Vehicle Data Analysis project

HOANG SAM BUI

07/12/2020

Abstract:

In this project, we discuss the relationships between the selling price of vehicles with different variables. We use the data in a vehicle data collection file to analyze the relationship between selling price with all variables; selling price with year, seller type, and finally, selling price with year, kilometer drove. We also discuss different perspectives from different types of people in the vehicle industry base on those relationships. During the analysis, we will use different methods such as multiple linear regression models, variable transformation, and cross-validation to help us analyze the dataset. Two out of three models can cover a decent amount of data and show different variables have significant impacts on our models.

Introduction

Vehicles are one of the important components of human life all around the world. With individual buyers and seller, selling price often gets people attention in the first place. Now the majority of people, as both buyers or sellers, often question what makes a significant impact on the selling price of all the vehicles. Does the amount of kilometers driven of a vehicle can affect its selling price. We also wonder if we have all the data that relate to vehicles, how accurate we can predict the selling price for the short run. In this project, we will discuss as well as analysis different methods to approach these huge dataset.

Data Description:

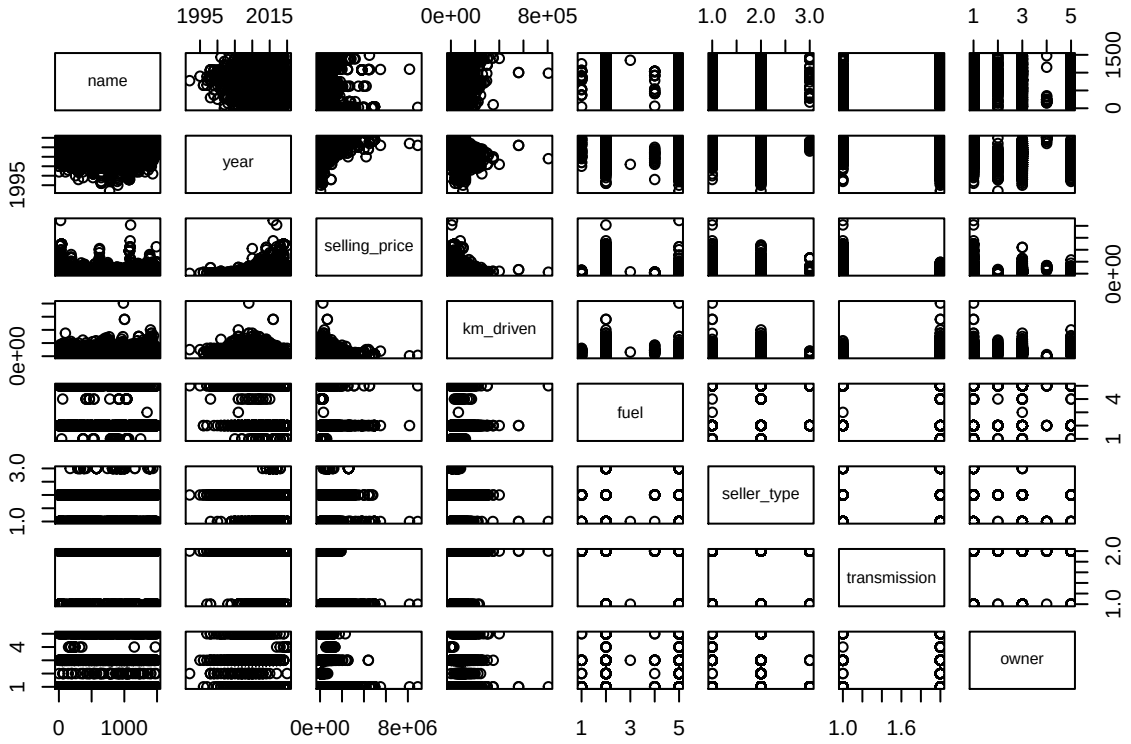
The whole data set will contain one response variable and six predicting variables with a total of 8129 observations. Several observations will have similar names but they are different in some criteria The response variable is the selling price. The six predictors are year, kilometers driven (km_driven), fuel, seller type, transmission, and owner. In the predictors, we have 4 quantitative variables and they are fuel, seller type, transmission, and owner. We will summarize the whole dataset to get an overview of the data.

```
##                       name year selling_price km_driven   fuel seller_type
## 1            Maruti 800 AC 2007         60000     70000 Petrol  Individual
## 2 Maruti Wagon R LXI Minor 2007        135000     50000 Petrol  Individual
## 3     Hyundai Verna 1.6 SX 2012        600000    100000 Diesel  Individual
## 4   Datsun RediGO T Option 2017        250000     46000 Petrol  Individual
## 5    Honda Amaze VX i-DTEC 2014        450000    141000 Diesel  Individual
## 6     Maruti Alto LX BSIII 2007        140000    125000 Petrol  Individual
##   transmission        owner
## 1       Manual  First Owner
## 2       Manual  First Owner
```

```
## 3       Manual  First Owner
## 4       Manual  First Owner
## 5       Manual Second Owner
## 6       Manual  First Owner
```

```
##                      name           year      selling_price
##  Maruti Swift Dzire VDI:  69   Min.   :1992   Min.   :  20000
##  Maruti Alto 800 LXI   :  59   1st Qu.:2011   1st Qu.: 208750
##  Maruti Alto LXi       :  47   Median :2014   Median : 350000
##  Hyundai EON Era Plus  :  35   Mean   :2013   Mean   : 504127
##  Maruti Alto LX        :  35   3rd Qu.:2016   3rd Qu.: 600000
##  Maruti Swift VDI BSIV :  29   Max.   :2020   Max.   :8900000
##  (Other)               :4066
##    km_driven            fuel                seller_type      transmission
##  Min.   :     1   CNG     :  40   Dealer          : 994   Automatic: 448
##  1st Qu.: 35000   Diesel  :2153   Individual      :3244   Manual   :3892
##  Median : 60000   Electric:   1   Trustmark Dealer: 102
##  Mean   : 66216   LPG     :  23
##  3rd Qu.: 90000   Petrol  :2123
##  Max.   :806599
##
##                   owner
##  First Owner         :2832
##  Fourth & Above Owner:  81
##  Second Owner        :1106
##  Test Drive Car      :  17
##  Third Owner         : 304
##
##
```

