Problem set 3

Matt He

2022-09-26

2. Explore the data and determine the number of variables and quantity of any missing values. If values are missing, prescribe a plan to deal with the problem.

```
glimpse(corrola)
```

```
## Rows: 1,436
## Columns: 39
## $ Id                <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1~
## $ Model             <chr> "TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors", "TO~
## $ Price             <dbl> 13500, 13750, 13950, 14950, 13750, 12950, 16900, 186~
## $ Age_08_04         <dbl> 23, 23, 24, 26, 30, 32, 27, 30, 27, 23, 25, 22, 25, ~
## $ Mfg_Month         <dbl> 10, 10, 9, 7, 3, 1, 6, 3, 6, 10, 8, 11, 8, 2, 1, 5, ~
## $ Mfg_Year          <dbl> 2002, 2002, 2002, 2002, 2002, 2002, 2002, 2002, 2002~
## $ KM                <dbl> 46986, 72937, 41711, 48000, 38500, 61000, 94612, 758~
## $ Fuel_Type         <chr> "Diesel", "Diesel", "Diesel", "Diesel", "Diesel", "D~
## $ HP                <dbl> 90, 90, 90, 90, 90, 90, 90, 90, 192, 69, 192, 192, 1~
## $ Met_Color         <dbl> 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1~
## $ Color             <chr> "Blue", "Silver", "Blue", "Black", "Black", "White",~
## $ Automatic         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ CC                <dbl> 2000, 2000, 2000, 2000, 2000, 2000, 2000, 2000, 1800~
## $ Doors             <dbl> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3~
## $ Cylinders         <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4~
## $ Gears             <dbl> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 5, 5~
## $ Quarterly_Tax     <dbl> 210, 210, 210, 210, 210, 210, 210, 210, 100, 185, 10~
## $ Weight            <dbl> 1165, 1165, 1165, 1165, 1170, 1170, 1245, 1245, 1185~
## $ Mfr_Guarantee     <dbl> 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0~
## $ BOVAG_Guarantee   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0~
## $ Guarantee_Period  <dbl> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 12, 3, 3, 3, 3, 3, 3, ~
## $ ABS               <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ Airbag_1          <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ Airbag_2          <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0~
## $ Airco             <dbl> 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ Automatic_airco   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0~
## $ Boardcomputer     <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0~
## $ CD_Player         <dbl> 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 1, 0~
## $ Central_Lock      <dbl> 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1~
## $ Powered_Windows   <dbl> 1, 0, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1~
## $ Power_Steering    <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ Radio             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1~
## $ Mistlamps         <dbl> 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0~
## $ Sport_Model       <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0~
## $ Backseat_Divider  <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0~
```

```
## $ Metallic_Rim      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0~
## $ Radio_cassette    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1~
## $ Parking_Assistant <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ Tow_Bar           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1~
```

```
summary(corrola)
```

```
##        Id              Model                Price          Age_08_04
##  Min.   :   1.0   Length:1436        Min.   : 4350   Min.   : 1.00
##  1st Qu.: 361.8   Class :character   1st Qu.: 8450   1st Qu.:44.00
##  Median : 721.5   Mode  :character   Median : 9900   Median :61.00
##  Mean   : 721.6                      Mean   :10731   Mean   :55.95
##  3rd Qu.:1081.2                      3rd Qu.:11950   3rd Qu.:70.00
##  Max.   :1442.0                      Max.   :32500   Max.   :80.00
##    Mfg_Month         Mfg_Year           KM          Fuel_Type
##  Min.   : 1.000   Min.   :1998   Min.   :     1   Length:1436
##  1st Qu.: 3.000   1st Qu.:1998   1st Qu.: 43000   Class :character
##  Median : 5.000   Median :1999   Median : 63390   Mode  :character
##  Mean   : 5.549   Mean   :2000   Mean   : 68533
##  3rd Qu.: 8.000   3rd Qu.:2001   3rd Qu.: 87021
##  Max.   :12.000   Max.   :2004   Max.   :243000
##        HP          Met_Color         Color              Automatic
##  Min.   : 69.0   Min.   :0.0000   Length:1436        Min.   :0.00000
##  1st Qu.: 90.0   1st Qu.:0.0000   Class :character   1st Qu.:0.00000
##  Median :110.0   Median :1.0000   Mode  :character   Median :0.00000
##  Mean   :101.5   Mean   :0.6748                      Mean   :0.05571
##  3rd Qu.:110.0   3rd Qu.:1.0000                      3rd Qu.:0.00000
##  Max.   :192.0   Max.   :1.0000                      Max.   :1.00000
##        CC            Doors         Cylinders     Gears       Quarterly_Tax
##  Min.   : 1300   Min.   :2.000   Min.   :4   Min.   :3.000   Min.   : 19.00
##  1st Qu.: 1400   1st Qu.:3.000   1st Qu.:4   1st Qu.:5.000   1st Qu.: 69.00
##  Median : 1600   Median :4.000   Median :4   Median :5.000   Median : 85.00
##  Mean   : 1577   Mean   :4.033   Mean   :4   Mean   :5.026   Mean   : 87.12
##  3rd Qu.: 1600   3rd Qu.:5.000   3rd Qu.:4   3rd Qu.:5.000   3rd Qu.: 85.00
##  Max.   :16000   Max.   :5.000   Max.   :4   Max.   :6.000   Max.   :283.00
##      Weight     Mfr_Guarantee    BOVAG_Guarantee  Guarantee_Period
##  Min.   :1000   Min.   :0.0000   Min.   :0.0000   Min.   : 3.000
##  1st Qu.:1040   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.: 3.000
##  Median :1070   Median :0.0000   Median :1.0000   Median : 3.000
##  Mean   :1072   Mean   :0.4095   Mean   :0.8955   Mean   : 3.815
##  3rd Qu.:1085   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: 3.000
##  Max.   :1615   Max.   :1.0000   Max.   :1.0000   Max.   :36.000
##       ABS            Airbag_1         Airbag_2          Airco
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.0000   Median :1.0000   Median :1.0000   Median :1.0000
##  Mean   :0.8134   Mean   :0.9708   Mean   :0.7228   Mean   :0.5084
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  Automatic_airco   Boardcomputer      CD_Player       Central_Lock
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.00000   Median :0.0000   Median :0.0000   Median :1.0000
##  Mean   :0.05641   Mean   :0.2946   Mean   :0.2187   Mean   :0.5801
```

