USED CAR PRICE PREDICTION
TOYOTA & FORD

GROUP 2: Joseph Onwukeme, Matthew Witschorke, Sam Slomowitz, Yarely Vargas

ABSTRACT

The purpose of our study is to help others determine the price of a used car based on the market price. The data, obtained from Kaggle, consists of web-scraped data on 100,000 used cars in 2020. We implemented a multiple linear regression model correcting for multicollinearity to predict the price based on a car's model, mileage, miles per gallon (mpg), engine size, fuel type, and transmission type. In this study, we focused on Toyota models, Aygo and Yaris, and Ford models, Fiesta and Focus, as they represented samples of data for Toyota and Ford. Further, we hypothesized that year and mpg would have a positive correlation with price while engine size and mileage would have negative correlations. We concluded that a prospective buyer should purchase a 2017 car for £11,000 with an engine size of 1.3.

INTRODUCTION

A research team from the University of Virginia has established that mileage drives the prices of used cars (Engers et al., 2009). A team of data scientists exploring a similar dataset on Kaggle, but on the German used car market, performed a Random Forest Analysis and determined that mileage, brand, and vehicle type were the primary predictors for the price (Pal et al., 2018). Further, Neural Network modeling techniques have been applied to the Toyota Avanza used cars in Medan, Indonesia with promising results for future machine learning investigations (Syahputra et al., 2019). However, one machine learning algorithm performs worse than an ensemble, of say, three machine learning algorithms (Gegic et al., 2019). In addition, one model may be optimized for one brand of cars while another model may be ideal for another brand of cars. Our study is most likely the first in literature to use this Kaggle dataset to predict the price based on simple and multiple linear regression via the Ordinary

Least Square regression model and write an analysis of the strengths and weaknesses of the implemented model.

OUR DATA STORY

We proposed that we are a person looking to buy a used car, and as data scientists, we set out to build a prediction model to gauge the market before investing in a long-term asset such as a used car. We know we prefer purchasing a Toyota or a Ford, and we prefer low mileage and high mpg. To help structure our study we developed three hypotheses:

1. There is a difference between sale price for models within brands.
2. The engine size and year of a used car impacts the sale price of the car.
3. There is a difference in sale price between Ford and Toyota

DATA CLEANING

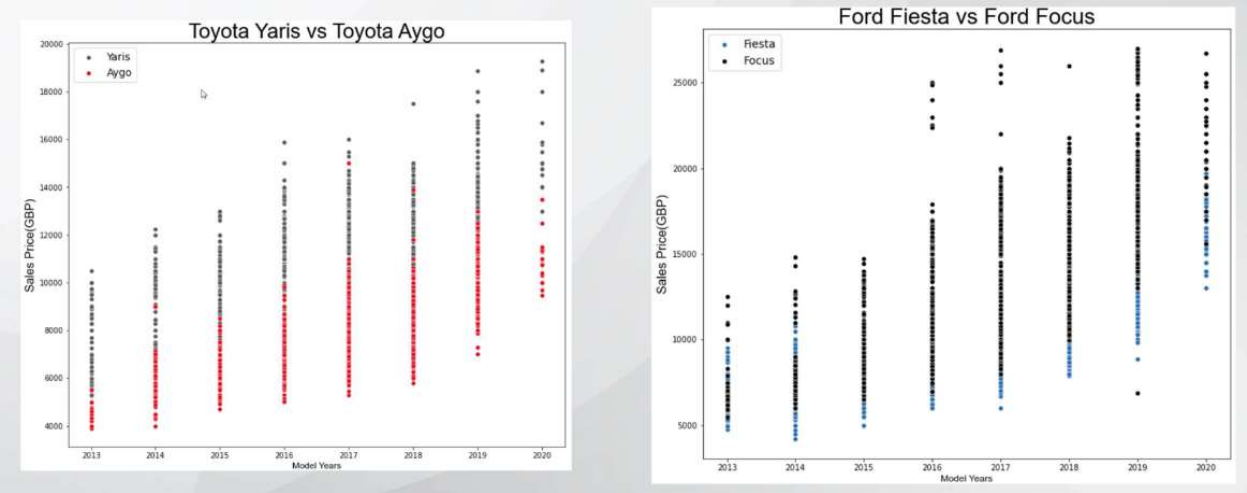
Our data cleaning step was the time-consuming, yet essential, part of our prediction model. To simplify our model, we focused our analysis on models Aygo and Yaris for the Toyota brand, and models Fiesta and Focus for the Ford brand. To remove outliers, we kept 95% of the data by removing the upper and lower bounds of price and mileage (mean plus or minus 2 standard deviations). Originally, the dataset had cars ranging from 1996 to 2060, but decided to keep only cars between 2013 and 2020, given it represents the bulk of our data. Further, we expected engine size of 0 to belong to electric or hybrid cars, however, most engine size zero belonged to petrol or diesel cars. We concluded that either the web scrapping method made an error or engine size was originally null but converted to 0 in the dataset and decided to remove all engine sizes of 0.

