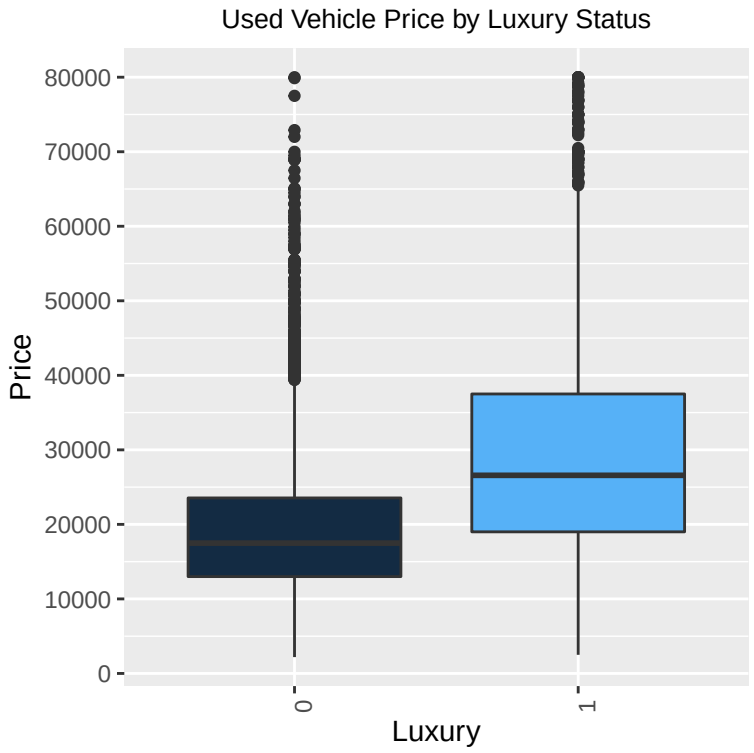
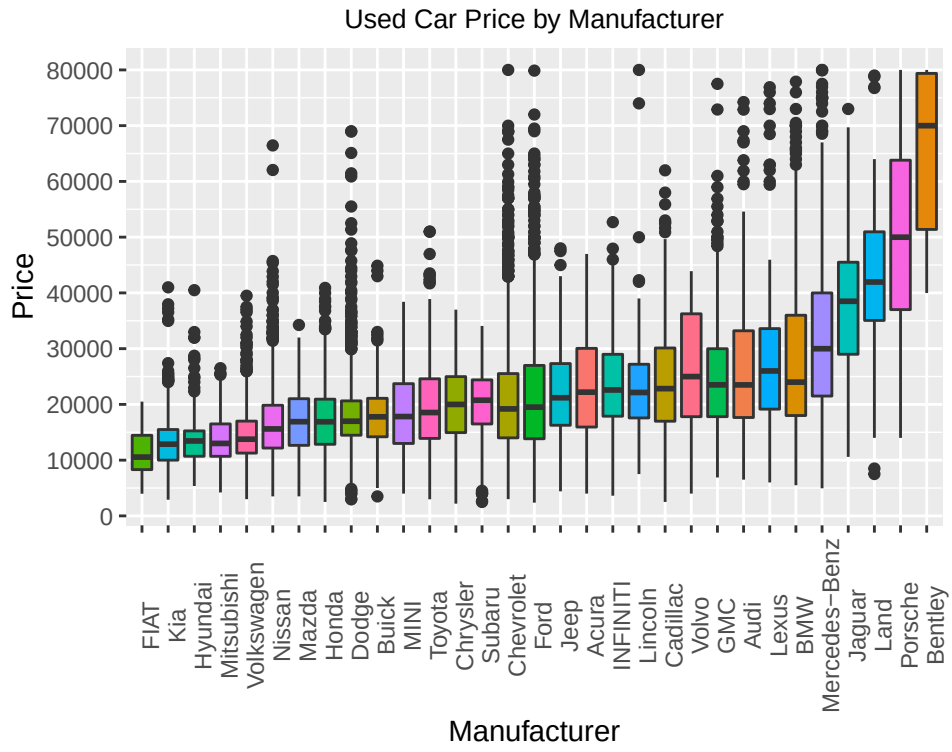
CarGurus Used Price Car Analysis

