

**School of Business**

**OPIM 5604 – Statistics in Business Analytics Project**

Auto Insurance

21stApril, 2018

**Team Members:**

Charles Steiz

Juhi Maheshwari

Kumud Ranjan

Malgorzata Pozniak

Ruying Xia

**Table of Contents**

Executive Summary

1. Problem Statement
2. Introduction
   1. Data Description
3. Methodology
   1. Data Preprocessing
      1. Missing Value
      2. Variable Transformations
   2. Descriptive Statistics & Visualizations
      1. Correlation Analysis
      2. Thematic elements analysis
4. Hypothesis Testing
5. Data Modeling Process for Objective 1
   1. Logistic Regression
6. Data Modeling Process for Objective 2
7. Conclusion & Recommendations
8. References

# Executive Summary

This paper describes detailed analysis on Auto Insurance dataset, a free, user-maintained, online resource of vehicle crash details for different type of auto vehicles with owned by different categories of people. Insurance companies tend to charge huge amount of money for the cars to be insured. The rates differ depending on the value of the car, driving experience and number of other factors. During our analysis we have discovered interesting insight that are being considered when selecting rate for the potential client. We divided our analysis into 5 different themes: physical, personal life, assets, career and risk. Our two target factors were MVR points which are the points received when violating the law while driving and claim frequency. Some of the questions we wanted to answer were whether an income or education have any impact on MVR points, are red cars more likely going to get into accidents, are single or married individuals more responsible drivers, do car type play significant role when measuring the risk or whether homeowners are responsible drivers.

Our goal was to determine what factors are significant in deciding whether the prospective customer is risky or not for the insurance company and how likely the car is going to get into accident. The objective has been achieved by running multiple models such as logistic regression and zero-inflated poisson model on Target\_flag and Clm\_Freq as a target variable, respectively.

The statistical testing and analysis included some visualization tools such as pie charts, bar graphs, histograms, mosaic plots, etc. For logistic regression model built on Target\_flag, the RSquare value that we are getting is 0.32, while for Ordinal logistic regression model built on Clm\_Freq, we are getting RSquare value of 0.42.

# Problem Statement

The average number of accidents happening per year are rapidly increasing. With the increase in car accidents, number of claims made by an individual to an auto-insurance company also eventually increase. The main idea of our analysis was to measure how each factor in the data set affect Car Crash and insurance claim frequency. Target\_flag and Claim frequency were our target variables. We also wanted to figure out the patterns in the data that lead us to interesting insights. We started with clear exploration and understanding of the data, followed by pre processing and dealing with missing values and outliers, and finally coming up with hypothesis and question we desired to have answered.

# 

# 

# 

# Introduction

The auto insurance company provides financial assistance when physical damage or bodily injury result from the accidents or any such incidents with vehicles. Through this project, we are trying to help the insurance industry to predict the claim loss of an individual based on few features. This will help the industry to target prospective customers keeping such traits in mind and taking appropriate steps on the existing customer base.

## Data Description

This is a dataset with records of customer base of an auto insurance company. The dataset has 8161 rows with 27 variables. It is about the insurance of different types of car.

The dataset is taken from kaggle. The source of the dataset is: [https://www.google.com/url?hl=en&q=https://www.kaggle.com/c/auto-insurance-fall-2017/data&source=gmail&ust=1524196769074000&usg=AFQjCNE4X4ASxQx7TIScXJ3YZFKtDJEStw](https://www.kaggle.com/c/auto-insurance-fall-2017/data)

Target\_flag and CLM\_Freq are the 2 target variables for the dataset, where the Target\_flag talks about whether the car was in crash or not and CLM\_Freq is the number of claims filed by an individual in past 5 years.

Following is the description of all the variables of the dataset:

|  |  |
| --- | --- |
|  |  |

# Methodology

There are two types of target variables for which we tried two types of regression models

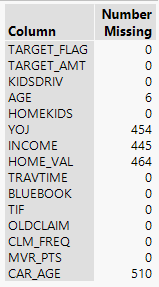
Logistic Regression: For identifying the factors affecting the car crash

Ordinal Logistic Regression: For identifying the factors affecting the Claim Frequency

## Data Preprocessing

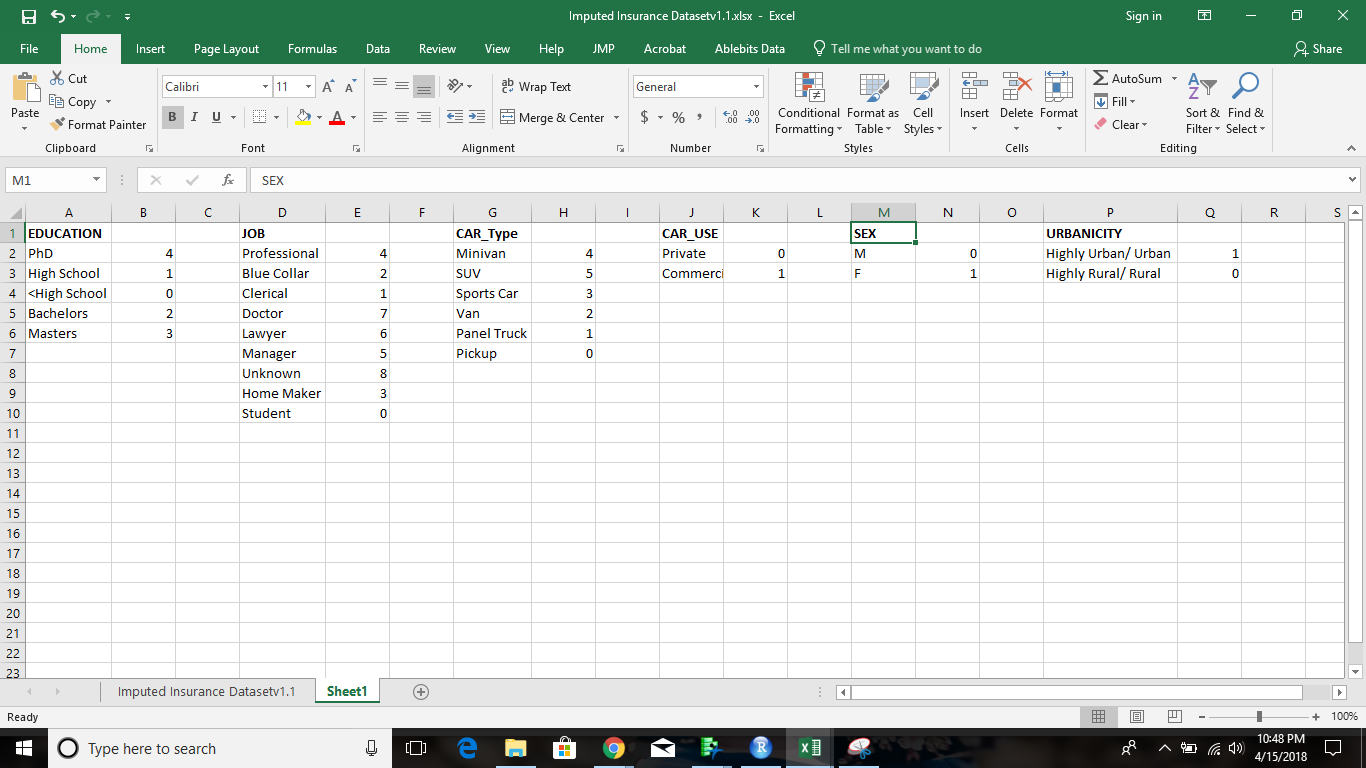
The overview of the dataset was done which lead to multiple insights and further exploration of the dataset. The dataset had multiple columns and rows with missing values. There were 6 columns including Job, Home\_val, Income, Car\_age, YOJ and age which had certain missing values. Here, Job is a categorical variable while Home\_val, Income, Car\_age, YOJ and age were the numerical variable. 526 rows had missing Job value. Getting rid of those rows would have lead to losing important trends that might be there in those rows. Hence all those rows ‘**unknown**’ value was given to Job column.

Below are the exact statistics of number of missing values for each numerical column

The AGE column had 6 missing values and looking at those records, we found no trends. Hence deleting those 6 rows was most appropriate solution.

YOJ, INCOME, HOME\_VAL and CAR\_AGE had 454, 445, 464 and 510 missing values respectively. Rather than putting in mean/median value, we decided to impute these values looking at different implied models/trends the dataset depicted. After imputing the missing values, we eye-balled the data to make sure the value-correctness of the data. Certain imputed values which did not make much sense, we modified the values to the most meaningful ones.

For the all columns with categorical values, their values were coded to the numerical values to make the modeling significantly relevant and easier. Below are the codes added for values of each categorical columns.

