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

Creation Date: Sunday, 28 Sep, 2025

Lim Boon Yan Edward

Student ID: 01548108

Email: Edward.lim.2025@mitb.smu.edu.sg

Content

Context

The coastal regions of Veneto and Emilia Romagna in Northern Italy are key economic areas driven by the hospitality industry but are increasingly vulnerable to climate change.

Rising sea levels, coastal erosion, storm surges, land subsidence, and urbanization threaten their stability. The sustainability of the hospitality sector depends on maintaining coastal environmental balance, with vulnerability reflected in natural and cultural dependence, physical exposure of infrastructure, and the economic impact on destination appeal.

The Task

Focus on the supply side by segmenting hotels using selected clustering variables that reflect the multidimensional aspects of coastal hospitality vulnerability with the key outcomes:

• Derive a set of clustering variables (at least five and not more than eight) by using appropriate data extracting and cleaning methods.

• Perform cluster analysis by using at least three different clustering methods.

• Interpret the analysis results and describe the characteristics of the clusters formed

Data Preparation

In this section, the following tasks were performed:

Step 1: Data Sanity Checks

A Data check have been conducted using summary statistics.

Variable Summary

Obs	Variable name	Width of the variable formatted value	Type of the raw values	Recommended level for analytics	Have more unreported levels	Number of levels	Number of missing values	Minimum numeric value	Maximum numeric value	Mean	Standard deviation
1	Name	76	C	ID	Y	20	0	.	.	.	.
2	latitude	12	N	INTERVAL	Y	20	0	43.95985153	45.66508758	44.733210275	0.6656948632
3	longitude	12	N	INTERVAL	Y	20	0	12.13347259	13.08028244	12.504214704	0.175502167
4	area	12	N	INTERVAL	Y	20	0	6.817076923	9191.599584	368.75316126	441.50308277
5	dist_waterway	12	N	INTERVAL	Y	20	0	0.019760544	15047.54328	1030.6332938	1294.4139382
6	dist_estuary	12	N	INTERVAL	Y	20	1	5.970919605	78435.4235	11126.725038	21608.902724
7	dist_sea	12	N	INTERVAL	Y	20	0	5.970919605	16041.2176	1109.1655154	1782.3882386
8	grond_z	12	C	ID	Y	20	0	.	.	.	.
9	slr_median_rcp45_2030	12	C	ID	Y	20	0	.	.	.	.
10	height	8	C	ID	Y	20	0	.	.	.	.
11	Rooms	4	C	ID	Y	20	0	.	.	.	.
12	Price	7	C	ID	Y	20	0	.	.	.	.
13	Ratings	3	C	ID	Y	20	0	.	.	.	.
14	encoded_Type	12	N	CLASS	N	8	0	1	8	6.2741550696	1.196945486
15	encoded_Access	12	N	CLASS	N	4	0	-1	3	1.5497017893	1.0397808778
16	encoded_EcoCertifd	12	N	CLASS	N	3	0	-1	2	0.8968190855	0.5153150815
17	encoded_Green	12	N	CLASS	N	3	0	-1	2	0.9980119284	0.6547591227
18	encoded_Pool	12	N	CLASS	N	4	0	-1	3	1.1149105368	0.7332109328
19	encoded_Sustainabl	12	N	CLASS	N	3	0	-1	2	0.5284294235	0.868511318

Fig 1.1 Initial summary statistic.

Changes done:

- **Data types were reclassified** from VARCHAR to Double using SAS Proc (Fig 1.2) for numerical variables, with **missing values (NA) converted to "."** (Fig 1.3) prior to the type conversion.
- **New columns** were created to map the encoded data descriptions (Fig 1.4).

Name	Type (Before)	Type (After)
grond_z	VARCHAR	Double
slr_median_rcp45_2030	VARCHAR	Double
height	VARCHAR	Double
Rooms	VARCHAR	Double
Price	VARCHAR	Double
Ratings	VARCHAR	Double

Fig 1.2 table of change

```
    /* Replace "NA" with .*/
    if strip(upcase(grond_z)) = 'NA' then grond_z_n = .;
    else grond_z_n = input(grond_z, best32.);

    if strip(upcase(slr_median_rcp45_2030)) = 'NA' then slr_n = .;
    else slr_n = input(slr_median_rcp45_2030, best32.);

    if strip(upcase(height)) = 'NA' then height_n = .;
    else height_n = input(height, best32.);

    if strip(upcase(Rooms)) = 'NA' then rooms_n = .;
    else rooms_n = input(Rooms, best32.);

    if strip(upcase(Price)) = 'NA' then price_n = .;
    else price_n = input(Price, best32.);

    if strip(upcase(Ratings)) = 'NA' then ratings_n = .;
    else ratings_n = input(Ratings, best32.);

    /* Keep only cleaned version and drop original*/
    drop grond_z slr_median_rcp45_2030 height Rooms Price Ratings;
    rename grond_z_n = grond_z
           slr_n     = slr_median_rcp45_2030
           height_n  = height
           rooms_n   = Rooms
           price_n   = Price
           ratings_n = Ratings;
```

Fig 1.3 SAS proc to replace NA to "."

```
/* Create norminal Category*/

length Type $20 Access $20 EcoCertifd $20 Green $20 Pool $20 Sustainabl $20;

    /* Map encoded values */
    if encoded_Type = 1 then Type = "Agriturismo";
    else if encoded_Type = 2 then Type = "Bed_and_breakfast";
    else if encoded_Type = 3 then Type = "Camping";
    else if encoded_Type = 4 then Type = "Guesthouse";
    else if encoded_Type = 5 then Type = "Holiday_rental_home";
    else if encoded_Type = 6 then Type = "Hostel";
    else if encoded_Type = 7 then Type = "Hotel";
    else Type = "NA";

    if encoded_Access = 1 then Access = "No";
    else if encoded_Access = 2 then Access = "Public access";
    else if encoded_Access = 3 then Access = "Private access";
    else Access = "NA";

    if encoded_EcoCertifd = 1 then EcoCertifd = "No";
    else if encoded_EcoCertifd = 2 then EcoCertifd = "Yes";
    else EcoCertifd = "NA";

    if encoded_Green = 1 then Green = "No";
    else if encoded_Green = 2 then Green = "Yes";
    else Green = "NA";

    if encoded_Pool = 1 then Pool = "No";
    else if encoded_Pool = 2 then Pool = "Shared pool";
    else if encoded_Pool = 2 then Pool = "Private pool";
    else Pool = "NA";

    if encoded_Sustainabl = 1 then Sustainabl = "No";
    else if encoded_Sustainabl = 2 then Sustainabl = "Yes";
    else Sustainabl = "NA";

run;
```

Fig 1.4 SAS proc to create named column

Step 2: Log10 Transformation for certain variables

A Log 10 transformation have been performed on the variables below and this was done using SAS procedure:

```
68      /* Log10 Transformation*/
69      log_dist_sea = log10(dist_sea + 1);
70      log_area = log10(area + 1);
71      log_price = log10(price + 1);
72      log_height = log10(height + 1);
73      log_rooms = log10(rooms + 1);
```

