Used car data analysis

# 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)

# load dataset  
file <- read.csv("Used\_Cars\_Data.csv")  
  
#to see the first few rows of the file  
head(file)

brand model year price transmission mileage fuelType tax mpg engineSize  
1 Audi A1 2017 12500 Manual 15735 Petrol 150 55.4 1.4  
2 Audi A6 2016 16500 Automatic 36203 Diesel 20 64.2 2.0  
3 Audi A1 2016 11000 Manual 29946 Petrol 30 55.4 1.4  
4 Audi A4 2017 16800 Automatic 25952 Diesel 145 67.3 2.0  
5 Audi A3 2019 17300 Manual 1998 Petrol 145 49.6 1.0  
6 Audi A1 2016 13900 Automatic 32260 Petrol 30 58.9 1.4

#to see the number of observations, number of variables, variable names and their data types  
str(file)

'data.frame': 59271 obs. of 10 variables:  
 $ brand : chr "Audi" "Audi" "Audi" "Audi" ...  
 $ model : chr " A1" " A6" " A1" " A4" ...  
 $ year : int 2017 2016 2016 2017 2019 2016 2016 2016 2015 2016 ...  
 $ price : int 12500 16500 11000 16800 17300 13900 13250 11750 10200 12000 ...  
 $ transmission: chr "Manual" "Automatic" "Manual" "Automatic" ...  
 $ mileage : int 15735 36203 29946 25952 1998 32260 76788 75185 46112 22451 ...  
 $ fuelType : chr "Petrol" "Diesel" "Petrol" "Diesel" ...  
 $ tax : int 150 20 30 145 145 30 30 20 20 30 ...  
 $ mpg : num 55.4 64.2 55.4 67.3 49.6 58.9 61.4 70.6 60.1 55.4 ...  
 $ engineSize : num 1.4 2 1.4 2 1 1.4 2 2 1.4 1.4 ...

# Data Preparation

I am interested in modelling the price of vehicles that have all of the following properties:

* mileage less than 65000
* Manual transmission
* Petrol engine (fuelType)
* Costing less than £175 in annual Vehicle Tax.

To model the price of used vehicles as needed, the used cars with the above conditions were selected using multiple filters on the data with symbols for the operators involved such as “AND(&)”,“less than(<)”, “equal to(==)”.

Once I select the rows of data with these properties, I will select a random sample of 2000 rows of the data by setting a seed and then sort the row numbers. Thereafter the rows of the ordered random row numbers will be selected to perform the rest of the analysis with.

There will be some variables in the data set that will no longer be useful information, such as the “transmission” and “fuel type” as its values will all be the same. I removed any redundant variables (where only one value remains in that variable).This subset of the data was used thereafter.

Finally the number of rows and number of columns in the new data frame can be checked with the original file and filtered data set, to confirm the new data frame to be used hereafter.

#to see how many rows are in the data set originally  
nrow(file)

[1] 59271

#filtering for conditions specified(sepal length>4.25 and species is setosa or versicolor)  
filtered\_file <- subset(file,mileage<65000 & transmission == "Manual" & fuelType == "Petrol" & tax < 175)  
  
#how many rows in the data set fit the conditions  
n\_rows <- nrow(filtered\_file)  
n\_rows

[1] 16987

#setting a seed to generate random numbers that can be replicable   
set.seed(21050385)  
  
#2000 is the number of random row numbers we want (for the new sample) from all row numbers   
#replace = FALSE ensures sampling without replacement  
sample\_rows <- sample(1:n\_rows,2000, replace=FALSE)  
  
#sorting the selected row numbers   
sample\_rows <- sort(sample\_rows)  
  
#choosing the rows that matches the row numbers above  
working\_data <- filtered\_file[sample\_rows,]  
  
#see the total number of rows in the new sample  
nrow(working\_data)

[1] 2000

#removing the redundant variables   
new\_df <- subset(working\_data, select = -c(transmission, fuelType) )  
  
#see a few rows in the new sample  
head(new\_df)

brand model year price mileage tax mpg engineSize  
10 Audi A1 2016 12000 22451 30 55.4 1.4  
32 Audi Q3 2017 16500 21369 125 51.4 1.4  
80 Audi A5 2017 18700 20052 145 49.6 2.0  
86 Audi Q3 2017 15600 36714 125 51.4 1.4  
88 Audi A1 2016 14500 12100 30 56.5 1.4  
119 Audi A3 2016 15500 25376 30 57.6 1.4

# Exploratory Data Analysis

## Descriptive Statistics

First the structure of the data set which includes the variable names in the data set along with their data types, total of rows and columns can be seen using the str() function. Then summary() can be used to see each variable’s(continuous variables) minimum, maximum, quartiles and the mean.The measures of central tendency, that is the mean and median in this case, be seen from the summary().

