STATISTICAL ANALYSIS OF TOYOTA MODEL CARS

A FINAL PROJECT REPORT SUBMITTED
IN FULFILMENT OF THE REQUIREMENTS FOR COURSE STAT 250 –
APPLIED STATISTICS
DEPARTMENT OF STATISTICS OF
METU

BY

DİLAY GÜMÜŞ 2361301
MELİSA ÜNYILMAZ 2361574
İLAYDA YILMAZ 2361657
OKAN ÖZHAYAT 2361459

June 2021

Abstract

Almost all individuals own at least one car. It is inevitable to examine this situation according to some factors, while cars are included in people's lives, and their prices are increasing day by day. For this reason, in this project, we examined the factors associated with the prices of 120 Toyota model vehicles. To understand this data properly, we do descriptive statistics about our Toyota car data in general, analyze some graphs, and do some tests that are simple, multiple linear regression, parametric, nonparametric tests, hypothesis test, ANOVA, hypothesis testing, and we do chi-square test. We obtain a surprising result from this analysis. While we expected a negative relationship between mileage and price, there was a strong positive relationship between them.

Data Description

The data consists of 120 Toyota model and their price in Euro, fuel type, mpg, type of gearbox, road tax, mileage, and age. Data includes two categorical variables, which are fuel type and transmission. Age, tax, and mileage can be given as discrete variables. Price and mpg can be given as continuous variables.
Target Population : Toyota Model Cars

Sampling Methods

We divided our data according to age as greater than 18 and less than 18. Then, we took two samples from all 4 sampling methods for our data which is less than 18 years old. We had 8 samples for those less than 18. The tax means of these samples are as follows; [Appendix A1]

```
## Means of Samples: 179.6667 139.1875 180.125 122.2222 148.75 108.5 181.3333 134.7647
```

These samples were found with simple random sampling, systematic sampling, stratified random sampling and cluster sampling, respectively. The standard errors of these samples are as follows; [Appendix A2]

```
## Sampling Errors: -31.31973 9.159439 -31.77806 26.12472 -0.4030612 39.84694 -32.98639 13.58223
```

The smallest sampling error is -0.4031 ,obtained from systematic random sampling.

The mean of the samples we found for those over the age of 18 are as follows; [Appendix A3]

```
## Means of Samples: 239.6667 138.2174 155.5 129.05 171.2353 152.4167 232.3333 121.8
```

And so are standard errors; [Appendix A4]

```
## Sampling Errors: -100.6948 0.7544397 -16.52817 9.921831 -32.26346 -13.44484 -93.3615 17.17183
```

We are interested in the mean value of tax from age. The smallest one of sampling mistakes belongs to the pattern that's over the age of 18 is the second one with 0.75 acquired from systematic random sampling, and also, the smallest sampling mistakes belongs to the pattern that's under age of 18 is the fifth one with -0.40 with this technique. Researchers use the systematic sampling technique to pick the pattern individuals of a populace at normal intervals. It calls for the choice of a start line for the pattern and pattern length that may be repeated at normal intervals. This sort of sampling technique has a predefined range, and as a result, this sampling method is the least time-consuming. For this data, the quality sampling method is systematic Random Sampling. We pick every fourth variable from one to the end as a systematic pattern from our data.

Summary of the dataset:[Appendix A5]

```
##      price        fueltype tranmission      mpg             tax
##  Min.   :  1074   1:27     1:29        Min.   :38.66   Min.   :  1.00
##  1st Qu.: 18208   2:33     2:37        1st Qu.:53.80   1st Qu.: 74.75
##  Median : 30271   3:27     3:26        Median :60.86   Median :126.00
##  Mean   : 36603   4:33     4:28        Mean   :61.50   Mean   :142.80
##  3rd Qu.: 52618                        3rd Qu.:69.94   3rd Qu.:223.75
##  Max.   :105633                        Max.   :85.88   Max.   :299.00
##     mileage           age
##  Min.   :  577   Min.   : 0.00
##  1st Qu.:11264   1st Qu.: 9.75
##  Median :18706   Median :20.00
##  Mean   :22613   Mean   :18.41
##  3rd Qu.:32573   3rd Qu.:27.00
##  Max.   :65394   Max.   :34.00
```

Estimating some parameters of the population

Estimating population car price and miles per gallon

The estimated population mean price for Toyota cars is 36602.62.
The estimated population variance of price for Toyota cars is 638010907.
The estimated population mean mpg for Toyota cars is 61.50352.
The estimated population variance of mpg for Toyota cars is 120.4122.

Properties of estimators :

They have minimum variance and they are unbiased estimators.

Does a linear relationship exist between mileage, miles per gallon, and the price of cars? Does a linear relationship exist between the mileage, and the price of cars?

