CCT COLLEGE

Strategic Thinking – Assessment Project

Dublin

2022

GROUP ID: DAB-FT-D921-004

Gabriela Fernandez Perez - 2021305
Mariana Yumi Yokoyama - 2021299
Nuno Alfredo Ribeiro Teixeira de Almeida - 2021310

PROJECT THEME

Engine Size as a price predictor in used cars in the UK.

Work developed in the analysis of a Dataset and research and study models to obtain results of the Strategic Thinking discipline of tHDip Data Analytics - FT - Sept 2021 at CCT course.

Lectures: James Garza

Dublin

2022

CCT College Dublin

Assessment Cover Page

Module Title:	Strategic Thinking
Assessment Title:	Final Presentation / Paper / Prototype.
Lecturer Name:	James Garza
Student Full Name:	Gabriela Fernandez Perez Mariana Yumi Yokoyama Nuno Alfredo Ribeiro Teixeira de Almeida
Student Number:	2021305 / 2021299/ 2021310
Assessment Due Date:	April 2021
Date of Submission:	Sunday, 22th May @ 23:59.
Links :	https://github.com/nunoaalmeida/StrategicThinking.git https://youtu.be/qMmurK34jdE

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

SUMMARY

1. INTRODUCTION

The Methodology used for this project is based on the Cross-Industry Standard Process for Data Mining (CRISP-DM) (Data Science Process Alliance, 2021). This is a model used since 1996 to standardize the data mining process between industries. It is divided into six phases which help to build a full project (Morrisey, 2021). Through this report, we will explain each of these phases and how we implemented them.

2. BUSINESS UNDERSTANDING

This is the first stage and it is very different from Data Understanding. In this section, we ask what is the requirement of the Company? Here we define the problem and the objectives.

In this project, we aim to know "***What is the impact of the engine size in the price of a used car in the UK?"***

This is a wide question that might be answered from different perspectives, so we had to define how we will structure the project to answer this question in the best way possible. This means that we need to narrow the spectrum of supposition to answer this, providing more questions that guide us into a narrower path.

1. What brand in specific are we taking into account?
2. What are the categories of these cars?
3. Are all these used cars in the same categories? E.g. Are all the cars Sedan or MPV?
4. Is there a difference in price between subcategories? E.g. Are Compact cars less expensive than mini-vans?
5. Does the Registration Year correlate with the price?

Based on those questions we defined that our study is based on **Ford used cars in the UK**.

However, it is important to emphasize the external factors that can influence the variation of used car prices. These factors can be colour, damage to the car, buyer psychology, among others. The lack of availability of this data is the main reason to not be taken into consideration for this project.

We take into consideration the model, year or tax that determine in a broader way the price of a car.

3. DATA UNDERSTANDING

This section is focused on the data set. There are four tasks to implement here. All of these steps are just to have a gross examination of the data we are dealing with.
First, we ***collect the data***, as our study is about used cars in the UK, we download the dataset from Kaggle, in **.csv format** into our Jupyter Notebook using pandas.
In the second place, we ***Describe the data*** just to have a general overview of it. In this step, we found that the data set we are working with contains **9 columns and 17,965 rows.** The types of data we have are **objects, floats and int.**
The Features we have are the following:

Figure 1 – Features by Data Set.

Variable	Description
Model	Model of the car
Year	Registration Year
Price	Price in Pounds
Transmission	Type of Gearbox
Mileage	Distance used
Fuel Type	Engine Fuel
Tax	Road tax
Mpg	Miles per gallon
EngineSize	Engine Size in litres

(Data visualization - Created by the Author, 2022)

However, we noticed that the data set is not divided into **categories**, this means we do not have any feature describing which class each car is, e.g. Cross-over / Subcompact.
We research the car models and their sizes to create these features in data preparation.
Subsequently, we ***Explore the data*** again, querying it and understanding the data types, how it is distributed, the values that it contains and the relation among the data.

Finally, in this stage, as we search for missing values or outliers we are ***Verifying the data quality.*** It is very important to document all the quality issues we might have with

the dataset. In our specific case, the added features are not well encoded and this might be an issue later in modelling.

List of results:

Figure 2 - Data Understanding Summary.

DATA UNDERSTANDING SUMMARY
Ford Data Set
17965 entries
9 columns
154 duplicate entries
0 missing values
4039 outliers
51 "engine size" values are zero value
43,22% of the cars are 1.0
86,37% of the cars are manual
Fiesta is the most frequent model
67,79% of the cars uses petrol
The most frequent price is 10.000 £
The most frequent registered year is 2017
Most of the Ford cars deliver a 65.7 mpg
The most frequent mileage is 10.

(Data visualization - Created by the Author, 2022)

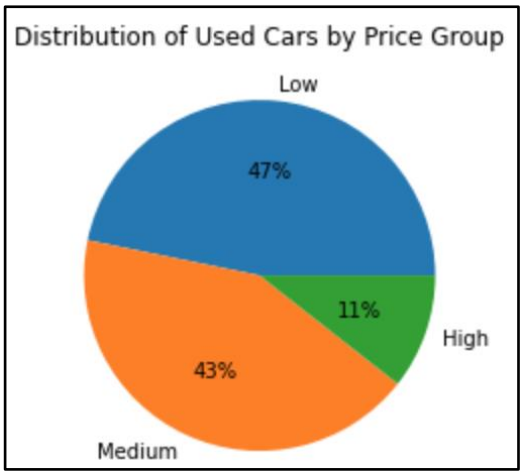
4. **DATA PREPARATION**

This is the most important and time-consuming phase of the project because here it is where we prepare the data set for the modelling. This phase is also called “Data Munging” (Data Science Process Alliance, 2021) In order to get a good data set that works accurately with the machine learning models we must follow these steps.

1. Select Data: We decided to keep all the features as they are important to compare in order to know if the engine size is the most important factor for the price of a used car. Although, if it is not the case, we would like to know then, which variables impact / affect this.

