

Bus Tours inc.


International bus tour company expansion proposal

Introduction

- An international bus tour company wants a way to expand to new cities.
- They are offering bus tours to popular trending locations of a given city over 5 days.
- In order to give them a starting point they want a model that will take the trending locations of each neighborhood and cluster them into similar clusters based on k-means.
- The bus tour consists of 5 days so they will need 5 clusters of venues, one for each day of the tour.

Acquire Neighborhood GPS location data

- The City of Vancouver was used as an example for this model
- The data was acquired from the city of Vancouver open data catalogue <https://data.vancouver.ca/datacatalogue/localareaboundary.htm>








CITY OF VANCOUVER

[Return to data catalogue index](#)
[Return to Open Data home page](#)
[Terms of Use](#)

Open Data Catalogue

Local area boundary

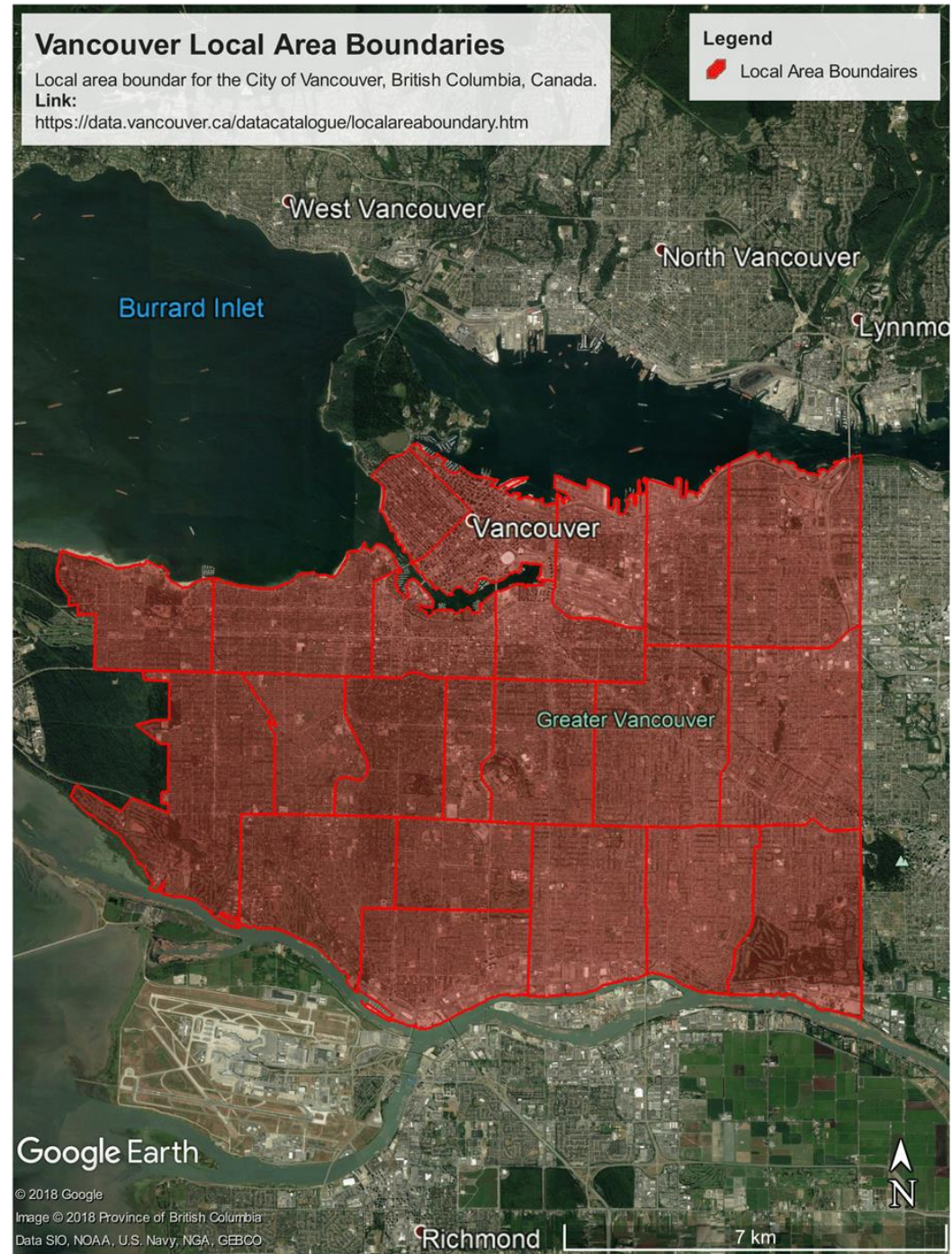
Data custodian	IT Applications - GIS and CADD Services Planning and Development Services - Research and Data
Data currency comments	These boundaries do not change.
Data set description	This data set contains the boundaries for the City's 22 local areas (also known as local planning areas).
Data accuracy comments	Local area boundaries generally follow street centrelines; centrelines are in the approximate centre of streets.
Attributes	Official name and boundaries
Coordinate system	N/A
Data set details	<ol style="list-style-type: none">1. Local Area Boundary (KML) 2. Local Area Boundary (SHP) 3. Local Area Boundary (Google Map) 4. Local Area Boundary (XLS) 5. Local Area Boundary (CSV)  <p>Note: .XLS and .CSV formats contain only names of the local areas</p>