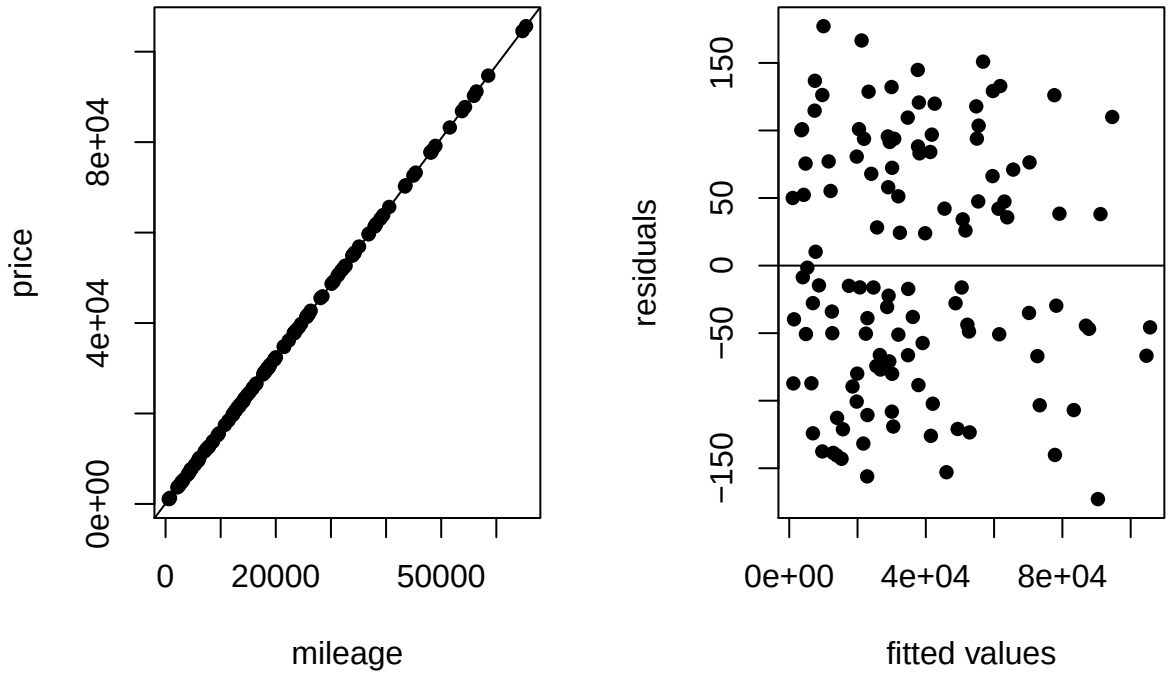
[Appendix B1]

```
## [1] 0.9999936
```

Correlation = 0.9999936. It is not meaningful because we expect that when the mileage of cars increases, the car's price is likely to decrease but let fit the model.

Assumptions

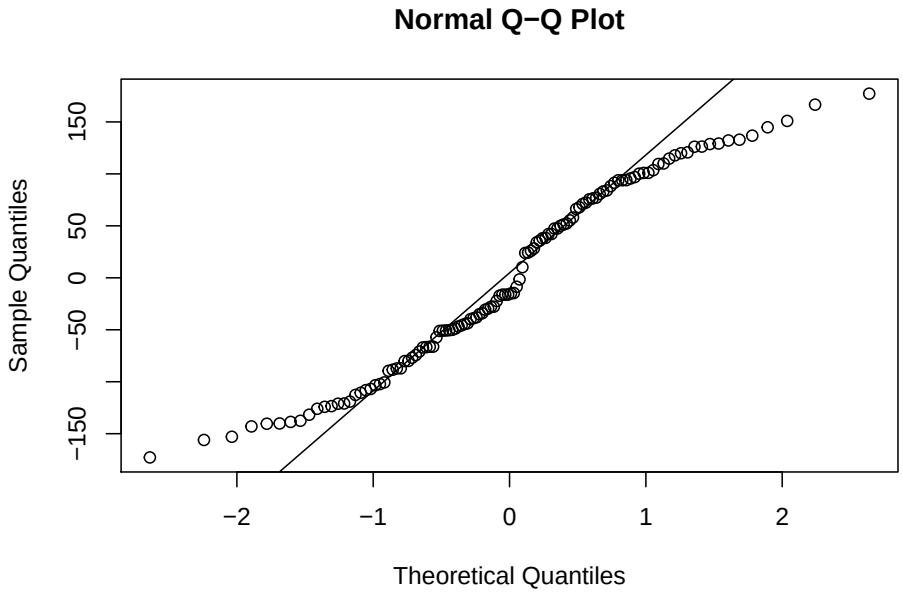
[Appendix B2][Figure 1]:



The linear relationship seems almost perfect. However, we will check the assumptions by using residuals.

The residuals bounce around the 0 line at random, indicating that the linear relationship assumption is valid. Furthermore, the plot shows no pattern, indicating that the error term has a constant variance and a 0 mean.

[Appendix B3][Figure 2]




```
## $p.value
## [1] 5.588349e-05
```

From the normality test p-value is equal to 5.588349e-05. The quantiles are not always on a straight line. This demonstrates that the residuals are not normally distributed. Hence, it is not logical to suppose that the error terms have normally distributed. Some of the assumptions are not satisfied, so we decide to try transformation.

