Data Science Capstone Project

Where to live in Vancouver?

1. Business Problem

Vancouver, a beautiful west coast city in British Columbia, Canada, is always well-known for its continuous present in the list of top cities in the world for quality of living. As a result, Vancouver is a very attractive destination for both immigrants and tourists. Although the city is quite small comparing to other giant metropolitan cities, it stills has 22 official neighborhoods (or areas) and the **decision of where to live (for immigrants) or where to stay (for tourists) is quite a challenging one**.

The purpose of this project is to use location data from Foursquare **to rank and recommend the neighborhoods in Vancouver that best fit to each specific preference of immigrants and tourists**. For example, an immigrant family with kids may value Education, Shop & Service and Outdoor Activities higher than other criteria. On the other hand, a young tourist couple may give Food, Transportation and Nightlife higher weights than other criteria. This project aims to provide recommendation for such specific preferences.

With such problem and solution, this project will offer values to the following stakeholders:

- Immigrants who want to live in Vancouver
- Tourists who want to visit Vancouver
- Real estate agents who want to give recommendations to their clients
- Real estate renting platforms such as Airbnb to help their users searching for neighborhoods that meet their references

2. Data

The following data will be used in this project:

- A list of all neighborhoods (or areas) in Vancouver. This can be retrieved from the city's official website at https://vancouver.ca/news-calendar/areas-of-the-city.aspx. There are 22 of them totally.
- A map of Vancouver splitting into neighborhoods. This is for reference only. Source: https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Vancouver#/media/File:Stadtgliederung_Vancouver_2008.png

- Latitude and longitude of each neighborhood. This can be retrieved by using the Python library ***geopy***. The data is something like this:

	Neighborhood	Latitude	Longitude
0	Arbutus Ridge	49.246305	-123.159636
1	Downtown	49.283393	-123.117456
2	Dunbar-Southlands	49.237864	-123.184354
3	Fairview	49.261956	-123.130408
4	Grandview-Woodland	49.275849	-123.066934
5	Hastings-Sunrise	49.277830	-123.040005
6	Kensington-Cedar Cottage	49.246790	-123.073475

The result is stored in a CSV file, which is also available on GitHub at: https://raw.githubusercontent.com/uclinux83/Coursera_Capstone/master/vancouver_neighborhood.csv

- Foursquare venues data retrieved by calling Foursquare APIs (mainly the ***/venues/explore*** API endpoint). The returned data will be in json format, which will be processed later using Python.


```
"venue": {
  "id": "49b6e8d2f964a52016531fe3",
  "name": "Russ & Daughters",
  "location": {
    "address": "179 E Houston St",
    "crossStreet": "btwn Allen & Orchard St",
    "lat": 40.72286707707289,
    "lng": -73.98829148466851,
    "labeledLatLngs": [
      {
        "label": "display",
        "lat": 40.72286707707289,
        "lng": -73.98829148466851
      }
    ],
```


- Foursquare venue category hierarchy and ID, which is available at: https://developer.foursquare.com/docs/build-with-foursquare/categories/.

  Categories in the list will be selected and mapped into the following 9 criteria (which will be used by the users to setup their preferences):

  - Entertainment
  - Education
  - Food
  - Nightlife
  - Outdoors
  - Health
  - Religion
  - Shop
  - Transportation

  These category IDs will be used as the input parameter ***categories*** when calling the ***/venues/explore*** API

3. Methodology

The overall approach is to get all venues within or around each Vancouver neighborhood and group them into the 9 “criteria” as listed above. Users can set their specific preferences by giving weight to each criteria. The data model will then rank the neighborhoods basing on the specified preferences. Users can use this ranking as recommendation on where they should live or stay in Vancouver basing on their individual references.

The whole process of data collection and data analysis can be divided into 4 parts. At the end of each part, data will be saved into files to make it easier to continue the next part.

Part 1: Retrieve the list of all neighborhoods in Vancouver and get their coordinates

First of all, we need to get all of the neighborhoods in Vancouver. Since Vancouver is quite a small city, it has only 22 official neighborhoods (or areas) and the list can be retrieved manually from its official website at https://vancouver.ca/news-calendar/areas-of-the-city.aspx.

