Project Step 3

Kiran Komati

2021-06-05

Introduction

In this term,for the project, I chose the price prediction of used cars based on the different attributes of a car. I have selected the carDekho's dataset from kaggle as it had a good rating for the data set i was searching for. It had three sets of data with different attributes related to the car such as the brand,year of purchase,selling price,no. of kms driven etc., My research is primarily based on understanding if there is any relationship of any car's attributes for the price of the car and if , what they are and by checking the covariance, correlation, R2 and others and by checking confidence intervals and performing various tests such as anova and Durbin Watson tests. Regresison can help us in estimating the value of the car there by helping the sellers to pay a fair price and also to estimate if it is a good buy or not.

The problem statement you addressed.

Through this project, I'm planning to address the general understanding that as the car becomes old, it's value is reduced and using linear regression if we can predict the price of the car based on the attributes of a car.

How you addressed this problem statement

I searched for different data sets that we can use to move further with our Analysis and found out the carDekho data set in Kaggle. This dataset is recommended by many people in that website and had a good rating. There are 3 csv files in the data set. Even though the data set do not have any missing values, I had to perform some cleaning to proceed and below are the steps that i have performed. We have learned a very important concept of linear regression in this term.

1. It is from 3 files and in csv format.We can import the csv data set into R easily by using read.csv function.
2. Data set 3 has additional columns. I need to create a new data set from this without these additional columns in order to be able to merge the 3 data sets without any issues.
3. Dataset 2 has one additional column,Current_price , I need to remove it before merging.
4. Dataset 1 has the price in multiples of 100k rupees. I need to convert to a price by multiplying it with 100000.
5. Rename data sets 1 and Data set 3 to match the names of Data set2.
6. I need to merge the 3 data sets into a single data set.
7. Perform Exploratory Data Analysis to understand more about the columns and if its normally distributed.

Once the above cleaning is complete, i had to do a covariance check to see if any variables are related and found that the variables selling price and the year are positively correlated and selling price and kms driven are negatively correlated. I used linear regression model to understand how much of the relationship our model can explain.

Analysis

```
library("tidyverse")
library('ggplot2')
library('dplyr')
CD.df1Raw <- read.csv("CarDekho1.csv")
CD.df2Raw <- read.csv('CarDekho2.csv')
CD.df3Raw <- read.csv('CarDekho3.csv')
CD.df1<-select(CD.df1Raw, -Present_Price) %>%
        rename(name=Car_Name,
                      year=Year,
                      selling_price=Selling_Price,
                      km_driven=Kms_Driven,
                      fuel=Fuel_Type,
                      seller_type=Seller_Type,
                      transmission=Transmission,
                      owner=Owner) %>%
        mutate(owner=recode(owner,`0`="First Owner",
                            `1`="Second Owner",
                            `3`="Fourth Owner"))

CD.df1$selling_price<- CD.df1$selling_price*100000

CD.df2 <- CD.df2Raw
CD.df3 <- select(CD.df3Raw,-engine,-max_power,-torque,-seats,-mileage)
CarData <- rbind(CD.df1,CD.df2,CD.df3)
```

