외국인 관광객 분석을 통한 관광산업 활성화 방안

20212095 김철현

Summary of questions and results

1. 코로나 전과 후의 방한 외국인 수는 어떻게 달라졌는가?
코로나로 인한 출입국에 제한이 생긴 2020년 2월부터 방한 외국인의 수는 급감했고 2023년 3월경 부터 증가하기 시작했다.

2. 한국을 방문하는 외국인은 무엇을 목적으로 방문하는가?
전체 방한 비율의 50%이상이 여가, 휴가 등을 위해 방문하였고 사업, 친구 혹은 친지 방문 등이 뒤를 이었다.

3. 외국인 입장에서 한국여행의 만족도는 어떠한가?
전반적으로 높은 만족도를 보여주지만 [관광지], [언어소통], [여행경비], [관광안내], [길 찾기] 부문의 만족도는 그다지 높지 않은 것을 확인 할 수 있었다.

4. 여행을 목적으로 한국을 방문하는 외국인들을 위해 관광안내소 설치는 어디에 하는게 좋을까?
대부분의 방한 외국인이 찾는 지역이 서울로 확인이 되었고 서울에서도 관광지가 몰려있는 곳에 안내소를 설치하는 것이 합리적이라고 판단했고 추려낸 결과 서울 강남구, 중구, 송파구가 선별되었다. 그리고 이 안에서도 적절한 위치에 관광 안내소 설치를 제안한다.

Motivation

위 그래프는 코로나 이전과 코로나 기간 중, 그리고 코로나 이후의 방한외국인수를 나타내는 그래프이다. 코로나 이후 다시 방한 관광객이 증가하면서 한국의 대한 만족도가 궁금해졌고 만족도가 낮은 부분이 있다면 해당 부분을 해결하기 위해 앞으로 어떻게 대처해야 할지 제시해보려 한다. 프로젝트는 서울의 외국인을 위한 관광안내소의 부족이 만족도 하락으로 이어진다는 가설을 세우고 최적의 장소에 관광안내소 설치를 제안한다.

Dataset

mdis > 교육·훈련/문화·여가 > 외래관광객조사 > 총괄(제공) > 2021

https://mdis.kostat.go.kr/mypage/extract/viewMyExtractListNew.do?curMenuNo=UI_POR_P9030#data_down

- 2021_총괄.csv
- 2019_총괄.csv

2021년 이후 데이터가 없어서 코로나 이전 방한 외국인의 만족도 조사, 한국 방문 목적, 그리고 주로 어느 도시를 방문했는지 알아보기 위해 사용한 데이터이다.
칼럼의 수가 많고 너무 방대한 데이터라 preprocess.ipynb에서 전처리를 진행했고 각각

- 2019만족도데이터.csv
- 권역방문데이터.csv

로 나누어 데이터 분석을 진행했다.

한국관광데이터랩 - 국가별 분석 > 방한여행 현황 > 방한여행 종합현황

https://datalab.visitkorea.or.kr/datalab/portal/nat/getForTourForm.do#

- 방한외국인수.csv

연월별로 한국을 방문한 외국인 수의 추이를 알아보기 위해 사용한 데이터이다.

한국관광데이터랩 - 지역별 분석 > 지역별 현황 > 지역별 관광 현황

https://datalab.visitkorea.or.kr/datalab/portal/loc/getAreaDataForm.do

- 서울전체.csv

	순위	관광지명	주소	분류	외지인 검색 수
0	1	코엑스	서울 강남구 영동대로 513-0	전시시설	570042
1	2	인터컨티넨탈호텔서울코엑스	서울 강남구 봉은사로 524-0	호텔	104867
2	3	임피리얼팰리스서울	서울 강남구 언주로 640-0	호텔	84505
3	4	SETEC	서울 강남구 남부순환로 3104-0	전시시설	82059

서울 전체에서 주요 관광지가 어느 구역에 몰려있는지 확인하기 위한 데이터이다.
구역을 특정하고 해당 구역의 관광지를 중심으로 다시 데이터 분석을 진행한다.

- 강남구.csv
- 중구.csv
- 송파구.csv

서울에서도 관광지가 몰려있다고 판단되는 구역이다. 이 구역 중에서도 관광지를 추려내서 데이터 분석을 진행한다.

Method

1. 서울 각 구역의 관광지 위치 정보를 경도, 위도로 변환한다.
2. 이 경도, 위도를 사용하여 클러스터링을 진행한다.
3. 클러스터링을 하기 위한 최적의 k를 찾기 위해 Elbow Method를 이용한다.
4. 사용하고자 하는 클러스터 범위를 지정한다. (1 ~ 10)
5. 각 클러스터를 WCSS방법으로 계산한다.

