25933337_CustomerSegmentationInRetailData_Report

SI 348 Assignment Report

Project 2: Customer Segmentation in Retail Data

Introduction

The data for Project 2 contains an eCommerce dataset for all sales between December 2010 and December 2011 for an online retail platform. The nature of this dataset makes it ideal for identifying key customer segments based on spending patterns, and the following report will perform an exploratory data analysis on this dataset in order to explore and identify how factors such as geographic location and time of year impact customer purchasing behavior for all sales at the online retail platform between December 2010 and December 2011. This will be achieved through the use of various tables, summary statistics, and visualisations derived from the dataset, with the objective of analyzing patterns in categories such as sales, returns, countries, quantities and invoice dates.

Data Cleaning and Preparation

A variety of steps were taken to ensure the dataset was meticulously clean and prepared for data analysis. The steps taken are outlined as follows.

Loading the Provided Dataset into R for Project 2

The provided dataset for Project 2 was first loaded into R Studio using a relative path to ensure the code is fully reproducible.

Performing the Initial Inspection

An initial inspection of the data was then performed. This entailed assessing the data in it's raw form in order to gauge the contents of it's columns, the formats of it's values, the names of it's columns, as well as the data types used for each column.

The table below presents a view of the first five rows of the data in it's raw form (Table 1).

Table 1: The Data in its Raw Form

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850	United Kingdom
536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850	United Kingdom
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850	United Kingdom

Inspecting Column Values

Once the the format of the data in it's raw form was gauged, the values of each column were inspected and filtered to detect the presence of any 'NA' values. This revealed the presence of 'NA' values in columns such as Description and CustomerID. Conducting this inspection revealed that the presence of 'NA' values in rows often indicated the presence of further missing or fixed values in the row. This is illustrated in Table 2 and Table 3 presented below.

Table 2: The First Five Rows with 'NA' Values in CustomerID

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536414	22139	NA	56	2010-12-01 11:52:00	0.00	NA	United Kingdom
536544	21773	DECORATIVE ROSE BATH-ROOM BOTTLE	1	2010-12-01 14:32:00	2.51	NA	United Kingdom

536544	21774	DECORATIV CATS BATH-ROOM BOTTLE	2	2010-12-01 14:32:00	2.51	NA	United Kingdom
536544	21786	POLKADOT RAIN HAT	4	2010-12-01 14:32:00	0.85	NA	United Kingdom
536544	21787	RAIN PONCHO RET-ROSPOT	2	2010-12-01 14:32:00	1.66	NA	United Kingdom

Table 3: The First 5 Rows where UnitPrice is 0

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536414	22139	NA	56	2010-12-01 11:52:00	0	NA	United Kingdom
536545	21134	NA	1	2010-12-01 14:32:00	0	NA	United Kingdom
536546	22145	NA	1	2010-12-01 14:33:00	0	NA	United Kingdom
536547	37509	NA	1	2010-12-01 14:33:00	0	NA	United Kingdom
536549	85226A	NA	1	2010-12-01 14:34:00	0	NA	United Kingdom

Additional filters were then added to filter for any obvious fixed values that arose which indicated untidy data, such as the presence of question marks in the Description column. This is illustrated in Table 4 below.

Table 4: The First 5 Rows where Description Contains '?'

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
537032	21275	?	-30	2010-12-03 16:50:00	0	NA	United Kingdom
538090	20956	?	-723	2010-12-09 14:48:00	0	NA	United Kingdom
539494	21479	?	752	2010-12-20 10:36:00	0	NA	United Kingdom
540100	22837	?	-106	2011-01-04 16:53:00	0	NA	United Kingdom
540558	21258	?	-29	2011-01-10 10:04:00	0	NA	United Kingdom

Addressing Any Missing Values

While inspecting the column values revealed that much of the dataset was either incomplete or 'untidy', these rows still contained data which would be useful for the exploratory data analysis process, and should therefore not simply be excluded from the analysis. Tidying this data required an iterative approach- the more data that was processed, cleaned, and tidied, the more data was discovered to be untidy.

Using the initial 'NA' values, fixed values, and missing values gleaned from the initial inspection of the columns as a starting point, regular expressions were then used in order to detect, categorise, and replace untidy data into values that were more meaningful and suitable for the data analysis process. As rows from the dataset were cleaned and the updates were confirmed by filtering to view the cleaned data in context, more categories of untidy data were found and addressed.

This resulted in the use of a total of eighty-two regular expressions in order to detect, categorise, and replace incomplete, flawed or untidy data in the dataset. This meant that fixed values in the dataset could now be addressed. Due to the nature of these fixed values and the eCommerce dataset, it was assumed that they were implemented as a data entry convenience, and could therefore be best dealt with using the "last observation carried forward" treatment.

In order to make use of this treatment, fixed values which had no other meaningful implications were converted to 'NA' values. Examples of these values included '?', '??', '???' and 0. Once these values were converted to 'NA' and then filtered by InvoiceNo, the "last observation carried forward" treatment could be applied and the data successfully cleaned. This was confirmed by filtering through the entire cleaned dataset for 'NA' values, confirming none were returned and the treatment was successful.

Categorizing the Data

Now that the data was successfully tidied, the rest of the data preparation process could be completed. This included categorizing the data and completing variable calculations.

Categorizing the data involved making use of the descriptions and categories that some of the regular expressions had assigned the rows, as well as some of the innate categorical values of the dataset. The default category for a row was set to 'Sale', due to the nature of the eCommerce dataset, while twenty one other categories could be assigned to each row depending on its value for columns such as StockCode, Description and InvoiceNo. An example row from each Category is shown in Table 5 below.

Table 5: An Example Row From Each Category

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Category
536941	22734	AMAZON	20	2010-12-03 12:08:00	4.25	14680	United Kingdom	AMAZON
546407	22719	BARCODE PROBLEM	-178	2011-03-11 16:24:00	2.95	17351	United Kingdom	BARCODE PROBLEM
C536383	35004C	ERROR	-1	2010-12-01 09:49:00	4.65	15311	United Kingdom	CANCELLED
536592	90051	DAMAGES	1	2010-12-01 17:06:00	7.22	14606	United Kingdom	DAMAGES
540564	22617	DISCARDED	-2600	2011-01-10 10:36:00	10.95	15100	United Kingdom	DISCARDED
C536379	D	DISCOUNT	-1	2010-12-01 09:41:00	27.50	14527	United Kingdom	DISCOUNT
536945	37464	DISPLAY	1	2010-12-03 12:24:00	1.25	14083	United Kingdom	DISPLAY

536544	DOT	DOTCOM	1	2010-12-01 14:32:00	569.77	16456	United Kingdom	DOTCOM
561249	DCGS0073	EBAY	-4	2011-07-26 11:51:00	0.62	16592	United Kingdom	EBAY
536365	22752	ERROR	2	2010-12-01 08:26:00	7.65	17850	United Kingdom	ERROR
539611	85135B	FOUND	53	2010-12-20 14:33:00	1.25	14606	United Kingdom	FOUND
540010	22501	INCORRECT	-100	2011-01-04 11:13:00	1.25	17315	United Kingdom	INCORRECT
547363	22459	LOST	-232	2011-03-22 12:33:00	1.69	13694	United Kingdom	LOST
536569	M	MANUAL	1	2010-12-01 15:35:00	1.25	16274	United Kingdom	MANUAL
550800	22458	MISSING	-65	2011-04-20 14:48:00	0.39	17115	United Kingdom	MISSING
536370	POST	POSTAGE	3	2010-12-01 08:45:00	18.00	12583	France	POSTAGE
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850	United Kingdom	SALE
549684	S	SAMPLES	1	2011-04-11 13:24:00	30.00	14273	United Kingdom	SAMPLE
537425	84968F	STOCK CHECK	-20	2010-12-06 15:35:00	1.25	15602	United Kingdom	STOCK CHECK
549527	21620	UNKNOWN	-1479	2011-04-08 16:03:00	0.95	13630	United Kingdom	UNKNOWN
558379	22618	UNSALEABL	-1681	2011-06-28 16:34:00	0.55	13418	United Kingdom	UNSALEABLE
553394	15058A	WATER DAMAGE	-30	2011-05-16 16:47:00	4.25	16131	United Kingdom	WATER DAMAGE

Variable Calculations

As the research question outlined in the introduction of this report refers to customer purchasing behavior, it was deemed essential to conduct a variable calculation and create a column in the dataset for a TotalPrice value for each row. This variable is the result of multiplying the Quantity value and the UnitPrice value for each row, so that the total amount spent per row could be taken into account when analyzing customer purchasing behavior. The dataset was then deemed fully tidied and sufficiently prepared for the exploratory data analysis process.

The updated columns for the dataset can be viewed in Table 6 below.

