Exam Template: Statistical Inference

21050352

Jan 2022: Sep21 run

# Instructions to students

You should only use the file Exam\_template.Rmd provided on blackboard and you should load this file from your scripts folder / directory.

Save this template as your studentID.Rmd; you will upload this file as your submission. Change the information on line 3 of this file – changing the author information to your **student ID**. Do not change the authorship to your name.

Ensure that you save your data into your data folder (as discussed in class). You may use the files mypackages.R and helperFunctions.R from blackboard. If you use these files, do not alter them. If you wish to create additional files for custom functions that you have prepared in advance, make sure that you upload these in addition to your .Rmd file and your compiled output file.

Your should knit this file to a document **Word** format.

Any changes that you make to the data (e.g. variable name changes) should be made entirely within R.

The subsubsections labelled **Answer:** indicate where you should put in your written Answers. The template also provides blank code chunks for you to complete your Answers; you may choose to add additional chunks if required.

# load dataset  
cars <- read\_csv(  
 "../data/Jan\_2022\_Exam\_Data.csv")

Rows: 52533 Columns: 10

-- Column specification --------------------------------------------------------  
Delimiter: ","  
chr (4): brand, model, transmission, fuelType  
dbl (6): year, price, mileage, tax, mpg, engineSize

i Use `spec()` to retrieve the full column specification for this data.  
i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

# Data description

This dataset is part of a larger dataset that has been collected to help to estimate the price of used cars.

It contains the following variables:

* brand (manufacturer)
* model (of car)
* year (of registration of the car)
* price (in GB pounds)
* transmission (type of gearbox)
* mileage (total distance covered by the car)
* fuelType (type of fuel used by the car)
* tax (annual cost of vehicle tax)
* mpg (miles per gallon - a measure of fuel efficiency)
* engineSize (size of the engine in litres)

# Question 1: Data Preparation (11 marks)

You are interested in modelling the price of vehicles that have all of the following properties:

* mileage less than 60000
* Manual transmission
* Petrol engine (fuelType)
* Costing less than £200 in annual Vehicle Tax.

Once you have selected the rows of data with these properties, then you must *use your studentID* to select a random sample of 2000 rows of the data to perform the rest of your analysis with.

You should remove any redundant variables (where only one value remains in that variable).

This subset of the data is what you should use for the rest of this assessment.

1. Explain what data preparation is required in order for the data in Jan\_2022\_Exam\_Data.csv to be suitable for this analysis.

**(4 marks)**

### Answer:

First of all, we need to change the data types of different variables as required. for example, the brand is a factor instead of a char data type. Similarly, the transmission is a factor type instead of a char type since we don’t wan’t to carry out statistical operation there. Engine size seems like a factor, however we are interested in average, median engine size data and its relationship with price of the car. Secondly, I create a R object which satisfies the conditions for which we are modeling (mileage less than 60000, Manual transmission, Petrol engine (fuelType), Costing less than £200 in annual Vehicle Tax.) I select only those rows from “cars” data into the new R data, say “cars2”. Thirdly, I remove any redundant columns, if it exists. Finally, I set seed using my student ID and select a sample of 2000 data for the analysis. Additionally, reading the missing data, there is no NA observations with codes to be later changed into NA.

1. Implement the required data preparation in the code chunk below:

**(7 marks)**

### Answer:

#changing datatypes to make data suitable for analysis in R  
cars$brand = as.factor(cars$brand)  
cars$model = as.factor(cars$model)  
cars$year = as.factor(cars$year)  
cars$transmission = as.factor(cars$transmission)  
cars$fuelType = as.factor(cars$fuelType)  
  
#creating a subset meeting the conditions  
cars2<- subset(cars, mileage< 60000 & transmission == "Manual" & fuelType == "Petrol" & tax<200)  
  
#creating a dataframe, just in case  
cars2<- data.frame(cars2)  
  
#final set with removed redundant columns  
cars2 <- subset(cars2, select = -c(transmission, fuelType))  
  
#setting seed  
  
set.seed(21050352)  
  
index <- sample(seq\_along(cars2$price), 2000)  
  
  
sample\_cars<- cars2[index,]  
  
#so my data is prepared.

# Question 2: Exploratory Data Analysis (22 marks)

## Descriptive Statistics

1. What descriptive statistics would be appropriate for this dataset? Explain why these are useful in this context.

**(2 marks)**

### Answer:

Since I am interested in the price of the car, the statistics I would be interested is the average value of cars that satisfies my conditions, compare with the median and also have the maximum price of the car and the minimum price within the conditions. This will help me make an informed decision if i wanted to purchase a car. The other information needed for decision making like, the average mpg i get within my budget, and the average amount of tax I need to pay annually. Additionally, I also get idea about the brand of cars (for some people who are brand savvy), are in the pool and average statistics for each brand. An Anova test might be required to decide if the difference in means of price between brands is significant for later analyses.

