Capstone Project: Portland Trip Planner

Applied Data Science Capstone by IBM/Coursera

Introduction

Do you like to travel? I do. Many people love traveling, and it is even a part of their life. Exploring a new place on your vacation is an interesting and exciting experience, nonetheless, planning the trip before setting off is quite different. It could be very time consuming and even stressful when your life is busy. This project aims at making planning an itinerary easier and more efficient.

What we are going to do or to see probably is the first question popping up when planning a trip. We can easily collect traveling information online. But the process of converting the information you collected into a trip plan is not a simple task. It is very common that you get a long list of what-to-see with online searching. The list is so long that we usually need to refer to ratings to pick an attraction. But at the same time we might find many places have the same rating. How to pick one among the same ratings?

After spending time making a what-to-see list, it is not yet to be relaxed though. Without the conception of geographical location, the itinerary still can not be finalized. We need geographical information to arrange the route and streamline the journey. Again we have to take time doing online search to figure out the locations and then planning the route along these places.

Last but not least, where to eat is an essential part of the trip. Don’t assume you can find food service or the kind of food you need in each attraction. It would be very beneficial if we can know whether the food service is available around the attractions when planning an itinerary. However, information of to-see and to-eat usually are separated into different sections on websites. Again we need to spend additional time searching for eating.

This project will try to provide a solution for the problems mentioned above. We will build a map that shows attractions with distinctive popularity by color. Available restaurants around the attractions will also be marked on the map. In this way, the popularity of the attractions, geographical locations and food services are integrated on

a map. It will shorten the time to spend on planning the journey. People who love traveling definitely can benefit from the idea.

Data

I live in Portland, Oregon, US. This project will use Portland as a destination to execute the idea.

To collect the attractions of Portland, I will use 'Things to do' section in Google travel website.(https://www.google.com/travel/things-to-do) The attractions on Google travel website are very comprehensive. It can make you travel like a local if you have sufficient time. Besides, almost every attraction has a rating and the number of reviews with it. These two numbers will be used in K-means clustering method to generate a label of popularity.

After creating the list of attractions, we will use geopy to find latitude and longitude for each place. With geographical coordinates in hand, we can use FourSquare API to search for the nearby restaurants around them. We will extract the name, geographical coordinates and category of food for each restaurant. Later we can use the data to mark these restaurants on the map.

Methodology

Data acquisition and cleaning

First we acquire data of attractions from Google travel website. We use regular expression(regex) to clean and extract the information we need. We take names, ratings and the number of reviews out of the data. After reviewing the descriptive statistics of the number of reviews, we find the range is very wide with mean of 1721 and standard of 3336. We decide to drop off the places having less than 173 reviews, the first percentile, which means these places gain limited attention. On the right you can find the descriptive information about the number of reviews.

	rating	reviews
count	93.000000	93.000000
mean	4.532258	1721.946237
std	0.292352	3336.139947
min	2.800000	1.000000
25%	4.400000	173.000000
50%	4.600000	598.000000
75%	4.700000	1650.000000
max	5.000000	24920.000000

With the attractions name in hand, we use geopy to search geographical coordinates for each place. Seven of the places fail to get geographical coordinates by geoby. We remove these seven places from the list. We complete the data we need for attractions and create a dataframe for it. There are totally 62 rows in the dataframe. Below you can find the first ten rows of the dataframe.

