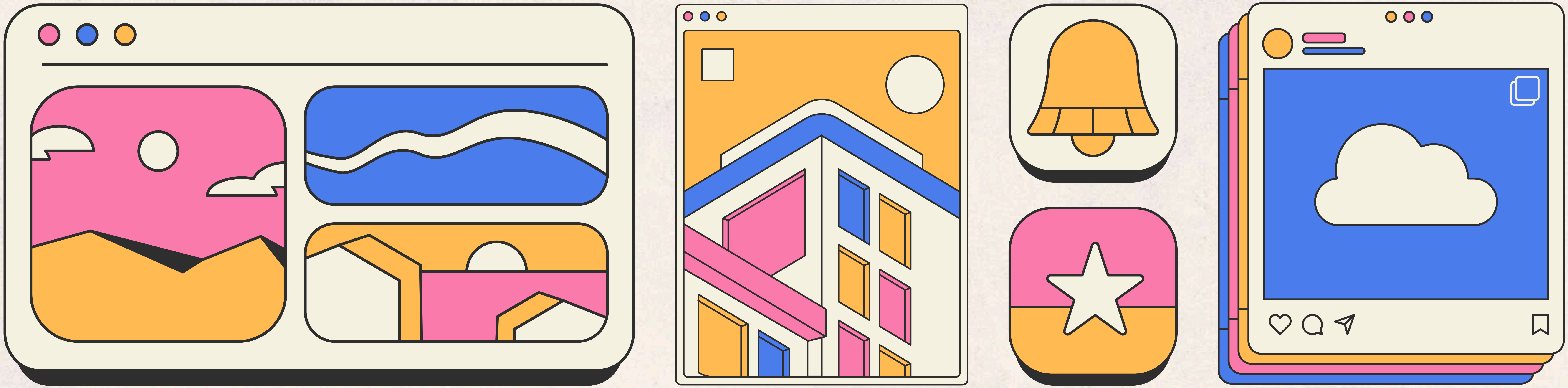
Clickstream Data

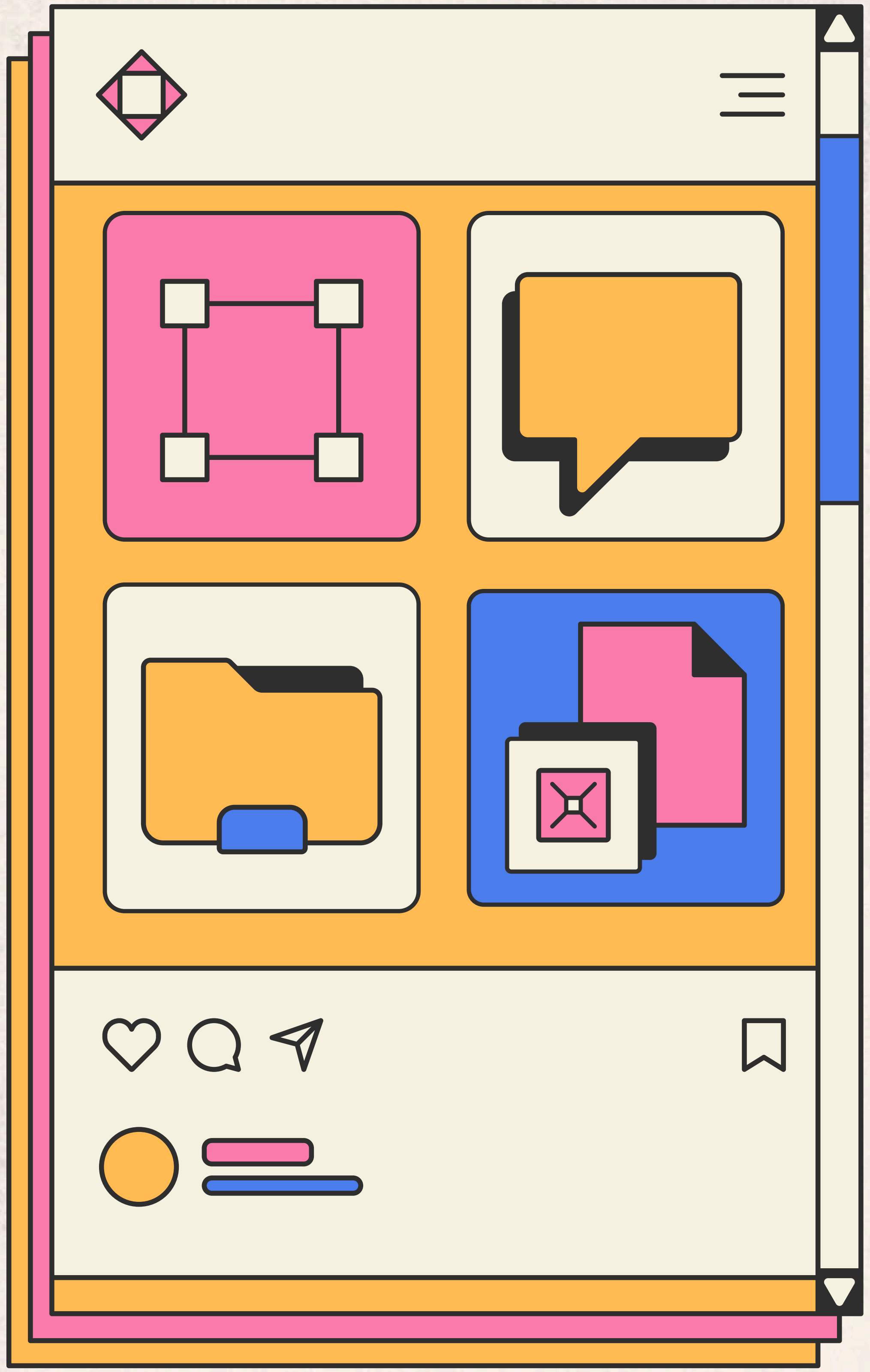
CSCI 113i - Final Project



Group 5 - Cheng, Montemayor, Shu Too

Report Outline

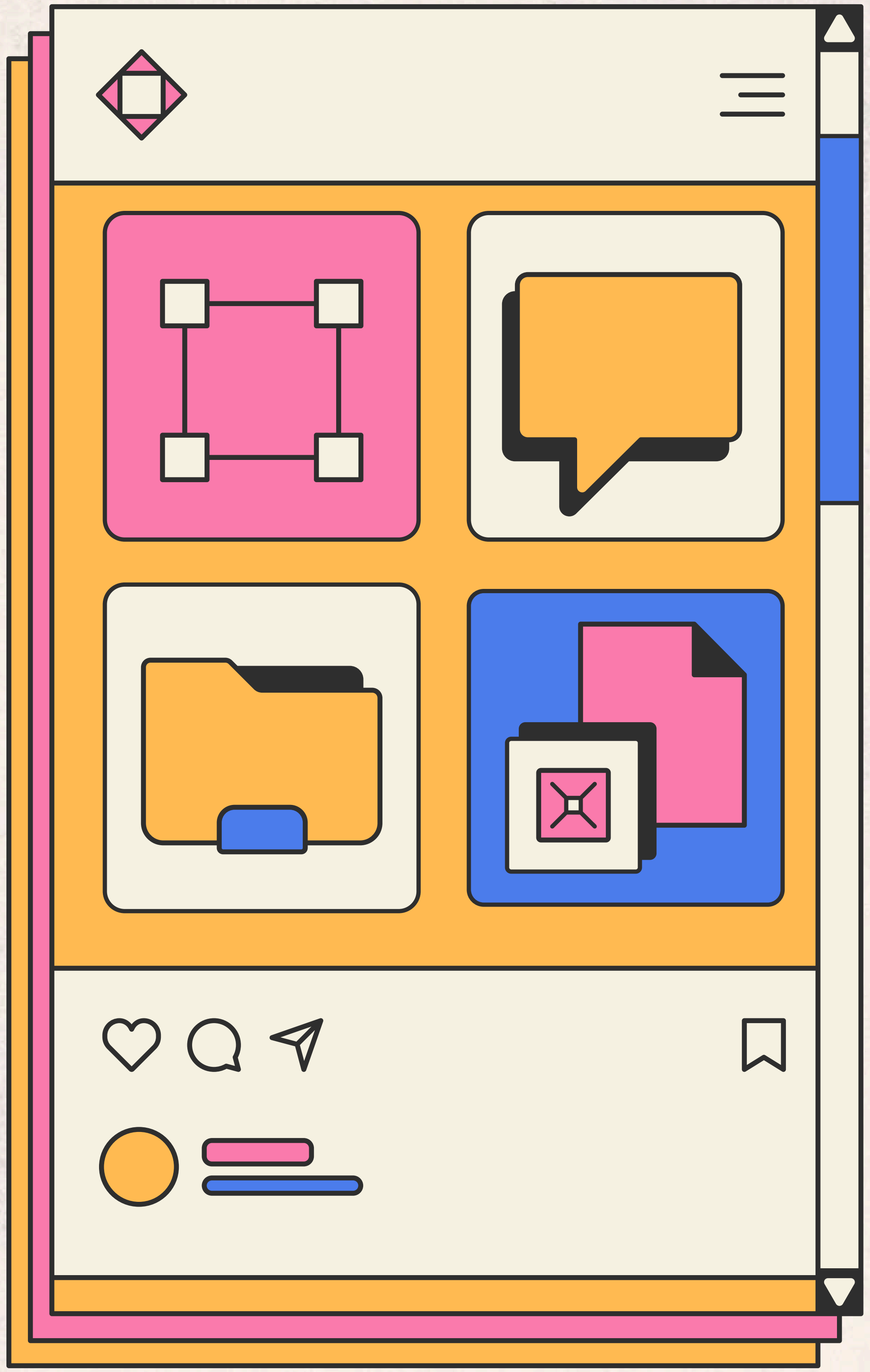
- **Introduction of Dataset and Problem Statement**
- **Dataset Preprocessing and EDA**
- **Machine Learning Models**
- **Results**



Overview & Methodology

Clickstream Data for Online Shopping

- Retrieved from the **UCI Machine Learning Repository**
- Log files from an **e-shop** in **Poland** offering clothing for pregnant women from **April to August 2008**
  - Spring & summer collections in the ff. categories: skirts & dresses, trousers, blouses, special offers
  - Local and international shipping
- Łapczyński & Białowąs 2013: Association rule mining and sequence analysis on blouses and tunics
- **165,474 instances with 14 features**
  - Instance: one clothing purchase



Problem Statement

Business Objective

- To create and implement **product placement strategies** (in terms of page number and location) ***for the international customers*** of the online store in order to increase overall profitability and revenue

Data Mining Objective

- To perform **unsupervised machine learning (clustering)** on the dataset in order to understand and analyze the behaviors of customers

Data Preprocessing

Column renaming

	year	month	day	order	country	session ID	page 1 (main category)	page 2 (clothing model)	colour	location	model photography	price	price 2	page
0	2008	4	1	1	29	1	1	A13	1	5	1	28	2	1
1	2008	4	1	2	29	1	1	A16	1	6	1	33	2	1
2	2008	4	1	3	29	1	2	B4	10	2	1	52	1	1
3	2008	4	1	4	29	1	2	B17	6	6	2	38	2	1
4	2008	4	1	5	29	1	2	B8	4	3	2	52	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
165469	2008	8	13	1	29	24024	2	B10	2	4	1	67	1	1
165470	2008	8	13	1	9	24025	1	A11	3	4	1	62	1	1
165471	2008	8	13	1	34	24026	1	A2	3	1	1	43	2	1
165472	2008	8	13	2	34	24026	3	C2	12	1	1	43	1	1
165473	2008	8	13	3	34	24026	2	B2	3	1	2	57	1	1

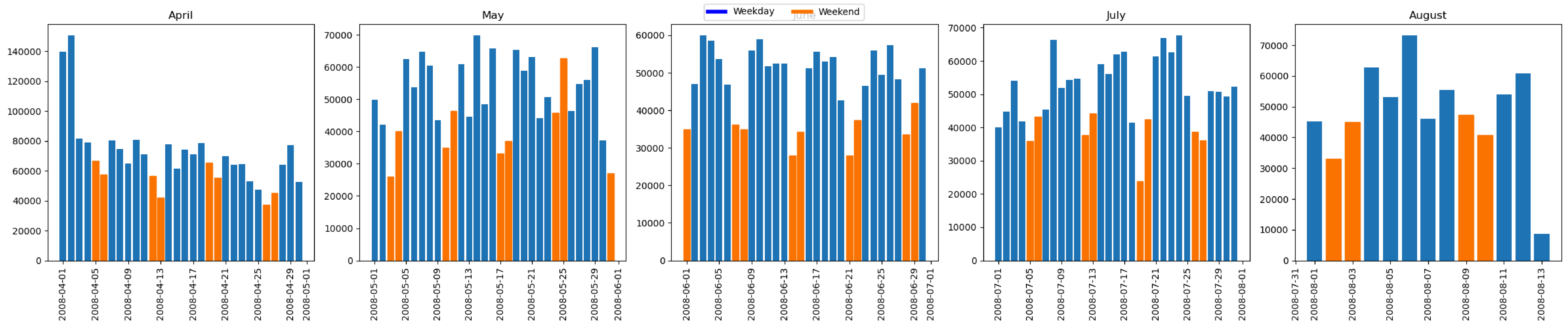
Data Preprocessing

Column dropping

	year	month	day	order	country	session ID	page 1 (main category)	page 2 (clothing model)	colour	location	model photography	price	price 2	page
0	2008	4	1	1	29	1	1	A13	1	5	1	28	2	1
1	2008	4	1	2	29	1	1	A16	1	6	1	33	2	1
2	2008	4	1	3	29	1	2	B4	10	2	1	52	1	1
3	2008	4	1	4	29	1	2	B17	6	6	2	38	2	1
4	2008	4	1	5	29	1	2	B8	4	3	2	52	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
165469	2008	8	13	1	29	24024	2	B10	2	4	1	67	1	1
165470	2008	8	13	1	9	24025	1	A11	3	4	1	62	1	1
165471	2008	8	13	1	34	24026	1	A2	3	1	1	43	2	1
165472	2008	8	13	2	34	24026	3	C2	12	1	1	43	1	1
165473	2008	8	13	3	34	24026	2	B2	3	1	2	57	1	1

Data Preprocessing

Year, Month, Day → Date → Weekday



Data Pre-Processing

Reduction of Categories

01

Countries to Regions (EuroVoc)

