Location Intelligence Data Clustering

Authors:
Aleksandra Kwiatkowska
Bogumiła Okrojek

Which museum do you recommend visiting?

Where is the best restaurant in town?

Where is the nearest ATM located?

Where is a good place to go shopping?

Is there a zoo in the city?

What are some must-see landmarks here?

Can you suggest a nice park for a picnic?

Where can I find a good coffee shop nearby?

Are there any famous theaters or concert halls around?

Problem Statement

Our team has developed a model that aims to provide answers to a wide array of questions within this domain. This model is designed to facilitate the clustering of places based on several criteria, including their type, rating, and geographical location.

Our primary recipient

United Arab Emirates

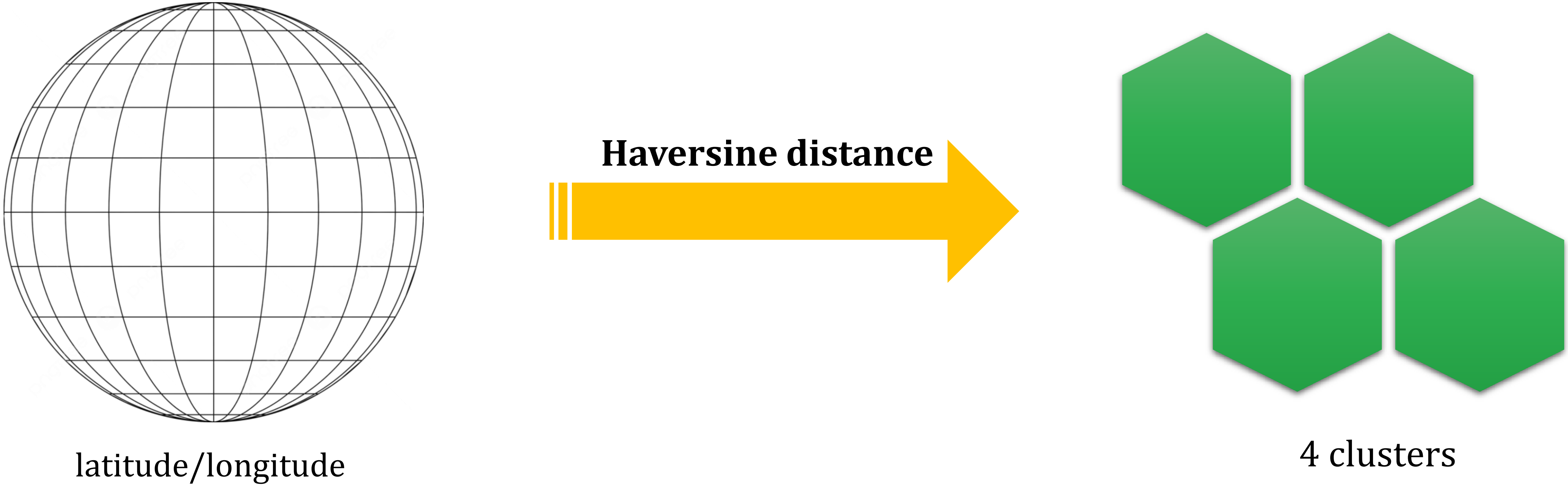
Dataset Overview

Google Maps

We worked with Google Places Comprehensive Business Dataset, which dataset has been scraped from Google Maps and presents extensive information about businesses across several countries. Each entry in the dataset provides detailed insights into business operations, location specifics, customer interactions and more.

What features were most important during the model construction?

