Hello!

Project 5 - Client

Khadija Conteh
Matt DeVay
Aidan Dominguez
ATX-DSI-11

GENERAL ASSEMBLY

Problem Statement/Overview

Create a feature than could enhance current image-based predictions of informal settlements based on the ratio of real estate adverts to population density.

Workflow

1. Research!
2. Collect real estate data via webscraping
3. Geocode real estate data
4. Regionalize population density data
5. Correlate real estate data to population data for each region

next

Guided Walk-Through:

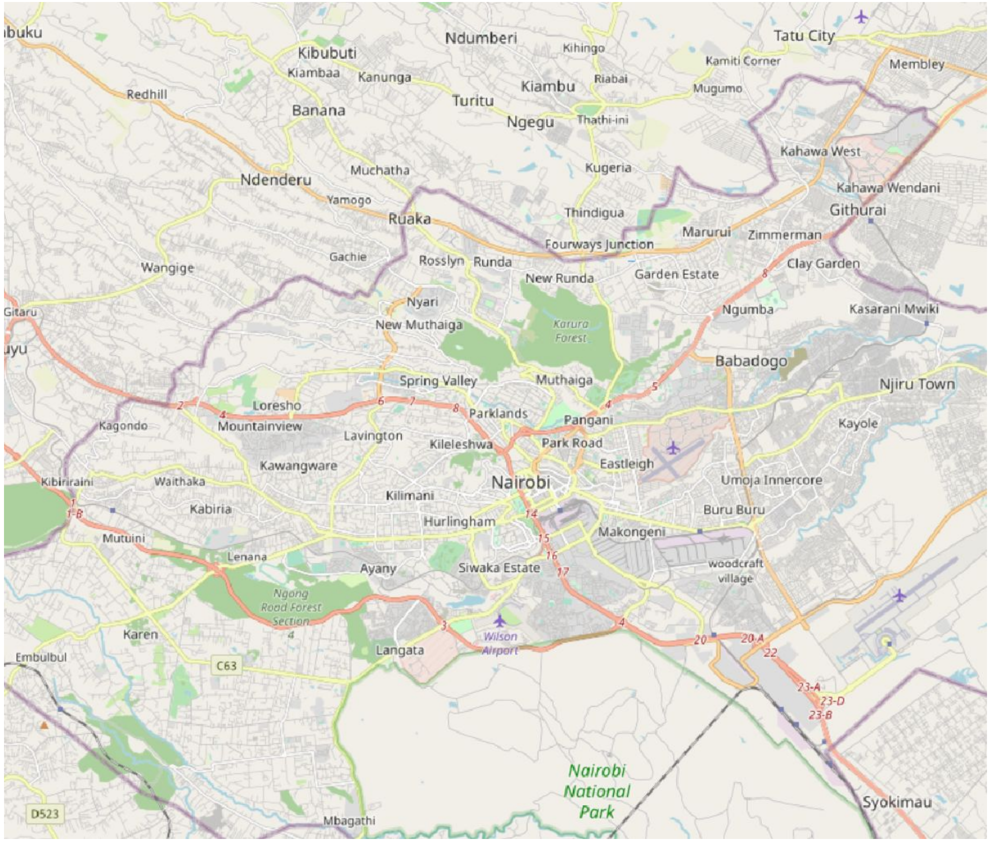
Technology Used

- Python with SKLearn, MatPlotLib, Pandas
- Selenium
- QGIS
- Google Earth Pro
- Google Maps
- Google Maps API
- Tableau

Data Collection

City Choice: Nairobi

Nairobi: Reference Map



City Background: Nairobi, Kenya

1. Modern, rapidly growing
2. English is common
3. Large and dynamic informal settlements
4. Accurate and recent population data available

Nairobi Demographics

- Population: 4.4 million residents
- Area: 269 sq mi
- Growth rate > 4%
- Poverty rate >20%
- High birth and immigration rates

