**A Guide to Living in Shanghai**

7/30/2019

Section 1: Introduction

Shanghai (上海), the most populous city in China as well as one of the most prosperous and modern cities in the world, is a splendid gem carved by generations of Chinese people and city planners with their remarkable wisdom. Over the past hundred years, Shanghai has thrown off shackles of pain and indignity imposed by [feudalists, warlords and imperialists](http://www.visitourchina.com/guide/culture/semi-colonial-and-semi-feudal-era.html), and has gradually transitioned itself into the economic and financial hub in the world. Known as “Pearl of the Orient” (东方之珠) or “Paris of the East” (东方巴黎), Shanghai at present days has attracted <millions> of adventurous people from the globe to study, tour, and live in this city.

Nevertheless, for people—especially “Westerners”—who are strangers to both Shanghai and mainland China at large, starting a new life on this mysterious piece of land is difficult. For them, questions like which neighborhood to live in, basic statistics of each neighborhood, where prominent venues are located at, to name a few, remain unanswered.

There is a growing body of articles and blogposts on the Internet providing rudimentary city-wise information to the audience who is unfamiliar with the nuts and bolts of Shanghai. Nevertheless, they are mainly qualitative and subjective. In essence, they are not quantitatively rigorous in terms of methodology. Therefore, they run the risk of being unreliable, at least from a data analyst’s perspective. This guide, namely ***A Guide to Living in Shanghai***, is deemed to answer the questions above and to provide an alternative, big data and machine learning-based solution to the potential trailblazers who are about to embark on a new chapter of life in Shanghai.

Section 2: Data

The project uses two datasets, namely *shanghai\_demo* and *shanghai\_data*.

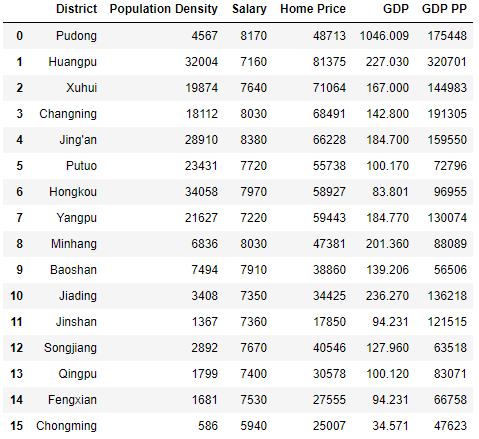
*shanghai\_demo* contains demographic information of all 16 districts (区) in Shanghai. It has 6 variables: *District*, *Population Density*, *Salary*, *Home Price*, *GDP* and *GDP Per Capita*.

In *shanghai\_demo* dataset, *District* is a character column. It lists Shanghai’s 16 districts. *Population Density* (person/square kilometer) is density of population in 2017 (data source: [Shanghai Statistical Yearbook 2018](http://www.stats-sh.gov.cn/tjnj/nje18.htm?d1=2018tjnje/E0202.htm)). *Salary* (RMB/month) stands for personal monthly salary in 2019 (data source: [Sohu](http://www.sohu.com/a/297378775_391502)). *Home Price* (RMB/square meter) is from the same source ([Sohu](http://www.sohu.com/a/297378775_391502)). *GDP* (billion RMB) column has district level GDP in 2018 (data source: [Sohu](http://www.sohu.com/a/314957795_612645)). Finally, *GDP Per Capita* (RMB) (year 2017 data) is sourced from [好金贵财经](https://www.haojingui.com/gdp/5058.html).

Note that all data are from different sources and are in different years. This is because it is extremely hard to find one single source that provides related district-level data. This is the most severe fallacy of this project, but this would not affect the results too greatly given that the data would not change too much within 1 to 2 years if the accuracy of data is fully guaranteed.

The full dataset of shanghai\_demo is:

Table 1: District Level Demographic Information



*shanghai\_data* lists prominent neighborhoods in Shanghai (in both English and Chinese) as well as the districts they belong to. It has three variables: *District*, *Neighborhood* and *Neighborhood Chinese Name*.

