Simple linear regression

Group 9

M6 ICA1

Introduction

This group assignment will use the Cars93 dataset from the MASS package and is inspired from the Lab 8 written by CMU's Professor Alexandra Chouldechova for her Programming in R for Analytics course.

Packages to be used:

1. `MASS`
2. `tidyverse`
3. `broom`

```
library(MASS)
library(tidyverse)
library(broom)
```

Data

The Cars93 dataset consists of data from 93 cars on sale in the USA in 1993. It has 27 variables.

```
attach(MASS::Cars93)
Cars93 <- as_tibble(MASS::Cars93)
glimpse(Cars93)
```

```
Rows: 93
Columns: 27
$ Manufacturer       <fct> Acura, Acura, Audi, Audi, BMW, Buick, Buick, Buick,~
$ Model              <fct> Integra, Legend, 90, 100, 535i, Century, LeSabre, R~
$ Type               <fct> Small, Midsize, Compact, Midsize, Midsize, Midsize,~
$ Min.Price          <dbl> 12.9, 29.2, 25.9, 30.8, 23.7, 14.2, 19.9, 22.6, 26.~
$ Price              <dbl> 15.9, 33.9, 29.1, 37.7, 30.0, 15.7, 20.8, 23.7, 26.~
$ Max.Price          <dbl> 18.8, 38.7, 32.3, 44.6, 36.2, 17.3, 21.7, 24.9, 26.~
$ MPG.city           <int> 25, 18, 20, 19, 22, 22, 19, 16, 19, 16, 16, 25, 25,~
$ MPG.highway        <int> 31, 25, 26, 26, 30, 31, 28, 25, 27, 25, 25, 36, 34,~
$ AirBags            <fct> None, Driver & Passenger, Driver only, Driver & Pas~
$ DriveTrain         <fct> Front, Front, Front, Front, Rear, Front, Front, Rea~
$ Cylinders          <fct> 4, 6, 6, 6, 4, 4, 6, 6, 6, 8, 8, 4, 4, 6, 4, 6, 6, ~
$ EngineSize         <dbl> 1.8, 3.2, 2.8, 2.8, 3.5, 2.2, 3.8, 5.7, 3.8, 4.9, 4~
$ Horsepower         <int> 140, 200, 172, 172, 208, 110, 170, 180, 170, 200, 2~
$ RPM                <int> 6300, 5500, 5500, 5500, 5700, 5200, 4800, 4000, 480~
$ Rev.per.mile       <int> 2890, 2335, 2280, 2535, 2545, 2565, 1570, 1320, 169~
$ Man.trans.avail    <fct> Yes, Yes, Yes, Yes, Yes, No, No, No, No, No, No, Ye~
$ Fuel.tank.capacity <dbl> 13.2, 18.0, 16.9, 21.1, 21.1, 16.4, 18.0, 23.0, 18.~
$ Passengers         <int> 5, 5, 5, 6, 4, 6, 6, 6, 5, 6, 5, 5, 5, 4, 6, 7, 8, ~
$ Length             <int> 177, 195, 180, 193, 186, 189, 200, 216, 198, 206, 2~
$ Wheelbase          <int> 102, 115, 102, 106, 109, 105, 111, 116, 108, 114, 1~
$ Width              <int> 68, 71, 67, 70, 69, 69, 74, 78, 73, 73, 74, 66, 68,~
$ Turn.circle        <int> 37, 38, 37, 37, 39, 41, 42, 45, 41, 43, 44, 38, 39,~
$ Rear.seat.room     <dbl> 26.5, 30.0, 28.0, 31.0, 27.0, 28.0, 30.5, 30.5, 26.~
$ Luggage.room       <int> 11, 15, 14, 17, 13, 16, 17, 21, 14, 18, 14, 13, 14,~
$ Weight             <int> 2705, 3560, 3375, 3405, 3640, 2880, 3470, 4105, 349~
$ Origin             <fct> non-USA, non-USA, non-USA, non-USA, non-USA, USA, U~
$ Make               <fct> Acura Integra, Acura Legend, Audi 90, Audi 100, BMW~
```

```
levels(AirBags)
```

