Online Retailer

Customer Segmentation

Data Science Capstone

Spring 2022 SECS 7259-01 22124

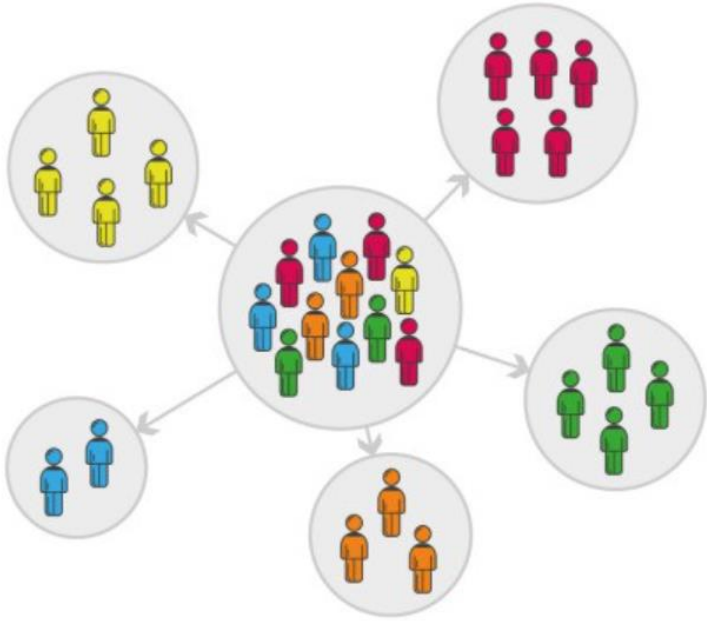
Online Retailer: Business Objective

- Need to explore: E-Commerce past infancy
  - Maximize efficiency leads to profits
  - Understand customers key to success

- Business Objective: Understand customer to enhance Product offering and maximize profit

- Scope and Design: Customer Segmentation
  - Data preparation: Data Exploration, Data Cleaning
  - Descriptive analysis: Visuals to derive insights
  - Unsupervised learning model: Clustering
    - Dimensionality reduction
    - Baseline: RFM Model
    - Elbow method and Silhouette score to find optimal clusters
    - K-Means Cluster analysis
    - Model iteration and best fit
  - Evaluating model
  - Model Implementation

- Stakeholders: Online Retailer, Customers, Support functions



Data Acquisition

- Transactions for a UK-based and registered, non-store online retail between 01/12/2009 and 09/12/2011

- Source:
  Dr. Daqing Chen, Course Director: MSc Data Science. School of Engineering,
  London South Bank University, London SE1 0AA, UK.
  https://archive.ics.uci.edu/ml/datasets/Online+Retail+II

- Reference:
  https://www.kaggle.com/datasets/lakshmi25npathi/online-retail-dataset?select=online_retail_II.xlsx

Online Retail: Year 2009-10

```
RangeIndex: 525461 entries, 0 to 525460
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   Invoice      525461 non-null  object
 1   StockCode    525461 non-null  object
 2   Description  522533 non-null  object
 3   Quantity     525461 non-null  int64
 4   InvoiceDate  525461 non-null  object
 5   Price        525461 non-null  float64
 6   Customer ID  417534 non-null  float64
 7   Country      525461 non-null  object
dtypes: float64(2), int64(1), object(5)
memory usage: 32.1+ MB
```

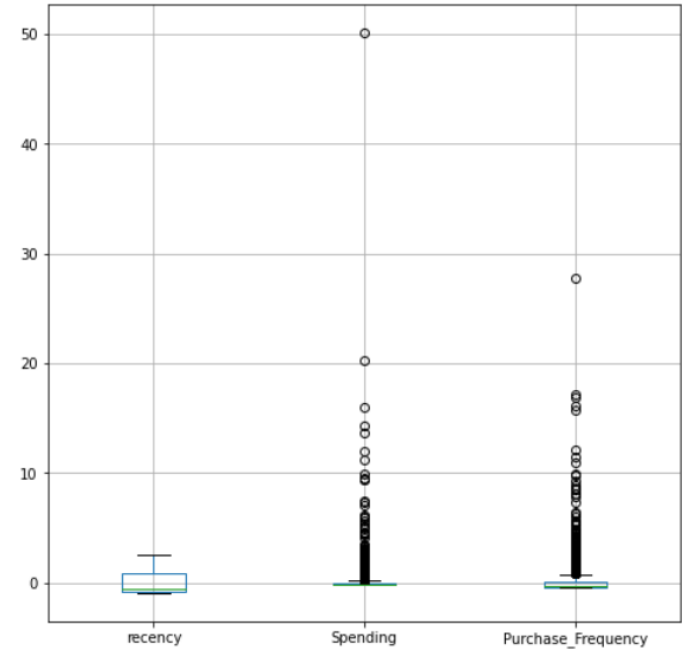
Online Retail: Year 2010-11

```
RangeIndex: 541910 entries, 0 to 541909
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   Invoice      541910 non-null  object
 1   StockCode    541910 non-null  object
 2   Description  540456 non-null  object
 3   Quantity     541910 non-null  int64
 4   InvoiceDate  541910 non-null  object
 5   Price        541910 non-null  float64
 6   Customer ID  406830 non-null  float64
 7   Country      541910 non-null  object
dtypes: float64(2), int64(1), object(5)
memory usage: 33.1+ MB
```

Online Retailer: Data Cleaning

- Step 1: Combined 2 worksheets:
  - Year 2009-10 and 2010-11
  - Total 1067371 entries
- Step 2: Removed entries without Customer ID
  - Reduced to 824364 entries
- Step 3: Outliers: Negative Quantity and Price
- Step 4: Normalization to scale
- Step 5: Removed Non UK dataset

