Project 5 - Used Car Prediction

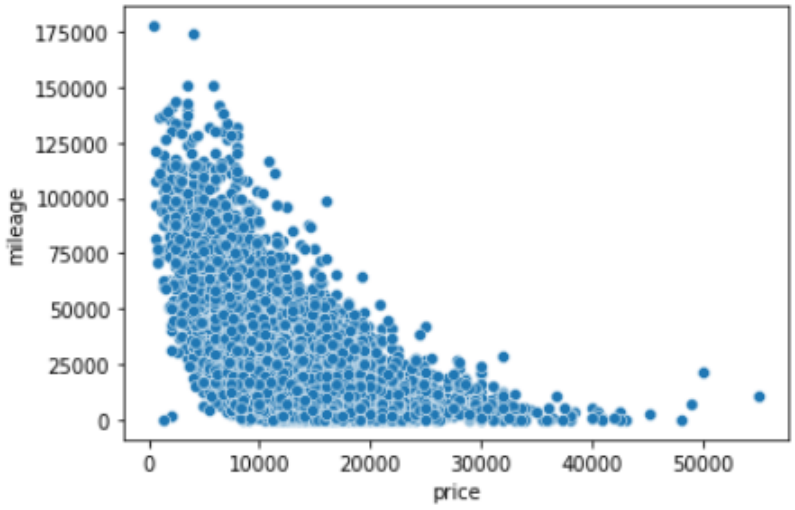
Introduction

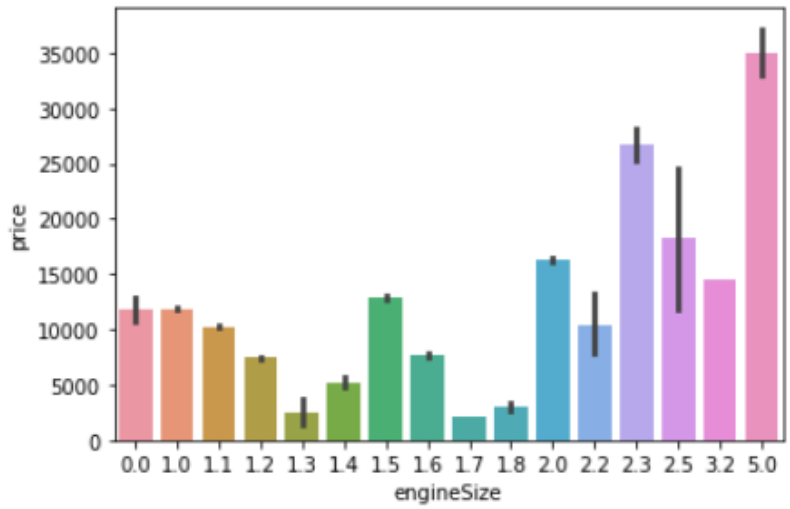
Since the pandemic began supply issues have caused massive issues for many industries. One industry that has suffered greatly in the news has been the automobile industry. Due to a semiconductor shortage manufacturers have not been able to complete production on new cars. This caused a cascade effect on the market and since there was a lack of supply in new cars many looked to the used car market. The sharp rise in demand caused used car prices to skyrocket leading to some people making a profit on their used cars. We want to use linear regression to try and predict the car market and also use it to tell just how inflated the market has gotten. This could be used by those looking to purchase a car to see what vehicles are suffering the least from this rapid rise in price. Based on prior knowledge of the car industry there are a couple of features that we would look at to be important for predictions of the industry. The main feature would be mileage, which is how many miles that a car has driven over the time of its creation. This normally plays a huge role because the higher the mileage, the older the car is and the closer it is to being scrapped. Another feature that may play a crucial role in the price is the engine size. Those with a larger engine size are normally worth more than those with a smaller

