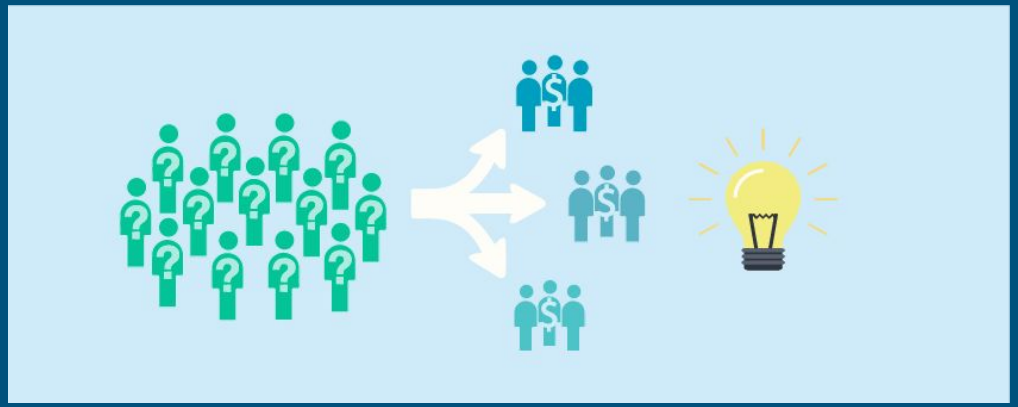
Customer Segmentation Using their spending personality

Nidhi Bodar, Foram Shah, Reem Ghabayen

Introduction

- Understanding the interaction patterns between online companies and their customers is important in the face of this online market growth. To develop personalized products and services, it is essential to differentiate between different types of customers
- The problem intends to develop segments that match individual customers' personalities based on their spending for each product in a particular period, whether promotion changes their purchasing power or not.



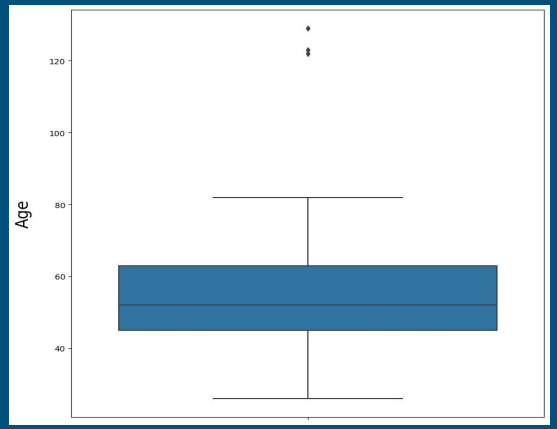
Data Set

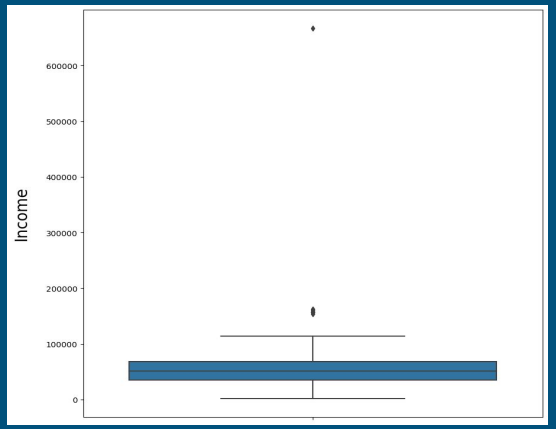
- The company's dataset fetched from the Kaggle website. With 2240 rows and 29 different features

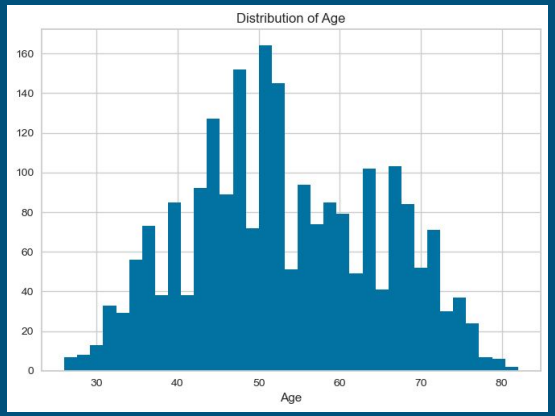
	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	...	NumWebVisitsMonth	AcceptedCmp3	Acce
0	5524	1957	Graduation	Single	58138.0	0	0	04-09-2012	58	635	...	7	0	
1	2174	1954	Graduation	Single	46344.0	1	1	08-03-2014	38	11	...	5	0	
2	4141	1965	Graduation	Together	71613.0	0	0	21-08-2013	26	426	...	4	0	
3	6182	1984	Graduation	Together	26646.0	1	0	10-02-2014	26	11	...	6	0	
4	5324	1981	PhD	Married	58293.0	1	0	19-01-2014	94	173	...	5	0	

