COVID19 데이터 분석

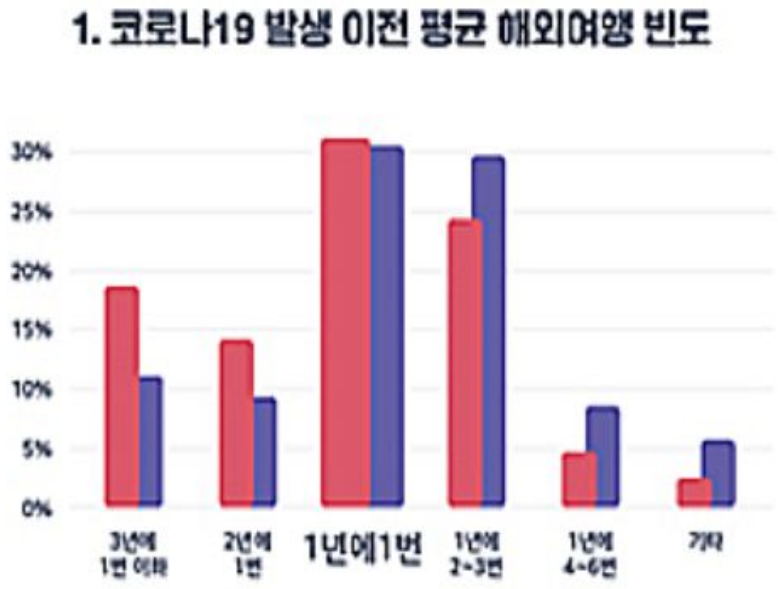
김용주

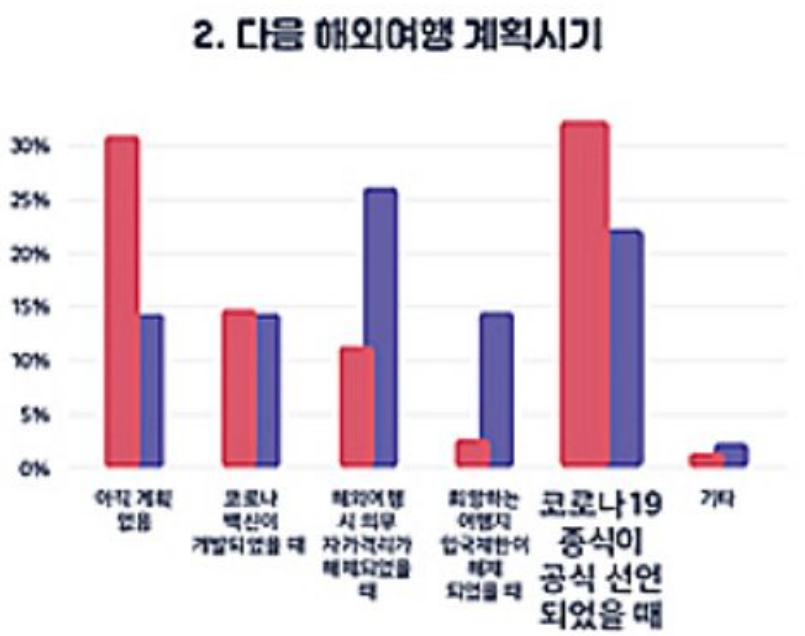
장은준

목차

Project Summary

Project background





Project background

Project purpose

1. 코로나 바이러스 데이터 시각화를 통해서 현 코로나 상황을 진단한다.

2. 코로나 바이러스 데이터 시각화를 통해서 현 코로나 백신 현황을 진단한다.

3. 현 코로나 상황과 백신 현황에 따라서 여행사업 추이를 진단한다.

4. 이러한 세 과정을 통해서 현 코로나 상황에 따른 여행사업 현황을 짐작한다.