Table 6: The First Five Rows From the Updated Dataset Including the Column 'TotalPrice

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	TotalPrice	CustomerID	Country	Category
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	15.30	17850	United Kingdom	SALE

536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	20.34	17850	United Kingdom	SALE
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	22.00	17850	United Kingdom	SALE
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	20.34	17850	United Kingdom	SALE
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	20.34	17850	United Kingdom	SALE

Exploratory Data Analysis

Summary Statistics

The first step in the exploratory data analysis process involved creating a variety of summary statistics for key variables from the dataset to explore the relationships both within and between variables. This was an iterative process, as certain summary statistics revealed interesting trends, while some were deemed irrelevant to the research question. Any notable patterns were then visualized using either scatterplots or bar graphs. The key variables relevant to the research question were identified as Quantity, TotalPrice, Country, and InvoiceDate.

Variation in Quantity by InvoiceDate

The first summary statistic which was considered summarized the variation in Quantity by InvoiceDate, revealing the total quantity of items purchased per day recorded in the dataset. This summary statistic relates directly to the research question, as it reveals patterns in how time of year impacts customer purchasing behavior through the quantity of items purchased. The table containing the first five rows for this summary statistic can be seen below (Table 7).

While the table presented below is merely a glimpse into the data, it is unsurprising that the invoice data on which the highest quantity of customer purchases took place falls in December-likely due to customers purchasing items ahead of Christmas Day on December 25th. While the months of the following four records are more novel, the relationship they represent can be better visualized using a scatterplot.

Table 7: The First Five Rows From the Quantity Variation by InvoiceDate Summary Statistic Dataframe, Arranged in Descending Order

InvoiceDate	TotalQuantity
2011-12-09	91560
2011-01-18	81164
2011-09-20	37292
2011-11-14	36870
2011-10-05	35868

Variation in Quantity by Country

The second summary statistic which was considered summarized the variation in Quantity by Country, revealing the total quantity of items purchased from December 2010 to December 2011 by each country listed in the dataset. This summary statistic also directly relates to the research question, as it reveals patterns in how geographic location impacts customer purchasing behavior through the quantity of items purchased. The table containing the first five rows of this summary statistic can be seen below (Table 8).

A glance at Table 8 reveals many relationships between Quantity variation and Country. One can infer from the data presented in the table below that the eCommerce platform is based in the UK, given majority of their sales take place there. Additionally, EIRE holding the spot of third highest quantity purchased is unsurprising, given the close relationship, both geographically and culturally, between the UK and EIRE. Finally, all top five results list countries listed in Europe. While customers from outside Europe do purchase from the eCommerce platform, obstacles such as shipping costs are likely what cause the highest quantity purchasing countries to be European.

Table 8: The First Five Rows From the Quantity Variation by Country Summary Statistic Dataframe, Arranged in Descending Order

Country	TotalQuantity
United Kingdom	3765746
Netherlands	144641
EIRE	107646
Germany	84018
France	82064

Variation in Quantity of Cancelled Transactions by InvoiceDate

The third summary statistic which was considered summarized the variation in the Quantity of cancelled purchases, known as returns, by InvoiceDate. This summary statistic reveals the total Quantity of Sales which were categorized as 'CANCELLED' per recorded InvoiceDate from December 2010 to December 2011. While the use of the 'CANCELLED' category instead of the 'SALE' category in this summary statistic is unique, the category had the second

highest amount of occurrences in the dataset behind 'SALE', and revealed interesting patterns in customer purchasing behavior. Thus, this summary statistic also directly relates to the research question while providing insight beyond only transactions categorized as 'SALE'. The table containing the first five rows for this summary statistic can be seen below (Table 9).

A glimpse at the data below reveals the linear relationship between sales and returns when viewed in conjunction with Table 7, as the first two rows for both Table 7 and Table 9 indicate the highest quantity of customer purchases occurred on the 9th of December 2011 followed by the 18th of January 2011. However, the following three records displayed in Table 9 below do not correspond with the dates of the similarly ranked rows in Table 7, revealing an interesting relationship between the two summary statistics.

Table 9: The First Five Rows From the Cancelled Quantity Variation by InvoiceDate Summary Statistic Dataframe, Arranged in Descending Order

InvoiceDate	Category	TotalQuantity	year
2011-12-09	CANCELLED	81029	2011
2011-01-18	CANCELLED	74244	2011
2010-12-02	CANCELLED	10287	2010
2011-04-18	CANCELLED	9247	2011
2011-10-11	CANCELLED	7825	2011

Variation in TotalPrice by InvoiceDate

The fourth summary statistic which was considered summarized the variation in TotalPrice by InvoiceDate, revealing the total value purchased by customers per day recorded in the dataset. This summary statistic directly relates to the research question in the way that it reveals patterns in how time of year impacts customer purchasing behavior through monetary value rather than quantity of goods purchased. The table containing the first five rows of this summary statistic can be seen below (Table 10).

Once again, the 9th of December 2011 takes the top ranking, corresponding with Tables 7 and 9 as well. However the following records do not correspond similarly, with the highest customer purchases by monetary value occurring on the 22nd of December 2010, three days before Christmas. This is presumably last minute Christmas purchasing by consumers. The following two records do correspond with Tables 7 and 9, with 22nd December 2010 and 18th of January 2011 appearing again- this is to be expected as there is also a direct relationship between the quantity of purchases made and the total monetary value of purchases made, however, this makes a lack of correspondence more intriguing.

Table 10: The First Five Rows From the TotalPrice Variation by InvoiceDate Summary Statistic Dataframe, Arranged in Descending Order

InvoiceDate	SumPrice
2011-12-09	192390.23
2010-12-22	122051.41
2011-01-18	90417.50
2011-09-20	89375.52
2011-11-14	85808.64

Variation in TotalPrice by Country

The fifth summary statistic which was considered summarized the variation in TotalPrice by Country, revealing the total value purchased by customers per country recorded in the dataset. This summary statistic directly relates to the research question as it reveals patterns in how geographic location impacts customer purchasing behavior through the total value of items purchased. The table containing the first five rows of this summary statistic can be seen below (Table 11).

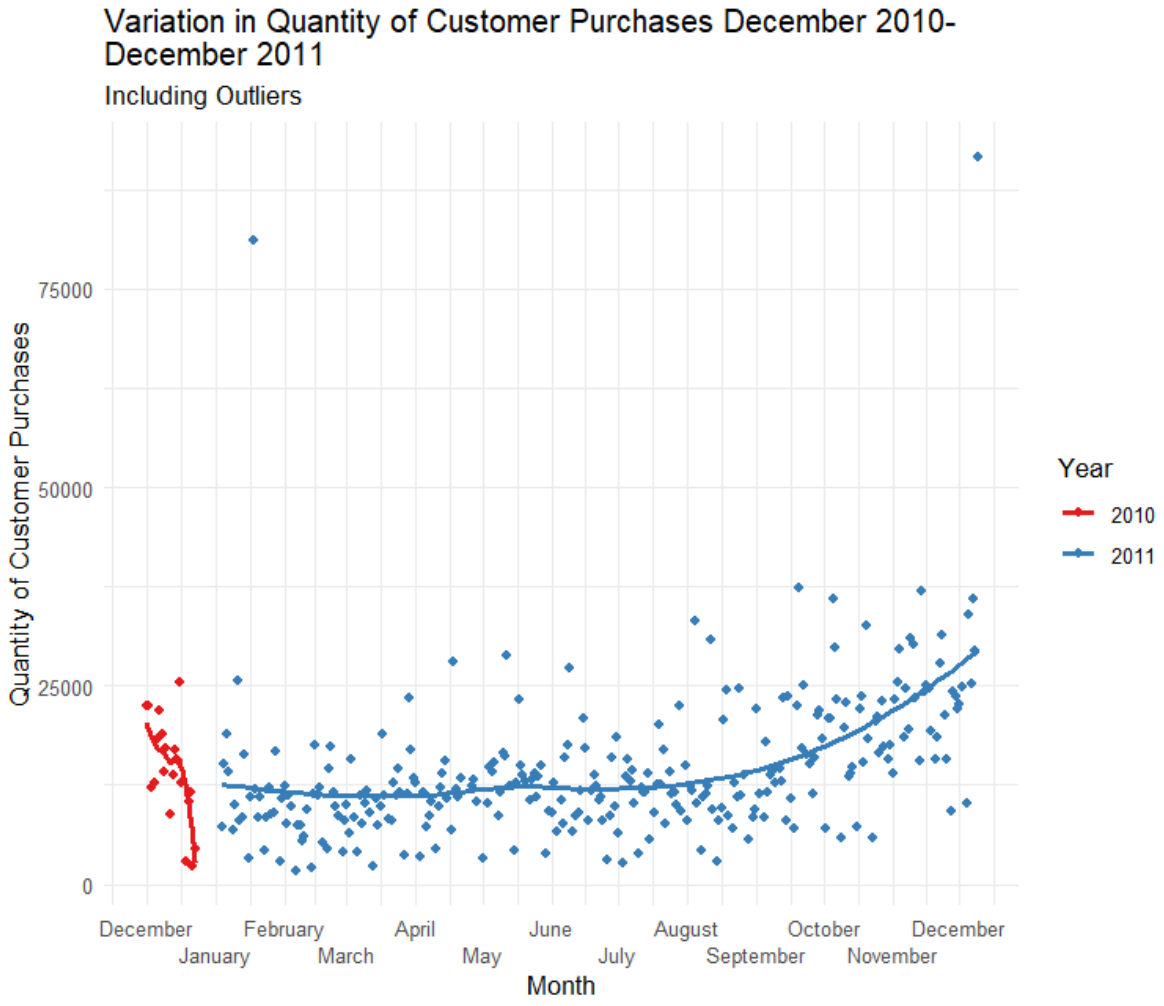
The results presented in the glimpse of the data frame aligns with the results presented in Table 8, as similar causes effect variation in Quantity as TotalPrice.

