# 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 4 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.