Logo

Description automatically generated

**College of Professional Studies**

**Northeastern University San Jose**

**MPS Analytics**

**Course: ALY6010 - Probability Theory and Introductory Statistics**

**Assignment:**

MODULE 5 PRACTICE ASSIGNMENT 5

**Submitted on:**

December 11, 2022

**Submitted to:**  **Submitted by:**

Professor: BEHZAD AHMADI NIKSHITA RANGANATHAN

# **INTRODUCTION**

Data analytics is important in any sector because it allows for businesses to make better decisions based on empirical evidence. Businesses can spot patterns and trends through data analysis that would otherwise go unnoticed. This allows businesses to optimize their operations, make better products and services, and identify new opportunities. Strategic growth in business is supported by the use of data analytics, which enables decision-makers to foresee consumer trends, improve productivity, and take measures based on reliable evidence. Many businesses, like the automotive industry, are using big data analytics more frequently. Big data analytics can be used to improve a number of aspects of the automotive industry, including sales, marketing, and service.

Regression analysis is a statistical method for figuring out the relations among two or more variables. It is used to determine the strength of the relationship between the variables and to assess the value of one variable based on the other. Dependent variables in regression are the variables that are predicted or explained by the model. Independent variables are those variables that are used to predict or describe the dependent variable.

In R, regression analysis can be performed using the lm() function. This function takes two arguments: a formula and a data frame. The formula specifies the relationship between the variables, while the data frame contains the data that will be used in the analysis. The lm() function returns a regression model object, which can then be used to make predictions or to analyze the results of the regression.

**About the dataset:**

Title of dataset – 100,000 UK Used Cars Data Set. I have considered only the Audi cars. The dataset has 10668 observations of car details with 9 attributes including model, year, price, transmission, mileage, fuelType, tax, mpg, and engineSize. The dataset contains numerical as well as categorical data (source-[*Kaggle*](https://www.kaggle.com/datasets/adityadesai13/used-car-dataset-ford-and-mercedes)*).*

**Purpose:**

The purpose of our project is to find the variables that affect used car prices and to predict and identify the strength of the relationship between the dependent (price) and independent variables. This project also aims to provide some insights and visualizations.

Below are the data descriptions of each variable that briefly describe the contents of the data set. The feature of the dataset is as follows:

|  |  |  |
| --- | --- | --- |
| **No** | **Feature** | **Dictionary** |
| 1 | model | Model of Audi car |
| 2 | year | Car registration year |
| 3 | price | Car price in £ |
| 4 | transmission | type of gearbox of the car (Manual / Automatic / Semi-Auto) |
| 5 | mileage | Car mileage/distance used |
| 6 | fuelType | Car engine fuel (Petrol / Diesel / Other) |
| 7 | tax | road tax paid for the car |
| 8 | mpg | Car miles per gallon |
| 9 | engineSize | Car Engine size in liters |

*Table 1: Features of the Used Car Data Set with their dictionary*

**ANALYSIS & INTERPRETATION**

* Importing the Audi CSV file

<cars> vector contains the information about all Audi used cars in UK

Graphical user interface, text, application

Description automatically generated

**Figure 1-read.csv()**

* Cleaning the dataset
  + N/A Values

gg\_miss\_which() and gg\_miss\_var() tell us about the missing values in the dataset.

miss\_var\_summary returns a summary table and a plot of the missing values in the data set. The summary table displays the percentage and number of missing values for every variable.

There are no NA and null values.



**Chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Table

Description automatically generated**

**A picture containing logo

Description automatically generated**

**Figure 2-<NA> values**

* Deleting duplicate rows

anyDuplicated() function in R will return a logical vector indicating which elements of a vector or data frame have duplicates. There are 103 duplicate rows.

**Table

Description automatically generated**

**Graphical user interface, text, application

Description automatically generated**

**Figure 3- duplicated() and anyDuplicated()**

* Changing datatypes

The data type of some variables is changed to factor using as.factor().

* Removing the outliers

Outliers can be caused by measurement errors, extreme values, or a subset of the data that is different from the rest of the data.

Data points outside the upper and lower quartiles are considered outliers in boxplots. In this case, we have only 1 extreme outlier in the price column.

Chart, box and whisker chart

Description automatically generated

**Figure 4- Outliers**

* Understanding the dataset

After data cleaning, there are 10564 observations and 9 features. The variables have the datatypes – numerical, integer, and factors.