The same can be extended to see the statistics for each brand.Then finding the mode for the categorical variables, brand and engine size would be interesting to look at. Categorical data types in this new data frame includes brand, model,and year as they take fixed set of values that can be grouped whereas others would be continuous. Year is thought to be categorical as we can group the data according to the years. A contingency table can be used to see how many cars of each brand were present in the data set in each of the registration years. A neater table which shows some main statistics can be drawn using flextable() too.

# to see variables along with their data types, total rows and columns  
str(new\_df)

'data.frame': 2000 obs. of 8 variables:  
 $ brand : chr "Audi" "Audi" "Audi" "Audi" ...  
 $ model : chr " A1" " Q3" " A5" " Q3" ...  
 $ year : int 2016 2017 2017 2017 2016 2016 2019 2016 2017 2017 ...  
 $ price : int 12000 16500 18700 15600 14500 15500 28985 11495 16995 12698 ...  
 $ mileage : int 22451 21369 20052 36714 12100 25376 824 32867 18068 11017 ...  
 $ tax : int 30 125 145 125 30 30 145 0 145 150 ...  
 $ mpg : num 55.4 51.4 49.6 51.4 56.5 57.6 40.9 67.3 52.3 67.3 ...  
 $ engineSize: num 1.4 1.4 2 1.4 1.4 1.4 1.5 1 1.4 1 ...

From str() it can be seen that the data set has 2000 rows and 8 variables. The brand and model variables are character data type, mpg and engine size is numerical and all other variables are integer data type.

summary(new\_df)

brand model year price   
 Length:2000 Length:2000 Min. :2001 Min. : 1795   
 Class :character Class :character 1st Qu.:2017 1st Qu.: 8945   
 Mode :character Mode :character Median :2017 Median :11200   
 Mean :2017 Mean :12420   
 3rd Qu.:2019 3rd Qu.:15290   
 Max. :2020 Max. :34259   
 mileage tax mpg engineSize   
 Min. : 4 Min. : 0.0 Min. :28.5 Min. :0.000   
 1st Qu.: 7942 1st Qu.: 30.0 1st Qu.:51.4 1st Qu.:1.000   
 Median :14876 Median :145.0 Median :57.7 Median :1.000   
 Mean :17753 Mean :111.2 Mean :56.3 Mean :1.202   
 3rd Qu.:24817 3rd Qu.:145.0 3rd Qu.:62.8 3rd Qu.:1.400   
 Max. :64447 Max. :165.0 Max. :69.0 Max. :3.000

With summary() it can be seen that the earliest registration year is 2001 while the latest is 2020.The minimum price of a used car is 1795 pounds and highest price is 34259 pounds and the mean price of a used car is 12420. The minimum mileage for a used car is 4mph while the highest is 64447mph and the mean mileage is 17753mph. The minimum annual cost for vehicle tax and minimum for engine size in liters is zero which can not be the case in real world so this must have been due to an error in data entry. Maximum vehicle tax is 165 while the mean is 111.2. Maximum engine size is 3 while mean is 1.202. lowest mpg is 28.5 while highest is 69 and the mean is 56.3. The difference between each of the variables mean and median is not very high so this indicates there are no gross outliers in the data set.

#to see the summary statistics for each brand   
by(new\_df, new\_df$brand, summary)

