# Intelligent Data Analysis - Homework 1.1

Panagiotis, Michael, Ignacio, Javier, Daniel

October 14th, 2018

## 1 Introduction

We have chosen the Mileage dataset, it contains data about the fuel consumption of different car manufacturers which had a new release every year between 1999 and 2008. The dataset contains 234 records of 11 different variables, which are:

- **Displ:** engine displacement or cylinder admission volume in liters, quantitative variable.

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|------|---------|--------|------|---------|------|
| 1,6  | 2,4     | 3,3    | 3,472 | 4,6    | 7    |

- **Cty:** city mileage in miles per gallon, quantitative variable.

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|------|---------|--------|------|---------|------|
| 9    | 14      | 17     | 16,83 | 19     | 35   |

- **Hwy:** highway mileage in miles per gallon, quantitative variable.

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|------|---------|--------|------|---------|------|
| 12   | 18      | 24     | 23,39 | 27     | 44   |

- **Manufacturer:** car manufacturer, categorical variable.

| Manufacturer | Nº Entries |
|---|---|
| Dodge | 37 |
| Toyota | 34 |
| Volkswagen | 27 |
| Ford | 25 |
| Chevrolet | 19 |
| Audi | 18 |
| Hyundai | 14 |
| Subaru | 14 |
| Nissan | 13 |
| Honda | 9 |
| Jeep | 8 |
| Pontiac | 5 |
| Land Rover | 4 |
| Mercury | 4 |
| Lincoln | 3 |
| **Total** | **234** |

- **Year:** year of manufacturing (1999 or 2008), categorical variable.

| Year | Nº Entries |
|---|---|
| 1999 | 177 |
| 2008 | 177 |

- **Class:** vehicle class, categorical variable

| Class | Nº entries |
|---|---|
| Suv | 62 |
| Compact | 47 |
| Midsize | 41 |
| Subcompact | 35 |
| Pickup | 33 |
| Minivan | 11 |
| 2seater | 5 |

- **Cyl:** number of cylinders, categorical variable.

| Nº gears | Nº entries |
|---|---|
| 4 | 81 |
| 6 | 79 |
| 8 | 70 |
| 5 | 4 |

- **Model:** car model name (38 different models), categorical variable.

- **Trans:** type of transmission (automatic or manual), categorical variable.

| Transmission | Nº entries |
|---|---|
| Auto | 156 |
| Manual | 77 |

- **Dvr:** drive type (front wheel, rear wheel, 4 wheel), categorical variable.

| Drive type | Nº entries |
|---|---|
| front-wheel | 105 |
| 4-wheel | 103 |
| rear-wheel | 25 |

- **Fl:** fuel type (petrol, diesel electric, ethanol, regular), categorical variable.

| Fuel type | Nº entries |
|---|---|
| Regular | 168 |
| Petrol | 52 |
| Ethanol | 8 |
| Diesel-elect | 5 |
| "c" | 1 |

# 2 Questions and data analysis plan

For this homework 1.1, we have decided to create a shiny dashboard application that allowed us to express our analysis in a more convenient way. We have included our results of the statistical tests for every question under each plot in the application.
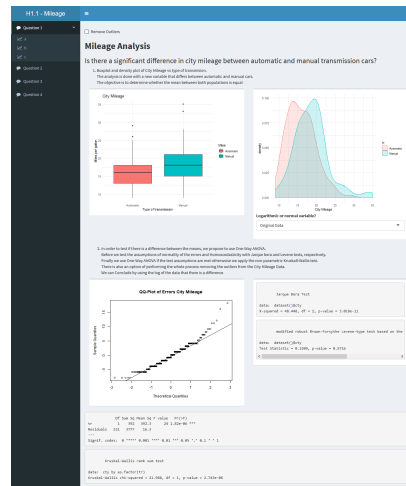