- ✗ Poland
- ✗ Region 6

02

Color Types

- Light neutrals
- Dark neutrals
- Light colored
- Dark colored
- Multi-colored

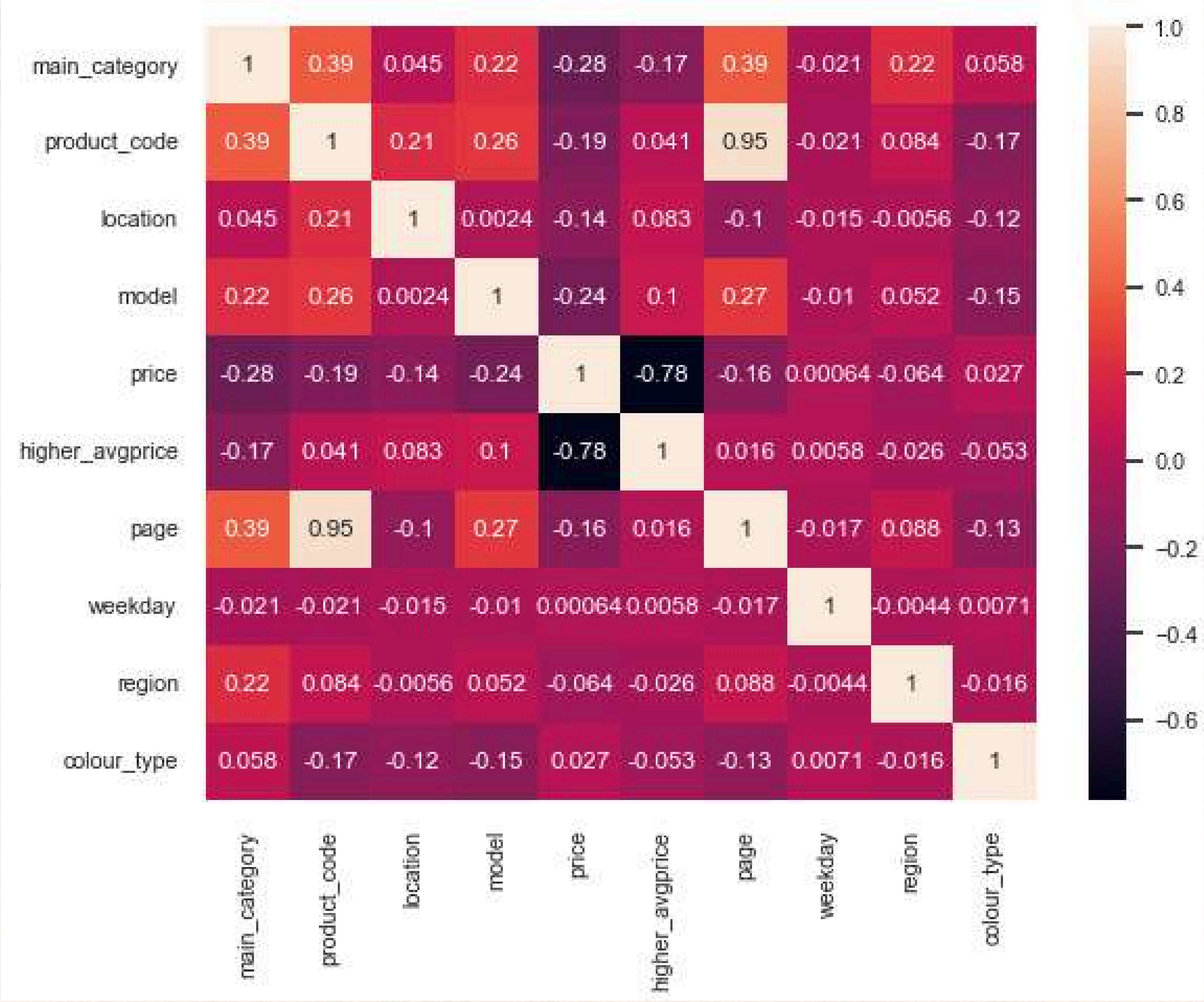
Data Preprocessing

Pre-Processing

- Product code: integer

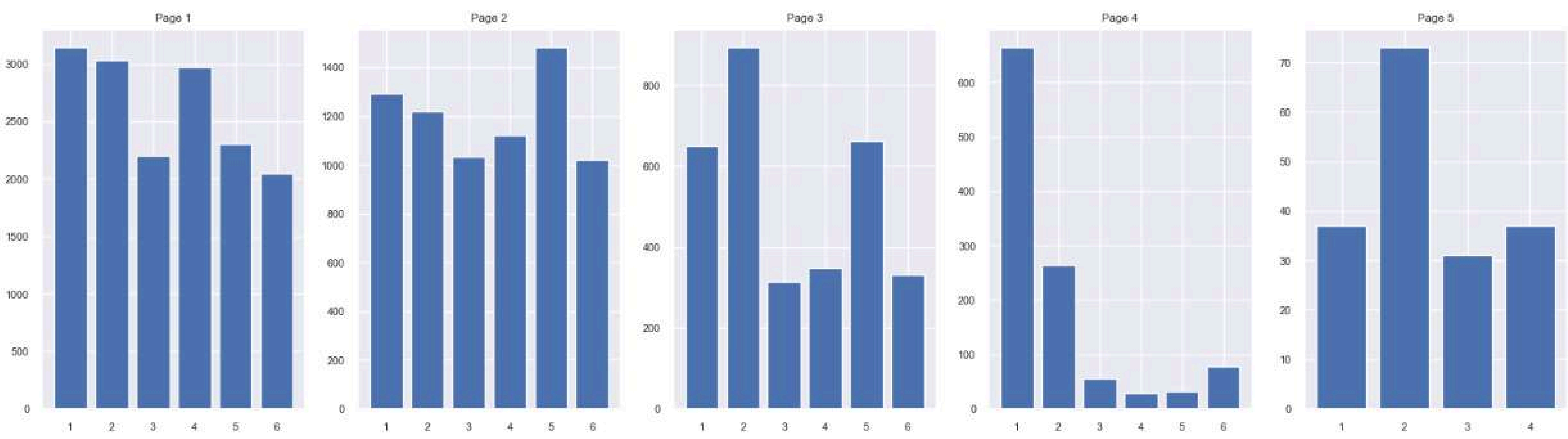
High Correlations

- ***product_code*** is highly correlated with the page (0.95).
- ***higher_avgprice*** is highly correlated with price (-0.78).



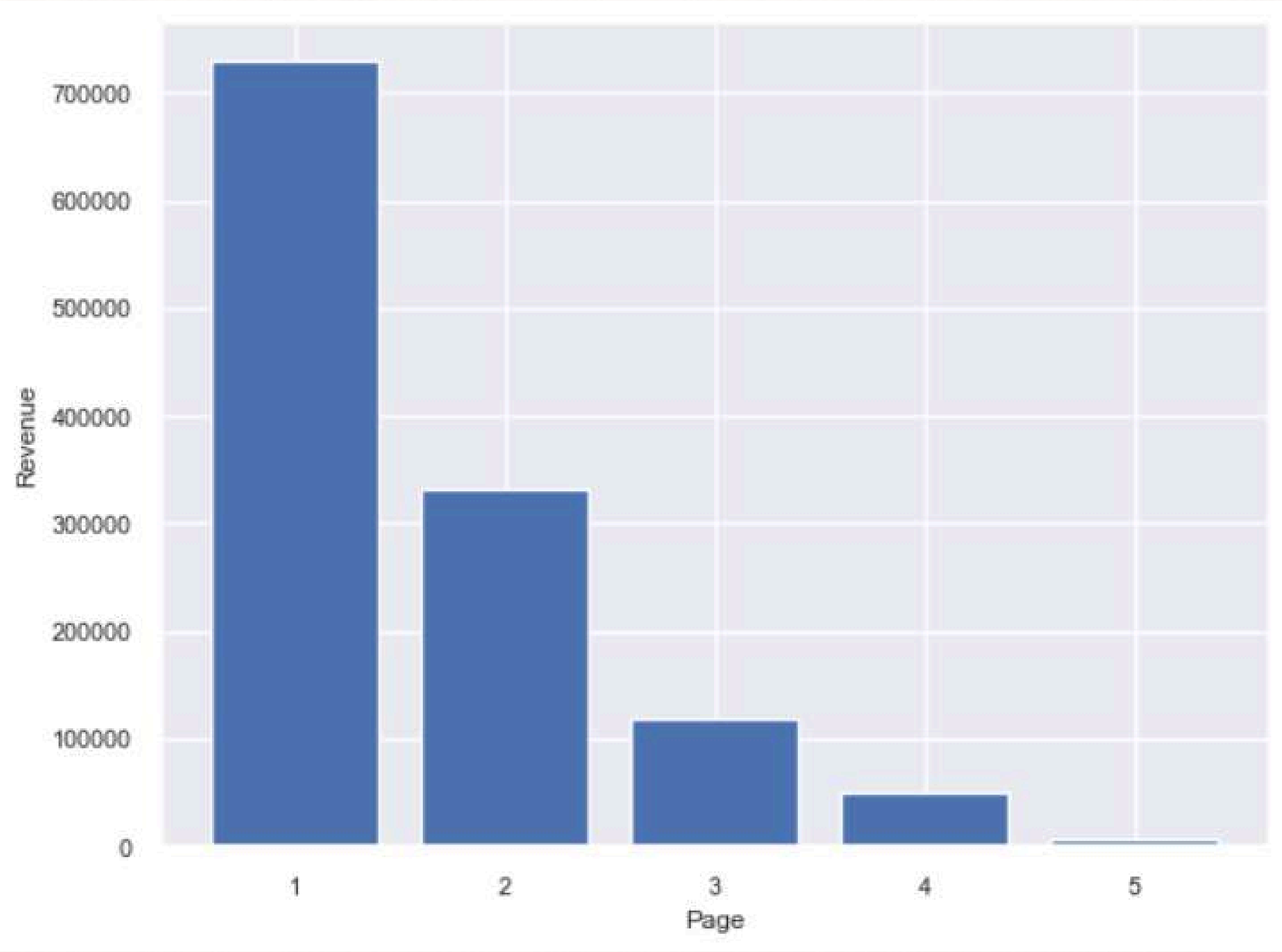
Initial EDA Results

Most Viewed Portion/Location of Each Page



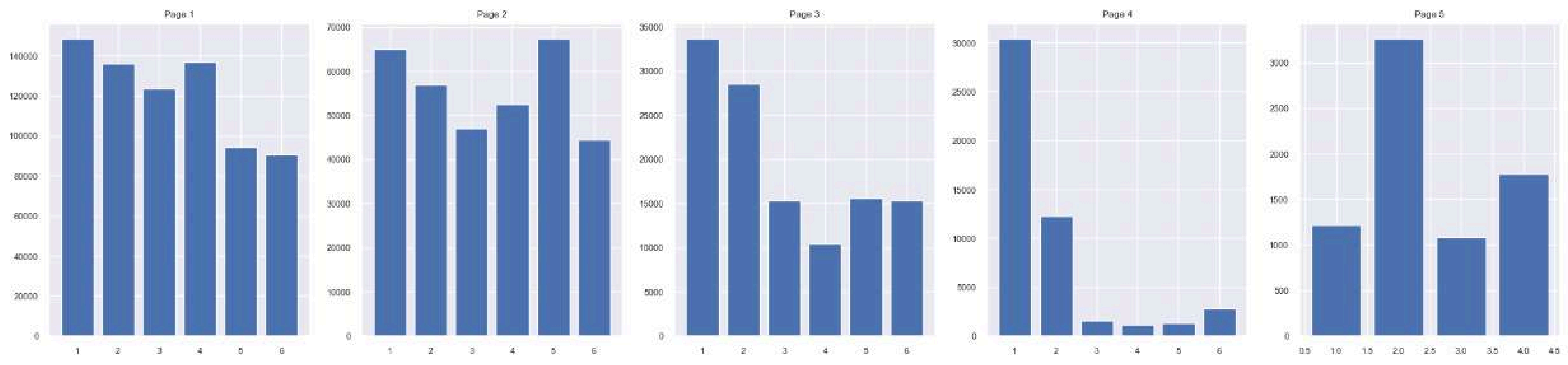
Initial EDA Results

Revenue per page



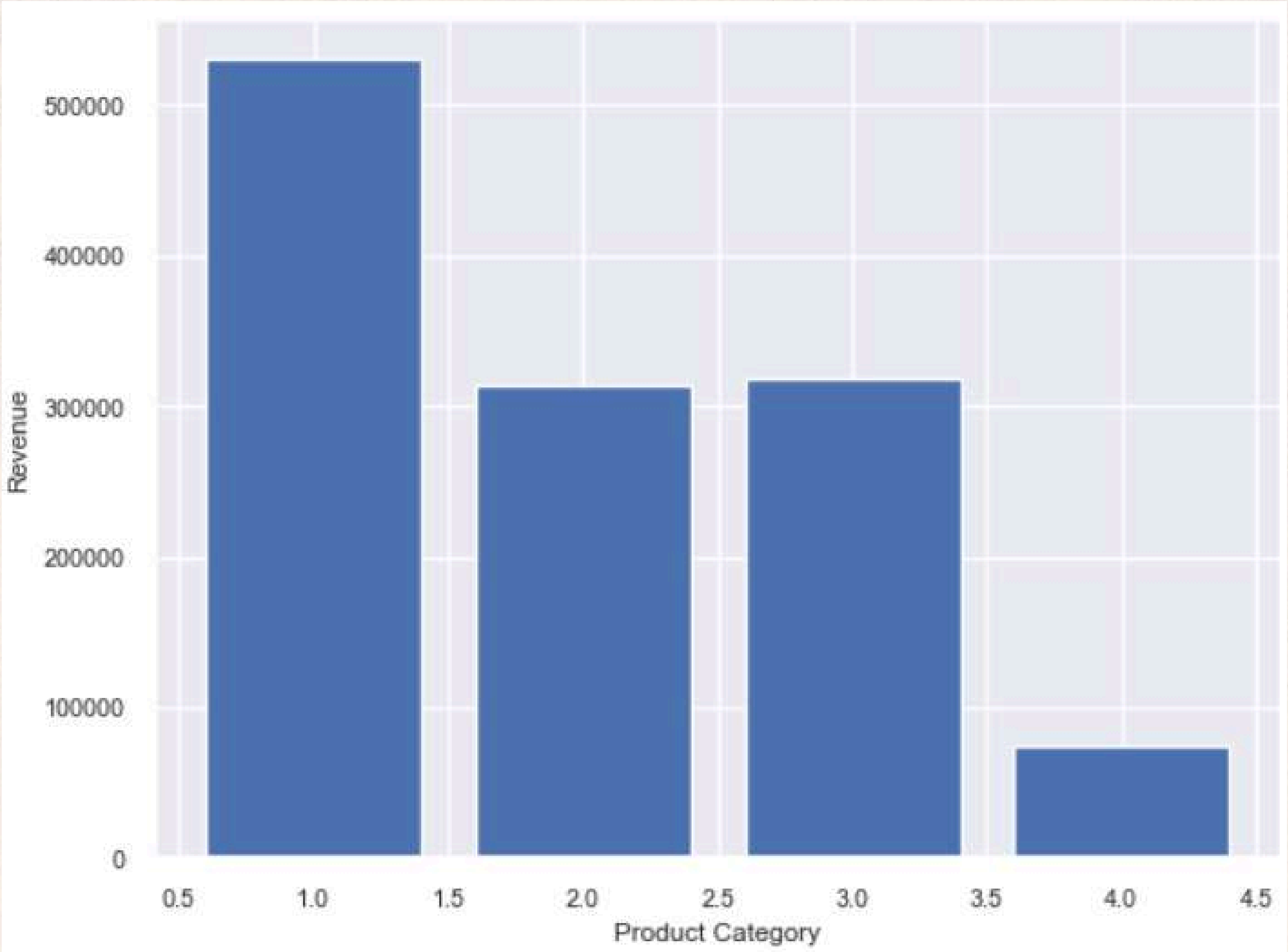
Initial EDA Results

Revenue per location per page



Initial EDA Results

Revenue per Main Category



Balancing the Dataset

Undersampling

	main_category	location	model	price	page	weekday	region	colour_type
19	2	6	2	38	1	1	4	1
20	3	2	1	48	1	1	4	2
21	3	3	1	48	1	1	4	4
22	3	4	2	28	1	1	4	5
23	3	6	1	48	1	1	4	1
...	...	...	...	...	...	...	...	...
165221	3	1	2	33	4	1	1	4
165470	1	4	1	62	1	1	1	3
165471	1	1	1	43	1	1	1	3
165472	3	1	1	43	1	1	1	3
165473	2	1	2	57	1	1	1	3

Balancing the Dataset (Undersampling)

- Idea: When there is class imbalance, **the model may be biased towards the majority class**, so you take a **random sample** of it to reduce this.
- Done using *RandomUnderSampler* under the *imbalanced-learn* library
  - Usually done in conjunction with **Synthetic Minority Oversampling Technique (SMOTE)**, which draws new samples under the minority class
- For this project, we decided to test the ff.:
  - Balancing Techniques: SMOTE, Undersampling, Combined
  - Target Variables: Location, Region
- Generally, **the balanced dataset performed similarly** in terms of silhouette scores in comparison to the imbalanced dataset.
- The dataset with an **under-sampled location column** worked best, but by a small margin. This shows that we can obtain similar results despite having fewer data points to analyze. In this manner, the model may run more efficiently.

