基于大数据的上海生活指南

一、前言

光怪陆离的上海，素来是冒险家的乐园。时光的车轮碾过，上海从一座毫不显眼的小渔村，蜕变为繁华的夜都市。几百年来，一代又一代冒险家们背上行囊，来沪闯荡。有人南下，有人北上，有人西进，也有人漂洋过海，从东而来。

对他们而言，若有一本《上海生活指南》在手，来沪开疆拓土或许会容易许多。如今网络和市面上随处可见各类“上海导游”、“上海指南”，它们多充斥着作者的主观色彩，而缺乏大数据支持。本人编写的《基于大数据的上海生活指南》，力求填补这一空白，为冒险家们提供科学缜密的上海旅居指南。

这份指南将使用探索性数据分析（Exploratory Data Analysis），向读者呈现上海各街区的资源信息。其次，运用机器学习（Machine Learning），将上海的街区进行分类，并向新人提供旅居建议。

指南聚焦于上海的“街区”（Neighborhood）。实际上，“街区”是舶来品，在中国的城市地理学中并不常见。反之，“街区”一词在美国大行其道。例如，纽约曼哈顿便有上西区（Upper West Side）、上东区（Upper East Side）、晨边高地（Morningside Heights）、熨斗区（Flatiron District）等街区。它们通常覆盖城市的大片范围，包含商业圈、住宅圈、生活圈，又因某种特色，而独立成为一片街区。例如，前述的晨边高地街区，坐落于哈德逊河东缘，普遍地势较高。

在上海的行政区划中，与美国“街区”相对应的大抵为“街道”（Subdistrict）。街道比“镇”小一层级。上海人耳熟能详的街道有黄浦区南京东路街道、杨浦区五角场街道、徐汇区徐家汇街道等。它们同样覆盖一片地理范围，有其显著特色。例如，浦东新区的陆家嘴街道是整座城市的金融心脏，而虹口区的提篮桥街道则是颇具历史的住宅街道，同时以提篮桥监狱闻名。

然而，“街道”并不完全等同于“街区”。有一些耳熟能详的区划，例如地处黄浦区的新天地，则不是街道。但是，以雅致、小资著称的新天地大可被认为是美国语境中的“街道”。

在指南中，我采用美式方法，主观选取了上海47处有名的街区。它们既包含官方“街道”，也包含诸如新天地这类的“街区”，前提是它们都能在Python的geopy包里被定位出来。这是由于本份指南会将街区通过经纬度在地图里展示出来。

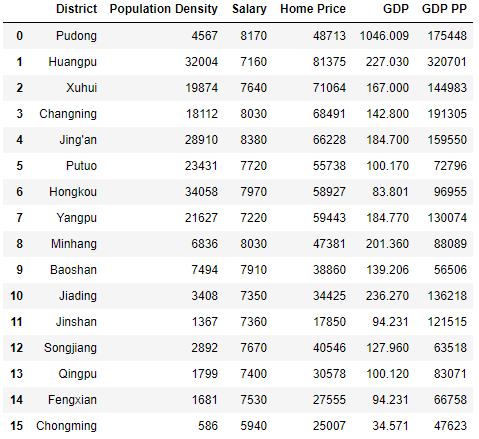
简言之，指南将呈现诸如街区的生活场所数量、上海最受欢迎生活场所等基本学习，并使用机器学习，将47处街区依据相似性进行分类， 最终为读者们提供旅居建议。

