Coursera Capstone

IBM Applied Data Science Capstone

Toronto, Ontario: A Shopping Mall venture

Humble endeavor of: Jay S

Introduction

With the advent of shopping malls as an internationalized idea of the European **piazza,** shopping became more socialized than ever, with people now being able to not only shop, like they used to when frequenting shopping streets or plazas, or dine at restaurants, but do all of that under the same roof, which served as the foundational vision for the creation of the first mall by the adept architect Victor Gruen, who was commissioned to build a shopping center in Edina, Minnesota. But, with the perpetually increasing consumer base and retail brands, it becomes an imperative for a retailer or a marketing strategist for brands to better assess the locale, and then decide which is the **right place to be in and what is the right time**. General baseline for all the retailers planning a market move, is usually, **central location** and **populace with ample buying capacity**. Furthermore, commercial undertakings like malls, also provides a steady rental income stream to the property developers, and with the city of Toronto, Ontario being the epicenter of Canada's top earning shopping areas, according to the **Canadian Shopping Center study** of 2019, with 8 shopping malls, out of the top 30 malls with the highest profitability per square foot according to the official records of Retail Council of Canada, located in Toronto, with the top spot actually going to the Yorkdale shopping center in Toronto with the average productivity of over $1900 per sq. ft , and thusly ,it is a jackpot of an opportunity for retailers and property developers alike, but before making the move the question of **WHERE** arises, and the main intent of this project is to tackle that, with the apt use of machine learning paradigms and data science.

Business Problem

The capstone project intends to ease off the load, of manually analyzing the different neighborhoods in the city of Toronto, ON, and undertake the tedious task of mining the data, to find informative pattern, by utilizing the various machine learning paradigms namely clustering algorithms and data science methodology, and to find answer to the question of "**Where is the right locality for a shopping mall venture**", and to assess "**If the venture would be profitable".**

Target audience of the project

The capstone aims to cater the cohort of property developers, retailers and marketing strategists, by providing elucidating insights about the viability of a locality in sustaining the growth of the mall, furthermore, investors planning to invest in the sector, can also be greatly benefitted by the same. To bolster the relevance of the project undertaken, the city of Toronto, Ontario, Canada, has been described by Canadian shopping center study, to be the epicenter of mercantile revolution , with the average earning grossing to $986 per square foot, but at the same time according to the same study also hinting that there is an unequal distribution of shopping complexes with majority of the outskirts relatively unreachable to the said malls. The Retail Council of Canada (RCC) also points out that there has been a massive increment in the investment in the shopping complex sector in the years and the trend is likely to be followed , therefore the opportunity of growth of a shopping mall in an area which has a paucity of well distributed, wholesome shopping complexes, together with the ever increasing investment into the sector, would definitely be an opportunity worth the toil for the said parties.

Data

To solve the issue, we will need the following data:

- List of neighborhoods in the city of Toronto, Ontario, Canada, also renowned as the “city of Big Smoke”, due to massive industrialization of the city
- Coordinate data for all the neighborhoods, constituting, the latitudes and the longitudes. It would be conducive to construct maps and get the venue data
- **Venue data,** pertaining to the shopping malls, using foursquare API, on which data clustering would be performed.

Source and extraction methodology of data

The first source of data is the interactive map provided by **https://www.scribblemaps.com/maps/view/Neighbourhoods_in_Toronto_Ontario/Unki2xMmX6**, and Wikipedia page “**https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M**”, which essentially contains a mapped layout of the 103 neighborhoods in the city of London, furthermore, the population data for the neighborhoods is to be scraped from the official website of Toronto city namely, https://www.toronto.ca/About-Toronto/community-statistics/neighbourhood-profiles, to append the population and the per capita income data into the database, where necessary for analysis, there will be a blend of strategies involved in the creation of a satisfactory dataset, which would include web scraping using the BeautifulSoup library, together with actually researching and appending population and per capita data, furthermore the coordinates of the same would be found out using the GeoPy library, and lastly the usage of RESTful APIs and Foursquare would yield the **results of the venue data pertaining to the shopping malls** for the neighborhoods which is required for further analysis, gauging the problem statement, one can easily deduce some crucial aspects which must be extracted from dataset, namely

the population of the neighborhood which **would be crucial in understanding if there is a large enough consumer base to instate a mall for** and the **per capita income** to better assess if the selected demographics is, as the business jargon describe**, a high return target or a low return target** . The business problem at hand would utilize an appropriate mix of data wrangling, data cleansing, data analysis, web scraping, data visualization using Folium library, researching and machine learning methodologies (K-Means Clustering) to be successfully solved. In the following section, we would be describing the plan of action and the steps taken to facilitate all the processes mentioned above.

