Genet Sebehat

MSDS692 - Practicum I

UK Online Retail Project Final Summary

**Problem statement**

This project examines the data of a United Kingdom (UK)-based online retail agency and aims to identify the agency’s potential business problems and providing solutions using data science methods. The big picture of the project is to study the company’s customer behavior in order to maintain the growth of customer loyalty and as a result expand the agency’s business. The methods used to identify the problem and provide the best solutions are data analysis and statistics techniques that include machine learning methods. In general, the company requires outstanding strategies to solve the common e-commerce problems that are hindering its business.

**Project Overview:**

Data for the UK-based non-store online retail data contains online transaction data. The UK has the most advanced e-commerce market in Europe. The online shopping landscape in the UK continues to grow, with more and more consumers choosing to make purchases online that involve only electronic transactions. Data for this project was acquired from the UCI Machine Learning Repository. This data represents online transactions. The goal of the project is to use data science techniques to identify, define, and analyze business problems. Moreover, my aim with this project is to gain analytics experience and to reconcile mathematical theory with business practices.

**Type of Data Science Learning**

This project will incorporate several different types of classification models using supervised and unsupervised learning. This project performs clustering to find interesting groups of customers and will use a few models such as random forests, KNN, and linear regression to find the best model that will be useful for the business and that they might be used to build their future model.

**Objective**

The objective of this project is to analyze the data to help the organization to improve its sales (total revenue).

**Key analysis of the project**

* Show an understanding of the data through descriptive statistics and visualization
* Use clustering to find interesting groups of customers
* Compare the performance of these methods
* Recommend strategies for the organization
* Produce a report

**Perform the analysis and discussion of this data set as per below sections:**

* About the Data Set
* Pre-processing of the Data set
* Exploratory and Descriptive Analysis
* Preparing a Data Frame for Clustering
* Clustering for Customer segmentation
* Exploratory and Descriptive Analysis of Three Clusters
* Create models and compare it

**Tools Used**

* I analyzed the data set using the R programming language with RStudio editor.

**About the Data Set**

This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.

***Source:***

Dr Daqing Chen, Director: Public Analytics group. chend ‘@’ lsbu.ac.uk, School of Engineering, London South Bank University, London SE1 0AA, UK.

**Data Set Information:**

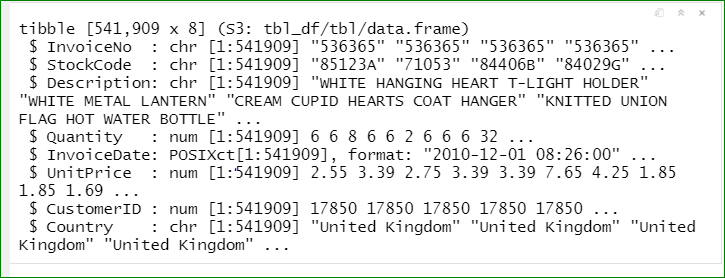
|  |  |
| --- | --- |
| Online Retail Data Set | Dataset Descriptions |
| Data derived | UCI Machine Learning Repository |
| Online Retail Dataset | 541,910 rows and 8 columns (attributes) |
| Data represents | Online transactions |
| Company | A UK-based and registered non-store online retail company |
| Products | Mainly all-occasion gifts |
| Transaction | Transition period 12/01/2010 – 12/09/2011 |

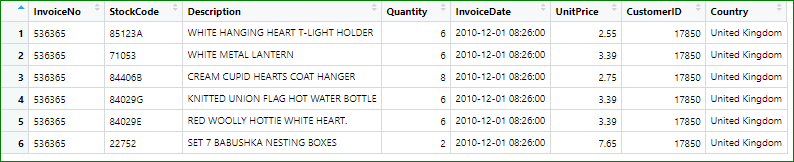
***Attribute Information:***

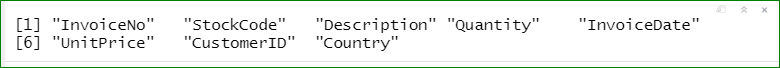
* InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter ‘c’, it indicates a cancellation.
* StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
* Description: Product (item) name. Nominal.
* Quantity: The quantities of each product (item) per transaction. Numeric.
* InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated.
* UnitPrice: Unit price. Numeric, Product price per unit in sterling.
* CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
* Country: Country name. Nominal, the name of the country where each customer resides.

Load “Online Retail” data set into statistical software “R” for our further analysis.







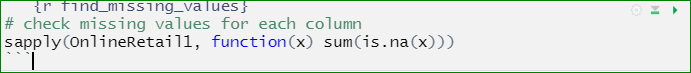


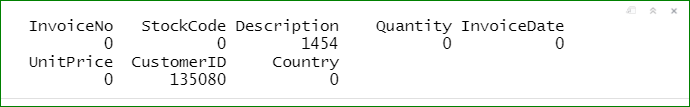
* **Pre-processing of the Data set**

Explore the data set and we encounter the need we will pre-process the data set to make it suitable for our further analysis of clustering of customers and prediction of sales.

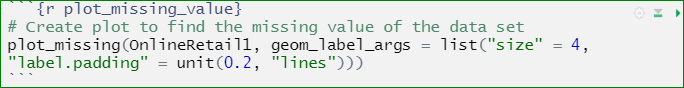
**Pre-processing\_1**

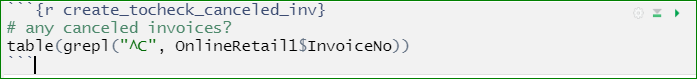
* check for any NAs in data set.
* any cancelled transactions.



