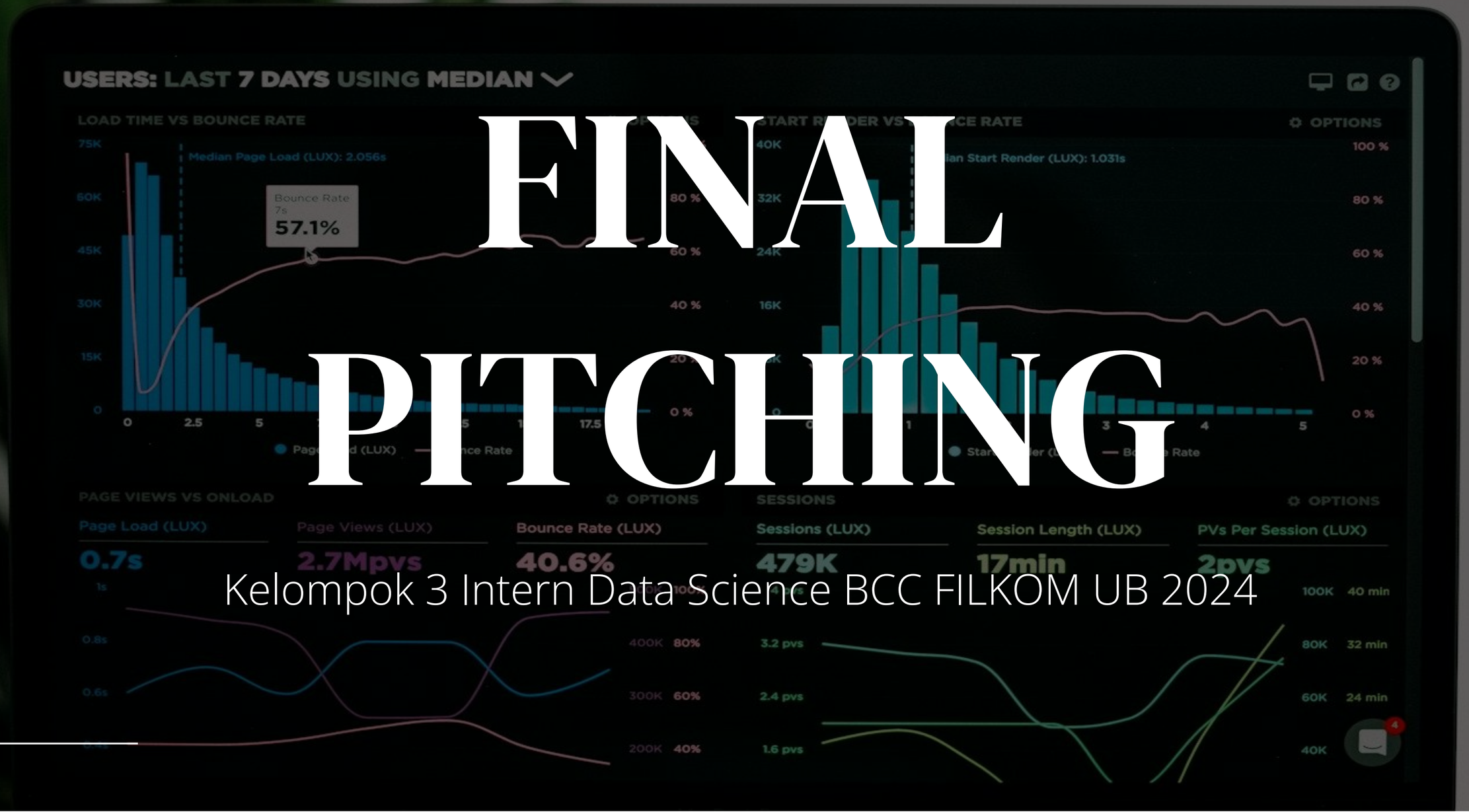
FINAL PITCHING

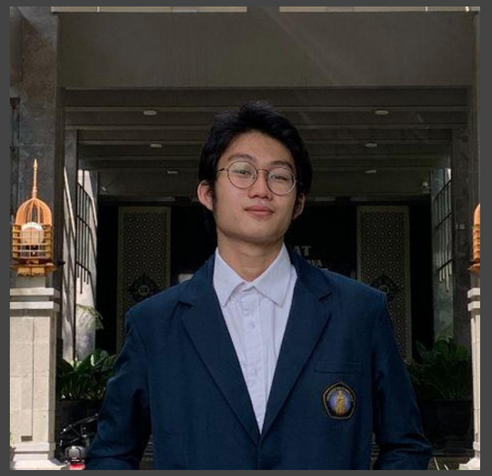
Kelompok 3 Intern Data Science BCC FILKOM UB 2024



OUR TEAM

Johanes Paulus Bernard Purek

Sistem Informasi ‘22



Pieter Christy Yan Yudhistira

Teknik Informatika ‘23

Theme & Topic

Consumer Segmentation Based on Behavior and Shopping Habits

Dataset

https://www.kaggle.com/datasets/zeesolver/consumer-behavior-and-shopping-habits-dataset?select=shopping_behavior_updated.csv

3900 Rows
18 Columns
1 Identifier
4 Numerical Column
13 Categorical Column

shopping_behavior_updated.csv (416.61 kB)

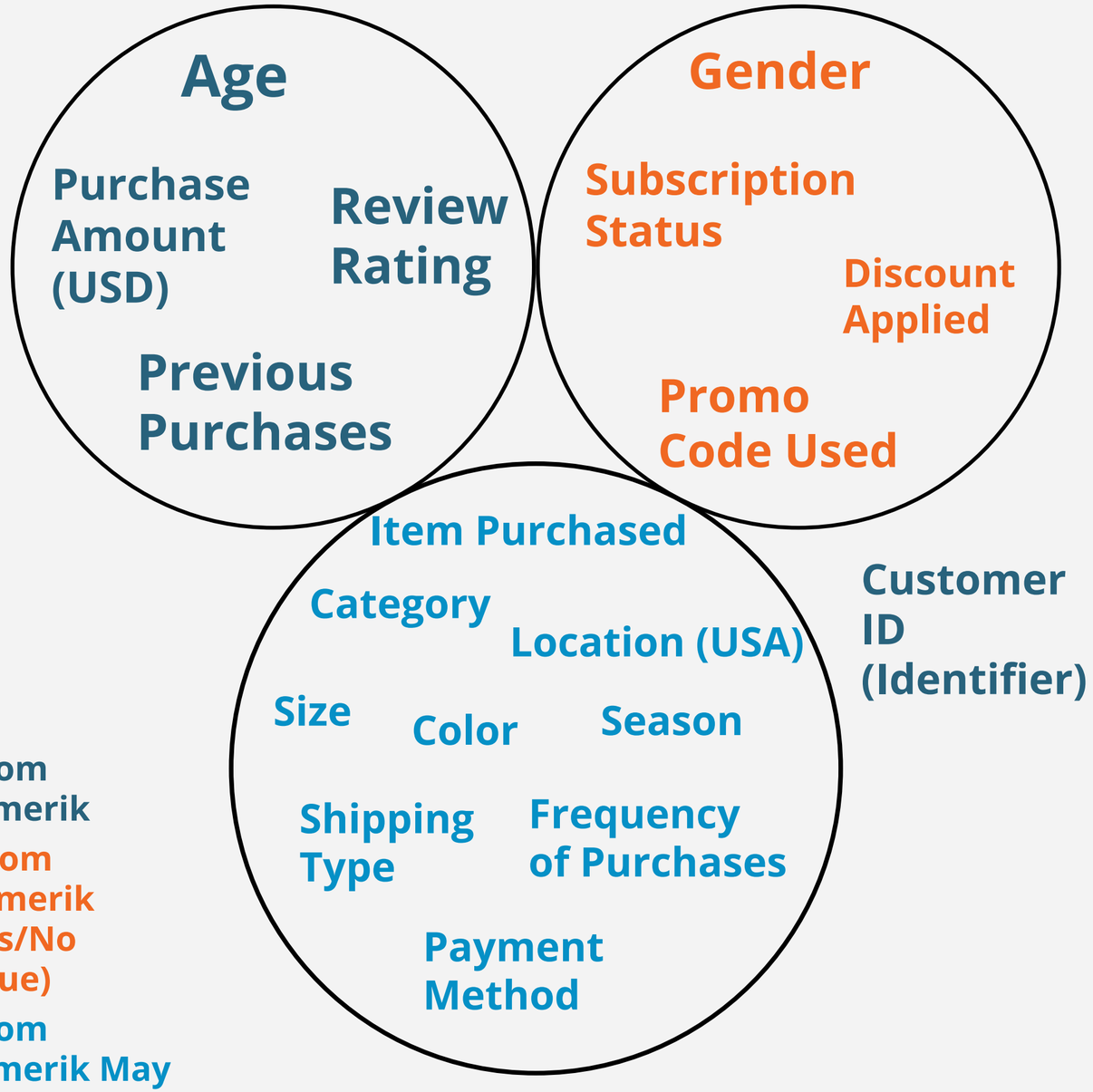
Detail | Compact | Column

18 of 18 columns

Customer ID	Age	Gender	Item Purch...	Category	Purchase ...	Location	Size	Color	Season	Revi
1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1
2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1
3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1
4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5
5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7
6	46	Male	Sneakers	Footwear	20	Wyoming	M	White	Summer	2.9
7	63	Male	Shirt	Clothing	85	Montana	M	Gray	Fall	3.2
8	27	Male	Shorts	Clothing	34	Louisiana	L	Charcoal	Winter	3.2
9	26	Male	Coat	Outerwear	97	West Virginia	L	Silver	Summer	2.6
10	57	Male	Handbag	Accessories	31	Missouri	M	Pink	Spring	4.8
11	53	Male	Shoes	Footwear	34	Arkansas	L	Purple	Fall	4.1
12	30	Male	Shorts	Clothing	68	Hawaii	S	Olive	Winter	4.9
13	61	Male	Coat	Outerwear	72	Delaware	M	Gold	Winter	4.5
14	65	Male	Dress	Clothing	51	New Hampshire	M	Violet	Spring	4.7
15	64	Male	Coat	Outerwear	53	New York	L	Teal	Winter	4.7
16	64	Male	Skirt	Clothing	81	Rhode Island	M	Teal	Winter	2.8
17	25	Male	Sunglasses	Accessories	36	Alabama	S	Gray	Spring	4.1
18	53	Male	Dress	Clothing	38	Mississippi	XL	Lavender	Winter	4.7

Acknowledgements
Sir Sourav Banerjee
Associate Data Scientist at CogniTensor Kolkata, West Bengal, India.

DATASET COLUMNS



- Kolom Numerik
- Kolom Numerik (Yes/No Value)
- Kolom Numerik May Unique Value

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3900 entries, 0 to 3899
Data columns (total 18 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Customer ID             3900 non-null   int64
 1   Age                     3900 non-null   int64
 2   Gender                  3900 non-null   object
 3   Item Purchased          3900 non-null   object
 4   Category                3900 non-null   object
 5   Purchase Amount (USD)   3900 non-null   int64
 6   Location                3900 non-null   object
 7   Size                    3900 non-null   object
 8   Color                   3900 non-null   object
 9   Season                  3900 non-null   object
 10  Review Rating           3900 non-null   float64
 11  Subscription Status     3900 non-null   object
 12  Shipping Type           3900 non-null   object
 13  Discount Applied        3900 non-null   object
 14  Promo Code Used         3900 non-null   object
 15  Previous Purchases      3900 non-null   int64
 16  Payment Method          3900 non-null   object
 17  Frequency of Purchases  3900 non-null   object
dtypes: float64(1), int64(4), object(13)
memory usage: 548.6+ KB
```

S.M.A.R.T ANALYSIS

	QUESTION	ANSWER
Spesific	Apa tujuan yang ingin dicapai dari pengerjaan Project ini?	Tujuan yang ingin dicapai dari project ini adalah untuk memprediksi segmentasi customer dalam kebiasaan berbelanja dengan melakukan beberapa teknik untuk melatih model Machine Learning sehingga mampu memprediksi secara akurat.
Measurable	Bagaimana cara menentukan bahwa Project ini dapat dicapai sesuai dengan rencana yang telah ditetapkan?	Project ini dapat dicapai apabila telah melakukan pemodelan untuk setiap segmentasi customer sehingga akan didapatkan hasil metrik yang akan disesuaikan dengan teknik ML.
Action Oriented	Apakah tujuan pengerjaan Project dapat dicapai secara maksimal? Apa yang perlu dilakukan agar mencapai hal tersebut?	Dapat dicapai secara maksimal dengan tahapan awal untuk memahami isi dari dataset, mengeksplorasi setiap fitur, melakukan tahap preprocessing, dan membuat model klasifikasi yang baik.
Relevant	Apakah pengerjaan Project ini selaras dengan kehidupan atau karir yang akan dijalani ke depan? Seberapa relevan pengerjaan Project ini?	Pengerjaan project ini selaras dengan kemampuan hard skill dalam bidang Data Science serta dapat membantu dalam memahami aspek bisnis mengenai perilaku seorang customer dalam kebiasaan belanja-nya.
Time-Bound	Kapan Batas Waktu Maksimal untuk Mencapai Tujuan Project ini?	Project ini bertujuan untuk dicapai secara maksimal pada tanggal 23 Maret 2024.

BCC
Basic Computing Community

HOW CAN WE MAKE A BETTER SEGMENTATION TO SEE CONSUMER BEHAVIOR AND THEIR SHOPPING HABITS?

BCC
Basic Computing Community

HOW CAN WE DETERMINE WHAT FACTOR IS THE MOST IMPORTANCE FOR A BETTER SEGMENTATION?

ALUR LANGKAH KERJA

01
Data Preparation

02
EDA and Visualization

03
Data Prepocessing

04
Modeling

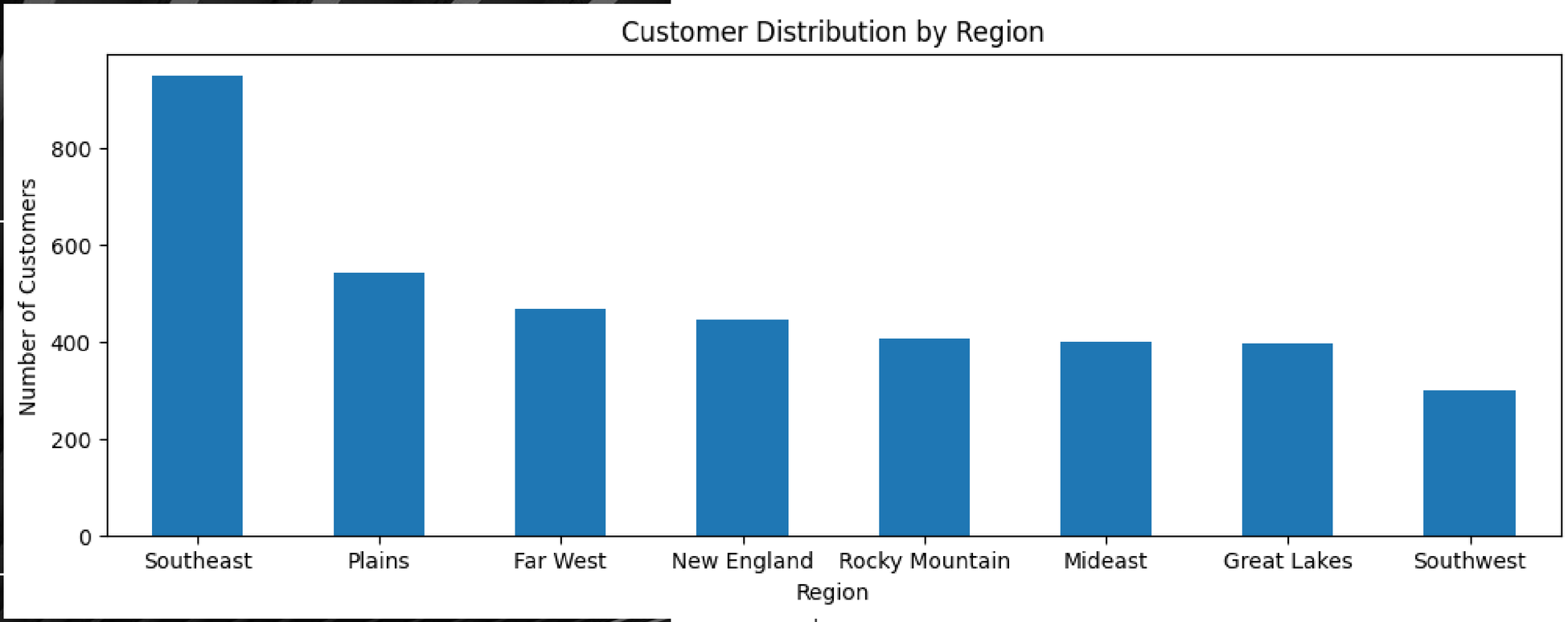
05
Performance Evaluation

06
Conclusion

BCC
Basic Computing Community

DATA PREPARATION

Classify Location into Region



Based on Bureau of Economic Analysis (BEA) regions

East Coast vs. West Coast: The impact of the Panama Canal's expansion on the routing of Asian imports into the United States

Preference for ethnic specialty produce by the Hispanics in the east coast of the USA

Sumber : "Preference for ethnic specialty produce by the Hispanics in the east coast of the USA"

https://apps.bea.gov/regional/docs/msalist.cfm?mlist=2

Sumber lain : "East Coast vs. West Coast: The impact of the Panama Canal's expansion on the routing of Asian imports into the United States"

Geographic Area

Sumber : http://www.bea.gov/newsreleases/regional/gdp_state/2012/_images/gsp_0612.png

Location Grouping | Age Grouping | Generation

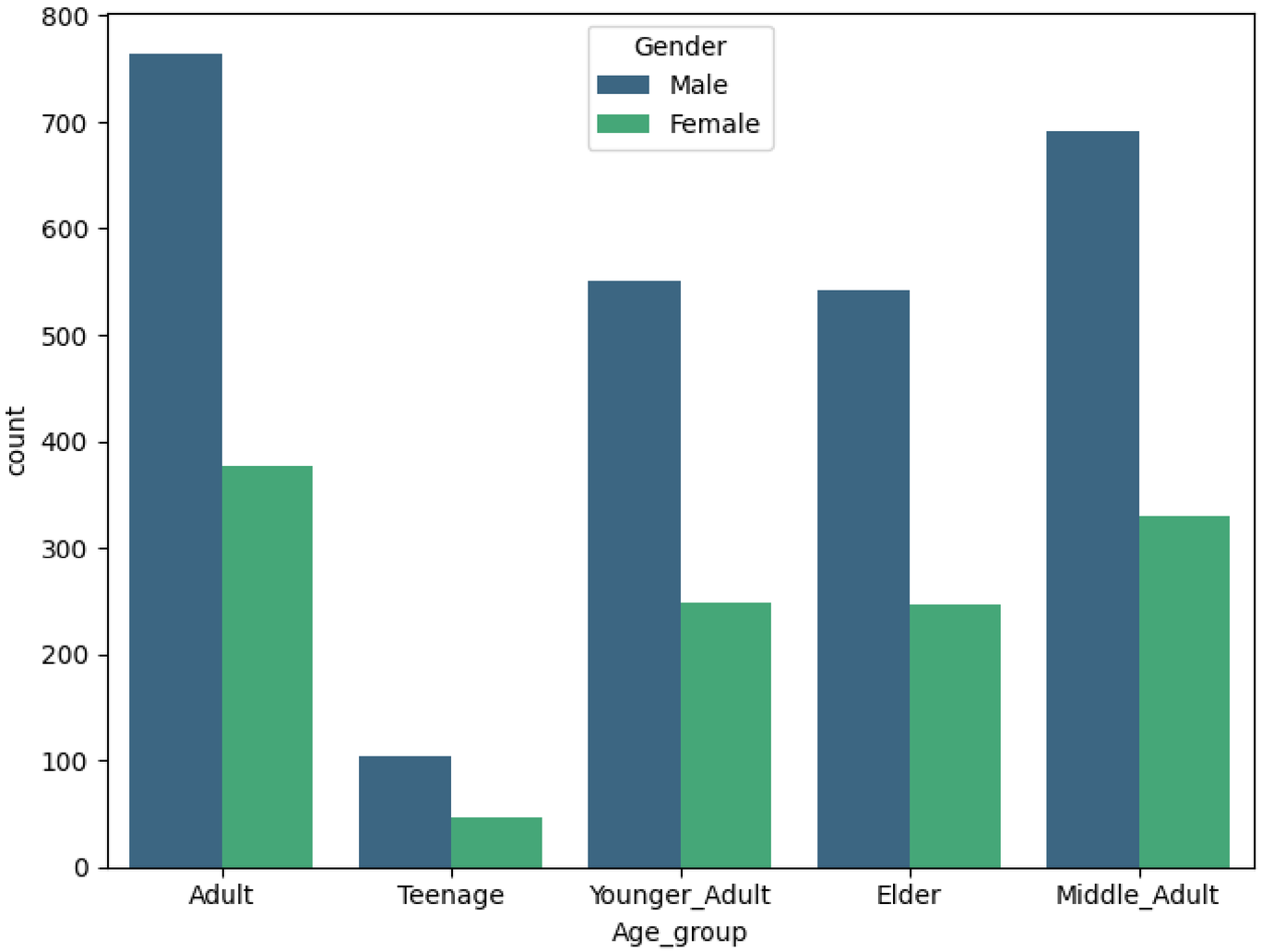
Grouping Age by Hierarchies

Age Group	Age Intervals
Baby	0~2
Young Adults	3~12
	13~19
	20~29
	30~39
Middle-aged Adults	40~49
	50~59
Old Adults	60~69
	70~79
	80~89
	90~99

"Usia pertengahan (middle age) atau disebut juga sebagai usia paruh baya yaitu kelompok usia dari 45-59 tahun."
Sumber : http://repo.unand.ac.id/397/3/bab%25201.pdf

"A teenager, or teen, is someone who is 13 to 19 years old."
Sumber : https://simple.wikipedia.org/wiki/Teenager

"In addition, while the older shoppers are more influenced by social support to adopt healthy shopping habits, the younger shoppers are more influenced by the relative price of products."
Sumber : https://dl.acm.org/doi/abs/10.1145/3209219.3209253



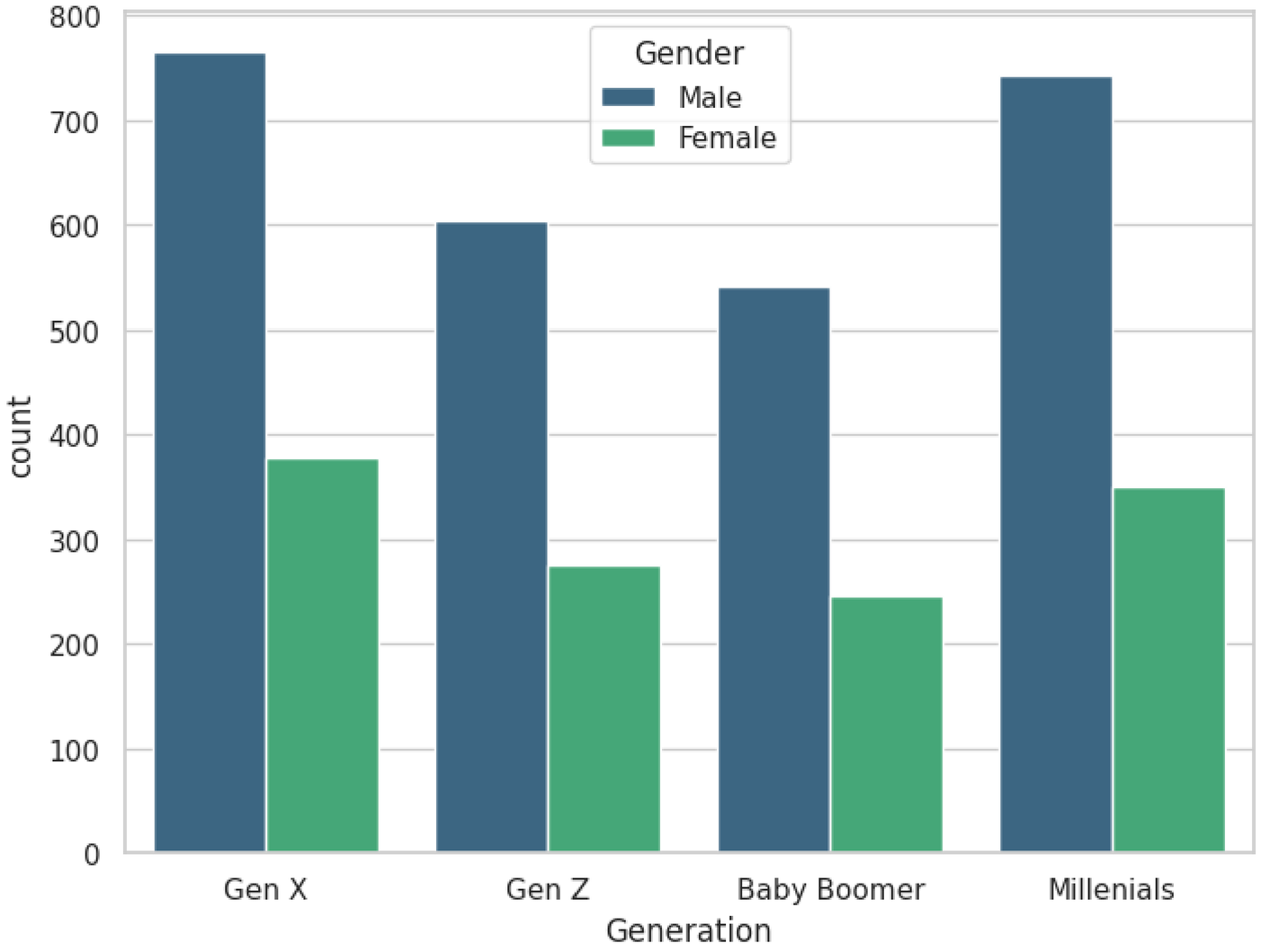
Grouping Age by Generation

- The Greatest Generation – born 1901-1924.
- The Silent Generation – born 1925-1945.
- The Baby Boomer Generation – born 1946-1964.
- Generation X – born 1965-1979.
- Millennials – born 1980-1994.
- Generation Z – born 1995-2012.
- Gen Alpha – born 2013 – 2025.

“The next generation is technologically adept; it's their job to program the VCR. And they're a more brand conscious group than any others. Whether they'll turn out to be Generation X squared or truly separate remains to be seen.”

“We found that Xers are for more competitive, far more materially driven than Boomers, and they're a little more realistic about what the future holds for them”.

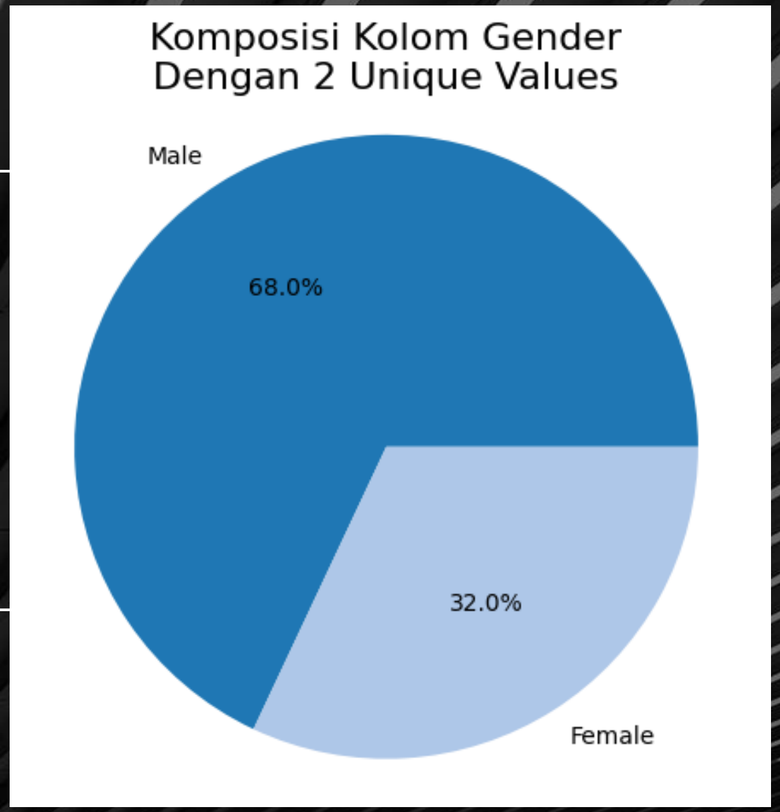
Higgins, K. T. (1998). Generational marketing. Marketing Management, 7(3), 6-9.

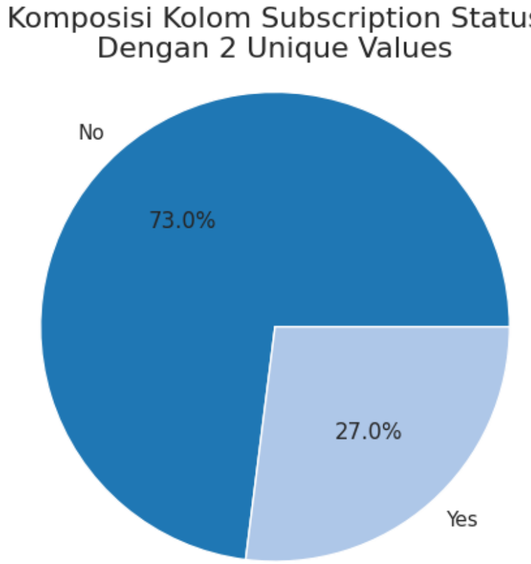


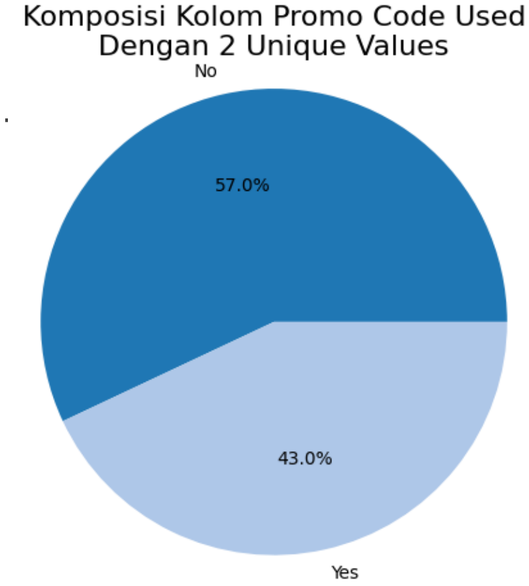
BCC
Basic Computing Community

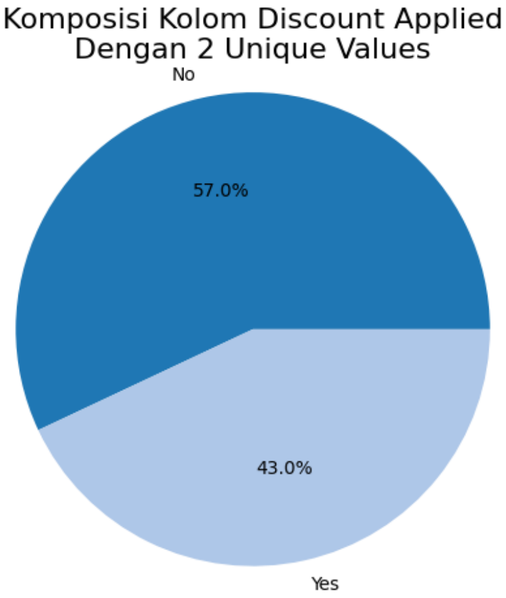
EDA
&
VISUALIZATION

Value Counts for Yes/No Column





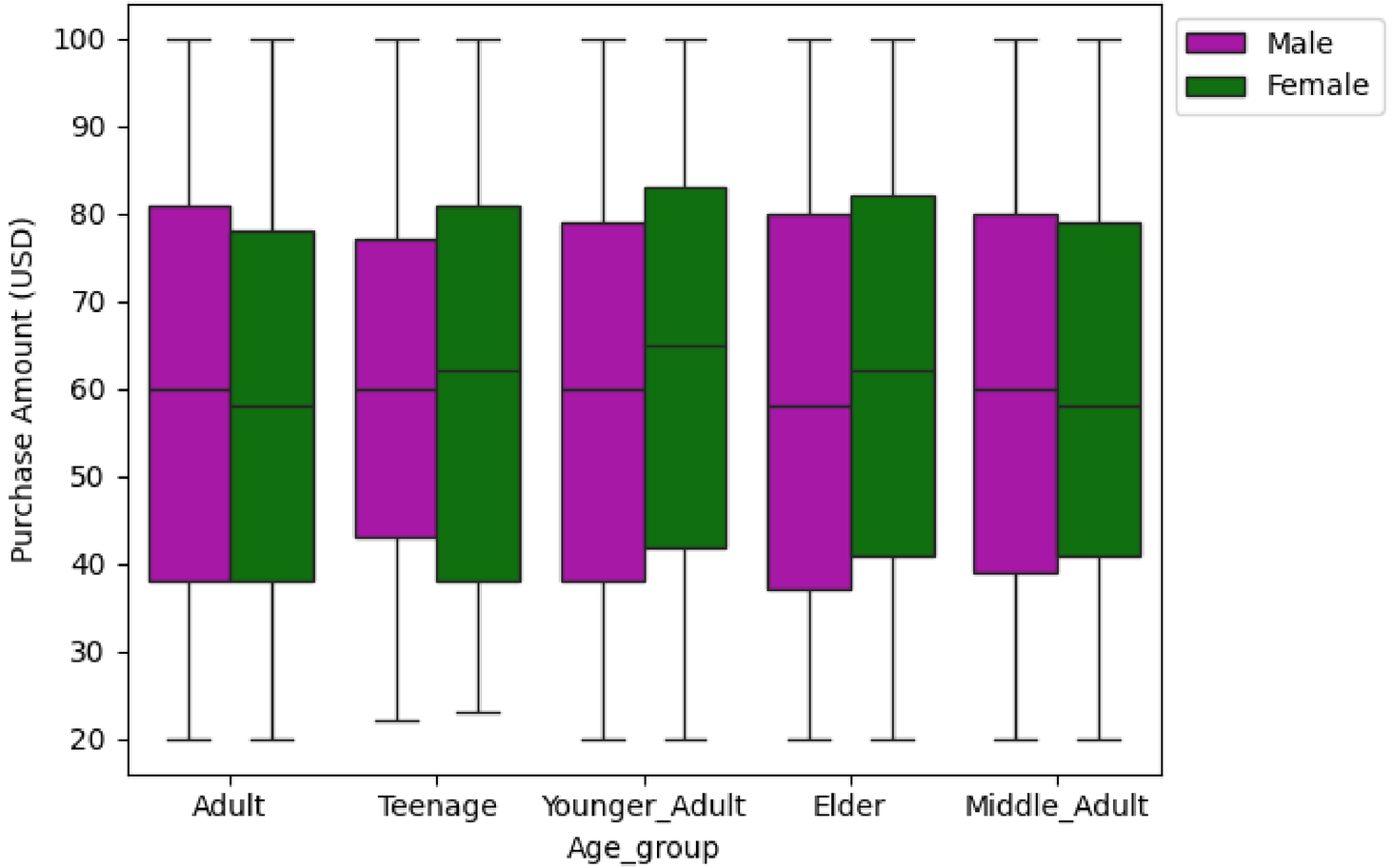




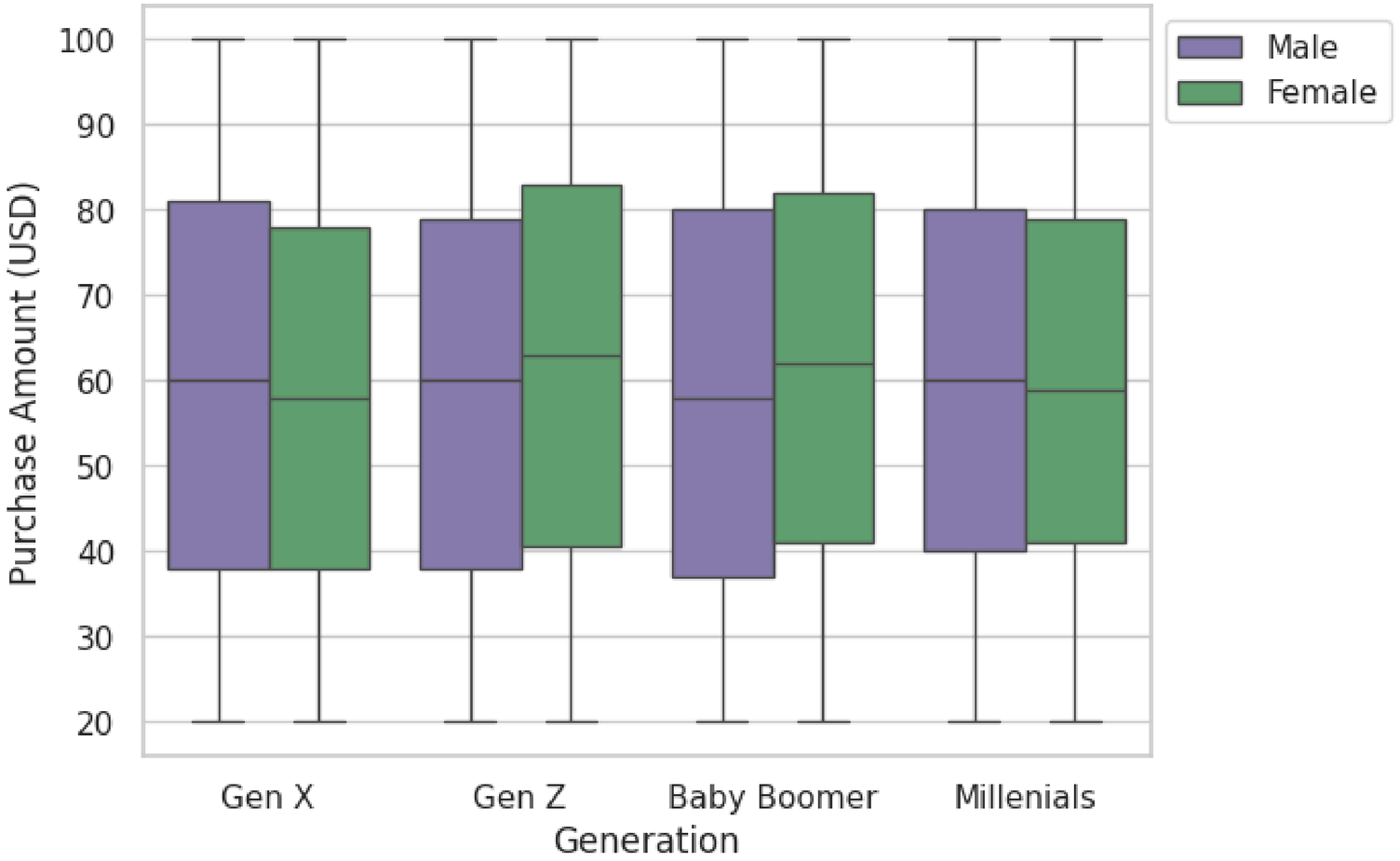
Kolom Promo Code Used dan Discount Applied Memiliki Persentase yang sama

Apa Hubungannya?

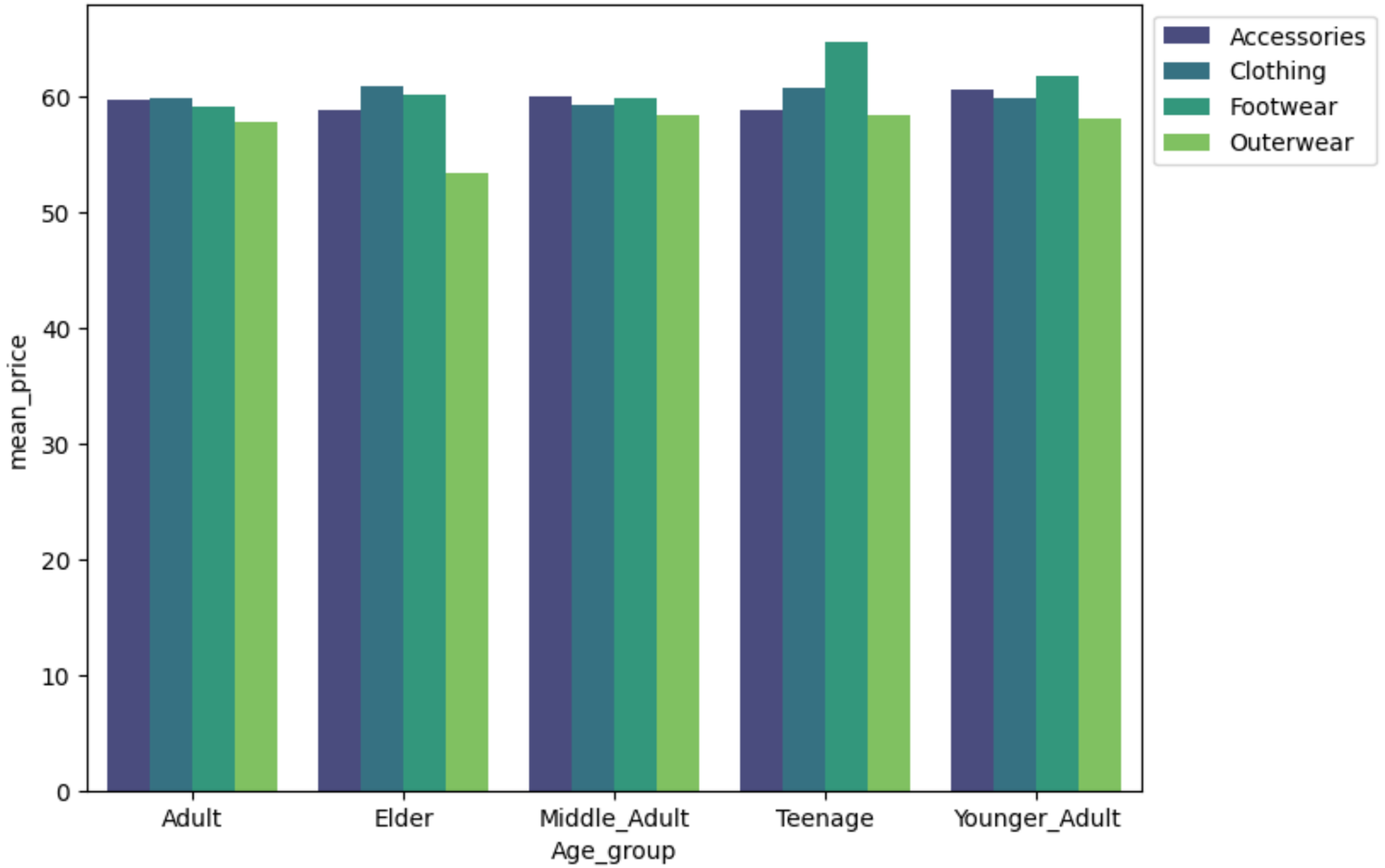
Analysis of Age Group and Purchase Amount



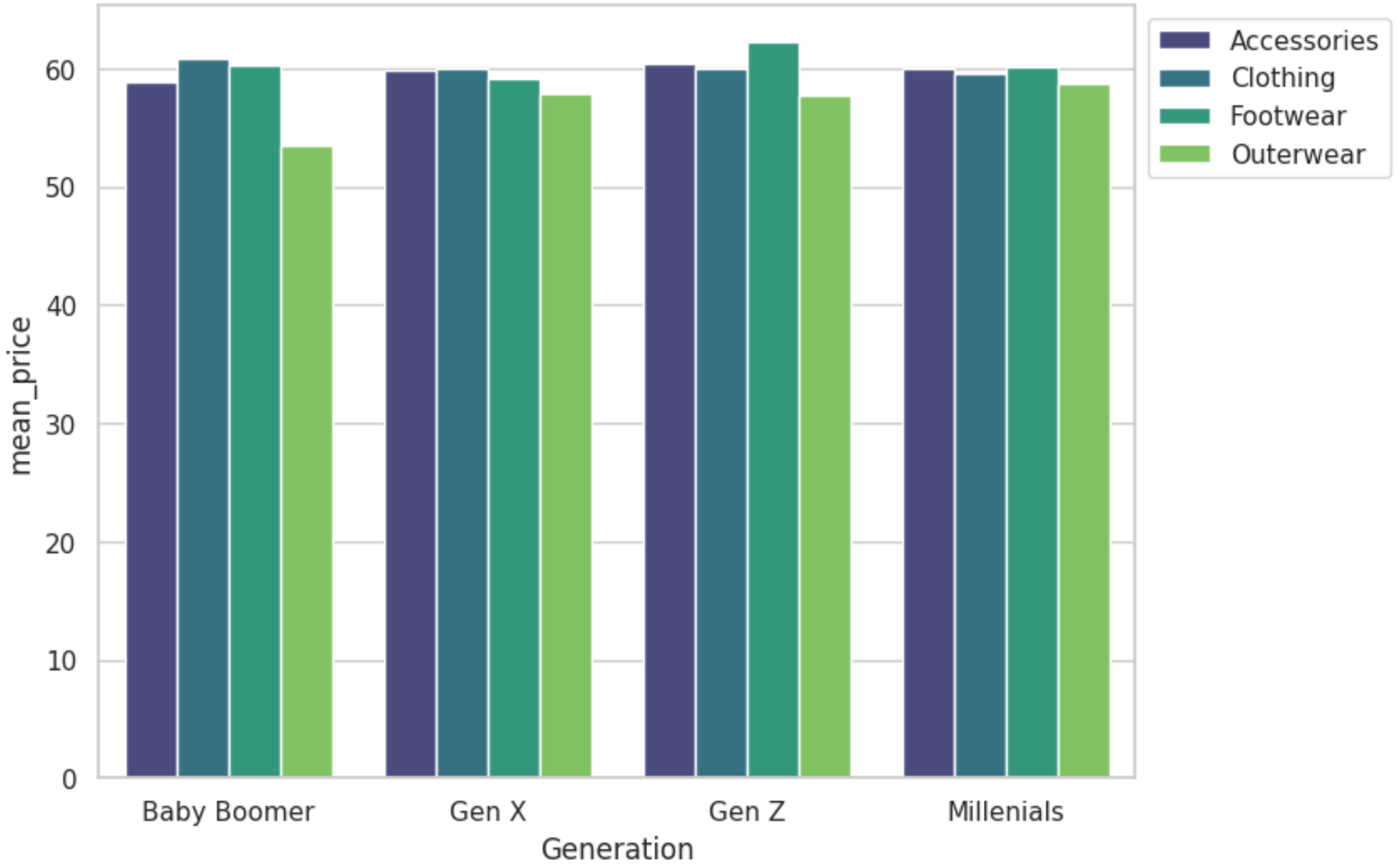
Analysis of Generation and Purchase Amount



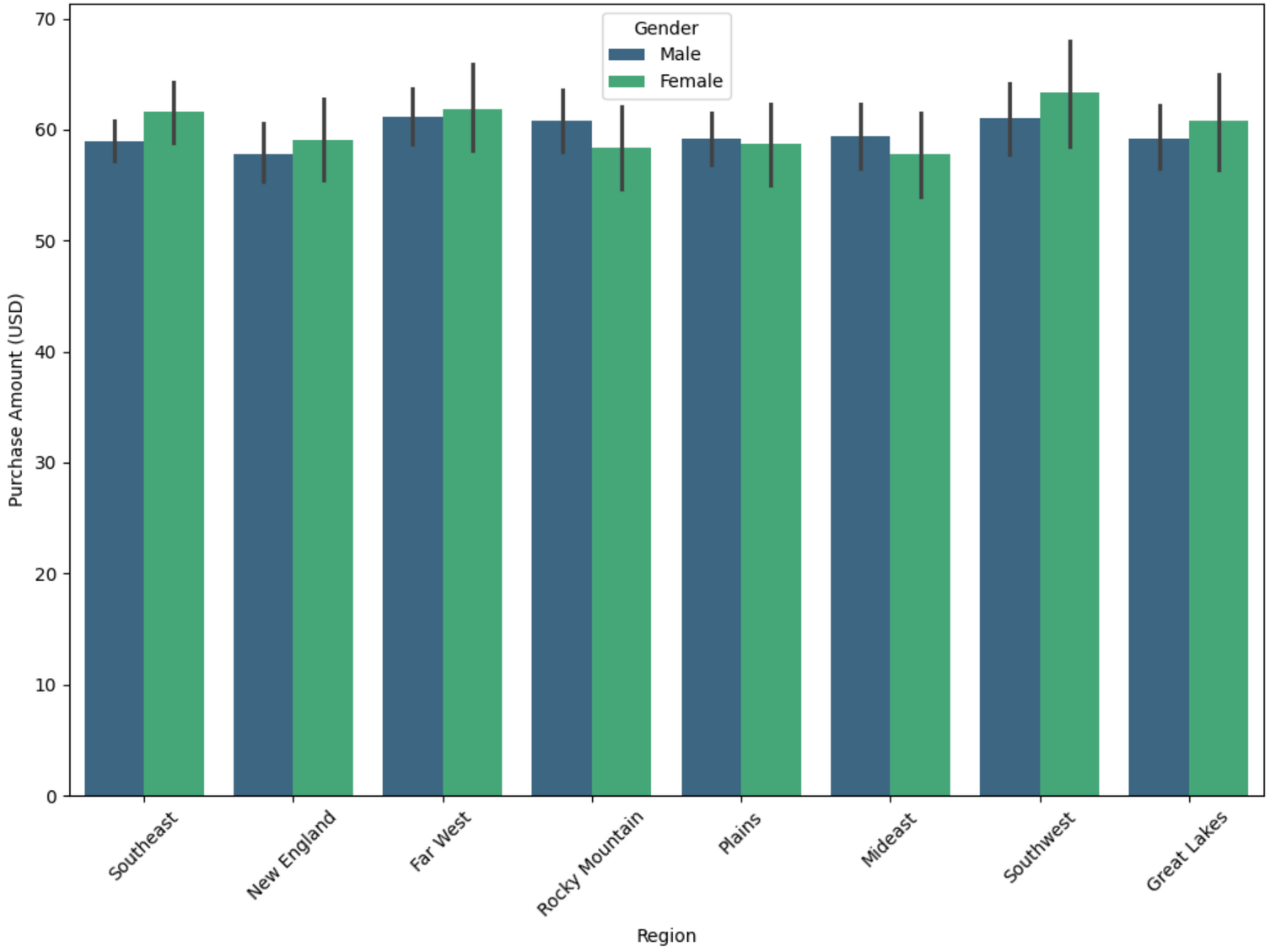
Analysis of Age Group, Category and Purchase Amount



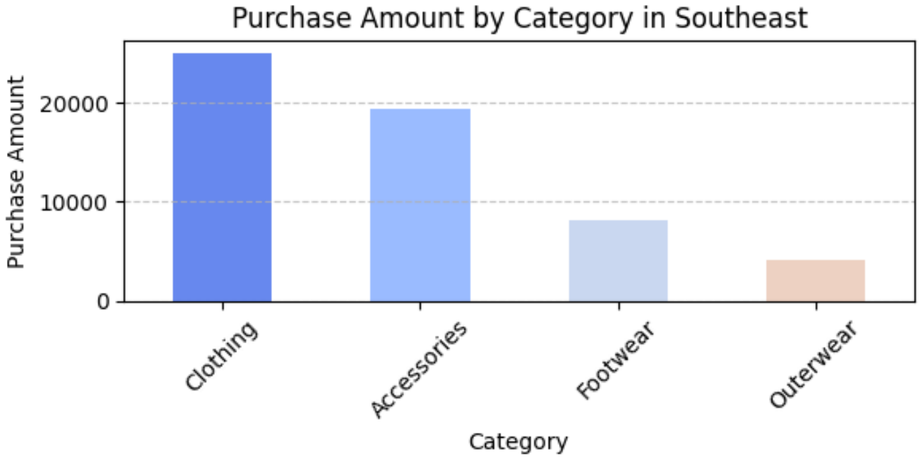
Analysis of Generation, Category and Purchase Amount

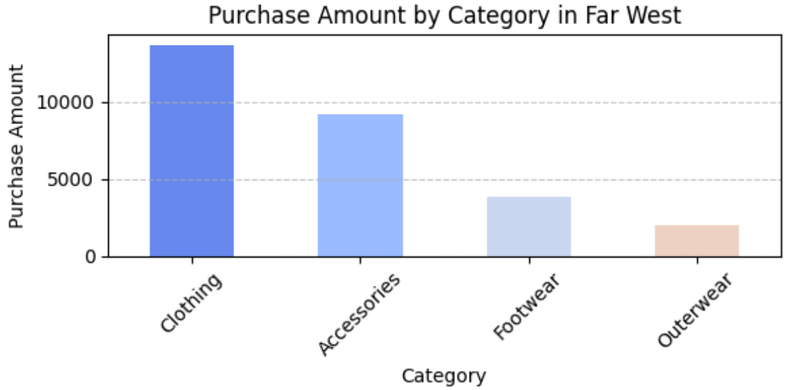


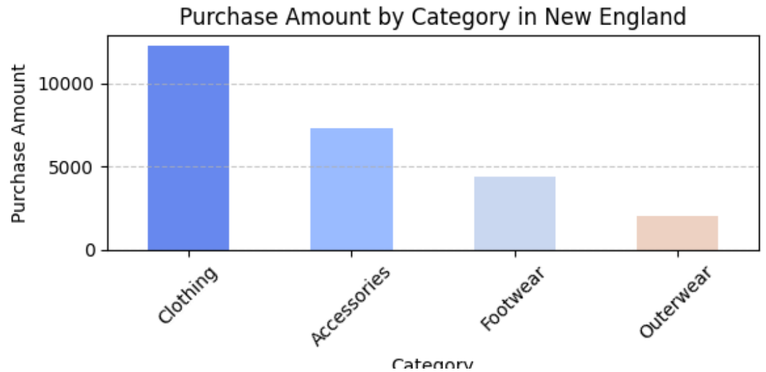
Top 10 Purchasing Region

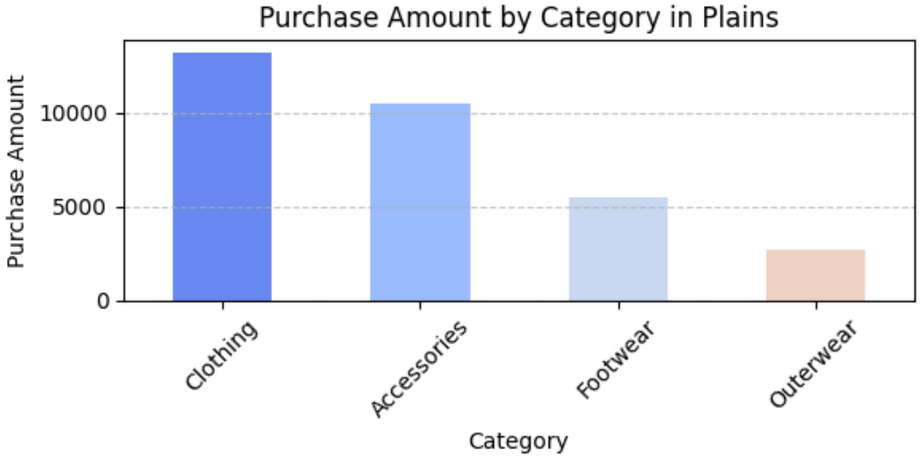


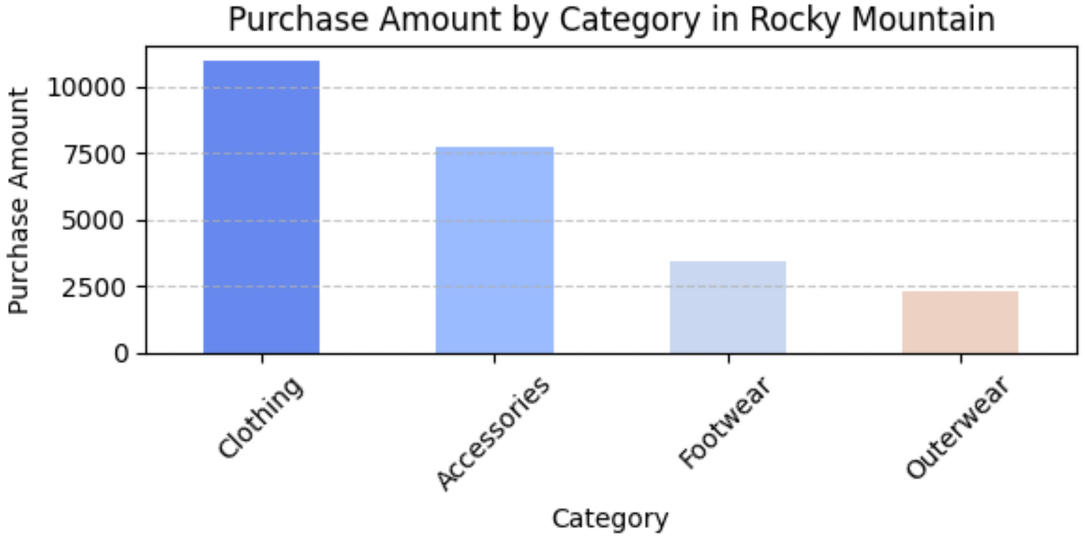
Top 5 Regions and Category



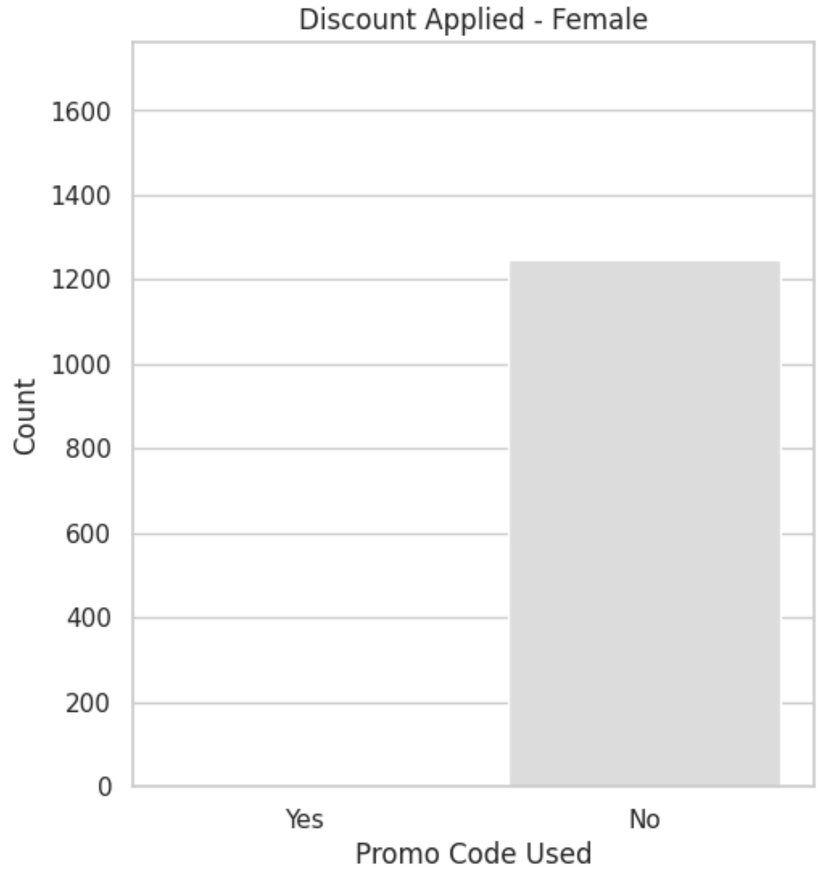


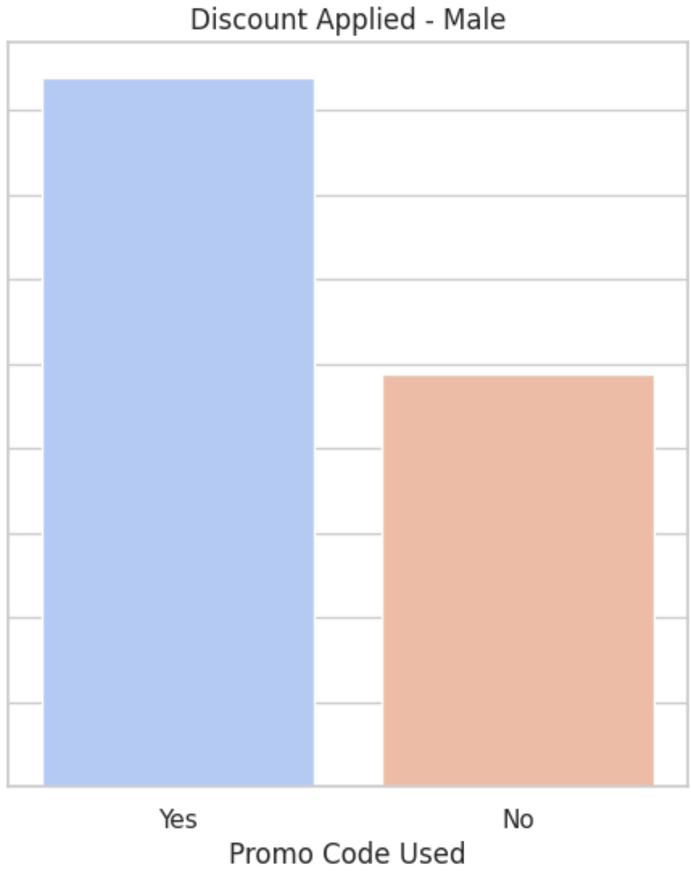




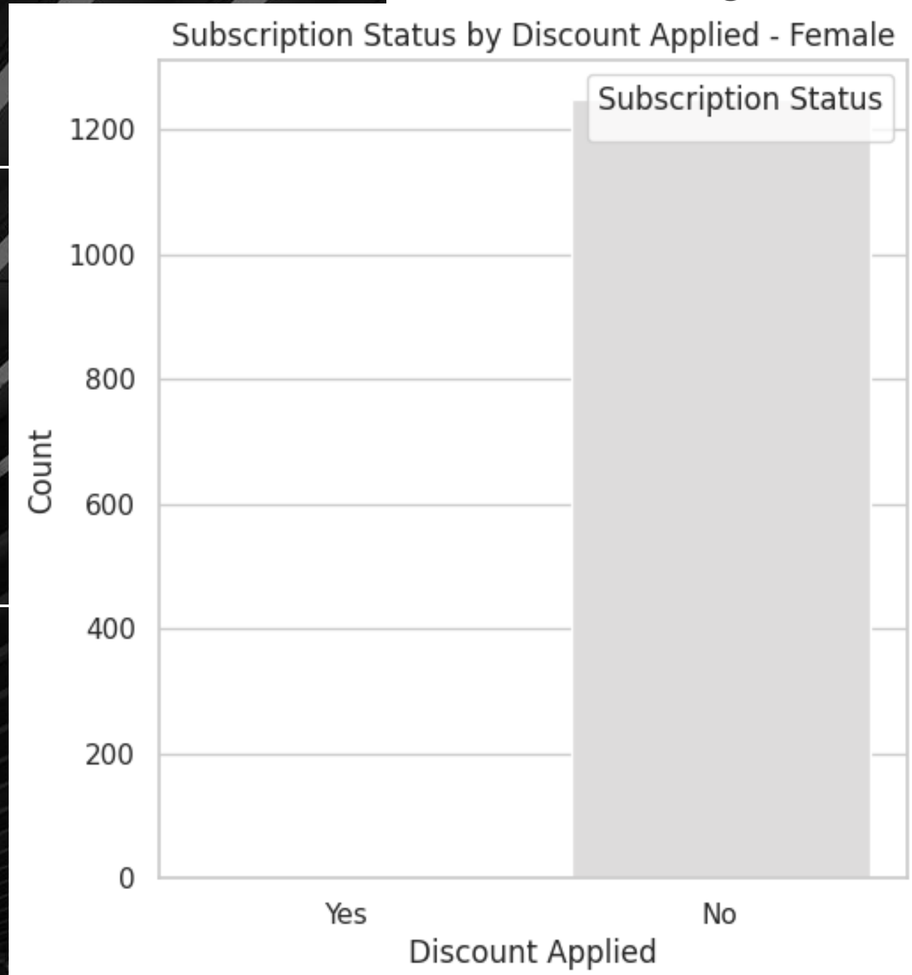


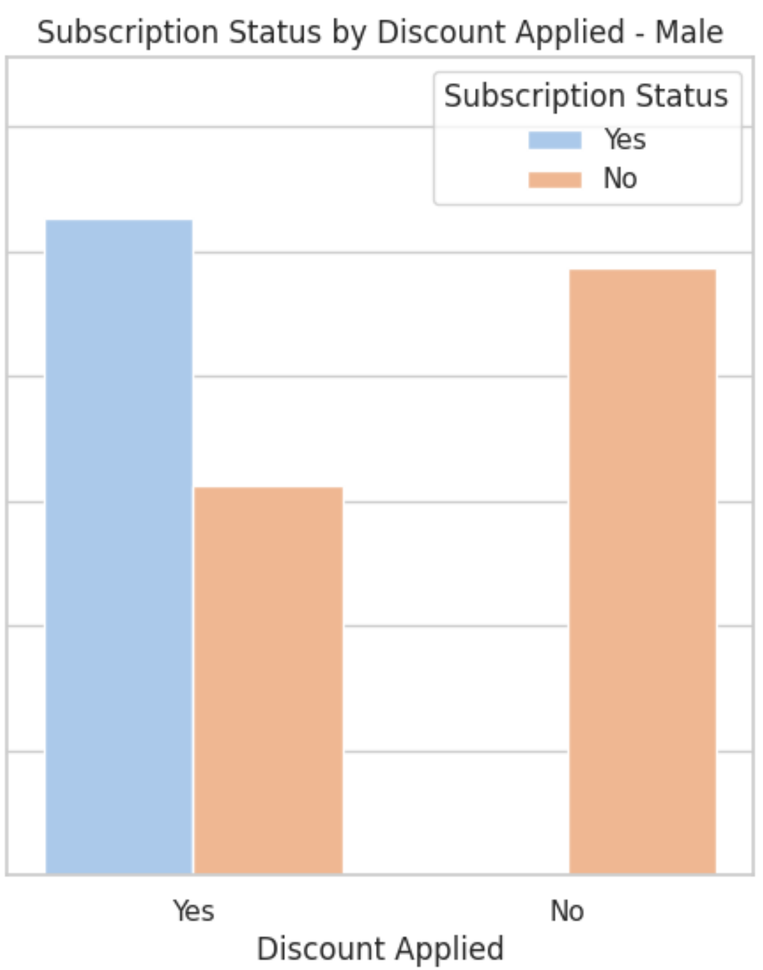
Discound Applied Used On Purchase Amount classified by Gender





Promo Code Used On Subscription Status classified by Gender





BCC
Basic Computing Community

DATA PREPROCESSING

MISSING VALUE OR DUPLICATE?