The majority of buyers prefer Diesel and Petrol fuel while CNG, LPG, and Electric are far less popular. We also notice that most people choose to buy vehicles from individuals instead of Dealer or Trustmark Dealer. As the same time, we can see that most people prefer to be the first owner. The difference in the number of Test Drive Car and Fourth & Above Owner is far less than other types of owners. After observing the dataset,We also want to see if there is an overall relationship between all the variable in the dataset:



By observing the plot above, we can see that selling price has a positive linear relationship with year and negative linaer relationship with km_driven. Others variable are quantitative variable so their plot is likely ambiguous.

Methods:

1.Multiple Linear Regression (MLR):

Linear regression model is a common model that helps analysts determine whether a predictor can have a significant impact on a response variable. Firstly, we express the simple linear regression model as $Y = \beta.X + \epsilon$ where Y, X represents the response variable and the predictor, respectively. The β represents the coefficient of the equation and ϵ represents the error which the equation cannot cover. But since datasets often contain multiple predictors so we have to expand our simple model into a Multiple Linear Regression Model. We express MLR as follow:

$$Y = \beta_0 + \beta_1.X_1 + \beta_2.X_2 + ... + \beta_n.X_n + \epsilon$$

The figure we will look at is the R^2 and $p - value$. The higher in R^2, the more data which the model can cover. We also consider how small the $p-value$ is. If the $p-value$ is less than the significant level ($\alpha = 0.05$), then we can conclude that there is a significant relationship between the predictor and the response variable. The interpretation of R^2 in multiple linear regression is still the same as in SLR. But we will consider the more relevant figure which the adjusted R^2. The adjusted R^2 can be computed as:

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1} \tag{1}$$

2.Variable Tranformation:

During the process of plotting the data, we often observe several relationships between two variables that are not closely linear. By adjusting our model, we expect a new model will be appeared to be more linear than the previous model. Let our first model be $Y = \beta_0 + \beta_1.X_1 + \beta_2.X_2 + \epsilon$. After plotting the model out, we observe that the distribution of the observation is quite linear so we want a new model that will appear to be more linear than the first model. we will create a new variable which is $W = log(Y)$ and $Z = log(X_1)$. Our new model will be as following:

$$W = b_0 + b_1.Z + b_2X_2 + \epsilon$$

3.Cross Validation:

After assigning all predicting variables into our model, we are interested in investigating the prediction performance of the model. Since there are difficulties in finding a new dataset so we decide to use some of the original data to predict the rest of the data. The method we will use in this section is data splitting. The procedure is to split the original data into two separate parts which are called the training sample and test sample. This technique is also called cross-validation. After the prediction process, we will compute the $R^2_{prediction}$. If the $R^2_{prediction}$ is above 0.8, then we know that our model is likely to be a good predictor for new observations.

Results:

1.The relationship between the selling price with all predictors:

Vehicles are already part of human life. For people who had experienced in the vehicles market, then they should aware that different variables can have a huge impact on the selling price of the vehicles. So now, we will discuss when all our variable in the dataset is included in the model, how does our response variable will be affected. We fit all the variables into the model as below:

```
##
## Call:
## lm(formula = selling_price ~ km_driven + year + fuel + seller_type +
##     owner + transmission, data = car.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1185295  -166741   -23884   114047  7547813
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -6.971e+07  3.872e+06 -18.006  < 2e-16 ***
## km_driven                   -9.591e-01  1.683e-01  -5.699 1.28e-08 ***
## year                         3.526e+04  1.918e+03  18.379  < 2e-16 ***
## fuelDiesel                   2.863e+05  6.818e+04   4.200 2.72e-05 ***
## fuelElectric                -6.059e+05  4.324e+05  -1.401 0.161172
## fuelLPG                      4.700e+04  1.117e+05   0.421 0.673889
## fuelPetrol                  -4.245e+03  6.823e+04  -0.062 0.950391
## seller_typeIndividual       -6.638e+04  1.648e+04  -4.029 5.70e-05 ***
## seller_typeTrustmark Dealer  1.675e+05  4.446e+04   3.768 0.000167 ***
## ownerFourth & Above Owner   -1.454e+03  4.986e+04  -0.029 0.976729
## ownerSecond Owner           -4.093e+04  1.668e+04  -2.454 0.014157 *
## ownerTest Drive Car          1.687e+05  1.048e+05   1.609 0.107656
```

```
## ownerThird Owner            -3.993e+04  2.778e+04  -1.437 0.150751
## transmissionManual          -8.703e+05  2.202e+04 -39.533  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 426100 on 4326 degrees of freedom
## Multiple R-squared:  0.4593, Adjusted R-squared:  0.4576
## F-statistic: 282.6 on 13 and 4326 DF,  p-value: < 2.2e-16
```

We observe that base on the $p-value$ in the column $Pr(>|t|)$, the variable year, km_driven, fuel, and seller type are all relatively less than 0.05 which implies that they have a significant impact on our model. The $p-value$ in this column also present how significant a variable has on the fitted model with the condition that all others variable are included in the model. For instance, we could claim that seller_type is significant to our response variable while the rest of the predictors are included in the model. Next, we also observe that the Adjusted R^2 is 0.4576 which illustrates that approximately 46% of the variability in demand is accounted for by the straight_line fit to the selling price.