2. Cleaning Data: “GIGO” -“Garbage in, Garbage-out” which means that if the data is not coherent, the output would be incorrect (APA, 2020), in other words, your results just can be as good as your data.
    a. We dropped **154 duplicated rows** as they were not adding anything new to the dataset.
    b. Replace all the **zero values in Engine Size** with the mean of that feature (1.35) and in Tax (113). This was very important as we are focused on knowing if the price is a consequence of this size.
3. Feature Engineering or Construct Data:
    a. We researched to know what class each model belongs to and we created these features in the dataset. (*see the Appendix E and F for more information*). These classes might be needed for **Classification Models.**
    b. With the function **“Cut”** we created the feature **Price Group** dividing the cars into three groups:
        i. Low: 4970.3 until 11556.667
        ii. Medium: 11556.667 until 18123.333
        iii. High: 18123.333 until 24690

Figure 3 - Pie Chart of the Distribution of the Used Cars by Price Group.



(Data visualization - Created by the Author, 2022)

4. Integrate Data: In this project we did not merge other data set from any other car brand as we are just focusing on Ford cars.

5. Format data: This is how we deal with the values:

a. Split categorical and numerical data:

Figure 4. Categorical and Numerical Data.

Numerical Features	Categorical Features
Year	Model
Mileage	Class
Tax	Class Size
MPG	Transmission
Engine Size	Fuel type
Price	Price Group

(Data visualization - Created by the Author, 2022)

Figure 5. Categorical and Numerical by ".head" function

```
1 ford_num.head()
```

	year	mileage	tax	mpg	engineSize	price
0	2017	15944	150	57.7	1.0	12000
1	2018	9083	150	57.7	1.0	14000
2	2017	12456	150	57.7	1.0	13000
3	2019	10460	145	40.3	1.5	17500
4	2019	1482	145	48.7	1.0	16500

```
1 ford_cat.head()
```

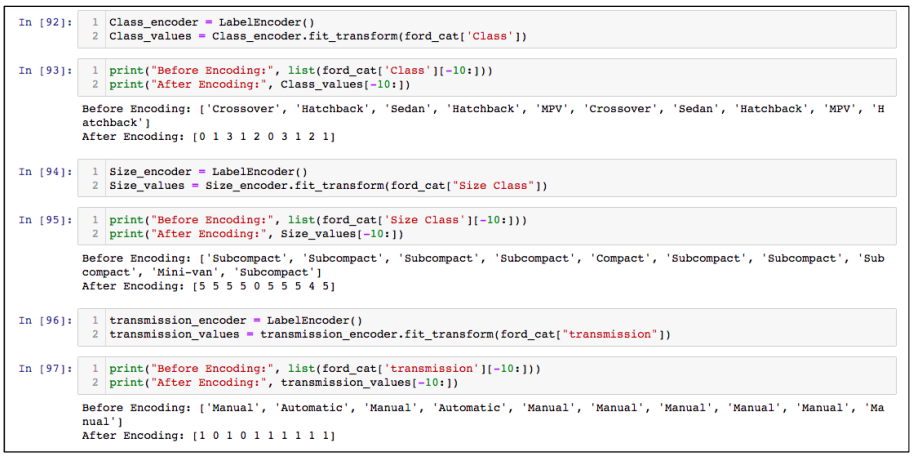
	model	Class	Size Class	transmission	fuelType	price_group
0	Fiesta	Hatchback	Subcompact	Automatic	Petrol	Medium
1	Focus	Hatchback	Compact	Manual	Petrol	Medium
2	Focus	Hatchback	Compact	Manual	Petrol	Medium
3	Fiesta	Hatchback	Subcompact	Manual	Petrol	Medium
4	Fiesta	Hatchback	Subcompact	Automatic	Petrol	Medium

(Jupyter Notebook - Created by the Author, 2022)

b. Processing Categorical Values:

i. Label Encoder: We transformed each categorical variable into numerical values.

Figure 6. Label Encoder function



```
In [92]:
Class_encoder = LabelEncoder()
Class_values = Class_encoder.fit_transform(ford_cat['Class'])

In [93]:
print("Before Encoding:", list(ford_cat['Class'][-10:]))
print("After Encoding:", Class_values[-10:])

Before Encoding: ['Crossover', 'Hatchback', 'Sedan', 'Hatchback', 'MPV', 'Crossover', 'Sedan', 'Hatchback', 'MPV', 'H
atchback']
After Encoding: [0 1 3 1 2 0 3 1 2 1]

In [94]:
Size_encoder = LabelEncoder()
Size_values = Size_encoder.fit_transform(ford_cat["Size Class"])

In [95]:
print("Before Encoding:", list(ford_cat['Size Class'][-10:]))
print("After Encoding:", Size_values[-10:])

Before Encoding: ['Subcompact', 'Subcompact', 'Subcompact', 'Subcompact', 'Compact', 'Subcompact', 'Subcompact', 'Sub
compact', 'Mini-van', 'Subcompact']
After Encoding: [5 5 5 5 0 5 5 5 4 5]

In [96]:
transmission_encoder = LabelEncoder()
transmission_values = transmission_encoder.fit_transform(ford_cat["transmission"])

In [97]:
print("Before Encoding:", list(ford_cat['transmission'][-10:]))
print("After Encoding:", transmission_values[-10:])

Before Encoding: ['Manual', 'Automatic', 'Manual', 'Automatic', 'Manual', 'Manual', 'Manual', 'Manual', 'Manual', 'Ma
nual']
After Encoding: [1 0 1 0 1 1 1 1 1 1]
```

(Jupyter Notebook - Created by the Author, 2022)

When we concat all the variables, we got several null values. This error might be because when we replace the object values by int from zero to *n* the variables with zero were not recognized as values. It is worth mentioning that this encode might affect results as it ranks the values based on the alphabet and this is why we chose another method.

- One Hot Encoder: This encoder worked perfectly, as the data is not ordinal and we do not have a large number of categories we have no issue with the dummy variables created. (Sethi, 2020). By this nature we increased our features from 12 to 38.
- Normalize Data: As we concat the variables, it is notable that the numerical columns have a larger input and this can affect the model because it prioritizes these higher values. To avoid this, we applied the MinMaxScaler.

This preserves the shape of the original distribution and scales all the data into a range [0,1] (Gogia, 2019) We did not choose StandarScaler as our data is not normal distributed and as we removed the outliers we did not apply RobustScaler.

- Train and Test Split: We split the label **Price** (y)**,** from our independent variables (X). As we had a leakage (explained below) we removed it from X as well.