new\_df$brand: Audi  
 brand model year price   
 Length:250 Length:250 Min. :2010 Min. : 6475   
 Class :character Class :character 1st Qu.:2016 1st Qu.:12996   
 Mode :character Mode :character Median :2017 Median :16498   
 Mean :2017 Mean :17511   
 3rd Qu.:2019 3rd Qu.:20569   
 Max. :2020 Max. :34259   
 mileage tax mpg engineSize   
 Min. : 10 Min. : 0.0 Min. :37.20 Min. :1.000   
 1st Qu.: 7856 1st Qu.: 30.0 1st Qu.:47.10 1st Qu.:1.400   
 Median :19831 Median :145.0 Median :52.30 Median :1.400   
 Mean :19754 Mean :109.7 Mean :52.09 Mean :1.418   
 3rd Qu.:28551 3rd Qu.:145.0 3rd Qu.:57.60 3rd Qu.:1.500   
 Max. :60000 Max. :160.0 Max. :67.30 Max. :2.000   
------------------------------------------------------------   
new\_df$brand: BMW  
 brand model year price   
 Length:85 Length:85 Min. :2001 Min. : 8490   
 Class :character Class :character 1st Qu.:2017 1st Qu.:14991   
 Mode :character Mode :character Median :2019 Median :19495   
 Mean :2018 Mean :18714   
 3rd Qu.:2019 3rd Qu.:22000   
 Max. :2020 Max. :32995   
 mileage tax mpg engineSize   
 Min. : 10 Min. : 20.0 Min. :36.20 Min. :0.000   
 1st Qu.: 2574 1st Qu.:145.0 1st Qu.:44.10 1st Qu.:1.500   
 Median : 6196 Median :145.0 Median :51.40 Median :1.500   
 Mean :11934 Mean :140.8 Mean :48.66 Mean :1.586   
 3rd Qu.:19995 3rd Qu.:150.0 3rd Qu.:52.30 3rd Qu.:1.500   
 Max. :57736 Max. :165.0 Max. :68.90 Max. :3.000   
------------------------------------------------------------   
new\_df$brand: Ford  
 brand model year price   
 Length:1213 Length:1213 Min. :2008 Min. : 1795   
 Class :character Class :character 1st Qu.:2017 1st Qu.: 8900   
 Mode :character Mode :character Median :2017 Median :10750   
 Mean :2017 Mean :11535   
 3rd Qu.:2018 3rd Qu.:13930   
 Max. :2020 Max. :28998   
 mileage tax mpg engineSize   
 Min. : 5 Min. : 0.0 Min. :34.40 Min. :0.000   
 1st Qu.: 8386 1st Qu.: 30.0 1st Qu.:54.30 1st Qu.:1.000   
 Median :14602 Median :145.0 Median :58.90 Median :1.000   
 Mean :17888 Mean :111.5 Mean :57.38 Mean :1.113   
 3rd Qu.:24195 3rd Qu.:145.0 3rd Qu.:64.20 3rd Qu.:1.200   
 Max. :64447 Max. :165.0 Max. :67.30 Max. :2.300   
------------------------------------------------------------   
new\_df$brand: Mercedes  
 brand model year price   
 Length:62 Length:62 Min. :2013 Min. : 9991   
 Class :character Class :character 1st Qu.:2017 1st Qu.:15892   
 Mode :character Mode :character Median :2018 Median :17770   
 Mean :2018 Mean :18754   
 3rd Qu.:2019 3rd Qu.:21659   
 Max. :2020 Max. :28200   
 mileage tax mpg engineSize   
 Min. : 19 Min. :125.0 Min. :28.50 Min. :1.300   
 1st Qu.: 3644 1st Qu.:145.0 1st Qu.:46.50 1st Qu.:1.300   
 Median :13280 Median :145.0 Median :49.60 Median :1.600   
 Mean :16834 Mean :141.4 Mean :46.85 Mean :1.544   
 3rd Qu.:26750 3rd Qu.:145.0 3rd Qu.:51.40 3rd Qu.:1.600   
 Max. :53000 Max. :160.0 Max. :53.30 Max. :2.000   
------------------------------------------------------------   
new\_df$brand: Toyota  
 brand model year price   
 Length:390 Length:390 Min. :2007 Min. : 3491   
 Class :character Class :character 1st Qu.:2016 1st Qu.: 7488   
 Mode :character Mode :character Median :2017 Median : 8995   
 Mean :2017 Mean : 9529   
 3rd Qu.:2019 3rd Qu.:10995   
 Max. :2020 Max. :31000   
 mileage tax mpg engineSize   
 Min. : 4 Min. : 0.0 Min. :33.20 Min. :1.000   
 1st Qu.: 8398 1st Qu.: 30.0 1st Qu.:52.30 1st Qu.:1.000   
 Median :15006 Median :145.0 Median :58.00 Median :1.000   
 Mean :17465 Mean :100.1 Mean :58.78 Mean :1.202   
 3rd Qu.:24287 3rd Qu.:145.0 3rd Qu.:68.90 3rd Qu.:1.500   
 Max. :63460 Max. :160.0 Max. :69.00 Max. :2.000

Each of the variables summary statistics can be seen next based on the different brands. For example, we can see the earliest and latest registration years or the range(maximum - minimum) of taxes or the variation in engine sizes for the brand “Toyota” if needed.

