Exam Template: Statistical Inference

21050307

Sep21 run RESIT

# 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 and read as a dataframe  
si\_2022 <- import("../data/Jan\_2022\_Exam\_Data.csv")  
head(si\_2022)

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

# check structure of dataset  
str(si\_2022)

'data.frame': 52533 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 ...

# Converting the years to factor format.  
si\_2022$year <- as.factor(si\_2022$year)

# 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 90000
* Manual transmission
* Diesel engine (fuelType)
* Costing less than £300 in annual Vehicle Tax.

Once you have selected the rows of data with these properties, then you must *use the last 4 digits of 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:

Inorder for the data provided to be suitable for analysis, the following need to be done:

1. Tools: the right tool to prepare the data would be needed. In this case we would be using, *R studio and the above imported packages* .
2. Importation & Profiling: the next stage is to *load and read the Jan\_2022\_Exam\_Data.csv data as a dataframe*. The correct file path leading to the data should be taken into consideration.
3. Formating: ensuring the data is in proper data type and structure after reading and profiling
4. Data selection: since are interested in modelling the price of vehicles that has *mileage less than 90,000, a manual transmission, diesel engine and has an annual tax of less than 300*, we need to creat a subset of the data that satisfies this. To finish the selection stage, we would be making use of my *student ID to select 2000 random data from the dataset using the set.seed()*.
5. Cleaning/Cleansing: here we would be making sure unnecessary and redundant variables and terms that has no effect on our model are removed
6. Data enrichment : enhance the data and model to make it optimal
7. Data publishing/analysis : publish data for analysis
8. Implement the required data preparation in the code chunk below:

**(7 marks)**

### Answer:

# select rows based on conditions  
si\_2022 <- si\_2022[si\_2022$mileage<90000 & si\_2022$transmission=="Manual" & si\_2022$fuelType=="Diesel"& si\_2022$tax<300,]  
  
# using setseed(), select random 2000 data   
set.seed(21050307)  
sample\_si\_2022 <- sample(1:nrow(si\_2022) , size=2000, replace= FALSE)  
sorted\_sample\_si\_2022 <- sort(sample\_si\_2022)  
sorted\_data <- si\_2022[sorted\_sample\_si\_2022,]  
  
# remove redundant variables   
final\_sorted\_data <- subset(sorted\_data, select = -c(fuelType, transmission))

# to check if there are na values  
final\_sorted\_data = na.omit(final\_sorted\_data)  
sum(is.na(final\_sorted\_data))

[1] 0

# 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:

Firstly, i will be taking a look at the central tendency of the the distribution of data by looking at the mean and median of price, mileage, tax,mpg and engine size

After that, i would be checking the spread of values of the price, mileage, tax,mpg and engine size around the central tendency by looking at the standard deviation and skewness

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

**(4 marks)**

### Answer:

describe(final\_sorted\_data)

vars n mean sd median trimmed mad min max  
brand\* 1 2000 2.49 0.94 3.0 2.49 0.00 1.0 4.0  
model\* 2 2000 17.65 10.58 21.0 17.68 11.86 1.0 43.0  
year\* 3 2000 22.34 1.82 22.0 22.44 1.48 15.0 26.0  
price 4 2000 13951.83 4732.72 13500.0 13640.21 4755.44 3200.0 34950.0  
mileage 5 2000 33082.24 20991.00 30004.0 31596.41 21336.10 1.0 89350.0  
tax 6 2000 86.41 65.46 125.0 87.70 37.06 0.0 260.0  
mpg 7 2000 64.16 9.92 64.2 64.53 11.42 31.7 88.3  
engineSize 8 2000 1.79 0.25 2.0 1.79 0.15 0.0 2.3  
 range skew kurtosis se  
brand\* 3.0 -0.38 -0.92 0.02  
model\* 42.0 -0.09 -1.17 0.24  
year\* 11.0 -0.63 0.96 0.04  
price 31750.0 0.68 0.56 105.83  
mileage 89349.0 0.56 -0.37 469.37  
tax 260.0 -0.13 -1.72 1.46  
mpg 56.6 -0.33 -0.07 0.22  
engineSize 2.3 -0.52 0.64 0.01