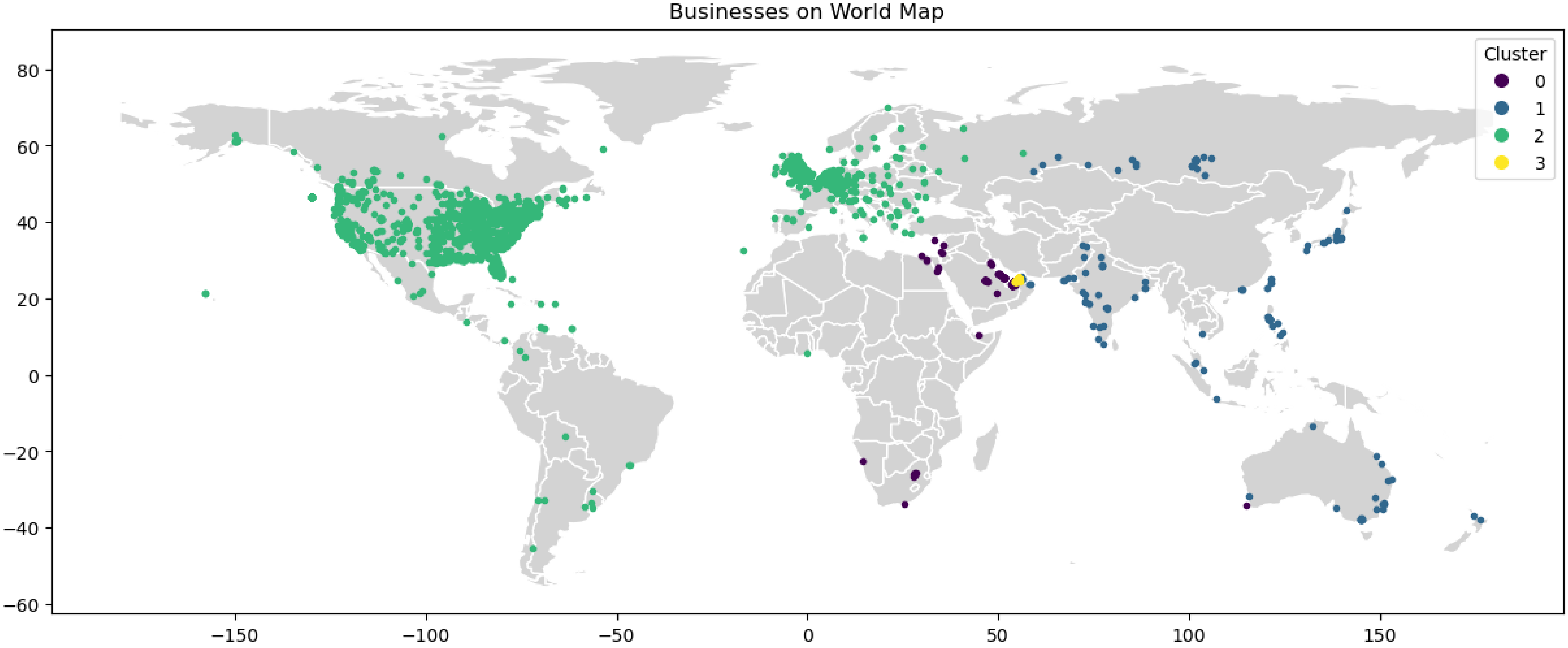
Location



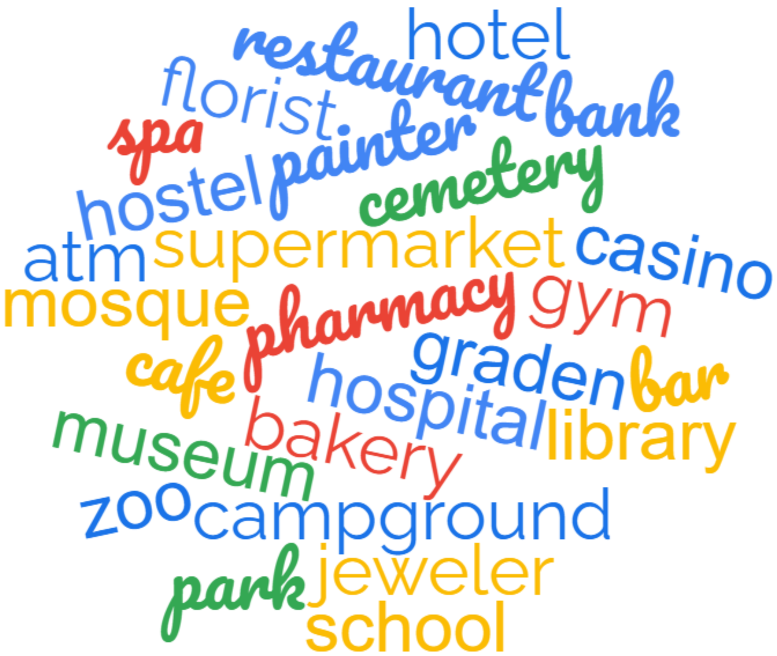
latitude/longitude

Haversine distance

