Segmenting the potential market for the E-Systems®software development company

Alaa Mahjoub

June 2020

1. Introduction

1.1 Background

E-Systems® is a software development company specialized in developing software systems for automating the business of hotels, coffeeshops and restaurants. The company develops three software products:

A- The Hotels Automation Software (HAS)

B- The Coffeeshops Automation Software (CAS)

C- The Restaurants Automation Software (RAS)

The Marketing Department of E-Systems® includes three Teams: A Hotel Automation Marketing Team, a Coffeeshop Automation Marketing Team and a Restaurant Automation Marketing Team.

The company is planning to expand its market by identifying potential overseas customers in relevant national capital cities across the whole world.

1.2 Problem

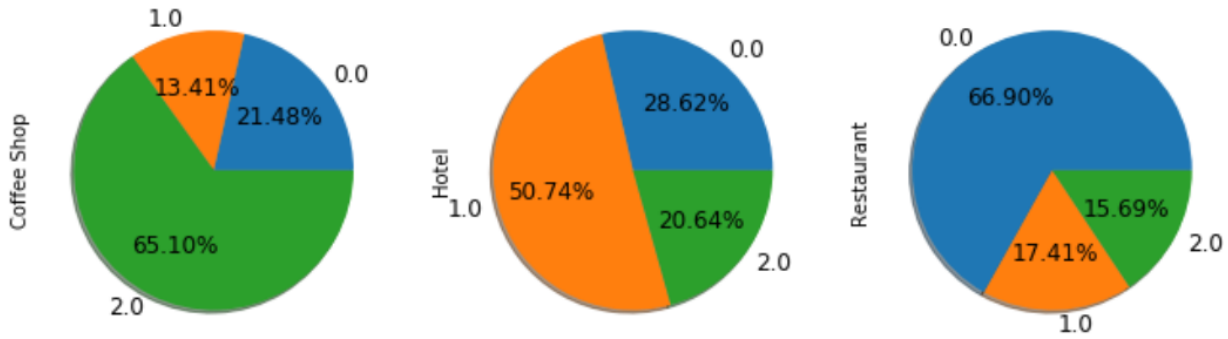
In order to expand its market, E-Systems® adopted a data driven approach and formulated a new market development strategy based on geo-demographic market segmentation. The data which will contribute to the market segmentation process includes:

- the national capital city name and its country name
- the national capital city geographical coordinates (i.e. the longitudes and latitude data of the city)
- the number and the category of potential customers in each national capital city (i.e. the number of hotels, the number of coffeeshops and the number of restaurants in the city)

This project is a data clustering project and it is aimed to segment the national capital cities into three marketing segments (i.e. 3 clusters). Although the above mentioned data will contribute in the segmentation process, the segmentation itself will be done according to the number and the category of potential customers in each city(i.e. the number of hotels, the number of coffeeshops and the number of restaurants). In this approach, each Marketing Team will lead the new market development efforts in one of these three market segments internationally.

Figure 1 below depicts an example of the potential market segmentation, and Figure 2 depicts an example of a geo-demographic market segmentation based on this descriptive analytics data clustering approach.

Figure 1 - An example of potential market segmentation

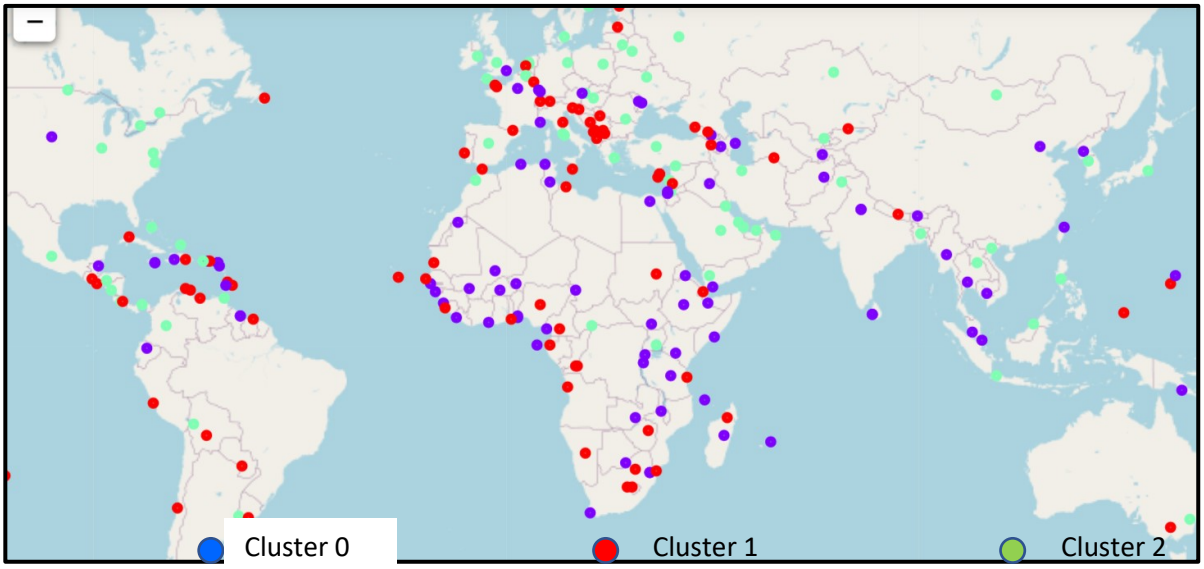


Market Segment 0
(Capital City Cluster 0)

Market Segment 1
(Capital City Cluster 1)

Market Segment 2
(Capital City Cluster 2)

Figure 2 - An example of geo-demographic market segmentation



1.3 Interest

Obviously, it is in the interest of E-Systems® marketing department to know which capital cities will be managed by which marketing team. The size of the new potential market in each segment is also of interest to the department (see Table 1 below for an example). This information will help the Marketing Department develop or acquire the appropriate human capital competencies necessary to manage the new international market development activities. Other audiences who care about this problem include the E-Systems® company Senior Management, E-Systems® Organization and Human Capital Planning Department and E-Systems® shareholders.

Table 1 - An example of potential market segment size (expressed in number of customers in each business line in each cluster)

Cluster Labels	Restaurant	Hotel	Coffee Shop
0.0	388.0	366.0	165.0
1.0	101.0	649.0	103.0
2.0	91.0	264.0	500.0

2. Data acquisition and cleansing

2.1 Data sources

Table 2 below describes the datasets used to build the clusters and their corresponding data sources.

Table 2 - the datasets and their data sources

No	Dataset	Description	Data Source
1	List of world-wide national capital cities	Data fields include City, Country and Notes. See Appendix I for an example of this dataset.	I scraped the following Wikipedia site to obtain this data https://en.wikipedia.org/wiki/List_of_national_capitals
2	Geo-Location data of each national capital city	Data fields include the longitude and latitude coordinates of each national capital. See Appendix II for an example of this dataset.	I obtained this data using the Python geocoding web services API.
3	Potential customers' data	Data fields include the venue name, category, longitude and latitude, See Appendix III for an example of this dataset.	I obtained this data by exploring the national capitals venues using the Foursquare API
4	The world map GIS data	Data of world map with the national capitals across the world. See Appendix IV for an example if this dataset.	I obtained this data using the Folium API

2.2 Data Cleansing

Data of national capitals are scraped from the Wikipedia page using Python. There were some missing data records which I discovered during searching the location data using the national capitals' names extracted from the Wikipedia. After investigation, I discovered that the missing data were due to some comments that were included in the Wikipedia page and put between round parentheses with some of the city names. So, I removed the parentheses and all data within them removed the parentheses and all data within them using the Pandas' based data cleansing module, and then I used the cleansed data to search the locations of the national capitals again. This time I got no missing data. However, I left this cleansing codes code which removes the parentheses and all data within them such that it can be used in future cases, should any update take place on the Wikipedia page.

Also, I have noticed that the column names in the Wikipedia page are not put in standard naming convention. Some column names use special characters, and this jeopardized the Python program code. So, I modified the column name to include only the standard alphabetic character set.

Then I inserted the latitude and longitude coordinate columns structure to the data frame structure of the table read from the Wikipedia page such that I can read the coordinates data from the geocoding web services and include it in the data frame.

I then obtained the national capitals' coordinates data using the geocoding web services. While doing that, I discovered that there are very few missing coordinates data that could not be retrieved by the API. So, I treated this data by displaying exception messages in the data acquisition software module, and then I dropped the rows with NaN values in latitude or longitude fields. This is quite acceptable since these missing data was associated with very few un-famous towns. I then combined the venue data with the location data and the master data acquired from the Wikipedia (see Table 3 below).

I then used Folium to create a World Map with all national capitals superimposed on top and used this map to visually verify the correctness of acquired data on the map (see Appendix IV). Having done all of that, the data quality became quite good and acceptable.