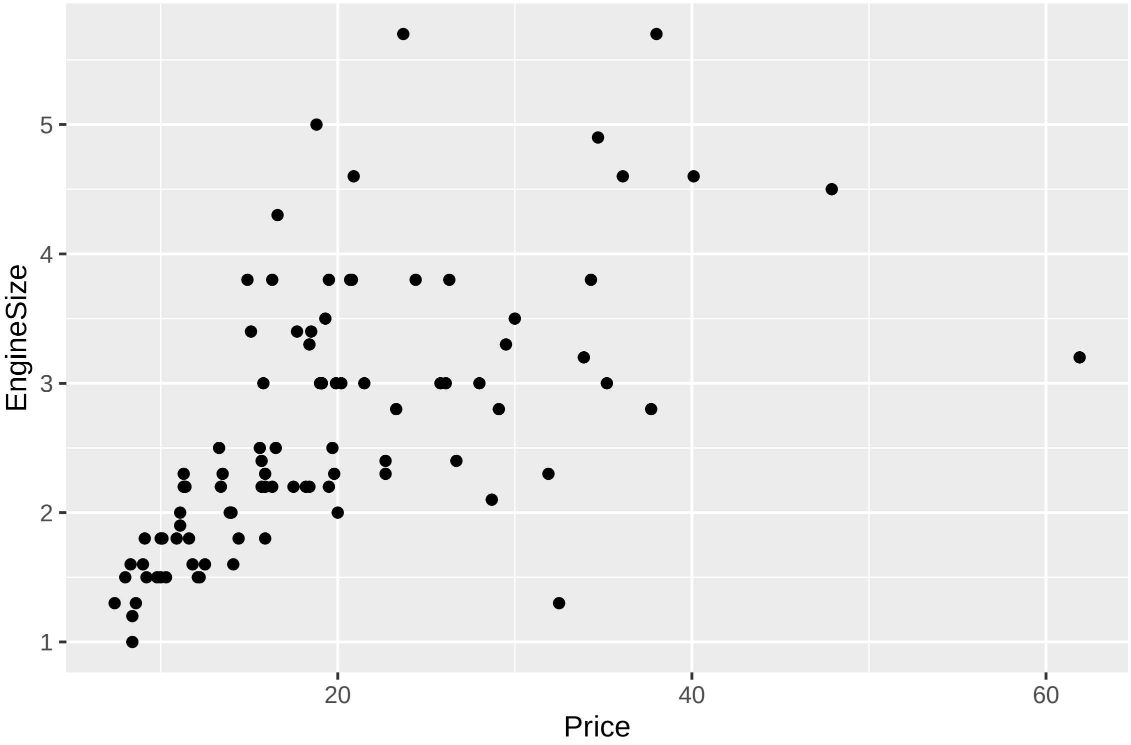
```
[1] "Driver & Passenger" "Driver only"        "None"
```