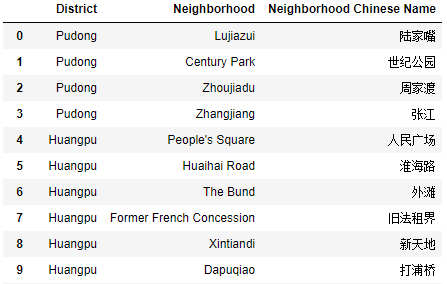
It is necessary to highlight that there is no such concept as “neighborhood” (社区) in Shanghai. “Neighborhood” is essentially a “western” concept and is not used to indicate the same thing in China. In the country, a “neighborhood” is more like a “residential community” (小区) that only has residential buildings rather than a large area that has shopping malls, stores, restaurants and attractions (and of course residential communities). An equivalent concept, in Shanghai, is in fact “subdistrict” (街道).

“Subdistrict,” however, is still not entirely the same as “neighborhood” in the western world. For instance, [Wujiaochang](https://en.wikipedia.org/wiki/Wujiaochang) (五角场) in Yangpu District is essentially a subdistrict (“五角场街道”) in Shanghai’s [township-level divisions hierarchy](https://en.wikipedia.org/wiki/List_of_township-level_divisions_of_Shanghai) and can be treated as a neighborhood to an extent. In contrast, the famous [Xintiandi](https://www.travelchinaguide.com/attraction/shanghai/xin-tian-di.htm) (新天地), an area full of delicacy, art, decent food and fashion in Huangpu District, is not a subdistrict but can be thought as a neighborhood.

Nevertheless, the report adopts the western convention and focuses on a total of 47 neighborhoods (subdistricts/towns) across 16 districts in Shanghai. The author built *shanghai\_data* based on his own discretion.

The top 10 rows of raw *shanghai\_data* is:

Table 2: Prominent Neighborhoods in Shanghai



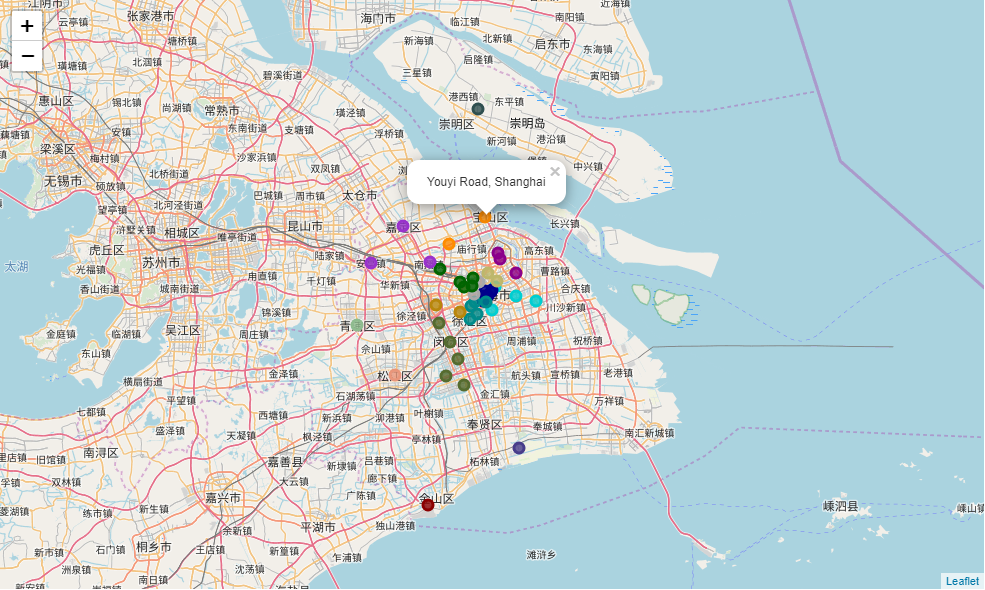
*shanghai\_data* does not have longitude/latitude information for each neighborhood. Thus, further manipulation with the help of Python’s geopy library is conducted. The resulting data frame with longitude/latitude information (top 10 rows) is:

Table 3: Prominent Neighborhoods in Shanghai (with Geographic Information)



With the geographic information of all neighborhoods in Shanghai at handy, we can create a map of Shanghai with all 47 prominent neighborhoods highlighted.