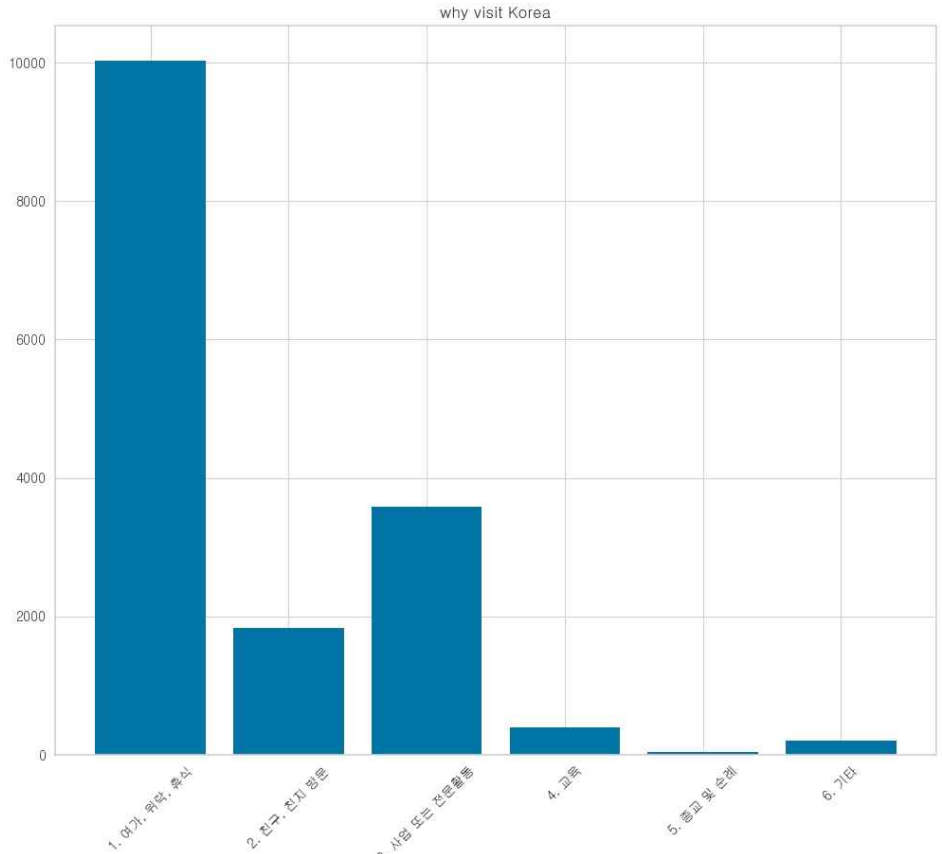
$$WCSS = \sum Pi \text{ in Cluster1 distance } (Pi\ C1)2 + \sum Pi \text{ in Cluster2distance } (Pi\ C2)2 + \sum Pi \text{ in CLuster3 distance } (Pi\ C3)2$$

WCSS는 각 데이터 포인트와 해당 클러스터의 중심 간 거리의 제곱 합을 의미한다.

6. WCSS값과 클러스터 K 갯수에 대한 커브선을 그린다. 이때 x축은 k 값, y축은 WCSS 값이다.
7. 그래프에서 "팔꿈치(elbow)"와 같은 지점을 찾는다. 이 지점은 WCSS 값이 급격히 감소하다가 완만해지는 지점으로, 최적의 k 값으로 간주한다.

Results

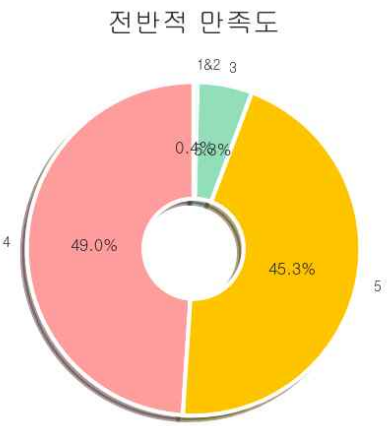
먼저 방한 외국인이 한국에 방문한 이유에 대해 분석해봤다.



