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

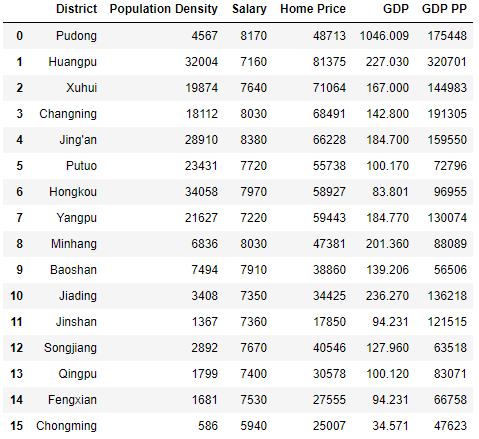
The project uses two datasets, namely *shanghai\_demo* and *shanghai\_data*.

*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:



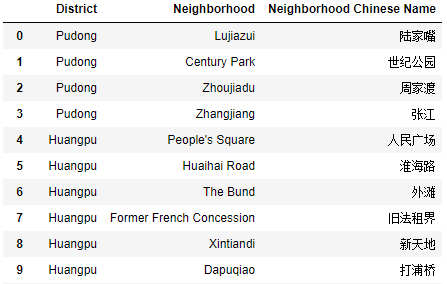
*shanghai\_data* lists prominent neighborhoods in Shanghai (in both English and Chinese) as well as the districts they belong to. It has three variables: *District*, *Neighborhood* and *Neighborhood Chinese Name*.

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:



*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:

