Online Retail Customer Behavior Analysis and Customer Segmentation

Springboard Data Science Career Track Capstone 3

Yuan Yin • 08.25.2021

Introduction

- **Dataset**:

  Two-year transactional data of an online **gift** retail, many customers of which are **wholesalers**

- **Goals of This Project**
  - To **analyze customer behavior** to help the company know the customers well
  - To **cluster customers** for better designing targeted marketing campaigns
  - To **A/B test** promotional campaigns to increase number of purchases

Data Wrangling

Main Steps:

- Concatenate two datasets and remove duplicates
- Keep UK-based transactions only
- Remove transactions with 0 or negative price(0.2%)
- Remove transactions with negative quantity(0.2%)

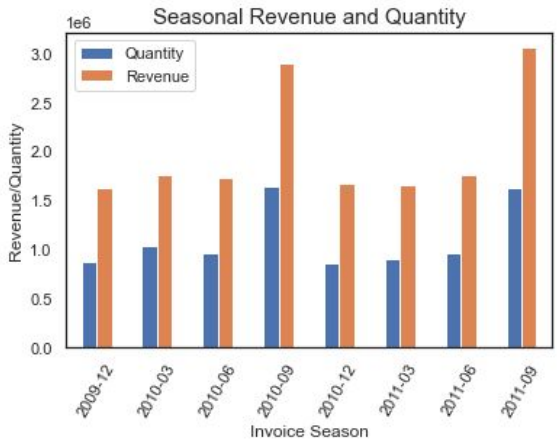
Key Performance Indicators

- Monthly/Seasonal Revenue Analysis
- Hypothesis Testing of Monthly Revenue

Monthly/Seasonal Revenue Analysis

- **Revenue and Sales Volume by Month**

- **Revenue and Sales Volume by Season***



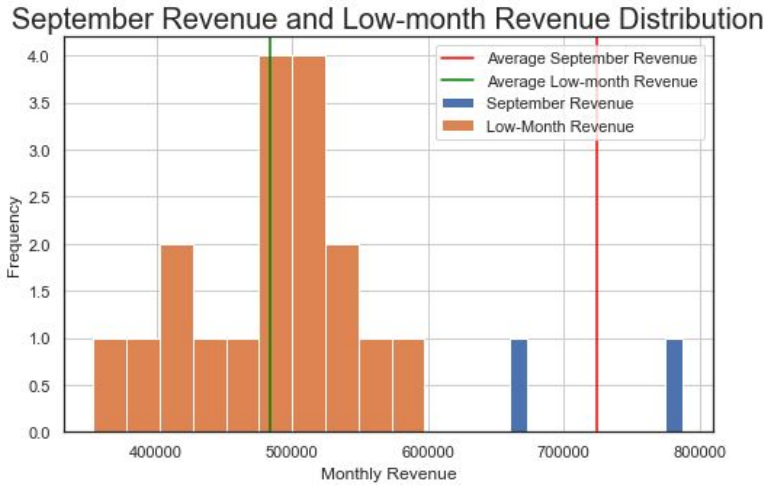
Season*: Dec-Feb, Mar-May, Jun-Aug, Sep-Nov

Hypothesis Testing of September Revenue

- **Null Hypothesis H_0:** the observed difference of monthly revenue between September and low-revenue months, i.e.,the other nine months except September, October, and November, is due to chance.

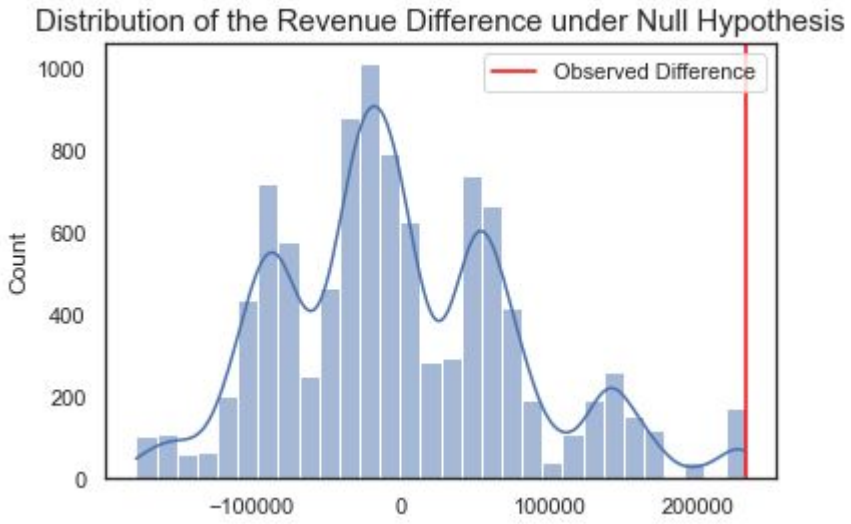
- **Alternative Hypothesis H_1:** the observed difference of monthly revenue between September and low-revenue months is not due to chance, in other words, due to pre-holiday purchase.

- **Method:** Permutation Test



Hypothesis Testing of September Revenue

- **Significant level:** 0.05
- **P Value:** 0.0014 (statistically significant)
- **Conclusion:** Reject null hypothesis



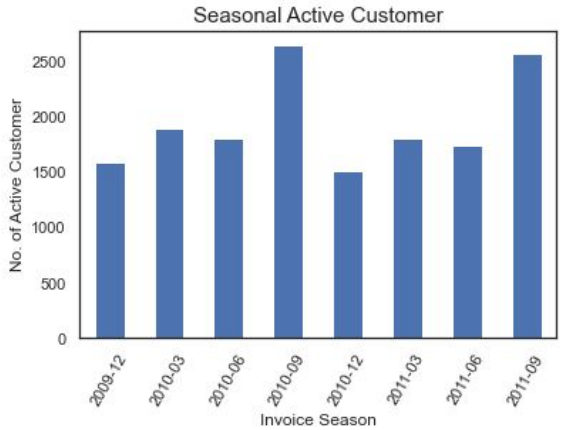
Customer Behavior Analysis

- Monthly/Seasonal Active Customer Analysis
- Time Cohort Analysis
- RFMT Analysis

Monthly/Seasonal Customers Analysis

- **Number of Active Customers by Month**

- **Number of Active Customers by Season**



Cohort Analysis

- **What is Cohort Analysis**

  Analyse data based on grouped customers(Cohort) rather than looking at the data as one unit.

- **Why Cohort Analysis**

  - Develop targeted marketing strategies
  - Provide insights on metrics across the customer lifecycle

- **Types of Cohort**

  - **Time Cohort**
  - Behavior Cohort
  - Size Cohorts

Time Cohort Analysis

- **Time Cohort**:

  customers who signed up for a service or purchased a product for the first time during a particular time frame