1. Produce those descriptive statistics in the code chunk below:

**(4 marks)**

### Answer:

#the function describe does it all for us.  
print(describe(sample\_cars))

vars n mean sd median trimmed mad min max  
brand\* 1 2000 2.67 0.77 3.0 2.80 0.00 1.0 4.0  
model\* 2 2000 33.37 13.74 35.0 33.13 1.48 1.0 92.0  
year\* 3 2000 23.17 1.68 23.0 23.34 1.48 7.0 26.0  
price 4 2000 12805.00 4960.45 11500.0 12245.07 4097.91 1294.0 41000.0  
mileage 5 2000 18360.20 12793.82 16236.0 17082.59 12131.37 1.0 59750.0  
tax 6 2000 115.05 54.37 145.0 124.68 0.00 0.0 165.0  
mpg 7 2000 56.01 8.07 57.6 56.78 7.86 20.9 85.6  
engineSize 8 2000 1.20 0.34 1.0 1.15 0.00 0.0 5.0  
 range skew kurtosis se  
brand\* 3.0 -1.37 0.73 0.02  
model\* 91.0 0.39 2.54 0.31  
year\* 19.0 -1.39 5.64 0.04  
price 39706.0 1.18 1.83 110.92  
mileage 59749.0 0.84 0.29 286.08  
tax 165.0 -1.33 -0.09 1.22  
mpg 64.7 -0.71 0.49 0.18  
engineSize 5.0 2.98 26.62 0.01

#this describeby function is for getting stats by each group which is brand in our case for knowing  
#the average price in each brand  
print(describeBy(sample\_cars, group="brand"))

Descriptive statistics by group   
brand: Audi  
 vars n mean sd median trimmed mad min max  
brand\* 1 311 1.00 0.00 1.0 1.00 0.00 1.0 1.0  
model\* 2 311 29.46 23.25 16.0 26.06 2.97 14.0 89.0  
year\* 3 311 23.13 1.56 23.0 23.18 1.48 18.0 26.0  
price 4 311 17339.70 5323.22 16500.0 16879.78 5203.93 8490.0 36950.0  
mileage 5 311 19318.54 13477.23 18419.0 18073.84 14098.04 10.0 58161.0  
tax 6 311 104.08 58.78 145.0 110.64 7.41 0.0 165.0  
mpg 7 311 52.76 8.15 52.3 52.63 7.86 34.9 70.6  
engineSize 8 311 1.41 0.29 1.4 1.39 0.00 0.0 2.0  
 range skew kurtosis se  
brand\* 0.0 NaN NaN 0.00  
model\* 75.0 1.13 -0.37 1.32  
year\* 8.0 -0.22 -0.06 0.09  
price 28460.0 0.73 0.12 301.85  
mileage 58151.0 0.69 -0.10 764.22  
tax 165.0 -0.76 -1.30 3.33  
mpg 35.7 0.02 -0.56 0.46  
engineSize 2.0 0.11 1.51 0.02  
------------------------------------------------------------   
brand: BMW  
 vars n mean sd median trimmed mad min max  
brand\* 1 102 2.00 0.00 2.0 2.00 0.00 2.0 2.0  
model\* 2 102 9.31 23.27 3.0 2.73 2.97 1.0 92.0  
year\* 3 102 23.27 2.43 24.0 23.60 1.48 7.0 26.0  
price 4 102 17545.31 5503.06 15986.0 17270.85 5931.88 8390.0 38776.0  
mileage 5 102 16471.12 14320.32 12570.5 14918.29 14825.26 10.0 55825.0  
tax 6 102 137.21 28.32 145.0 143.41 0.00 30.0 165.0  
mpg 7 102 49.08 5.47 52.3 49.60 2.82 32.1 57.6  
engineSize 8 102 1.62 0.36 1.5 1.55 0.00 0.0 3.0  
 range skew kurtosis se  
brand\* 0.0 NaN NaN 0.00  
model\* 91.0 3.13 8.11 2.30  
year\* 19.0 -3.20 17.72 0.24  
price 30386.0 0.70 0.66 544.88  
mileage 55815.0 0.73 -0.35 1417.92  
tax 135.0 -3.14 9.12 2.80  
mpg 25.5 -0.89 -0.04 0.54  
engineSize 3.0 1.43 9.13 0.04  
------------------------------------------------------------   
brand: Ford  
 vars n mean sd median trimmed mad min max  
brand\* 1 1519 3.00 0.00 3.0 3.00 0.00 3.0 3.0  
model\* 2 1519 36.38 6.12 35.0 35.37 0.00 22.0 75.0  
year\* 3 1519 23.14 1.64 23.0 23.33 1.48 14.0 26.0  
price 4 1519 11313.86 3822.50 10500.0 10930.55 2972.61 1294.0 41000.0  
mileage 5 1519 18390.08 12463.22 16044.0 17124.41 11288.52 1.0 59750.0  
tax 6 1519 114.66 55.16 145.0 124.31 0.00 0.0 165.0  
mpg 7 1519 57.58 7.46 58.9 58.48 7.86 20.9 85.6  
engineSize 8 1519 1.11 0.30 1.0 1.06 0.00 0.0 5.0  
 range skew kurtosis se  
brand\* 0.0 NaN NaN 0.00  
model\* 53.0 1.78 5.41 0.16  
year\* 12.0 -1.21 2.25 0.04  
price 39706.0 1.37 4.10 98.08  
mileage 59749.0 0.90 0.45 319.78  
tax 165.0 -1.33 -0.11 1.42  
mpg 64.7 -0.96 1.31 0.19  
engineSize 5.0 5.14 59.95 0.01  
------------------------------------------------------------   
brand: Mercedes  
 vars n mean sd median trimmed mad min max  
brand\* 1 68 4.00 0.00 4.0 4.00 0.00 4.0 4.0  
model\* 2 68 20.13 11.46 13.0 17.96 0.00 13.0 81.0  
year\* 3 68 23.75 1.55 24.0 23.86 1.48 19.0 26.0  
price 4 68 18264.31 4001.55 18250.0 18225.91 5077.90 10662.0 25890.0  
mileage 5 68 16143.28 14143.00 12811.0 14517.62 12909.00 9.0 57119.0  
tax 6 68 140.81 8.96 145.0 141.52 0.00 125.0 150.0  
mpg 7 68 46.30 7.51 48.7 47.51 4.00 28.5 53.3  
engineSize 8 68 1.52 0.19 1.6 1.51 0.00 1.3 2.0  
 range skew kurtosis se  
brand\* 0.0 NaN NaN 0.00  
model\* 68.0 2.60 9.72 1.39  
year\* 7.0 -0.70 -0.03 0.19  
price 15228.0 0.02 -1.03 485.26  
mileage 57110.0 0.95 0.11 1715.09  
tax 25.0 -1.11 -0.58 1.09  
mpg 24.8 -1.62 1.29 0.91  
engineSize 0.7 0.47 0.33 0.02

