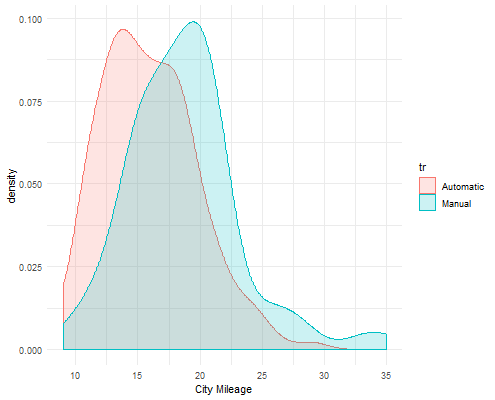
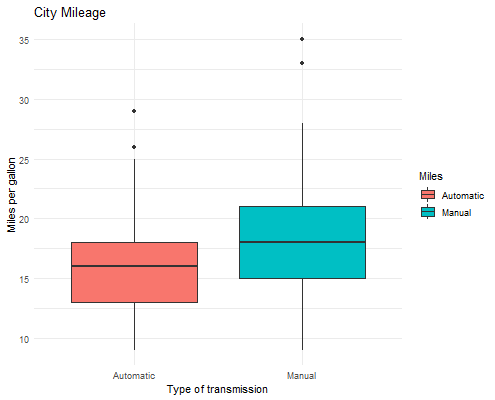
Intelligent Data Analysis - Homework 1.1

Panagiotis, Michael, Ignacio, Javier, Daniel

October 14th, 2018

# Q1: Is there a significant difference in City / Highway mileage or in Engine Displacement between automatic and manual transmission cars?

This question is devided in three parts depending on the three variables we are analysing: city milage, highway milage and engine displacement. On this page, only the plots for City mileage are presented:



*Boxplot and density plot of City Mileage vs type of transmission*.

The analysis is done with a new variable that differs between automatic and manual cars.   
The objective is to determine whether the mean between both populations is equal.

In order to test if there is a difference between the means, we propose to use One-Way ANOVA. Previously, we have to test if the variables fulfil some assumptions: Normality of errors and Homocedasticity. To achieve this, we are going to use Jarque Bera test and Levene test, respectively.

In case the assumptions above are met, we finally use the One-Way ANOVA test. Otherwise, we apply the non-parametric Kruskall-Wallis test. There is also an option of performing the whole process removing the outliers from the City Mileage Data.

The results of Jarque Bera test and Levene test are the following:

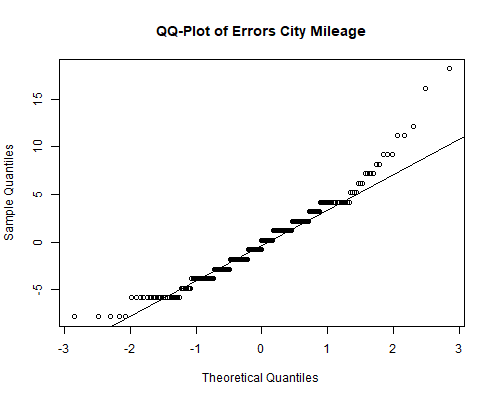
Jarque Bera Test

X-squared = 48.448, df = 2, p-value = 3.018e-11

Levene Test

Test Statistic = 0.3209, p-value = 0.5716

Due to the low value of p-value in the Jarque Bera test, we can reject the null hypothesis about the normality of the errors. Also, we can confirm this non-distribution by observing the following QQ plot:



At this point, we apply Kruskall-Wallis test instead of ANOVA:

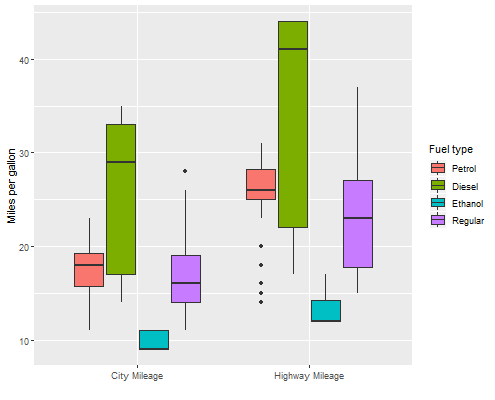
Kruskal-Wallis

chi-squared = 21.988, df = 1, p-value = 2.743e-06

By obtaining a low value very closed to zero as p-value with the test above, we can reject again the null hypothesis that the distribution of the means for automatic and manual transmission are similar.