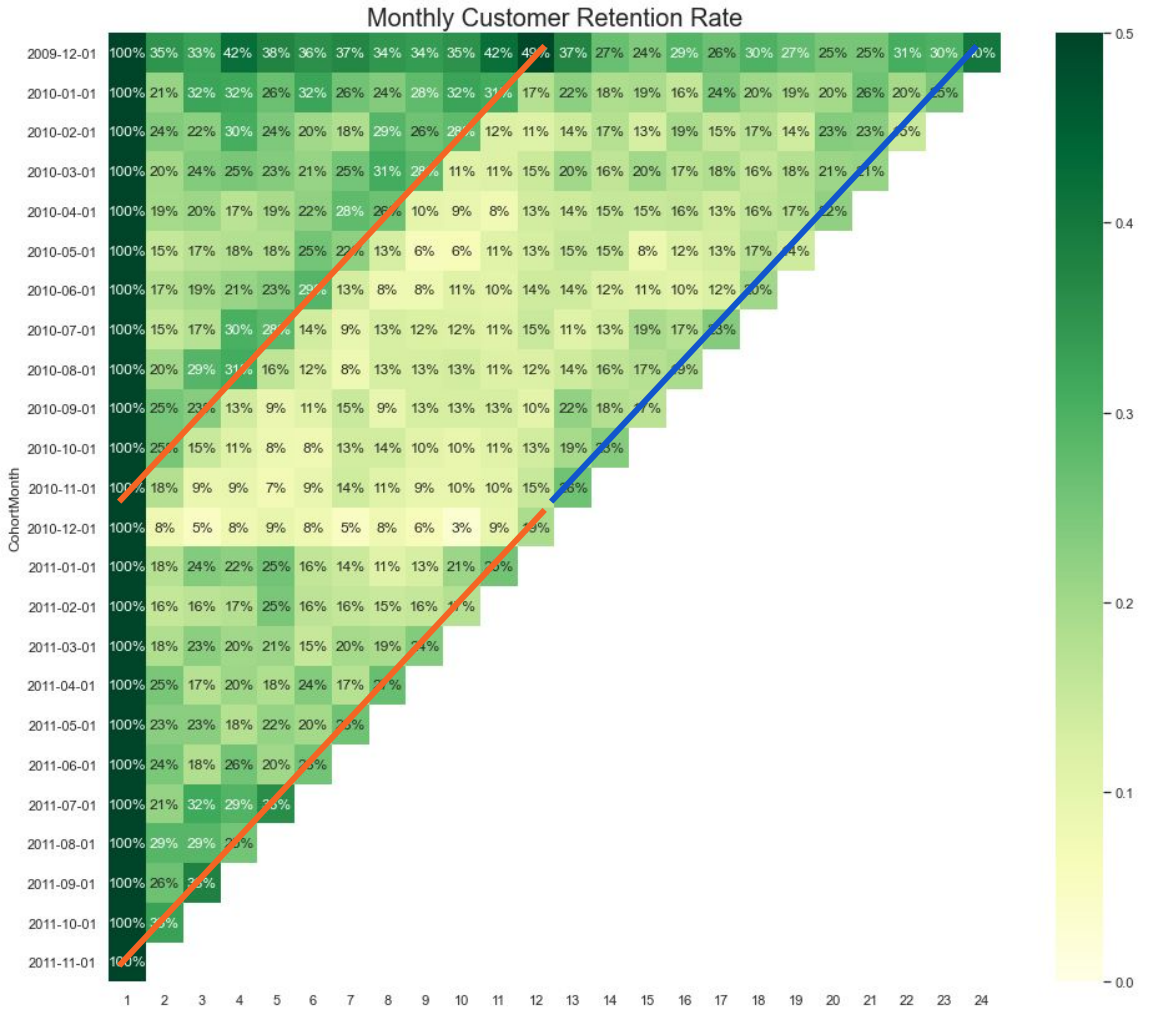
- The time may be monthly, quarterly, even daily

—

- **Retention Rate of Monthly Cohort**

Retention Rate: the ratio of how many customers came back in the subsequent months.

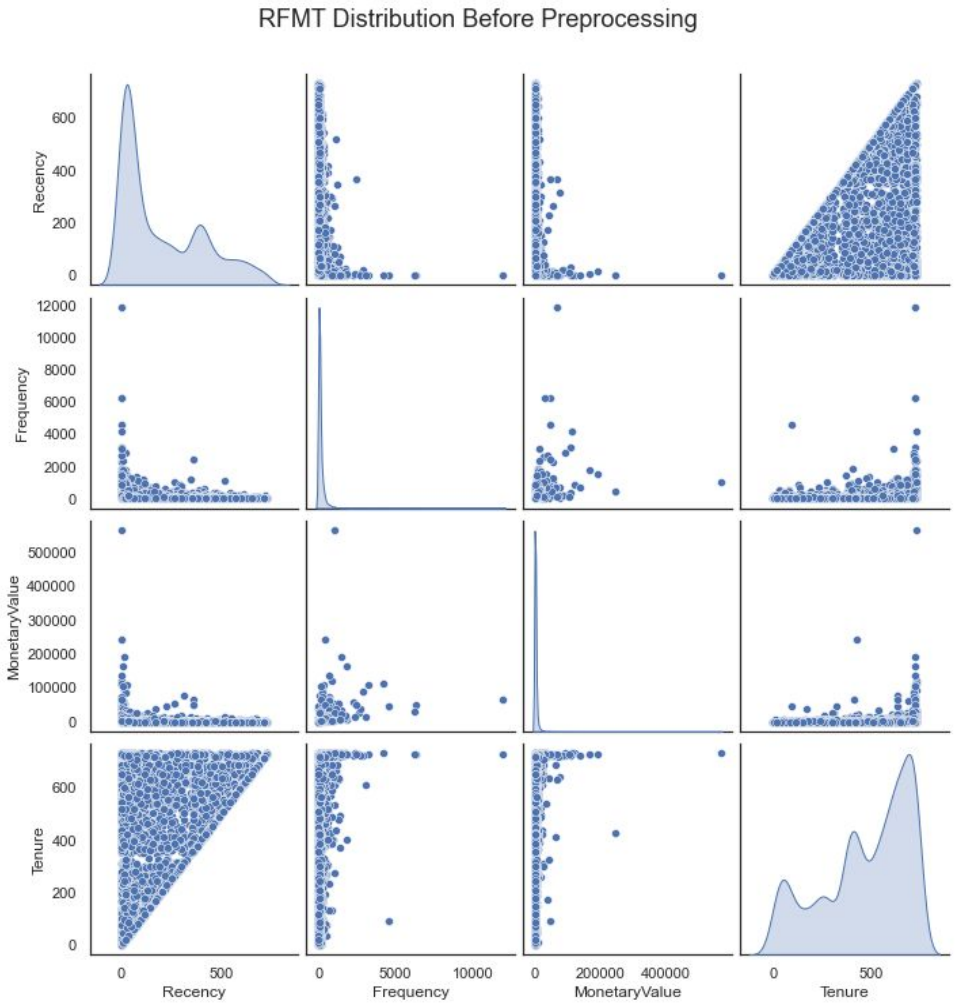
$$Retention\ Rate = \frac{Number\ of\ Active\ Customer}{Cohort\ Size}$$



RFM(T)

- **What is RFMT**
  - **R: Recency**
  - **F: Frequency**
  - **M: Monetary Value**
  - **T: Tenure**
- **RFM(T) Segmentation**