Data Preprocessing

Z-score scaling

	main_category	location	model	price	page	weekday	region	colour_type
19	2	6	2	38	1	1	4	1
20	3	2	1	48	1	1	4	2
21	3	3	1	48	1	1	4	4
22	3	4	2	28	1	1	4	5
23	3	6	1	48	1	1	4	1
...	...	...	...	...	...	...	...	...
165221	3	1	2	33	4	1	1	4
165470	1	4	1	62	1	1	1	3
165471	1	1	1	43	1	1	1	3
165472	3	1	1	43	1	1	1	3
165473	2	1	2	57	1	1	1	3

Data Preprocessing

One-hot encoding

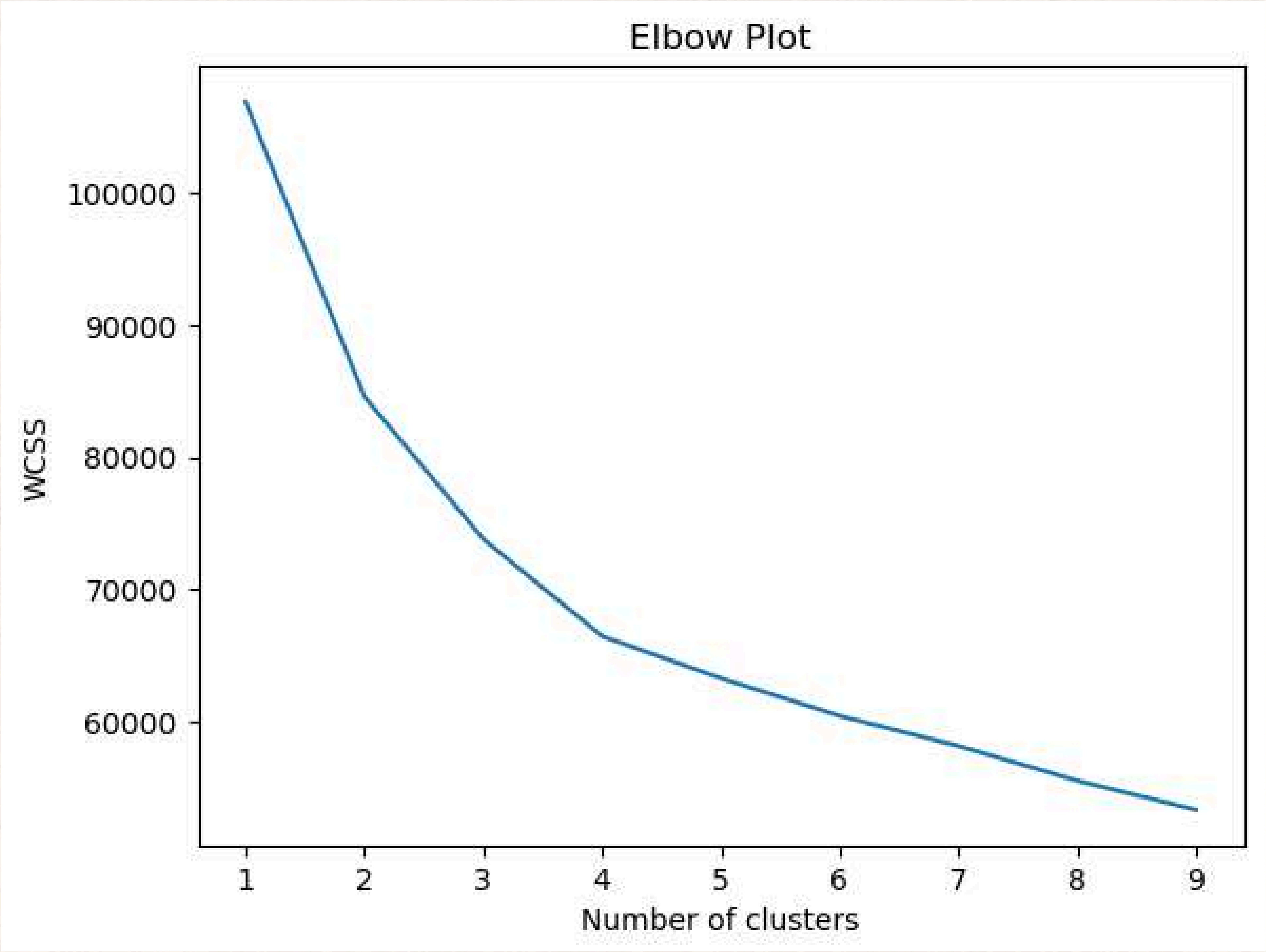
	main_category	location	model	price	page	weekday	region	colour_type
19	2	6	2	38	1	1	4	1
20	3	2	1	48	1	1	4	2
21	3	3	1	48	1	1	4	4
22	3	4	2	28	1	1	4	5
23	3	6	1	48	1	1	4	1
...	...	...	...	...	...	...	...	...
165221	3	1	2	33	4	1	1	4
165470	1	4	1	62	1	1	1	3
165471	1	1	1	43	1	1	1	3
165472	3	1	1	43	1	1	1	3
165473	2	1	2	57	1	1	1	3

Machine Learning (K-Means: Algorithm)

- Idea: The *n* data points are considered points in *d*-dimensional space.
- Process: The value of *k* (**number of clusters**) is a hyperparameter selected by the user.
  - *k* random points are selected, called **centroids**.
  - For each of the *n* data points, we pick the closest centroid using a distance metric.
  - Take the average of the points in each of the *k* clusters.
  - Repeat this until points do not change clusters.
- How to select *k*?
  - **Elbow Method**
  - **Silhouette Scores**

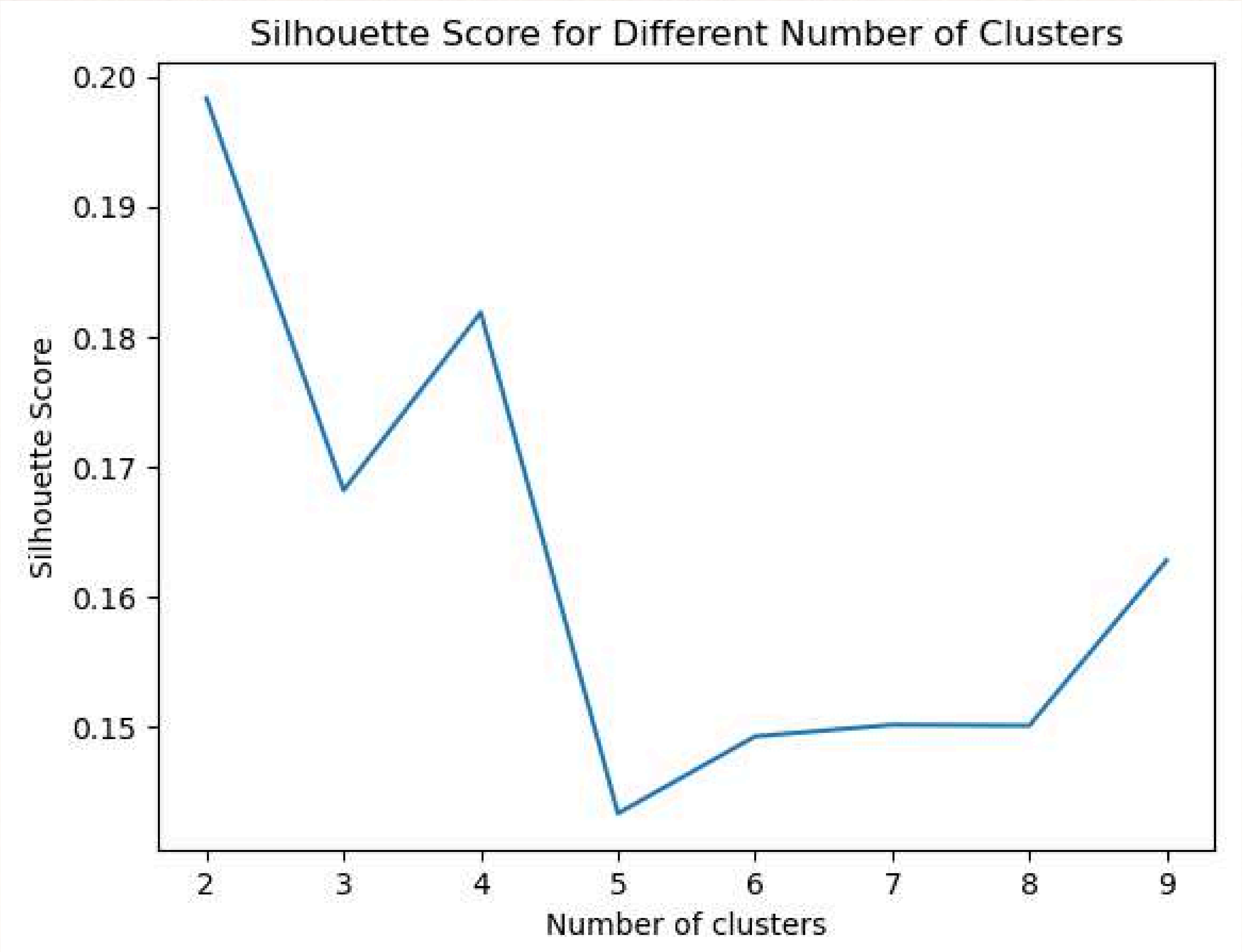
Machine Learning (K-Means: Evaluation)

- The elbow plot plots the **WCSS (within-cluster sum of squares)** which measures the sum of the squared distances between the data points in one cluster (i.e. how similar they are).
- We pick the number of clusters **at the “elbow” point of the plot** (i.e. when it suddenly “turns”).
- By the elbow plot, the optimal number of clusters is 4.



Machine Learning (K-Means: Evaluation)

- The silhouette score takes into account two things:
  - **Similarity of data points within one cluster**
  - **Dissimilarity of data points between multiple clusters**
- It takes on a value between -1 and 1, and **a higher score implies a better *k***.
- The plot suggests having two clusters, however, that's quite few for customer segmentation. We decided to choose 4.

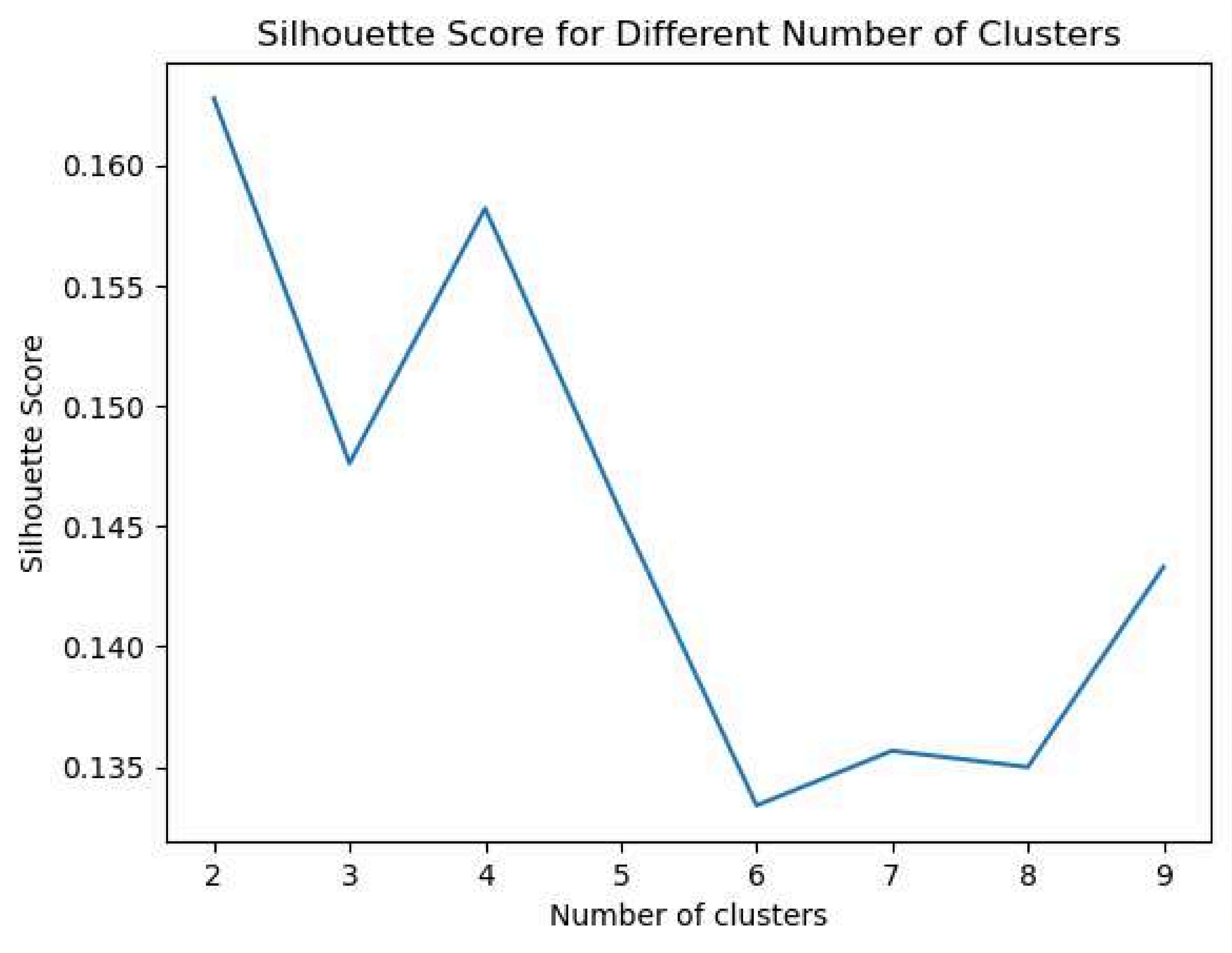


ML (Hierarchical Clustering: Algorithm)

- Process: The value of k (number of clusters) is a hyperparameter selected by the user.
  - Each of the n datapoints is initially considered its own cluster.
  - **Bottom-up approach**: the process is **agglomerative**.
  - Measure the distances between each pair of clusters, and merge the pair with the smallest distance.
  - Repeat until we get the desired number of clusters.
- How to select k? **Silhouette Scores**

ML (Hierarchical Clustering: Evaluation)

- The silhouette score follows the same definition from the previous discussion on K-Means.
- By similar reasoning, we chose 4 clusters also for this one.
    - It also allows **comparison between the two models**.

