

COVID19 데이터 분석

김용주

장은준

INDEX

1. Project Summary

- Project background and purpose
- Member and Role

2. Process

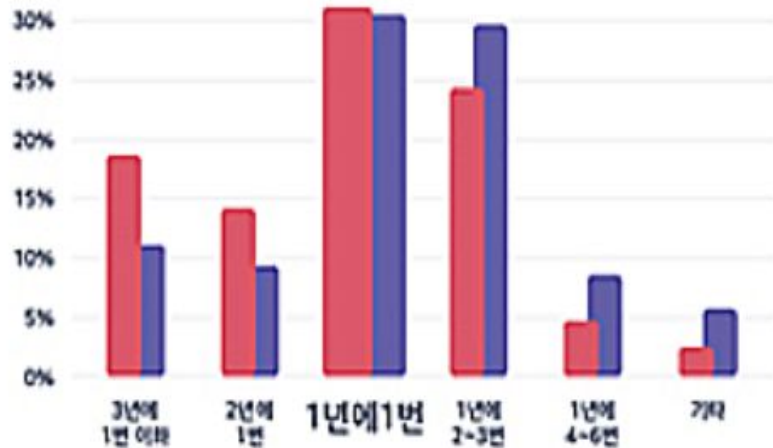
- Data used
- Preprocessing
- Data Analysis

3. Comprehensive Insight

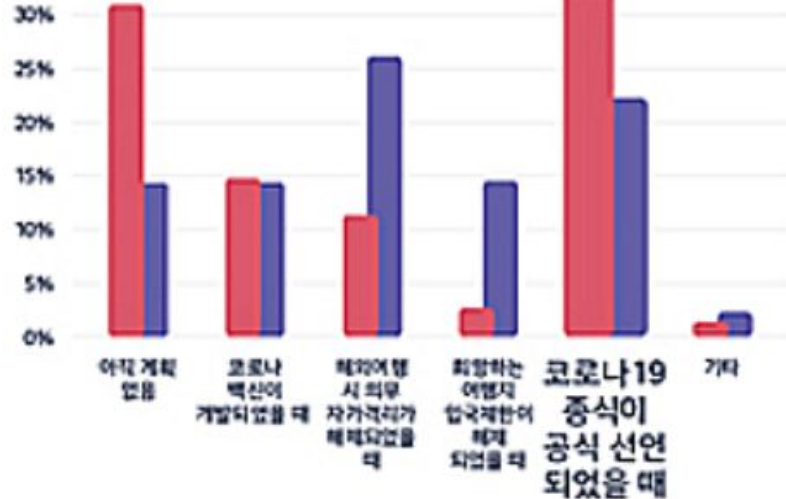
Project Summary

Project background

1. 코로나19 발생 이전 평균 해외여행 빈도



2. 다음 해외여행 계획시기



Project background

 아이뉴스24 **PICK** | 2021.08.08. | 네이버뉴스

포스트 코로나 준비하는 여행업계 "향후 폭발수요 노린다"

 뉴스스 | 2021.08.09. | 네이버뉴스

야놀자 조사, 응답자 80% "코로나 이후 여행 빈도 줄어"

 동아일보 |  B1면 1단 | 2021.08.10. | 네이버뉴스

코로나로 여행 횟수 줄어도 고급호텔 이용 185% 늘어

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 range(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/corona/coronaStatus.do')
corona_df.columns = for_columns[4].columns.tolist()
```

Crawling result

	연번	환자	확진일	거주지	여행력	접촉력	퇴원현황
0	<p class='corona19_no'>70870</p>	222111	2021-08-13	기타	-	감염경로 조사중	<b class='>-
1	<p class='corona19_no'>70869</p>	221232	2021-08-13	기타	-	감염경로 조사중	<b class='>-
2	<p class='corona19_no'>70868</p>	221356	2021-08-13	성북구	-	감염경로 조사중	<b class='>-
3	<p class='corona19_no'>70867</p>	221710	2021-08-13	동대문구	-	감염경로 조사중	<b class='>-
4	<p class='corona19_no'>70866</p>	220502	2021-08-13	타시도	-	기타 확진자 접촉	<b class='>-

여기에 용주꺼 수집된 데이터

Data used(3) - 여행사업 데이터

코로나19 여행관련 xlsx

		항 목	세 부 항 목	2021 년 06 월	2021 년 05 월	2021 년 04 월	2021 년 03 월	2021 년 02 월	2021 년 01 월	2020 년 12 월	2020 년 11 월	...	2010 년 10 월	2010 년 09 월	2010 년 08 월	2010 년 07 월	2010 년 06 월	2010 년 05 월	2010 년 04 월	2010 년 03 월	2010 년 02 월	2010 년 01 월
0	관 광/ 여 행	관 광/ 여 행	호텔 업	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	관 광/ 여 행	관 광/ 여 행	여관 업	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	관 광/ 여 행	관 광/ 여 행	여행 사업	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

#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
```

