Don’t Get Kicked: Predicting Whether Cars Bought at Auctions are of Poor Quality.

By Daniel Oskar Molent - s3661593 and
Adelia-Maria Guardado - s3604762

Source and Description

Car dealerships frequently purchase second-hand cars from the public to resell at a higher price to buyers as part of their business activities. Second-hand cars supplied by the public often come in varying qualities, models and ages. While most sellers are honest about the history and quality of the car they are selling, some sellers dishonestly inflate the sale price

of their vehicles. This manipulation can be achieved in various ways, including by illegally modifying their odometers to make their car appear younger than it is or by withholding critical information about inconspicuous mechanical issues. Such deception is of significant concern to car dealers as rectifying the issues is often time-consuming and expensive, and can cause dealerships to incur a loss on the transaction. In some instances, it may not be possible to resell the car at all. These purchases are colloquially known as “kicks”.

The “Don’t get kicked!” dataset (Carvana Co., 2011) contains detailed mechanical and market information about second-hand cars bought by dealerships at auctions. Most notably, the data set identifies whether cars purchased at the auctions were kicks or not. The data set contains 72,983 observations and 33 attributes. Of these 33 attributes, four relate to specific observations (e.g. ID variables) that are not useful for analysis purposes, 28 are predictor variables, and 1 is a binary target variable. The target variable for this project is ‘IsBadBuy’. The variable denotes whether a car was identified as a kick or not a kick. Values of 0 denote that a vehicle is not a kick, i.e. a typical vehicle, and values of 1 denote that a vehicle is a kick, i.e. it is of low quality and is not as marketed. It is known from publicly available information that the data set belongs to Carvana Co.; however, it is unclear where and how the data were collected.

Feature Name	**Data Type**	**Units**	**Description**
IsBadBuy	Binary	N/A	Vehicle is a kick or bad buy. 0 = Good buy. 1 = Bad buy.
PurchDate	Date	UNIX epoch timestamp	Date of purchase of the vehicle as a UNIX epoch timestamp.
Auction	Multinomial	N/A	Name of the auction the vehicle was sold at.
VehYear	Numeric	Year	Manufacture year of vehicle.
VehicleAge	Numeric	Years	Age of the vehicle in years.
Make	Multinomial	N/A	The manufacturer of the vehicle.
Model	Multinomial	N/A	Model of the vehicle.
Trim	Multinomial	N/A	The variant type for the style and features of the vehicle model.
SubModel	Multinomial	String	Submodel of the vehicle.
Color	Multinomial	N/A	Colour of vehicle.
Transmission	Binary	N/A	The transmission type of the vehicle, i.e. Automatic or Manual
WheelTypeID	Multinomial	N/A	Unique identifying value of the wheel type of the vehicle

WheelType	Multinomial	N/A	Type of wheels on the vehicle.
VehOdo	Numeric	Miles	The number of miles on the vehicle's odometer at the time of sale.
Nationality	Multinomial	N/A	The nationality of the vehicle.
Size	Multinomial	N/A	The size and structure of the vehicle.
TopThreeAmericanName	Multinomial	N/A	Vehicle was produced by a top three American manufacturer (Chrysler, Ford, GM).
MMRAcquisitionAuctionAveragePrice	Numeric	USD	Price for a vehicle of the same type in average condition at the time the vehicle was purchased.
MMRAcquisitionAuctionCleanPrice	Numeric	USD	Price for a vehicle of the same type in above average condition at the time the vehicle was purchased.
MMRAcquisitionRetailAveragePrice	Numeric	USD	Retail price for a vehicle of the same type in average condition at the time the vehicle was purchased.
MMRAcquisitonRetailCleanPrice	Numeric	USD	Retail price for a vehicle of the same type in above average condition at the time the vehicle was purchased.
MMRCurrentAuctionAveragePrice	Numeric	USD	Price for a vehicle of the same type in average condition at the current date.
MMRCurrentAuctionCleanPrice	Numeric	USD	Price for a vehicle of the same type in above average condition at the current date.
MMRCurrentRetailAveragePrice	Numeric	USD	Retail price for a vehicle of the same type in average condition at the current date.
MMRCurrentRetailCleanPrice	Numeric	USD	Retail price for a vehicle of the same type in above average condition at the current date.
PRIMEUNIT	Binary	N/A	Demand for vehicle is higher than usual.
AUCGUART	Multinomial	N/A	The guarantee amount provided on the vehicle.
BYRNO	Multinomial	N/A	Unique identifier of car buyer

VNZIP1	Multinomial	N/A	American zip code the vehicle was purchased in.
VNST	Multinomial	N/A	American state the vehicle was purchased in.
VehBCost	Numeric	USD	Price paid for vehicle.
IsOnlineSale	Binary	N/A	Vehicle was purchased online.
WarrantyCost	Numeric	USD	Cost of the vehicle's warranty.

Goals and Objectives

The financial cost of purchasing a car that cannot be resold due to mechanical issues is significant. As such, all car dealers need to minimise the number of kicks they purchase to reduce their financial losses. Car dealers must understand the factors and qualities of kicks so they can better identify vehicles with issues before committing to a purchase.

This project aims to provide a solution to the problem by automating and improving the accuracy of the process of identifying a kick. There are two main goals associated with this aim: accurately predicting whether an unknown vehicle is a kick, and better understanding the relationship between vehicle characteristics (both mechanical and market) and the likelihood of being a kick.

The first goal is to develop a logistic regression model that can accurately predict whether a new car at an auction is a kick from its mechanical and market characteristics. A logistic regression model provides the benefit of making predictions for many observations efficiently and automatically, which can save car dealers time and money on avoiding kicks. The model will ideally correctly predict whether a vehicle is a kick with an accuracy of at least 90% and will be trained using observations from the "Don't Get Kicked!" dataset (Carvana Co., 2011). The variable "isBadBuy" - which represents whether a vehicle is a kick - will be used as the target variable in the model and most of the remaining attributes relating to the mechanical and market properties, such as its resale value, will be used as predictor variables.

The second goal is to understand and quantify the relationships between vehicle attributes and kicks. By understanding these relationships, car dealers can better understand the factors associated with vehicles being kicks and make more educated decisions. The significance of the predictor variables on the log odds of a vehicle being a kick will be assessed from the coefficients of the logistic regression model. The coefficients not only indicate whether an attribute is significant but also quantify the increase in the log odds for a fixed change in a predictor. For example, the influence of a $1,000 increase in resale value on the log odds of a vehicle being a kick can be quantified. Furthermore, the influence of vehicle attributes on kicks will be explored using data visualisations. Some visualisations of interest include plotting histograms by the target for numeric variables and bar charts of categorical variables by the target variable. A significant difference between distributions of an attribute by target variable signifies that the attribute is useful in identifying kicks (and is

further likely to be a meaningful predictor in the logistic regression model). The relationships between predictor variables will also be examined from correlations and plots.

Data Cleaning & Preprocessing

RStudio was used to clean and preprocess the kick data. The packages "fastDummies", "openxlsx", "ggplot2", "plotly", "plyr", "plotrix", "ggpubr" and "dplyr" were firstly imported into R. The data set was then imported into R. Attributes which were not relevant to the analysis were then removed. These attributes were "PurchDate", "WheelTypeID", "BYRNO" and "Model". Of these four attributes, the first was a temporal attribute, the second and third were ID attributes, and the final contained too many values to encode as dummy variables. The dataset contained many question mark values representing missing values. These question mark entries were converted to "NA" so they could be handled appropriately. Similarly, the value "Not Avail" was converted to NA.

The total number of missing values and the number of missing values by each column was explored. Logistic regression models cannot handle missing values and as such must either be imputed or removed. There were 146,196 missing values in total. Approximately one-third of the attributes contained no missing values. Another third contained a negligible number of missing values (less than 20 observations missing). From the remaining third, half of the attributes had 315 missing values, which is small relative to the size of the data set. The variables "PRIMEUNIT" and "AUCGUART" were a significant issue as they contained 69,564 missing values each (95.3% of the total observations). These variables were dropped as they contained too many missing values. The attributes "Trim" and "WheelType" contained 2,360 and 3,174 missing values, respectively. These numbers were considered acceptable compared to the relatively large sample size of 72,983. All rows that contained any missing values were then dropped. In total 5,831 rows were dropped or 8.0% of the total observations, resulting in 67,152 observations.

The data types of each attribute were mostly correctly assigned by R; however, a few issues required rectification. Firstly, each MMR sale price variable was incorrectly interpreted as having a character data type and were converted to numeric values. The "VehBCost" was similarly interpreted as a character attribute and was converted to numeric. The binary features "IsOnlineSale", "Transmission" and "isBadBuy" were converted to factor types. Before this conversion, case-sensitivity issues in the transmission were fixed. The values of "AUTO" and "MANUAL" were changed to "1" and "0", respectively, to represent whether the vehicle has an automatic transmission or not.

The "SubModel" and "Trim" columns both contained valuable information but had too many unique values. Feature engineering was used to extract the most useful information out of these attributes so that the original columns could be dropped to prevent the data set and final model from having too many dummy variables. The SubModel attribute was searched against a list of terms. A new indicator variable was created for each term and assigned a value of 1 if the term was identified in the text value and assigned a value of 0 if it was not identified. As such, vehicles with no features (values for all feature columns are 0) were implicitly represented in the dummy variable encoding as being "Other". The lists of terms were identified by visually inspecting the unique column values for common phrases and

verifying whether the phrases occurred with sufficient frequency to construct an indicator variable. For this exercise, the necessary frequency was 1000. Extracting the car engine values of "V6" and "V8" from the model was also considered as an option. This process would have created misinformation as many V6 and V8 cars do not have the terms "V6" and "V8" respectively in their model name and would therefore not be identified, however. This misinformation was evident as only a few thousand vehicles had either of these engines in their model names despite these being commonly used engines.

The Trim attribute contained too many variables to encode as dummy variables in a logistic regression model. The value 1000 was defined as the minimum threshold frequency for values to be kept and encoded into a dummy variable. If fewer than 1,000 vehicles contained a particular trim, the trim was removed and replaced with "Other". This process reduced the number of trim types to 13. These 13 trims were dummy variables encoded with a baseline value of "Other".

