|  |  |
| --- | --- |
|  | E-commerce machine learning |
|  | Machine Learning |
|  | Name: Clare Kuan Sze Sing  Student Number: 220456793 |

TABLE OF CONTENTS

[**1.** **Introduction** 2](#_Toc194516031)

[**2.** **Dataset #1** 2](#_Toc194516032)

[2.1 Feature selection and modifications 2](#_Toc194516033)

[2.2 Exploratory Data Analysis (EDA) 3](#_Toc194516034)

[2.2.1 Numeric Features 3](#_Toc194516035)

[2.2.2 Categorical features 4](#_Toc194516036)

[**3. Unsupervised learning** 4](#_Toc194516037)

[3.1 Substantive issue 4](#_Toc194516038)

[3.2 Research questions 4](#_Toc194516039)

[3.3 Methodology 4](#_Toc194516040)

[3.4 Final modifications and analysis 4](#_Toc194516041)

[3.4.1 Principal Component Analysis 5](#_Toc194516042)

[3.4.2 K-means clustering 5](#_Toc194516043)

[3.4.3 Hierarchical clustering 6](#_Toc194516044)

[**4** **Classification** 6](#_Toc194516045)

[4.1 Substantive issue 6](#_Toc194516046)

[4.2 Research Questions 6](#_Toc194516047)

[4.3 Methodology 6](#_Toc194516048)

[4.4 Final modifications and analysis 6](#_Toc194516049)

[4.4.1 K nearest Neighbour (KNN) 7](#_Toc194516050)

[4.4.2 Logistic Regression 7](#_Toc194516051)

[**5. Dataset #2** 8](#_Toc194516052)

[5.1 Feature selection and modifications 8](#_Toc194516053)

[5.2 Exploratory Data Analysis (EDA) 8](#_Toc194516054)

[5.2.1 Categorical Features 8](#_Toc194516055)

[5.2.2 Numeric features 9](#_Toc194516056)

[**6. Regression** 9](#_Toc194516057)

[6.1 Substantive issue 9](#_Toc194516058)

[6.2 Research Questions 9](#_Toc194516059)

[6.3 Methodology 9](#_Toc194516060)

[6.5 Final modifications and analysis 10](#_Toc194516061)

[6.5.1 Linear Regression 10](#_Toc194516062)

[6.5.2 Regularisation models (Lasso, Ridge and Elastic net) 10](#_Toc194516063)

[6.5.3 Random Forest 11](#_Toc194516064)

[**APPENDIX** 12](#_Toc194516065)

### **Introduction**

Machine learning is an artificial intelligence tool that utilises datasets to train predictive models for various analyses and prediction. Our study focuses on e-commerce, an industry that has a potential to make up 33% of US retail revenue and add almost $1 trillion in this sector by 2027 [[1]](#footnote-1). With major corporations like JPMorgan and Walmart being more open to integrate generative AI to improve efficiency[[2]](#footnote-2), the interaction of AI and e-commerce is becoming more pronounced. Given this trend especially with the continuous advancements of AI tools, this study leverages machine learning algorithms to analyse 2 datasets to identify meaningful patterns:

dfonline\_sales (derived from online\_sales\_dataset.csv) which contains the global online transactions and purchasing behaviour

dfuk (derived from amz\_uk\_processed\_data.csv) which focuses on product specifically from the Amazon UK market due to its AI-driven infrastructure and vast dataset. Further explanations in 6.1.

### **Dataset #1**

### Feature selection and modifications

Before visualising the dataset and their relationships, modifications to the dataset are necessary.



Figure 1: dfonline\_sales

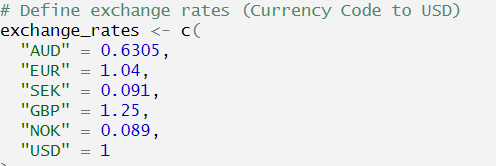
To scale down the dataset for both modification and the machine learning algorithms, we first remove any irrelevant variable(s) such as: InvoiceNo, StockCode, CustomerID, Description, InvoiceDate, ShipmentProvider and WarehouseLocation

Leaving us with 10 variables for further analysis. Next, we remove any NaN values that either resulted from column removal or were already present in the original dataset. In addition, we filtered out negative Quantity values for accuracy. This newly cleaned dataset is renamed to dfonline\_sales\_cleaned as a progress tracker.

Next, to standardise currency, UnitPrice, Discount and ShippingCost will be will be converted to USD. Firstly, we identity the unique countries present in the Country variable. 

Figure 2: The different countries present that has to be converted to USD

Afterwards, we proceed to map them out to their currency code and create a new column: CountryCode. Finally, we match the exchange rate of these currencies to USD and apply them to all the price related columns.



Lastly, to further condense the dataset, we remove the non-converted price related columns and CountryCode, leaving us with the final variables that will be used for Exploratory Data Analysis (EDA).

Figure 3: Currency conversion rates at the time of data processing

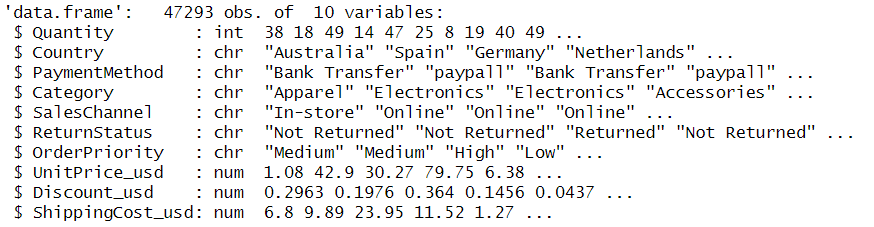


Figure 4: dfonline\_sales\_cleaned

### Exploratory Data Analysis (EDA)

### Numeric Features

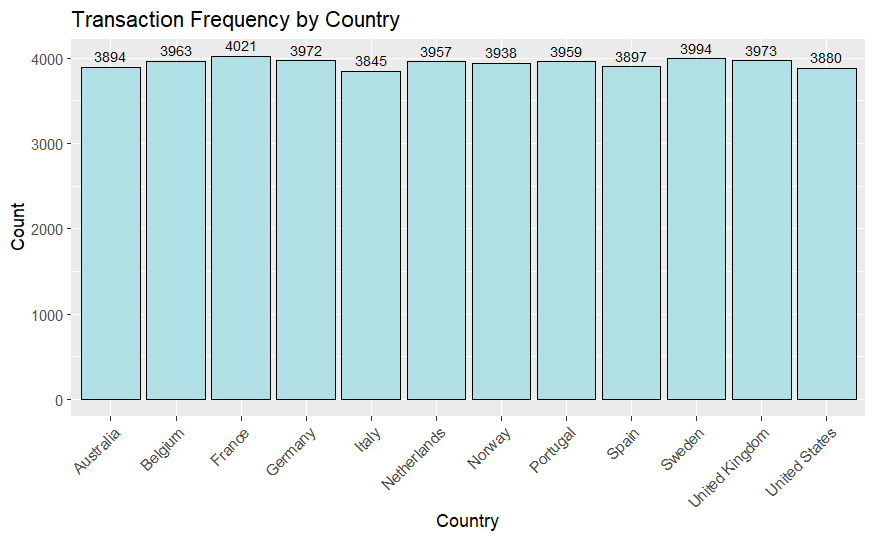
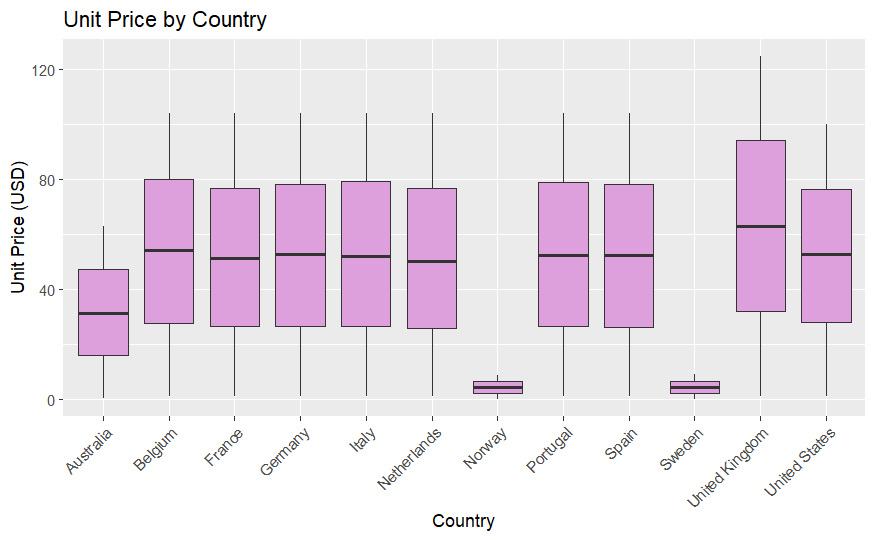
 

Figure 5.0: bar graph for Country

Figure 5.1: Boxplot of UnitPrice\_usd by Country

Figure 5.0 presents a bar graph of Countrywhich is well distributed, representing a well-balanced representation of consumers worldwide.

