**Overview**

In this study, we will we looking at different phenomenon and dimensions of car accidents that occurred in UK between 2005 to 2015 with respect to different variables. We will observe if there is association between certain variables as well as checking which factors can influence certain aspects of car accidents. We will also graphically illustrate the behaviors of the variables. We will try to see, if age of casualty and sex of casualty have an influence on how sever the casualty case from accident case is.

**Dataset**

We have a dataset of the UK car accidents between 2005 to 2015 from Kaggle. It is an open dataset. The main reason this dataset was collected because it gives a good idea about the accident phenomenon of UK from a relatively modern time.

**Variables**

The variables are, Sex of Casualty, Age of Casualty, Casualty Severity, Pedestrian Location, Casualty Type, Pedestrian Movement, Bus or Coach Passenger, Pedestrian Road Maintenance Worker.

**Data Cleaning**

After importing the csv file into R, we will summarize it to see the variables and if there is anything needed to be cleaned of edited. Usually, missing values are removed and irrelevant values are also removed as a part of the cleaning. Example of irrelevant values can be, negative values for age.

R-codes for initial observing are given below,

**x = read.csv( file.choose ( ) )**

**summary ( x )**

The output will be,

A picture containing table

Description automatically generated

We see from the output that, there are a lot of missing values and also variables like Age of causality and sex of causality has negative values. These type of values and missing values needs to be cleaned.

So, R codes for cleaning them are,

**x=replace ( x, x < 0, NA )**

**x = x[ complete.cases ( x ) , ]**

**summary( x )**

The output can be observed as,

Text

Description automatically generated

From the output, we can say that, there are no missing values and also no negative values. So, we may work with the data. The data is cleaned.

**Variables relevant for analysis**

The variables that are relevant for analysis are, Sex of Casualty, Age of Casualty, Casualty Severity, Pedestrian Location, Casualty Type, Car passenger, Bus or Coach Passenger

**Continuous variable:** Age of Casualty

**Categorical variable:** Sex of Casualty, Pedestrian Location, Casualty Type, Car passenger, Bus or Coach Passenger, Casualty Severity

**Descriptive Analysis**

Here, we will observe summary statistics for the continuous variables (Age of Casualty and Casualty Severity) and frequency of the categorical variables

To observe summaries the R codes are,

**summary ( x $ Age\_of\_Casualty )**

The Outputs are,



We see that, The mean age of the casualty is 36.42 years. The first quartile is 22 it means, 25% of the casualty has age less than or equal to 22. The second quartile or median is 33 it means, 50% of the casualty has age less than or equal to 33. The third quartile or median is 49 it means, 75% of the casualty has age less than or equal to 49. The minimum value of age is 0 which indicates less than 1 year and maximum indicates 104 years.

To observe frequencies for the categorical variables,

The R codes are,

**table( x $ Sex\_of\_Casualty )**

**table( x $ Pedestrian\_Location )**

**table( x $ Casualty\_Type )**

**table( x $ Car\_Passenger )**

**table( x $ Bus\_or\_Coach\_Passenger )**

**table( x $ Casualty\_Severity )**

The Outputs are,

Table

Description automatically generated

A picture containing text

Description automatically generated

We see that,

Most of the casualty are male (category 1) compared to female (category 2)

The pedestrian location has 10 category types. Most of them are not a pedestrian (category 0). The least are crossing in zig-zag exit lines (category 3).

There are 21 types of casualty types, the highest are within car occupants (category 9) and least are mobility scooter rider (category 9).

Car passenger has 3 category types and most are not a passenger (category 0).

The bus or coach passenger has 5 categories, most are not a bus or coach passenger (category 0).

Casualty severity has 3 categories. Most of them are slight casualty (category 3) and the minimum of them are fatal (category 1).

