CUSTOMER SEGMENTATION ON HOJO MOJO – A COFFEE SHOP

Group Members:

Abirami Balasubramayian

Dhinakar

Sai Siddharth

Shruthi Suresh

Surya KP

Vijaya Bharathi

PROBLEM STATEMENT

- HOJO MOJO a coffee chain in the New York city , observes that there is a frequent fluctuation in sales week by week and also with the number of people who visit the cafe .
- Objective 1: To help the owner in identifying the target customers by analysing customer behaviour who visit the cafe in order to retain the existing customers and attract new ones which would help in consistency of sales .
- Objective 2: To identify the key features that are responsible for the classification of customers.
- **Techniques :**

  EDA(Data visualization and Data pre-processing),

  Statistical Analysis,

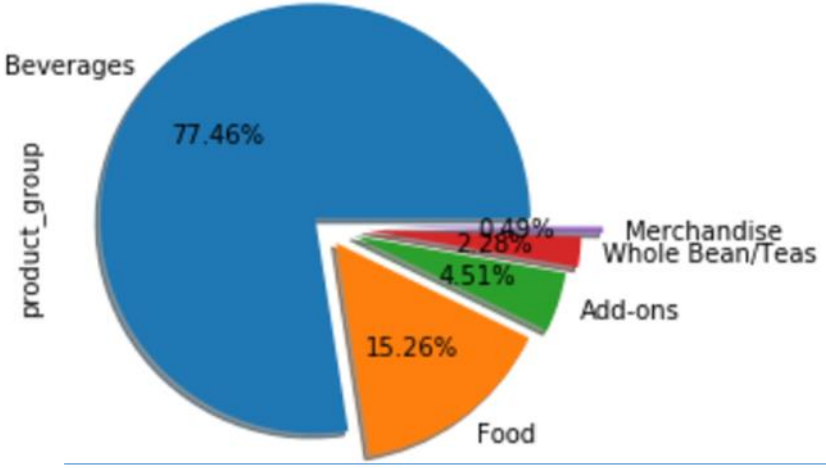
  Unsupervised Machine Learning,

  Supervised Classification Machine Learning algorithm

STATISTICAL ANALYSIS

greatlearning
Learning for Life

Product Segmentation



- Major product sold – Beverages
- Sales percentage – 77.46%

- Minimum Sale Percentage – 0.49%
- Product - Merchandise

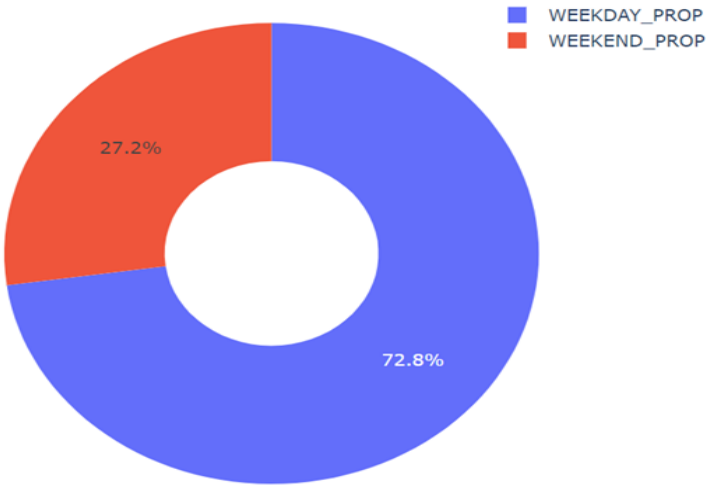
Most and Least Selling Products

- **Business or Marketing Strategy :**
  to increase profit or sale of the least selling product - exciting COMBO OFFERS!