```
Customer ID                0
Age                        0
Gender                     0
Item Purchased             0
Category                   0
Purchase Amount (USD)      0
Location                   0
Size                       0
Color                      0
Season                     0
Review Rating              0
Subscription Status        0
Shipping Type              0
Discount Applied           0
Promo Code Used            0
Previous Purchases         0
Payment Method             0
Frequency of Purchases     0
dtype: int64
```

Any Missing Value?

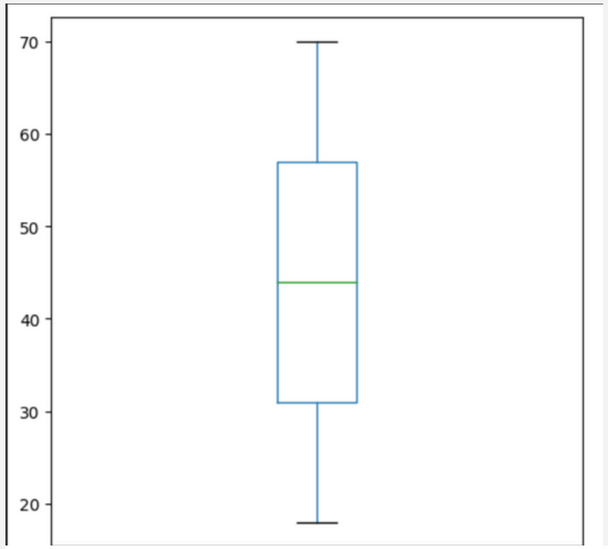
Any Duplicated Value?

```
[19] data.duplicated().sum()

     0
```

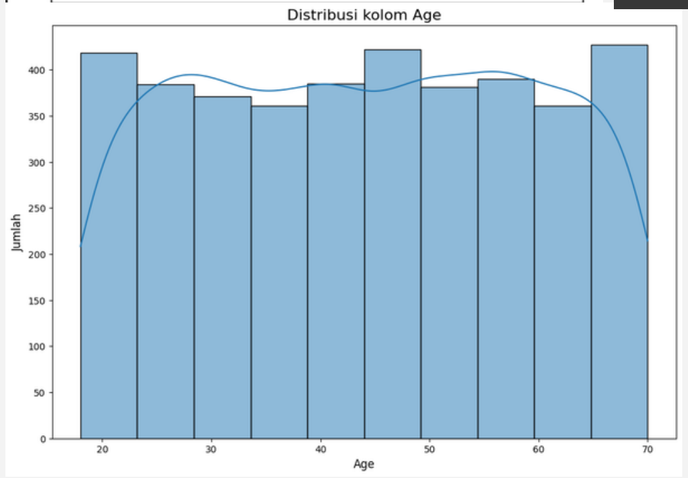
Even When Removed the Customer ID?

```
[20] duplicate_without_id = (data.drop(columns=["Customer ID"]))
     duplicate_without_id.duplicated().sum()

     0
```

CHECKING OUTLIER

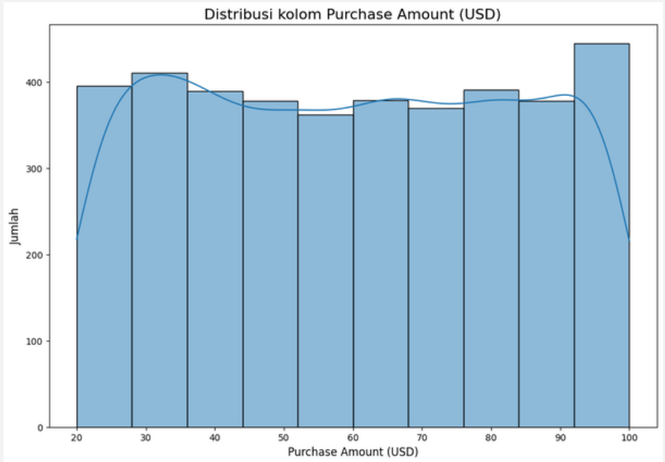


	Age
count	3900.000000
mean	44.068462
std	15.207589
min	18.000000
25%	31.000000
50%	44.000000
75%	57.000000
max	70.000000

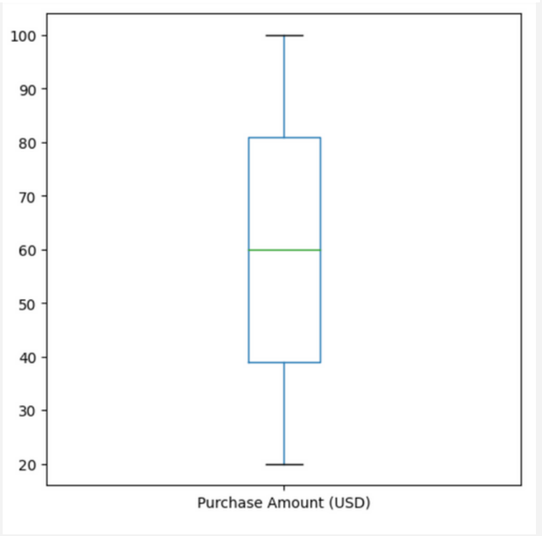


Age

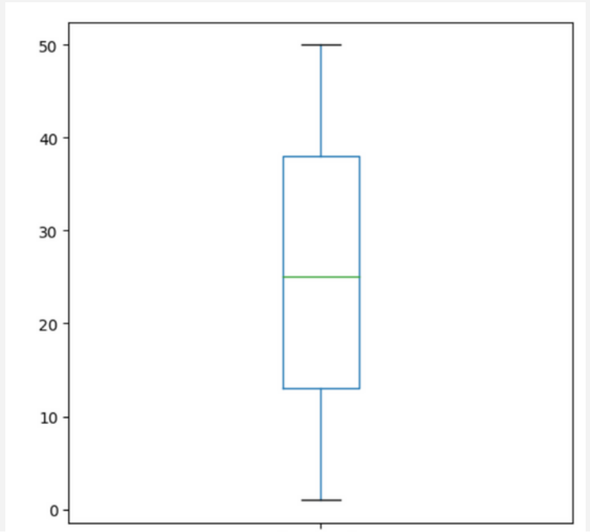
Purchase Amount (USD)



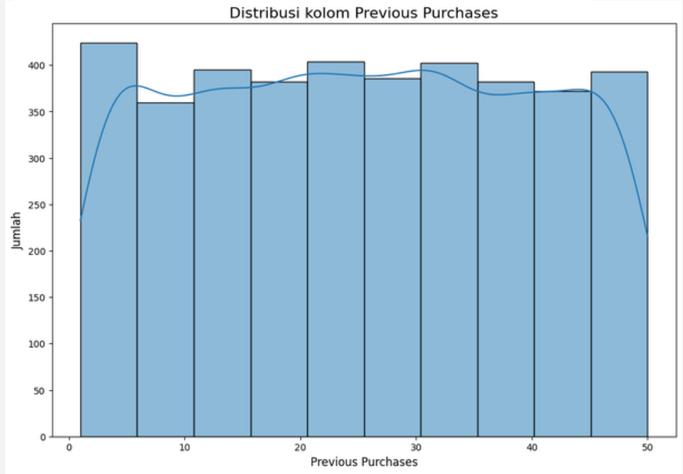
	Purchase Amount (USD)
count	3900.000000
mean	59.764359
std	23.685392
min	20.000000
25%	39.000000
50%	60.000000
75%	81.000000
max	100.000000



CHECKING OUTLIER

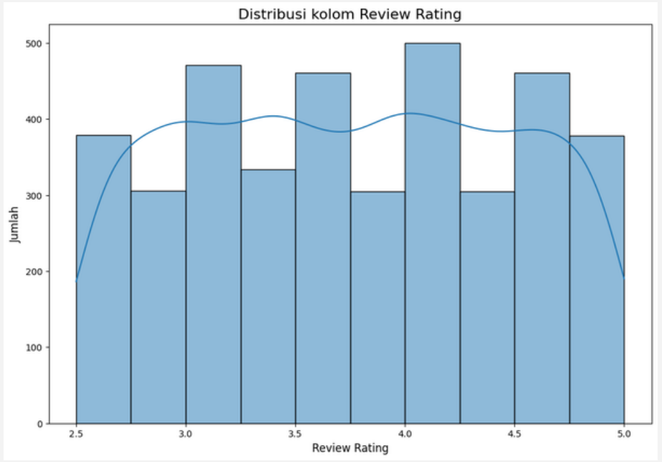


	Previous Purchases
count	3900.000000
mean	25.351538
std	14.447125
min	1.000000
25%	13.000000
50%	25.000000
75%	38.000000
max	50.000000

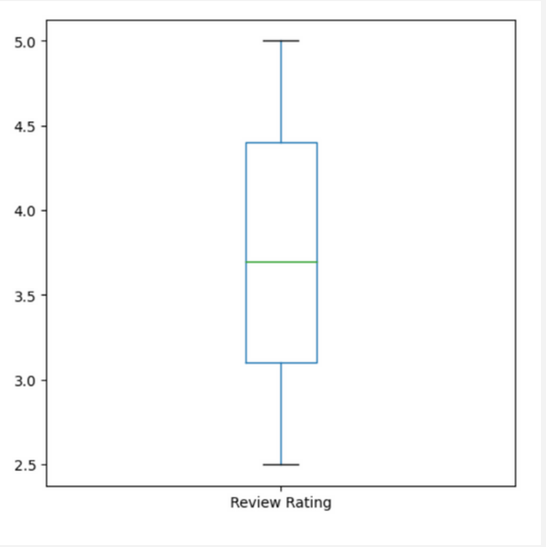


Previous Purchases

Review Rating



	Review Rating
count	3900.000000
mean	3.749949
std	0.716223
min	2.500000
25%	3.100000
50%	3.700000
75%	4.400000
max	5.000000



Categorical Column Handling

REMOVE USELESS FEATURE

```
nan_process_data = drop_column(data.copy(), ["Customer ID"])
nan_process_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3900 entries, 0 to 3899
Data columns (total 19 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Age                     3900 non-null   int64
 1   Gender                  3900 non-null   object
 2   Item Purchased          3900 non-null   object
 3   Category                3900 non-null   object
 4   Purchase Amount (USD)   3900 non-null   int64
 5   Location                3900 non-null   object
 6   Size                    3900 non-null   object
 7   Color                   3900 non-null   object
 8   Season                  3900 non-null   object
 9   Review Rating           3900 non-null   float64
 10  Subscription Status     3900 non-null   object
 11  Shipping Type           3900 non-null   object
 12  Discount Applied        3900 non-null   object
 13  Promo Code Used         3900 non-null   object
 14  Previous Purchases      3900 non-null   int64
 15  Payment Method          3900 non-null   object
 16  Frequency of Purchases  3900 non-null   object
 17  Region                  3900 non-null   object
 18  Age Group               3900 non-null   object
dtypes: float64(1), int64(3), object(15)
memory usage: 579.0+ KB
```

Type 1

Just removed the Customer ID

```
nan_encoding_data = processed_data = drop_column(data.copy(), ["Customer ID", "Item Purchased", "Color", "Location"])
processed_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3900 entries, 0 to 3899
Data columns (total 16 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Age                     3900 non-null   int64
 1   Gender                  3900 non-null   object
 2   Category                3900 non-null   object
 3   Purchase Amount (USD)   3900 non-null   int64
 4   Size                    3900 non-null   object
 5   Season                  3900 non-null   object
 6   Review Rating           3900 non-null   float64
 7   Subscription Status     3900 non-null   object
 8   Shipping Type           3900 non-null   object
 9   Discount Applied        3900 non-null   object
 10  Promo Code Used         3900 non-null   object
 11  Previous Purchases      3900 non-null   int64
 12  Payment Method          3900 non-null   object
 13  Frequency of Purchases  3900 non-null   object
 14  Region                  3900 non-null   object
 15  Age Group               3900 non-null   object
dtypes: float64(1), int64(3), object(12)
memory usage: 487.6+ KB
```

Type 2

Removed Unique Value Heavy Column

BCC Basic Computing Community

chase mount (USD)	Size	Season	Review Rating	Subscription Status	Shipping Type	Discount Applied	...	New England	Plains	Rocky Mountain	Southeast	Southwest	Adult	Elder	Middle_Age	Teenager	Young_Adult
53	0	3	3.1	1	1	1	...	0	0	0	1	0	0	0	1	0	0
64	0	3	3.1	1	1	1	...	1	0	0	0	0	0	0	0	1	0
73	2	1	3.1	1	2	1	...	1	0	0	0	0	0	0	1	0	0
90	1	1	3.5	1	3	1	...	1	0	0	0	0	0	0	0	0	1
49	1	1	2.7	1	2	1	...	0	0	0	0	0	0	0	1	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
28	0	2	4.2	0	0	0	...	0	0	0	1	0	1	0	0	0	0
49	0	1	4.5	0	5	0	...	0	1	0	0	0	0	0	1	0	0
33	0	1	2.9	0	4	0	...	0	0	0	0	0	0	0	1	0	0
77	2	2	3.8	0	1	0	...	0	1	0	0	0	1	0	0	0	0
81	1	1	3.1	0	5	0	...	0	0	0	0	0	0	0	1	0	0

Categorical Column Handling

One Hot Encoding

BCC
Basic Computing Community

Output Encoder

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	\
0	0.0	37.0	1	2	1	33.0	
1	1.0	1.0	1	23	1	44.0	
2	2.0	32.0	1	11	1	53.0	
3	3.0	3.0	1	14	2	70.0	
4	4.0	27.0	1	2	1	29.0	

	Location	Size	Color	Season	...	Subscription Status	Shipping Type	\
0	16	0	7	3	...	1	1	
1	18	0	12	3	...	1	1	
2	20	2	12	1	...	1	2	
3	38	1	12	1	...	1	3	
4	36	1	21	1	...	1	2	

	Discount Applied	Promo Code Used	Previous Purchases	Payment Method	\
0	1	1	13.0	5	
1	1	1	1.0	1	
2	1	1	22.0	2	
3	1	1	48.0	4	
4	1	1	30.0	4	

	Frequency of Purchases	Region	Age_group	Generation
0	3	6	0	1
1	3	3	3	2
2	6	3	0	1
3	6	3	4	2
4	0	0	0	1

