



# Customer Segmentation

01

# Background

An automobile company has plans to enter a new market with their 5 existing products. They've deduced that the behaviour of the new market is similar to their existing market where all customers have been classified into 4 segments: A, B, C, or D.

However, machine learning models may be able to segment customers more effectively, allowing for even more tailored marketing.



02

# Current market segments

Customer data includes the following features:

- ID
- Gender
- Marital status
- Age
- Graduate status
- Profession - categorised as healthcare, engineer, lawyer, entertainment, artist, executive, doctor, homemaker, or marketing
- Work experience - measured in years
- Spending score - categorised as low, average, or high
- Family size - including the customer
- Var 1 - an anonymised category

# Customer segments



A

Early 40s  
Works in arts or entertainment  
Graduate  
Married  
Spending score: Low  
Family size: 1-2



B

Late 40s  
Artist  
Graduate  
Married  
Spending score: Low to average  
Family size: 2



C

Late 50s  
Artist  
Graduate  
Married  
Spending score: Average  
Family size: 2+

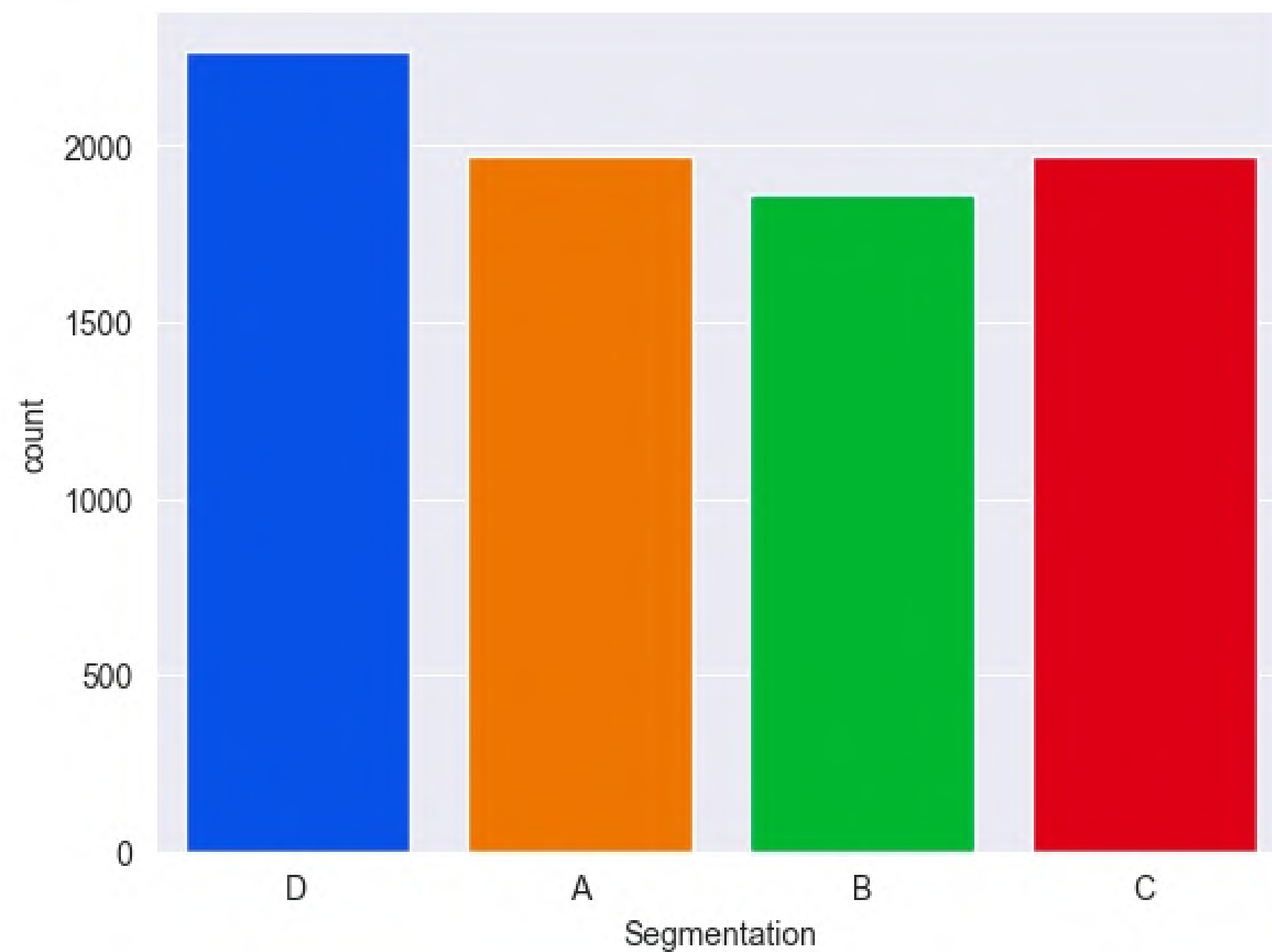


D

Early 20s  
Works in healthcare  
Non-graduate  
Unmarried  
Spending score: Low  
Family size: 3-4

