Task - 01

- **QUES 1.] Which Python libraries did you find most useful in loading and exploring the dataset?**
  ANS 1.] The most used python libraries in task 1 for me were, pandas and numpy for the purpose of loading and exploring dataset.

- **What preprocessing steps did you find necessary to apply to the dataset?**
  ANS 2.] The preprocessing steps included checking the null values and displaying the data types and then describing the dataset's mathematical summary for getting the insights from the data.

- **What metrics were used to evaluate the Classification problem and why?**
  ANS 3.] Confusion matrix and classification_report were used to evaluate the classification problem as they best explains the accuracy, precision, recall and f-1 score for the analysis and purpose.

- **How did you detect overfitting/underfitting in the model and what strategies did you use to mitigate it?**
  ANS 4.] K-fold and cross value scores were used to detect the difference between the training and testing dataset results as the machine learned on the training dataset and while predicting the test value or cases the results varied which might cause the faulty result.

- **If you had a choice to apply hyperparameter tuning to the models? Explain why you will apply?**
  ANS 4.] Hyperparameter Tuning is applied to the parameters in the parenthesis defined in the code block, it is applied so as to increase the performance or the accuracy of the model.

- **Describe the Hyperparameter tuning?**
  ANS 5.] Hyperparameter is the tuning or adjusting of the value of different parameters defined inside the functions used by various

python libraries to implement the tasks, so as to enhance and increase the model performance.

- **What hyperparameters will you use for the models? explain separately for every model?**
  ANS 6.]
  For DecisionTreeClassifier, the class_weight is set to be balanced.
  For KNeighborsClassifier, the number of neighbours is set to be 3.
  For RandomForestClassifier, the estimators are used to be a total of 10.

Task - 02

•What challenges did you encounter while preparing the dataset at preprocessing and EDA analysis?
ANS 1.] The dataset (i.e the Bitcoin transfer data) was unpredictable and understanding the meaning and relation between the various column entries was difficult to comprehend. As the unix and the volume of USD and BTC was unrelated it took some time to configure the real relation and to figure out the dependent and the independent variable.

•Describe the difference observed while modelling with linear and Random forest regressor.
ANS 2.] The linear regressor gives the direct outcome and is based on only finding the linear relation between the dependent and the independent variable whereas the random forest regressor generates the result by using the technique of ensemble modelling that gives the best or the most efficient model based on various different models.

•How did you evaluate the performance of the linear and random forest regressor?
ANS 3.] I evaluated the performance of linear and random forest regressor with the help of various evaluation metrics namely, MAE (Mean absolute error), MSE (Mean squared error) and R square metrics.

•How well did your model perform on the testing set compared to the training set?
ANS 5.]
The mean squared error of the model comes out to be : 1.0617761939775524e-21
The mean squared error of the model comes out to be : 3.25849074569432e-11

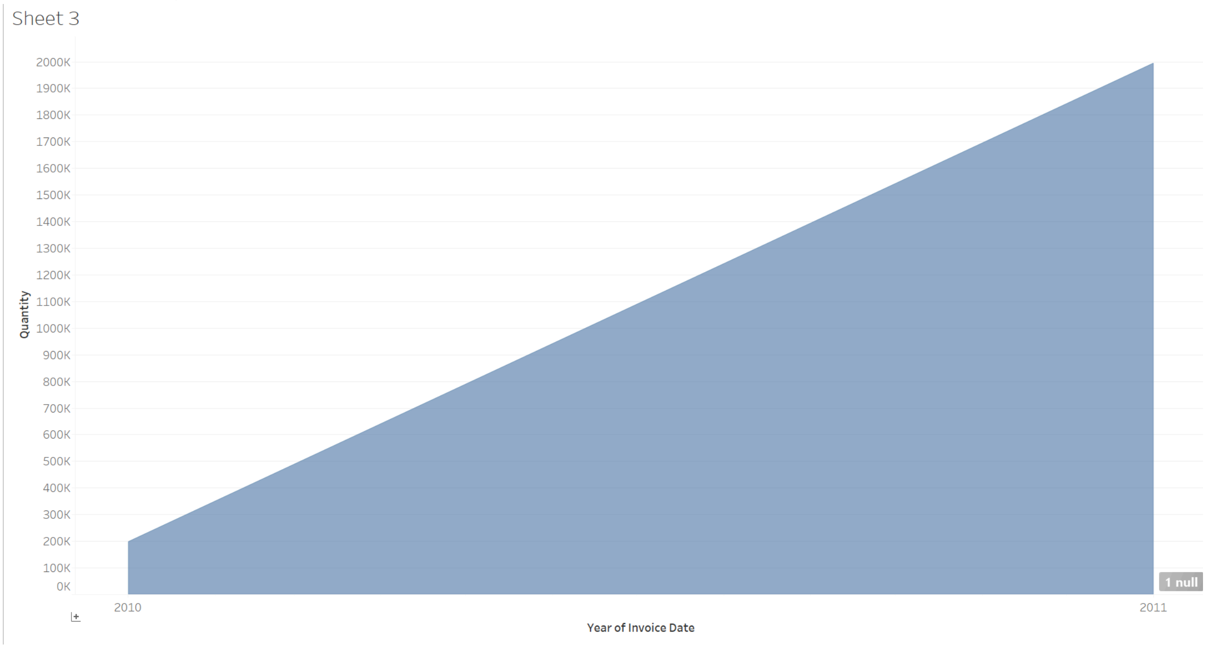
•Have you applied tuning and how did you identify the best tuned hyper-parameters?
ANS 6.]The hyper-parameters were analysed on the basis of R-square score As it tells the goodness of fit, here the best model is LinearRegression model.