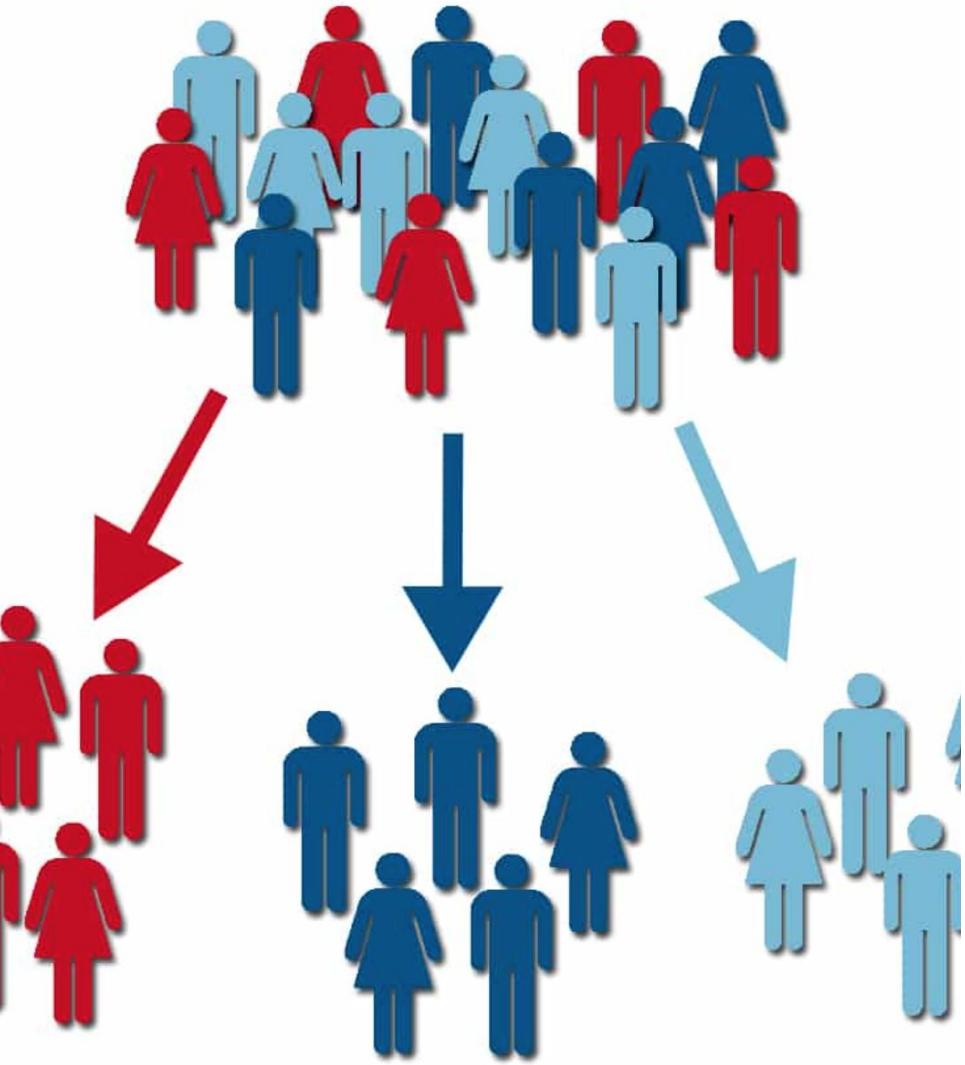
  Break customers into groups based on the percentile value of each metrics.



Customer Segmentation

— With KMeans Algorithm

- Metric Selection
- Data Preprocessing
- Number of Clusters Optimizing
- Customer Profiling



- **Customer segmentation** is the practice of dividing customers into groups based on the similarity of characteristics.
- Each group or segment is related to **a significant customer profile** so companies can design targeted marketing campaigns.

Metric Selection

- **Recency, Frequency, and Monetary Value (RFM)**

➔ Recency: **how recent** was each customer's last purchase
➔ Frequency: **how many** purchases the customer has done
➔ Monetary Value: **how much** has the customer spent

- **Recency, Monetary Value, and Tenure (RMT)**

➔ Tenure: **how long** the customer has been with the company since their first transaction

Data Pre-processing

➔ **Why Data Preprocessing**

KMeans Assumption:

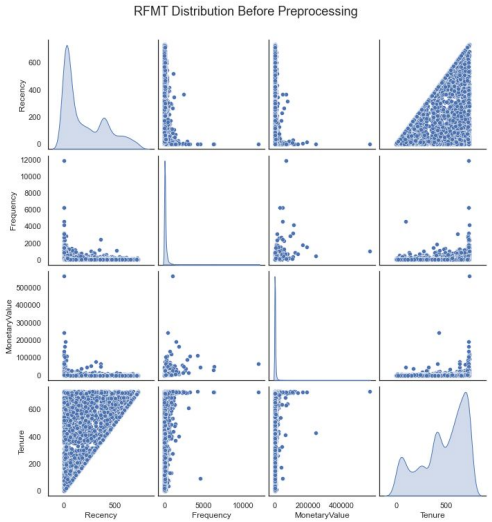
- Numeric features
- Symmetrical features
- Same mean values and same variance

