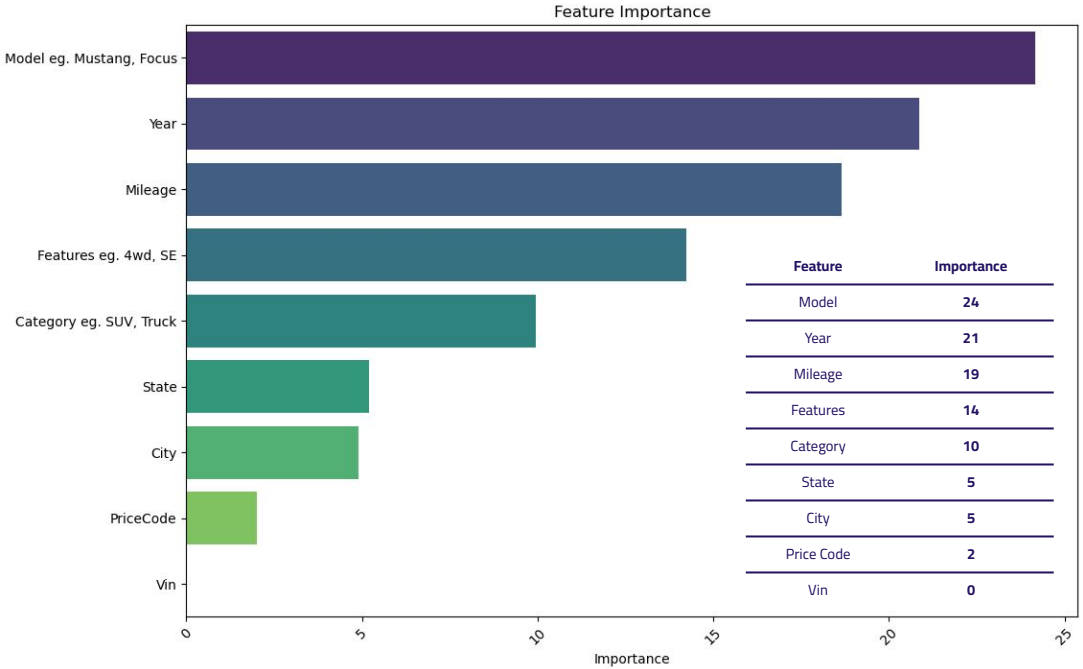
Synthetic Data Analysis

Angus J. McLean

Executive Summary

What variables appear to be most associated with resale price of used Ford cars?



Feature	Importance
Model	24
Year	21
Mileage	19
Features	14
Category	10
State	5
City	5
Price Code	2
Vin	0

A CatBoost Regression found that Model was by far the biggest predictor of resale price, followed by Mileage and Year.

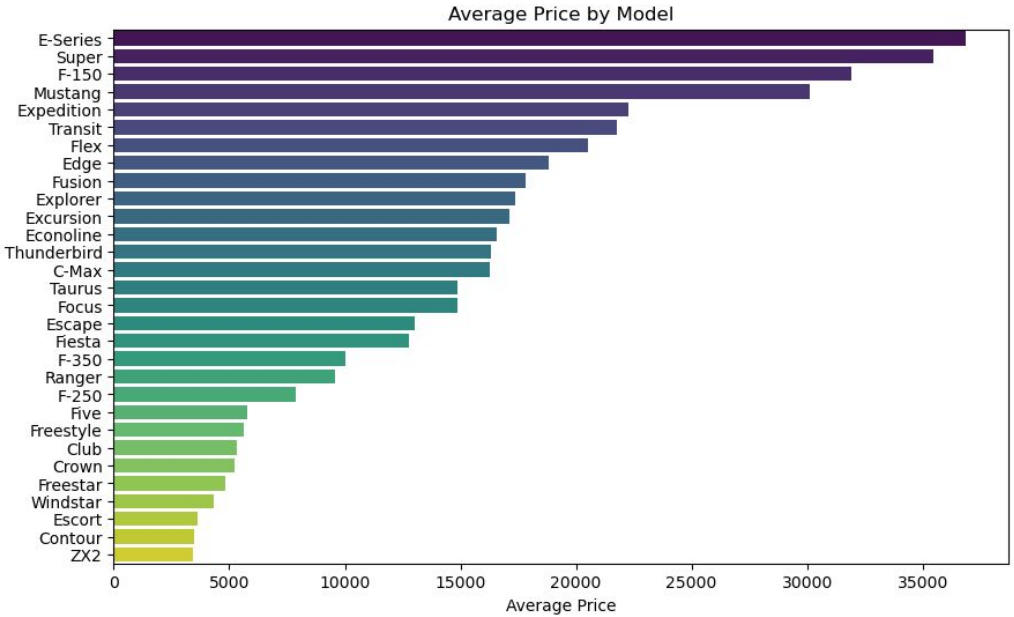
The model could predict the price within **$2324.**

