Project 2: Report

Customer Segmentation in Retail Data

Report by: 25090135

Introduction

This project report aims to investigate the complexity of customer segmentation in retail data. Through meticulous cleaning ,transformation of the given data-set ,and a comprehensive Exploratory Data Analysis(EDA) methodology the report will uncover insights that inform stakeholders on value points from the customer retail data to make inferences on the firms performance.

Exploratory Data Analysis (EDA)

This report's EDA process focused on answering the following questions:

1. Regional Distribution Trends:
   - How do product quantities vary across different regions?
   - Is there a key region where most products are sold?
2. Monthly Sales Trends:
   - What are the monthly sales movement trend ?
3. Item Popularity
   - Which products are the most popular?
4. Customer Segmentation (RFM Analysis)
   - How can we segment customers and analysis the proportions based on RFM scores?

Data Analysis

```
Warning: package 'ggplot2' was built under R version 4.3.2
```

```
Warning: package 'tidyr' was built under R version 4.3.2
```

```
-- Attaching core tidyverse packages ----------------------- tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ----------------------------------------- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom

Attaching package: 'scales'


The following object is masked from 'package:purrr':

    discard


The following object is masked from 'package:readr':

    col_factor
```

```
Warning: package 'ggridges' was built under R version 4.3.2
```

```
Warning: package 'tinytex' was built under R version 4.3.3
```

Cleaning the data

```
Rows: 541,909
Columns: 8
$ InvoiceNo   <chr> "536365", "536365", "536365", "536365", "536365", "536365"~
$ StockCode   <chr> "85123A", "71053", "84406B", "84029G", "84029E", "22752", ~
$ Description <chr> "WHITE HANGING HEART T-LIGHT HOLDER", "WHITE METAL LANTERN~
$ Quantity    <dbl> 6, 6, 8, 6, 6, 2, 6, 6, 6, 32, 6, 6, 8, 6, 6, 3, 2, 3, 3, ~
$ InvoiceDate <dttm> 2010-12-01 08:26:00, 2010-12-01 08:26:00, 2010-12-01 08:2~
$ UnitPrice   <dbl> 2.55, 3.39, 2.75, 3.39, 3.39, 7.65, 4.25, 1.85, 1.85, 1.69~
$ CustomerID  <dbl> 17850, 17850, 17850, 17850, 17850, 17850, 17850, 17850, 17~
$ Country     <chr> "United Kingdom", "United Kingdom", "United Kingdom", "Uni~
```

Through initial viewing of the data we observe a data set with 541 909 rows and 8 columns. As part of cleaning we analyse the level of NA values and interpret their significance. Through the inspection we recognise 136 534 rows with NA values, upon inspection of this these variable we drop these values as they will hamper analysis.

```
[1] 136534
```

```
 [1] "United Kingdom"       "France"               "Australia"
 [4] "Netherlands"          "Germany"              "Norway"
 [7] "EIRE"                 "Switzerland"          "Spain"
[10] "Poland"               "Portugal"             "Italy"
[13] "Belgium"              "Lithuania"            "Japan"
[16] "Iceland"              "Channel Islands"      "Denmark"
[19] "Cyprus"               "Sweden"               "Austria"
[22] "Israel"               "Finland"              "Bahrain"
[25] "Greece"               "Hong Kong"            "Singapore"
[28] "Lebanon"              "United Arab Emirates" "Saudi Arabia"
[31] "Czech Republic"       "Canada"               "Unspecified"
[34] "Brazil"               "USA"                  "European Community"
[37] "Malta"                "RSA"
```