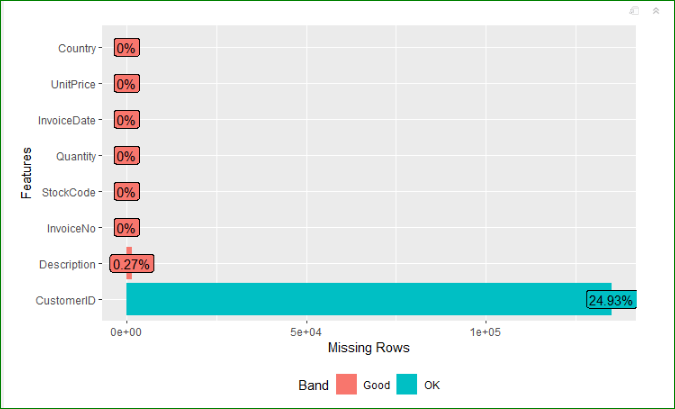


As the above result shows, Customer ID variable has 135080 NAs



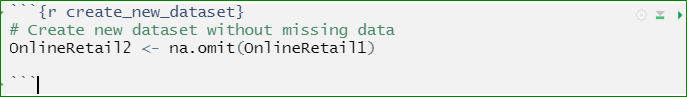






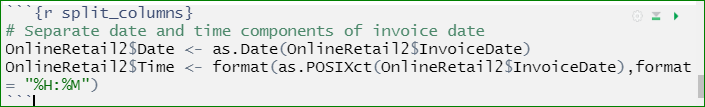
Looking at the size of the dataset and the missing value plot, it seems as if I can remove the missing values and still have a good-sized set of data to work on, I thus let me remove the missing values

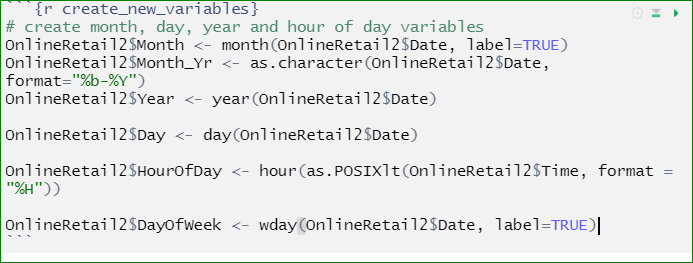
Remove NAs

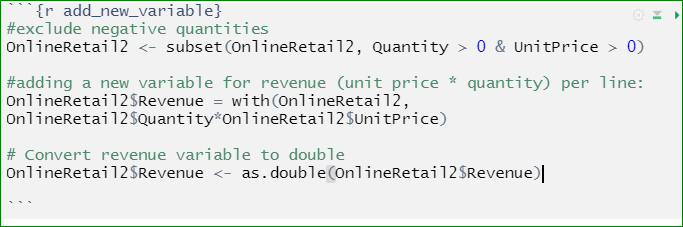


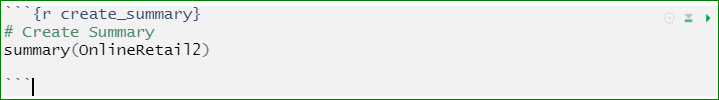
**Pre-processing\_2**

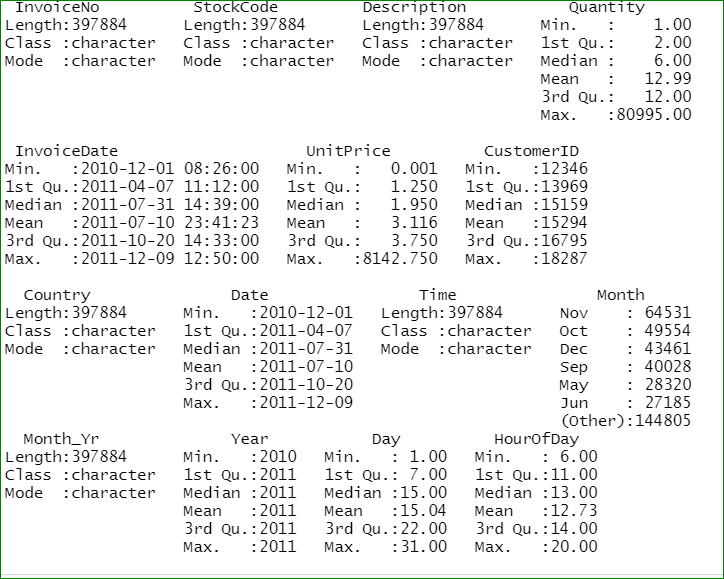
* format the InvoiceDate variable
* Create "Month", "Month\_Yr", Year", "Day", "HourOfDay" and "DayOfWeek" variables to work with future analysis
* add a new variable for “Revenue” (unit price \* quantity).













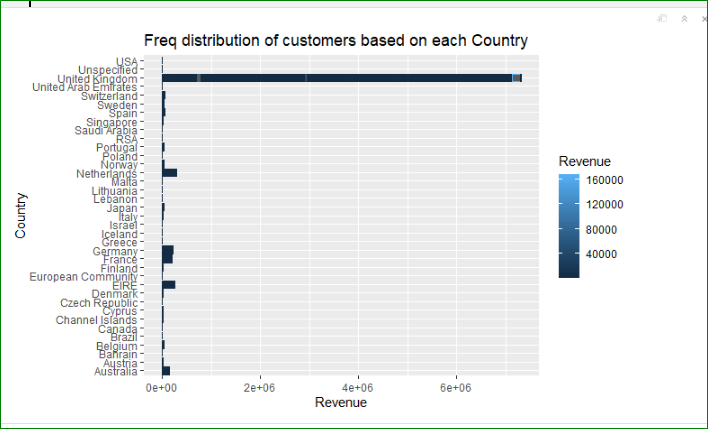
**Explanatory data analysis (EDA)**

* Exploring and visualizing data helps to validate the businesses assumptions via thorough investigation
* Avoids any potential anomalies in data to avoid feeding incorrect data to a machine learning model
* Clarify the model output and test the assumptions

