Predicting price of Cars Sales

Tyler Rappaport, Yanjing Wang, Ziru Wang

Ban 5600: Advanced Big Data Computing and Programming

Prof. Hamidreza Ahady Dolatsara

April 18, 2023

Table of Contents

Project Introduction

Since it's invention in the 1880s, cars have been a great equalizer in transportation and a great driver of economic growth across the World[1]. Seeing explosive evolution in a mere centuries time, cars have continually evolved to meet increasing standards on safety, speed, and fuel efficiency[2], while also placating to more subjective criteria like a 'coolness' factor[3]. With the automobile industry estimated at 66.7 million units sold in 2021[4], there is a great variety of both cars sold and consumer archetypes, meaning that the automobile industry can sell to many demographic groups with different lifestyles and purchasing power[5], and has room to experiment with new features such as electric engines[6]. Given the gigantic size of the industry and the large variety of vehicles for sale, purchasing a car either new or used can be a daunting task for a consumer bombarded with options[7].

For our project, our group is examining a car price data set available on kaggle[8], where inside this data set there are over 115k observations listing a car's price as well as other relevant factors such as brand and mileage. Our groups' project goal is to see which factors correlate most with a car's listed price, either positively or negatively, to determine what a car's price should be.

---

[1] *Automobile History - History*. https://www.history.com/topics/inventions/automobiles.
[2] Linkov, Jon. "Must-Have Features to Get in Your next New Car." *Consumer Reports*, 11 Mar. 2019, https://www.consumerreports.org/automotive-technology/must-have-features-to-get-in-next-new-car-a7878321576/.
[3] Dhimaan, Sidd. "Check out This List of Some of the Coolest Cars in the World." *TopSpeed*, 27 Nov. 2022, https://www.topspeed.com/the-coolest-cars-in-the-world/.
[4] Carlier, Mathilde. "Topic: Automotive Industry Worldwide." *Statista*, https://www.statista.com/topics/1487/automotive-industry/#topicOverview.
[5] *Your Car Lifestyle: How to Pick the Vehicle That Fits You Best*. https://knowhow.napaonline.com/car-lifestyle-pick-vehicle-fits-best/.
[6] "Ev Motors Explained." *Car and Driver*, 8 Sept. 2022, https://www.caranddriver.com/features/a39493798/ev-motors-explained/.
[7] *How to Buy a Car: 17 Steps to Ownership + Free Checklists - the Zebra*. https://www.thezebra.com/resources/driving/how-to-buy-a-car/.
[8] Baffour, George Jnr. "Used and New Cars Datasets." *Kaggle*, 9 Feb. 2023, https://www.kaggle.com/datasets/georgejnr/used-and-new-cars-datasets.

Our project would be of use to both consumers wanting to determine if they are getting a good value for their purchase, and for sellers as a reference when listing vehicles for sale. Given the extensive range of sources already available for this topic[9,10,11],we also are curious to see if our findings line up with these existing resources, or if we come to a contradicting conclusion.

Research Question and Significance

A car can be many things be it a utility, a luxury, or a lemon, but universally both buyers and sellers of cars need to ask 'what is the value of a car', and 'what price is a fair price'. Adding another dimension to this question is the dominance of used cars in annual car sales, meaning that in addition to factors like brand and car quality, the condition of the car can greatly impact what price a car can sell for, with new cars generally selling for higher prices compared to used cars[12].

Our main research questions include the following;

1. Does a car's brand significantly affect it's sales price? Do certain brands better maintain their value in comparison to others?
2. What is the price difference between new and used cars?
3. Using mileage as an indicator of condition, are cars of a lower condition priced proportionally lower than cars of a higher condition?

---

[9] "Edmunds TMV - True Market Value / True Car Value." *Edmunds*, https://www.edmunds.com/tmv.html.
[10] *New Car & Used Car Values | Get the Kelley Blue Book Value*. https://www.kbb.com/car-values/.
[11] "Getting a Fair Price on Your next Vehicle." *TrueCar Blog*, 3 June 2020, https://www.truecar.com/blog/getting-a-fair-price-on-your-next-vehicle/.
[12] -, Emmanuel Forge, et al. "New vs Used vs Very Used Cars (What's The Real Cost Difference?) - Not Waiting to Live." *Not Waiting to Live - What Are You Waiting for?*, 10 Oct. 2020, https://notwaitingtolive.com/new-vs-used/.

4. How does the age of a car affect its price? With there being an assumed relationship between mileage and age, does the inclusion of a car's condition change the outcomes of this question?

These questions are significant because not only will they help buyers and sellers navigate the short term car market, more specifically gauging what is a 'fair price' for a car, but can also help demonstrate the value of cars as a long term investment.

Literary Review

While it can be universally stated that a new car should be priced higher than it's used car counterpart, this begs the question of 'How much higher?'. While this greatly depends on the brand, model, and condition of said car, this alludes to a greater debate between whether or not the difference in price is enough to justify purchasing one over the other. Used cars are obviously cheaper but not only are their best performance years behind them, they can have damage to important features like the brakes and/or engines. Depending on their year and model, used-cars can also lack modern technology in demand such as enhanced safety features. New cars meanwhile are expensive even beyond their sticker price, as while they are easier to seek loans for, they are likewise pricier to insure[13]. Purchasing a new car from a car dealership can be seen as the safest and easiest option, as a buyer knows exactly what they are getting, and can feel safe knowing that if they run into any issues they bring them up to the dealership[14]. Beyond the sticker price, after purchasing a new car and driving it out of the dealership lot, there is a secret cost of about 15-20% to the cars value, meaning that if the buyer was to immediately resell then

[13] *New Cars vs. Used Cars | U.S. News*. https://cars.usnews.com/cars-trucks/advice/new-cars-vs-used-cars.
[14] *What Is the Easiest Way to Buy a Car? - Carsdirect*. https://www.carsdirect.com/auto-loans/getting-a-car-loan/what-is-the-easiest-way-to-buy-a-car.

they would have lost money on their investment[15]. A more modern issue that disproportionately affects new cars is the computer chip shortage, where a lack of chips means that not all modern features of a car can be accessed[16]. In addition to that, modern technical features are not universally valued, as some question if the technology could be used to infringe on their liberties[17], while others fear the implementation of microtransactions in modern vehicles where monthly fees would be required to access features like heated seats[18]. That is not to say however the implementation of this technology is a universal negative, quite the opposite, as modern features like backup-cameras can greatly increase the perceived value and utility of a vehicle[19]. Ultimately, it is up to the user to determine if a new car not only has desired features, but is worth the increased asking price.

In 2020 there were approximately 10 car crashes per minute, and it is a fact of the market that some of these cars with damages ranging from a minor fender bender to a complete totaling would be resold after the fact as used cars[20]. Measuring the price of a damaged used-car is not always cut and dry, prompting the development of a sub-industry devoted towards gauging both the full extent of the damages, and what effect on the cars value these damages have[21]. Beyond

---

[15] *Why Does a New Car Lose Value after It's Driven off the Lot?* https://www.carsdirect.com/used-car-prices/why-does-a-new-car-lose-value-after-its-driven-off-the-lot.