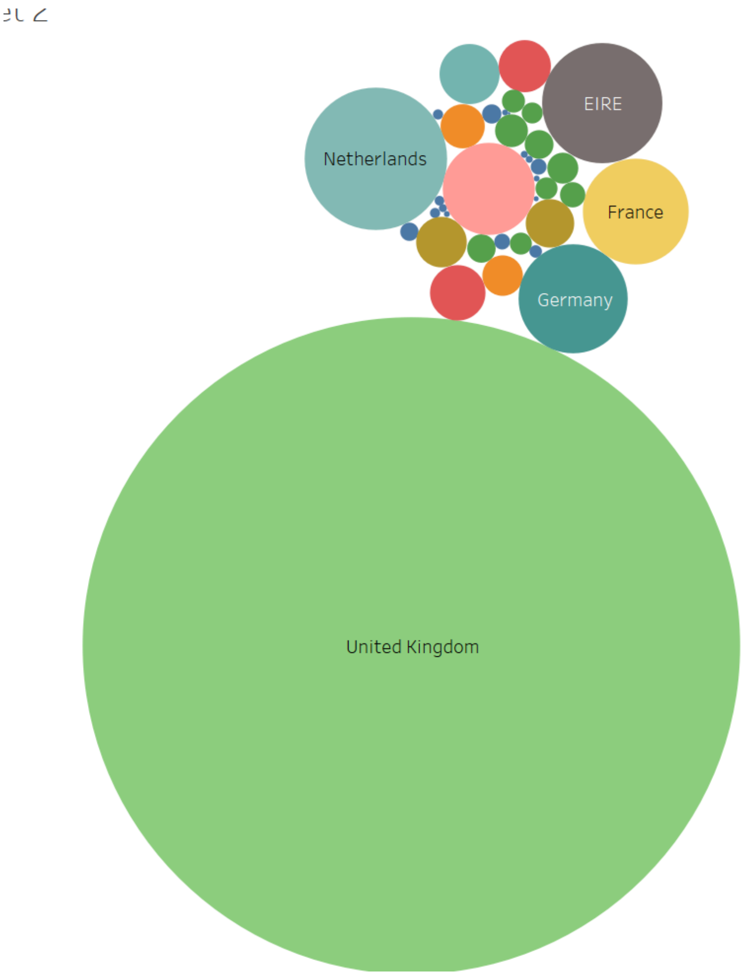
Major Project

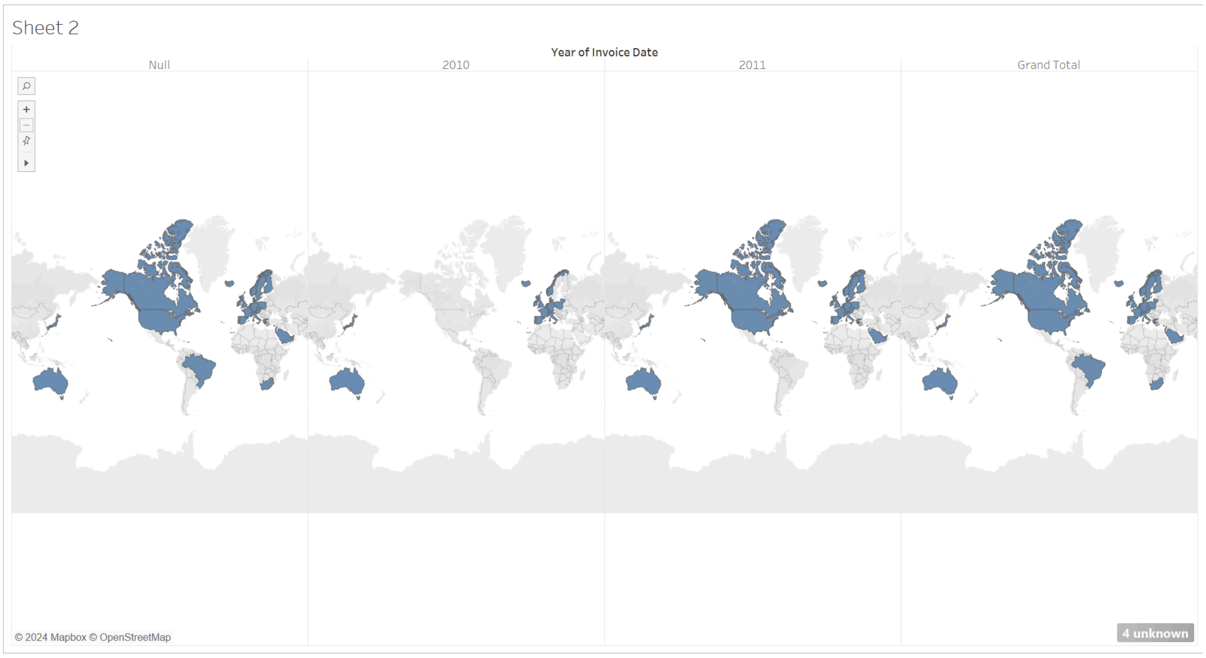
- **What is the business model of the e-commerce platform you're working with?**
  ANS 1.] The dataset is a transnational one, capturing every transaction made from December 1, 2010, through December 9, 2011, by a UK-based non-store online retail company.

