Section 1: Introduction

What matters to you most when you're looking for a car? Some people are most concerned with how it feels when they pull it off the lot, how it drives, how cool it looks. But some more savvy consumers are interested in a practical aspect - *how well does the car retain its value?* If they want to sell it in a few years, how much of their investment will they be able to recoup?

Other than direct customers, used car sellers would also be interested to know the answer to this question as it could help them better understand if a car is over or undervalued. In this project, we seek to examine what features predict how much a car will depreciate, and build a model to predict exactly that.

Section 2: Data

The dataset I started with was the Pakistan Used Cars dataset from Kaggle (https://www.kaggle.com/karimali/used-cars-data-pakistan). It contains information about used cars, and what price they were sold at. The columns include: (list columns here)

However, this dataset was missing some key information needed to calculate depreciation, which is what I wanted to predict. Most importantly, the price of the car when it was new.

The website that I used to scrape the missing information from was KBB.com (Kelley Blue Book), this had the majority of the cars that I was looking for but there were still some steps I had to take to ensure that dataset would be acceptable for later parts of the capstone project. After scraping the data from KBB.com its important to remember that the price that the cars were sold at in the original dataset from Kaggle are in Pakistan's currency (Pakistan Rupee), while the price listed on KBB.com is U.S dollars, so one of the first steps is to make a new column which has the price the car was sold at converted to dollars.

The first steps I took when cleaning the data was looking for the rows which had incomplete information, after I found out which ones weren't complete I looked at the information that was missing. If it was transaction type I kept the row as that wasn't information that I was interested in. My next step was to look at the car brands, there were some brands which were called classic and antiques, or other , or even had no brand in the column; These brands I deleted the rows all together as it just made any models I want to work on that much harder.

There were other brands which I could not use such as Changan and FAW as they aren't on kbb.com, so there is no information to scrape there. Another thing that I was aware of was that not every car goes by the same name in different regions, for example the Toyota Vitz also is called Yaris in the U.S. and the Toyota Prado is called the Lexus GX in the U.S. This meant that a decent part of the cleaning process was looking up the alternative names of the cars in the data set. As mentioned earlier sometimes the change was simply the model name while other times it was both the brand and model name that had to be changed.

Finally there were some outliers where there was no safety rating to be scrapped, in those cases I replace the NAN with zero so that they could still be used in the data story and the stats models later, the assumption why there might not have been any starts in the safety rating could be that the model was either to old so the information was no longer considered reliable or the car was to new and not enough reviews. Other outliers which had to be removed were models which were 2020 and 2019 models, simply because the exact date of the cars being sold is not given which makes figuring out how much the car depreciates by difficult and the last date the original dataset was updated was 2019. Had the exact date of when the cars were sold had been given then it would make the process of figuring the depreciation less prone to errors.

Calculating Compounded Depreciation Per Year

Before we begin exploring the data, it's important for us to define our target variable. Since we are hoping to discern what aspects of a car affect how much it depreciates, we'd like to know how much the cars in our dataset are depreciating on average every year rather than overall depreciation - since that is so affected by the number of years since purchase. Unfortunately, this isn't as simple as dividing depreciation by age of the vehicle, instead we used the following formula to calculate what we call Compounded Depreciation Per Year:

$$CDPY = -(-D + 1)^{(1/T)} - 1$$

D = depreciation overall = 1 - (used price/new price)

T = years since new

With this formula if you were to apply the CPDY every year for the length of the age of the car, you'd end up with the overall depreciation.

Section 3: Exploratory Data Analysis

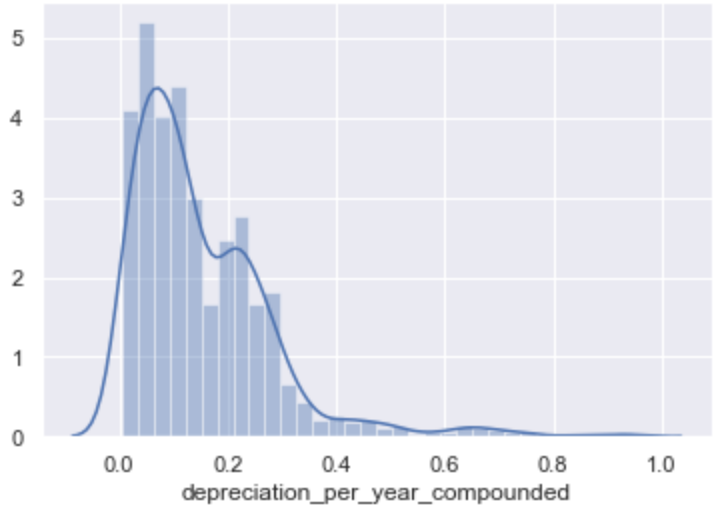


Fig 1. Distribution of depreciation per year across all vehicles

In Fig 1 we see the histogram for depreciation per year, the majority of the cars have a depreciation below .5, with only a very small outlier being above .5.

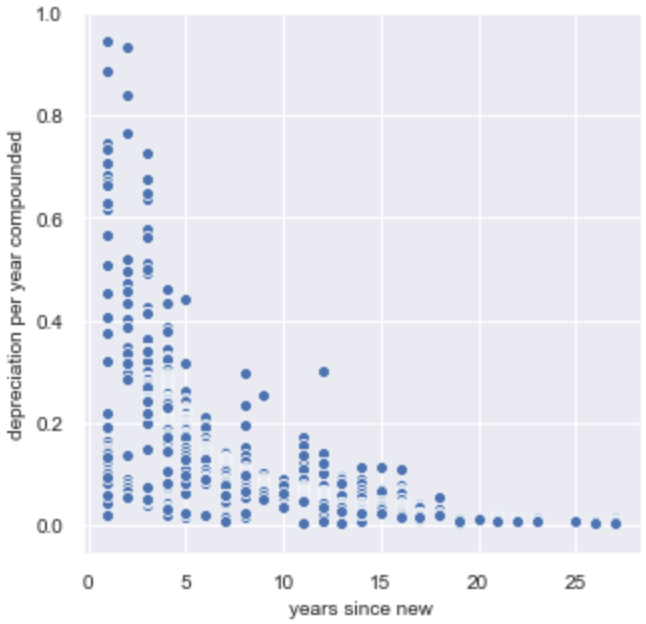


Fig 2. years since new vs. depreciation per year compounded

In Fig 2. we see that the cars depreciation per year compounded goes down the older the car is the less it will depreciate per year, for example there are cars which are 25 years old which will depreciate less than 10 percent a year, than there are cars which are less than 5 years old which depreciate more than 40 percent per year. As such we can say that the newer a car is, in general, the faster it depreciates.