**Data Visualization**

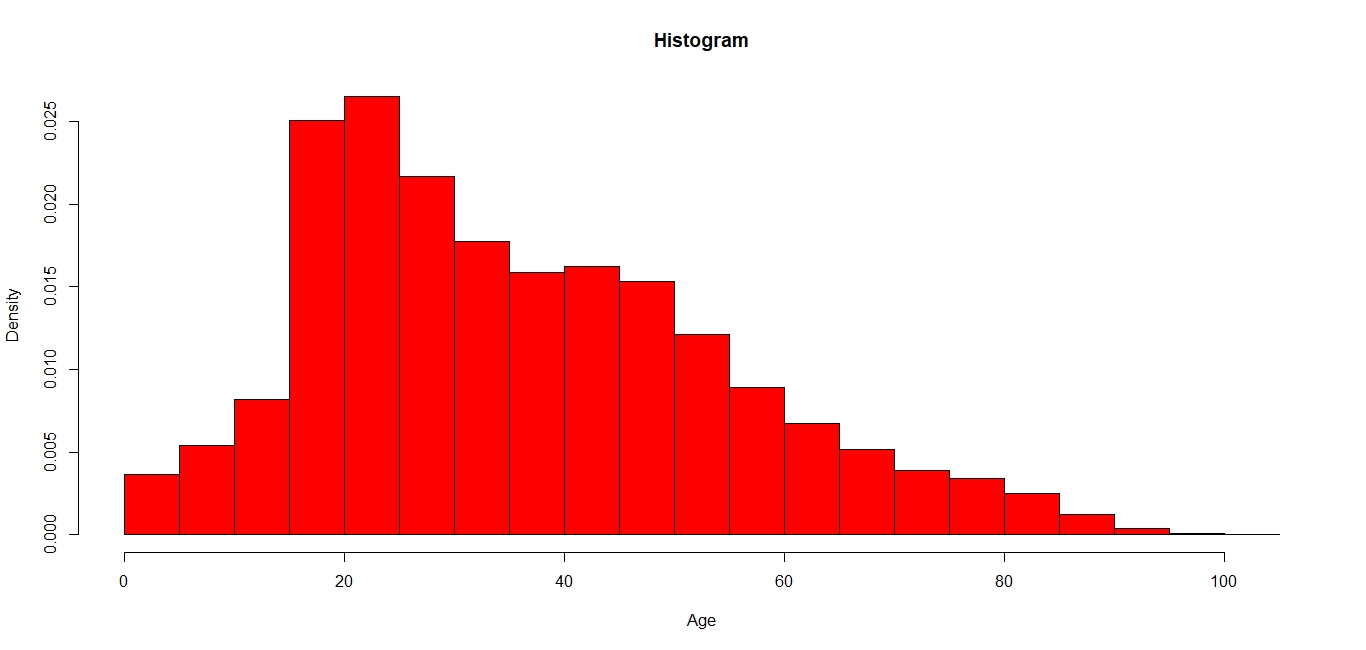
We will observe histogram and boxplot for the continuous variable age of casualty, to have an idea about its distribution and characteristic.

The R code will be,

**hist( x $ Age\_of\_Casualty , prob = T , col = ' red ' , xlab = ' Age ' , main = ' Histogram ' )**

**boxplot( x $ Age\_of\_Casualty, ylab = ' Age ', main = ' Boxplot' )**

The output will look like,



From the histogram, we see that, the distribution of age of casualty is approximately bell shaped. It’s not symmetric in fact it is positively skewed.

Chart, box and whisker chart

Description automatically generated

The boxplot also gives a similar output as the histogram, meaning that the distribution of the age of casualty is positively skewed.

