Singapore Visitors and Expatriates Venue Recommendation

PURPOSE

This document provides the details of my final peer reviewed assignment for the IBM Data Science Professional Certificate program – Coursera Capstone.

INTRODUCTION

Singapore is a small country and one of the most visited countries in Asia. There are a lot of websites where travelers can check and retrieve recommendations of places to stay or visit. However, most of these websites provides recommendation simply based on usual tourist attractions or key residential areas that are mostly expensive or already known for travelers based on certain keywords like "Hotel", or "Backpackers" etc. The intention on this project is to collect and provide a data driven recommendation that can supplement the recommendation with statistical data. This will also be utilizing data retrieved from Singapore open data sources and FourSquare API venue recommendations.

The sample recommender in this notebook will provide the following use case scenario:

- A person planning to visit Singapore as a Tourist or an Expat and looking for a reasonable accommodation.
- The user wants to receive venue recommendation where he can stay or rent an HDB apartment with close proximity to places of interest or search category option.
- The recommendation should not only present the most viable option, but also present a comparison table of all possible town venues.

For this demonstration, this notebook will make use of the following data:

- Singapore Median Rental Prices by town.
- Popular Food venues in the vicinity. (Sample category selection)

Note: While this demo makes use of Food Venue Category, Other possible categories can also be used for the same implementation such as checking categories like:

- Outdoors and Recreation
- Nightlife
- Nearby Schools, etc.

I will limit the scope of this search as FourSquare API only allows 50 free venue query limit per day when using a free user access.

DATA ACQUISITION

This demonstration will make use of the following data sources:

Singapore Towns and median residential rental prices.

Data will retrieved from Singapore open dataset from median rent by town and flattype from https://data.gov.sg website.

The original data source contains median rental prices of Singapore HDB units from 2005 up to 2nd quarter of 2018. I will retrieve rental the most recent recorded rental prices from this data source (Q2 2018) being the most relevant price available at this time. For this demonstration, I will simplify the analysis by using the average rental prices of all available flat type.

Singapore Towns location data retrieved using Google maps API.

Data coordinates of Town Venues will be retrieved using google API. I also make use of MRT stations coordinate as a more important center of for all towns included in venue recommendations.

Singapore Top Venue Recommendations from FourSquare API

(FourSquare website: www.foursquare.com)

I will be using the FourSquare API to explore neighborhoods in selected towns in Singapore. The Foursquare explore function will be used to get the most common venue categories in each neighborhood, and then use this feature to group the neighborhoods into clusters. The following information are retrieved on the first query:

- Venue ID
- Venue Name
- Coordinates : Latitude and Longitude
- Category Name

Another venue query will be performed to retrieve venue ratings for each location. Note that rating information is a paid service from FourSquare and we are limited to only 50 queries per day. With this constraint, we limit the category analysis with only one type for this demo. I will try to retrieve as many ratings as possible for each retrieved venue ID.

METHODOLOGY

Singapore Towns List with median residential rental prices.

The source data contains median rental prices of Singapore HDB units from 2005 up to 2nd quarter of 2018. I will retrive the most recent recorded rental prices from this data source (Q2 2018) being the most relevant price available at this time. For this demonstration, I will simplify the analysis by using the average rental prices of all available flat type.

Data Cleanup and re-grouping. The retrieved table contains some un-wanted entries and needs some cleanup.

The following tasks will be performed:

- Drop/ignore cells with missing data.
- Use most current data record.
- Fix data types.

