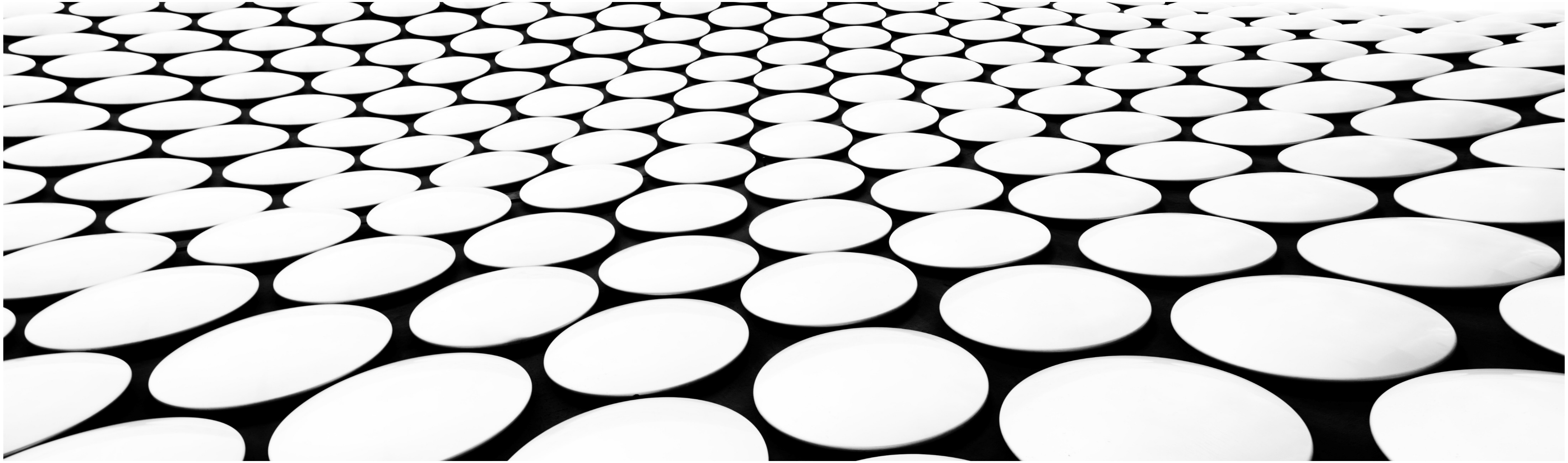
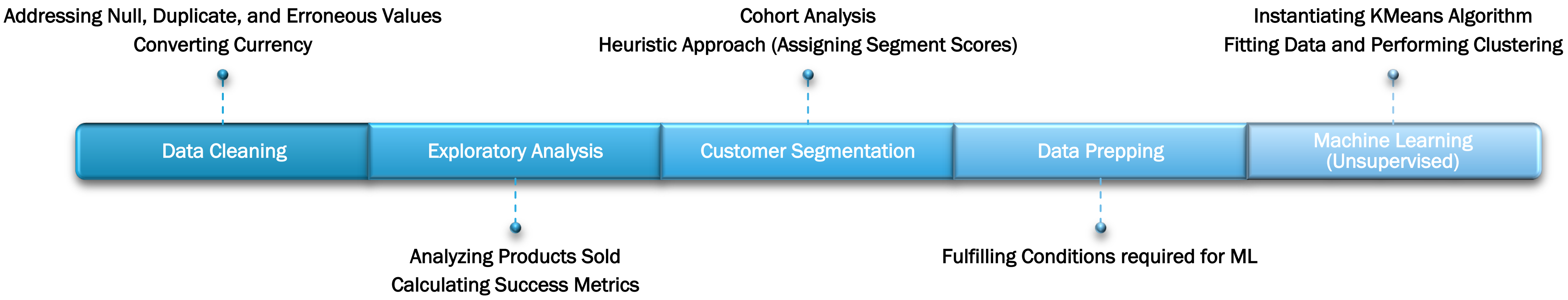
CUSTOMER SEGMENTATION FOR ONLINE RETAIL

CAPSTONE PROJECT BY DEA WANG



PROJECT OUTLINE

Sector: Retail (E-commerce/Online Retail)
Objective: Find the most profitable customer segment to target for marketing campaign/increase revenue



Addressing Null, Duplicate, and Erroneous Values
Converting Currency

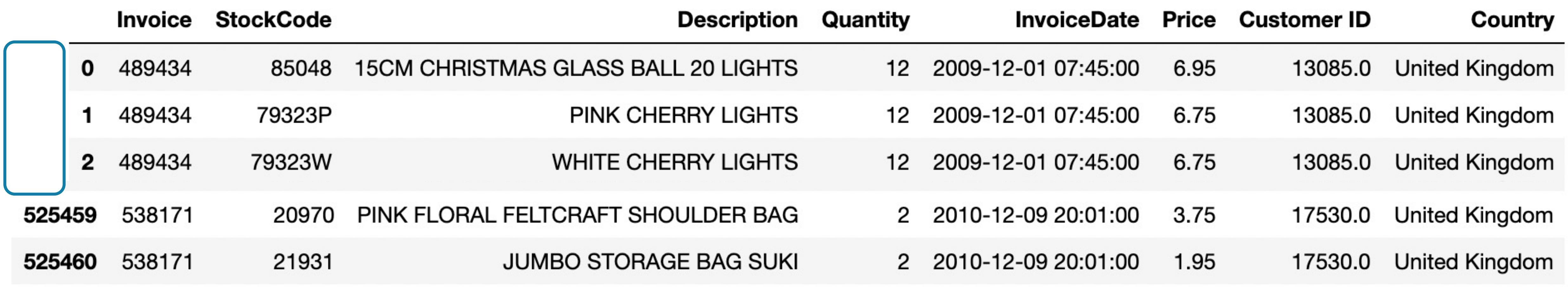
Cohort Analysis
Heuristic Approach (Assigning Segment Scores)

Instantiating KMeans Algorithm
Fitting Data and Performing Clustering

Analyzing Products Sold
Calculating Success Metrics

Fulfilling Conditions required for ML

Python Libraries:
numpy, pandas, datetime, os, exchangerateapi
matplotlib, seaborn, mpl_toolkits, wordcloud
sklearn: StandardScaler, KMeans, KElbowVisualizer



	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
0	489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	2009-12-01 07:45:00	6.95	13085.0	United Kingdom
1	489434	79323P	PINK CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
2	489434	79323W	WHITE CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
525459	538171	20970	PINK FLORAL FELTCRAFT SHOULDER BAG	2	2010-12-09 20:01:00	3.75	17530.0	United Kingdom
525460	538171	21931	JUMBO STORAGE BAG SUKI	2	2010-12-09 20:01:00	1.95	17530.0	United Kingdom

ONLINE RETAIL II DATESET FROM UCI ML REPO

HTTPS://ARCHIVE.ICS.UCI.EDU/ML/DATASETS/ONLINE+RETAIL+II

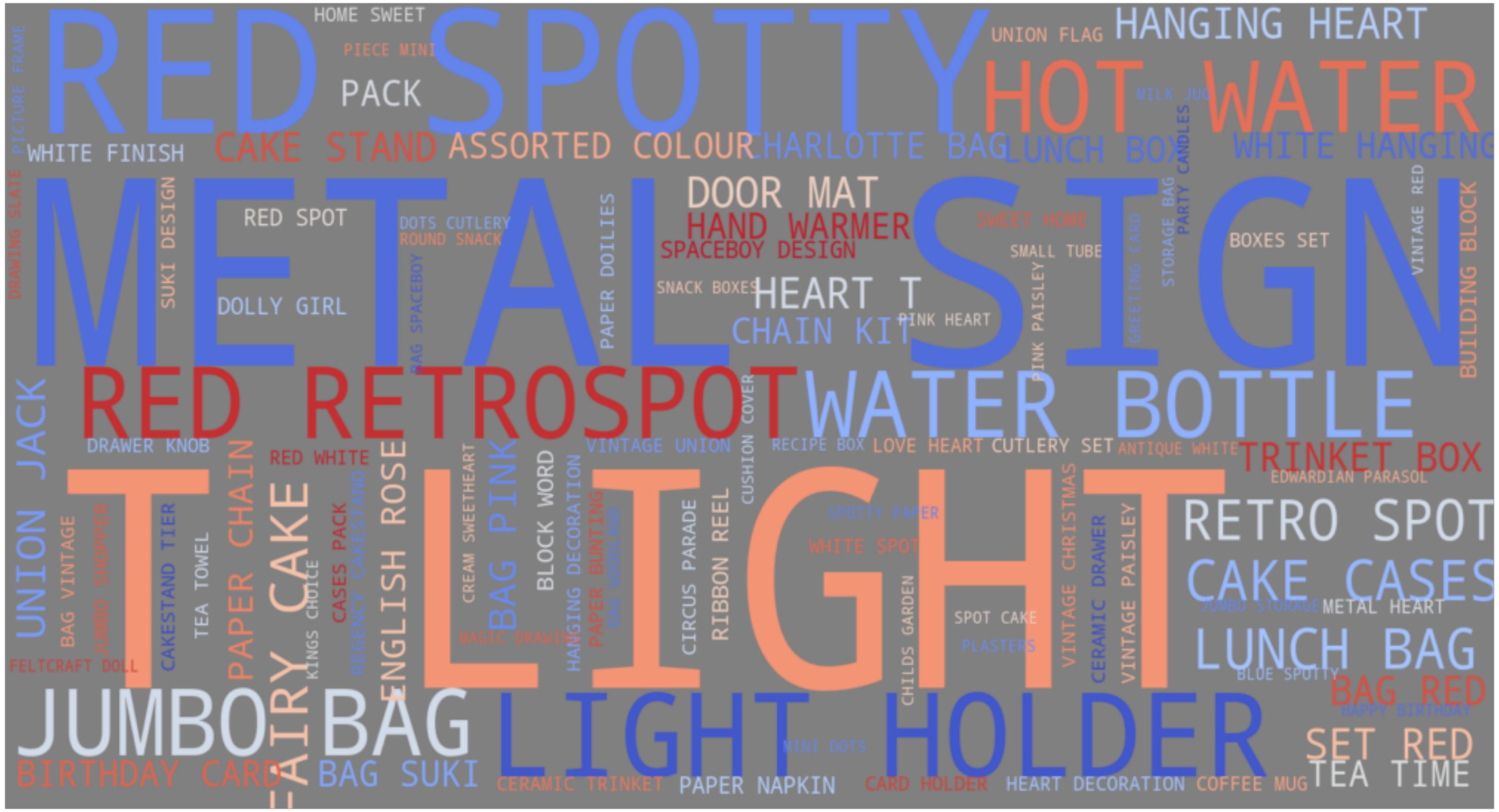
- 525,461 Invoice Listings
- 8 Features
- From 2009-12-01 to 2010-12-09
- File size = 25 MB