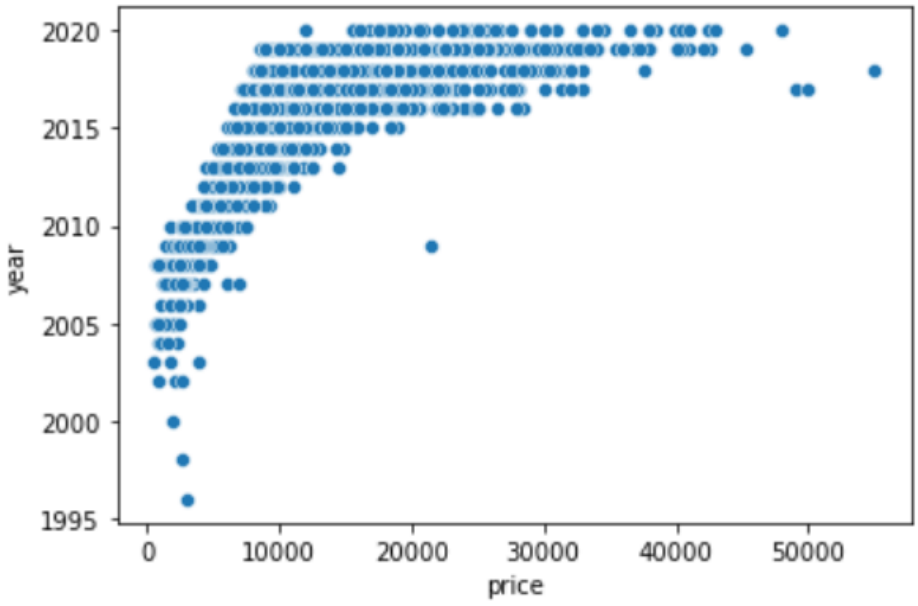
engine. The goal of the project in the end is to be able to predict the price of the used car industry.

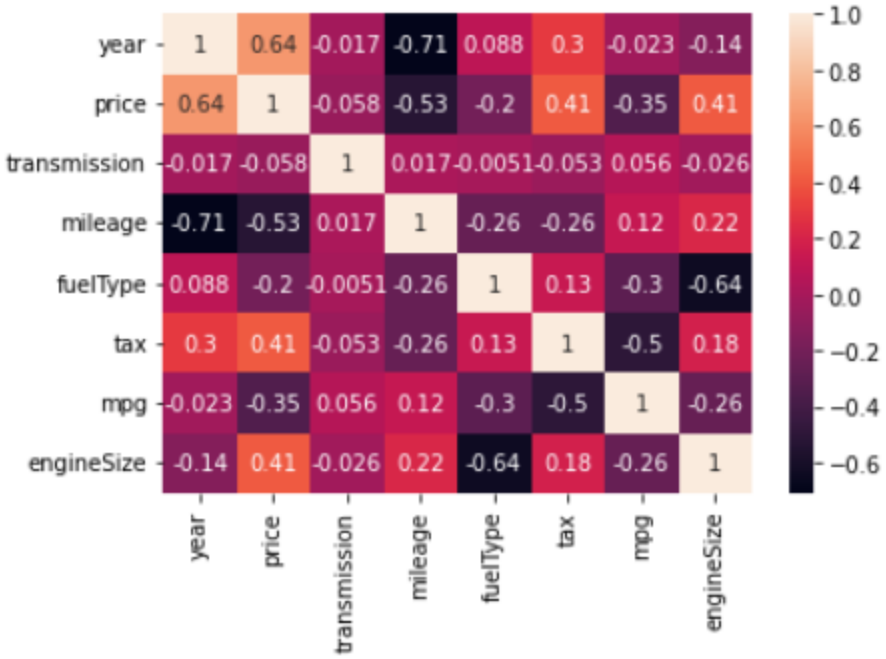
About the data

The data that we are going to be working with comes from kaggle.com and is hyperlinked for your convenience. The data comes from a Ford used car dataset, so the end result will be based off of the Ford brand. The data has features that include model, year, price in euros, transmission, mileage, fuel type, tax, miles per gallon, and engine size. The data also has the size of 910.94 kB. Below are visualizations that may be useful to help understand the data better.




price
35000
30000
25000
20000
15000
10000
5000
0
0.0 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 2.0 2.2 2.3 2.5 3.2 5.0
engineSize


year
2020
2015
2010
2005
2000
1995
0
10000
20000
30000
40000
50000
price



Methods

In our preprocessing we checked for null values to exclude, and we also checked for duplicate values in order to shrink the set down some. After we had removed the null values and duplicates we then changed any non numerical values into integers in order to perform analysis on them. Once we had finished cleaning the data up a little bit we made a heat map as

shown above to begin to check what columns may correlate, In order to decide which columns we would use in our linear regression analysis.

For the first model we decided to use all of the columns and see what the result was.

Evaluation

In our first model, we have an R-Square value of 0.7343 which indicates that about 73 percent of the variance in price of the ford vehicles can be attributed to the columns found in the data set after cleaning the data. From the predictions we can see that the model can predict the price of a vehicle within a few thousand dollars.