In practical point of view, we observe that kilomiters driven variable has a lowest negative coefficient. This shows that as more kilomiters a vehicles has driven, the more decrease in the its selling price. Coming before km_driven, we see that transmission_manual also have a negative slopes which are very close to km_driven's coefficient. This implies that as the vehicle has a high selling price, the buyers will prefer transmission automatic instead of manual.We also see the quantitative variable selle_type has a impressived outcome. The higher selling price that a vehicle has, the less likely buyer want to purchase from a individual and mostly likely prefer purchasing from a trustworthy company.

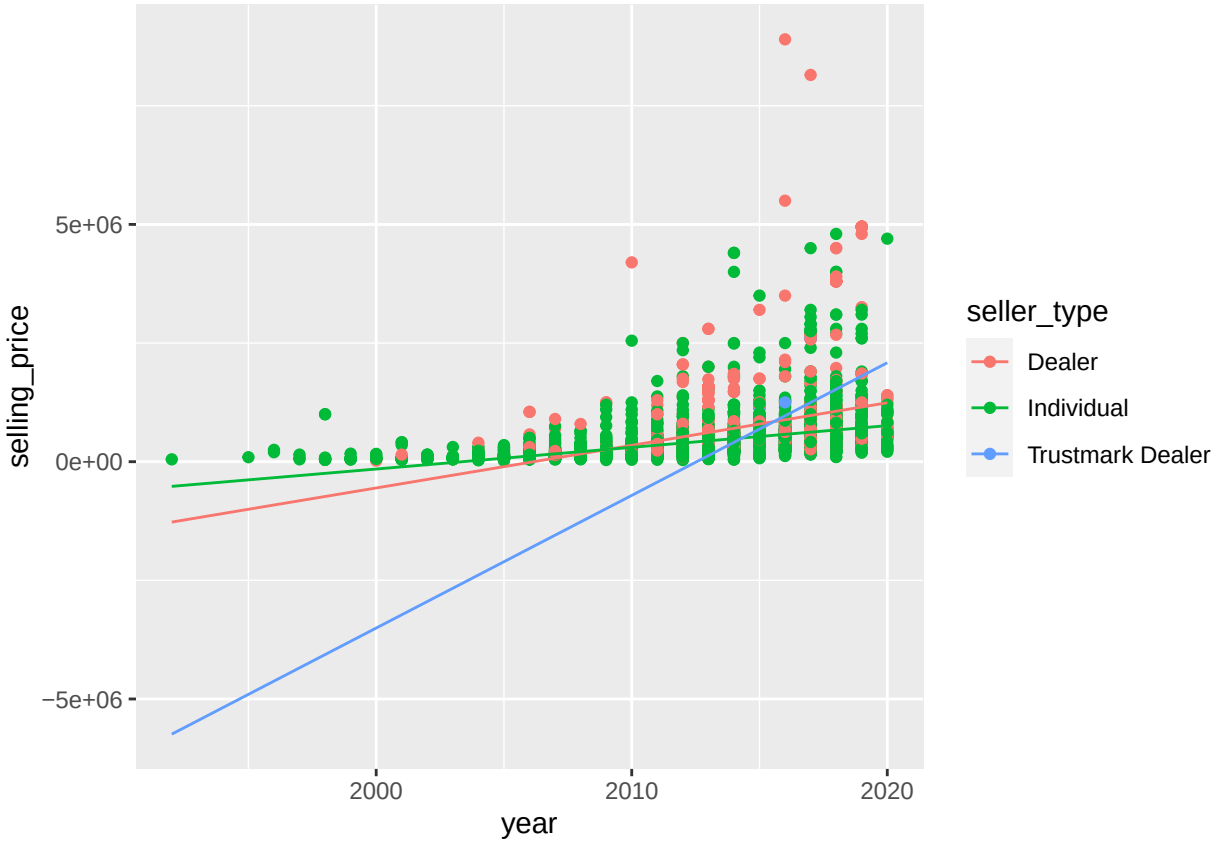
After having a general view of how some variable have significant impact on our model, we also discuss about how our predictive performance by using cross validation.

```
##   intercept     RMSE  Rsquared      MAE   RMSESD RsquaredSD    MAESD
## 1      TRUE 426221.8 0.4565325 230042.1 14366.08 0.05323861 9119.412
```

We observe the R^2 is almost stay the same as the R^2 from our above summary. We also observe large Root Mean Square Error and large Maximun Absolute Error due to the scale of the selling price.