t1<-t.test(sample\_cars$price, sample\_cars$mileage)  
print(t1)

Welch Two Sample t-test  
  
data: sample\_cars$price and sample\_cars$mileage  
t = -18.105, df = 2586.7, p-value < 2.2e-16  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 -6156.856 -4953.546  
sample estimates:  
mean of x mean of y   
 12805.0 18360.2

t2<-t.test(sample\_cars$price,sample\_cars$tax)  
print(t2)

Welch Two Sample t-test  
  
data: sample\_cars$price and sample\_cars$tax  
t = 114.4, df = 1999.5, p-value < 2.2e-16  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 12472.40 12907.48  
sample estimates:  
 mean of x mean of y   
12804.9950 115.0525

t3<-t.test(sample\_cars$price,sample\_cars$mpg)  
print(t3)

Welch Two Sample t-test  
  
data: sample\_cars$price and sample\_cars$mpg  
t = 114.94, df = 1999, p-value < 2.2e-16  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 12531.45 12966.51  
sample estimates:  
 mean of x mean of y   
12804.9950 56.0148

t4<-t.test(sample\_cars$price,sample\_cars$engineSize)  
print(t4)

Welch Two Sample t-test  
  
data: sample\_cars$price and sample\_cars$engineSize  
t = 115.43, df = 1999, p-value < 2.2e-16  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 12586.27 13021.32  
sample estimates:  
 mean of x mean of y   
12804.99500 1.19935

sample\_year<- subset(sample\_cars, year== "2018" | year == "2019")  
print(describe(sample\_year))

vars n mean sd median trimmed mad min max  
brand\* 1 876 2.77 0.68 3.0 2.91 0.00 1.0 4.0  
model\* 2 876 35.15 13.54 35.0 35.28 1.48 1.0 92.0  
year\* 3 876 24.43 0.50 24.0 24.41 0.00 24.0 25.0  
price 4 876 15058.41 4844.71 13998.5 14517.66 4449.28 7397.0 38776.0  
mileage 5 876 10998.95 7895.09 9350.5 10132.65 7014.92 5.0 49015.0  
tax 6 876 145.76 1.80 145.0 145.33 0.00 145.0 150.0  
mpg 7 876 54.63 8.90 57.7 55.46 9.34 20.9 67.3  
engineSize 8 876 1.18 0.36 1.0 1.12 0.00 0.0 5.0  
 range skew kurtosis se  
brand\* 3.0 -1.66 2.27 0.02  
model\* 91.0 0.26 2.97 0.46  
year\* 1.0 0.29 -1.92 0.02  
price 31379.0 1.05 1.13 163.69  
mileage 49010.0 1.20 1.96 266.75  
tax 5.0 1.92 1.71 0.06  
mpg 46.4 -0.81 0.07 0.30  
engineSize 5.0 3.56 29.96 0.01

1. What have those descriptive statistics told you – and how does this inform the analysis that you would undertake on this data or any additional data cleaning requirements?