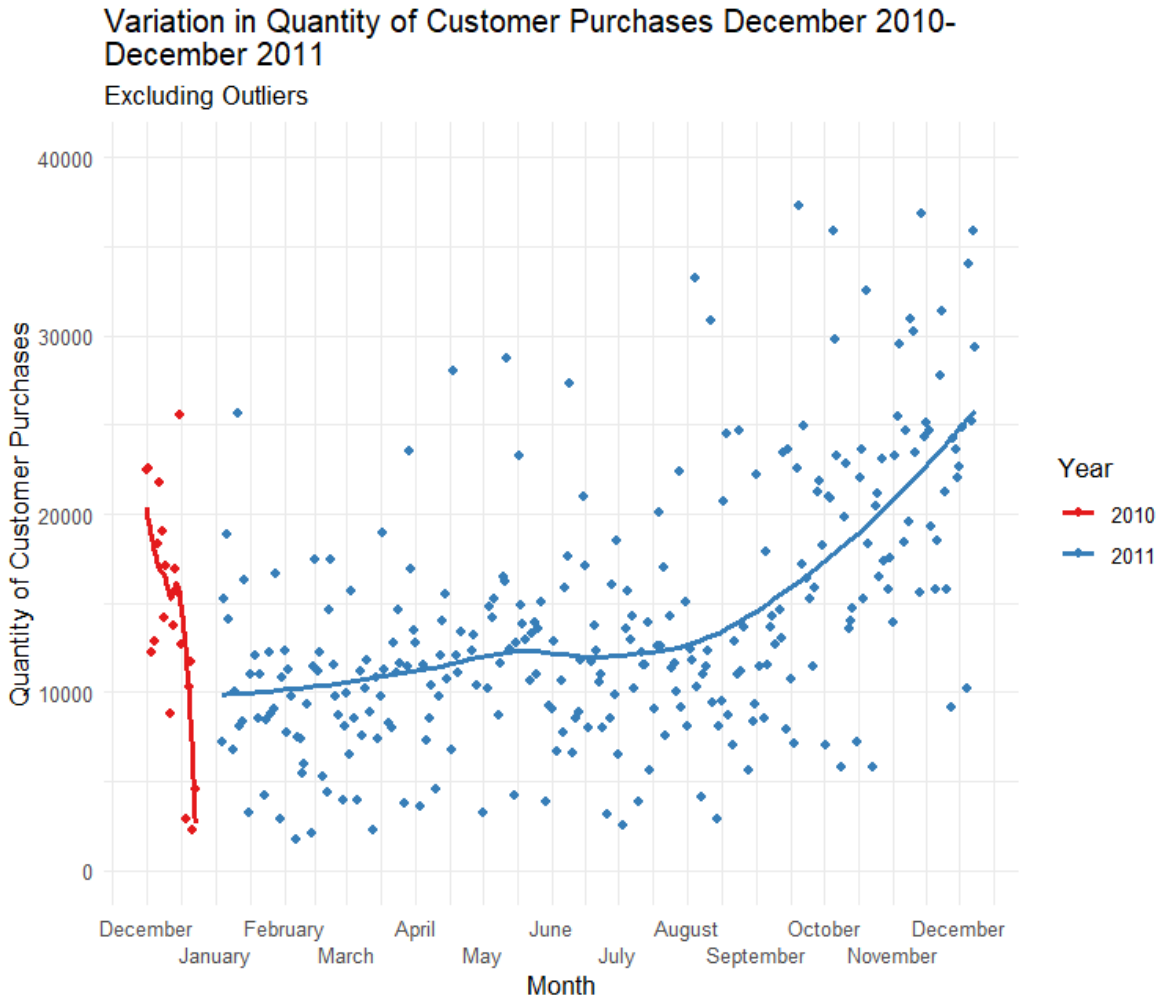
Table 11: The First Five Rows From the TotalPrice Variation by Country Summary Statistic Dataframe, Arranged in Descending Order

Country	SumPrice
United Kingdom	7417281.2
Netherlands	209665.2
EIRE	207182.3
Germany	146023.9
France	144725.4

