Neighborhood scouting for real estate

Coursera - Capstone Project - The Battle of Neighborhoods

2019

Interest of neighborhood scouting

Neighborhood scouting is very beneficial for a real estate company:

- Give reactivity to real estate agents to discover opportunities.
- Ensure that the future property will better fit the client's needs.
- This makes agents more active than passive in the real estate market.

Use case:

- Identify top-5 neighborhoods in NYC, Toronto, San-Francisco, Chicago.
- Criteria distributed in 2 priorities: High / Low
  - Low crime rate.
  - Close to services, transport, stadiums, medical venues, cultural venues, natural areas.
  - Wide variety of restaurants, bars, pubs.
  - Few burglaries.
  - Close to antique shops.

Data

- Source
  - Neighborhood names and locations from Coursera or Wikipedia + Geopy.
  - Crime rate (2018) from Kaggle and Toronto Police data portal.
  - Venues locations and categories from Foursquare (top-100 most common).
- Cleaning and processing
  - Drop missing values and duplicates.
  - Venue categories related to criteria were gathered in 8 groups.
  - Compute average or minimum distance between neighborhoods and venues.
- Features
  - Apply log transformation and scale data ([0-1]).
  - Apply 1-x transformation to crimes and distances > objective is to maximize all features.
  - Priority 1: 9 features (crimes, distinct venues, average/min. distances to venues).
  - Priority 2: 2 features (count burglaries, antique shop minimum distance).
  - 573 rows (neighborhoods).

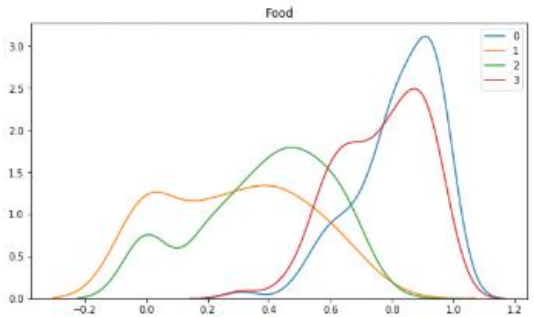
Methodology

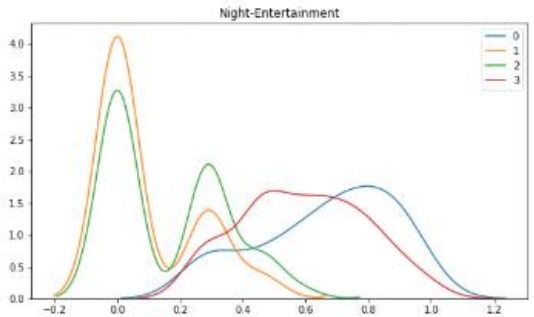
- Clustering neighborhoods by taking into account the 9 features of priority 1.
- Identify the cluster which best fits the future buyer's needs (priority 1).
- Select the top-5 best neighborhoods from the best cluster by minimizing the 2 features of priority 2 ➔ low burglaries + low minimum distance to antique shops.

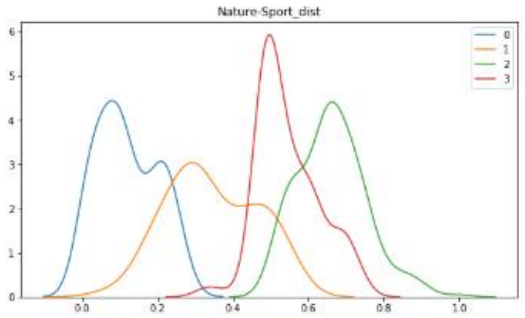
Clustering

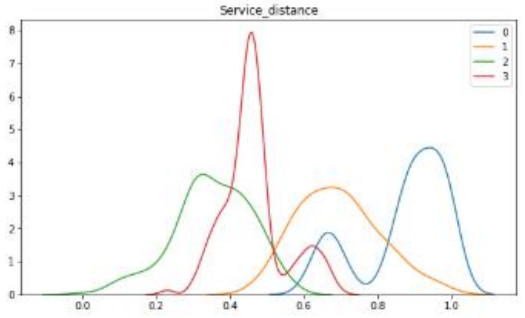
- Clustering neighborhoods into 4 groups.
- Compared to other values of k, feature distributions are quite well separated between clusters for k=4.

