Homework 1.2 Report

Monika Wysoczanska, 180817

Manuel Barbas, 180832

Diogo Oliveira, 180832

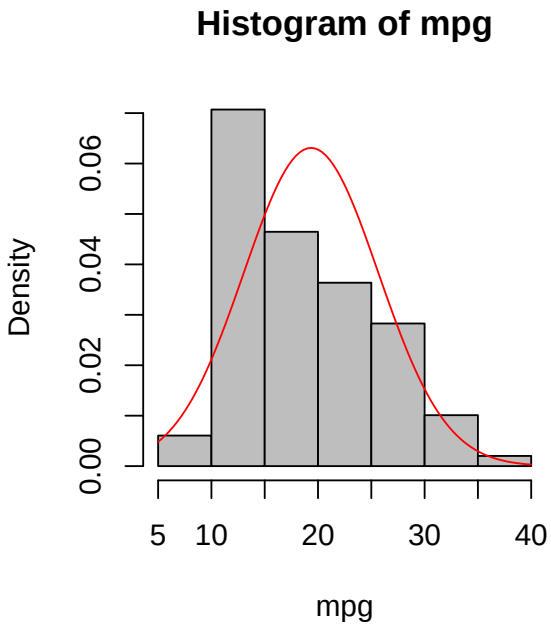
Cars Dataset

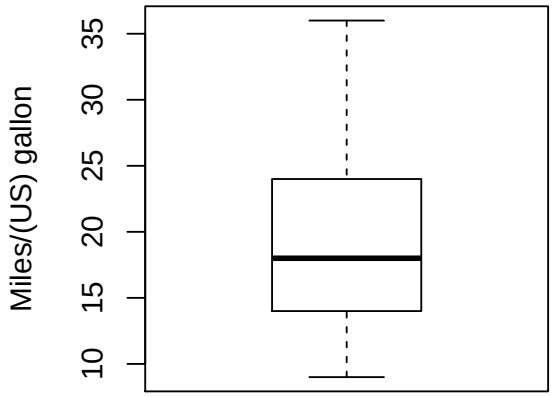
Description

Cars dataset consists of 99 observations of different car models. Each one of them is described by 9 different variables, 4 categorical: cylinders, car name, model year, origin, and 5 quantitative: engine displacement, horsepower, weight, acceleration, mpg.

Univariate analysis

The variable chosen for this analysis was the mpg or Miles Per Gallon. Let's take a closer look at it.





min	max	range	sum	median	mean	SE.mean	CI.mean.0.95	var	std.dev	coef.var
9	36	27	1915.5	18	19.348	0.635	1.261	39.972	6.322	0.327

```
    D'Agostino skewness test

data:  mpg
skew = 0.66097, z = 2.65280, p-value = 0.007982
alternative hypothesis: data have a skewness


    Anscombe-Glynn kurtosis test

data:  mpg
kurt = 2.52320, z = -0.99339, p-value = 0.3205
alternative hypothesis: kurtosis is not equal to 3


    Shapiro-Wilk normality test

data:  mpg
W = 0.93467, p-value = 0.0001004
```

Looking at the histogram, by the significant right-handed tail, we can already say that the data is positively skewed. It may also be concluded by the relation between mean and median (median < mean). We proved this fact by conducting the D'Agostino skewness test with the 0.66, and given p-value, which lets us reject the null hypothesis and accept the alternative hypothesis, that data is skewed indeed. On the other hand we see, that the data 'passes' the kurtosis test, and we cannot reject the hypothesis that the kurtosis value is