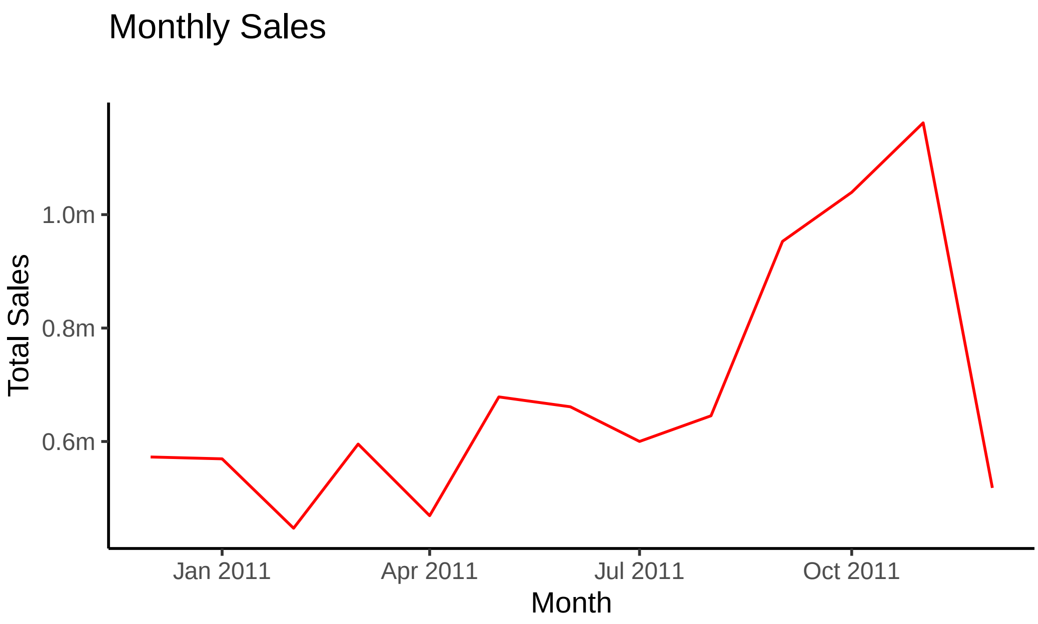
Furthermore we want to analyse unique variables to remove unwanted variables and group required variables in the data to analyse regions and clean our data for proper statistical view and analytical visualisation.

```
 [1] "United Kingdom"       "France"               "Australia"
 [4] "Netherlands"          "Germany"              "Norway"
 [7] "EIRE"                 "Switzerland"          "Spain"
[10] "Poland"               "Portugal"             "Italy"
[13] "Belgium"              "Lithuania"            "Japan"
[16] "Iceland"              "Channel Islands"      "Denmark"
[19] "Cyprus"               "Sweden"               "Finland"
[22] "Austria"              "Greece"               "Singapore"
[25] "Lebanon"              "United Arab Emirates" "Israel"
[28] "Saudi Arabia"         "Czech Republic"       "Canada"
[31] "Unspecified"          "Brazil"               "USA"
[34] "European Community"   "Bahrain"              "Malta"
[37] "RSA"
```

Transformations and Visualisations

1. Monthly Sales Trends

```
# A tibble: 13 x 2
   Month               TotalSales
   <dttm>                   <dbl>
 1 2010-12-01 00:00:00    572714.
 2 2011-01-01 00:00:00    569445.
 3 2011-02-01 00:00:00    447137.
 4 2011-03-01 00:00:00    595501.
 5 2011-04-01 00:00:00    469200.
 6 2011-05-01 00:00:00    678595.
 7 2011-06-01 00:00:00    661214.
 8 2011-07-01 00:00:00    600091.
 9 2011-08-01 00:00:00    645344.
10 2011-09-01 00:00:00    952838.
11 2011-10-01 00:00:00   1039319.
12 2011-11-01 00:00:00   1161817.
13 2011-12-01 00:00:00    518193.
```

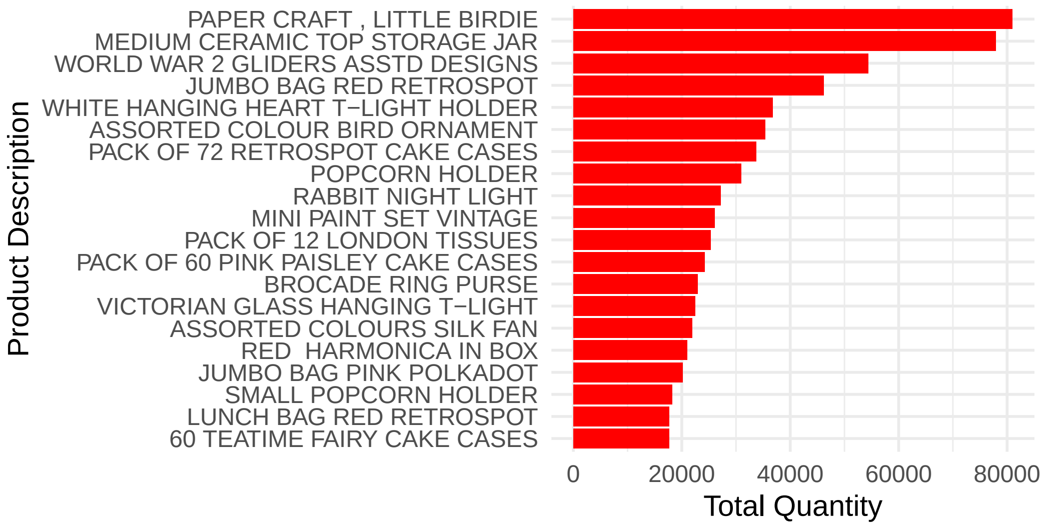


Data from: Online Retail Platfrom (Jan 2010 − Dec 2011)