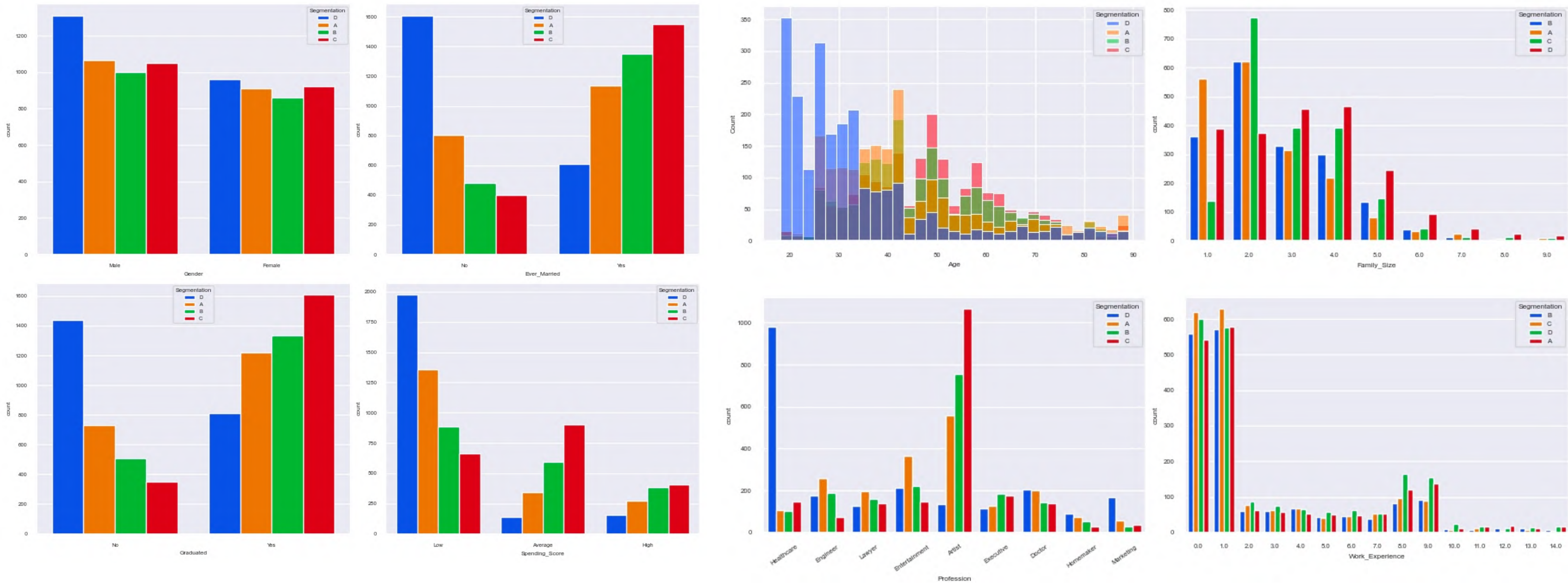
# Segment sizes

Analysing the training data,  
we see a balanced  
segmentation of customers



# Segment details

There's a lot of overlap in characteristics between each group



03

# A different way to segment

Due to the overlap between the segments, let's explore new ways of segmenting our target audience via machine learning.

Models used:

- K-Means Clustering
- Hierarchical Clustering

03

# But first, cleaning

Both datasets contain nulls, object variables, and more to deal with before our models can use the data.



# Steps

1

Turn categorical values to numerical values

Method:  
Value mapping

2

Scaling all values

Method:  
MinMax scaling

3

Filling null values

Method:  
KNNImputer

# Features we're using

All variables are ready for the machines

- Gender
- Marital status
- Age
- Graduate status
- Profession
- Work experience
- Spending score
- Family size
- Var 1