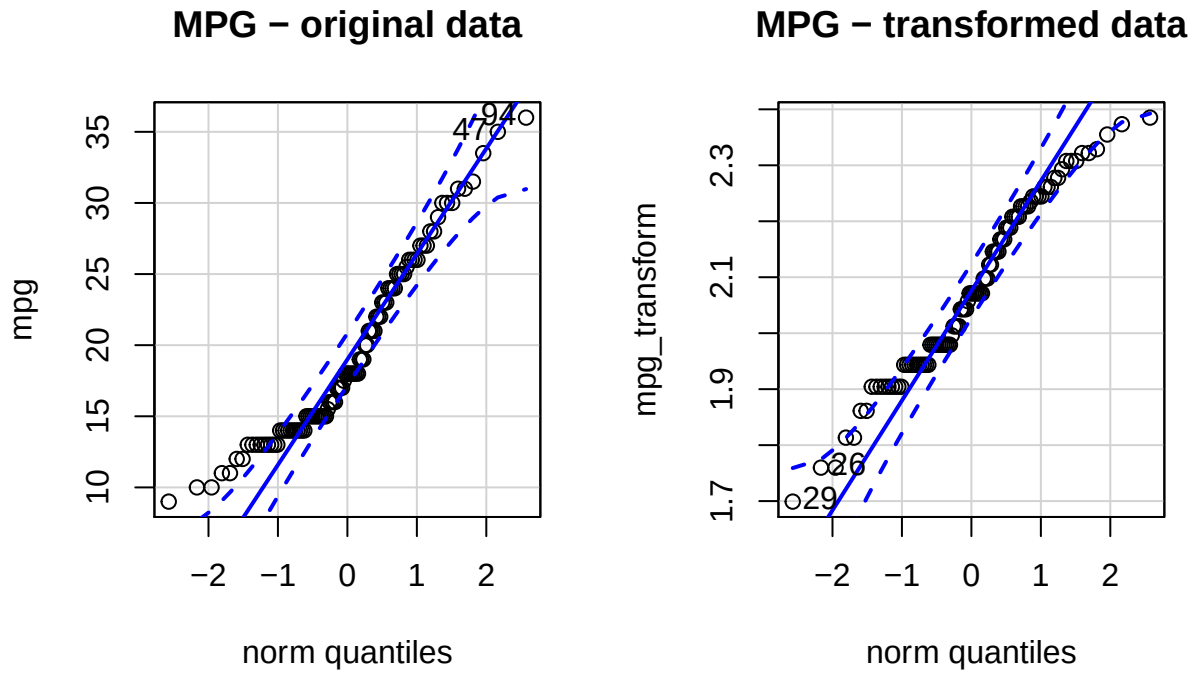
equal to 3 (the normal distribution characteristic). In order to give the final statement on the distribution of the data we conduct the Shapiro-Wilk normality test. We reject the null hypothesis, since we obtained a very low p-value (0.0001004), and conclude that our data isn't of normal distribution.
In the end, we studied the dispersion of the data with the coefficient variation around 33%, and no outliers detected, based on the boxplot above.

Next we tried to improve the normality of the data by applying Box-Cox tranformation. The lambda for the transformation we obtained:

```
Estimated transformation parameter
       mpg
-0.2450011
```

The analysis of the dataset after transformation is given below.

```
[1] 94 47
```




```
[1] 29 26
```

```
    D'Agostino skewness test

data:  mpg_transform
skew = 0.021572, z = 0.093308, p-value = 0.9257
alternative hypothesis: data have a skewness
```

```
    Anscombe-Glynn kurtosis test

data:  mpg_transform
kurt = 2.1934, z = -2.4357, p-value = 0.01486
alternative hypothesis: kurtosis is not equal to 3
```

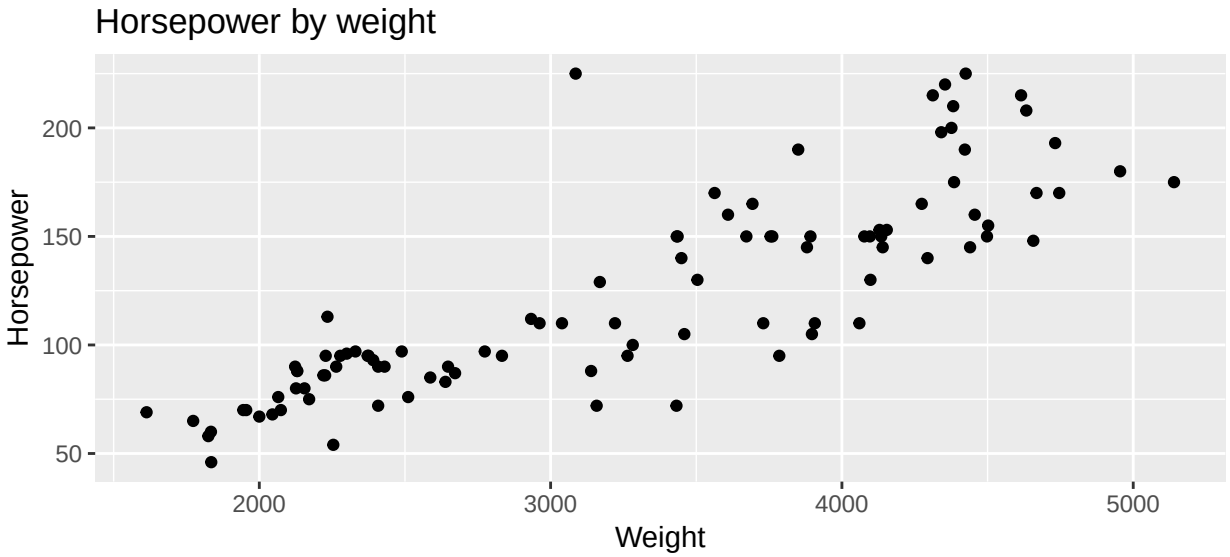
```
    Shapiro-Wilk normality test

data:  mpg_transform
W = 0.97143, p-value = 0.02976
```