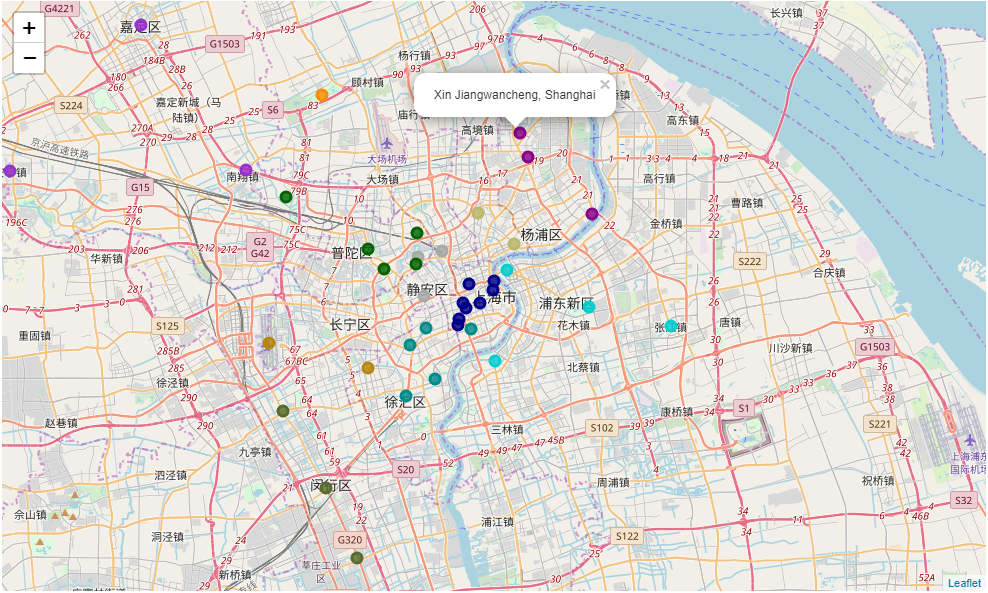
Figure 1: All Neighborhoods in Shanghai



From Figure 1, it is clear that Shanghai is huge. 47 neighborhoods are clustered at different corners of Shanghai (the northernmost point is Jianshe Zhen in Chongming District, and the southernmost point is Jinshanwei in Jinshan District). In Jupyter Notebook, the map is interactive and includes pop-ups (in Figure 1, Youyi Road in Baoshan is highlighted). Furthermore, all neighborhoods are colored by their corresponding districts. Same color indicates that neighborhoods are in the same district.

Shanghai’s breadth and reach make itself astonishing, but it also is not a bad idea to have a glance at Shanghai Proper (上海市区), where its beauty is hidden. Shanghai Proper comprises a series of districts: Huangpu, Xuhui, Changning, Jing’an, Putuo, Hongkou, and Yangpu. Figure 2 shows that most prominent neighborhoods are in fact located in Shanghai Proper (especially in Huangpu District).

Figure 2: All Neighborhoods in Shanghai (Zoomed In)



Section 3: Methodology

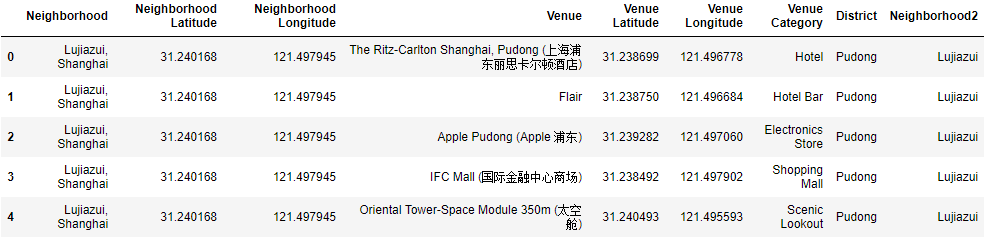
In order to answer questions that interest tourists and newcomers to the city, methods such as exploratory data analysis and machine learning (*k-means* clustering) are adopted.

