Classification of Greater Sydney's Suburbs under Urbanisation

IBM Coursera Data Science Final Capstone Project

Week 2 Submission

James Berry, 12/04/2020

Introduction and Project Overview

Sydney is a major commercial city in Australia with consistent annual population growth. The Transport Authority " Transport for New South Wales (TfNSW)" posted population projection data for the city from 2016-2056 which highlights spatial distribution of projected population and its implications.

As the city grows and develops, it becomes increasingly important to examine and understand it quantitatively. Entrepreneurs, Investors, City Planners and Developers have an interest in identifying early opportunities and growing urban footprint in underdeveloped neighborhoods.

We will use population projection data and foursquare API for following analysis:

1. Classifying neighborhoods as highly developed, downtown and less/under developed

2. Understanding how urban footprint of Sydney will expand

3. Exploring underdeveloped neighborhoods looking at remarkable population growth

4. Identifying business opportunities in urbanizing neighborhoods

Data

We need data from reliable sources for analysis. To understand our problem and quantify results we will use following data:

1. Population Projection https://opendata.transport.nsw.gov.au/dataset/population-projections

2. Foursquare Developers Access to venue data: https://foursquare.com/

Methodology

The methodology will include:

- Data retrieval, exploration and wrangling
- Performing K-means clustering algorithm to segment neighborhoods
- Visualizing population projections and neighborhood segments
- Understand growth pattern and urban shift

Data retrieval, exploration and wrangling

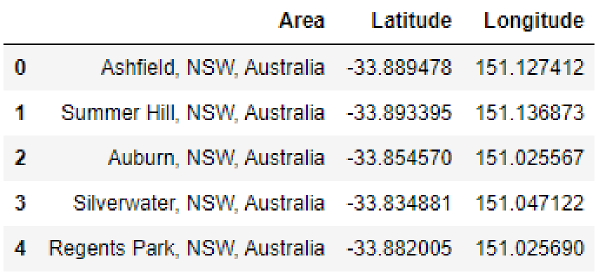
1. The population dataset, an excel file, is provided by TfNSW Open Data Hub under Creative Commons Attribution 4.0 International (CC BY 4.0) Licence. It aggregates population projections for different geographical divisions from 2016 to 2056. Spatial geographies used in this dataset are:

- **Sydney Greater Capital City Statistical Area (SGCCSA)** – This excludes areas of NewCastle and Wollongong from Sydney Greater Metropolitan Area. We will analyze only this area and henceforth refer to it as just Sydney.
- **Travel Zones (TZ)** – There are 2345 TZ's.
- **Statistical Areas 2 (SA 2)** – This geography is approximately the same size of a suburb and can be useful for reporting and reviewing of results at a local neighborhood level. This is the area we will use for our analysis and refer to it as neighborhoods. There are 249 neighborhoods in Sydney.
- **Statistical Areas 2 (SA 3)** – There are 45 SA3s.
- **Statistical Areas 4 (SA 2)** – Sub-regional geography used for collecting demographic data. There are 14 SA4's. However, it is too broad for our analysis to mean anything.
- **GSC District**- There are 6 districts.
- **Local Government Areas (LGA)** - These are political boundaries which may not always align with functional land use areas.

Data of Estimated Population Projection (ERP) from 2016- 2036 is retrieved in a Pandas DataFrame and grouped by SA2 (renamed as Areas).

From this data a new column is created containing % of population growth from 2016 to 2036.

Latitude and Longitude of each neighborhood is retrieved using Geocoder from Geopy Library of Python.

