**ITM 818 Data Management and Visualization**

Homework 4: Data Visualization and Modeling in R (60 points)

**Part I**. **Used car auctions**. One of the biggest challenges of an auto dealership purchasing a used car at an auto auction is the risk that the vehicle might have serious issues that prevent it from being sold to customers. The auto community calls these unfortunate purchases “kicks.”

Kicked cars often result when there are tampered odometers, mechanical issues the dealer is not able to address, issues with getting the vehicle title from the seller, or some other unforeseen problem. Kicked cars can be very costly to dealers for transportation cost, throw-away repair work, and market losses in reselling the vehicle.

Modelers who can figure out which cars have a higher risk of being kicked can provide real value to dealerships trying to provide the best inventory selection possible to their customers.

The goal of this problem is to predict if the car purchased at the auction is a Kick (bad buy).

**Dataset:** please find the CSV dataset and the Excel data dictionary. Each observation is a transaction. IsBadBuy is the response variable.

**Please use R (any packages such as ggplot2 and boot, etc.) to answer the following questions. (30 points)**

1. Do the proportions of vehicle **sizes** differ across **American manufacturers**? Which American manufacturers prefer van (in terms of percentage)? Please develop a single visualization to demonstrate the answer. (5 points)

**Code**:

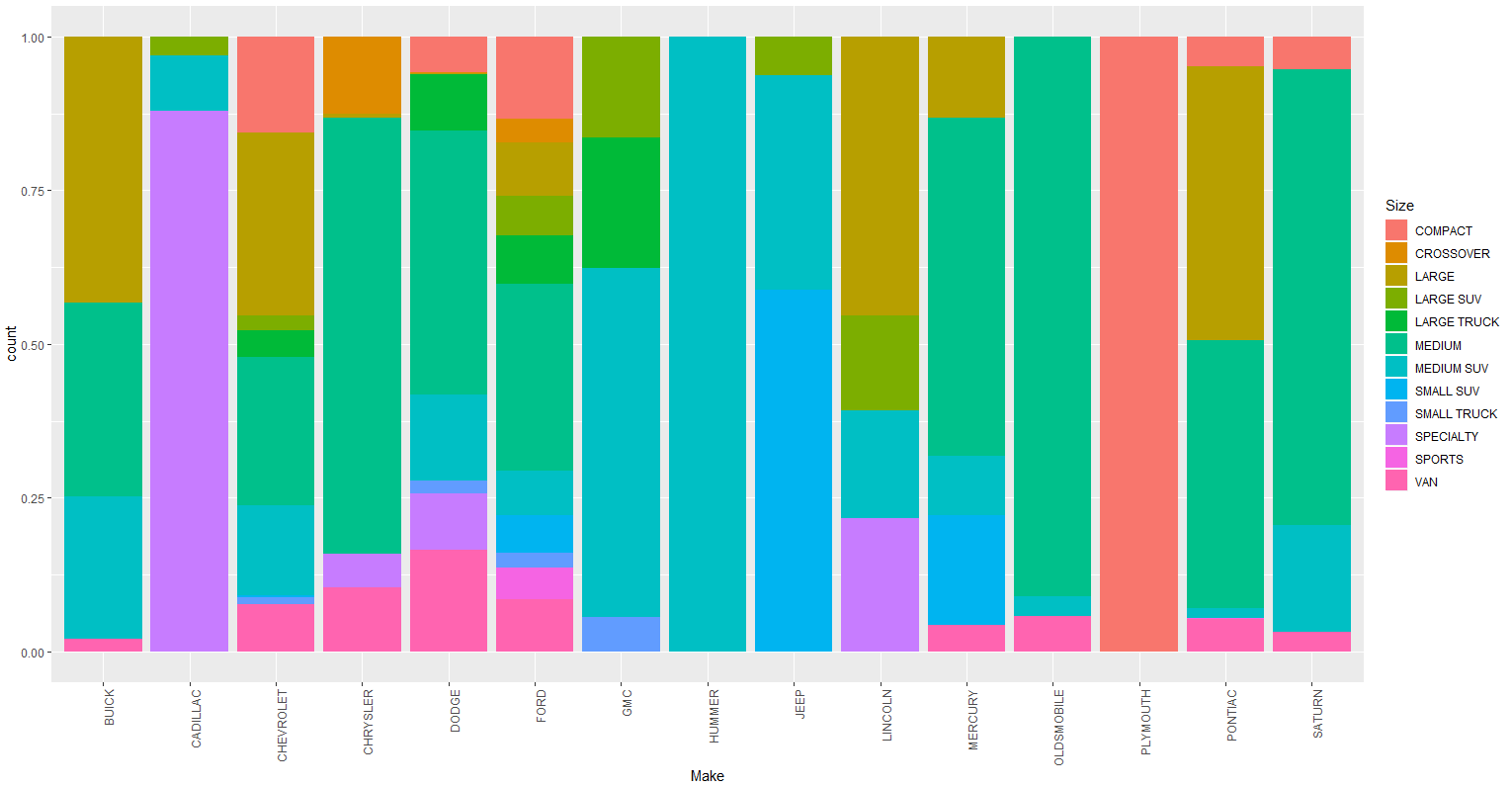
d1 = sqldf("select \* from kickedCars where Nationality = 'AMERICAN'")

q1 = ggplot(data = d1) + geom\_bar(mapping=aes(x=Make,fill=Size),position="fill") +

theme(axis.text.x = element\_text(angle = 90, hjust = 1))

