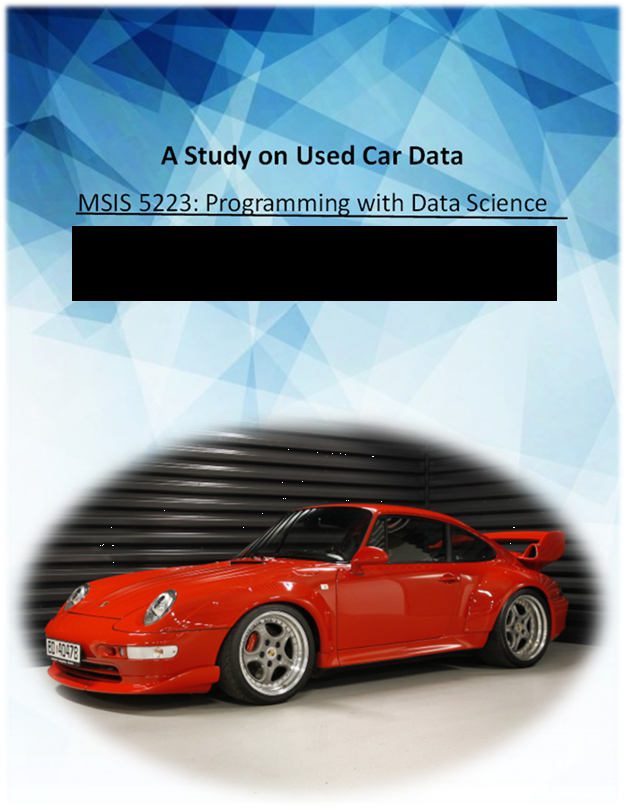
****

**Table of Contents**

Executive Summary 1

Statement of Scope 2

Project Schedule 5

Data Preparation 6

*1. Data Access* *6*

*2. Data Consolidation* *7*

*3. Data Cleaning* *8*

*4. Data Transformation* *14*

*5. Data Reduction* *18*

*6. Data Dictionary* *22*

Descriptive Statistics 24

**Executive Summary**

Selling used cars online, using eBay, in Germany is difficult. The seller must self-list and price their vehicle without professional help. Most people price their used vehicles using Kelley bluebook, online searches for similar vehicle prices, or use intuition to price their cars. This leads to both overpriced cars that don’t sell and underpriced cars that undermine owner’s profit. The research team believes that creating a predictive pricing model based on some vehicles attributes will make optimal pricing of used vehicles on eBay easier for private sellers. In order to create this model, the researchers used a database of 371,539 vehicles with 20 variables that were scraped from eBay-Kleinanzeigen between 3/5/2016 and 4/6/2016 (Leka 2016). This data was then cleaned, transformed and reduced to create a data set that included 292,425 observations and 16 variables. This data will be used to create the predictive pricing model that will offer pricing suggestions to eBay-Kleinanzeigen customers.

**Statement of scope**

This analysis is an investigation into what and how attributes of a used car listed on eBay-Kleinanzeigen in Germany affect a vehicle’s list price on the website. The goal of this analysis is to create a predictive pricing model using the available attributes scraped from eBay- Kleinanzeigen between 3/5/2016 and 4/6/2016.

The main “autos” dataset includes 371,539 vehicle listings records. The second dataset “cnt\_km\_year\_powerPS\_minPrice\_maxPrice\_avgPrice\_sdPrice” is an aggregation of the autos dataset sorted by horsepower (powerPS), year and odometer reading (km) which includes 1770 entries.

The investigation will be in 2 phases. The first phase is data preparation, which includes: access, consolidation, transformation, and reduction of data, along with descriptive statistics. The second phase, model design and implementation include model selection, data splitting/sub-sampling, model execution and model assessment.

**Project Objectives**

* Assess the effect of various used car attributes like mileage, horsepower, make, model type etc. on used vehicles listing price on eBay- Kleinanzeigen in Germany
* Build models that predict used vehicle listing price based on the statistically significant vehicle attributes.
* Choose best price prediction model based upon a models least squared values.

**Target Variable**

The target variable for this investigation is the “price” which is the listing price of a used vehicle on eBay- Kleinanzeigen. This variable takes on integer values between 0 and €99,999,999 which is the value of the vehicle in Euros €. The goal is to predict the list price using one or a combination of the predators listed below.

**Predictors**

* vehicleType
* year of registration
* gearbox
* powerPS
* model
* KM
* fuelType
* brand
* Repaired
* Age
* Country

The analysis of eBay-Kleinanzeigen used car data is threefold. First, analysis of what scraped attributes of the EBay-kleinanzeigen affect used car selling price in Germany is needed. This entails initial consolidation, cleaning, transformation and reduction of the dataset. The data reduction encompasses both statistical and logical data reduction. The statistical data reduction includes PCA and clustering analysis. While the logical data reduction removes any factor data that only consists of a single level making an analysis based on that attribute difficult. After this initial data reduction descriptive statistics on the remaining used car attributes will be analyzed and, if needed, further data reduction will take place. The descriptive statistics on the remaining attributes will guide model choice, design and impact the final decision on the used car attributes implemented in the final model. Before the data model is built the data will be split into testing and validation data sets. Once splitting is complete the model design and construction will begin. Once the model is complete it will be assessed using the validation dataset created previously.

**Project Schedule**

The project schedule is made up of two deliverable phases. The first is the data preparation phase and the second is the data modeling phase. The data preparation phase must be completed by 3/18/2018 and the modeling phase must be completed by 4/19/2018 and submitted on 4/20/2018. The group will hold project meetings at least once a week during MSIS 5223 class period. All other group communication will be done through WhatsApp and additional meetings will be scheduled as necessary. If the group falls behind schedule they do have some lag time available to complete any unfinished work before the hard due date of May 6th.

**Team Meetings**

3/6/2018

3/13/2018

3/15/2018

3/16/2018

4/17/2018

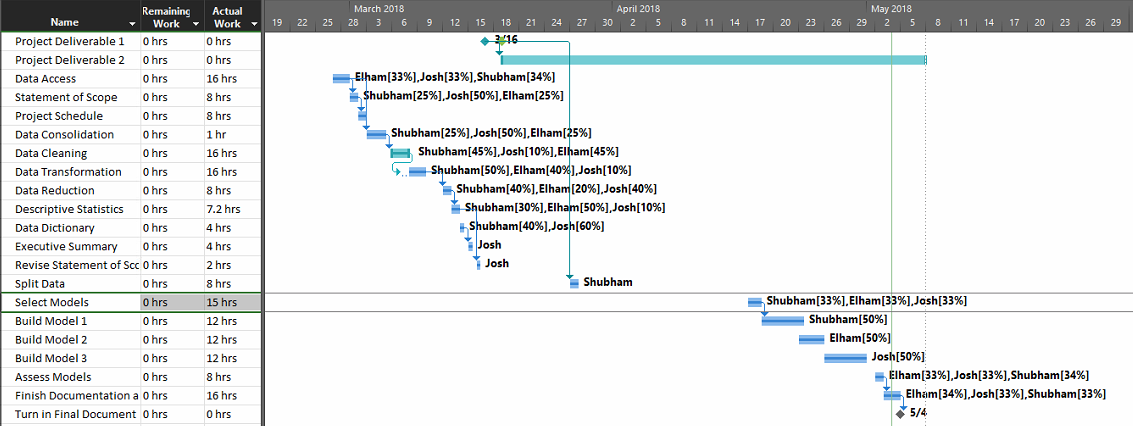
4/24/2018

4/27/2018

4/2/2018

4/3/2018

4/4/2018



**Data Preparation**

**Data Access**

The researchers were interested in predictive modeling of used vehicle prices on websites like eBay or Autotrader. The goal of this predictive model would be to provide eBay users an optimal selling price for their used car based on some attributes of the vehicle. The researchers searched Kaggle and various online data stores for online used vehicle datasets that included various vehicle attributes, including a couple of core attributes like make, model, vehicle age, and mileage. The researchers found used cars database on Kaggle that included the core vehicle attributes and met the researcher’s data needs (Leka 2016).

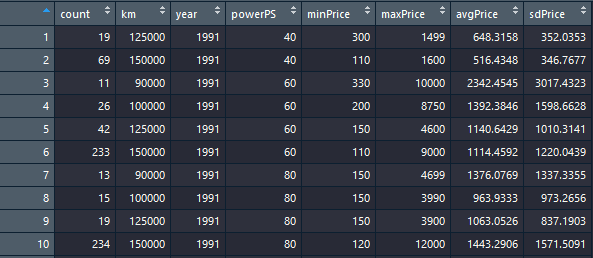
<https://www.kaggle.com/orgesleka/used-cars-database>

The Used cars database data set was scraped from eBay-Kleinanzeigen using Scrappy. eBay-Kleinanzeigen is the German arm of eBay that is specifically aimed at goods that would normally be sold using classified ads (eBay 2018).There are 2 different files included in the Used cars dataset. The first file “autos.csv” includes 371824 eBay-Kleinanzeigen used vehicle listings that were scraped from the website between 3/5/2016 and 4/6/2016 with 20 variables. This is the primary data source that the researchers intend to use for their investigation and predictive online used car pricing models.

The second file cnt\_km\_year\_powerPS\_minPrice\_maxPrice\_avgPrice\_sdPrice.csv is aggregate of the results of the autos.csv file grouped by powerPS, year, km. file contains 1770 records, 8 attributes. Four of the attributes are aggregate statistics which include count, maxPrice, avgPrice, sdPrice.

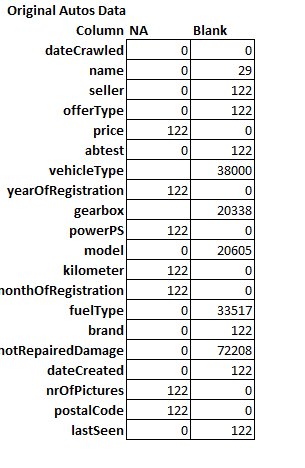
**Data Consolidation**

The only data consolidation that is necessary is dropping the cnt\_km\_year\_powerPS\_minPrice\_maxPrice\_avgPrice\_sdPrice.csv file. This does not risk any data loss because the file is aggregate data of the autos.csv data and can be recreated from the original source.

