characters in some unique values of column *repaircost* were removed
* All these assumptions were based on the existing values in the same column.
* These are the assumptions
  + approx 2~ is taken as approx 2k
  + approx 5!! is taken as approx 500
  + approx $\*0 is taken as approx 0
* All the preprocessing were done with the existing column itself.No need of extra variable

### 2. Features engineered to improve EDA visualisations.

* Age
  + Even though each individual value in the age provide information about fraudflag, its more convenient to group them to do more analysis.
  + So age values were converted into different age groups and made a new column called **AgeRange**

[1] "20 - 30" "50 - 70" "40 - 50" "30 - 40" "above 70" "below 20"



* Passenger1 & passenger2
  + Mainly 3 cases are here
    - Only one passenger is present
    - Two passengers are there
    - No passengers
  + So we can consider mainly two cases, either passenger is there or not
  + A new column is created called **HasPassenger** which contains *Yes* if passenger is present and *No* if passenger is not there.

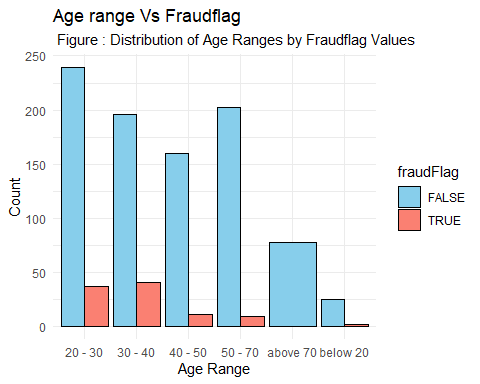
[1] TRUE NA

### 3. Removing unwanted columns

After data preprocessing, the unwanted columns age,passenger1,passenger2 are removed.

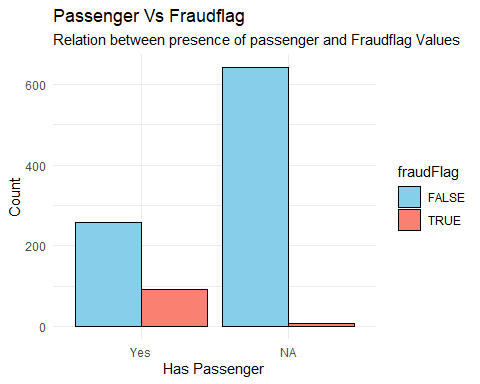
### 3. Exploratory analysis through visualisations.

**Distribution of fraud cases across different age groups**



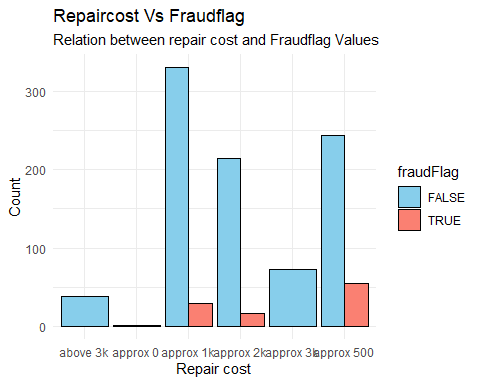
* Figure shows, most of the fraud cases are reported in between age 20 and 40, among which 30-40 reported more fraud claim cases.
* Above 70 age group doesn’t claimed any fraud claims.
* comparatively below 20 age group shows minimum number of fraud cases.
* 40-50 and 50-70 group shows similar trend.

**Relationship between presence of passenger and fraud**



* Fraud claims are more when passenger is present in the car
* If there is no passenger, fraud cases are rare.

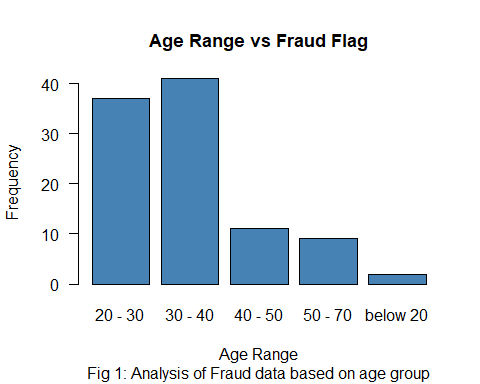
**Impact of repair cost on fraud likelihood**



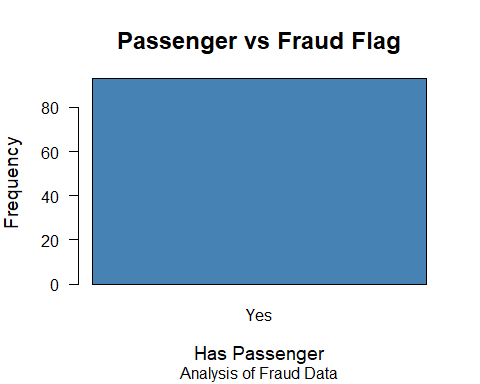
* Most of the drivers claimed for approximately 500 euros showed true flag in fraud
* No one tried to make fraud cases for repair cost greater than or equal to 3000 euros.
* Most of the fraud cases are in between 500 and 2000 euros.

