Segmenting and Clustering Socio-Economic Activities in Kenyan Counties

Sammy Ongaya

April 07, 2019

1. Introduction

1.1 Background

In 2010 the republic of Kenya introduced the units of devolved government called County governments. The counties of Kenya are geographical units lead by elected Governor and forms the local government of the particular county. Each county is composed of at least one town with different economic and social activities.

1.2 Problem

Each County government is obligated to generate its own revenue and supplement the income disbursed from the national government. Therefore the each Counties stakeholders and administrator sought to analyse the most viable economical activities that will increase revenue and improve the livelihood of its people. Also for any organization and stakeholders wishing to expand their business operation to different Counties will need to understand the best County to invest in.

The main consumer of this project are the County government's officials and business investors. This project analyses and segments the economic activities within each County and cluster the most common economical activities in each county. With this analysis the stakeholders are well informed on the most appropriate economical activities to invest in each county and make informed decision that will yield high ROI.

2. Data

This project is a data driven project and it utilizes data from 47 Counties in Kenya. The data is extracted from the wikipedia website through web scrapping using Python based library. The data has County name, Area in KM squared and Population Census as of 2009 among other details. We use the geopy library to get the coordinates of each county. We then create a map of Kenya and plot the coordinates of each county using the folium library.

After successfully extracting the relevant data we use the Foursquare API to extract relevant information such as facilities within a radius of 25km for all Counties and get the venue categories for each facility. We then explore the facilities for each county by retrieving the nearby venues. We analyse each county and identify which are the most dominant facilities in each area by displaying the top 5 most common venues.

We cluster the venues for all counties using the K-means clustering algorithm. K-means is unsupervised machine learning algorithm is used to segment data to identify pattern. We create a map using folium library to display the clusters. We finally analyse each cluster based on results obtained from the machine learning algorithm. This informs us on the business activities that can be viable at a given region in each county. The business investors can use this research to invest in activities that will have high return on investment (ROI) and also helping the county government in making strategic decisions on how to best improve and support business investor in their respective regions by providing necessary services

3. Methodology

3.1 Exploratory Data Analysis

We have a total of 47 Counties in Kenya geographically demacated based on population and sizes. The County with highest population is Nairobi while the least populated County is Isiolo.

Top 10 Counties by Population

	Code	County	Former Province	Area_km2	Population	Capital
46	47.0	Nairobi (County)	Nairobi (Province)	694.9	3138369	Nairobi (City)
36	37.0	Kakamega	Western	3033.8	1660651	Kakamega
21	22.0	Kiambu	Central	2449.2	1623282	Kiambu
31	32.0	Nakuru	Rift Valley	7509.5	1603325	Nakuru
38	39.0	Bungoma	Western	2206.9	1375063	Bungoma
11	12.0	Meru	Eastern	6930.1	1356301	Meru
44	45.0	Kisii	Nyanza	1317.9	1152282	Kisii
2	3.0	Kilifi	Coast	12245.9	1109735	Kilifi
15	16.0	Machakos	Eastern	5952.9	1098584	Machakos
8	9.0	Mandera	North Eastern	25797.7	1025756	Mandera

Least 10 Counties by Population

	Code	County	Former Province	Area_km2	Population	Capital
23	24.0	West Pokot	Rift Valley	8418.2	512690	Kapenguria
30	31.0	Laikipia	Rift Valley	8696.1	399227	Nanyuki
27	28.0	Elgeyo-Marakwet	Rift Valley	3049.7	369998	Iten
12	13.0	Tharaka-Nithi	Eastern	2409.5	365330	Kathwana
9	10.0	Marsabit	Eastern	66923.1	291166	Marsabit
5	6.0	Taita–Taveta	Coast	17083.9	284657	Mwatate
3	4.0	Tana River	Coast	35375.8	240075	Hola
24	25.0	Samburu	Rift Valley	20182.5	223947	Maralal
4	5.0	Lamu	Coast	6497.7	191539	Lamu
10	11.0	Isiolo	Eastern	25336.1	143294	Isiolo

The County with largest size is Turkana while Mombasa is the County with smallest size

