Smart Search: Canadian University Clustering

Adejokun Adedamola Emmanuel

May 31, 2020

adejokunontop@yahoo.co.uk

1. Introduction

Thousands of foreign nationals troop into Canada every year in search of University education. This is due largely to Canada's 'flexible' immigration policy, quality education, comparative lower tuition and consistent high ranking for overall quality of life. Prospective students are presented with a deluge of options from which to make a choice - no mean fit by any standard, as Canada has a plethora of reputable Universities across its provinces and territories. It would certainly be helpful if these young and enthusiastic ones were gifted a tool or platform to help, at the very least, point them in the right direction in their search.

This project offers a lifeline. Canadian Universities are organized into clusters predicated upon pre-selected features (or attributes). Each cluster comprises Universities having a unique combination of features. The clusters are described using the unique combination of features. Prospective students can then identify the cluster into which to focus their search based on the accompanying cluster description. Thus, the search experience is refined, effective and less time consuming.

2. Data

The dataset used for this project was assembled from scratch. The sources of the components of the final dataset are given below:

- **Canadian University List**: Sourced from the Wikipedia webpage List of Universities in Canada using the BeautifulSoup library.
- **Canadian University Ranking**: Sourced from the Wikipedia webpage Canadian University Ranking using the BeautifulSoup library.
- **Provincial Rent Rates**: Sourced from rentals.ca Provincial Rent Rates.
- **Geographical Coordinates of Universities**: The Geopy library was engaged. The University names are passed as arguments and the respective coordinates were returned.
- **Location Data (Recreational Index)**: Data for venues within 500m of the each University is obtained as location data with the Foursquare API. The Recreational Index for each University was then derived from the unique venue categories.

3. Methodology

Data Gathering

As detailed in the preceding section, the dataset used in this project was put together from different sources in order to fulfil the objective. The BeautifulSoup library was used to scrape the Canadian University List and Rankings from the stated webpage. The data for the Provincial Rent Rates was lifted from the referenced webpage. University coordinates were generated with the Geopy library. Lastly, the Recreational Index, which is a measure of the availability of ‘fun’ spots within 500m of the institution, was obtained as location data using the Foursquare API. A series of cleaning, merging and joining operations were involved in building the final dataset.

Feature Set Categorization

Prior the implementation of the machine learning algorithm, the feature set was selected. They included the following columns: ‘Rankings’, ‘Rental Rates’ and ‘Recreational Index’. These features were then converted to categorical variables upon which one-hot encoding was subsequently performed in order to facilitate analysis of algorithm output.

Feature	**Rankings**	**Rental Rates**	**Recreational Index**
Categories	High Rank, Low Rank, Mid Rank, No Rank	Cheap, Affordable, Expensive, Luxury	Exciting, Fun, Sparse

Table 1: Feature Categories

The above categorization was informed by exploratory data analysis performed on the feature set