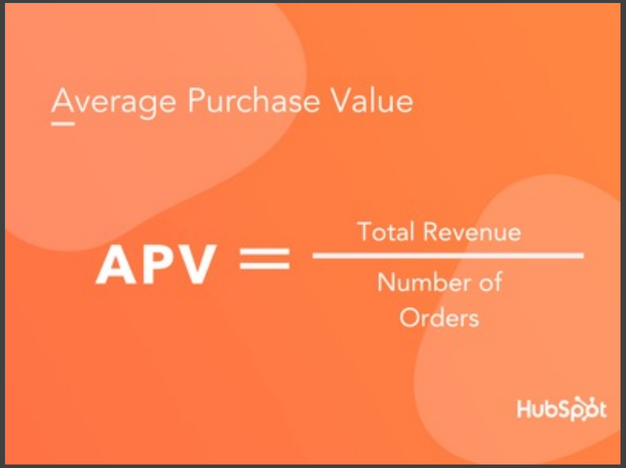
DATA CLEANING

- Fill or drop **Null** values where applicable
- Remove **Duplicate** values
- Remove **Erroneous** data: negative or zero values in Quantity and Price
- Convert **Price** from GBP in CAD

PRODUCT DESCRIPTION

- Common Products
- Description Word Cloud

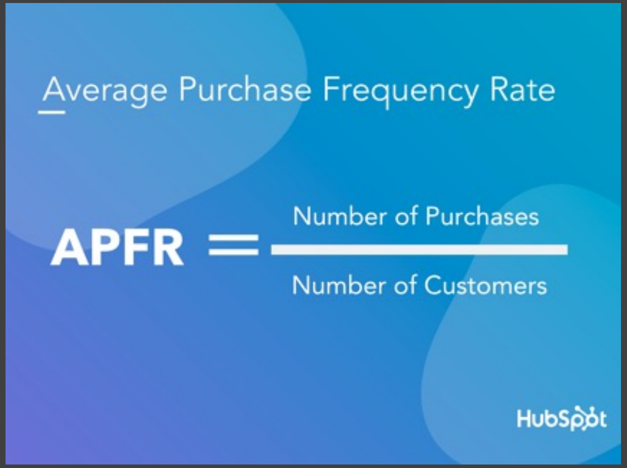




Average Purchase Value

$$APV = \frac{\text{Total Revenue}}{\text{Number of Orders}}$$

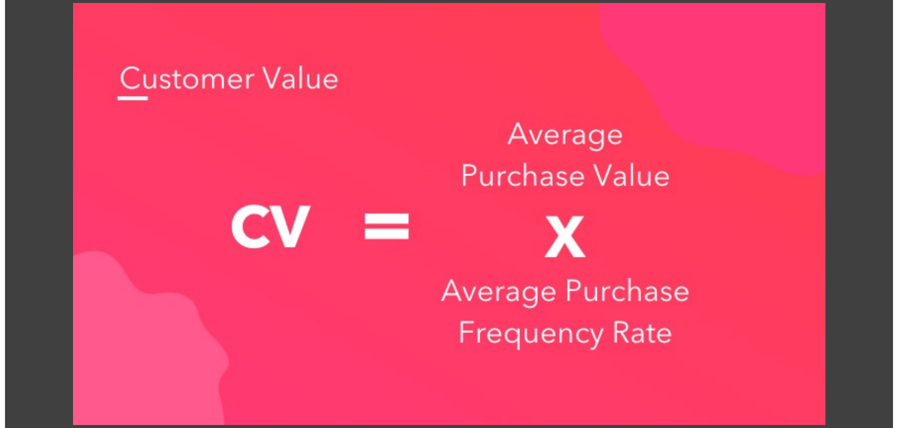
HubSpot



Average Purchase Frequency Rate

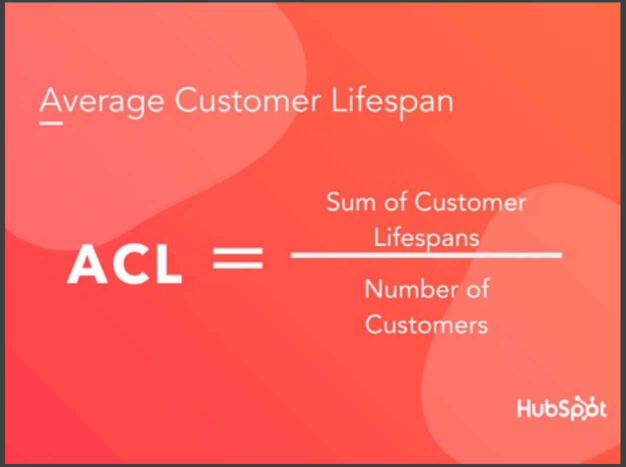
$$APFR = \frac{\text{Number of Purchases}}{\text{Number of Customers}}$$

HubSpot



Customer Value

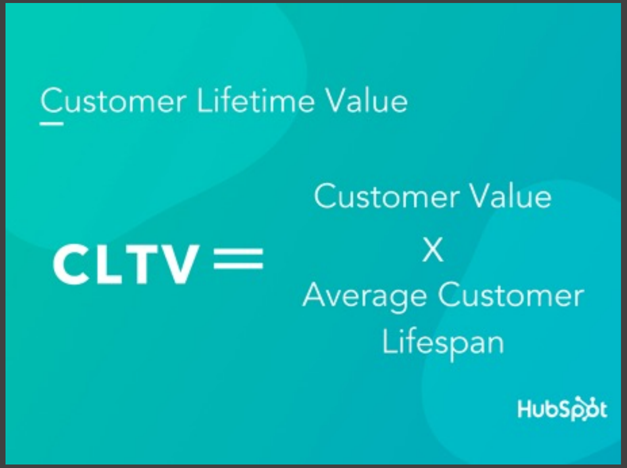
CV = Average Purchase Value X Average Purchase Frequency Rate