Member and Role

김용주

- 데이터 전처리 및 시각화 분석
- ppt 제작

장은준

- 데이터 전처리 및 시각화 분석
- ppt 제작

Process

Data used(1) - 크롤링 데이터

서울특별시 코로나 현황 사이트 데이터 크롤링 :
https://www.seoul.go.kr/coronaV/coronaStatus.do

확진자 현황 (30001번~현재) | 확진자 현황 (1번~30000번)

자치구 전체 | 100 | 검색

연번	환자	확진일	거주지	여행력	접촉력	퇴원현황
75136	236412	2021-08-22	은평구	-	감염경로 조사중	-
75135	237181	2021-08-22	영등포구	-	감염경로 조사중	-
75134	237252	2021-08-22	서초구	-	기타 확진자 접촉	-
75133	236423	2021-08-22	구로구	-	기타 확진자 접촉	-
75132	237182	2021-08-22	강남구	-	기타 확진자 접촉	-

Data used(1) - 크롤링 데이터

Crawling code

```
import pandas as pd
import numpy as np
import requests

def get_seoul_covid19_30001_current(page_num):
    start_no = (page_num - 1) * 100

    url = 'https://news.seoul.go.kr/api/27/getCorona19Status/get_status_ajax.php?draw={}'.format(page_num)
    url = '{}&order%5B0%5D%5Bcolumn%5D=0&order%5B0%5D%5Bdir%5D=desc&start={}&length=100&search%5Bvalue%5D=&search%5B

    response = requests.get(url)
    json_data = response.json()

    return json_data


page_list = []
all_page = 404

for page_num in trange(1, all_page + 1):
    one_page_json_data = get_seoul_covid19_30001_current(page_num)
    one_page_df = pd.DataFrame(one_page_json_data['data'])
    page_list.append(one_page_df)

    time.sleep(0.5)

corona_df = pd.concat(page_list)
for_columns = pd.read_html('https://www.seoul.go.kr/coronaV/coronaStatus.do')
corona_df.columns = for_columns[4].columns.tolist()
```

Crawling result

	연번	환자	확진일	거주지	여행력	접촉력	퇴원현황
0	70870	222111	2021-08-13	기타	-	감염경로 조사중	-
1	70869	221232	2021-08-13	기타	-	감염경로 조사중	-
2	70868	221356	2021-08-13	성북구	-	감염경로 조사중	-
3	70867	221710	2021-08-13	동대문구	-	감염경로 조사중	-
4	70866	220502	2021-08-13	타시도	-	기타 확진자 접촉	-

여기에 용주꺼 수집된 데이터

Data used(3) - 여행사업 데이터

코로나19 여행관련 xlsx

	항목	세부항목	2021년 06월	2021년 05월	2021년 04월	2021년 03월	2021년 02월	2021년 01월	2020년 12월	2020년 11월	...	2010년 10월	2010년 09월	2010년 08월	2010년 07월	2010년 06월	2010년 05월	2010년 04월	2010년 03월	2010년 02월	2010년 01월
0	관광/여행	호텔업	89.7	85.8	77.1	73.9	70.7	57.6	67.5	92.7	...	114.6	105.0	113.5	101.8	100.0	110.6	102.3	103.9	94.5	98.1
1	관광/여행	여관업	67.7	67.6	64.4	61.7	53.0	58.0	65.0	66.6	...	106.7	100.8	114.2	105.6	99.5	99.8	103.1	104.3	98.5	104.3
2	관광/여행	여행사업	22.5	17.4	17.7	16.7	13.3	13.4	18.0	18.3	...	98.5	78.9	85.6	89.7	87.7	95.0	82.6	63.6	67.0	75.3

Preprocessing(1) - 크롤링 데이터

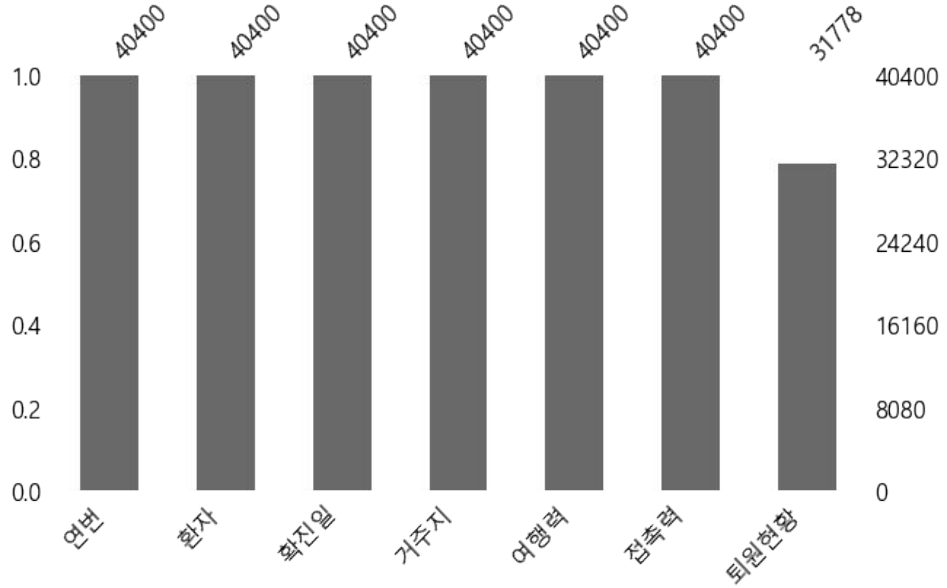
```python
#연번 전처리

def extract_number(num_string):
    if type(num_string) == str:
        num_string = num_string.replace("corona19", "")
        num = re.sub("[^0-9]","", num_string)

        #위의 정규화 식은 밑의 정규화 식을 생략하고 바로 함수화한 것.
        #r = re.compile("[^0-9]")
        #r.sub("",num_string)

        num = int(num)
        return num

    else:
        return num_string

corona_df['연번'] = corona_df['연번'].map(extract_number)


#퇴원현황 전처리

def extract_korean(text):
    extracted_text = re.sub("[^가-힣]","",text)

    return extracted_text

corona_df['퇴원현황'] = corona_df['퇴원현황'].map(extract_korean)
corona_df.loc[corona_df['퇴원현황'].isin([""]),'퇴원현황'] = np.nan
```

