IBM data science capstone assignment submission

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

Arron Hovingham

16/09/2019

Contents

**Section 1: Introduction2**

1.1. Background2

1.2. Problem definition2

1.3. Interest3

**Section 2: Data acquisition and cleaning3**

2.1. Data sources3

2.2. Data cleaning3

2.3. Feature selection4

**Section 3: Methodology4**

3.1. Sorting the information4

3.2. Using the Foursquare API5

3.3. Displaying the frequency of the top results6

3.4. Displaying the top ten results6

**Section 4: Results7**

4.1. Predictive modeling 7

4.2. Data visualization8

4.3. Breaking down the clusters8

**Section 5: Discussion9**

**Section 6: Conclusion10**

1. **Introduction**

The cities of New York and Toronto lie only 300 miles apart from one another with both being the central financial hubs of their countries. These cities boast a range of important cultural landmarks that forms a key part of their respective national identities. Between them over 12 million people reside in the cities and due to both cities having experienced historically elevated levels of immigration, a diverse multitude of languages is spoken across both cities; so what separates these cities from one another? What are the underlying factors that distinguish the cities between one another? Really just how similar are these cities? In this report, we will explore the neighbourhoods of New York and Toronto further using the Foursquare API and be performing an in-depth analysis on both the New York and Toronto csv files to come to a determination of how best to attract more young working professionals to both cities.

* 1. **Background**

As part of my IBM data science capstone project, I have been tasked with using my initiative to come up with an idea that leverages the Foursquare API to explore cities and neighbourhoods of my choosing. Since the course gave me access to rich and highly detailed neighbourhood data on the boroughs of New York and Toronto I wanted to make use of the datasets and really unlock the true value and potential of them. New York and Toronto are both major commercial hubs in wealthy advanced economically developed countries which are critical to the development of wealth in not only the North American region but to the wider global economy. Through performing an extensive and detailed analysis of both cities, we will gain unique insight into the factors that distinguish these two cities from one another, that on a broader scale may prove beneficial if we were to further this project into data driven methodologies to differentiate between other major commercial cities in the world, particularly those in a relatively close proximity to one another such as London and Paris or Tokyo and Seoul.

* 1. **Problem Specification**

The problem I have decided to focus on is with the creation of a case study scenario. Say you are working as part of an American Canadian tourist partnership trying to attract more young working professionals between the ages of 21 and 28 to move to the North American region. How would you try to distinguish between the cities to promote them to an audience of young working professionals try to convince them to move and take up residence in either city? What exactly are the features that distinguish the two cities which will allow you to promote themselves in a distinct way to a wide audience of young working professionals? How can the cities promote themselves in an attractive manner so they can to tempt young professional to move to them?

Specifically therefore we can define the problem as:

*“How can you best promote the cities of New York and Toronto in a distinct yet attractive manner that will tempt more young working professionals to move and take up residence within these cities based off of the features of each city.”*

Here, we would need to therefore leverage the Foursquare API to find venues and locations that would appeal to our target audience of young working professionals. We will need to leverage the API to find popular locations for young professionals like libraries, coffee shops, nightclubs, and restaurants. From this we will then need to compare between the two cities, to recommend potential data driven marketing strategies for each city so that they can be distinguishable and best promoted to our target audience.

* 1. **Interest**

Cities spend millions of dollars on both trying to distinguish themselves from each other and promoting themselves to attract the top talent worldwide. Hence understanding how cities distinguish themselves from one another is critical for future city planning especially when you take into consideration that more and more people are moving away from rural areas to urban centres. As the population continues to increase, understanding how to effectively promote cities to a younger generation is critical for how cities develop in the future. With people now living to better standards and an age rate that is higher than ever before young working professionals are a crucial element for a prosperous city development in maintaining the healthy development of a city. Recognising the features of a city that will attract and bring more young working professionals to it, is key therefore to enabling a city to realise and unlock its true future potential.

**2. Data acquisition and cleaning**

The acquisition and cleaning of our data is a critical process which will determine whether the outcome of the project will be successful. Without due consideration for how we acquire the data and process it, the results will likely be inaccurate so any conclusions that will be drawn 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, additional information will be utilising 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 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 not assigned to any specific borough. In cases where a postcode had been assigned to a borough but had no specific assigned neighbourhood, these neighbourhoods were classified by the name of the borough.

After a thorough cleansing and shaping of the data had taken place, there was 103 results left out of the original 288 results. Though this is only 35% of the original sample, there are still enough results to build up a detailed picture of the different neighbourhoods present in each borough of both cities. Lastly, the datasets were checked to see if they 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 it was 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 the cleansed dataset an extensive search was performed 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 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).

