Lab Machine Learning for Data Science

Sommer Semester 2023
Freie Universität Berlin

FREIE UNIVERSITÄT BERLIN

Project 1: Unsupervised Machine Learning

Project Authors:
Jan Jascha Jestel
Mustafa Suman
Gabriele Inciuraite

Presenter: Gabriele Inciuraite

14.7.2023, Berlin

Project Goals

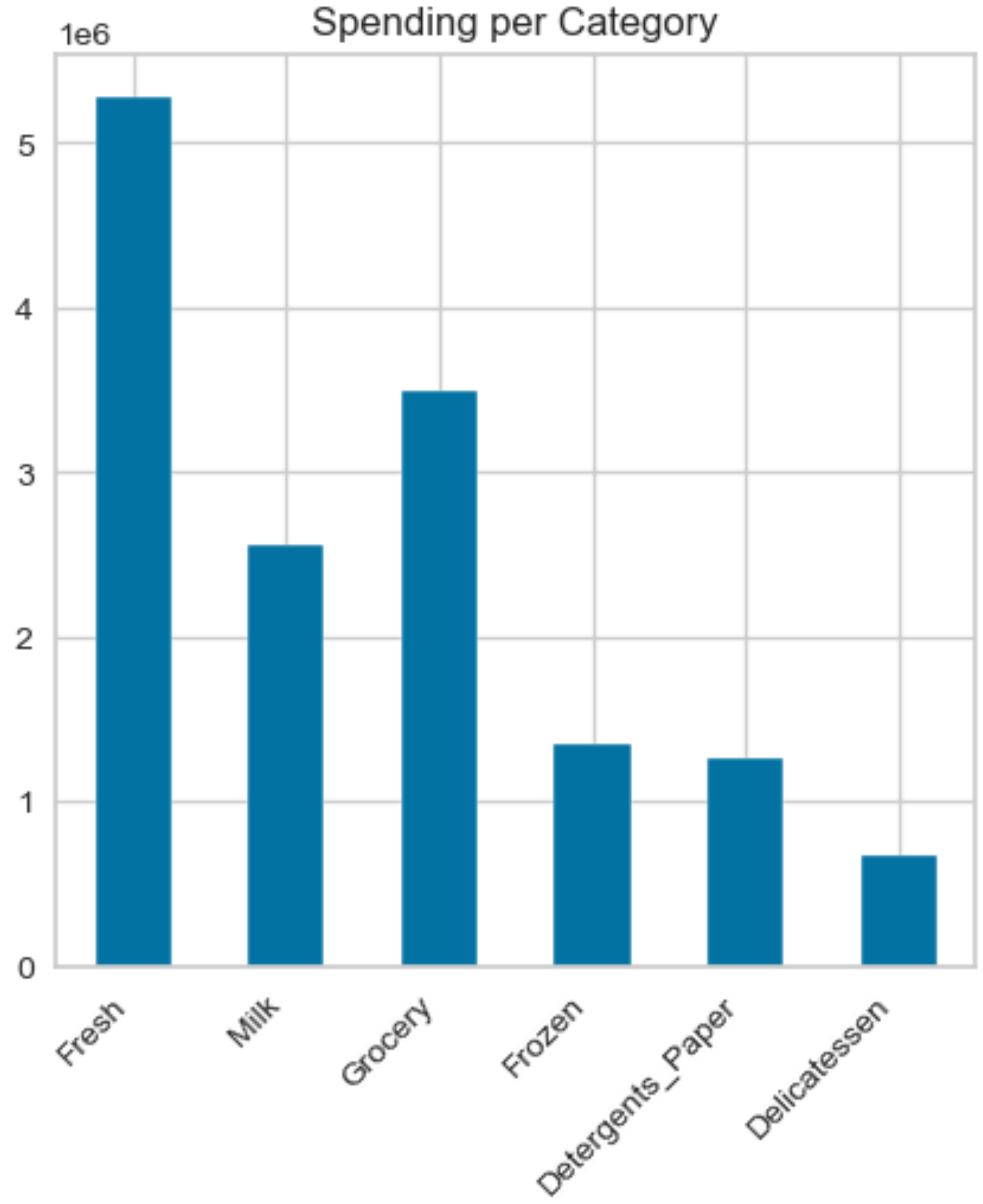
- UCI Wholesale customers dataset: annual spending on different product categories by wholesale customers located in Portugal

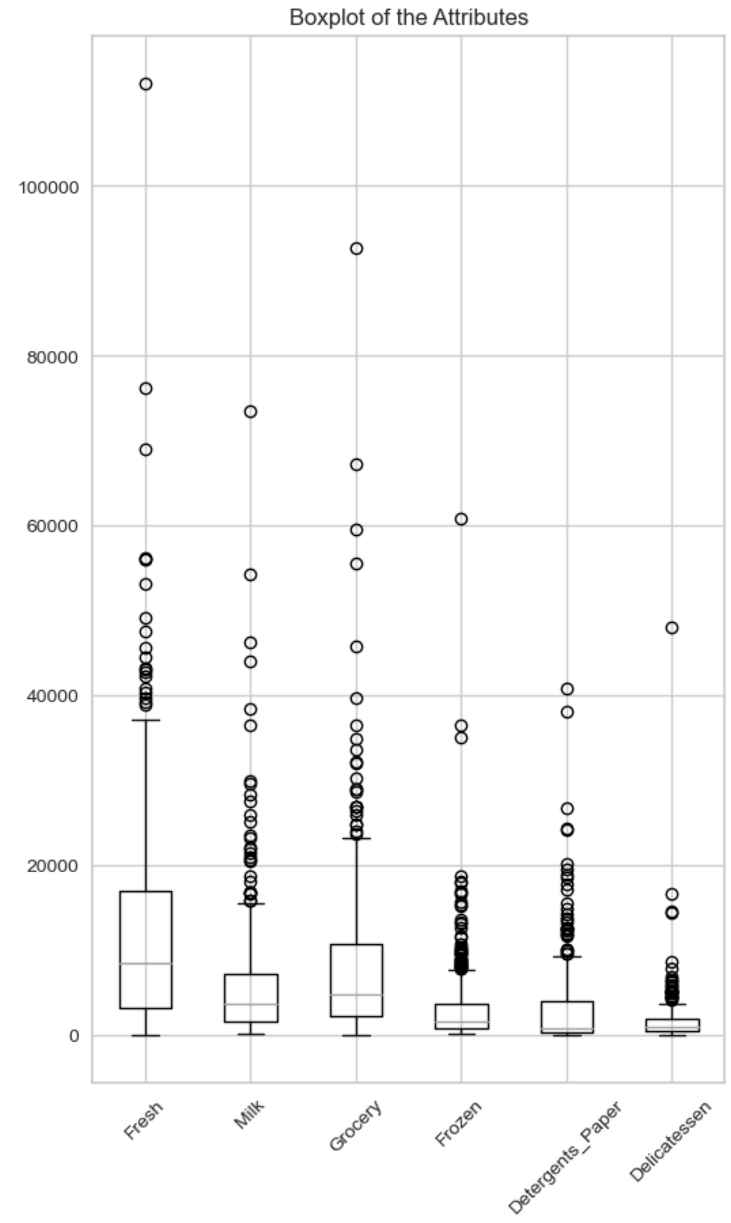
- → Identify instances with anomalous spending behaviour

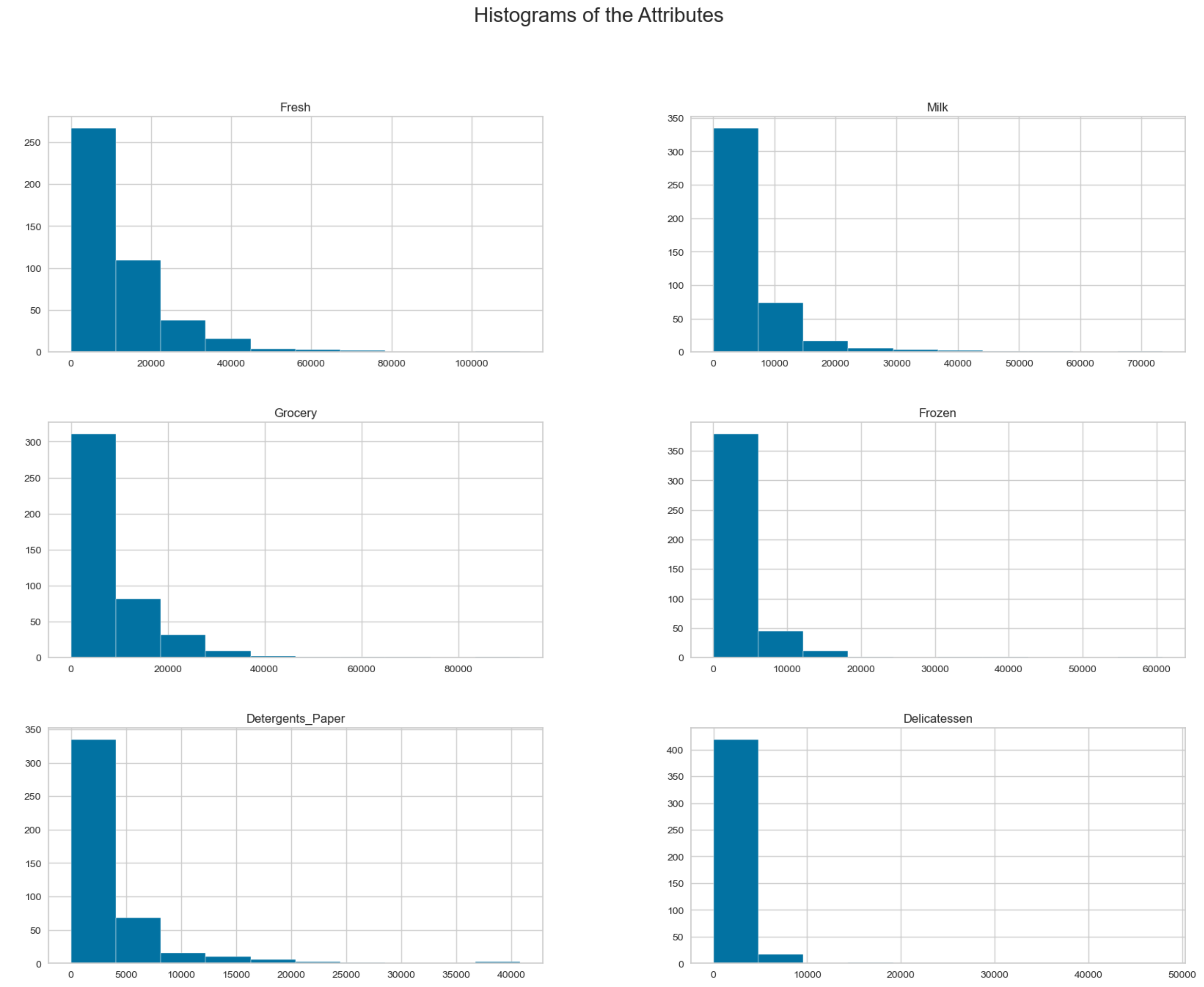
- → Identify clusters of similarly behaving wholesale customers

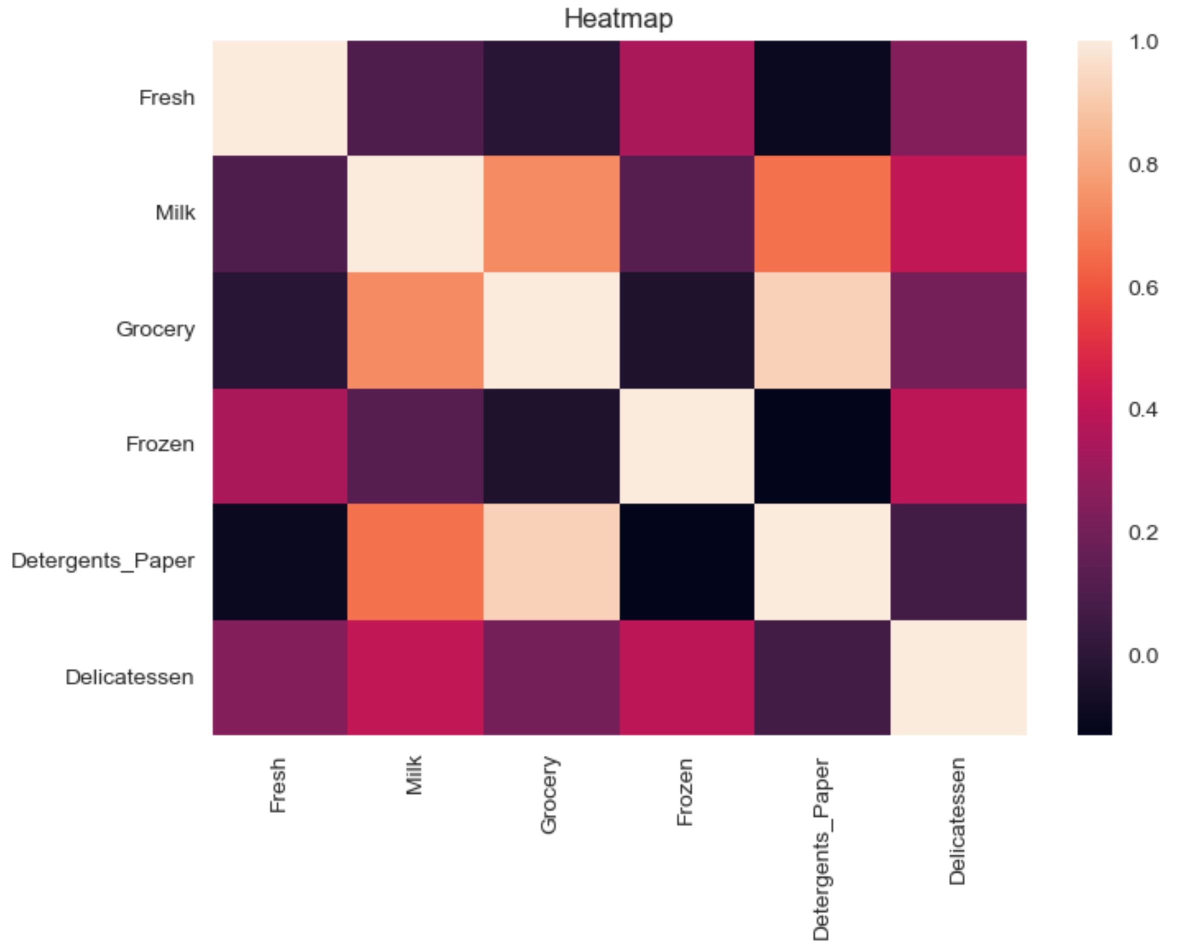
1. Initial Data Analysis and Preprocessing