count(final\_sorted\_data, 'model')

model freq  
1 1 Series 160  
2 2 Series 49  
3 3 Series 76  
4 4 Series 15  
5 5 Series 6  
6 A Class 116  
7 A1 42  
8 A3 130  
9 A4 69  
10 A5 36  
11 A6 27  
12 B Class 11  
13 B-MAX 13  
14 C Class 25  
15 C-MAX 56  
16 CL Class 10  
17 CLA Class 2  
18 EcoSport 40  
19 Edge 8  
20 Fiesta 89  
21 Focus 296  
22 Galaxy 34  
23 GL Class 4  
24 GLA Class 12  
25 Grand C-MAX 30  
26 Grand Tourneo Connect 11  
27 Kuga 296  
28 Mondeo 70  
29 Q2 35  
30 Q3 73  
31 Q5 15  
32 S-MAX 46  
33 Tourneo Connect 4  
34 Tourneo Custom 13  
35 TT 9  
36 V Class 21  
37 X-CLASS 1  
38 X1 38  
39 X2 7  
40 X3 2  
41 X4 1  
42 180 1  
43 200 1

count(final\_sorted\_data, 'year')

year freq  
1 2009 9  
2 2010 6  
3 2011 8  
4 2012 26  
5 2013 100  
6 2014 131  
7 2015 252  
8 2016 494  
9 2017 522  
10 2018 165  
11 2019 270  
12 2020 17

count(final\_sorted\_data, 'brand')

brand freq  
1 Audi 436  
2 BMW 354  
3 Ford 1006  
4 Mercedes 204

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:

a.Statistics given for brand, model and year should be ignored because they are categorical and factor variable, hence the asterisk. In order to summarize the variables suitably well, i used the count() to get the frequency of the data. b. mpg posses a symmetrical distribution (mean and median occur at same point) c. mileage and price is positively skewed, meaning most of the distribution occurs above the median d. engineSize,mpg and tax are negatively skewed, meaning most the distribution occurs below the median e) with the standard deviation of price and mileage been high, it shows that the values are spread across a wider range.

## Exploratory Graphs

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

**(2 marks)**

### Answer:

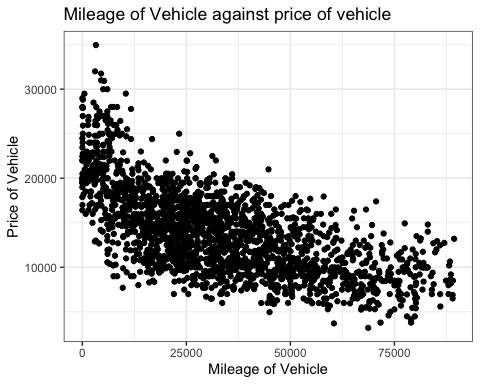
since the aim is to check what affects the price of a used manual transmission diesel vehicle

1. Scatter Plots:
2. mpg against price of vehicle
3. mileage against price of vehicle
4. brand against price of vehicle
5. Box Plot
   1. Model against price of vehicle
6. Now produce those exploratory graphs in the code chunk below:

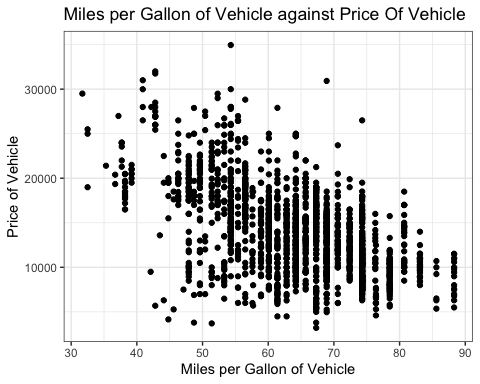
**(4 marks)**

### Answer:

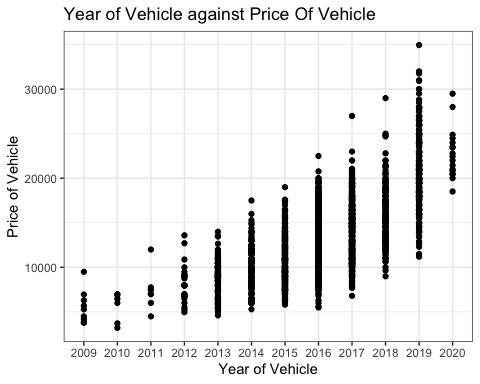
par(mfrow = c(3, 1))  
  