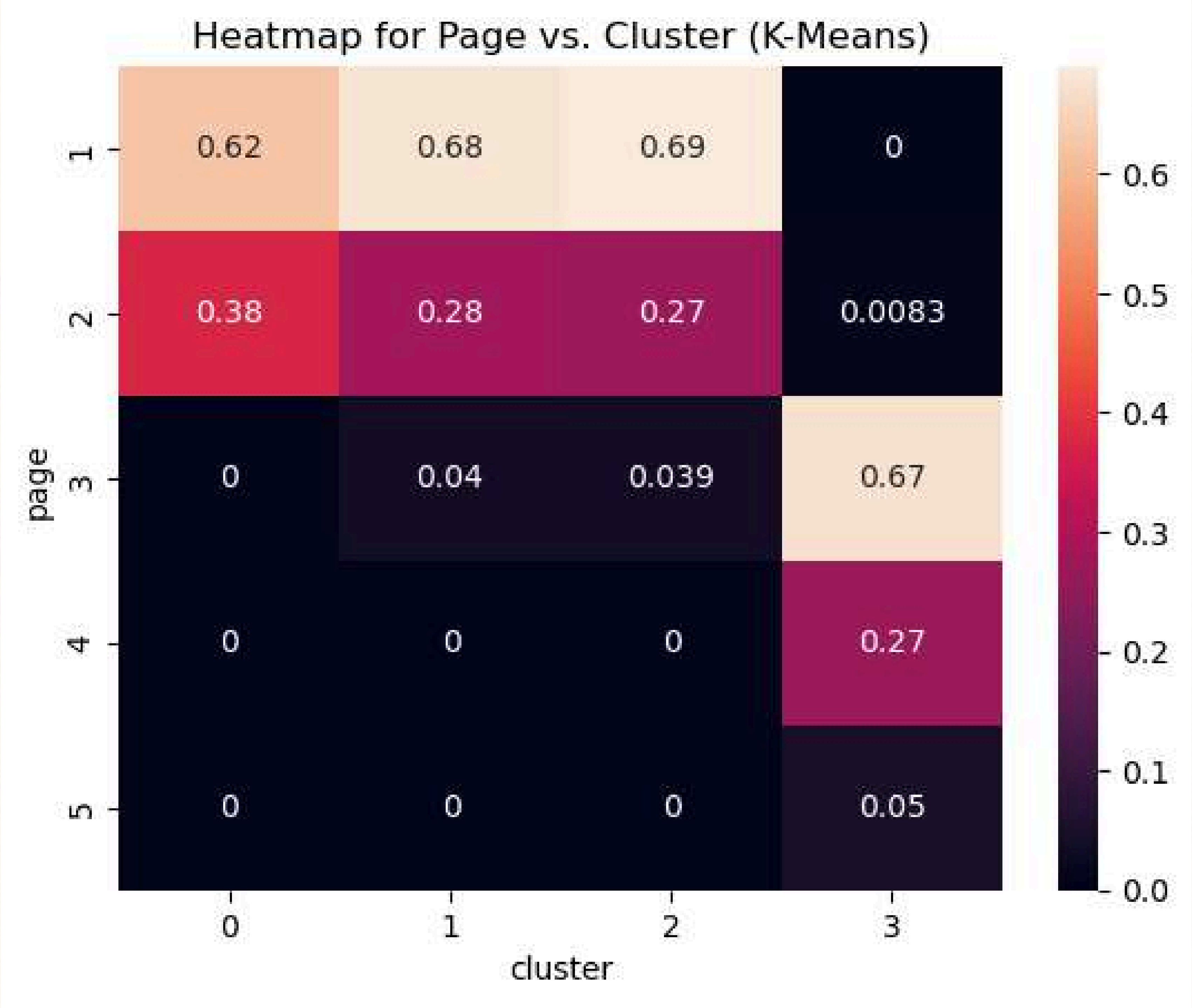


Results

- **Cluster vs. Categorical Variable**
  - Categorical variables include page, main category, region, location, weekday
  - **Cross-tabulation** was done, followed by **normalization** over the cluster variable.
    - This means that summing the numbers for one cluster results to 1.
    - Exception: Region, since the distribution over clusters is quite similar
- Cluster vs. Numerical Variable
  - Price is the only numerical variable: **Boxplots** were made to visualize data
- These were done for both algorithms.

Observations

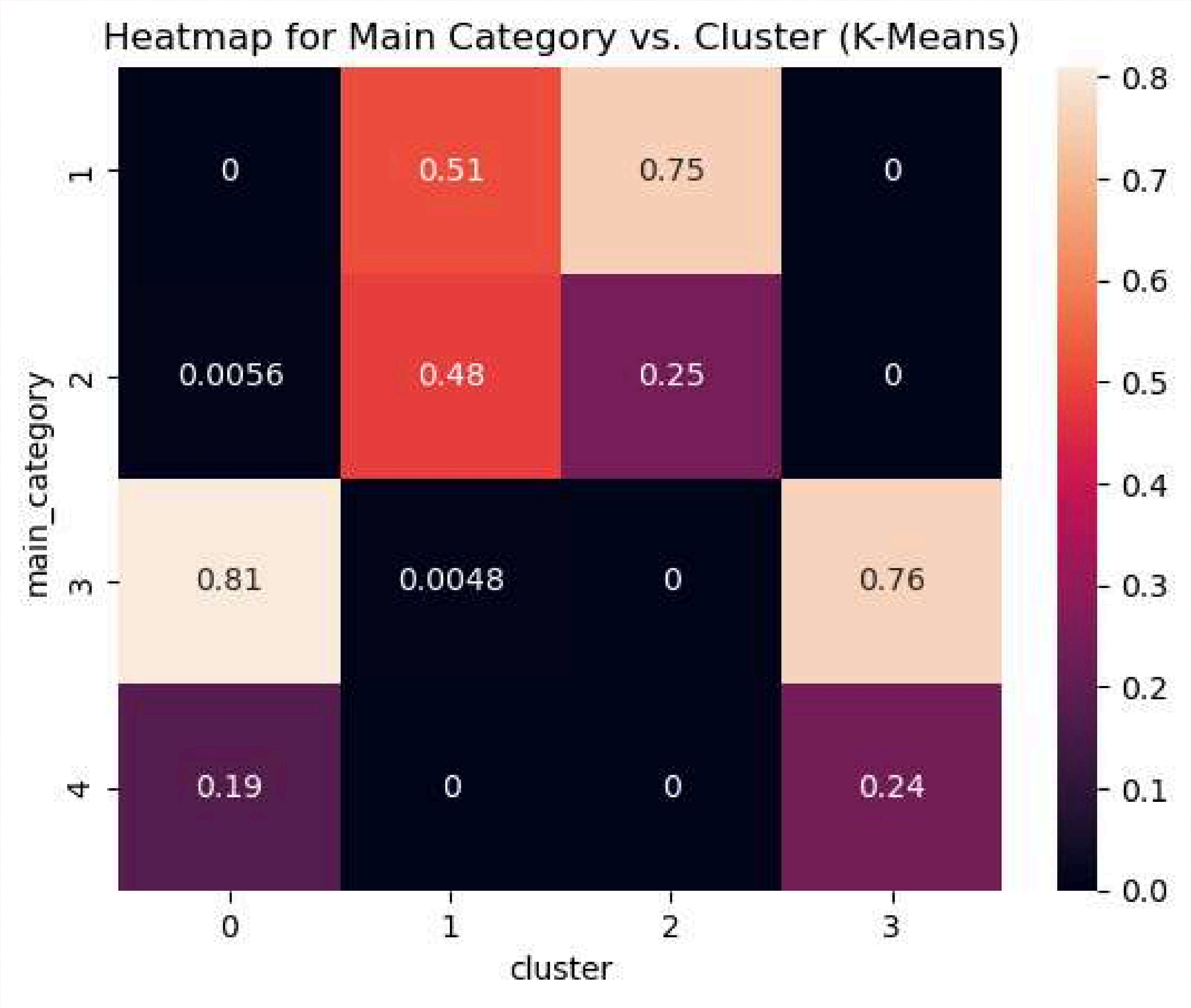
Page



0	Most purchases on Page 1, followed by Page 2
1	Most purchases on Page 1, followed by Page 2; A few purchases on Page 3
2	Most purchases on Page 1, followed by Page 2; A few purchases on Page 3
3	Most purchases on Page 3, followed by Page 4; A few purchases on Page 2 and 5

Observations

Main Category



0

Top purchase is blouses, followed by clothes on sale

1

Almost equal purchase of trousers and skirts

2

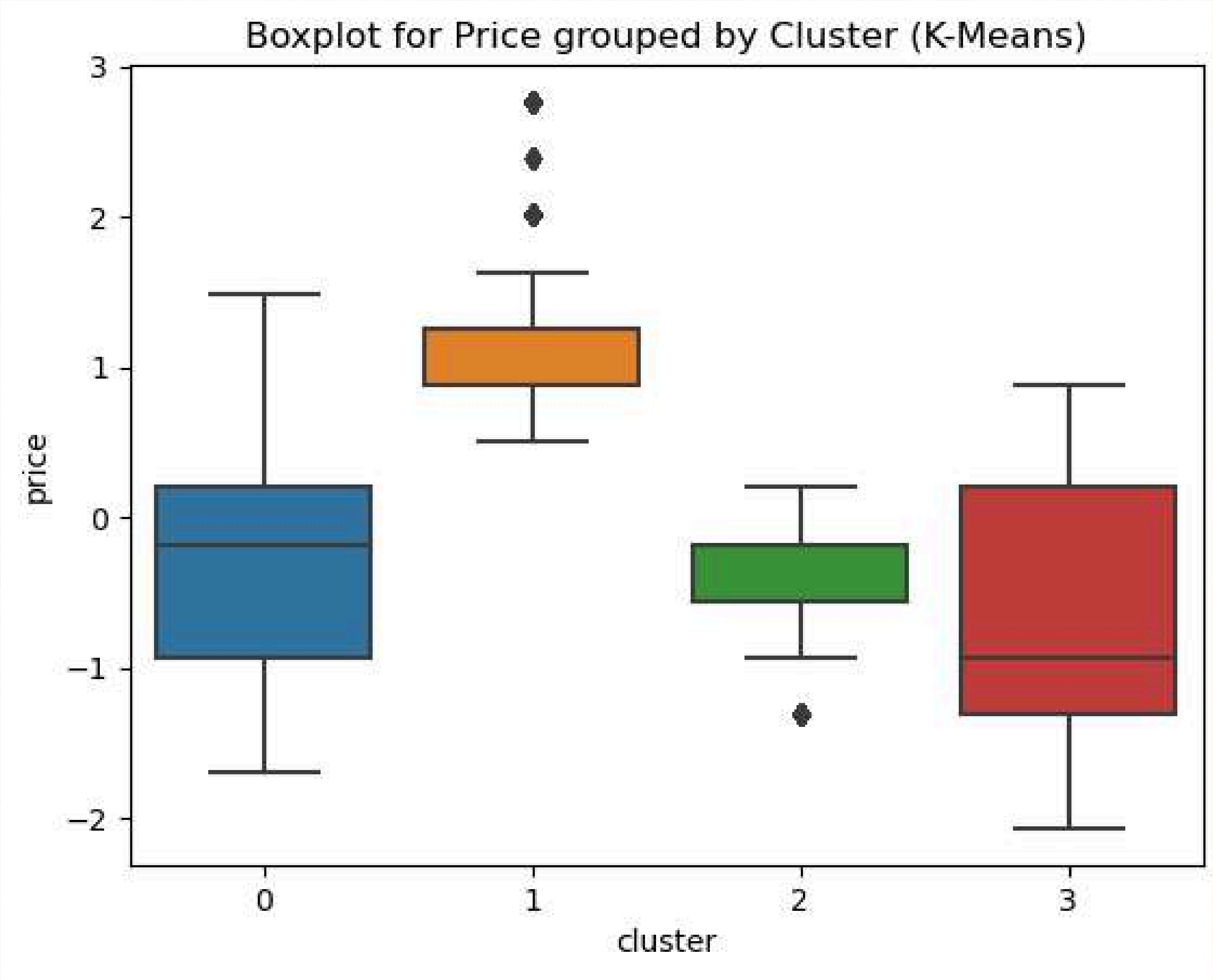
Top purchase is trousers, followed by skirts

