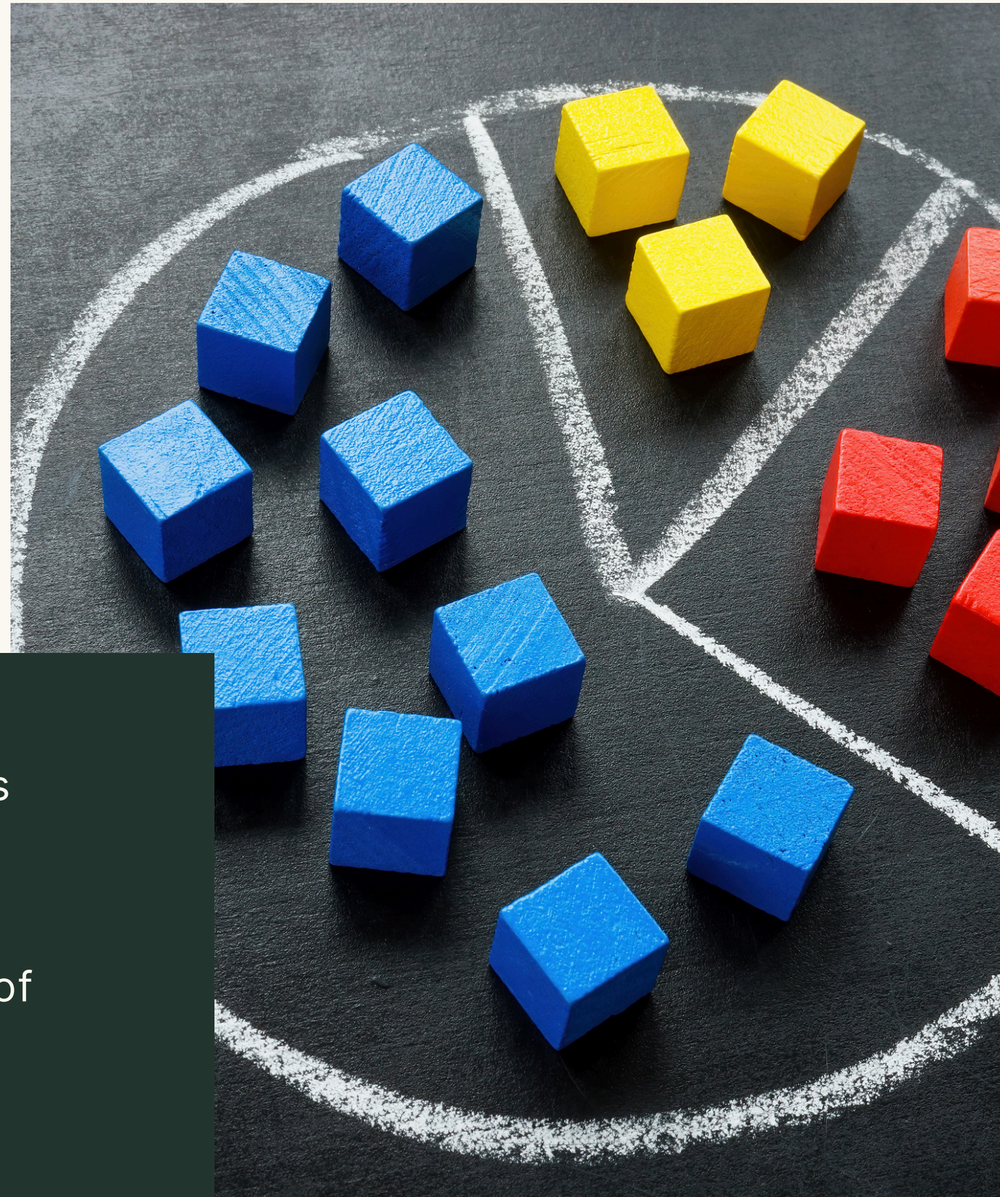
CUSTOMER CLUSTERING SEGMENTATION

INTRODUCTION



Customer segmentation is essential for many industries. it helps to segmentate the customer for easy management such as royal customer, the most active customer, occasional customer. Business can manage their marketing, sales and service for specific segment of customer.

OBJECTIVES

FIRST OBJECTIVE

To segment customer by their information and divide into distinct group

SECOND OBJECTIVE

To find the insights of each cluster and focus to manage relationship on specific segmentation

THIRD OBJECTIVE

To find the best model for Clustering

GET DATA

↓

DATA PREPARATION

↓

EXPLORATORY DATA ANALYSIS (EDA)

↓

DATA MODELING

↓

EVALUATION

GET DATA

THE DATA WAS COLLECTED BY

https://www.kaggle.com/datasets/akashdeepkuila/automobile-customer/data

THE DATA SET WAS CONSISTED BY THE COLUMN OF CUSTOMER INFORMATION FOLLOWING DETAILS :

CustomerID	Profession
Gender	WorkExperience
Married	SpendingScore
Age	Family_Size
Graduated	Category

	CustomerID	Gender	Married	Age	Graduated	Profession	WorkExperience	SpendingScore	FamilySize	Category
0	462809	Male	No	22	No	Healthcare	1.00	Low	4.00	Category 4
1	462643	Female	Yes	38	Yes	Engineer	NaN	Average	3.00	Category 4
2	466315	Female	Yes	67	Yes	Engineer	1.00	Low	1.00	Category 6

DATA PREPARATION

DATA VALIDATION FOR MODELING

- Total Data : 10695 Records
- Total Columns : 10 Columns
- Categorical Columns : 5 Columns
- Numerical Columns : 5 Columns
- Null values : 2108 values
- Duplicated values : 136 values

DATA CLEANSING

- Check any missing values and duplicated values
- Drop the Duplicated values
- Replace the null values by median and mode values
- Detect outlier values and remove

DATA PREPARATION

DATA TRANSFORMATION

- Drop the columns “CustomerID” and “Category”

```
cus_df = cus_df.drop(['CustomerID','Category'], axis=1)
```

- Drop the duplicated values

```
cus_df = cus_df.drop_duplicates()
print("Remaining duplicates:", cus_df.duplicated().sum())
```

