Choosing between Toronto and Montreal: A Clustering Approach

I. Introduction

A. Background and Problem

The decision to choose a place to live and settle can be a daunting experience, either with family or alone. Nowadays, many immigrants are directing their attention to Canada, as the country has been very receptive to them. Immigrants tend to perceive Canadians, and its immigration system, as more welcoming and friendly than the United States.

Canada is an outstanding country with an economy constantly growing and generating many opportunities. Canadian most popular cities are very attractive, truly vibrant, and dynamic. Of course, as an immigrant, choosing a city to live in is much more complex than appreciating the beauty of a city from a touristic perspective.

Toronto and Montreal are the two largest economies in Canada. Before COVID-19, both cities had very low unemployment rates and offered a high quality of life to their residents. Immigrants have a wide variety of factors to consider when deciding where to live. Among those factors, some are specific for each case (e.g. whether the person has a job offer or not, whether French is a barrier or not, etc.).

Nevertheless, this data science project aims to ease the analysis of the factors to consider when choosing between Toronto and Montreal. I will first use descriptive statistics of aggregated and disaggregated economic and demographic data. After better-understanding the economic situation of both cities, we will use a clustering approach to analyze how similar or different are both cities.

B. Audience

The audience of this data science are immigrants that face the decision to move either to Toronto or Montreal. The results of this study are not expected to be definitive. They are only going to provide information to these immigrants so that their decision is easier and can be concentrated on the factors that matter for them.

II. Data

I will use data from the OECD to make a comparison of the two cities and put them in the context of other Canadian cities. Other governmental and private organizations' data will be used to complete this first exploratory analysis.

For the clustering section of the analysis, I will use location and zip code data from geonames.org in addition to Toronto data provided by IBM and Coursera for a previous project. For this section of the project, I will cluster the boroughs of both cities, as separated datasets, according to the type of venues that exist in each city. This will allow me to understand if the clusters from one city are similar to the clusters from another city. Finally, I will cluster the boroughs of both cities as one dataset to see if the algorithm separates the boroughs in clusters isolated geographically because of being very dissimilar. If otherwise, clusters do not present geographic isolation it would mean that both cities are very similar in terms of the venues that they have.