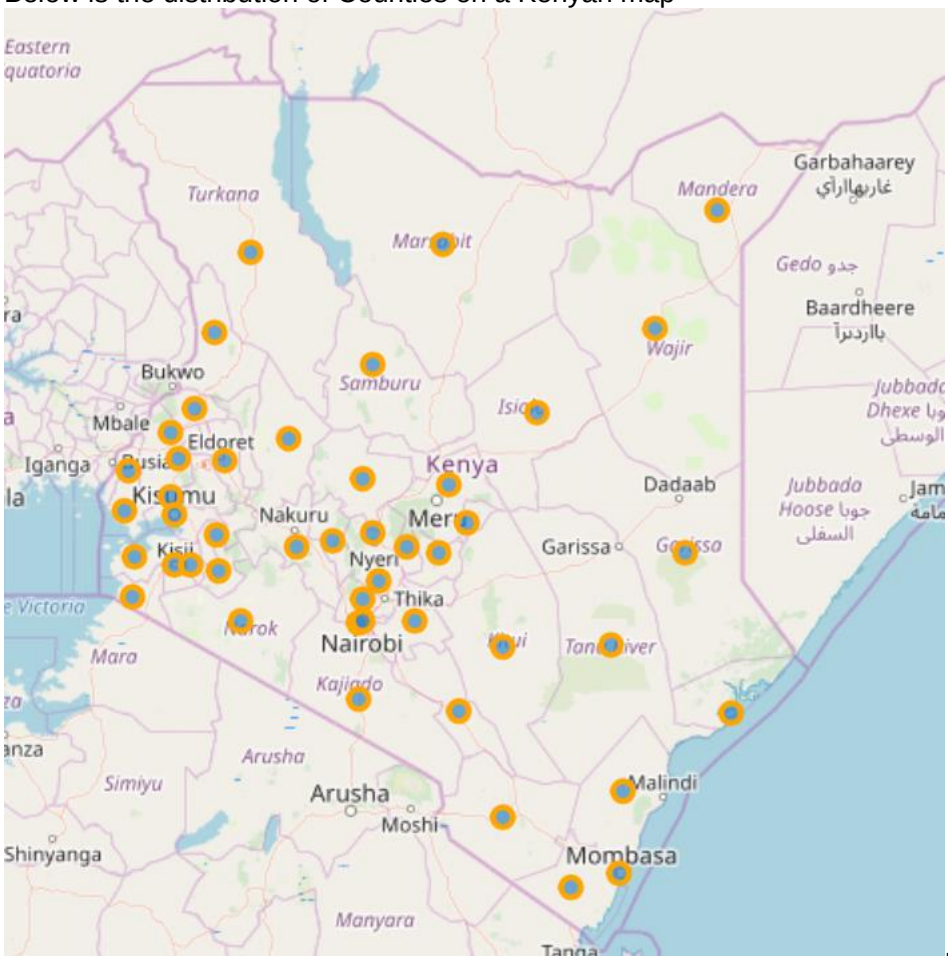
Top 10 Counties by Size

	Code	County	Former Province	Area_km2	Population	Capital
22	23.0	Turkana	Rift Valley	71597.8	855399	Lodwar
9	10.0	Marsabit	Eastern	66923.1	291166	Marsabit
7	8.0	Wajir	North Eastern	55840.6	661941	Wajir
6	7.0	Garissa	North Eastern	45720.2	623060	Garissa
3	4.0	Tana River	Coast	35375.8	240075	Hola
8	9.0	Mandera	North Eastern	25797.7	1025756	Mandera
10	11.0	Isiolo	Eastern	25336.1	143294	Isiolo
14	15.0	Kitui	Eastern	24385.1	1012709	Kitui
33	34.0	Kajiado	Rift Valley	21292.7	687312	Kajiado
24	25.0	Samburu	Rift Valley	20182.5	223947	Maralal

Least 10 Counties by Size

	Code	County	Former Province	Area_km2	Population	Capital
38	39.0	Bungoma	Western	2206.9	1375063	Bungoma
41	42.0	Kisumu	Nyanza	2009.5	968909	Kisumu(City)
35	36.0	Bomet	Rift Valley	1997.9	730129	Bomet
39	40.0	Busia	Western	1628.4	743946	Busia
44	45.0	Kisii	Nyanza	1317.9	1152282	Kisii
19	20.0	Kirinyaga	Central	1205.4	528054	Kerugoya / Kutus
45	46.0	Nyamira	Nyanza	912.5	598252	Nyamira
46	47.0	Nairobi (County)	Nairobi (Province)	694.9	3138369	Nairobi (City)
37	38.0	Vihiga	Western	531.3	554622	Vihiga
0	1.0	Mombasa (County)	Coast	212.5	939370	Mombasa (City)

