Term_Project (Laurence Cojocaru)

```
library(ggplot2)
library(car)
library(MASS)
```

Abstract:

Buying a vehicle can be a daunting task for an individual; especially if they do not possess an in depth understanding about all the variables that could affect the selling price. In this paper we look at data from CarDekho.com to ask the question how does: age of the vehicle, km driven, type of fuel, the type of seller the vehicle is being purchased from, the type of transmission, and the number of owners affects the selling price? What we found was: the age, km driven, and number of owners up until two has a negative correlation with selling price. We also found that dealers and trustmark dealers often sold cars at higher rates than individuals. Finally we found that vehicles that use manual transmission sold for a lower price than automatic vehicles, and that vehicle's that used diesel tended to be more expensive than vehicles that used other fuel types.

Introduction:

Buying a car or motorcycle can be a big deal for an individual, especially if the individual does not know a lot about cars. There are a lot of variables to consider when looking to purchase a car or motorcycle, and each of these variables affects the price of the vehicle differently, so for a person who's knowledge on cars or motorcycles may be limited it is hard for them to deduce if the price they are getting charged for that vehicle is ok or overpriced. In this paper we will perform a regression analysis of a dataset from CarDekho.com to ask the question how does: age of the vehicle, km driven, type of fuel, seller type, type of transmission, and number of owners affects the selling price of a vehicle?

Data Description:

Initially the Dataset contained the following variables: name, year, "selling_price", "km_driven", fuel, "seller_type", transmission, and owner. The variable name contained the name of the vehicle, and it was removed from our dataset as it was not needed for our current investigation. The variable year which contained the year the car was made; was also removed from our dataset, but it was replaced with a new unique variable called "age_of_veh". This variable was calculated by taking the current year (2020) and subtracting the year the car was made from it. This variable was added because we felt like knowing how the age of a vehicle affected its selling price was more important than knowing how the year it was made would affect its price. We also felt like this new variable would be easier for individuals who might not know a lot about cars and motorcycles to understand. The variable "selling_price" contains the price the vehicle is being sold for. The variable "km_driven" contains the amount of kilometric the car has been driven. We slightly modified this variable by making every 1km = 1000km we did this as we felt like it was

