How can you attract more young working professionals to the cities of New York and Toronto?

**2. Data acquisition and cleaning**

The acquisition and cleaning of our data is a critical process which will determine whether the outcome of our project will be successful. Without due consideration for how we acquire our data and process it, our results will likely be inaccurate so any conclusions that we draw from these results will be unreliable. Therefore, it is a major step in any data science-based project to explain both how we came to acquire the data and the steps we took to properly process it. In this section, we will explore further how we came to acquire the data and the measures that we took to process this data so that we can ready and reliable use of the information during our methodological section.

**2.1. Data sources**

For this project, the data sources that we will be making use of are the New York and Toronto datasets that we utilised in Week 3 of the course to compare between different neighbourhoods of New York and Toronto. We will utilize these datasets to find out how both cities can best develop and market themselves to attract more young working professionals. These two datasets contain all the information regarding the neighbourhoods that are present in both New York and Toronto, this includes the neighbourhood’s latitudinal co-ordinates, the neighbourhood’s longitudinal co-ordinates, the neighbourhood’s postcode, and what borough of the city that the neighbourhood is in. To further add to this picture, alongside the Foursquare API, we will be utilising additional information to build up a comprehensive and well-rounded picture of each city as to how they can best develop and distinguish themselves to attract young working professionals.

**2.2. Data cleaning**

The data that we obtained for the datasets was initially scraped from Wikipedia using the panda’s framework in Python. We then combined the datasets using the Python Geocoder package with the neighbourhoods latitudinal and longitudinal co-ordinates to give us a nice rich dataset that when combined with the versatile features of the Foursquare API can enable us to explore the neighbourhoods in an unpresented level of detail. To avoid the potential duplication of postcodes, we combined and grouped the boroughs and postcodes of the datasets so that all the neighbourhoods which contained the same postcode fell into one row. We also performed a thorough cleansing of the data by removing all the results displayed as “Not assigned” where the postcode was noy assigned to any specific borough. In cases where a postcode had been assigned to a borough but had no specific assigned neighbourhood, we classified these neighbourhoods by the name of the borough.

After a thorough cleansing and shaping of our data had taken place, we had 103 results left out of our original 288 results. Though this is only 35% of our original sample, it is still enough results for us to build up a detailed picture of the different neighbourhoods present in each borough of both cities. Lastly, I checked if our dataset had any unusual or extreme outliers and since each borough of both cities is partnered to a postcode, our datasets contained no extreme or unusual values. There was no missing, incomplete, or seemingly inaccurate datapoints and therefore I judged that the data was prepared and ready for the next stage of further exploratory and detailed analysis.

**2.3. Feature selection**

Having processed the data to a high standard, all that was left was to determine the features to select and look for when performing further detailed analysis. From our cleansed dataset we can perform an extensive search with the Foursquare API to build up an extensive view of each borough of the cities such as through what venues they contain or the most present and popular facilities of that each neighbourhood contains. The features that will be selecting for further analysis will help us to build up an extensive insight into what facilities that is present within each neighbourhood. From this we can determine and filter our results to those that are most likely to appeal and attract young working professionals. Features such as: fast food restaurants, pizza places, gym and fitness centres, coffee shops, pubs, will take precedence over those such as: parks, airports and shops (as many young professionals now are more likely to purchase goods online).

By creating a detailed oversight into each neighbourhood and each borough this will allows us to form a idea of where the features are in the city that will best appeal to our target audience and how we can best market our cities in a distinguishable way so they can attract those young working professionals. No features therefore will be dropped from our datasets as young professionals will be likely interested in all the features which are present in their neighbourhood especially if they are considering taking up a permanent residence in these areas. The features will simply be clustered and then sorted appropriately, which when combined with our findings and insights gained from the performance of advanced statistical techniques; will make us be able to best display the distinguishable features present between them that are likely to make young professionals want to move or reside in these cities when we come to present and submit our findings to our stakeholders.