Finding the Best City for Different Venue Categories Based On Popularity:

---

1. Introduction:

1.1 Background:

New York City, NY and Toronto, ON are often thought of as “sister” cities due to their common traits of being the biggest cities in their respective countries. Both cities are celebrated for their fast paced environments, extensive food scene, cultural diversity and being large financial capitals. Thus, it makes sense that Toronto is often thought of as the “New York of Canada” and vice versa.

1.2 Business Problem:

Both cities are vastly known for their large number of venues and venue types. Both cities are rich in their many venues; from pubs, restaurants, parks, museums and more. In fact, both cities draw hundreds of thousands of tourists every year to explore these venues. Therefore, it makes sense for a potential tourist who wants to visit either city (or both cities) to know which city’s are more well known for which *types* of venues, so they can plan their visits accordingly.

For example: say a tourist, John, wants to see which city is more popular for Korean restaurants vs Indian restaurants. He uses our program to discover that New York is more popular for Korean, and Toronto is more popular for Indian. Therefore, when he’s in Toronto, he visits an Indian restaurant and when he’s in New York, he visits a Korean restaurant.

1.3 Solution:

In this project, we will be observing the neighborhoods in both Toronto and NYC. We will observe the types of venues that are present in neighborhoods of each city, and through data analysis, determine which types of venues are more popular in which neighborhoods; and each city as a whole. We will then define a few functions that will tell us, for any given venue type, which city is more popular for that venue type, and *if* the user wants, which specific neighborhoods within that city are most popular for that venue type.

2. Data Acquisition and Cleaning:

2.1 Data Source and Initial Cleaning:

The New York City data will be provided by NYU and contains the neighborhood name, the borough they belong to and the respective latitudes and longitudes for each neighborhood. We will then be using the FourSquare API to get data about the different venues for each neighborhood, utilizing the latitude and longitude values to call it.

The Toronto data will be scraped off the wikipedia page for Toronto. We use the BeautifulSoup package in Python to grab the data from the wiki page and format it into a pandas dataframe. This data provides us with the postal code, borough name, list of neighborhoods within that postal code. We then use Nominatim package in Python to fetch the latitude and longitude coordinate average for each of the neighborhoods within a postal code region. We end up with a pandas dataframe which shows us the postal code, borough, neighborhood(s) within that postal code area, latitude and longitude. Since the Toronto data is not sorted by neighborhoods, but rather different postal code regions, for the sake of simplicity we will refer to each postal code region as a “neighborhood.”

3. Methodology:

3.1 – Step 1: Configuring the New York City Data:

- The first step was to configure the NYC data so it was ready to use for analysis. We used the NYU provided json data and converted it to a pandas dataframe which displayed: Borough, Neighborhood, Latitude, Longitude.
- We then used the Foursquare API to iterate through each neighborhood and find the venues nearby it.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
2	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
3	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

- Afterwards, we performed one-hot encoding on the venue category, to transform the dataframe to include columns of all the venue categories. Each row of our dataframe now had the neighborhood name, and a 1 or 0 as the value in each venue category column indicating whether or not the venue representing that row was of that category type.
- We then grouped the one-hot dataframe by neighborhood. So, each row of our dataframe represents a neighborhood in New York, and the columns would represent all the different venue categories/types, and the values would represent the average frequency of venues of that type within that neighborhood. Our data looked like:

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Terminal	American Restaurant	Antique Shop
0	Allerton	0.0	0.0	0.0	0.0	0.0	0.000000	0.0
1	Annadale	0.0	0.0	0.0	0.0	0.0	0.181818	0.0
2	Arden Heights	0.0	0.0	0.0	0.0	0.0	0.000000	0.0
3	Arlington	0.0	0.0	0.0	0.0	0.0	0.000000	0.0
4	Arrochar	0.0	0.0	0.0	0.0	0.0	0.000000	0.0

- Afterwards, we performed some data analysis on the dataframe and determined the top 10 most frequently occuring venue types for each neighborhood based on their frequency values. We ended up with a dataframe that shows each neighborhood, and the 10 most common venue categories found within each neighborhood.