➔ **How to Preprocess data**

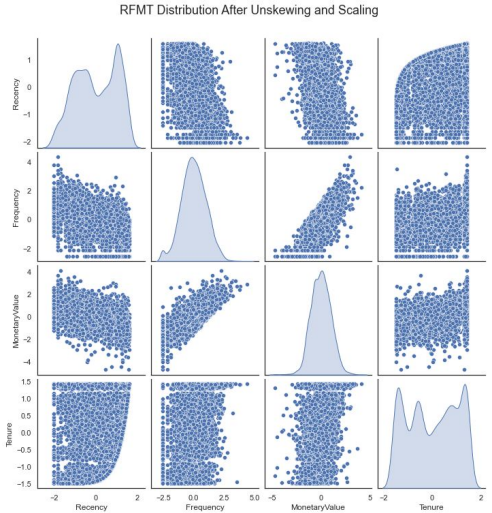
- Unskew Feature
- Scale Feature

Data Pre-processing

- **RFMT Distribution Before Preprocessing**

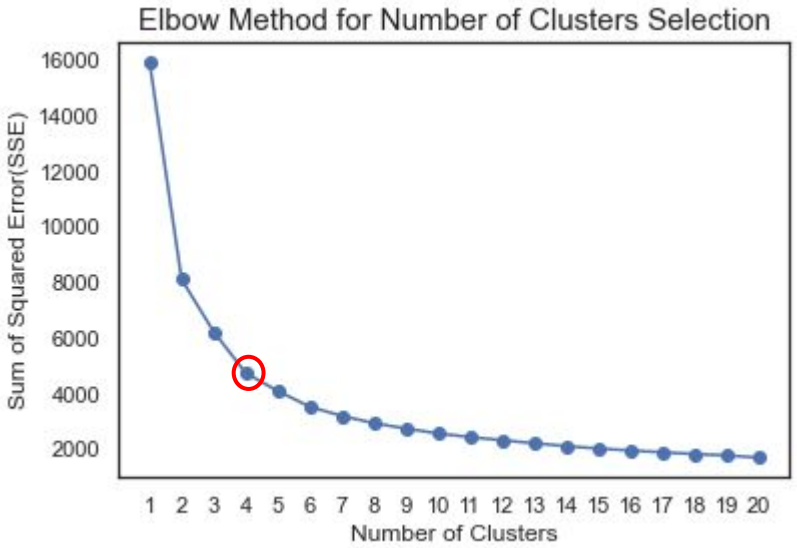


- **RFMT Distribution After Preprocessing**

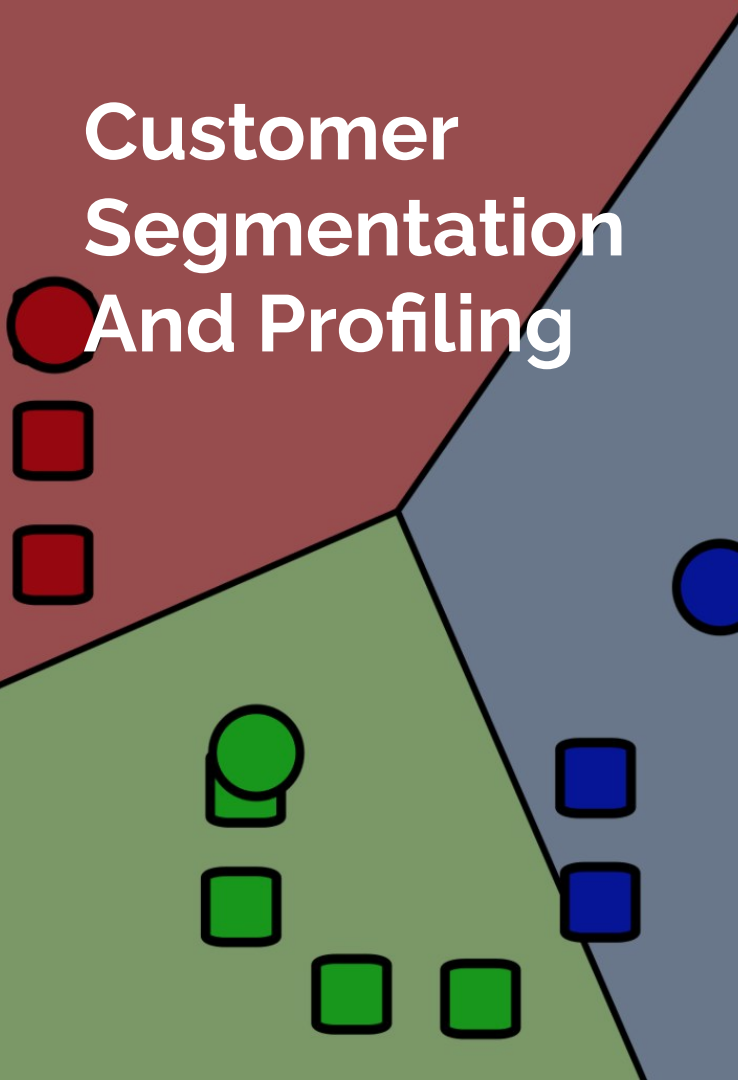


Number of Clusters Optimizing

- **How to Predefine the Number of Clusters, i.e., k**



Customer Segmentation And Profiling



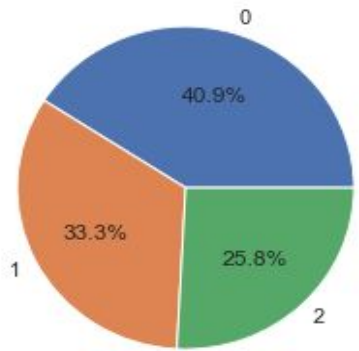
- In general, a well-formed set of clusters have two highlights:
  - ✓ good cohesion
  - ✓ good separation
- **Business Insights**

Customer Segmentation And Profiling

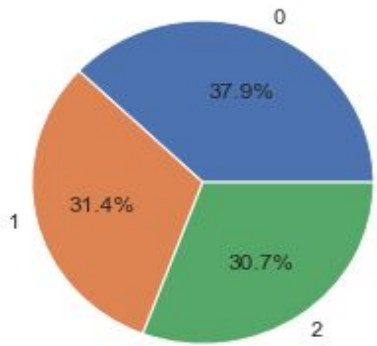
- **How to Identify Customer Profiles**
  - Summary Statistics
  - Relative Importance
  - Snake Plot
  - 2D kde Plot
  - 3D Scatter Plot

