# College and Pub Distribution in Bratislava and Wien

# (Comparative Study)

# Annamária Zagyi Takáts

# 4/30/2021

# Introduction

This report is part of the Capstone Project for IBM’s Applied Data Science Professional Certificate offered by Coursera. This is part of the final course in this 9-course series, week 5 assignment.

# Background

Bratislava (Slovakia) and Wien (Austria) are the two nearest capital cities in Europe, distancing from each other just around 70 kms. Both capitals are university cities with lot of colleges and of course pubs and bars are very often crowded especially in college surroundings.

The aim of this project is to explore the correlation and distances between these two objects and compare them based on city levels.

# Business problem

The objective of this Capstone project is to analyze and select the best location (in terms of city choice and average distances to the college) to open a pub.

This project helps to estimate the best localization to open a pub/bar and will give a good hint for a potential investor to decide whether to open a new pub in Wien or Bratislava and if it worth to open it close to a college. Potentially the owner of a pub can have great success and consistent profit. Opening a pub requires serious considerations and is more complicated than it seems from the first glance. In particular, the location of the restaurant is often the most important factor that will affect whether it will have success or a failure. This study can be used for a business plan afterwards.

On the other hand the final project explores the best choice for being a student taking into consideration the pub and college statistics in Bratislava and Wien. This report will be useful and handy for the students and lectors as well.

This material will attempt to answer these 2 questions :

“Where should the investor open a pub”

“How far should I go to the nearest pub from my college if I am a student”.

# Data

Using data science methodology and instruments as working with different Python libraries, working with API modules, Data Cleaning, Data Analyses, Visualization, this project aims to provide solutions to answer the business question: Whether in Wien or Bratislava and where in the cities, should the investor open a pub.

All data related to locations of venues are obtained via the FourSquare API utilized via the Request library in Python.

We need data about different venues in 2 different cities and states. In order to gain that information we use “Foursquare” API. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos. As such, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through their API.

We use Geopy library to receive location data for particular areas. Geopy is a Python client for several popular geocoding web services. Geopy makes it easy for Python developers to locate the coordinates of addresses, cities, countries, and landmarks across the globe using third-party geocoders and other data sources.

Geolocator from Python Geocoder package is used to gain geographical coordinates of Wien and Bratislava. This is required to plot the map and get the venue data.

Then we connect to the Foursquare API to gather information about venues inside each city. It provides many categories of the venue data, we were particularly interested in the pub and college data to solve the business problem defined above.

We limit the number of results returned for our Foursquare queries to 100 per area. For Wien we limit the radius to be 2 000 meters and for Bratislava 10 000 meters. This is because we want to achieve 2 datasets of approximately the same size. With these parameters we gather the following data sets:

* 32 university venues and 100 pub venues in Bratislava
* 68 university venues and 73 pub venues in Wien.

After normalizing the json files and some data cleaning we obtain 2 datasets for each city, containing college names with their geographical coordinates and pub names with their geographical coordinates respectively. This information is processed into Python dataframes. We are going to use this data to perform further analysis.

We apply Folium maps for visualizing the venues coordinates.

Based on definition of our problem, factors that will influence our decision are:

* Density of pubs and bars in college areas
* Minimum, maximum and average distances from each college to the nearest, utmost pub/bar

Geodesic and Math libraries are used for these calculations and the data is be visually assigned using graphing from Python libraries. Data results are plotted in Bar and Box plots to visualize statistical summaries of the obtained results.

# Methodology

When building up our dataset, we obtain the latitude and longitude values for Bratislava and Wien using Geolocator. The output is:

The geographical coordinate of Bratislava are 48.1516988, 17.1093063.

The geographical coordinate of Wien are 48.2083537, 16.3725042.

With these values we make a Foursquare query for University venues in 10 000 meters radius of Bratislava with the result limit set to 100. And another Foursquare query for Pub venues in 10 000 meters radius of Bratislava with the result limit set to 100.

We repeat the same Foursquare queries for Wien with radius of 2 000.

The obtained Json files are normalized, cleansed and structured into Pandas Dataframes with the following results:

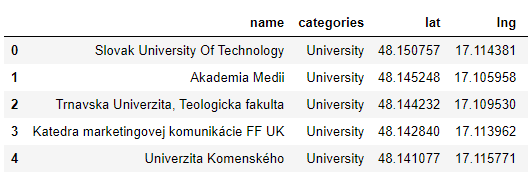
32 university venues in Bratislava were returned by Foursquare.

100 pub venues in Bratislava were returned by Foursquare.

68 university venues in Wien were returned by Foursquare.

73 pub venues in Wien were returned by Foursquare.