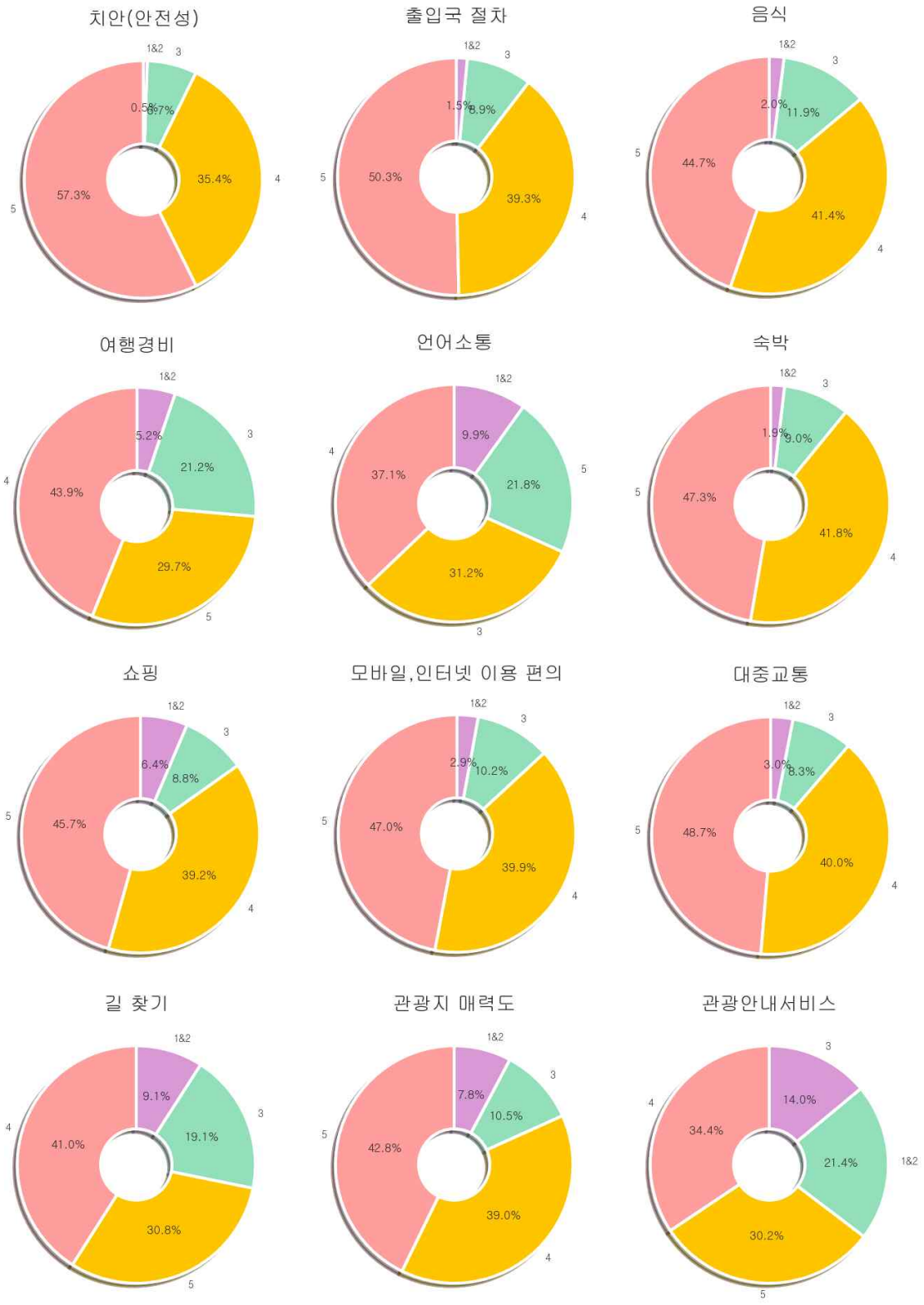
시각화 결과 대부분의 한국 방문의 목적이 여가, 휴식 즉 여행을 목적으로 방문하는 것을 확인할 수 있었다.

다음으로 한국을 방문하는 외국인의 만족도에 대해 시각화를 진행하였다.
그래프 해석은 다음과 같이 진행하면 된다.

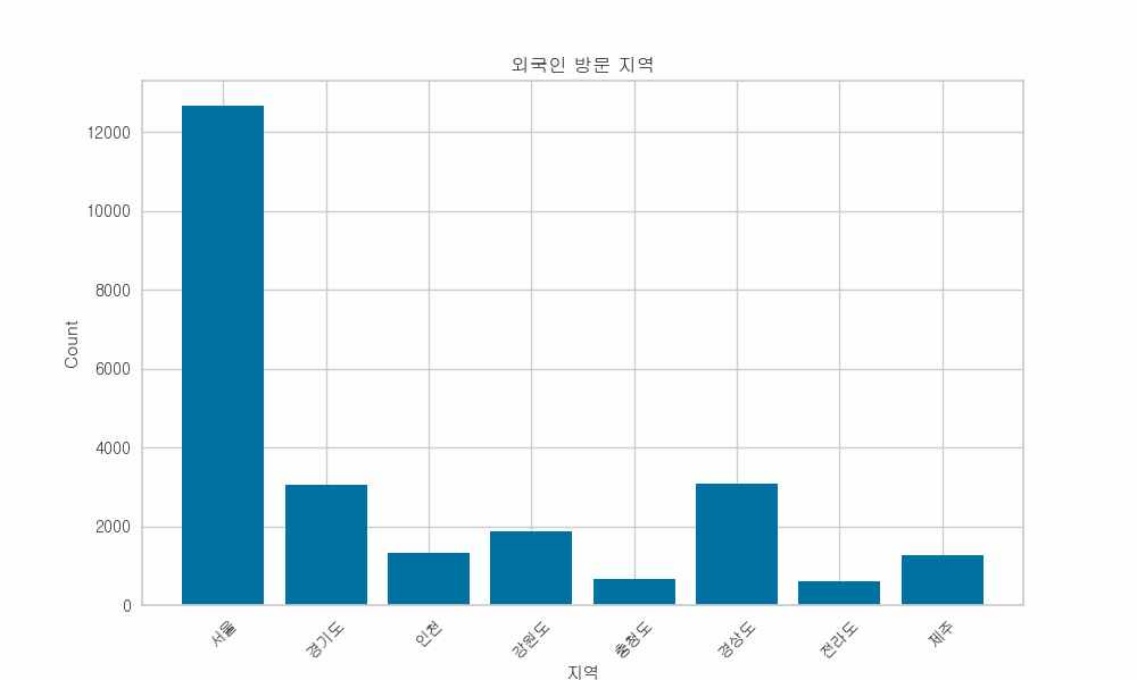
5 : 매우 만족
4 : 만족
3 : 보통
1&2 : 매우 불만족 & 불만족