[16] "Here Are the Models That Are Being Affected by the Chip Shortage." *Car and Driver*, 7 Feb. 2023, https://www.caranddriver.com/news/g36218381/car-models-affected-chip-shortage/.

[17] Getahun, Hannah. "Rick Ross Says He Would Never Ride in a Tesla Because He Fears the Self-Driving Car Could Take Him to the Authorities against His Will." *Insider*, Insider, 29 Jan. 2023, https://www.insider.com/rick-ross-fears-tesla-would-take-him-places-against-will-2023-1.

[18] Wayt, Theo. "BMW Owners Outraged over $18-a-Month Charge to Use Heated Seats." *New York Post*, New York Post, 12 July 2022, https://nypost.com/2022/07/12/bmw-owners-outraged-over-18-a-month-charge-to-use-heated-seats/.

[19] Gareffa, Peter. "What You Need to Know about Backup Cameras." *Edmunds*, 8 Nov. 2018, https://www.edmunds.com/car-technology/8-things-you-need-to-know-about-back-up-cameras.html.

[20] Christy Bieber, J.D. "Car Accident Statistics for 2023." *Forbes*, Forbes Magazine, 14 Mar. 2023, https://www.forbes.com/advisor/legal/car-accident-statistics/.

[21] "Shopping for a Used Car?" *CARFAX*, https://www.carfax.com/.

car accidents, cars can accumulate a general wear and tear from usage that can be measured in mileage, with it being assumed that with all else being equal a car with higher mileage will be in a worse condition than a car with lower mileage, thus commanding different prices[22]. Mileage does not affect all cars equally, with some cars having significantly longer lifespans than others owing to their design and intended usage[23]. For the purposes of this project, Mileage alone will be used as a gauge for a car's condition.

It is next to impossible to talk about cars without mentioning car brands, as to some a car's brand is more important than the actual car itself[24]. Car brands can court different audiences, with Fords being known for it's lineup of pickup trucks, Toyota for it's lineup of sedans[25], and Mercedes-Benz for it's luxury vehicles[26]. Not all brands are created equal, with a positive or negative reputation on reliability, safety, and utility all having a factor in determining what price a person would be willing to pay for it. The previously mentioned Mercedes-Benz is considered generally an unreliable car despite it's luxury brand status[27]. Car brands are also known to court loyalty in customers, typically due to a desire of the consumer to associate with the brand image and/or the customers having previous positive experiences with said brand[28],

---

[22] D'Allegro, Joe. "Just What Factors into the Value of Your Used Car?" *Investopedia*, Investopedia, 19 Dec. 2022, https://www.investopedia.com/articles/investing/090314/just-what-factors-value-your-used-car.asp.
[23] O'Neill, Rebecca. "10 Cars with the Longest Life Spans." *HotCars*, 25 Sept. 2019, https://www.hotcars.com/longest-life-span-cars/.
[24] Fugate, Jantzen. "Valuing a Brand: What's Your Brand Worth? How to Value a Brand." *Nav*, 8 Sept. 2022, https://www.nav.com/blog/valuing-a-brand-whats-your-brand-worth-how-to-value-a-brand-554217/.
[25] "37 Types of CAR Brands You Should Know." *Lemon Bin Vehicle Guides*, 5 Dec. 2021, https://lemonbin.com/types-of-car-brands/.
[26] Bradley, Michael. "15 Best Luxury Car Brands: Ranking of the Top Premium Vehicles." *Luxe Digital*, 23 Mar. 2023, https://luxe.digital/business/digital-luxury-ranking/best-luxury-car-brands/.
[27] https://www.autolist.com/guides/most-reliable-car-brands
[28] Levin, Tim. "These Are the 20 Car Brands with the Most Loyal Customers." *Business Insider*, Business Insider, https://www.businessinsider.com/car-buying-brands-most-loyal-customers-automotive-sales-loyalty-subaru-2020-7?op=1#20-acura-1.

meaning that brands with high levels of consumer loyalty can increase the value of their vehicles simply by adding their logo.

Running the Code

For this project, the code was generated in .ipynb files using Pyspark. This code can be run on both Databricks and on a google cloud with a Pyspark environment capable of running Hadoop and Yarn.

After importing our data onto the data bricks platform, we shared the code on GitHub and completed the cleaning of our respective data. Because the data existed with different types of characters, we changed the character type and also extracted the brand data from the text of the car type. Because there were significant outliers in the prices of the automotive goods, we used the box plot method to remove more than 75% of the outliers. After cleaning the data, we included a total of seven variables; Model, Year, Status, Mileage, price, MSRP, and Brand.

Dataset Description

The dataset on car sales was retrieved from kaggle, and contains information on over 115k cars with a great variety of prices, brand, and quality. Price is the dependent variable our group will analyze, with the independent variables including ‘Model’, ‘Year’, ‘Status’, ‘Mileage’ and ‘MSRP’. To give a full outline of the dataset, a table is included below to display all the variables used:

Data Name	Data Type	Data Desc.
Model	String	The design of the car, it contains the year of manufacture, the brand and the type of car.
Year	numeric	Year of manufacture of the car, ranging from 1923 to 2023.
Status	String	Condition of the car; used, certified, or new.
Mileage	numeric	Distance driven by used cars in miles. The distance traveled in a new car is not available in the data set and is assumed to be 0.
Price	numeric	The price is the price at which the car is offered for sale in the car market.
MSRP (Manufacturer's Suggested Retail Price)	numeric	MSRP means suggested car retail price; rows display quoted or discounted current selling price relative to MSRP.
Brand	String	A Car's specific brand

Descriptive Statistics

Below are the following data descriptions;

	Model	Year	Status	Mileage	Price	MSRP
0	2022 Acura TLX A-Spec	2022	New	Not available	$49,445	MSRP $49,445
1	2023 Acura RDX A-Spec	2023	New	Not available	$50,895	Not specified
2	2023 Acura TLX Type S	2023	New	Not available	$57,745	Not specified
3	2023 Acura TLX Type S	2023	New	Not available	$57,545	Not specified
4	2019 Acura MDX Sport Hybrid 3.0L w/Technology ...	2019	Used	32,675 mi.	$40,990	$600 price drop

After importing the data, we can see that the data is divided into "Model", "Year", "Status", "Mileage", "Price", "MSRP". Then we can see that there are multiple characters "Not available" in "Mileage" which are not integer type.

Data：

```
cars.dtypes
```

```
Unnamed: 0     int64
Model         object
Year           int64
Status        object
Mileage       object
Price         object
MSRP          object
dtype: object
```

```
cars.isnull().sum()
```

```
Unnamed: 0    0
Model         0
Year          0
Status        0
Mileage       0
Price         0
MSRP          0
dtype: int64
```

By looking at the data type and the number of data, we are fortunate that the data is complete without missing values. But the data type does not match what we expected. The prices, Mileage and MSRP are all in string form. In our data cleaning, we replace the invalid values with zeros and transform the string format to Integer.