Above steps are performed as part of the cleaning. Once the cleaning is done. I have selected few variables and performed covariance, correlation and the regression model.

```
CarData_SS <- select(CarData,selling_price,km_driven,year)
cov(CarData_SS)
```

```
##                selling_price     km_driven          year
## selling_price  537930657529 -8096713558.6 1231633.48424
## km_driven       -8096713559  2836334569.7  -89761.69959
## year                1231633      -89761.7      16.75862
```

```
cor(CarData_SS)
```

```
##                selling_price  km_driven       year
## selling_price     1.0000000 -0.2072845  0.4102033
## km_driven        -0.2072845  1.0000000 -0.4117116
## year              0.4102033 -0.4117116  1.0000000
```

```r
CarData_SS_LM1 <- lm(selling_price ~  year ,data = CarData_SS)
summary(CarData_SS_LM1)
```

```
##
## Call:
## lm(formula = selling_price ~ year, data = CarData_SS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -852100 -283131 -134145   39333 9158378
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -147392785    2911766  -50.62   <2e-16 ***
## year             73493       1446   50.82   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 668900 on 12767 degrees of freedom
## Multiple R-squared:  0.1683, Adjusted R-squared:  0.1682
## F-statistic:  2583 on 1 and 12767 DF,  p-value: < 2.2e-16
```

```r
CarData_SS_LM2 <- lm(selling_price ~  km_driven ,data = CarData_SS)
summary(CarData_SS_LM2)
```

```
##
## Call:
## lm(formula = selling_price ~ km_driven, data = CarData_SS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -751797 -342396 -162932   62052 9303418
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.822e+05  1.028e+04   76.08   <2e-16 ***
## km_driven   -2.855e+00  1.192e-01  -23.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 717500 on 12767 degrees of freedom
## Multiple R-squared:  0.04297,    Adjusted R-squared:  0.04289
## F-statistic: 573.2 on 1 and 12767 DF,  p-value: < 2.2e-16
```

```r
CarData_SS_LM <- lm(selling_price ~  year + km_driven,data = CarData_SS)
summary(CarData_SS_LM)
```

```
##
## Call:
## lm(formula = selling_price ~ year + km_driven, data = CarData_SS)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -847842 -284533 -133762   41589 9146037
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.405e+08  3.195e+06 -43.966  < 2e-16 ***
## year         7.008e+04  1.585e+03  44.211  < 2e-16 ***
## km_driven   -6.367e-01  1.218e-01  -5.226 1.76e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 668200 on 12766 degrees of freedom
## Multiple R-squared:   0.17,  Adjusted R-squared:  0.1699
## F-statistic:  1308 on 2 and 12766 DF,  p-value: < 2.2e-16
```

The above steps performed shows that the general myth of having less mileage will fetch a good price is accounted for less than 5% in the data set and that year is strongly positively correlated to the price of the car. I created a multiple linear regression model that has both the 'year' and 'km_driven' to predict 'selling_price' and the accuracy increased a little bit but not much.

confint(CarData_SS_LM)

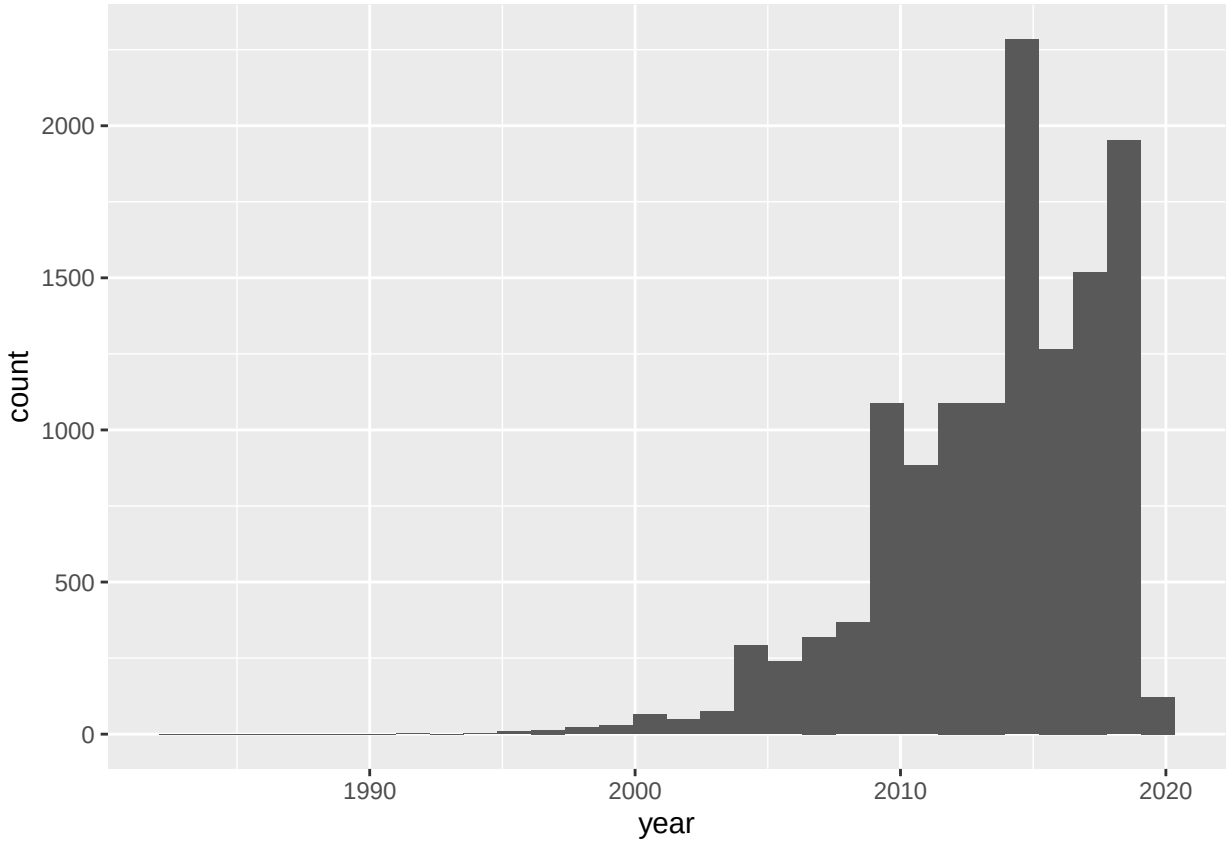
This confidence interval tells us that the predictor 'year' have very tight confidence intervals, indicating that the estimates for the current model are likely to be representative of the true population values. The interval for 'km driven' is wider (but still does not cross zero), indicating that the parameter for this variable is less representative, but nevertheless significant.

