IDS 570
Statistics for Management

AUTO INSURANCE COMPANY DATA ANALYSIS

Team Members:
Ashutosh Dwivedi (674351427)
Mohammed Rehan (668333717)
Srija Gupta (656331239)

Table of Contents

Executive Summary

The report contains the steps used to analyze sample of the claims data collected by auto-insurance provider, Indian Money, Bangalore, India. The purpose of this report is to predict the profitability of the company

Introduction

Dataset is a sample of data maintained by one of the auto-insurance providers named Indian Money, Bangalore, India. One of our friend is working for this organization and we have received the data from her.Data is collected by the organization during claims processing and reporting of claim data at the end of the year.

Dataset

Variable	**Datatype**	**Description**
Policy Number	Numeric	Unique Policy Identifier
Year	Factor	Year of manufacture
IDV	Numeric	Insured Declared Value of Car
City	Factor	City of registration of vehicle
State	Factor	State of registration
Cubic Capacity	Numeric	Capacity of engine
Mfr-Model	Factor	Manufacture and model of car
Premium	Numeric	Total Premium paid for the policy at the beginning of the term
Type	Factor	Source of lead
Gender	Factor	Male or Female
Channel	Factor	Lead generation channel
Age Group	Factor	Age Group of applicants

Payment Frequency	Factor	Annual payment or monthly instalments
ClaimsInd	Factor	Claims taken (0 - Not taken|1 - Taken)
Claim Amount	Numeric	Claims Amount

Research Question

- What factors affect the profitability of Indian Money Insurance company?
- Is there an impact of geography when it comes to the profitability of auto insurance?
- Does Age Group or Gender affect the profitability of auto Insurance?

Hypothesis

The profitability of Indian Money is higher for female drivers in North zone.

The profitability of Indian Money is higher for drivers having age more than 40 years.

The profitability of Indian Money is higher for vehicles with cubic capacity more than 2200.

Variables used for Analysis

Name	Data Type	Variable	Description
Year	Factor	CV	7 Years of Data
IDV	Numeric	IV	Insured Declared value of Car
Gender	Factor	IV	Male or Female

Zone	Factor	IV	Divided Regions of the Country
Age Group	Factor	IV	Age groups of Applicants
ClaimsInd	Factor	IV	Claims Take(0-not taken,1-Taken)
Vehicle Category	Factor	IV	Clubbed to Cubic Capacity size
Revenue	Numeric	DV	Derived from Premium minus Claim
Profit	**Numeric**	**DV**	**Profitability - Derived column(Revenue %)**

Detailed Explanation of variables.

Dataset has sample of the claims data collected by auto-insurance provider, Indian Money, Bangalore, India

Collected by the organization during claims processing and reporting of claim data at the end of the year

Dataset has 15 variables and 7702 observations

Derived 4 variables: Zone, Vehicle Category, Revenue, Profit

Dependent Variable: Premium and Claim Amount

Independent variables: Age Group, IDV, Gender, Zone, Vehicle Category

Zone: We have in total four geographical zones: North,South,East,West.

Vehicle Category: This is a derived column data according to the cubic capacity of the vehicles.

CC-large sized (cubic capacity >1800):

CC-medium sized (cubic capacity >1250 and <1800):

CC-small sized (cubic capacity <1250): ********

Premium: This is the total Premium paid for the policy at the beginning of the term.

Claim Amount: This is the amount claimed by the customers.

Revenue: This is a derived variable which we calculated by subtracting the Claim Amount from Premium Value.

Profit: This is the proportion of revenue calculated by the formula (Revenue/maximum Revenue)*100

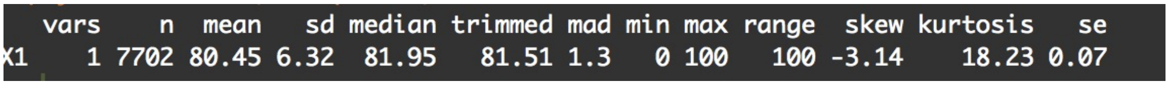
IDV: This is Insured Declared Value of the car. This is the valuation of the car according to the rules of the insurance company.

Age_group: We originally had various conflicting age groups which were re-levelled . And finally we had age groups: 18-24, 25-34, 35-44, 45-54, 55-64,65+

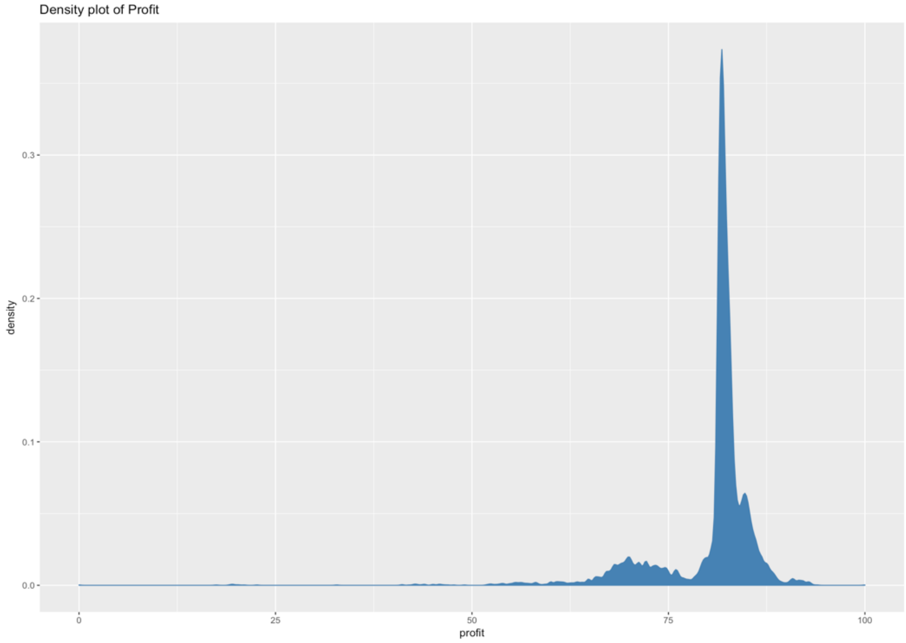
Gender: We have the genders Male and Female.

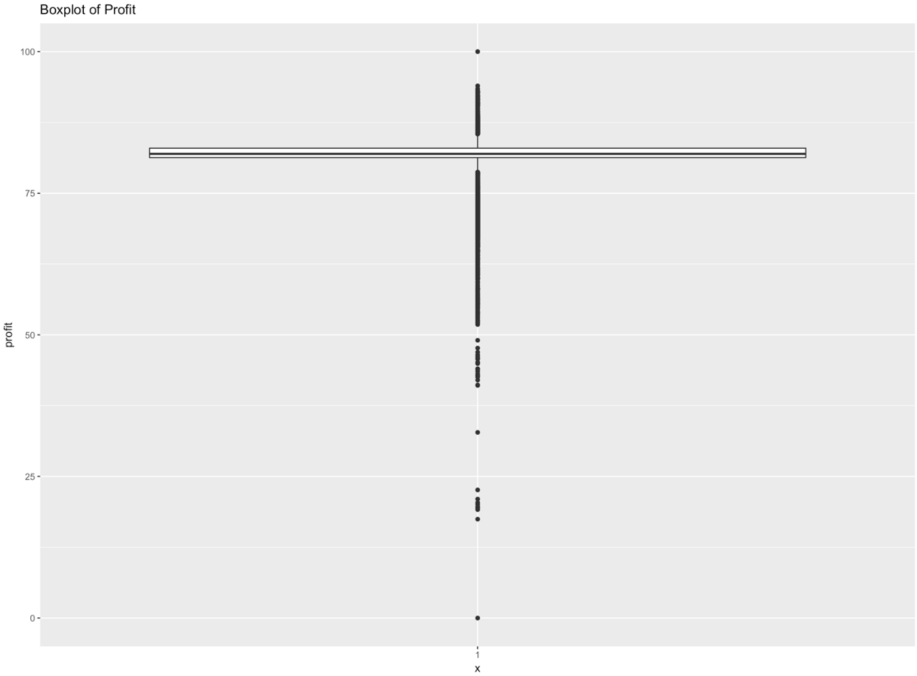
Univariate Analysis

Profit:



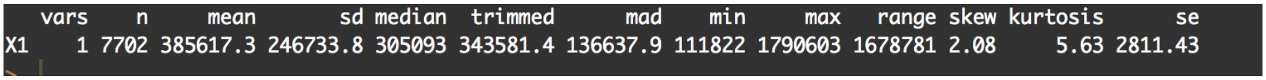
```
   vars    n  mean   sd median trimmed mad min max range  skew kurtosis   se
X1    1 7702 80.45 6.32  81.95   81.51 1.3   0 100   100 -3.14    18.23 0.07
```