**(4 marks)**

### Answer:

The first observation is that the average price of cars is approx 12800. However, when the descriptive statistics were computed for each group, it is observed that only ford cars (avg price = 11300) are below the average, and all other brands are 3000 more expensive than the average. This means the ford cars skew the mean by their sheer number of observations (n=1519). One decision can be made by this observation is to include only ford cars in later analysis. They also have low average tax and high mileage. The standard deviation is 4960.5 which is significant, this means the data is highly varied. We can also look at the min and max value and observe that there are significant outliers in the data. These outliers may affect later observations hence we need to remove them.

Also, the t-tests show that all the variables have statistically significant relationship with the price and hence we cannot ignore them.

#min\_acceptable = Q1-1.5\*range, max\_acceptable = Q3+1.5\*range  
#since we already have data from describeby.  
Q <- quantile(sample\_cars$price, probs=c(.25, .75), na.rm = FALSE)  
iqr= IQR(sample\_cars$price)  
min\_acc = Q[1]-1.5\*iqr  
max\_acc = Q[2] + 1.5\*iqr  
  
sample<- subset(sample\_cars, price<max\_acc & price>min\_acc)

Henceforth, I will use this data for further analysis.

## Exploratory Graphs

1. What exploratory graphs would be appropriate for this data set? Explain why these are useful in this context.

**(2 marks)**

### Answer:

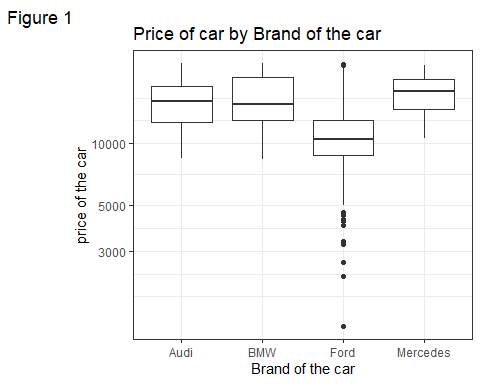
Box plots would be appropriate fot this data set as we can put mutiple observations based on car brands in a single plot and study. I would also carry out scatterpots of price against other variables. Scatter plots will give some idea about the distribution of price against other variables. I can give idea about correlation and linearity of the relations which is useful for analysis.

1. Now produce those exploratory graphs in the code chunk below:

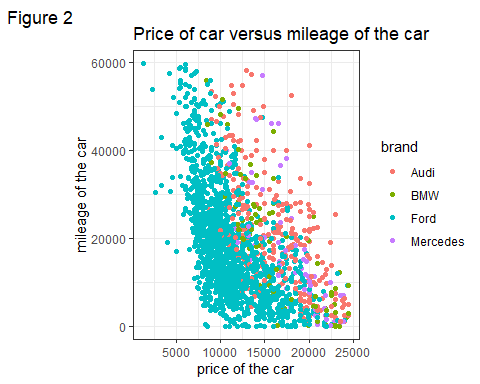
**(4 marks)**

### Answer:

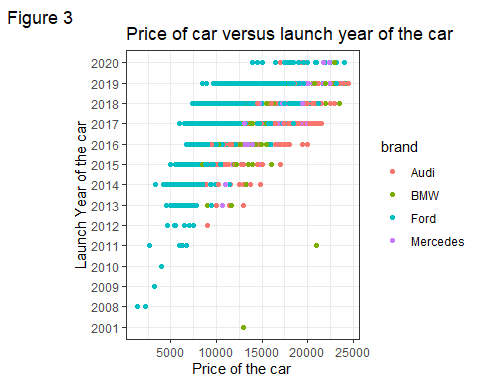
sample %>%  
 ggplot(aes(x=brand, y=price)) +   
 geom\_boxplot() +   
 scale\_y\_log10() +  
 labs(title = "Price of car by Brand of the car",  
 tag = "Figure 1") +   
 xlab("Brand of the car") +   
 ylab("price of the car") +   
 theme\_bw()



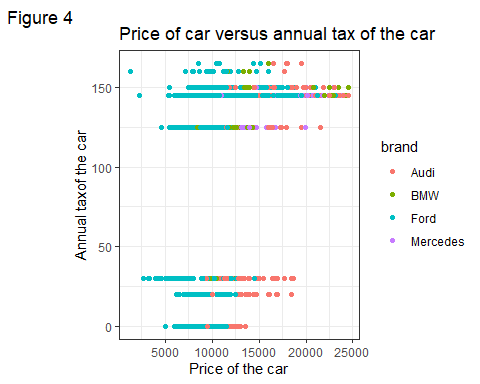
sample%>%  
 ggplot(aes(x=price, y=mileage, colour = brand))+geom\_point() +  
 labs(title = "Price of car versus mileage of the car",  
 tag = "Figure 2") +   
 xlab("price of the car") +   
 ylab("mileage of the car") +   
 theme\_bw()