The high unique observation count for the state of purchase (VNST) and zip code of purchase variables (VNZIP1) similarly needed to be reduced to limit the number of dummy variables in the logistic regression model. The range of possible values for these two columns was reduced down to 10 regions in America using the first digit of the VNZIP1 values. The first digit of American zip codes denotes one of 10 possible large zones that the zip code is from. For example, zip codes beginning with 9 belong to a zone containing states and cities near the Western coast of America, such as Seattle and San Francisco, and the state of Alaska. A dummy variable was created for each of these regions, excluding 0, which was the baseline.

Redundant attributes were removed. The attributes Trim and SubModel were removed as they were previously encoded, and both the state and zip code of purchase were removed as the geographical location of the sale is captured through the zip zone variable. The data types of the columns introduced were then corrected. The dummy variables - which were a logical data type - were converted to a binary factor. The character variables were also converted to factor data types. Finally, the categorical variables were dummy variable encoded. These variables were Auction, TopThreeAmericanName, Size, Nationality, WheelType, Make, Color and Size. The final data set had 132 columns.

Outlier identification:

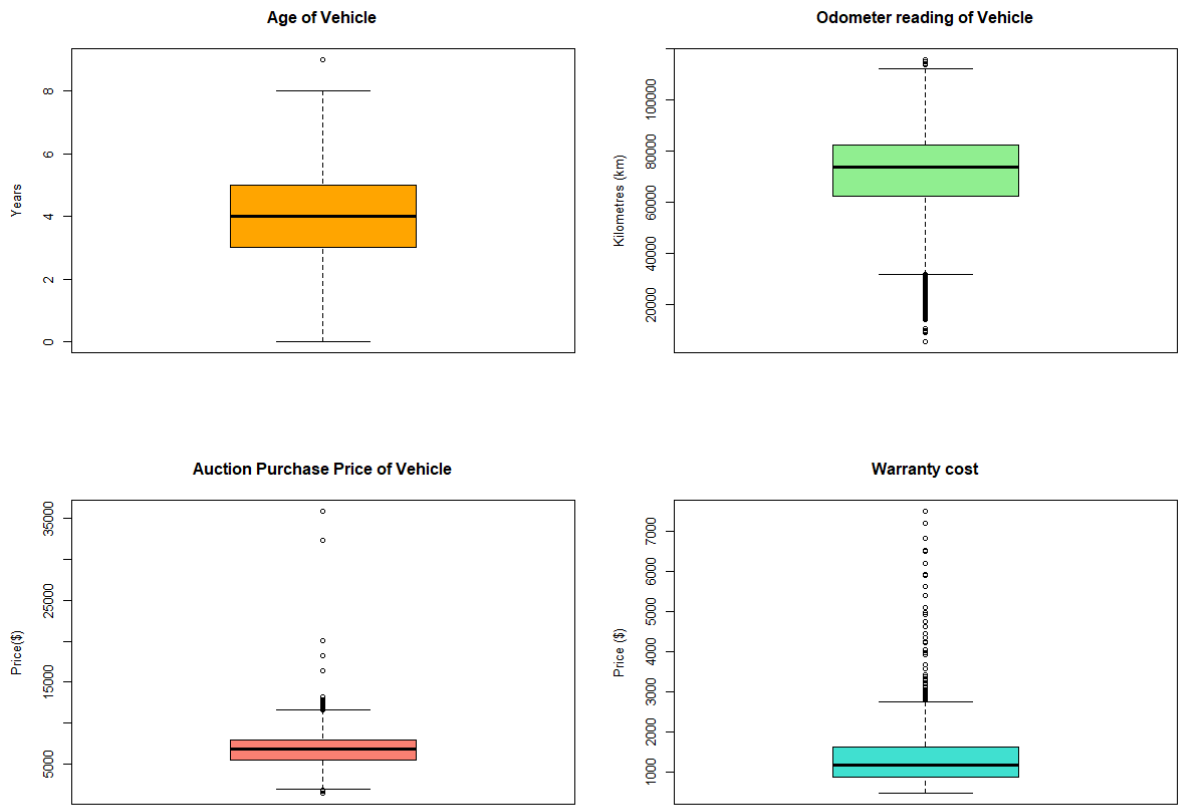

Figure 1 - Boxplot of outliers.

Of the 33 original variables, four numerical variables contained outliers. These variables contained less than 2000, mostly moderate outliers after dropping rows with missing values, which proportional to the total number of observations could be considered moderate. The kick dataset is unevenly proportioned as approximately 90% of the observations are not kicks, and 10% are kicks. If the outliers are disproportionately affecting kicks, however, they could have a significant influence due to the small number of values. This possibility was investigated in Figure 10 and was found not to be a significant issue. Both the vehicle odometer and warranty cost outliers are plausible values and as such were not removed as errors.

The attribute "VehicleAge" had approximately 565 observations that were above the average age of the vehicle, and none below. This could be further extracted as 0.7% of all non-kicks and approximately 2.29% of all kicks. The proportion of outliers that were kicks was slightly high, but was still within a feasible range. As such, these values were not removed from the data set. The number of outliers appears small overall but amounts to 2.29% of all kicks. The difference in proportions may demonstrate that old cars are more likely to be kicks.

The variable “VehOdo” had four observations that were above the upper outlier limit. Additionally, there were 278 observations in the lower outlier limit, where 0.4% were non-kicks, and approximately 0.36% were kicks. The Vehicle odometer reading is likely to be a significant variable, as vehicles with high mileage have become more damaged and are less likely to be bought than a vehicle that has a lower odometer reading. There was not a

significant number of outliers in the lower and upper limit, as the total amount of outliers only accounted for 0.4% of all observations.

The average vehicle purchase price (VehBCost) was $6754. There were only 3 outliers below the lower outlier limit. While only 166 observations were above the upper outlier limit. Overall the total amount of outliers accounted for 0.25% of all observations. This may not be a significant problem, and may not need to be removed. Due to the nature of auctions, these outliers may have been vehicles that were in higher demand than the average car, due to either being a collectors car or brand new by the manufacturer. Hence more demand equals higher bidding prices. For the 3 outliers below the lower outlier limit, these vehicles may have been total wrecks or very low in demand. It may be feasible to leave these outliers to improve the accuracy of logistic modelling.

The variable (WarrantyCost) had an average price of $1279 added to vehicles once purchased. Up to 696 observations were identified as outliers above the top outlier limit. Of these outliers, 1.85% were kicks, and 0.95% were non-kicks. While there were many large outliers, these values remained in a feasible range. They likely reflect the cost of covering the premium vehicles identified in the data set.

Overall most variables did not show signs of outliers caused by errors. However, some distributions were heavily skewed and require further examination through histograms. These distributions with outliers and abnormal distributions are concerning when using logistic modelling as they can interfere with parameter estimates. If issues with the parameter estimates are evident, or the performance of the model is bad, logarithmic transformations will be trialled to reduce the influence of these outliers.

Data Exploration & Visualisation

Summary statistics were firstly produced to explore the data. The target variable “IsBadBuy” contained 60,742 cars which were not kicks and 6,410 cars which were kicks. All of the numeric variables in the dataset had a mean and median relatively close together. This similarity suggests that although a number of extreme outliers were present, their influence is mitigated by the large quantity of data.

Variable	**Min**	**1st Quantile**	**Median**	**Mean**	**3rd Quantile**	**Max**
VehYear	2001	2004	2005	2005	2007	2010
VehicleAge	0	3	4	4.169	5	9
VehOdo	5368	62175	73533	71741	82539	115717
MMRAcquisitionAuctionAveragePrice	0	4311	6162	6162	7806	35722

MMRAcquisitionAuctionCleanPrice	0	5456	7380	7412	9049	36859
MMRAcquisitionRetailAveragePrice	0	6319	8498	8539	10710	39080
MMRAcquisitonRetailCleanPrice	0	7526	9870	9897	12155	40308
MMRCurrentAuctionAveragePrice	0	4311	6129	6167	7776	35772
MMRCurrentAuctionCleanPrice	0	5468	7389	7430	9045	36859
MMRCurrentRetailAveragePrice	0	6565	8811	8819	10973	39080
MMRCurrentRetailCleanPrice	0	7823	10178	10191	12373	40308
VehBCost	1400	5470	6750	6754	7911	35900
WarrantyCost	462	853	1169	1279	1623	7498

Table 1. Summary Statistics of Numeric Variables.

The MMR variables representing typical car prices of different qualities had many extreme outliers. The high and maximum values were manually investigated to determine whether they were mistakes. It was found by searching for the price of the car models corresponding to the high prices that the entries were valid. The models were rare and valuable cars that are well above the average price of the other second-hand vehicles. The high prices in the vehicle cost (VehBCost) were valid for similar reasons. Many cars also had values for the MMR variables of 0. This price is evidently incorrect as the average sale price for a vehicle model cannot realistically be $0USD. As such, these records were removed from the data set after the data visualisation stage. The maximum vehicle odometer value was considered realistic, so no outliers were removed from it.

Univariable plots:
Univariate plots were created to identify the asymmetric distributions, which may affect our logit modelling in phase 2.

