Capstone Project — The Battle of Neighborhoods: Where to open a new Tim Hortons in the city of Calgary, Alberta, Canada

1. Introduction/Business Problem

Tim Hortons Inc. is a fast food restaurant chain, specializing in coffee, doughnuts, and other fast food items. It is Canada's largest quick service restaurant chain; as of December 31, 2018, it had a total of 4,846 restaurants in 14 countries. The company has its headquarters in Toronto and uses franchise model for their stores. For Canadians, Tim Hortons is considered as one of the most iconic brands for the country, and it has been a very successful business. For potential Tim Hortons franchise owners; however, there is a problem – Tim Hortons is EVERYWHERE! It is difficult to find and choose their new Tim Hortons store location where they can minimize competitions from other Tim Hortons franchisees as well as other coffee shops. In this capstone project, I have created a Python code that utilizes Foursquare to find the best location for a new Tim Hortons in the city of my current residence - Calgary, Alberta Canada.

Audience:

Potential Tim Hortons franchise owners

Franchisee recruitment team in Tim Hortons head office

2. Data Requirements

Data source #1: Calgary's demographic data that contains a list of the city's 180 neighbourhoods, their latitude, longitude, median household income, median age, and population. The data I found is from 2014 – a little old – but good enough and still a good starting point. Source link: https://great-news.ca/demographics/

Data source #2: Foursquare APIs. By using this API, we will get all coffee shops and cafes in each neighborhood in the city.

3. Methodology

3.1 Data Preparation

3.1.1 Scraping Calgary Community Demographics

I first make use of Calgary Community Demographics page from https://great-news.ca/demographics/ to scrape the table to create a dataframe. For this, I used pandas to transform the data in the table on the page into a dataframe containing name of the 180 communities in Calgary, Newsletter Name, Median Household Income, Median Age, Population 2014, Dwellings 2014, City Quadrant, Median Home Sale Price. We start as below:

```
: df = pd.read_html("https://great-news.ca/demographics/")
  df = df[0] #1st table on the website
```

```
: df.head()
```

	Community	Newsletter Name	Median Household Income	Median Age	Population 2014	Dwellings 2014	City Quadrant	Median Home Sale Price
0	Abbeydale	-	$55,345	34	6071	2031	SE	$305,000
1	Acadia	Acadia	$46,089	42	10969	5067	SE	$447,000
2	Albert Park / Radisson Heights	-	$38,019	37	6529	2936	SE	$349,900
3	Altadore	The Source	$53,786	37	9518	4537	SW	$925,000
4	Applewood Park	-	$65,724	33	6864	2228	SE	$380,000

After removing columns not needed for analysis ("Newsletter Name", "Dwellings 2014", and "Median Home Sale Price"), and adding "Calgary" for community names (this is necessary for searching them later in Foursqure API because it is possible that communities with the same names exist in different cities elsewhere in the world), and latitude and longitude of the communities, the table looks like below.

[4]:	Community	Median Household Income	Median Age	Population 2014	City Quadrant	FullName	Lat	Lon
0	Altadore	$53,786	37	9518	SW	Altadore, Calgary	51.015104	-114.100756
1	CFB Lincoln Park	$30,411	35	2065	SW	CFB Lincoln Park, Calgary	51.074090	-114.087700
2	Acadia	$46,089	42	10969	SE	Acadia, Calgary	50.968655	-114.055587
3	Albert Park / Radisson Heights	$38,019	37	6529	SE	Albert Park / Radisson Heights, Calgary	51.044845	-113.990195
4	Douglasdale / Douglas Glen	$85,350	40	12952	SW	Douglasdale / Douglas Glen, Calgary	50.937830	-113.986829

```
[18]: WithCoord.dtypes
```

```
[18]: Community                 object
      Median Household Income   object
      Median Age                 int64
      Population 2014            int64
      City Quadrant             object
      FullName                  object
      Lat                      float64
      Lon                      float64
      dtype: object
```

3.2. Exploratory Data Analysis:

3.2.1 Using Foursquare Location Data

Getting communities' general characteristic

Let's make use of Foursquare API and get the top 100 venues for each community within a radius of 1000 meters.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Altadore	51.015104	-114.100756	River Park Off-Leash Area	51.013651	-114.095301	Dog Run
1	Altadore	51.015104	-114.100756	Monogram Coffee	51.010967	-114.100054	Coffee Shop
2	Altadore	51.015104	-114.100756	Garrison Pub	51.016131	-114.108674	Pub
3	Altadore	51.015104	-114.100756	My Favorite Ice Cream Shoppe	51.016239	-114.108888	Ice Cream Shop
4	Altadore	51.015104	-114.100756	The Trop	51.023080	-114.095147	Bar

Using onehot encoding, find the top 10 most common venues for each community

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Abbeydale	Convenience Store	Café	Sandwich Place	Filipino Restaurant	Frozen Yogurt Shop	Fried Chicken Joint	French Restaurant	Football Stadium	Fast Food Restaurant	Food Truck
1	Acadia	Breakfast Spot	Gym	Hardware Store	Grocery Store	Coffee Shop	Bank	Convenience Store	Gym / Fitness Center	Recreation Center	Gas Station
2	Albert Park / Radisson Heights	Asian Restaurant	Sandwich Place	Grocery Store	Indian Restaurant	Ethiopian Restaurant	Bowling Alley	Fast Food Restaurant	Falafel Restaurant	Gourmet Shop	Bank
3	Altadore	Dog Run	Park	Brewery	Coffee Shop	Café	Greek Restaurant	Bar	Hockey Rink	Liquor Store	Music Store
4	Applewood Park	Discount Store	Warehouse Store	Liquor Store	Park	Coffee Shop	French Restaurant	Football Stadium	Food Truck	Fried Chicken Joint	Fast Food Restaurant

Finding competitions (coffee shops or cafes) for each community

```
#Finding all coffee shops and cafes in Calgary
df_compe = Calg_venues[(Calg_venues["Venue Category"]=="Coffee Shop") | (Calg_venues["Venue Category"]=="Café")]
df_compe.head()
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
1	Altadore	51.015104	-114.100756	Monogram Coffee	51.010967	-114.100054	Coffee Shop
14	Altadore	51.015104	-114.100756	Bell's Bookstore Cafe	51.023094	-114.095614	Café
25	CFB Lincoln Park	51.074090	-114.087700	Starbucks	51.067270	-114.086350	Coffee Shop
29	CFB Lincoln Park	51.074090	-114.087700	Tim Hortons	51.067146	-114.082884	Coffee Shop
37	CFB Lincoln Park	51.074090	-114.087700	The Bullet Cappucino bar	51.082458	-114.092610	Coffee Shop

There are 333 coffee shops and cafes in the city

```
df_compe.shape
```

```
(333, 7)
```

However, not all competitors should be weighted equally. Tim Hortons stores get competitor score of 1 (direct competitor to future Tim Hortons franchisees), and coffee shops & cafes that are not Tim Hortons get competitor score of 0.5 (they

are competitors, but not quite, because of Tim Horton's strong customer loyalty and distinguished product offerings)

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Comp_score	Num_compe
1	Altadore	51.015104	-114.100756	Monogram Coffee	51.010967	-114.100054	Coffee Shop	0.5	0.5
14	Altadore	51.015104	-114.100756	Bell's Bookstore Cafe	51.023094	-114.095614	Café	0.5	0.5
25	CFB Lincoln Park	51.074090	-114.087700	Starbucks	51.067270	-114.086350	Coffee Shop	0.5	0.5
29	CFB Lincoln Park	51.074090	-114.087700	Tim Hortons	51.067146	-114.082884	Coffee Shop	1.0	1.0
37	CFB Lincoln Park	51.074090	-114.087700	The Bullet Cappucino bar	51.082458	-114.092610	Coffee Shop	0.5	0.5

Let's sum up the number of competitors for each community. The column is "Num_compe" column.

	Community	Median Household Income	Median Age	Population 2014	City Quadrant	Lat	Lon	Num_compe
0	Altadore	$53,786	37	9518	SW	51.015104	-114.100756	1.0
1	CFB Lincoln Park	$30,411	35	2065	SW	51.074090	-114.087700	2.0
2	Acadia	$46,089	42	10969	SE	50.968655	-114.055587	1.0
3	Albert Park / Radisson Heights	$38,019	37	6529	SE	51.044845	-113.990195	NaN
4	Douglasdale / Douglas Glen	$85,350	40	12952	SW	50.937830	-113.986829	NaN

"Num_compe" column shows 1, it could mean that the community has either one Tim Hortons or two coffee shops or cafes that are non-Tim Hortons. The above dataframe is showing the first 5 communities out of 180. We are seeing communities where there is no competition ("Num_compe" of "NaN" – which will be replaced with 0) – These communities might be good opportunities for future Tim Hortons franchisees. Having said so, there are a couple of other factors to consider besides the number of competitors in community:

"Population" – It is safe to assume that the more people in the community, the more potential customers.

"Number of competitors for each resident in the community" – If community A has 10,000 residents and two non-Tim Hortons coffee shops/cafes (ie. number of competitors of 1 for the community), it means the number of competitors for each resident in the community is 0.0001. This community would be less competitive than community B, which has 500 residents and one non-Tim Hortons coffee shop/café (ie. number of competitors of 0.5 for the community). The number of competitors for each resident in the community B is 0.001.
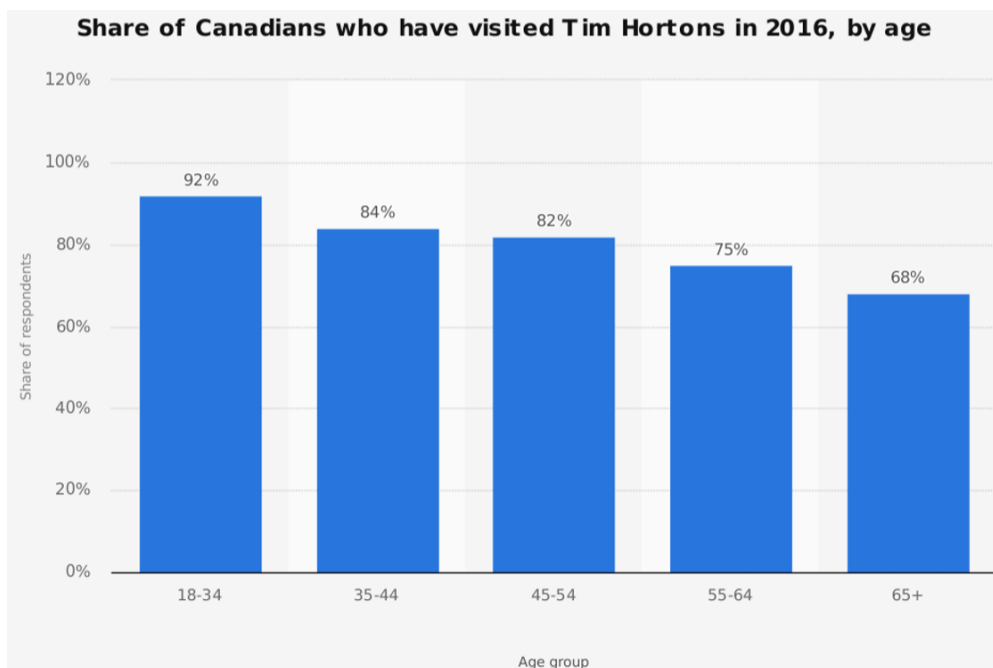
"Median Household Income" – the data from below shows that, generally speaking, people with higher income visit Tim Hortons more often. We can also assume that, people with higher income will spend more on each visit compared

with people with lower income. Data Source Link:
https://www.statista.com/statistics/575660/tim-hortons-visitors-by-age-canada/
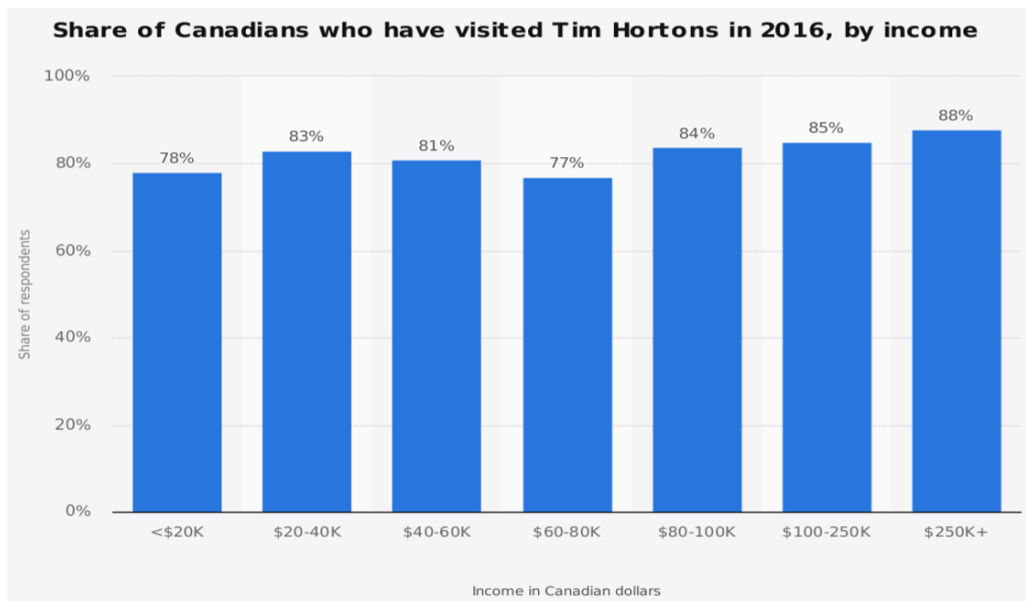
"Median Age" - the data from below shows that, Tim Hortons is more popular to younger people, and as people get older, they use Tim Hortons less. This makes sense as many products offered in Tim Hortons are comfort food, which, for health reasons, many people would reduce consumption as they get older. Data Source Link: https://www.statista.com/statistics/575660/tim-hortons-visitors-by-age-canada/

We now give rank each features for each community using pandas.DaraFrame.rank function.

"Median Household Income Ranking" – Community with the highest income gets 1. Ascending order (ie. the lowest income community gets a rank of 180)

"Median Age Ranking" – Community with the youngest median age gets 1. Ascending order. If more than one communities have the same Median Age, they get the same rank (we do this by using method ='dense' inside the rank function)

"Populating 2014 Ranking" – Community with the highest population gets 1. Descending order.

"Compe for each resident Ranking" – Community with the lowest number of competitors for each resident gets 1. Descending order.

Resulting dataframe (First 5 rows) is below. For example, the community of Bayview (the first line on the datafarme below), the community ranks 1st on the "Median Household Income Ranking" (great!), ranks 21st in "Median Age Ranking" (really good!), ranks 168th on "Populating 2014 Ranking" (horrible – almost at the bottom of the ranking), and ranks 121 on "Compe for each resident Ranking" (bad).

	Community	Median Household Income	Median Household Income Ranking	Median Age	Median Age Ranking	Population 2014	Population 2014 Ranking	City Quadrant	Lat	Lon	Num_compe	Compe_for_each_resident	Compe_for_each_resident Ranking	1st Mos Commor Venu
121	Bayview	184383	1.0	49	21.0	669	168.0	SW	50.9748	-114.1116	1.5	0.002242	121.0	Café
72	Rideau Park	153667	2.0	50	22.0	498	173.0	SW	51.0260	-114.0738	2.0	0.004016	125.0	Coffee Shop
66	Roxboro	153667	2.0	45	17.0	243	178.0	SW	51.0260	-114.0672	3.0	0.012346	129.0	Coffee Shop
101	Bel-Aire	138124	3.0	51	23.0	434	175.0	SW	50.9990	-114.0893	0.0	0.000000	1.0	Home Service
87	Elbow Park	135081	4.0	41	13.0	6252	82.0	SW	51.0220	-114.0824	0.5	0.000080	16.0	Parl

Now we assign weight to each of these factors to come up with the final ranking for each community.

Weights:

5% on "Median Household Income Ranking"

15% on "Median Age Ranking"

20% on "Population 2014 Ranking" (the more, the better)

60% on "Compe_for_each_resident Ranking"

Therefore, the math to come up with "Overall_Score" is: (5% * "Median Household Income Ranking") + (15% * "Median Age Ranking") + (20% * "Population 2014 Ranking") + (60% * "Compe_for_each_resident Ranking")

Finally, we sort "Overall_Score" by ascending order, and gives it a ranking. That becomes the "Oveall Ranking" (showing only the first 5 rows). The dataframe is also now combined with the top 10 most common venues for the communities to give views of the characteristics of the communities.

	Community	Median Household Income	Median Household Income Ranking	Median Age	Median Age Ranking	Population 2014	Population 2014 Ranking	City Quadrant	Lat	Lon	Num_compe	Compe_for_each_resident	Compe_for_each_resident Ranking	Overall_Score	Overall_Ranking	1st Mo Commo Venu
95	Panorama Hills	90859	30.0	32	4.0	25130	1.0	NW	51.1567	-114.0864	0.0	0.000000	1.0	2.90	1.0	Hom Servi
175	Tuscany	92453	27.0	34	6.0	19465	4.0	NW	51.1235	-114.2448	0.5	0.000026	2.0	4.25	2.0	Pharma
29	Cranston	91320	29.0	32	4.0	15354	9.0	SE	50.8891	-113.9809	0.0	0.000000	1.0	4.45	3.0	Piz Pla
172	Cougar Ridge	100908	21.0	33	5.0	13404	14.0	SW	51.0732	-114.2194	0.0	0.000000	1.0	5.20	4.0	I Crea Sho
11	Saddle Ridge	71101	68.0	30	2.0	17775	5.0	NE	51.1327	-113.9428	0.5	0.000028	3.0	6.50	5.0	Coffe Sho

4. Results & Discussion

We were able to rank communities for potential Tim Hortons franchisees thinking to open their stores in Calgary. Based on the criteria used, the best community would be Panorama Hills – it has high household income, residents of the community are young, population is large, and there is no competition in the community. However, in our analysis, we have ignored other factors such as the use of Tim Hortons related to ethnicity as well as residents' attitude toward a big national chain store etc. Since we do not have such data, I would say further investigation and local knowledge of each communities are needed before making the final decision.

5. Conclusion

In a fast-moving world, there are many real-life problems or scenarios where data can be used to find solutions to those problems. Like seen in the example above, data was used to find the best locations for a new business. However, this is just a reference. There are many other essential factors to consider for choosing locations to open businesses, and yet data have not been measured and simply do not exist. In those cases, we should avoid making important decisions sorely based on data available online. It is always a good idea to check and validate your analysis with experts of the areas you are exploring.