1. Mileage info:

```
cars["Mileage"].describe()
```

```
count   115110.00
mean     28279.32
std      38055.55
min          0.00
25%          0.00
50%      11484.00
75%      45778.25
max     974302.00
Name: Mileage, dtype: float64
```

```
cars["Mileage"].value_counts()
```

```
0         47458
310         101
23000        19
1000         18
35000        17
            ...
56045         1
144230        1
15207         1
62184         1
67391         1
Name: Mileage, Length: 49168, dtype: int64
```

Mileage is the main independent variable of our proposed problem, and by looking at its data information we can see that the median value of Mileage is 11484. The minimum value is 0, but the maximum value is 97402, with 97402 being a considerable outlier. The mileage is mainly concentrated in 0, which makes sense given that new cars would have a uniform mileage of 0.

2. Year info:

```
cars["Year"].value_counts()
```

```
2023    34744
2022    20520
2020    10104
2021     9774
2019     9748
        ...
1958        2
1962        2
1953        1
1949        1
1966        1
Name: Year, Length: 70, dtype: int64
```

The earliest model year for the cars being listed is 1966, with 2023 being the most prevalent. This makes sense, as with each passing year from year N, there would be less and less cars from year N being sold / resold.

3. Price info:

```
pd.options.display.float_format = '{:.2f}'.format
cars["Price"].describe()
```

```
count    115110.00
mean      51517.99
std       37931.87
min        1800.00
25%       29980.00
50%       43275.00
75%       61280.00
max     2499900.00
Name: Price, dtype: float64
```

```
cars["Price"].value_counts()
```

```
29995    267
34995    245
39995    217
72010    201
24995    196
        ...
50817      1
56962      1
54915      1
23547      1
10245      1
Name: Price, Length: 33885, dtype: int64
```

```
1  element_count_price = df.groupBy('Price').count().orderBy('count', ascending=False)
2  element_count_price.show(6)
3  price_count=df.select(col('Price')).count()
4  print('There are',price_count,'valid data in total')
```

```
▸ (4) Spark Jobs
▸ element_count_price:  pyspark.sql.dataframe.DataFrame = [Price: string, count: long]
+----------+-----+
|     Price|count|
+----------+-----+
|Not Priced|  652|
|   $29,995|  241|
|   $34,995|  226|
|   $72,010|  200|
|   $39,995|  192|
|   $24,995|  183|
+----------+-----+
only showing top 6 rows
```

Counting the price field, we found that 652 pieces of data do not record prices. Compared with the total data of 115,762 pieces, the proportion is very small, and the data can be deleted in the subsequent data processing. The average price of vehicles is $51517.99, the cheapest car price on record is 1800, and the most expensive car is the Porsche Carrera GT, priced at $2499900.

4. Brand info:

```
element_count_Brand = df.groupBy('Brand').count().orderBy('Brand', ascending=False)
element_count_Brand.show()
```

▸ (2) Spark Jobs

▸ element_count_Brand: pyspark.sql.dataframe.DataFrame = [Brand: string, count: long]

```
+-------------+-----+
|        Brand|count|
+-------------+-----+
|   Volkswagen| 9968|
|       Toyota| 5709|
|        Tesla| 9068|
|      Porsche| 9961|
|Mercedes-Benz|10100|
|        Lexus| 9965|
|     INFINITI| 8664|
|      Hyundai| 8280|
|         Ford| 4465|
|        Dodge| 9819|
|    Chevrolet| 9914|
|          BMW| 9827|
|        Acura| 9370|
+-------------+-----+
```

```
cars["Brand"].describe()
```

```
count             115110
unique                13
top       Mercedes-Benz
freq              10100
Name: Brand, dtype: object
```

Of the 115110 cars in the data set, Mercedes-Benz was the most prevalent, with Ford being the least prevalent.

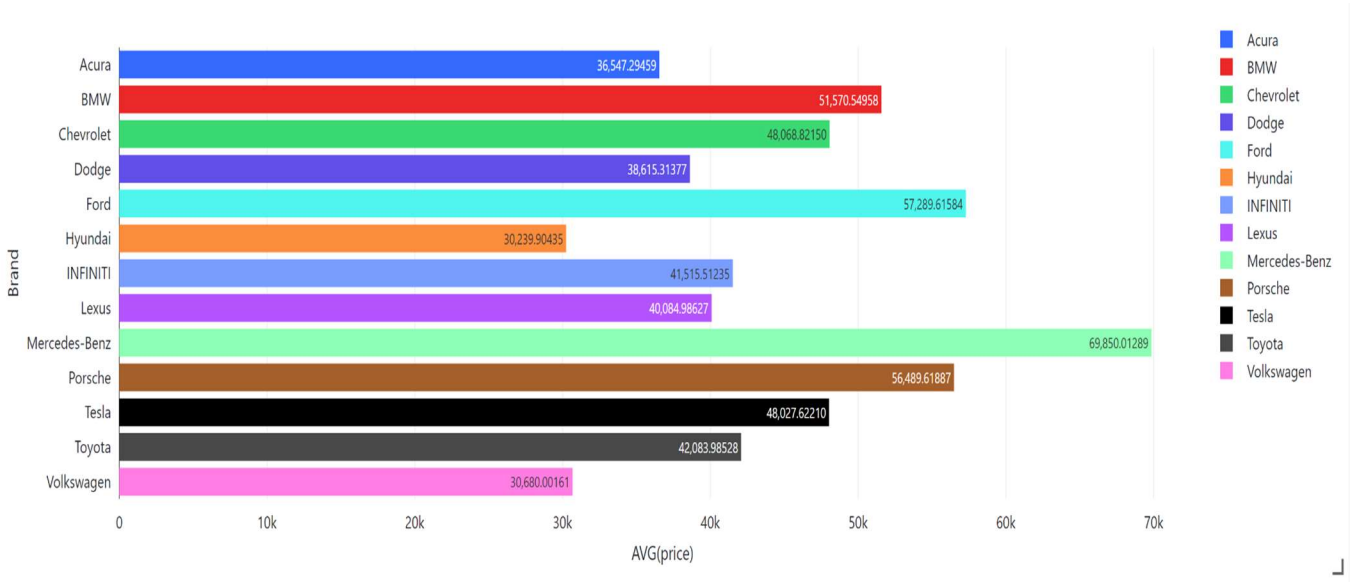
Data Visualizations

Using Pyspark to create our visualizations, our group demonstrates the relationships and overarching phenomenon in the below graphs;

1. Car Brand vs. Average Price

   In the first graph, the average price for car brands across all their models is shown. This graph displays the great variety in the car market, signaled by each brand having it's own average price range. Luxury cars like BMW and Mercedes-Benz tend to fall on the higher

