THE BATTLE OF NEIGHBORHOODS

Analyzing the cities in Germany using data science methodologies

Lucas Friedrich Meincke

June 22, 2021

ABSTRACT

Heinz has a dream: to open an open-air shop where is served all sort of coffees and ice creams in Germany. But Heinz also has a problem: he loves Stuttgart, the city he lives. Regarding the shop, as Heinz's idea is to serve the clients outside, he needs to pay attention to some factors such as the weather in the city he will be moving to, but also his future concurrence there. Additionally, as Heinz loves Stuttgart, the main city of his state, he wants that the new city to be a big city too (according to the German's pattern of a big city, of course) and as similar as possible to his hometown, regarding the venues that are there. As I learned many things in the last months during my attendance in IBM Data Science Course, I will try to help my friend Heinz to realize his dream. I will use data science techniques to find the best place for him to open his Coffee and Ice Cream Shop.

CONTENTS

LIST OF FIGURES

1 INTRODUCTION

In this section, the purpose of this report is described. It is expected that the reader is already familiarized with basic python programming and data science concepts as the explanation about the fundamentals of those topics are not explained in this report.

1.1 Problem Description

When an entrepreneur wants to create a business such as a coffee shop or a restaurant, one of the most difficult things to do is to find the best place for it to be started considering its future concurrence and the characteristics of the city. For example: is it good for a business to be started downtown or at the city boundary? The climate conditions of the city could somehow influence my business? Is my concurrence doing a good job, or the customers are not being well-served? This report has the purpose to answer some questions based on these issues.

1.2 Problem Delimitation

My friend Heinz wants to become an entrepreneur and open an open-air shop in a city that is similar to his hometown Stuttgart, regarding his preferred places. Yes, he wants to move and live in a new place, but that brings him the "home feeling". Heinz is a "cult guy" who likes art, theater, and museums, but he also likes to enjoy nature and to go to parks and forests. During our conversation, he gave me all his preferences that now are listed in section 1.3 of this report. The city where Heinz will be moving to shall also has a good weather condition as he plans to serve his clients outside. Additionally, it shall be a good place for his business in terms of concurrence. By using data science techniques I will try to find the best place for my friend Heinz to start his business, based on the objectives described in section 1.3. As this report is part of the final essay of the IBM Data Science course, it does not have the intent to show the most recent data science methodologies neither the most efficient algorithms to solve the proposed problem but to expose using a practical approach, part of the theory learned during the course.

1.3 Objective

Find the best place for Heinz to open his coffee and ice cream shop in Germany using data scraped from the internet and Foursquare databases. The decision shall be based on what follows:

- It has to be a **big city** in Germany;
- The **precipitation rate** in the city should be low during the spring, summer, and autumn, then Heinz may attend his clients outside as many days as possible during this period of the year;

- The **concurrence** should be weak based on the **score** from customers' reviews (venues scores);
- The city has to be **similar to Stuttgart** based on the Heinz' preferred venues, which are **Clubs, Concerts, Theaters, Museums, Pubs, Gardens, Forests, Hills, Lakes, Parks, Nature Places, and Rivers**.

2 DATA

In this section, the data collection, preparation, and analysis are shown, along with images of the data frames created from them using Jupyter Notebook and Python programming language.

2.1 Data Sources

Two main data sources were used to collect all the data needed. They were Wikipedia website and Foursquare database, as shown below:

- **Biggest cities in Germany**: the data was scraped from Wikipedia [1];
- **The precipitation rate in each city in Germany**: the data was scraped from the main Wikipedia's page of each city, through a loop, for example, for the city of Berlin from [2], and for the city of Stuttgart from [3];
- **The venues in each city and their scores**: once having the latitude and longitude information for each city in Germany, it was possible to collect the information about their venues through the Foursquare API [4]. It was used two types of calls: regular calls for venues' names, IDs, and categories, and premium calls for the score and ratings (customers' feedback).