The third figure is the correlation heatmap, this will help to determine the correlations between what characteristics influence the depreciation of the car. We will be looking at depreciation per year and the correlation to which characteristics as this will give us a somewhat better look at the depreciation, after all we went straight to depreciation it might not give us a clear of idea, as car which is 30 years most likely will have depreciate more than a car which is a couple of weeks old. The yoth streaming audio, bluetooth wireless, hands free, audio text messaging

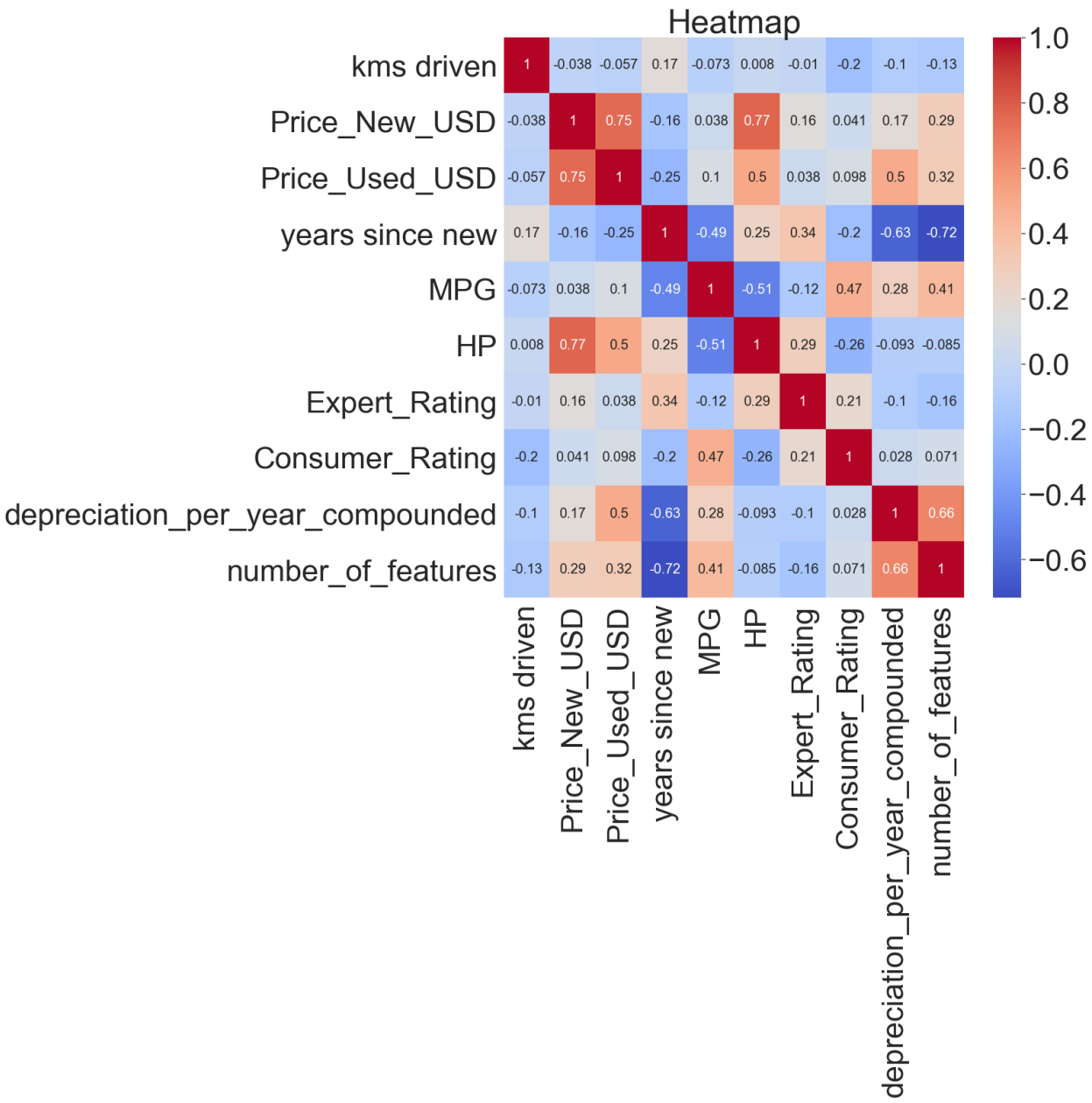


Fig 3. Correlation Heatmap

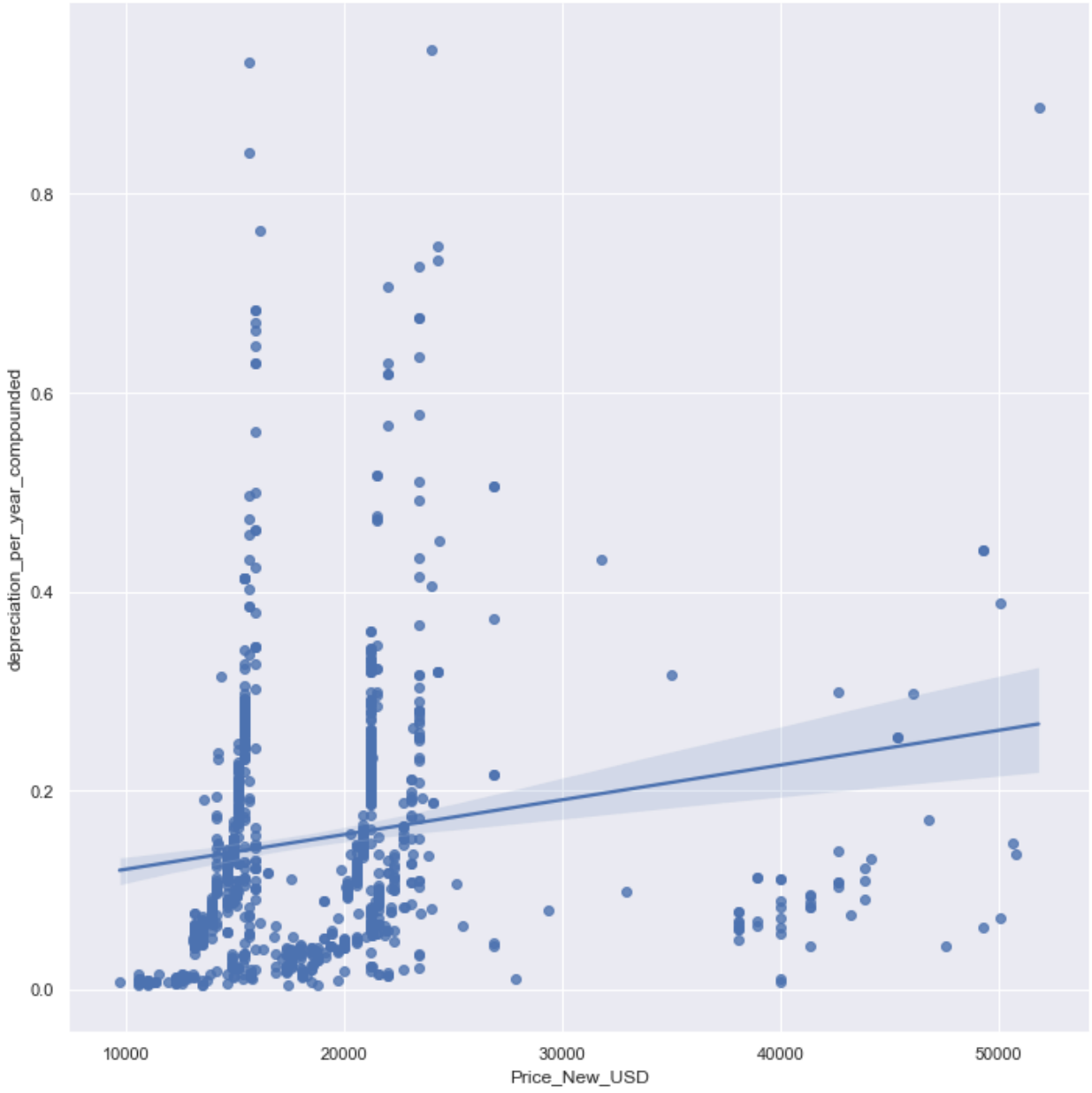


Fig 4. LM Plot Price of New Car

