Final Report

Ting Ho Marcus Cheung

April 29, 2024

Introduction

Most Canadian cities are very car-centric, which means that personal vehicles are necessities and core of day to day life. Ever since the global pandemic in 2020, the economy has been performing poorly throughout the continent: low economic growth, high inflation, and layoffs have been the norm for the past 4 years. Since cars are vital for day to day life here and are an expensive purchase, purchasing brand new vehicles have become out of reach for most people. Thus, the question arises:

How do we predict used car prices based on their year of production, body type, brand/model, engine displacement, drivetrain, and odometer kilometers?

Specifically, I want to find out the speed of depreciation on vehicles with time together with other factors. I believe answering this question will allow me to find the best time to purchase a used vehicle, find the most reliable (depreciation rate wise) vehicle, and evaluate the different aspects that affect prices of cars to make the best decisions financially. Lastly, by building a model of pricing with these factors, I can potentially help car sellers settle down with a price to sell their cars.

Methods

To answer this question, I will use a dataset I found of Kaggle: "Used cars listings for US & Canada" and use only the dataset for Canada. The link is attached to the bottom of this report. The dataset contains data of online listing of cars from that were manufactured in 1981 to 2022. As stated on Kaggle, "Individual listing records show year, make, model and trim, with VIN-level histories, showing the most recent time the car showed up online back to the earliest, with every change that occurred over that time". Furthermore, the dataset is collected across 8 years, where the latest was collected in 2022, which means that the first data was collected around 2014, which means that a car manufactured in 2013 might not be 9 years old when it was included in the dataset. Therefore, when looking at depreciation, I will strictly look at cars that appeared more than once. After exploring trends and building appropriate models using the data, I will parse through Autotrader to get come current listings to test the accuracy of the conclusion drawn from the Kaggle dataset.

Data exploration

I downloaded the data from kaggle and loaded it into an RMD file and stored it in the variable "canada_data". After running some summary functions such as ls, head, tail and nrow , I found that the earliest car ever produced was in 1981 and the latest was in 2022. The data has the 21 following variables, which are "body_type", "city", "drivetrain", "engine_block", "engine_size", "fuel_type", "id", "make", "miles", "model", "price", "seller_name", "state", "stock_no", "street", "transmission", "trim", vehicle_type", "vin", "year", and "zip". The variable names are very self explainatory except id, which is a unique identifier for the listing. I did find some problems, that I will need to address in the data cleaning and wrangling process. There are a total of 393603 rows in the data.

Data Cleaning and Wrangling

After looking at the older cars in the data, I noticed that a lot of them are not regular cars. For example, one of the oldest cars in the dataset is a Porsche Turbo, which is listed for 120,000 Canadian Dollars, which is very expensive. This is because it is considered a classic car, which means it is more of a collection than a means of transportation. Therefore, the first step I did was to exclusively select vehicles that were manufactured after the year 2005. Since I was performing modelling, I deducted 2005 from the year variable in the filtered dataset so it would start at 1 (2006) and end at 17 (2022) to avoid large numbers.

After that, I found some problems with the data as there are vehicles listed more than once in the dataset on the same year, same price, same mileage, but different dealers. For example, the 4th and 5th entry are both the same Acura NSX listed in Drive Autogroup and Acura Pickering. Therefore, I kept only 1 listing of the same vehicle (vin) at he same price and miles by using the distinct function.

Because I'm only interested in a selected few variables, I only kept the selected variables:

- vin
- price
- miles (is actually in kms since its Canada)
- year
- make
- model
- body_type
- drivetrain
- engine_size
- fuel_type

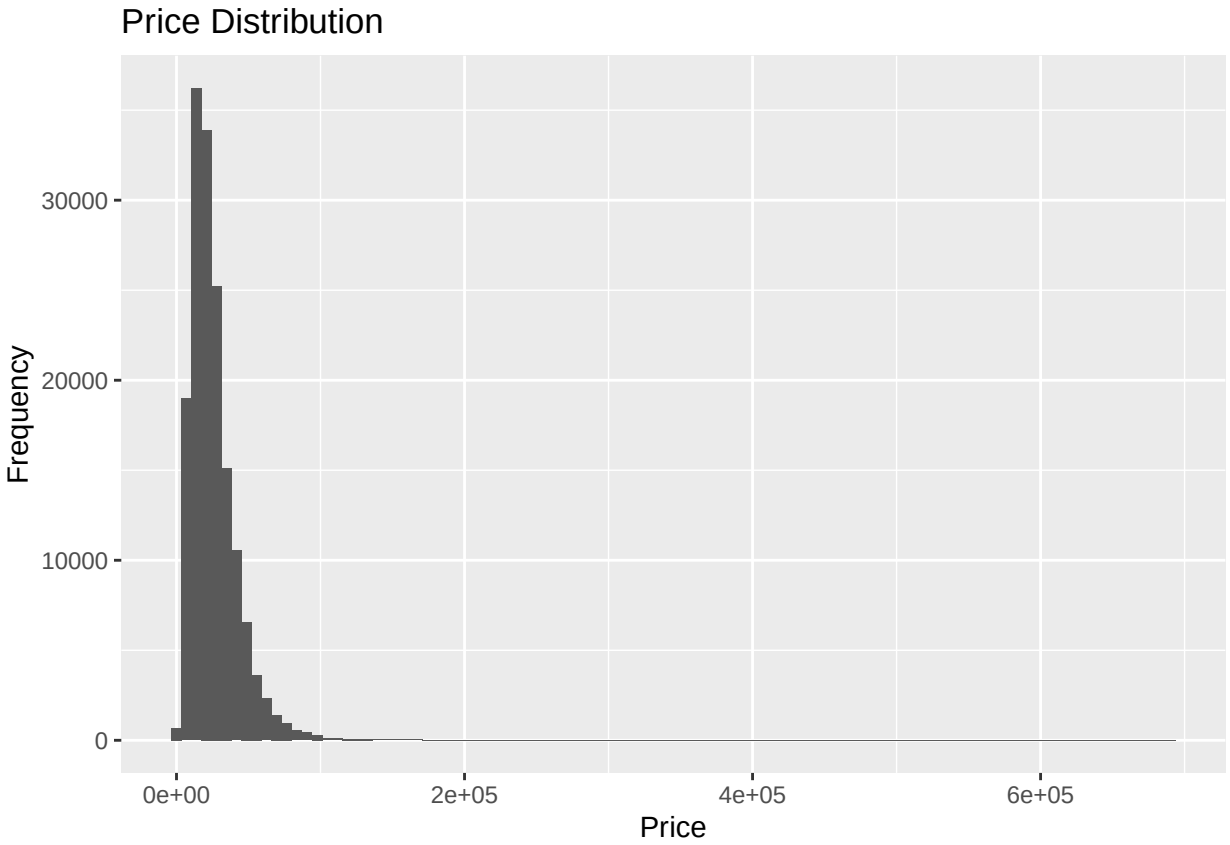
After performing further tests and checks for anomalies in the data, I found that many rows have missing data and they are represented as NA. Since I have large amounts of data, I removed all rows with NA. After this, there are only 157793 entries left.