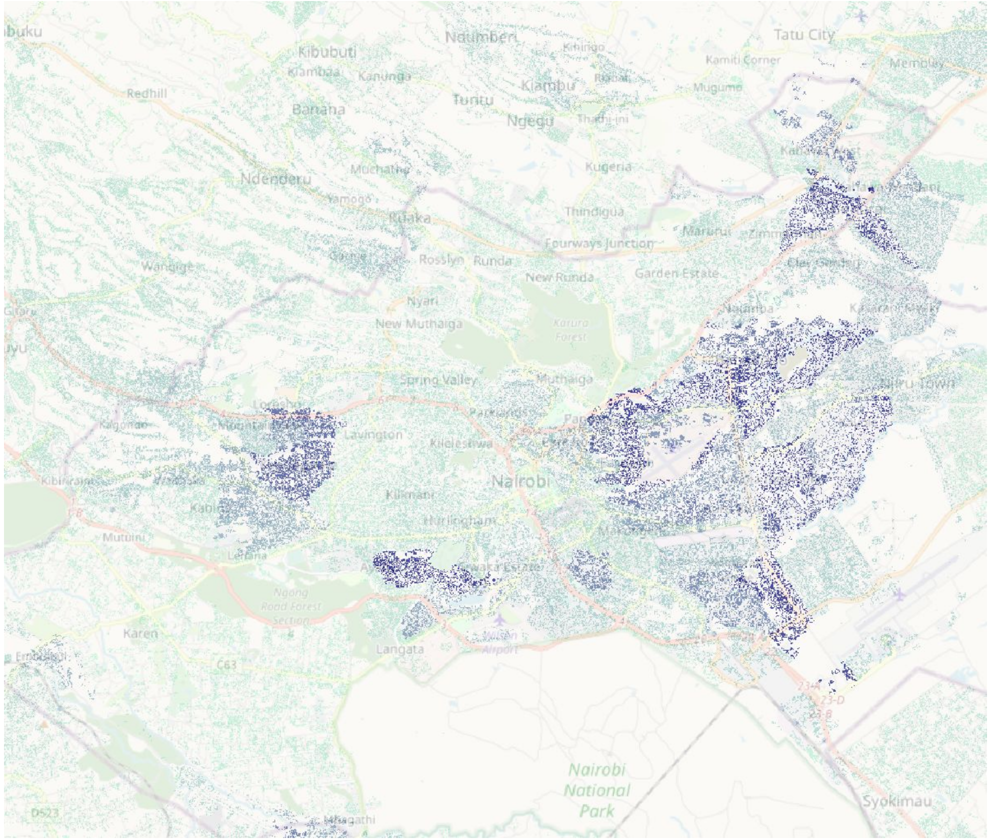
Population Density Data

Population Density Data (https://data.humdata.org/)

High resolution (1 arcsecond grid) population counts from the Humanitarian Data Exchange, created by Columbia University and FaceBook.

- High resolution (30m grid counts of population)
- Complete
- Updated frequently
- Convenient format
- REALLY REALLY BIG! More than 11 million data points for Kenya

Nairobi: Population Density Overlay



Webscraping

- Website: https://www.buyrentkenya.com/
- Date Scraped: May 08, 2020
- Total Advertisements scraped: 4,000
  - Rental Advertisements: 2,080
  - For Sale Advertisements: 1,920

