**Recommend the place to live in Phoenix Area by analyzing house price and venues data**

**A. Introduction**

A.1. Description of Background and Problem

A.1.1.

The United States economy is on the brink of hitting its longest-lasting period of growth ever — but no one can quite get the idea of a coming recession out of their heads. Several signals such as inverted yield curve and lead economic index has turn red, which making most people afraid that depression will not be too far.

During the recession, many economic elements, which intimately related to people's lives, such as gas price, traffic, salary, employment, have been greatly affected. The house price belongs to one of these economic elements. During last recession, the house price reached the peak and sharply dropped after the recession began.

A.1.2.

I love the place I live in the Columbus, Ohio. But I will move to Arizona, Phoenix Area. I plan to own a house, whether by renting or purchasing, and I hope that the community environment of the house, in which I will live, are pretty the same as I’m living right now. I hope the price of the house I will buy or long rent won’t drop significant during recession.

A.2. Data Description

To consider the problem we can list the data as below:  
1. House price data: will use the Five-Digit ZIP Codes Annual House Price Indexes (Developmental Index; Not Seasonally Adjusted) to represent the house price.  
2. US zip code with some basic data  
3. Venues data: will use Foursquare API to get the most common venues of given Zip code of Phoenix Area and Columbus.

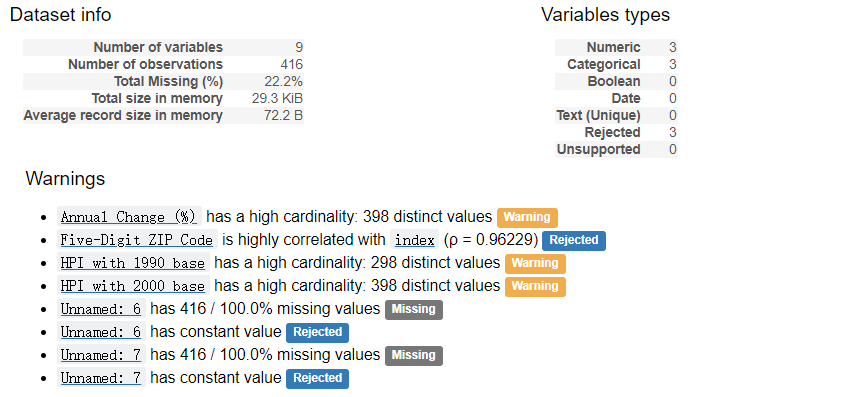
**B. Data preparation and format**

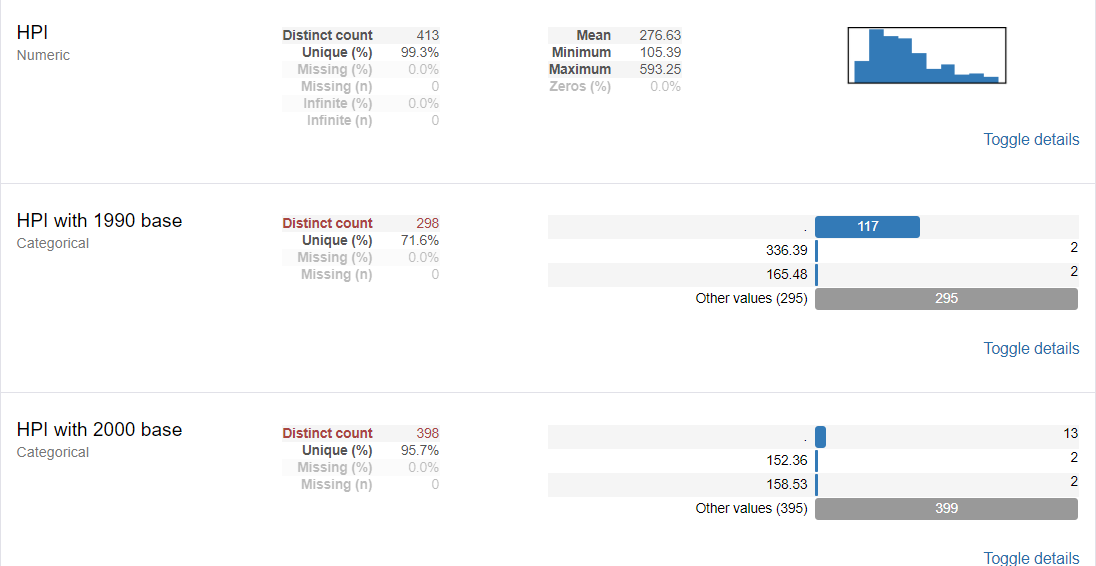
B.1 Data cleaning

The US zip code data and House price data contain many information.