전반적 만족도 조사의 경우 만족 혹은 매우 만족의 비율이 높은 것을 확인할 수 있었다.
그럼 세부적인 부문의 만족도는 어떨까? 아래 시각화 데이터를 확인해보자



여기서 주목해야 할 건 관광안내서비스의 만족도 그래프이다. 불만족, 매우 불만족의 비율이 21.4%로 가장 높고 그 뒤를 언어소통, 길 찾기 만족도가 잇는다. 이 분석을 통해 관광안내소의 증설과 관광안내소에서 다양한 언어의 책자와 주위 관광명소로 이동할 수 있는 대중교통이용방법 등을 안내할 것을 제안한다.

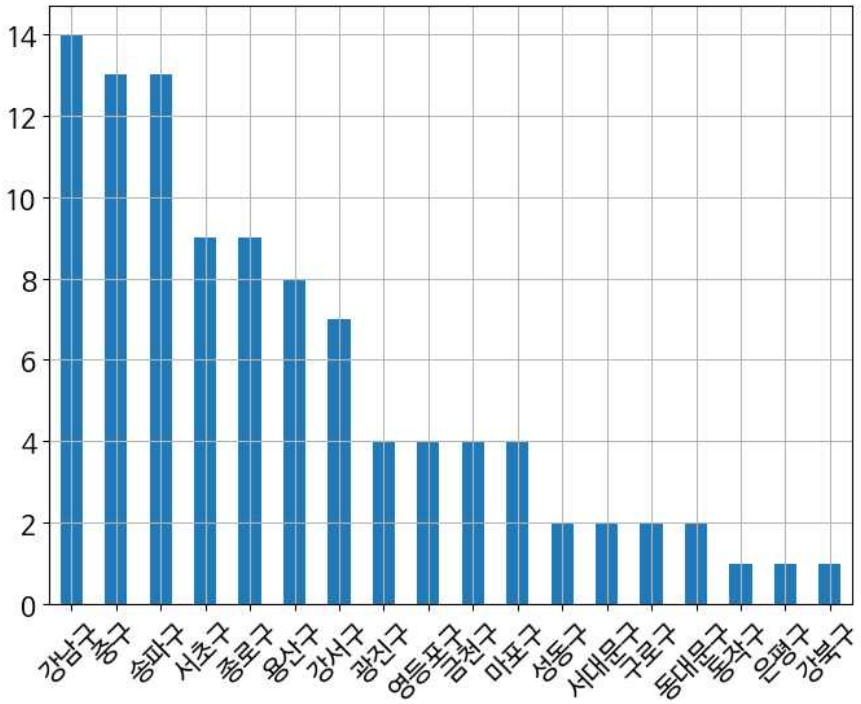


이 그래프에서는 외국인이 한국의 어느 지역을 많이 방문하는지 확인할 수 있다. 많은 비율의 외국인이 서울을 방문하는 것을 알 수 있다. 그렇기에 서울을 중심으로 관광안내소의 최적의 입지를 알아보려 한다. 그러기 위해 먼저 서울 내의 관광지의 분포를 알아보기 위해 아래 데이터를 분석한다.

	순위	관광지명	주소	분류	외지인 검색 수
0	1	코엑스	서울 강남구 영동대로 513-0	전시시설	570042
1	2	김포국제공항	서울 강서구 하늘길 112-0	교통시설	557109
2	3	김포국제공항국내선	서울 강서구 하늘길 111-0	교통시설	509869
3	4	타임스퀘어	서울 영등포구 영중로 15-0	쇼핑몰	393294
4	5	IFC몰	서울 영등포구 국제금융로 10-0	쇼핑몰	327154
...	...	...	...	...	...
95	96	서대문형무소역사관	서울 서대문구 통일로 251-0	전시시설	51755
96	97	CGV왕십리	서울 성동구 왕십리광장로 17-0	공연시설	51491
97	98	가든파이브툴	서울 송파구 충민로 10-0	쇼핑몰	51108
98	99	광화문	서울 종로구 사직로 161-0	역사유적지	50368
99	100	라마다호텔서울	서울 강남구 봉은사로 410-0	호텔	50153