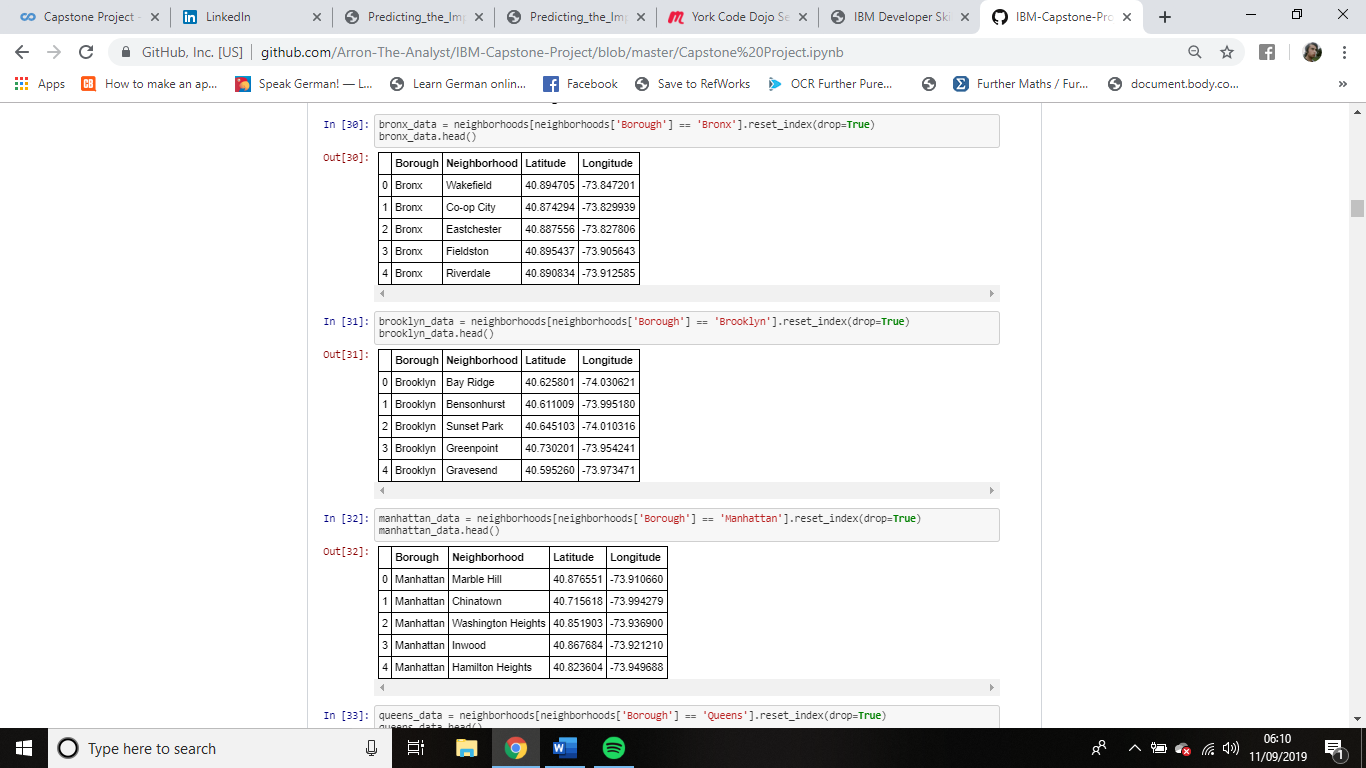
By creating a detailed oversight into each neighbourhood and each borough this will allow the formation of an idea of where the features are in the city that will best appeal to the target audience and how we can best market our cities in a distinguishable way so they can attract those young working 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.

**3. Methodology**

Having both acquired and cleansed the datasets, the next section of the process is to perform the methodology to allow for results to be gathered and collated. The methodological section is what underpins the quality and accuracy of the acquired results and therefore is an integral part of the overall project. In this section, we will be describing and discussing the exploratory data analysis that was performed so that results could be obtained through predictive modelling methods.

**3.1. Breaking down and sorting the information by borough**

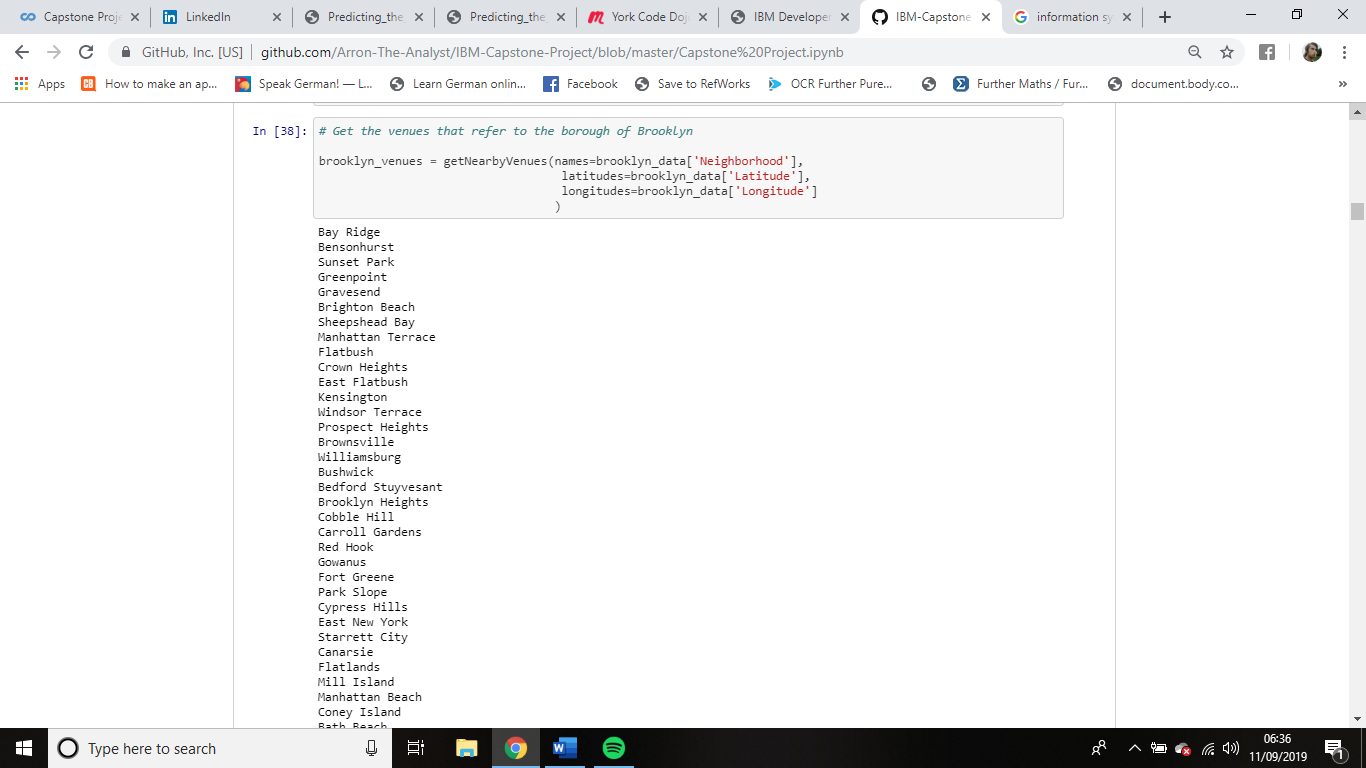
To start off with, the data was firstly sorted grouped by the different boroughs within each city. As New York has five different boroughs: the New York dataset was grouped into five separate data frames, and as the Toronto has 11 different boroughs: the Toronto dataset was grouped into 11 separate data frames. In Figure 1 we see an example of how the neighbourhoods were broken down into specific boroughs so that further in-depth analysis could take place.