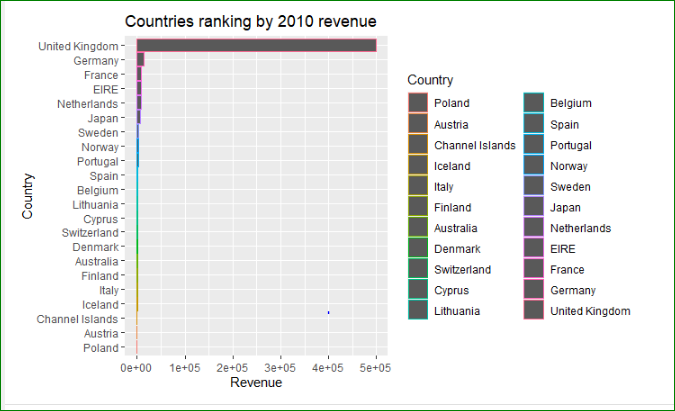


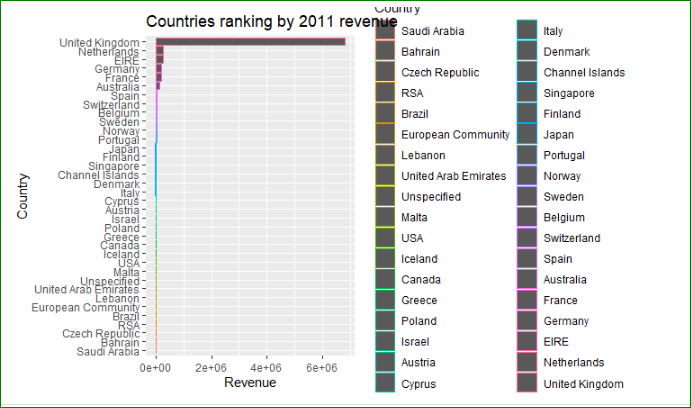
Based on the above output we know that the numbers of customers from Australia is 1182, from Austria is 398, from Bahrain is 17, from Germany is 9040, from Saudi Arabia is 9 and so on. So, the country with the most customers is in the United Kingdom with 354321 customers.

Plot frequency distribution of customers based on each Country:

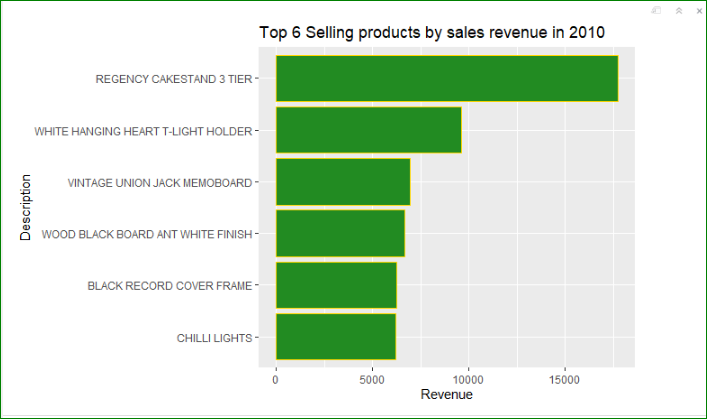


The topmost Countries contributing to revenue in 2010 and 2011:

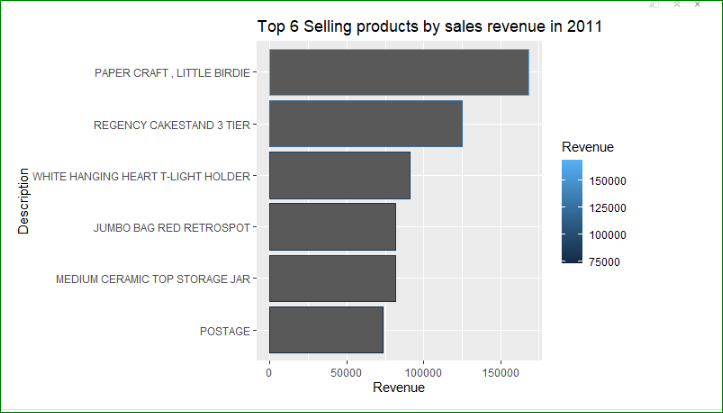




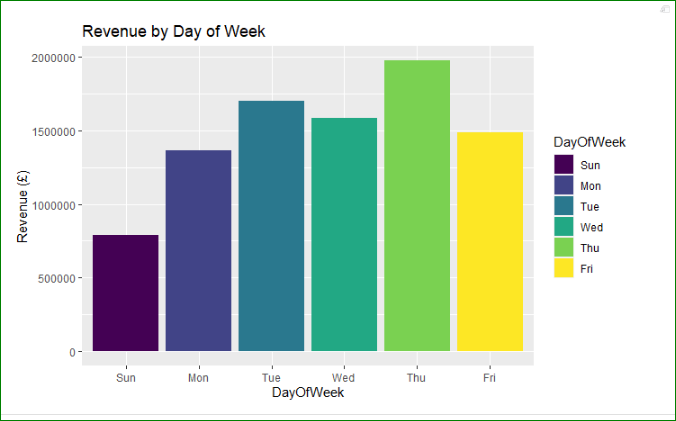
Top 6 Selling products by sales revenue in 2010 and 2011:



As the above ggplot shows, the “Regency cake stand 3 tier” has the most sales revenue in 2010.

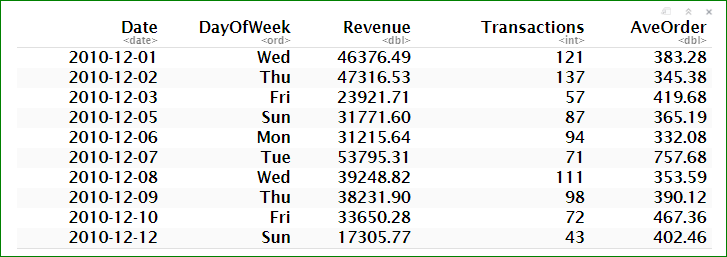
As the above ggplot shows, the “Paper craft, Little birdie” has the most sales revenue in 2011.

**Revenue by day of the week**



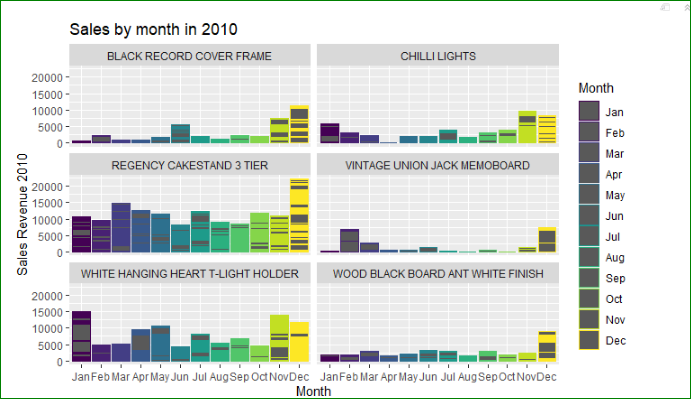
There were no transactions on Saturday throughout the whole period. I think behind the scenes there are an issue in the context of the metadata of the dataset. As shown in the above plot, a trend where the number of transaction increases from Sunday to Tuesday and decreases on Wednesday. Again, the trend for the number of transaction increases from Wednesday to Friday.