먼저 서울 내에서도 관광지가 어느 “구”에 속하는지 알아보기 위해 주소 칼럼에서 ㅇㅇ구 만을 검출해내서 새로운 데이터셋으로 만든다.

```
allSeoul['소속구'] = allSeoul['주소'].str.split(" ").str[1]
allSeoul
```



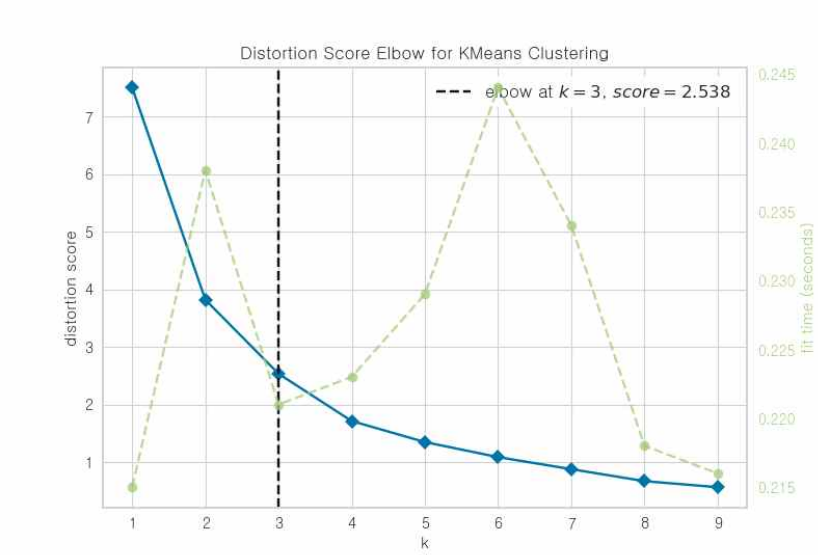
우선 서울 전체에서 주요 관광지 100개를 받아와서 주요 관광지의 소속구를 추려 내었다. 그 결과 **강남구->중구->송파구** 순으로 주요 관광지가 몰려 있는 것을 알 수 있었다.

이제 이 3개의 구역에서 관광지의 분포를 확인하고 위도와 경도를 각각 x,y 좌표로 이용하여 클러스터링을 진행한다.

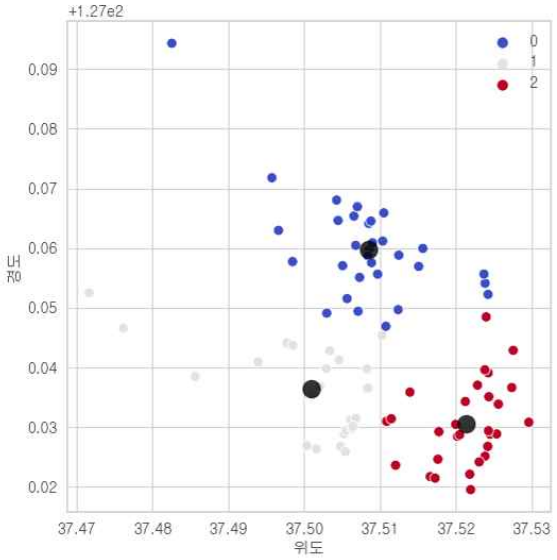
```
26      gmaps = googlemaps.Client(key=gmaps_key)
27
28      lat = []
29      lng = []
30      for name in allSeoul['주소']:
31          try:
32              tmpMap = gmaps.geocode(name, language='ko')
33              tmpLoc = tmpMap[0].get('geometry')
34              lat.append(tmpLoc['location']['lat'])
35              lng.append(tmpLoc['location']['lng'])
36          except:
37              print("{} Address Error".format(name))
38              lat.append(None)
39              lng.append(None)
40      allSeoul['위도'] = lat
41      allSeoul['경도'] = lng
```

그 전에 먼저 구글 Geocoding API 를 이용하여 각 지역의 관광지 주소를 경도, 위도로 변환하여 새로운 데이터 프레임으로 저장한다.