The variables

```
Rows: 93
Columns: 27
$ Manufacturer       <fct> Acura, Acura, Audi, Audi, BMW, Buick, Buick, Buick,~
$ Model              <fct> Integra, Legend, 90, 100, 535i, Century, LeSabre, R~
$ Type               <fct> Small, Midsize, Compact, Midsize, Midsize, Midsize,~
$ Min.Price          <dbl> 12.9, 29.2, 25.9, 30.8, 23.7, 14.2, 19.9, 22.6, 26.~
$ Price              <dbl> 15.9, 33.9, 29.1, 37.7, 30.0, 15.7, 20.8, 23.7, 26.~
$ Max.Price          <dbl> 18.8, 38.7, 32.3, 44.6, 36.2, 17.3, 21.7, 24.9, 26.~
$ MPG.city           <int> 25, 18, 20, 19, 22, 22, 19, 16, 19, 16, 16, 25, 25,~
$ MPG.highway        <int> 31, 25, 26, 26, 30, 31, 28, 25, 27, 25, 25, 36, 34,~
$ AirBags            <fct> None, Driver & Passenger, Driver only, Driver & Pas~
$ DriveTrain         <fct> Front, Front, Front, Front, Rear, Front, Front, Rea~
$ Cylinders          <fct> 4, 6, 6, 6, 4, 4, 6, 6, 6, 8, 8, 4, 4, 6, 4, 6, 6, ~
$ EngineSize         <dbl> 1.8, 3.2, 2.8, 2.8, 3.5, 2.2, 3.8, 5.7, 3.8, 4.9, 4~
$ Horsepower         <int> 140, 200, 172, 172, 208, 110, 170, 180, 170, 200, 2~
$ RPM                <int> 6300, 5500, 5500, 5500, 5700, 5200, 4800, 4000, 480~
$ Rev.per.mile       <int> 2890, 2335, 2280, 2535, 2545, 2565, 1570, 1320, 169~
$ Man.trans.avail    <fct> Yes, Yes, Yes, Yes, Yes, No, No, No, No, No, No, Ye~
$ Fuel.tank.capacity <dbl> 13.2, 18.0, 16.9, 21.1, 21.1, 16.4, 18.0, 23.0, 18.~
$ Passengers         <int> 5, 5, 5, 6, 4, 6, 6, 6, 5, 6, 5, 5, 5, 4, 6, 7, 8, ~
$ Length             <int> 177, 195, 180, 193, 186, 189, 200, 216, 198, 206, 2~
$ Wheelbase          <int> 102, 115, 102, 106, 109, 105, 111, 116, 108, 114, 1~
$ Width              <int> 68, 71, 67, 70, 69, 69, 74, 78, 73, 73, 74, 66, 68,~
$ Turn.circle        <int> 37, 38, 37, 37, 39, 41, 42, 45, 41, 43, 44, 38, 39,~
$ Rear.seat.room     <dbl> 26.5, 30.0, 28.0, 31.0, 27.0, 28.0, 30.5, 30.5, 26.~
$ Luggage.room       <int> 11, 15, 14, 17, 13, 16, 17, 21, 14, 18, 14, 13, 14,~
$ Weight             <int> 2705, 3560, 3375, 3405, 3640, 2880, 3470, 4105, 349~
$ Origin             <fct> non-USA, non-USA, non-USA, non-USA, non-USA, USA, U~
$ Make               <fct> Acura Integra, Acura Legend, Audi 90, Audi 100, BMW~
```

