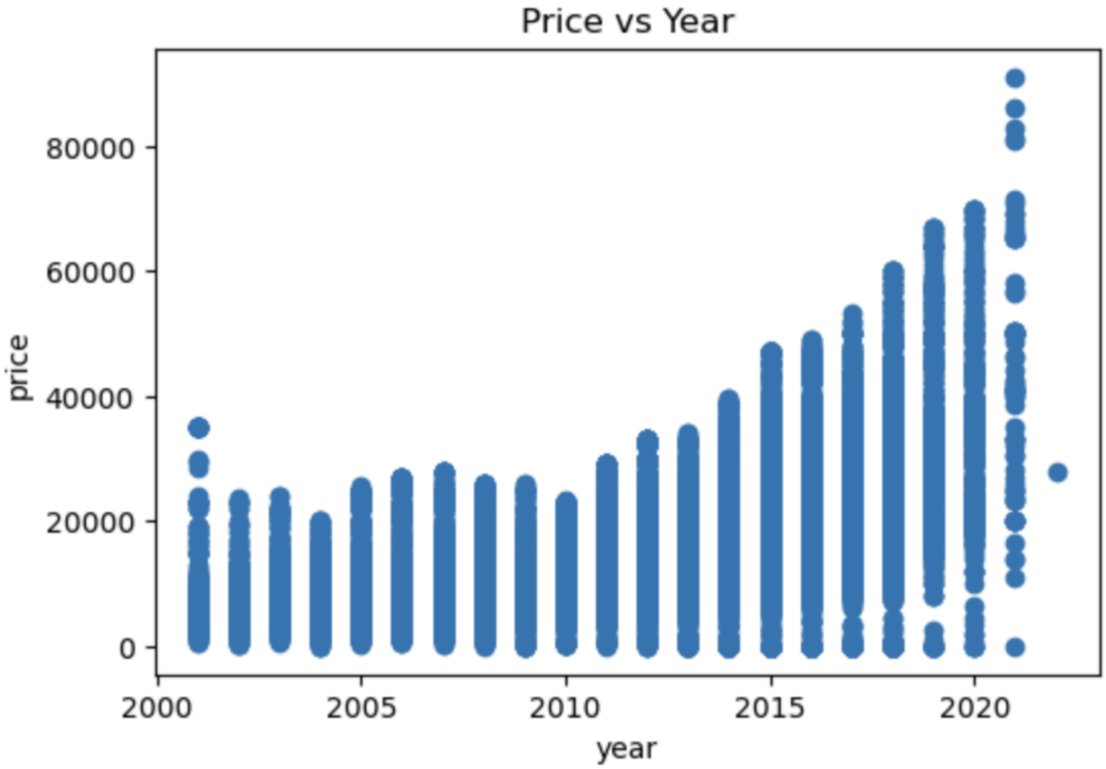
Overview

In this engagement, we were provided a dataset of used car prices that stretches from 1905 to 2022 and our task was to see if we can develop recommendations around the sorts of traits that the used car dealership community should use in select the procurement of used vehicles for resale.