Figure 5.1 presents a boxplot of UnitPrice\_usdagainstCountry that highlights 2 extreme outliers (Norway and Sweden) that represent transactions with exceedingly low unit prices compared to other countries.

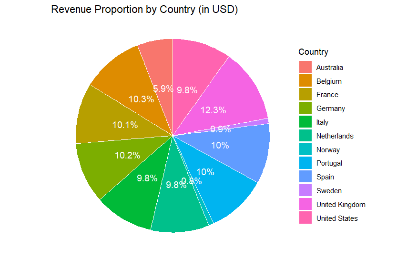


Figure 5.2: Pie chart of total revenue per unique countries in Country

Figure 5.2 presents a pie chart of sales revenue for each country. The calculations are:

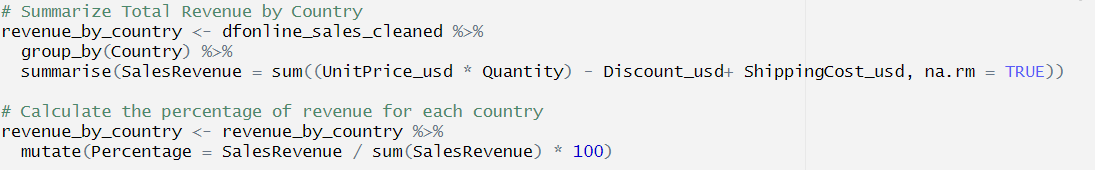


Figure 5.3: Total revenue calculations

In this chart, Norway and Sweden again are highlighted as the 2 extreme outliers. However, they might reflect a unique purchasing behaviour, hence we will keep it in our dataset.

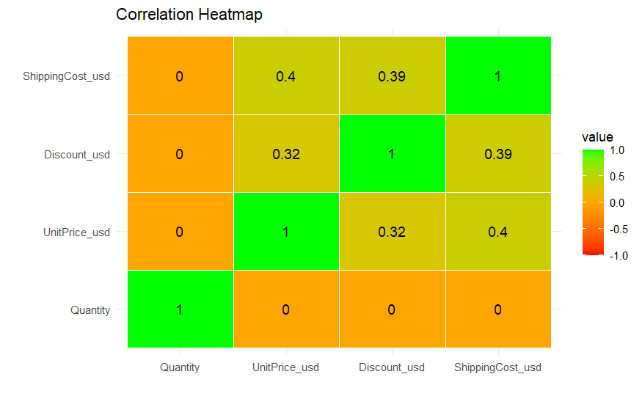


Figure 5.4 presents a heatmap of all the numeric features and it indicates that the features are generally correlated with each other except for Quantity.Itshows no significant correlation with any other variable, suggesting that it will not contribute to the predictive power of any model. As a result, Quantity will be removed from the dataset.

Figure 5.4: Heatmap of all numeric features

### 2.2.2 Categorical features

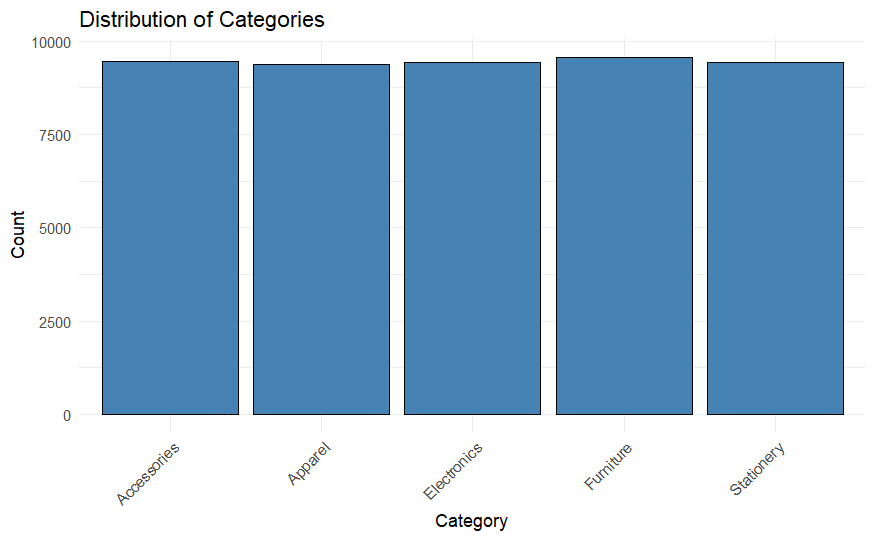
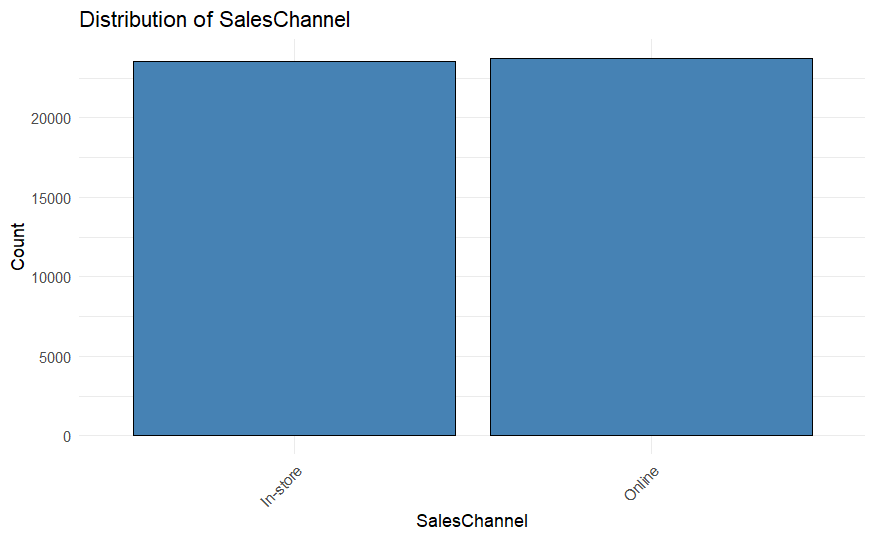
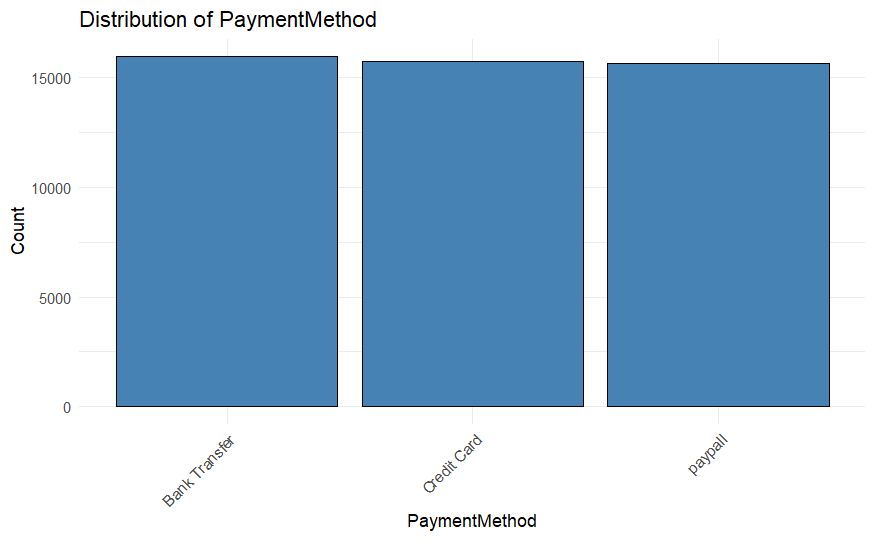
  