The fourth graph that we will look at is the Price in USD (dollars, the original currency in the dataframe was the Pakistan Rupee). From what we can see in this graph the more expensive a car is new , the faster it depreciates. To some degree this shouldn't be surprising as some cheaper

cars usually don't have the same quality as more expensive cars or they were not made to last over longer periods of time(they wear out faster). Both of these reasons can be the reason why the cheaper cars might depreciate at a faster rate.

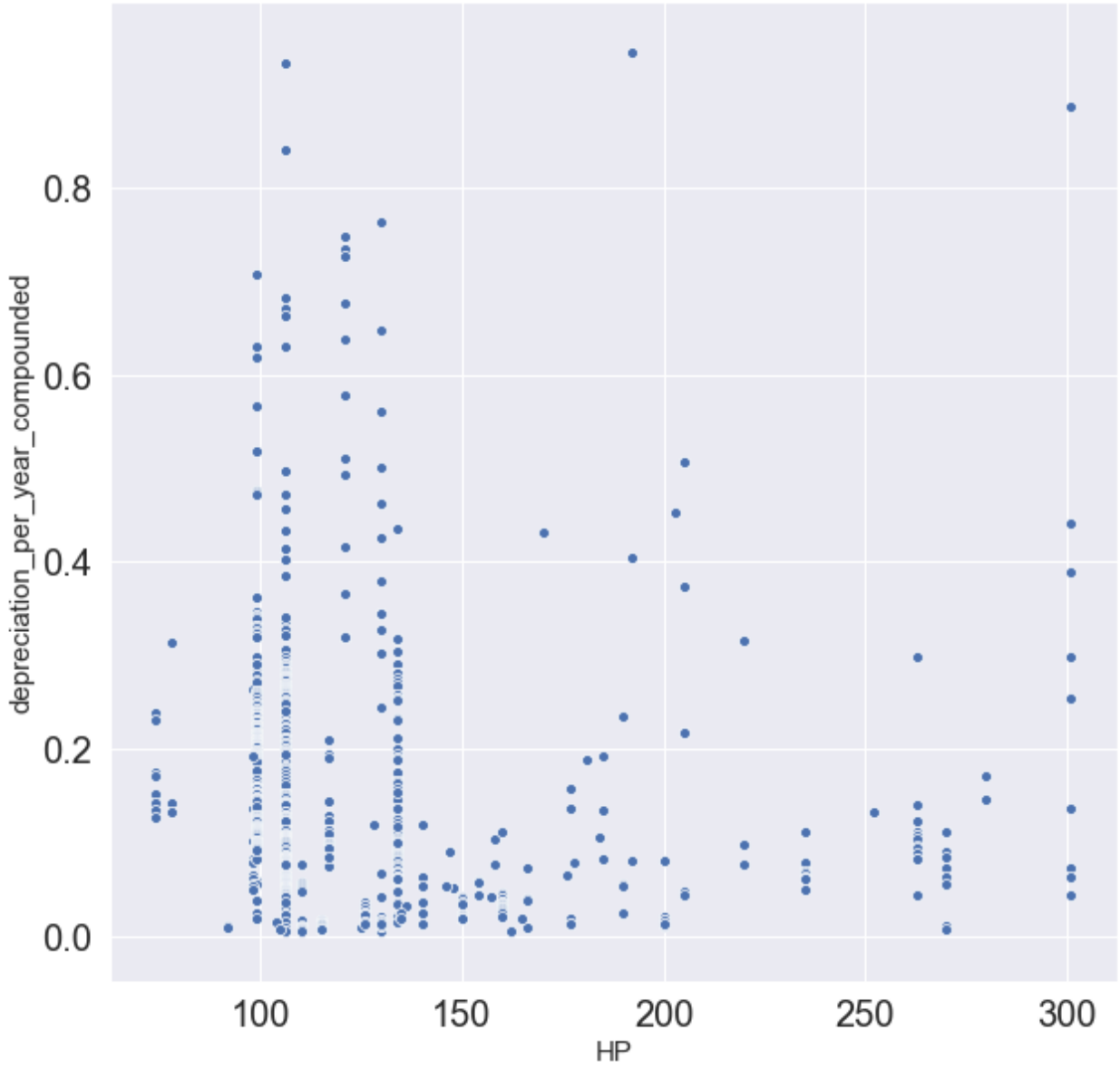


Fig 5. Horsepower vs. Depreciation Per Year Compounded

The next graph that we will look at is Horsepower (HP) in the x-axis and depreciation per year in the y-axis. From what we see here the less HP a car has the more it depreciates per year

($p < .001$), this shouldn't be much of a surprise as a more powerful car engine is generally more expensive.