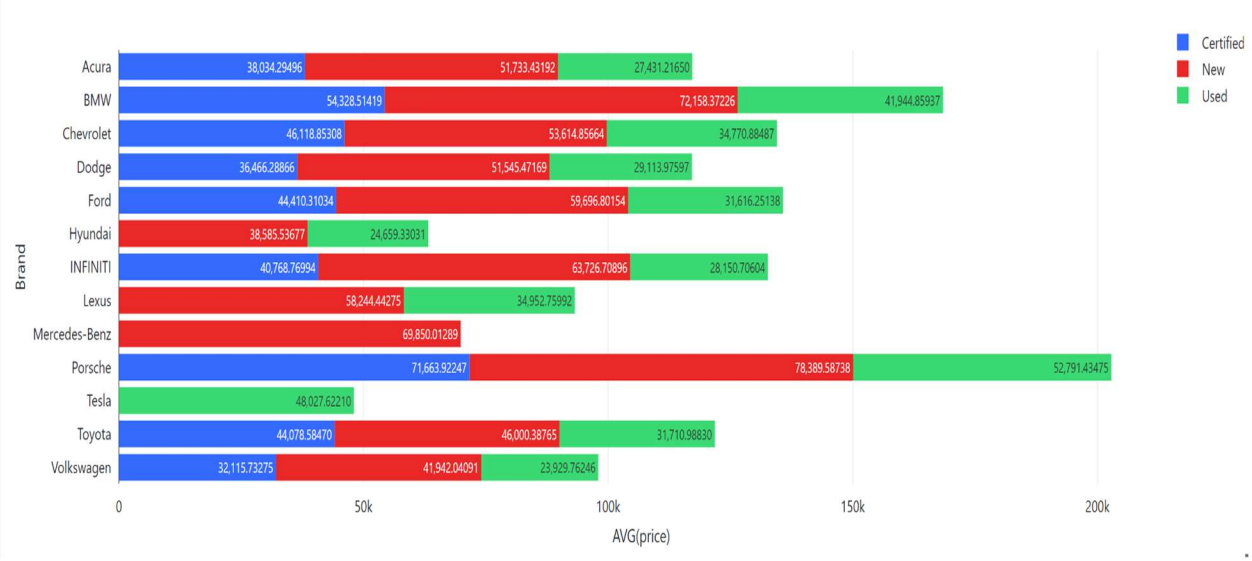
end of the price range, while more middle market cars like Lexus and INFINITI hover close to one another at a middle price range.



2. <u>Brand vs, Average Price, segmented by status</u>

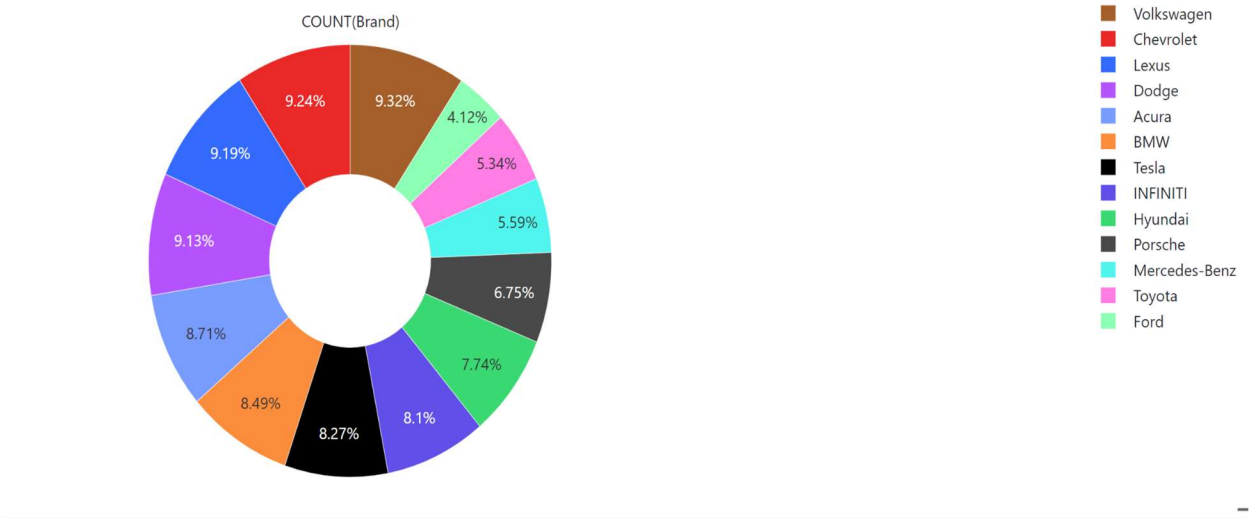
Expanding on the above graph, the Brand vs. Average price relationship is further segmented by car status, that being New, Used, or Certified. Regardless of brand, New cars tended to be the highest valued, followed by certified, and then used. Furthermore, some holes in the data make themselves known, as there are no tesla's recorded besides used, no Mercedes-Benz besides New, and no Certified Hyundais or Lexus' in the data set. As cars of these archetypes do exist in the real world, such as certified Lexus'[29], this can be considered a limitation of the data set.

[29] "LCertified." *Lexus*, https://www.lexus.com/lcertified.



3. Count of Brands

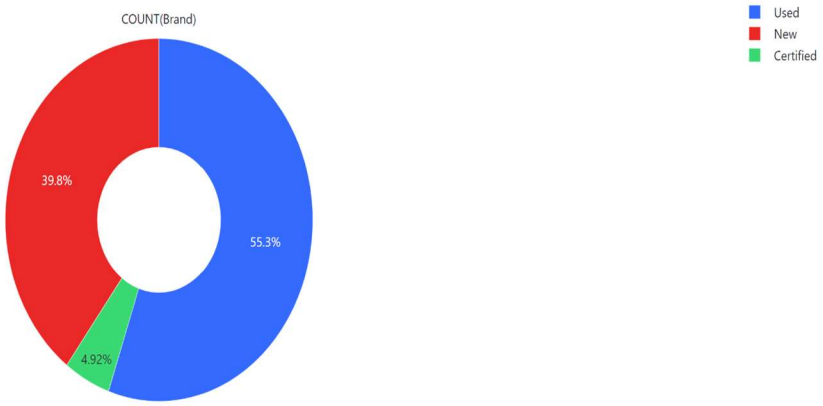
This pie chart show the percentage of observations out of roughly 107k each car brand had, with Volkswagaon having the most observations at 9.32% of the data set, and Ford having the least at 4.12% This disparity in percentages might have been a concern had it not been for the large number of base observations, meaning that Ford despite being the smallest still had roughly 4,400 observations.



