Auto Insurance Company Data Analysis

Team Members

Srija Gupta
Mohammed Rehan

Data

- 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

Business/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

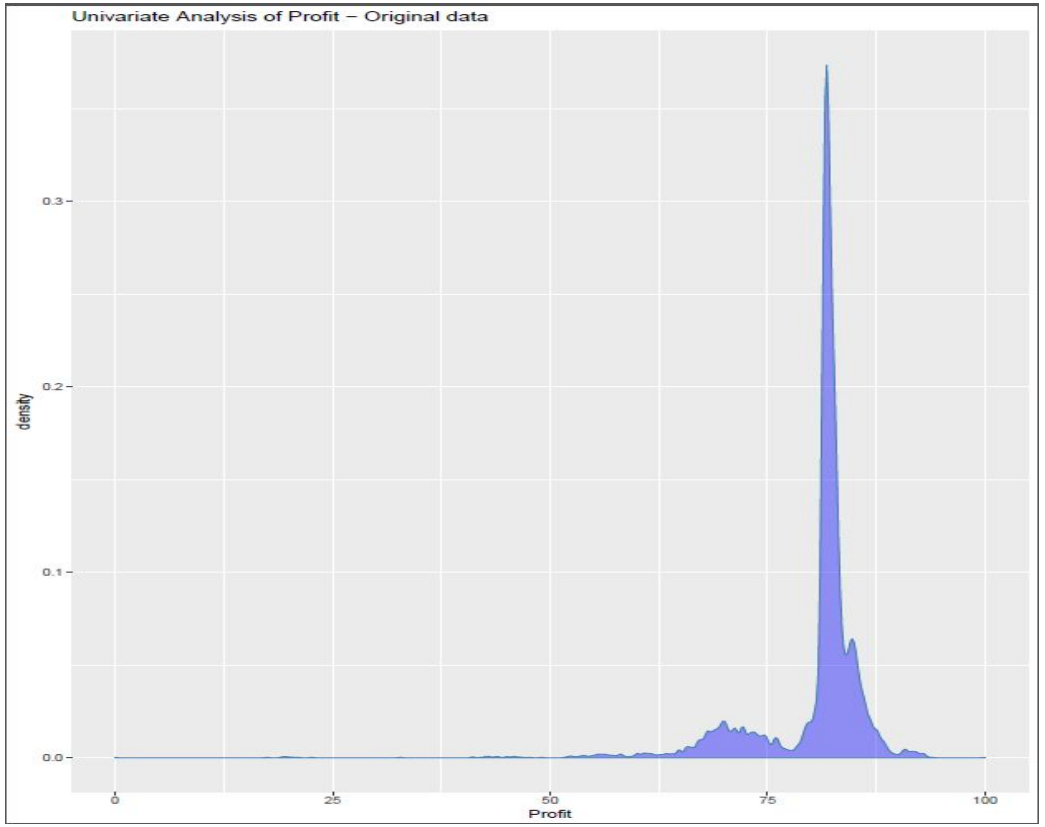
- The profitability of Indian Money is higher for female drivers in North zone.
- The profitability of Indian Money is higher for drivers having age more than 40 years.
- The profitability of Indian Money is higher for vehicles with cubic capacity more than 2200.

Variables Used For Analysis

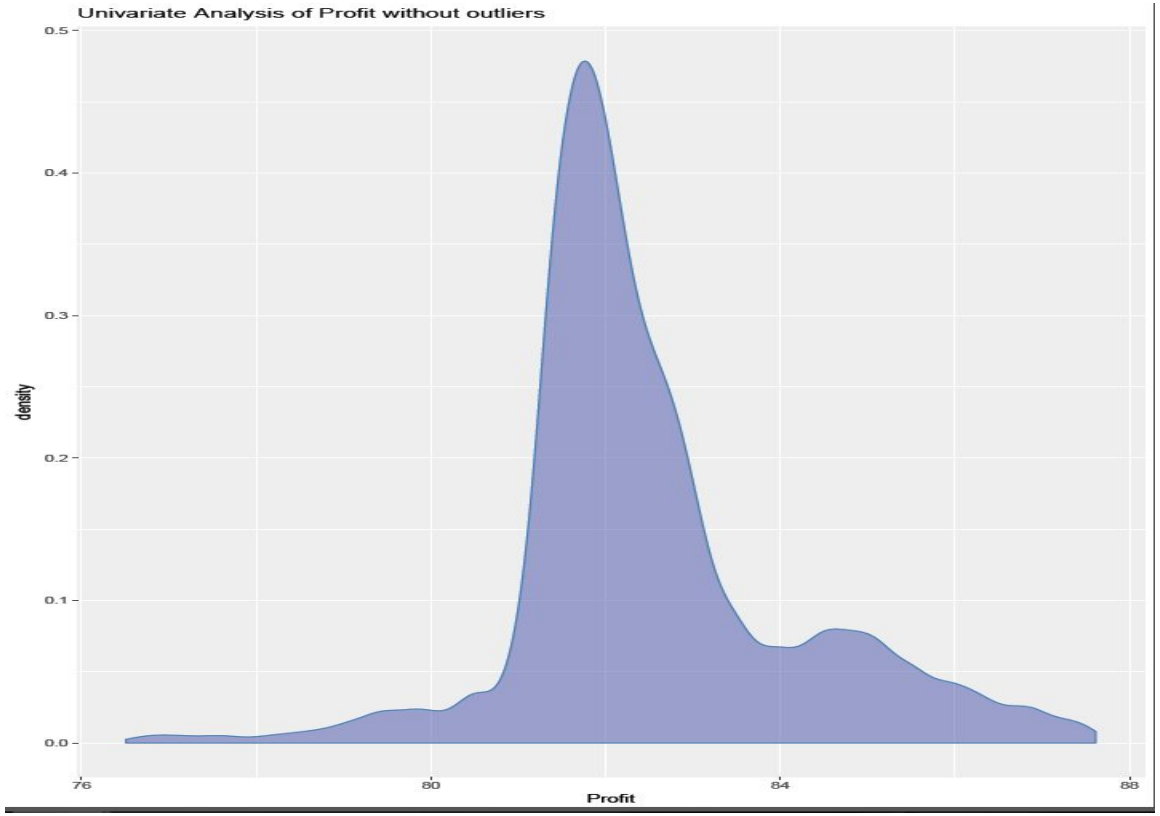
Name	Data Type	Variable	Description
Year	Factor	CV	7 Years of Data
IDV	Numeric	IV	Insured Declared value of Car
Gender	Factor	IV	Male or Female
Zone	Factor	IV	Divided Regions of the Country
Age Group	Factor	IV	Age groups of Applicants
ClaimsInd	Factor	IV	Claims Take(0-not taken,1-Taken)
Vehicle Category	Factor	IV	Clubbed to Cubic Capacity size
Revenue	Numeric	DV	Derived from Premium minus Claim
Profit	**Numeric**	**DV**	**Profitability - Derived column(Revenue %)**