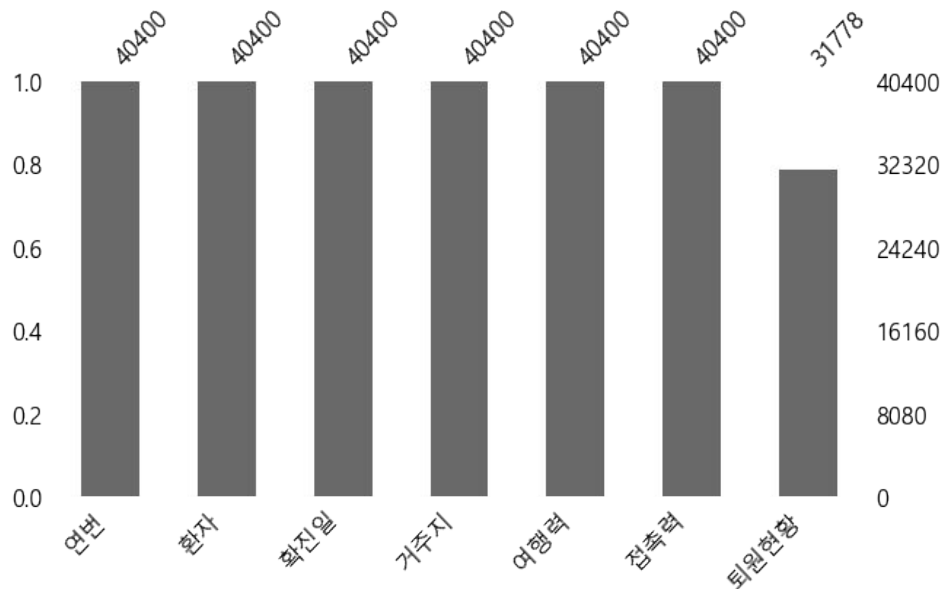
```
corona_df['퇴원현황'] = corona_df['퇴원현황'].map(extract_korean)  
corona_df.loc[corona_df['퇴원현황'].isin([""]), '퇴원현황'] = np.nan
```