Table 3 - Combined Wikipedia data, location data and venue data

	World Capital	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Abidjan	5.320357	-4.016107	Sofitel Abidjan Hôtel Ivoire	5.327097	-4.004801	Hotel
1	Abidjan	5.320357	-4.016107	Bao Café	5.348778	-3.996881	Coffee Shop
2	Abidjan	5.320357	-4.016107	Etoile du Sud	5.194655	-3.737721	Hotel
3	Abidjan	5.320357	-4.016107	Hotel Madrague	5.195719	-3.740848	Hotel
4	Abu Dhabi	24.474796	54.370576	Sofitel Abu Dhabi Corniche	24.499131	54.367792	Hotel
5	Abu Dhabi	24.474796	54.370576	Jumeirah at Etihad Towers (جميرا أبراج الاتحاد)	24.457974	54.321935	Hotel
6	Abu Dhabi	24.474796	54.370576	The Abu Dhabi EDITION	24.451979	54.336748	Hotel
7	Abu Dhabi	24.474796	54.370576	Jannah Burj Al Sarab	24.501516	54.373405	Hotel
8	Abu Dhabi	24.474796	54.370576	Cartel Coffee Roasters	24.458170	54.356326	Coffee Shop

2.3 Feature Selection

After data cleansing, there were 16,702 samples and to know the total number of features (i.e. the number of venue categories of the national capitals), I calculated the number of unique categories curated from all the returned national capital venues. They were 522 unique venue categories, however, in this market segmentation project, we need only three of these features. These are the features marked as ‘Kept’ in the Feature selection Table 4 below:

Table 4 - Feature selection during data cleaning

No	Feature	Type of variable	Kept/Dropped	Reason
1	hotel Category Venue	Categorical	Kept	We need it to build our market segmentation cluster
2	coffeeshop Category venue	Categorical	Kept	We need it to build our market segmentation cluster
3	restaurant category venue	Categorical	Kept	We need it to build our market segmentation cluster
4	All other categorical variables such as Auto Workshop, Supplement Shop, Women's Store, etc.	Categorical	Dropped	We do NOT need them to build our market segmentation cluster

3. Methodology

3.1 Exploratory data analysis

3.1.1 Exploring the master dataset

The master dataset includes the national capital cities, the country of each city as well as the national capital city coordinates (longitude and latitude). After combining the master dataset as explained in section 2, the master dataset was explored by printing the master dataset data frame, obtaining its summary information and displaying each city on the World Map. Table 5 shows a sample of the master dataset and its attributes. Figure 3 shows that the master dataset includes 260 cities, and Figure 4 depicts the location of each city on the world map.

Table 5 - Sample of the master dataset and the master dataset attributes

	City	Country	lat	lng
0	Abidjan	Ivory Coast	5.32036	-4.01611
1	Yamoussoukro	Ivory Coast	6.80911	-5.27326
2	Abu Dhabi	United Arab Emirates	24.4748	54.3706
3	Abuja	Nigeria	9.06433	7.4893
4	Accra	Ghana	52.4934	4.80368
5	Adamstown	Pitcairn Islands	-25.0667	-130.1
6	Addis Ababa	Ethiopia	9.01079	38.7613
7	Aden	Yemen	12.8333	44.9167
8	Sana'a	Yemen	15.3539	44.2059
9	Algiers	Algeria	36.7754	3.06019
10	Alofi	Niue	-19.0534	-169.919

Figure 3 – Master Dataset Information (260 national capital Cities)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 260 entries, 0 to 259
Data columns (total 4 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   City     260 non-null    object
 1   Country  260 non-null    object
 2   lat      260 non-null    object
 3   lng      260 non-null    object
dtypes: object(4)
memory usage: 8.2+ KB
```

Figure 4 –Location of each national capital city on the World Map

3.1.2 Exploring the venues dataset

The venues dataset of the national capital cities includes: the city name, city longitude, city latitude, the venue name, venue longitude, venue latitude and venue category. After combining the venue dataset with the master dataset as explained in section 2, the venue dataset was explored using the following descriptive statistics charts and summary information printouts:

A- A printout of sample venue dataset and its attributes (Table 6).

B- Venues raw dataset summary information printout (Figure 5). This dataset includes the venue information before dropping the duplicate rows. This figure shows that the venue raw dataset includes 16903 records.

C- Venues dataset summary information printout (Figure 6). This dataset includes the venue information after dropping the duplicate rows. This figure shows that the venue raw dataset includes 16101 records.

D- A bar chart that depicts the total number of venues for each national capital city (Figure 7)

E- Venues dataset descriptive statistics printout (Figure 8)

F- Venues dataset logarithmic scale histogram (Figure 9)

Table 6 - Sample of the venue dataset and its attributes

	World Capital	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Abidjan	5.320357	-4.016107	Sofitel Abidjan Hôtel Ivoire	5.327097	-4.004801	Hotel
1	Abidjan	5.320357	-4.016107	Norima	5.363668	-3.992067	American Restaurant
2	Abidjan	5.320357	-4.016107	Cap Sud	5.298763	-3.987246	Shopping Mall
3	Abidjan	5.320357	-4.016107	Bao Café	5.348778	-3.996881	Coffee Shop
4	Abidjan	5.320357	-4.016107	Pink Club	5.305360	-3.988696	Nightclub
5	Abidjan	5.320357	-4.016107	Nice Cream	5.291398	-3.982492	Ice Cream Shop
6	Abidjan	5.320357	-4.016107	Lifestar	5.324086	-4.015354	Nightclub
7	Abidjan	5.320357	-4.016107	Des Gateaux & Du Pain	5.360270	-3.989671	Bakery
8	Abidjan	5.320357	-4.016107	Di Sorrento	5.288542	-3.987629	Italian Restaurant
9	Abidjan	5.320357	-4.016107	Bushman Café	5.330411	-3.953990	African Restaurant

Figure 5 – Venues raw dataset summary information (16903 national capital cities before dropping the duplicate rows)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16903 entries, 0 to 16902
Data columns (total 7 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   World Capital    16903 non-null  object
 1   Latitude         16903 non-null  float64
 2   Longitude        16903 non-null  float64
 3   Venue            16903 non-null  object
 4   Venue Latitude   16903 non-null  float64
 5   Venue Longitude  16903 non-null  float64
 6   Venue Category   16903 non-null  object
dtypes: float64(4), object(3)
memory usage: 924.5+ KB
```

Figure 6 – Venues dataset summary Information (16101 national capital cities after dropping the duplicate rows)