Many adjusts had to be made in the data for them to be properly used for further analysis. All the data preparation process is shown in section 2.3.

2.2 Data Usage

The first step was to collect a list that contains the **biggest cities in Germany** and also their latitude and longitude information. By knowing the name of the biggest cities, it was possible, though a loop function, to scrap the weather information containing the **precipitation rate** of each city from Wikipedia. At this point, it is possible to filter only a few cities which have the lowest precipitation rate and thus narrow the options for Heinz to open his open-air coffee and ice cream shop. For this report, it was chosen 12 cities.
Once the best cities were filtered based on the climate condition, the **venues available** in each of these cities could be collected (through a regular call) and also their **scores** (through a premium call) using the Foursquare database. At this point, the data containing the availability of the venues was spliced: one data frame containing Heinz' preferred venues in each city to compare them to Stuttgart, and a second data frame containing only the venues that

would be Heinz' concurrence (coffees and ice cream shops) and also their scores. By using both data frames it was possible to select the city which would be the best option for him to open his shop based on both: the customers' feedback on the concurrence, and by using the k-clustering algorithm to find the most similar cities to Stuttgart.

2.3 Data Collection and Preparation

The data from the biggest cities in Germany scraped from the Wikipedia website, after some adjustment, were transformed into a data frame as shown in figure 1a. Once having the names of the biggest cities in Germany it was possible, though a loop, to scrap from the Wikipedia pages of each city and their weather conditions. The data returned, after some adjust, were transformed into a data frame as shown in figure 1b.

	City	State	Population	Latitude	Longitude
0	Berlin	Berlin	3520031	52.517	13.383
1	Hamburg	Hamburg	1787408	53.550	10.000
2	München	Bavaria	1450381	48.133	11.567
3	Köln	North Rhine-Westphalia	1060582	50.933	6.950
4	Frankfurt am Main	Hesse	732688	50.117	8.683
...	...	...	...	...	...
75	Moers	North Rhine-Westphalia	104529	51.459	6.620
76	Siegen	North Rhine-Westphalia	102355	50.883	8.017
77	Hildesheim	Lower Saxony	101667	52.150	9.950
78	Salzgitter	Lower Saxony	101079	52.150	10.333
79	Kaiserslautern	Rhineland-Palatinate	99845	49.733	7.117

80 rows × 5 columns

(a) Biggest cities in Germany

	Measure	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Sum/Avg	Total	City
0	Max. Temp. (°C)	2.9	4.2	8.5	13.2	18.9	21.6	23.7	23.6	18.8	13.4	7.1	4.4	Ø	13.4	Berlin
1	Min. Temp. (°C)	-1.9	-1.5	1.3	4.2	9.0	12.3	14.3	14.1	10.6	6.4	2.2	-0.4	Ø	5.9	Berlin
2	Precipitation (mm)	42.3	33.3	40.5	37.1	53.8	68.7	55.5	58.2	45.1	37.3	43.6	55.3	Σ	570.7	Berlin
3	Max. Temp. (°C)	3.5	4.4	8.0	12.3	17.5	19.9	22.1	22.2	17.9	13.0	7.5	4.6	Ø	12.8	Hamburg
4	Min. Temp. (°C)	-1.4	-1.2	1.1	3.3	7.4	10.5	12.7	12.5	9.6	6.0	2.4	0.0	Ø	5.3	Hamburg
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
173	Min. Temp. (°C)	-1.5	-1.6	0.3	3.8	7.4	10.6	13.0	12.1	8.9	5.4	3.0	0.0	Ø	5.2	Siegen
174	Precipitation (mm)	107.0	66.0	63.0	41.0	67.0	75.0	84.0	83.0	58.0	78.0	96.0	112.0	Σ	930.0	Siegen
175	Max. Temp. (°C)	4.0	5.0	10.0	13.0	19.0	22.0	25.0	25.0	20.0	15.0	9.0	5.0	Ø	14.4	Kaiserslautern
176	Min. Temp. (°C)	-1.0	-2.0	2.0	3.0	8.0	12.0	14.0	13.0	9.0	6.0	3.0	1.0	Ø	5.7	Kaiserslautern
177	Precipitation (mm)	65.0	59.0	65.0	53.0	69.0	64.0	64.0	63.0	59.0	74.0	66.0	81.0	Σ	782.0	Kaiserslautern