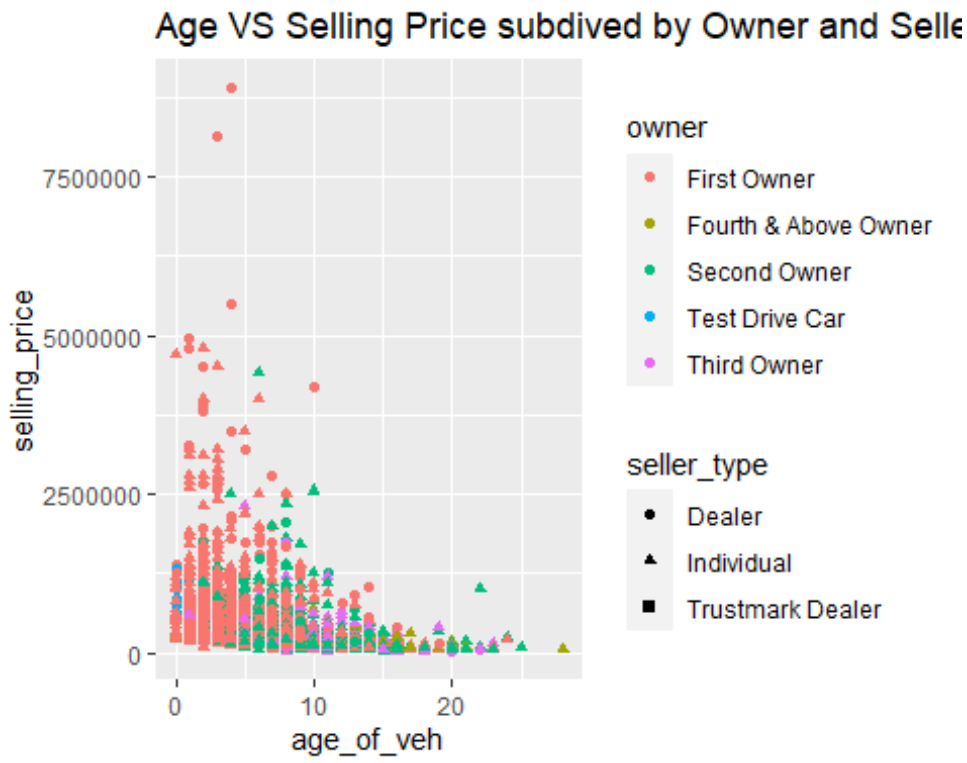
easier to read. The variable fuel informs us what type of fuel the vehicle uses; it is broken down into 5 different types of fuels: CNG, Diesel, electric, LPG and petrol. The variable "seller_type" informs us how the vehicle is being sold; it is divided into: dealer, individual and trustmark dealer. The variable transmission tells us what transmission the vehicle has, and it is subdivided into: automatic and manual. Finally the variable owner shows us how many people have driven the vehicle: 1, 2, 3, 4+ or whether the vehicle was used for test drives.

```
CarData <- read.csv("CAR DETAILS FROM CAR DEKHO.csv")



CarData <- subset(CarData, select = -c(name))
CarData$age_of_veh <- 2020 - CarData$year
CarData <- subset(CarData, select = -c(year))
CarData$km_driven <- CarData$km_driven/1000

summary(CarData)
```

```
##  selling_price       km_driven             fuel                seller_type
##  Min.   :  20000   Min.   :  0.001   CNG     :  40   Dealer          : 994
##  1st Qu.: 208750   1st Qu.: 35.000   Diesel  :2153   Individual      :3244
##  Median : 350000   Median : 60.000   Electric:   1   Trustmark Dealer: 102
##  Mean   : 504127   Mean   : 66.216   LPG     :  23
##  3rd Qu.: 600000   3rd Qu.: 90.000   Petrol  :2123
##  Max.   :8900000   Max.   :806.599
##     transmission                    owner        age_of_veh
##  Automatic: 448   First Owner         :2832   Min.   : 0.000
##  Manual   :3892   Fourth & Above Owner:  81   1st Qu.: 4.000
##                   Second Owner        :1106   Median : 6.000
##                   Test Drive Car      :  17   Mean   : 6.909
##                   Third Owner         : 304   3rd Qu.: 9.000
##                                               Max.   :28.000
```

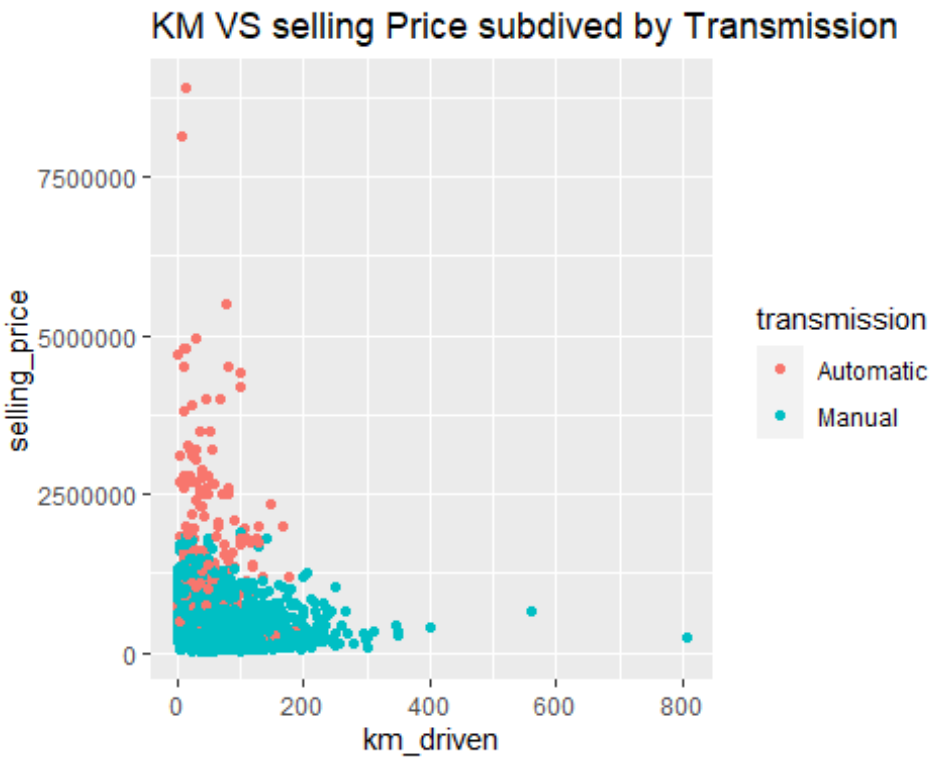
From the figure above we can see that the average selling price is roughly $500,000, and the average distances the vehicles have been driven is about 66,000km. We can also see that the most common type of fuel is diesel, but petrol is not far away. This figure also shows us the individuals sell the majority of vehicles, and that most vehicles are: manual, and only owned by 1 person, with an average age of close to 7 years.

```
ggplot(CarData, aes(age_of_veh, selling_price, color = owner, shape = seller_
type,)) + geom_point() + ggtitle("Age VS Selling Price subdived by Owner and
Seller Type")
```