Introduction

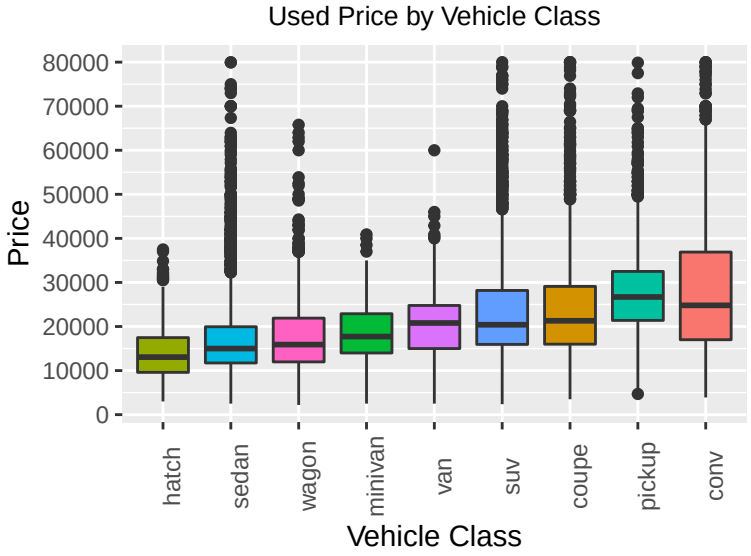
The goal off this project was to analyze used car listing from the website CarGurus in order to better understand how certain factors influence a used cars price. A secondary goal was to use the developed model in order to predict what a used car should be priced at or determine if a used car's price can be considered a 'Good Deal' or a 'Bad Deal'. The data attributes I was able to gather from Cargurus include, Price, Miles, Transmission Type (Automatic or Manual), Model, Make, Year, and Vehicle Class (Sedan, Mini-Van, Truck, etc.). I also included a field for whether or not the manufacturer was a luxury brand or not. I opted to only include cars from current manufacturers that have a relatively large volume of posting on CarGurus. Therefore manufactueres such as Mercury and Tesla were ommitted from the data. Furthermore, I decided to ommitt cars over $80,000 dollars and cars under $2000 dollars. Cars above $80,000 are luxury vehicles that may include many trim and powertrain options which help maintain the resale value regardless of some of the factors incorporated in the model. Cars under $2000 were ommitted because, in my experieince, essentially any car that can be driven off the lot can be sold for $1000 regardless of year, make, model etc. Lastly cars manufactured before 2006 were ommited because few cars older than this were posted on CarGurus. Lets now begin with some plots to visualise the data.



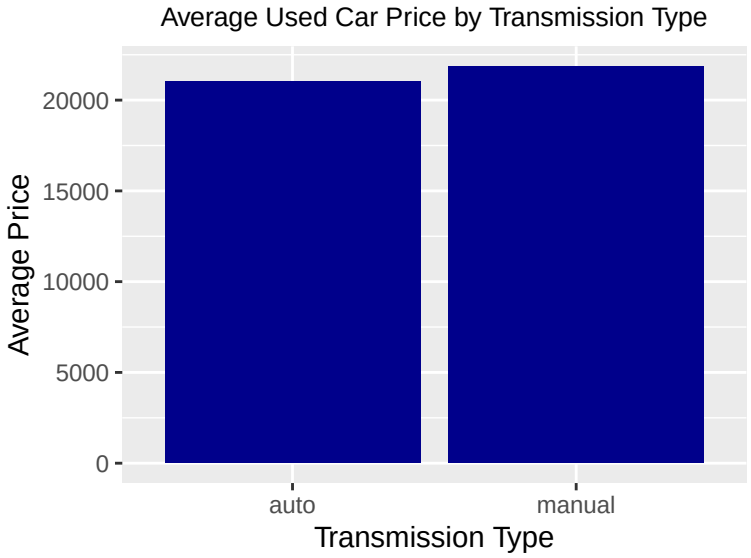
- The average used car price for luxury vehicles is over $10,000 more than that of non luxury vehicles.
- Both luxury and non-luxury categories are heavily skewed towards higher values. A log transformaton of price will be needed.