178 rows × 16 columns

(b) Weather condition in each city

Figure 1: Dataframes from web scrapping

As only the precipitation rate and average temperatures during the spring, summer, and autumn (the seasons when the clients will be attended outside

the shop) are needed, the data frame containing the weather conditions was summarized by city and merged into the main data frame, containing the top 12 cities with less precipitation rate plus Stuttgart, as shown in figure 2.

	City	Precipitation (mm)	Summed Avg. Temp. (°C)	Population	Latitude	Longitude
0	Magdeburg	347.6	12.96	235723	52.13	11.62
1	Leipzig	357.0	13.93	560472	51.33	12.38
2	Nürnberg	371.0	14.13	509975	49.45	11.08
3	Mainz	377.0	13.61	209779	50.00	8.27
4	Pforzheim	377.6	13.91	122247	48.90	8.72
5	Erfurt	384.0	12.03	210118	50.98	11.03
6	Dresden	387.0	14.74	543825	51.03	13.73
7	Berlin	396.2	13.37	3520031	52.52	13.38
8	Heidelberg	401.0	15.21	156267	49.42	8.72
9	Mannheim	401.0	15.21	305780	49.48	8.47
10	Potsdam	401.0	12.96	167745	52.40	13.07
11	Bremen	402.0	13.49	557464	53.08	8.80
12	Stuttgart	434.0	15.13	623738	48.78	9.18

Figure 2: The main dataframe df_main

The latitude and longitude information from the main data frame was used to explore the venues of each city by using a regular call to the Fourquare database generating a new data frame, that after being filtered with Heinz' only preferred venues as shown in figure 3, will be used to compare Stuttgart to the other cities. The venues' ID information was also collected to be used through a premium call to the Fourquare database, and return the customers' feedback (venue score) for each venue that would be concurrent to Heinz (Café, Koffee Shops, Ice Cream Shops, and Bistros). The rating information which is the number of feedbacks was also collected just to verify if a sufficient amount of feedbacks were posted to compose each score. To keep the data as clean as possible those four categories of concurrent venues were recategorized into only two (Café and Ice Cream Shop) Finally, the data frame with all concurrent venues and their scores was created and is shown in figure 4.

	City Name	Venue ID	Venue Name	Venue Latitude	Venue Longitude	Venue Category
0	Magdeburg	4bd464ebcfa7b713b4ee23da	Kulturhistorisches Museum	52.125470	11.629250	History Museum
1	Magdeburg	4d15ec556c8b5481406ee1cc	Stadtpark Rotehorn	52.116431	11.641090	Park
2	Magdeburg	505b6d8be4b03e4e73c3a84d	Opernhaus Magdeburg	52.137391	11.638469	Concert Hall
3	Magdeburg	53938089498e97c3af3dcb88	Schweizer Milchkuranstalt	52.122950	11.634850	Beer Garden
4	Magdeburg	4b98e02bf964a520a55335e3	Theater Magdeburg	52.137365	11.639108	Theater
...	...	...	...	...	...	...
174	Stuttgart	4b1d7ff5f964a520591124e3	Mercedes-Benz Museum	48.788169	9.234138	Museum
175	Stuttgart	4c8399d9dc018cfae1bed96c	Santiago-de-Chile-Platz	48.755496	9.172667	Park
176	Stuttgart	4bd08530b221c9b630d6d3d0	Biergarten Schlossgarten	48.784496	9.186104	Beer Garden
177	Stuttgart	532eb904498e2b589ca0a7a8	Gewächshäuser | Wilhelma	48.804081	9.207995	Botanical Garden
178	Stuttgart	5518fd68498e2008cf2021c6	Im Wizemann (Halle)	48.808133	9.201321	Concert Hall