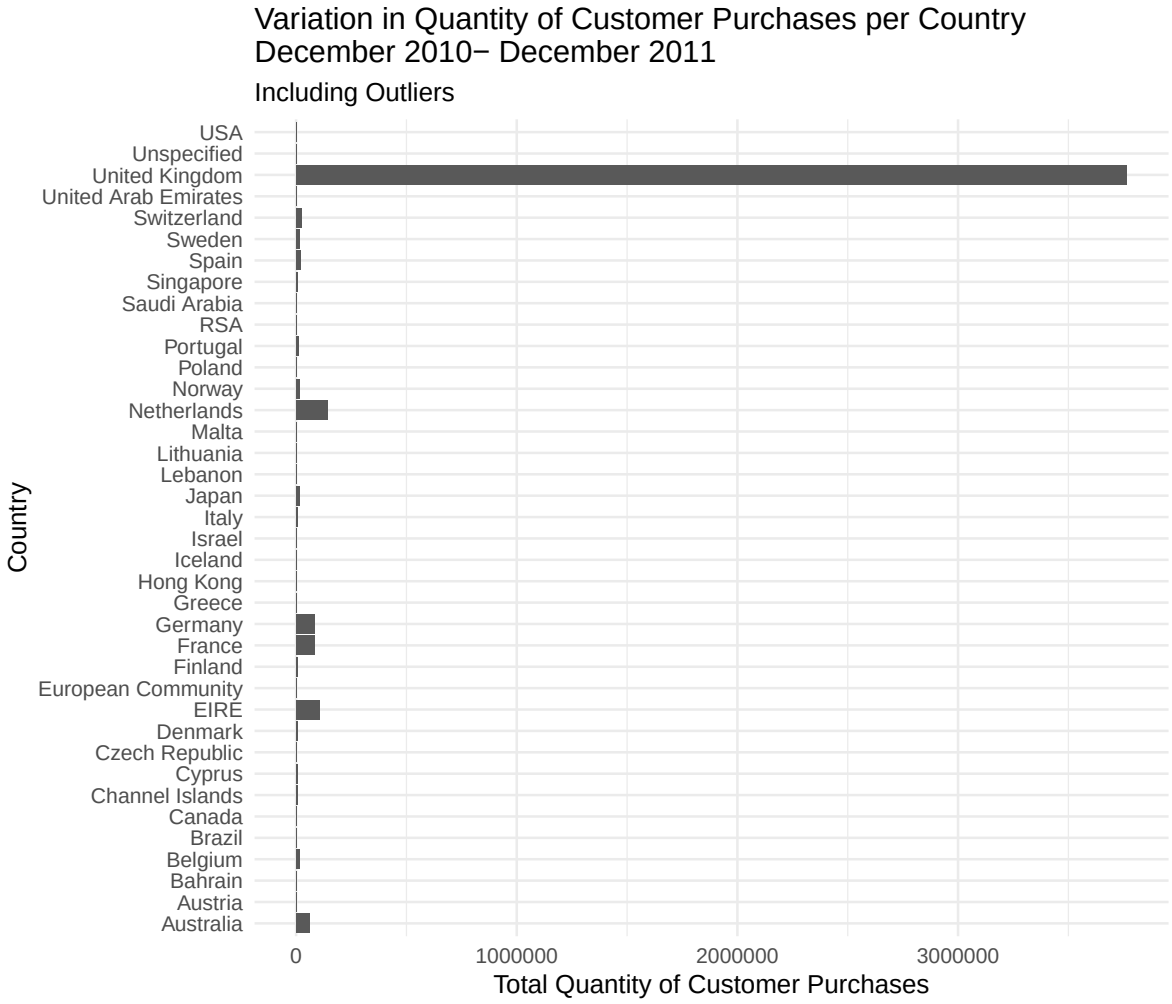
Visualizations

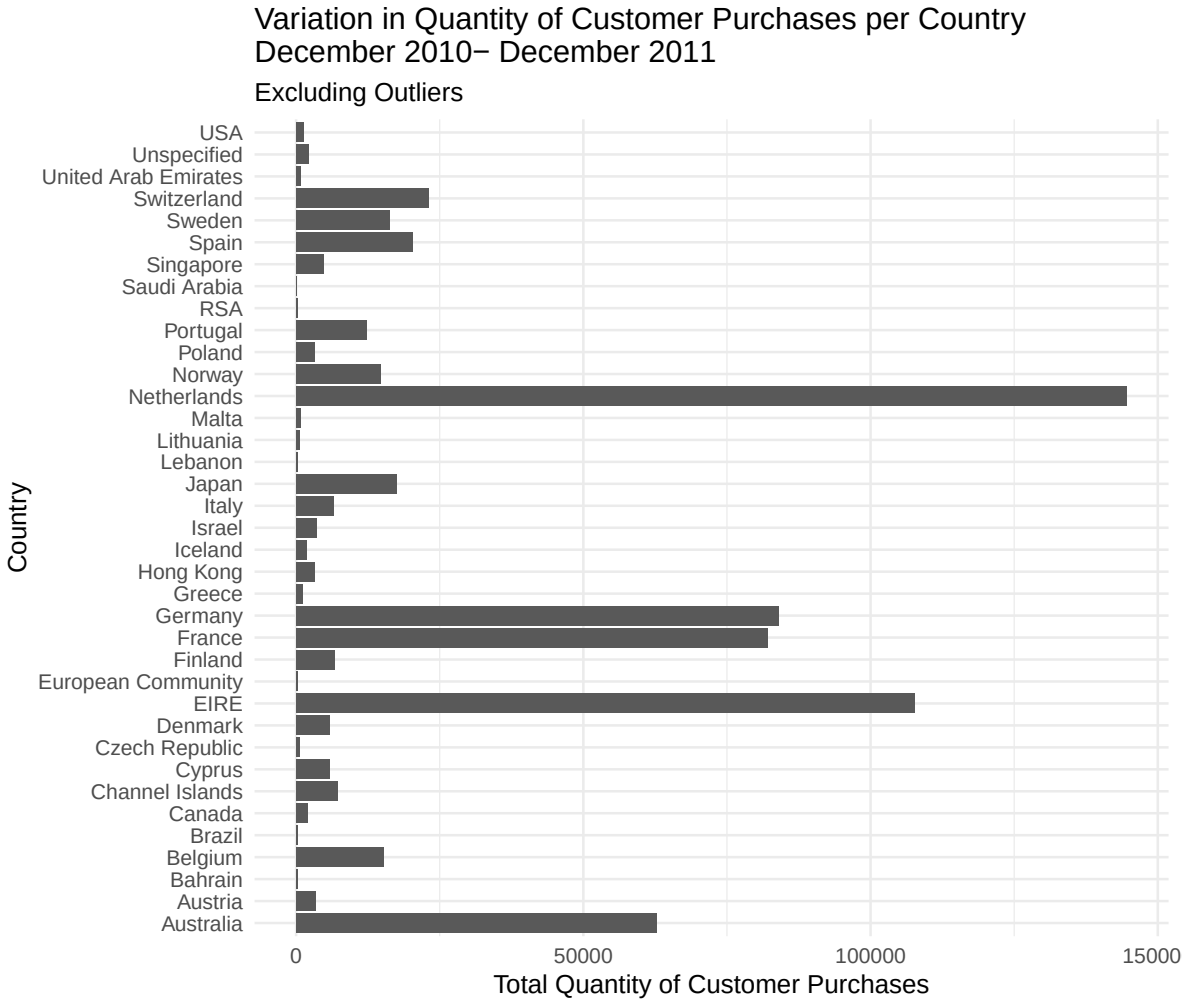
Variation in Quantity by InvoiceDate Scatterplots


Variation in Quantity of Customer Purchases December 2010-December 2011
Including Outliers
Quantity of Customer Purchases
75000
50000
25000
0
December
January
February
March
April
May
June
July
August
September
October
November
December
Month
Year
2010
2011

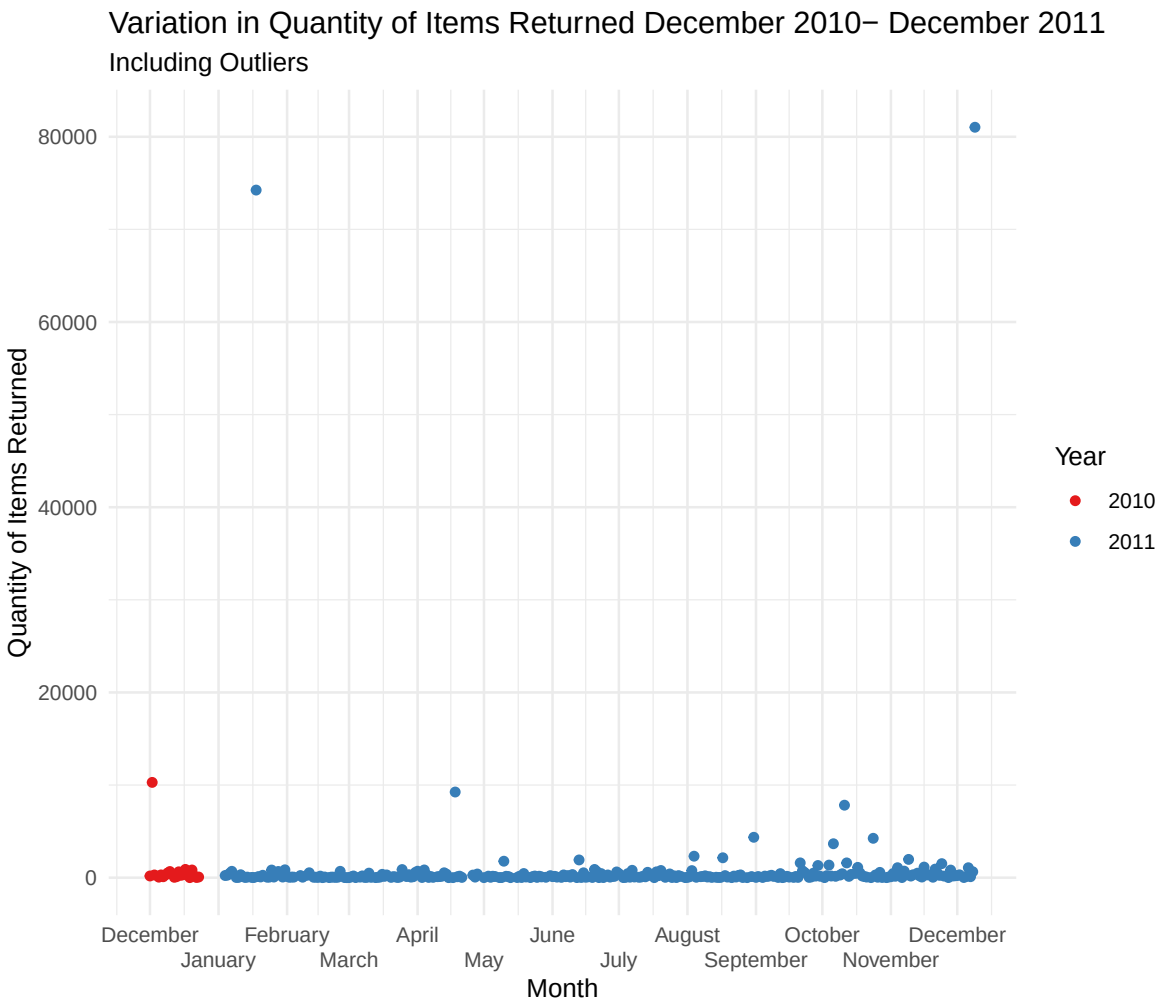


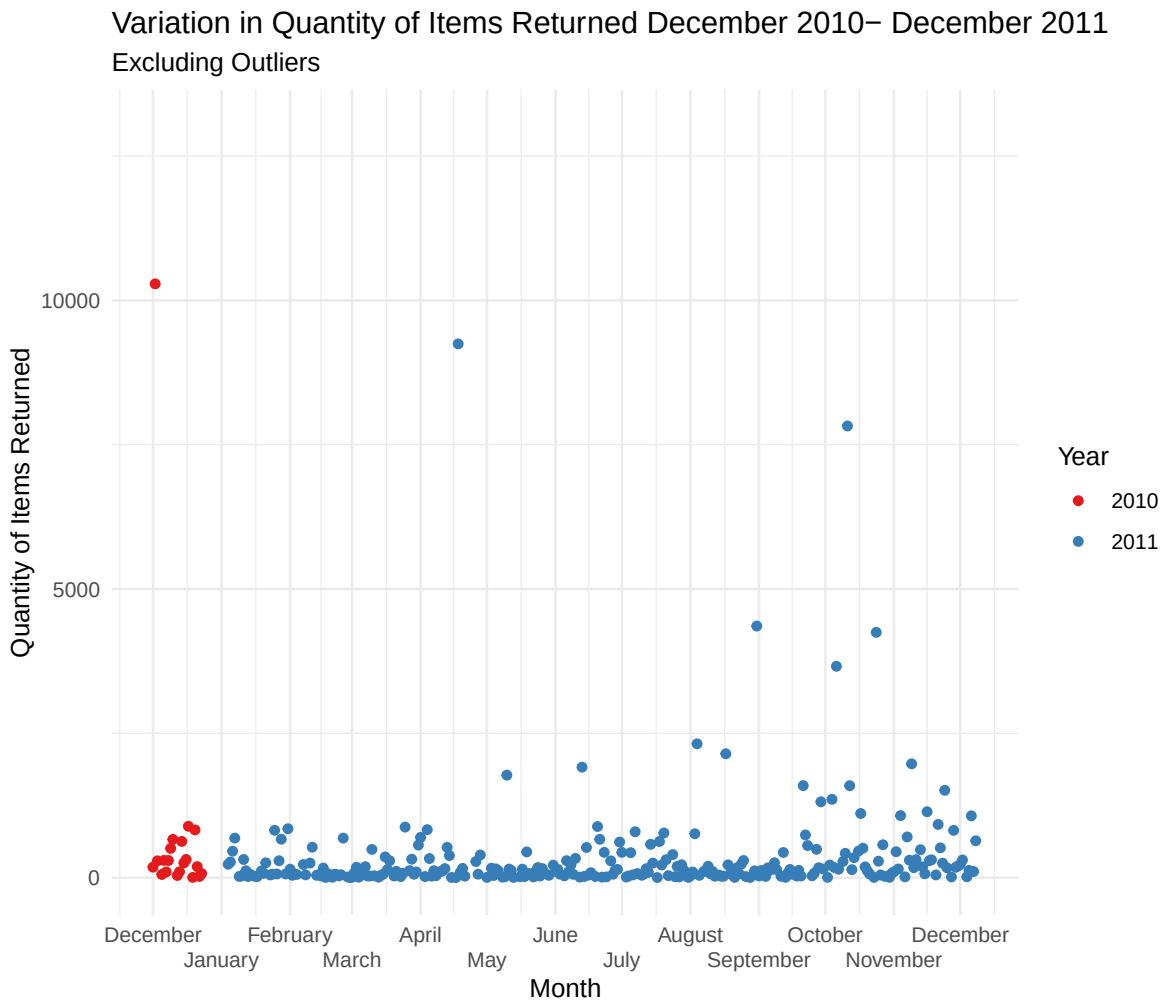
Variation in Quantity by Country Bar Graphs


