FINDING A PLACE TO LIVE

The power of analytics

FEBRUARY 16, 2021

JOACHIM ERIKSSON

Original work created for the Coursera Data Science Capstone assignment

Table of Contents

Introduction: 2
Problem statement: 2
Data: 3
Data Gathering and Preparation: 3
Fetching Boroughs from Wikipedia 3
Coordinates with GeoPy 3
Venues from Foursquare 4
Feature engineering 4
Features 4
Dimensionality reduction using Principal Component Analysis (PCA) 4
References: 5

Introduction:

One of the most sought-after places to live in Norway is Oslo with its cultural diversity, great business opportunities and wide offers in different activities and parks. Before I moved here, I was quite uncertain of where to settle down in Oslo and heard that there were some differences to the boroughs, however I was not able to find any information about this before moving. As we all know, moving can be quite stressful and reducing the time to find out the differences of each area could be quite helpful. In addition to having information about activity offers, the population density is also a factor that plays a role in the general activity level where you live, as more people could mean more noise. I therefore included population density to help highlight this and make it possible to do decisions based on multiple variables for each individual reading as preferences in where to live is highly subjective.

In short, I believe this analysis could be helpful for people moving to Oslo, so I wanted to use the Foursquare API to show venues, parks, and people density to make the difference of these boroughs more distinguishable.

Problem statement:

How do we distinguish the boroughs in Oslo so that a person moving to this city may get an understanding of activity offers and people density that differs each borough to aid him or her choosing where to live?

Data:

The dataset I used for this project is generated using GeoPy[1)] to retrieve coordinates for each borough, the Foursquare API[2)] for its venues and parks, and a table of boroughs with population in Oslo from Wikipedia[3)]. *Please visit references for details.*

Data Gathering and Preparation:

Fetching Boroughs from Wikipedia

All 15 boroughs where simply retrieved from Wikipedia using pandas read_html method and required renaming to English column headers and the population field to be converted into float values.

	Borough	Population	Square KM	Borough ID
0	Alna	49 801	137	12
1	Bjerke	33 422	77	9
2	Frogner	59 269	83	5
3	Gamle Oslo	58 671	75	1
4	Grorud	27 707	82	10
5	Grünerløkka	62 423	48	2
6	Nordre Aker	52 327	136	8
7	Nordstrand	52 459	169	14
8	Sagene	45 089	31	3
9	St. Hanshaugen	38 945	36	4
10	Stovner	33 316	82	11
11	Søndre Nordstrand	39 066	184	15
12	Ullern	34 569	94	6
13	Vestre Aker	50 157	166	7
14	Østensjø	50 806	122	13

Coordinates with GeoPy

With GeoPy I fetched all 15 Boroughs' corresponding latitude and longitude values to be used for the Foursqare API calls.

	Borough	Population	Square KM	Borough ID	lat	lng
0	Alna	49801.0	137.0	12	59.932417	10.835276
1	Bjerke	33422.0	77.0	9	59.940668	10.808725
2	Frogner	59269.0	83.0	5	59.922224	10.706649
3	Gamle Oslo	58671.0	75.0	1	59.899237	10.734767
4	Grorud	27707.0	82.0	10	59.961424	10.880549
5	Grünerløkka	62423.0	48.0	2	59.923856	10.757889
6	Nordre Aker	52327.0	136.0	8	59.953638	10.756412
7	Nordstrand	52459.0	169.0	14	59.864561	10.786143
8	Sagene	45089.0	31.0	3	59.936887	10.755306
9	St. Hanshaugen	38945.0	36.0	4	59.927950	10.738958
10	Stovner	33316.0	82.0	11	59.962140	10.922823
11	Søndre Nordstrand	39066.0	184.0	15	59.835944	10.798496
12	Ullern	34569.0	94.0	6	59.925567	10.655798
13	Vestre Aker	50157.0	166.0	7	59.958300	10.670319
14	Østensjø	50806.0	122.0	13	59.887563	10.832748

Venues from Foursquare

I used the /explore endpoint of the Foursquare API in order to retrieve max 50 venue names and categories per borough as seen below:

	index_col	Borough	venue_name	venue_category	venue_lat	venue_lng
423	8	Sagene	Bombay Cuisine	Indian	59.929423	10.760781
331	6	Nordre Aker	Mat & Mer	Deli / Bodega	59.940046	10.759173
12	0	Alna	en til! pizza	Pizza	59.942593	10.814542
593	11	Søndre Nordstrand	Godt Brød	Bakery	59.907029	10.757393
716	14	Østensjø	Ulvøya	Beach	59.869724	10.772059
596	11	Søndre Nordstrand	Østre Greverud Idrettshall	Athletics & Sports	59.773655	10.815591
28	0	Alna	Harald Huysman Karting	Racetrack	59.919503	10.836280
357	7	Nordstrand	Fiskevollbukta	Beach	59.842372	10.777234
711	14	Østensjø	Fuglen Coffee Roasters Oslo	Coffee Shop	59.906185	10.774646
279	5	Grünerløkka	East Kitchen	Asian	59.920755	10.757422

The shape of this dataframe is 750 rows long and 6 columns wide.

Feature engineering

Features

With the current data on different venues I decided to one-hot encode venue categories to create features and find the frequency of all features per borough.

	Borough	index_col	venue_lat	venue_lng	Advertising Agency	Amphitheater	Apparel	Art Gallery	Art Museum	Arts & Crafts	...	Theme Park	Theme Restaurant	Track	Trail	Train Station	Wat
0	Alna	0	59.930589	10.821124	0.0	0.0	0.00	0.00	0.00	0.02	...	0.0	0.00	0.0	0.0	0.0	
1	Bjerke	1	59.939794	10.787494	0.0	0.0	0.00	0.00	0.00	0.00	...	0.0	0.00	0.0	0.0	0.0	
2	Frogner	2	59.920573	10.714704	0.0	0.0	0.00	0.00	0.02	0.00	...	0.0	0.00	0.0	0.0	0.0	
3	Gamle Oslo	3	59.908831	10.739617	0.0	0.0	0.02	0.04	0.00	0.00	...	0.0	0.02	0.0	0.0	0.0	
4	Grorud	4	59.944661	10.858087	0.0	0.0	0.00	0.00	0.00	0.02	...	0.0	0.00	0.0	0.0	0.0	

The dataset then became 15 rows long and 127 columns wide.

Dimensionality reduction using Principal Component Analysis (PCA)

As the dataset became quite rich in features, but with a low number of samples I wanted to see if I could extract the most important features from the data so the model would perform better. PCA therefore became a choice in reducing the dimensionality and number of features was reduced from 127 to 15 where 95% of the variance was explained. The data was also scaled using sklearn StandardScaler beforehand to normalize values for a more accurate model.

References:

1) **Geopy Python Library:** https://geopy.readthedocs.io/en/stable/#
2) **Foursquare API:** https://developer.foursquare.com/docs/api-reference/venues/explore/
3) **Wikipedia, List of Boroughs in Oslo:** https://no.wikipedia.org/wiki/Liste_over_Oslos_bydeler