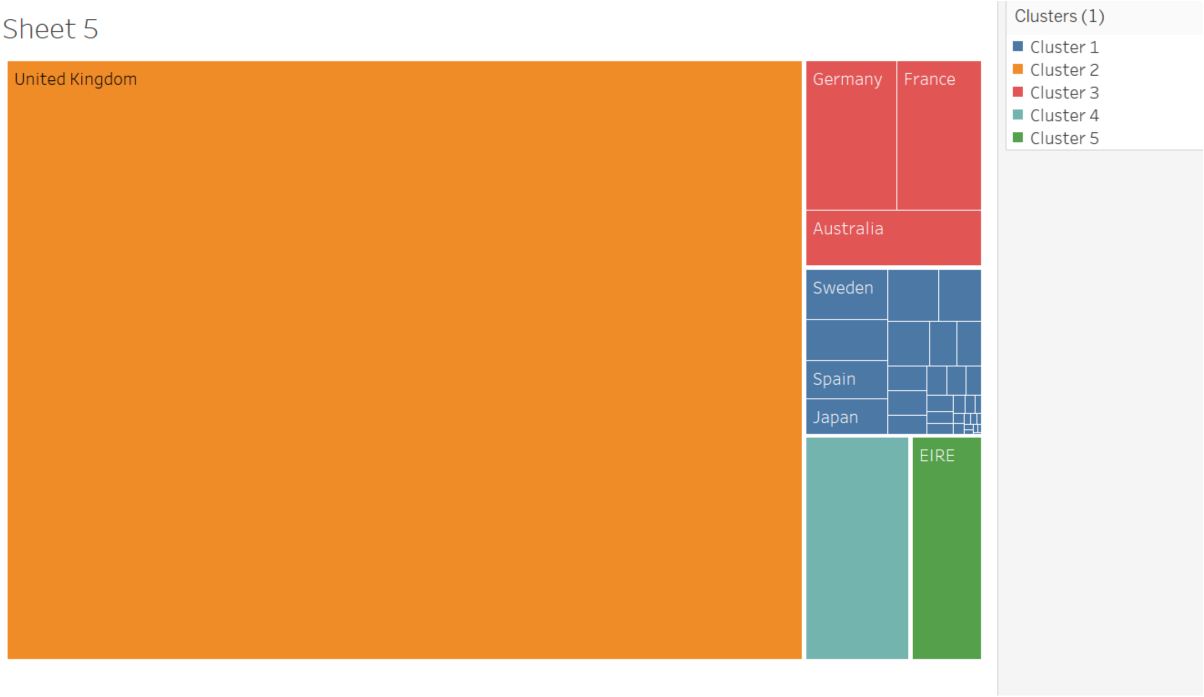
- **What kind of data preprocessing and cleaning was required for the Online Retail II Dataset?**
  ANS 2.]The dataset was checked for the null values and they were removed, later on the data was checked and all the categorical data types were converted to the integer type. Then a new column was created based on the feature selection.