Average Customer Lifespan

$$ACL = \frac{\text{Sum of Customer Lifespans}}{\text{Number of Customers}}$$

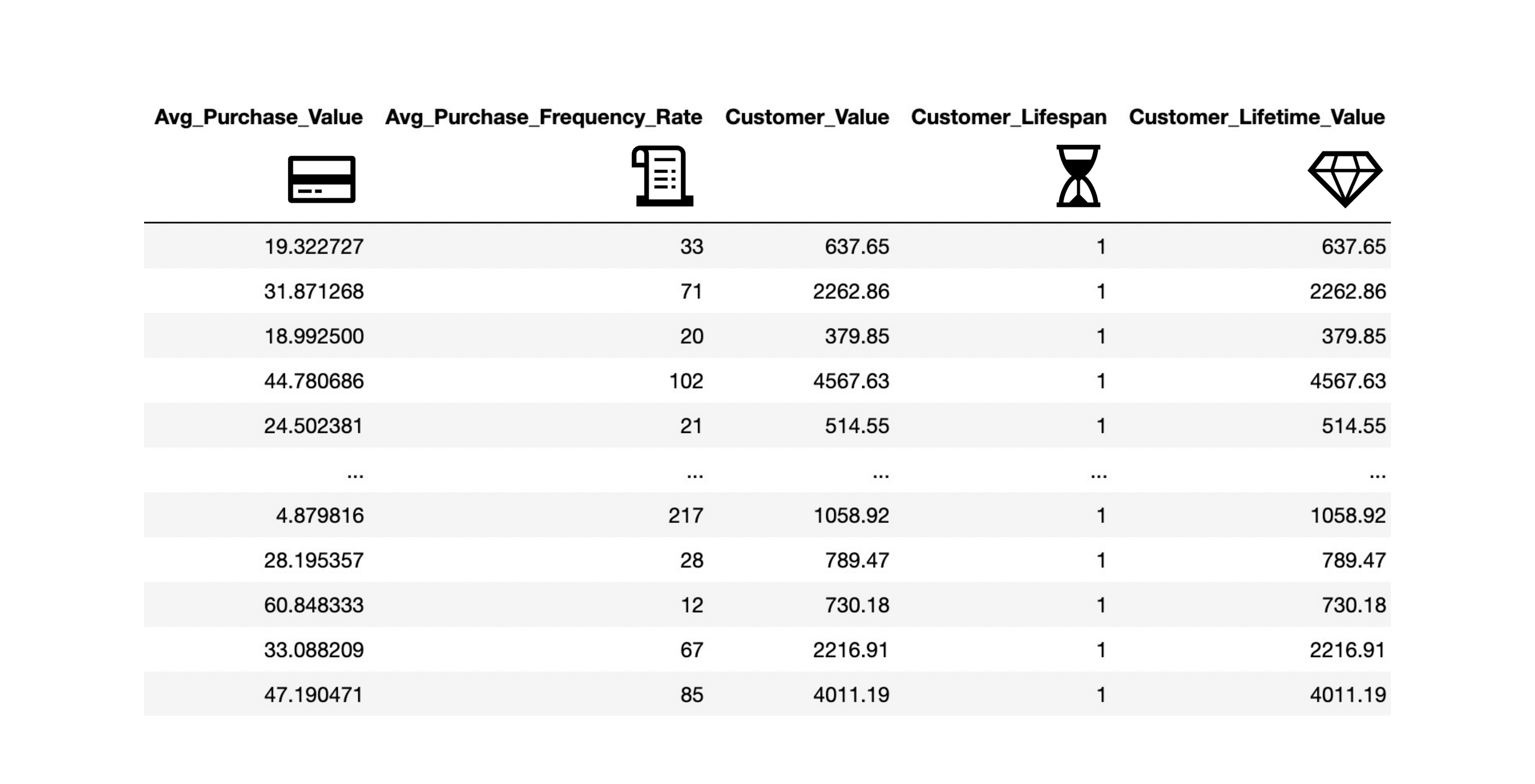
HubSpot



Customer Lifetime Value

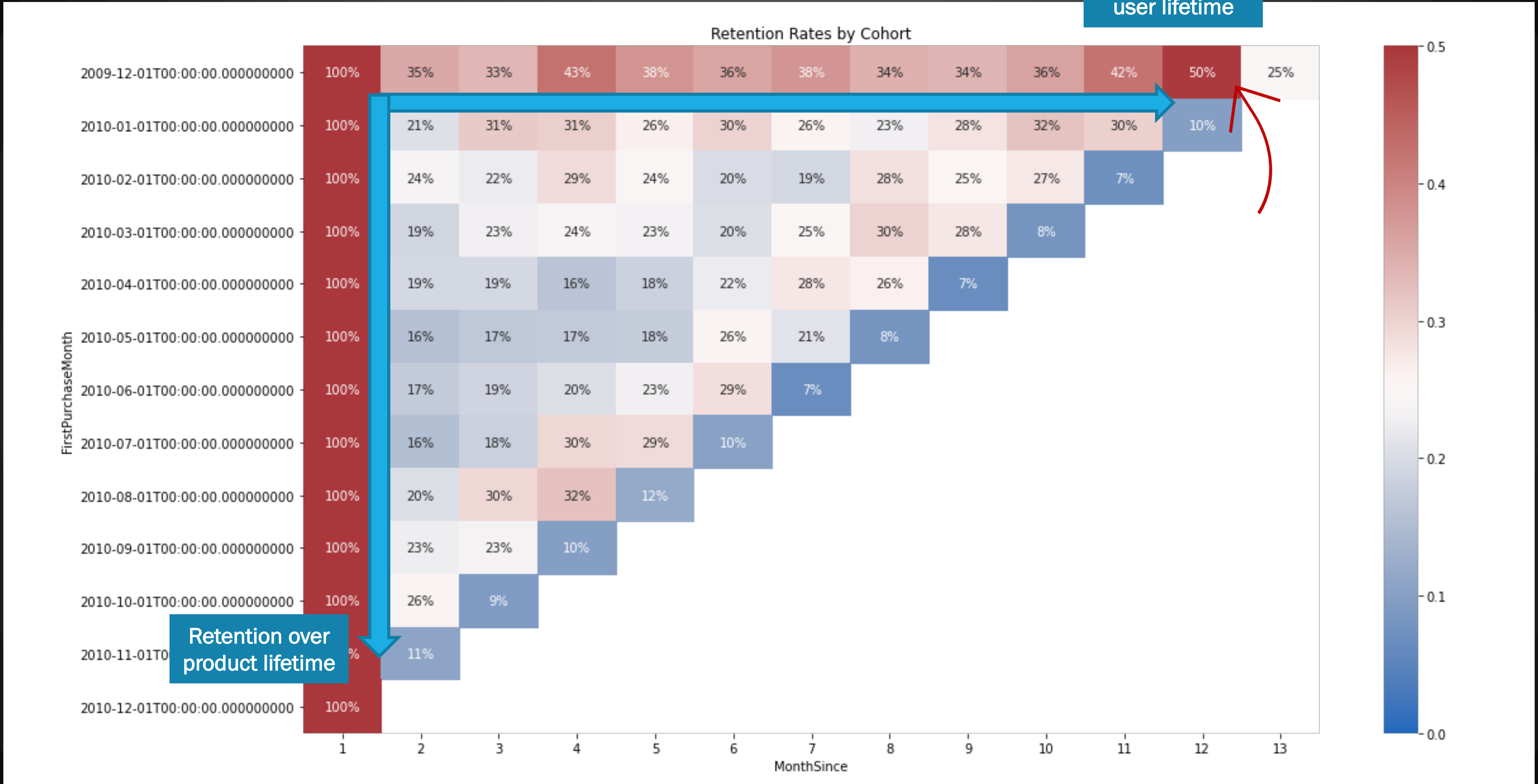
CLTV = Customer Value X Average Customer Lifespan

HubSpot



Avg_Purchase_Value 💳	Avg_Purchase_Frequency_Rate 🧾	Customer_Value	Customer_Lifespan ⌛	Customer_Lifetime_Value 💎
19.322727	33	637.65	1	637.65
31.871268	71	2262.86	1	2262.86
18.992500	20	379.85	1	379.85
44.780686	102	4567.63	1	4567.63
24.502381	21	514.55	1	514.55
...	...	...	...	...
4.879816	217	1058.92	1	1058.92
28.195357	28	789.47	1	789.47
60.848333	12	730.18	1	730.18
33.088209	67	2216.91	1	2216.91
47.190471	85	4011.19	1	4011.19

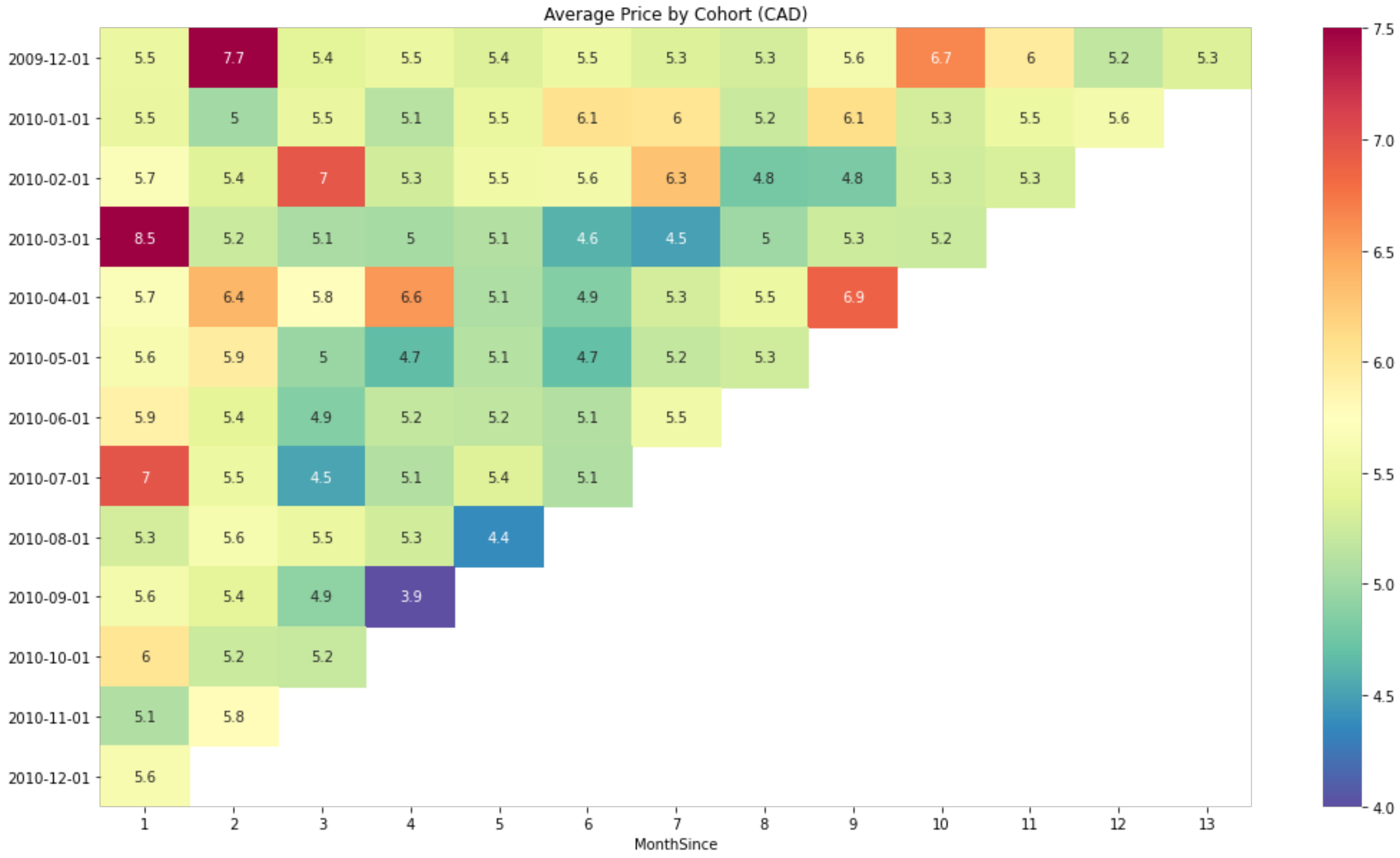
Retention over user lifetime



FirstPurchaseMonth	1	2	3	4	5	6	7	8	9	10	11	12	13
2009-12-01T00:00:00.000000000	100%	35%	33%	43%	38%	36%	38%	34%	34%	36%	42%	50%	25%
2010-01-01T00:00:00.000000000	100%	21%	31%	31%	26%	30%	26%	23%	28%	32%	30%	10%	
2010-02-01T00:00:00.000000000	100%	24%	22%	29%	24%	20%	19%	28%	25%	27%	7%		
2010-03-01T00:00:00.000000000	100%	19%	23%	24%	23%	20%	25%	30%	28%	8%			
2010-04-01T00:00:00.000000000	100%	19%	19%	16%	18%	22%	28%	26%	7%				
2010-05-01T00:00:00.000000000	100%	16%	17%	17%	18%	26%	21%	8%					
2010-06-01T00:00:00.000000000	100%	17%	19%	20%	23%	29%	7%						
2010-07-01T00:00:00.000000000	100%	16%	18%	30%	29%	10%							
2010-08-01T00:00:00.000000000	100%	20%	30%	32%	12%								
2010-09-01T00:00:00.000000000	100%	23%	23%	10%									
2010-10-01T00:00:00.000000000	100%	26%	9%										
2010-11-01T0	%	11%											
2010-12-01T00:00:00.000000000	100%												