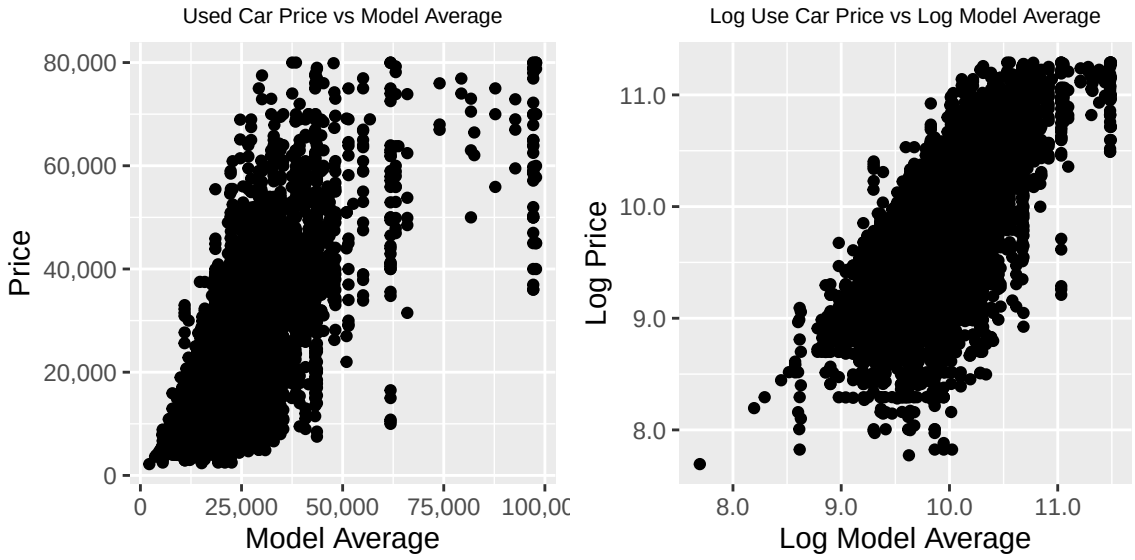
- As expected and somewhat obvious, manufacturers typically associated with budget options have a lower average price and more luxury brands have a higher average price.
- More interesting is the variablility in luxury prices. Audi, BMW, Infinitty, Lincoln, Volvo, Acura, Cadillac, and Lexus are typically considered luxury brnads however their average used prices are signifigantly lower than Porsche and Bentley's average used price. Mercedes, Jaguar and Land seem to occupy a middle ground between these two groups but are still over $10,000 more expense on average than Audi, BMW and the rest.
- GMC has an average used price similar to luxury brands. This is most likely because GMC mainly sells Pick-Up which are typiclly more expensive than other classes of vehicles.



- From the plot above we see how average used car price varies by vehicle class. Hatchbacks appear to be the budget class followed by sedans, wagons, and minivans. Intuitively, Pick-up trucks are relatively expensive, most likely due to their size, robust suspension, and large powertrain. Convertables are the most expensive presumaby because they are often luxury vehicles.