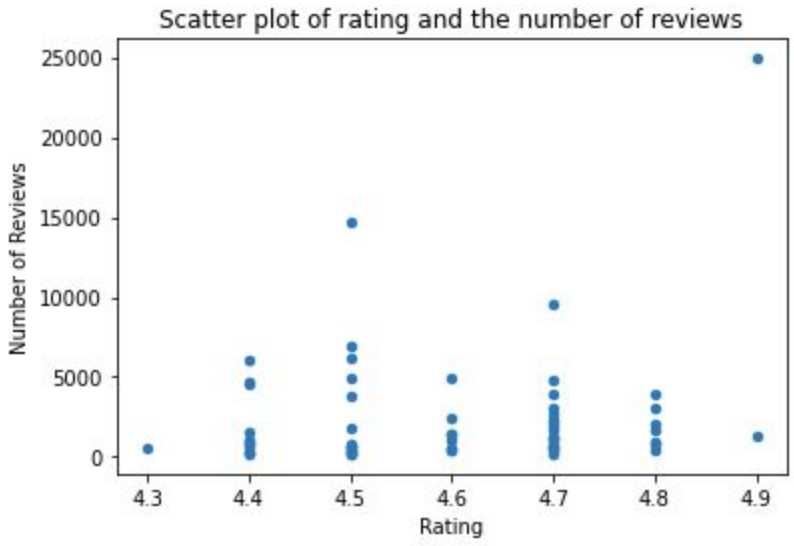
	place	rating	reviews	latitude	longitude
0	Pittock Mansion	4.6	4965	45.5252	-122.716
1	Portland Japanese Garden	4.5	4967	45.5187	-122.708
2	Lan Su Chinese Garden	4.6	2488	45.5257	-122.673
3	OMSI	4.5	6932	45.5083	-122.666
4	Oregon Zoo	4.5	14743	45.5098	-122.713
5	International Rose Test Garden	4.7	4753	45.5191	-122.705
6	Portland Art Museum	4.7	3964	45.5162	-122.684
7	Washington Park	4.7	9618	45.5155	-122.706
8	Pioneer Courthouse Square	4.4	6062	45.5189	-122.679
9	Governor Tom McCall Waterfront	4.5	6210	45.5204	-122.67

Next I use the latitude and longitude of each attraction to search FourSquare for the nearby restaurants that are within 1000 meters. There are a total of 665 restaurants extracted through the FourSquare API. Please find the first ten rows of the restaurant data in dataframe below.

	name	categories	lat	lng	id
0	Khao Soy Thai Restaurant	Thai Restaurant	45.524726	-122.699428	55c6d381498e9a5c8f034964
1	수라 Korean Restaurant	Korean Restaurant	45.525406	-122.698562	510dd94be4b01d20ece508a7
2	Thai Bloom! Portland: Restaurant and Catering	Restaurant	45.525288	-122.698662	5be4088b625a66002cf315fe
3	August Moon Chinese	Chinese Restaurant	45.525643	-122.698497	4aa30efaf964a5202c4320e3
4	Hobo's Restaurant & Lounge	Bar	45.524300	-122.673258	40b13b00f964a5208bf51ee3
5	Chen's Good Taste Restaurant	Chinese Restaurant	45.523531	-122.674297	4ad781f6f964a520950b21e3
6	Al-Amir Lebanese Restaurant	Middle Eastern Restaurant	45.520157	-122.673887	41e46880f964a520d51e1fe3
7	Scooter McQuade's Restaurant & Bar	Bar	45.522329	-122.684916	40b13b00f964a520f1f51ee3
8	Wilf's Restaurant	American Restaurant	45.528633	-122.676461	4bf19a6c189f0f47d875b762
9	Kells Irish Restaurant & Pub	Irish Pub	45.521579	-122.672568	4a915f09f964a520011a20e3

Analysis

From the scatter plot on the right, you can find the range of ratings for attractions are narrow, between 4.3 and 4.9. A lot of places have the same rating, but the number of reviews they got may have a big discrepancy. Take rating of 4.9 as example, two attractions have the same rating, but one has almost 25000 reviews and another has less than 2000 reviews. I sliced out the data of these two places as below.



	place	rating	reviews	latitude	longitude
26	Powell's City of Books	4.9	24920	45.5233	-122.682
50	Portland Oregon Temple	4.9	1360	45.4254	-122.742

K-Means Cluster

Therefore, solely checking the rating of an attraction is not enough to decide its popularity. I will use K-Means cluster to include both rating and the number of reviews to generate a label of popularity.