	연번	환자	확진일	거주지	여행력	접촉력	퇴원현황
0	70870	222111	2021-08-13	기타	-	감염경로 조사중	NaN
1	70869	221232	2021-08-13	기타	-	감염경로 조사중	NaN
2	70868	221356	2021-08-13	성북구	-	감염경로 조사중	NaN
3	70867	221710	2021-08-13	동대문구	-	감염경로 조사중	NaN

Preprocessing(1) - 크롤링 데이터

Null value handling

1. ‘퇴원현황’ value handling
2. corona_data['퇴원현황'].fillna('퇴원전')



2. ‘여행력’ value handling

```
#"-" 를 여행력 없음으로 간주, 보기 쉽게 바꿀 예정

counted_null2 = len(corona_data[corona_data['여행력'] == '-'])
print('여행력 column null 값 개수 : ', counted_null2)
corona_data[corona_data['여행력'] == '-'].sample(5)
```

Preprocessing(2) - 용주꺼 전처리작업

Preprocessing(3) - 여행사업 데이터

Data array code

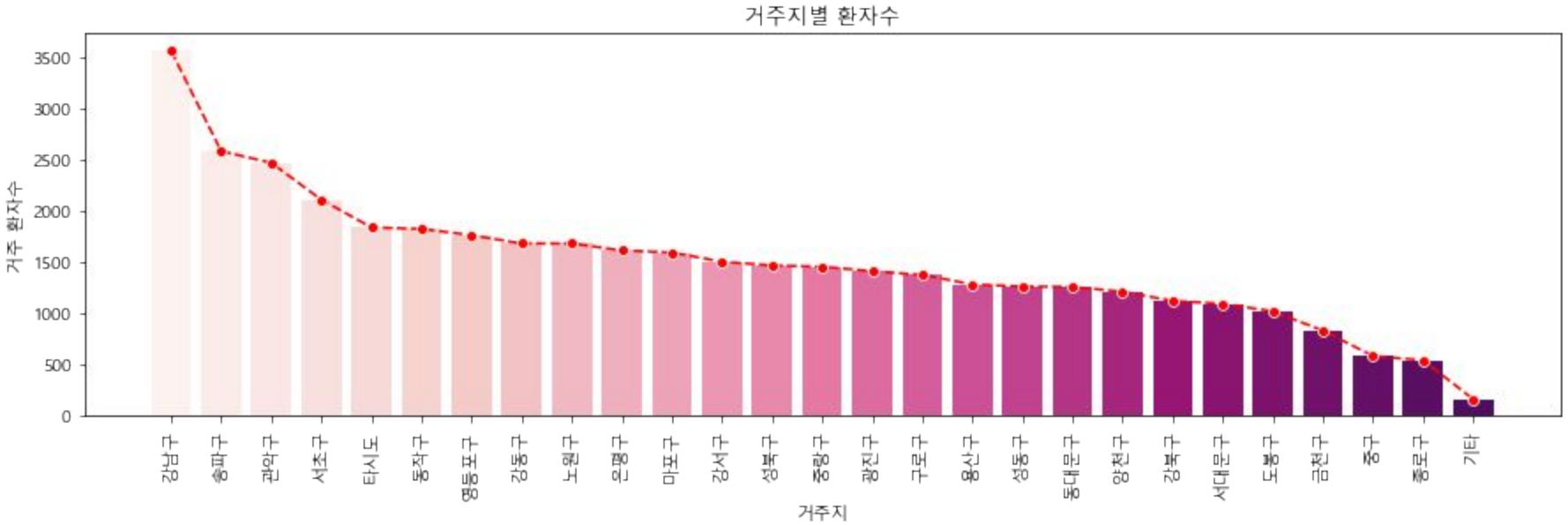
```
trip_industry_df.T

wanted_data = trip_industry_df.T.iloc[2:20,2]
visualization_df = wanted_data.reset_index()
visualization_df['results'] = visualization_df['results'].astype(int)
```