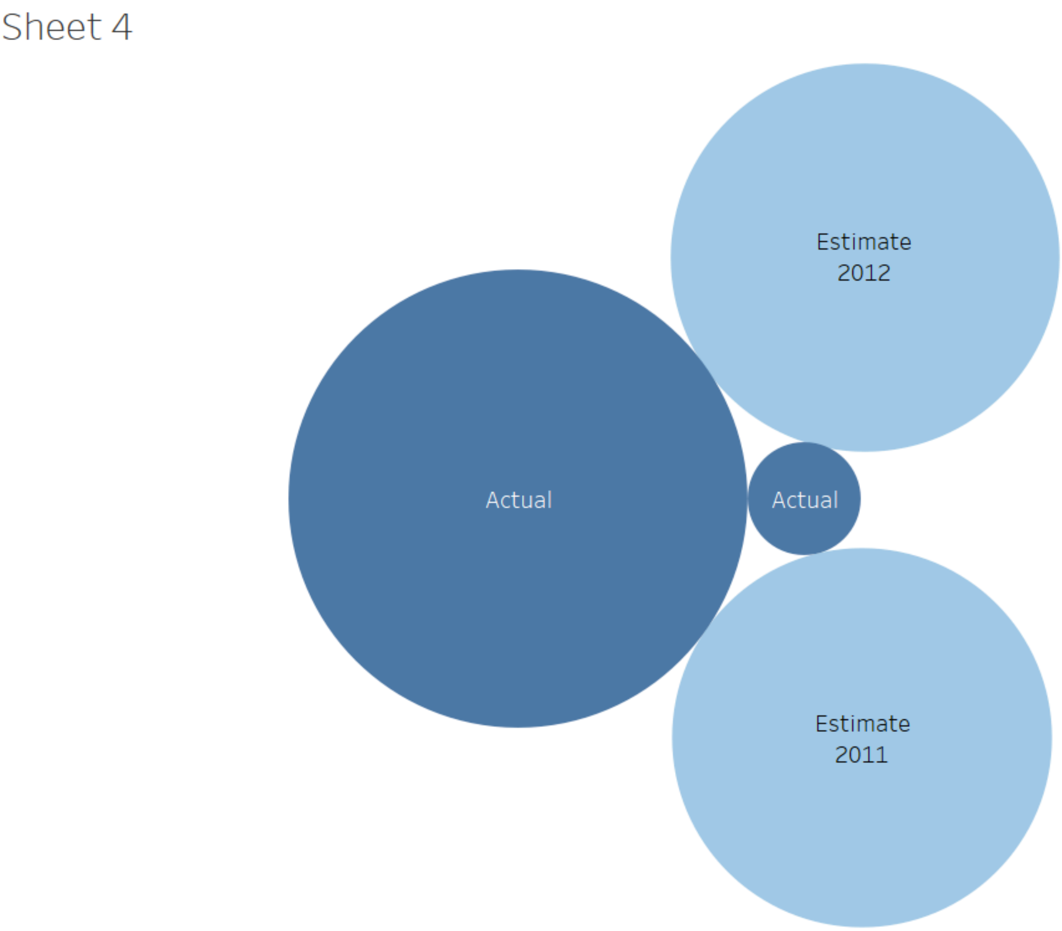
- **How did you visualise and interpret the data distributions and relationships using Power BI/Tableau?**
  ANS 3.] I used Tableau personally for analysing and finding the relationship between the datasets.