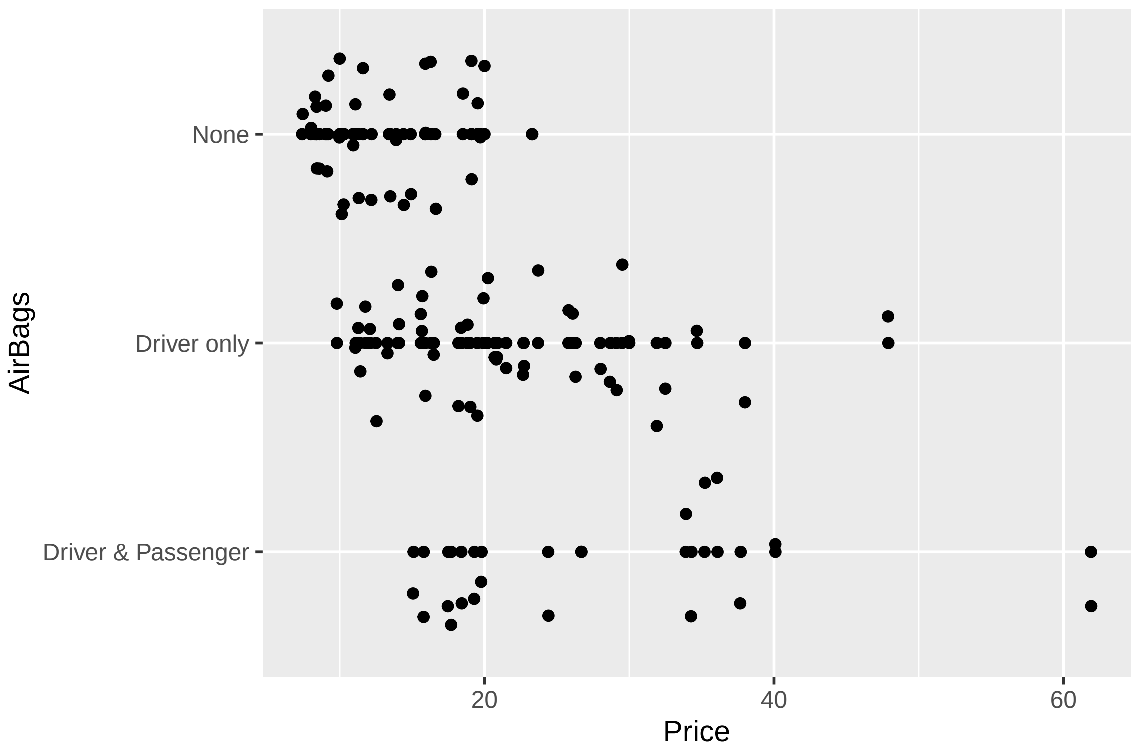
Exploratory data analysis

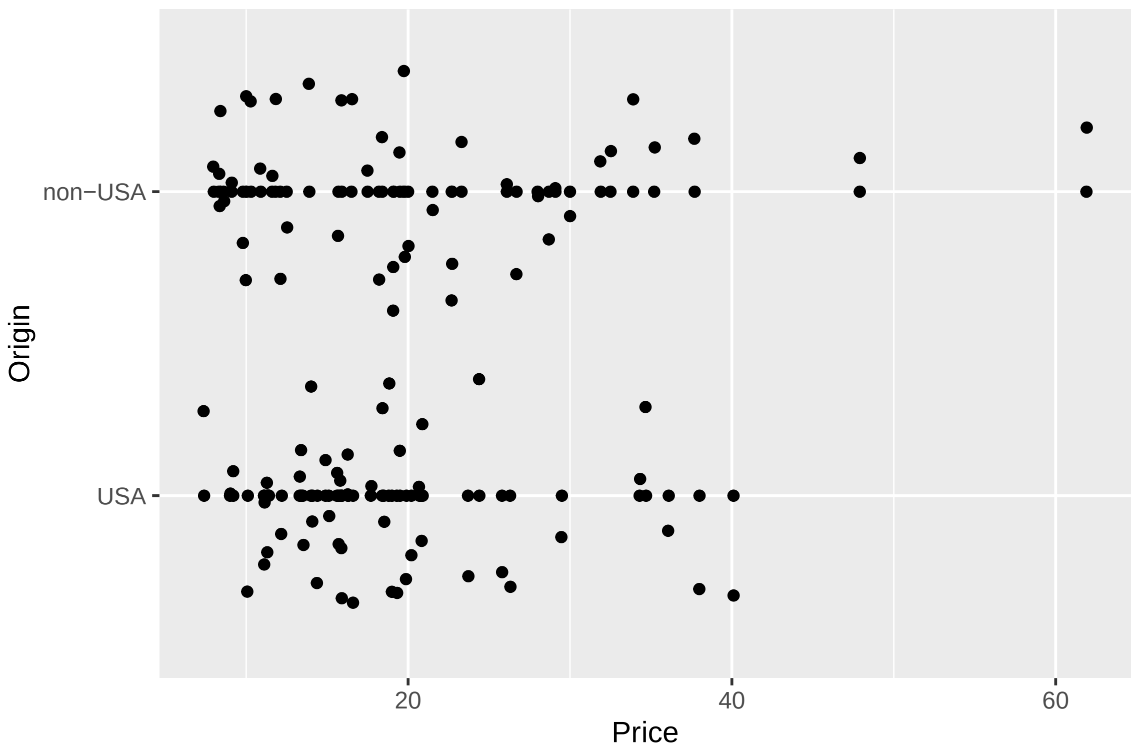
Question 1 Create a scatter plot between `Price` and the `EngineSize`. Comment on the relationship. What do you think the correlation is between the variables?



