ISSS602_AY2025-26Aug_Assign2_ES

Creation Date: Sunday, October 12, 2025, 01:03:39 PM

Author: edward.lim.2025

10 Pages

Cover Page

Executive Summary_v.1
Creation Date: Sunday, 28 Sep, 2025
Author: edward.lim.2025

ISSS602 Data Analytics Lab Assignment 2: Hospitality Segmentation with Cluster Analysis

---

Objective:

To perform cluster analysis by using at least three different clustering methods with a focus on supply side by segmentation of hotels.

Visuals:

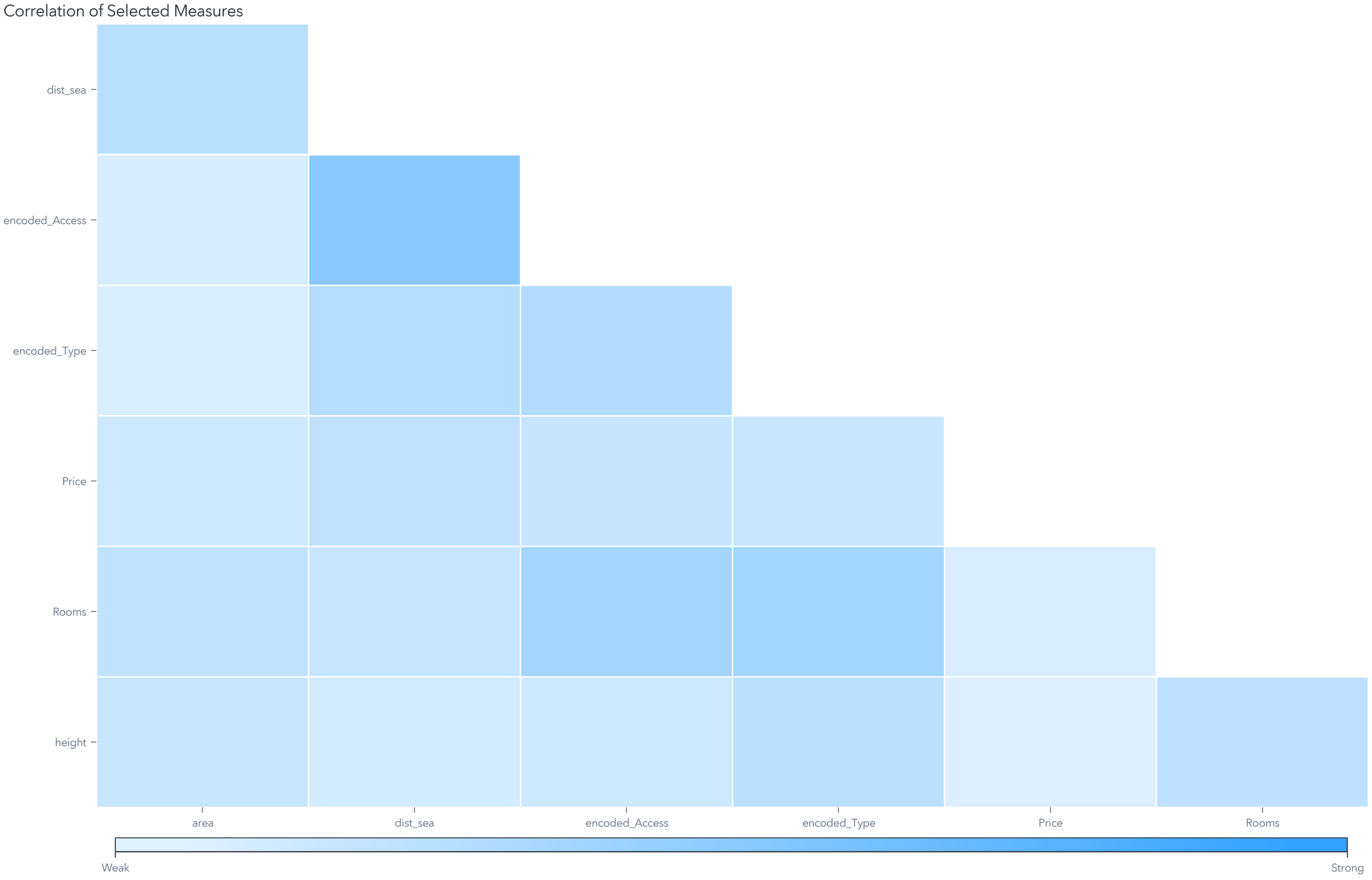
Visualization techniques such as correlation matrix, bar charts, histogram, pie chart and box plots are employed.

Data:

Vilane Gonçalves Sales (2025) "A coastal hospitality sector database for vulnerability assessments in Veneto and Emilia-Romagna, Italy" Data in Brief, Vol. 62.

Clustering Models:

1. K-means Clustering
2. Hierarchical Clustering – Ward's minimum distance
3. Hierarchical Clustering – Complete linkage (Maximum linkage)


Correlation of Selected Measures
dist_sea
encoded_Access
encoded_Type
Price
Rooms
height
area
dist_sea
encoded_Access
encoded_Type
Price
Rooms
Weak
Strong

Cluster Model 1: Kmeans clustering

Number of Clusters: 6
Variables Used for Clustering: price, height, dist_sea, area, room.

Cluster 1:

-Cluster 1 contains the largest number of observations at **727**.
-It has an **RMS Std Dev of 0.43**, indicating a tighter and more homogeneous cluster compared to the others.
-The **Within-Cluster SS** is the highest at 135.9, primarily because of the large cluster size.
-RMS Std Dev is at 0.43, the lowest among all clusters, indicating that the observations within this cluster are relatively tight and homogeneous.
-The **nearest neighbour** to Cluster 1 is Cluster 6, with a centroid distance of **0.50**, suggesting that both clusters may share similar profiles.

Cluster 2:

-The **RMS Std Dev is 0.58**, indicating that this cluster is more dispersed compared to the others.
-The **Within-Cluster SS is 131.9**, the second highest after Cluster 1, despite having nearly half the number of observations (382).
-This higher value is likely due to the presence of **outliers/ positive skewness** in area, Room, and Price.
-Although the nearest cluster to it is Cluster 1, the centroid distance is relatively high at 0.94.
-This difference is mainly attributed to variations in Rooms, height, and Price.

Cluster 3:

-The RMS Std Dev is 0.58, indicating that this cluster is only more dispersed compared to the others. Overall its still moderately tight.
-The **nearest neighbour to Cluster 3 is Cluster 5**, but the centroid distance is relatively high at 0.94.
-Cluster 3 is located **farther from the sea** and is characterized by larger area and taller height.
-The main differences between Cluster 3 and Cluster 5 are primarily attributed to variations in **height and number of rooms**.