© 2019 City of Vancouver

[Terms of Use](#) | [Privacy policy](#) | [Website accessibility](#)

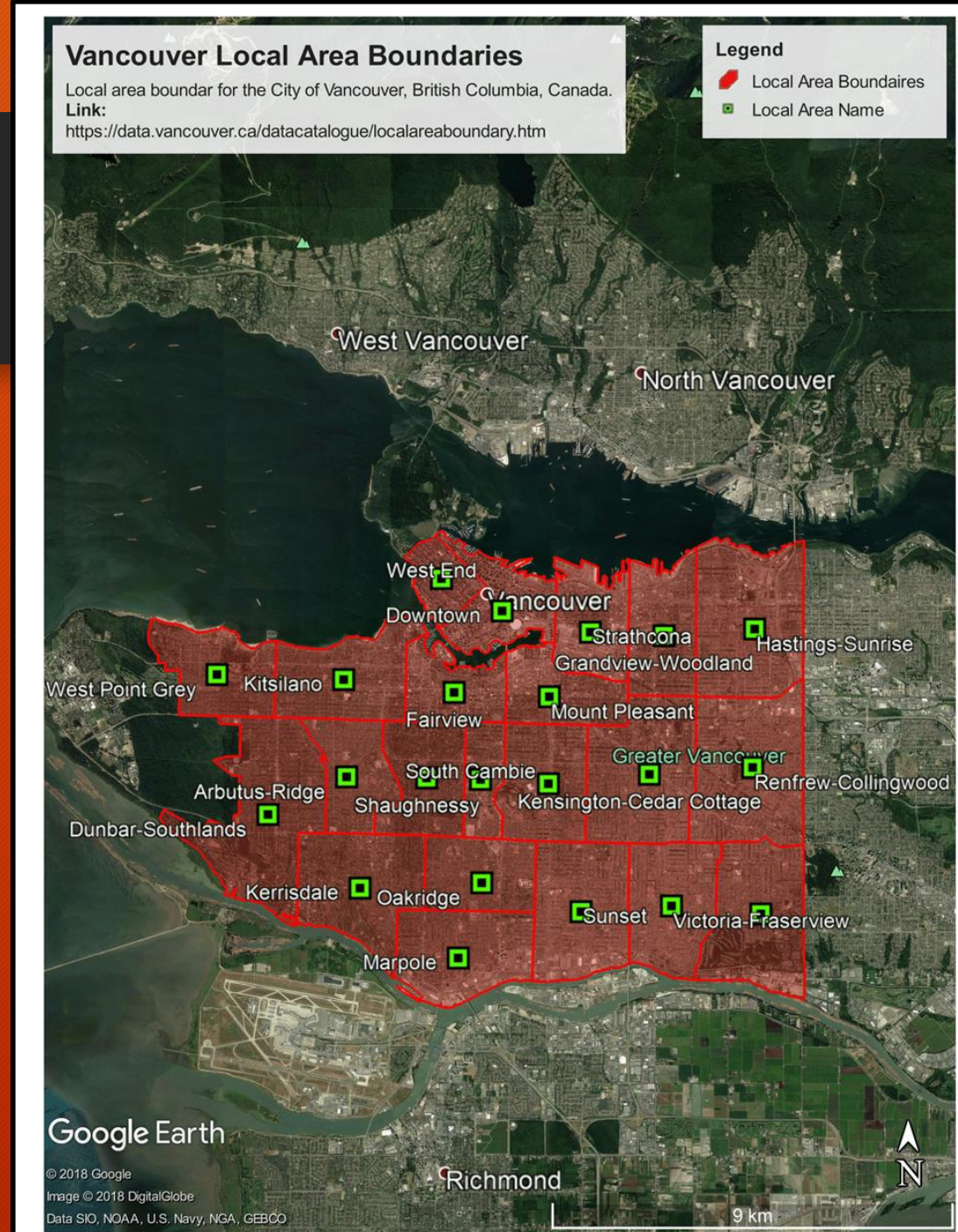
Raw data, Vancouver

- Raw data acquired from City of Vancouver Website website.



Clean the data

- find the center points to be used in the model



convert to useable format

- <http://www.gpsvisualizer.com/>
- use website to convert KML in to CSV

```
Van_data.head()
```

	latitude	longitude	name
0	49.246316	-123.163438	Arbutus-Ridge
1	49.279594	-123.115711	Downtown
2	49.238770	-123.187580	Dunbar-Southlands
3	49.263254	-123.130439	Fairview
4	49.274615	-123.065973	Grandview-Woodland

Use Foursquare To Find Trending Venues

- use the Foursquare call “Explore Section = Trending” end points
- Use one hot encoding and group by most common venue

	name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Arbutus-Ridge	Coffee Shop	Sandwich Place	Grocery Store	Sushi Restaurant	Chinese Restaurant
1	Downtown	Coffee Shop	Japanese Restaurant	Restaurant	Seafood Restaurant	Sandwich Place
2	Dunbar-Southlands	Golf Course	Sushi Restaurant	Coffee Shop	Bakery	Park
3	Fairview	Japanese Restaurant	Restaurant	Coffee Shop	Bakery	Café
4	Grandview-Woodland	Coffee Shop	Brewery	Pizza Place	Sushi Restaurant	Café

