Starbucks Customer Segmentation using K-Means Cluster Analysis

Emily Siegel
DSI Capstone
September 21, 2021

Starbucks Rewards & Customer Data

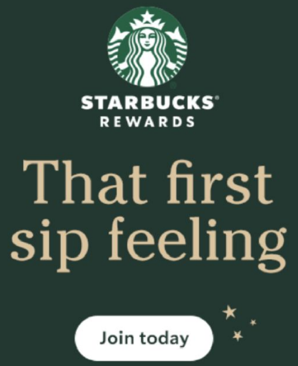
"As of October 2020, the Starbucks Rewards program has over **19.3 million members** and generates nearly **50% of their revenue**."

306,534 "events"

17,000 customers

10 promo types

30 days



Problem Statement

Conducting analysis on a company's customer base and sending personalized campaigns to high value targets has massive benefits in any industry. Using unsupervised learning, I will implement a K-Means cluster analysis for customer segmentation and targeted marketing outreach for Starbucks. This type of analysis can be used by Starbucks to automate promotional outreach and reward fulfillment, as well as measurement and tracking of spending behaviors and other KPIs.

Workflow

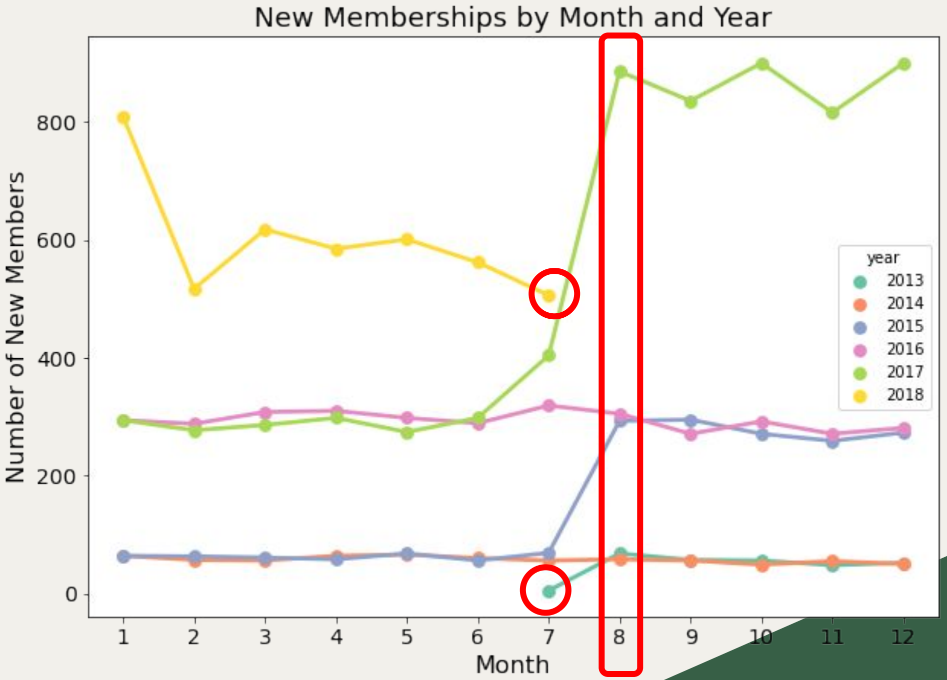
Data Wrangling
Impute nulls, Explore outliers, One-Hot Encoding, Aggregating data where needed & Merging all data

EDA
Customer Profiles, Offer Types & Transaction Data, Feature Engineering & RFM Metrics

Cluster Analysis
Feature Scaling, Dimensionality Reduction & Clustering using K-Means and an attempt at DBSCAN