The above graph deplicts the trend of monthly sales from the firm. The data indicates that sales within the firm experienced exponential growth from January to October, however we notice a massive spike nearing December approximately in November. This massive spike can be explained as a result of festive season sales in retail such as Black-Friday and Cyber-Monday which add an additional 20% *(as of 2023)* increase in a hyper purchase season (Absa, 2023). The decline van thus be attributed to consumers maximising the lower price season and stores often being closed for days in the December holiday season and lowered operating times.

2. Top 20 Most Sold Items

```
# A tibble: 20 x 2
   Description                        TotalQuantity
   <chr>                                      <dbl>
 1 PAPER CRAFT , LITTLE BIRDIE                80995
 2 MEDIUM CERAMIC TOP STORAGE JAR             77916
 3 WORLD WAR 2 GLIDERS ASSTD DESIGNS          54415
 4 JUMBO BAG RED RETROSPOT                    46181
 5 WHITE HANGING HEART T-LIGHT HOLDER         36725
 6 ASSORTED COLOUR BIRD ORNAMENT              35362
 7 PACK OF 72 RETROSPOT CAKE CASES            33693
 8 POPCORN HOLDER                             30931
 9 RABBIT NIGHT LIGHT                         27202
10 MINI PAINT SET VINTAGE                     26076
11 PACK OF 12 LONDON TISSUES                  25345
12 PACK OF 60 PINK PAISLEY CAKE CASES         24264
13 BROCADE RING PURSE                         22963
14 VICTORIAN GLASS HANGING T-LIGHT            22433
15 ASSORTED COLOURS SILK FAN                  21876
16 RED  HARMONICA IN BOX                      20975
17 JUMBO BAG PINK POLKADOT                    20165
18 SMALL POPCORN HOLDER                       18252
19 LUNCH BAG RED RETROSPOT                    17697
20 60 TEATIME FAIRY CAKE CASES                17689
```



Top 20 Most Sold Items

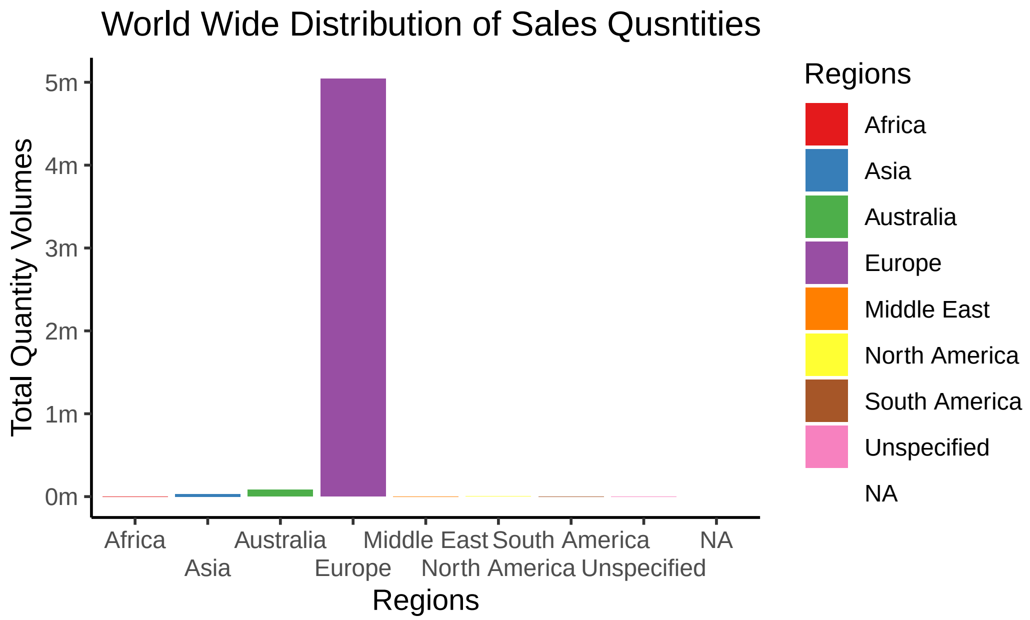
Data from: Online Retail Platfrom (Jan 2010 − Dec 2011)