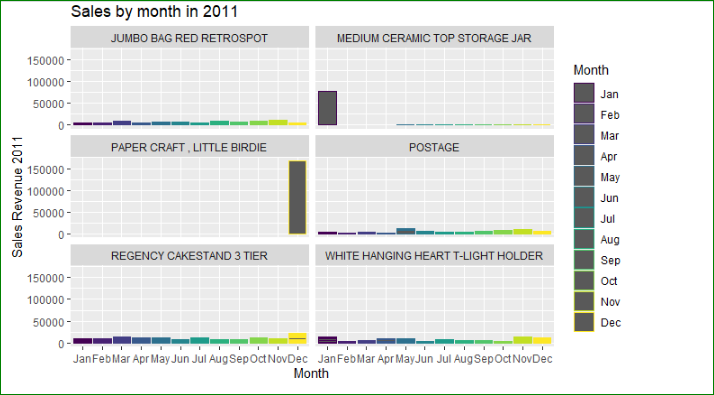
Below I create a new data frame that can be used to look at what's going on at the day-of-the week level in a bit more detail.



As shown in the above table, the greatest number of transactions (137) was made on Thursday.

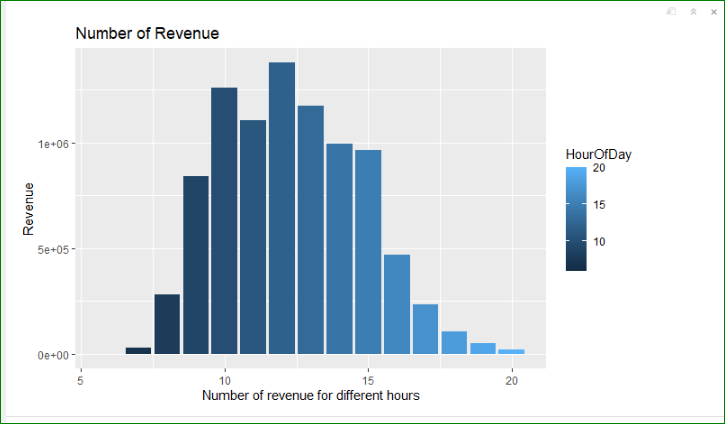
Monthly performance of the top six-selling products in 2010

Monthly performance of the top six-selling products in 2011

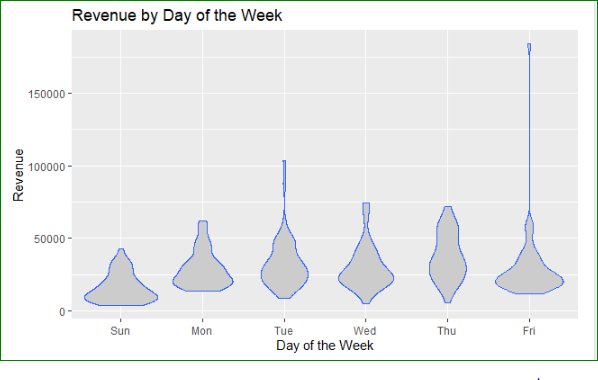


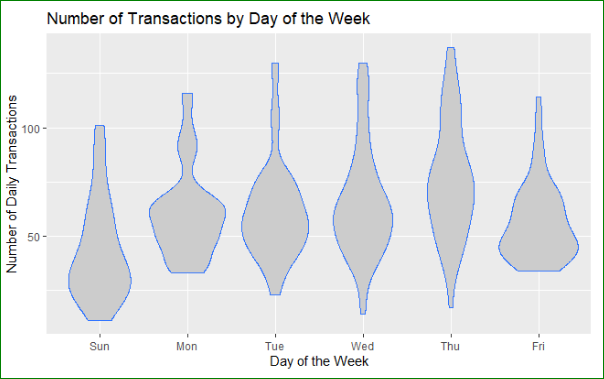
As the above plot shows, the sales of products changes over time. In terms of months, there were high number of transactions in December. This could be partly due to the fact that most customers make purchases during the holiday season.

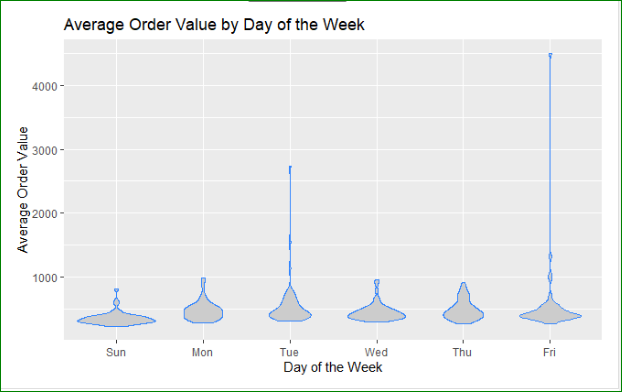
Amount of revenue by hour:



In terms of hours, there are no transactions after 8:00pm until the next day at 7:00am. The busiest hour of the day is around 12:00 p.m. One of the reasons could be due to the fact that most customers make purchases during lunch time and also mid-morning (in tea breaks) hour around 10:00 a.m.