	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

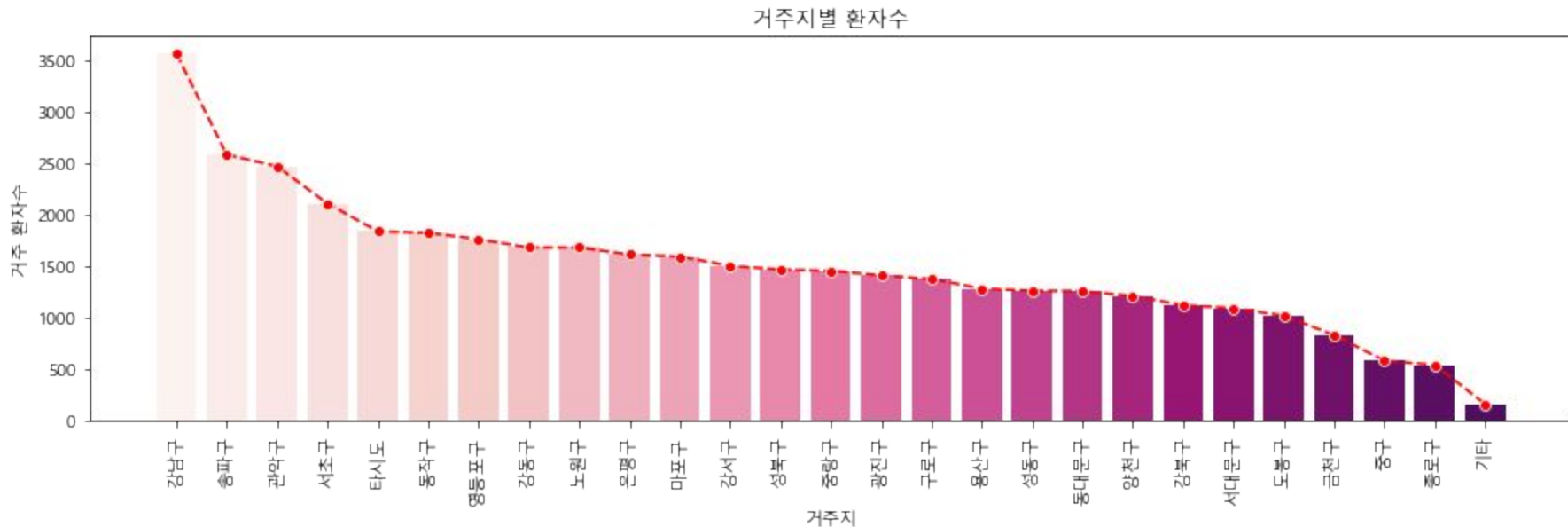
```
: trip_industry_df.T  
  
wanted_data = trip_industry_df.T.iloc[2:20,2]  
visualization_df = wanted_data.reset_index()  
visualization_df['results'] = visualization_df['results'].astype(int)