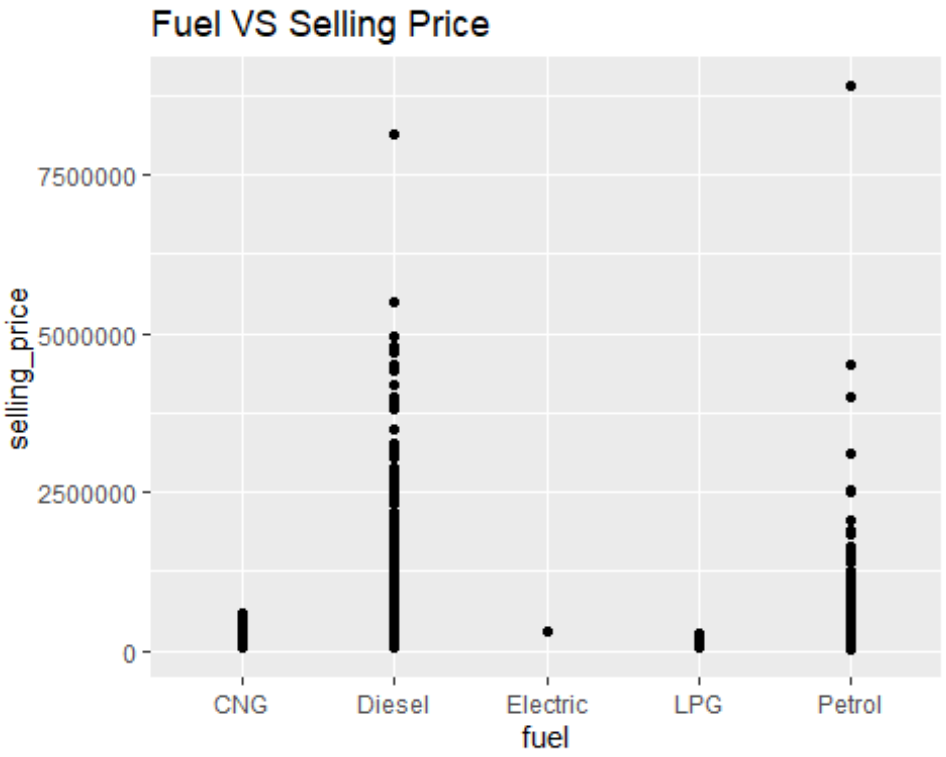
From this graph we can easily make out two patterns. The first being that age has a negative correlation with selling price, and second that as the number of owners increases the price of the vehicle tends to decrease. This graph also lets us see that the 3 most expensive vehicles are dealer sold and have only had one owner

```
ggplot(CarData, aes(km_driven, selling_price, color = transmission)) + geom_p
oint() + ggtitle("KM VS selling Price subdived by Transmission")
```



From this graph we can also make out two patterns. The first is the negative correlation between selling price and km driven, and the second is that automatic vehicles have a much higher selling price than manual vehicles do.

```
ggplot(CarData, aes(fuel, selling_price)) + geom_point() + ggtitle("Fuel VS S
elling Price")
```



Our last graph shows us that vehicles fuelled with CGN, Electric, and LGP don't have high selling prices; however vehicles fuelled with petrol and diesel tend to have wide ranging selling prices, with diesel being a little more expensive on average than petrol.

