The Battle of Neighborhoods: Shopping Mall Location Selection in St. Louis IBM Data Science Capstone

Ivan YU
Washington University

May 2020

Content

- Background
- Business Problem
- Description of Data
- Methodology
- Results
- Limitations and Suggestions

Background

- Something about St. Louis
- Why selecting a proper location for new shopping mall is important?
  - Property Developers' prospect
  - Government's prospect

Business Problem

- Where to build a new shopping mall?
- What factors should be taken into account?
- How to add these factors into our models?

Description of Data

- **The list of neighbourhoods in St. Louis**: This defines the scope of this project which is confined to the city of St. Louis.
  - **Data Source**: St. Louis (Wikipedia)
- **Latitude and longitude coordinates of those neighbourhoods**: This is required in order to plot the map and also to get the venue data.
  - **Data Source**: Geocoder package
- **Venue data**: Data related to shopping malls, restaurants, and cafe. We will use this data to perform clustering on the neighbourhoods.
  - **Data Source**: Foursquare API

Methodology

Data Retrieving

List of Neighbourhoods

We will use *requests* and *beautifulsoup* packages to help extract the data from the Wikipedia page.

Latitude and Longitude

We will use *Geocoder* package to help retrieve the information about the latitude and longitude.

Venues Data

Foursquare is a location data provider. Using the RESTful API to retrieve data from Foursquare database is pretty easy. We can just simply create a uniform resource identifier, or URI, and append it with extra parameters depending on the data that we are seeking from the database.

K-Means Clustering

K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for different venues.

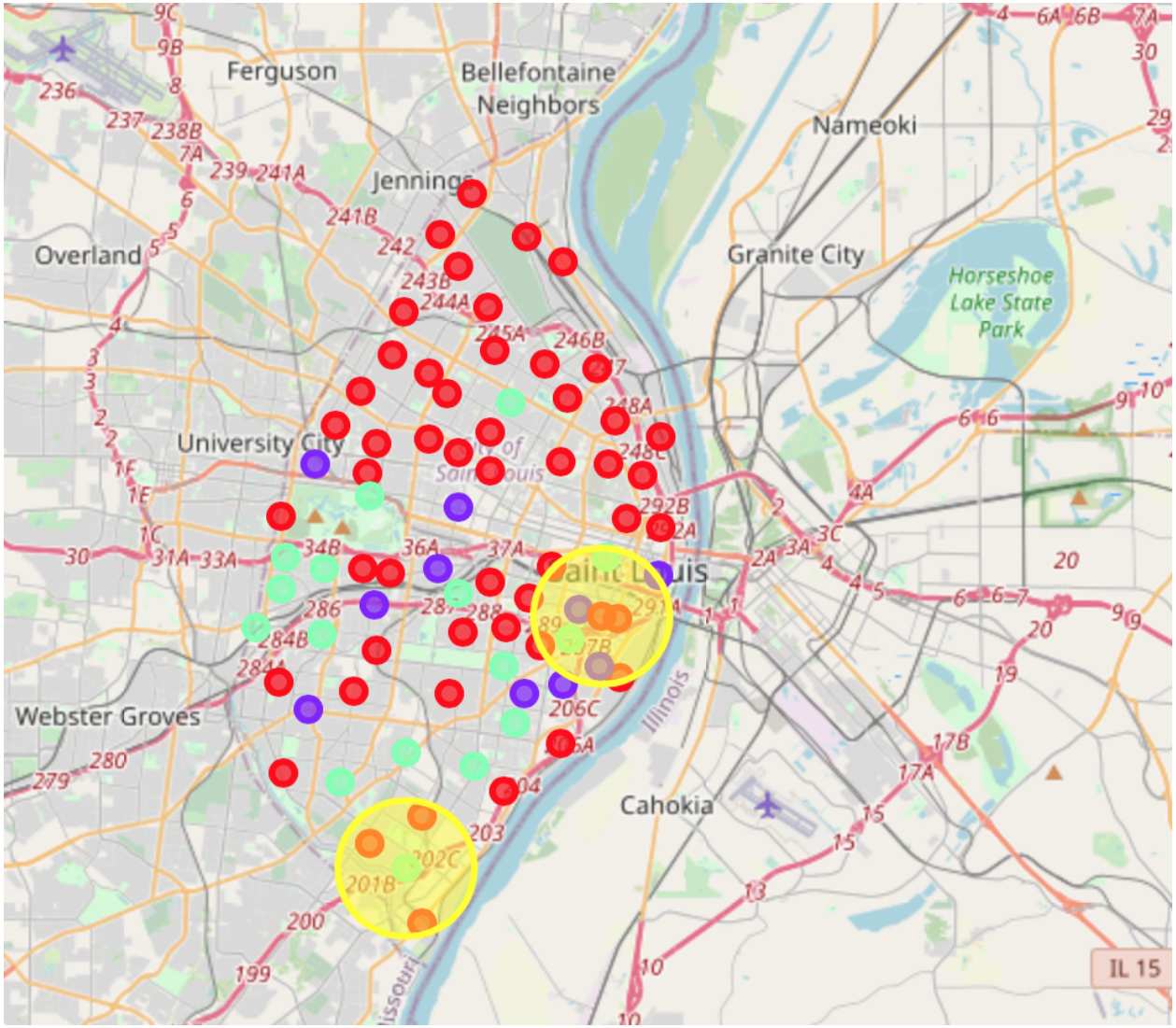
Methodology

Decision-making Philosophy

Clustering Features

We will use the result to identify which neighbourhoods have higher concentration of shopping malls while which neighbourhoods have fewer number of shopping malls. Which neighbourhoods have higher concentration of restaurant and some other entertainment facilities while which neighbourhoods have fewer. With these information, we can answer the question: where is the most suitable location to open a new shopping mall we raised at the beginning.

Results


Ferguson
Bellefontaine Neighbors
Jennings
Nameoki
Granite City
Horseshoe Lake State Park
Overland
University City
Webster Groves
Cahokia
Illinois
IL 15

Results

- **Caterings**: Restaurant, Burger, etc.
- **Entertainments**: Bar, Club, Theater, etc.
- **Cafes**: Cafe, Breakfast, Dessert, etc.

Results

- **Cluster 0**: Least flourishing neighbourhoods (Red)
- **Cluster 1**: Most flourishing neighbourhoods (Purple)
- **Cluster 2**: In the middle of **Cluster 0** and **Cluster 1** (Mint Green)
- **Scope of Influence**: The competition factor of the existing Shopping Malls and Supermarkets (Yellow Circles)

Recommendation

We should choose Purple point outside the Yellow circles as the ideal location to open a new shopping mall.

Limitations and Suggestions

- **Subjectivity Define Parameters**
- **Robust Check**
- **Radius of 'Scope of Influence' Circles**

Thanks for Watching