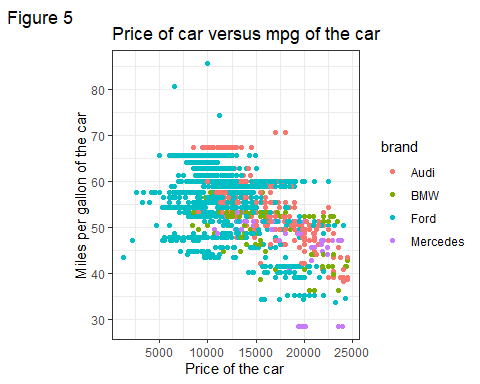
sample%>%  
 ggplot(aes(x=price, y=year, colour = brand))+geom\_point() +  
 labs(title = "Price of car versus launch year of the car",  
 tag = "Figure 3") +   
 xlab("Price of the car") +   
 ylab("Launch Year of the car") +   
 theme\_bw()



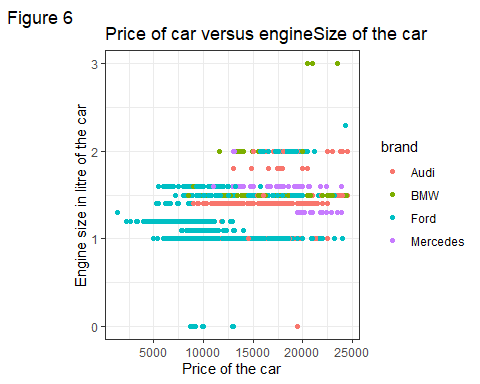
sample%>%  
 ggplot(aes(x=price, y=tax, colour = brand))+geom\_point() +  
 labs(title = "Price of car versus annual tax of the car",  
 tag = "Figure 4") +   
 xlab("Price of the car") +   
 ylab("Annual taxof the car") +   
 theme\_bw()



sample%>%  
 ggplot(aes(x=price, y=mpg, colour = brand))+geom\_point() +  
 labs(title = "Price of car versus mpg of the car",  
 tag = "Figure 5") +   
 xlab("Price of the car") +   
 ylab("Miles per gallon of the car") +   
 theme\_bw()



sample%>%  
 ggplot(aes(x=price, y=engineSize, colour = brand))+geom\_point() +  
 labs(title = "Price of car versus engineSize of the car",  
 tag = "Figure 6") +   
 xlab("Price of the car") +   
 ylab("Engine size in litre of the car") +   
 theme\_bw()



1. Interpret these exploratory graphs. How do these graphs inform your subsequent analysis?

**(4 marks)**

### Answer:

The box plots say that the Audi, BMW, Mercedes data seem to be more normally distributed, and the ford data has many outliers. We might want to remove the outliers in this data for better analysis results. The scatterplots show the linearity between price and other variables. We can observe that the price is negatively related to mileage and mpg, while there seems to be minimal relation with engine size and tax (which has a weird jump). We can also see that as the year increases, the price goes up albeit it isn’t linear. The ford cars have lowest of the prices and it tends to decrease with the mileage. The annual tax has a significant jump. However, the almost horizontal distribution means, it does not affect the price of car significantly. Same is the case of egnine size, which seems to have almost no effect in the price of the car.

## Correlations

1. What linear correlations are present within this data?

**(2 marks)**

### Answer:

I can observe that the linear correlation of price is highest with mpg followed by mileage. however, they are negative correlation, meaning price increases with decrease in mileage and mpg. Similarly, price tend to increase with tax and engine size. However, the correlation isn’t strong in either cases. These information tells me to conduct our inferential analysis of price in mpg and mileage.

cor\_data <- subset(sample, select= c(price, mileage,tax, mpg, engineSize))  
print(cor(cor\_data))

price mileage tax mpg engineSize  
price 1.0000000 -0.52137838 0.4009465 -0.4906590 0.38867294  
mileage -0.5213784 1.00000000 -0.3982898 0.1178890 0.08396627  
tax 0.4009465 -0.39828982 1.0000000 -0.3825407 0.18088613  
mpg -0.4906590 0.11788900 -0.3825407 1.0000000 -0.65399195  
engineSize 0.3886729 0.08396627 0.1808861 -0.6539920 1.00000000

# Question 3: Bivariate relationship (14 marks)

1. Which of the potential explanatory variables has the strongest linear relationship with the dependent variable?

**(1 mark)**

### Answer:

Mileage has the strongest linear relationship with the dependent variable, which is price. This is shown in both the scatter plots and the correlation matrix.

1. Create a linear model to model this relationship.

**(2 marks)**

### Answer:

model1 <- lm(price~mileage, data= sample)  
print(model1)

Call:  
lm(formula = price ~ mileage, data = sample)  
  
Coefficients:  
(Intercept) mileage   
 15593.4734 -0.1731

1. Explain and interpret the model:

**(3 marks)**

### Answer:

The model is

This model is intrepreted as a linear equation which is price = 15593.51- 0.1731 \* mileage. For example, if a car’s mpg is 1 mpg, it’s price would be 15593.51 - 0.17 which is 15593.34. The intercept is the value “a” and the mpg is the value “b” in the linear equation model x = a+b\*y, where x is the price and y is the mpg.

1. Comment on the performance of this model, including comments on overall model fit and the validity of model assumptions. Include any additional code required for you to make these comments in the code chunk below.