As price increases engine size also increases. However, there is greater variation between engine sizes as price increases. There is a positive correlation between both variables.

Question 2 Create scatter plots between `Price` and categorical variables `Airbags` and `Origin`. Comment on the relationship. Add `geom_jitter()` to your plot. What do you notice about the variability in `Price` between different levels?





Generally, as the price increases the variability at all levels increases. In regards to the variability between levels, non-USA (in the origin plot) and driver and passenger (in the airbag plot) have more variability due to outlines present in the graph.

Model fit

Question 3 Fit a linear model with function lm() between variables `Price` and `EngineSize`. Save it as an object mod.fit. What are the values of b_0 and b_1? Write out the model. *Hint*: `tidy()`

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)   1.42     0.196        7.21 1.58e-10
2 Price         0.0642   0.00903      7.11 2.59e-10
```

EngineSize =1.416+ 0.0642 x Price b_0= 1.416 b_1= 0.0642 #### Question 4

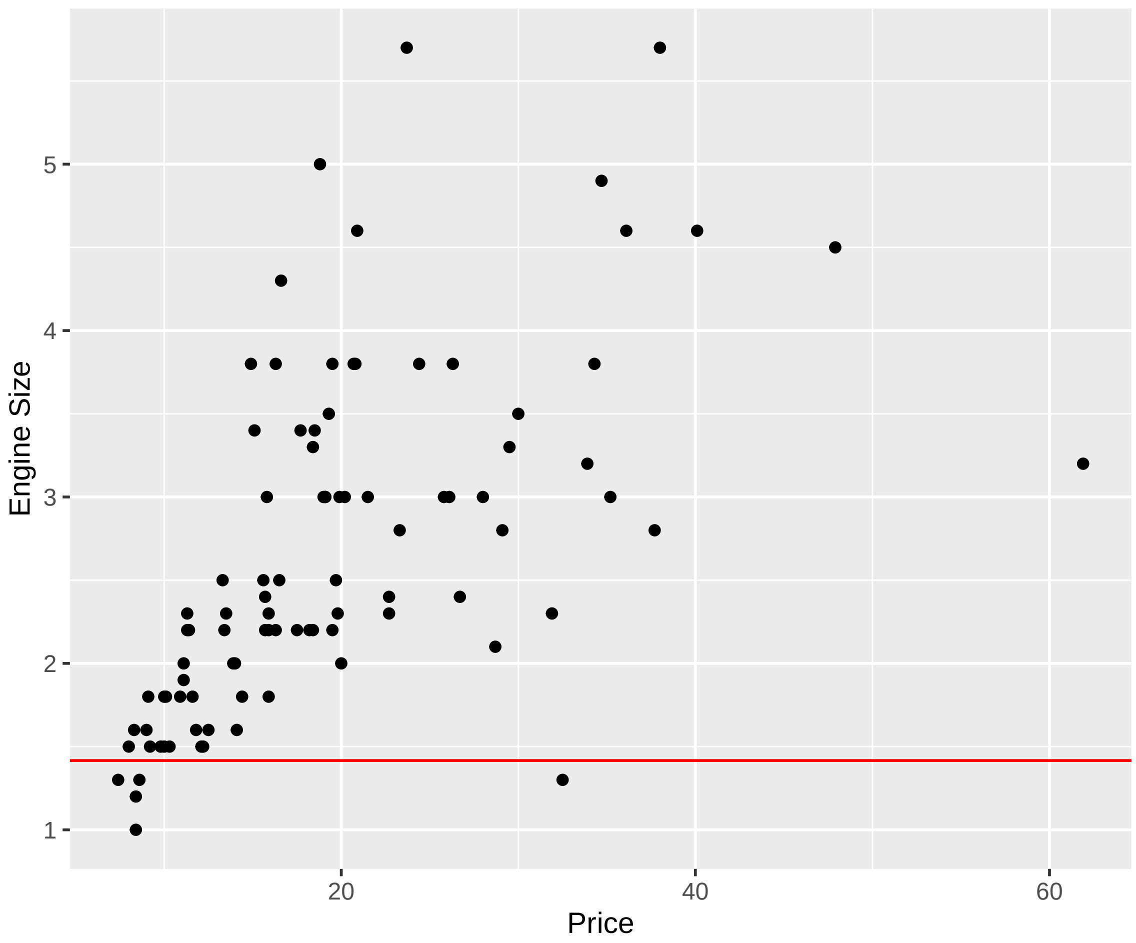
Interpret the slope of the fitted model. Does the intercept have any practical meaning within the scope of our data?

For each additional increase in price we would expect, on average, that the Engine size would increase by 0.0642 units.

If there was no price attached, the engine size would be 1.416 which a realistic situation.

Question 5 Plot the residual plot and comment on the model fit. *Hint*: `augment()`