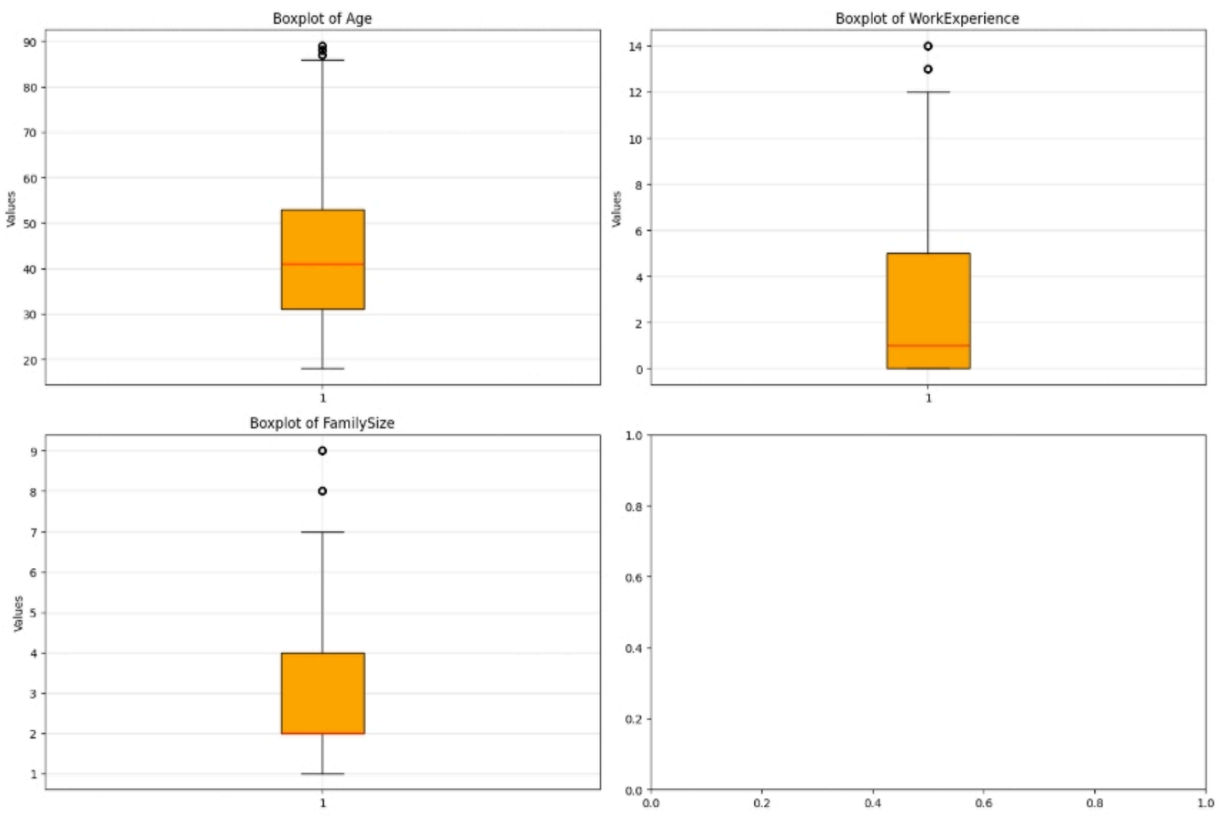
```
Remaining duplicates: 0
```

- Fill the null values in numerical columns with median values and fill the mode values in categorical columns

```
if 'WorkExperience' in cus_df.columns and cus_df['WorkExperience'].isnull().any():
  median_exp = cus_df['WorkExperience'].median()
  cus_df['WorkExperience'].fillna(median_exp, inplace=True)

if 'FamilySize' in cus_df.columns and cus_df['FamilySize'].isnull().any():
  mode_fam = cus_df['FamilySize'].mode()[0]
  cus_df['FamilySize'].fillna(mode_fam, inplace=True)

for col in category_cus_df:
  if col in cus_df.columns and cus_df[col].isnull().any():
    mode_col = cus_df[col].mode()[0]
    cus_df[col].fillna(mode_col, inplace=True)
```