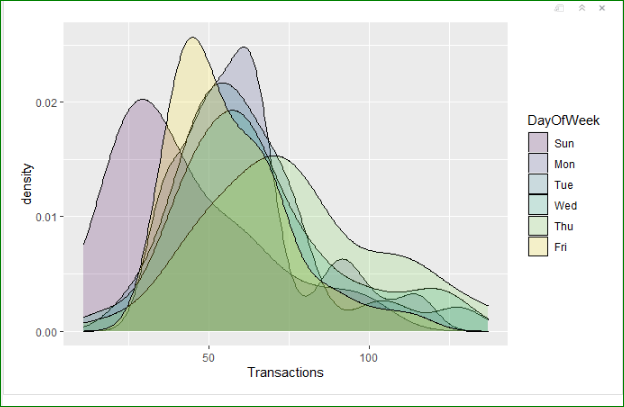






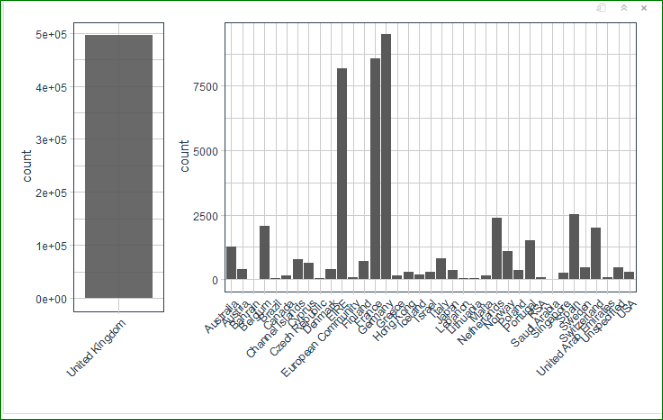
As shows the above plots, the amount of revenue of each day of the week is different. Thus, this difference is determined by a difference in the number of transactions, rather than the average order value.

Density plot to check how these data are distributed:

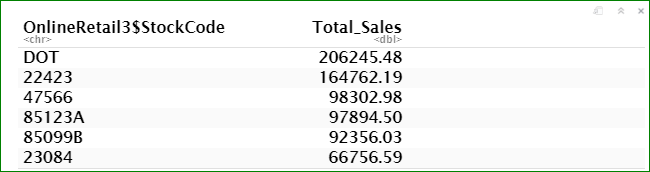


As shown in the above density plot, the tails point to the right indicating a positive skew. There is a reasonable amount of skewness in the distributions.

Transactions by country: The online retailer is UK-based, but its customers come from all over the world. However, the plots below tell us very quickly that the main customer base is from the UK, followed by Germany and France.

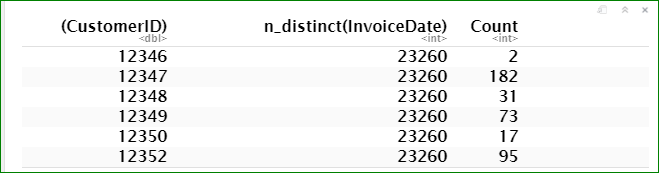


**Which products bring in the most revenue?**



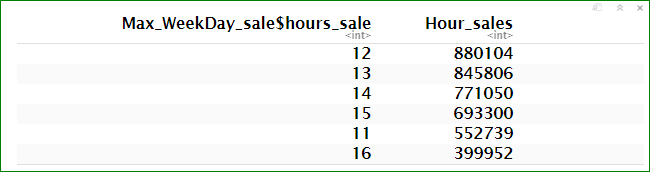
Of the various types of products sold there are several products that provide the largest revenue for the company, 6 of which are the selling product code of DOTCOM POSTAGE (DOT), selling at 206245.48, Code 22423 sold at 164762.19, cold 47566 sold at 98302.98, code 85123A sold at 97894.50, code 85099B sold at 92356.03, and code 23084 sold at 66756.59.

Which customers are repeat purchasers?



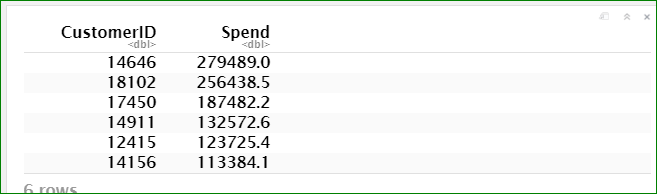
As with above output, this shows that the six top customers in terms of repeat purchases: were customers 12346, 12347, 12348,12349,12350, and 12352.

Which hours are most crowded?



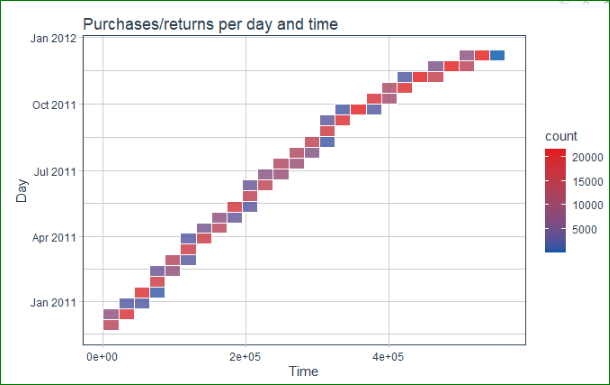
Based on the above output, we know that the most trafficked hour is 12 pm, with 880104 sales, and this activity continues until 3 pm. Sales are also high at 11 and 10.

Who are the top 6 customers with the most purchases?



As in the above output, the customers who make the most purchases are customer who makes the most purchases is customer 14646. The five others are customers 18102, 17450, 14911, 12415, and 14156.

Create ggplot to plot purchases/returns per day and time



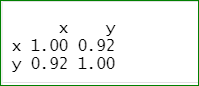
As the above plot shows, more transaction (purchases and returns) occurred between November and January because people purchase products to prepare for Christmas and New Year. That is what accounted for the volume of purchases and returns over the holiday season, especially for Christmas. People like to use their lunch hours to shop online.

**Recommendations**: Based on the results of the analysis, I would make the following recommendations to the UK online retail or (e-commerce) company as follows:

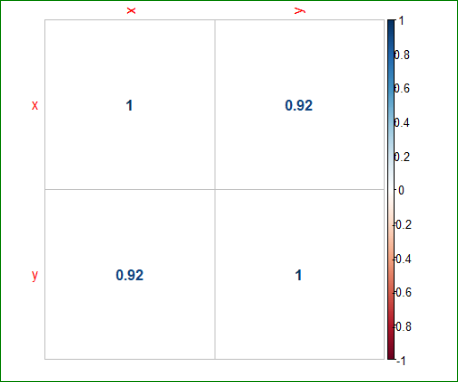
* Provide a bonus or door prize for customers with the highest number of purchases.
* Increase the number of staff to help customers, especially for UK
* Increase the number of staff with shifts on Thursday, especially at 12 pm.
* Increase stock of the highest selling products.
* The company may need to hire temp employee over the holiday season.
* Update their websites and database frequently.
* Ensure that the website (platform) is as user-friendly as possible.