Data Preparation Summary:

Data Cleaning:

- Deleting duplicate rows

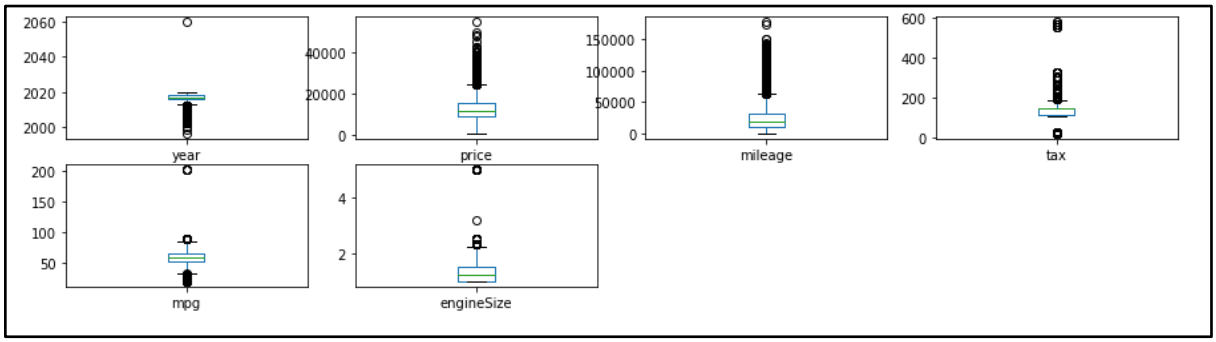
- Look for missing values
- Replace zero values with Mean values of each feature (Tax and Engine Size)

Feature Engineering:

- Class
- Size Class
- Price Groups

5. DATA VISUALIZATION

Figure 7. Outliers of the variable price for Ford used cars.



(Jupyter Notebook - Created by the Author, 2021)

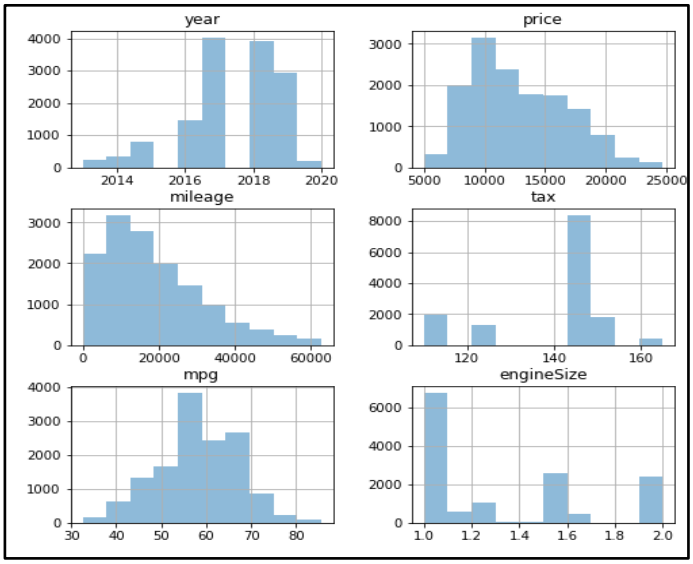
With the function above we can take a look at all numerical variables and outliers. Before removing duplicate rows, we had a total of 17965 rows, now removing outliers it has a total of 13926 rows.

With the histogram below we can confirm that most of the used cars´ **prices** are closer to the value of £10,000. The variable **'Mileage'** follows a right-skewed distribution with most of the values clustered from 0 to 25,000.

Moreover, we are able to verify in the figure that in the variable **'year'** most of the data are clustered between 2016 and 2020 and the graph presents a left tail.

In engine size, 1.0 is the most frequent followed by 1.5 and 2.0. And finally, most of the tax values are clustered between the values 140 and 160.

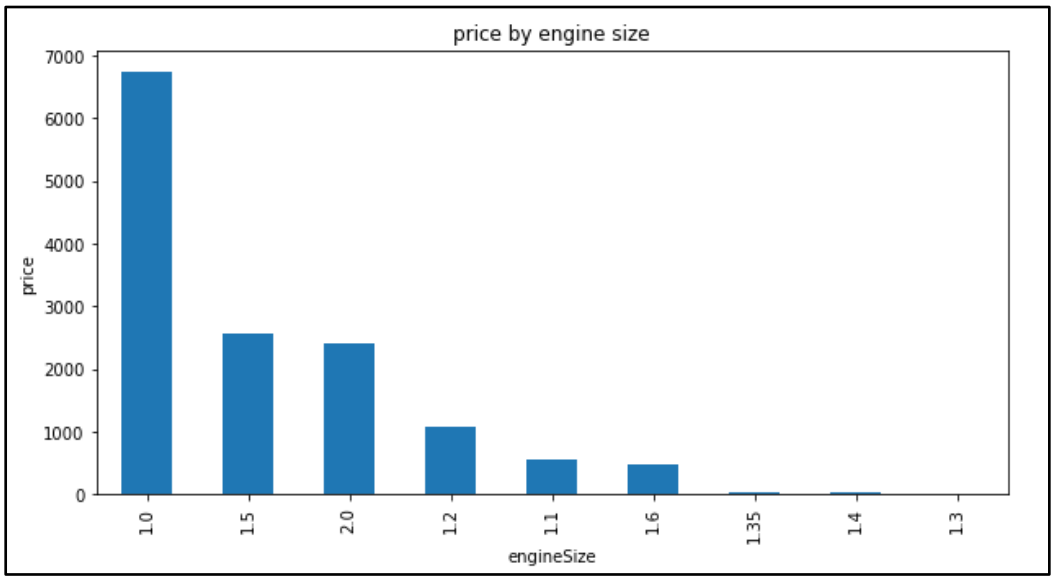
Figure 8. Histogram of the variable price for Ford used cars.



(Data Visualization - Created by the Author, 2021)

As we are studying the correlation between the engine size and the price, we can look for a relationship. With the graph below we can see that E.S. 1 is the one that contains the higher value prices.