Here is a sneak peak of the cleaned data where I removed VIN since it takes up too much space:

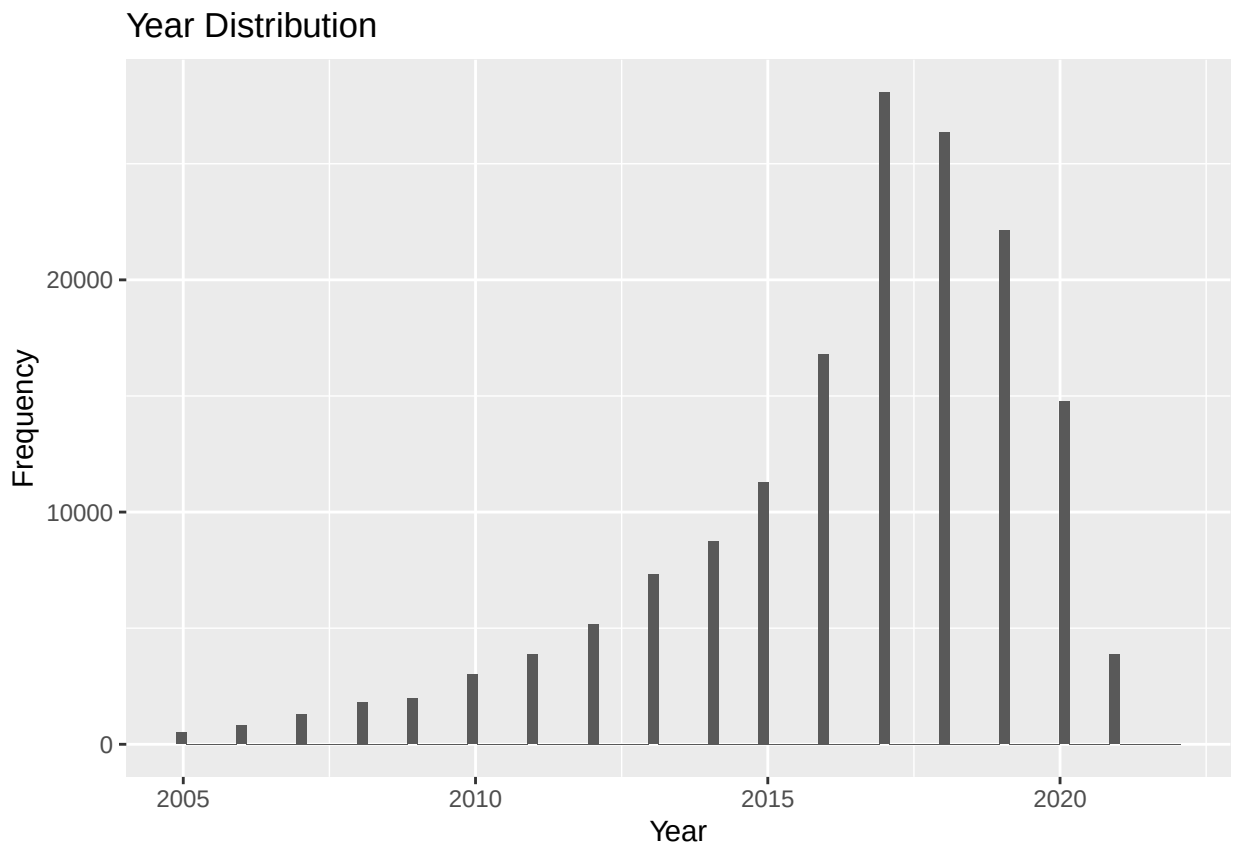
price	miles	year	make	model	body_type	drivetrain	engine_size	fuel_type
179999	9966	12	Acura	NSX	Coupe	4WD	3.5	Electric / Premium Unleaded
179995	5988	12	Acura	NSX	Coupe	4WD	3.5	Electric / Premium Unleaded
168528	24242	12	Acura	NSX	Coupe	4WD	3.5	Electric / Premium Unleaded
220000	6637	15	Acura	NSX	Coupe	4WD	3.5	Electric / Premium Unleaded
155771	18281	12	Acura	NSX	Coupe	4WD	3.5	Electric / Premium Unleaded
169998	17207	12	Acura	NSX	Coupe	4WD	3.5	Electric / Premium Unleaded

Data distribution

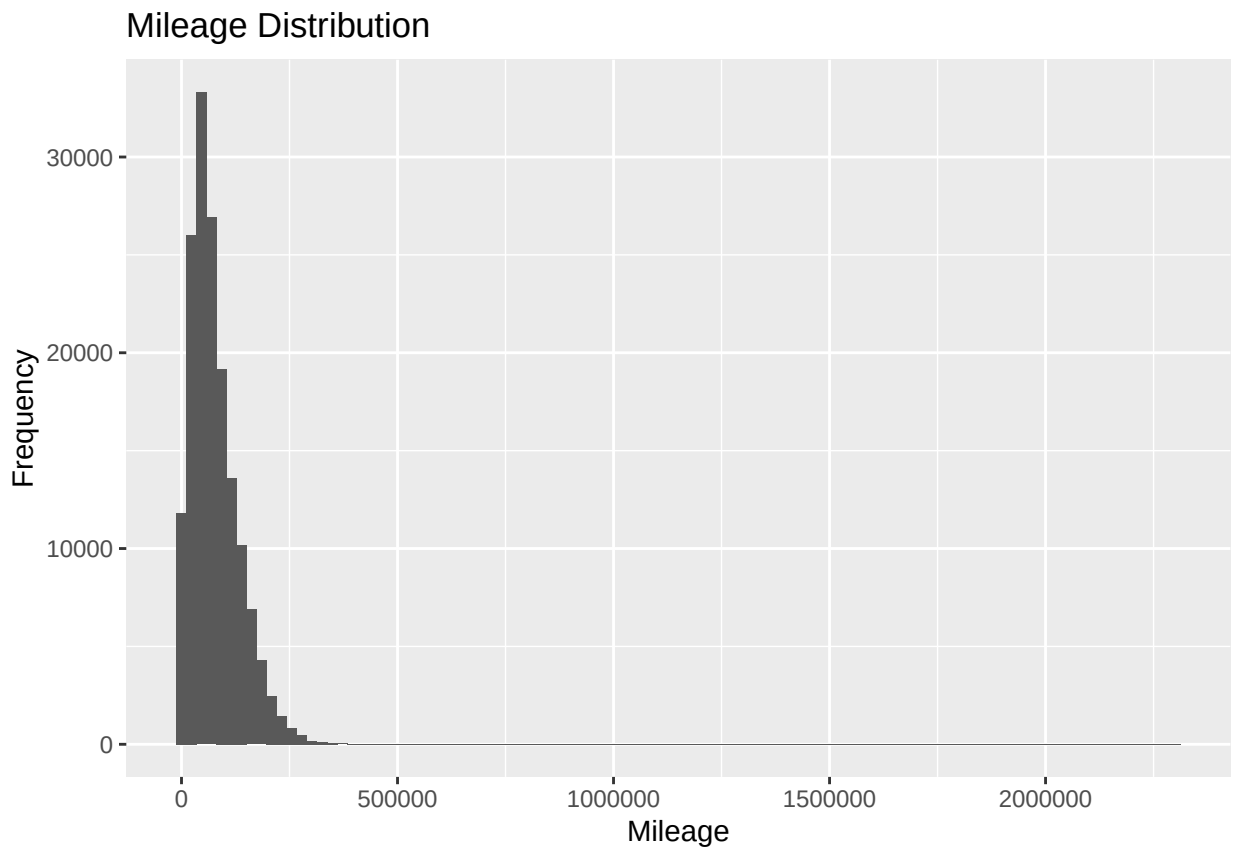
I will plot some graphs to figure out the distribution of my cleaned data first.