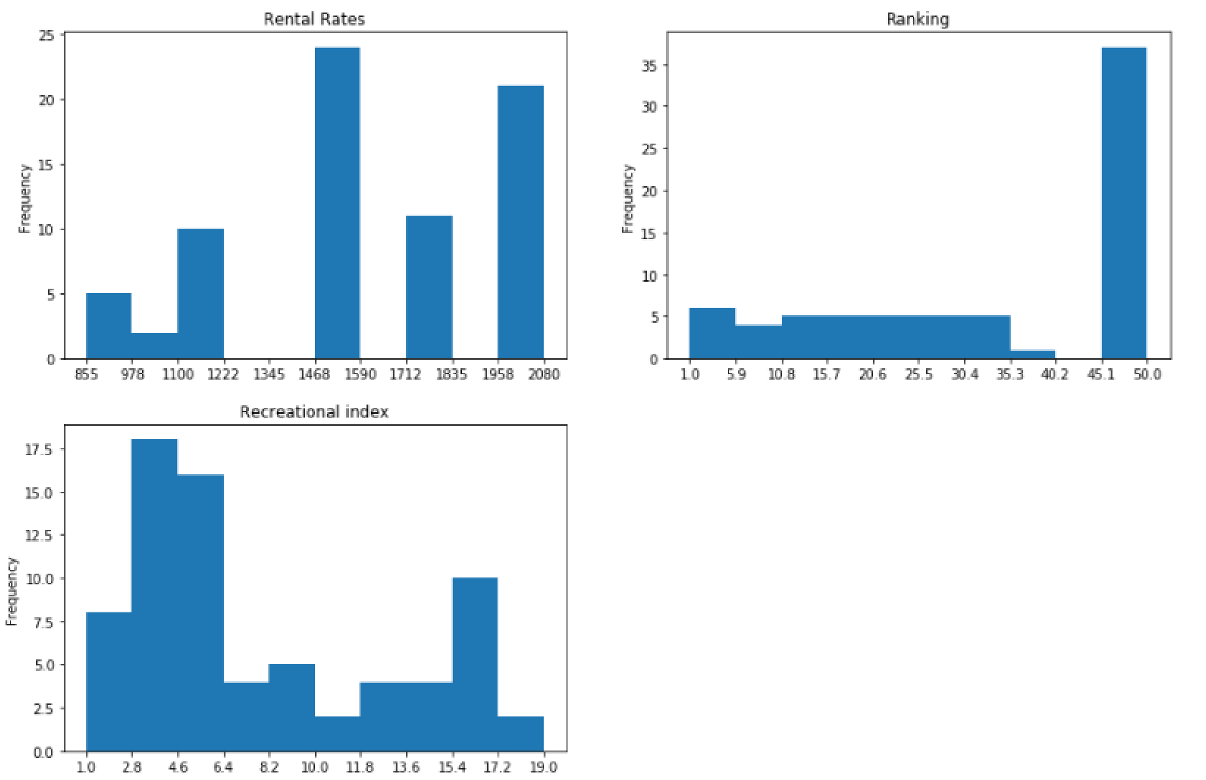


Fig 1: Histogram plot of Feature Set

Machine Learning Algorithm

After the dataset had been assembled, one hot encoding was implemented on the feature set. This ensured the data was transformed into a structure better suited for the implementation of the machine learning algorithm. Next, the k-means clustering algorithm was implemented on the encoded dataset.

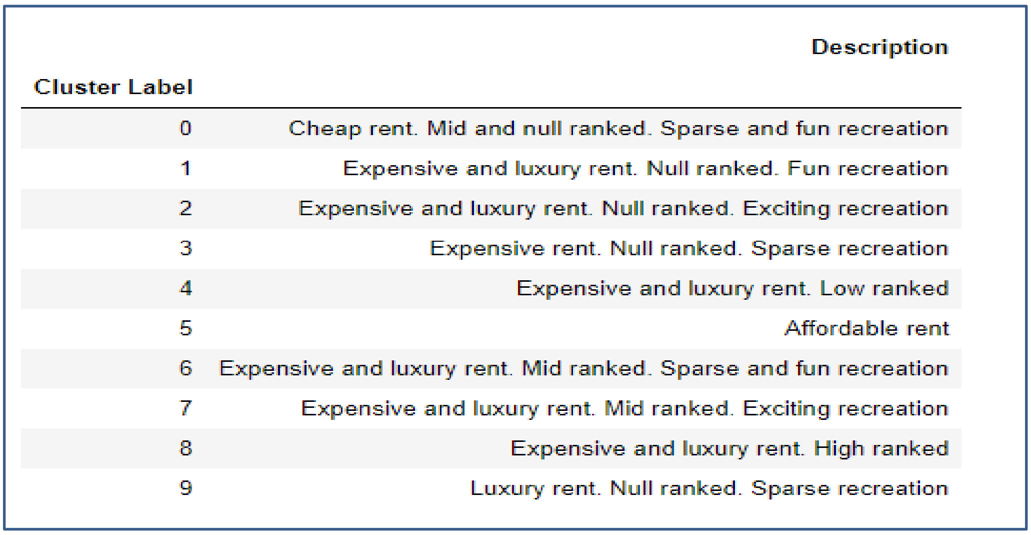
Assumptions/Limitations

Only publicly run Universities are considered in the study. An arbitrary ranking of **50** was assigned to Universities without an official ranking to facilitate numeric data handling. The Provincial rent rates employed are representative of the median monthly rates within the specified Province. Thus, actual rent rates across cities within the Province may differ from those stated.