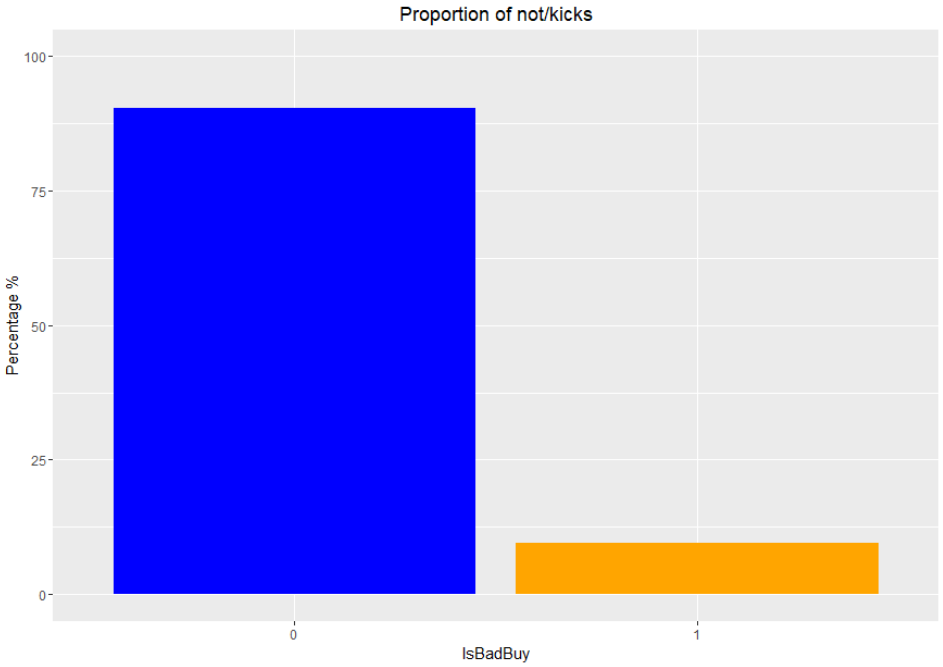

Figure 2. Kick data set target variable: IsBadBuy

In Figure 2, an uneven proportion of kicks to non-kicks of approximately 90% to 10%, respectively, is evident. Unfortunately, a relatively small proportion of the total observations were kicks; however, the number is still likely to be high enough to predict effectively using logistic regression. Suppose too few kicks are present to determine whether a vehicle is a kick accurately. In that case, judgement from subject matter experts from Manheim and Adesa in conjunction with a semi-accurate model may be sufficient to classify cars with an acceptable level of accuracy correctly.

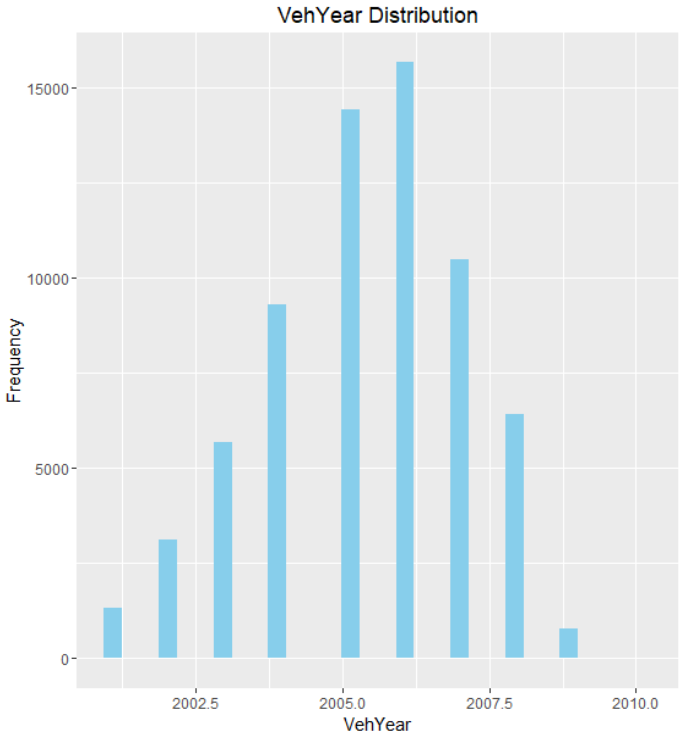

Figure 3. Kick dataset explanatory variable: VehYear

For the purpose of logistic regression, Figure 3 had a symmetrical distribution, and no outliers. The average manufactured year was 2005, and about 49.6% of all 67,152 kick dataset observations were above the mean.

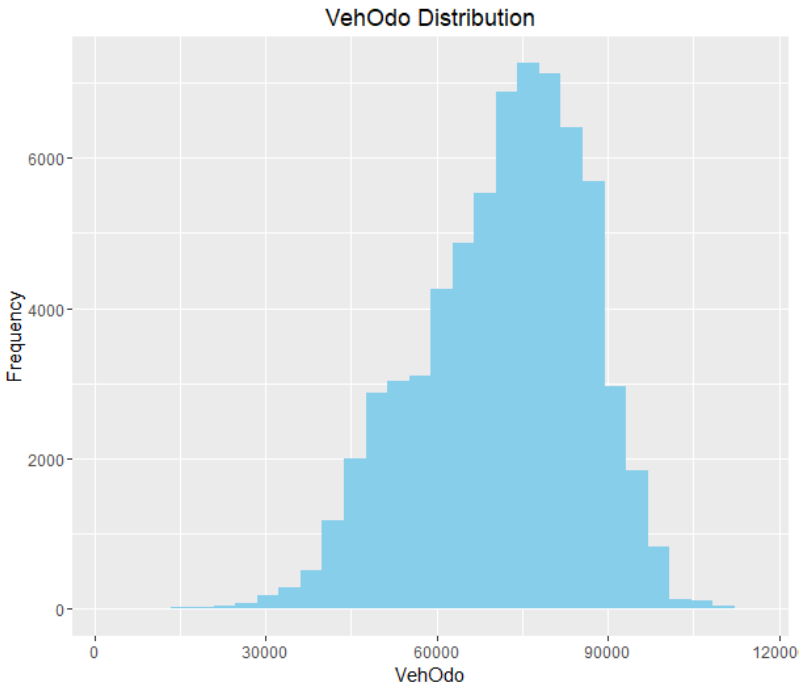

Figure 4. Kick dataset explanatory variable: VehOdo

The variable 'VehOdo' was another distribution that was almost symmetrical with approximately 55% of observations above the odometer mean of 71,730 miles. It was one of the few variables that contained outliers. Yet these outliers only accounted for 0.4% of all kick dataset observations, and may not cause a concern for logistic regression.

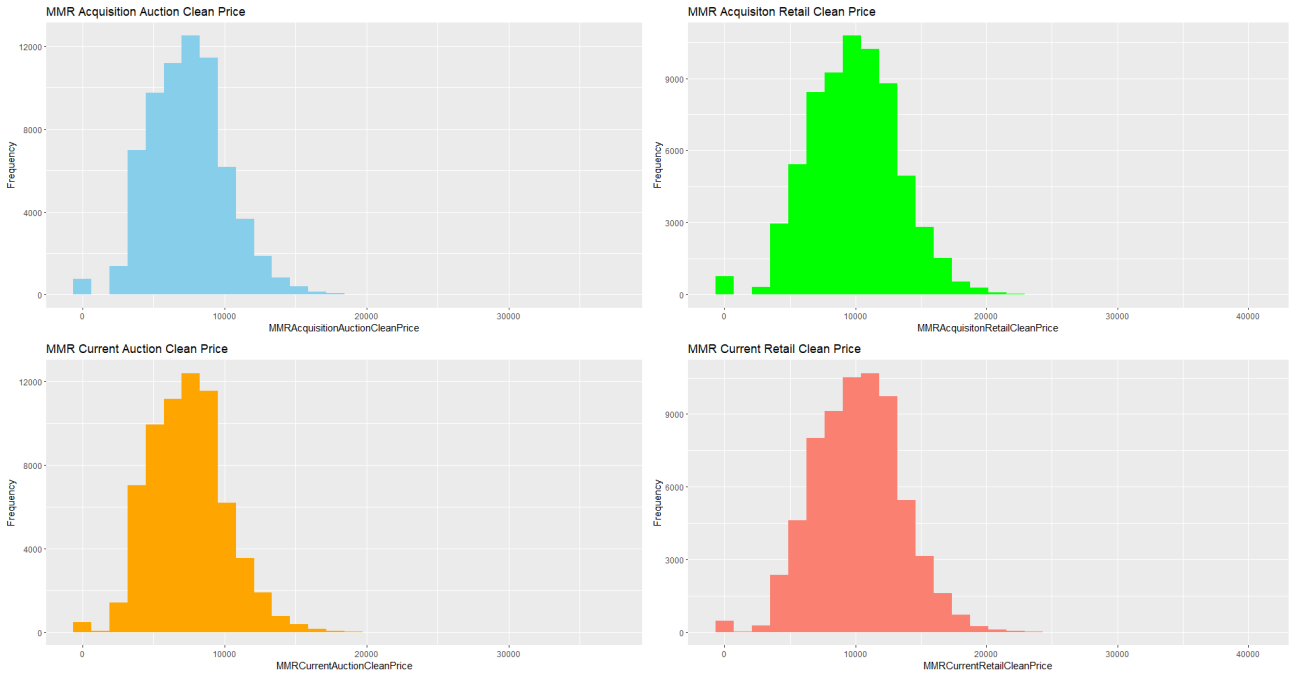

Figure 5. Kick dataset explanatory variables.

Most of the MMR variables had heavily skewed distributions. As such, they may be issues in the logistic regression modelling as previously discussed and addressed. Similarly, the high proportion of 0 values was previously identified as a concern, and observations containing this error were removed.

Bivariate plots

The bivariate graphs aim to identify the relationships between the binary target “IsBadBuy” and other independent predictors. The data was subset into kicks and non-kicks with the intent to compare the distributions to determine whether they differed by the target variable. A difference in distribution is indicative of a variable being a good predictor.

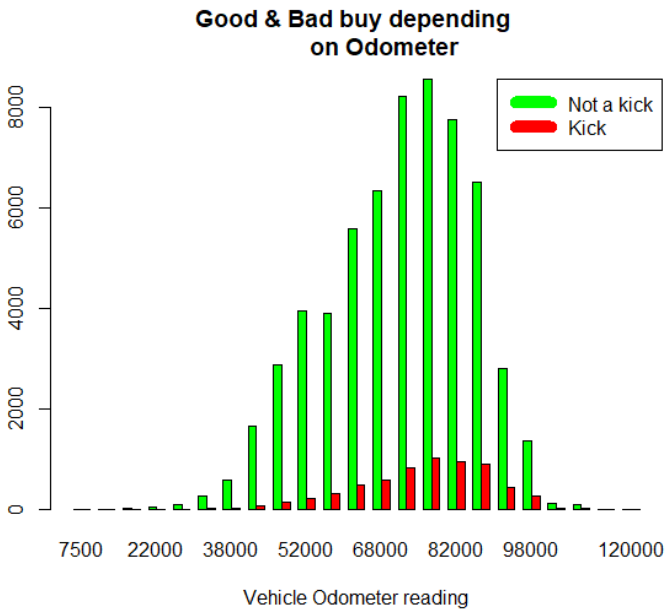

Figure 6. Kick dataset variables: IsBadBuy vs VehOdo.