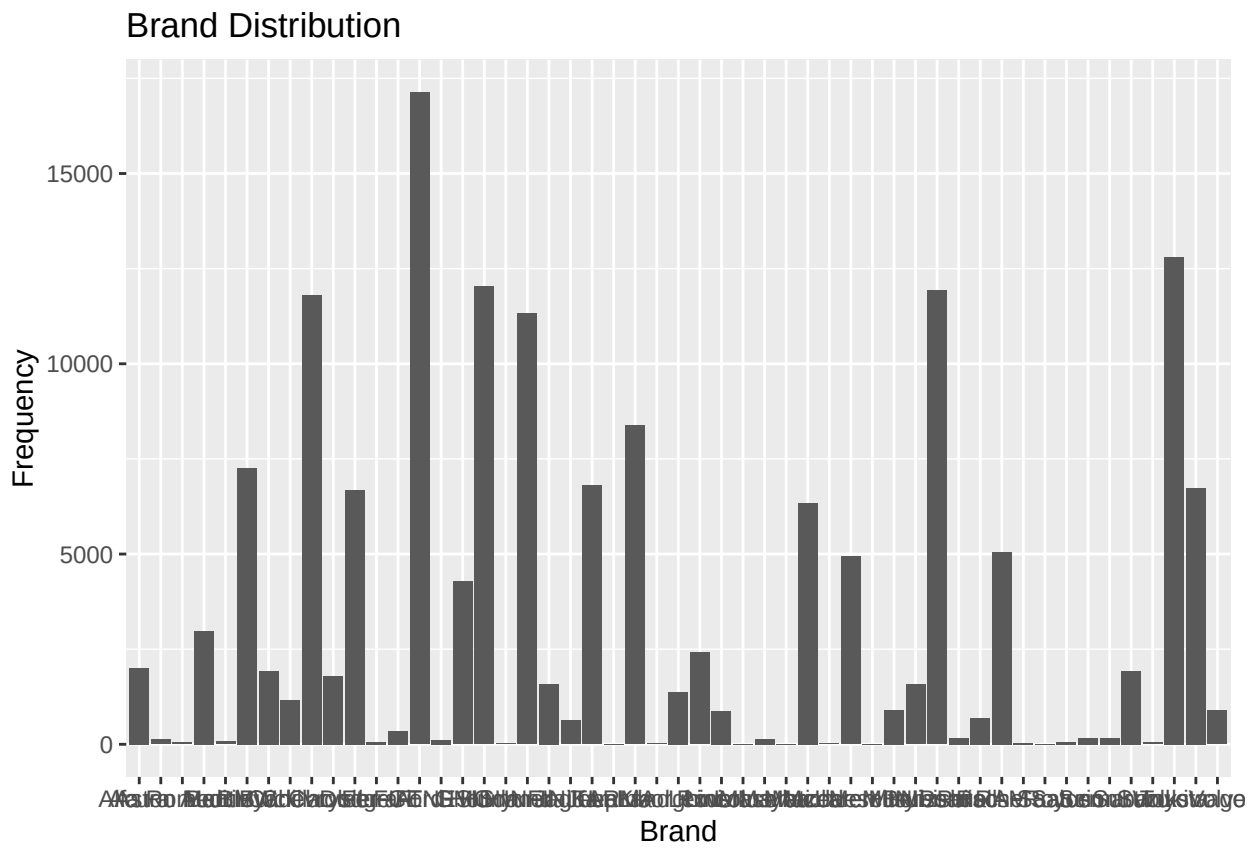
As shown in the price distribution histogram, as expected, most cars fall in between 0 to 100,000 dollar range. Specifically, between 10,000 to 40,000 CAD.



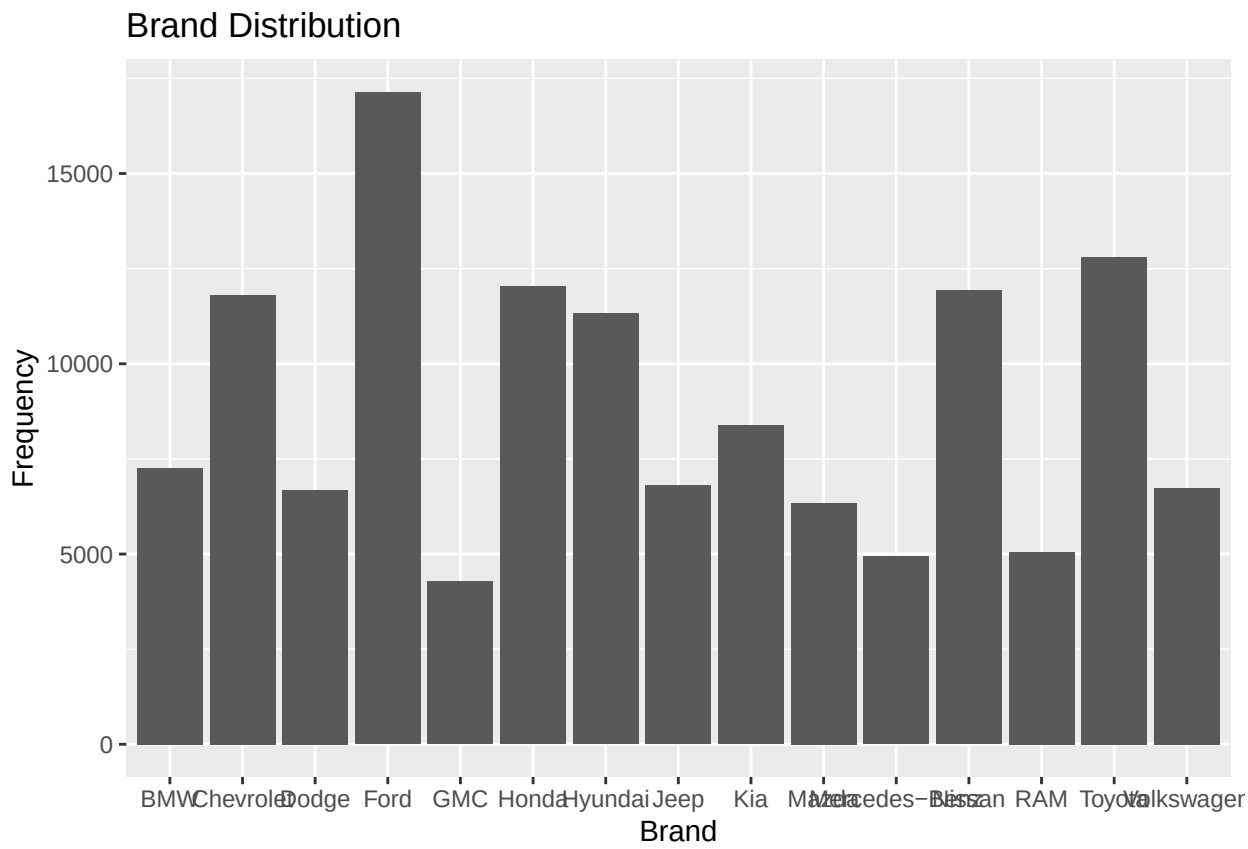
As mentioned earlier, the earliest year in my cleaned data is 2005 and the latest is 2021. The most common years are the years between 2015 and 2020, as this data set was fetched in those years.