Attempting transformation for normality

p-value of normality test [Appendix B4]:

```
## $p.value
## [1] 0.3117674
```

Result: Transformed data are normal.
[Appendix B5]

```
##              Pr(>F)
## mileage   < 2.2e-16 ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Firstly, we define the hypothesis about the significance of the model.
p-value $<$ 2.2e-16 which is smaller than α 0.05 so we rejected the null hypothesis. It can be concluded that the model is significant. Conduct a hypothesis test for testing $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ and testing $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$.

[Appendix B6]

From the summary of linear model, the p-value for the significance of the coefficients $\hat{\beta}_0$ is smaller than 0.05,then we rejected the null hypothesis and we can say that $\hat{\beta}_0$ is significant. The p-value for the significance of the coefficients $\hat{\beta}_1$ is less than 0.05 so we rejected H_0 and we can say that $\hat{\beta}_1$ is significant.

$\hat{\beta}_1 = 0.00337134$: If the mileage increases by one unit, we expect the mean price of cars to increase by 0.00337134 units.
$\hat{\beta}_0$: 68.35513489 : If mileage = 0 , the mean of the distribution of the response y = price = 68.35513489.

We calculate a 90% confidence interval for β_1, and we also calculate a 90% confidence interval for β_0

90% confidence interval for $\hat{\beta}_1$ and $\hat{\beta}_0$ is

[Appendix B7]

```
##                       5 %         95 %
## (Intercept) 65.025280475 71.684989315
## mileage      0.003250074  0.003492604
```

The confidence interval for β_0 is (65.025280475 ,71.684989315) which does not include 0 so the β_0 is significant.The confidence interval for β_1 is (0.003250074 ,0.003492604) which does not include 0 so the β_1 is significant.

The coefficient of correlation between mileage(x) and price(y):[Appendix B8]

```
## [1] 0.9733333
```

The coefficient of correlation is 0.9733333 which points out to a strict positive linear relationship.

Determination of coefficient:

According to output of summary(fit), the R^2 is 0.9474. This shows that the regression model explains about 94.74% of the total variation.

Predicting the mean price of cars when mileage is equal to 10000.[Appendix B9]

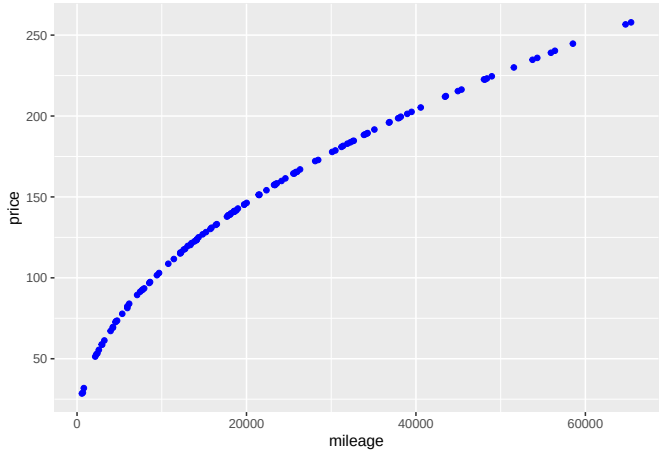
```
##        1
## 102.0685
```

Multiple linear regression part

[Appendix B10][Figure 3]

```
## The correlation coefficient between mpg and price: -0.03324994
```

```
## The correlation coefficient between mileage and price: 0.9733333
```



The linear relationship between price and mileage seems strong and positive. Moreover, according to the output from the fitted model, we can verify that by looking at the p-value of the estimated coefficient of mileage.

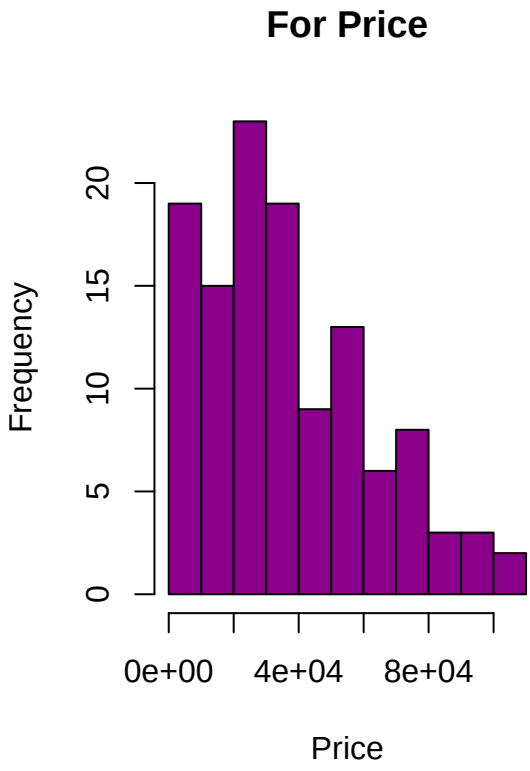
[Appendix B11][Figure 4]

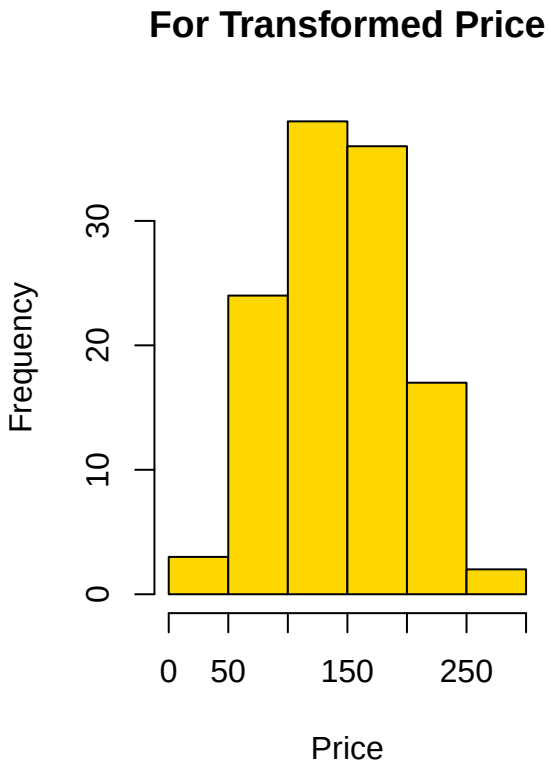
The linear relationship between price and mpg seems to be weak and negative. Moreover, according to the output from the fitted model, we can verify that by looking at the p-value of the estimated coefficient of mpg.

Diagnostic Checks

1. Checking whether the response is normally distributed or not

[Appendix B12][Figure 5]






