An Exploration of Washington D.C for

Isabel Laurenceau

Introduction

Large cities are often a popular relocation destination for young adults.

Washington D.C is a not only the capital of the United States but has also been ranked one of the best cities for Black women[1] .

[1] https://www.essence.com/news/best-metro-areas-black-women/ && https://www.bloomberg.com/news/articles/2020-01-09/the-best-and-worst-cities-for-black-women && https://travelnoire.com/best-worst-cities-black-women-relocate

Introduction

But there are over 130 official named neighborhoods[1]

[1] https://opendata.dc.gov/datasets/neighborhood-labels

Solution

fast-track the process of looking at the city by dividing it up based on venues into three main groups: *Residential, Tourist, and Industrial*

Data

OpenData DC

Neighborhood Names

Location Data

https://opendata.dc.gov/datasets/neighborhood-labels

FourSquare

Venue Data

https://foursquare.com

Cleaned Data Labels

Residential	**Residential & Tourist**	**Tourist**	**Industrial**
Grocery	Shopping	Lodging	Business
Health/Wellness	Bar	Tourism	Infrastructure
Home	Restaurant		College
Athletics & Sports	Culture		
	Banking		
	Entertainment		
	Transportation		

Methodology

Comparison of clustering methods- Manual

```
# each neighborhood has an overall label
# for each neighborhood
    #each neighborhood has an overall residential, industrial and tourist count
    #count how many residential, industrial and tourist establishments there are.
    #if an establishment is labeled as “rtboth” it gets a point in both residential and tourist.

#neighborhood is assigned label based on majority of establishments.
# cluster neighborhoods based on labels
```