Post Processed Singapore towns list with and median residential rental prices

	Town	m
0	ANG MO KIO	2033.333333
1	BEDOK	2087.500000
2	BISHAN	2233.333333
3	BUKIT BATOK	1962.500000
4	BUKIT MERAH	2162.500000
5	BUKIT PANJANG	1737.500000
6	CENTRAL	2450.000000
7	CHOA CHU KANG	1933.333333
8	CLEMENTI	2263.333333

	Town	m
9	GEYLANG	2166.666667
10	HOUGANG	1962.500000
11	JURONG EAST	2150.000000
12	JURONG WEST	1975.000000
13	KALLANG/WHAMPOA	2300.000000
14	MARINE PARADE	1950.000000
15	PASIR RIS	2066.666667
16	PUNGGOL	1825.000000
17	QUEENSTOWN	2162.500000
18	SEMBAWANG	1883.333333
19	SENGKANG	1900.000000
20	SERANGOON	2187.500000
21	TAMPINES	2075.000000
22	TOA PAYOH	2210.000000
23	WOODLANDS	1762.500000
24	YISHUN	1900.000000

- Adding geographical coordinates of each town location.

Segmenting and Clustering Towns in Singapore

Retrieving FourSquare Places of interest.

Using the Foursquare API, the **explore** API function was be used to get the most common venue categories in each neighborhood, and then used this feature to group the neighborhoods into clusters. The *k*-means clustering algorithm was used for the analysis. Fnally, the Folium library is used to visualize the recommended neighborhoods and their emerging clusters.

In the ipynb notebook, the function **getNearbyVenues** extracts the following information for the dataframe it generates:

- Venue ID
- Venue Name
- Coordinates : Latitude and Longitude
- Category Name

The function **getVenuesByCategory** performs the following:

1. **category** based venue search to simulate user venue searches based on certain places of interest. This search extracts the following information:

- Venue ID
- Venue Name
- Coordinates : Latitude and Longitude
- Category Name

2. For each retrieved **venueID**, retrive the venues category rating.

The generated data frame in the second function contains the following column:

Search Venues with recommendations on : Food Venues (Restaurants,Fastfoods, etc.)

To demonstrate user selection of places of interest, We will use this Food Venues category in our further analysis.

- This Foursquare search is expected to collect venues in the following category:
- category
- Food Courts
- Coffee Shops
- Restaurants
- Cafés
- Other food venues

List of my FourSquare Data collection saved in Github can be found in the following location:

- https://github.com/crismag/Coursera_Capstone/tree/master/saved_data

I used the FourSquare API to retrieve venue scores of locations. Note that there is max query limit of 50 in FourSquare API for free subscription. So use or query carefully.

Data cleanup uneeded entries

- Eliminate possible venue duplicates.
- Improve the quality of our venue selection by removing venues with no ratings or 0.0

Column Name	Description
Town	Town Name
Town Latitude	Towns MRT station Latitude
Town Longitude	Town MRT station Latitude
VenueID	FourSquare Venue ID
VenueName	Venue Name
score	FourSquare Venue user rating
category	Category group name
catID	Category ID
latitude	Venue Location - latitude
longitude	Venue Location - longitude