```
## [1] "p-value for the price:"

## $p.value
## [1] 5.588349e-05

## [1] "p-value for the transformed price:"

## $p.value
## [1] 0.3117674
```

From Shapiro-Wilk test, the value of p is less than 0.05, so we rejected H_0, and it means that the response is non-normal. We could apply a transformation on price.

2. Checking whether the distribution of residuals are normal with mean 0 and constant variance.

Now we obtain residual plots:[Appendix B13][Figure 6]

The residuals do not mainly bounce around the 0 line in a random way. This indicates the assumption that the relationship is linear is not logical. The residuals do not look to form a horizontal line around the zero line. This proposes that constant variance does not exist. The plot's basic random pattern has some residuals that stand out. This proposes that there may exist outliers. From the Normal QQ plot, residuals are not mostly lined on the straight dashed line. Thus, normality seems to be not satisfied. Homoscedasticity seems to be not satisfied as the line is not straight. From Leverage plot, we can decide that there seem no influential observations. The fitted versus residual plot verifies that there seems to be a problem with this model's constant variance assumption.

3. Checking whether the multicollinearity exists.

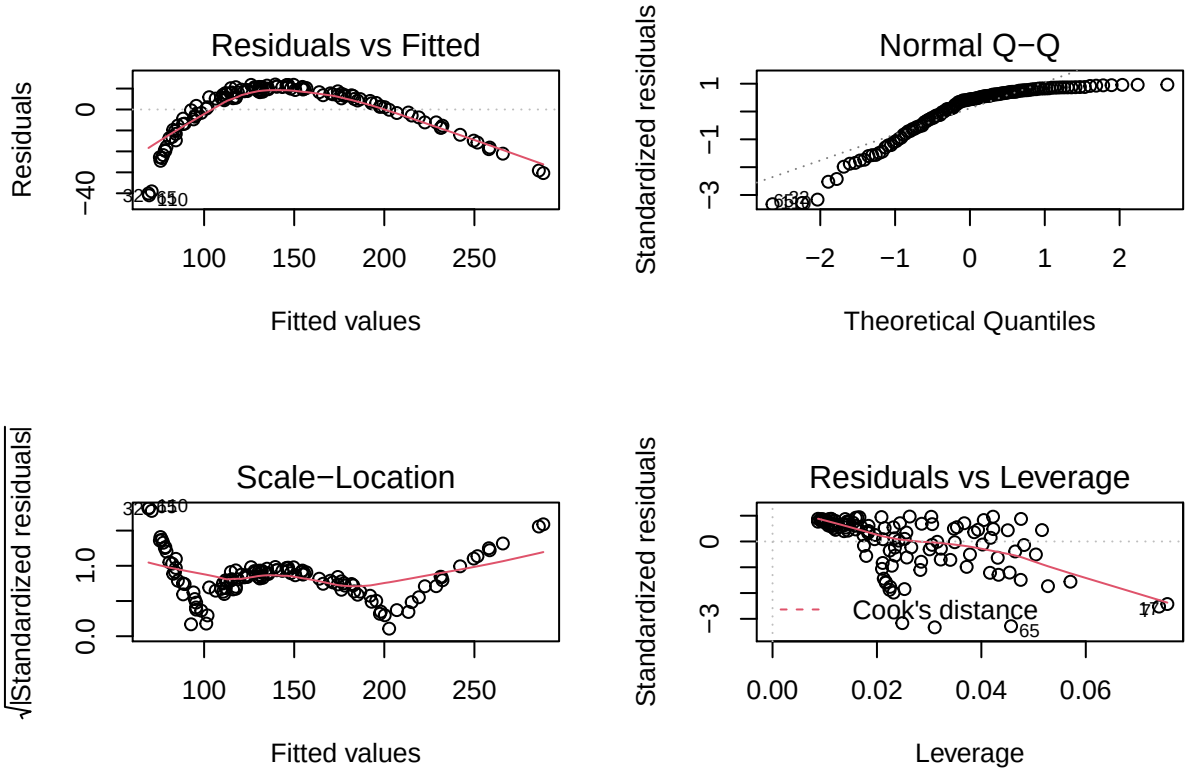
[Appendix B14]