Fig 1.5 SAS procedure to create Log transformed columns

Variable Summary

Obs	Variable name	Width of the variable formatted value	Type of the raw values	Recommended level for analytics	Have more unreported levels	Number of levels	Number of missing values	Minimum numeric value	Maximum numeric value	Mean	Standard deviation
1	Name	76	C	ID	Y	20	0	.	.	.	.
2	latitude	12	N	INTERVAL	Y	20	0	43.95985153	45.66508758	44.733210275	0.6656948632
3	longitude	12	N	INTERVAL	Y	20	0	12.13347259	13.08028244	12.504214704	0.175502167
4	area	12	N	INTERVAL	Y	20	0	6.817076923	9191.599584	368.75316126	441.50308277
5	dist_waterway	12	N	INTERVAL	Y	20	0	0.019760544	15047.54328	1030.6332938	1294.4139382
6	dist_estuary	12	N	INTERVAL	Y	20	1	5.970919605	78435.4235	11126.725038	21608.902724
7	dist_sea	12	N	INTERVAL	Y	20	0	5.970919605	16041.2176	1109.1655154	1782.3882386
8	encoded_Type	12	N	CLASS	N	8	0	1	8	6.2741550696	1.196945486
9	encoded_Access	12	N	CLASS	N	4	0	-1	3	1.5497017893	1.0397808778
10	encoded_EcoCertifd	12	N	CLASS	N	3	0	-1	2	0.8968190855	0.5153150815
11	encoded_Green	12	N	CLASS	N	3	0	-1	2	0.9980119284	0.6547591227
12	encoded_Pool	12	N	CLASS	N	4	0	-1	3	1.1149105368	0.7332109328
13	encoded_Sustainabl	12	N	CLASS	N	3	0	-1	2	0.5284294235	0.868511318
14	grond_z	12	N	INTERVAL	Y	20	470	-2.416205205	10.99923992	1.4706921362	1.421140906
15	slr_median_rcp45_2030	12	N	INTERVAL	Y	20	834	-2.645362433	10.901739	1.3631330052	1.4047721491
16	height	12	N	INTERVAL	Y	20	742	-1	30.5136	9.8597264459	4.3483099379
17	Rooms	12	N	INTERVAL	Y	20	1372	1	3000	24.775560416	67.506458102
18	Price	12	N	INTERVAL	Y	20	875	1	4536	204.21730999	275.62458132
19	Ratings	12	N	INTERVAL	Y	20	1395	1	5	4.3137276479	0.5270464192
20	Type	20	C	CLASS	N	8	0	.	.	.	.
21	Access	20	C	CLASS	N	4	0	.	.	.	.
22	EcoCertifd	20	C	CLASS	N	3	0	.	.	.	.
23	Green	20	C	CLASS	N	3	0	.	.	.	.
24	Pool	20	C	CLASS	N	3	0	.	.	.	.
25	Sustainabl	20	C	CLASS	N	3	0	.	.	.	.
26	log_dist_sea	12	N	INTERVAL	Y	20	0	0.8432900741	4.2052644029	2.6387119373	0.5468910627
27	log_area	12	N	INTERVAL	Y	20	0	0.8930443856	3.9634383433	2.4273701934	0.3401923533
28	log_price	12	N	INTERVAL	Y	20	875	0.3010299957	3.6567687793	2.1467065256	0.3386619513
29	log_height	12	N	INTERVAL	Y	20	747	0	1.4984980182	0.9914678381	0.2221648528
30	log_rooms	12	N	INTERVAL	Y	20	1372	0.3010299957	3.4772659954	1.0832608394	0.5499159693

Fig 1.6 post log transformation

After replacing the NA values with "." in Step 1, the summary statistics now display missing values.

The base 10 log transformation helped reduce skewness, as shown in the bottom chart of Figure 1.7, which presents the distribution after transformation.

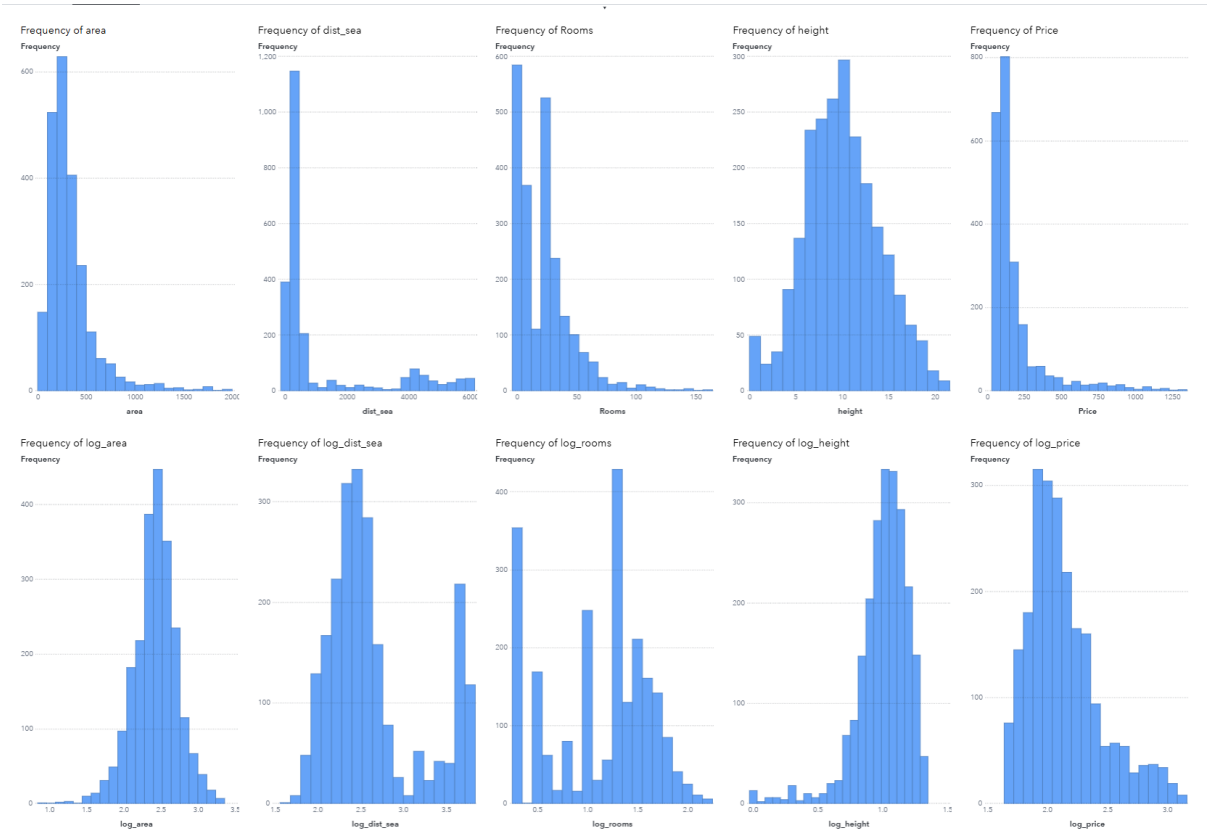

Fig 1.7 variables distribution

Step 3: Count for duplicates

In this step, PROC SQL (Fig 1.9) was used to check for duplicate entries in the Name column (Fig 1.8).

