# Coursera Capstone Project: IBM Data Science

Exploring and Analyzing Schools in Lusaka

**Mwimbi Shindola**

mwimbi.shindola@gmail.com

Lusaka, Zambia

June 15, 2020

# Introduction

**Background**

- Lusaka is the capital and largest city of Zambia. It is one of the fastest developing cities in southern Africa
- It is important to analyze location of schools  in Lusaka neighborhoods so as to properly plan how institutions of learning are located

**Problem**

- Where investors can deploy new schools
- Which neighborhoods should government target for building schools
- What neighborhoods can parents with school going children target to live

**Interest**

- Government, Parents with school going children and, Private investors for schools

# Data acquisition and cleaning

**Data sources**

- Wikipedia page for Lusaka
- Macpro
- Created CSV file and uploaded to assets of coursera project

**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
```

# Data acquisition and cleaning contd

**Geocoding**

- 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 datafram<sup>e</sup>

# Data acquisition and cleaning contd

**Venue Data Feature selection**

- 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.

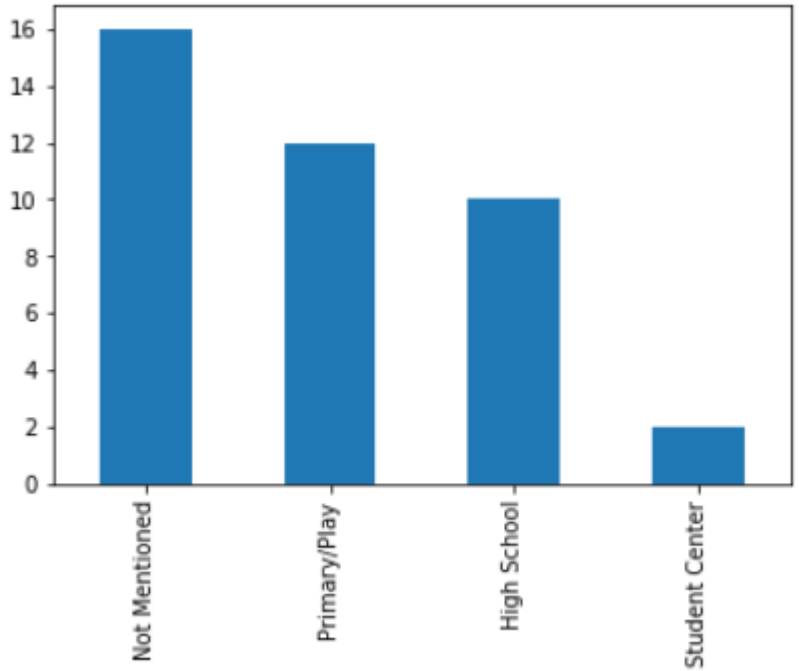
```
limit=100
radius=2000

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_
```

# Methodology

- we analyze the data by exploring the various categories to which the schools belong
- we explore the distribution in each neighborhood in a more explanatory way using choropleth maps for visualization
- cluster this data using K-Means.

# Analysis

- **Categories**



# Analysis contd

**Clustering the neighborhoods**

- the neighborhoods were clustered using K-Means clustering algorithm into 3 clusters
- Also I have counted the number of schools in each neighborhood and created a dataframe

| | 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 |

# Results and Discussion

- We observe that schools providing education only at High School level are comparatively less.
- 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
- Neighborhoods under category label 1 are a great potential for building school

# Conclusions

- In this analysis, the distribution of schools in Lusaka has been analyzed.
- 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.
- The various categories of schools have been observed.