Summarize the implications to the consumer (target audience) of your analysis.

From the above Analysis we can safely say that the car price is positively correlated the year of the purchase for the used cars. THe confidence interval also confirms the same.

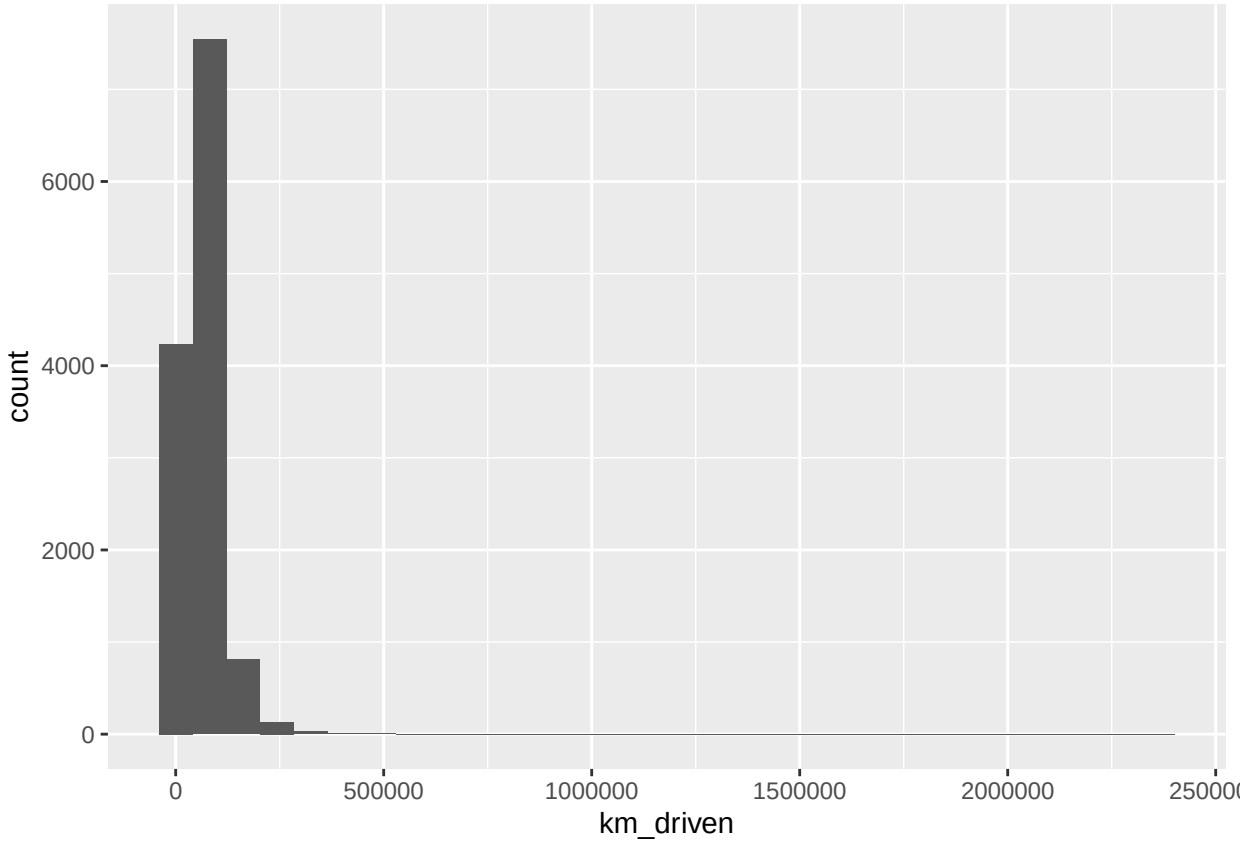
```
ggplot(CarData, aes(x=year)) + geom_histogram()
```