Variation in Quantity of Customer Purchases per Country
December 2010− December 2011
Including Outliers
Country
USA
Unspecified
United Kingdom
United Arab Emirates
Switzerland
Sweden
Spain
Singapore
Saudi Arabia
RSA
Portugal
Poland
Norway
Netherlands
Malta
Lithuania
Lebanon
Japan
Italy
Israel
Iceland
Hong Kong
Greece
Germany
France
Finland
European Community
EIRE
Denmark
Czech Republic
Cyprus
Channel Islands
Canada
Brazil
Belgium
Bahrain
Austria
Australia
0
1000000
2000000
3000000
Total Quantity of Customer Purchases

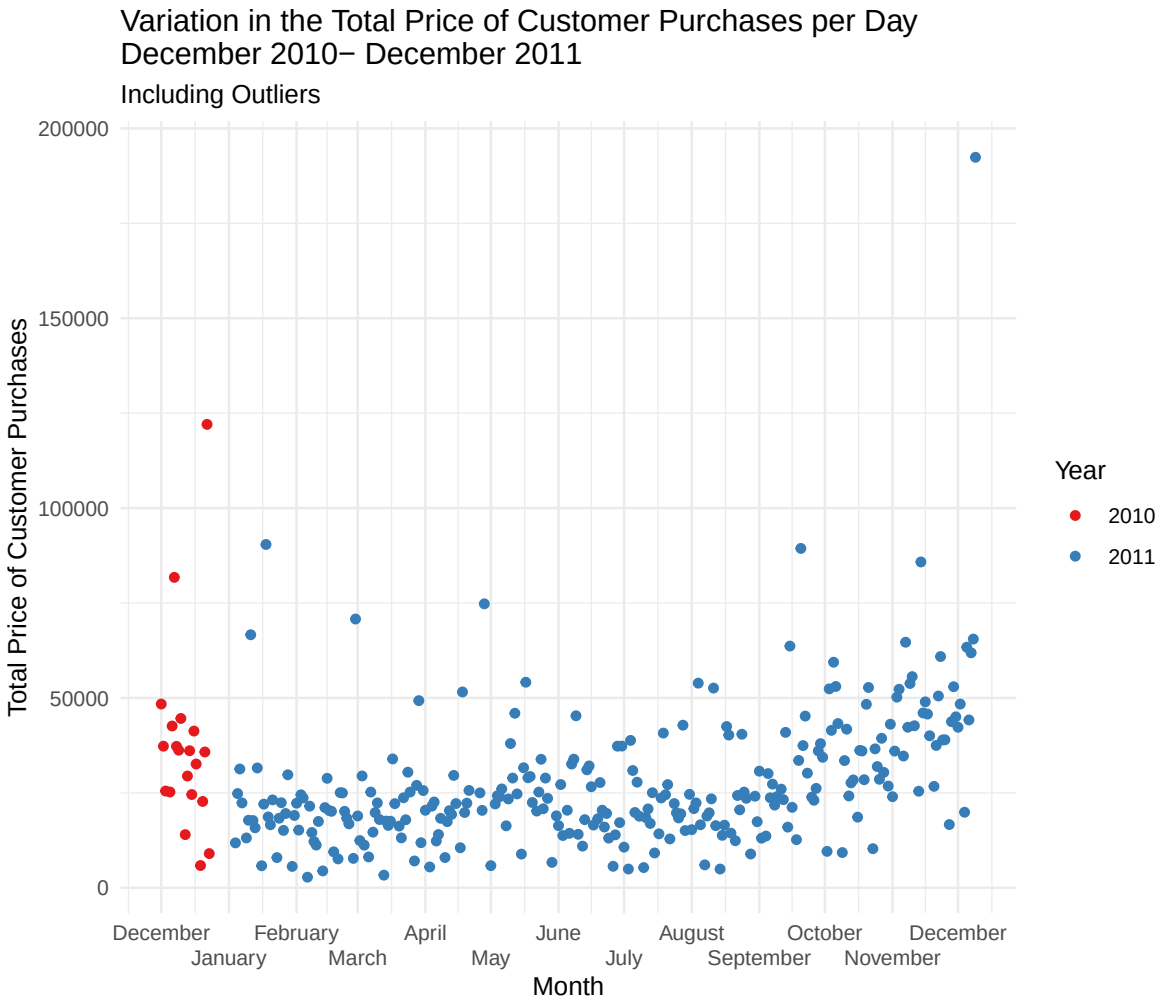


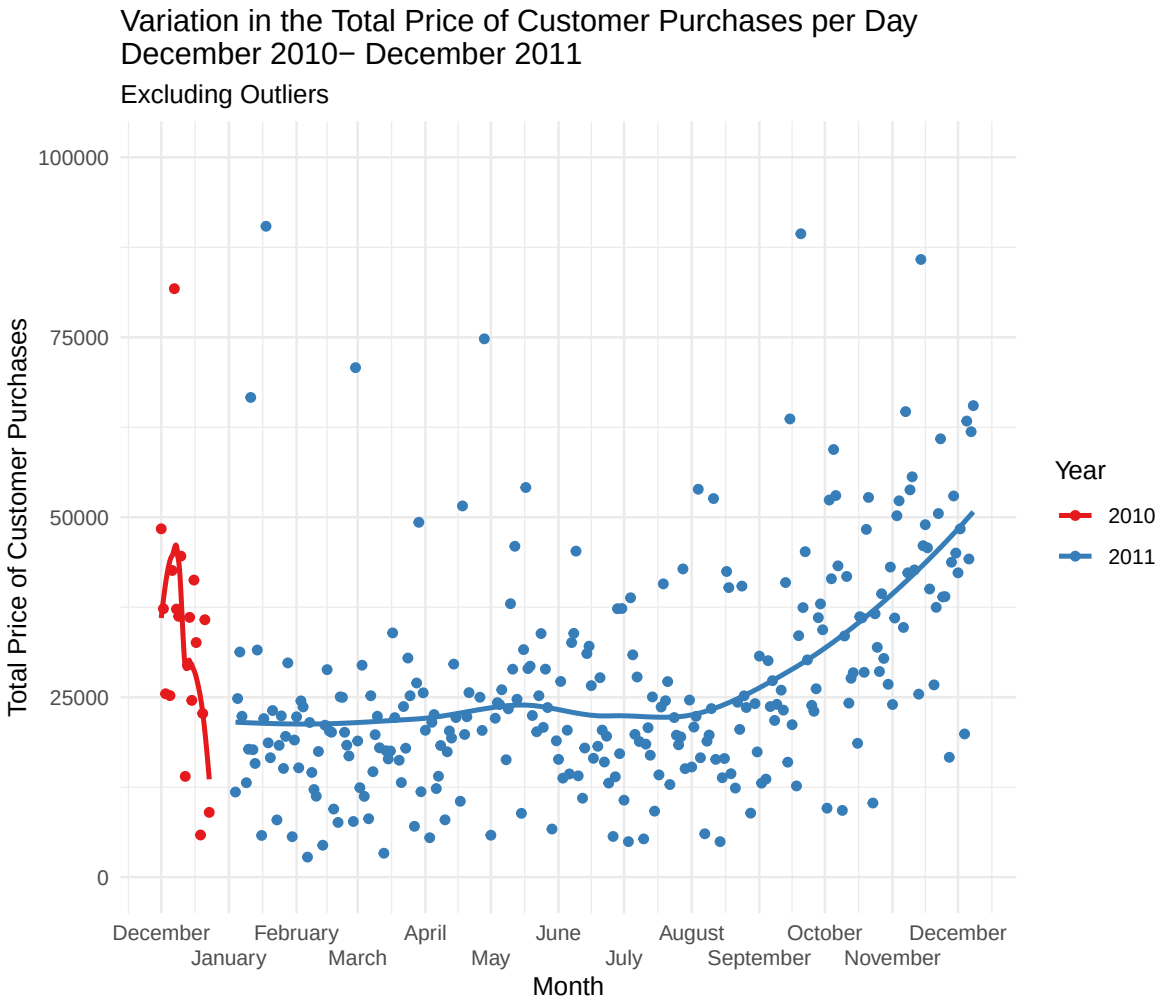
Variation in Quantity of Cancelled Transactions by InvoiceDate Scatterplots