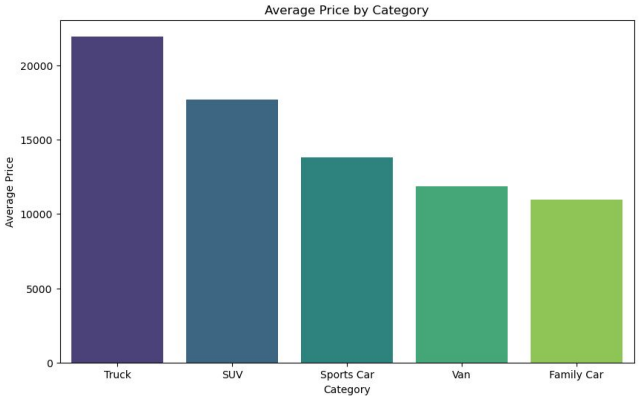
Overall the model performed well, with an 88.5% of the variance in the target variable being explained by the model. (R2)

However, when there were errors, they were large. This is likely due to the large price variance in newer models. (MSE)

Model affects the price of used Ford.

Model is the most important predictor of price.



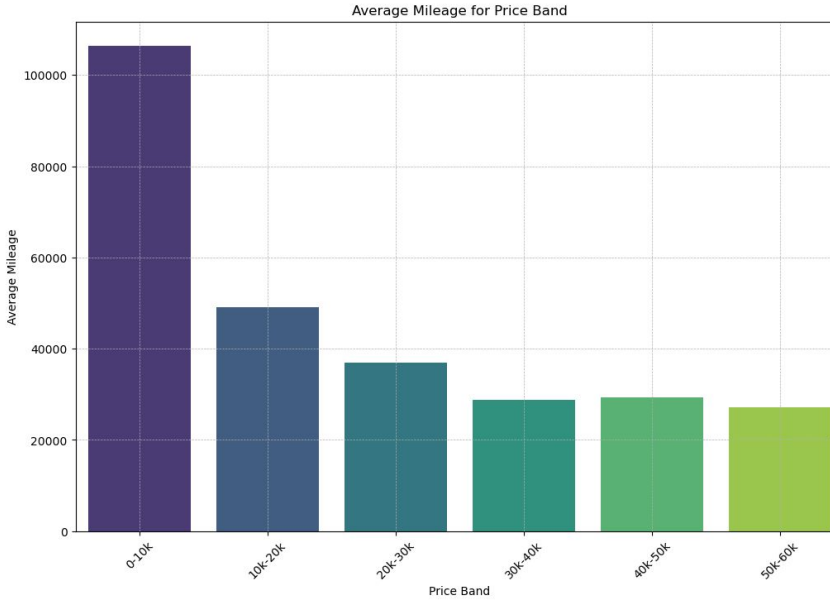


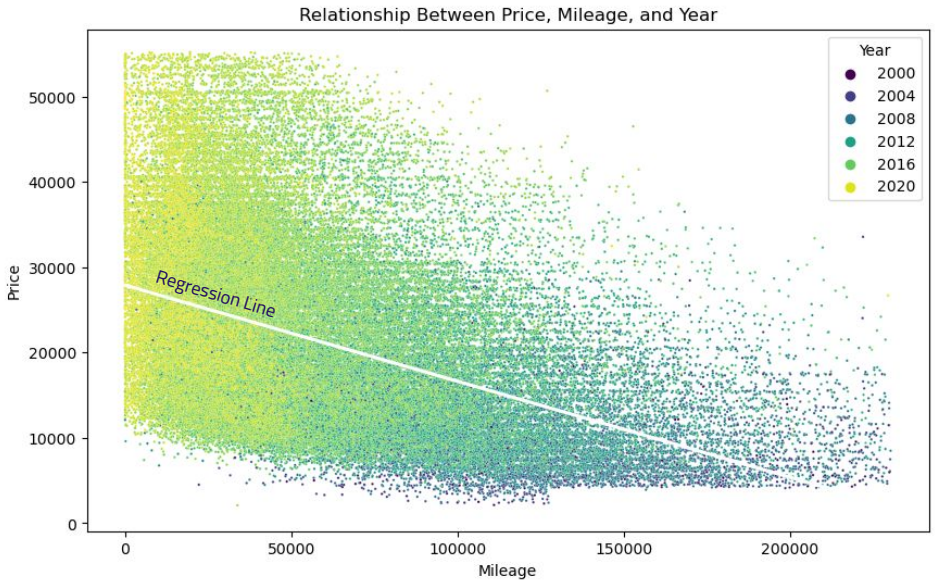
The most expensive category was Trucks, followed by SUV's.