We will create pie charts and bar plots for some of the categorical variables,

For Sex of casualty we will draw pie chart, the R code will be,

**pie( table ( x $ Sex\_of\_Casualty ), main = ' Pie chart of Car Passenger ' )**

The output will be,

Chart, pie chart

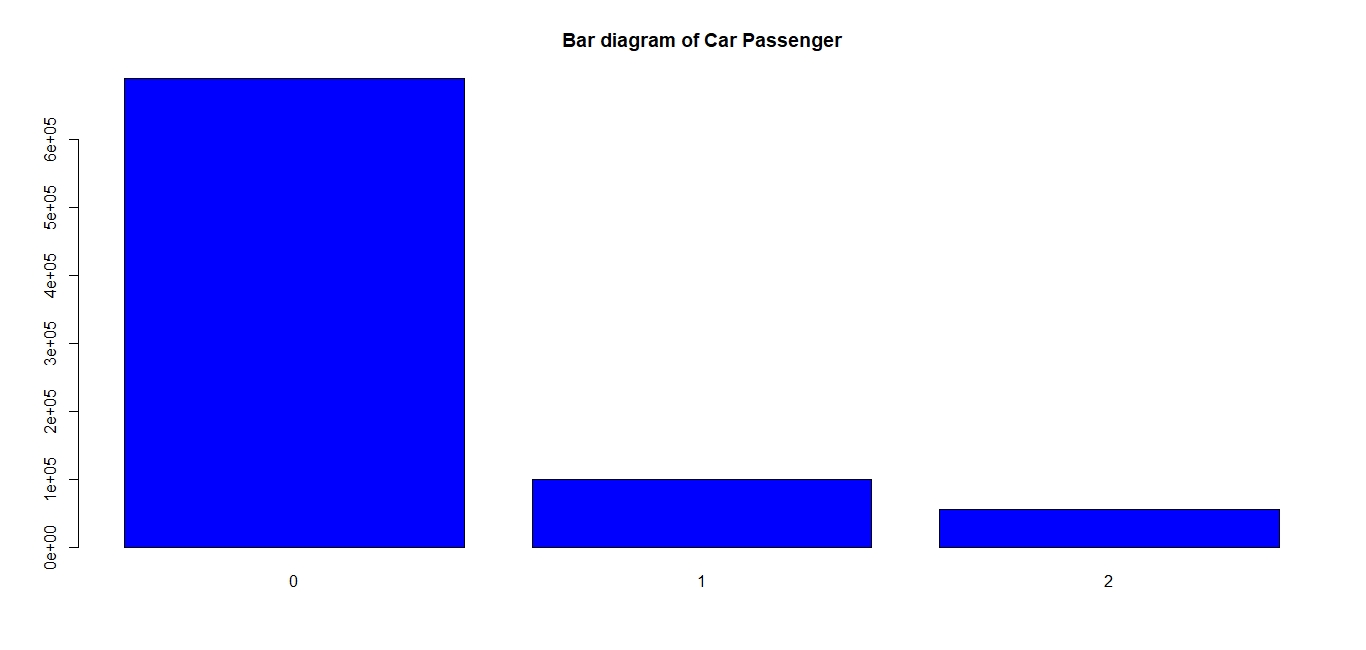
Description automatically generated

Here, we observe that there are more males (category 1) than female (category 2).

For car passenger we will do a bar diagram,

The R code will be,

**barplot( table ( x $ Car\_Passenger ) , main = ' Bar diagram of Car Passenger ' , col = ’ blue ’ )**

****

Here we see that, not a car passenger (category) has the most values compared to other categories.

We will observe another pie chart for casualty severity,

The R codes will be,

**pie( table ( x $ Casualty\_Severity ) , main = ' Pie Chart of Casualty Severity ' )**

The output will be,

Chart, pie chart

Description automatically generated

We see that, most of the observations of casualty belong to slight casualty (category 3). Second most belong to serious casualty (category 2) and least belong to fatal casualty (category 1).

We will also draw a clustered bar diagram of sex of casualty on basis of different levels of casualty severity.

The R codes will be,

**t = table ( x $ Sex\_of\_Casualty , x $ Casualty\_Severity )**

**barplot ( t , beside = T , legend = rownames(x$Sex\_of\_Casualty) , main = ' Clustered barplot of sex on different levels of Casualty severity ' )**

The output will be,

Chart, waterfall chart

Description automatically generated

We can say that, both sex of casualty has the highest frequency in slight casualty (category type 3) of case severity. In all the categories male (category 1) has more values compared to female (category 2) which is also common with the findings we got from the pie chart we drew earlier.

**Hypothesis testing on measure of association between categorical variables**

In this section we would like to analyze if there is significant association between some of the categorical variables.

Our first hypothesis will be,

Ho: Sex of casualty and casualty severity do not have a statistically significant association

HA: Sex of casualty and casualty severity have a statistically significant association

To test this, we will use chi-square test,

We will consider 5% level of significance. So, value of α will be 0.05.

The R code will be,

**Casualty\_Severity = as.factor( x $ Casualty\_Severity ) # Transforming the numeric variable into factor variable**

**Sex\_of\_Casualty = as.factor(x $ Sex\_of\_Casualty) #Transforming the numeric variable into factor variable**

**chisq.test ( Casualty\_Severity , Sex\_of\_Casualty )**

The R output will be,

Text

Description automatically generated

Here we see, the p-value is less than 0.05.

Comment: We may reject out null hypothesis. So, we can say that, Sex of casualty and casualty severity have a statistically significant association.

We will test another hypothesis,

Ho: Sex of Casualty and Casualty type do not have a statistically significant association

HA: Sex of Casualty and Casualty type have a statistically significant association

To test this, we will use chi-square test,

We will consider 5% level of significance. So, value of α will be 0.05.

The R code will be,

**Casualty\_Type = as.factor ( x $ Casualty\_Type) # Transforming the numeric variable into factor variable**

**chisq.test ( Sex\_of\_Casualty ,Casualty\_Type )**

The R output will be,

A picture containing graphical user interface, text

Description automatically generated

Here we see, the p-value is less than 0.05.

Comment: We may reject out null hypothesis. So, we can say that, Sex of Casualty and Casualty type have a statistically significant association.

The third one we will test will be,

Ho: Bus or Coach Passenger and Casualty severity do not have a statistically significant association

HA: Bus or Coach Passenger and Casualty severity have a statistically significant association

To test this, we will use chi-square test,

We will consider 5% level of significance. So, value of α will be 0.05.