Variation in Quantity of Items Returned December 2010− December 2011
Including Outliers
Quantity of Items Returned
80000
60000
40000
20000
0
December
January
February
March
April
May
June
July
August
September
October
November
December
Month
Year
2010
2011

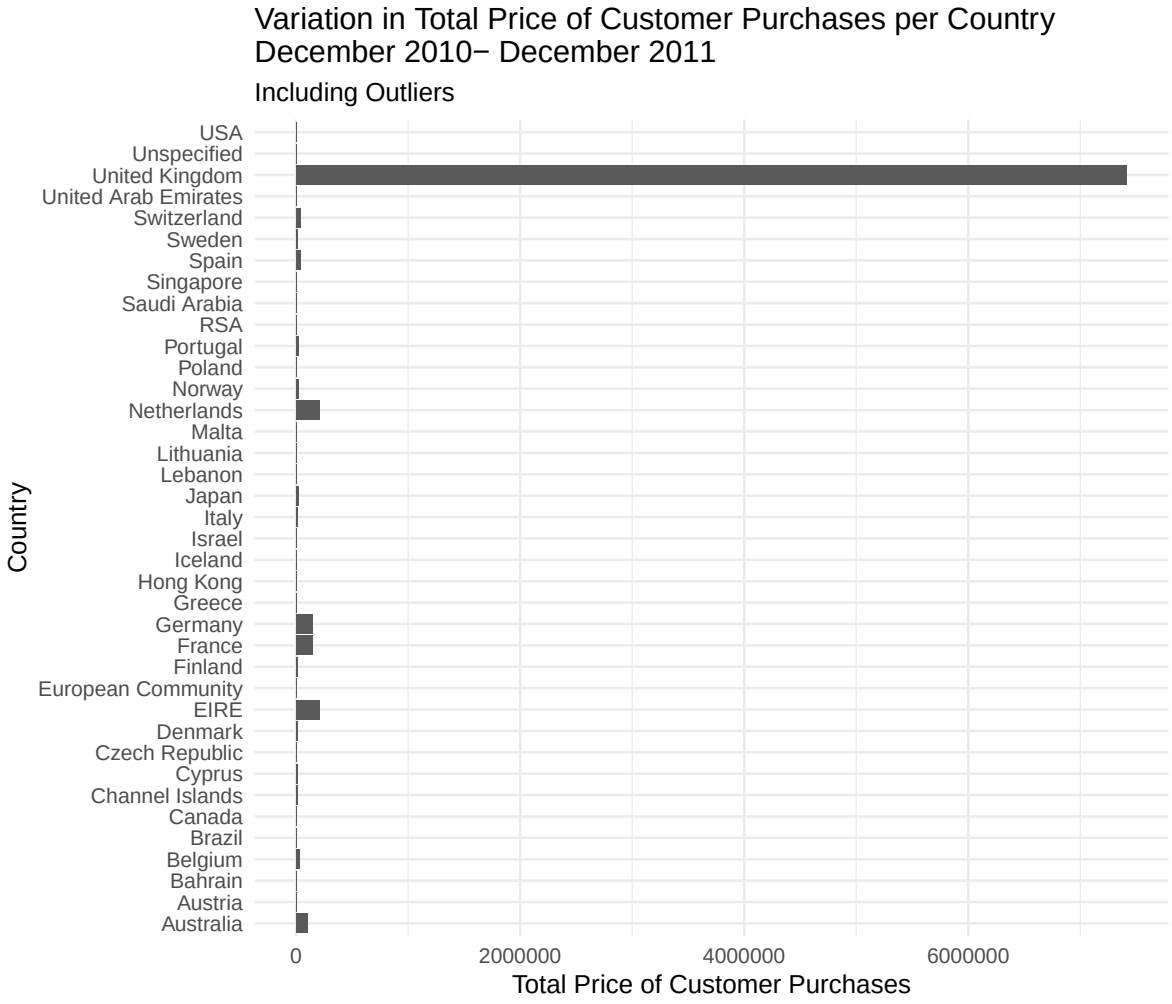


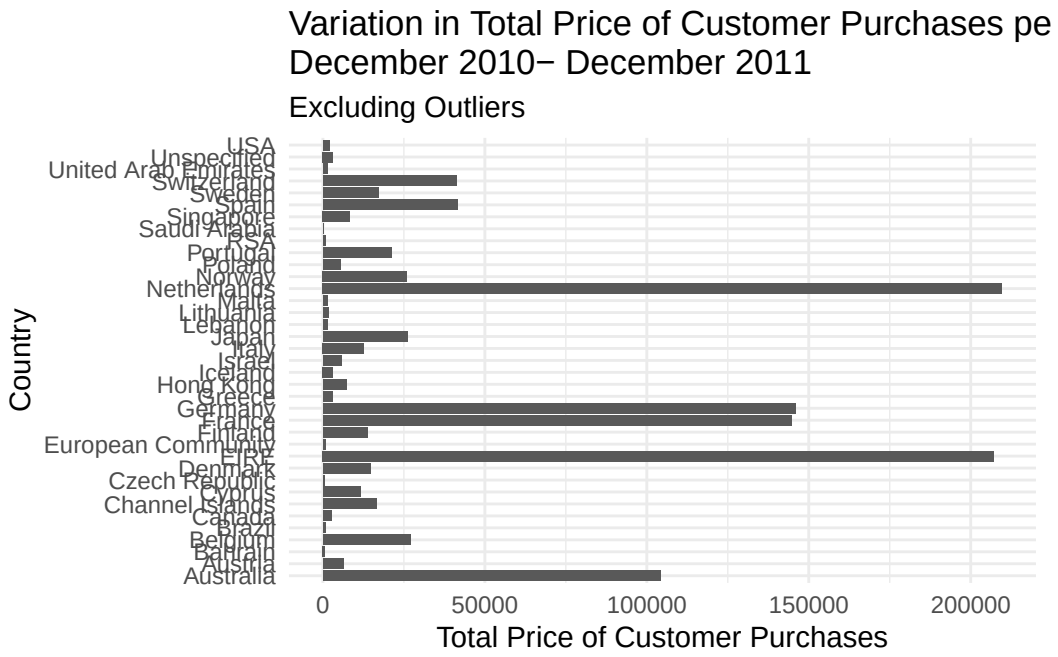
Variation in TotalPrice by InvoiceDate Scatterplots


Variation in the Total Price of Customer Purchases per Day
December 2010− December 2011
Including Outliers
Total Price of Customer Purchases
200000
150000
100000
50000
0
December
January
February
March
April
May
June
July
August
September
October
November
December
Month
Year
2010
2011



Variation in TotalPrice by Country Bar Graphs

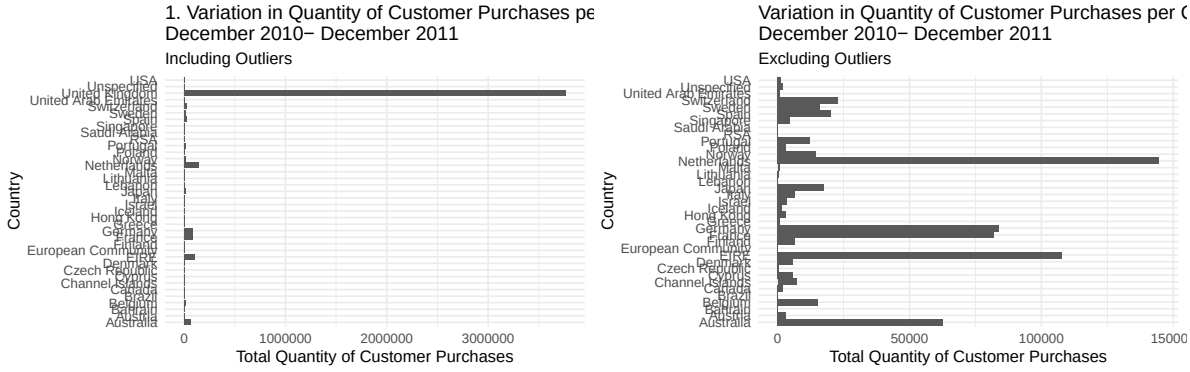

Variation in Total Price of Customer Purchases per Country
December 2010− December 2011
Including Outliers
Country
USA
Unspecified
United Kingdom
United Arab Emirates
Switzerland
Sweden
Spain
Singapore
Saudi Arabia
RSA
Portugal
Poland
Norway
Netherlands
Malta
Lithuania
Lebanon
Japan
Italy
Israel
Iceland
Hong Kong
Greece
Germany
France
Finland
European Community
EIRE
Denmark
Czech Republic
Cyprus
Channel Islands
Canada
Brazil
Belgium
Bahrain
Austria
Australia
0
2000000
4000000
6000000
Total Price of Customer Purchases



Reporting

When considered in conjunction, the summary statistics presented and their associated visualisations clearly demonstrate the direct correlation between factors, such as geographic location and time of year, and their impact on customer purchasing behavior between December 2010 and December 2011 at an online retail platform. The findings associated with these correlations are presented below.