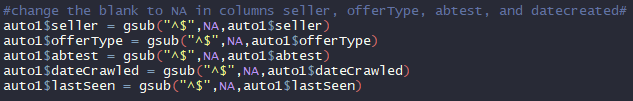


**Data Cleaning**

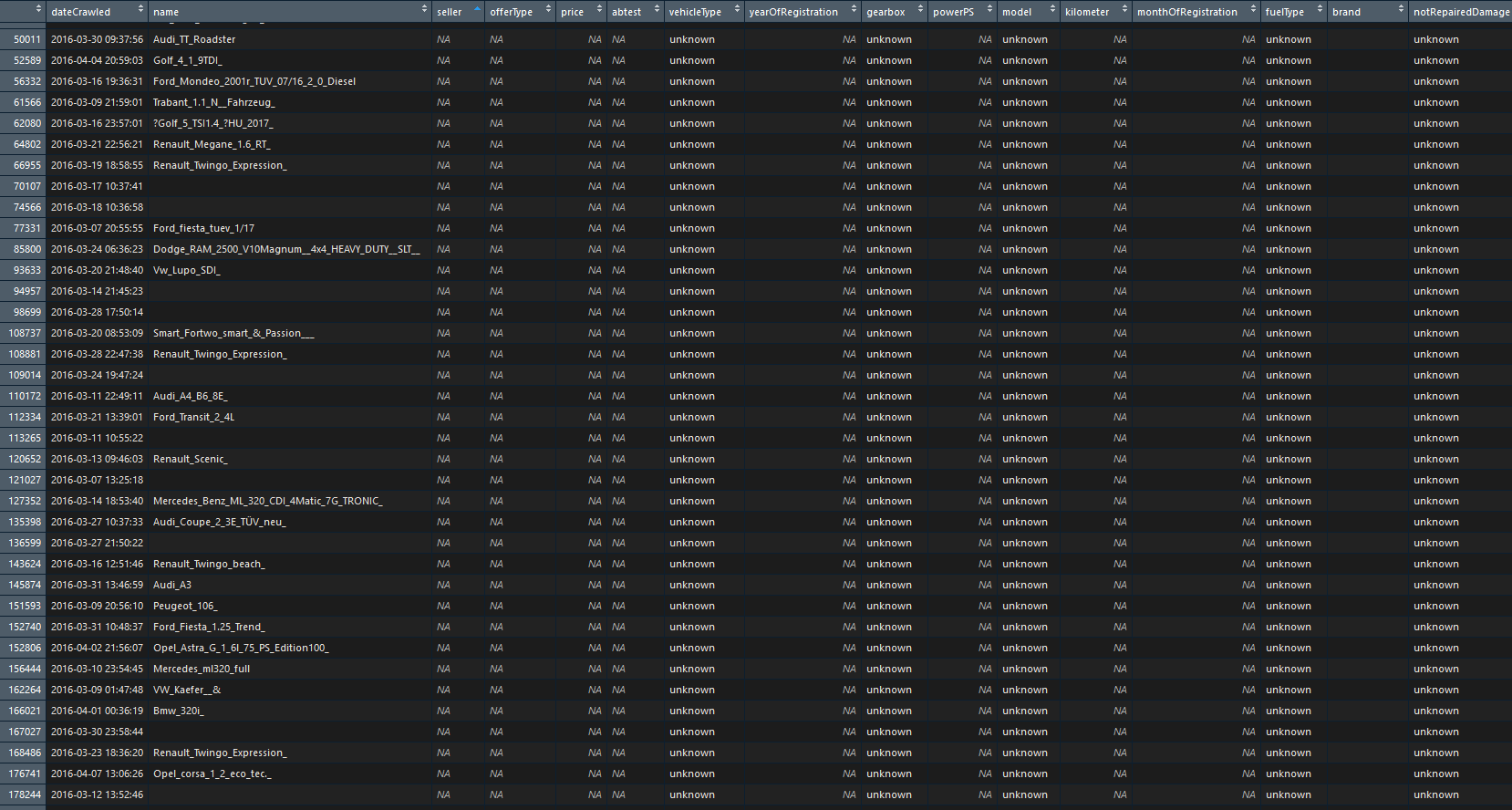
The autos dataset required cleaning before it was ready for analysis. Most columns either had several null or blank values, as shown below.

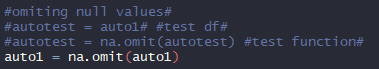


The research team chose to deal with blank values in one of two ways. If a column had blank values that summed to equal 122, which is the same amount as the sum of the NA values in any column. The blank values would be changed to NA and all 122 NA values would be dropped. The blanks in columns: seller, offerType, abtest, dateCrawled and lastSeen summed to equal 122 and were changed to NA. Subsequently all NA values were dropped.



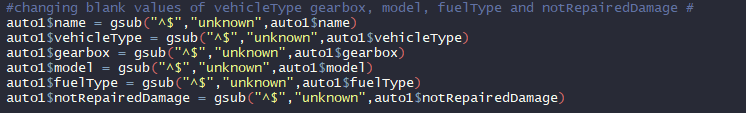
Once all the values were changed the researchers found that 122 rows of data only contained data in the columns name and dateCrawled. Because these rows of data lack the target variable price and most other variables they were omitted from the dataset using the omit.na function. Which will bring the total number of tuples to 371,824 – 122 = 371,702 that leaves the researchers with enough data to continue their research with the dataset.







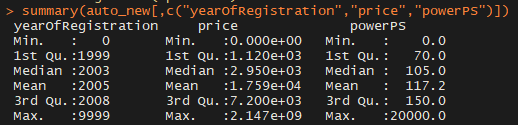
If a column had the sum of blank data that was greater than or less than 122 and was categorical data, the blank value was changed to unknown. The reason for this is that the absence of a value in certain categorical data sets may affect the list price of a vehicle and by removing the data due to NA values may change the analysis. For example, people may pay less for a vehicle online if the type of gearbox is unknown. The blanks in columns: name, vehicleType, gearbox, model, fuelType, and Repaired were changed to unknown.



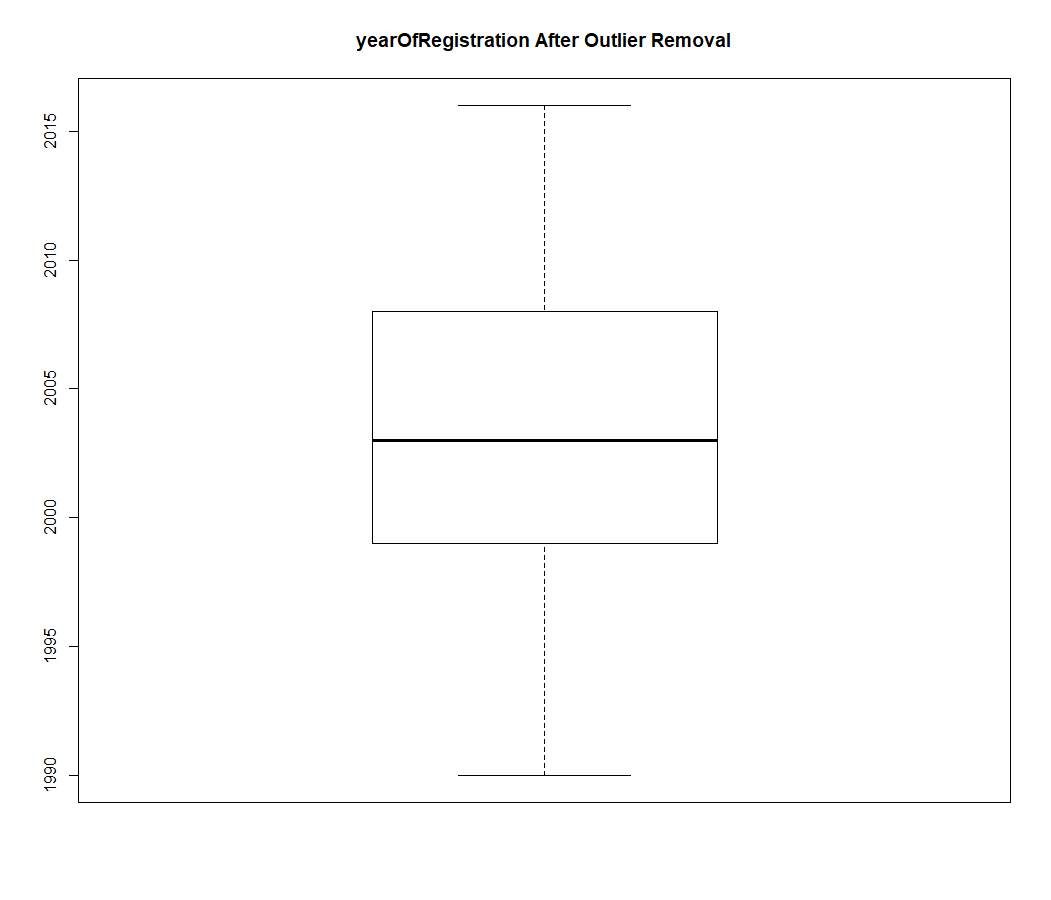
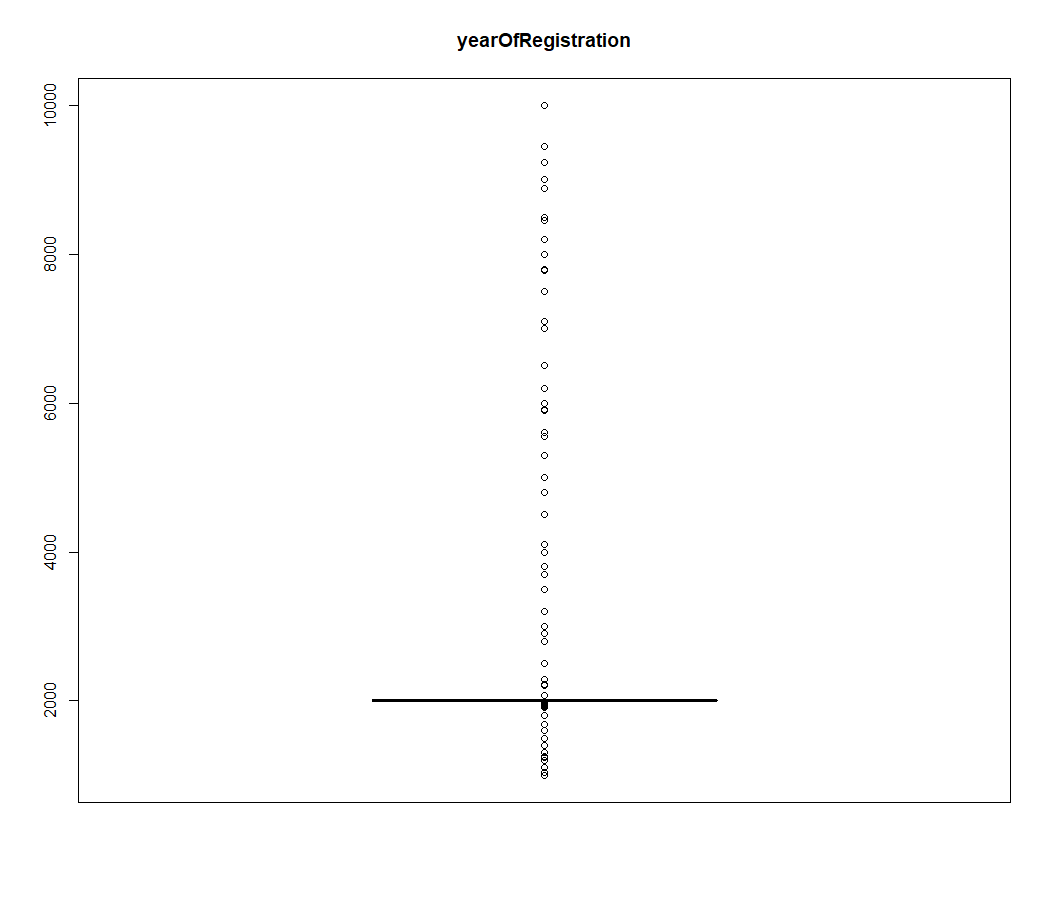
After the NA and blank values were corrected the team removed any **duplicate values** using the drop\_duplicate function in python. Which removed an additional 8932 records from the dataset.