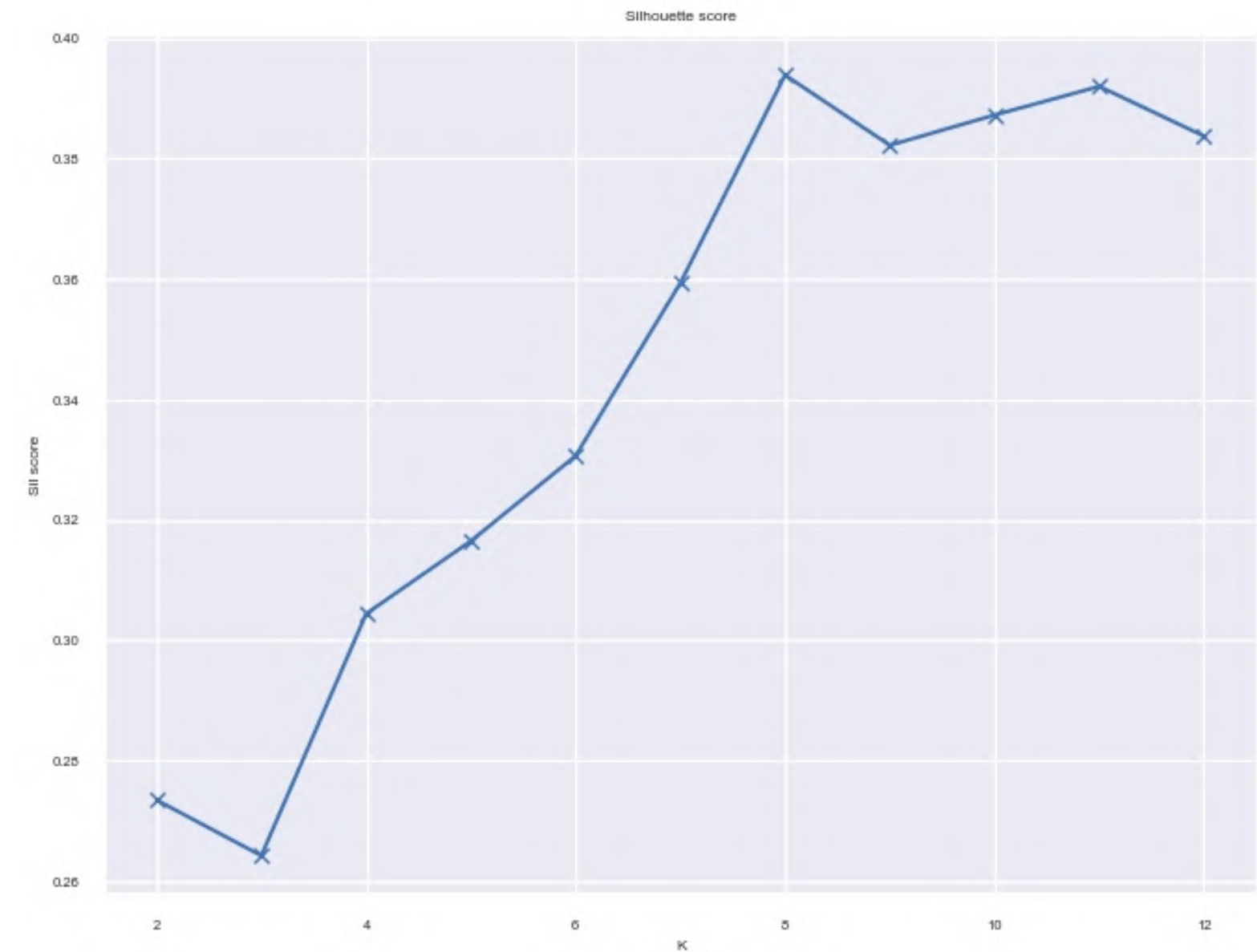