How Does Variation in Quantity Illustrate the impact of Geographic Location on Customer Purchasing Behavior at an Online Retail Platform?

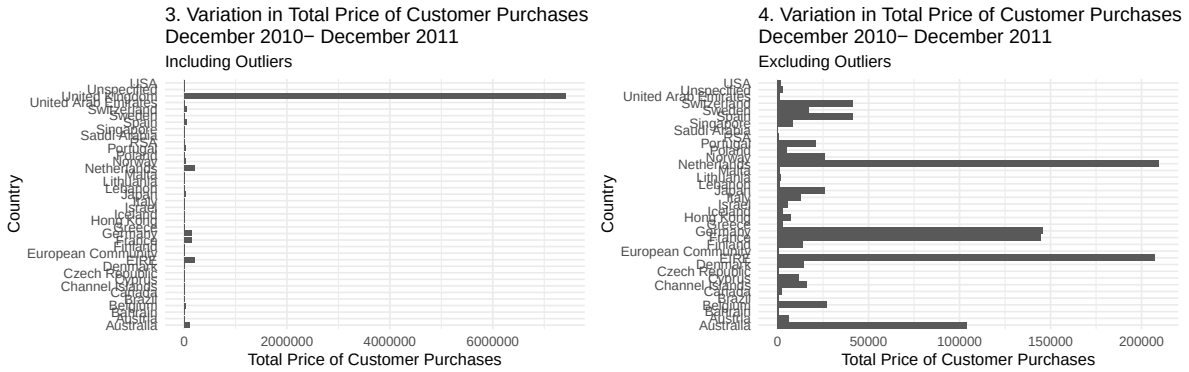


The above visualisations both refer to the total Quantity of customer purchases by Country. However, one includes outliers while the other excludes outliers. This is not because the outlier is unimportant to the data analysis process, but rather so that trends other than those which occur most often can be better observed. With that in mind, the visualisations above illustrate that the eCommerce platform is likely based in the United Kingdom, due to the overwhelming Quantity of purchases originating from the country.

This assumption remains consistent when the United Kingdom is treated as an outlier, as is the case in the second visualization, as the closest contenders in Quantity are overwhelmingly European. It can be inferred that this pattern is likely due to the direct proximity between the countries, due to factors that affect customer purchasing behaviors such as shipping costs and duration. However, Australia seems to disrupt this pattern, ranking 6th in terms of total Quantity of customer purchases. There could be numerous explanations for this, such as Australia and the UK's close cultural ties, that account for this total Quantity despite the distance between the two countries.

The above visualisations clearly illustrate the impact of factors, such as geographic location, on customer purchasing behavior at an eCommerce platform based in the UK. This is evident by the prevalence of European countries having the highest total Quantities of customer purchases, indicting a correlation between a country's proximity to the base of an eCommerce platform and customer purchasing behavior.

How Does Variation in Total Price Illustrate the impact of Geographic Location on Customer Purchasing Behavior at an Online Retail Platform?