Methodology

Firstly, we need to scrape data pertaining to the various boroughs and neighborhoods in the city of Toronto, which was available in the following Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M), as well as a more interactive map to bolster the data set, provided by Scribble Maps, (**https://www.scribblemaps.com/maps/view/Neighbourhoods_in_Toronto_Ontario/Unki2xMmX6**)**.** We proceed to undertake web scraping using the Python API requests and the BeautifulSoup packages, but the result of the scraping is just a text extraction of an LXML scraper, we need to transform the textual data into geographical coordinates, which is in turn done by the GeoPy library, and of particular relevance the Geocoder library, which essentially provides coordinates for an textual address entry. After the data wrangling, and cleansing, we populate the data into a pandas DataFrame, which is a python data frame, which eases out computational complexities, during coding. However as described, data science is all about the prowess of telling story, and to do so we transform the textual and almost dreadfully incomprehensible textual data into map, using the powerful Folium library, to elucidate the scope of data to the viewer, and providing sanity checks as to whether the data plotted is actually relevant or not

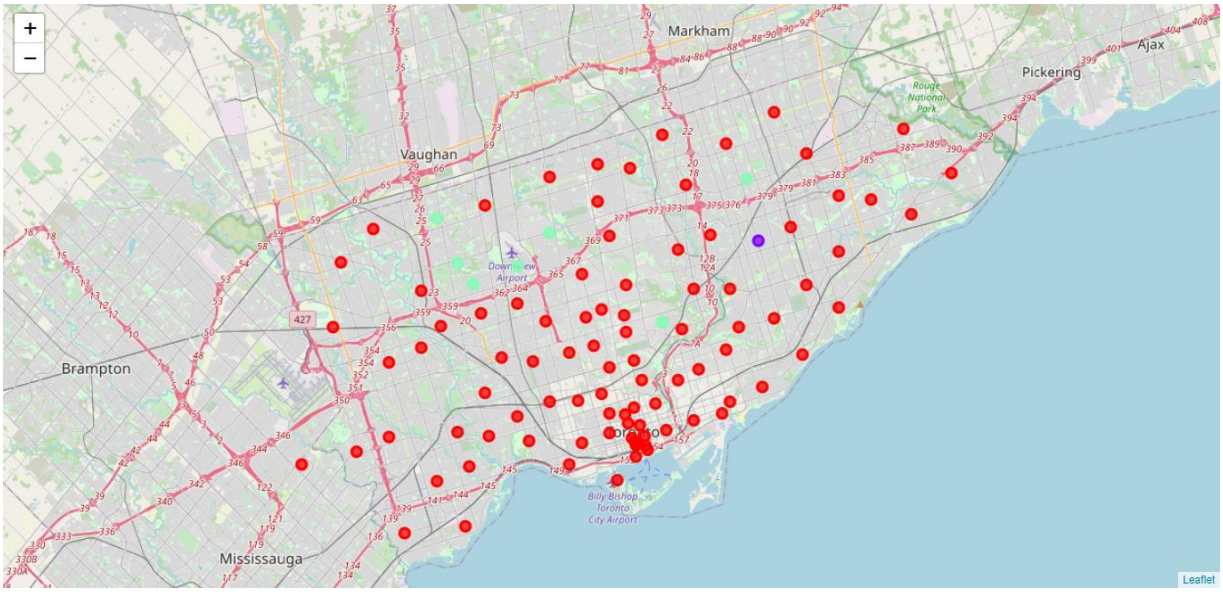
We then proceed to use the FourSquare API, to get the top 100 venues situated within the 500 meter vicinity of the said neighborhoods, but to be able to make API calls, an account creation is necessary, for the creation of Foursquare Secret ID and Secret Key, which are passed as perimeters to verify the authenticity of the API call. We iteratively call the FourSquare API with the coordinates of the different boroughs and neighborhoods using a python loop. The API returns the data request in JSON format, from which extraction of relevant information, namely **"Venue Name, Venue Category, Venue latitude, Venue Longitude"**, is done using apt data slicing techniques. With the dataset at hand, we then proceed to find out the unique number of venue categories grouped by neighborhood, to get a clearer picture of the venue preferences in each area, we then proceed to take the mean of the frequency of the said venues, to get a sorted list of popular venues, for the scope of this capstone, we analyze the "shopping mall" data in the venue category section.

Conclusively, we perform K-Means clustering on the shopping mall data. K-means clustering works by segmenting the data into K (a user defined value for number of clusters) clusters based upon the distance (Euclidian preferably) between the cluster center or **centroids** and the data points, with relatively shorter distanced centroids data points kept in the same cluster, an example of unsupervised learning model where the machine finds patterns in the data without any intervention, which is particularly apt for solution to the problem. We cluster the data into 3 clusters based upon the frequency of occurrence of "shopping malls", together with relevant population and average household income data, which would provide greater insight to the solution to the problem of "**Where and IF"** pertaining to shopping mall construction.