**(4 marks)**

### Answer:

This model can only be used as a estimate on where the price may be in. However, it is not the exact value because the model is not in very strong linear correlation. Another flaw in this model is that, if we put 0 mpg in the model, it gives 15539.3 for the price is car, which is absolutely not optimal. Who would pay such money for a car that just doesn’t run? (maybe collectors?).

Also, checking the model below with graphs, we can say that the linearity assumption (i.e there is linear relationship between price and mpg) is valid. The homogeneneity of variance graph shows why. we have significant outliers in or dataset.

We can also observe that there are few influential observations that can skew our model and the residues fall within the normality line. However, the distribution seem to follow the normal distribution (albeit inperfectly).

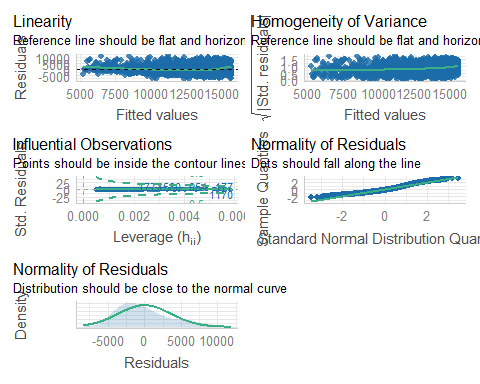
This model shows the price of car for a close estimate given the mileage.

The second strongest indepedent variable from checking the corelation of residuals is the engine size.

print(model1)

Call:  
lm(formula = price ~ mileage, data = sample)  
  
Coefficients:  
(Intercept) mileage   
 15593.4734 -0.1731

check\_model(model1)



cor\_data$residuals<- residuals(model1)  
cor(cor\_data)

price mileage tax mpg engineSize  
price 1.0000000 -5.213784e-01 0.4009465 -0.4906590 0.38867294  
mileage -0.5213784 1.000000e+00 -0.3982898 0.1178890 0.08396627  
tax 0.4009465 -3.982898e-01 1.0000000 -0.3825407 0.18088613  
mpg -0.4906590 1.178890e-01 -0.3825407 1.0000000 -0.65399195  
engineSize 0.3886729 8.396627e-02 0.1808861 -0.6539920 1.00000000  
residuals 0.8533256 -4.375577e-18 0.2265101 -0.5029665 0.50678327  
 residuals  
price 8.533256e-01  
mileage -4.375577e-18  
tax 2.265101e-01  
mpg -5.029665e-01  
engineSize 5.067833e-01  
residuals 1.000000e+00

## Bootstrap

1. Use bootstrapping on this model to obtain a 95% confidence interval of the estimate of the slope parameter.

**(4 marks)**

### Answer:

Nbootstrap<- 200   
  
slope\_model<-rep(NA,Nbootstrap)  
  
set.seed(21050352)  
  
for(i in seq\_len(Nbootstrap)){  
 index<-sample(seq\_along(sample$price),size=length(sample$price),replace=TRUE)   
   
 bootstrap.sample<-sample[index,]  
   
 slope\_model[i]<- confint(lm(price~mileage, data= bootstrap.sample), 'mileage', 0.95 )  
 }  
  
CI<- quantile(slope\_model,probs=c(.025,0.975))  
c(print("the 95% confidence interval for means is:"), print(CI))

[1] "the 95% confidence interval for means is:"  
 2.5% 97.5%   
-0.2001807 -0.1729691

"the 95% confidence interval for means is:"   
 2.5%   
 "-0.200180690335917"   
 97.5%   
 "-0.172969145178164"

# Question 4: Multivariable relationship (10 marks)

Create a model with all of the appropriate remaining explanatory variables included:

modelall<- lm(price~mileage+engineSize+mpg+tax, data= sample)  
summary(modelall)

Call:  
lm(formula = price ~ mileage + engineSize + mpg + tax, data = sample)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-8324.8 -1796.2 -69.2 1866.0 8701.9   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1.753e+04 1.059e+03 16.560 < 2e-16 \*\*\*  
mileage -1.622e-01 5.937e-03 -27.320 < 2e-16 \*\*\*  
engineSize 3.933e+03 3.240e+02 12.141 < 2e-16 \*\*\*  
mpg -1.306e+02 1.276e+01 -10.235 < 2e-16 \*\*\*  
tax 5.317e+00 1.450e+00 3.667 0.000252 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 2991 on 1934 degrees of freedom  
Multiple R-squared: 0.4998, Adjusted R-squared: 0.4988   
F-statistic: 483.2 on 4 and 1934 DF, p-value: < 2.2e-16

1. Explain and interpret the model:

**(4 marks)**

### Answer:

The model in equation form is

The 0.49 r squareed means that, the independent variables don’t satisfactorily explain the price of the car. There still are unknowns which significantly affect the price of the car. This means that estimates of mpg, mileage, tax and engineSize. The model estimates the price of a car given the values of independent variables, albeit not so accurately. for example, if we opt for a car with mileage = 50000 miles, tax = 150 per year, mpg = 50 miles per gallon, engine size = 1.5 litres, the price estimate of the car will be 9586.