#To find the mode for Brands and engine size where number of occurrences for each unique value is taken and sorted from highest to lowest.  
sort(table(new\_df$brand), decreasing = TRUE)

Ford Toyota Audi BMW Mercedes   
 1213 390 250 85 62

sort(table(new\_df$engineSize), decreasing = TRUE)

1 1.5 1.2 1.4 2 1.1 1.3 1.6 1.8 2.3 0 3   
1061 285 214 122 78 72 70 70 14 8 4 2

With the sort() function descending ordering, it can be concluded that the modal brand of used cars in the data set is Ford (61% as 1213 of the 2000 used cars is Ford brand). It can also be seen that mode for engine size in litres is 1 (53% as 1061 used cars out of 2000).

#two qualitative variables year and brand used to create a contingency table  
xtabs(~ new\_df$year + new\_df$brand)

new\_df$brand  
new\_df$year Audi BMW Ford Mercedes Toyota  
 2001 0 1 0 0 0  
 2007 0 0 0 0 1  
 2008 0 0 1 0 0  
 2009 0 0 1 0 2  
 2010 1 0 1 0 0  
 2011 0 0 1 0 0  
 2012 2 0 2 0 0  
 2013 7 0 37 1 7  
 2014 6 3 49 1 13  
 2015 13 3 90 2 25  
 2016 48 9 111 8 52  
 2017 84 17 314 14 119  
 2018 21 5 325 11 71  
 2019 58 46 263 17 95  
 2020 10 1 18 8 5

#to see the summary statistics rounded to 2 decimal places, of continuous columns  
Summary\_Stats<-round(describe(new\_df[,c(4,5,6,7,8)]),2)  
  
#to see only the 5 statistics: mean, median,sd, min and max values for each continuous variable  
Summary\_Stats<-Summary\_Stats[,c("mean","median","sd","min","max")]  
Summary\_Stats<-cbind(rownames(Summary\_Stats),Summary\_Stats)  
colnames(Summary\_Stats)[1]<-"Variable"  
flextable(Summary\_Stats)

| Variable | mean | median | sd | min | max |
| --- | --- | --- | --- | --- | --- |
| price | 12,419.97 | 11,200.0 | 4,954.09 | 1,795.0 | 34,259 |
| mileage | 17,753.07 | 14,875.5 | 13,220.59 | 4.0 | 64,447 |
| tax | 111.22 | 145.0 | 57.38 | 0.0 | 165 |
| mpg | 56.30 | 57.7 | 8.25 | 28.5 | 69 |
| engineSize | 1.20 | 1.0 | 0.29 | 0.0 | 3 |

The contingency table shows the registration yearsand how many of the used cars of each brand were registered for each year. In 2001 only 1 BMW car was registered in our sample. Then from 2007 to 2012 only very few cars were registered. 2014 on wards registration of all cars have increased. Ford has increased in higher proportion from 2016 to 2019 and however dropped drastically in 2020.

## Exploratory Graphs

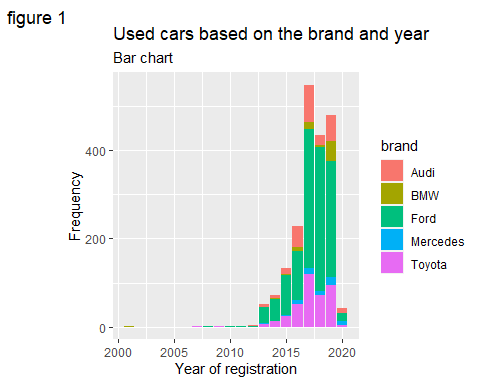
Dependent variable for this data set can be price as we are trying to estimate it. The fact that price shows response to changes in any other (predictor) variable implies that it is the dependent variable.

Bar charts can be used to see the distribution of variables and to see how many observations it has. This can be used to see how many cars of each brand were registered in the years. The bars can be stacked to show the two categorical variables, year and brand.

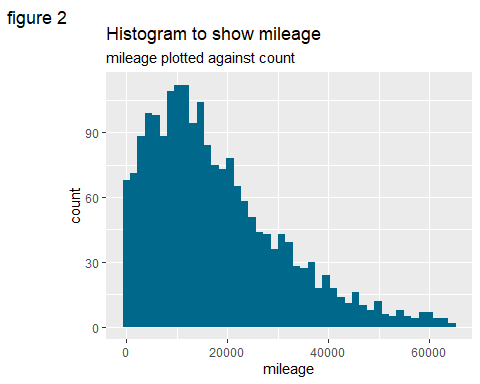
A histogram can be drawn to represent mileage to see the shape of its distribution.Box plots can be drawn to show the variation of data, including range and IQR so miles per gallon can be compared for each brand.