5 rows × 29 columns

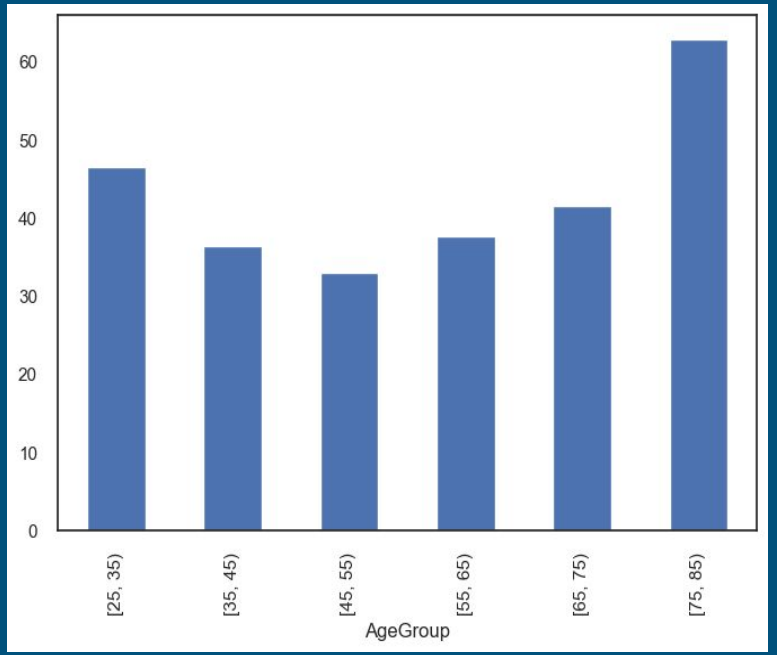
Data Preprocessing



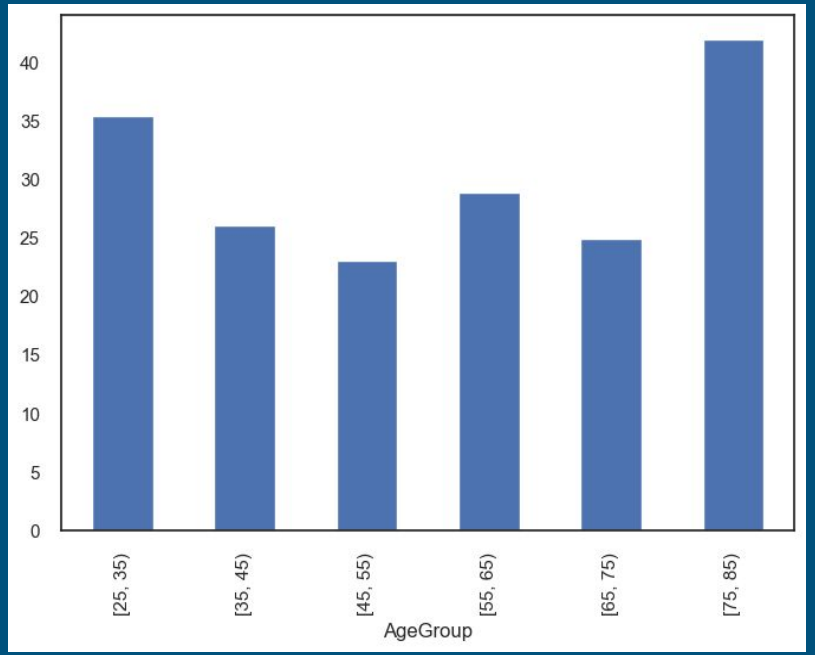




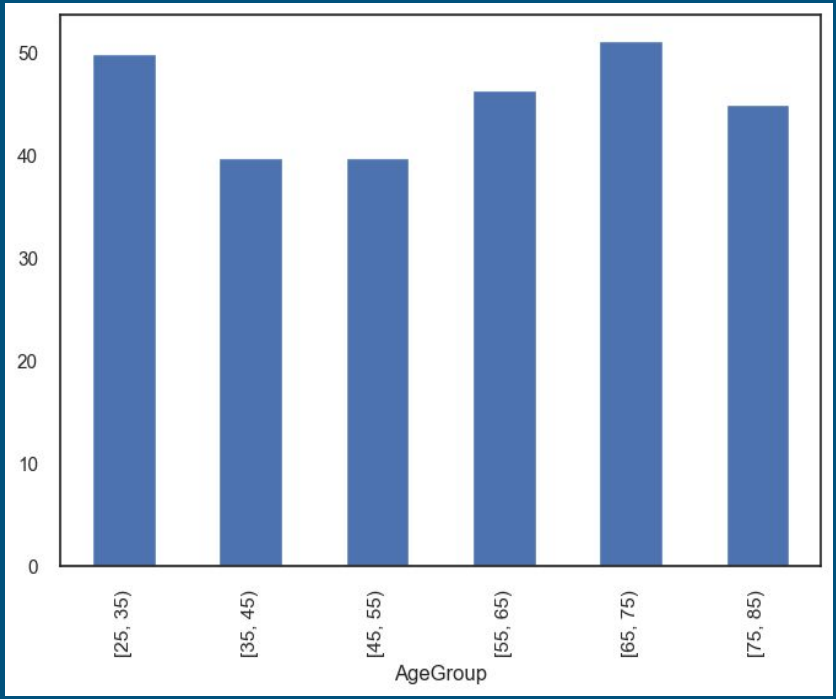
fish&agegroup



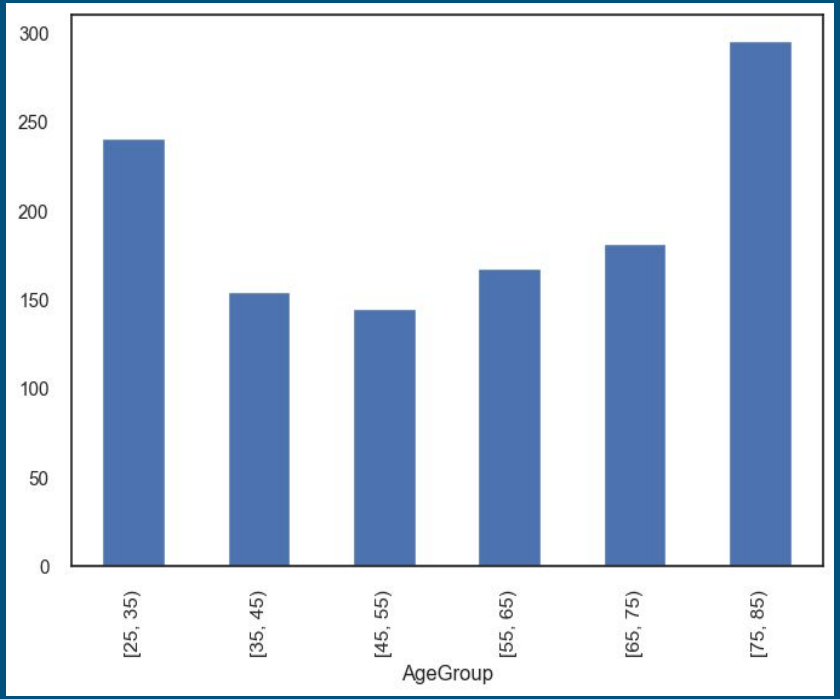
