## Introduction

Dave Harper is the quintessential New Yorker; he loves his state. There is no corner in the state of New York where he hasn’t been already, he knows the best places for the perfect occasions. Mr. Harper is known everywhere in the state because of the reputation he earned through the years, he even got a nickname from all the knowledge he has about New York, he is called “York”. He is faster than some platforms to recommend the same places to go and enjoy.

## Business Problem

Mr. Harper finally managed to get the job of his dreams as a Data Scientist in one of the top Video Game companies of the century. The job has one bad side, he needs to relocate to one of the available offices distributed around the globe. The states with available openings are:

* California
* Nevada
* Washington
* Florida

Since Dave is not going to let this opportunity of a lifetime go to waste, he decided to move to one of the proposed states with one condition: “The State needs to be the most similar to New York”. The employer could not answer his question, so he decided to use his amazing knowledge of New York and Data Science skills in order to pick a territory that will become his new “New York”.

Dave asked the employer for a couple of weeks so he could do a more detailed analysis, but the Video Game company needed an answer on the next two days, otherwise, they would consider another person for the role. Mr. Harper did not have much time to transform all his knowledge into a numeric form to measure or compare all the cities within each region, so he decided to get the data from the platform he knew it was the most similar to his knowledge to NY, the name of the platform is ‘Foursquare’.

After a couple of hours thinking, Dave defined the KPI’s in which he would base his analysis to conclude that a state is ‘similar’ to New York:

* Distribution of venues
* Categories available
* Quality of venues

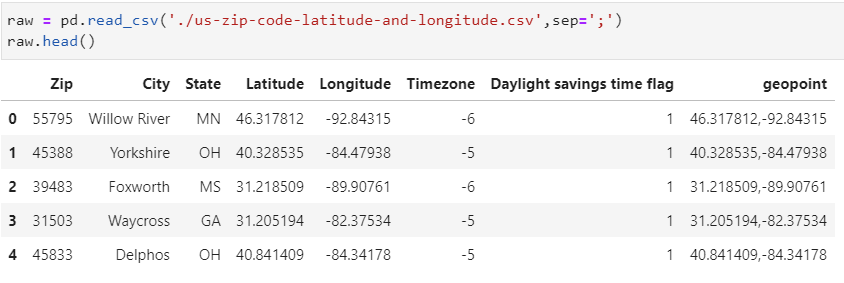
***Distribution of Venues:* There has to be a similar distribution of venues along side the territory. This means that the concentration of venues does not has to be only in one place, but it needs to have some of them placed alongside the rest of the territory.**

***Categories Available:* The destination must have at least eighty percent of the categories that New York possess to be enjoyable for Mr. Harper. A state will have the category if there is at least one venue marked within it.**

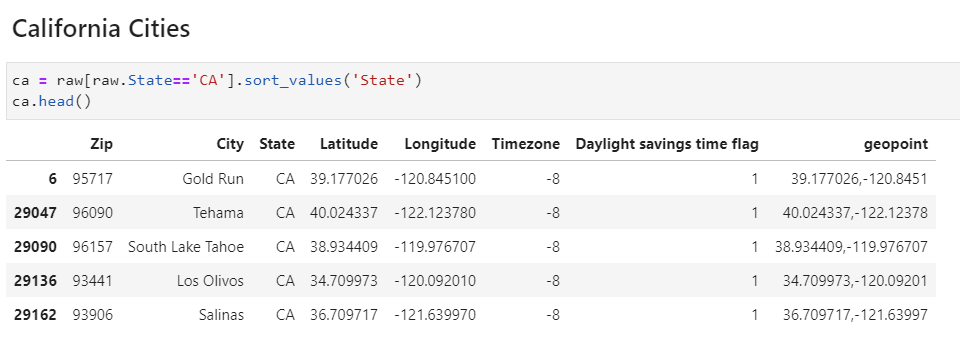
***Quality of Venues:* There needs to be a similar percentage of likes given to the top hundred venues. The percentage will be defined by the likes of the top venues divided by the total likes that the state contains.**

## **Data**

The data that will be used for this project was extracted from public.opendatasoft.com and foursquare.com. The first dataset contains data about all of the geographical locations in the US, it has the following columns: Zip, City, State, Latitude, Longitude, Timezone, Daylight savings time flag and geopoint.



The main dataset will be divided into five datasets which will contain the data from states of: New York, California, Nevada, Washington, and Florida. For example:



After the division, the dataset will be merged with data from the available venues in the area for each zip code. There are 5193 queries need to be done to get all the data required to complete the analysis, there is the need of a strategy to get them all of them.