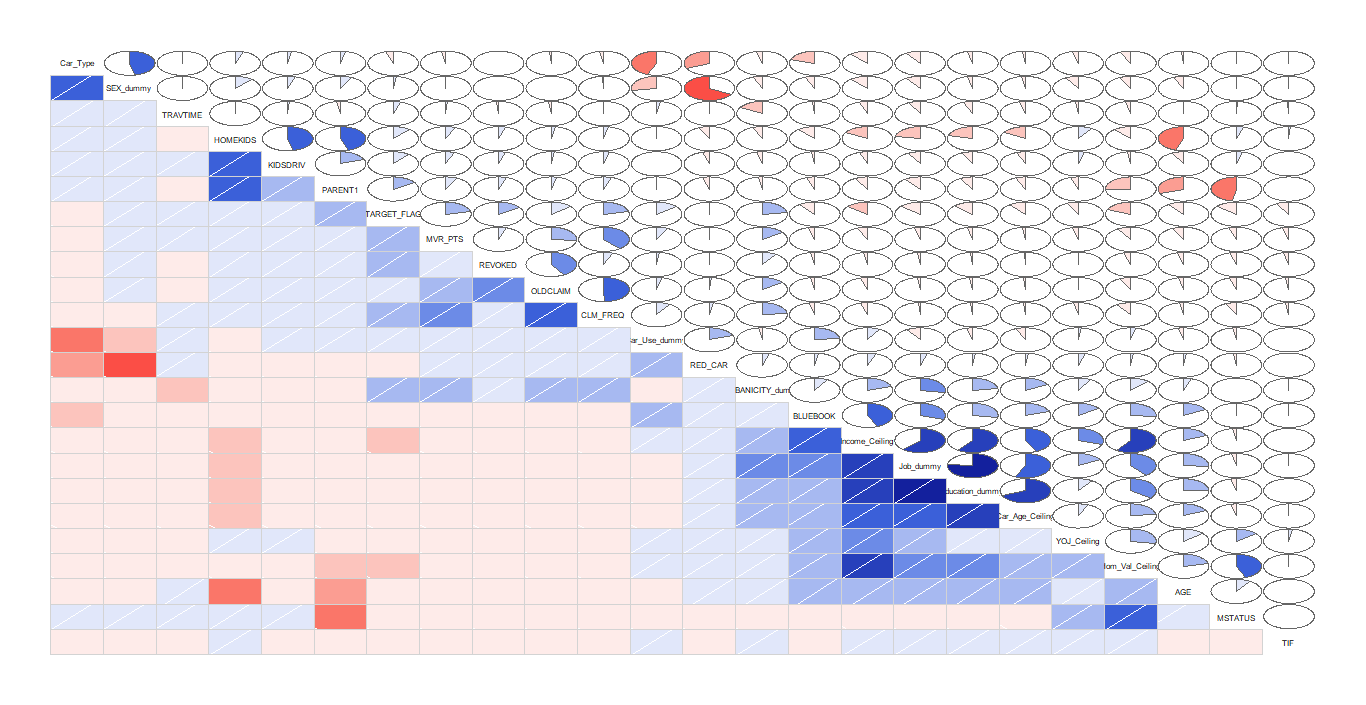


All the column values with yes/no values were coded to 0 and 1 respectively, where ‘0’ being No and ‘1’ depicting Yes.

## Descriptive Statistics & Visualizations

### Correlation Analysis

The correlations among the variables is assessed using corrgram package in R. The results can be visualised and interpreted from the below plot.

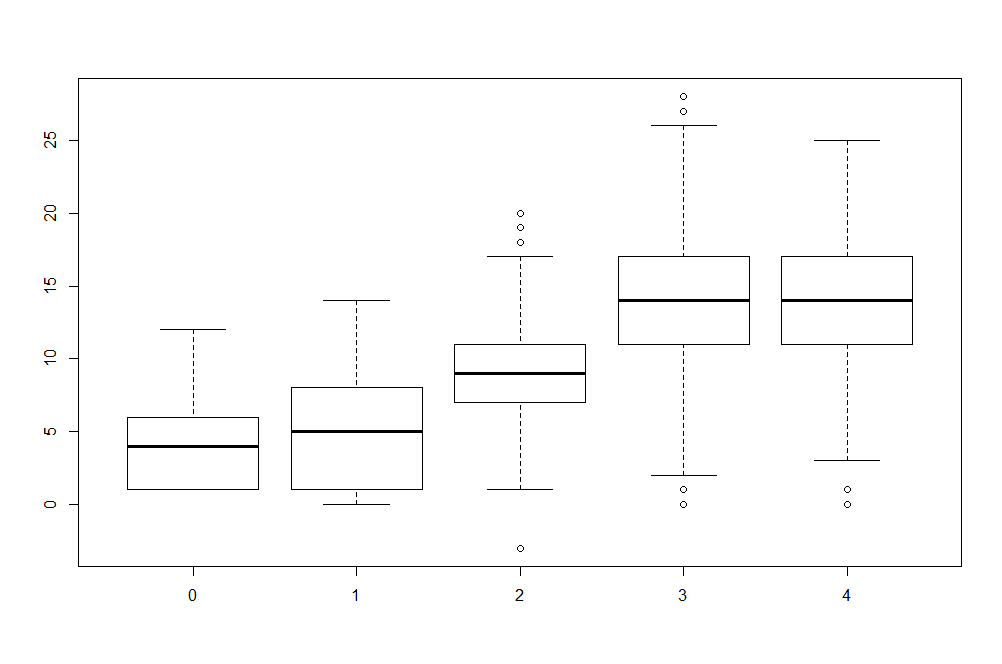


Some of the highly correlated variables were found to be as belows. The correlation between continuous variable were tested using cor() function using all the three methods e.g Pearson, Kendall and Spearman.

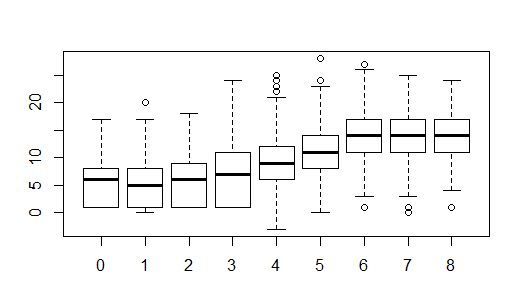
**Hom\_Val\_Ceiling and Income\_Ceiling:** The correlation coefficient is 0.59

**Education\_dummy and Job\_dummy:** Both are highly correlated (coef=0.76) and dependent on each other

**Car\_Age\_Ceiling and Education\_dummy:** Since one is continuous variable and another is an ordinal variable, We have created box plot to see the correlation. The mean as well median value increases as we move from 0 to 4 i.e. <High School to PhD.

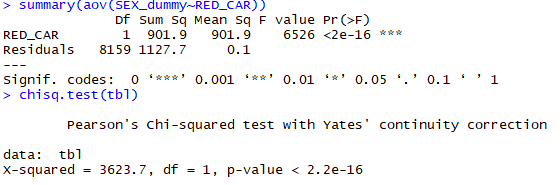


**Job\_dummy and Car\_Age\_Ceiling:** Since one is continuous variable and another is an ordinal variable, We have created box plot to see the correlation. The mean as well median value increases as we move from 0 to 8.



Some of the negative correlations were found to be:

**SEX\_dummy and RED\_CAR:** We performed chi-squared contingency table tests and one way anova test. Since p-value <0.05, we reject the null hypothesis i.e Sex(gender) and RED\_CAR is independent. Both depend on each other.



Also, we calculate correlation using cor() function which gives correlation as -0.66.

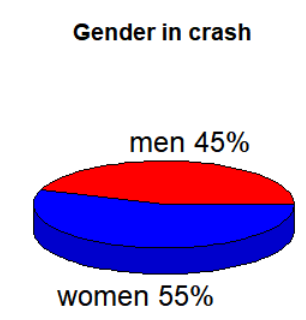
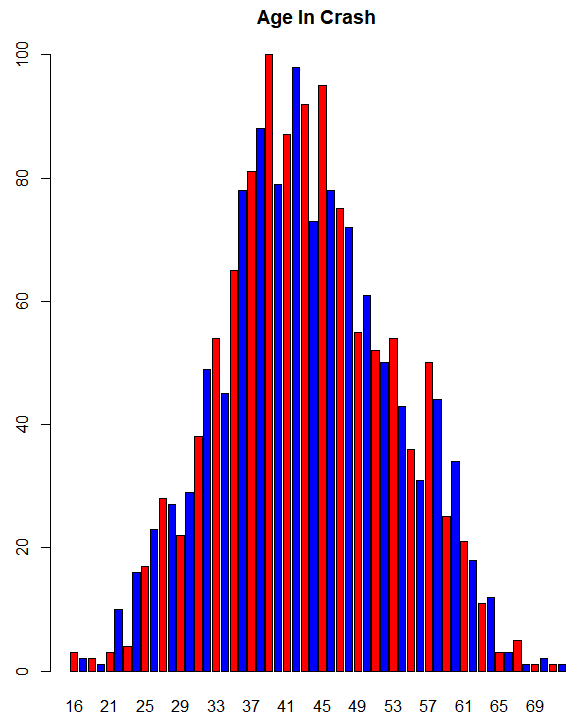
This means both are negatively correlated i.e more males (0) have red cars(1).

### Thematic elements of our dataset

We elected to group our variables into thematic groups. We hoped to take advantage of any real world relationships which might exist between the variables. We believed this would aid us in focusing on specific business aspects of the data. We grouped the variables in our dataset into five thematic areas: Physical, Personal Life, Career, Assets and Risk. A brief description of the theme and a list of the associated variables follows.

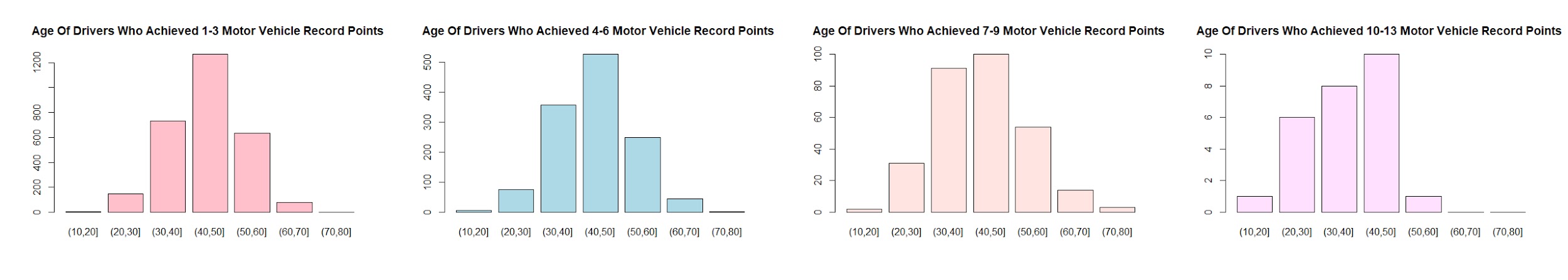
#### **Physical:** Physical characteristics of the policyholder

* Policyholder's age (AGE)
* Policyholder's gender (SEX\_dummy)

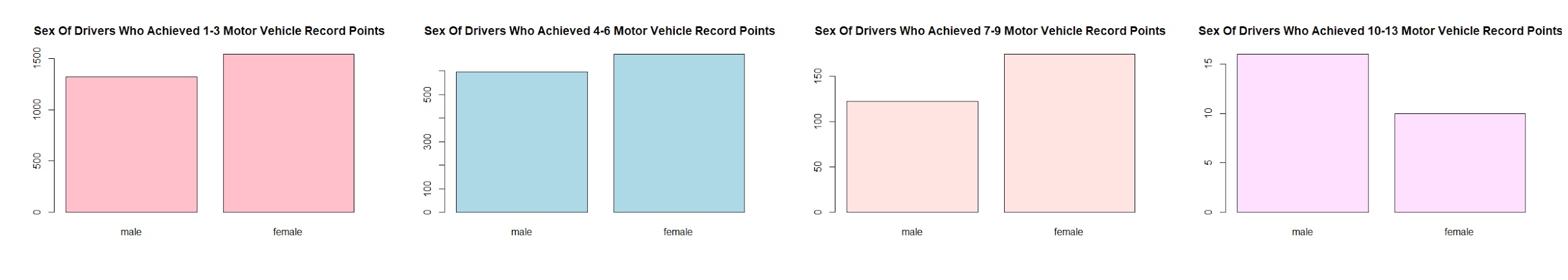


Age in Crash Distribution represents the age distribution of drivers who were in a car crash. According to this plot, we can see that drivers aged between 40 and 45 have highest rate of car accident.