```
##  mileage      mpg
## 1.002792 1.002792
```

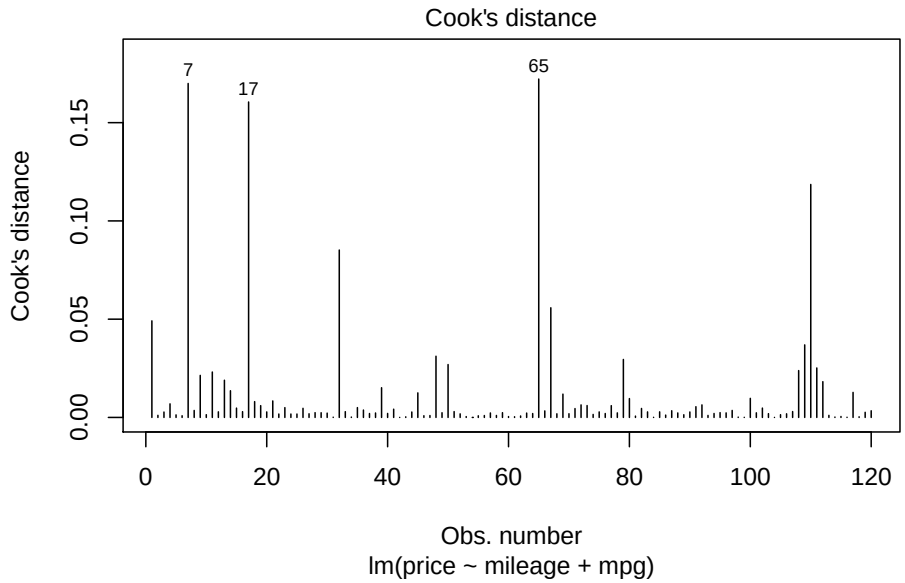
It seems we do not have such a problem since all VIF's are less than 5.[Appendix B15]

```
## $adj.r.squared
## [1] 0.9468128
```

Adjusted R-squared: 0.9468
Scale-location plot: Variances of the residuals seem to stay constant with the fitted value, proposing constant variances in the residuals errors.[Appendix B16][Figure 7]



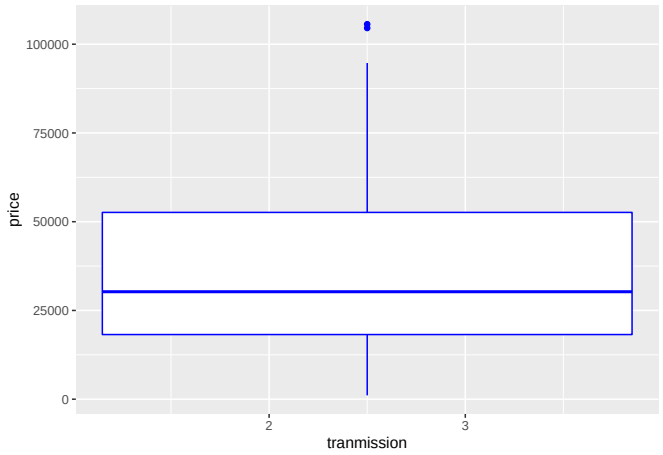
[Appendix B17][Figure 8]



However, all of these are within the 0.5 band. Thus there are no influential observations in this data.

Is the average price of cars significantly less in automatic transmission compared to manual tranmission at 1% significance level?

PARAMETRIC APPROACH
Boxplot [Appendix C1][Figure 9]:

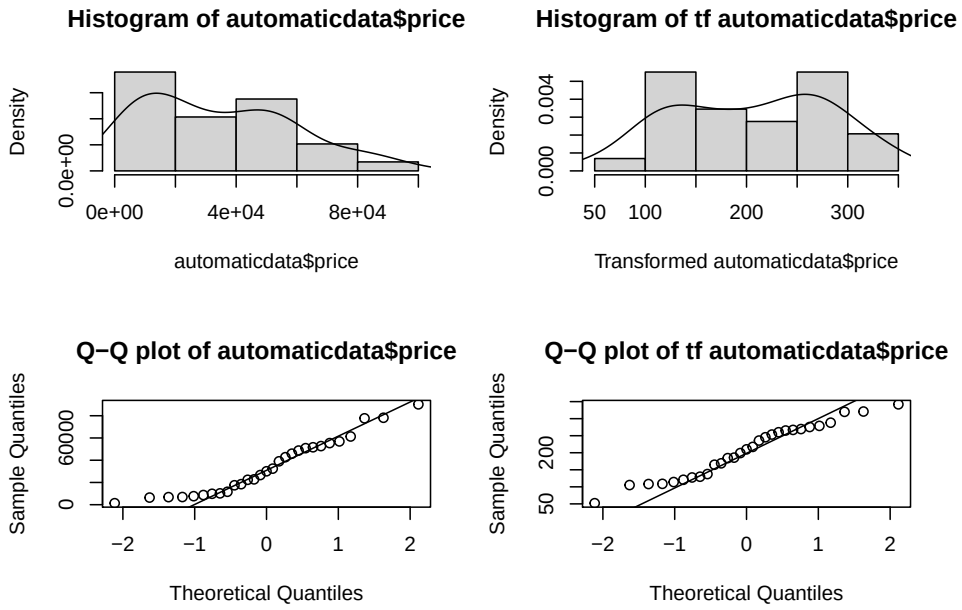


We can see that the normality assumption is not satisfied. However, we use the Shapiro-Wilk Normality test to be sure.

[Appendix C2]