DATA PREPARATION

- Check the outlier values in numerical columns

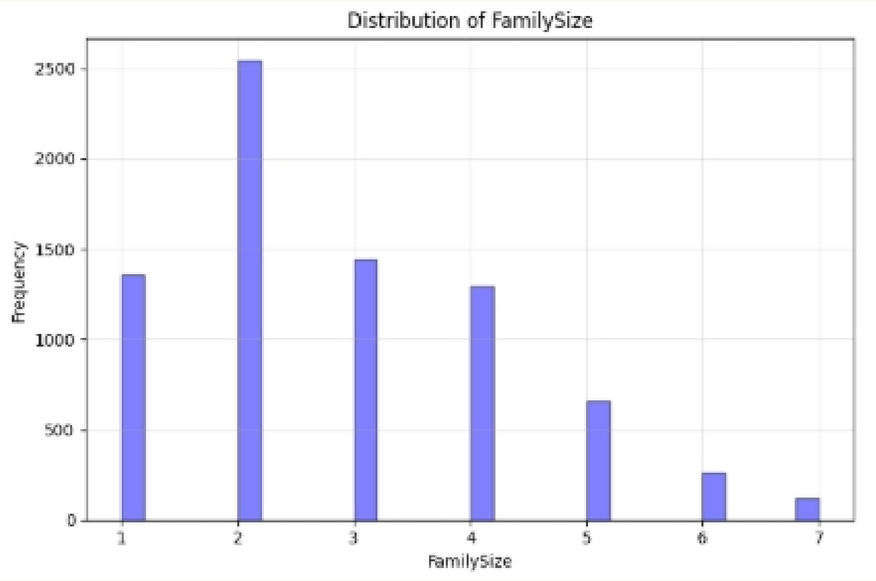


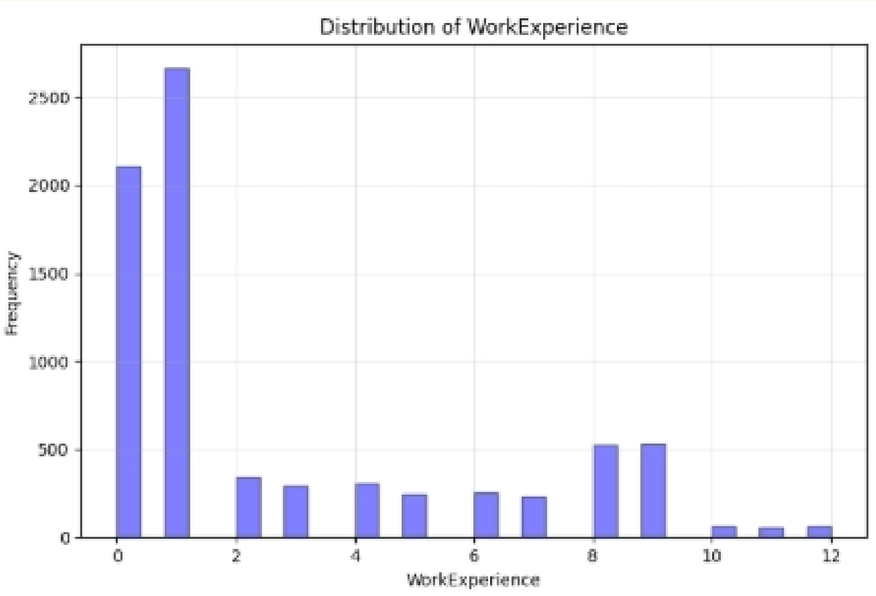
- Drop the outlier values

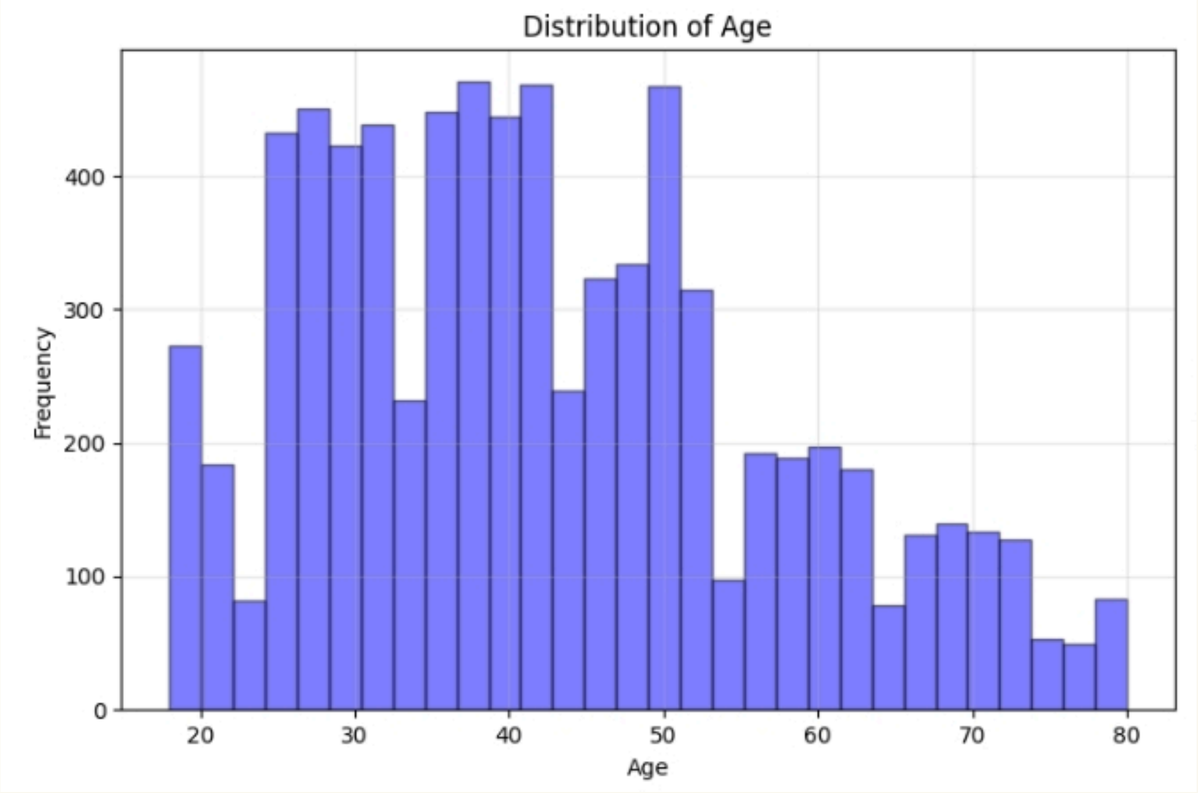
```
for col in numberical_cus_df:
    if col in cus_df.columns:
        Q1 = cus_df[col].quantile(0.25)
        Q3 = cus_df[col].quantile(0.75)
        IQR = Q3 - Q1
        lower_bound = Q1 - 1.5 * IQR
        upper_bound = Q3 + 1.5 * IQR

        cus_df = cus_df[(cus_df[col] >= lower_bound) & (cus_df[col] <= upper_bound)]
```

```
[383] cus_df = cus_df.drop(cus_df[cus_df['Age'] > 80].index)
```

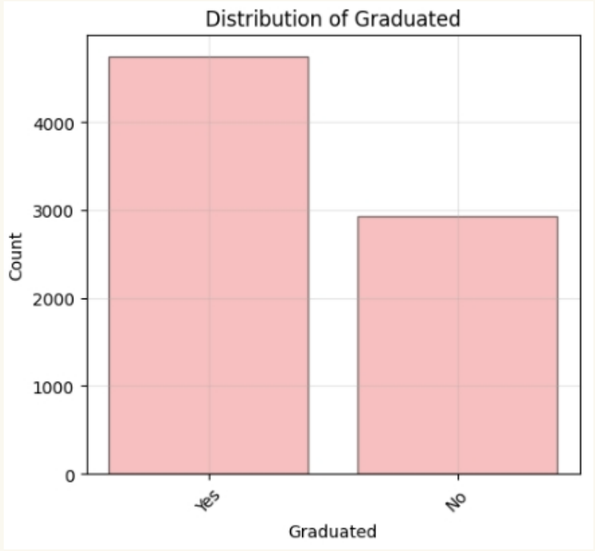
EXPLORATORY DATA ANALYSIS (EDA)

