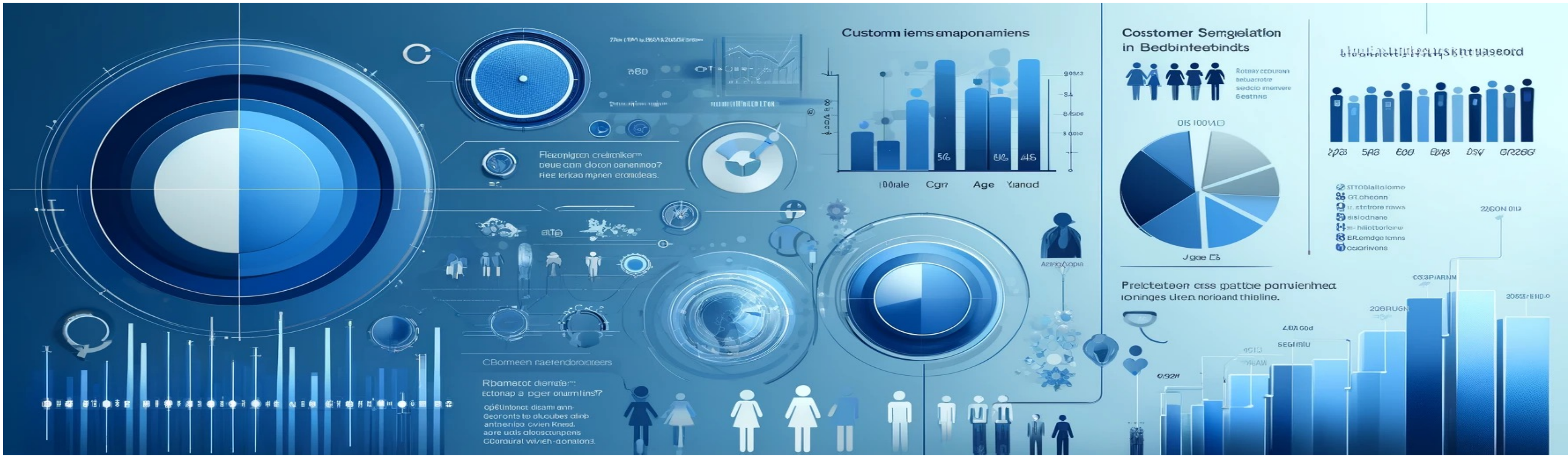


CUSTOMER SEGMENTATION

FINAL PROJECT:

UNSUPERVISED ML MODEL TO SEGMENT CUSTOMERS BASED ON DEMOGRAPHICS AND SPENDING BEHAVIOUR

CUSTOMER SEGMENTATION

The primary business problem this model aims to solve is the need for **effective customer segmentation** based on key behavioral and demographic factors. Specifically, the business requires a model that can:

1. **Identify Distinct Customer Segments:** businesses can better understand the varying needs and preferences of different customer groups.
2. **Optimize Marketing Strategies:** with well-defined customer segments, the business can develop targeted marketing campaigns that resonate with each specific segment.
3. **Improve Customer Retention and Satisfaction:** businesses can tailor its products and services to meet the specific needs of each group.
4. **Maximize Revenue and Profitability:** Effective segmentation allows the business to allocate resources more efficiently, focusing on high-value segments with the potential for increased revenue.

DATA LOADING, ANALYSIS AND CLEANING

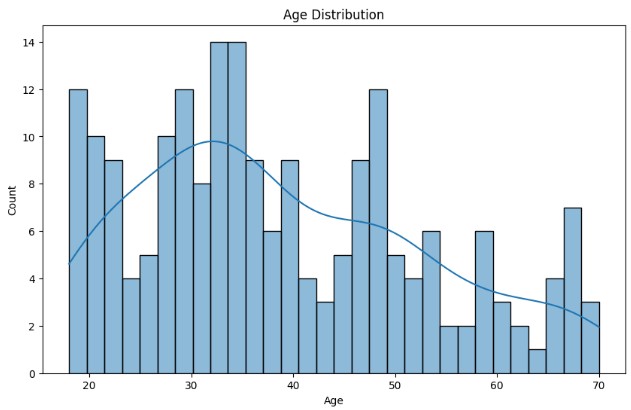
1. The csv file from Kaggle contains the customer dataset that will be used in this project. It can be found here https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python