It would be interesting to see how the prices and vehicle tax have changed with the mileage. So a scatter plot can be drawn using aesthetics.

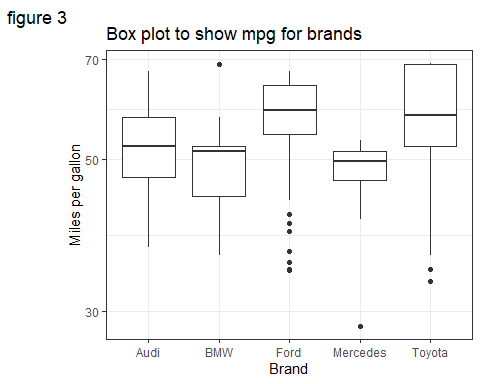
ggplot(new\_df, aes(year, fill = brand))+   
 geom\_bar()+  
 labs(title= "Used cars based on the brand and year",  
 subtitle= "Bar chart",  
 tag="figure 1")+  
 xlab("Year of registration")+  
 ylab("Frequency")+  
 theme\_gray()

 Figure 1 is to get an insight of how many observations from the dataset fall to each brand of used cars based on the year of registration. From mid year 2012 most of the cars in our sample data set has been registered and the number of cars registered in total has increased over the years upto mid 2017, specially the brands Ford and Audi. Then total registrations have reduced but increased again shortly. However there is a sharp drop in the number of cars(of every brand) that were registered in the year 2020.

ggplot(new\_df, aes(x = mileage)) +  
 geom\_histogram(bins = 45, fill = 'deepskyblue4') +  
 labs(title= "Histogram to show mileage",  
 subtitle= "mileage plotted against count",  
 tag="figure 2")+  
 xlab("mileage")

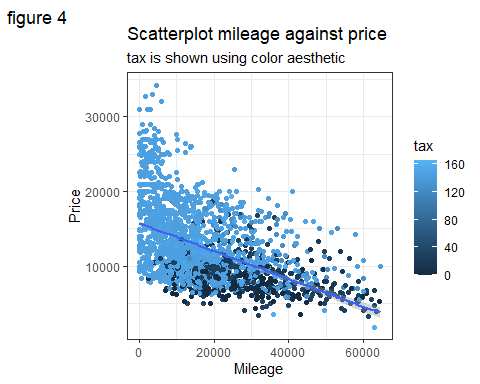
 Figure 2 mileage is one of the most important when considering used cars. However from the histogram it can be seen that the data is positively skewed for our sample dataset.This can be confirmed With the summary statistics obtained earlier(Q3 -Q2> Q2 - Q1).

ggplot(new\_df)+   
 geom\_boxplot(mapping = aes(x= brand, y=mpg))+  
 scale\_y\_log10()+  
 labs(title= "Box plot to show mpg for brands",  
 tag="figure 3")+  
 xlab("Brand")+  
 ylab("Miles per gallon")+  
 theme\_bw()

 Figure 3 Miles per gallon gives information about the fuel efficiency. So with the box plot it can be seen that in used cars, Toyota cars seem to have the highest fuel efficiency with a bigger range and IQR. It can also be seen that the data is positively skewed for Toyota, almost symmetrical for Audi and Ford, and negatively skewed for BMW, Mercedes.

ggplot(data = new\_df, aes(x=mileage, y=price))+  
 geom\_point(aes(color=tax),)+  
 geom\_smooth(method = "lm")+  
 labs(title= "Scatterplot mileage against price",  
 subtitle= "tax is shown using color aesthetic",  
 tag="figure 4")+  
 xlab("Mileage")+  
 ylab("Price")+  
 theme\_bw()

`geom\_smooth()` using formula = 'y ~ x'

 Figure 4 shows that as mileage increases, prices reduces and the line of best fit has a negative slope, both suggesting a negative correlation. Also the color aesthetic suggests that vehicle taxes are generally higher for used cars with less mileage.