179 rows × 6 columns

Figure 3: Dataframe df_venues_pref containing Heinz's prefered venues in each city

	City Name	Venue ID	Venue Name	Venue Latitude	Venue Longitude	Venue Category	Score	Rating
0	Magdeburg	561a856e498ee6fc8b49cafb	Herzstück - Das Kuchenatelier	52.133728	11.615929	Café	8.3	9.0
1	Magdeburg	4dc53e2345dd2645526a7f15	Eis-Konditorei Bortscheller	52.106756	11.640792	Ice Cream Shop	8.4	17.0
2	Magdeburg	4cb5ddce1b0af04d8a64cc25	Il Capitello	52.126257	11.634307	Café	7.5	14.0
3	Leipzig	51e3cd828bbde0ebc2f0d32f	Handbrotzeit	51.341338	12.378071	Café	8.5	80.0
4	Leipzig	548137ca498eb80c2a45fc33	Espresso Zack Zack	51.332784	12.404315	Café	8.7	45.0
...	...	...	...	...	...	...	...	...
149	Stuttgart	53e89bfe498e2fc08ba33412	Kantinchen	48.763924	9.176687	Café	8.5	9.0
150	Stuttgart	4ea478f47beb98c09bcc1547	Starbucks	48.776407	9.175892	Café	7.9	280.0
151	Stuttgart	563e1dd8cd1028d37ef044b1	Schwarzmahler	48.785164	9.209637	Café	8.5	13.0
152	Stuttgart	57bb5dfb498e86d2c115234b	misch misch	48.765881	9.169148	Café	8.1	50.0
153	Stuttgart	5ad8762a8c812a57d7ca4663	Taraba	48.792606	9.201570	Café	8.2	34.0

154 rows × 8 columns

Figure 4: Dataframe df_venues_conc containing the concurrent venues

A summary of the data frames that will be used from now on is shown below:

- **df_main**: data frame containing the top 12 big cities with the lowest precipitation rate (not counting the winter) plus Stuttgart, their latitudes and longitudes, but also their summed average temperature and population (these last 2 only for consult);
- **df_venues_pref**: data frame containing Heinz' preferred venues in each of the 12 cities plus Stuttgart so that they can be clustered;
- **df_venues_conc**: data frame containing the venues in each of the 12 cities that would be the concurrence for Heinz (Cafés and Ice Cream Shops) and their scores.

3 METHODOLOGY

In this section, data science techniques and tools such as one-hot encoding, k-mean clustering, and folium map for data visualization are used on the data collected to give Heinz the best help possible on his decision. The idea is to start clustering the cities regarding their venues to find those which are most similar to Stuttgart. After narrowing the list of cities only with those in the same cluster as Stuttgart, it was possible to select the one which has the concurrence with the lowest average score. The procedure and data manipulation are shown in the following topics.

3.1 One Hot Encoding

One hot encoding is a method used to convert categorical values into numbers so that is possible to use machine learning algorithms on them. The method was used in the **df_venues_pref** data frame shown in figure 3, resulting in the data frame shown in figure 5.