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

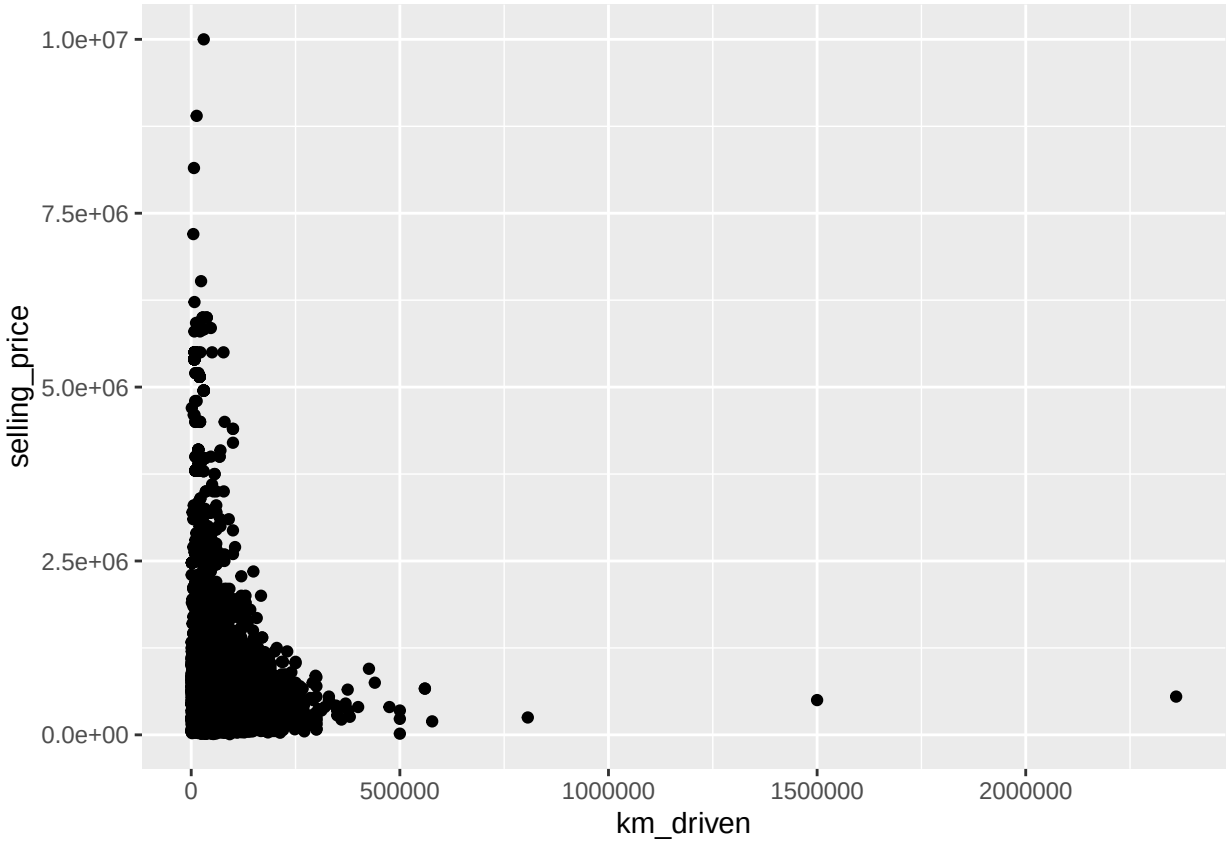



```
ggplot(CarData, aes(x=km_driven)) + geom_histogram()
```