Basic Solution: 3-Segment Model Based on RFM/RMT

- Segment Size Ratio of RFM Model

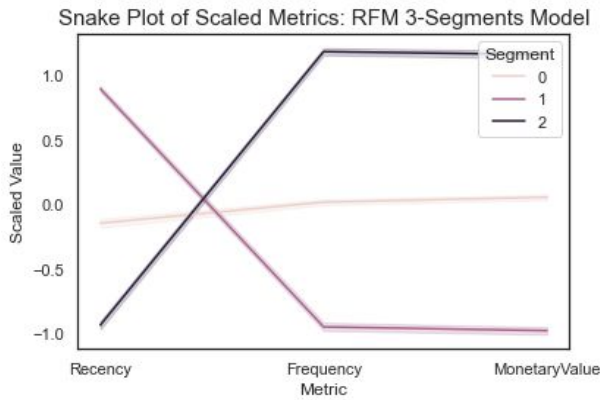


- Segment Size Ratio of RMT Model

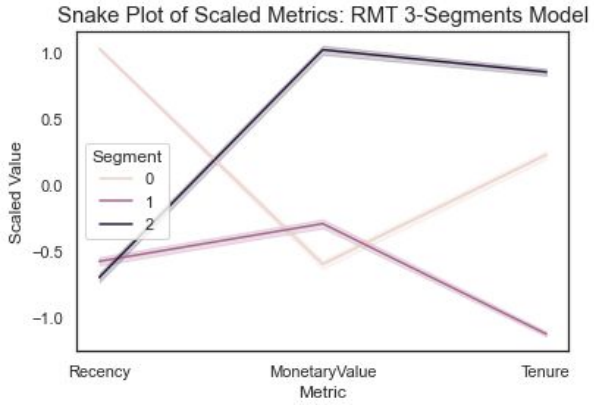


Basic Solution: 3-Segment Model Based on RFM/RMT

- Snake Plot of RFM Model

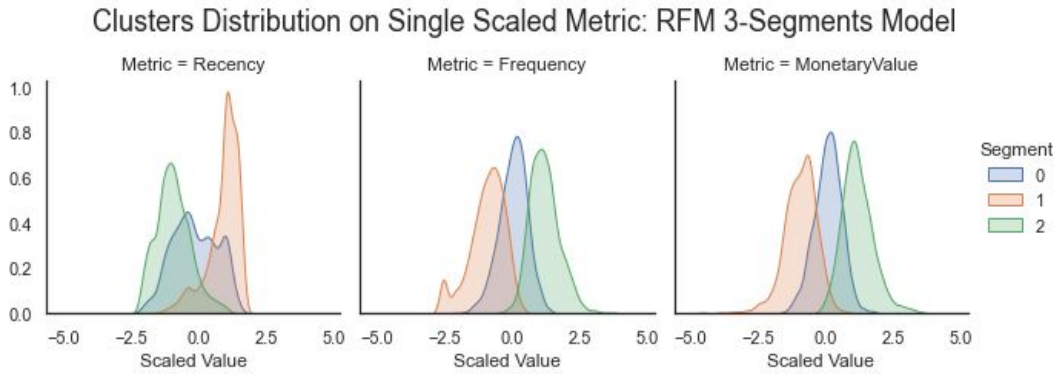


- Snake Plot of RMT Model

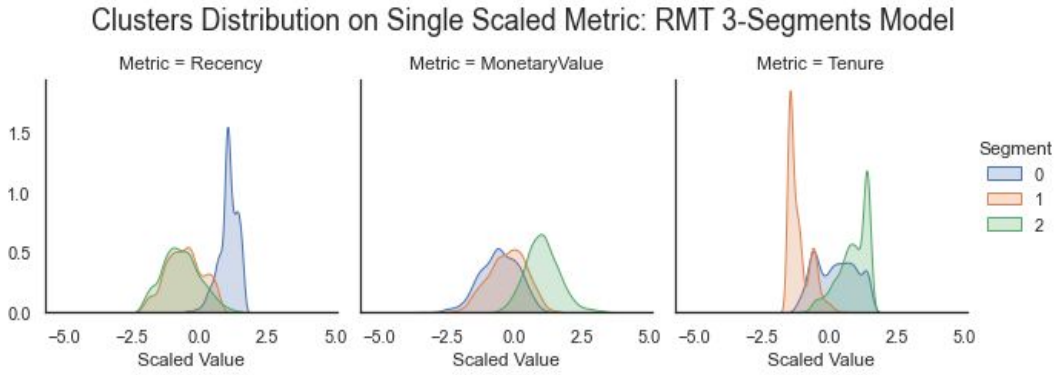


Basic Solution: 3-Segment Model Based on RFM/RMT

- **2D kde Plot of RFM Model**

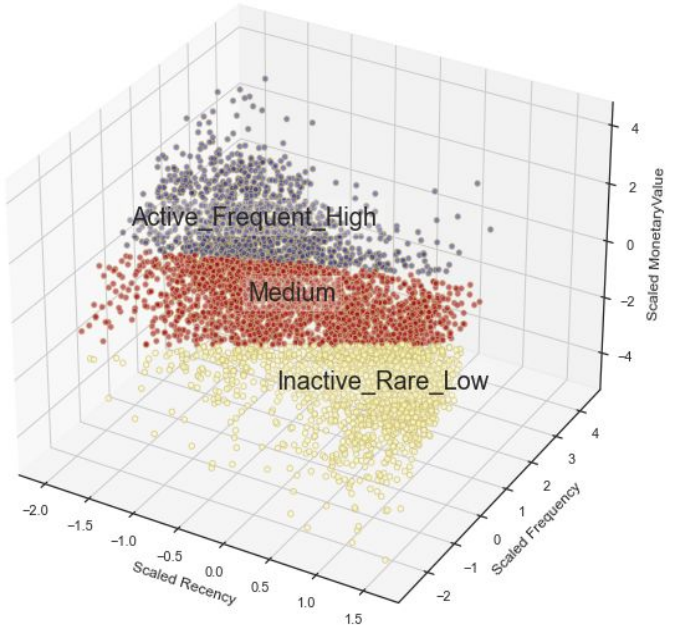


- **2D kde Plot of RMT Model**

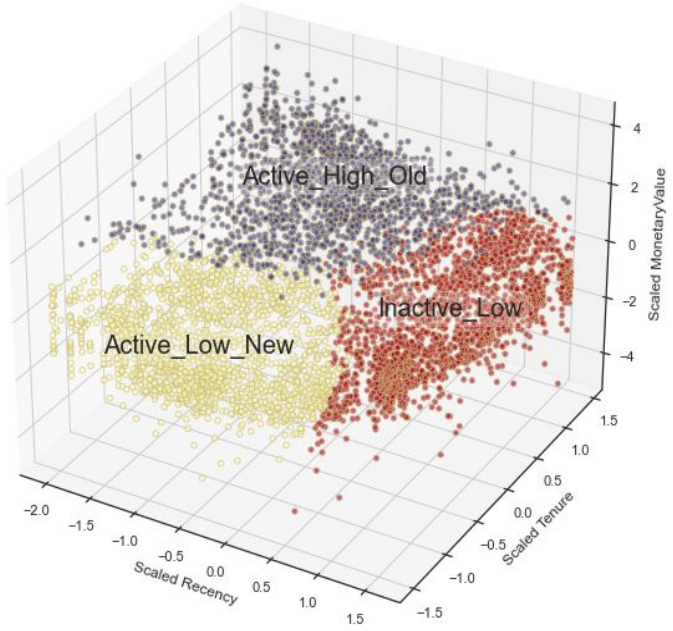


Basic Solution: 3-Segment Model Based on RFM/RMT

- **Scatter Plot of RFM Model**

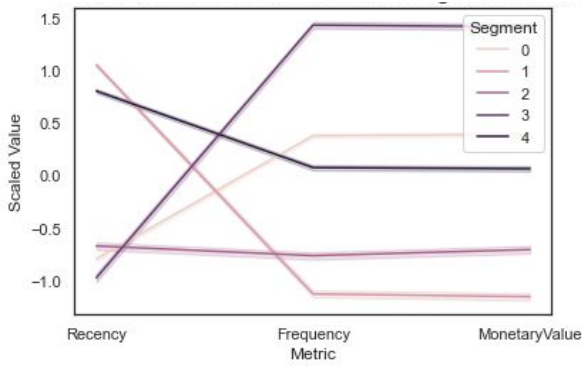


- **Scatter Plot of RFM Model**

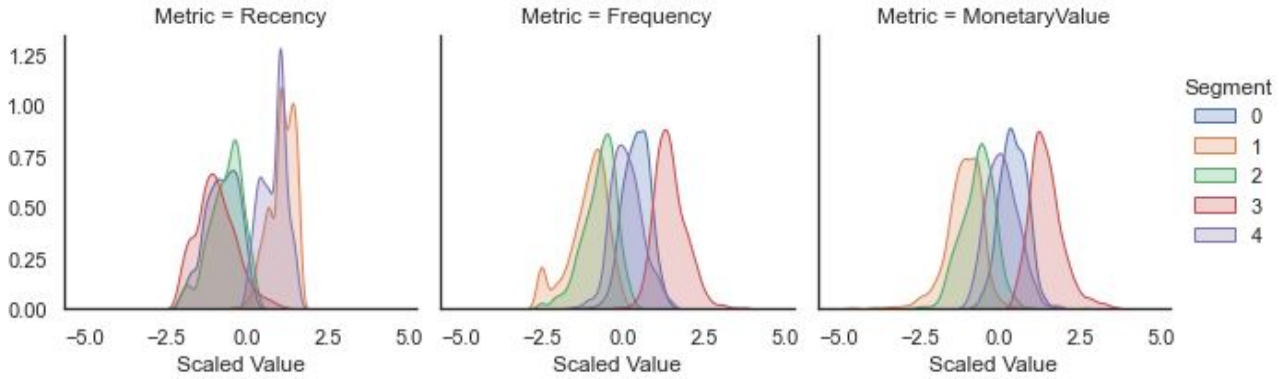


Improved Solution: 5-Segment RFM Model

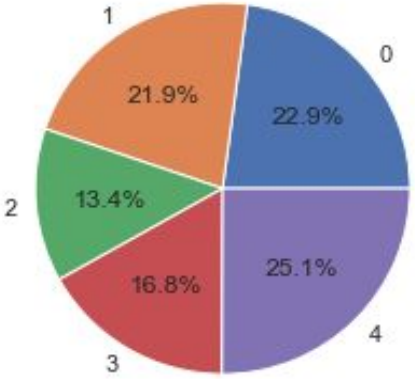
- **Snake Plot**



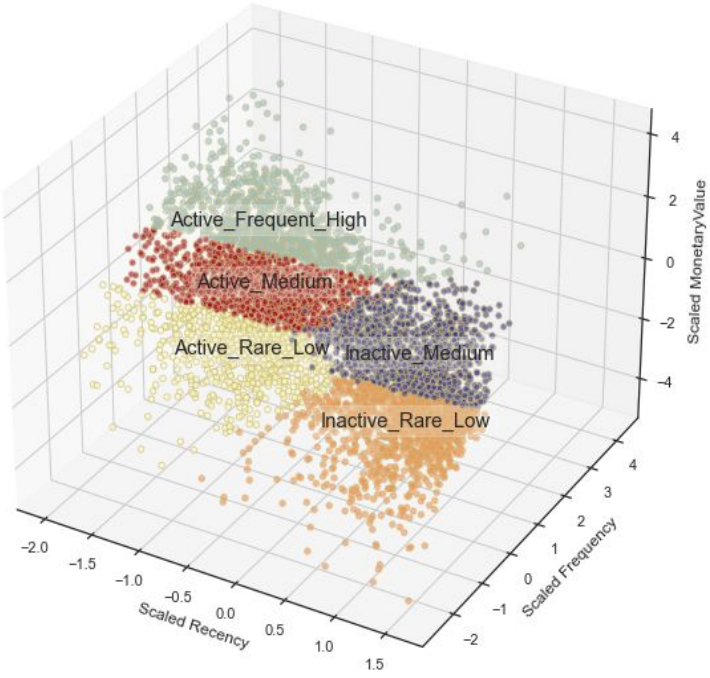
- **2D kde Plot**



Improved Solution: 5-Segment RFM Model



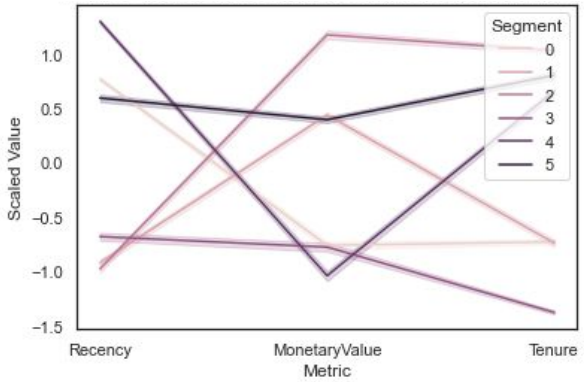
- **Segment Size Ratio**



