Coursera Capstone Project - Week 5
Finding US cities similar to Phoenix

Bhaumik Choksi

February 18, 2019

Introduction

A city can be defined broadly as a collection of various commercial, governmental and residential properties. Identifying similar cities can have many potential applications. Lawmakers, businesses and even common people can use these relationships to draw relevant insights.

This project aims to leverage location and venue data in order to identify cities similar to a given city (Phoenix, Arizona in this case).

Business Problem

Given a city, can we use the data about the nearby venues to identify other cities that are similar to the given city in terms of the composition and availability of these venues.

More specifically, *what cities in the US are similar to Phoenix in terms of venues*?

Description of the data

I will be using data from two different sources for this project - Foursquare and the US Cities Dataset. The Foursquare API provides data about the venues whereas the US Cities Dataset provides city names and location coordinates.

Foursquare API

Website: https://developer.foursquare.com/docs/api/venues/details
The Foursquare Places API allows users to get details about nearby venues for a given location. I've used this API to get the following details about 10 nearby venues for a given city:

1. Venue Name
2. Venue Latitude
3. Venue Longitude
4. Venue Category (Short name)

US Cities Dataset

- Original Name: 1000 Largest US Cities By Population With Geographic Coordinates
- Website: opendatasoft.com
- Source: Click here
- Reference: Miserlou on GitHub
- Disclaimer: I do not own any part of this dataset. All copyrights belong to their respective owners.

I will be using the following columns from this dataset:

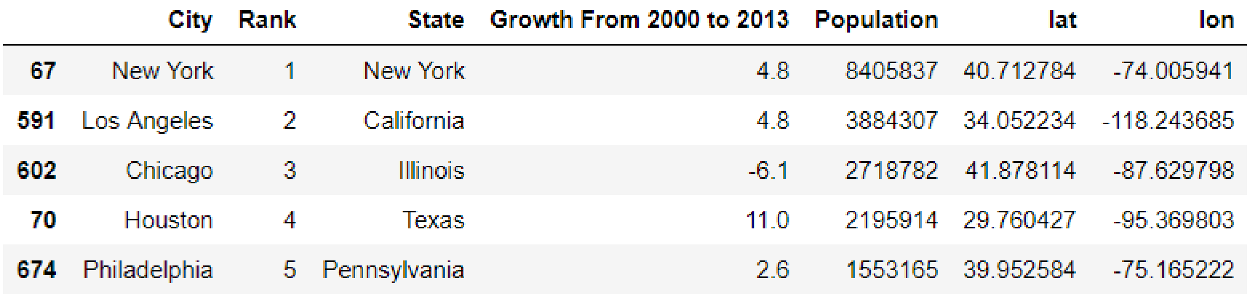
1. City
2. Rank
3. Coordinates

Methodology

The implementation involves three main phases - pre-processing, clustering, and PCA. Each phase has it's own set of plots and other visualizations.

Pre-processing

We start by importing the US cities dataset obtained from opendatasoft.com. We read the CSV file into a pandas dataframe. It contains details about 1000 cities in the United States. In order to simplify plots and to make sure we don't exceed the free API usage limit, I only select the top 50 cities based on population.



	City	Rank	State	Growth From 2000 to 2013	Population	lat	lon
67	New York	1	New York	4.8	8405837	40.712784	-74.005941
591	Los Angeles	2	California	4.8	3884307	34.052234	-118.243685
602	Chicago	3	Illinois	-6.1	2718782	41.878114	-87.629798
70	Houston	4	Texas	11.0	2195914	29.760427	-95.369803
674	Philadelphia	5	Pennsylvania	2.6	1553165	39.952584	-75.165222

Figure 1: Cities Data

Next, I use the latitude and longitude of each city to obtain the nearby venues using the Foursquare API. I obtain the top 10 venues for each city and extract the name and category of each venue. I store these in a list that is later converted to a dataframe.

In order to process this categorical data, I one-hot encode the venue categories so that each category is represented as a column. I take the mean of this dataframe to ensure that each category value represents the fraction of the venues that belong to that category for that city.

Clustering

In order to identify similar cities, we must cluster the data. Our target here is Phoenix, so we'll be looking for cities that are assigned the same label is Phoenix.

I start by importing Kmeans from the scikit-learn package. I cluster the one-hot encoded data into 5 clusters. I decided to pick k=5 since we're working with 50 cities, so that would give us 5 clusters with approximately 10 cities per cluster.

The kmeans labels are stored in the dataframe along with the one-hot values.

PCA

Results

Discussion

Conclusion