3

Top purchase is blouses, followed by clothes on sale

Observations

Price



0

Tends to spend around the average price

1

Tends to buy expensive items

2

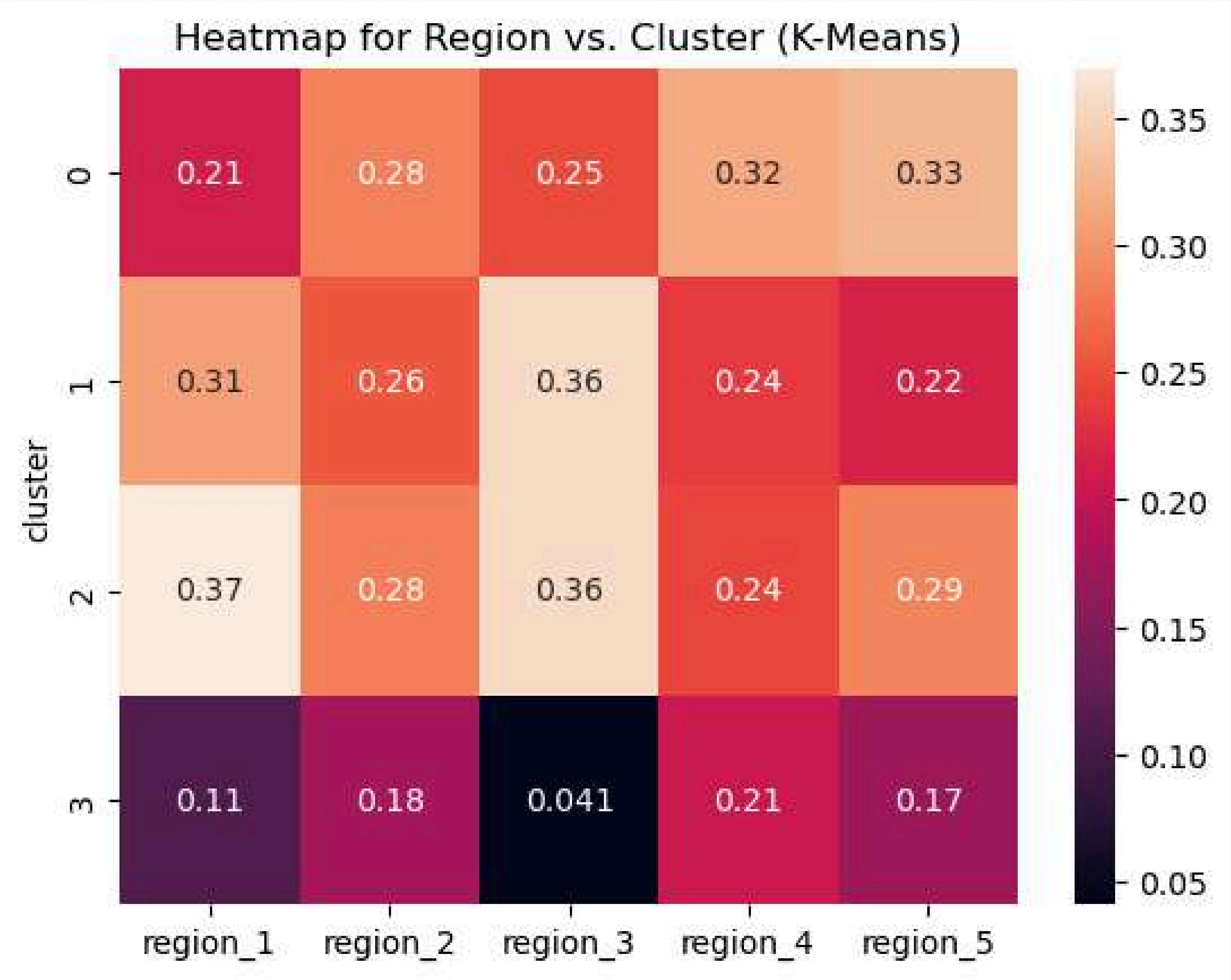
Tends to spend less than the average price

3

Tends to spend less than the average price

Observations

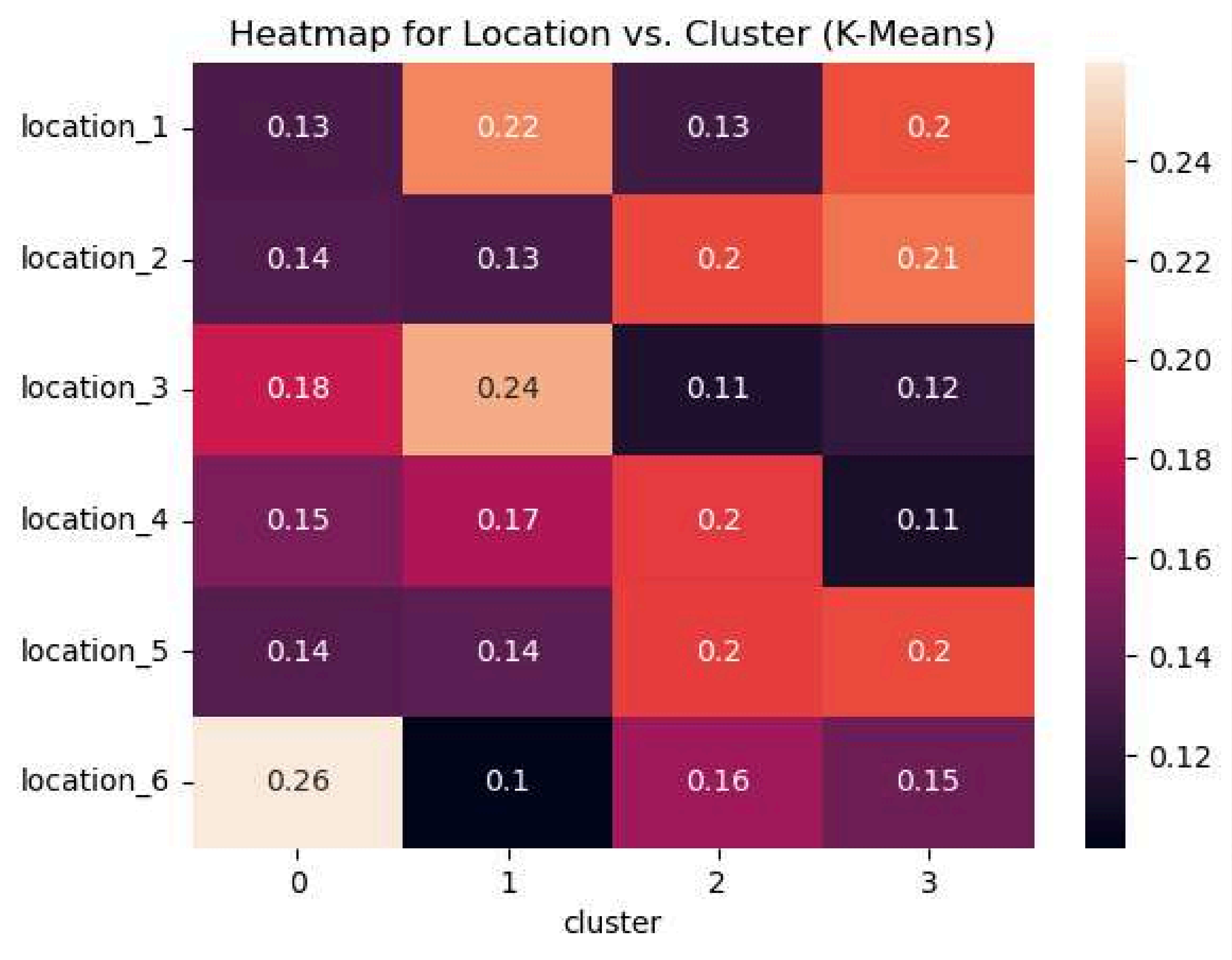
Region



1. Mainly in Cluster 1 and 2, followed by Cluster 0
2. Mainly in Cluster 0, 1, and 2, followed by Cluster 3
3. Mainly in Cluster 1 and 2, followed by Cluster 0
4. Mainly in Cluster 0, followed by Cluster 1, 2, and 3
5. Mainly in Cluster 0 and 2, followed by Cluster 1 and 3

Observations

Location



0 Mostly purchasing from the bottom right

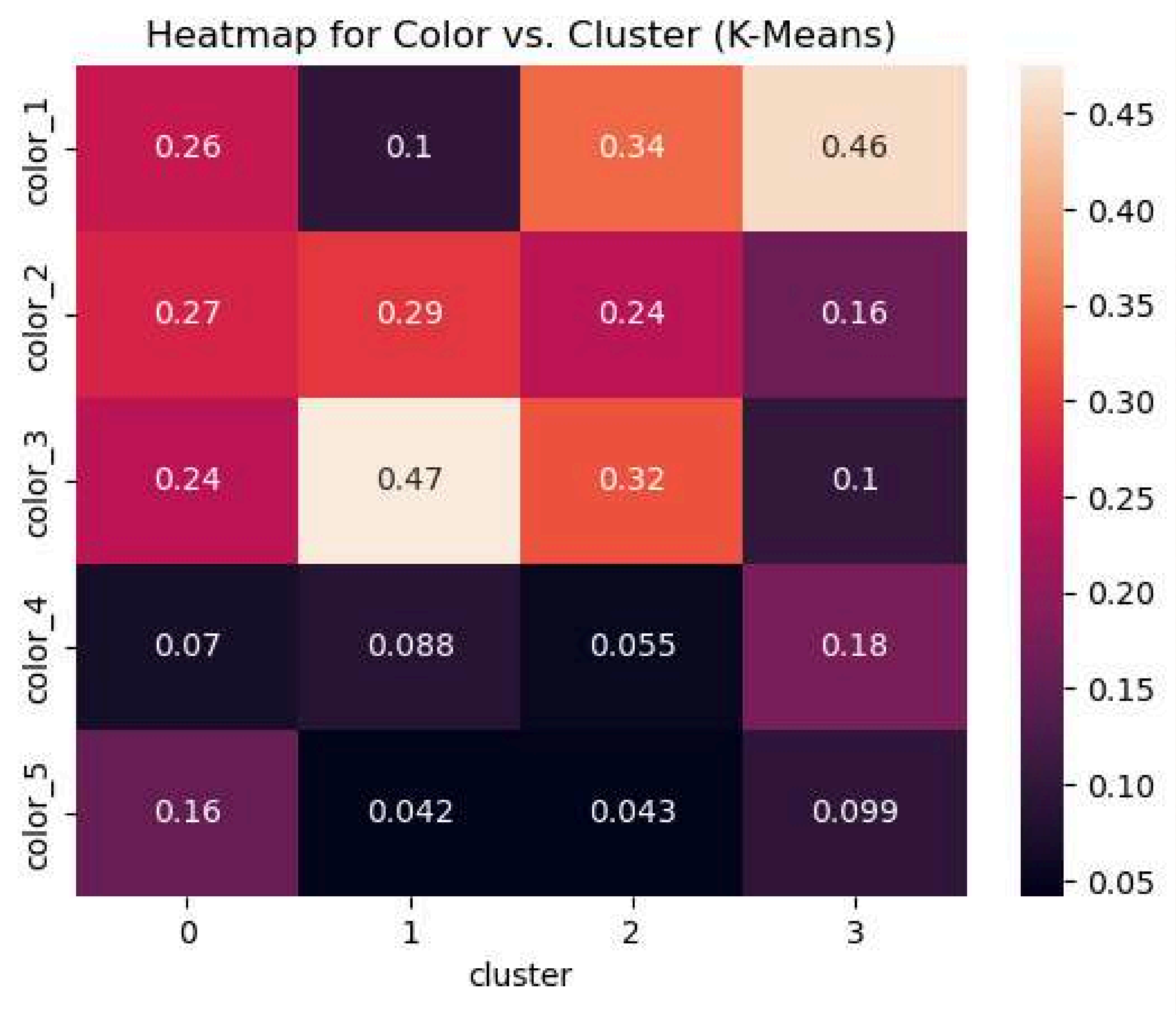
1 Mostly purchasing from the top corners

2 Mostly purchasing from the bottom left and middle columns

3 Mostly purchasing from the top left and middle columns

Observations

Color



0 Top purchases are light and dark neutrals and light colors

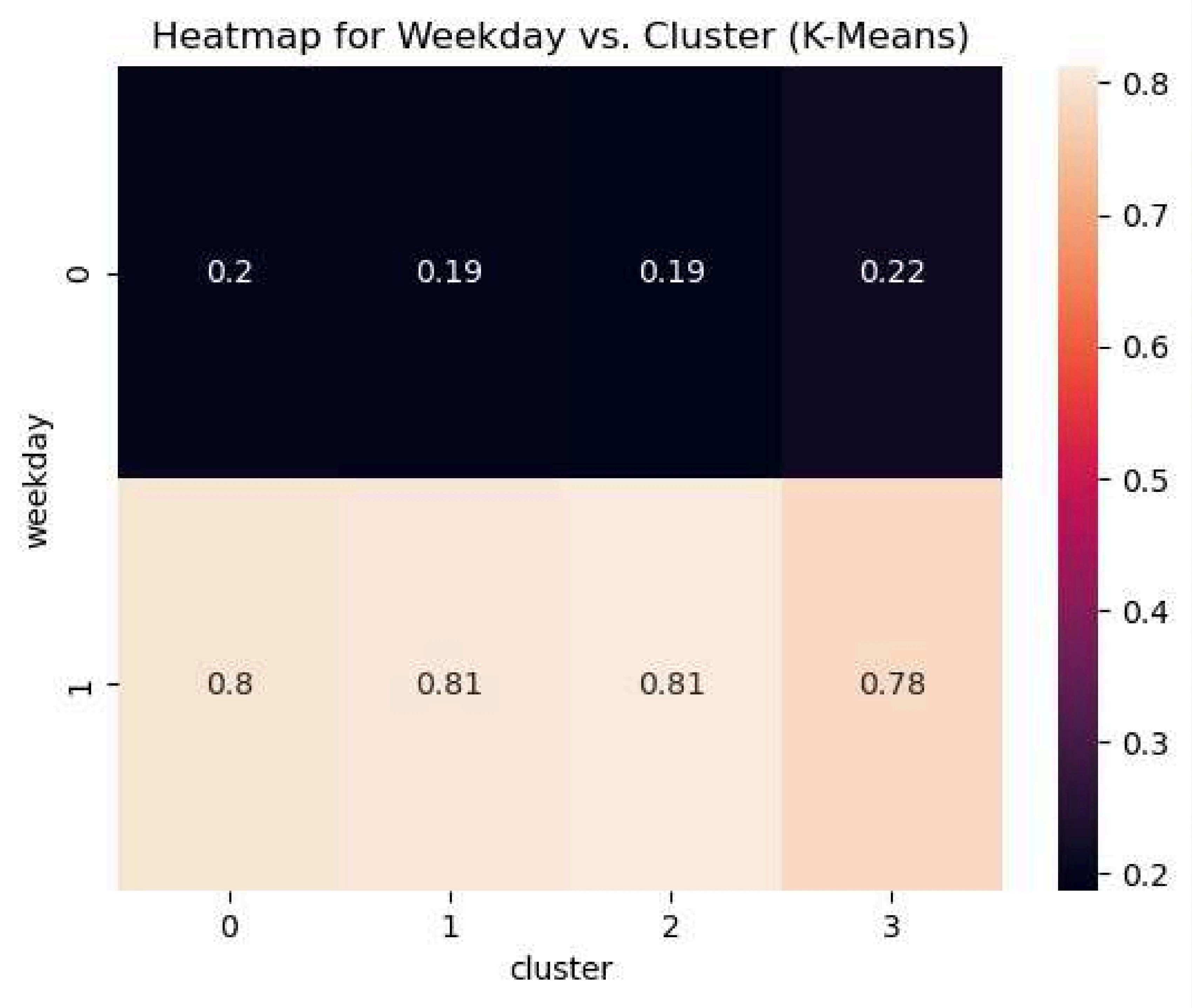
1 Top purchases are light colors followed by dark neutrals

