Interpretation of Data - EDA for fraud claim dataset

Ganesh Nimkar (K00277641)

4/15/2022

## Problem Statement

In this assignment we need to provide insight and correlation between other attributes with a fraud flag attribute.This insight helpful for developing a predictive model to predict the claim is a fraudulent or real by exploring motor fraud cases for an insurance company based on a given data set.

## Solution Summary

To accomplish this task we are using Exploratory Data Analysis (EDA). Exploratory Data Analysis (EDA) is a key process in the development of the model to detect the fraud based on features like age,address and repair cost,etc.

##1. Loading Libraries

We are loading R libraries such as, tidyverse, forcats,dplyr,skimr,stringr. which we are using in our EDA process,

## 2. Loading the Data set

The provided data is in CSV format.

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

## 3.Exploring the Dataset

## 'data.frame': 1000 obs. of 7 variables:  
## $ driver : chr "JOSEPH MCGRATH" "MARY BRENNAN" "JOSEPH COLLINS" "ROBERT WALSH" ...  
## $ age : int 24 53 48 40 27 38 34 34 33 21 ...  
## $ address : chr "3 CO$RIB VIEW" "21 BLACKWATER VIEW" "7 SLANEY LODGE" "12 LIFFEY GROVE" ...  
## $ passenger1: chr "JOSEPH GRIFFIN" "" "" "" ...  
## $ passenger2: chr "" "" "" "" ...  
## $ repaircost: chr "approx 3k" "approx 2k" "approx 2k" "approx 500" ...  
## $ fraudFlag : chr "FALSE" "FALSE" "FALSE" "FALSE" ...

Data summary

|  |  |
| --- | --- |
| Name | data\_set |
| Number of rows | 1000 |
| Number of columns | 7 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 6 |
| numeric | 1 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| driver | 0 | 1 | 10 | 17 | 0 | 641 | 0 |
| address | 0 | 1 | 10 | 19 | 0 | 760 | 0 |
| passenger1 | 0 | 1 | 0 | 17 | 648 | 296 | 0 |
| passenger2 | 0 | 1 | 0 | 17 | 890 | 93 | 0 |
| repaircost | 0 | 1 | 8 | 10 | 0 | 10 | 0 |
| fraudFlag | 0 | 1 | 4 | 5 | 0 | 4 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| age | 0 | 1 | 41.89 | 16.28 | 20 | 28 | 39 | 53 | 85 | ▇▆▅▂▂ |

## 3.1 Overview of Dataset

We have our dataset of 1000 rows and 7 columns such as,

1. Driver: Represents the First and Last name of the driver.
2. Age: Age of driver.
3. Address: Contains House number,Address Line1 , Address Line2 of the driver.
4. Passenger 1: Represents to the passenger who was with the driver when the accident had happened or the incident for was submitted for claim.
5. Passenger 2 : Represent the second passenger who was with the driver.
6. Repaircost : Represent the repair cost of the claim.
7. FraudFlag : Indicates the claim is a fraud or genuine.

## 4. Data Wrangling

Most of the values in the “passenger1” and “passenger2” attributes are missing. these attributes are in the character form. This attributes store the passenger names, and missing values mean that no passengers were present. As a result, other columns are examined and no missing values are discovered.

As a part of cleaning the address column , we have removed the house number and we fix the punctuation errors in the address by locating the index and typing in the correct address

## new\_address  
## 1 CORRIB VIEW  
## 2 BLACKWATER VIEW  
## 3 SLANEY LODGE  
## 4 LIFFEY GROVE  
## 5 BLACKWATER GLADE  
## 6 BARROW GLADE

There are some spelling mistakes in fraudflag column. we are correcting them

## fraudFlag  
## 1 FALSE  
## 2 FALSE  
## 3 FALSE  
## 4 FALSE  
## 5 FALSE  
## 6 FALSE

We are removing special characters from diver name and repair cost column.