Figure 6 describes the relationship between the kick and non-kick groups and the Vehicle odometer reading. It was assumed that a vehicle being a kick was highly correlated with a higher odometer reading as vehicles typically depreciate as they gain miles. However, in Figure 6, there does not appear to be a significant difference in distributions between kicks and not kicks. Car dealers may want to note that odometer reading may not be a concern when choosing cars apart from kicks when purchasing from auctions such as Manheim and Adesa. However, it should be noted that non-kicks had lower average miles than kicks with 71,320 miles compared to 75,640 miles for bad buys. Most non-kicks can be seen to be placed on the left tail, indicating again that lower mileage is preferred. In contrast, most kicks were towards the higher end of odometer reading.

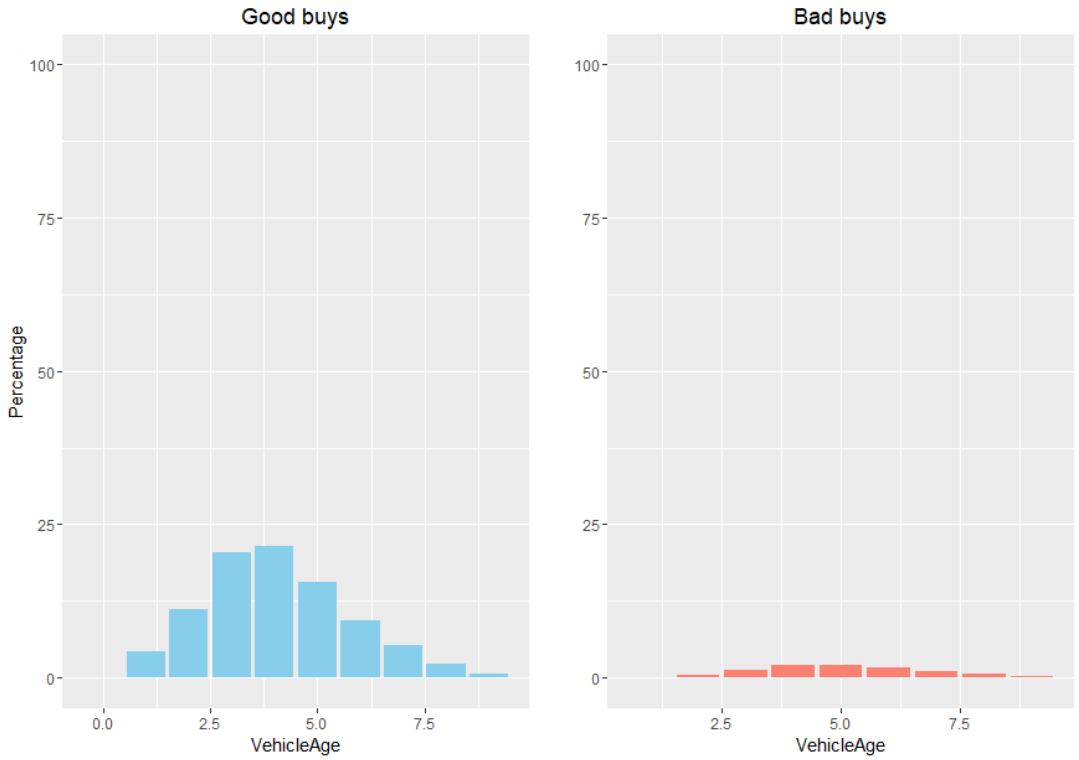

Figure 7. Kick dataset variables: VehicleAge group by IsBadBuy.

In Figure 7, vehicle age was grouped into kicks and non-kicks. The non-kicks had an average age of 4.075 years while kicks had a slightly higher average age of 5.059. There is not a massive distinction between non/kicks and vehicle age. However, most non-kicks tended towards the lower values of vehicle age, while most kicks tended towards a higher vehicle age, though the difference in means was not significant. Car buyers should still pay close attention to the age of a vehicle, specifically in small differences of age.

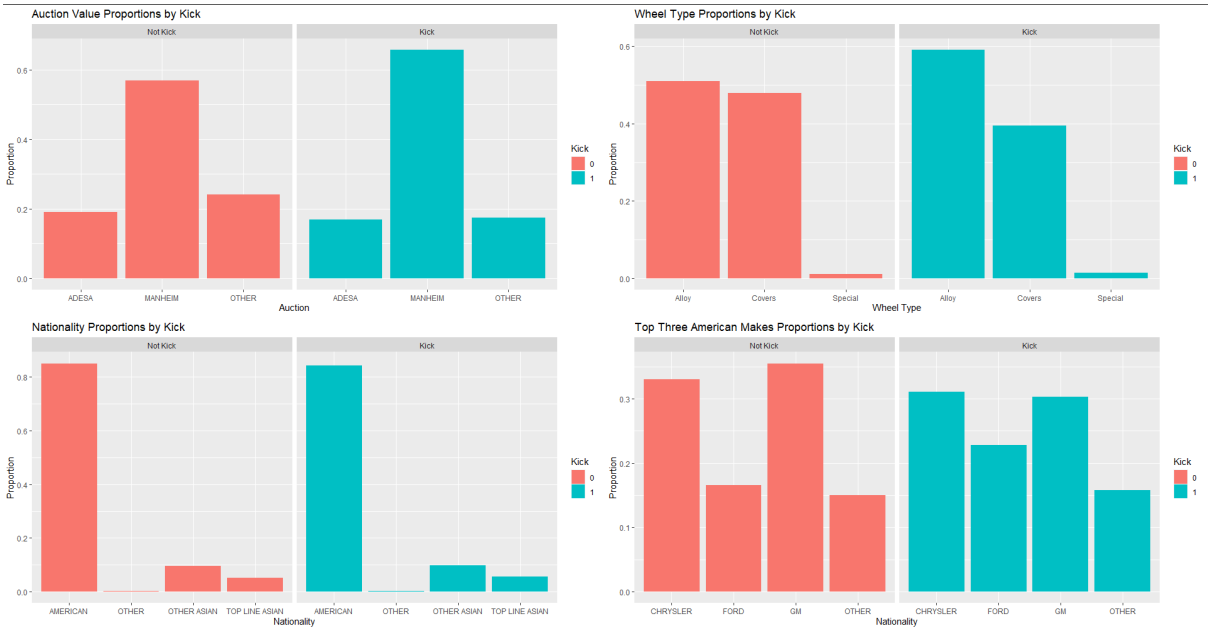

Figure 8. Proportional frequencies of categorical variables by kick.

Bar plots of the proportional frequencies of attributes were created and faceted by kick to examine the differences in their distributions. Marginal differences were present between kicks and non-kicks. Most notably, the proportion of cars with the “Alloy” wheel type was higher for kicks than non-kicks. Similarly, the proportion of cars with the “Covers” wheel type was lower for kicks than non-kicks. As such, a random car with an “Alloy” wheel type is marginally more likely to be a kick and a random car with a “Covers” wheel type is slightly more likely not to be a kick. The “Manheim” auction was also found to sell a higher proportion of kicks than the “Adesa” and other auctions. Car salespeople may therefore need to exercise more caution at the “Manheim” auction. Finally, there was a slight relationship between the top three American car manufacturers and kicks. GM vehicles were slightly less likely to be a kick, and “Ford” vehicles were slightly more likely to be a kick. No visually significant differences in proportion were observed for the vehicle nationality and the remaining categorical variables in the kick data set. While these differences - and the differences in many of the plots in this report - may be small, collectively they may be powerful enough to discriminate between kicks and non-kicks accurately.

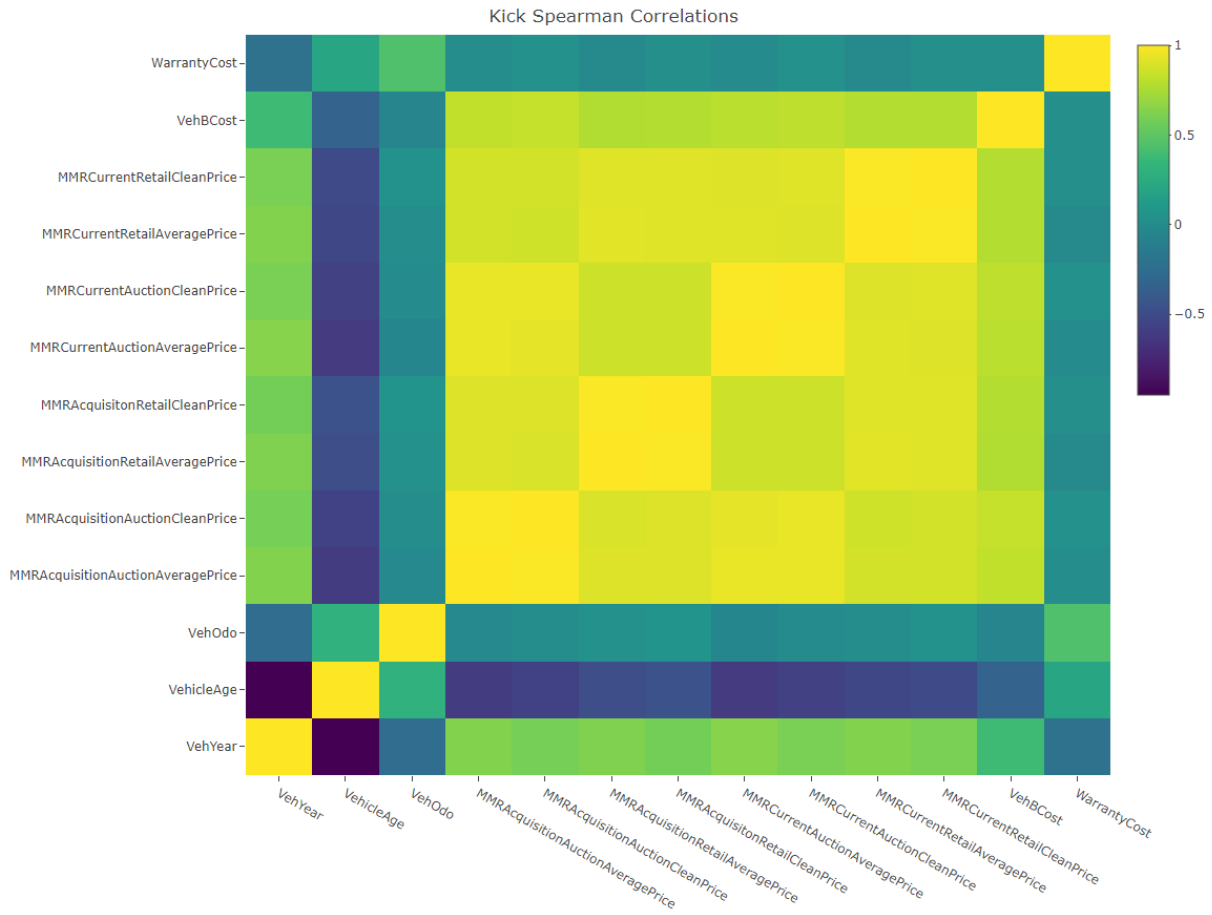

Figure 9. Spearman correlations of numeric variables.