Cluster 4:

-Cluster 4 has the smallest number of observations at **163**.
-The **RMS Std Dev is 0.54**, indicating a moderately tight and homogeneous cluster compared to the others.
-The nearest neighbour to Cluster 4 is **Cluster 6**, with a centroid distance of **0.59**, suggesting that these two clusters **may share similar profiles**.
-However, Cluster 4 shows **higher prices** than Cluster 6 (Fig 2.12), which may indicate that it represents higher-end or luxury accommodations located closer to the sea.

Cluster 5:

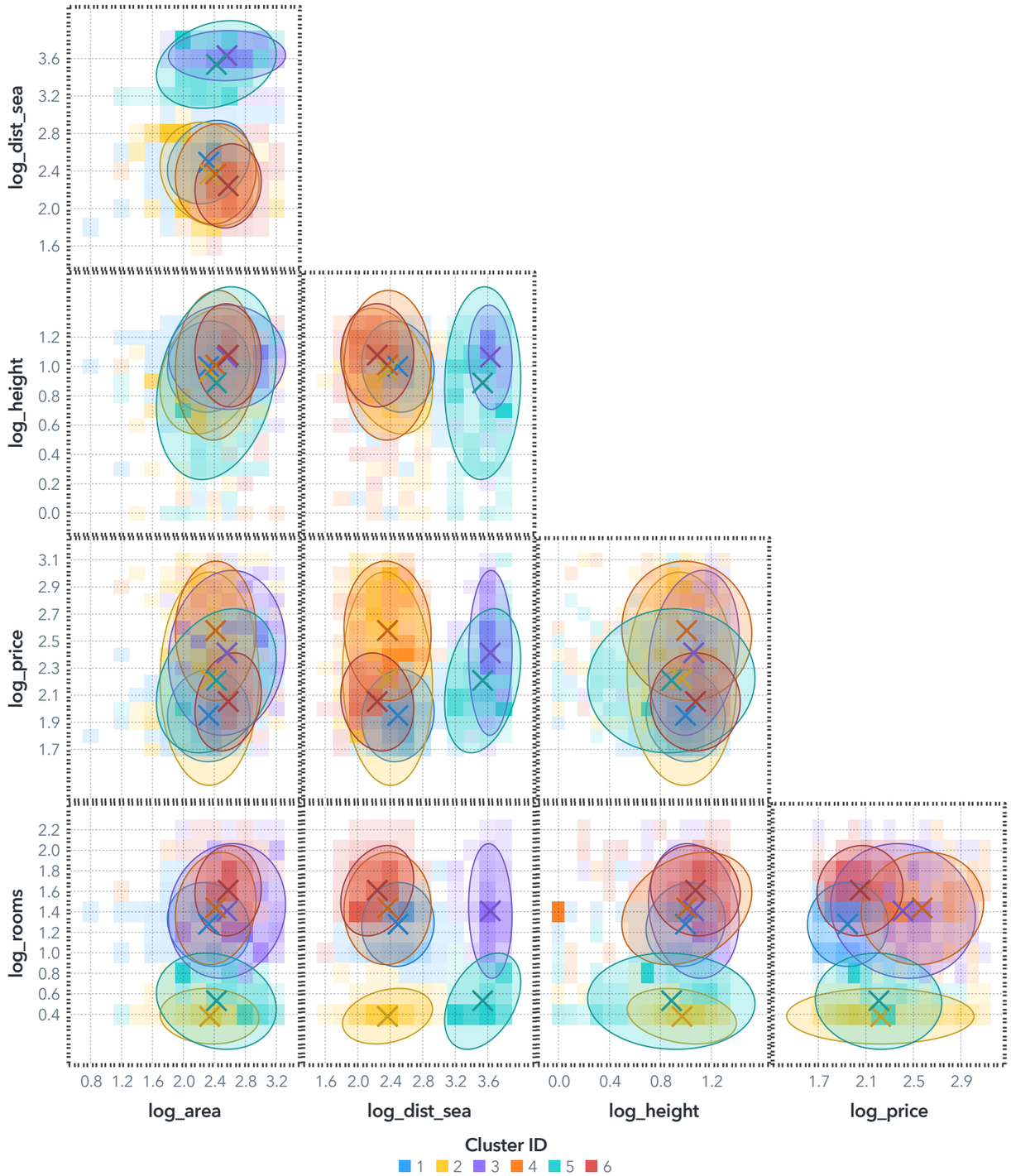
-The nearest neighbour to Cluster 5 is **Cluster 3**, with a relatively large centroid distance of **0.93**.
-Cluster 5 is mainly characterised by its **proximity to the sea**, **smaller number of rooms**, and generally **lower prices**.

Cluster 6:

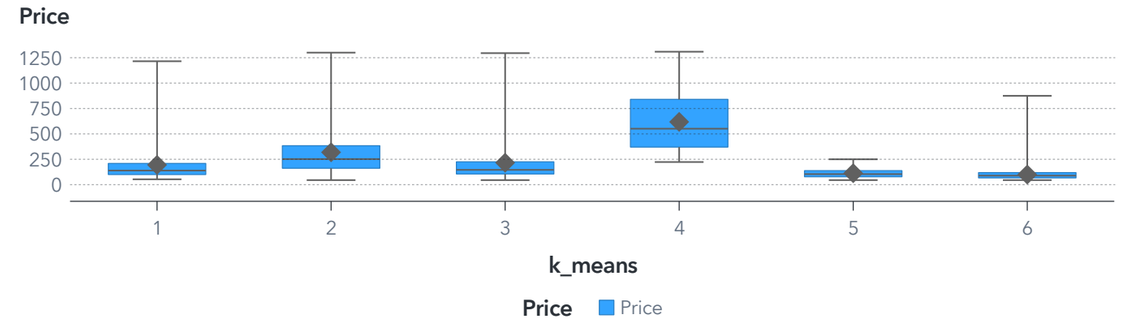
-Cluster 6 has **502** observations, making it the second largest cluster after Cluster 1.
-The nearest neighbour to Cluster 6 is **Cluster 1**, with a centroid distance of **0.50**.
-Both clusters share **similar characteristics**, with the main difference being Cluster 6 is located **close to the sea**.