q1

**Output:**



**The Manufacturer “DODGE” prefers VAN which amounts to around 15% of total size options**

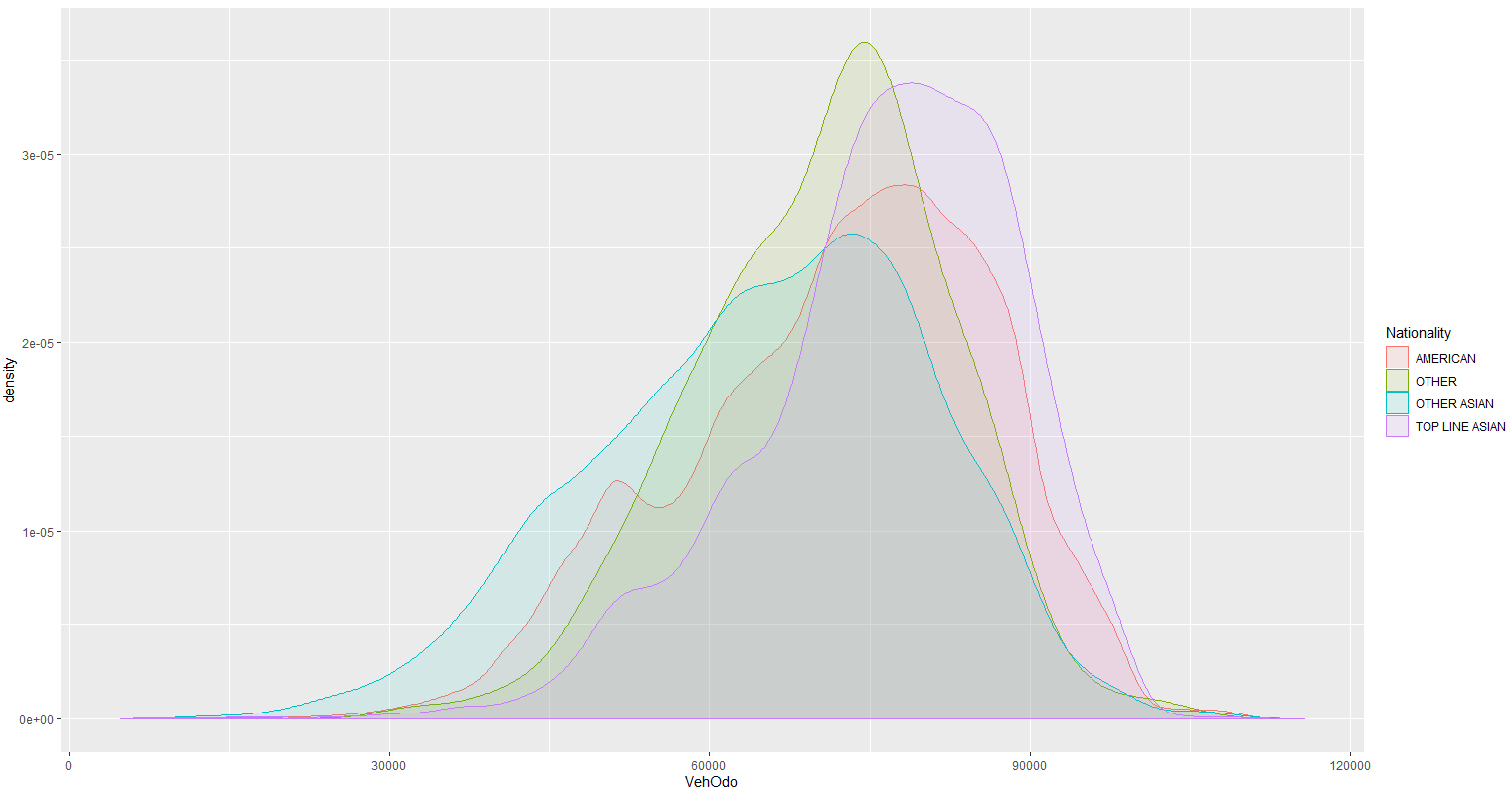
2. Compare the distributional density of continuous attribute **odometer** across different **nationalities**. Are the distributions very different from each other? (5 points)

**Code**:

q2=ggplot(data=kickedCars)+ geom\_area(aes(x=VehOdo,fill = Nationality, color=Nationality),stat="density",alpha=0.1,position="identity")

q2

**Output:**



3. Use a **stacked bar chart** to show the **total number** of “bad buys” for each manufacturer and year. List the top 3 manufacturers that have the largest total number of bad buys. (Hint: notice that it is total number of transactions that are “bad buy,” not the total number of transactions.) (5 points)

**Code**:

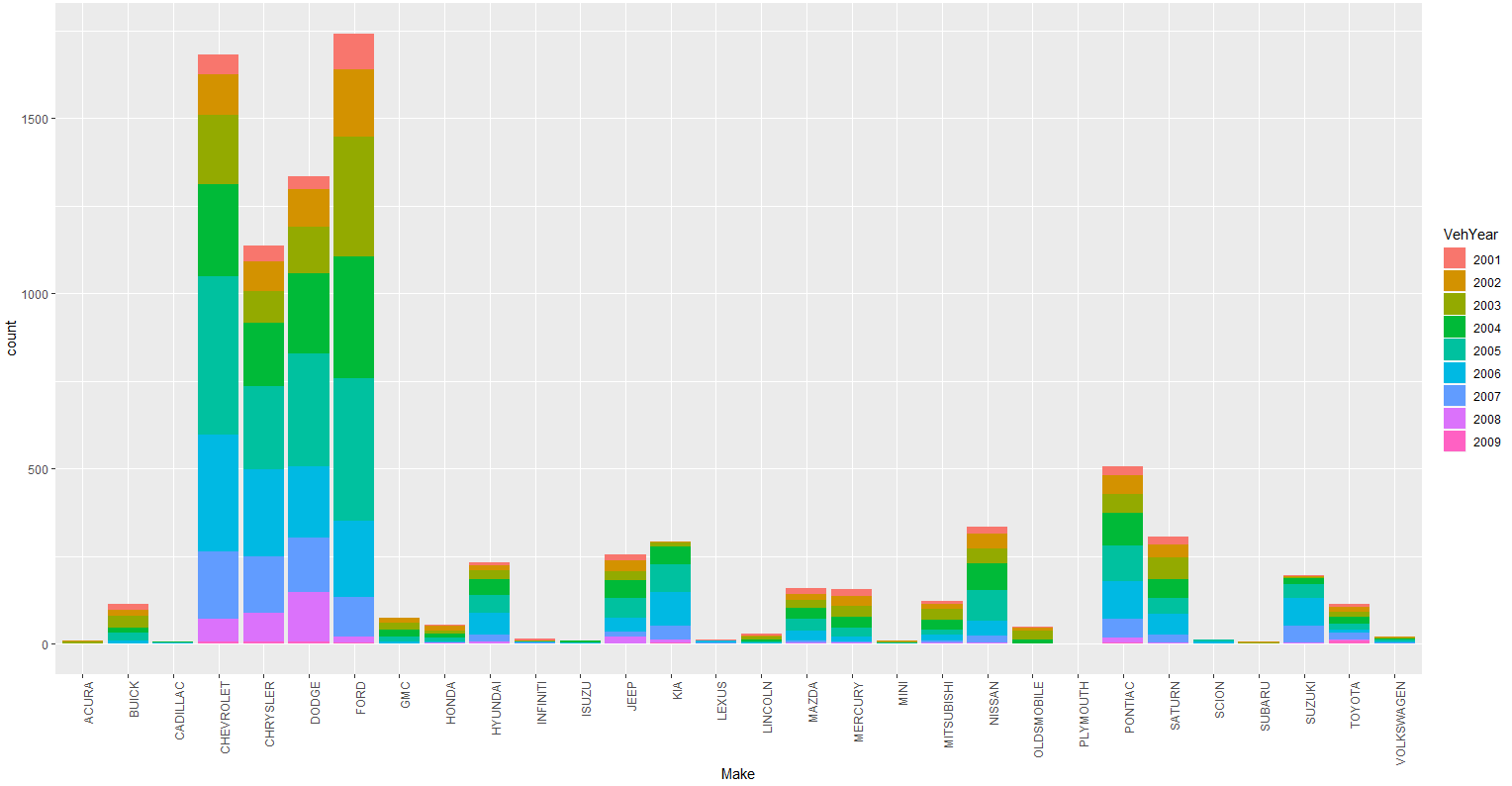
d3 = sqldf("select \* from kickedCars where IsBadBuy = '1' ")

q3 = ggplot(data=d3)+geom\_bar(mapping=aes(x= Make,fill=VehYear),position="stack") +

theme(axis.text.x = element\_text(angle = 90, hjust = 1))

q3

**Output:**



**top 3 manufacturers that have the largest total number of bad buys:**

1. **Ford**
2. **Chevrolet**
3. **Dodge**

4. Use **slide-by-slide boxplots** show the distribution of “MMRCurrentRetailAveragePrice” for each vehicle year. Draw the same graph for “MMRCurrentRetailCleanPrice” as well. Comparing the two graphs, which group has higher average prices? Is this conclusion consistent for different years? Hint: vehicle year must be converted to a factor data. (5 points)