```
# A tibble: 93 x 8
   EngineSize Price .fitted .resid   .hat .sigma  .cooksd .std.resid
        <dbl> <dbl>   <dbl>  <dbl>  <dbl>  <dbl>    <dbl>      <dbl>
 1        1.8  15.9    2.44 -0.636 0.0123  0.838 0.00364      -0.765
 2        3.2  33.9    3.59 -0.391 0.0349  0.840 0.00409      -0.476
 3        2.8  29.1    3.28 -0.483 0.0215  0.840 0.00374      -0.584
 4        2.8  37.7    3.83 -1.03  0.0493  0.834 0.0417       -1.27
 5        3.5  30      3.34  0.159 0.0236  0.841 0.000448      0.193
 6        2.2  15.7    2.42 -0.223 0.0124  0.841 0.000455     -0.269
 7        3.8  20.8    2.75  1.05  0.0109  0.834 0.00881       1.26
 8        5.7  23.7    2.94  2.76  0.0128  0.788 0.0717        3.33
 9        3.8  26.3    3.10  0.697 0.0161  0.838 0.00578       0.840
10        4.9  34.7    3.64  1.26  0.0376  0.830 0.0459        1.53
# ... with 83 more rows
```



Question 6 How much of the variability in the outcome variable `Price` is explained by the regression equation? Comment on the R^2. *Hint*: `glance()`

```
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
1     0.357         0.350 0.836      50.5 2.59e-10     1  -114.  235.  242.
# ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

About 35.69% of variation of Engine Size is explained by Price. The variability based on Price is modertly low. #### Question 7

Compute the sum of squared residuals. Extract the residuals using `augment()`. Match these residuals to the original data and sort the residuals. Is there manufacturers that on averages produces larger or smaller residuals? What do you notice about the residual patterns with respect to the manufacturers?

```
# A tibble: 1 x 1
  sumResid
     <dbl>
1     63.7
```