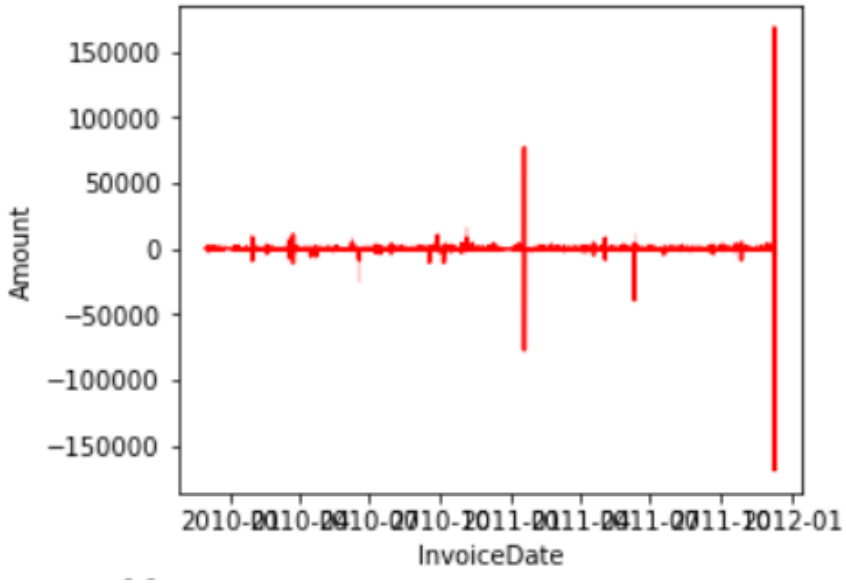
```
Data columns (total 8 columns):
 #   Column       Non-Null Count    Dtype
---  ------       --------------    -----
 0   Invoice      1067371 non-null  object
 1   StockCode    1067371 non-null  object
 2   Description  1062989 non-null  object
 3   Quantity     1067371 non-null  int64
 4   InvoiceDate  1067371 non-null  datetime64[ns]
 5   Price        1067371 non-null  float64
 6   Customer ID  824364 non-null   float64
 7   Country      1067371 non-null  object
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
memory usage: 65.1+ MB
```



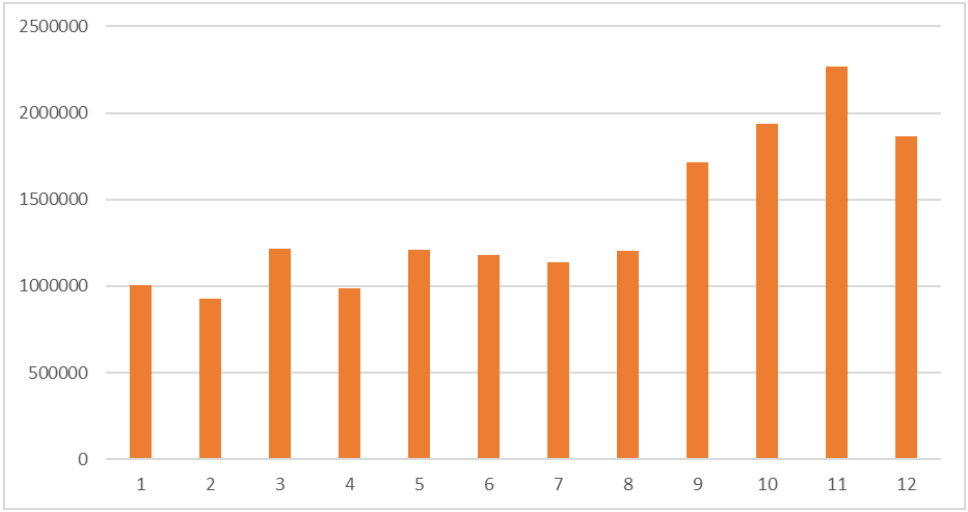
Online Retailer: Exploratory Data Analysis

- 5942 customers spent on average 2801.80 currency
- Each customers had purchased on average approx. 7.5 times in 2 years, at least once and at most 508 times
- On average, 293.97 spent by a customer on a purchase
  - 75% of customer purchases with 357.41
- Most Customers and most revenue generated from United Kingdom
  - Ireland, Netherlands, Germany, France, and Australia - top Revenue generating countries/ markets
- Seasonality in revenue generated
  - Peak months: Q4 – Oct to Dec