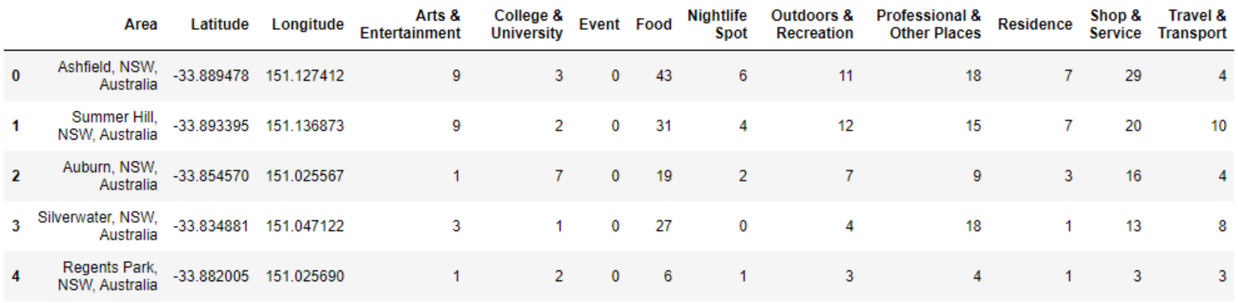


	Area	Latitude	Longitude
0	Ashfield, NSW, Australia	-33.889478	151.127412
1	Summer Hill, NSW, Australia	-33.893395	151.136873
2	Auburn, NSW, Australia	-33.854570	151.025567
3	Silverwater, NSW, Australia	-33.834881	151.047122
4	Regents Park, NSW, Australia	-33.882005	151.025690

2. Foursquare API is used to explore types of venues in each area. Foursquare identifies 10 top level categories. There are multiple sub categories which will not be used it for the time-

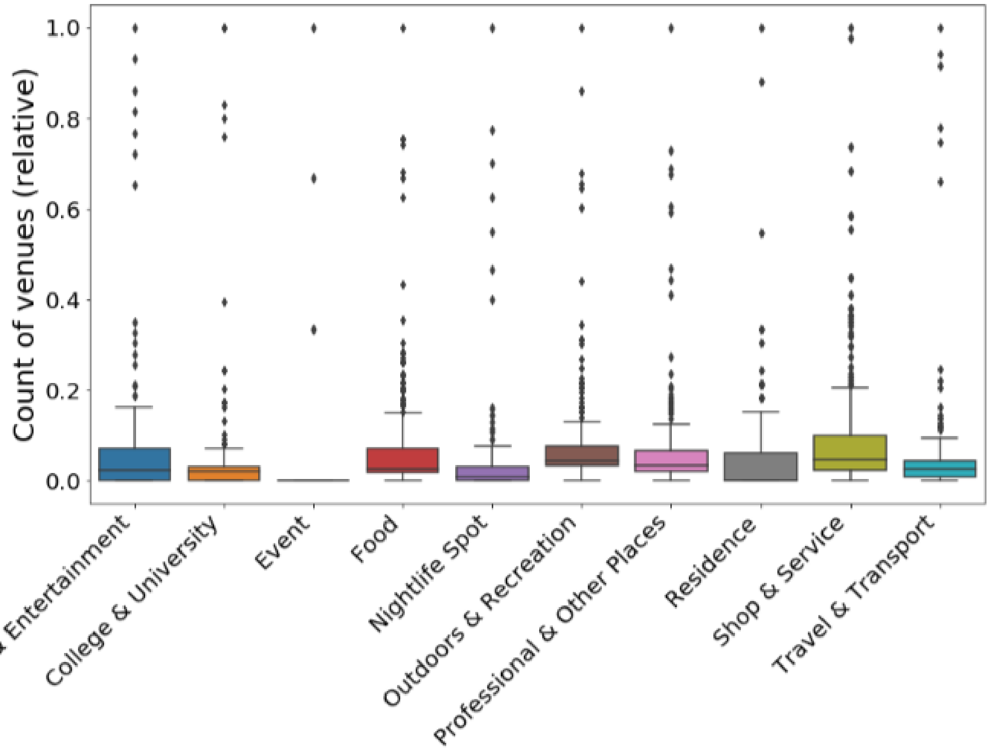
Arts & Entertainment
College & University
Event
Food
Nightlife Spot
Outdoors & Recreation
Professional & Other Places
Residence
Shop & Service
Travel & Transport