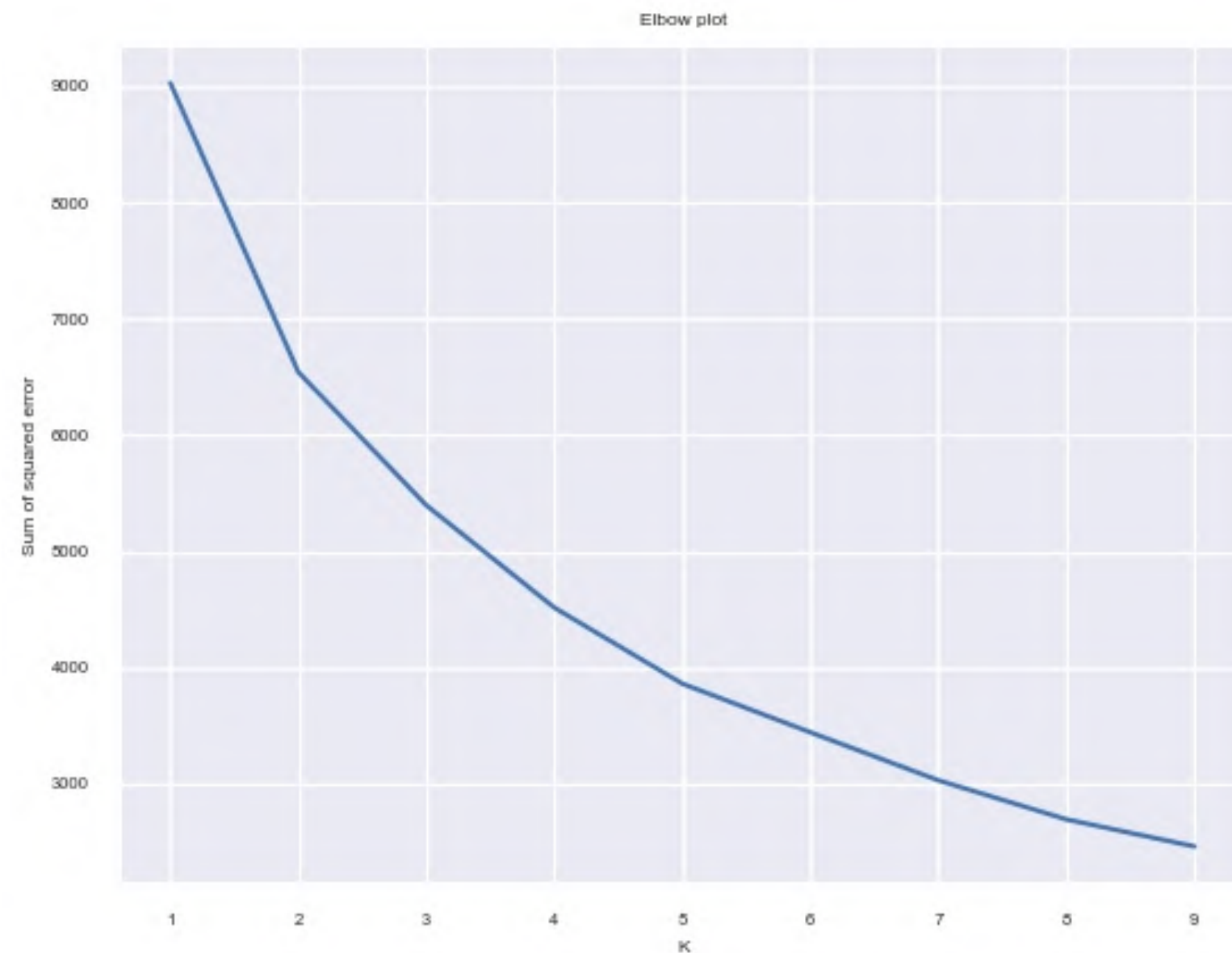
03