二、数据

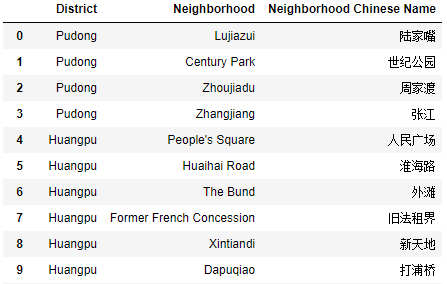
指南将用到的基本数据有两个，一为上海各区经济数据，另一则为上海47个街区列表。

上海各区经济数据有六个变量：区名、人口密度（人/平方公里）、人均收入（元/月）、平均房价（元/平方米）、GDP（元）、人均GDP（元）。所有变量均在区一级。

上海各区经济数据一览：



上海街区数据有三个变量：区名、街区英文名和街区中文名。这个数据包含的都是字符变量。上海街区数据的前10行是：

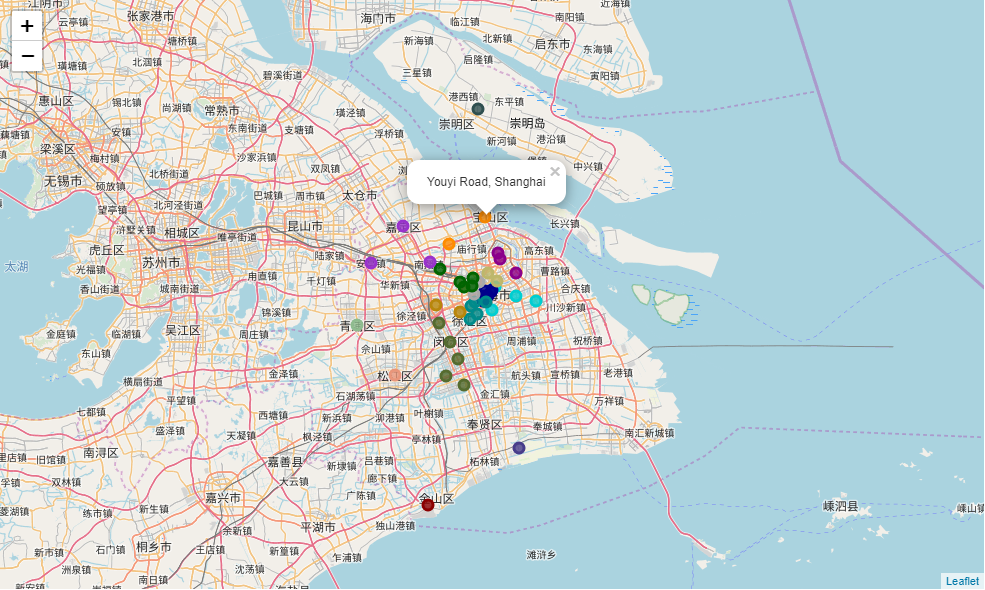


将上海47街区在地图上绘制出来的关键步骤是生成街区经纬度。运用Python的geopy包便可得到：

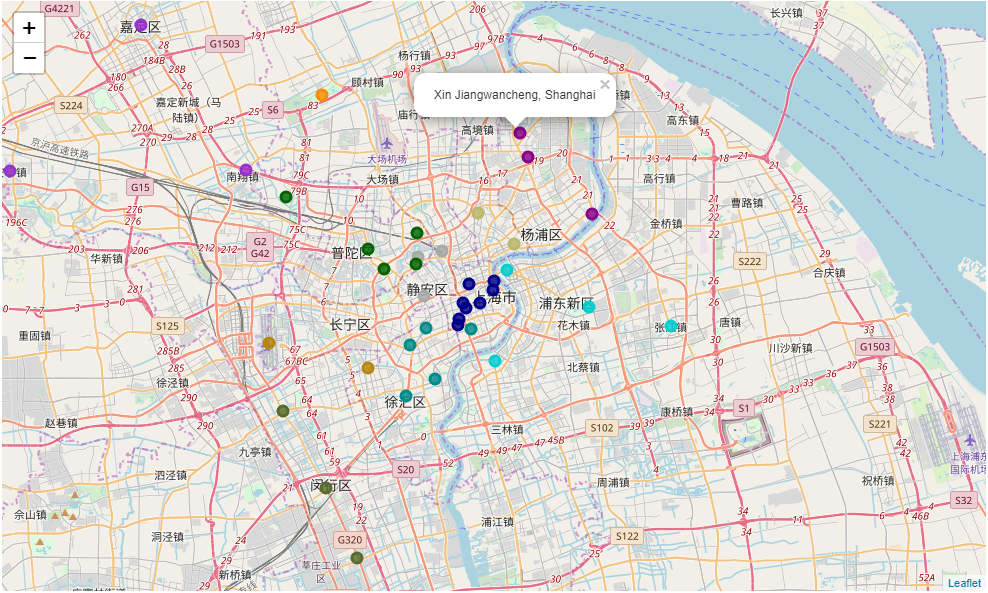




这样一来，便能在Python中把47个街区在上海地图里标识出来。如下图所示，各街区以所在区上色（相同颜色代表街区在同一个区）。