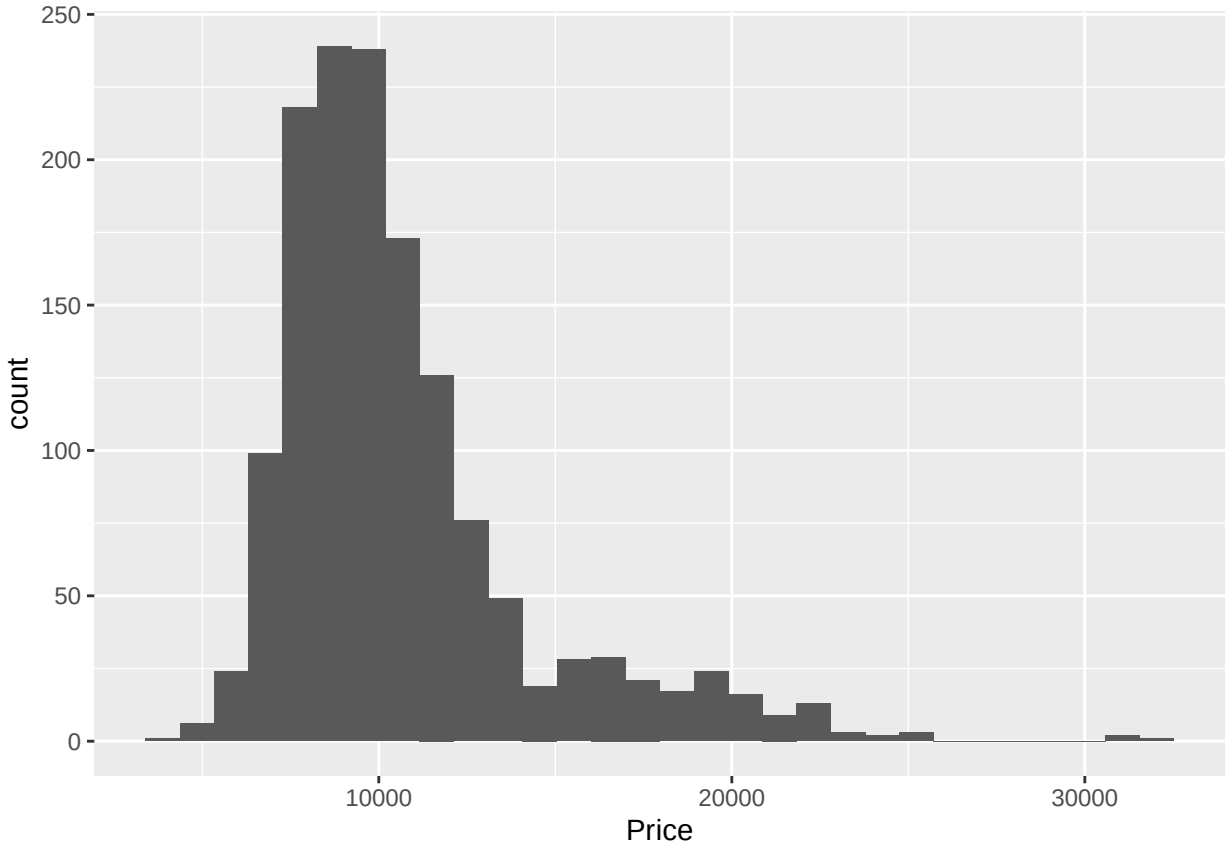
```
##  3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  Powered_Windows Power_Steering       Radio          Mistlamps
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.000   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.000
##  Median :1.000   Median :1.0000   Median :0.0000   Median :0.000
##  Mean   :0.562   Mean   :0.9777   Mean   :0.1462   Mean   :0.257
##  3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.000
##  Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.000
##   Sport_Model     Backseat_Divider  Metallic_Rim    Radio_cassette
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :1.0000   Median :0.0000   Median :0.0000
##  Mean   :0.3001   Mean   :0.7702   Mean   :0.2047   Mean   :0.1455
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  Parking_Assistant     Tow_Bar
##  Min.   :0.000000   Min.   :0.0000
##  1st Qu.:0.000000   1st Qu.:0.0000
##  Median :0.000000   Median :0.0000
##  Mean   :0.002786   Mean   :0.2779
##  3rd Qu.:0.000000   3rd Qu.:1.0000
##  Max.   :1.000000   Max.   :1.0000
```