Post Hoc Analysis
Customer Segments & Insights

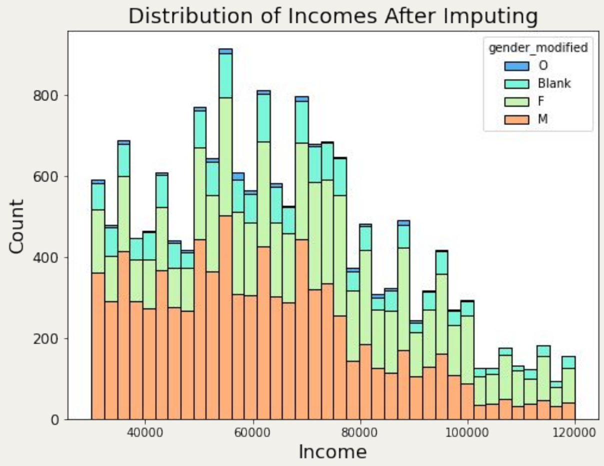
Rewards Program Membership

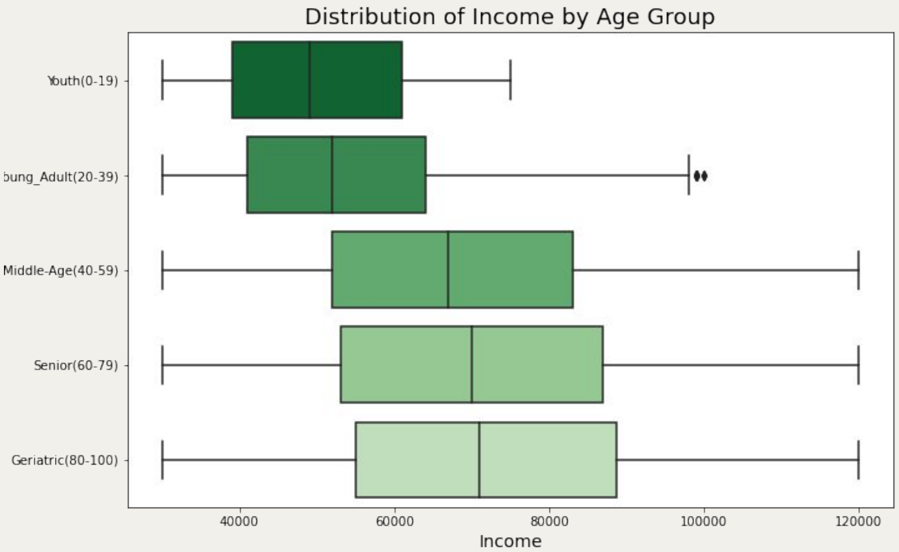


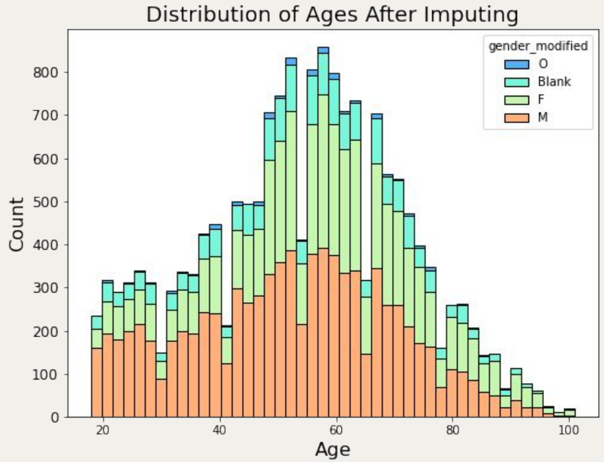
August stands out with large jump in new memberships

Customer Demographics

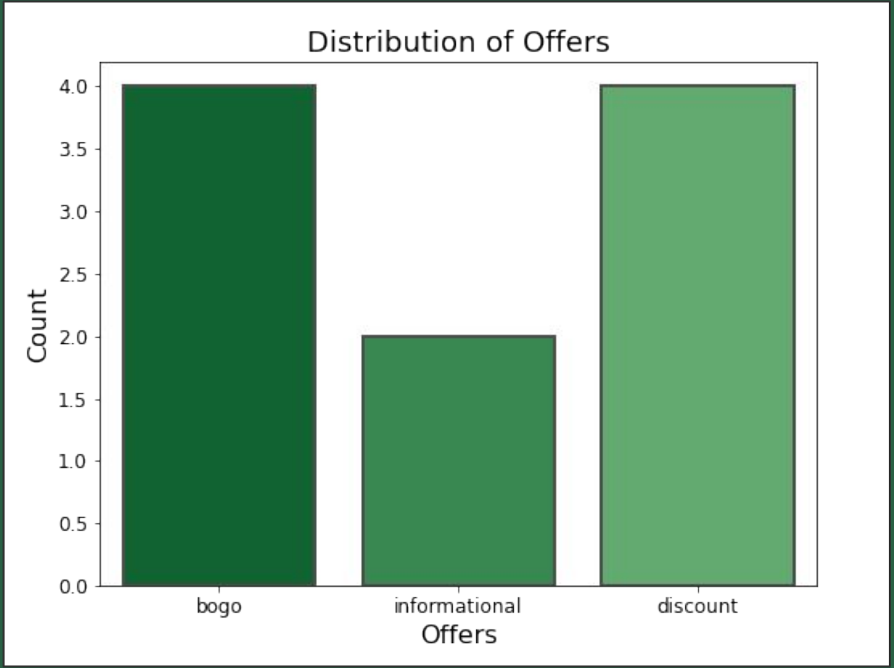
Male: 49.9%
Female: 36.1%
Blank: 12.8%
Other: 1.2%