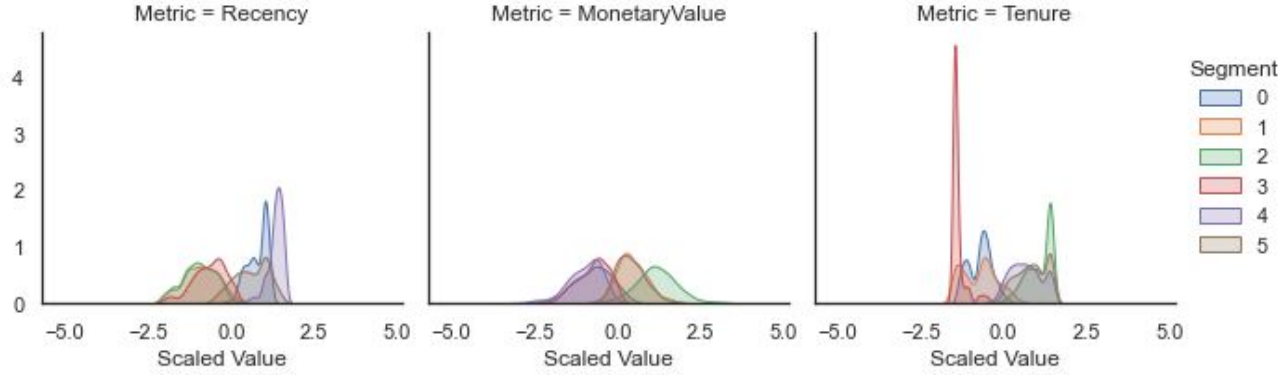
- **3D Scatter Plot**

Improved Solution: 6-Segment RMT Model

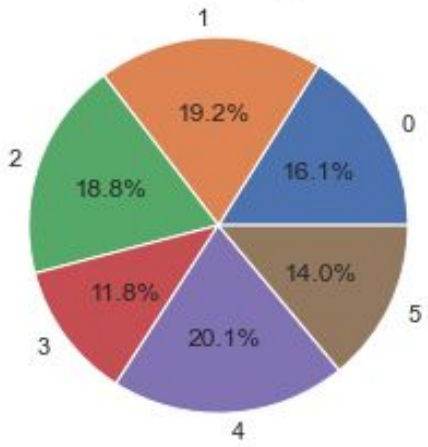
- **Snake Plot**



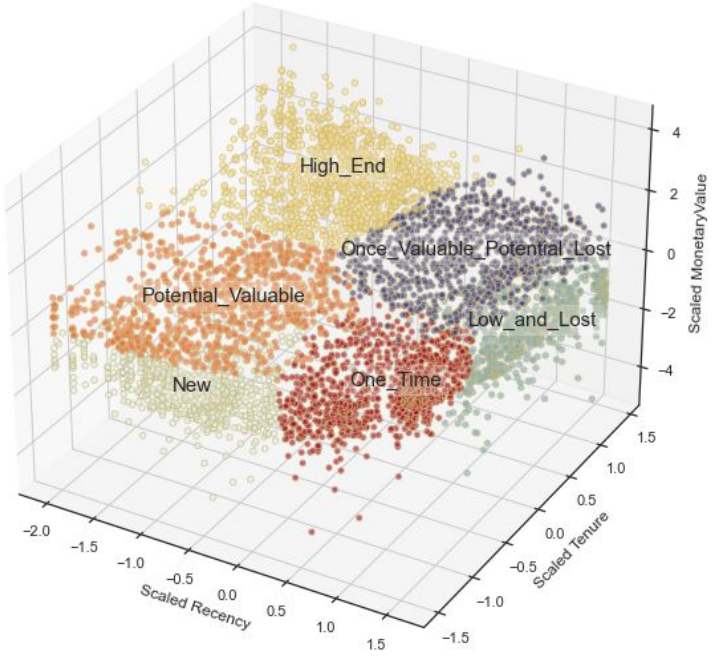
- **2D kde Plot**



Improved Solution: 6-Segment RMT Model



- **Segment Size Ratio**



- **3D Scatter Plot**

Improved Solution: 6-Segment RMT Model

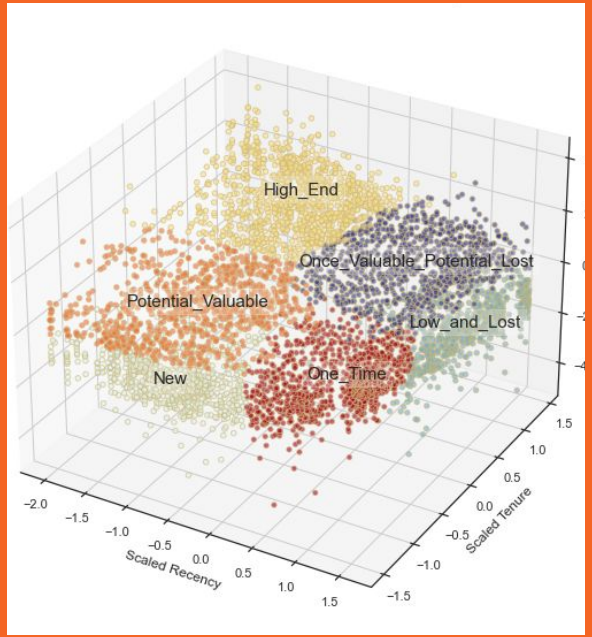
- **Summary Statistics**

	Recency					MonetaryValue			
	mean	median	min	max	std	mean	median	min	max
Segment									
0	315.0	358	37	576	111.0	528.0	432.0	96.0	4512.0
1	31.0	18	1	521	41.0	9286.0	4600.0	462.0	564101.0
2	33.0	23	1	227	32.0	1604.0	1126.0	211.0	44534.0
3	90.0	58	1	326	83.0	262.0	212.0	3.0	1862.0
4	259.0	231	6	730	172.0	1787.0	1276.0	96.0	77347.0
5	538.0	557	42	730	135.0	236.0	195.0	3.0	1437.0

	Tenure					Size
	std	mean	median	min	max	std	
Segment							
0	387.0	388.0	401	105	587	97.0	857
1	24047.0	663.0	686	93	730	79.0	1020
2	2354.0	285.0	297	1	659	170.0	998
3	196.0	133.0	84	1	639	122.0	628
4	2778.0	651.0	658	403	730	64.0	1066
5	169.0	591.0	603	350	730	98.0	744