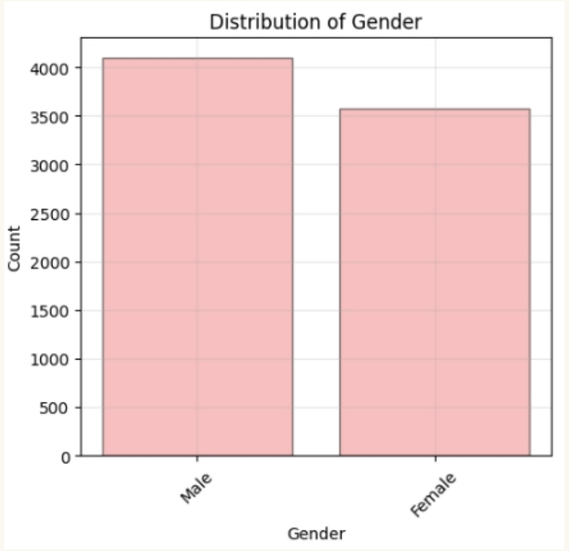


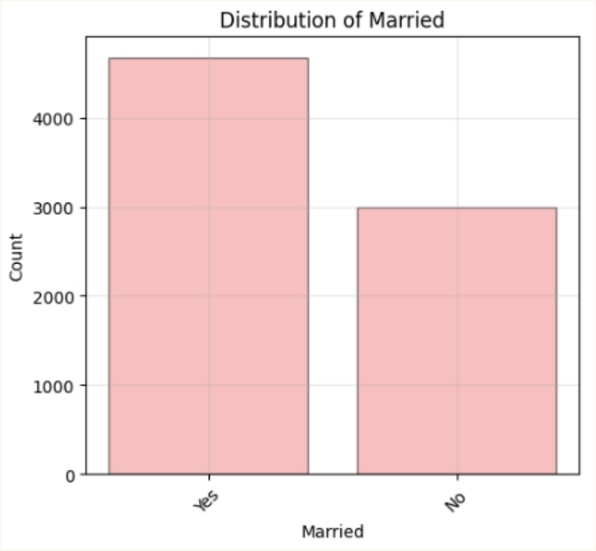


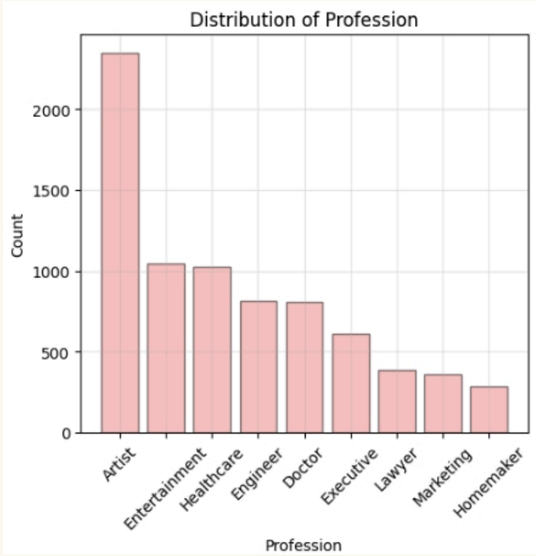


EXPLORATORY DATA ANALYSIS (EDA)

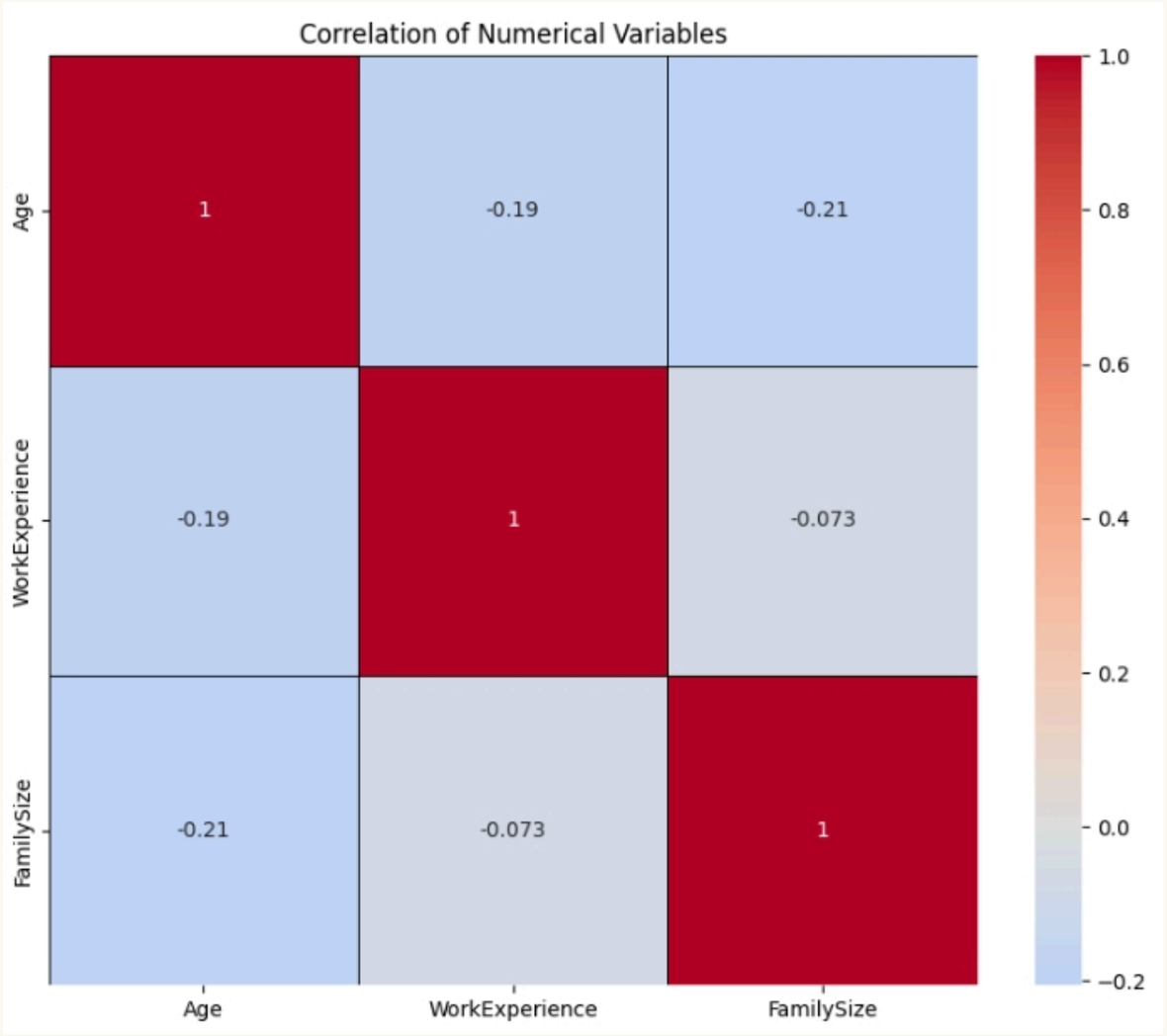








EXPLORATORY DATA ANALYSIS (EDA)