# Q2: Has fuel type (or cylinders) any influence on highway mpg or city mpg?

For this question we have created Boxplots comparing the City Mileage to the Highway Mileage with respect to the fuel type of each automobile and the number of cylinders each automobile has. Here is the example of the fuel type:



The boxplot above shows that each type of fuel represents a different amount of variation in the miles per gallon variable, so we can observe that there is much overlap of values between some types of fuel such as Petrol, Diesel and Regular. Thus, differences in the means could be produced by chance. At this point we can use ANOVA to compare the variation among samples with the ones within groups.

Df Sum Sq Mean Sq F value Pr(>F)

Fl 3 802 267.3 18.18 1.29e-10 \*\*\*

Residuals 229 3367 14.7

---

Signif. codes: 0 \*\*\* 0.001 \*\* 0.01 \* 0.05 . 0.1 1

By obtaining low F value and p-value we can accept the alternative hypothesis that there is a significant relationship between the fuel type and miles per gallon. However, with ANOVA we only know that not all the means are equal but not which groups are different from the others. To achieve this we can apply the Tukey test:

diff lwr upr p adj

e-d -15.8500000 -21.507492 -10.1925081 0.0000000

p-d -8.2346154 -12.881191 -3.5880397 0.0000438

r-d -8.8619048 -13.365566 -4.3582440 0.0000044

p-e 7.6153846 3.846514 11.3842552 0.0000023

r-e 6.9880952 3.396900 10.5792900 0.0000057

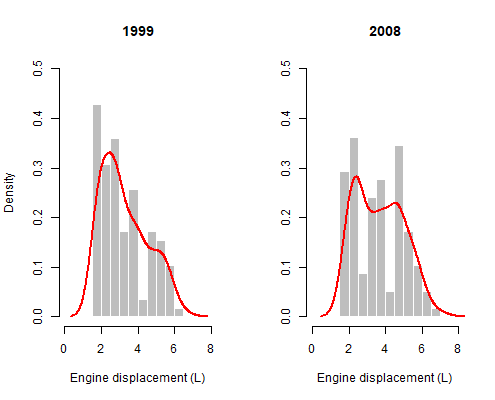
r-p -0.6272894 -2.202132 0.9475537 0.7315324

With the table above, we can conclude that there is a significant difference in the milles per gallon between all types of fuels but petrol and ethanol.

# Q3: Is there any difference in car requests between the years 1999 and 2008?

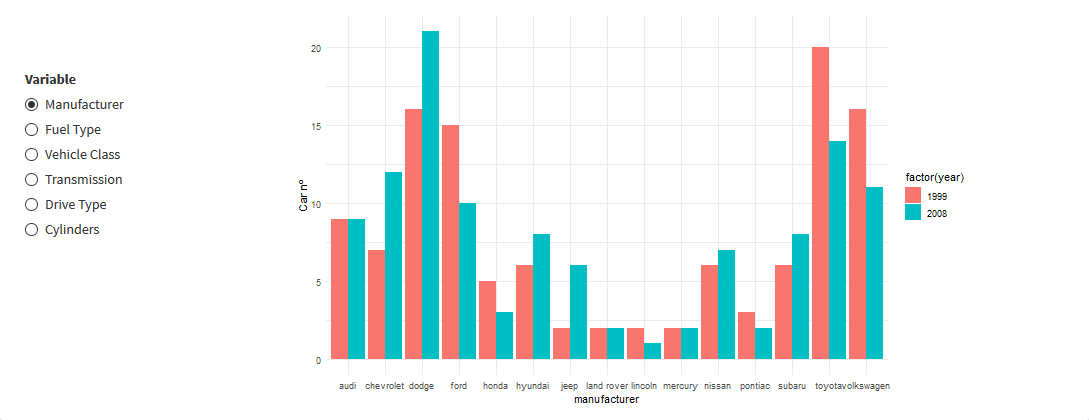
For this question we compared the distribution of some of the variables in years 1999 and 2008 in order to determine whether the requests on automobiles have been changed.