In the data set, there are 39 variables and 1436 observations. In these variables, there are no missing values.

3. Analyze wheter the Price variable is appropriate for a linear regression model and discuss its distribution. Are there any transformations that we might apply to the price variable?

```
corrola %>%
  ggplot(aes(Price))+
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

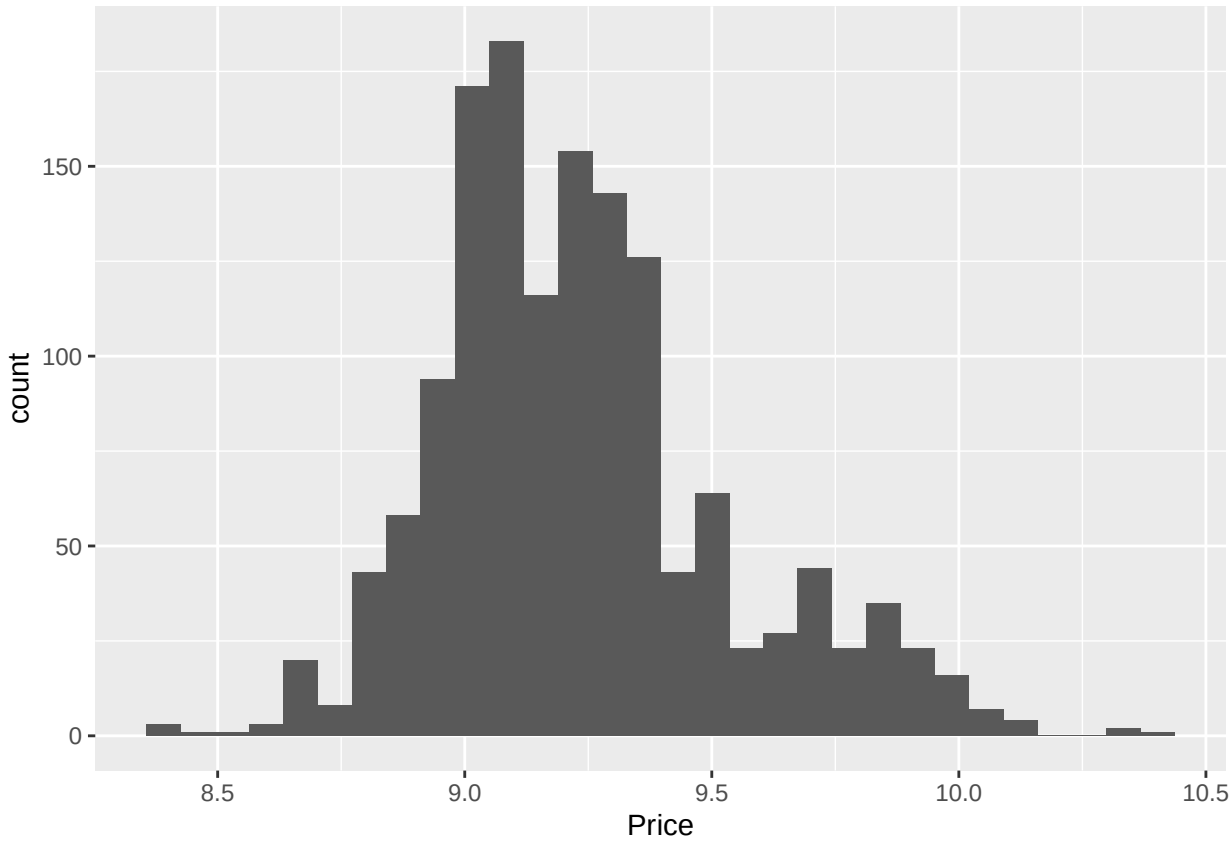


The Price variable is right skewed. And distribution is from 4350 to 32500. The range is large, therefore, I think we need a transformation.