1. 강남구

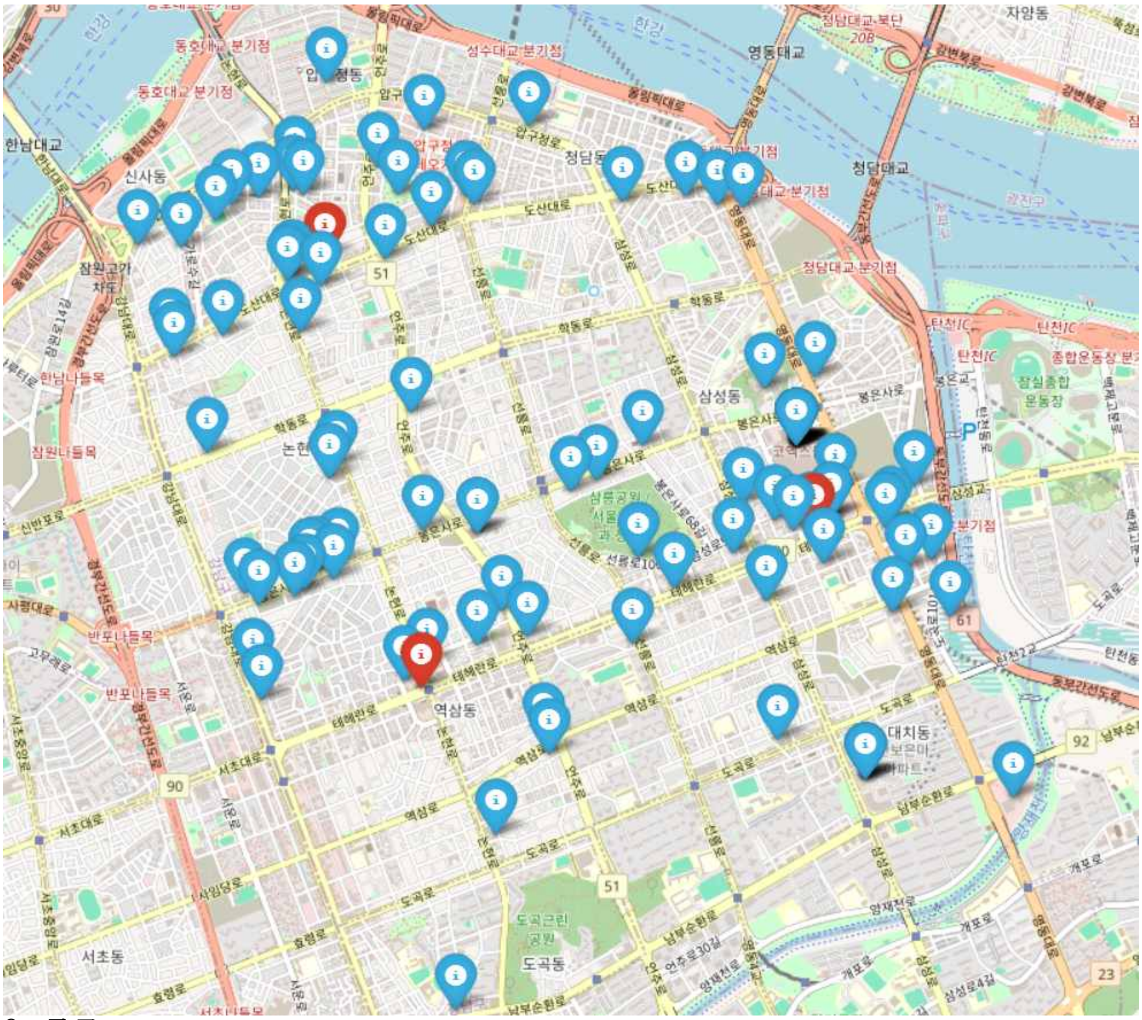


Elbow Method를 확인해보았다. 2~3 지점에서 그래프가 확 꺽이고 최적의 k값이 3이라는 것을 알 수 있다.



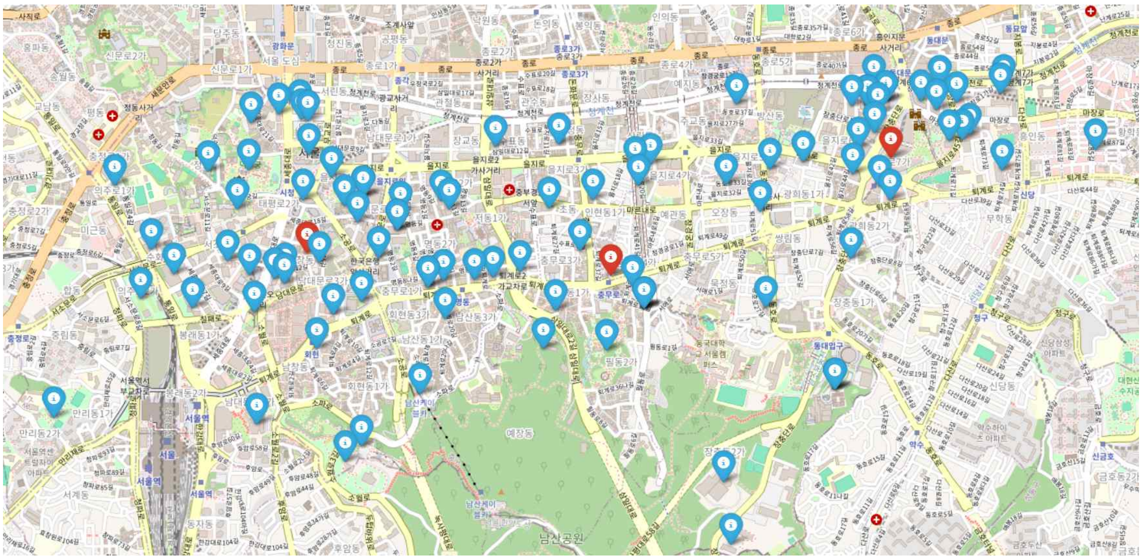
클러스터링 한 결과는 다음과 같다. 이제 각 군집의 중심에 있는 경도와 위도 값을 따로 저장하고 이를 구글맵 위에 표시해보기로 한다.