Comparison of clustering methods- Machine Learning

KMeans

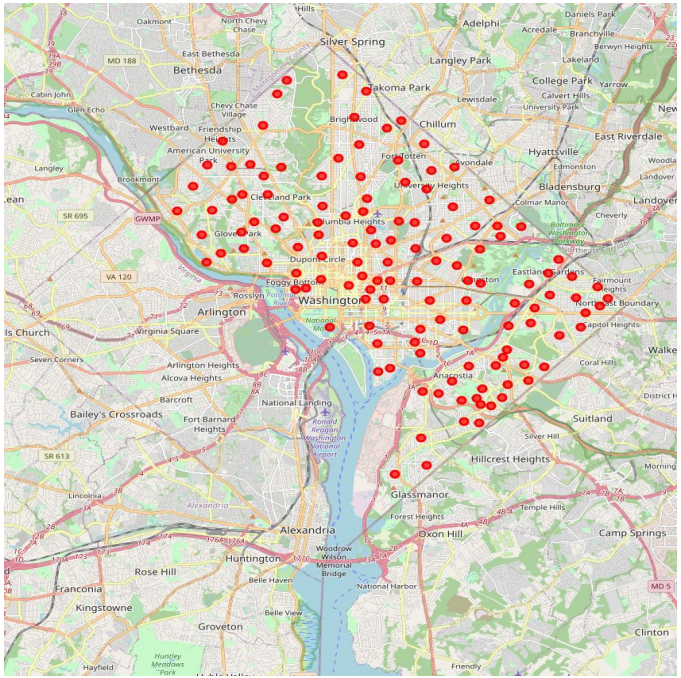
Scikit-learn implementation

Clusters{3-10}

Fit on cleaned and labeled data

Results

Manual

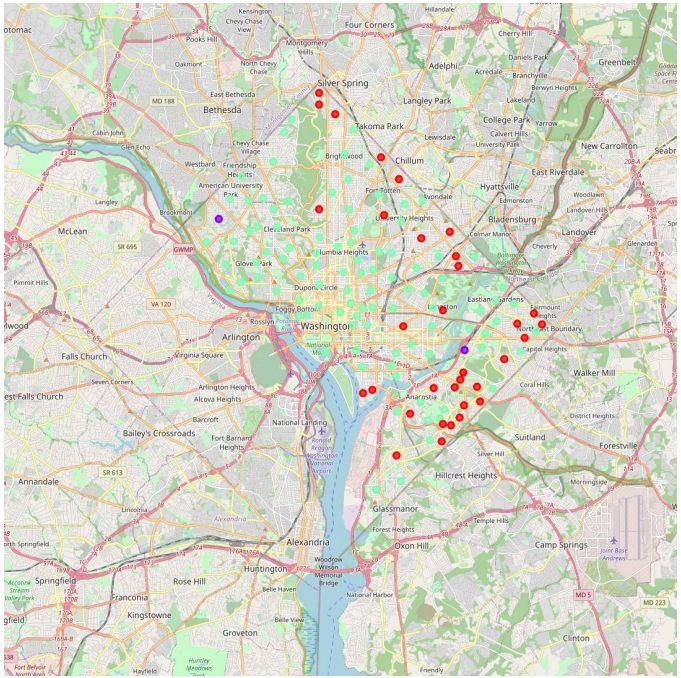


The manual algorithm ended up producing only one cluster for the neighborhoods

Supported by clustering by Kmeans:

Mainly in one cluster until k=8

Kmeans

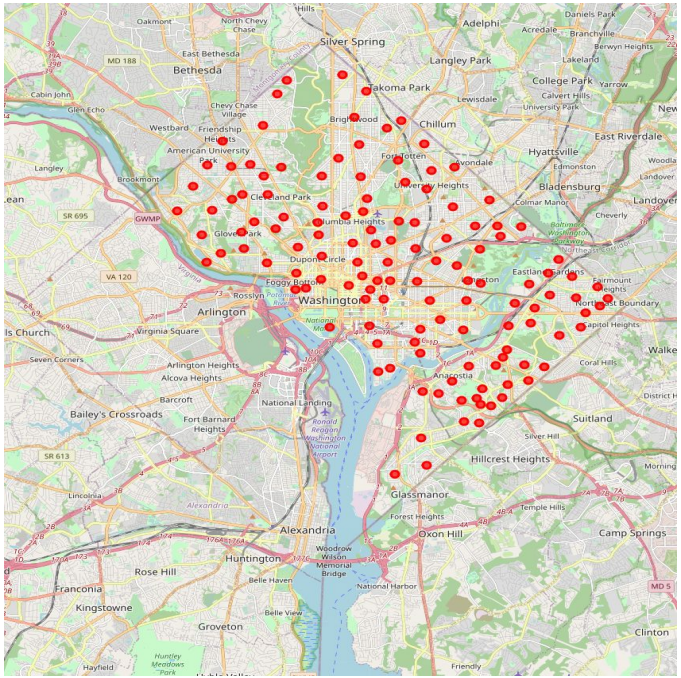


Kmeans always produces correct number of clusters

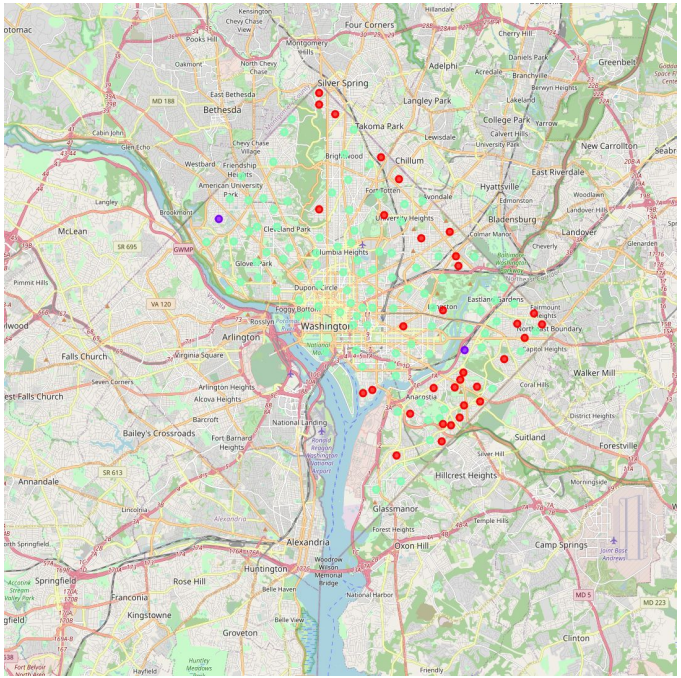
However very unbalanced

Comparison

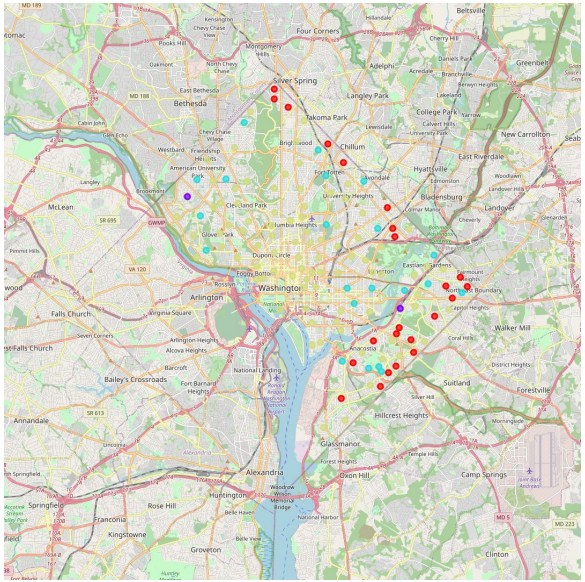
Manual



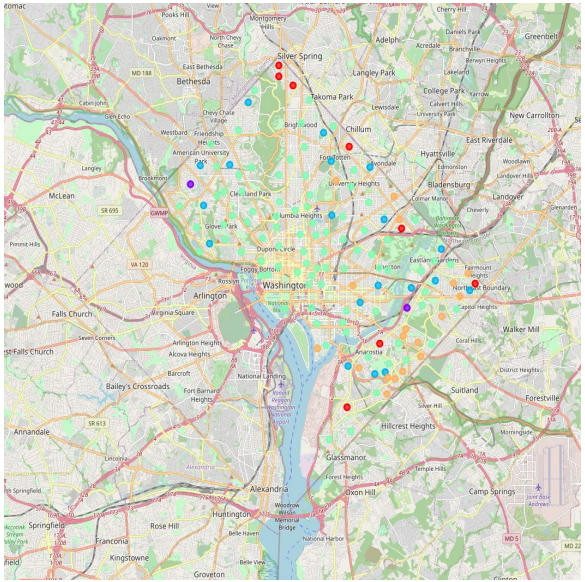
KMeans



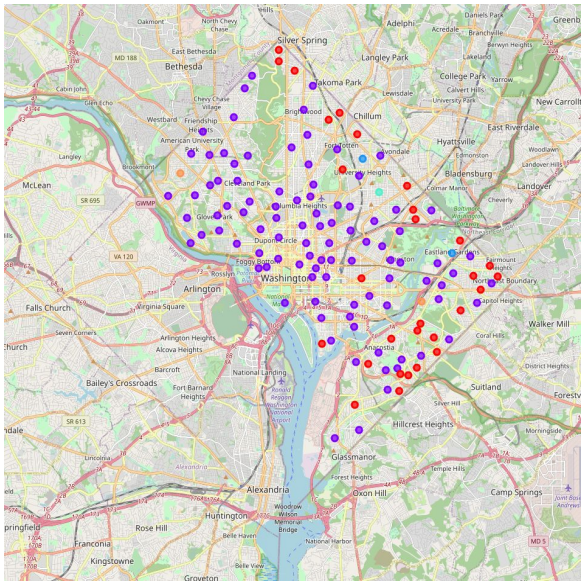
KMeans (k=4-6)