- The distributions are heavy tailed → apply the log function, so that the distribution gets compressed for large values and expanded for small values

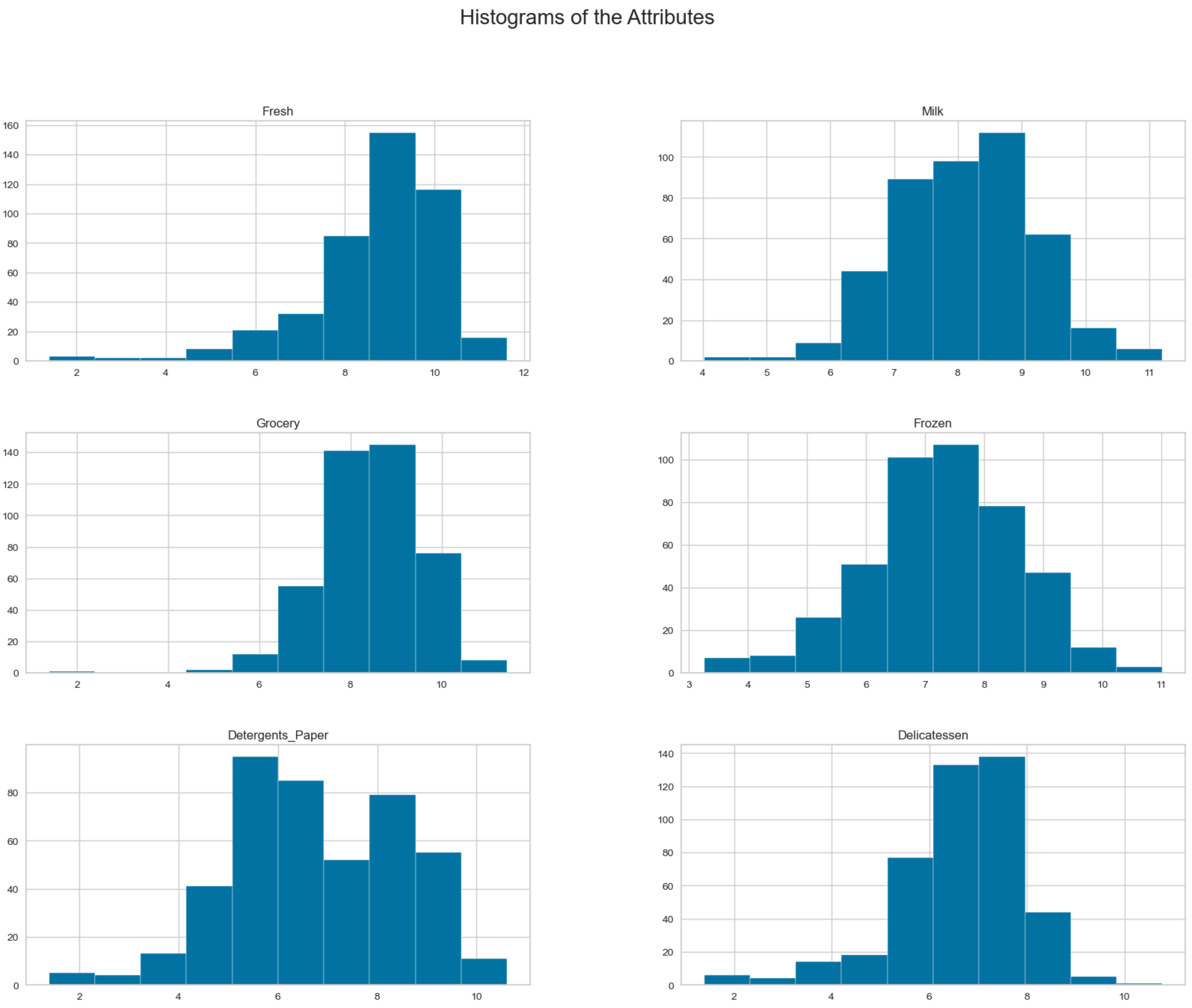


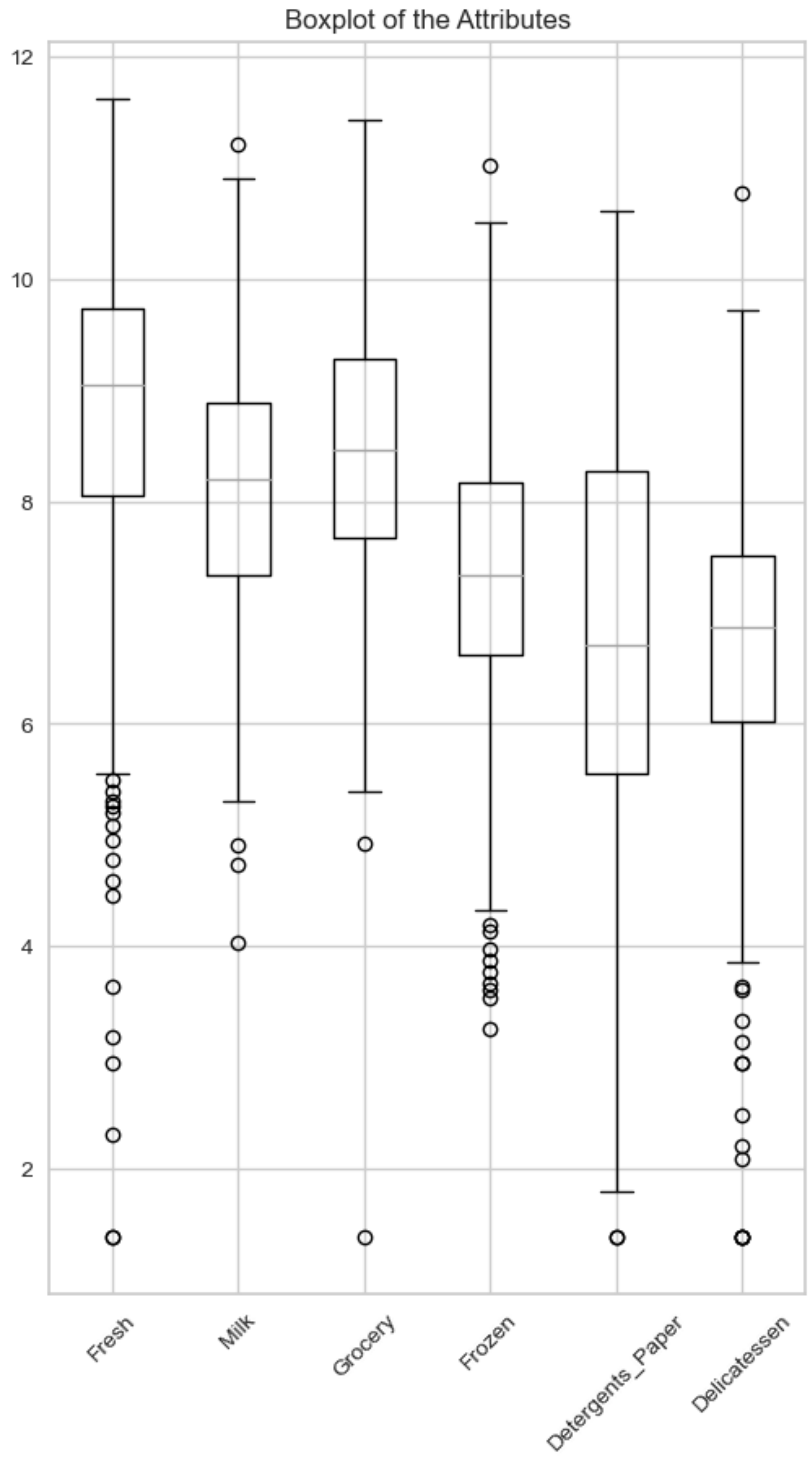






- → More normally distributed attributes
- → Fewer high spending customers possess extreme values





2. Detecting Anomalies

2.1. Hard-Min Score

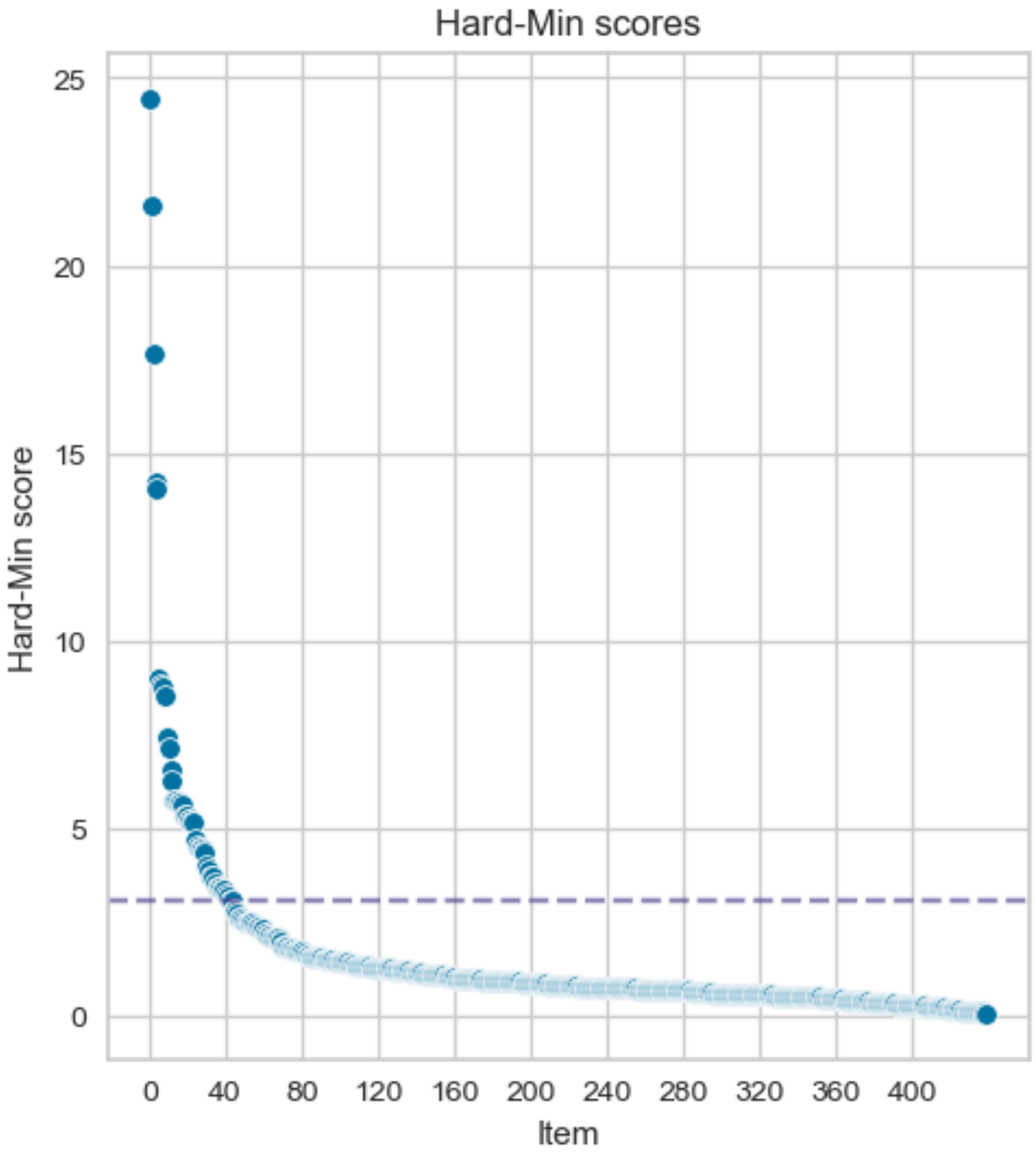
Creating Artificial Ground Truth

- **Hard-Min**: nearest neighbour distance per instance as outlier score
- For a more robust outlier score → apply **bootstrapping** with replacement, compute Hard-Min Scores for each sample
- Average over the scores per sample → 440 x 10000 measurements