Revenue Generated in 2 years

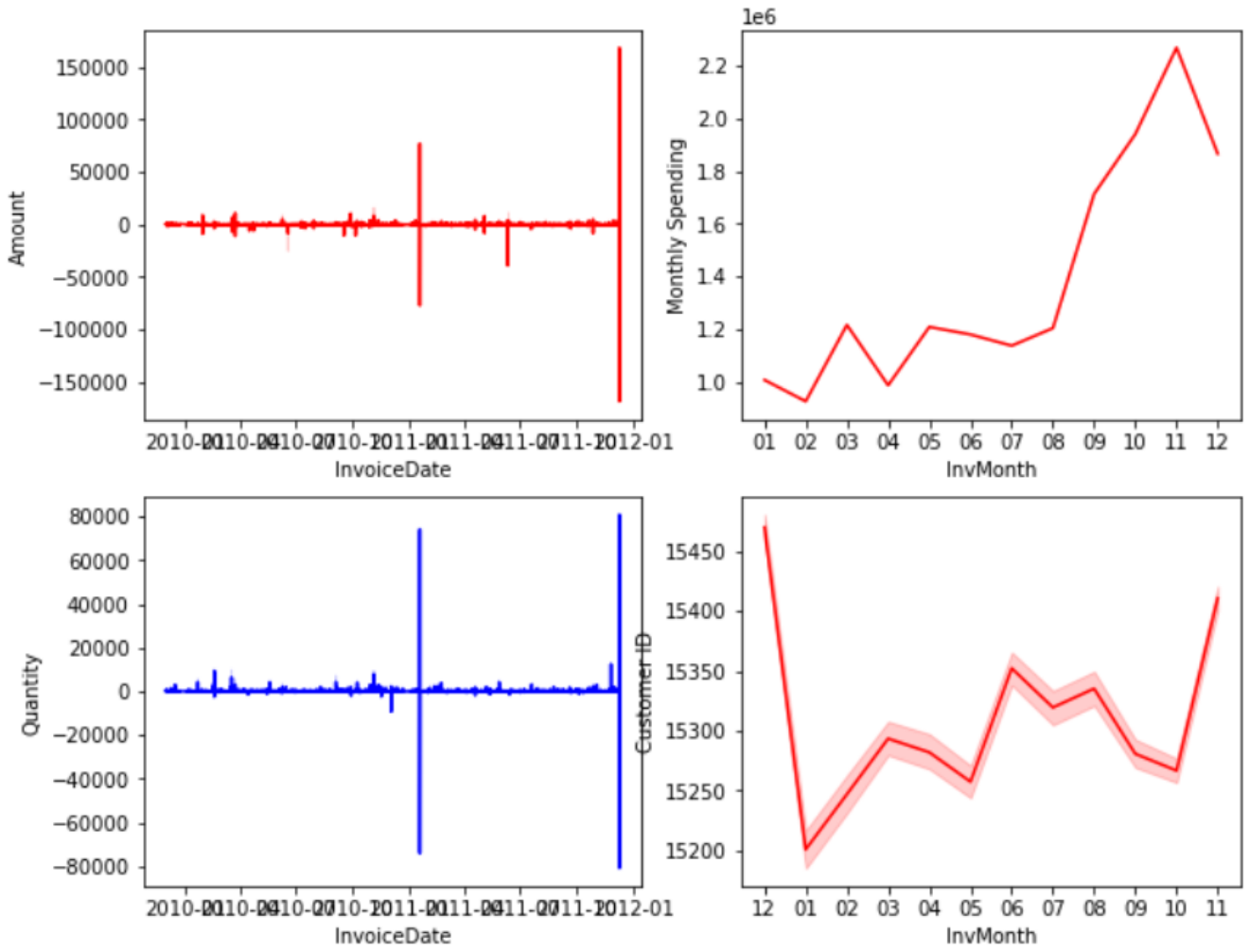


Month over Month: Revenue Generated



Online Retailer: Exploratory Data Analysis

- Further review of Cancel Orders:
  - Mainly coincided with peak order days
  - About a third of customers cancelled orders
  - Equates to just 8% of revenue
  - Top 3 Most cancels by Countries that order revenue with in Top 5:
    - United Kingdom
    - EIRE
    - France
  - About 49% of products have had cancelled orders



Online Retailer: Baseline

- Scatter plot of revenue to time distribution indicated 3 clusters
  - Cluster 0: Orders with moderate revenue generation
  - Cluster 1: Orders that were cancelled and resulted in loss of revenue
  - Cluster 2: Orders with high revenue generation

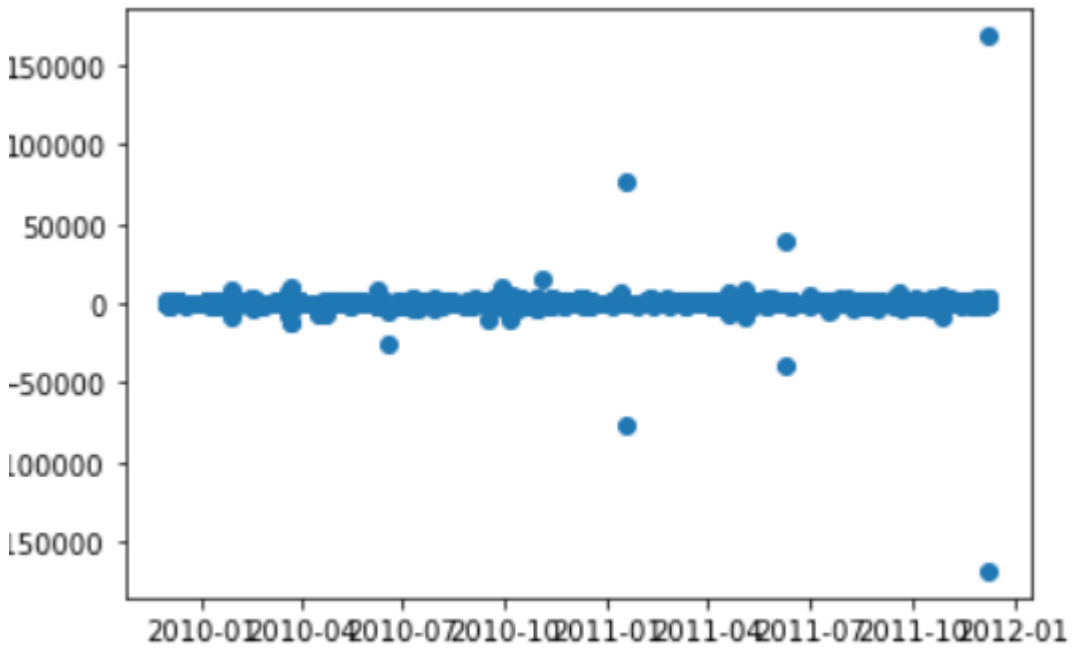
- K Means with 3 clusters shows similar composition as scatter plot clusters

```
1 online_retailer_2['clusters'].value_counts()
```

```
0    824339
1        23
2         2
Name: clusters, dtype: int64
```

- Segment of customer who is the big spender but what if they purchased only once or how recently they purchased?