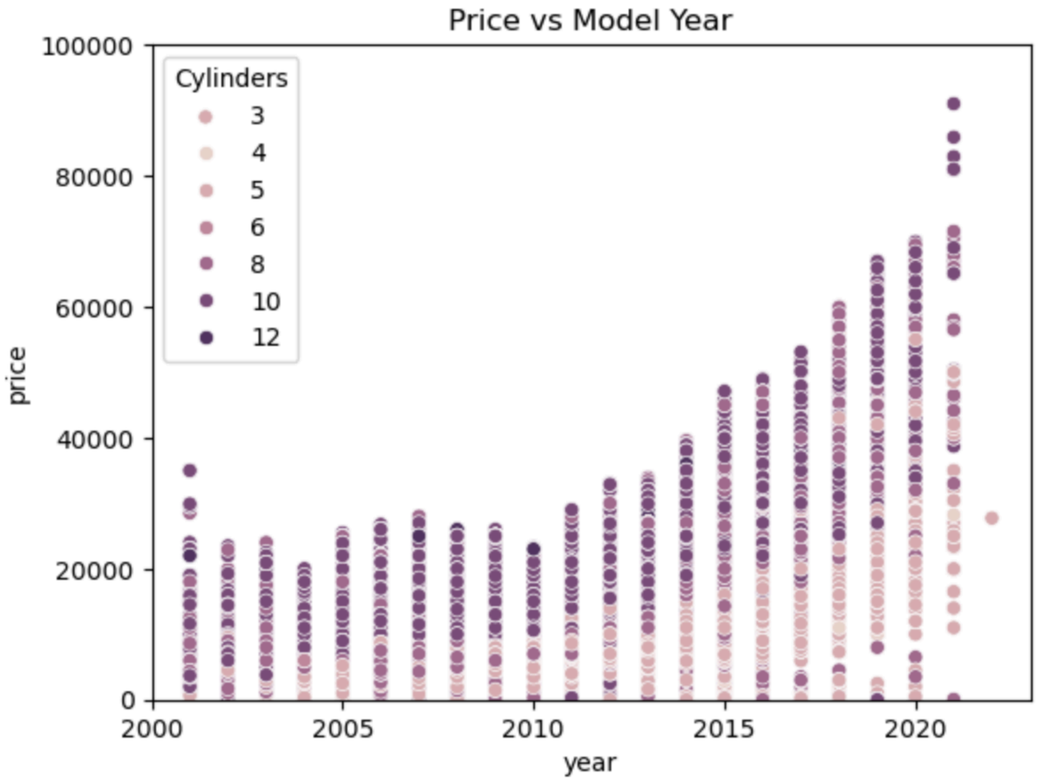
Summary Findings

To the used car dealership community, based off our analysis and findings we have determined the top 5 drivers of used car pricing which holds across linear regression and grid search approaches. Focusing on the top 5 drivers of used car pricing, we recommend that the used car dealership community focus on:

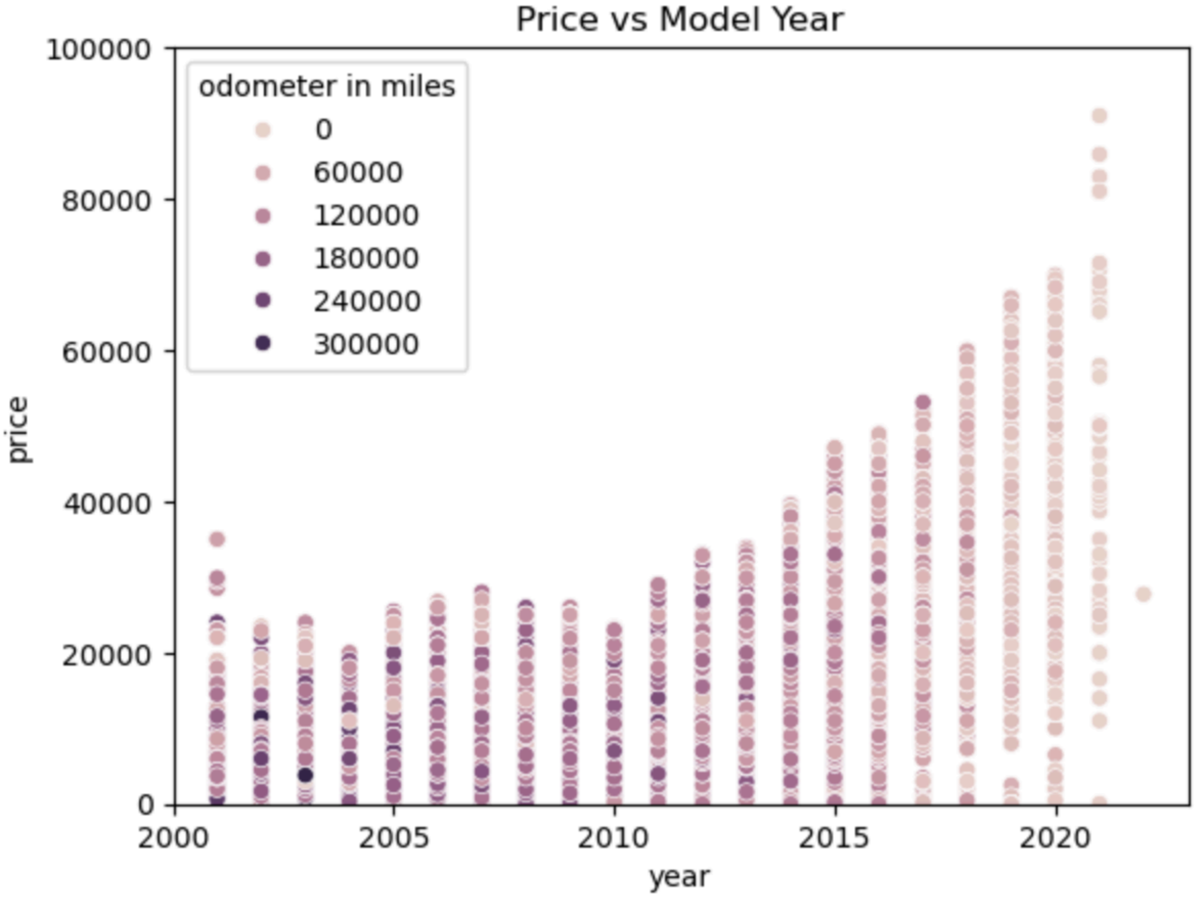
- Model Year - Sourcing late model year used vehicles as this has a large positive impact on pricing



