1. Introduction

1.1 Background

A well-established Chinese real estate company is planning to enter the United States market, and it need some helpful information that can be obtained from location data in order to determine the strategies. China and United States are two very different countries in many ways including real estate market. For example, Chinese residence types are mostly apartments, but United States residence types are mostly single house and town house. Chinese commercial malls tend to be buildings with high floors while United States has more low floors buildings. The differences in real estate market are caused by multiple reasons including culture and geolocation, and the Chinese real estate company believes that the locations data can help them discover the similarities and differences between two markets. The findings can benefit the Chinese real estate company for reducing costs and making better decisions.

1.2 Problem

The Chinese real estate company plan to use location data such as city and venue to find the cities which are similar with the Chinese cities where the Chinese real estate company has already succeeded. Then, the Chinese real estate company can apply similar strategies to those United States cities in order to reduce costs and obtain bigger success rate.

2. Data Acquisition and Cleaning

2.1 Data Sources

The city, longitude, and latitude data can be obtained from the web, and the datasets contains most of the major cities of China and United States. The cities' venues data can be obtained from Foursquare API.

2.2 Data Cleaning

The dataset that contains city, longitude, and latitude data is well organized and collected from professional data collectors. All that must be done is to obtain 100 most common venues for each city from the Foursquare API. Some small cities may not have enough common venues, so those cities will be eliminated.