B.1.1 House Price data:

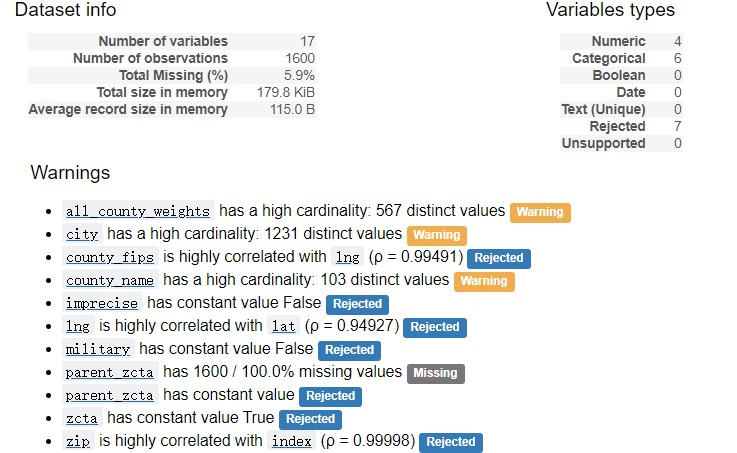
The House of an place(zip) are described by HPI: house price index, and it also contain several useless columns like: Unnamed:6, Unnamed:7. As the HPI with 1990 base and HPI with 2000 base I decide to only use HPI to evaluate the house price of each target place.





B.1.2 Location data

It contains many columns that may not need. I believe, to describe a place whether good or not fit for someone to live, population, race, security, service, traffic and so on.

The service data can be generated by the category data getting from Foursquare, and the others can be got by some other API, but in this report, I will only use the density, population, which are free from the Internet, only.

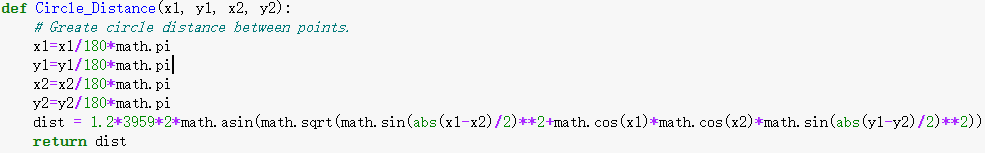
B.2 Feature engineering by getting data from Foursquare

B.2.1 Distance calculation

To descirbe a place, I believe the distance that can normally reach should be considered.

1. Taking 50km/h as the seed to travel in car, 15 min will take me to the 12km away from I live.

2. We should consider Manhattan distance and great circle distances as the distance between to place on the earth because of road design. To simplify the calculation, I set 1.2 times of the great circle distance. The function to calculate the distance between two point on the earth (ignore the sea level distance) will be the following:

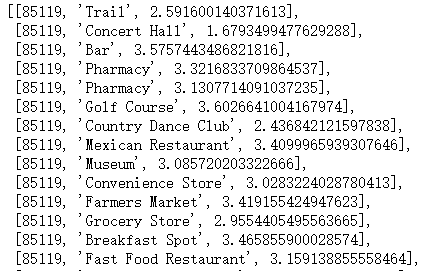


Then I make a dictionary to store the neighbors of each place with distance less than 6 km.