As we can see from above pie chart, women drivers were found to be involved in 55 percent of all crashes, while men drivers were involved in 45 percent of all.



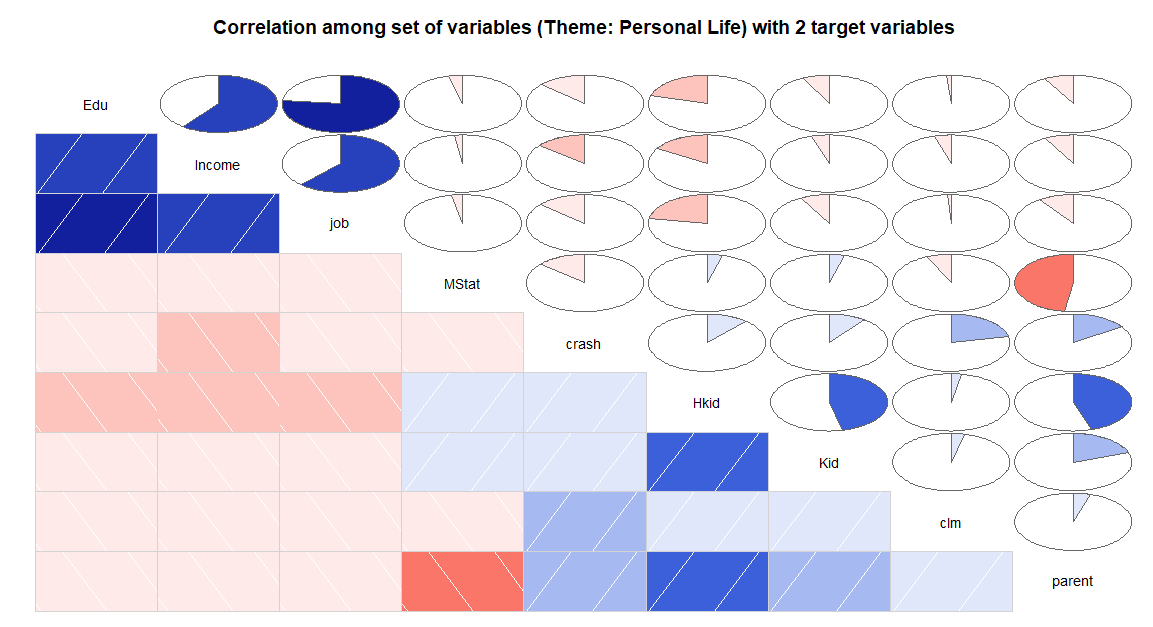
Above picture is a series of barplots about motor vehicle record points of drives grouped by age. We can conclude that among drivers who achieved motor vehicle record points, people aged between 40 and 50 are in the majority, which indicates that they are more likely to have illegal behaviours when driving a car. People aged between 30 and 40 are in the second place.



Above picture is a series of barplots about motor vehicle record points of drives grouped by sex. We can conclude that among drivers who achieved 1-9 record points, women drivers are in the majority. However, among drivers who achieved 10-13 record points, men drivers are in the first place. This result indicates that compared to women drivers, men drivers might commit more offences or more severe offences.

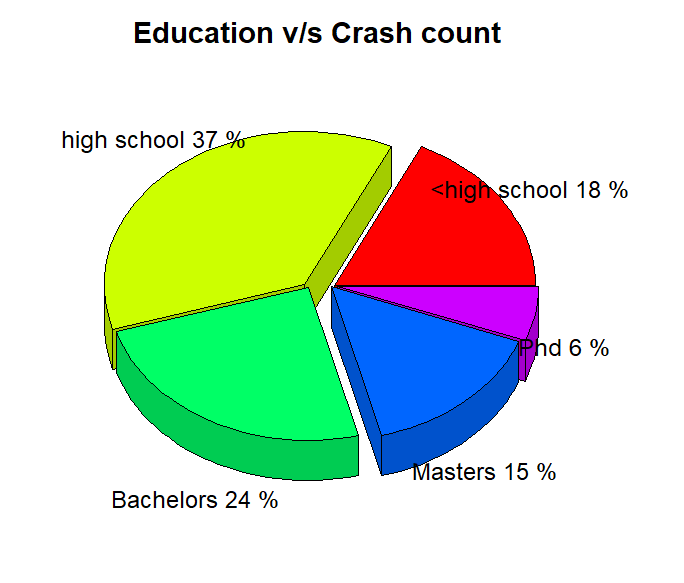
#### **Personal Life:** Details of the policyholder's life relevant to the analysis

* Policyholder's level of education (Education\_dummy)
* Policyholder's marital status (MSTATUS)
* Is the Policyholder a parent? (PARENT1)
* Policyholder has children of driving age? (KIDSDRIV)



Where,

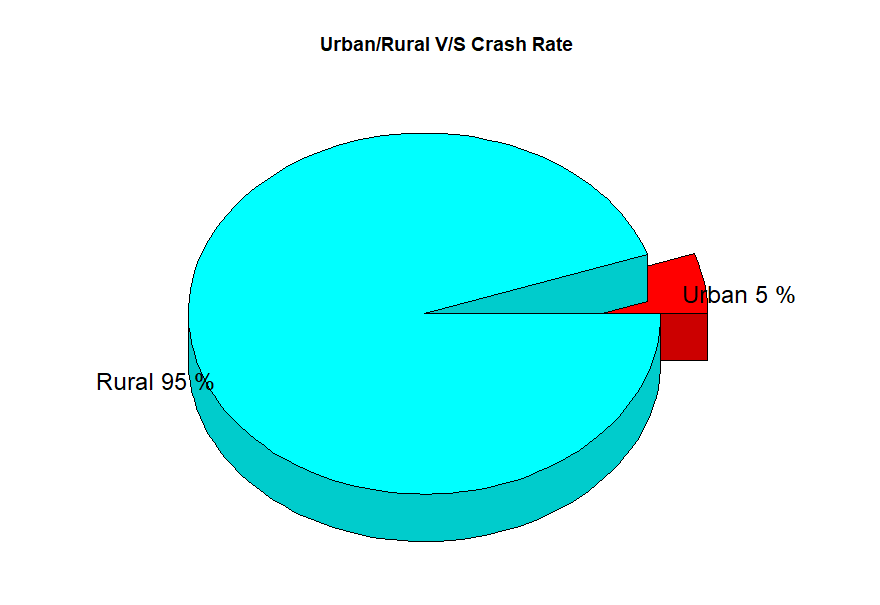
|  |  |
| --- | --- |
| Edu = Education of people | Hkid = Number of children @ home |
| Income = Income of people | Kid = Number of driving children |
| Job = Type of Job of an individual | Clm = Target variable: Number of claims in past 5 years |
| MStat = Marital Status of an individual | Parent = Whether parent is single or not |
| **Crash** **= Target variable:** Whether a car was in crash or not | |



The chart here represents the education level of a person v/s whether an individual got into crash or not. It seems that people with higher level of education, any type of college degree, are less likely to get into car accident. As the education level of an individual increases, the rate of people getting into crash decreases.

Analyzing further about the education and the crash rate, we found that the education is not only the factor that affects the crash rate.

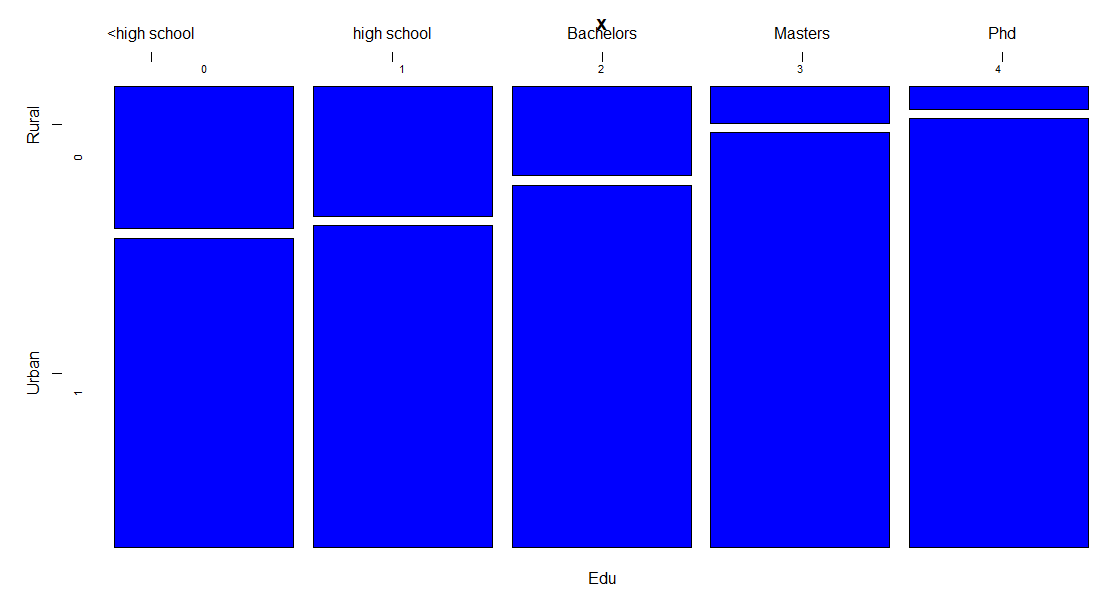
The below graph shows that of all the people who get into accident 95% of people live in Rural area while only 5 % of people live in the Urban area.