Fruits&Agegroup



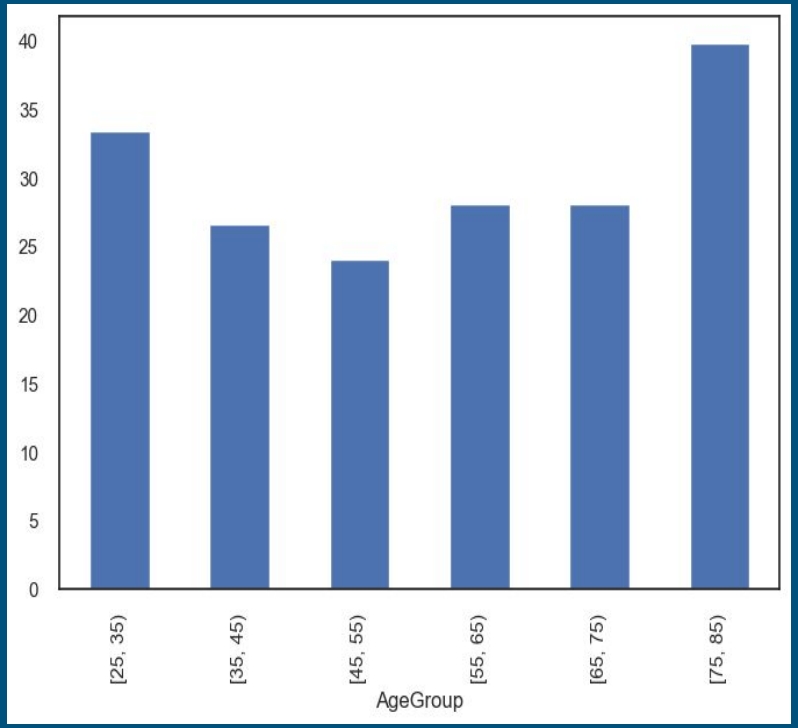
Gold&agegroup



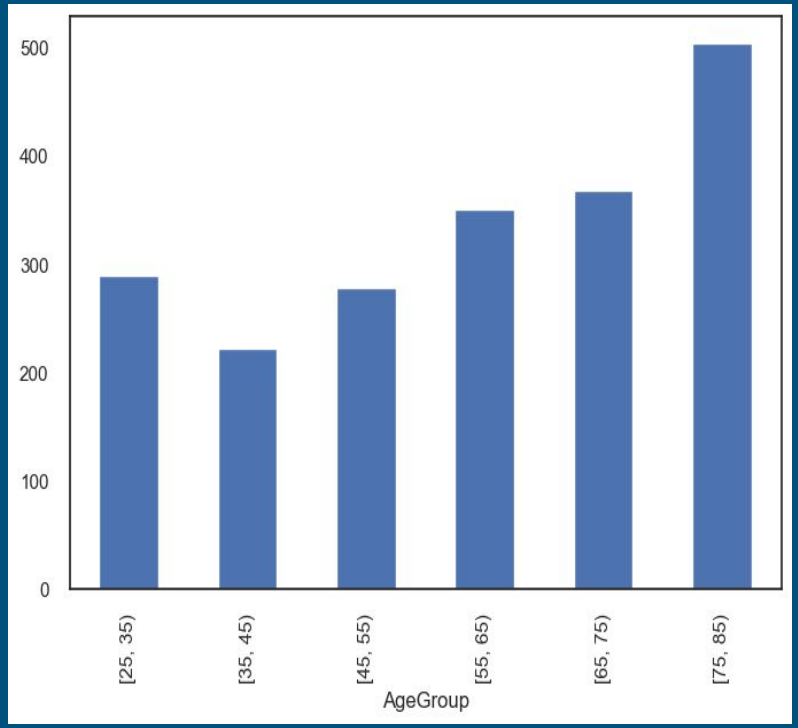
meat&agegroup



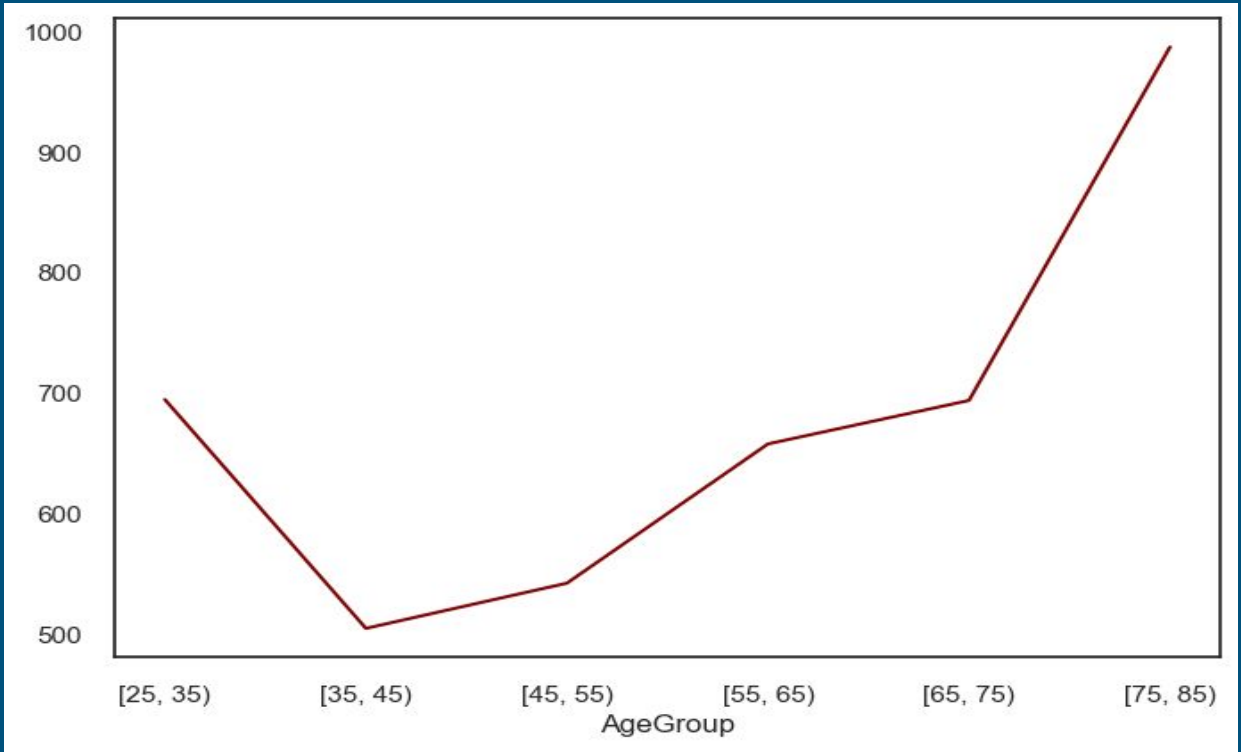
Sweets&agegroup

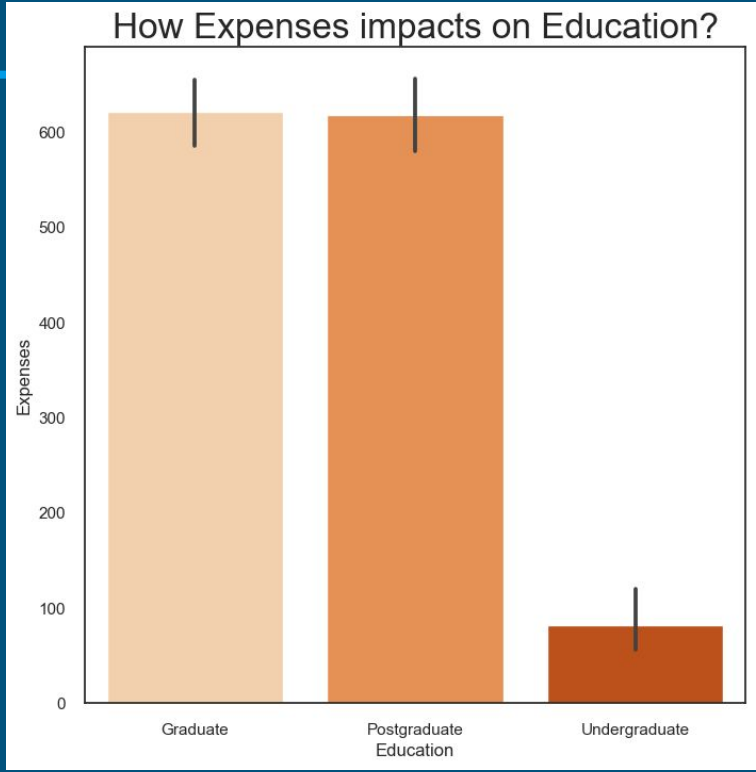