**Correlation:** Correlation analysis is used to investigate the association between two or more variables.

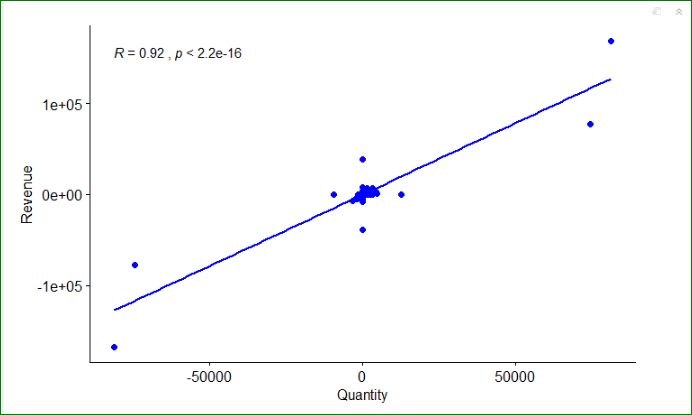
Correlation matrix: x = Quantity and y = Revenue



**Create correlation plot**



**Correlation plot**



**Clustering**

* Clustering is the procedure of partitioning a set of observations into a set of meaningful subclasses or clusters.
* For example, clustering can help business to discover distinctive group of customers for whom they might develop targeted marketing programs.

**Clustering methods**

* Partitioning techniques
* Finding centers of clusters among the observations and assigning each one to the cluster that has the closest center. (e.g., Kmeans, which is an unsupervised learning technique).
* Hierarchical techniques
  + Mode-base methods
    - Using probabilistic distribution to create clusters (e.g., mixture models).

Customer segmentation for this analysis was displayed based on customers’ purchasing behavior. The purchasing behavior features to be analyzed were Recency, Frequency and Monetary Value (RFM)

For the scope of this project, RFM was defined as follows:

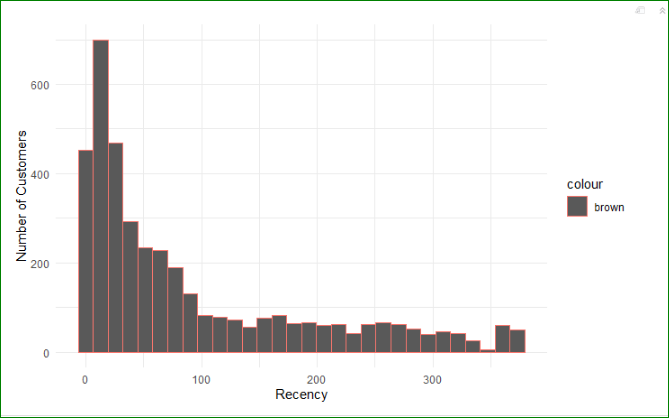
* Recency: Number of days the user has been inactive, from the moment of last purchase to the latest time in the dataset.
* Frequency: Total number of transactions completed by a customer in a year.
* Monetary Value: Total revenue generated by the customer in a year.

**The recency statistic indicates how long a customer has been inactive.**



As the above summary of statistics shows, 50% of the site’s users show fewer than 49 days of inactivity. On average, however, customers went three months without making a single purchase, and a small number of customers had not made a single transaction in over a year. 50% of the users had fewer than 49 days of inactivity.

**Recency Histogram**



Based on the above histogram, most clients have been active over the last 90-100 days. There are no benchmarks to compare this with, but three months of customer inactivity does not sound terrible, especially after it was discovered in the previous analysis that the company has not been acquiring users as quickly as it used to. However, some of these customers have gone over 140 days without making a single purchase.

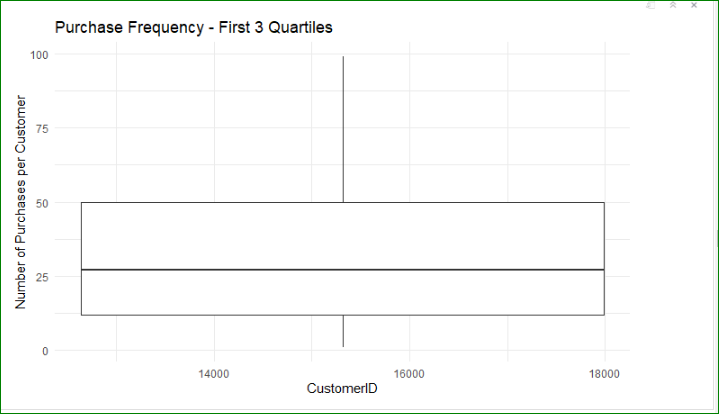
**Frequency Statistics**



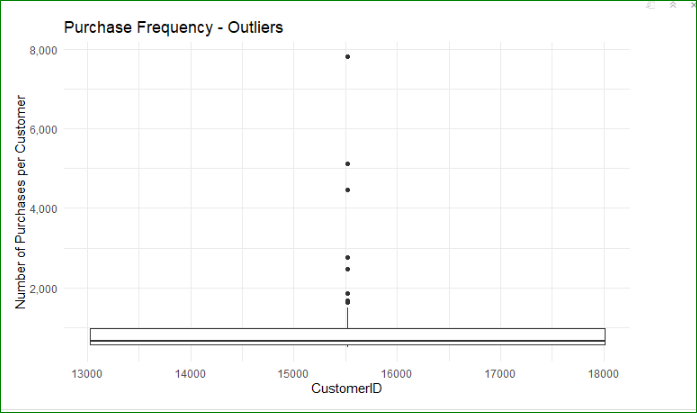
The above results show how often customers purchased from the online retailer. Customers make an average of 90 transactions a year, with 75% of users making fewer than 100 purchases a year. As shown in the above statistical summary, there are big difference between the 3rd quartile (99) and the maximum number of purchases (7,812).

Frequency Boxplot: Since there is a remarkable difference between the above frequency statistics of 3rd Quartile and Max value, the boxplot graph will be divided in two.