The study says that people with less education often live in poorer neighborhood and rural communities. These areas generally have bad conditioned roads which is one of the factor, a person gets his car into an accident.

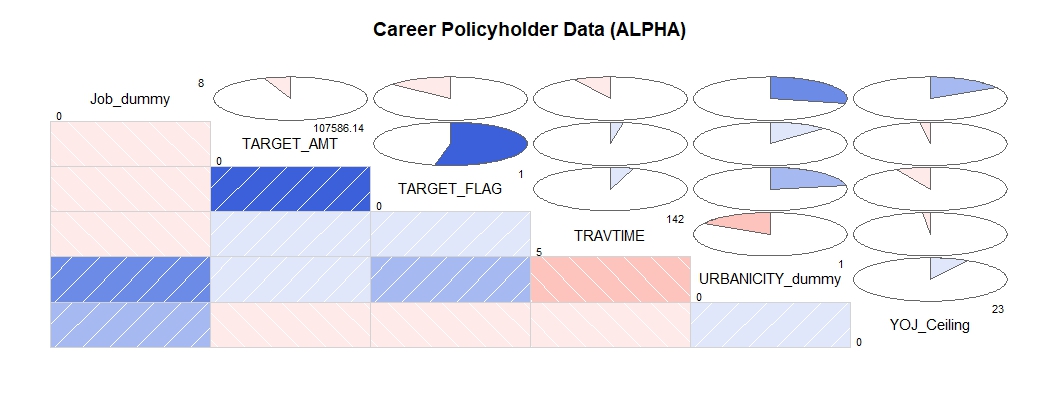
Applying this knowledge to the dataset, we found a similar kind of result. As the level of education increases, the percentage of people living in urban area increases.

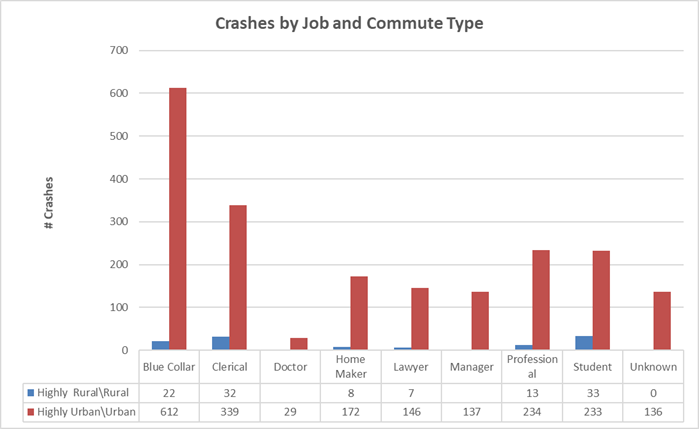
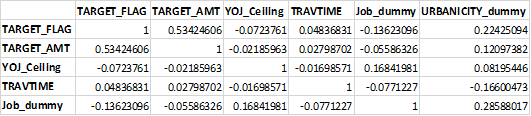
Below is the mosaic plot which will clearly explain the result.



#### **Career** **:** Details of the policyholders job and commute

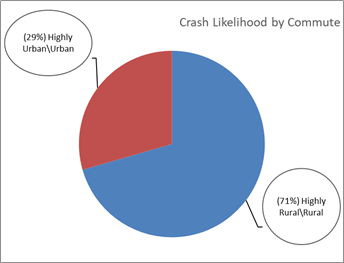
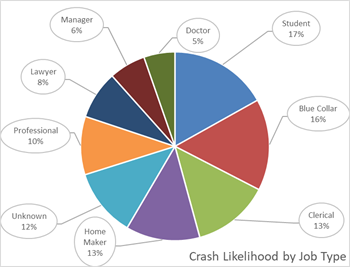
* What does policyholder currently do for a living? (Job\_dummy)
* How long has policyholder been employed in current role? (YOJ\_ceiling)
* What type of environment does the policyholder commute through? (URBANICITY\_dummy)
* How much time does the policyholder spend commuting daily? (TRAVTIME)

Examining the correlation between the career variables and the crash count and claim amount, we did not note any especially strong correlations within the career variables. 



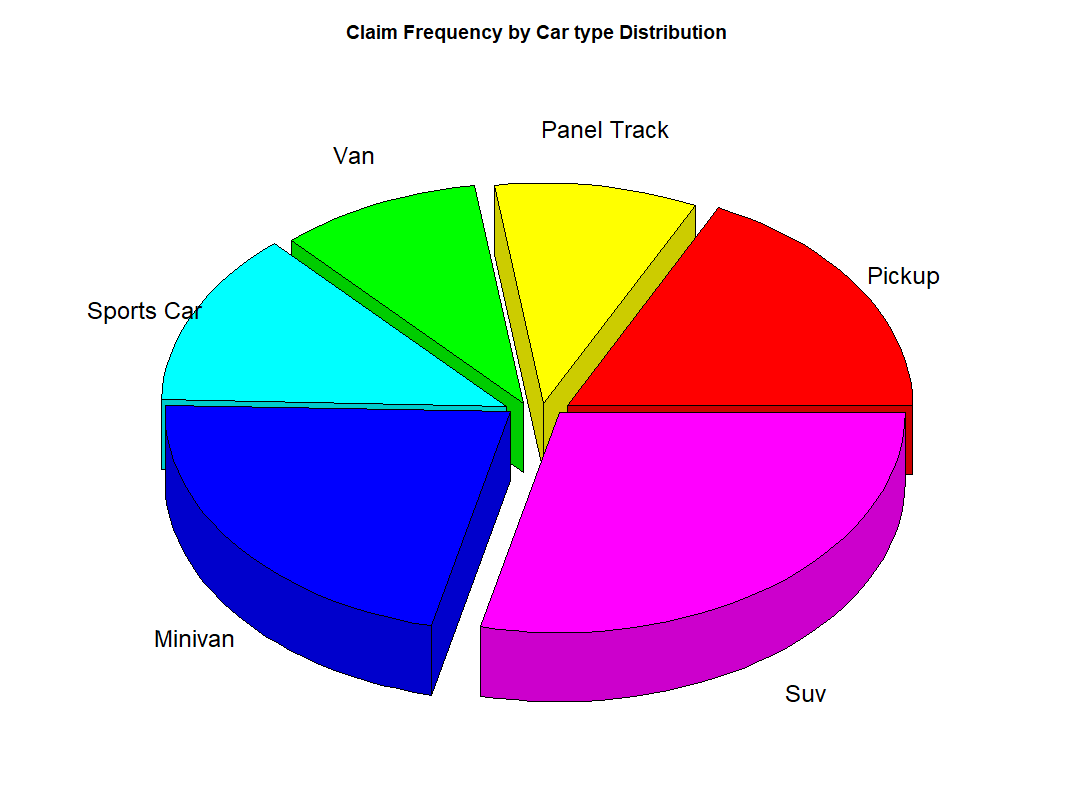
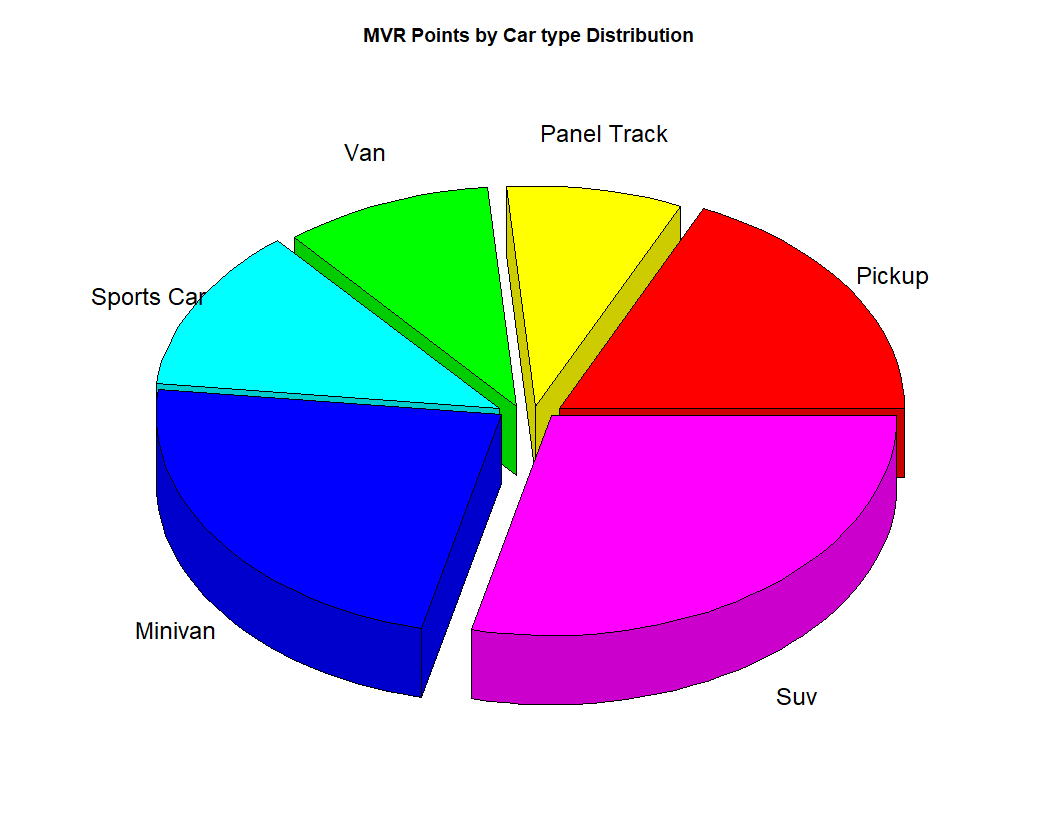
We looked at the total crashes based on the type of commute, broken out by the commuters’ job types. We saw that the Urban environments definitively posed the greatest accident risk. Blue Collar and Clerical workers had the 1st and 2nd most Urban crashes and the 2nd and 3rd most Rural crashes. Students had the highest total number of rural crashes and fourth highest number of Urban crashes. Rounding out the top four for both environments, people classed as Professionals had the third highest number of Urban crashes and fourth highest number of Rural crashes.