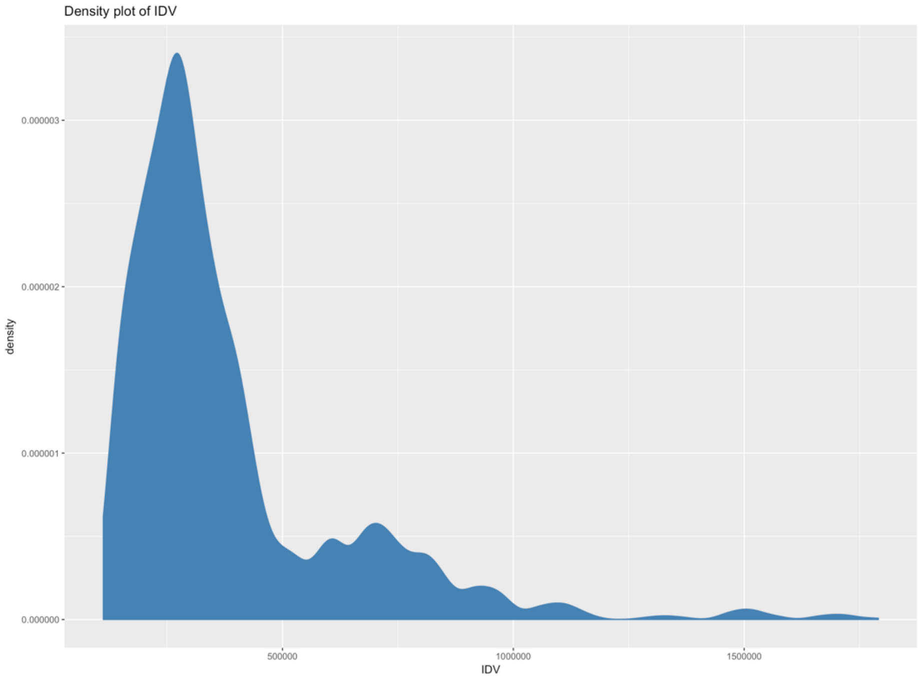


Analysis: The profit data does not show a normal distribution and it is negatively skewed towards the left.

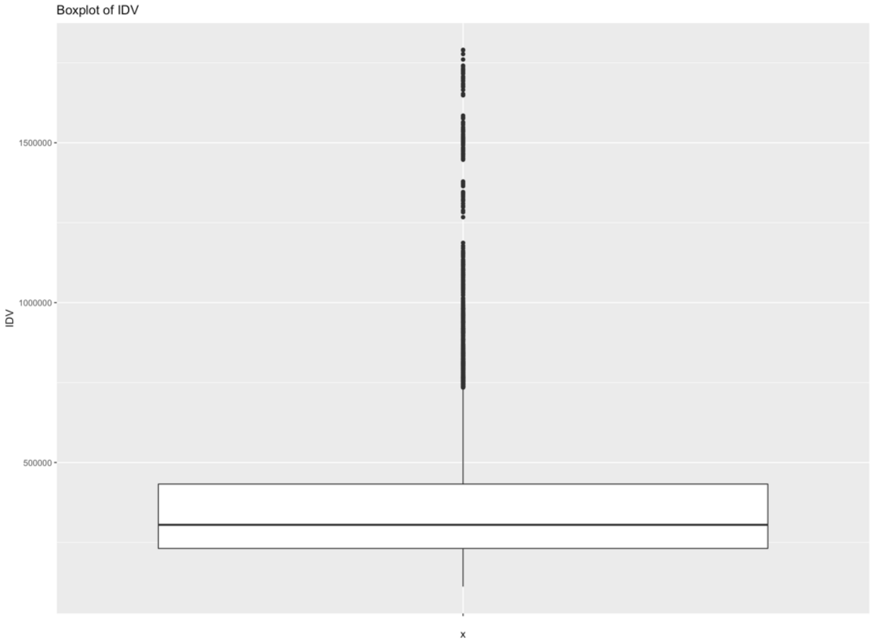
IDV:



```
   vars    n     mean       sd median  trimmed      mad    min     max   range skew kurtosis      se
X1    1 7702 385617.3 246733.8 305093 343581.4 136637.9 111822 1790603 1678781 2.08     5.63 2811.43
```



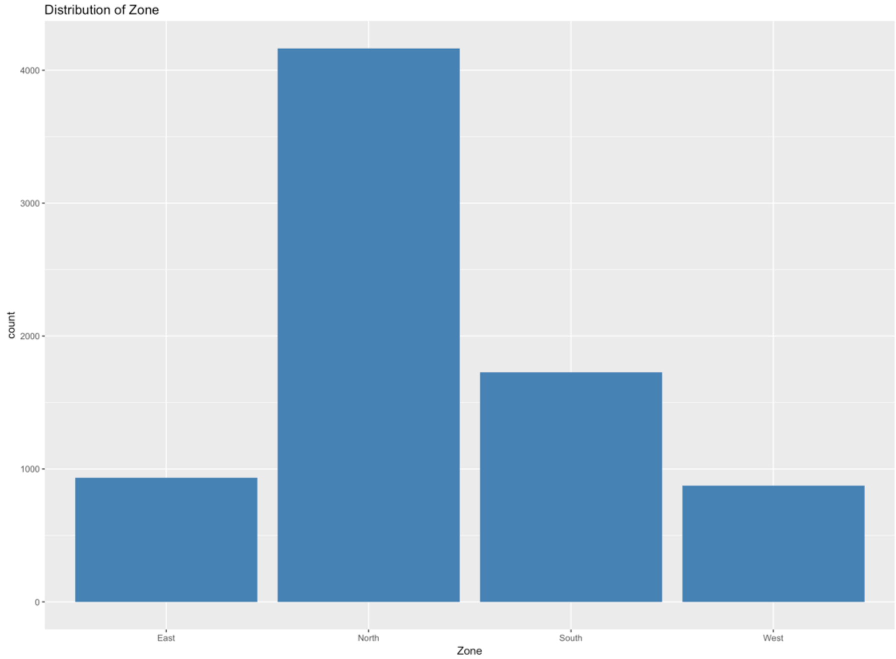
Density plot of IDV



Boxplot of IDV

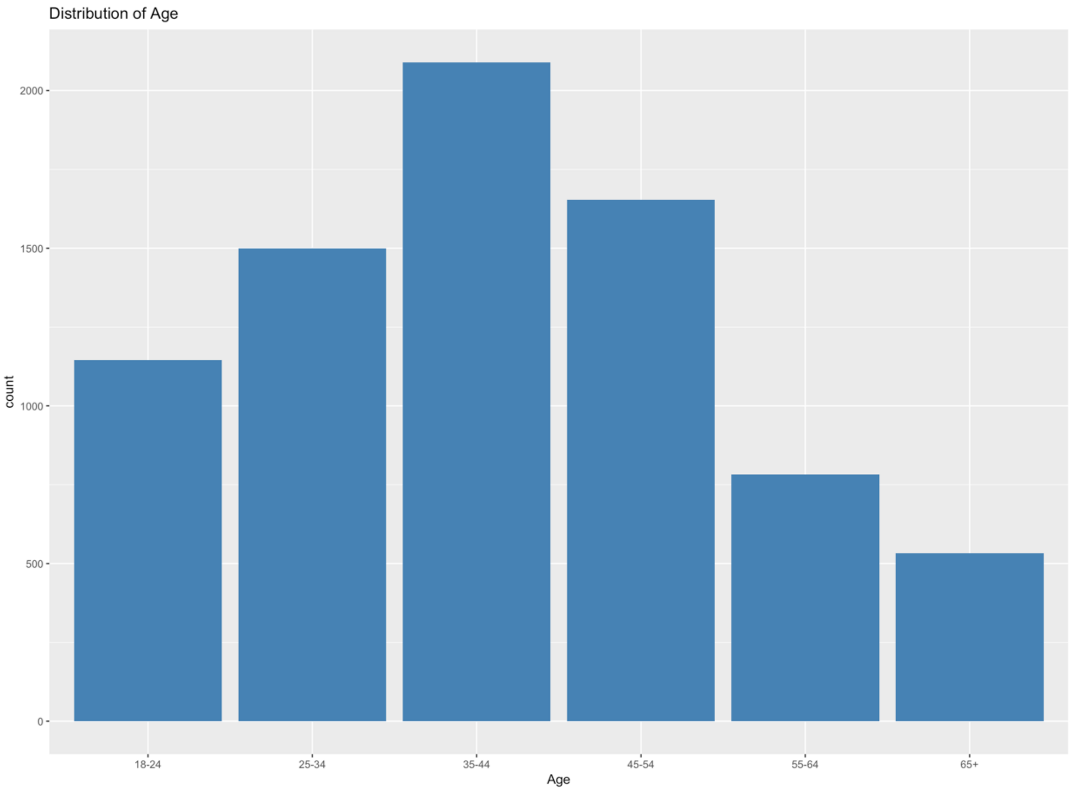
Zone:

East	935	12.361185
North	4163	55.037017
South	1728	22.845056
West	876	11.37367



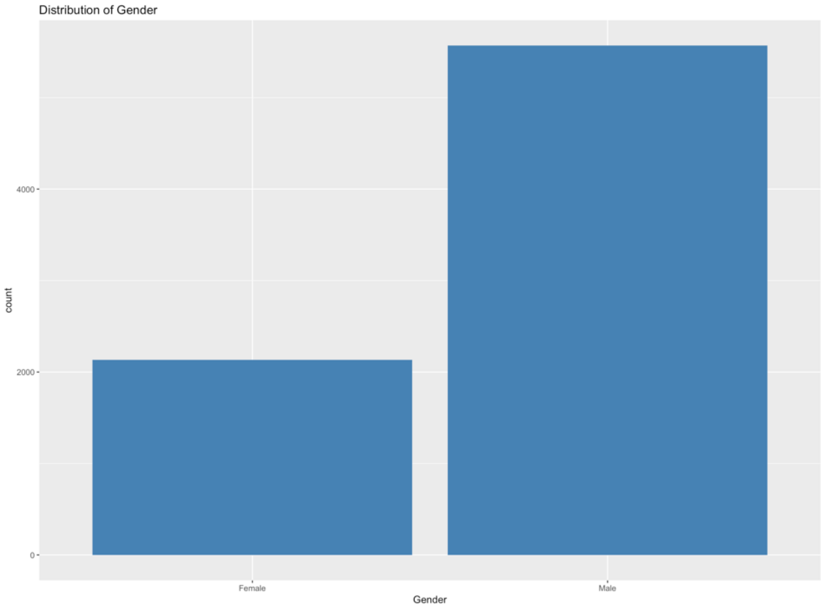
Age Group:

18-24	1145	14.866269
25-34	1500	19.475461
35-44	2089	27.122825
45-54	1654	21.474942
55-64	782	10.153207
65+	532	6.907297



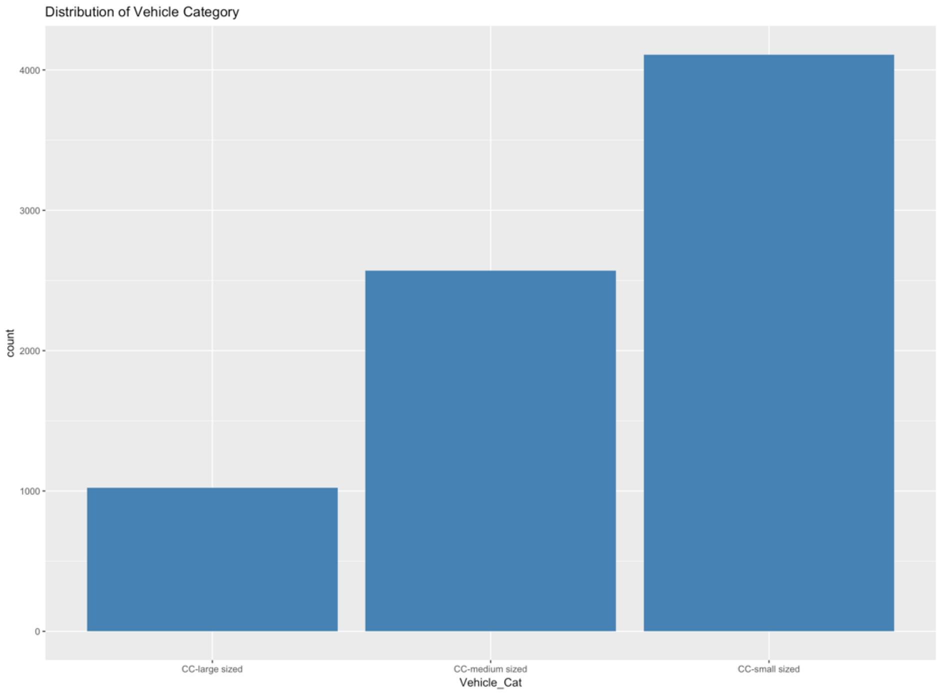
Gender:

Female	2134	27.70709
Male	5568	72.29291



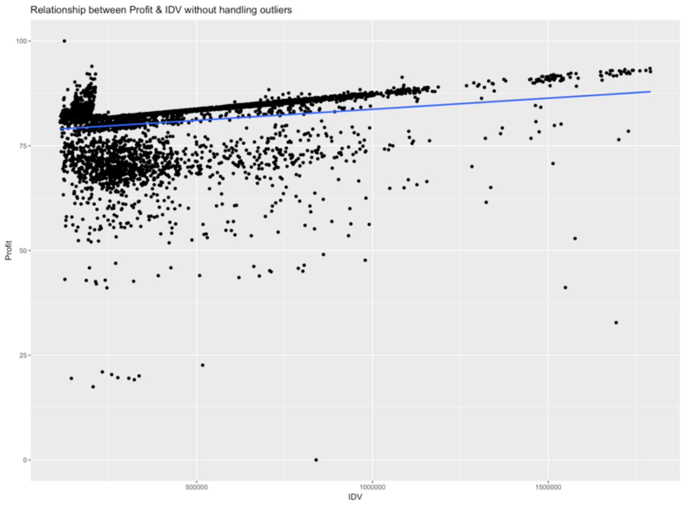
Vehicle Category:

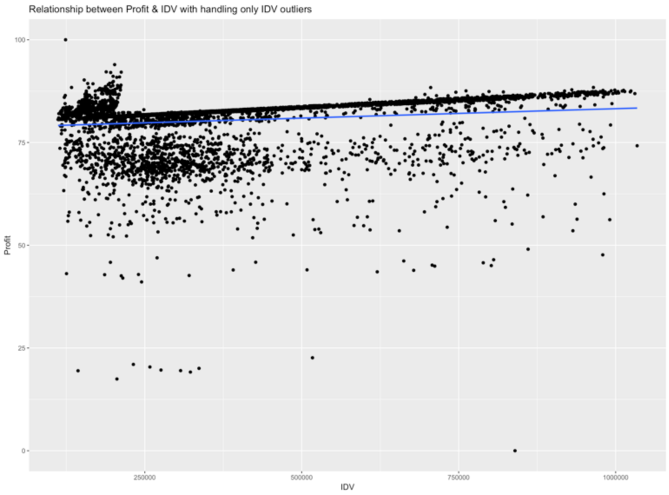
CC-large sized	1023	13.28226
CC-medium sized	2571	33.38094
CC-small sized	4108	53.33680

