'Coffee in Cape Town'

IBM Data Science Professional Certificate
Capstone Project
(a.k.a. 'Battle of the Neighborhoods')

by Laura Tissing
January 2020

Table of Contents

Introduction

A client wants to explore the possibilities of opening a Coffee Shop in Cape Town.
Cape Town has many suburbs, spread around the central area within a distance of approximately 50-60 kilometers. Like many other cities in the world, Cape Town's suburbs are becoming more and more attractive to South Africans and tourists, and a lot of development projects are taking place. The client wants to open the Coffee Shop in an area where citizens and tourists are mingling, away from but within a reasonable distance of the busy city centre.

Target Audience
Apart from this specific client, the studied factors and collected information can serve as a starting point to make recommendations to other clients who want to start a business in Cape Town.

Task Description
A Coffee Shop is a business that sells coffee, tea, and some food items to customers. A business like a Coffee Shop normally has to have fast and efficient operations in place since part of the customers will order on the go. On the other hand, a Coffee Shop needs to be a place where people can sit for a moment, meet up with friends or read a book and escape the daily stresses of life. Looking at the suburbs of Cape Town, what is a recommendable location to open such a Coffee Shop?

The location of a Coffee Shop is an important part of its success, therefore various factors need to be studied before a recommendation can be done. These factors can include:

- location of competitors
- surrounding venues
- surrounding businesses
- surrounding attractions
- distance to the city centre
- accessibility to and from the city centre
- population and demographics
- tourism
- rental and/or m² prices for commercial property

Data Description

Data sources

The following data will be used to address the task description:
Website on South Africa's postal codes (www.postoffice.co.za/Tools/postalcodes.html)
Geographical coordinates from Google Maps
Geolocator geographical coordinates
Foursquare API with data on Venue names, locations, id's and categories
Wikipedia
Websites on Cape Town's suburbs (SA-venues.com, blog.rawson.co.za/spotlight-on-maitland-cape-town)

The data will be used to

1. Divide the broad area of Cape Town's suburbs into reasonably sized areas based on Street Codes (equal to Postal Codes)
2. Retrieve information on venues (category, location, total number, etc) in each Street Code area
3. Cluster the Street Codes based on the occurrence of Coffee Shops
4. Analyze the venues, clusters, Street Codes and suburbs to be able to make a recommendation

Methodology

Research Methods

- Visualization with maps, WordClouds and bar charts
- One-hot encoding
- K-means clustering
- Webscraping

Part 1 – Data collection and visualization

I start by searching the web on information about Cape Town and its suburbs. I use the information on the website of the Postal services in South Africa since the suburbs are provided with Street Codes, which have the same function as Post Codes. This way, I can merge the suburbs into borough-like areas.
I download the information on suburbs and Street Codes and put them in a dataframe. I have to clean the dataframe since it contains all Street Codes in South Africa and NaN values.
After selecting the Street Codes of Cape Town only, I get the geographical coordinates of every Street Code with the help of Google Maps and Geolocator.

I combine all the information and end up with a dataframe with information on Street Codes, suburbs in these Street Codes and the geographical coordinates of the Street Codes:

	STR-CODE	SUBURB	Latitude	Longitude
0	7100	EERSTE RIVER, EERSTERIVIER	-33.987809	18.667149
1	7405	MAITLAND, PAARDEN ISLAND, PAARDENEILAND, PINEL...	-33.926604	18.498313
2	7441	BOTHASIG, EDGEMEAD, MELKBOS, MELKBOSSTRAND, MI...	-33.844994	18.522595
3	7455	LANGA, LANGA ZONE 1, LANGA SONE 1, LANGA ZONE ...	-33.944727	18.532317
4	7460	MONTE VISTA	-33.908873	18.550382
5	7490	MATROOSFONTEIN	-33.955570	18.587988

For visualization of the dataset, I use Folium to superimpose markers with the Street Codes on StreetMaps:

Part 2 – using Foursquare API to retrieve venue information

In this part of the project, I use Foursquare API to gain access to information on venues in the different Street Code areas. Through Foursquare I retrieve information on the venues like geographical coordinates, category, name and id.
I retrieve the venue information from every Street Code within a 2km radius and put that in a dataframe:

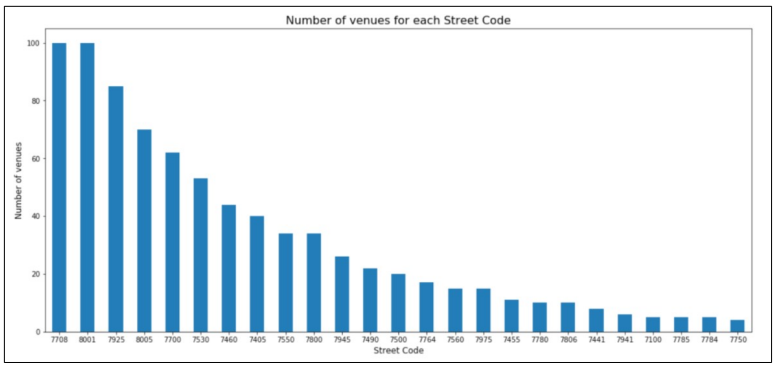
	Street Code	Street Code lat	Street Code lng	Venue	Venue lat	Venue lng	Category
0	7100	-33.987809	18.667149	Airport Mall	-33.994398	18.668121	Arts & Crafts Store
1	7100	-33.987809	18.667149	Engen Hindle Road Service Station	-33.979858	18.657894	Gas Station
2	7100	-33.987809	18.667149	Shoprite	-33.986895	18.680509	Supermarket
3	7100	-33.987809	18.667149	Allrich Trading (Pty) Ltd	-33.984661	18.654280	Print Shop
4	7100	-33.987809	18.667149	Wimpy	-33.979030	18.651890	Burger Joint
5	7100	-33.987809	18.667149	Shoprite Liquor Shop	-33.995441	18.650288	Liquor Store
6	7405	-33.926604	18.498313	Merrypak & Print	-33.928847	18.500087	Arts & Crafts Store
7	7405	-33.926604	18.498313	Magica Roma	-33.940149	18.497787	Italian Restaurant

It can be interesting to have a look at the top 10 venue categories around the Cape Town and its suburbs:

	Category	Count
0	Coffee Shop	49
1	Fast Food Restaurant	39
2	Hotel	36
3	Café	35
4	Grocery Store	34
5	Shopping Mall	30
6	Pizza Place	22
7	Gas Station	22
8	Steakhouse	20
9	Burger Joint	18

… it shows Coffee Shops are the most common venues.

Next, I explore the number of venues in each Street Code and visualize the result with a bar chart:



Part 3 – Clustering

To be able to use a model on the data, one-hot encoding is necessary to transform the categorical variables into numerical variables.
After the one-hot encoding, the dataframe shows a '1' if a venue category is present in a Street Code and a '0' it a venue category is not present. After grouping by Street Code and taking the

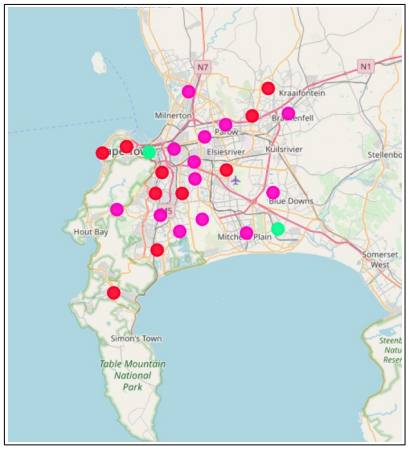
mean of the venue occurence I have a new dataframe showing the mean frequency occurence for every venue category in each Street Code.
Now I am ready to compile a new dataframe on the mean frequency occurence of Coffee Shops in every Street Code and use that for clustering:

	Street Code	Coffee Shop
0	7100	0.000000
1	7405	0.000000
2	7441	0.000000
3	7455	0.000000
4	7460	0.000000
5	7490	0.083333
6	7500	0.000000
7	7530	0.074074

I cluster the Street Codes using K-means. The objective of K-means is to group similar data points based on similarities and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (*k*) of clusters in a dataset and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

To determine the fixed number of clusters, I use the 'elbow method' which will give an optimal *k* in its output.

After each Street Code is allocated to a cluster I create a new dataframe and add the geographical coordinates columns. Now I can visualize the clusters:



pink = cluster 0
red = cluster 1
green = cluster 2

Part 4 – Recommendation

With the clusters in place, I can determine which Street Codes are suitable for recommendation (target Street Codes) within a radius of 15km from Cape Town's city centre.

Cluster 0 contains the Street Codes with no existence of Coffee Shops so I focus on those Street Codes:

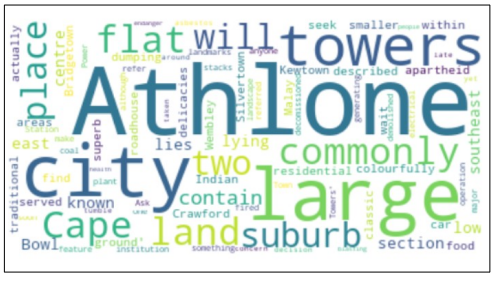
	Street Code	Cluster Labels	Coffee Shop	Latitude	Longitude	Distance(km)
0	7405	0	0.0	-33.926604	18.498313	7.542683
1	7441	0	0.0	-33.844994	18.522595	13.030715
2	7455	0	0.0	-33.944727	18.532317	10.949512
3	7460	0	0.0	-33.908873	18.550382	12.427999
4	7764	0	0.0	-33.968128	18.533450	11.906517
5	7800	0	0.0	-34.019579	18.474911	12.060983
6	7806	0	0.0	-34.012013	18.400462	10.072664

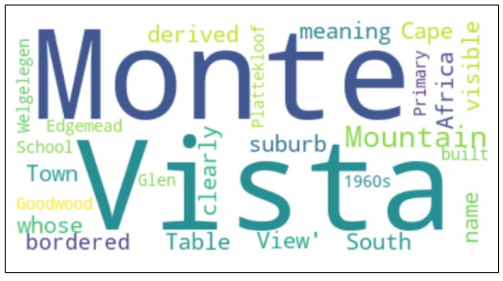
To get an idea of the target Street Code areas I merge the target Street Codes with their 10 most common venues:

THE FINAL DATAFRAME WITH STREET CODES THAT ARE SUITABLE FOR RECOMMENDATION (TARGET STREET CODES):

	Street Code	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	7405	-33.926604	18.498313	Furniture / Home Store	Grocery Store	Gym	Train Station	Burger Joint	Café	Fast Food Restaurant	Seafood Restaurant	Pharmacy	Playground
1	7441	-33.844994	18.522595	Pizza Place	Gay Bar	Shopping Mall	Convenience Store	Stadium	Garden Center	Chinese Restaurant	Fast Food Restaurant	Dance Studio	Warehouse Store
2	7455	-33.944727	18.532317	Convenience Store	Gas Station	Seafood Restaurant	Fast Food Restaurant	Farmers Market	Steakhouse	Business Service	Shopping Mall	Pizza Place	Design Studio
3	7460	-33.908873	18.550382	Fast Food Restaurant	Portuguese Restaurant	Gas Station	Steakhouse	Casino	Bowling Alley	Diner	Hotel	Seafood Restaurant	Skating Rink
4	7764	-33.968128	18.533450	Fast Food Restaurant	Pizza Place	Nightclub	Convenience Store	Burger Joint	Bakery	Stadium	Steakhouse	Indian Restaurant	Seafood Restaurant
5	7800	-34.019579	18.474911	Grocery Store	Fast Food Restaurant	Golf Course	Pharmacy	Pizza Place	Café	Thai Restaurant	Bakery	Performing Arts Venue	Portuguese Restaurant
6	7806	-34.012013	18.400462	Vineyard	Café	Flower Shop	Playground	Trail	Restaurant	French Restaurant	Tapas Restaurant	Wine Bar	Event Space

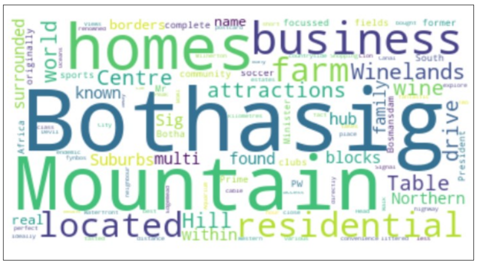
Finally, to expand my knowledge on the target Street Codes I explore every *first* suburb in these Street Codes using webscraping with BeautifulSoup and visualize the results through WordClouds:






township
Langa
hostels
first
centre
people
schools
accessible
kilometres
joins
birth
Set
Bunga
Fassie
Avenue
resting
infamous
oldest
superstar
Brenda
workers
place
queen
squatter
singles
outrageous
migrant
deprivation
apartheid
quarters
visiting
tourists
extent
taxi
one
list
rank
pop
take
tour
vast
five
era
area
high
percent
settlers
Hills
twenty
camp
endeavour
made


Constantia
popular
original
Alphen
vineyards
walk
Luxury
leafy
accommodation
South
Africa
operative
planning
style
holiday
verdant
guests
old
memorable
farmland
recently
lies
discerning
stately
area
homesteads
money
lined
description
long
find
still
driveways
streets
stretching
oak
suburb
outing
belts


Bothasig
Mountain
homes
business
located
residential
farm
Winelands
wine
attractions
Centre
world
surrounded
known
Suburbs
multi
found
blocks
hub
family
drive
Table
Northern
Hill
within
real
Sig
borders
complete
name
focussed
fields
former
sports
community
soccer

Results

Clustering the Street Codes using the mean frequency occurence of Coffee Shops gives me the following result:

- Cluster 0
  Street Codes with few, or no presence of Coffee Shops

- Cluster 1
  Street Codes with a high number of Coffee Shops

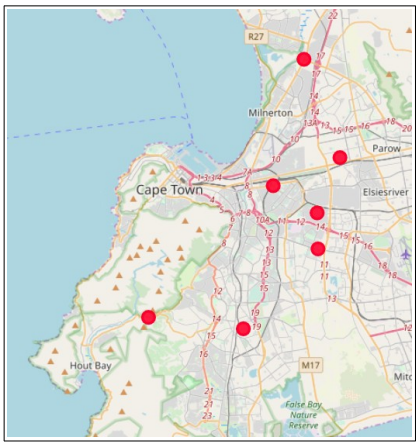
- Cluster 2
  Street Codes with a moderate number of Coffee Shops

I select the Street Codes of Cluster 0, with few or no presence of Coffee Shops, so there will be as little competition as possible. Furthermore, these Street Codes must be within a radius of 15km from the city centre to satisfy the desire of the client that the target location must lie within a reasonable distance from the city centre.

These criteria filter out the following target Street Codes:

	Street Code	Cluster Labels	Coffee Shop	Latitude	Longitude	Distance(km)
0	7405	0	0.0	-33.926604	18.498313	7.542683
1	7441	0	0.0	-33.844994	18.522595	13.030715
2	7455	0	0.0	-33.944727	18.532317	10.949512
3	7460	0	0.0	-33.908873	18.550382	12.427999
4	7764	0	0.0	-33.968128	18.533450	11.906517
5	7800	0	0.0	-34.019579	18.474911	12.060983
6	7806	0	0.0	-34.012013	18.400462	10.072664

Locations of these Street Codes:



The WordClouds on every first suburb in the target Street Codes reveal that these areas all have a different atmosphere and characteristics. Together with the information on the 10 most common venues in the target Street Codes, this forms a good foundation for further exploration.

Conclusion & Discussion

Most of the target Street Codes seem to have good opportunities to start a business like a Coffee Shop since there are no or very few competitors. I would state that especially Street Codes 7800 and 7806 are showing potential in terms of characteristics and types of venues that can be found there. On the other hand, Street Code 7455 is a township which might not be recommendable. Further research is essential be able to recommend more specific locations.
Furthermore, it might be interesting to see if the outcome returns even better options if the Street Codes in Cluster 1 (with moderate competition) are included in the research.

Up to this point I didn't include factors like rental or m^2 prices for commercial properties, accessibility to and from the city centre and demographic structures, so these might be knowledge gaps that need to be filled since they can influence the location decision as well.

Examples of limitations of this project are the limited availability of open data sources and outdated data sources. For more in depth research one might need to invest in paid services to obtain more, and more up to date, data.

All of the above needs to be explored and researched in a next phase to work towards a final recommendation.