IBM Data Science Professional

Coursera Capstone Project

Districts in Seoul

22-04-2020

Table of Contents

1. 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. Given that South Korea is small, with Seoul a mere four-hour train ride from even the most remote area, it's easy to go full holiday mode and explore the country at large. Jejudo, which may just be the most popular destination for holidays. Jejudo is part of 12 Unesco World Heritage Sites roster for the region. These national treasures include royal palaces, graves, shrines and well-preserved villages in the hanok.

Due to the natural isolation of the Korean Peninsula, the country's tradition hasn't modified as a lot as in different regions. Still today, you can stand in awe of some of the hugest cultural heritage websites of the country, such as Gyeongbokgung Palace, the main palace located in Seoul. This makes Seoul one of the most targeted travel location.

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.

2. 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 a lot 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.

- How does price vary for different districts?
- How does transportation affect districts?
- Do you prefer visiting historical or modern venues?
- Which district has more entertainment activities?

- Which districts are more safe?

2.1 Target Audience

- **Travelling Solo**
  The people that are travelling alone, looking for an experience away from their comfort zone.

- **Travelling with family or friends**
  The people you're traveling with or lack of it can make some places and experiences more feasible than others. You have to wonder what kind of place everyone traveling will enjoy.

3. Data

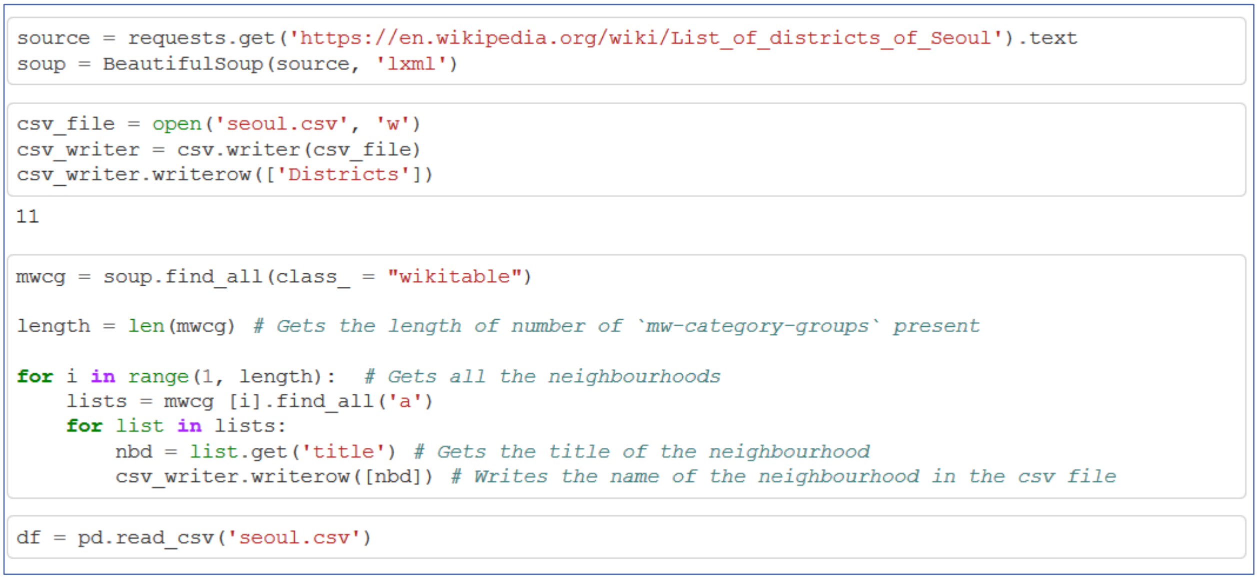
Following are the datasets that are used in the project.

3.1 Seoul Districts Dataset

This dataset available from Wikipedia. This is the core dataset that will be used.