Sample data for Bratislava:



Table

Description automatically generated

Visualizing them in the following manner: green dots stand for Colleges and yellow ones for Pubs in Slovakia:

Map

Description automatically generated

The same for Wien with black (College) and red (Pubs) dots :

Chart

Description automatically generated

We create the final datasets for Bratislava by combining the Pubs and Colleges datasets into one data frame by merging them over a fictive key. After calculating all possible distances via Geodesic and grouping by Universities, we get these results:

* Universities with corresponding shortest distance to a Pub:

Graphical user interface, text

Description automatically generated

* Universities with corresponding longest distance to a Pub



* Universities with corresponding mean distance to a Pub

Graphical user interface, text

Description automatically generated

We repeat the same process for Wien as well. Detailed datasets for both Bratislava and Wien can be found in the attached Jupyter notebook.

# Results

According to the base datasets, we can see that in Bratislava we have 100 pubs serving 32 universities while in Wien the same ratio is just 73/68. It means a better opportunity for investing into pub or bar opening in Wien than in Bratislava.

Based on the final 6 datasets we produce the following results on College – Pub distances:

* Statistical distribution of minimum distances in Bratislava looks like:

Table

Description automatically generated

* Statistical distribution of minimum distances in Wien looks like:

Table

Description automatically generated

This gives us an insight that in Bratislava there are longer distances between the studied entities than in Wien.

By following the above idea of opening a pub in Wien it is highly recommended to choose the place for the new place ulterior to college areas as the near surroundings are already taken up by existing Bars and Pubs.

We can prove this recommendation by visualizing the resulting data sets into bar charts. We can visually compare the shortest ways to Pubs in Bratislava and Wien, then the longest routes in our samples and finally the average distances in these two cities.

A picture containing chart

Description automatically generated

Chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

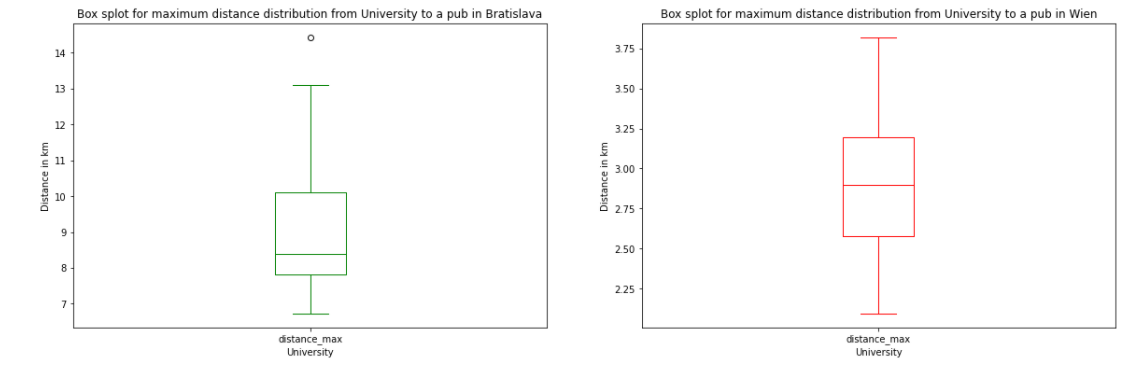
Chart, bar chart

Description automatically generated

The statistical distance distributions are plotted with box charts:

A picture containing shape

Description automatically generated



Chart

Description automatically generated

It can be clearly seen that despite the fact that in Bratislava there is a higher percentage of Pub and Bar occurrence in a given College neighborhood, their accessibility in terms of distances has lower range.

# Discussion

We need to point out the fact that the selection criteria for venues in these two cities differ in the range of examined radius. We had to narrow it down for Wien to avoid getting too much entities when compared to Bratislava. This can be led to biases when determining the results of distances between recognized Colleges and Pubs.

Finding big enough and still comparable data sets might be a recommended approach for further studies in this area. Also needs to take into consideration the geographical parameters of these two cities as Wien is situated in a largish area than Bratislava that might affect the geographical distribution of the above analyzed venues and entities.

# Conclusion

The main approach of this report was to answer the following two questions stated in the introduction section:

“Where should the investor open a pub”

“How far should I go to the nearest pub from my college if I am a student”.

To answer the first one, a potential investor should invest in opening a new bar or pub rather in Wien than in Bratislava. Our sample data shows a clear fact that there are more Pubs and Bars in College neighborhoods in Bratislava than in Wien. On the other side they are situated more far from each other in Bratislava than in Wien. As the shortest distances are smaller in Wien, the recommendation would be to open new bars and pubs outside the college areas.

Students looking for the nearest refreshment can find the answers in the results sets of the attached Jupyter notebbok containing the shortest way from each university in the sample data.