```
# Exclude 'Customer_ID' from the encoding process
cols_to_encode = ['Gender', 'Item Purchased', 'Category', 'Location', 'Size', 'Color', 'Season',
                  'Subscription Status', 'Shipping Type', 'Discount Applied', 'Promo Code Used',
                  'Payment Method', 'Frequency of Purchases', 'Region', 'Age_group', 'Generation']
```

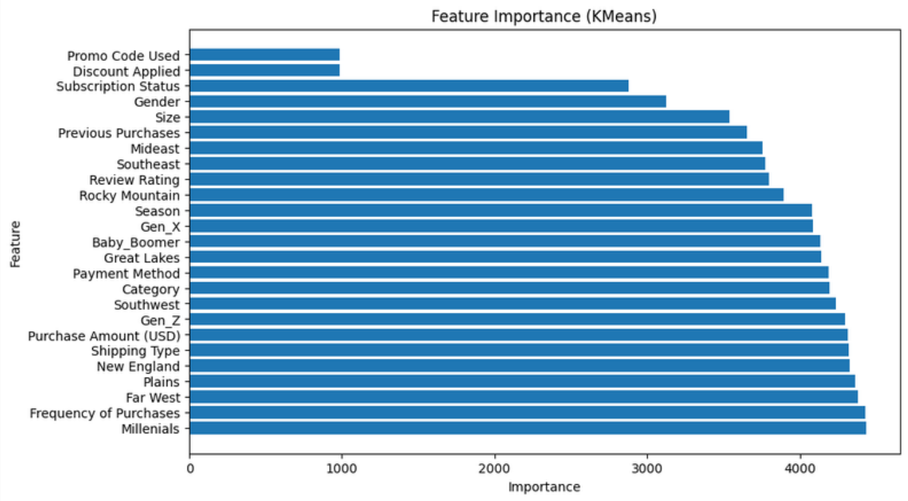
Categorical Column Handling

Using LabelEncoder and OrdinalEncoder For Object Columns

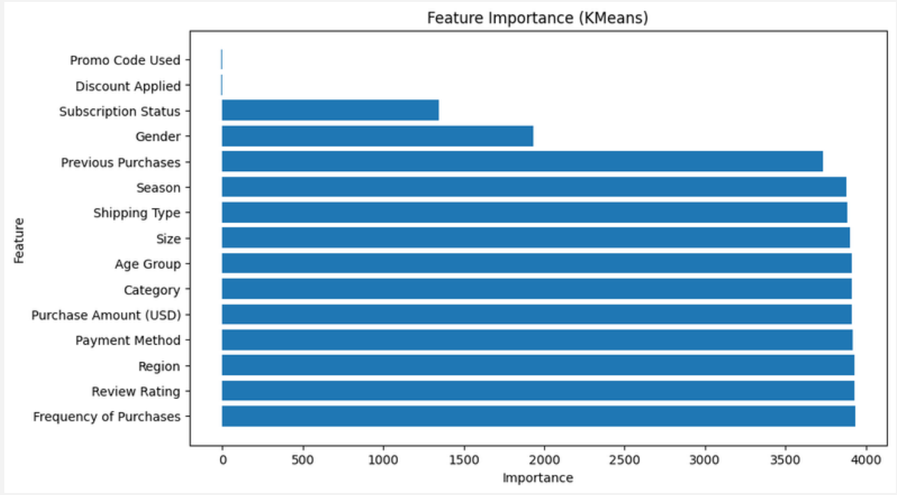
BCC
Basic Computing Community

FEATURE SELECTION

Model 1



Model 2

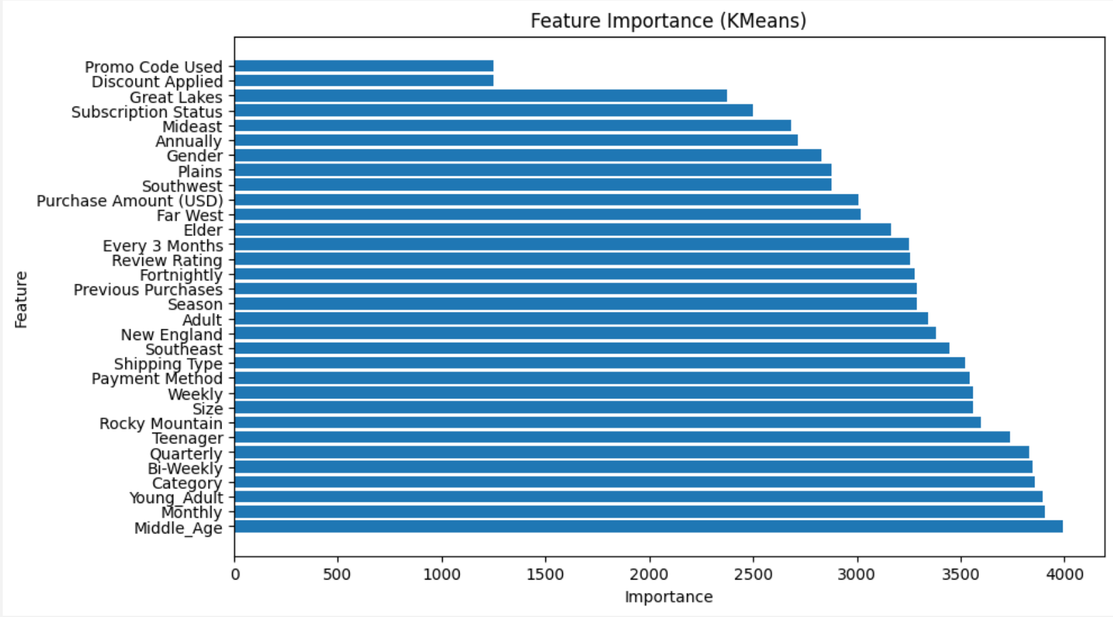


Feature Selection using KMeans Embedded

Yes/No Perform Lowest while Encoded Column Perform Better

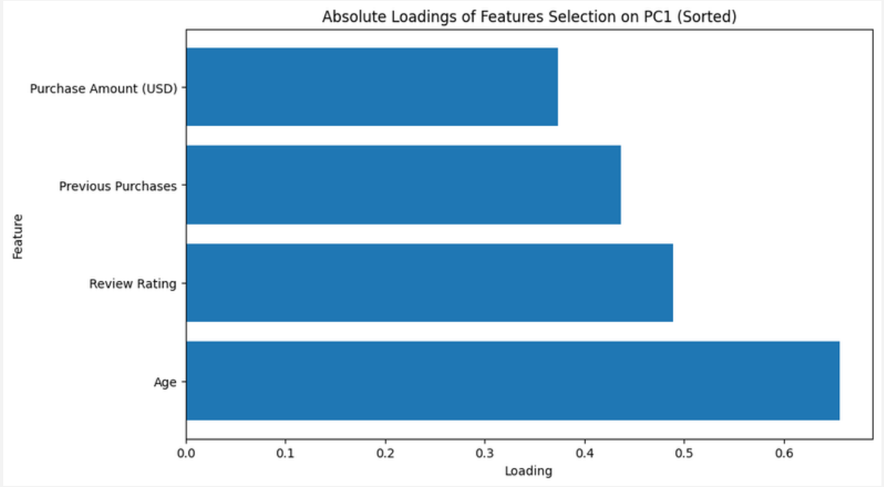
With “Frequency Of Purchases and Category Perform well

Model 3



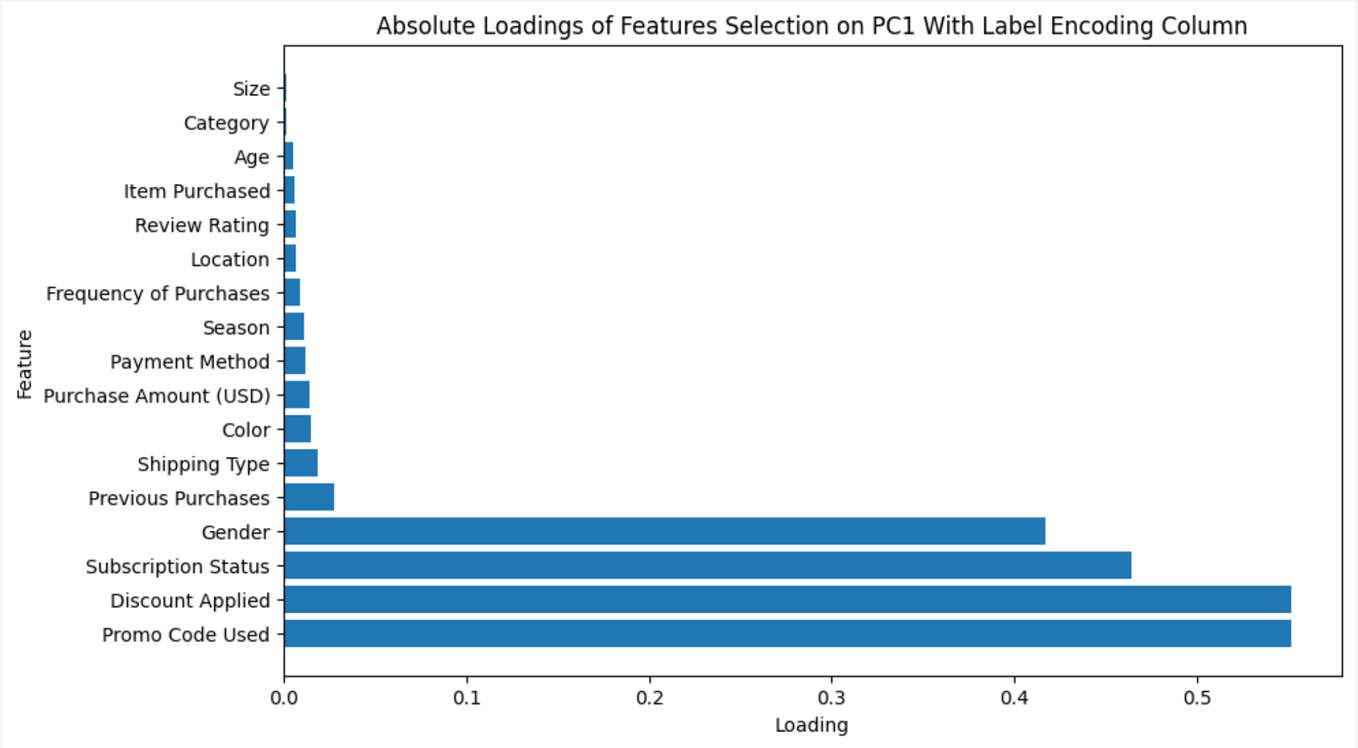
Basic Computing Community

Numerical Data



When loaded with encoded Column, the higher load PCA is Yes/No Colum like Promo Code and Gender

With Encoded Column

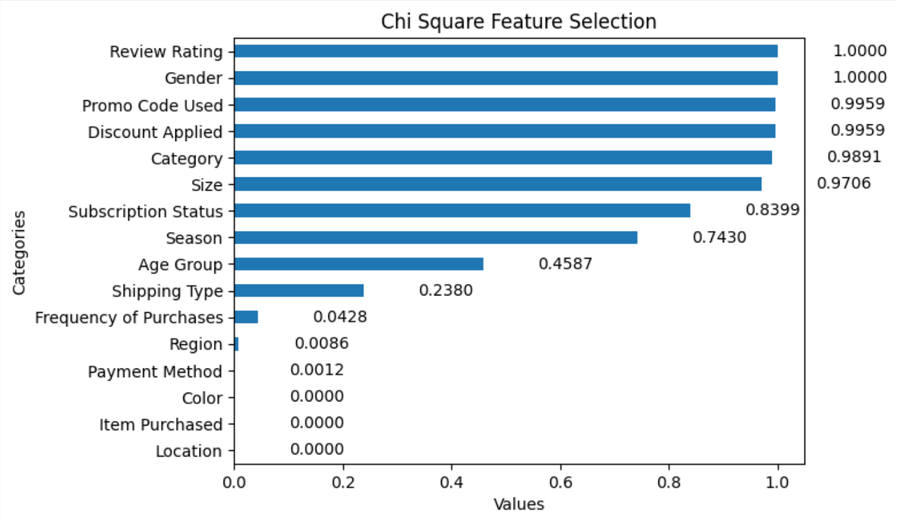


Feature Selection using PCA (Principal Component Analysis)

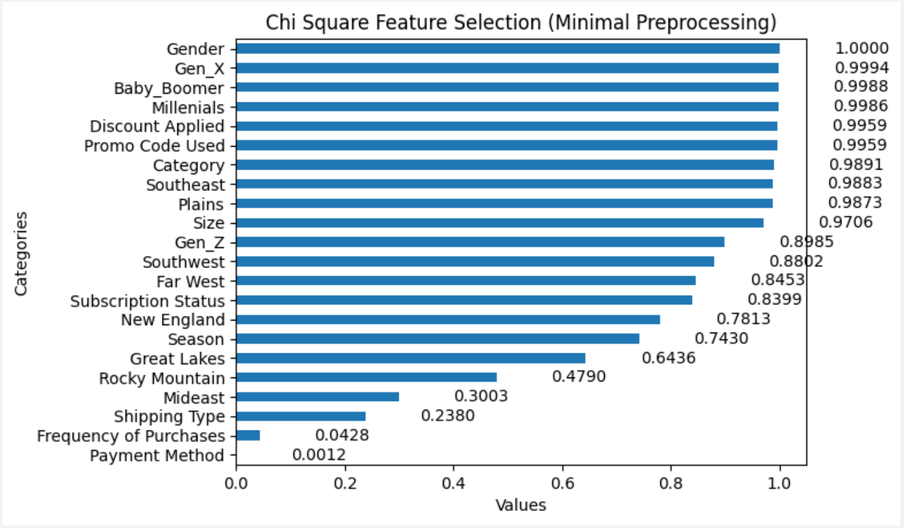
BCC
Basic Computing Community

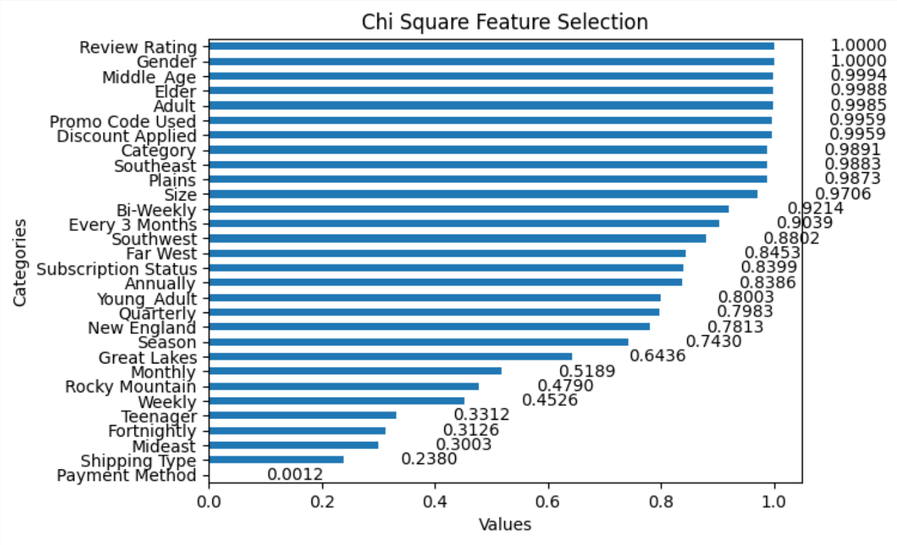
Feature Selection using Chi Square

Model 1



Model 2





Heavy Unique Column Perform Least While Column Yes/No Perform Well .

Especially Gender and Rating perform maximum

Feature like Category and Size also perform suprisingly well for 4 unique column

Model 3

BCC
Basic Computing Community

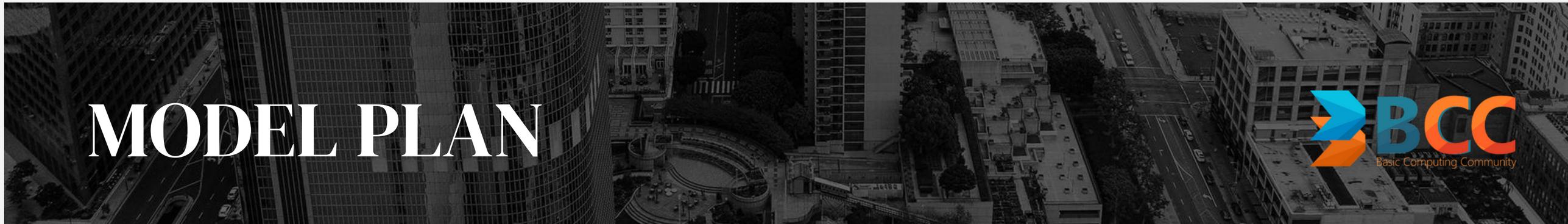
MODELLING

K-Means Clustering

Agglomerative Clustering

Gaussian Mixture Model Clustering

Extra Algorithm : K-Modes, K-Prototypes and
BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)



SUMMARY MODEL

MODEL 1 - 3

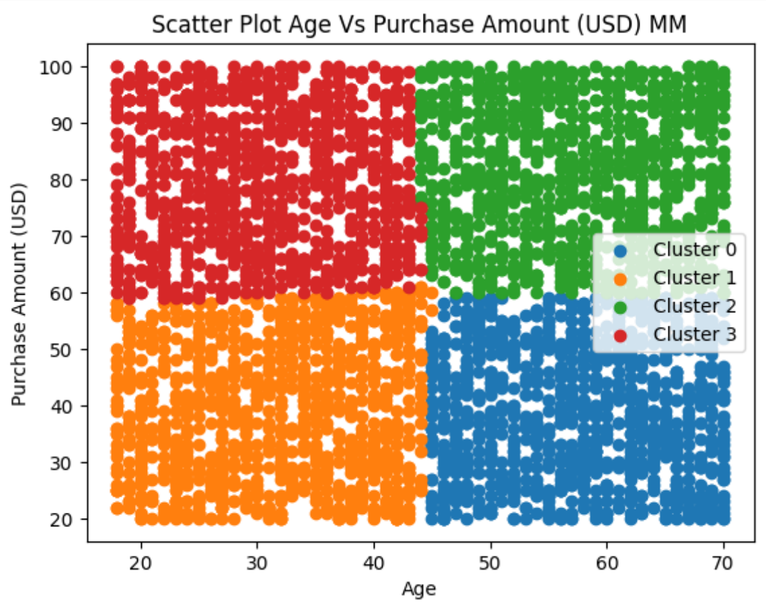
AGE and PURCHASE AMOUNT (USD)

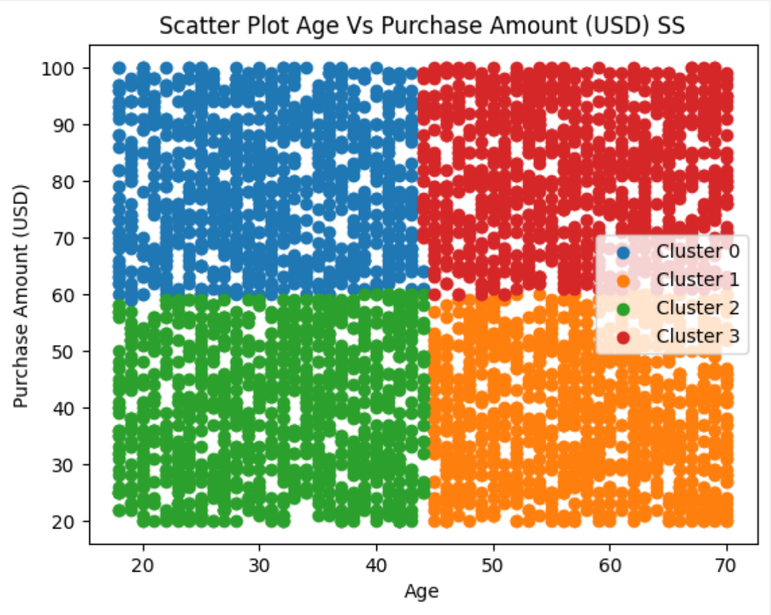
- Modelling with only two Numerical Column
- Explore minimal preprocessing, with and without One Hot Encoding
- All Defaulted n cluster = 4

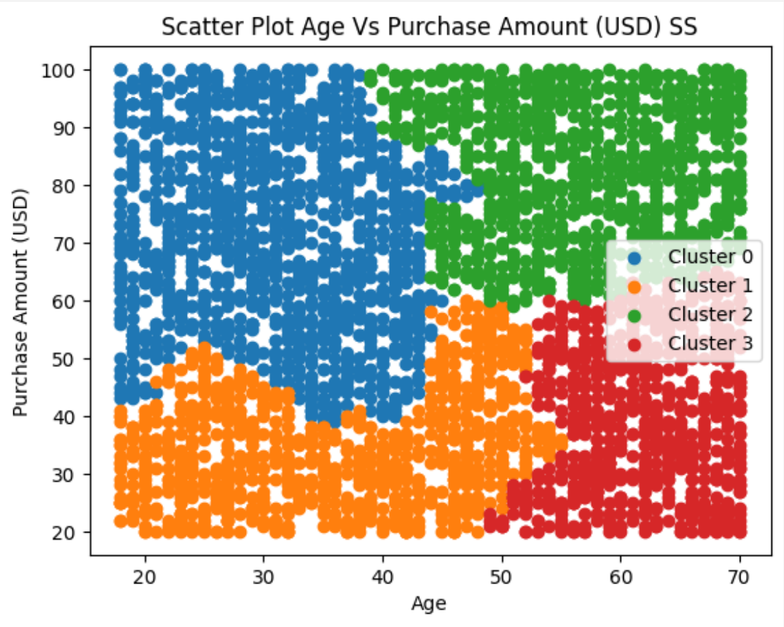
Basic Computing Community

Silhouette Score

	Model	Silhouette Score
0	K-Means Age Purchase Amount (USD) SS	0.413762
1	K-Means Age Purchase Amount (USD) MM	0.413725
4	Gaussian Age Purchase Amount (USD) SS	0.413624
5	Gaussian Age Purchase Amount (USD) MM	0.413615
2	AHC Age Purchase Amount (USD) SS	0.372396
3	AHC Age Purchase Amount (USD) MM	0.350095







CLUSTERING WITH ALL COLUMN

index	Model	Silhouette Score
13	(Nan Encoding Data) K-Means Age Purchase Amount (USD) MM	0.18108622717425715
17	(Nan Encoding Data) Gaussian Age Purchase Amount (USD) MM	0.14438121566990897
15	(Nan Encoding Data) AHC Age Purchase Amount (USD) MM	0.14024832660495704
7	(Nan Process Data) K-Means Age Purchase Amount (USD) MM	0.13985167740712495
11	(Nan Process Data) Gaussian Age Purchase Amount (USD) MM	0.13985167740712495
9	(Nan Process Data) AHC Age Purchase Amount (USD) MM	0.13985167740712495
12	(Nan Encoding Data) K-Means Age Purchase Amount (USD) SS	0.11644489929812246
14	(Nan Encoding Data) AHC Age Purchase Amount (USD) SS	0.11332625361246117
3	AHC Age Purchase Amount (USD) MM	0.10393431501712494
16	(Nan Encoding Data) Gaussian Age Purchase Amount (USD) SS	0.10116824307314544
6	(Nan Process Data) K-Means Age Purchase Amount (USD) SS	0.0952208825949715
10	(Nan Process Data) Gaussian Age Purchase Amount (USD) SS	0.08815866375583821
8	(Nan Process Data) AHC Age Purchase Amount (USD) SS	0.08779767459237334
1	K-Means Age Purchase Amount (USD) MM	0.08067032165953238
5	Gaussian Age Purchase Amount (USD) MM	0.07831641993729559
0	K-Means Age Purchase Amount (USD) SS	0.07110544092481591
2	AHC Age Purchase Amount (USD) SS	0.05697828500236625
4	Gaussian Age Purchase Amount (USD) SS	0.05567523670020354

- Nan Process Data (Model 1) : KMeans MinMax (0,139851677)
- Nan Encoding Data (Model 2) : KMeans MinMax (0,181086227)
- Process Data (Model 3) : AHC MinMax (0,103934)

CLUSTERING 5 SELECTED FEAT

index	Model	Silhouette Score
5	Gaussian Age Purchase Amount (USD) MM	0.38351937855911594
3	AHC Age Purchase Amount (USD) MM	0.34936552860693065
2	AHC Age Purchase Amount (USD) SS	0.31054216772224486
4	Gaussian Age Purchase Amount (USD) SS	0.279555022594238
1	K-Means Age Purchase Amount (USD) MM	0.26255164661723746
0	K-Means Age Purchase Amount (USD) SS	0.2302898888290238
17	(Nan Encoding Data) Gaussian Age Purchase Amount (USD) MM	0.13031039842186648
13	(Nan Encoding Data) K-Means Age Purchase Amount (USD) MM	0.12259522262309908
12	(Nan Encoding Data) K-Means Age Purchase Amount (USD) SS	0.1224045386584888
7	(Nan Process Data) K-Means Age Purchase Amount (USD) MM	0.11611496649685828
11	(Nan Process Data) Gaussian Age Purchase Amount (USD) MM	0.11147024466443575
16	(Nan Encoding Data) Gaussian Age Purchase Amount (USD) SS	0.10490328300530806
6	(Nan Process Data) K-Means Age Purchase Amount (USD) SS	0.09744135648669207
10	(Nan Process Data) Gaussian Age Purchase Amount (USD) SS	0.09541149986272633
15	(Nan Encoding Data) AHC Age Purchase Amount (USD) MM	0.08566164239264461
9	(Nan Process Data) AHC Age Purchase Amount (USD) MM	0.07017308434592164
14	(Nan Encoding Data) AHC Age Purchase Amount (USD) SS	0.06850392391454273
8	(Nan Process Data) AHC Age Purchase Amount (USD) SS	0.05321860058175387

- Nan Process Data (Model 1) : KMeans MinMax (0.11611496649)
- Nan Encoding Data (Model 2) : Gaussian MinMax (0.1303103984)
- Process Data (Model 3) : Gaussian MinMax (0.38351937)

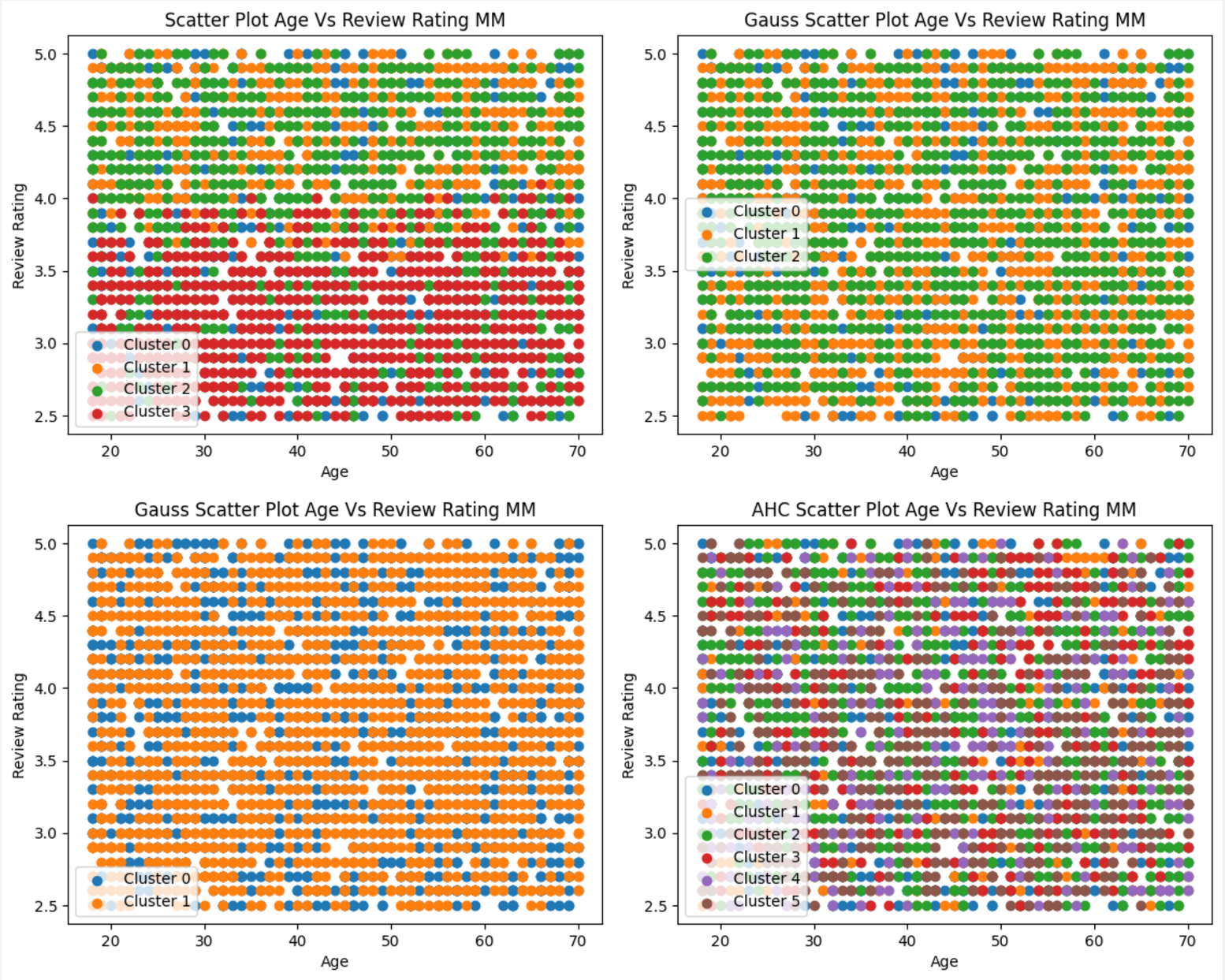
BCC
Basic Computing Community

MODEL 4

- Very Minimal Preprocessing and LabelEncoder
- Trying Fine Tuning N clusters

	Model	Silhouette Score
3	AHC Age Review Rating MM	0.433141
5	Gaussian Age Review Rating MM	0.433141
9	Gaussian Age Review Rating MM (3 Cluster)	0.387032
2	AHC Age Review Rating SS	0.282650
4	Gaussian Age Review Rating SS	0.282650
7	AHC Age Review Rating MM (6 Cluster)	0.251563
1	K-Means Age Review Rating MM	0.248276
8	Gaussian Age Review Rating SS (3 Cluster)	0.187877
0	K-Means Age Review Rating SS	0.175852
6	AHC Age Review Rating SS (6 Cluster)	0.124075

Silhouette Score

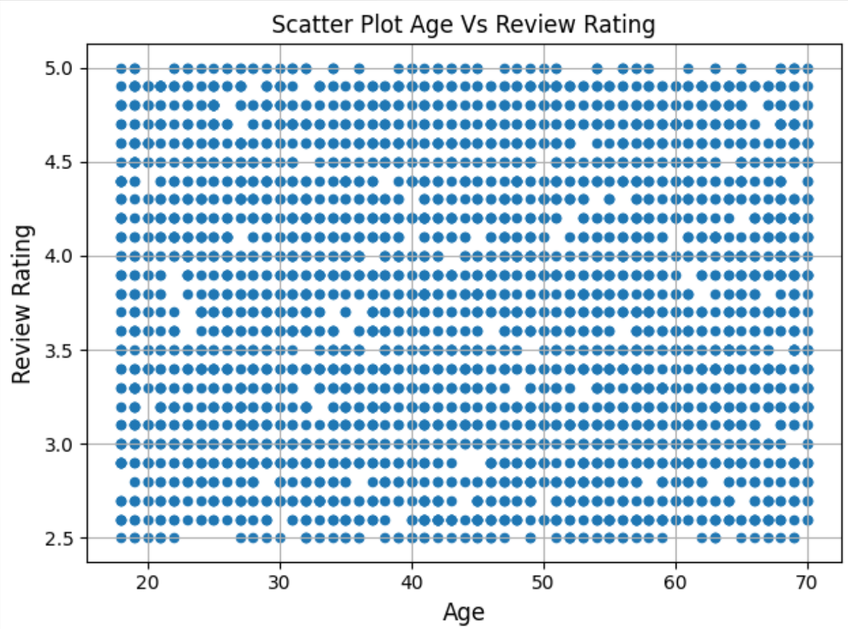


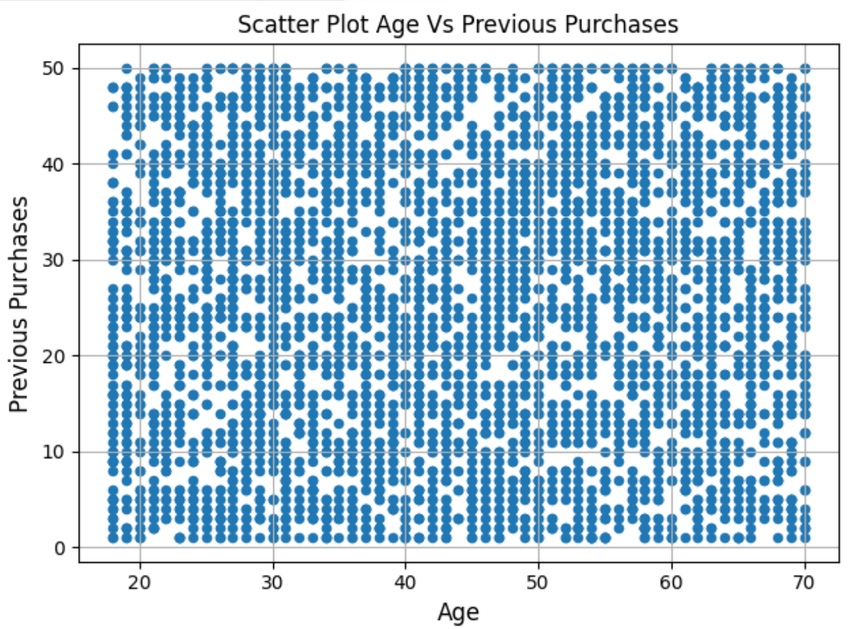
With Scaling (MinMax or Standard

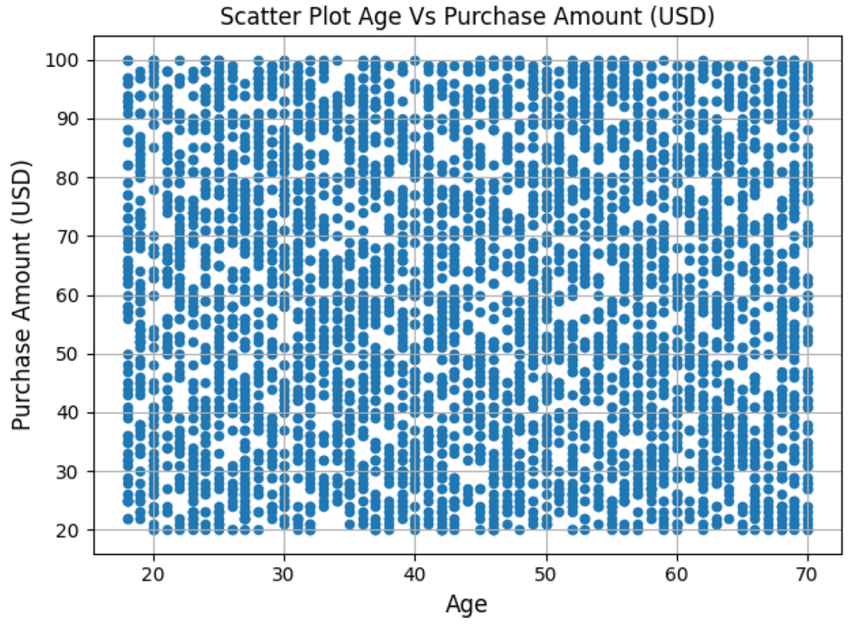
	Model	Silhouette Score
3	AHC Age Review Rating MM	0.433141
5	Gaussian Age Review Rating MM	0.433141
9	Gaussian Age Review Rating MM (3 Cluster)	0.387032
2	AHC Age Review Rating SS	0.282650
4	Gaussian Age Review Rating SS	0.282650
7	AHC Age Review Rating MM (6 Cluster)	0.251563
1	K-Means Age Review Rating MM	0.248276
8	Gaussian Age Review Rating SS (3 Cluster)	0.187877
0	K-Means Age Review Rating SS	0.175852
6	AHC Age Review Rating SS (6 Cluster)	0.124075

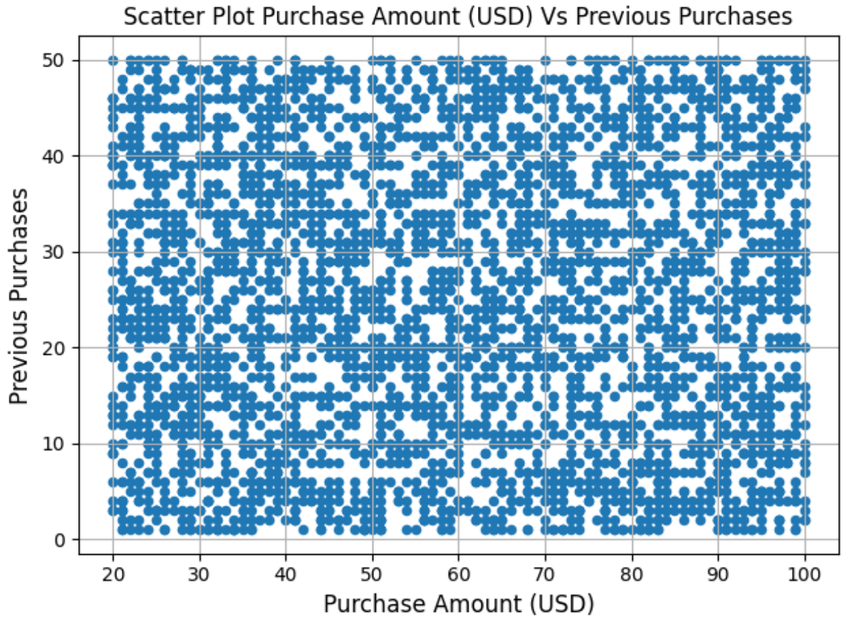
Without Scaling

	Model	Silhouette Score
7	BIRCH Age Review Rating	0.521878
0	K-Means Age Review Rating SS	0.516945
1	K-Means Age Review Rating MM	0.516945
4	Gaussian Age Review Rating SS	0.510841
5	Gaussian Age Review Rating MM	0.510841
2	AHC Age Review Rating SS	0.498850
3	AHC Age Review Rating MM	0.498850
9	BIRCH Age Review Rating (4)	0.497267
8	BIRCH Age Review Rating (All Column)	0.208245
6	KModes Age Review Rating SS	0.010154









Numerical Column Analysis

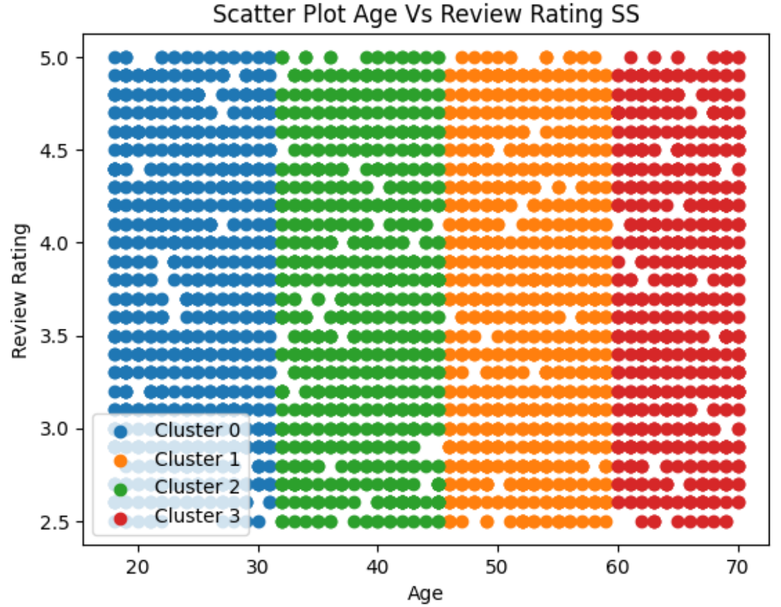
BCC
Basic Computing Community

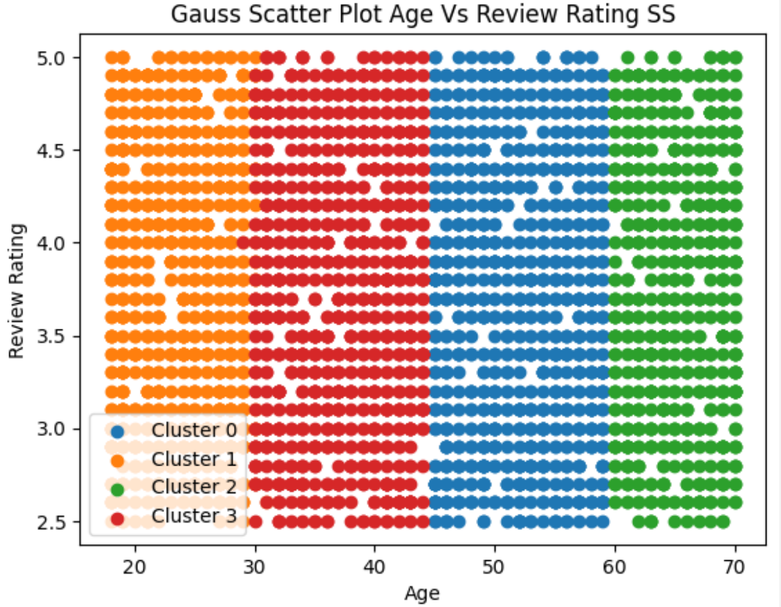
MODEL 5

- One Hot Encoding Region and Generation
- LabelEncoder
- Without Scaler

SELECTED FEATURES

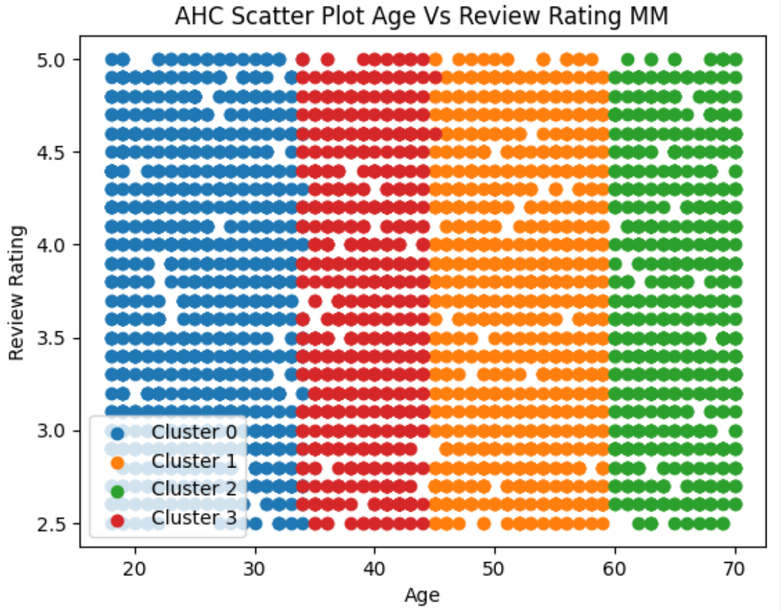
```
"Gender",
"Gen_X",
"Southeast",
"Category",
"Size",
"Age",
"Review Rating"
```





Silhouette Score

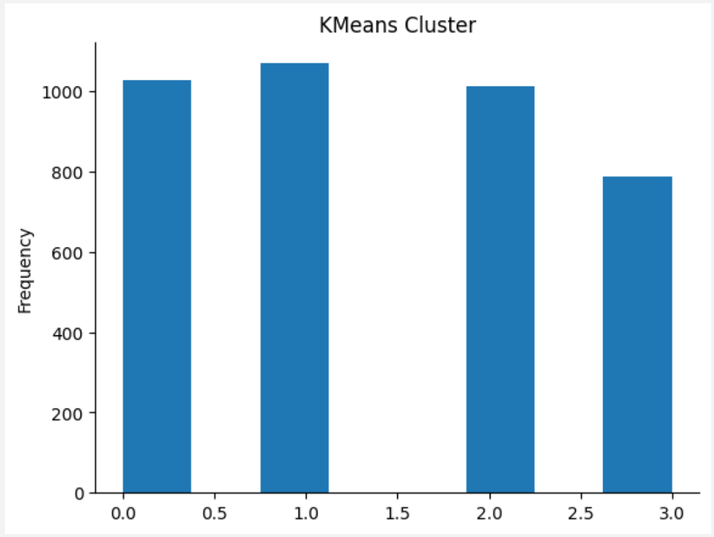
	Model	Silhouette Score
7	BIRCH Age Review Rating	0.521878
0	K-Means Age Review Rating SS	0.516945
1	K-Means Age Review Rating MM	0.516945
4	Gaussian Age Review Rating SS	0.510841
5	Gaussian Age Review Rating MM	0.510841
2	AHC Age Review Rating SS	0.498850
3	AHC Age Review Rating MM	0.498850
9	BIRCH Age Review Rating (4)	0.497267
8	BIRCH Age Review Rating (All Column)	0.208245
6	KModes Age Review Rating SS	0.010154



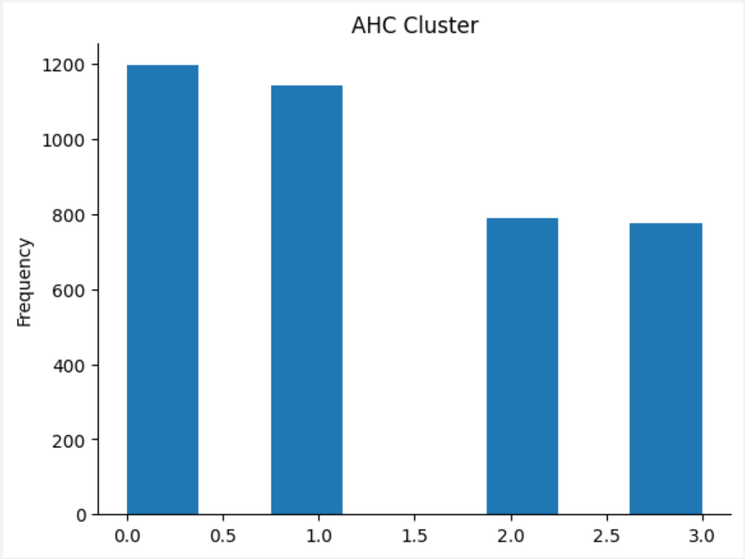
MODEL 5

SELECTED FEATURES