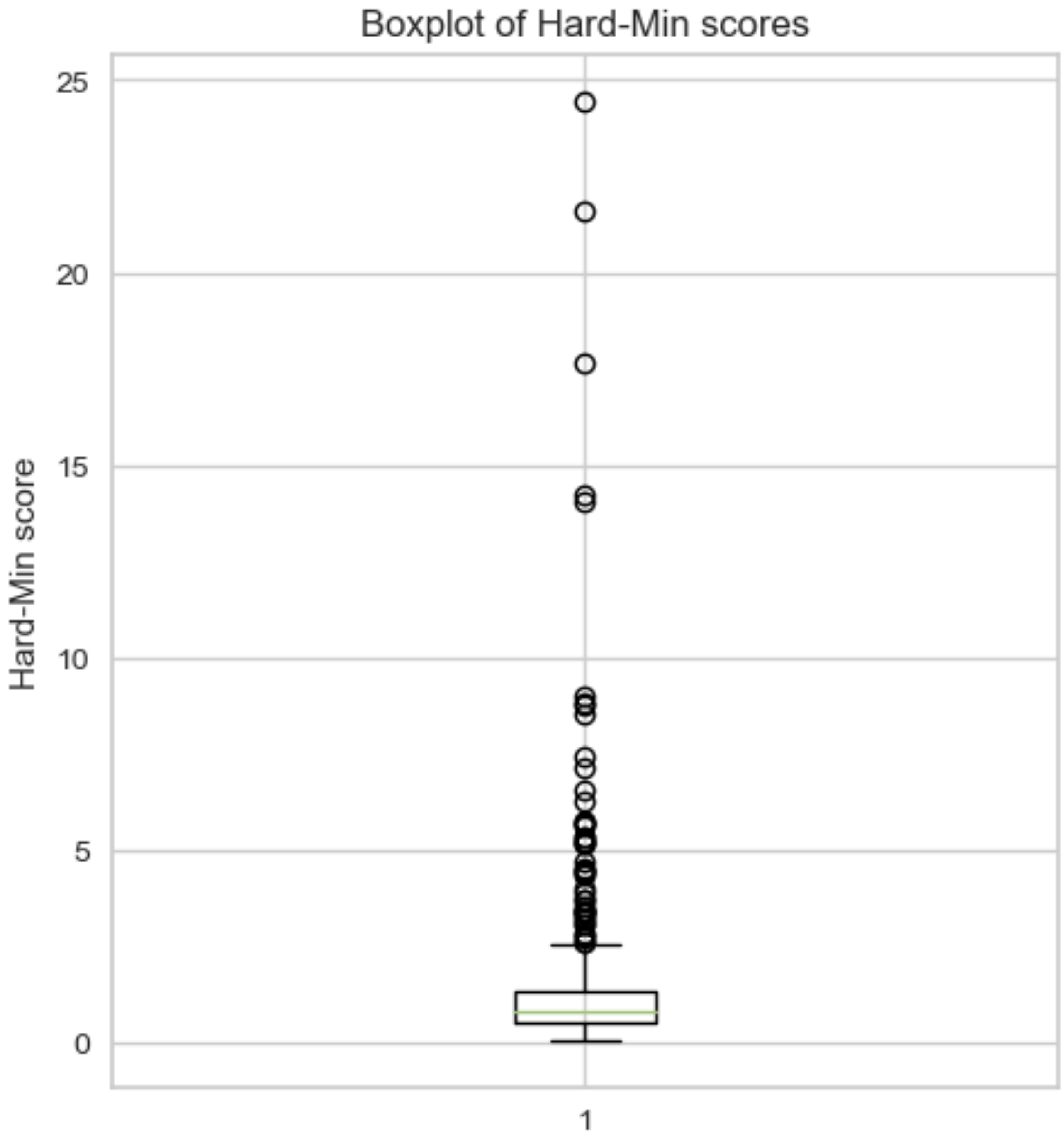
2.1. Hard-Min Score

Considering 10 % Most Extreme Points as Outliers

- 44 outliers with Hard-Min score above 3 (elbow)



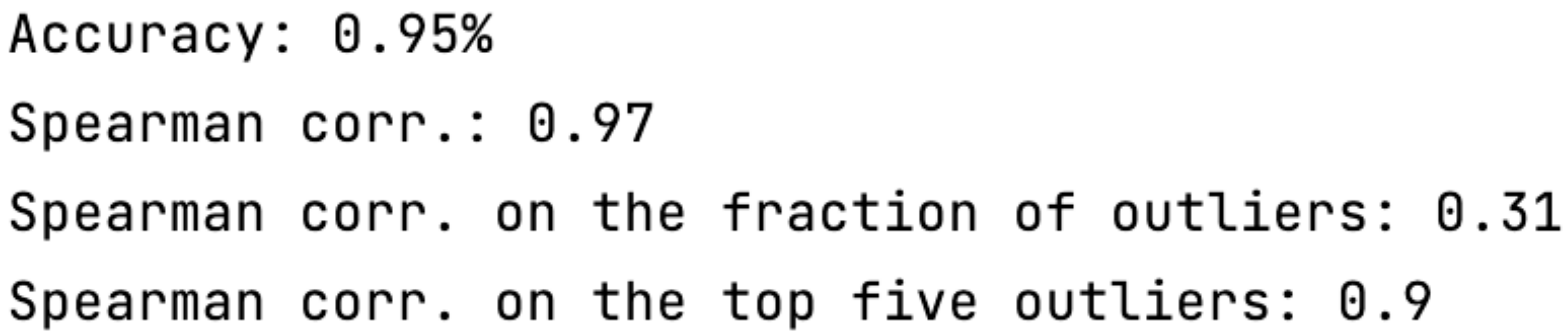
- 51 extreme values in the Boxplot



2.1. Hard-Min Score

Evaluation: Biasedness

- → Spearman’s ranking correlation
- → Accuracy of classifying the same set of outliers



```
Accuracy: 0.95%
Spearman corr.: 0.97
Spearman corr. on the fraction of outliers: 0.31
Spearman corr. on the top five outliers: 0.9
```

2.2. Soft-Min Score

Measure outlierness based on multiple neighbours

- **Soft-Min** = related to log-likelihood, predicted by a kernel density estimator of the rest of the data
- **γ**: the inverse of the bandwidth or variance of the used Gaussian distributions → small γ leads to more robust estimates, but with the cost of introducing bias
- The Hard-Min and Soft-Min score distributions are similar, but they “operate” on different scales.
- → comparison challenging
- Due to the 1/γ factor, anomaly scores decrease for increasing γ values
- → comparison between different γ values challenging

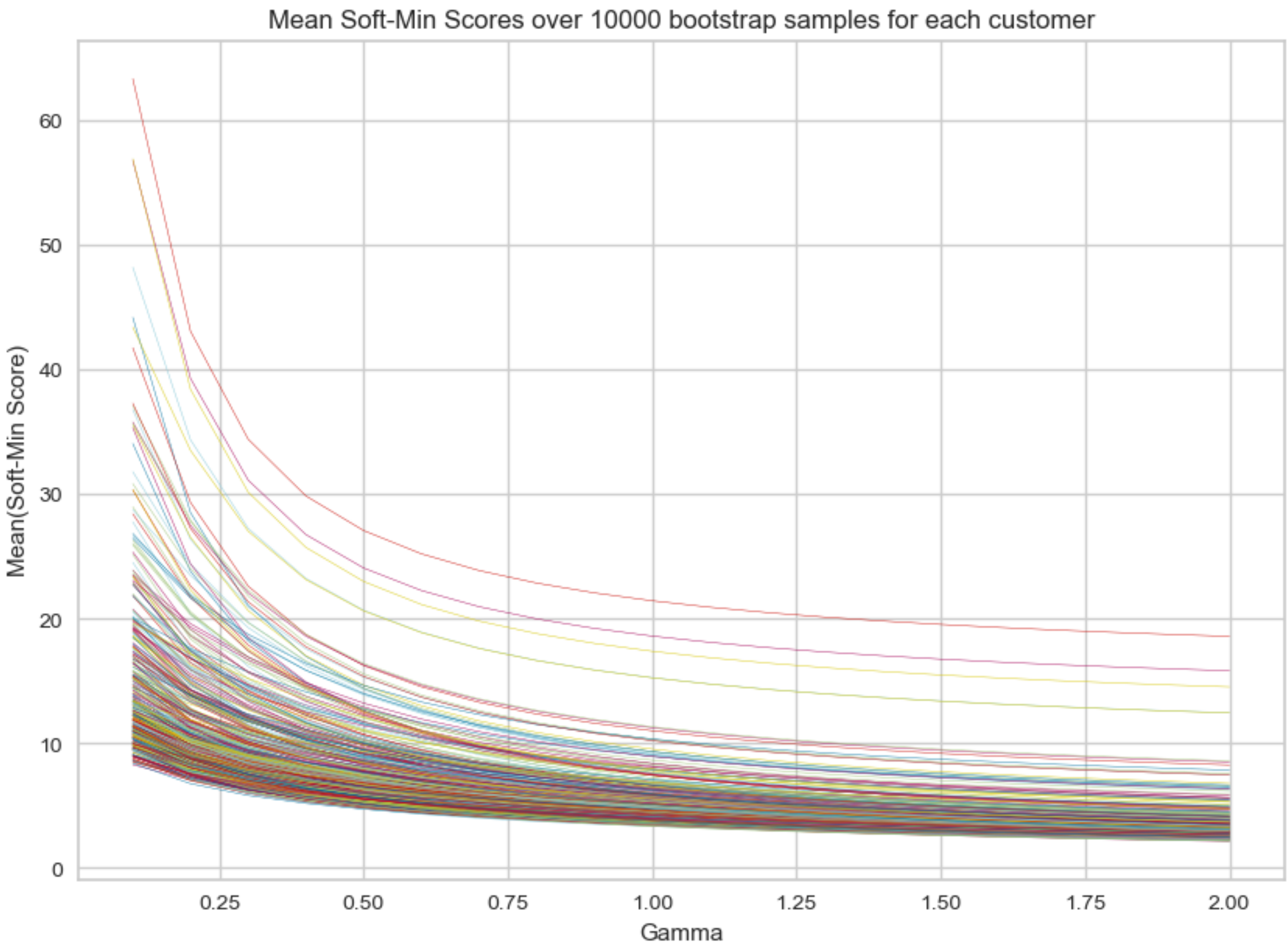
2.2. Soft-Min Score

Gamma Tuning

- Apply **bootstrapping** with replacement and compute Soft-Min scores of 20 γ values in the range [0.1, 20) for each sample.
- Average over the Soft-Min scores per sample → 440 x 10000 x 20 measurements

2.2. Soft-Min Score

Evaluation: Between Instance Variance

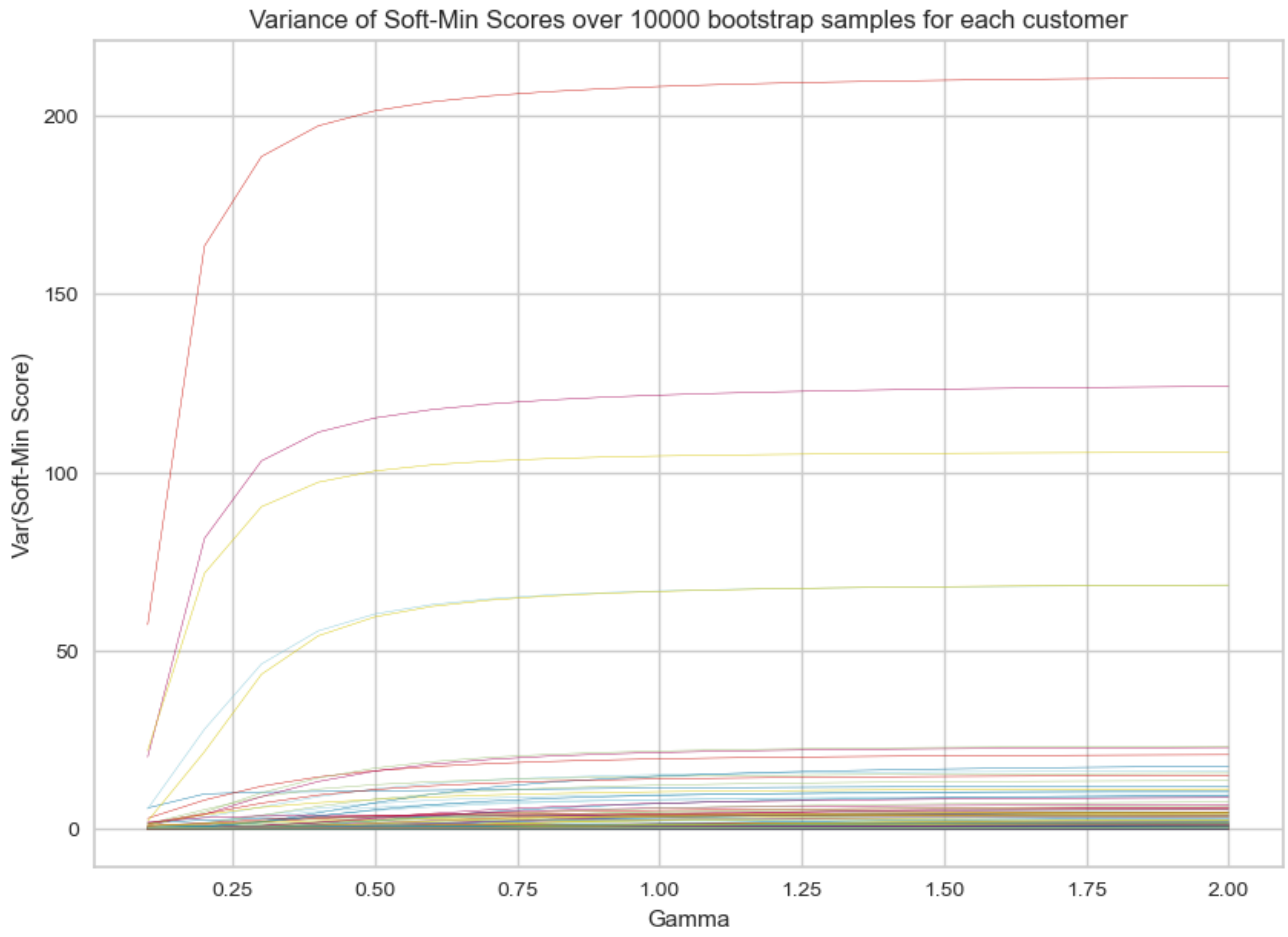


- Soft-Min scores reduce with increasing γ values
- The ranking appears to not change much
- **→ Not a good measure for discriminating ability**