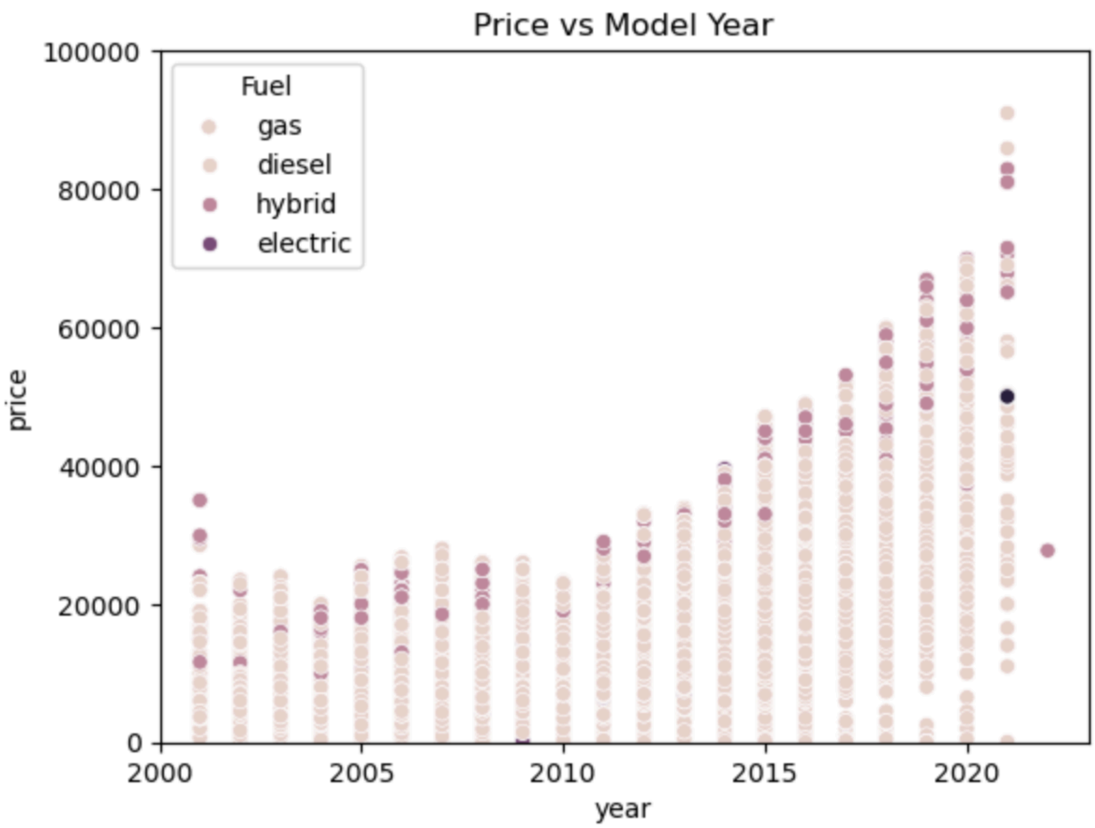
- Engine Cylinder Size - Sourcing larger cylinder engine cars, typically meaning high performing vehicles with higher horsepower and torque, will result in higher selling prices


Price vs Model Year
price
100000
80000
60000
40000
20000
0
2000
2005
2010
2015
2020
year
Cylinders
3
4
5
6
8
10
12

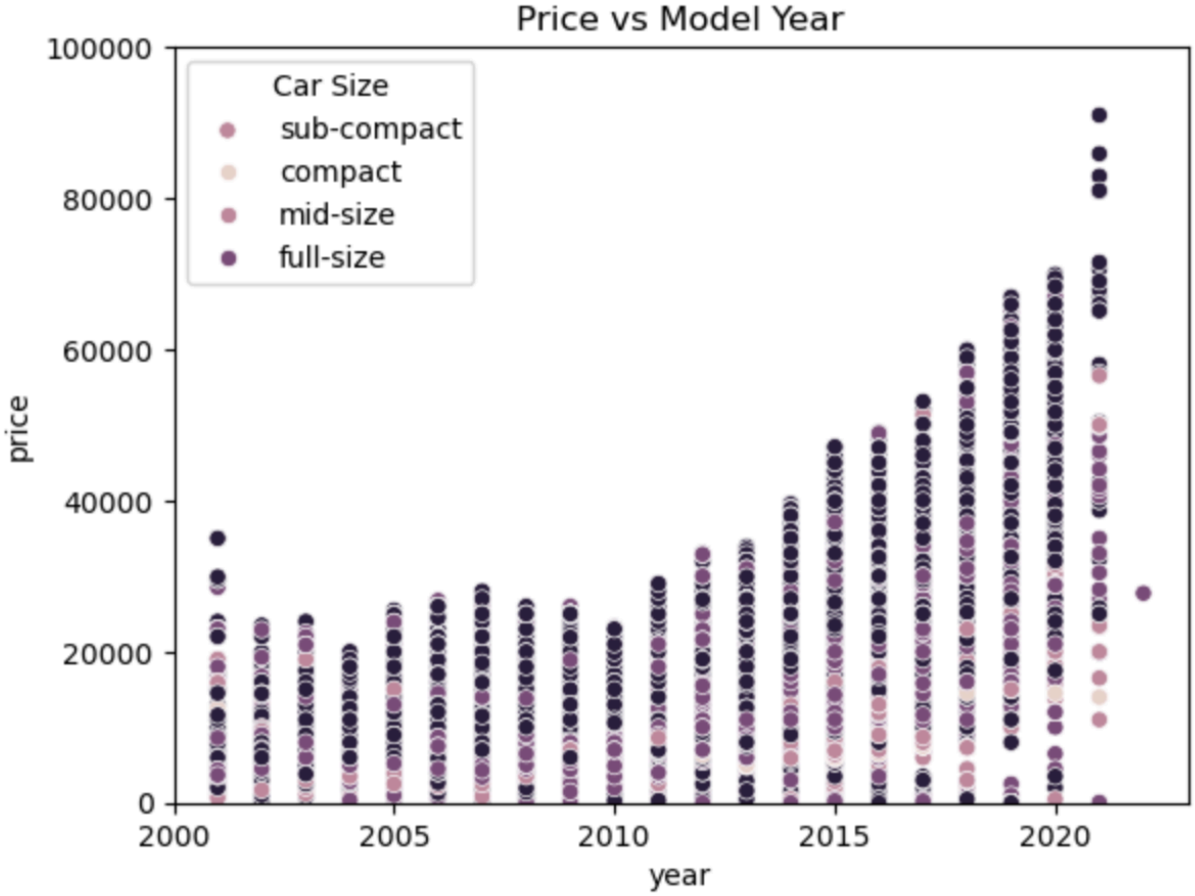
- Odometer - Sourcing low mileage vehicles is recommend as the higher the mileage, the lower the price



