**Columbus Neighborhoods and Houses Assessment**

Sheng Guo

6/27/2020

# Introduction

## Background

When we move in a house or apartment, we live not only in a building, but also in a neighborhood. We not only care about conditions and price of a house, but also safety, convenience, and good schools for kids. We need to evaluate the neighborhood of selected houses based on nearby venues, schools and crimes. Based on clients' preference, I design a ranking system and evaluate each neighborhood. I return a report to help clients to have a better assessment of the houses' environment and narrow down options.

The audiences whom this project aims at are real estate buyers and sellers.

## Problem

Assess neighborhoods of selected houses in Columbus Ohio by looking at nearby venues, schools and crimes, and use the assessment to help clients to make decision of purchasing a house.

# Data Acquisition and Cleaning

## Data Sources

**Houses Information**: The data was from Realtor.com. I selected 10 houses with the following filters: Location: Columbus OH; Price: from $180,000 to $220,000; Bedroom 3+; Bathroom 2+; Property Type: Single Family. The data collected includes address, the number of bedrooms and bathrooms, the inner and outer dimensions (sqft) and the price.

**Nearby Venues**: The data was obtained through Foursquare API. A list of recommended venues within a mile from the specified location was obtained through Foursquare API’s Venue Recommendations (https://api.foursquare.com/v2/venues/explore).

**Nearby Schools Information**: The data was from greatschools.org. The website provides schools’ information and different scores to evaluate each school. In this project only public schools (including district and charter) within 5 miles from each selected house were considered. I selected the GreatSchools’ Summary Rating from the website as the main evaluation each school. The Top 1 school's score, Top 3 schools' average scores and the total number for elementary, middle and high schools respectively are collected. The Top 1 and Top 3 average scores represent the best choice of nearby schools and might be important for parents who want their children to go to top schools. The total number of schools represent the education resources available in the area.

**Crimes Information**: The data is from Location Inc’s website <https://locationinc.com/claim-probability-tools/residential-crime-claim/>. The CrimeRisk™ Score and Probability of a Crime Claim from the website were collected. The CrimeRisk™ Score is calculated by Location Inc to indicate the insecurity of the communities, more details can be found in <https://locationinc.com/crime-risk-data/>.

**Clients’ Preference Weight**: To personalize the assessment for clients, I designed a Clients’ Preference Weight form to collect clients’ preference. Clients are asked to put weight (from 0 to 10) on the significance of each specific category related to nearby venues, nearby schools in their consideration.

## Data Cleansing

I grouped the nearby venues data into different categories and obtained the counts for each category. The categories considered are: Food, Other, Shopping, Recreation, Grocery/Market, Gas\_Station, Convenience/Pharmacy, Auto, Home\_Supply, Bar, Bank/Atm, Pet, Beauty.

# Data Analysis

## Data Exploratory Analysis

I first looked into the correlation of houses’ information. I did not find any useful information.

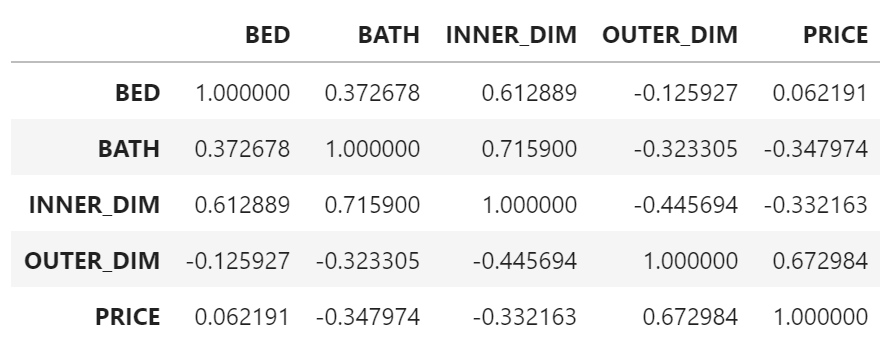


Figure 1

Then I looked into the correlation of crimes information. Crime Risk Score is high correlated with probability of crimes claim. Crime Risk Score will be will be solely used evaluate safety of the neighborhoods.



