

Finding the optimal location in the US for an Aviation Consultant based in Stuttgart, Germany

Alina Henne, November 2nd 2020

1. Introduction

1.1 Background

A fictional aviation consulting company with headquarters in Stuttgart, Germany wants to expand to the US. They have been consulting European aviation businesses successfully for many years and now want to enhance American partnership and move closer to their clients overseas.

The company's culture is largely based on employee satisfaction, so in order to achieve the best possible working atmosphere for its employees and to synchronize the surroundings of the new company location with the surroundings of the headquarters (that have proved to be popular), a key criterium for choosing the location of their new offices in the US is that the surrounding venues should be as similar as possible.

1.2 Problem

An analysis of surrounding venues in Stuttgart compared to venues in different US cities with airports (for example Atlanta, Los Angeles, Chicago, New York and Seattle) is supposed to give insight into the question: based on similarity of surrounding venues, which US city with an airport as busy as possible is best for expanding the Stuttgart-based consulting firm?

2. Data

For the analysis I will be using Foursquare location data for Stuttgart and US cities, including columns:

- Venue type
- Venue longitude/latitude
- city longitude/latitude

I will be comparing the venue types of the ten most common venues across the different cities and cluster them in order to find the cluster that includes Stuttgart (thus the choices of similar cities). To pinpoint the best choice, the city with the busiest airport of the cluster including Stuttgart will be chosen, as a busy airport indicates a good economy and many potential customers.

In order to find the list of US cities that shall be included in the analysis, I will also have a data set on US international airports and in what city they are. In order to not over-complicate this, I will only take the top 40 US airports (in terms of passenger numbers). The data set will be scraped from: <https://www.world-airport-codes.com/us-top-40-airports.html>. Since the ranks of airport business rarely change and the pandemic has potentially skewed data, it will be legitimate to use the ranks as of 2016.

3. Methodology

3.1 Data preparation

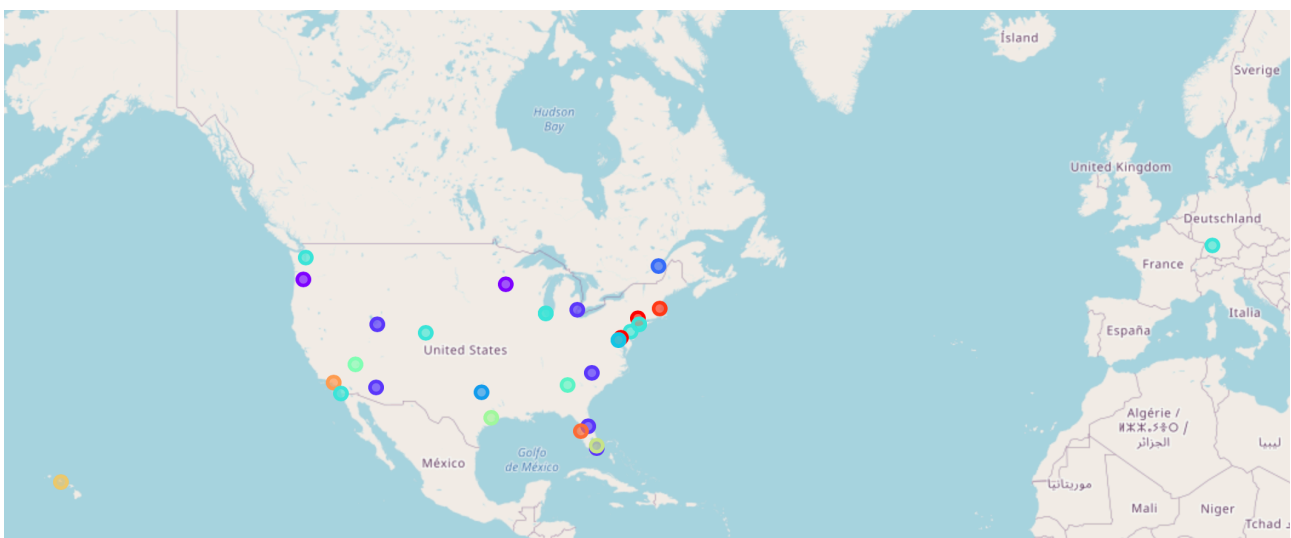
The list of airports was first of all scraped from the above mentioned source. Irrelevant columns were deleted, so only the airport's city and state remained. The coordinates for each city of the 40 busiest US airports were added as well as another row for Stuttgart, including coordinates.