	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
0	Allerton	Pizza Place	Chinese Restaurant	Deli / Bodega	Spa	Supermarket	Fried Chicken Joint	Donut Shop
1	Annadale	Pizza Place	American Restaurant	Dance Studio	Restaurant	Food	Train Station	Diner
2	Arden Heights	Pharmacy	Deli / Bodega	Coffee Shop	Bus Stop	Pizza Place	Flea Market	Factory
3	Arlington	Deli / Bodega	Coffee Shop	Bus Stop	Home Service	Boat or Ferry	Grocery Store	Yoga Studio
4	Arrochar	Bus Stop	Deli / Bodega	Italian Restaurant	Hotel	Middle Eastern Restaurant	Pharmacy	Bagel Shop

3.2 – Step 2: Configuring the Toronto Data:

- All the same general steps as 3.1 were followed to obtain and clean up the Toronto data, with some key differences:
    - To obtain the initial data, we used BeautifulSoup python package to scrape a wikipedia page which displayed the postal code regions for Toronto, the associated neighborhoods, and their respective buroughs. To add the latitude and longitude fields, we utilized the Nominatim package to extract the different neighborhoods within each postal code region; and average their latitude/longitude values.

- Our final dataframe considers each 'neighborhood' as a postal code region. This is because of how the data was initially formatted for Toronto. Therefore, our final dataframe shows all the top 10 most common venue types present in each *postal code* region rather than specific neighborhoods. For the sake of simplicity, when referring to Toronto in this project, we will use 'neighborhood' term in palce of postal code region.

	PostalCode	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	
0	M1E	Electronics Store	Yoga Studio	Discount Store	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Event Space	F
1	M1G	Korean BBQ Restaurant	Yoga Studio	Fish & Chips Shop	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Event Space	F
2	M1H	Caribbean Restaurant	Thai Restaurant	Hakka Restaurant	Athletics & Sports	Bank	Farmers Market	Falafel Restaurant	
3	M1J	Playground	Yoga Studio	Diner	Farmers Market	Falafel Restaurant	Event Space	Ethiopian Restaurant	
4	M1M	American Restaurant	Motel	Yoga Studio	Discount Store	Farmers Market	Falafel Restaurant	Event Space	F

3.3 Step 3 - Compare NYC and Toronto neighborhoods by common venues only:

- First, we determined which common venue categories existed between NYC and Toronto. We created a function to determine this, and discovered that there are 202 venue categories common between the two.
- We then modified both our NYC and Toronto dataframes to only include the common venue categories shared by both cities, and dropped the other columns of venue categories.
- Now, we have two dataframes that are ready for comparison – one for NYC and one for Toronto – which display the top 10 most popular venue categories for each neighborhood in the respective cities, only considering the venue categories that are common amongst both cities.

3.4 Step 4 - Determine the best city and neighborhoods for a particular venue category:

We created a function called best_city which takes in two parameters:
- **venue_type (String):** the category of the venue we are interested in (e.g: 'Fast Food Restaurant')
- **view (Bool):** True if we want to get a the neighborhoods which are most popular for this venue type, False if we just want to know which city is the most popular.

Our function takes in a venue category, and performs data analysis to determine which city is more popular for that venue category. Then, if the user wants, the function can return a dataframe of all of the neighborhoods in the more popular city which have the highest frequencies for that venue category.

<u>How the function works:</u>

- (1) determine the popularity score for the specified venue category in NYC and Toronto. (For example: we want to determine which city is more popular for 'Korean Restaurant')
    - We call another function called city_score. This function grabs the dataframe for each city that contains the top 10 most common venues in each neighborhood. It then filters out the rows in such a way that only the neighborhoods which contain our specified venue category in its top 10 most common venues remain. Then, we will find out *how* popular (i.e is it the first most common venue category? Second most? Third..? etc) that venue category is for each neighborhood in the city, and come up with an average popularity score for each city.
    - For example:
        - we would end up with a dataframe for NYC which only contains the neighborhoods which have 'Korean Restaurant' as one of their top 10 most common venue categories. Our algorithm then goes through each neighborhood and sees which index 'Korean Restaurant' is found within the top 10 most common venue category types. For sake of example, let's say there are only 4 neighborhoods in NYC that have korean restaurant within their top 10 venues list. In neighborhood 1 – korean restaurant is the #1 venue type, in neighborhood 2 – it's the #2 venue type, neighborhood 3 - #3 venue type and neighborhodo 4 – it's the #4