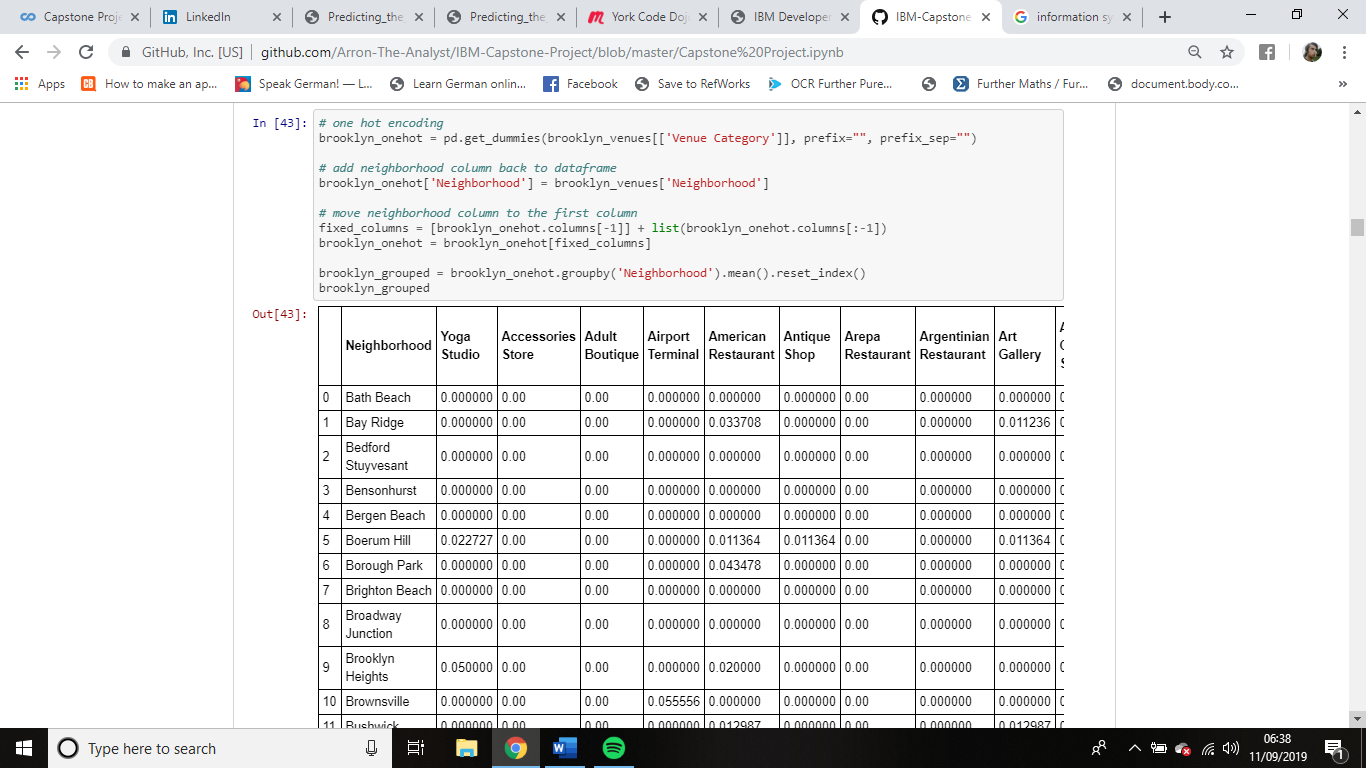
*Figure 1. The original dataset being broken down by every different borough.*

**3.2. Using the Foursquare API to get information on the venues in each borough.**

Having sorted the information into the relevant boroughs, the Foursquare API was then used to get the knowledge about every location that is in each borough of the datasets. (five for the New York dataset and 11 for the Toronto dataset). In Figure 2, we see the code that was used to retrieve these facts with. Once this knowledge was gathered, we then created a frequency table to display how frequently these venues appear in every neighbourhood within each borough using one hot encoding. In Figure 3, we see the results that were generated from this process allowing further detailed analysis to be undertaken into the makeup of every borough.