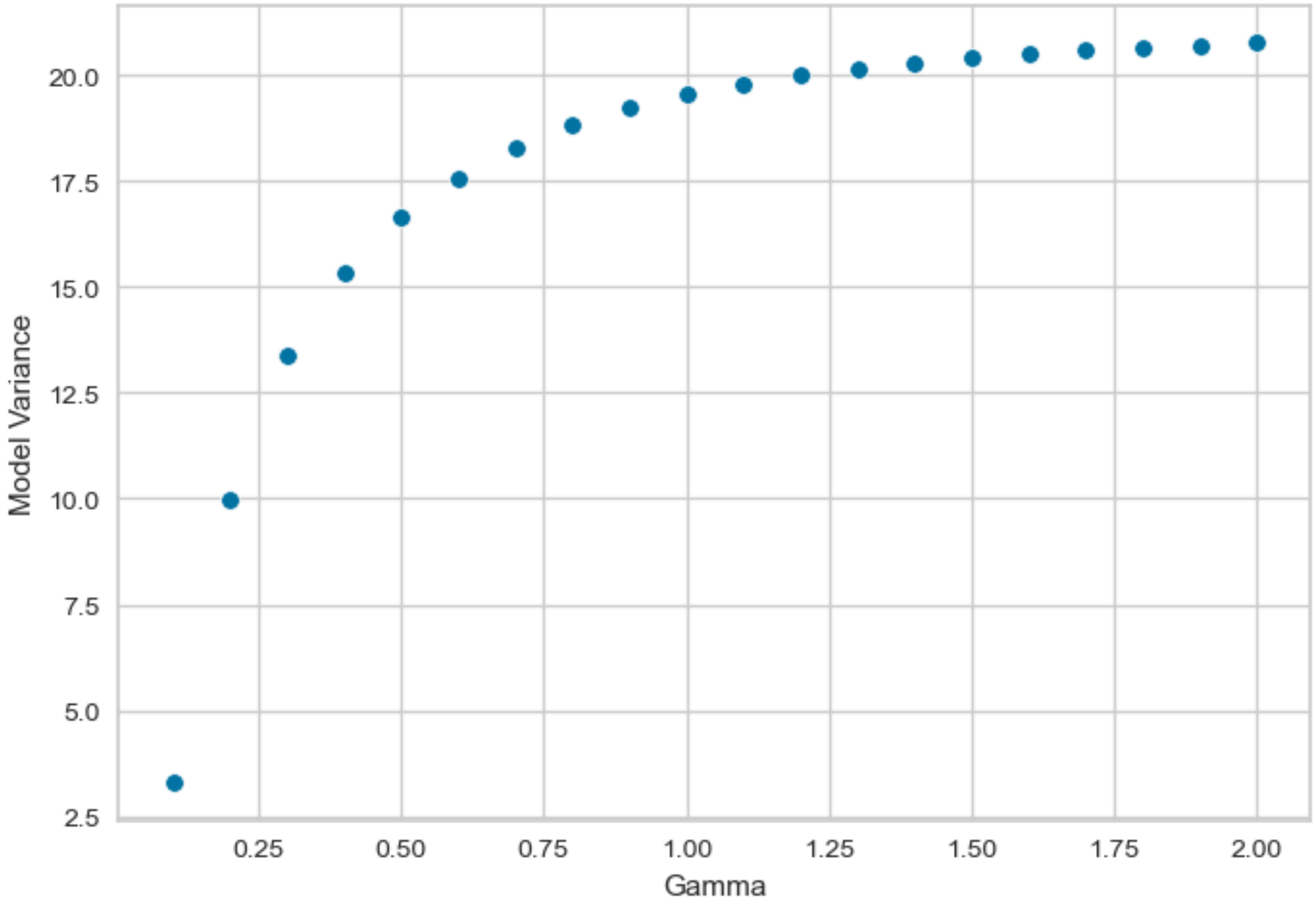
2.2. Soft-Min Score

Evaluation: Spread (Within Instance Variance)

- The variance of the model increases with increasing γ values



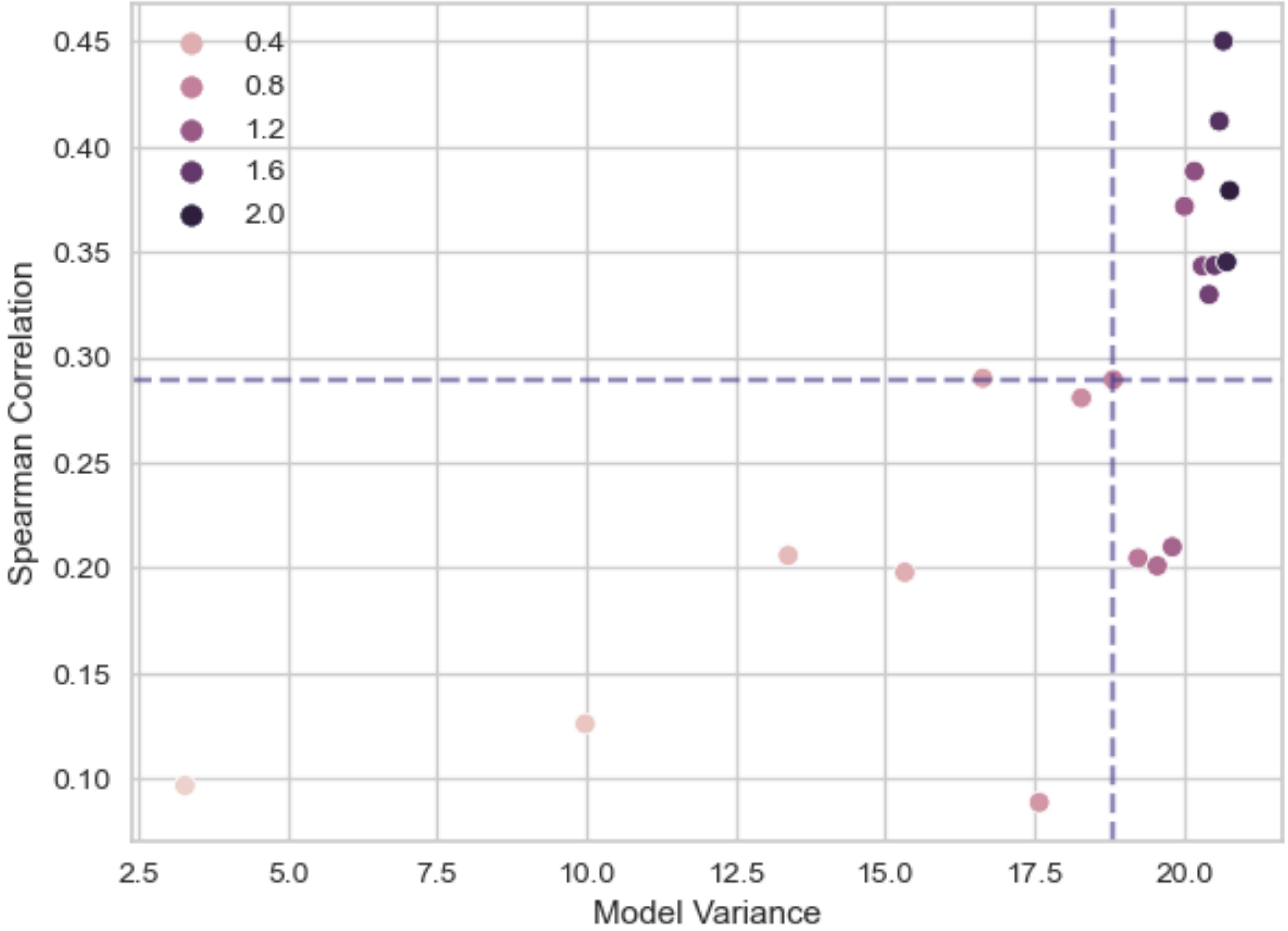
- Average over the variance of the outliers → spread evaluation metric