- Car Fuel - Sourcing alternative fuel cars such as hybrids, electric cars along with higher mileage diesel cars tend to have a higher value over gas counterparts



- Car Size - Sourcing large cars is advisable, full-size cars sell at a higher price than sub-compact counterparts

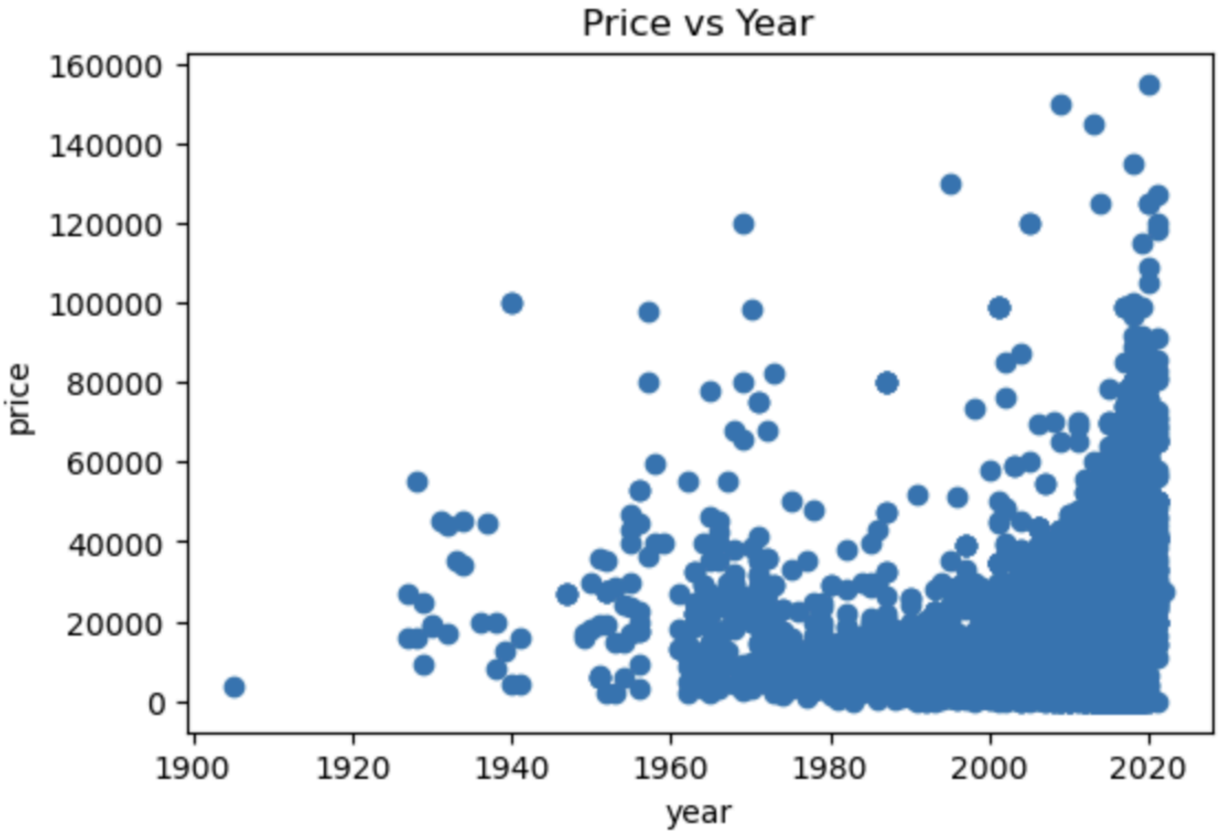


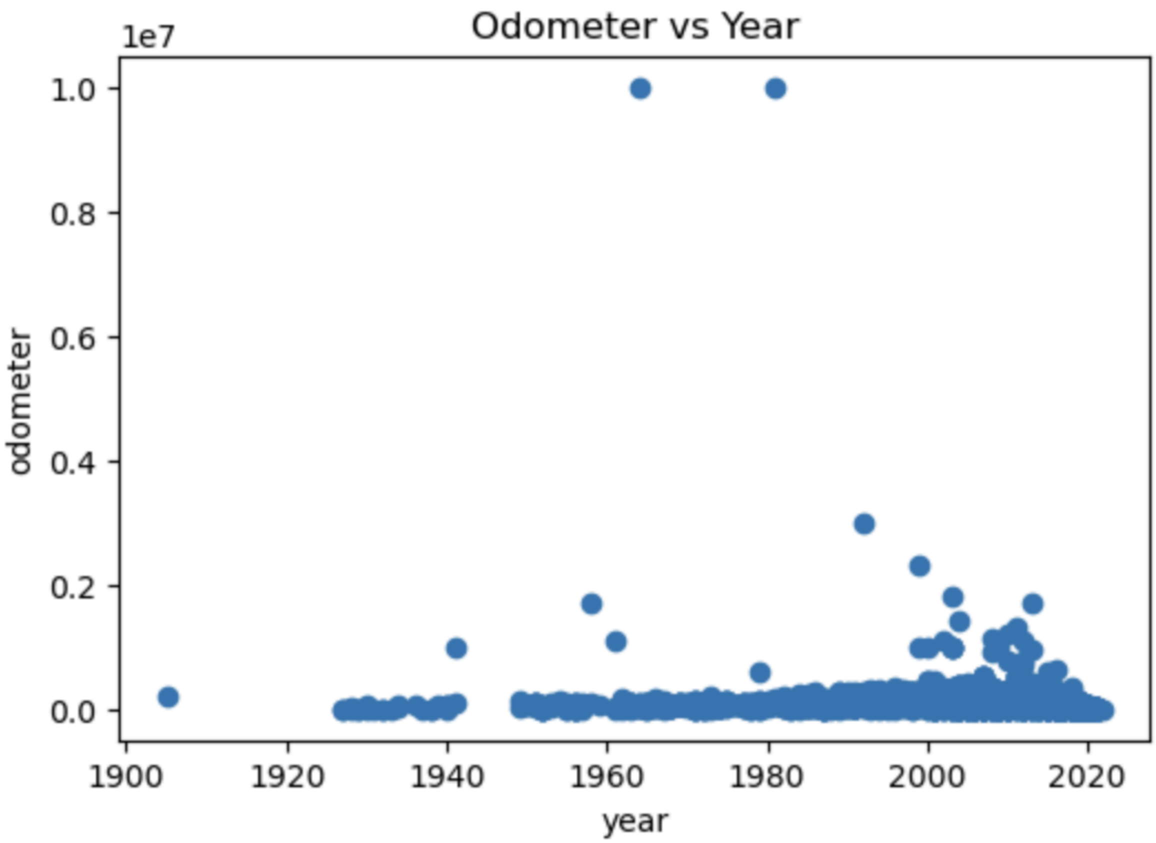
Methodology

Data Understanding