2. The relationship between the selling price with year and seller type predictors:

We use different statistical tools and methods to discuss how would all our predictors will affect the selling price when we fit all variables into the model. We discuss many variables have different impacts on the selling price but that discussion might be suitable for people who are in the vehicles industry for long enough. With a new buyer, they might find quite an amount of difficulties to approach that discussion. So if we look at a perspective of a buyer, especially for people with little experience in the vehicle industry, the most concerns that grab people's attention mostly likely be the age of the car and which seller should they approach. He/she may want a recent year new car or a decent old car with a reasonable price and whether they put quality is the priority. That why we will discuss if two predictors: year and seller_type will have a significant impact on the selling price. Now, we fit the two variables: year and seller_type into the model. Since we have a continuous variable (year) and a quantitative variable (seller_type) in our model, we set up our model equation differently. Our model can be express as $Y = \beta_0 + \beta_1.X_1 + \beta_2.X2 + \beta_3.X_1.X_2 + \epsilon$. we will take a look at the overall diagram for the relationship between year, seller_type, and selling price:



By observing the diagram, the majority of color is green which implies that the majority in seller type likely individual. This is reasonable since when buying from an individual, it's easier to negotiate the selling price. We also see that the slope of the blue line for Trustmarker Dealer is around 30 degrees and it is much higher than the slope of the two others. This tells us that as the age of the vehicles decreases, the selling price in Trustmark Dealer rises more significantly than comparing the selling price of individual and dealer. This is a situation when the buyers want to purchase cars from a company with trustworthy service such as car insurance, car equipment, etc. FOr a more closer look, we can see that when 1 unit of year increase, there will be an increase of 279,507.68 in selling price of Trustmark Dealer. With Dealer and Individual, the increase are only 89503.53 and 45,697.62 in amount of price,respectively. This comparison show that Trustmark worth huge amount of money.Next, We also observe that Dealer has some outliner points represent some exclusive vehicles in the market. But overall, those Outliners does not change our linear line. After discussing the diagram, we will observe how the statistical figure will show us. Below is the summary of the fitted model:

```
##
## Call:
## lm(formula = selling_price ~ year * seller_type, data = car.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -931524 -222594  -86316   64781 8016549
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -180160460    9830523 -18.327  < 2e-16 ***
## year                             89804       4881  18.400  < 2e-16 ***
## seller_typeIndividual         88610950   10649757   8.320  < 2e-16 ***
```