Numerous high correlations between the MMR variables, and MMR variables and vehicle price were evident. As such, these variables must be modelled with interaction terms in the logistic regression model. Furthermore, the age of a vehicle and its price demonstrate a high negative correlation. This relationship must similarly be captured through an interaction term.

Three Variable Plots

Several three variable plots were created to further identify relationships between vehicle and market characteristics and kicks. The plots produced consisted of scatter plots between numeric variables coloured by kicks and boxplots of numeric variables split by two categorical variables. The exploration was conducted by selecting two and one numeric variables of interest for these plots, respectively, and iterating through different meaningful combinations of the categorical variables. The numeric variables of interest were the cost of the vehicle (VehBCost), age of the vehicle (VehicleAge) and the current average quality auction price (MMRCurrentAuctionAveragePrice). The remaining price variables did not produce significantly different plots due to the high amount of correlation with MMRCurrentAuctionAveragePrice. Overall, few distinct relationships were identified between the predictor variables and kicks. The distributions and relationships varied marginally between kicks and non-kicks; however, it is likely that many of these differences, albeit small, are likely to be significant due to the large sample size. Notwithstanding this, there were some intriguing relationships in the data.

Firstly, a scatter plot by the vehicle price, and the number of miles on the vehicle odometer was produced and coloured by kick. Lines of best fit calculated using least-squares linear regression were also fit for kicks and non-kicks. A random sample of 7.5% of the total observations was selected for plotting, and the transparency of the points reduced to 50% to avoid overplotting issues. This amount equates to x observations.

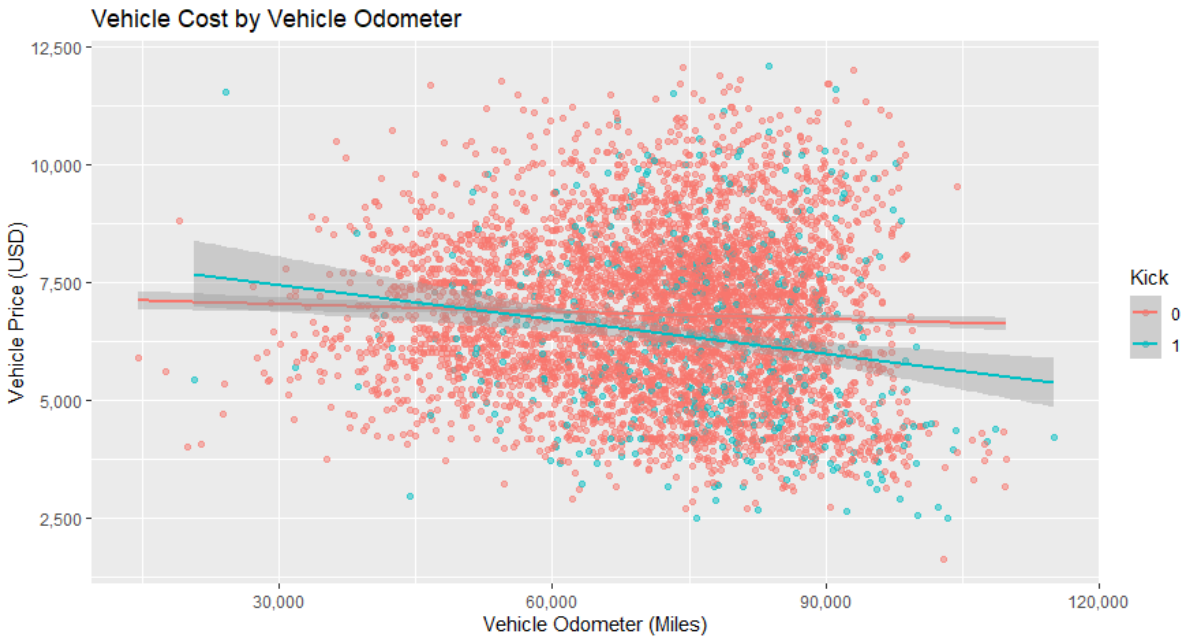

Figure 10. Scatter plot of Vehicle Odometer and Vehicle Price by Kick.

As the usage of a vehicle increased, the auction sale price slowly decreased for ordinary vehicles (i.e. not a kick). However, for kicks, the price declined more quickly for vehicles with more usage. The visualisation suggests that kicks are likely to have a higher vehicle price at the beginning of their lifespan, and a lower vehicle price at the end of their lifespan than ordinary vehicles. A plausible explanation for this relationship may be that the owners of kicks do not treat their vehicles with the same standard of care and respect as regular car owners. A lack of care would result in more damage and depreciation of a vehicle, which would reduce the value of the car. It is important to note, however, that the relationships have very weak correlations. The trend lines were also calculated using all observations to confirm that the observed trends did not result from the random sampling process.

A boxplot of vehicle costs was then created. This boxplot was split by vehicle size (Size) and kicks (IsBadBuy) to explore differences between the distributions for each combination of levels. The range of the vehicle cost was restricted to $15,000 USD to increase the size of the boxplots in the figure. A total of 5 observations were excluded from the results as a result of this limit. The car models of these outliers were searched and confirmed as valid observations.

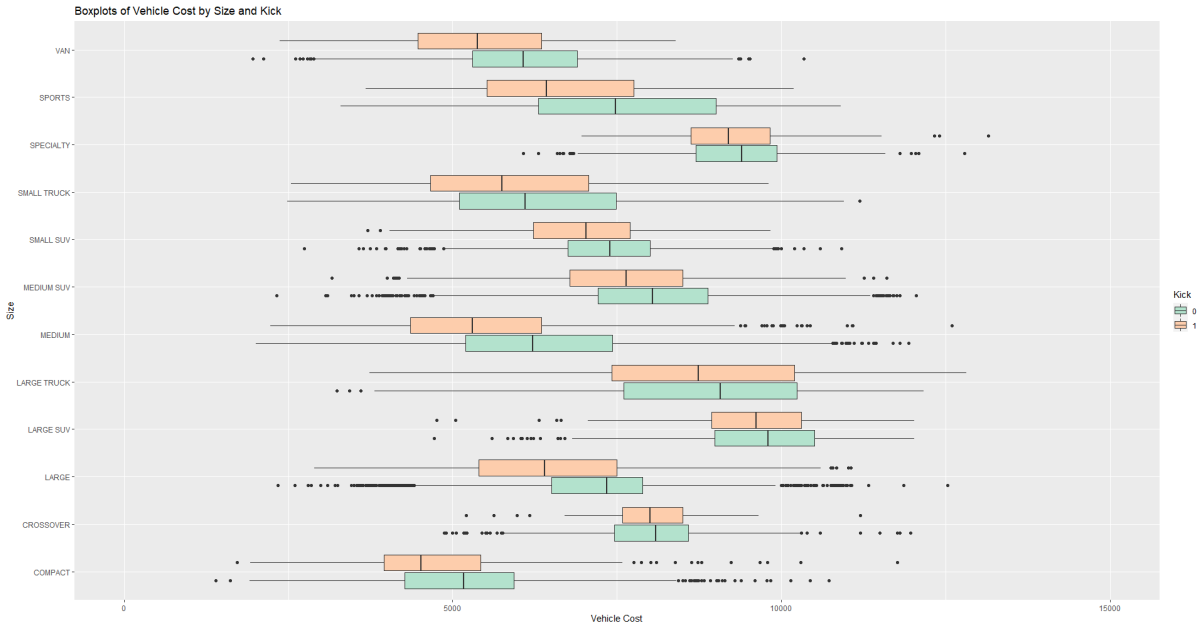

Figure 11. Boxplots of Vehicle Cost by Size and Kick.

The distributions for many of the truck sizes including “Specialty”, “Large Truck”, “Large SUV” and “Crossover” were approximately identical for kicks and non-kicks. There were several variables, however, that differed significantly between kicks and non-kicks. These attributes were most notably “Sports”, “Medium”, “Large” and “Compact”. The distributions of vehicle cost for these attributes were centred around a higher average vehicle cost for non-kicks and a lower vehicle cost for kicks. This difference was particularly pronounced for the “Large” size. As such, the results suggest that low-cost vehicles of sizes “Sports”, “Medium”, “Large” and “Compact” have a higher chance of being a kick. Conversely, higher prices in these sizes increase the probability of not-being a kick. Car dealers should, therefore be cautious and conduct further research before purchasing low-cost vehicles from these categories. This information will likely be useful for predicting the probability of a vehicle being a kick, and as such, these variables should contribute significantly in the final logistic regression model.

