Battle of Neighborhood and House

IBM Data Science Capstone Project

Abhishek Padala

1. December 2020

# **1.Introduction**

## Background

There are thousands of students, working professionals, Individuals moving to different countries each year. Understanding of jobs, housing, neighborhoods and venues of it is kind of difficult. I created hypothesis as I wanted to move to Canada and find data science related job. Canada government offering different kind of immigration opportunities, as I understood it is very easy to take a step ahead with immigration process, when a one has a clear idea of province, city and working opportunities. I understand that choosing a province or city based on other people suggestions is a bad idea, so I decided to make a project for a clarity on why I choose a particular province, city, neighborhood or housing to move.

## Problem

* Find the Province & City where the "Data science" related job postings are high.
* Find the companies who have Data science related positions.
* Find the Rental Apartments near to the companies.
* find the best neighbourhoods for amazon company. Neighbourhoods should have following venues in 1.5 to 3km radius.
  + a. Gym
  + b. Coffee shop
  + c. Bar and Restaurant
  + d. Grocery store etc.
  + e. dance school
  + f. yoga centre

## Interest

This project would be very interested in accurate finding perfect province, city, neighbourhood and housing for immigrants(e.g.: students, working professionals) and also helpful for Immigration consultants to help their clients.

# 2. Data Acquisition and cleaning

## 2.1 Data Sources

* This project is designed for leaning purposes, I scraped data from Indeed.ca website for job postings to find out job opportunities in different provinces and cities.
* Rental website called Rentals.ca is popular to find housing, I have scraped data from it.
* I have used Geopy Nominatim to get the rental locations postal codes.
* Scraped <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M> to get the postal codes and neighbourhoods.
* Foursquare API used to get the venues for each neighbourhood.
* Used google maps API to find latitude and longitude of Companies.
* Geopy.distance Geodesic library used to find the distance between companies and neighbourhood, companies and housing locations.

## 2.2 Data Cleaning

Data downloaded or scraped from multiple web sources were combined into one excel sheet. The location of jobs column is split into 2 columns i.e City and Province, There were null values in the jobs data frame, removed null values.

After finding the province and city, removed rest of the data which are related to other provinces and cities. Unique job posting companies are 28 companies which are posted 627 jobs.

Rental locations data scraped with rental price and location name. The address of the housing is not clear, First I have used google maps geocode to find out latitude and longitude of the address by adding city and province name to location. Second I have used Rental locations geocodes to find the full address of the housing/rental locations.

Neighbourhoods added to each rental location, unique neighbourhoods were retrieved from the same dataframe to understand how many neighbourhoods have housings.

## 2.3 Feature Selection

After data cleaning of Job postings data set having 4 features, they are Job Title, Company, Province and City. These 4 features are very important to understand the job market in province and city, Company feature is will reveal which companies offering jobs and how many jobs available.

After cleaning the Rental dataset, there were multiple unwanted features removed and the useful features are Rent, Address, latitude, longitude, neighborhood. Geocodes will be useful to show the locations on map, rent will be helpful to understand the rent price in Particular areas, Address is the identity of all other features.

Venues dataset is created with neighborhood feature and other 100 most common venue features to understand the activities of people and popular venues of that area.

Distance is of the important feature to understand the distance between company and neighborhood, company and housings.

# 3. Exploratory Data Analysis

During the EDA and data investigation of job listing dataset I have found that Ontario province has highest job postings and 2nd highest is Quebec province. This packed bubble solved province selection.

Chart, bubble chart

Description automatically generated

**Packed bubble with job listings - province wise**

From job listing features the side by side bar chart is created which solved the city selection problem. Toronto is city in Ontario province which has highest number of job listings.

Chart

Description automatically generated

**Side by side bar chart with job listing data – city wise**

Map

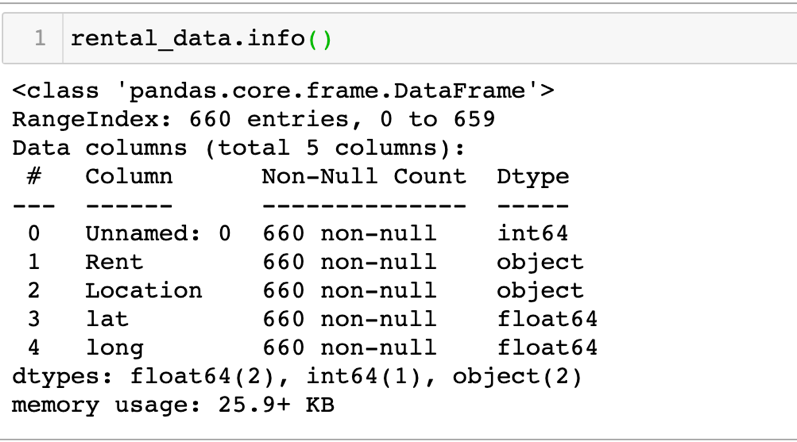
Description automatically generated

**Folium Map showing companies geo locations.**

Based on the above findings of province and city, we are going back to data gathering to get the data of housing, neighborhoods then cleaning the data set.