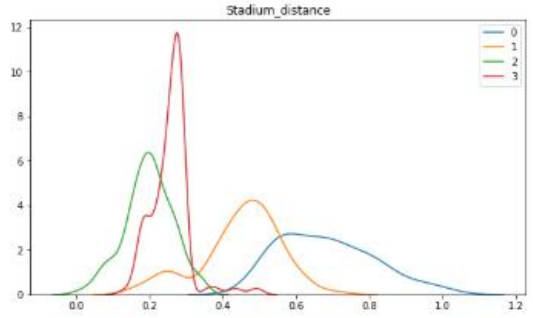
Feature distribution for each cluster:

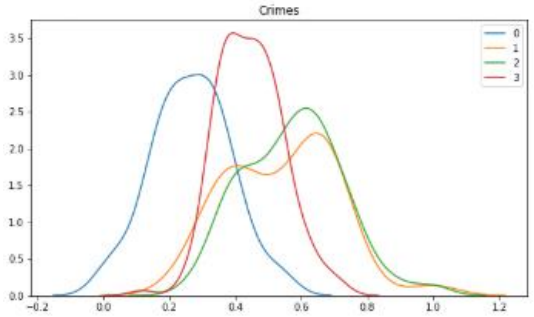


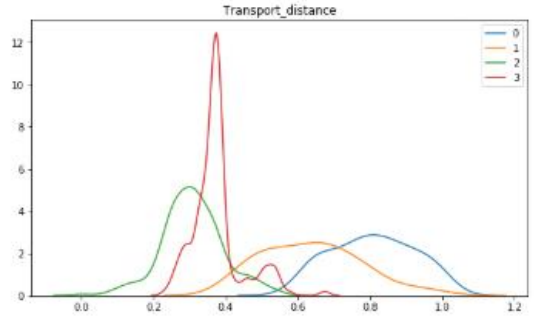


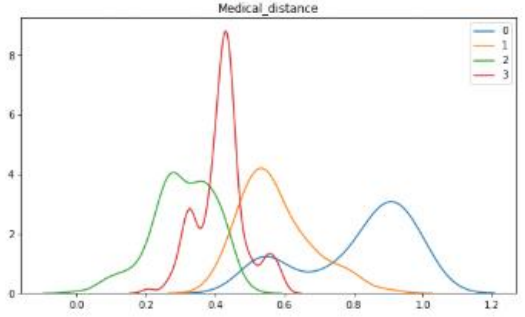


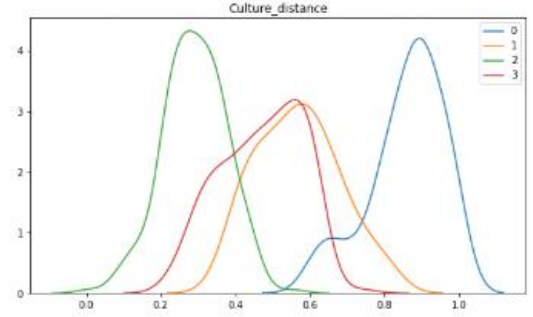




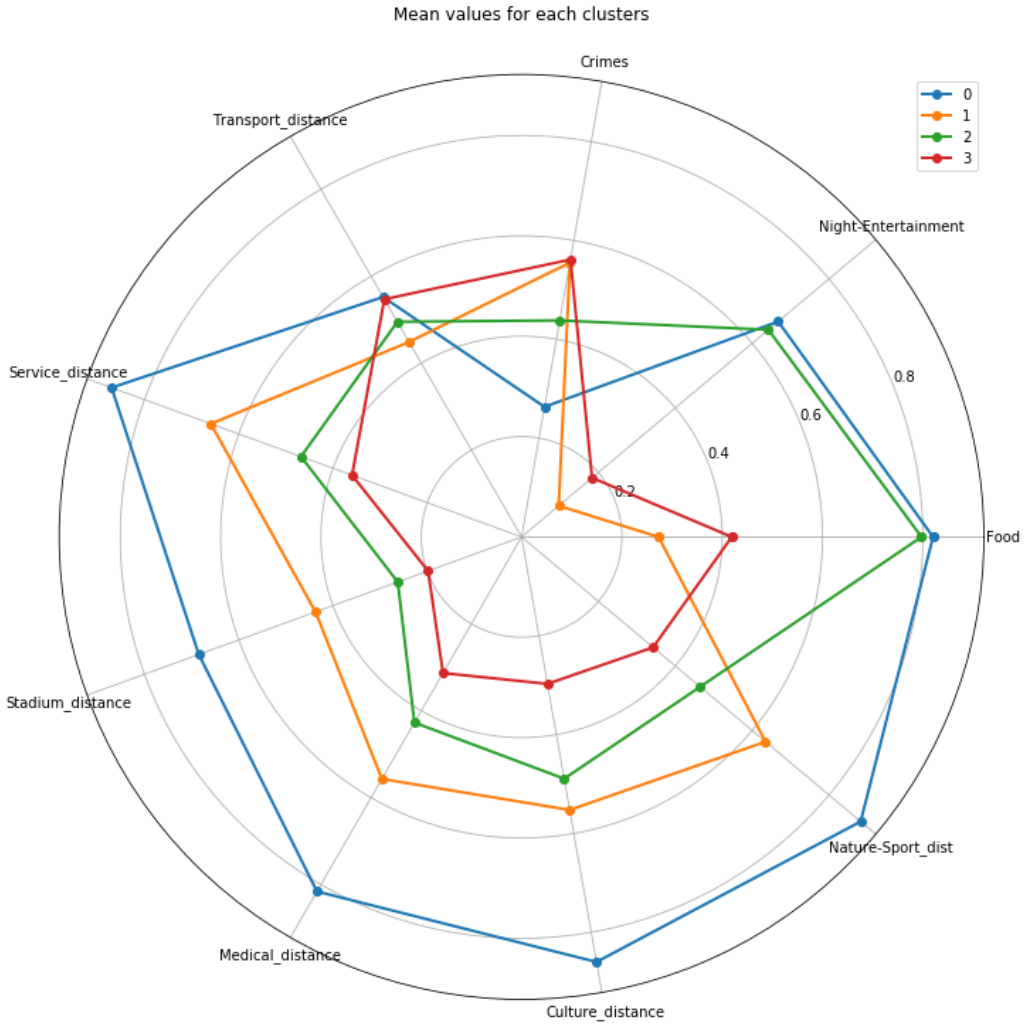








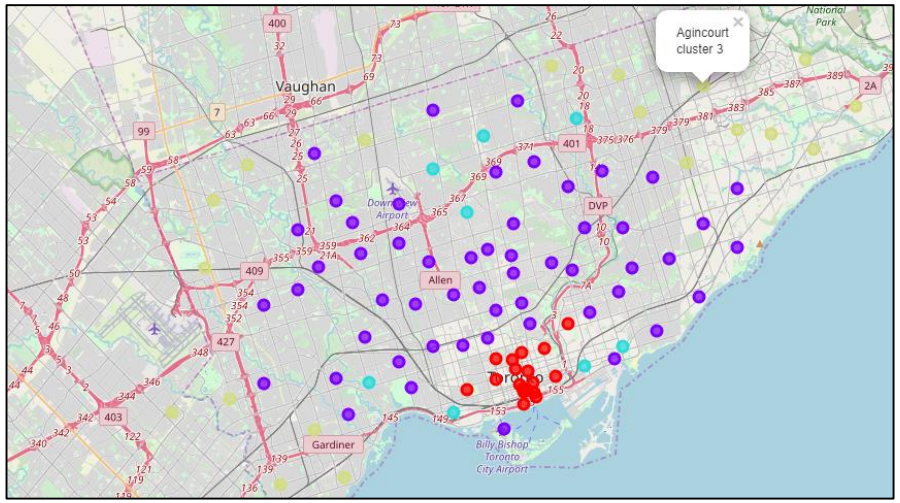
Clustering



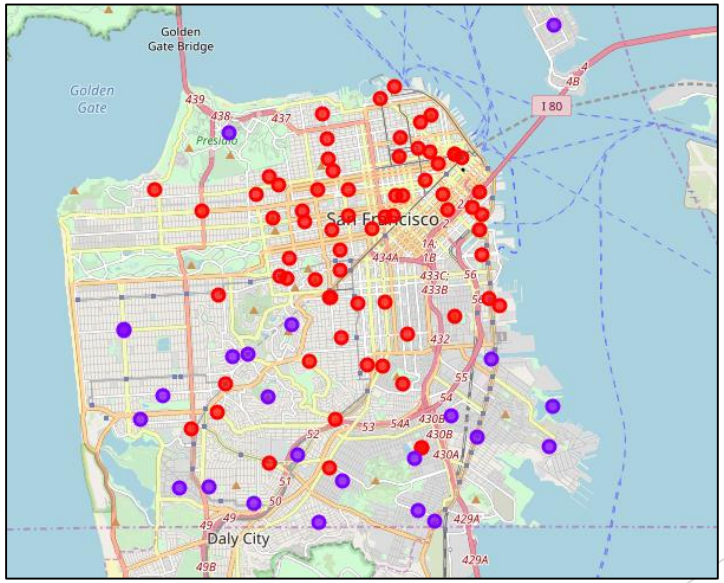
- Group neighborhoods by cluster and compute the mean for each feature
- Best cluster maximize all features.
- Group 0 was considered as the best compromise that meets the customer's needs
- 87 neighborhoods in cluster 0.
- Disadvantage of group 0: higher crime rate.