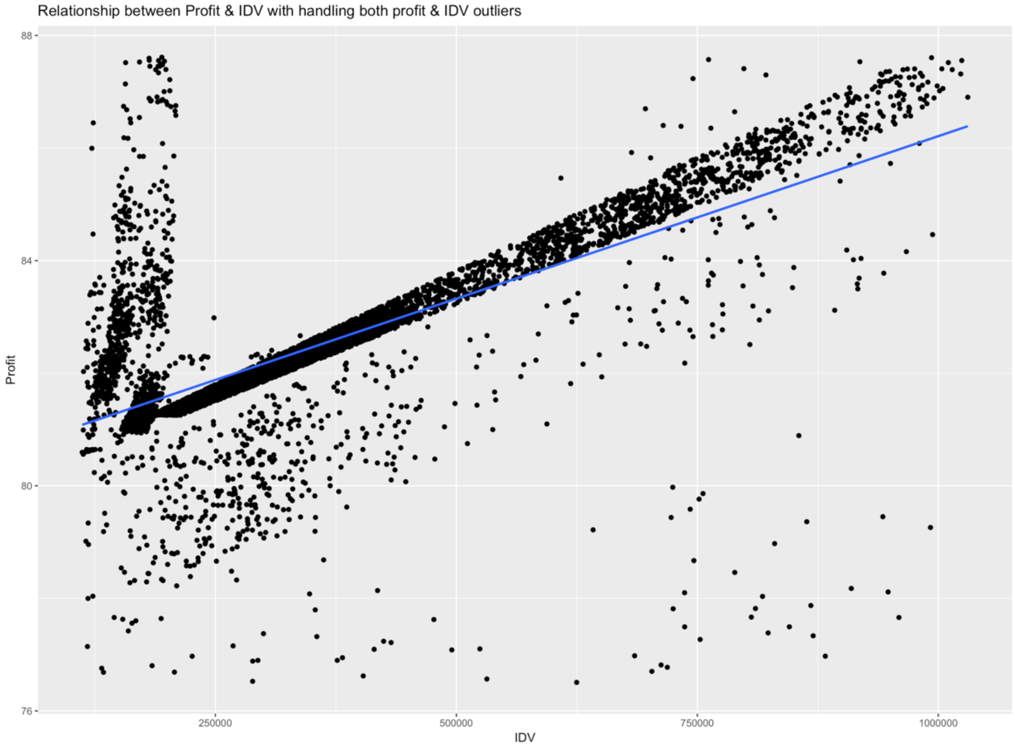


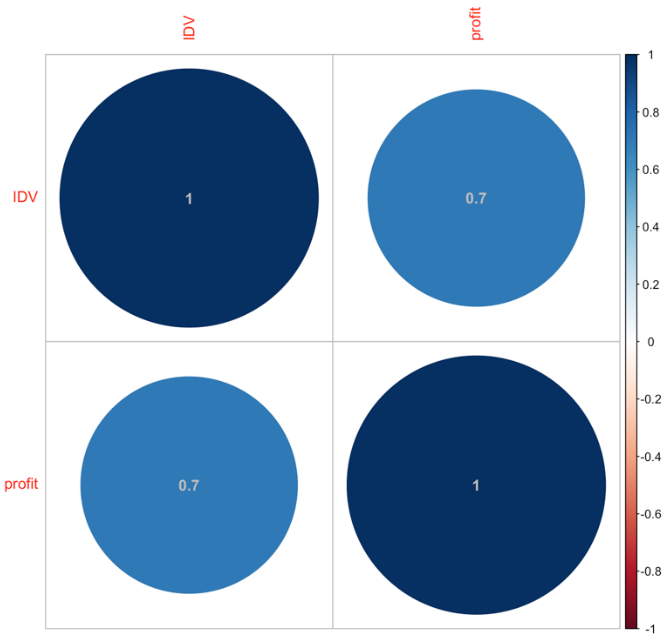
Bivariate Analysis

Profit with IDV:



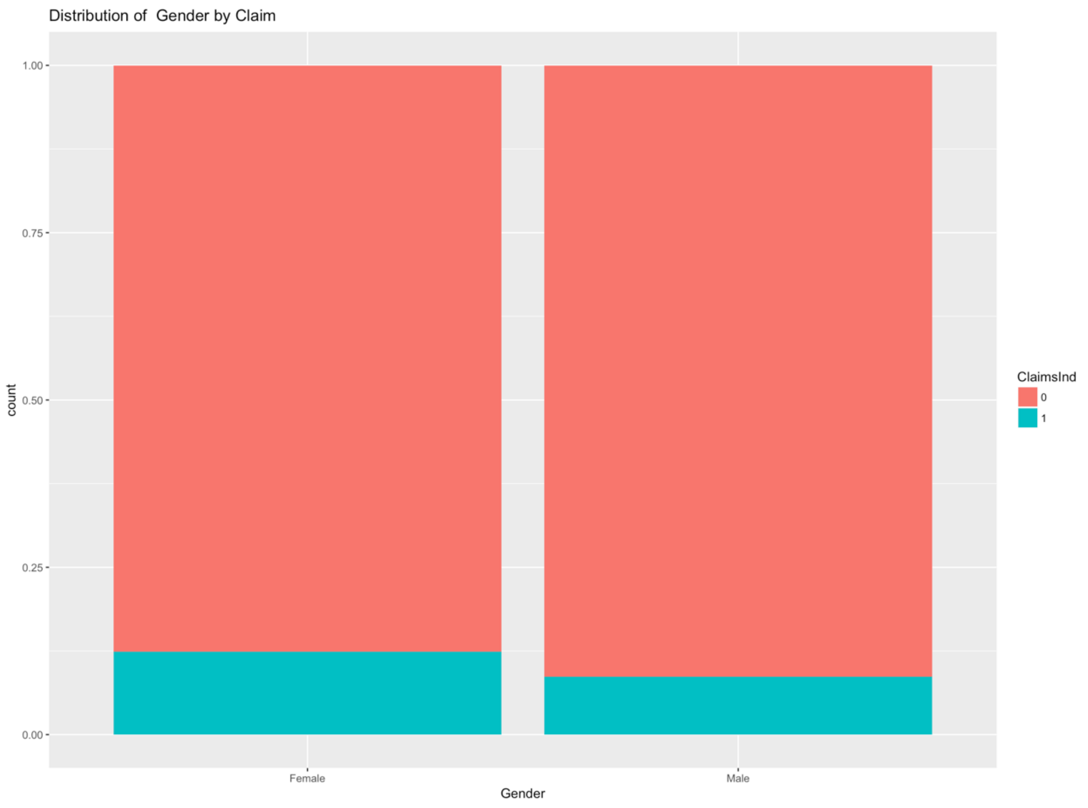




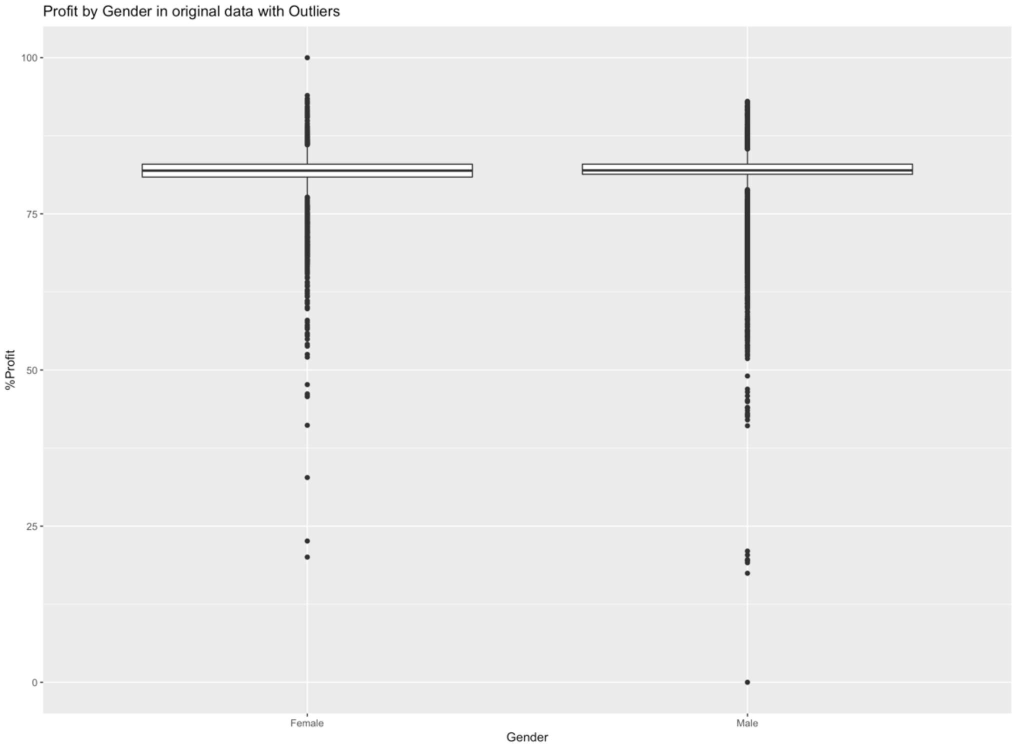


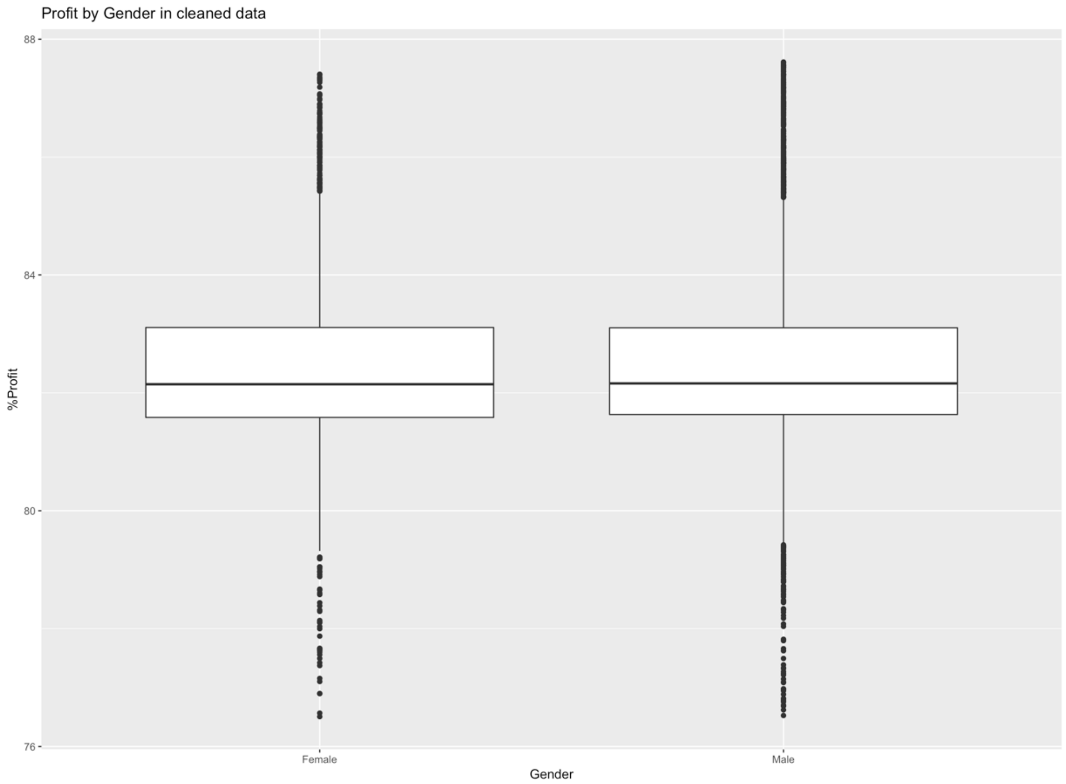
Analysis: Profit has a strong positive correlation with IDV.After performing the cor test we got the p-value less than 0.01, which means that with 99% confidence interval we can reject the null hypothesis that there is no relation between profit and IDV.

Claim and Gender:



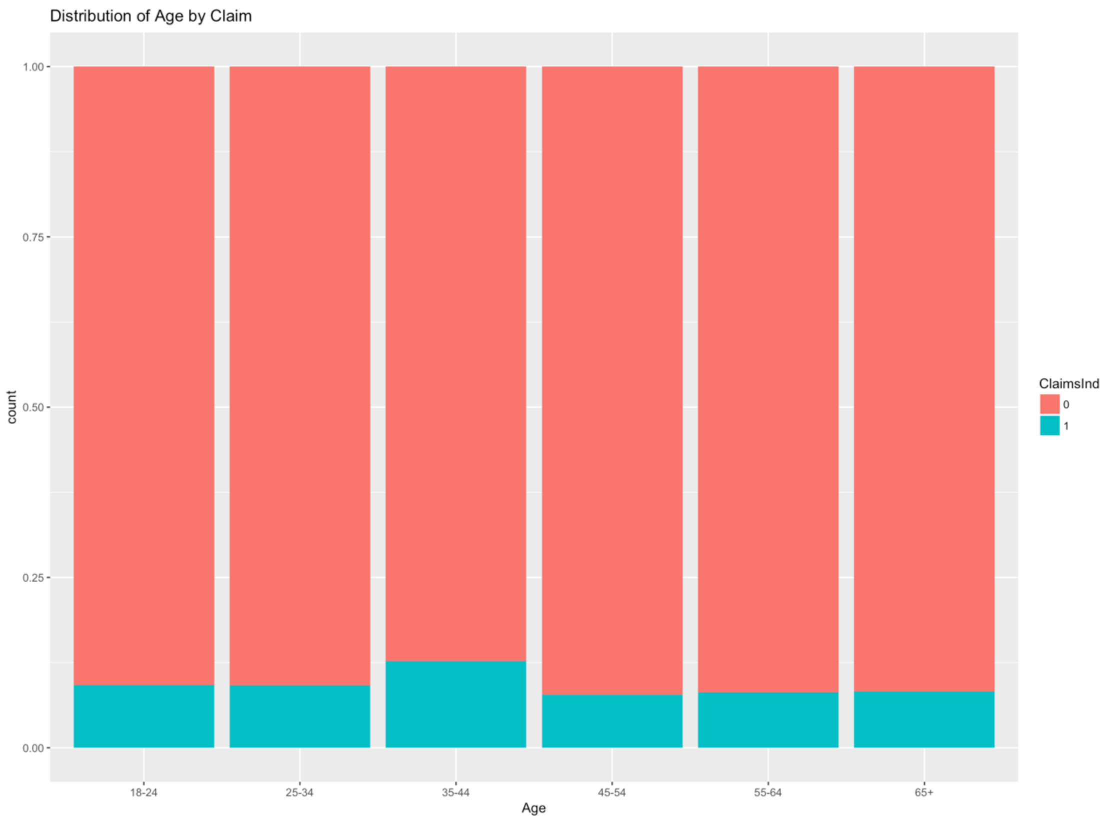
Profit with Gender:



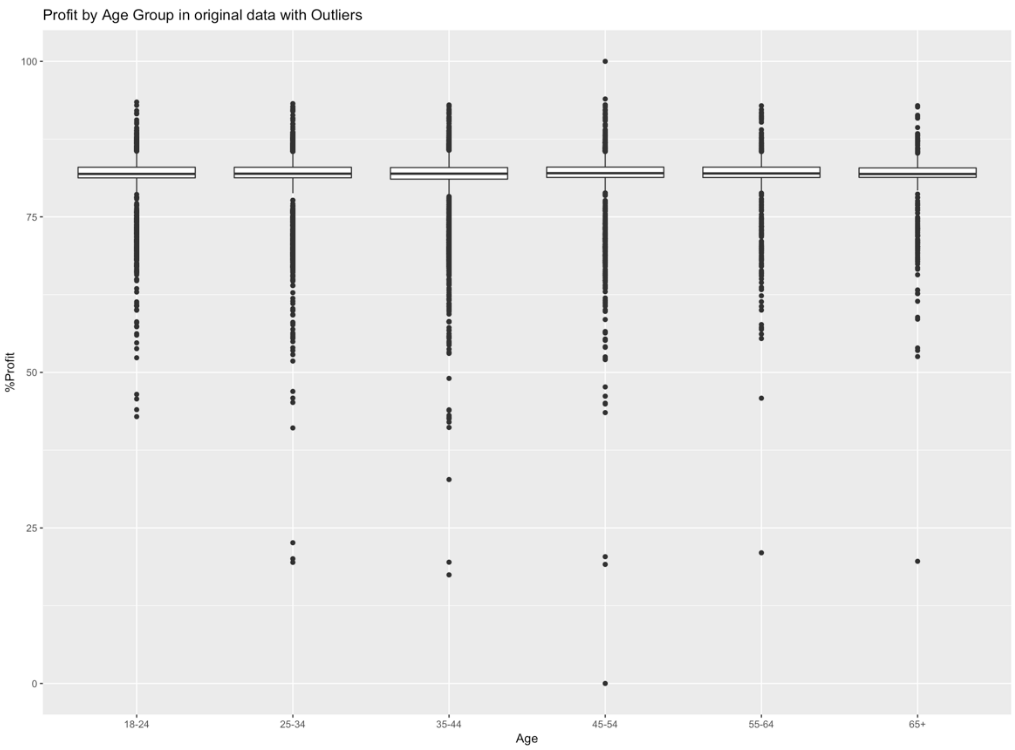


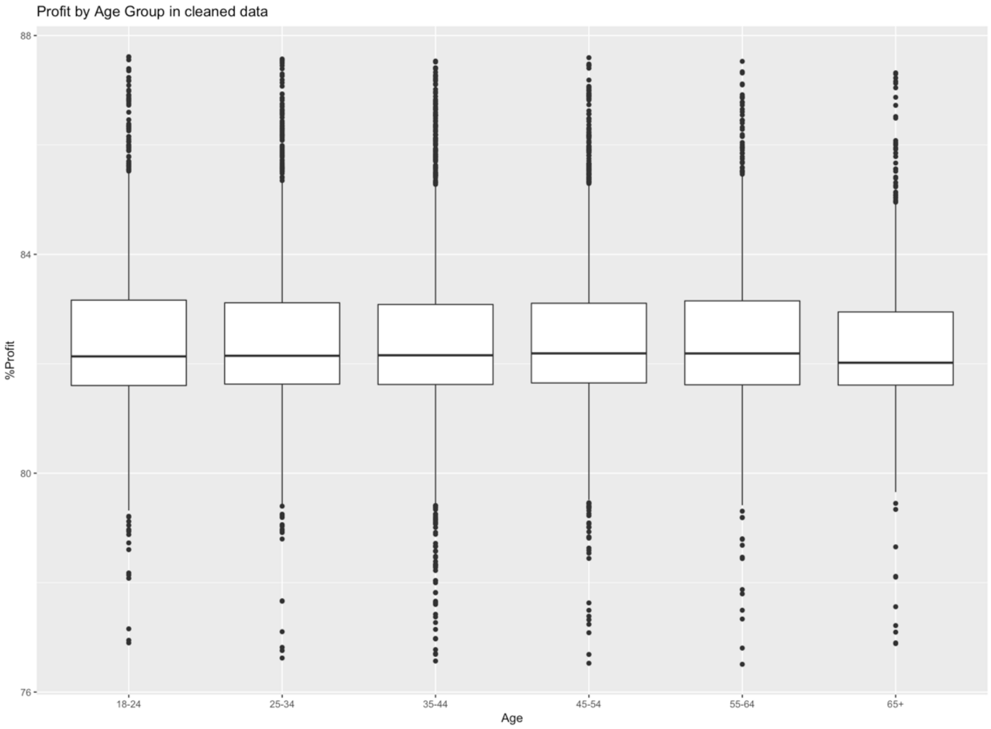
Analysis: From the box plot we see that there's not much difference in profit by gender as the mean for both the data are equal. After performing the Anova testing we got the p-value is greater than 0.05, which means that with 95% confidence interval we cannot reject the null hypothesis that there is no relation between profit and IDV. So, our Profit is not affected by gender.

Claim and Age Group:



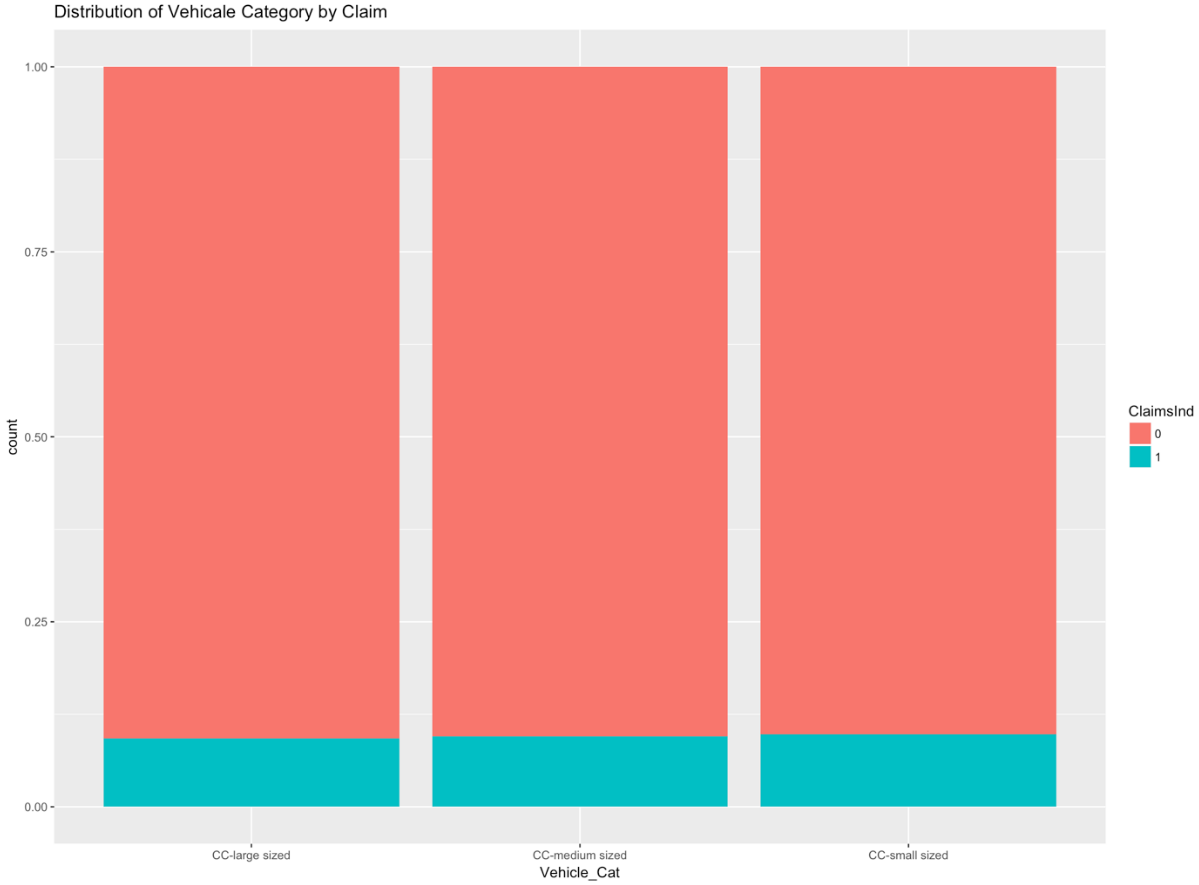
Profit with Age Group:



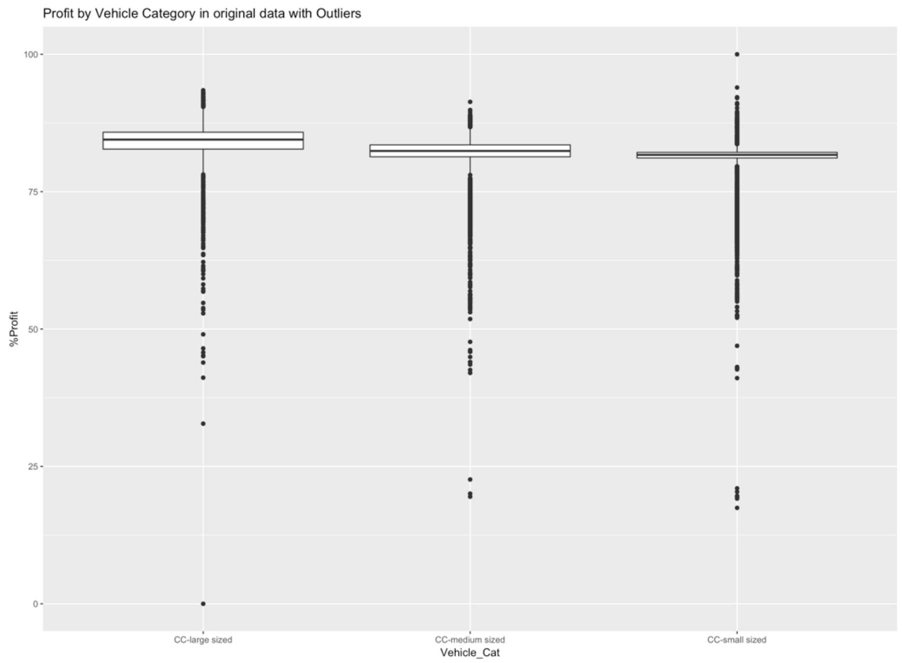


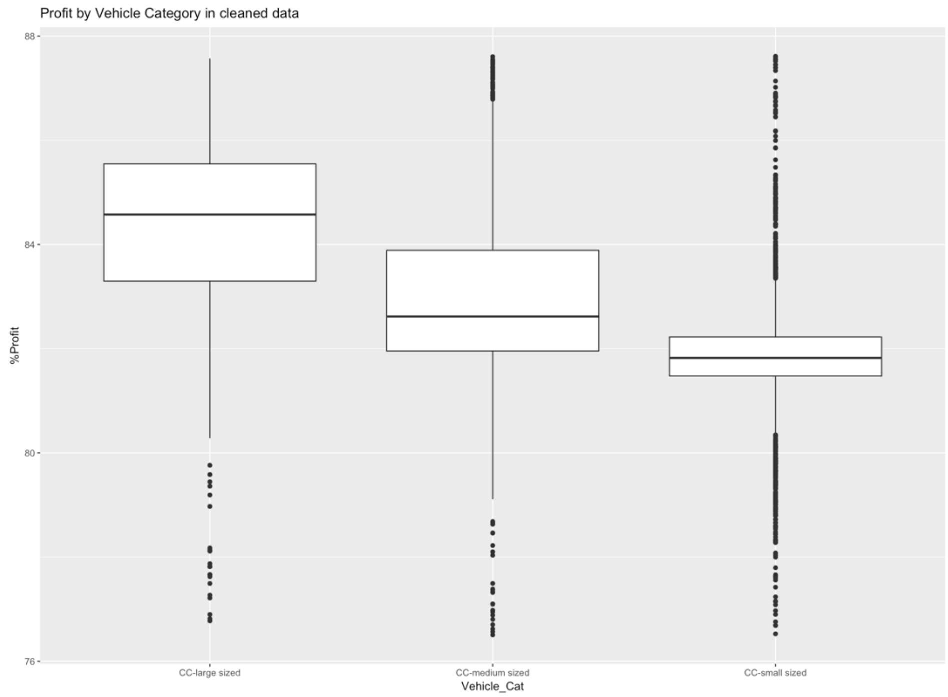
Analysis: From the box plot we see that there's not much difference in profit by zone, the mean for both the data are equal. After performing the Anova testing we got the p-value is greater than 0.05, which means that with 95% confidence interval we cannot reject the null hypothesis that there is no relation between profit and Age Group . So, our Profit is not affected by Age Group.

