Pablo Salar Carrera

MSD325

Simple Linear Regression Project

1. **Using R**, explore the relationships between two variables from your dataset. Select one variable as dependent (that makes sense) and one as independent and explore whether there is any linear relationship. Investigate this for more than one pairs of variables.

```
Call:
lm(formula = CarPrice$Price ~ CarPrice$Mileage, data = CarPrice)

Residuals:
   Min     1Q Median     3Q    Max
-19179 -11985  -3767   4803 853782

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.920e+04  1.832e+02  104.79   <2e-16 ***
CarPrice$Mileage -1.370e-02  7.985e-04  -17.16   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19650 on 19202 degrees of freedom
Multiple R-squared:  0.0151,    Adjusted R-squared:  0.01505
F-statistic: 294.4 on 1 and 19202 DF,  p-value: < 2.2e-16
```

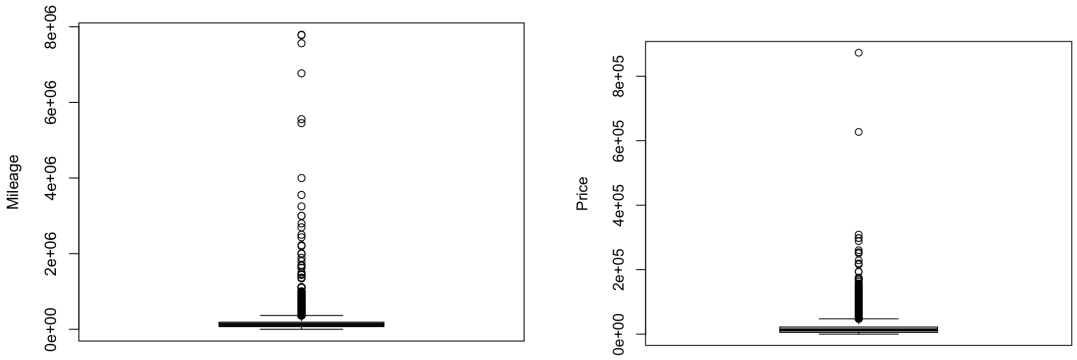
```
Call:
lm(formula = CarPrice$Price ~ CarPrice$Cylinders, data = CarPrice)

Residuals:
   Min     1Q Median     3Q    Max
-30167 -11844  -3645   5405 849767

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          9198.2      560.7   16.41   <2e-16 ***
CarPrice$Cylinders   1747.7      118.3   14.77   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19690 on 19202 degrees of freedom
Multiple R-squared:  0.01123,   Adjusted R-squared:  0.01118
F-statistic: 218.1 on 1 and 19202 DF,  p-value: < 2.2e-16

Call:
lm(formula = CarPrice$Price ~ CarPrice$Engine.volume, data = CarPrice)

Residuals:
   Min     1Q Median     3Q    Max
-59727 -11463  -3495   5315 850713

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             10350.0      398.3   25.98   <2e-16 ***
CarPrice$Engine.volume   2970.7      161.3   18.42   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19620 on 19202 degrees of freedom
Multiple R-squared:  0.01737,   Adjusted R-squared:  0.01731
F-statistic: 339.3 on 1 and 19202 DF,  p-value: < 2.2e-16
```

Yes, there is a linear relationship between all the pairs of variables.

2. Interpret the **outcomes of each analysis** you performed in step #1. Identify any outliers that your data might have and explain how you will deal with these outliers in your analyses. Apply a simple linear regression model and interpret the outcomes (e.g., model fit and comparison, outputs/coefficients, hypothesis tests, statistical significance, etc.).

From the outcomes shown in step #1, all three linear regressions have a p-value<2.2e-16, meaning they are statistically significant. Moreover, all variables were statistically significant. The highest R2 (Multiple R-squared) was given by the linear regression of the variable engine volume as the predictor and the variable price as the outcome. However, these values are really low, under 0.1, meaning there are better models to fit our data.

Doing some boxplots, I found some outliers in the price and mileage variables:



The next steps I will have to make are getting rid of the values larger than 6e+5 and the mileage values larger than 4e+6, so I will get a better linear regression model.

3. Based on your interpretations from step #3, identify which pair of variables can be used for a simple linear regression model in your dataset.

From the previous R results, all pairs of variables could be used for a simple linear regression model since all the variables are statistically significant with p-values<2.2e-16.