3.2 Neighborhoods

The data of the neighborhoods will be gathered using scraping technique by **BeautifulSoup**. It's a Python based library. Beautiful Soup is a Python bundle for parsing HTML and XML reports. It makes a parse tree for parsed pages that can be utilized to extricate information from HTML, which is helpful for web scratching. It is accessible for Python 2.7 and Python 3.



```
source = requests.get('https://en.wikipedia.org/wiki/List_of_districts_of_Seoul').text
soup = BeautifulSoup(source, 'lxml')
```

```
csv_file = open('seoul.csv', 'w')
csv_writer = csv.writer(csv_file)
csv_writer.writerow(['Districts'])
```

```
11
```

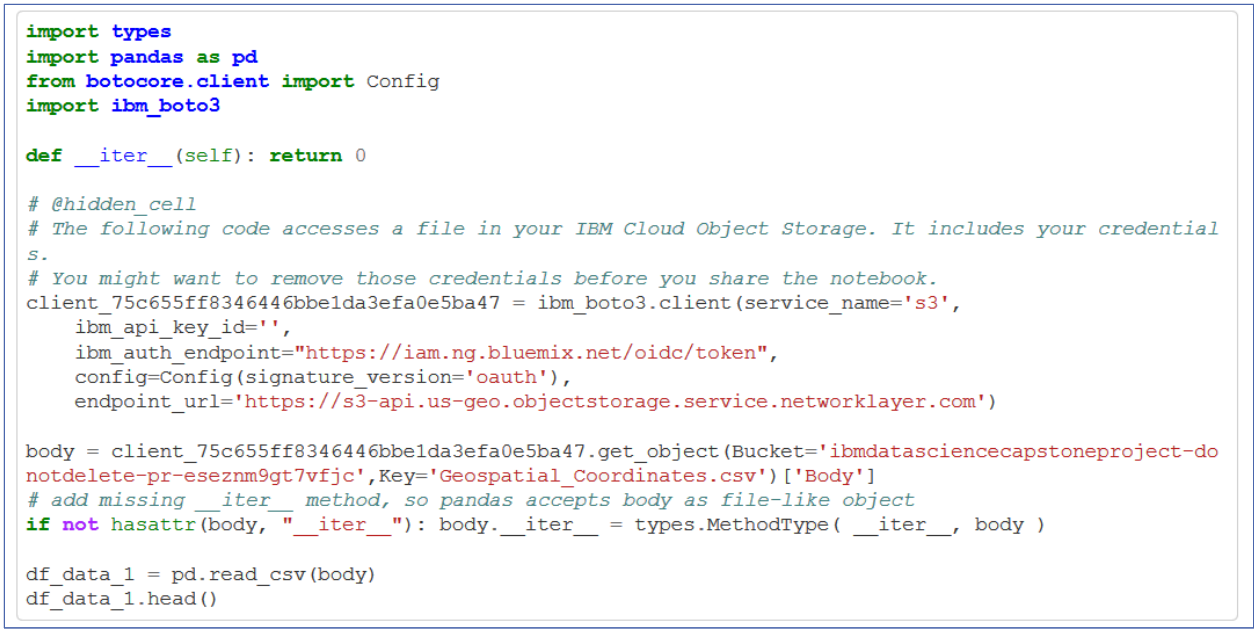
```
mwcg = soup.find_all(class_ = "wikitable")

length = len(mwcg) # Gets the length of number of `mw-category-groups` present

for i in range(1, length):  # Gets all the neighbourhoods
    lists = mwcg [i].find_all('a')
    for list in lists:
        nbd = list.get('title') # Gets the title of the neighbourhood
        csv_writer.writerow([nbd]) # Writes the name of the neighbourhood in the csv file
```

```
df = pd.read_csv('seoul.csv')
```