County Locations

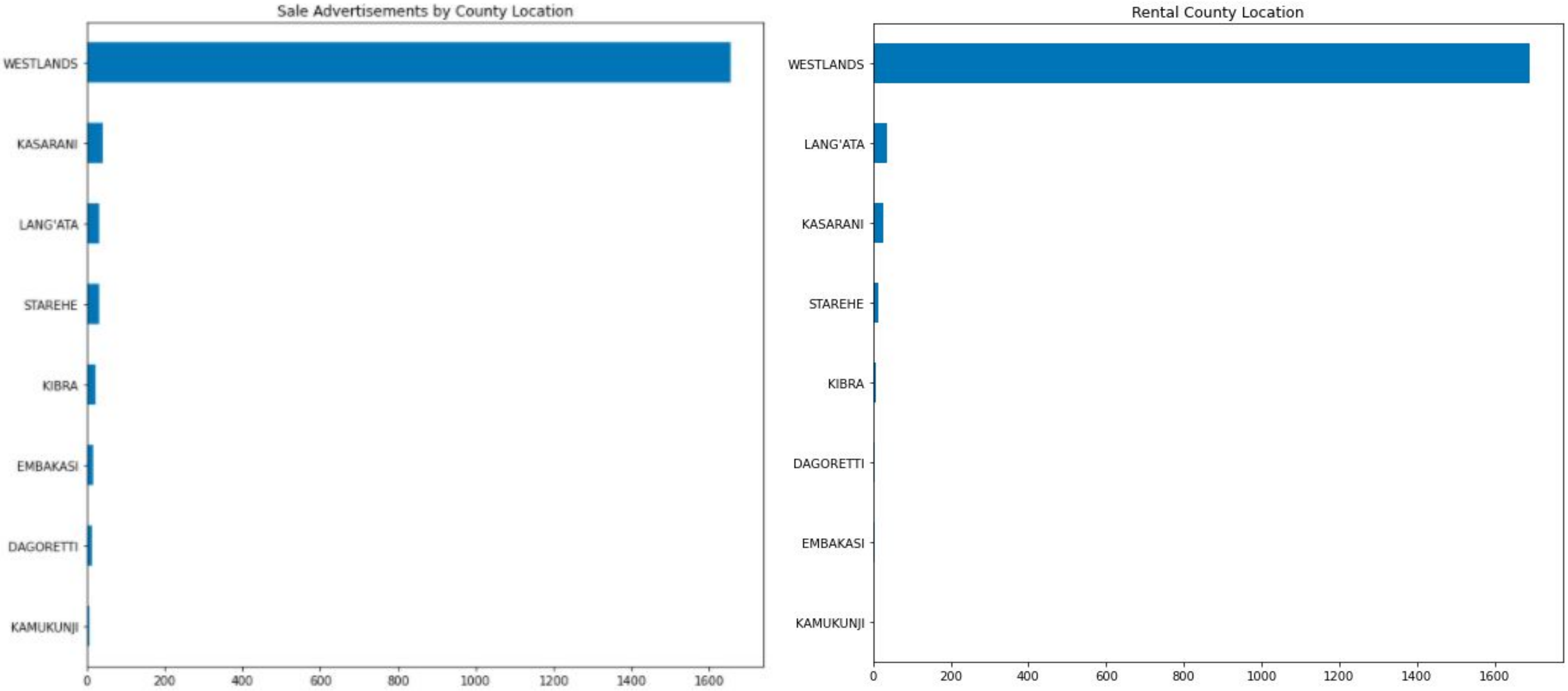
Housing price by County

For Sale	
Kamukunji	288,246 USD
Westlands	231,958 USD
Starehe	162,791 USD
Kasarani	115,341 USD
Lang'ata	115,176 USD
Embakasi	102,595 USD
Dagoretti	94,269 USD
Kibra	71,005 USD

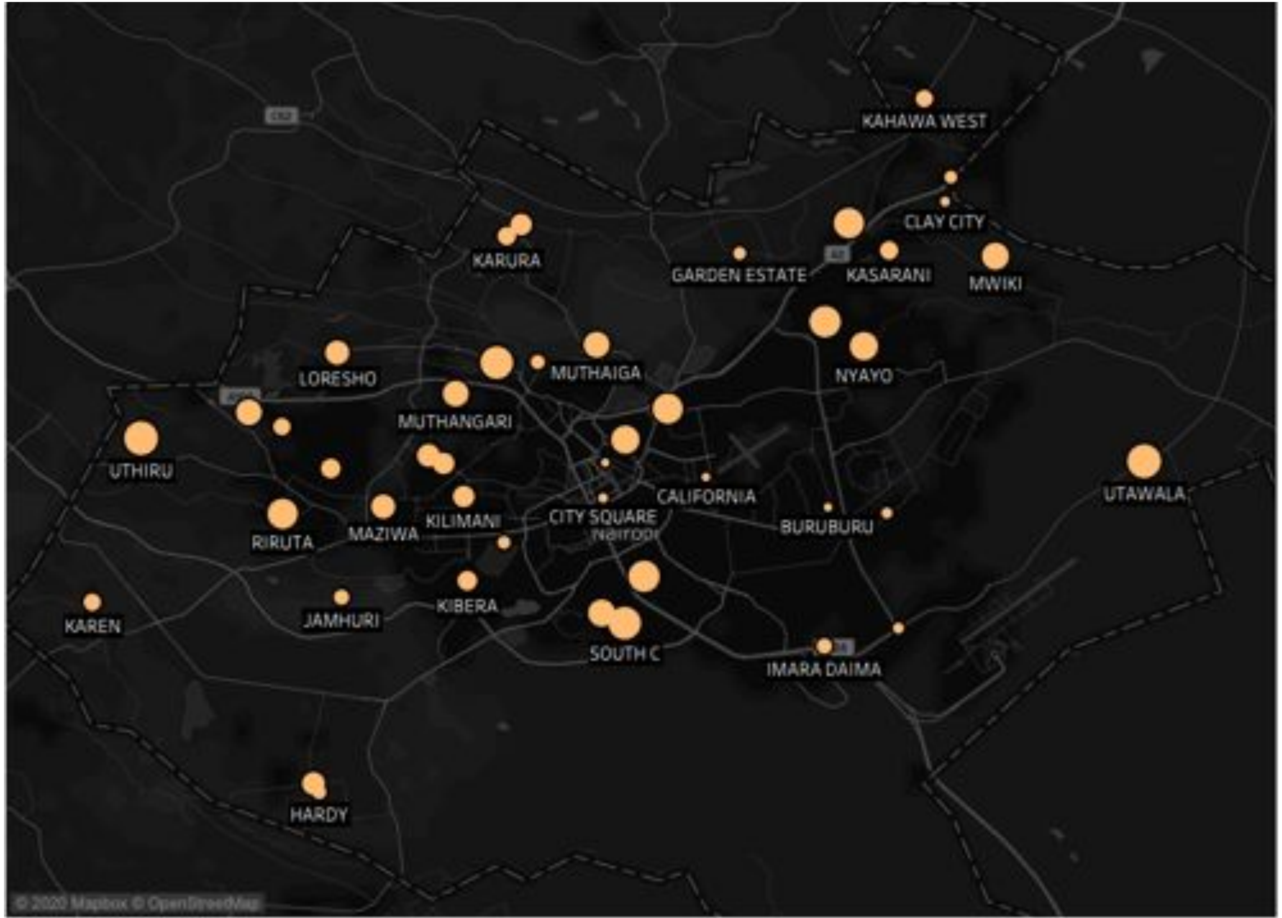
Rentals	
Westlands	1,427 USD
Lang'ata	730 USD
Kasarani	509 USD
Embakasi	502 USD
Starehe	496 USD
Kibra	433 USD
Kamukunji	424 USD
Dagoretti	325 USD