# Determining optimum segmentation

Multiple methods were used to find the ideal number of clusters — the segments containing the most similar customers in terms of characteristics.

# K-Means cluster selection

The elbow plot was unclear but the silhouette score peaked at 8.

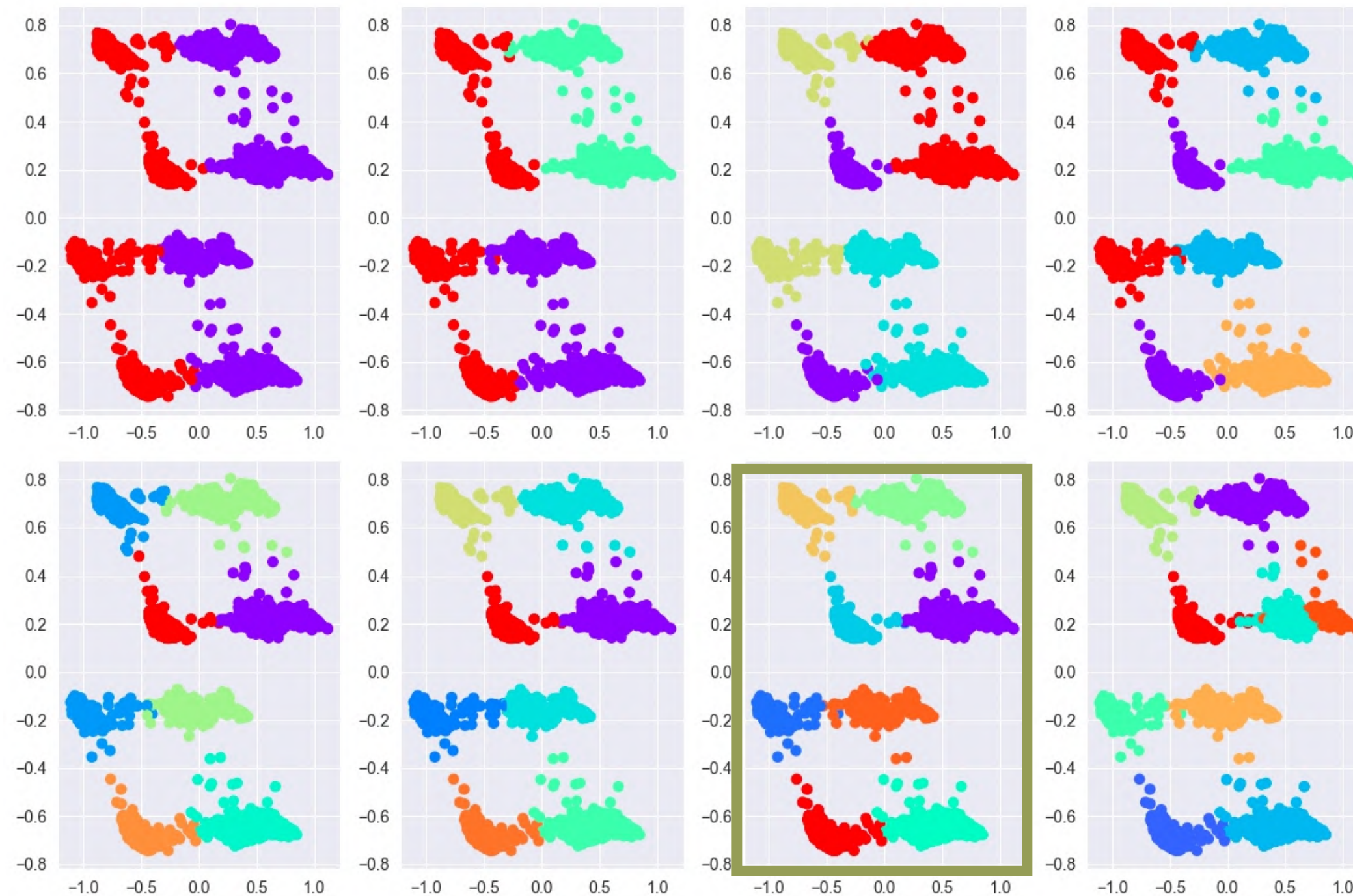