```

Result Df

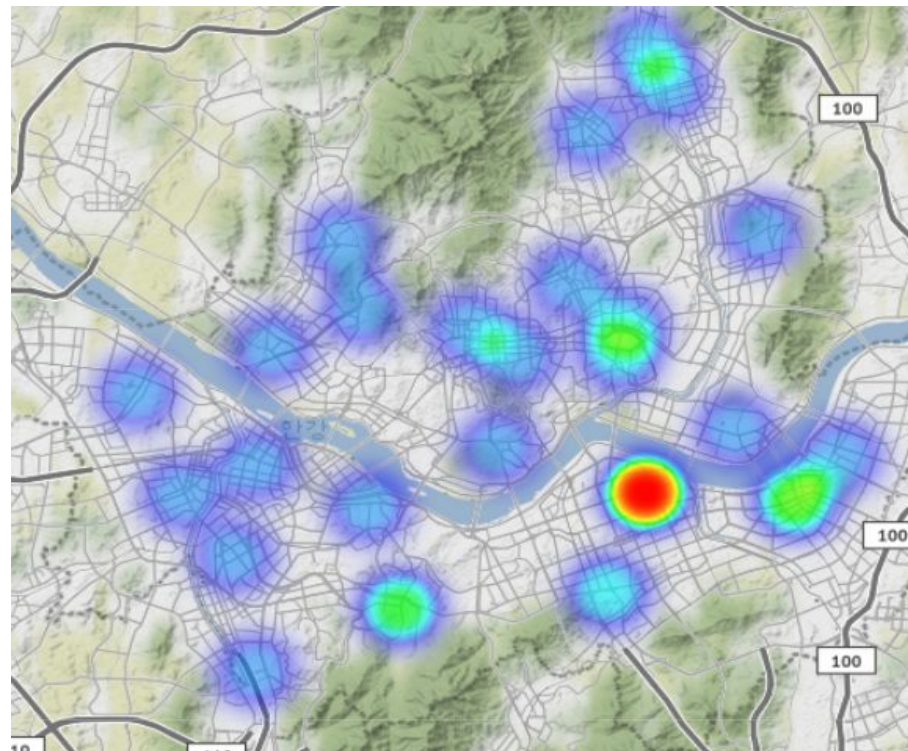
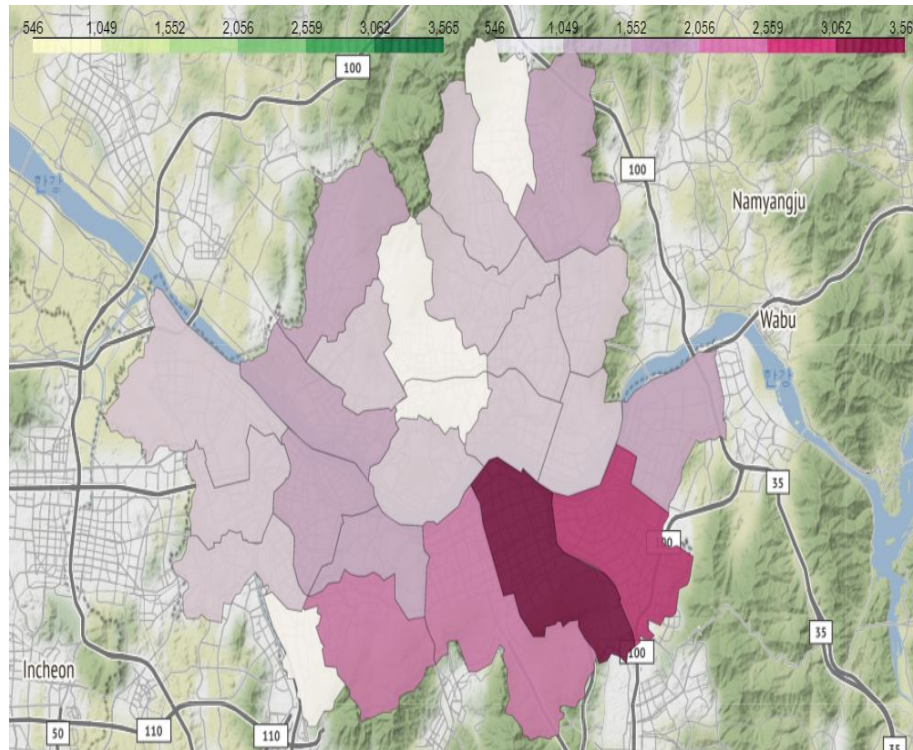
	날짜	results
17	2020년 01월	110.8
16	2020년 02월	53.6
15	2020년 03월	25.6
14	2020년 04월	12.7
13	2020년 05월	12.1
12	2020년 06월	14.1
11	2020년 07월	12.4