```
## $p.value
## [1] 0.04476206
```

p-value = 0.04476 < 0.05 (significance level = 0.05 = alpha), data are not normally distributed. We need transformation. We can apply boxcox transformation on the data.
[Appendix C3][Figure 10]




```
## $p.value
## [1] 0.2831648
```

After transformation from normality test, p.value = 0.2831648 > 0.05. So, transformed data are normal.
[Appendix C4]

```
## $p.value
## [1] 0.05165291
```

p-value = 0.05165 > 0.05, so we assume that data are normally distributed. Since we apply transformation to the automaticdata , we apply transformation to manueldata even if it is normally distributed. [Appendix C5]

```
## $p.value
## [1] 0.6054241
```

The hypothesis is:
μ_1 : average price of automatic transmission car
μ_2 : average price of manual transmission car

$H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 < \mu_2$
[Appendix C6]

Shapiro Wilk:
Mymanuel data and myautomaticdata are normally distributed because the p-value of both of them is higher than 0.05.
[Appendix C7]

```
## $p.value
## [1] 2.836284e-10
```

p-value = 2.836e-10 from F test to compare two population variances is less than 0.05 (significance level), we can suppose the variances of two population are not equal.
[Appendix C8]

```
##
##  Welch Two Sample t-test
##
## data:  myautomaticdata and mymanueldata
## t = -9.1238, df = 42.41, p-value = 7.356e-12
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -163.901
## sample estimates:
## mean of x mean of y
##  91.35602 292.29086
```

We rejected the null hypothesis since the p-value is smaller than α=0.05. We can infer that there is a significant difference between the mean price of automatic transmission car and mean price of manual transmission car.

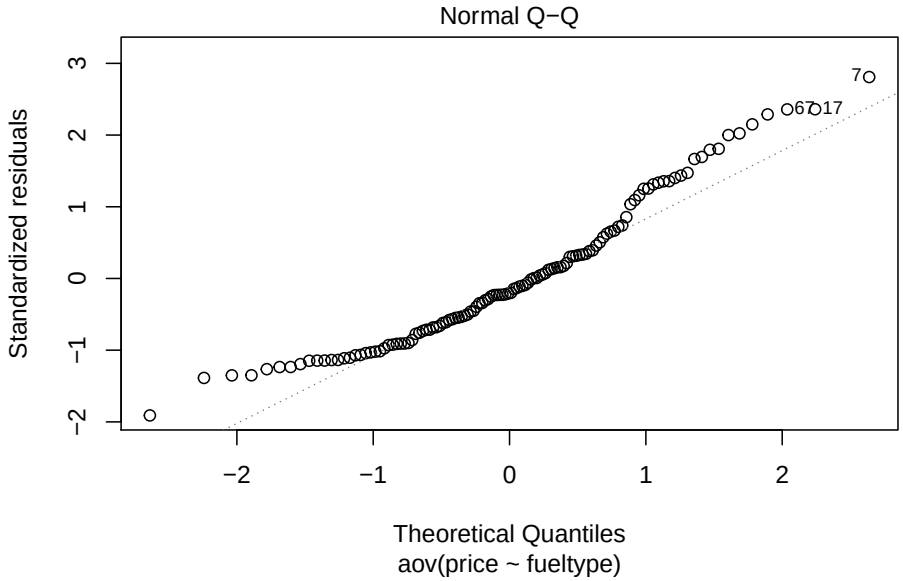
Is there a significant difference among diesel, petrol, hybrid on mean prices of cars?

Assumption Checks

Normality of The Residuals [Appendix D1]

```
## $p.value
## [1] 8.550863e-05
```

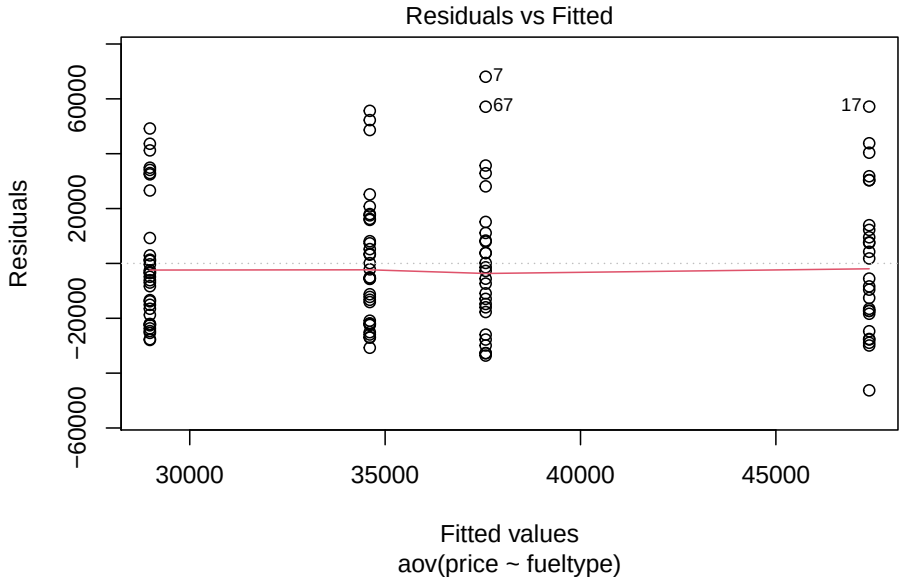
The p from Shapiro-Wilk test shows us that the distribution of residuals is not normal.
[Appendix D2][Figure 11]



The quantiles do not always lie on a straight line, indicating that residuals are not normally distributed, as shown by the QQ-plot.
[Appendix D3]