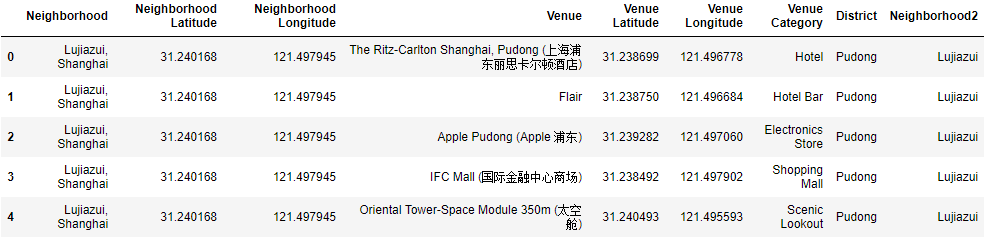
街区最北可至崇明区的建设镇，最南可至金山区的金山卫和奉贤区的海湾镇。而大部分街区，都聚集于上海市区，尤以黄浦区居多，如下图所示：



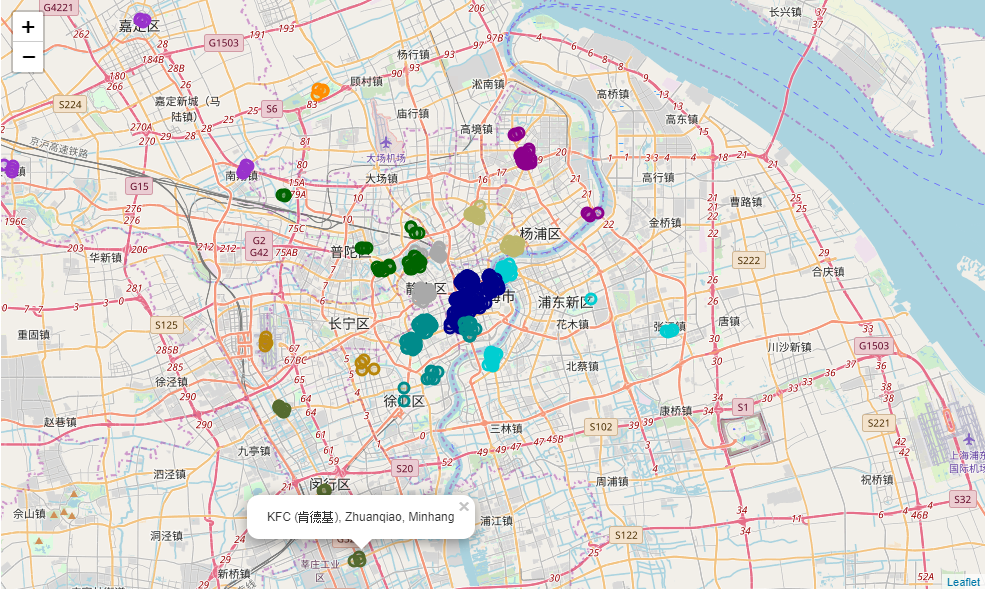
三、方法

探索性数据分析

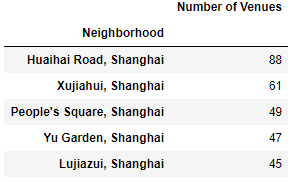
在这一部分，使用Foursquare API得到各街区方圆五百米的所有生活场所及其经纬度。以下是结果的前五行。它举例展示了在陆家嘴街区，方圆500米会遇见的5个场所，有酒店、商场、商店、景点等。

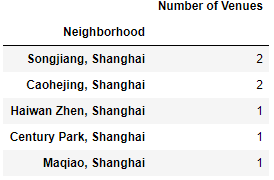


Foursquare API为47个街区生成800个左右的生活场所。我们同样可以将这些街区方圆500米步行可及的生活场所在地图里标识出来。同样，这些生活场所按所在区上色。