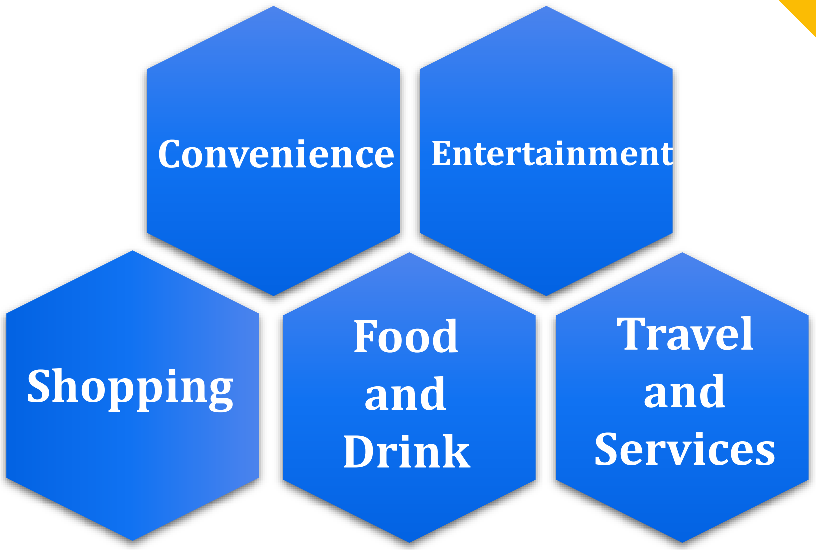
4 clusters


Businesses on World Map
Cluster
● 0
● 1
● 2
● 3
80
60
40
20
0
−20
−40
−60
−150
−100
−50
0
50
100
150