Scatter plot of Revenue Generated



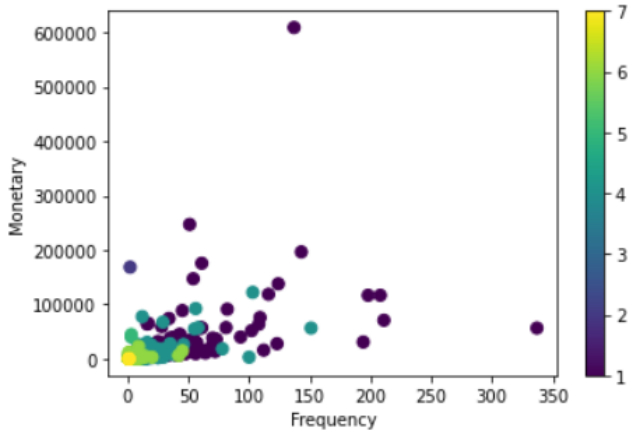
Clusters Composition: v1

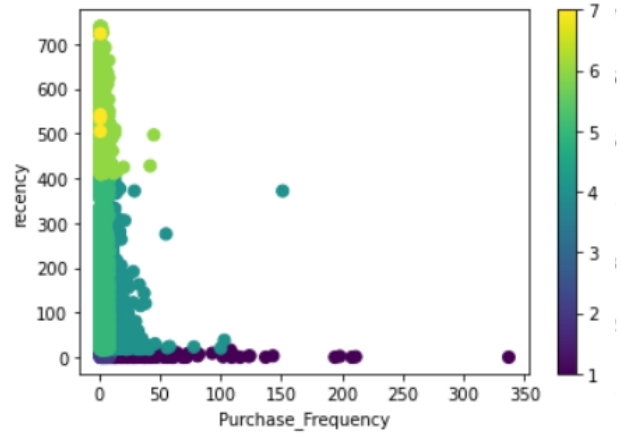
clusters	Quantity	Amount	Price
0	12.414962	19.869580	2.811637
1	-6748.652174	-15767.043478	9648.739130
2	77605.000000	122826.000000	1.500000

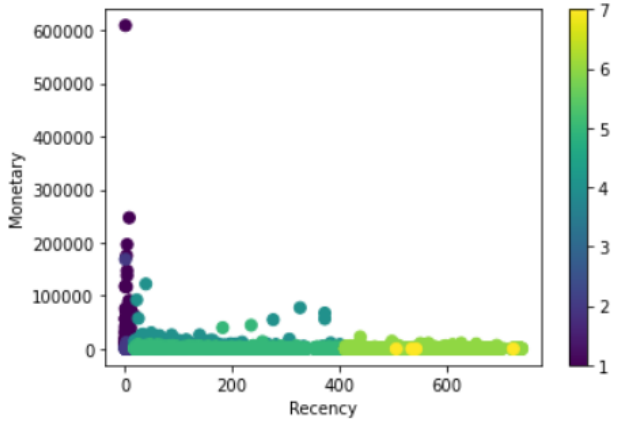
Online Retailer: RFM Model

- Customer behavior segmentation technique
  - Recency,
  - Frequency,
  - Monetary
- Each component split into quantiles
- Scores for each component assigned from quantiles
- Overall Score generated based on
  - how recently customer bought products,
  - how often customer buys, and
  - how much money was spent by customer,
- Based on each component and with Recency at higher weights, model designed with 7 clusters

