SMART SEARCH:

CANADIAN UNIVERSITY CLUSTERINIG

INTRODUCTION

- Thousands of foreign nationals troop into Canada every year in search of University education
- Prospective students are presented with a deluge of options from which to make a choice
- Imagine for a second that a provision existed that these young and enthusiastic ones could engage to, at the very least, point them in the right direction in their search
- This project seeks to address this need

INTRODUCTION

- Canadian Universities are organized into clusters predicated upon pre-selected features (or attributes
- Each cluster comprises Universities having a unique combination of features
- A description is given to each cluster consistent with the prevalent feature-combination
- Prospective students can then identify the cluster into which to focus their search based on the accompanying cluster description

DATA GATHERING

- The dataset used in this project was put together from different sources and comprised of: Canadian Universities and their rankings, University coordinates, Provincial Rent Rates and The Recreational Index
- The BeautifulSoup library was used to scrape the Canadian University List and Rankings from List of Universities in Canada and Canadian University Ranking respectively
- The data for the Provincial Rent Rates was obtained from Provincial Rent Rates
- University coordinates were generated with the Geopy library
- 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

FEATURE SET CATEGORIZATION

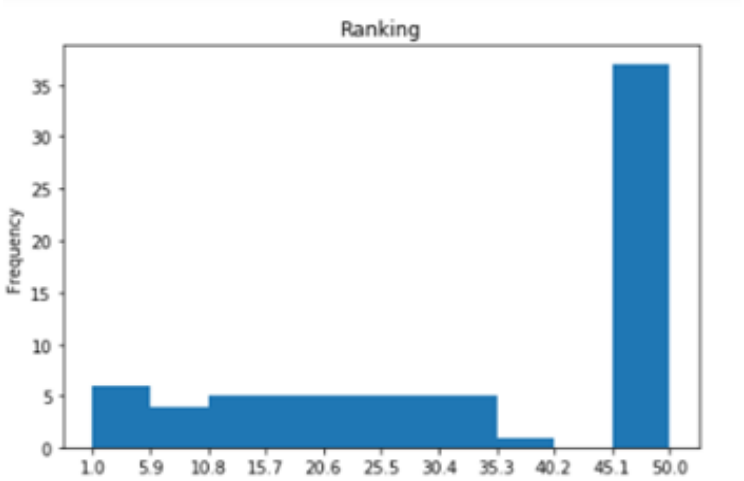
- The feature set was selected to include the following columns in dataset: 'Rankings', 'Rental Rates' and 'Recreational Index'
- These features were then converted to categorical variables upon which one-hot encoding was subsequently performed in order to facilitate analysis of machine algorithm output

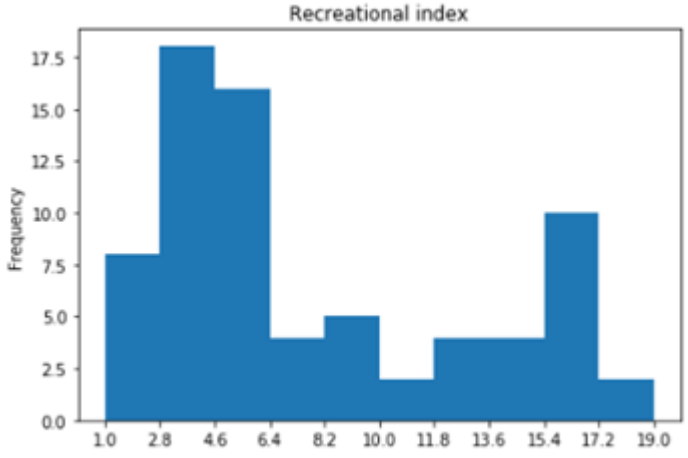
Feature	Rankings	Rental Rates	Recreational Index
Categories	High Rank, Low Rank, Mid Rank, No Rank	Cheap, Affordable, Expensive, Luxury	Exciting, Fun, Sparse