# K-Means cluster selection

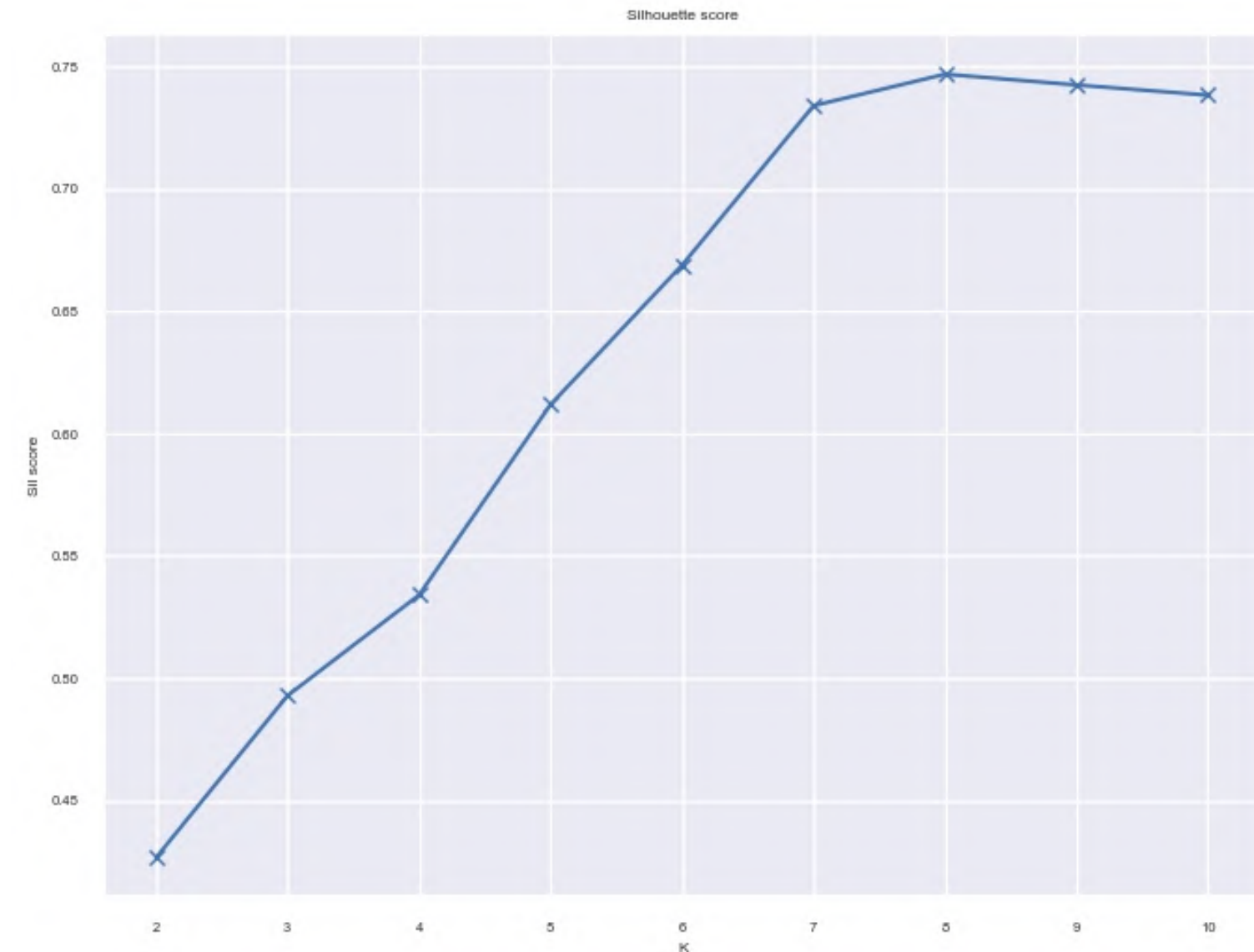
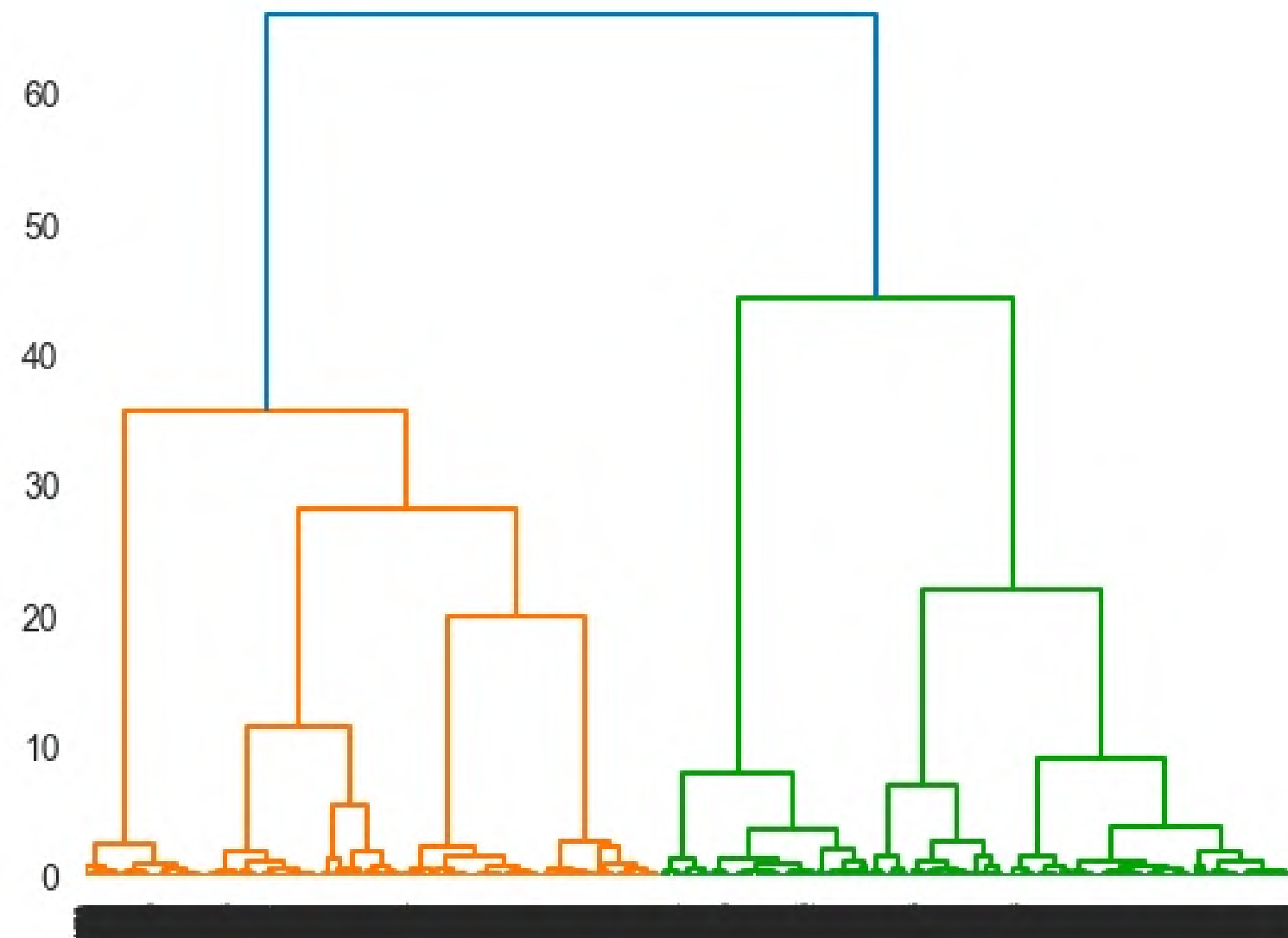
8 also looked like the optimum number from the scatter plots.

**Clusters selected: 8**



# Hierarchical cluster selection

The dendrogram showed an optimal 2 clusters while the silhouette score peaked at 8.