**Text, letter

Description automatically generated**

**A screenshot of a computer

Description automatically generated with medium confidence**

**A screenshot of a computer

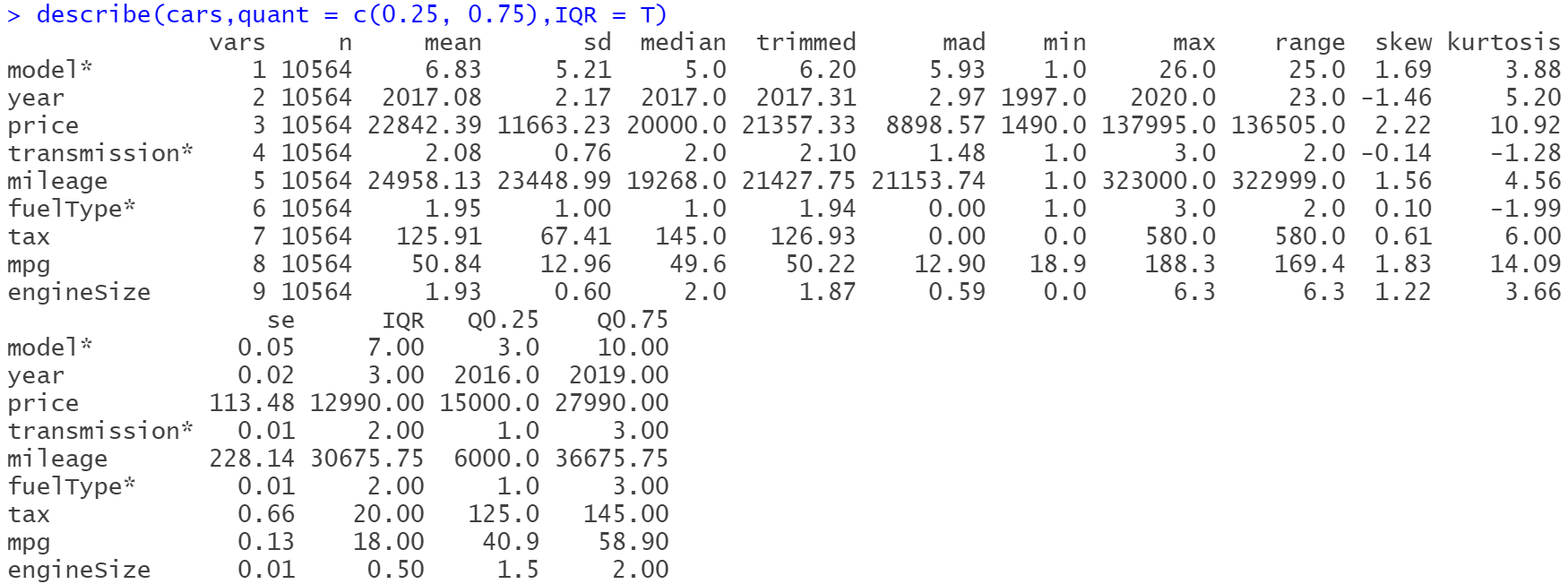
Description automatically generated with medium confidence**

**Figure 5 – str(),summary() and headTail()**

I have used the describe() function from the "PSYCH" package to get the descriptive statistics of the features in the data set.

I have included the quartiles value (0.25 and 0.75) and IQR(Interquartile range) in the describe function.

The mean of the price feature is around 22842.39£ with a standard deviation of 11663.23£. The lower Quartile value is 15000£ and the higher Quartile value is 27990£.

****The price column has a skew value greater than 1 means it is right-skewed and that the mean is greater than the median. A kurtosis value of 10.92 means data is highly peaked.

**Figure 6-describe()**

describeBy() returns the statistical summary of grouped data. In this case, it is by Transmission type (Automatic, Manual and Semi-Auto).

The mean price of automatic (28157.13£) is more than that of semi-auto (27100.93£) and manual (16024.31£).