more common venue type. We then calculate an 'average' popularity score:

popularity score = $\frac{(sum\ of\ all\ popularity\ indexes)}{number\ of\ neighborhoods}$

in our example, the popularity score for NYC would be (1+2+3+4)/4 =2.5 for NYC.

- We would repeat this process for Toronto.

- Once we have both scores, we check to see which city has a higher score. We call that the "win city"

- (2) Return dataframe – if specified
  - Sometimes the user might just want to know the name of better city for that venue type and that's it. In those cases, they would pass in 'False' in the original function, and they would just get the result as the name of the city which is more popular, alongside that city's popularity score.
  - If the user wants to know more details – i.e not just the city that's more popular, but which neighborhoods within that city are most popular for that venue category, they would input 'True' in the original function. This would then call another function which filters out the city's dataframe and returns only the rows where the specified venue type is found anywhere within the top 10 categories columns.

4. Results:

While we can use the best_city function to tell us which city and neighborhoods within that city are best for a specific venue type, we can also use this function to generate a dataframe which shows us for ALL venue categories what the best city is.

Here is a snippet of the dataframe:

From this snippet, we can see that New York is more popular than Toronto for: Accessories, Adult boutique, Arepa restaurants, art galleries, arts & crafts stores, asian restaurants, athletics & sports and bakeries. Meanwhile Toronto is more popular than New York for airport terminals, american restaurants, BBQ joints.

	Category	Most Popular City
0	Accessories Store	New York
1	Adult Boutique	New York
2	Airport Terminal	Toronto
3	American Restaurant	Toronto
4	Arepa Restaurant	New York
5	Art Gallery	New York
6	Arts & Crafts Store	New York
7	Asian Restaurant	New York
8	Athletics & Sports	New York
9	BBQ Joint	Toronto
10	Bakery	New York

The whole list of venue categories that are more popular in Toronto: TOTAL = 69

```
 'Airport Terminal',                        'French Restaurant',
  'American Restaurant',                     'Garden',
  'BBQ Joint',                               'Gas Station',
  'Bank',                                    'Gastropub',
  'Bar',                                     'Gay Bar',
  'Beer Bar',                                'Gluten-free Restaurant',
  'Beer Store',                              'Golf Course',
  'Bike Shop',                               'Grocery Store',
  'Boat or Ferry',                           'Gym / Fitness Center',
  'Bookstore',                               'Health & Beauty Service',
  'Brewery',                                 'History Museum',
  'Burrito Place',                           'Ice Cream Shop',
  'Bus Line',                                'Italian Restaurant',
  'Bus Stop',                                'Jazz Club',
  'Camera Store',                            'Malay Restaurant',
  'Candy Store',                             'Market',
  'Cheese Shop',                             'Middle Eastern Restaurant',
  'Comfort Food Restaurant',                 'Modern European Restaurant',
  'Concert Hall',                            'Moroccan Restaurant',
  'Cosmetics Shop',                          'Movie Theater',
  'Cupcake Shop',                            'Photography Studio',
  'Dance Studio',                            'Polish Restaurant',
  'Deli / Bodega',                           'Record Shop',
  'Design Studio',                           'Salad Place',
  'Dessert Shop',                            'Seafood Restaurant',
  'Dim Sum Restaurant',                      'Shopping Plaza',
  'Tech Startup',                            'Smoothie Shop',
  'Tennis Court',                            'Social Club',
  'Thai Restaurant',                         'Sporting Goods Shop',
  'Theater',                                 'Sports Bar',
  'Turkish Restaurant',                      'Tea Room',
  'Wings Joint'
```