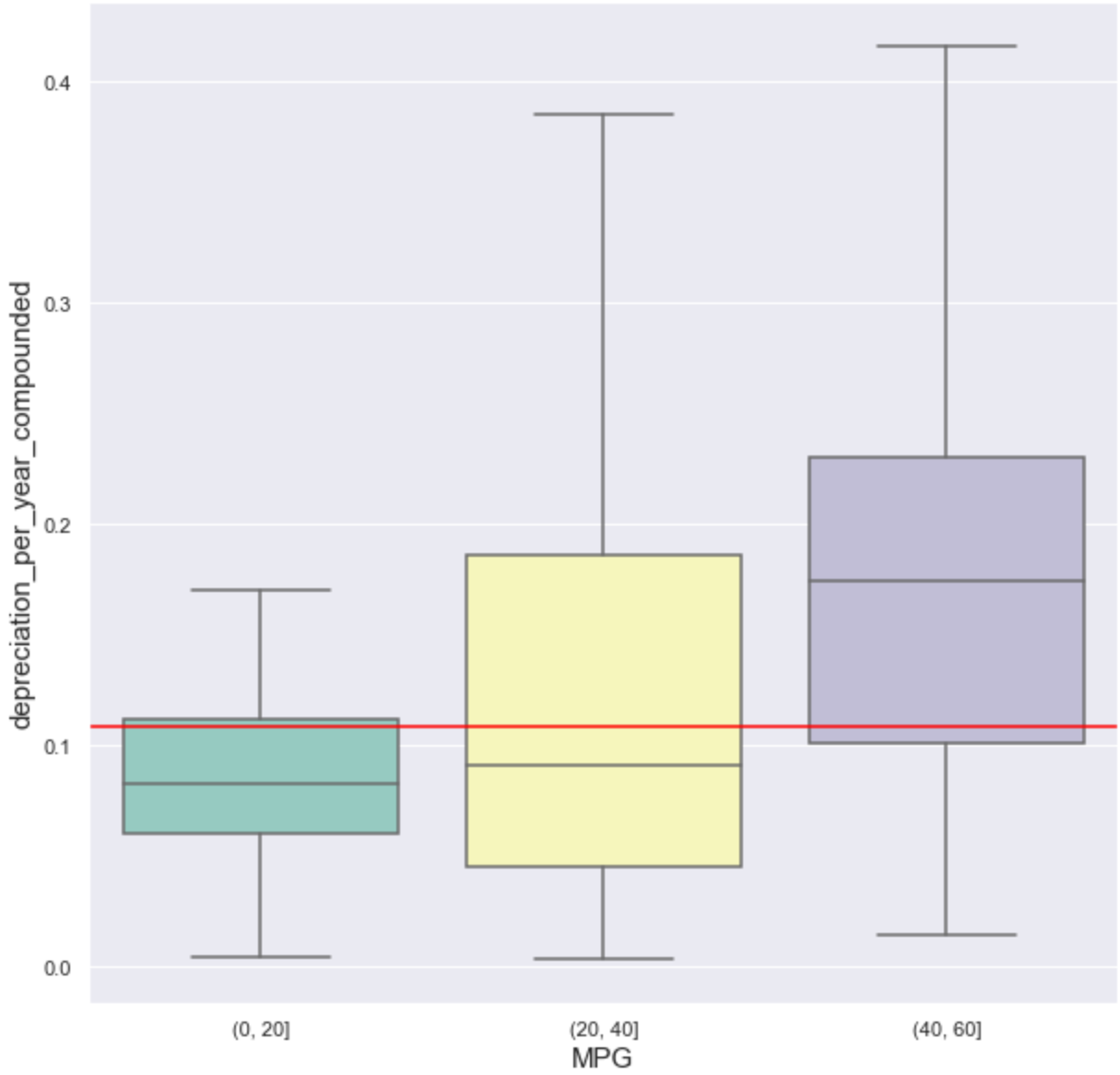


Fig 6. MIles Per Gallon vs. Depreciation Per Year Compounded

We can see that cars with a higher MPG depreciate at a faster rate. This could be for a number of reasons. It's possible that owners are more likely to put miles on a car with low mpg,

which would increase depreciation. It's also possible that the cars could simply be more expensive on average which we've seen can increase depreciation as well.

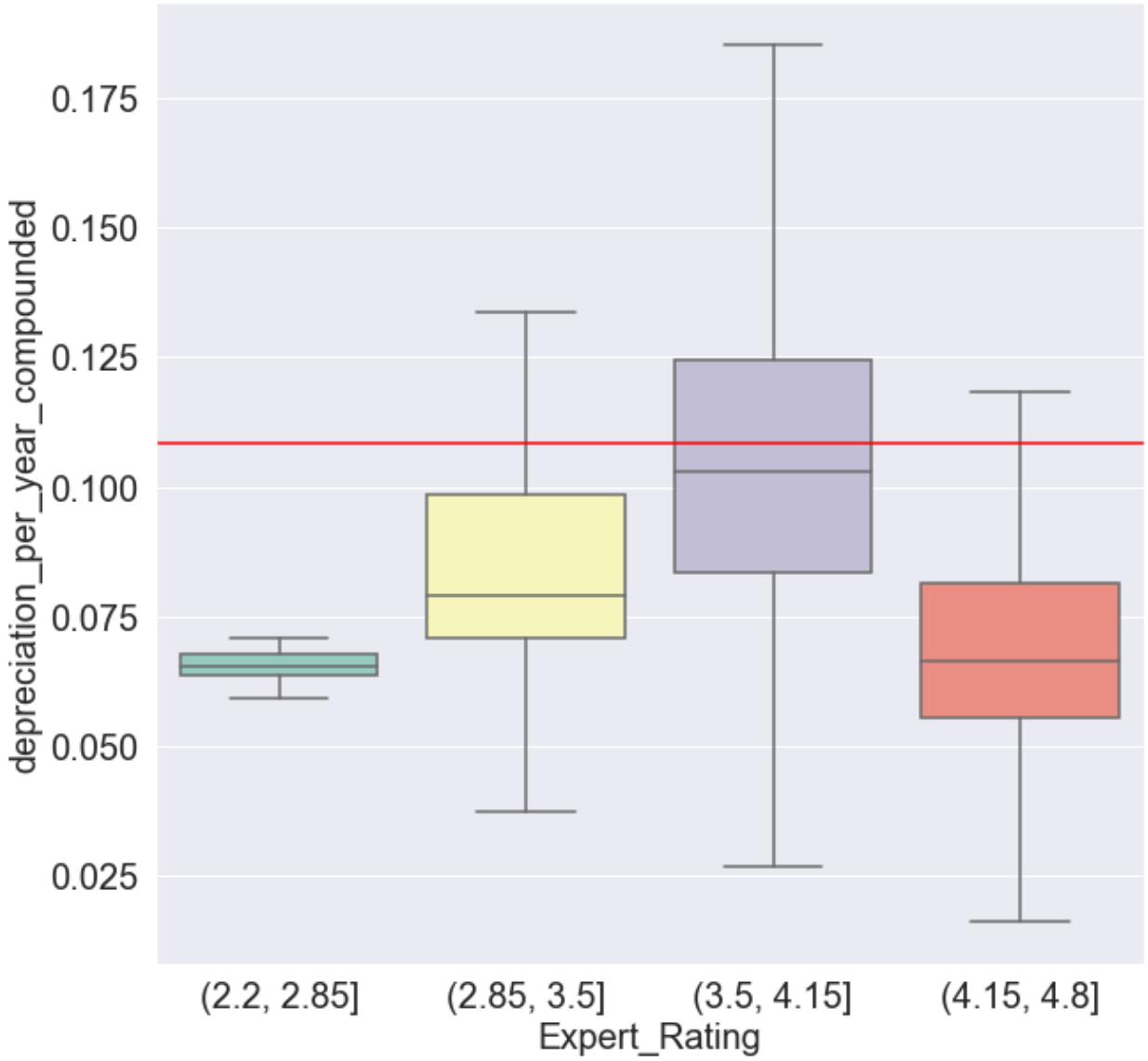


Fig 7. Expert Rating vs. Depreciation Per Year Compounded

For expert rating we see that the highest rated category of cars has the lowest depreciation. This is unexpected given that higher valued cars tend to depreciate more and indicates that the expert rating may be a good indicator of the actual longevity of the car. Of