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 16101 entries, 0 to 16430
Data columns (total 7 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   World Capital    16101 non-null  object
 1   Latitude         16101 non-null  float64
 2   Longitude        16101 non-null  float64
 3   Venue            16101 non-null  object
 4   Venue Latitude   16101 non-null  float64
 5   Venue Longitude  16101 non-null  float64
 6   Venue Category   16101 non-null  object
dtypes: float64(4), object(3)
memory usage: 1006.3+ KB
```

Figure 7 – Sample of the total number of venues for each national capital city (after dropping the duplicate rows)

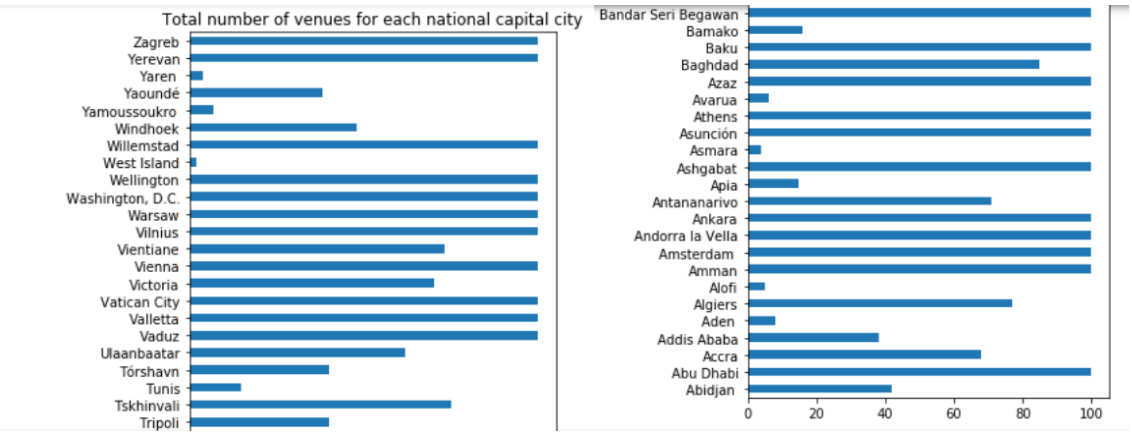
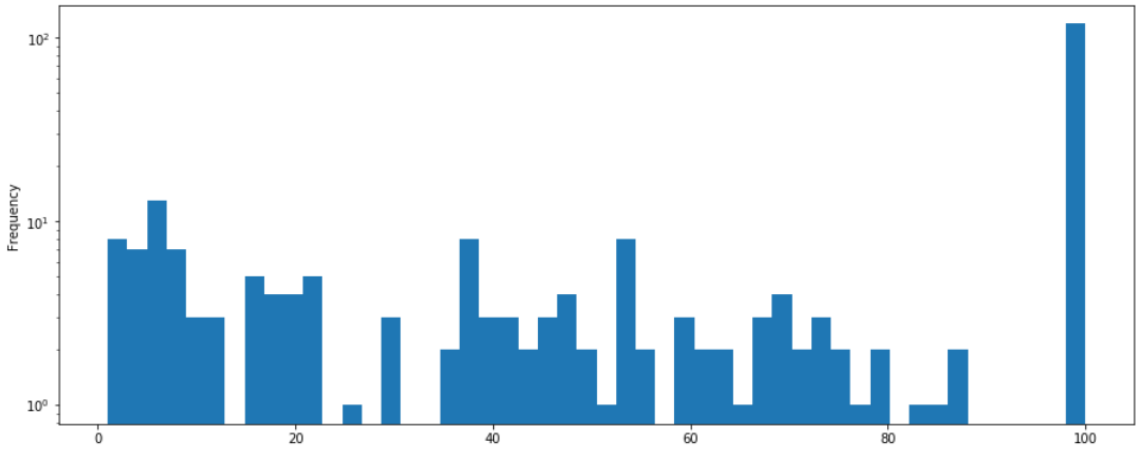


Figure 8 – Venues dataset descriptive statistics (after dropping the duplicate rows)

	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude
count	249.000000	249.000000	249.000000	249.000000	249.000000
mean	65.746988	65.746988	65.746988	65.746988	65.746988
std	37.808252	37.808252	37.808252	37.808252	37.808252
min	1.000000	1.000000	1.000000	1.000000	1.000000
25%	30.000000	30.000000	30.000000	30.000000	30.000000
50%	79.000000	79.000000	79.000000	79.000000	79.000000
75%	100.000000	100.000000	100.000000	100.000000	100.000000
max	100.000000	100.000000	100.000000	100.000000	100.000000

Figure 9– Venues log-scale histogram (after dropping the duplicate rows)

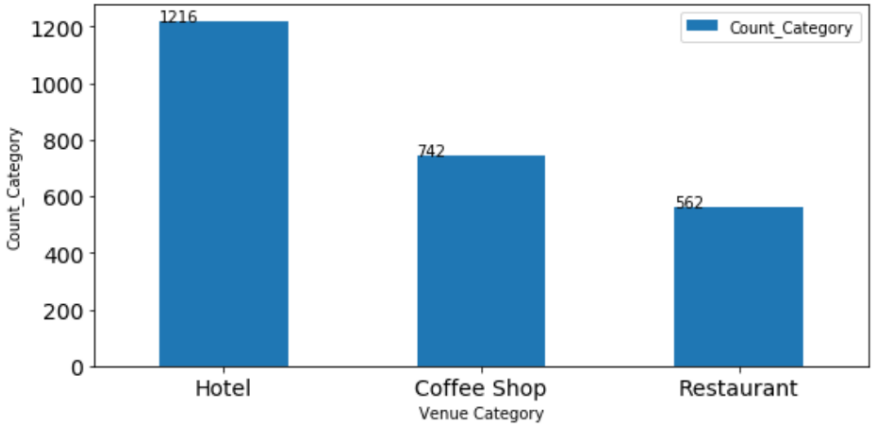


3.1.3 Exploring the features dataset

The features dataset is a subset of the venues' dataset. It includes the hotels venues, the coffeeshop venues and the restaurant venues. It consists of 2520 cleansed data records. These records were obtained by filtering the venues dataset to obtain only the venues belonging to the hotels, coffeeshop and the restaurant categories. The selected features dataset was explored using the following descriptive statistics charts and summary information printouts:

A- A bar chart that shows the total number of each venue category in the selected features dataset (i.e. the hotels, coffeeshops and restaurants), and a printout of the total number of all venues in the selected features dataset (Figure 10).

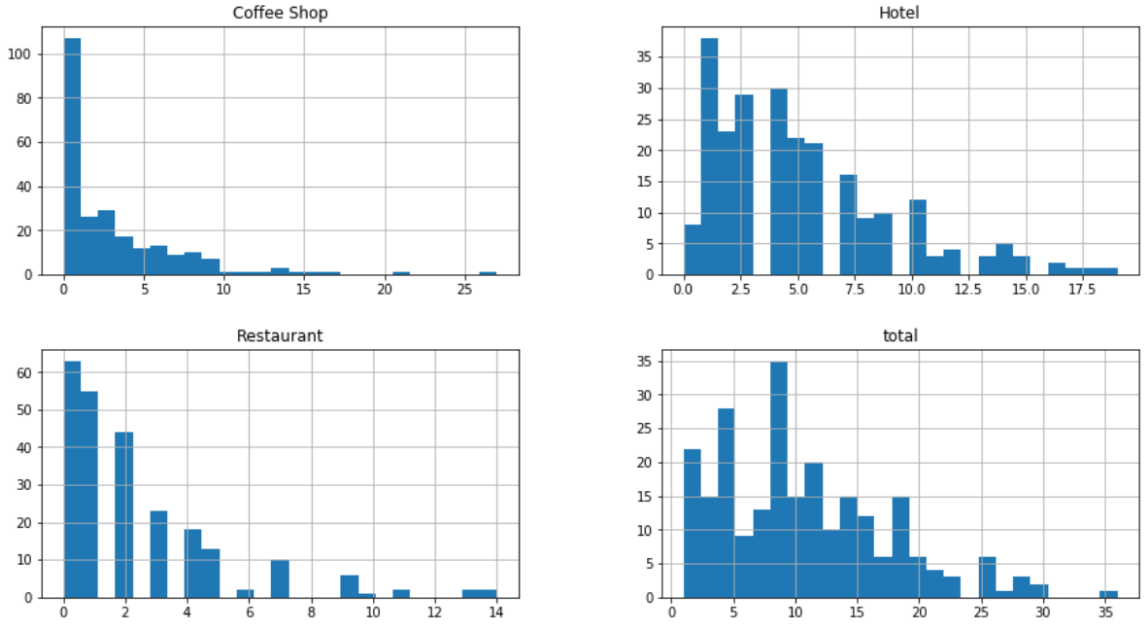
Figure 10 - Number of venues in each feature category, and total number of venues in the features dataset (2220)