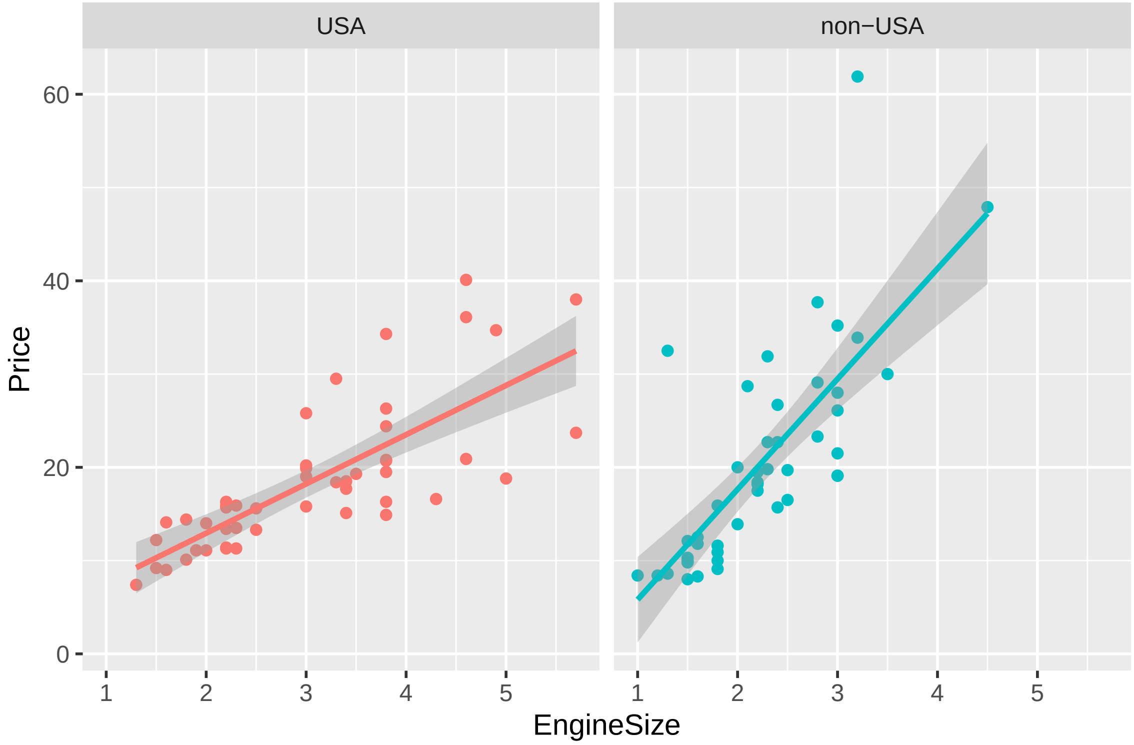
```
# A tibble: 32 x 2
```

```
   Manufacturer meanResid
   <fct>            <dbl>
 1 Acura           -1.03 
 2 Audi            -1.52 
 3 BMW              0.159
 4 Buick            4.29 
 5 Cadillac         1.87 
 6 Chevrolet        8.14 
 7 Chrylser         0.703
 8 Chrysler         0.562
 9 Dodge            0.160
10 Eagle            0.147
# ... with 22 more rows
```

Larger USA car companies have much larger poistive residuals while non-USA companies have negative residuals.Chevrolet, Buik, Ford, etc all have large residuals on average. Toyota, Honda, and Mazda all have negative or very low residuals.

Exploring additional variables

Below is figure showing how Price varies with EngineSize in the Cars93, with accompanying regression lines. There are two plots, one for USA cars, and one for non-USA cars.



Question 8 Use the `lm()` function and fit two models to regress Price on EngineSize with respect to the `Origin` (US and non-US). *Hint*: `filter()`

```
# A tibble: 2 x 5
```

```
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    0.997    0.276       3.61 7.59e- 4
2 Price          0.111    0.0137      8.11 2.01e-10
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)   1.27     0.149        8.48 1.01e-10
2 Price         0.0476   0.00639      7.45 2.92e- 9
```

Question 9 Compare and contrast the fitted models in Question 8 (1-4 lines).

Engine Size of USA cars increases more dramatically as price goes us than foreign cars. If the price was zero, foreign cars would have larger Engine Size than US cars.

Question 10 Multiple regression and categorical predictors were briefly introduced in the class. Do you think analyzing the price based on the size of the engine and the origin will give us more information than the model fitted in Question 3? Justify your answer.

Yes, by adding in more variables and creating multiple regression and categorical predictors, we will be given more information to analyze based on the price. More information and details in regards to the variables will provide a greater understanding and analysis of the data at hand.

References

1. http://www.andrew.cmu.edu/user/achoulde/94842/