Clustering

- Group 0 is present only in San-Francisco & Toronto
- Higher geographical extent in SF for cluster 0.
- Cluster 0 mainly in very urbanized areas.

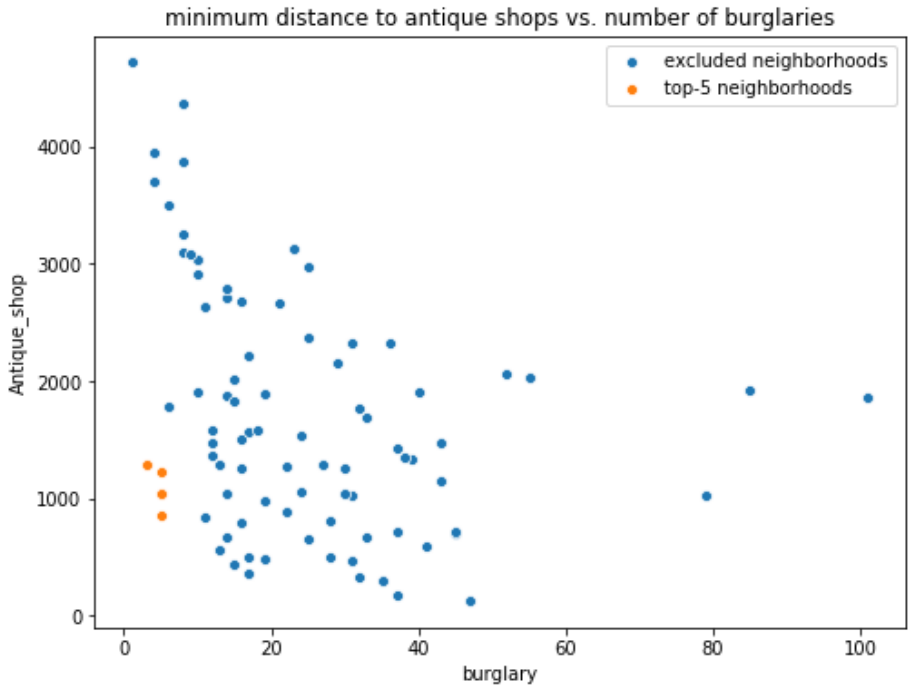


Toronto



San-Francisco

Top-5 neighborhoods selection



- 87 neighborhoods of group 0 were plotted.
- Select the 5 neighborhoods with the lowest min. distance to antique shops and the lowest burglary rate.
- Top-5 neighborhoods are (orange markers):

	Neighborhood	Latitude	Longitude	City
372	First Canadian Place, Underground city	43.648429	-79.382280	Toronto
498	Embarcadero	37.792864	-122.396912	San-Francisco
502	Financial District	37.793647	-122.398938	San-Francisco
503	Financial District South	37.793647	-122.398938	San-Francisco
504	Fisherman's Wharf	37.809167	-122.416599	San-Francisco

Conclusion

Clustering was a good approach to filter neighborhoods by criteria of priority 1.

Top-5 neighborhoods were then easily found among the 87 with a simple graph visualization.

Future directions:

- Neighborhoods have not the same population density:
  ➔ divide the number of crimes by the population.
- To improve the clustering, define a better approach to determine the best k.
- Investigate other features.
- With a free Foursquare account: limited by top-100 venues.
  Try to go further with a professional account.