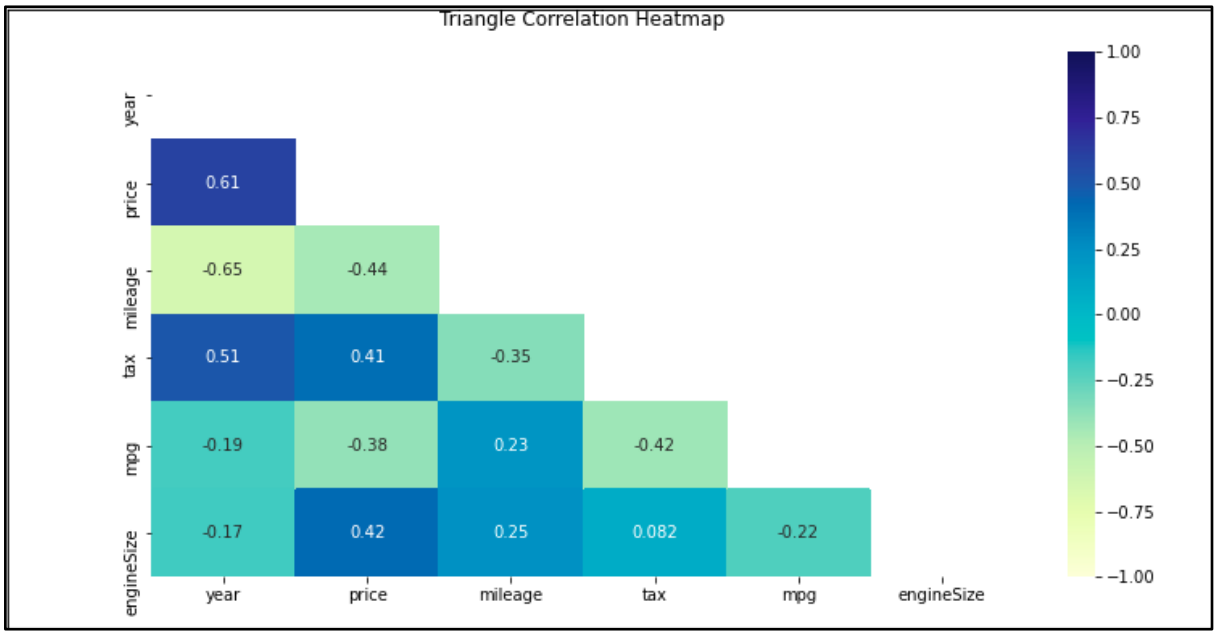
Figure 9. Histogram of 'price by engine size' for Ford used cars



(Data Visualization - Created by the Author, 2021)

In this heatmap, we can see that the correlation between 'price' and 'year' are the ones with the greater relationship and that 'engine size' and 'tax' are the ones with the lowest relationship. It also suggests that 'price' and 'engine size' have a moderate correlation between them as they present a value of 0.42, according to the values and the strength of the correlation Louzan (2021).

Figure 10. Heatmap of the variables for Ford used cars



(Data Visualization - Created by the Author, 2021)

6. **MODELLING**

In this section we explain the chosen models and why.

We split the dataset into the default splitting settings of 75% training and 25% for testing for the Random Forest Feature Selection, Linear Regression and Supported Vector Regressor.

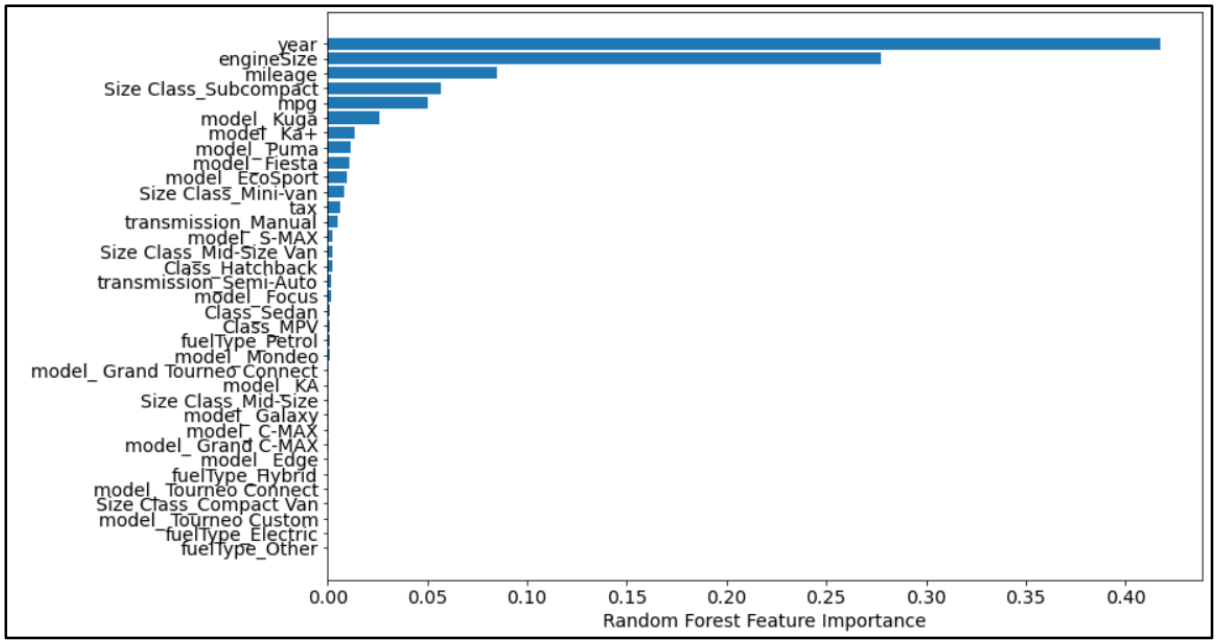
6.1 RANDOM FOREST FEATURE SELECTION

We chose this supervised model because each tree calculates the importance of a feature according to its ability to increase the pureness of the leaves. (Malato, 2021). This technique gives a score to input characteristics depending on how valuable the feature is to predict the target variable, this may increase the efficiency and efficacy of

a predictive model on the problem. (Brownlee, 2020). We expect to have the most important predictive features here and determine if Engine Size is one of them.

The graph below demonstrates that Engine Size is not the top predictor of price, but it is in the top three predictors. The best variable to predict price is Year, followed by Mileage.

Figure 11. Random Forest Feature Importance.



(Data Visualization - Created by the Author, 2021)

Once the most important features are defined, we proceed to apply the models.