STATISTICAL ANALYSIS

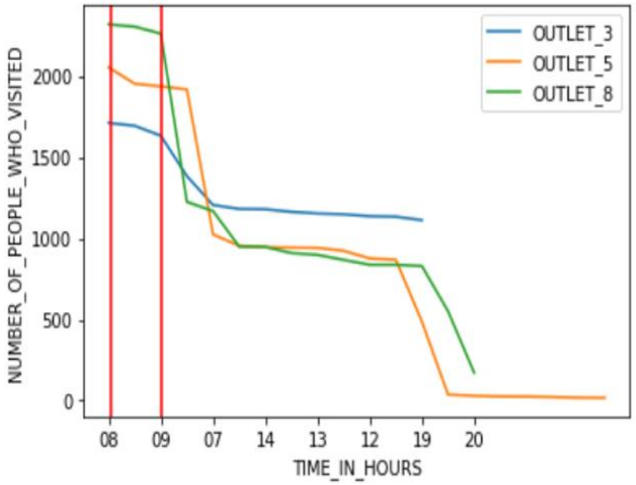
greatlearning
Learning for Life

Sales Analysis



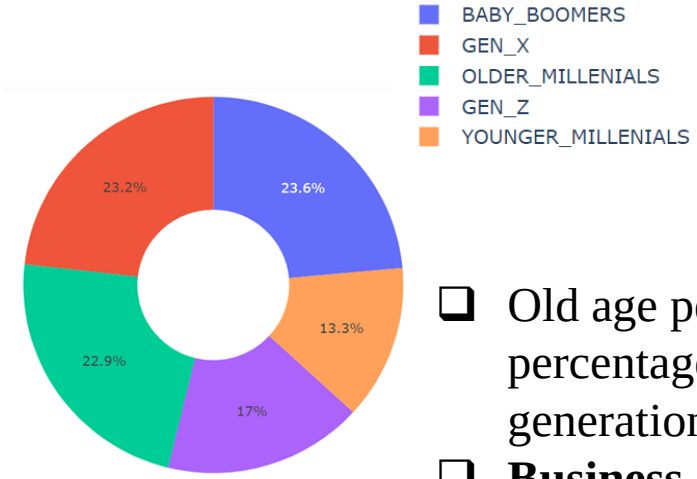
- No. of People visited during Weekday are higher than weekend

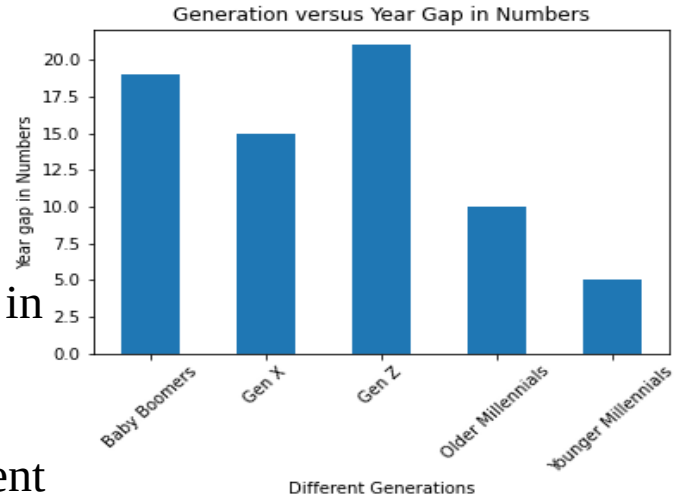
Peak Hour Analysis



- During morning at 8am and 9am(prime time)

Proportion of People among Different Generations





- Old age people are higher in percentage than younger generation people
- **Business Strategy :** Student coupons