2 Top purchases are light neutrals and light colors followed by dark neutrals

3 Top purchase are light neutrals, followed by dark neutrals and colors

Observations

Weekday



0 Most purchases were made on a weekday

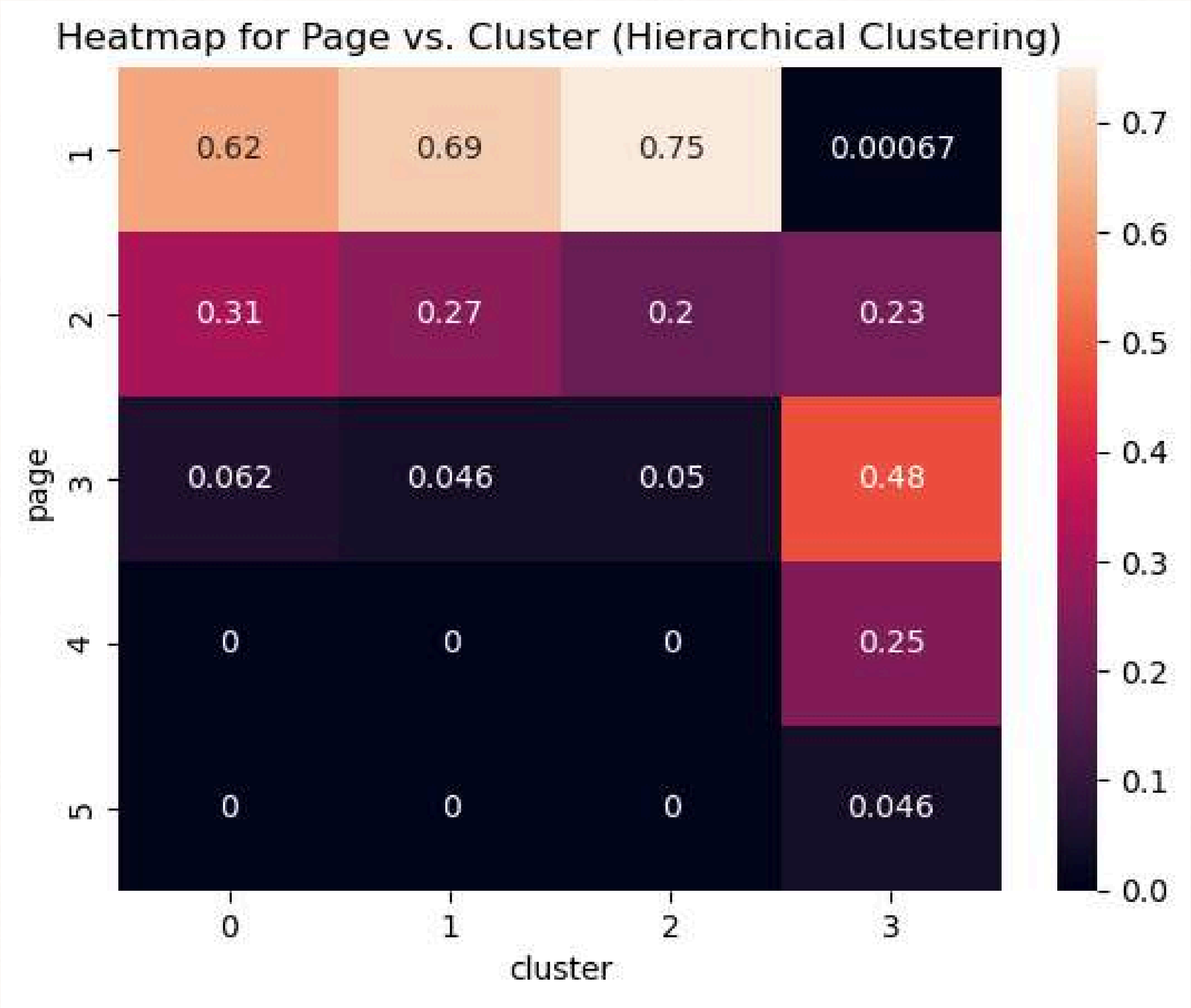
1 Most purchases were made on a weekday

2 Most purchases were made on a weekday

3 Most purchases were made on a weekday

Observations

Page



0 Most purchases on Page 1, followed by Page 2

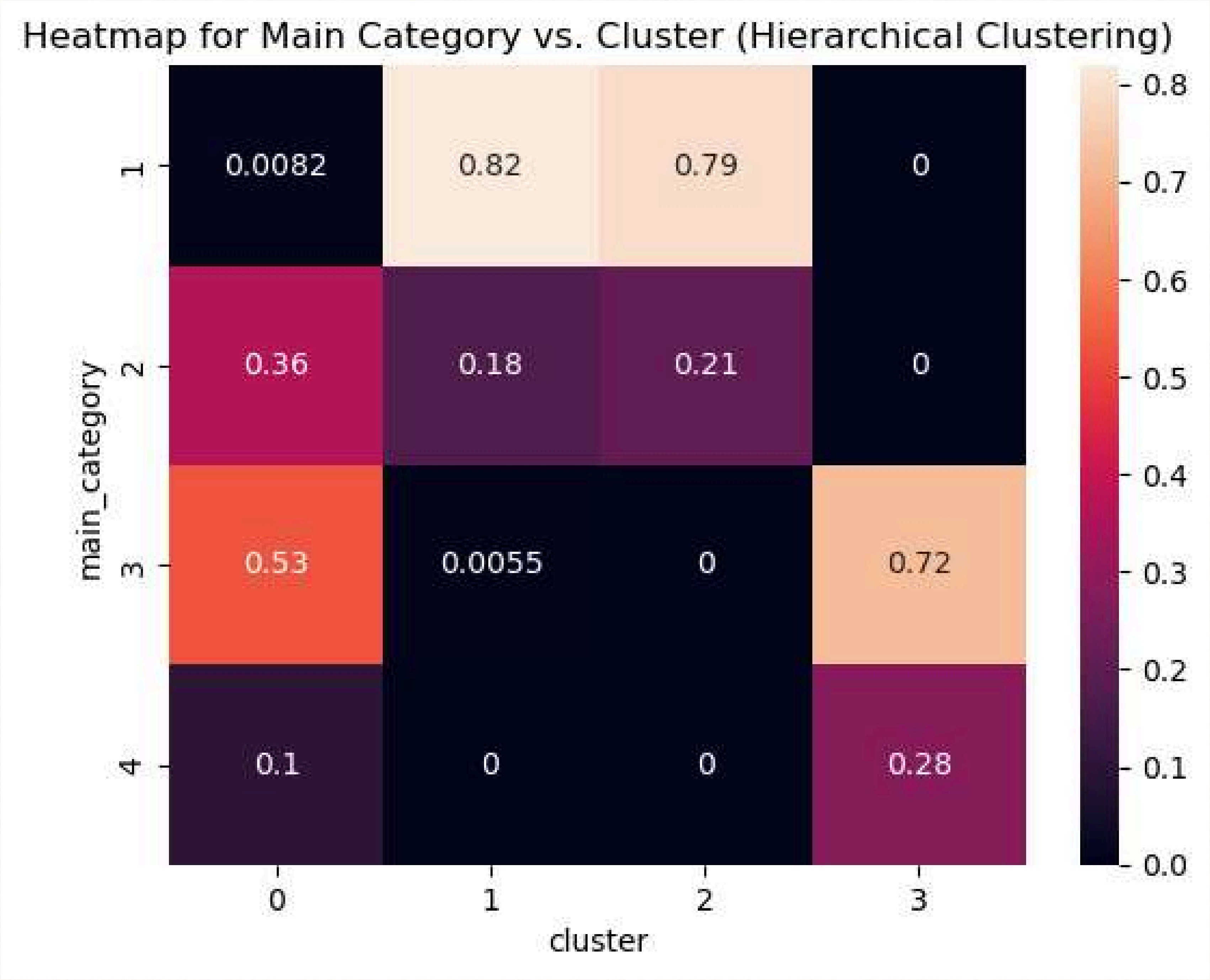
1 Most purchases on Page 1, followed by Page 2

2 Most purchases on Page 1, followed by Page 2

3 Most purchases on Page 3, followed by Page 2 and 4

Observations

Main Category



0 Top purchase is blouses, followed by skirts

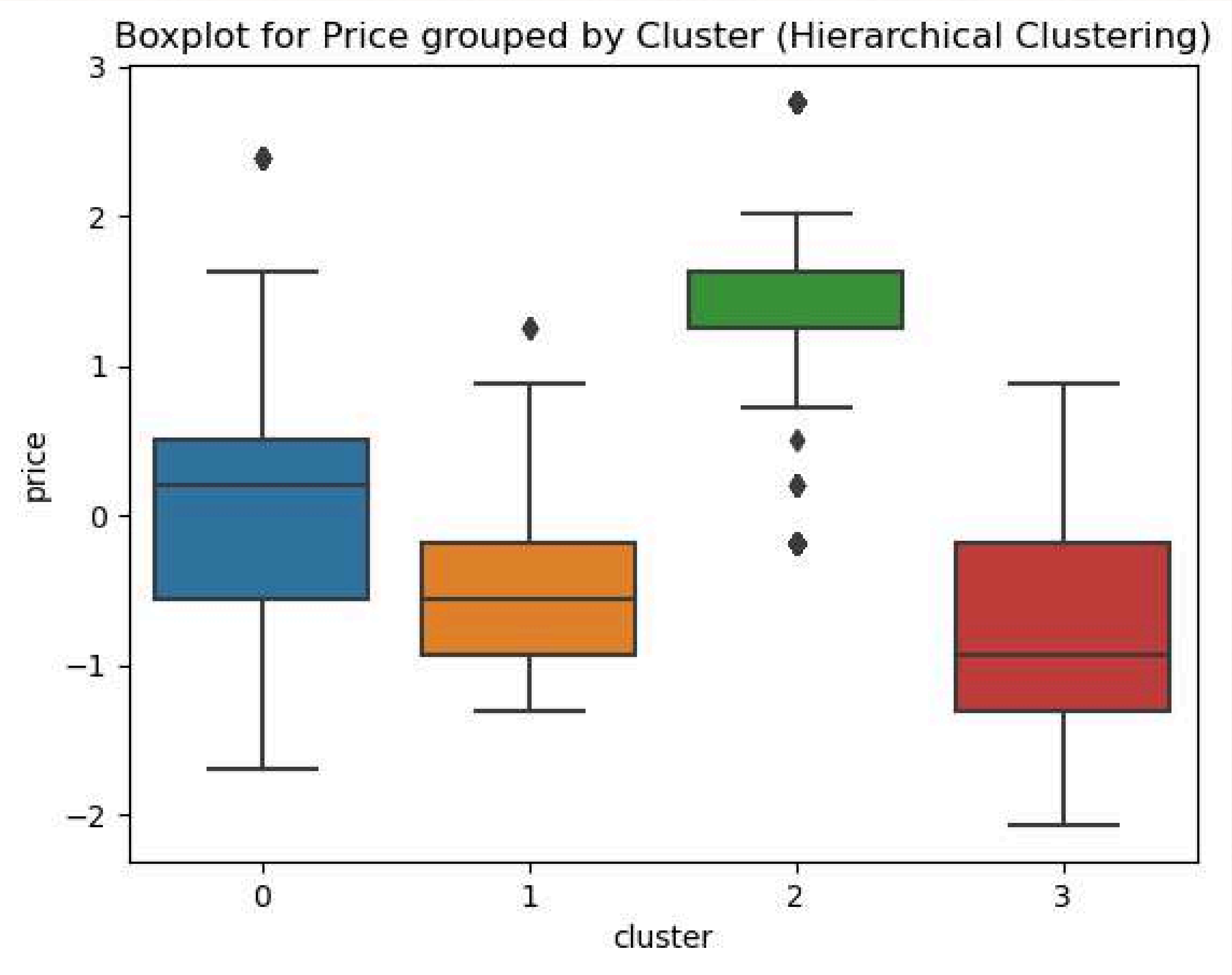
1 Top purchase is trousers, followed by skirts

2 Top purchase is trousers, followed by skirts

3 Top purchase is blouses, followed by clothes on sale

Observations

Price



0 — Tends to spend around the average price

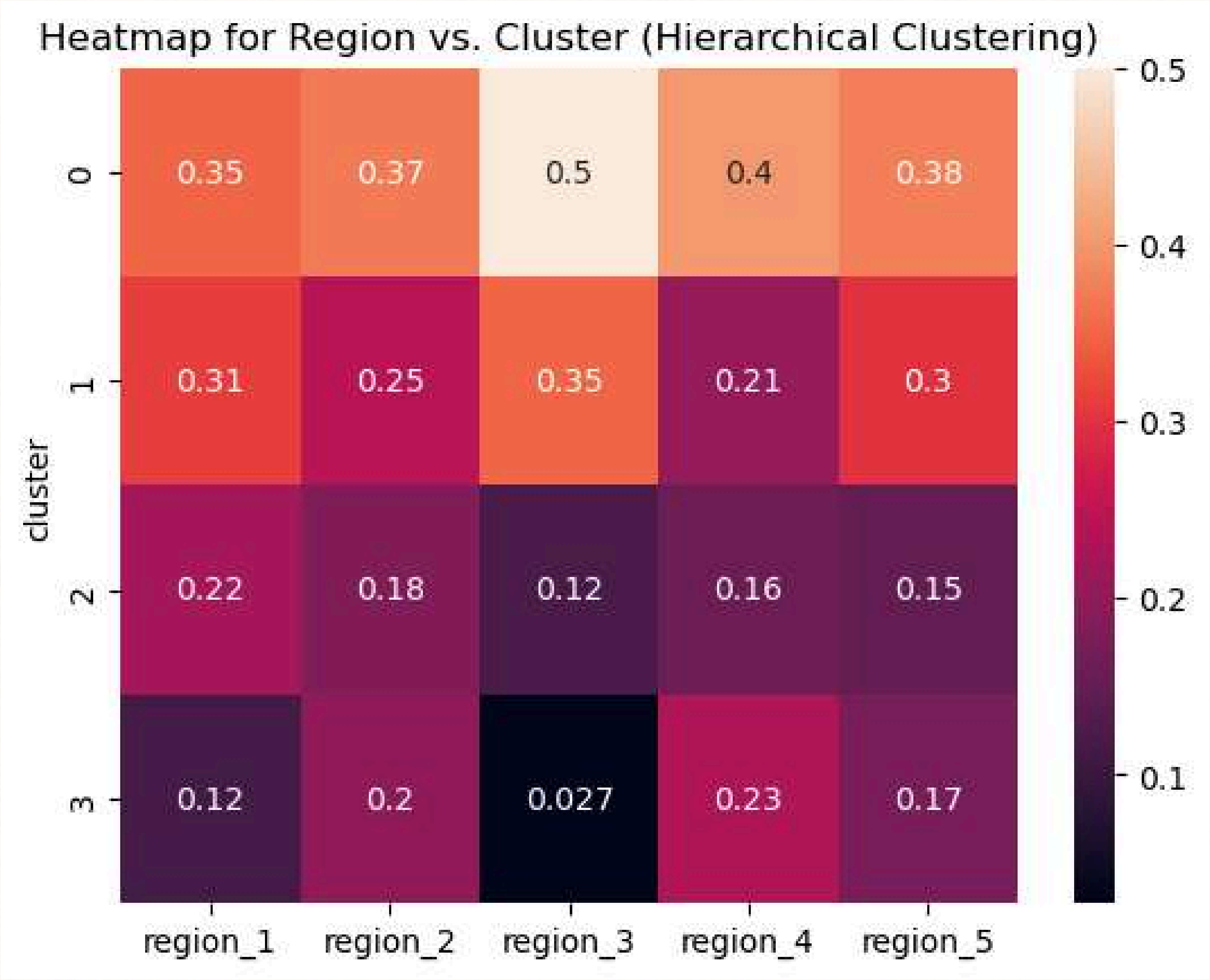
1 — Tends spend less than the average price