3.1 Exploratory Data Analysis

The first step is to use Foursquare API to explore neighborhoods in Shanghai. Essentially, Foursquare API enables developers to get trending venues around a location (in this case, around a neighborhood in Shanghai). This answers the key questions of newcomers: what can I access in the neighborhood I live in? Later, various venues’ relative frequency data in each neighborhood would also add value to the machine learning algorithm.

After constructing a URL, setting parameters (number of venues returned by Foursquare API is 200, radius is 500), and sending a request to the API, the following results are returned:

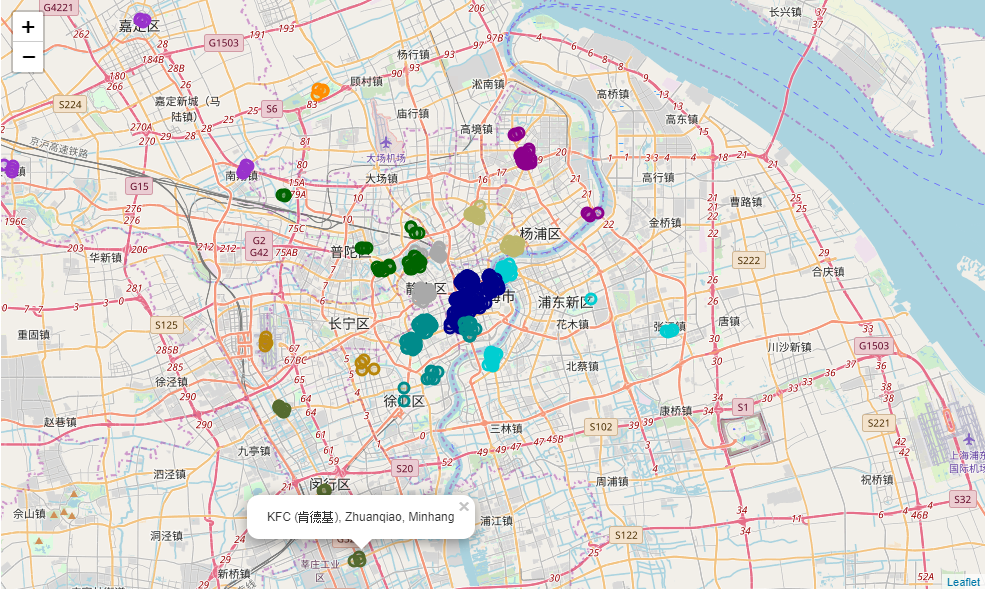
Table 4: Venues Information within each Neighborhood (Top 5 Rows)



The algorithm returns recorded venues (upper limit is 200 venues) within 500-meter-radius of a certain neighborhood. For instance, Table 4 shows 5 venues that are within 500-meter-radious of Lujiazui neighborhood in Pudong.

Notice that venues’ geographic information is also generated by Foursquare API. Therefore, we can create a venue map, as is put below. Again, all venues are colored by the corresponding districts they are affiliated with.

Figure 3: All Venues in Shanghai



Next, with venues data, we can check top 5 neighborhoods that have the highest number of venues as well as top 5 neighborhoods that have the lowest number of venues. Unsurprisingly, the most touristy neighborhoods are renowned Huaihai Road (淮海路), Xujiahui (徐家汇), People’s Square (人民广场), Yu Garden (豫园) and Lujiazui (陆家嘴). These are places where normal Shanghainese people would hangout during their free time. On the other hand, for newcomers, it is advised to avoid living in the neighborhoods that have very few venues.