## driver  
## 1 JOSEPH MCGRATH  
## 2 MARY BRENNAN  
## 3 JOSEPH COLLINS  
## 4 ROBERT WALSH  
## 5 KEVIN OCONNELL  
## 6 BRIAN CULLEN

## repaircost  
## 1 approx 3k  
## 2 approx 2k  
## 3 approx 2k  
## 4 approx 500  
## 5 approx 500  
## 6 approx 2k

# 5. Feature Engineering

Feature engineering is an important aspect of model development. Here, we are developing new features from the data set that will help us in the aid of data exploration.

Dividig the whole ages into specific groups such as Seniors,Adults,Youth and Childeren

## agegroup  
## 1 Youth  
## 2 Adults  
## 3 Adults  
## 4 Adults  
## 5 Youth  
## 6 Adults

splitting new\_address column into two part such as Address Line 1 and Address Line 2

## AddLine\_1 AddLine\_2  
## 1 CORRIB VIEW  
## 2 BLACKWATER VIEW  
## 3 SLANEY LODGE  
## 4 LIFFEY GROVE  
## 5 BLACKWATER GLADE  
## 6 BARROW GLADE

A new column was introduced to know the how many passengers are in car at the time of accident.

## passengernumber passenger1 passenger2  
## 1 1 JOSEPH GRIFFIN   
## 2 0   
## 3 0   
## 4 0   
## 5 0   
## 6 1 MICHAEL BROWNE

We then create a new column that contains both Passenger Names

## passengers  
## 1 JOSEPH GRIFFIN  
## 2   
## 3   
## 4   
## 5   
## 6 MICHAEL BROWNE

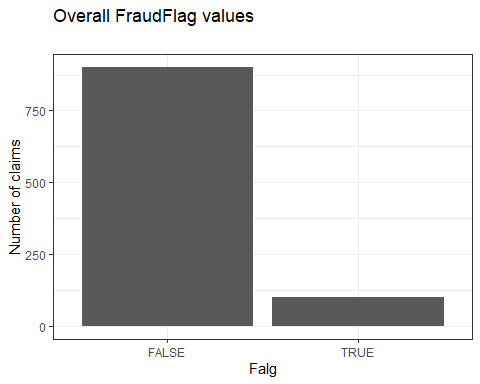
A new binary feature was feature was introduced as to check if the driver name has been appeared in passenger names.

1- if the name was appeared, 0- if the name did not appear

## driveraspassenger  
## 1 0  
## 2 0  
## 3 0  
## 4 0  
## 5 0  
## 6 0

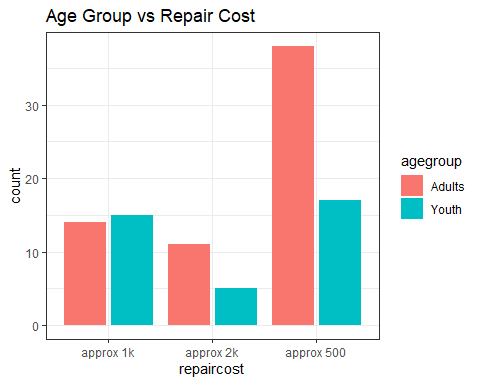
## 6 Data Exploration using visualisation

### 6.1 Dataset Skewness



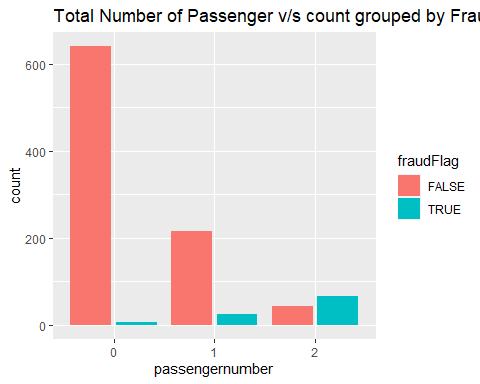
The data that has been given to us is highly skewed as out of 1000 cases of claim,only nearby 100 cases are Fraud cases.This gives us difficulty in predicting the accuracy of results.