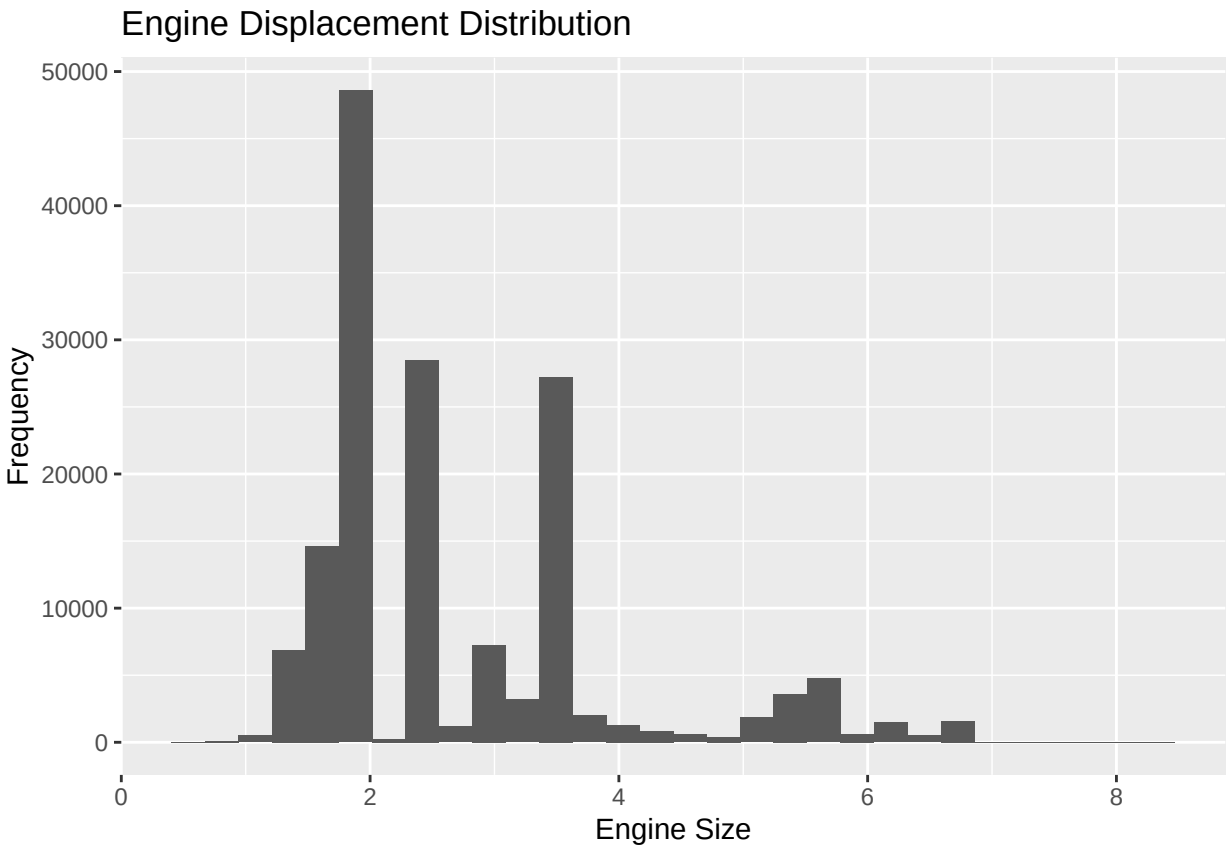
As shown, most of the cars have low mileage.



There are many brands, but only around 15 of them have decent quantity. In the modelling, I will only train the model based on this data.

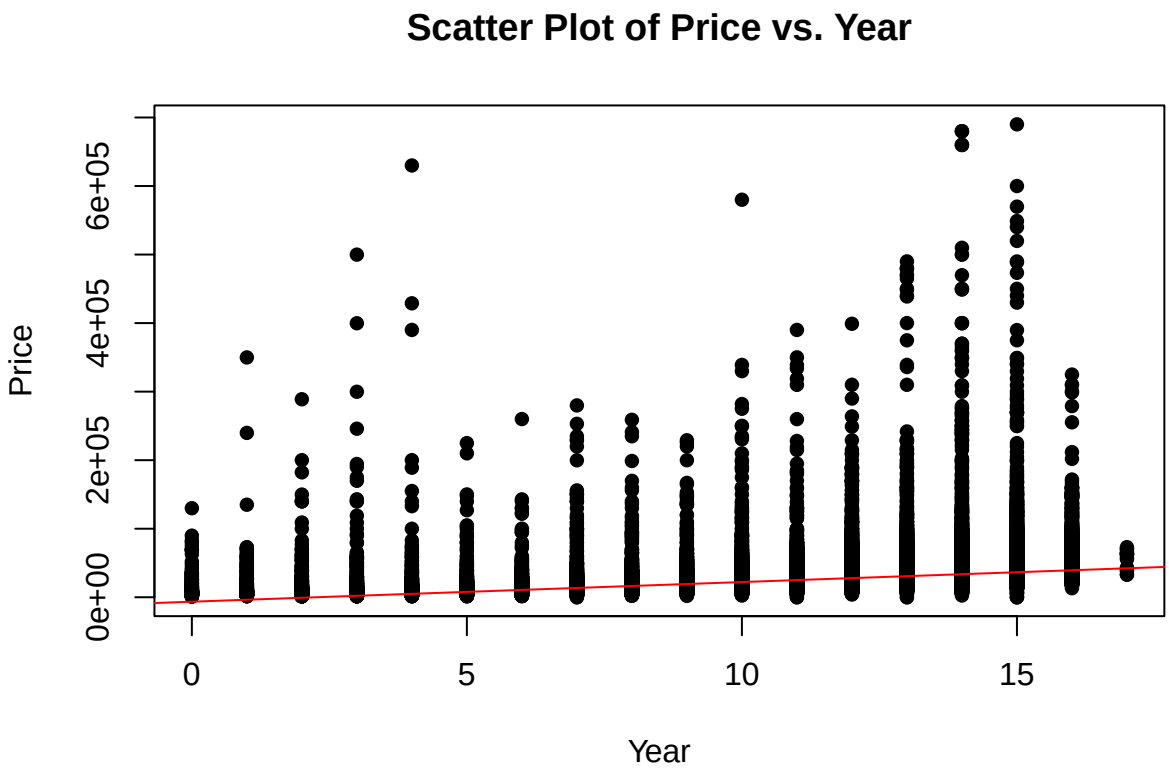


As shown in the brand bar graph, these are the most popular 15 brands. Unsurprisingly, Ford is the most popular brand.