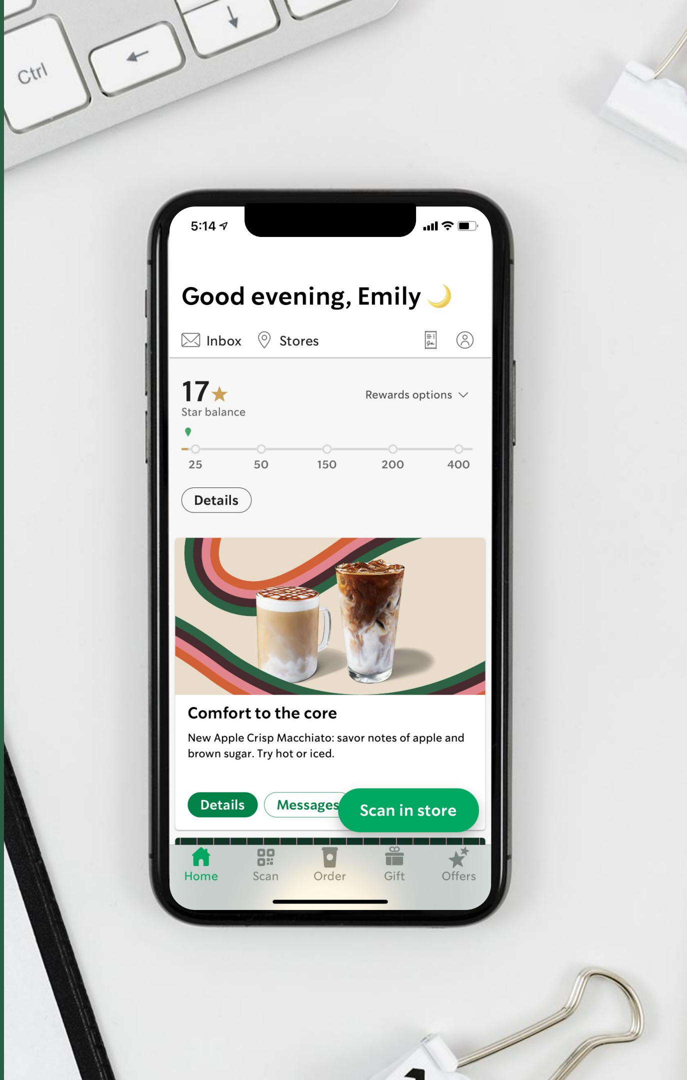




Promotional Offers



Majority of offers use all 4 channel types (email, mobile, web, social) for promotion.

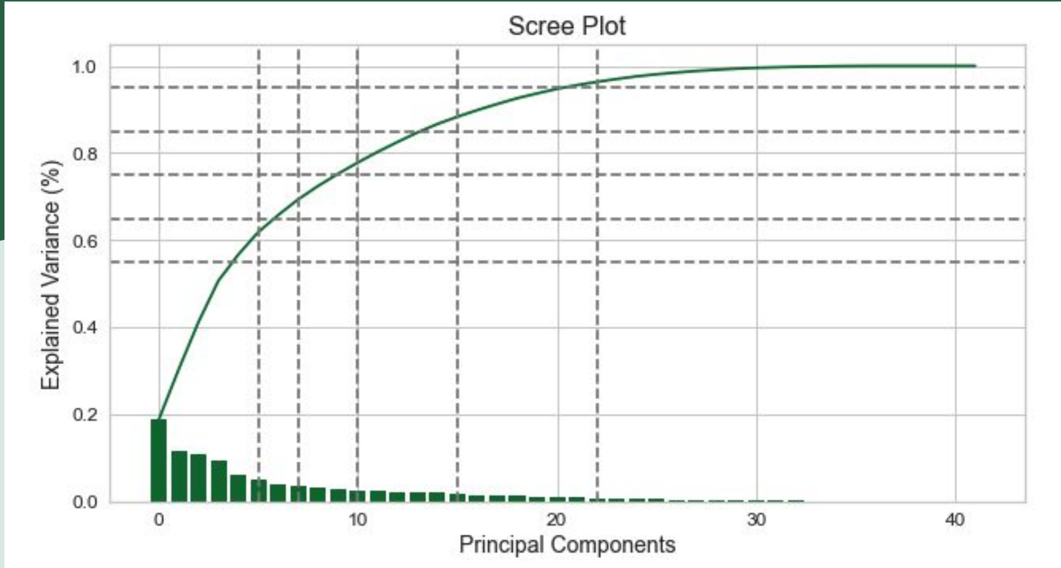


PCA

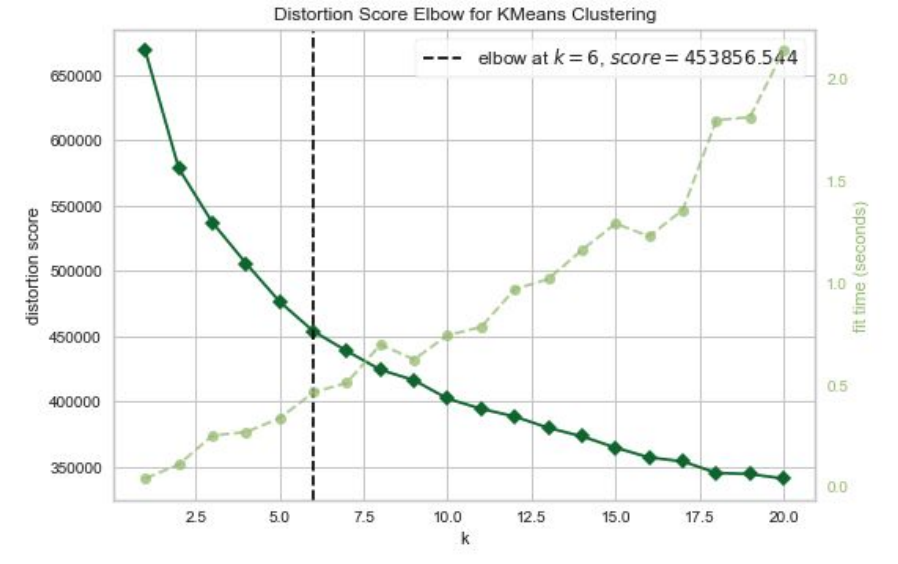
A large set of possibly correlated variables can be summarized with a smaller number of variables that explain most of the variability in data.