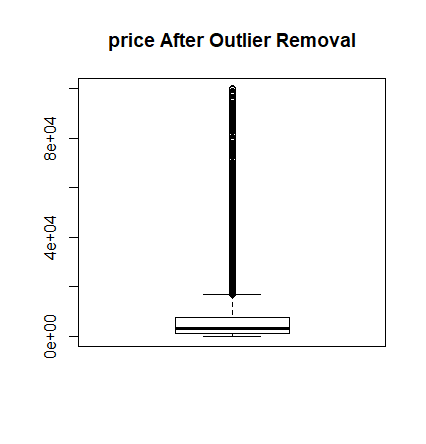
Once the NA’s, blank values and duplicates were corrected, the team looked at the data for **outliers** using the summary function in R. The summary data revealed that price was highly right-skewed and had major outliers including



The detection of outliers in numerical data was completed using boxplots. YearOfRegistration had a number of outliers. These outliers were resolved by setting the minimum year at 1990 and a maximum year of 2016. 1990 was chosen because we didn’t want to have classic car prices skewing the dataset. This is because classic car pricing is not based on the same attributes of the vehicle as regular vehicles. The maximum of 2016 was chosen because our data was gathered in 2016 and any cars built after 2016 would be entry anomalies in the dataset. After the outliers were removed the yearOfRegistration became roughly normally distributed.

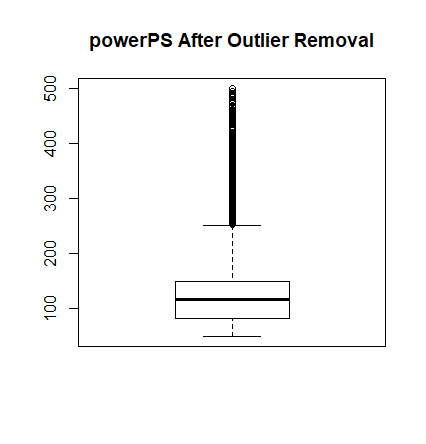
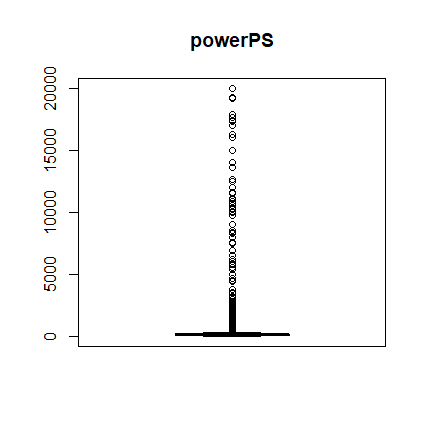


Price has a number of outliers, the worst being a single car with a price of €2147483647 and 15 cars having price €99,999,999. These high price points need to be removed because they are either unrealistic prices set by the seller, pricing mistake made by the seller or rare vehicle whose price is not affected by the given variables the researchers are studying. To resolve the outliers, the maximum price for a vehicle was set at €100,000 and the minimum was set at €100. The minimum was set at €100 to eliminate any vehicles that were being sold as scrap.



Still, after reducing the range between 100 and 100,000 it is visible that Price variable is still right-skewed. One possible alternative is to log transform Price.

PowerPS had a number of outliers. This is due to the variety of vehicles listed which includes sports cars and the additional need for some to drive at the high speeds found on the Audubon. The minimum and maximum powerPS values were selected to remove motorcycles and high-end sports cars from the data to a minimum power level of 50 and maximum power level of 500 was chosen. The removal of outliers was done to remove the noise from high end sports cars and non-car vehicles from the dataset.



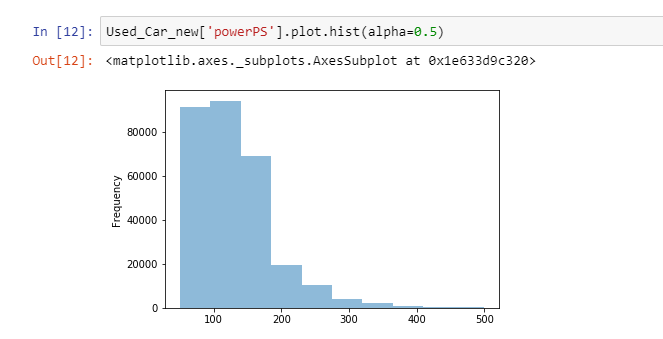
Still, after reducing the range between of power PS it is visible that the variable is still right-skewed. One possible alternative is to log transform powerPS.

**Data Transformation**

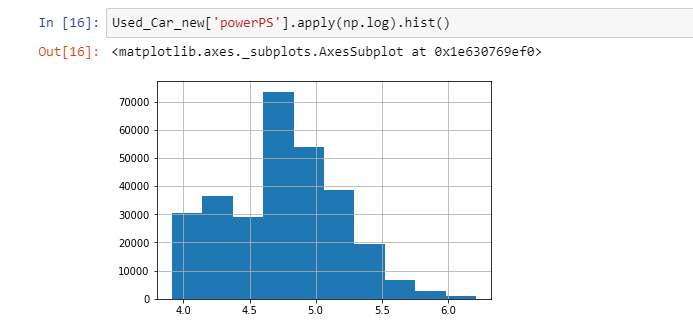
A number of data transformation processes were necessary to increase the usability of the data for statistical analysis.

1. **Logarithmic Transformations**
   1. **powerPS**

The logarithmic transformation normalized the powerPS variable from extreme right skewness.

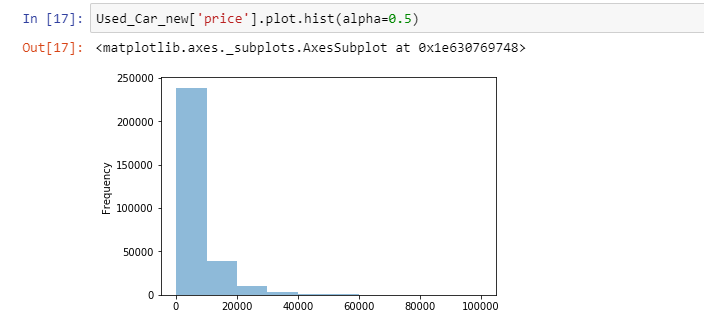


Log Transform:

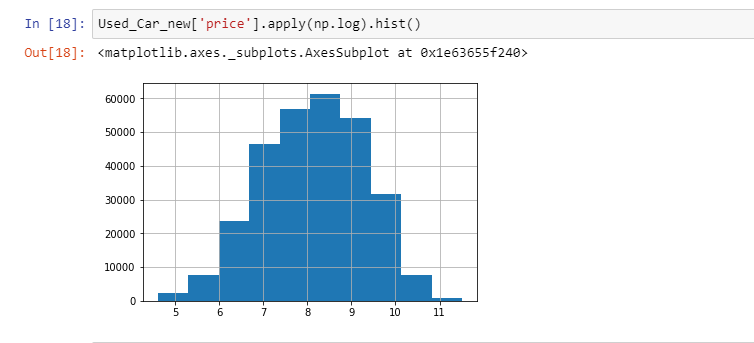


1. **Price**

The logarithmic transformation normalized the price variable from right skewness.



Log Transform:



1. **Derived/New Columns** 
   1. **Age**

Age=2016 (year data scraped) – the year of Registration

This allows for simpler analysis and understanding on how the age of a vehicle affects is online selling price vs yearOfRegistration variable.

* 1. **Country**

Allows analysis based on whether a vehicle is an import or German made affects its resale value. An example of this would be BMW is headquartered in Germany. This column was manually created in Excel based on the brand name of the vehicle.

1. **Binning**
   1. **Kilometer**

The kilometer variable is binned at 1, 75,000, 100,000, 125,000, 149,999 and 999,999 to normalize the variable.



**Data Reduction**

**Dropping single Level variables**

There were a number of variables that needed to be dropped from the dataset due to the lack of variability in their values.

1. **Seller** only had 4 variables that were non-private.

4/371539= .00001076\*100=.0001076% chance of a non-private seller which shows that the variable is nearly homogenous.