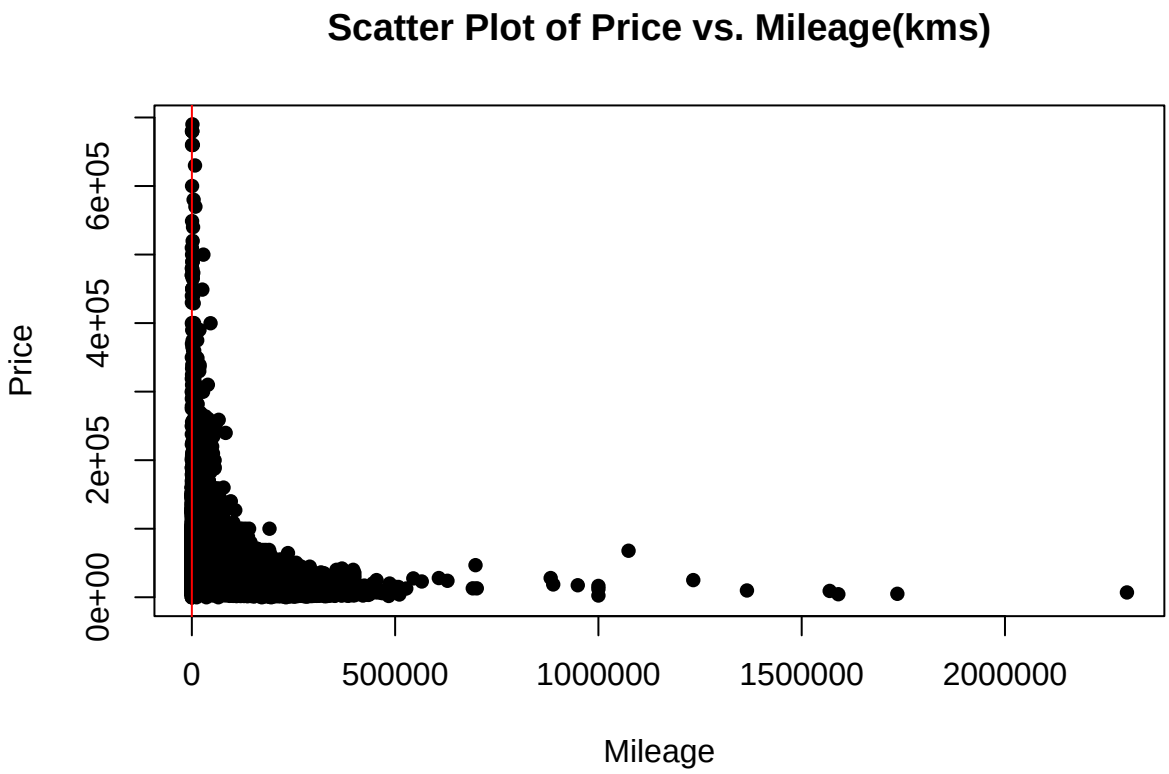


As shown in histogram, the most frequent engine size is around 2, this could be because 2 liter engines are exceedingly popular nowadays.

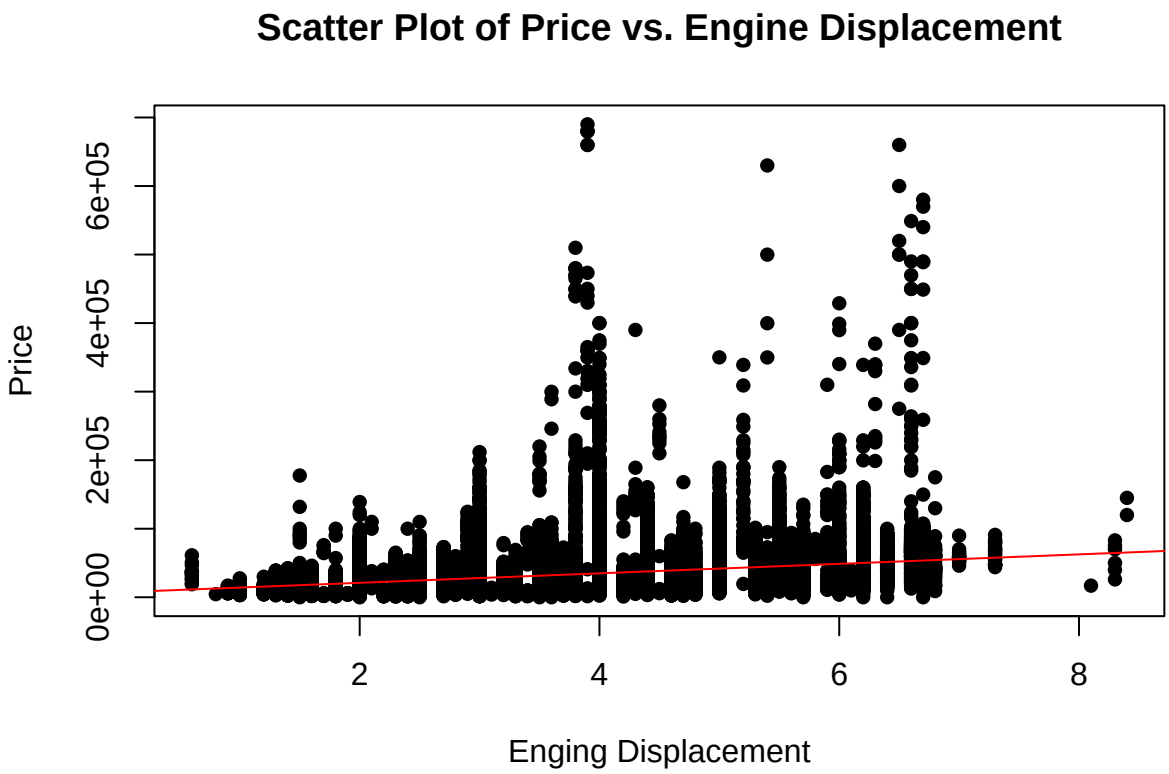
First, I will address the elephant in the room by plotting a graph to see the relationship between the year of production vs the price:



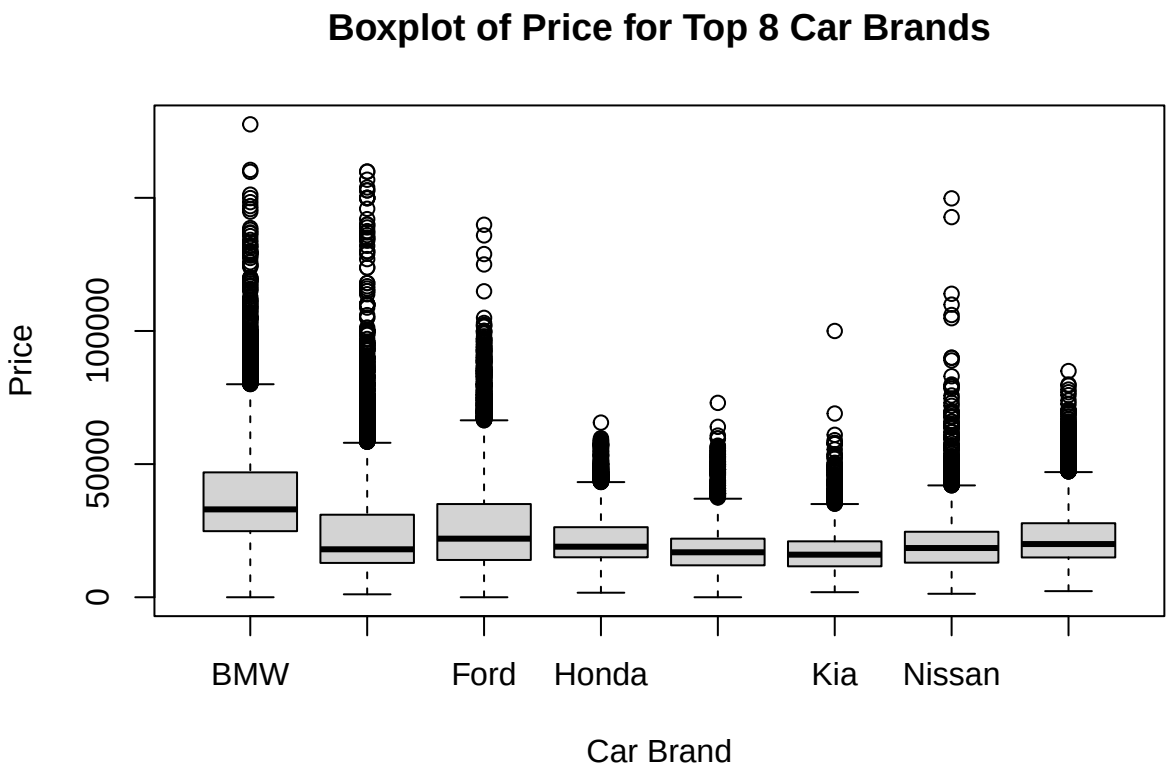
As shown from the red line, there is a trend that price increases as the model gets newer. This is expected because of depreciation of old cars and inflation in cost of things.



