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 litres, 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 milles 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.

Transm	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, etanol, regular), categorical variable.

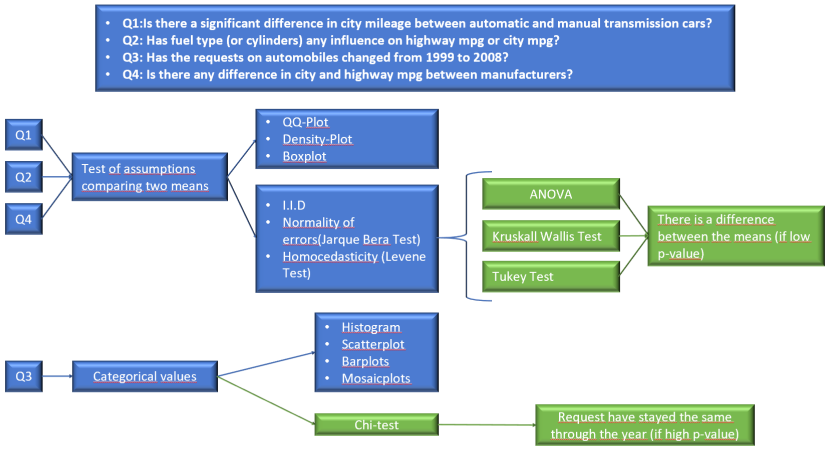
Fuel type	nº entries
Regular	168
Petrol	52
Ethanol	8
Diesel-elect	5
"c"	1

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.

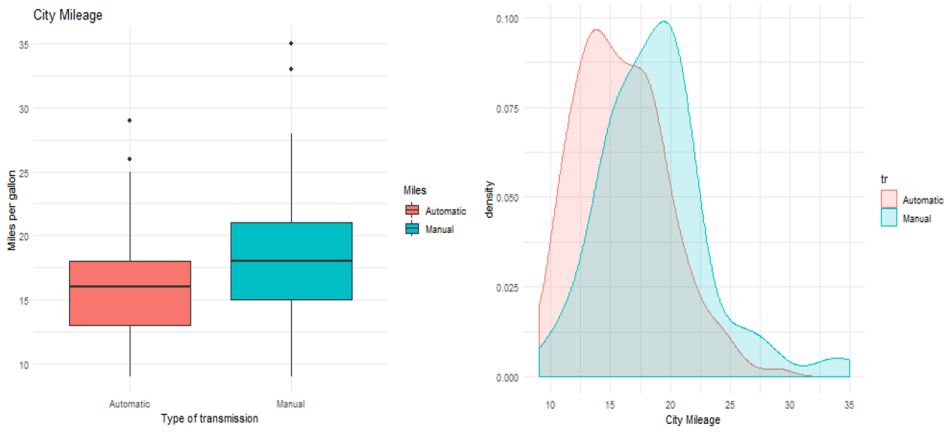
2.1 Data Analysis Plan

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 analysing: city milage,highway milage and engine displacement. The following is the example for City mileage:

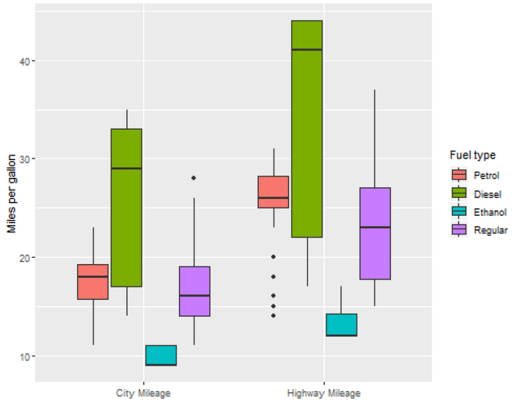


Boxplot and density plot of City Mileage vs type of transmision. The analysis is done with a new variable that differs between automatic and manual cars. The objective is to determine whether the mean between both populations is equal.

In order to test if there is a difference between the means, we propose to use One-Way ANOVA. Before we test the assumptions of normality of the errors and Homoscedasticity with Jarque bera and Levene tests, respectively. Finally we use One-Way ANOVA if the test assumptions are met otherwsise we apply the non-parametric Kruskall-Wallis and Tukey tests. There is also an option of performing the whole process removing the outliers from the City Mileage Data. We can Conclude by using the log of the data that there is a difference.

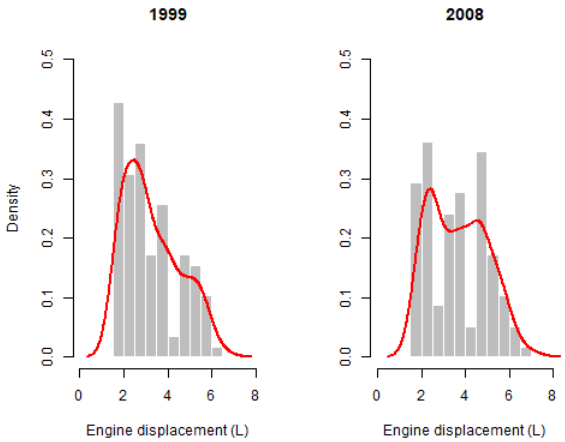
2.3 Q2, Has fuel type (or cylinders) any influence on highway mpg or city mpg?

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:

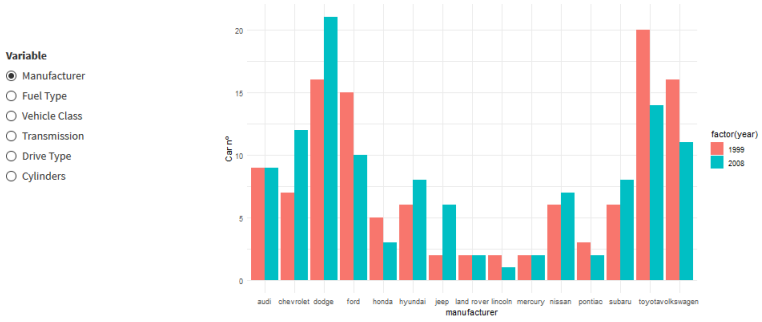


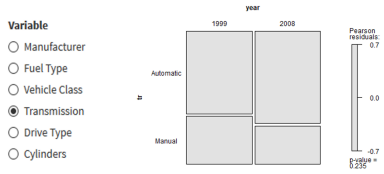
2.4 Q3, Is there any difference in car requests between the years 1999 and 2008?

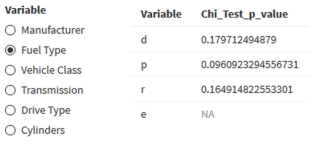
For this question we compared the distribution of some of the variables in years 1999 and 2008 in order to determine whether the requests on automobiles have been changed. On the histogram plot we compare the distribution of the engine displacement for each automotive throught the years.



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.







Variable	Chi_Test_p_value
d	0.179712494879
p	0.0960923294556731
r	0.164914822553301
e	NA

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.