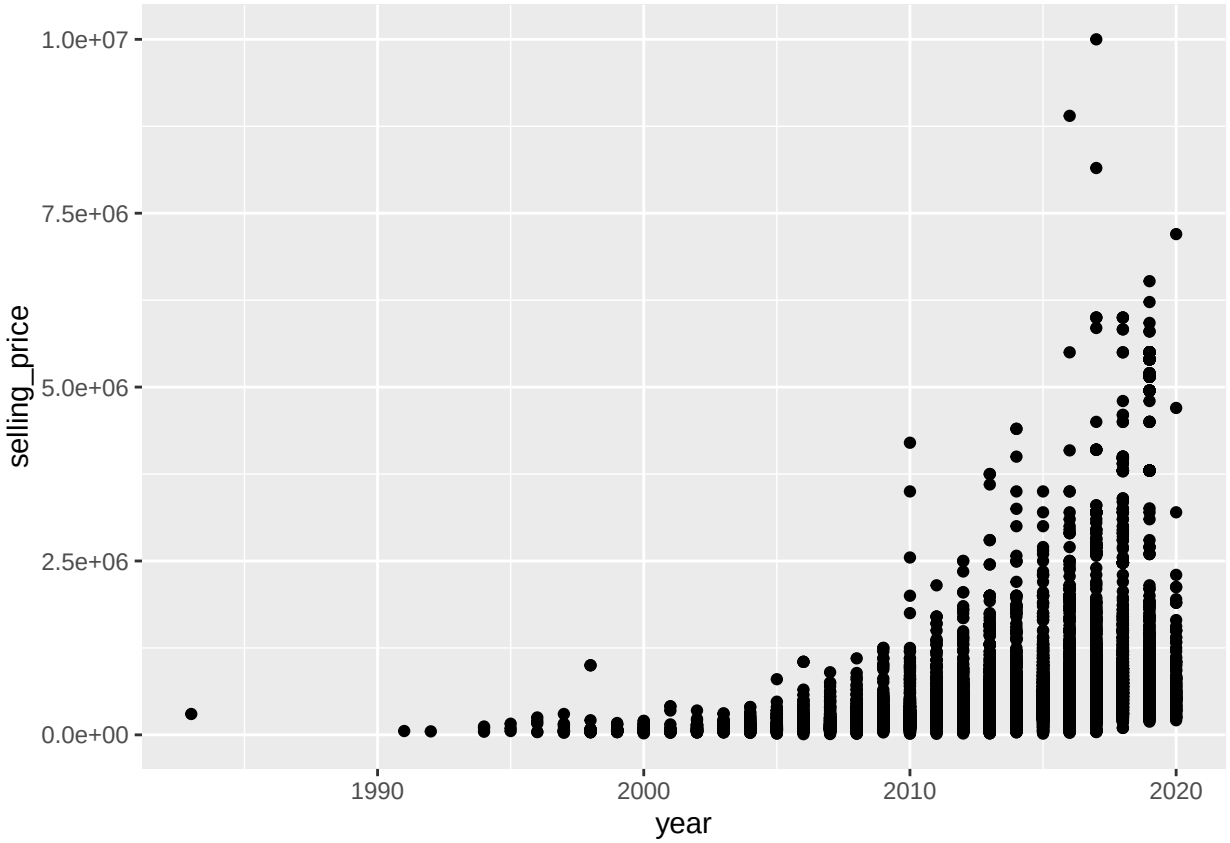
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