Cluster

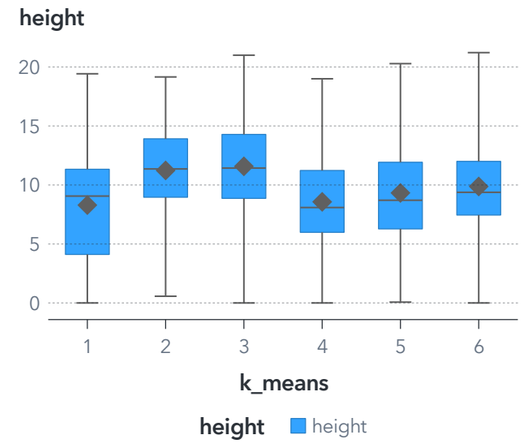
Observations: **2.3K of 2.3K** Polylines: **1.6K**



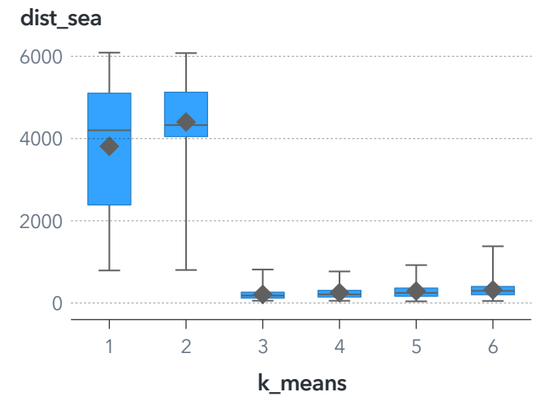
Price by k_means



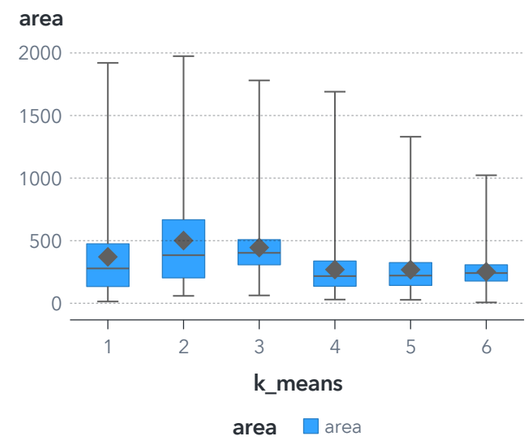
height by k_means



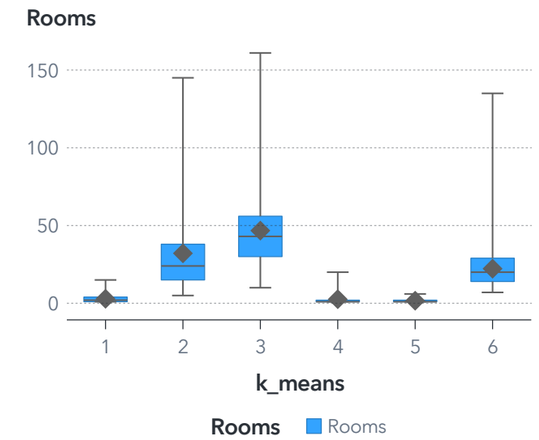
dist_sea by k_means



area by k_means



Rooms by k_means



Cluster Model 2: Hierarchical Clustering – Ward's minimum distance

Number of Clusters: 3
Variables Used for Clustering: price, height, dist_sea, area, room.
Category variable used: Access, Type

Cluster 4 is the largest, followed by Cluster 3 and Cluster 7.

CL3:

-This cluster is located furthest from the sea compared to the others.
-It has the highest median price among all clusters.
-Most accommodations in this cluster do not have direct beach access.
-Majority of properties are hotels and holiday rental homes.

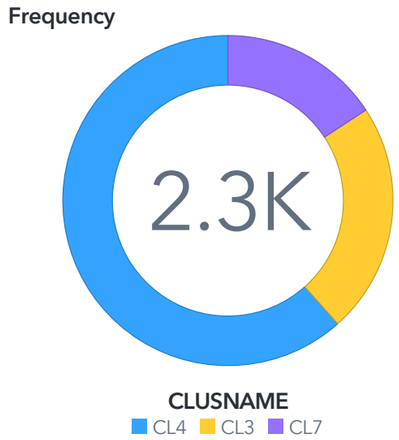
CL4:

-This cluster is located very close to the sea, similar to Cluster 7.
-It has the lowest median price among all clusters.
-Most accommodations have either public or private beach access.
-It also has the highest number of rooms compared to the other clusters.
-The majority are hotels, with a small number of properties classified as bed and breakfast or holiday rental homes.

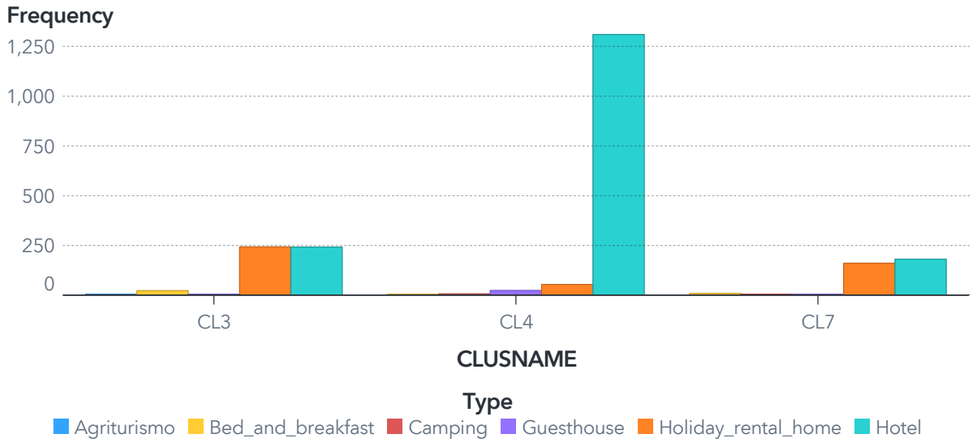
CL7:

-This cluster is located very close to the sea.
-It generally has a slightly shorter median height compared to the other clusters.
-Most accommodations have either public or no beach access.
-The Majority of properties are hotels and holiday rental homes.