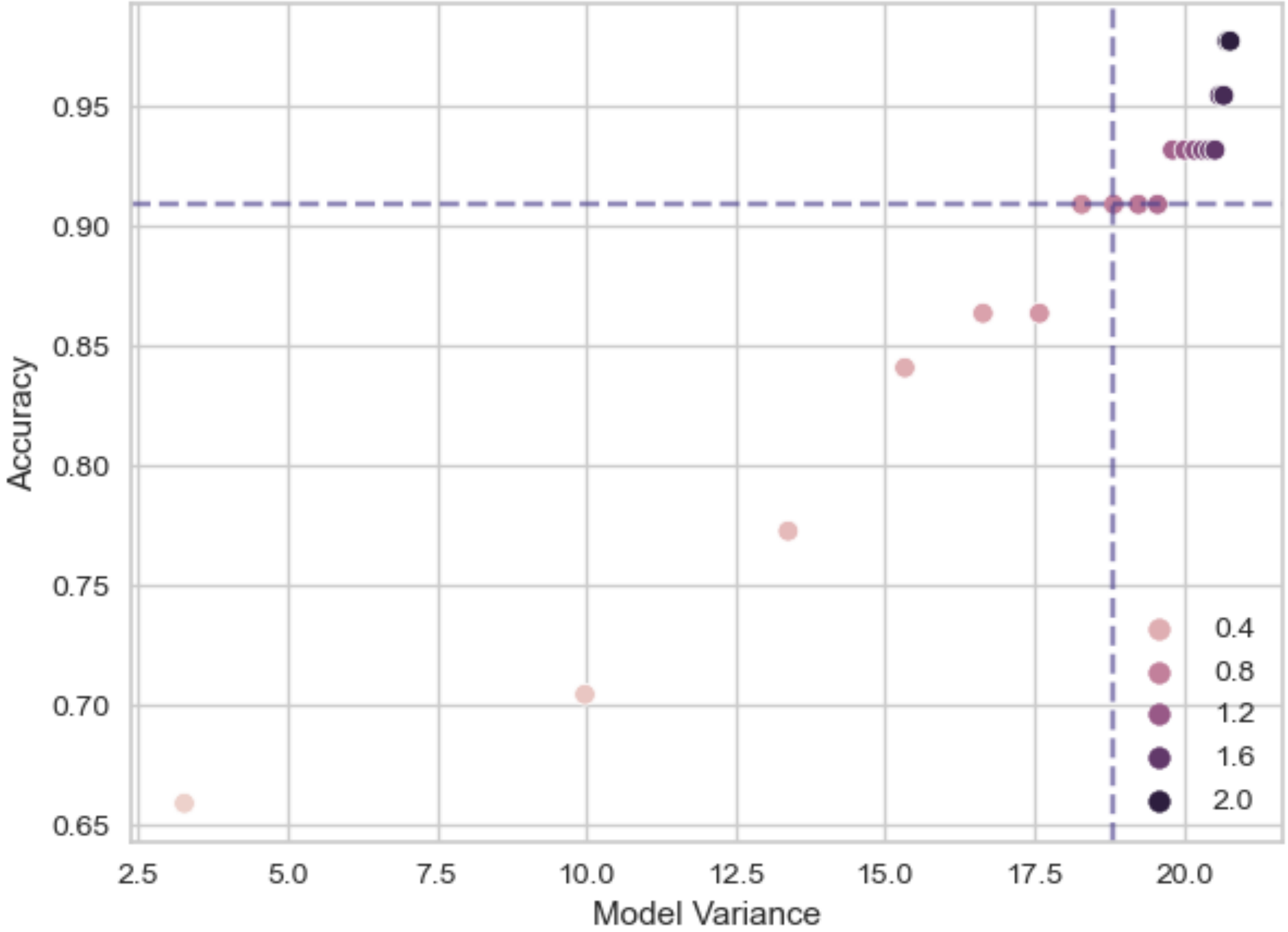


2.2. Soft-Min Score

Gamma Choice: γ = 0.8

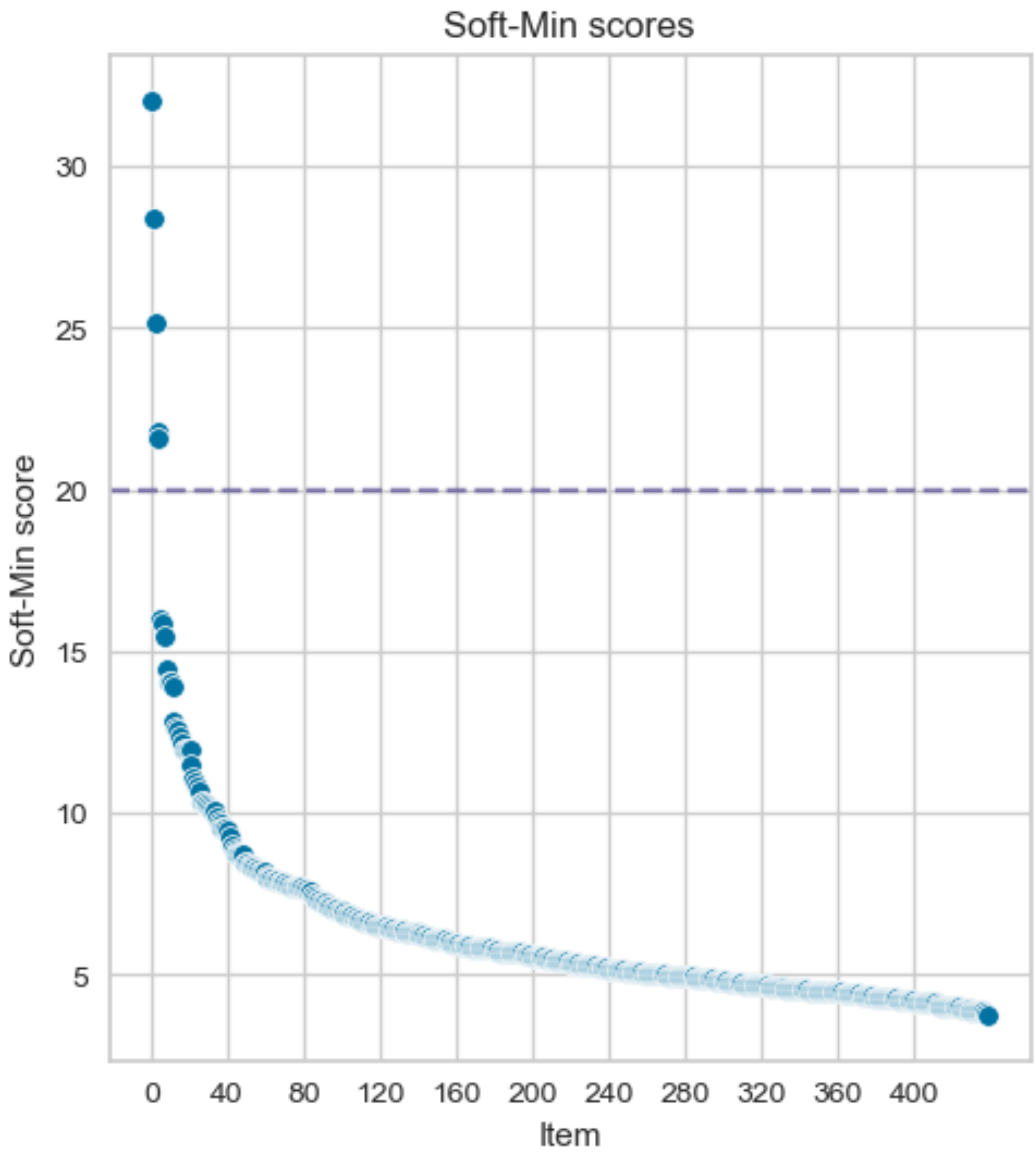
```
Accuracy: 0.91%
Spearman corr.: 0.88
Spearman corr. on the fraction of outliers: 0.41
Spearman corr. on the top five outliers: 0.9
```

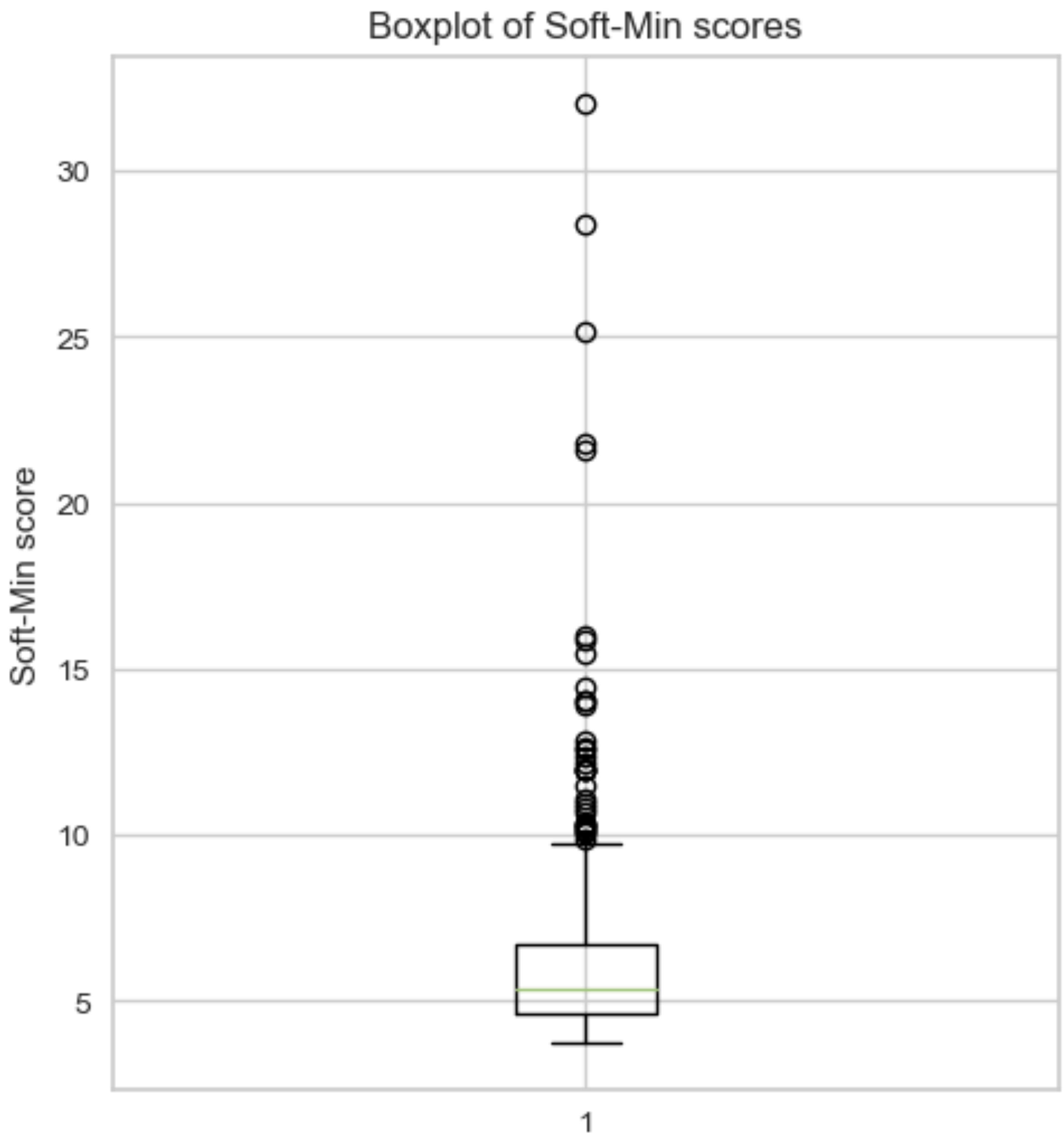




2.5. Outlier Selection

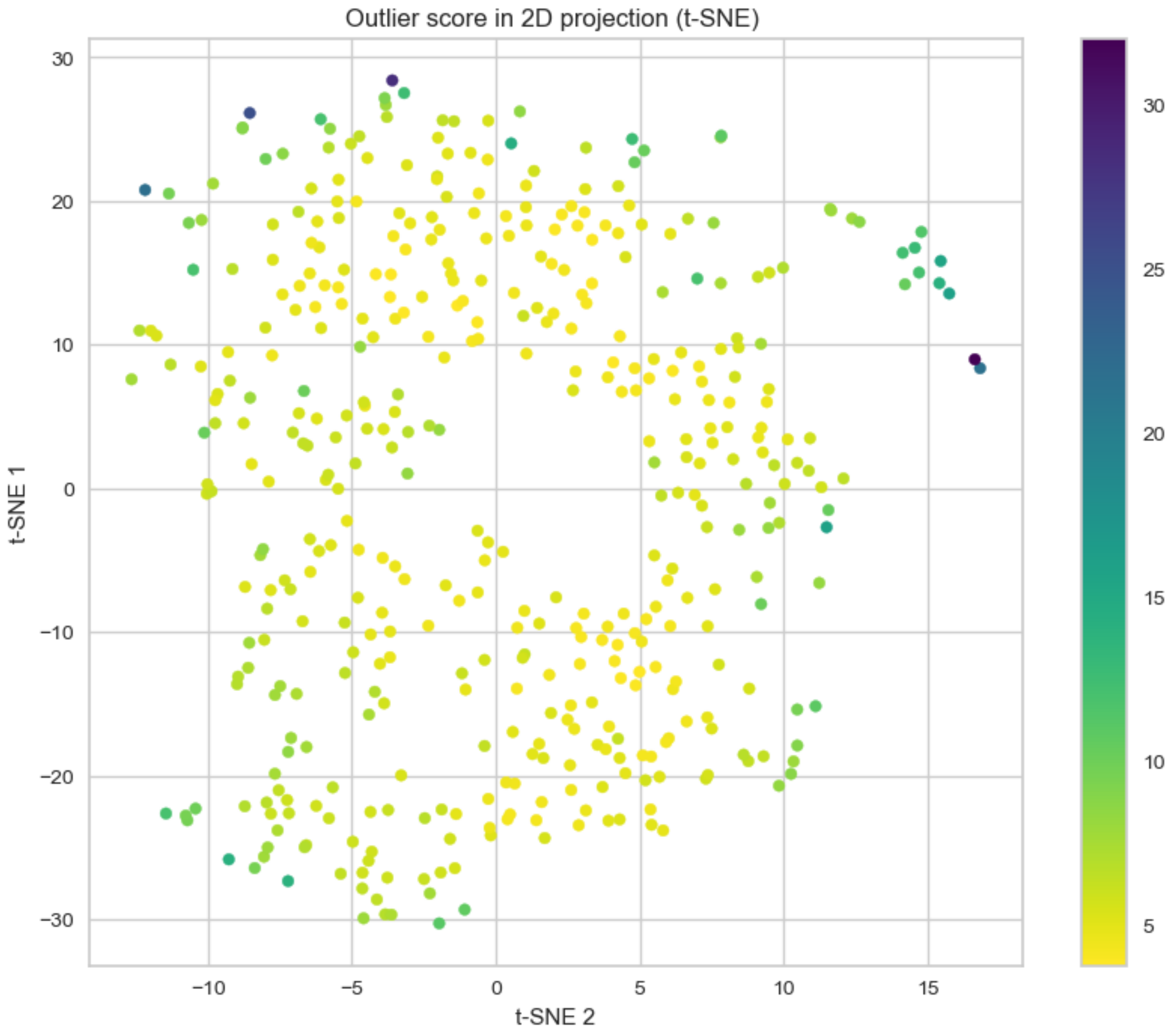
```
5 outliers above 20
8 outliers above 15
35 outliers above 9.71
```