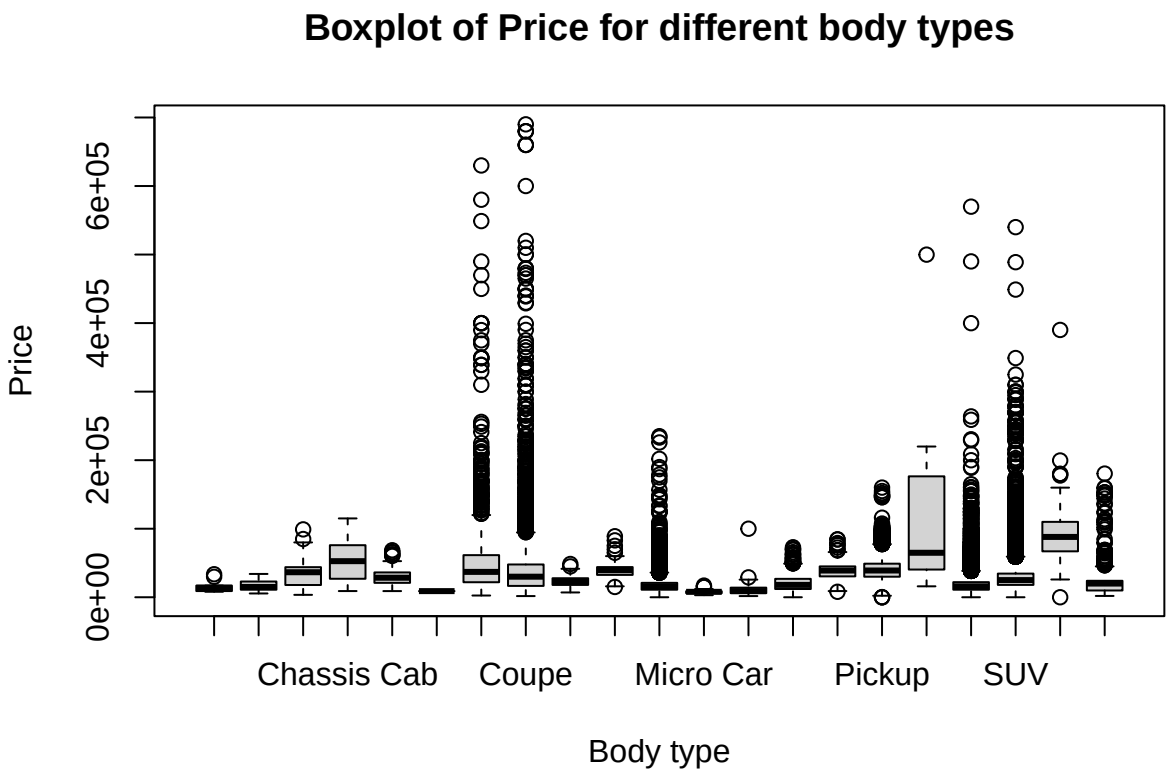
There is a strong negative relationship betwewen miles and price as shown in the scatterplot. This is intuitive because cars depreciate the more you drive them.



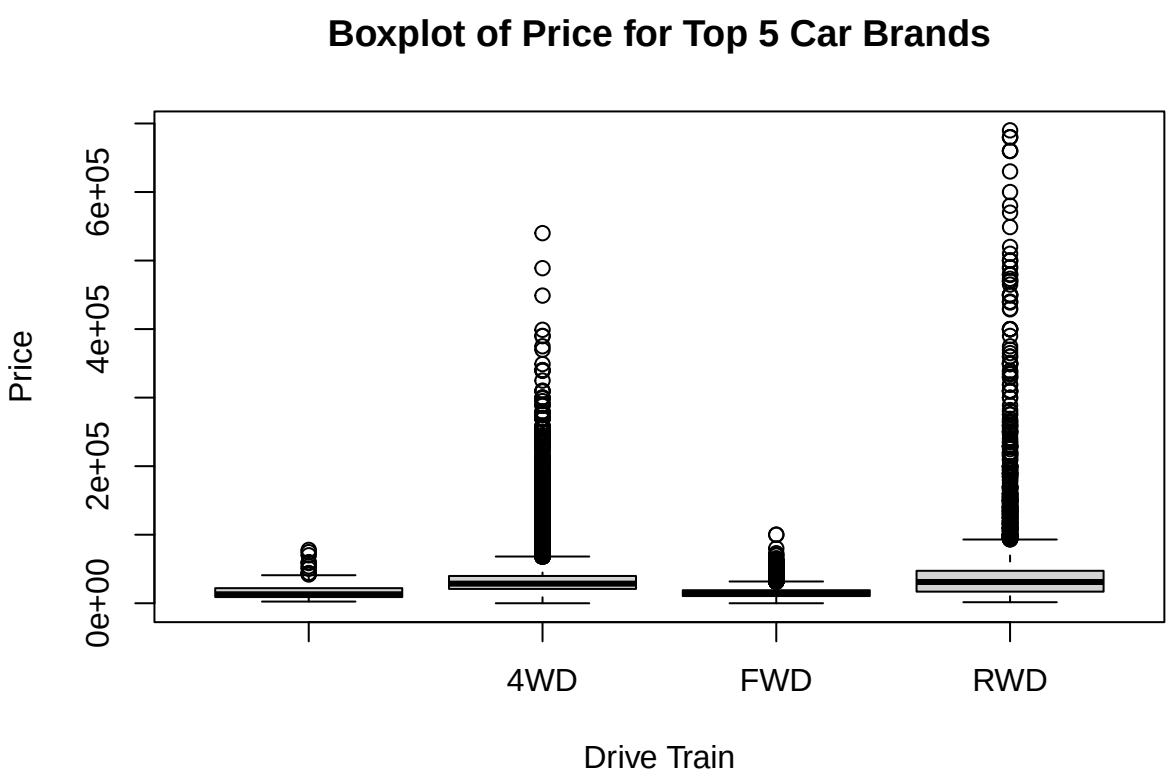
As shown in scatterplot, there is a positive relationship between the engine displacement and the year. Something to look out for, is the cluster of engine displacement of 0, because they are electric cars, which may need to be addressed in the final project.