Scatter plot: RFM Segments







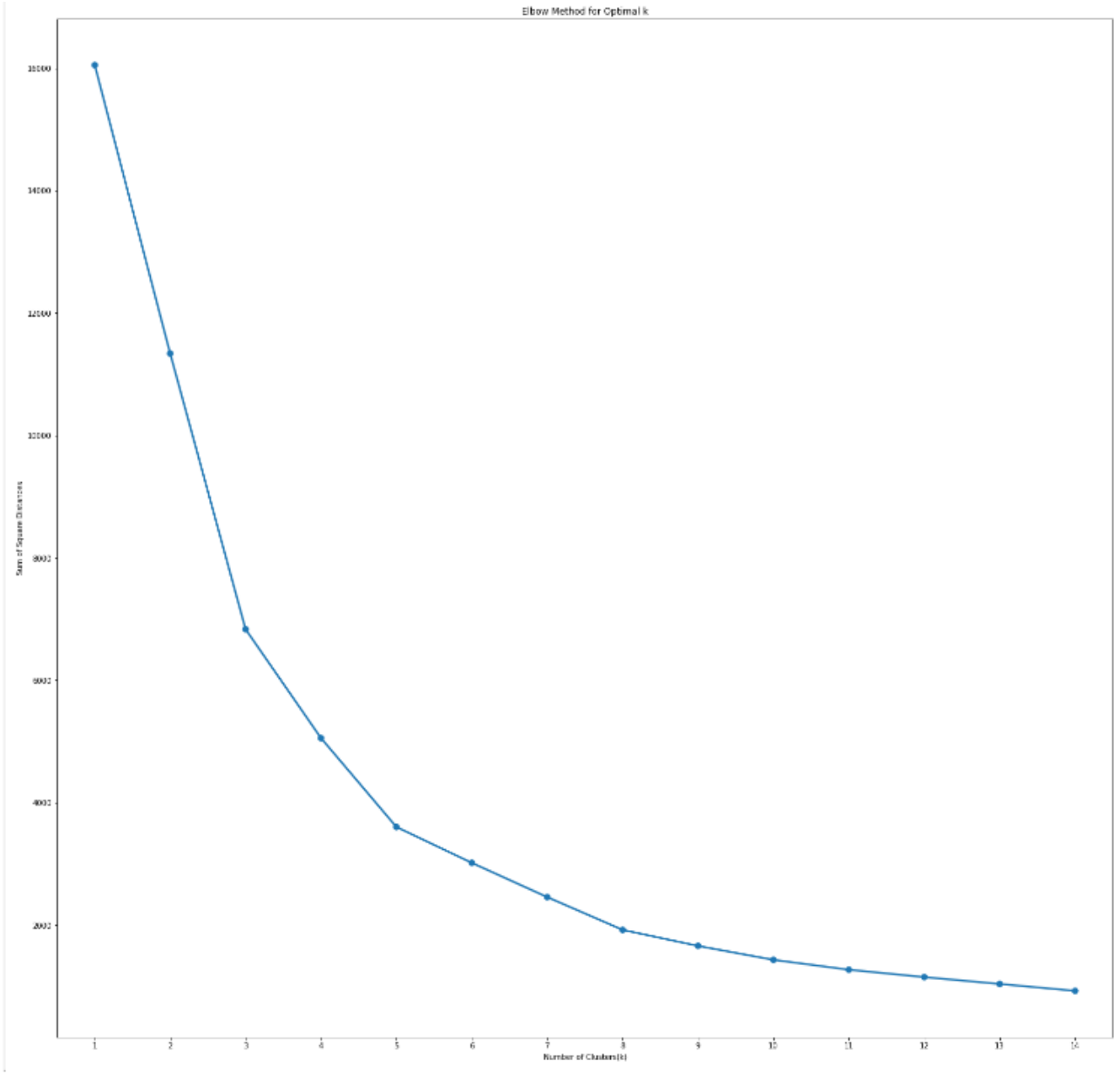
RFM and scores

	Customer ID	recency	Spending	Purchase_Frequency	r_score	f_score	m_score	rfm	rfm_scores	Segment_name	cluster
4	12747.0	2.0	9276.54	26	3	3	3	333	9	champion	1
5	12748.0	1.0	56599.39	337	3	3	3	333	9	champion	1
6	12749.0	4.0	6897.36	9	3	3	3	333	9	champion	1
9	12820.0	3.0	2689.52	11	3	3	3	333	9	champion	1
15	12826.0	3.0	2955.75	12	3	3	3	333	9	champion	1

Online Retailer: Elbow Method

- Most popular methods to determine Optimal value of k – Elbow method
- It is a plot of explained variation for a range of number of clusters
- Using the "elbow" as a cutoff point for number of clusters so that adding another cluster doesn't give much better modeling
- Elbow method to find optimal K indicates 5 clusters

Elbow method



Online Retailer: Silhouette Score

Silhouette score and Clusters

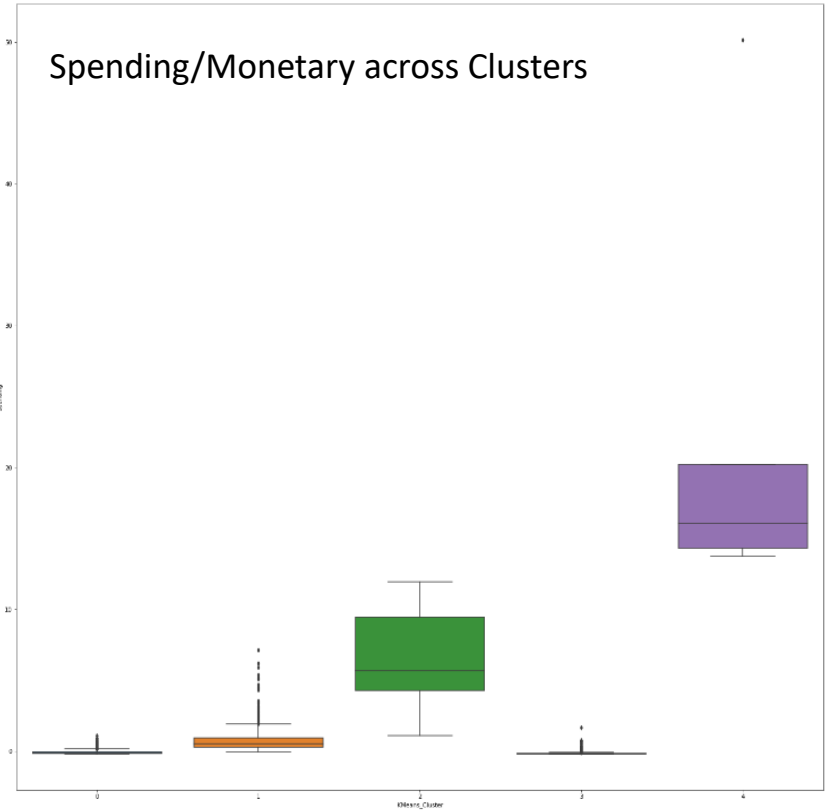
```
#3. Silhouette Analysis
for num_clusters in range(2,15):
    kmeans=KMeans(n_clusters=num_clusters,max_iter=50)
    kmeans.fit(rfm4_scaled)
    cluster_labels=kmeans.labels_
    #silhouette score
    silhouette_avg=silhouette_score(rfm4_scaled,cluster_labels)
    print("For n_clusters={0},the silhouette score is {1}".format(num_clusters,silhouette_avg))
```