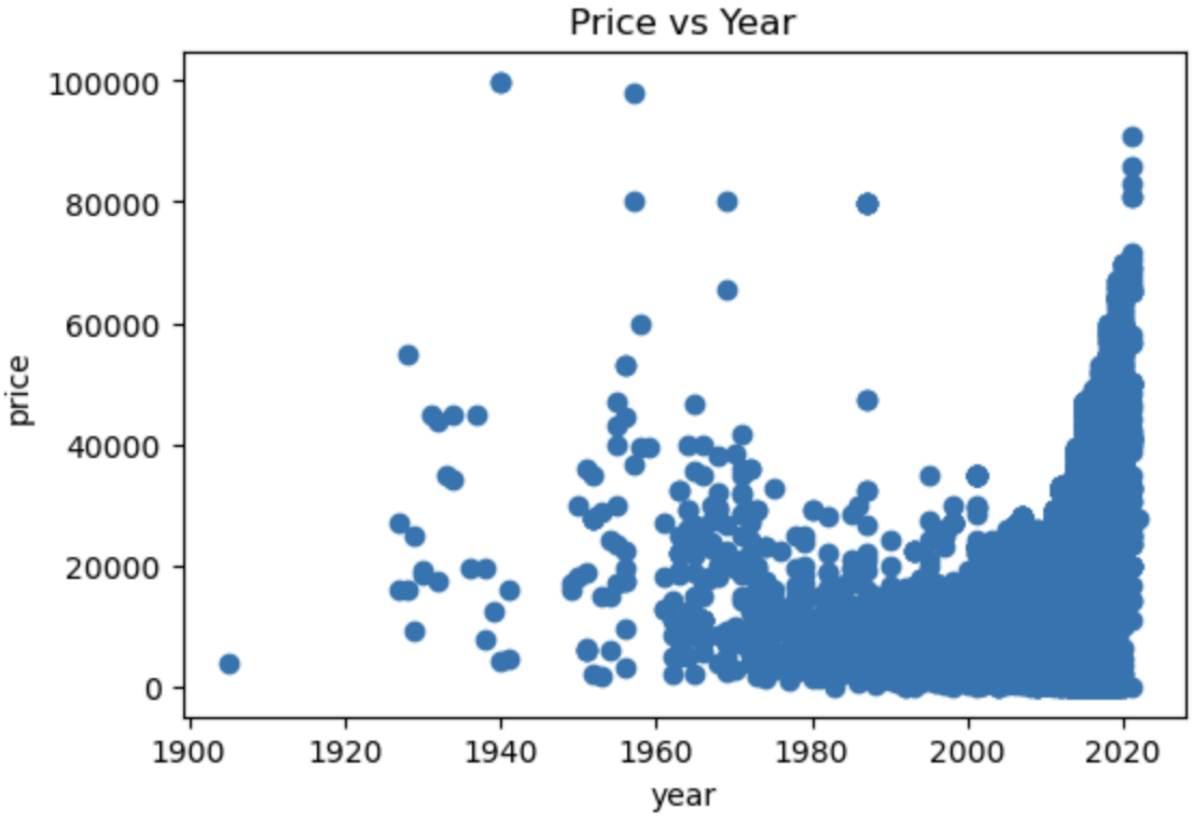
In order to run my analysis, I spent considerable time understanding the raw underlying dataset. Using value_counts was extremely helpful in understanding the various values associated with the columns of the used car data set. It immediately became apparent that there is a lot of qualitative data that requires transformation. There is also