4. Result and Discussion

The output from the clustering algorithm was examined. Unique descriptions were then assigned to each cluster. These descriptions were predicated upon the combination of features prevalent within each cluster, and served as markers to prospective students as to which cluster to 'smartly select' for further exploration.

The table below shows the clusters and their respective descriptions:



	Description
Cluster Label	
0	Cheap rent. Mid and null ranked. Sparse and fun recreation
1	Expensive and luxury rent. Null ranked. Fun recreation
2	Expensive and luxury rent. Null ranked. Exciting recreation
3	Expensive rent. Null ranked. Sparse recreation
4	Expensive and luxury rent. Low ranked
5	Affordable rent
6	Expensive and luxury rent. Mid ranked. Sparse and fun recreation
7	Expensive and luxury rent. Mid ranked. Exciting recreation
8	Expensive and luxury rent. High ranked
9	Luxury rent. Null ranked. Sparse recreation

Table 2: Unique Clusters and Descriptions

Further examination of the clusters revealed the following:

- Universities in Provinces with cheap rents do not have accompanying 'exciting' recreation (i.e. the recreational index is either 'fun', or 'sparse')
- Universities in Provinces with 'expensive' and 'luxury' rents have the full spectrum for rankings and recreation as options
- None of the low ranking or unranked Universities has accompanying 'exciting' recreation
- Every high ranking University is situated in a Province with either an expensive or luxury rent

The following charts provide more insight into the clusters:

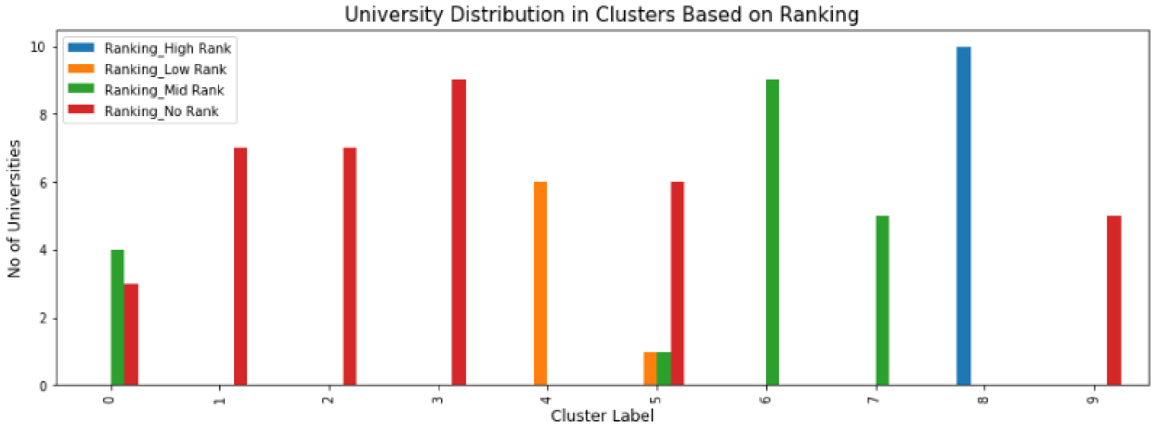


Fig 2: University Distribution in Clusters Based on Ranking Feature

Gleaning insight from the **Fig 2** above, students looking to study in high ranking Universities would directly explore the options in Cluster 8.