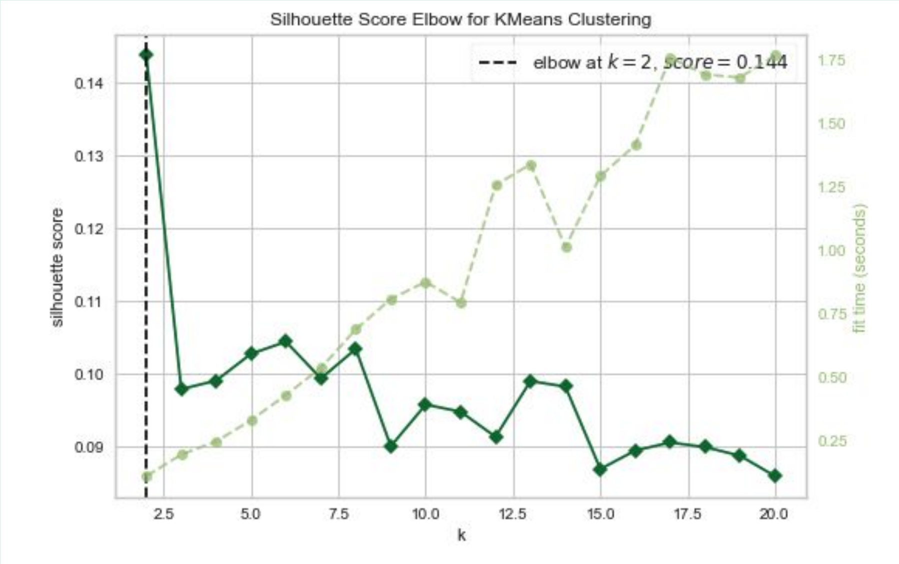
Scree plot - A visual approach to selecting the number of principal components to keep

20 Components explained over 90% of the variability



How many clusters for K-Means???





Distortion Score:
The sum of squared distances from each point to its assigned center.

Silhouette Score:
The mean Silhouette Coefficient of all samples. Cohesion of a point to its cluster compared to other clusters (separation).

K-Means Clustering

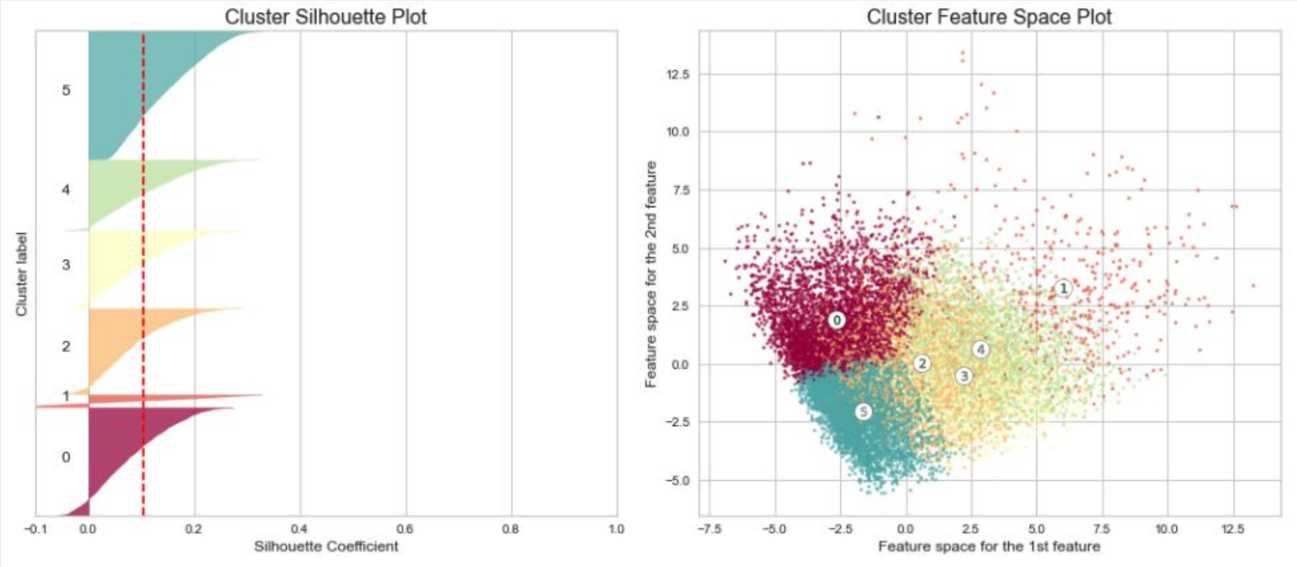
Avg. Silhouette Score for 6 Clusters: 0.1043