The most popular subcategories were 4 Wheel Drive (20.6k), Special Edition (9.8k), 4 Door (7.9k), 2 Wheel Drive (6.3k) and Sedan (5.7k).

Mileage affects the price of used Ford.

For a used Ford to sell for over 20k, the mileage would typically be under 40k.



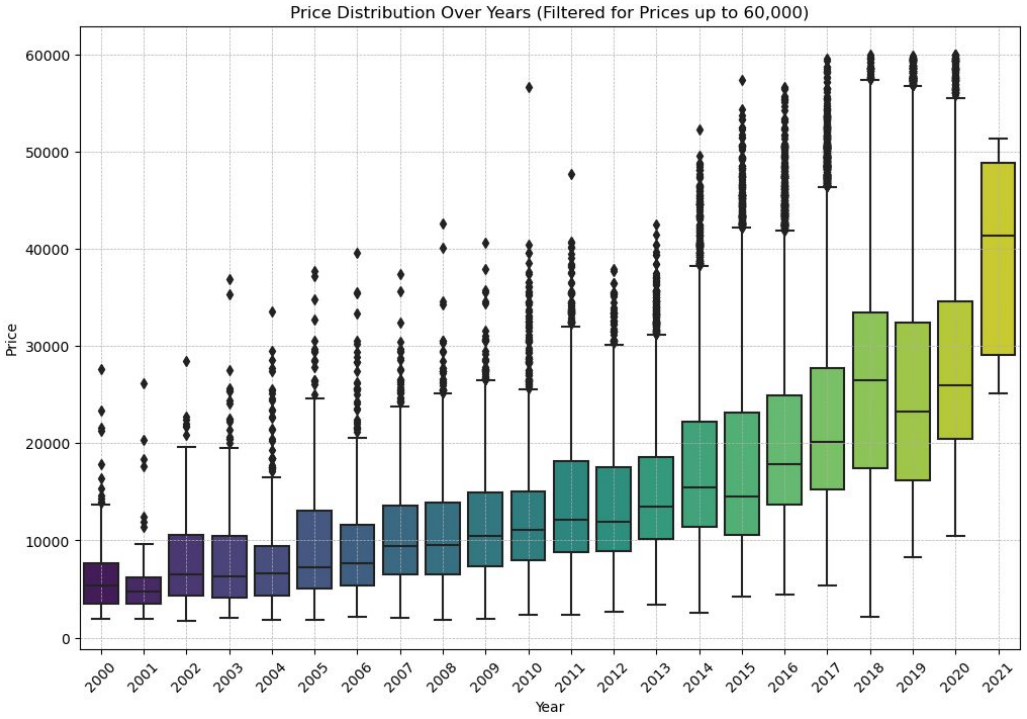


If the mileage is over 50k, the price is typically under 20k.

Younger cars tend to have higher prices and lower mileages.

[Milage and Years are closely correlated which is likely why they have similar Feature Importance.]

Age affects the price of used Ford.