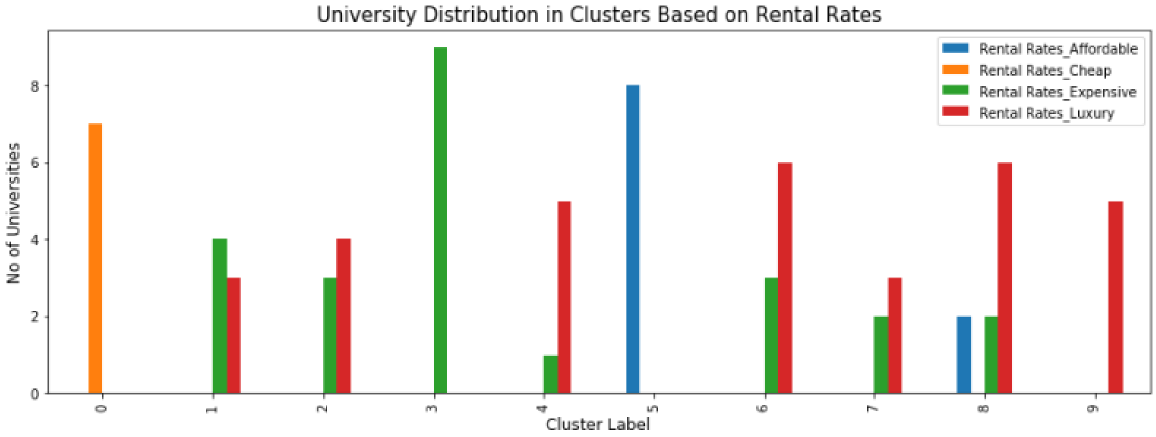


Fig 3: University Distribution in Clusters Based on Rental Rates Feature

Prospective students whose primary consideration is cheap rent would readily explore Universities in Cluster 0, based on **Fig 3** above.

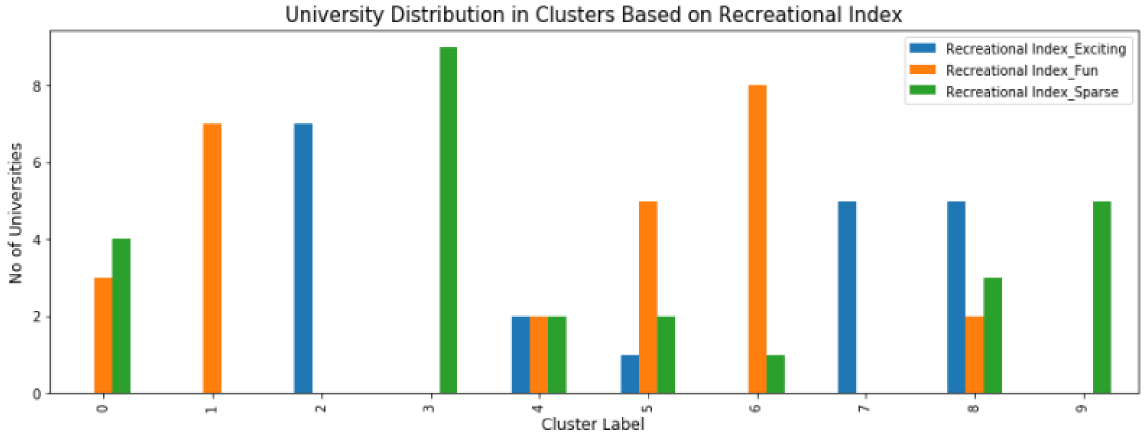


Fig 4: University Distribution in Clusters Based on Recreational Index Feature

The figure above (Fig 4) shows that for prospective students who consider recreation as an integral component of the educational experience, Clusters 2 and 7 offer exclusively 'Exciting' Universities. Clusters 4, 5 and 8 also present further options.

It should however be noted that no single chart would be effective in isolation, as prospective students would usually have at least two features they consider imperative to their University education experience. Consequently, the charts should be used in combination for the best results.

5. Conclusion

Canadian Universities have been clustered with respect to three features (University ranking, Provincial rent rates and Recreational Index). The clusters were obtained using the popular k-means machine learning clustering algorithm, while the location data upon which the novel Recreational Index was derived, was generated using the Foursquare API. The clusters are uniquely characterized, each distinctly comprising Universities having a combination of the defined features. Prospective students can now confidently rely on the clusters to streamline their search for their dream Canadian University.