Below is the distribution of Counties on a Kenyan map



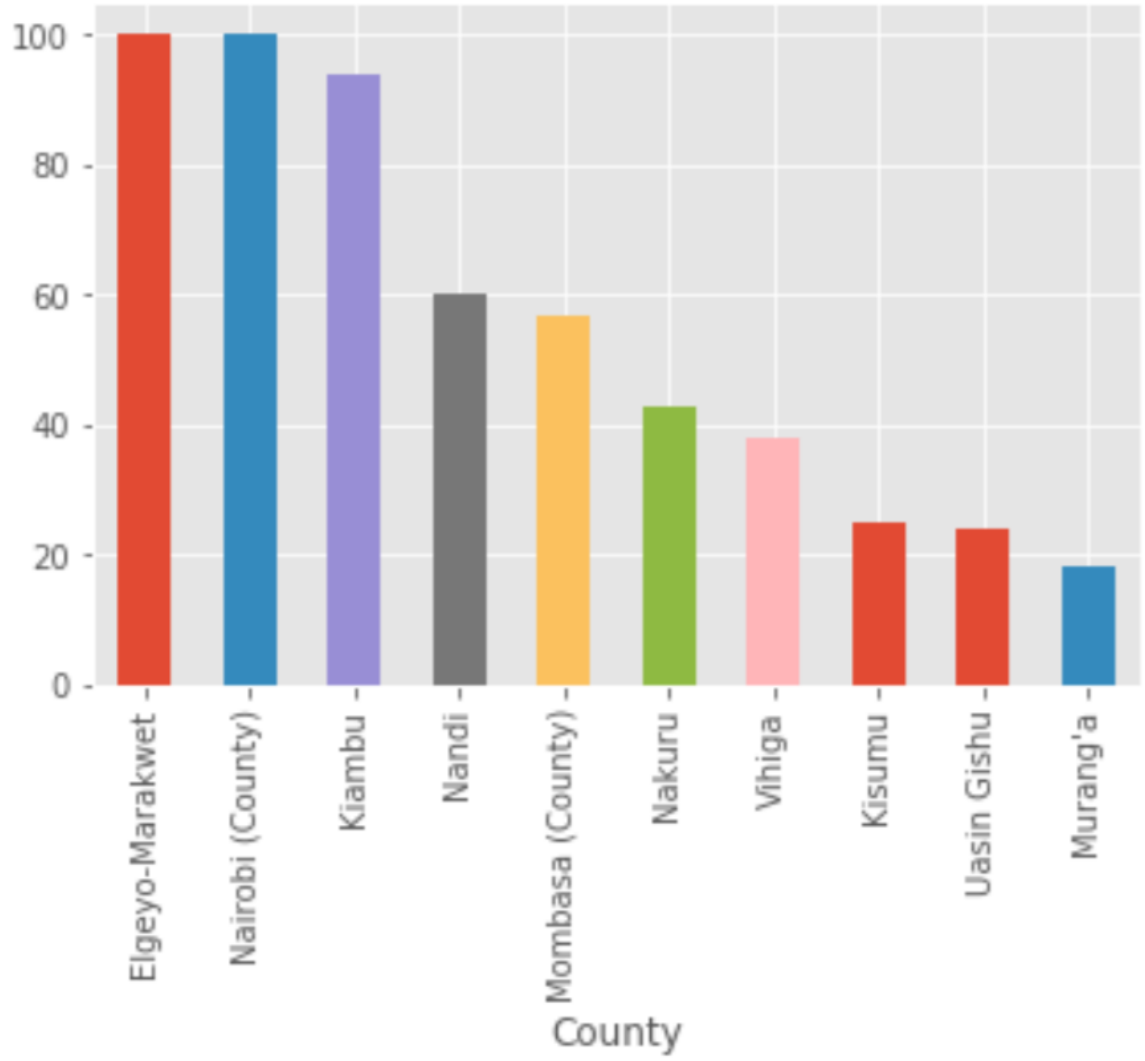
Using the Foursquare API we retrieve a list of up to 100 venues for each county within a radius of 25KM. We get a total of 703 venues returned by the Foursquare API. Elgeyo-Marakwet and Nairobi Counties 100 have the highest venues count of 100 each while Kilifi and Kwale are the least with only 1 venue from the data retrieved from Foursquare API

15 Counties with Most Venues		15 Counties with Least Venues	
Elgeyo-Marakwet	100	Lamu	5
Nairobi (County)	100	Trans-Nzoia	5
Kiambu	94	Nyamira	4
Nandi	60	Kericho	4
Mombasa (County)	57	Nyandarua	4
Nakuru	43	Homa Bay	4
Vihiga	38	Siaya	4
Kisumu	25	Busia	4
Uasin Gishu	24	Machakos	4
Murang'a	18	Kajiado	3
Nyeri	11	Tharaka-Nithi	3
Kirinyaga	10	Bungoma	3
Makueni	10	Kitui	1
Kakamega	9	Kwale	1
Bomet	8	Kilifi	1

We have a total of 129 Venue categories returned by Foursquare API. Top venues categories are Hotel with 82 venues and Shopping Mall with 38 venues. The least categories are Dessert

Shop and Harbor / Marina each with 1 venue

Top Venue Categories	
Hotel	82
Shopping Mall	38
Café	35
Lounge	35
African Restaurant	31
Resort	29
Nightclub	28
Restaurant	25
Bar	22
Coffee Shop	21
Least Venue Categories	
Gift Shop	1
General Travel	1
Gaming Cafe	1
French Restaurant	1
Farm	1
English Restaurant	1
Duty-free Shop	1
Dumpling Restaurant	1
Dessert Shop	1
Harbor / Marina	1



We also computed the frequency of occurrence of each venue in each County as seen from Notebook While limiting our count to 5 venues per County. Additionally we analysed the top 10 most common venues in each County