Run K-Means Clustering

- 5 clusters
- random state = 0

```
In [24]: # set number of clusters
kclusters = 5

Van_grouped_clustering = Van_grouped.drop('name', 1)

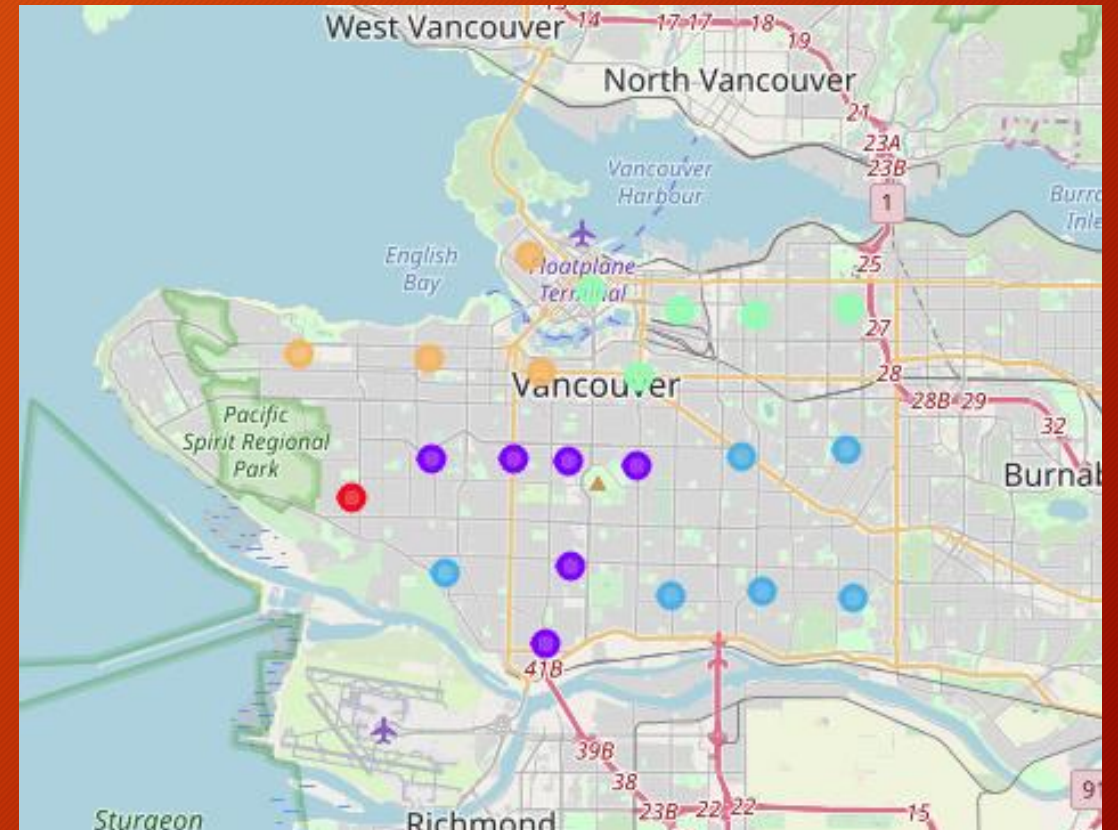
# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(Van_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[:]
```

```
Out[24]: array([1, 3, 0, 4, 3, 3, 2, 2, 2, 4, 1, 3, 1, 2, 1, 1, 1, 3, 2, 2, 4, 4],
              dtype=int32)
```


Results

- Display Clusters on Folium Map
- 5 clusters
- One cluster for each day of the bus tour



Discussion

- A bus tour company can apply this model to any city and produce a starting point for a bus tour company with no prior knowledge of the city.
- Future refinement of the model could use the foursquare API to acquire the '**top picks**' for each cluster, allowing for better decision making when planning out the route of the bus.
- One problem with this model could be that that the foursquare **explore section = trending** end point category may not be a good representation of what tourists want to see. Using other endpoints such as **explore section = outdoors** or **explore section = sites** may be a better solution.

Conclusion

- This model could be applied to any city where the GPS locations of a neighborhood are known. As it stands the model breaks the neighborhoods into 5 clusters of similar trending venues. The bus tour will be 5 days long and each day will be spent in a cluster.
- Further refinement of the model could help choose what venues to visit in each cluster, by using the top picks end point. This model will cut down on research time and allow a company to expand faster than the competitors in theory. This could also be applied to individuals who would like to go traveling, but are unfamiliar with a given city.

Resources

- Vancouver Neighborhoods data:
<https://data.vancouver.ca/datacatalogue/localareaboundary.htm>
- Van_hood.csv data used in this model:
https://github.com/lmuller92/Van_hoods
- Code for bus tour model: <https://github.com/lmuller92/IBM-Final-Capstone-Project/blob/master/Trending.ipynb>