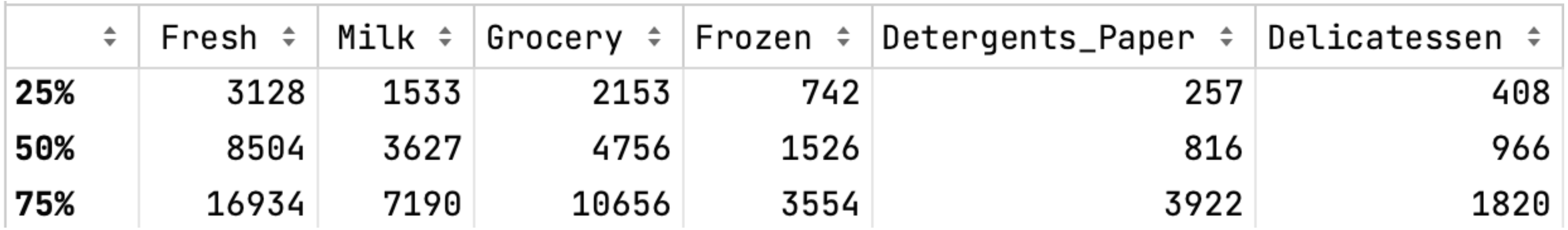




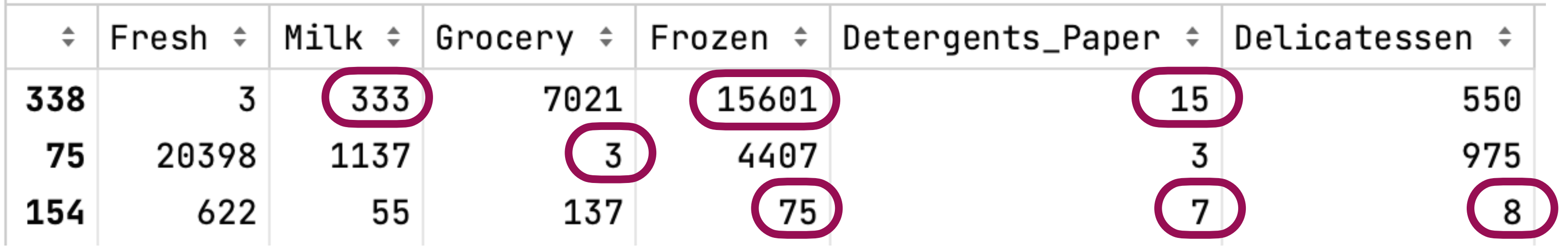
3. Explaining Anomalies

3.1. Layer-wise relevance propagation

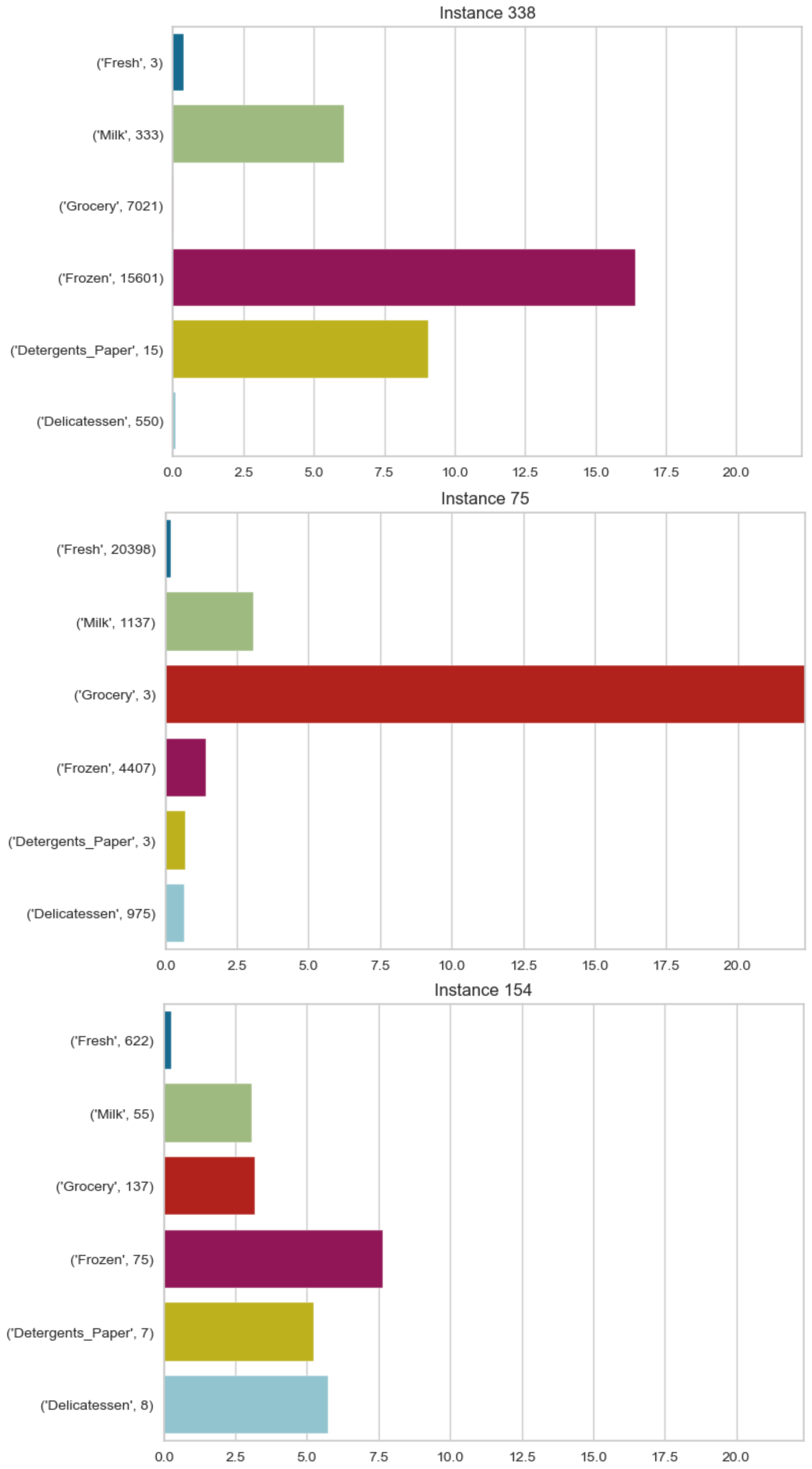




	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
25%	3128	1533	2153	742	257	408
50%	8504	3627	4756	1526	816	966
75%	16934	7190	10656	3554	3922	1820

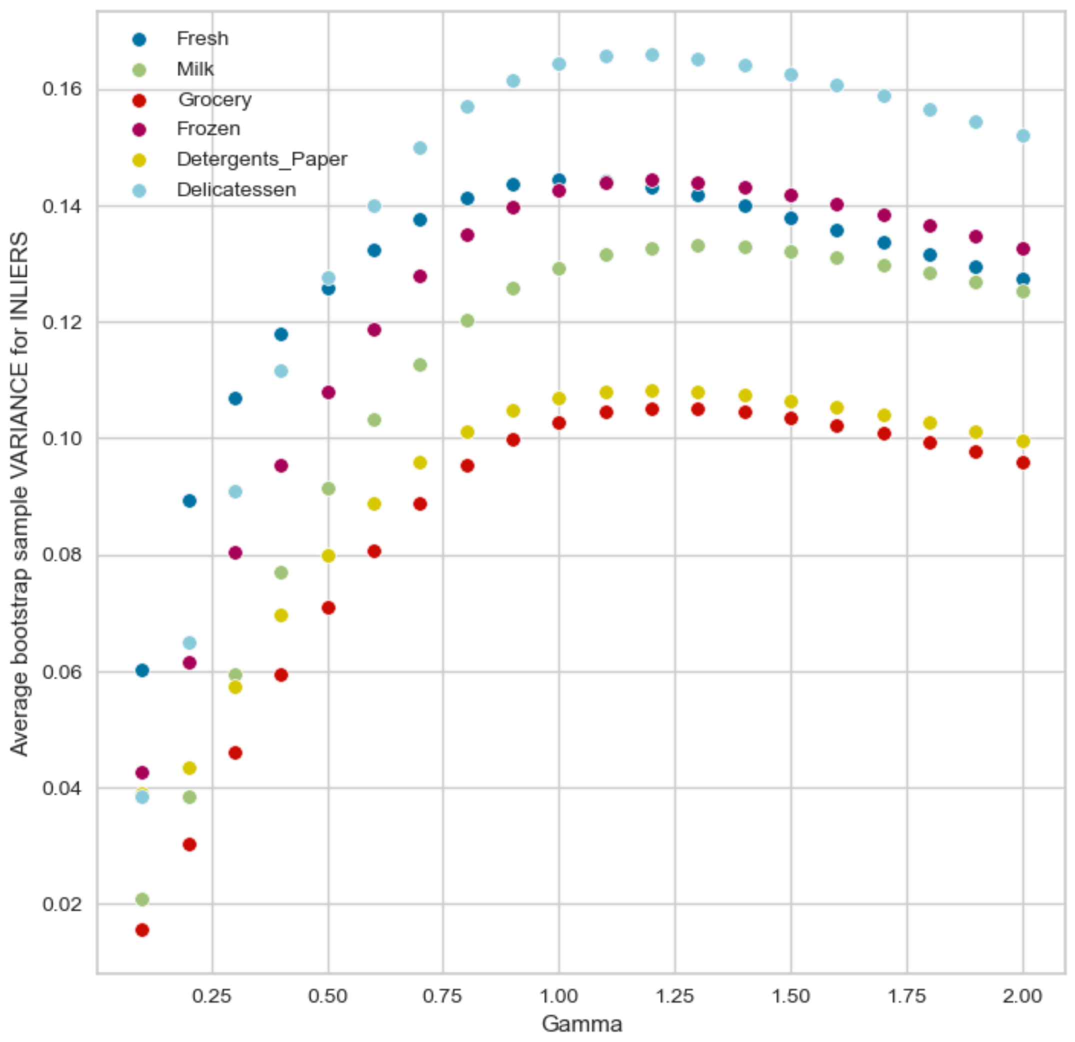


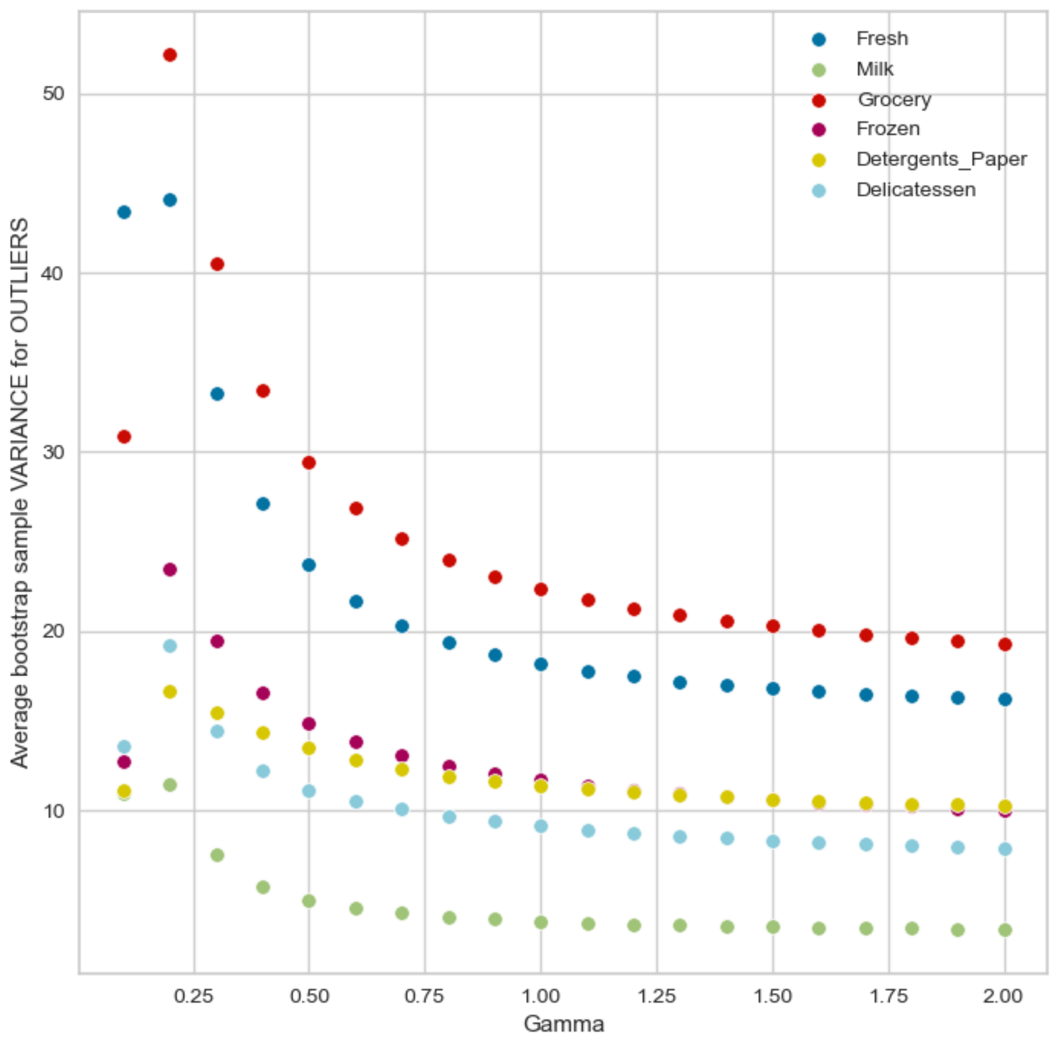
	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
338	3	333	7021	15601	15	550
75	20398	1137	3	4407	3	975
154	622	55	137	75	7	8



3.2. Spread and Biasednessof the explanations

- Bootstrapping with replacement → 440 x 1000 x 20 x 6 measurements





```
Fresh: all data: 0.95
       outliers: 0.7
Milk: all data: 0.96
       outliers: 1.0
Grocery: all data: 0.91
       outliers: 0.9
Frozen: all data: 0.98
       outliers: 0.9
Detergents_Paper: all data: 0.96
       outliers: 0.9
Delicatessen: all data: 0.98
       outliers: 0.9
```

4. Cluster Analysis

4.1. K parameter for K-means

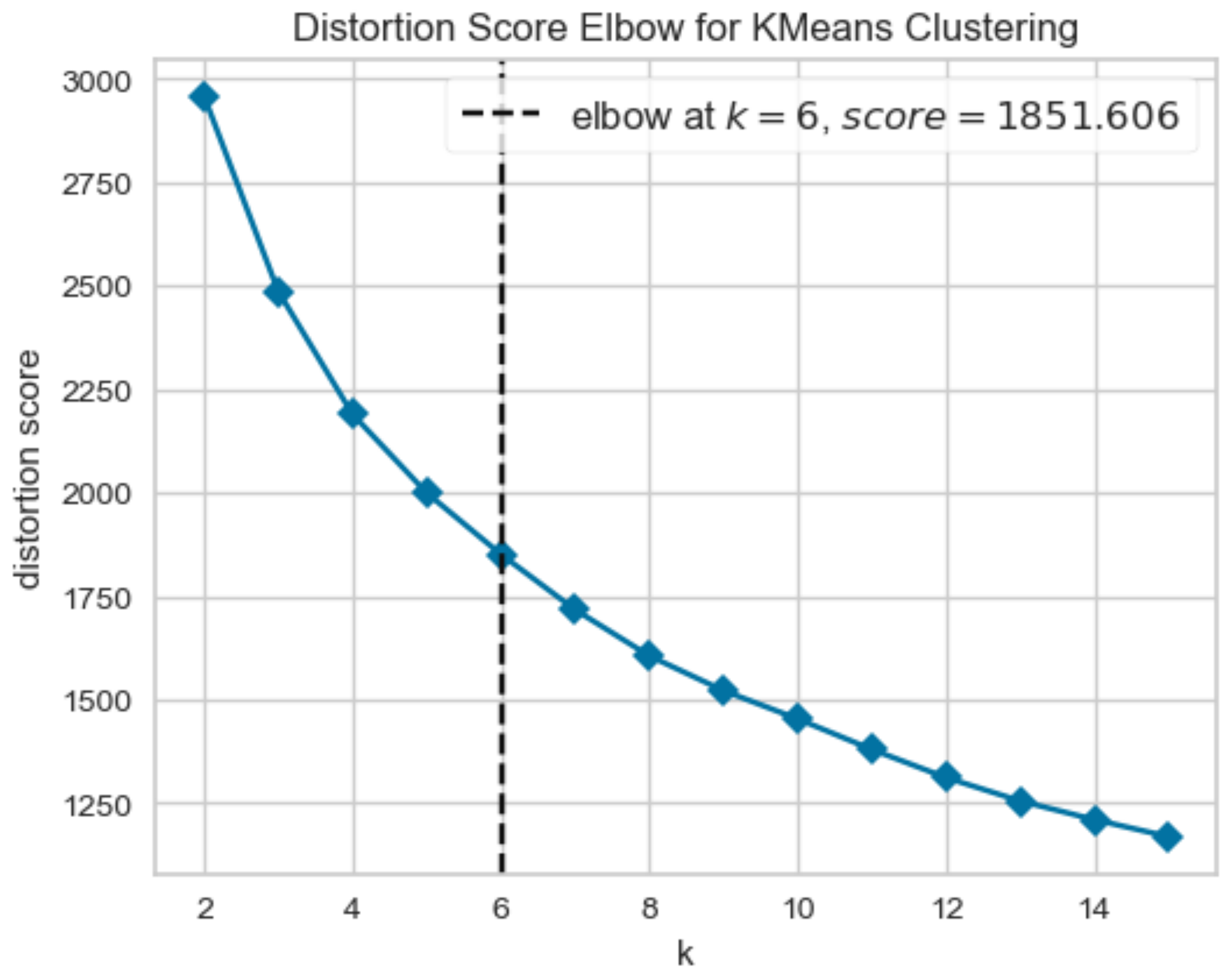
Intro

- No natural cluster formations
- → Apply K-means clustering algorithm with greedy k-means++ algorithm over 100 initialisations
- Goal: partition customers into groups of similar size that share tendencies in their purchases