Figure 2

Next I used PCA (principal components analysis) to study nearby venues and nearby schools. I used K-means to divided the houses into 3 or 4 clusters. But the classification and separation seem not provide much insight.

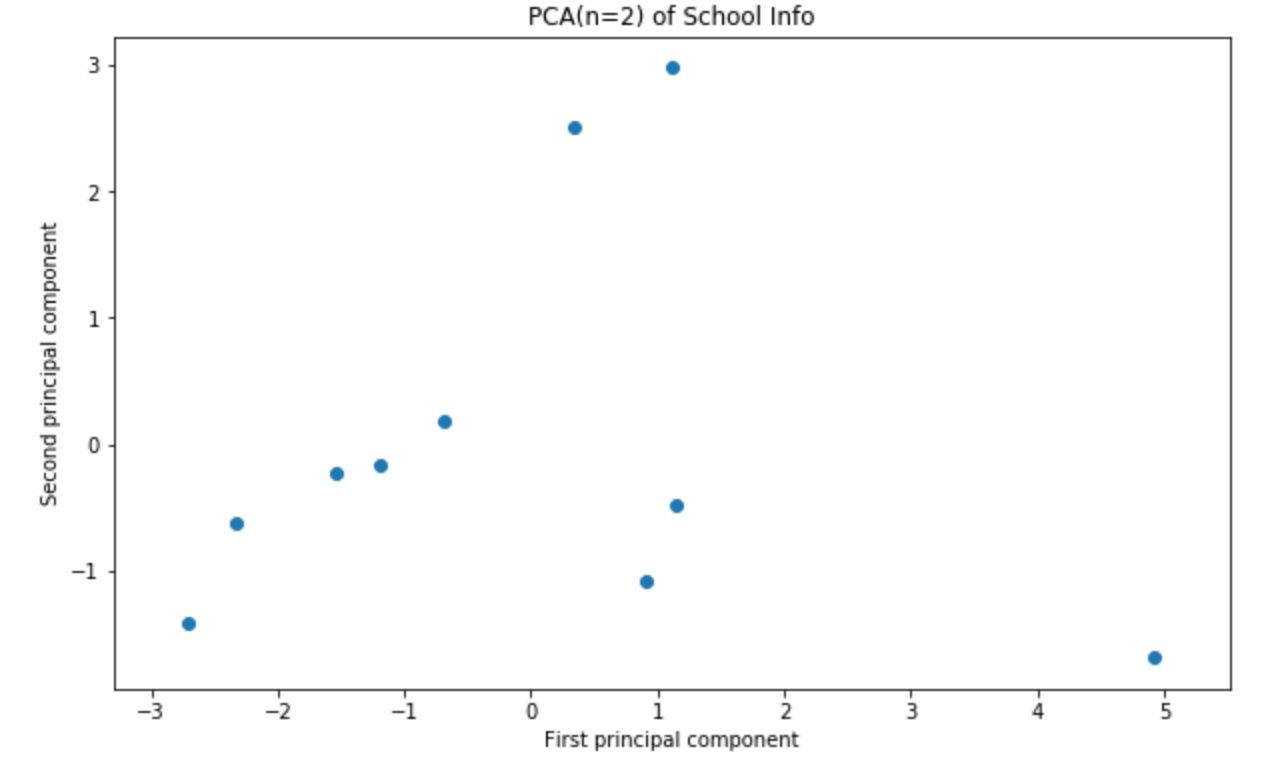


Figure 3

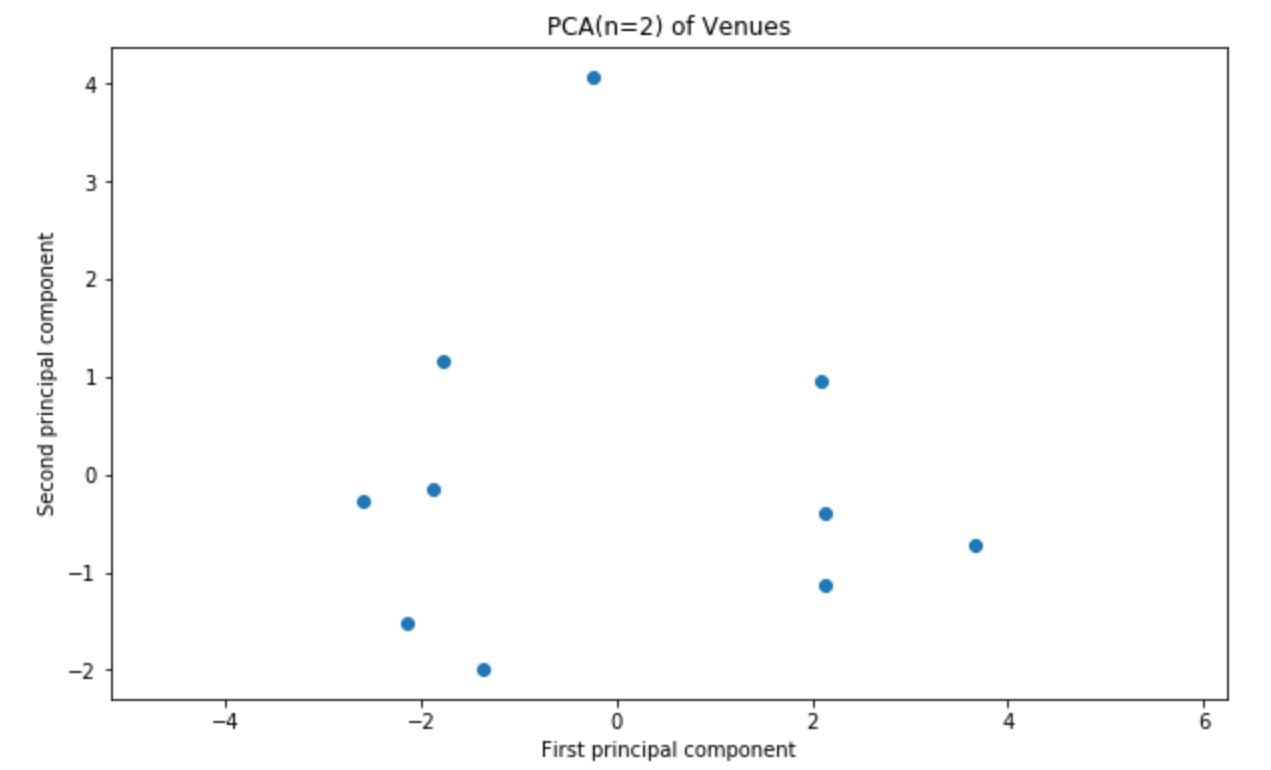


Figure 4

## Evaluation of neighborhoods

I first normalized the data with Z-scale (i.e. mean = 0, std = 1).

I used the counts of venues of different categories to evaluate nearby venues for each house. I used the following formula to obtain a score:

To evaluate nearby schools for each house, I used the following formula:

To evaluate safety for each house, I used the following formula:

RawScore = (1 / CrimeRisk™ Score)

Then I used the following formula to obtain the regularized score:

Score = RawScore × 20 + 50

From above formula, the average score is 50. The further above from 50, the better; the further below from 50, the worse. The neighborhoods score is the average of the scores of nearby venues, nearby schools, and safety.

# Results

# I obtained the evaluation of neighborhoods as follows:

# 

Figure 5

# I used the following bar and line graphs to visualize the results:

# 

Figure 6

# 

Figure 7

# Discussion and Conclusion

# Combining price and neighborhoods assessment, I will strongly recommend the clients to first consider House 1, 6, 7 and 9 based on Figure 6 and Figure 7. These houses have relatively high neighborhoods score while relatively low unit price.

# Since I could not get free access to many data sources related to real estate data, schools data and crimes data, I could not get a good classification of neighborhoods and housing. I believe better data accessibility can help to improve neighborhoods assessment. Even though I only chose 10 houses in this report to reduce the amount of work, but real estate related API can be used to access large amount of housing data. With similar data analysis above, we can offer clients with better options and more persuasive suggestions in a larger pool of housing in a reasonably short time.