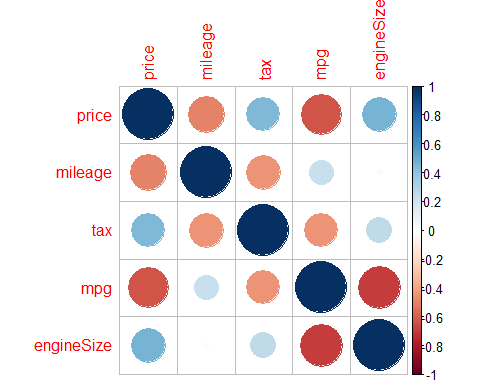
Skewness is detected in many variables, and the tail parts can have outliers. Outliers affect the performance of a model. So care should be taken when performing linear regression models.

## Correlations

#setting the continuous variables only as the dataframe to be used forward  
new\_cont\_df <- new\_df[-c(1:3)]  
  
#to see the correlation matrix   
cor(new\_cont\_df)

price mileage tax mpg engineSize  
price 1.0000000 -0.49711650 0.4417856 -0.6219929 0.46593263  
mileage -0.4971165 1.00000000 -0.4491412 0.2288482 -0.01969536  
tax 0.4417856 -0.44914124 1.0000000 -0.4410553 0.25433996  
mpg -0.6219929 0.22884822 -0.4410553 1.0000000 -0.69492850  
engineSize 0.4659326 -0.01969536 0.2543400 -0.6949285 1.00000000

#corrplot to see which variables are strongly correlated  
corrplot(cor(new\_cont\_df))



#to see the correlation coefficients of each variable with the dependent variable   
cor(new\_cont\_df)[-1,1]

mileage tax mpg engineSize   
-0.4971165 0.4417856 -0.6219929 0.4659326

#to see which variable has strongest correlation with the dependent variable  
names(which.max(abs(cor(new\_cont\_df)[-1,1])))

[1] "mpg"

mpg has the strongest linear relationship with the dependent variable,price.

#creating a linear relationship between price and mpg  
model\_1 <- lm(price ~ mpg, data=new\_cont\_df)   
summary(model\_1)

Call:  
lm(formula = price ~ mpg, data = new\_cont\_df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-14622.1 -2508.4 -16.6 2133.4 16839.5   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 33455.51 598.76 55.88 <2e-16 \*\*\*  
mpg -373.65 10.52 -35.51 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 3880 on 1998 degrees of freedom  
Multiple R-squared: 0.3869, Adjusted R-squared: 0.3866   
F-statistic: 1261 on 1 and 1998 DF, p-value: < 2.2e-16

The linear regression equation is, price = 33455.51-373.65\*mpg

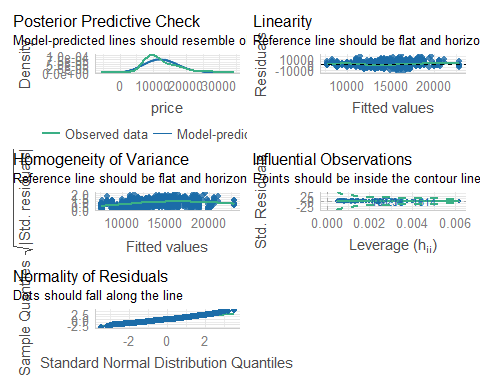
For every unit increase in miles per gallon, Price decreases by 373.65. When miles per gallon is theoretically zero,the price is 33455 GB pounds. The coefficients for intercept and gradient is significant as their p values are less than 0.05(indicated by the stars). Hence the null hypothesis stating that there is no relationship between mpg and price can be rejected as there is significant evidence of a relationship between price and mpg.

In real world this may not be the case, but in our model, as mpg is positively correlated with mileage and mileage is negatively correlated with price, it can cause the negative correlation between price and mpg.

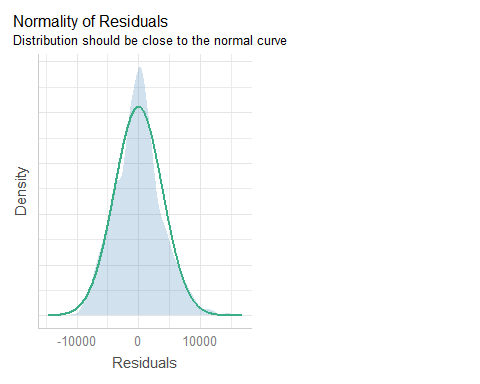
The output R-squared = 0.3869 means that the explanatory variable mpg, explains 38.69% of variation in price of used cars.R -squared is used to measure the relationship between price and mpg. The overall model is also significant at 5% level as p value is less than 0.05.