Methods:

The first step we took was running a simple regression model on all our variables of interest. In order to construct this regression model we constrained R to use the levels with the highest number of observations as its reference instead of letting R just select the first level by default. We did this primarily because of the fuel variable. We did not feel that selecting the first level was the best idea as there are not many observations in that level so the model might not do a good job comparing between the different levels. Since we did this for fuel we felt that it was important to be consistent and do it for the other categorical variables as well. Once the model was constructed we took a look at its coefficients and its ANOVA table. Following this step we run a plot function on the regression model to gain access to the residuals vs fit line and the normal Q-Q plot. We then started to look for problematic data points and possible outliers. We did this by plotted a leverage plot and an influence plot, and running an outlier test; once all the problematic points were found we created a copy of the main dataset without these points. We then ran another simple regression model on our new dataset to see what would happen once the outliers and problematic points were removed. We again ran the summery of this new model, as well as plotted its residuals vs fit line and normal Q-Q plot to see what had changed. After that we once more searched for outliers and problematic points the same way as we did before, and again created a new data set with them removed. We ran one last simple regression model on this new data set, and plotted the residuals vs fit line and normal Q-Q plot, however we

saw no real change here so we did not go any further after this. After we finished with the simple regression models we decided to run a robust regression model with Huber's psi function on the original dataset, just to see if this model would be any better. We concluded by calculating the MSPE for all four of our constructed models, and comparing them to see what model did the best job in predicting.

Results:

```
contrasts(CarData$fuel) <- contr.treatment(5, base = 2)
contrasts(CarData$transmission) <- contr.treatment(2, base = 2)
contrasts(CarData$seller_type) <- contr.treatment(3, base = 2)
LM <- lm(selling_price~., CarData)
summary(LM)

##
## Call:
## lm(formula = selling_price ~ ., data = CarData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1185295  -166741   -23884   114047  7547813
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  859540.3    17508.2  49.094  < 2e-16 ***
## km_driven                      -959.1      168.3  -5.699 1.28e-08 ***
## fuel1                        -286332.0    68177.0  -4.200 2.72e-05 ***
## fuel3                        -892265.6   427124.5  -2.089  0.03677 *
## fuel4                        -239327.4    89547.9  -2.673  0.00755 **
## fuel5                        -290577.3    14227.6 -20.424  < 2e-16 ***
## seller_type1                  66383.3    16477.5   4.029 5.70e-05 ***
## seller_type3                 233924.3    43393.9   5.391 7.39e-08 ***
## transmission1                870336.6    22015.3  39.533  < 2e-16 ***
## ownerFourth & Above Owner     -1454.4    49857.0  -0.029  0.97673
## ownerSecond Owner            -40932.2    16678.1  -2.454  0.01416 *
## ownerTest Drive Car          168678.6   104824.6   1.609  0.10766
## ownerThird Owner             -39927.6    27782.7  -1.437  0.15075
## age_of_veh                   -35256.9     1918.3 -18.379  < 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
```

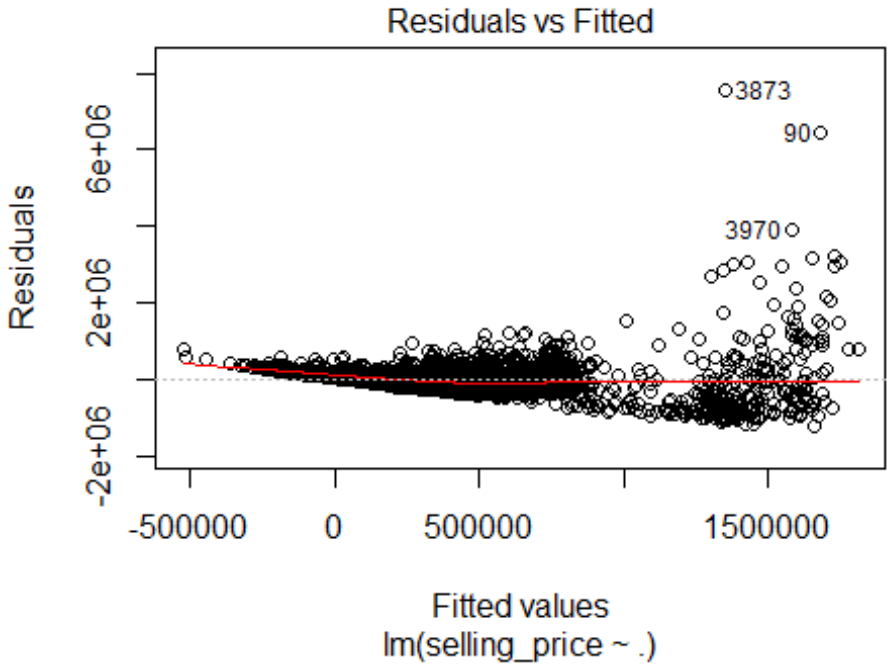
Above is the summary of the simple linear model on the whole data set. The first thing we can see is that the adj R^2 is 0.4576, this is not the best, but also not the worst. The second thing we can see from this is that given all the variables are in the model, all but: "ownerFourth & Above owner", "ownerTest Drive Car", and "ownerThird owner" has a significant effect on the model.

```
anova(LM)
```

```
## Analysis of Variance Table
##
## Response: selling_price
##                Df     Sum Sq    Mean Sq  F value    Pr(>F)
## km_driven       1 5.3700e+13 5.3700e+13  295.803 < 2.2e-16 ***
## fuel            4 1.8049e+14 4.5122e+13  248.552 < 2.2e-16 ***
## seller_type     2 4.6354e+13 2.3177e+13  127.669 < 2.2e-16 ***
## transmission    1 3.0610e+14 3.0610e+14 1686.107 < 2.2e-16 ***
## owner           4 1.9034e+13 4.7585e+12   26.212 < 2.2e-16 ***
## age_of_veh      1 6.1324e+13 6.1324e+13  337.797 < 2.2e-16 ***
## Residuals    4326 7.8535e+14 1.8154e+11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table shows us that all the variables in this dataset are important in predicting the selling price. The rest of the ANOVA tables showed the same results so they will not be included in the results.

```
plot(LM, which = 1)
```



The residual plot for the simple regression model on the complete data set doses show that the data could be modeled well by a linear model. As there are no real patterns in this plot, and the points do seem to be well scattered. This plot also dose show 3 points of interest: 90, 3970, and 3873.

```r
outlierTest(LM)
```

```
##       rstudent unadjusted p-value Bonferroni p
## 3873 18.421877         5.1019e-73   2.2137e-69
## 90   15.619485         1.5192e-53   6.5920e-50
## 3970  9.301670         2.1400e-20   9.2856e-17
## 556   7.612903         3.2703e-14   1.4190e-10
## 575   7.612903         3.2703e-14   1.4190e-10
## 594   7.612903         3.2703e-14   1.4190e-10
## 613   7.612903         3.2703e-14   1.4190e-10
## 901   7.612903         3.2703e-14   1.4190e-10
## 920   7.612903         3.2703e-14   1.4190e-10
## 1024  7.612903         3.2703e-14   1.4190e-10
```

The outlier test reviled ten possible outliers, three of which were the points of interest from the residual plot. All ten points were removed from the data set

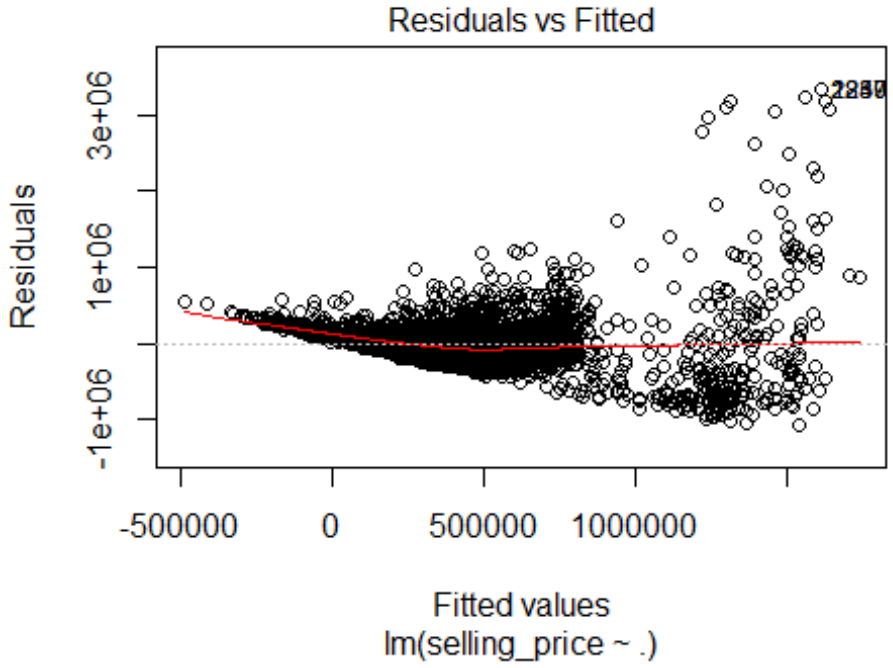
```r
DatanoOut <- CarData[-c(3873,90,3970,556,575,594,613,901,920,1024,1244),]
contrasts(DatanoOut$fuel) <- contr.treatment(5, base = 2)
contrasts(DatanoOut$transmission) <- contr.treatment(2, base = 2)
contrasts(DatanoOut$seller_type) <- contr.treatment(3, base = 2)
LM.noOut <- lm(selling_price~., DatanoOut)
summary(LM.noOut)
```

```
##
## Call:
## lm(formula = selling_price ~ ., data = DatanoOut)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1066752  -157858   -20898   108708  3334814
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  852766.5    15469.1  55.127  < 2e-16 ***
## km_driven                      -979.1      153.4  -6.383 1.92e-10 ***
## fuel1                        -280417.1    59348.0  -4.725 2.37e-06 ***
## fuel3                        -794124.3   371793.3  -2.136  0.03274 *
## fuel4                        -243310.6    77945.9  -3.122  0.00181 **
## fuel5                        -283983.5    12497.4 -22.723  < 2e-16 ***
## seller_type1                  37311.3    14386.0   2.594  0.00953 **
## seller_type3                 247249.7    37778.6   6.545 6.66e-11 ***
## transmission1                788309.1    19303.8  40.837  < 2e-16 ***
## ownerFourth & Above Owner    -11641.4    43406.0  -0.268  0.78856
## ownerSecond Owner            -39970.2    14528.0  -2.751  0.00596 **
## ownerTest Drive Car          204881.4    91251.3   2.245  0.02480 *
## ownerThird Owner             -43015.1    24204.8  -1.777  0.07562 .
## age_of_veh                   -33827.6     1677.8 -20.161  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 370900 on 4315 degrees of freedom
## Multiple R-squared:  0.4881, Adjusted R-squared:  0.4866
## F-statistic: 316.5 on 13 and 4315 DF,  p-value: < 2.2e-16
```

The summary for the new regression model that was constructed once the outliers were removed shows us that removing the outliers increased the adj R^2 of the data to 0.4866, as well as made "ownerTest Drive Car" relevant to the model.

```
plot(LM.noOut, which = 1)
```



From this new residual plot we can see that removing the outliers, definitely improved the fit of the linear model, as the points are even more scattered, and uniform than before.