```
"Gender",
"Gen_X",
"Southeast",
"Category",
"Size",
"Age",
"Review Rating"
```




```
array([1028, 1070, 1014,  788])
```




```
array([1196, 1141,  788,  775])
```

```
array([1142,  906,  788, 1064])
```

MODEL 5

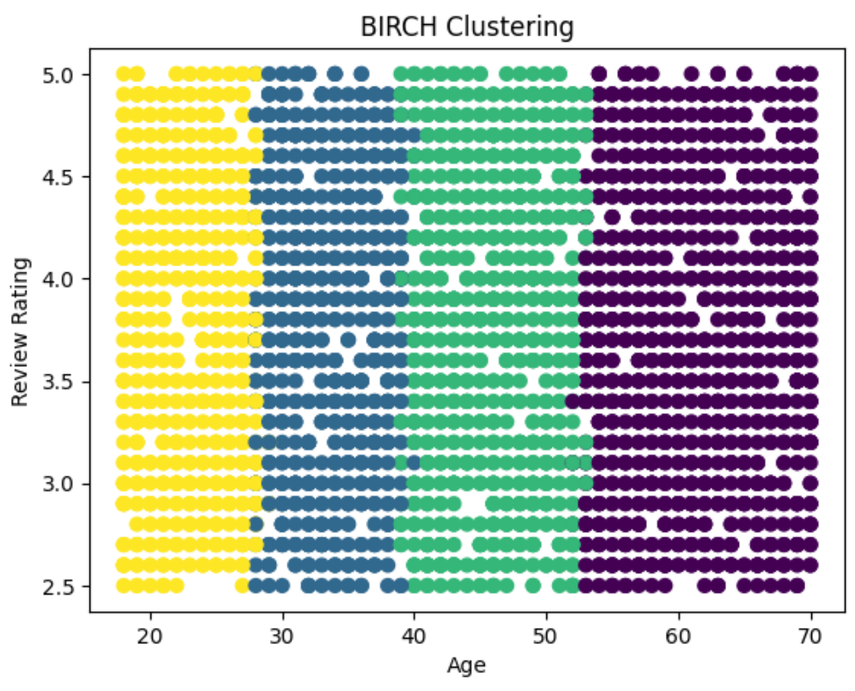
SELECTED FEATURES

```
"Gender",
"Gen_X",
"Southeast",
"Category",
"Size",
"Age",
"Review Rating"
```

K-Modes


```
array([2264,  770,  540,  326])

Silhouette Score: 0.010154210565084898
```

BIRCH




```
array([1318,  820,  992,  770])

Silhouette Score: 0.49726741691851695
```

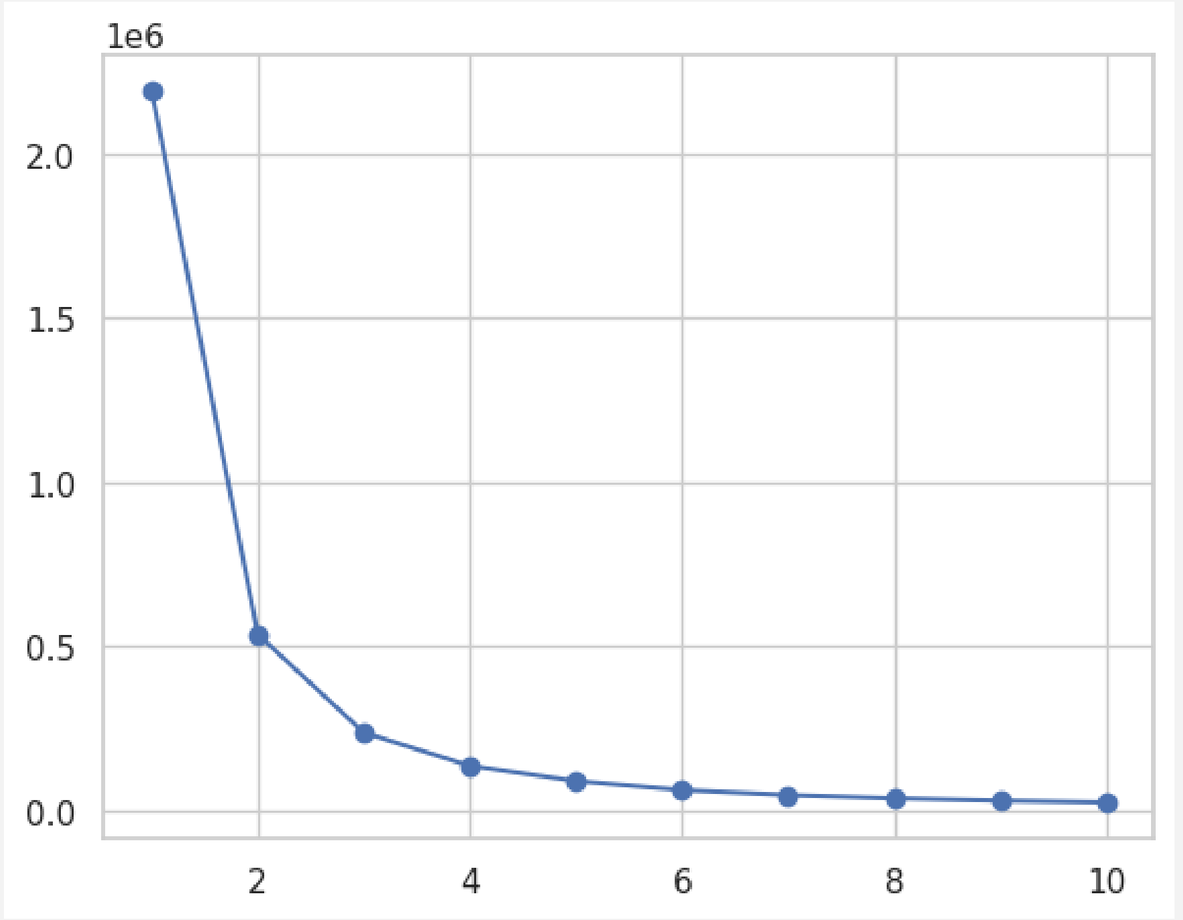
Modelling (6) for Generation and Purchase Amount(No-Scaling)

```
[249] X_column = 'Generation'
      y_column = 'Purchase Amount (USD)'

      clus_score = pd.DataFrame(columns=["Model", "Silhouette Score"])
```

```
X = df_encoded[['Generation', 'Purchase Amount (USD)']]
X.head()
```

	Generation	Purchase Amount (USD)
0	1	33.0
1	2	44.0
2	1	53.0
3	2	70.0
4	1	29.0

Elbow Method to Determine the k for Number of Clusters



K-Means Algorithm

```
[259] km = KMeans(n_clusters=4, random_state=45, n_init="auto")
      km.fit(X)

      y_km = km.predict(X)
```

```
[260] pd.DataFrame(y_km).value_counts()
```

```
1    1024
2     987
3     971
0     918
dtype: int64
```

Cluster Segmentation

Silhouette Score

	Model	Silhouette Score
0	K-Means No Scaling	0.560821

	KMeans Cluster with No Scaling	Gender	Generation	Category	Purchase Amount (USD)	Region
0	0	Male	Gen X	Clothing	53	Southeast
1	2	Male	Gen Z	Clothing	64	New England
2	2	Male	Gen X	Clothing	73	New England
3	1	Male	Gen Z	Footwear	90	New England
4	0	Male	Gen X	Clothing	49	Far West
...	...	...	...	...	...	...
3895	3	Female	Millenials	Clothing	28	Southeast
3896	0	Female	Gen X	Accessories	49	Plains
3897	3	Female	Gen X	Accessories	33	Mideast
3898	2	Female	Millenials	Footwear	77	Plains
3899	1	Female	Gen X	Accessories	81	Far West

[3900 rows x 6 columns]

Agglomerative Algorithm

```
[264] agg_model = AgglomerativeClustering(n_clusters=4)

      agg_model.fit(X)

      y_agg = agg_model.labels_
```

```
[265] pd.DataFrame(y_agg).value_counts()

      0    1513
      1     977
      2     743
      3     667
      dtype: int64
```

Cluster Segmentation

Silhouette Score

	Model	Silhouette Score
0	Agglomerative No-Scaling	0.501314

	Agglomerative Cluster with No Scaling	Gender	Generation	Category	Purchase Amount (USD)	Region
0	1	Male	Gen X	Clothing	53	Southeast
1	3	Male	Gen Z	Clothing	64	New England
2	0	Male	Gen X	Clothing	73	New England
3	0	Male	Gen Z	Footwear	90	New England
4	1	Male	Gen X	Clothing	49	Far West
...	...	...	...	...	...	...
3895	2	Female	Millenials	Clothing	28	Southeast
3896	1	Female	Gen X	Accessories	49	Plains
3897	2	Female	Gen X	Accessories	33	Mideast
3898	0	Female	Millenials	Footwear	77	Plains
3899	0	Female	Gen X	Accessories	81	Far West

[3900 rows x 6 columns]

GMM Algorithm

```
[294] gmm = GaussianMixture(n_components=4, random_state=42)

      gmm.fit(X)