final\_sorted\_data %>%  
ggplot(aes(x=mileage,y=price))+  
 geom\_point()+  
 labs(title = "Mileage of Vehicle against price of vehicle") +   
 xlab("Mileage of Vehicle") +   
 ylab("Price of Vehicle") +   
 theme\_bw()



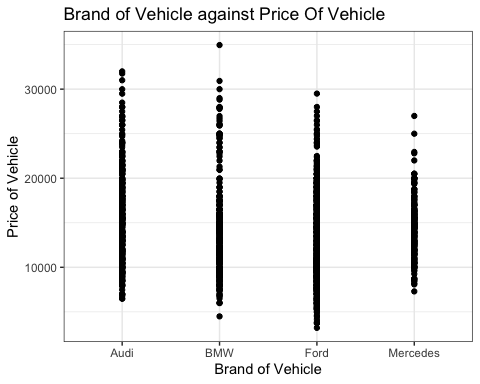
final\_sorted\_data %>%  
ggplot(aes(x=mpg,y=price))+  
 geom\_point()+  
 labs(title = "Miles per Gallon of Vehicle against Price Of Vehicle") +   
 xlab("Miles per Gallon of Vehicle") +   
 ylab("Price of Vehicle") +   
 theme\_bw()



final\_sorted\_data %>%  
ggplot(aes(x=year,y=price))+  
 geom\_point()+  
 labs(title = "Year of Vehicle against Price Of Vehicle") +   
 xlab("Year of Vehicle") +   
 ylab("Price of Vehicle") +   
 theme\_bw()

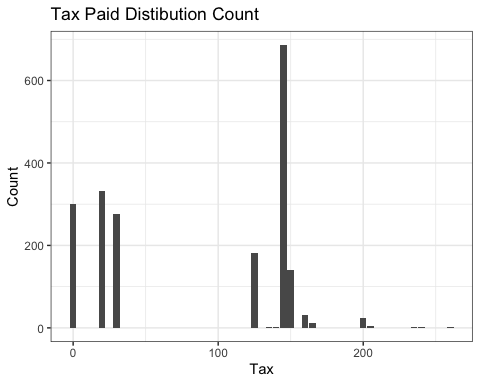


final\_sorted\_data %>%  
ggplot(aes(x=brand,y=price))+  
 geom\_point()+  
 labs(title = "Brand of Vehicle against Price Of Vehicle") +   
 xlab("Brand of Vehicle") +   
 ylab("Price of Vehicle") +   
 theme\_bw()

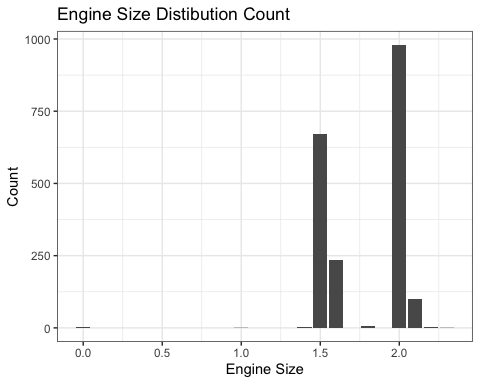


par(mfrow=c(1, 1))

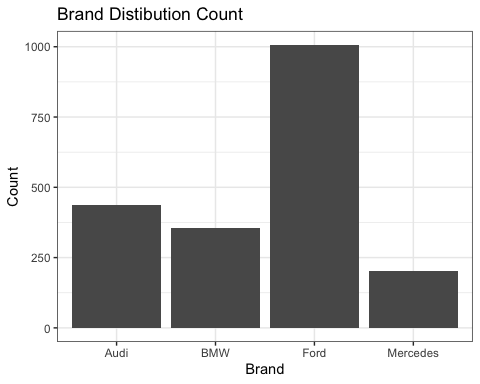
par(mfrow = c(3, 1))  
  
final\_sorted\_data %>%  
ggplot(aes(x=tax))+  
 geom\_bar()+  
 labs(title = "Tax Paid Distibution Count")+   
 xlab("Tax") +   
 ylab("Count") +   
 theme\_bw()



final\_sorted\_data %>%  
ggplot(aes(x=engineSize))+  
 geom\_bar()+  
 labs(title = "Engine Size Distibution Count") +   
 xlab("Engine Size") +   
 ylab("Count") +   
 theme\_bw()