1. **OfferType** only had 1 record at the second level (Gesuch)

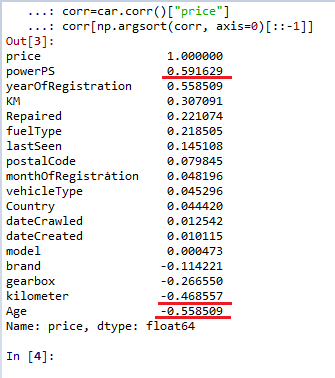
1/371539=.00000269\*100=.0000269% chance of Gesuch offerType and shows that the variable practically speaking only has one level.

1. **Abtest** variable was removed from the dataset because it only contained data for splitting the data into the control and test datasets and the researchers will split the dataset themselves later with their own splitting parameters.
2. **nrOfPictures** was removed because it only had one integer value 0 and a number of null values, hence it was removed.



**Correlation among Target Variables and each predictor**

After performing correlation analysis which compared each variable against the target variable price we came to 2 conclusions. The first is that variables KM and Repaired have the strongest positive correlations with the price. YearOfRegistration is excluded because yearOfResistration and Age effectively equivalent. The second conclusion is that the Age and Kilometer variables have the strongest negative correlations with the price. Third, the variables dateCreated, dateCrawled, vehicleType,monthOfRegistration are the only variables that may be dropped based on their low correlation with price on an individual level. It may make sense to drop dateCrawled because that variable has to do with the gathering of the data and not the data itself. That decision will be made in phase 2 during model design and construction.



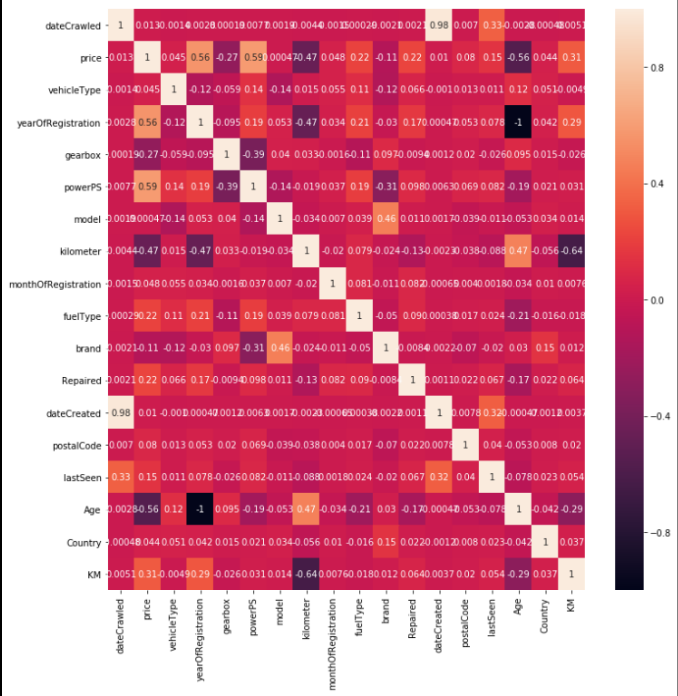
**Correlation Table**

Correlation table analysis was performed on the test and predictor variables. The analysis found that 4 pairs of variables had the highest correlations among their peers. It was not surprising that the four pairs of variables had a higher correlation with each other than normal for three reasons. The first is that KM is a binned variable version of the kilometer. The second is that older cars generally have more miles driven on them. This explains- kilometer and yearOfRegistration; Age and kilometer. The third is that models of vehicles are directly related to their brands (e.g. Only Honda can make an Accord). However, none of the correlations among these variables were high enough to cause multicollinearity anomalies later in the model design and construction phase.

1. **KM and kilometer**
2. **Kilometer and yearOfRegistration**
3. **Age and kilometer**
4. **Brand and model**

**Correlation Matrix Analysis**

Due to the direct measurement of variables in the dataset and the lack of behavioral factors. The need to reduce using PCA analysis is not necessary for the dataset. Instead, correlation matrix analysis is used to reduce the number of variables.



**Initial Dataset after Initial Cleaning and Data Reduction**

**New Dataset that was left after these changes is called: Used\_Car\_Data\_new.csv**

**Regression Dataset after Cleaning and Data Reduction For Regression Analysis**

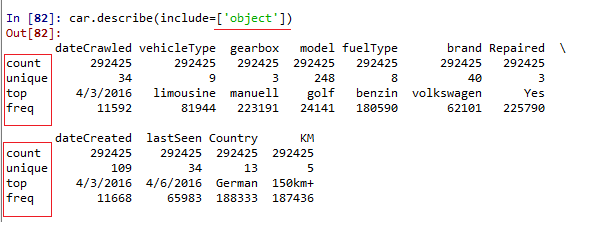
**Data Dictionary**



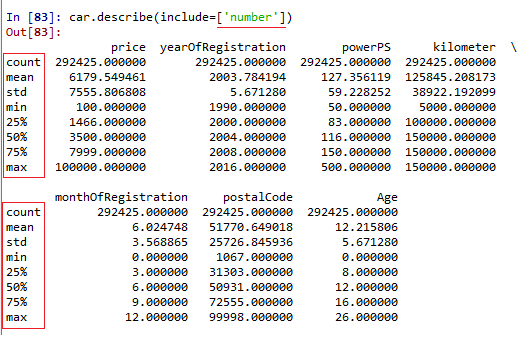


**Descriptive Statistics**

**Summary of all Object Statistics**

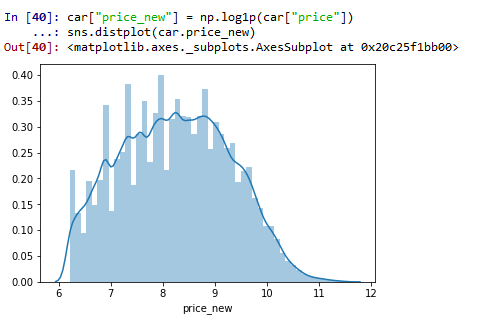
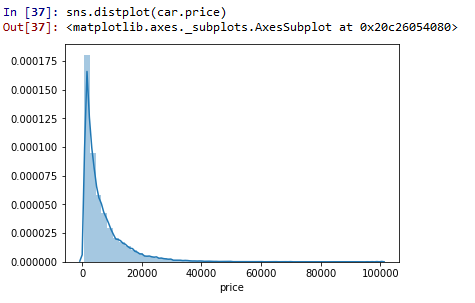
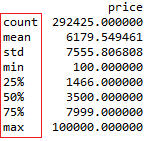


**Summary of all Numeric Statistics**

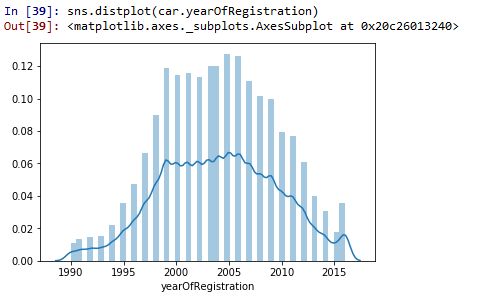
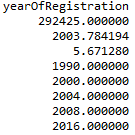


1. **Target variable price**

Price variable is shown to be extremely right skewed and needs to be transformed in order to normalize the variable. This was accomplished using a logarithmic transformation. The range of the data is (100, 100000) which is expected due to the min and max levels selected by the researchers during data cleaning. Now that the target variable is normally distributed model generation using the price target variable is possible.

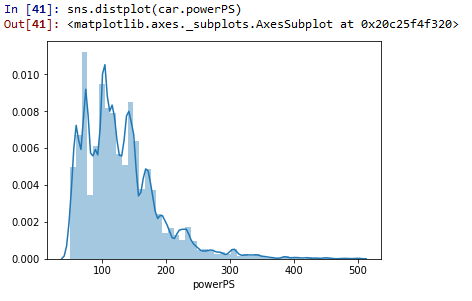
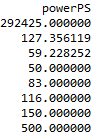


1. **yearOfRegistration**



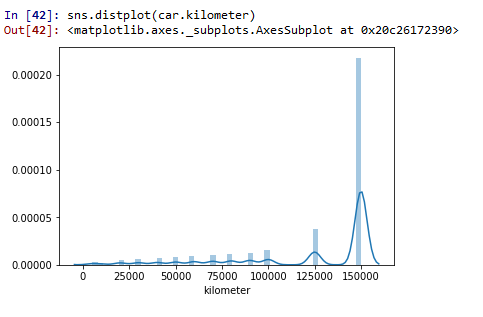
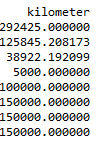
1. **powerPS**

PowerPS is a key predictor variable because there is relatively high correlation between the price of a car and the vehicles horsepower. The variable is right skewed, which is not surprising because most vehicles are under 200 horsepower which fits that most vehicles are not built for speed or high torque. As horsepower increases you are usually dealing with sports cars whose price is normally higher than traditional cars, suv, crossovers.



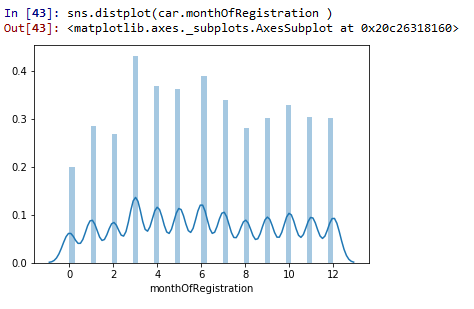
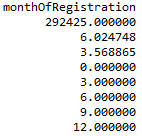
1. **kilometer**

This is a histogram of the kilometer variable. It is left skewed and was binned to normalize the variable as will be seen later.



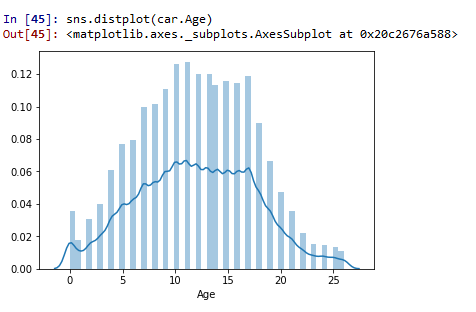
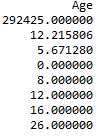
1. **monthOfRegistration**

This is a histogram of monthOfRegistration. It is relatively uniformly distributed because vehicles are produced and sold all year long with some minor variability.