**Frequency boxplot graph showing customers in the third quartile**

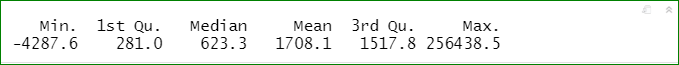


**Frequency boxplot graph showing customers in the fourth quartile**



Monetary Value: Finally, the last calculation to build before creating the cluster segmentation model is Monetary Value. This refers to the total sum of revenue generated by a given user over the course of a year.

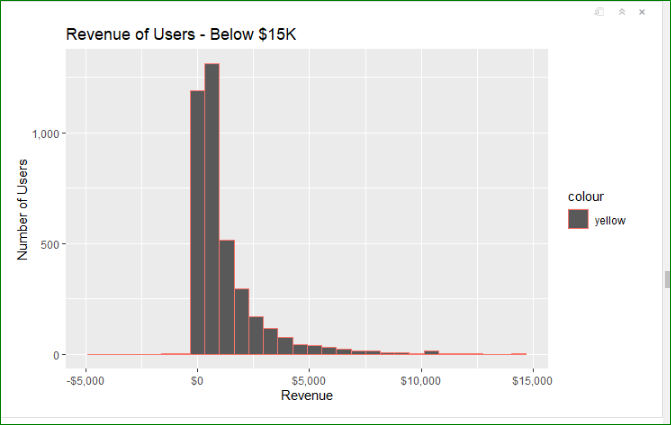
**Monetary Value Statistics**



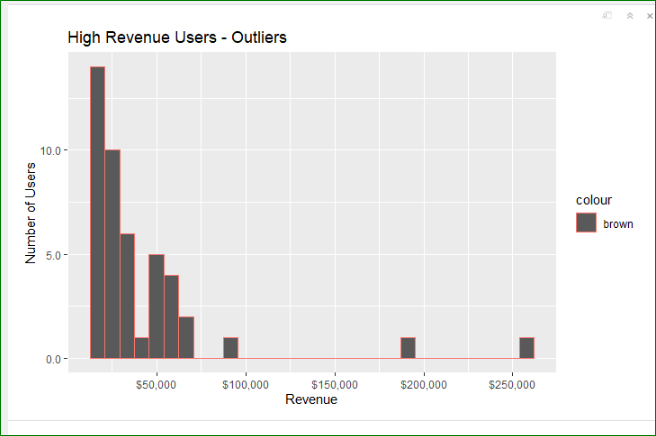
As the above results show, there are customers who have negative revenue that is, apparently there are negative numbers in the price column. This could suggest purchase returns, and this is an important assumption to keep in mind as we continue to work on the segmentation.

**Monetary Value Histograms**

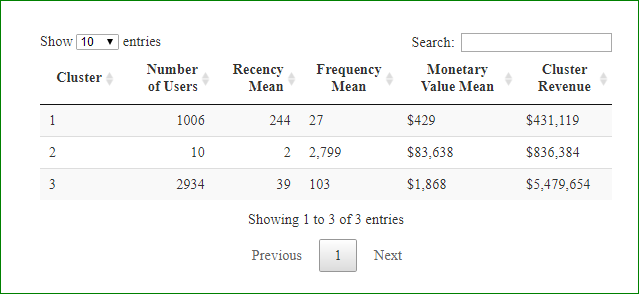
Plot two sets of users. The first histogram shows Customers below the 3rd Quartile



**Customers with revenue greater than 15k a year**



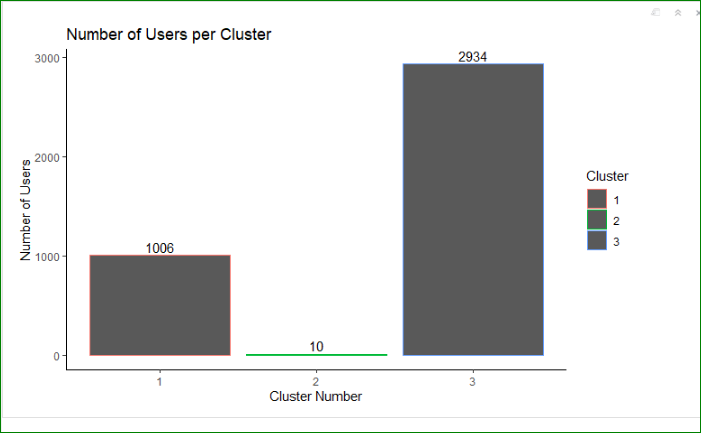
There are three different clusters based on customers’ behavior with the Online Retailer. We can examine the differences in these clusters in order to identify high value customers, medium value customers, and low value customers.



Below is a description of the above three customer segments:

|  |  |  |
| --- | --- | --- |
| Low Value Customers - Cluster 1 | Medium Value Customers - Cluster 3 | High Value Customers - Cluster 2 |
| 1,006 customers | 2,934 customers | 10 customers |
| Avg 244 days of inactivity | Avg 39 days of inactivity | Avg 2 days of inactivity |
| Avg number of Purchases, 27 a year | Avg Number of Purchases, 103 | Avg Number of Purchases, 2,799 |
| Avg Monetary Revenue of $429 a year | Avg Revenue of $1,868 a year | Avg Revenue of $83,638 a year |

Create histogram that displays number of clusters:



**Summary - KMeans Segmentation**: Customers of the business have been clearly identified into three groups using the KMeans clustering algorithm.

These clusters can help the business to better understand its customers in terms of revenue generated, frequency of purchases, and days of inactivity. This will help inform their marketing strategies, allowing them to maintain and/or improve the profitability of different type of customers.

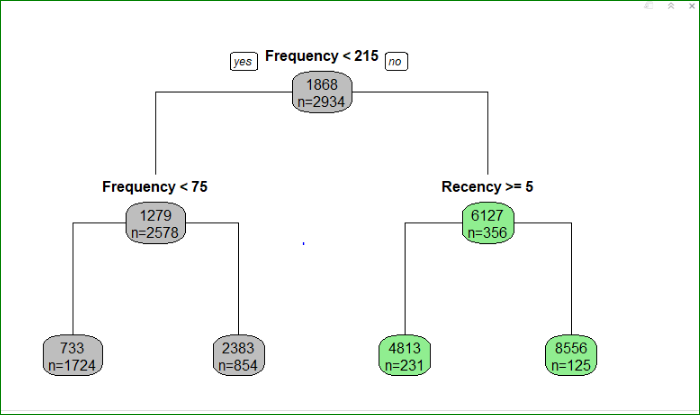
**Extra Segmentation - Hierarchical Clustering**

To improve this clustering analysis, it was decided to further segment of the largest cluster of customers in the first segmentation (Cluster 3). This further sub-segmentation was performed using hierarchical clustering to further understand the customer characteristics of this group.