Table 5: Top 5 Neighborhoods Having the Highest Number of Venues

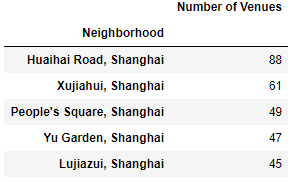
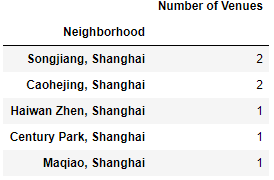


Table 6: Top 5 Neighborhoods Having the Lowest Number of Venues



Next, what are the venue categories a person would most/least frequently see in Shanghai? The tables below tend to answer this question.

Table 7: Most Popular Venue Types

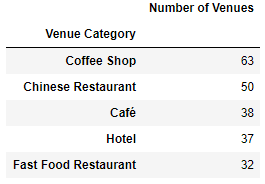
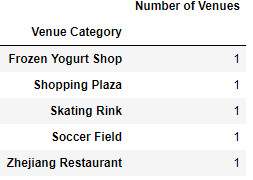


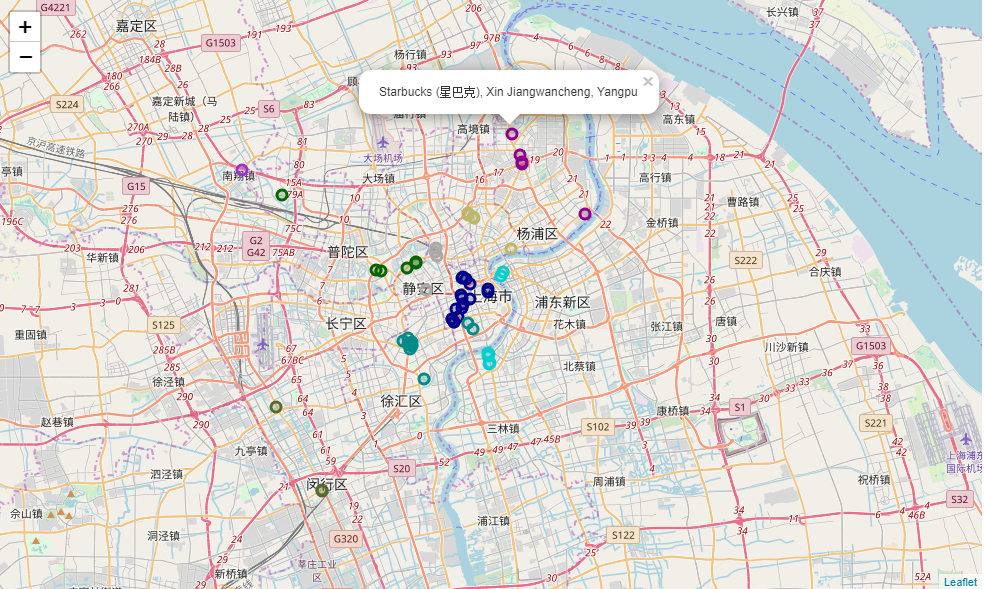
Table 8: Least Popular Venue Types



The outside world usually regards Shanghainese people enjoying a “yuppie” (小资) lifestyle. The big data technique seems to have proved this point. Among all venue types, coffee shops (including café) are most commonly seen. Therefore, why not take your time and chill in one of the coffee shops on Hengshan Road on a warm Sunday afternoon?



Figure 4: All Coffee Shops in Shanghai



Other destinations such as shopping malls and parks are also visualized.