Profit Univariate Analysis

Original Data- With Outliers



Without Outliers

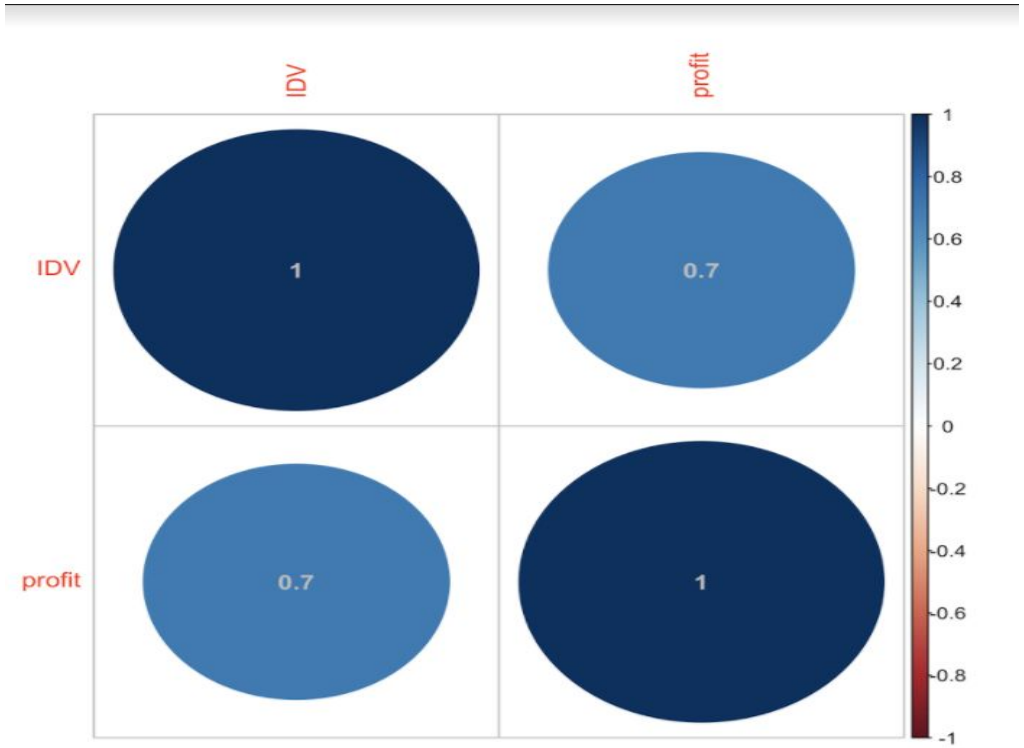


Relevant Analysis

- Profit with IDV
- Profit with Gender
- Profit with Zone
- Profit with Age Group
- Profit with Vehicle Category
- Linear Model

Profit with IDV

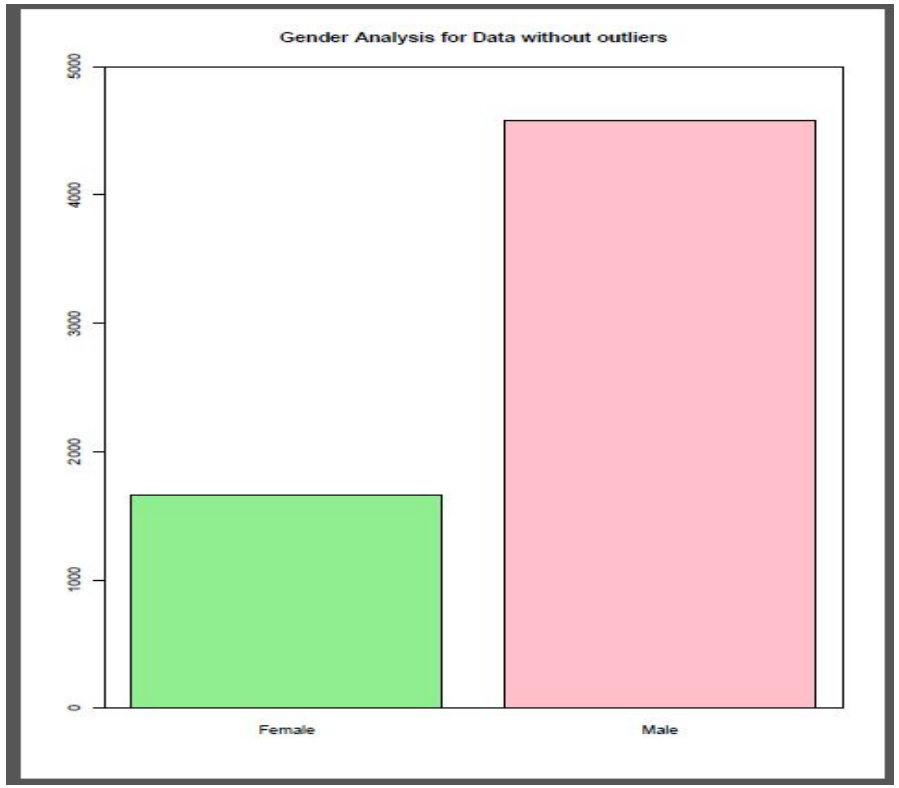
Profit has a strong correlation with IDV