Figure 6.1: Distribution of SalesChannel

Figure 6.2: Distribution of PaymentMethod

Figure 6.0: Distribution of Categories

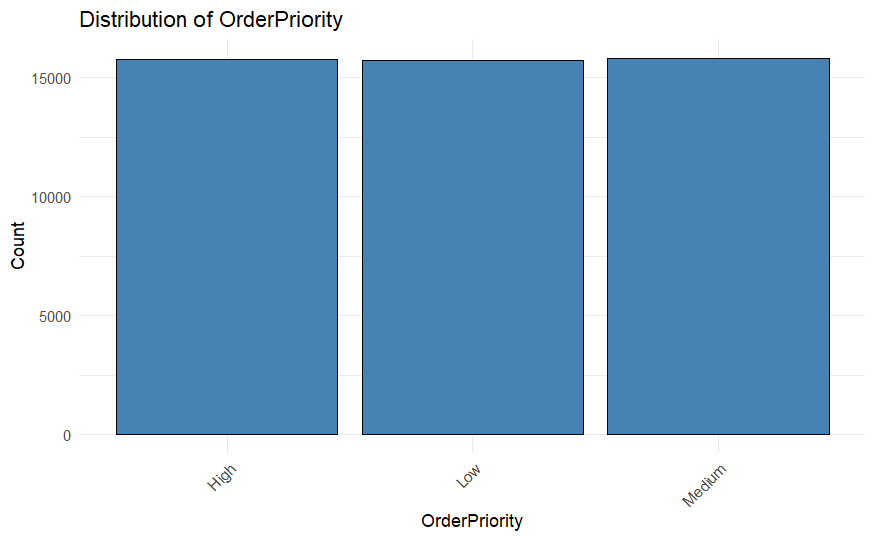
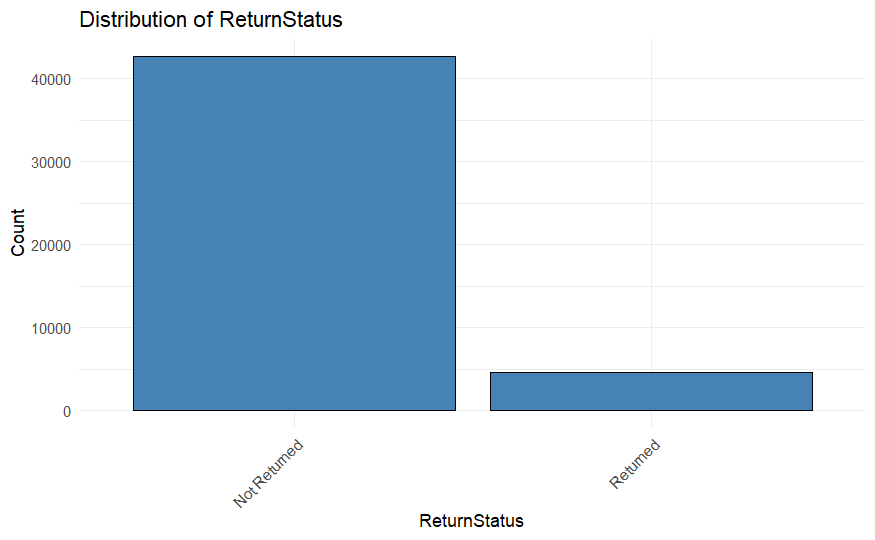
 

Figure 6.4: Distribution of ReturnStatus

Figure 6.3: Distribution of OrderPriority

Figure 6.0 to 6.4 present a bar graph for the different categorical variables, representing a well-distribution all across the different features. Figure 6.4 however, shows that the ‘Not returned’ input is highly skewed and thus ReturnStatus will be removed in our predictive models to prevent biases.

### **3. Unsupervised learning**

### 3.1 Substantive issue

In today’s highly competitive e-commerce landscape, businesses must go beyond traditional strategies to attract 1st time buyers and retain loyal customers. Understanding consumer behaviour is crucial in optimising marketing strategies and enhancing customer satisfaction. However, a major challenge lies in uncovering hidden patterns and segmenting customers effectively based on diverse variables present in a pool of data. This portion of the report specifically aims to identify customer clusters using unsupervised learning models to gain insight into purchasing trends and push for a more data-driven decision-making process.

### Research questions

The research questions (RQs) to be addressed

* RQ 1: How effective can the clustering techniques group consumers?
* RQ 2: What are the optimal clusters?

### Methodology

To aid in the research, the algorithms used for unsupervised learning will be

1. Principal Component Analysis
2. K-means clustering
3. Hierarchical clustering

The Country column will be removed for unsupervised learning to prevent having too many one-hot encoded variables and will be reintroduced in classification.

### Final modifications and analysis

Before applying our algorithms, we converted and encoded the categorical variables for unsupervised and classification. However, since Country contains too many unique values, we exclude it from unsupervised for clearer cluster visualisation. We also remove the non-encoded columns to prevent duplication and mismatch clusters. Finally with the removal of Quantity and ReturnStatus, our dataset now renamed to df\_processed is ready for analysis.

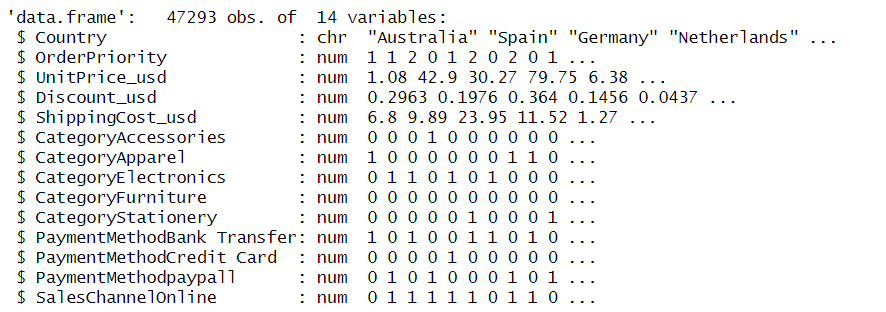
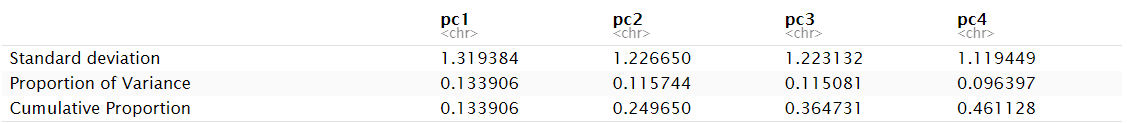


Figure 7: df\_processed

### Principal Component Analysis

Principal Component Analysis (PCA) is a linear dimensionality reduction technique to reduce number of variables in a data while retaining essential information and variance. These reduced variables are called principal components (PC) which captures and narrows down to the most significance patterns in the data whilst reducing complexity.



Based on figure 8.0, the standard deviation indicates that PC1 captures the most significance amount of variation in the data.

The proportion of variance indicates that PC1 explains 13.39% of the total variation.

The cumulative proportion indicates after PC4, 46.11% of the variation can be explained by the first 4 PCs. Overall, the PCA does reduce dimensions while retaining a good portion of the variance but the percentage could be higher.

Figure 8.0 Statical result of the PCA

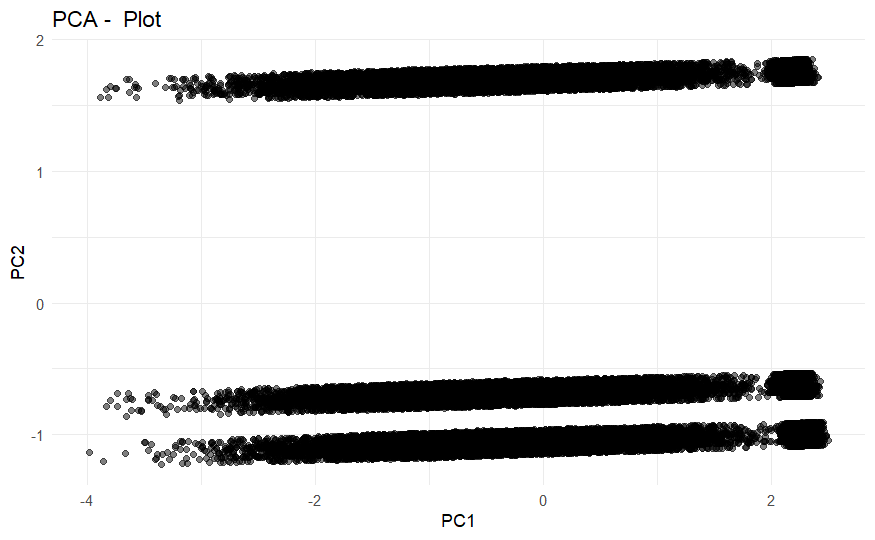


Figure 8.1 PCA plot

### K-means clustering

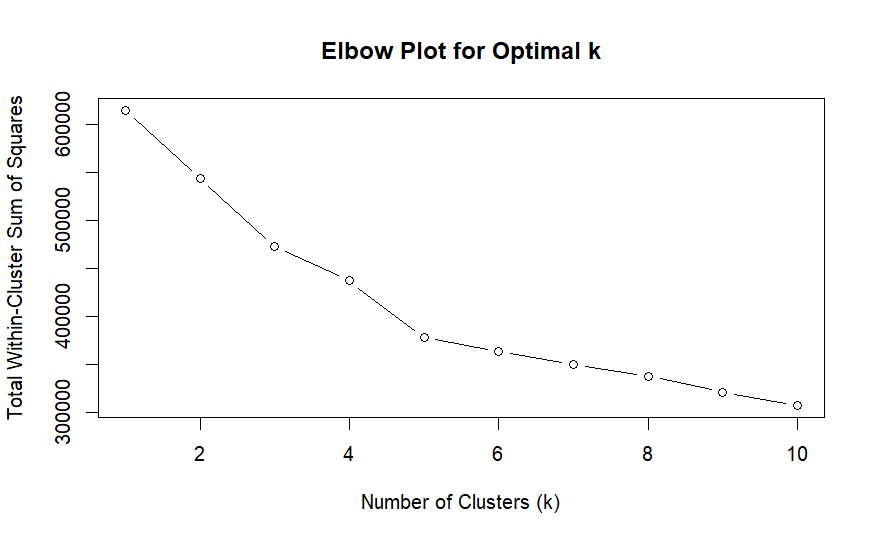
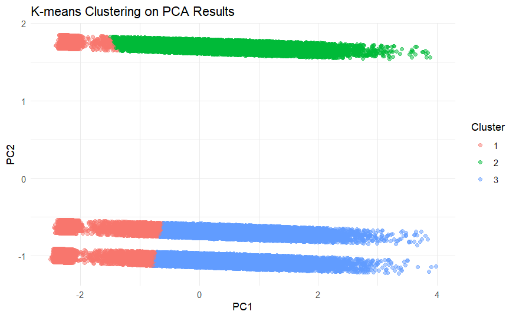
K-means clustering is an unsupervised algorithm that divides the dataset up into k clusters base on their similarity and proximity to each cluster’s centroid. It repeatedly assigns points to the nearest centroid and adjusts the centroids to minimise the squared distance. 