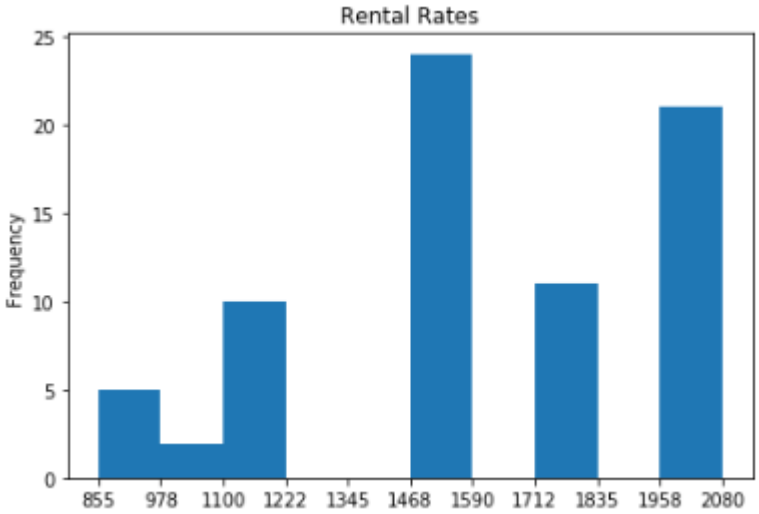
- The above categorization was informed by exploratory data analysis performed on the feature set

FEATURE SET CATEGORIZATION

- Histogram Plot of Feature Set







University Clustering

- After the dataset had been correctly 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

RESULT AND DISCUSSION

- Unique descriptions were then assigned to each cluster
- These descriptions were predicated upon the combination of features prevalent within each cluster
- The descriptions served as markers to prospective students as to which cluster to 'smartly select' for further exploration.

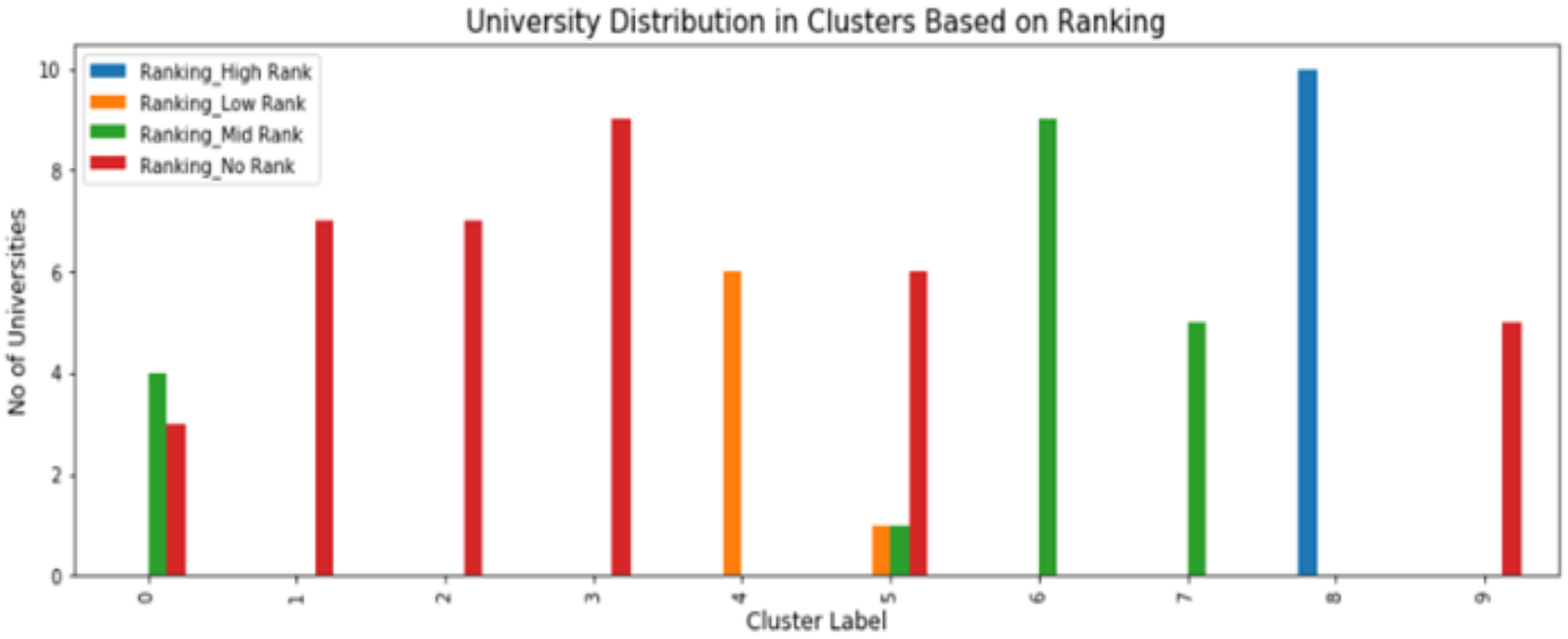
Cluster Label	Description
0	Cheap rent. Mid and null ranked. Sparse and fun recreation
1	Expensive and luxury rent. Null ranked. Fun recreation
2	Expensive and luxury rent. Null ranked. Exciting recreation
3	Expensive rent. Null ranked. Sparse recreation
4	Expensive and luxury rent. Low ranked
5	Affordable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