Types



TF-IDF Vectorizer



Convenience

Entertainment

Shopping

Food and Drink

Travel and Services

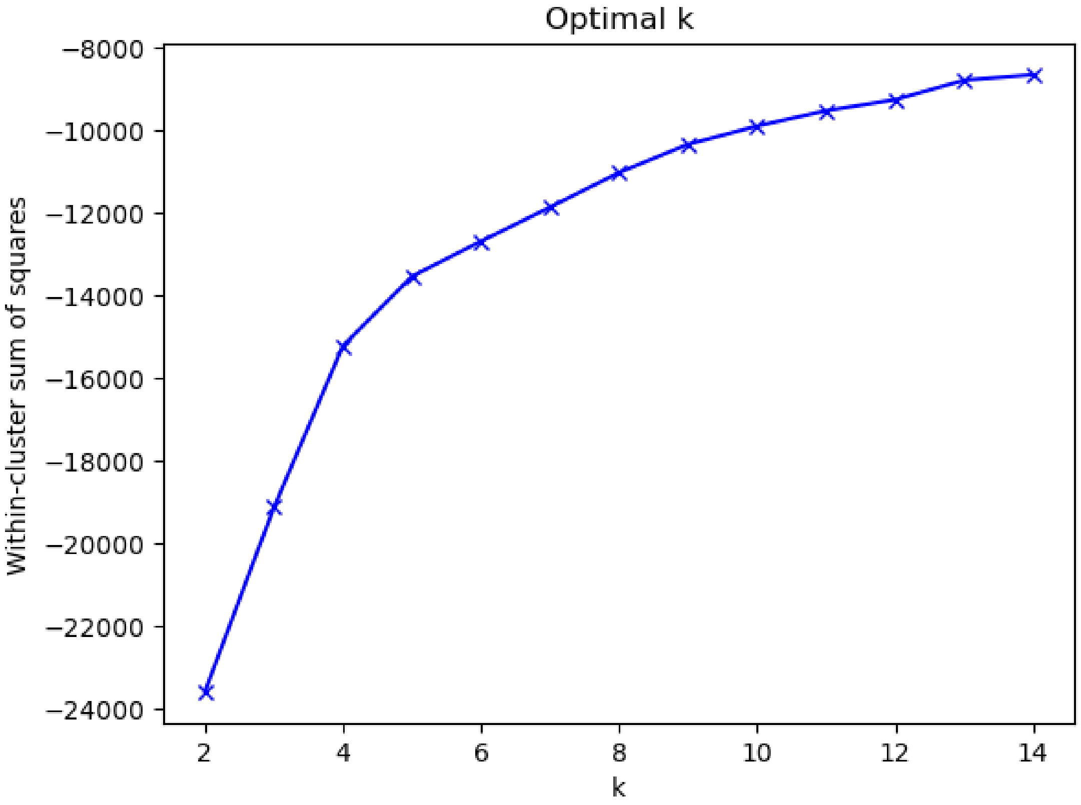
5 clusters

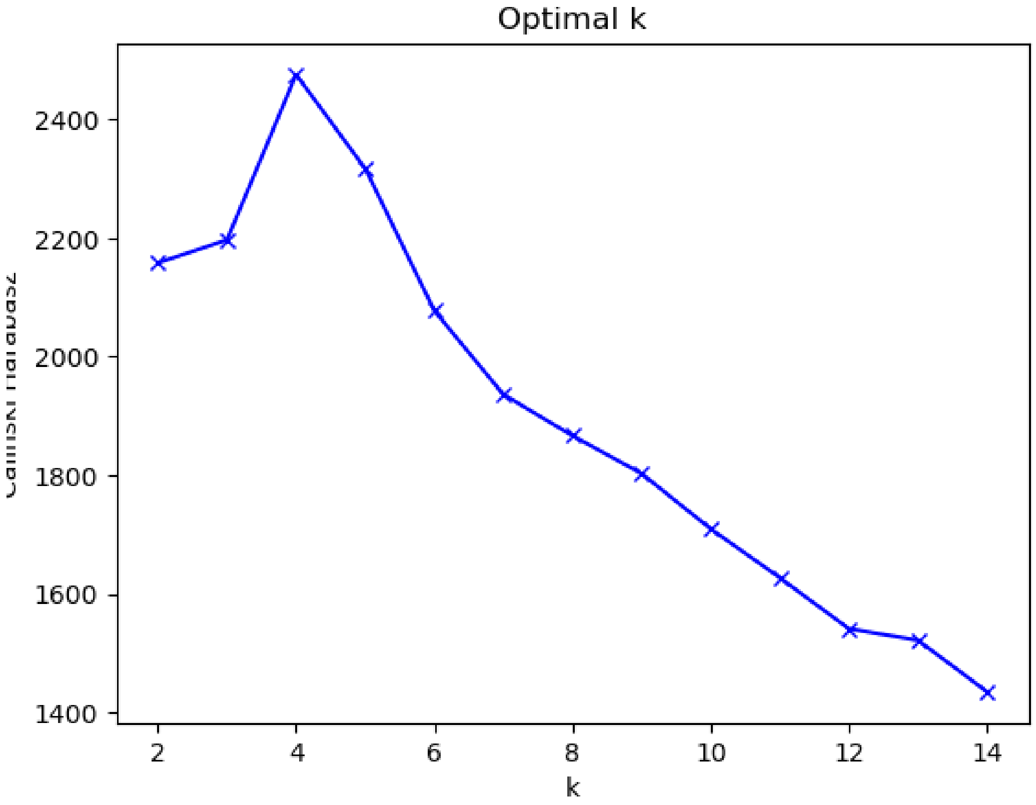
Rating

- Rating
- Number of review
- Having a website
- Being verified

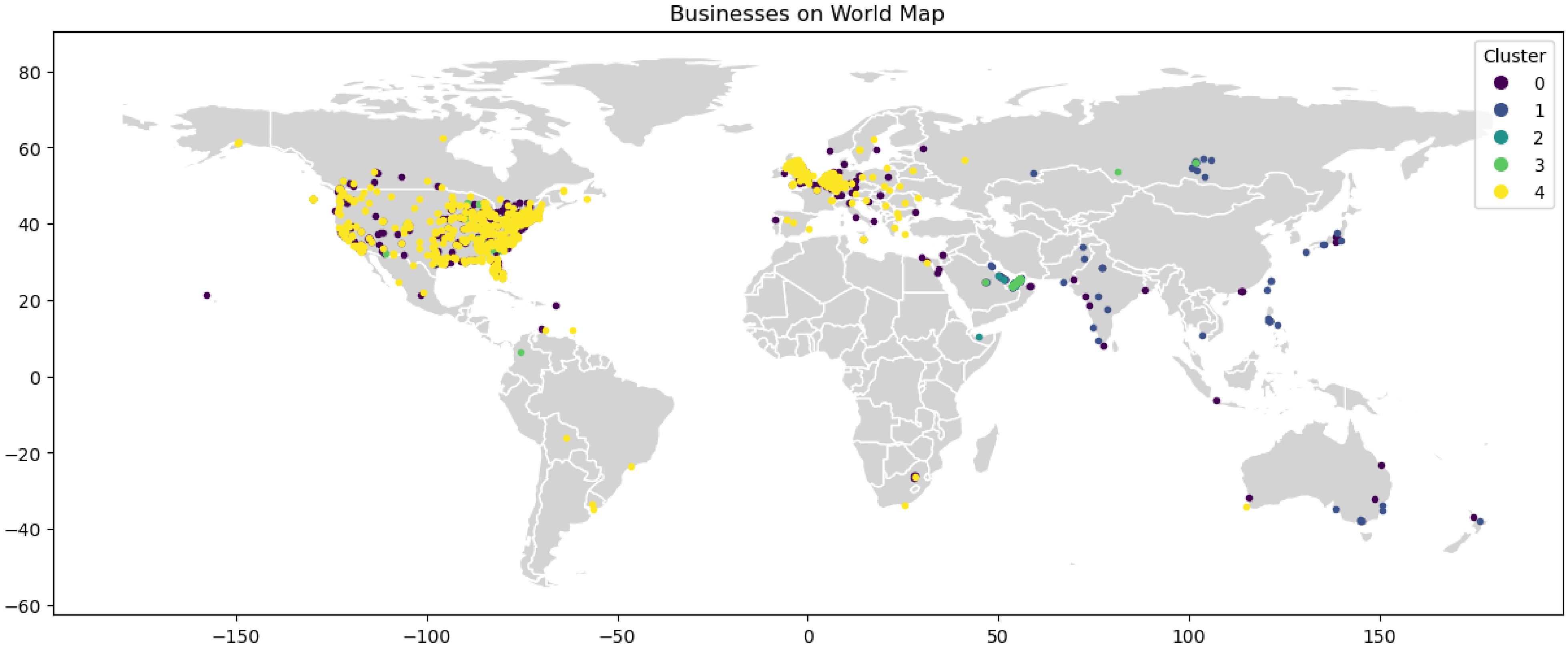
K-means

K-means

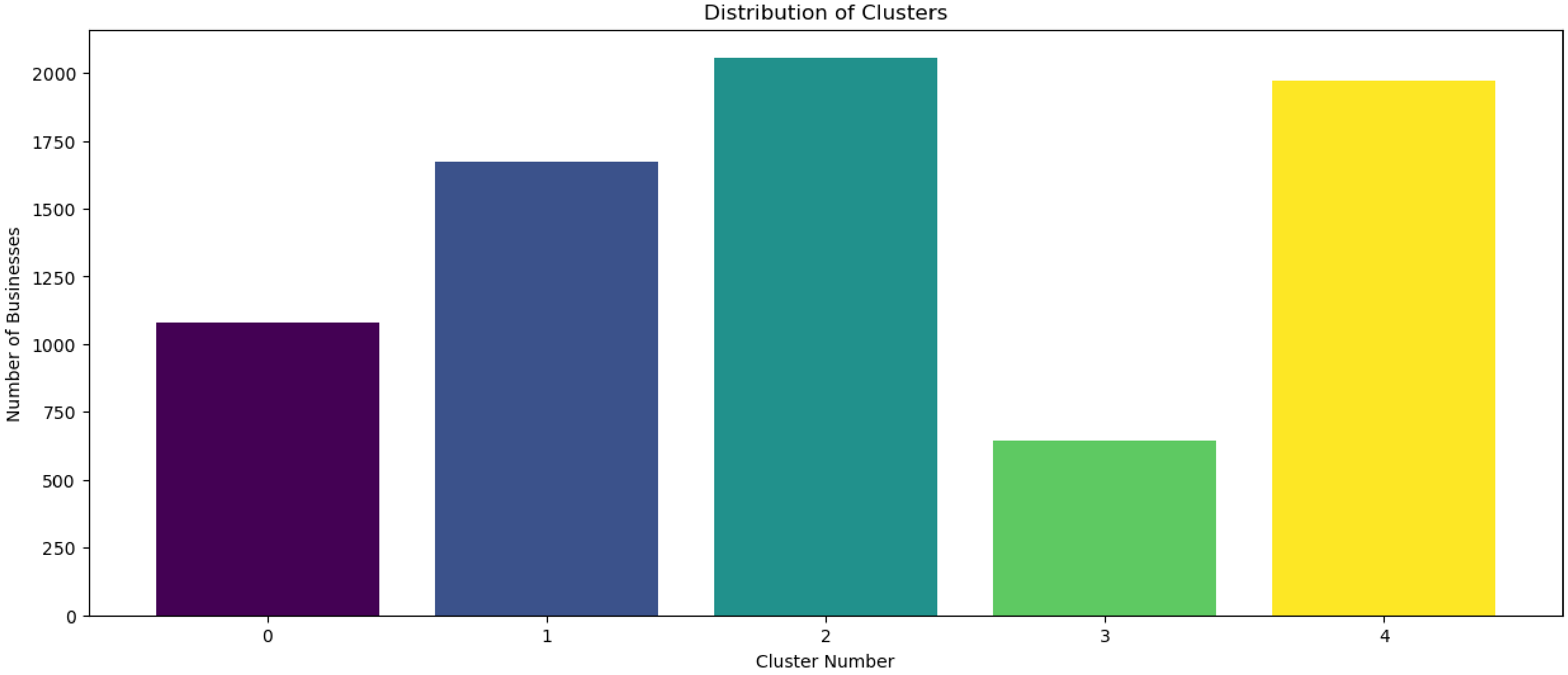

Optimal k
−8000
−10000
−12000
−14000
−16000
−18000
−20000
−22000
−24000
Within-cluster sum of squares
2
4
6
8
10
12
14
k


Optimal k
2400
2200
2000
1800
1600
1400
Calinski Harabasz
2
4
6
8
10
12
14
k

K-means