4. Brands segmented by Status

A companion pie chart to the one above, this further segments the share of the data set by the vehicle status. Intresignly, Used cars dominated the data set, while certified is only a tiny sliver. In the cleaned data set, the real numbers that correspond to the pie chart are the following;

- New Cars= 42,514
- Used Cars= 59,139
- Certified Cars= 5,258

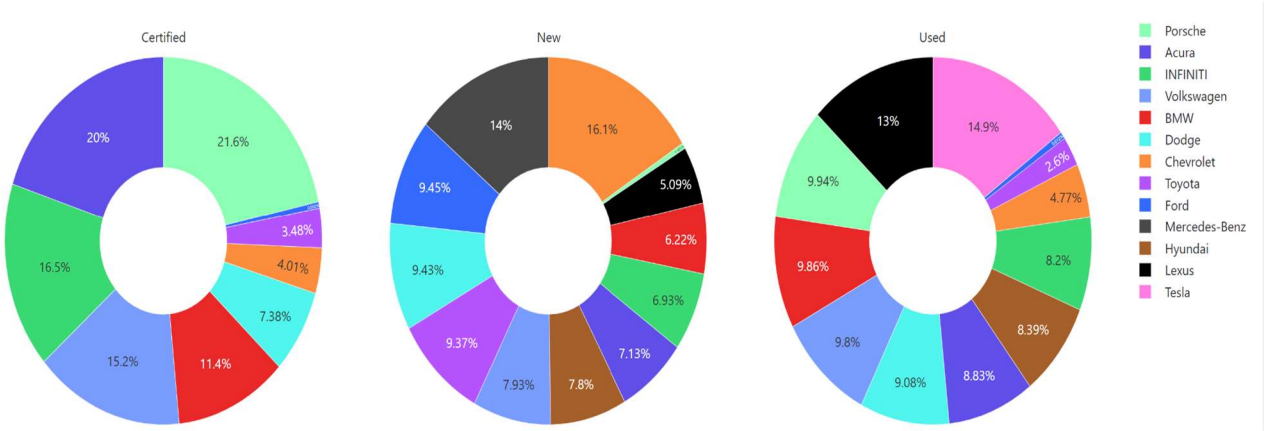


---

5. Count of Brand segmented by status

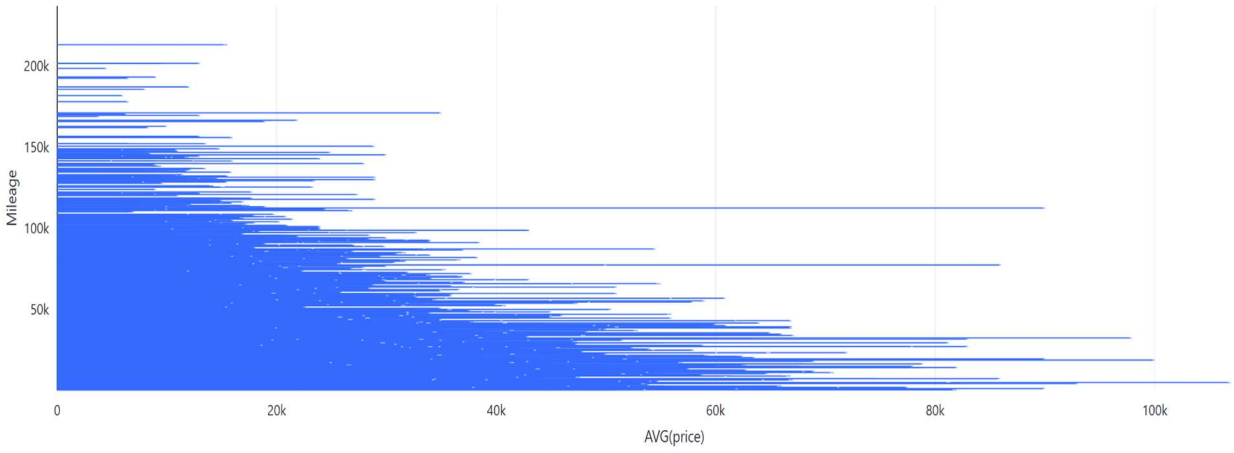
Another companion pie chart, this further segments the share of the data set by the vehicle status. This gives some additional data to flesh out the relationship of brand and status, like how the majority of new cars are Chevrolet, and that Porsche leads the certified market. To find the real number in each pie chart, the following formula can be used: (Size of Status **X** * Percentage of Staus **Y**).

For example, the number of New Chevrolets is: (42514 * 16.1%) = 6,845



6. Mileage vs. Average Price

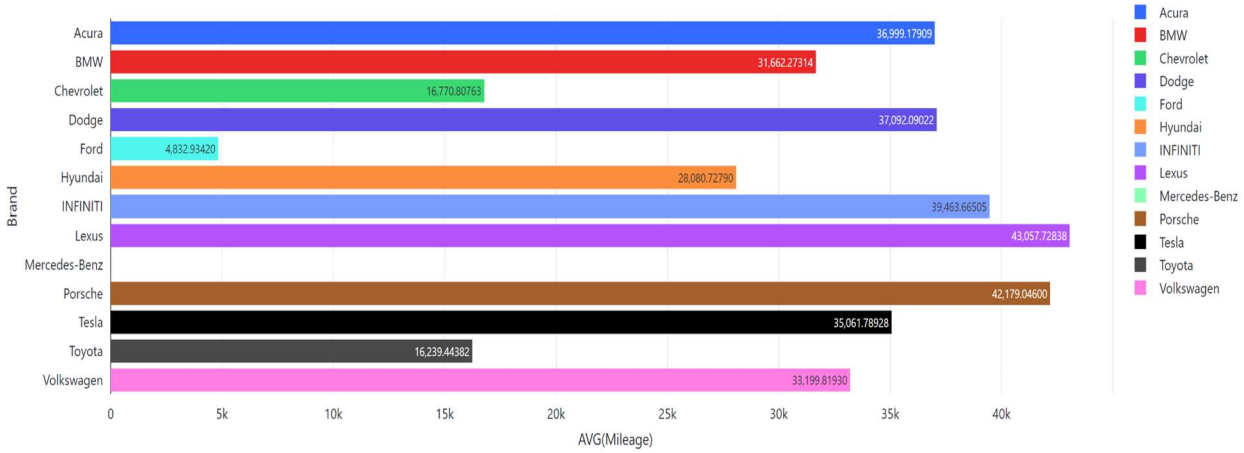
In this graph, the relationship with mileage and price is shown. As can be seen, generally the average price for vehicles tends to diminish as the mileage increases, with there being some outliers. These outliers can be explained by the fact that there are viewer observations at the higher mileages, meaning that the data can more easily be skewed, and that certain vehicles may overperform in regards to holding their value.