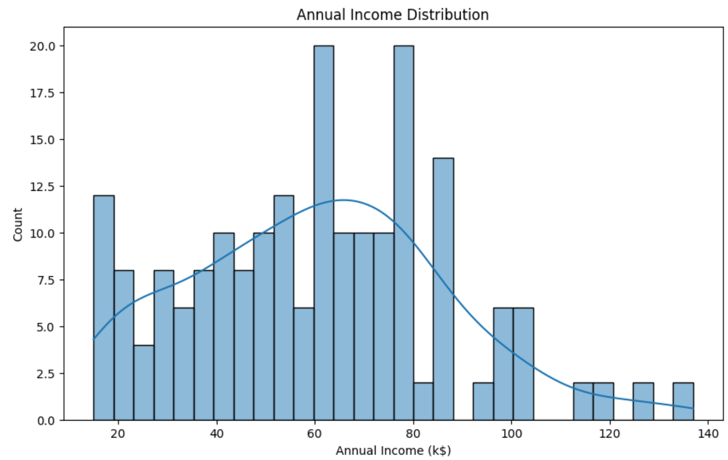
2. Data frame contains 200 entries and 5 attributes

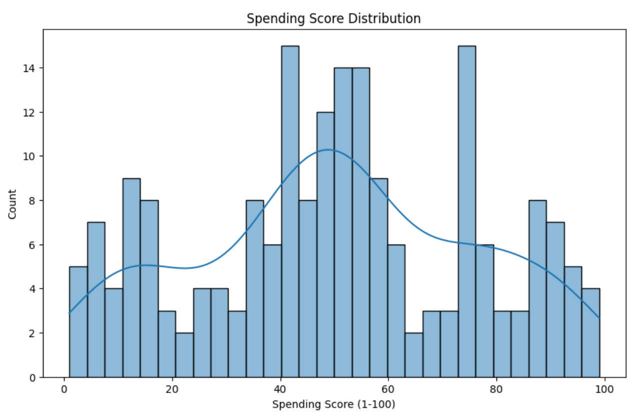
```
   CustomerID  Gender  Age  Annual Income (k$)  Spending Score (1-100)
0           1    Male   19                  15                      39
1           2    Male   21                  15                      81
2           3  Female   20                  16                       6
3           4  Female   23                  16                      77
4           5  Female   31                  17                      40
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   CustomerID              200 non-null    int64
 1   Gender                  200 non-null    object
 2   Age                     200 non-null    int64
 3   Annual Income (k$)      200 non-null    int64
 4   Spending Score (1-100)  200 non-null    int64
dtypes: int64(4), object(1)
memory usage: 7.9+ KB
None
       CustomerID         Age  Annual Income (k$)  Spending Score (1-100)
count  200.000000  200.000000          200.000000              200.000000
mean   100.500000   38.850000           60.560000               50.200000
std     57.879185   13.969007           26.264721               25.823522
min      1.000000   18.000000           15.000000                1.000000
25%     50.750000   28.750000           41.500000               34.750000
50%    100.500000   36.000000           61.500000               50.000000
75%    150.250000   49.000000           78.000000               73.000000
max    200.000000   70.000000          137.000000               99.000000
```

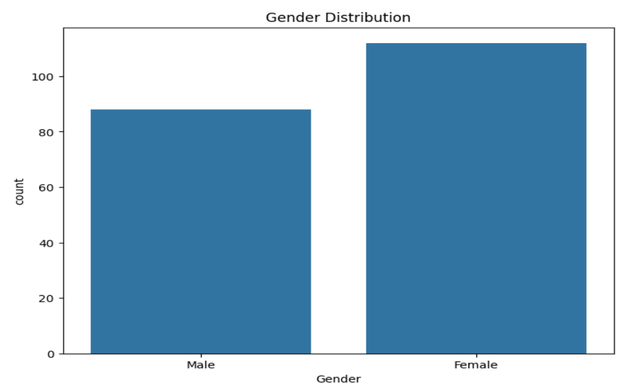
3. No cleanup needed, all attributes are informed and the are no anomalies.