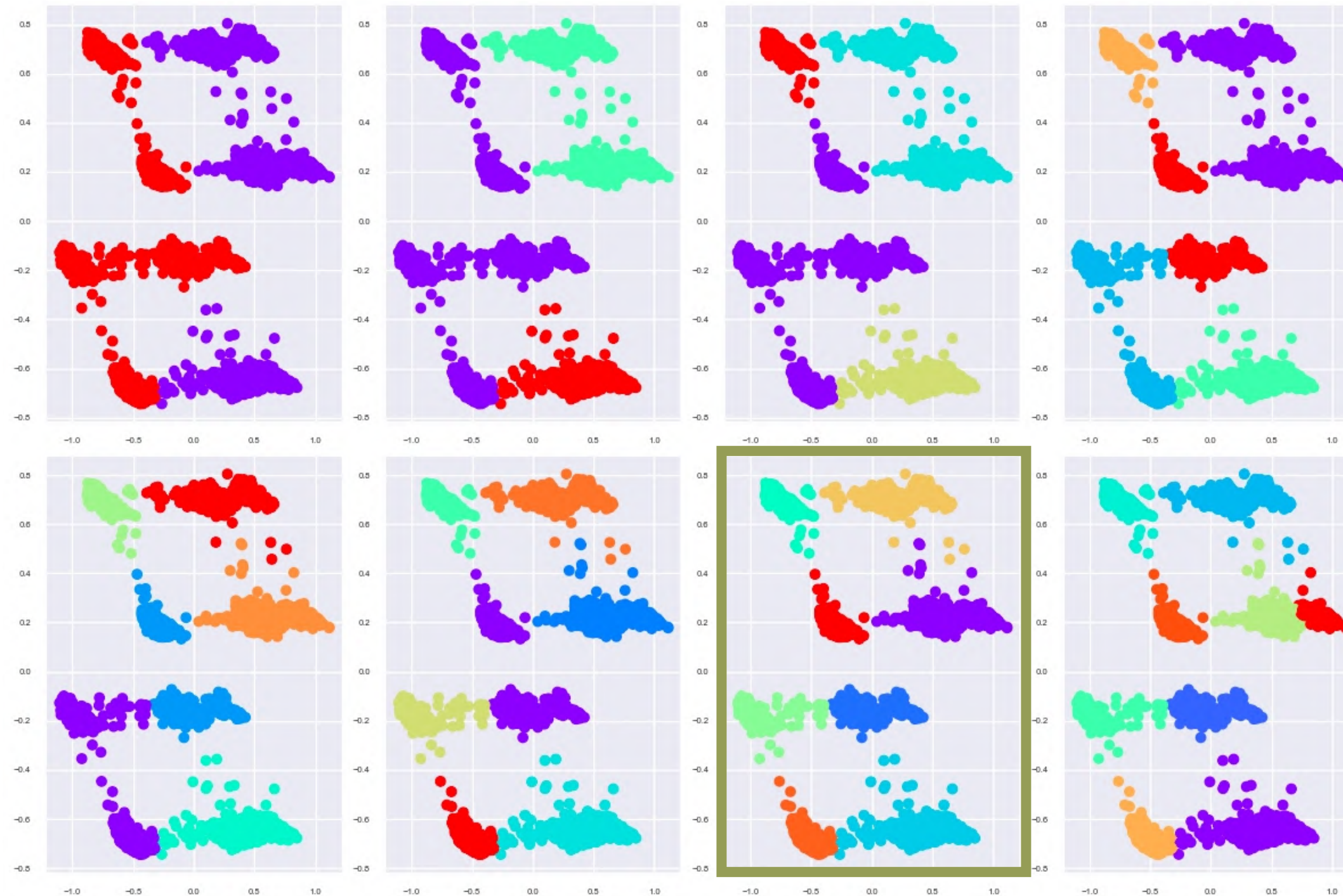




# Hierarchical cluster selection

Again, 8 also looked like the optimum number from the scatter plots.