Silhouette Coefficients

+1: sample is far away from neighboring clusters

0: sample is on or very close to decision boundary between neighboring clusters

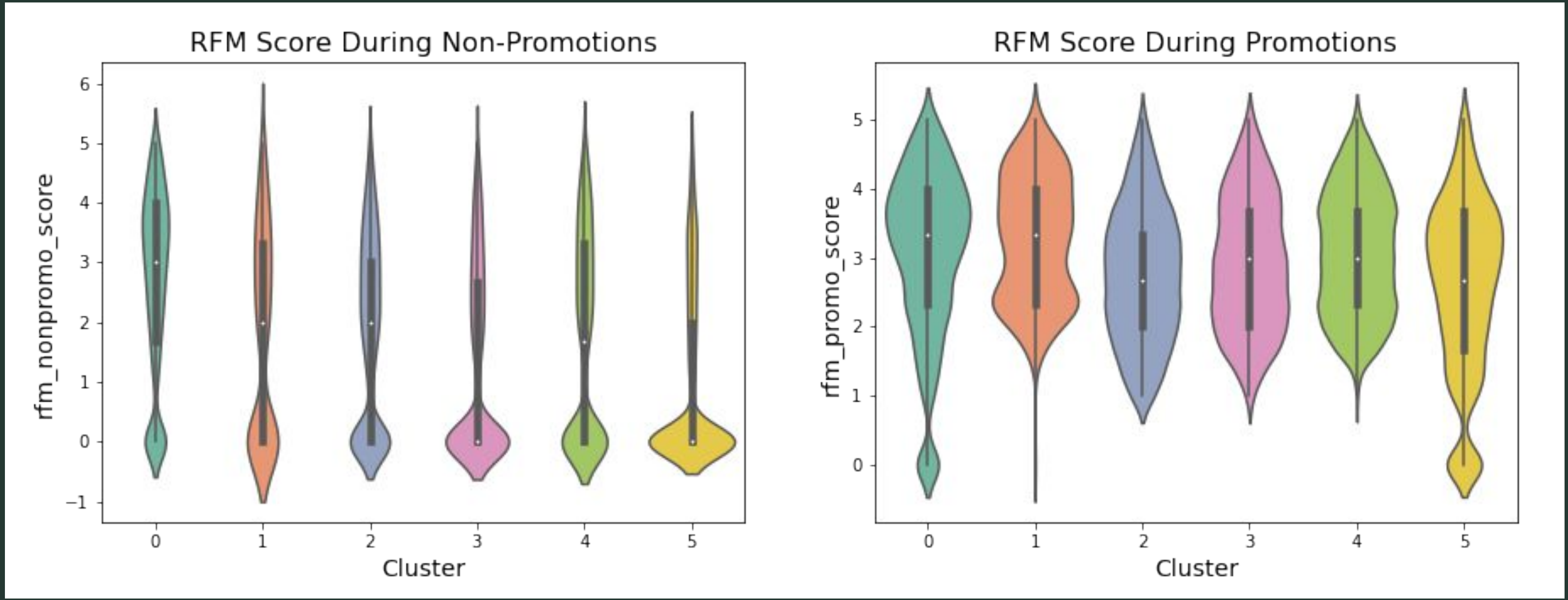
-1: samples might have been assigned to the wrong cluster



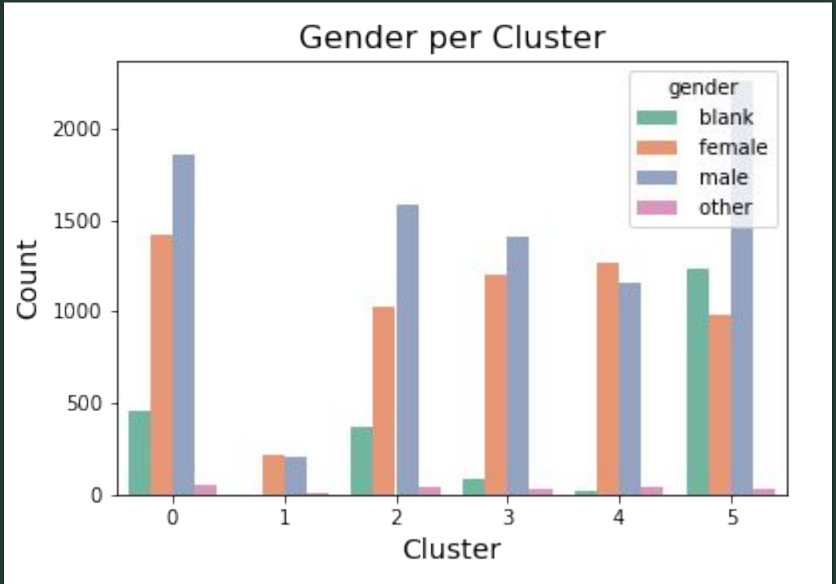
Visually inspected the separation distance between resulting clusters for all combos of features