Monthly sales data represents the revenue generated month-to-month, howevereven more value can be derived from analysing which products generate the most monetary value. The above

3. Analysis of Region Sales Volumes

```
# A tibble: 9 x 2
  Regions       TotalQuantity
  <chr>                 <dbl>
1 Europe              5043267
2 Australia             84209
3 Asia                  31257
4 <NA>                  13537
5 North America          5221
6 Unspecified            1789
7 Middle East            1708
8 South America           356
9 Africa                  352
```

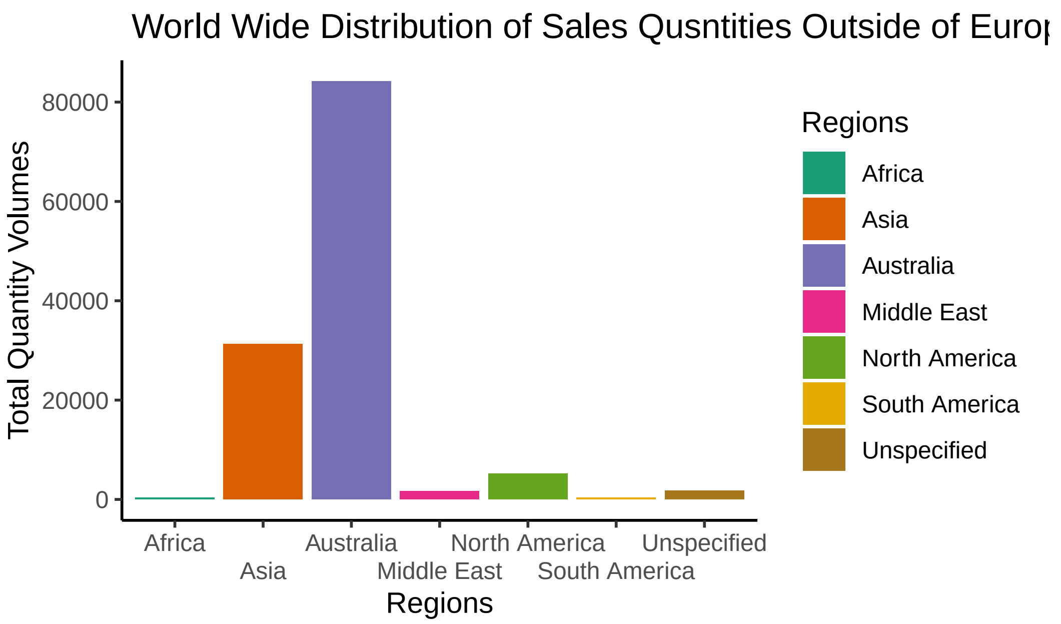
In the previous section we noticed the sales trends present within the firm's monthly sales volumes and most sold items. For further analysis we can view the region which generates the most sales for the firm to determine their base operations based on the volume of quantity moved within that region. In this data we can deduce that the firm primary region is Europe, with the nation generation the most quantities being the United Kingdom.



Data From: Online Retail Platfrom (Jan 2010 − Dec 2011)

4. Region Quantities Outside of Europe

```
# A tibble: 7 x 2
  Regions       TotalQuantity
  <chr>                 <dbl>
1 Australia             84209
2 Asia                  31257
3 North America          5221
4 Unspecified            1789
5 Middle East            1708
6 South America           356
7 Africa                  352
```


World Wide Distribution of Sales Qusntities Outside of Euro

Data From: Online Retail Platfrom (Jan 2010 − Dec 2011)

In order to get a better view of quantity distribution throughout different regions provided by (online retail platform, 2024), we eliminate Europe to analyse the value added by other regions. The graph shows that the firm's next most valuable source of sales is the Australian region followed by Asia.