```
corrola <- corrola %>%
  mutate(Price = log(Price))
```

```
corrola %>%
  ggplot(aes(Price))+
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



After transforming the Price variable, we can see the distribution is more bell curved, even thought it is still quite right skewed, and there is a low spot in the middle. I tried other two kind of transformation, but log has the better result comparing the distribution.

4. Is there a relationship between any of the feature in the data and the Price feature? Perform some exploratiory analysis to determine some features that are related using a feature plot.

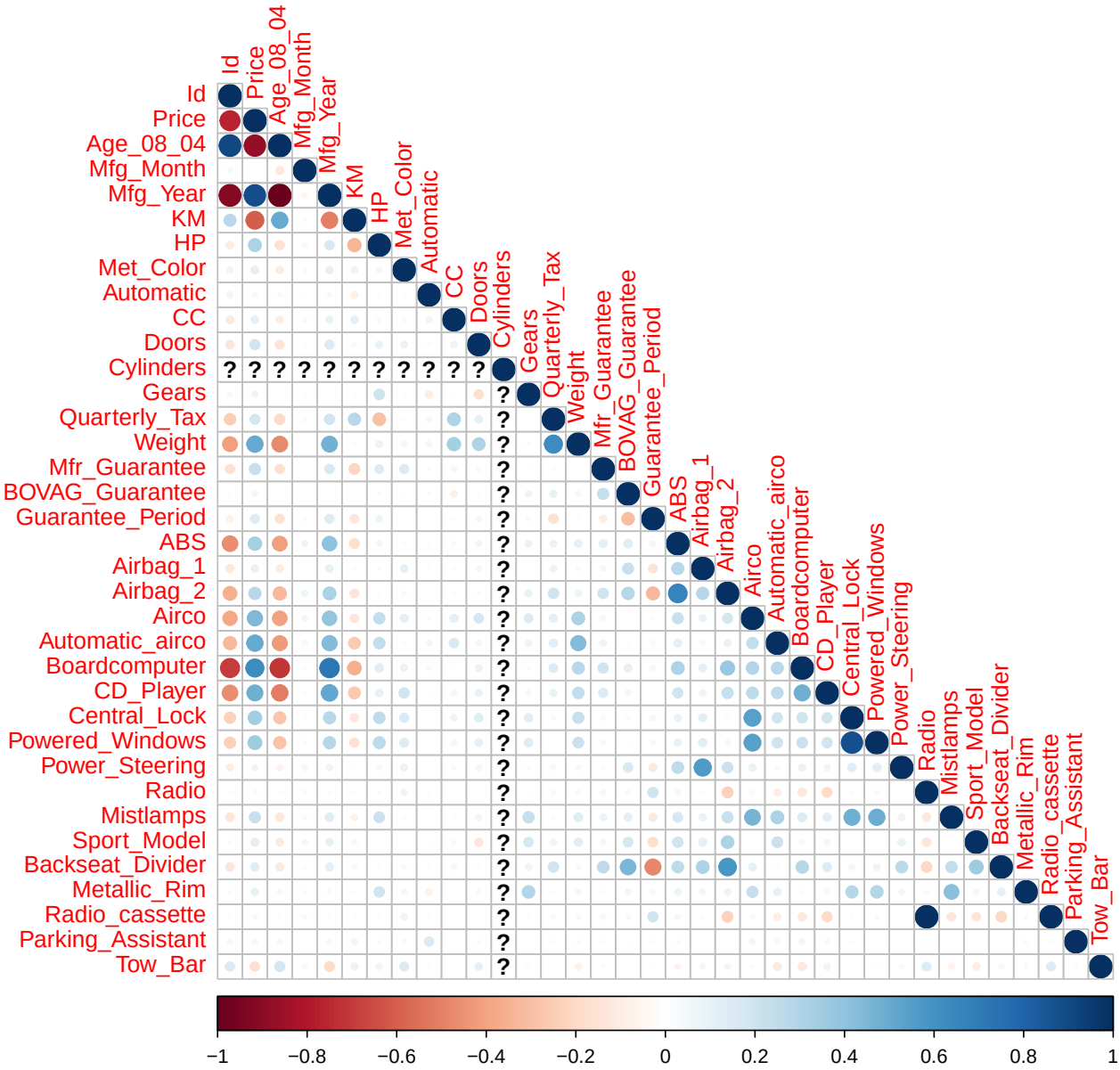
```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
corrMatrix = corrola %>%
  keep(is.numeric) %>%
  cor()
```

```
## Warning in cor(.): the standard deviation is zero
```