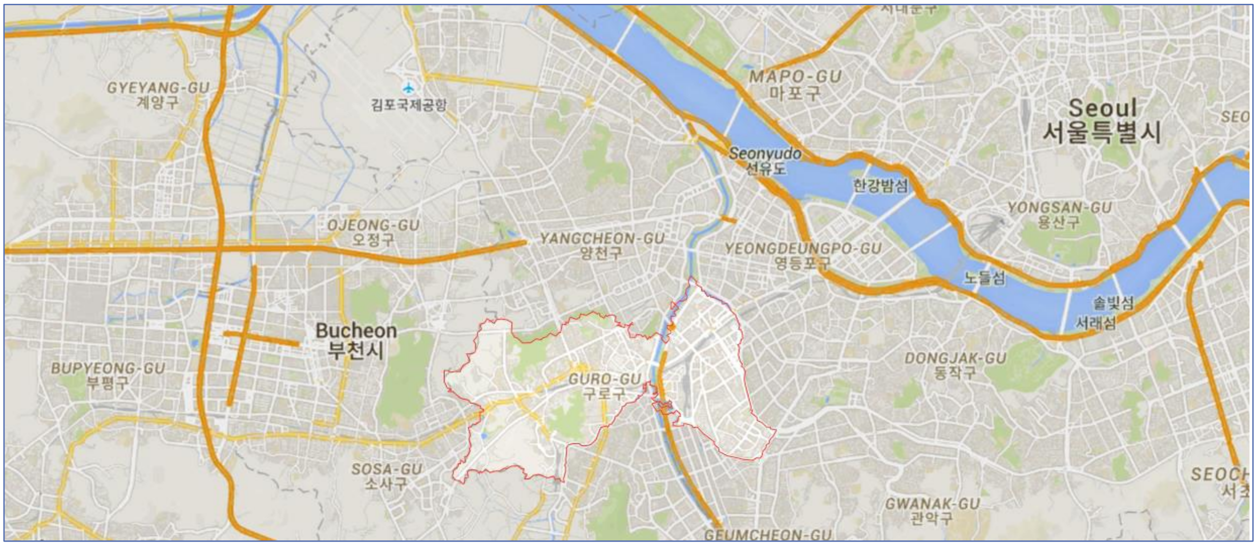
3.3 Geocoding | Geocoding API

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**.

In the underlying advancement stage with Geocoder API, then quantity of wrong outcomes was 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.

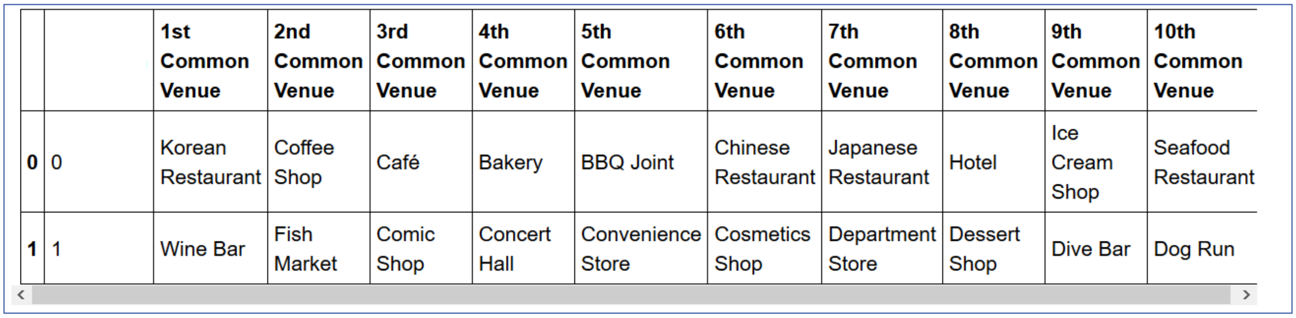


```
import types
import pandas as pd
from botocore.client import Config
import ibm_boto3

def __iter__(self): return 0

# @hidden_cell
# The following code accesses a file in your IBM Cloud Object Storage. It includes your credential
s.
# You might want to remove those credentials before you share the notebook.
client_75c655ff8346446bbe1da3efa0e5ba47 = ibm_boto3.client(service_name='s3',
    ibm_api_key_id='',
    ibm_auth_endpoint="https://iam.ng.bluemix.net/oidc/token",
    config=Config(signature_version='oauth'),
    endpoint_url='https://s3-api.us-geo.objectstorage.service.networklayer.com')

body = client_75c655ff8346446bbe1da3efa0e5ba47.get_object(Bucket='ibmdatasciencecapstoneproject-do
notdelete-pr-eseznm9gt7vfjc',Key='Geospatial_Coordinates.csv')['Body']
# add missing __iter__ method, so pandas accepts body as file-like object
if not hasattr(body, "__iter__"): body.__iter__ = types.MethodType( __iter__, body )

df_data_1 = pd.read_csv(body)
df_data_1.head()
```