Data Analysis(1) - 크롤링 데이터

1. 거주지별 환자수

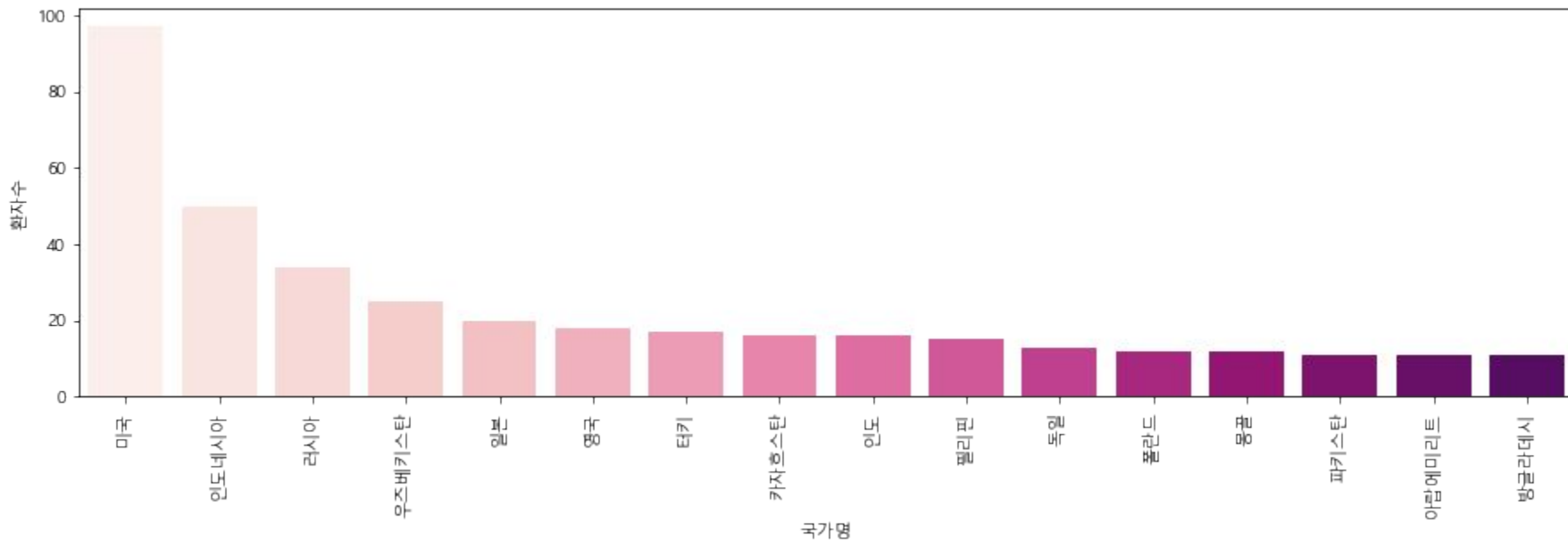


Data Analysis(1) - 크롤링 데이터



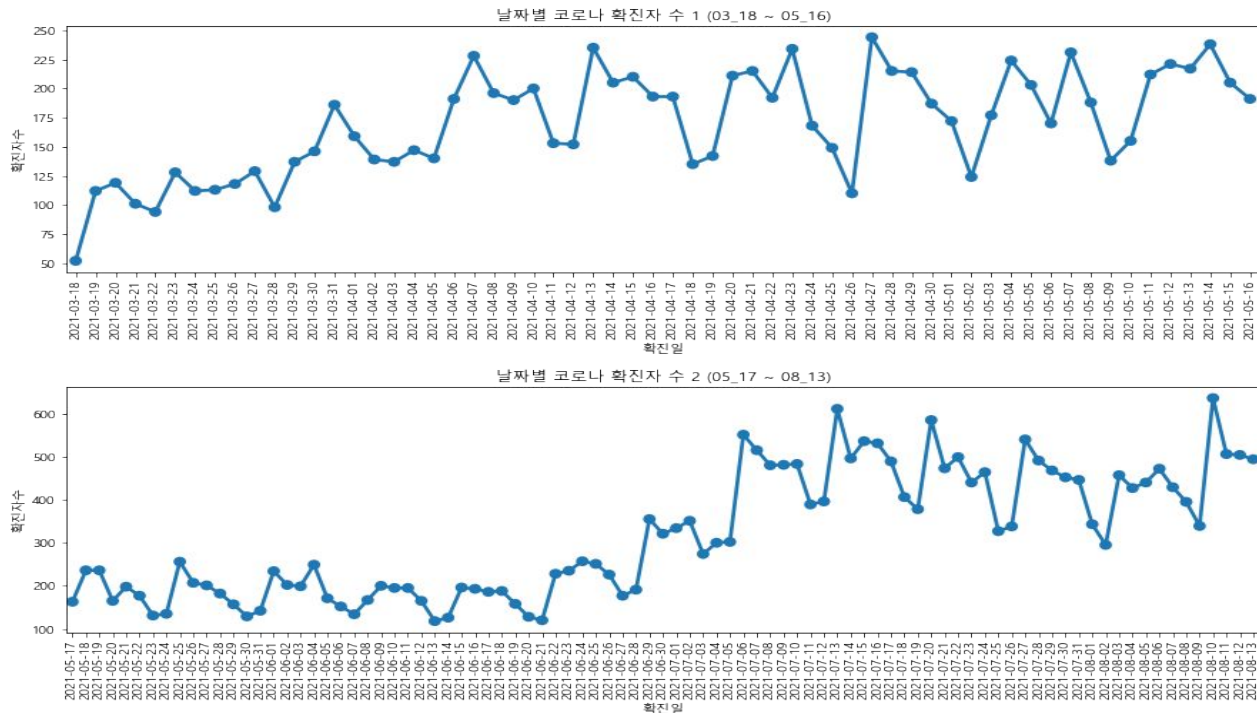
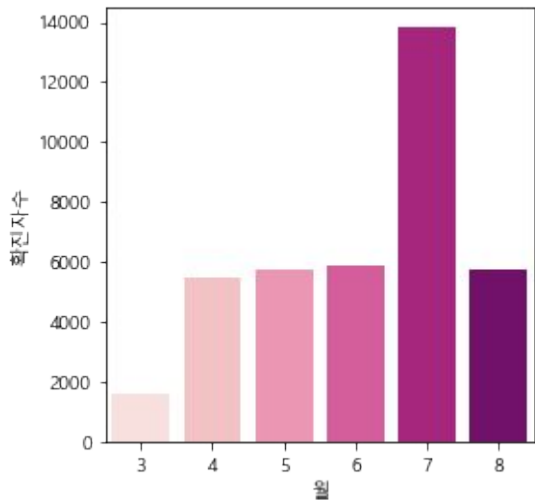
Data Analysis(1) - 크롤링 데이터