The whole list of venue categories that are more popular in NYC: TOTAL = 139

```
 'Accessories Store',
  'Adult Boutique',
  'Arepa Restaurant',
  'Art Gallery',
  'Arts & Crafts Store',
  'Asian Restaurant',
  'Athletics & Sports',
  'Bakery',
  'Baseball Field',
  'Basketball Court',
  'Bistro',
  'Board Shop',
  'Boutique',
  'Bowling Alley',
  'Breakfast Spot',
  'Bubble Tea Shop',
  'Building',
  'Burger Joint',
  'Bus Station',
  'Butcher',
  'Café',
  'Caribbean Restaurant',
  'Chinese Restaurant',
  'Chocolate Shop',
  'Clothing Store',
  'Cocktail Bar',
  'Coffee Shop',
  'Colombian Restaurant',
  'Comic Shop',
  'Construction & Landscaping',
  'Convenience Store',
  'Creperie',
  'Cuban Restaurant',
  'Department Store',
  'Diner',
  'Discount Store',
  'Dog Run',
  'Donut Shop',
  'Dumpling Restaurant',
  'Electronics Store',
  'Ethiopian Restaurant',
  'Event Space',
  'Falafel Restaurant',
  'Farmers Market',
  'Fast Food Restaurant',
  'Field',
  'Fish & Chips Shop',
  'Fish Market',
  'Flower Shop',
  'Food Court',
  'Food Truck',
  'Fountain',
 'Pub',
  'Ramen Restaurant',
  'Restaurant',
  'Salon / Barbershop',
  'Sandwich Place',
  'Shoe Store',
  'Shopping Mall',
  'Skating Rink',
  'Snack Place',
  'Soup Place',
  'Spa',
  'Speakeasy',
  'Steakhouse',
  'Strip Club',
  'Supermarket',
  'Supplement Shop',
  'Sushi Restaurant',
  'Taco Place',
  'Tailor Shop',
  'Tanning Salon',
  'Theme Restaurant',
  'Thrift / Vintage Store',
  'Toy / Game Store',
  'Trail',
  'Vegetarian / Vegan Restaurant',
  'Video Game Store',
  'Vietnamese Restaurant',
  'Warehouse Store',
  'Whisky Bar',
  'Wine Bar',
  'Wine Shop',
  "Women's Store",
  'Yoga Studio'
 'Fried Chicken Joint',
  'Fruit & Vegetable Store',
  'Furniture / Home Store',
  'Gaming Cafe',
  'General Entertainment',
  'Gift Shop',
  'Gourmet Shop',
  'Greek Restaurant',
  'Gym',
  'Hardware Store',
  'Hobby Shop',
  'Home Service',
  'Hookah Bar',
  'Hotel',
  'Hotel Bar',
  'IT Services',
  'Indian Restaurant',
  'Indie Movie Theater',
  'Intersection',
  'Irish Pub',
  'Japanese Restaurant',
  'Jewelry Store',
  'Juice Bar',
  'Korean Restaurant',
  'Lake',
  'Latin American Restaurant',
  'Liquor Store',
  'Lounge',
  'Martial Arts School',
  'Mediterranean Restaurant',
  "Men's Store",
  'Mexican Restaurant',
  'Miscellaneous Shop',
  'Mobile Phone Shop',
```

```
'Monument / Landmark',
'Motel',
'Museum',
'Music Venue',
'New American Restaurant',
'Noodle House',
'Office',
'Opera House',
'Organic Grocery',
'Other Great Outdoors',
'Park',
'Performing Arts Venue',
'Pet Store',
'Pharmacy',
'Pizza Place',
'Playground',
'Plaza',
'Poke Place',
'Pool',
'Print Shop',
```

5. Discussion / Conclusion:

In conclusion, we can observe that Toronto and New York City, though they are considered 'sister cities' due their vast similarities – they each have their unique subsets of venues which are more popular (i.e have more neighborhoods which these venues are more frequently prominent). Of the 202 venue types that are shared between the two cities, New York City is more popular for 139 of them, and Toronto for 69 of them. The specific venues that each city is more 'popular' for can be seen in the results section. The differences between the cities have many possible contributing factors, such as the different demographics of people living there and the overall differences in city populations. NYC itself houses over 8.4 million people, while Toronto's population is just shy of 3 million.