K = 4

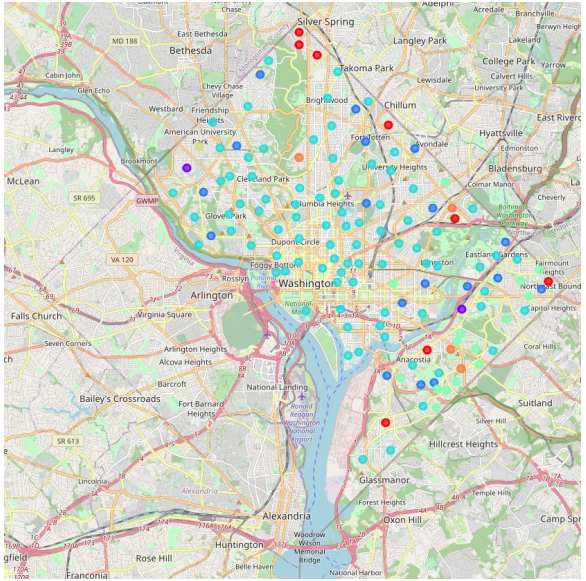


K = 5

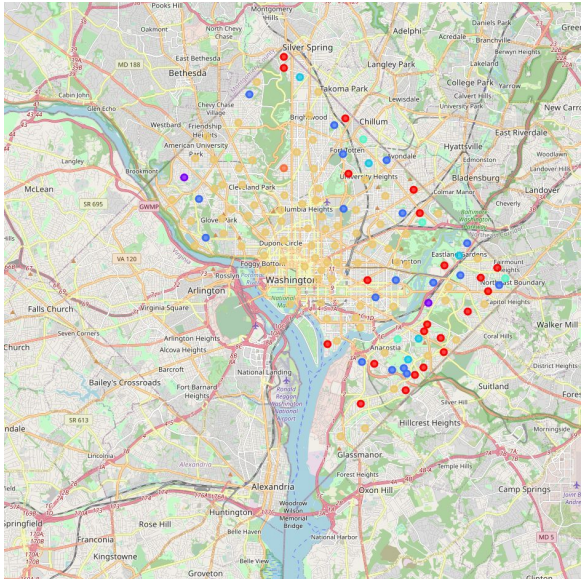


K = 6

KMeans (k=7-8)

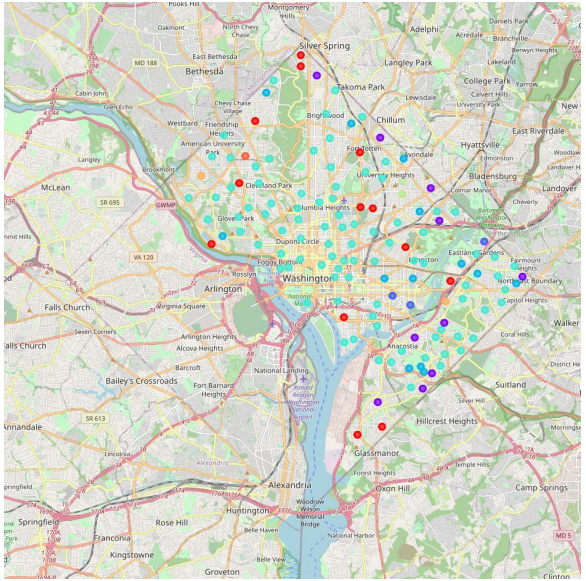


K = 7

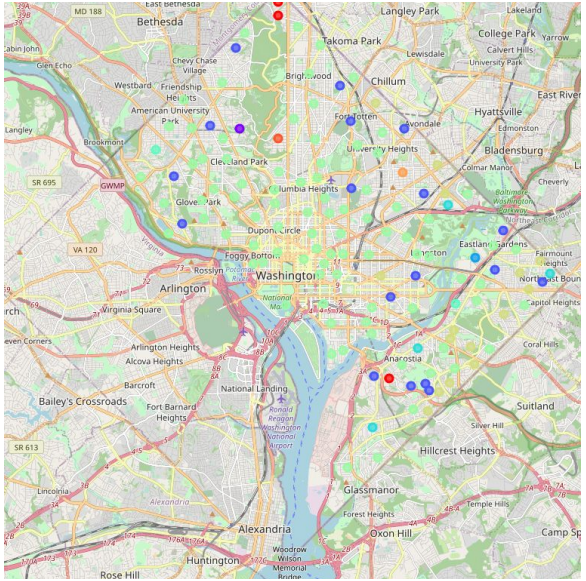


K = 8

KMeans (k=9-10)



K = 7



K = 8

Conclusion

neighborhoods were very ubiquitous in terms of Foursquare data

More data from different sources needed to create proposed system