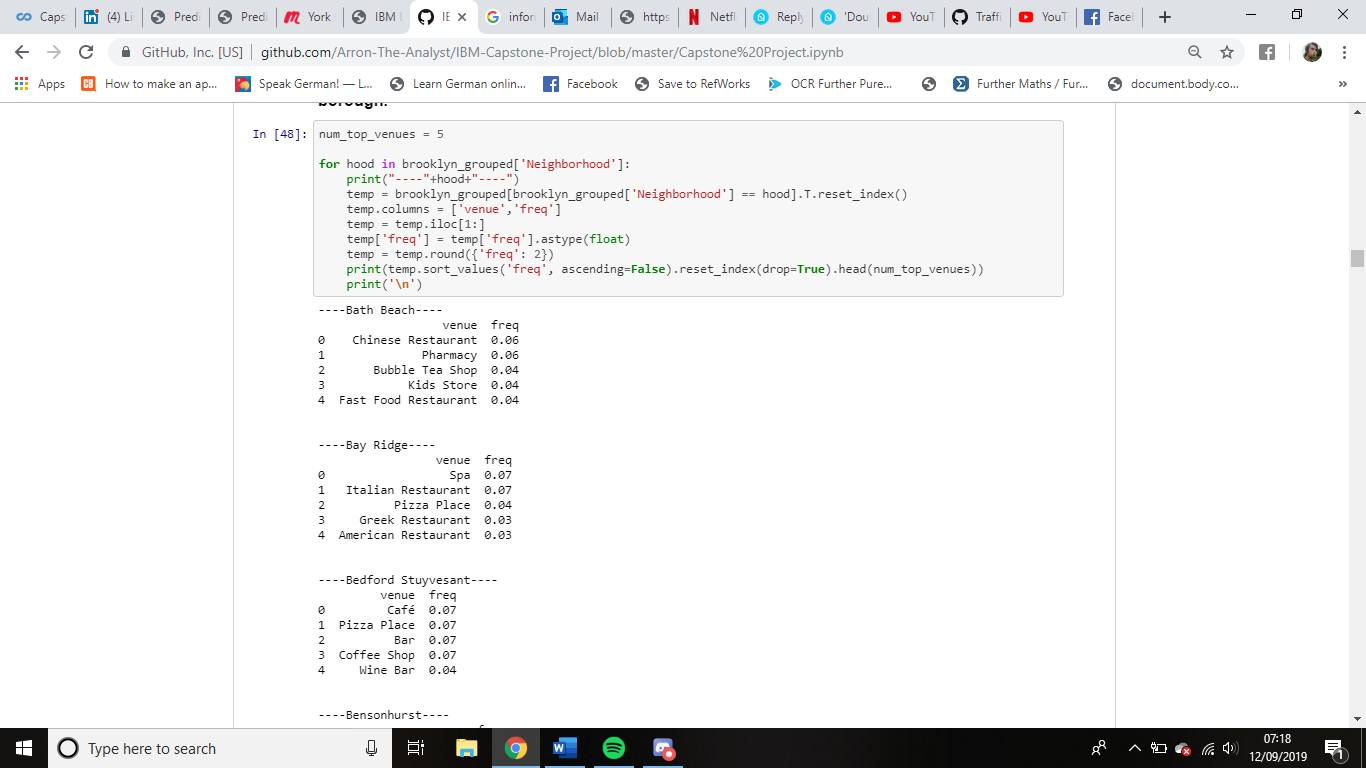
*Figure 2. Retrieving the data on each venue in every borough using the Foursquare API*



*Figure 3. A grouped frequency table displaying the frequency of every venue in each neighbourhood*