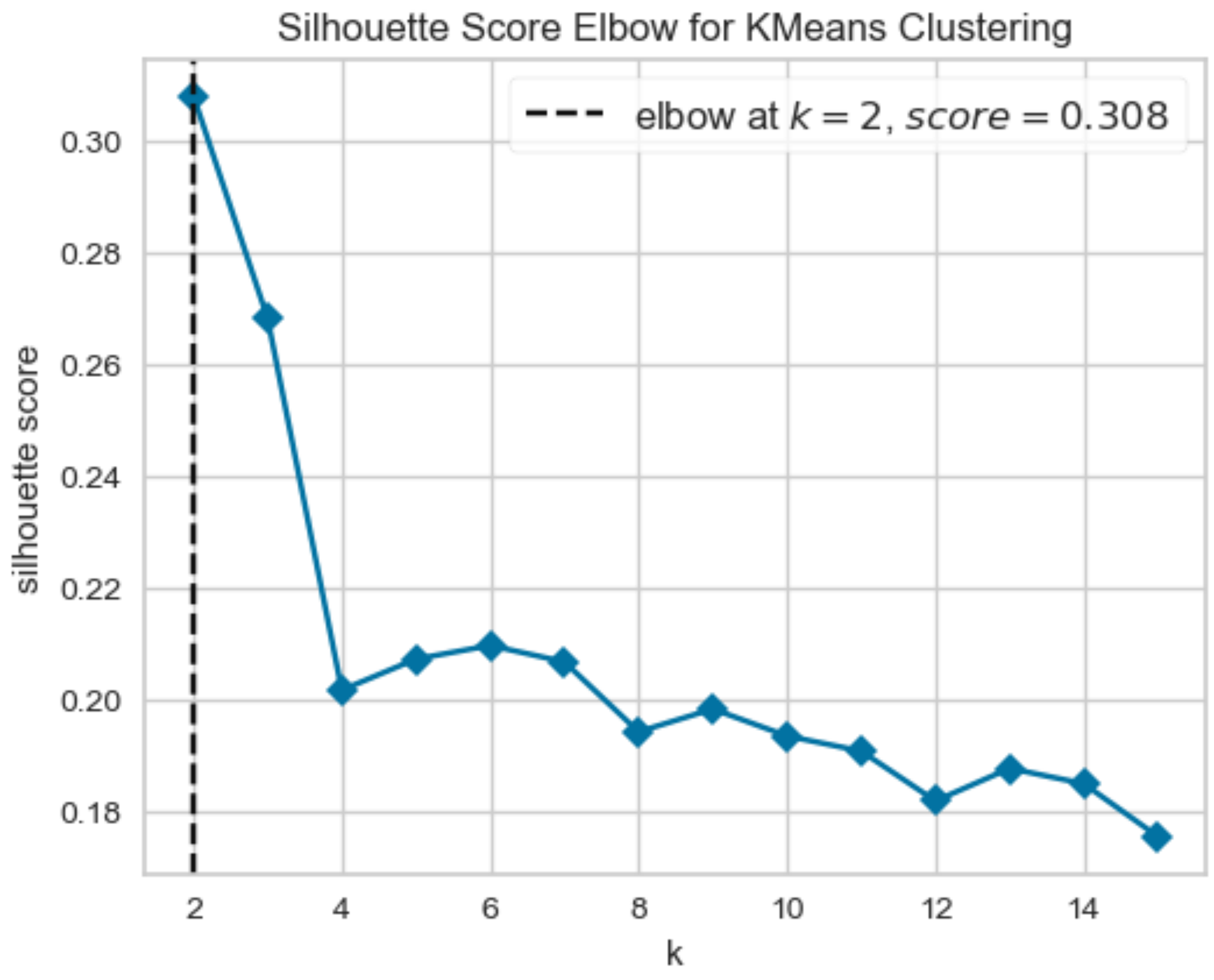
4.1. K parameter for K-means

Optimal Inflection Point for K in [2, 15]

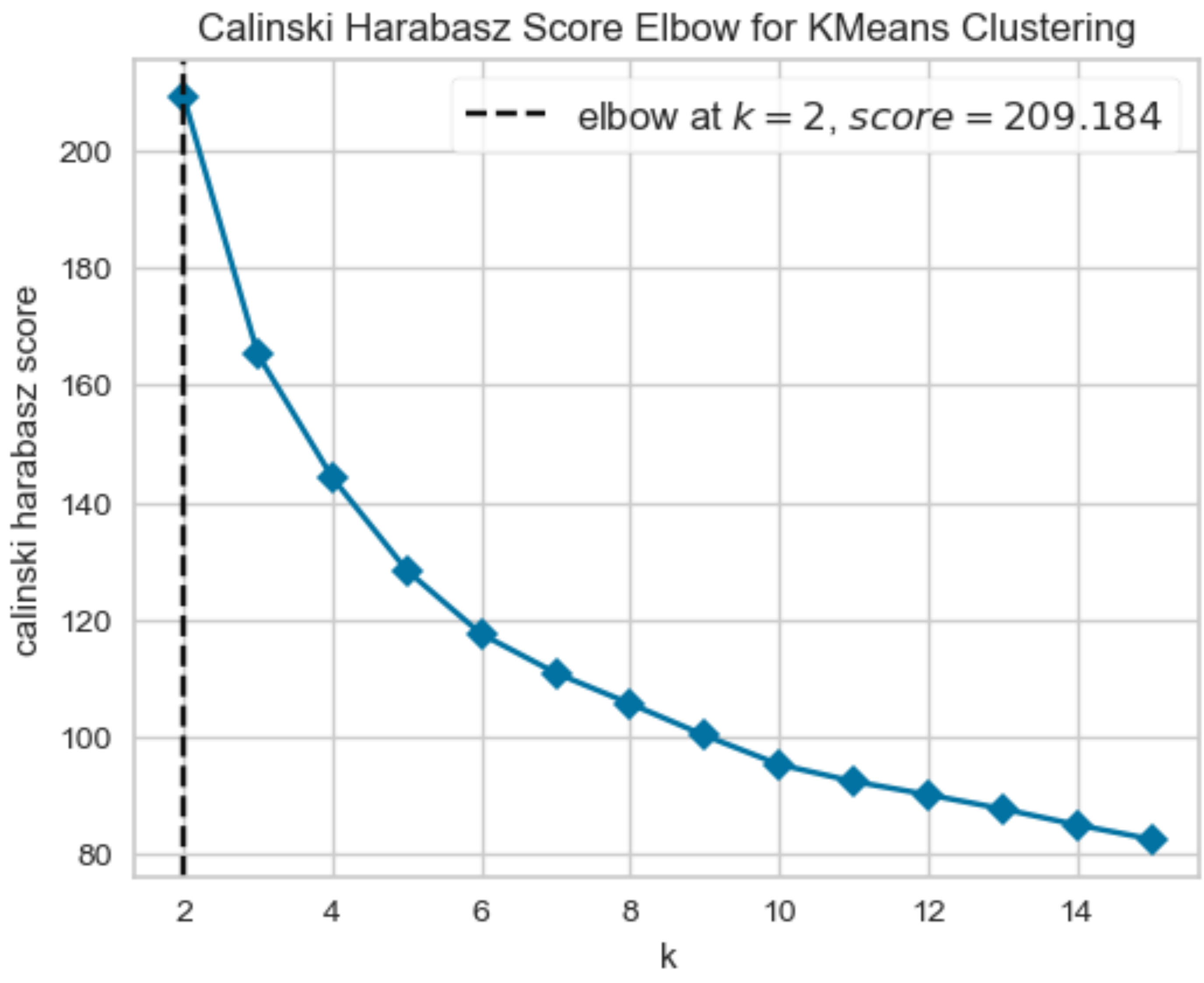
- **Elbow (distortion score)**: the sum of squared distances from each point to its assigned center
- **Silhouette score**: the mean Silhouette Coefficient of all samples
- **Calinski-Harabasz score**: the ratio of dispersion between and within clusters