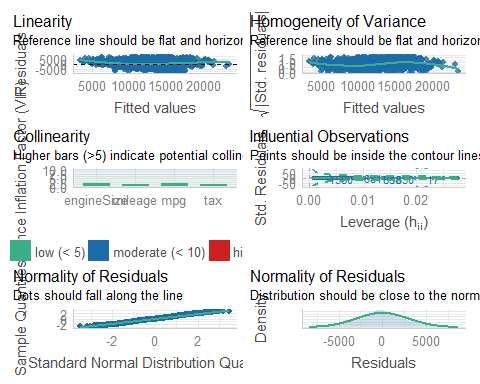
1. Comment on the performance of this model, including comments on overall model fit and the validity of model assumptions. Include any additional code required for you to make these comments in the code chunk below.

**(4 marks)**

### Answer:

This model describes the data better than only mpg. As we can see from the linearity and homogenity of variance plots, the reference line is close to flat, compared to a single regression. There still exist some outliers which are affecting the model. There are no issue with colinearity, this means that none of the independent variables (viz. mileage, mpg, tax,engineSize) are linear with each other and are not dependent with each other. All the residues fall in the referene line and the distribution is suffeciently close to normal distribution. This is a better model for price that model1.

check\_model(modelall)



1. What general concerns do you have regarding this model?

**(2 marks)**

### Answer:

The main concern for this model is the linearity. None of the explanatory variables had strong linear correlation with price. This is also seen in the R squared where this model isn’t best suited to predict the price. The second concern is the sample data that we took from the population. Since the data was randomly sampled, there might be conditions of not being homogenous. The condition of normalcy is however met.

# Question 5: Model simplification (8 marks)

1. What approaches for model simplification would you consider implementing and why?

**(4 marks)**

### Answer:

Firstly, I would get AIC for my linear model to find insignificant interactions and remove them,

model3<- lm(price~mileage+engineSize, data= sample)  
glance(model1)

# A tibble: 1 x 12  
 r.squared adj.r.squared sigma statistic p.value df logLik AIC BIC  
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 0.272 0.271 3606. 723. 1.29e-135 1 -18631. 37269. 37285.  
# ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>

glance(modelall)

# A tibble: 1 x 12  
 r.squared adj.r.squared sigma statistic p.value df logLik AIC BIC  
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 0.500 0.499 2991. 483. 5.39e-289 4 -18267. 36547. 36580.  
# ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>

glance(model3)

# A tibble: 1 x 12  
 r.squared adj.r.squared sigma statistic p.value df logLik AIC BIC  
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 0.460 0.460 3106. 825. 6.59e-260 2 -18341. 36690. 36713.  
# ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>

print(model3)

Call:  
lm(formula = price ~ mileage + engineSize, data = sample)  
  
Coefficients:  
(Intercept) mileage engineSize   
 8189.9710 -0.1853 6481.6843

The AIC and BIC decrease as we increase or explanatory variables. However, the percent decrease is 1.55% if we add energySize into the model. However, if I add more variables to the model, the AIC decrease is just 0.34% This decresae doesn’t justify complicating the model. I would also test the difference in the estimate with a simplified model and a complex model. Let’s take values of explanatory variables to be as follows: mpg=50, mileage = 50000, tax = 150, engineSize = 1.5

In modelall, the price estimate would be 17530 - 130.6 \* 50- 0.1622 \* 50000+5.31 \* 150+ 3933 \* 1.5 = 9586. And in model3, the simplified one, the estimate would be 8189.9710-0.1853*50000+6481.6843*1.5 = 8647.5.

The estimate difference is almost 10%. I would accept this simplification if the estimate difference is only 10%.

1. What are the potential advantages of simplifying a model?

**(2 marks)**

### Answer:

The obvious advantage of simplifying a model is less complex operations and formations. The estimation of parameter (price of car in this case), can be more faster with less loss of information if we simplify the model. Our current model has only a few variables and hence the difference is not visible. However, in a data with large set of variables, the robustness of a model makes a difference in time saving and estimaion.

1. What are the potential disadvantages of simplifying a model?

**(2 marks)**

### Answer:

The most prominent disadvantage of such simplification will be loss of vital information in model and parameter estimation. In critical scientific or other data observations, these lost information make a significant difference. For example, in quantum mechanics calculation, simplifyng a model will be detrimental.

# Question 6: Reporting (35 marks)

A client is looking to purchase a used Skoda Superb (registration year either 2018 or 2019, manual transmission, petrol engine) and wants to understand what factors influence the expected price of a used car, (and how they influence the price).

Write a short report of 300-500 words for the client.

Furthermore, include an explanation as to which model you would recommend, and why you have selected that model.

Comment on any suggestions for alterations to the model that would be appropriate to consider.

Highlight what may or may not be directly transferable from the scenario analysed in Questions 1 to 5.

