Data Exploration

JENNIFER AIGBUGWILI

18/12/2021

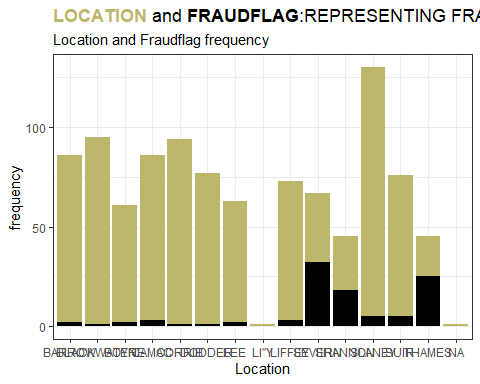
## 0.1 Problem Statement

The project task is to explore and make insight for a motor fraud case which contains different rows and columns, this dataser requires some cleaning, we also carried out some Exploratory Data Analysis (EDA) which is a very important processing the analysis phase. here we replaced missing variable, converting from charachter to numerical value when needed, replacing specail characher, and sepaating and spliting of columns.

## # A tibble: 1,000 x 3  
## # Groups: Location [15]  
## driver Location fraudFlag  
## <chr> <chr> <lgl>   
## 1 JOSEPH MCGRATH CORRIB FALSE   
## 2 MARY BRENNAN BLACKWATER FALSE   
## 3 JOSEPH COLLINS SLANEY FALSE   
## 4 ROBERT WALSH LIFFEY FALSE   
## 5 KEVIN OCONNELL BLACKWATER FALSE   
## 6 BRIAN CULLEN BARROW FALSE   
## 7 HELEN PHELAN BARROW FALSE   
## 8 SEAN KAVANAGH LEE FALSE   
## 9 PATRICK GRIFFIN SHANNON FALSE   
## 10 MARK MCNAMARA SLANEY FALSE   
## # ... with 990 more rows

## 0.2 Solution Summary

### 0.2.1 Exploratory Visualisations

A range of visualisations is developed to analyse and gain insight in respect to different location and different number of passenger and insurence cost for the motor fraud case. 

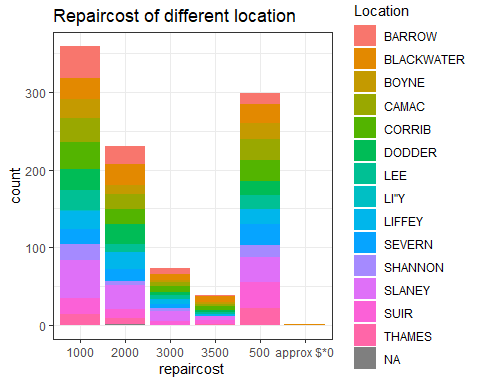


Figure 2: Repaircost allocated to different location

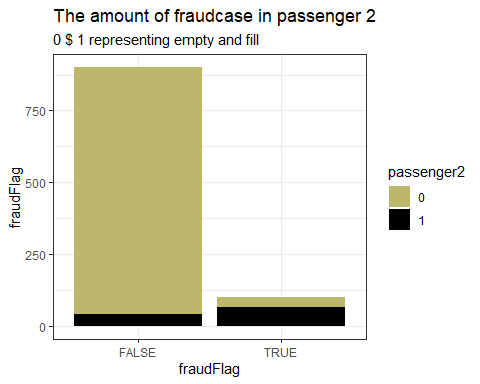


Figure 3: The amount of fraudcase in passenger 2

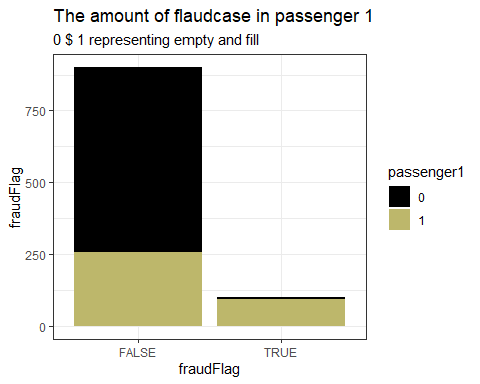


Figure 4: The amount of fraudcase in passenger 1

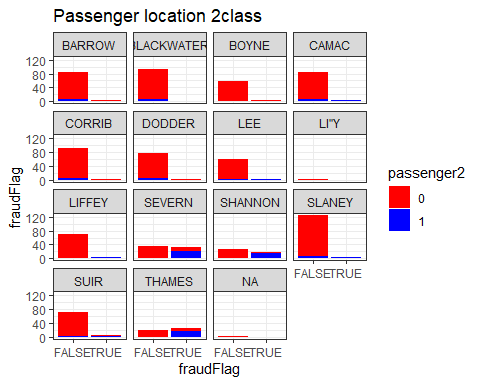


Figure 5: fraudcase from passenger2 in different location

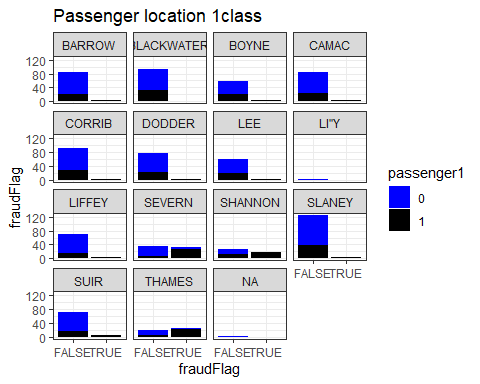
 The dataset shows the level of fraudcase in different location as well as showing its repaircost for each individual location, it also domonstrate the amount of passenger each driver had either passenger 1 or passenger 2 or both, according to locations and age range. it also identify the age range with the highest fraudflag and case. the extent to which eash age and repaircost relate which will help give more insight on the amount of fraudflags located in the dataset.

Figure @ref(fig:repaircost\_location) highlights the repaircost for different location. Figure ?? provides information regarding the amount of fraudcases in different location. variable across the weekend and helps understand the characteristics of the repaircost in each location and how it affect the fraudflag.

### 0.2.2 Engineered Features

Features can be engineered based on the insights to improve predictive model performance. Feature engineering can be viewed as the process of transforming raw data into features that better represent the underlying problem to predictive models. Important features or variables are engineered from the data to better represent the underlying analysis context to the predictive models.

You can represent the insights from the EDA process to the respective model through a binary encoding process.

### 0.2.3 Conclusions from Exploratory Analysis

The conclusions from the analysis are as follows:

* The fraudflag of each location stating the highest and lowest fraudcase, severn hiving the highest and blackwater has the least. \*The amount of fraudflags identified in passenger1 column is higher than the passenger2.
* The location BOYNE has no passenger identified in the passenger 2 location
* Severn, Shannon and Thames had the highest number of fraudflags identified when using the passenger1 and passenger2 column.
* Feature engineering can be used to represent the data to the model. \*The visualization alos demonstrate the number of passenger present in each location, where 0 represent an empty cell and 1 represent a cell with passenger availability.