## Performance of this model

check\_model(model\_1)



check\_model(model\_1,check="normality")

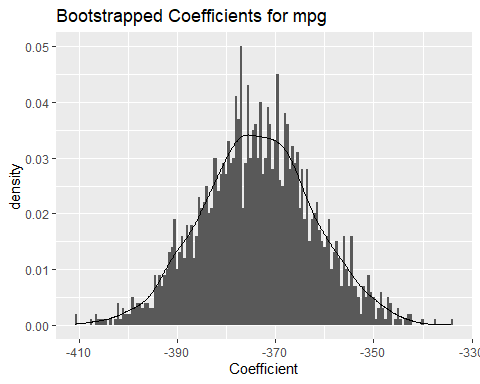
 The overall model fit and the validity of model assumptions can be seen from above.From the output for check\_model() it can be seen that linearity assumption is not violated as an almost flat line can be seen. Homogeneity of variance shows about the linearity of the residuals versus the fitted values and that is also almost a flat line. None of the predictions are outside of the contour lines indicating that there are no high leverage points. The normal density plot of normality of residuals graph is also almost a symmetrical normal curve.So all these graphs above indicate that no further adjustments are needed for the model.

However,there are some dots that do not fall along the line in normality of residuals graph indicating that the model may need to be changed slightly for those points.

## Bootstrap

Bootstrapping on was done on this model to obtain a 95% confidence interval of the estimate of the slope parameter.

# Set up bootstrap  
Nbootstrap<- 2000   
coeff\_of\_mpg <-rep(NA,Nbootstrap) #empty vector  
  
# Perform bootstrap and set seed to make sample reproducible  
set.seed(4363)   
  
for(i in seq\_len(Nbootstrap)){  
 usevalues <- sample(seq\_along(new\_cont\_df$price),size=length(new\_cont\_df$price),replace=TRUE)   
 bootstrap.sample <- new\_cont\_df[usevalues,]  
   
 new\_model\_1 <- lm( price ~ mpg, data=bootstrap.sample)  
   
 #trace = false wont show output each time loop is run  
 model\_boot <- stepAIC(new\_model\_1, trace=FALSE)  
   
 #collecting the coefficient called mpg and storing it as mpg[i]  
 coeff\_of\_mpg[i] <- model\_boot$coef["mpg"]   
}  
  
coeff\_of\_mpg[is.na(coeff\_of\_mpg)] <- 0 #setting all the missing values to zero  
  
Bootstrap <- data.frame(Coefficient=coeff\_of\_mpg)  
ggplot(Bootstrap, aes(x=Coefficient)) + geom\_histogram(aes(y = after\_stat(density)),binwidth = .5)+  
 geom\_density()+  
ggtitle("Bootstrapped Coefficients for mpg")+  
theme\_gray()



# 95% bootstrap confidence interval  
quantile(coeff\_of\_mpg,c(.025,0.975))

2.5% 97.5%   
-395.9558 -350.8136

# Multivariable relationship

Creating a model with all of the appropriate remaining explanatory variables included.

#creating a linear model with all independent variables and price  
model\_all <- lm(price ~ . , data=new\_cont\_df)   
summary(model\_all)

Call:  
lm(formula = price ~ ., data = new\_cont\_df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-9028.0 -2267.4 83.2 1986.7 14079.9   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 2.402e+04 1.184e+03 20.290 < 2e-16 \*\*\*  
mileage -1.423e-01 6.506e-03 -21.864 < 2e-16 \*\*\*  
tax 4.748e+00 1.599e+00 2.970 0.00302 \*\*   
mpg -2.347e+02 1.393e+01 -16.850 < 2e-16 \*\*\*  
engineSize 3.006e+03 3.758e+02 7.999 2.11e-15 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 3376 on 1995 degrees of freedom  
Multiple R-squared: 0.5366, Adjusted R-squared: 0.5357   
F-statistic: 577.5 on 4 and 1995 DF, p-value: < 2.2e-16

For every unit increase in mileage, Price decreases by 1.423e-01. For every unit increase in tax price increases by 4.748e+00. For every unit increase in miles per gallon, price decreases by 2.347e+02. For every unit increase in engine size, price increases by 3.006e+03. When all independent variables are theoretically zero,the price is 2.402e+04 GB pounds.