Results: Venue Count Per Town

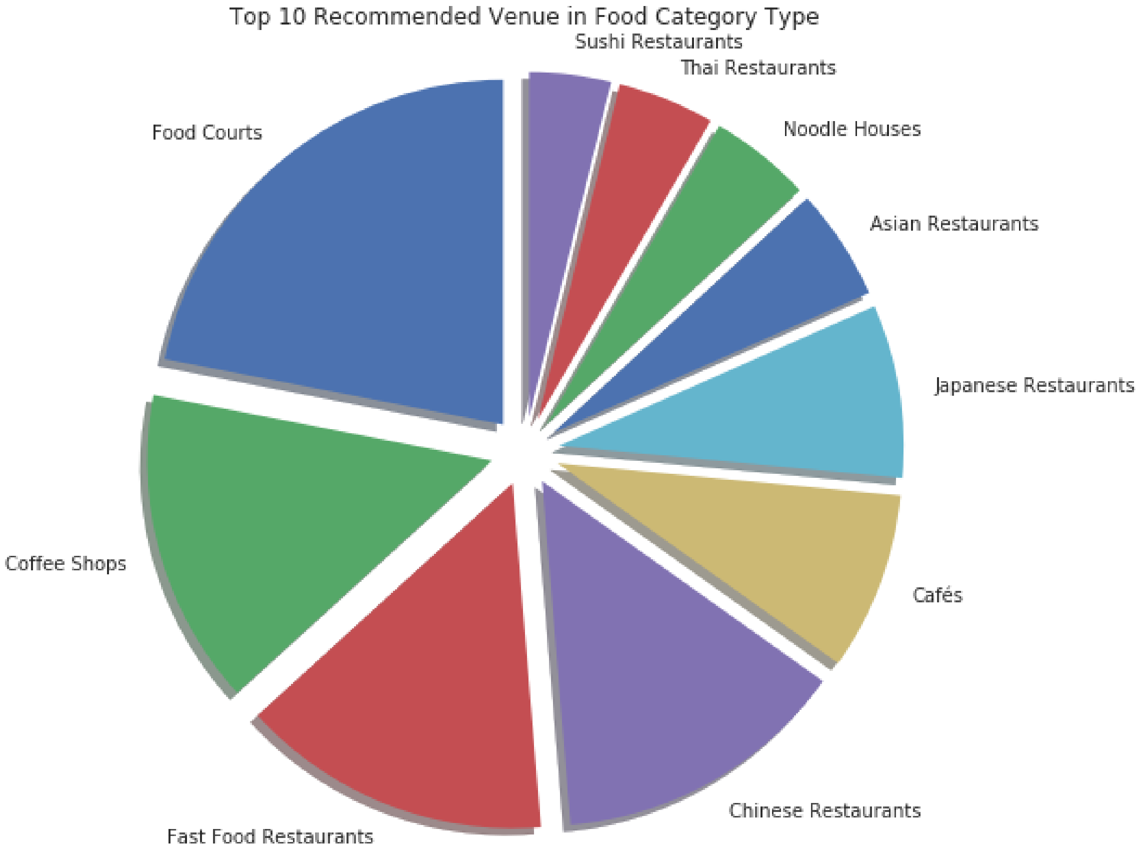
	Town Latitude	Town Longitude	VenueID	VenueName	score	category	cat
Town							
ANG MO KIO	34	34	34	34	34	34	34
BEDOK	29	29	29	29	29	29	29
BISHAN	36	36	36	36	36	36	36
BUKIT BATOK	22	22	22	22	22	22	22
BUKIT MERAH	9	9	9	9	9	9	9
BUKIT PANJANG	15	15	15	15	15	15	15
CENTRAL	46	46	46	46	46	46	46
CHOA CHU KANG	27	27	27	27	27	27	27
CLEMENTI	34	34	34	34	34	34	34
GEYLANG	25	25	25	25	25	25	25
HOUGANG	26	26	26	26	26	26	26
JURONG EAST	39	39	39	39	39	39	39
JURONG WEST	31	31	31	31	31	31	31
KALLANG/WHAMPOA	15	15	15	15	15	15	15
MARINE PARADE	21	21	21	21	21	21	21
PASIR RIS	17	17	17	17	17	17	17

	Town Latitude	Town Longitude	VenueID	VenueName	score	category	cat
Town							
PUNGGOL	25	25	25	25	25	25	25
QUEENSTOWN	8	8	8	8	8	8	8
SEMBAWANG	18	18	18	18	18	18	18
SENGKANG	17	17	17	17	17	17	17
SERANGOON	42	42	42	42	42	42	42
TAMPINES	25	25	25	25	25	25	25
TOA PAYOH	34	34	34	34	34	34	34
WOODLANDS	31	31	31	31	31	31	31
YISHUN	18	18	18	18	18	18	18

RESULTS: How many unique categories can be curated from all the returned venues?

```
* There are 67 uniques categories.
```