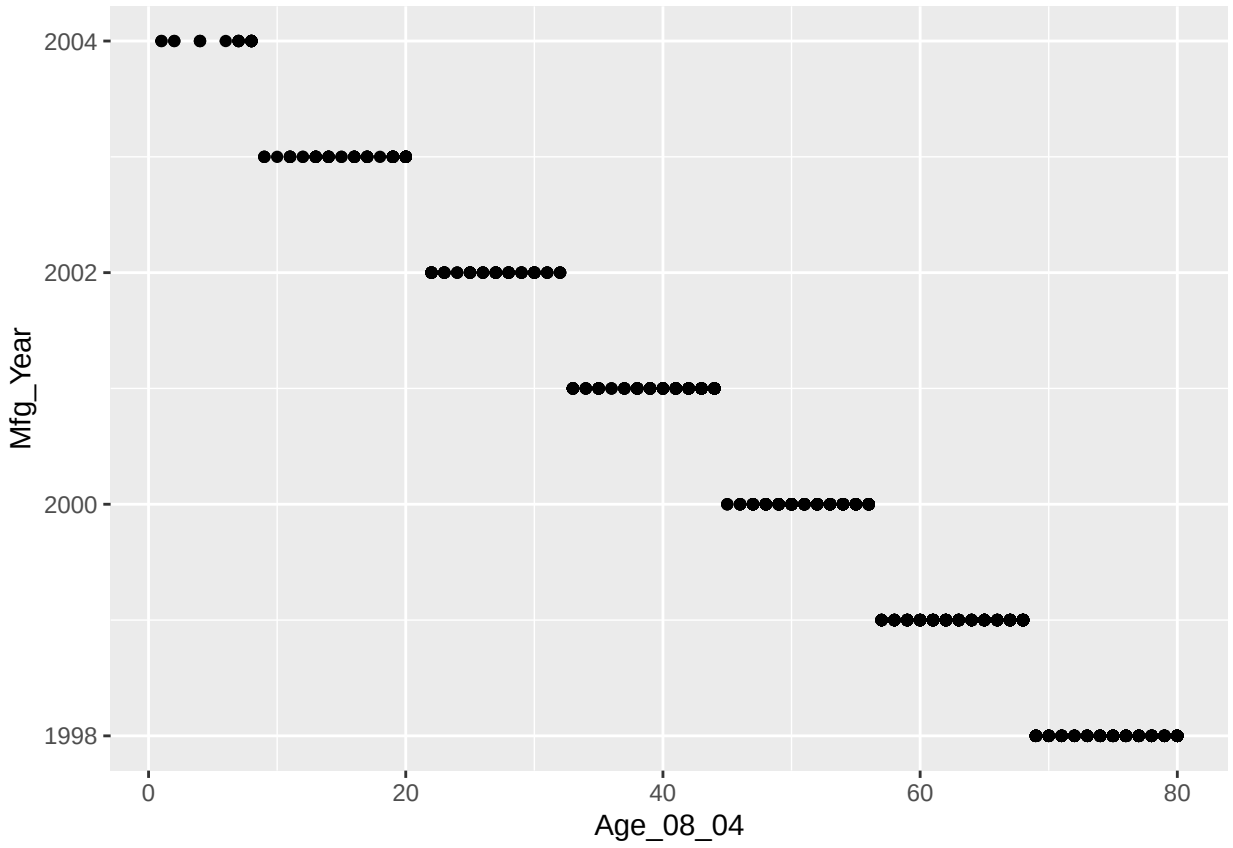
```
corrplot(corrMatrix,type = 'lower')
```