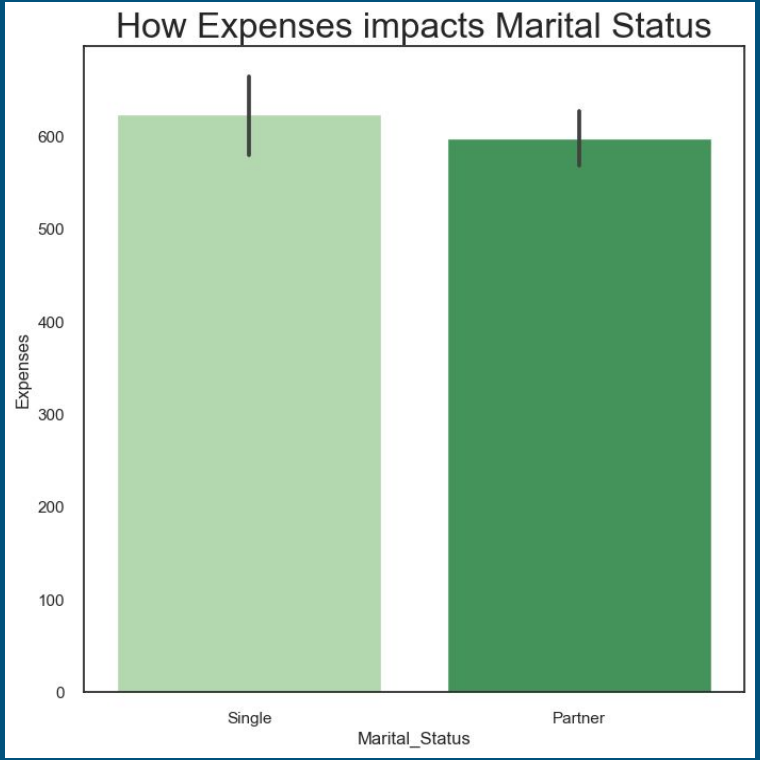
wine$agegroup

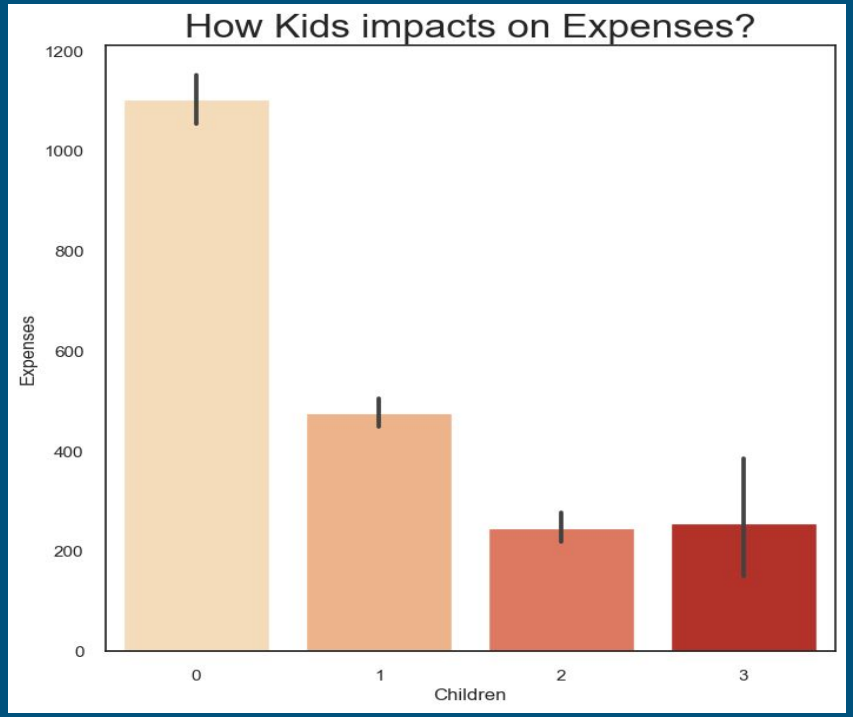


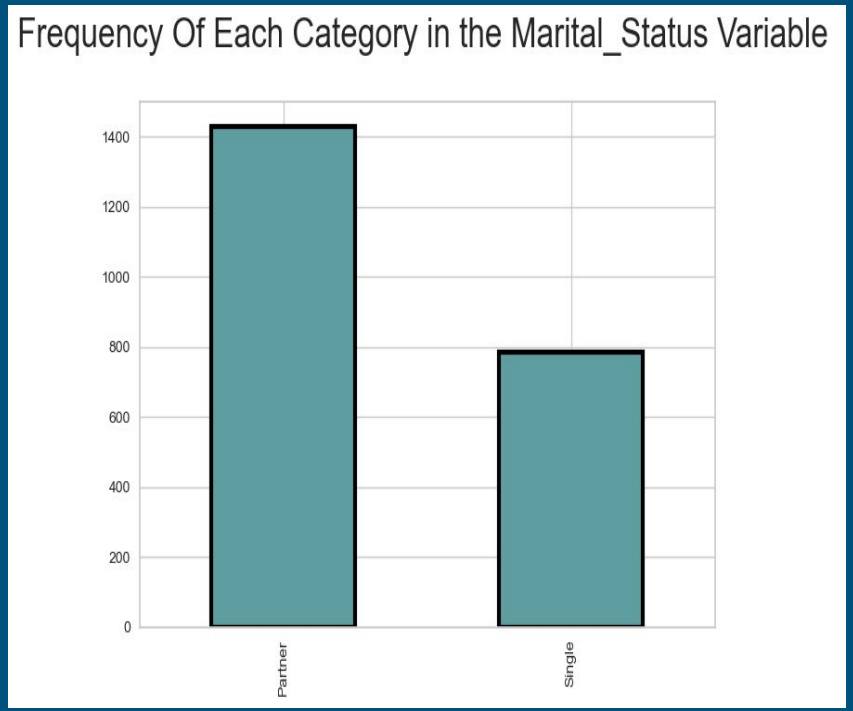
Agegroup & Expenses

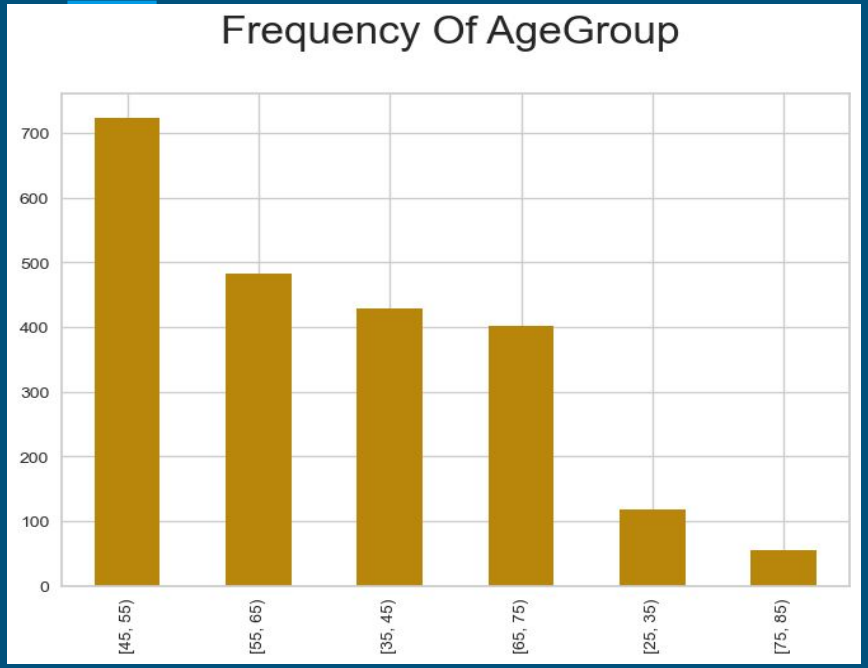


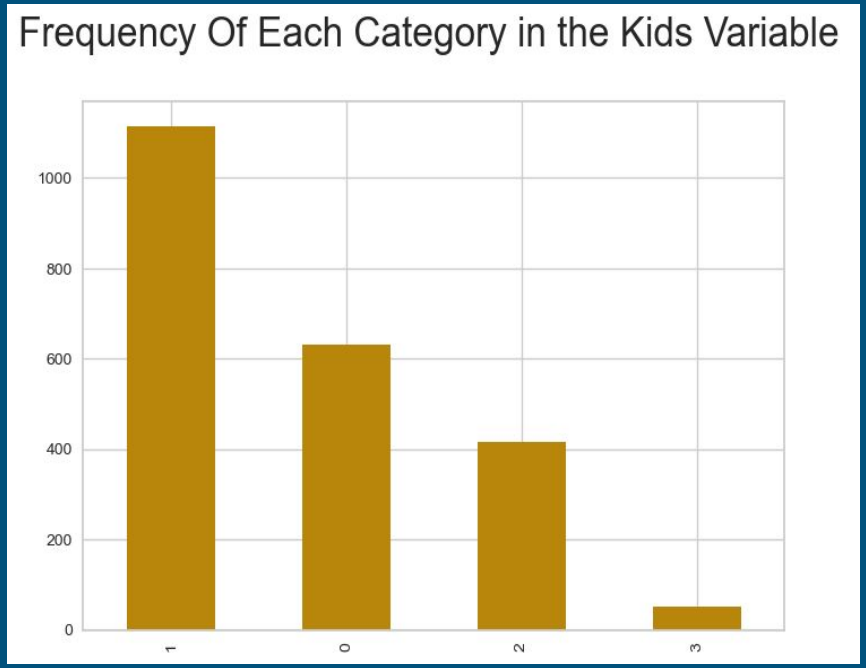