course, this only occurs in the highest category, whereas in the lower ratings we still see that the higher the rating is, the greater depreciation is, as might be expected.

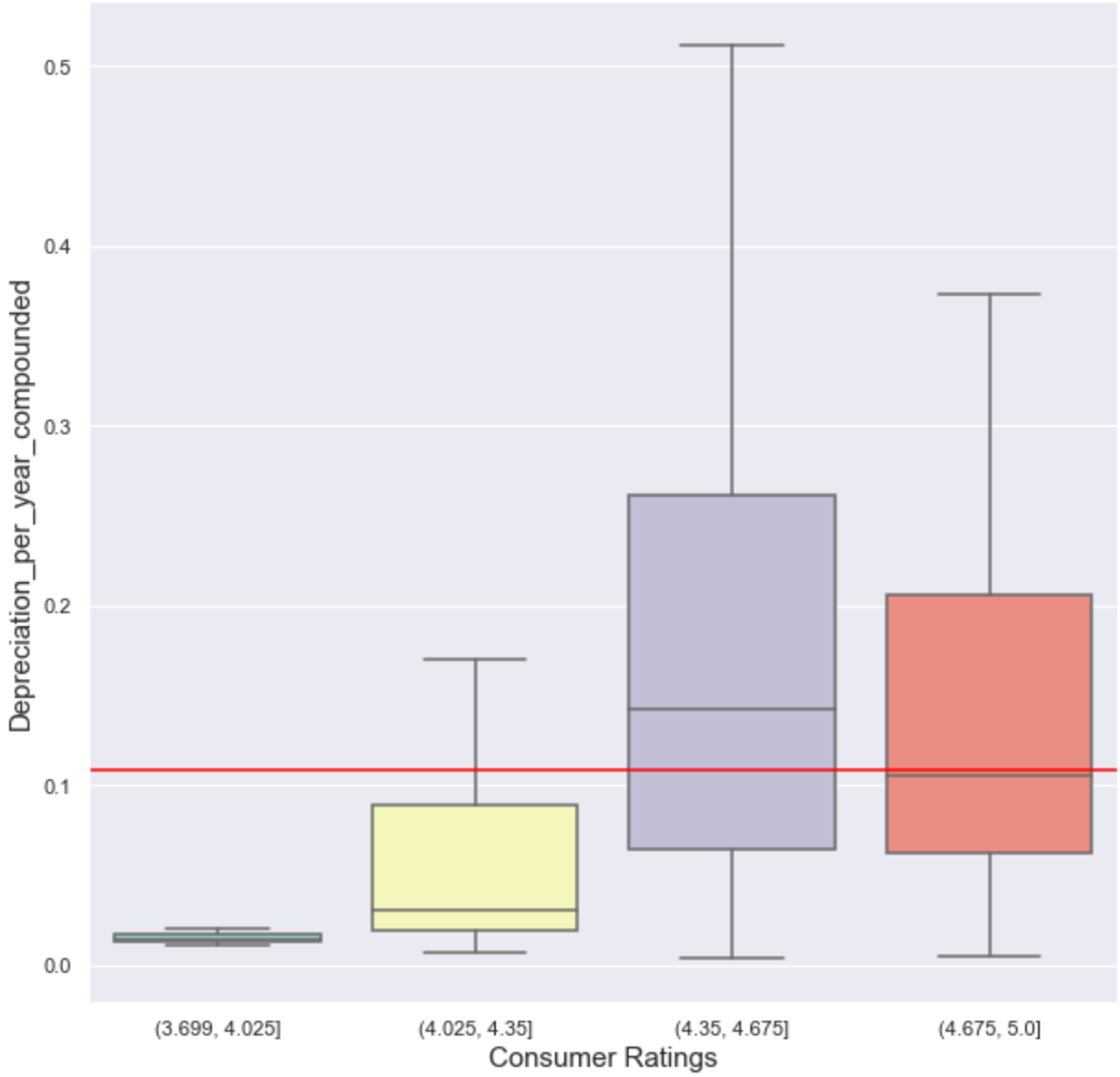


Fig 8. Consumer Ratings vs. Depreciation Per Year Compounded

Unlike with expert ratings, consumer ratings show a consistently faster depreciation with higher ratings even in the highest ratings category. In fact, the group with a 5 (a perfect score) has the most significant depreciation. Again we could attribute this to higher depreciation being associated with more valuable vehicles.

