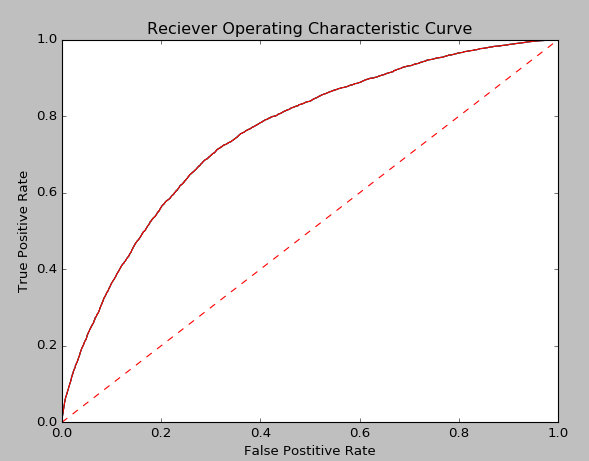
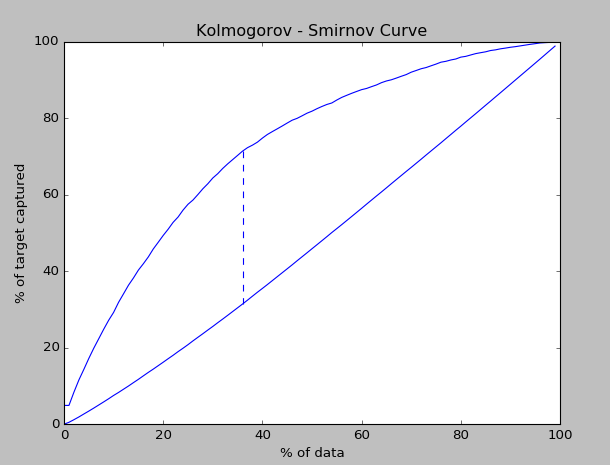
**Normal unweighted logistic Regression**

