Project: Capstone Project - The Battle of Neighborhoods

A full report consisting of all of the following components (15 marks):

Business Problem / Who is Interested **2**
Introduction where you discuss the business problem and who would be interested in this project

Data Review **3**
Data where you describe the data that will be used to solve the problem and the source of the data

Methodology Data Review **3**
Methodology describing any exploratory data analysis, inferential statistical testing, and what machine learnings were used and why.

Results **6**
Results section where you discuss the results.

Observations and Recommendations **8**
Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.

Conclusion **9**
Conclusion section where you conclude the report.

Business Problem / Who is Interested

Introduction where you discuss the business problem and who would be interested in this project.

Problem Exploration: With current COVID-19 issues, we are faced with an issue as warmer weather transitions to cooler and out-right cold climates. This will push us to spend more time indoors, especially in a climate like Toronto.

Background of the Solution: Due to less ventilation and tighter spaces, health experts everywhere, even in areas that have already flattened their initial outbreaks, predict a reoccurrence once activities are forced indoors.

There is a two-approach solution that can be implemented separately or together:
1) Bigger or More Space - Expand the area of the business for social distancing
2) Improved COVID-19 Ventilation – HVAC filers have been identified that are more effective than standard use HVAC filters. There are also design features that can be built into a portable unit, like a two-layer ceiling that better allows for increased airflow. Additionally, more efficient heating systems complement the air filtering capabilities to work in the coldest climates.

Problem/Solution Overview: This new business opportunity for "Pop-up Outside Spaces" would be an extension to the current portable structure products offered by modular builders and higher-end tent systems, such as those used for wedding receptions and corporate retreats.

The building units would have to meet COVID-19 prevention standards providing necessary additional space to meet social distancing requirements with featuring heating and ventilation that met new COVID-19 standards. The pop-up spaces will be modular to allow for adjustments from larger to smaller spaces, as needed. They can be used for additional rooms for schools or offices that need more space due to social distancing requirements.

Three product offerings and target businesses:
1) **Require improved ventilation**: Gym or fitness businesses that require improved ventilation and airflow
2) **Event spaces**: Large event rooms that would provide space for conventions, banquets, parties, and concerts
3) **Large empty spaces**: Public spaces that are unlikely to be used in winter can be built into community gathering centers

Economy Assumptions: The data we are using is a blend of geo-information for postal codes and FourSquare, a location-based business listings, and reviews directory. It does not include business performance (i.e., revenue vs. expenses) or planning information. This would include projected occupancy rates of the local area, commerce activity, city revenues, economic development plans, and health department reports – to name some of the data outside of this project scope.

One primary assumption that is the foundation of this project that needs to be acknowledged is that COVID-19 is a significant disruptor to local economies. So analysis using pre-2020 data to rank the "most popular businesses in each neighborhood" will likely swing considerably, especially for small local-owned.

Data Review

Data where you describe the data that will be used to solve the problem and the source of the data.

Using the Data: The goal for clustering data, in this case, is to identify neighborhoods that have the right business or locations in place that can benefit from "Pop-up Outside Spaces," since the pop-up structures are an add-on to a current site.

It can also help identify the type of marketing plan based on the number of locations that match up for each product/neighborhood mix. For groups with a larger number of locations, like gyms, the plan would be to use direct and digital marketing. For smaller target groups, like parks and hotels, it would identify that using a personal sales manager may be the better choice.

The first step would be to build the target businesses by identifying venue categories that would match up for each product. For instance, the categories "gym" and "Gym/fitness" are just starters; we also need to find other businesses where people do physical exercise.

Data analysis approach to the data:
There are two approaches to using the data. First, analyze neighborhoods against each other using all 229 categories. If our target categories show up at the top of these lists, then these will be are Tier 1 targets since they are the top for this category in the city. Then, second, analyze the neighborhoods using a sub-set of the categories that we selected. This will still identify key target areas but at a Tier 2 level.

What to do with the data next:
The key to clustering data is that it provides insights into what neighborhoods to focus marketing efforts based on the highest concentration and the importance of the targeted business types.

One Final Note – ALL of the categories selected for each targeted group can be debated that is acceptable. However, to achieve the goal of identifying the right neighborhoods the plan was to get a good sample rather than to spend too much time trying to achieve 100% accuracy. The real goal was to create a sample selection that could generate actionable data.

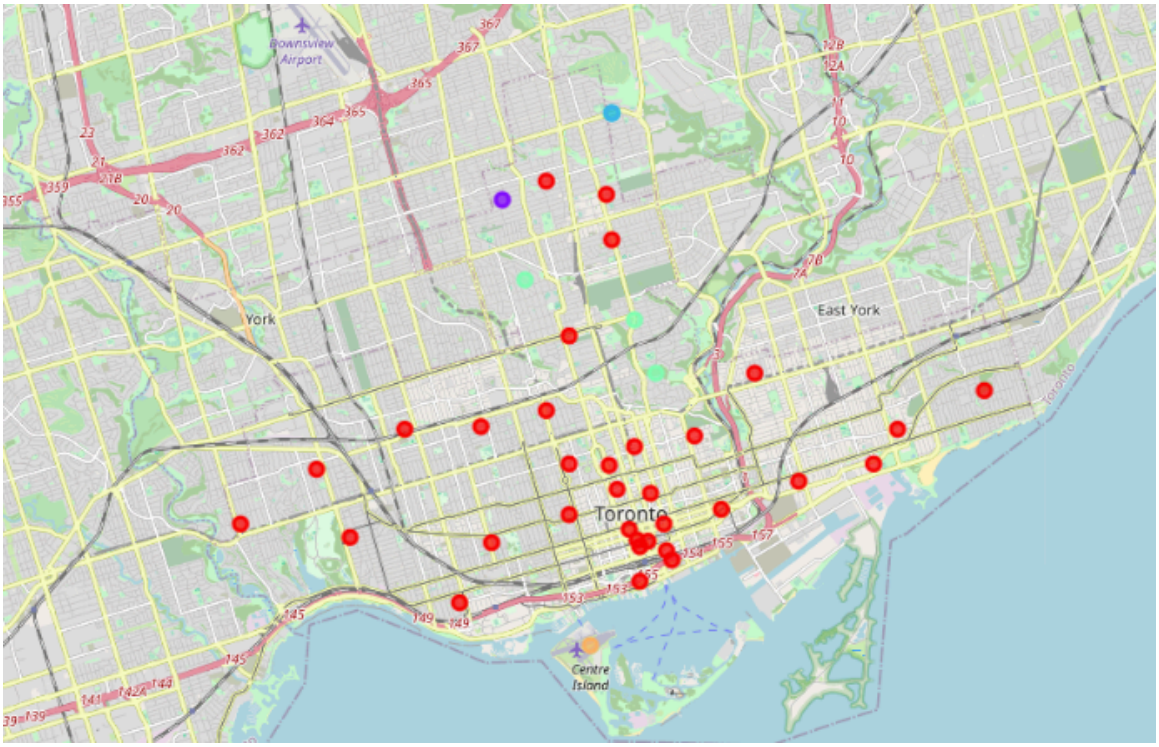
Methodology Data Review

Methodology describing any exploratory data analysis, inferential statistical testing, and what machine learnings were used and why.

Source Evaluation – Focusing on "Contains 'Toronto'"
The dataset for the source of our data was a blend of a geo-information (postal codes with latitude and longitude) and FourSquare location-based business listings. Considering the business plan is very geo-targeted, it requires installing a product at the customer location, this would be sufficient. The data targets 39 neighborhoods in the downtown area of Toronto.

The sub-Toronto target contains the areas of Downtown, Central, East & West Toronto is based on the time and resource needs of selling, designing, getting permits, and building, also knowing that each of those has substeps. The estimate was that if they could

successfully 40-50% of the venue locations in results, they would be 90-120% scheduled through the end of the year. So targeting the whole city was not necessary, and potential trouble if they could not deliver on time.



Clustering using all 229 venue categories in the target area of Downtown, Central, East & West Toronto (k=5)

Reviewing the Venue Categories

The business plan is a response to the COVID-19 crisis. One of the unfortunate outcomes of the crisis that is visible by walking through the neighborhood, but not yet captured in available data research is the economic impact the crisis is having now and will have one-year, five-years, etc. down the road.

The one advantage of the business plan is that it can target businesses that may have better odds of surviving due to government backing or support (parks and conventions spaces) or there the business (gyms and fitness spaces) are in a high enough quantity that a core group will survive the business slowdown and would rely on this advanced ventilation product to reopen.

Creating Target Tiers

The first step was to identify Tier 1 prospects and determine where they were located for marketing or lobbying purposes. This was done by analyzing neighborhoods against each other, using all 229 venue categories. If our target categories show up at the top of these lists, then these will be are Tier 1 targets since they are the top for this category in the city. Then, second, analyze the neighborhoods using a sub-set of the categories that we selected. This will still identify key target areas but at a Tier 2 level.

```
In [190]: toronto_grouped
Out[190]:
```

	Neighborhood	Yoga Studio	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	...	Theme Restaurant	Toy / Game Store	Tra
0	Berczy Park	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0000
1	Brockton, Parkdale Village, Exhibition Place	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0000
2	Business reply mail Processing Centre, South C...	0.066667	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0000
3	CN Tower, King and Spadina, Railway Lands, Har...	0.000000	0.071429	0.071429	0.071429	0.142857	0.214286	0.142857	0.000000	0.000000	...	0.000000	0.000000	0.0000
4	Central Bay Street	0.015152	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0000
5	Christie	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0000
6	Church and Wellesley	0.027397	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.013699	0.000000	...	0.013699	0.000000	0.0000
7	Commerce Court, Victoria Hotel	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.040000	0.000000	...	0.000000	0.000000	0.0000
8	Davisville	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.029412	0.0000
9	Davisville North	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0000
10	Dufferin, Dovercourt Village	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0000

Portion of the full Venue Category X Neighborhood table with indicating which Venue Categories are in each each Neighborhood. This uses all 233 venue categories for 39 Neighborhoods (k=10)

Next was to identify the venue categories that fill-out the target businesses. For instance, expand the number of venue categories for the ventilation product beyond the obvious "gym and "gym/fitness. " Below, you can see how each was expanded.

1) **Require improved ventilation**: Gym or fitness businesses that require improved ventilation and airflow -- Yoga Studio, Gym, Gym / Fitness Center, College Gym, Martial Arts Dojo, Skate Park, Skating Rink, Tennis Court, Spa

2) **Large gathering locations**: Large event rooms that would provide space for conventions, banquets, parties, and concerts -- Hotel, Convention Center, Baseball Stadium, Basketball Stadium, College Auditorium, College Rec Center, Event Space, General Entertainment, Performing Arts Venue

3) **Large empty space**: Public spaces that are unlikely ever used in winter that can be built into community gathering centers -- Park, Trail, Garden

Results

Results section where you discuss the results.

Running Neighborhood vs Neighborhood Comparison

A) Tier 1 Analysis: Clustering data of 229 Venue Categories across 39 Neighborhoods

Findings:

1) Two clusters (groups 2 & 3) were created that had "Park" as the #1 Common Venue, these area will be targeted with marketing about the benefits of outdoor space in the winter. Since it likely that these parks are managed by a government group, the sales team will also lobby the neighborhood elected officials.

```
In [195]: toronto_merged.loc[toronto_merged['Cluster Labels'] == 2, toronto_merged.columns[[1] + list(range(5, toronto_merged.shape[1]))]]
```

Out[195]:

	Postalcode	Cluster Labels	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	M4N	2	Park	Photography Studio	Bus Line	Swim School	Deli / Bodega	Eastern European Restaurant	Dumpling Restaurant	Donut Shop	Doner Restaurant	Dog Run

```
In [196]: toronto_merged.loc[toronto_merged['Cluster Labels'] == 3, toronto_merged.columns[[1] + list(range(5, toronto_merged.shape[1]))]]
```

Out[196]:

	Postalcode	Cluster Labels	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
4	M4T	3	Park	Gym	Tennis Court	Trail	Distribution Center	Department Store	Dessert Shop	Diner	Discount Store	Doner Restaurant
7	M5P	3	Park	Jewelry Store	Trail	Sushi Restaurant	Deli / Bodega	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Donut Shop	Doner Restaurant
9	M4W	3	Park	Playground	Trail	Dance Studio	Eastern European Restaurant	Dumpling Restaurant	Donut Shop	Doner Restaurant	Dog Run	Distribution Center

Tier 1 Cluster results for "Park" using all 229 (k=5). The 4 neighborhoods are listed below.

Top Venue = "Park"	
Lawrence Park	Moore Park, Summerhill East
Forest Hill North & West, Forest Hill	Rosedale

2) Other Target Groups – the clustering exercise did not surface any apparent pointers to help with the hotel/convention or the gym/fitness offerings. It's likely that there will be too much noise with 229 venue categories, and in this situation, success can be achieved with a smaller sample set.

B) Tier 2 Analysis: Narrowing the Venue Categories:

1) Try 1: Narrowed to just 8 Venue Categories for the 3 products
   Result - There wasn't enough variation for any clustering using K=3-5
2) Try 2: Expanded to 22 Venue Categories and tried clustering at K=3-7
   Result – There wasn't much variation between the different K values. At k=5, 4 of the clusters had 1 or 2 entries, leading 30 in the k=1 cluster
3) Try 3: Expanded to K=10 using the 22 Venue Categories

Result – It was a better breakdown though still a heavy focus on the k=1 cluster. However each entry in the K=1 has at least 4 of 5 top categories being in the gym fitness segment.

```
In [205]: toronto_merged.loc[toronto_merged['Cluster Labels'] == 0, toronto_merged.columns[[1] + list(range(5, toronto_merged.shape[1]))]]
```

Out[205]:

	Postalcode	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
3	M4S	0	Gym	Tennis Court	Park	Spa	Basketball Stadium
5	M4V	0	Spa	Video Game Store	Gym	Gym / Fitness Center	Hotel
8	M5R	0	Martial Arts Dojo	Park	Spa	Basketball Stadium	Gym
10	M4X	0	Park	Gym / Fitness Center	General Entertainment	Spa	Basketball Stadium
11	M4Y	0	Yoga Studio	Hotel	Gym	Martial Arts Dojo	Park
13	M5B	0	Hotel	Spa	Gym	Gym / Fitness Center	Park
14	M5C	0	Gym	Hotel	Park	Performing Arts Venue	Spa
15	M5E	0	Gym	Hotel	Park	Basketball Stadium	Spa
16	M5G	0	Spa	Park	Gym / Fitness Center	Hotel	Yoga Studio
17	M5H	0	Gym	Hotel	Gym / Fitness Center	Event Space	Spa
18	M5J	0	Hotel	Baseball Stadium	Park	Gym	Skating Rink
19	M5K	0	Hotel	Spa	Basketball Stadium	Gym	Gym / Fitness Center
20	M5L	0	Hotel	Gym	Gym / Fitness Center	Park	Spa
21	M5S	0	Yoga Studio	College Gym	Video Game Store	Basketball Stadium	Gym
22	M5T	0	Park	Spa	Video Game Store	Gym	Gym / Fitness Center
23	M5V	0	Spa	Video Game Store	Gym	Gym / Fitness Center	Hotel
24	M5W	0	Gym	Hotel	Park	Yoga Studio	Basketball Stadium
25	M5X	0	Hotel	Gym	Gym / Fitness Center	Spa	Video Game Store
27	M7A	0	College Auditorium	General Entertainment	Gym	Park	Yoga Studio
29	M4K	0	Spa	Trail	Yoga Studio	Skating Rink	Baseball Stadium
31	M4M	0	Yoga Studio	Gym / Fitness Center	Park	Video Game Store	Gym
93	M6J	0	Yoga Studio	Park	Video Game Store	Gym	Gym / Fitness Center
94	M6K	0	Gym	Performing Arts Venue	Spa	Basketball Stadium	Gym / Fitness Center

Ranking of Venue Categories by Neighborhood using just the 12 venue categories (k=10)

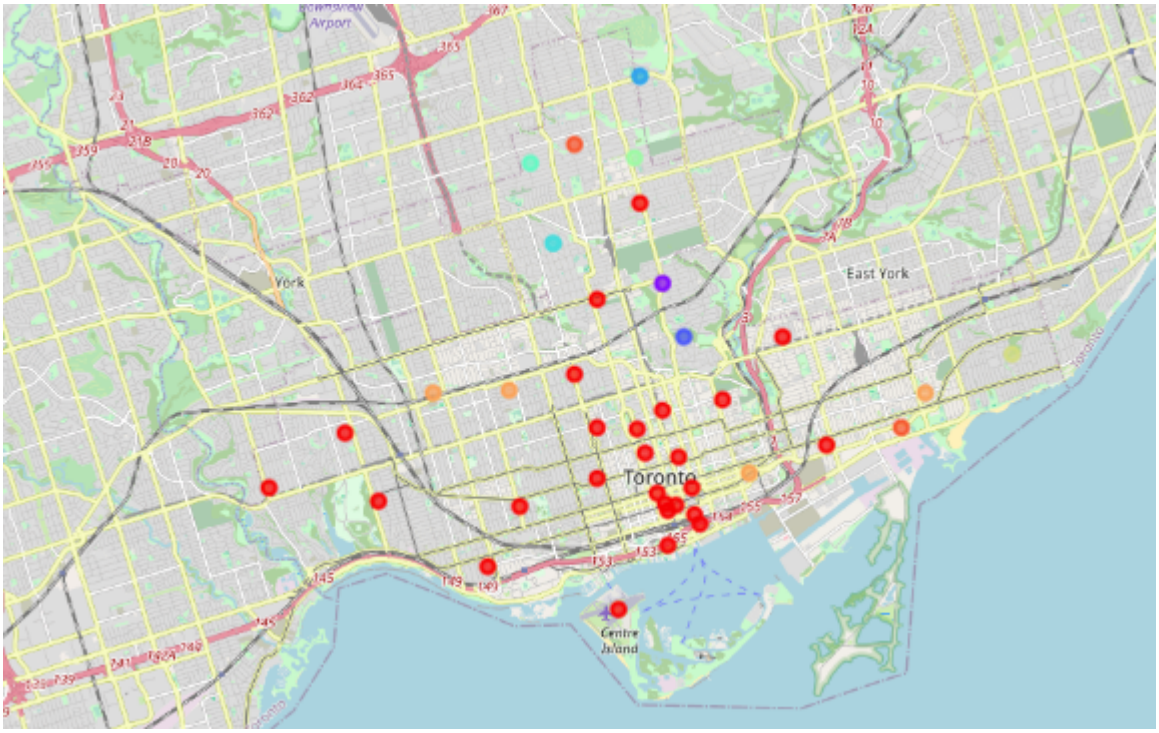
Special Ventilation - Gyms	**Large Gathering - Hotel**
• Davisville • Summerhill West, Rathnelly, South Hill, Fore • The Annex, North Midtown, Yorkville • St. James Town, Cabbagetown • Church and Wellesley • Garden District, Ryerson • St. James Town • Berczy Park • Central Bay Street • Richmond, Adelaide, King • Harbourfront East, Union Station, Toronto Islands • Toronto Dominion Centre, Design Exchange • Commerce Court, Victoria Hotel • University of Toronto, Harbord • Kensington Market, Chinatown, Grange Park • CN Tower, King and Spadina, Railway Lands, • Stn A PO Boxes	• Church and Wellesley • Garden District, Ryerson • St. James Town • Berczy Park • Richmond, Adelaide, King • Harbourfront East, Union Station, Toronto Islands • Toronto Dominion Centre, Design Exchange • Commerce Court, Victoria Hotel • Stn A PO Boxes • First Canadian Place, Underground city

• First Canadian Place, Underground city • Queen's Park, Ontario Provincial Government • The Danforth West, Riverdale • Studio District • Little Portugal, Trinity • Brockton, Parkdale Village, Exhibition Place • High Park, The Junction South • Parkdale, Roncesvalles • Runnymede, Swansea	

Observations and Recommendations

Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.

The importance of the clustering data is that it provides insights into where marketing efforts should focus based on the highest concentration of the businesses types listed below.



Cluster map pf the Tier 2 venues provides target of direct marketing and planned drop-in meetings (k=10)

From Step A – Tier 1 results – For the "**Park**" neighborhoods, a campaign targeting area residents with a "Wouldn't it be nice to have a safe place for the community to gather?" campaign, and there will be direct reach out to lobby government officials.

From Step B – Tier 2 results – For the **gym ventilation** targets, the list of businesses will be much longer than the others, so it will start with a direct media campaign and then use the cluster to do drop-in visits to see if they are open. For the **Hotel** targets, the list will be much shorter so I'll reach out directly.

Conclusion

Conclusion section where you conclude the report.

Even when working with limited datasets for a neighborhood solution, the data review and the attempt to get to a workable plan is never cut and dry. More dimensions appear on top of what the original review gave.

In this case, the analysis could have continued with additional datasets being pulled to determine which businesses are successful, which are in higher-income or more densely populated neighborhoods, but that would have been more than needed.

Since the company had done an early review of their production capability, they could draw a box around the solution and know their capabilities. They knew starting the project in that they were only capable of completing X# of new installations, based on a formula of A# days for selling/planning/permitting. They also projected that they could only add B# of new workers since any additional new hires, and they couldn't on-board and safely train them in time.