Coursera Capstone Project: IBM Data Science
Exploring and Analyzing Schools in Lusaka

Mwimbi Shindola

mwimbi.shindola@gmail.com
Lusaka, Zambia

June 15, 2020

1. Introduction

1.1 Background

Lusaka is the capital and largest city of Zambia. It is one of the fastest developing cities in southern Africa. According to Macprotrends.net, as of 2020, the city's population was about 2.7 million. Lusaka is the center of both commerce and government in Zambia. In view of the nature of Lusaka's population and level of activity, it is important to analyze location of schools in Lusaka neighborhoods so as to properly plan how institutions of learning are located.

1.2 Problem

The problem this project attempts to resolve to do with where new schools can be deployed by government or private investors. Also being resolved is the challenges young families may have in choosing a neighborhood to settle in to make it easy for their school going children.

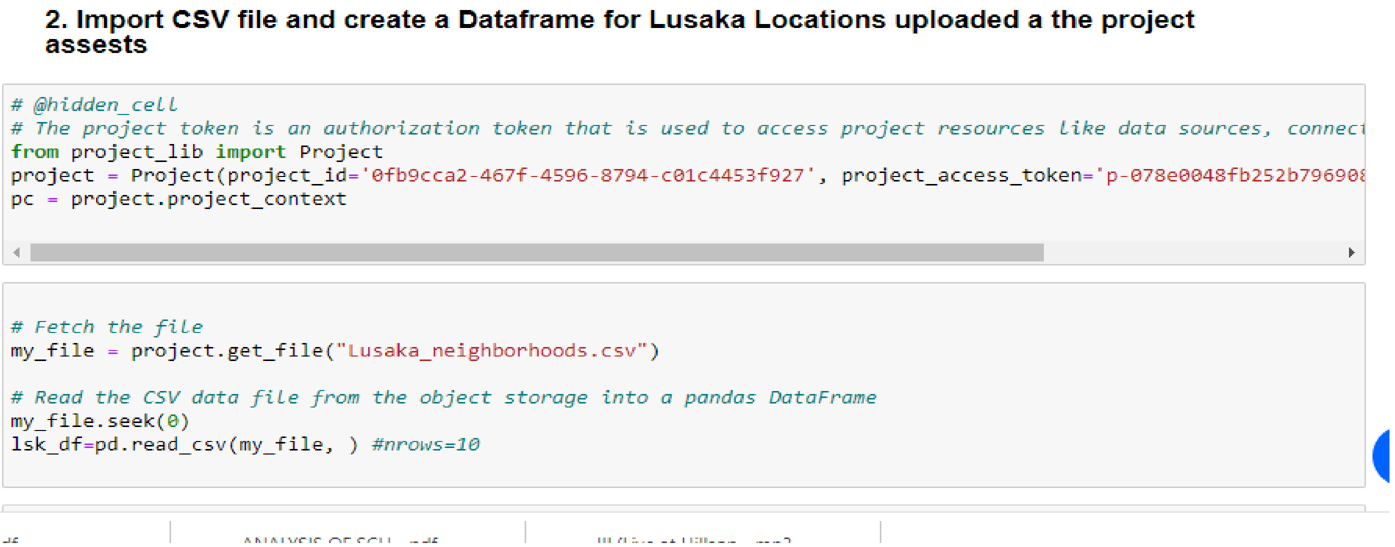
1.3 Interest

Government would be very interested to understand how the schools available to the school going citizens are located. Others who may be interested are private investors who intend to venture into private schools and parents with school going children when choosing a neighborhood to settle in.

2. Data acquisition and cleaning

2.1 Data sources

The data of the Lusaka neighborhoods was collected from Lusaka Wikipedia page here. A second data source macpro here for neighborhoods was used because the data from Wikipedia had some missing neighborhoods. The information for from these two was used to create a csv file that was uploaded to the project Watson studio assets section. The Csv was then imported to the notebook as shown in the image below.



2. Import CSV file and create a Dataframe for Lusaka Locations uploaded a the project assests

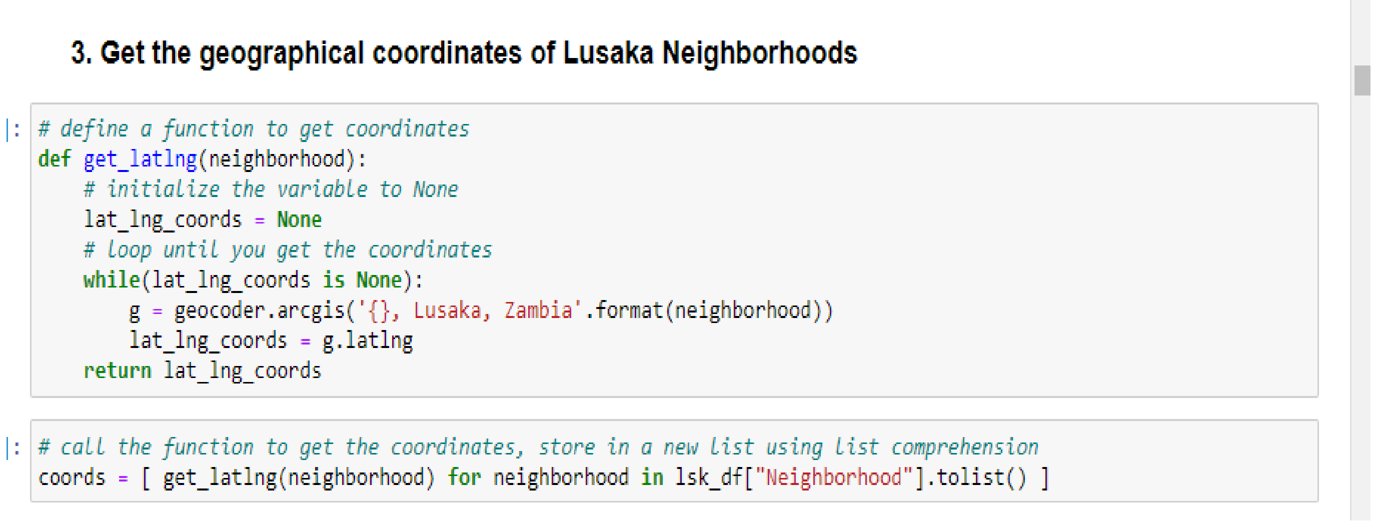
```
# @hidden_cell
# The project token is an authorization token that is used to access project resources like data sources, connect
from project_lib import Project
project = Project(project_id='0fb9cca2-467f-4596-8794-c01c4453f927', project_access_token='p-078e0048fb252b796908
pc = project.project_context
```

```
# Fetch the file
my_file = project.get_file("Lusaka_neighborhoods.csv")

# Read the CSV data file from the object storage into a pandas DataFrame
my_file.seek(0)
lsk_df=pd.read_csv(my_file, ) #nrows=10
```

2.2 Geocoding

The file contents from Lusaka_nieghborhoods.csv were stored in a Pandas DataFrame. The latitude and longitude of the neighborhoods were retrieved using Google Maps Geocoding API. The geometric location values were then stored into the a dataframe which was later merged with the neighborhoods list dataframe.



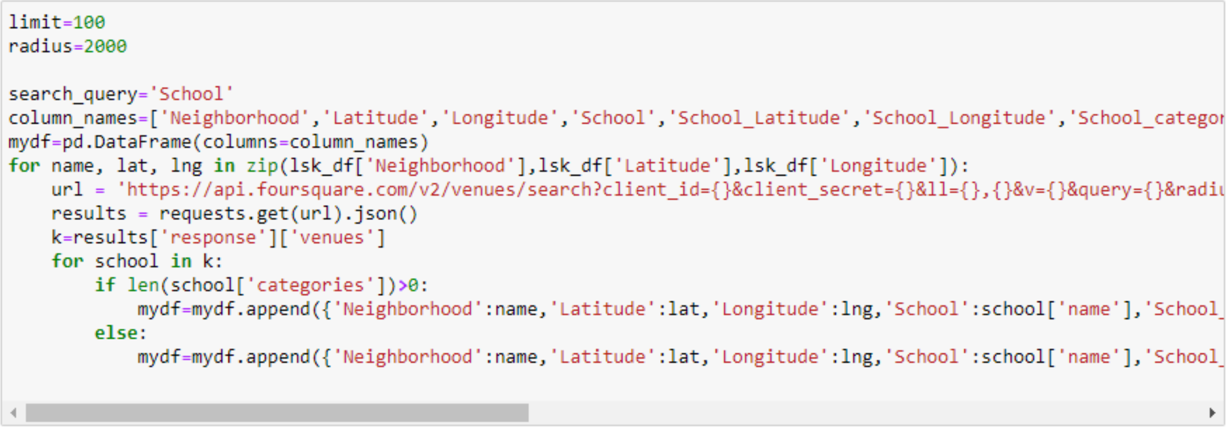
3. Get the geographical coordinates of Lusaka Neighborhoods

```
# define a function to get coordinates
def get_latlng(neighborhood):
    # initialize the variable to None
    lat_lng_coords = None
    # loop until you get the coordinates
    while(lat_lng_coords is None):
        g = geocoder.arcgis('{}, Lusaka, Zambia'.format(neighborhood))
        lat_lng_coords = g.latlng
    return lat_lng_coords
```

```
# call the function to get the coordinates, store in a new list using list comprehension
coords = [ get_latlng(neighborhood) for neighborhood in lsk_df["Neighborhood"].tolist() ]
```

2.3 Venue Data Feature selection

From the location data obtained, the venue data is found out by passing in the required parameters to the FourSquare API, and creating another DataFrame to contain all the venue details along with the respective neighborhoods.
Code



```
limit=100
radius=2000

search_query='School'
column_names=['Neighborhood','Latitude','Longitude','School','School_Latitude','School_Longitude','School_categor
mydf=pd.DataFrame(columns=column_names)
for name, lat, lng in zip(lsk_df['Neighborhood'],lsk_df['Latitude'],lsk_df['Longitude']):
    url = 'https://api.foursquare.com/v2/venues/search?client_id={}&client_secret={}&ll={},{}&v={}&query={}&radiu
    results = requests.get(url).json()
    k=results['response']['venues']
    for school in k:
        if len(school['categories'])>0:
            mydf=mydf.append({'Neighborhood':name,'Latitude':lat,'Longitude':lng,'School':school['name'],'School_
        else:
            mydf=mydf.append({'Neighborhood':name,'Latitude':lat,'Longitude':lng,'School':school['name'],'School_
```

3. Methodology

First, the data was retrieved and cleaned as explained previously. The data consists of schools within a radius of 2km for each neighborhood with a limit set to 100.

Next, we analyze the data by exploring the various categories to which the schools belong. After this we explore the distribution in each neighborhood in a more explanatory way using choropleth maps for visualization. This gives us a good idea of the distribution of schools. Then for those neighborhoods where there are no schools we find the distance to the closest school by incrementing the radius by 100 meters in each iteration until a school is found using Folium.

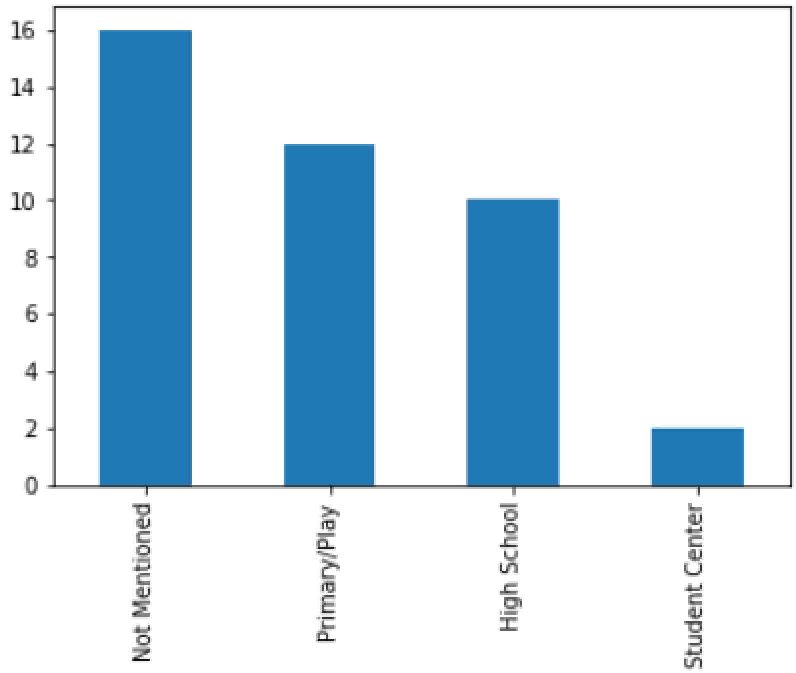
On obtaining this refined data, consisting of the distance to nearest school for each neighborhood and the number of schools in each neighborhood we cluster this data using K-Means. This helps us identify those regions where there is a need for improvement and those neighborhoods which have good facilities in terms of schools.

4. Analysis

Based on the data obtained from Folium about schools, the categories to which schools Belong were explored

4.1 Categories

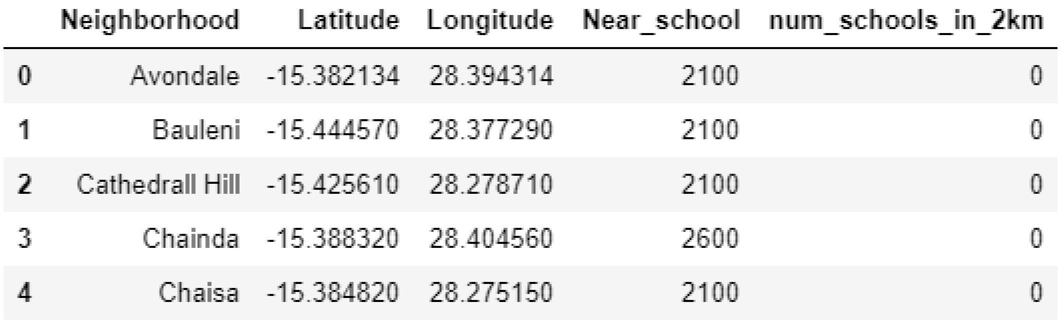
The various categories of schools are as follows:



As we can see most of the schools have not mentioned their category. An interesting thing to note is that a lot of these schools provide education from kindergarten level to high school. Hence there is no surprise in finding only a few schools providing only primary/play level education as the big schools already cover this. Also "Student Center" categorized schools provide education from kindergarten level to high school
(included).

4.2 Clustering the neighborhoods

In this section the neighborhoods were clustered using K-Means clustering algorithm into 3 clusters based on distance to nearest school and number of schools in each neighborhood. For those regions that have no schools within the radius of 1km, I have identified the distance to nearest schools using Folium by incrementing the radius in steps by 100 meters until 2km. Also I have counted the number of schools in each neighborhood and created a dataframe based on this information. In this method I have taken nearest distance to school to vary from 1km to 2km. For all those neighborhoods where the schools may be closer than 1km radius, I have assumed to start from 1km.
The first five elements of this dataframe are shown:

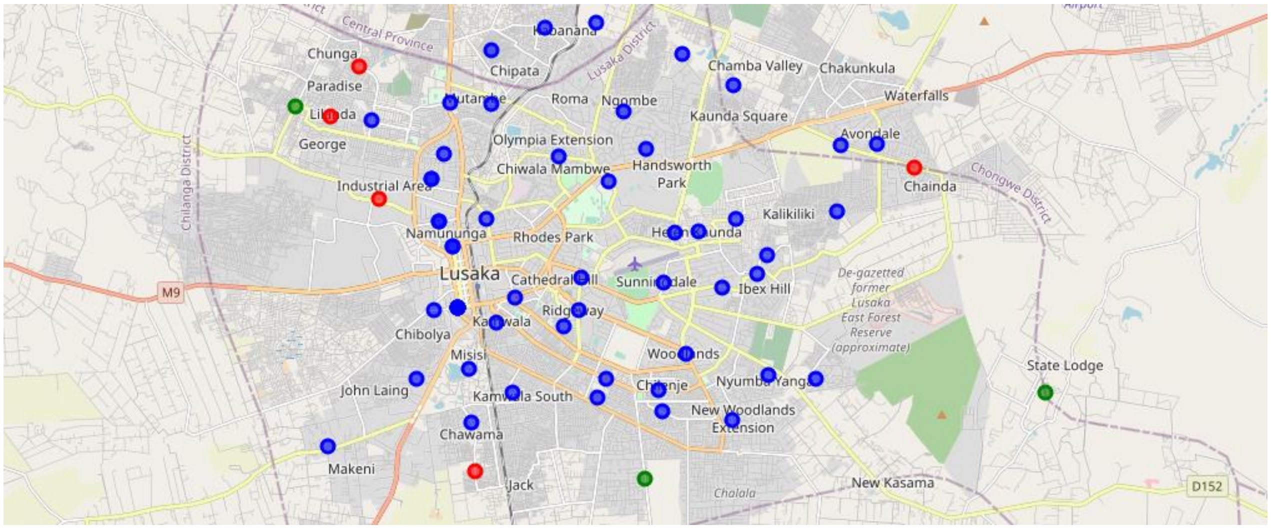


	Neighborhood	Latitude	Longitude	Near_school	num_schools_in_2km
0	Avondale	-15.382134	28.394314	2100	0
1	Bauleni	-15.444570	28.377290	2100	0
2	Cathedrall Hill	-15.425610	28.278710	2100	0
3	Chainda	-15.388320	28.404560	2600	0
4	Chaisa	-15.384820	28.275150	2100	0

The average of the distance to nearest school and the number of schools in each neighborhood has been obtained when grouped by the label assigned in the clustering algorithm. From the clusters it is shown that the closest schools are in cluster0 that are around 1.1km average distance. Cluster 1 has the furthest school and hence requires a lot of improvement.

.

The plot of each neighborhood on map using Folium is as follows:



The blue circles correspond to label 0, the green ones to 1 and the red ones to 2.

5. Results and Discussion

We have observed the various categories the schools belong to in Lusaka, with most of them being more of general schools which provide education from kindergarten level itself till 12th grade. Also we observe that schools providing education only at High School level are comparatively less.

Finally, we have clustered the neighborhoods into 3 clusters and have observed that most of the neighborhoods fall under category with label 0, i.e., the average distance to nearest school for these neighborhoods are 2100 meters. These regions consist of potential neighborhoods for the families with school going children to live because school is nearer compared to other clusters.

Neighborhoods under category label 1 are a great potential for building schools by private investors or government because thy have an average distance of more than 3km from the center to schools requiring improvement which is the motive of this analysis. Government can also concentrate on these neighborhoods for building new schools
The full list of neighborhoods in each cluster can be found in the Jupyter notebook I have attached.

6. Conclusions

In this analysis, the distribution of schools in Lusaka has been analyzed. Also the various neighborhoods have been clustered based on their access to school facilities. The neighborhoods requiring improvement have been identified along with those neighborhoods which have potential for profitable schools. Also the various categories of schools have been observed.