**Urban Diversity Within Seattle’s Neighborhoods**

Natalia Quintero

March 2019

**Contents**

Introduction

Methodology

Data processing cycle

* Obtain
* Scrub
* Exploration
* Model: exploration and clustering
* Results

**Introduction**

What makes a neighborhood? Not only residents, usually venues tend to attract diverse groups of people. A neighborhood with a variety of places may turn to more social interactions but it will also attract people who may change it in the long term. When looking for a new place to live, you are not just buying or renting a home; you are becoming a part of that neighborhood and looking to match your needs at that point in time. There are common components to all neighborhoods, some people may be attracted to areas with lots of restaurants and nightlife while others would like a more residential and quiet area. With this project I want to explore and have a better understanding of the neighborhoods in Seattle and their local amenities, and hopefully help those who are looking to move in the near future find a place that is in sync with their current lifestyle.

**Methodology**

Unsupervised learning involves models that describe data without reference to any labels and partitions it into distinct groups of similar items. This project attempts to categorize and segment data in Seattle’s neighborhoods; identify different venues and what a neighborhood offers for newcomers or those who are looking to live close to specific places.

**Data Processing Cycle**

OSEMN or OSEMiN is an acronym that stands for Obtain, Scrub, Explore, Model and iNterpret. It is a list of tasks that a data scientist should be familiar with, not an expert on each of them. It is also considered as the blueprint for data exploration. This steps don't follow one another, as you go through the analysis is normal to come back to any stage and move back and forth.

**Obtain** Understand requirements, gather information about the problem to be solved, adopt and understand the tools that will be most suitable to do the job. In this case, the *categories* function of the Foursquare API will be used to retrieve the venues and the Seattle Department of Neighborhoods website to obtain a list of the city’s neighborhoods. (<https://www.seattle.gov/neighborhoods/neighborhoods-and-districts>)

Raw data sets generated

* **sea\_neigh\_list.csv**, data set generated from extracting the name of the neighborhoods by zone from the website
* **coordinates.csv**, data set with the geographical coordinates of each neighborhood.

A screenshot of a cell phone

Description automatically generated

From sea\_neigh\_list.csv

**Scrub (clean)** In general, this is where most of the time is spent; data cleaning has an impact on the accuracy of the results. This is where you check for missing or null values, replace or remove them, extract columns or format data types. At this point the dataset presented should be read in a clean manner, free of irrelevant characters and in a usable format. I have removed the zone column

* **coordinates\_shaped.csv**, data file ready to be used without the zone column

A screenshot of a cell phone

Description automatically generatedData set ready to use

**Explore** This is where exploratory data analysis happens, where you get to know the data you are working with. No hypotheses are tested and no predictions are made here. The focus is to understand the distribution of the columns, check for multicollinearity, and make sure the dataset meets what is necessary for the type of model you will apply later on.

A close up of a map

Description automatically generated

Map of Seattle with neighborhoods defined in the data set

256 unique venues found

A screenshot of a cell phone

Description automatically generated

**A screenshot of a cell phone

Description automatically generated**10 most popular venues for the first 5 neighborhoods. From venues\_sorted set

**Model** It refers to the process of using probabilistic methods to try to predict the outcome of an event. Apply a technique(s) or algorithm(s) to predict and interpret, adjust the ones you have to try and increase results.

A close up of a map

Description automatically generated

Model with 5 clusters

**Interpret** Draw conclusions, evaluate the meaning of the results, and use visualization tools to communicate in an understandable manner

Cluster 1 (purple) dominant throughout the city, with venues like parks and coffee shops. Refer to the notebook to find the full list