Figure 9.1 K-means cluster plot based on PCA results

Figure 9.0 Elbow plot

Both Figure 9.0 and 9.1 shows that the optimal number of clusters is 3. The similarity between Figure 9.1 and 8.1 suggest that the PCA plot has successfully transformed the dataset into structure where distinguishable patterns are shown. When K-means was added ontop of the PCA\_transformed space, the algorithm was able to segment the data into 3 defined clusters. The fact that these clusters align with the PCA distribution confirms that the algorithms has managed to capture meaningful patterns in the data, making segmentation reliable.

### Hierarchical clustering

Hierarchical clustering is an unsupervised algorithm that builds a hierarchy of clusters by either merging or splitting. The resulting dendrogram (a tree-like structure), visualises the relationships and distance among data points.

To determine the optimal clustering structure, we tested both Manhattan and Euclidean distance metrics across four linkage methods (Ward, Average, Complete, Single) setting clusters (k) to 3, resulting in a total of 8 hierarchical clustering dendrograms. We then used the silhouette method to evaluate each of the 8 dendrograms and selected the two best-performing models per distance: Ward’s linkage with Manhattan distance and Average linkage with Euclidean distance.

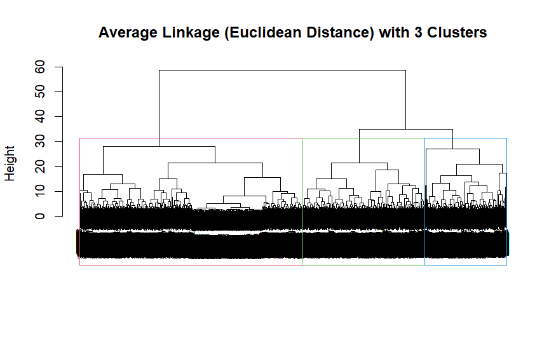
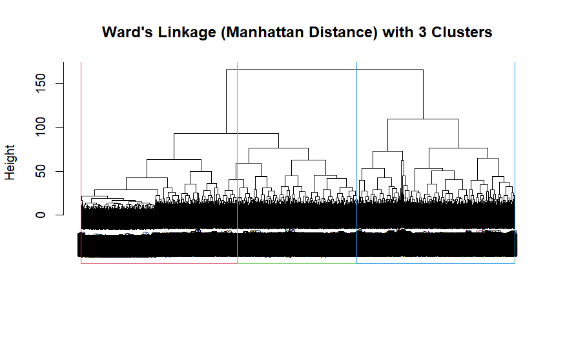


Figure 10.1 Average Linkage in Euclidean Distance

Figure 10.0 Ward’s Linkage in Manhattan Distance

These dendrograms illustrate how hierarchical clustering divides the data into three clusters. Figure 10 produces similar clusters, while Figure 10.1 allows for more flexible cluster shapes. The consistency between these two methods suggests a robust and reliable clustering structure like the previous algorithms used.

### **Classification**

### 4.1 Substantive issue

Consumer traffic generates vast amount of raw data, offering any business valuable insights into the market’s purchasing patterns and power. One way to harness said data is through classification models that predicts a consumer's demographic based on transactions details. This study applies machine learning classification to identify patterns in consumer’s country origin, helping businesses understand and accurately target regional markets.

### Research Questions

The research questions (RQs) to be addressed

* RQ 1: How accurately can a classification model predict a consumer's country of origin based on the data provided?
* RQ 2: Which of the 2 models are better in prediction?

### Methodology

To aid in the research, the algorithms used for unsupervised learning will be

1. K Nearest Neighbour (KNN)
2. Logistic Regression

For these algorithms, the df\_processed dataset will be used again but with the addition of the Country feature to serve as a target variable.

### Final modifications and analysis

Before implementation, we encode and convert country for usability similarly in 3.4. The dataset is renamed to df\_newlyprocessed to distinguish it from the one used for unsupervised and prevent the need of rerunning the entire process if an error arises.

### K nearest Neighbour (KNN)

K-Nearest Neighbours (KNN) is a supervised algorithm that classifies data points based on the majority class of their nearest neighbours in the same proximity. It works by measuring the distance between data points and assigning a label based on the most common label among the nearest neighbours to make predications

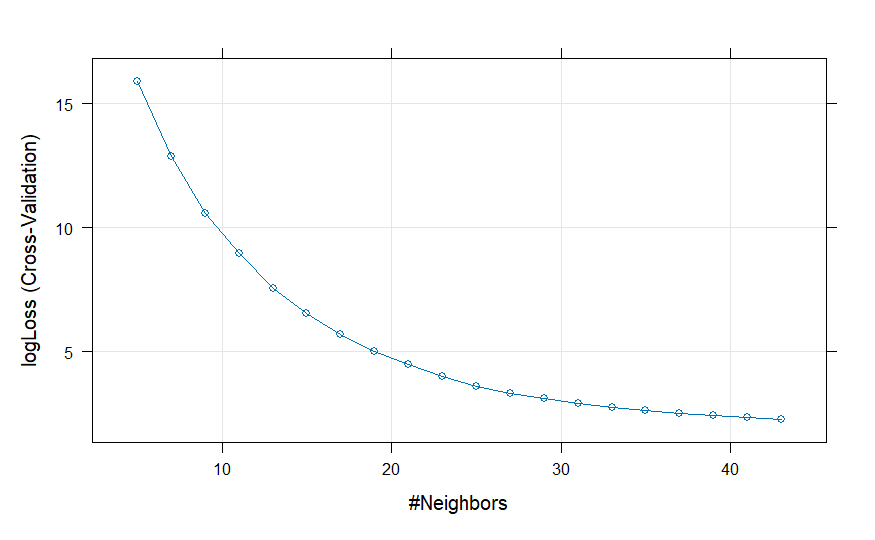
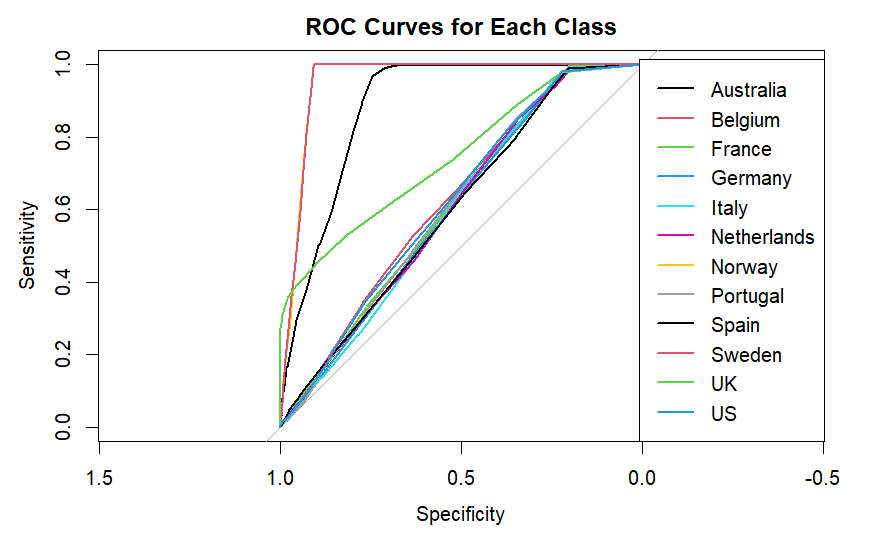
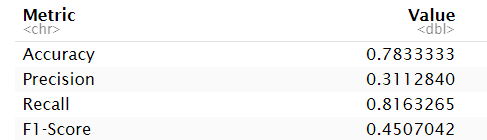
 

Figure 11.0 KNN with logloss

Figure 11.2 KNN’s ROC curve for each class

Figure 11.1 KNN’s confusion matrix evaluation metrics

Based on Figure 11.0, it indicates that the model is improving at predicting correct probabilities as k increases particularly when k > 20. In Figure 11.1, accuracy is 78% and recall is 81% which shows that KNN performs well in predicting the correct country and identifying the correct instances of the target class. However, the precision having 31% indicates a high number of wrongly identified countries. From Figure 11.2, all the Country’s ROC curve are generally performing well as its higher than the benchmark, indicating that the model is better than random guessing. Among them, Australia, Belgium and France have the steepest curve which shows high AUC score and model performance. However, most of the ROC are relatively flatter which indicate that the model is not performing optimally.

### Logistic Regression

Logistic Regression is a model that is used for binary classification tasks, where it predicts the probability of an event occurring. It uses a logistic function to model the relationship between the input features and the binary outcome, producing values between 0 and 1, which are then classified based on a threshold which are usually 0.5.

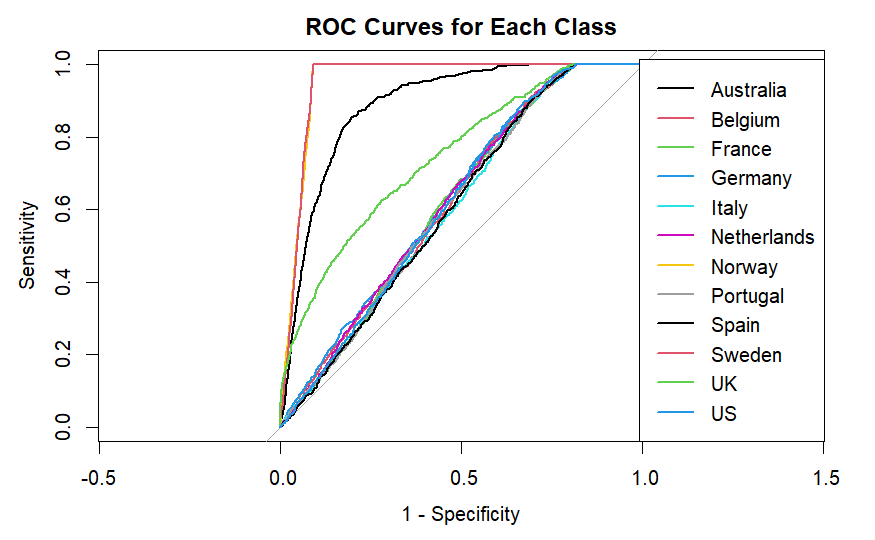
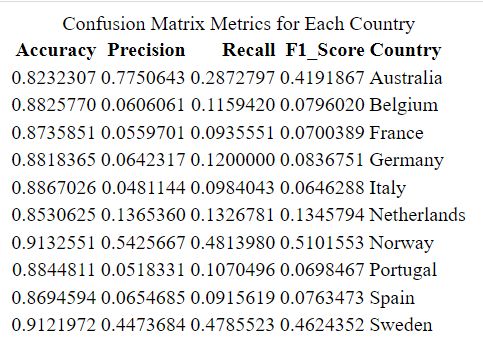
 

Figure 12.0 Logistic Regression’s ROC curve for each class

Figure 12.1 Logistic Regression’s confusion matrix metric

From Figure 12.0, all the Country’s ROC curve are generally performing well as its higher than the benchmark, indicating that the model is better than random guessing. Among them, Australia, Belgium and France have the steepest curve which shows high AUC score and model performance.

However, most of the ROC are relatively flatter which indicate that the model’s performance is suboptimal

Interestingly, these curves are generally similar to the KNN’s ROC, reflecting consistent performance across the board

In Figure 12.1, accuracy values are generally >80% across all countries, indicating high correct country prediction. Precision and recall values are relatively low which indicates high wrongly predicted countries and that the model struggles to correctly identify instances in the correct country. Overall, the Logistic Regression model performs better for countries like Australia and Norway in both model evaluations. However, in general, the model performs less optimally than the KNN based on their confusion matrix metrics, particularly in terms of precision and recall.

### **5. Dataset #2**

### Feature selection and modifications

This is the original dataset:

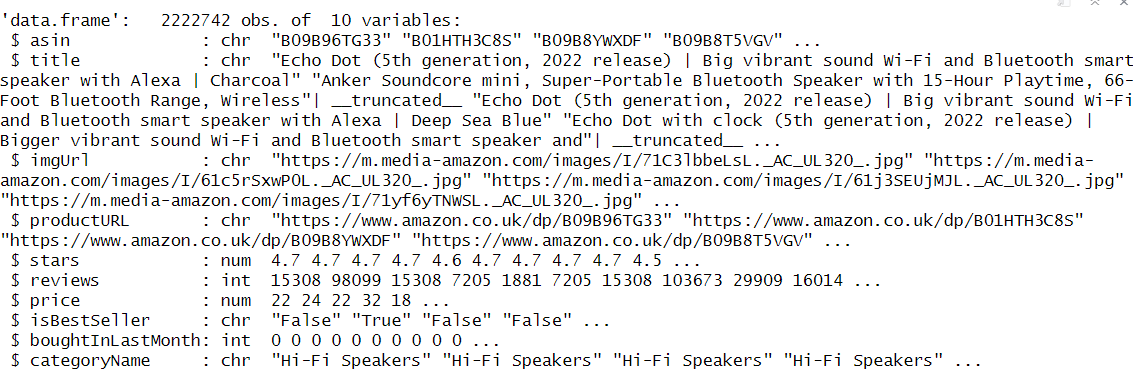


Figure 13: Total number of variables and data points in dfuk

First, we handle any missing values and then remove asin, title, imgurl, and productURL as they are irrelevant to our models. The refined dataset is labelled dfuk\_cleaned.

Next, we will modify the categoryName by grouping its 296 unique values into 5 subcategories obtained from Category (Dataset #1) for consistency and ease. Due to our large dataset, we prioritise mapping high-frequency counts while excluding mismatched and low-frequency counts. We have overall selected the top and 3rd most frequency counts to shrink the dataset.

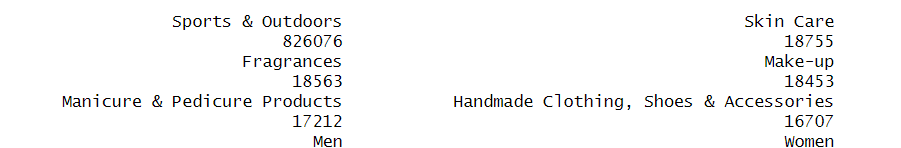


Figure 14: The frequency counts of unique names in descending order

For example, Handmade Clothing, Shoes & Accessories is mapped to Apparel as it matches that board subcategory and has the highest frequency count of that subcategory.

After mapping, we filter out the dataset to retain the mapped subcategories and rename it to dfuk\_modified to track changes. Continuing on with our dfuk\_modified, similar to dataset #1 we will convert the price from pounds to USD, storing it as price\_usd while removing the original column to avoid mistakes. Finally, we do the last check for any NaN values.

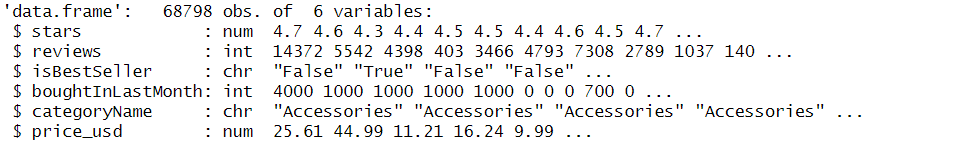


Figure 15: dfuk\_modified

### 5.2 Exploratory Data Analysis (EDA)

### 5.2.1 Categorical Features

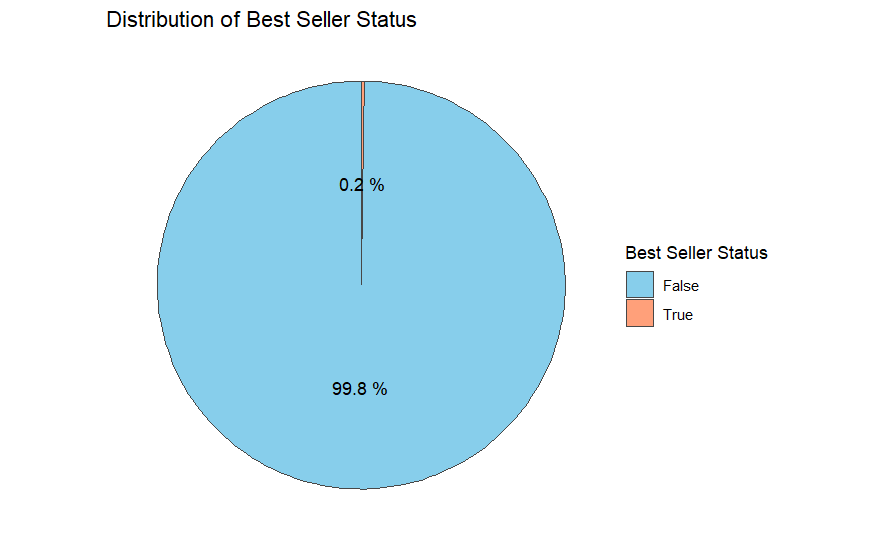
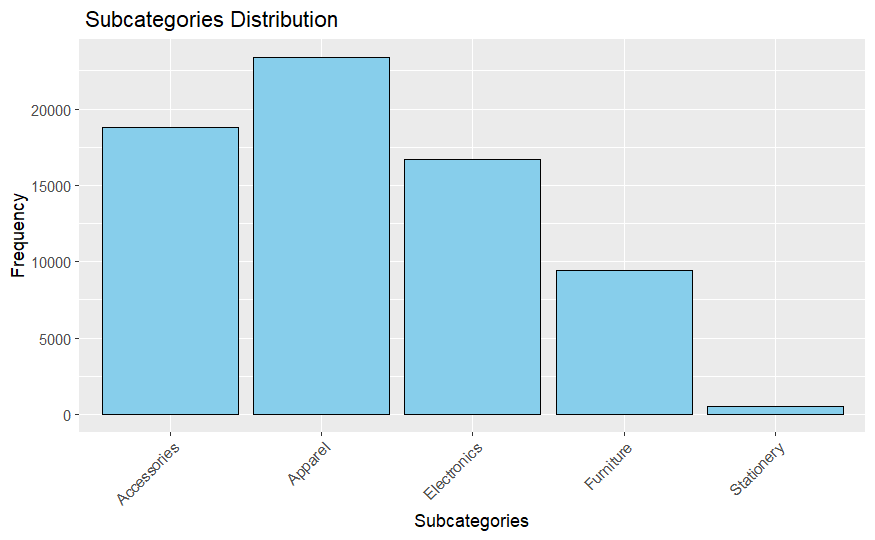
 

Figure 16.1: Distribution of the 5 subcategories mapped out

Figure 16.0: Piechart of whether product is a BestSeller or not

Based on both figures, it is evident that BestSeller is highly skewed towards ‘False’ and the frequency count of Stationery is extremely low compared to the rest. Hence, we will remove the BestSeller feature and the ‘Stationery’ subcategory from the dataset.

### 5.2.2 Numeric features

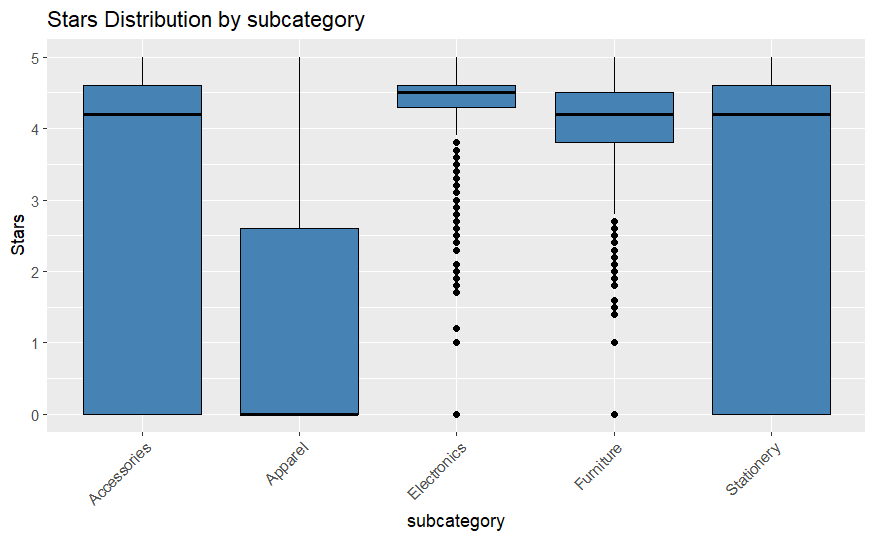
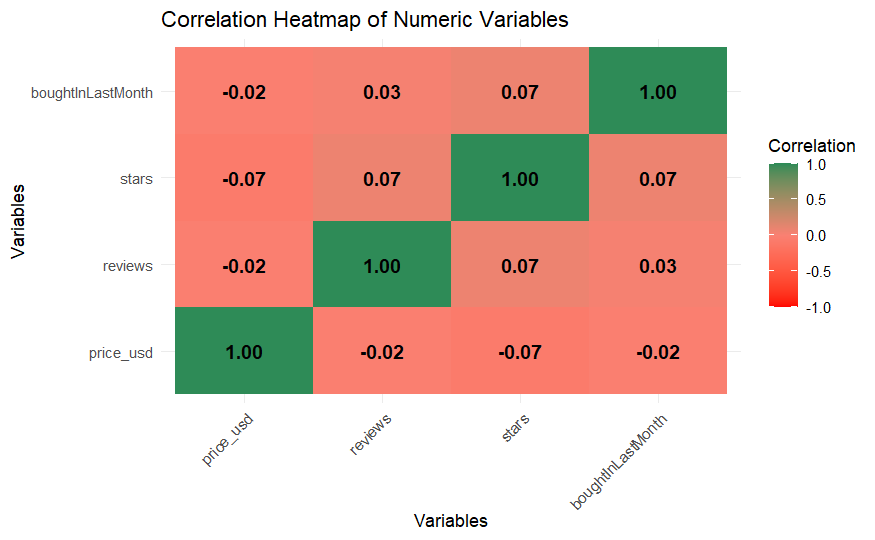


Figure 17.0 Correlation Heatmap of all numeric features

Figure 17.1: Boxplot of subcategories against stars

From Figure 17.0, all numeric variables exhibit some level of correlation with each other. Since there are no extreme outlier that lacks correlation with other variables, all of the features will be kept for Regression.

Figure 17.1, shows that Accessories and Stationery have highest star ratings while Apparel has the lowest but with more variation indiciated by longer whiskers.Overall, these 3 subcategories shows variability due to their larger box sizes. Electronics and Furniture have many outliers, suggesting inconsistent ratings. Their small box sizes also implies that most ratings a concentrated around a similar value. As such, extreme outliers (ratings of 0-2 for Electronics and 0-1.5 for Furniture) will be removed.

### **6. Regression**

### 6.1 Substantive issue

Instead of spending hours analysing raw data, customer satisfaction and purchasing trend can often lie in the product ratings. This regression model predicts product ratings based on various transaction and product-related features, focusing on the UK – a key Amazon market, where they plan to invest £8 billion in the UK to ramp up its cloud and AI infrastructures[[3]](#footnote-3) , highlighting the region’s strategic importance. Hence, by identifying the key drivers behind consumer ratings, businesses can better assess product performance and make informed decisions according to the market demands.

### 6.2 Research Questions

The research questions (RQs) to be addressed

* RQ 1: How accurately can a regression model predict a product’s rating based on transaction and product-related features?
* RQ 2: Do Regularisation models do help in improving prediction?

### 6.3 Methodology

To aid in the research, the algorithms used for unsupervised learning will be

1. Linear Regression
2. Regularisation models (Lasso, Ridge, Elastic net)
3. Random Forest

### 6.5 Final modifications and analysis

Similar to dataset #1, we will only convert and encode the categoryName variable and the renamed dfuk\_modified\_cleaned will be used that has undergone all the modifications carried over from 5.2.

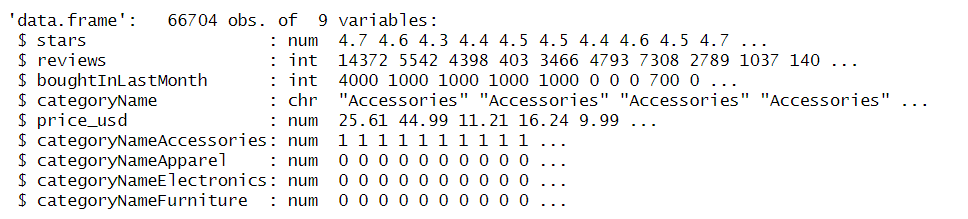


Figure 18: dfuk\_modified\_cleaned

### 6.5.1 Linear Regression

Linear regression is a supervised algorithm that shows the relationship between a dependent variable (no. of stars) and >1 independent variables by fitting a linear equation.

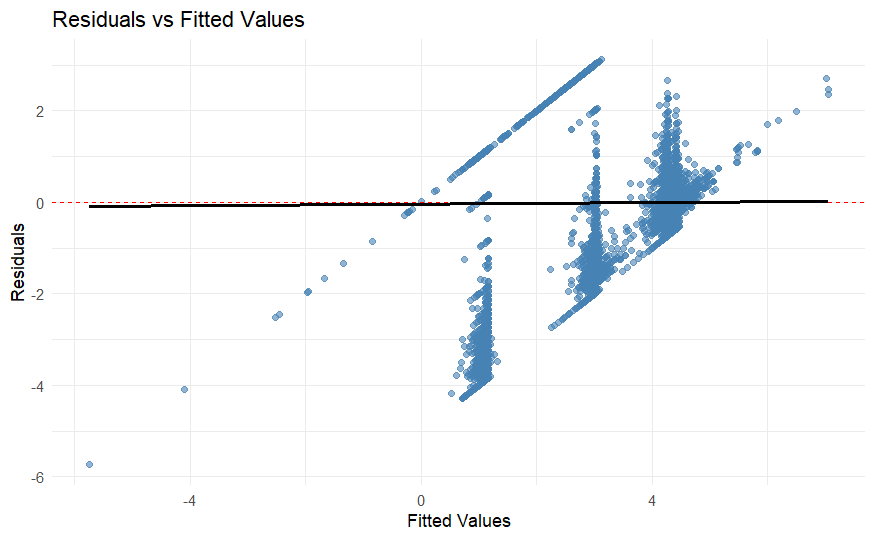
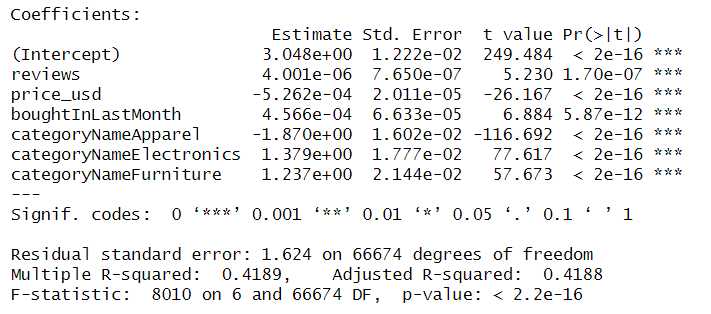


Figure 19.1 Residuals vs Fitted values

Figure 19.0 Statistics of the Linear Regression

Figure 19.0, price\_usd has a negative coefficient, higher prices tend to decrease star ratings. While reviews and boughtInLastMonth have positive coefficients, more reviews and products bought directly increases star ratings. It is observable that categoryNameAccessories has been used to compare with other features in categoryName hence its absence in the statistic table. categoryNameApparel has a negative coefficient while the other 2 have positive coefficients. This means, apparel products have lower star ratings than accessories while electronics and furniture receive higher ratings.

The Adjusted R-square indicates that 41.89% of the variation can be explained by the model’s predictors which could be better. Since the F-statistic is large while the p- value is extremely small, we can conclude that the model is significant. Figure 19.1 shows that most of the residuals are evenly distributed around the fitted values, indicating good fit of the model. However, it also shows possible non-linearity and presence of influential outliers that were not filtered.

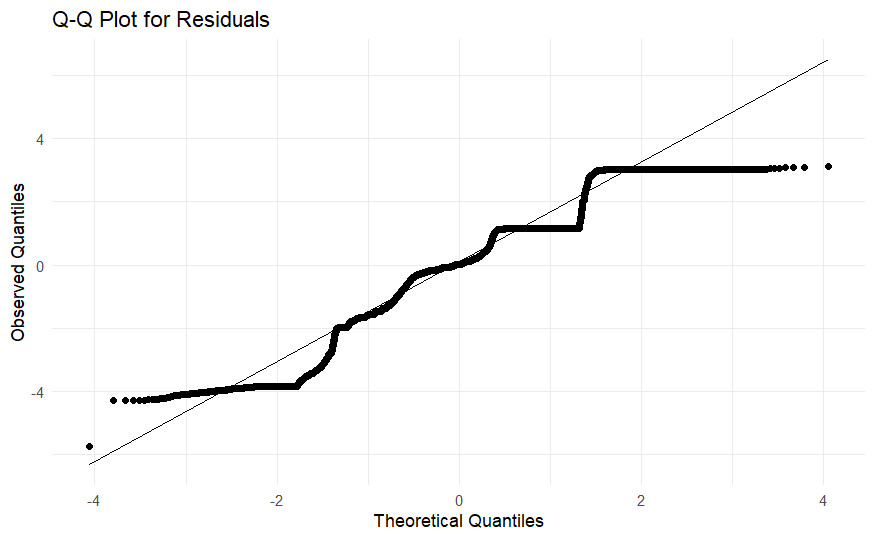


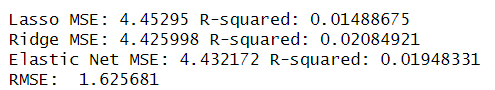
Figure 19.2 Q-Q plot for residuals

Figure 19.2 shows that the curve is slightly zig-zagged but the residuals mostly do follow the diagonal line, so they are roughly normal. There is also no extreme curvature which means no big problems with prediction.

### Regularisation models (Lasso, Ridge and Elastic net)

Regularisation models, like Lasso, Ridge, and Elastic Net, are techniques to help prevent overfitting by adding penalties to the Regression model. Lasso performs feature selection, Ridge shrinks coefficients of less important features, and Elastic Net combines both approaches.

Figure 20.0 Statistics of the Regularisation Models



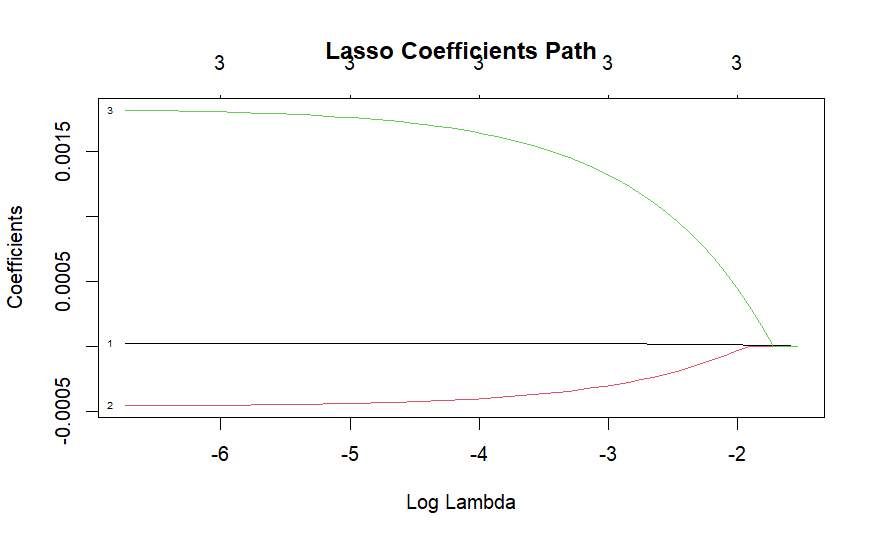
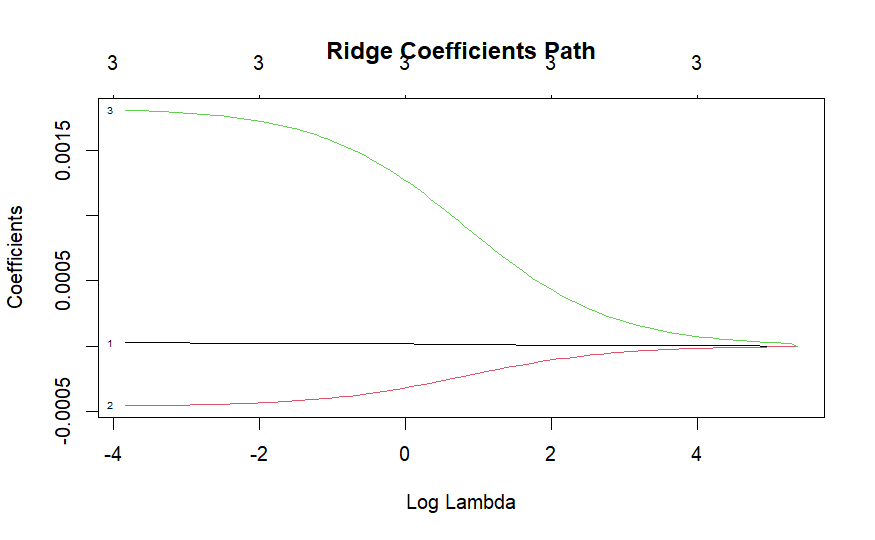
 

Figure 20.1 Lasso

Figure 20.2 Ridge

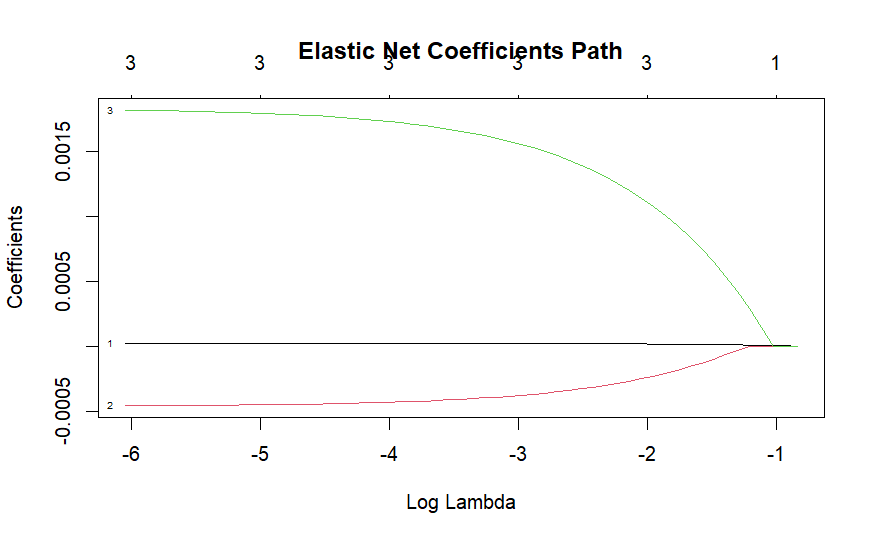


Figure 20.0 indicates that all 3 models don’t explain much variation due to the low R-squared and high RMSE. However, Ridge does slightly outperform the other 2 regression models, leaving room for more improvements in their predictive power. Figure 20.1 to 20.3 shows that Lasso and Elastic net take longer for regularisation to occur, as seen from how flat their curves are for negative log lambda values. While Ridge regularisation is generally faster as their coefficients do shrink as log lambda increases. Overall, these Regularisation models do not drastically improve our prediction accuracy but it does still help to prevent overfitting.

Figure 20.3 Elastic Net

### Random Forest

Random Forest is an ensemble learning technique that combines multiple decision trees to make and improve prediction.

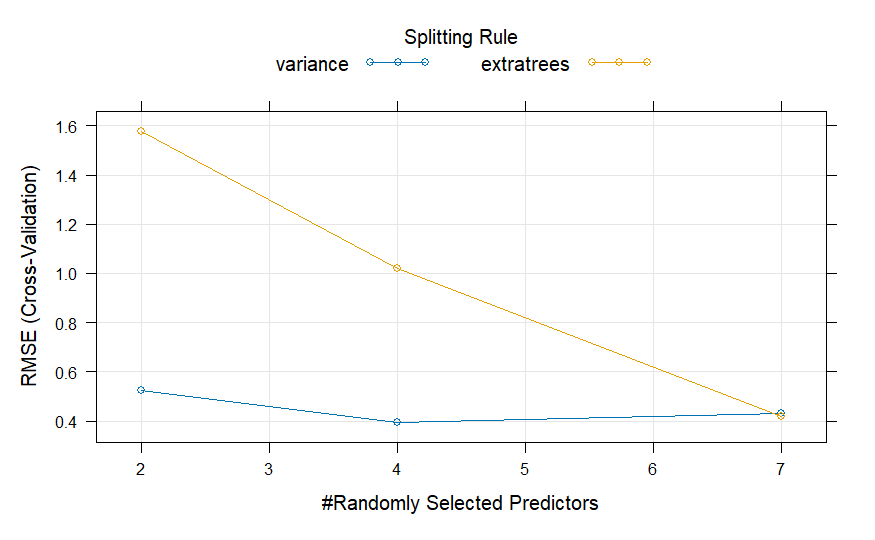
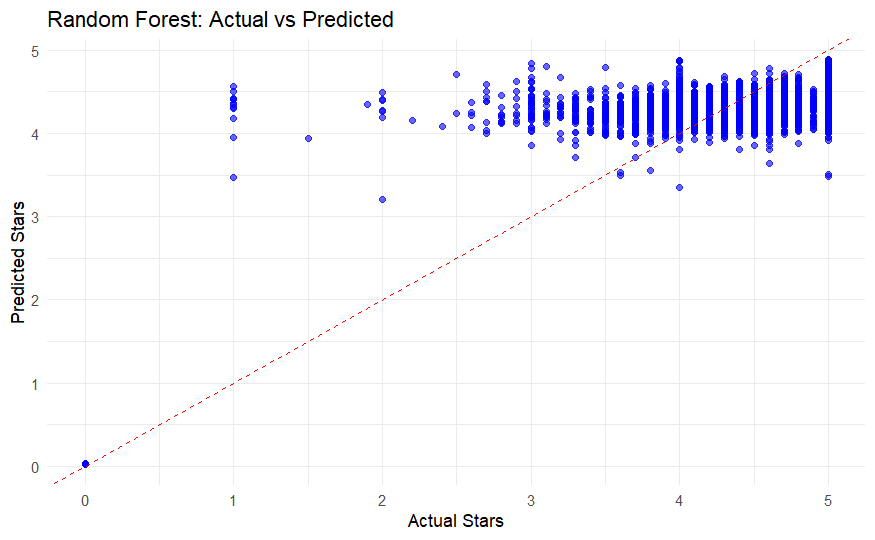
 

Figure 21.0 Random Forest

Figure 21.1 Actual vs Predicted

Figure 21.0 shows that the extratrees method have a negative linear trend while the variance method is generally a straight line. Both methods result in a low RMSE meaning that they are both fit models. Overall, the variance method is better if stability is prioritised as it continuously has a low RMSE while extratrees method may be preferred for maximum performance.

Figure 21.1 shows that while, the Random Forest captures general trends in star ratings, it has a tendency to overestimate low ratings and underestimate high ratings. Hence suggesting that the model is decently effective but tends to gravitate to mid-ratings predictions. Figure 21.2 indicates that most of the points are generally near the 0 range which means the model is performing relatively well. However, there are significant deviations which could mean that the model is not performing to its maximum ability and needs to be improved.

Overall, while neither model is perfect, Random Forest appears to perform better than Linear Regression in capturing complexity. However, it does need extra tuning to avoid mid-range biases.

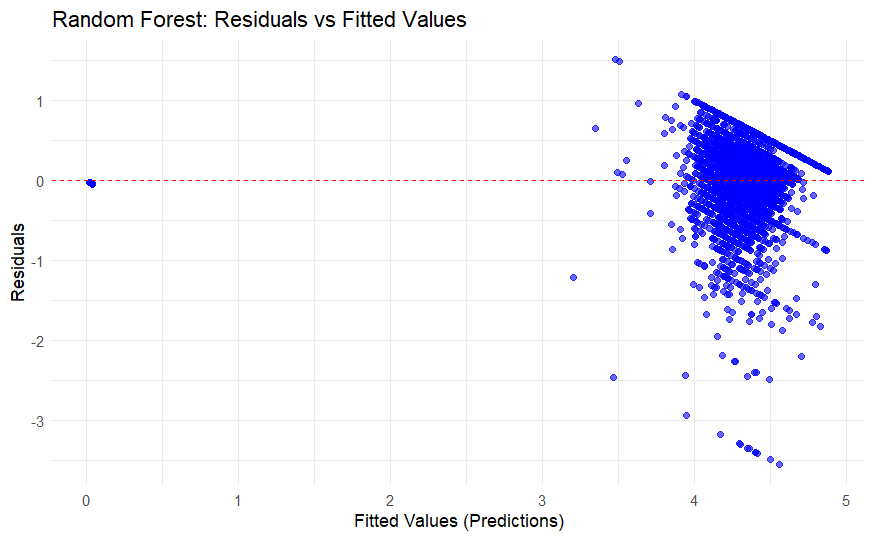


Figure 21.2 Residuals vs Fitted

### **APPENDIX**

1. (Goyal, September 26, 2023) E-commerce surge can add $1 trillion to US retail sales by 2027. Available at: <https://www.bloomberg.com/professional/insights/trading/e-commerce-surge-can-add-1-trillion-to-us-retail-sales-by-2027/>

2. (Curry, Aug 28 2024) Why companies including JPMorgan and Walmart are opting for internal gen AI assistants after initially restricting usage. Available at: <https://www.cnbc.com/2024/08/28/why-jpmorgan-and-walmart-are-opting-for-internal-gen-ai-assistants.html>

3. (Kharpal, Sep 10 2024) Amazon makes £8 billion UK investment to build cloud and AI infrastructure. Available at: <https://www.cnbc.com/2024/09/10/amazon-makes-8-billion-uk-investment-to-build-cloud-and-ai-infrastructure.html>

1. https://www.bloomberg.com/professional/insights/trading/e-commerce-surge-can-add-1-trillion-to-us-retail-sales-by-2027/ [↑](#footnote-ref-1)
2. https://www.cnbc.com/2024/08/28/why-jpmorgan-and-walmart-are-opting-for-internal-gen-ai-assistants.html [↑](#footnote-ref-2)
3. https://www.cnbc.com/2024/09/10/amazon-makes-8-billion-uk-investment-to-build-cloud-and-ai-infrastructure.html [↑](#footnote-ref-3)