6-Segment RMT Model Business Insights

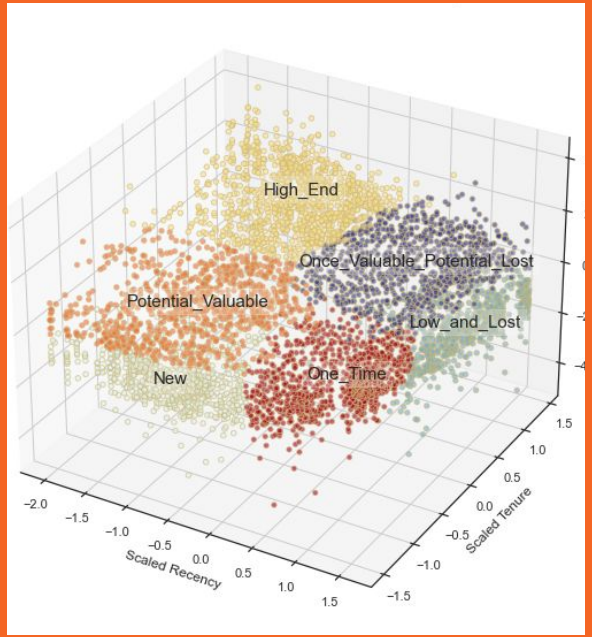
➔ Segment 0: **One_Time.** The customers within this segment purchased products about one year ago, for the first and last time.



6-Segment RMT Model Business Insights

➔ Segment 1: **Potential_Valuable.** The customers within this segment are relatively new but have generated considerable revenue. Besides, this segment is pretty active.

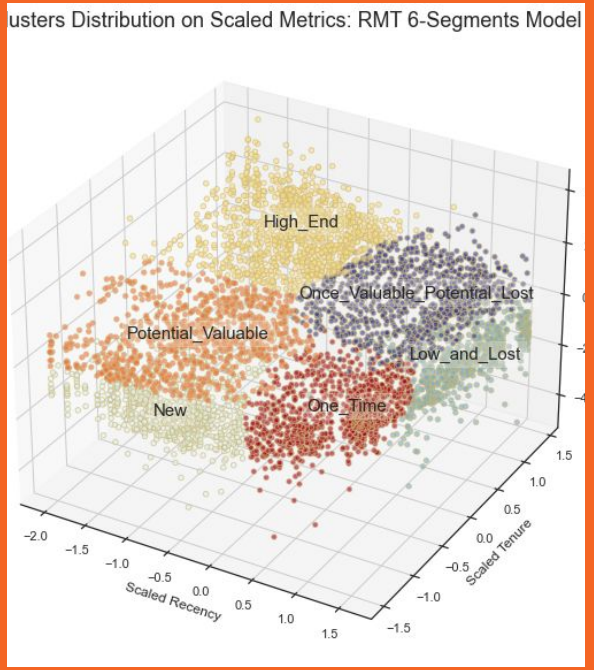
This segment is worthy to highlight



6-Segment RMT Model Business Insights

➔ Segment 2: **High_End**. The customers within this segment are active, loyal, and have generated much revenue.

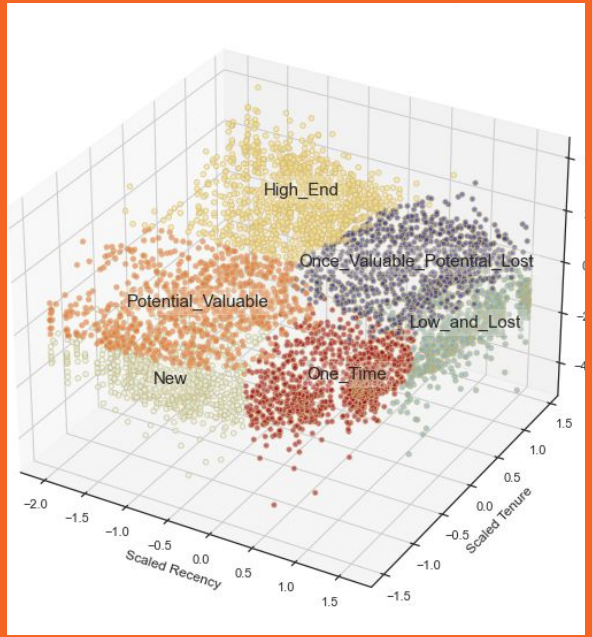
The most important segment for the company



6-Segment RMT Model Business Insights

➔ Segment 3: **New**. The customers within this segment are new to the company. The median value of Tenure is only 64 days, which is much lower than the other segments

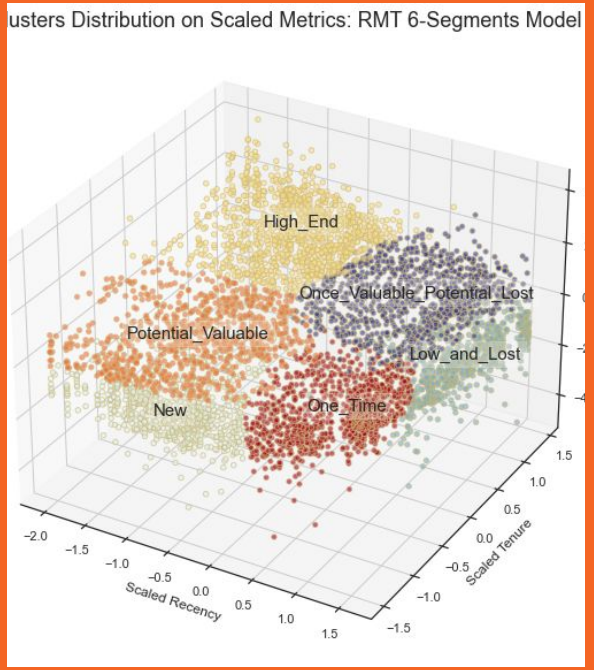
Based on the Time Cohort Analysis, the customers in this segment are easy to churn



6-Segment RMT Model Business Insights

➔ Segment 4: **Low_and_Lost.** The customers in this segment generated very low revenue since most of them have not purchased anything for more than one year and a half.

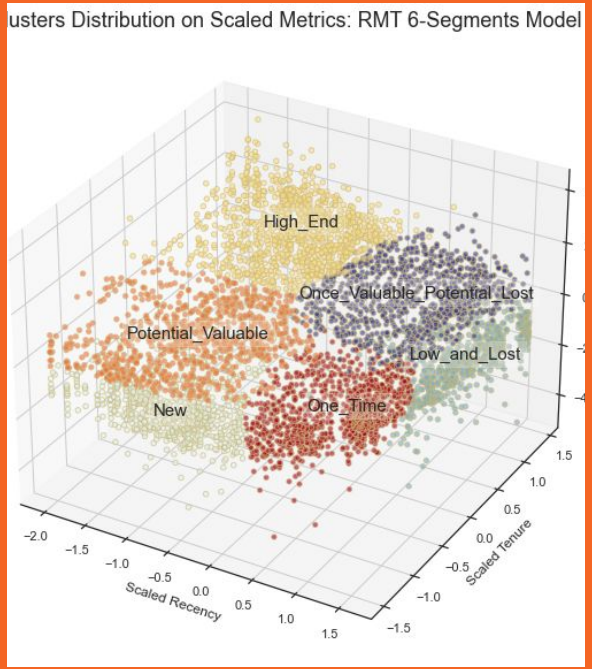
No need to pay close attention to this segment.