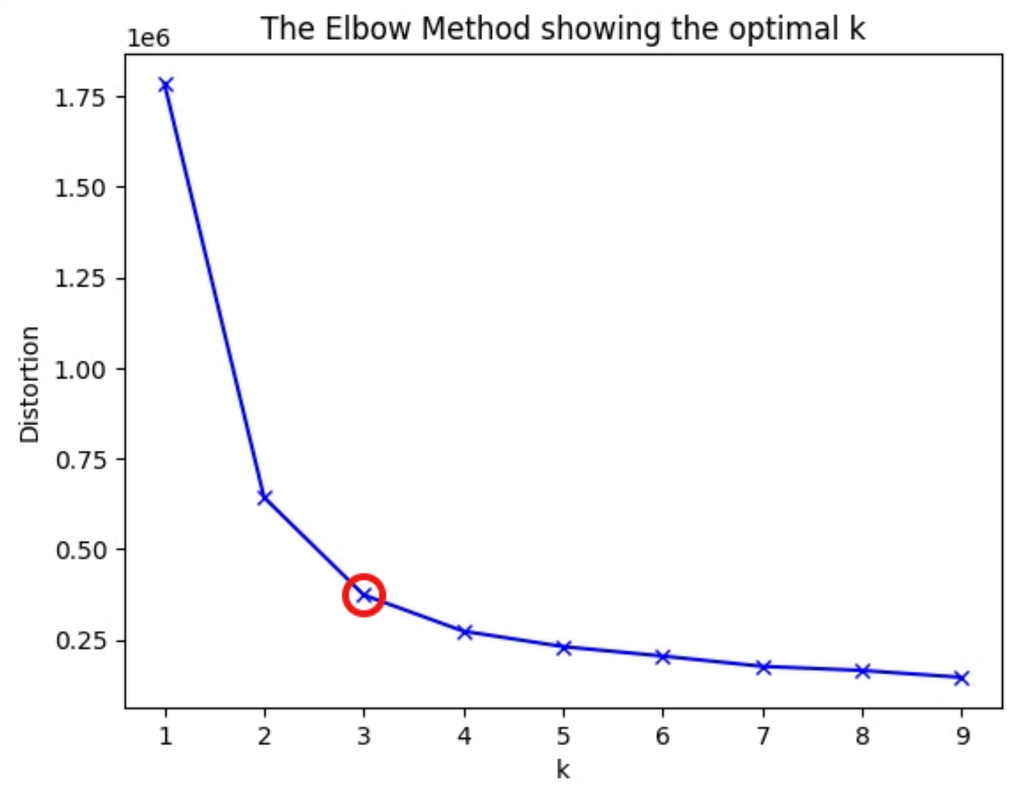
DATA MODELING

- Turn the categorical values into numerical values

	Gender	Married	Age	Graduated	Job	WorkExperience	SpendingScore	FamilySize
0	1	0	22	0	5	1.00	2	4.00
1	0	1	38	1	2	1.00	0	3.00
2	0	1	67	1	2	1.00	2	1.00
3	1	1	67	1	7	0.00	1	2.00
4	0	1	40	1	3	1.00	1	6.00
...	...	...	...	...	...	...	...	...
7671	0	0	27	0	8	8.00	2	4.00
7672	0	1	42	1	0	0.00	2	2.00
7673	0	0	35	1	3	1.00	2	2.00
7674	0	0	53	1	3	1.00	2	2.00
7675	0	0	43	1	5	9.00	2	3.00

DATA MODELING

- Use the Elbow method to predict the number of cluster



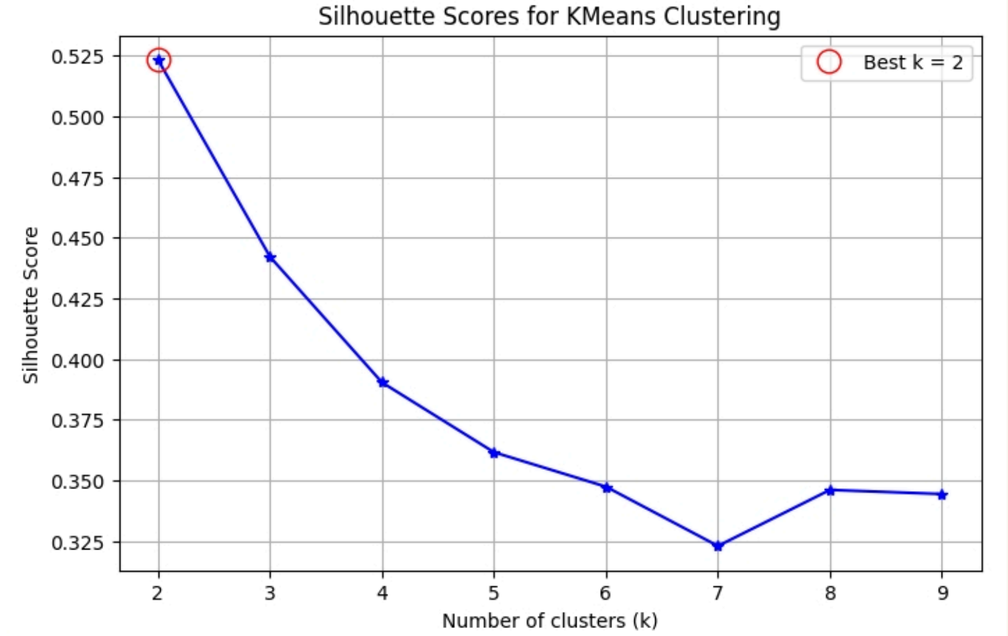
The best number of cluster is

k = 3

DATA MODELING

- Use the Silhouette score to find the best score of cluster

```
Silhouette Coefficient for k == 2: 0.5233
Silhouette Coefficient for k == 3: 0.4421
Silhouette Coefficient for k == 4: 0.3904
Silhouette Coefficient for k == 5: 0.3616
Silhouette Coefficient for k == 6: 0.3474
Silhouette Coefficient for k == 7: 0.323
Silhouette Coefficient for k == 8: 0.3461
Silhouette Coefficient for k == 9: 0.3444
```