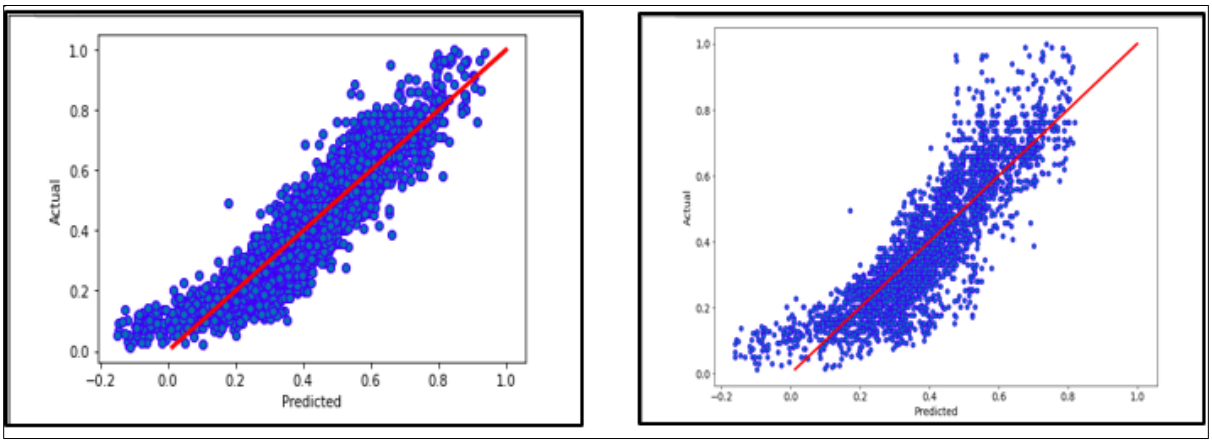
6.2 LINEAR REGRESSION

The predictions are made based on the value of another variable and the model draws a line that minimizes the squared distance of each point from the line of best fit. (Chen, 2021). Here the number of features will be analysed and determine the projection of the prediction, with the best line. As Müller and Guido (2017) mention, if we have only one feature our result will be a line, when analysing two features we will obtain a plane and when higher dimensions are implemented our result will be a hyperplane.

This model was tested with all the features and with the ones selected by RF. The difference between both is minimal, as we can see in the graphs below. However, the

first model could have been overfitted and the selected features give us a model more accurate.

Figure 12. Linear Regression All Features and Feature Selection



(Data Visualization - Created by the Author, 2021)

6.3. SUPPORT VECTOR REGRESSOR (SVR)

Based on the similarity this model has with Linear Regression, and according to Sethi (2020), this model provide a better prediction fit since it allows to select those values that fall within a decision boundary and will be characterised by a lower error rate.

7. EVALUATION

Here is a comparison of the performance of the models with their MSE and Cross Validation.

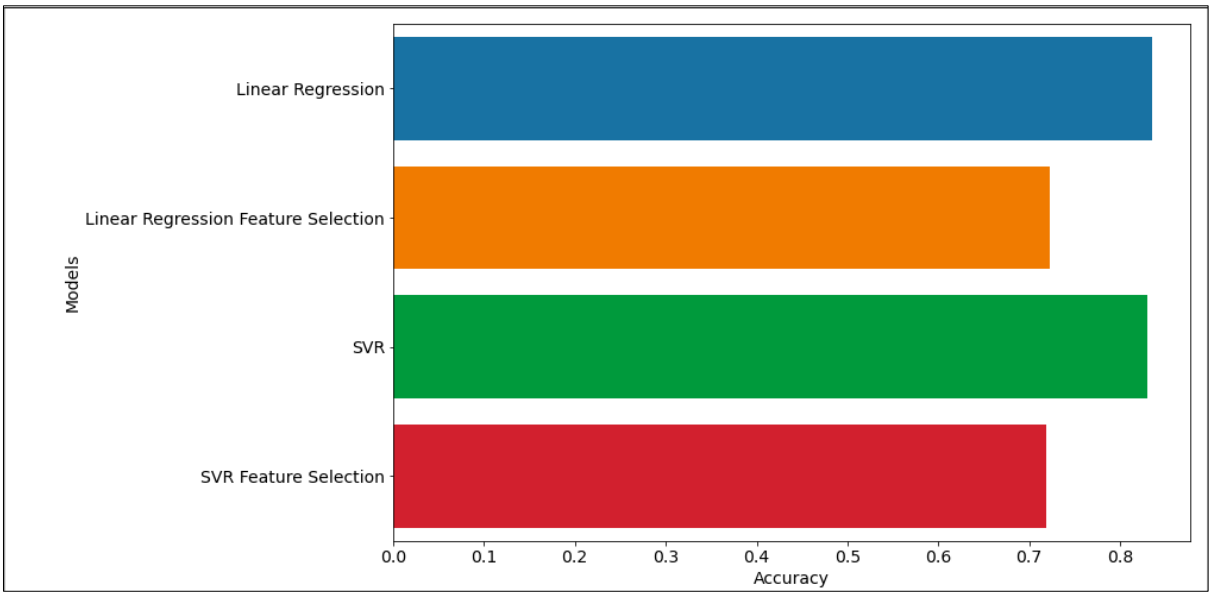
Figure 13. Table of Results

	Models	Accuracy	MSE	Mean Cross Validation
0	Linear Regression	0.841042	0.006209	0.840803
1	Linear Regression Feature Selection	0.740715	0.010129	0.730673
2	SVR	0.827285	0.006747	0.829016
3	SVR Feature Selection	0.737389	0.010259	0.727844

(Jupyter Notebook - Created by the Author, 2021)

The table above demonstrates how the accuracy of the models drop when the Feature Selection was implemented. Despite the fact the models with the feature selection have a lower accuracy, we believe they are better as they are less likely to be overfitted.

Figure 13. Models Evaluation



(Data Visualization - Created by the Author, 2021)

As we apply regression models, we used the Cross Validation Technique to evaluate the performance of the model. We used a K-Fold of 10 (which means the function mix the data selected for training and testing 10 times and gives a result of each iteration). In the table we chose the mean of these iterations and the results are quite similar.