```
Venue Category    HotelCoffee ShopRestaurant
Count_Category                          2520
dtype: object
```

B- Four histograms that depict characteristics of the selected features dataset, i.e. the hotels, coffeeshops, restaurants and the total number of selected features (Figure 11).

Figure 11 - Selected features dataset histograms



3.1.4 Exploring the relationship between the cities and the features

The relationship between the national capital cities and each feature in the features dataset is explored using the following descriptive statistics charts, maps and summary information printouts:

A- A bar chart that shows the relationship between the cities and each feature in the featuresdataset (i.e. the number of hotels, the number of coffeeshops and the number of restaurants in the city). This bar chart is depicted in Figure 12.

B- A map that shows the density of total number of venues in each national capital city (i.e. the number of hotels + the number of coffeeshops + the number of restaurants). In order to show the density of venue distribution across the cities, the data was classified into three type of densities (low density, medium density and high density) and a colour coding was used to depict each type of density on the map (Figure 13).

Figure 12 - Relationship between the cities and each selected feature

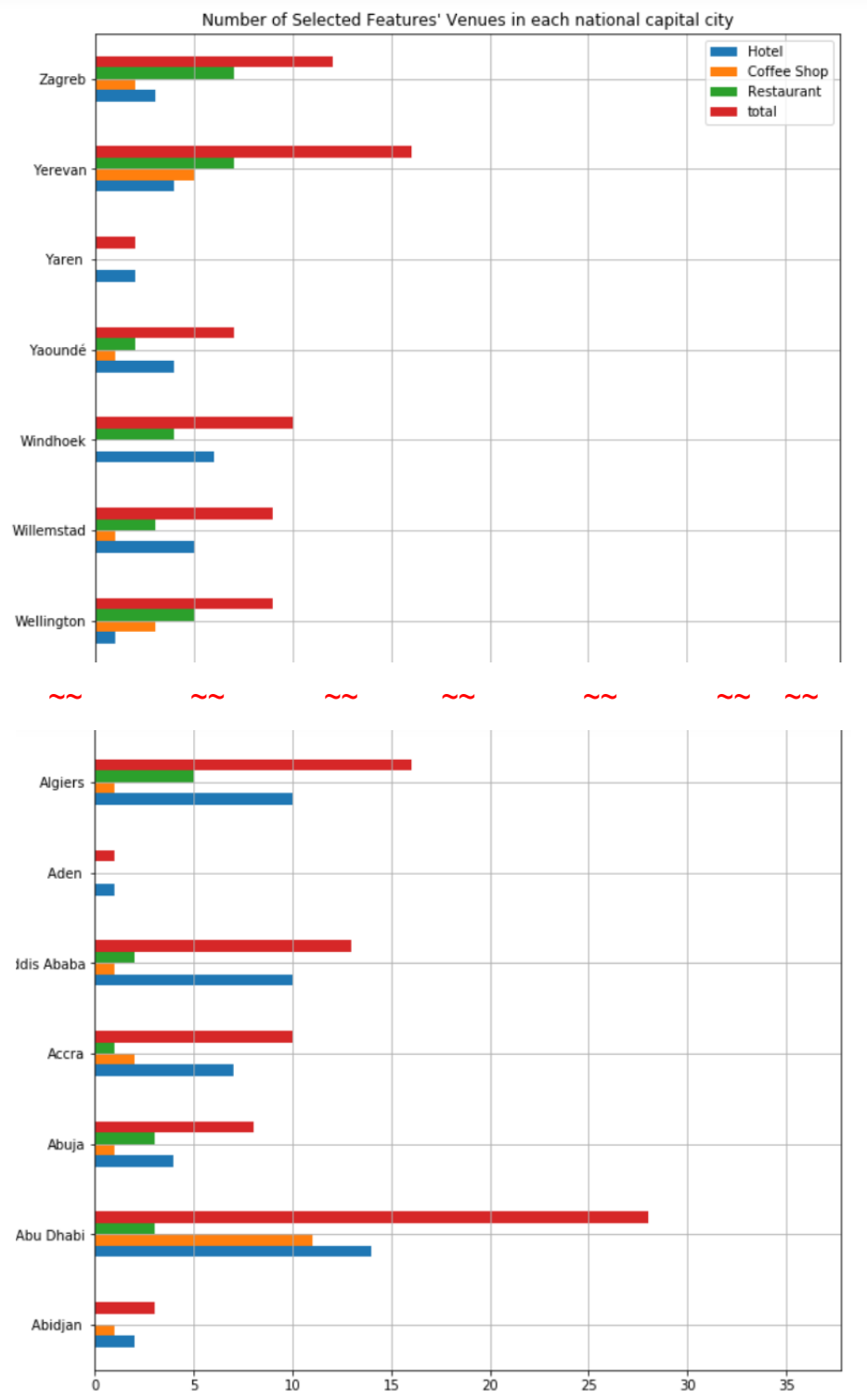
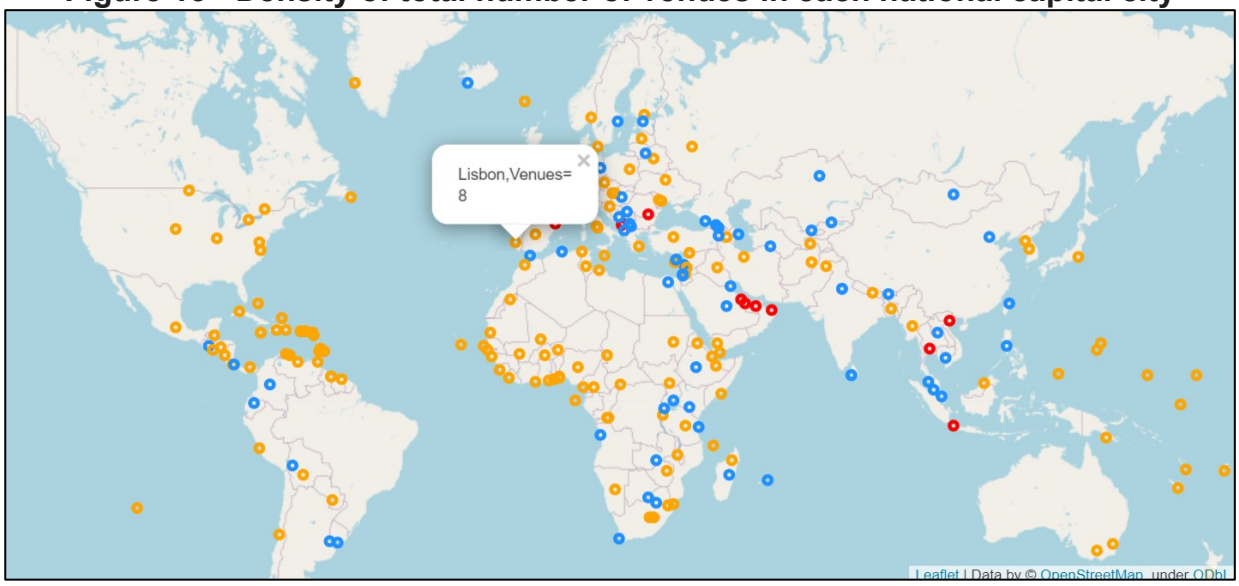
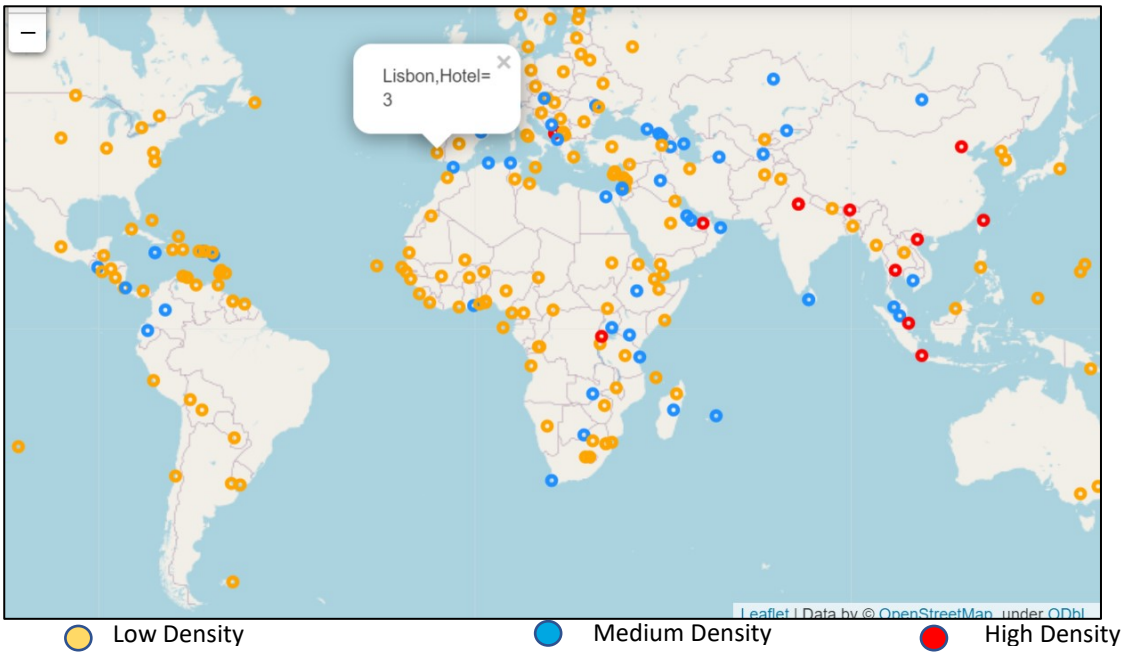


Figure 13 - Density of total number of venues in each national capital city



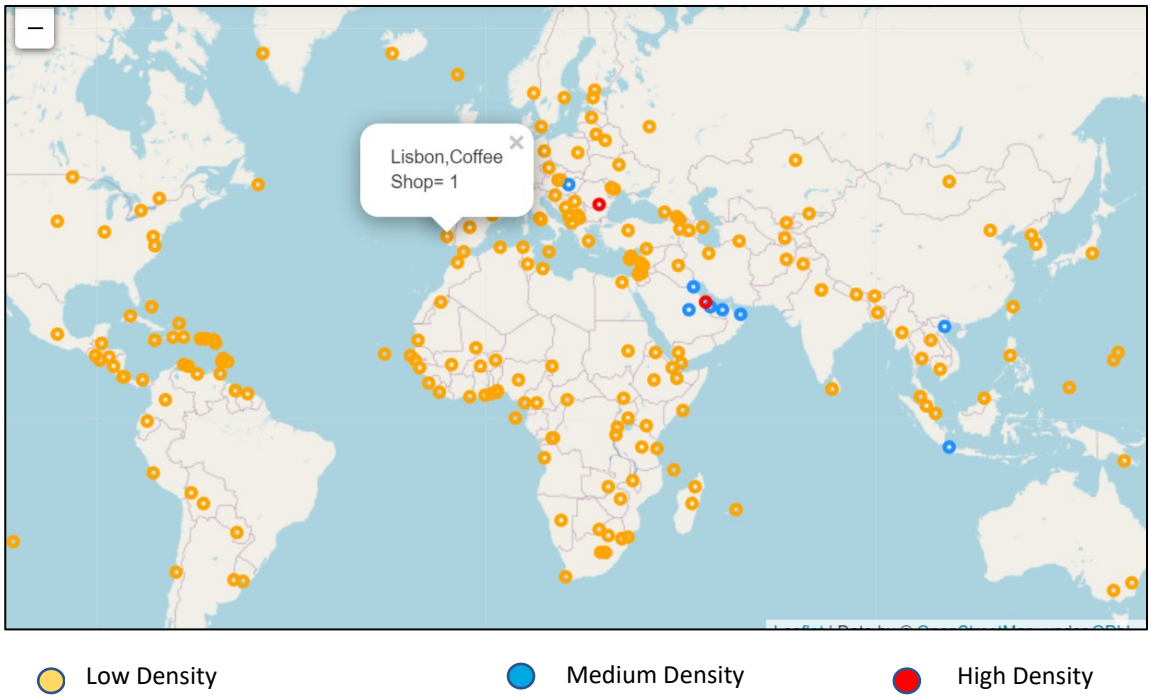
C- A map that shows the density of the hotels in each national capital city. In order to show the density of the hotels' distribution across the cities, the data was classified into three type of densities (low density, medium density and high density) and a colour coding was used to depict each type of density on the map (Figure 14).

Figure 14 - Density of total number of hotels in each national capital city



D- A map that shows the density of the coffeeshops in each national capital city. In order to show the density of the coffeeshops distribution across the cities, the data was classified into three type of densities (low density, medium density and high density) and a colour coding was used to depict each type of density on the map (Figure 15).

Figure 15 - Density of total number of coffeeshops in each national capital city



E- A map that shows the density of the restaurants in each national capital city. In order to show the density of the coffeeshops distribution across the cities, the data was classified into three type of densities (low density, medium density and high density) and a colour coding was used to depict each type of density on map (Figure 15).

F- To verify the correctness on the data displayed on the above mentioned four maps, the features data of Lisbon city was extracted from the features dataset and displayed in the form of a data frame and compared with the corresponding data displayed on the maps. Figure 17 depicts the features data of Lisbon city.

Figure 16 - Density of total number of restaurants in each national capital city

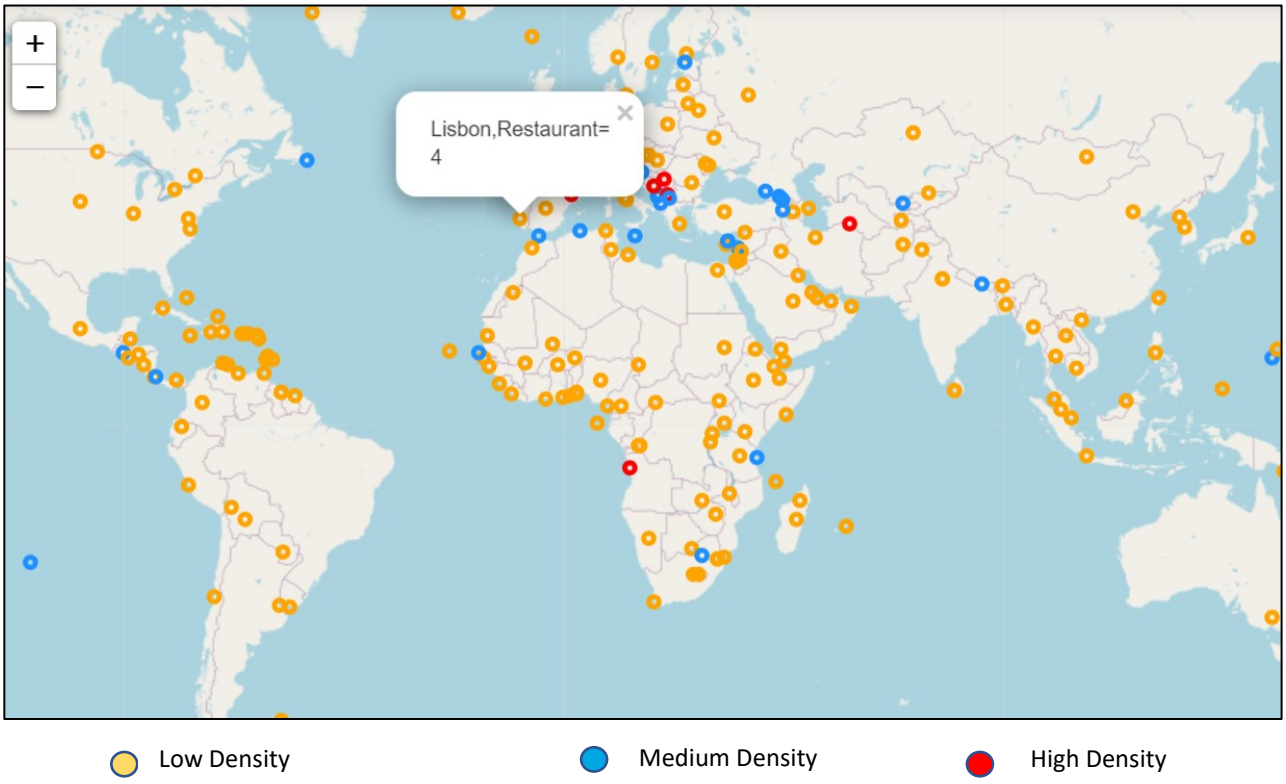


Figure 17 – The features data of Lisbon city

	World Capital	Coffee Shop	Hotel	Restaurant	total	City	Country	lat	lng	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	marker_color
108	Lisbon	1	3	4	8	Lisbon	Portugal	38.7078	-9.13659	0.0	Restaurant	Hotel	Coffee Shop	#FFA500

3.2 Inferential statistical testing

No inferential statistical testing was needed since the required dataset was fully acquired through the Internet. The dataset was then cleansed, filtered and prepared to generate the full master dataset and the full features dataset required for building the clustering models.

3.3 Model selection

3.3.1 Selecting the Machine Learning Technique

Table 7 summarizes the most important predictive and descriptive machine learning techniques. It also shows why I selected the clustering technique to solve this business problem.

Table 7 – Justification of selecting the clustering technique

No	ML Technique	Analytical Approach	Description	Selected?	Reason
1	Regression	Predictive	Supervised learning technique used for predicting a continuous value	No	A- The business problem solution is descriptive in nature. B- The data is unlabelled, and the process must be unsupervised