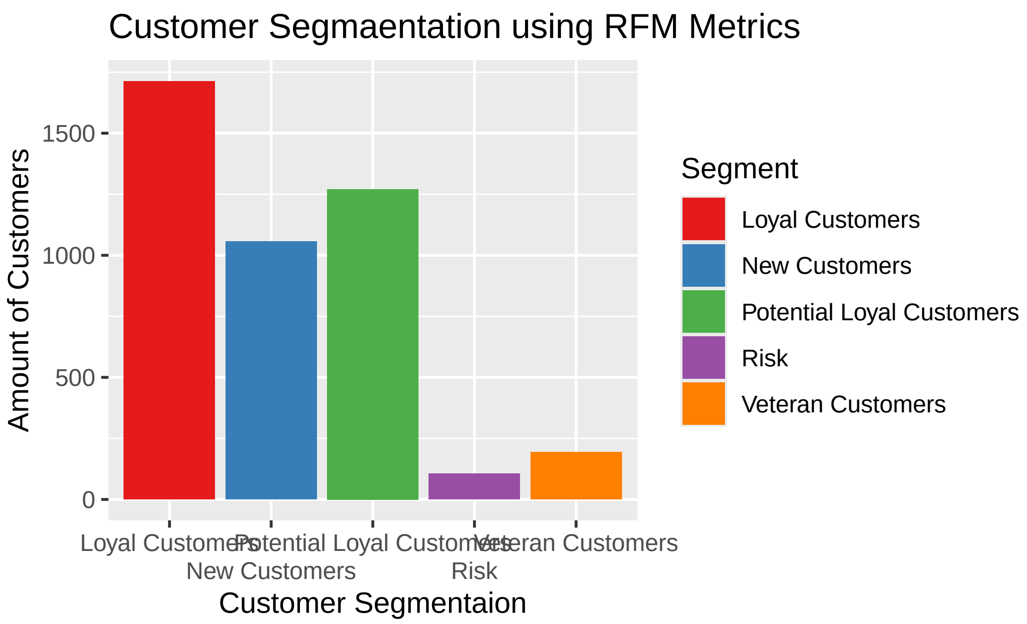
5. Customer Segmentation: RFM Analysis

```
# A tibble: 4,339 x 8
   CustomerID Recency Frequency Monetary RecScore FreqScore MonScore RFM_Score
        <dbl>   <dbl>     <int>    <dbl>    <int>     <int>    <int>     <int>
 1      12346  -5017.         1   77184.        5         1        5        11
 2      12347  -4693.         7    4310         1         5        5        11
 3      12348  -4766.         4    1797.        4         4        4        12
 4      12349  -4710.         1    1758.        2         1        4         7
 5      12350  -5001.         1     334.        5         1        2         8
 6      12352  -4727.         8    2506.        3         5        5        13
 7      12353  -4895.         1      89         5         1        1         7
 8      12354  -4923.         1    1079.        5         1        4        10
 9      12355  -4905.         1     459.        5         1        2         8
10      12356  -4714.         3    2811.        2         3        5        10
# i 4,329 more rows
```

In RFM Analysis we aim to analyse the value in which customers add to the firm in terms of their spending within the Firm. We are able to see that using the frequency in which they purchase from the firm, coupled with their last purchase and the amount in which customers spend (Septia, A. 2024). The analysis using mathematics to extract this information which can serve as a rich insight for firms and provide them with great insight on the scale of their customer retention.

```
# A tibble: 4,339 x 9
   CustomerID Recency Frequency Monetary RecScore FreqScore MonScore RFM_Score
        <dbl>   <dbl>     <int>    <dbl>    <int>     <int>    <int>     <int>
 1      12346  -5017.         1   77184.        5         1        5        11
 2      12347  -4693.         7    4310         1         5        5        11
 3      12348  -4766.         4    1797.        4         4        4        12
 4      12349  -4710.         1    1758.        2         1        4         7
 5      12350  -5001.         1     334.        5         1        2         8
 6      12352  -4727.         8    2506.        3         5        5        13
 7      12353  -4895.         1      89         5         1        1         7
 8      12354  -4923.         1    1079.        5         1        4        10
 9      12355  -4905.         1     459.        5         1        2         8
10      12356  -4714.         3    2811.        2         3        5        10
# i 4,329 more rows
# i 1 more variable: Segment <chr>
```

The table above shows us the distribution of the firm’s customers based on their RFM scores and through analysis we can deduce the spread through visualisation in a bar chart.