*USD price based on KES May 08, 2020 conversion rate (1 USD = 106 KES)

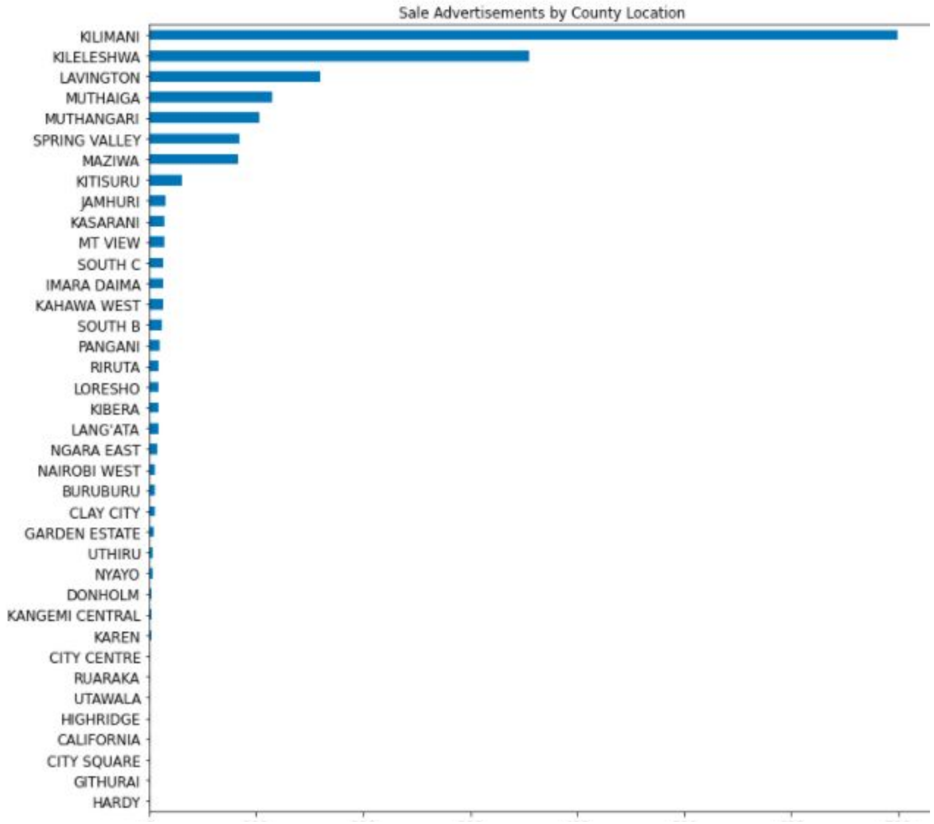
County Sale and Rental Advertisements

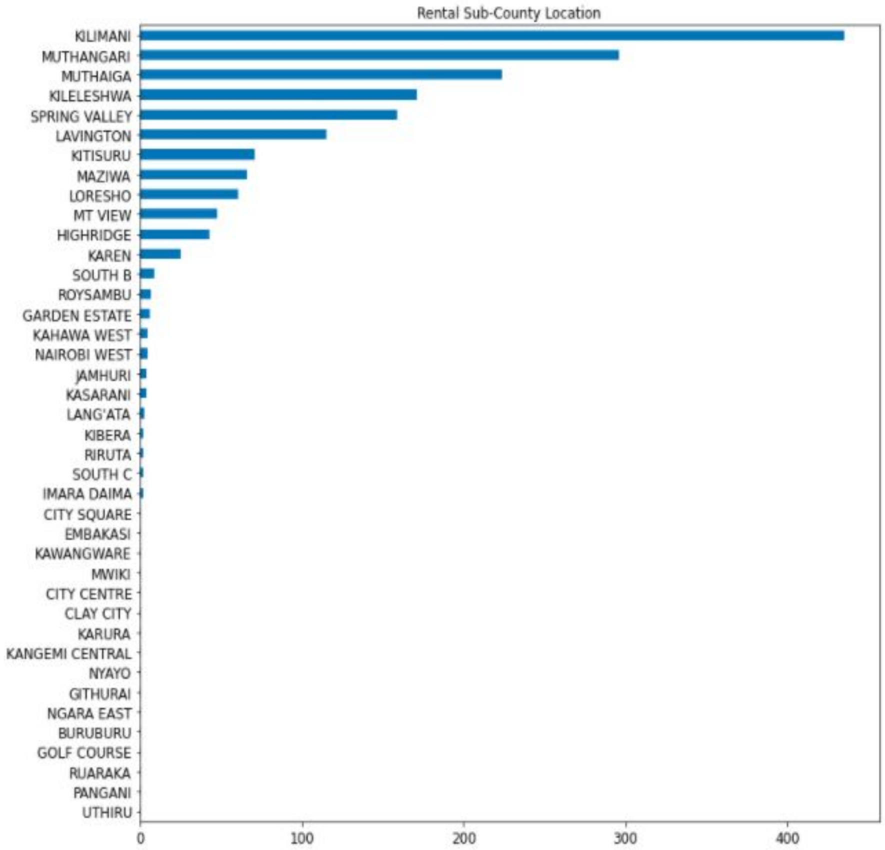


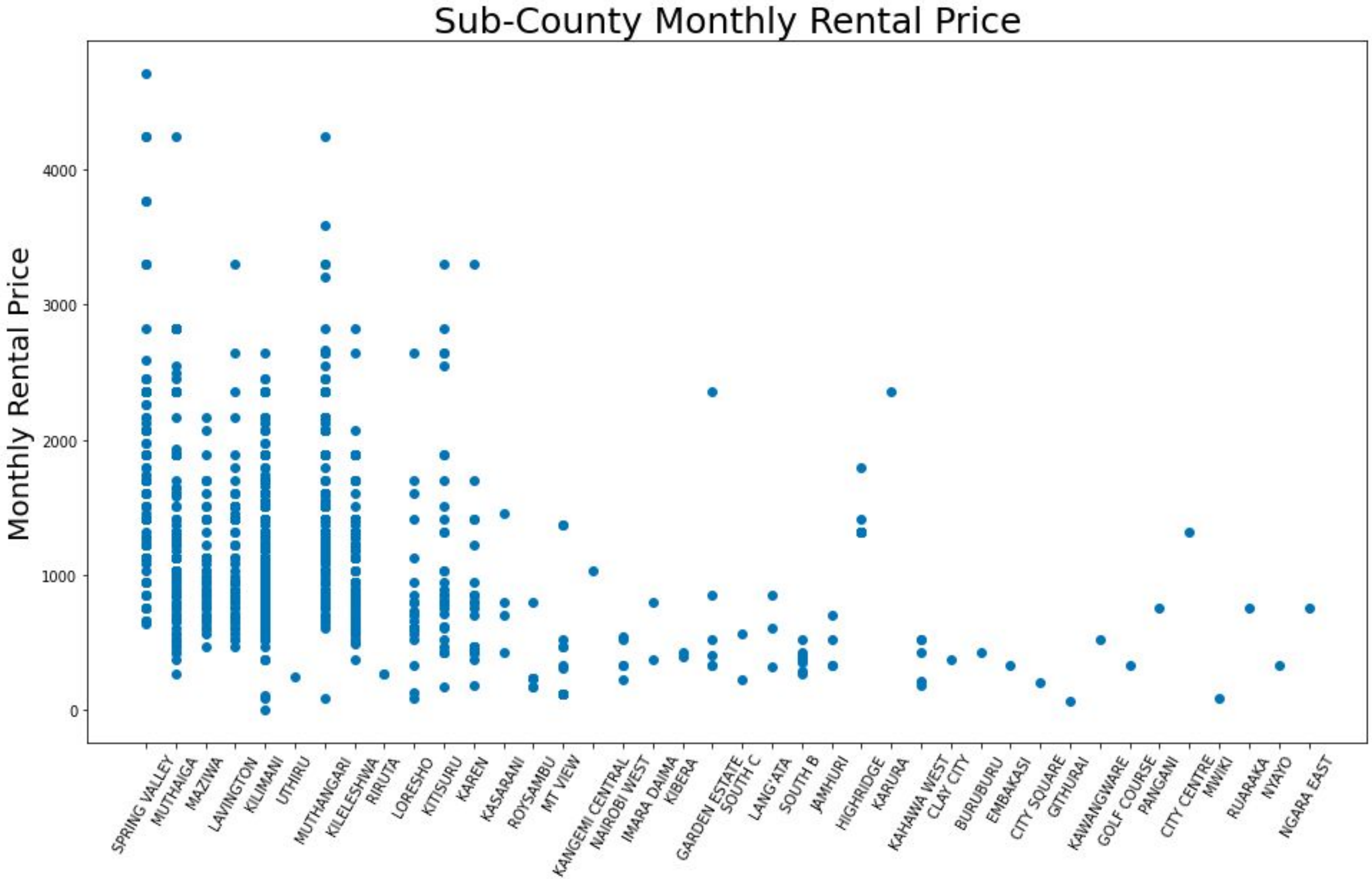
Sub-County location



Sub-County Sale and Rental Advertisements







Sub-County Monthly Rental Price
Monthly Rental Price
0
1000
2000
3000
4000
SPRING VALLEY
MUTHAIGA
MAZIWA
LAVINGTON
KILIMANI
UTHIRU
MUTHANGARI
KILELESHWA
RIRUTA
LORESHO
KITISURU
KAREN
KASARANI
ROYSAMBU
MT VIEW
KANGEMI CENTRAL
NAIROBI WEST
IMARA DAIMA
KIBERA
GARDEN ESTATE
SOUTH C
LANG'ATA
SOUTH B
JAMHURI
HIGHRIDGE
KARURA
KAHAWA WEST
CLAY CITY
BURUBURU
EMBAKASI
CITY SQUARE
GITHURAI
KAWANGWARE
GOLF COURSE
PANGANI
CITY CENTRE
MWIKI
RUARAKA
NYAYO
NGARA EAST

Challenges

- Selecting real estate site to scrape
- Lack of date reference for advertisements
- Generalized or missing advertisements locations
- Misclassification of advert locations
- Price currency conversion

Informal Settlements...
what are they anyway?

Informal Settlements

Informal settlements are unplanned areas where housing is constructed on land that the occupants have no legal claim to, or occupy illegally.*

As these locations fall outside of government regulation these areas usually lack, or are cut off from basic public services (i.e. schools, water, sanitation etc.)

*source: Glossary of Environment Statistics, Studies in Methods, Series F, No. 67, United Nations, New York, 1997.

Can real estate and public services data serve to indicate probable areas for informal settlements?

Real Estate Data & Primary School locations

Real Estate Data & Nairobi City Water and Sewerage locations

Geocoding

Geocoding - What is it?

- Geocoding is assigning Latitude and Longitude coordinates to locations
- Several ways of doing it:
  - Geocoding regions and correlating real estate data directly to the region
  - Geocoding the address itself and mapping it
- Keep in mind: Population Density Data is already geocoded and is extremely granular (~30m squares), and also needs to be regionalized

Geocoding - Challenges

- Full addresses or GPS coordinates were not available for real estate data
- Local place names and sub-county districts are not accurately mapped
- Neighborhoods often overlap
- Complex map boundary data is not easy to work with in python
- No familiarity with local place name conventions or usage
- Hard to automate based on generally available data