Limitations and suggestion for further improvement:

In performing this analysis, there are a few limitations that hinder the accuracy of the results. The primary limitation is the definition of 'popularity.' For the sake of simplicity with the data that was acquired for this project, I calculated popularity of a venue category by observing how frequently a venue type appeared in each neighborhood relative to other venue types within that same neighborhood. When comparing the two cities, the more 'popular' city for a venue type would be based off of how many neighborhoods within that city had that venue type most frequently present *and,* amongst all of those neighborhoods, what the average frequency of that venue type was. This assumption would not be valid in real life due to some key factors that need to be considered:

1 – Not all neighborhoods are considered 'equal' and should have equal weighting. Different neighborhoods within the cities have different populations and sizing, but in this project's calculations they are all considered equal and have an equal affect on the popularity score of the venue types, when in reality the average calculations hould be a weighted average based on metrics like area size and population. For further improvement, data on area size and population should be included and factored in.

2 – our basis of popularity is simply on the frequency, however, in reality if someone wants to know which city is more well known or popular for a certain venue category – they might want to not only know frequency, but also how good those venues are. This can be done utilizing the Fousquare API and making premium calls to not only capture the venues, but see additional features like ratings and reviews to see the quality of the venues themselves. This additional data analysis was not performed in this project due to limitations on the number of premium calls for free users of the API, but should heavily be considered for anyone who wants to create a more accurate finding on popularity between the two cities.

Visualizations of the cities based on the different venue categories:

There are 202 common venue categories between NYC and Toronto. We want to display all the venues present in Toronto and NYC, but be able to easily distinguish what venue type/category each venue belongs to. We do this by using distinct colours for each marker that belongs to a certain venue type.

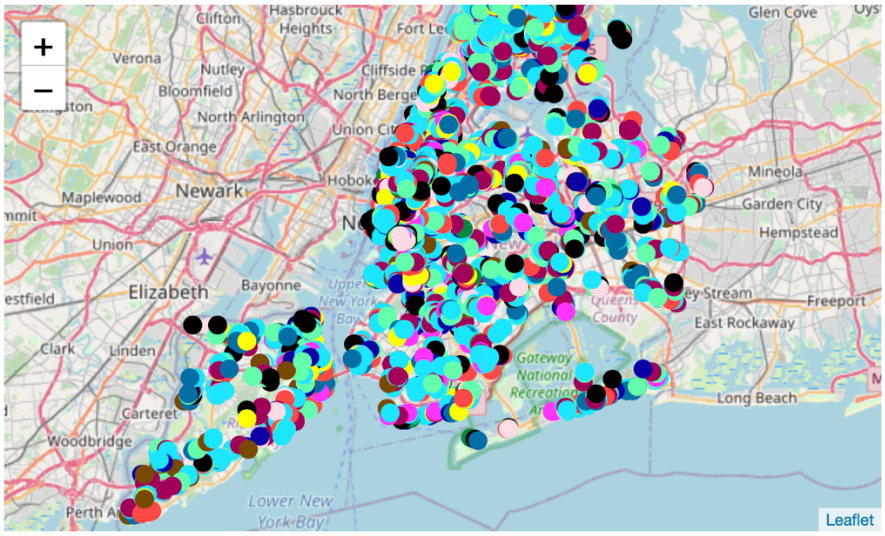
For simplicity, we grouped the 202 common venue categories into 13 unique categories, which we call the 'overall category':

Overall categories = [Outdoor, Café, Restaurant, Health & Beauty, Fitness, Arts & Culture, Travel & Hospitality, Shopping, Nightlife, Services, Entertainment, Food Shopping, Business]

Each of the above 13 categories are represented by a different color. If we click on the marker we can see the name of the venue and its overall category. By doing this, we can easily visualize which neighborhoods in each city are more dominated by certain venue categories, and the overall distribution of these categories throughout the city.

New York:

Out[180]:



Toronto:

Out[179]: