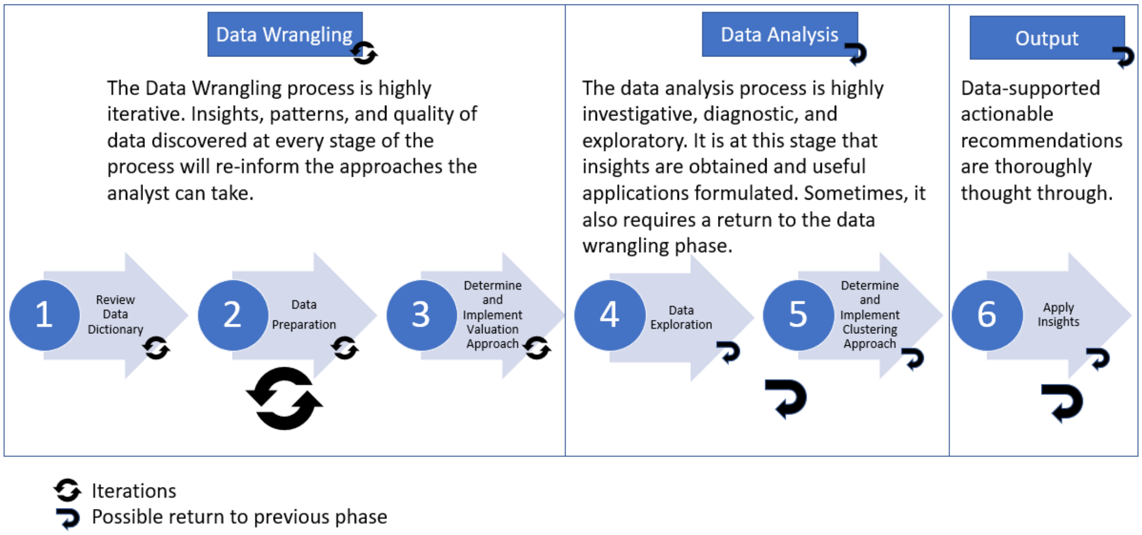
How We Do It: Data Mining Transactions to Know Our Customers

We used an E-Commerce transactional dataset (Ecommerce dataset) and a data visualiser (Tableau).

Our method looks like this:



Data Wrangling

The Data Wrangling process is highly iterative. Insights, patterns, and quality of data discovered at every stage of the process will re-inform the approaches the analyst can take.

Data Analysis

The data analysis process is highly investigative, diagnostic, and exploratory. It is at this stage that insights are obtained and useful applications formulated. Sometimes, it also requires a return to the data wrangling phase.

Output

Data-supported actionable recommendations are thoroughly thought through.

1 Review Data Dictionary

2 Data Preparation

3 Determine and Implement Valuation Approach

4 Data Exploration

5 Determine and Implement Clustering Approach

6 Apply Insights

⟳ Iterations

⤵ Possible return to previous phase

Each of the six stages includes the following actions

1. **Data Wrangling**
   a. **Review Data Dictionary**
      - Examine datasets
      - Outline possible approaches
      - Solidify analyses objectives

   b. **Data Preparation**
      - Remove extraneous dimensions
      - Create new useful dimensions
      - Add new data

   c. **Determine Valuation Approach**
      - Identify the method that would best meet the objective
2. **Data Analysis**
   a. **Data Exploration**
      Identify patterns and trends that would affect our analysis or provide insights

   b. **Determine Clustering Approach**
      Decide the clustering approach that would best extract useful insights based on the information gathered
3. **Output**

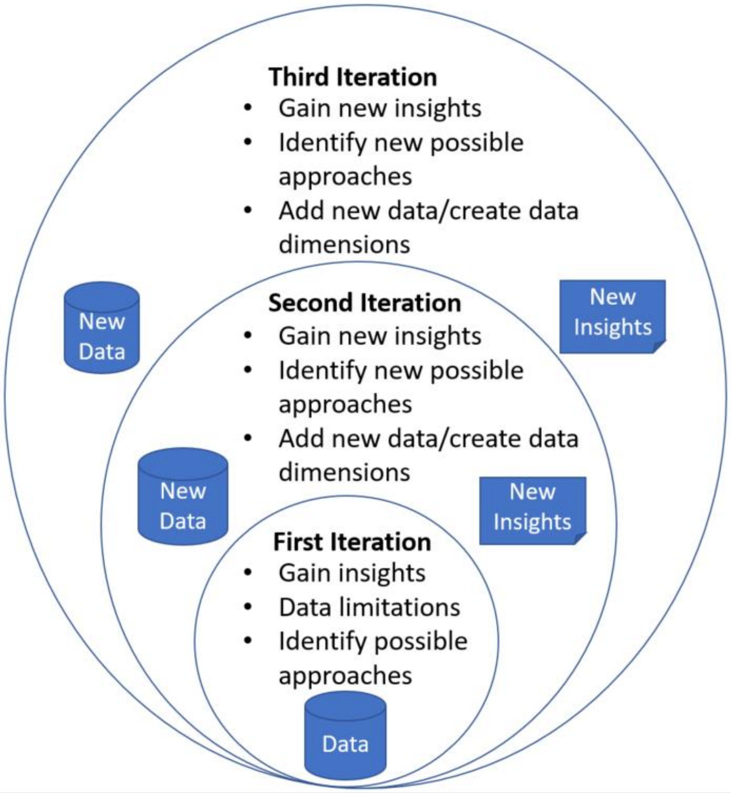
a. **Implement Approach**
    - Segment customers based on determined approach

b. **Apply insights**
    - Suggest actions based on findings
    - Plan for future iterations of same experiment to obtain better insights

PHASE 1: DATA WRANGLING (REVIEW, PREPARE, DETERMINE)

Data wrangling typically takes up more than 70% of the time invested in any project. It is typically reiterative, and I describe it as the “Cycle of all Cycles” because it may never end.

The Cycle of all Cycles


Third Iteration
• Gain new insights
• Identify new possible approaches
• Add new data/create data dimensions
New Data
New Insights
Second Iteration
• Gain new insights
• Identify new possible approaches
• Add new data/create data dimensions
New Data
New Insights
First Iteration
• Gain insights
• Data limitations
• Identify possible approaches
Data

And it goes on and on……. until:
- You are satisfied with the dimensions
- Limited by the data at the current time
- Constrained by time
- You are tired

This also means there will be a variety of ways to meet the same objective depending on factors like the number of iterations done, the experience of the analyst, the creativity and flexibility applied on the data and so on. Below I describe the procedure I ended up with.

1A. REVIEW DATA DICTIONARY

At this point, it is useful to think about the bigger picture. I typically ask:

- What is the objective of this exercise?
- What are the possible approaches?
- Which approach is the most efficient and should be explored first?

- Which approach is the most reliable/accurate/ business-friendly?
- Will the approach incur loss of information?
- What other data is available?
- What is the nature of the data?

These questions will be repeatedly asked and answered as we iterate through this data wrangling phase. Let's define our parameters and objectives for this exercise.

Objective
To identify and prioritise business-applicable customer segments.

Possible Approaches
Valuation Framework
Clustering Method

Nature of Business
B2C

Datatype
Transactional procurement/purchasing data

1B. PREPARE DATA

A peak into the dataset's data dictionary shows these 8 variables:

1. Invoice No.
2. Stock Code
3. Description
4. Quantity
5. Invoice Date
6. Unit Price
7. Customer ID
8. Country

I removed these observations as they would detrimentally affect our analysis:

- **December data as it was incomplete**
- **Transactions with no Customer ID as our objective is customer ID orientated**
- Credit Notes and their corresponding Invoices
- Adjustments to Bad Debt
- Discounts
- Free Gifts

- Bank Charges
- Fees
- Stock Taking adjustments (E.g. damages, lost, discarded, etc.)
- Promotional Gifts
- Missing Customer (Useful but not for the objective of this article)

In addition, I created a few more dimensions to interpret the data:

1. Cutoff Date – The date for which our analysis is cutoff
2. Difference between Deadline and Invoice Date (days)
3. Invoice Month (MM-YYYY)
4. Invoice Day of the week (Mon, Tues, etc.)
5. Time Period (Morning, Afternoon, etc.)
6. Unit Price Group (Small Purchase, Medium Purchase, etc.)
7. Quantity Group (Small Quantities, Medium Quantities, etc.)
8. Total revenue per invoice

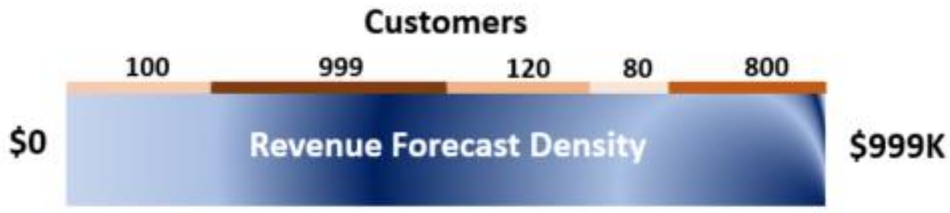
These will help us slice up the dataset into analysable segments.

1C. Evaluate Data Distribution & Data Dissection with Ratio Significance

Ratio Significance is essentially examining the distribution of data across each variable and evaluating if there are certain concentrations of values that suggest it should be classified on its own.

This helps us statistically capture and classify similar data into categories.

For example



5 Groups. Useful for fine-grained splits.

Side note: Each data column must be examined in its own context. For this scenario, I did not use the percentile method, which is a “crowd favourite,” because of the bell curve effect; i.e. classifying the top 25 as “high” amongst a huge number of “high” items would not fairly represent the data.

1C. DETERMINE VALUATION APPROACH

We need to create useful customer segments. To do that, we have to score each segment within a framework. This segmentation score is then applied to each customer to "bucket" them into the appropriate customer segments.

But what scoring method would be the most useful? There are many scoring frameworks like:

1. RFD (Recency, Frequency, Duration)
2. RFE (Recency, Frequency, Engagement)
3. RFM (Recency, Frequency, Monetary Value)
4. RFM-I (Recency, Frequency, Monetary Value – Interactions)
5. RFMTV (Recency, Frequency, Monetary Value, Time, Churn Rate)
6. Customised Framework

Depending on the business context and the objective of each exercise, different scoring matrixes would be useful.

Remember: We have to identify and prioritise business-applicable customer segments for an E-Commerce business.

Simple is always good. Especially when we have to explain our findings to stakeholders. Hence, I chose the basic RFM ("Recency, Frequency, Monetary") framework:

1. Revenue - Total revenue generated from customer
2. Frequency - Total transactions performed by each customer
3. Recency – No. of days from last transaction by each customer

Using the same methodology as above; i.e. evaluating the distribution of each column and dissecting the data by evaluating their ratio significance, I created the scoring matrix below which we'll use on our customer dataset.

Scoring	Total Revenue ($)	Frequency (n)	Recency (day)
5 points	>87,783.50	>5000	<30
4 points	<25,049.80	<1000	<90
3 points	<6,770.02	<500	<180
2 points	<1,028.57	<100	<360
1 points	<485.035	<10	>720

To summarise, we have created segments based on Revenue, Frequency and Recency scores. We score each customer based on these metrics. Then use their scores to classify customers into the right segment.

PHASE 2: DATA ANALYSIS (EXPLORE, CLUSTER)

The valuation approach (FRM, RFM-I etc.) will affect our actual customer segmentation. In preparing the data, I considered these possible clustering methods:

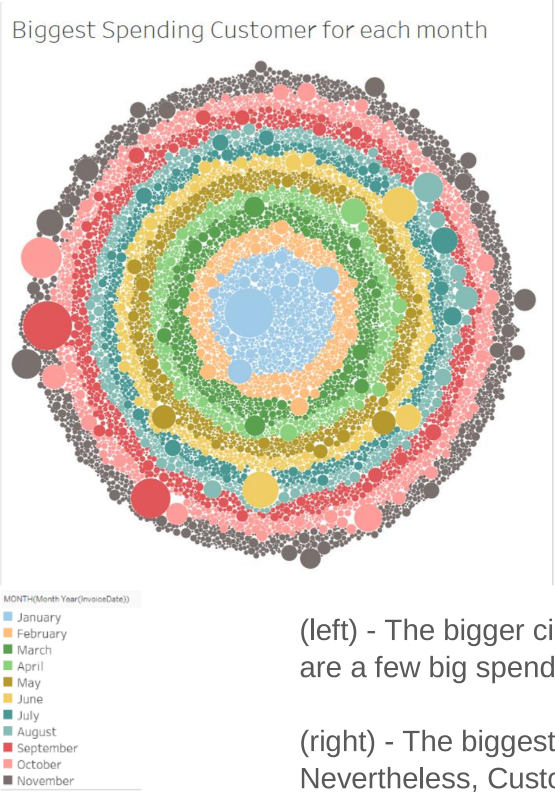
1. K-Means Clustering
2. Decision Tree Analysis
3. Multivariate Analysis
4. Latent Class Analysis
5. Mathematical formulas
6. Transformed Scoring

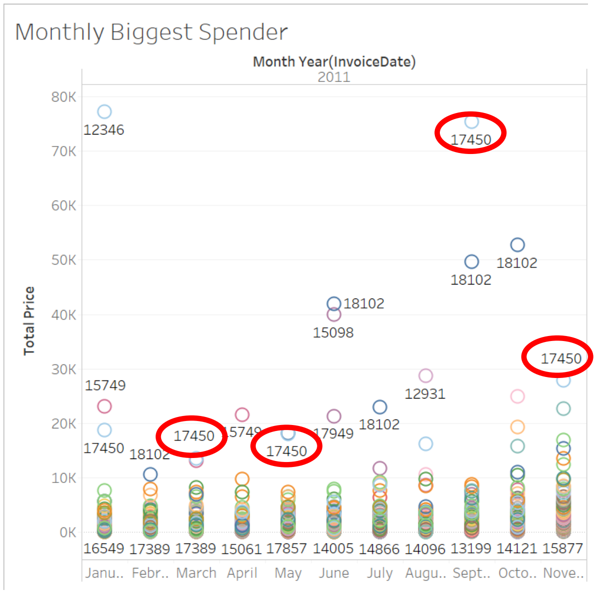
But which method should I use? Now, I explore the data to make an informed decision.

2A. DATA EXPLORATION

We determined that the RFM approach should be taken. Hence, we explore the data from these perspectives.

Monetary

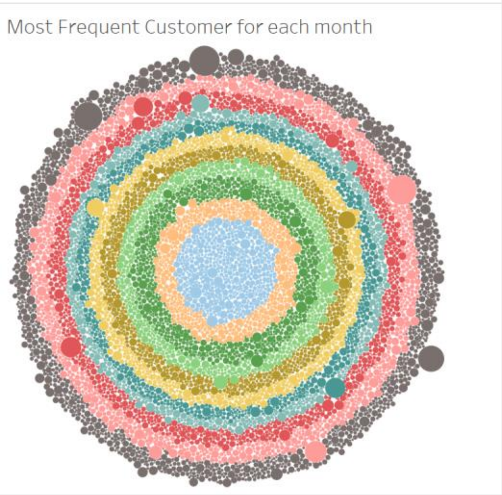


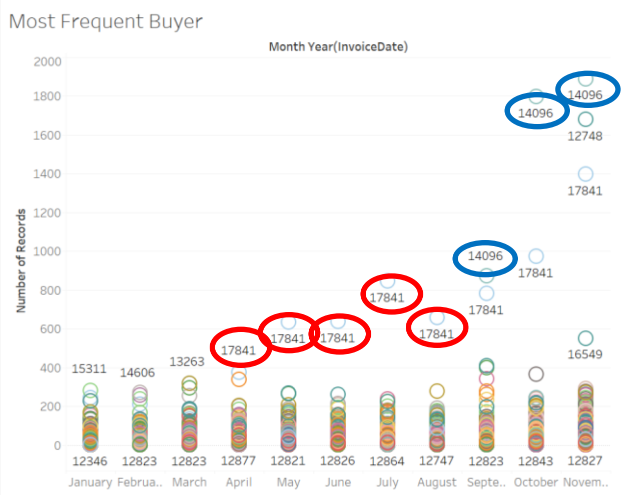


(left) - The bigger circles in different colors indicate that there are a few big spenders every month.

(right) - The biggest spender every month differs. Nevertheless, Customer ID 17450 appeared as top spender 4 times; in March, May, September and November (**Circled red**).

Frequency





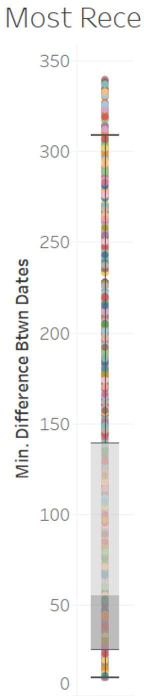
MONTH(Month Year(InvoiceDate))

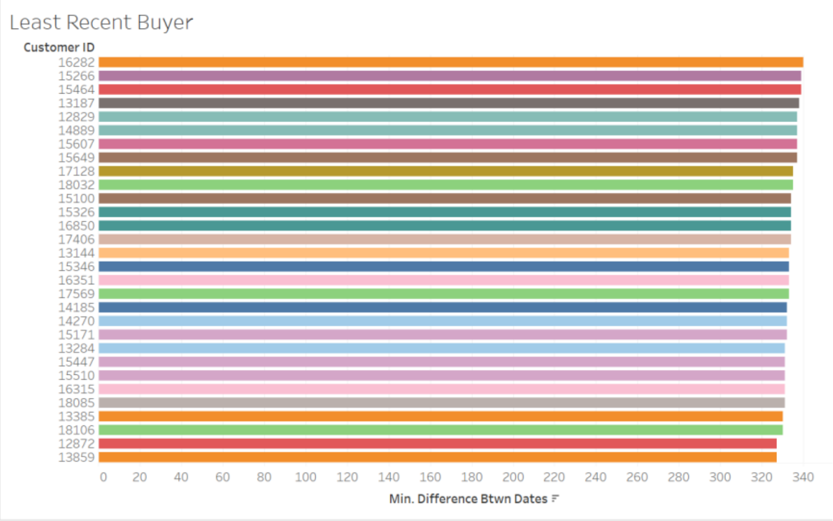
- January
- February
- March
- April
- May
- June
- July
- August
- September
- October
- November

(left) - The bigger circles in different colors indicate that there are a few high quantity purchase customers every month.

(right) - The highest quantity customer differs each month. Nevertheless, Customer ID 14096 purchased significantly higher quantity of items towards the end of the year (**Circled in blue**) and Customer ID 17841 appeared as the highest quantity customer for 5 months; April, May, June, July and August (**Circled in red**).

Recency

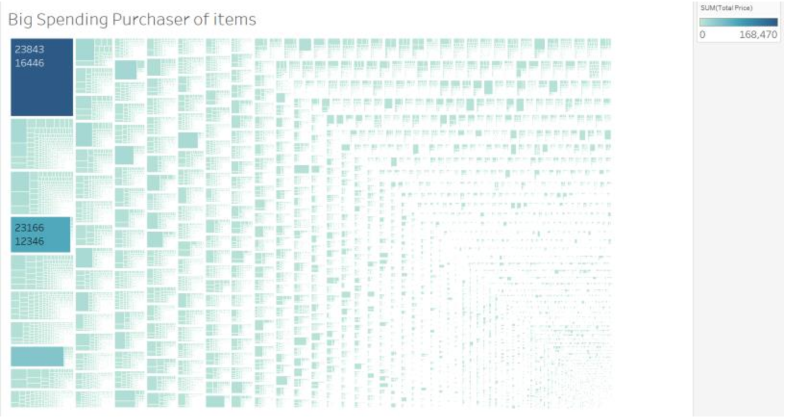




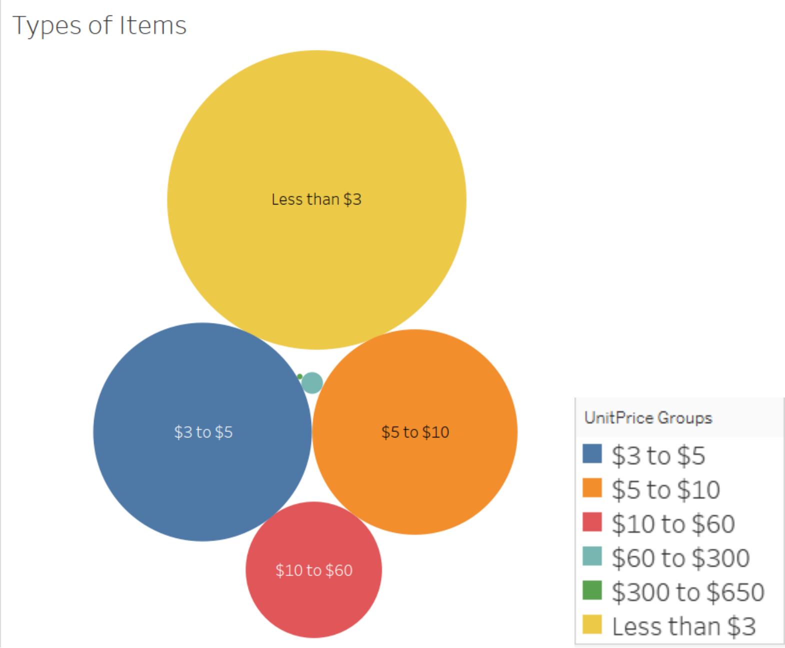
(left) - The mean and mode appear in the lower half of the box plot. This is good as it means most customers have made purchases recently with mean at 55 days. However, there is a fair number of customers that last made purchases very long ago. We don't want to lose these customers!

(right) These guys last made purchases more than 320 days ago; almost a year!

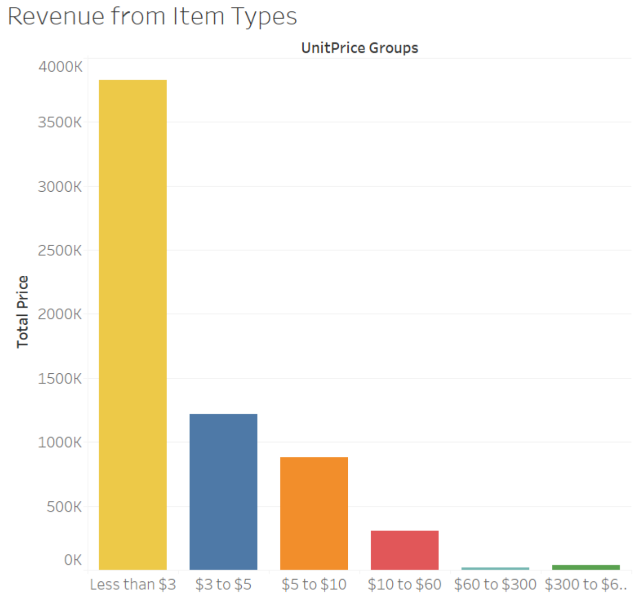
Product point of view.



Here, we can see that Item 23843 is bought only by Customer ID 16446. Furthermore, it was the top selling item. Item 23166 came in as the fourth highest revenue generated and it was bought only by Customer ID 12346.

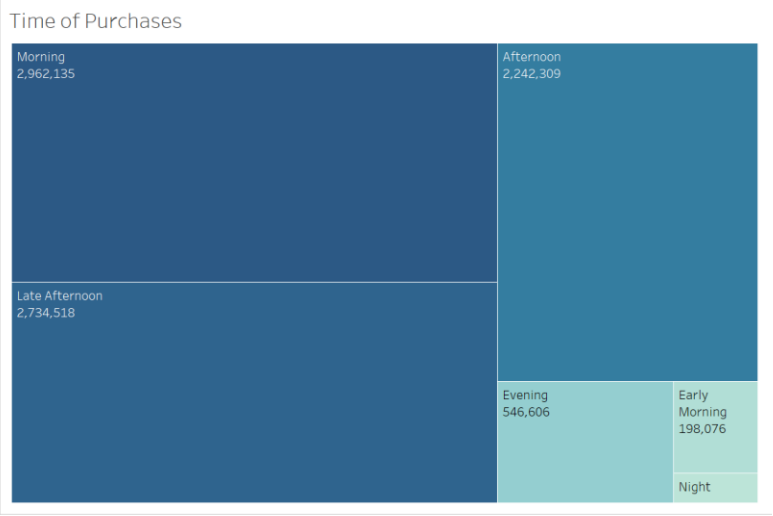


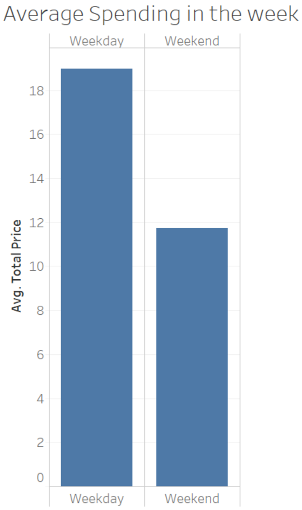
Here, we see that the cheapest items had the most variety and the variety of the most expensive items were significantly lesser.



The revenue generated from the cheapest items generated the most revenue.

The revenue generated from the most expensive items generated the least revenue.

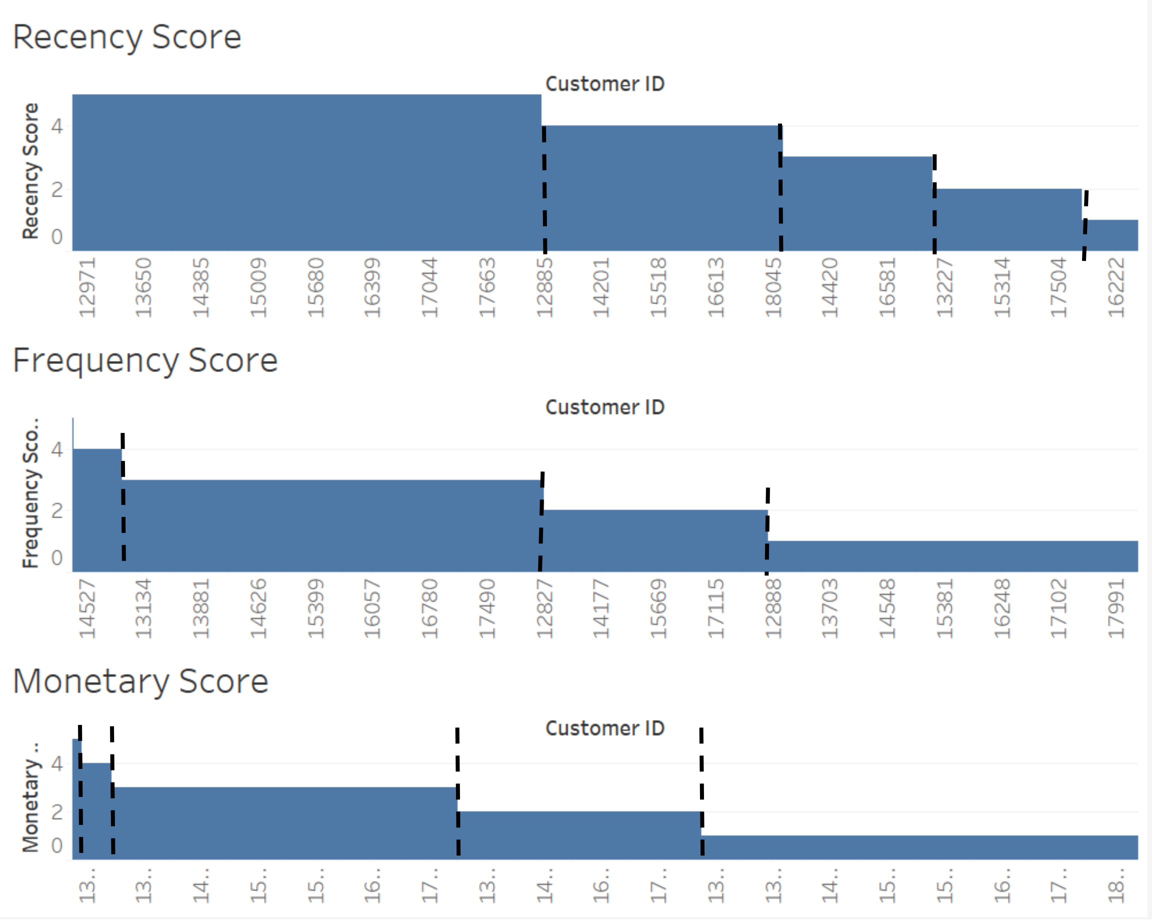




(left) - Most purchases occurred during office hours.

(right) - Most transactions took place on weekdays.

Proportion of Scores



The distribution of score for Monetary, Frequency and Recency differs in proportion.

Data Exploration Summary

From the visualisations, we have identified a few patterns that we should take into consideration when segmenting customers. These insights include:

1. The proportion of scores for Recency, Frequency and Monetary differs and we have to structure our scoring matrix to accommodate this.
2. Some items are only bought by certain customers and the revenue generated from these sales are quite significant.
3. There are more purchases on weekdays. Should we segment customers based on the date of purchases?
4. Small purchases generate the most revenue and big purchases generated the least revenue.

PS: We have identified several other insights unrelated to the customer segmentation. Hence, we've not included them in this article.

2B. DETERMINE CLUSTERING APPROACH

As explained above, there are a few methods that we can use to segment customers:

1. K-Means Clustering
2. Decision Tree Analysis
3. Multivariate Analysis
4. Latent Class Analysis
5. Mathematical formulas

These methods can be classified into:

1. Supervised Learning
   We know the outcomes. We simply apply user-supervised Machine Learning processes to this dataset, compare the outcomes from the Machine Learning process, and adjust accordingly.

2. Unsupervised Learning
   We don't know the outcomes. So, we use a bunch of mathematical algorithms to identify patterns in the data with minimal user interference.

Through data exploration, we have already identified several patterns that we can use, hence, a Supervised Learning method would provide more business-applicable results as it allows us to control the classification in a manner that would be the most useful in a business context. Either the Latent Class Analysis or a simple logical method comes to mind.