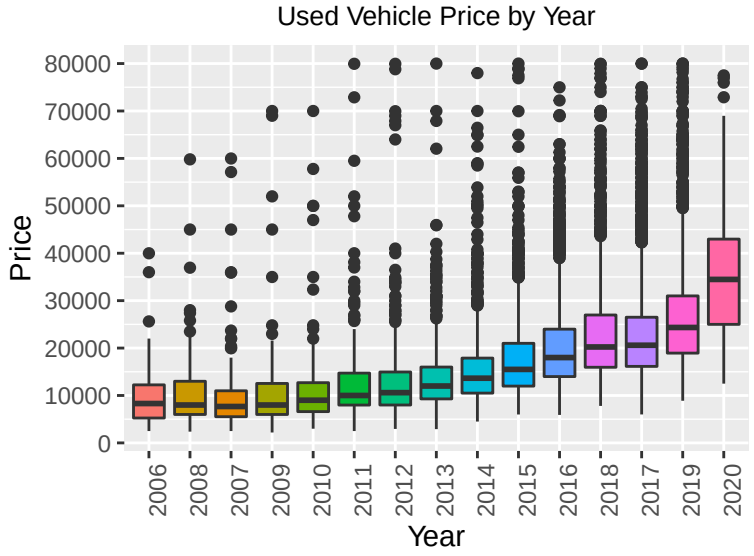


- The average used price for manual transmission cars is a couple thousand dollars more than that of automatic transmission cars. Traditionally however, manual transmission cars are cheaper than automatic cars. The difference observed here may be because not many manual transmission cars are sold anymore and the ones that are tend to be performance vehicles with higher end components. However, this difference does not appear signifigant and transmission type will not be included in the model.

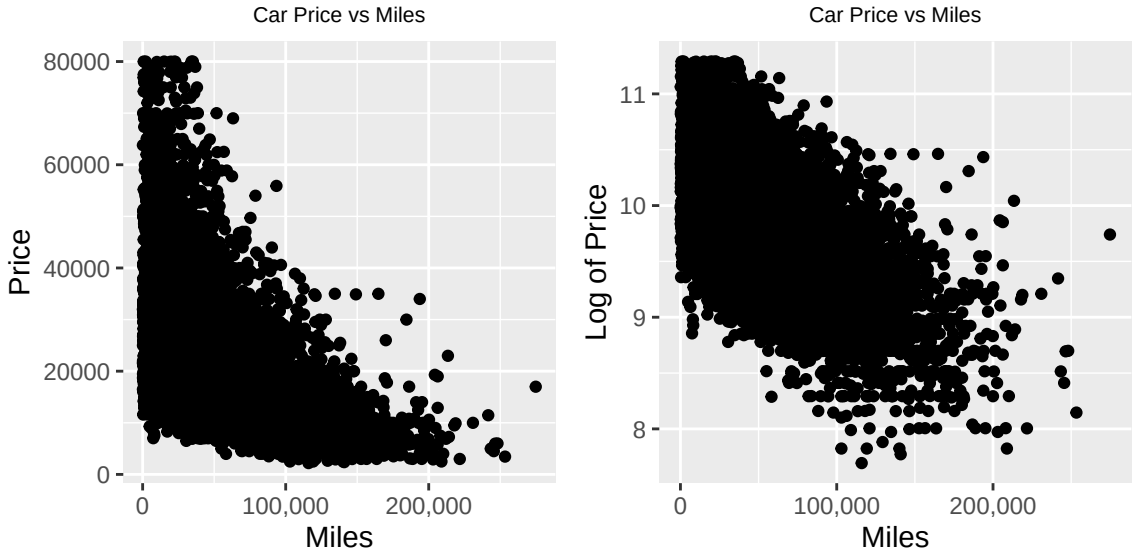


- To incorporate make and model into the model, I decided to target encode the variables because there are many different levels. Target, or mean encoding, creates a new variable out of the mean responses by level of the original variable. Given the first graph, the relationship between average price by

make and model and price does not appear linear. However taking the log of both price and the average price by make and model appears to linearize the relationship, as seen in the second graph.