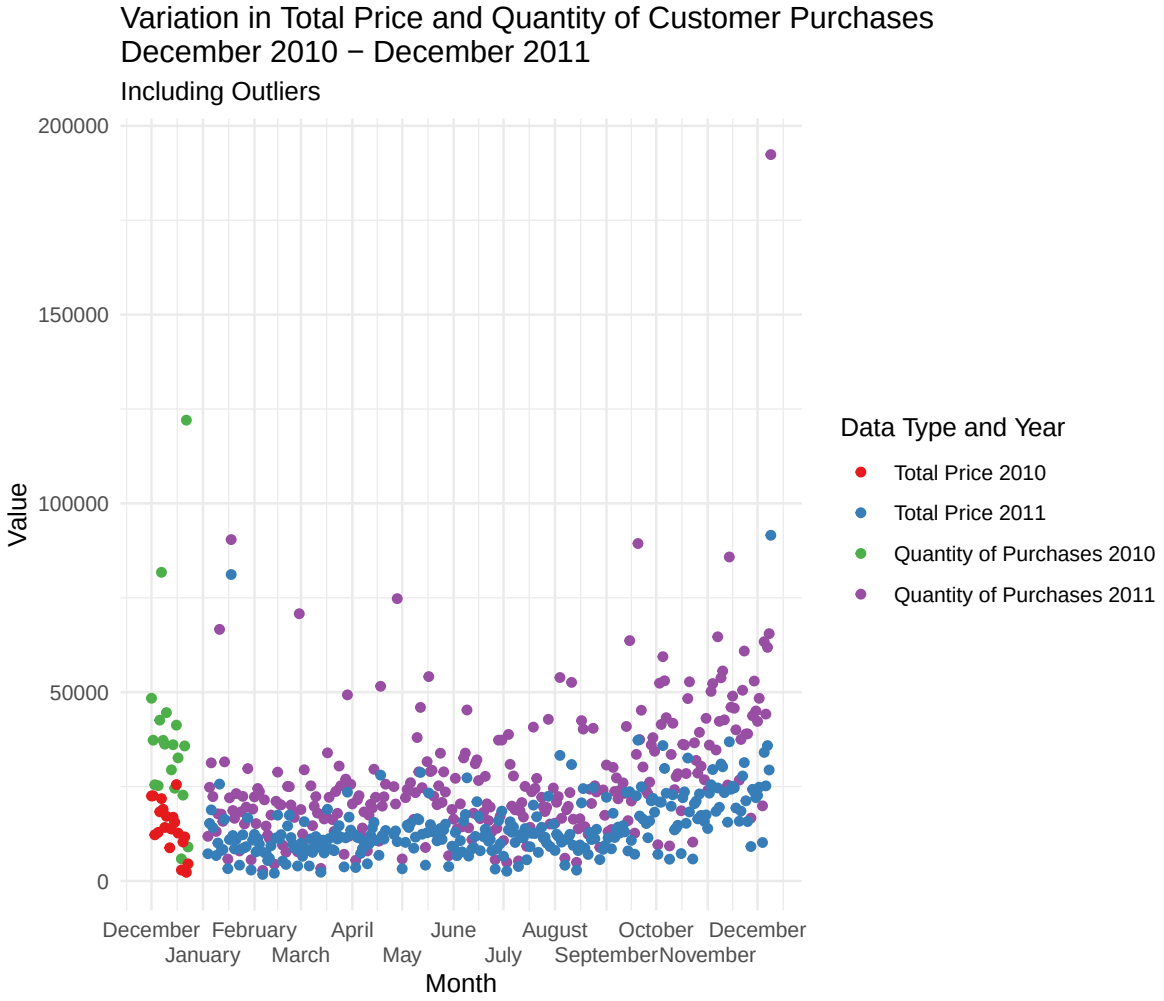


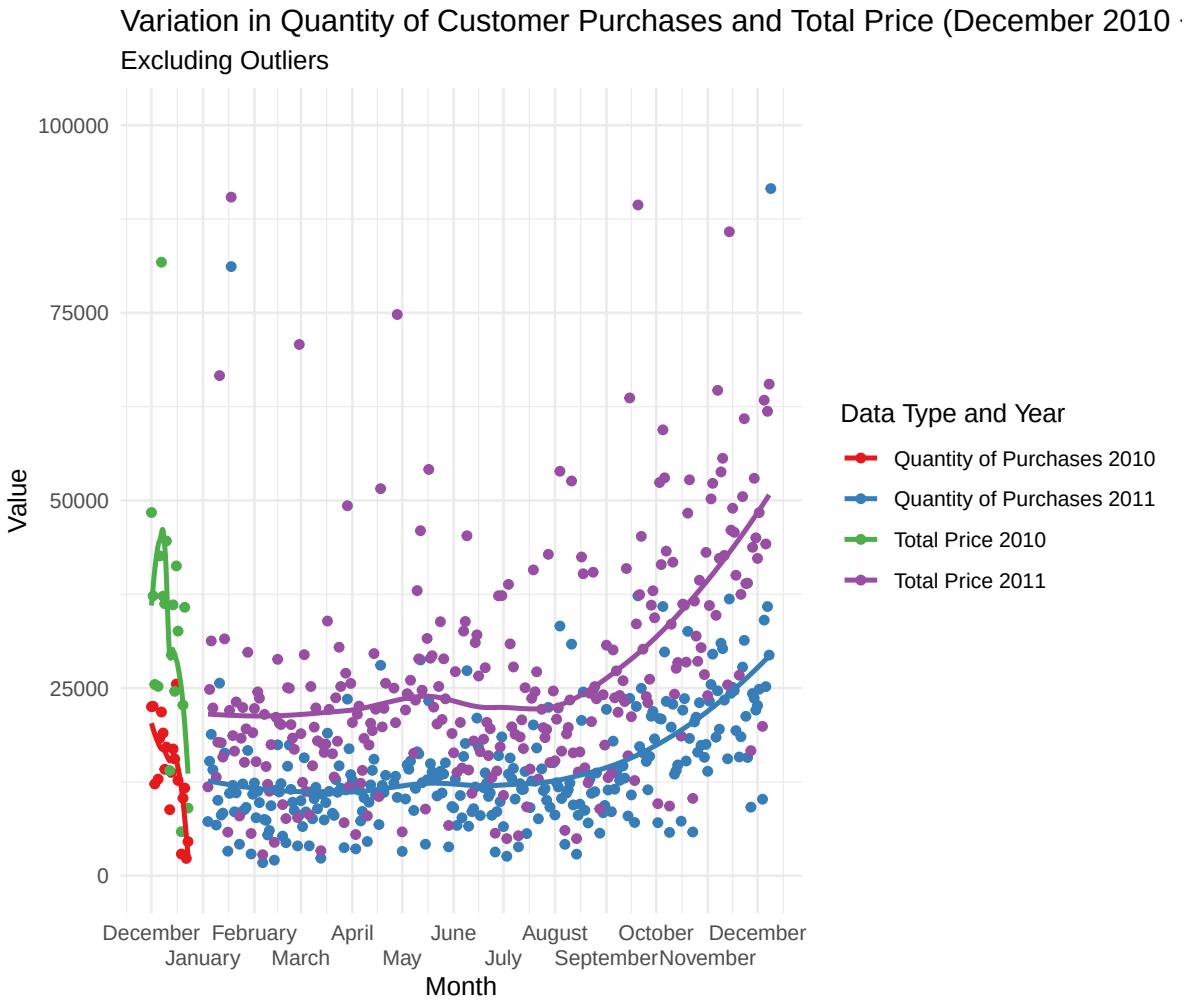
The above two visualisations handle the question of geographic location's impact on customer purchase behavior in a similar manner to the initial two visualisations presented, however, they instead make use of the TotalPrice of customer purchases per Country. Once again, one visualization includes outliers while the other does not- not because the outlier is insignificant, but rather to make other patterns in the data more legible.

Unsurprisingly, the United Kingdom again leads. This is fully anticipated, as there is a direct, linear correlation between the total Quantity of customer purchases and the TotalPrice of

customer purchases. It is in this way that the visualisations presented above further confirm the relationship between geographic location and customer purchasing behavior, with customers in European countries spending the largest Total Price of customer purchases behind the UK.

How Does Variation in Quantity and Total Price Illustrate the impact of Time of Year on Customer Purchasing Behavior at an Online Retail Platform?





The visualisations presented above refer to the variation in the quantity and total price of customer purchases recorded per day on an eCommerce platform from December 2010 to December 2011. Additionally, one visualization includes outliers while the other excludes outliers. This was done in order to make preexisting trends in the data more visible, and not to exclude the outliers from data analysis. Data points were also colored according to year and type for legibility, and lines of best fit were added to emphasize patterns in the data excluding outliers.

There are clear trends highlighted by the lines of best fit. In the visualization which includes outliers, the quantity of customer purchases peaked mid December 2010, and then sharply declined towards the end of the month. The quantity of customer purchases then plateaued at a median value from the beginning of January 2011, gradually increasing to a peak again in December 2011. There is a slight increase in the quantity of customer purchases around June 2011, and this is better illustrated in the above visualization which does not contain outliers.

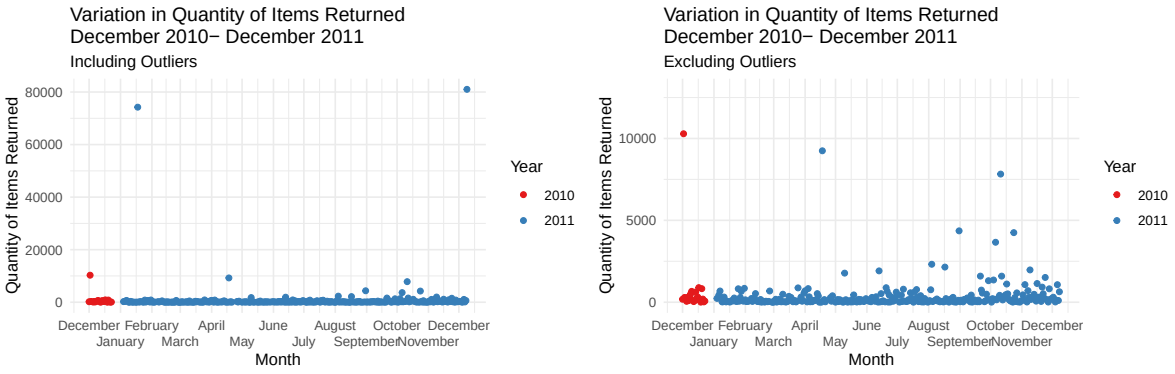
The largest outliers which arise in the visualization occur in February and December. While the December outlier is more expected, it is assumed that the occurrence of the outlier in February is likely due to post-holiday season clearance sales.

The trends in both December 2010 and December 2011 are anticipated, as customer purchasing behavior is expected to increase towards the end of year, and peak in December due to holiday spending, followed by a decline before New Year. While the the slight uptick in customer purchasing behavior in mid May is more unforeseen, there are a number of reasons which could account for the increase. One such reason could be the end of the financial year in the UK occurring at the beginning of April, triggering End of Financial Year sales as soon as May through to June.

Both visualisations indicate that the factor of time of year does have an impact on customer purchasing behavior, with customer purchasing behavior increasing as the months of the year progress. There are variations to this, evident in the uptick in the quantity of customer purchases made in May and significant decline which occurs after mid December.

While the general trend in total price spending is similar to the general trend in the quantity of purchases, the outliers fall differently. In the above visualization containing outliers, it is clear to see that the outlier which occurs at the end of December 2011 is consistent when viewing trends in both total price and Quantity, likely due to belated holiday shopping. However, the variation in the total price of customer purchases per day shows a wider spread of points along the y-axis, indicating that high points in the total price of customer purchases occurred fairly regularly at the end of most months. While this is consistent with expectations for usual end-of-month spending habits, a comparison between both types reveals there may be a correlation with customers spending more on less items- indicating customers spend more purchasing fewer, more expensive items at the end of each month.

How Does Variation in Quantity of Items Returned Illustrate the impact of Time of Year on Customer Purchasing Behavior at an Online Retail Platform?



The above two visualisations tackle customer purchasing behavior in a a different manner, as they focus on the quantity of returns made instead of sales. However, this is still an interesting

aspect of customer purchasing behavior which should not be excluded. As is expected, the amount of returns made peaked in a similar pattern to the quantity and total price of sales. Additionally, the four highest outliers on the graph all occur on or near the cusp of a month, indicating that returns peak around the beginning or end of most months.

Conclusion

The report presented above explored and identified how factors such as geographic location and time of year impact customer purchasing behavior in terms of both sales and returns at an online retail platform between December 2010 and December 2011. This was achieved through the performance of an exploratory data analysis in order to investigate patterns and trends in the dataset through the use of various visualisations, tables and summary statistics.

In order to perform this exploratory data analysis process, the eCommerce dataset first needed to be cleaned and prepared. This involved a variety of steps, such as inspecting, detecting, and replacing incorrect, missing, or fixed values and applying the "Last Value Carried Forward Treatment. Now that the eCommerce dataset had been successfully tidied, it could be categorized according to the nature of each row. This entailed the implementation of a column 'Category', followed by the addition of a TotalPrice column to complete variable calculations determining the total price of each row.

As part of the first step of the exploratory data analysis process, summary statistics were created for the key variables of the data set, such as Quantity, TotalPrice, Country, InvoiceDate, and Category. Each statistic revealed interesting patterns and trends in the data that were further explored through the use of visualisations.

These visualisations revealed correlations between the key variables of the dataset, aside from the direct linear relationships between variables such as Quantity and TotalPrice. These correlations included the ways in which geographic proximity to the UK was shown to effect the quantity and total price of customer purchases made, confirming the impact of geographic location on customer spending patterns. Additionally, the visualisations illustrated the ways in which time of year impacted the quantity and total price of customer purchasing behavior in stable patterns.

While the exploratory data analysis process was successful, there were limitations- such as the level of disarray of the eCommerce dataset prior to being cleaned and prepared. Multiple rows with significant quantities and total prices had to be classified into categories such as 'UNKNOWN' or 'ERROR' due to the lack of information denoting the nature of the transaction, and much of this data could have been significant for the data analysis process. Additionally, every care was put into sufficiently tidying the data with a systematic approach, however, the shortcomings of the dataset may have inevitably crept into the final analysis.

The exploratory data analysis process also faced limitations in terms of some of the given values. This particularly applied to the Country variable, as an area of further research could

pertain to the impact of city level geographic locations on customer purchasing behavior per country, which would further explore the impacts of geographic location.