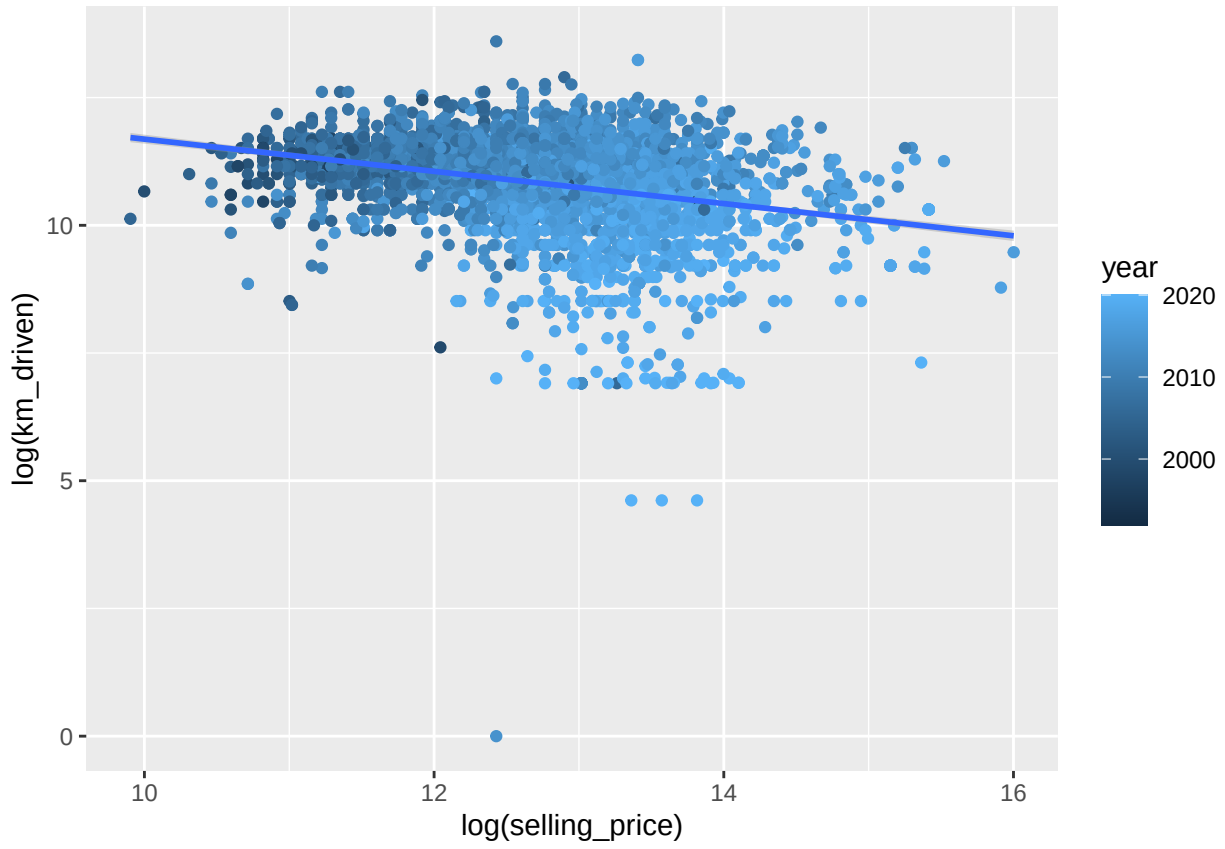
```
## seller_typeTrustmark Dealer      -382360011  103027517  -3.711 0.000209 ***
## year:seller_typeIndividual           -44106       5288  -8.341  < 2e-16 ***
## year:seller_typeTrustmark Dealer     189704      51110   3.712 0.000208 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 512500 on 4334 degrees of freedom
## Multiple R-squared:  0.2163, Adjusted R-squared:  0.2154
## F-statistic: 239.2 on 5 and 4334 DF,  p-value: < 2.2e-16
```

Base on the summary of the model, the $p-value$ of all the predictors: year, seller_type Individual, seller type Dealer are simultaneously very low (far less than 0.05). This show us every individual variable in our model is significant to the response. We also notice the adjusted R^2 is only 21% which tell us that our model can only cover 21% of the dataset.

3. Relationship between Kilometer Driven, year and Selling price:

We discuss how 2 variables: year and seller_type will affect the selling price from the perspective of the buyers. Next, we look at the perspective of a new seller. This is the type of person who has less experience in negotiating and selling their first vehicles. For the new sellers, they often adjust their selling price of the vehicles based on how old is the vehicles and how many kilometers have driven. For the purpose of helping sellers have a basic understanding of how a reasonable selling price should be, we discuss how selling price change when there are only 2 predictors in our model; that is year and kilometers driven. Before we jump into plotting the relationship between these 3 variables, we process a variable transformation. We will fit $log(sellingprice)$ with $log(kmdriven) + year$. Our new diagrams will be below:

```
## `geom_smooth()` using formula 'y ~ x'
```