The results indicate a total of 514 duplicate rows.

duplicate_groups	duplicate_rows
503	514

Name	dup_count
a tribute to music	2
affittacamere giabetta	2
agora? park hotel	2
ai mori d`oriente	2
ai patrizi venezia	2
al redentore di venezia	2

Fig 1.8 duplicated data

```
proc sql;
select count(*) as duplicate_groups,
         sum(dup_count - 1) as duplicate_rows
   from (
      select Name, count(*) as dup_count
      from casuser.COASTAL_DATASET_1
      group by Name
      having count(*) > 1
   ) as dup_summary;

   select Name, count(*) as dup_count
   from casuser.COASTAL_DATASET_1
   group by Name
   having count(*) > 1;
quit;
```

Fig 1.9 SQL procedure to identify duplicated record

Step 4: Removal of duplicated record

Initially, the dataset contained 5030 rows, and after removing the duplicates, 4516 rows remained.

This process was carried out using SAS Proc (Fig 1.10).

```
proc sort data=casuser.COASTAL_DATASET_1
          out=casuser.COASTAL_DATASET_2
          nodupkey;
   by Name;
run;
```

Fig 1.10 SAS proc to remove identified duplicates

Step 5: Removal of 99th and above percentile numerical data

Data exceeding the 99th percentile were removed to eliminate extreme values, following a pessimistic approach (Fig 1.11).

Column	99^{tt} percentile (Cutoff value)
Price	1,309
Rooms	161
area	2,128
height	21
dist_sea	6,091
dist_estuary	75,030
dist_waterway	5,167

Fig 1.11 99th percentile cut off for data column

```
proc univariate data=casuser.COASTAL_DATASET_3 noprint;
   var Price Rooms area height dist_sea dist_estuary dist_waterway;
   output out=casuser.percentiles
          p99 = P99_Price P99_Rooms P99_Area P99_Height
                P99_Dist_Sea P99_Dist_Estuary P99_Dist_Waterway;
run;

data _null_;
   set casuser.percentiles;
   call symputx('p99_price', P99_Price);
   call symputx('p99_rooms', P99_Rooms);
   call symputx('p99_area', P99_Area);
   call symputx('p99_height', P99_Height);
   call symputx('p99_dist_sea', P99_Dist_Sea);
   call symputx('p99_dist_estuary', P99_Dist_Estuary);
   call symputx('p99_dist_waterway', P99_Dist_Waterway);
run;

data casuser.COASTAL_DATASET_4;
   set casuser.COASTAL_DATASET_3;
   if Price <= &p99_price and
      Rooms <= &p99_rooms and
      Area <= &p99_area and
      Height <= &p99_height and
      dist_sea <= &p99_dist_sea and
      dist_estuary <= &p99_dist_estuary and
      dist_waterway <= &p99_dist_waterway;
run;

proc casutil;
    save casdata= 'COASTAL_DATASET_4'
    casout='COASTAL_DATASET_4.sashdat'
    outcaslib='CASUSER'
    replace;

quit;
```

Fig 1.12 SAS Proc to remove 99th percentile cut data

STEP 6: Dropping missing value and set range limit

Rows with missing values in selected columns were removed, while others were replaced as summarised.

Certain columns were retained to preserve as much data as possible.

Records falling outside the defined range were also excluded.

After the cleaning process, the dataset was reduced to **2273** rows.

Column Name	Action done
Price	Drop missing data, data lesser than 45 were removed
Rooms	Drop missing data
height	Drop missing data
Ratings	Missing value replaced with 0

Fig 1.13 Summary of data treatment

```
data casuser.COASTAL_DATASET_Final;
   set casuser.COASTAL_DATASET_4;
    if missing(Price) then delete;
    if missing(Rooms) then delete;
    if missing(height) then delete;
    if missing(Ratings) then Ratings = 0;
    if Price < 45 then delete;

run;
proc casutil;
    save casdata= 'COASTAL_DATASET_Final'
    casout='COASTAL_DATASET_Final.sashdat'
    outcaslib='CASUSER'
    replace;

quit;
```

Fig 1.14 SAS proc to further cleanup data

Variable Summary

Obs	Variable name	Width of the variable formatted value	Type of the raw values	Recommended level for analytics	Have more unreported levels	Number of levels	Number of missing values	Minimum numeric value	Maximum numeric value	Mean	Standard deviation
1	Name	76	C	ID	Y	20	0	.	.	.	.
2	latitude	12	N	INTERVAL	Y	20	0	43.96044014	45.54237758	44.579646705	0.6377327266
3	longitude	12	N	INTERVAL	Y	20	0	12.23328511	12.751232	12.489144246	0.1435735202
4	area	12	N	INTERVAL	Y	20	0	6.817076923	1974.199678	331.05688179	248.52116375
5	dist_waterway	12	N	INTERVAL	Y	20	0	0.019760544	5151.582741	958.59865032	1109.403289
6	dist_estuary	12	N	INTERVAL	Y	20	0	25.25193926	74949.50952	13046.120157	23118.111195
7	dist_sea	12	N	INTERVAL	Y	20	0	39.15005461	6091.313856	1114.7073876	1710.1583694
8	encoded_Type	12	N	CLASS	N	6	0	1	7	6.447866256	1.0738718596
9	encoded_Access	12	N	CLASS	N	4	0	-1	3	1.7751869776	0.7657157845
10	encoded_EcoCertifd	12	N	CLASS	N	3	0	-1	2	1.0193576771	0.1719147838
11	encoded_Green	12	N	CLASS	N	3	0	-1	2	1.1526616806	0.3776469941
12	encoded_Pool	12	N	CLASS	N	4	0	-1	3	1.262208535	0.468987012
13	encoded_Sustainabl	12	N	CLASS	N	3	0	-1	2	0.4302683678	0.9181267135
14	grond_z	12	N	INTERVAL	Y	20	173	-1.972686046	10.99862957	1.6183439076	1.4965888413
15	slr_median_rcp45_2030	12	N	INTERVAL	Y	20	353	-2.529299092	10.901136	1.511355413	1.4749804631
16	height	12	N	INTERVAL	Y	20	0	0	21.2174	9.9875660801	4.0384059366
17	Rooms	12	N	INTERVAL	Y	20	0	1	161	21.82358117	22.062653423
18	Price	12	N	INTERVAL	Y	20	0	45	1309	184.38531896	193.58788458
19	Ratings	12	N	INTERVAL	Y	20	0	0	5	2.8558732952	2.0892649926
20	Type	20	C	CLASS	N	6	0	.	.	.	.
21	Access	20	C	CLASS	N	4	0	.	.	.	.
22	EcoCertifd	20	C	CLASS	N	3	0	.	.	.	.
23	Green	20	C	CLASS	N	3	0	.	.	.	.
24	Pool	20	C	CLASS	N	3	0	.	.	.	.
25	Sustainabl	20	C	CLASS	N	3	0	.	.	.	.
26	log_dist_sea	12	N	INTERVAL	Y	20	0	1.6036861403	3.7847822687	2.644995686	0.5468430545
27	log_area	12	N	INTERVAL	Y	20	0	0.8930443856	3.2956110061	2.4248498428	0.2948083217
28	log_price	12	N	INTERVAL	Y	20	0	1.6627578317	3.1172712957	2.1369456339	0.3026530524
29	log_height	12	N	INTERVAL	Y	20	0	0	1.3466932341	1.00263516	0.2038459326
30	log_rooms	12	N	INTERVAL	Y	20	0	0.3010299957	2.2095150145	1.1238836583	0.5046761109

Fig 1.15 Summary Statistic after action

Variables Selection

The selected variables are as below:

Variables
dist_sea
area
height
Price
encoded_Type
encoded_access
Rooms

Insights and Findings

All data analysis was performed in **SAS Viya**.