- The relationship between year an average used is plotted above. As expected, the average values generally increase year after year. Where this is not the case may be the product of sampling error. Furthermore from 2006 to 2010, the average used vehicle price doesn't not increase much at all. After 2010, the increase in average price accelaerates until 2020. The difference between the 2019 and 2020 price is nearly $10,000. The drastic increase from 2019 to 2010 could be because 2020 cars have the most up to date features and may be considered new.

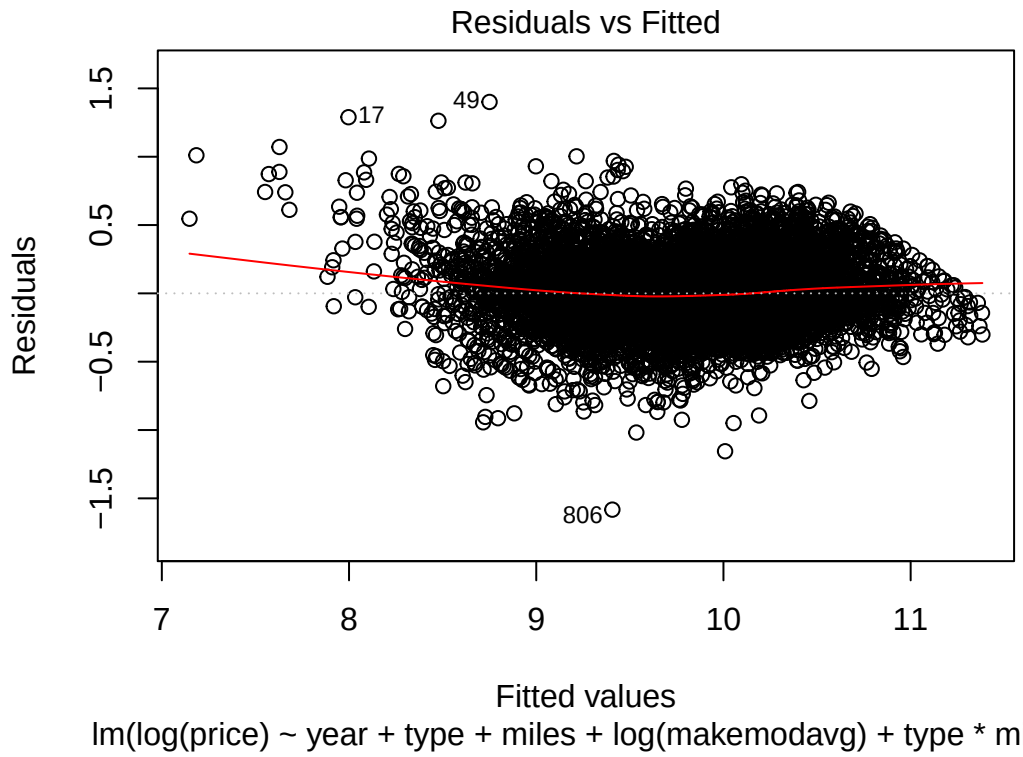


- There appears to be an inverse exponential relationship between price and miles. The price of a vehicle seems to decrease exponentially as the amount of miles increases. Taking the log of price linearizes this relationship.