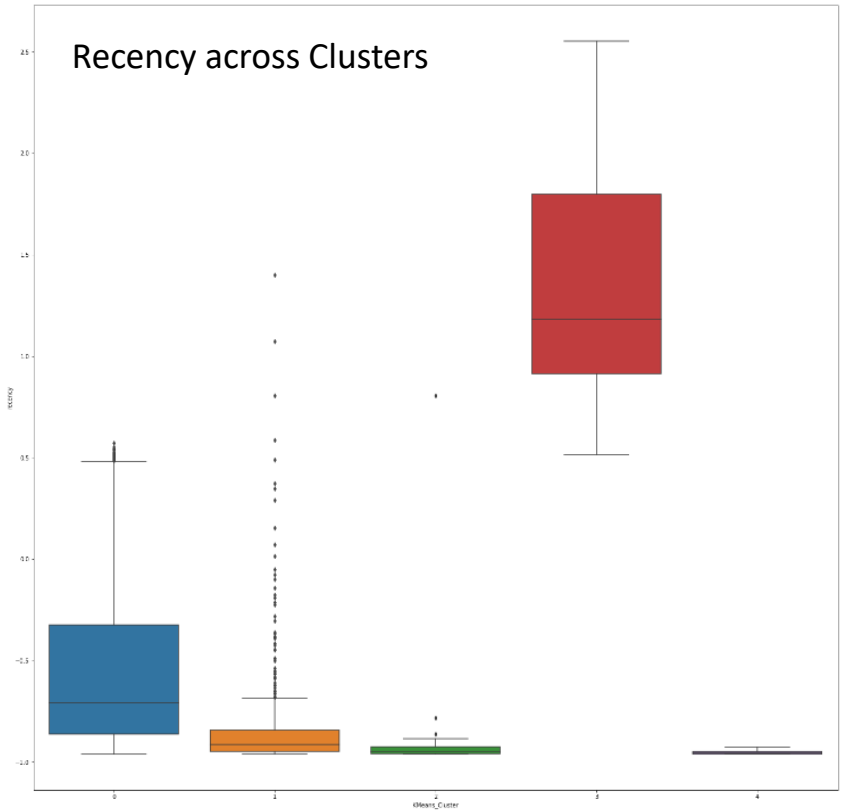
```
For n_clusters=2,the silhouette score is 0.5427383933423554
For n_clusters=3,the silhouette score is 0.5709408998925913
For n_clusters=4,the silhouette score is 0.5756265375828242
For n_clusters=5,the silhouette score is 0.5999459927700813
For n_clusters=6,the silhouette score is 0.5937624905154639
For n_clusters=7,the silhouette score is 0.5350453656163757
For n_clusters=8,the silhouette score is 0.5346261659994758
For n_clusters=9,the silhouette score is 0.4764734096744537
For n_clusters=10,the silhouette score is 0.4973912941046593
For n_clusters=11,the silhouette score is 0.4938293671036302
For n_clusters=12,the silhouette score is 0.4777990330924042
For n_clusters=13,the silhouette score is 0.49099801625143963
For n_clusters=14,the silhouette score is 0.468996663562025
```

- Silhouette Analysis for Clusters
  - Silhouette Coefficient or score is a metrics to calculate goodness of clustering technique
    - 1=> clusters are well apart from each other and clearly distinguished.
    - 0 => clusters are indifferent, or we can say that the distance between clusters is not significant.
    - -1 => clusters are assigned in the wrong way
  - Applying on Online Retailer dataset, 5 clusters highest Silhouette coefficient