Correlation Matrix

Based on the selected variables, a correlation matrix has been used to check for any strong correlation **> 0.8** to minimise of multi collinearity issue. None of the variables fall under the multicollinearity.

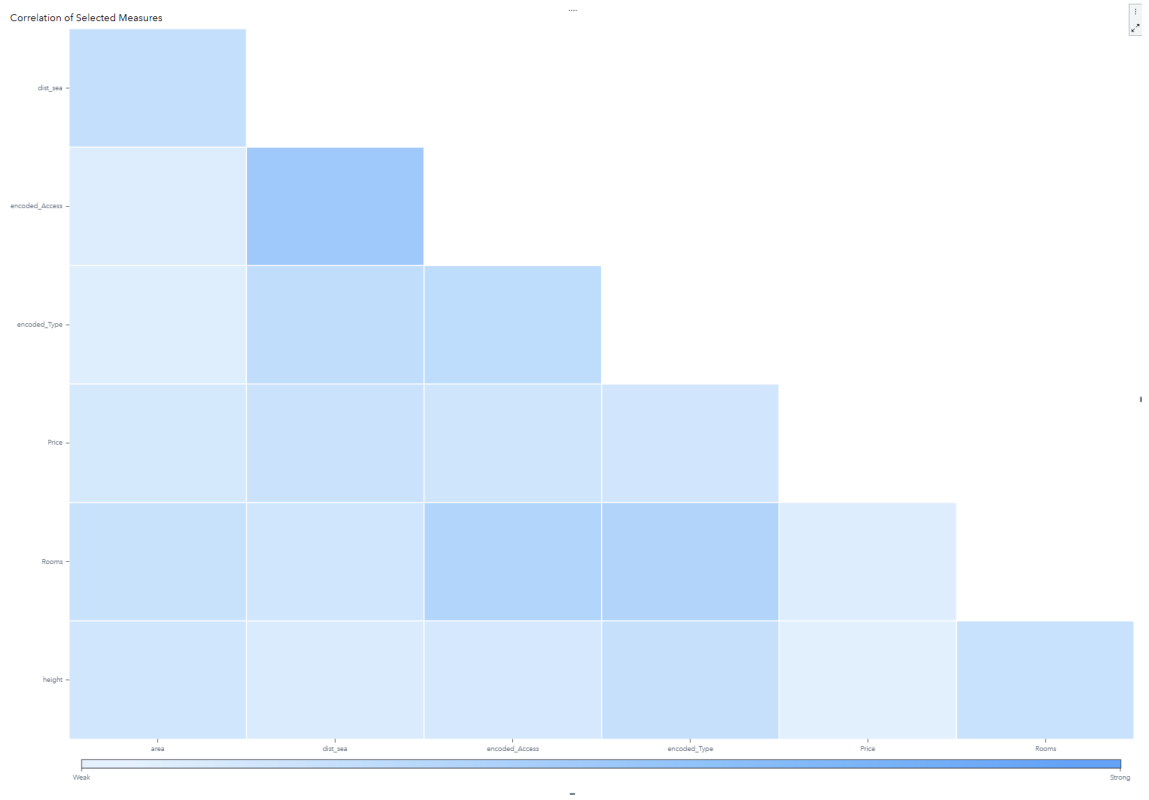

Fig 2.1 Correlation Matrix

Cluster Model 1: Kmeans clustering

A K-means clustering analysis was conducted using the selected variables (Fig 2.2).

Cluster observation was also carried out to determine the optimal number of clusters (Fig 2.4).

After identifying the ideal cluster range (6), the final clustering was executed in SAS Visual Explorer.

Variables
Price
Room
Area
height
dist_sea

Fig 2.2 variable used for Kmeans clustering

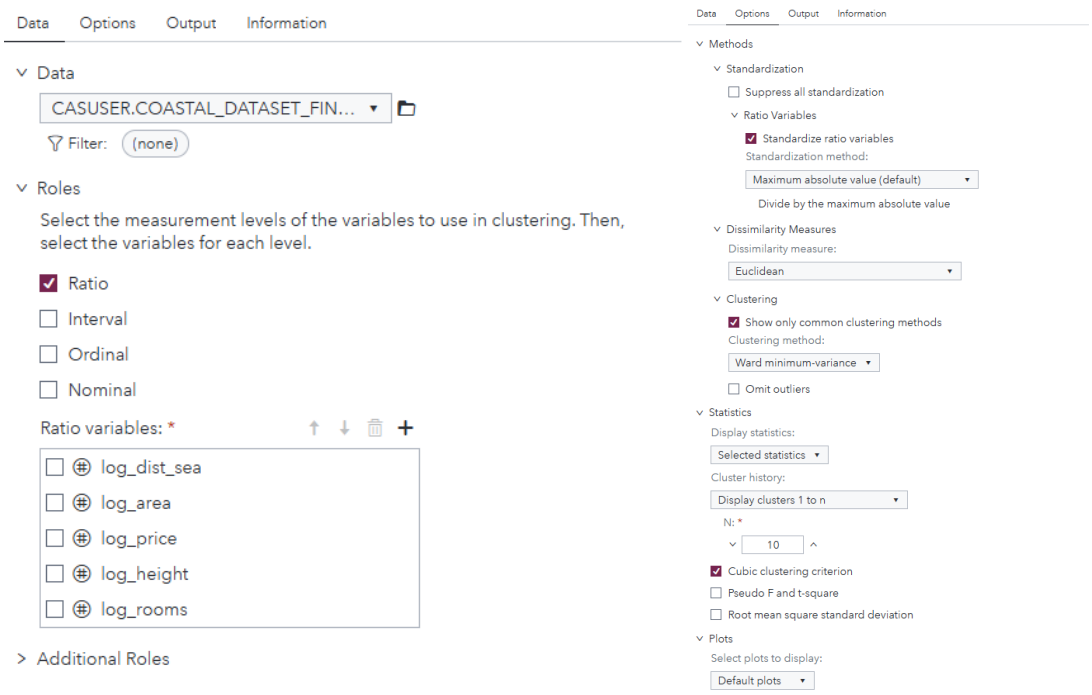

Fig 2.3 set up for CCC to identify ideal clustering number

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	0.05607837	0.03408959	0.4946	0.4946
2	0.02198877	0.00245013	0.1939	0.6885
3	0.01953864	0.01065544	0.1723	0.8608
4	0.00888320	0.00198442	0.0783	0.9392
5	0.00689879		0.0608	1.0000