For better analysis of true fraud flags, only **true cases** are filtered and analysed.

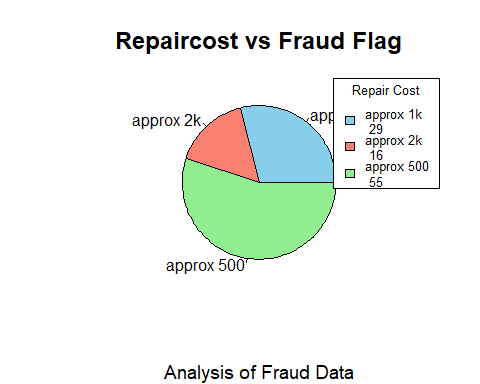
**Distribution of fraud cases across different age groups**



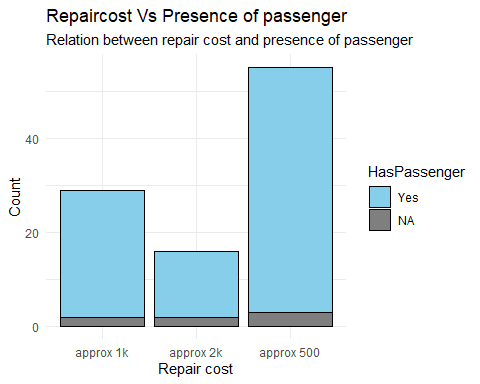
**Relationship between presence of passenger and fraud**



**Impact of repair cost on fraud likelihood**

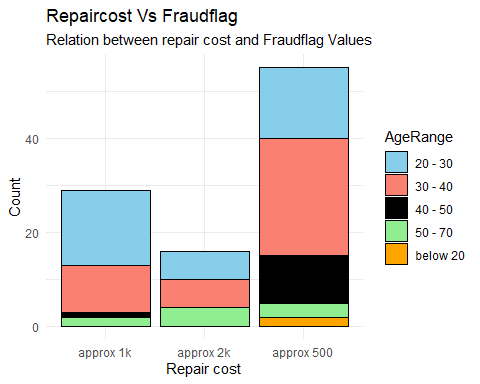


**Relationship between presence of passenger and repair cost**



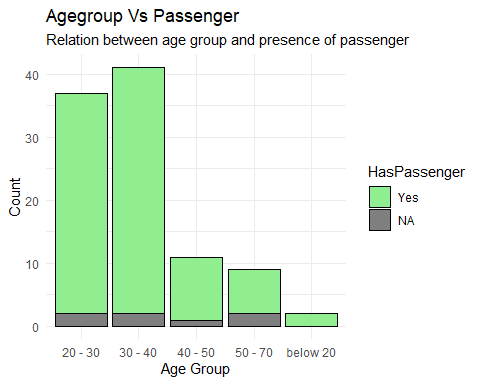
* Most of the cases claimed for 500 euros in which passengers were present.
* least number of fraud cases were associated with repair cost 2000.

**Relationship between age range and repair cost**



* 30-40 group claimed for 500 euros more, compared to others
* Fraud claim cases related to below 20 age group can only be seen in 500 euro repair cost.
* 40-50 group claimed for below 1000 euros in fraud cases.

**Relationship between age range and presence of passenger**



* If passenger is present, all age group shows more fraud flag
* Younger drivers have the tendency to show fraud flag if there is passenger, otherwise they aren’t.

Younger drivers with higher repair costs are more likely to have flagged fraud cases. Claims with no passengers tend to exhibit lower fraud rates compared to those with passengers. High repair costs often correspond to a small percentage of cases but have a notable fraud impact.

### Conclusions from Exploratory Analysis

The conclusions from the analysis are as follows:

* Old people above 70 years doesn’t try to do fraud cases with claim
* younger drivers below 20 years also shows minimum cases of fraud claim
* The one who showed fraud cases more is in 30-40 age group
* Fraud claims are more when passenger is present in the car
* Most of the drivers claims for a lesser amount when they are in fraudflag