From the first qqPlot we can be sure that the distribution is not normal because of the slightly positive distribution and the soft nonpeaked distribution for the original data. However, the transformation didn't improve much the data normality as it is shown on the qqPlot on the right. We confirm our observation by conducting another set of kurtosis, skewness and normality tests. The transformation improved the skewness of the data, as we obtained the value closer to 0. Nonetheless, the data still fails two other tests, for kurtosis as well as final Shapiro-Wilk test for normality.

Bivariate analysis

We start with basic scatterplot to see the distribution of the data for the two chosen variables, which are: 1. Horsepower 2. Weight



We observe the general positive relationship between the two variables, but by the given plot we cannot say anything about bivariate normality yet. What we can conclude by now, is we'll probably be dealing with some outliers in the data. First of all we perform Mardia's multivariate normality test.

Test	Variable	Statistic	p value	Normality
Shapiro-Wilk	Column1	0.9370	1e-04	NO
Shapiro-Wilk	Column2	0.9347	1e-04	NO

Test	Statistic	p value	Result
Mardia Skewness	21.2222121214886	0.00028610719730013	NO
Mardia Kurtosis	1.3379526217569	0.180911881729716	YES
MVN	NA	NA	NO

As we can see both variables fail the univariate normality test. They also fail bivariate normality test, because of the skewness. We try to apply the Box-Cox transformation so as to improve bivariate normality, with the parameters given below:

```
Estimated transformation parameters
         Y1          Y2
-0.07851897  0.46412207
```

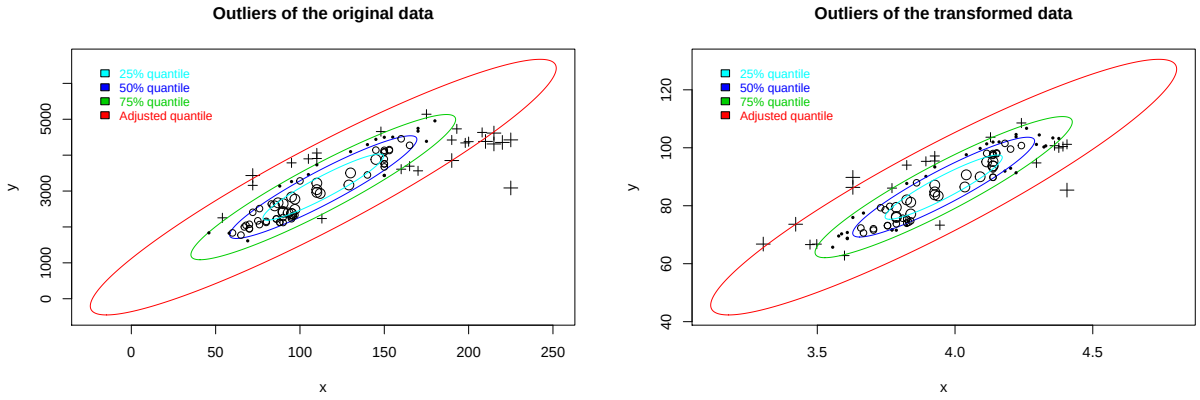
After applying the transormation we conduct bivariate normality analysis the same way as before.

Test	Statistic	p value	Result
Mardia Skewness	0.477156091683814	0.975686394361116	YES
Mardia Kurtosis	-0.443162058283486	0.657648520358104	YES
MVN	NA	NA	YES

As we can see, the normality has improved as data after Box-Cox transform passes both tests, for kurtosis as well as for skewness.

Outliers detection

Another thing we want to conduct during our bivariate analysis is the outliers detection. To achieve this we use 'mvoutlier' package. Firstly, we apply 'pcout' method on the original dataset.



We detected 18 outliers in the original dataset based on the bivariate analysis.