2	Classification	Predictive	Supervised learning technique used for predicting the class or category of a case	No	A- The business problem solution is descriptive in nature. B- The data is unlabelled, and the process must be unsupervised
3	Recommender systems	Predictive	Supervised or unsupervised learning technique used to offer relevant suggestions to users. Categorized as either a collaborative filtering or a content-based system	No	A- The business problem solution is descriptive in nature. B- The data is unlabelled, and the process must be unsupervised C- No data for similar companies' performance is available

No	ML Technique	Analytical Approach	Description	Selected?	Reason
4	Clustering	Descriptive	Unsupervised learning technique used for finding groups of similar cases, for example, can be used for customer segmentation	**Yes**	A- The business problem needs to find similar national capital cities (i.e. market segments) in order to target them by specific marketing teams in the company B- The data is unlabelled, and the process must be unsupervised
5	Association	Descriptive	A learning technique (commonly unsupervised) used for finding items or events that often co-occur	No	A- The business problem does not need finding items that often co-occur
6	Anomaly detection	Descriptive	Supervised, unverified or semi-supervised learning technique used for discovering abnormal and unusual cases	No	A- The business problem does not need to detect anomalies
7	Sequence mining	Descriptive	A learning technique (commonly unsupervised) used for determining sequential patterns in data	No	A- The business problem does not need to detect sequential patterns in data

No	ML Technique	Analytical Approach	Description	Selected?	Reason
8	Dimension reduction	Descriptive	Unsupervised learning technique used for reducing the size of data	No	A- No need to reduce data size. Data size is not too large (260 cities and 2520 venues).

3.3.2 Selecting the Machine Learning Model

Having selected the clustering technique, the second step in my model selection approach was to decide which clustering model is suitable for solving the business problem.

Table 8 summarizes the most important clustering models and shows why I selected the K-Means clustering model and the agglomerative clustering model to solve the business problem.

Table 8 – Justification of clustering models selection

No	Clustering Model	Description	Selected?	Reasons
1	K-Means clustering	It divides the data into K non-overlapping subsets or clusters without any cluster internal structure or labels (unsupervised algorithm)	**Yes**	A- It is an easy and simple model, with fewer hyperparameters than the other clustering algorithms (we mainly need to specify the K value). B- In our case, there is no difficulty in determining the value of K since a specific number of market segments (3) is already predefined by the management of E-Systems® C- We have a medium size dataset (260 cities and 2520 venues), and this algorithm is optimal for dealing with medium or large size datasets D- In our case, there is no large computation cost during runtime especially because the sample size is not large (260 cities and 2520 venues). E- In our case, we do not need to care about the complexity normally associated with providing a proper scaling (for fair treatment among features) since all the features are homogeneous and hence no scaling is needed (each feature represents

No	Clustering Model	Description	Selected?	Reasons
				the number of specific category of venues in the city). F- In our case, we are not interested in the notion of outliers (since every single city must be classified in a distinct market segment). Therefore, using this algorithm which has no notion of outliers, does not represent a problem in our case.
2	Hierarchical clustering	It builds a hierarchy of clusters where each node is a cluster consisting of the clusters of its daughter nodes.	**Yes**	A- Hierarchical clustering is normally used for small size datasets. For the size of our dataset (260 cities), it becomes difficult to use the dendrogram. However, this does not present an inhibitor to use this model in our business problem. B- For the size of our dataset, K-means is more efficient. Hierarchical clustering takes longer computation times in comparison with K-means. However, in our case, the algorithm would not take too long computation times because the size of the dataset is not too big.