**3.3. Displaying the frequency of the top five venues within each neighbourhood**

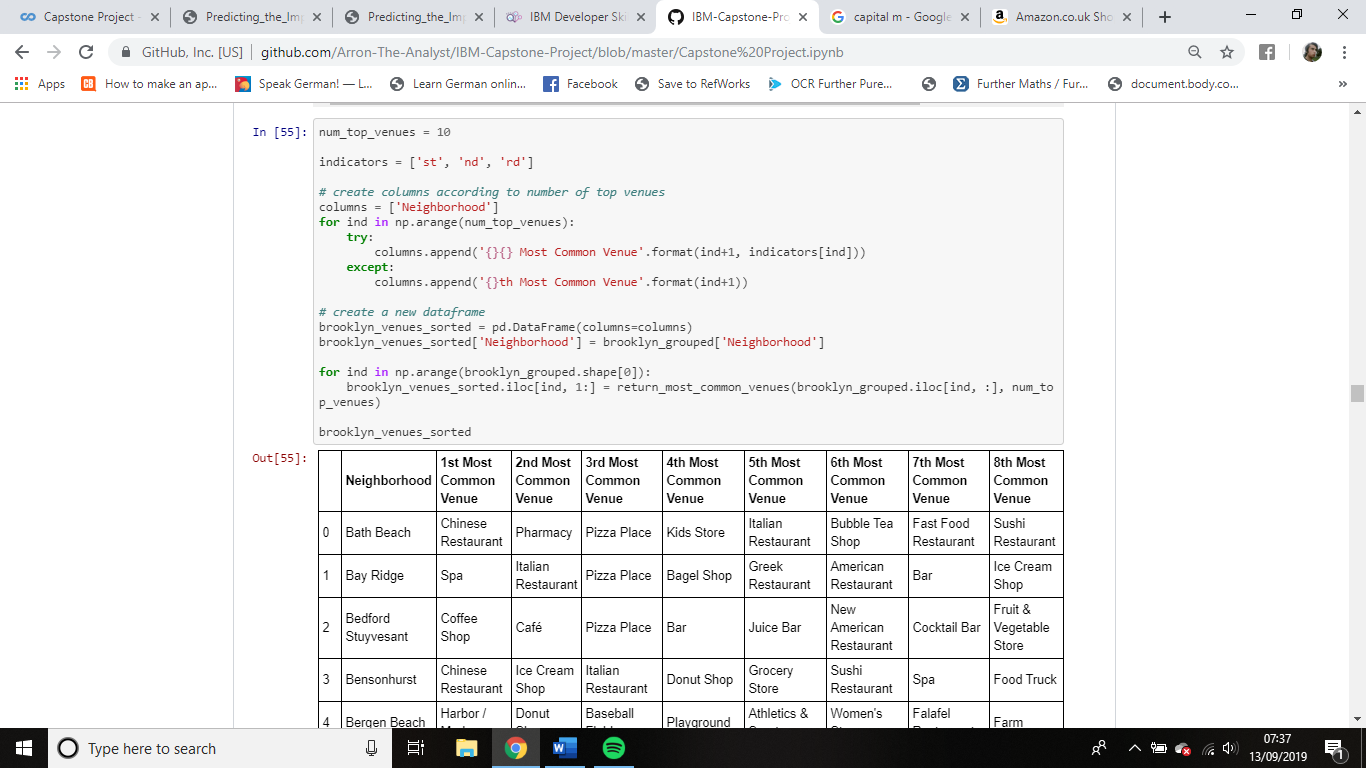
Having obtained the frequencies for every venue type in each neighbourhood, next the results were filtered so that only the top five venues in each neighbourhood were displayed. What this does is allow for us to explore further the most common venues types in each neighbourhood enabling us to create a comprehensive understanding of what neighbourhoods might appeal to young working professionals. In figure 4, we see the code used to display these grouped frequency tables.



*Figure 4. A grouped frequency table displaying the top five venues in each neighbourhood*

**3.4. Displaying the top ten venues within each neighbourhood**

To make the calculated frequencies for how often the top five venues appear within each neighbourhood for every borough within the New York and Toronto datasets easier for us to visualise, a