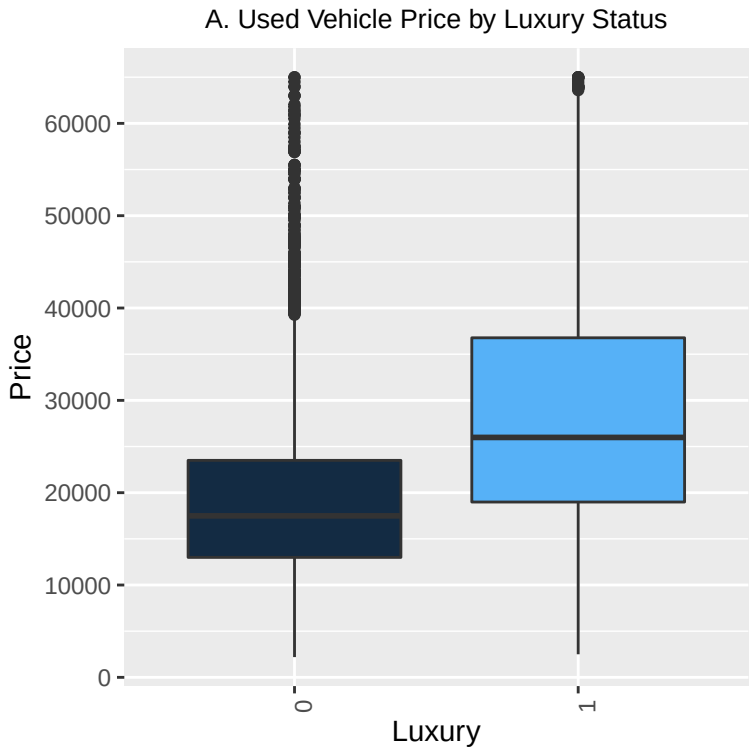
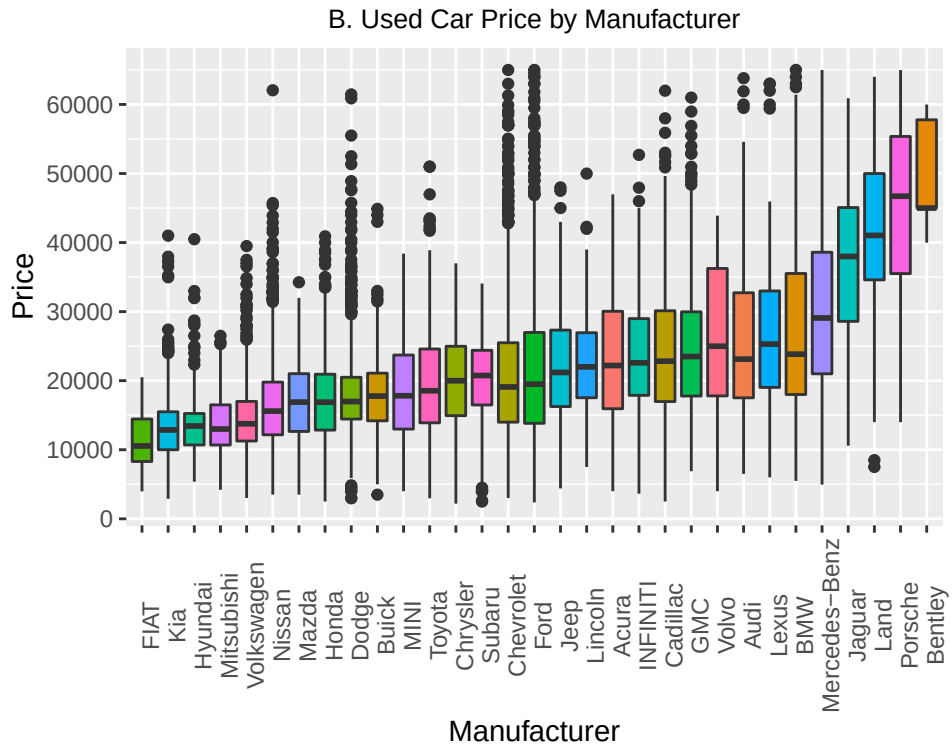
CarGurus Used Vehicle Price Analysis

Introduction

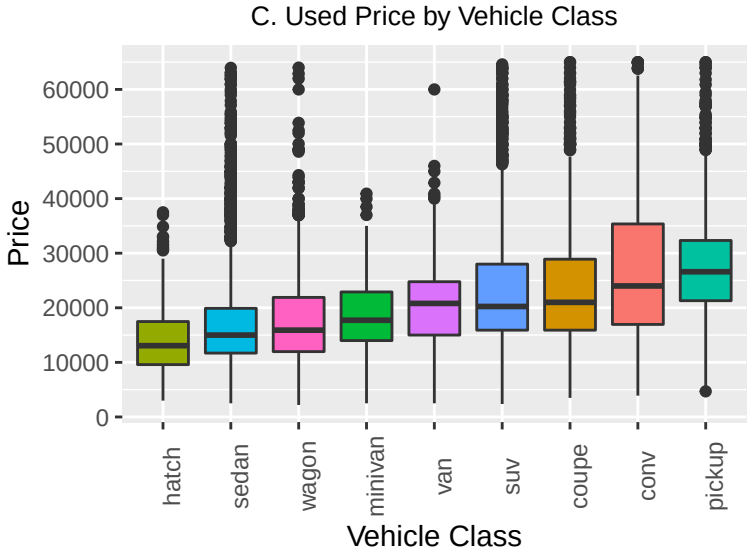
The goal off this project was to analyze used car listings from the website CarGurus in order to better understand how certain factors influence a used cars price. A secondary goal was to use the developed model in order to predict what a used car should be priced at or determine if a used car's price can be considered a 'Good Deal' or a 'Bad Deal'. The data attributes I was able to scrape from CarGurus using Python 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 omitted from the data. Furthermore, I decided to ommitt cars over $65,000 dollars and cars under $2000 dollars. Used cars above $65,000 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 omited because few cars older than this were posted on CarGurus. Lets now begin with some plots to visualise the data.



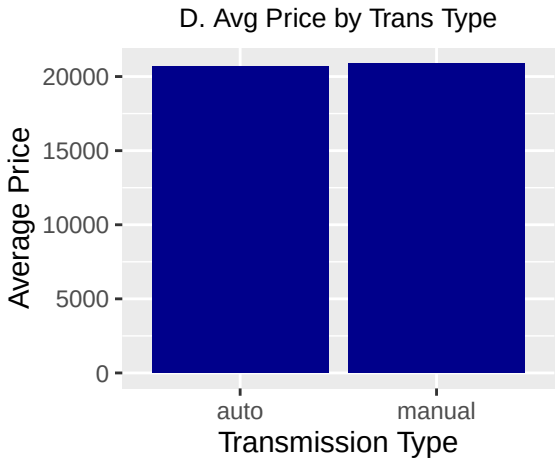
- The average used car price for luxury vehicles is over $10,000 more than that of non luxury vehicles.
- Both luxury and non-luxury categories are heavily skewed towards higher values. A log transformaton of price will be needed.



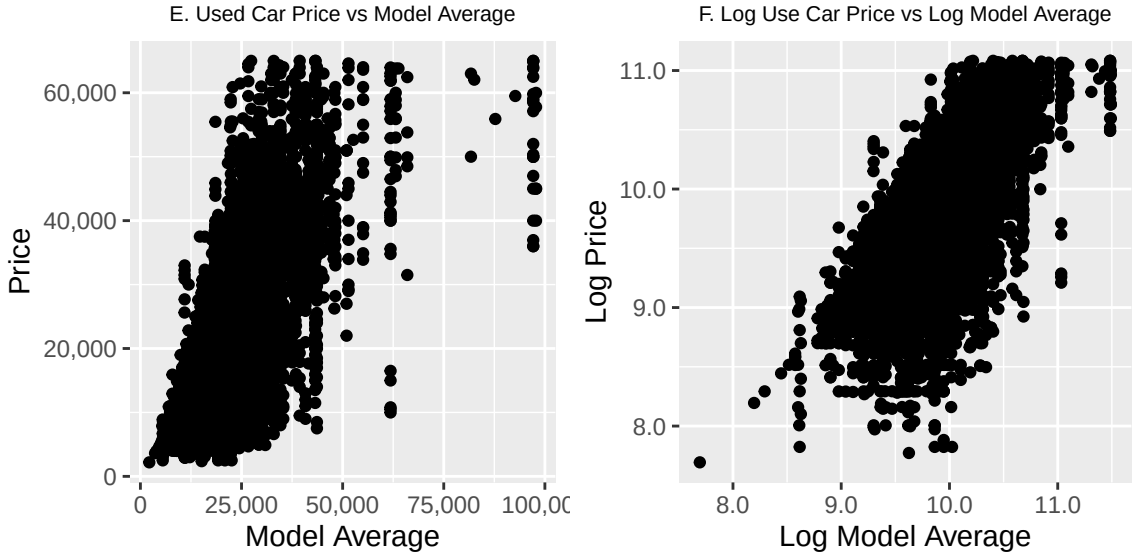
- As expected and somewhat obvious, manufacturers typically associated with budget options have a lower average price and more luxury brands have a higher average price.
- More interesting is the variablility in luxury prices. Audi, BMW, Infinitty, Lincoln, Volvo, Acura, Cadillac, and Lexus are typically 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 trucks, which are typicly more expensive than other classes of vehicles.



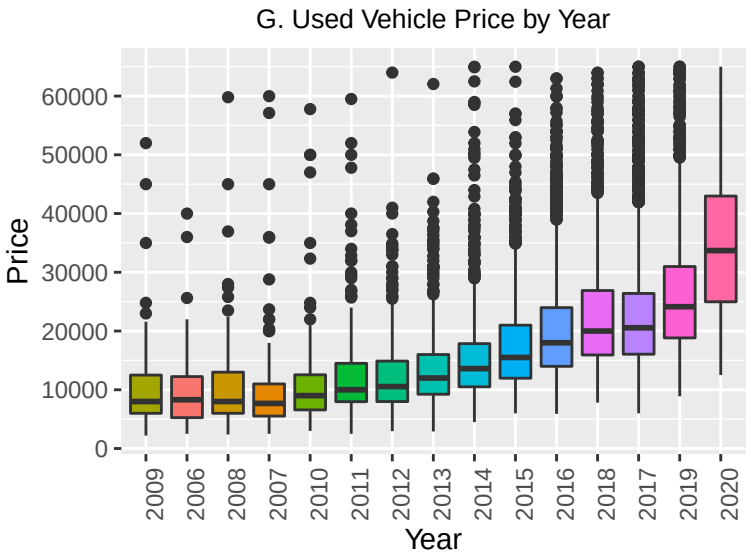
- From Figure C. above we see how average used car price varies by vehicle class. Hatchbacks appear to be the budget class followed by sedans, wagons, and minivans. Intuitively, Pick-up trucks are the most expensive, most likely due to their size, robust suspension, and large powertrain. Convertables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



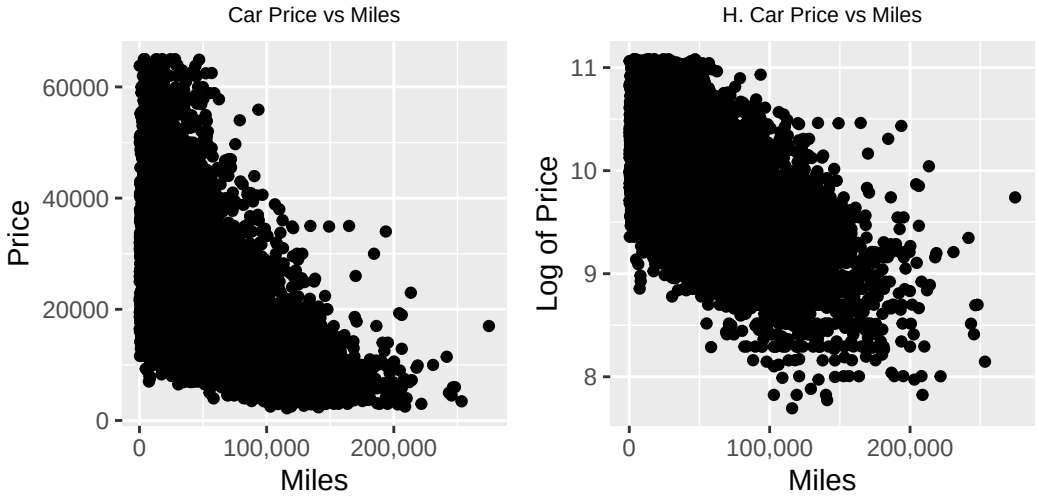
- From Figure D. the average price for manual transmission cars is a approximately two thousand dollars more than automatic transmission cars. Traditionally however, manual transmission cars are cheaper than automatic cars. The difference observed may be because not many manual transmission cars are sold anymore and the ones that are tend to be performance vehicles with higher end components. However, this difference does not appear large 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 Figure E., 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 Figure F.



- The relationship between model-year and average used is plotted in Figure G. 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 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 are considered 'new'.

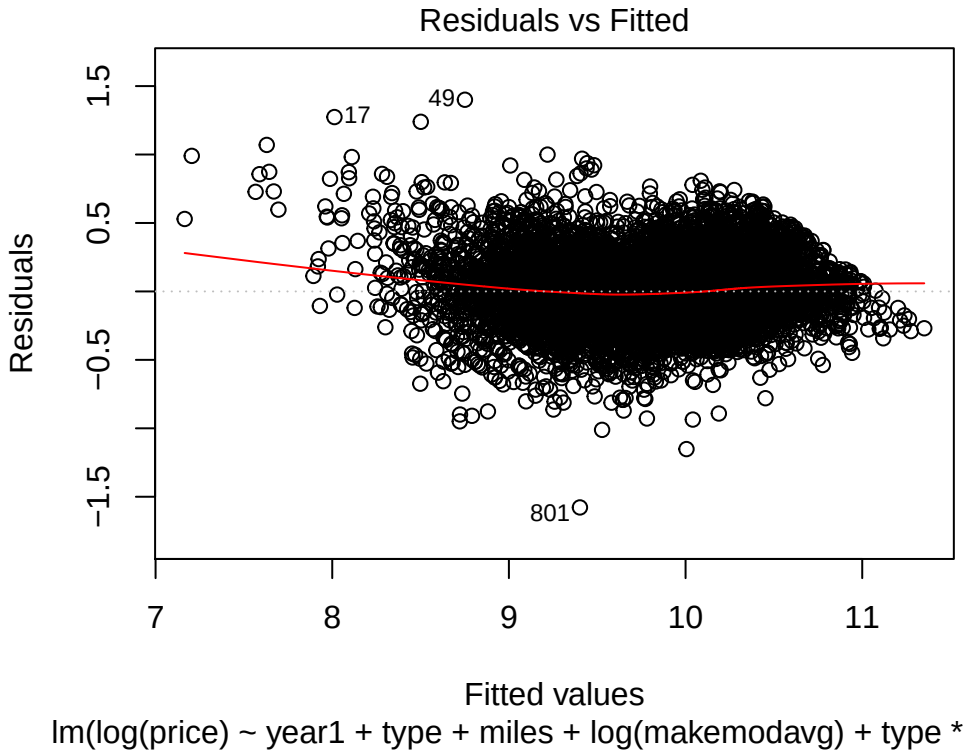


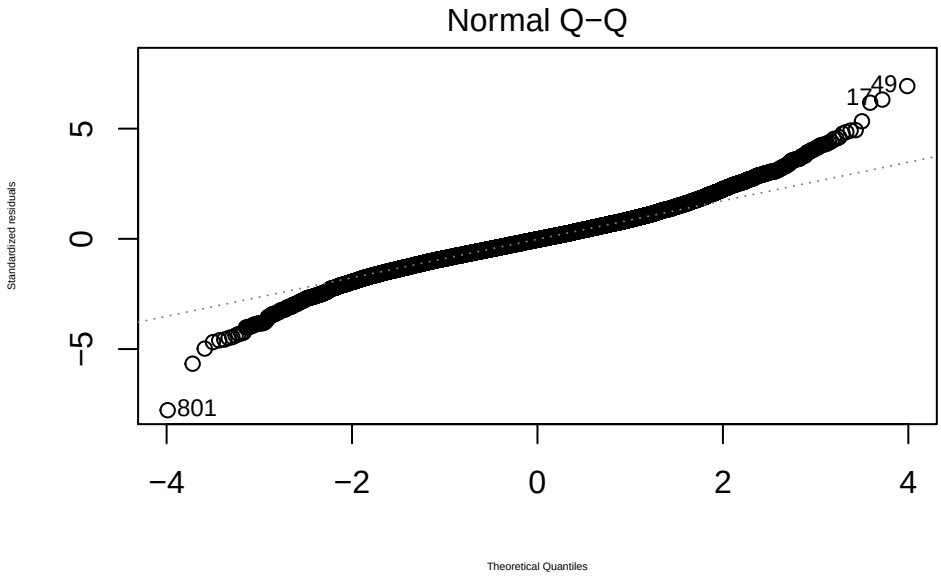
- There appears to be an exponential relationship between price and miles in Figure H. The price of a vehicle seems to decrease exponentially as the amount of miles increases. Taking the log of price linearizes this relationship.

Model Summary

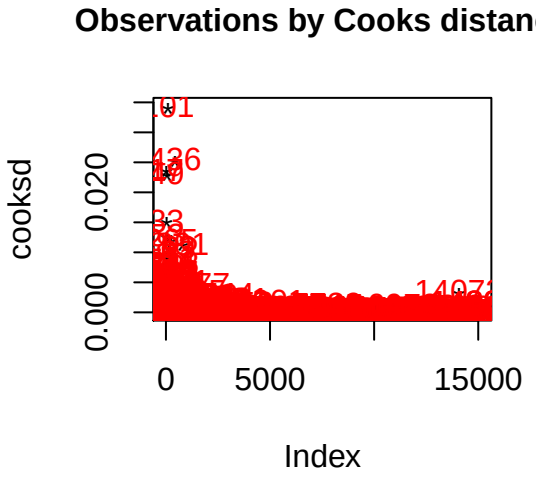
The model is specified as log(price)~year+type+miles+log(makemodavg)+type*miles. The model achieves a high R-Squared of 82%. The residuals plots below indicate some assumptions of linear regression have been violated. There appears to be some non-linearity structured in the residuals. The tails also appear heavy. However, since the number of observations is large, non-normal errors should not lead to biased

estimators. The most import assumption is that error variance is constant to ensure that P-Value estimates are accurate. Regardless, I would like to build a model with all regression errors assumptions satisfied.



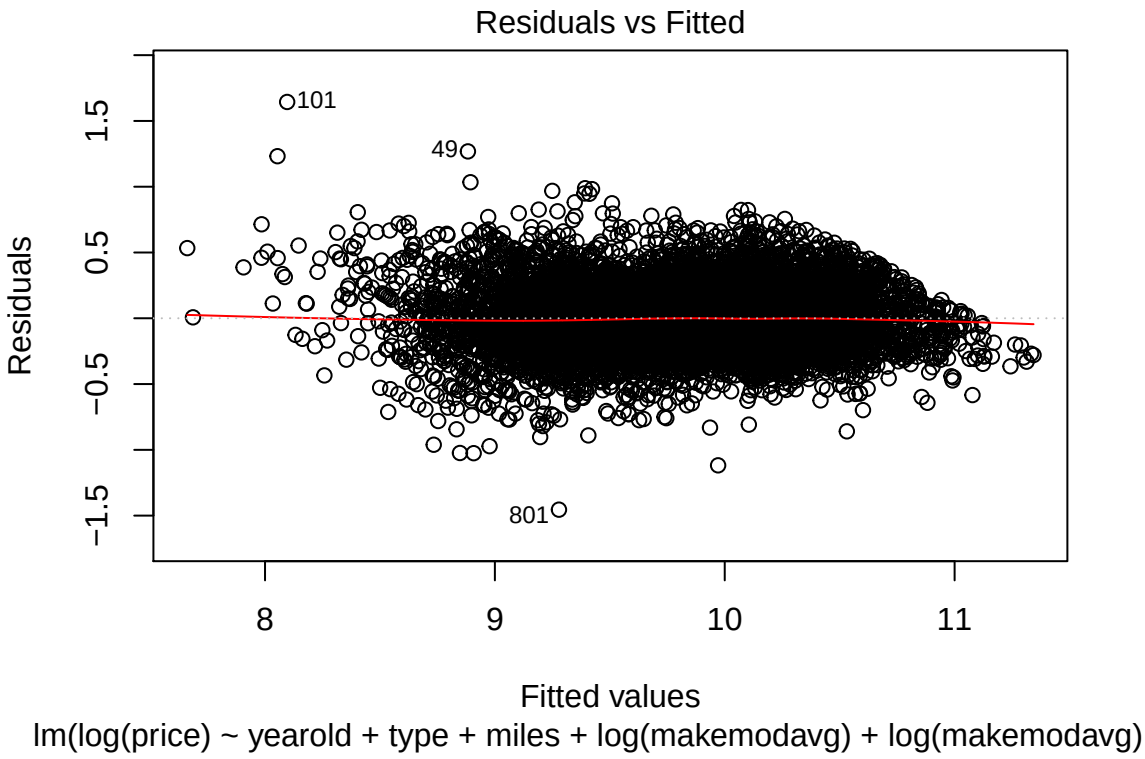


A plot of the observation index number vs the Cook's Distance reveals the most influential observations occur with lower indexes. Since the data is sorted by model-year, we can extraploate that older model-year vehicles are distorting the model. Instead of treating year as a factor, lets create a "Years Old" variable which can then be transformed or used to create interaction terms without heavily effecting the model degrees of freedom.



The new model is specified as the following: log(price)~yearold + type+miles + log(makemodavg) + (log(makemodavg) x yearold) + (yearold x miles). The R-Squared term has increased but more importantly the non linearity in the residuals has been remediated. Furthermore, every main and interaction effect is signifigant at the .05 level.

```
##                              Estimate Std. Error    t value  Pr(>|t|)
## (Intercept)                 0.4787746  0.1022339   4.683132 0.0000028
## yearold                     0.2558047  0.0152117  16.816346 0.0000000
## typecoupe                  -0.0794928  0.0096592  -8.229782 0.0000000
## typehatch                  -0.1345745  0.0104892 -12.829761 0.0000000
## typeminivan                -0.0322632  0.0088676  -3.638325 0.0002753
## typepickup                  0.0681960  0.0089852   7.589791 0.0000000
## typesedan                  -0.1025956  0.0079824 -12.852761 0.0000000
## typesuv                    -0.0618272  0.0075706  -8.166733 0.0000000
## typevan                    -0.0640848  0.0094238  -6.800350 0.0000000
## typewagon                  -0.0810373  0.0091013  -8.903930 0.0000000
## miles                      -0.0000042  0.0000001 -34.210905 0.0000000
## log(makemodavg)             0.9971784  0.0101780  97.974171 0.0000000
## yearold:log(makemodavg)    -0.0306859  0.0015230 -20.148542 0.0000000
## yearold:miles              -0.0000001  0.0000000  -6.715751 0.0000000
```



Model Interpretation

- Years Old: A one unit increase in years old predicts a 25.6% increase in vehicle price.
- Miles: Every additional mile is associated with a .00042% decrease in vehicle price.
- Model Average: A 1% increase in the average model price predicts a 1% increase vehicle price.
- Years Old x Miles: For every year older a vehicle is, the effect miles has on price decreases by .00001%
- Year Old X Model Average: For every year older a vehicle is, the effect model average has on price decreases by 3%.
- Coupe: Coupes are predicted to be 7.9% less expensive than convertibles.
- Hatch: Hatches are predicted to be 13.5% less expensive than convertibles.
- Minivans: Minivans are predicted to be 3.2% less expensive than convertibles.
- Pick-Ups: Pick Ups are predicted to be 6.8% more expensive than convertibles.
- Sedans: Sedans are predicted to be 10.3% less expensive than convertibles.
- SUV: SUVs are predicted to be 6.2% less expensive than converibles.
- Van: Vans are predicted to be 6.4% less expensive than convertibles.
- Wagon: Wagons are predicted to be 8% less expensive than convertibles.

Conclusion

- Overall the model was able to explain almost 83% of the used car price on CarGurus listings. Since the constant variance assumption of the error was satisfied, the P-Values for the coefficients are accurate and the model can be used for inferential purposes. The average absolute prediction error for the test data is $2,830. While this may seem high, it is important to remember data such as trim package, condition, horsepower and vehicle accident history were not available from CarGurus.

- $8,986 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 appears consistent with the search results on CarGurus for vehicles with the same specifications. It is important to note the model does not attempt to predict the true market value for each vehicle, rather the estimated price it would be listed for on CarGurus.
- The model is able to predict used car listing prices with impressive accuracy despite lacking detailed information about the vehicle. I feel the model has exhausted any possible variation that could reasonably explained by the collected attributes. Collecting data for other fields such as trim, horsepower, and accident history would likely greatly improve the model.