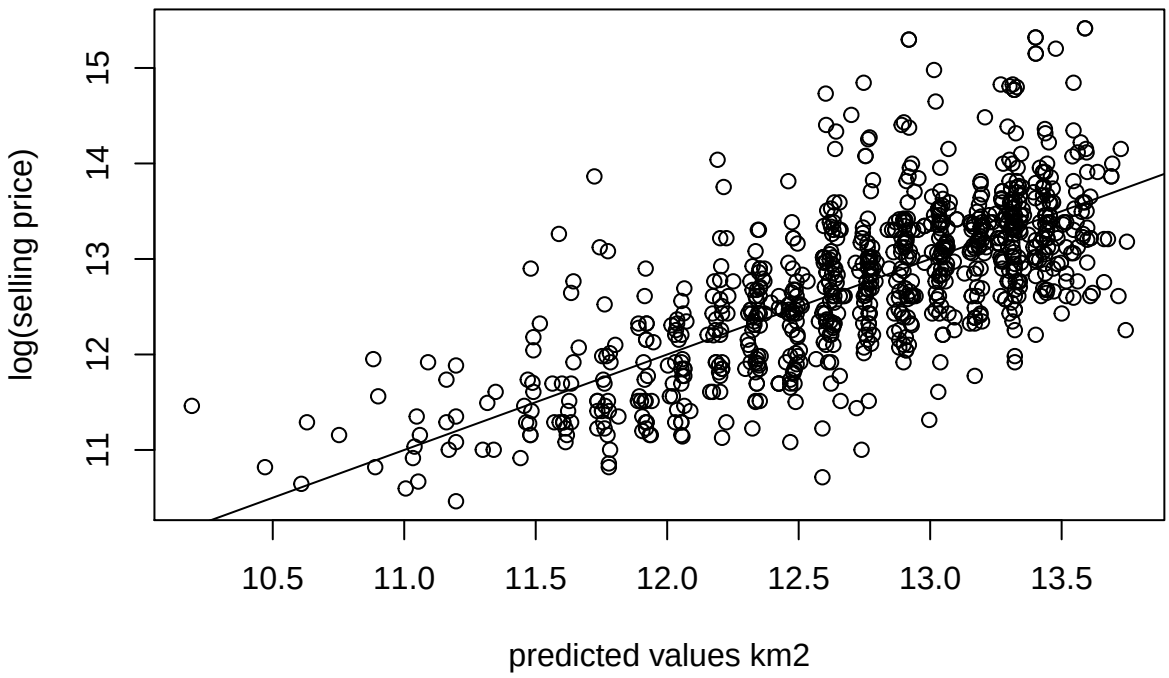


There distribution of observation seems much more linear. We observe that there is a negative slope in a straight line which shows that the fewer kilometers a vehicle has driven, the more expensive it is. It's similar to the year when the diagrams show us that majority of buyers prefer the year after around 2012. Vehicles that have a year value that less than 2010 are offered at a really low price. Now we will use the statistical figure to help us discuss the impact of kilometers driven and year with a selling price:

```
##
## Call:
## lm(formula = log(selling_price) ~ log(km_driven) + year, data = car.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1300 -0.3957 -0.0271  0.3105  3.2231
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -2.754e+02  5.018e+00 -54.884  < 2e-16 ***
## log(km_driven)  4.332e-02  1.153e-02   3.758 0.000174 ***
## year            1.429e-01  2.462e-03  58.032  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.602 on 4337 degrees of freedom
## Multiple R-squared:  0.4857, Adjusted R-squared:  0.4855
## F-statistic:  2048 on 2 and 4337 DF,  p-value: < 2.2e-16
```

We observe that both kilometers driven and year are individually significant with the selling price because their $p - value$ is far lower than signifcant level. The R^2 is approximately 48% which show that our model is covering quite decent amount of data.