Result Df

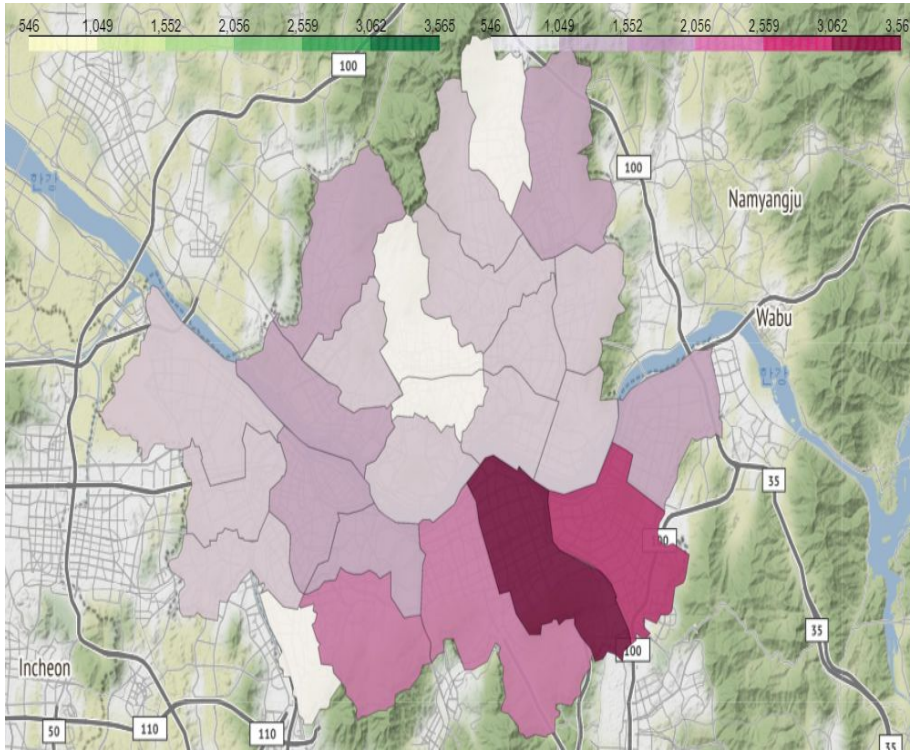
	날짜	results
17	2020년 01월	110.8
16	2020년 02월	53.6
15	2020년 03월	25.6
14	2020년 04월	12.7
13	2020년 05월	12.1
12	2020년 06월	14.1
11	2020년 07월	12.4