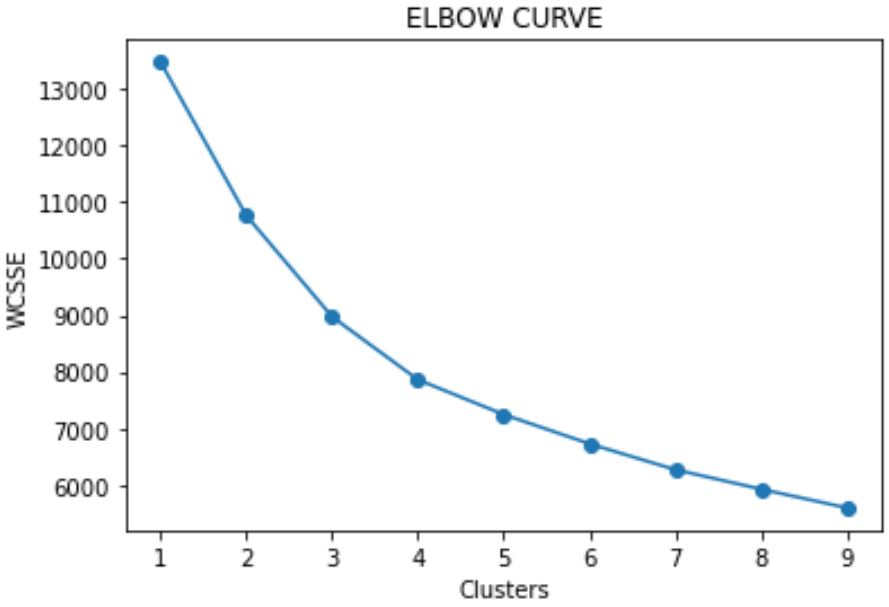
FEATURE ENGINEERING

- Feature 1: “Quantity Purchased”
  - Group by customer
  - Sum of the product purchased by each customer during the month of April 2019.
- Feature 2: “Frequency”
  - Group by customer
  - Count the number of times they visited the café.
- Feature 3: “Recency”
  - Difference between most recent visit of the customer and 1st May2019.
- Feature 4: “Customer_Since”
- Feature 5,6,7,8: “Weekly Purchase”
  - Group by customer
  - Sum of amount spent
- Feature 9: “Star purchase”
  - Group by customer
  - Maximum amount spent by the customer during their visit in the month of April 2019.
- Feature 10: “Preferred product type”
  - Group by customer
  - Most frequently purchased product.