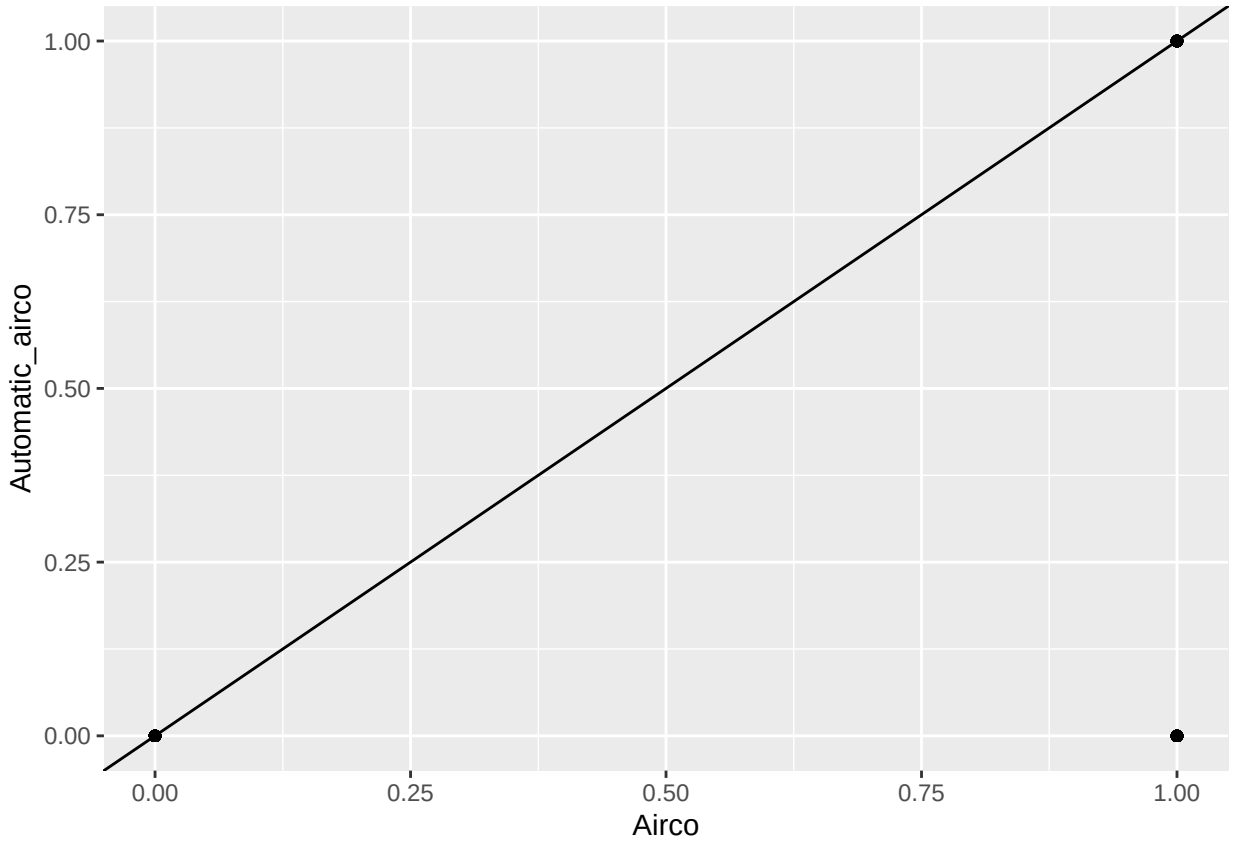
According to the correlation feature plot, we can see the age, weight, and Boardcomputer have strong correlation with Price.