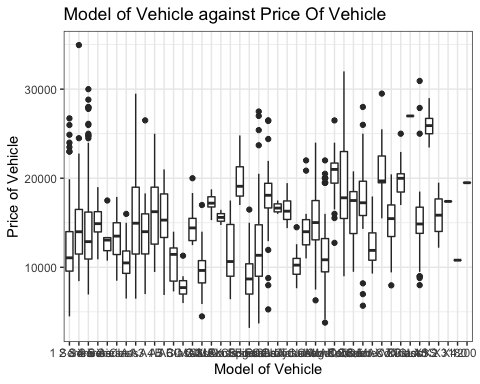


final\_sorted\_data %>%  
ggplot(aes(x=brand))+  
 geom\_bar()+  
 labs(title = "Brand Distibution Count") +   
 xlab("Brand") +   
 ylab("Count") +   
 theme\_bw()



par(mfrow=c(1, 1))

final\_sorted\_data %>%  
ggplot(aes(x=model,y=price))+  
 geom\_boxplot()+  
 labs(title = "Model of Vehicle against Price Of Vehicle") +   
 xlab("Model of Vehicle") +   
 ylab("Price of Vehicle") +   
 theme\_bw()



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

**(4 marks)**

### Answer:

Diagram 1-4: a) There looks to be a relationship between mileage and price. From left to right, we can see as mileage increases, the price lowers. It can also be seen that the relationship is not definite. On further analysis, we could see that other factors like brand and year of the car affects the price.

1. There is a relationship between mpg and price. From left to right, we can see as mpg increases, the price lowers. It can be seen the decline in price isnt constant as brand and year of vehicle can be a big factor as luxury brands which are more expensive tend to have lower mpg. It can also be seen that the relationship is not definite. On further analysis, we could see that other factors like brand and year of the car affects the price.
2. There is a strong relationship between the year a car was made and the price. We can see that from left to right, as year increases, the price of the vehicle increases too. A vehicle made in 2019 had the highest price. This could be a luxury brand.
3. It can be seen that each brand have different cars at different prices. Ford had more models/make at the lower end of the price spectrum while BMW had more expensive cars.

Diagram 5-7 a) most of the vehicles in the dataset had engine size of 1.25 - 2.25 b) there was a disparity between taxes paid but bulk of the tax was between 130 - 250.

## Correlations

1. What linear correlations are present within this data?

**(2 marks)**

### Answer:

we would be checking if there is a correlation between price and other variables

1. price of a vehicle decreases as mileage or mpg increases hence the negative correlation.
2. price of a vehicle increases as tax or engine size decreases hence the positive correlation.

cor(final\_sorted\_data[,unlist(lapply(final\_sorted\_data, is.numeric))])

price mileage tax mpg engineSize  
price 1.0000000 -0.6500129 0.5352730 -0.5474798 0.3496872  
mileage -0.6500129 1.0000000 -0.3817466 0.2017684 0.0821694  
tax 0.5352730 -0.3817466 1.0000000 -0.6097958 0.2859365  
mpg -0.5474798 0.2017684 -0.6097958 1.0000000 -0.5498712  
engineSize 0.3496872 0.0821694 0.2859365 -0.5498712 1.0000000

# 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:

Based on the test done below, its mileage with 0.6500129

numeric\_data<-dplyr::select\_if(final\_sorted\_data, is.numeric) # selecting only numeric variables  
numeric\_data<-cor(numeric\_data)  
n<-nrow(numeric\_data)  
# Getting the correlation between the dependent and the independent variables.  
abs(numeric\_data[1,2:(n)])

mileage tax mpg engineSize   
 0.6500129 0.5352730 0.5474798 0.3496872

1. Create a linear model to model this relationship.

**(2 marks)**

### Answer:

model1<-lm(price~mileage,data=final\_sorted\_data)  
summary(model1)