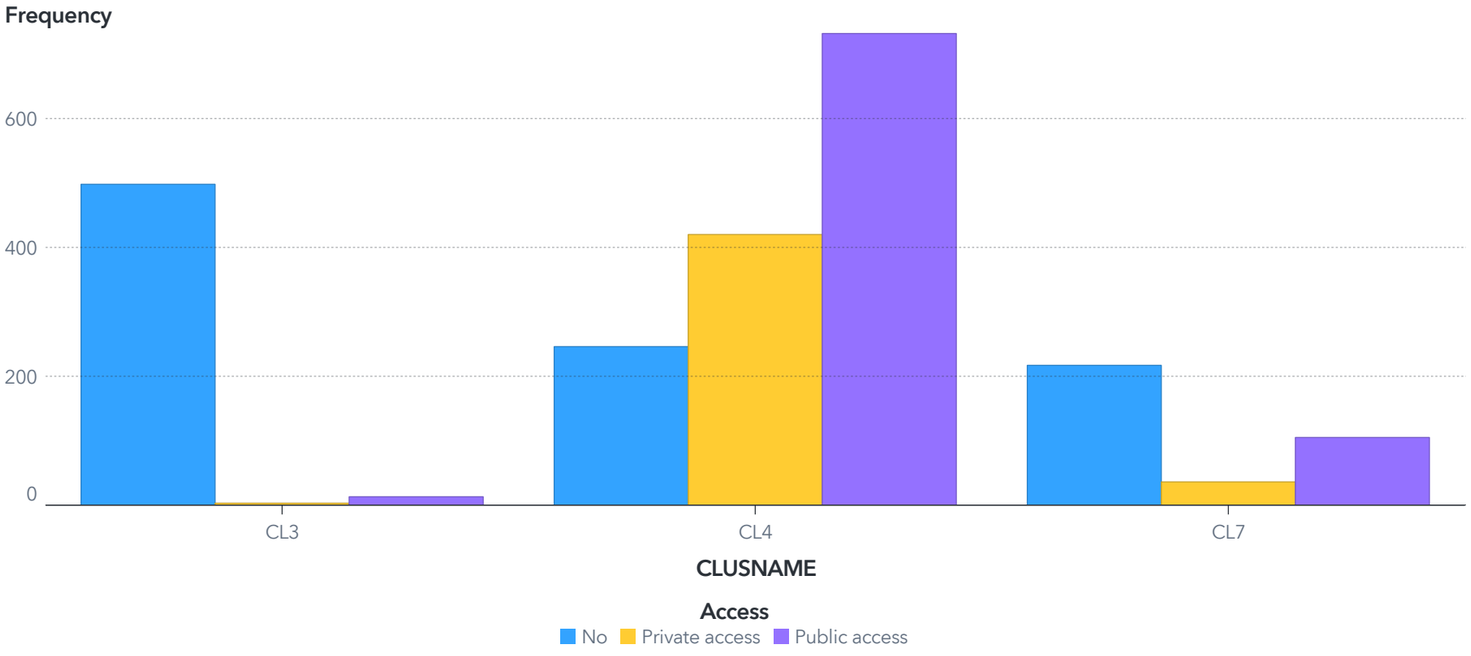
Frequency of CLUSNAME



Frequency of CLUSNAME grouped by Type

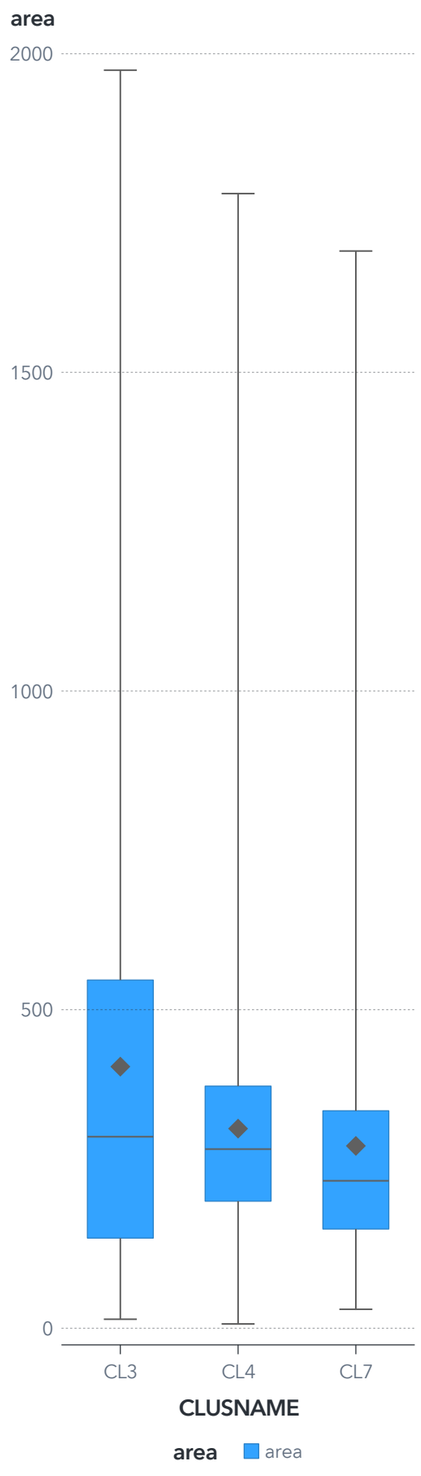


Frequency of CLUSNAME grouped by Access

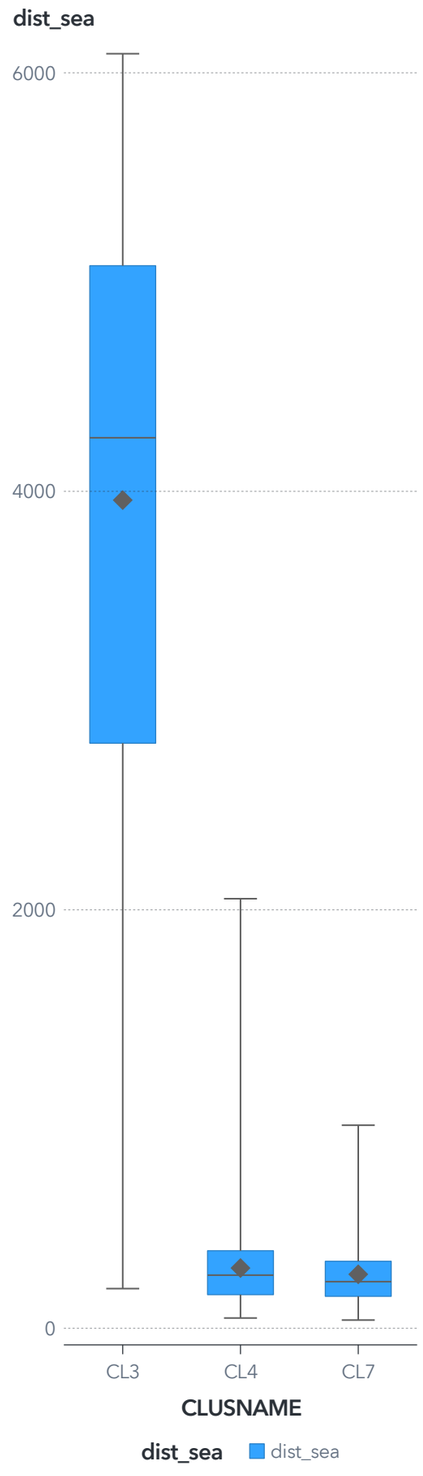


A1.1

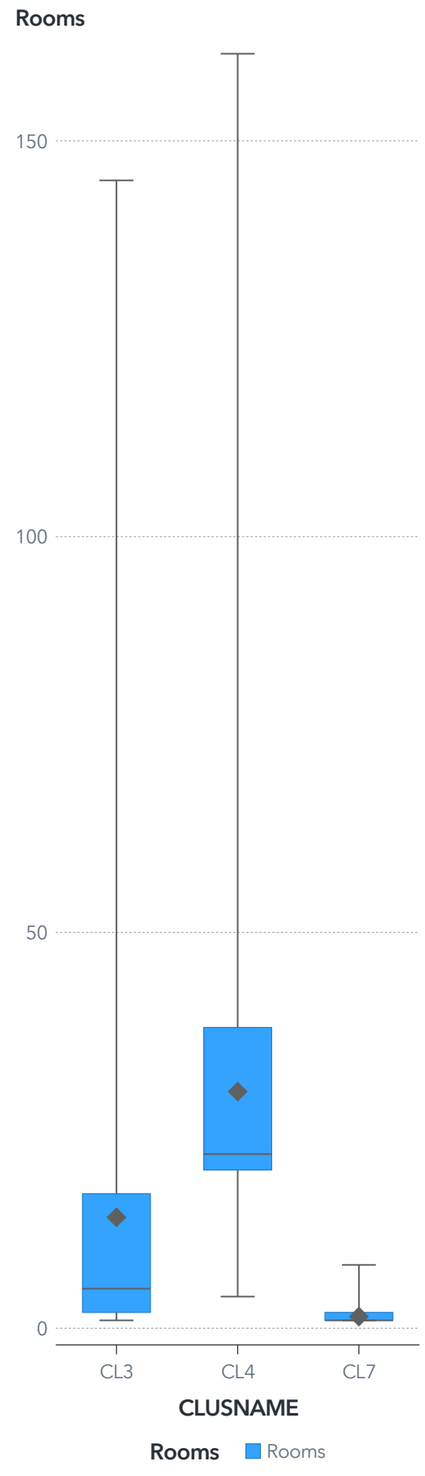
area by CLUSNAME



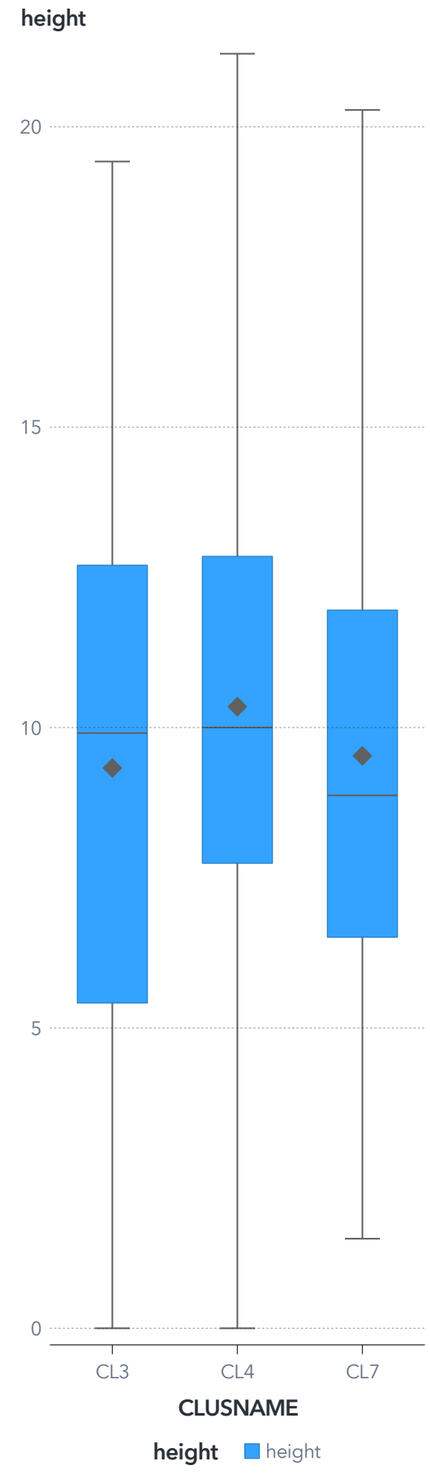
dist_sea by CLUSNAME



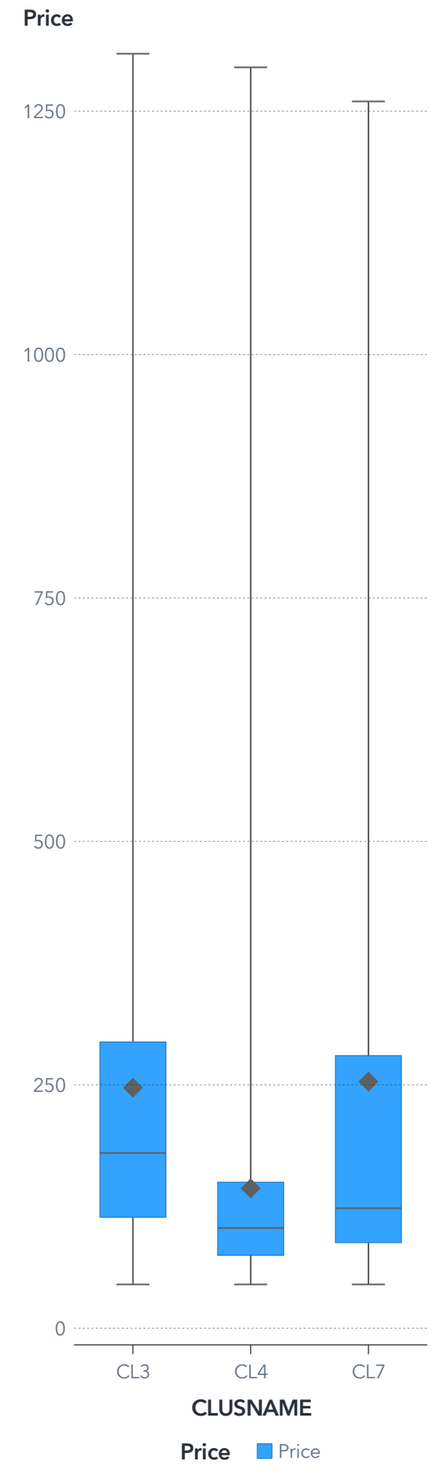
Rooms by CLUSNAME



height by CLUSNAME



Price by CLUSNAME



Cluster 3: Hierarchical Clustering – Complete linkage (Maximum linkage)

Number of Clusters: 5
Variables Used for Clustering: price, height, dist_sea, area, room.
Category variable used: Access, Type

CL10:

-This cluster has the highest median distance from the sea and the highest median accommodation height.
-It mainly consists of hotels, most of which have no beach access.

