Final Report

IBM Data Science Capstone Project
Analysing the best location to open a cafe in Bali

MICKAEL GRATIA
JULY 2020

Introduction

Description of the problem

Starting a business in Bali combines two of every entrepreneur's greatest desires: living in paradise while making a steady profit. The number of businesses springing up in Bali over the last few years has turned this island paradise into something more lucrative than just another Southeast Asian holiday destination.

Starting a business in Bali by opening a coffee shop is a good choice because of its simplicity and the handsome profit that will come with it. Plus, who doesn't like coffee? You can do this in major parts of the island. Good news is, opening a coffee shop in Bali will not require a huge start-up capital.

We know many foreigners planning on opening a coffee shop in Bali have no idea how and where to start at first. We will skip the "how" part for thés Capstone project and I propose we focus on the "where".

To help entrepreneurs in their choice of location, we will leverage the Foursquare location data and define which district is worth considering. The main indicator will be competition, going to a place where the market is not already submerged with coffee shops.

Data and its use

Bali island is divided into several districts. Bali can be described as a mix from the bustling streets and white-sand beaches of Kuta, Legian and Seminyak to the upscale resorts of Nusa Dua, the cultural charms of Ubud and the cool highlands of Bedugul. Each district has specific attractions, some are also being developed. So for this project we will need the following data:

- **Bali data on its districts**
  - Data source: Wikipedia https://en.wikipedia.org/wiki/Category:Districts_of_Bali
  - Description: this data set contains the list of districts composing Bali. It will be used to explore the various businesses

- **Coordinates:** Latitude and longitude of each district.
  - Data source: Python geocoded package
  - Description: this data set will be required to be able to plot each district onto a map

- **Venue data**
  - Data source: Foursquare API
  - Description: this data set will be obtained via an API to collect all the venues in each district. It will also be filtered on coffee shops.

Methodology

In this section we will discuss and describe any exploratory data analysis, any inferential statistical testing, and what machine learnings used.

We first start by collecting the Bali districts data from the Wikipedia page https://en.wikipedia.org/wiki/Category:Districts_of_Bali

The second step consists in doing web scraping using Python requests and beautiful soup packages to extract the list of districts data.

The list collected is a set of names as presented below

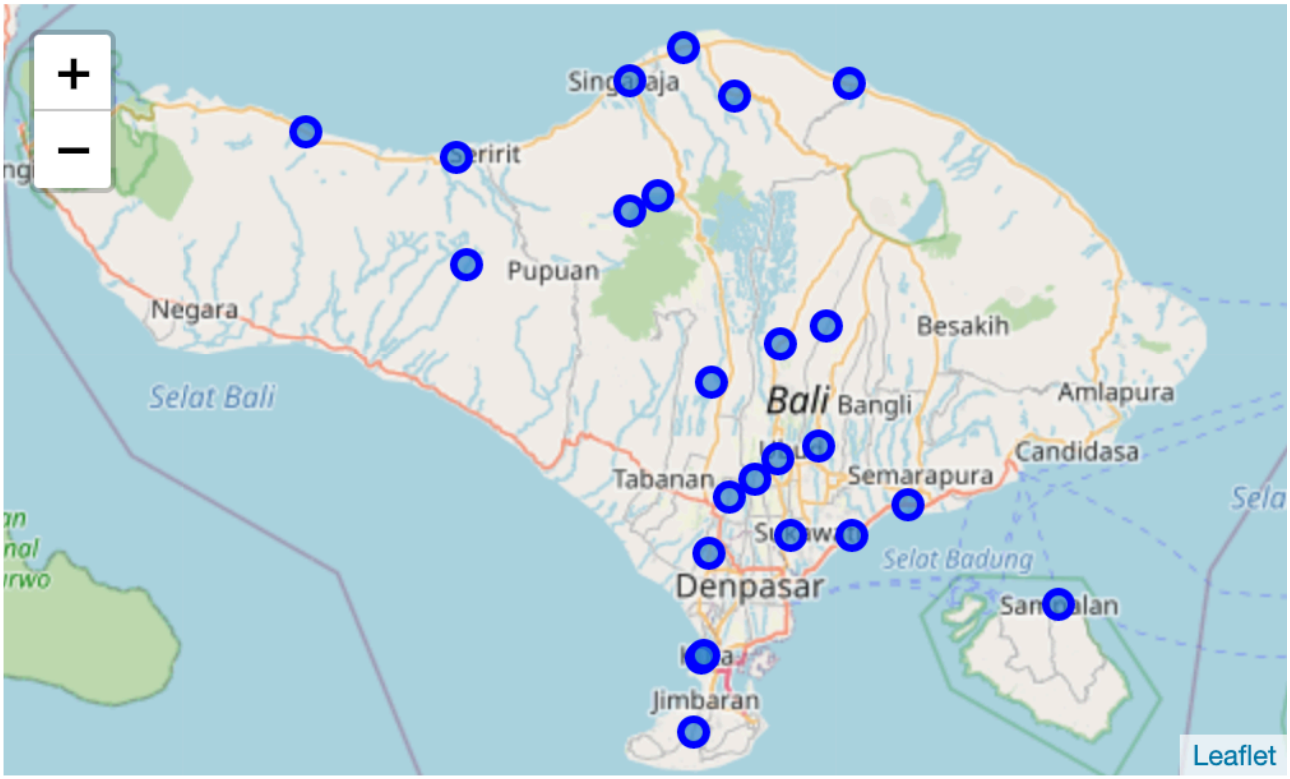
	Neighborhood
19	Sukasada, Buleleng
20	Sukawati
21	Tegallalang
22	Tejakula, Buleleng
23	Ubud District

The second step consists in getting the coordinates of each district so we can use them in the Foursquare API. This is achieved using the Geocoder package. It converts addresses into geographical coordinates being in the form of latitude and longitude.

A panda dataframe is then created by merging the list of districts to the coordinates.

	Neighborhood	Latitude	Longitude
0	List of districts of Bali	-8.72969	115.16812
1	Abiansemal District	-8.54097	115.22325
2	Banjar, Buleleng	-8.25522	115.09030
3	Banjarangkan	-8.56762	115.39053
4	Blahbatuh	-8.59972	115.32788
5	Buleleng, Bali	-8.11591	115.09037
6	Busung Biu, Buleleng	-8.31146	114.91358
7	Gerokgak, Buleleng	-8.17042	114.74011
8	Tampaksiring	-8.50366	115.29287

The below map is created:



In order to define existing competition, we use the Foursquare API to get the top 100 venues that are within a radius of 10km. We target a large radius due to the distance sometimes between key districts.

We use the registered Foursquare developer account as learned in the course and make an API call to collect the venue data under the form of a JSON file. We do consider in the extraction the venue name, category, latitude and longitude.

Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
-8.72969	115.16812	Odysseys Surf School	-8.720849	115.169901	Surf Spot
-8.72969	115.16812	Young Spa	-8.722417	115.175280	Spa
-8.72969	115.16812	Cara Cara Inn	-8.722761	115.173320	Hotel
-8.72969	115.16812	Sheraton Bali Kuta Resort	-8.717966	115.169126	Hotel
-8.72969	115.16812	Discovery Kartika Plaza Hotel	-8.729493	115.166609	Hotel

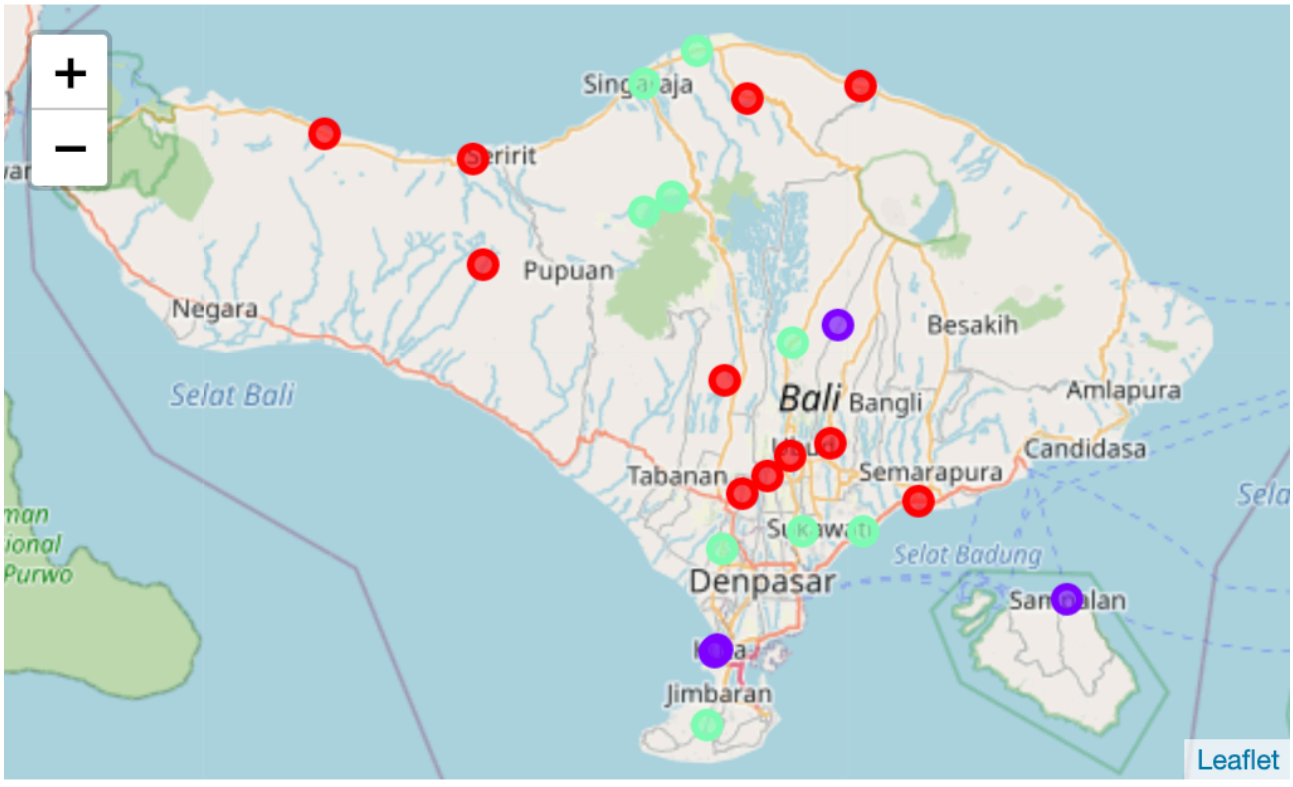
We then prepare the data by filtering on the business we are focusing on for this case: Coffee Shop. Grouping them by location we obtain a ratio / occupancy rate that will define whether the competition is high or low

	Neighborhoods	Coffee Shop
0	Abiansemal District	0.010000
1	Banjar, Buleleng	0.020408
2	Banjarangkan	0.000000
3	Blahbatuh	0.030000
4	Buleleng, Bali	0.021277

Finally our point is to group the locations. Therefore we perform a clustering on the data using k-mean. This algorithm identifies the k number of centroids and then allocates every data point to the nearest cluster. It is an unsupervised machine learning algorithm and is very well suited to solve cases similar to the one presented here.

It is decided to divide the pool into 3 clusters. The outcome shall help us identify which neighbourhood has higher number of coffee shops and therefore support the entrepreneur in her/his decision with regards to location

Results



In the above map:

- red cluster - Low competition
- green cluster - Medium competition
- purple cluster - High competition

Discussion

We can observe a high concentration of coffee shop in well known places such as Kuta or the rice fields north of Ubud. Interesting enough the island of Nusa Penida has seen a huge business increase over the past 2 to 3 years and indeed the number of coffee shops is now quite important.

Green markers refer to either expanding Denpasar area due to the new highway opened and that links the airport to the north part of the main city of Denpasar but also to the fact people are now taking houses more towards the east part of the island.

To note the north with black sand beaches of Lovina gets a medium concentration of coffee shops

Less touristic areas marked in red have a low concentration of coffee shops which makes sense. Also access to some of those location can be a bit time consuming. However interesting enough is the area of Tabanan which is marked as low in terms of competition and is a good tourist area, well known for a Unesco rice fields, temple and strawberries producer area.

Conclusion

For this project we would therefore recommend the entrepreneur to go for Tabanan if the focus is on tourism. North of Denpasar otherwise due to the medium competition.