Claim and Vehicle Category:



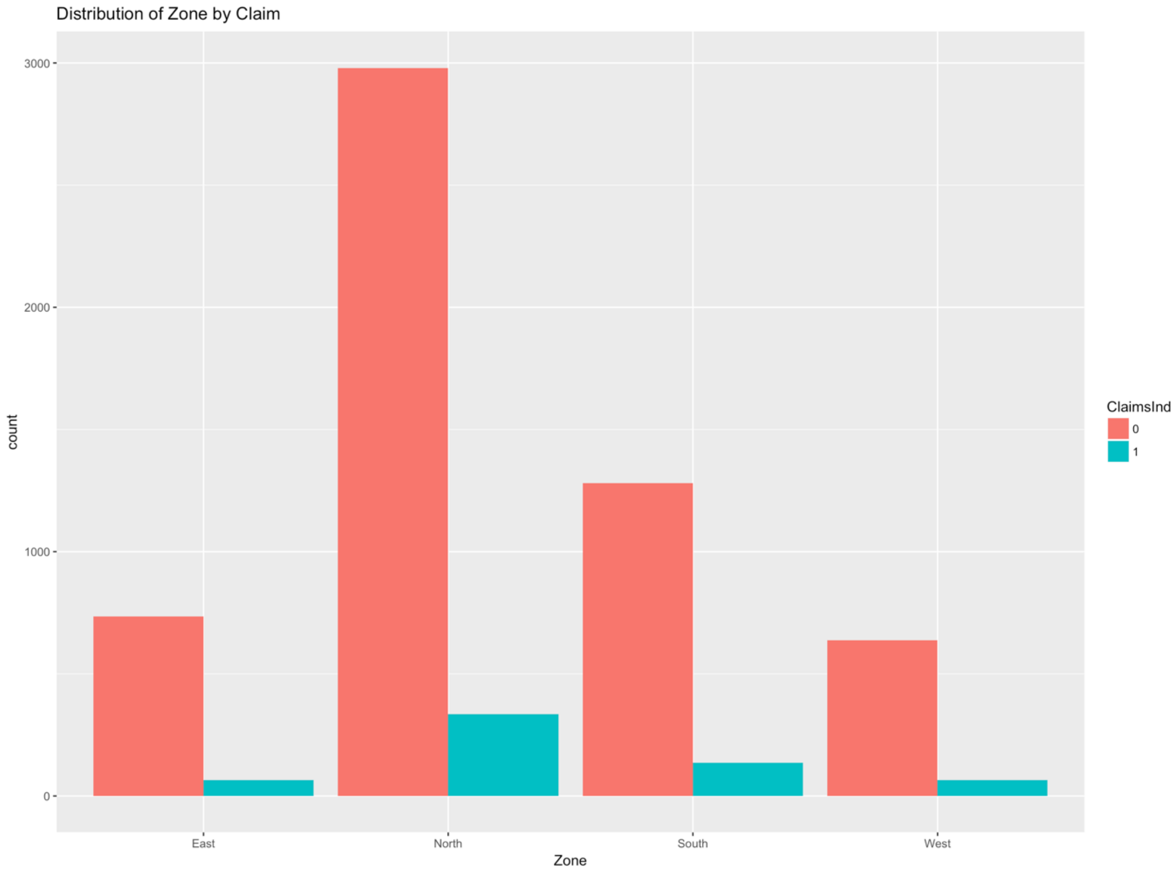
Profit with Vehicle Category:



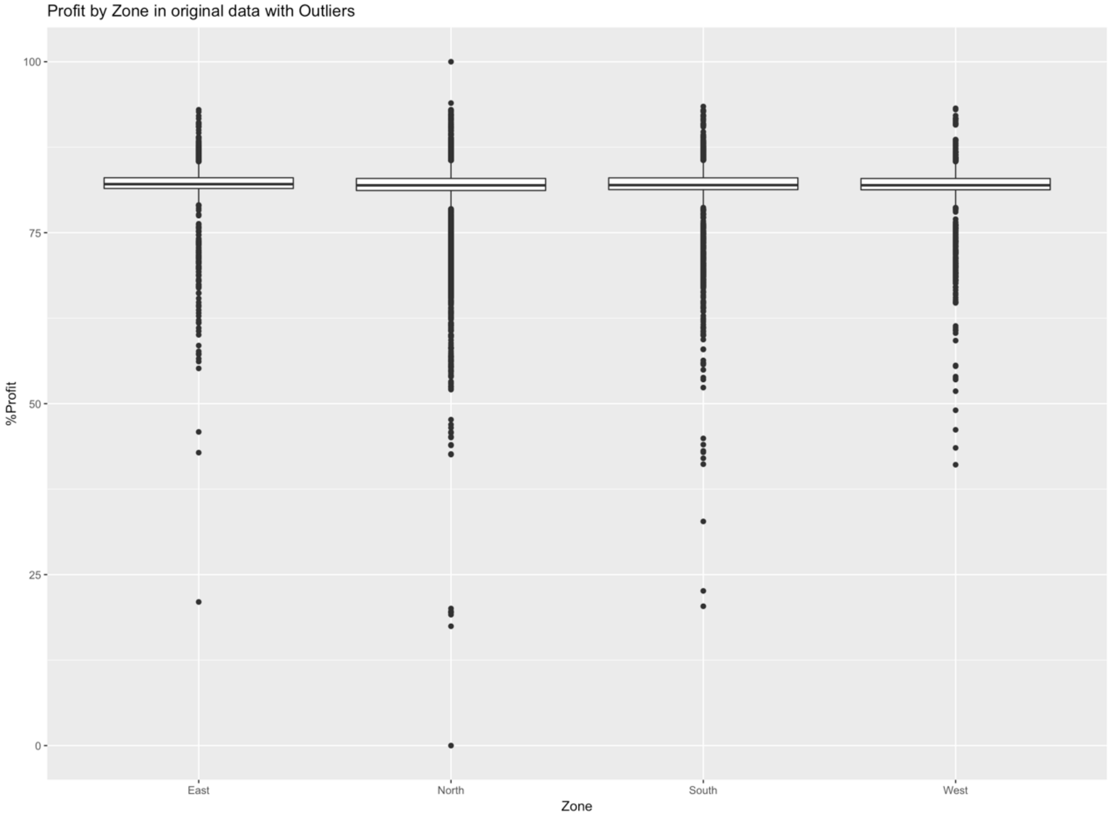


Analysis: From the box plot we see that there's not much difference in profit by zone, the mean for both the data are equal. After performing the Anova testing we got the p-value is greater than 0.05, which means that with 95% confidence interval we cannot reject the null hypothesis that there is no relation between profit and Age Group . So, our Profit is not affected by Age Group.