Root-Mean-Square Total-Sample Standard Deviation	0.150591

Root-Mean-Square Distance Between Observations	0.47621

Cluster History

Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Tie
10	CL28	CL21	74	0.0134	.690	.702	-4.3	
9	CL27	CL11	383	0.0143	.676	.688	-4.1	
8	CL14	CL18	573	0.0148	.661	.673	-3.8	
7	CL12	CL22	789	0.0150	.646	.654	-2.5	
6	CL40	CL10	119	0.0204	.626	.631	-1.6	
5	CL7	CL8	1362	0.0413	.585	.601	-4.3	
4	CL9	CL13	617	0.0573	.527	.559	-8.0	
3	CL5	CL15	1537	0.0748	.452	.483	-6.5	
2	CL4	CL6	736	0.0948	.358	.371	-2.4	
1	CL3	CL2	2273	0.3576	.000	.000	0.00	

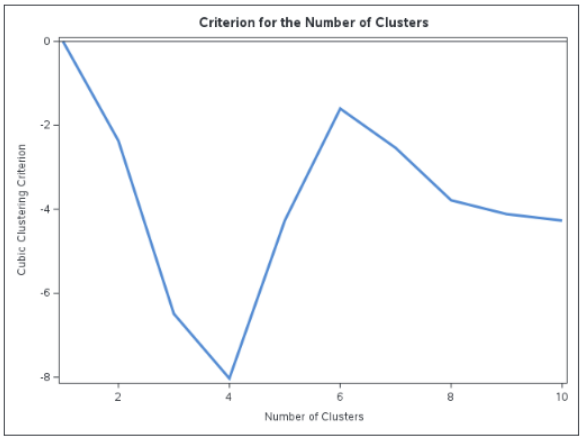

Fig 2.4 output of the analysis

Cluster Observation

Based on the ccc criteria, a cluster of 6 have been identified as the optimal number.

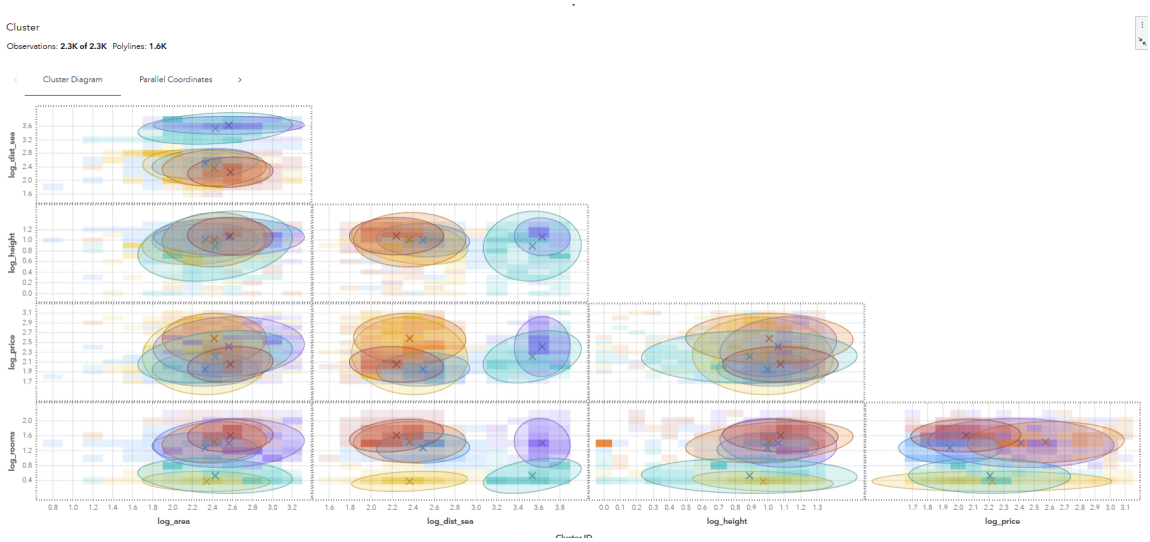

Fig 2.5 Cluster Diagram

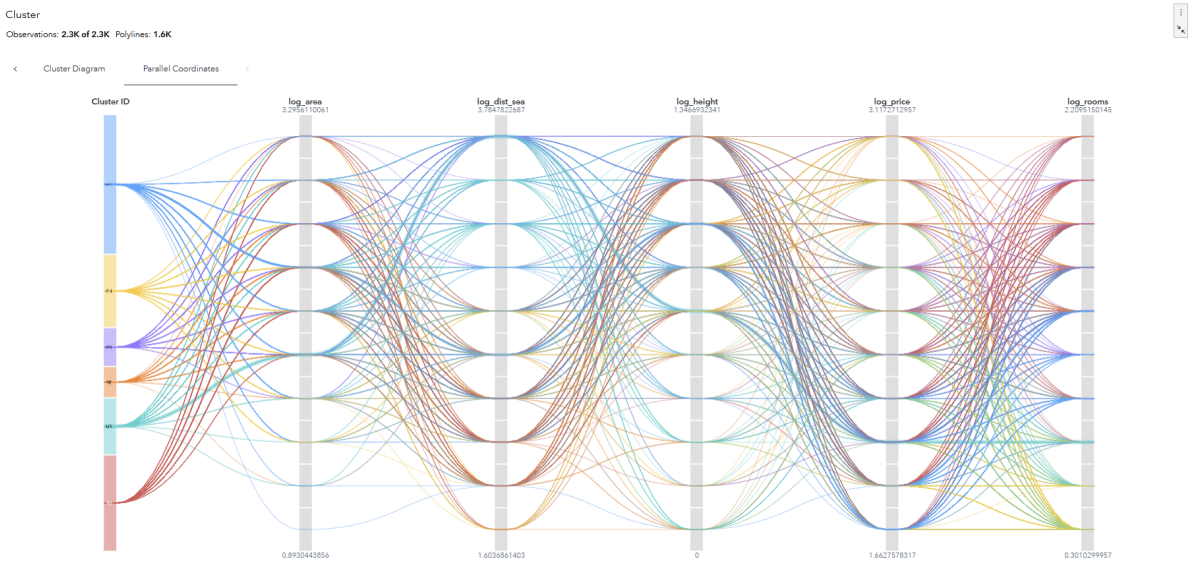

Fig 2.6 Parallel Coordinates

Centroids | Cluster Summary | Model Information | Within Cluster Statistics | Iteration History | Interval information | Parallel Coordinates Plot

Cluster ID	Observations	RMS of STD	Within cluster SS	Min centroid-to-observation	Max centroid-to-observation	Nearest Cluster	Centroid Distance	Average Distance
1	727	0.4325126594	135.99785482	0.0885500817	1.5707816889	6	0.5031490407	0.3992262883
2	382	0.5875472877	131.87091343	0.0950222095	1.4071152407	1	0.9445493573	0.5469061998
3	202	0.5833067699	68.729851146	0.1335499574	1.1298649888	5	0.9362259561	0.5478307137
4	163	0.5469524235	48.762583441	0.1874453534	1.2861678452	6	0.5930956172	0.5054810841
5	297	0.6163216368	112.81615091	0.2227654816	1.3381838887	3	0.9362259561	0.5834465885
6	502	0.4254278775	90.856417235	0.1027756457	1.225072505	1	0.5031490407	0.3980189532

Fig 2.7 Cluster Summary

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 (Fig 2.8).