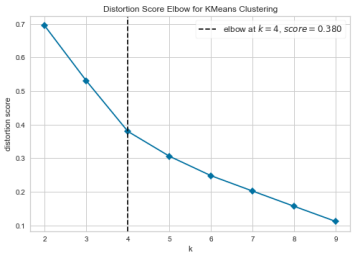
	City Name	Art Museum	Beer Garden	Botanical Garden	Comedy Club	Concert Hall	Forest	Garden	Garden Center	History Museum	...	Lake	Movie Theater	Museum	Nature Preserve	Park
0	Berlin	0.100000	0.000000	0.0000	0.000000	0.200000	0.000000	0.150000	0.000000	0.050000	...	0.000000	0.000000	0.000000	0.000000	0.400000
1	Bremen	0.062500	0.062500	0.0000	0.000000	0.062500	0.000000	0.000000	0.000000	0.000000	...	0.125000	0.000000	0.000000	0.000000	0.187500
2	Dresden	0.058824	0.058824	0.0000	0.000000	0.000000	0.000000	0.058824	0.000000	0.000000	...	0.000000	0.058824	0.117647	0.000000	0.235294
3	Erfurt	0.000000	0.000000	0.0000	0.000000	0.000000	0.000000	0.000000	0.142857	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.714286
4	Heidelberg	0.142857	0.000000	0.0000	0.000000	0.000000	0.000000	0.142857	0.142857	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.142857
5	Leipzig	0.045455	0.045455	0.0000	0.000000	0.090909	0.000000	0.000000	0.000000	0.045455	...	0.000000	0.000000	0.090909	0.045455	0.272727
6	Magdeburg	0.066667	0.066667	0.0000	0.000000	0.066667	0.000000	0.000000	0.000000	0.133333	...	0.000000	0.000000	0.000000	0.000000	0.333333
7	Mainz	0.000000	0.000000	0.0000	0.000000	0.090909	0.090909	0.000000	0.090909	0.090909	...	0.000000	0.090909	0.000000	0.000000	0.363636
8	Mannheim	0.111111	0.000000	0.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.666667
9	Nürnberg	0.000000	0.105263	0.0000	0.052632	0.052632	0.000000	0.000000	0.000000	0.157895	...	0.052632	0.000000	0.000000	0.000000	0.473684
10	Pforzheim	0.250000	0.000000	0.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.250000
11	Potsdam	0.062500	0.000000	0.1250	0.000000	0.000000	0.000000	0.000000	0.000000	0.062500	...	0.000000	0.000000	0.062500	0.062500	0.375000
12	Stuttgart	0.000000	0.062500	0.0625	0.000000	0.125000	0.000000	0.000000	0.000000	0.062500	...	0.000000	0.000000	0.062500	0.000000	0.500000

13 rows × 24 columns

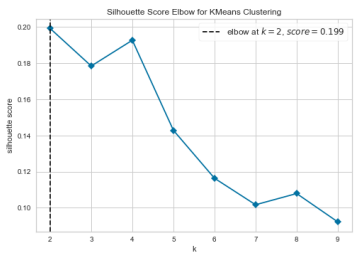
Figure 5: One hot enconding

3.2 Finding the Best k Value

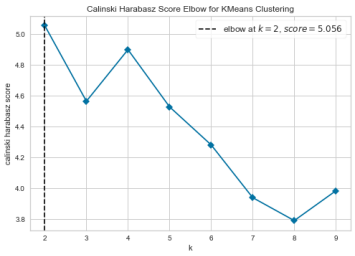
K-means is the most popular algorithm used in unsupervised Machine Learning problems. The main question when using the k-mean algorithm is about the value of the k, which is the number of clusters that a data frame is grouped into. Therefore, it is necessary to find the best k value before using the k-mean clustering model. There are different algorithms to do so, but in this report, four of the most known methods were tried: the elbow curve, the silhouette method, the Calinski-Harabasz criteria, and the dendrogram diagram. All of them were implemented with the help of the Yellowbrick library [5]. It is possible to see in figure 6a that the "elbow" pattern was formed right on the value 4. With the silhouette method shown in figure 6b, the k value given was 2, but the value 4 has also a high score. The same occurs when using the Calinski-Harabasz criteria as shown in figure 6c. The dendrogram diagram shown in figure 6d highlights 3 distinct clusters. You may find more information about those four methods here [6].