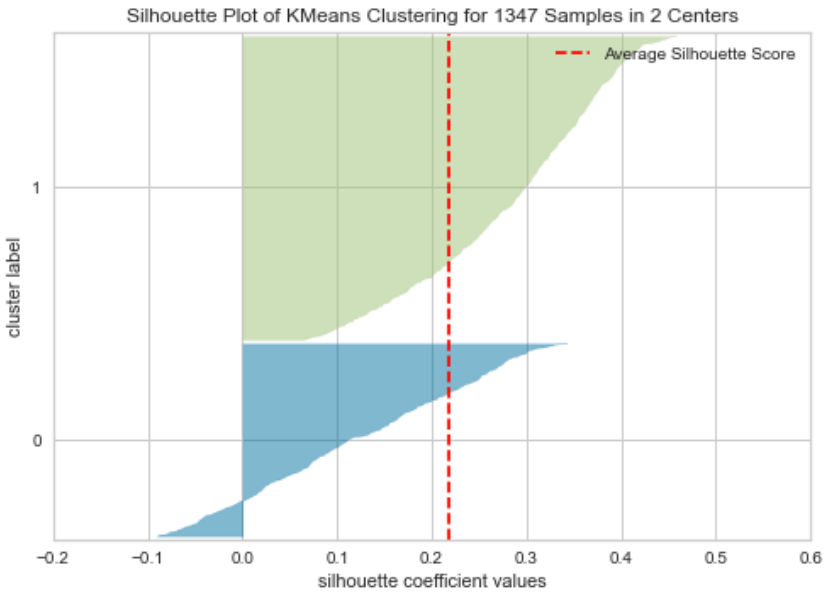
CLUSTERING MODEL

greatlearning
Learning for Life

ELBOW Curve

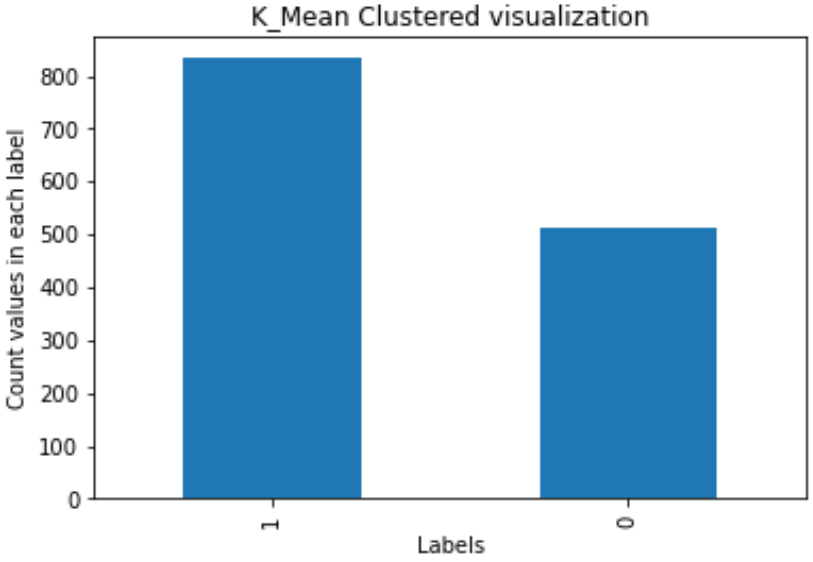


Silhouette Visualizer



- Unsupervised Machine Learning model
- Centroids plays a major role
- Distance Based Computation
- Metrics to find number of Cluster:
  - ELBOW curve
  - Silhouette Visualizer