## 2.1 Data Analysis Plan

The following is the scheme of the data analysis plan we followed:
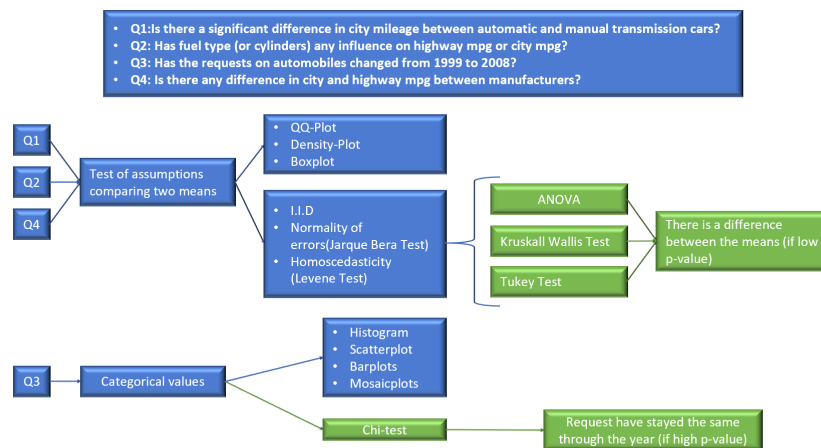
## 2.2   Q1, Is there a significant difference in City / Highway mileage or in Engine Displacement between automatic and manual transmission cars?

This question is divided in three parts depending on the three variables we are analyzing: city mileage, highway mileage and engine displacement. The following is the example for City mileage:
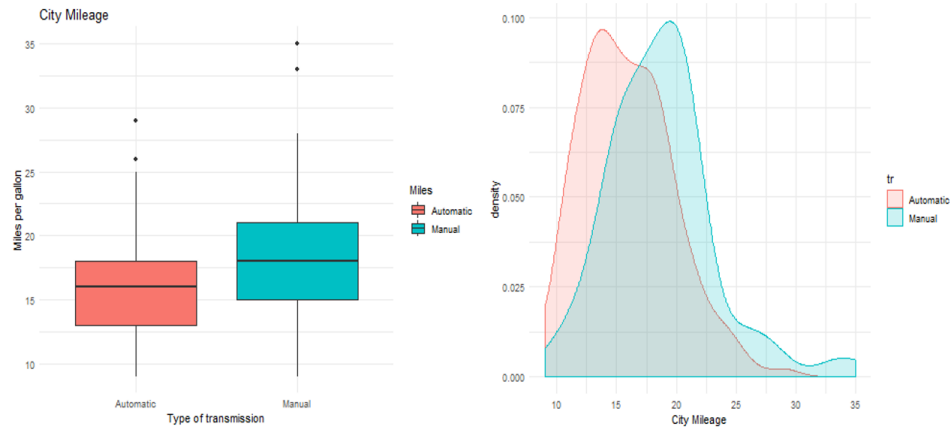


The analysis is done with a new variable that differs between automatic and manual cars. The objective is to determine whether the mean between both populations is equal.

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 fulfill some assumptions: Normality of errors and Homoscedasticity (as we assume the variables are i.i.d). 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:

**QQ-Plot of Errors City Mileage**
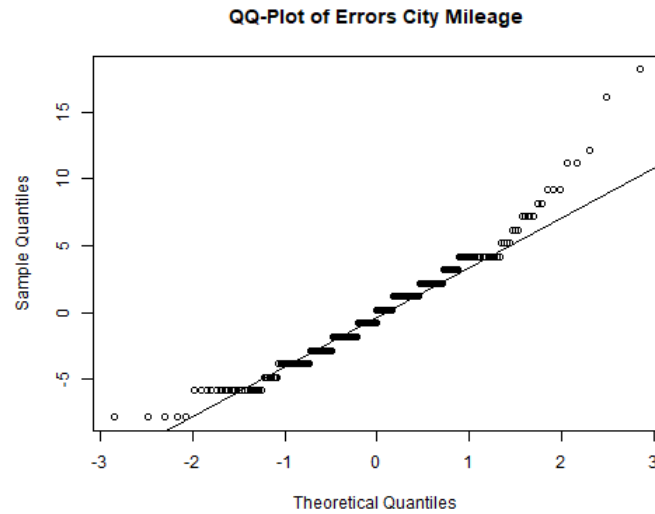


At this point, we apply Kruskall-Wallis test instead of ANOVA:

```
Kruskal−Wallis
chi−squared = 21.988, df = 1, p−value = 2.743e−06
```

By obtaining a low value very close to zero as the p-value above, we can reject again the null hypothesis that the distribution of the means for automatic and manual transmission are similar. There is also a possibility of repeat the process with the log of the data in the shiny app.

## 2.3 Q2, Has fuel type (or cylinders) any influence on high-way 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:
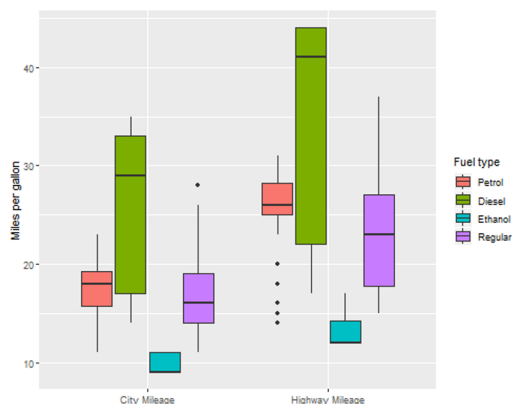


The boxplot above shows that each type of fuel represents a different amount of variation in the miles per gallon variable, so we can observe that there is much overlap of values between some types of fuel such as Petrol, Diesel and Regular. Thus, differences in the means could be produced by chance. 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 | $1.29\,\mathrm{e}{-10}$ | *** |
| 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:

6

|     | **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 miles per gallon between all types of fuels but petrol and ethanol.

## 2.4 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.
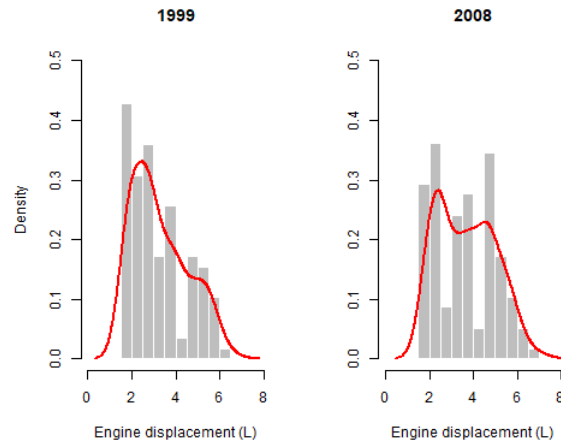
We have used a reactive bar chart to visualize how the different levels of the categorical variables are distributed between 1999 and 2008. Following with a reactive Mosaic plot and Chi-test table.
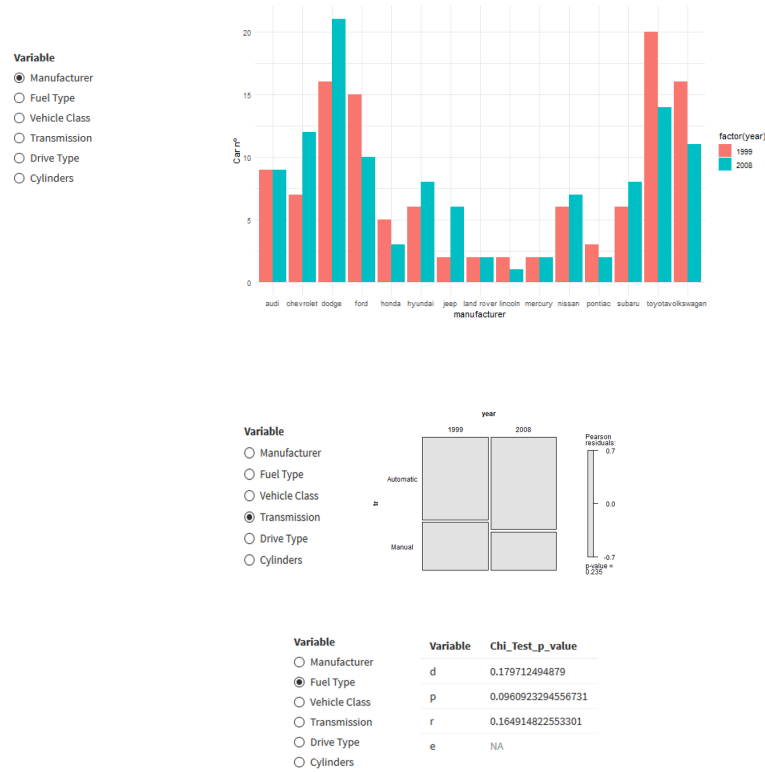






## 2.5 Q4, Is there any difference in city and highway mpg between manufacturers?

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. The tests used are similar as the used in the previous questions (see the shiny app):