new data frame was created displaying the top ten venues of every neighbourhood for each borough. In figure 5, we see the code that enabled us to achieve this, allowing for further analysis to take place for us to better assess which areas might appeal to young working professionals in both cities.



*Figure 5. A simple frequency table that displays the top ten venues for every neighbourhood*

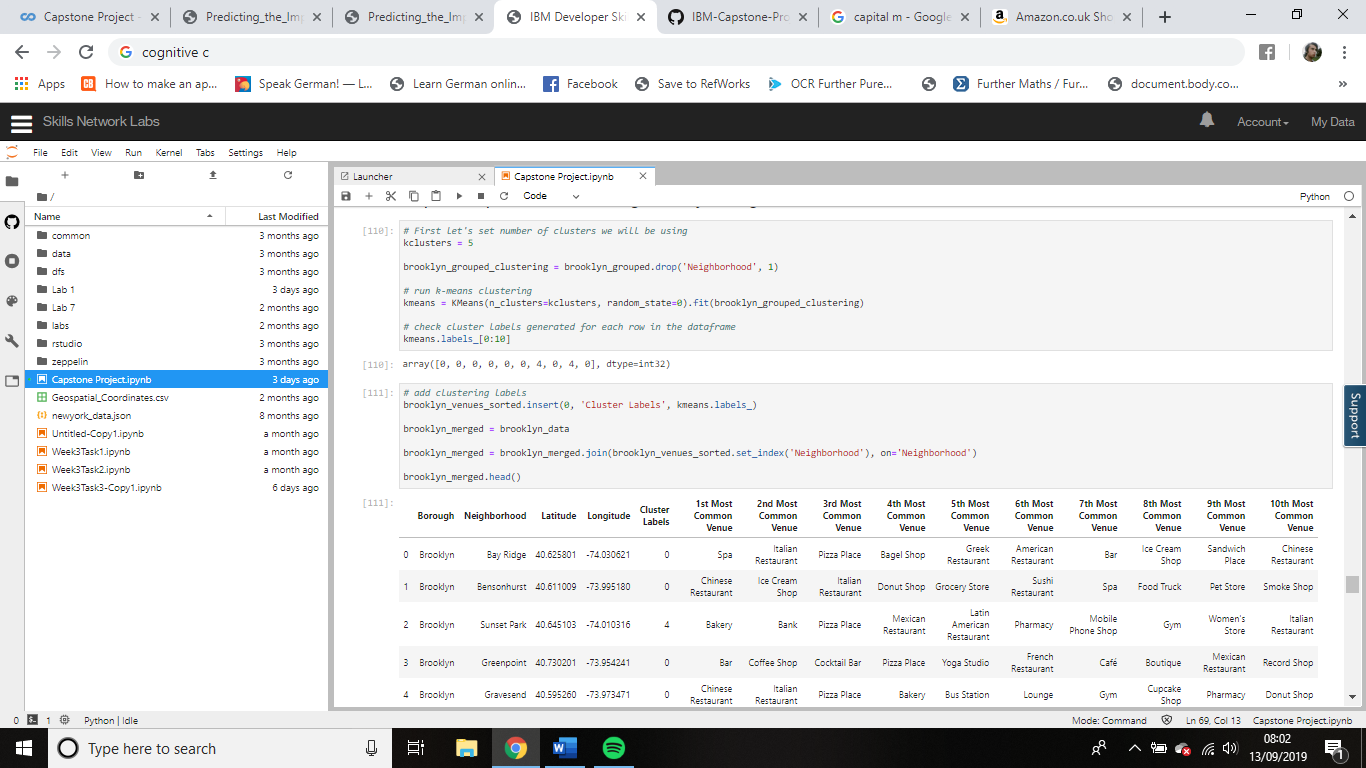
**4. Results**

Having completed the exploratory data analysis through the methodological section, the next stage of the process is to perform predictive modelling so that results can be obtained for us to discuss and be able to conclude appropriately. In this section, we will outline the steps taken to achieve this by using inferential statistical testing and machine learnings techniques alongside discussion surrounding the rational and reasoning behind why these techniques were chosen to be utilized.