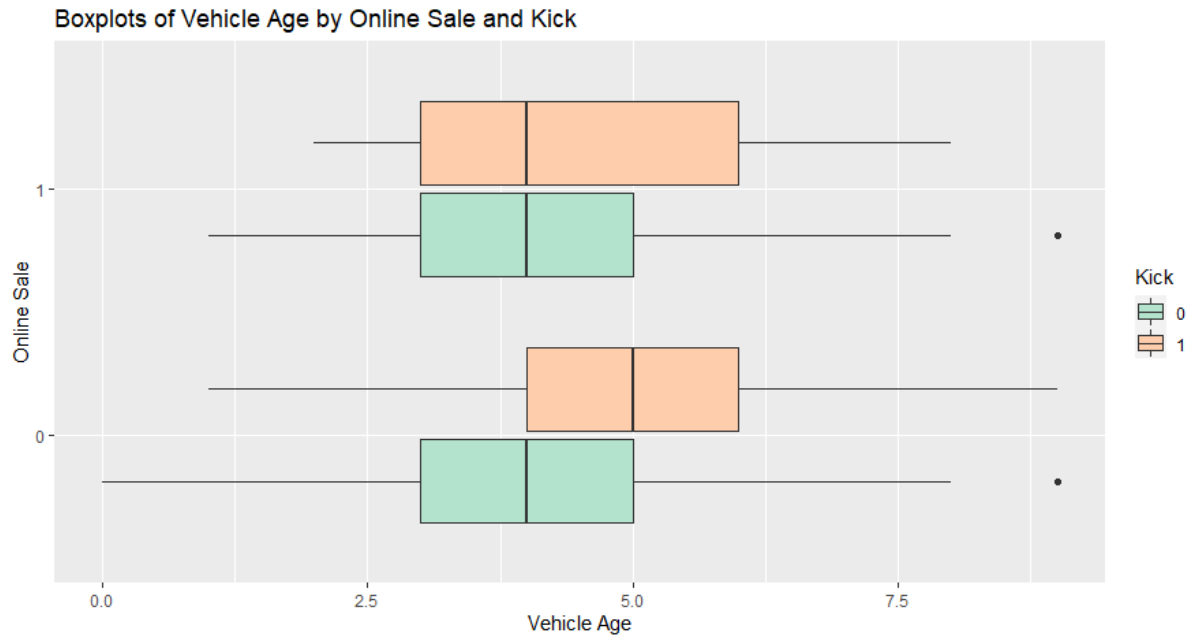

Figure 12. Boxplots of Vehicle Cost by Online Sale and Kick.

Similarly, a difference in distributions of vehicle age by online sale and kick was evident. While cars from an online sale had an approximately identical distribution for kicks and non-kicks, kicks from offline sales tended to have a higher vehicle age. As such, car dealers may want to take extra caution when purchasing older vehicles in an offline setting. It is important to note that the number of online sales was relatively low but was higher than 1000 and as such, should produce a stable and accurate boxplot.

References

Carvana Co. (2011). *Don't Get Kicked! Predict if a car purchased at auction is a lemon* [Data Set]. Kaggle. https://www.kaggle.com/c/DontGetKicked/overview

Appendix

Import libraries and the data set.

```r
#Import relevant libraries
library(fastDummies)
library(openxlsx)
library(ggplot2)
library(plotly)
require(plotrix)
library(dplyr)
library(gridExtra)

#Settings
#This setting skips lengthy visualisation loops
#The script contains loops which iterate through a list of variables and create plots for each one
skipVisualisations = FALSE

#Import data:
#1) Save R script and data to the same directory
setwd(dirname(rstudioapi::getActiveDocumentContext()$path)) #set working directory to R script location
filename = "kick.csv"
filePath = file.path(getwd(), filename)
kick = read.csv(filePath)
```

Data Processing(1/2)

```r
#Drop ID columns, temporal variables and variables with too many values to encode
kick = subset(kick, select = -c(PurchDate, WheelTypeID, BYRNO, Model))

#Replace question marks and "NOT AVAIL" in data set with null values
kick[kick == "?"] = NA
kick[kick == "NOT AVAIL"] = NA

#Check data types of all columns
str(kick)

#Count missing values
colSums(is.na(kick)) #Missing values by column
paste("Total NA:", sum(is.na(kick))) #Total missing values

#Drop PRIMEUNIT and AUCGUART variables as they have an unacceptably
#high number of missing values
kick = subset(kick, select = -c(PRIMEUNIT, AUCGUART))

#Remove all rows with missing values - for logistic regression
kick = kick[complete.cases(kick), ]

#Convert all variables beginning with MMR from char to int
cols = colnames(kick)

for (i in 1:ncol(kick)){
  #If column name begins with MMR
  if(grepl("^MMR", cols[i])){
    #Convert to numeric
    kick[,i] = as.numeric(kick[,i])
  }
}

#Convert VehBCost to int
kick[,"VehBCost"] = as.numeric(kick[,"VehBCost"])

#Convert IsBadBuy and IsOnlineSale to factor
kick[,"IsBadBuy"] = as.factor(kick[,"IsBadBuy"])
kick[,"IsOnlineSale"] = as.factor(kick[,"IsOnlineSale"])

#Change manual to MANUAL in transmission
kick[kick[["Transmission"]] == "Manual","Transmission"] = "MANUAL"
#Convert transmission to binary: 1 = AUTO, 0 = MANUAL
kick[kick[["Transmission"]] == "AUTO","Transmission"] = 1
kick[kick[["Transmission"]] == "MANUAL","Transmission"] = 0
```

Feature Engineering

```
#List of terms to extract from SubModel as features
features = c("CONVERTIBLE", "COUPE", "EXT CAB", "HATCHBACK","CUV", "SEDAN",
             "SPORT UTILITY","SUV","UTILITY","WAGON", "MINIVAN", "PASSENGER", "QUAD CAB",
             "REG CAB", "Limited")

#Iterate through features and construct indicator variables
for (f in features){
  columnName = gsub(" ", "", f, fixed = TRUE) #remove spaces from feature name for new column
  columnName = paste0("subModel", columnName) #Add SUBMODEL_ prefix to column name
  kick[, columnName] = grepl(f, kick[["SubModel"]]) #Create indicator variable
}

#Correct SUBMODEL_UTILITY column (Searching for "Utility" picks up "Sports Utility" as well)
for (i in 1:nrow(kick)){
  #if the submodel contained both "Sports Utility" and "Utility",
  #then the type should only be "Sports Utility"
  if(kick[i,"subModelSPORTUTILITY"] & kick[i,"subModelUTILITY"]){
    kick[i, "subModelUTILITY"] = FALSE #set utility to false
  }
}

#Set infrequent trims to other to reduce the number of values, i.e. < 1000 frequency.
trimCounts = table(kick["Trim"])
threshold = 1000

for (val in names(trimCounts)){
  #if frequency of trim does not meet threshold; replace with other
  if(trimCounts[val] < threshold){
    kick[kick[["Trim"]] == val, "Trim"] = "Other" #set values to other
  }
}

#Create dummy variables for all trims except "other" which is the baseline level
for (f in unique(kick[["Trim"]])){
  if(f != "Other"){
    columnName = gsub(" ", "", f, fixed = TRUE)
    kick[, columnName] = grepl(f, kick[["Trim"]])
  }
}

#Convert VNZIP1 to character.
kick[, "VNZIP1"] = as.character(kick[, "VNZIP1"])

#Create a dummy variable for the first digit of each zip code (the zip zone)
#Excludes 0 which serves as the baseline
for (i in 1:9){
  searchTerm = paste0("^[",as.character(i),"]") #First character of value is i
  columnName = paste0("ZIP", as.character(i))
  kick[, columnName] = grepl(searchTerm, kick[["VNZIP1"]])

}
```

Data Processing (2/2)

```
#Create a dummy variable for the first digit of each zip code (the zip zone)
#Excludes 0 which serves as the baseline
for (i in 1:9){
  searchTerm = paste0("^[",as.character(i),"]") #First character of value is i
  columnName = paste0("ZIP", as.character(i))
  kick[, columnName] = grepl(searchTerm, kick[["VNZIP1"]])

}

#Drop columns which features were extracted from
kick = subset(kick, select = -c(Trim, SubModel, VNST, VNZIP1))
#drop state

#Convert logical variables to factor
#convert character variables to factor
for (col in colnames(kick)){
  #if column is logical or character data type
  if(is.character(kick[[col]])){
    kick[[col]] = as.factor(kick[[col]])
  }
  else if(is.logical(kick[[col]])){
    kick[[col]] = as.factor(as.numeric(kick[[col]])) #convert to 0/1 binary factor
  }
}

#Remove variables with only one unique value
for (col in colnames(kick)){
  #if there is only one unique value
  if(length(unique(kick[[col]])) == 1){
    kick[[col]] = NULL
  }
}
```

Summary Statistics

```
#Summary Statistics
summary(kick)
```

One variable visualisations

```
#Plot distribution of all variables in data set
if(!skipVisualisations){
  #iterate through all columns of data set
  for (i in 1:ncol(kick)){
    columnName = colnames(kick)[i] #name of column being visualised
    column = kick[,i] #column values

   #Numeric columns: Histogram
    if(is.numeric(column)){
      #Plot histogram of variable
      hist = ggplot(kick, aes_string(x = columnName)) +
        geom_histogram(fill="skyblue", na.rm = TRUE) +
        xlab(columnName) + ylab("Frequency") +
        ggtitle(paste(columnName, "Distribution"))
      print(hist)
    }
    #Nominal columns: Bar chart of value counts
    else if(is.factor(column)){
      #Plot bar chart if less than 25 values in column
      if(length(unique(column)) < 25){
        bar = ggplot(kick, aes_string(x = columnName)) +
          geom_bar(stat = 'count') +
          xlab(columnName) + ylab("Frequency") +
          ggtitle(paste(columnName, "Value Counts"))
        print(bar)
      }
      else{
        print("Did not plot because of high number of unique values (>= 25)")
      }
    }

    #Wait for user to continue
    readline("Please select any key to continue (console must be active)")
  }
}

#Plot distribution of MMRAcquisitionAuctionCleanPrice
MMRa = (ggplot(kick, aes_string(x = kick$MMRAcquisitionAuctionCleanPrice)) +
     geom_histogram(fill="skyblue", na.rm = TRUE) +
     xlab("MMRAcquisitionAuctionCleanPrice") + ylab("Frequency") +
     ggtitle("MMR Acquisition Auction Clean Price"))

#Plot distribution of MMRAcquisitionRetailCleanPrice
MMRb = (ggplot(kick, aes_string(x = kick$MMRAcquisitonRetailCleanPrice)) +
      geom_histogram(fill="green", na.rm = TRUE) +
      xlab("MMRAcquisitonRetailCleanPrice") + ylab("Frequency") +
      ggtitle("MMR Acquisiton Retail Clean Price"))

#Plot distribution of MMRCurrentAuctionCleanPrice
MMRc = ggplot(kick, aes_string(x = kick$MMRCurrentAuctionCleanPrice)) +
  geom_histogram(fill="orange", na.rm = TRUE) +
  xlab("MMRCurrentAuctionCleanPrice") + ylab("Frequency") +
  ggtitle("MMR Current Auction Clean Price")

#Plot distribution of MMRCurrentRetailCleanPrice
MMRd = ggplot(kick, aes_string(x = kick$MMRCurrentRetailCleanPrice)) +
  geom_histogram(fill="salmon", na.rm = TRUE) +
  xlab("MMRCurrentRetailCleanPrice") + ylab("Frequency") +
  ggtitle("MMR Current Retail Clean Price")

#Combine MMR plots
ggpubr::ggarrange(MMRa, MMRb, MMRc, MMRd)
```

```
#Filter data set down to numeric columns
kickNumeric = Filter(is.numeric, kick)

#Calculate spearman correlations between numeric variables
corr = round(cor(kickNumeric, method = "spearman", use = "complete.obs"), 2)
names = colnames(kickNumeric)

#create heatmap
heatmap = plot_ly(x = names, y = names, z = corr, type = "heatmap") %>%
  layout(title = "Kick Spearman Correlations")

heatmap
```