The best number of cluster is k = 2 , but for the interpretation of cluster , we will use k = 3 according to the Elbow method for more cluster to segment

EVALUATION

We will use 3 models for clustering

- KMeans clustering
- Agglomerative Clustering
- DBScan Clustering

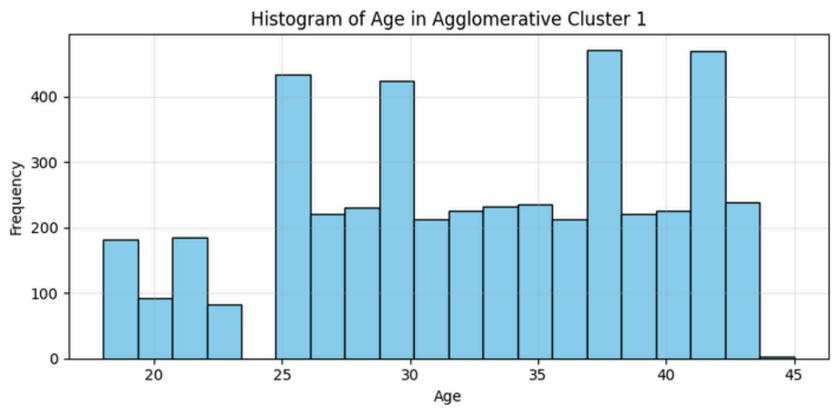
- The best Silhouette score for clustering model is **Agglomerative**

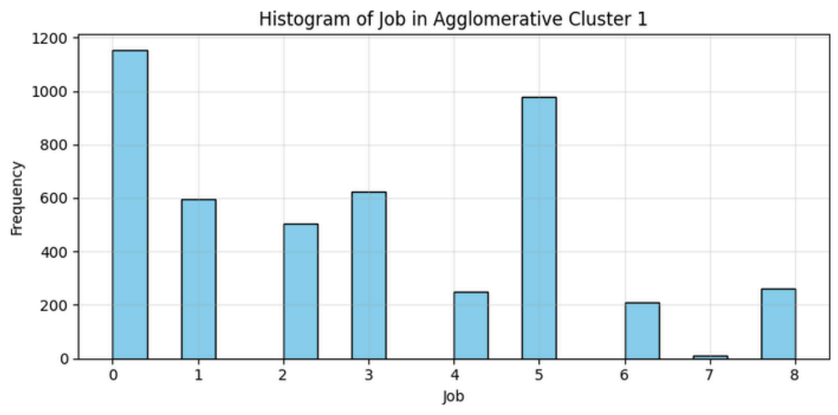
```
KMeans Silhouette Score: 0.293
Agglomerative Silhouette Score: 0.303
DBSCAN Silhouette Score: -0.114
```

EVALUATION

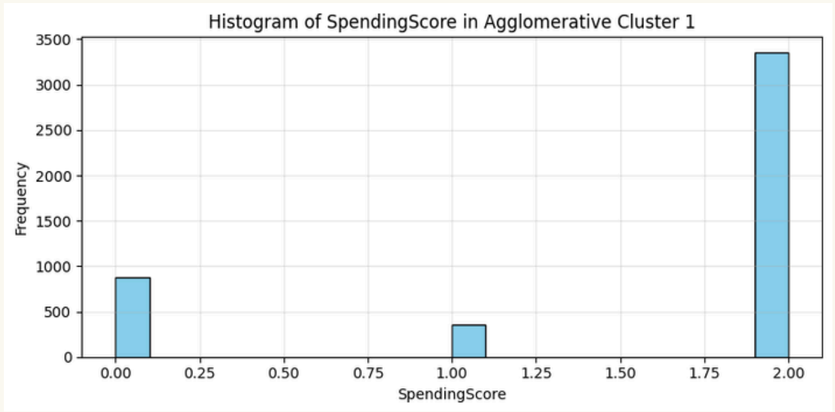
Let’s see the insights of **Agglomerative** cluster

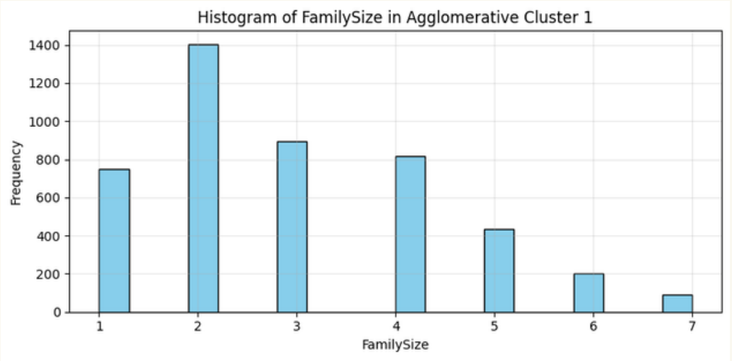
The First clustering insights is





	Gender	Married	Graduated
count	4,590.00	4,590.00	4,590.00
mean	0.51	0.43	0.56
std	0.50	0.49	0.50
min	0.00	0.00	0.00
25%	0.00	0.00	0.00
50%	1.00	0.00	1.00
75%	1.00	1.00	1.00
max	1.00	1.00	1.00





EVALUATION

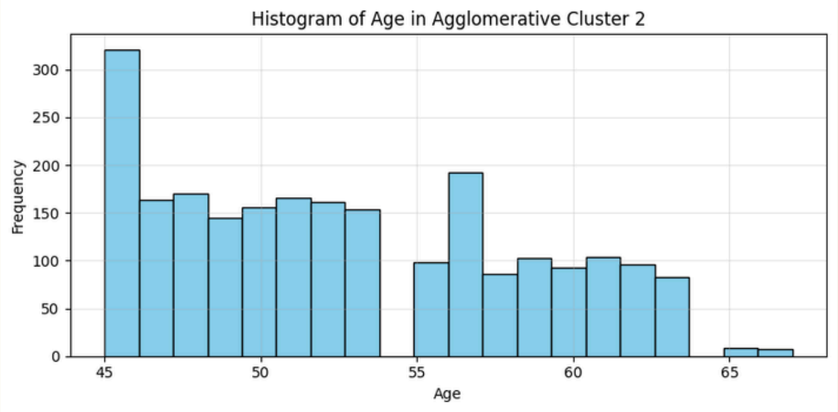
Let’s see the insights of **agg_cluster1**

Gender	Male: 51% Female: 49%
Married	Yes: 57% No: 43%
Age	Mostly: 37 yrs **Mean: 32 yrs** Min: 23 yrs
Graduated	Yes: 56% No: 44%
Job	Mostly: Artist **Mean**: Healthcare, Doctor, Engineer, Entertainment Min: Lawyer
WorkExperience	Mostly: 0–1 yr **Mean**: 2–9 yrs Min: 10–12 yrs
SpendingScore	Mostly: Low **Mean**: Average Min: High
FamilySize	Mostly: 2 **Mean**: 1, 3, 4 Min: 5–6

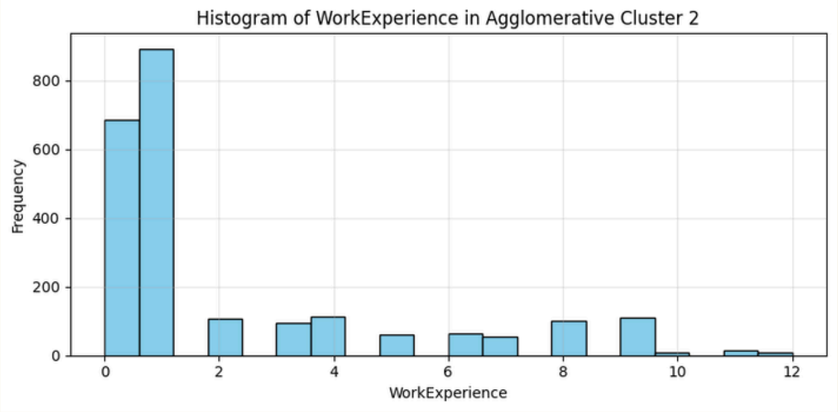
EVALUATION

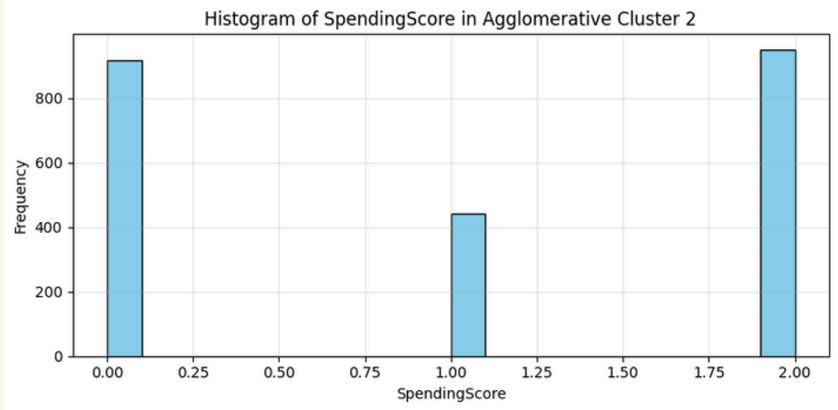
Let’s see the insights of **Agglomerative** cluster

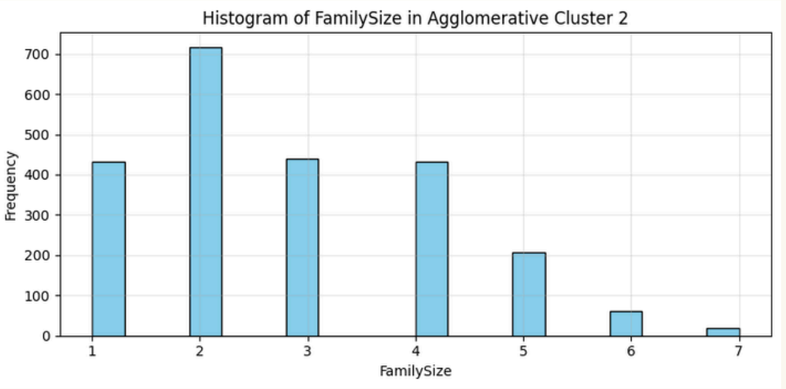
The Second clustering insights is



	Gender	Married	Graduated
count	2,308.00	2,308.00	2,308.00
mean	0.56	0.86	0.74
std	0.50	0.35	0.44
min	0.00	0.00	0.00
25%	0.00	1.00	0.00
50%	1.00	1.00	1.00
75%	1.00	1.00	1.00
max	1.00	1.00	1.00







EVALUATION

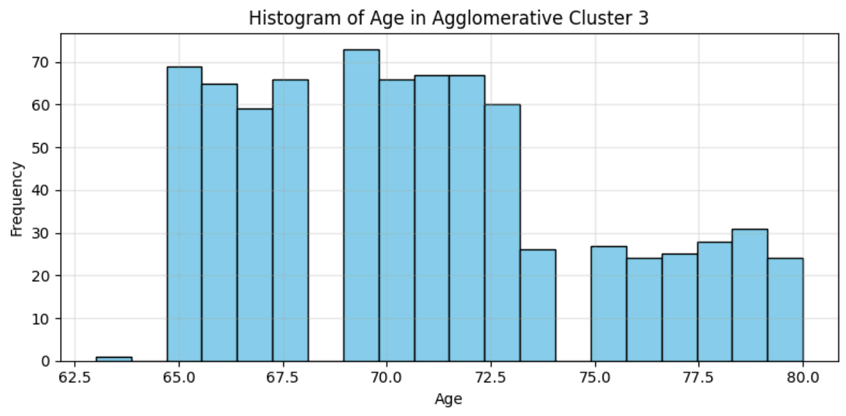
Let’s see the insights of **agg_cluster2**