Businesses on World Map
Cluster
0
1
2
3
4
80
60
40
20
0
−20
−40
−60
−150
−100
−50
0
50
100
150

K-means


Distribution of Clusters
Number of Businesses
0
250
500
750
1000
1250
1500
1750
2000
0
1
2
3
4
Cluster Number

K-means

Train	
Score name	**Score value**
Silhouette Score	0.305531
Calinski-Harabaz Index	2769.311557
Davies-Bouldin Index	1.204398

Val	
Score name	**Score value**
Silhouette Score	0.299523
Calinski-Harabaz Index	1161.897084
Davies-Bouldin Index	1.237226

K-means

Feature	Importance
num__review_count	0.253232
num__rating	0.178003
cat__cluster_geo_1	0.160070
cat__cluster_geo_3	0.113183
cat__continent_Asia	0.104198

Features related to business evaluation have the greatest impact on the k-means model, followed by features associated with geographic location.

K-means

Cluster	num_review_count	num_rating
0	52	4.5
1	45	4.4
2	46	4.5
3	515	4.4
4	6	3

The table next to it presents the average values of the number of reviews and ratings in a given cluster.

As can be seen from the adjacent table, cluster 3 contains places with a large number of reviews, whereas cluster 4 consists of poorly rated places with a small number of reviews.