Call:  
lm(formula = price ~ mileage, data = final\_sorted\_data)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-9679.4 -2524.6 -85.9 2282.8 16626.7   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1.880e+04 1.502e+02 125.19 <2e-16 \*\*\*  
mileage -1.466e-01 3.833e-03 -38.23 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 3597 on 1998 degrees of freedom  
Multiple R-squared: 0.4225, Adjusted R-squared: 0.4222   
F-statistic: 1462 on 1 and 1998 DF, p-value: < 2.2e-16

anova(model1) # Coefficients w/inferential tests

Analysis of Variance Table  
  
Response: price  
 Df Sum Sq Mean Sq F value Pr(>F)   
mileage 1 1.8918e+10 1.8918e+10 1461.8 < 2.2e-16 \*\*\*  
Residuals 1998 2.5857e+10 1.2941e+07   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

coef(model1) # Coefficients (same as reg1)

(Intercept) mileage   
18800.1857368 -0.1465546

confint(model1)

2.5 % 97.5 %  
(Intercept) 18505.6814286 19094.6900450  
mileage -0.1540719 -0.1390373

1. Explain and interpret the model:

**(3 marks)**

### Answer:

The estimated average price of car is 18,800, but, for every increase in mileage, the price of the vehicle decreases by 14.660. Also, the R-squared value shows that mileage greatly affects price by 42.25%.

# Residuals(difference between the observed and predicted values)

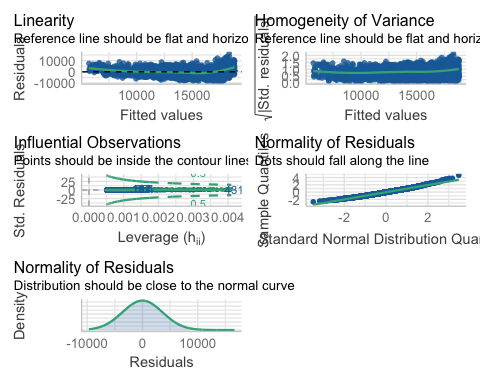
Min: -9679.4 is the furthest point below the regression line 1Q: 25% of the residuals are below -2524.6 Median: The media is -85.9 3Q: 25% of the residuals are greater than -2282.8 Max: 16626.7 is the furthest point above the regression line The Min and Max are not close in magnitude

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:

check\_model(model1)



1. Though the reference line doesnt start perfectly straight and flat, it begins to form a horizontal line showing linearity and homogeneity of variance
2. This model shows variables are normally distributed as dots fall along the lines and distribution close to the normal curve
3. All our observations are perfectly influential as all points fall inside the contour lines, therefore any deletion from would have an effect on the data set.

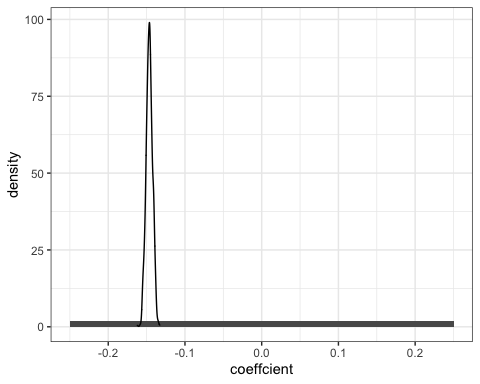
## Bootstrap

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

**(4 marks)**

### Answer:

n<- 1000   
mil <-rep(NA,n)  
  
set.seed(21050307)  
for(i in seq\_len(n)){  
 sample\_size\_data<-sample(seq\_along(final\_sorted\_data$mileage),size=length(final\_sorted\_data$mileage),replace=TRUE)   
 sample\_boot<- final\_sorted\_data[sample\_size\_data,]  
 model\_boot <- lm(price ~ mileage, data=sample\_boot)  
 mil[i] <- model\_boot$coef["mileage"]  
}  
  