CL5:

-Cluster 5 consists mainly of **holiday rental homes and hotels** and likely contains luxury accommodations.
-A small number of properties have private beach access.
-It has a **smaller median area** compared to the other clusters, with the smallest accommodation area overall.
-The **median number of rooms is very low**, typically between one and two.
-The median height of the accommodations is also **relatively lower** than the other clusters.
-This cluster has the **highest median price** and a **wider interquartile range** compared to the others.

CL6:

-Cluster 6 consists **mainly of holiday rental homes**.
-Most accommodations **do not have beach access**, likely because the **median distance from the sea is relatively far**.
-These properties have the **largest area but a small number of rooms**.
-Prices are moderate compared to the other clusters.

CL7:

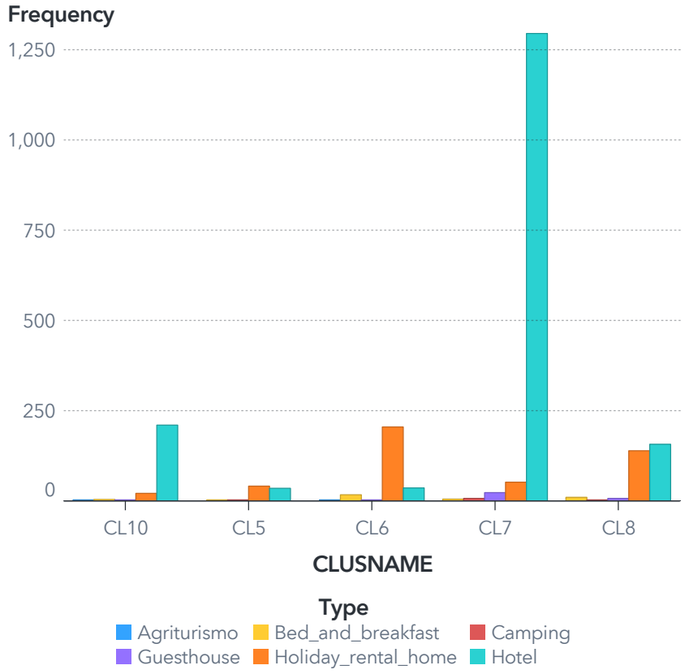
-Cluster 7 **consists mainly of hotels**.
-Being **close to the sea**, most accommodations have **public beach access**, followed by private and no-access options.
-It has the **highest number of rooms**, which is expected given the dominance of hotels.
-In terms of **price**, this cluster ranks among the **lowest**.

CL8:

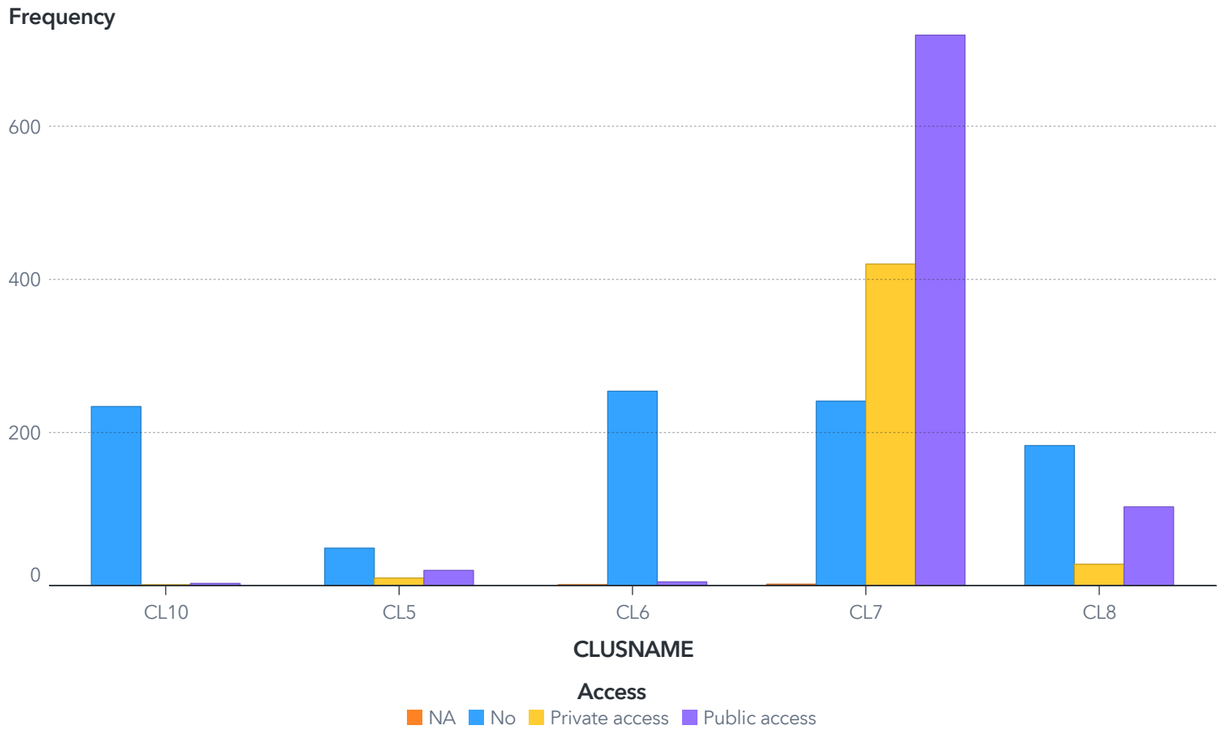
Cluster 8 has a fairly balanced mix of **hotels and holiday rental homes**.
Most accommodations have no beach access, while the remaining are mainly those with public access.
The **number of rooms is relatively low**.
In terms of price, this cluster is among the more **affordable** compared to the others.