      labels = gmm.predict(X)
```

```
pd.DataFrame(labels).value_counts()
```

```
3    123
2    119
0     91
1     63
dtype: int64
```

Cluster Segmentation

BCC
Basic Computing Community

	GMM Cluster with No Scaling	Gender	Generation	Category	Purchase Amount (USD)	Region
0	3	Male	Gen X	Clothing	53	Southeast
1	1	Male	Gen Z	Clothing	64	New England
2	3	Male	Gen X	Clothing	73	New England
3	1	Male	Gen Z	Footwear	90	New England
4	1	Male	Gen X	Clothing	49	Far West
..	...	...	...	...	...	...
391	0	Male	Millenials	Clothing	86	Plains
392	3	Male	Gen Z	Clothing	82	Rocky Mountain
393	1	Male	Millenials	Footwear	65	New England
394	0	Male	Millenials	Accessories	29	Plains
395	0	Male	Millenials	Footwear	65	New England

[396 rows x 6 columns]

Silhouette Score

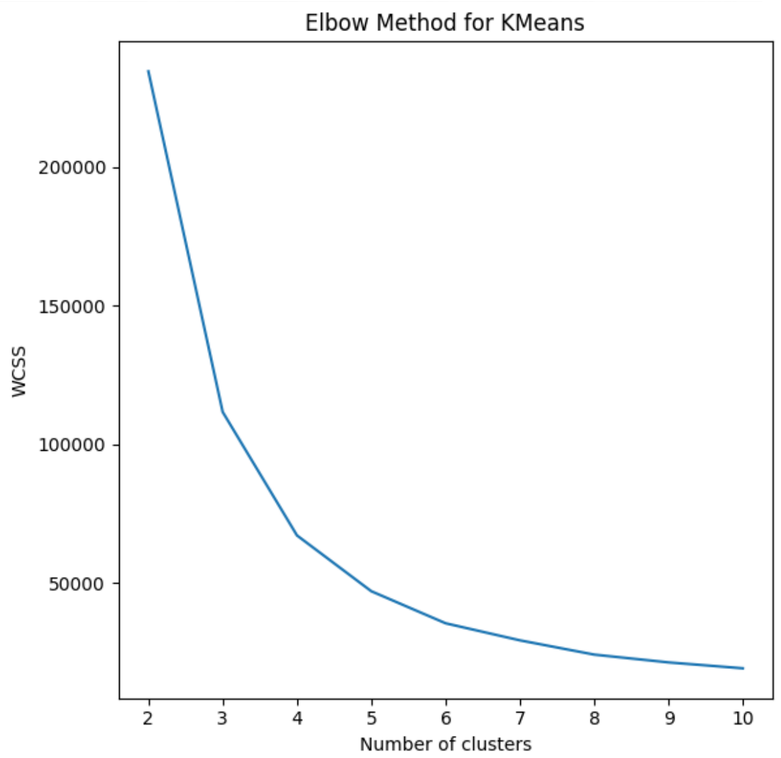
	Model	Silhouette Score
1	Gaussian Generation Purchase Amount (USD) No-S...	0.559333

BCC
Basic Computing Community

METRIC EVALUATION

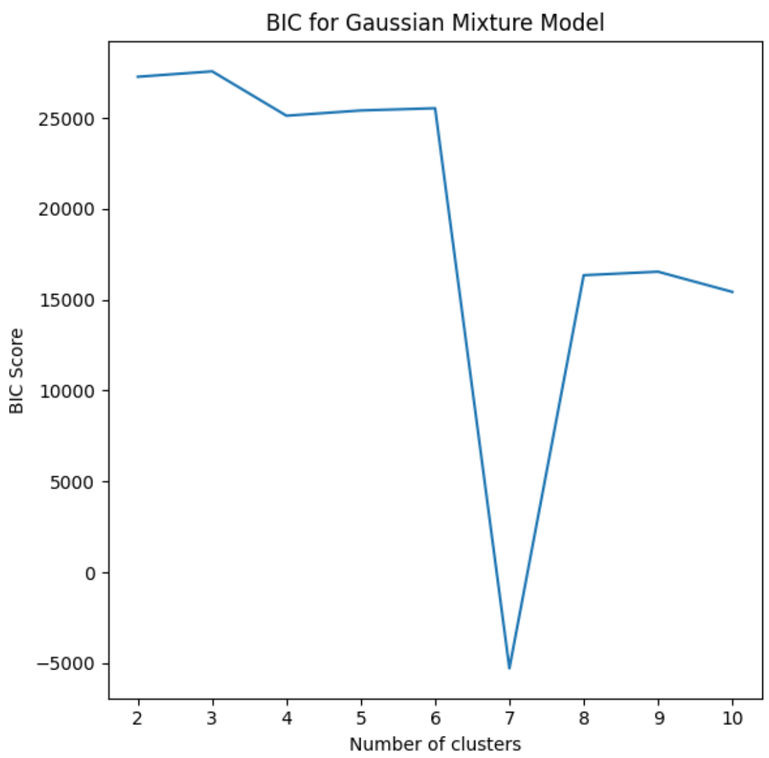
Model 5 Fine Tuning

Elbow Method KMeans



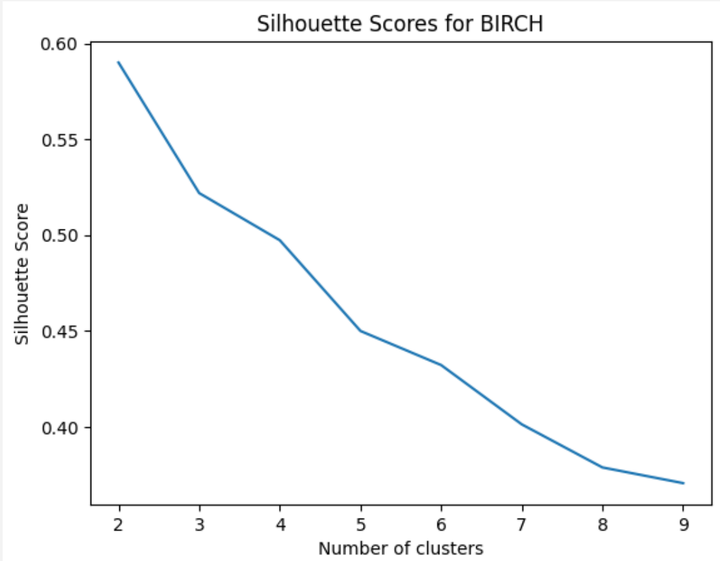
Silhouette Score Selection

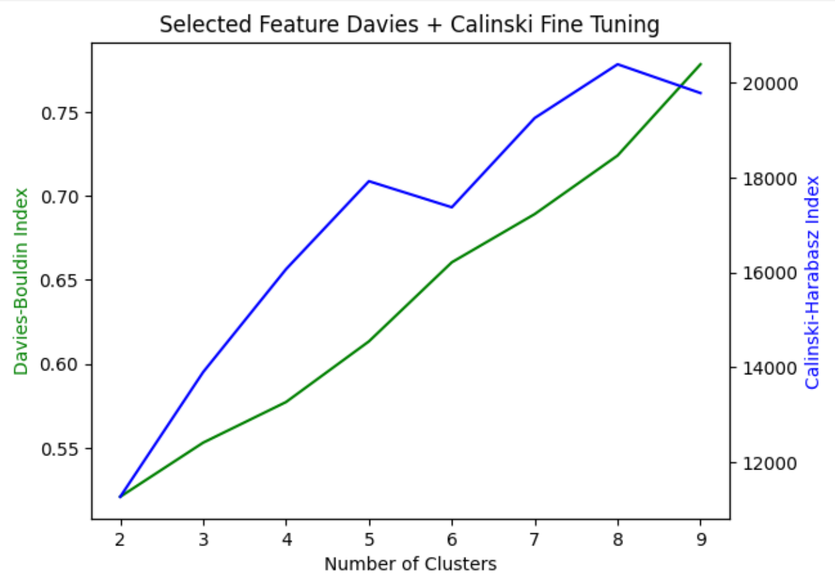
BIC GMM

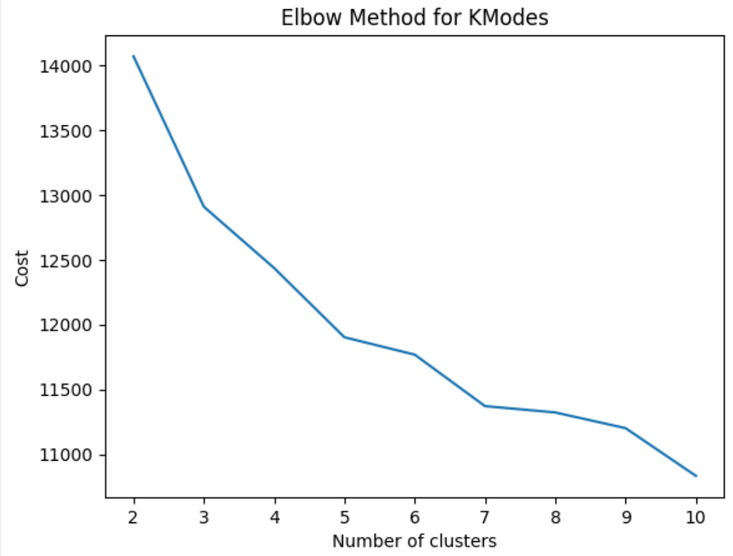


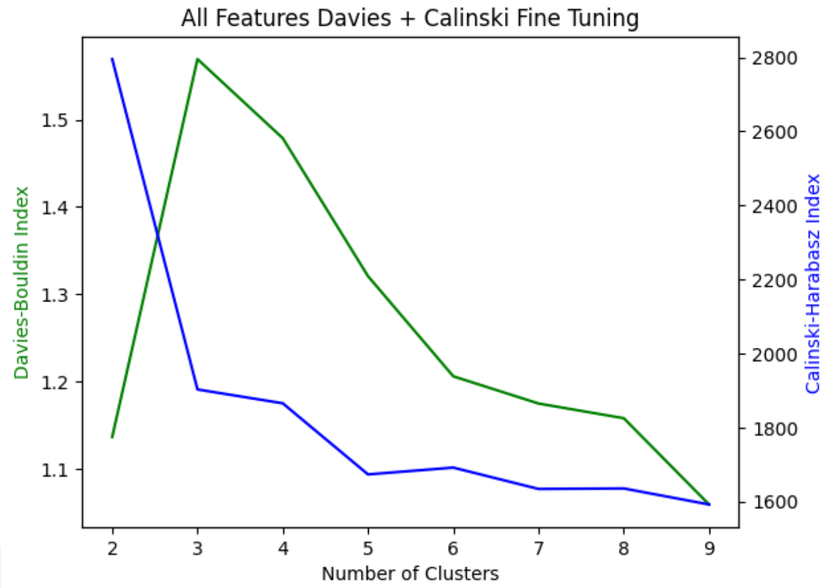
Model 5 Fine Tuning

- Silhouette Score fr BIRCH and K-Modes
- Davies and Calinski Index





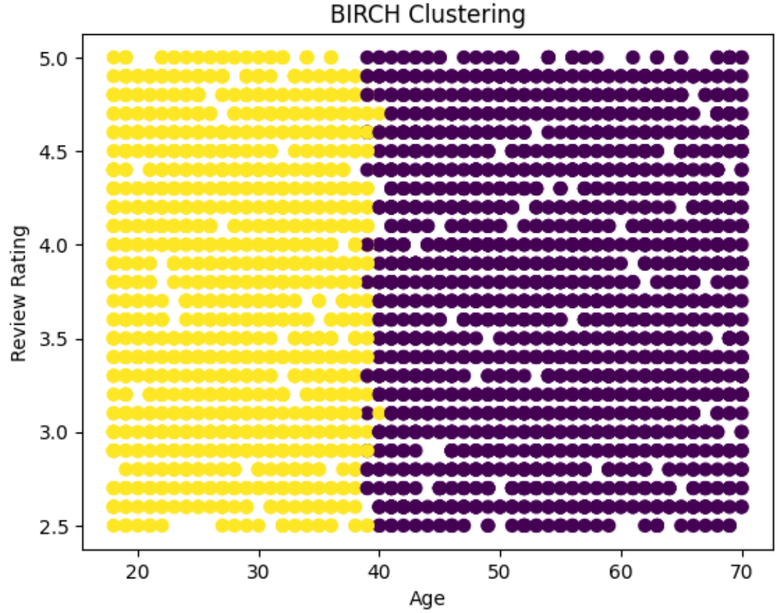


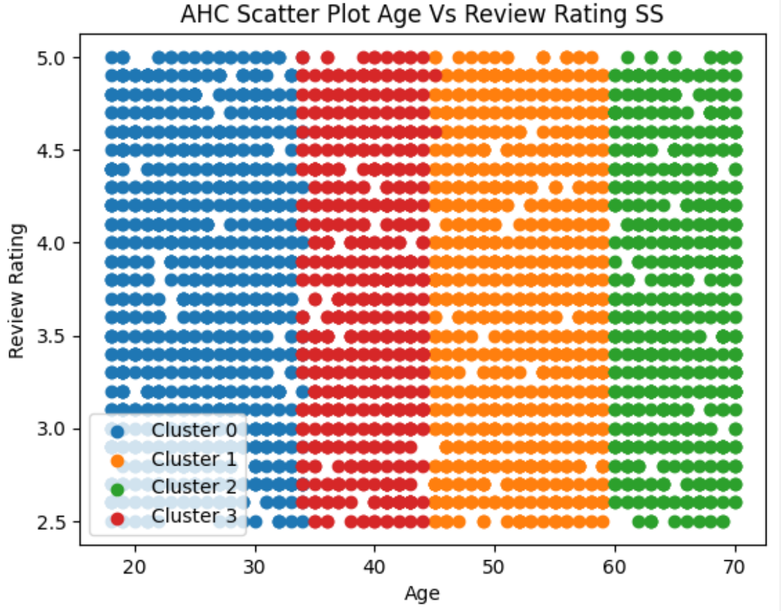


BCC
Basic Computing Community

TUNED MODEL 5

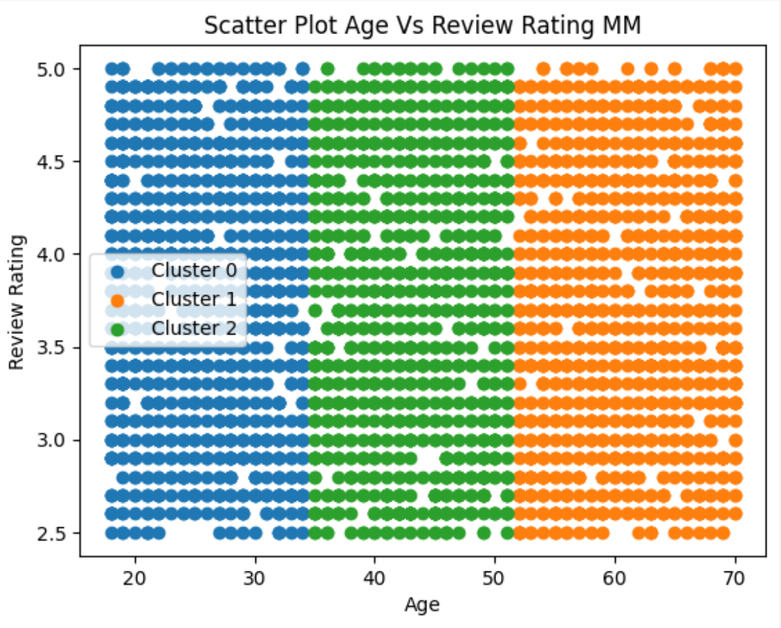
```
"Gender",
"Gen_X",
"Southeast",
"Category",
"Size",
"Age",
"Review Rating"
```





	Model	Silhouette Score
8	BIRCH Age Review Rating (4)	0.589970
0	K-Means Age Review Rating SS	0.552705
2	AHC Age Review Rating SS	0.498850

Silhouette Score of Our Best Model



BCC
Basic Computing Community

BCC
Basic Computing Community

CONCLUSION & EVALUATION

Consumer Segmentation with K-Means No-Scaling Model

Cluster 0:

	KMeans Cluster	Gender	Age	Generation	Item Purchased	Category	Purchase Amount (USD)	Location	Region
0	0	Male	55	Gen X	Blouse	Clothing	53	Kentucky	Southeast
4	0	Male	45	Gen X	Blouse	Clothing	49	Oregon	Far West
13	0	Male	65	Baby Boomer	Dress	Clothing	51	New Hampshire	New England
14	0	Male	64	Baby Boomer	Coat	Outerwear	53	New York	Mideast
18	0	Male	52	Gen X	Sweater	Clothing	48	Montana	Rocky Mountain

Cluster 1:

	KMeans Cluster	Gender	Age	Generation	Item Purchased	Category	Purchase Amount (USD)	Location	Region
3	1	Male	21	Gen Z	Sandals	Footwear	90	Rhode Island	New England
6	1	Male	63	Baby Boomer	Shirt	Clothing	85	Montana	Rocky Mountain
8	1	Male	26	Gen Z	Coat	Outerwear	97	West Virginia	Southeast
15	1	Male	64	Baby Boomer	Skirt	Clothing	81	Rhode Island	New England
19	1	Male	66	Baby Boomer	Pants	Clothing	90	Rhode Island	New England

Cluster 2:

	KMeans Cluster	Gender	Age	Generation	Item Purchased	Category	Purchase Amount (USD)	Location	Region
1	2	Male	19	Gen Z	Sweater	Clothing	64	Maine	New England
2	2	Male	50	Gen X	Jeans	Clothing	73	Massachusetts	New England
11	2	Male	30	Millenials	Shorts	Clothing	68	Hawaii	Far West
12	2	Male	61	Baby Boomer	Coat	Outerwear	72	Delaware	Mideast
21	2	Male	31	Millenials	Pants	Clothing	62	North Carolina	Southeast

Cluster 3:

	KMeans Cluster	Gender	Age	Generation	Item Purchased	Category	Purchase Amount (USD)	Location	Region
5	3	Male	46	Gen X	Sneakers	Footwear	20	Wyoming	Rocky Mountain
7	3	Male	27	Gen Z	Shorts	Clothing	34	Louisiana	Southeast
9	3	Male	57	Gen X	Handbag	Accessories	31	Missouri	Plains
10	3	Male	53	Gen X	Shoes	Footwear	34	Arkansas	Southeast
16	3	Male	25	Gen Z	Sunglasses	Accessories	36	Alabama	Southeast

- **Cluster 0 : Male, 45-65, Gen X/Baby Boomer, Clothing, 48-53**
- **Cluster 1 : Male, 63-66, Gen Z/Baby Boomer, Clothing, 81-97**
- **Cluster 2 : Male, 31-50, Gen X/Millenials, Clothing, 64-73**
- **Cluster 3 : Male, 25-46, Gen X/Gen Z, Footwear, 20-36**

Consumer Segmentation with AHC No-Scaling Model

Cluster 0:

	Agglomerative Cluster	Gender	Age	Generation	Item Purchased	Category	Purchase Amount (USD)	Location	Region
2	0	Male	50	Gen X	Jeans	Clothing	73	Massachusetts	New England
3	0	Male	21	Gen Z	Sandals	Footwear	90	Rhode Island	New England
6	0	Male	63	Baby Boomer	Shirt	Clothing	85	Montana	Rocky Mountain
8	0	Male	26	Gen Z	Coat	Outerwear	97	West Virginia	Southeast
12	0	Male	61	Baby Boomer	Coat	Outerwear	72	Delaware	Mideast

Cluster 1:

	Agglomerative Cluster	Gender	Age	Generation	Item Purchased	Category	Purchase Amount (USD)	Location	Region
0	1	Male	55	Gen X	Blouse	Clothing	53	Kentucky	Southeast
4	1	Male	45	Gen X	Blouse	Clothing	49	Oregon	Far West
7	1	Male	27	Gen Z	Shorts	Clothing	34	Louisiana	Southeast
13	1	Male	65	Baby Boomer	Dress	Clothing	51	New Hampshire	New England
14	1	Male	64	Baby Boomer	Coat	Outerwear	53	New York	Mideast

Cluster 2:

	Agglomerative Cluster	Gender	Age	Generation	Item Purchased	Category	Purchase Amount (USD)	Location	Region
5	2	Male	46	Gen X	Sneakers	Footwear	20	Wyoming	Rocky Mountain
9	2	Male	57	Gen X	Handbag	Accessories	31	Missouri	Plains
10	2	Male	53	Gen X	Shoes	Footwear	34	Arkansas	Southeast
24	2	Male	18	Gen Z	Jacket	Outerwear	22	Florida	Southeast
25	2	Male	18	Gen Z	Hoodie	Clothing	25	Texas	Southwest

Cluster 3:

	Agglomerative Cluster	Gender	Age	Generation	Item Purchased	Category	Purchase Amount (USD)	Location	Region
1	3	Male	19	Gen Z	Sweater	Clothing	64	Maine	New England
11	3	Male	30	Millenials	Shorts	Clothing	68	Hawaii	Far West
21	3	Male	31	Millenials	Pants	Clothing	62	North Carolina	Southeast
27	3	Male	56	Gen X	Shorts	Clothing	56	Kentucky	Southeast
32	3	Male	36	Millenials	Jacket	Outerwear	67	Kansas	Plains

- **Cluster 0 : Male, Gen Z/Baby Boomer, Clothing, 72-97**
- **Cluster 1 : Male, 45-65, Gen X/Baby Boomer, Clothing, 34-53**
- **Cluster 2 : Male, 46-53, Gen X/Gen Z, Footwear, 20-34**
- **Cluster 3 : Male, 30-36, Millenials/Gen Z, Clothing, 56-68**

Consumer Segmentation with GMM No-Scaling Model

Cluster 0:

	GMM Cluster	Gender	Age	Generation	Item Purchased	Category	Purchase Amount (USD)	Location	Region
25	0	Male	18	Gen Z	Hoodie	Clothing	25	Texas	Southwest
26	0	Male	38	Millenials	Jewelry	Accessories	20	Nevada	Far West
28	0	Male	54	Gen X	Handbag	Accessories	94	North Carolina	Southeast
31	0	Male	33	Millenials	Dress	Clothing	79	West Virginia	Southeast
37	0	Male	35	Millenials	Jeans	Clothing	45	Indiana	Great Lakes

Cluster 1:

	GMM Cluster	Gender	Age	Generation	Item Purchased	Category	Purchase Amount (USD)	Location	Region
1	1	Male	19	Gen Z	Sweater	Clothing	64	Maine	New England
3	1	Male	21	Gen Z	Sandals	Footwear	90	Rhode Island	New England
4	1	Male	45	Gen X	Blouse	Clothing	49	Oregon	Far West
10	1	Male	53	Gen X	Shoes	Footwear	34	Arkansas	Southeast
14	1	Male	64	Baby Boomer	Coat	Outerwear	53	New York	Mideast

Cluster 2:

	GMM Cluster	Gender	Age	Generation	Item Purchased	Category	Purchase Amount (USD)	Location	Region
5	2	Male	46	Gen X	Sneakers	Footwear	20	Wyoming	Rocky Mountain
6	2	Male	63	Baby Boomer	Shirt	Clothing	85	Montana	Rocky Mountain
7	2	Male	27	Gen Z	Shorts	Clothing	34	Louisiana	Southeast
9	2	Male	57	Gen X	Handbag	Accessories	31	Missouri	Plains
11	2	Male	30	Millenials	Shorts	Clothing	68	Hawaii	Far West

Cluster 3:

	GMM Cluster	Gender	Age	Generation	Item Purchased	Category	Purchase Amount (USD)	Location	Region
0	3	Male	55	Gen X	Blouse	Clothing	53	Kentucky	Southeast
2	3	Male	50	Gen X	Jeans	Clothing	73	Massachusetts	New England
8	3	Male	26	Gen Z	Coat	Outerwear	97	West Virginia	Southeast
15	3	Male	64	Baby Boomer	Skirt	Clothing	81	Rhode Island	New England
17	3	Male	53	Gen X	Dress	Clothing	38	Mississippi	Southeast

- **Cluster 0 : Male, 18-33, Millenials/Gen Z, Clothing, 25-45**
- **Cluster 1 : Male, 21-45, Gen Z/Gen X, Clothing, 64-90**
- **Cluster 2 : Male, 57-63, Gen X/Baby Boomer, Clothing, 20-31**
- **Cluster 3 : Male, 50-55, Gen X/Baby Boomer, Clothing, 53-73**

Consumer Segmentation Model 5

BCC
Basic Computing Community

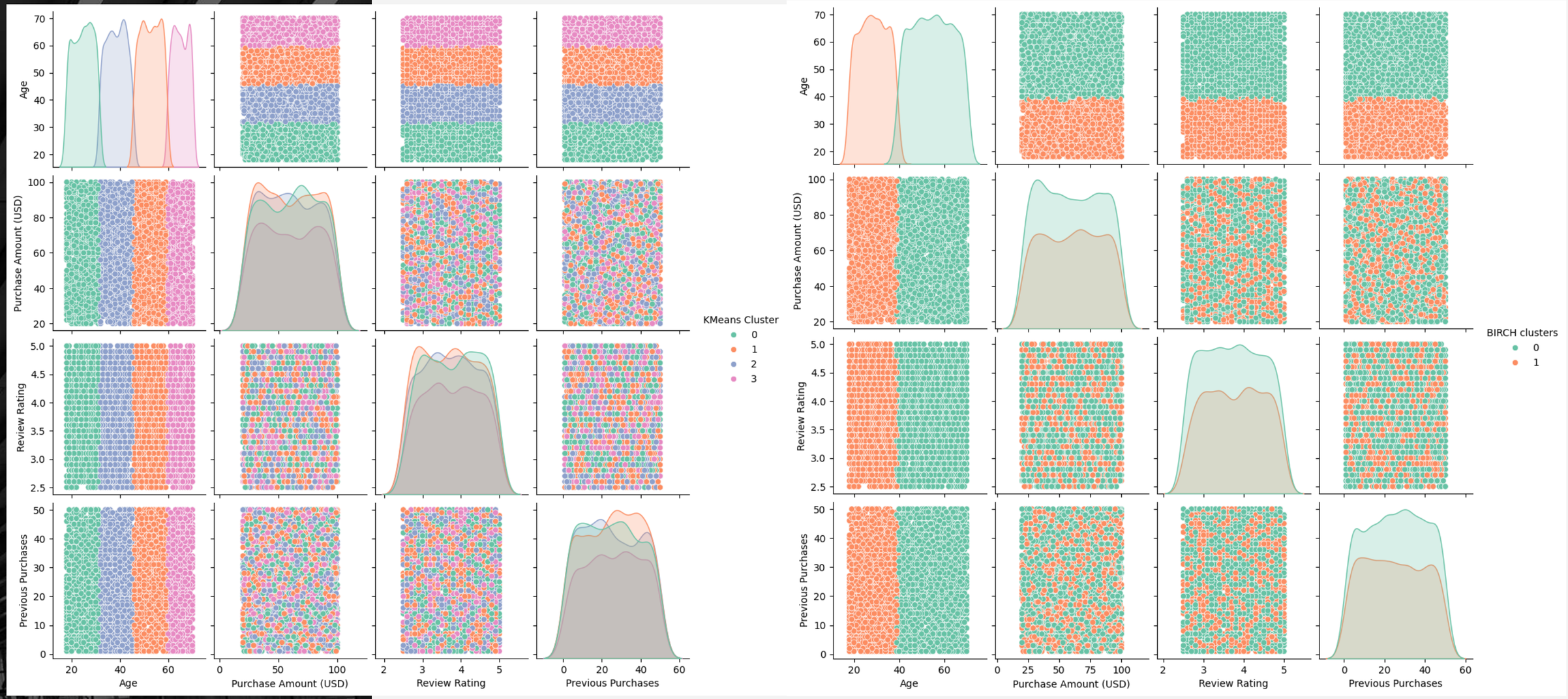
	Customer ID	Age	Gender	Item Purchased	Category
0	1	55	Male	Blouse	Clothing
1	2	19	Male	Sweater	Clothing
2	3	50	Male	Jeans	Clothing
3	4	21	Male	Sandals	Footwear
4	5	45	Male	Blouse	Clothing
...	...	...	...	...	...
3895	3896	40	Female	Hoodie	Clothing
3896	3897	52	Female	Backpack	Accessories
3897	3898	46	Female	Belt	Accessories
3898	3899	44	Female	Shoes	Footwear
3899	3900	52	Female	Handbag	Accessories

Overview

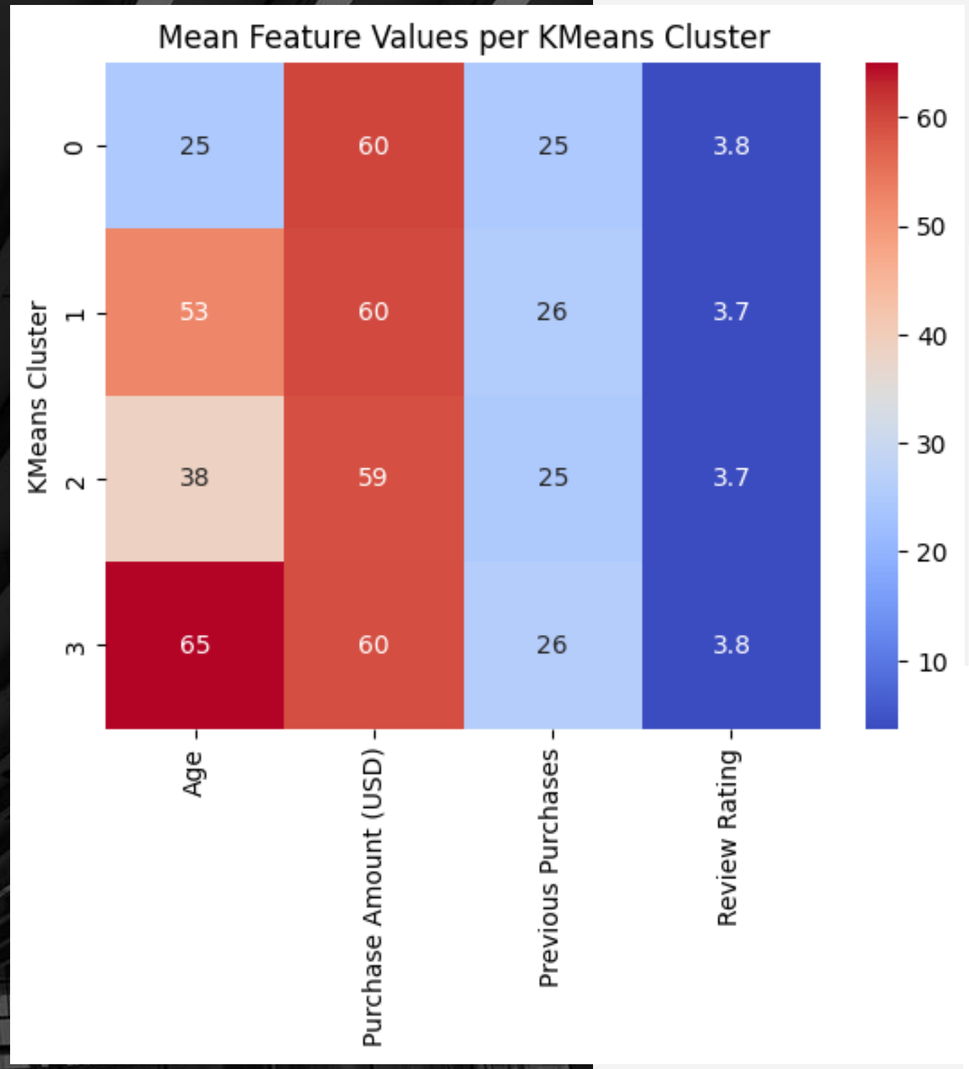
	Payment Method	Frequency of Purchases	Region	Generation	KMeans Cluster	BIRCH clusters
0	Venmo	Fortnightly	Southeast	Gen_X	1	0
1	Cash	Fortnightly	New England	Gen_Z	0	1
2	Credit Card	Weekly	New England	Gen_X	1	0
3	PayPal	Weekly	New England	Gen_Z	0	1
4	PayPal	Annually	Far West	Gen_X	2	0
...	...	...	...	...	...	...
3895	Venmo	Weekly	Southeast	Millenials	2	0
3896	Bank Transfer	Bi-Weekly	Plains	Gen_X	1	0
3897	Venmo	Quarterly	Mideast	Gen_X	1	0
3898	Venmo	Weekly	Plains	Millenials	2	0
3899	Venmo	Quarterly	Far West	Gen_X	1	0

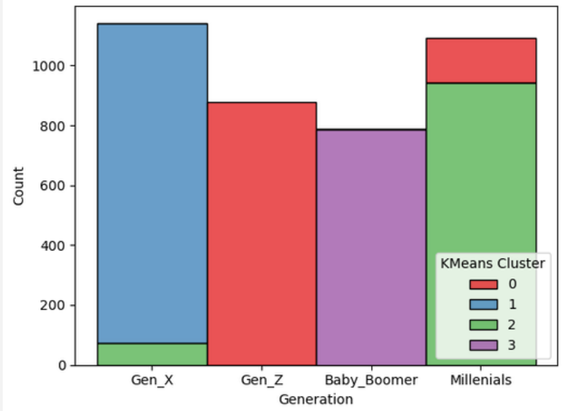
Consumer Segmentation

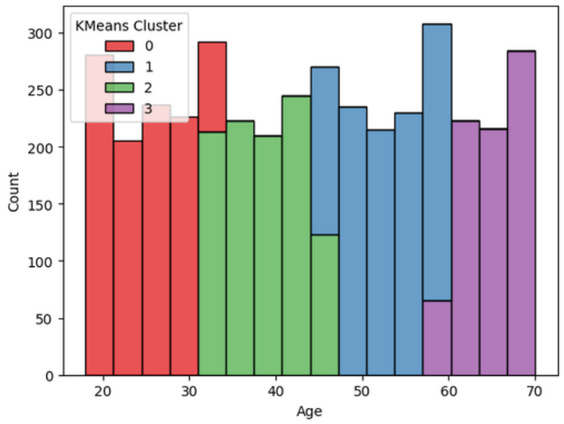
Model 5

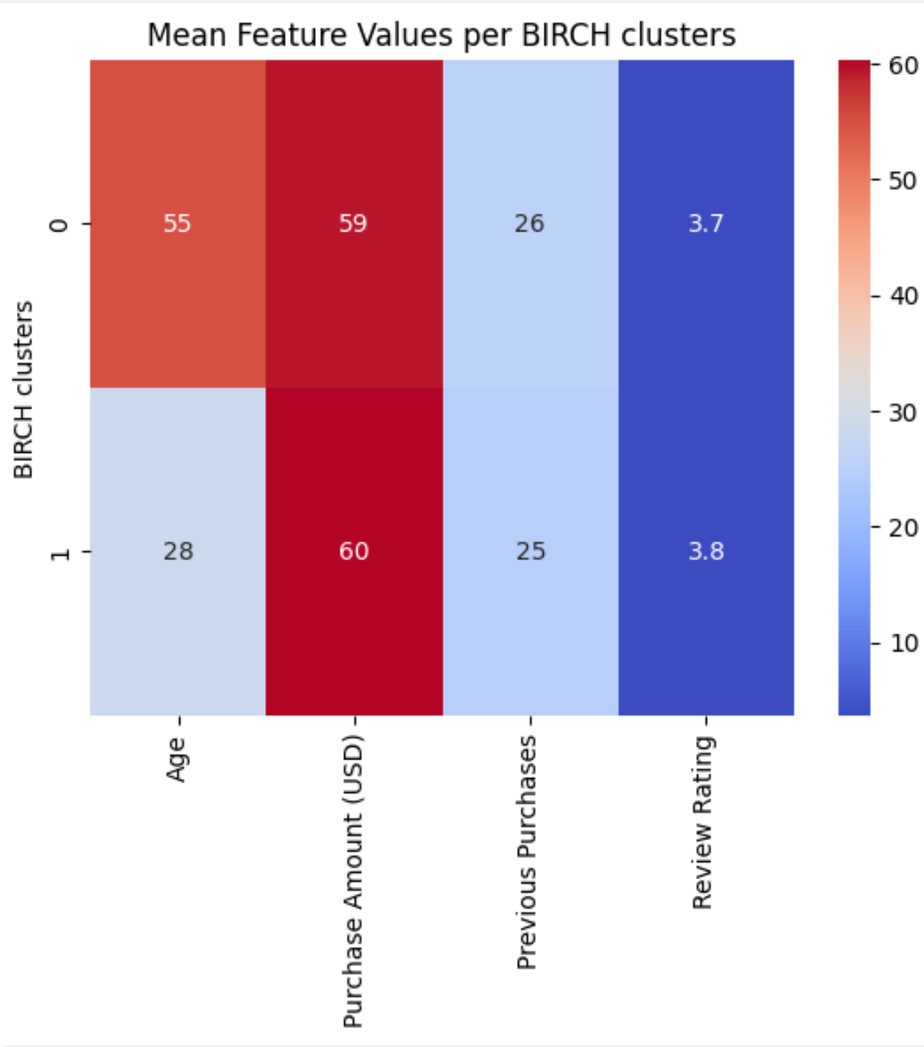


Consumer Segmentation Model 5



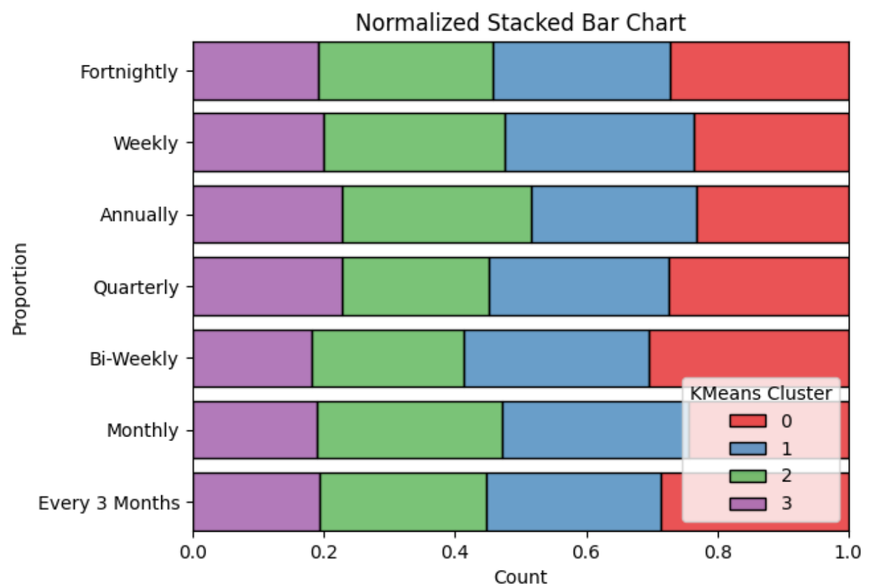


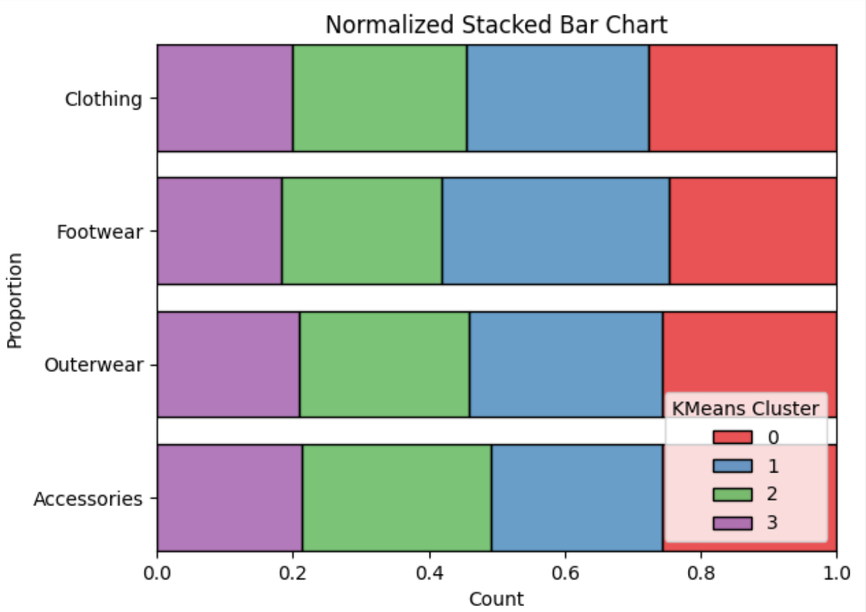


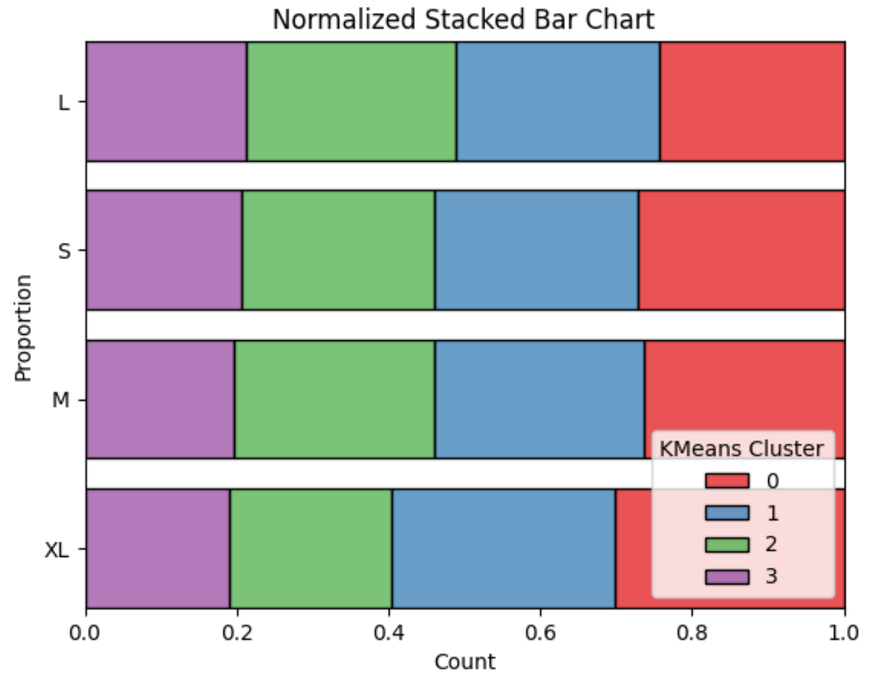


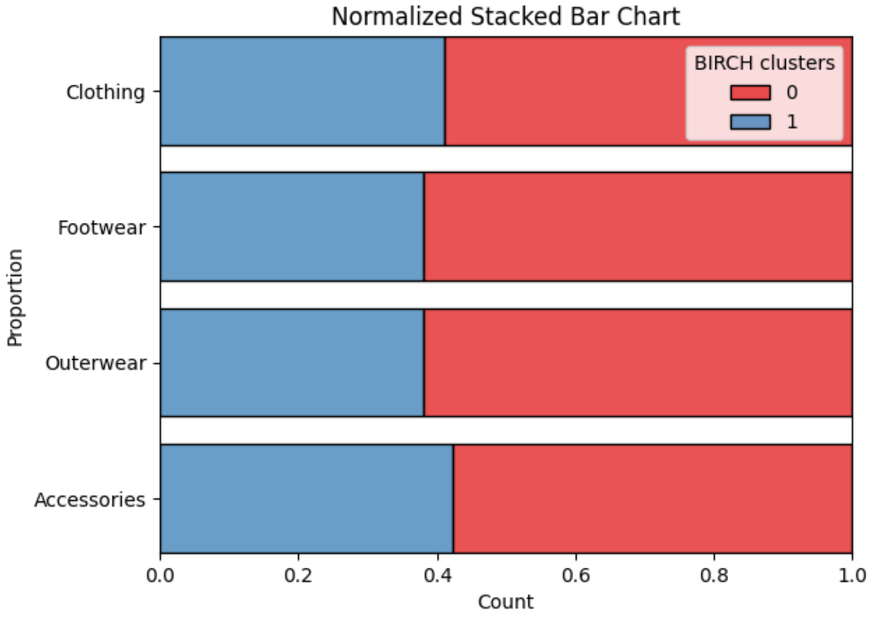
Basic Computing Community

Consumer Segmentation Model 5

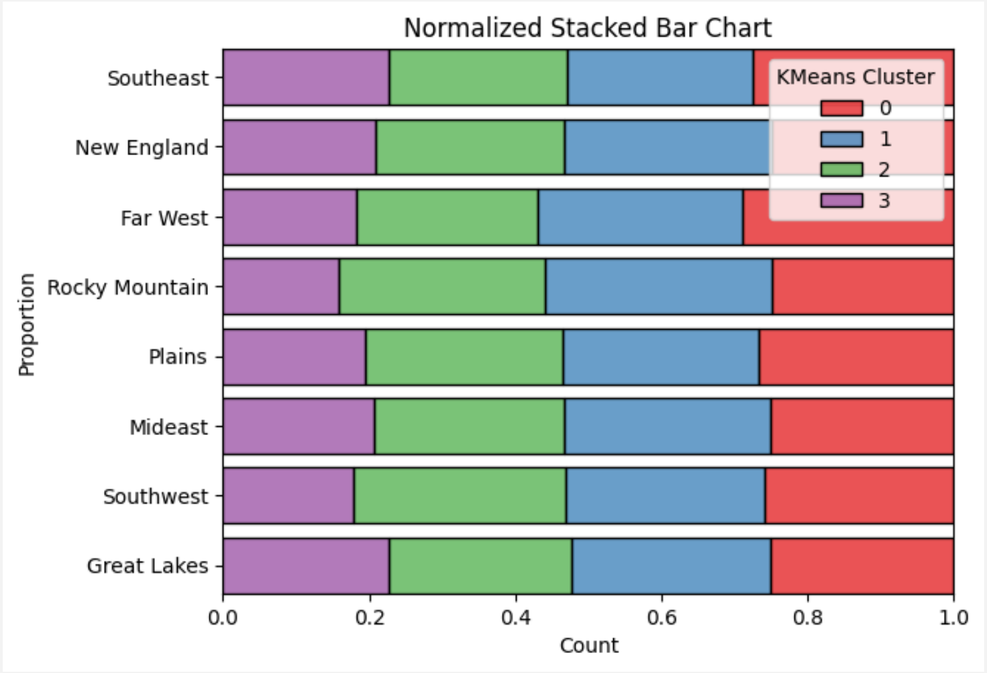




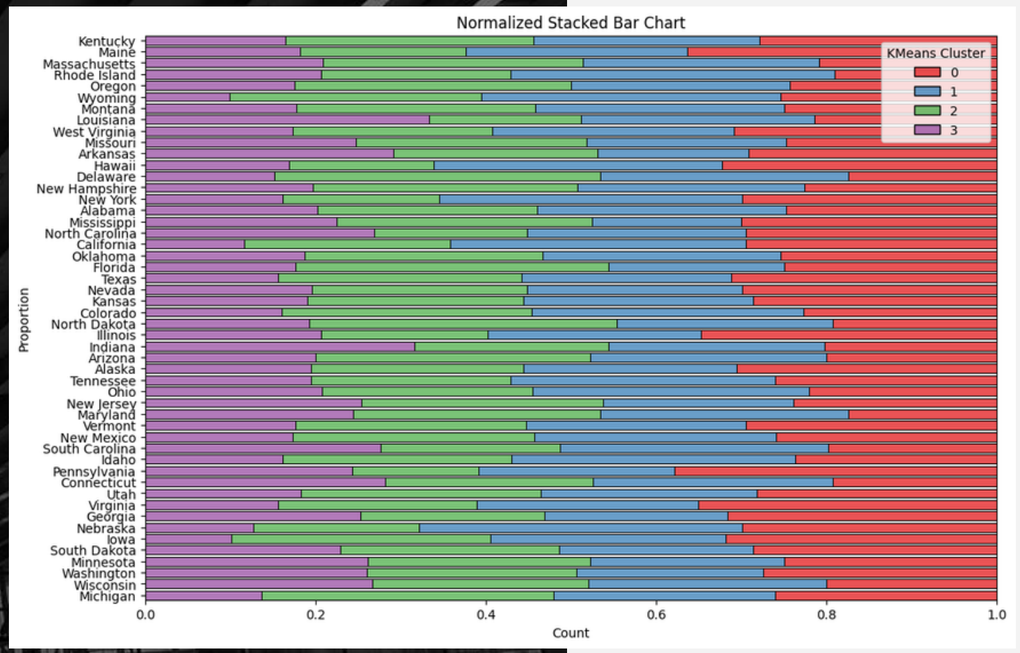


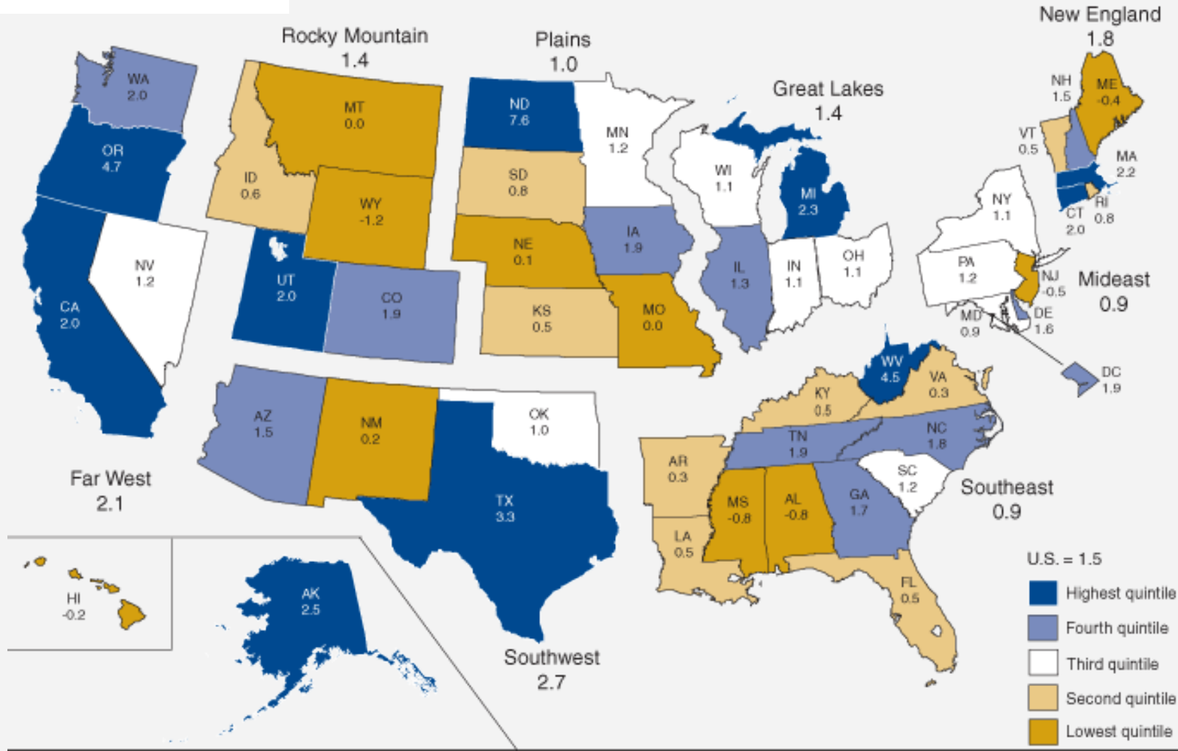


Consumer Segmentation Model 5

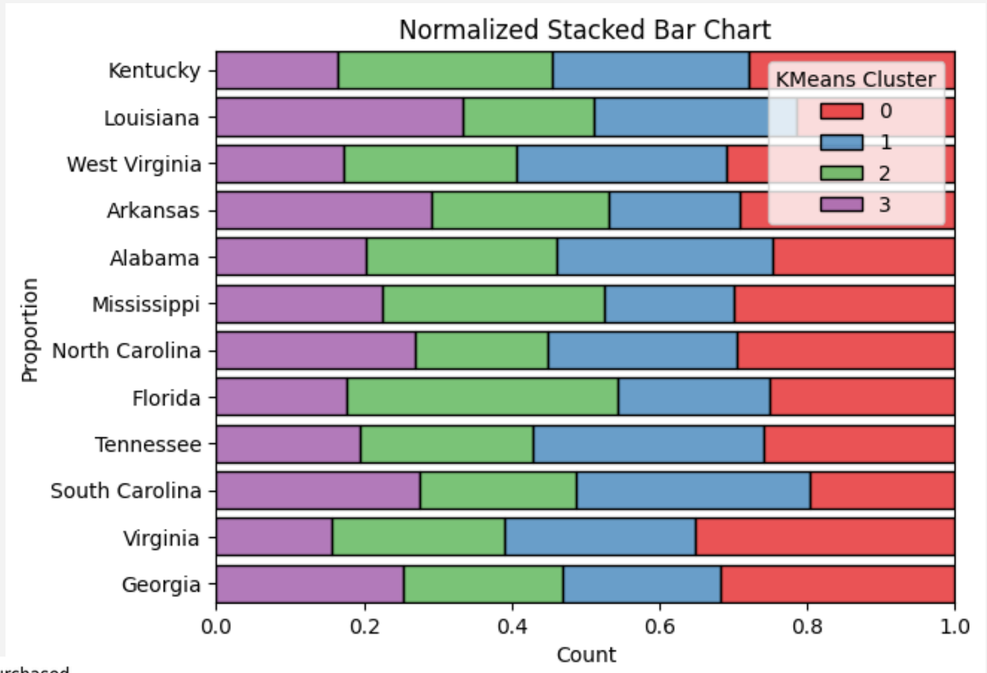


- **Cluster 0** : Far West, Plains and Southeast
- **Cluster 1** : New England, Rocky Mountain
- **Cluster 2** : Rocky Mountain, SouthWest
- **Cluster 3** : Southeast, Mideast and Great Lakes

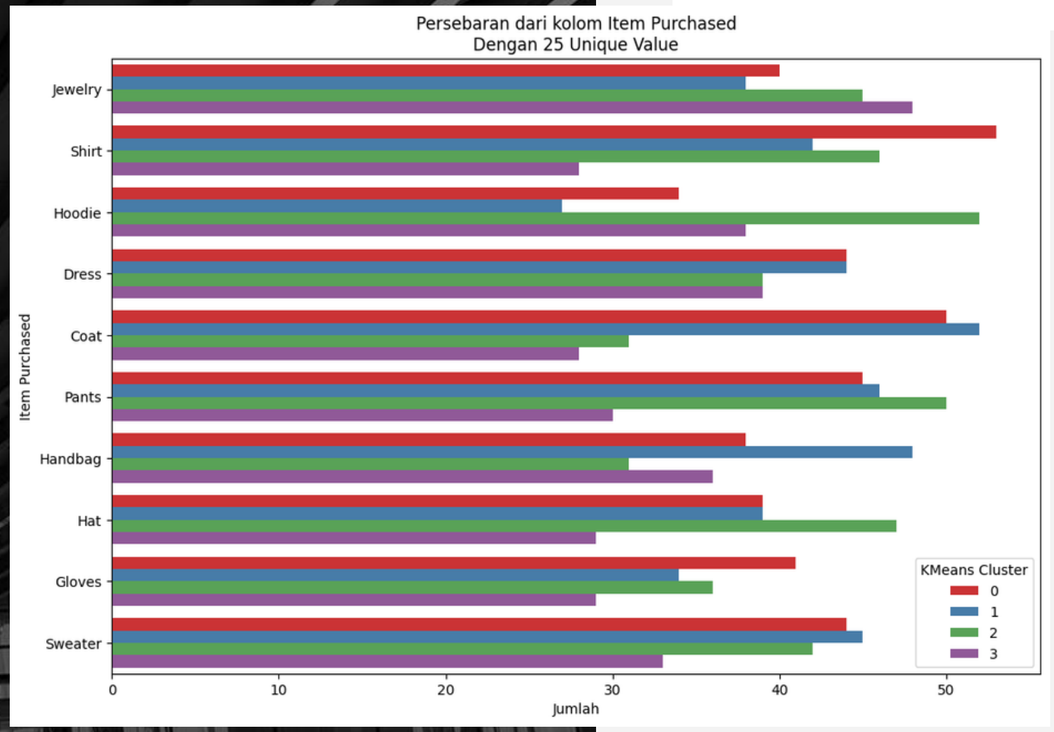


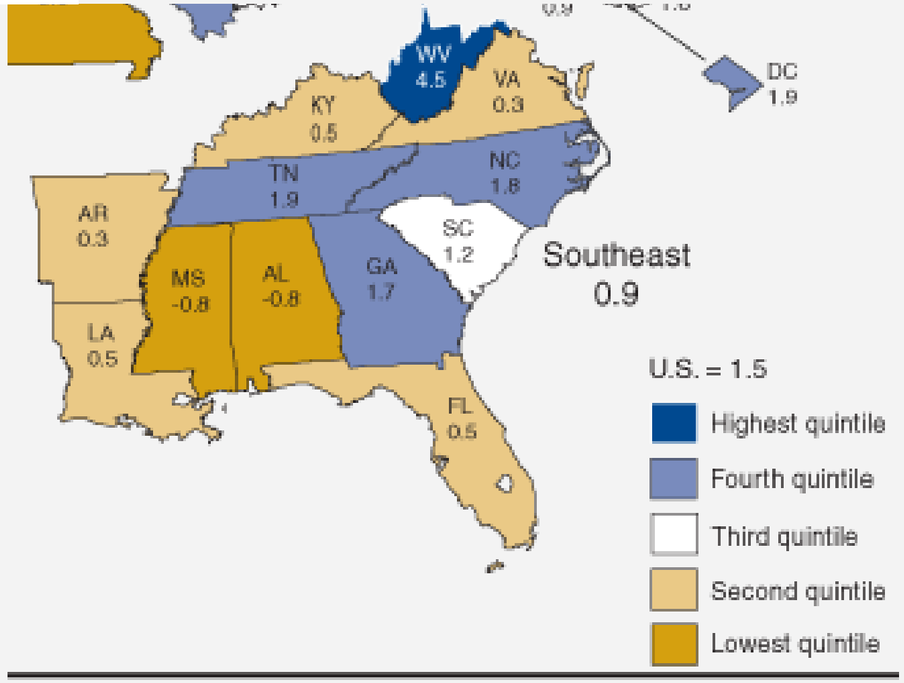


Consumer Segmentation on “Southeast”



- **Cluster 0** : Virginia, Georgia and West Virginia
- **Cluster 1** : Louisiana, Tennessee and South Carolina
- **Cluster 2** : Florida, Kentucky and Missisippi
- **Cluster 3** : ouisiana, Arkansas and North Carolina

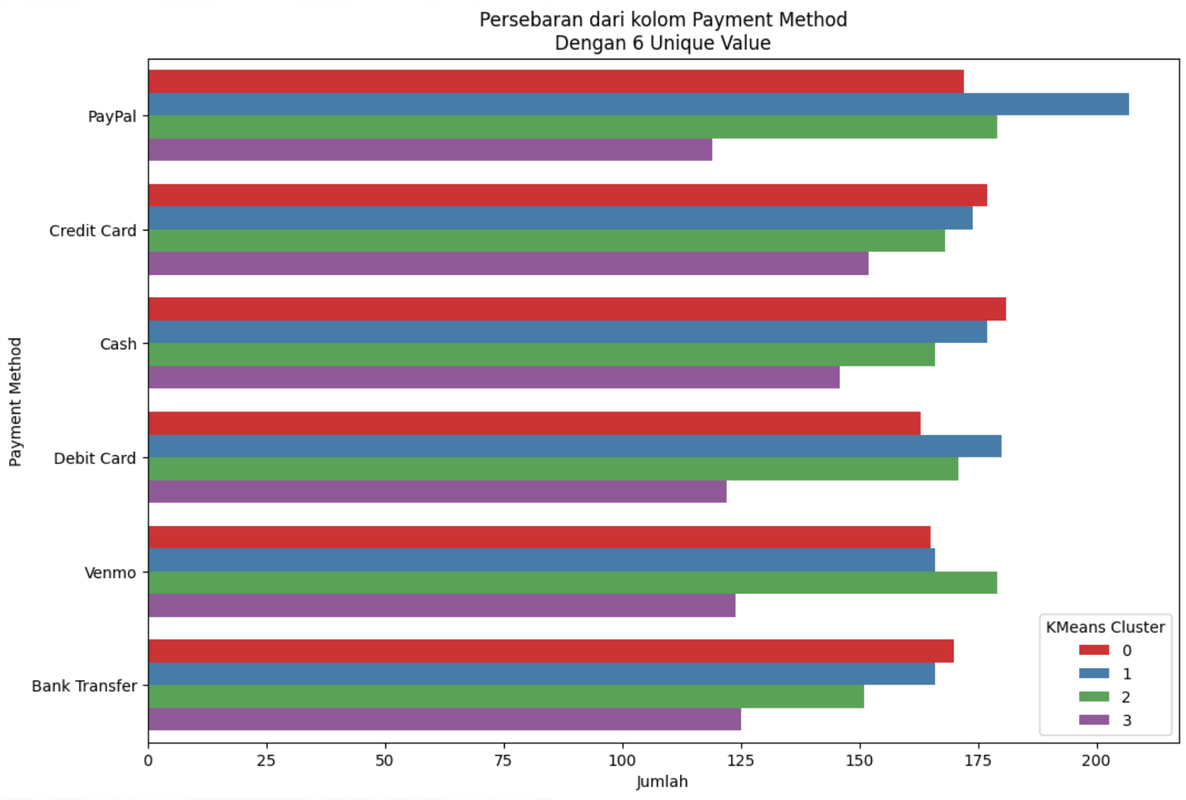


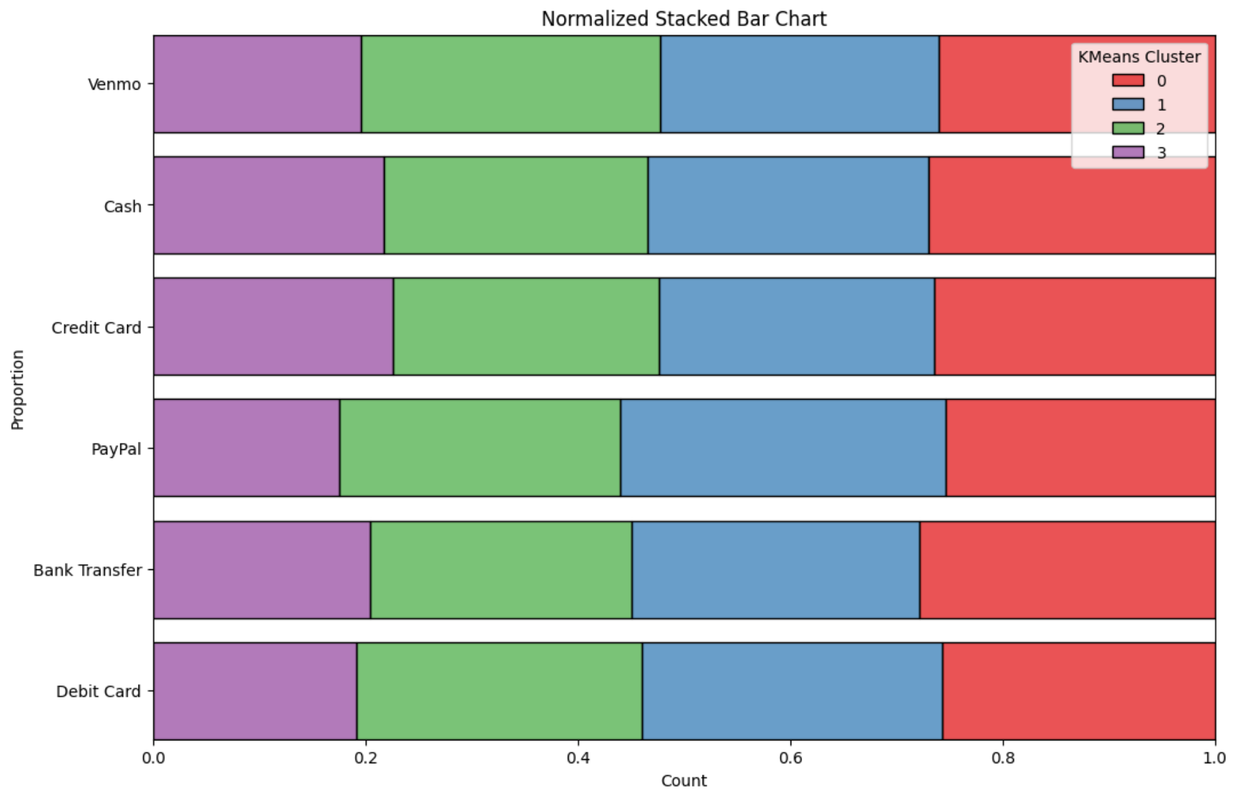


Consumer Segmentation Model 5

Payment Method

- **Cluster 0 :** Prefer Bank Transfer and Cash
- **Cluster 1 : Prefer PayPal and Debit Card**
- **Cluster 2** : Prefer Venmo and Paypal
- **Cluster 3** : Prefer Cash and Credit Card

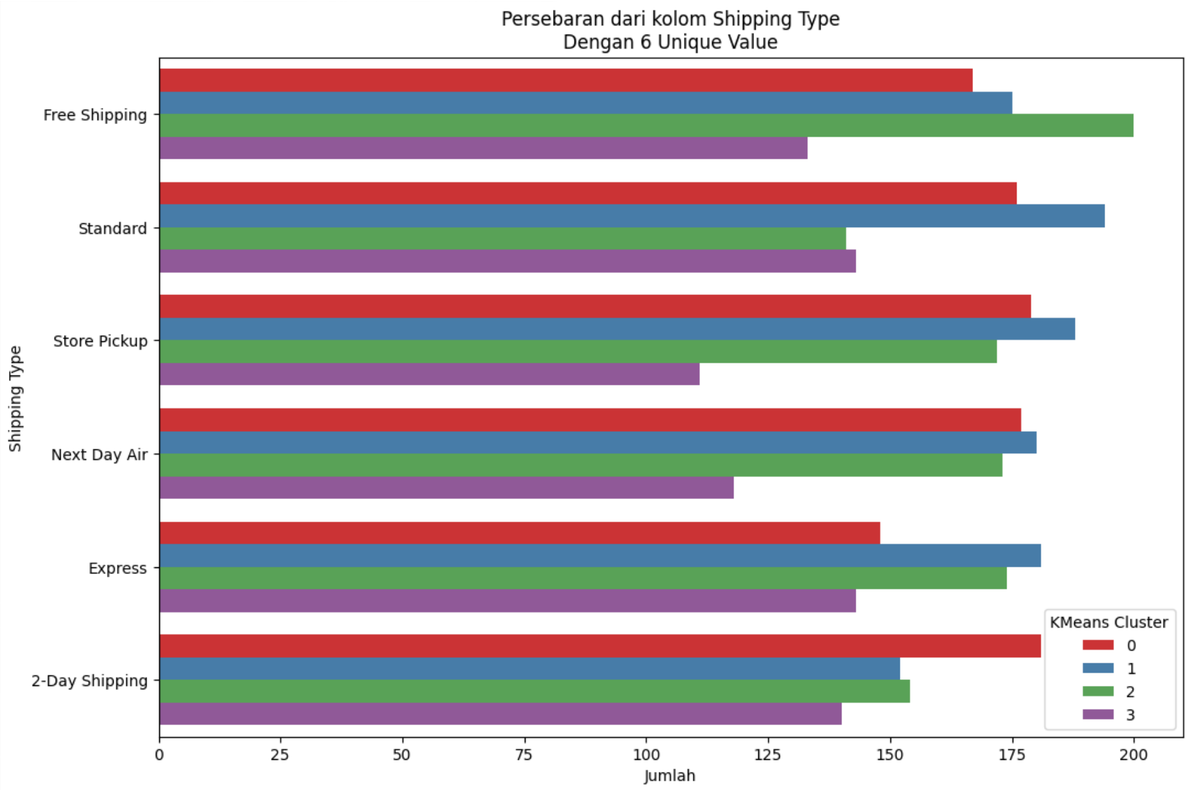


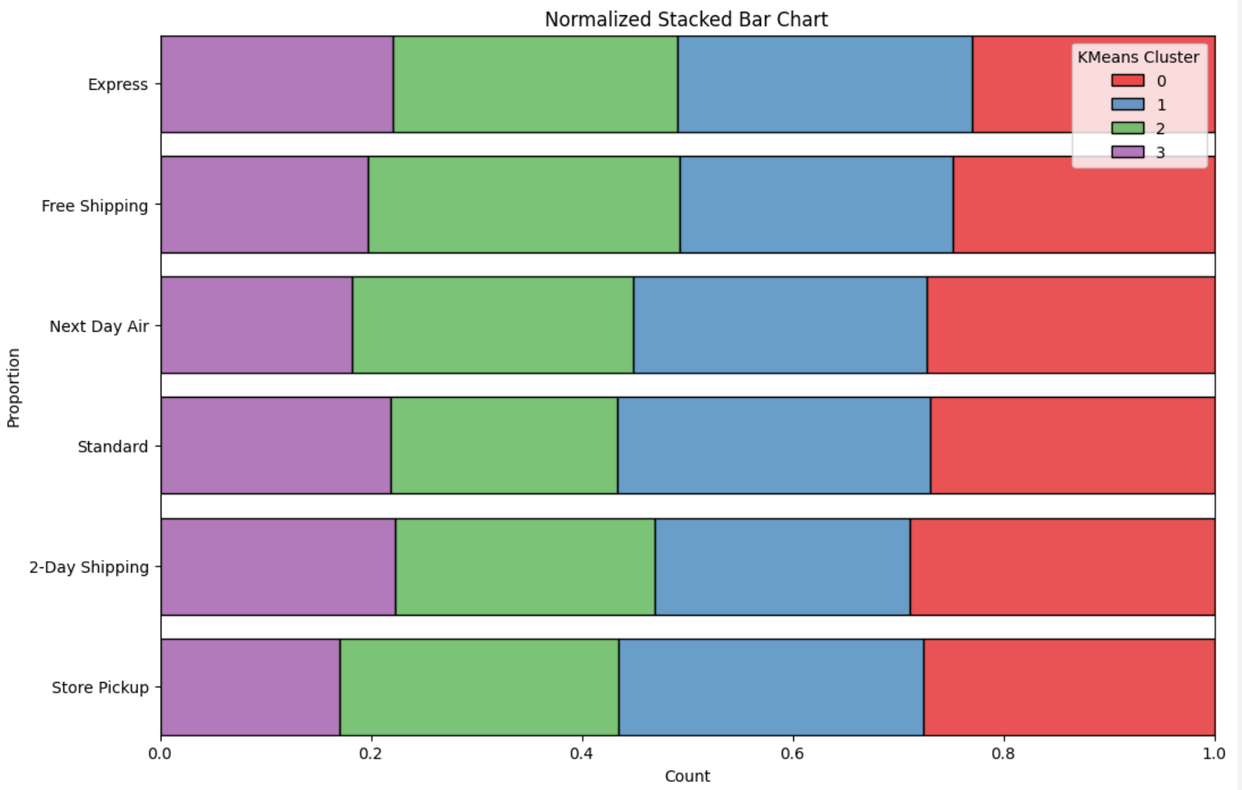


Consumer Segmentation Model 5

Shipping Type

- **Cluster 0 :** Prefer 2 Days Shipping and Store Pickup
- **Cluster 1 :** Prefer Standard and Store Pickup
- **Cluster 2** : Prefer Free Shipping and Express
- **Cluster 3** : Prefer 2 Days Shipping and Standard





Conclusion

- Untuk proses modelling tanpa menggunakan scaling terhadap kolom numerik, ketika menggunakan model K-Means dan Agglomerative sudah baik dalam proses segmentasinya sedangkan untuk Gaussian Mixture sedikit kurang perform.
- Untuk eksplorasi algoritma K-Mode and K-Prototype, algoritma tersebut perform lebih buruk jika preprocess tanpa scaling, sedangkan algoritma BIRCH perform lebih baik tanpa scaling.
- Untuk Fine tuning, jumlah cluster terbaik didominasi cluster dengan n yang kecil seperti 2 dan 3. Tetapi model dengan algoritma Gaussian Mixture Model kurang baik saat finetuning berbasis BIC.

THANK YOU

Feel free if you have any questions

BCC
Basic Computing Community

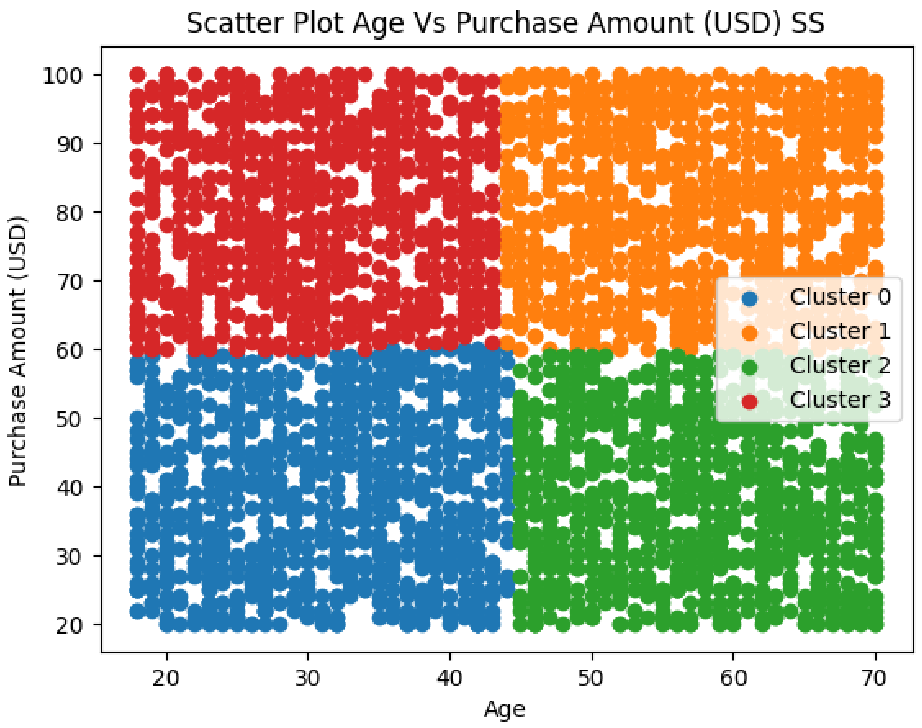
APPENDIX

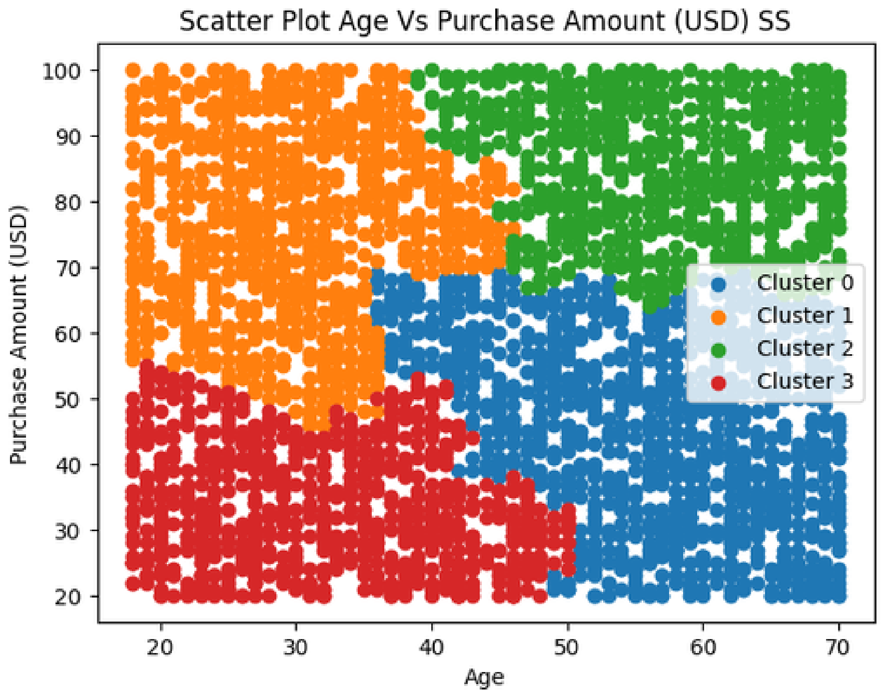
MODEL 1-3

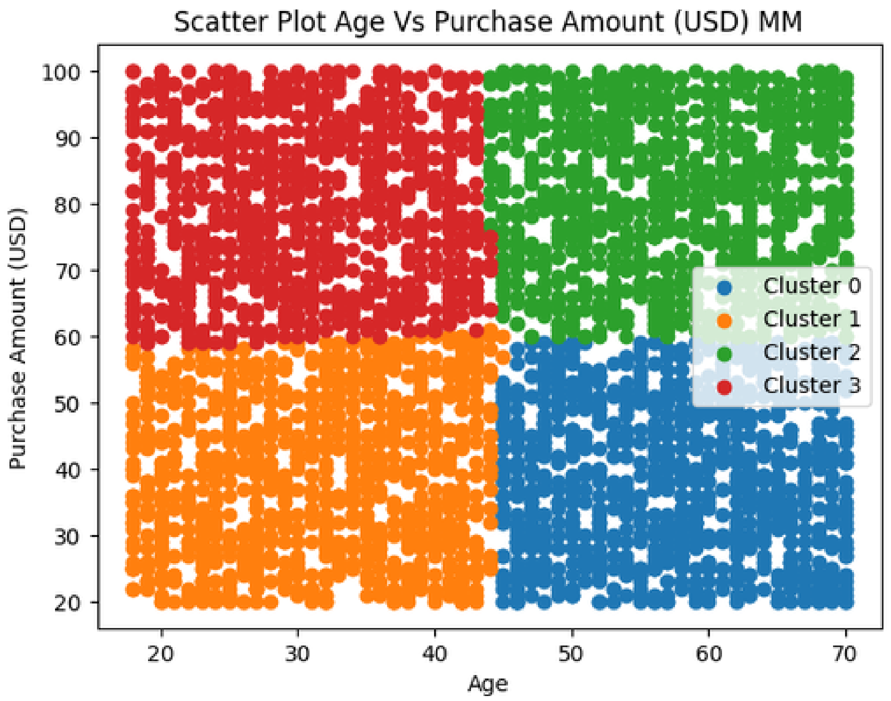
- X_Column = Age
- y_Column = Purchase Amount (USD)
- Different Preprocessing, same results

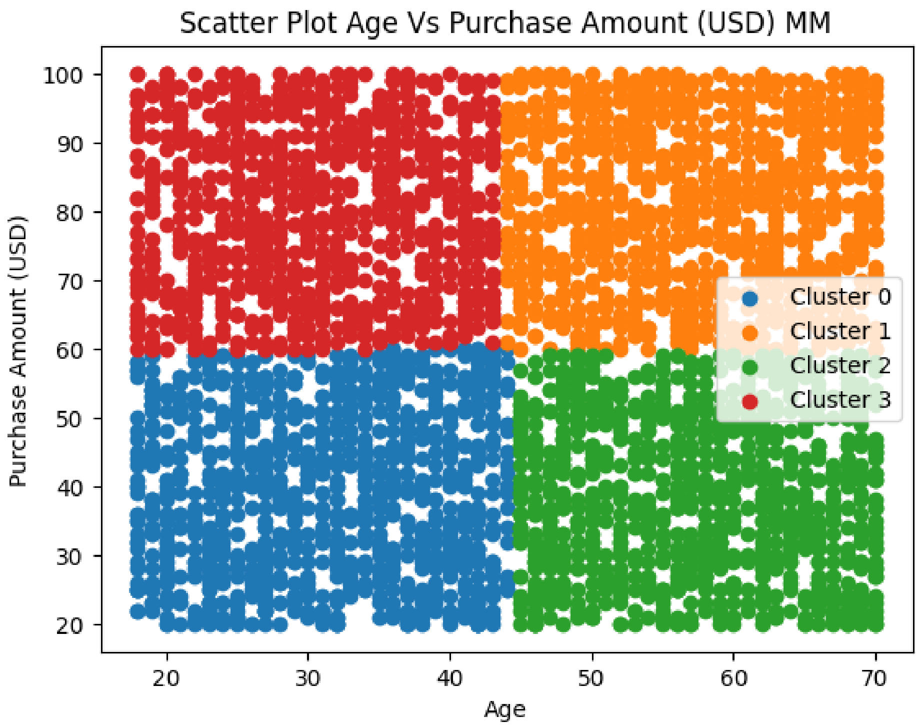
Just X_Column and y_column

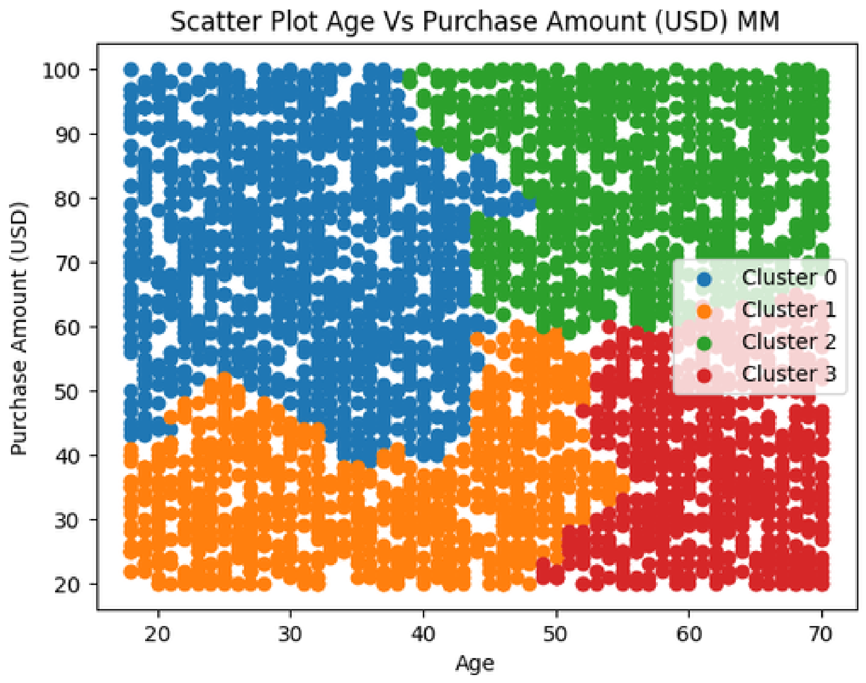
	Model	Silhouette Score
0	K-Means Age Purchase Amount (USD) SS	0.413762
1	K-Means Age Purchase Amount (USD) MM	0.413725
4	Gaussian Age Purchase Amount (USD) SS	0.413624
5	Gaussian Age Purchase Amount (USD) MM	0.413615
2	AHC Age Purchase Amount (USD) SS	0.372396
3	AHC Age Purchase Amount (USD) MM	0.350095

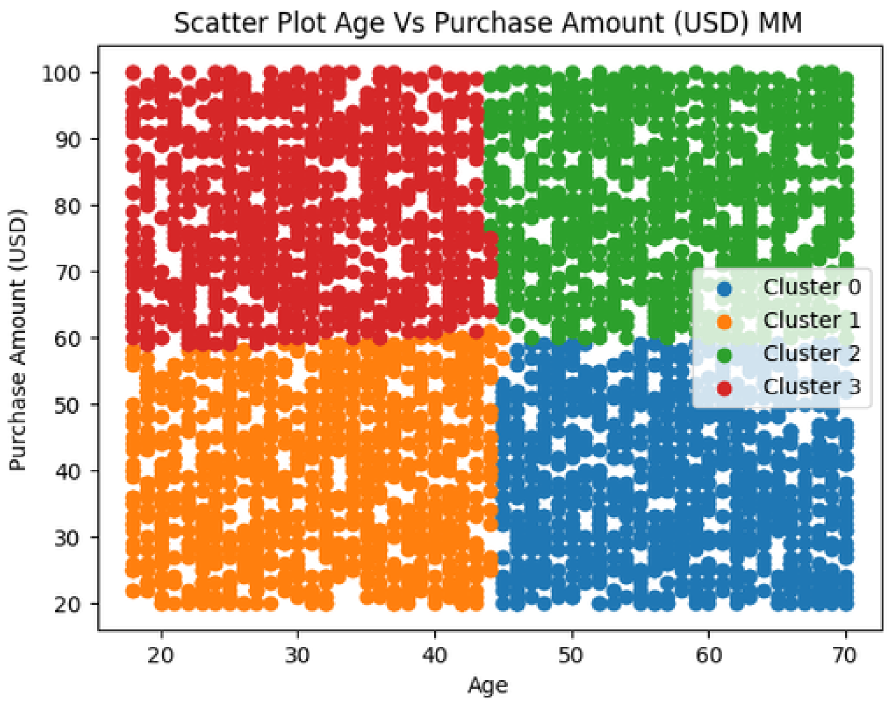












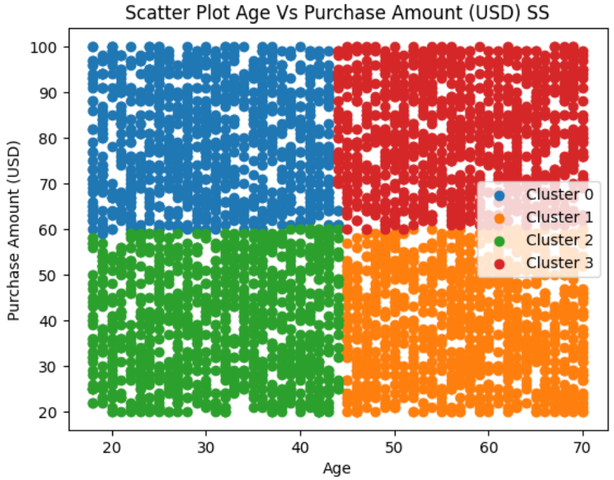
MODEL 4

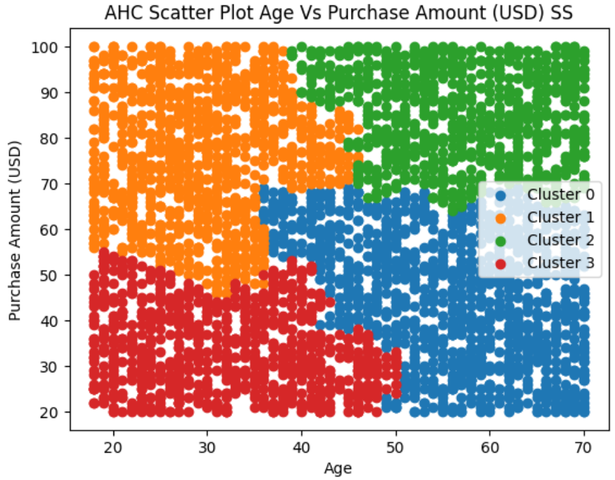
- Minimal Preprocessing Data PP 4
- Without Feature Engineering
- LabelEncoder all Column

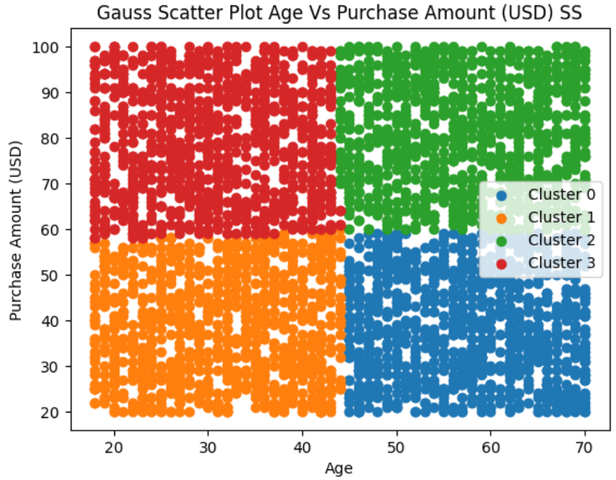
- X_Column = Age
- y_Column = Purchase Amount (USD)

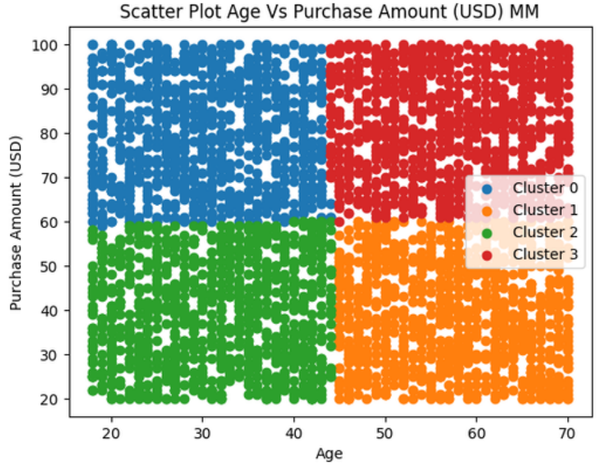
	Model	Silhouette Score
0	K-Means Age Purchase Amount (USD) SS	0.413719
1	K-Means Age Purchase Amount (USD) MM	0.413657
4	Gaussian Age Purchase Amount (USD) SS	0.413624
5	Gaussian Age Purchase Amount (USD) MM	0.413615
2	AHC Age Purchase Amount (USD) SS	0.372396
3	AHC Age Purchase Amount (USD) MM	0.350095

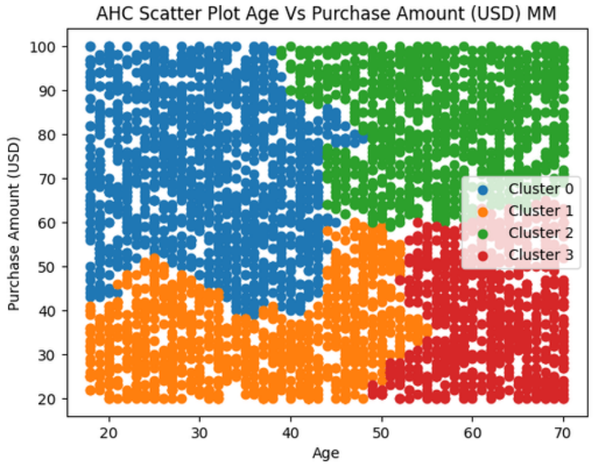
Just X_Column and y_column

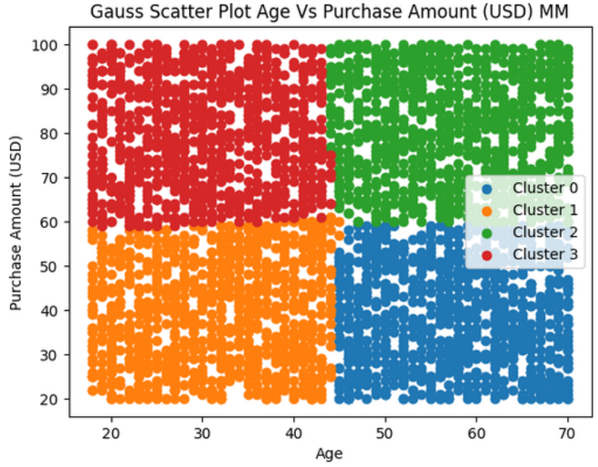












All Column

	Model	Silhouette Score
1	K-Means Age Purchase Amount (USD) MM	0.159789
3	AHC Age Purchase Amount (USD) MM	0.159789
5	Gaussian Age Purchase Amount (USD) MM	0.159789
0	K-Means Age Purchase Amount (USD) SS	0.068506
2	AHC Age Purchase Amount (USD) SS	0.065037
4	Gaussian Age Purchase Amount (USD) SS	0.064747

Selected Feature