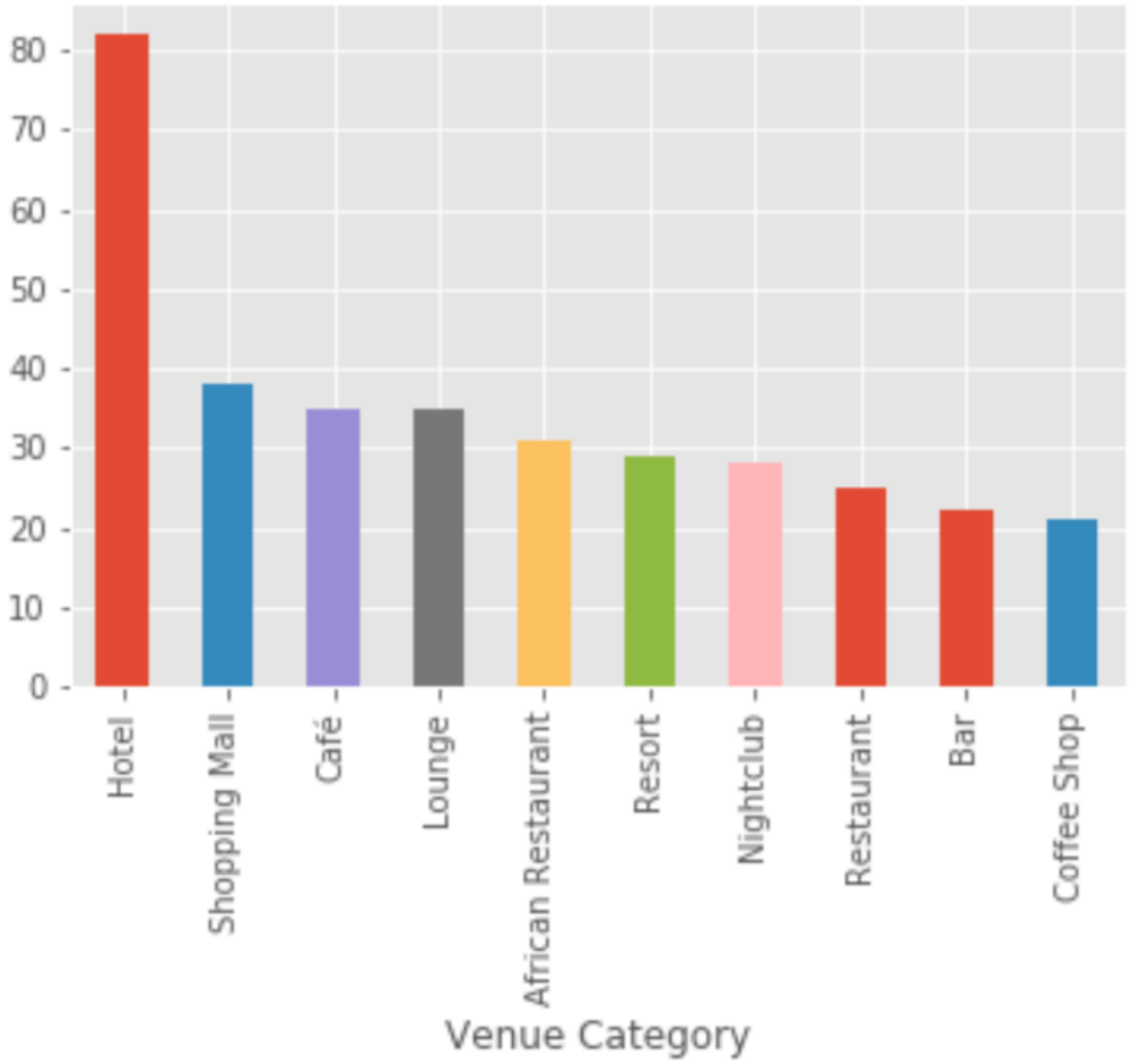
	County	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Baringo	Lake	Hotel	Zoo	Diner	English Restaurant	Electronics Store	Eastern European Restaurant	Duty-free Shop	Dumpling Restaurant	Dessert Shop
1	Bomet	African Restaurant	Café	Bar	River	IT Services	Vineyard	Travel & Transport	Lounge	Convenience Store	Convention Center
2	Bungoma	Historic Site	Shop & Service	Food	Zoo	Dessert Shop	Eastern European Restaurant	Duty-free Shop	Dumpling Restaurant	Diner	Department Store
3	Busia	Vegetarian / Vegan Restaurant	Cocktail Bar	Bus Station	River	Zoo	Dessert Shop	Eastern European Restaurant	Duty-free Shop	Dumpling Restaurant	Diner
4	Elgeyo-Marakwet	Hotel	Lounge	Café	African Restaurant	Italian Restaurant	Coffee Shop	Burger Joint	Bar	Shopping Mall	Chinese Restaurant
5	Embu	Beer Garden	Shopping Mall	Nightclub	Bus Station	Convenience Store	Diner	Electronics Store	Eastern European Restaurant	Duty-free Shop	Dumpling Restaurant
6	Homa Bay	Plaza	Beach	Hotel	Restaurant	Zoo	Diner	Electronics Store	Eastern European Restaurant	Duty-free Shop	Dumpling Restaurant
7	Kajiado	Campground	Mountain	Zoo	Ethiopian Restaurant	Coffee Shop	Concert Hall	Convenience Store	Convention Center	Department Store	Dessert Shop
8	Kakamega	Hotel	Lounge	Food	Department Store	Restaurant	Electronics Store	Convenience Store	Nightclub	English Restaurant	Concert Hall