Retention over product lifetime

Average Price by Cohort (CAD)



	1	2	3	4	5	6	7	8	9	10	11	12	13
2009-12-01	5.5	7.7	5.4	5.5	5.4	5.5	5.3	5.3	5.6	6.7	6	5.2	5.3
2010-01-01	5.5	5	5.5	5.1	5.5	6.1	6	5.2	6.1	5.3	5.5	5.6	
2010-02-01	5.7	5.4	7	5.3	5.5	5.6	6.3	4.8	4.8	5.3	5.3		
2010-03-01	8.5	5.2	5.1	5	5.1	4.6	4.5	5	5.3	5.2			
2010-04-01	5.7	6.4	5.8	6.6	5.1	4.9	5.3	5.5	6.9				
2010-05-01	5.6	5.9	5	4.7	5.1	4.7	5.2	5.3					
2010-06-01	5.9	5.4	4.9	5.2	5.2	5.1	5.5						
2010-07-01	7	5.5	4.5	5.1	5.4	5.1							
2010-08-01	5.3	5.6	5.5	5.3	4.4								
2010-09-01	5.6	5.4	4.9	3.9									
2010-10-01	6	5.2	5.2										
2010-11-01	5.1	5.8											
2010-12-01	5.6												

ANALYZING CUSTOMER VALUE

RECENCY

Earliest Invoice Date – First Purchase Date

- Number of days since a customer made the last purchase
- Or last visit day or the last login time
- Lower the better

FREQUENCY

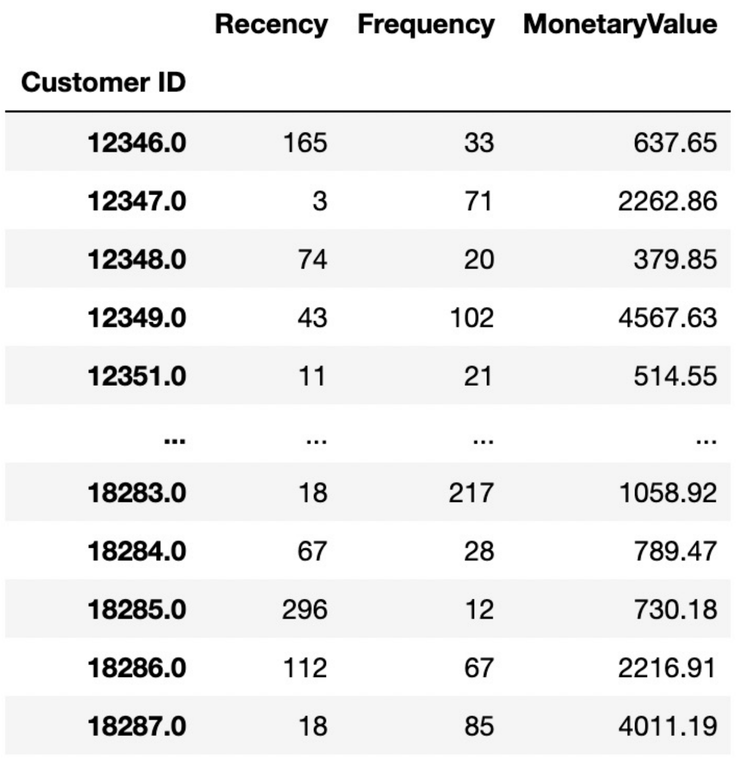
Number of Invoices grouped by Customer ID

- The number of purchases made in a given period
- How often a customer use our products/services
- Higher the better

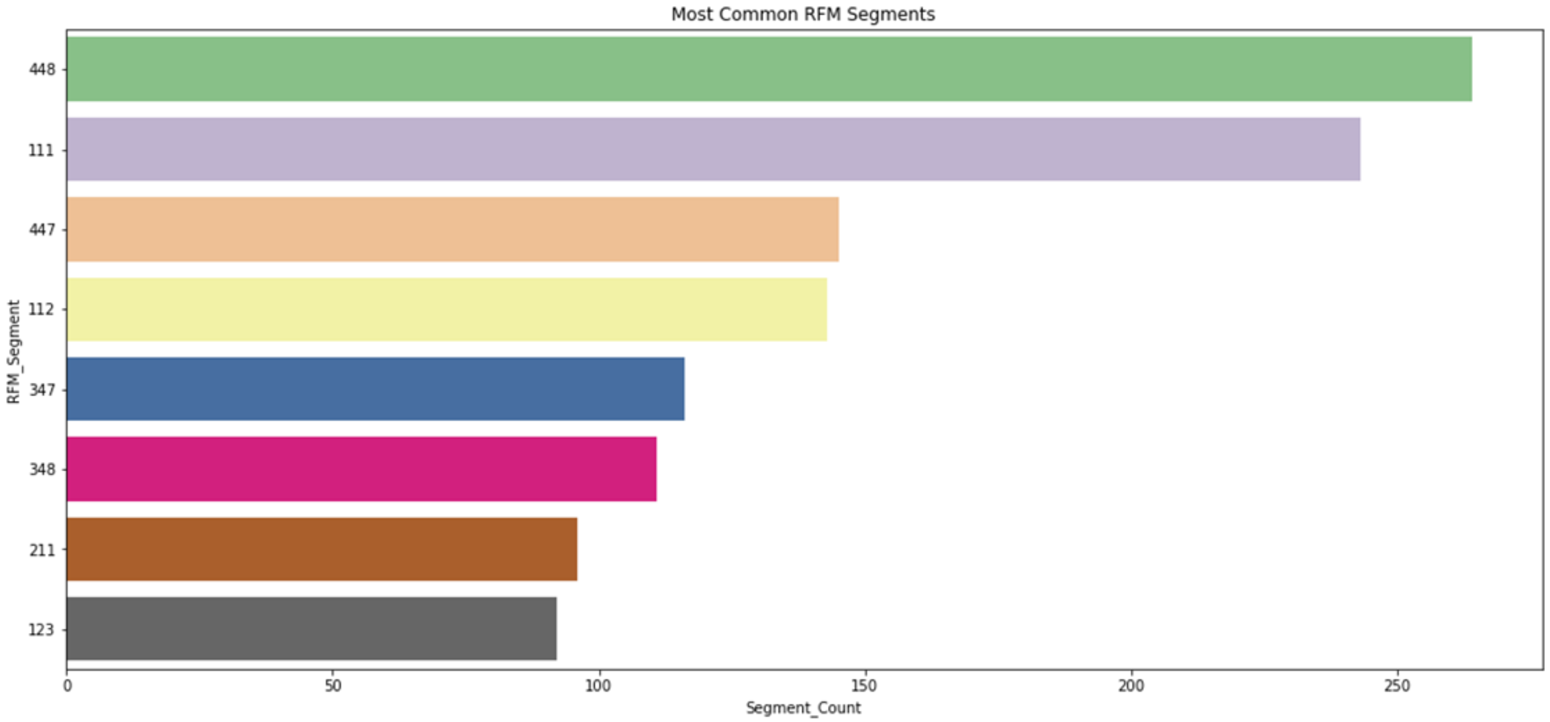
MONETARY VALUE

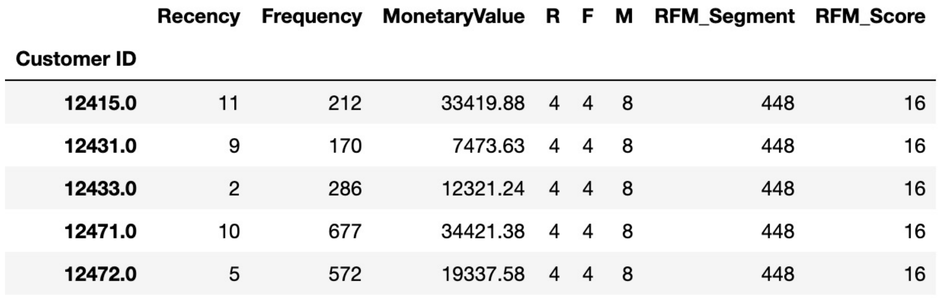
Unit Price * Quantity grouped by Customer ID

- Total amount of money a customer spent in a given period
- Useful for recognizing opportunities to upsell
- Higher the better



	Recency	Frequency	MonetaryValue
Customer ID			
12346.0	165	33	637.65
12347.0	3	71	2262.86
12348.0	74	20	379.85
12349.0	43	102	4567.63
12351.0	11	21	514.55
...	...	...	...
18283.0	18	217	1058.92
18284.0	67	28	789.47
18285.0	296	12	730.18
18286.0	112	67	2216.91
18287.0	18	85	4011.19