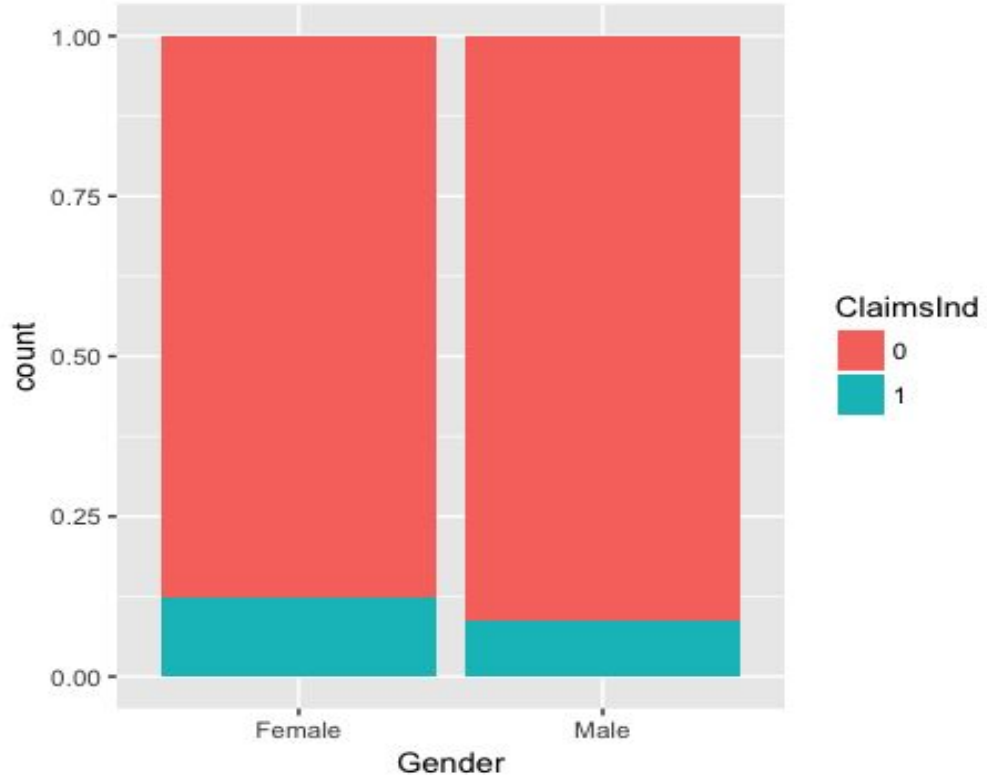



```
> cor.test(ins_rm_ex_IDV_pro_r$IDV,ins_rm_ex_IDV_pro_r$profit)

        Pearson's product-moment correlation

data:  ins_rm_ex_IDV_pro_r$IDV and ins_rm_ex_IDV_pro_r$profit
t = 82.823, df = 6223, p-value < 0.00000000000000022
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7120742 0.7357153
sample estimates:
      cor
0.7241074
```