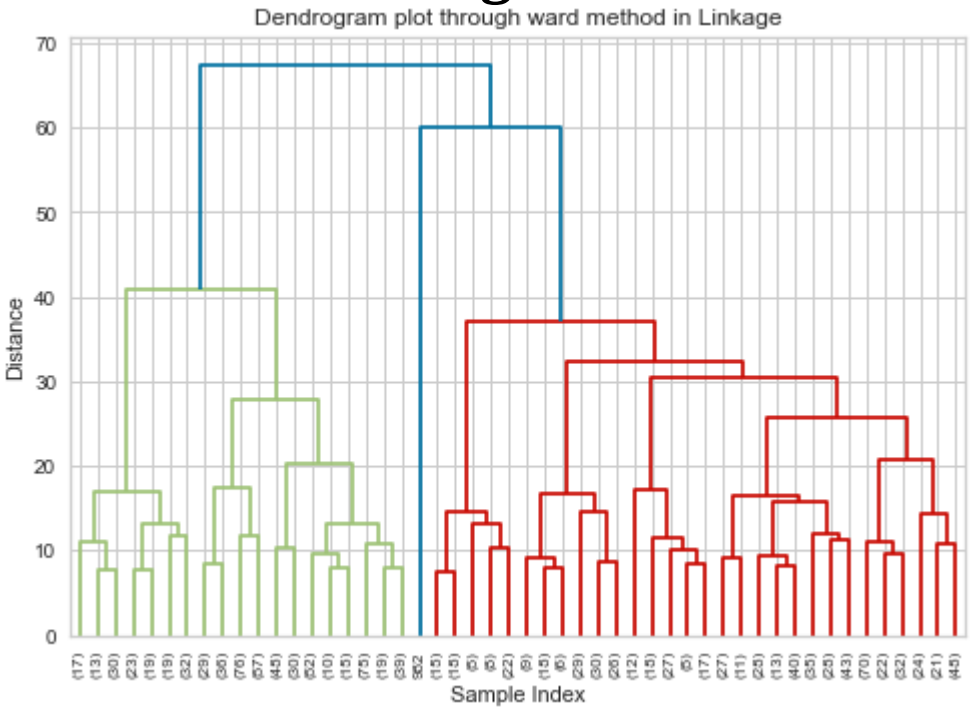
K_Mean clustering



CLUSTERING MODEL

greatlearning
Learning for Life

Dendrogram Plot

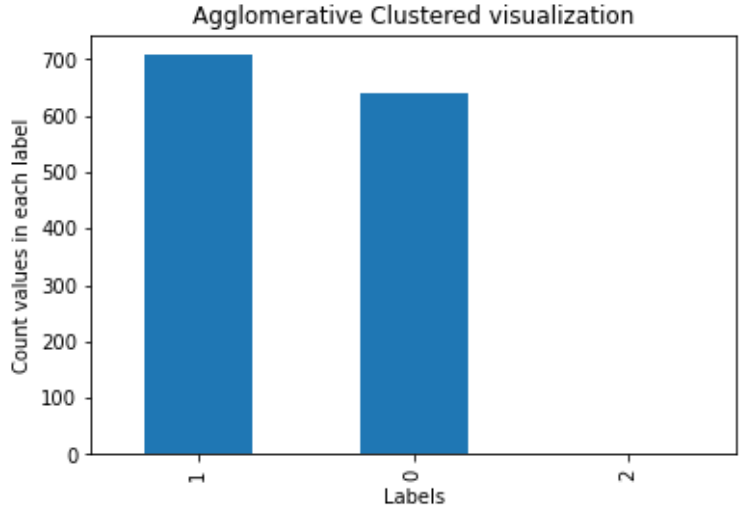


"n" Calculation

```
The number of cluster for the distance of 40 is 4
The number of cluster for the distance of 42 is 3
The number of cluster for the distance of 44 is 3
The number of cluster for the distance of 46 is 3
The number of cluster for the distance of 48 is 3
The number of cluster for the distance of 50 is 3
The number of cluster for the distance of 52 is 3
The number of cluster for the distance of 54 is 3
The number of cluster for the distance of 56 is 3
The number of cluster for the distance of 58 is 3
The number of cluster for the distance of 60 is 2
The number of cluster for the distance of 62 is 2
The number of cluster for the distance of 64 is 2
The number of cluster for the distance of 66 is 2
The number of cluster for the distance of 68 is 1
The number of cluster for the distance of 70 is 1
The number of cluster for the distance of 72 is 1
The number of cluster for the distance of 74 is 1
The number of cluster for the distance of 76 is 1
The number of cluster for the distance of 78 is 1
```

Agglomerative clustering

- Unsupervised Machine Learning model
- Hierarchal Method:
  - Top_Down approach
  - Bottom_Up approach
    - Agglomerative Clustering
- Drawback:
  - Computation is heavy



CLASSIFICATION MODEL

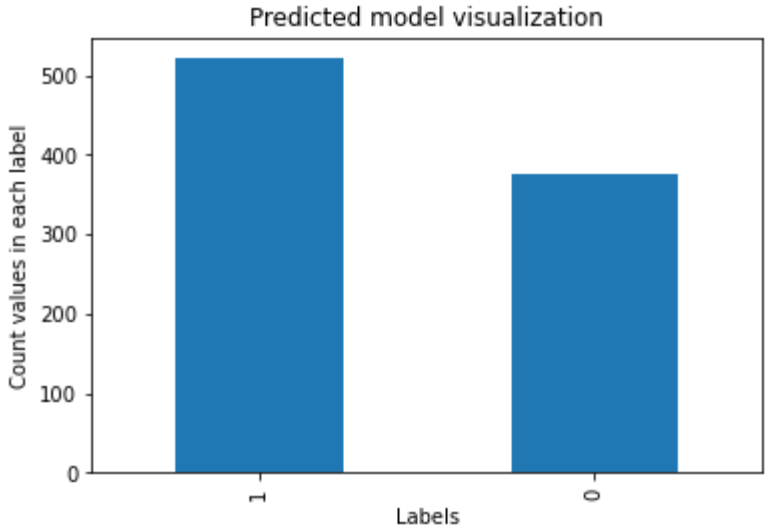
greatlearning
Learning for Life

- Why Classification not Regression?
  - Target : Nature is discrete
- Algorithms used :
  - Decision Tree
  - Random Forest
  - Gaussian Naïve Bayes
  - K_Neighbors
  - Logistic Regression
- Metrics to analyse model performance:
  - Precision
  - Recall
  - F1_Score
  - Accuracy