Figure 5: All Shopping Malls in Shanghai

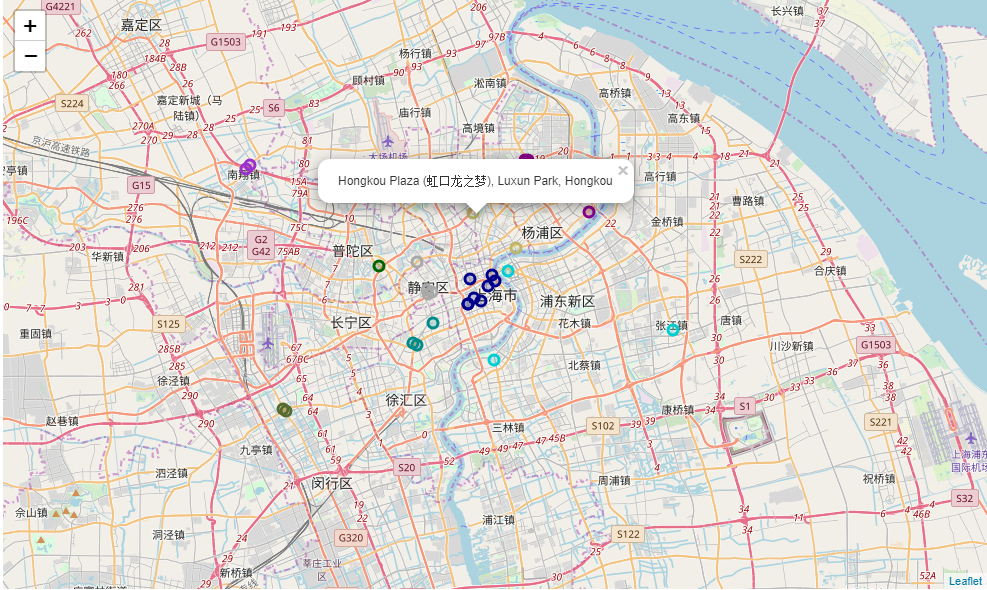
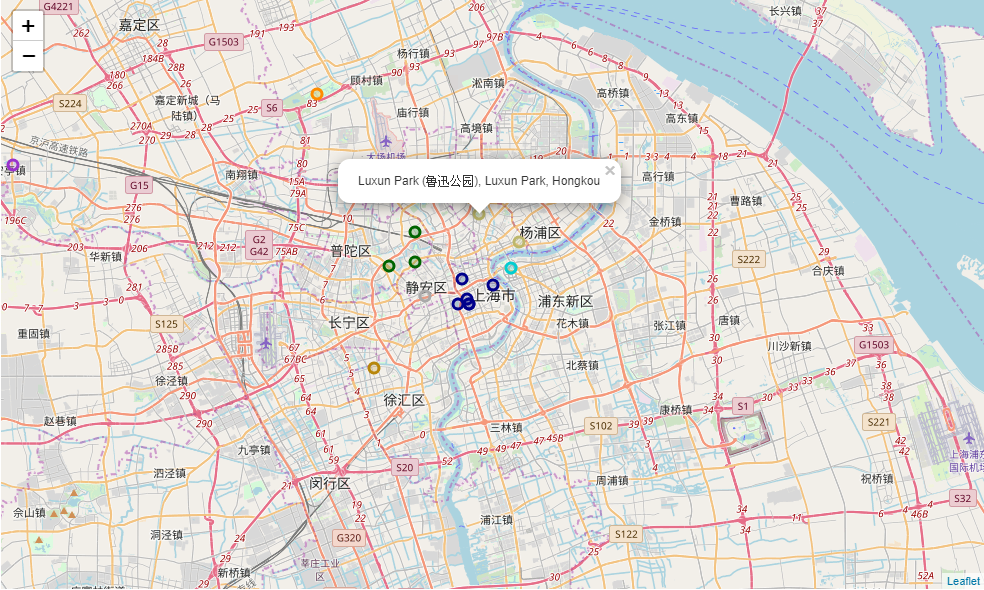


Figure 6: All Parks in Shanghai



The final part of exploratory data analysis is the most popular venue type within each neighborhood. To get the result, we essentially transform the venue information table (Table 4) from long form to wide form. Table 5 is the new data frame (wide form) that has relative frequency of each venue type for each neighborhood. For instance, 0.074074 in row 5 and *Bakery* column means that of all venue types in Changshou Road neighborhood, about 7.4% are bakeries. Table 10 is a table of most popular venue category for each neighborhood.