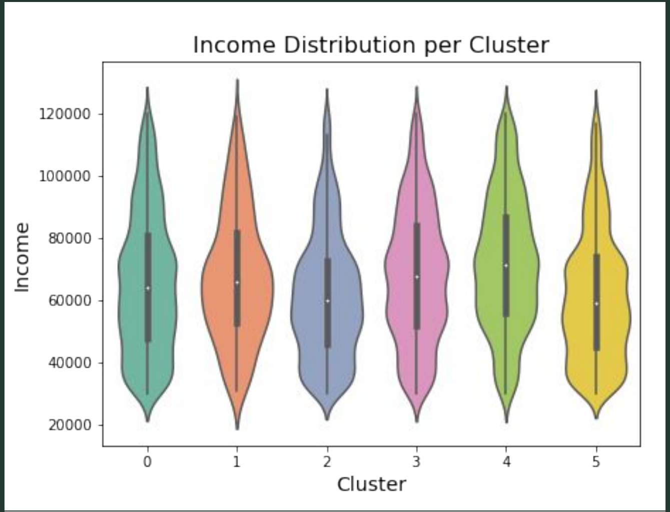
RFM Metrics

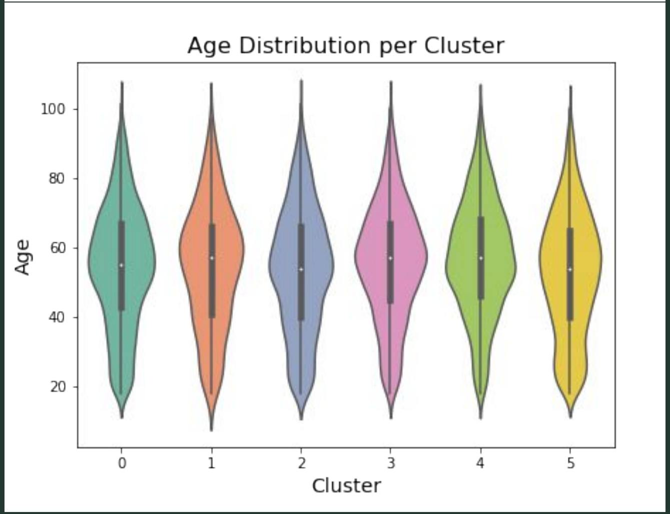
R = **Recency** (days since last purchase)
F = **Frequency** (number of purchases)
M= **Monetary** (total $ spent by customer)