This is a snapshot of resulting DataFrame



	Area	Latitude	Longitude	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	Ashfield, NSW, Australia	-33.889478	151.127412	9	3	0	43	6	11	18	7	29	4
1	Summer Hill, NSW, Australia	-33.893395	151.136873	9	2	0	31	4	12	15	7	20	10
2	Auburn, NSW, Australia	-33.854570	151.025567	1	7	0	19	2	7	9	3	16	4
3	Silverwater, NSW, Australia	-33.834881	151.047122	3	1	0	27	0	4	18	1	13	8
4	Regents Park, NSW, Australia	-33.882005	151.025690	1	2	0	6	1	3	4	1	3	3

Following Box Plot shows frequencies of each type of Venue



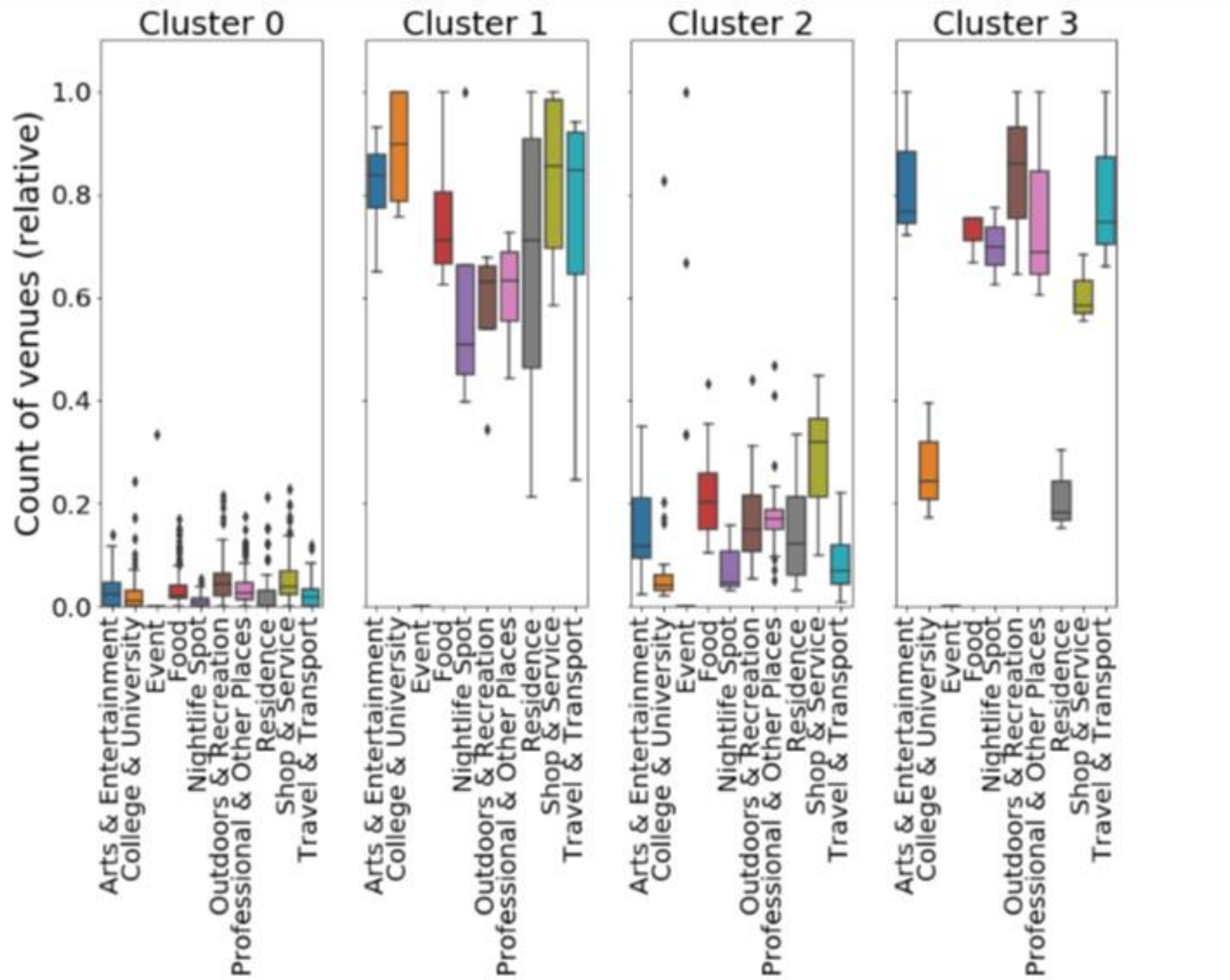
As can be seen, the are a substantial amount of outliers in all categories, which shows a highly skewed development pattern .