Then we applied the same method for the transformed dataset and found only 5 outliers, which are:

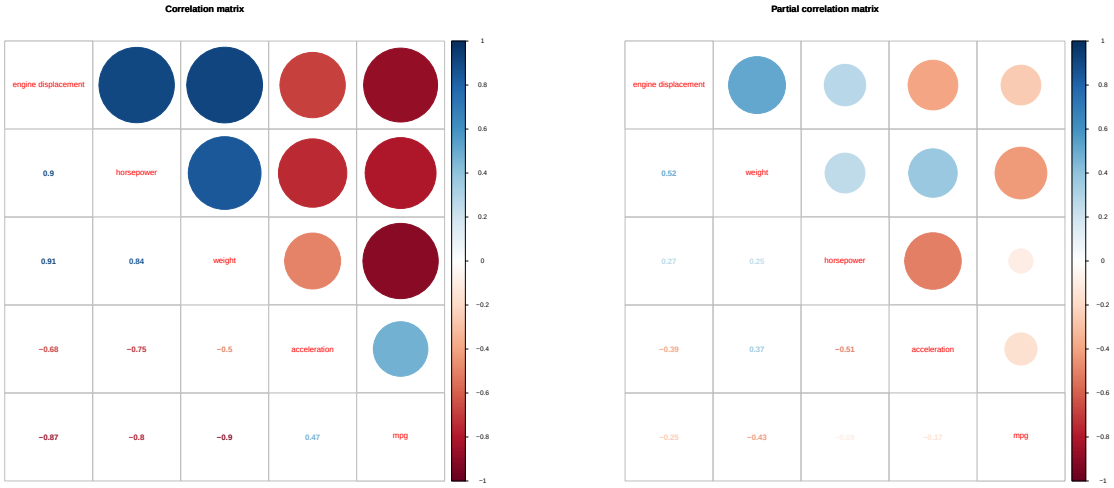
	name	horsepower	weight
14	buick estate wagon (sw)	225	3086
20	volkswagen 1131 deluxe sedan	46	1835
52	volkswagen type 3	54	2254
76	mercury monarch	72	3432
77	ford maverick	72	3158

As we analyzed, more than 18% of the original dataset has been classified as outliers. Depends on the type of each outlier, and obviously the main objectivity of our analysis, sometimes we may consider outlier removal. In case of our dataset this is not an option, since it consists of only 99 observations.
After normality improvement we qualified about 5% samples as the outliers, and none of them seems to be a typing mistake. They should be taken into account in further analysis.

Linear relationship between multiple variables

The variables chosen for this analysis were:
1.mpg
2.engine displacement
3.horsepower
4.weight
5.accelaration

We start investigating linear relationship between variables by analyzing the correlation matrix and comparing the results with the partial correlation matrix.



The strongest linear correlation we observe between 'weight' and 'engine displacement' (positive relation), which is equal to 0.91, although looking at the parial correlation between the two variables, we see that the estimate significantly drops. It means that the correlation between 'weight' and 'engine displacement' also relies on some other variable, which in this case is the 'horsepower', as we can conclude from the plot.

We also analyzed coefficient of determination.

Table 1: Coefficient of determination

	x
mpg	0.832
engine displacement	0.916
horsepower	0.866
weight	0.898
acceleration	0.675

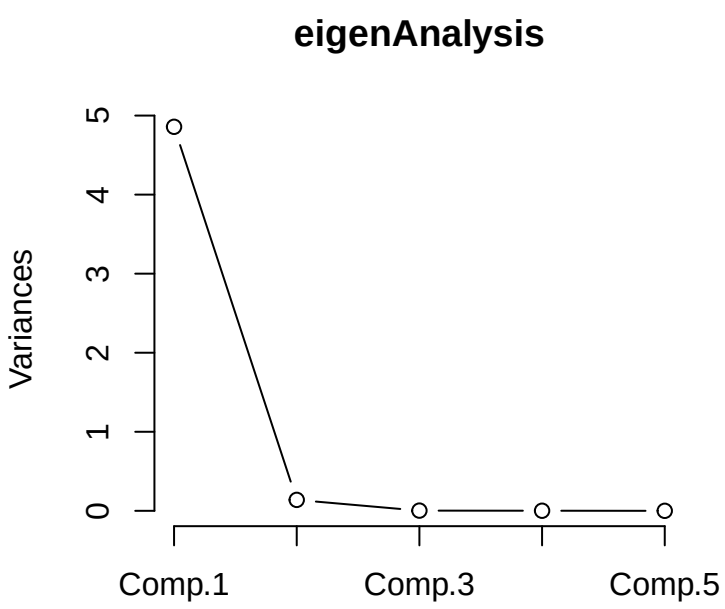
From this data we can say that engine displacement is the best linearly explained by others (R^2 = 0.91), followed by weight (R^2 = 0.89) and horsepower (R^2 = 0.86). The worst linearly explained by the others is acceleration (R^2 = 0.67), which is still the high value, meaning that the relation between all of the variables is pretty strong. This conclusion is easily proved by the determinant of the correlation matrix given below.

```
[1] 0.001688829
```

The determinant of the correlation matrix is low enough to say that the linear pairwise correlation between variables is strong, but there is none of the is a linear combination of the other (det not equal to 0).