有了生活场所信息，能够很容易得出哪些区有最多/最少的生活场所：

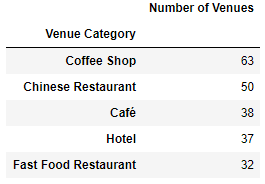


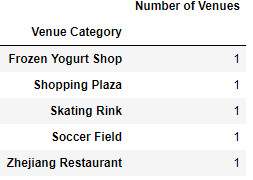


毫不意外，占据榜单前五位（拥有最多生活场所）的街区是耳熟能详的淮海路、徐家汇、人民广场、豫园和陆家嘴。而榜单末五位（拥有最少生活场所）的街区是松江、漕河泾、海湾镇、世纪公园和马桥。对于初来上海的人，可以前往榜单前五位游玩，那里有形形色色的生活场所。反过来说，避免去到榜单末五位的街区。

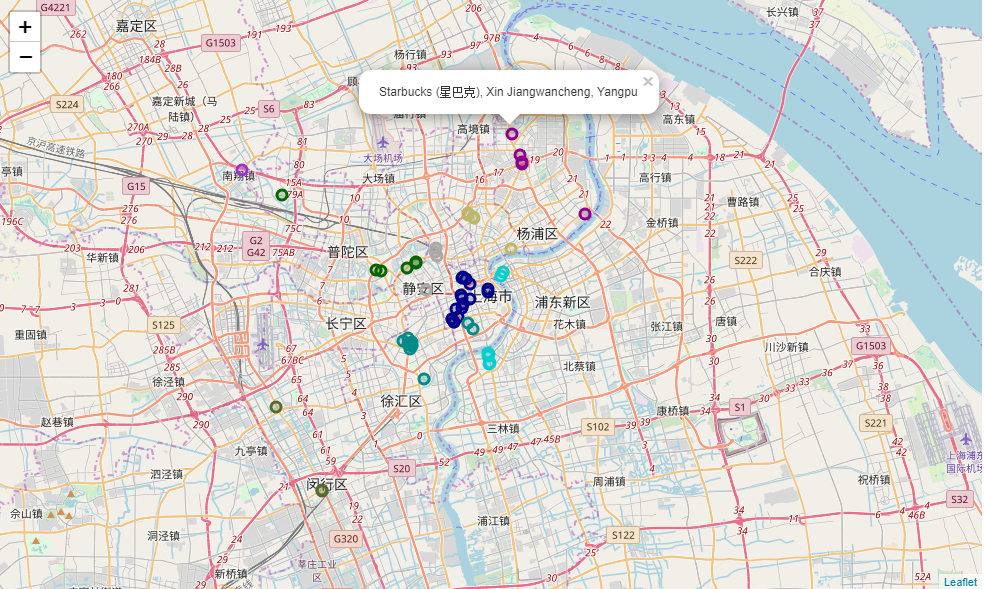
榜单的结果实则漏洞百出，试想，再怎么穷乡僻壤，一个街区不至于只有一个生活场所（例如，在榜单最后一名的闵行区马桥镇，不可能只有Foursquare API生成的马桥体育场这一处生活场所）。这一问题将在指南末尾稍作解释。

其次，哪些是上海最常见/最不常见的生活场所？以下两张表格给出了答案：

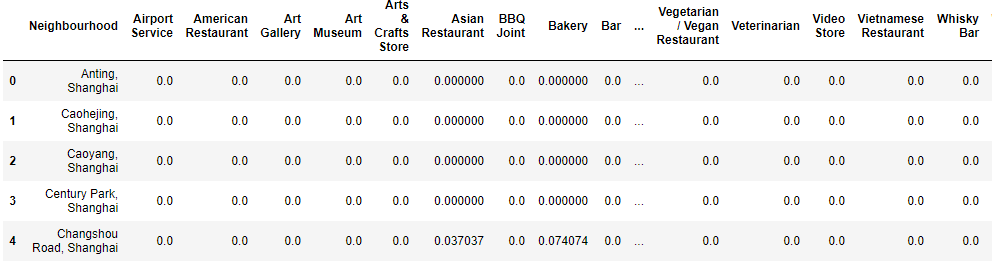




根据Foursquare API结果，上海最常见的场所是咖啡馆（63处Coffee Shop，38处Cafe）。外界通常认为上海“小资”，表格印证了这一说法。如果你也想享受片刻小资生活，莫不如在一个温暖的星期日午后，在徐汇区衡山路独享一杯咖啡。不用发愁找不到咖啡馆，指南已经将上海的咖啡馆标识了出来：