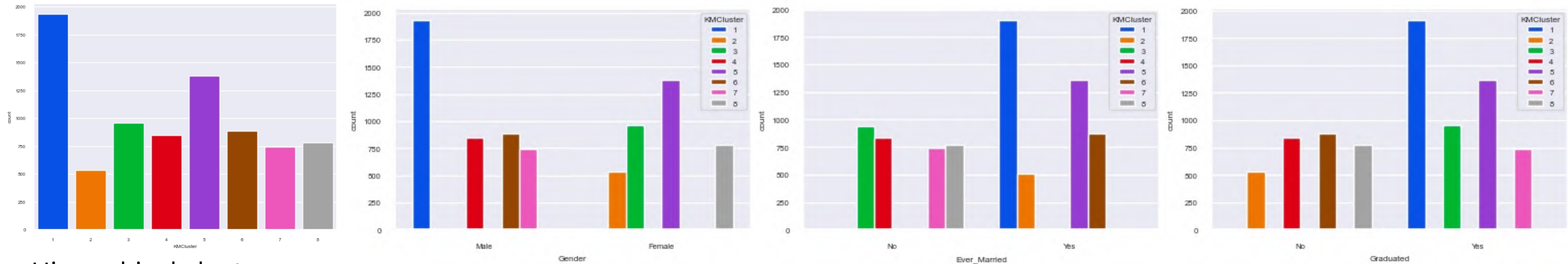
**Clusters selected: 8**



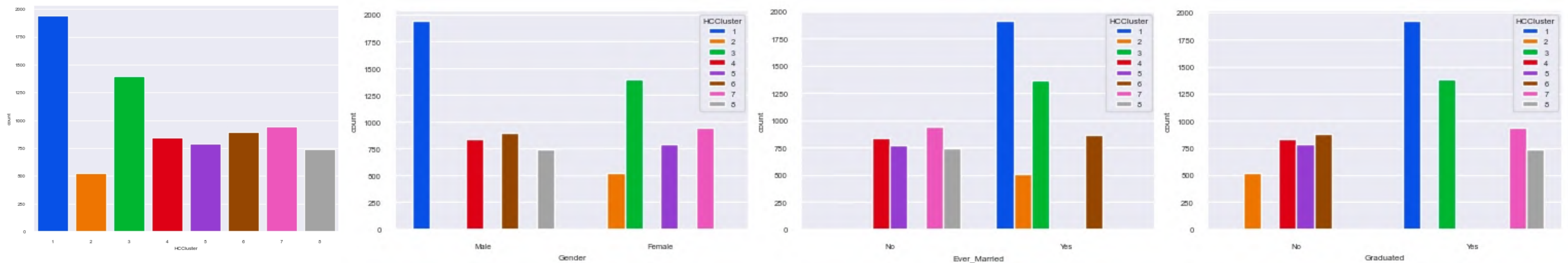
# Model output

Both models produced near-identical results, just with different labels

## K-Means clusters



## Hierarchical clusters













04

# Our new customer segments

Taking the output of both models, we can categorise and define the 8 segments amongst our customers.

# New customer segments

<div>A</div> <div><ul style="list-style-type: none"><li>• Largest segment</li><li>• 50s</li><li>• Male</li><li>• Graduate</li><li>• Primarily artists</li><li>• Married</li><li>• Family size: 2</li><li>• Spending score: Average, but falls into other categories</li></ul></div> <div></div>	<div>B</div> <div><ul style="list-style-type: none"><li>• 50s</li><li>• Female</li><li>• Graduates</li><li>• Primarily artists</li><li>• Married</li><li>• Family size: 2</li><li>• Spending score: Average, but falls into other categories</li></ul></div> <div></div>	<div>C</div> <div><ul style="list-style-type: none"><li>• Smallest segment</li><li>• 25-50</li><li>• Female</li><li>• Non-graduates</li><li>• Primarily engineers and lawyers</li><li>• Married</li><li>• Family size: 2</li><li>• Spending score: Average, but falls into other categories</li></ul></div> <div></div>	<div>D</div> <div><ul style="list-style-type: none"><li>• 50s</li><li>• Male</li><li>• Non-graduates</li><li>• Primarily executives</li><li>• Married</li><li>• Family size: 2+</li><li>• Spending score: falls evenly into each category</li></ul></div> <div></div>
<div>E</div> <div><ul style="list-style-type: none"><li>• 30s</li><li>• Male</li><li>• Graduates</li><li>• Primarily work in arts, healthcare, and entertainment</li><li>• Unmarried</li><li>• Family size: 1-3</li><li>• Spending score: Low</li></ul></div> <div></div>	<div>F</div> <div><ul style="list-style-type: none"><li>• 30s</li><li>• Female</li><li>• Graduates</li><li>• Primarily work in arts or healthcare</li><li>• Unmarried</li><li>• Family size: 1</li><li>• Spending score: Low</li></ul></div> <div></div>	<div>G</div> <div><ul style="list-style-type: none"><li>• 20s</li><li>• Female</li><li>• Non-graduates</li><li>• Primarily work in healthcare</li><li>• Unmarried</li><li>• Family size: 3-4</li><li>• Spending score: Low</li></ul></div> <div></div>	<div>H</div> <div><ul style="list-style-type: none"><li>• 20s</li><li>• Male</li><li>• Non-graduates</li><li>• Primarily work in healthcare</li><li>• Unmarried</li><li>• Family size: 4</li><li>• Spending score: Low</li></ul></div> <div></div>

05

# Conclusion and recommendations

8 clusters are considered the optimum to split these markets into more defined segments, allowing for more tailored marketing and outreach.

The models performed well at separating binary categories such as gender, marital status and graduate status. They were also good at identifying segments with a low spending score.

However, the segments can be joined together by age, gender, spending score etc. based on company needs and strategy.