In [38]:
```
y_pred = lr.predict(x_test)
```

In [40]:
```
lr_diff = pd.DataFrame({'Actual Value':y_test, 'Predicted Value':y_pred})
lr_diff.head()
```

Out[40]:

	Actual Value	Predicted Value
12532	15970	17731.733822
13512	19499	16856.963324
7454	8100	11950.005303
9668	10178	12292.487596
10397	8750	9382.359381

One thing that must be considered when looking at the price of vehicles is that it is not just as cut and dry as this is what the manufacturers state the price is and then factor in depreciation for use. There are many market factors that are in play when it comes to vehicles. This may be a factor in how close to 1 we can get in our R-Squared value. For example, the supply shortage in new vehicles causing used vehicles prices to rise as supply and demand trends rapidly change cannot be accounted for with data that tells us more about the vehicle. We would need to have data on the amount of people who bought and sold vehicles over time to try and predict those trends and place that into our model.

Storytelling and Conclusion

Based on our findings it is evident that the price of used cars is affected by a wide range of factors. As expected, mileage and year were the two most impactful variables in determining the price of a used car. Most used car shoppers would agree that the age and mileage of a car are very important to their decision, but there are 3 other variables that have a bigger impact than might be expected. Tax, MPG, and engine size all had a significant impact on the price that a used car sells for.

Despite the small number of variables that were used to train our model we were able to predict car values to within a couple thousand dollars of the actual sell price. This model is useful for owner's of used cars to determine

roughly how valuable their vehicle is. It could also be useful information for someone looking to buy a used car. This model can help the user understand the tradeoff between the age of the car and the price and other factors. One thing our model does not take into account is the change in demand due to external factors. As we discussed covid-19 caused complete upheaval in the automotive industry and resulted in used cars being worth much more than normal. This kind of rapid change is difficult to model.

It is also important to evaluate the model to further understand the reasoning behind consumer listing prices, and see if there are any noticeable factors that caused the model to either predict higher or lower than the listing price. Here we added the predicted prices back to the original dataframe, removed the null predictions, and also created a price parity column to see how far off the model was.

```
df['predicted_price'] = lr_diff['Predicted Value']
```

```
df = df[df['predicted_price'].notna()]
```

```
indexNames = df[df['predicted_price'] < 200 ].index
df.drop(indexNames , inplace=True)
```

```
df['price_parity'] = df['predicted_price'] - df['price']
```

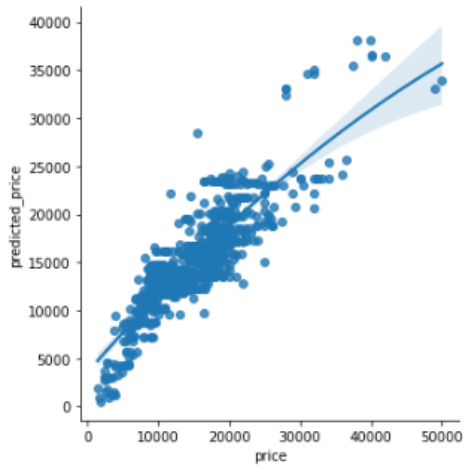
```
df.head()
```

	model	year	price	transmission	mileage	fuelType	tax	mpg	engineSize	predicted_price	price_parity
3	5	2019	17500	1	10460	4	145	40.3	1.5	18243.813564	743.813564
23	13	2014	8995	1	59000	0	160	47.9	2.0	12841.850641	3846.850641
61	13	2018	17998	1	12162	4	145	39.2	1.5	17089.743550	-908.256450
63	6	2019	17498	1	2714	4	150	49.6	1.0	13510.102132	-3987.897868
75	13	2017	17920	0	15815	4	235	38.2	1.5	15904.109372	-2015.890628

As we will discuss in the impact section later, it appears that the higher value cars have more price parity on the upside. These higher value cars are typically the higher mpg mustangs or trucks.

```
sns.lmplot(
    data = df,
    x = "price",
    y = "predicted_price",
    order = 2
)
```