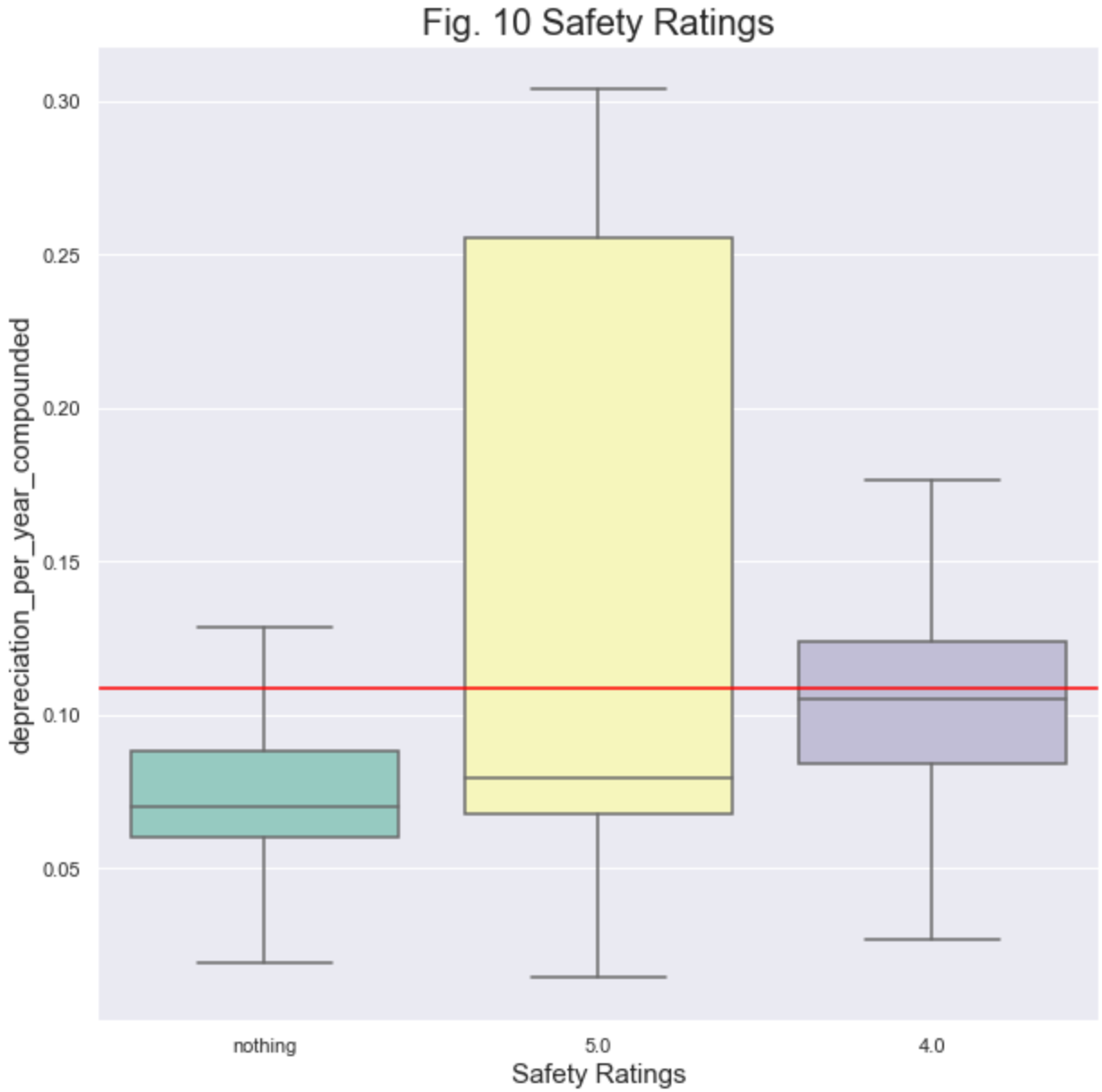


Fig 9. Safety Ratings vs. Depreciation Per Year Compounded

As with the previous two graphs, safety rating tells a story where the more valued group depreciates at a faster rate. Again as with the last graph, the group with a 5 depreciates faster than any other group. This could becauase thecar was initially valued so much more than the others and a large usage of the vehicle would naturally depreciate the car, while the cars with low ratings were already considered less than ideal so the usage might not have affected them as much. The statistics for these groups (miles per gallon, safety ratings, consumer ratings, and expert ratings) were fairly similar; the statistical test used to check the null hypothesis for these columns was Pearson correlation coefficient. The results from all of these tests were p-values

larger than .05, therefore we fail to reject the null hypotheses that cars with lower MPG depreciate more per year and that cars with higher safety, consumer, and expert ratings depreciate more per year.

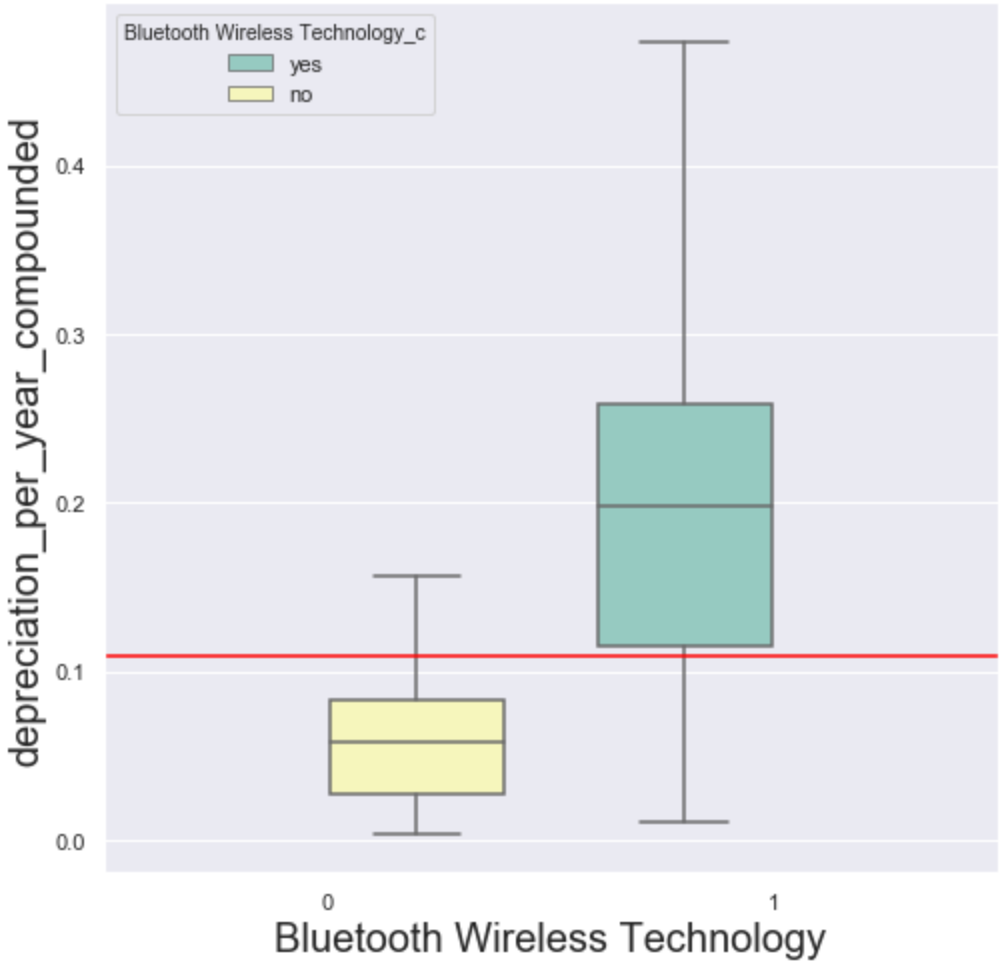


Fig 10. Bluetooth Wireless vs. Depreciation Per Year Compounded

A few features in particular had a high correlation with depreciation per year. Cars which have bluetooth streaming audio depreciate faster than those without it ($p < .001$) , the reason for this could be the bluetooth in cars will become out of date faster and while the cars that never had one to begin with could always have the latest bluetooth installed into the car. This feature could also be due to these cars being more expensive which is our highest predictor of depreciation. The next graph is about bluetooth wireless and it has a similar story to that of bluetooth wireless audio, so cars with bluetooth depreciate at a faster rate. The cars with bluetooth initially had a

higher value so they would depreciate faster over time, also the technology would become outdated rather quickly while either car could have the newest bluetooth technology installed. The technological features had a similar situation as the bluetooth wireless technology, cars with the technological feature depreciated at a faster rate per year. As for the statistical tests done the results for features like Bluetooth had p-values significantly lower than .05, in the Bluetooth Streaming Audio for example had a p-value of < .001. Therefore, we can reason that the null hypothesis is not statistically significant, the null hypothesis in this case is that cars with the features would depreciate more per year than cars without the feature.