7. Brands vs. Average Miles

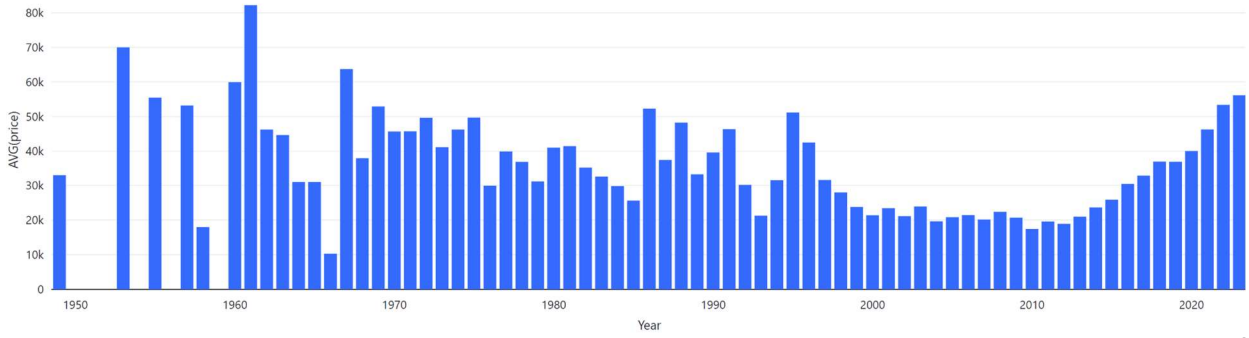
Showing the relationship between brands and average miles, this graph can speak to a brands general durability, with it being assumed that a higher average means that a car has a

longer lifespan. Lexus proved to be the leader in this regard, while Ford fell significantly behind the competition. Interestingly, as only new Mercedes-Benz cars were listed, the average number of miles is 0.



8. Model Year vs. Average Price

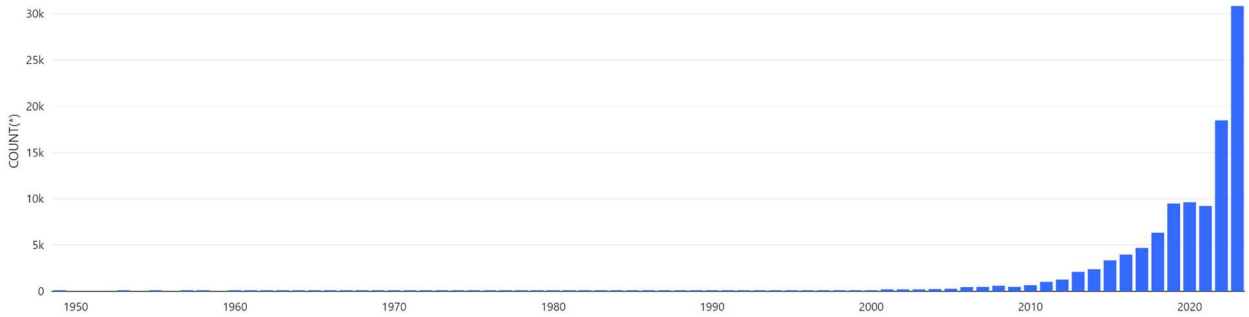
In the below graph, there is almost a c-shaped curve beyond 1996, where cars that are the oldest and cars that are the newest seem to hold the most value. Cars made prior to 1996 have wildly different pricing, with the highest average price of cars being in 1961



9. Count of Cars per Model Year

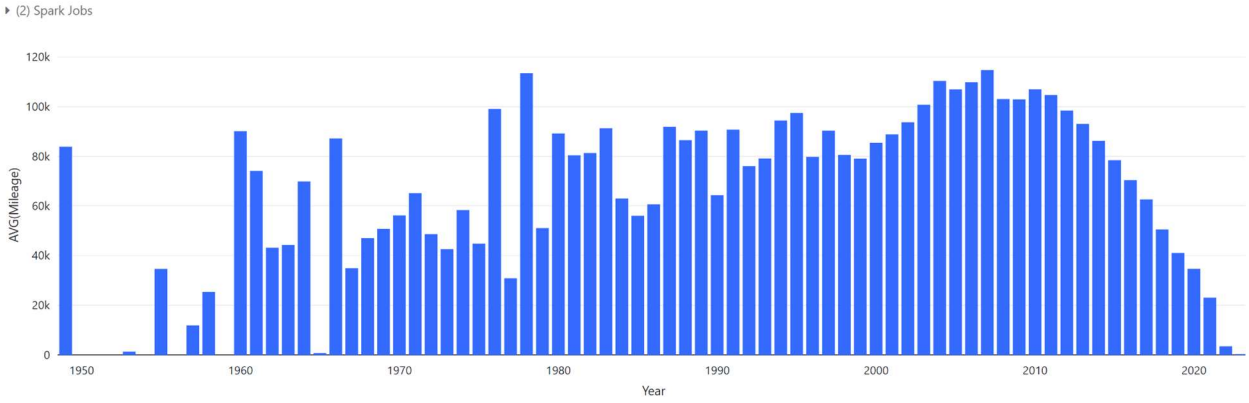
Further explaining the above graph, this graph displays how the majority of the cars in the data set were made in the past decade. The reason for the fluctuating prices pre 1996 is due to a shortage of observations, meaning that they are more susceptible to being skewed towards

outliers. As an example of this, there were only three cars listed for the 1961 range, with one of these three being priced relatively high at 159,900. Given that 1999 is the last year holding observations >50, the years 1998 and prior have to be taken with a pinch of salt



10. Model Year vs. Mileage

This graph shows the average mileage per model year. Given the above information, it is hard to confirm any relationship pre 1999 given the low number of observations. Looking at 1999 and onwards, a clear pattern emerges, where cars with more recent model years have lower mileages, which generally makes sense as they would have had less time to accumulate miles.



Data Analysis

In the data analysis part, to predict vehicle prices we used three models: Linear Regression, Random Forest and Stacked Ensemble, then from there determining which model was the strongest. The dataset was imported into the spark environment and cleaned using pyspark.sql. Linear regression and random forest use pyspark.ml to complete the machine learning and adjust the parameters to output the models. For Stacked Ensemble we used H2O, a fully open-source distributed in-memory machine learning platform with excellent algorithms and performance. H2O is more convenient and efficient than pyspark.ml for model building and can output model test results synchronously, so we convert the cleaned dataset from a spark dataframe to a H2O data frame and then completed the model construction.

Linear Regression

We first selected two baseline models to try to fit the fluctuation pattern of prices. By using a simple linear model, we can see if there is a linear relationship between dependent variable price and independent variables vehicle age, mileage, and status. The 'status' feature at 0 is not significantly contributing to the prediction of car price in the model.Note that in the below code, The coefficients are listed in the order ‘Age’, ‘Mileage’, and ‘Status’.

```
# Print the coefficients and intercept for linear regression
print("Coefficients: {} Intercept: {}".format(model.coefficients,model.intercept))
```

```
Coefficients: [-315.84256608918884,-0.29440577788989764,0.0] Intercept: 53962.328205819256
```

RMSE (Root Mean Squared Error) is a measure of the prediction error of a model, usually used in regression models. Here, RMSE=16657 means that the root mean squared error between the model prediction and the actual value in the test set is 16657. Here, the model does not predict

very well, the error is relatively large, and the coefficient of "Status " is 0, a feature that does not contribute to the model

The R^2, coefficient of determination, of the derived multiple regression model is a statistical indicator that represents the proportion of the variation in the dependent variable explained by the independent variables in the regression model. An R^2 score of 0.357 means that 35.7% of the variation in the dependent variable can be explained by the independent variables in the regression model. In other words, the model can only capture 35.7% of the total variation in the dependent variable, which indicates that there is still a lot of unexplained variation in the data and the model leaves much unknown.

```
evaluator = RegressionEvaluator(labelCol="price", predictionCol="prediction", metricName="r2")
r2_score = evaluator.evaluate(predictions)

print("R2 score on test data: {:.3f}".format(r2_score))
```

```
▸ (1) Spark Jobs
R2 score on test data: 0.357
```