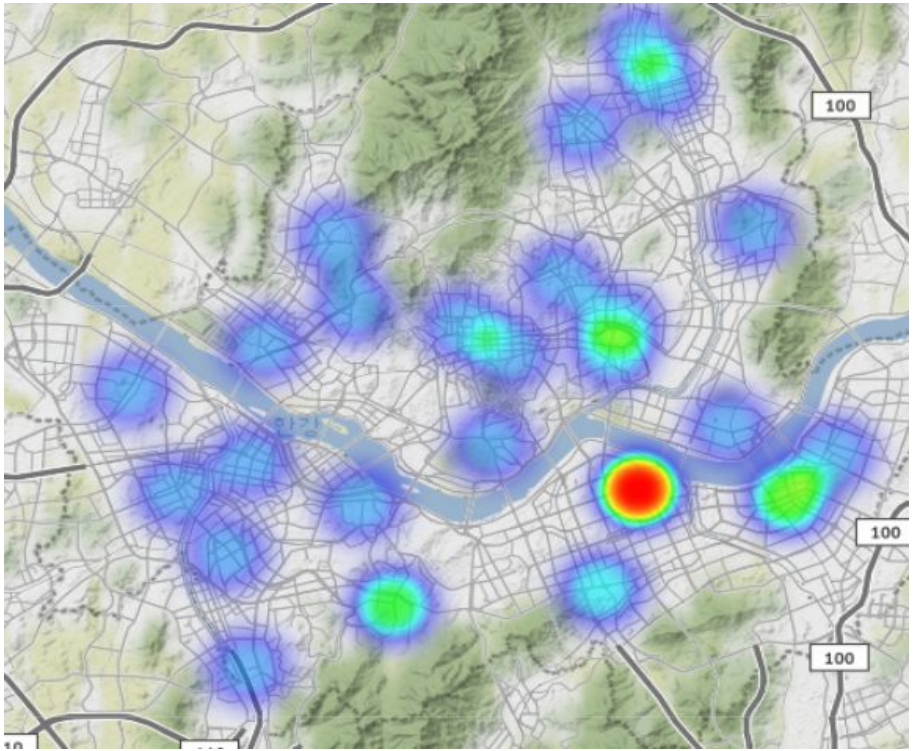
Data Analysis(1) - 크롤링 데이터

1. 거주지별 환자수



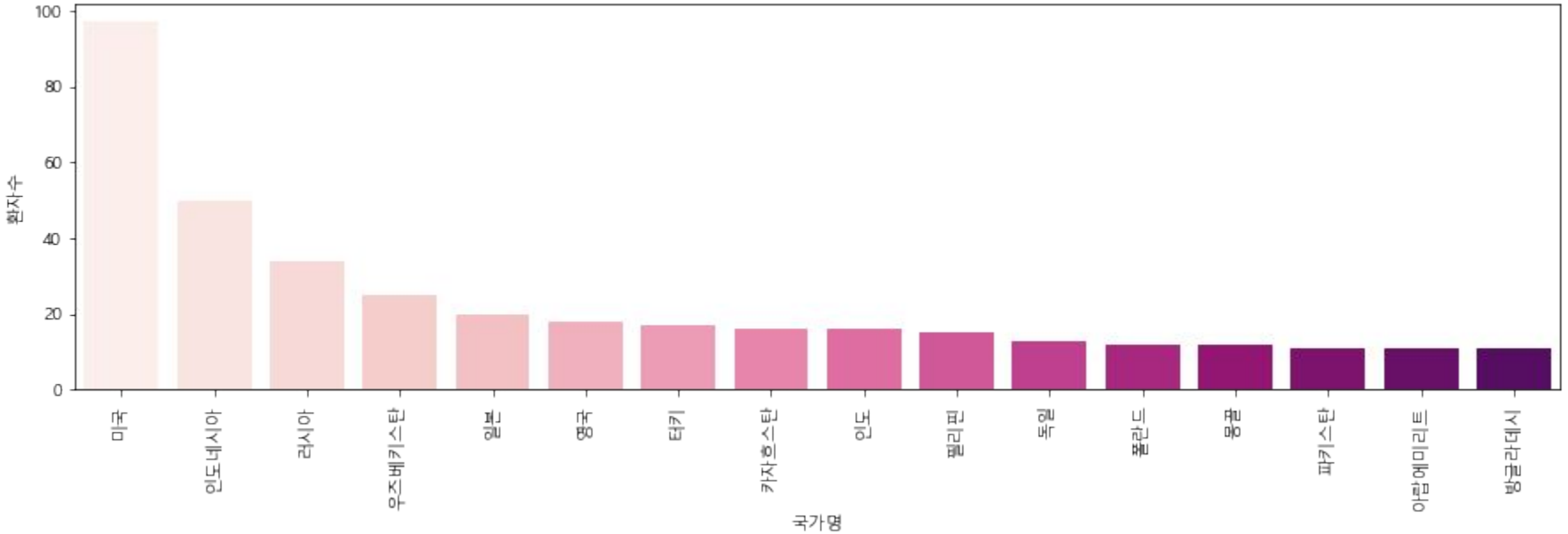
Data Analysis(1) - 크롤링 데이터





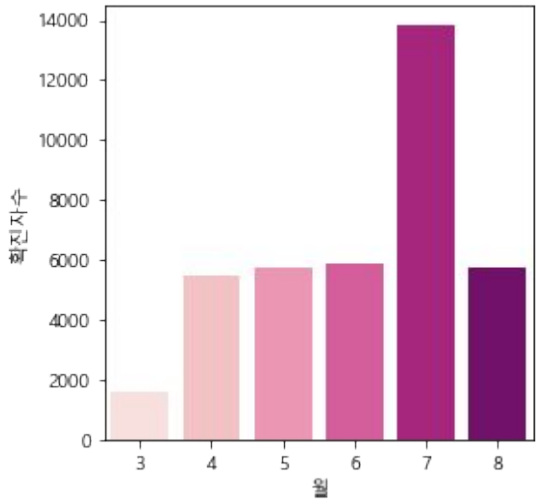
Data Analysis(1) - 크롤링 데이터

2. 해외유입 환자



Data Analysis(1) - 크롤링 데이터