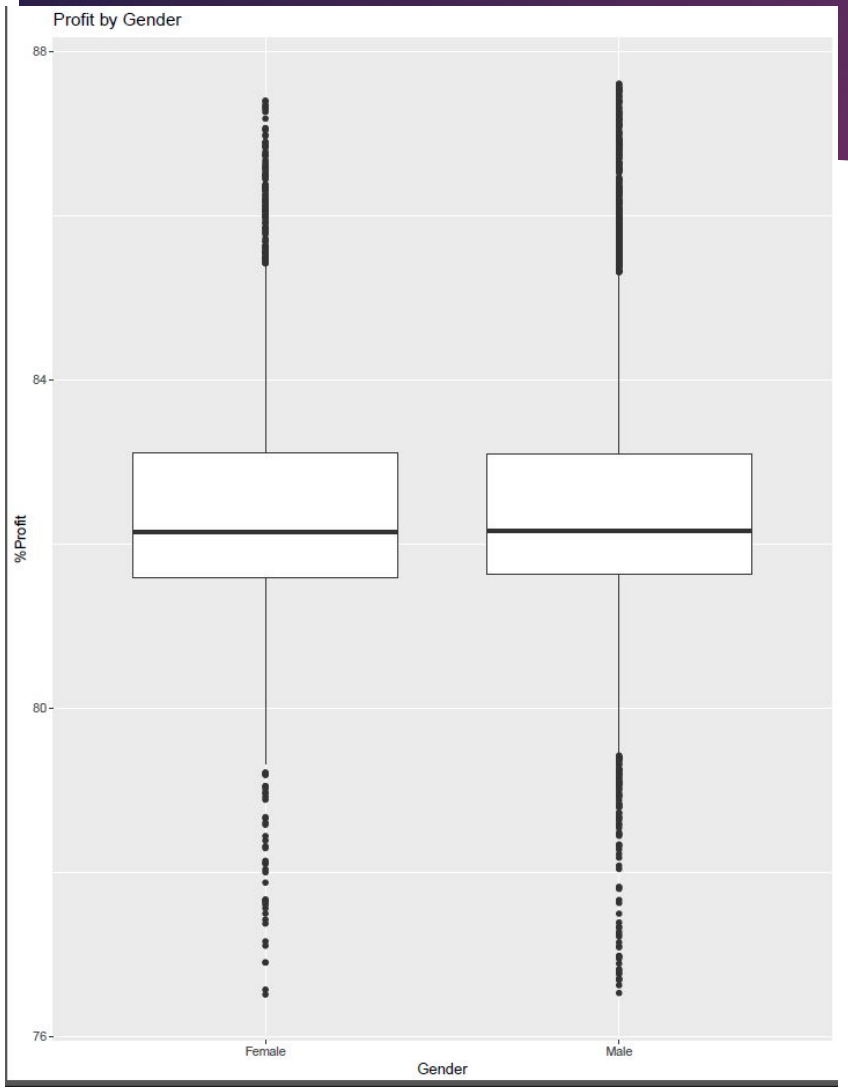
Analysis - Gender

Our Hypothesis - Profit more for Female as to Male





Profit with Gender

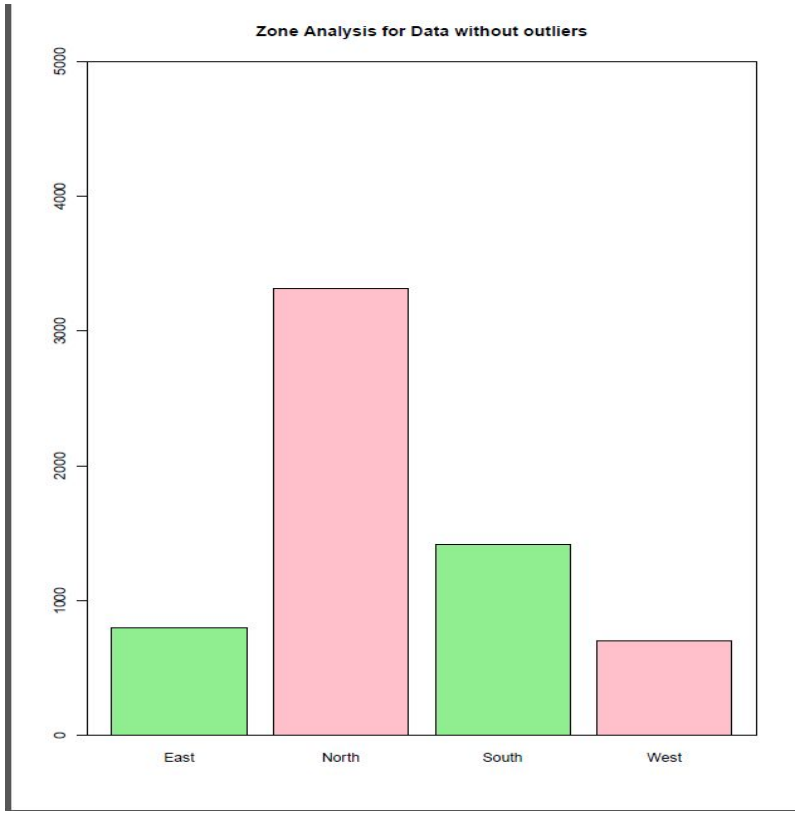


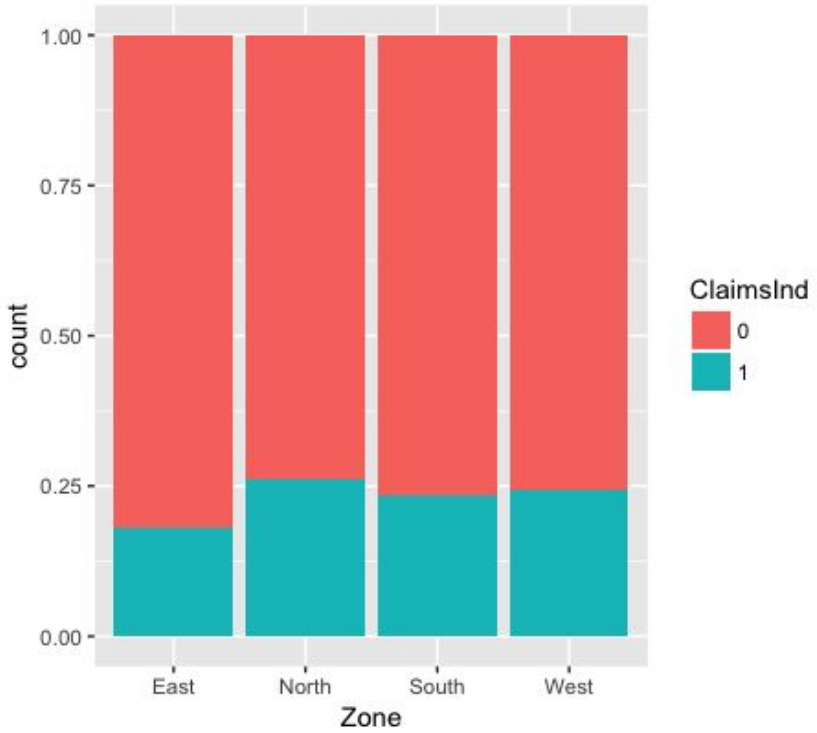

```
> gender_No_outlier_anova <- aov(Profit~Gender,data=ins_rm_ex_IDV_pro)
> summary(gender_No_outlier_anova)
              Df Sum Sq Mean Sq F value Pr(>F)
Gender         1      5   5.244   2.022  0.155
Residuals   6233  16165   2.593
> T_gender_No_outlier_anova <- TukeyHSD(gender_No_outlier_anova)
> T_gender_No_outlier_anova
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Profit ~ Gender, data = ins_rm_ex_IDV_pro)

$Gender
                  diff         lwr       upr     p adj
Male-Female 0.06563905 -0.02485273 0.1561308 0.1550896
```

Conclusion - We fail to reject Null Hypothesis and thus there is no relationship of Profit by Gender