When we visualized the data on urbanicity and job type related to crashes at an aggregate level, we were interested to note some changes in the appearance of the results. Viewing the data in pie charts helped us to see that fewer rural drivers actually meant that more rural drivers per capita were getting into accidents.

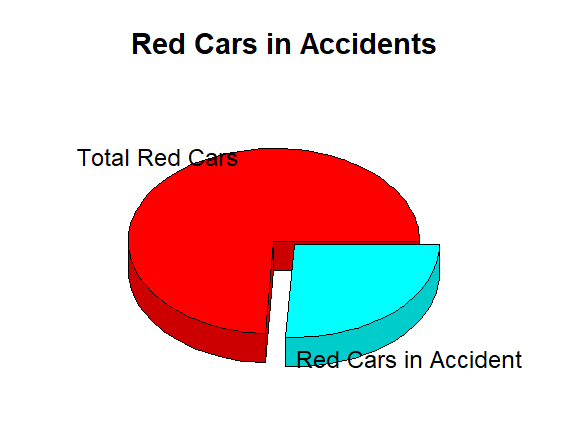
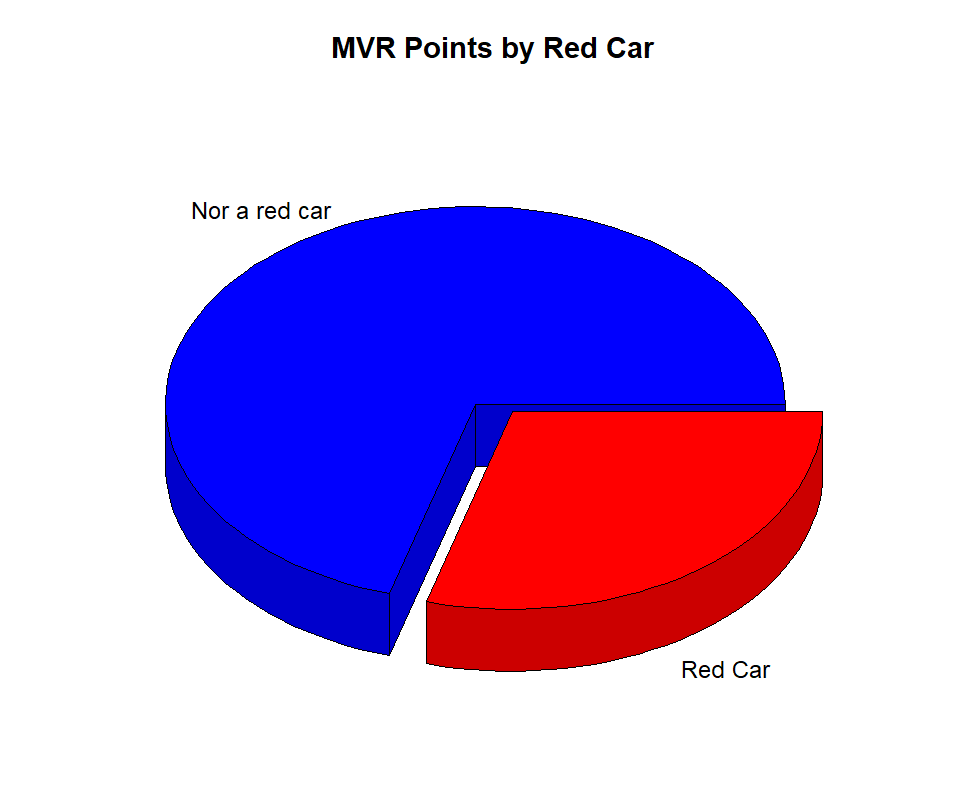


#### **Assets:** Details of the major physical belongings of the policyholder (home and vehicle)

* What is the value of the policyholder's home? (Hom\_Val\_Ceiling)
* What type of car does the policyholder drive? (Car\_Type)
* What is the current Kelly Blue Book value of the vehicle? (BLUEBOOK)
* How does the policyholder make use of the vehicle (Car\_Use\_dummy)
* Is the car red? (RED\_CAR)



MVR Points by Car type Distribution represents the number of MVR points in the data distributed by the car type. According to the results SUV drivers commit driving felony the most often and are considered the least responsible drivers. SUV driver are responsible for 30% of the points in the data, followed by Minivans with 23%, followed by Pickup Trucks with 18% of the data. Surprisingly, sports car are listed 4th and are responsible for only 12% points in the data. The distribution among claim frequency shows similar results. SUV’s hold 30% of the claims, followed by Minivans with 21%, followed by Pickup Trucks with 18%. Sports car are listed only with 13% of the claim frequency among entire dataset.

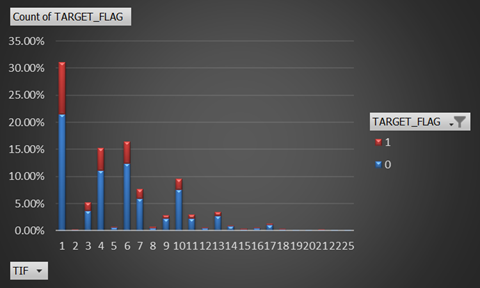


Another interesting insight we have have discovered during our analysis is about the red car and it’s crash frequency and how save red car drivers are. Among entire data set, red car driver are responsible of the 30% of the MVR points. However, among all red car drivers, almost 30% have been in an accident.

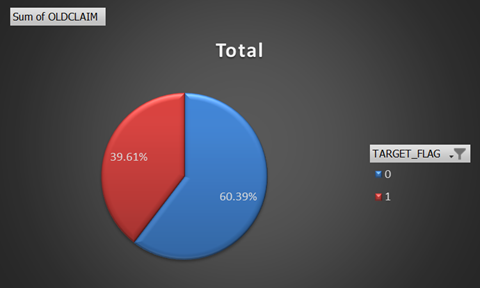
#### **Risk:** Details of the policyholder's life which impact the riskiness of the policy.

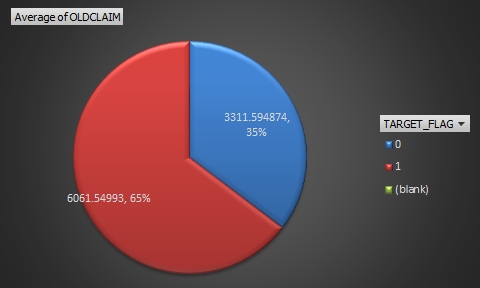
* Policyholder's accumulated points on their driver's license in last 5 years. (MVR\_PTS)
* Policyholder's license has been revoked in the last 7 years. (REVOKED)
* Length in years that this policy has been in force with the carrier (TIF)
* The total amount in dollars of claims over the last five years. (OLDCLAIM)
* Number of previous accident claims the policyholder has filed in the last five years (CLM\_FREQ)

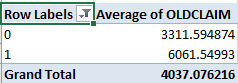
**Time in Force(TIF):** Time in Force is the Length in years that this policy has been in force with the carrier . Our analysis reveals that 89.2% of our policy are having TIF less than or equal 10 years. Also,the most of the crash belong to TIF=1,3,4,6,7 and 10 years.

****

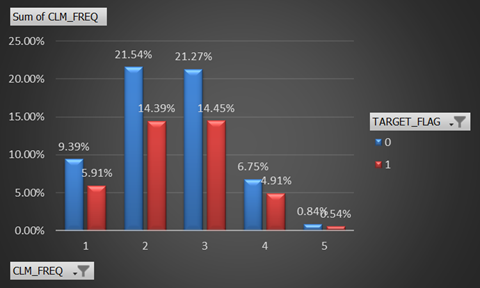
**OLDCLAIM:** This is the total amount in dollars of claims over the last five years. Data reveals, that 39.6% claim amount where involved in crash whereas 60.4% claim amount was not involved in crash. However,the average claim amount for crash was 6061 i.e. 65% of total average.



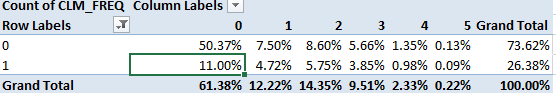




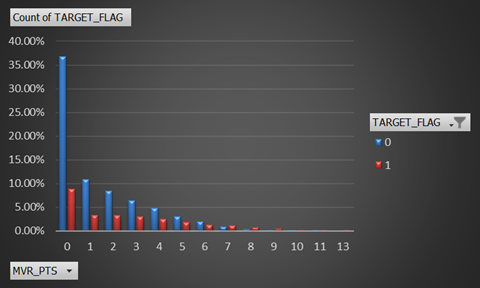
**CLM\_FREQ:** It is the Number of previous accident claims the policyholder has filed in the last five years . From the data, it is shown that 40% of claims were filed in case of crash.