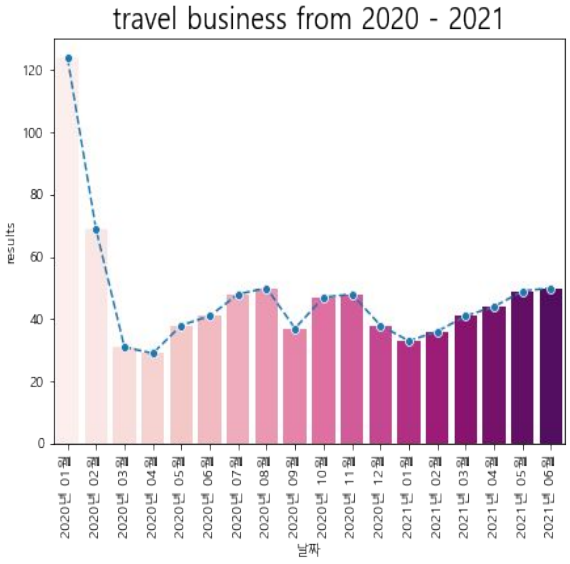
3. 날짜별 감염자 수



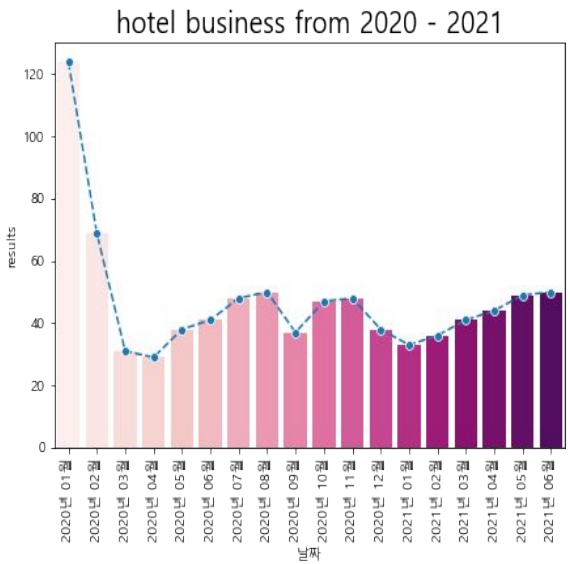
여기 사이는 용주꺼 내용이 들어가고

Data Analysis(3) - 여행사업 데이터

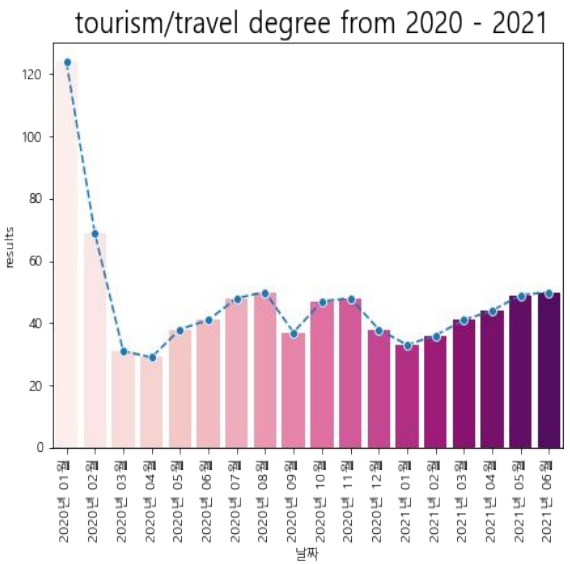
1. 여행사업 추이



2. 호텔사업 추이



3. 관광/여행 지수 추이



Comprehensive Insight

여기서 종합