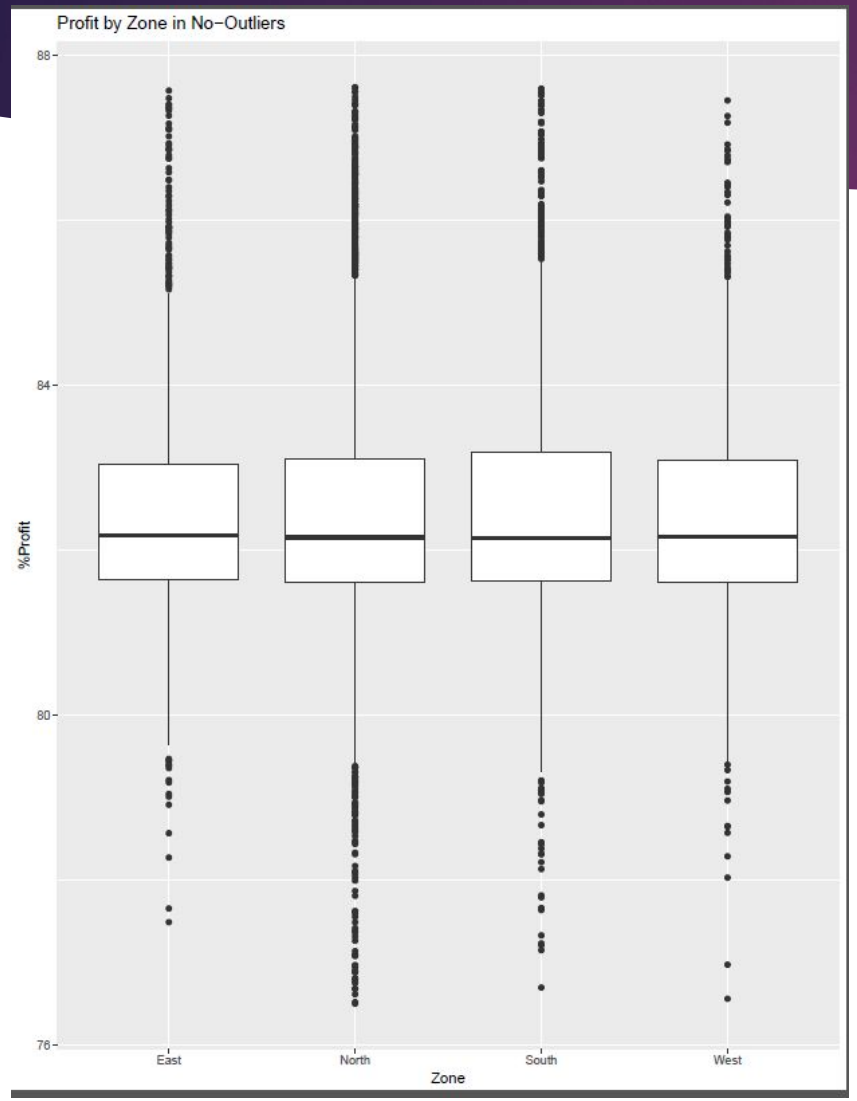
Analysis - Zone

Our Hypothesis - Profit more for North Zone as to other zones





Profit with Zone

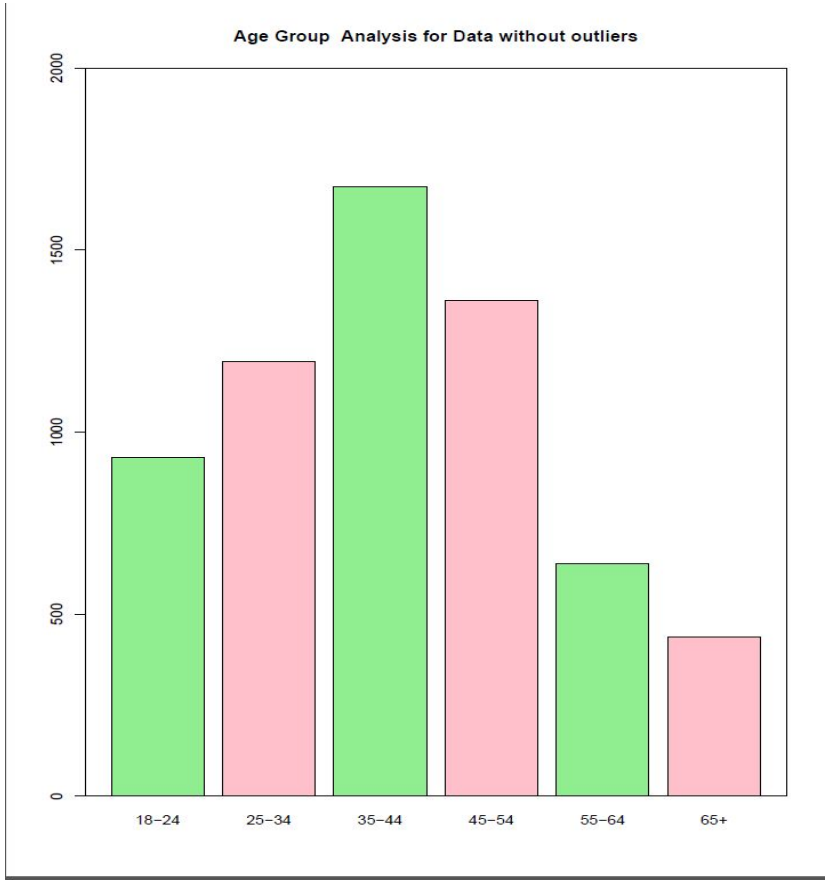


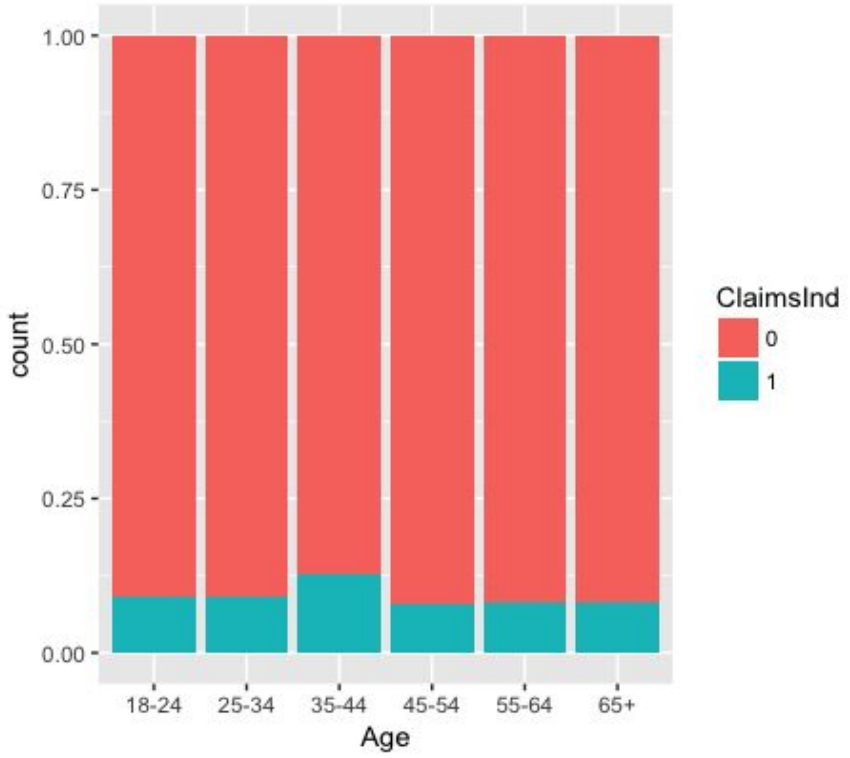

```
> summary(zone_No_outlier_anova)
             Df Sum Sq Mean Sq F value Pr(>F)
Zone          3      6   1.872   0.722  0.539
Residuals   6231  16165   2.594
> T_zone_No_outlier_anova <- TukeyHSD(zone_No_outlier_anova)
> T_zone_No_outlier_anova #no relation
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = profit ~ Zone, data = ins_rm_ex_IDV_pro)

$Zone
                    diff         lwr        upr     p adj
North-East  -0.0676722936 -0.23071534 0.09537075 0.7099404
South-East  -0.0083692591 -0.19138525 0.17464673 0.9994201
West-East   -0.0686146540 -0.28258235 0.14535304 0.8431045
South-North  0.0593030345 -0.07203768 0.19064375 0.6519885
West-North  -0.0009423604 -0.17280773 0.17092301 0.9999990
West-South  -0.0602453949 -0.25116298 0.13067219 0.8493551

>
```