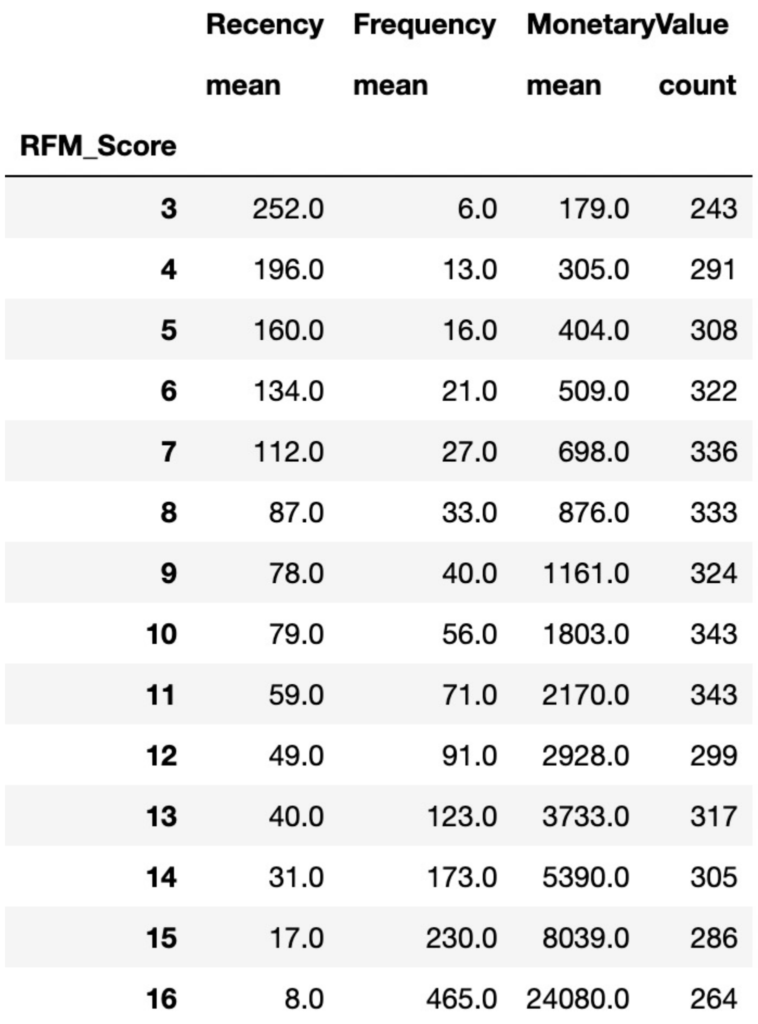




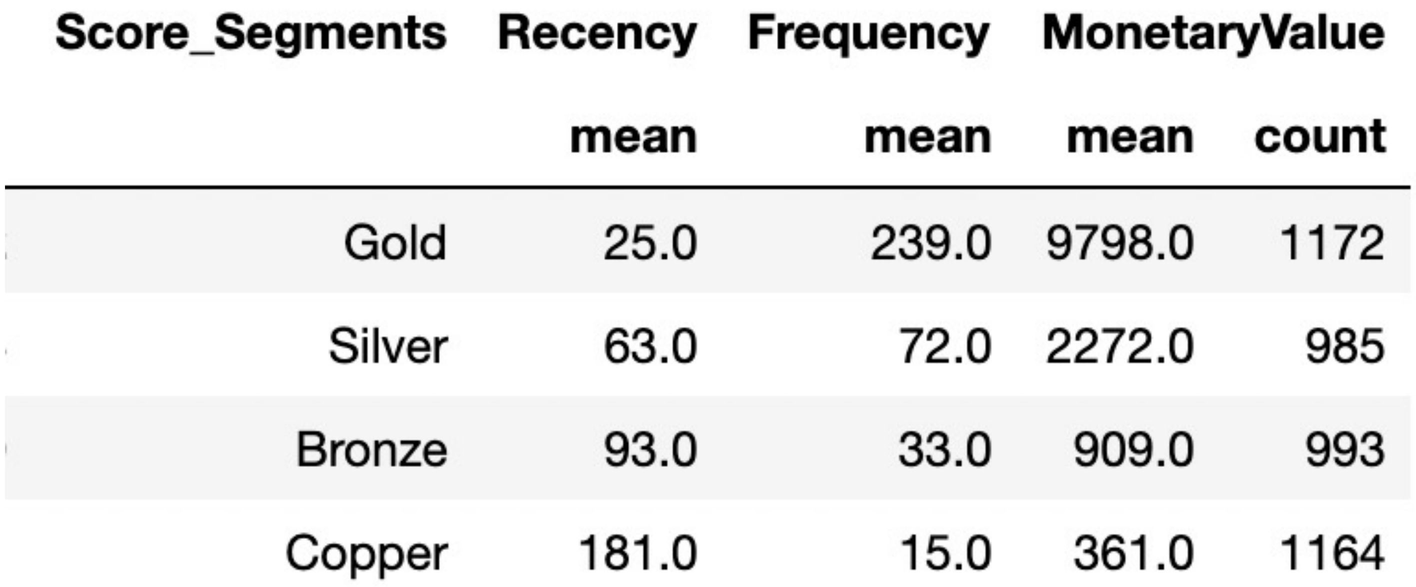
	Recency	Frequency	MonetaryValue	R	F	M	RFM_Segment	RFM_Score
Customer ID								
12415.0	11	212	33419.88	4	4	8	448	16
12431.0	9	170	7473.63	4	4	8	448	16
12433.0	2	286	12321.24	4	4	8	448	16
12471.0	10	677	34421.38	4	4	8	448	16
12472.0	5	572	19337.58	4	4	8	448	16

CUSTOMER SEGMENTATION BY COHORTS

- RFM Score: 3 to 16 (14 Segments Scores)
- RFM Segment Categories: 4x4x8 (128 Unique Categorical Segments)



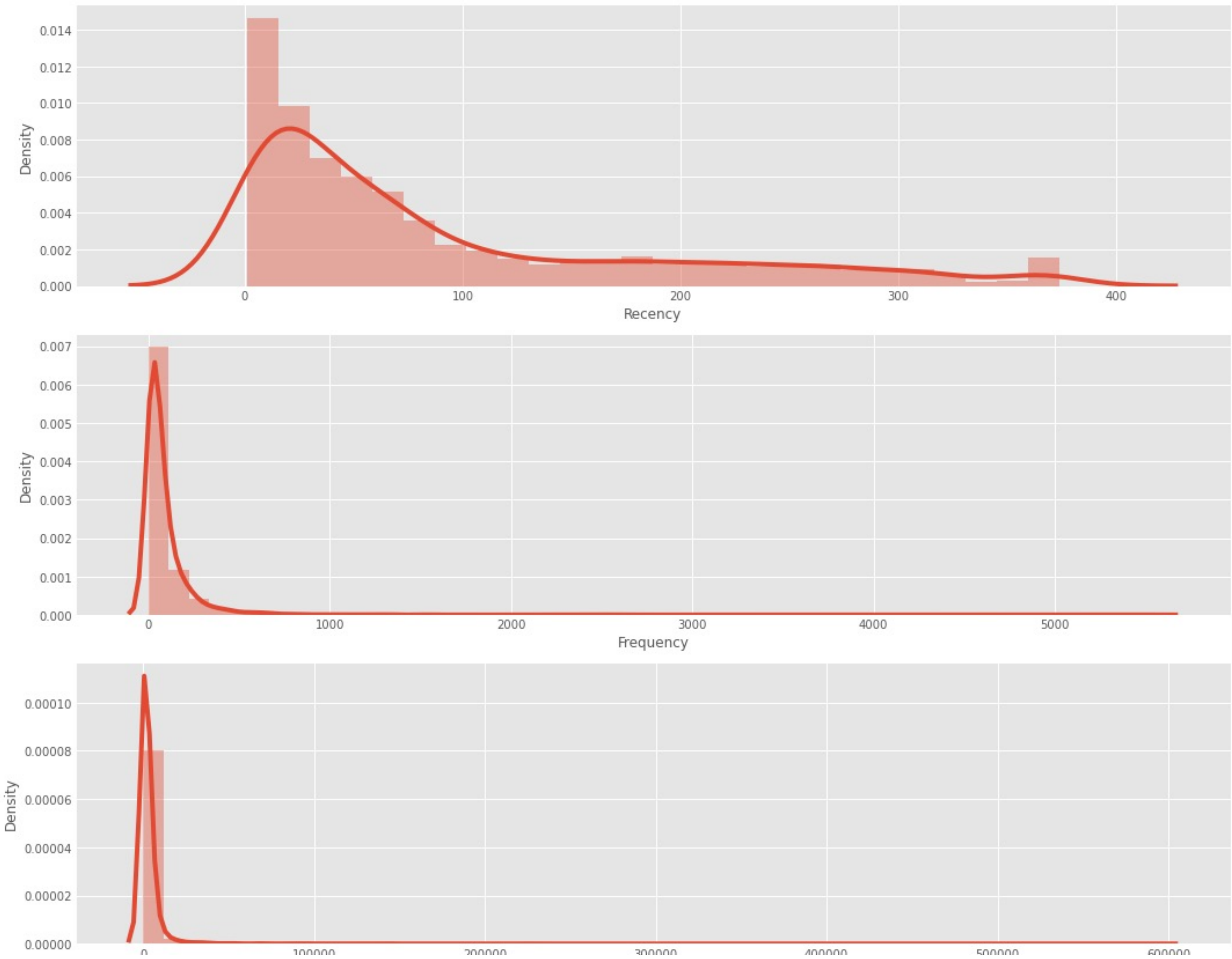
	Recency	Frequency	MonetaryValue	
	mean	mean	mean	count
RFM_Score				
3	252.0	6.0	179.0	243
4	196.0	13.0	305.0	291
5	160.0	16.0	404.0	308
6	134.0	21.0	509.0	322
7	112.0	27.0	698.0	336
8	87.0	33.0	876.0	333
9	78.0	40.0	1161.0	324
10	79.0	56.0	1803.0	343
11	59.0	71.0	2170.0	343
12	49.0	91.0	2928.0	299
13	40.0	123.0	3733.0	317
14	31.0	173.0	5390.0	305
15	17.0	230.0	8039.0	286
16	8.0	465.0	24080.0	264



Score_Segments	Recency	Frequency	MonetaryValue	
	mean	mean	mean	count
Gold	25.0	239.0	9798.0	1172
Silver	63.0	72.0	2272.0	985
Bronze	93.0	33.0	909.0	993
Copper	181.0	15.0	361.0	1164