Data from: Online Retail Platform (Jan 2010 − Dec 2011)

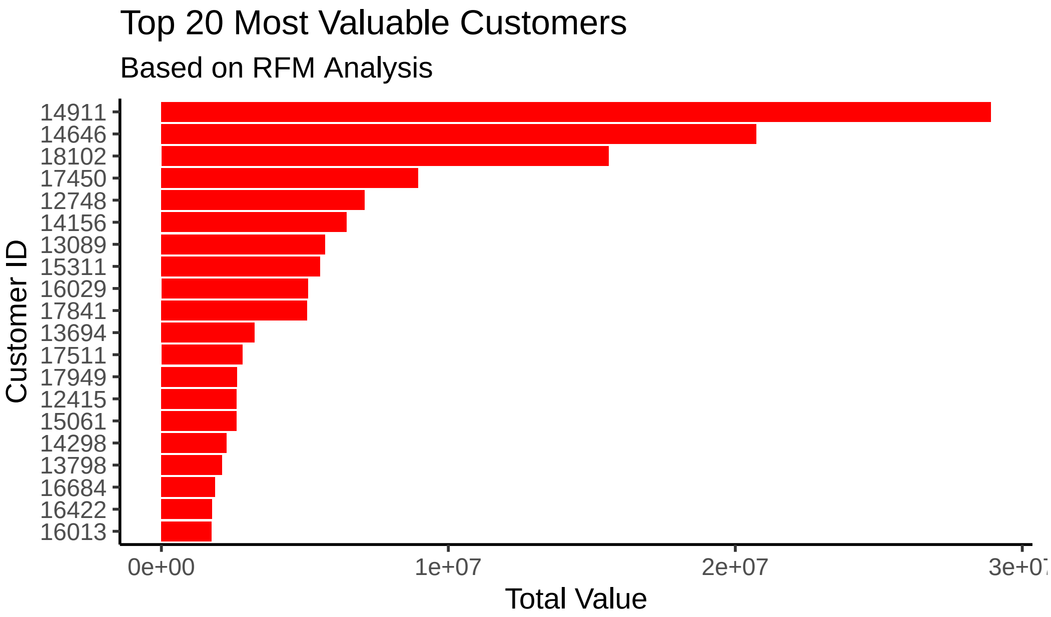
The bar graph above indicates there's a positive distribution as the firm has most of it's customers falling under their loyal bracket and the second highest peak being their potential loyal customers, and the amount of risk customers is low indicating that the firm has positive relations with their consumers and their current business processes are positive and giving good returns on investment.

6. Customer Segmentation: Customer Value

We will then analyse the total customer value to derive what each individual customer added in their spend and look at the top 20 most valuable customers.

```
# A tibble: 4,339 x 10
   CustomerID Recency Frequency Monetary RecScore FreqScore MonScore RFM_Score
        <dbl>   <dbl>     <int>    <dbl>    <int>     <int>    <int>     <int>
 1      12346  -5017.         1   77184.        5         1        5        11
 2      12347  -4693.         7    4310         1         5        5        11
 3      12348  -4766.         4    1797.        4         4        4        12
 4      12349  -4710.         1    1758.        2         1        4         7
 5      12350  -5001.         1     334.        5         1        2         8
 6      12352  -4727.         8    2506.        3         5        5        13
 7      12353  -4895.         1      89         5         1        1         7
 8      12354  -4923.         1    1079.        5         1        4        10
 9      12355  -4905.         1     459.        5         1        2         8
```

```
10      12356  -4714.         3    2811.        2         3        5        10
# i 4,329 more rows
# i 2 more variables: Segment <chr>, CustomerValue <dbl>
```



Data from: Online Retail Platform (Jan 2010 − Dec 2011)

The above graph is indicative to show which customers spent the most, which is a valuable stat to have in order to know which consumers to focus most on and keep communication most with for relationship building as they add the most value to the firm.

Conclusion

The online retail data platform data in it's wholeness provided invaluable insight into how the firm can grow it's relationships with customers as well as the sources and time frames for which most value is added to the firm. The study was then successful in asweing it's EDA questions

References

Absa. 2024. Black Friday Results 24-27 November. *ABSA*. [Online] Avaialable: https://cib.absa.africa/wp-content/uploads/2023/12/2023-18553-BlackFriday2023-updated.pdf (Accessed: 5 0ctober 2024)

Septia, Andina. 2024. Inroduction to RFM Analysis in R. *Supertype.* [Online] Available: https://supertype.ai/notes/rfm-analysis-r-examples/ (Accessed : 5 October 2024)

WICKHAM, Hadley, ÇETINKAYA-RUNDEL, Mine and GROLEMUND, Garrett, 2023. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. Second edition. Beijing ; Sebastopol, CA: O'Reilly [Online] Available : https://r4ds.hadley.nz

https://stackoverflow.com/questions/73657563/how-to-select-range-of-unique-character-values-in-dplyr

https://stackoverflow.com/questions/28953934/how-to-create-a-simple-heatmap-in-r