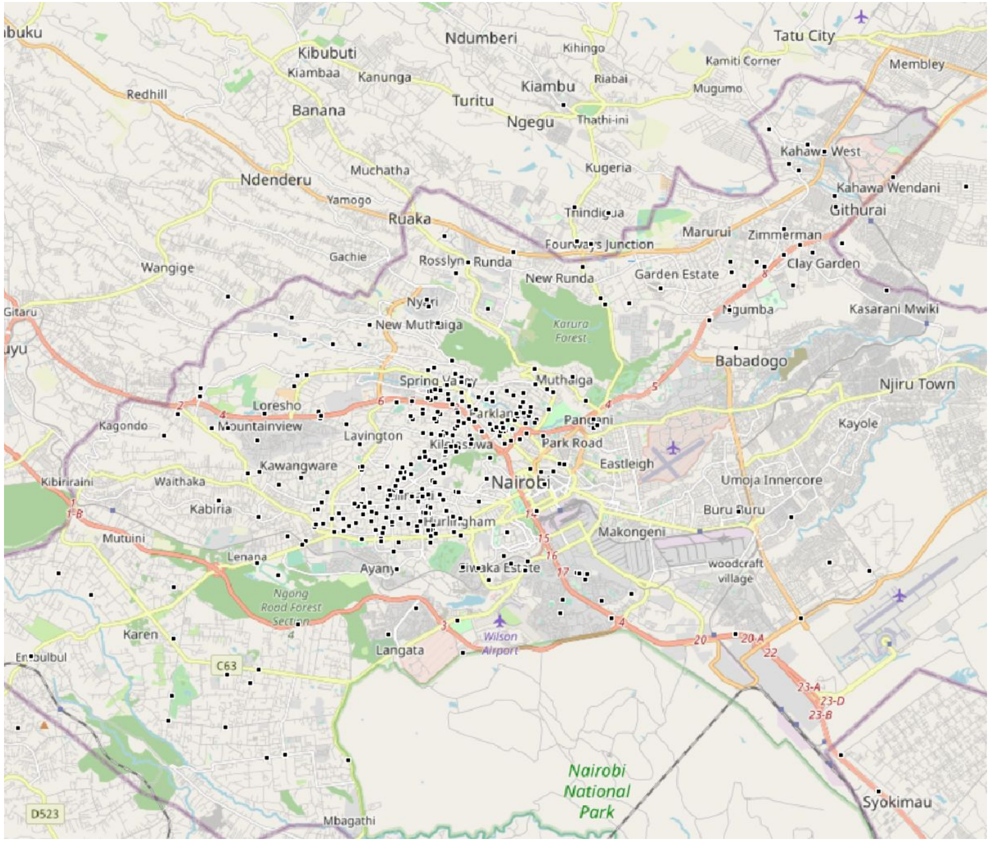
Geocoding - Enter Google

- Google Maps API can geocode just about anything
- Simple to utilize and automate
- Cheap

Geocoding - Google Caveats

- GIGO
- It may be cheap, but it’s not free
- Even automated, it can take a while (1-2 seconds per record)
- Even good addresses likely return geocodes with significant error
- Errors are impossible to reasonably quantify
- Regionalizing Issues

Geocoding: Mapped Results



Modeling

Models

KMeans & DBSCAN

- Filtered by different population thresholds
  - 2, 4, 6, 7, 8
- Scaled
- K-Means