The last part of our experiment is the Eigenanalysis. The eigen values and the corresponding eigen vectors are given below.

4.085	**0.644**	**0.121**	**0.091**	**0.058**
0.449	-0.402	0.699	0.384	-0.030
-0.481	0.049	0.126	0.443	0.744
-0.471	-0.146	0.576	-0.652	-0.005
-0.461	0.359	0.225	0.457	-0.632
0.363	0.828	0.336	-0.153	0.214



We can observe that the variables mostly involved in the overall linear dependence are 'engine displacement' and 'weight', because their variances are high in the least significant component (the one with the lowest eigenvalue). We already know that these two variables are also mostly explained by other variables and even mostly correlated. At the end of our Cars dataset analysis, we wanted to indicate how many components we would use in the case of the necessity of dimentionality reduction. We used the simple 'elbow rule', and looking at the plot above, we see that when choosing only 2 main components, we could still represent our data without the significant loss of information.

Restaurant Tips Dataset

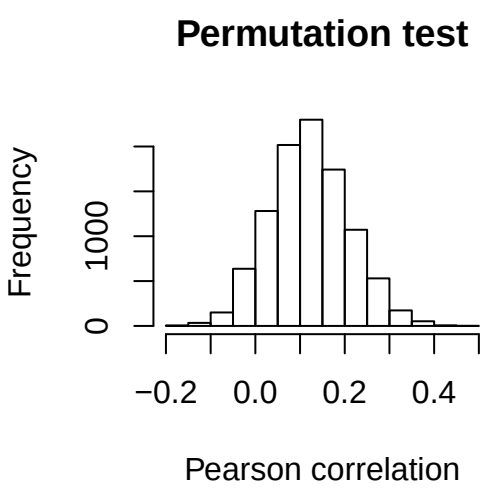
Permutation test

In this case we want to see relationship between the size of the bill and tip percent. First of all, we try to investigate this relationship by visualization.

Looking at the scatterplot given above we cannot actually say much about the correlation between these two variables. In order to analyze it, we conducted a permutation test running 1000 simulations.

We analyze the observed correlation between the variables (**Bill** and **PctTip**) by the histogram given below.

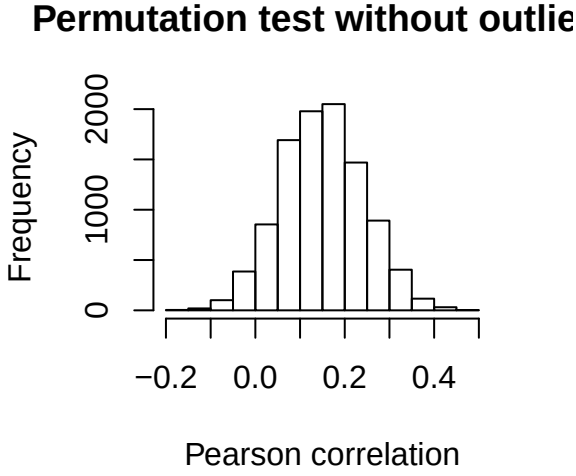
The following histogram represents the different values of the correlation between the variables after the permutation test



Using the vector of the correlation values calculated before was possible to do the test described in the exercise sheet (upper-tail test). The upper-tail test is a statistical test in which the critical area of a distribution is one-sided so that it is either greater than or less than a certain value, but not both. We obtained the following value:

When we look at the correlation values obtained through the test we can see that the values never exceed the value **0.4** (positive way) neither **-0.2** (negative way). Supported with the following figure, the values obtained with the permutation test infer that the strength of the association is small.

We ran the analysis once again excluding the bills with the tip above 30%, as we believe those generous customers might be considered as outliers. The histogram of correlation coefficients for this experiment is given below.



After outliers removal we can observe the slight shift to positive correlation between variables, as now the peak of our histogram fails closer to 0.2.

In order to state our final conslusion we conducted the upper-tailed test and we obtained the p-value equal to **0.4347** for the original data and **0.5585** after outliers removal. There is a weak evidence against the null

hypothesis, so we fail to reject it. To sum it up, we cannot say that percentage of the tips and the total amount of bill is not correlated, althogh this association isn't strong.