5. Are there any predictore variables in the data that are potentially too strongly related to each other? Make sure to use reference any visualizations, tables, or numbers to show this.

```
corrola %>%
  ggplot(aes(x = Age_08_04, y = Mfg_Year))+
  geom_point()+geom_abline()
```




```
corrola %>%
  ggplot(aes(x = Airco, y = Automatic_airco))+
  geom_point()+geom_abline()
```



According to two graphs, the relationship between Age_08_04 and Mfg_Year is too strong, and Airco and Automatic_Airco. So we shoud not include some of them. We can also ignore some not important variables like Id, Quarterly_Tax and model since they are not likely to affect prediction.

```
corrola <- corrola %>%
  dplyr::select(., -Id,-Mfg_Month, -Mfg_Year, -Quarterly_Tax, -Model, -Airco)%>%
  mutate(Fuel_type = as.factor(Fuel_Type))%>%
  mutate(Color = as.factor(Color))
```

6. Partition your data into a training set with 70% of the observations and a test set with the remaining 30%.

```
set.seed(1234)
samp = caret::createDataPartition(corrola$Price, p = 0.7, list = FALSE)
training = corrola[samp,]
testing = corrola[-samp,]
```

7.

```
ctrl = caret::trainControl(method = 'cv', number = 10)
tree_1 = caret::train(Price~.,
                      data = training,
                      method = "rpart",
                      trControl = ctrl,
                      tuneGrid = expand.grid(cp = seq(0.0, 0.4, 0.005)))
```

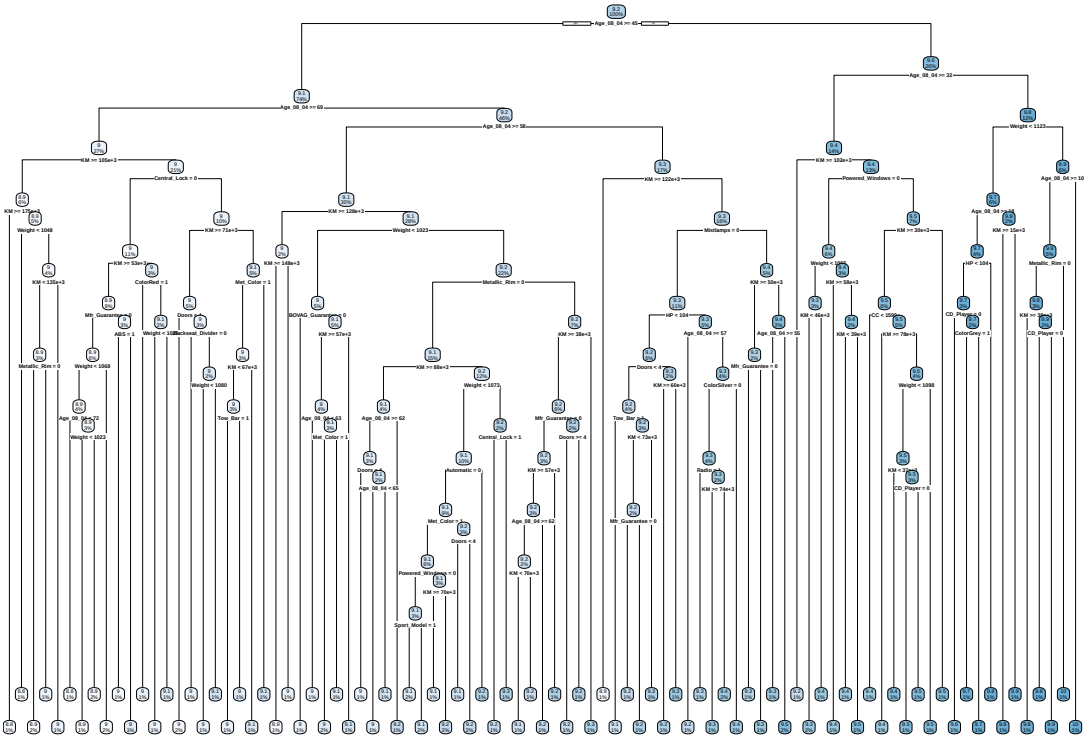
```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(rpart.plot)
rpart.plot(tree_1$finalModel)
```

```
## Warning: labs do not fit even at cex 0.15, there may be some overplotting
```



The plot shows very complex tree. The depth is 10. The role of pre and post pruning is to prevent overfitting, so our model does not go too complex.