data that would be very onerous to transform to numerical data such as the region of the car and the specific model of the car and those were flagged for consideration to drop from the data set.

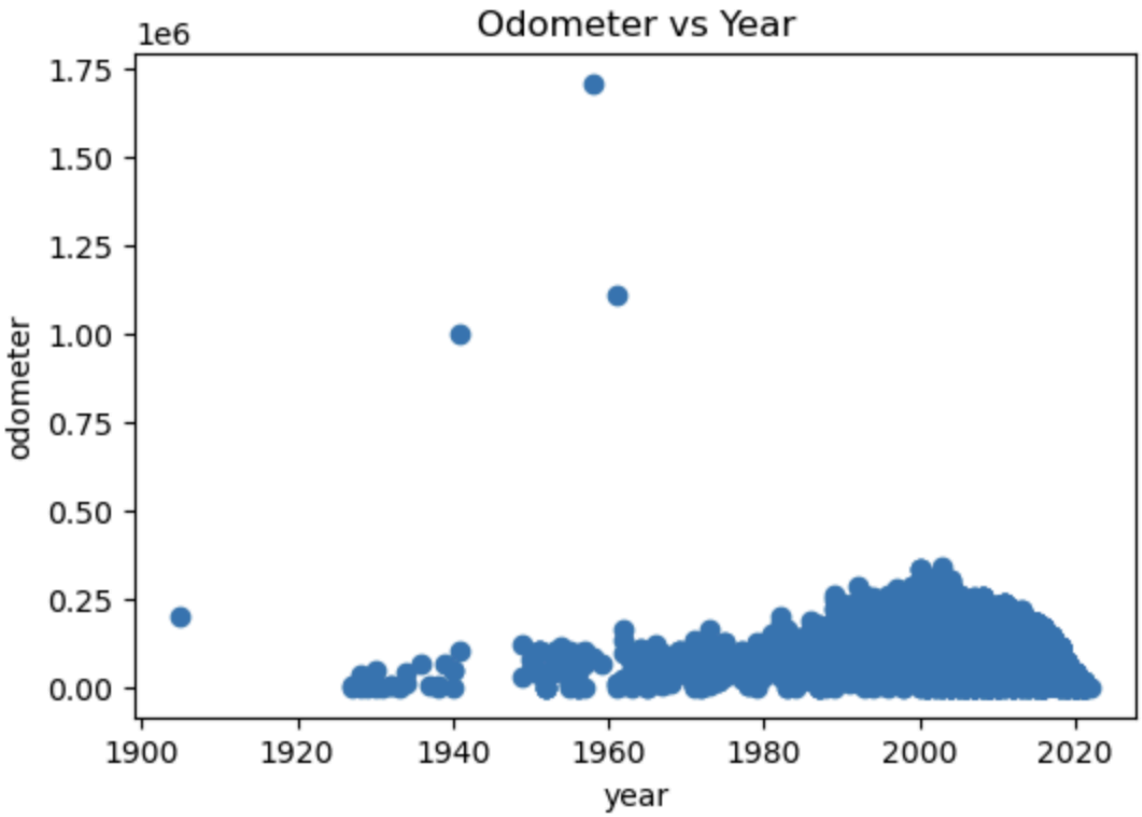
I created some scatterplots to understand the shape of the price data vs year of the automobile and the same for odometer vs year and quickly saw outliers.