Table 9: Relative Frequency of Each Venue Type for All Neighborhoods

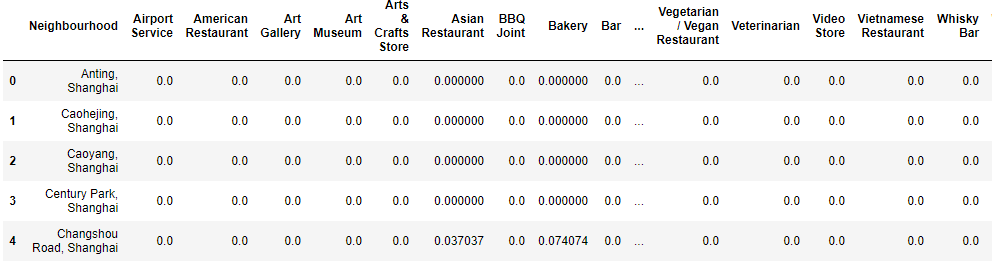


Table 10: Most Popular Venue Type for Each Neighborhood

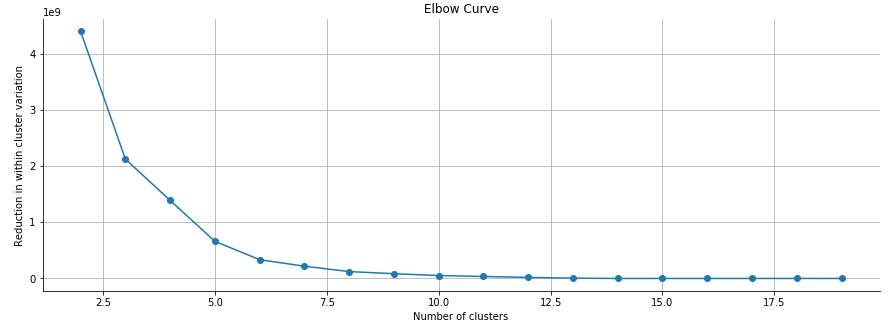


3.2 Machine Learning

We use machine learning to categorize all 47 neighborhoods in Shanghai. Dependent variable is GDP per capita, and all remaining demographic and economic variables as well as relative frequency of various venue types are independent variables used for predicting clusters.

First, the elbow curve tells that 5 is the optimal number of clusters.

Figure 7: Elbow Curve

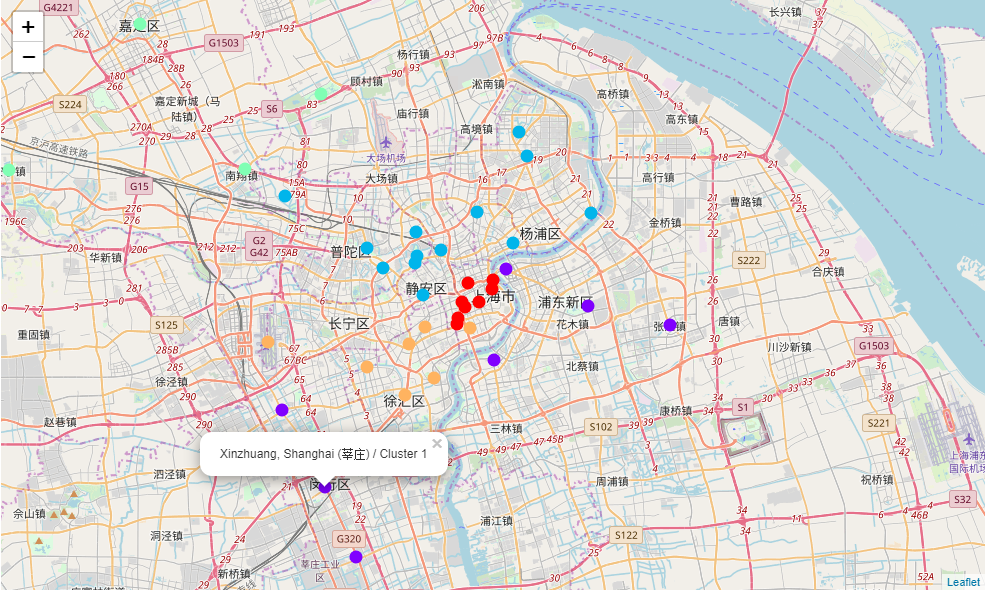


Then, we fit a K-means clustering model with 5 as the number of clusters. The results are illustrated in the next section.