Results

The results of K-means clustering depict one very important aspect of shopping complex market in Toronto, which is the uneven distribution of wholesome shopping complex across Toronto, as is pointed out by the Retail Council of Canada, under the supervision of Government of Canada, furthermore, we also see a categorization of datapoints into three clusters based upon the frequency of occurrence of “shopping mall”

- Cluster 0: Depicting areas with relatively lower presence of shopping complexes and centers, implying lower market competition for the establishment of a new shopping mall venture. It is tagged with red circle marker.
- Cluster 1: Depicting area with relatively moderate presence of shopping complexes and centers, which implies a considerable market competition and difficulties in establishment of a new shopping mall venture. It is tagged with blue circle marker.
- Cluster 2: Depicting areas with relatively high presence of pre-existing shopping complexes, implying significant competition and difficulties in establishment of a new shopping mall venture. Tagged with green circle markers.



Discussions

As is evident from the map presented above, As is evidently visible the cluster 0 has a paucity in the sector of shopping malls and complexes and there is slight to almost negligible competition from the existing shopping mall complexes, and this presents a wonderful opportunity for construction of new shopping complexes, furthermore with areas such as **Scarborough**, with **population being 632,000** and **median income being around $59,000** , **Etobicoke**, with **population being 345,000 and median income being $70,500** and the areas of **York University** would definitely be a boon for upcoming mercantile businesses , following up is the cluster 1, which has a moderate presence of competitor shopping malls, the area of **Wexford and Maryvale** in Scarborough borough, estimated 30% of the population in Wexford being categorized as young and with a population of around 28,000 people and the maximum cohort earning in the tune of about $35,000 to $52,000, it is certainly going to be challenging for a new mall venture to be established here unless the developers and retail strategists have a unique USP to stand out in the competition.

Lastly we have the extremely competitive market of cluster 2 with the inclusive boroughs and neighborhoods, being among the more financial successful ones in Toronto, of particular notice is again the borough of North York and East York, the neighborhoods of Downsview, having a population of around 36,000 people and despite being a medium earning neighborhood is sprawling with low to mid-size shopping complexes, with the borough of North York with the average household income of around $82,000.

Therefore from the cluster analysis above, it becomes an imperative for any retail strategist or property developer to target the cluster 0, and more specifically the neighborhoods in the areas of **Scarborough and Etobicoke** and Downtown, as they house greater population as well as a higher income earning populace too, and with slim competition in the area, it becomes an even more lucrative investment area,

Cluster 1, with its medium competitive presence may pose certain challenges in the establishment of a new shopping mall venture, and while certainly not impossible, the retail strategists and the property developers alike might have to come up with USPs plausible enough to stand out from the crowd, and the NO GO zone becomes cluster 2 as it is already rife with competition in the area, and with such well-established competitors catering the neighborhood, new establishment here is highly unadvised.

Limitations and Scope

During the implementation of the project, we could not fully consider the other plausible factors which can affect decision pertaining to the opening of a new shopping mall venture on the part of property developers and retailers, namely age cohort, recent data on population earnings. However, to the best of knowledge, concrete and novel data for the same was not available with many sources providing conflicting data. Future research could bolster the work by the inclusive of said factors into the K-Means clustering algorithm to provide a more robust analysis, furthermore, the project was created using the ”Free Sandbox Tier” account of FourSquare API, which limits the number of calls, type and results of calls per day, and thus future research work could be undertaken using paid accounts for more insightful analysis.

Conclusion

During the completion of this project, we implemented the methodologies of data science problem identification, data procurement, data cleansing as well as understanding and implementing machine learning paradigm K-Means Clustering, to analyze interesting pattern in the data, and finally data visualization to depict the findings in a way which is elucidating to the business customer, furthermore, by provision of relevant insight into the Toronto Commercial Marketplace and explanation as to why Cluster 0 should be the prime choice for undertaking new shopping mall venture, what would

be the challenges faced if considering moderately competitive cluster 1 together with suggested recommendation to stand out, and lastly the banes of choosing cluster 2 as a location choice for the mall, this capstone aims to provide valuable input to the meticulous task of decision making when it comes to multi million dollar investments such as shopping malls and complexes.

References

- Toronto Boroughs and Neighborhoods: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- Scribble Map: Toronto, Canada: https://www.scribblemaps.com/maps/view/Neighbourhoods_in_Toronto_Ontario/Unki2xMmX6
- Neighborhood Profiling: https://www.toronto.ca/About-Toronto/community-statistics/neighbourhood-profiles
- Retail Council of Canada, Study: https://www.retailcouncil.org/wp-content/uploads/2018/12/RCC-Canadian-Shopping-Centre-Study-2018_EN_Final-Rev1.pdf