How to read the graph below:

* Read the statement at the top of the model and move right for “no” or left for “yes”.
* The top number in the bubble represents the average monetary value of the group
* n = represents the number of variables that are part of the group

**Hierarchical Segmentation**



**Summary of hierarchical segmentation**

This sub-segmentation of Cluster 3 divided the segment into four smaller clusters.

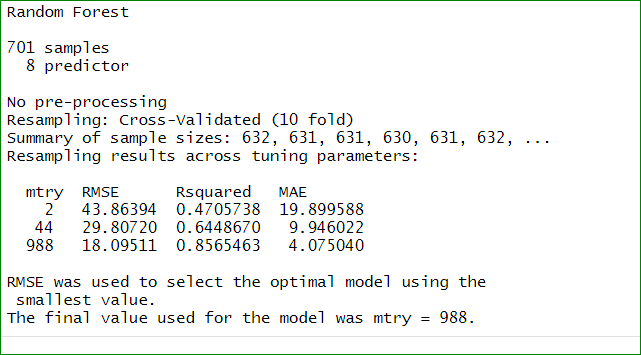
The above results show (from low value to high value customers):

* 1,724 customers who purchased less than 75 times, for an average monetary value of $733.
* 854 customers who purchased more than 75 times, for average monetary value of $2,383 (significantly higher than the previous group).
* 231 customers who purchased more than 215 times and have not purchased over the last five or more days. Average Monetary Value of $4,813.
* 125 customers that purchased more than 215 times and had not made any purchases with in the previous five days. Highest average monetary value of $8,556.
* This last sub-segment of 125 customers represents the most valuable customers within Cluster 3.

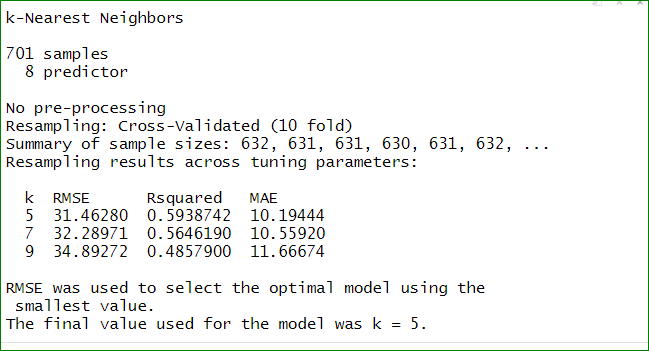
I think the executive and management team can take further strategic actions to increase the average monetary value of lower sub-segments within this cluster of customers.

**Create different models**

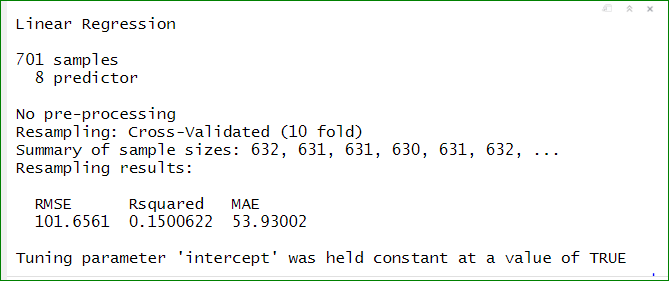
* I built and compared the following three models based on the results: a random forest model, a knn model, and a linear regression model.
* Random forest model



* KNN model



* Linear regression model



Comparing Models

* RMSE (Root Mean Squared Error)
* MAE (Mean Absolute Error)
* Rsquared (R-squared): represents the proportion of variation in the outcome explained by the predictor variables included in the model
* The lower the RMSE and the MAE, the better the model
* The higher the R-squared, the better the model

|  |  |  |  |
| --- | --- | --- | --- |
| Model | RMSE | Rsquared | MAE |
| random forest | 18.09511 | 0.8565463 | 4.075040 |
| K-Nearest Neighbors | 31.46280 | 0.5938742 | 10.19444 |
| Linear regression | 101.6561 | 0.1500622 | 53.93002 |

* As the above table shows, random forest is the best of the three models because it has the lowest RMSE and MAE and the highest R-squared value.
* Random forest can be used for classification or regression. I preferred to use regression

**Conclusion**

In general, this project allowed me to gain experience on how to define and use problem statements in order to create a clear and concise description of the issue that needs to be addressed by doing this project. I also gained experience in using project milestones in order to break down the timeline of a project. Moreover, I was able to practice and utilize data visualization and machine learning technique that I have learned so far in the Data Science course programs. In addition to that, I have learned to present project updates every week. In my life outside of school, I mostly work as a developer, and I communicate only with my immediate supervise. So, this has been a great experience for me. Overall, this is my first time working on this kind of practical project. It was very interesting, and I got the chance to practice with and use data visualization and machine learning techniques.

References

DataCamp.com. (2020). R Documentation. Corrplot. Retrieved from <https://www.rdocumentation.org/packages/arm/versions/1.10-1/topics/corrplot>

Statistics How To. (2020). What is Cohen’s Kappa Statistic? Retrieved from <https://www.statisticshowto.com/cohens-kappa-statistic/>

STHDA. (2018). Statistical tools for high-throughput data analysis. Retrieved from <http://www.sthda.com/english/articles/38-regression-model-validation/157-cross-validation-essentials-in-r/>

Yihui Xie (2017). The R package tinytex. Helper Functions to Manage TinyTeX, and Compile LaTeX Documents. Retrieved from <https://yihui.org/tinytex/r/>

Zhao, Y. (2011-2020). RDM. RDataMining.com: R and Data Mining. Retrieved from <http://www.rdatamining.com/>

Source:

Dr. Daqing Chen, Course Director: MSc Data Science. chend '@' lsbu.ac.uk, School of Engineering, London South Bank University, London SE1 0AA, UK.