As shown in the box plot of the price distribution of the top 5 brands, I notice that different brands have different distributions. BMW has higher prices in general as its mean and quantiles are all higher than the other 7 brands.



As shown in the boxplot above, different body types correspond to different price ranges.



For drivetrain, RWD and 4WD are more expensive than 4wd in general. Furthermore, there is a large amount of listings that didn't label the drivetrain of the car.

Summary of data

I have found substantial support that there is correlation between different factors of a car and its listed price, which in turns mean I may be able to build a model to predict the price based on other information given.

However, there are some problems already identified in my midterm report.

Firstly, electric cars have an engine displacement of 0, which can scew results.

Secondly, when exploring the relationships, particularly year, engine displacement, and mileage, I fitted a linear model and looked at their relationship. The true relationship might not be linear. Nonetheless, the linear plots give an idea of how the variables affects price.

Lastly, there may be multicollinearity and confounding in the variables. For example, year definitely affects mileage and brand affects engine displacement (luxury brands usually have bigger engines by intuition).

Modelling

I will use machine learning to predict car prices. From previous experience and the nature of the data, I will use random forest, gradient boosting, and extreme gradient boosting to predict car prices. For each model, I will try different parameters to fine tune the model. After that, I will compare the performance of the models by comparing their mean squared errors. Besides predictions of car prices, by interpreting the

variance importance plot generated by these models, we can also understand how important each variable is in prediction.

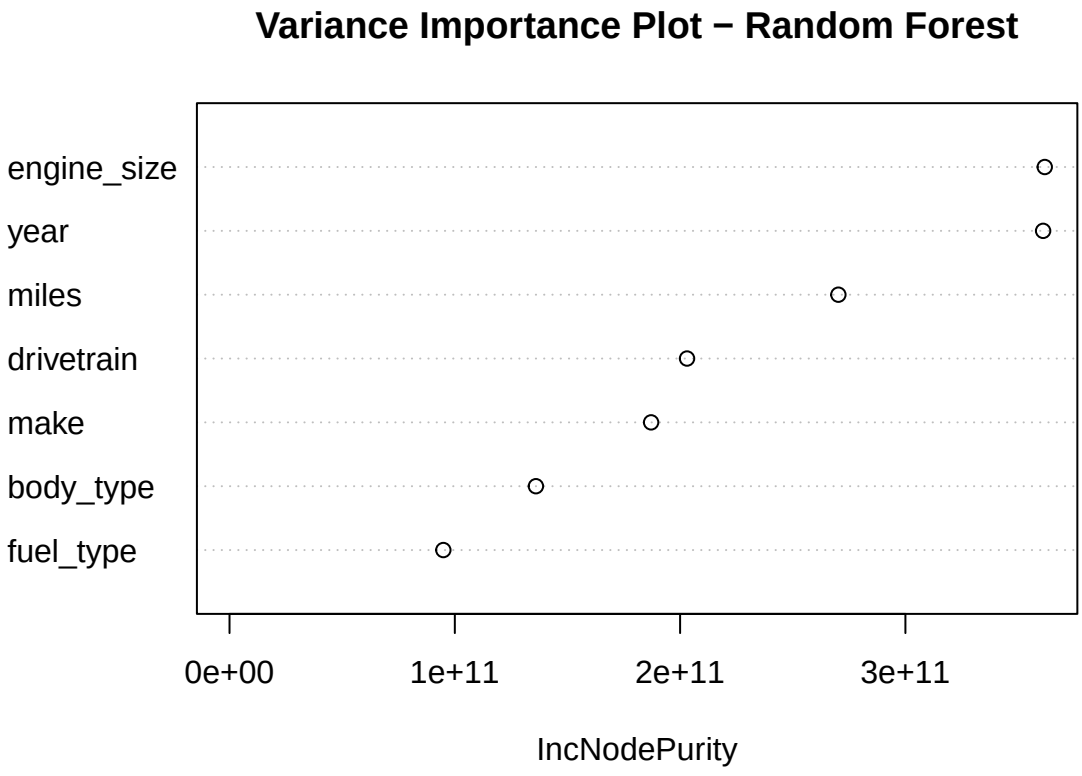
For these models, I will use miles, year, make, body_type, drivetrain, engine_size, and fuel_type to make predictions. I chose to ommit VIN since it serves as a primary key in the dataset, I also chose to ommit model since there are too many unique models to the point where I think it is possible. Furthermore, I limited the make of the cars to the top 15 cars since the sample size is too small for some exotic car brands. I will split the data into training and testing data.

Random forest

I originally tried to train the model using the entire cleaned dataset consisting the top 15 brands. However, due to its enormous size, R studio would crash whenever I tried to train the random forest model using the full dataset. Therefore I set the seed and sampled 10000 rows to train this model.

I used 500 trees to train this random forest model because 500 is a reasonable number of trees that balances out accuracy and number of trees.

Below is variable importance plot:

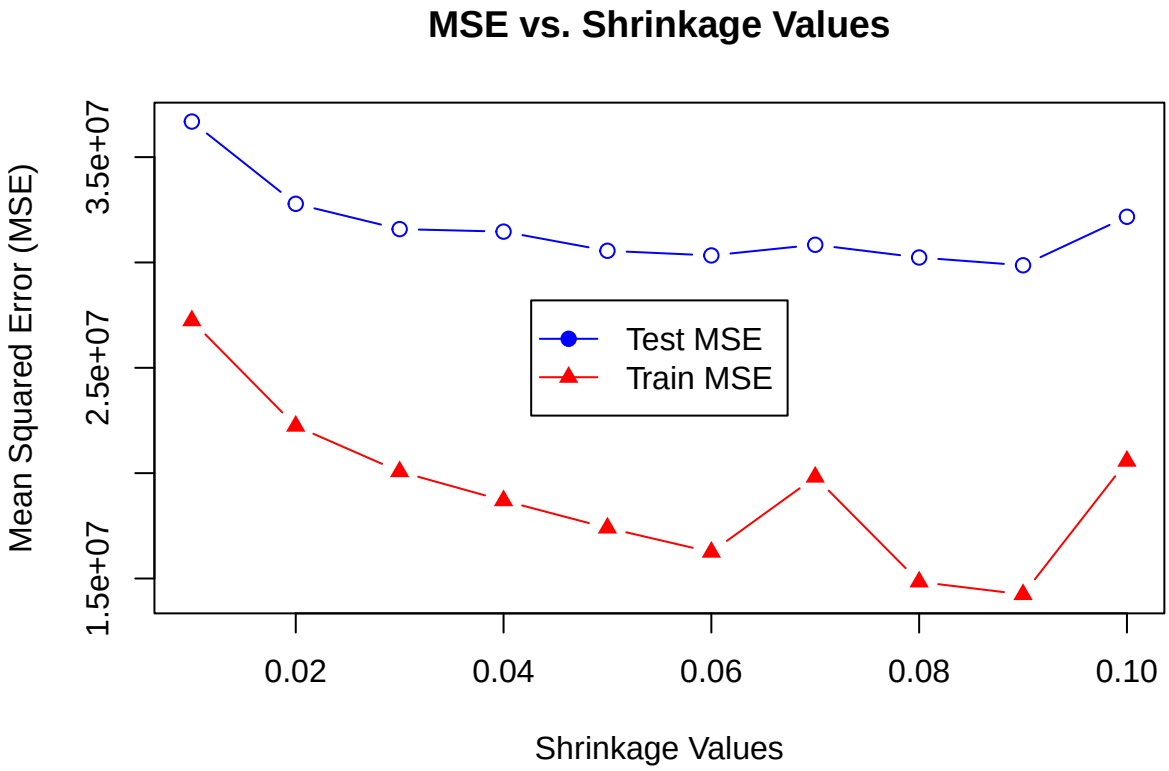


As shown in the Variance importance plot, engine_size, years are the most important variables in determining the price of a vehicle. Miles, drivetrain, and make all have moderate importance. Body type and fuel type are the least important.

