## Borough Suggestion for living in Porto Alegre

## Coursera capstone Daniel Omar Soria Santos – [danielomar.santo@gmail.com](mailto:danielomar.santo@gmail.com) – 07/

## Table of contents

1. Introduction: Analysis Problem
2. Data
   * 1. Data Collection
     2. Data Preparation
3. Methodology
4. Discussion
5. Results
6. Conclusion

## Introduction: Analysis Problem

The objective to this project is to **recommend** boroughs so a user could choose a borough to live in the city Porto Alegre, even without knowing any carecteristic of the neighborhoods.

To better recommend the borough we will use some information previously collected by the user such as previous borough that the user have lived, the job location in the city of Porto Alegre and users preferences for venues type. More information about the data collected will be explained in the Data section.

## Data

## Data Collection

Based on definition of our problem, factors that will influence our recommendation are:

* Distance from the boroughs to the user workplace in Porto Alegre;
* Types of venues in the boroughs;
* Characteritics of the boroughs, such as population and income;

Data sources extracted are:

* Boroughs in the city of Porto Alegre/RG - Brazil [[https://pt.wikipedia.org/wiki/Lista\_de\_bairros\_de\_Porto\_Alegre]](https://pt.wikipedia.org/wiki/Lista_de_bairros_de_Porto_Alegre%5D" \t "_blank);
* IDH for each Borough of Porto Alegre and IDH for the user previous boroughs from [[http://www.atlasbrasil.org.br/2013/pt/consulta/]](http://www.atlasbrasil.org.br/2013/pt/consulta/%5D" \t "_blank)
* Boroughs and workplace latitude and longitude from **Geopy lib**;
* Venues type for each borough from **Foursquare API**;
* Data set from user research with ratings for venues preference;

## Data Preparation

The Data Preparation focused in cleaning data and aggregating all the relevant information from the boroughs into one single dataset. Some of the preparation consisted into removing rows, removing columns with high correlation and aggregating columns into a single one to create a new feature;

Data Preparation for borough\_df imported from **wikipedia**:

* Delete row last row - TOTAL;
* Clear DF NaN;
* Drop irrelevant columns;
* Rename columns;

Data Preparation for location with **Geopy lib**:

* Collect latitude and longitude for each Borough in the DF and for workplaces;
* Calculate distance between boroughs and workplace;
* Remove the boroughs with distance higher than the mean;
* Aggregate data to the df;

A close up of a mans face

Description automatically generated

Data Preparation for venues with **Foursquare API**:

* Collect the boroughs venues for each category rated by user in df\_ratings;
* Collect the venus from the previous borough that the user lived;
* Aggregate data to df;

Data Preparation for IDH with **Atlas Base**:

* Collect the boroughs data from Atlas Base;
* Create the ratio between population aged 19- 35 years and the total population;
* Remove columns with high correlation;
* Aggregate data to df;

A square white text

Description automatically generated

* Plot borough map;

A close up of a map

Description automatically generated

## Methodology

Methodology relies on 2 algorithims, sklearn-recommender library and k-means clustering. The analysis by the sklearn-recommender lib first consisted in applying a df transformation and normalization using the skr similarity transformer. Then finding the top 5 boroughs similar to the previous user borough “Saúde”.

The second step of analysis is based on the k-means method. Another transformation was applied to the df using skr standard scaler. Then k-means model were generated for k=1 up to k=10 and plotted into an Elbow graphic. Using the elbow method the optimal k choosen were k=7. With that we generated the clusters using k=7 and find the cluster with the boroughs aggregated with the previous user borough “Saúde”.

## The intersection of both algorithms are the boroughs recommended to the user to live in the new city Porto Alegre.

A close up of a person

Description automatically generated

## Discussion

Since we have only one borough that the user lived previously, and no data for any other users that are somehow similar to the user we are unable to use methodologies for content-based and collaborative filtering recommender systemns. For that the methodology to recomend a borough to the user used in this project relies on 2 algorithms, the sklearn-recommender library and k-means model. After using the 2 algorithms separately the boroughs recommended to the user is the intersection of both results;

## Results

Using both algorithm we could recommend 4 borough for the user to live.

* Rio Branco;
* Moinhos de Vento;
* Mont'Serrat;
* Auxiliador;

A picture containing text, map

Description automatically generated

## Conclusion

## The sklearn-recommender library resulted in a top 5 similar boroughs and the k-means a cluster with 12 boroughs, the intersection were 4 boroughs that were present in the results of both algorithms. It’s definitely a different approach to the problem but considering the cold start issue and the lack of user data the result its quite satisfactory.