# Capstone Project
# Battle of the Neighborhoods

## Planning a trip based on city names

10/06/2021

**Table of Contents**

# 1. Introduction

The problem that I will be solving during this capstone project is regarding a trip to the United States. I hypothetically have saved a lot of money, such that I can go there, but my goal is very specific. My first name is “Anne”, so the goal of my trip is to visit all cities with the name Anne in their name. To make it even more specific, I really want to let my family know that I have visited those places, so that is why I am looking for a post office in each of these cities. This is the point where data science will come in since my goal is to make a list of all cities that include the name Anne and to find the post offices within these cities. Because based on this, I can plan my trip to the United States.

One last note, I will make this algorithm in a very general way, such that people with similar interests can use it to plan their trips too. For example, when people want to visit all restaurants in cities that include an ‘X’.

# 2. Data

## 2.1. Data Acquisition

The main data source that will be used is from the website OpenDataSoft (https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/table/) and this data source includes 43.191 cities in the United States with their respective zip codes, timezone, daylight savings, and longitude and latitude values. This data source will be downloaded in the form of an xlsx file and imported into Jupyter Notebook.

## 2.2. Data Cleaning

After data acquisition, the data cleaning process will start. First, the unnecessary information will be deleted to keep an overview of the available data. This includes dropping the columns “zip”, “Daylight savings time flag”, and “Timezone”. Afterwards, the data needs to be sorted based on the name of the cities, such that only the name of the string that is being looked for, Anne in this case, is in the new data frame. This is done by looping over all rows of the data frame and appending the rows that include “Anne” in the city name to a new data frame.

# 3. Methodology

The methodology includes several for-loops, for example, over each city in the dataframe that has been established as described above. For each city, I will make an API call to FourSquare to get a JSON file. This JSON file will be organized in such a way that it is a data frame with the venue id, category, and address. Based on this new dataframe, all the venues with a certain category can be extracted to answer the final question of what the addresses of certain types of venues in certain cities are.

# 4. Results

The result is a dataframe with the addresses of all postal offices in cities that contain the name “Anne”. For this specific example, these are:

- 31676 Eden Allen Rd (Rt 13)', 'Eden, MD 21822', 'United States
- 702 Saint Charles St', 'Beaverville, IL 60912', 'United States
- 2560 County Road 4', 'Minter, AL 36761', 'United States
- 8545 County Road 59', 'Furman, AL 36741', 'United States

However, this can be repeated with any type of string and any type of venue. So for example all restaurants in cities with Anne are:

- Seafood restaurants:
    - 12138 Carol Ln', 'Princess Anne, MD 21853', 'United States
- Fast Food restaurant:
    - 30617 Backbone Road, Student Services Center', 'Princess Anne, MD 21853', 'United States
    - 30362 Mount Vernon Rd', 'Princess Anne, MD 21853', 'United States'
    - '12112 Brittingham Ln', 'Princess Anne, MD 21853', 'United States
- Italian Restaurant:
    - 12302 Somerset Ave', 'Princess Anne, MD 21853', 'United States'
- American Restaurant:
    - 30361 Mount Vernon Rd', 'Princess Anne, MD 21853', 'United States'

# 5. Discussion

The results show that we can visit a specific type of venue in a specific type of city. One point worth mentioning is that it takes venues based on a certain radius from the city center of this city. At the moment, this radius is set to be 10 km. However, as you can see for the postal offices, this does not mean that all the postal offices are in cities with “Anne”, but within a radius of 10 km within those cities. I have set this radius at this range, since most cities with “Anne” in their name are very small and otherwise there would only be one venue. However, it is good to take this into consideration when repeating similar assignments with larger cities.

# 6. Conclusion

It is possible to make an algorithm to determine a specific type of venue in a city with a certain name. In case for my holiday to all cities with Anne, I can send postcards from:

- 31676 Eden Allen Rd (Rt 13)', 'Eden, MD 21822', 'United States
- 702 Saint Charles St', 'Beaverville, IL 60912', 'United States
- 2560 County Road 4', 'Minter, AL 36761', 'United States
- 8545 County Road 59', 'Furman, AL 36741', 'United States

Thus, my goal for this capstone project is fulfilled!