Conclusion - We fail to reject Null Hypothesis and thus there is no relationship of Profit by Zone

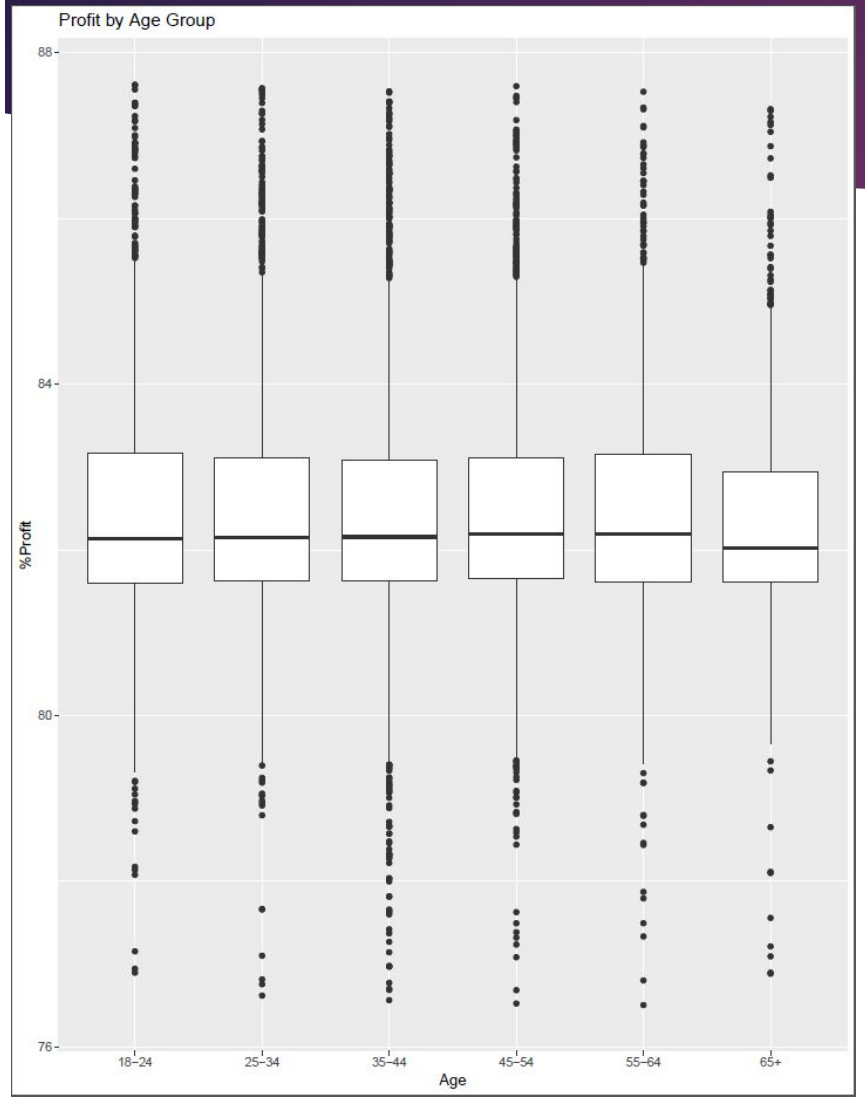
Analysis - Age Group

Our Hypothesis - Profit more for Age Group over 40 years than rest





Profit with Age Group

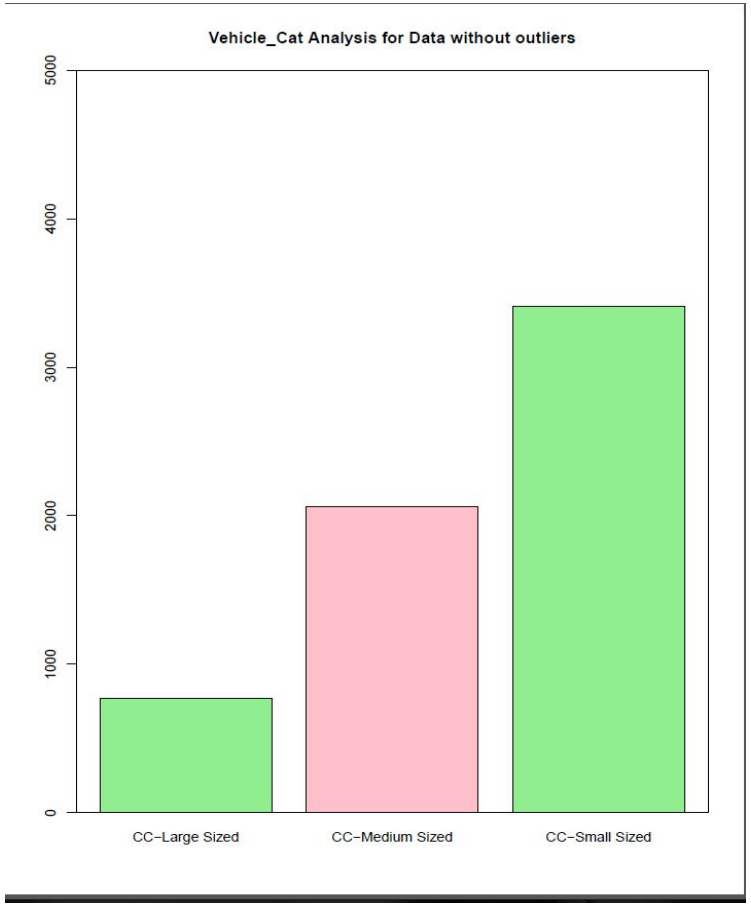


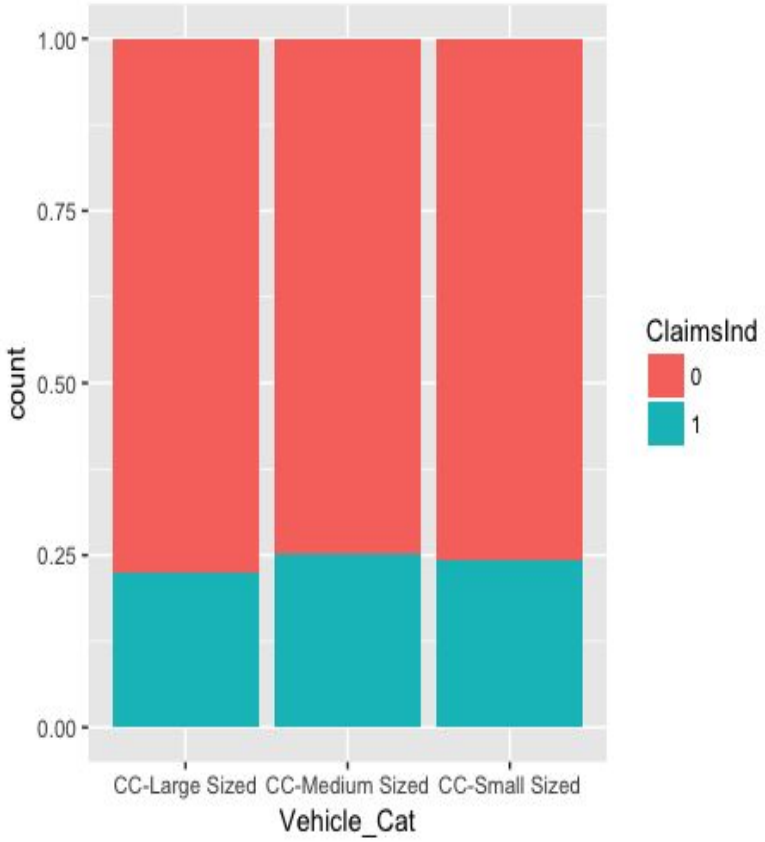

```
> summary(Age_No_outlier_anova)
             Df Sum Sq Mean Sq F value Pr(>F)
Age           5     17   3.427   1.322  0.252
Residuals  6229  16153   2.593
> T_Age_No_outlier_anova <- TukeyHSD(Gender_No_outlier_anova)
> T_Age_No_outlier_anova #no relation
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = profit ~ Gender, data = ins_rm_ex_IDV_pro)

$Gender
                  diff        lwr       upr     p adj
Male-Female 0.06563888 -0.02485266 0.1561304 0.1550896
```

Conclusion - We fail to reject Null Hypothesis and thus there is no relationship of Profit by Age Group

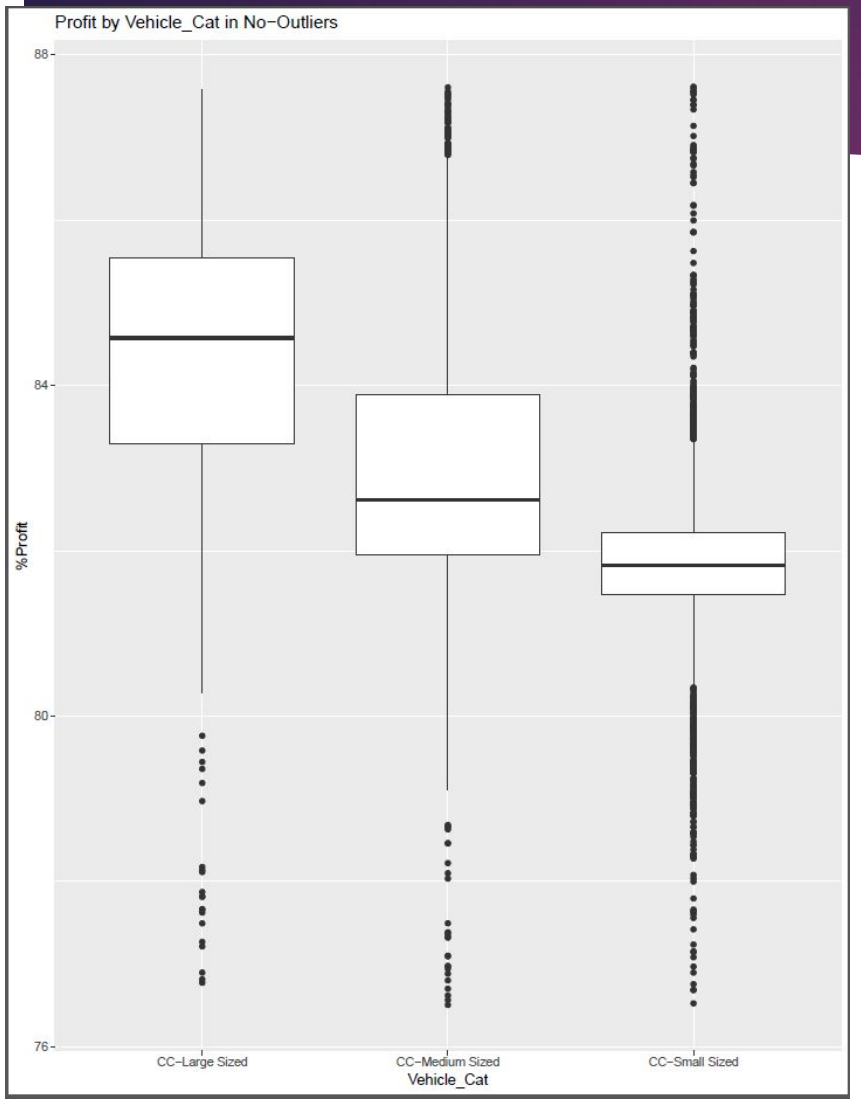
Analysis - Vehicle Category

Our Hypothesis - Profit more for Large Size CC vehicles that rest of categories





Profit with Vehicle Category




```
> VehicleCat_No_outlier_anova <- aov(Profit~Vehicle_Cat,data=ins_rm_ex_IDV_pro)
> summary(VehicleCat_No_outlier_anova)
              Df Sum Sq Mean Sq F value               Pr(>F)
Vehicle_Cat    2   4272  2136.0    1119 <0.0000000000000002 ***
Residuals   6232  11898     1.9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> T_VehicleCat_No_outlier_anova <- TukeyHSD(VehicleCat_No_outlier_anova)
> T_VehicleCat_No_outlier_anova
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Profit ~ Vehicle_Cat, data = ins_rm_ex_IDV_pro)

$Vehicle_Cat
                                       diff       lwr        upr p adj
CC-Medium Sized-CC-Large Sized    -1.436567 -1.573591 -1.2995437     0
CC-Small Sized-CC-Large Sized     -2.467407 -2.596858 -2.3379569     0
CC-Small Sized-CC-Medium Sized    -1.030840 -1.121248 -0.9404321     0
```