Result

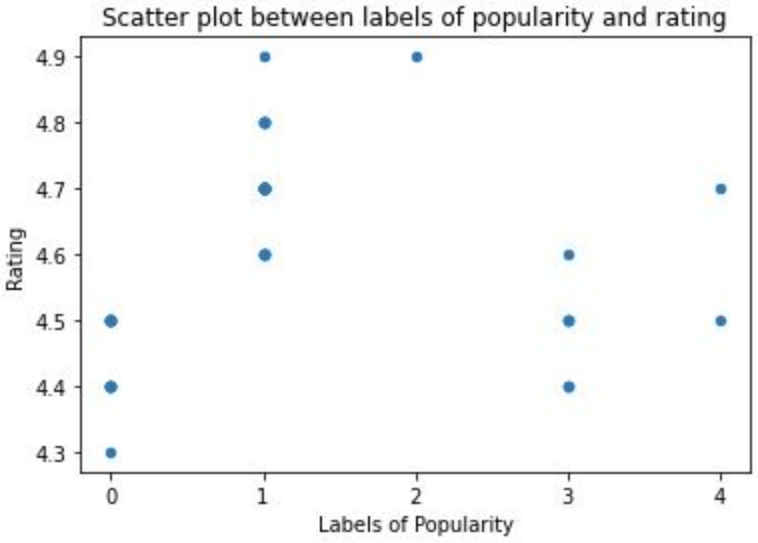
Labels of Popularity

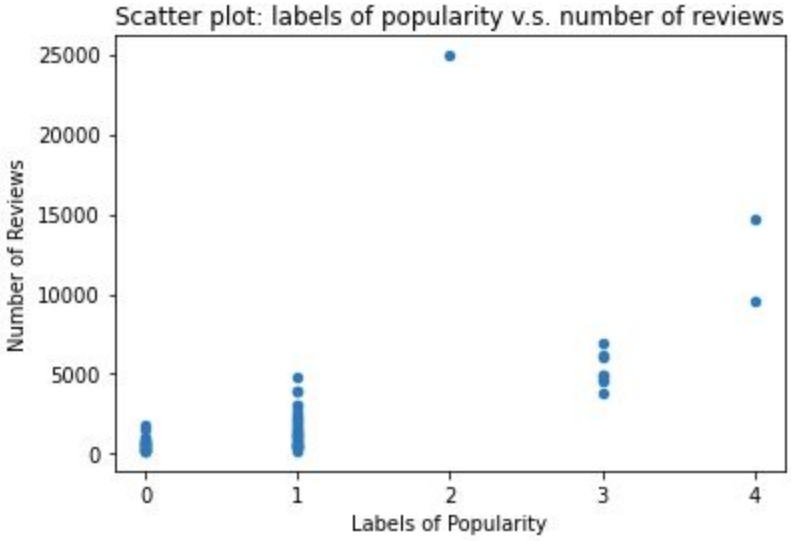
The label of popularity generated from the K-Means cluster is added into the dataframe at the last column as below.

	place	rating	reviews	latitude	longitude	popularity
0	Pittock Mansion	4.6	4965	45.5252	-122.716	3
1	Portland Japanese Garden	4.5	4967	45.5187	-122.708	3
2	Lan Su Chinese Garden	4.6	2488	45.5257	-122.673	1
3	OMSI	4.5	6932	45.5083	-122.666	3
4	Oregon Zoo	4.5	14743	45.5098	-122.713	4
5	International Rose Test Garden	4.7	4753	45.5191	-122.705	1
6	Portland Art Museum	4.7	3964	45.5162	-122.684	1
7	Washington Park	4.7	9618	45.5155	-122.706	4
8	Pioneer Courthouse Square	4.4	6062	45.5189	-122.679	3
9	Governor Tom McCall Waterfront	4.5	6210	45.5204	-122.67	3

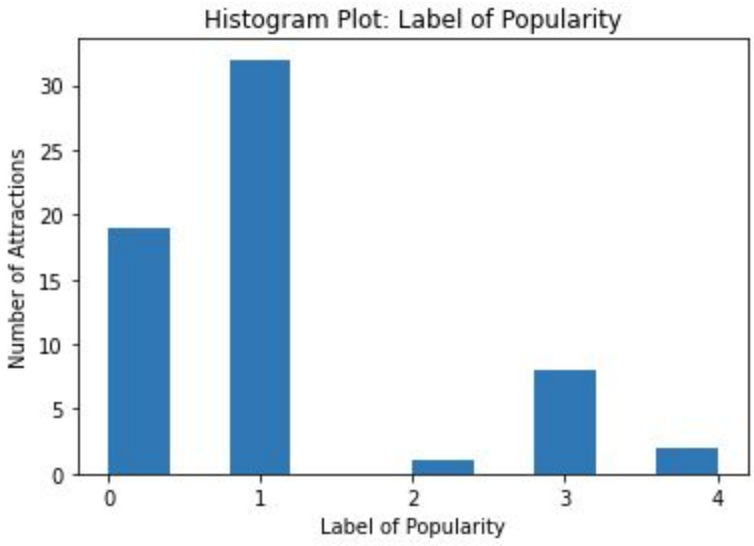
Let's see how the labels of popularity mean by scatter plots. The first scatter plot on the right is labels of popularity v.s. Ratings. The second one is labels of popularity v.s. The number of reviews. From these two plots, we can draw the conclusion:

Label 0: low rating and low number of reviews
Label 1: higher rating but lower number of reviews
Label 2: very high rating and very high number of reviews
Label 3: lower rating but higher number of reviews
Label 4: average rating but very high number of reviews



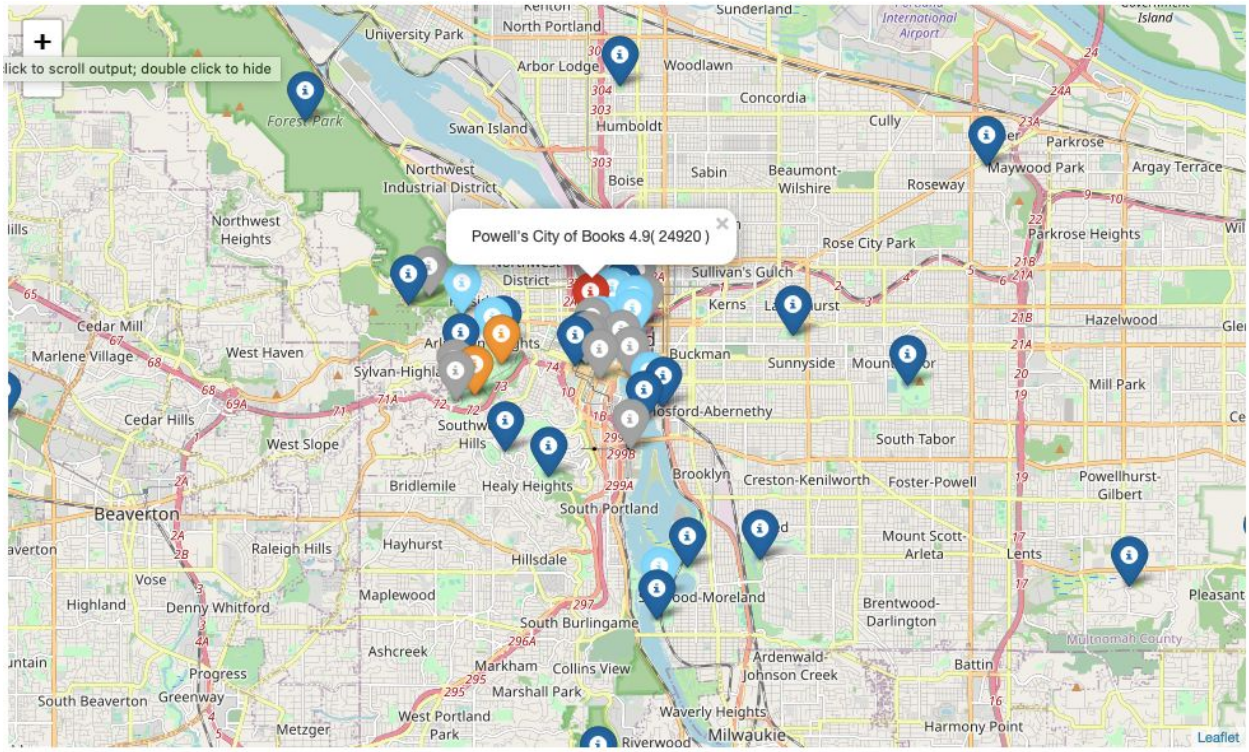


The histogram plot of labels of popularity is as below. More than 30 attractions fall in label 1 which is highest. The second high is label 0, a little lower than 20.