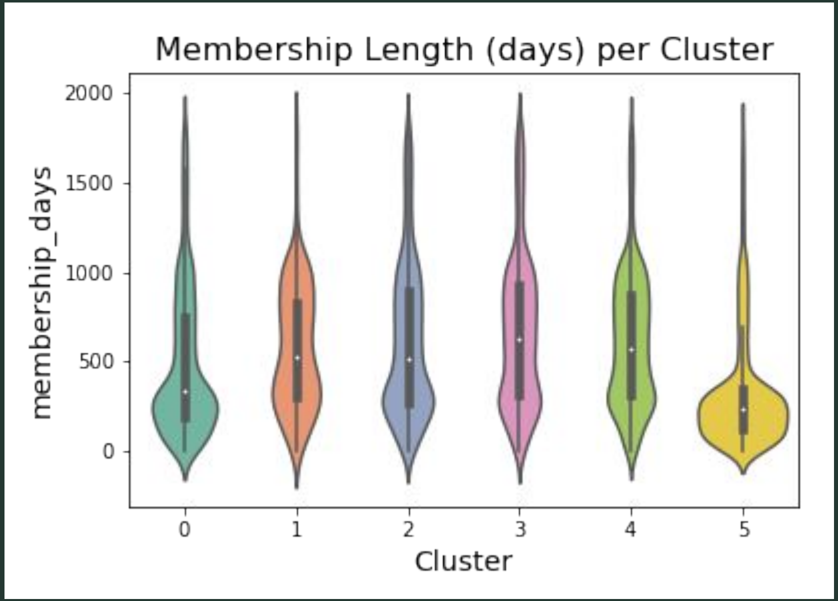
Demographics per Cluster

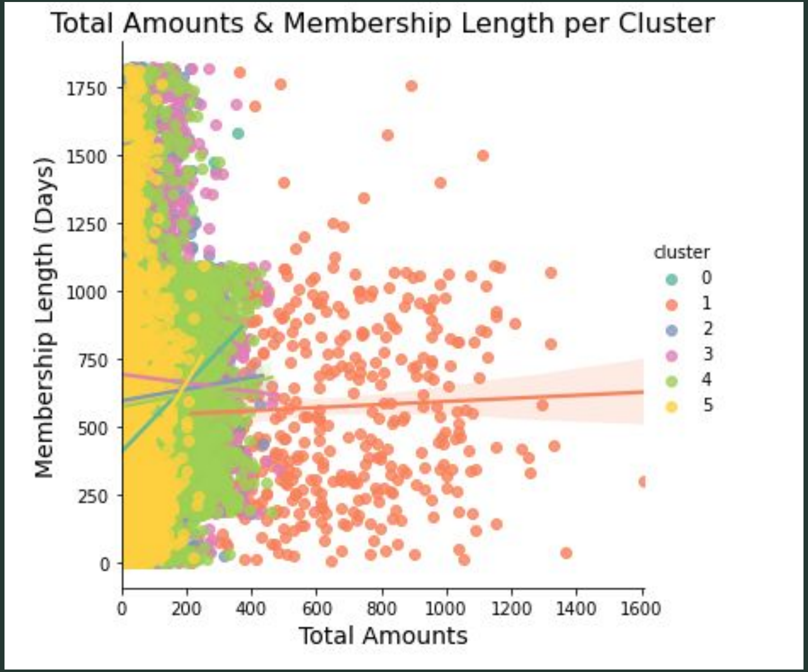




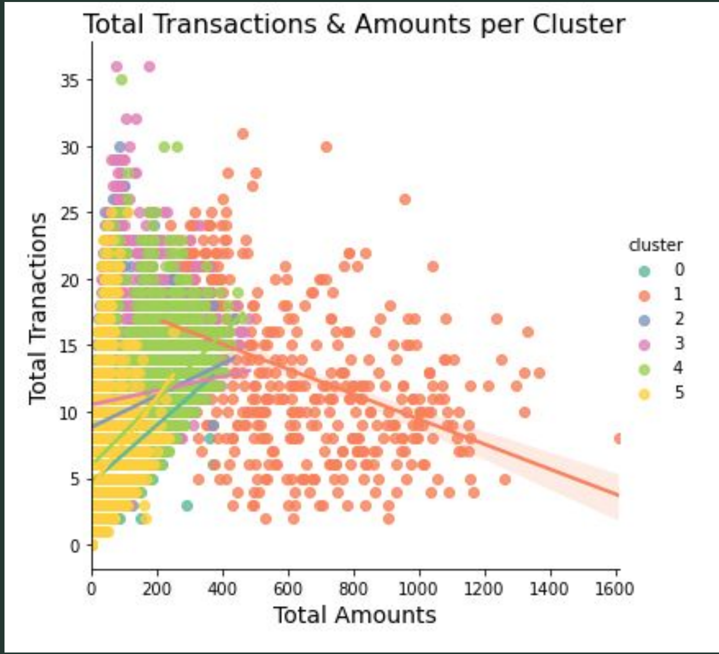


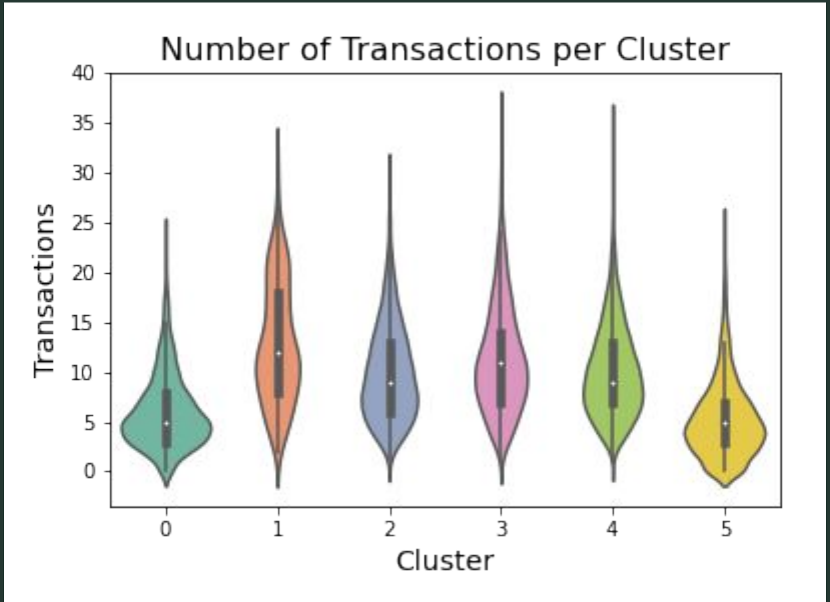
Membership Length per Cluster



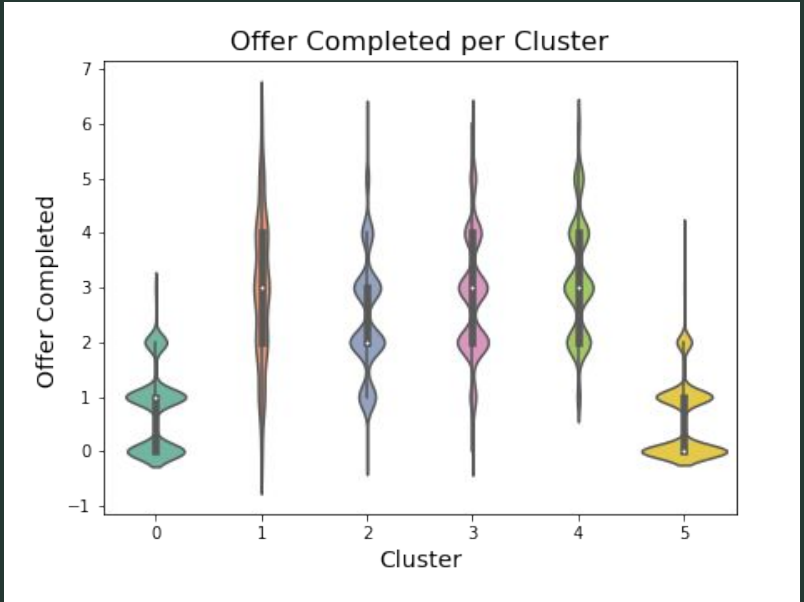


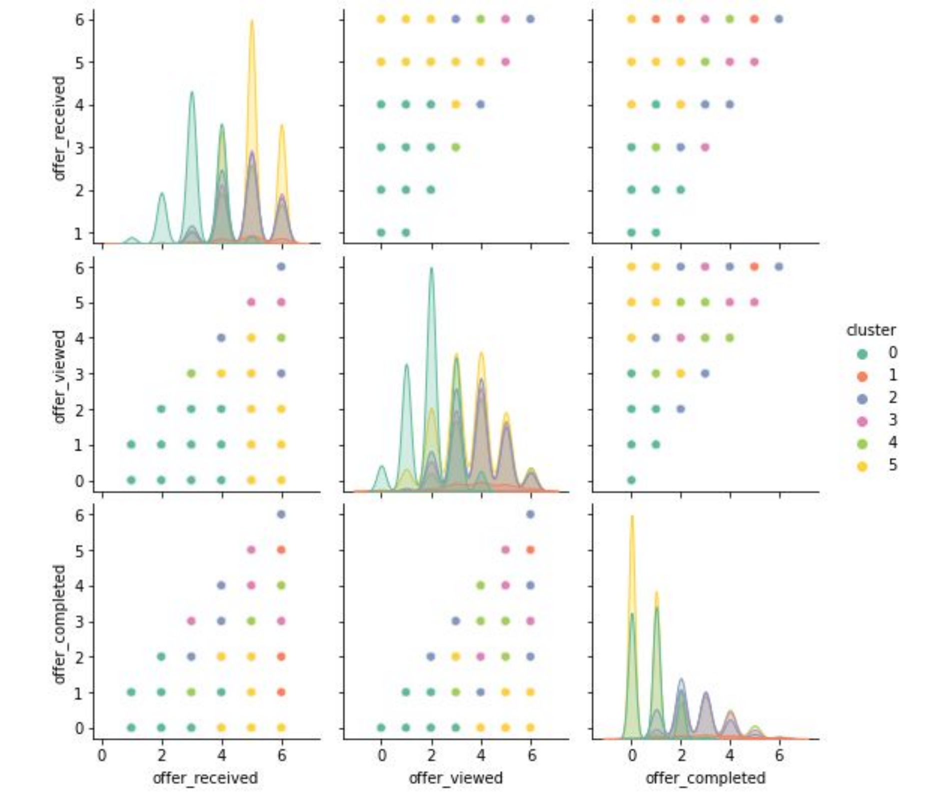
Transactions per Cluster





Offers per Cluster





6 Customer Segments

Segment 2
- Highest response to informational offers
- Shorter durations, lower difficulty & lower reward offers unless they're informational

Segment 4
- Lower number of overall transactions
- Higher response to BOGO offers
- Higher RFM scores during promos but not highest
- Highest avg and median incomes
- Oldest avg and median ages
- Only segment with more women
- Most likely to redeem rewards

Segment 0
- Most valuable during Non-promos
- More transactions during non-promos than promos which stood out
- Least likely to view offers

Segment 1
- Most valuable during Promos
- Highest amt per transaction
- Smallest segment - Only 2.6 % of customers
- Higher incomes

Segment 3
- Highest response to discount offers
- Lower RFM scores during non-promos
- Customers with the longest memberships
- Higher incomes but not highest
- Higher difficulty offers

Segment 5
- Longer duration promos
- Lowest number of transactions
- Least valuable in both non-promos and promos
- More recent customers
- Lowest avg and median incomes
- A lot of the people who left gender, income and age blank

Recommendations

- **Segment 0:** Don't bother with many promotions.
- **Segment 1:** Do well with promos, and have high potential with their incomes and highest $/trans. but they only represent small % of customers.
- **Segment 2:** Informational offers - if targeting them with a different type of promo make it shorter duration, lower difficulty, but offer a lower reward.