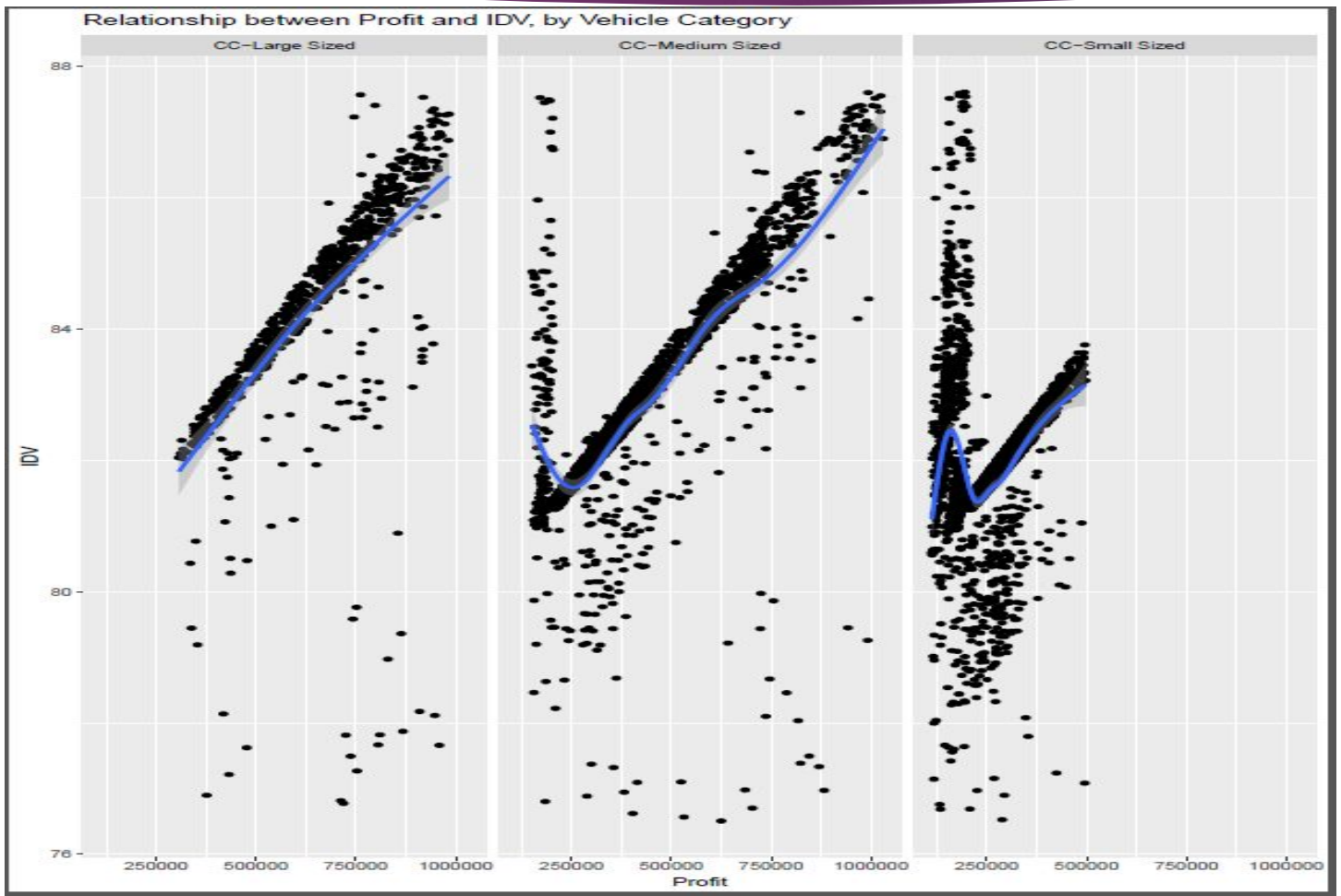
Conclusion - We are successful to reject the Null Hypothesis.

Profit increase as the CC size of Vehicle increases.

Linear Model - Profit by IDV + Vehicle Category

```
> mod_r<- lm (profit~IDV+Vehicle_Cat, data=ins_rm_ex_IDV_pro_r) #value improves to 52.52%
> summary(mod_r)

Call:
lm(formula = profit ~ IDV + Vehicle_Cat, data = ins_rm_ex_IDV_pro_r)

Residuals:
    Min      1Q  Median      3Q     Max
-7.9886 -0.2483 -0.0822  0.2215  6.1873

Coefficients:
                             Estimate   Std. Error  t value            Pr(>|t|)
(Intercept)                80.648942064  0.076745076 1050.868 < 0.0000000000000002 ***
IDV                         0.000005724  0.000000100   57.227 < 0.0000000000000002 ***
Vehicle_CatCC-Medium Sized -0.188144371  0.051833393   -3.630             0.000286 ***
Vehicle_CatCC-Small Sized  -0.217725677  0.059638505   -3.651             0.000264 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.103 on 6221 degrees of freedom
Multiple R-squared:  0.5255,    Adjusted R-squared:  0.5252
F-statistic:  2296 on 3 and 6221 DF,  p-value: < 0.00000000000000022
```

Linear Model - Profit by IDV + Vehicle Category ...



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.

"Company should focus on insuring vehicles with higher cubic capacity".

*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.

Questions ?

or

Suggestions.