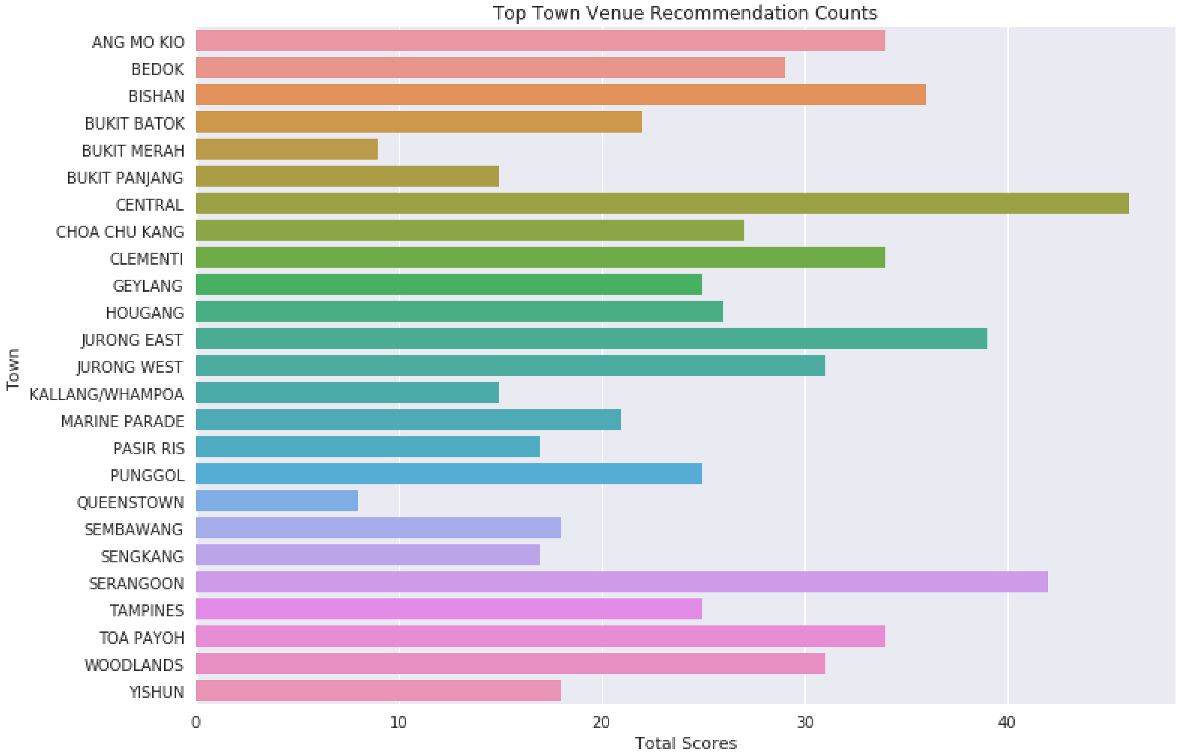
What are the top 20 most common venue types?