**4.1. Predictive modelling**

There are two distinct types of predictive modelling methods which could been undertaken to model the most common venues within each neighbourhood so that they can be best presented such that detailed analysis can take place surrounding which areas of both cities might appeal best to young working professionals, regression models and classification models. Regression models can provide additional information on quantitative values that can enable further differentiation between the neighbourhoods, while classification models focus more on the overall picture to predict which neighbourhood will appeal best to the target audience. In both instances the underlying algorithms are similar between regression and classification models, however as classification models are more interpretable in this study only classification modelling was undertaken.

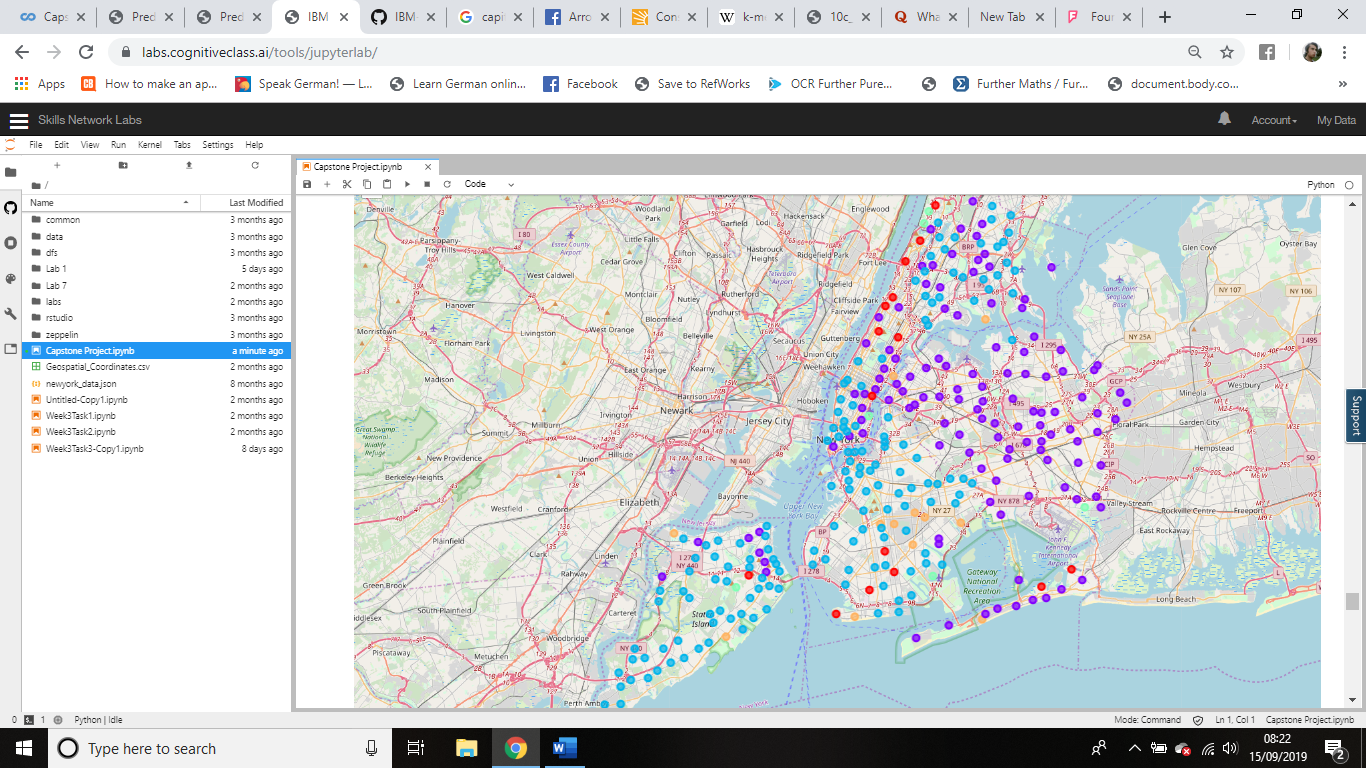
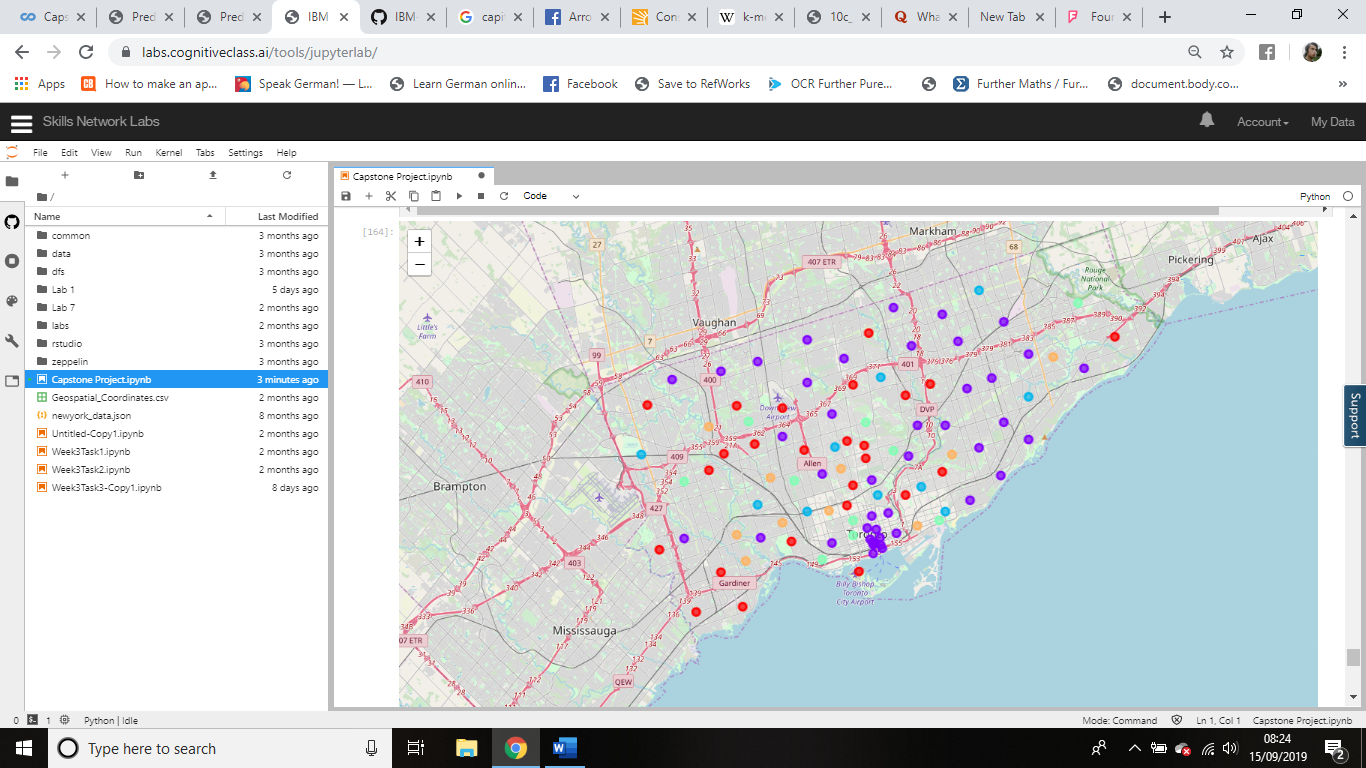
Specifically, this study opted to use the advanced machine learning technique of K means clustering to predict how similar or dissimilar each neighbourhood is from one another. Though other classification options were considered such as Support Machine Vectors (SVM), random forests, and mixed gradient boosting methods, after careful deliberation it was clear that K means clustering was the best choice of method to use as it is simple to implement, scales well to large datasets, and easily adapts to new examples. This technique was then performed using five different clusters on every borough within both the New York and Toronto datasets. In figure 6, we see the code that was used and the results that were gathered from the process.



