Data Utilized

For my project, I used 2 basic categories of data. The first is the list of neighborhoods and their corresponding longitude and latitude. For New York, I was able to use data I already had from an earlier module in the IBM Data Science Course. For Toronto, the neighborhood data containing the postal codes was obtained by using Beautiful Soup to scrape data from a Wikipedia page. I also had a csv file from a Coursera course which linked the postal codes to longitude and latitude data. I used Pandas to merge together the 2 data sources and create a single dataframe, which I saved to file using to_pickle().

The second, and more significant source of data was the FourSquare API, which provided me with 2 types of information. The Explore endpoint gave me a large amount of simple venue data based on a radius around a point of longitude and latitude. Below is a sample of the data obtained from the Explore API (I will explain the highlighted area next), after being merged with the neighborhood data for Toronto:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue_Id	Venue	Venue Latitude	Venue Longitude	Venue Category	SimpleCategory
0	Lawrence Park	43.728020	-79.388790	50e6da19e4b0d8a78a0e9794	Lawrence Park Ravine	43.726963	-79.394382	Park	Recreation
1	Lawrence Park	43.728020	-79.388790	5082ef77e4b0a7491cf7b022	Zodiac Swim School	43.728532	-79.382860	Swim School	Recreation
3	Davisville North	43.712751	-79.390197	4ba011c2f964a5204a5737e3	Sherwood Park	43.716551	-79.387776	Park	Recreation
5	Davisville North	43.712751	-79.390197	4adb2fd3f964a520c42421e3	Homeway Restaurant & Brunch	43.712641	-79.391557	Breakfast Spot	CheapMeal
8	Davisville North	43.712751	-79.390197	4b0b3691f964a520c62e23e3	Subway	43.708474	-79.390674	Sandwich Place	CheapMeal
9	Davisville North	43.712751	-79.390197	4c3f2724db3b1b8d635e6695	900 Mount Pleasant - Residents Gym	43.711671	-79.391767	Gym / Fitness Center	Recreation
10	Davisville North	43.712751	-79.390197	4e03639dc65b8061424e3495	Provocative Pizza Series	43.708293	-79.389546	Pizza Place	Ethnic Food
11	North Toronto West	43.715383	-79.405678	51606062e4b0878cf540f4a2	Barreworks	43.714070	-79.400109	Yoga Studio	Recreation
13	North Toronto West	43.715383	-79.405678	515f21e5e4b03c1e772c36c7	Sushi Shop	43.713861	-79.400093	Restaurant	CheapMeal
14	North Toronto West	43.715383	-79.405678	4bdd97185b31c9b6f00d9d16	Nailsense	43.717467	-79.400653	Spa	Recreation

The Explore endpoint does a great job providing basic information about venues based on a geographical point. But for my purposes in this analysis, I found one part problematic. This is the fact that, by the time I had gathered all of the venues for all of the neighborhoods in Toronto and New York, I found that I had **over 440 different values** in the “Venue Category” column. There were 112 different values in the category I call “Ethnic Food” alone. This made my project much more difficult, because, in reality, I was not trying to distinguish between many types of places which were not actually different in terms of understanding “fun venues” in each city.

I needed to come up with something which would enable me to work with this data more easily, so I manually categorized all of the Venue Category values, and broke them down into 13 categories that I could work with:

Entertainmer	Recreation	CheapMeal	CoffeeDesse	Ethnic Food	Fancy Food	Bar	FunStore	NotFunCheapMeal	Store	FoodStore	Travel	Misc
Theater	Park	Café	Coffee Shop	Pizza Place	Seafood Restau	Bar	Bookstore	Fast Food Restaurant	Clothing Store	Fish Market	Hotel	Bank
Concert Hall	Gym	Restaurant	Bakery	Italian Restaurant	Steakhouse	Pub	Art Gallery	Airport Food Court	Pharmacy	Cheese Shop	Rental Car Loca	Intersection
Movie Theater	Gym / Fitness Center	Sandwich Place	Dessert Shop	Japanese Restaurant	French Restaura	Gastropub	Antique Shop	College Cafeteria	Sporting Goods	Grocery Store	Light Rail Statior	Salon / Barbershop
Aquarium	Yoga Studio	Breakfast Spot	Tea Room	Sushi Restaurant	New American R	Liquor Store	Flea Market		Department Stor	Beer Store	Train Station	Plaza
Jazz Club	Spa	Burger Joint	Juice Bar	Thai Restaurant	Bistro	Beer Bar	Candy Store		Convenience St	Supermarket	Airport Service	Office
Museum	Athletics & Sports	Deli / Bodega	Bubble Tea Sho	American Restaurant	Molecular Gastr	Brewery	Comic Shop		Furniture / Home	Farmers Market	Bus Line	Gas Station
Music Venue	Playground	Fried Chicken Joint	Donut Shop	Diner		Cocktail Bar	Record Shop		Electronics Stor	Gourmet Shop	Bus Station	Distribution Center
Basketball Stadi	Trail	Food Court	Cupcake Shop	Vegetarian / Vegan Restaurant		Lounge	Shopping Mall		Cosmetics Shop	Food & Drink Sh	General Travel	Construction & Landscapin
Performing Arts	Skating Rink	Bagel Shop	Smoothie Shop	Mexican Restaurant		Wine Bar	Arts & Crafts Store		Pet Store	Health Food Sto	Metro Station	IT Services
Event Space	Baseball Field	Food Truck	Chocolate Shop	Greek Restaurant		Nightclub	Thrift / Vintage Store		Discount Store	Butcher	Airport Terminal	Home Service

This is the origin of the **SimpleCategory** column shown highlighted in green above. It is a form of binning, in that I am reducing a very large number of category types to only 13 of them. But for my purposes, I needed to go one

step further. I needed to apply one more filter which would help me determine how fun a neighborhood was. So, from my 13 Simple Categories, I chose these to represent the Fun Venue Types:

- Entertainment
- Recreation
- Cheap Meal
- Coffee Dessert
- Ethnic Food
- Fancy Food
- Bar
- Fun Store

The next step in preparing my data involves using the **Venues endpoint** at FourSquare. The Venues endpoint provides a large amount of detail for a single venue when you pass in a Venue_Id. However, this is a Premium Endpoint, which means that for a developer account, you can only access the API 500 times within 24 hours. It took me over a week to finally pull in all the data I needed, which included details on 2400 venues. I actually had far more venues in Toronto and New York data pulled from the Explore endpoint, but for practical reasons, I chose to focus on **only the Top 10 neighborhoods** in terms of the number of Fun venues each has. By limiting the number of venues to only the best neighborhoods in each city, I was able to realistically get all the details I needed for each neighborhood.

The Venues endpoint provides a huge amount of data, but after studying it in detail, and making several dozen premium calls to the endpoint to see what kind of data I was actually getting, I realized that only certain columns were really going to be useful in my analysis. There were some data points which I was excited about, but I realized soon that this data was not actually being returned by the API in any of the calls. I ended up using the following data in my analysis:

- Rating
- RatingSignals (how many people provided a rating)
- LikesCount
- TipsCount
- PriceTier

	Venue_Id	Name	TipsCount	LikesCount	Rating	RatingSignals	PriceTier	VoteSum
0	537befcc498edb1da559269b	Symposium Cafe Restaurant & Lounge	11	30	6.4	52	4	123
1	4b75ce6af964a5201b262ee3	Subway	2	0	5.9	5	1	7
6	4b5fcadff964a520eecc29e3	Wingporium	24	44	8.0	69	2	181
7	4bc9f9b6b6c49c7469688f91	South St. Burger	11	25	8.1	35	1	96
8	4b1d492af964a520370e24e3	Artisano Bakery Café	24	34	6.4	72	2	164
9	4c116455d41e76b09552310d	Subway	0	1	6.2	3	1	5
14	4b017e51f964a520d84222e3	LCBO	8	29	7.4	41	NaN	107
15	5cebda3300b068002dca9d81	Starbucks	1	2	6.9	3	1	8
16	4b251b4cf964a520586c24e3	Cafe Sympatico	6	7	6.5	13	1	33
17	4cb4e17052edb1f7745763fe	Pizza Hut	0	1	6.4	1	1	3

There are 2 things to point out in the example above. One is that PriceTier is null in some cases, and the other is that I have added a VoteSum column, which is basically adding up the TipsCount, LikesCount x 2, and the RatingSignals. VoteSum was added to help me get a general idea of how popular and “likely to be fun” a venue was. The lack of PriceTier data for some venues was something I found interesting, as a person who generally cares how much things cost. This is why I decided to use Multiple Linear Regression to see if I could accurately predict the PriceTier from the City, Neighborhood, and other Venue details