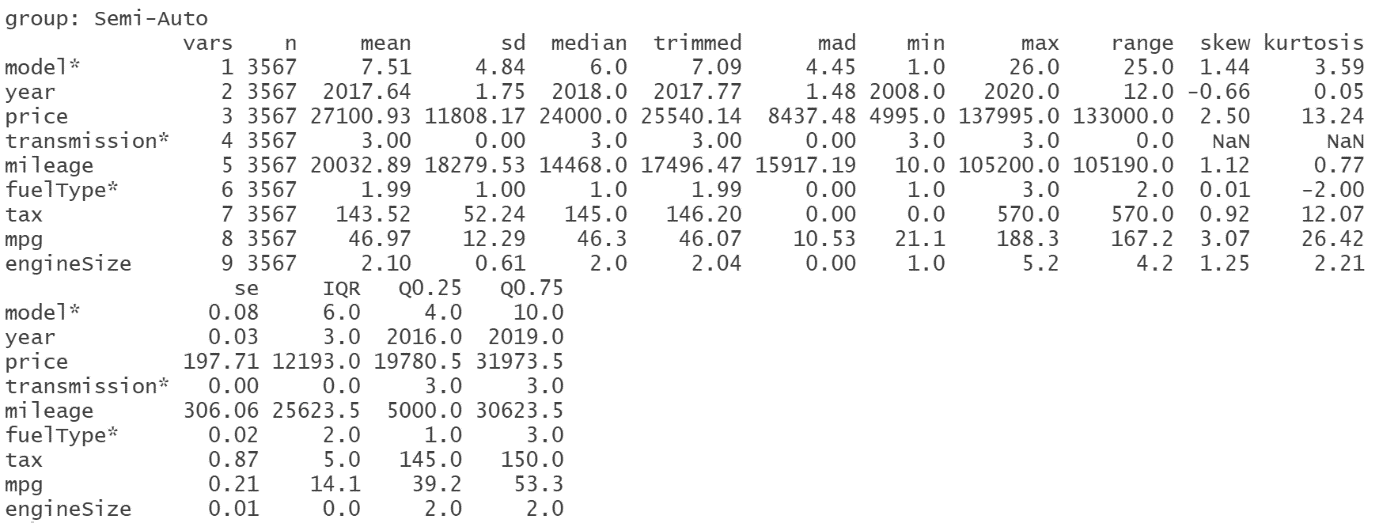
There are 2680, 4317, and 3567 observations in Automatic, Manual and Semi-auto categories respectively.

**A picture containing calendar

Description automatically generated**

**A picture containing text, receipt, screenshot

Description automatically generated**

****

**Figure 7-describeBy()**

* Normality tests
  + Kolmogorov-Smirnov test

This statistical test is used to examine whether a sample of data fits into a normal distribution. It works by comparing the cumulative distribution of the sample data to the cumulative distribution of a normal distribution. Shapiro test is appropriate for smaller sample sizes and the Kolmogorov test is for larger sample sizes.

The data is not normal because all of the tests have p values below 0.05.

**Text

Description automatically generated with medium confidence**

**Text

Description automatically generated**

**Text

Description automatically generated**

**Text

Description automatically generated**

**Text

Description automatically generated**

**Figure 8- Kolmogorov Smirnov test**

* + Density plots

The below plots does not follow a bell-shaped curve. Instead, the plot shows a skewed or bimodal distribution.

**Chart, waterfall chart

Description automatically generatedChart

Description automatically generated**

**Chart

Description automatically generatedChart

Description automatically generated**

**Chart

Description automatically generated**

**Figure 9- Density Plots**

**PART-1**

* Correlation
  + Correlation Table

This statistics table displays the relationships between the attributes. The correlation coefficient, which is a metric for measuring the strength of a relationship, is included in the table.

**Text

Description automatically generated with medium confidence**

**Figure 10- Correlation Table**

**Question -** A correlation chart is diagnostic and should not be larger than 5 variables for reporting purposes. Why is this?

To determine relationships between variables, a correlation chart is used. When there are too many variables, it can be hard to identify meaningful relationships. Additionally, having too many variables can make the chart visually cluttered and difficult to read. Therefore, it is best practice to limit the number of variables in a correlation chart to no more than five for reporting purposes.

A correlogram is a visual representation of a dataset's correlation matrix.

The correlation coefficient ranges from -1 to 1, where:

* -1 to -0.8: Very Strong negative
* -0.79 to -0.6: Strong negative
* -0.59 to -0.4: Moderate negative
* -0.39 to -0.2: Weak negative
* -0.19 to -0.01: Very weak negative
* 0 to 0.19: Very weak positive
* 0.2 to 0.39: Weak positive
* 0.4 to 0.59: Moderate positive
* 0.6 to 0.79: Strong positive
* -1 to -0.8: Very Strong positive

Price has a weak positive relationship with Tax (Correlation Coefficient – 0.36)

Price has a moderate positive relationship with Engine Size (Correlation Coefficient – 0.59)

Price has a moderate negative relationship with Mileage (Correlation Coefficient - -0.54)

Price has a strong negative relationship with Miles per gallon (Correlation Coefficient - -0.6)

A picture containing diagram

Description automatically generated

**Figure 11- corrgram()**

A chart of a correlation matrix can be shown using the method chart.Correlation() [in the package PerformanceAnalytics.