Performing K-means clustering algorithm to segment neighborhoods

These were preliminary results of K-means algorithm with different number of clusters:

- 2 clusters only show the uptown/downtown divide
- 3 clusters add clustering within the downtown
- 4 clusters add clustering within uptown (residential and professional) and downtown
- 5 and more clusters are difficult to interpret

Four clusters are a perfect fit for this study. Following box plots give us a visual representation of clusters

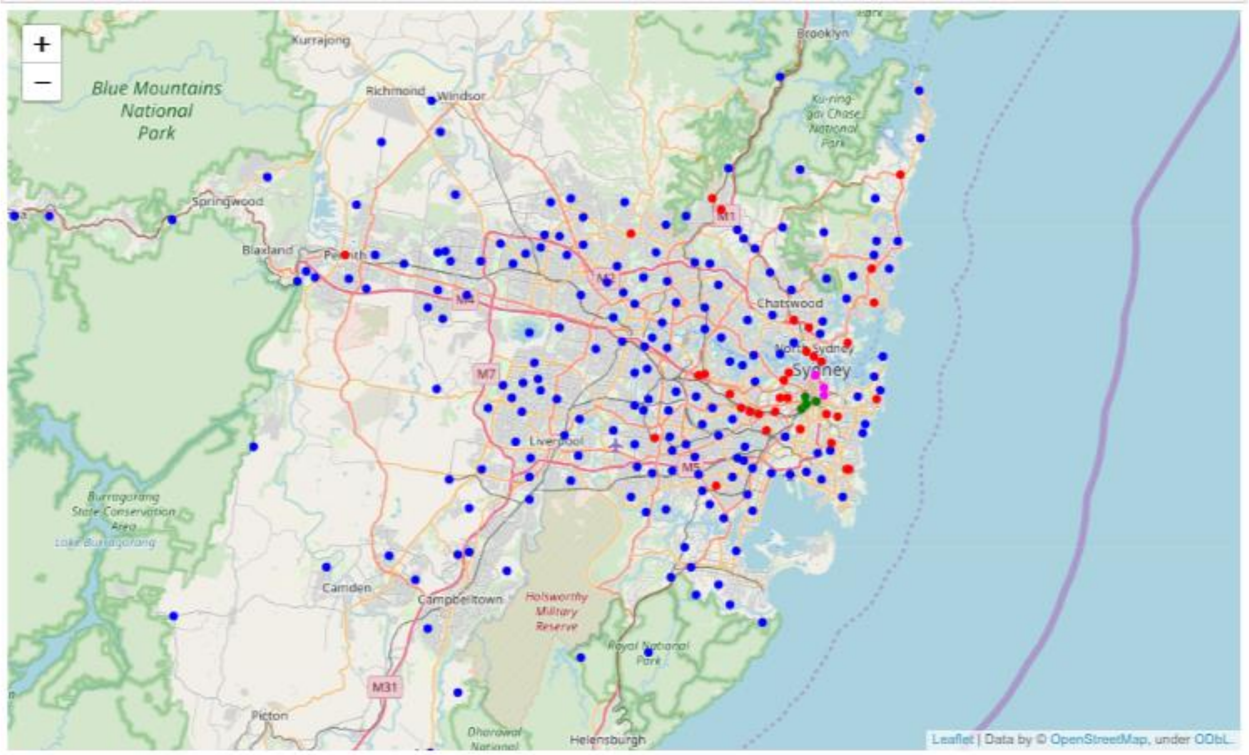


The interpretation of these clusters based on Box Plots above is:

- Cluster 0 (blue) has low frequencies for all venue categories. These appear to be underdeveloped neighborhoods