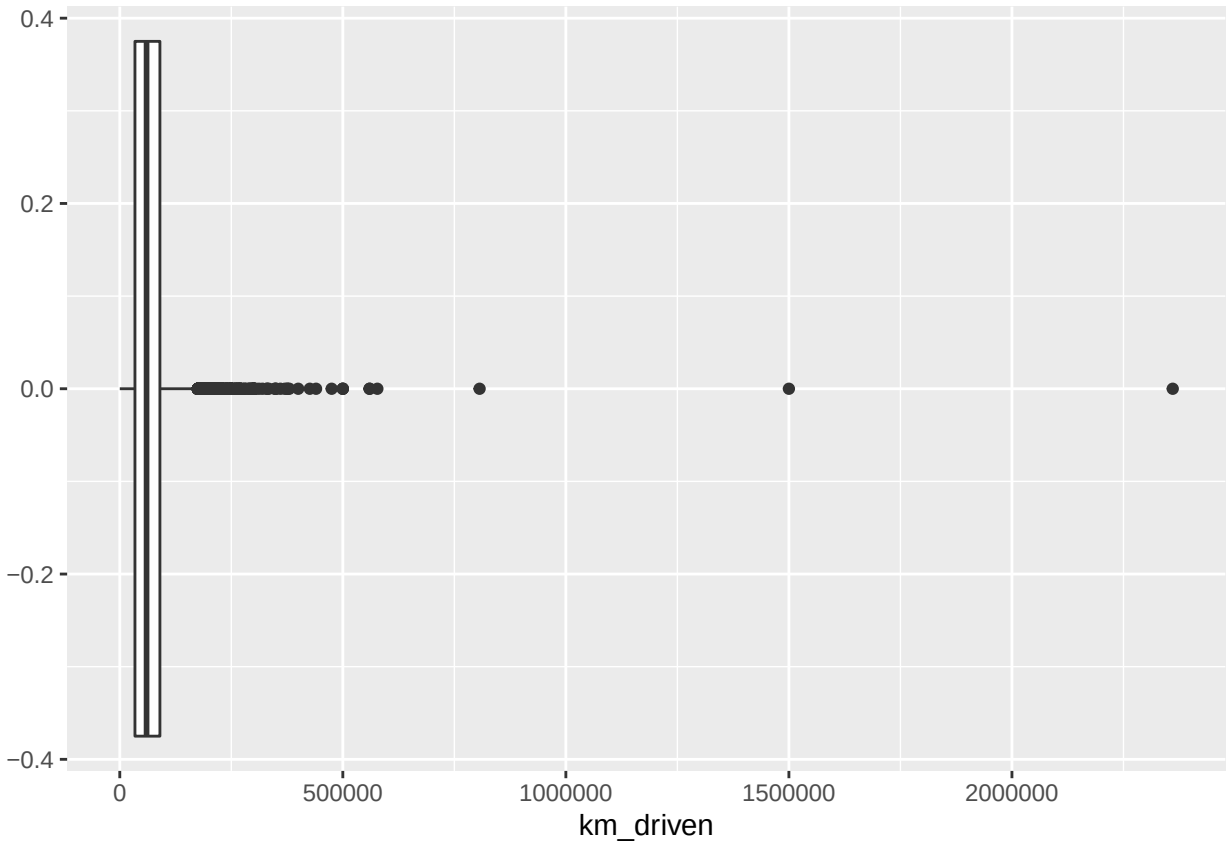

```
ggplot(CarData, aes(x=km_driven, y=selling_price)) + geom_point()
```