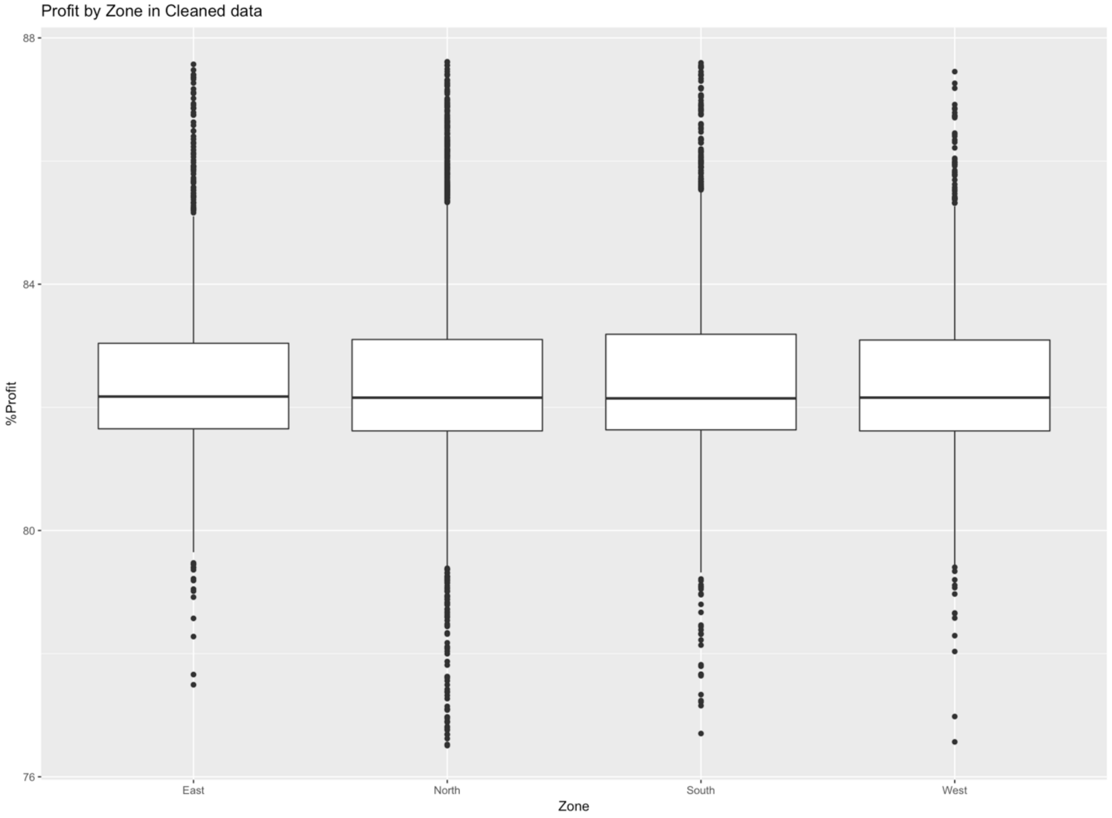
Claim and Zone:



Profit with Zone:



Profit by Zone in original data with Outliers



Profit by Zone in Cleaned data

Analysis: From the box plot we see that there's not much difference in profit by zone, the mean for both the data are equal. After performing the Anova testing we got the p-value is greater than 0.05, which means that with 95% confidence interval we cannot reject the null hypothesis that there is no relation between profit and Zone. So, our Profit is not affected by Zone.

Conclusion

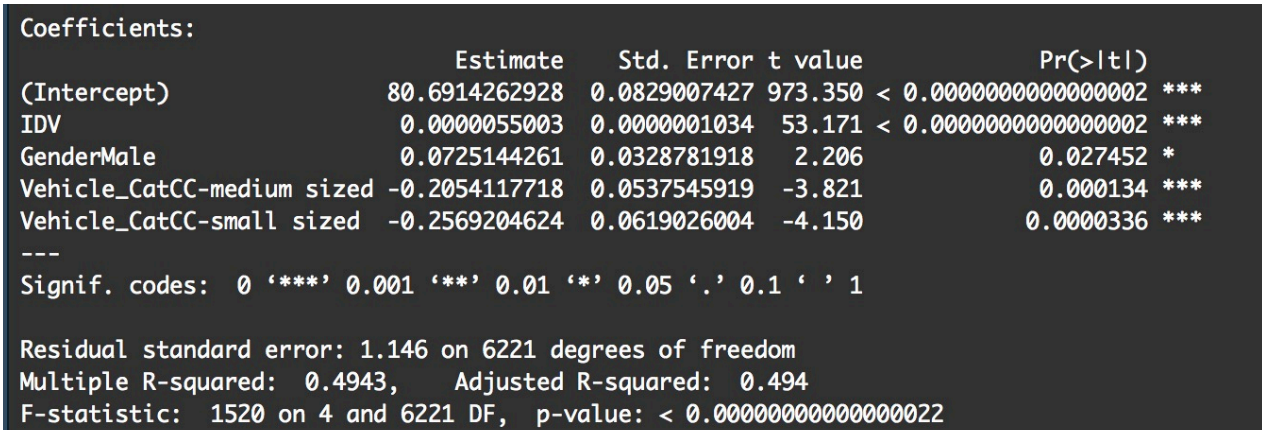
As per our analysis based on the available dataset,

Profit is **not dependent** on age & gender of the driver.

Profit is **not dependent** on any geographical zone.

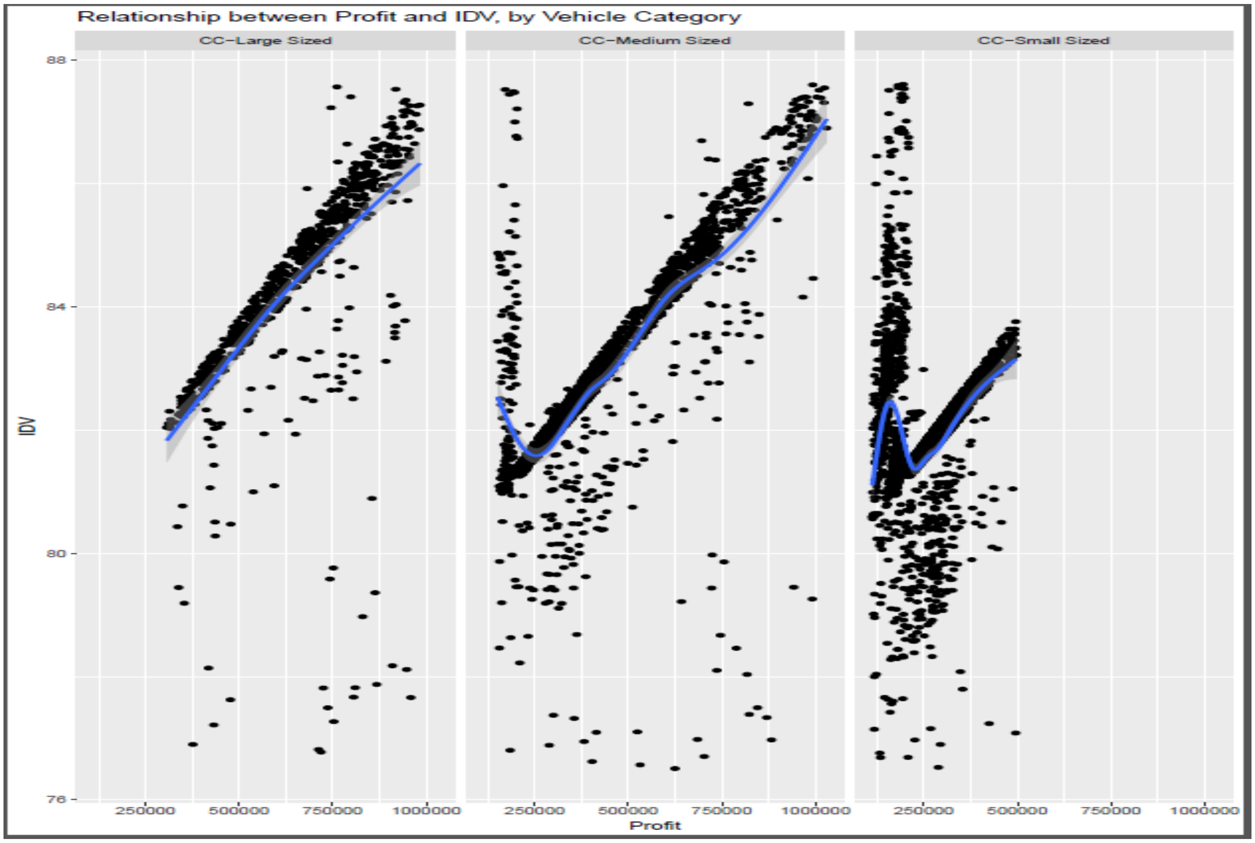
Profit of the company **increases** with the increase in cubic capacity of the vehicle.

Linear Regression Model



```
Coefficients:
                              Estimate    Std. Error t value             Pr(>|t|)
(Intercept)              80.6914262928  0.0829007427 973.350 < 0.0000000000000002 ***
IDV                       0.0000055003  0.0000001034  53.171 < 0.0000000000000002 ***
GenderMale                0.0725144261  0.0328781918   2.206             0.027452 *
Vehicle_CatCC-medium sized -0.2054117718  0.0537545919  -3.821             0.000134 ***
Vehicle_CatCC-small sized  -0.2569204624  0.0619026004  -4.150            0.0000336 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.146 on 6221 degrees of freedom
Multiple R-squared:  0.4943,	Adjusted R-squared:  0.494
F-statistic:  1520 on 4 and 6221 DF,  p-value: < 0.00000000000000022
```



Limitations

There are many other factors like experience of driver, driver's state of mind and severity of accident which may further affect the profitability and affect the models (r-square) value.