IBM Data Science Professional- Capstone Project

Seoul Districts

CONTENTS

Introduction

Seoul is South Korea's capital, and largest city. The town sits on the Han River in the northwest region of South Korea. Compared to this, Han River was used as a trading route to China. Han River is no longer used for navigation today, since its estuary is at the North-South Korean border.

Korea has 2,413 kilometers of seaside, with large coastal plains to the west and south. Also, 3,000 remote uninhabited offshore islands. Seoul has twenty five districts, which makes it difficult to decide where to visit and stay.

Business Problem

- As Seoul have 23 districts, it is difficult for travelers, to choose which district to visit. District reviews are subjective and differ from people, you can't just depend on that. It is more important to consider other aspects like price, distance, venues and entertainment, that can highly influence one's experience. There are many of aspects to consider.
- For example, if you are traveling for less than two weeks and have a fixed schedule, it is recommended to book accommodation for the duration of your trip if it gives you peace of mind. Other aspects include the following but are not limited to. The main objective is to find ideal venues where travelers and tourists can find the best suitable to them.

Data

- Neighborhoods

The data of the neighborhoods will be gathered using scraping technique by **BeautifulSoup** . It's a Python based library.

- Geocoding

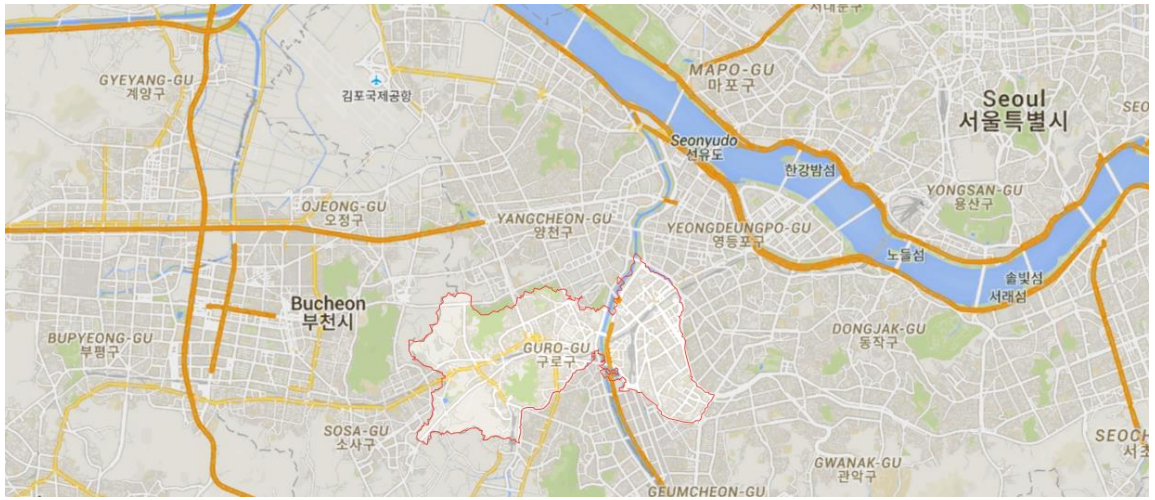
The data generated will in the csv file seoul.csv will be retrieved in a Panda DataFrame. Both latitude and longitude will be established using Google Maps and Geocoding API. Both will be stored into the DataFrame.

Methodology

- Geocoding API

In the underlying advancement stage with Geocoder API, the quantity of wrong outcomes were of a calculable sum, which prompted the improvement of a calculation to break down the exactness of the Geocoding API utilized.

In the calculation created, Geocoding API from different suppliers were tried, and at long last, Google Maps Geocoder API ended up having minimal number of impacts (mistakes) in our examination.



- Venues

Top 10 most regular Venues Due to high assortment in the scenes, just the best 10 basic scenes are chosen and another DataFrame is made, which is utilized to prepare the K-implies Clustering Algorithm.

		1st Common Venue	2nd Common Venue	3rd Common Venue	4th Common Venue	5th Common Venue	6th Common Venue	7th Common Venue	8th Common Venue	9th Common Venue	10th Common Venue
0	0	Korean Restaurant	Coffee Shop	Café	Bakery	BBQ Joint	Chinese Restaurant	Japanese Restaurant	Hotel	Ice Cream Shop	Seafood Restaurant
1	1	Wine Bar	Fish Market	Comic Shop	Concert Hall	Convenience Store	Cosmetics Shop	Department Store	Dessert Shop	Dive Bar	Dog Run