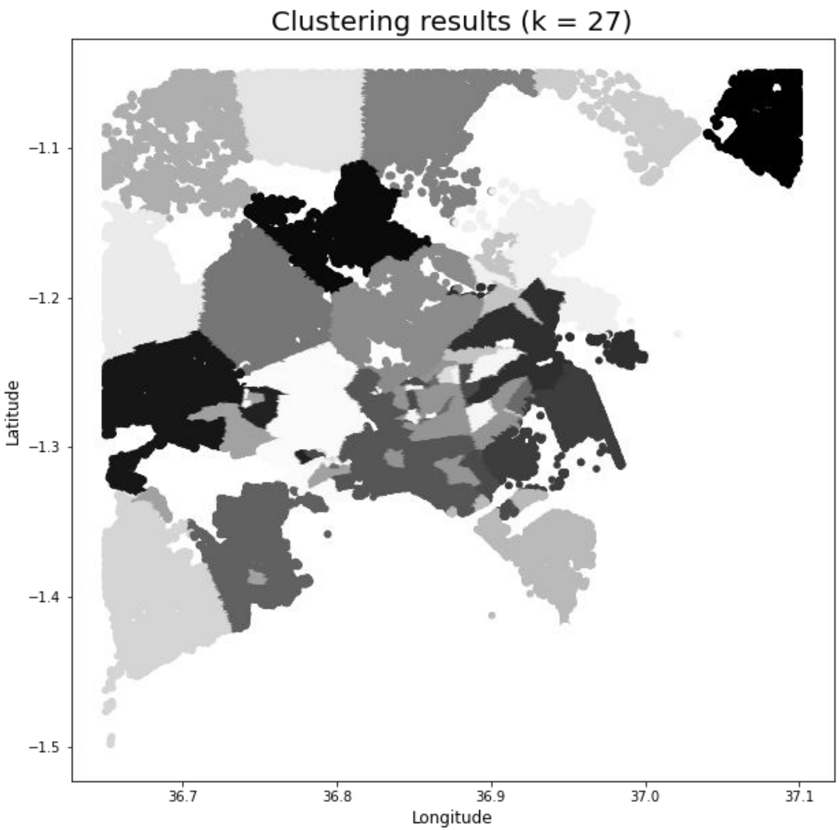
- DBSCAN
  - Inefficient

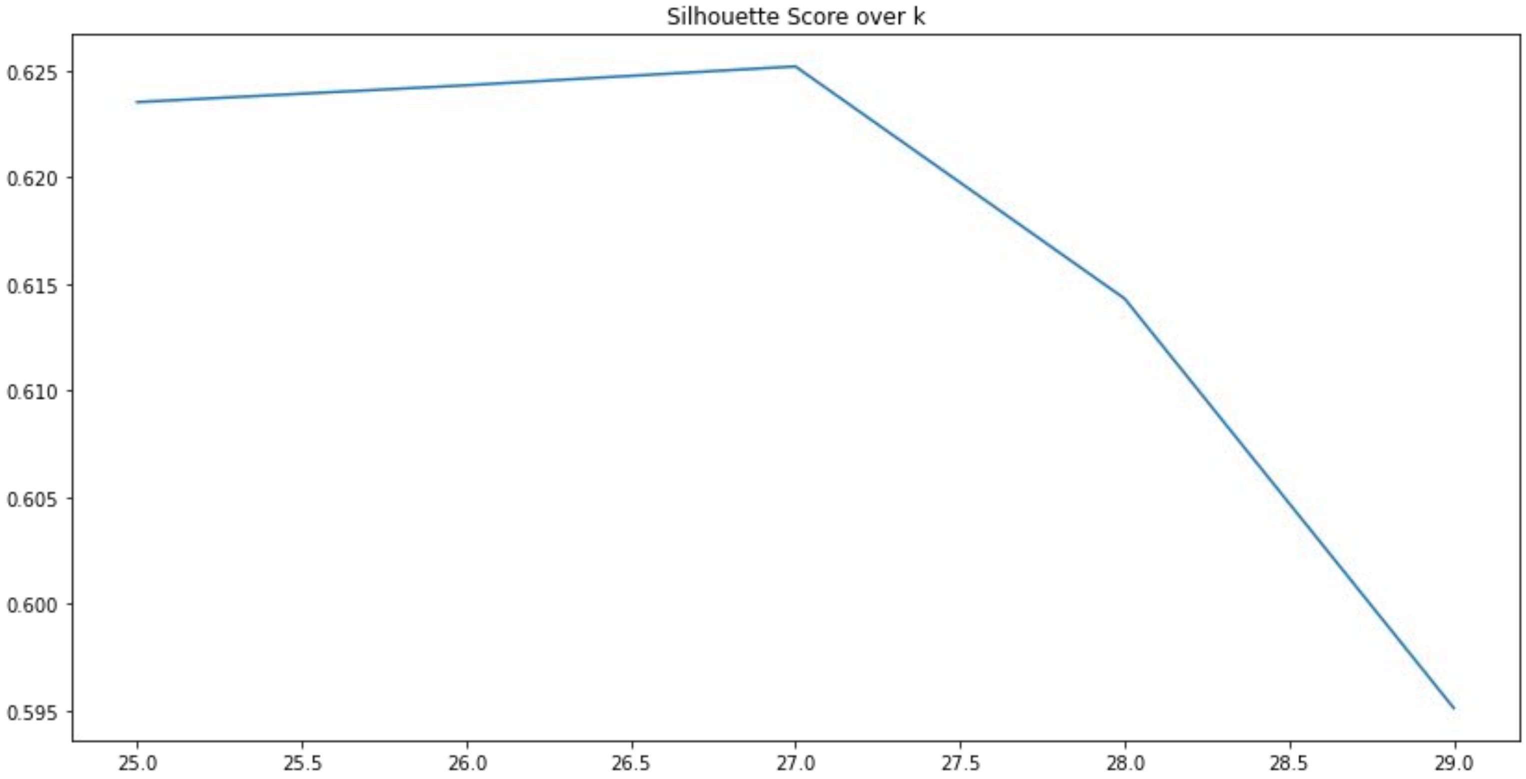
Metrics

Silhouette Score

- Random sample of 20,000 points
- Score: 0.5404

Clusters


Clustering results (k = 27)
−1.1
−1.2
−1.3
−1.4
−1.5
Latitude
36.7
36.8
36.9
37.0
37.1
Longitude



Silhouette Score over k
0.625
0.620
0.615
0.610
0.605
0.600
0.595
25.0
25.5
26.0
26.5
27.0
27.5
28.0
28.5
29.0

Populations

clusters	Population
12	521.097123
20	375.180868
2	270.542778
13	186.906127
23	123.269341
6	89.214281
11	81.746511
18	61.653474
14	42.249595
21	31.883960
1	23.766944
16	22.951393
0	13.384745
5	13.249545
8	11.840499
10	11.612038
27	10.902056
3	9.534588
26	8.762661
25	8.576788
17	8.421819
24	8.347612
15	7.645986
9	7.336330
4	6.843999
22	6.304768
7	6.280758
19	5.575023

Ratios

clusters	advert ratio
25	38.010932
17	28.080450
18	20.143562
10	19.072199
12	17.667609
3	15.484693
7	13.905433
16	11.486705
24	10.304435
6	8.677806
13	7.531674
9	7.285023
5	7.013444
22	6.514870
15	4.454998
26	4.089884
8	3.891865
23	2.897026
21	2.691377
19	2.662652
11	1.893319
14	0.552008
1	0.413862
2	0.388768
20	0.330199
4	0.186971

Conclusion

Conclusion

- Real estate data extracted spanned affluent neighborhoods in Nairobi, with majority of advertisements located in Westlands county.
- Using mapping and modeling techniques, we created a feature that would help infer probable locations for informal settlements.

Next Steps

- Expand webscraping to other languages and sites
- Develop historical/time series real estate data
- Leverage other data to refine feature
- Test feature with existing machine vision models
- Create a model to predict informal settlements based on generated feature(s) for comparison

Thank You!

Khadija Conteh
Matt DeVay
Aidan Dominguez