The thought process behind the classifications method described above can be intuitively explained using this logical process. It is also easy to realise that if an unsupervised mathematical algorithm is forced upon the data, some business insights may be lost. Hence, for the reasons described above, I went with the simple logical process.

PHASE 3: OUTPUT

First, I filtered out the items that were bought solely by single customers (**From "Big Spending Purchaser by items" Chart**). There is no use including these data as part of the classification process because nobody else would purchase these items but the specific customer. These items should be evaluated on their own.

Next, I evaluated the customers who would buy the big purchases (**From "Revenue from Item Types" Chart**). These transactions generated the least revenue. I looked at all the $300 to $650 items and noted that only Customer ID 15098 was purchasing these items. Upon closer inspection, Customer ID 15098 only made 3 purchases over the year, but the quantities were huge; a bulk purchaser. Let's exclude this customer from the classification exercise as he was the only one purchasing expensive items and clustering him together with the rest would not be useful for business-applicability reasons.

A more thorough exploration can be performed to identify more anomalies similar to the above 2 examples, but I think the point is clear now; Bulk classification would group anomalies together within a cluster/segment, thereby compromising accuracy over efficiency. Depending on the business context, the usefulness of the final classification will then be affected. Ever received marketing calls selling specific items that were totally irrelevant to you? You were probably misclassified.

Based on this sample, let's classify everyone else now.

3A. Apply Clustering Approach

3 data columns were used:

- Difference between deadline and invoice date for each transaction
- Total number of Invoices per Customer
- Total amount spent per customer

They were applied in the following manner to obtain the respective scores:

- **Recency Score**
  The smallest "Difference Between Deadline and Transaction Date" for each customer determined their Recency Score
- **Frequency Score**
  The total number of invoices per customer determined their frequency score
- **Monetary Score**
  The total amount spent per customer determined their monetary score

The below segmentation framework was built with the business context in mind after considering the distribution of scores for Recency, Frequency and Monetary.

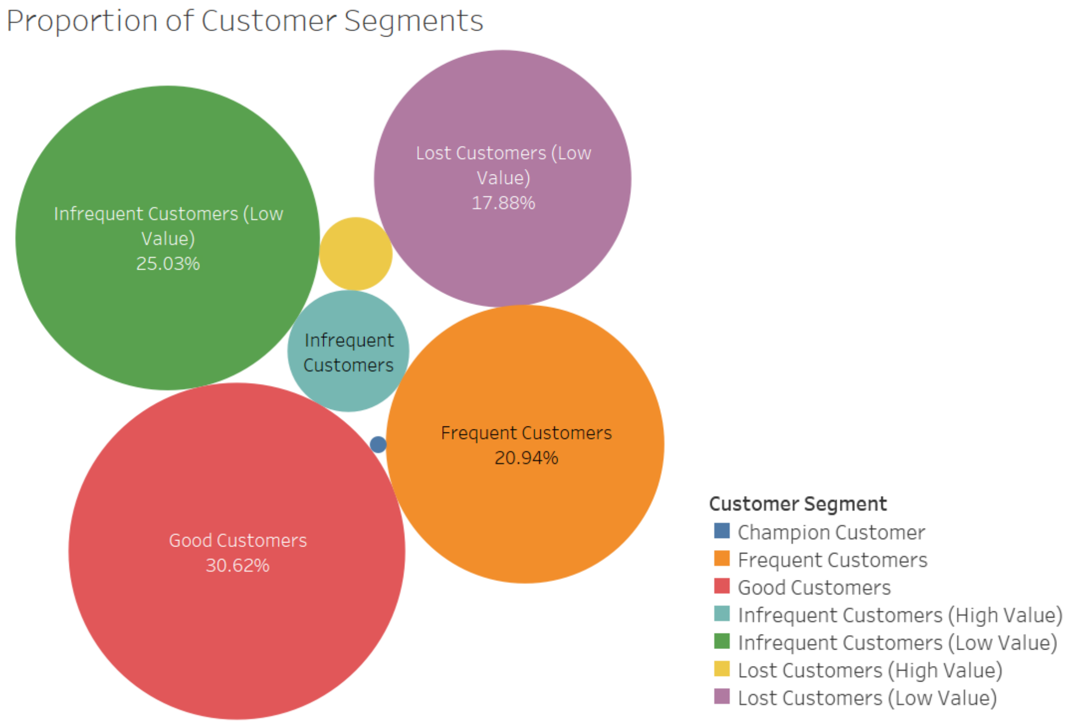
Sidenote: It was created using the same scoring methodology as above; i.e. evaluating the distribution of each column and dissecting the data by evaluating their ratio significance.