### 6.1.1 Age Vs Repair cost



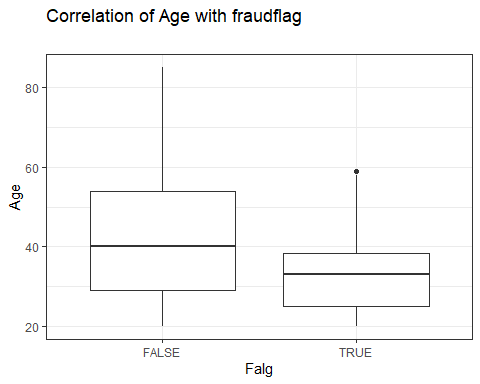
graph shows thatthe age group of Adults are dominant in all price ranges expect approx 1k,Where the Youth are more.

### 6.1.2 Totalnum of passengers vs frraudflag



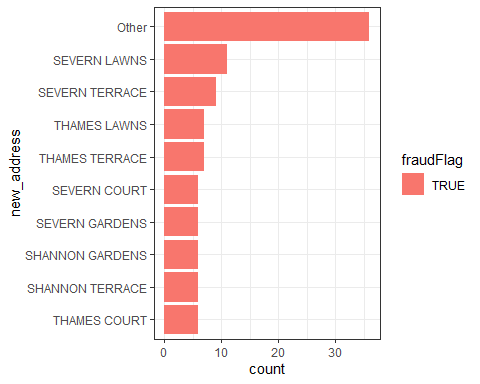
This graph describes,The passenger number increases the fraud cases also increases. We can see the highest number of fraud cases are from the claim where two passengers traveling.

### 6.1.3 Age count grouped by fraud flag



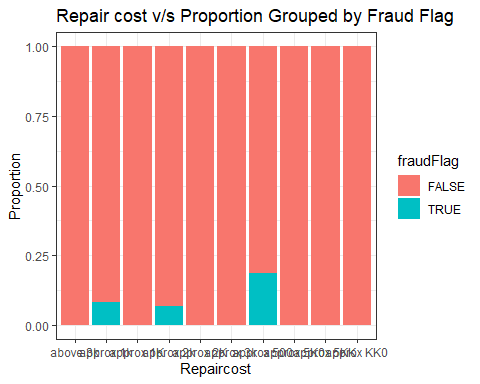
This box plot shows, In the genuine claim cases age group mainly from 30 to 55 ,where as for the Fraud cases it varies from 25 to 40.This shows that fraud cases are mainly attempted by Youth and Adults.

### 6.1.4 New\_address v/s count grouped by fraud flag = true



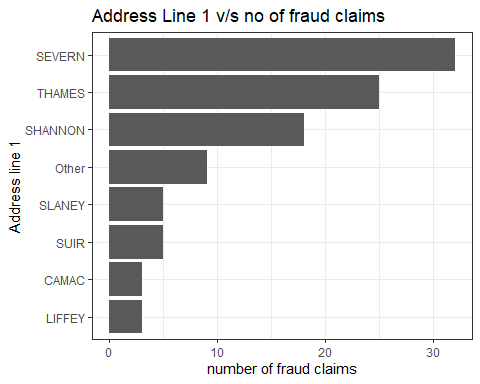
The graph tells us that most of the fruad cases are deteced from Seven Lawns, followed by Severn Terrace .

### 6.1.5 Repair cost vs Fraud Flag



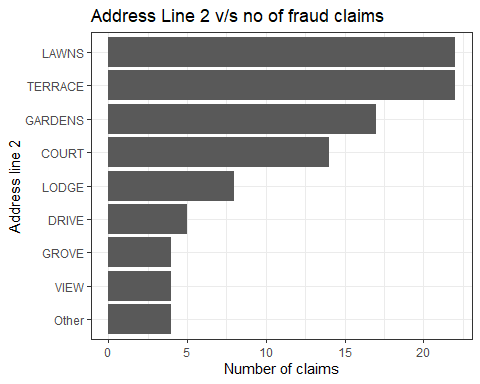
This bar graph shows that, the most of the fraud cases has been attempted for repair cost that approximately 500 and the followed by 1000