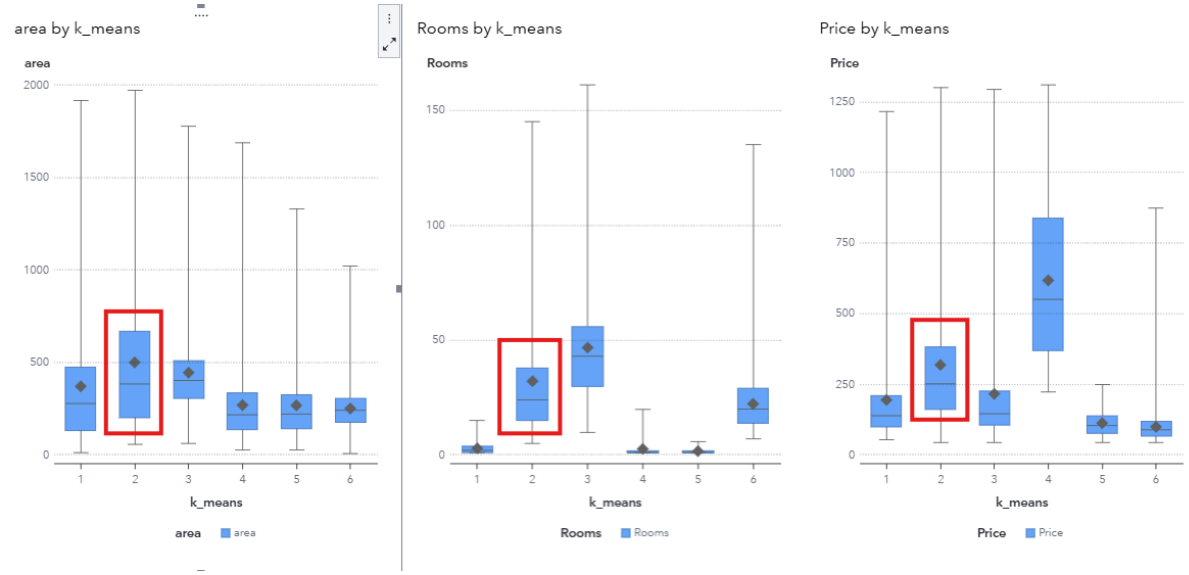

Fig 2.8 Boxplot between area, Rooms and price

- 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 (Fig 2.9).

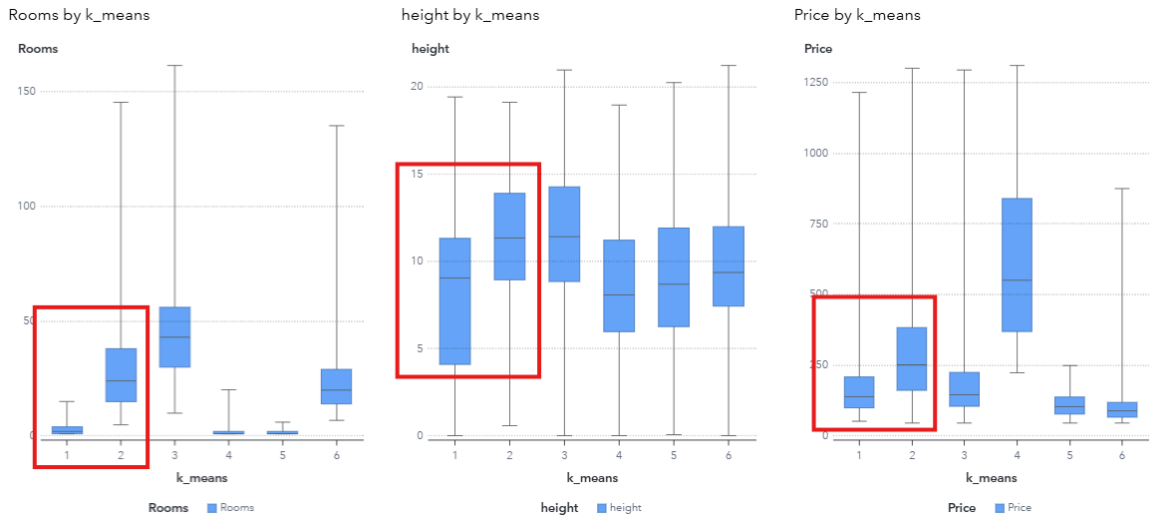

Fig 2.9 Boxplot between area, height and price

<u>Cluster 3:</u>

- **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 (Fig 2.10).
- The main differences between Cluster 3 and Cluster 5 are primarily attributed to variations in **height and number of rooms** (Fig 2.10).

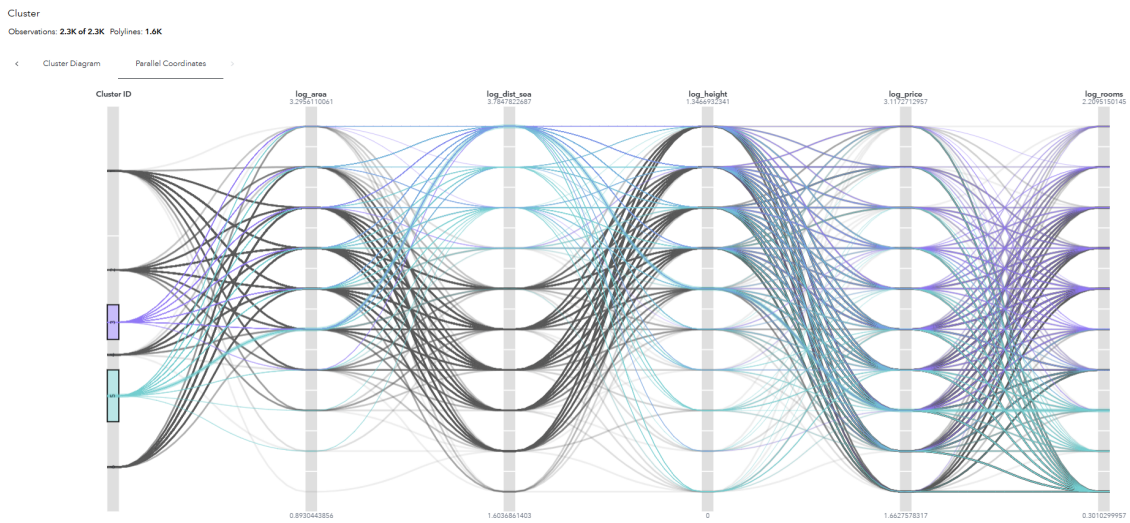

Fig 2.10 Parallel Coordinates Cluster 3 v 5

<u>Cluster 4:</u>

- 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 (Fig 2.11).