```
##
##  Shapiro-Wilk normality test
##
## data:  data6$price
## W = 0.94171, p-value = 5.588e-05
```

Homogenity Of The Variance
[Appendix D4][Figure 112]



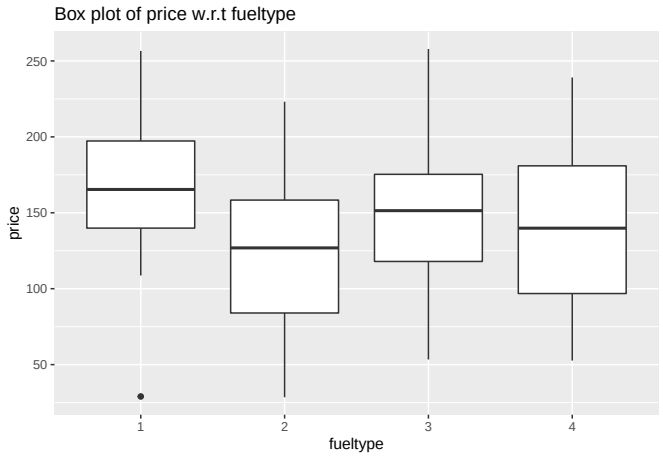
When we look at the fitted values vs. residual plot, the variances of each group are close to each other because of the similarity of the spread of observations. Hence, groups' variances are close. However, to make sure, we may use the Bartlett test, which is testing the homogeneity variance.

```
##
##  Bartlett test of homogeneity of variances
##
## data:  price by fueltype
## Bartlett's K-squared = 0.9242, df = 3, p-value = 0.8196
```

Since p value > 0.05 = alpha(significance level), homogeneity of variances are satisfied.

Transformation on price variable

[Appendix D5][Figure 13]



The means of price for diesel and hybrid are almost similar, but diesel cars' mean price is higher than the others. We can apply ANOVA to check whether there is any price difference between diesel, petrol, and hybrid or not. [Appendix D6]

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## fueltype      3  25940    8647   3.101 0.0295 *
## Residuals   116 323450    2788
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to Anova test, we rejected $H_0 : \mu_1 = \mu_2 = \mu_3 = 0$ because the p value = 0.0295 is less than 0.05. It claims that the means for some groups are different. However, we have to find which pairs are different. So, we need to conduct multiple pairwise comparisons. The hypotheses for the TUKEY test are
$H_0 : \mu_i = \mu_j$
$H_1 : \mu_i \neq \mu_j$
[Appendix D7]

```
##           diff       lwr       upr      p adj
## 2-1 -41.575960 -77.29476 -5.857159 0.01555674
## 3-1 -21.076922 -58.53912 16.385272 0.46097573
## 4-1 -25.279203 -60.99800 10.439597 0.25787403
## 3-2  20.499038 -15.21976 56.217838 0.44323243
## 4-2  16.296756 -17.58907 50.182586 0.59418645
## 4-3  -4.202282 -39.92108 31.516519 0.98994293
```

The average price of a car differs markedly between petrol and diesel, since their p-value is less than 0.05. Conditions of normality and homogeneity of variances are satisfied. In addition, we can apply a non-parametric approach to discuss the results.

Non Parametric Approach :

Kruskal Wallis Test [Appendix D8][Figure 1]

```
## $p.value
## [1] 0.04028422
```

Since p-value = 0.04028 < significance level = 0.05, we can infer that there are significant differences among diesel, petrol, hybrid on mean prices of cars.

From the Tukey multiple comparisons of means table, The average price of a car differs markedly between petrol and diesel. since their p-value is less than 0.05.
[Appendix D9]

```
##
##  Pairwise comparisons using Wilcoxon rank sum exact test
##
## data:  data6$price and data6$fueltype
##
##   1     2     3
## 2 0.037 -     -
## 3 0.268 0.268 -
## 4 0.153 0.288 0.791
##
## P value adjustment method: BH
```

Wilcoxpairwise test shows that petrol and diesel on mean prices of cars are significantly different because p-value = 0.037 < 0.05.

Is the average age of cars significantly less than 20 at 5% significance level?

Significance level is 5 percent. (alpha = 0.05)
$H_0 : \mu_0 = 20$ (mean age of cars is 20)
$H_1 : \mu_0 < 20$(mean age of cars is less than 20)

From normality test: [Appendix E1]

```
## $p.value
## [1] 9.564669e-05
```

Age variable is not normally distributed because p value of the Shapiro-Wilk test is less than α=0.05. Therefore, we reject the null hypothesis. However, the sample size of agedata is sufficiently large enough so we can claim that the sample mean's distribution is normal by CLT. In this reason, we use t.test statistic because sd of the population is not known. Agedata is not normally distributed but since we have sample size = 60 > 30 , we can say that the sample mean's distribution is approximately normal.
[Appendix E2]