*Figure 6. The code used and results obtained from performing K-means clustering on the datasets of every borough.*

**4.2. Data Visualisation**

Once we had clustered each dataset using the k means clustering method, next the results from the process were plotted onto a map of each city. In figures 7 and 8 we see the results from this process, with each colour representing a separate cluster type.

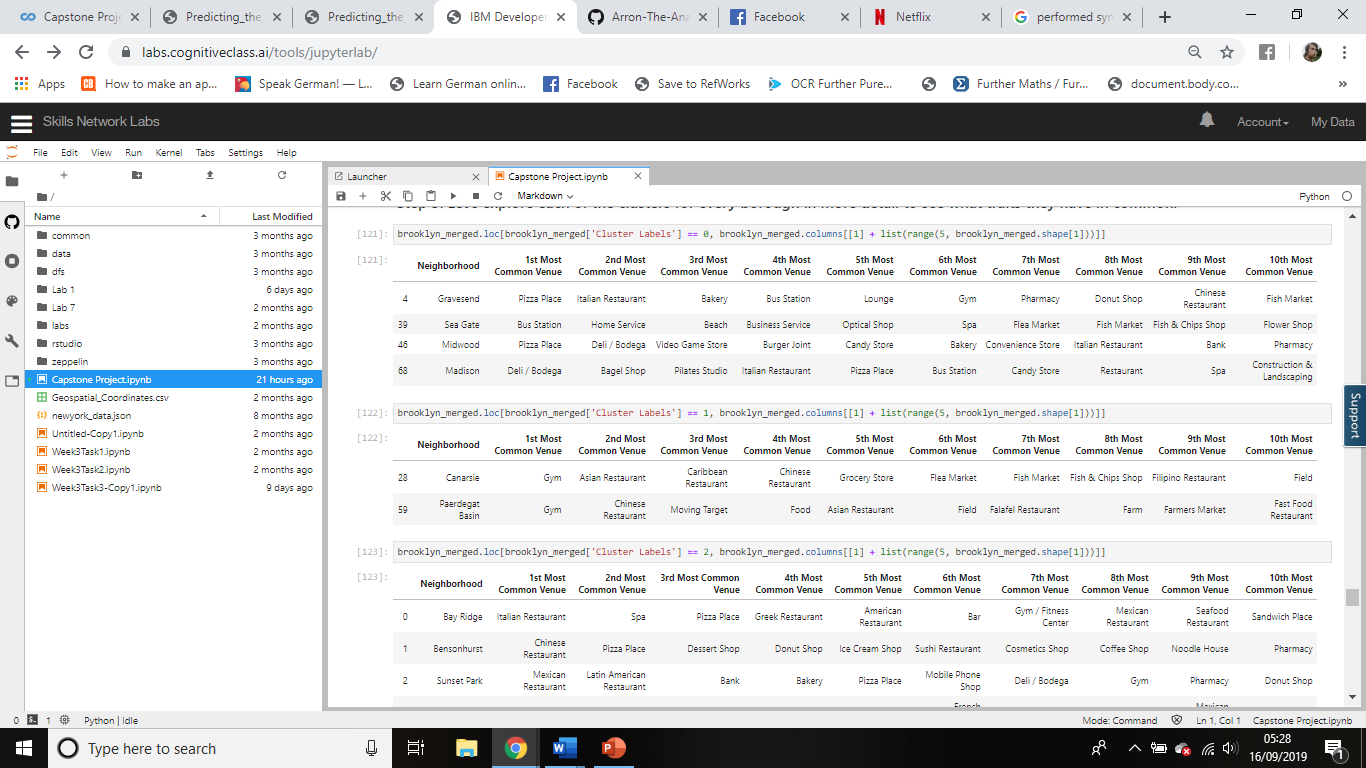


**‘**

*Figure 7. A map of New York City displaying the generated clusters Figure 8. A map of Toronto displaying the generated clusters*

**4.3. Breakdown of the clusters**

Lastly, having clustered the data and visualised it in an appropriate fashion, the last step is to break down every cluster type so that of the similarities and difference between each cluster type. From this we will be able to draw judgements surrounding which cluster type, neighbourhoods, and boroughs of both cities will best appeal and be suited towards the needs of a young working professional person. In figure 9, we see the code used to gather this information, which was repeated for every cluster generated within every borough for both cities datasets.



*Figure 9. A breakdown of code used to show the neighbourhoods within each cluster for every borough in our datasets.*

**5. Discussion**

In this study, I analysed the makeup of neighbourhoods within the cities of New York and Toronto to see which best appeal to young working professionals looking to move to these cities. This was achieved through the construction of a sophisticated classification model that made use of the k means clustering technique to predict which neighbourhoods would best attract young working professionals. By doing so, the aim of this project was to use these generated models to help marketers best promote both cities in a contrasting way to get more young working professionals moving to North America so that they wouldn’t clash with one another.

More and more, young working professionals are away from rural areas and into urbanised centres. Therefore collecting and analysing data around the makeup of our urban centres is critical to the future development and strategies of our cities. These datasets of our urban surroundings can be extraordinary useful to use in giving us detailed insights surrounding various parts of cities. For example, they can be used to predict the future crime rate of neighbourhoods, predict population increases and decreases within neighbourhoods, understand the most popular languages that are being spoken in regions of our cities. In short, datasets which present accurate information on our surroundings, allow us to better strategize and develop appropriate to what an area really needs.

Once the results were gathered, analysis showed that the areas predicted as the ones that will most likely appeal to young working professionals, are those with a high density of bars, cafes, restaurants, and social spaces such as libraries. There are many reasons as to why this might be but ultimately it is because most young working professionals like to be able to socialise after work, and therefore areas which have more spaces in which they can do this, will attract them more than spaces which cannot though this is not an entire representation of the whole demographic. To further add to our analysis the foursquare API was utilized to add further dimensional value and be able to relay insights in real time allowing us to see which venues were trending at which times among young people. From this, we were therefore able to deduce that spaces which were clustered as orange are the most likely to appeal to young working professionals as these spaces contained the highest density of social spaces likely to appeal to our target audience.

**6. Conclusion**

In this report, we set out to compare the cities of New York and Toronto and see how each city could best attract young working professionals and convince them to move to the North American continent. This was done through the performance of detailed analysis and the construction of a highly accurate clustering model which was able to conclude that areas highlighted in orange are the best areas for young working professionals to relocate to within both cities.

However, where we could have improved within our project could be by considering further datasets to get a heuristic view on why specific areas are best for young working professionals. For example, in our analysis we didn’t consider crime rate data, and so areas which we could have suggested may in fact turn out to be the worst areas of the city for knife crime or burglaries. Alongside this, a further classification model could have been built to contrast with our K means models such as through generating an SVM model which would allow us to see how good of a model is using K means clustering, and as no optimization techniques were utilised this to could be an additional way of how the model could have been further improved upon.

Had we had more time, there are many ways in which we could improve and further develop our project to take it to an even higher level. By making use of more diverse types of data this would help improve model performances significantly. Models in this study mainly focused on specific features but they could have been extended and tailored towards specific individual needs when looking to relocate to an area within both cities, such as if our young working professional had specific dependencies like children. Added to this, we could have analysed further cities in close geographical proximity to see and contrast how they compete and make themselves stand out differently, such as with the cities of London and Paris, or Shanghai and Hong Kong.

Overall though, I am very satisfied with the work I have completed as part of my capstone project. The work that has been done, especially within the coding section, is to an extremely high standard and is some of the best code I have developed yet. The course itself has been very entertaining but challenging and has tested and pushed the limits my technical abilities on several occasions. I hope to continue building on the techniques I have learnt from this course as I start my master’s course in Data Science and Statistics at the University of Bath in the coming days.