**Code**:

q4a = ggplot(data=kickedCars,mapping=aes(x=VehYear,y=MMRCurrentRetailAveragePrice))+

stat\_boxplot()

q4b = ggplot(data=kickedCars,mapping=aes(x=VehYear,y=MMRCurrentRetailCleanPrice))+

stat\_boxplot()

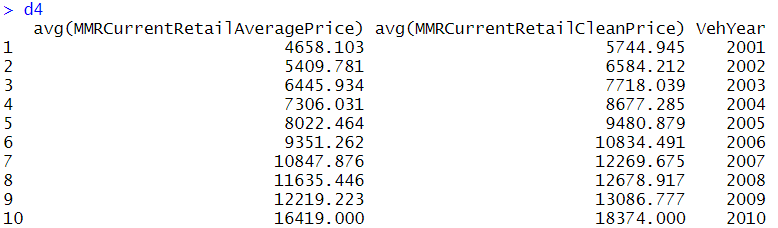
grid.arrange(q4a,q4b, ncol = 2)

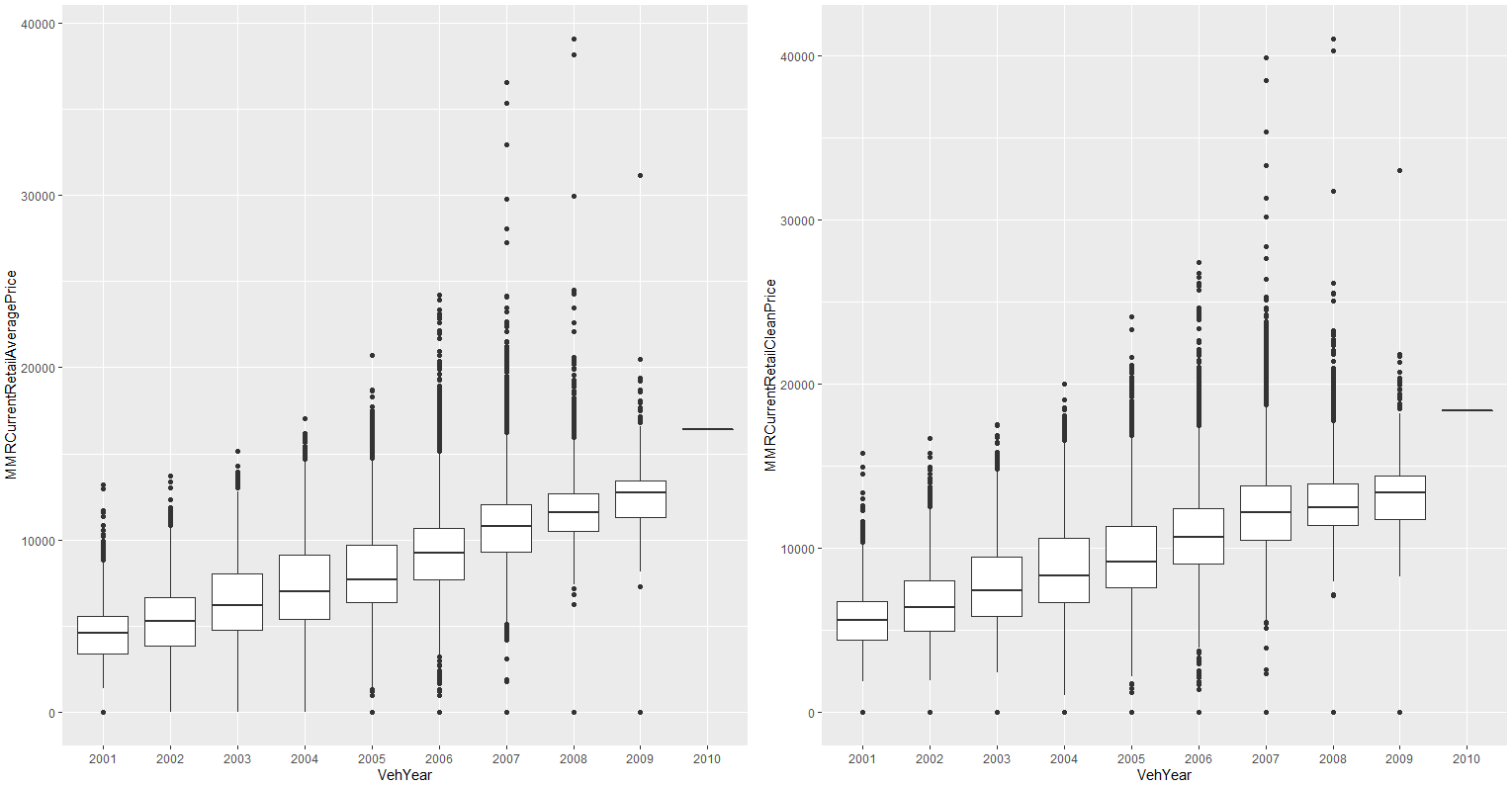
d4 = sqldf("Select avg(MMRCurrentRetailAveragePrice), avg(MMRCurrentRetailCleanPrice), VehYear from kickedCars group By VehYear")

**Output:**

Comparing the two graphs, MMRCurrentRetailCleanPrice group has higher average price.

Yes This conclusion is consistent for different years.





5. Focus on those car manufacturers that had more than 300 bad buys, and then show the “**proportion of** **bad buy”** for each manufacturer and vehicle year, but in a different visualization tool (do not use bar chart) Hint: use a chart for summary statistics. (5 points)

**Code**:

d5 = sqldf("Select Make, IsBadBuy, count(Make) as badcount from kickedCars

where IsBadBuy = 1 group by Make,IsBadBuy having badcount >300")

lst = d5$Make

d5a = kickedCars[kickedCars$Make %in% lst,]

d5b = sqldf("select Make, VehYear, count(Make) as totcount from d5a group by Make,VehYear")

d5c = sqldf("select Make, VehYear, count(Make) as badcount from d5a where IsBadBuy = 1

group by Make, VehYear")

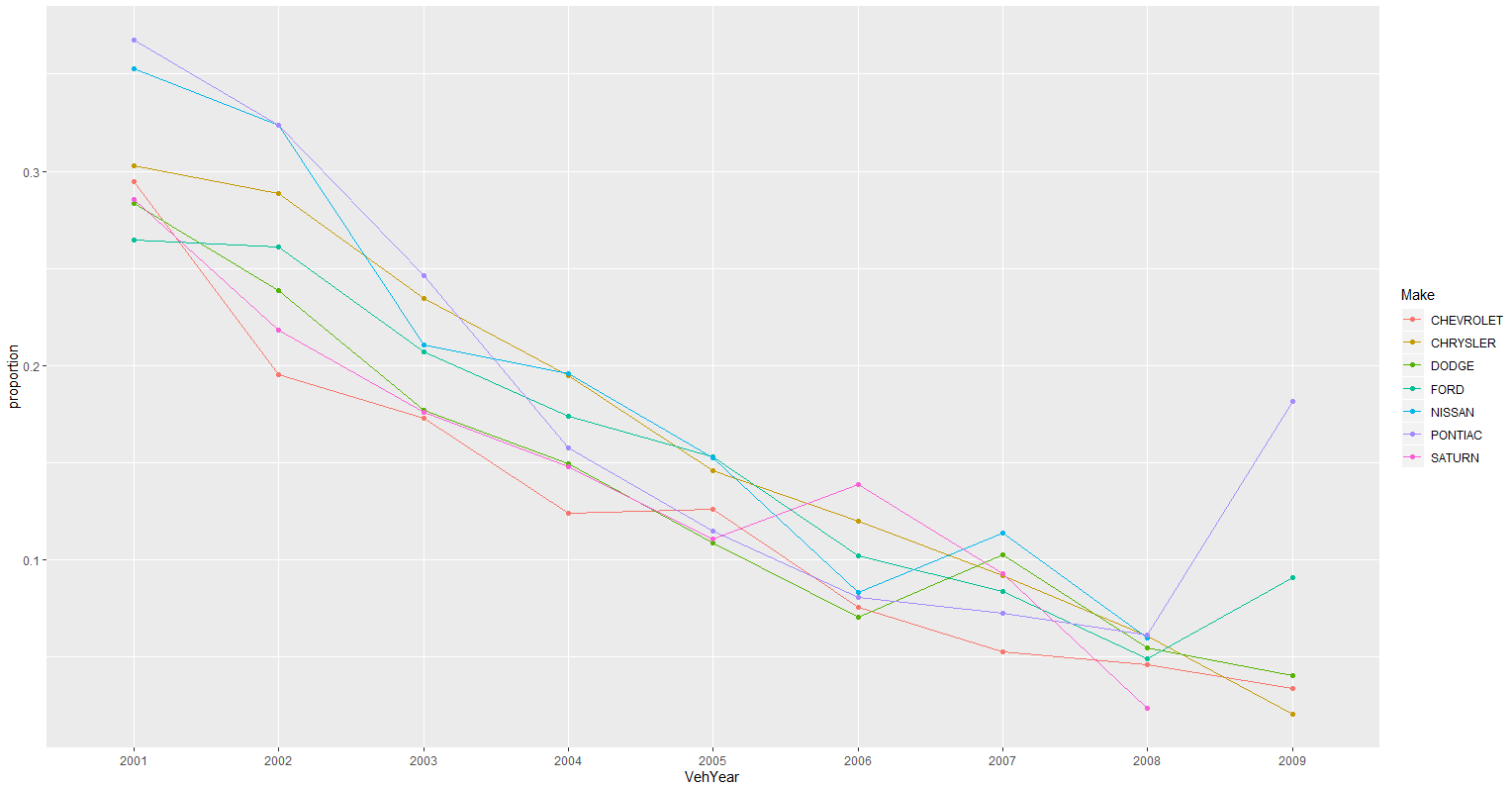
final = sqldf("select d5c.Make, d5c.VehYear, badcount, totcount from d5c, d5b

where d5c.Make = d5b.Make and d5c.VehYear = d5b.VehYear")

final$proportion = final$badcount/final$totcount

ggplot(data=final, aes(x=VehYear, y=proportion, group=Make)) + geom\_line(aes(color=Make)) + geom\_point(aes(color=Make))

**Output:**



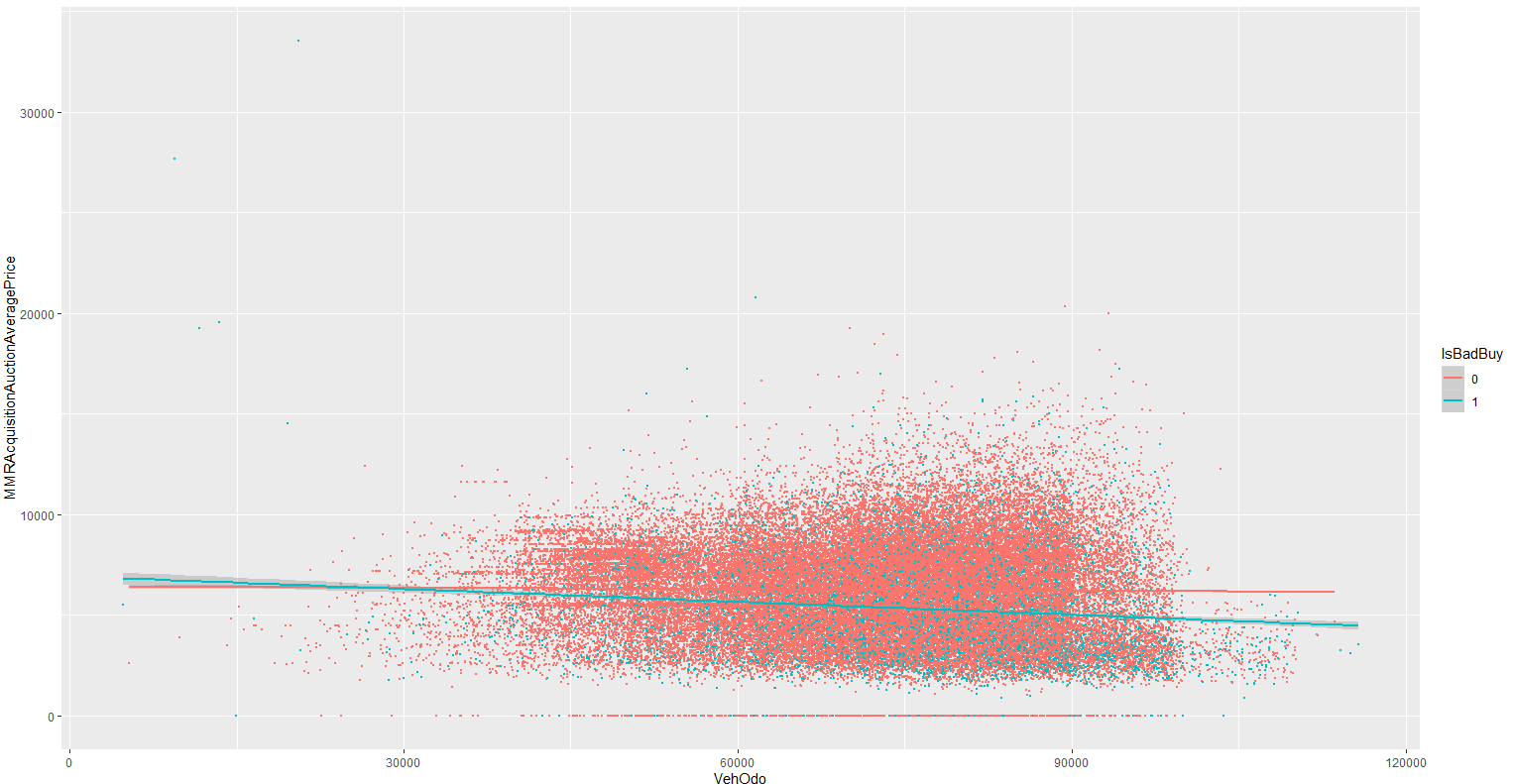
6. For vehicles of **top three American manufactures (**see attribute **TopThreeAmericanName)**, what is the relationship between **odometer** and auction price of the vehicle in average condition at the time of purchase (**MMRAcquisitionAuctionAveragePrice**)? Is that relationship different depending on whether the transactions are “bad buys?” Please use a single **scatter plot** to show the relationship and create two “linear model” smoothed lines with confidence intervals (standard errors), one for “good buys” and one for “bad buys.” Hint: use a small point size, so that the scatter plot is easier to view. (5 points)

**Code**:

ggplot(data=d1,mapping=aes(x=VehOdo,y=MMRAcquisitionAuctionAveragePrice, color = IsBadBuy))+ geom\_point(size = 0.005) + geom\_smooth(method = "lm")

**Output:**

Yes, the relationship is different depending on whether the transactions are “bad buys”



7. We then decide to use logistic regression to predict **IsBadBuy** using the following predictors: **VehicleAge**, **TopThreeAmericanName**, **WheelType**, **VehOdo**, **Size**, **VehBCost**, **IsOnline**, and **WarrantyCost**. Please apply the validation set approach to evaluate performance of the model in terms of **accuracy**, **precision** and **recall**. Hint: use sample() function to generate a vector of TRUE/FALSE with probabilities 0.7 and 0.3:

train.index=sample(c(TRUE,FALSE),size=nrow(data),prob=c(0.7,0.3),replace=TRUE).

Given that the proportions of bad buy and good buy are 87.7% and 12.3%, respectively, what can we comment on the prediction performance of the logistic regression model? (10 points)

**Code**:

set.seed(1)

train.index=sample(c(TRUE,FALSE),size=nrow(kickedCars),prob=c(0.7,0.3),replace=TRUE)

train=kickedCars[train.index,]

valid=kickedCars[-train.index,]

logit7=glm(IsBadBuy~VehicleAge + TopThreeAmericanName + WheelType + VehOdo + Size + VehBCost + IsOnlineSale + WarrantyCost,family=binomial(link="logit"),data=train)

summary(logit7)

valid$prob=predict(logit7,valid,type="response")

#myROC = roc(valid$IsBadBuy, valid$prob)

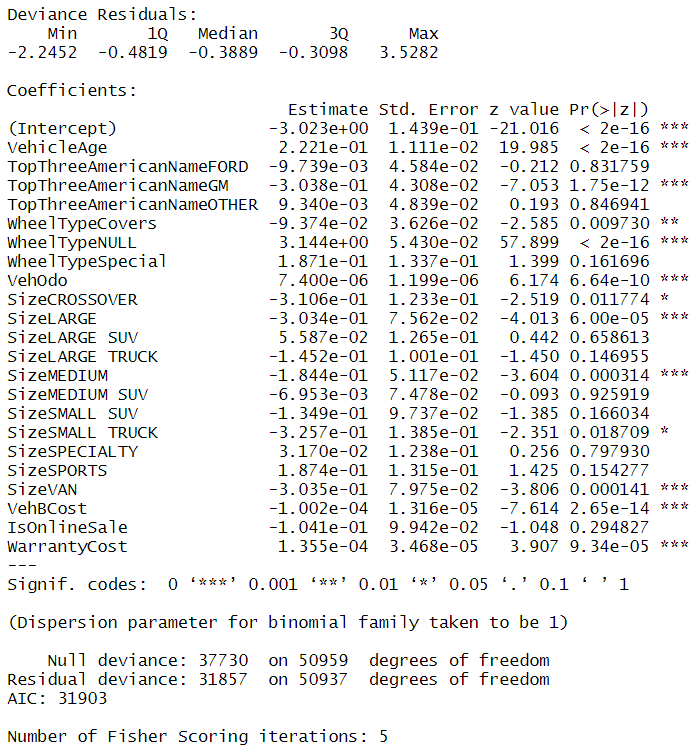
#a = coords(myROC, "best")

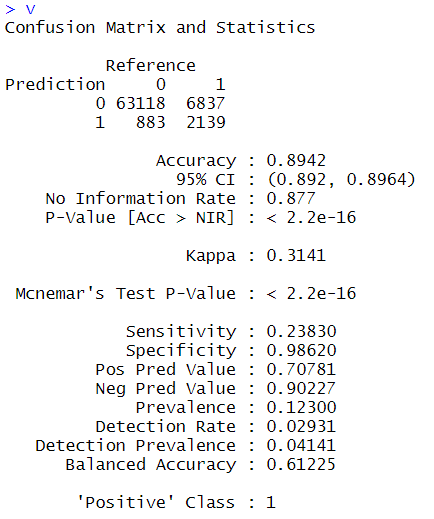
valid$pred=ifelse(valid$prob>0.5,1,0)

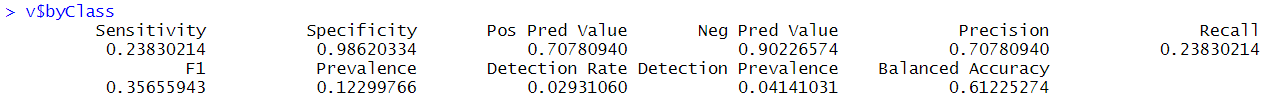
v = confusionMatrix(as.factor(valid$pred), as.factor(valid$IsBadBuy), positive = "1")

v$byClass

Output:







Due to the imbalance in data where, the proportions of bad buy and good buy are 87.7% and 12.3%, respectively We can say that the model is better at predicting bad buy.

8. Finally, we apply a logistic regression model with the same set of predictors as (7), using **10-fold cross-validation** to evaluate its predictive performance. Please calculate the cross-validation-based (average) **accuracy**, **precision,** and **recall**. Hint: use the “boot” package. (10 points)

Code:

set.seed(1)

logit8=glm(IsBadBuy~VehicleAge + TopThreeAmericanName + WheelType + VehOdo + Size + VehBCost + IsOnlineSale + WarrantyCost,family=binomial(link="logit"),data=kickedCars)

summary(logit8)

cost.error=function(r, pi = 0) mean(abs(r-pi)>0.5)

cv.error=boot::cv.glm(kickedCars,logit8,cost=cost.error,K=10)

cost.accuracy=function(r, pi = 0) mean(abs(r-pi)<0.5)

cv.accuracy=boot::cv.glm(kickedCars,logit8,cost=cost.accuracy,K=10)

cost.precision=function(r, pi = 0) {

TP=sum((pi>0.5)&(r==1))

FP=sum((pi>0.5)&(r==0))

return(TP/(TP+FP))

}

cost.recall = function(r, pi = 0){

TP = sum((pi > 0.5) & (r == 1))

FN = sum((pi < 0.5) & (r == 1))

return(TP/(TP+FN))

}

cv.precision=boot::cv.glm(kickedCars,logit8,cost=cost.precision,K=10)

cv.recall=boot::cv.glm(kickedCars,logit8,cost=cost.recall,K=10)

c(cv.error$delta[1],cv.accuracy$delta[1],cv.precision$delta[1],cv.recall$delta[1])

**Output:**



9. In question (7) and (8), we only add a subset of the predictors. Discuss the possibility of including additional predictors into the model. Which variables should or should not be included? Why? This is an open question; you should discuss at least three variables. (10 points)

**Code:**

# Base Model with added Multiple variables

summary(kickedCars)

logit9=glm(IsBadBuy~Auction + VehicleAge+ Make + Color + Transmission

+ WheelTypeID + WheelType + VehOdo + Nationality + Size + TopThreeAmericanName +

MMRAcquisitionAuctionAveragePrice+ MMRAcquisitionAuctionCleanPrice + MMRAcquisitionRetailAveragePrice +

MMRAcquisitonRetailCleanPrice + MMRCurrentAuctionAveragePrice + MMRCurrentAuctionCleanPrice +

MMRCurrentRetailAveragePrice + MMRCurrentRetailCleanPrice + PRIMEUNIT + AUCGUART +

VehBCost + IsOnlineSale + WarrantyCost,family=binomial(link="logit"),data=kickedCars)

summary(logit9)

# Reduced, Based on P-Valus

logit9a=glm(IsBadBuy~Auction + VehicleAge+

WheelType + VehOdo + Size + TopThreeAmericanName +

MMRAcquisitionAuctionAveragePrice+ MMRAcquisitionRetailAveragePrice +

MMRCurrentAuctionAveragePrice +

VehBCost + IsOnlineSale + WarrantyCost,family=binomial(link="logit"),data=kickedCars)

summary(logit9a)

# Exhaustive Search

search=leaps::regsubsets(IsBadBuy~Auction + VehicleAge+

WheelType + VehOdo + Size + TopThreeAmericanName +

MMRAcquisitionAuctionAveragePrice+ MMRAcquisitionRetailAveragePrice +

MMRCurrentAuctionAveragePrice +

VehBCost + IsOnlineSale + WarrantyCost,data=kickedCars,nbest=1,nvmax=27,

method="exhaustive")

res=summary(search)

names(res)

res$which

plot(1:length(res$adjr2),res$adjr2,type="b")

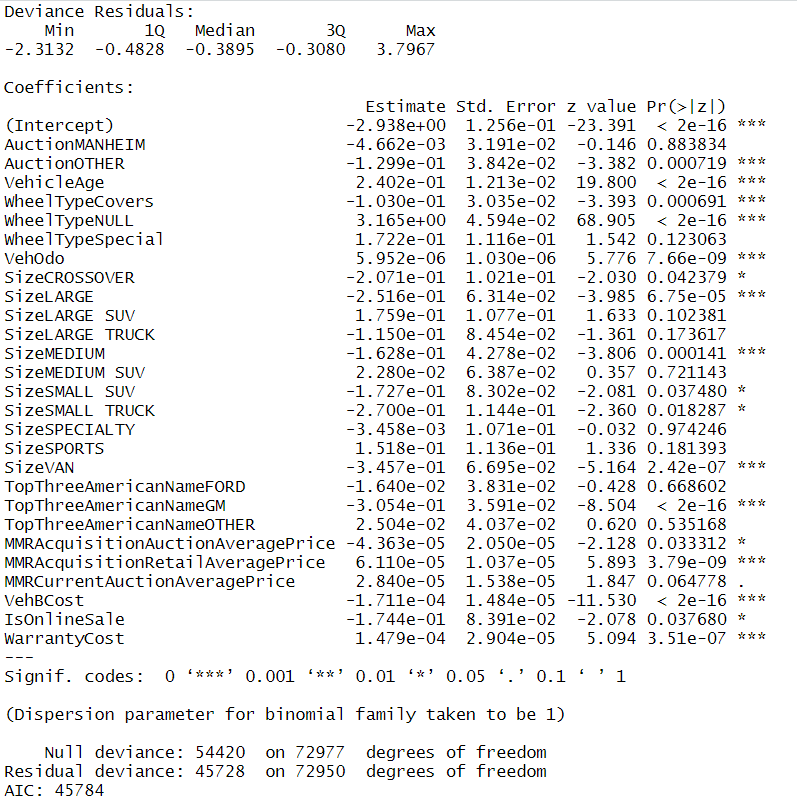
plot(1:length(res$adjr2),res$bic,typ="b",col="red")

**Output:**

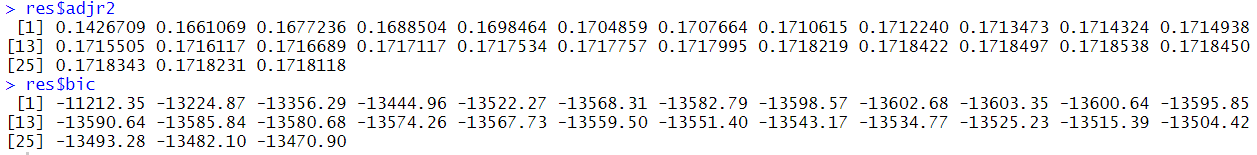
# Checking P values

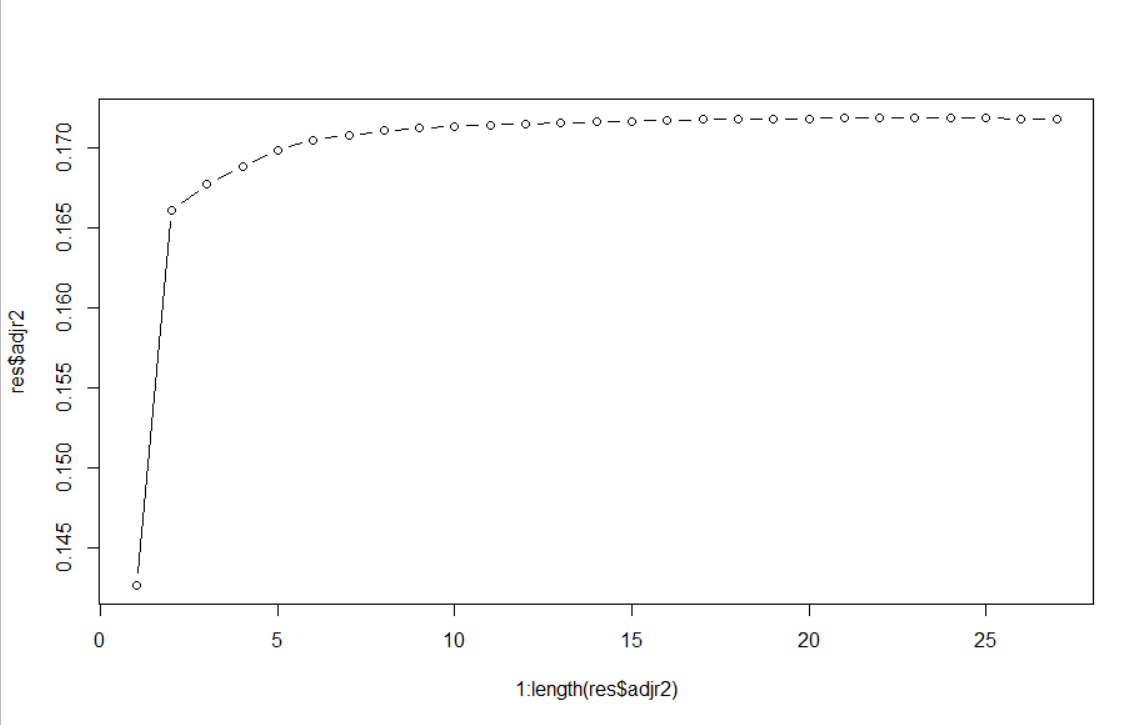
|  |
| --- |
| Coefficients: (8 not defined because of singularities) |
| Estimate Std. Error z value Pr(>|z|) |
| (Intercept) -4.114e+00 5.976e+02 -0.007 0.994508 |
| AuctionMANHEIM 4.390e-03 3.228e-02 0.136 0.891826 |
| AuctionOTHER -1.163e-01 3.894e-02 -2.986 0.002828 \*\* |
| VehicleAge 2.354e-01 1.357e-02 17.353 < 2e-16 \*\*\* |
| MakeBUICK -1.184e+01 2.654e+02 -0.045 0.964417 |
| MakeCADILLAC -1.145e+01 2.654e+02 -0.043 0.965597 |
| MakeCHEVROLET -1.181e+01 2.654e+02 -0.044 0.964517 |
| MakeCHRYSLER -1.131e+01 2.654e+02 -0.043 0.966014 |
| MakeDODGE -1.157e+01 2.654e+02 -0.044 0.965234 |
| MakeFORD -1.149e+01 2.654e+02 -0.043 0.965464 |
| MakeGMC -1.196e+01 2.654e+02 -0.045 0.964064 |
| MakeHONDA -1.196e+01 2.654e+02 -0.045 0.964076 |
| MakeHUMMER -2.225e+01 5.976e+02 -0.037 0.970297 |
| MakeHYUNDAI -3.466e-01 4.687e-01 -0.740 0.459518 |
| MakeINFINITI 7.092e-01 5.815e-01 1.220 0.222599 |
| MakeISUZU -9.746e-01 5.912e-01 -1.648 0.099263 . |
| MakeJEEP -1.132e+01 2.654e+02 -0.043 0.965977 |
| MakeKIA -1.655e-01 4.679e-01 -0.354 0.723511 |
| MakeLEXUS 1.310e+00 6.187e-01 2.117 0.034285 \* |
| MakeLINCOLN -1.085e+01 2.654e+02 -0.041 0.967387 |
| MakeMAZDA -8.884e-02 4.715e-01 -0.188 0.850561 |
| MakeMERCURY -1.136e+01 2.654e+02 -0.043 0.965871 |
| MakeMINI -1.039e+01 2.654e+02 -0.039 0.968791 |
| MakeMITSUBISHI -6.057e-01 4.746e-01 -1.276 0.201822 |
| MakeNISSAN -1.127e+01 2.654e+02 -0.042 0.966123 |
| MakeOLDSMOBILE -1.160e+01 2.654e+02 -0.044 0.965148 |
| MakePLYMOUTH -1.057e+01 2.655e+02 -0.040 0.968227 |
| MakePONTIAC -1.161e+01 2.654e+02 -0.044 0.965109 |
| MakeSATURN -1.165e+01 2.654e+02 -0.044 0.964981 |
| MakeSCION -5.601e-01 5.765e-01 -0.972 0.331229 |
| MakeSUBARU 6.964e-01 7.046e-01 0.988 0.322973 |
| MakeSUZUKI 1.862e-01 4.718e-01 0.395 0.693070 |
| MakeTOYOTA -1.164e+01 2.654e+02 -0.044 0.965011 |
| MakeTOYOTA SCION -1.179e+01 5.354e+02 -0.022 0.982430 |
| MakeVOLKSWAGEN -1.139e+01 2.654e+02 -0.043 0.965779 |
| MakeVOLVO -2.250e+01 2.792e+02 -0.081 0.935754 |
| ColorBLACK 5.058e-02 9.354e-02 0.541 0.588681 |
| ColorBLUE -3.873e-02 9.121e-02 -0.425 0.671073 |
| ColorBROWN 1.470e-01 1.765e-01 0.833 0.405043 |
| ColorGOLD 7.364e-02 9.525e-02 0.773 0.439454 |
| ColorGREEN -1.086e-01 1.029e-01 -1.056 0.291045 |
| ColorGREY 2.431e-02 9.288e-02 0.262 0.793504 |
| ColorMAROON 3.442e-02 1.114e-01 0.309 0.757319 |
| ColorNOT AVAIL -2.529e-01 3.380e-01 -0.748 0.454276 |
| ColorNULL 7.486e+00 5.354e+02 0.014 0.988845 |
| ColorORANGE 1.349e-02 2.106e-01 0.064 0.948937 |
| ColorOTHER -1.262e+00 2.538e-01 -4.972 6.61e-07 \*\*\* |
| ColorPURPLE 1.493e-01 1.849e-01 0.807 0.419602 |
| ColorRED 1.087e-01 9.414e-02 1.154 0.248324 |
| ColorSILVER 5.151e-02 8.873e-02 0.581 0.561512 |
| ColorWHITE 4.005e-02 8.968e-02 0.447 0.655180 |
| ColorYELLOW -2.246e-01 2.226e-01 -1.009 0.312849 |
| TransmissionAUTO 1.091e+01 5.354e+02 0.020 0.983741 |
| TransmissionManual -4.762e-01 7.572e+02 -0.001 0.999498 |
| TransmissionMANUAL 1.077e+01 5.354e+02 0.020 0.983948 |
| TransmissionNULL NA NA NA NA |
| WheelTypeID 8.308e-02 5.626e-02 1.477 0.139760 |
| WheelTypeCovers -1.632e-01 6.146e-02 -2.656 0.007906 \*\* |
| WheelTypeNULL 3.272e+00 7.562e-02 43.267 < 2e-16 \*\*\* |
| WheelTypeSpecial NA NA NA NA |
| VehOdo 9.167e-06 1.089e-06 8.415 < 2e-16 \*\*\* |
| NationalityOTHER NA NA NA NA |
| NationalityOTHER ASIAN -1.124e+01 2.654e+02 -0.042 0.966229 |
| NationalityTOP LINE ASIAN NA NA NA NA |
| SizeCROSSOVER -3.485e-01 1.087e-01 -3.207 0.001339 \*\* |
| SizeLARGE -3.241e-01 6.586e-02 -4.921 8.63e-07 \*\*\* |
| SizeLARGE SUV 1.177e-01 1.150e-01 1.023 0.306214 |
| SizeLARGE TRUCK -1.548e-01 9.046e-02 -1.711 0.087107 . |
| SizeMEDIUM -2.295e-01 4.568e-02 -5.024 5.06e-07 \*\*\* |
| SizeMEDIUM SUV 2.069e-02 6.936e-02 0.298 0.765477 |
| SizeSMALL SUV -3.250e-01 9.337e-02 -3.481 0.000500 \*\*\* |
| SizeSMALL TRUCK -2.431e-01 1.163e-01 -2.090 0.036591 \* |
| SizeSPECIALTY -1.073e-01 1.181e-01 -0.909 0.363611 |
| SizeSPORTS 1.629e-01 1.178e-01 1.383 0.166732 |
| SizeVAN -3.549e-01 7.020e-02 -5.055 4.30e-07 \*\*\* |
| TopThreeAmericanNameFORD NA NA NA NA |
| TopThreeAmericanNameGM NA NA NA NA |
| TopThreeAmericanNameOTHER NA NA NA NA |
| MMRAcquisitionAuctionAveragePrice -3.335e-04 8.007e-05 -4.165 3.11e-05 \*\*\* |
| MMRAcquisitionAuctionCleanPrice 9.960e-05 6.992e-05 1.424 0.154304 |
| MMRAcquisitionRetailAveragePrice 2.555e-04 4.983e-05 5.127 2.94e-07 \*\*\* |
| MMRAcquisitonRetailCleanPrice -5.434e-05 4.600e-05 -1.181 0.237428 |
| MMRCurrentAuctionAveragePrice 1.532e-04 7.798e-05 1.964 0.049477 \* |
| MMRCurrentAuctionCleanPrice 5.879e-05 6.920e-05 0.850 0.395591 |
| MMRCurrentRetailAveragePrice -9.483e-05 5.029e-05 -1.886 0.059351 . |
| MMRCurrentRetailCleanPrice -5.922e-05 4.596e-05 -1.289 0.197572 |
| PRIMEUNITNULL 1.530e+00 9.792e-02 15.629 < 2e-16 \*\*\* |
| PRIMEUNITYES 1.707e+00 4.462e-01 3.825 0.000131 \*\*\* |
| AUCGUARTNULL NA NA NA NA |
| AUCGUARTRED 4.036e-01 3.922e-01 1.029 0.303394 |
| VehBCost -1.498e-04 1.591e-05 -9.415 < 2e-16 \*\*\* |
| IsOnlineSale -1.607e-01 8.495e-02 -1.892 0.058511 . |
| WarrantyCost 1.086e-04 3.099e-05 3.503 0.000459 \*\*\* |

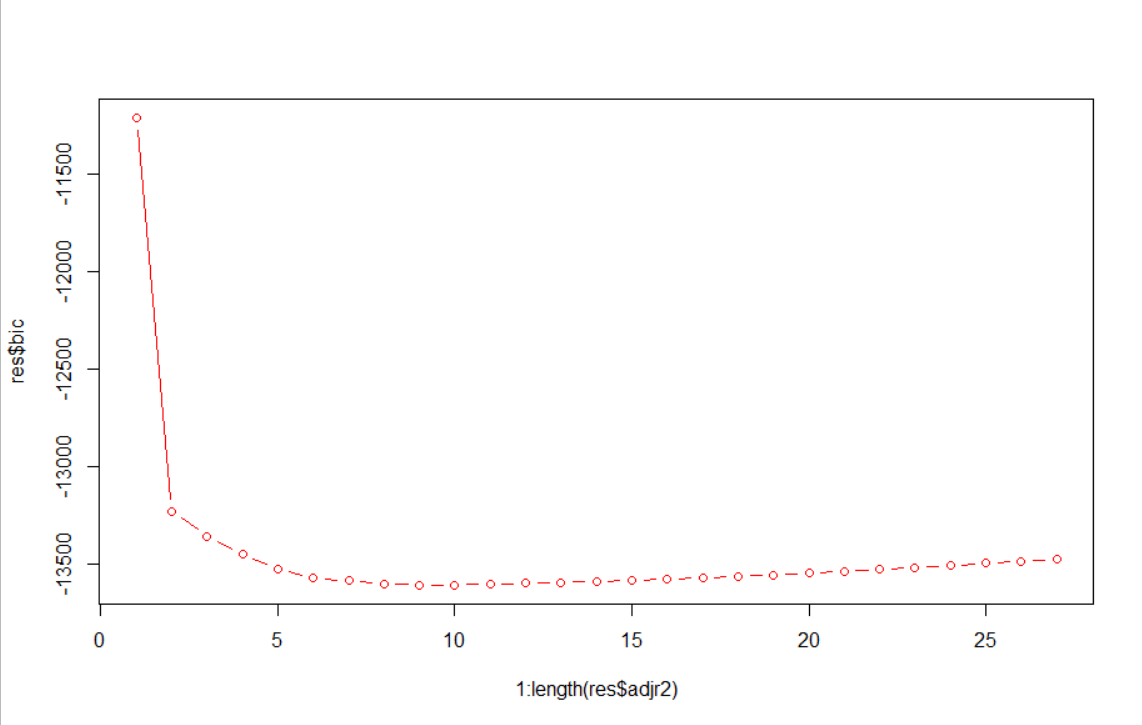
# Checking the model with reduced variables



# Output of exhaustive search:







We will take the model with 9 variables which is suggested to be one of the best number of predictors according to the above graph

Generating multiple models and then checking the p-values of the variables I would like to discuss the following variables in depth:

**$ MMRAcquisitionRetailAveragePrice:** This Variable should be added in our model as the p-value in 2 glm models for this is less than 0.05 giving us enough evidence that this will have a significant impact on the output target.

**$ PRIMEUNIT, $ AUCGUART**: These 2 Variables should not be added as there are too many null values even though the p-value is less than 0.05

**$ MMRAcquisitionAuctionAveragePrice**: This variable has p-value <0.05 but is highly related to MMRAcquisitionRetailAveragePrice thus we should avoid adding this as to avoid multicollinearity.

**$ Make, $ Color:** These two variables do not have significant P-values except for 1 level each. There are 33 and 17 levels for these two variables, which will increase the complexity and cost of the model. Thus, should not be included