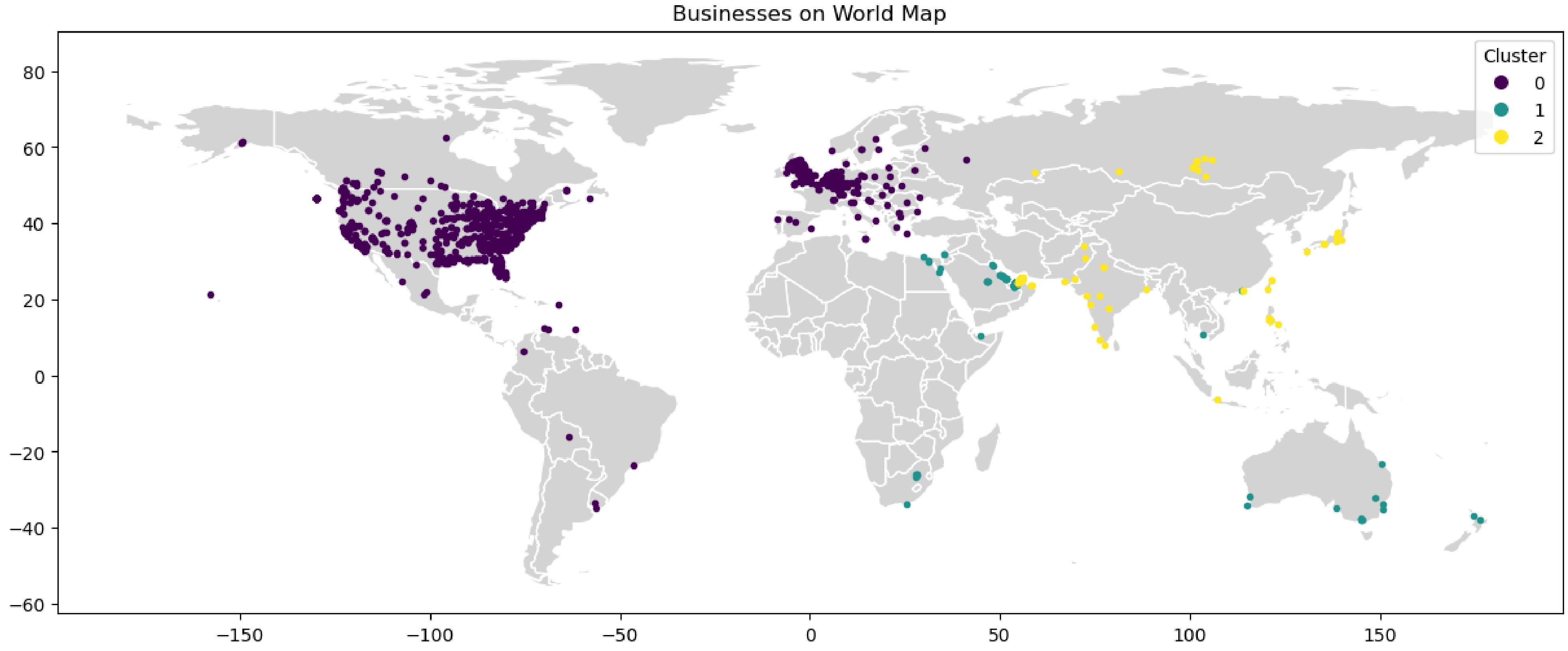
K-means

Cluster	cat__cluster_type_0	cat__cluster_type_1	cat__cluster_type_2	cat__cluster_type_3	cat__cluster_type_4
0	1%	10%	8%	**58%**	13%
1	23%	13%	32%	5%	26%
2	15%	**41%**	23%	7%	26%
3	15%	8%	2%	8%	9%
4	**46%**	29%	35%	22%	26%