Frequency of CLUSNAME

Frequency of CLUSNAME grouped by Type

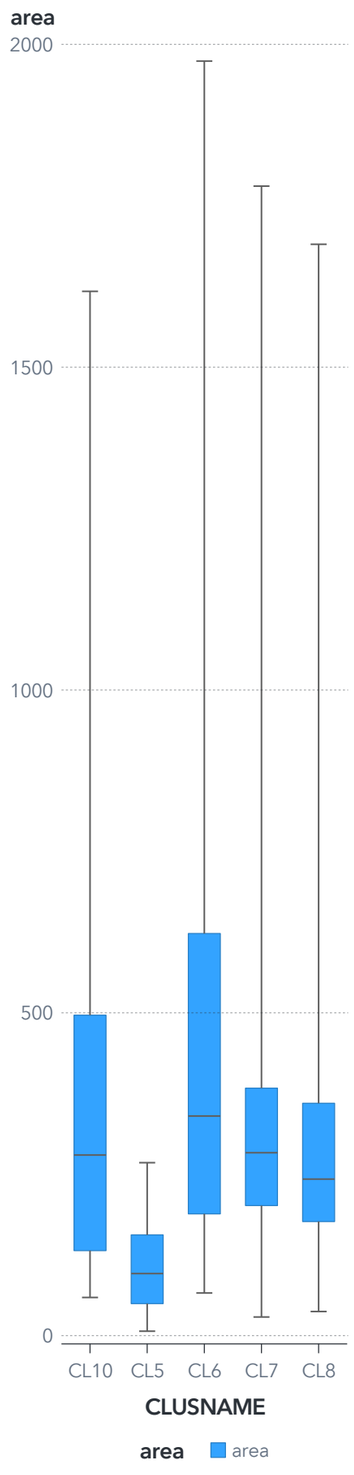


Frequency of CLUSNAME grouped by Access

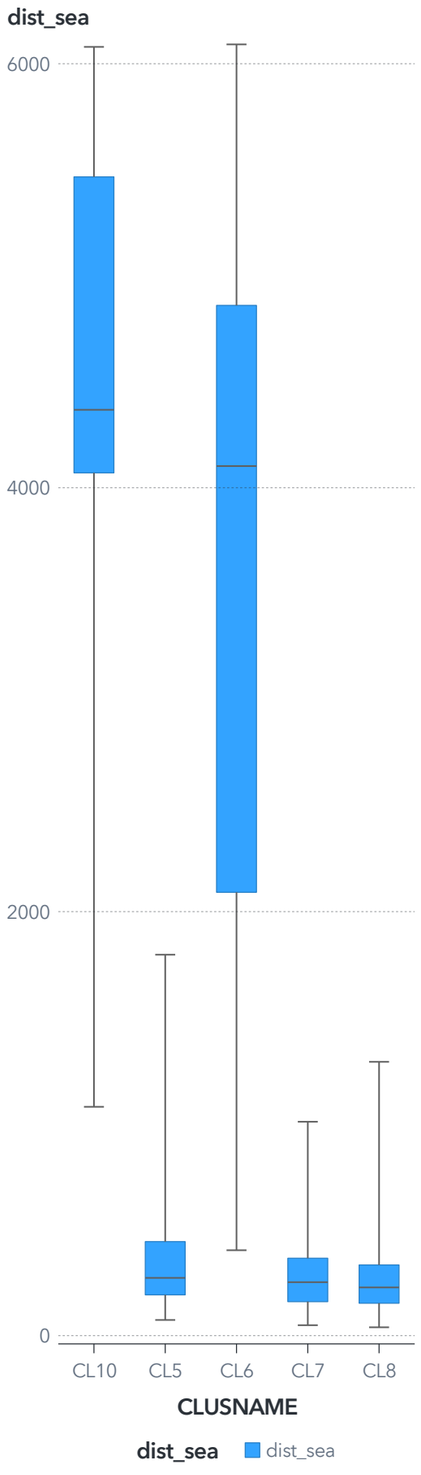


Complete_linkage_max (1)