3.4 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

3.5 One hot encoding

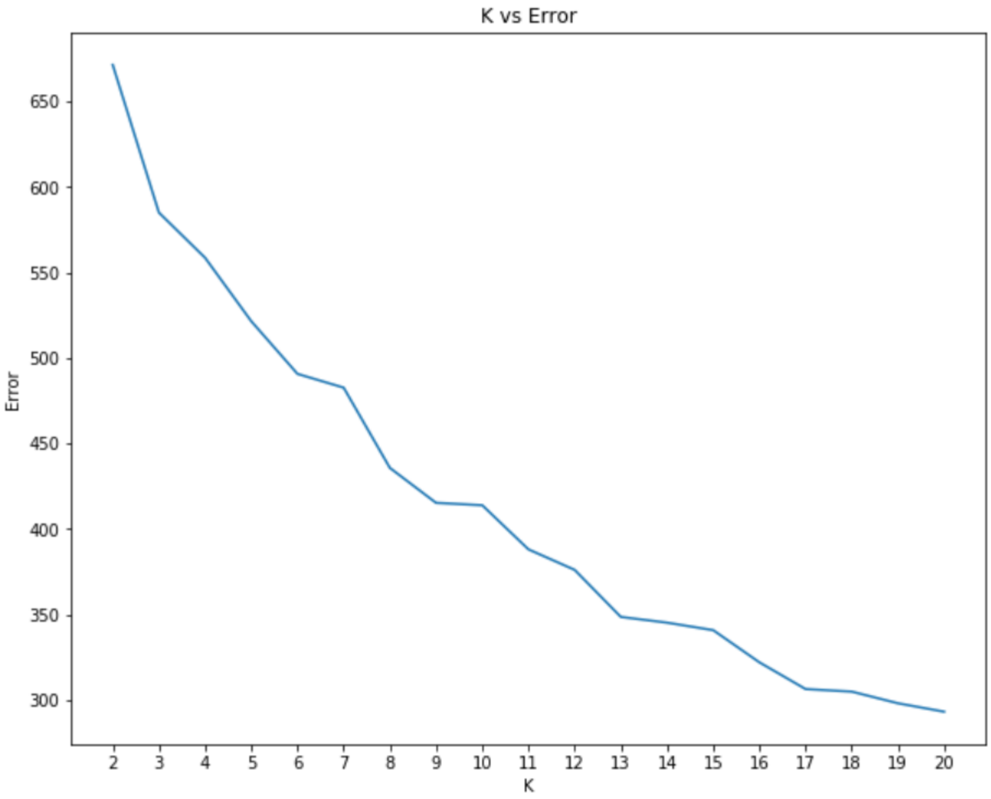
One hot encoding is a procedure by which all out factors are changed over into a structure that could be given to ML calculations to make a superior showing in forecast.

In spite of the fact that mark encoding is straight however it has the hindrance that the numeric qualities can be confounded by calculations as having a type of chain of importance/request in them. This requesting issue is tended to in another basic elective methodology called 'One-Hot Encoding'.

3.6 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.

In contrast to administered getting the hang of, bunching is viewed as a solo learning technique since we don't have the ground truth to think about the yield of the grouping calculation to the genuine names to assess its exhibition. We just need to attempt to research the structure of the information by gathering the information focuses into unmistakable subgroups.



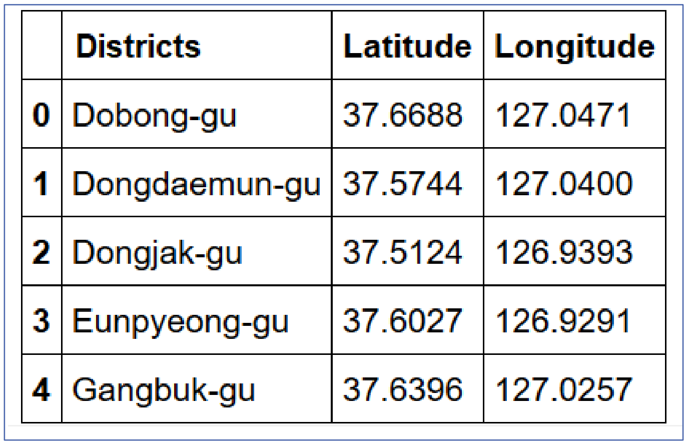
From the above vizual, we can see that the error reduces starting from K=6

4 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.

5 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