- Cluster 1 (green) has consistently high frequencies for all venue categories. This is the most diversely developed part of city

- Cluster 2 (red) has moderate scores with shops and services being the most popular. These are developed residential suburbs
- Cluster 3 (magenta) has high frequencies but with less residential places and more professional places. These are the developed professional or industrial suburbs.

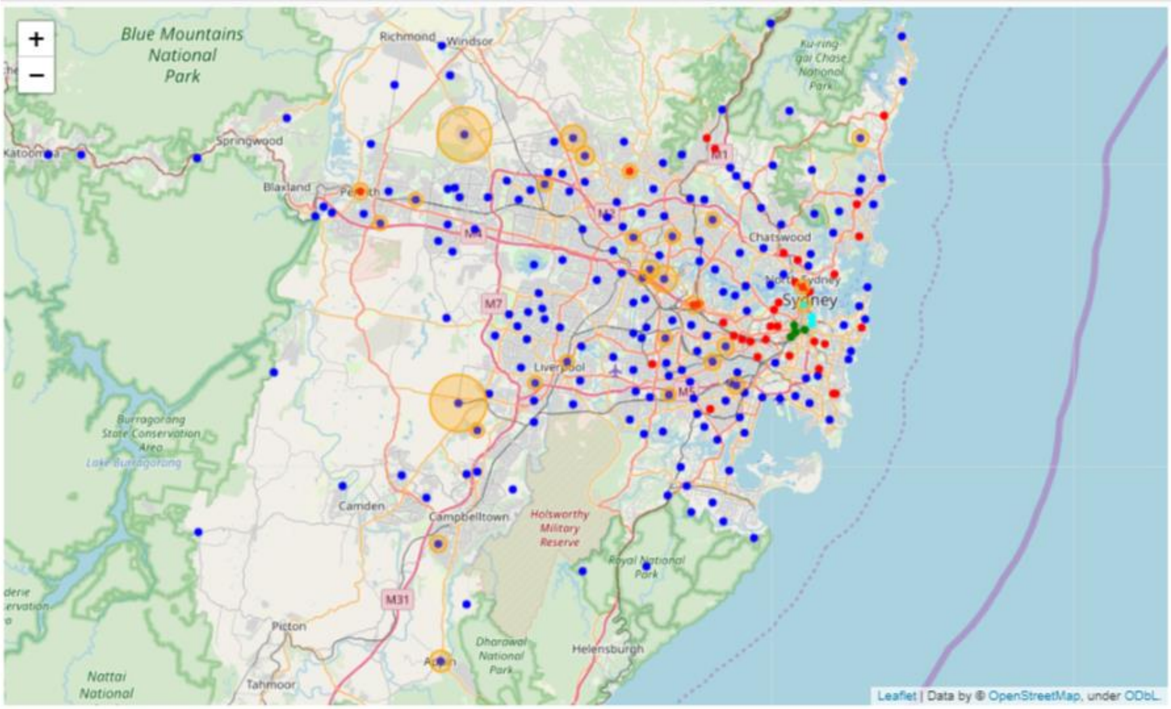
Plotting latitude and longitude and visualising the data set's point confirms that urban development of the city is concentrated in a small zone while most of the outer areas are considerably less developed.