파란 핀은 강남구에 위치하는 인기 명소를 나타내고 그 개수는 100개이다. 빨간 핀은 3개의 군집 각각의 중심 위치를 의미한다.
이 빨간 핀을 중심으로 관광안내소 설치를 제안한다.
중구와 송파구는 결과 이미지만을 보인다.

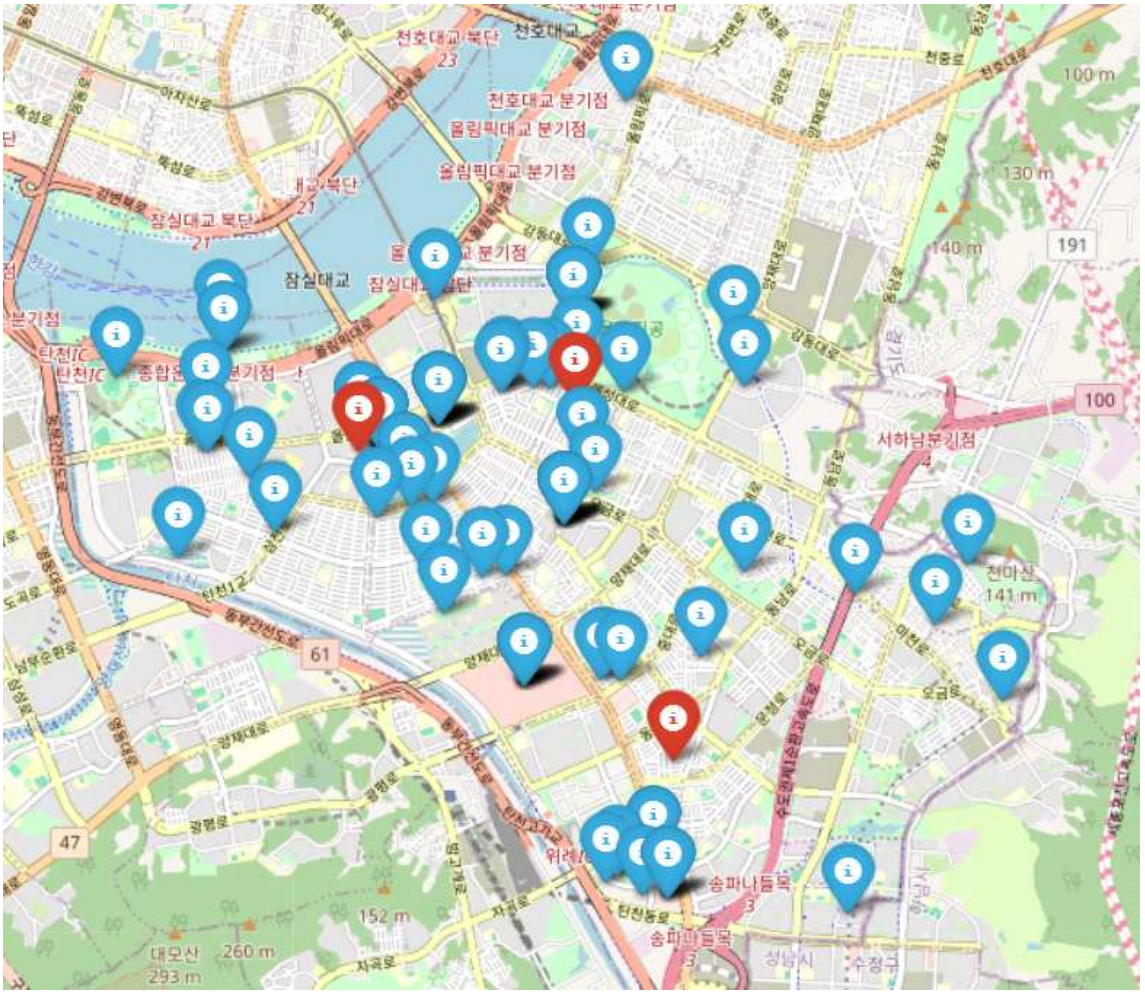


동호대교 분기점
성수대교 분기점
청담대교 북단
영동대교
자양동
한남대교
신사동
압구정동
청담동
도산대로
청담대교
청담대교 분기점
잠원나들목
한남나들목
학동로
삼성동
봉은사로
논현동
탄천IC
잠실종합운동장
코엑스
삼릉공원
선릉로
신반포로
사평대로
반포나들목
테헤란로
역삼동
역삼로
대치동
남부순환로
서초대로
서초동
도곡로
도곡동
도곡근린공원
양재천로
효령로

2. 중구



3. 송파구



Impact and Limitations

한국을 방문하는 외국인 관광객들 중에서 약 76.4%가 서울을 방문하고 있고 그 다음은 경기도 14.9%, 부산 14.1% 순으로 방문하고 있다(문화체육관광부, 2019) 따라서 외국인 관광객의 방문이 가장 많은 서울을 중심으로 분석을 진행하였다.