```
## $p.value
## [1] 0.0425555
```

Conclusion : From the t-test for one sample, since p-value = 0.04256 < 0.05 , we rejected the null hypothesis. We can infer that the average age of cars is significantly less than 20 at a 5 % significance level.
[Appendix E3]

```
## Point Estimation : 18.40833
```

```
## [1] 0.9167364
```

```
## [1] 1.9801
```

```
## Interval Estimation : 16.5931 , 20.22356
```

True population mean inside the 95% confidence interval.

Do the data provide enough evidence to determine that age and price are associated at a 5% significance level?

H_0 : Rows and columns of this table are independent. H_1 : They are not independent

[Appendix F1]

```
##              Age <18 Age >18
## Price <36000      26      42
## Price >36000      23      29
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  contingecy_table
## X-squared = 0.22537, df = 1, p-value = 0.635
```

```
##          Price <36000 Price >36000
## Age <18    0.5306122    0.4693878
## Age >18    0.5915493    0.4084507
```

χ^2 critical value = 3.841. We found that X-squared test = 0.22537
χ^2 test = 0.22537 < χ^2 critical value = 3.841 , so we fail to reject H0.
There is no enough evidence to claim that age and price are associated at a 5% significance level.

Cramer's V:[Appendix F2]

```
## Cramer V
##  0.06044
```

Since Cramer's V = 0.06044 is very close to 0 , there is very weak association between ages and prices of cars.

Conclusion

For first question, we analyze that if there is a relationship between mileage and the price of cars with a simple linear regression model. There is a positive correlation which is 0.999, between mileage and the price of a car. Also, assumptions Figure1 proved this relationship. For normality, we check the QQ-plot and Shapiro-Wilk test. The p-value of the Shapiro-Wilk test is too small, so we can not say that residuals are normally distributed. For this reason, we have to transform the data. By multiplying residual by 0.48, we made the data normal. We found that the model was significant by performing a hypothesis test. As a result, our model became y=68.35513+0.00337x_1. When we created a 90% confidence interval for β_0 and β_1 in simple linear regression, we concluded that the interval of β_0 was (65.025280475,71.684989315) and that of β_1 was (0.003250074,0.003492604). We found that β_0 and β_1 were significant. And then, we analyze that if there is a relationship between mileage, miles per gallon, and the price of cars with a multiple linear regression model. There is a strong and positive relationship between mileage and price from Figure 3, but there is a weak relationship between mpg and price from Figure 4. The p-value of the Shapiro-Wilk test is too small, so we can not say that residuals are normally distributed. We have to transform the data. By multiplying residual by 0.48, we made the data normal. We interpreted our residual Figure 6. Also, there is no multicollinearity problem. Adjusted R-squared, which is 0.8704, looks good. Mileage and price are statistically significant because their p-values are all less than 0.05. But mpg is not statistically significant because its p-value is less than 0.05. For the 2nd question, we did both parametric and nonparametric tests to find whether the average price of cars significantly less in automatic transmission than manuel transmission or not. According to the parametric approach, since transmission is not normally distributed, we did boxcox transformation to our data, and then it distributed normally. From the hypothesis test results, we can say that the average price of cars for automatic is less than for manuel. For the 3rd question, we examined whether there is a significant difference among diesel, petrol, hybrid on the mean prices of cars or not. Firstly, we checked some assumptions. The result of assumptions says that there is no normality, so we did transformation on the price variable and then did an ANOVA test. According to the ANOVA test, some of the group means are different, but we conducted multiple pairwise comparisons to determine which one is different. Consequently, there is a significant difference between petrol and diesel in the mean prices of cars. And we also did a nonparametric test to be sure. Wilcoxpairwise test indicates the same result of the parametric test. From the one-sample t-test, we found that p-value = 0.04256, which is less than 0.05 significance level. Therefore, we rejected the null hypothesis and said that we have sufficient evidence to conclude that the mean age of cars is significantly less than 20 at 5 % significance level. Moreover, we found that the point estimation inside the 95% confidence interval.Since we found that χ^2 test = 0.22537 < χ^2 critical value = 3.841 and p-value is higher than 0.05 (significance level), we could not reject the null hypothesis . As a result, we can conclude that we do not have sufficient evidence to say that specified age and price groups are associated at a 5% significance level. Also, we found that Cramer's V = 0.06044, close to 0, so; we concluded that the association between specified age and price groups is very weak.

References

F-Test: Compare Two Variances in R - Easy Guides - Wiki - STHDA. (n.d.). STHDA. Retrieved June 25, 2021, from http://www.sthda.com/english/wiki/f-test-compare-two-variances-in-r

Non Parametric Data and Tests (Distribution Free Tests). (2021, May 31). Statistics How To. https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/parametric-and-non-parametric-data/

Parametric and Non-parametric tests for comparing two or more groups. (2017, January 8). Health Knowledge. https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1b-statistical-methods/parametric-nonparametric-tests

Unpaired Two-Samples Wilcoxon Test in R - Easy Guides - Wiki - STHDA. (n.d.). STHDA. Retrieved June 25, 2021, from http://www.sthda.com/english/wiki/unpaired-two-samples-wilcoxon-test-in-r