**AUC = 0.756913087216**

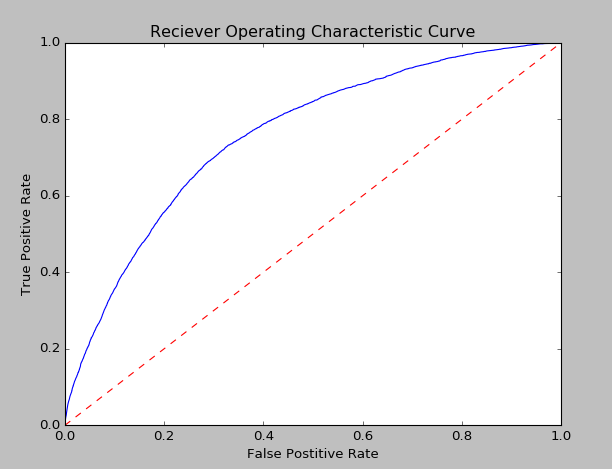


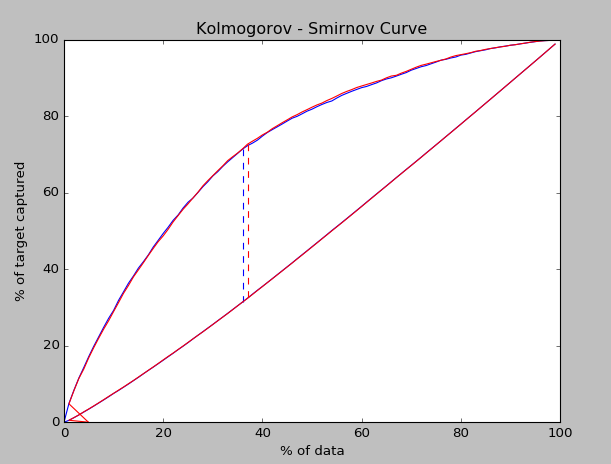
**K-S Point = 0.399**



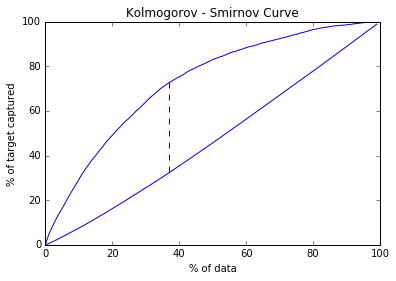
**Weigh- balanced Logistic Regression**

**AUC: 0.757971406736**

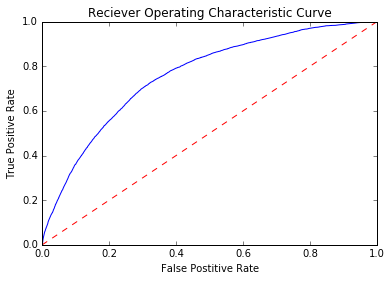


**K- S Point = 0.4031** 

**Bootstrap aggregation on original data**

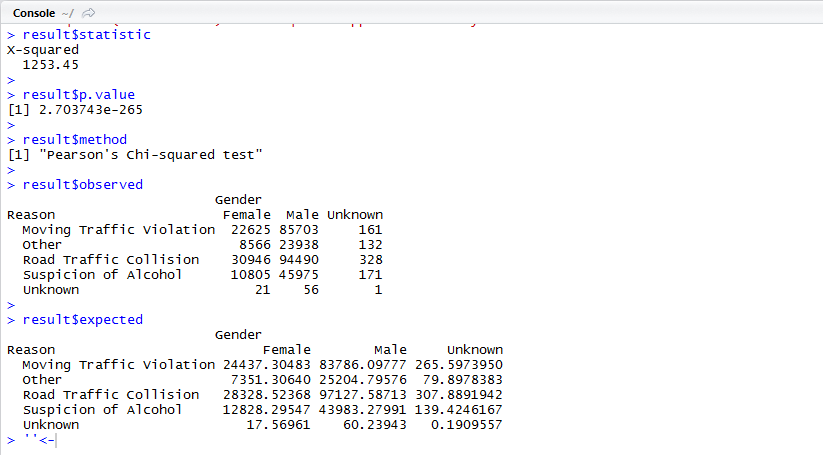


K-S point = 40.427

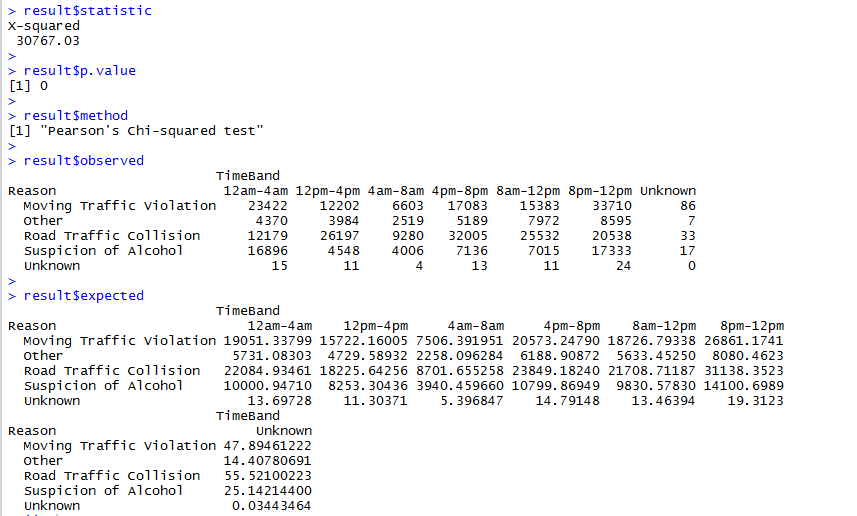


AUC = 0.7597

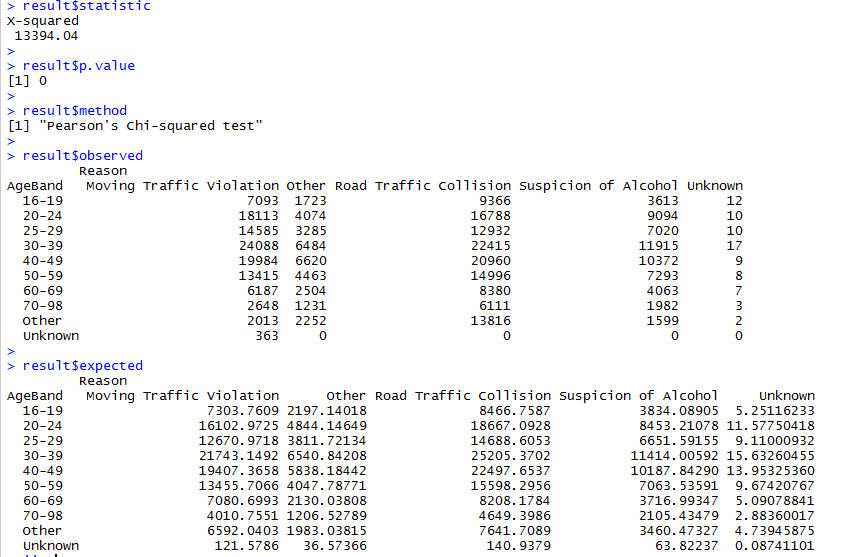
**Results of the Pearson's Chi Square Test of Independence:**



First, the Pearson’s Chi Squared Test of Independence, was carried out on the variables Reason and Gender. Now, the null hypothesis in this test is that the variables Gender and Reason are independent. The result of the test shows a p value of nearly zero. This means, at a 0.05 significance level, we reject the null hypothesis, and conclude that, the variables Gender and Reason are indeed dependent. The expected and observed values of the combinations of these variables shed some light on the details of the dependencies. For example, the number of cases for females due to suspicion of alcohol should have been about 12800, whereas the actual figure is about 10800. The corresponding figure for male was expected to be about 44000, but the actual figure is 46000. This clearly suggests that, the male drivers are more likely to be detained by the Police under suspicion and asked to undergo the test. Also, the expected number of cases for females due to road traffic collisions was around 28300, but the actual figure is close to 31000. The same figure for males was expected around 97200, but is close to 94000. This indicates that females are comparatively less adept at driving and are more likely to be involved in a road traffic collision