MMR Outlier Removal and Dummy variables:

```
#Remove MMR values of 0
for(col in colnames(kick)){
  #If column begins with MMR
  if(grepl("^MMR", col)){
    kick = kick[kick[[col]] != 0, ]
  }
}

#Create new kick dataset with dummy vars
kickWDumVars =  dummy_cols(kick,select_columns = c("Auction",
                    "TopThreeAmericanName","Size", "Nationality",
                    "WheelType","Make", "Color", "Size"))
```

Heatmap:

```
corr = round(cor(kick, method = "spearman", use = "complete.obs"), 2)
names = colnames(kick)
heatmap = plot_ly(x = names, y = names, z = corr, type = "heatmap") %>%
  layout(title = "Kick Spearman Correlations")
heatmap
```

Box Plots:

```
#Explore outliers
a=boxplot(kick$VehicleAge, main="Age of Vehicle", ylab ="Years",col='orange')
upperwhisker = quantile(kick$VehicleAge,.75)+(1.5*IQR(kick$VehicleAge))
lowerwhisker= quantile(kick$VehicleAge,.25)-(1.5*IQR(kick$VehicleAge))
table(kick$VehicleAge > upperwhisker)#565 rows


#VehOdo
boxplot(kick$VehOdo, main="Odometer reading of Vehicle", ylab="Kilometres (km)",col='light green')
summary(kick$VehOdo)
upperwhisker = quantile(kick$VehOdo,.75)+(1.5*IQR(kick$VehOdo))
lowerwhisker= quantile(kick$VehOdo,.25)-(1.5*IQR(kick$VehOdo))
table(kick$VehOdo > upperwhisker) #4 rows above upper whisker
table(kick$VehOdo <lowerwhisker) #278 rows below lower whisker

#vehBcost
boxplot(kick$VehBCost, main="Auction Purchase Price of Vehicle", ylab="Price($)",col='salmon')
summary(kick$VehBCost)
upperwhisker = quantile(kick$VehBCost,.75)+(1.5*IQR(kick$VehBCost))
lowerwhisker= quantile(kick$VehBCost,.25)-(1.5*IQR(kick$VehBCost))
table(kick$VehBCost<lowerwhisker) #1772 rows below lower whisker


#WarrantyCost
boxplot(kick$WarrantyCost, main="Warranty cost", ylab="Price ($)",col='turquoise')
summary(kick$WarrantyCost)
upperwhisker = quantile(kick$WarrantyCost,.75)+(1.5*IQR(kick$WarrantyCost))
lowerwhisker= quantile(kick$WarrantyCost,.25)-(1.5*IQR(kick$WarrantyCost))
table(kick$WarrantyCost > upperwhisker) #696 rows above upper whisker
```

Subsetting Outliers:

```
#Vehicle age
#Subset outliers into good/bad buys above upperwhisker
#418 rows
kicYeargb <- subset(kick,kick$IsBadBuy==0 & kick$VehicleAge > upperwhisker)
#147 rows
kicYearbb <- subset(kick,kick$IsBadBuy==1 & kick$VehicleAge > upperwhisker)


#Vehicle odometer
#Subset outliers into good/bad buys above upper whisker
#1 good buys
kicYeargb <- subset(kick,kick$IsBadBuy==0 & kick$VehOdo > upperwhisker)
#3 bad buys
kicYearbb <- subset(kick,kick$IsBadBuy==1 & kick$VehOdo > upperwhisker)
#Subset outliers into good/bad buys below lowerwhisker
#255 rows
kicYeargb <- subset(kick,kick$IsBadBuy==0 & kick$VehOdo<lowerwhisker)
#23 rows
kicYearbb <- subset(kick,kick$IsBadBuy==1 & kick$VehOdo <lowerwhisker)


#VehBCost
#Subset outliers into good/bad buys below lowerwhisker
#1631 rows
kicYeargb <- subset(kick,kick$IsBadBuy==0 & kick$VehBCost<lowerwhisker)
#141 rows
kicYearbb <- subset(kick,kick$IsBadBuy==1 & kick$VehBCost<lowerwhisker)


#Warranty Cost
#Subset outliers into good/bad buys
#577 rows
kicYeargb <- subset(kick,kick$IsBadBuy==0 & kick$WarrantyCost > upperwhisker)
#119 rows
kicYearbb <- subset(kick,kick$IsBadBuy==1 & kick$WarrantyCost > upperwhisker)
```

Scatterplots of correlated terms:

```
#2 Var plots(highly correlated terms)
#0.55
(ggplot(kick)+aes(y=MMRAcquisitonRetailCleanPrice,x=VehYear)
  +geom_point(color="Purple",size=2)
  +xlab("Year Vehicle Made")+ ylab("MMRAcquisitonRetailCleanPrice")
  +scale_x_discrete(limits=kick$VehYear,expand = c(0.1, 0))
  +theme(axis.text.x = element_text(angle = 45, hjust = 1)))

#0.52
(ggplot(kick)+aes(y=MMRCurrentRetailCleanPrice,x=VehYear)
  +geom_point(color="red",size=2)+xlab("Year Vehicle Made")
  +ylab("MMR Current Retail Clean Price")
  +scale_x_discrete(limits=kick$VehYear,expand = c(0.1, 0))
  +theme(axis.text.x = element_text(angle = 45, hjust = 1)))

#0.61
(ggplot(kick)+aes(y=kick$MMRCurrentRetailAveragePrice,x=VehOdo)
  +geom_point(color="blue",alpha = 2/10)
  +xlab("Vehicle's Odometer reading")+ ylab("MMR Current Retail Average Price")
  +theme(axis.text.x = element_text(angle = 45, hjust = 1)))

#0.59
(ggplot(kick)+aes(y=kick$MMRCurrentAuctionCleanPrice,x=kick$VehicleAge)
  +geom_point(color='dark green')
  +xlab("Vehicle Age")+ ylab("MMR Current Auction Clean Price")
  +scale_x_discrete(limits=kick$VehicleAge,expand = c(0.1, 0))
  +theme(axis.text.x = element_text(hjust = 1)))

#0.52
(ggplot(kick)+aes(y=kick$MMRCurrentRetailCleanPrice,x=kick$VehBCost)
  +geom_point(color="blue",alpha = 2/10)+xlab("Price paid for vehicle")
  +ylab("MMR Current Retail Clean Price"))

#0.72
(ggplot(kick)+aes(y=kick$MMRCurrentAuctionAveragePrice,x=kick$VehBCost)
  +geom_point(color="darkturquoise",alpha = 2/10)+xlab("Price paid for vehicle")
  +ylab("MMR Current Auction Average Price"))
```

```
#0.81
(ggplot(kick)+aes(y=kick$MMRAcquisitonRetailCleanPrice,x=kick$MMRCurrentRetailCleanPrice)
  +geom_point(color="darkturquoise",alpha = 2/10)+xlab("MMR Current Retail Clean Price")
  +ylab("MMR Acquisiton Retail Clean Price"))

#0.52
(ggplot(kick)+aes(y=kick$MMRCurrentAuctionAveragePrice,x=kick$MMRAcquisitonRetailCleanPrice)
  +geom_point(color="salmon",alpha = 2/10)+xlab("MMR Acquisiton Retail Clean Price")
  +ylab("MMR Current Auction Average Price"))

#-0.95
(ggplot(kick)+aes(y=kick$MMRAcquisitionAuctionAveragePrice,x =kick$MMRAcquisitionAuctionCleanPrice)
  +geom_point(color="purple",alpha = 2/10)+xlab("MMR Acquisition Auction Clean Price")
  +ylab("MMR Acquisition Auction Average Price") +ggtitle("Distribution"))
```