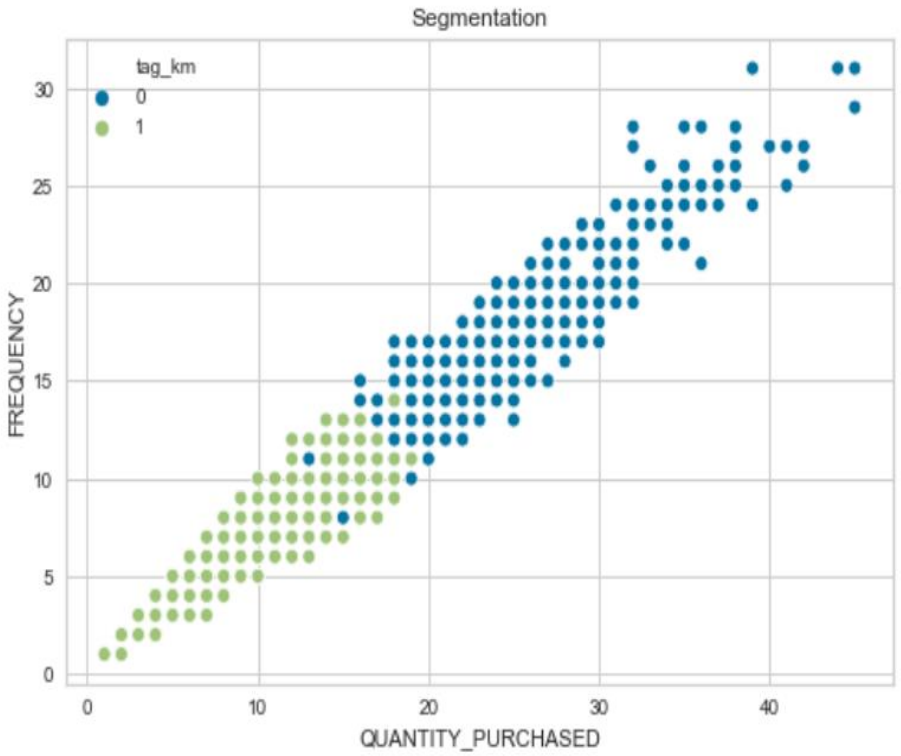
ALGORITHM	ACCURACY _SCORE	CO.EFF OF VARIANCE ERROR
DT	97.01	0.07
RF	97.21	0.09
GNB	95.34	1.4
LR	91.07	4.3
KNN	91.25	1.6

Classification Report

	precision	recall	f1-score	support
0	0.92	0.94	0.93	150
1	0.96	0.95	0.96	255
accuracy			0.95	405
macro avg	0.94	0.95	0.94	405
weighted avg	0.95	0.95	0.95	405



INFERENCES



Quantity_Purchased has strong relationship with Frequency of Purchased

Feature Importance

	Features	Feature_importance
0	QUANTITY_PURCHASED	0.443033
1	FREQUENCY	0.301927
4	WEEK1_AMOUNT_SPENT	0.069187
5	WEEK2_AMOUNT_SPENT	0.064747
7	WEEK4_AMOUNT_SPENT	0.028221
8	STAR_PURCHASE	0.025563
2	RECENCY	0.020356
6	WEEK3_AMOUNT_SPENT	0.019214
9	PRODUCT_PREFERENCE	0.014050
3	CUSTOMER_SINCE	0.013703

Quantity_Purchased plays major role on customer_segmentation

BUSINESS INSIGHTS

- To retain the Loyal customers - more offers and discounts (on special occasions like birthdays , anniversaries)
- Cluster 0 : not loyal customers
    - Solution to be a frequent visitors - free samples of products .
    - “Catchy cash back offers” along with terms and conditions
- To attract new customer - varieties in the menu can be increased.
- Proportion of “gen-z” and “Younger Millennials” are less compared to the older generation .
    - Students coupons – attract younger generation people.

CONCLUSION

- A unsupervised machine learning algorithm was build using K_Mean in order to analyze the customer behavior of people who visit HOJO MOJO.
- The transaction level data was converted to customer level data through feature extraction.
- A classification model was build in order to understand the key features that played a role in customer segmentation.

LIMITATION AND FUTURE SCOPE

- Drawback : Only one month outlet data is present.
- Future Scope:
    - Time-Series
    - Recommendation System

THANK YOU