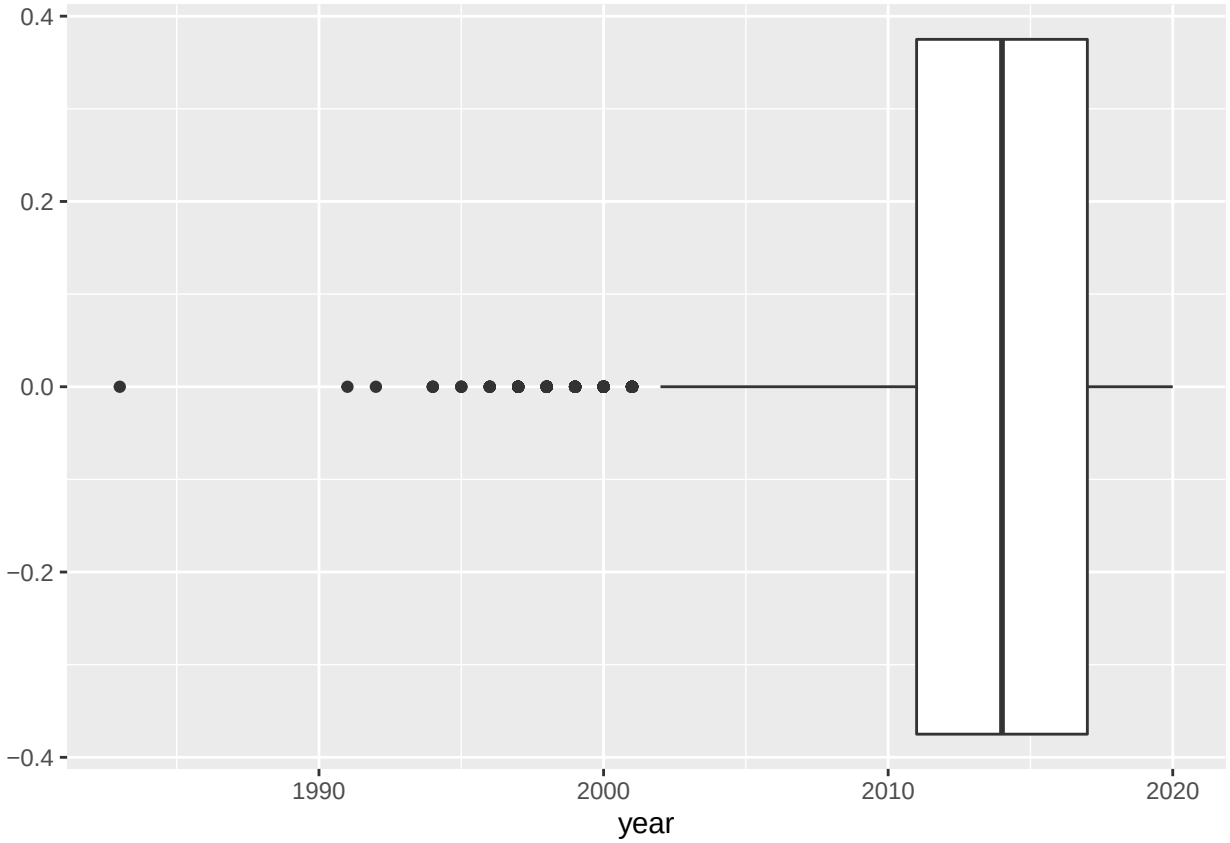

```
ggplot(CarData, aes(x=year, y=selling_price)) + geom_point()
```




```
ggplot(CarData, aes(x=km_driven)) + geom_boxplot()
```




```
ggplot(CarData, aes(x=year)) + geom_boxplot()
```



THe histograms depict that the year histogram is left skewed where as the km_driven histogram is right skewed with values concentrated heavily on the right and left respectively and that the data is not noramlly distributed. With the box plots showing the anomalies.

Limitations.

The results we got account for around 17% of the data. There's a fair percentage of amount that is unanswered. Rigorous research needs to be done if there are other attributes that are not present in our dataset that may affect our results. We have eliminated few fields because they are missing from the other 2 data sets. We can try to include those fields as well and perform a research to figure out the relationship of selling price with other variables.

Concluding Remarks

based on the research, we can conclude that multiple regression can be used to understand if there are any attributes that affect the selling price of used cars. But there are other factors that needs to be considered and more extensive research needs to be performed to overcome the limitations mentioned above.