Model Summary

The model is specified as log(price)~year+type+miles+log(makemodavg)+type*miles

```
##
## Call:
## lm(formula = log(price) ~ year + type + miles + log(makemodavg) +
##     type * miles, data = car)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.58253 -0.12439 -0.00658  0.11631  1.40067
##
## Coefficients:
##                       Estimate    Std. Error t value             Pr(>|t|)
## (Intercept)        1.5586783554  0.0632594499  24.639 < 0.0000000000000002 ***
## year2007          -0.0554747051  0.0311347815  -1.782              0.07481 .
## year2008           0.0203288329  0.0302643971   0.672              0.50178
## year2009           0.0721784420  0.0304879355   2.367              0.01792 *
## year2010           0.1164649462  0.0287784038   4.047   0.0000521464077910 ***
## year2011           0.2052787228  0.0270644095   7.585   0.0000000000000352 ***
## year2012           0.2872819630  0.0265062099  10.838 < 0.0000000000000002 ***
## year2013           0.3182314561  0.0261675418  12.161 < 0.0000000000000002 ***
## year2014           0.3862409538  0.0259905875  14.861 < 0.0000000000000002 ***
## year2015           0.4348291294  0.0258755022  16.805 < 0.0000000000000002 ***
## year2016           0.4747429411  0.0258443263  18.369 < 0.0000000000000002 ***
## year2017           0.5282185145  0.0258946088  20.399 < 0.0000000000000002 ***
## year2018           0.5475969810  0.0261866855  20.911 < 0.0000000000000002 ***
## year2019           0.6126775140  0.0263767640  23.228 < 0.0000000000000002 ***
## year2020           0.7234688484  0.0292382555  24.744 < 0.0000000000000002 ***
## typecoupe         -0.0989759666  0.0160425764  -6.170   0.0000000007021863 ***
## typehatch         -0.1875566825  0.0170180101 -11.021 < 0.0000000000000002 ***
## typeminivan       -0.0838892943  0.0148646919  -5.644   0.0000000169595669 ***
## typepickup        -0.1002498097  0.0145294621  -6.900   0.0000000000054163 ***
## typesedan         -0.1480265633  0.0129619474 -11.420 < 0.0000000000000002 ***
## typesuv           -0.1519758851  0.0123147716 -12.341 < 0.0000000000000002 ***
## typevan           -0.2015536310  0.0148446636 -13.578 < 0.0000000000000002 ***
## typewagon         -0.1039454476  0.0152885110  -6.799   0.0000000000109356 ***
## miles             -0.0000069435  0.0000002542 -27.315 < 0.0000000000000002 ***
## log(makemodavg)    0.8255656660  0.0054343550 151.916 < 0.0000000000000002 ***
## typecoupe:miles    0.0000006149  0.0000003320   1.852              0.06404 .
## typehatch:miles    0.0000010843  0.0000003417   3.173              0.00151 **
## typeminivan:miles  0.0000015097  0.0000002811   5.372   0.0000000792316807 ***
## typepickup:miles   0.0000036227  0.0000002754  13.155 < 0.0000000000000002 ***
## typesedan:miles    0.0000010635  0.0000002664   3.992   0.0000658811496497 ***
## typesuv:miles      0.0000022803  0.0000002589   8.808 < 0.0000000000000002 ***
## typevan:miles      0.0000033272  0.0000002855  11.653 < 0.0000000000000002 ***
## typewagon:miles    0.0000006401  0.0000003084   2.075              0.03798 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2052 on 15097 degrees of freedom
## Multiple R-squared:  0.8319, Adjusted R-squared:  0.8316
## F-statistic:  2335 on 32 and 15097 DF,  p-value: < 0.00000000000000022
```



- Overall the model was able to explain 83% of the used car price on CarGurus listings. However, since the errors are heteroskadestic, the beta p-values are no longer valid. For this reason, interpreting the effects and signifigance of the model will not lead to accurate conclusions. The model however, can still be used to make predictions. The average absolute error of predicted listing price on the test data set was $2850. While this may seem high, it is important to remember data such as trim package, condition, horsepower and vehicle accident history were not able to be collected of CarGurus.
- $8,282 is the model's prediciton for the listing price of my own personal car on CarGurus. I own a 2013 Ford Fusion with 100,000 miles and an automatic transmission. The model's prediction of appears consist with the search results on CarGurus.
- Despite not being very interpretable, the model is able to predict used car listing prices with impressive accuracy despite lacking detailed information about the vehicle. Further investigation and resolution of the heteroscedasticity would make the model and effects of the individual predictors more interpretable. One possible remedy is to implement weighted least squares.