EXPLORING DATA

Models within brands

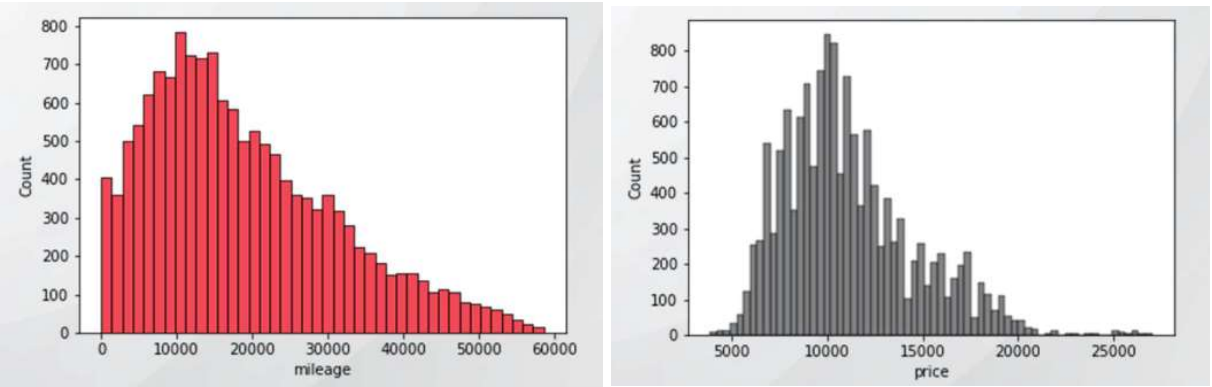
We explored the differences in price over the years for our chosen models in Toyota and Ford. Both brands had a positive trend showing and increase of sale price for newer cars. We

found particularly interesting Ford Focus model has a higher sale price for the years 2016 and 2017.



Price & Mileage

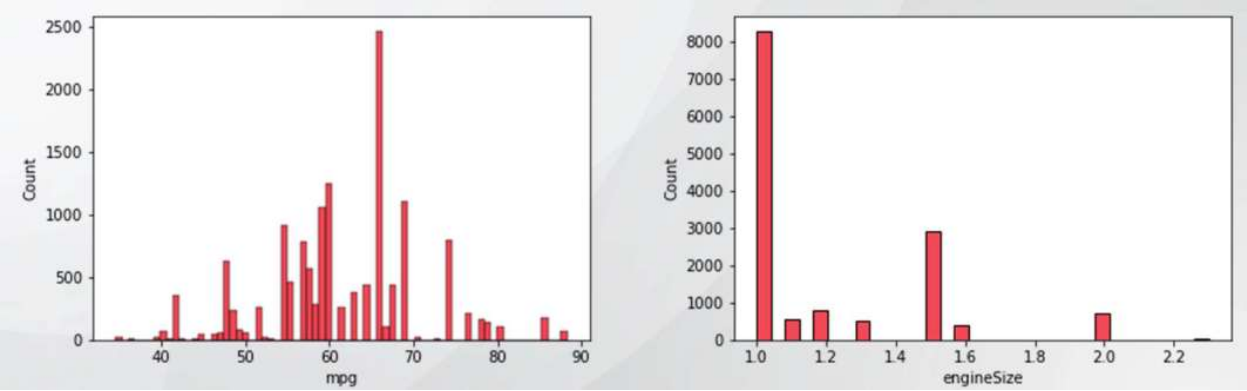
Price and Mileage are skewed to the right even after the data cleaning step. Trimming mileage and price to achieve a normal distribution created a different variable. We decided not to trim further because those upper bound data were integral to the dataset (i.e., higher mileage is typical of used cars). We found interesting that most cars in our dataset had a sale price around £10,000, or mileage of at least 10,000 miles.



MPG & Engine Size

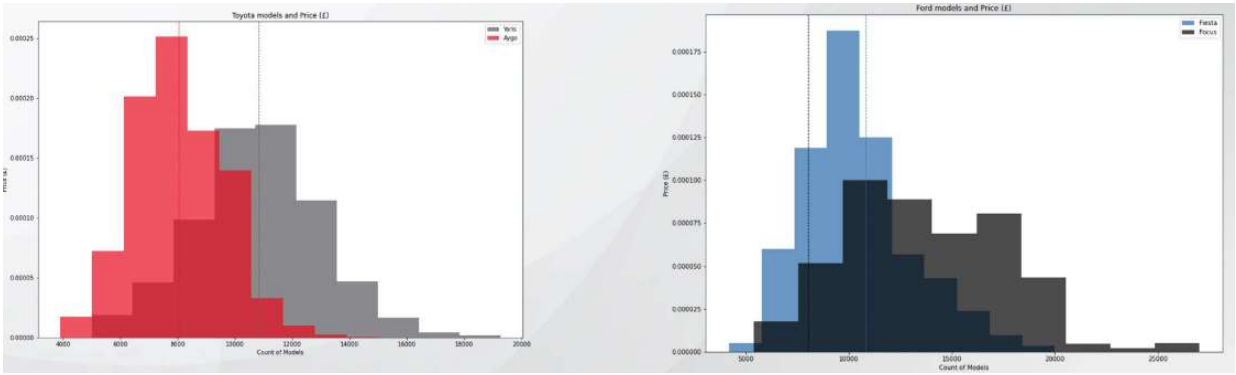
Mpg and engine size were not normally distributed. Mpg could have been binned as <50, 51-60, 61-70, and >70. Engine size could have been binned to size of 1 and size of >1 given

that most of the data could fall under those two bins. We did find interesting that most of the cars in our dataset had an engine size of either 1 or 1.5.



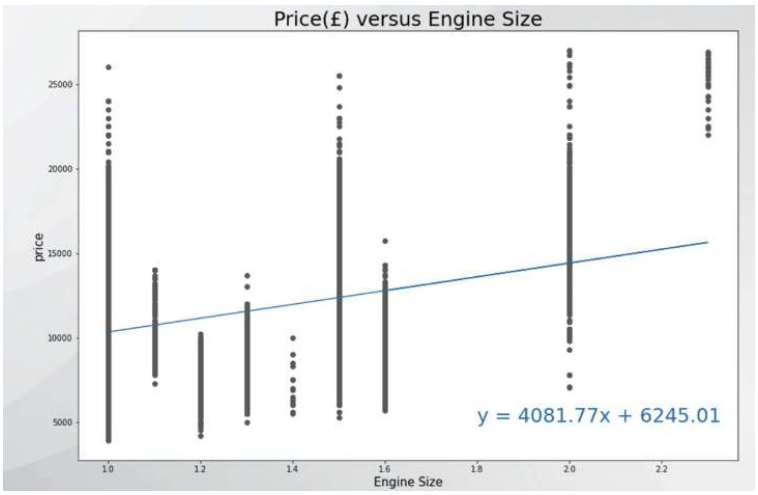
Difference in Sale price

After completing an independent t-test on both brands we found that there is no difference in price for models within brands (p-value < 0.0001). We found interesting that although Ford's Fiesta seemed to have more cars sold at a higher price range than Focus, the t-test proved that it is not much different from sale price of a Ford Focus.



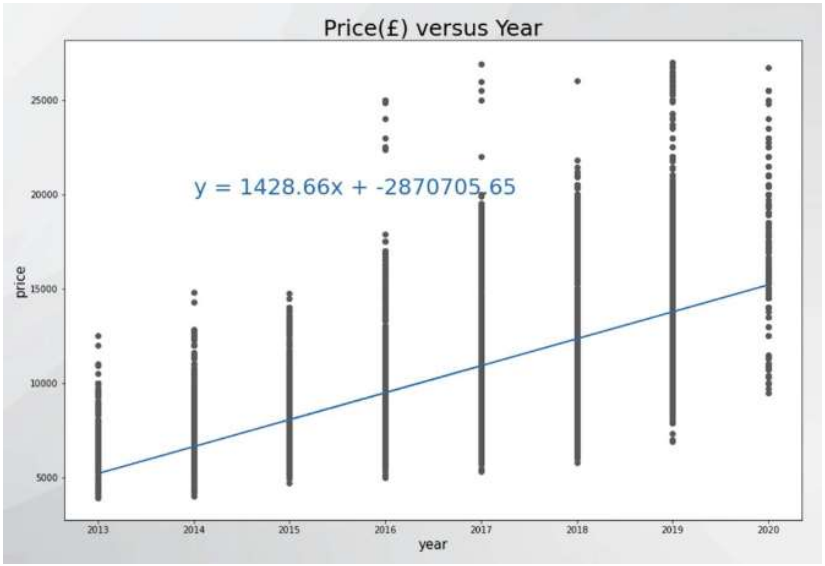
Relationship between Engine size and Price

Engine size positively predicted car price (r-squared = 0.1164). There was a weak but positive correlation between price and engine size. This leads us to believe that as engine size increase, we can expect the car to sell at a higher price.



Relationship between Year and Price

Year positively predicted car price (r-squared = 0.41). There was a strong positive correlation between year and price. As we expected, the newer the car the higher the sales price.



OLS Regression Results

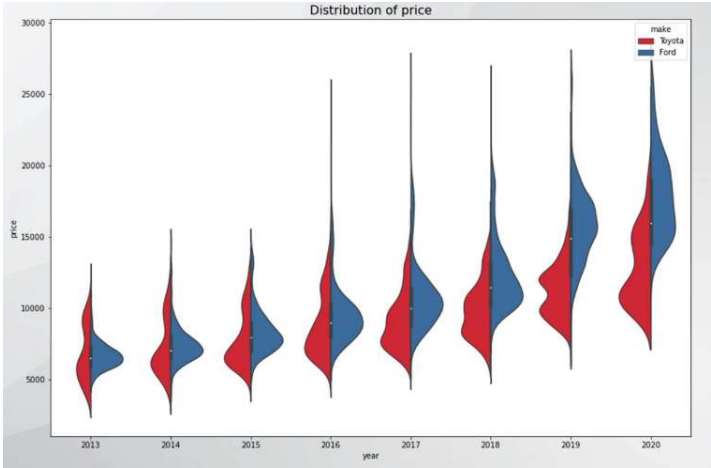
Dep. Variable:	price	**R-squared:**	0.406
Model:	OLS	**Adj. R-squared:**	0.406
Method:	Least Squares	**F-statistic:**	9725.
Date:	Thu, 17 Feb 2022	**Prob (F-statistic):**	0.00
Time:	08:08:04	**Log-Likelihood:**	-1.3236e+05
No. Observations:	14233	**AIC:**	2.647e+05
Df Residuals:	14231	**BIC:**	2.647e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.871e+06	2.92e+04	-98.236	0.000	-2.93e+06	-2.81e+06
year	1428.6633	14.487	98.617	0.000	1400.267	1457.060

Omnibus:	2245.635	**Durbin-Watson:**	1.028
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	4524.491
Skew:	0.963	**Prob(JB):**	0.00
Kurtosis:	4.980	**Cond. No.**	2.66e+06

Ford and Toyota

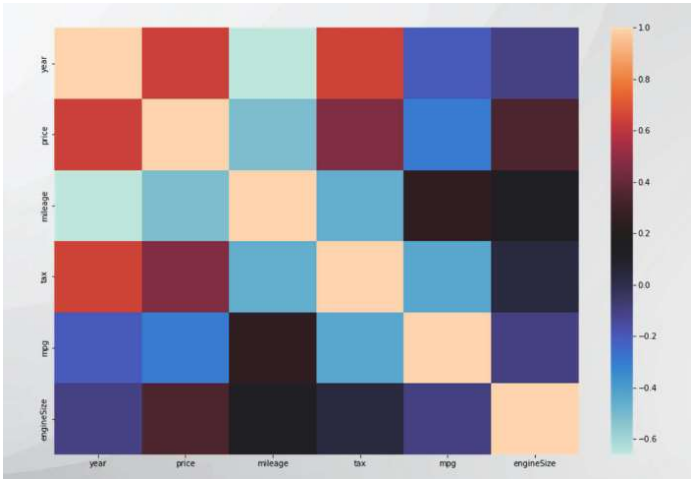
There was a significant difference between brands and car price (ANOVA p-value < 0.0001). A violin plot revealed that the Toyota data was bimodal with Yaris being more expensive than Aygo. The Ford dataset was unimodal, however, Focus tended to be more expensive than Fiesta.



MULTIPLE LINEAR REGRESSION

Correlation heatmap

Our correlation heatmap revealed that mpg was negatively correlated to price and engine size was positively correlated to price, contrary to our predictions



Analysis

Model, fuel type, and type of transmission were dummy-coded. Year, price, mileage, tax, mpg, and engine size were scaled. “Other” transmission type, year, “Other” fuel type, manual transmission type, Petrol fuel type, and semi-auto transmission type were dropped due to multicollinearity. Training and testing sets were divided into 70%/30%, respectively. The r-squared of the training set was 0.691 while the testing set had an r-squared of 0.697. The data was unscaled manually to facilitate the price prediction in units of GBP.

The screenshots below show the code used to perform a prediction on a car sale price based on unscaled values. In this example, we predicted the sale price for a Ford Fiesta with 10,000 miles, 58 mpg, engine size of 1.5 and tax of 125 to be £13,740.57.

```
y = intercept  + mile_coeff*10000 + tax_coeff*125 + mpg_coeff*58 + ES_coeff*1.5 + fiesta_coeff*1 + focus_coeff*0 + yaris_coeff*0 + hybrid_coeff*1
```

```
y = "{:,.2f}".format(y)
print(f"Your car is valued at £{y}")
```

```
Your car is valued at £13,740.57
```

CONCLUSION

Call to action

Based on the data, our car prediction fell within the £10,000 to £18,000 range. Thus, to retain value over time and to not overpay for the engine size, a 2017, £11,000, and 1.3 engine should be purchased by the respective consumer, based on the data.

Limitations & Future Work

Our model is not generalizable to other countries or regions. There are other variables for use car prices such as crude oil price, consumer confidence, and the economy. Further, our simple linear regression has relatively small r-squared values. Future work should focus on exploring events in 2020 such as Covid-19, crude oil price, used car supply shortage, labor shortage, and the general state of the economy. In addition, others should consider web-scrapping 2022 data and comparing them with 2020. APIs should be considered to build and merge datasets, including data from new cars. Our code is open-source and should be applied to U.S. used cars for comparison and to the other CSV files, with appropriate citations of our work.

REFERENCES

Background:

Engers, M., Hartmann, M. and Stern, S. (2009), Annual miles drive used car prices. J. Appl. Econ., 24: 1-33. https://doi.org/10.1002/jae.1034

Gegic, Enis, et al. “Car Price Prediction Using Machine Learning Techniques.” *TEM Journal*, UIKTEN - Association for Information Communication Technology Education and Science, Aug. 2019, https://www.ceeol.com/search/article-detail?id=746689.

Pal N., Arora P., Kohli P., Sundararaman D., Palakurthy S.S. (2019) How Much Is My Car Worth? A Methodology for Predicting Used Cars’ Prices Using Random Forest. In: Arai K., Kapoor S., Bhatia R. (eds) Advances in Information and Communication Networks. FICC 2018. Advances in Intelligent Systems and Computing, vol 886. Springer, Cham. https://doi.org/10.1007/978-3-030-03402-3_28

Saputra, Yuris Mulya, et al. “Energy Demand Prediction with Federated Learning for Electric Vehicle Networks.” *IEEE Xplore*, IEEE, 2019, https://ieeexplore.ieee.org/abstract/document/9013587/citations#citations.

Dataset:

https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes

Multiple liner regression:

https://towardsdatascience.com/multiple-linear-regression-model-using-python-machine-learning-d00c78f1172a