2. 국가별 해외유입 환자



Data Analysis(1) - 크롤링 데이터

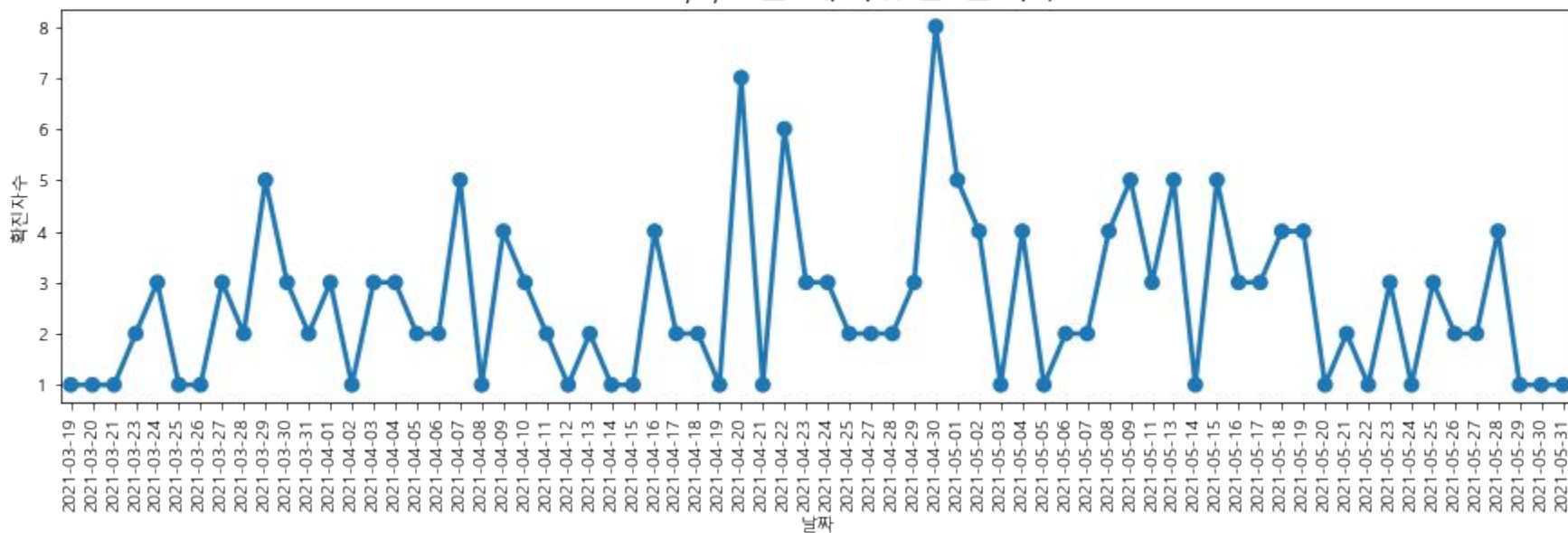
3. 날짜별 감염자 수



Data Analysis - 크롤링 데이터

4. 날짜별 해외유입 환자수

2021. 3,4,5 월 해외유입 환자수

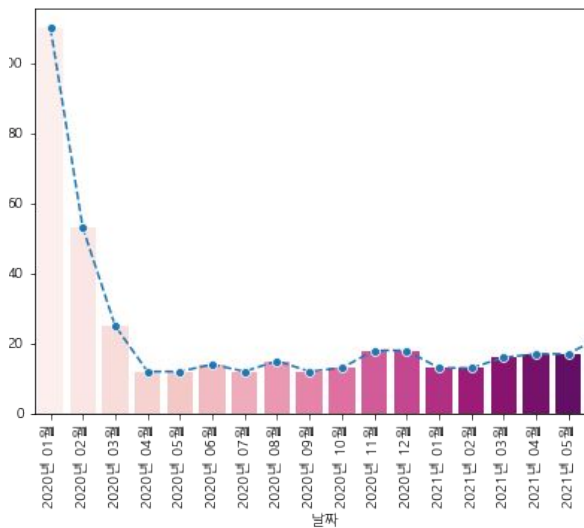


여기 사이는 용주꺼 내용이 들어가고

Data Analysis(3) - 여행사업 데이터

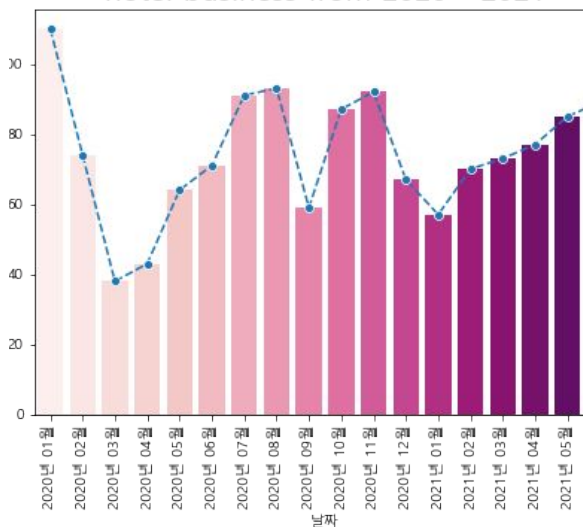
1. 여행사업 추이

travel business from 2020 - 2021



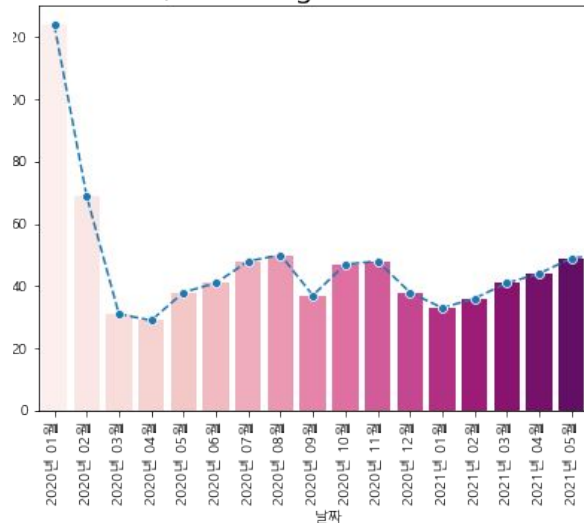
2. 호텔사업 추이

hotel business from 2020 - 2021



3. 관광/여행 지수 추이

tourism/travel degree from 2020 - 2021



Comprehensive Insight

Insight(1)

1. 날짜별 해외유입 환자수 선그래프를 살펴보면, 4.27일을 기점으로 해외유입 환자수가 증가함.
2. 여행사업 데이터의 여행사업과 관광/여행 지수를 살펴보면 4월부터 상승함. 이 사실을 통해서 사람들이 4월부터 여행을 많이 다님을 알 수 있으며 여행을 많이 다니기 때문에 해외유입 환자수가 증가했음을 유추 가능.
3. 4.27일을 기점으로 하여 국내 확진자 수를 보니, 급격한 상승세를 보인다. 이를 통해서 사람들이 확진자가 많이 나오더라도 상관없이 여행을 많이 다님을 알 수 있다.

코로나가 장기화 되더라도 여행에 대한 사람들의 수요는 높기 때문에, 여행사업이 우려와는 달리 침체될 것으로 보이지 않음.

여기서 종합