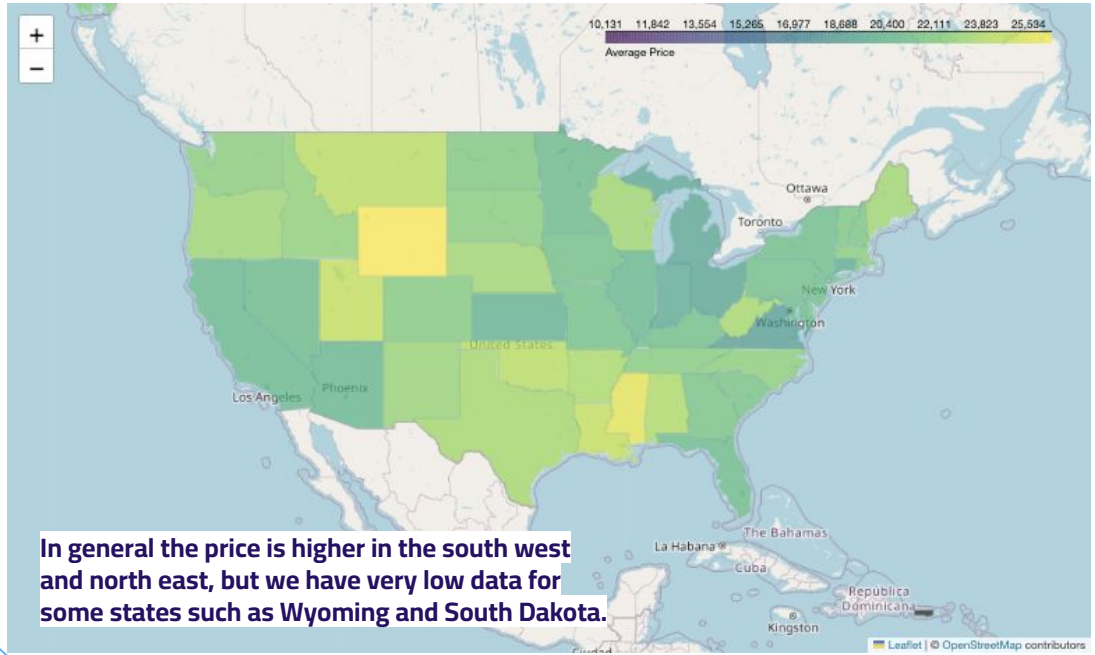


As the age of a Ford *increases*, its price *decreases*.

However, as a Ford gets older, its price becomes more predictable.

The younger the Ford, the wider the possible price range.

How do resale prices of used Ford vary across US states?

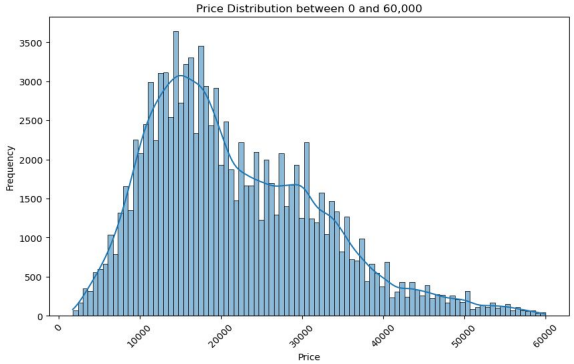


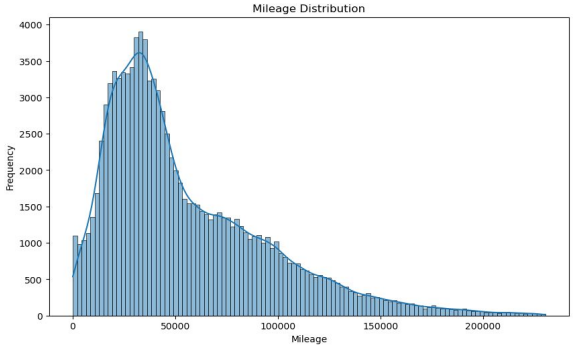
In general the price is higher in the south west and north east, but we have very low data for some states such as Wyoming and South Dakota.