Diagram

Description automatically generated

**Figure 12- chart.correlation()**

**PART-2**

* Regression Models

**Question -** How does regression analysis differ from correlation analysis?

Correlation analysis measures the strength of the relationship between two variables, whereas regression analysis is used to predict the value of a dependent variable. The purpose of correlation is not to make predictions about the value of one variable based on the value of another.

**Model 1 – Price with tax**

**Chart, scatter chart

Description automatically generated**

**Text, letter

Description automatically generated**

The R2 value of 0.12 indicates that the model explains 12.75% of the variance in the data. This is a relatively low R2 value, which suggests that the model is not a good fit.

**Chart, histogram

Description automatically generatedDiagram

Description automatically generated**

**Chart

Description automatically generated with low confidenceChart

Description automatically generated**

**Figure 13- Regression Model 1**

**Model 2 – Price with mileage**

**Chart, scatter chart

Description automatically generated**

**Text, letter

Description automatically generated**

The R2 value of 0.2878 indicates that 28.78% of the variation in the data can be explained by the model. This is a relatively low R2 value, which means that the model is not a very good fit for the data.

**Chart

Description automatically generated**

**Chart, line chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Chart

Description automatically generated with medium confidence**

**Figure 14- Regression Model 2**

**Model 3 – Price with engine size**

**Chart, line chart, scatter chart

Description automatically generated**

**Text, letter

Description automatically generated**

**Chart

Description automatically generated**The R2 value of 0.352 indicates that approximately 35.2% of the variability in the data can be explained by the model.

**Chart

Description automatically generated**

**A picture containing diagram

Description automatically generatedChart

Description automatically generated**

**Figure 15- Regression Model 3**

**Model 4 – Price with Miles per gallon**

**Chart, scatter chart

Description automatically generated**

**Text, letter

Description automatically generated**

The R2 value of 0.3597 indicates that the model is only able to explain approximately 36% of the variance in the data. This suggests that the model is not a very good fit for the data.

**Chart, histogram

Description automatically generatedChart, scatter chart

Description automatically generated**

**Chart, line chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Figure 15- Regression Model 4**

**Model 5 – Price with Miles per gallon, Mileage, Engine size, and Tax**

**Text

Description automatically generated**

The R2 value of 0.7243 indicates that the model explains 72.43% of the variance in the data. This is considered a good fit.

**Chart, line chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated**

**Figure 16- Regression Model 5**

**CONCLUSION**

In this assignment, I initially performed normality tests, then plotted the correlation matrix and used regression models to determine the factors that influence the price of the cars.

The key findings are :

* The variables are not normally distributed, according to Kolmogorov-Smirnov tests.
* To evaluate the characteristics' visual normality, density plots can be used. It is evident that none of the features exhibit a normal distribution.
* Using a correlation matrix, we could comprehend the relationship between features.
* The R-square value of the final model is 0.7243 which explains that other variables in the model are strongly related to the dependent variable (price).

**REFERENCES**

Bluman, A. G. (2018). Elementary Statistics, 10th ed. McGraw Hill.

Kabacoff, R. I. (2011). R in action: Data analysis and graphics with R. Manning Publications Co.

Zach. (2021, February 16). How to Add a Regression Equation to a Plot in R. Statology. <https://www.statology.org/add-regression-equation-to-plot-in-r/>

**APPENDIX: CODE**

#---------------------- Week\_5\_Module\_5 R Script ----------------------#

print("Author : Nikshita Ranganathan")

print("Week 4 Assignment - Module 4 R Practice")

print("Course Name - ALY6010: Probability Theory and Introductory Statistics")

# Importing the dataset

getwd()

cars<-read.csv("audi.csv")

# Loading the libraries

library(psych)

library(naniar)

library(ggpubr)

library(dplyr)

library(skimr)

library(corrplot)

library(corrgram)

library(ggplot2)

library(ochRe)

library(PerformanceAnalytics)

library(wesanderson)

library(gridExtra)

library(Hmisc)

#------------------- Data Cleaning -------------------#

# Visualization and Checking on N/A values

gg\_miss\_which(cars)

gg\_miss\_var(cars)

miss\_var\_summary(cars)

sum(is.na(cars))

sum(is.null(cars))

# Checking duplicate values and removing it

duplicated(cars)

anyDuplicated(cars)

cars<-cars[!duplicated(cars), ]

anyDuplicated(cars)

# Changing the datatypes

cars$transmission<-as.factor(cars$transmission)

cars$fuelType<-as.factor(cars$fuelType)

cars$model<-as.factor(cars$model)

# checking outliers

boxplot(cars$price)

cars<-subset(cars,price!=145000)