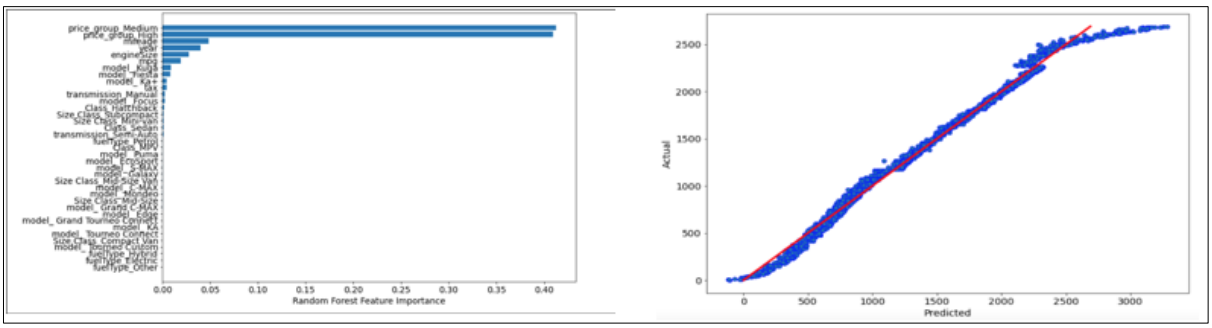
We also apply the Mean Squared Error to measure if the predictions match the observed data (Zach, 2020) The closer this number is to zero the better. As we can see in the table above, the MSE for the models without the FS are smaller, however, the result for those with the FS are still good results.

7.1 DISCUSSION

It is worth mentioning that all the models above were also implemented with the variables price group. This approach is incorrect because we notice a ***leakage*** in the data.

Leakages occurs when some information or data that has “extra” information and is used to train the data that you are trying to predict (Singh, 2022). In other words, the model is learning the same pattern and understanding that this extra information (variable) is the one that matches with the assumed result. In our case, as we tried to predict price, the variables that contain information about this were highly correlated and bias the model.

Figure 14. Feature Selection Leakage



(Data Visualization - Created by the Author, 2021)

As expected, the Random Forest chose Medium and High Price as the most important variables to predict Price. As well as the Linear Regression seems to have an excellent fit, which help to visualize something is not correct.

8. DEPLOYMENT

It is important to remember we worked without outliers and a future step could be to create models that include them and compare results.
Regarding the models, Decision Tree can also be implemented and the parameters of SVR can be changed to compare the results, as the C parameter contribute into the margin of the overall error.
Furthermore, classification models can be a good approach as we created the classes and price groups variables.

9. CONCLUSIONS

We tried to predict if Engine Size is a predictor of price in Used Cars in the UK. Based on the models, we determined that it is not the most important predictor, as it is Year of the car. But we found that Engine Size is one of the three most important predictors.

It is increasingly evident that the entire process is important, decision making and testing, according to this emphasis, as the conclusion of the results, we can leave explained individually, an option of all participants for the success of this. With great emphasis on group work carried out throughout the semester. Following below the opinion of members, their perceptions and lessons learned.

Gabriela: In this project, we learnt CRISP-DM and its application in projects. In Particular, I learnt how to organize the projects and how to divide it into sections. About the code, I understand now that I do not need to explain it, I need to explain the result I get with it, because there is the answer and the analysis giving an easy and general explanation of the data. Documentation of all the approaches and discoveries are essential to not confuse the process and to get a deeper understanding of the data for a better analysis and modelling. Finally, I am still learning how to work as a team. This means that sometimes I have to understand that others have the correct answer even though it can be a last minute call.

Mariana: In that, I understood with CRISP-DM that we must pay more attention and focus on documentation, on understanding what we want to analyse and respond with our project. And also that at various moments we as a group need to be flexible to restart, rearrange and restructure our project. We also had to know and study better statistical techniques and machine learning so that we could improve our analyses and to choose wisely which models would be the most suitable for what we want to deliver.

Nuno: Given the knowledge acquired from the other materials and used in this whole project, the most important were Data Understanding and Machine Learning in their initial preparation stage for a future prediction. CRISP-DM has given us a better view of structuring reports and reports to develop them for readers. Each step required a lot of research and team understanding, which would not have been possible without it. A clear understanding of Machine Learning models and explanation of projects as examples were essential. So we hope to deliver the results and explanations clearly.

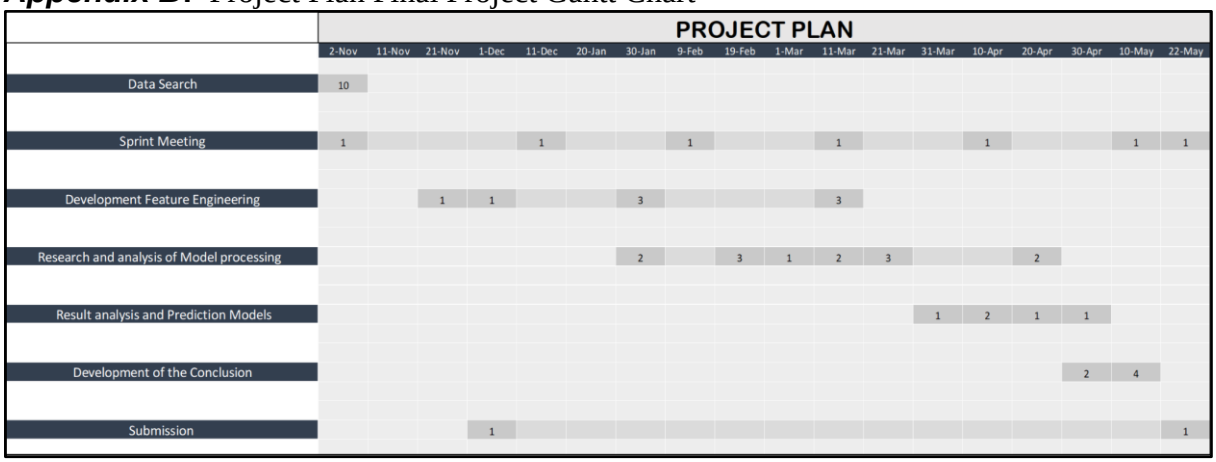
APPENDIX

***Appendix A*:** Team Contribuition Final Project

Team Contribuition	
Member	Participation
Gabriela Fernandez Perez - 2021305	Research and storytelling of the analysis (report), as well as providing guidance to the rational sequence to apply the results and elaborate the presentation. Development and analysis of results to be discussed in order to present them to the team.
Mariana Yumi Yokoyama - 2021299	Codification and first inputs for the python code, as well as experimentation about the clustering algorithms. In Machine Learning tested models, and alternative strategies for improved data selection.
Nuno Alfredo Ribeiro Teixeira de Almeida - 2021310	Research and correction of codes, as well as summarising the key outcomes in Data Preparation in order to achieve the main goal. Validation of models to make the results and outcomes with more accurate analysis.

Source: Elaborated by Author.

***Appendix B*:** Project Plan Final Project Gantt Chart



Source: own elaboration based on Subjectmoney, 2013.

***Appendix C*:** Ford Fiesta UK market prices, 2016 by region

Region	Average Mileage (K)	Average Price
Northern Ireland	22	£7,547
Scotland	20	£7,782
North West	21	£7,949
North East	23	£8,006
Wales	24	£8,106
Yorkshire	21	£8,180
East Midlands	22	£8,192
West Midlands	21	£8,197
South West	23	£8,265
Greater London	21	£8,354
South East	21	£8,425
East of England	21	£8,663

Source: HPI, 2016

***Appendix D*:** Ford Focus UK market prices, 2016 by region

Region	Average Mileage (K)	Average Price
North East	72	£9,013
Wales	55	£10,295
Yorkshire	45	£10,699
South West	49	£10,876
North West	36	£11,006
Scotland	33	£11,136
Greater London	30	£11,449
East Midlands	58	£11,460
East of England	34	£11,716
West Midlands	39	£11,848
South East	35	£12,523
Northern Ireland	51	£12,984

Source: HPI, 2016.

Appendix E: Create new columns Class and Size Class

```
ford.insert(1,"Class", ford["model"])
```

```
ford["Class"] = ford["Class"].replace([" Fiesta"," Focus"," Kuga"," EcoSport"," C-MAX"," Ka+"," Mondeo"," B-MAX",
                                       " S-MAX"," Grand C-MAX"," Galaxy"," Edge"," KA"," Puma"," Tourneo Custom",
                                       " Grand Tourneo Connect"," Mustang"," Tourneo Connect"," Fusion"," Streetka",
                                       " Escort"," Ranger"," Transit Tourneo"],
                                      ["Hatchback","Hatchback","Crossover","Crossover","Sedan","Sedan","Sedan","MPV",
                                       "MPV","MPV","MPV","Crossover","Hatchback","Crossover","MPV","MPV","Sports Car",
                                       "MPV","MPV","Hatchback","Sedan","Pickup truck","MPV"])
```

```
ford.insert(2,"Size Class", ford["model"])
```

```
ford["Size Class"] = ford["Size Class"].replace([" Fiesta"," Focus"," Kuga"," EcoSport"," C-MAX"," Ka+"," Mondeo",
                                                 " B-MAX"," S-MAX"," Grand C-MAX"," Galaxy"," Edge"," KA"," Puma",
                                                 " Tourneo Custom"," Grand Tourneo Connect"," Mustang"," Tourneo Connect",
                                                 " Fusion"," Streetka"," Escort"," Ranger"," Transit Tourneo"],
                                                ["Subcompact","Compact","Compact","Subcompact","Subcompact","Subcompact",
                                                 "Mid-Size","Mini-van","Mini-van","Compact","Mini-van","Mid-Size Van",
                                                 "Subcompact","Subcompact","Mid-Size Van","Compact","Muscle Car",
                                                 "Compact Van","Mid-Size Van","Subcompact","Compact","Mid-Size Pickup",
                                                 "Compact"])
```

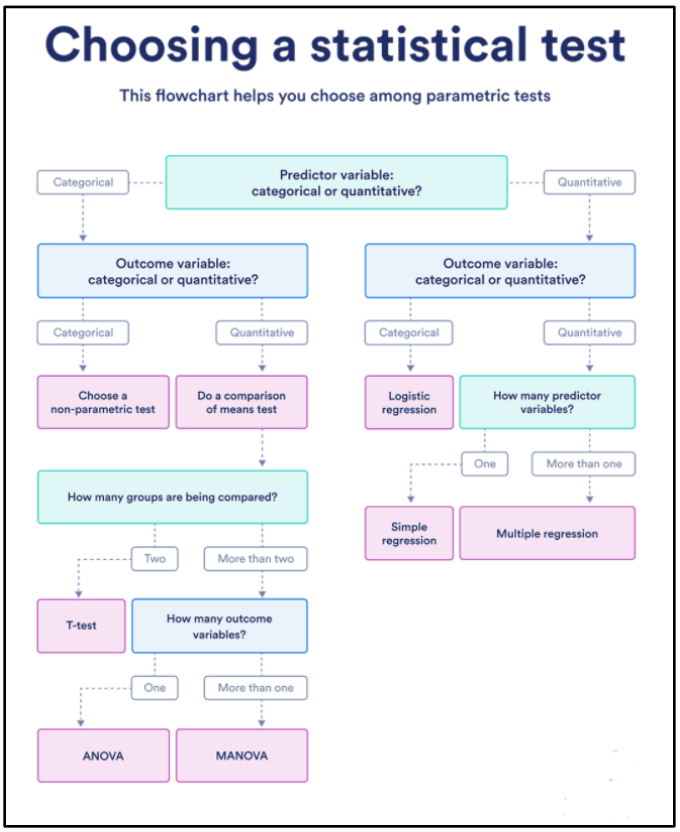
Source: Jupyter Notebook

***Appendix F*:** List of Ford Vehicles according to the class and Size Class

A	B	C
Model	**Class**	**Size Class**
Fiesta	Hatchback	Subcompact
Focus	Hatchback	Compact
Kuga	Crossover	Compact
EcoSport	Crossover	Subcompact
C-MAX	Sedan	Subcompact
Ka+	Sedan	Subcompact
Mondeo	Sedan	Mid-Size
B-MAX	MPV	Mini-van
S-MAX	MPV	Mini-van
Grand C-MAX	MPV	Compact
Galaxy	MPV	Mini-van
Edge	Crossover	Mid-Size Van
KA	Hatchback	Subcompact
Puma	Crossover	Subcompact
Tourneo Custom	MPV	Mid-Size Van
Grand Tourneo Connect	MPV	Compact
Mustang	Sports Car	Muscle Car
Tourneo Connect	MPV	Compact Van
Fusion	MPV	Mid-Size Van
Streetka	Hatchback	Subcompact
Escort	Sedan	Compact
Ranger	Pickup truck	Mid-Size Pickup
Transit Tourneo	MPV	Compact

Source: Wikipedia, 2021

Appendix G: Flowchart: choosing a statistical test Source:



Source: Bevans, 2020

References:

Aditya (2020) *100,000 UK Used Car Data set*. Available at: https://kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes (Accessed: 19 November 2021).

Afonja, T. (2018) *Kernel Functions*, *Medium*. Available at: https://towardsdatascience.com/kernel-function-6f1d2be6091 (Accessed: 19 November 2021).

Amat, R. (2017) *Máquinas de Vector Soporte (Support Vector Machines, SVMs)*. Available at: https://www.cienciadedatos.net/documentos/34_maquinas_de_vector_soporte_support_vector_machines (Accessed: 19 November 2021).

APA, D. of P. (2020) *garbage in, garbage out – APA Dictionary of Psychology*. Available at: https://dictionary.apa.org/garbage-in-garbage-out (Accessed: 6 December 2021).

APD, R. (2019) '¿Cuáles son los tipos de algoritmos del machine learning?', *APD España*, 4 April. Available at: https://www.apd.es/algoritmos-del-machine-learning/ (Accessed: 18 November 2021).

Bevans, R. (2020) *Statistical tests: which one should you use?*, *Scribbr*. Available at: https://www.scribbr.com/statistics/statistical-tests/ (Accessed: 6 December 2021).

Brownlee, J. (2017) 'Difference Between Classification and Regression in Machine Learning', *Machine Learning Mastery*, 10 December. Available at: https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/ (Accessed: 19 November 2021).

Brownlee, J. (2020) 'How to Calculate Feature Importance With Python', *Machine Learning Mastery*, 29 March. Available at: https://machinelearningmastery.com/calculate-feature-importance-with-python/ (Accessed: 22 May 2022).

Chen, J. (2021) *Line Of Best Fit*, *Investopedia*. Available at: https://www.investopedia.com/terms/l/line-of-best-fit.asp (Accessed: 22 May 2022).

Data Science Process Alliance (2021) 'CRISP-DM', *Data Science Process Alliance*. Available at: https://www.datascience-pm.com/crisp-dm-2/ (Accessed: 6 December 2021).

Geek for Geeks (2020) 'Normalization vs Standardization', *GeeksforGeeks*, 8 June. Available at: https://www.geeksforgeeks.org/normalization-vs-standardization/ (Accessed: 18 May 2022).

Gogia, N. (2019) *Why Scaling is Important in Machine Learning? | by Nishesh Gogia | Analytics Vidhya | Medium*. Available at: https://medium.com/analytics-vidhya/why-scaling-is-important-in-machine-learning-aee5781d161a (Accessed: 22 May 2022).

Gyoza (2918) *python - How to change the space between histograms in pandas*, *Stack Overflow*. Available at: https://stackoverflow.com/questions/52359595/how-to-change-the-space-between-histograms-in-pandas (Accessed: 25 November 2021).

HongGit (2020) *Using Data Contracts - WCF*. Available at: https://docs.microsoft.com/en-us/dotnet/framework/wcf/feature-details/using-data-contracts (Accessed: 30 November 2021).

Iqbal, M. (2019) 'Gaussian Naïve Bayes and Support Vector Machine.'

Jiang, L. (2019) *What are kernels in machine learning and SVM and why do we need them?*, *Quora*. Available at: https://www.quora.com/What-are-kernels-in-machine-learning-and-SVM-and-why-do-we-need-them (Accessed: 19 November 2021).

Jones, T. (2018) 'Supervised learning models', *IBM Developer*, 26 February. Available at: https://developer.ibm.com/articles/cc-supervised-learning-models/ (Accessed: 18 November 2021).

Lopez, F. (2018) *Teorema de Bayes - Definición, qué es y concepto*, *Economipedia*. Available at: https://economipedia.com/definiciones/teorema-de-bayes.html (Accessed: 18 November 2021).

Lyashenko, V. (2020) *Cross-Validation in Machine Learning: How to Do It Right*, *neptune.ai*. Available at: https://neptune.ai/blog/cross-validation-in-machine-learning-how-to-do-it-right (Accessed: 18 May 2022).

Majumder, P. (2020) *Gaussian Naive Bayes*, *OpenGenus IQ: Computing Expertise & Legacy*. Available at: https://iq.opengenus.org/gaussian-naive-bayes/ (Accessed: 19 November 2021).

Malato, G. (2021) 'Feature selection with Random Forest', *Your Data Teacher*, 11 October. Available at: https://www.yourdatateacher.com/2021/10/11/feature-selection-with-random-forest/ (Accessed: 22 May 2022).

Morrisey, M. (2021) 'CRISP-DM stage four - modelling'.

Oxford (2021) *garbage in garbage out*, *Oxford Reference*. doi:10.1093/oi/authority.20110803095842747.

Rombauts, W. (2021) 'List of Ford vehicles', *Wikipedia*. Available at: https://en.wikipedia.org/w/index.php?title=List_of_Ford_vehicles&oldid=1057606756 (Accessed: 30 November 2021).

Sethi, A. (2020a) ‘Categorical Encoding | One Hot Encoding vs Label Encoding’, *Analytics Vidhya*, 5 March. Available at: https://www.analyticsvidhya.com/blog/2020/03/one-hot-encoding-vs-label-encoding-using-scikit-learn/ (Accessed: 22 May 2022).

Sethi, A. (2020b) ‘Support Vector Regression In Machine Learning’, *Analytics Vidhya*, 27 March. Available at: https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/ (Accessed: 6 December 2021).

Singh, P. (2022) *Data Leakage in Machine Learning: How it can be detected and minimize the risk*, *Medium*. Available at: https://towardsdatascience.com/data-leakage-in-machine-learning-how-it-can-be-detected-and-minimize-the-risk-8ef4e3a97562 (Accessed: 22 May 2022).

Stack, O. (2021) *python - Add column to dataframe with constant value*, *Stack Overflow*. Available at: https://stackoverflow.com/questions/29517072/add-column-to-dataframe-with-constant-value (Accessed: 23 November 2021).

VanderPlas, J. (2021) *4. Visualization with Matplotlib - Python Data Science Handbook [Book]*. Available at: https://www.oreilly.com/library/view/python-data-science/9781491912126/ch04.html (Accessed: 25 November 2021).

ZACH (2020) ‘An Easy Guide to K-Fold Cross-Validation’, *Statology*, 4 November. Available at: https://www.statology.org/k-fold-cross-validation/ (Accessed: 22 May 2022).