Online Retailer: Final model

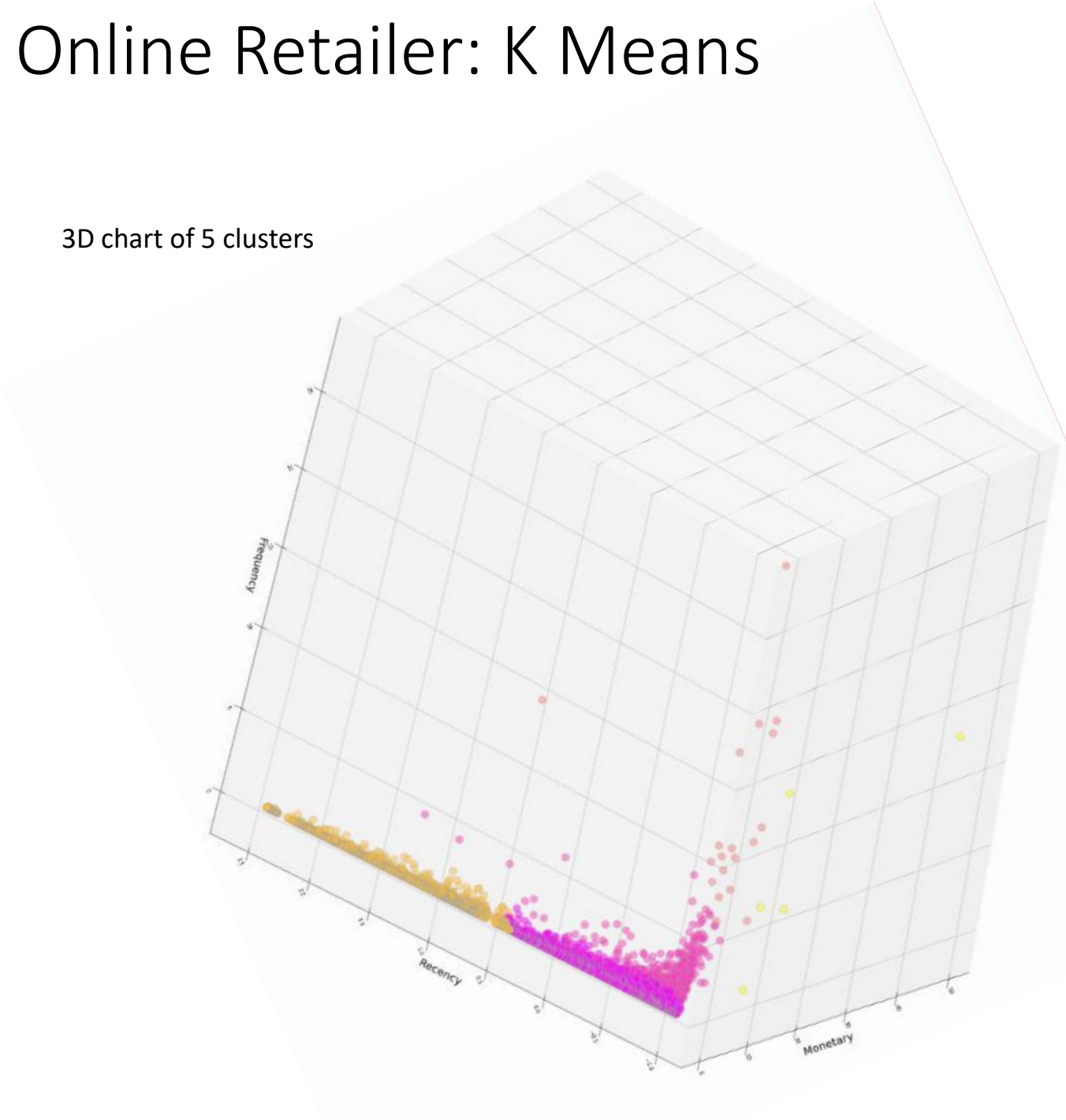
- K-Means identified 5 optimal Customer Segments or clusters
  - Cluster 0: Customer group who bought fairly recently, but fewer number of times and spending lower compared to others
  - Cluster 1: Customers who have recently purchased fairly high value products but did not repeat purchasing frequently
  - Cluster 2: Customers who have very high repeat purchases and higher spending, and recently purchased
  - Cluster 3: These Customers have not purchased in a very long time and have high churn while spending very little
  - Cluster 4: This is the best group of customers who buy very frequently and spent the most. Also they bought very recently





Online Retailer: K Means

3D chart of 5 clusters



K Means Clusters: 5 optimal

	Customer ID	Segment_name	Spending	Purchase_Frequency	recency	cluster	KMeans_Cluster
4	12747.0	champion	0.540141	1.660052	-0.957423	2	1
5	12748.0	champion	4.457811	27.798019	-0.962186	1	2
6	12749.0	champion	0.343179	0.231288	-0.947896	0	0
9	12820.0	champion	-0.005172	0.399378	-0.952659	0	0
15	12826.0	champion	0.016868	0.483423	-0.952659	0	0
28	12839.0	champion	0.605972	1.660052	-0.952659	2	1
30	12841.0	champion	0.396436	2.752636	-0.943133	2	1
62	12877.0	champion	0.016820	1.155782	-0.947896	2	1
85	12901.0	champion	1.233722	1.828142	-0.924080	2	1
97	12913.0	champion	0.326911	0.567468	-0.943133	0	0
105	12921.0	champion	2.745933	5.610163	-0.924080	2	1
119	12935.0	champion	0.123506	0.735558	-0.957423	0	1
121	12937.0	champion	0.028612	0.231288	-0.895500	0	0
132	12948.0	champion	-0.035975	0.231288	-0.890737	0	0
135	12951.0	champion	0.038225	0.735558	-0.924080	0	1
139	12955.0	champion	0.166000	0.399378	-0.962186	0	0
141	12957.0	champion	0.492989	0.567468	-0.919316	0	1
147	12963.0	champion	0.070335	0.483423	-0.928843	0	0
154	12970.0	champion	-0.101580	0.315333	-0.933606	0	0
155	12971.0	champion	1.094497	8.887914	-0.947896	2	2

Online Retailer: RFM Segment vs. K Means Clusters

```
KMeans_Cluster
0    3270
1     448
2      19
3    1608
4       5
Name: Customer ID, dtype: int64
KMeans_Cluster
0    61.121495
1     8.373832
2     0.355140
3    30.056075
4     0.093458
Name: Customer ID, dtype: float64
```

KMeans_Cluster	Segment_name	
0	champion	185
	needing_attention1	360
	needing_attention2	2205
	potential1	520
1	champion	298
	lost1	2
	needing_attention1	146
	needing_attention2	2
2	champion	16
	needing_attention1	3
3	lost1	1078
	lost2	4
	needing_attention1	15
	needing_attention2	511
4	champion	4
	potential1	1