분석 결과 방한 외국인들의 만족도를 높이고, 방문 목적이 음식 탐방, 쇼핑이 아닌 다른 분야로도 넓히는데에 가장 효과적인 방법을 적절한 위치에 관광안내소 설치라고 보았다.

문화체육관광부는 관광안내를 “관광객에게 관광과 관련된 정보를 제공하여 관광객의 의사결정을 도와주고 편의증진을 도모함으로써, 낯선 지역의 방문에서 오는 불안감을 해소시켜 편안하고 풍부한 관광체험을 하도록 돕는 종합 인적 및 물적 서비스”라고 정의하였다. 이와 같이 관광안내는 관광서비스 향상에 중요하다. 적절한 위치에 관광안내소를 설치하고 외국인들이 많이 사용하는 언어로 보다 다양한 한국의 문화와 관광지를 소개하다보면 자연스럽게 관광객 유치에 도움이 되고 관광사업도 코로나 이전보다도 더 활성화 될 것이다.

그러나 이 프로젝트는 한계가 있다. 위와 같이 군집의 중심에 관광안내소를 설치한다는 것은 무리가 있는데, 먼저 뜬금없는 장소가 최적의 장소로 선정되었을 가능성이 있으며, 관광객의 입장에서 지하철이나 버스 등 대중교통의 위치도 고려해야 한다. 또한 관광지 데이터 셋이 단순히 검색한 횟수를 기준으로 순위를 매겨서 만들어진 데이터 셋이며 그 수도 100행까지 밖에 없어 일반화하기에는 무리가 있다. 마지막으로 프로젝트를 진행하며 느꼈던 가장 큰 문제는 한 군집이 원의 형태로 가장자리에 넓게 분포되어 있고 그 중심을 관광안내소 설치 장소로 지정했을 때, 정작 관광안내소 근처에는 관광지가 없는 문제를 발견할 수 있었다.

따라서 좀 더 볼륨이 큰 데이터셋을 활용하고, 단순히 관광지의 위치 뿐만 아니라 대중교통의 위치, 주택들의 위치를 고려하며 현장조사까지 진행한 뒤 관광안내소를 증설한다면 외국인 관광객의 만족도를 높일 수 있지 않을까 라고 결론 내린다.

Testing

```
16  gmaps_key = 'AIzaSyACFxEfY-a2bR_B7h86pyw07qTz4RKCCIc'  # API설정할 때 얻은 key
17  gmaps = googlemaps.Client(key=gmaps_key)
18
19  def clustering_path(file_path, save_file_name, k_count):
20      allSeoul = pd.read_csv(file_path, encoding='cp949')
21
22      allSeoul['소속구'] = allSeoul['주소'].str.split(" ").str[1]
23      allSeoul
24
25      gmaps = googlemaps.Client(key=gmaps_key)
26
27      lat = []
28      lng = []
29      for name in allSeoul['주소']:
30          try:
31              tmpMap = gmaps.geocode(name, language='ko')
32              tmpLoc = tmpMap[0].get('geometry')
33              lat.append(tmpLoc['location']['lat'])
34              lng.append(tmpLoc['location']['lng'])
35          except:
36              print("{} Address Error".format(name))
37              lat.append(None)
38              lng.append(None)
39      allSeoul['위도'] = lat
40      allSeoul['경도'] = lng
41
42      data = allSeoul[['위도', '경도']]
43      data = data.dropna()
44
```

API 키를 이용해서 위도, 경도를 받아올 때 오류가 종종 발생하여 다음과 같은 코드를 작성했다.

모든 데이터셋의 주소에서 정상적으로 위도, 경도를 받아오는 것을 확인 했으나 가끔 API를 받아오는데 문제가 발생해 모든 행의 위도, 경도 값이 None으로 채워지고 dropna()에 위해 모든 데이터값이 삭제되는 문제가 있었다.

이를 사전에 알기 위해 어떤 주소에서 문제가 발생했는지를 출력하도록 try except를 활용했다. 만약 비정상적으로 많은 주소에 대해 Address Error문이 출력되고 코드가 종료된다면 정상적으로 API를 받아오지 못한 것이므로 재실행을 하여 해결한다.

이외에도 test.py를 작성했다. 모든 module 함수에 대해 정상적으로 실행 되는지 try except문을 이용하여 결과를 출력하도록 한다.