2 — Tends to buy more expensive items

3 — Tends to spend less than the average price

Observations

Region



1 Mainly in Cluster 0 and 1, followed by Cluster 2

2 Mainly in Cluster 0, followed by Cluster 1, 2, and 3

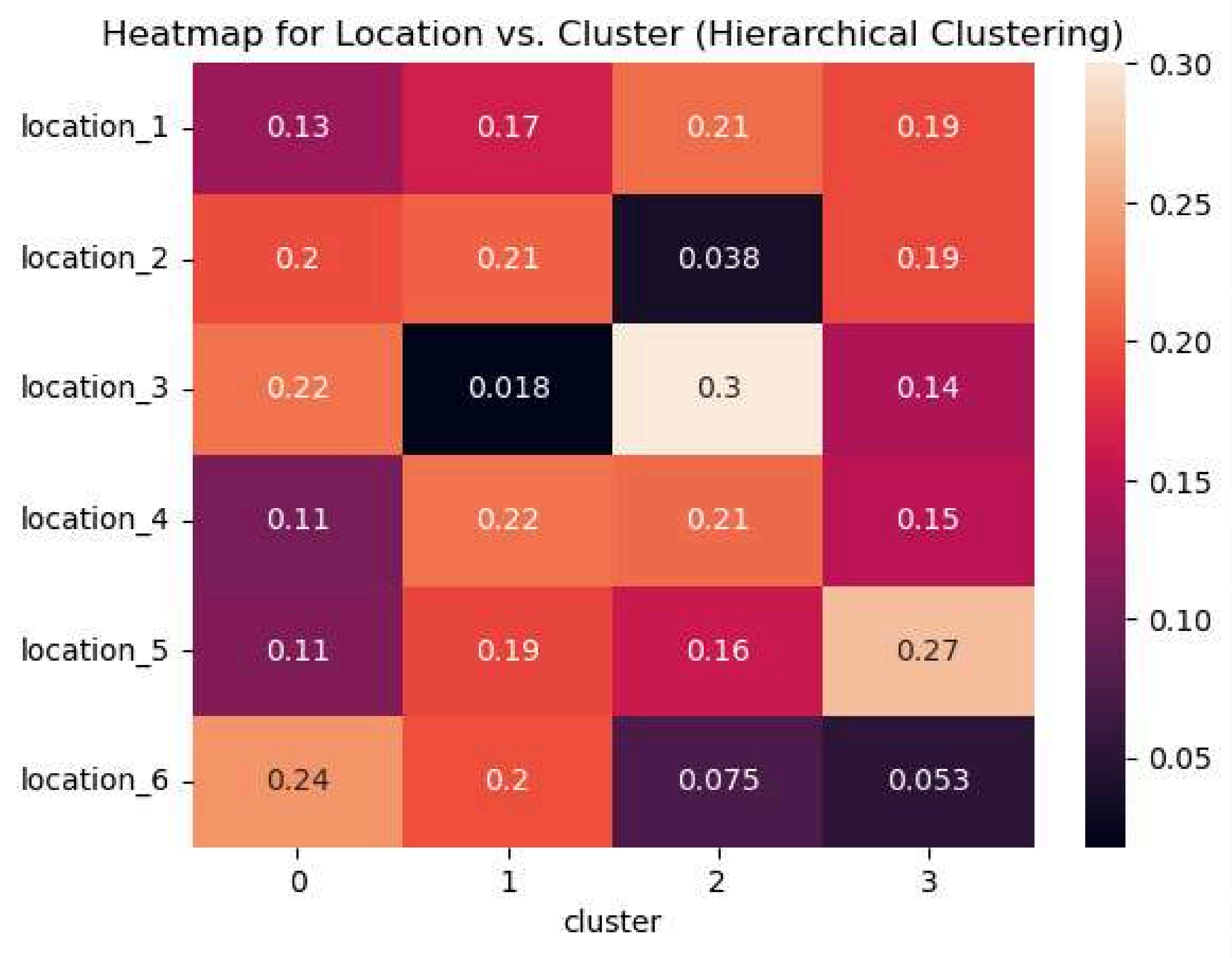
3 Mainly in Cluster 0, followed by Cluster 1

4 Mainly in Cluster 0, followed by Cluster 1 and 3

5 Mainly in Cluster 0 and 1, followed by Cluster 2 and 3

Observations

Location



0 Mostly purchasing from the top middle and right column

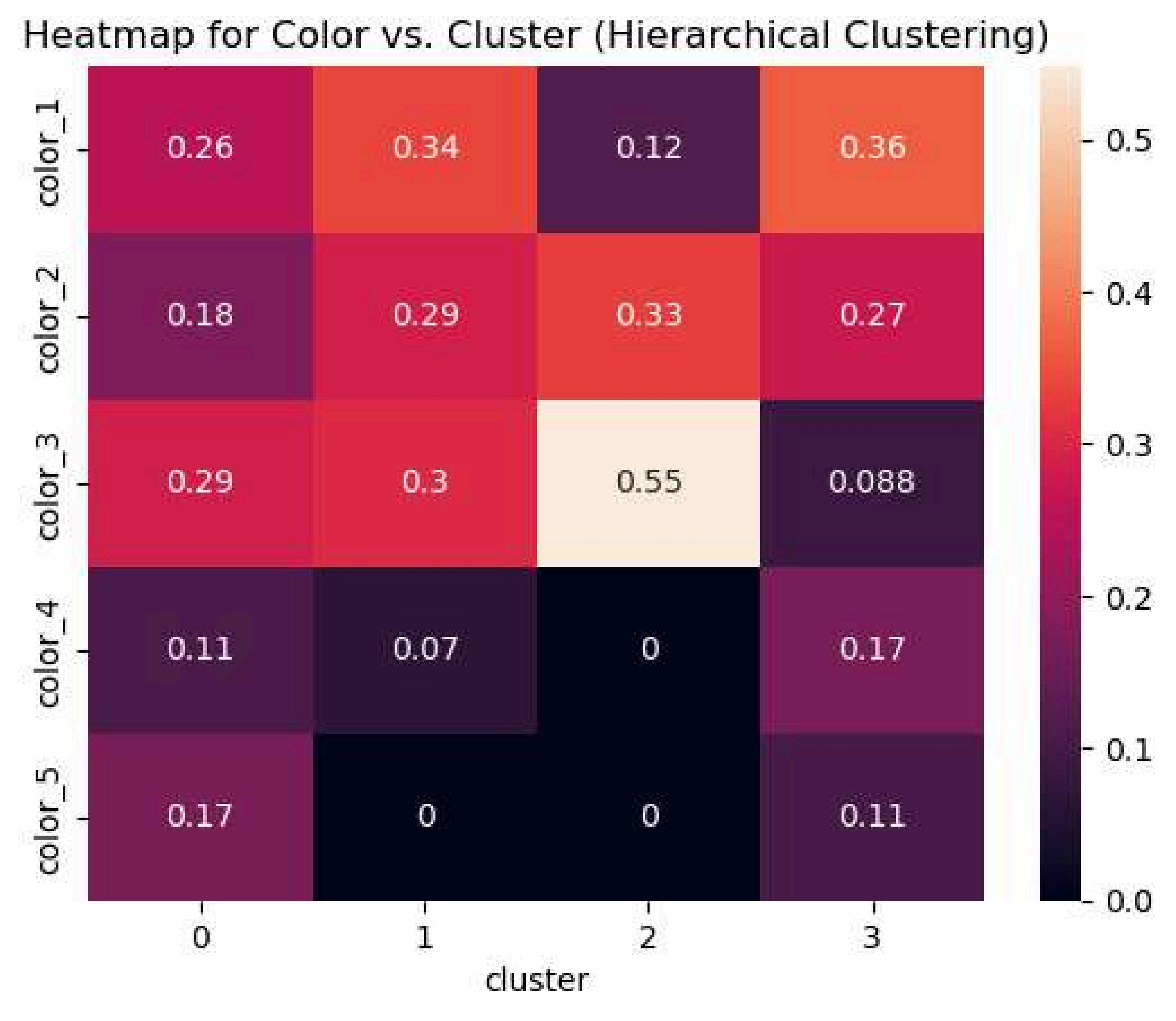
1 Mostly purchasing from the bottom row and the top middle

2 Mostly purchasing from the top corners and the bottom left

3 Mostly purchasing from the bottom middle, followed by top left and middle

Observations

Color



0 Top purchases are light neutrals and light colors

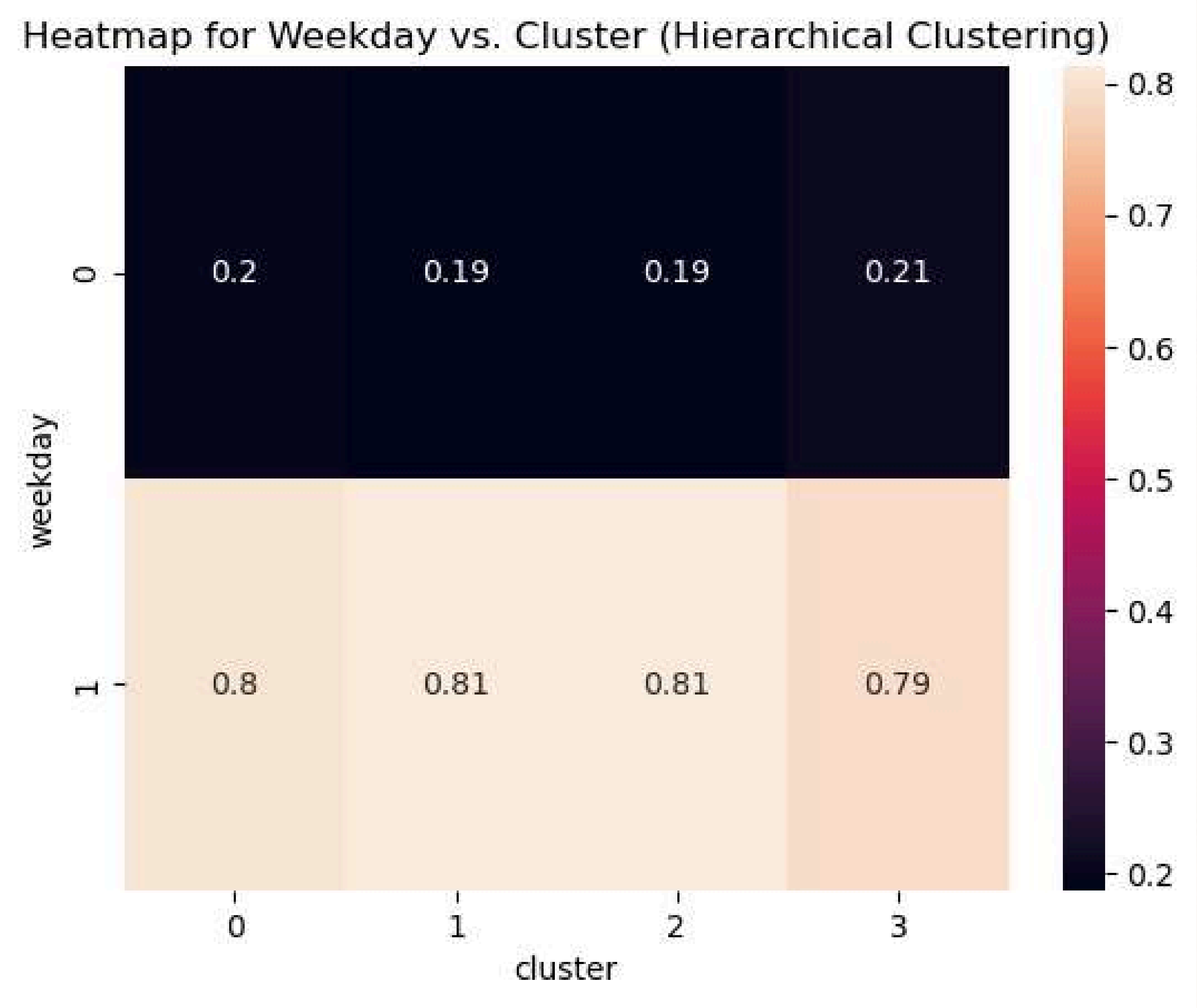
1 Top purchases are light and dark neutrals, and light colors

2 Top purchases are light colors followed by dark neutrals

3 Top purchases are light and dark neutrals

Observations

Weekday



0 — Most purchases were made on a weekday

1 — Most purchases were made on a weekday

2 — Most purchases were made on a weekday

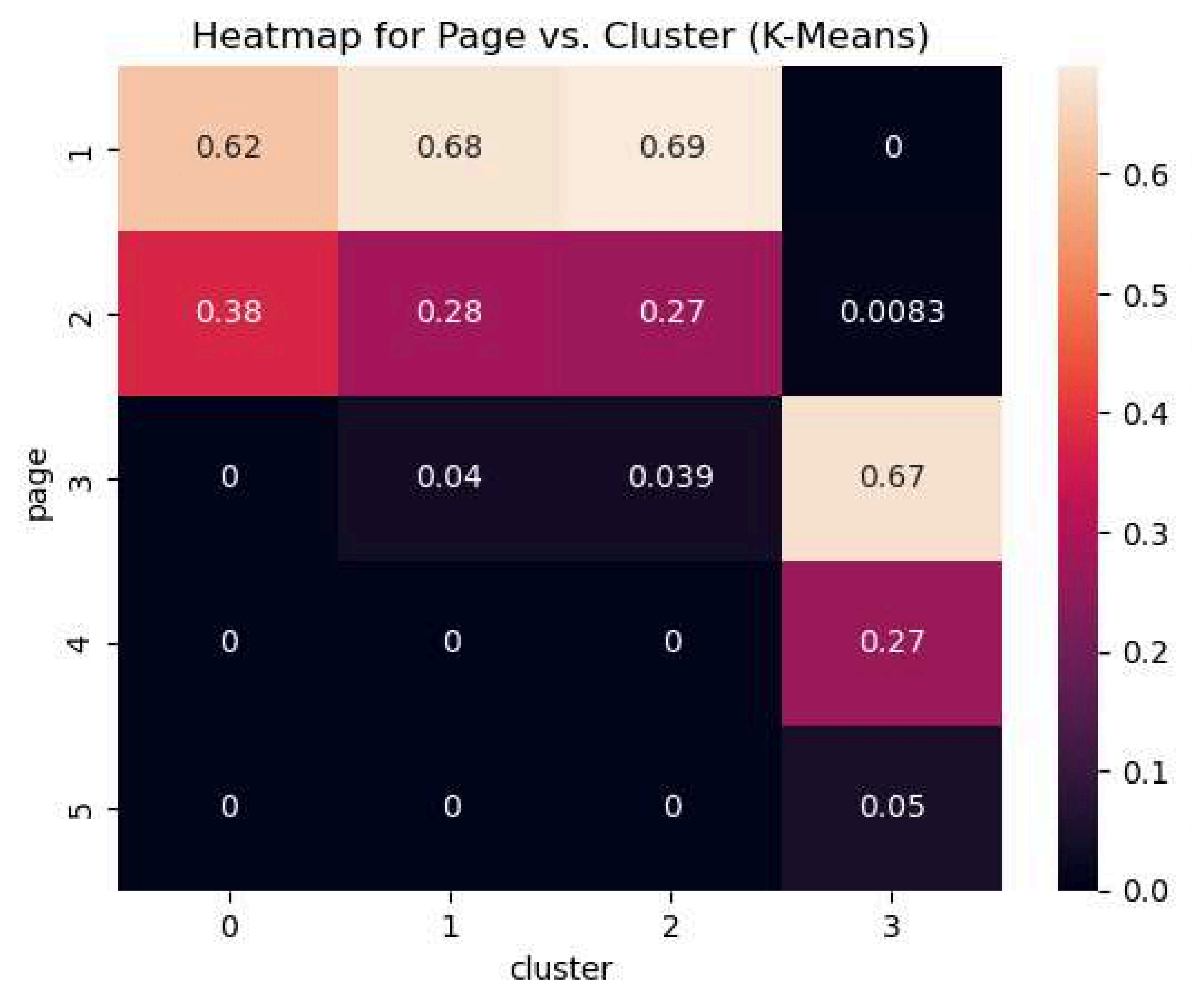
3 — Most purchases were made on a weekday

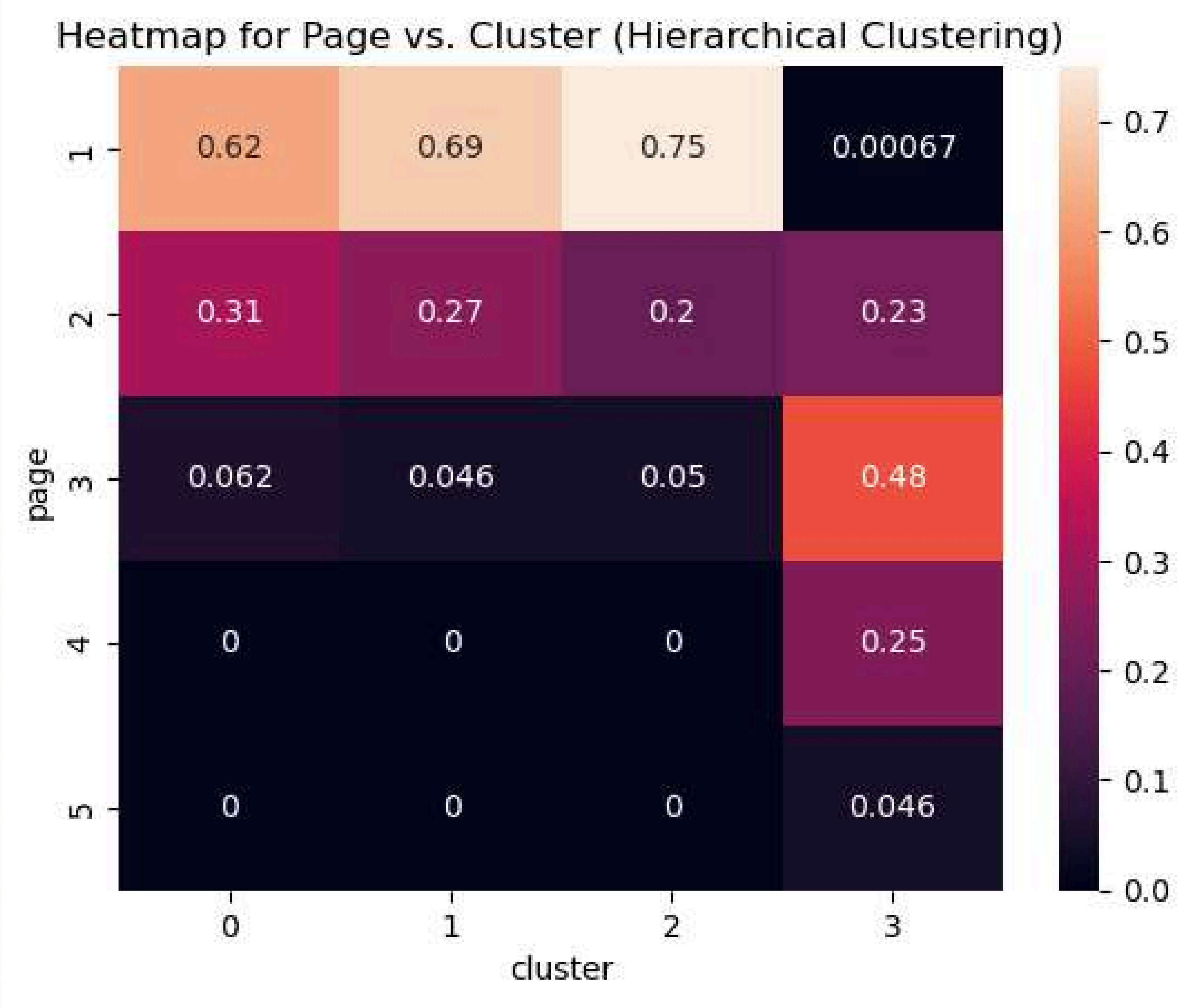
Initial Insights

K-Means has a higher silhouette score but by a small margin (0.18 vs 0.16). However, the K-Means clustering has more clear-cut results in terms of page numbers and locations, which are our most important variables for this study.

Comparison Between Models

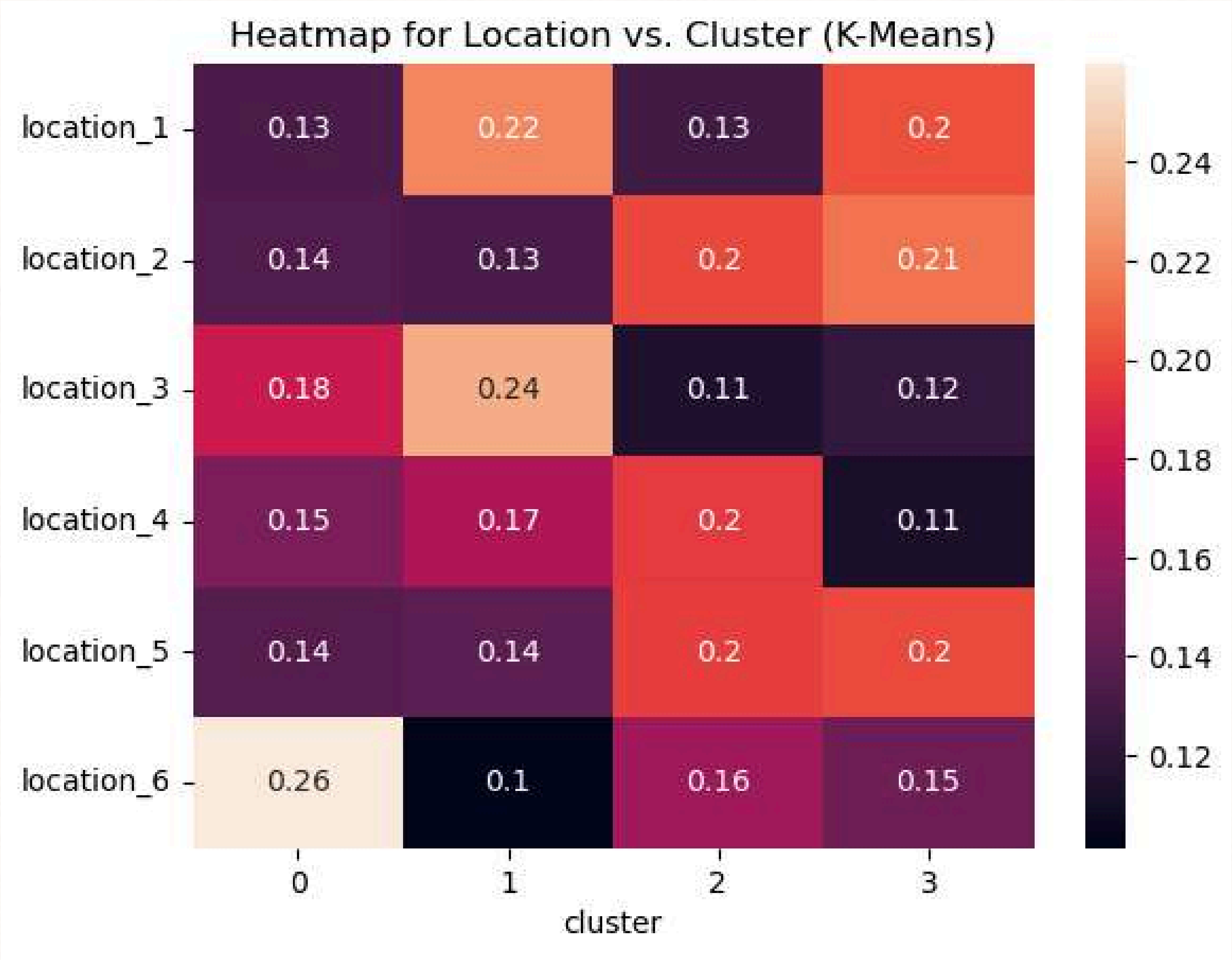
Page

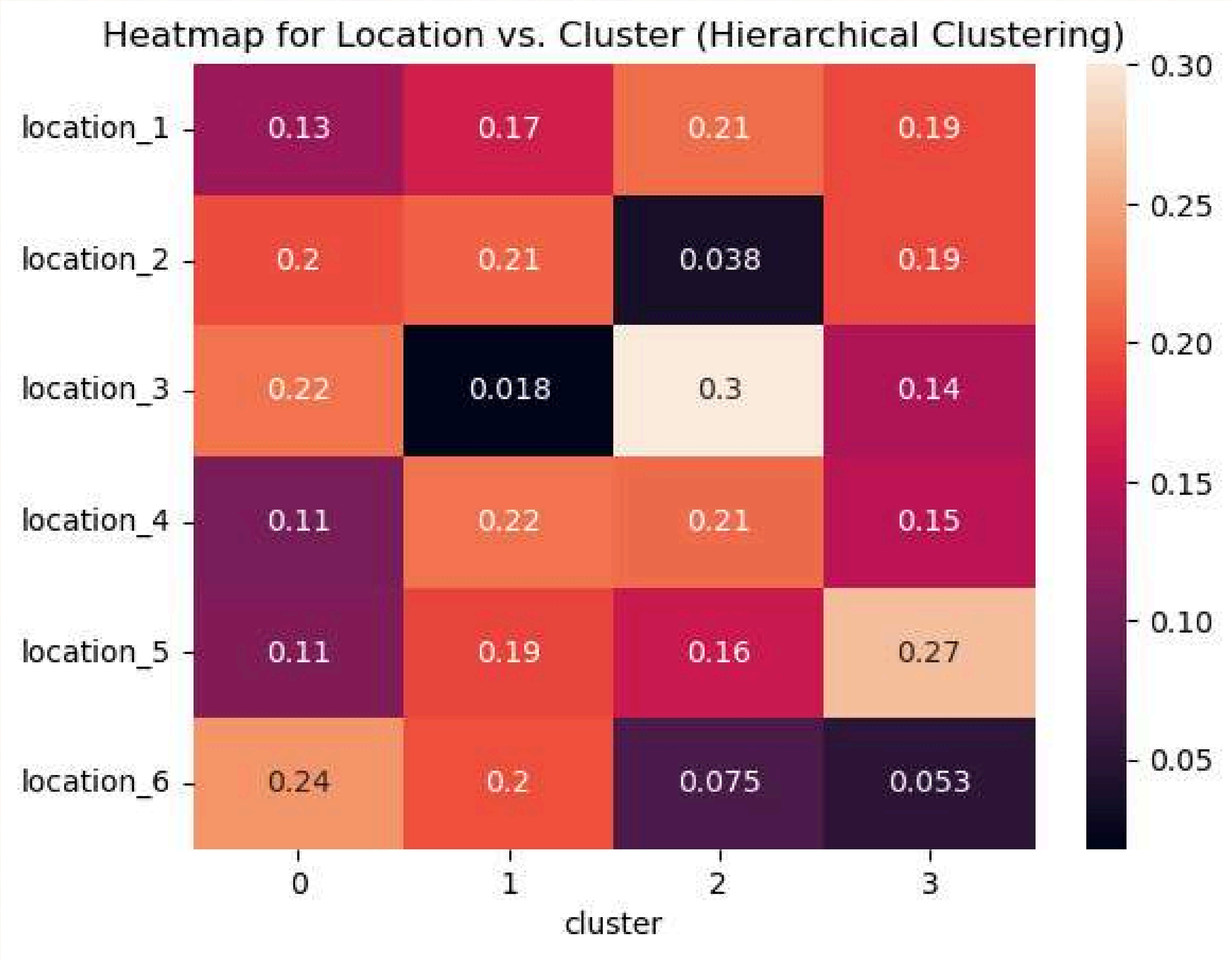




Comparison Between Models

Location





Revenue Generation

For K-Means, Cluster 1 generates the most revenue (i.e., median higher than 0).

- Purchases from Page 1 and 2
- Purchases trousers and skirts
- Purchasing from the top corners
- Purchases light colored and dark neutral clothing

Revenue Generation

Additionally, Cluster 0 also spends around the average price.

- Purchases from Page 1 and 2
- Purchases blouses and clothes on sale
- Purchasing from the bottom right
- Purchases neutrally colored clothing

Insights

- The distribution of clusters per region is not quite clear-cut.
- If there is a future model with more clear-cut insights:
  - Suppose majority of Region 1 is composed of Cluster 1, and Cluster 1 purchases more blouses.
  - Advertising in Region 1 can focus on blouses.

Conclusions

Customer segmentation is a useful technique to understand the behavior of the online store's customers.

Our insights can be used to implement personalized product placement strategies based on several factors that result in higher revenues from each cluster. In doing so, we are able to increase the overall revenue of the company.

Recommendations

- Explore other unsupervised machine learning models (e.g., other clustering methods, association rules).
  - DBSCAN
- The number of items purchased per session (which we can extract via order) can also be considered.
- Consider purchases made in Poland.
- Test the accuracy of the model by considering the clusters as labels.

Personal Reflection

Sted Cheng

I realized the importance of model interpretability in the business context. In BI, we prioritize generating insights that are actionable over ML models which are sophisticated but unexplainable to the common stakeholder.

Personal Reflection

Annika Montemayor

This experience highlighted the importance of always going back to your "Why?" Each step of the pipeline should always be anchored on what goals you and your stakeholders are trying to achieve.

Personal Reflection

Kaitlyn Shu Too

Beyond technical Data Science skills, communication skills play a key factor in doing business analytics. Working in a team, it’s important to communicate in order to create more effective strategies and insights. In reporting to the business, it’s important to know how to express the insights that are useful to the business goals.

Thank you!