```
X = min_df[
    "Gender",
    "Category",
    "Size",
    "Promo Code Used",
    "Discount Applied",
    "Frequency of Purchases",
    "Age",
    "Review Rating"
```

	Model	Silhouette Score
5	Gaussian Age Review Rating MM	0.262356
3	AHC Age Review Rating MM	0.253236
1	K-Means Age Review Rating MM	0.248276
0	K-Means Age Review Rating SS	0.175852
4	Gaussian Age Review Rating SS	0.172418
2	AHC Age Review Rating SS	0.170444

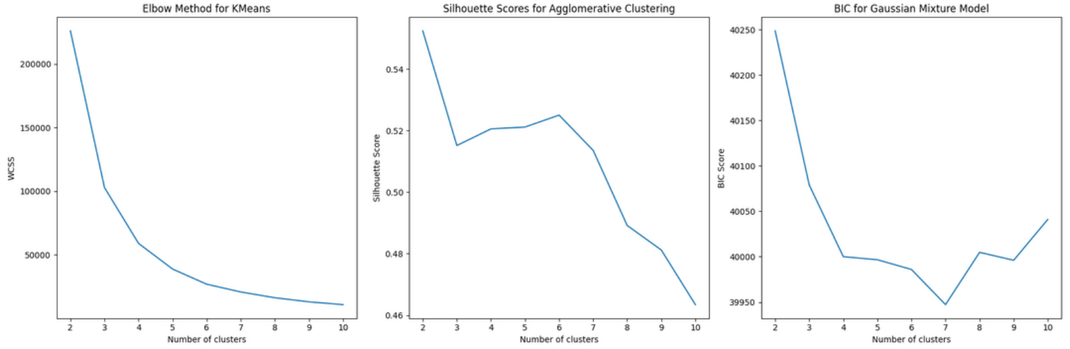
MODEL 4

Fine Tuning

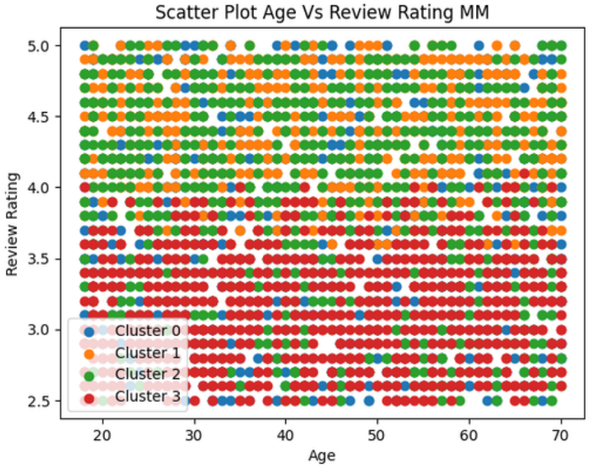
	Model	Silhouette Score
3	AHC Age Review Rating MM	0.433141
5	Gaussian Age Review Rating MM	0.433141
9	Gaussian Age Review Rating MM (3 Cluster)	0.387032
2	AHC Age Review Rating SS	0.282650
4	Gaussian Age Review Rating SS	0.282650
7	AHC Age Review Rating MM (6 Cluster)	0.251563
1	K-Means Age Review Rating MM	0.248276
8	Gaussian Age Review Rating SS (3 Cluster)	0.187877
0	K-Means Age Review Rating SS	0.175852
6	AHC Age Review Rating SS (6 Cluster)	0.124075

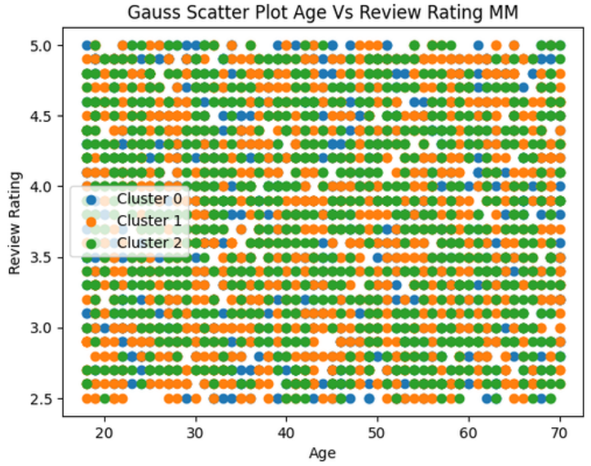
Age Vs Purchase Amount (USD)

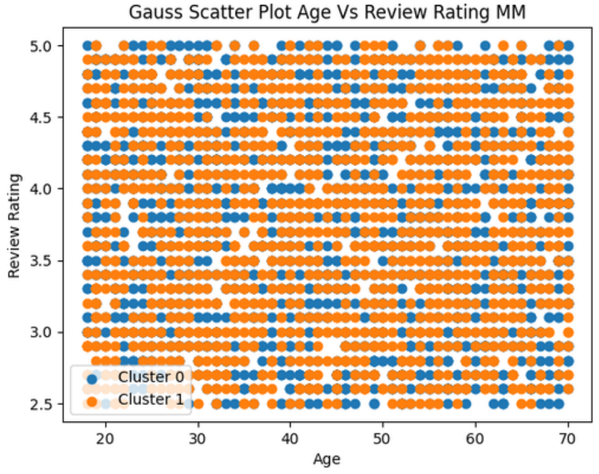
- Elbow Method, n = 4
- AHC best, n = 2
- Gaussian Mixture Model, n = 3-5, n = 3

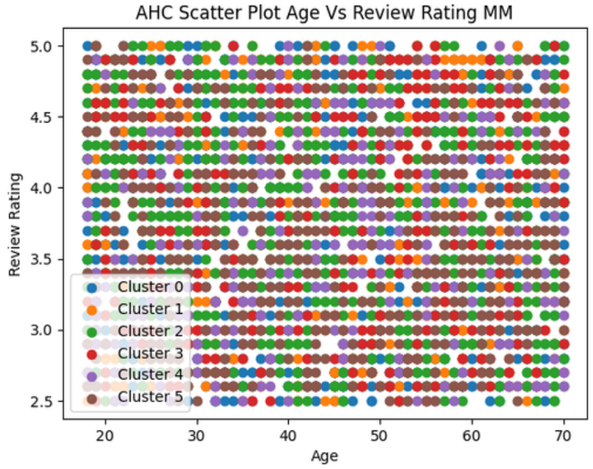


- Age Vs Review Rating
- Elbow Method, n = 4
- AHC best, n = 2
- Gaussian Mixture Model, n = 2

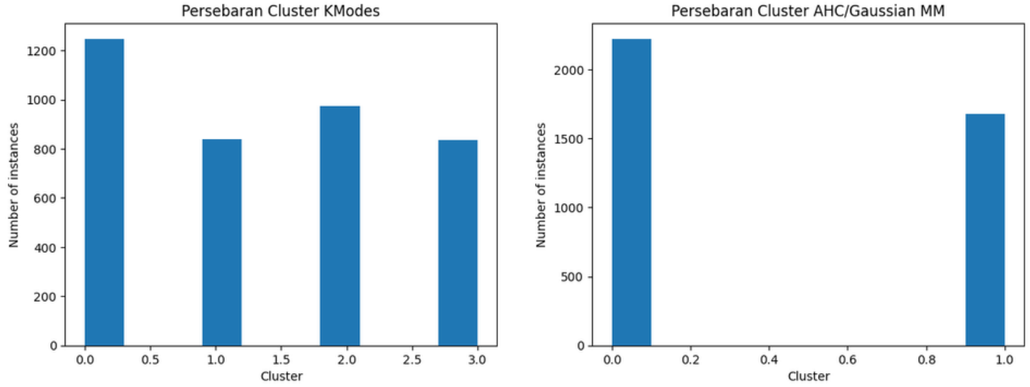








Histogram Persebaran Cluster



MODEL 5

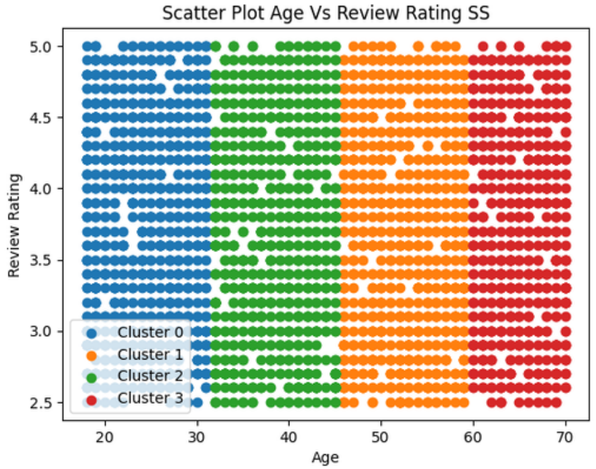
- Fully Preprocess
- Without Scaling
- Feature Engineer (Generation + Region)
- One Hot Encoding + LabelEncoder

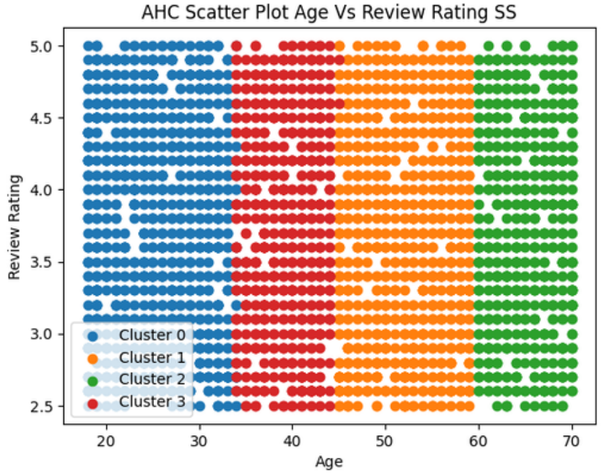
- X_Column = Age
- y_Column = Review Rating

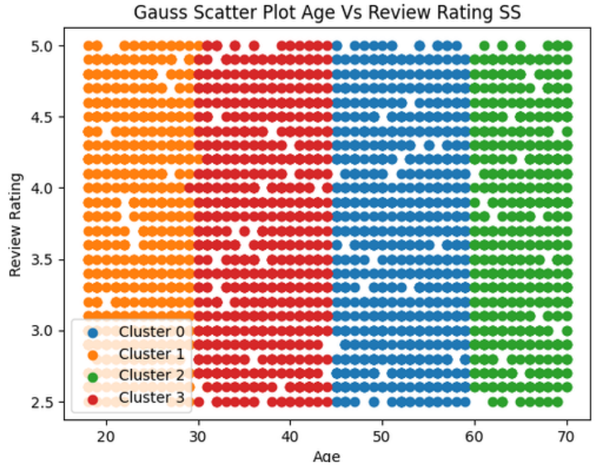
Selected Feature

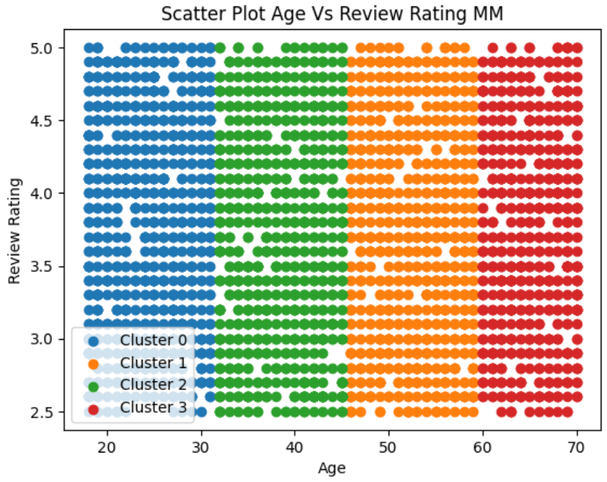
```
"Gender",
"Gen_X",
"Southeast",
"Category",
"Size",
"Age",
"Review Rating"
```

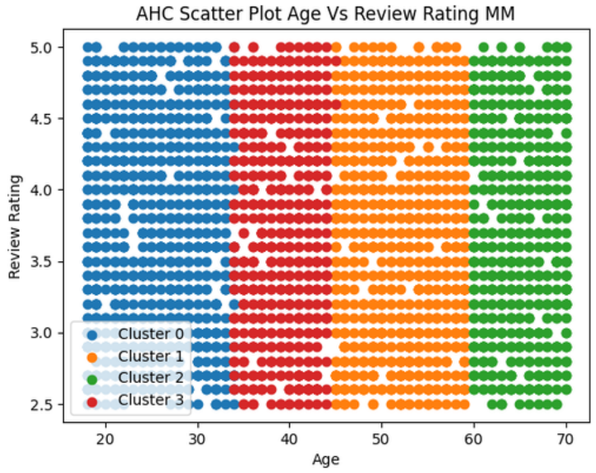
	Model	Silhouette Score
7	BIRCH Age Review Rating	0.521878
0	K-Means Age Review Rating SS	0.516945
1	K-Means Age Review Rating MM	0.516945
4	Gaussian Age Review Rating SS	0.510841
5	Gaussian Age Review Rating MM	0.510841
2	AHC Age Review Rating SS	0.498850
3	AHC Age Review Rating MM	0.498850
9	BIRCH Age Review Rating (4)	0.497267
8	BIRCH Age Review Rating (All Column)	0.208245
6	KModes Age Review Rating SS	0.010154

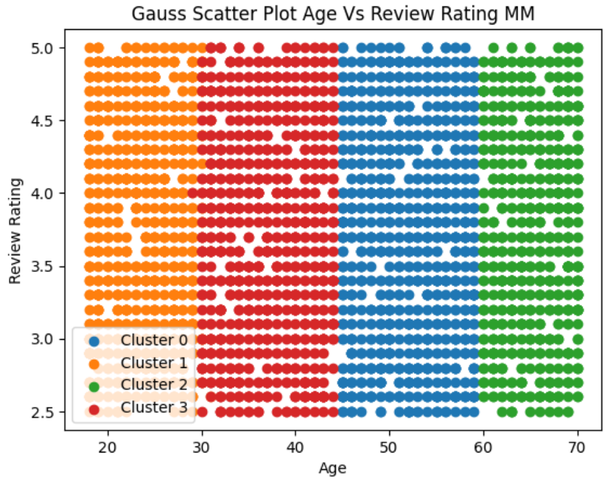










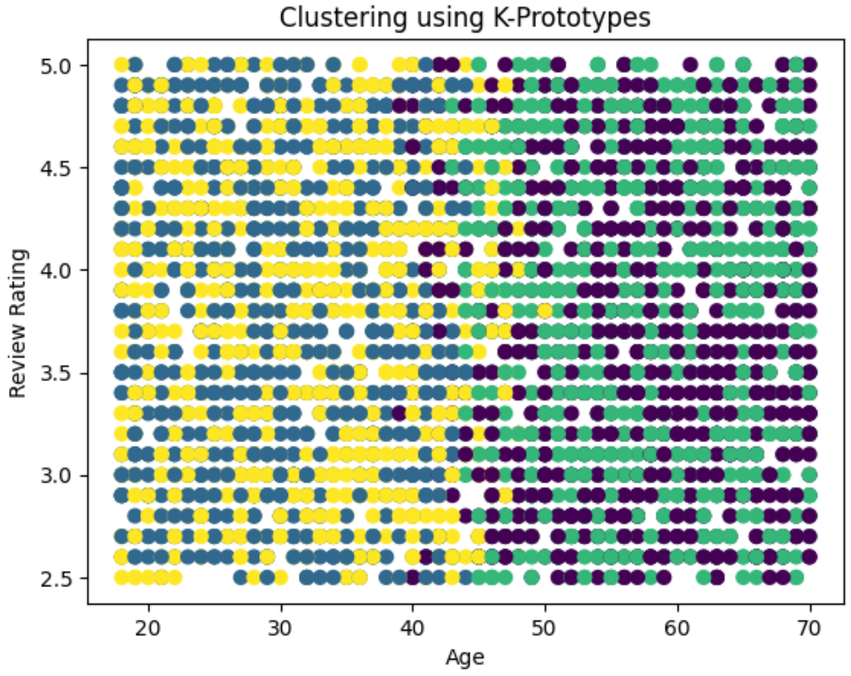


MODEL 5

Selected Feature

- Fully Preprocess
- Without Scaling
- Feature Engineer (Generation + Region)
- One Hot Encoding + LabelEncoder

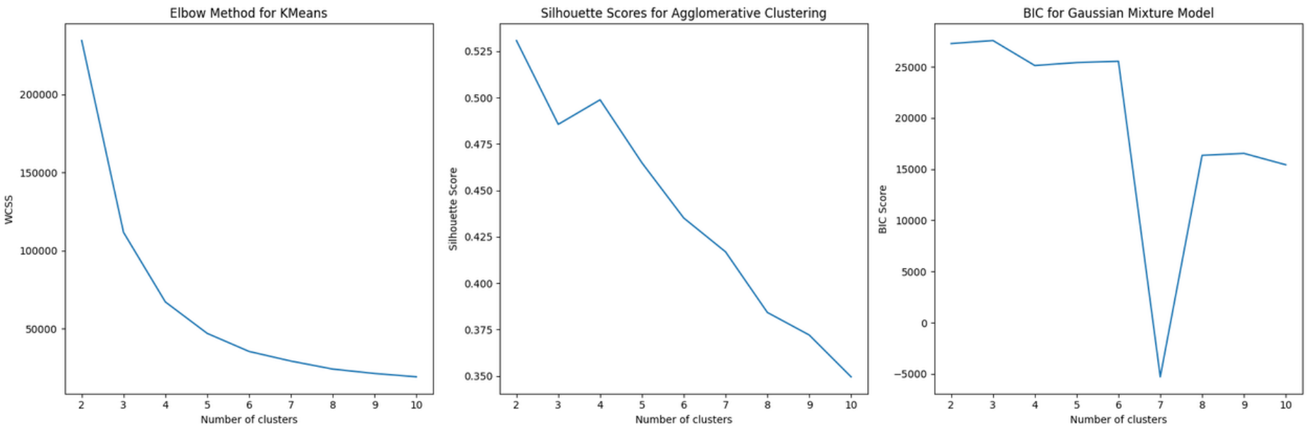
- X_Column = Age
- y_Column = Review Rating



MODEL 5

Fine Tuning

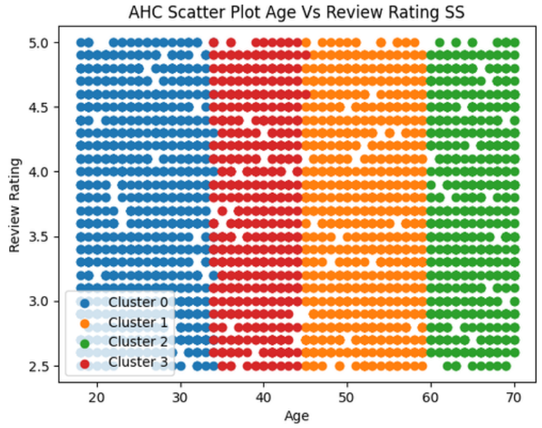
	Model	Silhouette Score
8	BIRCH Age Review Rating (4)	0.589970
0	K-Means Age Review Rating SS	0.552705
1	K-Means Age Review Rating MM	0.552705
2	AHC Age Review Rating SS	0.498850
3	AHC Age Review Rating MM	0.498850
7	K-Prototypes Age Review Rating	0.246099
9	BIRCH Age Review Rating (All Column)	0.208245
4	Gaussian Age Review Rating SS	0.186911
5	Gaussian Age Review Rating MM	0.186911
6	KModes Age Review Rating SS	0.010154

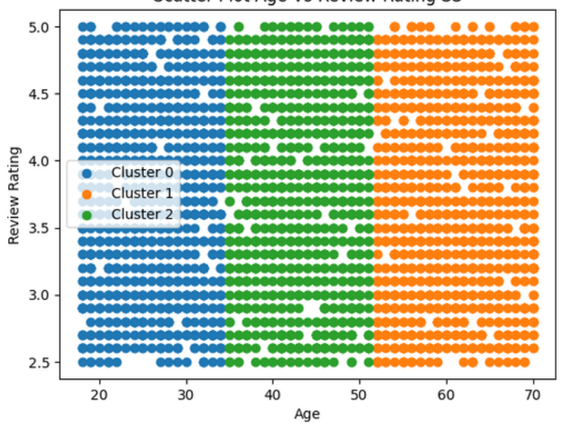


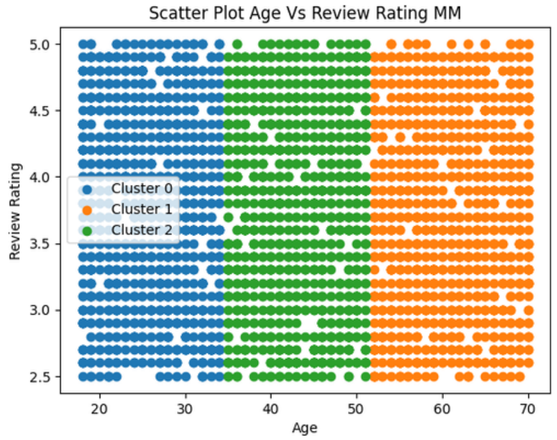
- K-Modes, n = 4
- BIRCH, n = 2

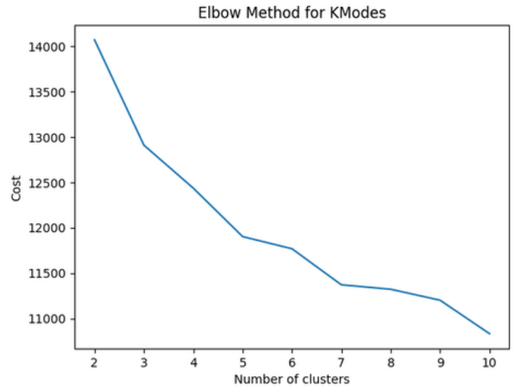
Age Vs Review Rating

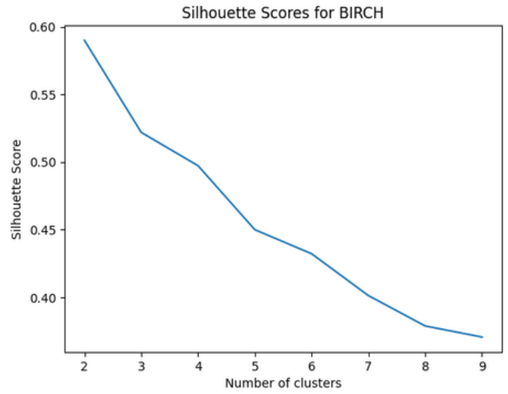
- Elbow Method, n = 3
- AHC best, n = 4
- Gaussian Mixture Model, n = 3

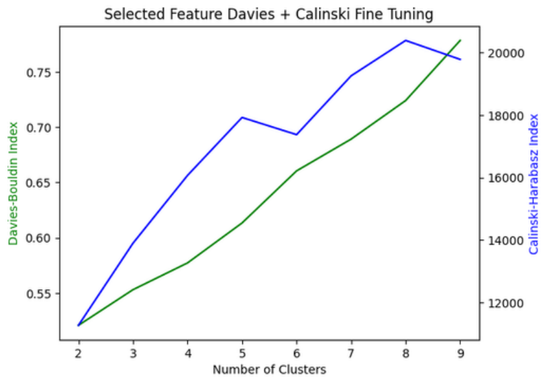


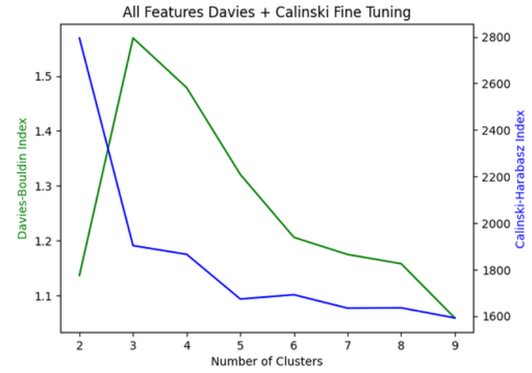








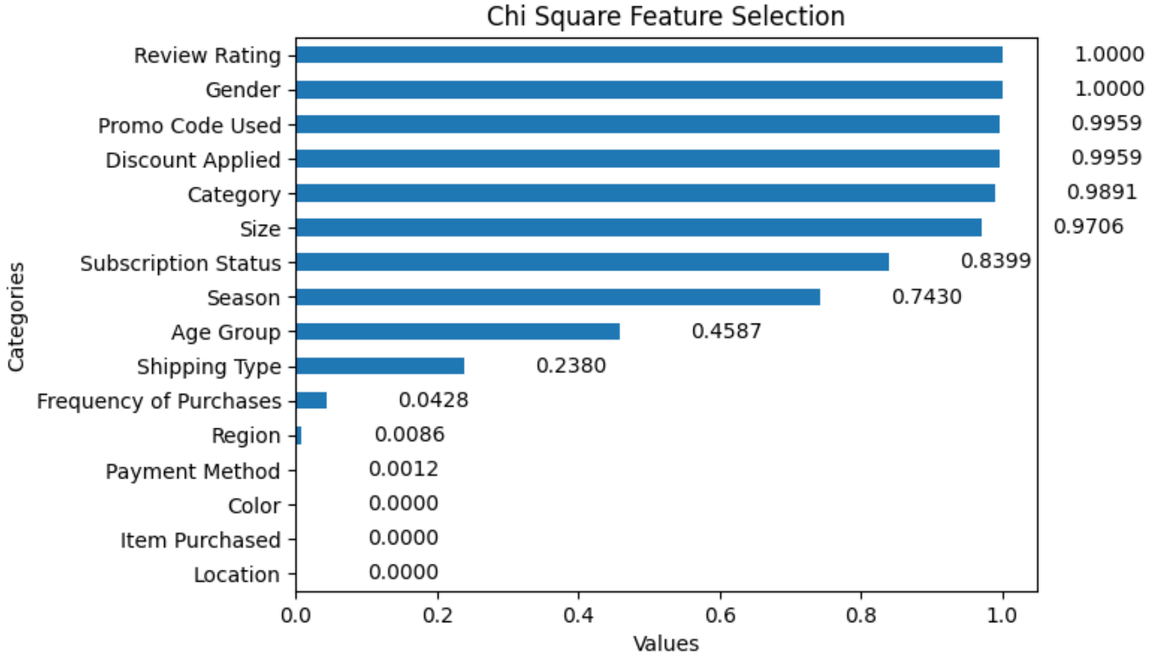


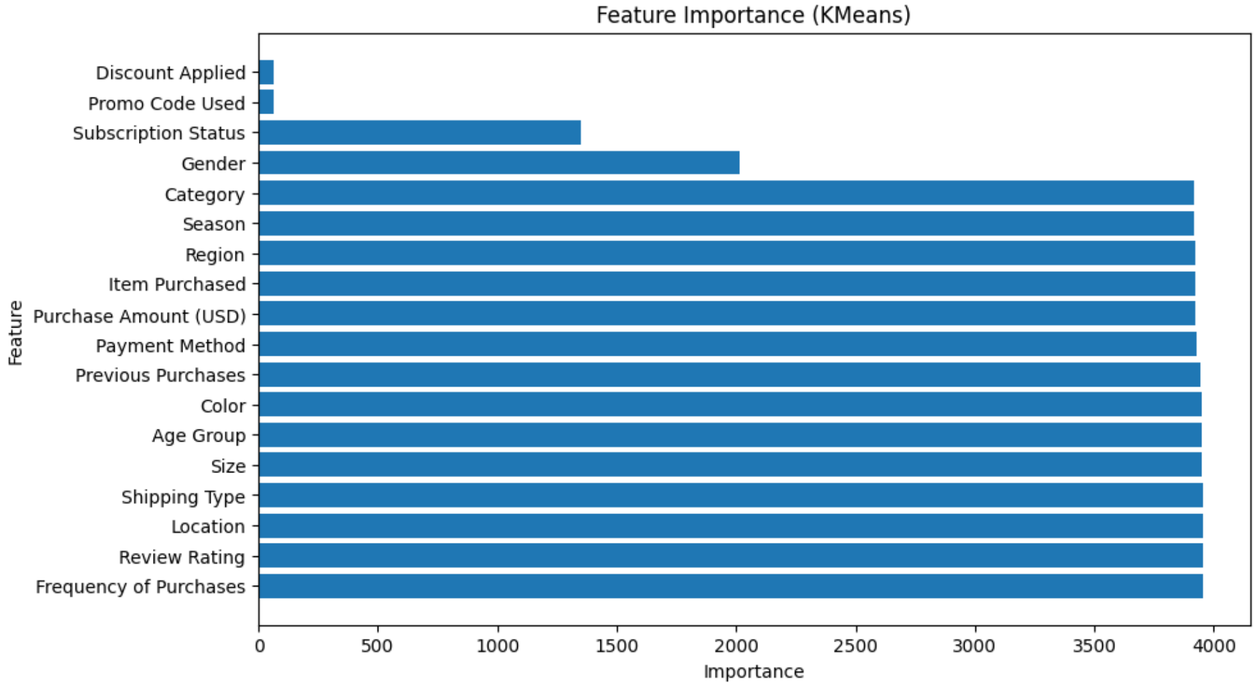


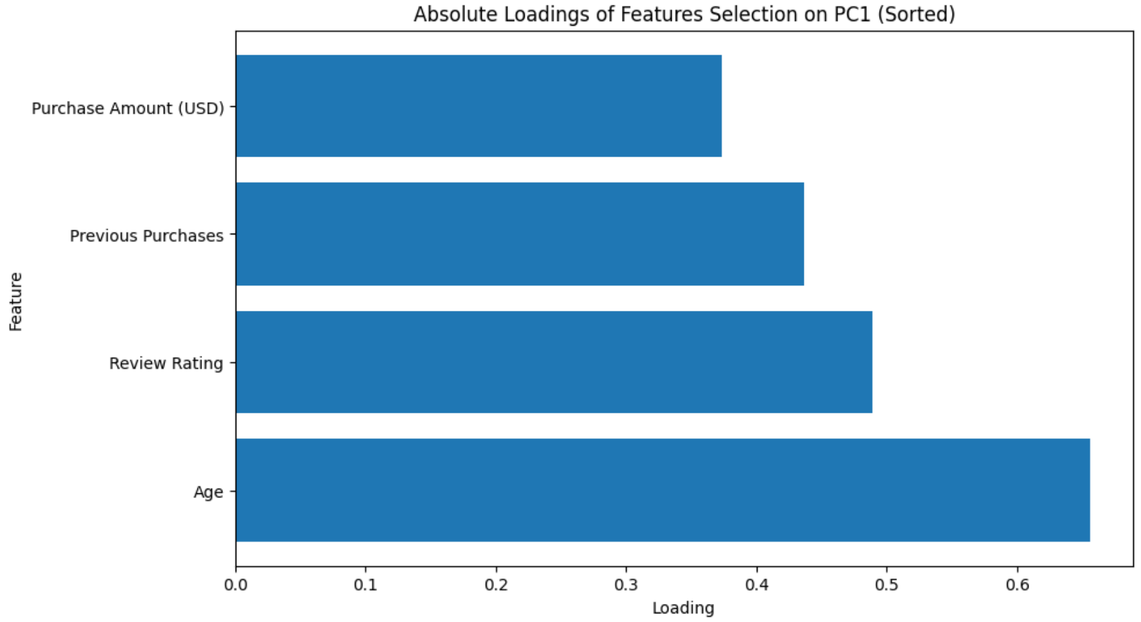
PREPROCESSING 1

FEATURE SELECTION

- Minimal Preprocessing
- Feature Engineering Age Group & Region
- LabelEncoder (Without One Hot Encoding)





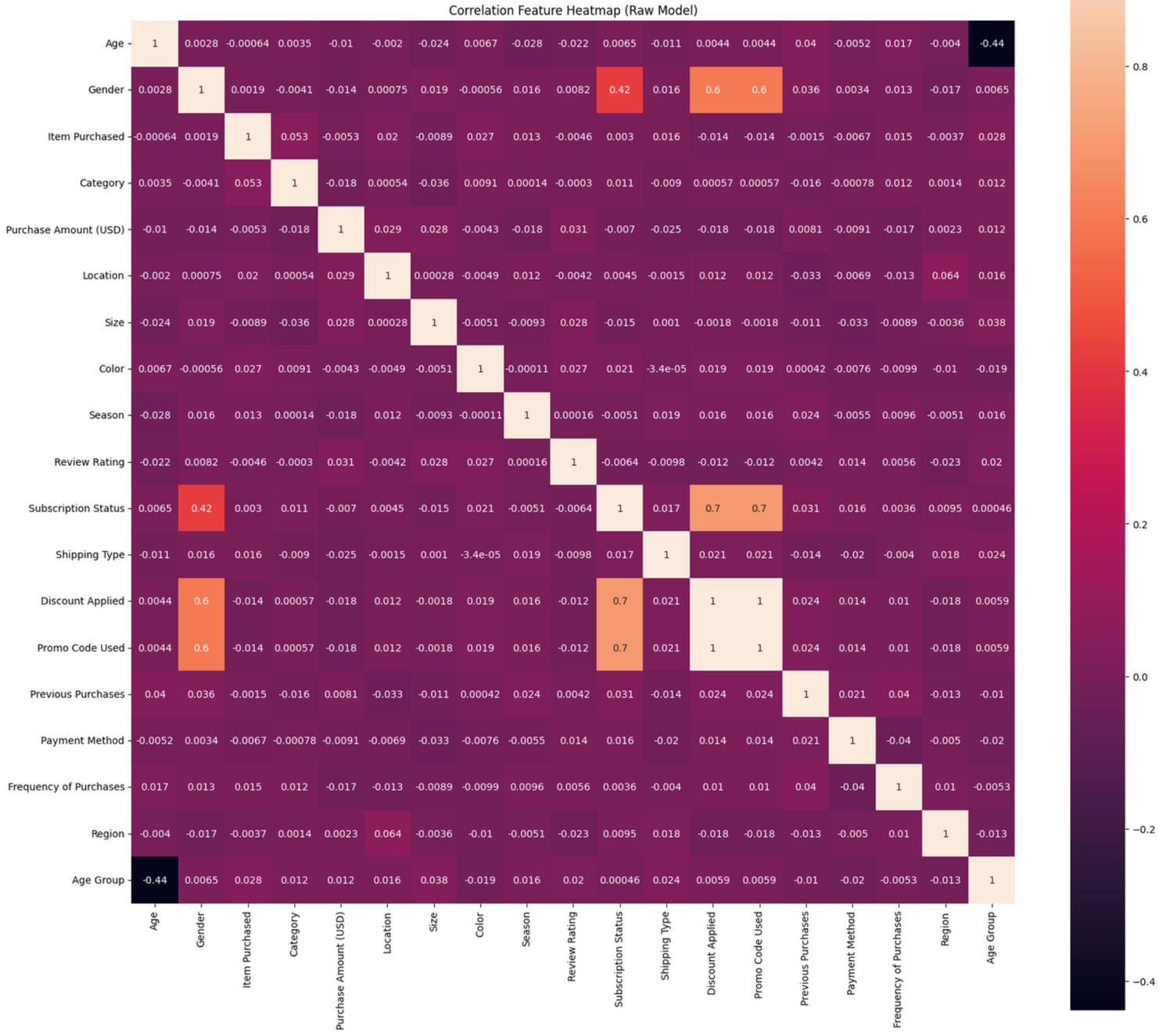


PREPROCESSING 1

FEATURE SELECTION

- Minimal Preprocessing
- Feature Engineering Age Group & Region
- LabelEncoder (Without One Hot Encoding)

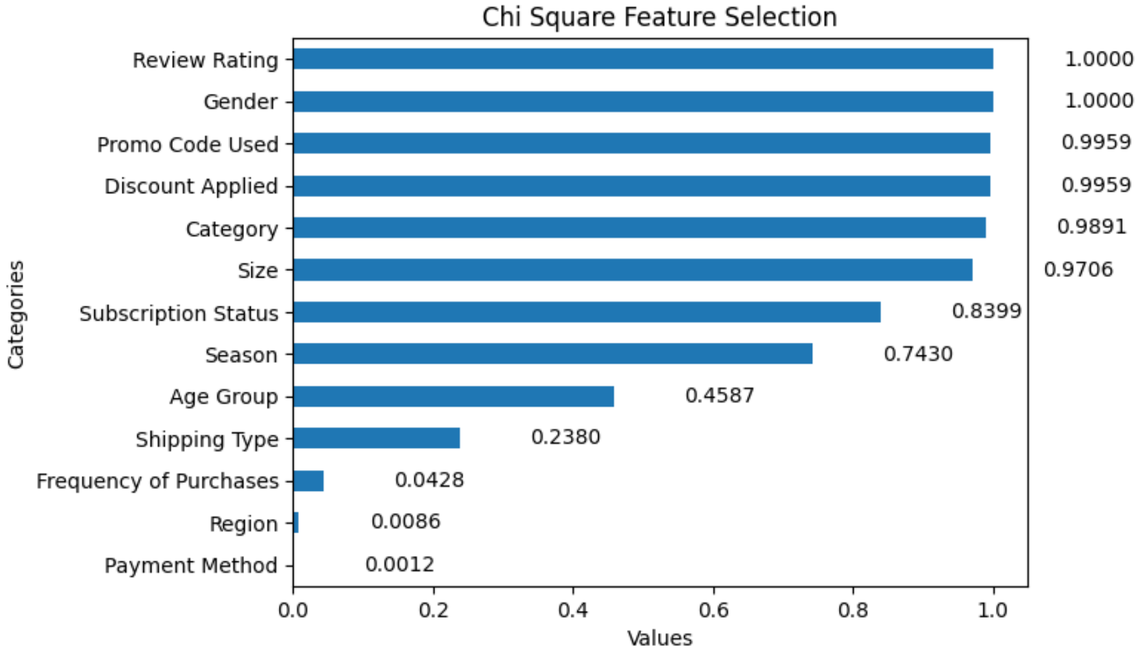
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3900 entries, 0 to 3899
Data columns (total 19 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Age                     3900 non-null   int64
 1   Gender                  3900 non-null   object
 2   Item Purchased          3900 non-null   object
 3   Category                3900 non-null   object
 4   Purchase Amount (USD)   3900 non-null   int64
 5   Location                3900 non-null   object
 6   Size                    3900 non-null   object
 7   Color                   3900 non-null   object
 8   Season                  3900 non-null   object
 9   Review Rating           3900 non-null   float64
 10  Subscription Status     3900 non-null   object
 11  Shipping Type           3900 non-null   object
 12  Discount Applied        3900 non-null   object
 13  Promo Code Used         3900 non-null   object
 14  Previous Purchases      3900 non-null   int64
 15  Payment Method          3900 non-null   object
 16  Frequency of Purchases  3900 non-null   object
 17  Region                  3900 non-null   object
 18  Age Group               3900 non-null   object
dtypes: float64(1), int64(3), object(15)
memory usage: 579.0+ KB
```