### Answer:

While selecting a used car, the brand of the car, mileage, engine size, miles per gallon and annual tax affect the price of the car. The year of registration also affects the price, older cars being cheaper.

The average price of car registered in year 2018-2019 that have less than 60000 miles and are manual transmission petrol fueled is $15058.

The most influential factor that influences the price of a car is the mileage, meaning how many miles has the car run before being sold. Increase in mileage will decrease the price of a car. The registration year has lesser difference if we consider only 2018 and 2019, with cars registered in 2019 slightly more expensive. The other influential factor is the engine size of the car. The higher the engine size, the higher the price of the car. Miles per gallon has negative effect in the price of the car. Generally, cars with higher miles per gallon are priced lower. The brand of car has a huge effect in the pricing of the car as well. For example, Ford car are cheaper in average than BMW or Audi.

I would recommend a simple linear equation to calculate the price of the car:

to get a general idea about the price of the car. However, this estimate might not work in lower values of mileage.

If we consider a car with mileage of 50000, and 1.5 litre engine size, the estimated price will be $8911.

The main flaw with this model is that it does not incorporate all the factors of pricing. The model has excluded the mpg of the car which can be a factor in pricing. The dents and cracks might decrease the price or the profit margin the agent expects or simply inability of owners to estimate their depreciation can also affect the price of the car.

# Session Information

Do not edit this part. Make sure that you compile your document so that the information about your session (including software / package versions) is included in your submission.

sessionInfo()

R version 4.1.1 (2021-08-10)  
Platform: x86\_64-w64-mingw32/x64 (64-bit)  
Running under: Windows 10 x64 (build 19042)  
  
Matrix products: default  
  
locale:  
[1] LC\_COLLATE=English\_United States.1252   
[2] LC\_CTYPE=English\_United States.1252   
[3] LC\_MONETARY=English\_United States.1252  
[4] LC\_NUMERIC=C   
[5] LC\_TIME=English\_United States.1252   
  
attached base packages:  
[1] stats graphics grDevices utils datasets methods base   
  
other attached packages:  
 [1] broom\_0.7.11 performance\_0.8.0 outliers\_0.14 skimr\_2.1.3   
 [5] psych\_2.1.9 summarytools\_1.0.0 forcats\_0.5.1 stringr\_1.4.0   
 [9] dplyr\_1.0.7 purrr\_0.3.4 readr\_2.0.2 tidyr\_1.1.4   
[13] tibble\_3.1.5 ggplot2\_3.3.5 tidyverse\_1.3.1   
  
loaded via a namespace (and not attached):  
 [1] nlme\_3.1-152 matrixStats\_0.61.0 fs\_1.5.0 lubridate\_1.8.0   
 [5] bit64\_4.0.5 insight\_0.14.5 httr\_1.4.2 repr\_1.1.4   
 [9] tools\_4.1.1 backports\_1.3.0 utf8\_1.2.2 R6\_2.5.1   
[13] mgcv\_1.8-36 DBI\_1.1.1 colorspace\_2.0-2 withr\_2.4.2   
[17] tidyselect\_1.1.1 mnormt\_2.0.2 bit\_4.0.4 compiler\_4.1.1   
[21] cli\_3.0.1 rvest\_1.0.2 see\_0.6.8 xml2\_1.3.2   
[25] labeling\_0.4.2 bayestestR\_0.11.5 scales\_1.1.1 checkmate\_2.0.0   
[29] digest\_0.6.28 rmarkdown\_2.11 base64enc\_0.1-3 pkgconfig\_2.0.3   
[33] htmltools\_0.5.2 dbplyr\_2.1.1 fastmap\_1.1.0 highr\_0.9   
[37] rlang\_0.4.11 readxl\_1.3.1 rstudioapi\_0.13 pryr\_0.1.5   
[41] generics\_0.1.1 farver\_2.1.0 jsonlite\_1.7.2 vroom\_1.5.5   
[45] magrittr\_2.0.1 rapportools\_1.0 Matrix\_1.4-0 patchwork\_1.1.1   
[49] Rcpp\_1.0.7 munsell\_0.5.0 fansi\_0.5.0 lifecycle\_1.0.1   
[53] stringi\_1.7.5 yaml\_2.2.1 plyr\_1.8.6 grid\_4.1.1   
[57] ggrepel\_0.9.1 parallel\_4.1.1 crayon\_1.4.1 lattice\_0.20-44   
[61] splines\_4.1.1 haven\_2.4.3 pander\_0.6.4 hms\_1.1.1   
[65] magick\_2.7.3 tmvnsim\_1.0-2 knitr\_1.37 pillar\_1.6.4   
[69] tcltk\_4.1.1 codetools\_0.2-18 reprex\_2.0.1 glue\_1.6.0   
[73] evaluate\_0.14 modelr\_0.1.8 vctrs\_0.3.8 tzdb\_0.2.0   
[77] cellranger\_1.1.0 gtable\_0.3.0 datawizard\_0.2.1 assertthat\_0.2.1   
[81] xfun\_0.29 ellipsis\_0.3.2