# plotting a bootstrap histogram  
boot <- data.frame(coeffcient=mil)  
ggplot(boot, aes(x=coeffcient)) +  
geom\_histogram(aes(y = ..density..),binwidth = .5)+  
 geom\_density()+  
theme\_bw()



# getting the 95% bootstrap confidence interval  
quantile(mil,c(.025,0.975))

2.5% 97.5%   
-0.1546614 -0.1384399

# Question 4: Multivariable relationship (10 marks)

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

model2 <- lm(price ~ mileage + mpg, data = final\_sorted\_data)  
summary(model2)

Call:  
lm(formula = price ~ mileage + mpg, data = final\_sorted\_data)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-9422.4 -1967.8 32.7 1864.9 15174.0   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 3.143e+04 4.371e+02 71.92 <2e-16 \*\*\*  
mileage -1.268e-01 3.245e-03 -39.08 <2e-16 \*\*\*  
mpg -2.071e+02 6.868e+00 -30.16 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 2983 on 1997 degrees of freedom  
Multiple R-squared: 0.6032, Adjusted R-squared: 0.6028   
F-statistic: 1518 on 2 and 1997 DF, p-value: < 2.2e-16

model2 is a huge improvement on model1 as we see an increase in the multiple and adjusted-rsquares.

model3 <- lm(price ~ mileage + mpg + tax, data = final\_sorted\_data)  
summary(model3)

Call:  
lm(formula = price ~ mileage + mpg + tax, data = final\_sorted\_data)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-10191.4 -1875.2 47.7 1875.1 15193.1   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 2.891e+04 6.422e+02 45.006 < 2e-16 \*\*\*  
mileage -1.207e-01 3.418e-03 -35.318 < 2e-16 \*\*\*  
mpg -1.806e+02 8.437e+00 -21.402 < 2e-16 \*\*\*  
tax 7.240e+00 1.355e+00 5.345 1.01e-07 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 2962 on 1996 degrees of freedom  
Multiple R-squared: 0.6088, Adjusted R-squared: 0.6082   
F-statistic: 1035 on 3 and 1996 DF, p-value: < 2.2e-16

model3 is a slight improvement on model2 as we see very little increase in the multiple and adjusted-rsquares.

model4 <- lm(price ~ mileage + mpg + tax+ engineSize, data = final\_sorted\_data)  
summary(model4)

Call:  
lm(formula = price ~ mileage + mpg + tax + engineSize, data = final\_sorted\_data)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-10320.2 -1898.0 -15.6 1661.0 16821.8   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1.695e+04 1.025e+03 16.532 < 2e-16 \*\*\*  
mileage -1.316e-01 3.337e-03 -39.443 < 2e-16 \*\*\*  
mpg -1.137e+02 9.252e+00 -12.292 < 2e-16 \*\*\*  
tax 7.156e+00 1.289e+00 5.554 3.17e-08 \*\*\*  
engineSize 4.497e+03 3.096e+02 14.525 < 2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 2818 on 1995 degrees of freedom  
Multiple R-squared: 0.6462, Adjusted R-squared: 0.6455   
F-statistic: 911 on 4 and 1995 DF, p-value: < 2.2e-16

model4 shows a better but slight improvement on model3.

1. Explain and interpret the model:

**(4 marks)**

### Answer:

Model4 uses the price , mileage, mpg, tax and engine size. It is the most optimal model because it has the highest mutiple and adjusted r-squared. We are taking a very high r-squared because the high value shows that the variations in data of the dependent variable that can be explained by the independent variables. However due to the this, model4 has too many independent variable

model4 having p-value less than 0.05 means that there is significant relationship between variables in the model.