Name: KMeans_Cluster, dtype: int64

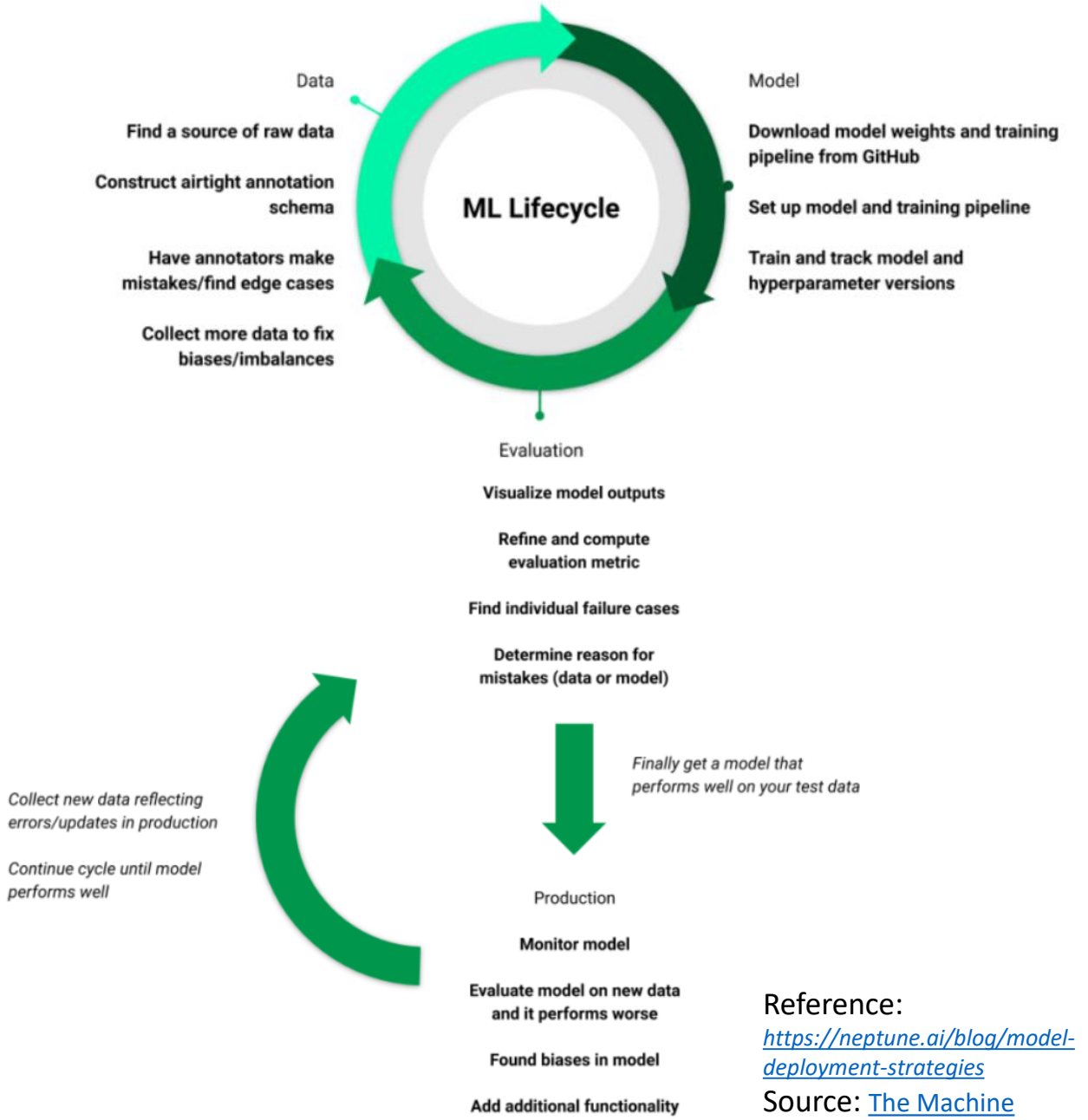
- 61% of customers are in K-Means Cluster 0
  - Overlap in Segments identified with RFM Scores vs. K Means Clusters
    - Champions: high Recency, high Frequency, high Spending spread across 4 K-Means clusters
      - 59% in K Means Cluster 1
      - 37% in Cluster 0
      - 3% in Cluster 2
      - 1% in Cluster 4

Online Retailer: Conclusion

- 61% of customers in Cluster 0
  - Grow this group with Marketing plan to drive more frequent purchase utilizing recommenders, coupon offerings for related past purchases
- 30% of customers not purchased recently.
  - Likely to churn
  - Nurture campaigns to bring back the group
- Cluster 1 is 8% of customers
  - Purchased fairly recent, bought high value products, did not purchase frequently
  - Likely dissatisfied customers
  - Feedback campaigns to increase engagement

Online Retailer: Next Steps

- Model deployment and Validation
  - Canary deployment strategy: Gradually add customer country
  - Internal Validation: Evaluating the fitness and Stability of clustering
    - Geometrical Properties of cluster like Compactness, Separation, Connectedness: Average Silhouette Width, and Bayesian information criterion (BIC)



Reference: *https://neptune.ai/blog/model-deployment-strategies*

Source: The Machine Learning Lifecycle in 2021

Online Retailer: Model Profiling

- Risks of model:
  - Every clustering algorithm will find clusters in a dataset, even if there is no cluster structure in it
  - Outliers
    - Country specific
    - No Cancel orders
    - Extreme outliers will skew clusters

- Stakeholders for project:
  - Online Retailer, Customers, Support functions
    - Marketing team for building nurture campaigns via emails, texts
    - Finance team for determining profitable offers and discounts
    - Product purchase and maintenance team for stocking products
    - Data Science team for building and maintenance of Customer segmentation model

Next Steps:

- Optimize product offerings by customer segments
- Design Recommenders to assist customer segments