### 6.1.6 Address Line 1 v/s no of claims grouped by fraud flag = true



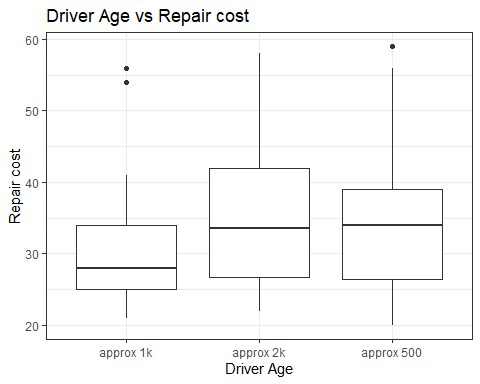
This graph shows that, majority of the fraud cases are from Severen followed by Thames and Shannon

### 6.1.7 Address Line 2 v/s no of claims grouped by fraud flag = true



We can see that Highest number i.e over 20+ of Fraud cases has been deteced from Terrace followed by Terrace and Gardens

### 6.1.8 Driver Age vs repair cost



This boxplot shows some interesting facts such as,

Repair costs Driver age 1k 25-35 2k,500 33-34

This data confirms our findings with that of previous plots.

### 6.1.9 Drivers as passengers

From the feature engineered variable , we can see that

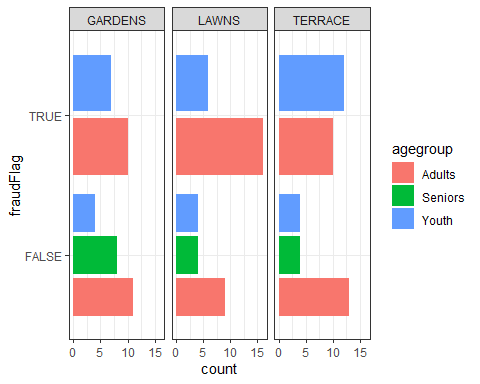
## [1] 0

There are 210 drivers name that comes in the passenger list.

Even with the high skewness of data, it’s shown people who appears as passengers in different cases has high chance of being Frauds.

### 6.1.10 Address vs Agegroup ,Group ed by Fraud Flag

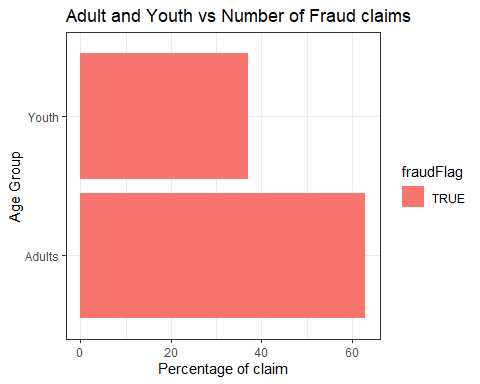
we wan to check which age group people are attempting fraud cases



This graph clearly shown, Citizens belonging to age group of Adults and Youth mainly commit fraud cases as well as mainly adults commits fraud cases in Gardens and Lawns, where as it is youth in Terrace.

### 6.1.11 Adult and youth age group grouped by fraud flag

Previous data tells us that Youth and Adults has major contribution in fraud flags ,We can cross verify whole scenario.



In this graph, we can see that over 60% of the fraud cases are commited by the Adults and nearly 40% are by Youth.

## 7. Findings

In this EDA process we found below findings,

1. Drivers which are under Youth and Adults age group are commiting more Fraud cases,So we can low prioritize Senior citizens from suspicius cases.
2. Most of the Fraud cases fall under for the price range of 500 and then followed by 1k.
3. In Address Line 1 - SEVERN ,THAMES or SHANNON and In Address Line 2 - LAWNS,TERRACE,GARDENS are more prone to fraud cases. So, We can say claims from these addresses to a suspicious list.
4. The fraud’s are mainly detected by the number of passengers, as the number of passenger increases is more chances for fraud claim.
5. After doing feature engineering graphs plotted, it is clear that the customers who are repeatedly appears in passenger list in for other cases are susceptible for frauds.

We can conclude that,By keeping the above things in consideration we can categories upcoming cases for detect if its a fraud case or not as well as what are the potential factors for the fraud cases.