The R code will be,

**Bus\_or\_Coach\_Passenger = as.factor( x $ Bus\_or\_Coach\_Passenger ) # Transforming the numeric variable into factor variable**

**chisq.test(Casualty\_Severity,Bus\_or\_Coach\_Passenger)**

The R output will be,

Text

Description automatically generated

Here we see, the p-value is less than 0.05.

Comment: We may reject out null hypothesis. So we may conclude that, Bus or Coach Passenger and Casualty severity have a statistically significant association.

**Modelling**

From the overview section, we have seen that, we want to see if age of casualty and sex of casualty

have a significant effect on how sever the casualty case from the accident is.

Here, our independent variable will be age of casualty and sex of casualty. Our dependent variable will be casualty severity.

**Method:** The regression model we will be considering the conduct this analysis will be, multinomial logistic regression.

**Reason:** Our dependent variable here is casualty severity. It is a categorical variable with more than two categories. So, the best possible method will be multinomial logistic regression in this case.

**Procedure:** We would consider category 1 of the casualty severity as the reference category meaning our model interpretation will be mainly with respect to this category,

We will need to install package “nnet” in R to conduct the analysis.

The R codes for conduction the analysis is given below,

**library ( nnet )**

**levels(Casualty\_Severity)=list(Fatal="1",Serious="2",Slight="3")**

**Casualty\_Severity =relevel ( Casualty\_Severity , ref= "Fatal" )**

**# making category 1/ Fatal the reference category for analysis**

**levels( Sex\_of\_Casualty)= list(Male="1",Female="2")**

**Sex\_of\_Casualty=relevel(Sex\_of\_Casualty, ref="Female")**

**#creating reference variable of Female**

**Age\_of\_Casualty = x $ Age\_of\_Casualty**

**model = multinom ( Casualty\_Severity ~ Sex\_of\_Casualty + Age\_of\_Casualty )**

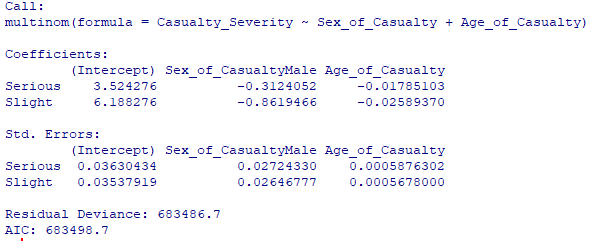
**summary( model )**

The output is as follows,

Text

Description automatically generated

Here we see that, it took 20 iterations to get a final model.



Here we find the coefficients and standard error for the model. But we will not just fit the model. We will check the p-value of the independent variable coefficients and decide which variables should we add into our final model.

Here we are considering 5% level of significance, meaning if the p-value is less than 0.05, we will conclude that the independent variable have a statistically significant effect on the dependent variable.

The R code for getting the p-value is given below,

**z = summary ( model ) $ coefficients / summary ( model ) $ standard.errors # calculation of test statistic value by diving the coefficients with the standard error**

**p =(1-pnorm(abs(z)))\*2**

**p**

The R output will be,

A picture containing text

Description automatically generated

It is clear that, all the variables have a p-value of less than 0.05 for both the categories of the dependent variable. So, both of them will be included in our model.

Our estimated model can be written as,

And

**Interpretation:**

From the first equation, we can write that, For sex of casualty male compared to sex of casualty female, the log odds of casualty severity being serious compared to being fatal is negative and it has the value of -0.312. For age of casualty, the log odds of casualty severity being in serious compared to fatal is negative and it has the value of -0.0178.

From the second equation, we can write that, For sex of casualty male compared to sex of casualty female, the log odds of casualty severity being slight compared to being fatal is negative and it has the value of -0.862. For age of casualty, the log odds of casualty severity being slight compared to fatal is negative and it has the value of -0.026.

**P-value interpretation:** Since, p-value is smaller than 0.05 for all the coefficients of the independent variables and for all the categories of the dependent variable, we can write that, Sex of casualty male compared to female and age of casualty both have a significant impact on log odds of casualty severity being less serious ( serious and slight) compared to being fatal.

In an easy language, we can say that, males have more chance of being in a comparatively fatal accident compared to females as their coefficient is negative. Also, a person with more age also has more chance of being in a comparatively fatal accident compared to less aged person because their coefficient is also negative.

**Conclusion:**

From the above discussion we have gotten a very good idea on the accident phenomenon of UK from 2005 to 2015. We found many aspects of the subject such as, the distribution of age of the casualty, the association between certain factors which can be accountable for this mishappening. Also, we have found two important factor that can account for the severity of an accident.