Now we want to see our predictive performance:




```
## [1] "|R-square  |Root Mean Square Prediction Error| Maximun Absolute prediction error"
```

```
## [1] 0.4969793 0.6078225 0.4560041
```

```
## [1] " The viability is "
```

```
## [1] 0.7094239
```

We observe our prediction graph and it's clear that many points are a little far from the predicted line. This shows that the 2 variables are not predicting our model really well. Our R^2 is only 42% and the viability is 0.76. This means we are not predicting well. Our maximum absolute prediction error is 0.44 so that means the furthest distance from a point to the line is not really high.

Conclusion:

We discuss three different relationships from three different perspectives. Our discussion moves from the perspective of experienced individuals in the vehicle industry to a new vehicle buyer perspective and then ends up in a new vehicle seller point of view. In general, all the models show us how significant a variable can affect our response variable which is the selling price. The first model and the third model can cover a decent amount of our dataset while the second model does not cover our dataset very well. None of our models have good prediction performance which shows us that we need more data or observations in order to improve our model. These three models also help viewers, buyers, and sellers have a general understanding of what they should expect the selling price when they are involved in the vehicle industry.

Appendix:

```r
#list of library needed:
library(ggiraphExtra)
library(devtools)
library(ggeffects)
library(tidyverse)
library(caret)
library(psych)
library(car)
#Getting the vehicle dataset.
car.data<- read.csv(file = "C:\\Users\\SAM\\OneDrive\\Desktop\\STAT350\\Project\\CAR DETAILS FROM CAR DE
                    ,header = TRUE
                    ,sep = ',',)
head(car.data)
summary(car.data)
pairs(car.data)
#Fit all the predictors to the model:
rgln.allvar <- lm(selling_price ~  km_driven + year + fuel + seller_type + owner + transmission,
                  data = car.data)
summary(rgln.allvar)
#Prediction Performance of Model 1:
data.ctrl <- trainControl(method = "cv", number = 6)
model.caret <- train(selling_price ~ year + km_driven + seller_type + fuel + owner + transmission,
                     data = car.data,
                     trControl = data.ctrl,
                     method = "lm",
                     na.action = na.pass)
model.caret$results
#Fit only year, seller type with selling price in Model 2
fit.ysller <- lm(selling_price ~  year*seller_type, data = car.data)
ggPredict(fit.ysller,interactive=FALSE)
summary(fit.ysller)
#Fit only year, log(km_driven) with selling price in Model 3
sel <- lm(log(selling_price) ~ log(km_driven) + year, data = car.data)
ggplot(car.data,aes(x=log(selling_price),y=log(km_driven),color=year))+
  geom_point()+
  stat_smooth(method="lm",se=TRUE)
summary(sel)
#Process of creating training sample & test sample
nsample = ceiling(0.8*length(car.data$km_driven))
train_sample = sample(c(1:length(car.data$km_driven)),nsample)
train_sample = sort(train_sample)
train_sample.km = car.data[train_sample, ]
test_sample.km = car.data[-train_sample, ]

train.km <- lm(log(selling_price) ~ log(km_driven) + year, data = train_sample.km)
pred.km2 <- predict(train.km, test_sample.km)

#plot out the prediction performance
plot(pred.km2,log(test_sample.km$selling_price),
     xlab = "predicted values km2",
     ylab = "log(selling price)")
```

```
abline(c(0,1))
R.sq2 = R2(pred.km2, log(test_sample.km$selling_price))
RMSPE.k2 = RMSE(pred.km2,log(test_sample.km$selling_price))
MAPE.k2 = MAE(pred.km2, log(test_sample.km$selling_price))
print('|R-square  |Root Mean Square Prediction Error| Maximun Absolute prediction error')
print(c(R.sq2,RMSPE.k2,MAPE.k2))
viability = RMSPE.k2/sd(log(test_sample.km$selling_price))
print(' The viability is ')
viability
# this R markdown chunk generates a code appendix
```

All the data and code are embedded during the analysis and displayed in the appendix.