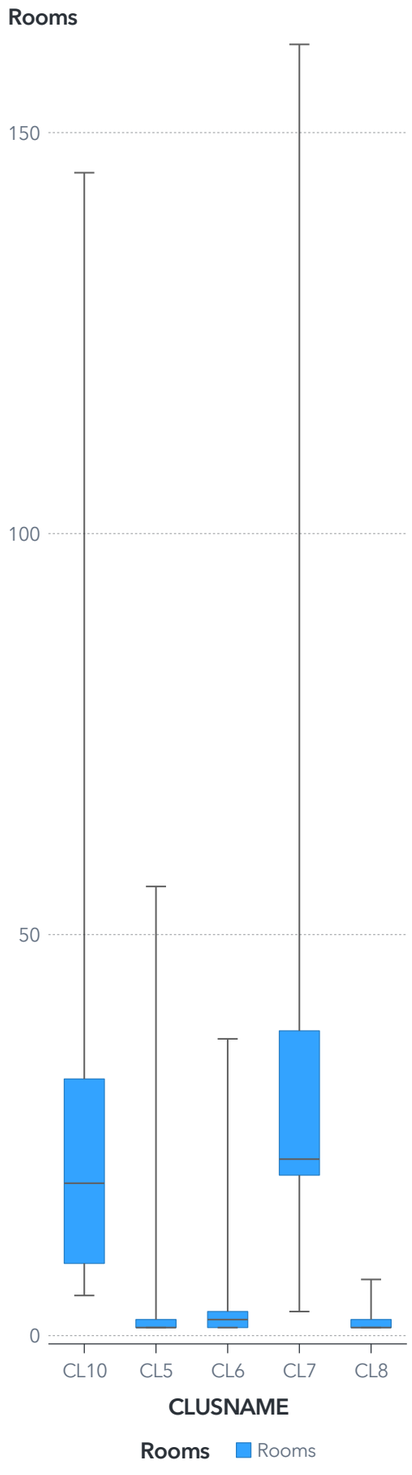
area by CLUSNAME



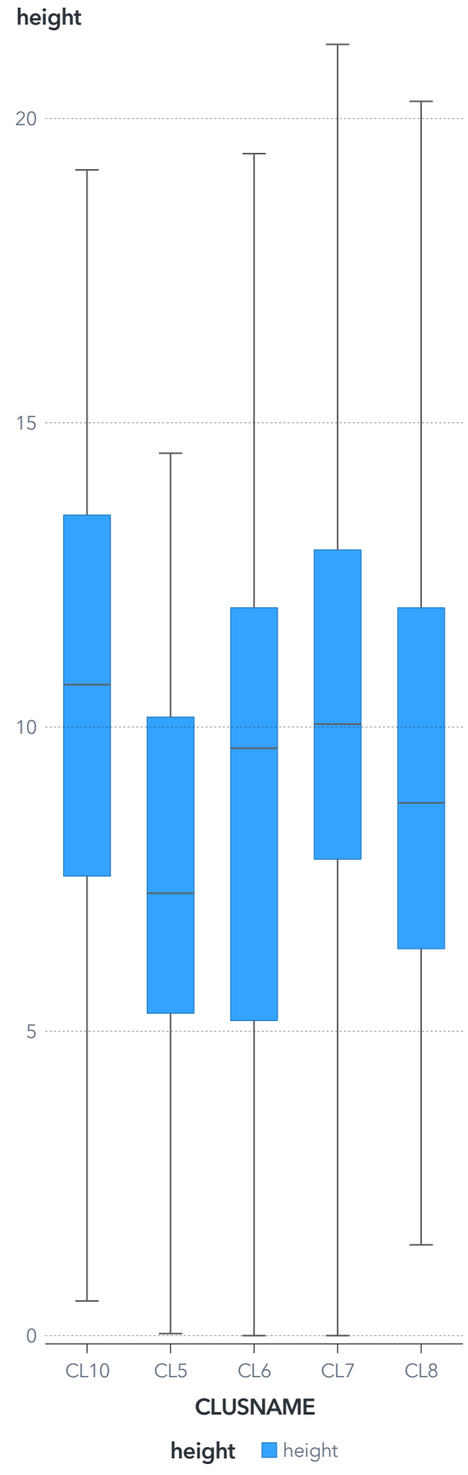
dist_sea by CLUSNAME



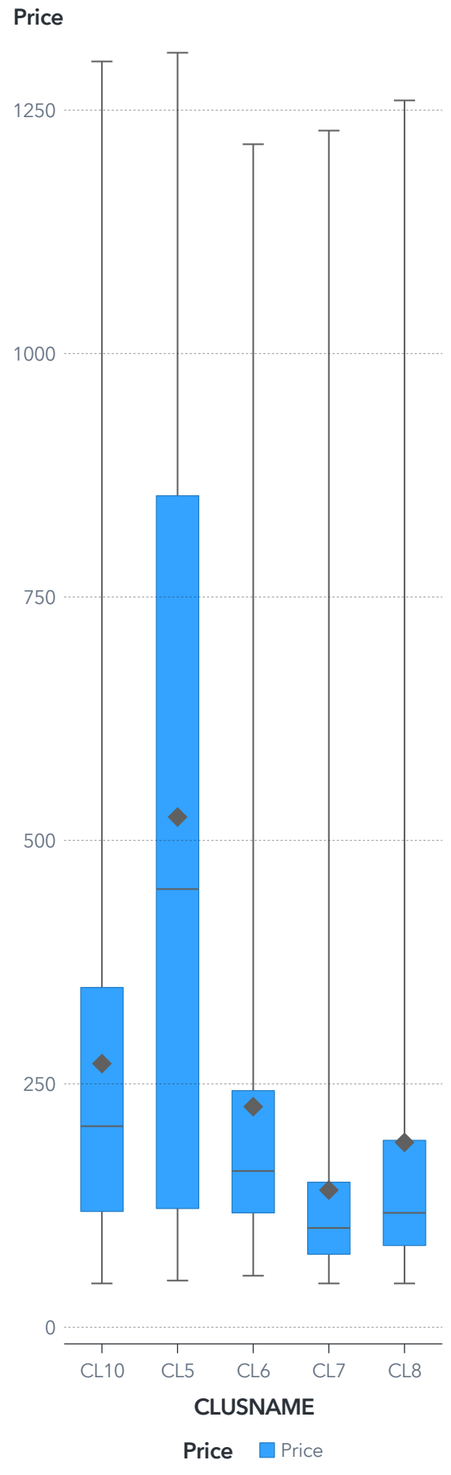
Rooms by CLUSNAME



height by CLUSNAME



Price by CLUSNAME



Managerial Communication

The data preparation and cleaning process helps to mitigate on the risk of erroneous data, but this does not warrant a complete accurate dataset.
Several key actions were implemented to enhance data quality and analytical validity:

Outlier and Duplication Management

-Duplicate records were identified and removed to prevent bias from repeated entries.
-During the review, some records shared the same name but had slight variations in other attributes.
-For this process, it was assumed that these duplicates represented the same entity and that minor differences (for example, small variations in latitude or longitude for "Hotel XYZ") were negligible.

Removal of NA variables and converting to proper data type

-Some of the variables are having "NA" values not captured in the summary statistics and this have been removed accordingly.

Major missing data issue

-Several columns contain missing data, with some variables such as ratings showing a high number of missing values. To retain as much information as possible, only essential rows or columns with minimal missing data were removed.

Removal of 99th and above percentile data

-Observations exceeding the 99th percentile were excluded under a conservative approach, mitigating the impact of extreme outliers on model accuracy and stability.

Model Comparison:

When comparing the different clustering models used in this analysis, the **Complete Linkage Hierarchical** Clustering method produced clusters with the **most distinct and clearly defined traits**.

In contrast, **Ward's Minimum Variance Hierarchical** Clustering resulted in an optimal three-cluster solution; however, the characteristics of these clusters were **less distinct compared to those generated by the complete linkage method**.

The **K-Means Clustering model** appeared **overly complex**, with several overlapping clusters that were difficult to distinguish visually.
Nevertheless, **interpretation can be improved** by referring to the cluster summary information and using boxplot diagrams to illustrate differences in variable distributions across clusters.

Additional Data & Deeper analysis:

Further data segmentation before clustering:
-During the analysis, it was observed that some clusters contained a wide range of values for certain variables, particularly Price.
-This variation is probably due to the different accommodation types with distinct price structures.
Therefore, it is recommended to further segment the data into smaller, more homogeneous subsets before performing clustering. Such segmentation would help reduce the influence of extreme price differences and allow for more meaningful cluster formation.

Inclusion of room occupancy rate:
-Although the dataset provides extensive information on available accommodations, it lacks visibility into their actual rental or occupancy rates.
-Including this variable would enhance the analysis by allowing the removal of inactive accommodations (e.g., those with zero occupancy) and enabling deeper investigation into how occupancy rate varies with other factors such as price, area, or location. Moreover, the occupancy rate could be analysed through a regression model to explore its relationship with other variables, providing further insights into accommodation performance.

Appendix

A1.1 Frequency of CLUSNAME grouped by Access

Filters: (Access IN { 'No'; 'Private access'; 'Public access' }) OR Access MISSING