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)**

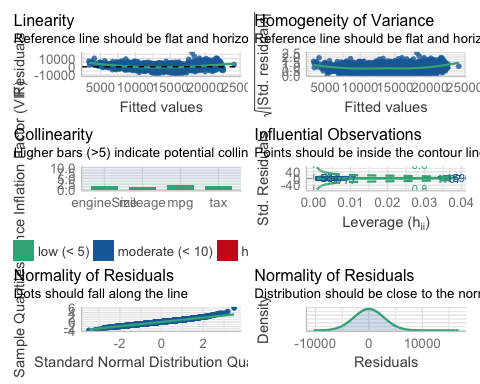
### Answer:

due to the high r-squared value, this model (model4) is suited for predicting what affects price of vehicles

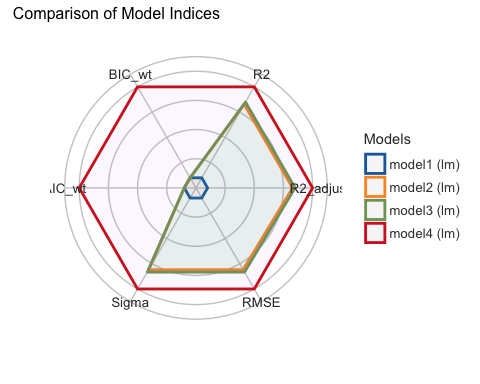
model 4 shows the following:

1. Though the reference line doesnt start perfectly straight and flat, it begines to form a horizontal line showing linearity and homogeneity of variance
2. Collinearity bars are less than 5. This indicates lower chances of multicollinearity.
3. This model shows variables are normally distributed as dots fall along the lines and distribution close to the normal curve
4. All our observations are perfectly influential as all points fall inside the contour lines, therefore any deletion from would have an effect on the data set.

check\_model(model4)



plot(compare\_performance(model1,model2, model3,model4, rank = TRUE))



anova ( model1, model4)

Analysis of Variance Table  
  
Model 1: price ~ mileage  
Model 2: price ~ mileage + mpg + tax + engineSize  
 Res.Df RSS Df Sum of Sq F Pr(>F)   
1 1998 2.5857e+10   
2 1995 1.5841e+10 3 1.0016e+10 420.47 < 2.2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

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

**(2 marks)**

### Answer:

1. there are perfect multicollinearity.
2. some identifying outliers seen can have an effect on the model
3. since there are no straight line relationship between the dependent and independent variables, the model foregoes variables that may be relevant

# Question 5: Model simplification (8 marks)

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

**(4 marks)**

### Answer:

for model simplification i would go through the following steps: Step 1: find/use maximum model. in this case, which is model1 Step 2: Remove non-significant explanatory variables. Step 3: Make list of valid terms to drop and drop least significant terms Step 4: Compare two models(using F test , anova()) and if its not significantly worse, go through Step 1, if it is remove next significant valid term if available and if not, use as the minimum adequate model

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

**(2 marks)**

### Answer:

* a simplified model produces a model that is better and most optimal , whether it has more independent variables or not.

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

**(2 marks)**

### Answer:

* The process is not straight forward and can cause more damage
* removing a term from the model can affect the model for the worst. This can cause r-squared value to be removed or not derived which affects the model and subsequently EDA.

# Question 6: Reporting (35 marks)

A client is looking to purchase a used Skoda Superb (registration year either 2018 or 2019, manual transmission, diesel 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 statistical model you would recommend, and why you have selected that statistical model.

Comment on any suggestions for alterations to the statistical 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:

Based on the previous reviews being made by analysts, it has been found that there are various factors that greatly affect the prices of certain cars more than others. It is erroneously believed that the brand and year of a car is the single most determining factor that affects the price of a car, but in the context of used cars, other determining factors should be considered to ensure a buyer gets the best value for their money. A lot of thought should go into buying a used car as there are factors that influence their performance and prices. For instance, when purchasing a new car, one does not have to worry about the mileage passed, but in purchasing a used car, buyers would typically want a car with less than 100,000miles on the odometer. The buyer should also consider the tax on the vehicle, as the engine size and year of the car would determine the amount of tax to be paid.

In relation to your choice of either a 2018 or 2019 Skoda Superb with manual transmission and a petrol engine, research indicates that the mileage, miles per gallon, year, the brand of car, and tax would all influence its price. Based on our analysis we defined the following: The price of a vehicle decreases when: - mileage and miles per gallon increases The price of a vehicle increases: - depending on the year the car was created. i.e. later models are more expensive – for instance, the BMW brand has more expensive cars while the Ford brand has more affordable cars. The newer a brand of car, the more expensive it is likely to be.