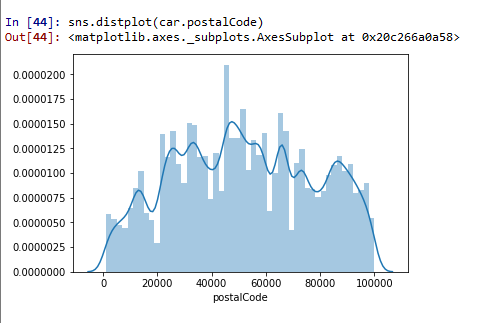
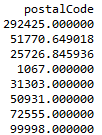
1. **Age**

Age is normally distributed and another key variable for our analysis which negatively correlates with price.



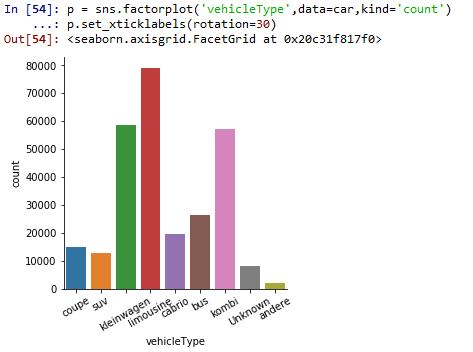
1. **postalCode**

This is a histogram of the postalCode variable and is distributed normally.



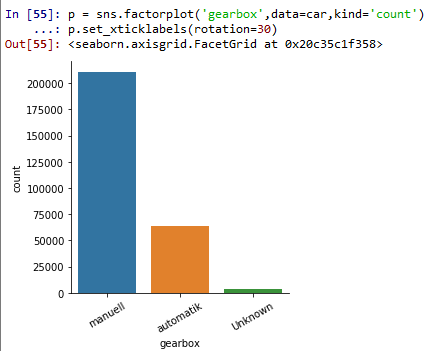
1. **vehicleType**

The distribution of vehicleType variable.



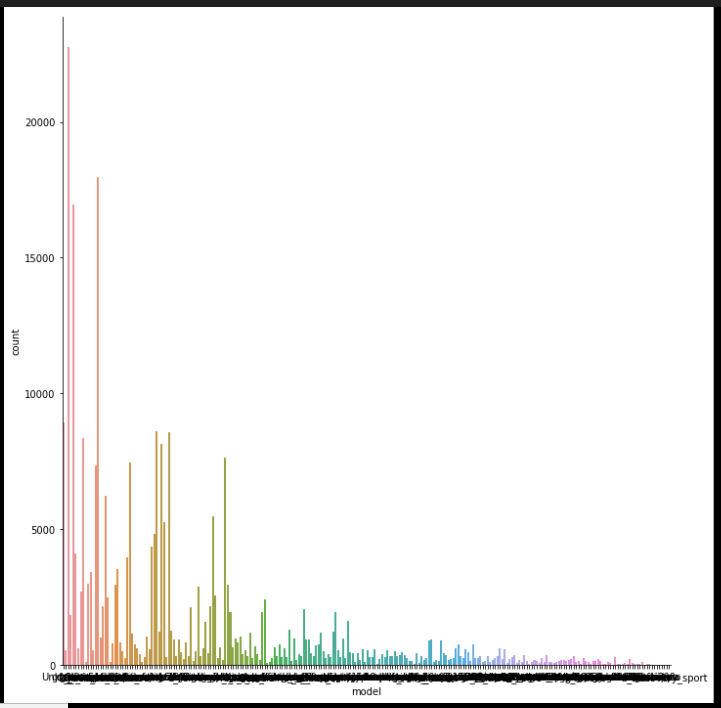
1. **gearbox**

Below is the distribution of gearbox variable. It is interesting to note that most of the used cars that are listed are manual which is in stark contrast to the US used car sales where most vehicles feature automatic transmissions. May also indicate that there are more sports cars listed as sports cars tend to have more manual transmissions compared to other types of vehicle.



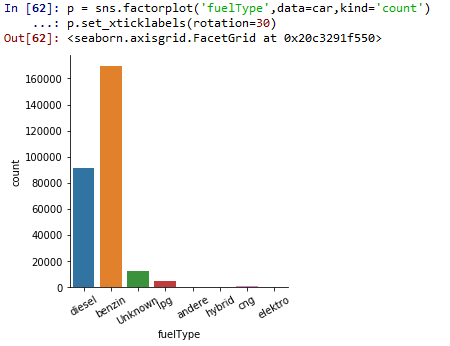
1. **model**

The distribution for the model variable is right skewed.

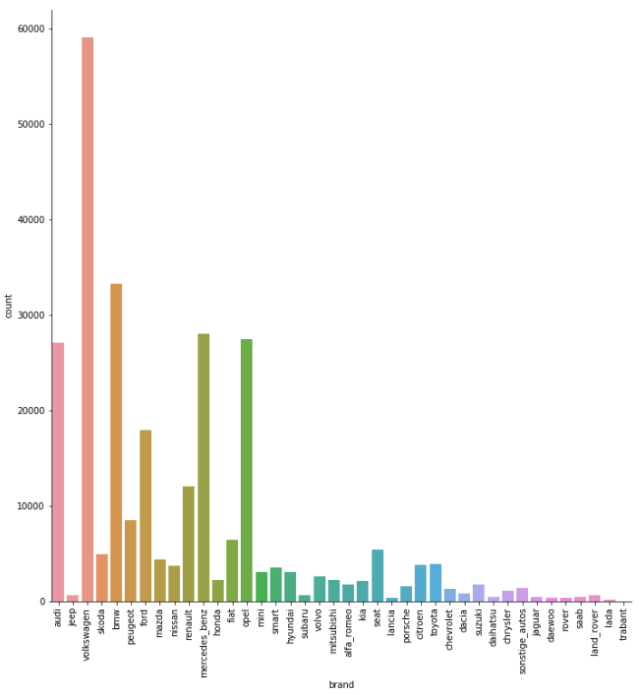


1. **fuelType**

Below is the distribution for fuelType

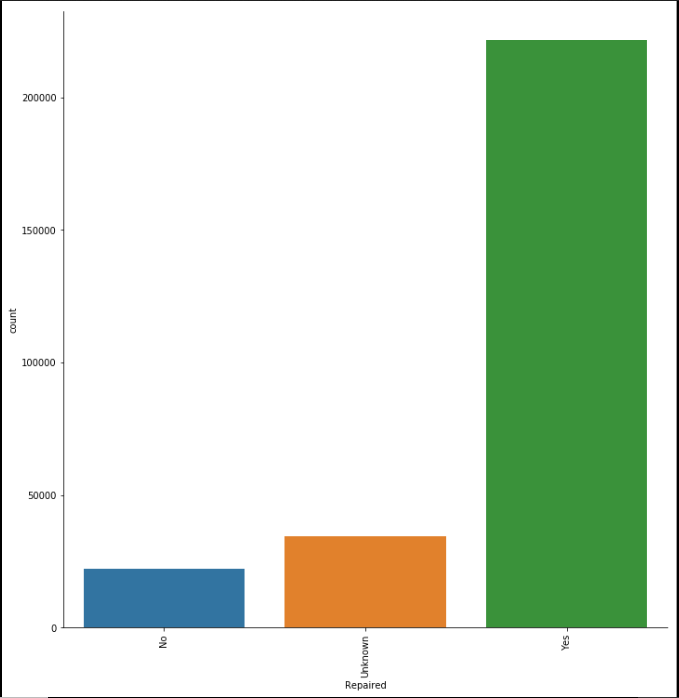


1. **brand**Below is the distribution for brand.

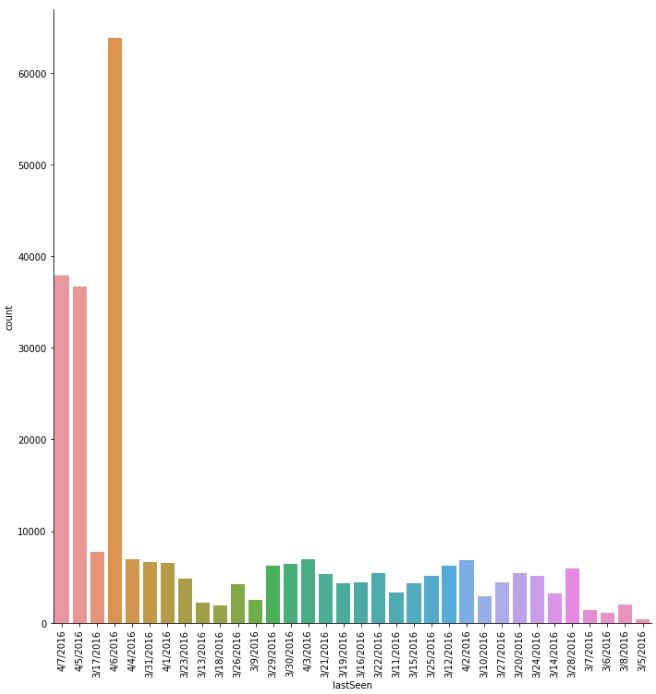


1. **Repaired**

The repaired distribution is displayed below.

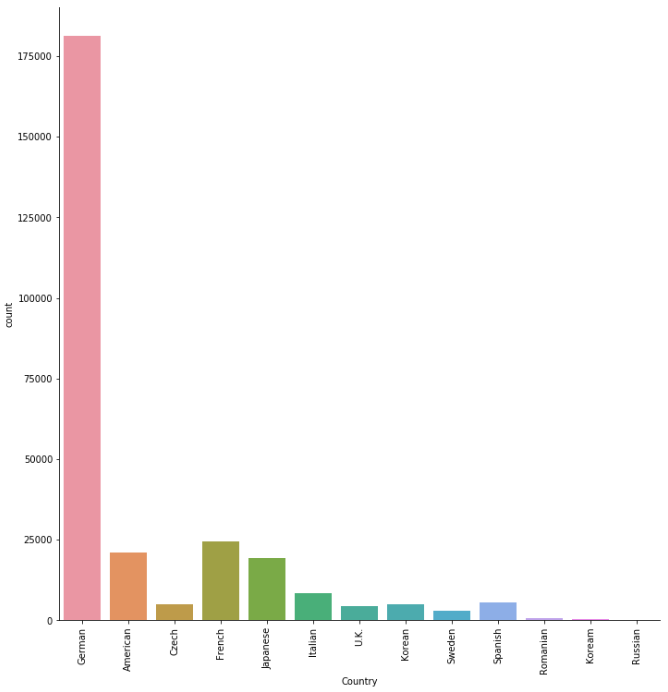


1. **lastSeen**



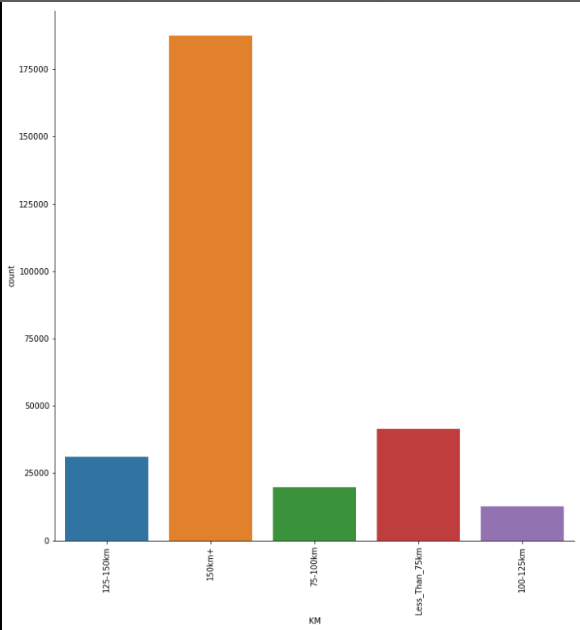
1. **Country**

Country variable shows the distribution between domestically and foreign cars in Germany. We see that German cars clearly are the dominant vehicle choice for German drivers.



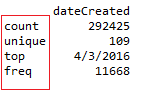
1. **KM**

Below is the distribution of the binned kilometer variable that is now called KM. The majority of the used cars listed have over 150,000 km.



1. **dateCreated**

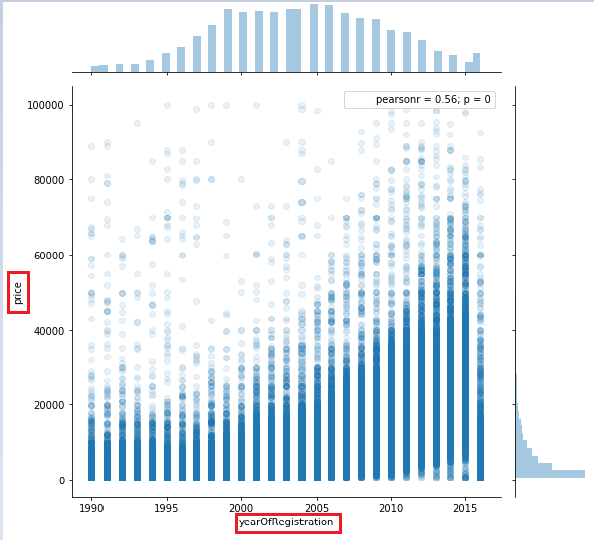
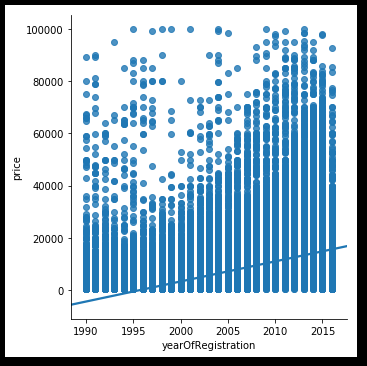
Below are the descriptive statistics for the dateCreated variable.



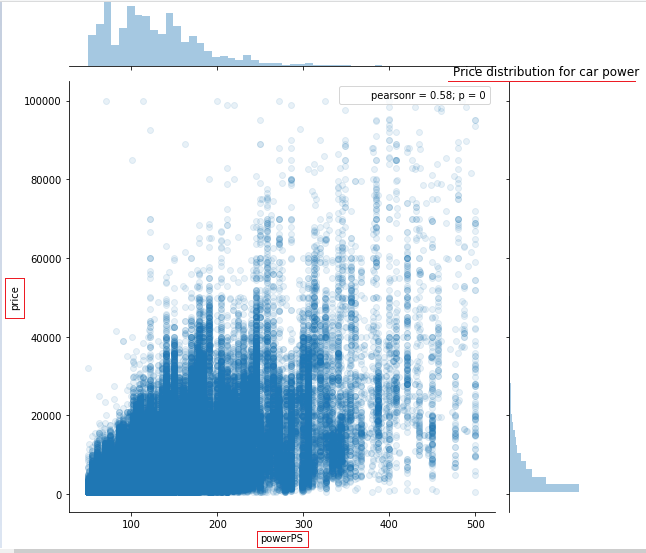
**Relationships Between various variables and price**

Here we see that powerPS and Year of Registration/ Age variable has direct impact on the price. Further analysis will be done in phase 2 of model deliverables.

**Price and YearOfRegistration**

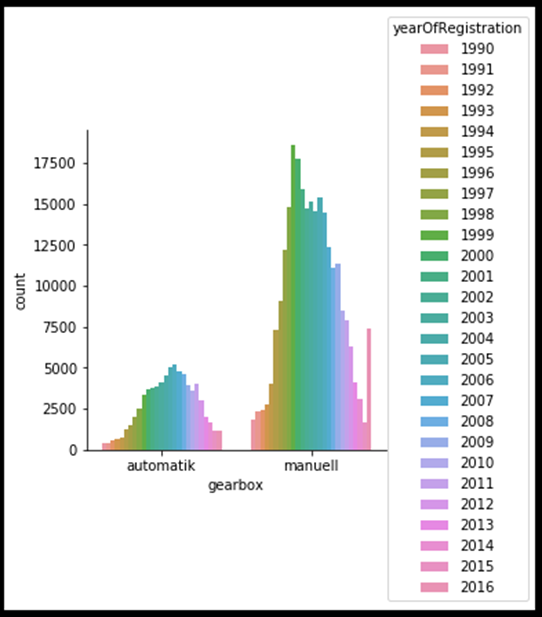


**Price and PowerPS**



**Gearbox and YearOfRegistration**

Most ads seem to involve vehicles with manual transmission systems .It could be due to people tending to upgrade their old cars and replace them with newer models with automatic transmission



**Selected Modeling Techniques**

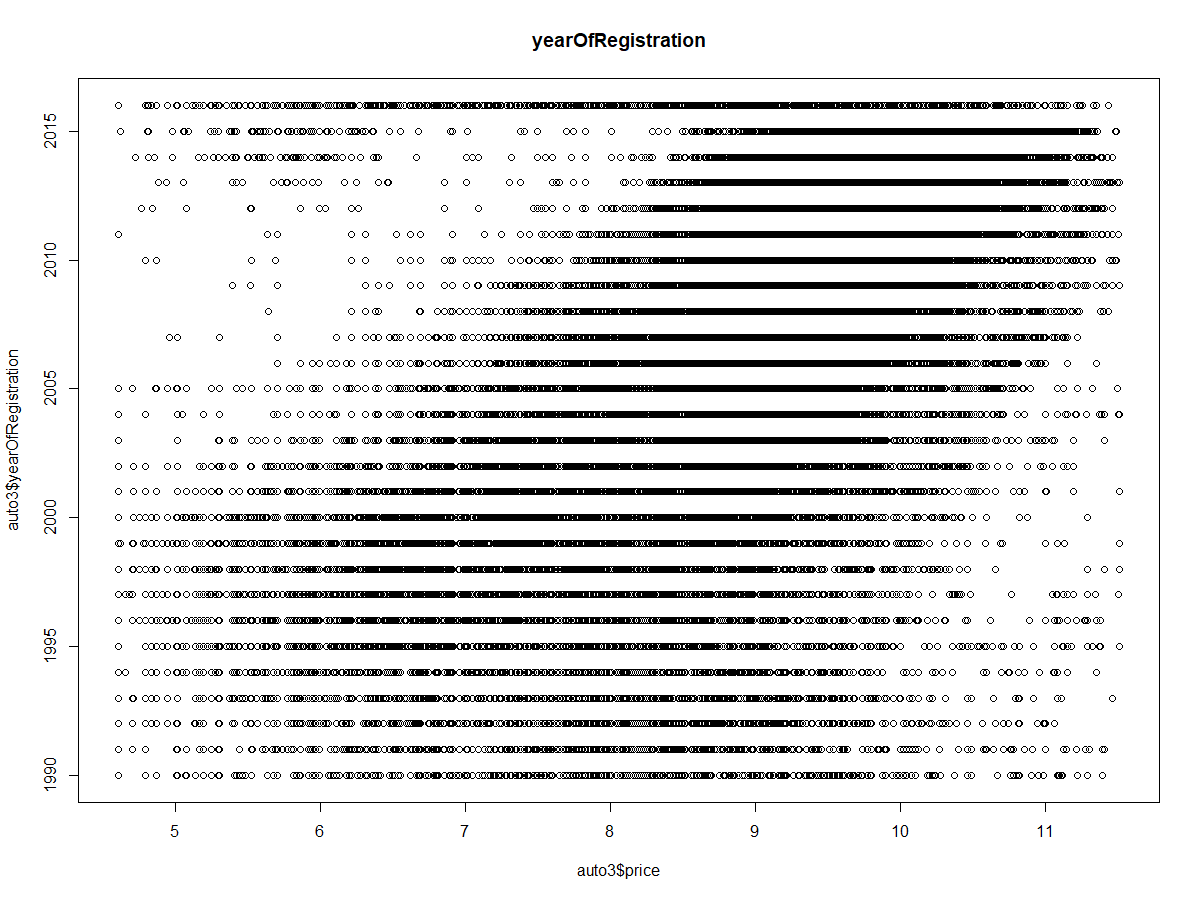
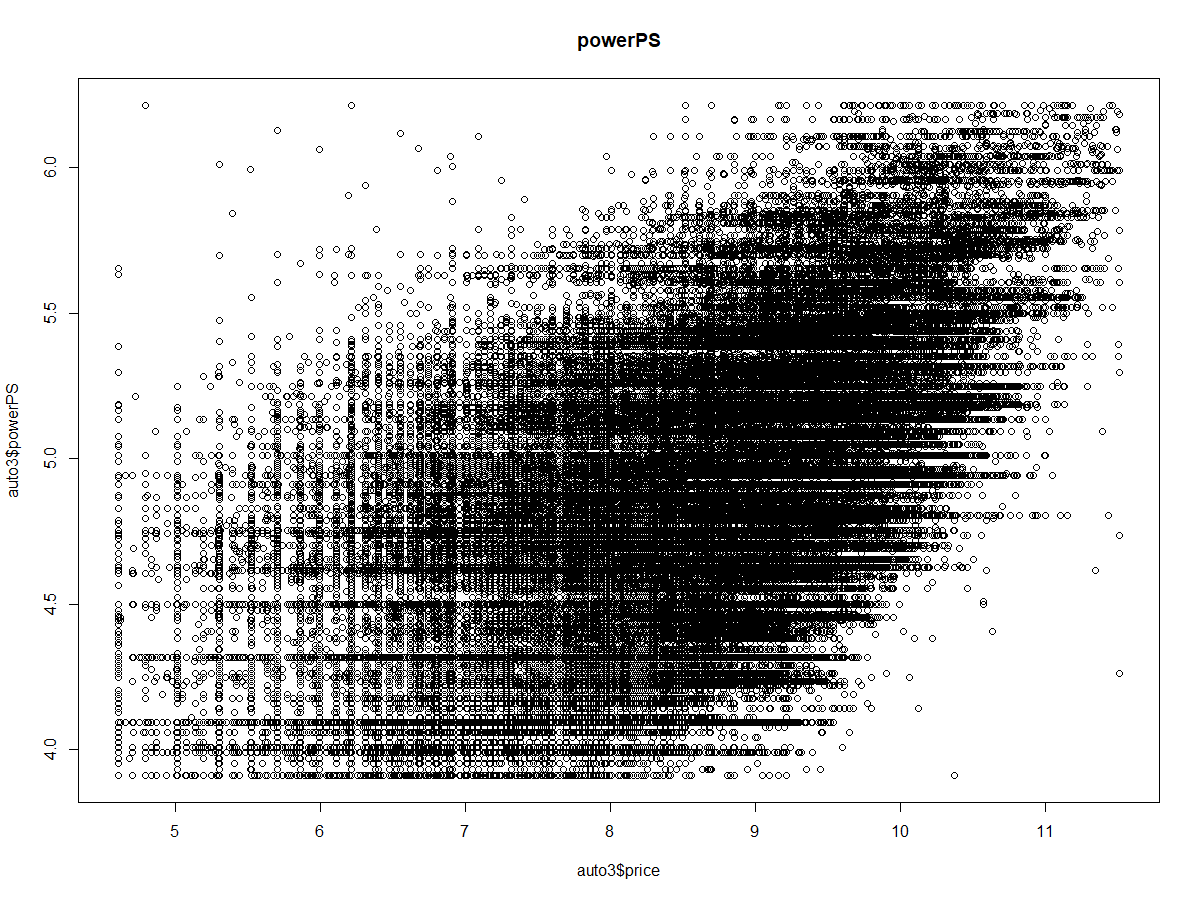
The two modeling techniques chosen were multiple regression and neural network. These analysis were chosen because they allow prediction of numerical multiplicative dependent variables with a combination of both categorical and numeric independent variables.

**Regression Goals**

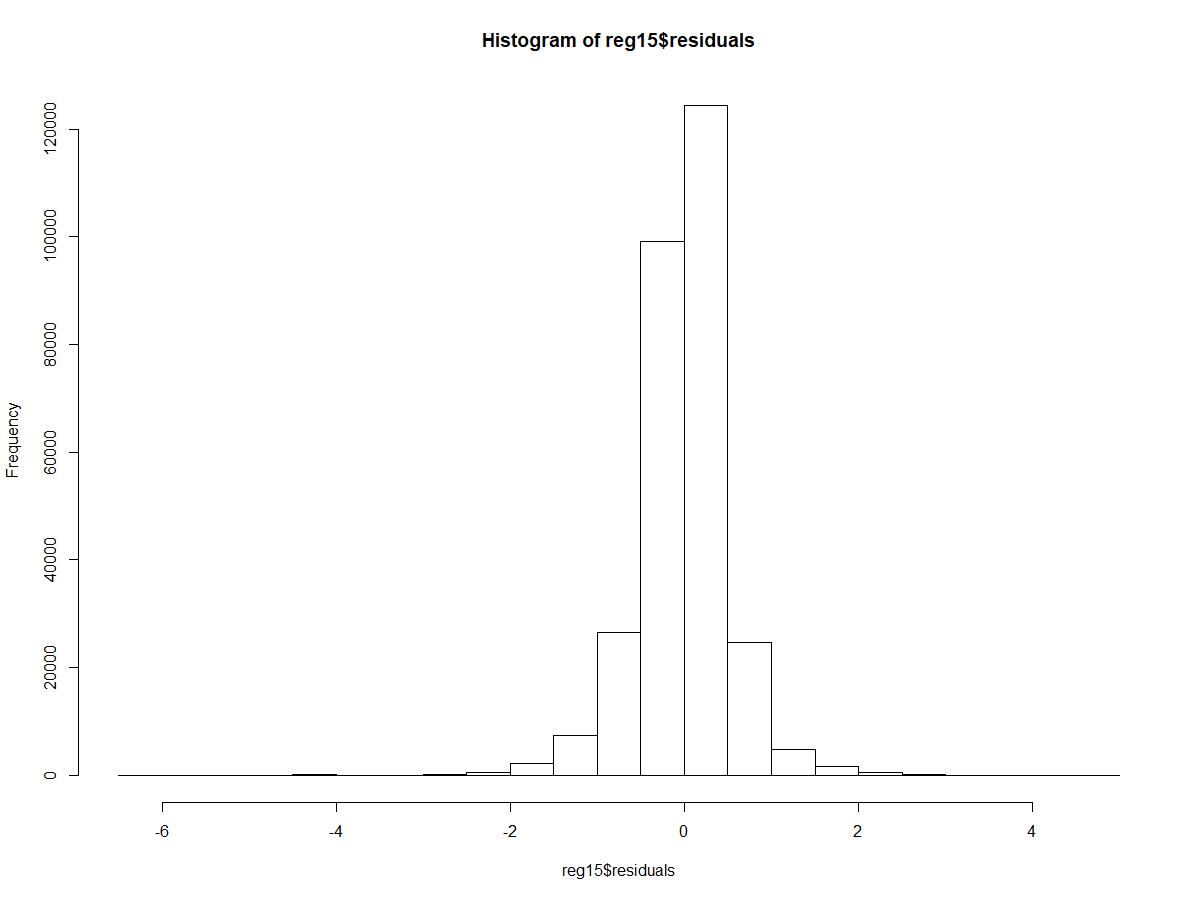
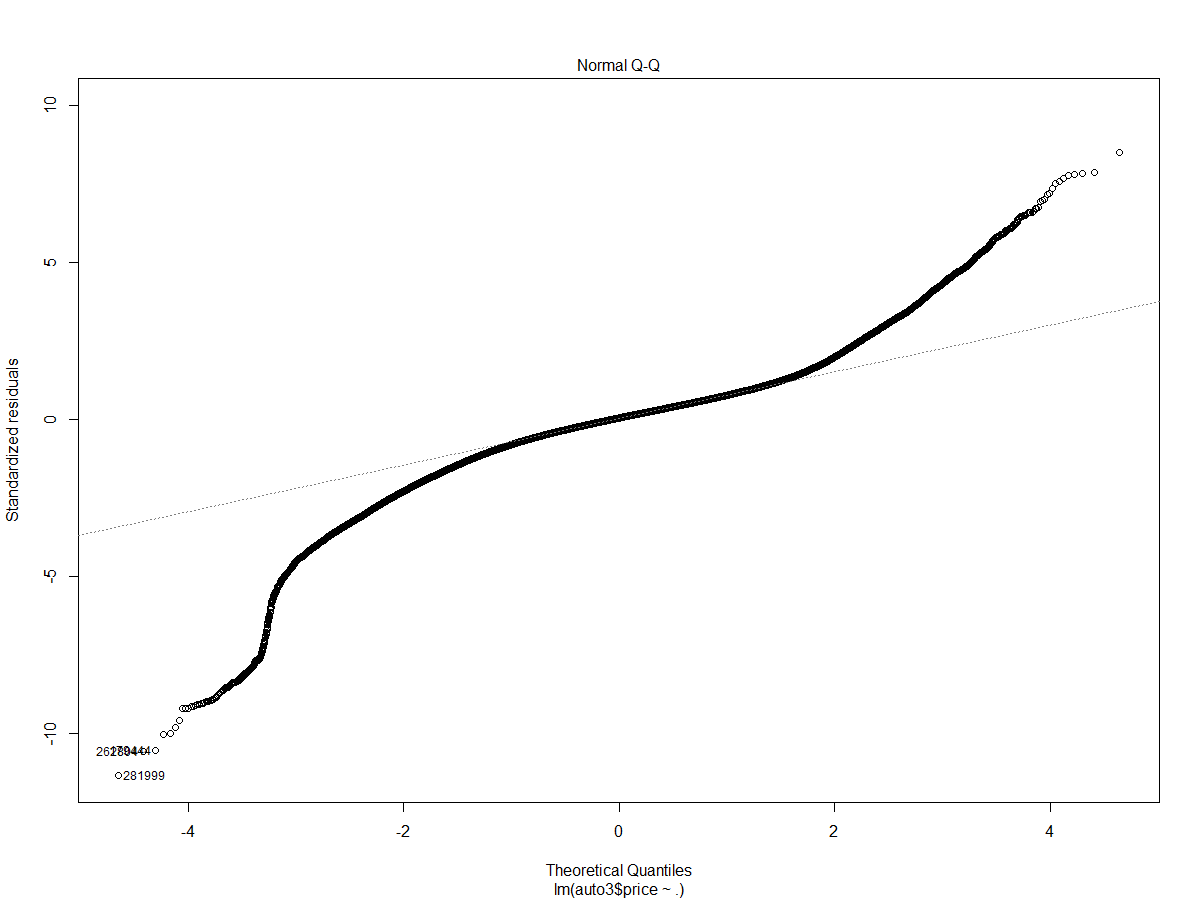
The goal of regression is to create a model using weighted values of the various categorical and numeric independent variables in the dataset. What this means is that we take a single unit of PowerPS and multiply it by its weight and see how much that single unit increase in PowerPS increases the price of a vehicle. Then we add that weight to the other weighted variables to come up with a predicted price of a vehicle.

**Regression assumptions**

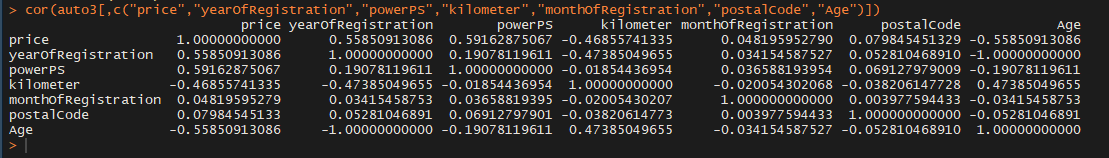
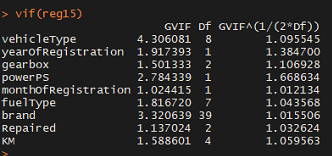
**Linearity:** The linearity assumption is the assumption that the relationship between an independent variable and the dependent variable are linear. Any variables that do not display linear relationship must be transformed into linear relationships. All of the numeric variables that were used in the regression analysis are linear or were transformed into a linear relationship. Below are graphs that plot yearOfRegistration, powerPS and Kilometer vs Price.

**Normality:** linear regression analysis requires all variables to be normal. All of the variables included in the dataset are normally distributed except powerPS and Price which required a log transformation in order to be normally distributed. The normality of the residuals after regression also need to be normally distributed and while the QQPlot seems to show that the residuals are not normally distributed.

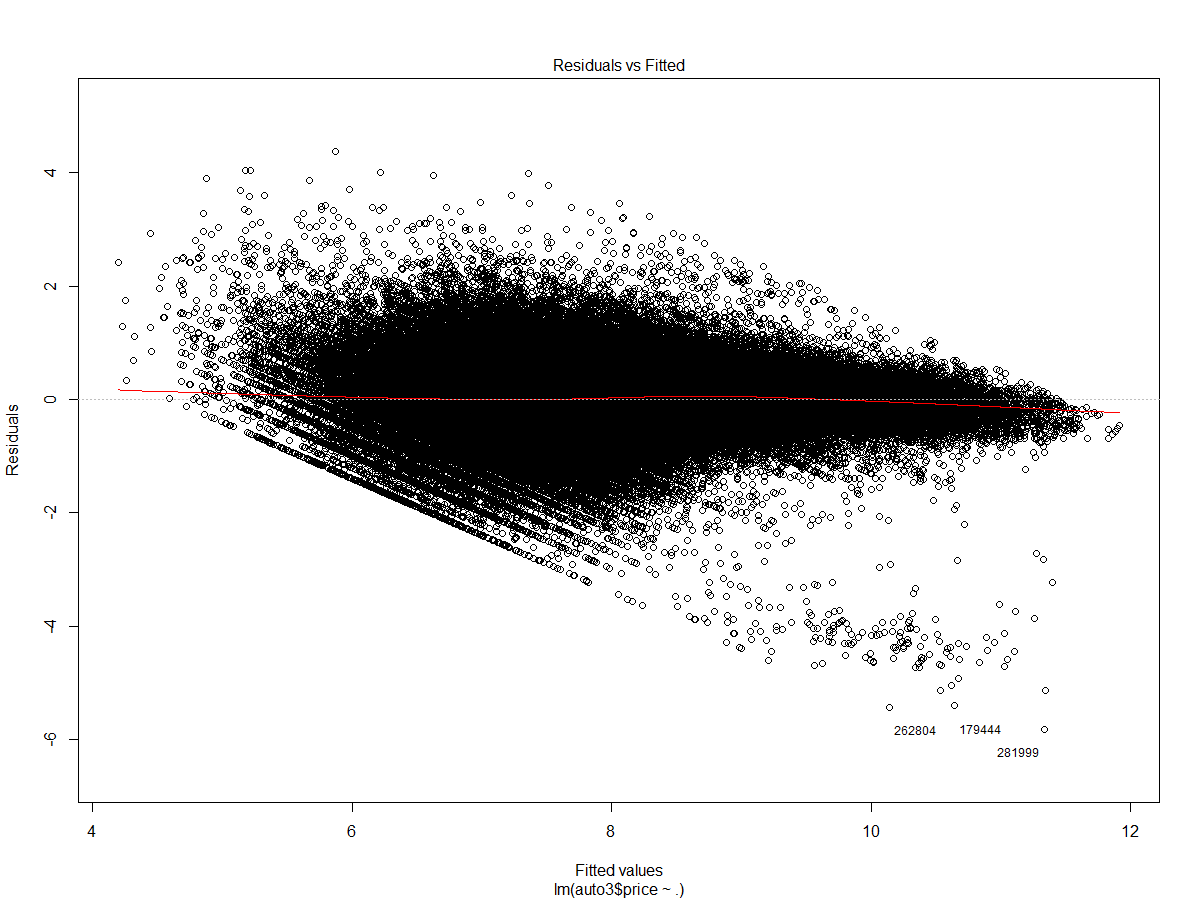
 

**Multicollinearity:** This assumes that the independent variables are only significantly correlated with the dependent variable and not correlated with other independent variables. Multicollinearity is only an issue for numeric variables and none of the numeric variables in the dataset display multicollinearity as shown is the correlation matrix below.

**Homoscedasticity:**

The homoscedasticity assumption is violated by the dataset because the plot of the residuals shows that they are not evenly distributed over price. Transformation we were not able to improve the homoscedasticity of the residuals.

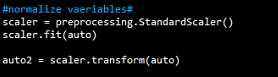


**Neural Network Goals**

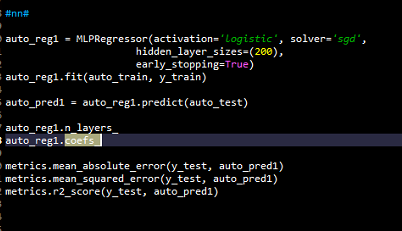
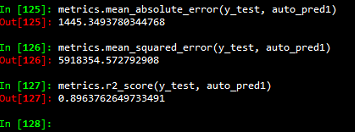
Neural Networks goals are similar to those of regression in that the independent variables are being multiplied by a weight and then summed to create a predicted price value. However, the neural network uses a complex series of interactions between various variables and processing nodes in the neural network to create these weights and values that are then summed.

**Neural Network Assumptions**

The main assumption of neural networks is that all of the variables need to be numeric and all of the variables need to be standardized to not throw off the weighting process. In order to process the data through the neural network numeric dummy variables were created for all of the categorical variables. Then all of the variables were standardized using the code below.

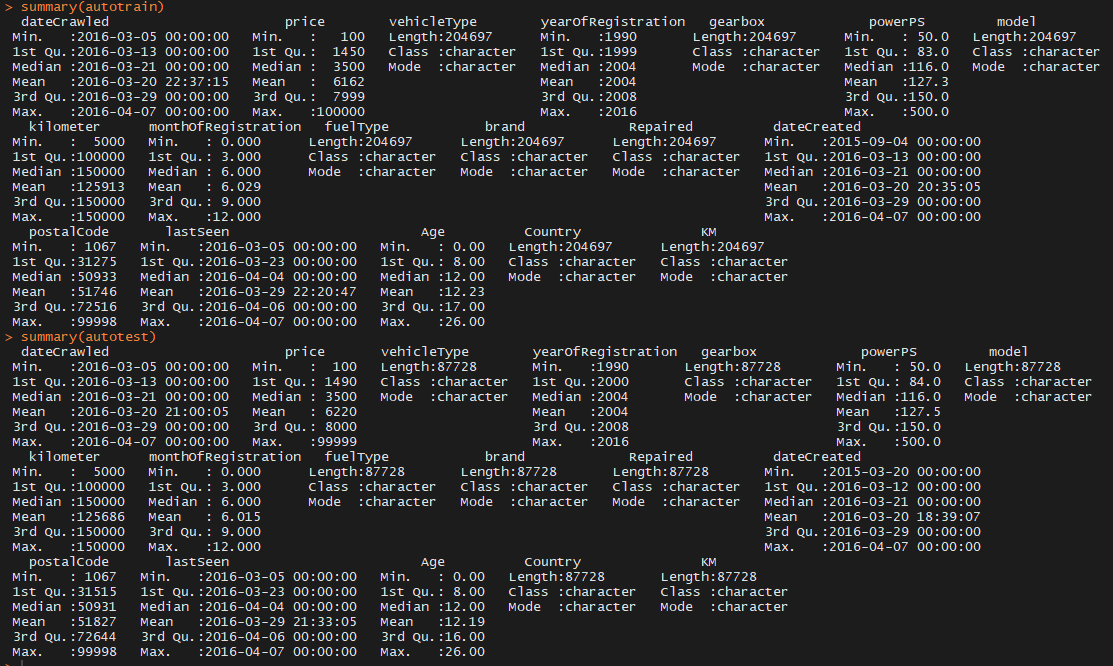


This code changed all of the variable values into standard normal Values which will all influence the neural networks weighting based solely on their correlation with price and not the extremeness of their values.

The Neural networks works similarly to regression except that it uses a much more complicated algorithm to determine the weights used for each independent variable. The benefit of neural network over regression is that the use of dummy variables is not needed. However leveling of numeric variables is necessary to allow the weighting algorithm to assign weights based upon a variables importance to the prediction of price vs the size of the independent variables values. After training the neural network predicting values is simply a matter of inputting variables into the trained neural network and receiving a prediction.  

**Splitting Data**

The data was split into two groups. The first group was training data and the second group was validation data. The data was split between the group with 70% going to the training data and 30% going to the validation data. The reason for assigning so much of the data to the training data set was that increasing the size of the training data set improves the training of the neural network. Thus we chose to use most of the data for training and less for validation. The data splitting was kept the same across all three data models.



After comparing the mean, medians and standard deviation of the two split datasets to each other there seems to be no significant difference in the makeup of the two split data groups. The data looks to have successfully been randomly split.

**Reference**

EBay. (n.d.). EBay-Kleinanzeigen. Retrieved March 15, 2018, from https://www.eBay-kleinanzeigen.de/

Leka, O. (2016, November 28). Used cars database. Retrieved February 14, 2018, from <https://www.kaggle.com/orgesleka/used-cars-database>

The Car Spy. (2011, December 10). 1996 Porsche 911 993 GT2 [Photograph].