8.

```
library(iml)
library(patchwork)
pred = iml::Predictor$new(tree_1, data = training)
imp = iml::FeatureImp$new(predictor = pred, loss = 'rmse', compare = 'ratio', n.repetitions = 10)
imp$results
```

```
##              feature importance.05 importance importance.95 permutation.error
## 1          Age_08_04      3.740183   3.838615      3.905932        0.34889114
## 2                 KM      1.497375   1.548133      1.592204        0.14070956
## 3             Weight      1.250497   1.261261      1.267850        0.11463586
## 4       Metallic_Rim      1.057841   1.071790      1.082012        0.09741481
```

```
## 5       Central_Lock      1.059268   1.069062      1.085312        0.09716692
## 6                 HP      1.050532   1.060388      1.066742        0.09637850
## 7    Powered_Windows      1.039253   1.043990      1.047893        0.09488806
## 8          Mistlamps      1.032495   1.041855      1.060551        0.09469403
## 9      Mfr_Guarantee      1.027839   1.036034      1.047992        0.09416495
## 10             Doors      1.028529   1.034515      1.039965        0.09402690
## 11         Met_Color      1.016963   1.021456      1.029265        0.09283995
## 12         CD_Player      1.011810   1.014661      1.018787        0.09222238
## 13                CC      1.009033   1.013453      1.018769        0.09211258
## 14             Color      1.006735   1.010489      1.015109        0.09184319
## 15           Tow_Bar      1.005040   1.010321      1.014194        0.09182797
## 16         Automatic      1.004653   1.006780      1.010605        0.09150608
## 17  Backseat_Divider      1.005406   1.006074      1.007222        0.09144196
## 18   BOVAG_Guarantee      1.002947   1.004350      1.009262        0.09128526
## 19               ABS      1.002089   1.003554      1.007779        0.09121291
## 20             Radio      1.001651   1.003419      1.005697        0.09120059
## 21       Sport_Model      1.000171   1.002115      1.006127        0.09108208
## 22         Fuel_Type      1.000000   1.000000      1.000000        0.09088985
## 23         Cylinders      1.000000   1.000000      1.000000        0.09088985
## 24             Gears      1.000000   1.000000      1.000000        0.09088985
## 25  Guarantee_Period      1.000000   1.000000      1.000000        0.09088985
## 26          Airbag_1      1.000000   1.000000      1.000000        0.09088985
## 27          Airbag_2      1.000000   1.000000      1.000000        0.09088985
## 28   Automatic_airco      1.000000   1.000000      1.000000        0.09088985
## 29     Boardcomputer      1.000000   1.000000      1.000000        0.09088985
## 30    Power_Steering      1.000000   1.000000      1.000000        0.09088985
## 31    Radio_cassette      1.000000   1.000000      1.000000        0.09088985
## 32 Parking_Assistant      1.000000   1.000000      1.000000        0.09088985
## 33         Fuel_type      1.000000   1.000000      1.000000        0.09088985
```

In this prediction, we probably do not want weight since noone really cares car's weight when buying.

9.

```
training <- training %>%
  dplyr::select(., -Weight)
ctrl_2 = caret::trainControl(method = 'cv', number = 10)
tree_2 = caret::train(Price~.,
                      data = training,
                      method = "rpart",
                      trControl = ctrl_2,
                      tuneGrid = expand.grid(cp = seq(0.0, 0.4, 0.005)))
```

10.

```
price_pred <- predict(tree_2, testing,type = 'raw')
pred_table <- table(testing$Price, price_pred)
pred_table
```

```
sum(diag(pred_table))/nrow(testing)
```

```
## [1] 0.02564103
```

Here I used the method from textbook. However, I got a quite small number and the table looks weird. I am pretty sure the method that textbook uses is for classification trees that predict category of variable. Here, our price is numeric, so this way does not work.