To improve the interpretation of linear regression and the accuracy of the model, we continued to fit the training data using a multinomial regression model.

```
from pyspark.ml.feature import VectorAssembler, PolynomialExpansion
# Then, let's apply a polynomial expansion to create polynomial features
poly_expansion = PolynomialExpansion(inputCol="features", outputCol="poly_features", degree=2)
df_expanded = poly_expansion.transform(df_clean)
```

The model contains all possible combinations of the original features, where the maximum number of polynomials is 2. This process is called polynomial expansion, which increases the complexity of the model and thus fits the data better and avoids underfitting. The results of the linear regression model were not satisfactory as now the model explained 38% of the rate of change of the dependent variable, which while still an improvement still left many open

questions. Given this, it can be said that there was no linear relationship between the status and the dependent variable. From here, we proceeded to train the fitted data with other machine learning models.

Random Forest

The clean dataset is passed to df1 to construct the random forest model. We mapped all the string-type variables to values using StringIndexer, and stored the mapped variables in a new intermediate dataframe, "indexed," with the same rules. We then combined all the variables into vector "features" using VectorAssembler.

```
Python
1  df1=df_clean
2  modelIndexer = StringIndexer(inputCol="Model", outputCol="ModelIndex")
3  statusIndexer = StringIndexer(inputCol="Status", outputCol="StatusIndex")
4  brandIndexer = StringIndexer(inputCol="Brand", outputCol="BrandIndex")
5  indexed = modelIndexer.fit(df1).transform(df1)
6  indexed = statusIndexer.fit(indexed).transform(indexed)
7  indexed = brandIndexer.fit(indexed).transform(indexed)
8  assembler = VectorAssembler(inputCols=["Year", "Mileage", "StatusIndex", "BrandIndex",
   "ModelIndex"], outputCol="features")

▸ (6) Spark Jobs
▸ df1:  pyspark.sql.dataframe.DataFrame = [Model: string, Year: long ... 5 more fields]
```

When constructing the random forest model, we chose a maximum tree depth of 10 and a maximum node splitting discretization of 9000, doing so because we introduced the model variable and needed to ensure that the number of nodes was greater than the number of variable types in the model variable. As you can see, our RMSE looks large, but because of the large amount of test data and the large order of magnitude of the predictor variable "price", the RMSE is considered acceptable here, and the R-squared shows that our model fit is 0.729 with a relatively high accuracy.

```
1  data = assembler.transform(indexed).select("features", "price")
```

```
Command took 0.13 seconds -- by wzr971008@gmail.com at 2023/4/15 20:36:24 on Ziru Wang's Cluster-2
```

```
Cmd 47
```

Python

```
1  (trainingData, testData) = data.randomSplit([0.7, 0.3], seed = 1)
2  rf = RandomForestRegressor(featuresCol="features", labelCol="price", numTrees=10,maxBins=9000)
3  model = rf.fit(trainingData)
4  predictions = model.transform(testData)
```

▸ (8) Spark Jobs

```
1  from pyspark.ml.evaluation import RegressionEvaluator
2
3  evaluator = RegressionEvaluator(labelCol="price", predictionCol="prediction",
   metricName="rmse")
4
5  rmse = evaluator.evaluate(predictions)
6  print("RMSE = %g" % rmse)
7
8  evaluator2 = RegressionEvaluator(labelCol="price", predictionCol="prediction",
   metricName="r2")
9
10 r2 = evaluator2.evaluate(predictions)
11 print("R-squared = %g" % r2)
```

▸ (2) Spark Jobs

```
RMSE = 10806.7
R-squared = 0.729684
```

In this model, it can be seen that the variable model has the greatest impact on price, followed by the mileage, status and year of the vehicle. Since the brand is a variant extracted from the model, it is a collection of models and has a strong correlation, with the brand having a strong influence on the price of the car.

```
importances = model.featureImportances.toArray()

# Create a dictionary of feature names and their corresponding importance scores
feature_importances = {}
for i in range(len(["Year", "Mileage", "StatusIndex", "BrandIndex","ModelIndex"])):
    feature_importances[["Year", "Mileage", "StatusIndex", "BrandIndex","ModelIndex"][i]] =
    importances[i]

# Print the feature importance scores in descending order
print("Feature Importance Scores:")
for feature, score in sorted(feature_importances.items(), key=lambda x: x[1], reverse=True):
    print("{}: {}".format(feature, score))
```

```
Feature Importance Scores:
ModelIndex: 0.6690082108804108
Mileage: 0.0910190390098326
BrandIndex: 0.0841503414764426
StatusIndex: 0.07792579806397547
Year: 0.0778966105693384
```

Stacked Ensemble

Stacked ensemble is an integrated learning method that first trains multiple base models and uses the found data of these base models to build a metamodel, which then generates the final prediction results based on the prediction structure of the base models. We expect to build a more powerful model with the stacked ensemble.

We first converted the dataset from a spark dataframe to pandas dataframe and then to an H2O data frame and initial H2O environment. We also divided the training set and test set in an 8:2 ratio.

```
import h2o
from h2o.estimators import H2OXGBoostEstimator
from h2o.automl import H2OAutoML
from h2o.estimators import H2OGradientBoostingEstimator
```

```
Command took 0.09 seconds -- by wzr971008@gmail.com at 2023/4/15 20:37:44 on Ziru Wang's Cluster-2
```

```
Cmd 53
```

```
# Initialize H2O
h2o.init()
```

For the base model, we choose the gradient boosting tree and XGBoost model, which have stronger performance compared to random forest. Due to the excessive amount of data, we set the maximum 50 trees, depth 5, and cross-validation folds to 5. Also, since the base model of the stacked ensemble model needs to be validated, all the base models should be added to the keep_cross_validation_predictions=True tag on the base model.

```
gbm = H2OGradientBoostingEstimator(seed=123, ntrees=50, max_depth=5, nfolds=5,
keep_cross_validation_predictions=True)
gbm.train(x=predictors, y=target_col, training_frame=train, validation_frame=valid)

xgb = H2OXGBoostEstimator(seed=123, ntrees=50, max_depth=5, nfolds=5,
keep_cross_validation_predictions=True)
xgb.train(x=predictors, y=target_col, training_frame=train, validation_frame=valid)
```

We then used these two models as the basis for constructing the stacked ensemble model and chose the automatic metamodel construction method to find the best metamodel algorithm. The final metamodel was created using the linear method, and the model RMSE = 8444.5537, R-Squared = 0.8277. The importance of each variable is shown below. In this model, mileage and year are more important, but since our base models are GBM and XGB, the importance of each car model is calculated separately in the tree-based model. Hence, the conclusions of the Stacked Ensemble model's importance and the Random Forest conclusions testify to each other and do not conflict. Thus, the stacked ensemble is our best-performing model, fulfilling our original purpose of using the stacked ensemble.

```
stack = H2OStackedEnsembleEstimator(base_models=[gbm, xgb], seed=123)
stack.train(x=predictors, y=target_col, training_frame=train, validation_frame=valid)
```