RESULT AND DISCUSSION

Further examination of the clusters revealed the following:

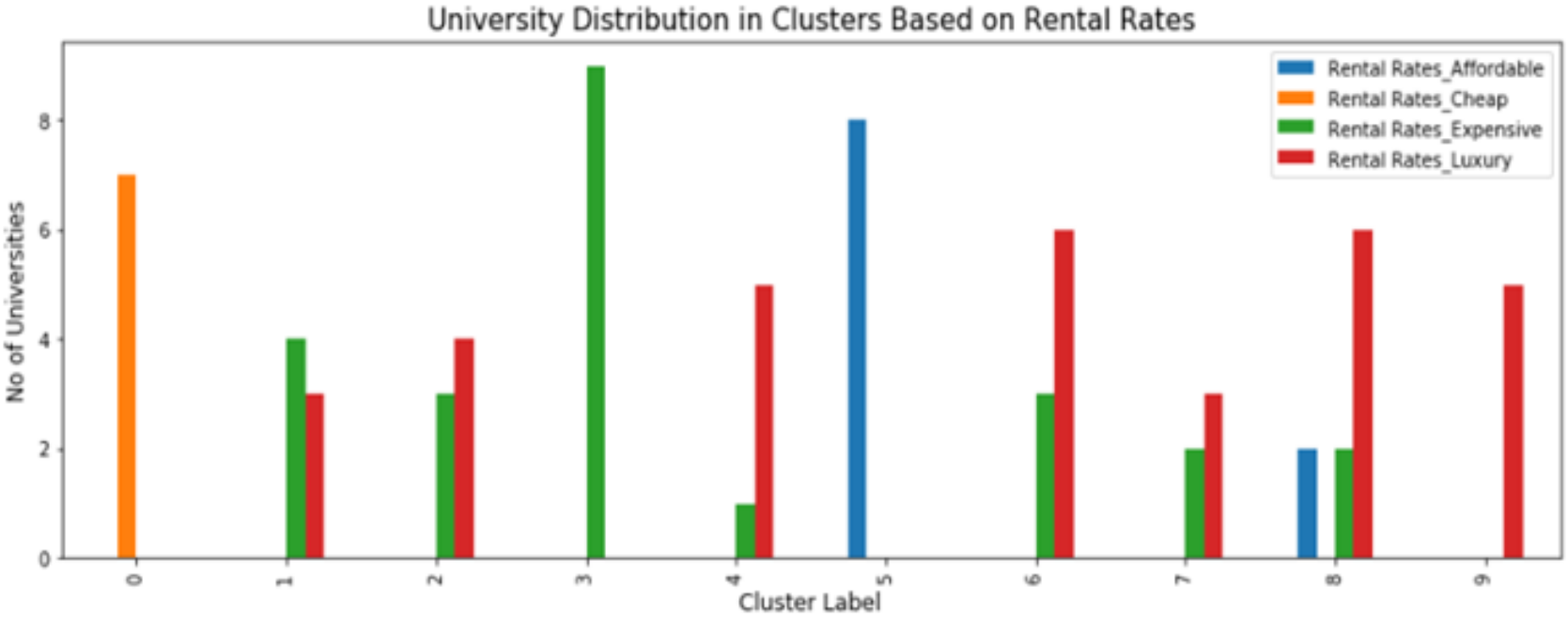
- Universities in Provinces with cheap rents do not have accompanying 'exciting' recreation (i.e. the recreational index is either 'fun', or 'sparse')
- Universities in Provinces with 'expensive' and 'luxury' rents have the full spectrum for rankings and recreation as options
- None of the low ranking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