Visualizing population projections data and neighborhood segments

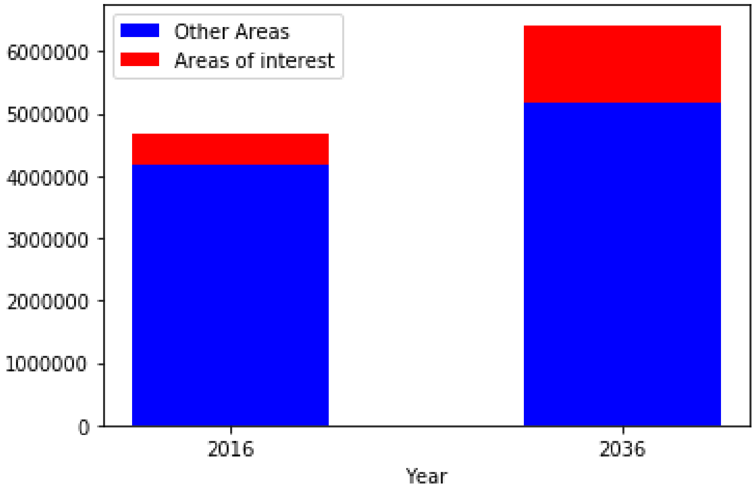
Areas showing more than 60% rise in population are separated into a new DataFrame and selected for analysis. These are named expanding areas.

The following map shows expanding areas as bubbles of varying size, quantifying the amount of population growth on top of clustered neighborhoods in blue, green, red and cyan color. Areas of interest are blue dots with rising population. There are 26 such neighborhoods.



Understand growth pattern and urban shift

How does population growth in areas of interest compare to rest of the city? To understand this we calculate total population in 26 neighborhoods classified in previous step for years 2016 and 2036 and plot a stacked bar graph with total population of the city.



There is a noticeable increase in portion of total population dwelling in areas of interest. This means population is not just growing but also expanding towards these neighborhoods which hints at a shift of urban footprint.

Results

The analysis enabled us to discover and describe visually and quantitatively:

1. Very few neighborhoods of Sydney can be classified as highly developed based on venues data from Foursquare API. However, with population projection data we see that urban footprint of Sydney is expected to expand as new growth areas establish to the **North West and South West** of Sydney. Small scale incremental developments will provide for population growth within the existing urban footprint.

2. We identified 26 underdeveloped neighborhoods, mostly to North West, whose contribution to the city's total population will rise from 10.7% to around 20%. This shows not just growth but also a shift of population towards these areas

3. Mardsen Park and **Leppington** emerge as the major growing areas and might even see a transformation from underdeveloped to gentrified desirable suburbs as its occupancy increases.

4. Predicting this growth pattern leads us to identify early business and service opportunities in currently underdeveloped areas. However, to understand its full impact we must categorize population data further. However, growth on this scale will have fundamental implications on venues belonging to categories **Residence, Shop & Service and Food** which benefit positively from population shift and expansion.

Discussion and Recommendations

Our aim was to identify opportunities in underdeveloped neighborhoods in Greater Sydney that are expected to grow in the future.

It would be interesting to further study how population segments disaggregate by private and non-private dwellings, age-sex, household types, enrolments and workforce status. ABS census data and ABS Laborforce Survey Data is available which can be used for further analysis and obtaining deeper insights into the future of these neighborhoods.

Conclusion

Using a combination of datasets from the New South Wales Open Data project and Foursquare venue data we were able to analyze, discover and statistically describe neighborhoods and population expansion quantitatively to identify some neighborhoods of interest.

More data from the NSW Open Data about employment, workforce, population segmentation should be used for true valued quantitative analysis and predictive analytics which would be most valued by investors and developers to guide them in their investment appraisal and decision making process.