探索性分析的最后一步是想要得到每个街区最常见的生活场所。我们将各街区生活场所表格从长型（long form）转换成宽型（wide form），这样每一行都是独特的街区，每一列都是某一生活场所在该街区出现的频率：



例如，在上表中。第五行/Bakery列的0.074074代表的是在普陀区长寿路街区，面包店在所有生活场所中占比7.4%。如果我们取这张表所有列的最大值，便能得到某街区最常见的生活场所，如下表所示。这样一来，游客能知道自己所处街区的最大特色（例如，如果一个西北人生活在黄浦区老西门，那么吃面食不用愁）。

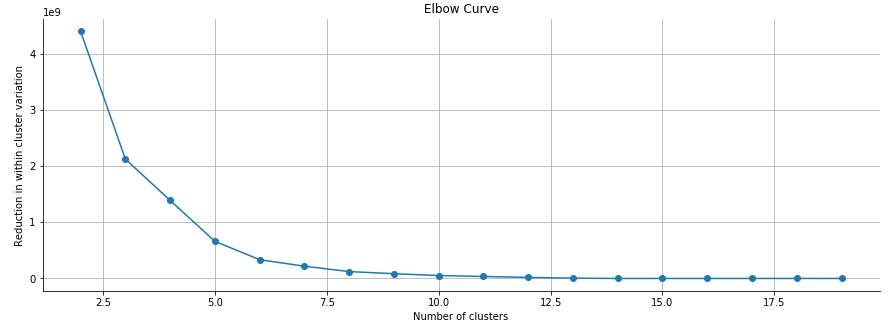


机器学习

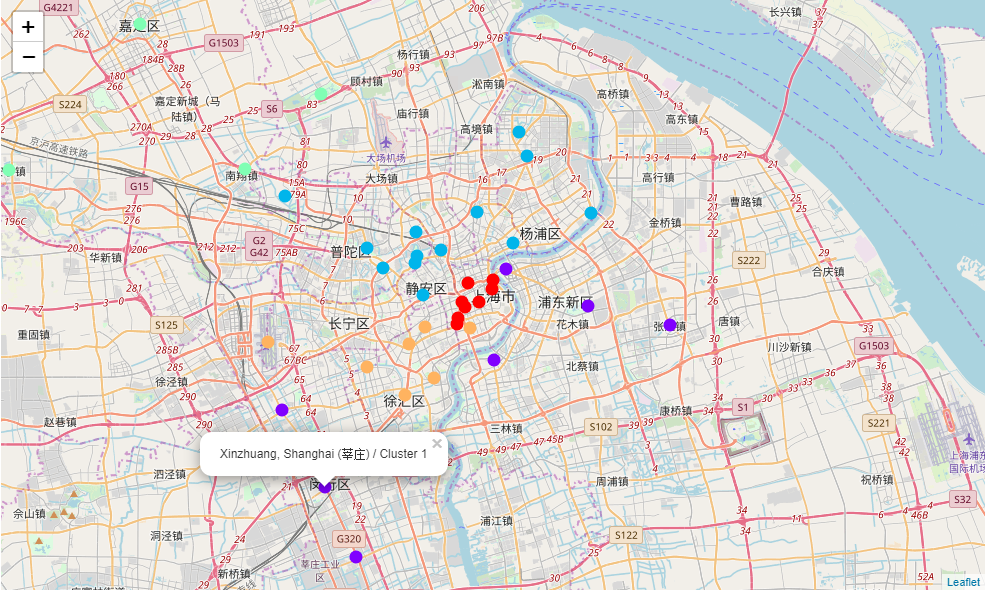
指南的最终目标是将全部47街区根据生活场所出现频率、各类经济数据进行聚类（clustering），相似的街区会被分为一类。

首先，绘制Elbow Curve，确定理想的聚类为5类。





接着，使用k-means clustering（此处k=5），依据生活场所出现频率、各类经济数据，对上海的47个街区进行分类。分类结果如下：



结果并不出人意料。地图中一共有五种颜色，代表五大类。居于最核心的红色类（黄浦的街区），往外依次是蓝色类（杨浦、静安、虹口、长宁、普陀的街区）、橙色类（徐汇的街区）、紫色类（闵行、浦东的街区）、绿色类（郊区）。从内至外，聚类变得分散，发达程度也降低。

四、讨论与总结

根据上一张地图，