3	Density-based spatial clustering of applications with noise (DBSCAN)	It groups together points that are closely packed together (points with many nearby neighbors)	No	A- DBSCAN does not fit our business problem (in which every single city must be classified in a distinct market segment) since DNSCAN has the notion of noise (outliers) and it ignores less dense areas or noises based on the two parameters: radius and minimum points B- DBSCAN needs a careful selection of its parameters. The radius and the minimum points parameters are indeterministic in our case, since no constraints are imposed by the company management on them. C- DBSCAN is much slower than K-Means D- DBSCAN doesn't work well over clusters with different densities

3.4 Applying the K-Means clustering model to the business problem

Based on the justification described in the previous section, I applied the K-Means clustering model to segment the E-Systems® potential market into three distinct segments. The model was applied using the parameters shown in Table 9 and the datasets described in Table 10. The obtained results are described in section 4.

Table 9 – Parameters used in the K-Means Clustering model

N0	Parameter	Value
1	Number of clusters (K)	K=3 i.e. we have 3 clusters (since three market segments are specified as business requirement in the problem definition section (Section 1)
2	Distance calculation method	Euclidean Distance (default)

Table 10 – Description of the used datasets

N0	Dataset	Dataset Type	Description	Number of records
1	National capital cities (raw)	Master data	Raw data of the national capital cities (extracted from the Wikipedia website)	260
2	National capital cities (used)	Master data	Data of national capital cities that have venue information in the Four-Square database	239
3	National capital cities venues	Features data	Data of the venues belonging to the 239 national capital cities	2520

3.5 Applying the hierarchical agglomerative clustering model to the business problem

Based on the justification described in the previous section, I have also applied the hierarchical agglomerative clustering model to segment the E-Systems® potential market into three distinct segments. The model was applied using the parameters shown in Table 11 and the same datasets described before in Table 10. The obtained results are described in section 4.

Table 11 – Parameters used in the hierarchical agglomerative clustering model

N0	Parameter	Value
1	Number of clusters (K)	K=3
2	Distance calculation method	- single - complete - average

4. Results

4.1 Results of the K-Means clustering model

Table 12 shows the number of cities in each of the three clusters (i.e. in each market segment), and Table 13 shows the centroid value of each of them.

Table 12 – Number of cities in each market segment

Clus_km	World Capital	Coffee Shop	Hotel	Restaurant
0	77	77	77	77
1	80	80	80	80
2	82	82	82	82

Total = 239 National capital cities

Table 13 – Centroid value of market segments

Clus_km	Coffee Shop	Hotel	Restaurant
0	0.085506	0.820830	0.093664
1	0.105860	0.456108	0.438032
2	0.563100	0.309666	0.127234

Figures 18,19, 20 and 21 show the distribution of cities based on the frequency of occurrence of coffeeshop, hotel and restaurant venues in each city.

Figure 18 – Distribution of cities based on the frequency of occurrence of coffeeshop and hotel venues

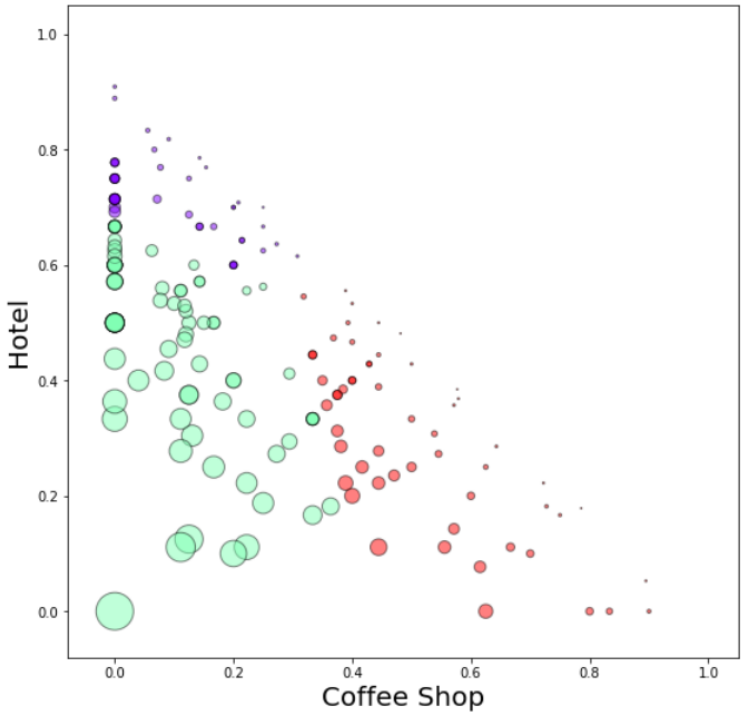


Figure 19 – Distribution of cities based on the frequency of occurrence of hotel and restaurant venues

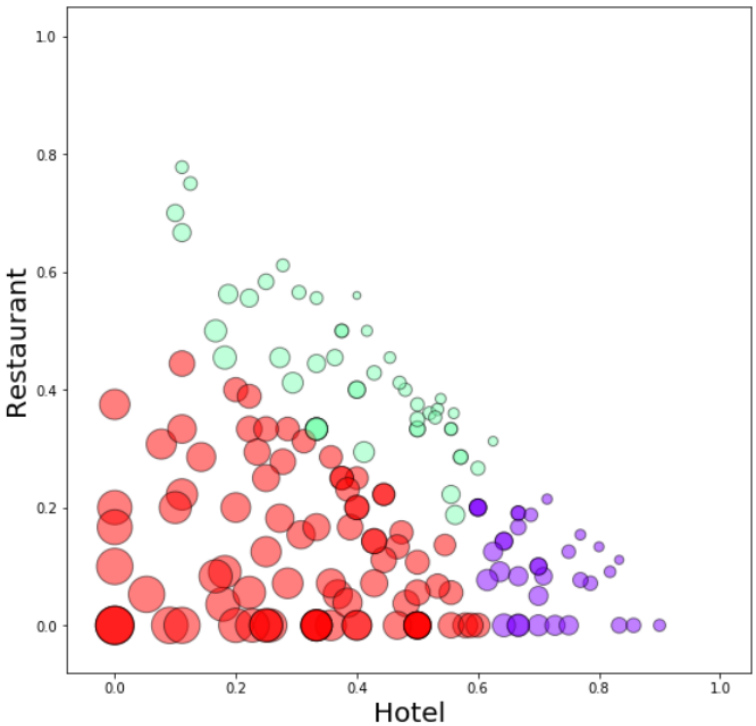


Figure 20 – Distribution of cities based on the frequency of occurrence of coffeeshop and restaurant venues

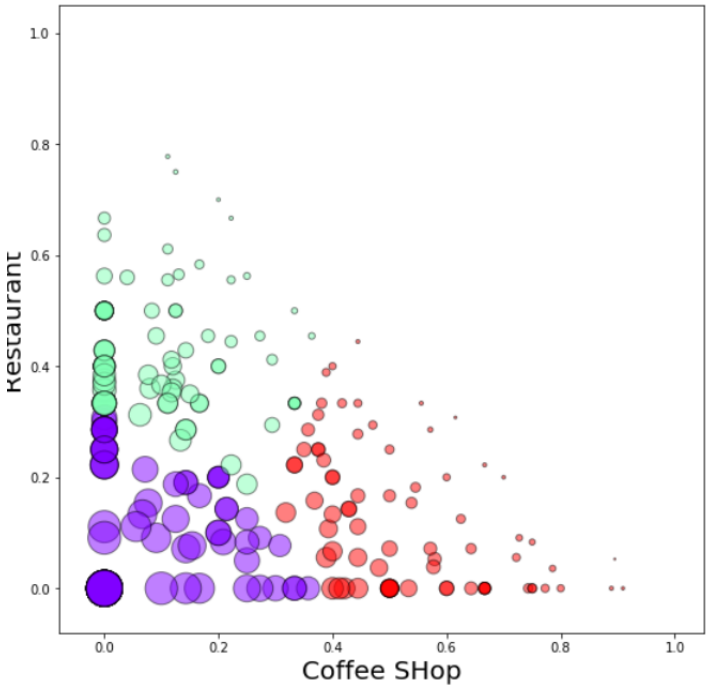


Figure 21 – Distribution of cities based on the frequency of occurrence of coffeeshop, hotel and restaurant venues

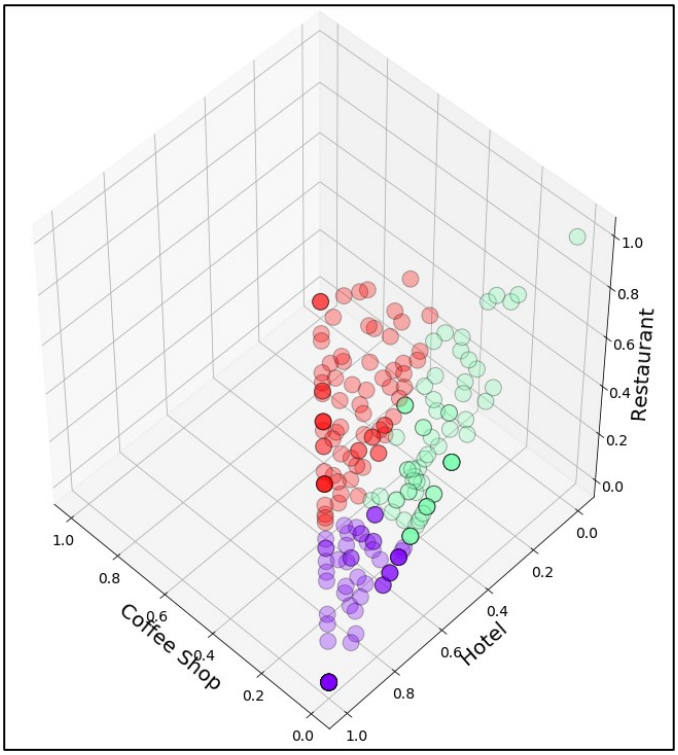


Table 14 shows the number of coffeeshops, hotels and restaurants in each market segment. The same results are illustrated by the bar chart in Figure 22. Figure 23 depicts the density of coffeeshops, hotels and restaurants in each market segment