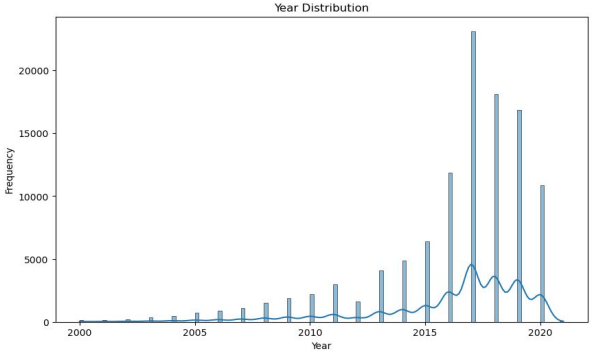
	State	Price	Count
1	Wyoming	26390	101
2	Mississippi	25657	1307
3	Utah	24602	1714
4	Louisiana	24583	979
5	Montana	24262	385
6	Oklahoma	24185	1583
7	Alabama	23851	1861
8	Arkansas	23650	763
9	West Virginia	23574	374
10	Texas	23477	122410
11	Nebraska	23335	908
12	Maine	23326	398
13	Wisconsin	23232	2175
14	Oregon	23188	1537
15	New Mexico	22978	763
16	Idaho	22725	641
17	North Carolina	22473	5263
18	Washington	22437	3198
19	Massachusetts	22388	2130
20	South Dakota	22365	171

Dataset

Here we can see the shape of our cleaned dataset:







Mean price is $21’654 with a standard deviation of $10’247.
Prices range from $2’106 to $55’222.
Ages range from the oldest, from 2000, to the newest from 2021
On average the cars were 7 years old (2017)*
On average mileage is 55’504 miles with a standard deviation of 40’068 miles.
Most Common Model was the F-1504WD.

	Price	Year	Mileage
mean	21654.0	2016.0	55504.0
std	10247.0	3.0	40068.0
min	2106.0	2000.0	5.0
25%	13726.0	2015.0	26396.0
50%	19435.0	2017.0	42971.0
75%	28521.0	2018.0	77242.0
max	55222.0	2021.0	230329.0

Quality Issues (1)

There were numerous problems with the data:

The first problems I encountered the null and negative values for **mileage**, there were also some extreme mileage values, that while plausible (1m+) was best to remove. My initial thought was to use a catboost to impute the missing values, which I tried and it appeared accurate (within 600m). As many of the negative miles seemed like plausible positive values, and because a car can't have negative values I converted these to positive. *However, going back over the dataset, I decided to alter my approach and went with IQR which removed 2% of the data, this also had the effect of limiting the max price.*

The second problem was that the data was missing the **state** column, so we couldn't differentiate between cities, this was solved with a left join using 'Vin' as the key. However, there were still duplicates that needed to be grouped. It would also be useful to have population values, so we could see them in relation to count per state.

This was not a quality issue, but after I looked up one or two of the **'Vin'** numbers up online, they to be fake which meant we couldn't parse them into more meaningful units - this was confirmed in the Catboost where feature importance = 0.

Quality Issues (2)

The **model** column was concatenated, there were a number of different approaches we could have taken here, but as there were only 150 values and complex strings I fed them into an LLM to correctly break them up into Main Categories and Subcategories. I did this using the OpenAI Playground, as the data is not used to train models, and is therefore safe to use with proprietary data, however, this should be confirmed at an organisational and client level - I could have also used the API if I had more (or more complicated) requests. I then hand checked the results.

The **price** column was also concatenated, I resolved this by splitting the column into 'Currency', 'New Price', and 'Price Code', after then identifying the prices were all in dollars I removed 'Currency', but retained the 'Price Code' as I thought it may contain some yet unidentifiable but relevant information.

The **state** names appeared to contain repetitions, and inconsistencies in capitalisation, because my GeoJSON contained the full state names, eg. California, I used a dictionary to convert these, this still left duplicates, which I could then group, being sure to combine the means correctly when merging. Next time I might correct the capitalization first then merge before calculating the mean, but initially I wasn't sure if the capitalisation was for differentiation.

We don't have any data for the original prices, it would be great if we could get the original price data so we could calculate depreciation adjusted for inflation, as well as date of resale.