Price vs Year
160000
140000
120000
100000
80000
60000
40000
20000
0
price
1900
1920
1940
1960
1980
2000
2020
year

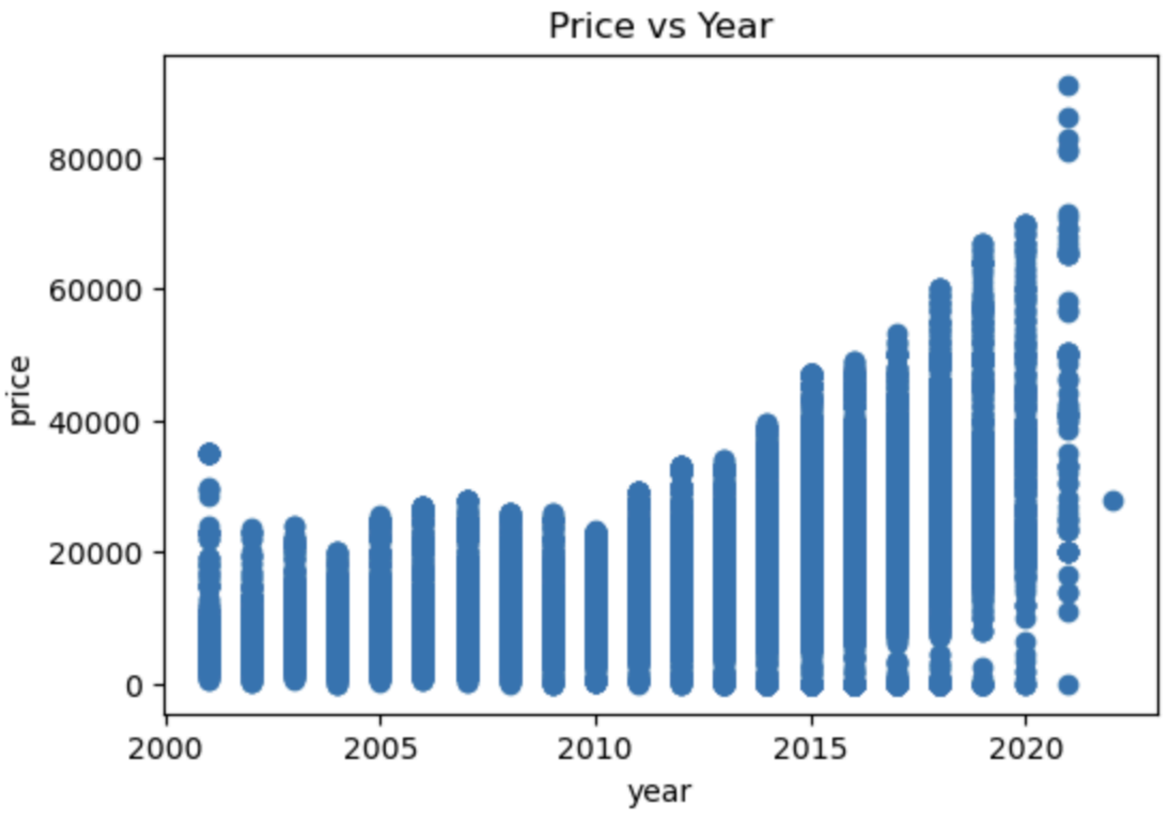


For each year, I decided to remove data for cars that had a price and odometer reading that exceeded two standard deviations from the mean for that year.


Price vs Year
price
100000
80000
60000
40000
20000
0
1900
1920
1940
1960
1980
2000
2020
year



This helped to clean up the data a bit, but it was still pretty noisy for years prior to 2000, especially for the ‘Price vs Year’ graph, no doubt due to exotic/classic cars being part of the data set. Removing data prior to 2001 really helped to clean up the data set.



Data Preparation

Technically, some amount of data preparation was already completed above (elimination of two standard deviations above mean for price and odometer and filtering for data after 2001) by the time I started to even further prepare the data. The biggest task in the data preparation phase was to map the qualitative data to numerical values. I tried to do this in a manner for each parameter with the lower number being the likely low price value and the higher one being the higher value, for example for cylinders, less cylinders had a lower numerical value, high cylinders had a higher numerical value, or full-size car had the higher numerical value vs sub-compact being the lower numerical value.