Graphical user interface, text

Description automatically generated

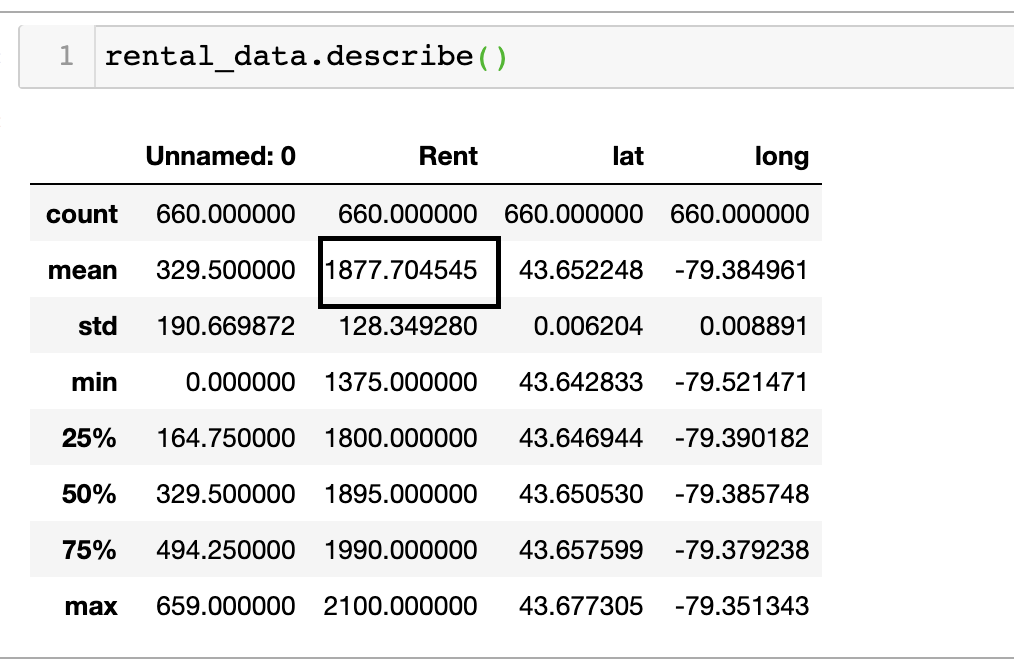


To find out rent average rent the Rent column should be numeric, so I removed dollar sign and comma from the Rent feature and converted the data type to integer to estimate the average housing rent.

Graphical user interface, text, application

Description automatically generated

Average rent mean is 1877.



Map

Description automatically generated

**Neighborhoods geo locations on folium map**

# 4. Methodology

* The Strategy: Finding the Province and City based on the job postings data. I have mapped data into packed bubbles for province selection, side by side bars chart to find the city based on highest job postings.
* After selection of the city, rental locations and job posted companies, Neighborhoods of the selected city will be gathered, wrangled and mined.
* Venues of the neighborhoods will be gathered and clustered. Distance is one of the important features. So, by geolocation the distance will be calculated between company & neighborhood, company & rental apartments.
* K-means Clustering is a process of grouping similar things into one group. Once the clusters revealed, find out the venues which are important, this selection may vary based on activities and hobbies.
* Finally, the available rental locations will be displayed on map based on neighborhoods selection.

# 5. Results and Oberservations

Table

Description automatically generated

Figure: Unique Neighborhoods

A picture containing application, table

Description automatically generated

Figure : Job posting companies and distance to neighborhoods

A picture containing table

Description automatically generated

Figure: Neighborhood venues

Table

Description automatically generated

Figure : Neighborhood Clustering

* Cafe, Restaurants, Coffee shops, food business are very famous in every Neighborhood.
* 9 neighborhoods are with in 1.5km to select company (Amazon).
* Based on cluster 0 Neighborhood’s venues mapped rental locations to find out better transport and happening place to live.
* The final selection of rental location and neighborhood changes person to person.
* I am okay with cluster 0 neighborhoods they are very near to the company and one rental location which is suitable to me, i.e. 1 king West, 1 king street west, Old Toronto, rent is 1800 CAD.

Map

Description automatically generated

Map: Rental locations from selected neighborhoods

Graphical user interface, application, map

Description automatically generated

Map: Finalized rental place to move

Finalized rental Place is :

* Near to metro
* Near to company
* Near to grocery, bank, coffee ships

The venues I wanted are with in 1.5 km like dance school, pool and shopping malls, yoga center