This next graph is the first one which isn't about features but the car models, unfortunately there weren't many cars which had more than 20 instances in the dataframe, we were left with only 7 models. The model which saw the most depreciation was the Sentra, while it has the most depreciation it also had the least variation in terms of depreciation. the Accord was close second in terms of depreciation (sort descending for the graph)

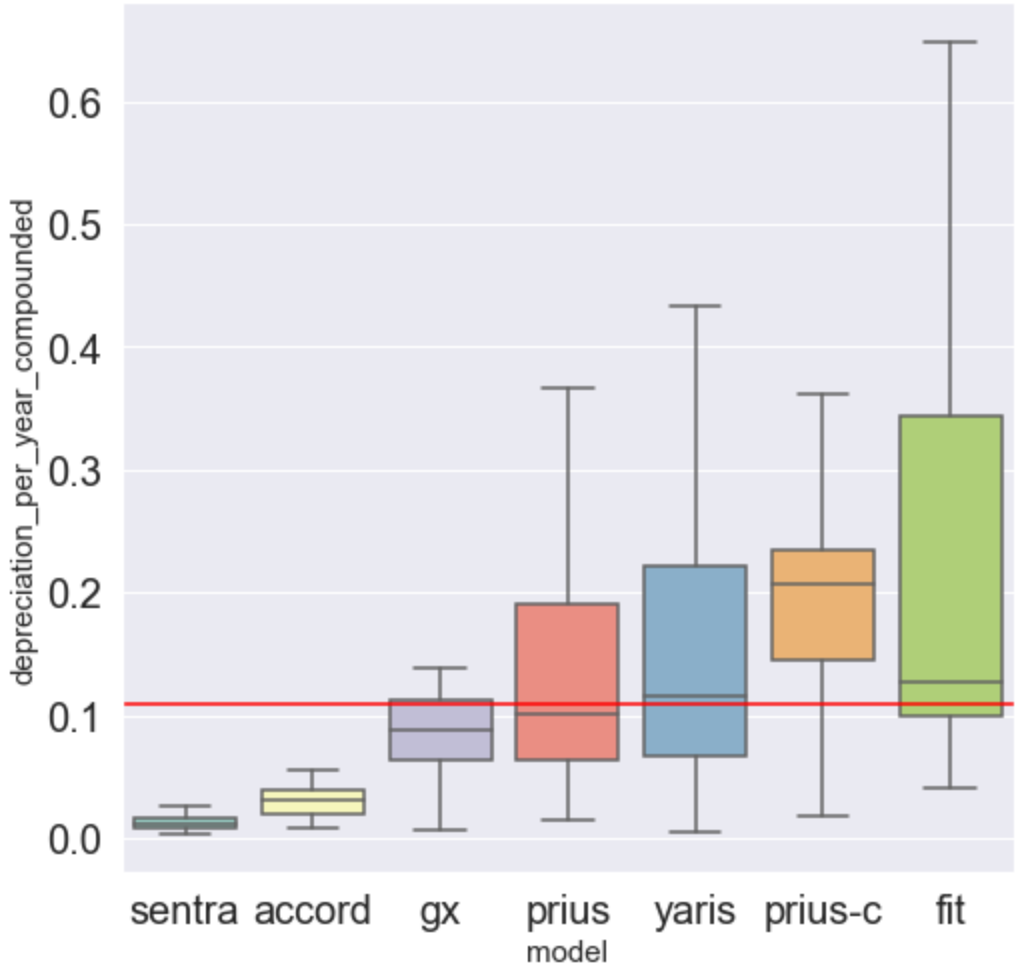


Fig 11. Model vs. Depreciation Per Year Compounded

From the graph at the bottom we see the linear relationship between depreciation and years since new. There is a positive correlation between the two showing that the older the car is the greater the depreciation will be.

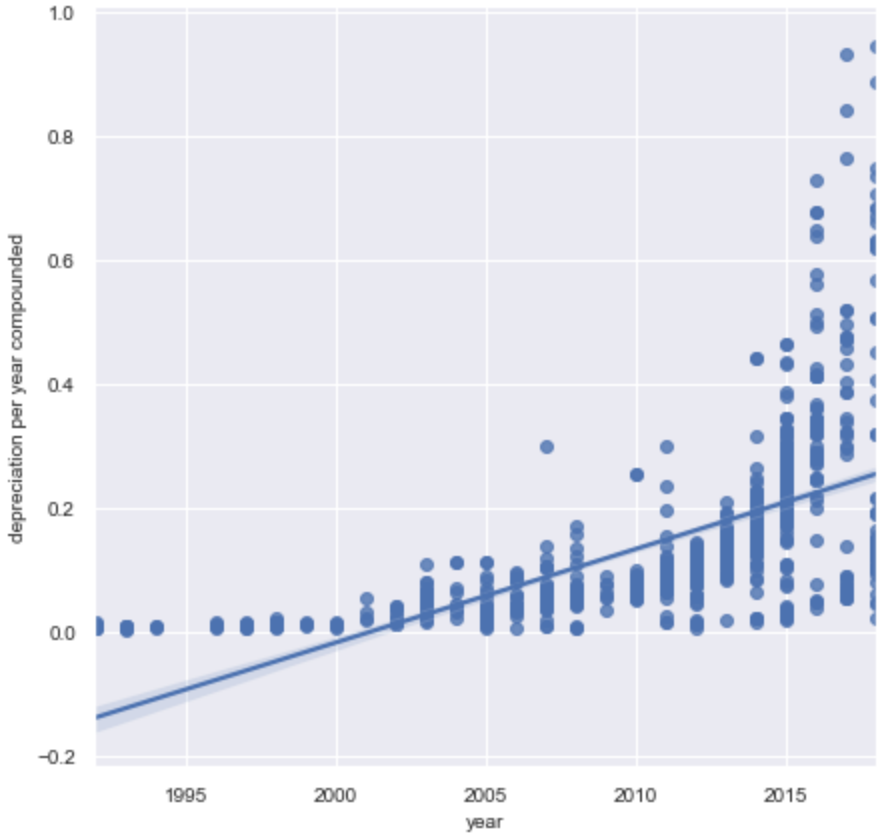


Fig 12. Years Since New vs. Depreciation Per Year Compounded

The final graph dealing with depreciation again shows that there is a positive correlation between depreciation and years since new, we see here that the cars have greater depreciation the older the car is.