No.	Segment Name	Recency Score	Frequency Score	Monetary Score	Description
1.	Lost Customers (Low value)	1 to 2	1 to 5	1 to 2	These guys last transacted a long time ago; they didn't spend much though
2.	Lost Customers (High value)	1 to 2	1 to 5	3 to 5	These guys last transacted a long time ago and they spent a significant amount
3.	Frequent Customers	3 to 5	3 to 5	1 to 2	These guys buy frequently but don't spend that much

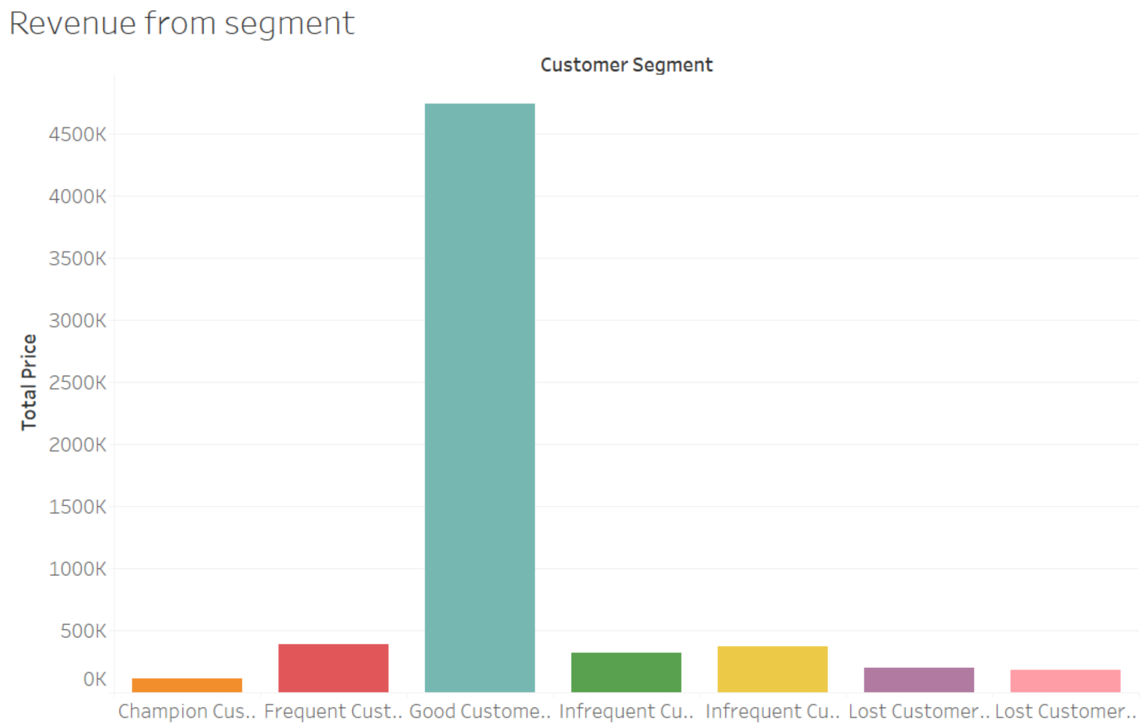
No.	Segment Name	Recency Score	Frequency Score	Monetary Score	Description
4.	Good Customers	3 to 5	3 to 5	3 to 5	These guys buy frequently and spend a good amount
5.	Infrequent Customers (Low Value)	3 to 5	1 to 2	1 to 2	These guys rarely make purchases and when they do, they don't spend much.
6.	Infrequent Customers (High Value)	3 to 5	1 to 2	3 to 5	These guys rarely make purchases but when they do, they spend a good amount.
7.	Champion Customers	5	5	5	Champion customer. They spent the most, frequent us the most and have been loyal.

3B. APPLY INSIGHTS

I set up an "if else" loop to capture the above framework and obtained the below result.



The bigger customer segments would require more investment in resources as they would encompass a bigger number of people.



The revenue generated from each segment is shown above.

The graph above clearly shows that “Good Customers” generated the most revenue for the organisation. This is generally in line with the Pareto principle that states, “for many events, roughly 80% of the effects come from 20% of the causes.”.

With this information, we can now devise relevant marketing strategies for each customer segment. For example:

- **Champion Customer**: Give Annual Gift Cards to show appreciation.
- **Frequent Customer**: Execute a survey to investigate how we could improve their customer experience.
- **Good Customer**: Examine purchasing data to identify purchase behavioral patterns.

Use Data to Know Your Customer

We have successfully shown a simple data-centric approach to Knowing Your Customer and it enables the E-Commerce business to apply further activities as the segmentation were business-applicable.

In addition, we can add more data types to find more segments. For example, if we included demographic data, we could determine the types of people within each customer segment, for example "Career High Flyer," "House Matters-In-Charge" and so on.
Insights gathered at this juncture will then re-inform the direction and possibilities of gathering more forms of data to gain further insights, resulting again in the "The Cycle of all Cycles."

Hopefully, we've shown how a simple data-centric method of Knowing your Customers can enable businesses to gain valuable insights that would feed even more strategic activities.

Data@Construct: Forge Better Decisions
We run weekly Data Science Experiments to help marketers use data to increase their Digital Marketing ROI and Effectiveness. Sometimes we'll do wacky stuff, sometimes we'll focus on common business and marketing problems, but always, we'll share our learnings here on our blog.
Please subscribe to our Email Updates for weekly updates and learnings (or just head over to our Big List of Data Experiments).
If you've got a marketing data problem that needs cracking, we'd love to hear from you.