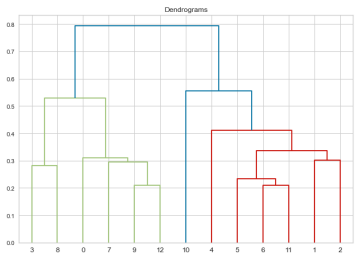
(a) The elbow curve



(b) The silhouette method



(c) The Calinski-Harabasz criteria



(d) Dendograms diagrams

Figure 6: Methods to find the K value

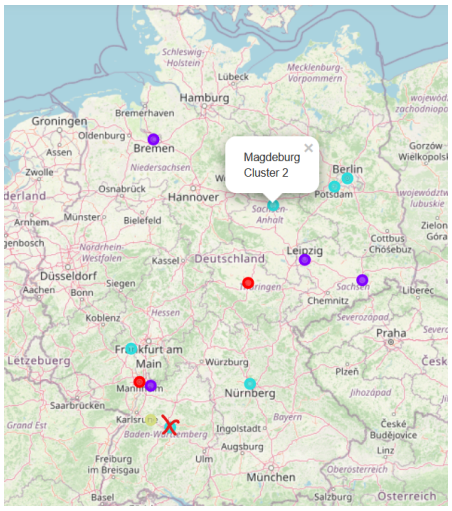
It was assumed 4 for the k value as it was the value returned by the elbow curve method and the second value with the highest returned score by the silhouette method and by the Calinski-Harabasz criteria. It is also not far from the number of clusters returned by the dendrogram diagram.

3.3 Clustering the Cities

Once having 4 as the k value, the cities were clustered using the k-means algorithm, and a map was plotted using the Folium library [7]. It is possible to see in figure 7 that six cities are similar to Stuttgart (marked with a red X on the map) regarding Heinz's preferred venues and are listed in cluster 2. They are: **Magdeburg, Nürnberg, Mainz, Berlin, and Potsdam**.

	City	Latitude	Longitude	Cluster
0	Magdeburg	52.13	11.62	2
1	Leipzig	51.33	12.38	1
2	Nürnberg	49.45	11.08	2
3	Mainz	50.00	8.27	2
4	Pforzheim	48.90	8.72	3
5	Erfurt	50.98	11.03	0
6	Dresden	51.03	13.73	1
7	Berlin	52.52	13.38	2
8	Heidelberg	49.42	8.72	1
9	Mannheim	49.48	8.47	0
10	Potsdam	52.40	13.07	2
11	Bremen	53.08	8.80	1
12	Stuttgart	48.78	9.18	2

(a) Dataframe containing the cluter of each city



(b) Folium map with the clustered cities

Figure 7: The four clusters

3.4 Concurrence and their Scores

The venues in each city listed in the data frame **df_venues_conc** shown in figure 4 were grouped and with that the mean of the scores given by the customers to the local coffee and ice cream shops could be verified. The cluster information was merged, resulting in a final data frame that was narrowed filtering only the cities labeled with cluster 2. The resulting data frame is shown in figure 8. Now we know the cities that are most similar to Stuttgart regarding Heinz's preferences and also the mean of the scores given to Heinz's concurrence in each city.

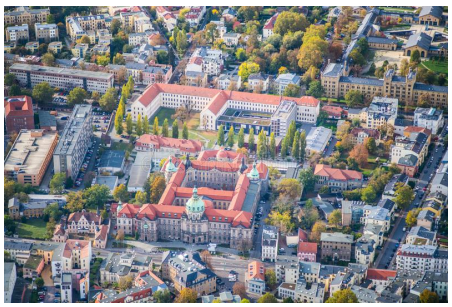
	City	Cluster	Score
0	Potsdam	2	7.91
1	Magdeburg	2	8.07
2	Mainz	2	8.18
3	Stuttgart	2	8.46
4	Nürnberg	2	8.52
5	Berlin	2	9.03

Figure 8: The cities in the cluster 2 and their scores

4 RESULTS AND DISCUSSION

By the analysis done so far, the city of Potsdam could be declared as a good city for Heinz to start his business as it belongs to the same cluster as Stuttgart, and has a concurrence with a low score according to the customers' feedback. In figure 9a a map with the concurrent venues in Potsdam is shown, where the black points refer to coffee shops and the blue points refer to ice cream shops. The map helps as a good start point to analyze in which part of the city would be the best place to open the shop, taking into account, for example, the place with less concurrence in terms of quantity, or the place which has the lowest rent price.