On the histogram plot we compare the distribution of the engine displacement for each automotive through the years.



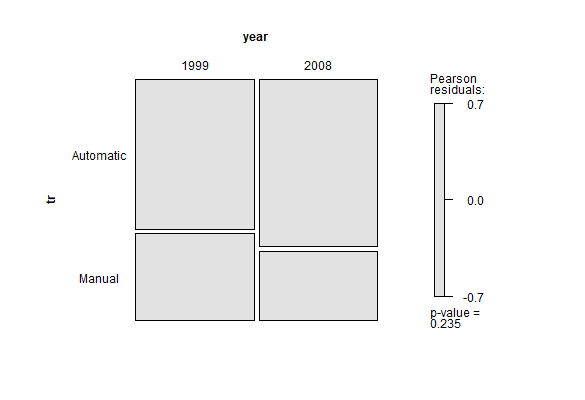
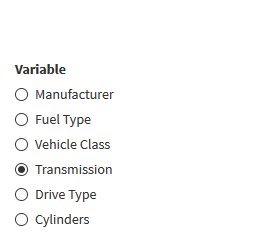
As we can observed in the histogram above, in 2008 there has been a growing demand on vehicles with higher engine displacement.

We have used a reactive bar chart to visualize how the different levels of the categorical variables are distributed between 1999 and 2008.

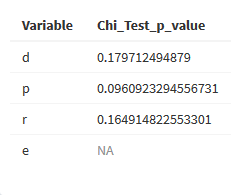
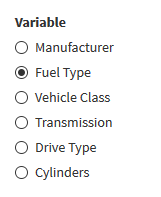


In addition, we have used a reactive mosaic plot for each of the categorial variables showing in the list below and the categorical variable of the year. By using this plot, we can observe a pair of highlighted facts:

* The use of ethanol has appeared as a strong option as fuel type.
* It seems to be a reasonable tendency of displacement of manual transmission for automatic one.
* It might be a slightly tendency of stop using front-wheel drive type.
* The use of five cylinders in the engine has started to be strongly considered.

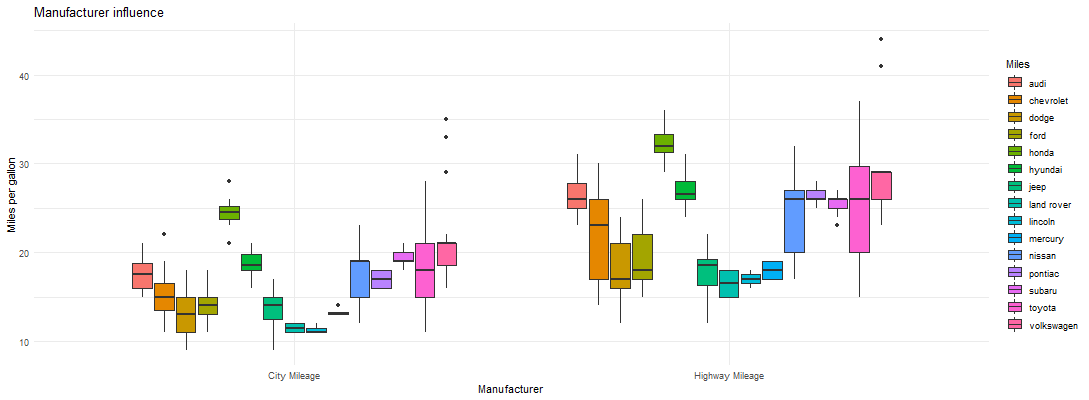


At this point, we have included a reactive chi-square test for each categorical variable in the list below and the categorical variable of the year. Given that none of the p-values obtained are lower than the significance level of 0.05, we cannot reject the null hypothesis of non-independence between the variables, therefore, we cannot conclude if the variables are related to each other or not.



# Q4: Is there any difference in city and highway mpg between manufacturers?

For this question we have used a double box plot to show how the different manufacturers engines behave in terms of consumption for City driving and Highway driving.



In order to test if there is a difference between the means, we propose to use One-Way ANOVA. Previously, we have to test if the variables fulfil some assumptions: Normality of errors and Homocedasticity. To achieve this, we are going to use Jarque Bera test and Levene test, respectively.

In case the assumptions above are met, we finally use the One-Way ANOVA test. Otherwise, we apply the non-parametric Kruskall-Wallis test. There is also an option of performing the whole process removing the outliers from the City Mileage Data.

The results of Jarque Bera test and Levene test are the following:

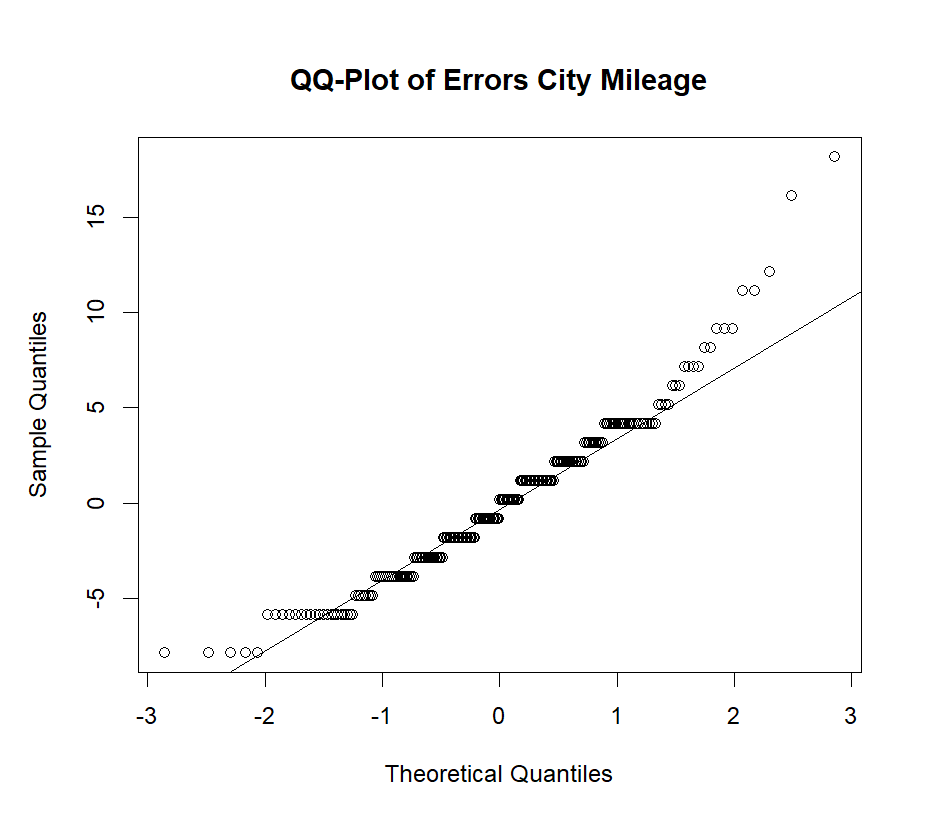
Jarque Bera Test

X-squared = 48.448, df = 2, p-value = 3.018e-11

Levene Test

Test Statistic = 2.7775, p-value = 0.000806

Due to the low value of p-value in the Jarque Bera test, we can reject the null hypothesis about the normality of the errors. Also, we can confirm this non-distribution by observing the following QQ plot:



At this point, we apply Kruskall-Wallis test instead of ANOVA:

Kruskal-Wallis

chi-squared = 141.14, df = 14, p-value < 2.2e-16

By obtaining a low value very closed to zero as p-value with the test above, we can reject again the null hypothesis that the distribution of the means in the behavior of manufacturers in both city and highway is similar.

# Data Analysis Plan