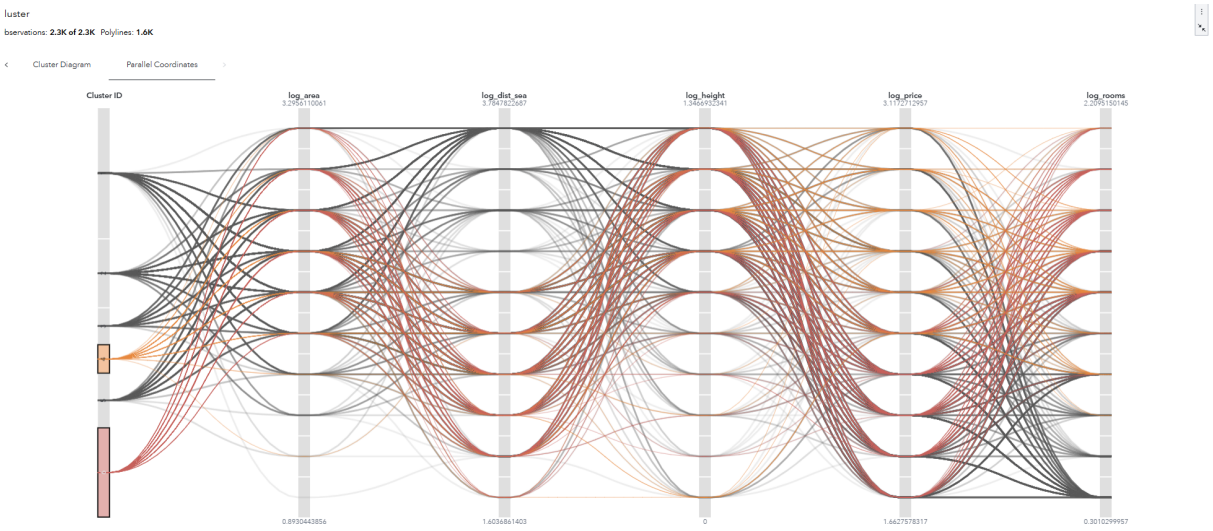

Fig 2.11 Parallel Coordinates Cluster 4 v 6

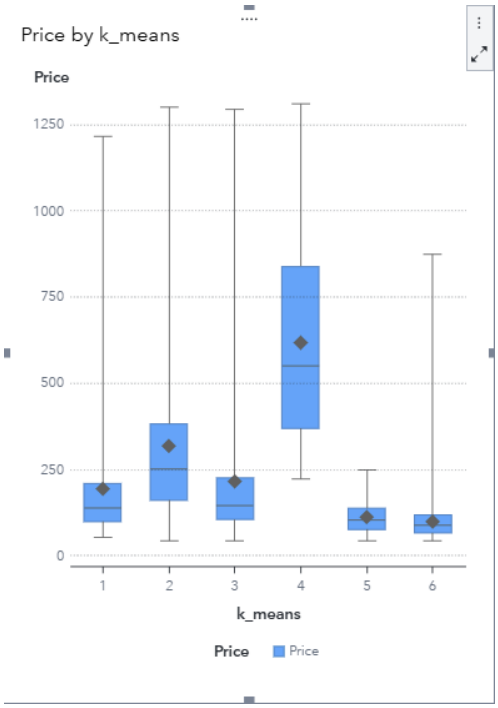

Fig 2.12 Boxplot of Price

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** (Fig 2.13).

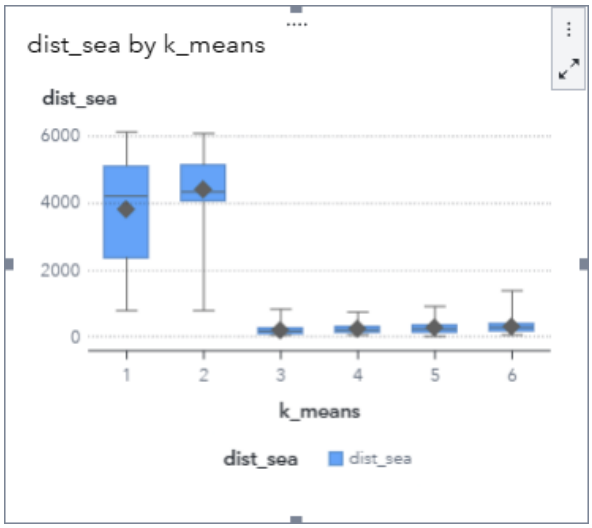

Fig 2.13 Boxplot of dist_sea

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

- The hierarchical clustering are being using SAS procedure while taking some reference from **task>Cluster Observation** (Fig 3.1) and number of clusters selected is 3 (Fig 3.2).
- At 3 clusters, approx. 85% of the eigenvalues are explainable.

```
ods noproctitle;

proc cluster data=casuser.coastal_dataset_final
             method=ward print=10
             ccc
             outtree=work.Cluster_Tree
             plots(only)=(ccc dendrogram);
   var log_dist_sea log_area log_price log_height log_rooms;
   id Name;
run;

proc tree data=work.Cluster_Tree nclusters=3 out=casuser.min_ward_Clusters;
   id Name;
run;

proc casutil;
    save casdata='min_ward_Clusters'
    casout='min_ward_Clusters.sashdat'
    outcaslib='CASUSER'
    replace;
quit;
```

Fig 3.1 SAS procedure

Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	0.34898619	0.12900247	0.4510	0.4510
2	0.21998373	0.12597552	0.2843	0.7353
3	0.09400821	0.02037960	0.1215	0.8568
4	0.07362861	0.03643606	0.0952	0.9519
5	0.03719255		0.0481	1.0000

Root-Mean-Square Total-Sample Standard Deviation	0.393395

Root-Mean-Square Distance Between Observations	1.244025

Cluster History								
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Tie
10	CL25	CL13	192	0.0118	.712	.706	2.11	
9	CL12	CL19	861	0.0167	.695	.692	0.99	
8	CL31	CL20	180	0.0174	.678	.676	0.51	
7	CL18	CL24	358	0.0197	.658	.657	0.38	
6	CL11	CL35	539	0.0205	.638	.633	1.47	
5	CL15	CL8	323	0.0207	.617	.602	4.65	
4	CL6	CL9	1400	0.0431	.574	.560	3.60	
3	CL10	CL5	515	0.0512	.523	.497	5.69	
2	CL4	CL7	1758	0.1801	.343	.339	0.73	
1	CL2	CL3	2273	0.3426	.000	.000	0.00	

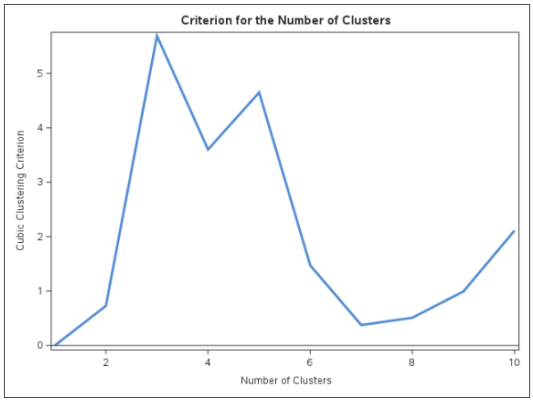

Fig 3.2 Ward's minimum results

Observation

Fig 3.3 Overall variable classification using clustering results

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.