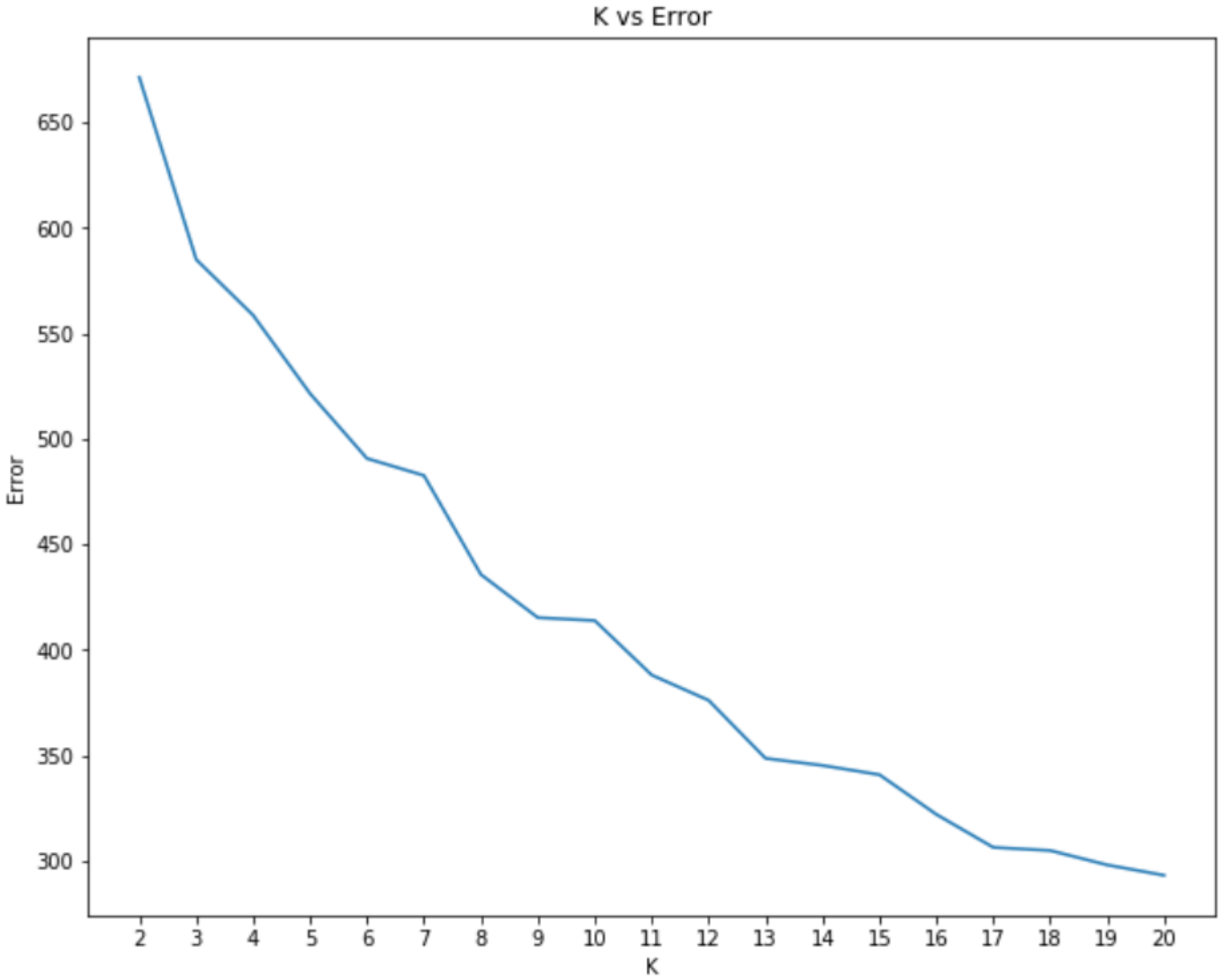
- One hot encoding

One hot encoding is a procedure by which all out factors are changed over into a structure that could be given to ML calculations to make a superior showing in forecast.

- Clusters / K-means

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. The outline ranges from - 1 to +1, where a high worth shows that the item is all around coordinated to its own group and ineffectively coordinated to neighboring bunches.

Figure 3: K vs Error



From the above vizual, we can see that the error reduces starting from K=6

Result

The areas are partitioned into N bunches where n is the number of groups discovered utilizing the ideal methodology. The bunched neighborhoods where most venues are imagined utilizing various hues in order to make them discernable.

Conclusion

The five districts Donong-gu, Dongdaemun-gu, Dongjak-gu, Eunpyeong-gu and Gangbuk-gu fall in the outskirts of Seoul, hence these are the districts with the most Venues.

	Districts	Latitude	Longitude
0	Dobong-gu	37.6688	127.0471
1	Dongdaemun-gu	37.5744	127.0400
2	Dongjak-gu	37.5124	126.9393
3	Eunpyeong-gu	37.6027	126.9291
4	Gangbuk-gu	37.6396	127.0257