After defining a venue search function, it was used on all cities in the initial list. For all 41 cities, the venue types were collected and matched. With one hot encoding, grouping and sorting methods, the final data frame for cluster analysis was found. The data frame now showed the 41 cities and the ten most common venue types for each city.

3.2 Clustering

Via k-means clustering, the cities were put into 15 clusters. This amount of clusters was chosen after an exploratory analysis, in which also 5, 10 and 20 clusters were created. For 5 and 10 clusters, the cluster set was too large, so it would have not been specific on which city according to its venues would actually be similar to the venues in Stuttgart, whereas 20 clusters was already too specific, as Stuttgart was the only one in the cluster. Hence, 15 clusters was the best choice.

	City	Latitude	Longitude	Cluster	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Atlanta, GA	33.749099	-84.390185	7	Intersection	Breakfast Spot	College Basketball Court	Food Court	Music Venue	Boat or Ferry	Caribbean Restaurant	Tapas Restaurant	Event Space	Art Gallery
1	Los Angeles, CA	34.053691	-118.242767	12	Coffee Shop	Museum	Breakfast Spot	Shopping Mall	Sushi Restaurant	Supermarket	Gift Shop	Speakeasy	Bookstore	Men's Store
2	Chicago, IL	41.875562	-87.624421	6	Coffee Shop	Sandwich Place	Hotel	Pub	Boutique	Pizza Place	Bookstore	Museum	Concert Hall	Garden
3	Dallas, TX	32.776272	-96.796856	4	Hotel	Coffee Shop	Café	Plaza	Bistro	IT Services	Furniture / Home Store	Mobile Phone Shop	Food Truck	Department Store
4	Denver, CO	39.739236	-104.984862	6	Sandwich Place	Mexican Restaurant	Coffee Shop	Yoga Studio	Café	Pub	Pizza Place	Noodle House	Exhibit	Convenience Store



There were another 8 cities in Stuttgart's cluster.

4. Results

In order to be able to give a specific response to the question: „which US city is best for expanding?“, another factor had to be regarded. As busy airports indicate a strong economy as well as many potential customers, the specific choice of the 8 given cities was refined by finding the city with the busiest airport.

This was done by simply taking the first result, as the cities had already been ordered by busiest to less busy airports. Hence, the choice for the best city is Chicago, IL.

5. Discussion

The reason to use data on venue types of a city for a choice of where to expand to is simply that of employee motivation: if an employee of the Stuttgart company finds the same type of food and other venues in the city of his or her new office, he or she may feel more at home. While employee motivation is key, trying to copy the feeling of home may not be the most productive approach.

In pandemic times, we have all learned the relevancy of home office, virtual meetings and online alternatives. If this consultancy was real, I would suggest an approach where you gather data on potential customers and compare them to your portfolio. In other words, clustering potential customers and choosing the cluster that is the most similar to the successfully handled past customers would be the best choice.

However, I do not have any data on potential customers and to create fictional data would be too big an effort for this purpose alone. Hence, the choice of a new hub was limited to venue data, and within that framework, the analysis was successfully executed and the result satisfactory.

6. Conclusion

In this study, I analyzed the similarity of venue types taking a city in Germany and the cities of the busiest US airports. The fit of the choice was discussed in section 5, the approach given in section 3. To not repeat myself, I will only give a short statement on the lessons learned that I achieved within this project and list the skills I was able to put to use.

First of all, I was able of course to use my basic skills like grouping, filtering and altering data frames. Handling data with pandas was essential and this project helped me deepen those basic skills. I have in this course learned to scrape data from dynamic sources, which I have done with an html file in this project. Also, I have used my gained knowledge on how to simply find coordinates and use Foursquare data, namely for finding surrounding venues. Another essential skill used was the method for (k-Means) clustering.

Lastly, the story of finding the best solution as well as a discussion and reflection was given, which also shows the storytelling skill that was used and that is vital for data scientists.

Thank you for reading this paper. I am grateful for being able to finish this project successfully.