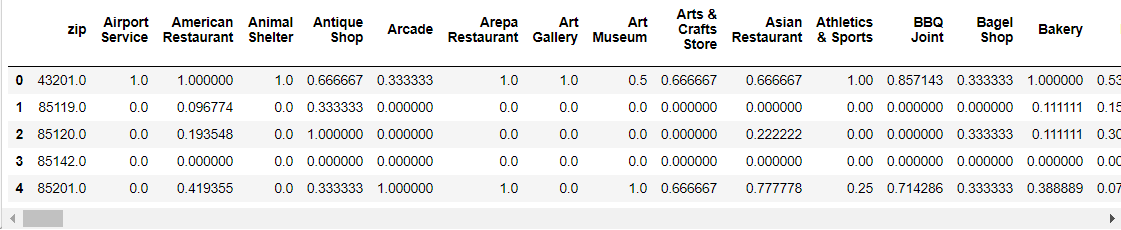
B.2.2 Foursquare data

Utilized the Foursquare API to explore the boroughs and segment the neighbor. To get the almost all the insight of the neighbor, I set the limit as 500 venues. And, also, to cover the place that I can reach by car I set the radius 5000 meter for each neighbor from their given latitude and longitude information.

For each target I will get a straight forward dictionary contains only the zip, category name, and distance of the category to the zip in it.



This will be used to identify a target place. Then merge this data to the Location data to generate the evaluate vector for each zip (all the factor have been min-max-scaled, in this report I consider all the factor have the same weight, but to consider real word, it should give weight by evaluating personal preferences ):



B.3 House Price Data

B.3.1

As to understand the change of the price, we should find the lowest price, the highest price and so on. To find the lowest price, I really surprise that all target got their lowest in 2011. So, the way of their change is useless. To compare the house price, I have to treat the house as an investment. Then is easy to give them a score. I choose 4 time points to calculate--2006,2011,2014 and 2018. All the 3 points after 2006 will use interest rate 3% as the expected investment response rate to calculate the money:

'score' = ('2014'-‘2006’)\*pow(1/z,2014 - 2006 +1) + (‘2011’-‘2006’)\*pow(1/z,2011 - 2006 +1) + pow(1/z,2018 - 2006 +1)\*( '2018'-‘2006’)/( '2006')\*( 'Pay')-( 'Pay')\*pow(z,2018 - 2006 +1)

Then min-max-scaled the data to get the final score.

**C. Evaluate the similarity and Give recommendation.**

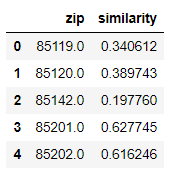
C.1 Calculate the similarity of the place I live now with the target place.

There are some different kind of way to calculate the similarity of two data, after evaluated the advantage and disadvantage of the function, I choose cosine similarity.

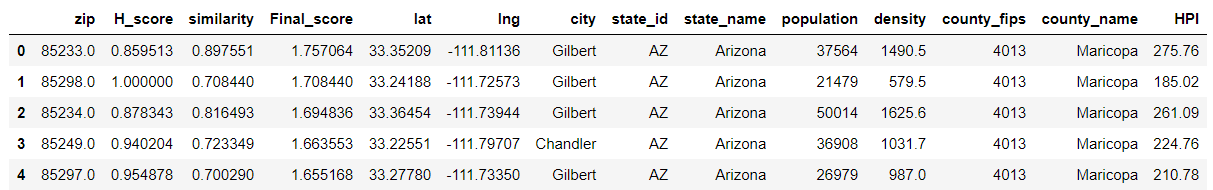
Cosine similarity:

<https://en.wikipedia.org/wiki/Cosine_similarity#targetText=Cosine%20similarity%20is%20a%20measure,(0%2C%CF%80%5D%20radians.>

I directly use scikit-learn.metrics.pairwise.cosine\_similarity package to calculate. The following are the sample of result:



C.2. Give the final Score by combine the House price score with Location Similarity score(min-max-scaled)

I chose the first five place as the recommendation: 

**D. Further Thinking**

D.1. Improving

1. Category can be clustered and be weighted.

2. More detail data like safe level, tax and other factor should be involved.

3. Keep thinking!!!!

D.2. Expansion

This project can be used to give recommendation to people who want change city or place to live. The zip code can be change to single house with adding the detail of the house.