Section 4: Results

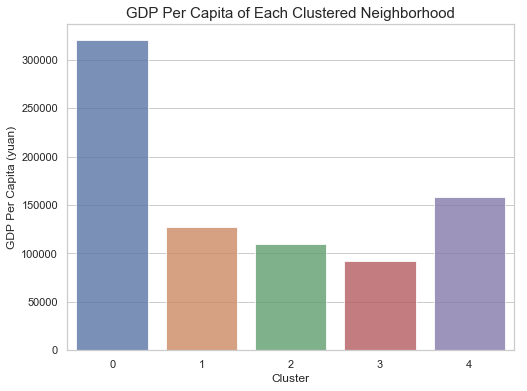
Now that all neighborhoods are assigned specific clusters, we create a new map of neighborhoods colored by clusters.

Figure 8: All Neighborhoods in Shanghai (Clustered)



The results are expected and intuitive. The very central part of Shanghai (Huangpu District, red color) can be interpreted as the center cluster. It is Shanghai’s political and economic center. It is surrounded by the remaining four clusters. As one can imagine, the farther the cluster is away from the center, the more scattered and less developed it is. Therefore, for newcomers, they are recommended to tour and live in the center cluster (red, orange and blue clusters). The pros is that neighborhoods in these clusters are full of attractions and decent dining places, and are “richer” to an extent. The cons, on the flip side, is that it might be costly to live. The bar charts below help illustrate this point. For instance, Cluster 0/Huangpu District (i.e. red-color cluster in Figure 8) has the highest GDP Per Capita. Meanwhile, it has the highest home price among all five clusters (and surprisingly, lowest monthly salary), meaning that it would be extremely hard for a starter to afford the cost of daily living in this cluster.

Figure 9: Bar Charts

Section 5: Discussion and Conclusion

From the exploratory data analysis part in Section 4, one can easily locate her favorite venues in Shanghai and have a rough idea on which neighborhoods are touristy. The machine learning part, however, mainly suggests which district is suitable for a newcomer to live in.

We already know that Cluster 0/Huangpu District (i.e. red-color cluster in Figure 8) is unsuitable. Based on the statistics, it turns out that Cluster 1/Minhang and Pudong District (i.e. purple-color cluster) might be an appropriate option.

Table 11: Statistics of Cluster 1/Minhang and Pudong District



This area is not rich in terms of GDP Per Capita ranking, but people in this cluster tend to earn the most and have relatively less stress to live. Honestly, it is not a big deal for newcomers who want to live in this cluster while having fun during the weekend. The reason is: public transportation is extremely convenient in Shanghai; say if a person lives in Minhang District (purple points at bottom left in Figure 8), she is able to arrive at the center cluster (Cluster 0, Huangpu District) in just 30 minutes by taking Shanghai Metro Line 1 (上海地铁1号线), or arrive at Cluster 4/Xuhui and Changning District (i.e. orange-color cluster in Figure 8) in just 15 minutes by taking the same metro line.

It is essential to mention potential problems of this project. First, like has already been discussed in the Data section, the economic data of each district is not consistent; they are not in the same year and are not from the same source. Unfortunately, only *Population Density* variable is from the authority.

Second, venues in Shanghai generated by Foursquare API are problematic and are limited in numbers. In plain words, it only returns renowned venues (e.g. KFC and McDonald’s for fast food category, and Starbucks for café category). It fails to return less renowned or “Eastern” venues. For example, it is completely untrue that there is only KFC in Zhuanqiao neighborhood, Minhang District. This is probably due to the fact that remote areas in Shanghai has never been a focus of Foursquare API, and the system fails to take into account location information of many venues.

To conclude, this guide presents venues information for each neighborhood across Shanghai, categorizes all prominent neighborhoods, and visualizes them in the map of Shanghai on a big data basis. The guide also recommends newcomers to begin their new lives in either Minhang District or Pudong District.