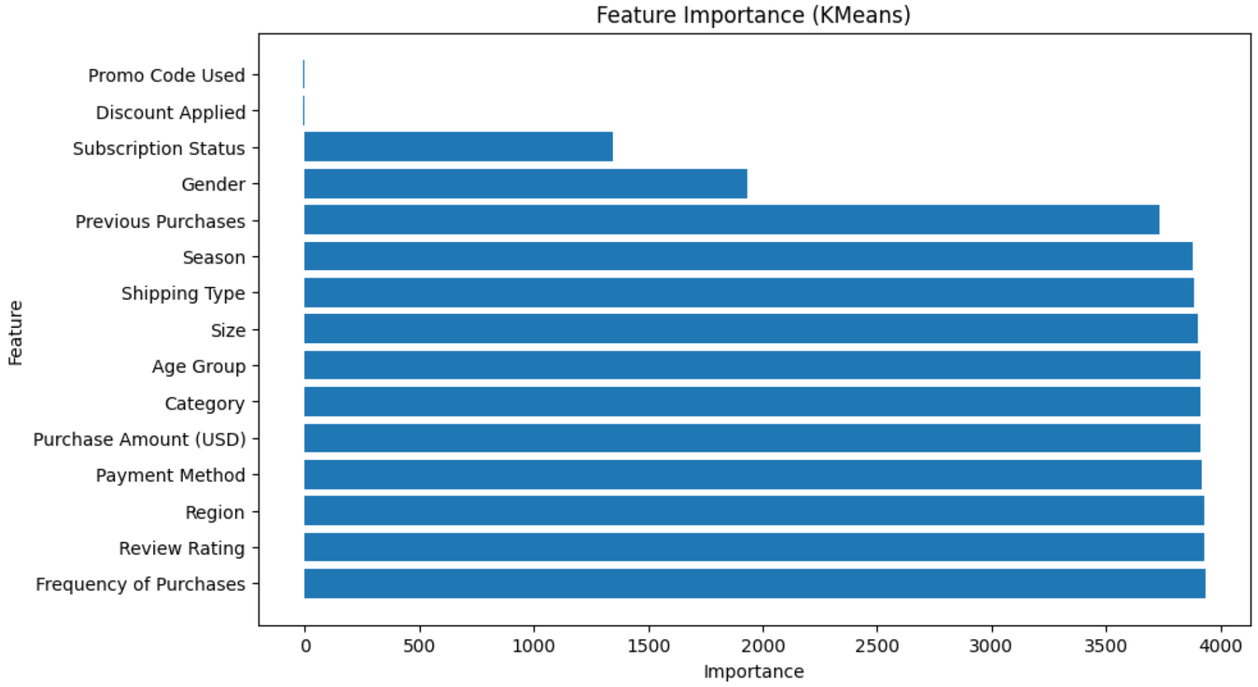


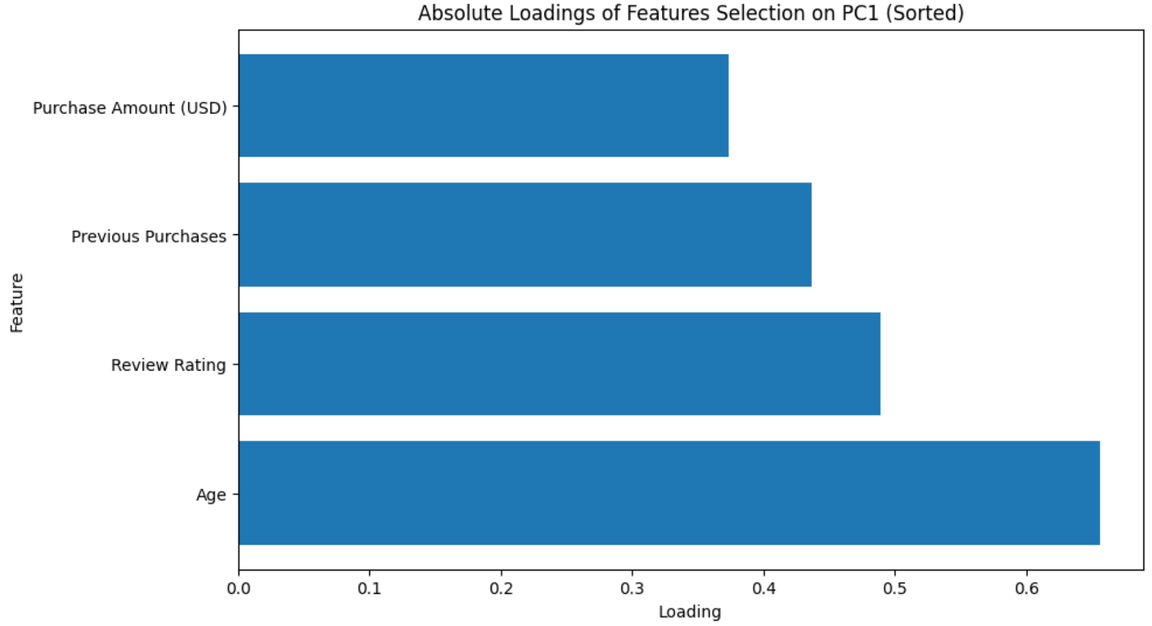
PREPROCESSING 2

FEATURE SELECTION

- Nan Encoding Preprocessing
- Feature Engineering Age Group & Region
- Drop Heavy Unique Column (Color, Location, Item Purchased)
- LabelEncoder (Without One Hot Encoding)





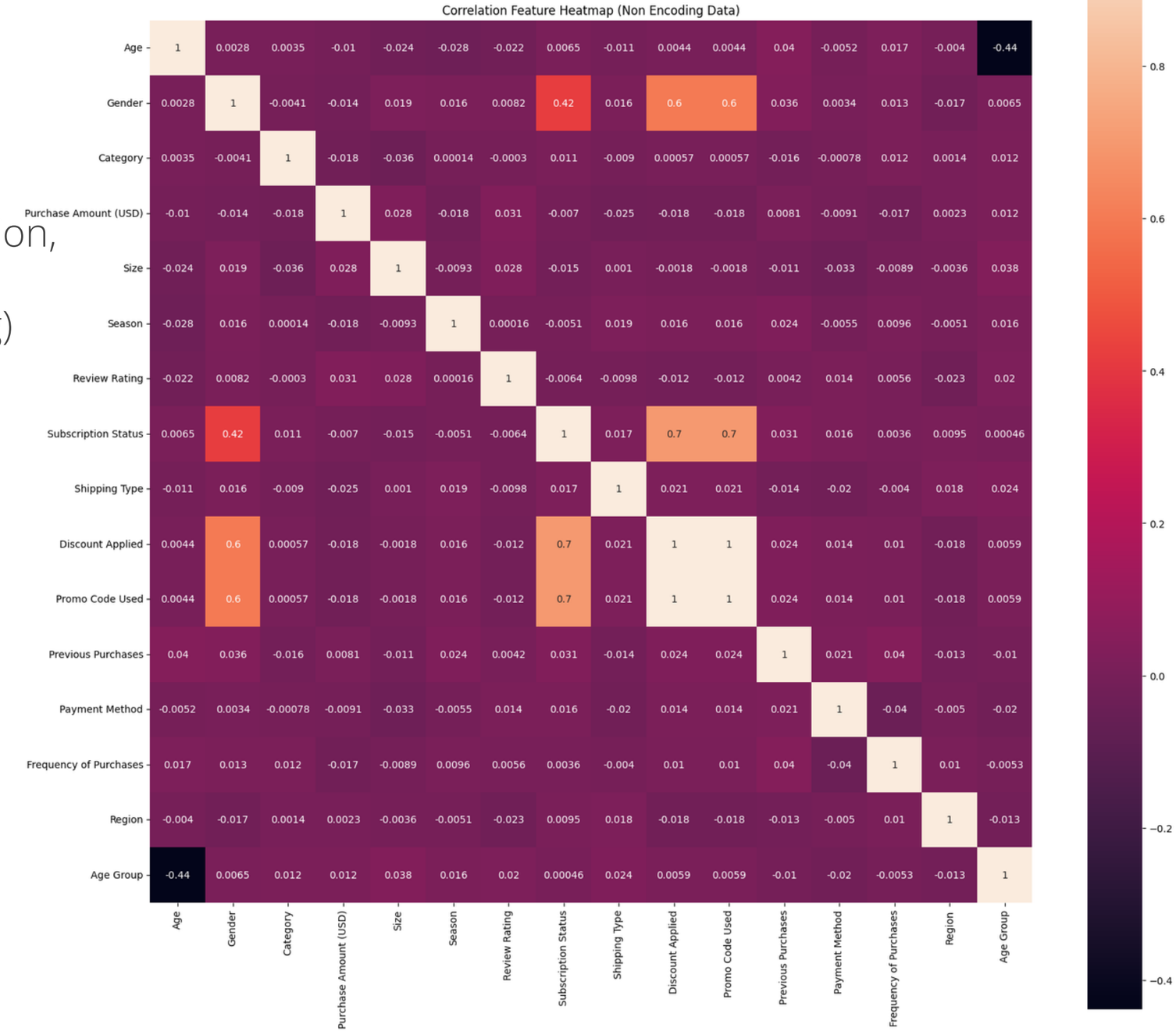


PREPROCESSING 2

FEATURE SELECTION

- Nan Encoding Preprocessing
- Feature Engineering Age Group & Region
- Drop Heavy Unique Column (Color, Location, Item Purchased)
- LabelEncoder (Without One Hot Encoding)

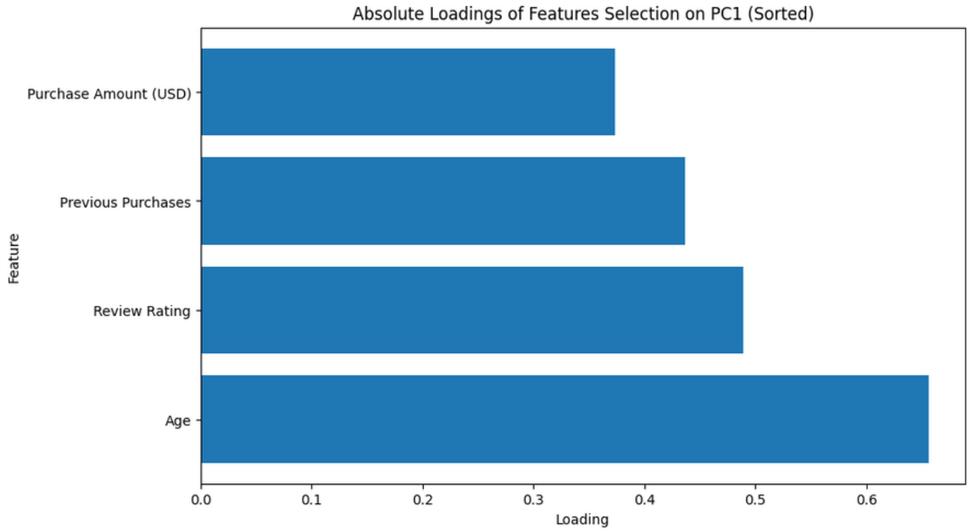
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3900 entries, 0 to 3899
Data columns (total 16 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Age                     3900 non-null   int64
 1   Gender                  3900 non-null   object
 2   Category                3900 non-null   object
 3   Purchase Amount (USD)   3900 non-null   int64
 4   Size                    3900 non-null   object
 5   Season                  3900 non-null   object
 6   Review Rating           3900 non-null   float64
 7   Subscription Status     3900 non-null   object
 8   Shipping Type           3900 non-null   object
 9   Discount Applied        3900 non-null   object
 10  Promo Code Used         3900 non-null   object
 11  Previous Purchases      3900 non-null   int64
 12  Payment Method          3900 non-null   object
 13  Frequency of Purchases  3900 non-null   object
 14  Region                  3900 non-null   object
 15  Age Group               3900 non-null   object
dtypes: float64(1), int64(3), object(12)
memory usage: 487.6+ KB
```

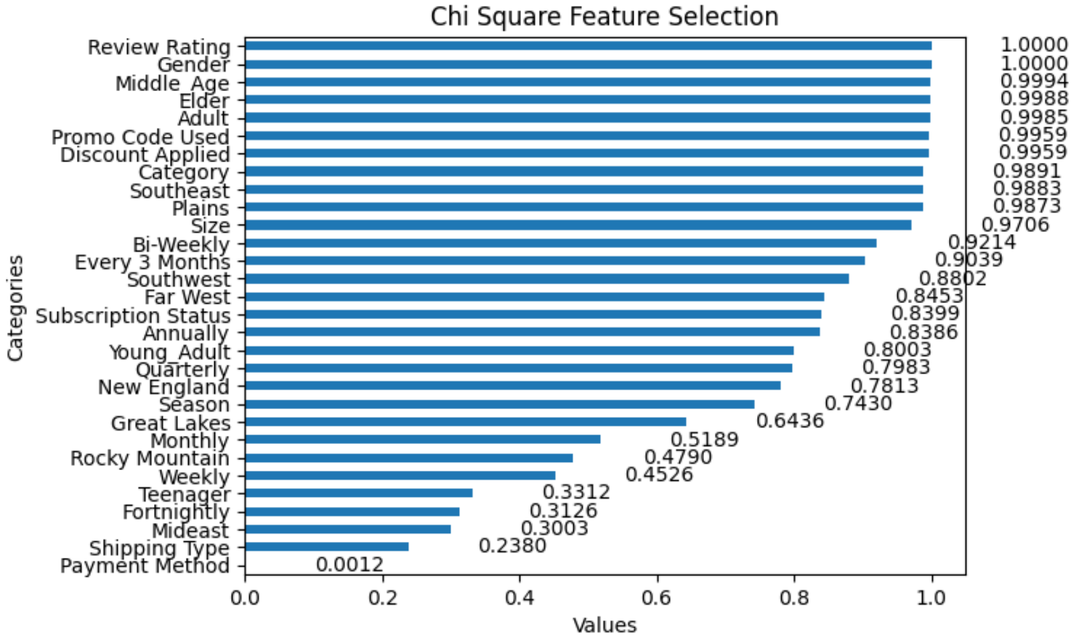


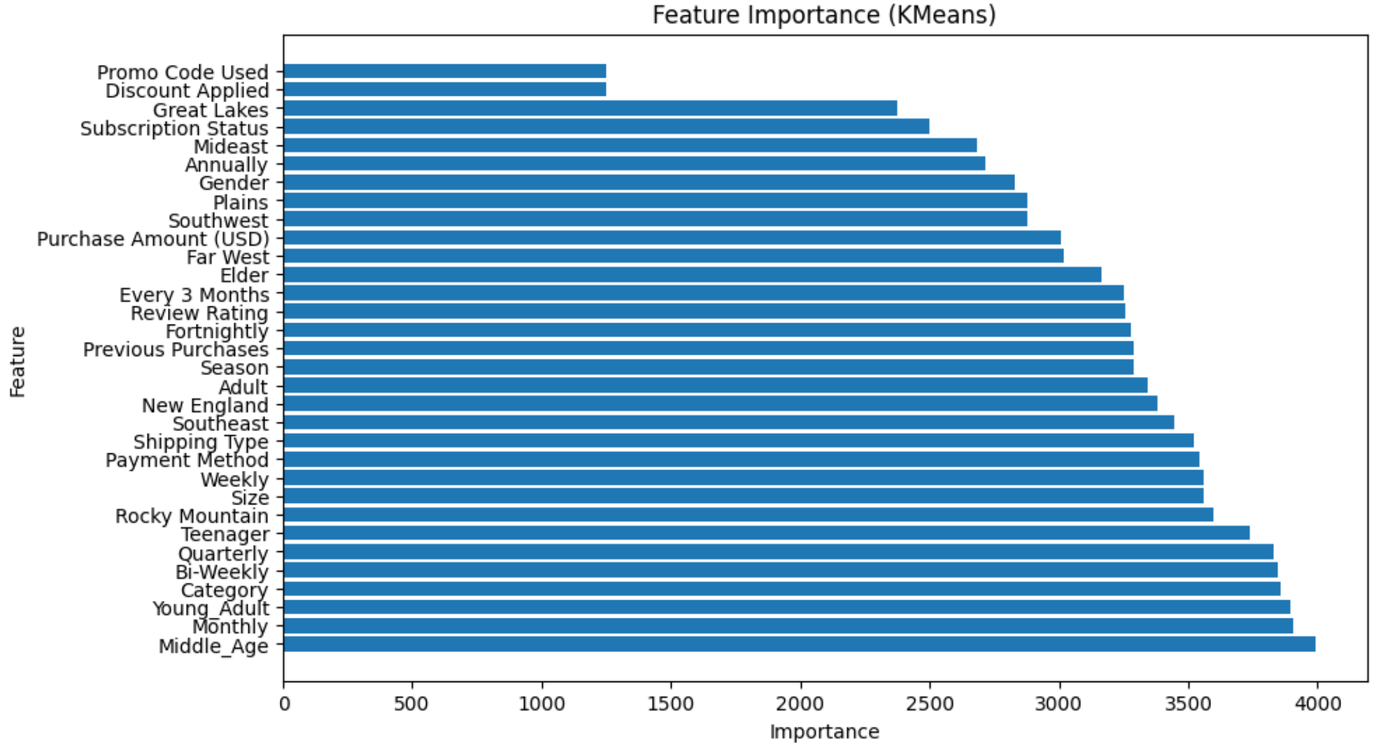
PREPROCESSING 3

FEATURE SELECTION

- Full Preprocessing
- Feature Engineering Age Group & Region
- Drop Heavy Unique Column (Color, Location, Item Purchased)
- LabelEncoder + One Hot Encoding







PREPROCESSING 3

FEATURE SELECTION

- Full Preprocessing
- Feature Engineering Age Group & Region
- Drop Heavy Unique Column (Color, Location, Item Purchased)
- LabelEncoder + One Hot Encoding

```python
# Method for preprocessing
def prep(df):

  # Convert kategorikal ke variabel indikator
  df_dummies_freq = pd.get_dummies(df['Frequency of Purchases'])
  df_dummies_reg = pd.get_dummies(df['Region'])
  df_dummies_age = pd.get_dummies(df['Age Group'])

  # Menggabungkan DataFrame asli dengan DataFrame hasil get_dummies
  df = pd.concat([df, df_dummies_freq, df_dummies_reg, df_dummies_age], axis=1)

  # Hapus kolom yang sudah di convert
  df.drop(columns=['Frequency of Purchases', 'Region', 'Age Group'], axis=1, inplace=True)

  return df
```

	Annually	Bi-Weekly	Every 3 Months	Fortnightly	Monthly	Quarterly	Weekly
0	0	0	0	1	0	0	0
1	0	0	0	1	0	0	0
2	0	0	0	0	0	0	1
3	0	0	0	0	0	0	1
4	1	0	0	0	0	0	0

Frequency Of Purchase (7 Unique Value)

Region (8 Unique Value)

Adult	Elder	Middle_Age	Teenager	Young_Adult
0	0	1	0	0
0	0	0	1	0
0	0	1	0	0
0	0	0	0	1
0	0	1	0	0

Far West	Great Lakes	Mideast	New England	Plains	Rocky Mountain	Southeast	Southwest
0	0	0	0	0	0	1	0
0	0	0	1	0	0	0	0
0	0	0	1	0	0	0	0
0	0	0	1	0	0	0	0
1	0	0	0	0	0	0	0

Age Group (5 Unique Value)

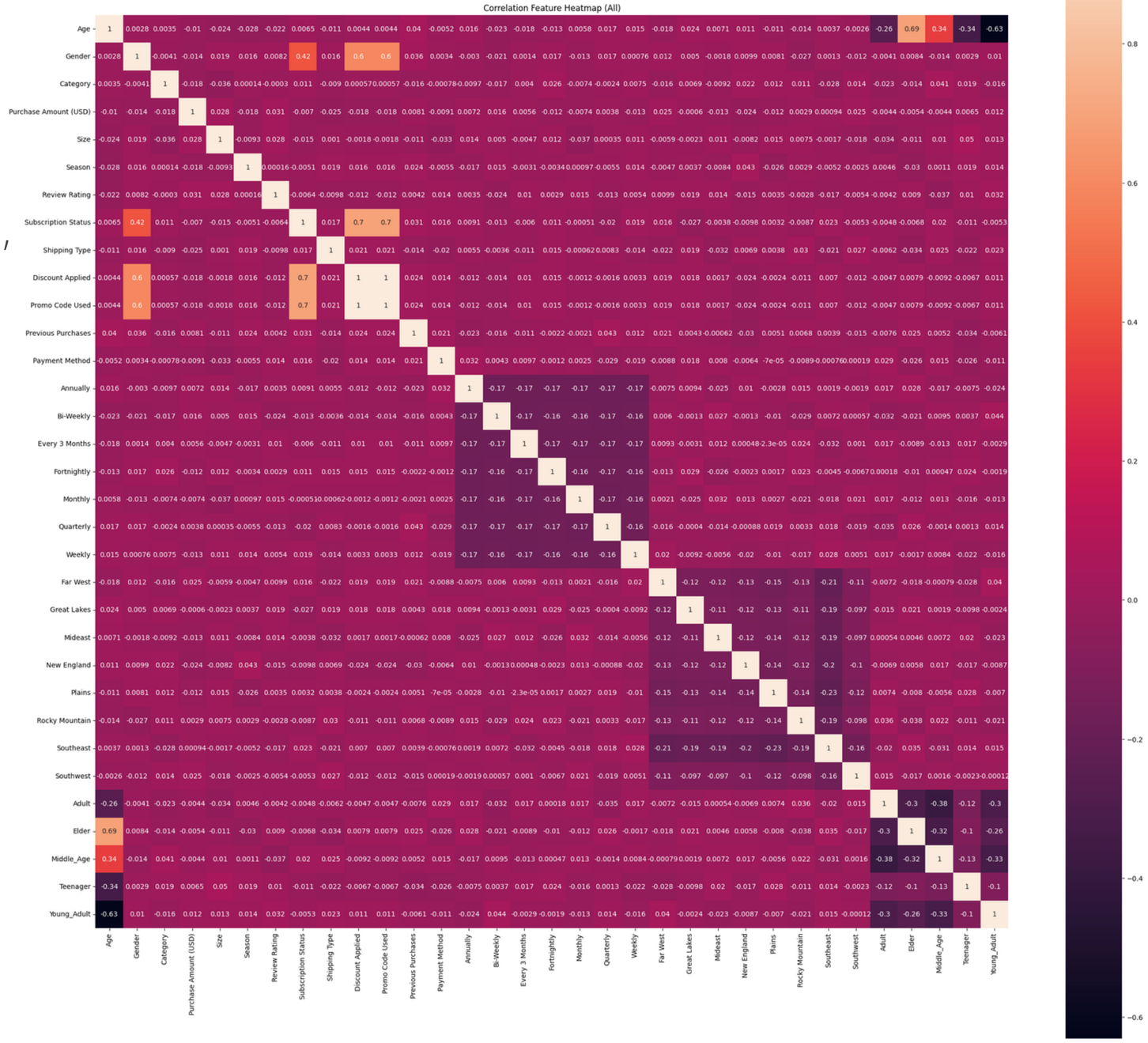
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3900 entries, 0 to 3899
Data columns (total 20 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Annually        3900 non-null   uint8
 1   Bi-Weekly       3900 non-null   uint8
 2   Every 3 Months  3900 non-null   uint8
 3   Fortnightly     3900 non-null   uint8
 4   Monthly         3900 non-null   uint8
 5   Quarterly       3900 non-null   uint8
 6   Weekly          3900 non-null   uint8
 7   Far West        3900 non-null   uint8
 8   Great Lakes     3900 non-null   uint8
 9   Mideast         3900 non-null   uint8
 10  New England     3900 non-null   uint8
 11  Plains          3900 non-null   uint8
 12  Rocky Mountain  3900 non-null   uint8
 13  Southeast       3900 non-null   uint8
 14  Southwest       3900 non-null   uint8
 15  Adult           3900 non-null   uint8
 16  Elder           3900 non-null   uint8
 17  Middle_Age      3900 non-null   uint8
 18  Teenager        3900 non-null   uint8
 19  Young_Adult     3900 non-null   uint8
dtypes: uint8(20)
memory usage: 76.3 KB
```