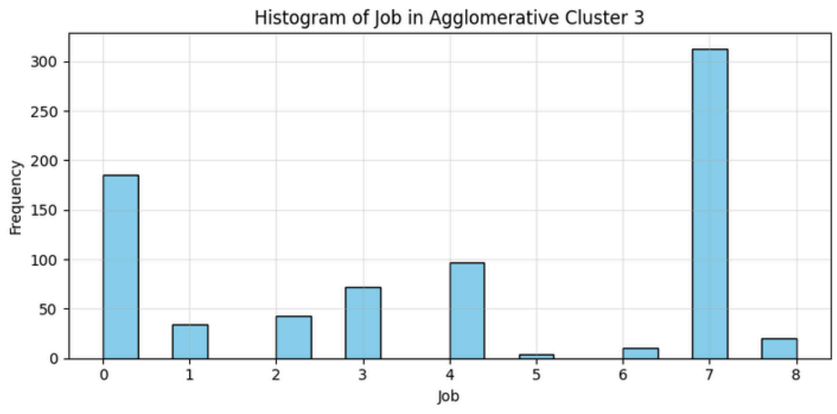
Gender	Male: 56% Female: 44%
Married	Yes: 86% No: 14%
Age	Mostly: 46 yrs **Mean: 47–54 yrs** Min: 65–66 yrs
Graduated	Yes: 74% No: 26%
Job	Mostly: Artist **Mean**: Engineer, Entertainment, Executive, Doctor Min: Healthcare
WorkExperience	Mostly: 0–1 yr **Mean**: 2–9 yrs Min: 10–12 yrs
SpendingScore	Mostly: Low, Average Min: High
FamilySize	Mostly: 2 **Mean**: 1, 3, 4 Min: 5–6

EVALUATION

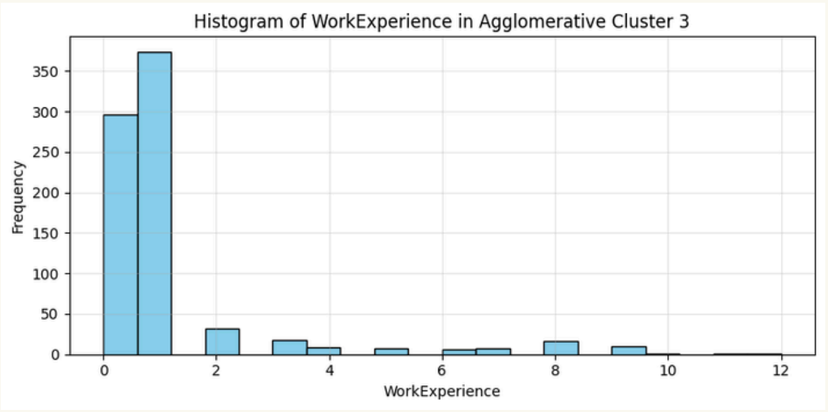
Let’s see the insights of **Agglomerative** cluster

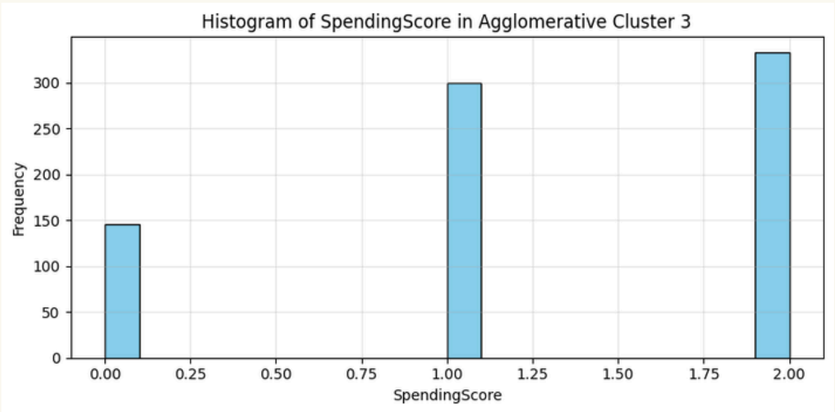
The Third clustering insights is

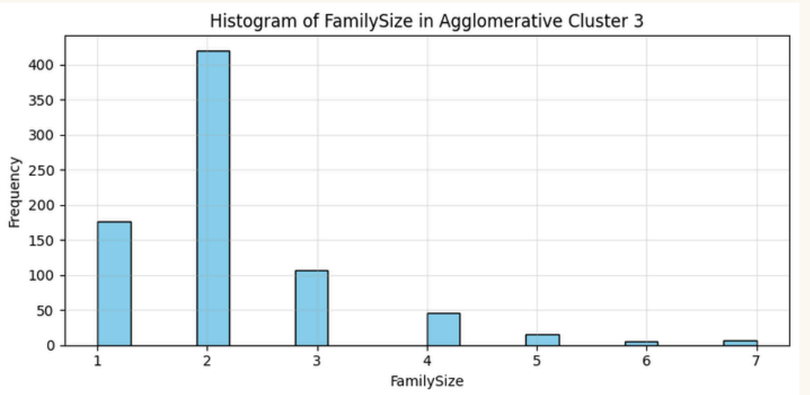




	Gender	Married	Graduated
count	778.00	778.00	778.00
mean	0.60	0.94	0.63
std	0.49	0.24	0.48
min	0.00	0.00	0.00
25%	0.00	1.00	0.00
50%	1.00	1.00	1.00
75%	1.00	1.00	1.00
max	1.00	1.00	1.00







EVALUATION

Let’s see the insights of **agg_cluster3**

Gender	Male: 60% Female: 40%
Married	Yes: 94% No: 6%
Age	Mostly: **65–72 yrs** Mean: 73–80 yrs
Graduated	Yes: 63% No: 37%
Job	Mostly: Lawyer **Mean**: Artist, Entertainment, Executive Min: Healthcare
Work Experience	Mostly **0–1 yrs**
Spending Score	Mostly: Low Mean: High Min: Average
Family Size	Mostly: 2 Mean: **1, 3** Min: 4

End