The Chi Squared test was then carried out on the variables Reason and TimeBand. The resulting p value is less than 0.05, and hence we conclude that the variables Reason and TimeBand are also dependent. From the result of the test, it can be seen that the number of breath test cases due to suspicion of alcohol in the time band 12 midnight to 4 am, at about 17000, is way higher than the expected 10000. The same figure for the time band 8 am to 12 noon, is about 7000, much lower than the expected about 10000. This shows that , the a person driving in the time period between midnight to 4 am, is more likely to be asked to undergo the test on the grounds of suspicion of alcohol, than when driving the morning hours of 8 am to 12 noon.



The same test was further conducted on the variables Reason and AgeBand , and Reason and WeekType. Both of these tests show that there is a relationship between these variables. It can be seen from the results of the tests, that the numbers of cases due to moving traffic violations, are higher than expected for individuals in the age group 20 to 24, 25 to 29 and 30 to 39, but are more or less as expected for other age groups. This means that the drivers in these age groups are at a higher risk of committing a traffic violation, as compared to drivers in other age groups. Surprisingly, the individuals in these very age groups are less likely than others to be involved in a traffic collision, as the numbers of cases due to collisions for these, are less than expected. This can be explained by the fact that people in these groups, i.e., young people below the age 39, generally have good reflexes and are able to maneuver through difficult situations, where others might find themselves in a traffic collision.

**Visualizations:**

****

****

****

****