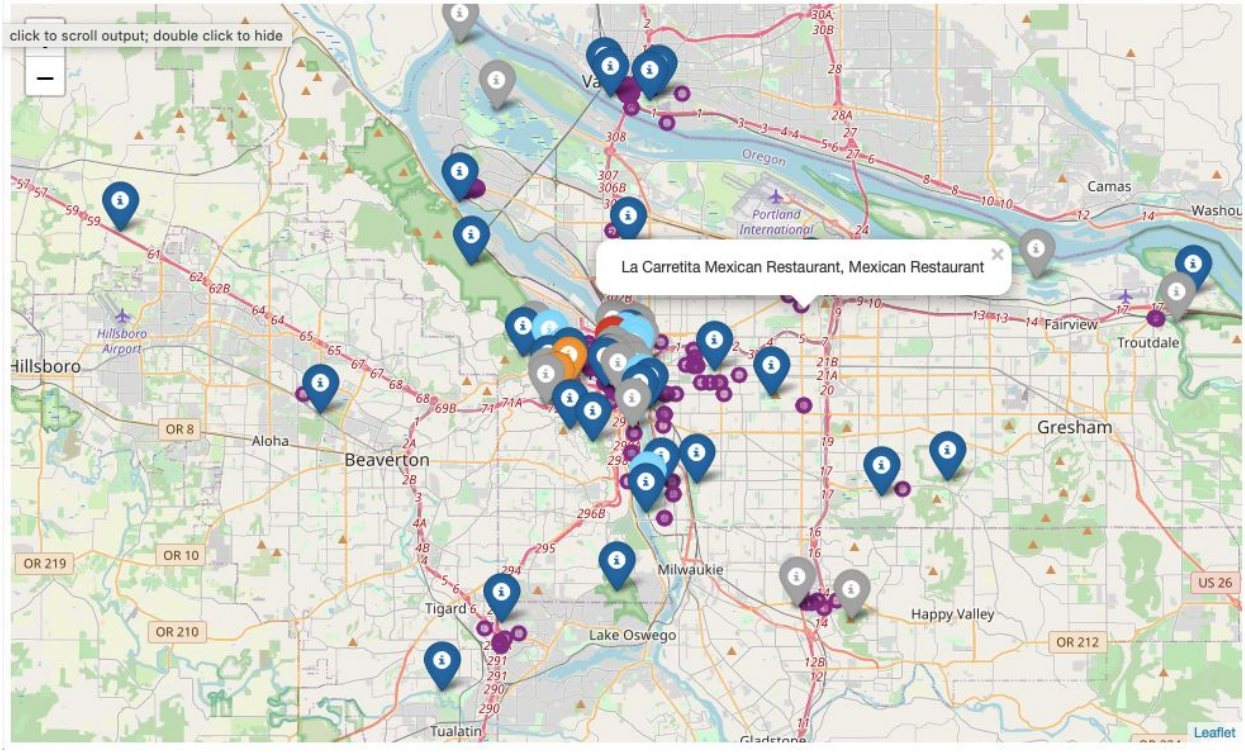


Folium Map

This map shows only the attractions. The markers of attractions are in different colors for distinguishing popularity. **Label 0 is light gray. Label 1 is dark blue. Label 2 is red. Label 3 is light blue. label 4 is orange.** When you click the marker, it would show you the name, rating and the number of reviews.



The purple circles in the second map are restaurants. When you click the restaurant markers, it will show you the name and the type of food.



Discussion

According to the labels of popularity generated by K-Means cluster, label 2 is the most popular attraction. It gets the highest rating and highest number of reviews. We have only one attraction belonging to this label, which is Powell's city of books. Powell's city of books is Portland's icon and it is really one of a kind. I think it deserves the highest mark.

Attractions of label 0 are low rating and low number of reviews. There are 19 attractions belonging to this label. If you want to skip some of the attractions, obviously label 0 is the first group to be removed.

Label 4 has good ratings and high number of reviews, I would consider this group as the second best attraction to visit. There are two attractions belonging to label 4.

I think it is hard to choose between label 1 and label 3. Label 1 has higher rating but lower number of reviews while label 3 has lower rating but higher number of reviews. I

think attractions in these two groups are not dominating places when planning an itinerary. You can add attractions of label 1 and 3 into your itinerary when they are nearby or close to your route.

Conclusion

I think labels of popularity generated by K-Means cluster do help to pick attractions from a long list. Besides it identifies an outstanding attraction like Powell's city of books, it makes us confidently remove the attractions belonging to label 0. By skipping label 0, the attractions list is shortened by 30%, from 62 attractions to 43 attractions. It definitely is easier to make choices from a shorter list.

A map showing both attractions and nearby restaurants indeed solves three essential questions when planning a trip. To-see, to-eat and geographical locations are integrated on a map. I think this is really neat!