The Python library ***geopy*** is used to get geolocation coordinates of all those 22 neighborhoods. The results are stored in the following dataframe:

	Neighborhood	Latitude	Longitude
0	Arbutus Ridge	49.246305	-123.159636
1	Downtown	49.283393	-123.117456
2	Dunbar-Southlands	49.237864	-123.184354
3	Fairview	49.261956	-123.130408
4	Grandview-Woodland	49.275849	-123.066934
5	Hastings-Sunrise	49.277830	-123.040005
6	Kensington-Cedar Cottage	49.246790	-123.073475
7	Kerrisdale	49.220985	-123.159548
8	Killarney	49.218012	-123.037115
9	Kitsilano	49.269410	-123.155267
10	Marpole	49.209223	-123.136150
11	Mount Pleasant	49.264048	-123.096249
12	Oakridge	49.226615	-123.122943
13	Renfrew-Collingwood	49.248577	-123.040179
14	Riley Park	49.244854	-123.103035
15	Shaughnessy	49.246305	-123.138405
16	South Cambie	49.246464	-123.121603
17	Strathcona	49.277693	-123.088539
18	Sunset	49.219093	-123.091665
19	Victoria-Fraserview	49.218980	-123.063816
20	West End	49.284131	-123.131795
21	West Point Grey	49.268102	-123.202643

We then save the dataframe into a CSV file to make it more convenient to be used later. The file is also uploaded to GitHub and available at: https://raw.githubusercontent.com/uclinux83/Coursera_Capstone/master/vancouver_neighborhood.csv

Part 2: Get Foursquare data for each neighborhood

Before calling Foursquare APIs to get all of the relevant venues for each neighborhood in Vancouver, we need to know which Foursquare categories are corresponding to which of the 9 criteria that we defined earlier. Fortunately, Foursquare lists all of its supported categories here: https://developer.foursquare.com/docs/build-with-foursquare/categories/. With that information, we can manually map our criteria with Foursquare category IDs as the following:

Criteria	Foursquare Category ID(s)
Entertainment	4d4b7104d754a06370d81259
Education	4d4b7105d754a06372d81259, 4bf58dd8d48988d13b941735
Food	4d4b7105d754a06374d81259
Nightlife	4d4b7105d754a06376d81259
Outdoors	4d4b7105d754a06377d81259
Health	4bf58dd8d48988d104941735
Religion	4bf58dd8d48988d131941735
Shop	4d4b7105d754a06378d81259
Transportation	4bf58dd8d48988d129951735, 52f2ab2ebcbc57f1066b8b4f, 4bf58dd8d48988d1fe931735, 4bf58dd8d48988d12d951735

Now we call Foursquare APIs to get venues for each criteria around each neighborhood. Since each neighborhood is quite small, a radius of 2km (2,000 meters) is sufficient enough to cover the whole neighborhood. Also because Foursquare API limit only 50 venues to be returned from each API call, we will need to use the ***offset*** parameter to make sure that we get all of the venues. The result is a list of 8,811 venues, which is stored in a dataframe like this:

	Neighborhood	Criteria	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Arbutus Ridge	Entertainment	Dunbar Theater	49.245613	-123.185428	Indie Movie Theater
1	Arbutus Ridge	Entertainment	Dance Co	49.248822	-123.154979	Dance Studio
2	Arbutus Ridge	Education	Point Grey Secondary	49.237441	-123.153967	School
3	Arbutus Ridge	Education	Carnarvon Community School	49.256532	-123.173862	School
4	Arbutus Ridge	Education	St. John's School	49.262445	-123.153701	School

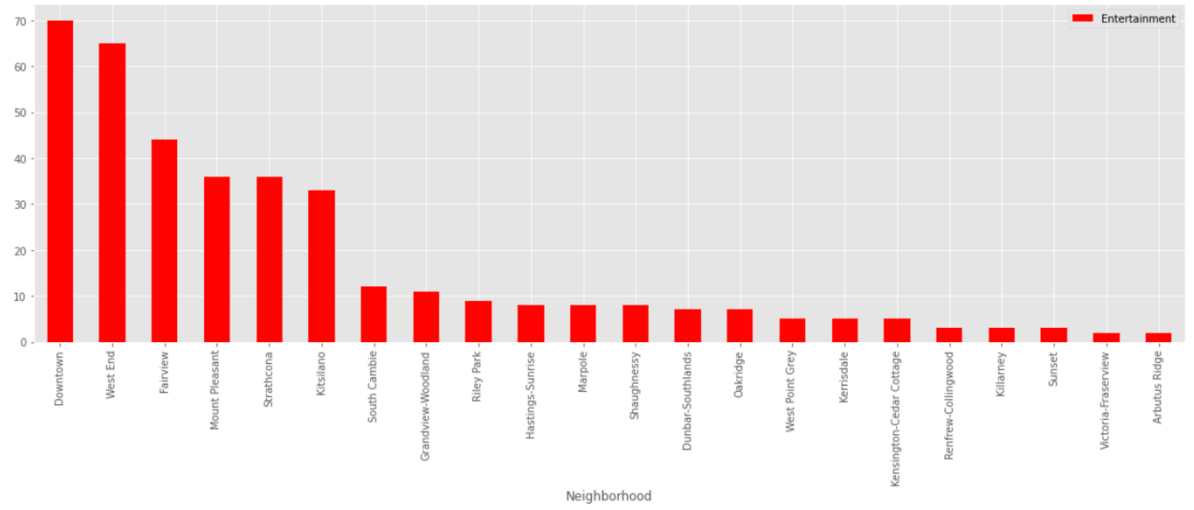
We then save the dataframe into a CSV file to make it more convenient to be used later. The file is also uploaded to GitHub and available at: https://raw.githubusercontent.com/uclinux83/Coursera_Capstone/master/vancouver_venues_details.csv

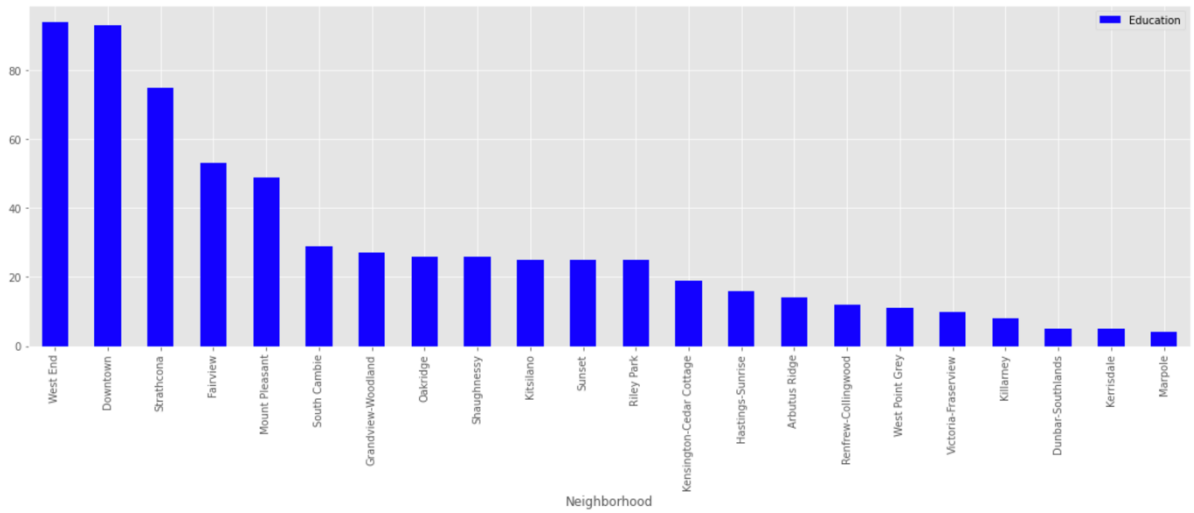
Part 3: Create a summary dataframe to make it easier to calculate score and rank the neighborhoods later

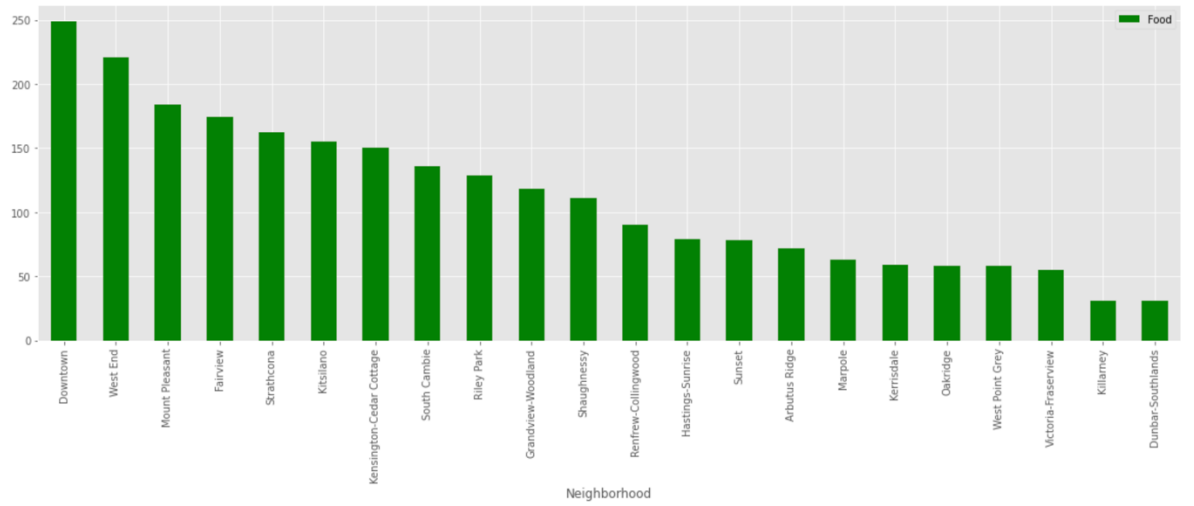
Using the detailed dataframe from the previous part, we create a summary dataframe that count all venues in each criteria for each neighborhood:

	Neighborhood	Entertainment	Education	Food	Nightlife	Outdoor	Healthcare	Religion	Shop	Transportation
0	Arbutus Ridge	2	14	72	4	28	21	8	43	11
1	Downtown	70	93	249	169	148	61	20	184	111
2	Dunbar-Southlands	7	5	31	5	16	4	5	33	4
3	Fairview	44	53	174	60	99	67	12	129	72
4	Grandview-Woodland	11	27	118	30	46	26	9	71	31
5	Hastings-Sunrise	8	16	79	12	31	14	10	50	26
6	Kensington-Cedar Cottage	5	19	150	6	35	29	9	39	46
7	Kerrisdale	5	5	59	5	18	9	6	57	15
8	Killarney	3	8	31	6	17	9	4	43	9
9	Kitsilano	33	25	155	36	86	34	5	119	55
10	Marpole	8	4	63	6	21	7	4	76	30
11	Mount Pleasant	36	49	184	56	79	43	22	109	66
12	Oakridge	7	26	58	5	30	21	15	70	26
13	Renfrew-Collingwood	3	12	90	12	31	17	8	63	31
14	Riley Park	9	25	129	19	49	48	19	78	44
15	Shaughnessy	8	26	111	7	37	25	15	72	33
16	South Cambie	12	29	136	16	50	52	18	87	42
17	Strathcona	36	75	162	74	96	35	12	85	69
18	Sunset	3	25	78	6	18	14	10	48	29
19	Victoria-Fraserview	2	10	55	5	11	15	6	49	17
20	West End	65	94	221	164	142	68	16	191	113
21	West Point Grey	5	11	58	6	29	8	5	41	7

From this summary dataframe, we can rank the neighborhood basing on each criteria. For example, the following charts rank the neighborhoods on each Entertainment, Education and Food criteria:







We can save this summary dataframe to a CSV file to make it more convenient to be used later. The CSV file is also uploaded to GitHub and available at: https://raw.githubusercontent.com/uclinux83/Coursera_Capstone/master/vancouver_venues_summary.csv

Part 4: Take user's preference, calculate score and rank the neighborhoods

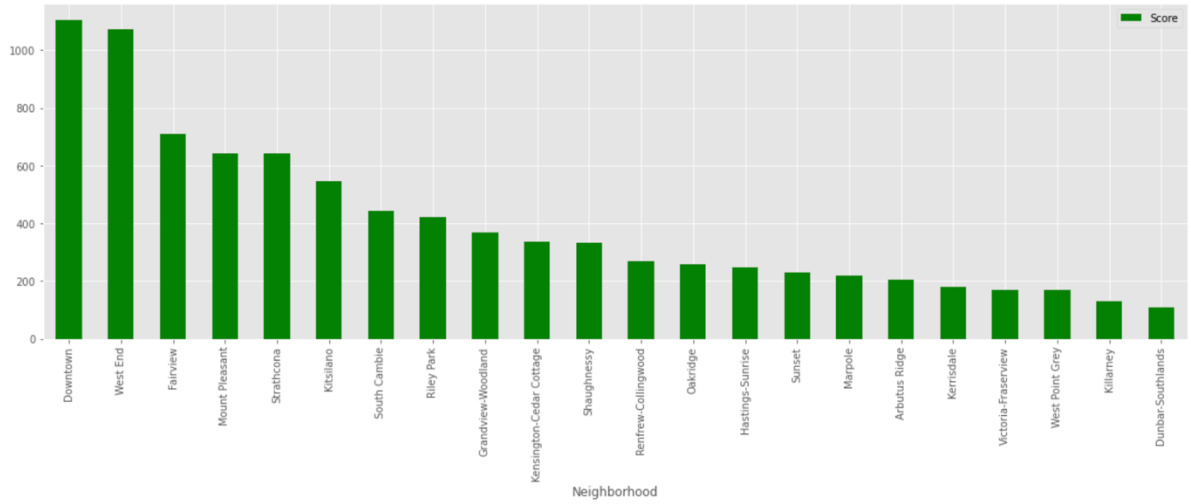
With the summary dataframe created from the previous part, we can now take the preferences from users and calculate score for each neighborhood. User can create a preference by giving weights for each of the 9 criteria. For example, the following preference give equal weight (1) to all of the criteria:

preference = {"Entertainment": 1, "Education": 1, "Food": 1, "Nightlife": 1, "Outdoor": 1, "Healthcare": 1, "Religion": 1, "Shop": 1, "Transportation": 1}

Score of each neighborhood is calculated and added to the dataframe as the following:

	Neighborhood	Entertainment	Education	Food	Nightlife	Outdoor	Healthcare	Religion	Shop	Transportation	Score
1	Downtown	70	93	249	169	148	61	20	184	111	1105
20	West End	65	94	221	164	142	68	16	191	113	1074
3	Fairview	44	53	174	60	99	67	12	129	72	710
11	Mount Pleasant	36	49	184	56	79	43	22	109	66	644
17	Strathcona	36	75	162	74	96	35	12	85	69	644
9	Kitsilano	33	25	155	36	86	34	5	119	55	548
16	South Cambie	12	29	136	16	50	52	18	87	42	442
14	Riley Park	9	25	129	19	49	48	19	78	44	420
4	Grandview-Woodland	11	27	118	30	46	26	9	71	31	369
6	Kensington-Cedar Cottage	5	19	150	6	35	29	9	39	46	338
15	Shaughnessy	8	26	111	7	37	25	15	72	33	334
13	Renfrew-Collingwood	3	12	90	12	31	17	8	63	31	267
12	Oakridge	7	26	58	5	30	21	15	70	26	258
5	Hastings-Sunrise	8	16	79	12	31	14	10	50	26	246
18	Sunset	3	25	78	6	18	14	10	48	29	231
10	Marpole	8	4	63	6	21	7	4	76	30	219
0	Arbutus Ridge	2	14	72	4	28	21	8	43	11	203
7	Kerrisdale	5	5	59	5	18	9	6	57	15	179
19	Victoria-Fraserview	2	10	55	5	11	15	6	49	17	170
21	West Point Grey	5	11	58	6	29	8	5	41	7	170
8	Killarney	3	8	31	6	17	9	4	43	9	130
2	Dunbar-Southlands	7	5	31	5	16	4	5	33	4	110

The following chart show the ranking of Vancouver neighborhood using this user's preference:



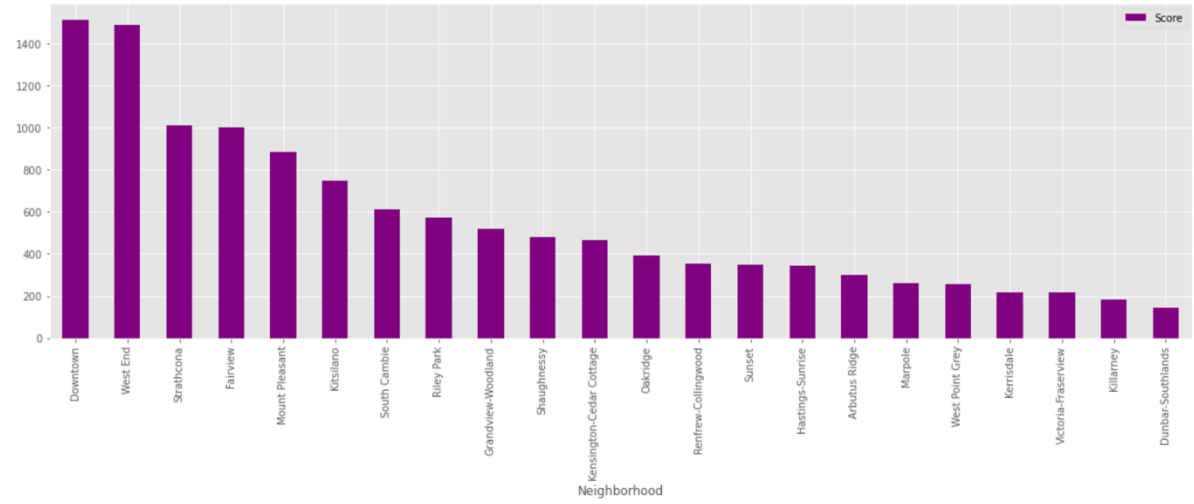
4. Results

By using this data model, stakeholders (immigrants, tourists, real estate agents, etc…) can create their own preference by setting weight for each of the 9 criteria as the input. The model will then calculate scores for each of the 22 official neighborhoods in Vancouver and rank them. Basing on such ranking, the stakeholder can make decision of where to live or stay in Vancouver or give recommendation to their clients.

For example, a family with kids may give high weight to Education, Outdoors, Food and Shop with the following preference settings:

preference = {"Entertainment": 0, "Education": 5, "Food": 1, "Nightlife": 0, "Outdoor": 3, "Healthcare": 1, "Religion": 0, "Shop": 1, "Transportation": 1}

The model will have the following ranking for the neighborhoods:

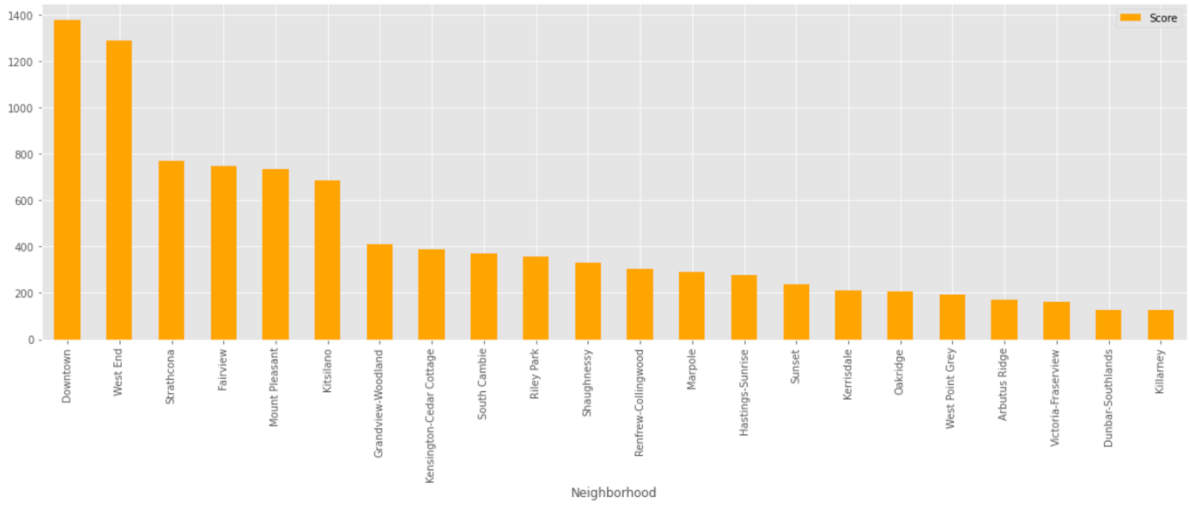


It is clear from the result that for this particular stakeholder, Stracona should be considered before Fairview and Mount Pleasant, which is different than the previous chart (where all criteria weights have the same value of 1).

Another example is a young tourist couples who give high weights to Entertainment and Nightlife but doesn't want to stay near spiritual centers and hospitals. Their preference may look like this:

preference = {"Entertainment": 3, "Education": 0, "Food": 2, "Nightlife": 2, "Outdoor": 1, "Healthcare": -3, "Religion": -2, "Shop": 1, "Transportation": 2}

The model will provide the following ranking:



For this young tourist couple, Grandview-Woodland is more recommended than South Cambie. This is different than what the model suggested for the family with kids who value Education and Outdoors more.

5. Discussion

One good thing about this model is that weight for each criteria can also be set to 0 (in case users don't care about that criteria) or even a negative number (in case users don't like to have such types of venues around their neighborhood).

One observation is that the venues in Vancouver are not equally distributed across the neighborhoods:

	Neighborhood	Entertainment	Education	Food	Nightlife	Outdoor	Healthcare	Religion	Shop	Transportation	Score
1	Downtown	70	93	249	169	148	61	20	184	111	1105
20	West End	65	94	221	164	142	68	16	191	113	1074
3	Fairview	44	53	174	60	99	67	12	129	72	710
11	Mount Pleasant	36	49	184	56	79	43	22	109	66	644
17	Strathcona	36	75	162	74	96	35	12	85	69	644
9	Kitsilano	33	25	155	36	86	34	5	119	55	548
16	South Cambie	12	29	136	16	50	52	18	87	42	442
14	Riley Park	9	25	129	19	49	48	19	78	44	420
4	Grandview-Woodland	11	27	118	30	46	26	9	71	31	369
6	Kensington-Cedar Cottage	5	19	150	6	35	29	9	39	46	338
15	Shaughnessy	8	26	111	7	37	25	15	72	33	334
13	Renfrew-Collingwood	3	12	90	12	31	17	8	63	31	267
12	Oakridge	7	26	58	5	30	21	15	70	26	258
5	Hastings-Sunrise	8	16	79	12	31	14	10	50	26	246
18	Sunset	3	25	78	6	18	14	10	48	29	231
10	Marpole	8	4	63	6	21	7	4	76	30	219
0	Arbutus Ridge	2	14	72	4	28	21	8	43	11	203
7	Kerrisdale	5	5	59	5	18	9	6	57	15	179
19	Victoria-Fraserview	2	10	55	5	11	15	6	49	17	170
21	West Point Grey	5	11	58	6	29	8	5	41	7	170
8	Killarney	3	8	31	6	17	9	4	43	9	130
2	Dunbar-Southlands	7	5	31	5	16	4	5	33	4	110

As we can see from the summary table, most of the venues (in almost all criteria) are in Downtown and West End so in most cases, these neighborhoods are in the top 2, no matter what the user's preference is. However, it is also very obvious that real estates in Downtown and West End are more expensive so this model can help stakeholders to consider other neighborhoods basing on their specific preference.

One limitation of this data model is that it just takes into account the quantity of venues but not the quality of them. Foursquare has premium APIs that can be used to get the ratings of each venues. So this can be a future improvement for this model.

6. Conclusion

In conclusion, this data model can be used by immigrants, tourists and real estate agents to compare and rank the 22 neighborhoods in Vancouver basing on specific "preferences", which are the combinations of nine criteria: Entertainment, Education, Food, Nightlife, Outdoors, Healthcare, Religion, Shop and Transportation. This scoring and ranking system can

help the stakeholders to make decision on where to live or stay in Vancouver, a beautiful west coast city in Canada.

In the future, this data model can be improved by not being limited to only the quantity of the venues but also taking into account the quality of each venue. This can be done by calling Foursquare premium APIs to get ratings and reviews of each venue. With such extra data, a more complex algorithm can be implemented to calculate the score for each neighborhood, basing on both quantity and quality dimensions. A web application can also be put on top of this to make it easier to use the model.