VISUALIZATIONS - DISTRIBUTIONS

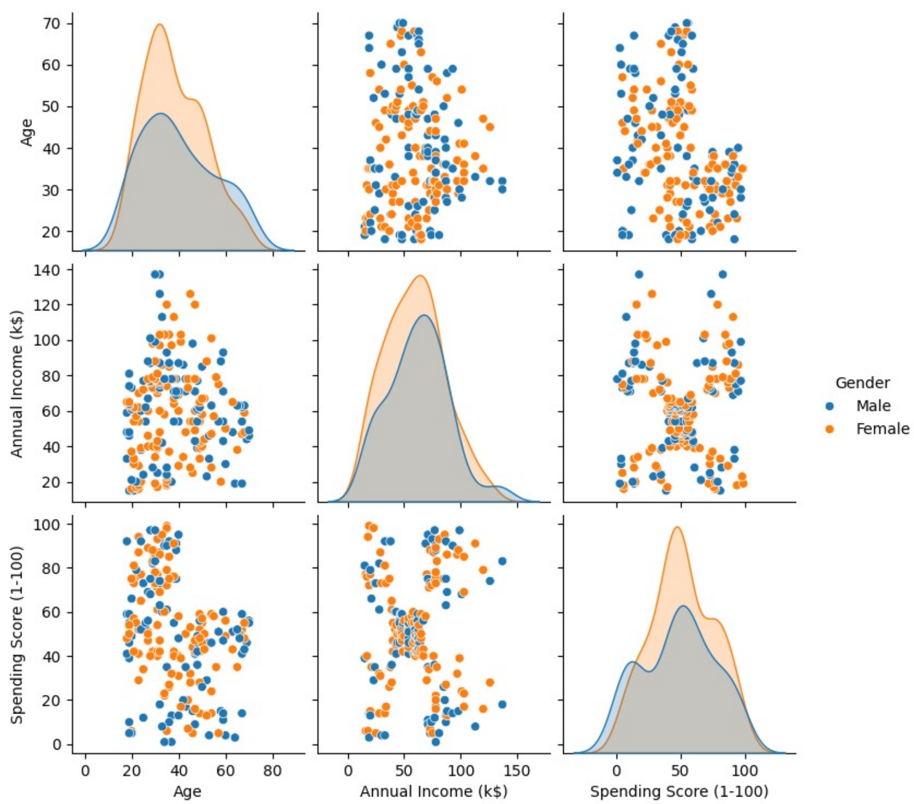








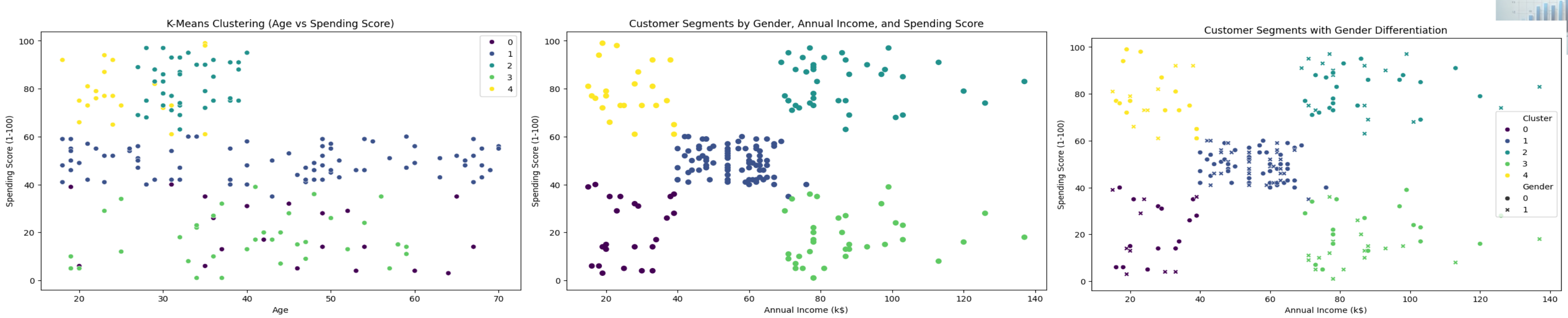
VISUALIZATIONS - RELATIONSHIPS



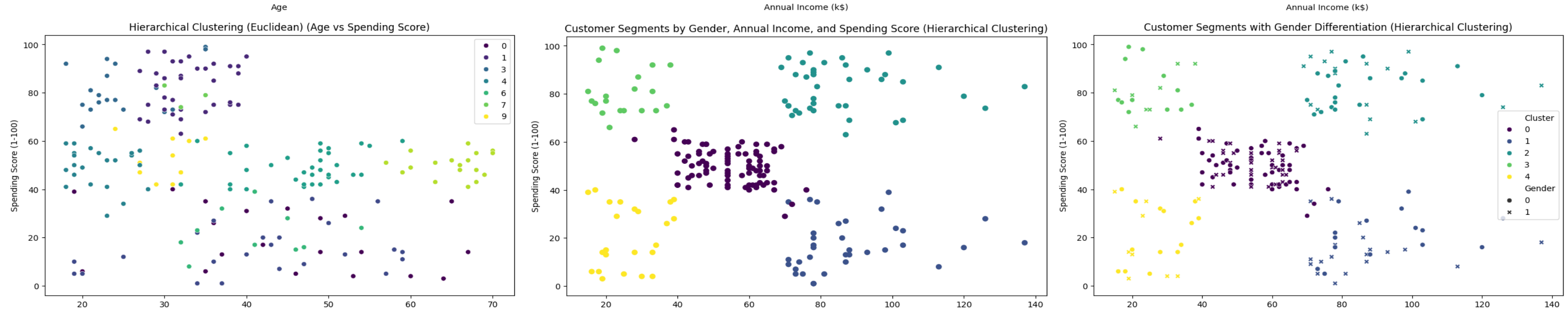
- No linear correlation between the variables considered in this study
- Clustering: it looks like the annual income vs Spending score will provide the best way to interpret the data
- # Clusters : well defined clusters with concentrations around 5 points in the two dimensional plots

