# ProblemSet2_new

Ke-li Chiu & Diego Mamanche Castellanos

06/02/2020

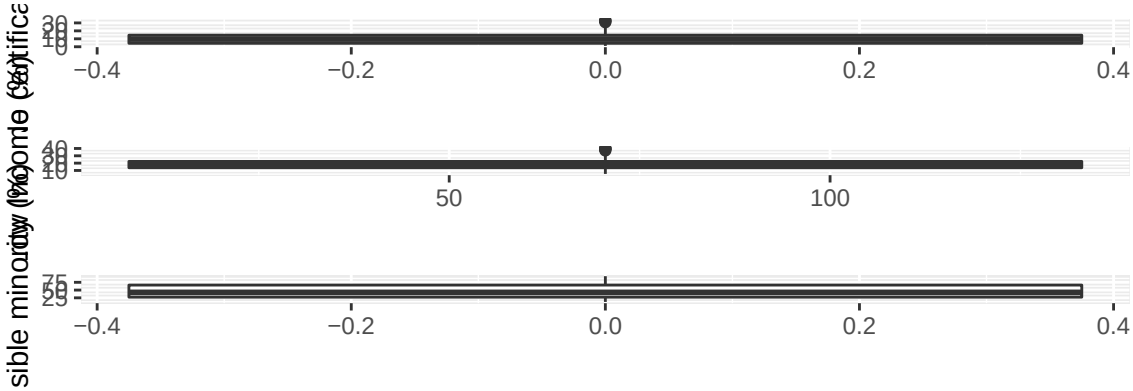
**Abstract**

Abstract nnnnn nnnnnn

# Dataset and Method Description

## Dataset

To address the research question, the city of Toronto, through its portal Open Toronto Data, offers a comprehensive dataset called Neighbourhood Profiles. It contains several categories divided by topics, that in turn, are broken down into different characteristics, all presented in a total of 2383 rows. As for the neighborhoods, all of them are displayed as columns. For this analysis, the variables "no certificate, diploma, or degree 2", "18 to 64 years %", and "total visible minority population" will feed into the model to answer the research question. Descriptive statistics about the variables are shown in the figure 1 and table1.



| | | | |
|---|---|---|---|
| *range* | 32.36 | 35.1 | 82.84 |
| *sum* | 1462.46 | 2587.3 | 6507.23 |
| *median* | 9.67 | 17.75 | 42.52 |
| *mean* | 10.45 | 18.48 | 46.48 |
| *SE.mean* | 0.61 | 0.6 | 1.86 |
| *CI.mean.0.95* | 1.2 | 1.18 | 3.67 |
| *var* | 51.3 | 49.92 | 482.36 |
| *std.dev* | 7.16 | 7.07 | 21.96 |
| *coef.var* | 0.69 | 0.38 | 0.47 |

## Method

Expose the most vulnerable neighborhoods in the city of Toronto is a classification case. The goal is to group all of them by sets based on how similar they are. To achieve this, the classification method to be used will be clustering, an unsupervised machine learning technique that partitions a set of objects with similar characteristics into subsets.

Different approaches were analyzed. Clustering techniques such as K-means, Density-based classification, and Hierarchical clustering were tested out with the dataset, finding the last one the most suitable for this investigation. Hierarchical clustering allows to view at once, each possible number of clusters (k) through the dendrogram, which is a tree representation of those clusters. Moreover, the hierarchical method needs no k in advance. It is important in this study because the purpose is to find the most vulnerable neighborhoods without any bias. Lastly, by using the Elbow curve, the best number of clusters (k) can be impartially calculated.

# Introduction

# Dataset and Method Description

## Limitation of the Dataset and Approach

## Ethical Considerations

# Clustering Toronto's Neighbourhoods by Feautures of Education, Income and Visible Minority Status

## Euclidean Distance Between Toronto's Neighbourhoods

## Determine optimal number of clusters

## Assessing Characteristics of Neighbourhoods

## Clusters Distribution

## Toronto's Neighbourhoods Clusters Map

https://www.datanovia.com/en/lessons/dbscan-density-based-clustering-essentials/