RESULT AND DISCUSSION



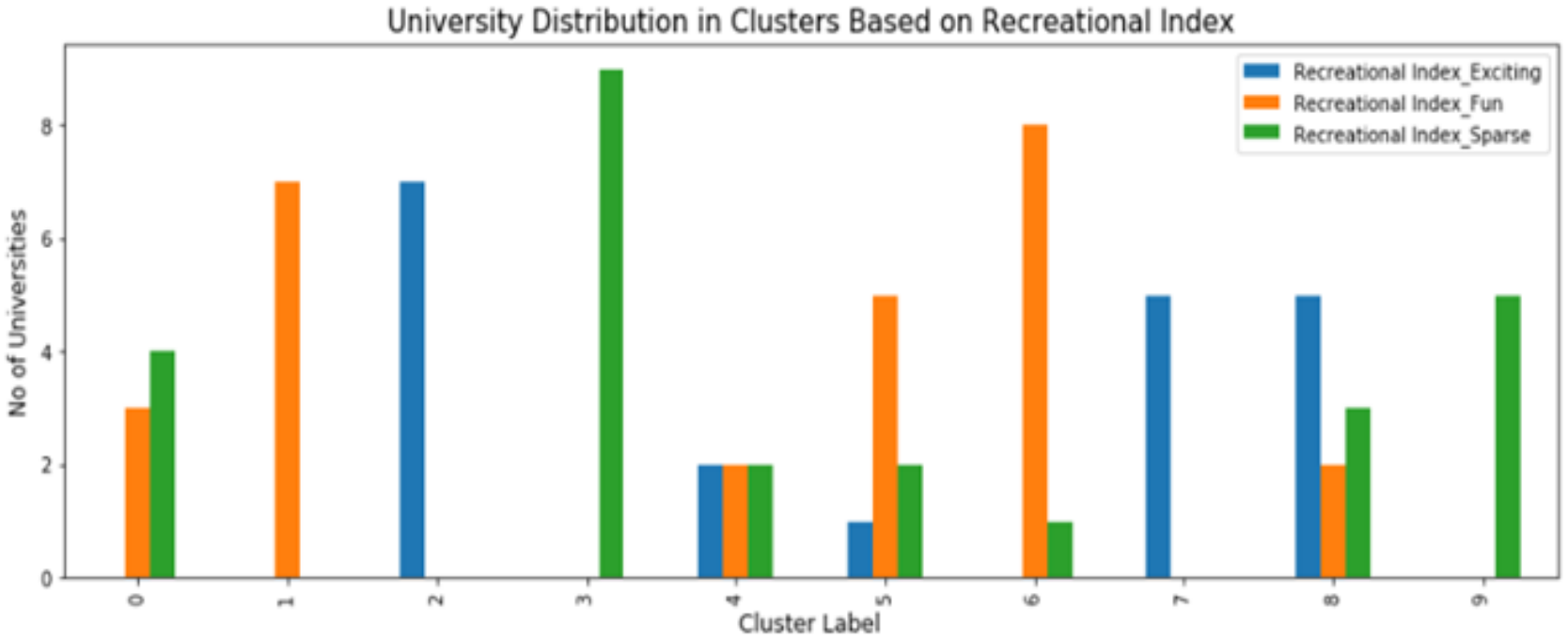
- Gleaning insight from the chart, students looking to study in high ranking Universities would directly explore the options in Cluster 8. These descriptions were predicated upon the combination of features prevalent within each cluster

RESULT AND DISCUSSION



- Prospective students whose primary consideration is cheap rent would readily explore Universities in Cluster 0.

RESULT AND DISCUSSION



- The figure 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

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