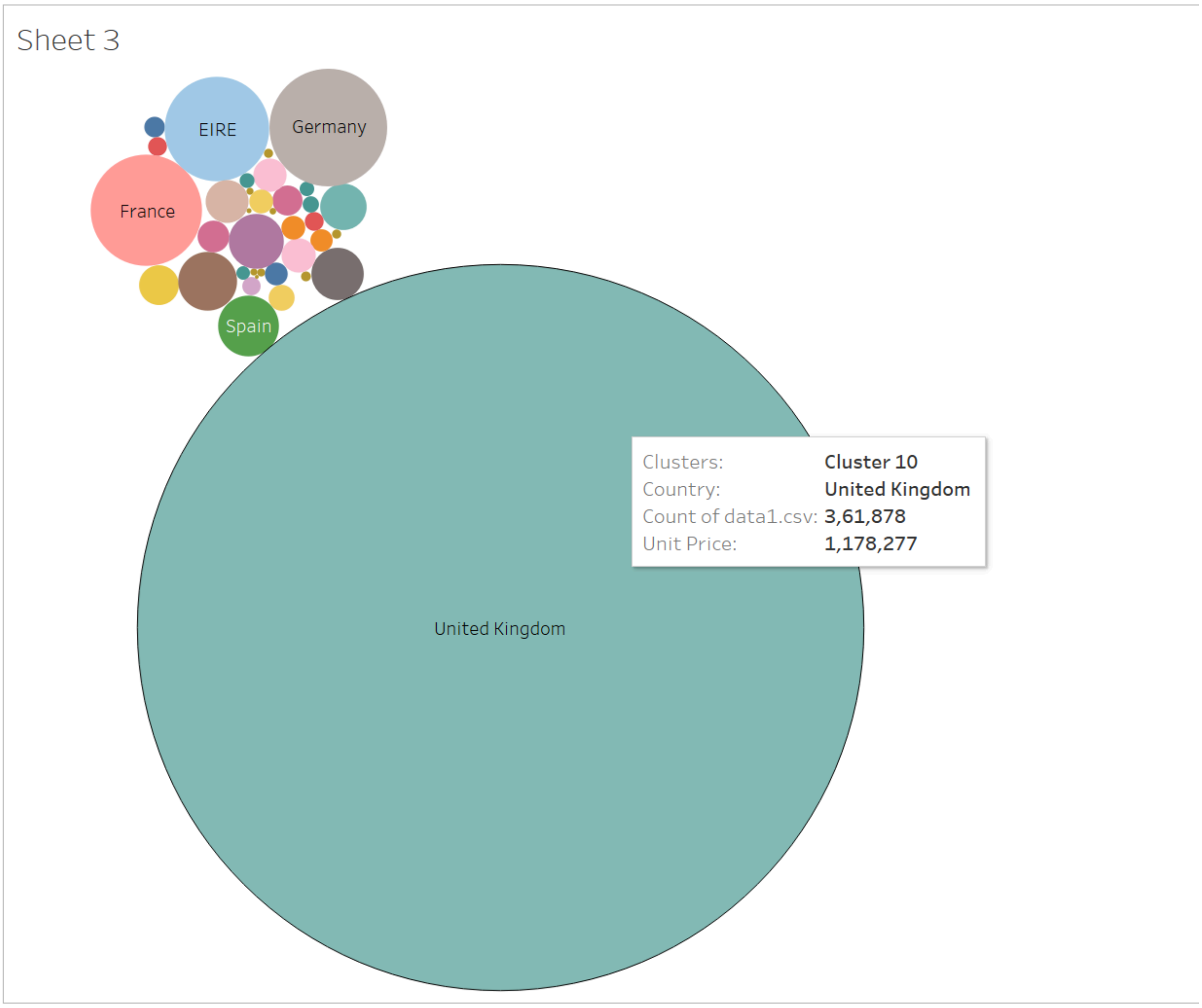



et 2
EIRE
Netherlands
France
Germany
United Kingdom


Sheet 2
Year of Invoice Date
Null
2010
2011
Grand Total
© 2024 Mapbox © OpenStreetMap
4 unknown


Sheet 5
United Kingdom
Germany
France
Australia
Sweden
Spain
Japan
EIRE
Clusters (1)
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5


Sheet 4
Estimate
2012
Actual
Actual
Estimate
2011



Clusters: Cluster 10
Country: United Kingdom
Count of data1.csv: 3,61,878
Unit Price: 1,178,277

- **What new features did you engineer from the existing dataset and why?**
  ANS 4.] I introduced a new column named total_cost.

- **Which regression models did you test for predicting the annual spending of a customer?**
  ANS 5.] I used LinearRegression, DecisionTreeRegressor, and RandomForestRegressor for this purpose.

- **What metrics did you use to evaluate the performance of the predictive models?**
  ANS 6.] The mean absolute error, mean squared error, root mean squared error and r square metrics were used to evaluate all the predictive models.

- **How did you apply clustering techniques for customer segmentation? What were the results?**
  ANS 7.]Clustering technique was applied by using the kmeans clustering method.

- **How would you interpret the results obtained from the model in a business context?**
  ANS 8.]The results show that the maximum of the order were placed in the UK, Germany, France and Spain. And the no. of orders increases annually linearly.

- **How can the insights derived from this project be beneficial for the e-commerce platform's business strategy?**
  ANS 9.]We can derive insights from the data by using various visualisation techniques such as seaborn, matplotlib, Tableau, and Power BI etc.

- **What did you learn about the data science project lifecycle throughout this project?**
  ANS 10.] I learned about the analysis and importance of various techniques used throughout the lifecycle of the project including the data preprocessing, feature selection, predictive modelling, and classification tasks and later on performing of visualisation and analysis.