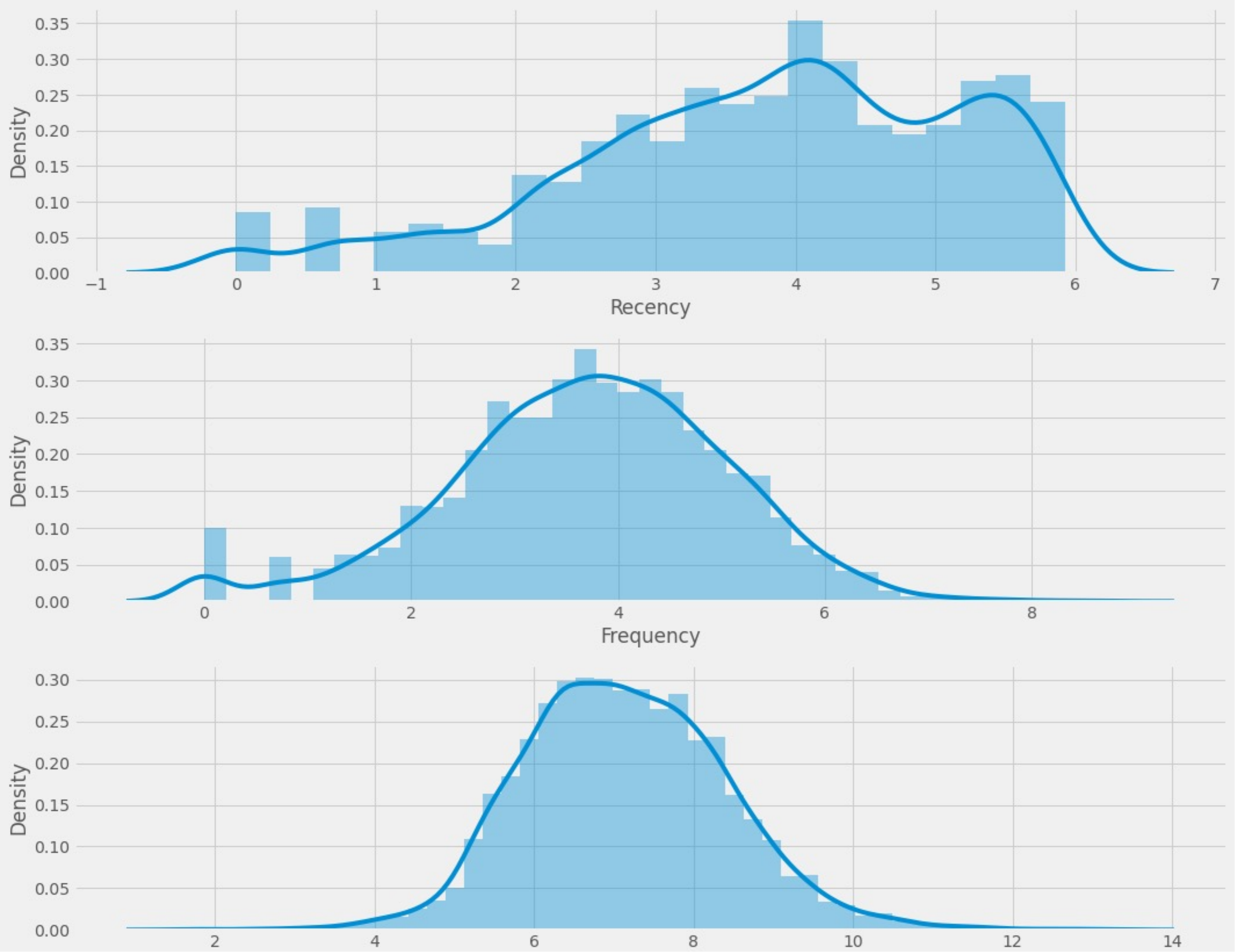
HEURISTIC APPROACH

- 4 Segments of similar size defined by having similar characteristics

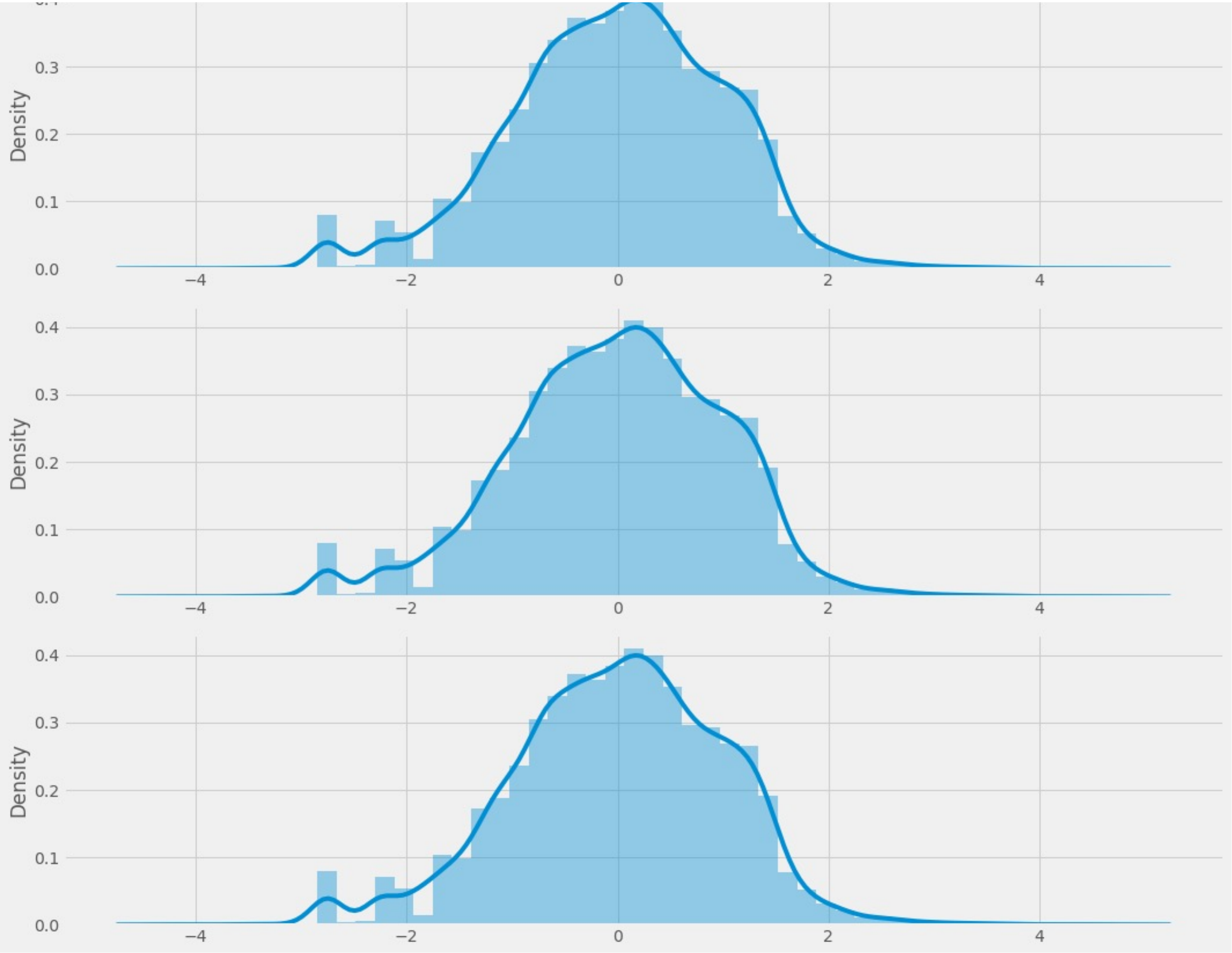


KMEANS ASSUMPTIONS

- Distribution of each variable is spherical (SSE is the right objective to minimize)
- All variables have the same mean (SSE)
- All variables have the same variance (variables are of equal importance)