2 variable plots:

```
#NUMERICAL COLUMNS
#Age
gbAge=kick$VehicleAge[kick$IsBadBuy==0]
bbAge=kick$VehicleAge[kick$IsBadBuy==1]
hist= multhist(list(gbAge,bbAge),col=c("Green","Red"),main="Good & Bad buy depending on Age",
               xlab="Vehicle Age",y="Frequency",breaks=seq(0,9,1),xap=c(0,9,1))
hist + legend("topright", c("Not a kick","Kick" ), col=c("light Green","Red"), lwd=10)


#VehOdo vs Bad Buy
gbOdo= kick$VehOdo[kick$IsBadBuy==0]
bbOdo=kick$VehOdo[kick$IsBadBuy==1]
summary(kick$VehOdo)
hist =multhist(list(gbOdo,bbOdo),col=c("Green","Red"),main="Good & Bad buy depending
        on Odometer",xlab="Vehicle Odometer reading",y="Frequency")
hist + legend("topright", c("Not a kick","Kick"), col=c("Green","Red"), lwd=10)


#Vehicle Cost vs Bad Buy
gbVeh= kick$VehBCost[kick$IsBadBuy==0]
bbVeh=kick$VehBCost[kick$IsBadBuy==1]
hist=multhist(list(gbVeh,bbVeh),col=c("Green","Red"),main="Good & Bad buy depending
        on Vehicle Cost",xlab="Vehicle Price",y="Frequency")
hist + legend("topright", c("Not a kick","Kick"), col=c("Green","Red"), lwd=10)


#Warranty vs Bad Buy
gbWar= kick$WarrantyCost[kick$IsBadBuy==0]
bbWar=kick$WarrantyCost[kick$IsBadBuy==1]
multhist(list(gbWar,bbWar),col=c("Green","Red"),main="Good & Bad buy depending
        on Warranty",xlab="Warranty Price",y="Frequency")
hist + legend("topright", c("Not a kick","Kick"), col=c("Green","Red"), lwd=10)
```

```r
#NOMINAL COLUMNS
#Subset kick data into good and bad buys

kickgb <- subset(kick,kick$IsBadBuy==0)#60796 90% good buys
kickbb <- subset(kick,kick$IsBadBuy==1)#6416 10% bad buys

#IsbadBuy vs Top 3 American Names
bb = kickbb %>% group_by(TopThreeAmericanName) %>% summarize(DF="Kick", n=n())
gb = kickgb %>% group_by(TopThreeAmericanName) %>% summarize(DF="Not Kick", n=n())
DF <- rbind(bb, gb)
(ggplot(DF, aes(x=TopThreeAmericanName, y=n, fill=DF)) + geom_bar(stat="identity", position="dodge")
  +ylab("Frequency")+xlab("Top Three American Manufacturers")
  +ggtitle('Is Bad Buy depending on Manufacturer')+labs(fill = ""))


#IsbadBuy vs Transmissiom
bb = kickbb %>% group_by(Transmission) %>% summarize(DF="Kick", n=n())
gb = kickgb %>% group_by(Transmission) %>% summarize(DF="Not a kick", n=n())
DF <- rbind(bb, gb)
(ggplot(DF, aes(x=Transmission, y=n, fill=DF)) + geom_bar(stat="identity", position="dodge")
  +ylab("Frequency")+xlab("Transmission Type")
  +ggtitle('Is a kick depending on Transmission')+labs(fill = ""))


#Is bad buy vs Color
bb = kickbb %>% group_by(Color) %>% summarize(DF="Kick", n=n())
gb = kickgb %>% group_by(Color) %>% summarize(DF="Not a kick", n=n())
DF <- rbind(bb, gb)
(ggplot(DF, aes(x=Color, y=n, fill=DF)) + geom_bar(stat="identity", position="dodge")
  +ylab("Frequency")+ggtitle('Is a  Kick depending on Color')
  +theme(axis.text.x = element_text(angle = 45, hjust = 1))+labs(fill = ""))


#AUCTION vs IsBadBuy

bb = kickbb %>% group_by(Auction) %>% summarize(DF="Kick", n=n())
gb = kickgb %>% group_by(Auction) %>% summarize(DF="Not a kick", n=n())
DF <- rbind(bb, gb)
(ggplot(DF, aes(x=Auction, y=n, fill=DF)) + geom_bar(stat="identity", position="dodge")
  +ylab("Frequency")+ggtitle('Is a kick depending on Auction Type')
  +theme(axis.text.x = element_text(angle = 45, hjust = 1))+labs(fill = ""))

#WheelType vs IsBadBuy

bb = kickbb %>% group_by(WheelType) %>% summarize(DF="Kick", n=n())
gb = kickgb %>% group_by(WheelType) %>% summarize(DF="Not a kick", n=n())
DF <- rbind(bb, gb)
(ggplot(DF, aes(x=WheelType, y=n, fill=DF)) + geom_bar(stat="identity", position="dodge")
  +ylab("Frequency")+ggtitle('Is a kick depending on Wheel Type')
  +theme(axis.text.x = element_text(angle = 45, hjust = 1))+labs(fill = ""))


#Nationality IsBadBuy

bb = kickbb %>% group_by(Nationality) %>% summarize(DF="Kick", n=n())
gb = kickgb %>% group_by(Nationality) %>% summarize(DF="Not a kick", n=n())
DF <- rbind(bb, gb)
(ggplot(DF, aes(x=Nationality, y=n, fill=DF)) + geom_bar(stat="identity", position="dodge")
  +ylab("Frequency")+ggtitle('Is a kick depending on Wheel Type')
  +theme(axis.text.x = element_text(angle = 45, hjust = 1))+labs(fill = ""))


#Size
bb = kickbb %>% group_by(Size) %>% summarize(DF="Kick", n=n())
gb = kickgb %>% group_by(Size) %>% summarize(DF="Not a kick", n=n())
DF <- rbind(bb, gb)
(ggplot(DF, aes(x=Size, y=n, fill=DF)) + geom_bar(stat="identity", position="dodge")
  +ylab("Frequency")+ggtitle('Is a kick depending on Wheel Type')
  +theme(axis.text.x = element_text(angle = 45, hjust = 1))+labs(fill = ""))
```

2 variable percentage plots:

```
#PERCENTAGE GRAPHS
#Change ColumnName to any as you please for percentage comparison graphs
#goodbuy
kickgb %>%
  count(columnName) %>%
  mutate(perc = n / nrow(kick)) -> kickgb2

a=ggplot(kickgb2, aes(x = ColumnName, y = perc*100)) + geom_bar(stat = "identity",fill="skyblue")
  +ggtitle("Good buys")+ ylab("Percentage")+ylim(0,100)

#bad buys
kickbb %>%
  count(columnName) %>%
  mutate(perc = n / nrow(kick)) -> kickbb2

b=ggplot(kickbb2, aes(x = columnName, y = perc*100)) + geom_bar(stat = "identity",fill="salmon")
  +ggtitle("Bad buys")+ylab("") +ylim(0,100)
grid.arrange(a,b,nrow = 1)
```

```
#Plot Auction Value Proportions bar chart by Kick
barAuction = ggplot(data = kick) +
        stat_count(aes(x = Auction, y=..prop.., group = IsBadBuy, fill = IsBadBuy)) +
        facet_grid(~IsBadBuy, labeller = as_labeller(c(`0` = "Not Kick", `1` = "Kick"))) +
        xlab("Auction") + ylab("Proportion") + ggtitle("Auction Value Proportions by Kick") +
        labs(fill = "Kick")

#Plot Wheel Type Proportions bar chart by Kick
barWheel = ggplot(data = kick) +
  stat_count(aes(x = WheelType, y=..prop.., group = IsBadBuy, fill = IsBadBuy)) +
  facet_grid(~IsBadBuy, labeller = as_labeller(c(`0` = "Not Kick", `1` = "Kick"))) +
  xlab("Wheel Type") + ylab("Proportion") + ggtitle("Wheel Type Proportions by Kick")+
  labs(fill = "Kick")

#Plot Nationality Proportions bar chart by Kick
barNationality = ggplot(data = kick) +
  stat_count(aes(x = Nationality, y=..prop.., group = IsBadBuy, fill = IsBadBuy)) +
  facet_grid(~IsBadBuy, labeller = as_labeller(c(`0` = "Not Kick", `1` = "Kick"))) +
  xlab("Nationality") + ylab("Proportion") + ggtitle("Nationality Proportions by Kick") +
  labs(fill = "Kick")

#Plot Nationality Proportions bar chart by Kick
barAmerican = ggplot(data = kick) +
  stat_count(aes(x = TopThreeAmericanName, y=..prop.., group = IsBadBuy, fill = IsBadBuy)) +
  facet_grid(~IsBadBuy, labeller = as_labeller(c(`0` = "Not Kick", `1` = "Kick"))) +
  xlab("Nationality") + ylab("Proportion") + ggtitle("Top Three American Makes Proportions by Kick") +
  labs(fill = "Kick")

#Combine bar charts into a single figure
ggpubr::ggarrange(barAuction, barWheel, barNationality, barAmerican)
```

3 variable plots

```r
set.seed(1) #set seed to make results reproducible

#randomly sample data set to reduce overplotting
proportion = 0.075
kickSample = sample_frac(kick, proportion)

#Scatter Plot of Vehicle Cost and Vehicle Odometer by Kick
MMRbyOdo = ggplot(data = kickSample, aes(x = VehOdo, y = VehBCost, colour = IsBadBuy)) +
  geom_point(alpha = 0.5) + geom_smooth(formula = y ~ x, method = "lm") +
  scale_x_continuous(labels = scales::comma) + scale_y_continuous(labels = scales::comma) +
  ggtitle("Vehicle Cost by Vehicle Odometer") +
  xlab("Vehicle Odometer (Miles)") + ylab("Vehicle Price (USD)") +
  labs(colour = "Kick")
MMRbyOdo

#Box plot of Vehicle Cost by Size and Kick
bpSizeKick = ggplot(data = kick, aes(x = VehBCost, y = Size, fill = IsBadBuy)) +
          geom_boxplot() +
          scale_fill_brewer(palette = "Pastel2") +
          xlab("Vehicle Cost") + ggtitle("Boxplots of Vehicle Cost by Size and Kick") +
          xlim(0.0,15000) +
          labs(fill = "Kick")
bpSizeKick

#Box plot of Vehicle Cost by TopThreeAmericannName and Kick
bpNationalKick = ggplot(data = kick, aes(x = VehBCost, y = TopThreeAmericanName, fill = IsBadBuy)) +
        geom_boxplot() +
        scale_fill_brewer(palette = "Pastel2") +
        xlab("Vehicle Cost") + ggtitle("Boxplots of Vehicle Cost by TopThreeAmericanName and Kick") +
        xlim(0.0,15000)
bpNationalKick

#Box plot of Vehicle Cost by IsOnlineSale and Kick
bpOnline = ggplot(data = kick, aes(x = VehicleAge, y = IsOnlineSale, fill = IsBadBuy)) +
  geom_boxplot() +
  scale_fill_brewer(palette = "Pastel2") +
  xlab("Vehicle Age") + ylab("Online Sale") + ggtitle("Boxplots of Vehicle Age by Online Sale and Kick") +
  labs(fill = "Kick")
bpOnline


#Warraty Cost and Vehicle Odometer by Kick and Transmission
warrantyOdo = ggplot(data = kickSample, aes(x = VehOdo, y = WarrantyCost, colour = IsBadBuy)) +
  geom_point(alpha = 0.5) + geom_smooth(formula = y ~ x, method = "lm") +
  scale_x_continuous(labels = scales::comma) + scale_y_continuous(labels = scales::comma) +
  ggtitle("Warranty Cost and Vehicle Odometer by Kick and Transmisson") +
  xlab("Vehicle Odometer") + ylab("Warranty Cost") +
  facet_wrap(~ Transmission) +
  labs(colour = "Kick")
warrantyOdo

set.seed(1)
```