****

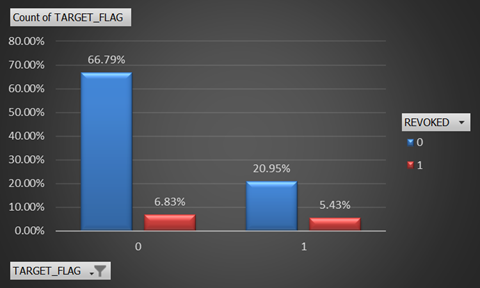
Also, 11% of crashes did not result into any claim.



**MVR\_PTS: It is the** Policyholder's accumulated points on their driver's license in last 5 years. For the graph below, it is clear that number of crashes decreases with MVR Points.Also, the average MVR point for vehicle involving in crash is 2.48



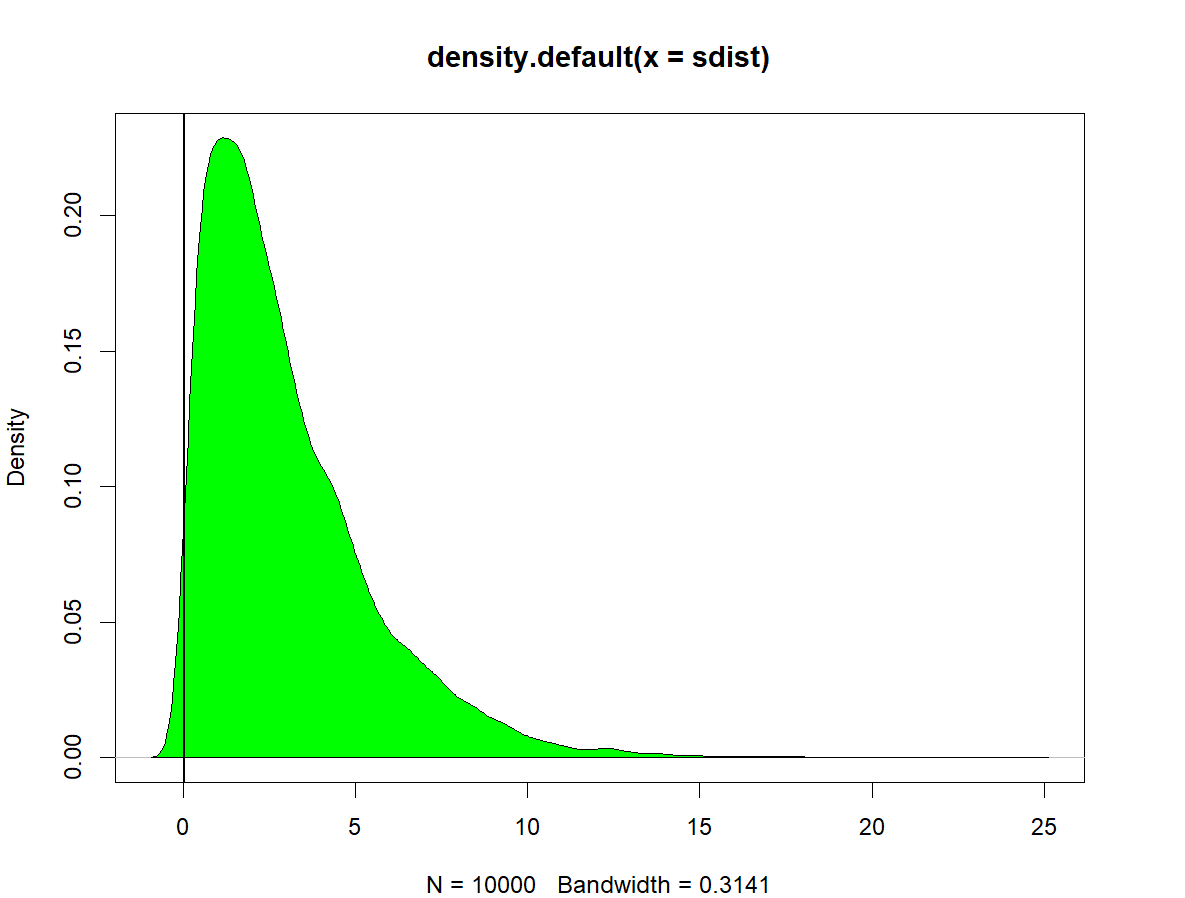
**REVOKED:** This is a binary variable (1=Yes, 0-No). It means if Policyholder's license has been revoked in the last 7 years. The data shows us that out of 12.25% of license revoked cases ,5.43% met with a crash. Also, 44.3% cases having license revoked met with a crash.

****

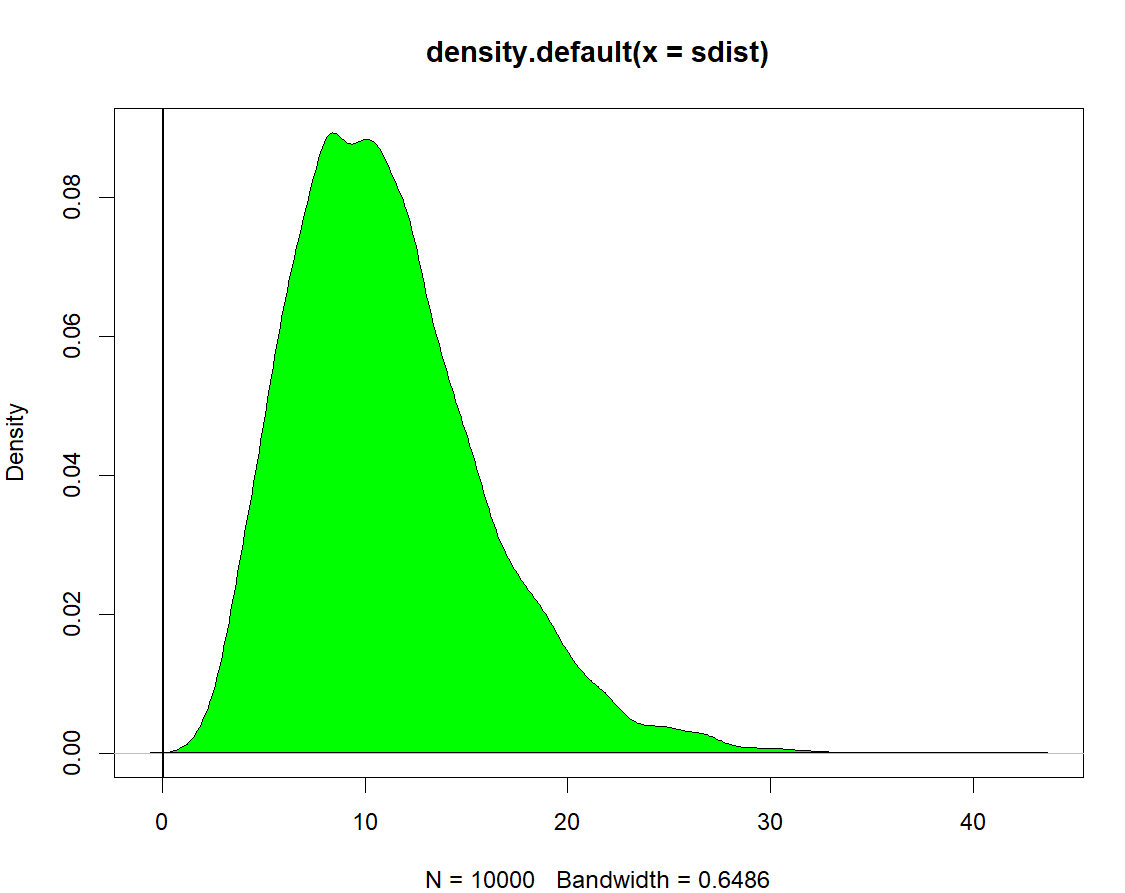
# 

# Hypothesis Testing

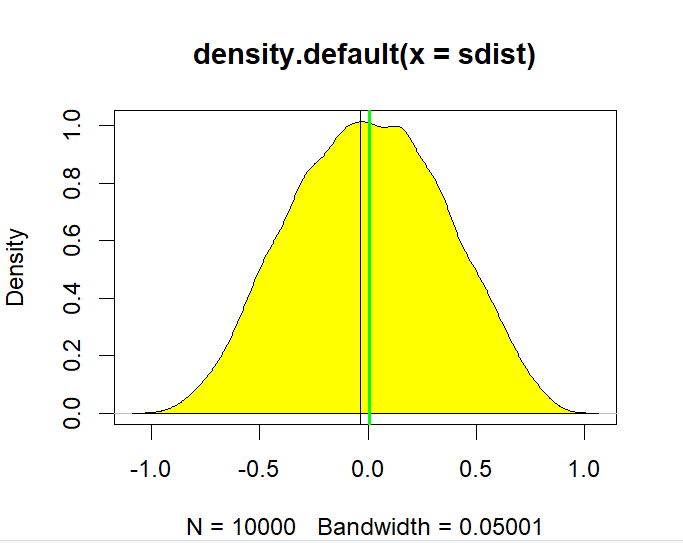
The first hypothesis we have analyzed is whether red car driver is more likely to get into accident. The hypothesis was that red car has no impact on potential crash. We computed two sample Test with two categorical variables: RED\_CAR (“yes”, “no”) and TARGET\_FLAG (0, 1). According to high pvalue of 0.9988, we failed to reject our hypothesis and concluded that the color of the car has no impact on crash.



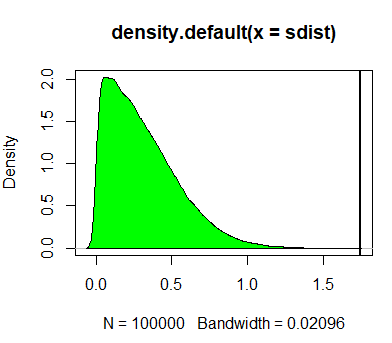
Additionally to the color of the car, we have analyzed whether car type have any impact on crash potential. Our hypothesis was that there is relation between car type and crash. We computed two sample test with two categorical variables: CAR\_TYPE (0, 1, 2, 3, 4, 5) and TARGET\_FLAG (0, 1). According to very high pvalue of 1 we failed to reject our hypothesis, and concluded that there is relation between car type and risk of crashing.