LOGARITHMIC TRANSFORMATION

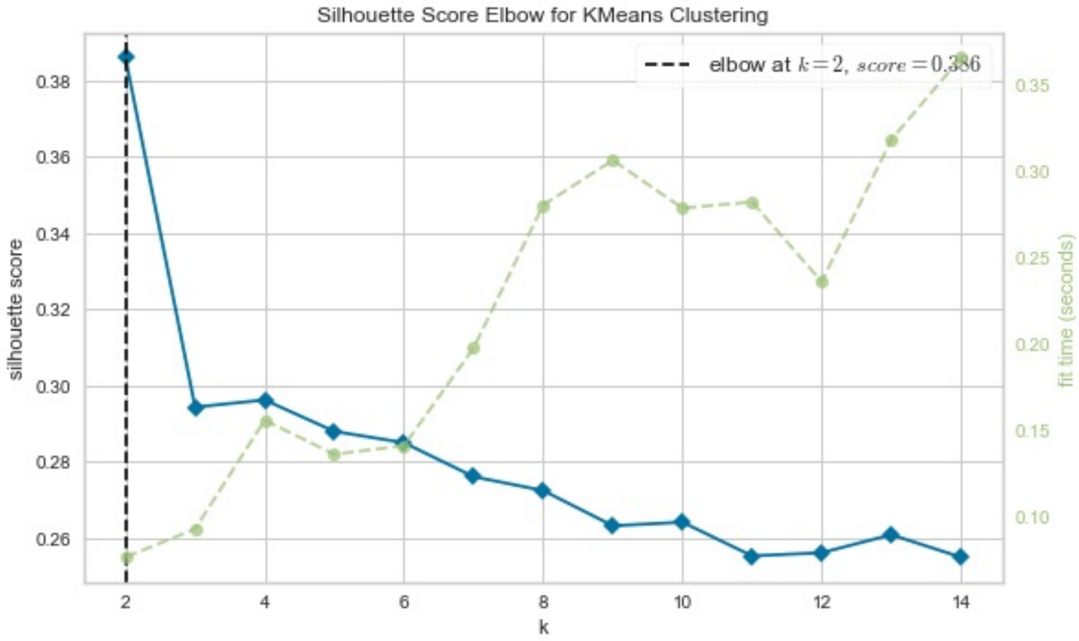


STANDARDIZING

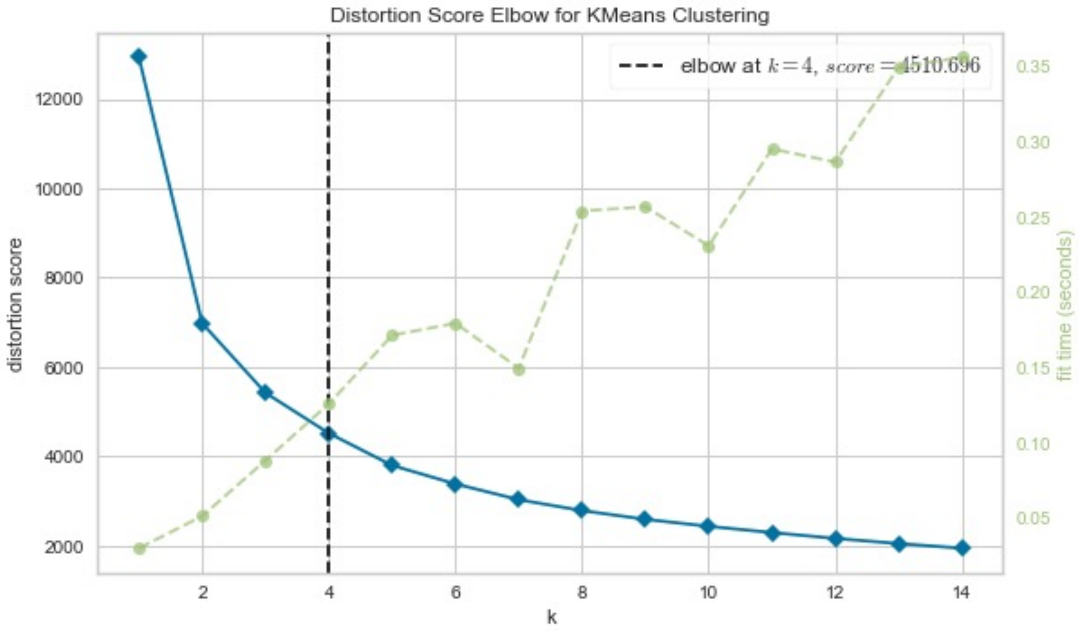
- Applied Standard Scalar
- Standardize to the same mean
- Scale to the same standard deviation

DETERMINING BEST K VALUE

from yellowbrick.cluster import KElbowVisualizer

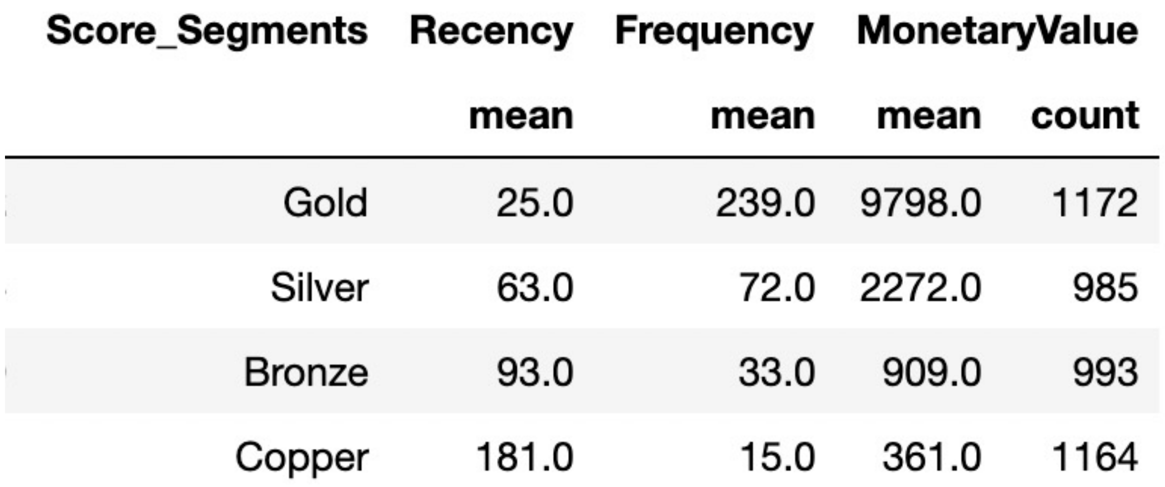


SILHOUETTE METHOD

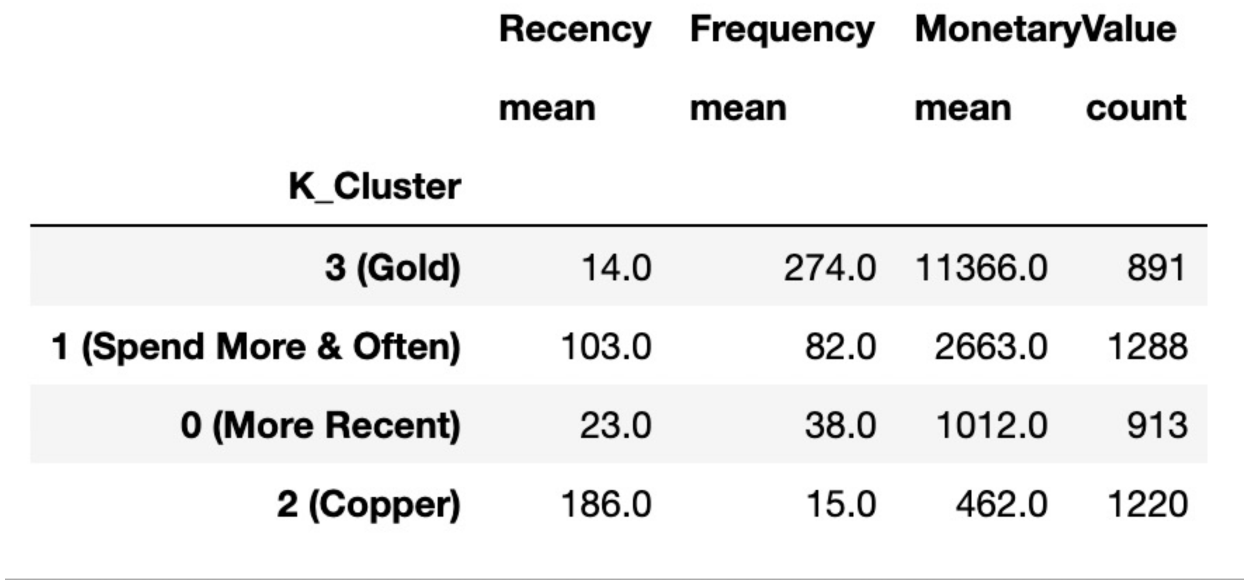


ELBOW METHOD

HEURISTIC GROUPING VS KMEANS

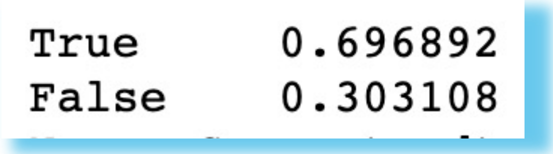


Score_Segments	Recency	Frequency	MonetaryValue	
	mean	mean	mean	count
Gold	25.0	239.0	9798.0	1172
Silver	63.0	72.0	2272.0	985
Bronze	93.0	33.0	909.0	993
Copper	181.0	15.0	361.0	1164



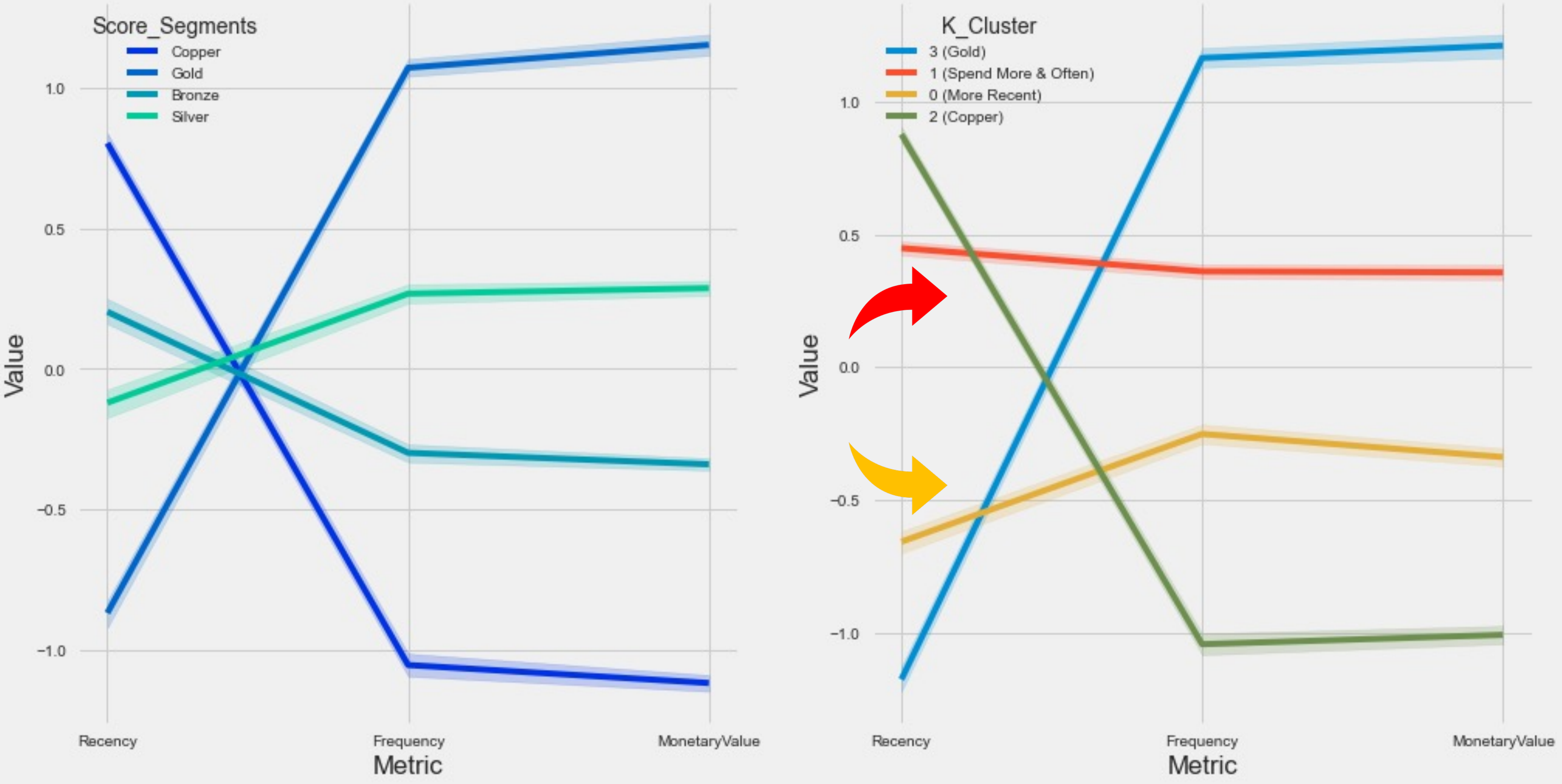
	Recency	Frequency	MonetaryValue	
	mean	mean	mean	count
K_Cluster				
3 (Gold)	14.0	274.0	11366.0	891
1 (Spend More & Often)	103.0	82.0	2663.0	1288
0 (More Recent)	23.0	38.0	1012.0	913
2 (Copper)	186.0	15.0	462.0	1220