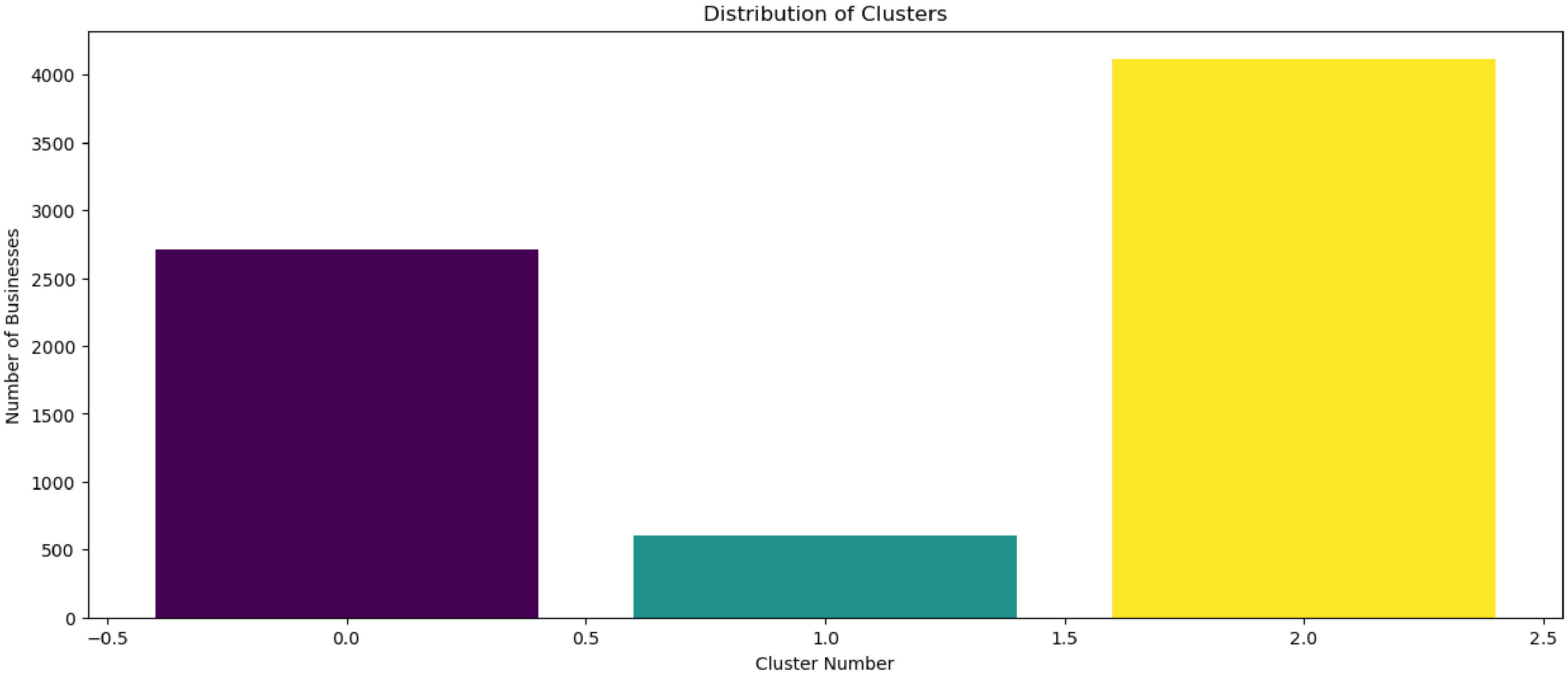
Cluster	Description
0	This group includes places mostly form North America and Europe, half of these places are tourist attractions. This places have a good rating and an average number of reviews.
1	This group includes locations situated in Asia. This places have a good rating and an average number of reviews.
2	This group includes places located in the Arabian Peninsula, most of this places belong to the gastronomic sector. This places have a good rating and an average number of reviews.
3	Places from this group are located in United Arab Emirates, they have high numer of good rating.
4	The group includes places mostly form North America and Europe, almost half of these places are conviences stores, this places are poorly rated and don’t have many reviews.

Gaussian Mixture Models

Gaussian Mixture Models



Gaussian Mixture Models



Gaussian Mixture Models

Train	
Score name	**Score value**
Silhouette Score	0.120660
Calinski-Harabaz Index	947.990349
Davies-Bouldin Index	4.000534

Val	
Score name	**Score value**
Silhouette Score	0.120867
Calinski-Harabaz Index	422.689053
Davies-Bouldin Index	4.000985

Gaussian Mixture Models

Feature	Importance
cat__continent_Asia	0.270495
cat__cluster_geo_2	0.231331
cat__continent_Europe	0.124585
cat__cluster_geo_0	0.084741
cat__cluster_geo_3	0.052553

The most important features for the Gaussian Mixture Models (GMM) were those related to geographic location, followed by features related to the type of place.

Gaussian Mixture Models

Cluster	cat__cluster_type_0	cat__cluster_type_1	cat__cluster_type_2	cat__cluster_type_3	cat__cluster_type_4
0	51%	36%	57%	48%	35%
1	**49%**	4%	**43%**	**51%**	4%
2	0%	**60%**	0%	1%	**61%**

Gaussian Mixture Models

Cluster	num_review_count	num_rating
0	38	4.3
1	378	4.3
2	63	4.3

The table next to it presents the average values of the number of reviews and ratings in a given cluster.

As can be noticed, the rating score did not influence our clustering, whereas the number of reviews did. Places in group 1 have the highest number of reviews.

Cluster	Description
0	The group includes locations situated in North America, South America, and Europe.
1	The group includes locations situated in Asia. In this group, many places are from the entertaiment, convenience and travel sector. Places in this group have the highest number of reviews.
2	The group includes locations situated in Africa and Australia. In this group, many places are from the shopping and gastronomic sector.

Thank you

Authors:

Aleksandra Kwiatkowska

Bogumiła Okrojek