**Himanshu Prajapati**

Write a simple R script to execute the following data preprocessing and statistical analysis. Where required show analytical output and interpretations.

**Preprocessing**

1. Load into an object the data from the "Craigs List Cars" worksheet in the " Cars list Data.xlsx " spreadsheet file. This file contains information on 46,484 vehicles in the United States listed for sale on Craig’s List. This is the master data set.

**rm(list=ls())**

**library(rio)**

**Cr\_list\_cars = import("Cars list Data.xlsx ")**

**colnames(Cr\_list\_cars) = tolower(make.names(colnames(Cr\_list\_cars)))**

**attach((Cr\_list\_cars))**

1. Using the numerical portion of the U number as a random number seed, select a random sample of n=200 autos from the master data set. This will be your primary data set.

**set.seed(56390726)**

**my\_primary\_data\_set = Cr\_list\_cars[sample(1:nrow(Cr\_list\_cars),200,replace=FALSE),]**

**Analysis**

Using your primary data set:

1. Conduct a simple regression analysis using your sample data with the dependent variable being "price" and the independent variable being "odometer".

**p\_out = lm(price~odometer,data=my\_primary\_data\_set)**

**summary(p\_out)**

**Call:**

**lm(formula = price ~ odometer, data = my\_primary\_data\_set)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-20928 -6221 -1250 3423 41547**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 2.192e+04 1.317e+03 16.647 < 2e-16 \*\*\***

**odometer -8.207e-02 1.131e-02 -7.258 8.73e-12 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 8986 on 198 degrees of freedom**

**Multiple R-squared: 0.2101, Adjusted R-squared: 0.2062**

**F-statistic: 52.68 on 1 and 198 DF, p-value: 8.735e-12**

1. Give verbal interpretations of all beta coefficients in your regression model. Make certain the language you use is understandable to a reasonably competent lay person shopping for a car on Craig's List.

**From the above summary, we can say that The intercept represents the estimated price of a vehicle when its odometer reading is exactly zero.  
Odometer** **-8.207e-02 ,the estimated price of the vehicle decreases by approximately 8.2 cents. So, as the mileage on the vehicle increases, the estimated price tends to decrease.**

**And the model explains approximately 21.01% of the variability in vehicle prices. The relationship between odometer reading and price is statistically significant (p-value < 0.001), suggesting that mileage is a meaningful predictor of vehicle price in our dataset.**

**We can reject the null hypothesis as *p value* is extremely smaller.**

**The t-value is approximately -7.258 for the "odometer" variable or T value for slope and intercept are high, we can reject null hypothesis**

**Multiple R squared is 0.2101 and Residual Standard Error: 8986**

1. Evaluate and interpret both the p value and the confidence interval on the "odometer" coefficient in your regression model.

**confint(p\_out)**

|  |
| --- |
| **confint(p\_out)**  **2.5 % 97.5 %**  **(Intercept) 19325.8185050 2.451965e+04**  **odometer -0.1043631 -5.976815e-02** |
|  |
| |  | | --- | |  | |

1. Run appropriate diagnostics on your regression model to determine if it is in conformity with the LINE assumptions of regression.

**#Linearity**

**plot(my\_primary\_data\_set$price,p\_out$fitted.values,pch=19,**

**main = "Between Actual price and Fitted")**

**abline(0,1,col="red", lwd=3)**

****

**It looks like model is not so linear**

**#Normality**

**qqnorm(p\_out$residuals,pch=19, main="O&G Normality Plot")**

**qqline(p\_out$residuals,col="red",lwd=3)**

****

**#Hist**

**hist(p\_out$residuals,col="red",**

**main="Histogram of Residuals")**

****

**#Equality of Variances**

**plot(p\_out$fitted.values,scale(p\_out$residuals),**

**pch=19,**

**main="price and Odometer St residual")**

**abline(0,0,col="red",lwd=3)**

****

1. Ms. Trayla Parks is considering offering her Toyota for sale on Craig's List. The Toyota currently has an odometer reading of 78,521 miles. Use your regression model to predict the price of the vehicle on Craig's List. Determine and verbally interpret the appropriate confidence interval on this prediction. If Trayla kept her vehicle one more year until the odometer showed 98,000 by how much would your model predict the price of her car would change?

**> newdata=data.frame(odometer=78521)**

**> predict(p\_out,newdata,interval="predict")**

**fit lwr upr**

**1 15478.86 -2294.104 33251.82**

**> predict(p\_out,newdata,interval="confidence")**

**fit lwr upr**

**1 15478.86 14120.73 16836.99**

**According to the model, the cost the car with an odometer reading of 78,521 would be $15,478.86 which is in between the range of $14120.73 and $16836.99**

**If we want to do for 98000**

**> predict(p\_out,newdata,interval="predict")**

**fit lwr upr**

**1 13880.3 -3885.166 31645.77**

**> predict(p\_out,newdata,interval="confidence")**

**fit lwr upr**

**1 13880.3 12624.05 15136.56**

**There is drop in price of the car if we changes the value of odometer.**