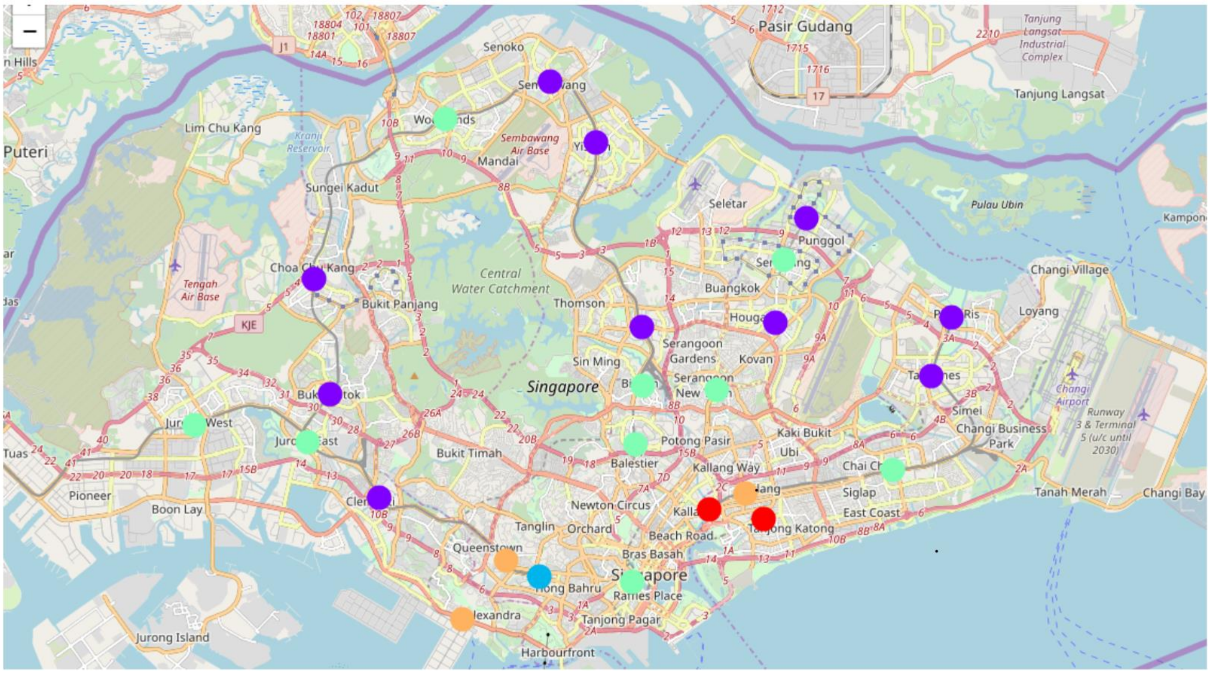


```
category
Food Courts              92
Coffee Shops             62
Fast Food Restaurants    61
Chinese Restaurants      58
Cafés                    35
Japanese Restaurants     34
Asian Restaurants        22
Noodle Houses            20
Thai Restaurants         19
Sushi Restaurants        16
Seafood Restaurants      15
Italian Restaurants      13
Sandwich Places          13
Indian Restaurants       11
```

```
    Bubble Tea Shops          11

    American Restaurants       8

    Burger Joints              8

    Dim Sum Restaurants        8

    Dessert Shops              8

    Fried Chicken Joints       7
```

What are the top 20 venues given with highest score rating?





Discussion and Conclusion

On this notebook, Analysis of best town venue recommendations based on Food venue category has been presented. Recommendations based on other user searches like available outdoor and recreation areas are also available. As singapore is a small country with a whole host of interesting venues scattered around the town, the information extracted in this notebook present on the town areas, will be a good supplement to web based recommendations for visitors to find out nearby venues of interest and be a useful aid in deciding a place to stay or where to go during their visits.

Using Foursquare API, we have collected a good amount of venue recommnedations in Singapore Towns. Sourcing from the venue recommendations from FourSquare has its limitation, The list of venues is not exhaustive list of all the available venues is the area. Furthermore, not all the venues found in the the area has a stored ratings. For this reason, the number of analyzed venues are only about 50% of all the available venues initially collected. The results therefore may significantly change, when more information are collected on those with missing data.

The generated clusters from our results shows that there are very good and interesting places located in areas where the median rents are cheaper. This kind of results may be very interesting for travelers who are also on budget constraints. Our results also yielded some interesting findings. For instance, The initial assumption among websites providing recommendations is that the Central Area that have the highest median rent also have better food venues. The results however shows that while Marine Parade, a cheaper location has better rated food courts. Result shows that most popular food venue among Singaporeans, residents and visitors are **Food Courts, Coffee Shops and Fast Food Restaurants**. The highest rated Food Courts are located in **Marine Parade**, and in **Central Area**.

I will be providing a other supplementary Inferential Statics in the future about on these data collected and also update in a new notebook using other categories. For now, this completes the requirements for this task.