Distortion Score Elbow for KMeans Clustering



Silhouette Score Elbow for KMeans Clustering

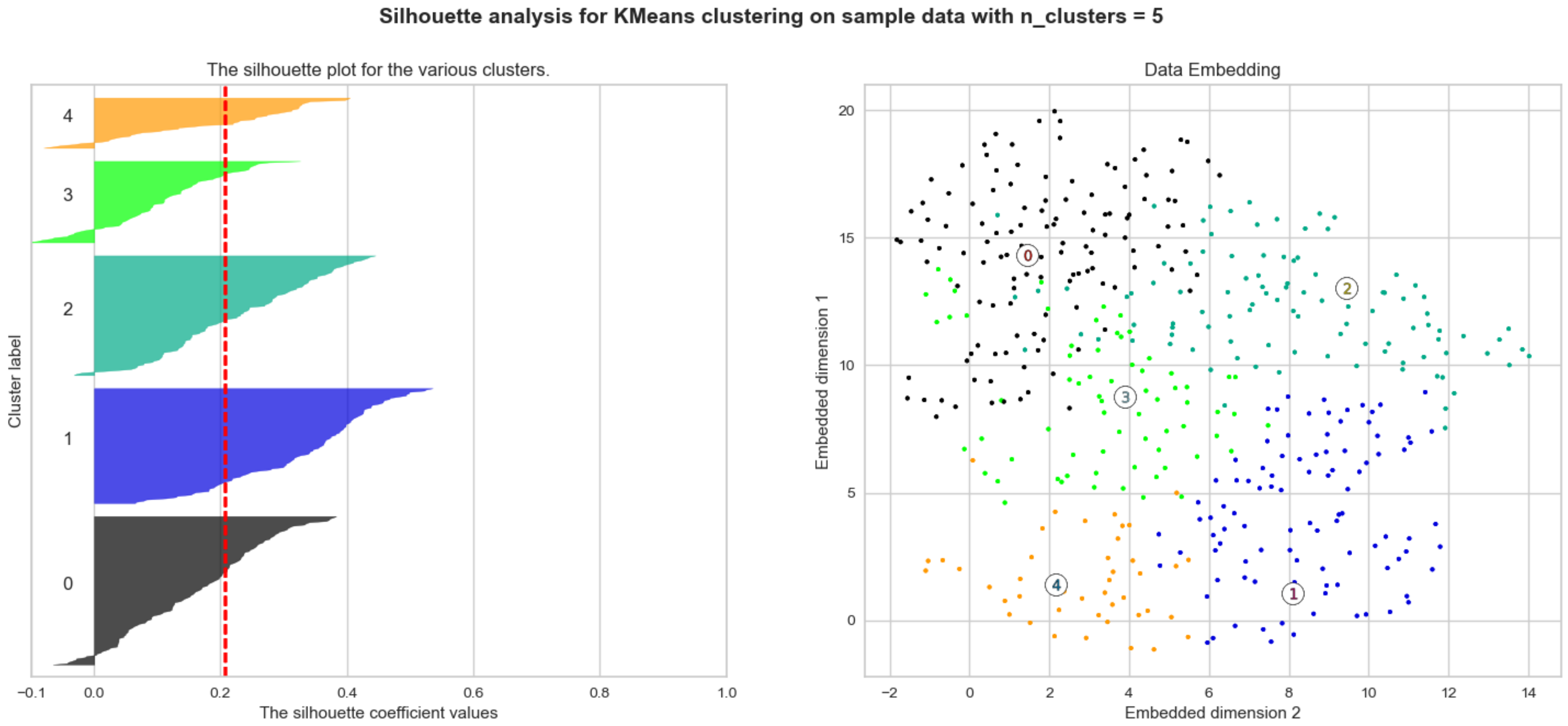


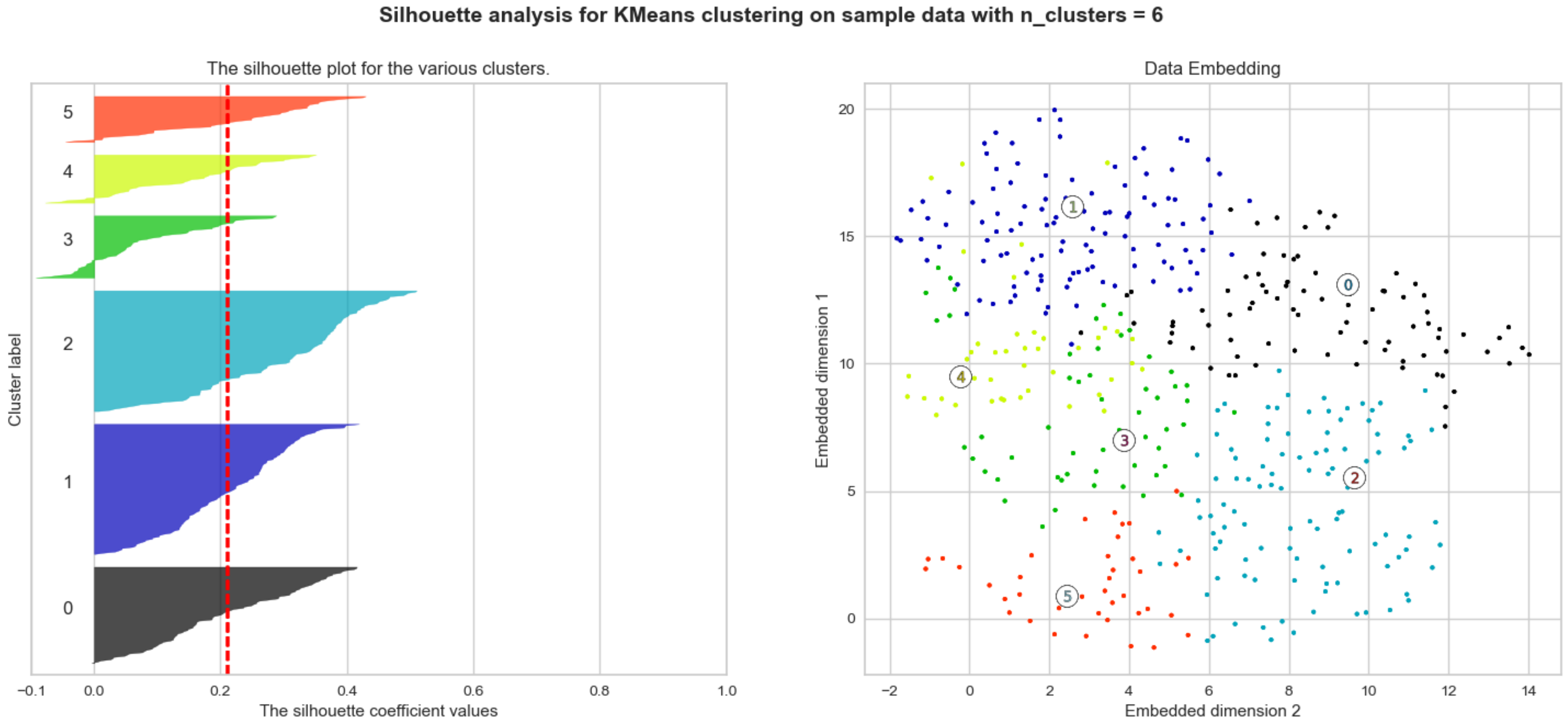
Calinski Harabasz Score Elbow for KMeans Clustering

4.1. K parameter for K-means

Silhouette Plots for K in range [5, 6]

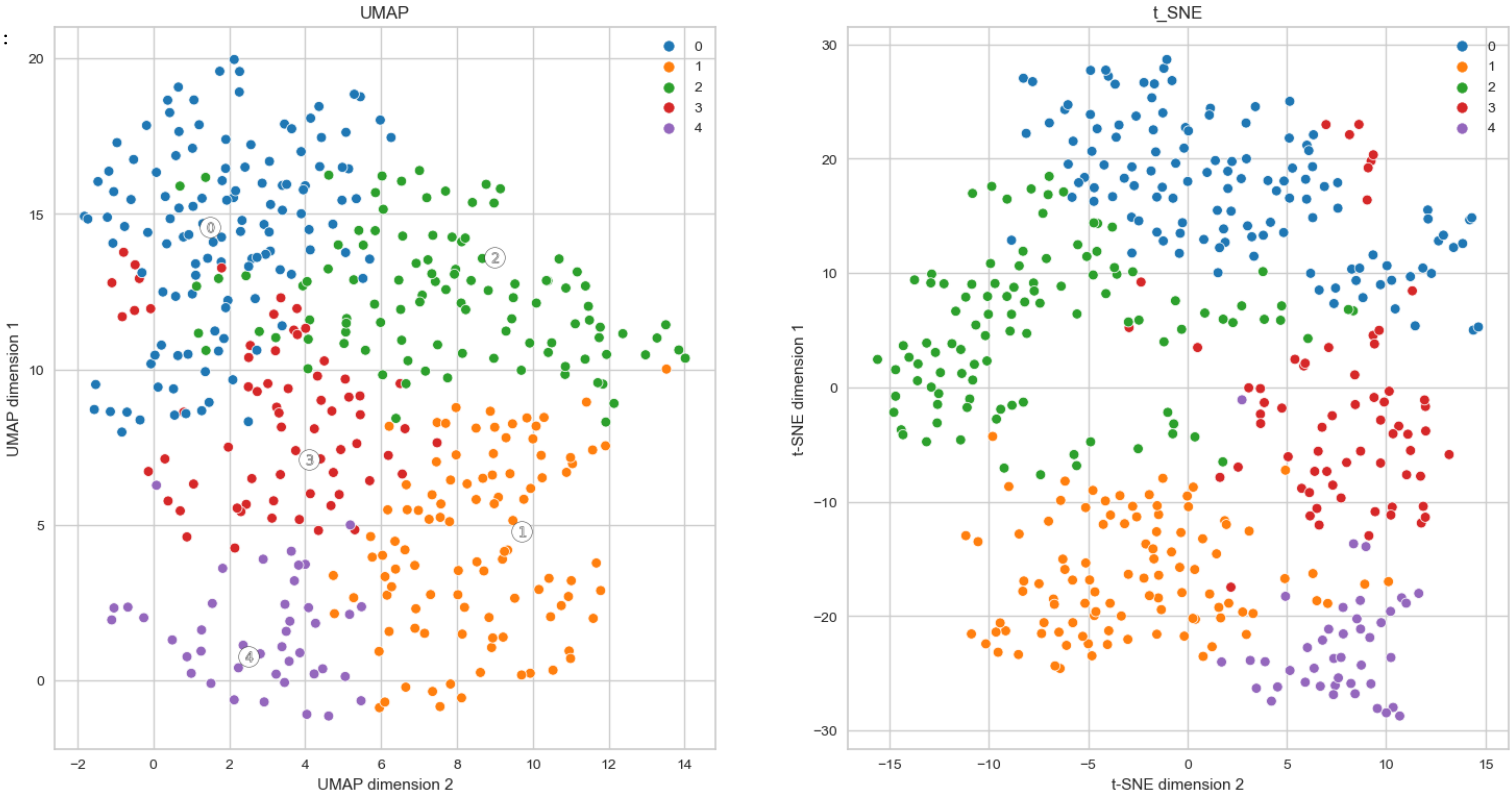
```
For n_clusters = 5 The average silhouette_score is : 0.20699
For n_clusters = 6 The average silhouette_score is : 0.21196
```



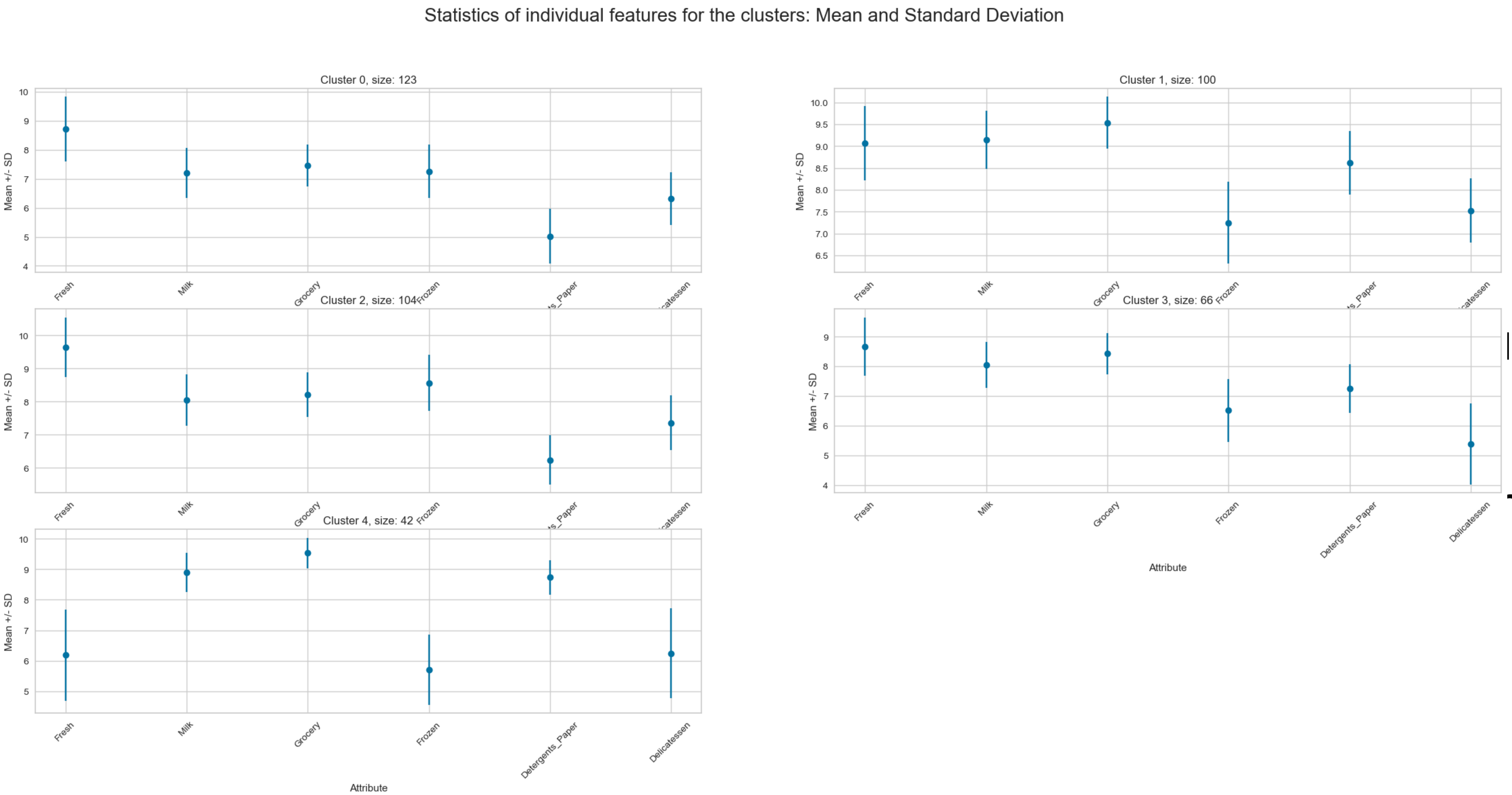


4.2. Clustering

```
k-Means cluster sizes:
Cluster: 0 : 123
Cluster: 1 : 100
Cluster: 2 : 104
Cluster: 3 : 66
Cluster: 4 : 42
```



4.3. Interpretation of the Clustering

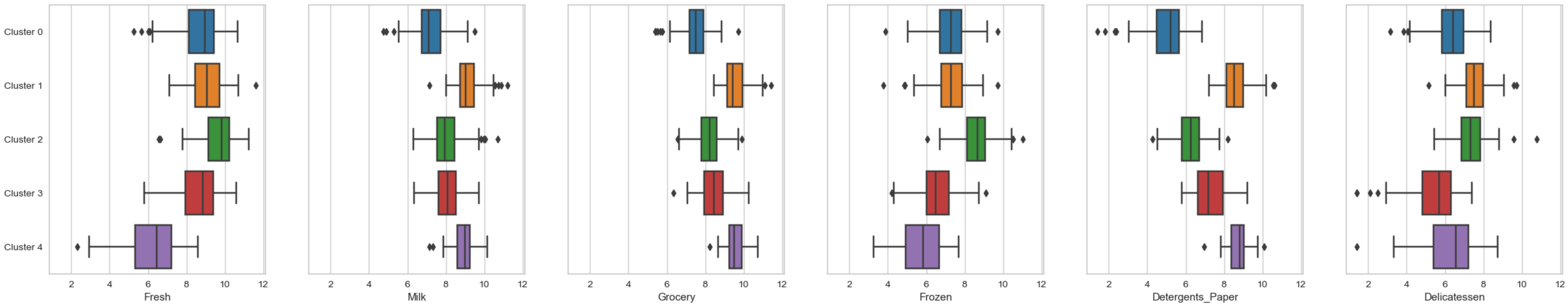

Statistics of individual features for the clusters: Mean and Standard Deviation

- For ar ster memb

- We cc n each cluste

4.3. Interpretation of the Clustering

Statistics of individual features for the clusters: Boxplots


Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Fresh
Milk
Grocery
Frozen
Detergents_Paper
Delicatessen