How Expenses impacts on Education?
Expenses
0
100
200
300
400
500
600
Graduate
Postgraduate
Undergraduate
Education

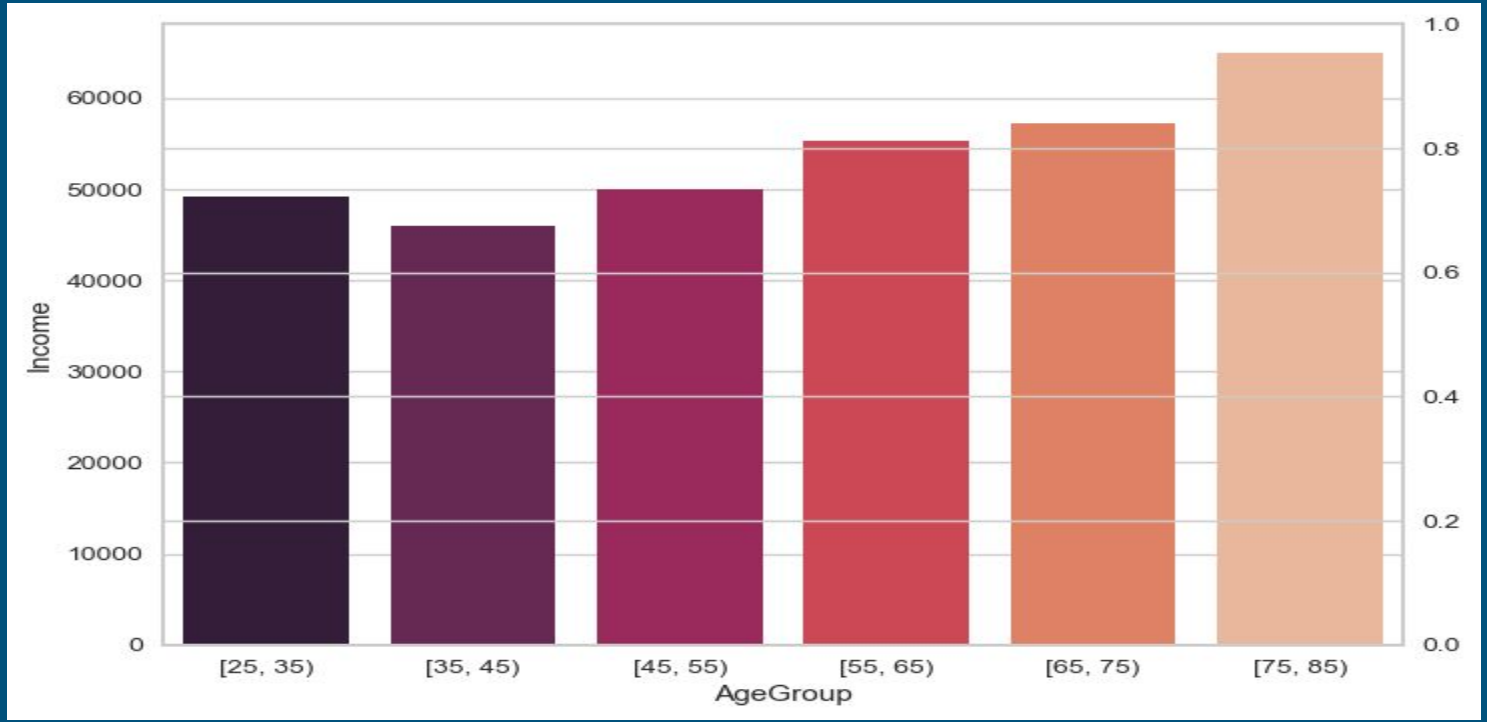

How Expenses impacts Marital Status
Expenses
0
100
200
300
400
500
600
Single
Partner
Marital_Status


How Kids impacts on Expenses?
1200
1000
800
600
400
200
0
Expenses
0
1
2
3
Children


Frequency Of Each Category in the Marital_Status Variable
1400
1200
1000
800
600
400
200
0
Partner
Single

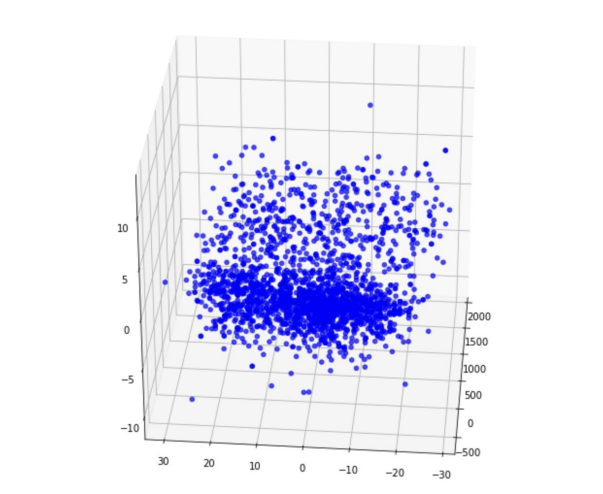

Frequency Of AgeGroup
700
600
500
400
300
200
100
0
[45, 55)
[55, 65)
[35, 45)
[65, 75)
[25, 35)
[75, 85)

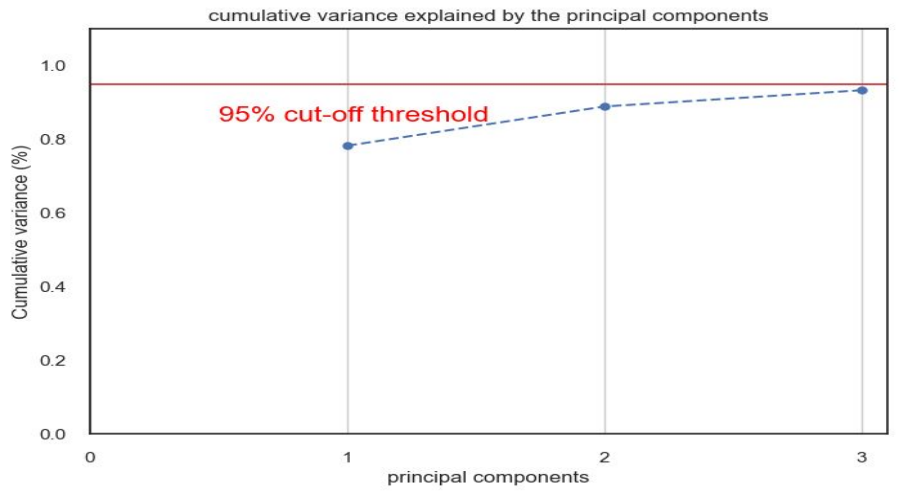

Frequency Of Each Category in the Kids Variable
1000
800
600
400
200
0
1
0
2
3


Income
60000
50000
40000
30000
20000
10000
0
1.0
0.8
0.6
0.4
0.2
0.0
[25, 35)
[35, 45)
[45, 55)
[55, 65)
[65, 75)
[75, 85)
AgeGroup

Dimensionality Reduction

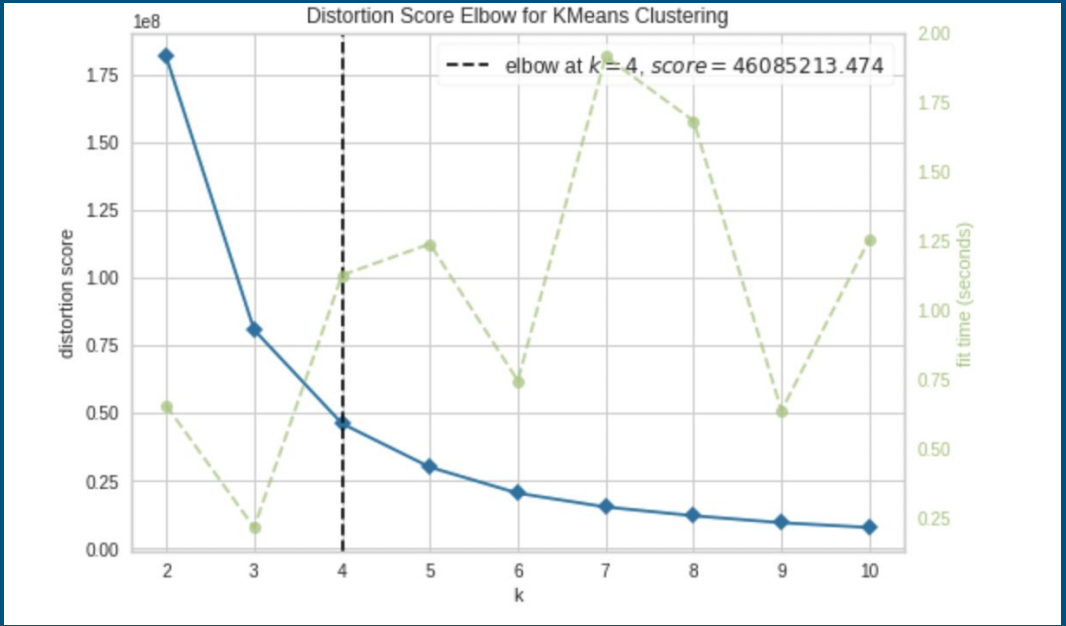
As the dataset consists of 29 features, Dimensionality Reduction techniques such as PCA was used to reduce it.





Elbow Method

Elbow Method was Used Pre K means to determine the optimal Number of Clusters : Optimal Was 4.