Model Summary for Stacked Ensemble:

key	value
Stacking strategy	cross_validation
Number of base models (used / total)	1/2
# GBM base models (used / total)	1/1
# XGBoost base models (used / total)	0/1
Metalearner algorithm	GLM
Metalearner fold assignment scheme	AUTO
Metalearner nfolds	0
Metalearner fold_column	None
Custom metalearner hyperparameters	None

```
ModelMetricsRegressionGLM: stackedensemble
** Reported on validation data. **

MSE: 71310487.53336556
RMSE: 8444.553720201297
MAE: 5482.719436145965
RMSLE: 0.2051909741879166
Mean Residual Deviance: 71310487.53336556
R^2: 0.8277158387733412
Null degrees of freedom: 2096
Residual degrees of freedom: 2095
Null deviance: 868467125081.6254
Residual deviance: 149538092357.4676
AIC: 43876.143872440436
```

Variable Importances:

variable	relative_importance	scaled_importance	percentage
Mileage	2332983033856.0000000	1.0	0.4377605
Year	615770030080.0000000	0.2639411	0.1155430
Brand.Porsche	363956174848.0000000	0.1560046	0.0682927
Brand.Hyundai	240909910016.0000000	0.1032626	0.0452043
Brand.Volkswagen	238156791808.0000000	0.1020825	0.0446877
Brand.Mercedes-Benz	223239503872.0000000	0.0956884	0.0418886
Brand.BMW	164876107776.0000000	0.0706718	0.0309373
Brand.Tesla	109355655168.0000000	0.0468737	0.0205195
Brand.Toyota	48141070336.0000000	0.0206350	0.0090332
Model.2023 INFINITI QX80 SENSORY	46981156864.0000000	0.0201378	0.0088155
---	---	---	---
Model.2022 Volkswagen Taos 1.5T SE	733083392.0000000	0.0003142	0.0001376
Model.2011 BMW 1 Series M Base	643838592.0000000	0.0002760	0.0001208
Model.1966 Dodge Coronet DELUXE	640519424.0000000	0.0002745	0.0001202
Model.1974 Porsche 911 S	319927808.0000000	0.0001371	0.0000600
Model.2020 Hyundai Palisade SEL	306273152.0000000	0.0001313	0.0000575
Model.2020 Volkswagen Passat 2.0T SE	174999552.0000000	0.0000750	0.0000328
Model.2012 Chevrolet Corvette Base	92664960.0000000	0.0000397	0.0000174
Model.1987 Volkswagen Vanagon GL	71283920.0000000	0.0000306	0.0000134
Status.Certified	61060488.0000000	0.0000262	0.0000115
Model.2016 Volkswagen Passat 2.5L S	27260944.0000000	0.0000117	0.0000051

[181 rows x 4 columns]

To better explain the data, separate importance tests were carried out for GBM and XGBoost, with their importance scores then averaged out. In this calculation, the final importance of factors like model and brand were calculated as a group, and not by individual models or brands. Listed in descending order, the features with the highest average importance are Mileage, Model, Brand, Year, and then finally Status. While Mileage is still on top, Model and Brand were able to rank above Year given their group weight. Status is still relatively less important.

variable	relative_importance_gbm	relative_importance_xgb	relative_importance_total	percentage
Mileage	6.432276e+12	2.225651e+12	4.328963e+12	38.905167
Model	6.304103e+12	9.538880e+11	3.628996e+12	32.614433
Brand	4.191227e+12	1.507103e+12	2.849165e+12	25.605959
Year	5.458844e+09	5.979128e+11	3.016858e+11	2.711305
Status	3.343560e+10	2.868864e+09	1.815223e+10	0.163137

Project Conclusion

From our data analysis, we can determine that "Mileage" is the most important factor for influencing the price. Regarding our four research questions, we can make the following conclusions;

1. We found that a car's brand does affect it's price, as in the example of the brand Porsche having a high positive correlation to prices. This makes sense, as Brands tend to favor differ car markets, (I.E. trucks vs luxury vehicles), with these differing car markets command different price ranges amongst their targeted demographics. Additionally, certain cars would have a higher brand loyalty in comparison to other allowing them to command a higher price, but since we did not explicitly test the loyalty of each car brand, this is at most a point of speculation.
2. Interestingly, we found that in our data set that Status, (I.E. if a car is New, Used, or Certified), was not a significant factor in relation to price. Contradictory to both common knowledge and some of our visuals, this can be explained as that while the cars status

itself does not determine the price, the implied effects of status does. For example, a used car with a high mileage will have a lower price. In this sense, Status can still be used as a stop gap measure for determining a price, however to get a truly accurate measure one would need to look at the more specific details of the car itself.

3. We found that cars with a higher mileage and a presumed lower condition did sell for lower prices, which goes with common knowledge. While certain cars can better keep their value with a high mileage, a high mileage is never a boon and is always an obstacle to be overcome.
4. While our linear regression model leaves a lot to be desired, it can be used as a reference for the relationship between the age (model year) of the car and it's price. Our group found that generally cars' price will decrease by about $326.69 for each additional age. While we are less confident about this answer given the low R^2 for our linear regression, we can say that generally a higher age should correlate with a reduction in price. Like with mileage, certain models might be better at keeping their value at an old age, but this is always an obstacle to overcome.

Business Recommendations and Retrospective

We can predict from the Stacked Ensemble regression model the price of the car with an accuracy of 82.77%. As this model has a fairly high accuracy for predicting car prices, we can make the following business recommendations:

- Price optimization of the model: We can use the model to predict the prices of cars, which then can help merchants best assign prices. If their prices are found to be too low,

they can increase them, and if they are found to be high they can consider lowering them if they have difficulty selling their product.

- Comparing price ranges: We can analyze a car seller to find if their prices fall within our predicted model. We can then review these car sellers and tell others if they are getting a good value for their purchase, or if they would be better served making a purchase at a different location. In this scenario, cars that are priced below our models prediction would be considered to be a great value, cars priced at the model a good value, and cars priced over a bad value.
- Consumer price comparison: Knowing that mileage is the most important factor followed by the year of the car, when comparing car prices a consumer can investigate to see if the listed price of a car reflects it's predicted value. For example, a car that is highly priced with a low mileage would make sense, but a car with a high price and mileage should be placed under higher scrutiny.

For our retrospective, one missing element in this paper was car damages. As car damages were not listed in the data set, we had to assume the quality of the car by factors like mileage. In a future project, we would ideally like to be able to include damages as a factor, such as ranking the quality of a car on a scale of 1-10, with 10 being pristine condition, and a 1 meaning that the car is totaled beyond repair. Additionally, we were not able to get into more specific detail about individual models or car features, as that would be outside the scope of both the data set and the project. Other factors like engine type and color would also be interesting to investigate.