I also wanted to scale my data set because odometer readings would be orders of magnitude larger than any other mapped parameter. I used the **MinMaxScaler** approach to scale values to between 0 and 1. I also created a train/test split using the **train_test_split** function

Modeling

I ran a total of 3 models for this analysis. My X input was

X Field	X Field Descriptor	Transformation(s) Applied

year	Vehicle Year	MinMaxScaler
odometer	Vehicle Odometer Reading	MinMaxScaler
type_code	Vehicle Type	MinMaxScaler, Type numerical mapping *
size_code	Vehicle Size	MinMaxScaler, Size numerical mapping *
condition_code	Vehicle Condition	MinMaxScaler, Condition numerical mapping *
fuel_code	Vehicle Fuel Type	MinMaxScaler, Fuel numerical mapping *
cylinder_code	Vehicle Cylinders	MinMaxScaler, Cylinder numerical mapping *
title_code	Vehicle Title	MinMaxScaler, Title numerical mapping *
paint_code	Vehicle Paint Color	MinMaxScaler, Paint numerical mapping *
state_code	Vehicle State	MinMaxScaler, State numerical mapping *
manufacturer_code	Vehicle Manufacturer	MinMaxScaler, Manufactuer numerical mapping *

* see workbook for details of the mapping

And my Y output was

Y Field	**Y Field Descriptor**	**Transformation(s) Applied**
price	Vehicle Price	MinMaxScaler

Scaled Linear Regression Model

For the linear regression model I ran LinearRegression() on the X_train and y_train data, ran predictions on the X_train and X_test data, and computed mean standard errors using the mean_squared_error on both the training predicted values and the test predicted values.

Scaled Polynomial Model

For the polynomial model, I created a pipeline that ran with PolynomialFeatures that iterated from degree 1 in 2 step increments to 9 and Ridge set to alpha = 1 (effectively a linear regression on polynomial data). I ran it on the X_train and y_train data, ran predictions on the X_train and X_test data, and computed mean standard errors using the mean_squared_error on both the training predicted values and the test predicted values.

Scaled Ridge Polynomial Model

For the scaled ridge polynomial model I created a Ridge Pipeline and set a parameter containing multiple alpha values initially set to a range of 0.1 to 100 with increments of 10x. After a few trial and error runs, the model converged to an optimal value for alpha of 41.1 based off the 'rank_test_score' parameter and its offset into the ‘params’ array. This set did not need a training or test data set, instead I ran it over the entire data set, dropping price for the X values, and price for the Y values.

Evaluation

For each run, I looked at the lowest test MSE value to evaluate which model performed best since it would indicated better model performance compared to unseen data.

Scaled Linear Regression Model

The linear regression model had training and test MSEs that were in pretty close proximity to each other, indicating that the modeling did not overfit the training data.

```
training mse is  0.006532390659139967
test mse is  0.006707181010956457
```

Scaled Polynomial Model

The scaled polynomial model had training and test MSEs that were in pretty close proximity to each other, indicating that the modeling did not overfit the training data and the model performance for the test MSE kept improving as it stepped from degree 1, 3, 5, 7 and 9. The runs began to take a long time so I was unable to draw a conclusion on the most effective model when alpha was set to 1

```
training mse is  [0.006532409452755361, 0.0037617290035583087,
0.0032764962672685374, 0.002910872309874212, 0.0026088447043674046]
test mse is  [0.006707586737730332, 0.003919616658322686, 0.00354342786587832,
0.003316024046471655, 0.0031673219645035173]
```

Scaled Ridge Polynomial Model

For the scaled ridge polynomial model, it has the worst MSE performance, examining its mean_test_score array and taking the array's mean

```
Mean_test_score array: [0.5313442 , 0.5313442 , 0.5313442 , 0.53134418]
mean of array is 0.5313441941879545
```

Deployment and Conclusion

The scaled polynomial model performed the best on the test MSE front having the lowest value of all models but I couldn't converge on an optimal for the chosen alpha as the model kept improving for higher polynomial values and the runs began to take quite a long time to complete. I thus returned to the scaled linear regression model with the second lowest test

MSE value and compared the top 5 parameters for that model versus the optimal scaled ridge regression model and both models gave the same insights. The top 5 drivers of used car pricing were Model Year, Engine Cylinder Size, Odometer, Car Fuel and Car Size.

Additional Factors

Additional factors that influence used car pricing include

- Car Title - Having a clean car title has a positive impact on the value of the car, specifically we recommend a clean title.
- Car Condition - The car condition has a positive impact on the value of the car, specifically we recommend new, like new, or excellent condition cars.
- Car Type - The car type has a positive impact on the value of the car. Larger cars such as SUVs, mini-vans and vans will have higher resale values.
- Car Manufacturer - Interestingly, marquee/luxury manufacturers tended to have a negative impact on pricing. This is consistent with depreciation being very high for luxury cars.
- State - it was hard to draw conclusions on the impact of state on pricing as the mapping tried to map higher value resale states to higher numerical values but still we got a slightly negative impact from the scaled linear regression model
- Paint - certain color paints (white or black) tended to have a positive impact on pricing, but it was slight.

Resources

Link to notebook: · link to the notebook