Table 14 – Number of coffeeshops, hotels and restaurants in each market segment

Cluster Labels	Coffee Shop	Hotel	Restaurant	total
0.0	535.0	337.0	141.0	1013.0
1.0	92.0	365.0	353.0	810.0
2.0	81.0	521.0	76.0	678.0

Figure 22 – Number of coffeeshops, hotels and restaurants in each market segment

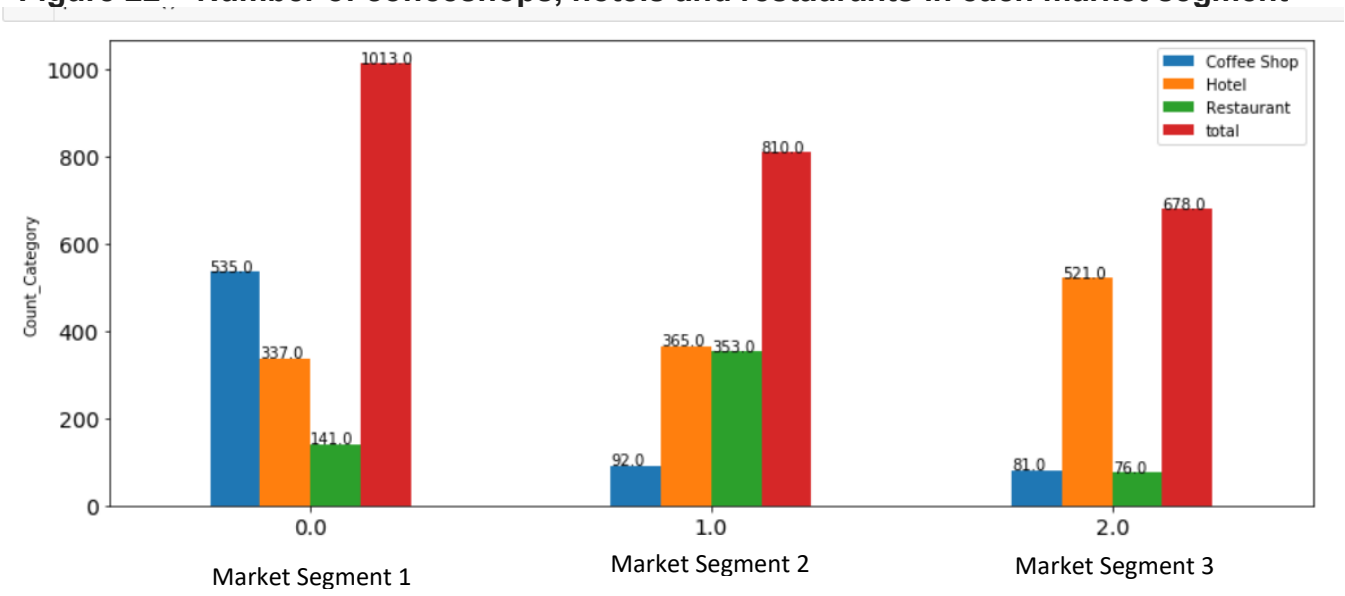


Figure 23 – Density of coffeeshops, hotels and restaurants in each market segment

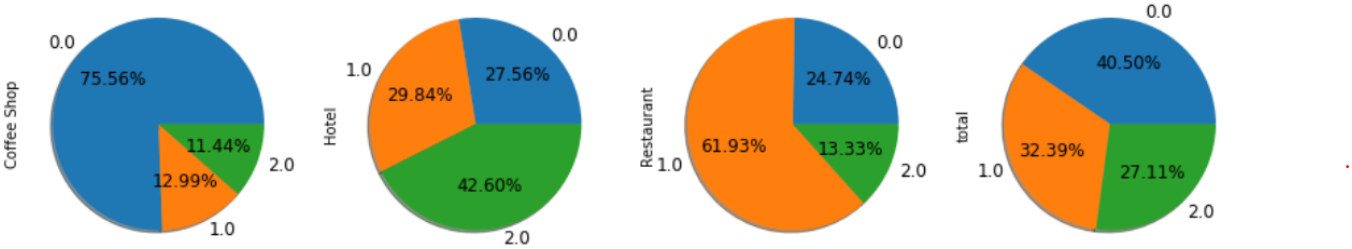


Figure 24 shows the of national capital cities on the three market segments, and Tables 15, 16 and 17 show sample reports of the national capitals of each market segment.

Figure 24 – Distribution of national capital cities on the three market segments

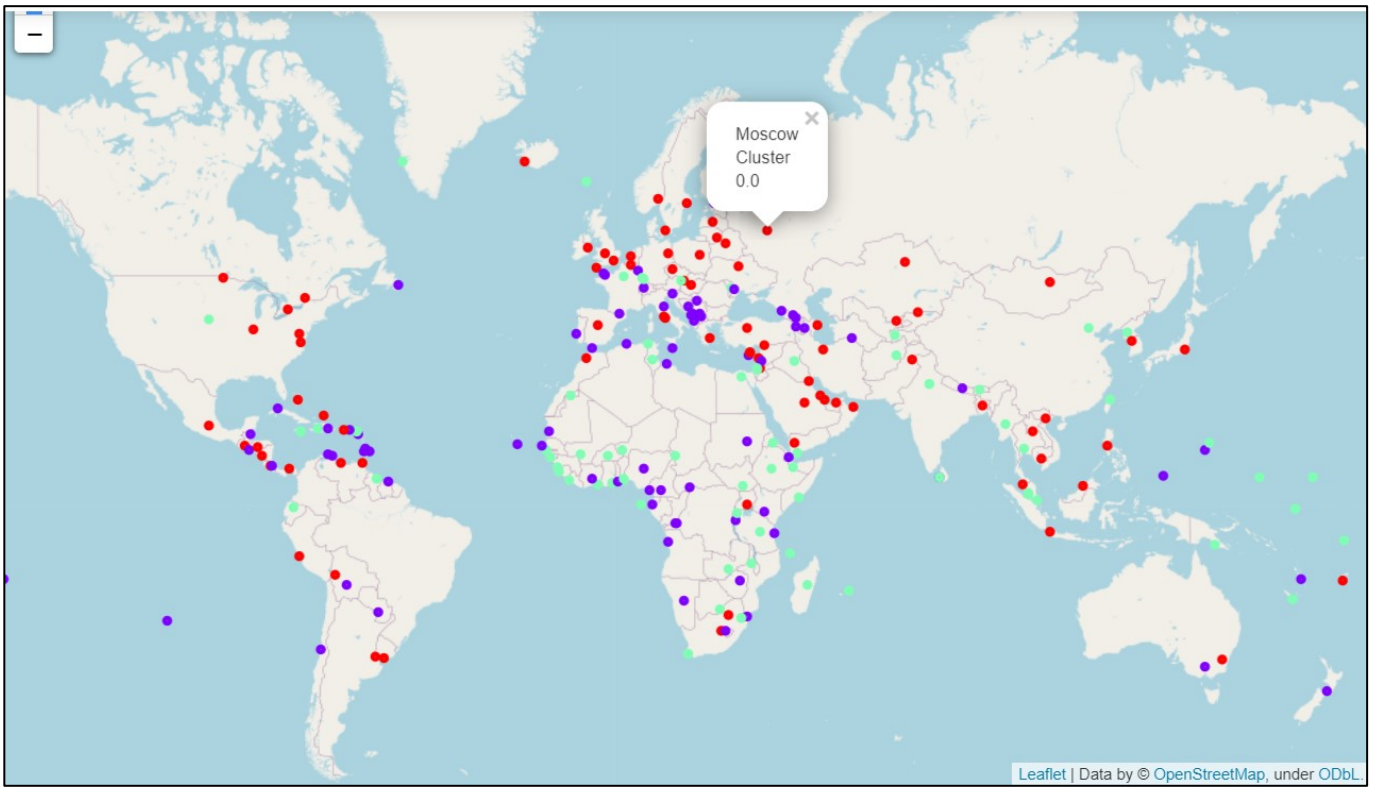


Table 15 – Report sample: National capitals of market segments 1

	World Capital_x	Country	Coffee Shop	Hotel	Restaurant	total	Venue	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Manama	Bahrain	26	9	0	35	100	0.0	Coffee Shop	Hotel	Restaurant
1	Jakarta	Indonesia	12	16	2	30	100	0.0	Hotel	Coffee Shop	Restaurant
2	Abu Dhabi	United Arab Emirates	11	14	3	28	100	0.0	Hotel	Coffee Shop	Restaurant
3	Doha	Qatar	12	12	4	28	100	0.0	Hotel	Coffee Shop	Restaurant
4	Hanoi	Vietnam	13	13	1	27	100	0.0	Hotel	Coffee Shop	Restaurant
5	Muscat	Oman	15	10	1	26	100	0.0	Coffee Shop	Hotel	Restaurant
6	Baku	Azerbaijan	7	12	3	22	100	0.0	Hotel	Coffee Shop	Restaurant
7	Pretoria	South Africa	8	6	7	21	100	0.0	Coffee Shop	Restaurant	Hotel
8	Guatemala City	Guatemala	7	8	5	20	100	0.0	Hotel	Coffee Shop	Restaurant
9	Riyadh	Saudi Arabia	16	4	0	20	100	0.0	Coffee Shop	Hotel	Restaurant

Table 16 – Report sample: National capitals of market segments 2

	World Capital_x	Country	Coffee Shop	Hotel	Restaurant	total	Venue	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	St. Peter Port	Guernsey	3	16	11	30	100	1.0	Hotel	Restaurant	Coffee Shop
1	Podgorica	Montenegro	3	12	10	25	100	1.0	Hotel	Restaurant	Coffee Shop
2	Cetinje	Montenegro	2	14	9	25	100	1.0	Hotel	Restaurant	Coffee Shop
3	Andorra la Vella	Andorra	1	10	14	25	100	1.0	Restaurant	Hotel	Coffee Shop
4	St. Helier	Jersey	3	13	9	25	100	1.0	Hotel	Restaurant	Coffee Shop
5	Ashgabat	Turkmenistan	3	7	13	23	100	1.0	Restaurant	Hotel	Coffee Shop
6	Sarajevo	Bosnia and Herzegovina	0	8	14	22	100	1.0	Restaurant	Hotel	Coffee Shop
7	Sukhumi	Abkhazia	0	12	7	19	82	1.0	Hotel	Restaurant	Coffee Shop
8	Belgrade	Serbia	4	4	10	18	100	1.0	Restaurant	Hotel	Coffee Shop
9	Luanda	Angola	2	5	11	18	53	1.0	Restaurant	Hotel	Coffee Shop
10	Tirana	Albania	2	9	6	17	100	1.0	Hotel	Restaurant	Coffee Shop

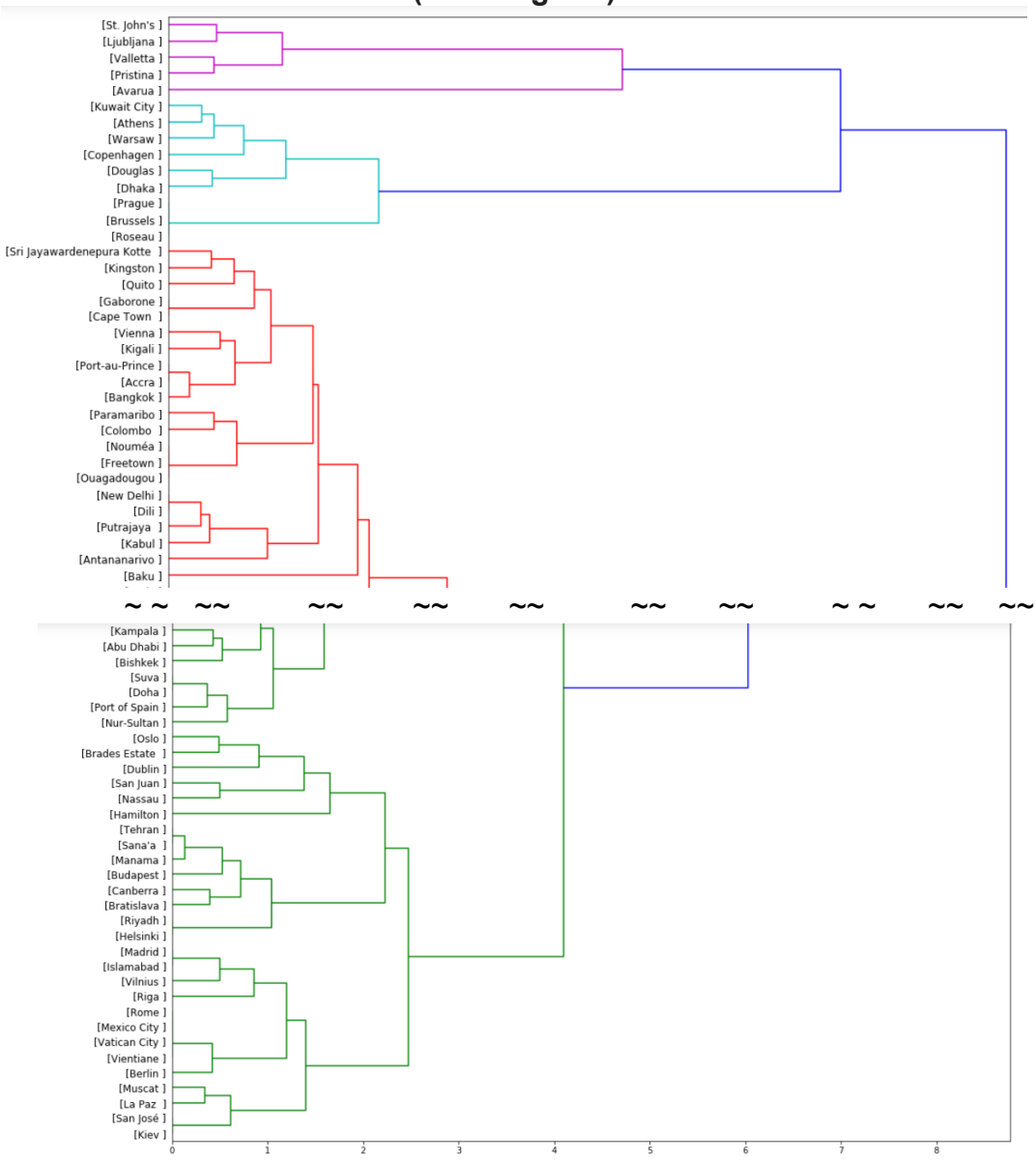
Table 17 – Report sample: National capitals of market segment 3

	World Capital_x	Country	Coffee Shop	Hotel	Restaurant	total	Venue	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Bangkok	Thailand	5	17	2	24	100	2.0	Hotel	Coffee Shop	Restaurant
1	Thimphu	Bhutan	0	20	2	22	51	2.0	Hotel	Restaurant	Coffee Shop
2	New Delhi	India	3	14	4	21	100	2.0	Hotel	Restaurant	Coffee Shop
3	Dili	East Timor	3	14	4	21	100	2.0	Hotel	Restaurant	Coffee Shop
4	Singapore	Singapore	2	18	0	20	100	2.0	Hotel	Coffee Shop	Restaurant
5	Kigali	Rwanda	5	14	1	20	51	2.0	Hotel	Coffee Shop	Restaurant
6	Beijing	China	3	15	0	18	100	2.0	Hotel	Coffee Shop	Restaurant
7	Putrajaya	Malaysia	2	11	3	16	100	2.0	Hotel	Restaurant	Coffee Shop
8	Sri Jayawardenepura Kotte	Sri Lanka	4	10	2	16	100	2.0	Hotel	Coffee Shop	Restaurant

4.2 Results of the hierarchical agglomerative clustering model

When I applied the hierarchical agglomerative clustering model on this business problem, I obtained the same results generated from the K – Means clustering model. Also, the results of the hierarchical agglomerative clustering model did not change when I used different options for the distance calculation method (i.e. the single, complete and average distance calculation methods). Figure 25 shows the results of applying the hierarchical agglomerative clustering model to the business problem (in the form of a dendrogram.)

Figure 25 – Result of the hierarchical agglomerative clustering model (Dendrogram)



5. Discussion

The results obtained from the K-Means clustering algorithm and the agglomerative clustering algorithm are identical, and both algorithms have provided a good solution to the business problem. As shown in Figures 22 and 23 we can notice that:

A- Market segment 1 (cluster 0) is a “coffeeshop oriented” market segment: About 75% of the world-wide coffeeshops’ target market belongs to the cities of this market segment. The remaining 25% of coffeeshop world-wide target market is almost equally distributed between the other two market segments.

B- Market segment 2 (cluster 1) is a “restaurant oriented” market segment: About 62% of the world-wide restaurants’ target market belongs to the cities of this market segment. The remaining world-wide restaurants’ target market is shared between the market segments 1 and 3 with a share of 24% and 13% respectively.

C- Market segment 3 (cluster 2) is a “hotel oriented” market segment About 42% of the world-wide hotels’ target market belongs to the cities of this market segment. The remaining 58% of the world-wide hotels’ target market is almost equally shared between market segments 1 and 2.

It is also worth noting that the total number of customers in each market segment is somewhat different (1013, 810 and 670 for market segments 1,2, and 3 respectively). Based on this data, the management of E-Systems® is advised to consider adjusting its organization structure and uplifting its human capital capabilities in order to be able to successfully implement the new marketing strategy and cope with the requirements of the new world-wide market.

6. Conclusion

In this study, I used the K-Means and the Agglomerative clustering machine learning techniques to segment the new world-wide market of E-Systems®. I identified the frequency of occurrence of hotel, restaurants and coffeeshops as the most important features that affect the segmentation of this potential market. I built both K-Means clustering model and Agglomerative clustering model to build the market segments. These models can be very useful in helping E-Systems® management in several ways. For example, it could help develop a new organization chart and plan the human capital and competencies necessary to implement the company’s new marketing strategy.

Appendix I – Example of dataset 1
the Wikipedia List of world-wide national capitals

City/Town	Country/Territory	Notes
Abidjan (former capital; still has many government offices)	Ivory Coast	
Yamoussoukro (official)		
Abu Dhabi	United Arab Emirates	
Abuja	Nigeria	Lagos was the capital from 1914 to 1991.
Accra	Ghana	
Adamstown	Pitcairn Islands	British Overseas Territory.
Addis Ababa	Ethiopia	
Aden (de facto, temporary)	Yemen	Sana'a has been occupied by Houthis rebels since February 2015. Aden is Yemen's acting capital. See also: Yemeni Civil War (2015–present).
Sana'a (de jure)		
Algiers	Algeria	
Alofi	Niue	Self-governing in free association with New Zealand.
Amman	Jordan	
Amsterdam (official)		The Dutch constitution refers to Amsterdam as the "capital".

Appendix II – Example of dataset 2
Geo-Location data of each national capital from the geocoding web services

	City	Country	lat	lng
0	Abidjan	Ivory Coast	5.32036	-4.01611
1	Yamoussoukro	Ivory Coast	6.80911	-5.27326
2	Abu Dhabi	United Arab Emirates	24.4748	54.3706
3	Abuja	Nigeria	9.06433	7.4893
4	Accra	Ghana	52.4934	4.80368
5	Adamstown	Pitcairn Islands	-25.0667	-130.1
6	Addis Ababa	Ethiopia	9.01079	38.7613
7	Aden	Yemen	12.8333	44.9167
8	Sana'a	Yemen	15.3539	44.2059
9	Algiers	Algeria	36.7754	3.06019
10	Alofi	Niue	-19.0534	-169.919

Appendix III – Example of dataset 3

National capitals important venues from Foursquare API.

	World Capital	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Abidjan	5.320357	-4.016107	Sofitel Abidjan Hôtel Ivoire	5.327097	-4.004801	Hotel
1	Abidjan	5.320357	-4.016107	Norima	5.363668	-3.992067	American Restaurant
2	Abidjan	5.320357	-4.016107	Cap Sud	5.298763	-3.987246	Shopping Mall
3	Abidjan	5.320357	-4.016107	Bao Café	5.348778	-3.996881	Coffee Shop
4	Abidjan	5.320357	-4.016107	Pink Club	5.305360	-3.988696	Nightclub
5	Abidjan	5.320357	-4.016107	Nice Cream	5.291398	-3.982492	Ice Cream Shop
6	Abidjan	5.320357	-4.016107	Lifestar	5.324086	-4.015354	Nightclub
7	Abidjan	5.320357	-4.016107	Des Gateaux & Du Pain	5.360270	-3.989671	Bakery
8	Abidjan	5.320357	-4.016107	Di Sorrento	5.288542	-3.987629	Italian Restaurant

Appendix IV – Example of dataset 4

The world map GIS data from Folium