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

- The hierarchical clustering are being using SAS procedure while taking some reference from **task>Cluster Observation** (Fig 4.1) and number of clusters selected is 5 (Fig 4.2).
- At 5 clusters, approx. 100% of the eigenvalues are explainable.

```
ods noproctitle;

proc cluster data=casuser.coastal_dataset_final
             method=complete print=10
             ccc
             outtree=work.Cluster_Tree
             plots(only)=(ccc dendrogram);
   var log_dist_sea log_area log_price log_height log_rooms;
   id Name;
run;

proc tree data=work.Cluster_Tree nclusters=5 out=casuser.complete_max_Clusters;
   id Name;
run;

proc casutil;
    save casdata='complete_max_Clusters'
    casout='complete_max_Clusters.sashdat'
    outcaslib='CASUSER'
    replace;
quit;
```

Fig 4.1 SAS procedure

Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	0.34898619	0.12900247	0.4510	0.4510
2	0.21998373	0.12597552	0.2843	0.7353
3	0.09400821	0.02037960	0.1215	0.8568
4	0.07362861	0.03643606	0.0952	0.9519
5	0.03719255		0.0481	1.0000

Root-Mean-Square Total-Sample Standard Deviation	0.393395

Mean Distance Between Observations	1.141813

Cluster History									
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Norm Maximum Distance	Tie
10	CL25	CL12	238	0.0120	.655	.706	-16	1.7373	
9	CL17	CL18	354	0.0104	.645	.692	-15	1.7495	
8	CL15	CL23	314	0.0074	.637	.676	-12	1.7501	
7	CL9	CL16	1382	0.0251	.612	.657	-14	1.8855	
6	CL13	CL30	260	0.0104	.602	.633	-9.5	1.9056	
5	CL14	CL11	79	0.0107	.591	.602	-3.4	1.9478	
4	CL6	CL8	574	0.1116	.479	.560	-20	2.3013	
3	CL7	CL10	1620	0.1914	.288	.497	-38	2.3493	
2	CL3	CL4	2194	0.2487	.039	.339	-46	2.6789	
1	CL2	CL5	2273	0.0394	.000	.000	0.00	2.8559	

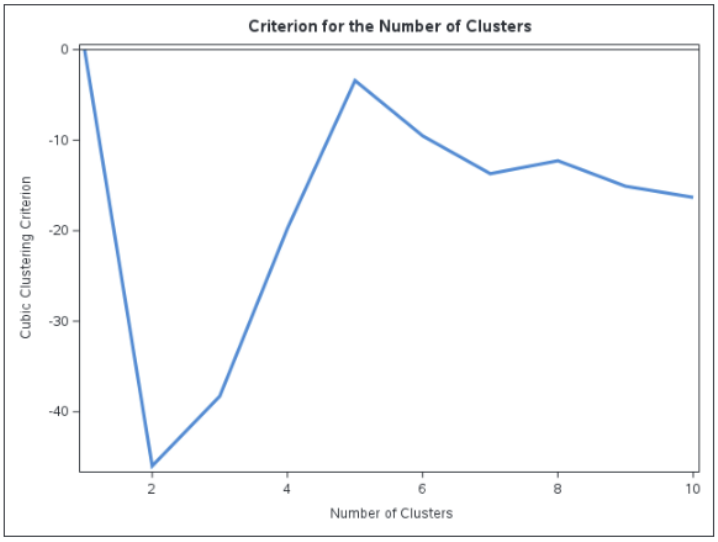

Fig 4.2 Maximum linkage results

Observation

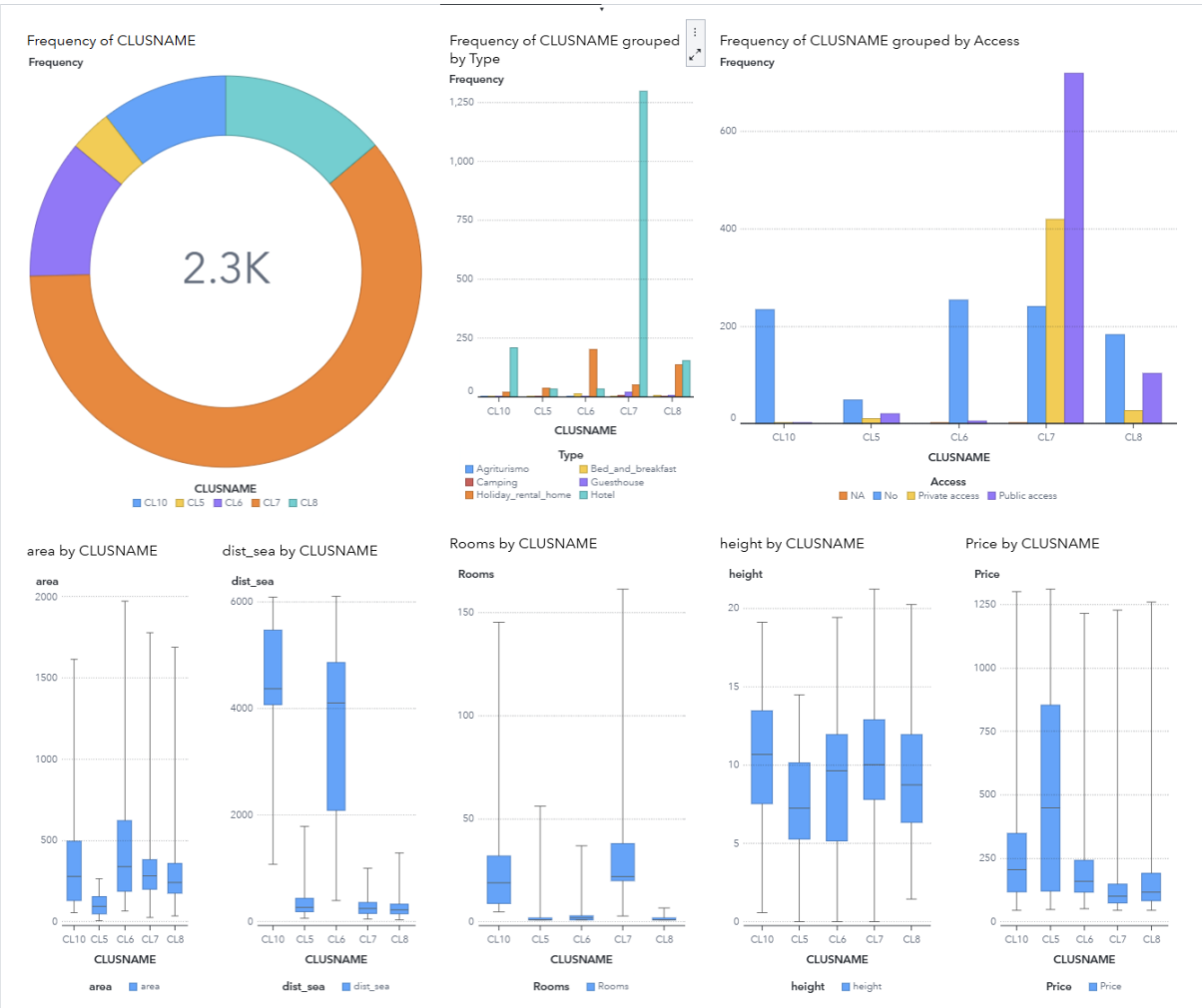

Fig 4.3 Overall variable classification using clustering results

Cluster 7 is the largest, followed by Cluster 8 and Cluster 6.

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.

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:

1. **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.

2. **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.
3. **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.
4. **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:

1. **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.

2. **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.

References

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

https://doi.org/10.1016/j.dib.2025.111921