Gradient boosting model

I will use the same training and testing data as the previous model. For this model, I will continue to use 500 trees. Due to the large abundance in training data, I will iterate through different shrinkage values between

0.01 to 0.1 and check which yields most accurate prediction. Furthermore, I set cross validation folds to 5 and interaction depth to 7.



Variance Importance Chart:

variable	rel.inf
engine_size	25.119385
year	19.551015
drivetrain	16.744341
miles	15.933598
make	10.802827
fuel_type	6.293206
body_type	5.555629

Looking at the variance importance chart, the ranking of the relative influence of the variables is very similar to the one produced by the random forest model.

Extreme Gradient Boosting

I will repeat the above with Extreme Gradient Boosting. For the parameters of the model, I will set the maximum depth to (6, 7, 8, 9), use the same shrinkage values, and subsample ratio of columns of (0.8, 0.9, 1). I will use a tuning grid to search for the optimal parameter: it will be run for 10 rounds with iterations ranging incremented by 50 from 50 to 500 and find the optimal shrinkage value.

After searching through the tune grid, I have narrowed the parameters to tree depth of 9, nrounds = 200, learning rate = 0.06, subsample of ratio of columns to be 0.6. After training the model, the variable importance is as follows.

	Overall
engine_size	100.000000
miles	76.698169
year	65.568145
drivetrainFWD	55.711772
fuel_typeUnleaded	10.572475
drivetrain4WD	9.282231
fuel_typePremium Unleaded	8.896107
body_typePickup	8.461758
makeMercedes-Benz	4.871735
body_typeSUV	4.399986

Similarly, year, engine size, and miles are the variables with the highest importance in predicting price. Surprisingly, make has the lowest importance here.

Results

In order to validate the models, I used them to predict the test data using the parameters and calculated the mean of the absolute difference between the price and the prediction. Here are the results:

Model	Mean Absolute Error
Random Forest	3095.312
Gradient Boosting	3164.283
Extreme Gradient Boosting	2978.323

As shown in table, extreme gradient boosting is the most accurate model, since it's predictions are closest to the real price. The improvement of using extreme gradient boosting, however, is not significant compared with other models.

I then tested the extreme boosting tree on the larger complete set of data consisting of the popular brands and I got these results.

Statistic	Value
Mean	3006.54
Maximum	554387.58
Minimum	0.06

For the complete dataset, the mean of the absolute error is 3006, minimum is approximately 0, and largest is 554387. Taking a closer look at predictions that were very accurate and very inaccurate, I noticed a pattern.

Table 6: most inaccurate predictions

	price	miles	year	make	model	body_type	drivetrain	engine_size	fuel_type
223159	630000	7788	4	Mercedes-Benz	SLR McLaren	Convertible	RWD	5.4	Premium Unleaded
223162	499900	28356	3	Mercedes-Benz	SLR McLaren	Roadster	RWD	5.4	Premium Unleaded
221798	428888	4570	4	Mercedes-Benz	SL-Class	Coupe	RWD	6.0	Premium Unleaded
223161	399800	46000	3	Mercedes-Benz	SLR McLaren	Convertible	RWD	5.4	Premium Unleaded
223158	349900	10669	1	Mercedes-Benz	SLR McLaren	Coupe	RWD	5.4	Premium Unleaded

Table 7: most accurate predictions

	price	miles	year	make	model	body_type	drivetrain	engine_size	fuel_type
163913	23885	41269	13	Hyundai	Tucson	SUV	4WD	2.0	Unleaded
93039	15825	61550	12	Honda	Civic	Sedan	FWD	2.0	Unleaded
88865	18997	83942	11	GMC	Terrain	SUV	4WD	2.4	Unleaded
133960	23495	9709	15	Nissan	Rogue	SUV	FWD	2.5	Unleaded
147598	16995	93737	10	Hyundai	Santa Fe	SUV	4WD	2.0	Unleaded

As shown in the above two tables, the extreme gradient boosting algorithm predicts prices of regular vehicles well but fail catastrophically when predicting prices of high end luxury vehicles. This, I believe, is due to the nature of the entire dataset consisting of mostly luxury vehicles.

I tried to scrape data off Autotrader and us it for validation to see whether or not the model fits today's market. However, the data that is on Autotrader is incomplete as it does not specify engine size and fuel type of the vehicle. Same goes with Kijiji. Therefore, I manually scraped a few listings of F150 trucks off Autotrader and run them through the model.

Price	Predicted
64995	62740.78
26495	32562.53
36995	44205.83
95788	64074.98
18880	24400.44

As shown in the table, the prediction of Ford F-150 trucks on the market isn't accurate, but it does give an idea of the price range of the vehicle.

Conclusion

With the car listing dataset I downloaded off Kaggle, I trained 3 different models that used mileage, model year, brand, body type, drivetrain, engine size, and fuel type to predict the price of the vehicle. The most accurate model came out to be the Extreme Gradient Boosting model, where the average error it's prediction of vehicles in the entire dataset was about 3000 dollars. The model tells us the most important variables

that can be used to predict the variables are engine size followed by mileage, drivetrain, fuel type, body type, then finally make. The model performs very well on normal consumer vehicles but fails on luxury exotic vehicles. Lastly, I tested the model on 5 listings on the current market, and the results weren't accurate, but it did give a good estimate of the price range of the vehicle.

Future extension

The model was built on a small sample of a large dataset consisting of many different brands and models of cars. This lead to the model making poor predictions of exotic vehicles but good results of regular vehicles. This problem may be solved by filtering the dataset and training the model exclusively for the specific segment or brand or even model of vehicles.

Data

data: https://www.kaggle.com/datasets/rupeshraundal/marketcheck-automotive-data-us-canada

listing1: https://www.autotrader.ca/a/ford/f-150/north%20york/ontario/5_62125027_20200826183433557/?showcpo=ShowCpo&ncse=no&ursrc=boost_hl&orup=10_15_2719&sprx=100

listing2: https://www.autotrader.ca/a/ford/f-150/cayuga/ontario/5_62041135_20211217192838090/?showcpo=ShowCpo&ncse=no&ursrc=boost_hl&orup=5_15_2719&sprx=100

listing3: https://www.autotrader.ca/a/ford/f-150/mississauga/ontario/5_62118553_on20080331102336047/?showcpo=ShowCpo&ncse=no&ursrc=boost_hl&orup=6_15_2719&sprx=100

listing4: https://www.autotrader.ca/a/ford/f-150/waterloo/ontario/5_61957896_20180810152841763/?showcpo=ShowCpo&ncse=no&ursrc=boost_hl&orup=7_15_2719&sprx=100

listing5: https://www.autotrader.ca/a/ford/f-150/georgetown/ontario/5_61883577_on20080211122257851/?showcpo=ShowCpo&ncse=no&ursrc=pl&urp=2&urm=8&sprx=100