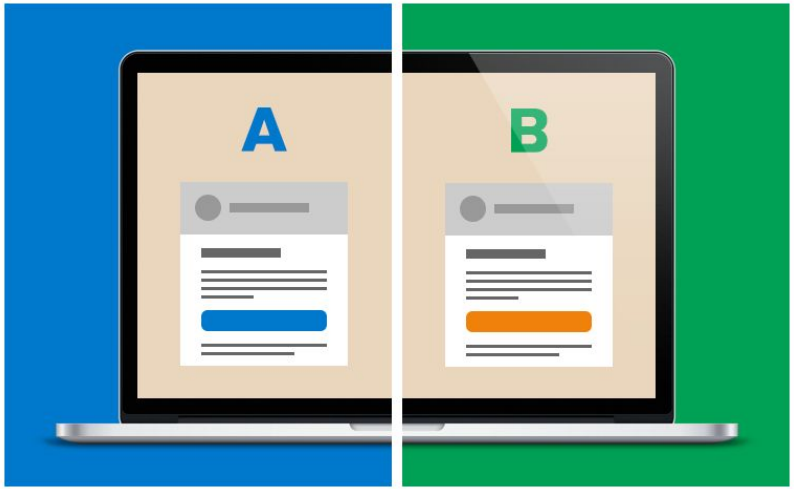
6-Segment RMT Model Business Insights

- Segment 5: Once_Valuable_Potential_Lost. The customers within this segment did not purchase recently but they have generated pretty high revenue.

The company may take special actions to prevent these customers from being really lost.



A/B Testing



- Design
- Implementation
- Analysis

A/B Testing

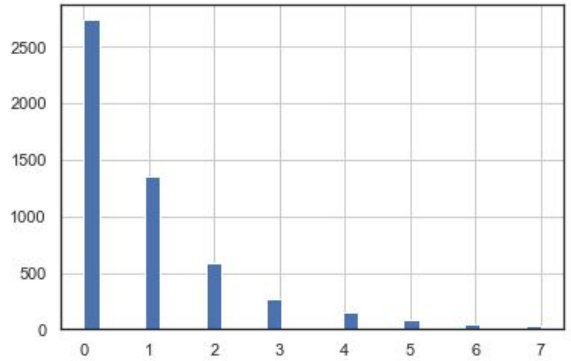
- **Problem Statement**

  How to raise the number of purchases by 20% in the high-revenue months through promotional campaigns?

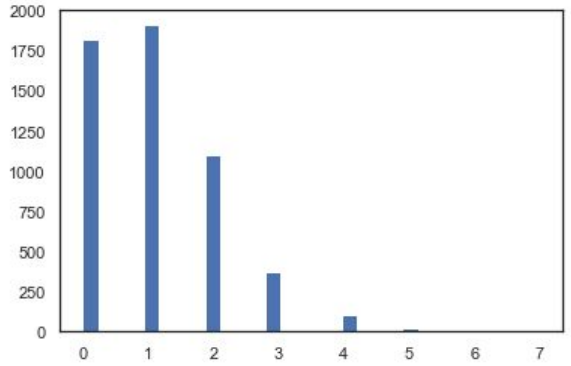
- **Metrics**
  - **Revenue/Profit Margin:**
    The most important KPI
  - **Purchase:**
    Related to revenue and easy to measure

Number of Purchases by Customer

- Baseline

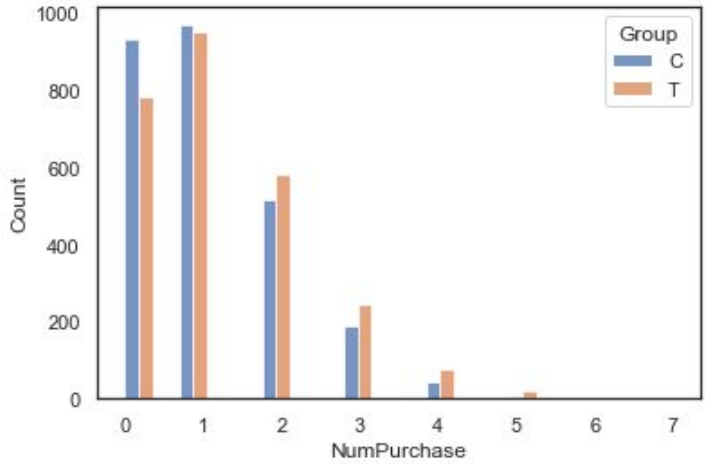


- Poisson Distribution



Dummy A/B Testing

Group	count	mean	std	min	25%	50%	75%	max
C	2656.0	1.045181	1.010785	0.0	0.0	1.0	2.0	6.0
T	2657.0	1.229959	1.111362	0.0	0.0	1.0	2.0	7.0



Hypothesis Testing

- **Null Hypothesis H_0**:

  There is no significant difference between the control group and the treatment group, i.e., the lift of the average number of purchase is due to chance.

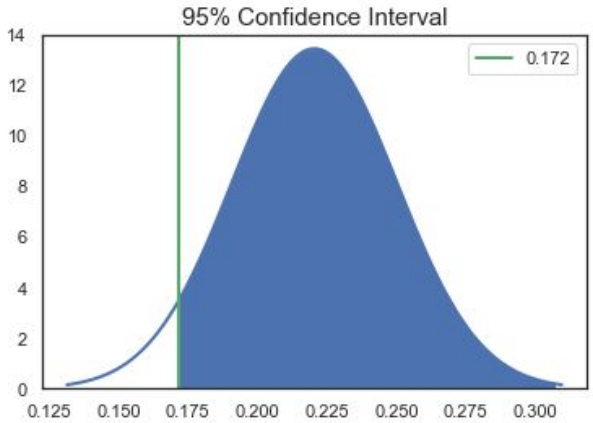
- **Alternative Hypothesis H_1**:

  There is a significant improvement of number of purchase in treatment group compared with control group.

- **Significant Level:** 0.5

- **Method:** t-test
- **P value**: 0
- **95% Confidence Interval**:

  There is a 95% chance that the increase in the average numbers of purchase is greater than 0.172 through promotional campaigns



Future Work

- In this project, we segment customers using KMeans. We can try to identify meaningful segments using other clustering algorithms such as Non-negative Matrix Factorization (NMF) or Hierarchical Clustering
- Except for customer segmentation, we can do market basket analysis, product segmentation, Customer Lifetime Value prediction, next transaction prediction, customer churn prediction, and so on.

Permutation Test

	InvoiceDate	Revenue	Sep	Permutational
0	2009-12-31	715672.58	N	550370.41
1	2010-01-31	476816.63	N	570654.58
2	2010-02-28	443363.23	N	544839.20
3	2010-03-31	665085.04	N	874645.60
4	2010-04-30	553739.84	N	417288.59
5	2010-05-31	550370.41	N	704092.24
6	2010-06-30	599382.08	N	723023.80
7	2010-07-31	548533.93	N	443363.23
8	2010-08-31	577565.06	N	553739.84
9	2010-09-30	723023.80	Y	476816.63
10	2010-12-31	704092.24	N	574652.33
11	2011-01-31	544839.20	N	604242.08
12	2011-02-28	417288.59	N	577565.06
13	2011-03-31	570654.58	N	617446.61
14	2011-04-30	464783.50	N	548533.93
15	2011-05-31	617446.61	N	665085.04
16	2011-06-30	604242.08	N	577412.95
17	2011-07-31	574652.33	N	715672.58
18	2011-08-31	577412.95	N	599382.08
19	2011-09-30	874645.60	Y	464783.50