```
robLM <- rlm(selling_price~., CarData, psi = psi.huber)
summary(robLM)
```

```
##
## Call: rlm(formula = selling_price ~ ., data = CarData, psi = psi.huber)
## Residuals:
##     Min      1Q  Median      3Q     Max
## -683021 -116489   -8162  113045 7985245
##
## Coefficients:
##                             Value        Std. Error   t value
## (Intercept)                  724592.7637    8032.3498      90.2093
## km_driven                      -549.0198      77.2087      -7.1109
## fuel1                        -211437.8918   31278.0136      -6.7600
```

```
## fuel3                       -440185.0662  195954.7193      -2.2464
## fuel4                       -197704.9927   41082.4977      -4.8124
## fuel5                       -191725.5244    6527.2838     -29.3729
## seller_type1                  42360.4171    7559.5068       5.6036
## seller_type3                 240098.2580   19908.0857      12.0603
## transmission1                465393.1351   10100.0930      46.0781
## ownerFourth & Above Owner    -26834.6112   22873.2176      -1.1732
## ownerSecond Owner            -32571.3064    7651.5146      -4.2568
## ownerTest Drive Car          278367.9681   48091.0876       5.7883
## ownerThird Owner             -39257.6259   12746.0567      -3.0800
## age_of_veh                   -29682.1941     880.0706     -33.7271
##
## Residual standard error: 171400 on 4326 degrees of freedom
```

The summary of the robust residual model seems to match pretty close with the summaries of the other two models.

The results of the third linear model that was done on the data which had two sets of outliers removed will not be shown in the results as the results don't differ much from the results of the second model

```
pred.LM <- predict(LM, newdata = CarData)
pred.LM.noOut <- predict(LM.noOut, newdata = DatanoOut)
pred.LM.RobLM <- predict(robLM, newdata = CarData)

(MSPE.1 = mean((CarData[,"selling_price"] - pred.LM)^2))

## [1] 180955157231

(MSPE.2 = mean((DatanoOut[,"selling_price"] - pred.LM.noOut)^2))

## [1] 137096838085

(MSPE.4 = mean((CarData[,"selling_price"] - pred.LM.RobLM)^2))

## [1] 206570828636
```

To check the strength of our prediction models we calculated the MSPE of all four of our models (the third regression model with be in the appendix) as it did not change drastically from that of the second regression model. We can see that the second regression model has the lowest MSPE of the three models shown

Discussion/Conclusion:

From the results of the ANOVA table we can see that all the variables are important to the in modeling this data. The means that: km driven, type of fuel, seller type, type of transmission, and number of owners, as well as age of the vehicle do play a role in the selling price of said vehicle. From all three of summary reports for the models run we can see that km driven, and age of the vehicle have a negative correlation with the selling price given that all the other variables are in the model. We can also see that fuel types: 1,3,4,5 (CNG, Electric LPG, and petrol respectively) reduce the selling price of the vehicle when

compared to diesel given all other variables are in the model. An interesting finding from the summary reports of the models is that if the vehicle has had two owners the price will decrease when compared to a vehicle with only one owner; however if the car was a test drive car then the price of that vehicle will increase when compared to a vehicle with only one owner. Also having three or more owners does not meaningfully affect the selling price of that vehicle. The last finding from the summary reports is that a vehicle sold by seller types 1 and 3 (deal, and trustmark dealer) were sold at higher prices than vehicles sold by individuals.

Our MSPE tests showed that the third regression model (see appendix) had the best MSPE of all four models tested; however it was not much better than that of the second regression model run, and due to the fact that the second regression model is simpler and simpler models tend to be better we have chosen that model to be our prediction model if predictions need to be made. Future studies could be done to determine if this model is the best at predicting sells price or if other models like a GAM model could do a better job.

Appendix:

Code that was also run on our data set, but results of which are not shown in the results section. This part is added for reproducibility

```
#plot(LM)

#leveragePlots(LM)
#influencePlot(LM, main="Influence Plot", sub="Circle size is proportial to C
ook's Distance" )



#anova(LM.noOut)

#plot(LM.noOut)
#outlierTest(LM.noOut)



#DatanoOut.2 <- DatanoOut[-c(1837,2240,2259,2740,3454,102,2581,3884,540,2179)
,]
#LM.noOut.2 <- lm(selling_price~., DatanoOut.2)
#summary(LM.noOut.2)
#plot(LM.noOut.2)



#pred.LM.noOut.2 <- predict(LM.noOut.2, newdata = DatanoOut.2)
#(MSPE.3 = mean((DatanoOut.2[,"selling_price"] - pred.LM.noOut.2)^2))
```