MODELS FOR CLUSTERING

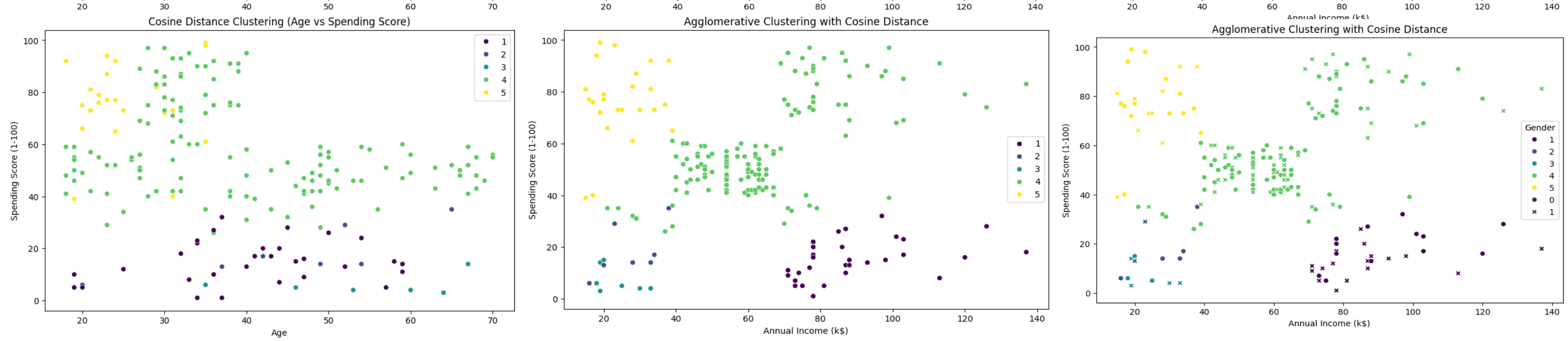
K-MEANS



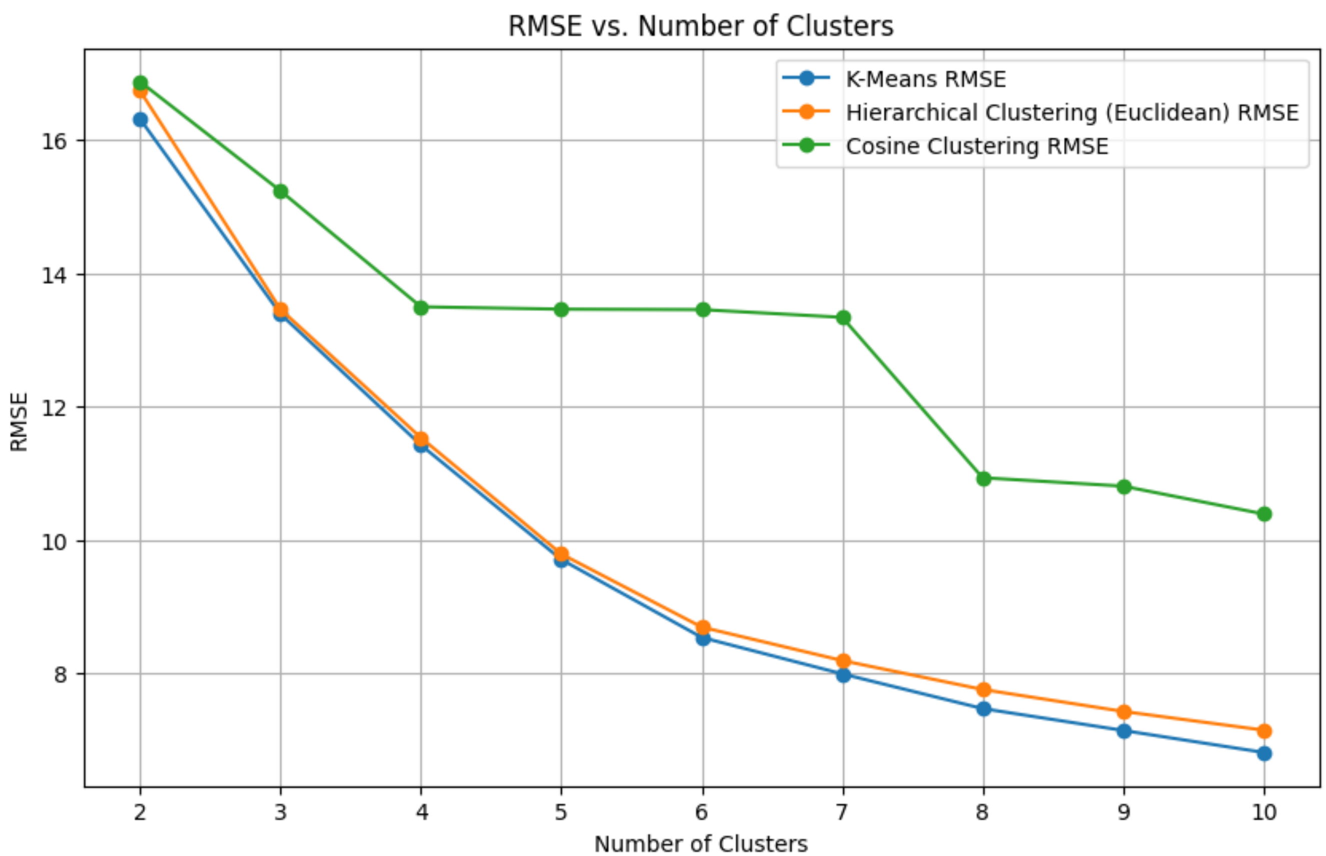
HIERARCHICAL



SIMILARITY
COSINE

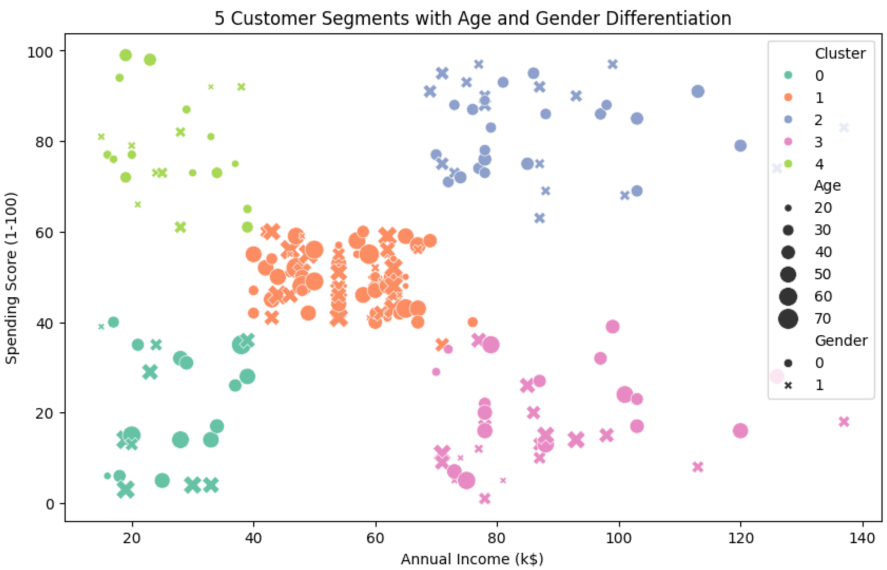


MODELS COMPARISON



- K-means is the best performing model. The cosine distance similarity does not seem to yield good results with the data analyzed in this project.
- Number of clusters: the optimal value seems to be 5 as the RMSE starts to decrease much less for greater values. 6 and 7 could also be considered.
- For 8 and greater values the RMSE decreases very slowly. Adding complexity not worthy

RESULTS DISCUSSION



5-clusters model

1. Older low earners, low spenders (dark green)
2. Younger low earners, high spenders (light green)
3. Middle earners, middle spenders (orange)
4. Older high earners, low spenders (pink)
5. Younger high earners, high spenders (purple)

Gender does not appear to play any role
Age emerges as a meaningful variable to distinguish spending

7-clusters model
1. Older low earners, low spenders (yellow)
2. Younger low earners, high spenders (light green)
3. Older predominantly female middle earners, middle spenders (orange)
4. Younger, predominantly male middle earners, middle spenders (dark green)
5. Older high earners, low spenders (pink)
6. Younger moderately high earners, high spenders (purple)
7. Younger, predominantly male highest earners, high spenders (beige)

Gender creates two segments in the middle earners group and in the split of high earners

CONCLUSIONS

1. 3 methods tested: K-Means, Hierarchical Clustering (Euclidean distance), and Cosine Distance Clustering. Through evaluation of RMSE, **K-Means model with 5 clusters** offers the most effective and actionable segmentation

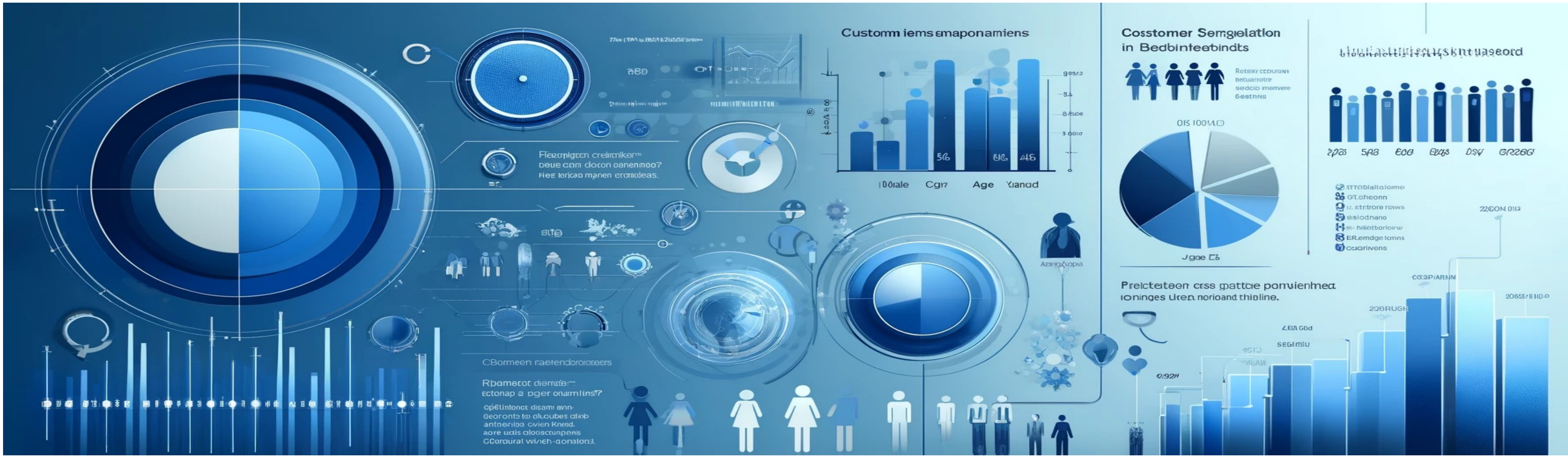
2. **Annual income and age** as the most influential variables in the segmentation process. Customers are naturally divided into distinct economic tiers, with age further refining the segments within these tiers. Notably, younger customers within the same income range tend to spend more than older customers.

3. Gender Analysis: The visualization of clusters considering gender indicated that **gender does not significantly influence the cluster formation**, only in the 7-cluster model there is some minor effect.

4. Comparison with the **7-Cluster Model**: introduced additional complexity by segmenting middle earners and high earners based on gender. This complexity **did not yield substantial new insights** or business value.

5. Final Recommendation : we recommend adopting the **5-cluster K-Means model for customer segmentation.** This model provides clear and actionable segments based on income and spending patterns, complemented by age as a secondary factor.

6. Further research : incorporating additional data, exploring advanced techniques, and validating the segmentation with real-world testing, the business can refine its understanding of the customer base and further enhance its marketing effectiveness.



END

FINAL PROJECT:

UNSUPERVISED ML MODEL TO SEGMENT CUSTOMERS BASED ON DEMOGRAPHICS AND SPENDING BEHAVIOUR