Clustering

Clustering Algorithms Used with their Silhouette and Calinski Results Compared::

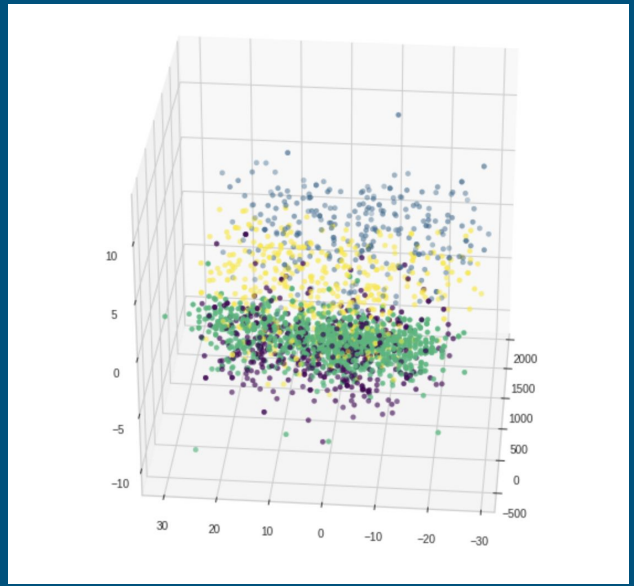
K Means (Had the Highest Scores)

Hierarchical Clustering

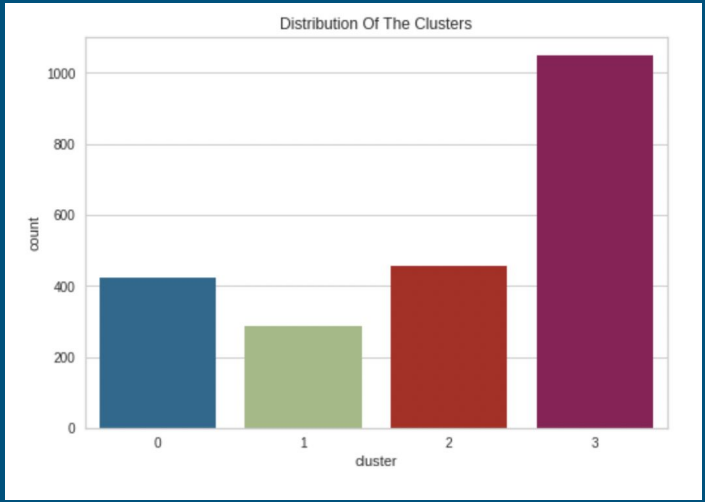
GMM

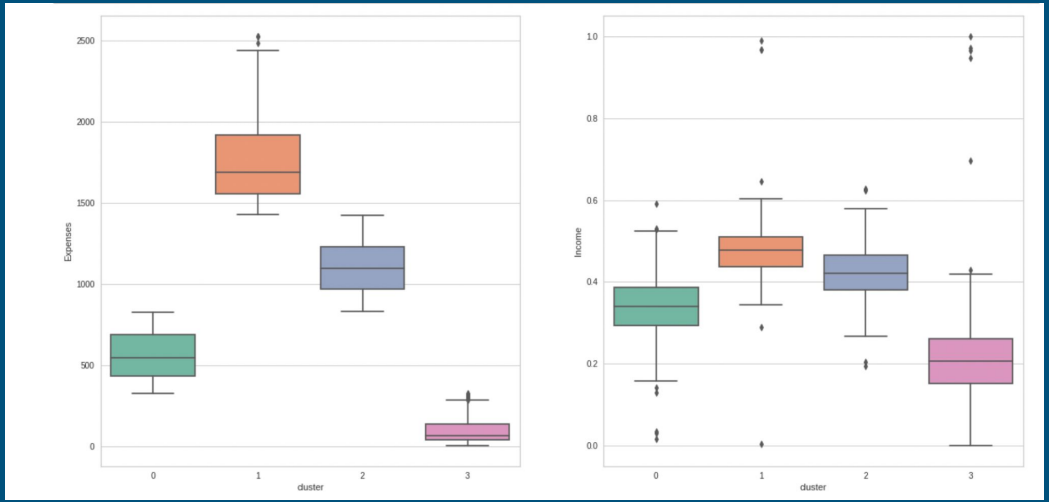
DBSAN

Fast Clustering Using Cosine Similarity



Modeling Clusters





Profiling the Clusters

Cluster 0 : (Lowest Expenses In general, density Very High, lowest Income)

Definitely a parent with more number of kids at home(teen and children,, On the older Age group , Majority have a partner, Above Undergraduates Mostly old Customers

Cluster 1: (Highest Expenses in general, Density Very sparse, Highest Income)

One to Zero Kids at Home, average age group, mostly have a partner, No undergraduates, Majority Old Customers

Cluster 2:(Second Highest Spending Group)

One to Zero Kids, Average Age group, Mostly have a partner , Mostly Graduate, Zero Undergraduates, Average Customers

Cluster 3:(Second to last spending group)

Majority 1 Kid at Home, Average age group, Mostly Have partner, half having graduate degree half with no degree, Average Customers

Variable Percentage Allocation Per Cluster

index	cluster	variable	Mean_Column	percentage
0	0	Year_Birth	1970.9083175803403	47.878473662856926
1	0	Income	0.21088261513972417	32.2637700751864
2	0	Recency	0.4950831567088657	47.82396178214718
3	0	MntWines	0.030379958394372224	7.106217125527177
4	0	MntFruits	0.025044884156130365	9.05375937913154
5	0	MntMeatProducts	0.013596887756499821	6.7163775137358925
6	0	MntFishProducts	0.02748319477998117	9.043096098656324
7	0	MntSweetProducts	0.020512561508824082	9.50407006869808
8	0	MntGoldProds	0.04531267482877821	15.838170515828908
9	0	NumDealsPurchases	2.114366729678639	43.504472967716836

Conclusion

- We used unsupervised clustering in this project.
- We used the elbow technique to determine the optimal number of clusters
- K Means, DBSCAN, GMM, and hierarchical clustering utilizing PCA's dimensionality reduction.
- We selected the Kmeans variation to develop a quicker clustering method that allows us to train huge datasets in exponentially less time with little accuracy loss.
- Further consumer profiling and EDA allowed us to determine their characteristics, such as expenses, incomes, and family groups, which helped the marketing organizations further divide the customers into groups and target the correct audience.