#------------------- Exploratory Data Analysis -------------------#

# Understanding cars dataset

str(cars)

summary(cars)

headTail(cars)

describe(cars,quant = c(0.25, 0.75),IQR = T)

describeBy(cars,group=cars$transmission,quant = c(0.25, 0.75),IQR = T)

# Kolmogorov-Smirnov test

ks.test(cars$price,"pnorm")

ks.test(cars$mileage,"pnorm")

ks.test(cars$tax,"pnorm")

ks.test(cars$mpg,"pnorm")

ks.test(cars$engineSize,"pnorm")

# Density Plots

ggplot(cars, aes(x = `price`, fill = `transmission`)) +geom\_density(alpha = 0.6) +scale\_fill\_ochre(palette="williams\_pilbara")+xlab("Price")+ggtitle("Density plot A")

ggplot(cars, aes(x = `mileage`, fill = `transmission`)) +geom\_density(alpha = 0.6) +scale\_fill\_ochre(palette="williams\_pilbara")+xlab("Mileage")+ggtitle("Density plot B")

ggplot(cars, aes(x = `mpg`, fill = `transmission`)) +geom\_density(alpha = 0.6) +scale\_fill\_ochre(palette="williams\_pilbara")+xlab("Miles per gallon")+ggtitle("Density plot C")

ggplot(cars, aes(x = `tax`, fill = `transmission`)) +geom\_density(alpha = 0.6) +scale\_fill\_ochre(palette="williams\_pilbara")+xlab("Tax")+ggtitle("Density plot D")

ggplot(cars, aes(x = `engineSize`, fill = `transmission`)) +geom\_density(alpha = 0.6) +scale\_fill\_ochre(palette="williams\_pilbara")+xlab("Engine Size")+ggtitle("Density plot E")

# Correlation

cor<-cars %>% select(price,mileage,tax,mpg,engineSize)

table = rcorr(as.matrix(cor))

table

corrgram(cor, order=TRUE, lower.panel=panel.conf, upper.panel=panel.pie, text.panel=panel.txt,main="Correlogram")

chart.Correlation(cor, histogram=TRUE, pch="+")

# Models

cars %>% ggplot(aes(y=tax,x=price,color=transmission))+geom\_point(size = 0.7) +geom\_smooth(method="lm",se=FALSE,fullrange=TRUE,size = 0.5, color="black") +scale\_color\_brewer(palette = "Dark2")+ylim(0, 1000)+stat\_regline\_equation(label.y = 1000, aes(label = ..eq.label..))+ stat\_regline\_equation(label.y = 950, aes(label = ..rr.label..)) +facet\_wrap(~transmission)+theme(legend.position = "none")

model1<-lm(price~tax,data=cars)

summary(model1)

plot(model1)

cars %>% ggplot(aes(y=mileage,x=price,color=transmission))+geom\_point(size = 0.7) +geom\_smooth(method="lm",se=FALSE,fullrange=TRUE,size = 0.5,color="black")

+ylim(0, 175000)+scale\_color\_brewer(palette = "Set1")+stat\_regline\_equation(label.y = 175000, aes(label = ..eq.label..))+ stat\_regline\_equation(label.y = 165000, aes(label = ..rr.label..))+facet\_wrap(~transmission)+theme(legend.position = "none")

model2<-lm(price~mileage,data=cars)

summary(model2)

plot(model2)

cars %>% ggplot(aes(y=engineSize,x=price,color=transmission))+geom\_point(size = 0.7) +geom\_smooth(method="lm",se=FALSE,fullrange=TRUE,size = 0.5,color="black")

+ylim(0, 8)+scale\_color\_manual(values= wes\_palette("Darjeeling1", n = 3))

+stat\_regline\_equation(label.y = 8, aes(label = ..eq.label..))+ stat\_regline\_equation(label.y = 7.5, aes(label = ..rr.label..))+facet\_wrap(~transmission)+theme(legend.position = "none")

model3<-lm(price~engineSize,data=cars)

summary(model3)

plot(model3)

cars %>% ggplot(aes(y=mpg,x=price,color=transmission))+geom\_point(size = 0.7)+ylim(0, 300)+geom\_smooth(method="lm",se=FALSE,fullrange=TRUE,size = 0.5,color="black")+scale\_color\_manual(values= wes\_palette("GrandBudapest1", n = 3))+stat\_regline\_equation(label.y = 300, aes(label = ..eq.label..))+ stat\_regline\_equation(label.y = 280, aes(label = ..rr.label..))+facet\_wrap(~transmission)+theme(legend.position = "none")

model4<-lm(price~mpg,data=cars)

summary(model4)

plot(model4)

model5<-lm(price~mpg+engineSize+mileage+tax,data=cars)

summary(model5)

plot(model5)