- **Segment 3:** Discount offers - Loyalty in terms of membership lengths, feel free to send them more difficult offers to complete but with higher rewards.
- **Segment 4:** BOGO offers but generally have lower # of trans. possibly because of this. Older members with higher incomes, they are likely to redeem rewards so choose difficulty and reward levels accordingly.
- **Segment 5**: Recent members (<1 year) and not a lot going on. Possibly need more time & demographic info (left a lot of this blank) to better understand their spending habits..

Final Insights

- Clusters are separated by relevant metrics for related marketing decisions to be made with each of the segments by Starbucks.
- Gender, Income and Age seemed to have less of a direct impact compared to amount $ spent, number of transactions, promotional offers, and length of membership.
- This work highlighted that the bulk of a data science problem is often the data cleaning and wrangling.
- There is longer term potential for this project to keep developing these customer segments and iterating through this workflow with even more customer data.

THANK YOU!

Questions?

Emily Siegel
LinkedIn
GitHub

References

1. Data Source: https://www.kaggle.com/ihormuliar/starbucks-customer-data
2. https://seifip.medium.com/starbucks-offers-advanced-customer-segmentation-with-python-737f22e245a4
3. https://www.barilliance.com/rfm-analysis/
4. https://formation.ai/blog/how-starbucks-became-1-in-customer-loyalty/
5. https://lifetimes.readthedocs.io/en/latest/
6. https://www.natasshaselvaraj.com/customer-segmentation-with-python/
7. https://towardsdatascience.com/find-your-best-customers-with-customer-segmentation-in-python-61d602f9eee6
8. https://www.datacamp.com/community/tutorials/introduction-customer-segmentation-python
9. https://clevertap.com/blog/rfm-analysis/