The Skoda Superb, however, is not included in the data given. Our model derived would in a degree of correctness, be able to make suggestions. With this, our model dictates as all cars, getting the Skoda Superb 2018 would be cheaper than the 2019 model as the lesser engine size found in 2018 would consume less fuel hence more miles per gallon. However, if finding an alternative is an option to be considered, a Ford would be preferable as the Ford brand has multiple options in the affordable region compared to other top brands, whilst giving you all the specifications you require. Logistic Regression is the statistical model that we have chosen to model our data on. I would recommend going with model2 (price ~ mileage + mpg) as model 3 and 4 is not a significant improvement on model 2 as model 2 is on model 1. Also, model 2 has lesser independent variables as indicated in our model above.

# 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: aarch64-apple-darwin20 (64-bit)  
Running under: macOS Monterey 12.4  
  
Matrix products: default  
BLAS: /Library/Frameworks/R.framework/Versions/4.1-arm64/Resources/lib/libRblas.0.dylib  
LAPACK: /Library/Frameworks/R.framework/Versions/4.1-arm64/Resources/lib/libRlapack.dylib  
  
locale:  
[1] en\_US.UTF-8/en\_US.UTF-8/en\_US.UTF-8/C/en\_US.UTF-8/en\_US.UTF-8  
  
attached base packages:  
[1] stats graphics grDevices utils datasets methods base   
  
other attached packages:  
 [1] corrplot\_0.92 plyr\_1.8.6 patchwork\_1.1.1 see\_0.6.8   
 [5] performance\_0.8.0 psych\_2.1.9 scales\_1.1.1 tidyr\_1.1.4   
 [9] stringr\_1.4.0 shiny\_1.7.1 rmarkdown\_2.11 rio\_0.5.29   
[13] plotly\_4.10.0 ggplot2\_3.3.5 lubridate\_1.8.0 httr\_1.4.2   
[17] ggvis\_0.4.7 boot\_1.3-28 dplyr\_1.0.7 pacman\_0.5.1   
  
loaded via a namespace (and not attached):  
 [1] splines\_4.1.1 jsonlite\_1.7.2 viridisLite\_0.4.0 tmvnsim\_1.0-2   
 [5] datawizard\_0.2.2 assertthat\_0.2.1 highr\_0.9 cellranger\_1.1.0   
 [9] yaml\_2.2.1 bayestestR\_0.11.5 pillar\_1.6.4 lattice\_0.20-45   
[13] glue\_1.6.0 digest\_0.6.29 promises\_1.2.0.1 colorspace\_2.0-2   
[17] Matrix\_1.4-0 htmltools\_0.5.2 httpuv\_1.6.5 pkgconfig\_2.0.3   
[21] haven\_2.4.3 purrr\_0.3.4 xtable\_1.8-4 openxlsx\_4.2.5   
[25] later\_1.3.0 tibble\_3.1.6 mgcv\_1.8-38 generics\_0.1.1   
[29] farver\_2.1.0 ellipsis\_0.3.2 withr\_2.4.3 lazyeval\_0.2.2   
[33] mnormt\_2.0.2 effectsize\_0.6.0 magrittr\_2.0.1 crayon\_1.4.2   
[37] readxl\_1.3.1 mime\_0.12 evaluate\_0.14 fansi\_1.0.2   
[41] nlme\_3.1-155 forcats\_0.5.1 foreign\_0.8-82 tools\_4.1.1   
[45] data.table\_1.14.2 hms\_1.1.1 lifecycle\_1.0.1 munsell\_0.5.0   
[49] zip\_2.2.0 compiler\_4.1.1 rlang\_0.4.12 grid\_4.1.1   
[53] parameters\_0.16.0 htmlwidgets\_1.5.4 labeling\_0.4.2 gtable\_0.3.0   
[57] curl\_4.3.2 R6\_2.5.1 knitr\_1.37 fastmap\_1.1.0   
[61] utf8\_1.2.2 insight\_0.15.0 stringi\_1.7.6 parallel\_4.1.1   
[65] Rcpp\_1.0.8 vctrs\_0.3.8 tidyselect\_1.1.1 xfun\_0.29