(a) Folium map of Potsdam showing the coffee and ice cream shops



(b) Overview of Potsdam

Figure 9: The chosen city

Other cities belong to the same cluster as Stuttgart, that is, have a similar distribution of venues that Heinz likes to frequent. Furthermore, other cities have good weather conditions, similar to Potsdam. Even Potsdam being the best city for Heinz to open his business taking into account the objectives stipulated in section 1.3, the final decision will be taken by him. To help him a little bit more, a final data frame was created as shown in figure 10 summarizing all the important information collected so far. The last five columns were added, containing the five most recurrent venues in each city of cluster 2.

	City	Cluster	Score	Precipitation (mm)	Summed Avg. Temp. (°C)	Population	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Potsdam	2	7.91	401.0	12.96	167745	Park	Botanical Garden	Pub	Art Museum	History Museum
1	Magdeburg	2	8.07	347.6	12.96	235723	Park	Theater	History Museum	Pub	Beer Garden
2	Mainz	2	8.18	377.0	13.61	209779	Park	Theater	Movie Theater	Sculpture Garden	Concert Hall
3	Stuttgart	2	8.46	434.0	15.13	623738	Park	Concert Hall	Botanical Garden	Pub	History Museum
4	Nürnberg	2	8.52	371.0	14.13	509975	Park	History Museum	Beer Garden	Irish Pub	Lake
5	Berlin	2	9.03	396.2	13.37	3520031	Park	Concert Hall	Garden	Art Museum	Science Museum

Figure 10: The final summarized results

5 CONCLUSION

The purpose of this report is to show how it is possible to use data science methodologies and tools to help us in the analysis of the most diverse problems. In the problem exposed in this report, it was stipulated four objectives to be reached to help Heinz with his issue of opening an open-air coffee and ice cream shop in a big city in Germany similar to Stuttgart and that has good weather. By choosing Potsdam as the new city, all the objectives proposed were satisfied.
Further analysis could be done using this report as an initial point, such as discovering the best neighborhood in Potsdam, with the lowest rental cost, or the neighborhood which has the most similarity with the one Heinz lives in Stuttgart.
This report, as part of the final essay of the IBM Data Science course, had the intent of exposing some of the data science methodologies and tools practically.

REFERENCES

[1] *List of Cities in Germany*. URL: https://en.wikipedia.org/wiki/List_of_cities_in_Germany_by_population. [Last accessed on 15 May 2021]
[2] *Main page of Berlin*. URL: https://de.wikipedia.org/wiki/Berlin. [Last accessed on 15 May 2021]
[3] *Main page of Stuttgart*. URL: https://de.wikipedia.org/wiki/Stuttgart. [Last accessed on 15 May 2021]
[4] *Foursquare Library*. URL: https://foursquare.com/. [Last accessed on 15 May 2021]
[5] *Yellowbrick Library*. URL: https://www.scikit-yb.org/en/latest/. [Last accessed on 15 May 2021]
[6] Indraneel Dutta Baruah. *Cheat sheet for implementing 7 methods for selecting the optimal number of clusters in Python*, 2020. URL: https://towardsdatascience.com/cheat-sheet-to-implementing-7-methods-for-selecting-optimal-number-of-clusters-in-python-898241e1d6ad. [Last accessed on 15 May 2021]
[7] *Folium Library*. URL: https://python-visualization.github.io/folium/. [Last accessed on 15 May 2021]