3.2 Machine Learning

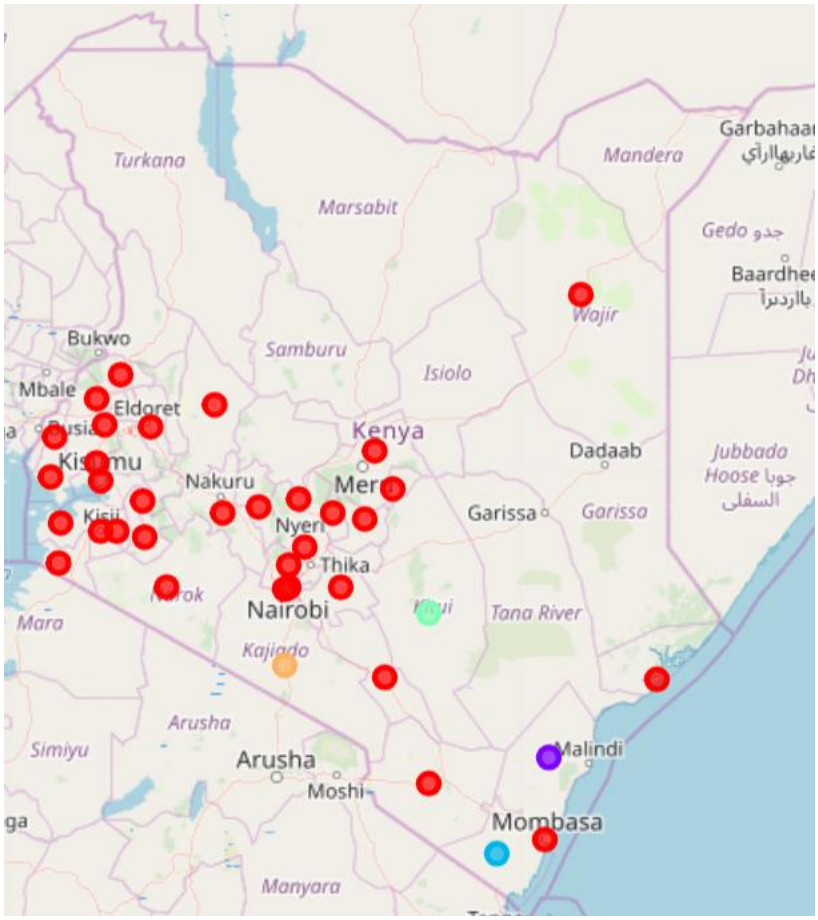
This problem is a form of unsupervised leaning since we don't have labelled data which is the case of supervised learning. We used the K-Means algorithm. K-means is unsupervised learning algorithm that falls into the Clustering classes of unsupervised algorithms. In clustering problem the algorithms tends to group data points into different clusters with each cluster having similar features. We use the K-Means clustering algorithm in this project because our data is not labelled and our objective is to segment the venues into 5 cluster each having similar characteristics.

4. Results

From this project we see that most venues are for hotels and shopping malls while the least venues are the Desert shop and Marina. This is because the fast consumer moving goods are essential to most people and are found anywhere unlike rare venues such as Desert shops and Marina's. Few Counties have Habours and Marina's because they are land locked and most don't have water bodies such as lakes.



We cluster our data into 5 clusters and observe that most venues fall into the first cluster forming the largest cluster. This is because of the Similarities of the venues providing similar services such as Food shops, hotels, resorts, restaurants and coffee shops which are grouped into one cluster.



The size of the County does not influence the number of venues in a County. This can be seen in County such as Wajir which is the third largest County in terms of geographical boundaries with a size of 55840.6KM squared but has it has very few venues (5 venues). While Nairobi County with a size of 694.9 KM squared but has a total of 100 venues.

Population seems to influence the presence of venues in our analysis. We can see that for County such as Nairobi with the highest population of 3138369 Million people has the highest number of venues 100 venues. While Counties such as Lamu with a population size of 191539 persons has a total of 5 venues.

5. Discussion

The results from our analyses shows that most of the venues are food related venues such as hotels, restaurants, food courts and related venues. For counties with natural sceneries such lake, mountain, national park, game reserves and historical sites tourist activities can suffice well. This can be seen in Counties such as Bungoma, Kilifi, Lamu, Baringo and Mombasa among others. The business activities that are in demand for Cities and major towns such as Nairobi, Kisumu, Mombasa Nakuru and Uasin Gishu are hotels, restaurants, resorts, shopping malls. This is due to high population.

From the above analysis we can see that most venues in Kenya are food and restaurant facilities which lies in Cluster 1. From cluster 2 we find Coffee and Cocktail Shops which are more dominant in Kilifi County.

Our data is limited to the data retrieved from the Foursquare API. This data is not adequate for analysis to Counties whose venues are not adequately collected by the foursquare provider, however, it provides a general overview of the economic activities position of Kenyan Counties.

6. Conclusion

For Counties that have major towns in Kenya such as Nairobi, Mombasa, Kisumu and Nakuru the most dominant business activity is related to consumer food facilities such as hotels, restaurants, and cafe's e.t.c. For business investors the most promising business is the hotel and restaurants business. For administrative purpose the Counties and relevant administrative authorities should invest and improve the agricultural production industry, build good roads to transport farm inputs to the urban areas. For Counties with Tourist attraction sites such as Mountains, wildlife, historical sites, lakes and beach they should invest in tourist attraction economical activities as well as food courts. However, our resaarch didn't cover 11 Counties including Turkana, Mandera among other 11 due to unavailability of data. For the full source code and analysis of this project visit my github account here.