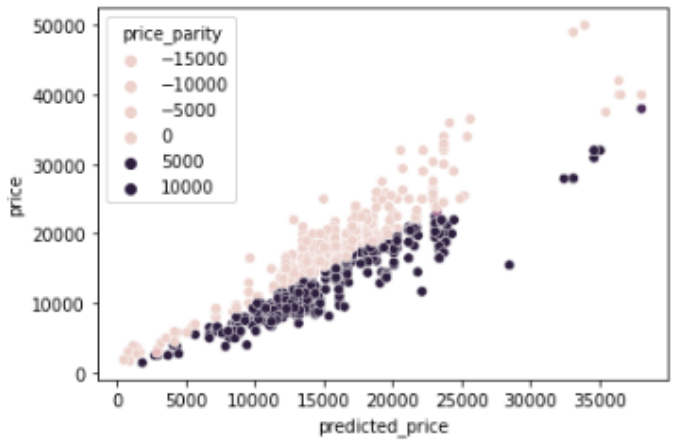
```
<seaborn.axisgrid.FacetGrid at 0x2c984a991f0>
```



The model that predicted prices lower than the listing price almost all ranged around the fifteen to twenty-five thousand dollar range. When it comes to the value of the larger trucks and sports cars the model still vast majority ofthem are listed for way higher than the model predicted.

```
sns.scatterplot(x="predicted_price", y="price", hue="price_parity", hue_norm=(0,100), data=df)
```

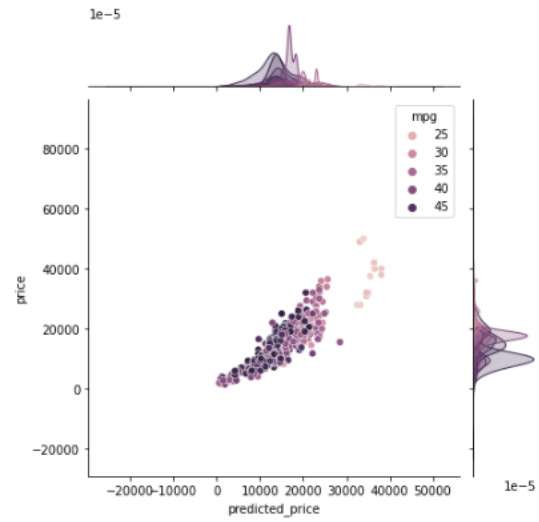
```
<AxesSubplot:xlabel='predicted_price', ylabel='price'>
```



The most overpriced models were the Focus, Edge, and the Ford escort. These are the international models and have a higher supply of newer models. The linear regression model deemed these cars to all be overpriced and did it consistently. The most underpriced models (listed by sellers) were the Ford mustangs and Trucks.

```
: sns.jointplot(x='predicted_price', y='price', data=df,   hue='mpg')
```

```
: <seaborn.axisgrid.JointGrid at 0x2c98c108250>
```



Impact

The economic impact that this could have is something that needs to be considered. The increase in the used car industry has been a factor in terms of the inflation that has been a part of this country recently according to cnbc. Also, when you think about the rise in used car prices, it makes you think about the people who are in the market for a car. If someone is looking for a car but is not looking to spend the money on a new car, if available, they are going to look towards buying a used one. It is what many peoples first cars are, but what is the point if the price of a used car is similar to those of a new car? This is something that could plague those who are not able to afford these prices. There is also a flipped side though in this industry. Those who are selling their cars could benefit from this trend. In the introduction, it was discussed that people could benefit from being able to make money off of their cars by selling them.

Ford's price distribution is heavily influenced by the engine size. This is because of their F150 line being the most popular truck in America. Over the past couple of years ford cut production of new trucks including their newly released ford bronco. Light trucks makeup 57.2% of total U.S. vehicles in operation in 2021, in 2018 the share was 56.0%. Even with the recent

increase in oil prices, the demand for trucks by Americans continues to increase, especially as auto makers create more fuel-efficient large vehicles like trucks and SUVs. This is interesting because even though small cars sales numbers are higher than trucks, the demand for larger engines vehicles continues to rise. This leads me to believe that Americans aren't valuing their cars by their effect on the environment, therefore the benefits of a more fuel efficient car do not affect the overall price in the used car market. Ford is unique in the automotive sector because of this.

Ford recently cut production of all their sedan models, leaving only the mustang as the lone sedan to continue production. This is interesting because when analyzing used Fords, their prices are not correlated to their MPG. As we discussed earlier, the main reason for the increase in used cars was due to the cutting of production of the newer models. The lower MPG the more expensive a lot of fords are. And the regression models we used to predict prices were not as accurate when using mpg as a predictor. This leaves Ford in a unique situation as they try to navigate this increase in demand for trucks throughout this shortage.

Code

LeifAndersenGH/ITCS3162-GROUP-PROJECT (github.com)