Another hypothesis we have analyzed, was whether the value of the car have any relation with crash frequency. Our hypothesis stated that the car does have an impact. We have computed correlation test with two continuous variables: CLM\_FREQ and BLUEBOOK. High pvalue of 0.9175 indicated that the hypothesis was correct and car value do have an impact on claim frequency.



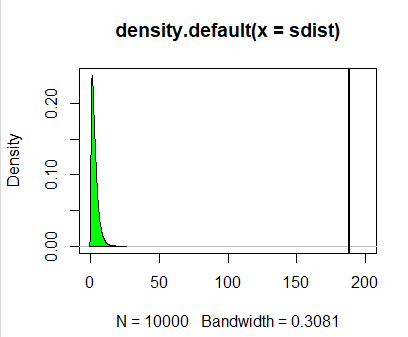
We also tried to test whether the distance to work i.e. distance travelled by drivers to his/her workplays has any role to play in any crash.Our null hypothesis was that it should play no role. To test this hypothesis,we did Two sample test with numeric data (TRAVTIME) and categorical data (TARGET\_FLAG).Since P value is 0<0.05, which means we will reject the null hypothesis.The distance to work does have an impact on cars getting into accident.People who drive longer tend to get into crashes more often than others.



We have also tried to see if licence revoked and involvement into crash are independent events.

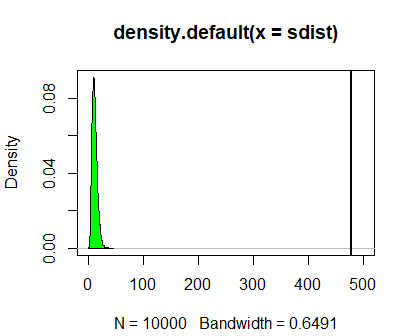
Our null hypothesis is : License Revoked in the Past 7 Years and Involvement in crash are independent.For this we did chi square test between categorical variables TARGET\_FLAG,REVOKED.

Since P value is 0<0.05, which means we will reject the null hypothesis.Both are dependent. A car involved in crash is likely to have license revoked in the past.



We have also tried to see whether probability of crash has any dependence on the no of claims made by a driver in past. Our Null hypothesis:Claim frequency in past and involvement in crash are not dependent.for this we have done chi square test on the variables CLMFREQ and TARGET\_FLAG.

The Low p-value indicates to reject null hypothesis. Since p value is 0, we reject null hypo. We can conclude that having claimed once increases chances to crash again.

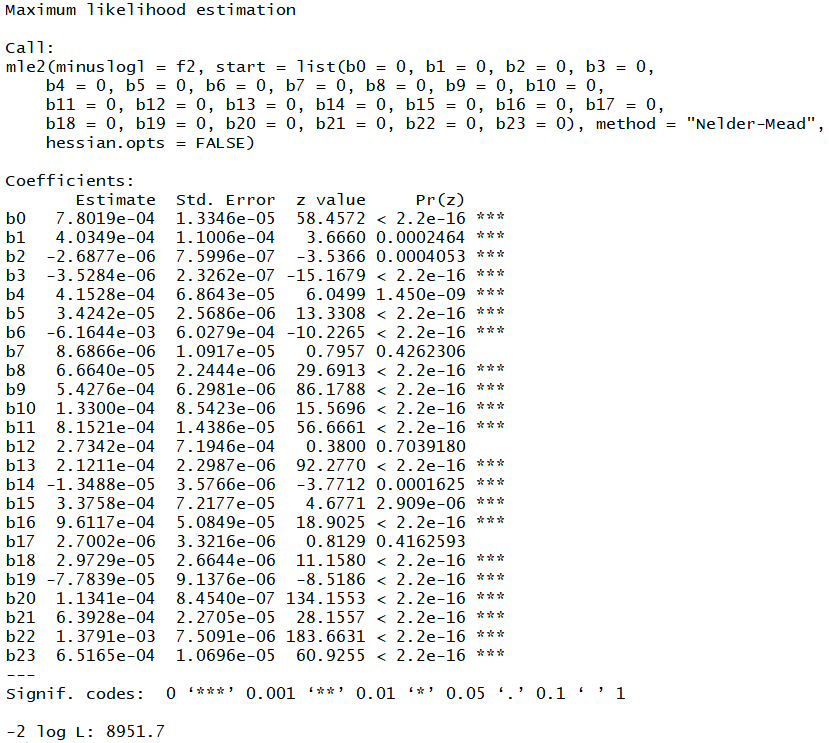


# Data Modeling Process for Objective 1

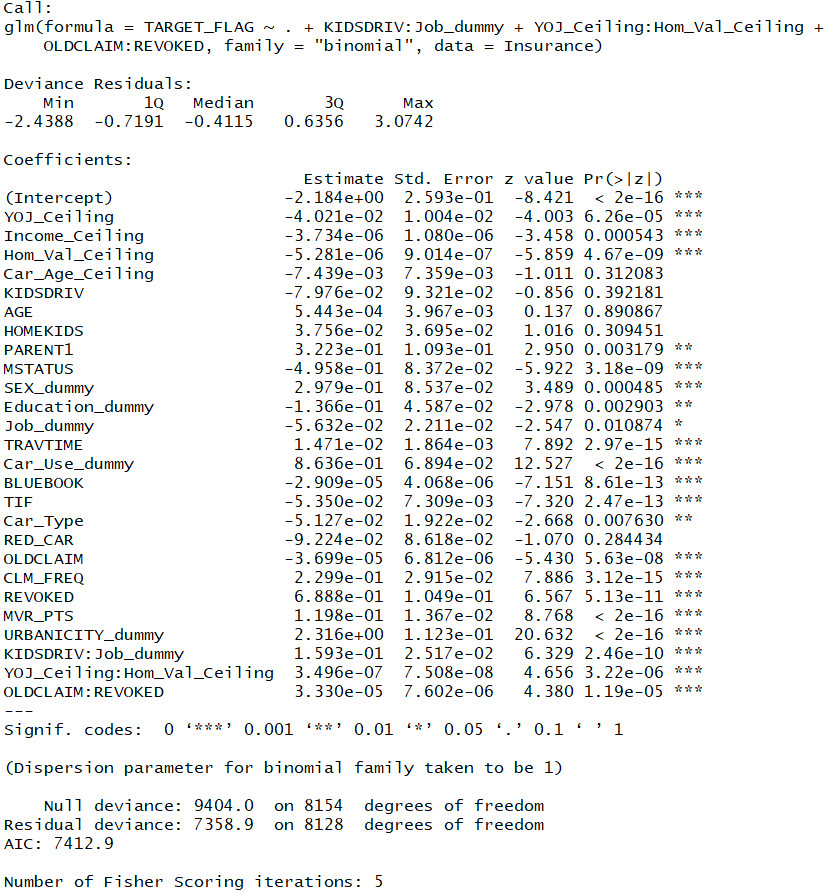
## Logistic Regression

The estimation aims to answer which predictors have the most impact on the target variable that is a binary variable (consisting of 1 and 0). To achieve this, we conducted a logistic regression using mle function that  
explains the impact of each variable. We included all predictors except INDEX to fit model.

The combination of predictors can be shown as log(probability(car was in a crash)/probability(car wasn’t in a crash))=b0+b1\*Insurance$YOJ\_Ceiling+b2\*Insurance$Income\_Ceiling+b3\*Insurance$Hom\_Val\_Ceiling+b4\*Insurance$Car\_Age\_Ceiling+b5\*Insurance$KIDSDRIV+b6\*Insurance$AGE+b7\*Insurance$HOMEKIDS+b8\*Insurance$PARENT1+b9\*Insurance$MSTATUS+b10\*Insurance$SEX\_dummy+b11\*Insurance$Job\_dummy+b12\*Insurance$TRAVTIME+b13\*Insurance$Car\_Use\_dummy+b14\*Insurance$BLUEBOOK+b15\*Insurance$TIF+b16\*Insurance$Car\_Type+b17\*Insurance$RED\_CAR+b18\*Insurance$OLDCLAIM+b19\*Insurance$CLM\_FREQ+b20\*Insurance$REVOKED+b21\*Insurance$MVR\_PTS+b22\*Insurance$URBANICITY\_dummy+b23\*Insurance$Education\_dummy. Then we used the function mle2() to find the maximum likelihood parameter value. The summarized  
coefficient is shown as follow:



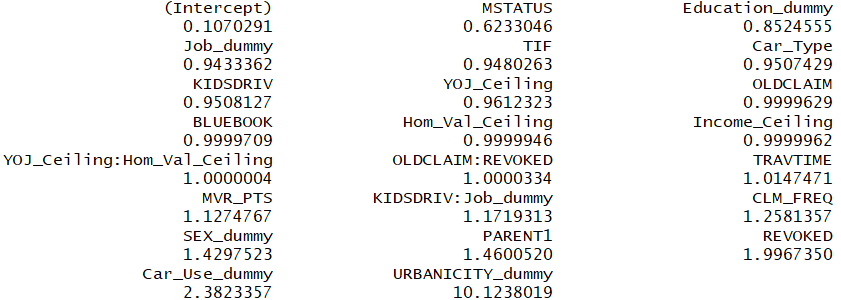
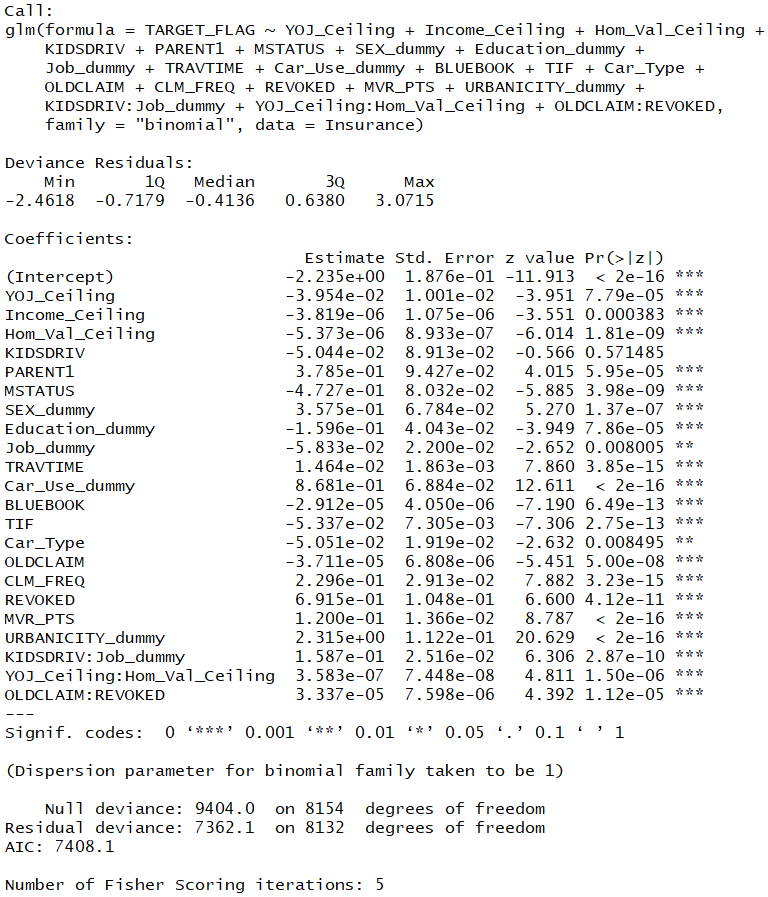
We can conclude that HOMEKIDS, RED\_CAR and TRAVTIME are not significant.

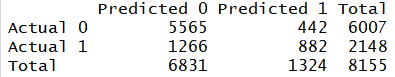
**Regression model using glm() function:** After finding logistic results using mle2(), we built another model based on logistic regression function glm(). First, we excluded the first column to exclude INDEX variable. And then we applied glm() function to do the logistic regression of outcome, the function is shown as follows.

The result shows AIC as 7412.9. Car\_Use\_dummy is the top variable that impacts the outcome.

To optimize the model, we used the function step(regression model, ~.^2) to evaluate all possible interaction terms. The interaction terms that we chose to include in our model are KIDSDRIV:Job\_dummy, YOJ\_Ceiling:Hom\_Val\_Ceiling, OLDCLAIM:REVOKED. From this, we built reg2 and we found a significant  
improvement in AIC from 7491.014 to 7412.888.

Afterwards, we used step <- stepAIC(reg2, direction="both") function to further optimize the model. According to result, we removed four predictors: AGE, Car\_Age\_Ceiling, HOMEKIDS, RED\_CAR and got final model. AIC drops from 7412.888 to 7408.1. The summary and odds ratios of final model are shown as follows.



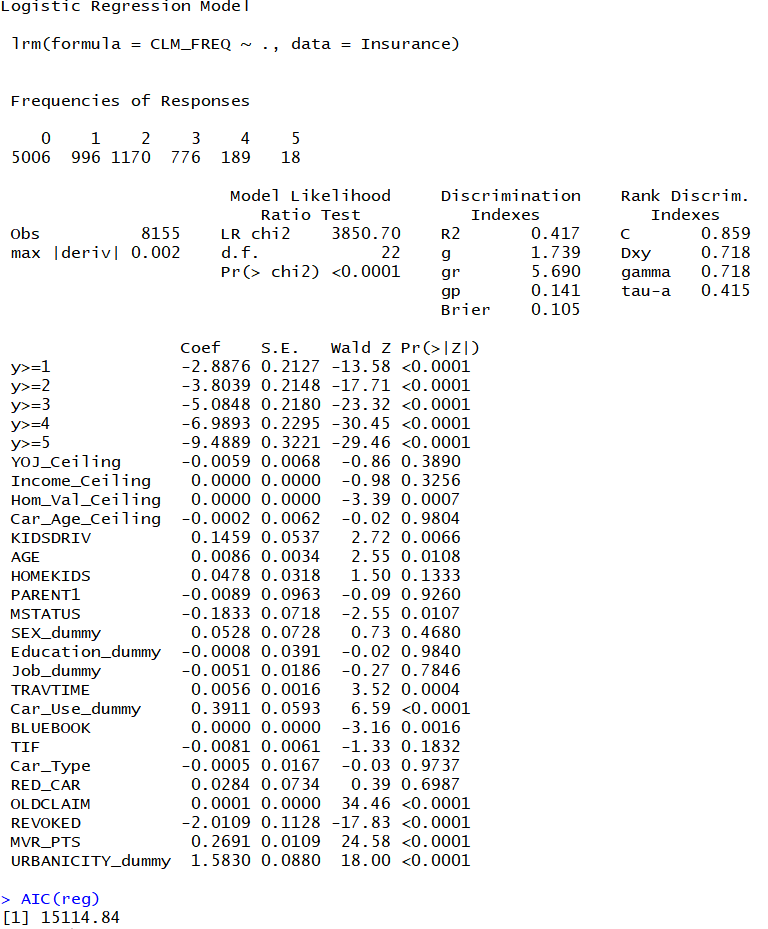


The above is confusion matrix of the third model. Overall accuracy is 79.06% and Precision is 66.62%.

# Data Modeling Process for Objective 2

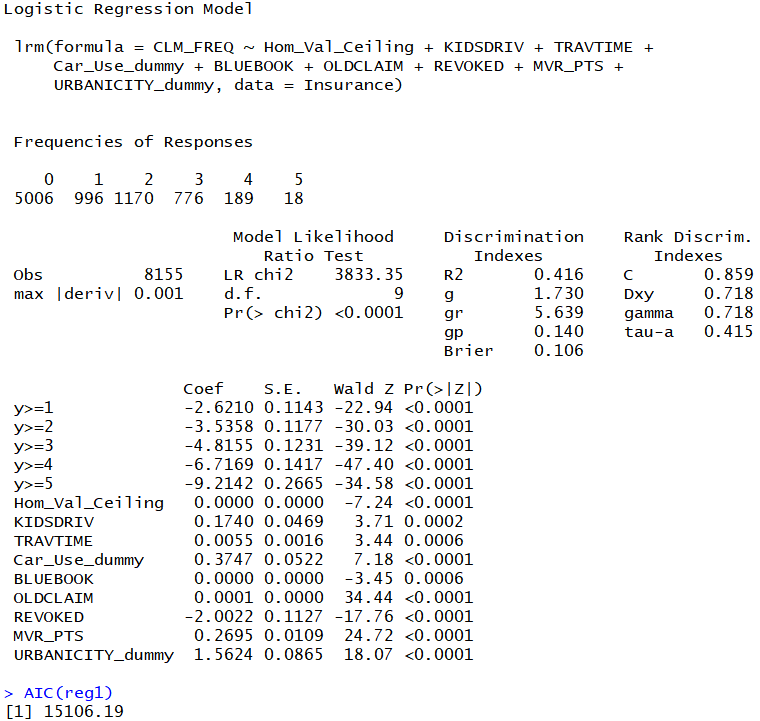
**Ordinal Logistic Regression**

Our aim is to answer which predictors have the most impact on the claim frequency in past five years(CLM\_FREQ). So we built an ordinal logistic regression model using lrm() function because this CLM\_FREQ is an ordinal dependent variable. We included all predictor except INDEX and TARGET\_FLAG to fit the model. Report is shown as follows.



The result show AIC is 15114.84. The top significant predictor is REVOKED.

And then we used fastbw() function to further optimize the model. According to the result, we removed some predictors and 9 predictors remain. All remaining predictors are significant and AIC drops from 15114.84 to 15106.19.



# 

# Conclusion & Recommendations

R-Omatics performed statistical testing and observed patterns for independent variables and assessed their impact on:

* occurrence of accidents
* frequency of accident claims

We noted that cumulative license points in the last five years (MVR\_PTS) and the revocation of a driver’s license in the last 7 years were positively correlated with crashes. We established that certain “common knowledge” about auto insurance did not hold up to statistical analysis. Example: Our analysis supported the hypothesis that drivers of red cars are not at greater risk of a crash.

The riskiest customers tend to be the youngest (16), oldest (70+) and middle aged (40-60). Drivers with a history of moving vehicle violations (license points and revoked licenses) also demonstrated a tendency to crash. People with four categories of professions demonstrated the greatest chance of crashes: blue collar workers, clerical workers, students and professionals.

We found that drivers with SUVs, pickups and minivans had both the greatest chances of receiving points from moving vehicle violations and filing claims over the last five years. We also dispelled the myth that red car drivers were more prone to accidents. Drivers with high school and undergraduate degrees show the greatest tendency for accidents, possibly in line with the distribution of this education in the population. We established a link between the travel distance of the policyholder and the chance of crashing.

We noted that certain car types (SUV’s, minivans and pickup trucks) were more strongly correlated with having points on a driver’s license for motor vehicle infractions and the number of claims a policyholder had in the last five years. Barring collinearity, these car types would also be associated with crashes. We also tested our hypotheses of the association between car value and crashes. We confirmed that the value of the car was associated with the occurrence of a crash.

We propose the following applications of our models:

* Selective discount pricing for low risk policyholders
* Premiums for history of revoked license
* One-time drivers courses to address risk factors

References

<https://www.mcminnlaw.com/the-relationship-between-education-and-accident/>

<https://www.firstdependable.com/blog/2016/09/5-statistics-that-illustrate-the-current-auto-insurance-industry>

<https://www.statista.com/topics/3087/car-insurance-in-the-united-states/>

<https://visual.ly/community/infographic/transportation/auto-insurance-facts-and-interesting-statistics>

<https://www.cheapcarinsurance.net/auto-insurance-statistics-2015/>

<http://www.insuranceqna.com/auto-insurance/curious-auto-insurance-statistics.html>