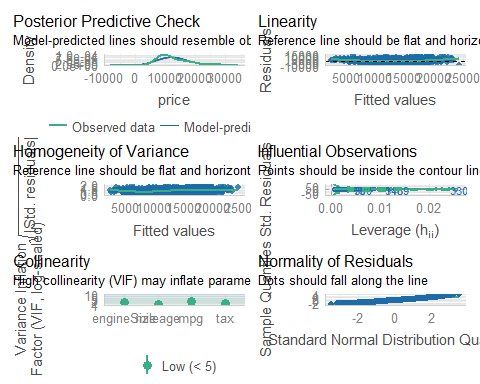
All the coefficients are significant at 5% significance level as they all have pvalues less than 0.05(shown by the stars). Similarly the overall model is significant. This makes us reject the null hypothesis as there is significant evidence of a relationship between price and all other predictor variables collectively.

## Performance, model fit and validity of model assumptions

From previous output it can be seen that R squared has increased with the multiple regression model,as expected, from 0.3869 to 0.5366 implying that all the predictor variables together explains 53.66% of variation in price. Adjusted R squared is preferred in this case as it accounts the variation (by adding all predictor variables). Adjusted R squared is 0.5357.

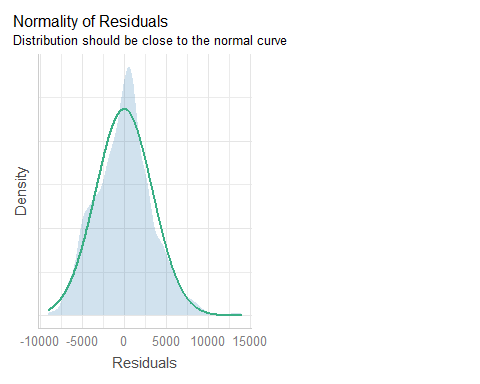
check\_model(model\_all)

Variable `Component` is not in your data frame :/



check\_model(model\_all,check="normality")

Variable `Component` is not in your data frame :/



From the output for check\_model() it can be seen that linearity assumption is appropriate because the line is almost flat line. Homogeneity of variance graph that shows the linearity of the residuals versus the fitted values is almost a flat line. None of the predictions are outside of the contour lines indicating that there are no high leverage points. The normal density plot of normality of residuals graph is also almost a symmetrical normal curve.There is no issue of multicollinearity as well. So all these graphs above indicate that no further adjustments are needed for the model.

However,some dots do not fall along the line in normality of residuals graph. That means that the model may have to be inspected for those points.

## Concerns regarding this model

The model takes into account some important predictor variables to predict the changes in the response variable, however, other factors like car interior condition, repair history and service history can affect the price of a used cars. Also as some points do not fall along the line in normality residual graph, it raises concerns for the performance of the model. Also some of the relationships may be non linear. Linear regression model may not be accurate if there are outliers or if there are data entry errors.

# Scenario of Reporting to a client

A client is looking to purchase a used VW Polo (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).

The following include an explanation as to which statistical model I would recommend, and why I have selected that statistical model. I would further comment on any suggestions for alterations to the statistical model that would be appropriate to consider and highlight what may or may not be directly transferable from the scenario analysed in the codes above.

### Report:

Since the client wants a used car registered in 2018 or 2019, with manual transmission and petrol engine, the statistical analysis done will be appropriate. This is because, the sample dataset analysed is for the used cars registered from the year 2001 to 2020 where data was filtered to have only manual transmission and petrol engines.

This dataset consisted of five different brands of cars namely BMW, Toyota,Mercedes, Ford, Audi. Unfortunately it did not include VW Polo. However the conditions for VW can be compared in relation to those brands in the dataset. From summary statistics it was found that the mean price(which may be affected by outliers) of a used car was 12,420 pounds while median price(better measure of central tendency as it is not affected by outliers) was 11,200 pounds.

The (second) multiple linear regression model is better to explain which factors affect the price of a used car as that model is with all explanatory variables with a higher multiple R squared value, and hence explains more of the changes in price than with only mpg with lower R squared value.

With the analysis done on correlation such as the correlation matrix it was found that as the mileage and miles per gallon increase, the price of the used car will decrease. As taxes and enginesize increased in value, the price of the used car also increased. So if the price needs to be reduced, it is better to check for used cars with higher mileage and miles per gallon, and for used cars with lower vehicle tax and engine size.

An alteration to the statistical model to be considered would be to introduce some observations for the brand VW so that a more relevant analysis can be done to meet your requirements. A graph such as column chart, showing the variation in price for the years 2018 and 2019 can be plotted to help you decide on which year of registration to look for when buying a used car.