Section 4: Machine Learning

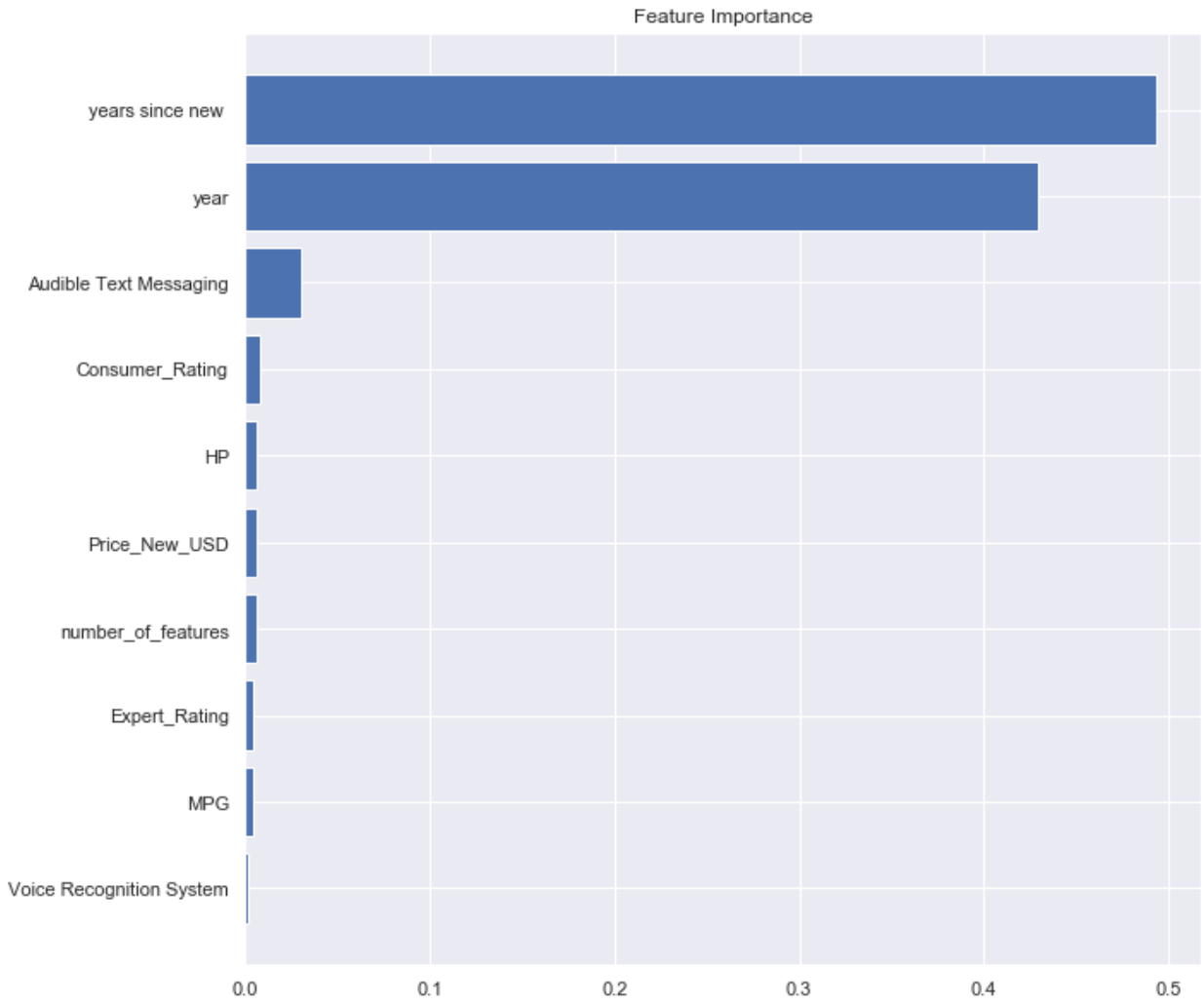
The next step was to build a ML model to predict the depreciation per year for a given used car as accurately as possible. I decided to try tree-based methods because they work well

when there is a non-linear relationship between the features and the target variable as is the case with a few of our features here (such as “years since new”). I used Gridsearch with cross-validation for each model to optimize the hyperparameters before determining the error against the test set. The results are below:

Comparing machine learning models

Model	Parameters	R square results
Random Forest	Max_depth= 20	0.693616599681034
	n_estimators= 50	
Gradient Boosting Regressor	Max_depth= 3	0.6797173263251267
	'n_estimators': 750	

At last, it was time to build a model to predict depreciation per year. I decided to try out two different tree-based models: Random Forest Regressor, and Gradient Boosting Regressor. For Gradient Boosting Regressor the parameters that we wanted to search were once again max_depth and n_estimators. For max-depth we tested 1, 3, 5, and 7 to see which would be the optimal number, and for n_estimators we tested 500, 750, and 1,000. The result of the grid search was max_depth = 7, and n_estimators = 1000; these were the optimal parameters. The r-squared score for the Gradient Boosting Regressor was 0.83619; this was the best r square score of the models and indicates it has the least error of the two models. With an RMSE of .069, my ultimate model is able to predict depreciation per year to within ~7% of the actual value of a used car on average.



Feature importance

If there is one characteristic of why the vehicles retain their value it is the age of the car when sold. Therefore, the largest piece of advice to anyone looking to buy a car with value-retention in mind is to buy an old used car. Looking at other features, we see Audible Text Messaging as the third most important with a number of features not too far behind, this indicates that cars with the newest technology is likely to depreciate more quickly, so value-conscious customers looking at either new or used cars should avoid these upgraded models. Lastly, it seems that while both consumer and expert rating are important features, consumer rating seems to matter more. So overall, value conscious customer should seek to buy a car that is:

- Ideally, an older used vehicle

- Lacking the newest technology
- Low horsepower
- Low in price when new
- Rated highly by consumers

Conclusion

Ultimately, using this data I was able to build a model that could predict depreciation per year on a new vehicle with 7% error on average. I found that, in general, the more desirable a car is (how much new technology, horsepower, how new it was) the faster it would depreciate. My model could be used both by buyers, who are looking for cars that will retain value over time, and also by companies like Kelly Blue Book looking to help customers do more thorough research into which cars to buy. In order to give a more complete picture of why cars retain value over time, next steps might include gathering data on how many kilometers the cars were driven, and what wear and tear there has been on the cars, as these are notable blindspots in the data used here.