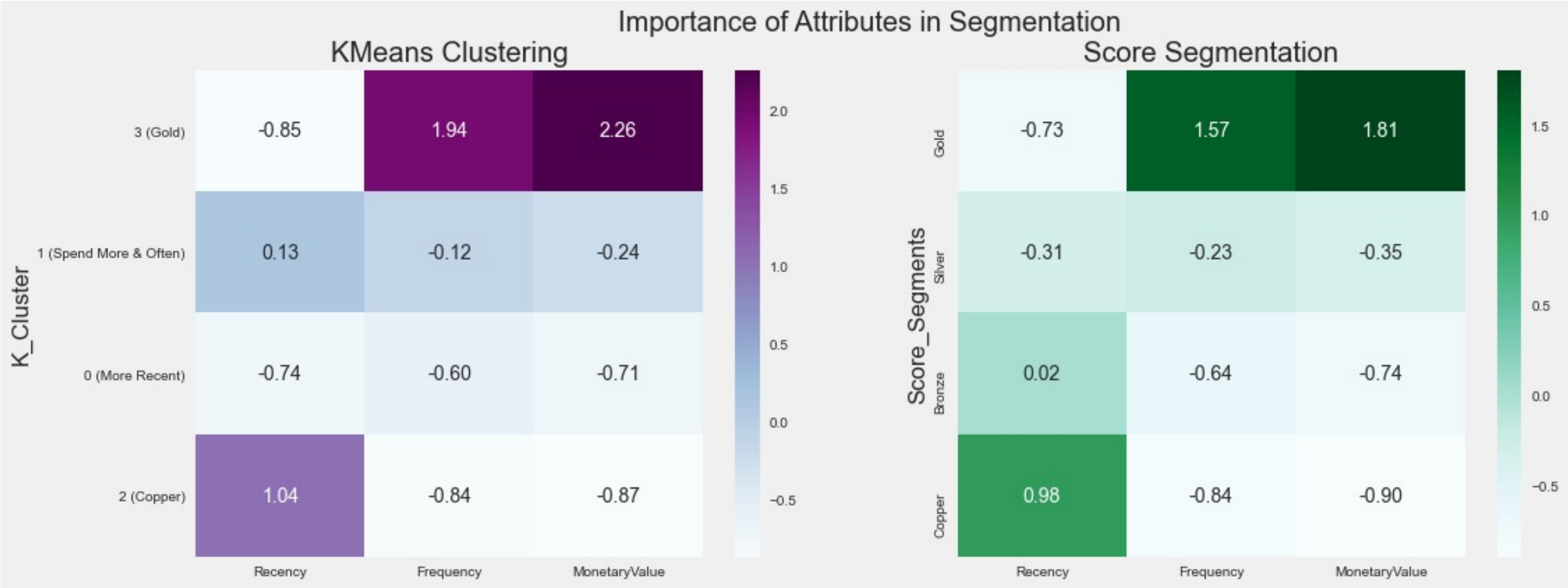
ERROR RATE = 30.3%



```
True        0.696892
False       0.303108
```

Segment Attributes





Importance of Attributes in Segmentation

KMeans Clustering

K_Cluster	Recency	Frequency	MonetaryValue
3 (Gold)	-0.85	1.94	2.26
1 (Spend More & Often)	0.13	-0.12	-0.24
0 (More Recent)	-0.74	-0.60	-0.71
2 (Copper)	1.04	-0.84	-0.87

Score Segmentation

Score_Segments	Recency	Frequency	MonetaryValue
Gold	-0.73	1.57	1.81
Silver	-0.31	-0.23	-0.35
Bronze	0.02	-0.64	-0.74
Copper	0.98	-0.84	-0.90

WHO ARE OUR CUSTOMERS?

&

WHERE DO OUR OPPORTUNITIES LIE?

Golden Goose

- 20.5% of customers *(891 customers)*
- Spends on average $11,000 CAD per year
- Repurchases 2 weeks after First Purchase
- Makes 270 orders per year

Loyal Legion

- 30% of customers *(1288 customers)*
- Spends on average $2,600 CAD per year
- Repurchases 3 months after First Purchase
- Makes 80 orders per year

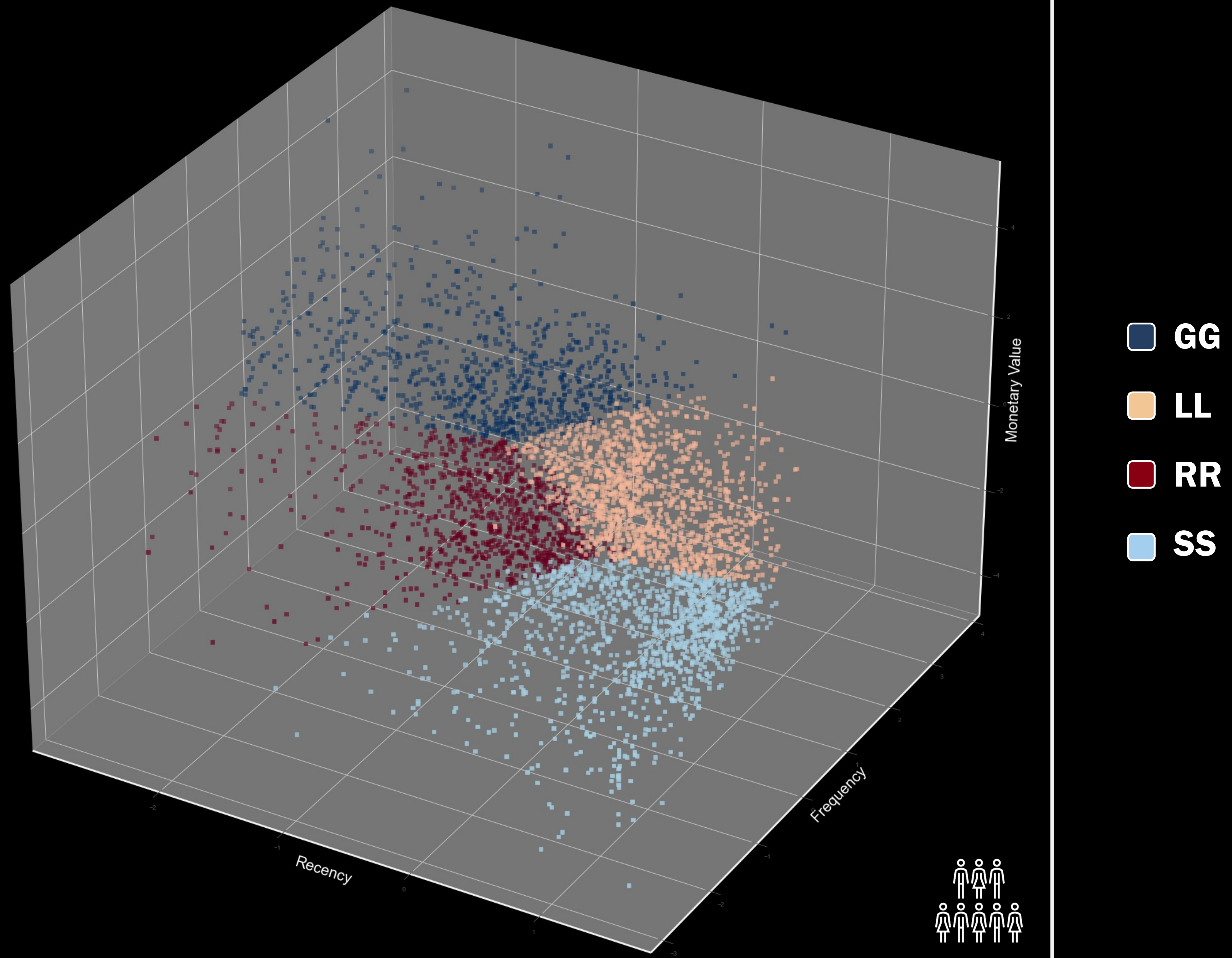
Recent Revisitor

- 21.2% of customers *(913)*
- Spends on average $1,000 CAD
- Repurchases 3 weeks after First Purchase
- Makes on average 38 orders

Slow Spender

- 28.3% of customers *(1220)*
- Spends on average $460 CAD
- Purchases again over 6 months later
- Makes on average 15 orders

3D
K-CLUSTERS
VISUALIZATION



QUESTIONS, SUGGESTIONS, & RETROSPECTION

THANK YOU TO ROGELIO AND SONIA FOR THE WONDERFUL LESSONS ❤️

- Limitations of K-Means
- Alternatives to K-Means
- Business Impact
- Additional Dataset – Combine multi-year data to follow entire Customer Lifespan
  - Combine with Customer Table to segment further by customer demographics