PREPROCESSING 3

FEATURE SELECTION

- Full Preprocessing
- Feature Engineering Age Group & Region
- Drop Heavy Unique Column (Color, Location, Item Purchased)
- LabelEncoder + One Hot Encoding

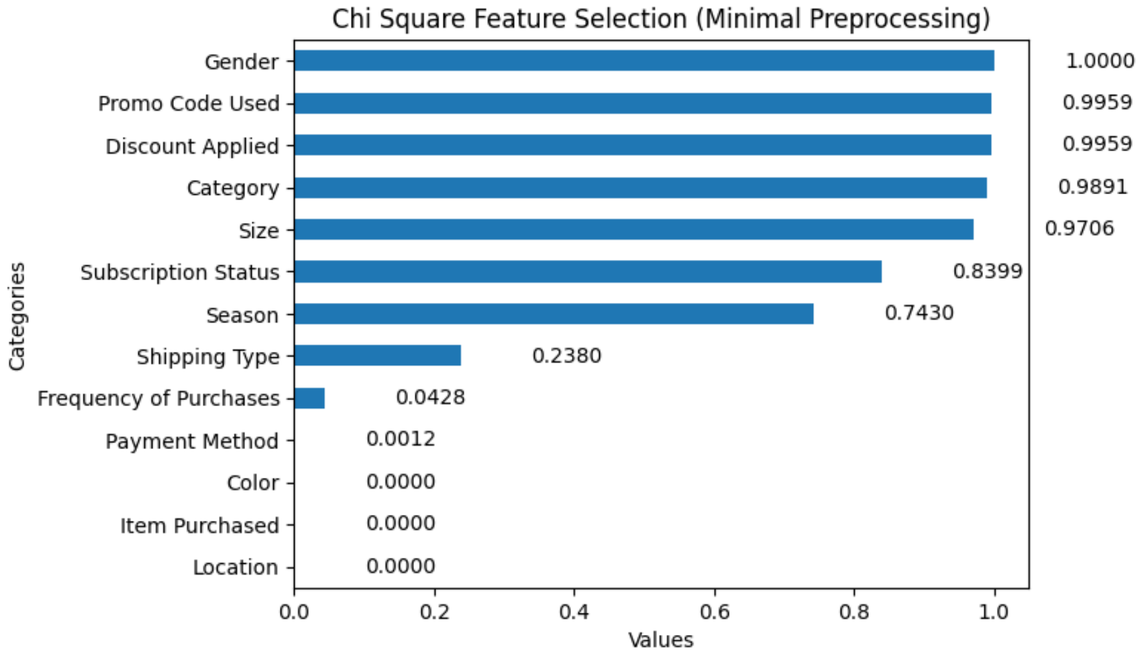
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3900 entries, 0 to 3899
Data columns (total 16 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   Age                     3900 non-null   int64  
 1   Gender                  3900 non-null   object 
 2   Category                3900 non-null   object 
 3   Purchase Amount (USD)   3900 non-null   int64  
 4   Size                    3900 non-null   object 
 5   Season                  3900 non-null   object 
 6   Review Rating           3900 non-null   float64
 7   Subscription Status     3900 non-null   object 
 8   Shipping Type           3900 non-null   object 
 9   Discount Applied        3900 non-null   object 
 10  Promo Code Used         3900 non-null   object 
 11  Previous Purchases      3900 non-null   int64  
 12  Payment Method          3900 non-null   object 
 13  Frequency of Purchases  3900 non-null   object 
 14  Region                  3900 non-null   object 
 15  Age Group               3900 non-null   object 
dtypes: float64(1), int64(3), object(12)
memory usage: 487.6+ KB
```

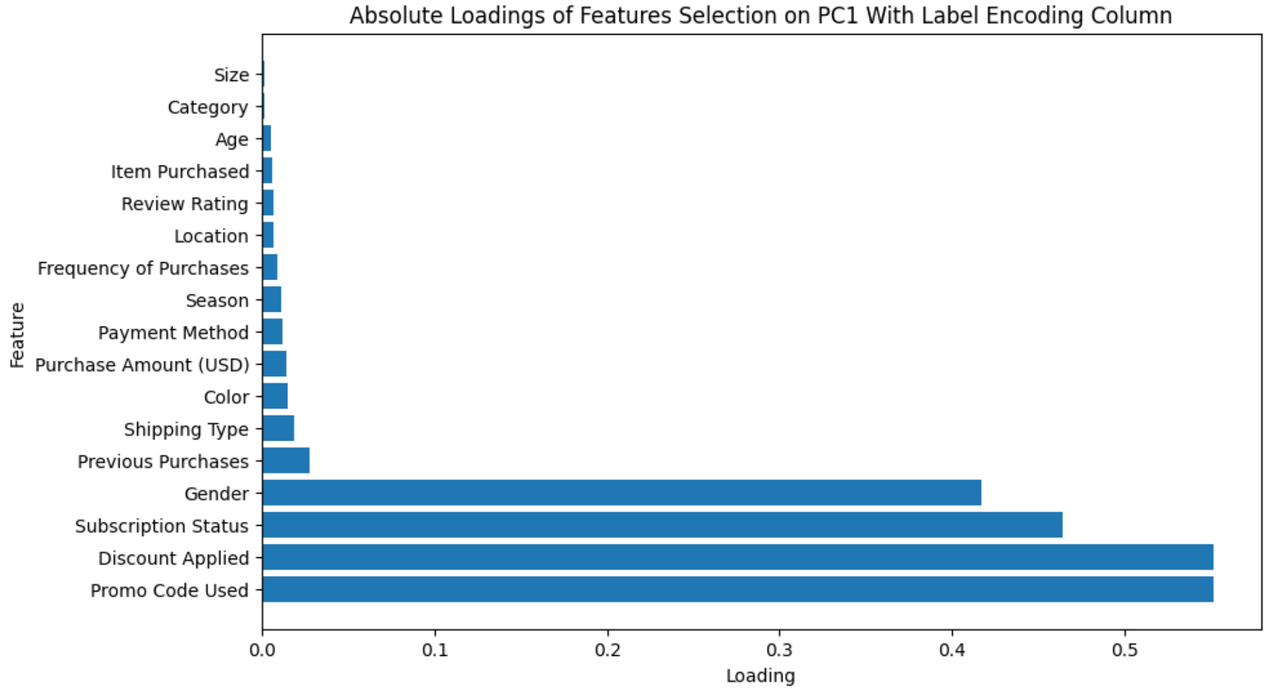


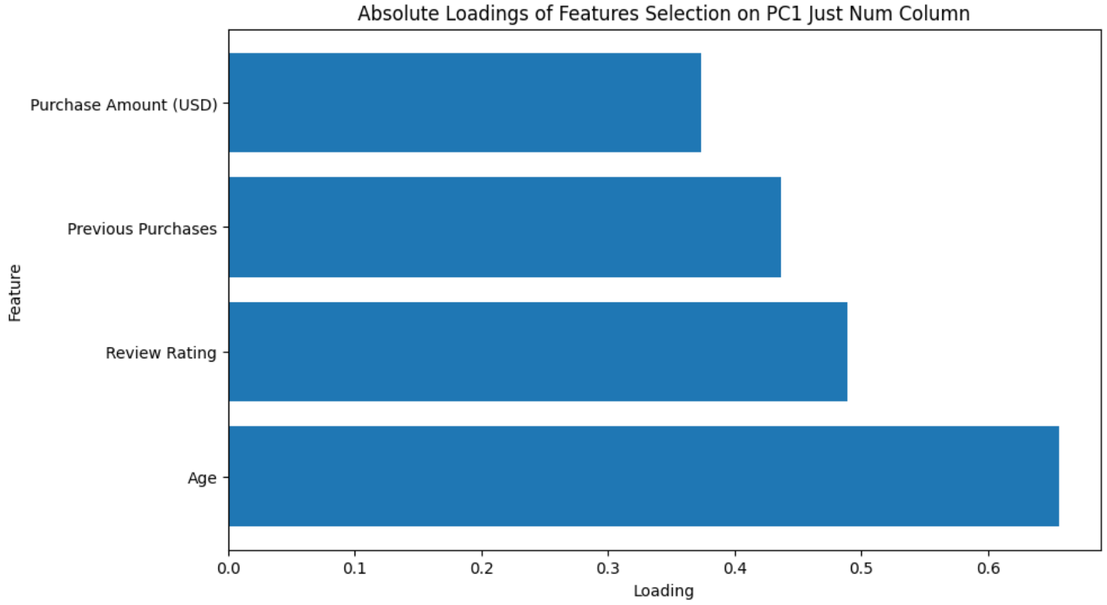
PREPROCESSING 4

FEATURE SELECTION

- Minimal Preprocessing
- Without Feature Engineering
- LabelEncoder All Features
    - Column With Higher Unique Value are more lower on the list compared to smaller one





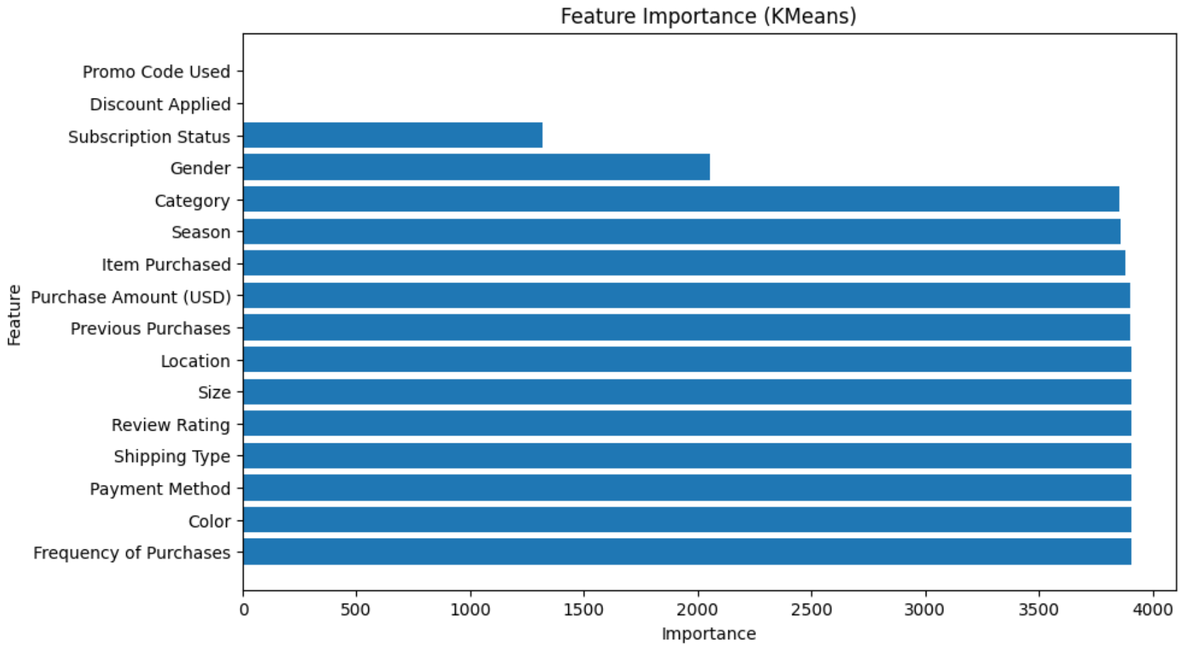


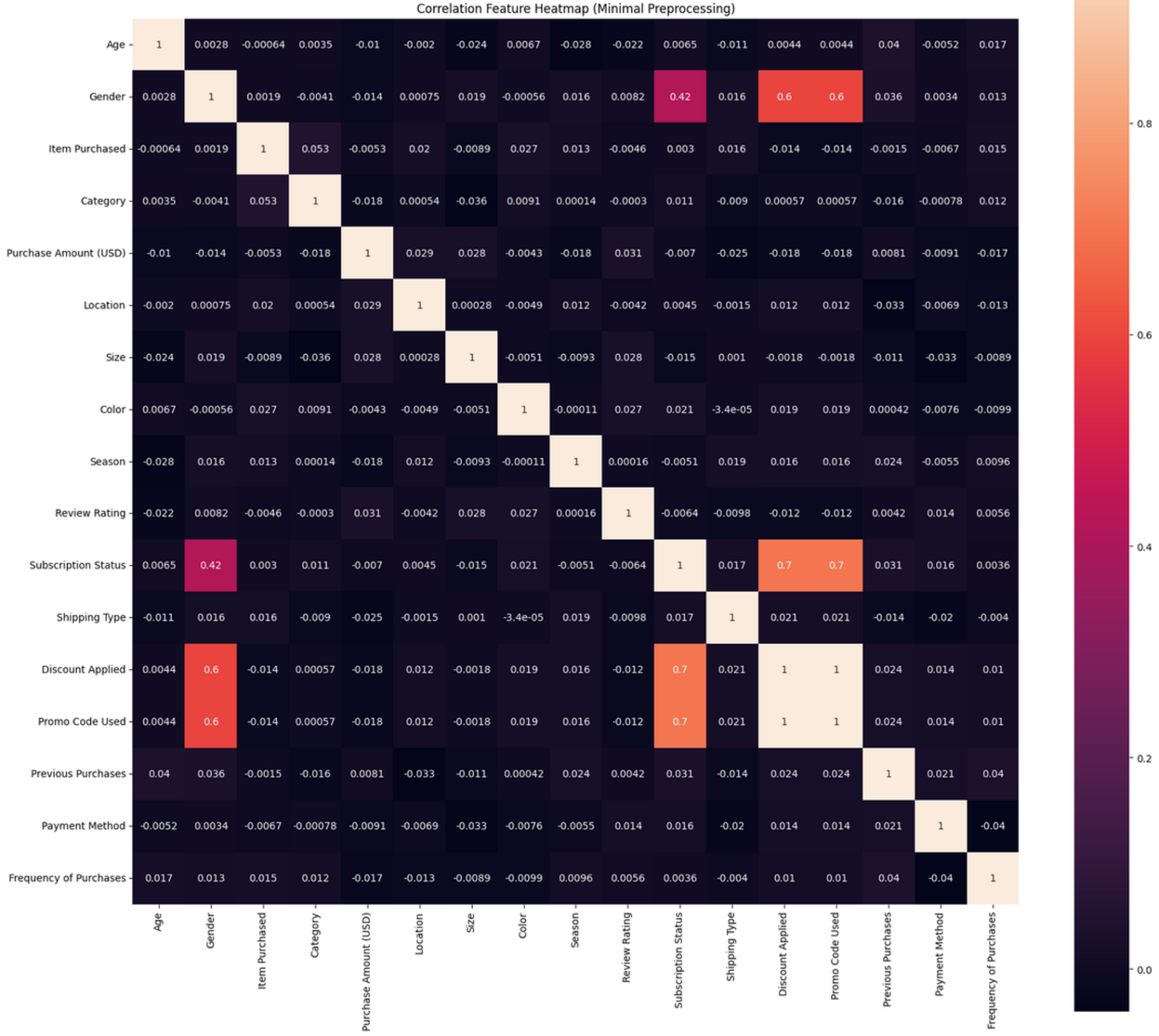
PREPROCESSING 4

FEATURE SELECTION

- Minimal Preprocessing
- Without Feature Engineering
- LabelEncoder All Features

Feature Selection KMeans

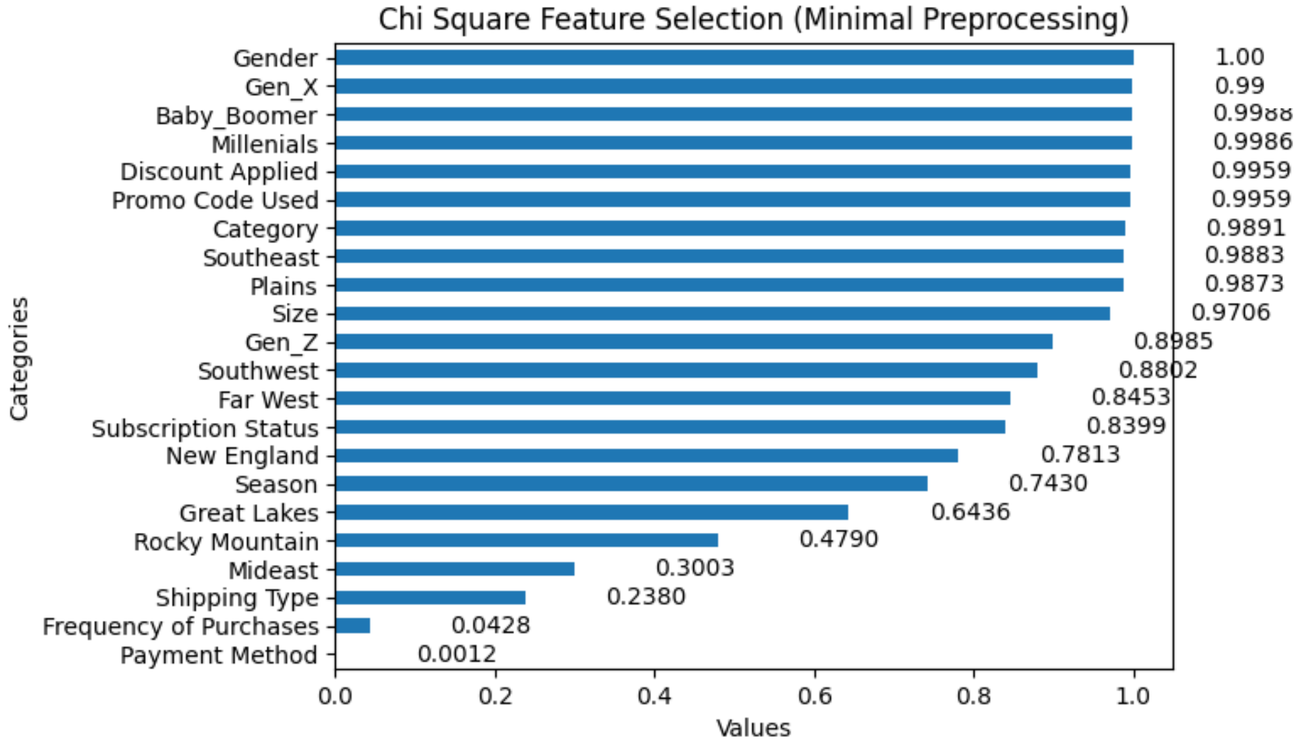


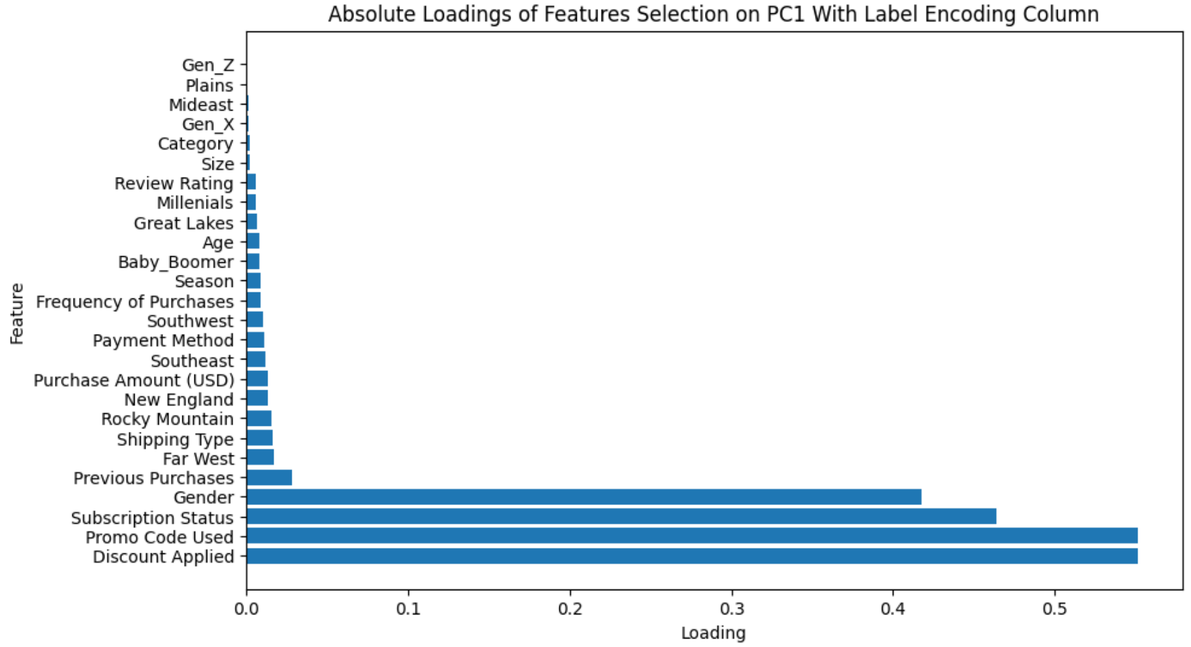


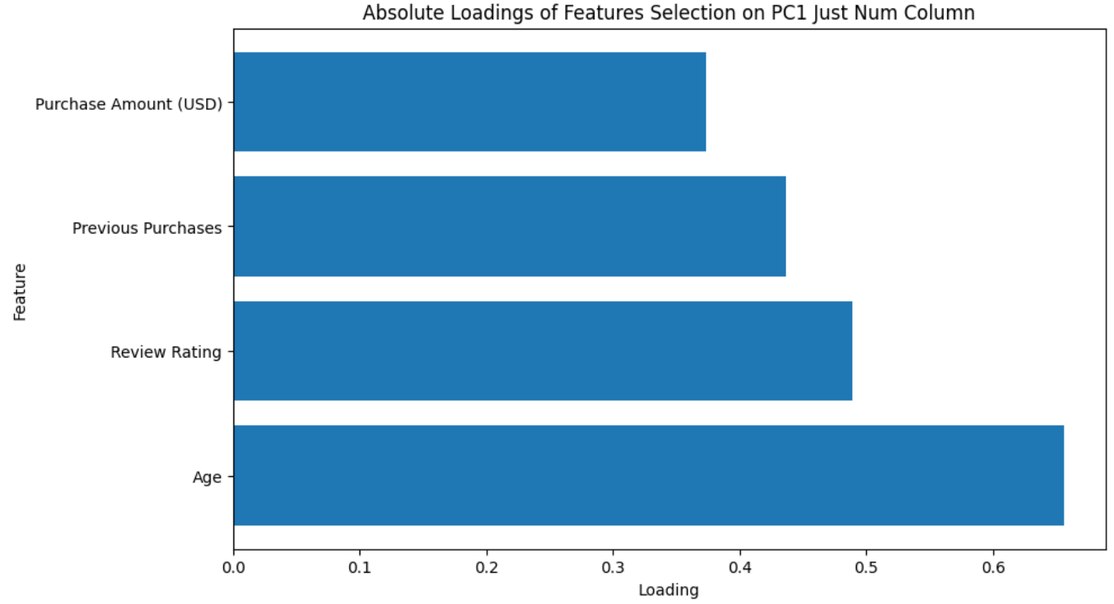
PREPROCESSING 5

FEATURE SELECTION

- Fully Preprocess
- Without Scaling
- Feature Engineer (Generation + Region)
- One Hot Encoding + LabelEncoder





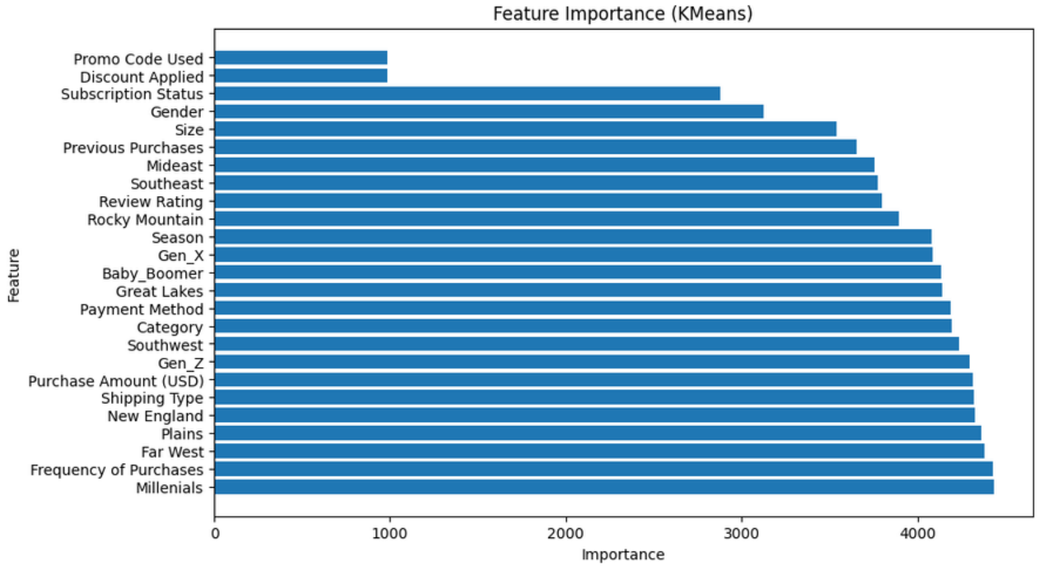


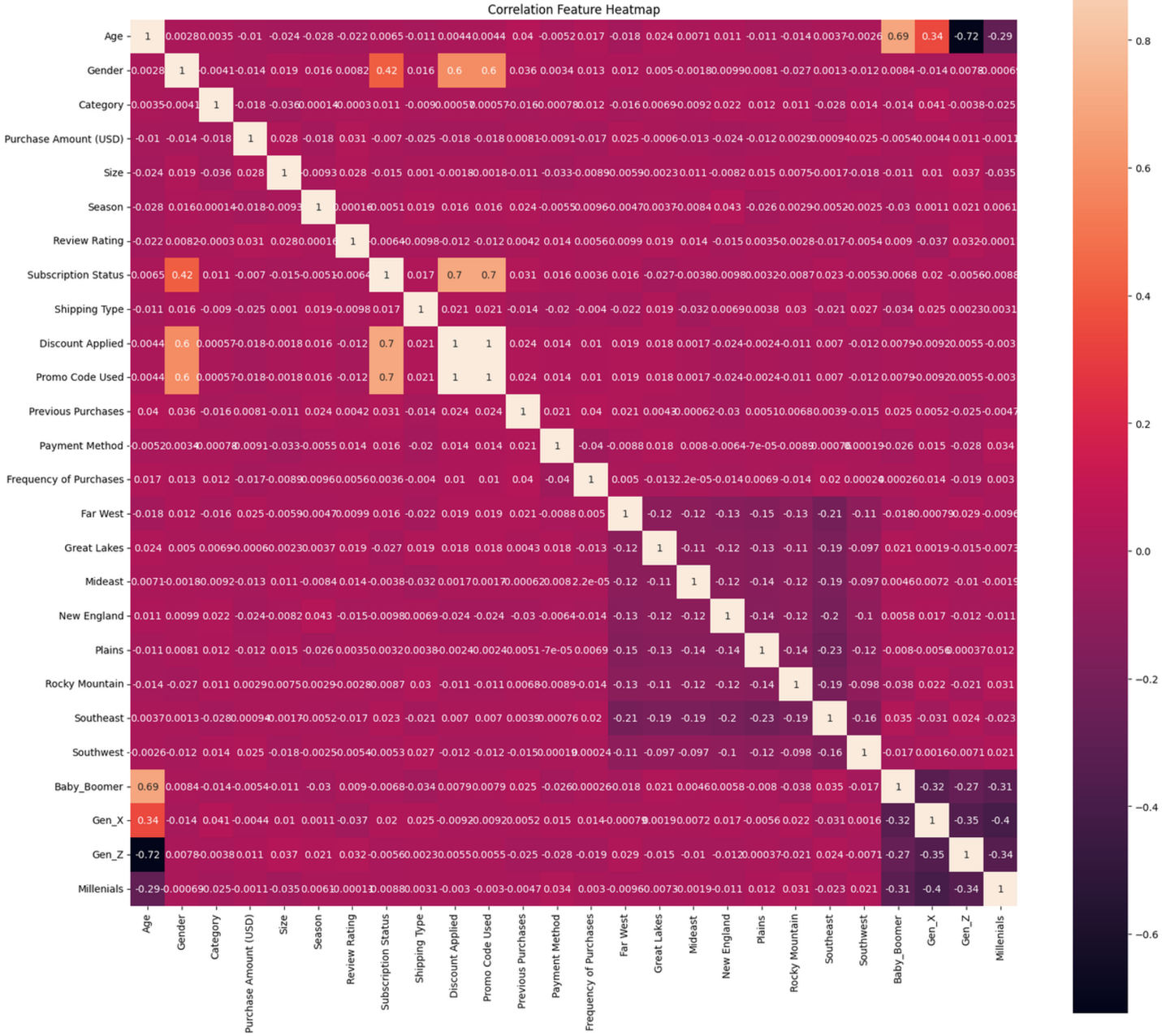
PREPROCESSING 4

FEATURE SELECTION

- Minimal Preprocessing
- Without Feature Engineering
- LabelEncoder All Features

Feature Selection
KMeans





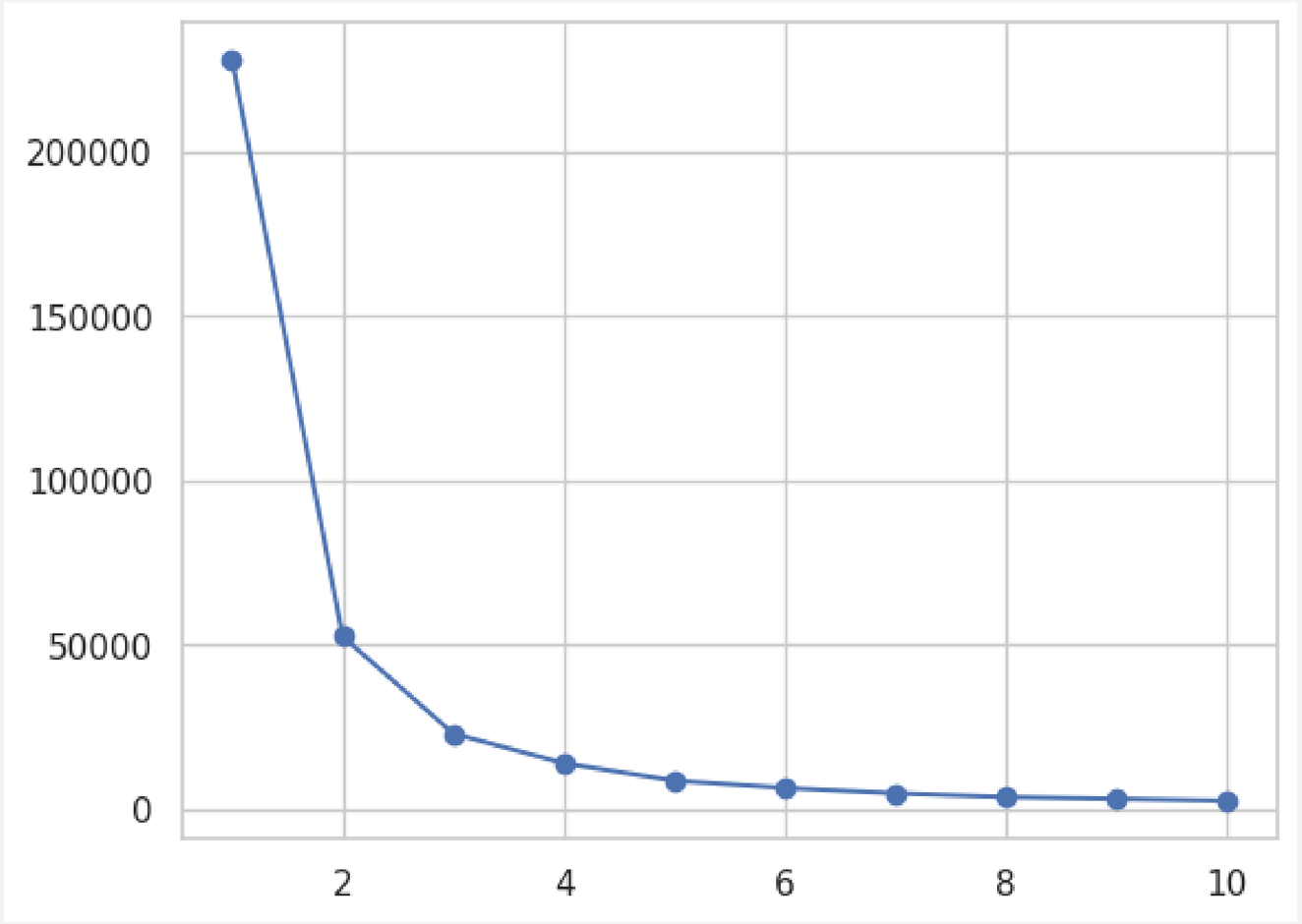
Modelling for Generation and Purchase Amount [Southeast](No-Scaling)

```
[276] southeast = df_encoded[df_encoded["Region"] == 1]
```

```
X = southeast[['Generation', 'Purchase Amount (USD)']]
X.head()
```

	Generation	Purchase Amount (USD)
36	3	49.0
37	3	25.0
43	2	49.0
74	1	38.0
87	1	33.0

Elbow Method to Determine the k for Number of Clusters



K-Means Algorithm

```
[281] km = KMeans(n_clusters=4, random_state=45, n_init="auto")
      km.fit(X)

      y_km = km.predict(X)
```

```
      pd.DataFrame(y_km).value_counts()
```

```
0    126
1     98
3     90
2     82
dtype: int64
```

Cluster Segmentation

Silhouette Score

	Model	Silhouette Score
0	K-Means No-Scaling	0.56858

	KMeans Cluster with No-Scaling	Gender	Generation	Category
0	1	Male	Gen X	Clothing
1	2	Male	Gen Z	Clothing
2	1	Male	Gen X	Clothing
3	2	Male	Gen Z	Footwear
4	2	Male	Gen X	Clothing
..	...	...	...	...
391	3	Male	Millenials	Clothing
392	2	Male	Gen Z	Clothing
393	2	Male	Millenials	Footwear
394	3	Male	Millenials	Accessories
395	3	Male	Millenials	Footwear

	Purchase Amount (USD)
0	53
1	64
2	73
3	90
4	49
..	...
391	86
392	82
393	65
394	29
395	65

[396 rows x 5 columns]

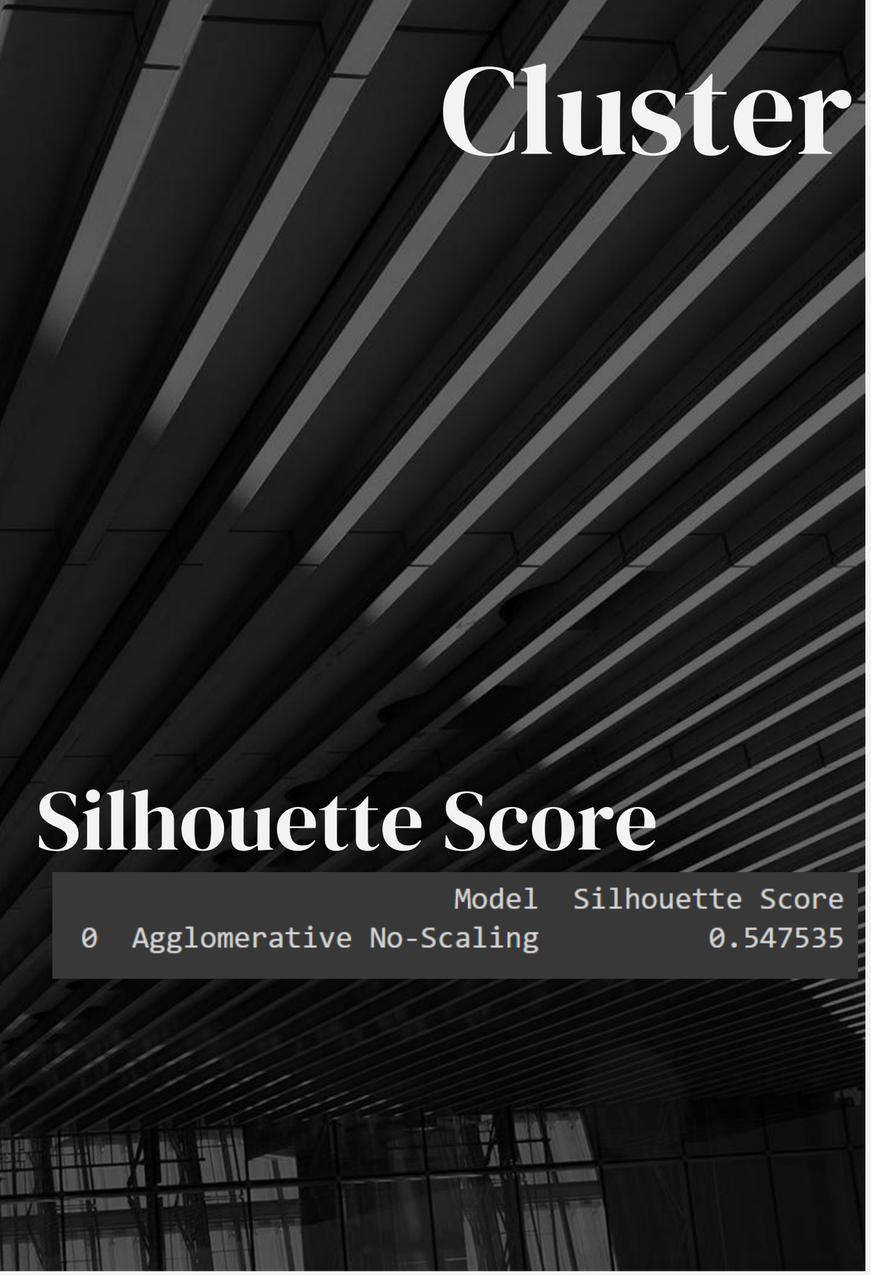
Agglomerative Algorithm

```
agg_model = AgglomerativeClustering(n_clusters=4)

agg_model.fit(X)

y_agg = agg_model.labels_
```

```
pd.DataFrame(y_agg).value_counts()
```

```
0    122
1    116
2     98
3     60
dtype: int64
```

Cluster Segmentation



Silhouette Score

	Model	Silhouette Score
0	Agglomerative No-Scaling	0.547535

	Agglomerative Cluster with No-Scaling	Gender	Generation	Category
0	0	Male	Gen X	Clothing
1	3	Male	Gen Z	Clothing
2	0	Male	Gen X	Clothing
3	0	Male	Gen Z	Footwear
4	3	Male	Gen X	Clothing
..	...	...	...	...
391	2	Male	Millenials	Clothing
392	0	Male	Gen Z	Clothing
393	3	Male	Millenials	Footwear
394	2	Male	Millenials	Accessories
395	2	Male	Millenials	Footwear

	Purchase Amount (USD)	Region
0	53	Southeast
1	64	New England
2	73	New England
3	90	New England
4	49	Far West
..	...	...
391	86	Plains
392	82	Rocky Mountain
393	65	New England
394	29	Plains
395	65	New England

[396 rows x 6 columns]

GMM Algorithm

```
[297] gmm = GaussianMixture(n_components=4, random_state=42)

      # Fit the model to the data
      gmm.fit(X)

      # Predict cluster labels
      labels = gmm.predict(X)
```

```
pd.DataFrame(labels).value_counts()
```

```
3    123
2    119
0     91
1     63
dtype: int64
```

Cluster Segmentation

Basic Computing Community

	GMM Cluster	Gender	Generation	Category	Purchase Amount (USD)	Region
0	3	Male	Gen X	Clothing	53	Southeast
1	1	Male	Gen Z	Clothing	64	New England
2	3	Male	Gen X	Clothing	73	New England
3	1	Male	Gen Z	Footwear	90	New England
4	1	Male	Gen X	Clothing	49	Far West
..	...	...	...	...	...	...
391	0	Male	Millenials	Clothing	86	Plains
392	3	Male	Gen Z	Clothing	82	Rocky Mountain
393	1	Male	Millenials	Footwear	65	New England
394	0	Male	Millenials	Accessories	29	Plains
395	0	Male	Millenials	Footwear	65	New England

[396 rows x 6 columns]

Silhouette Score

	Model	Silhouette Score
0	Agglomerative No-Scaling	0.547535
1	Gaussian Generation Purchase Amount (USD) No-S...	0.545607