=

데이터 분석 및 시각화 프로젝트 보고서

통계학과 2017110499 장문주
통계학과 2017110515 이용휘

1. 데이터 설명

case - 감염 유형에 관해 나타낸 데이터이다. case_id는 감염 위치 고유번호, province는 행정구역, city는 시군구, group은 집단 감염 여부, infection_case는 감염장소, confirmed는 확진자 수를 나타낸다. 위 데이터를 통해 감염장소별 확진자 비율을 알아보고자 하였고, 이에 필요한 infection_case와 confirmed에 관한 데이터만 사용하였다.

patient_info - 2020년 상반기 코로나 환자의 정보에 관한 데이터이다. patient_id는 환자 ID, sex는 성별, age는 연령대, country는 국적, province는 행정구역, city는 시군구, infection_case는 감염장소, infection_by는 감염한 사람의 ID, contact_number는 사람과 접촉 수, symptom_onset_date는 증상개시 날짜, confirmed_date는 확진 날짜, released_date는 완치날짜, deceased_date는 사망 날짜, state는 상태를 나타낸다. Patient_Info.csv 파일에서 내국인만을 대상으로 조사하기 위해 국적이 Korea인 대상만 나타내도록, 그리고 연령대와 성별을 기준으로 분석했으므로 그 부분에서 결측치는 제거하도록 전처리하였다. 또한, 위 데이터를 통해 성별 및 연령대에 따른 감염장소별 확진자 수를 알아보는데 필요한 변수인 sex, age, infection_case만 사용하였다.

searchtrend - 한국의 2016년 상반기 ~ 2020년 상반기까지 일별 코로나 및 감기 등에 관한 검색 추세 데이터이다. date는 날짜, cold는 '감기' 검색량 수, flu는 '독감' 검색량 수, pneumonia는 '폐렴' 검색량 수, coronavirus는 '코로나바이러스' 검색량 수를 나타낸다.

time_gender - 2020년 3월 ~ 6월까지 성별에 따른 누적 환자에 대한 데이터이다. date는 날짜, time은 시간, sex는 성별, confirmed는 누적 확진자 수, deceased는 누적 사망자 수를 나타낸다. 위 데이터를 통해 시간의 흐름에 따른 확진자 대비 사망자 비율을 확인하기 위해 치사율에 관한 변수인 Fatality rate를 추가하였고, 주마다 변화를 보여주기 위해 일별 데이터를 주별로 전처리하였다.

time_age - 2020년 3월 ~ 6월까지 연령대에 따른 누적 환자에 대한 데이터이다. date는 날짜, time은 시간, age는 연령대, confirmed는 누적 확진자 수, deceased는 누적 사망자 수를 나타낸다. 위 데이터를 통해 시간의 흐름에 따른 확진자 대비 사망자 비율을 확인하기 위해 치사율에 관한 변수인 Fatality rate를 추가하였고, 주마다 변화를 보여주기 위해 일별 데이터를 주별로 전처리하였다.

corona_month - 2020년 2월 ~ 12월까지 covid19 확진자 수를 월 단위로 나타낸 데이터이다. date는 날짜, confirmed는 확진자 수를 나타낸다.

infectious_disease - 2013년~2019년까지 여러 질병들의 감염자 수를 월 단위로 나타낸 데이터이다. date는 날짜, hepatitis A는 A형 간염, hepatitis B는 B형 간염, epidemic parotitis는 유행성이하선염, chicken pox는 수두, scarlet fever는 성홍열, scrub typhus는 쯔쯔가무시증, hemorrhagic fever with renal syndrome은 신증후군출혈열, tuberculosis는 결핵의 감염자 수를 나타낸다.

swineflu_month - 2009년 5월~2010년 5월까지 신종인플루엔자 감염자 수를 월 단위로 나타낸 데이터이다. date는 날짜, swineflu는 감염자 수를 나타낸다.

High-end Restaurant Revenue_age - 2019년, 2020년의 고급레스토랑 매출을 연령별로 나타낸 데이터이다. date는 년도, age는 연령대, sales는 매출을 나타낸다. 연령대는 20대~60대로 나뉘어 있다.

High-end Restaurant Revenue_gender - 2019년, 2020년의 고급레스토랑 매출을 성별로 나타낸 데이터이다. date는 년도, gender는 성별, sales는 매출을 나타낸다.

space of leisure - 자신이 가장 많이 가는 여가장소를 고르고 장소마다 선택 된 비율을 나타낸 데이터이다. date는 년도, sample size는 표본 수, age는 연령대, vacant lot in apartment는 아파트 내 공터, movie theater는 영화관, restaurant는 식당, large supermarket은 대형마트, cafe는 카페, a park in the living area는 생활권에 있는 공원, health club은 헬스 클럽이 선택된 비율을 나타낸다.

sales - 2015년, 2019년, 2020년의 삼성카드 결제 건수를 소비업종, 성별, 연령별로 나누어 나타낸 데이터이다. 코로나 전 후의 상황을 비교하는 것이 목적이기 때문에 2015년 데이터는 제거하였다.

< 출처 >

▶ kaggle (https://www.kaggle.com/kimjihoo/coronavirusdataset)

▶ kosis

▶ KDX데이터거래소

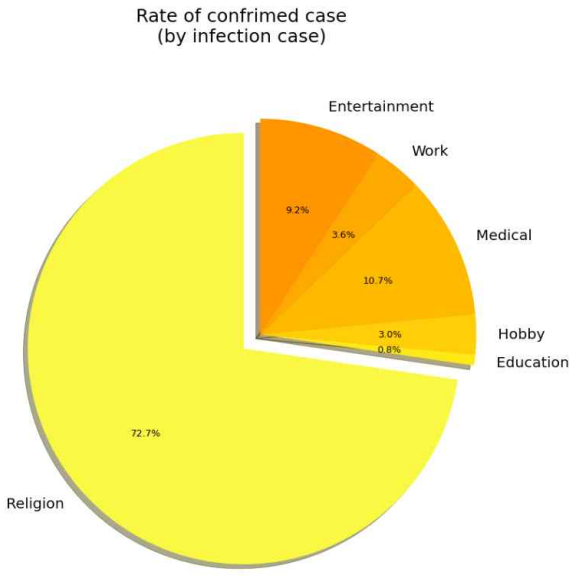
2. EDA 목적

인터넷과 뉴스를 통하여 covid19 확진자 수와 집단 감염 사례 등은 쉽게 알 수 있다. 이에 더 나아가 확진자 수를 장소별로 나타낸 데이터를 통해 확진자가 많이 나오는 장소가 있는지에 대해 파악해 보고자 한다. 예를 들어, 종교 시설, 교육 시설, 유흥 시설 등으로 나누어 확진자의 비율을 구하여 비교하여 많은 확진자가 나오는 특정 장소를 찾아보고 싶다. 더 구체적으로는 특정 연령대 또는 성별마다 확진자가 많이 나오는 장소가 다른가에 대해서도 알아보고 싶다.

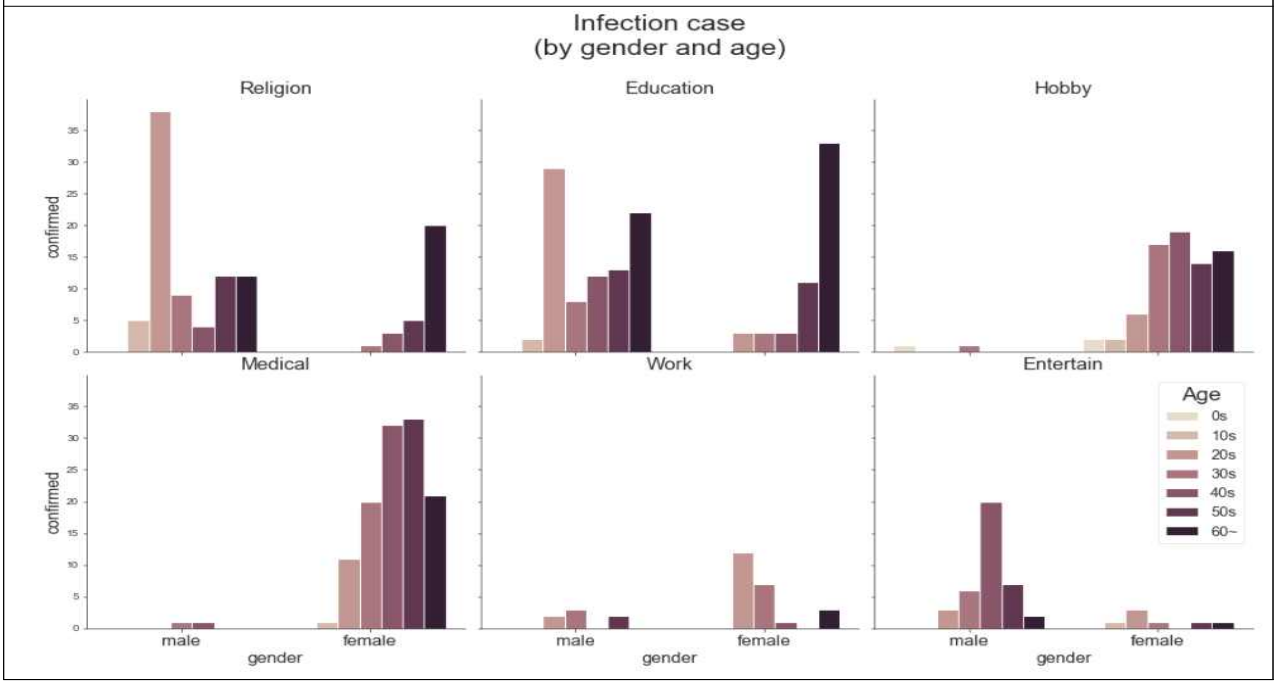
또 covid19와 현재까지 지속해서 감염자가 나오는 질병들(간염, 수두, 결핵 등)의 감염 추세를 비교하여 covid19와 비슷한 추세를 보이는 질병이 있는지, 만약 있다면 그 질병을 더 분석하여 앞으로의 covid19 감염 추세가 어떻게 될 것인지 그 질병을 통하여 간단한 예측을 해보고 싶다.

그밖에도 covid19가 발생하기 전과 후의 상황을 여러 지표를 통해 비교하여 covid19가 우리에게 어떠한 영향을 미치는지도 알아보고 싶다.

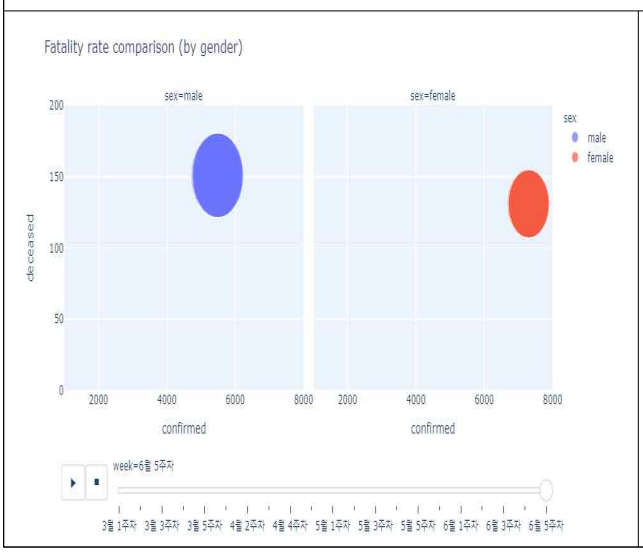
3. EDA/시각화를 통해 얻은 통찰

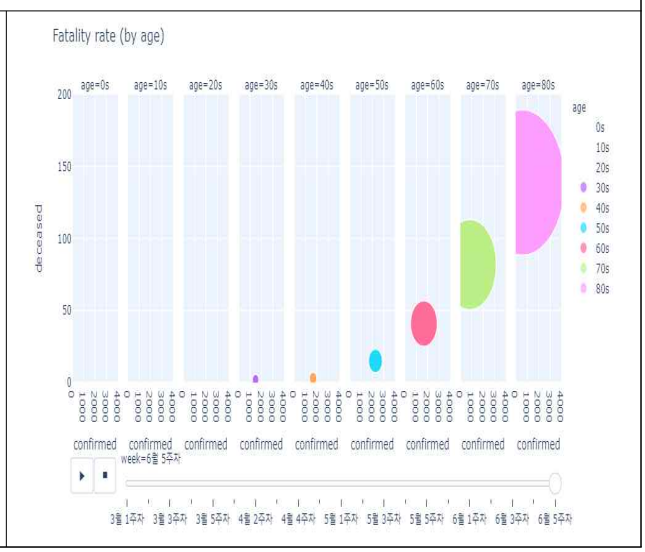
1. 파이차트를 통해 종교 관련 시설에서 다른 시설에 비해 많은 확진자가 나왔음을 알 수 있었다.	

2. 성별, 연령별로 확진자가 많이 나오는 장소가 있는지 파악해 보려고 했으나 결측치와 sample size 문제로 답을 내리기 어려웠다.

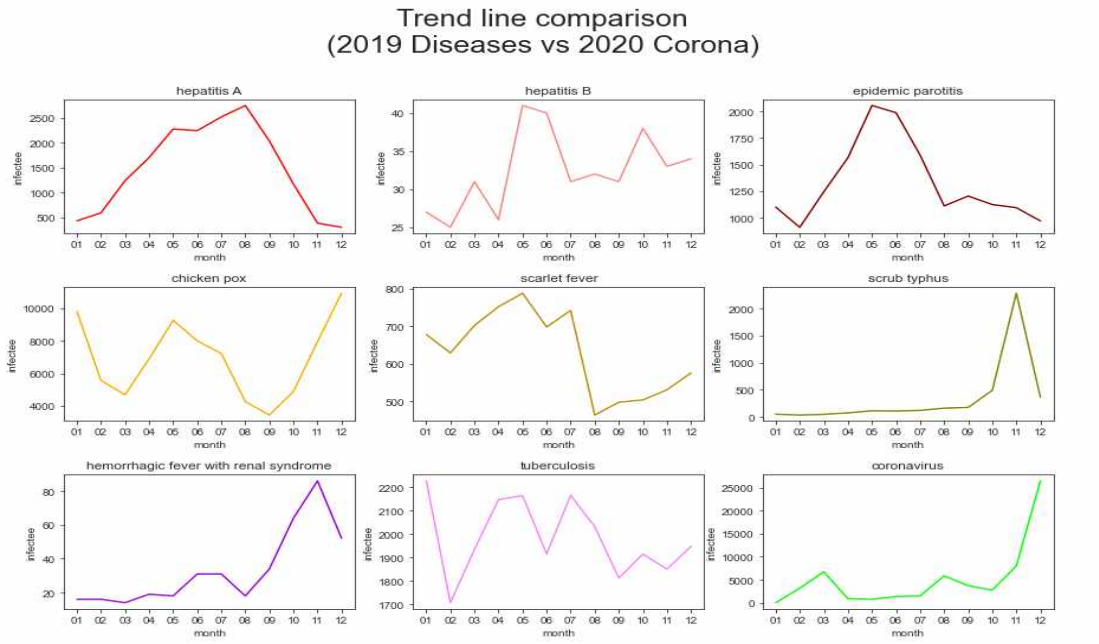


3. 치사율을 비교해본 결과 여자보다는 남자, 연령대는 높을수록 치사율이 높은 것을 알 수 있었다.

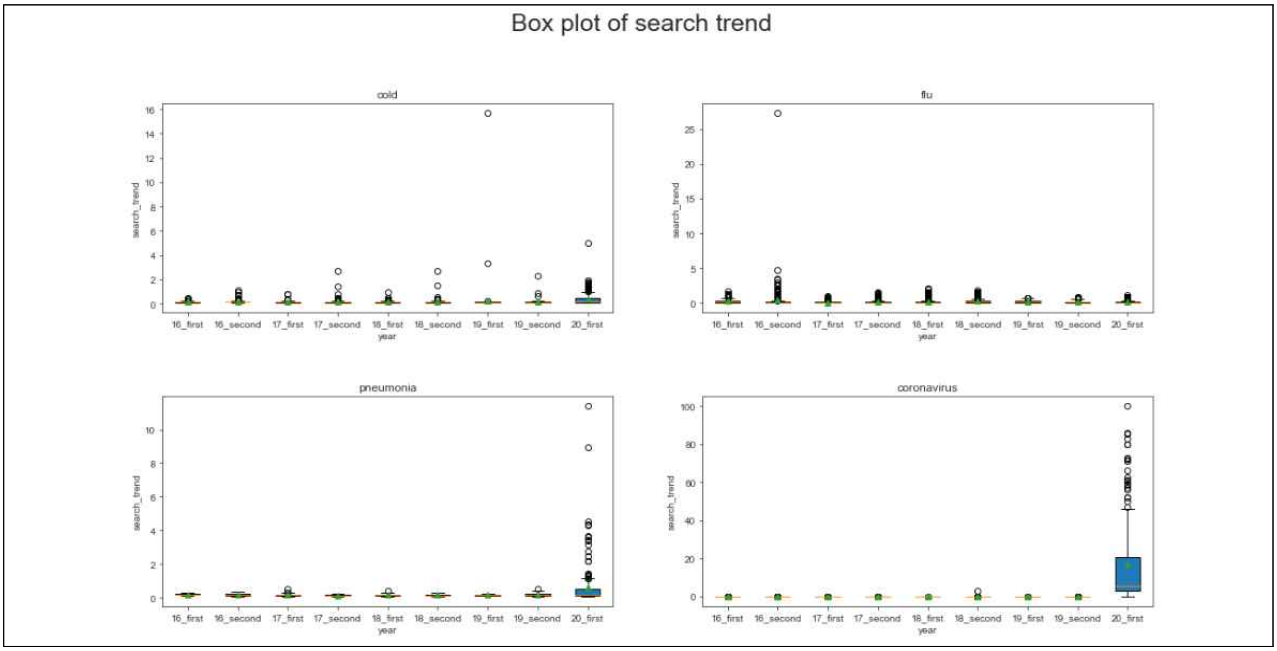




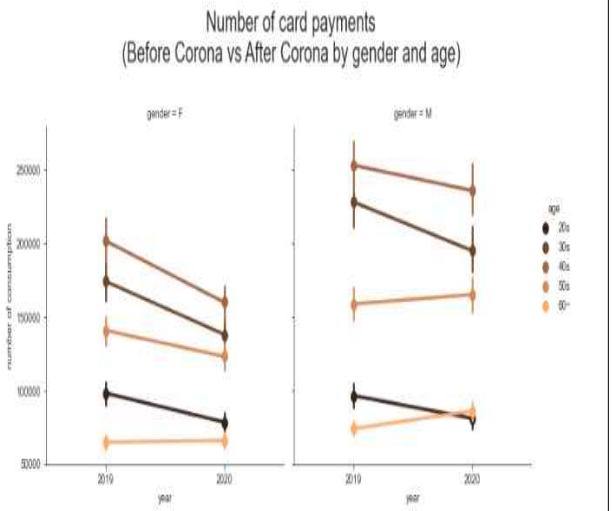
4. 다른 질병들의 2013년부터 2019년까지 감염 추세와 covid19의 감염 추세를 비교해 본 결과 covid19와 비슷한 추세를 보이는 질병은 없었다. 월별이 아닌 일별로 수집된 감염자 데이터가 있었다면 더 섬세하게 비교할 수 있었을 것 같다.

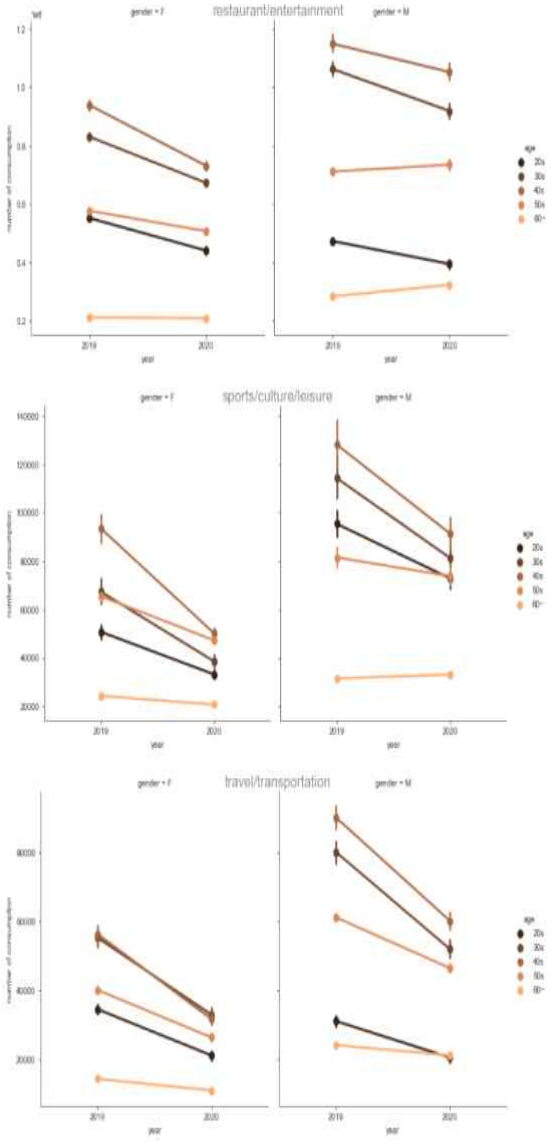


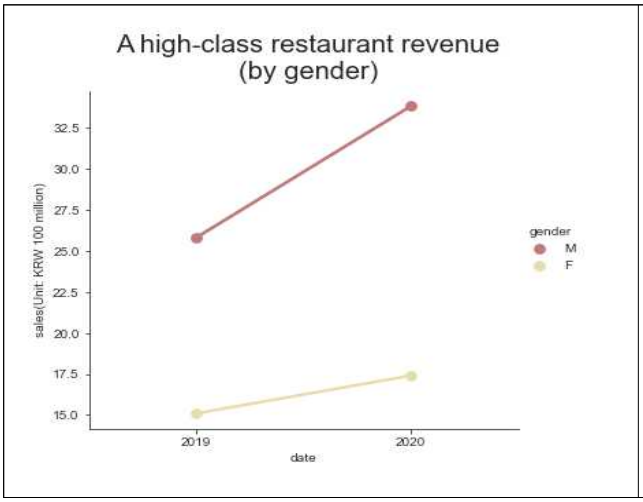
5. 검색 추세는 감기와 플루는 covid19 발생 전과 후의 차이가 없지만, 폐렴의 검색 수는 현저하게 증가한 것이 눈에 띈다. covid19의 증상이 폐렴과 비슷하고, covid19가 '우한 폐렴'으로 불리는 것도 이에 영향을 미친 것으로 생각된다.

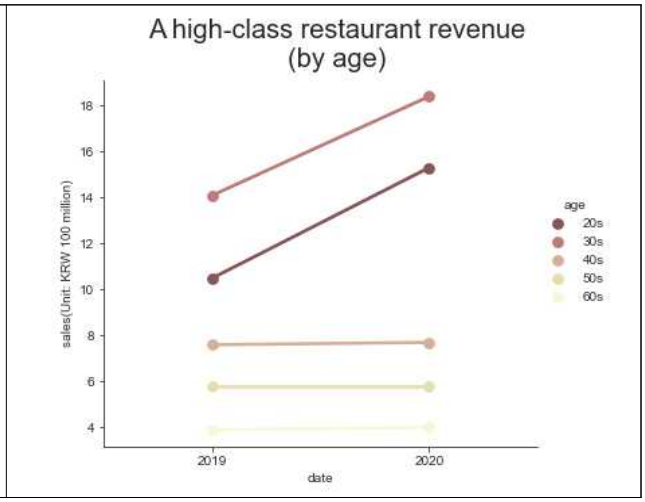


6. 카드 결제 건수는 전체적으로 보았을 때와 업종 별로 나누어 보았을 때 모두 대부분 연령 대에서 감소한 것을 확인할 수 있었다. 요식업 부분에서도 감소하였지만 고급 레스토랑의 매출은 증가하였다. 이를 이상하다고 느껴 조사를 해 본 결과 여행이나 다른 여가 생활에 쓰는 돈이 줄어들고, 고급 레스토랑에 쓰는 돈이 많아졌다고 한다. 일반 식당은 매출이 줄었다고 한다.

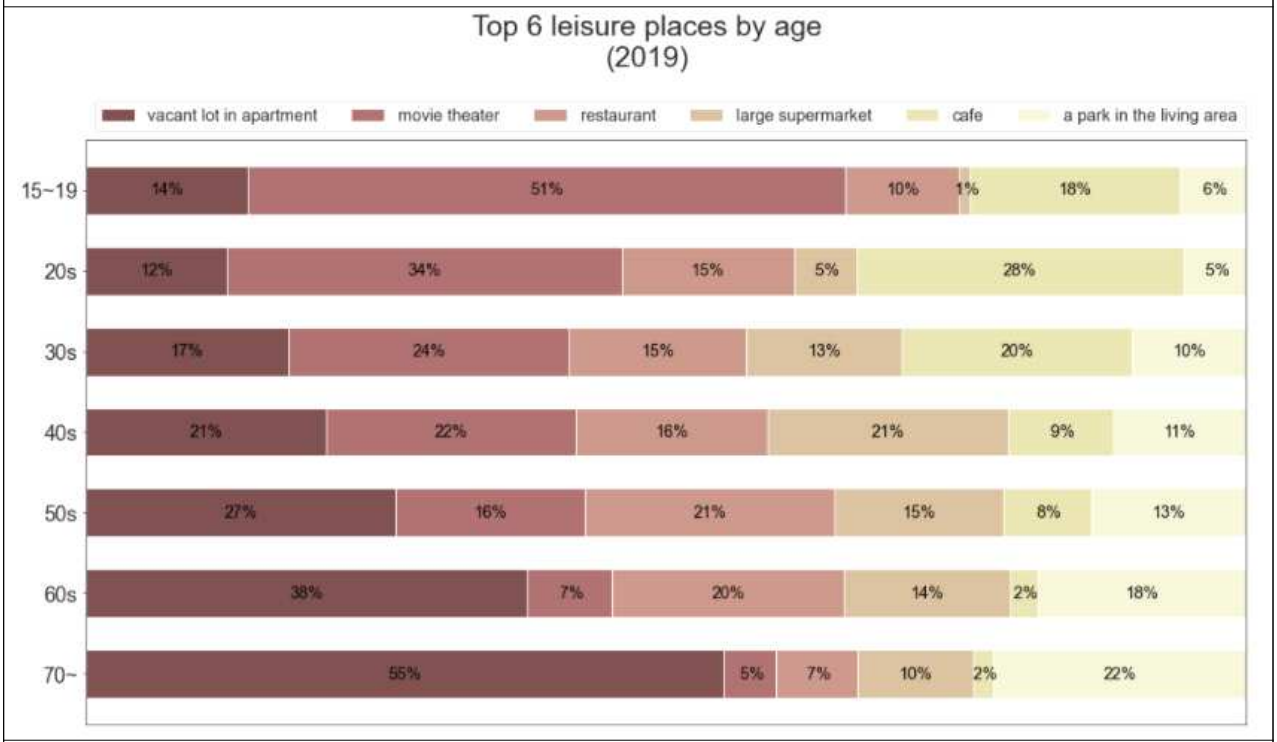


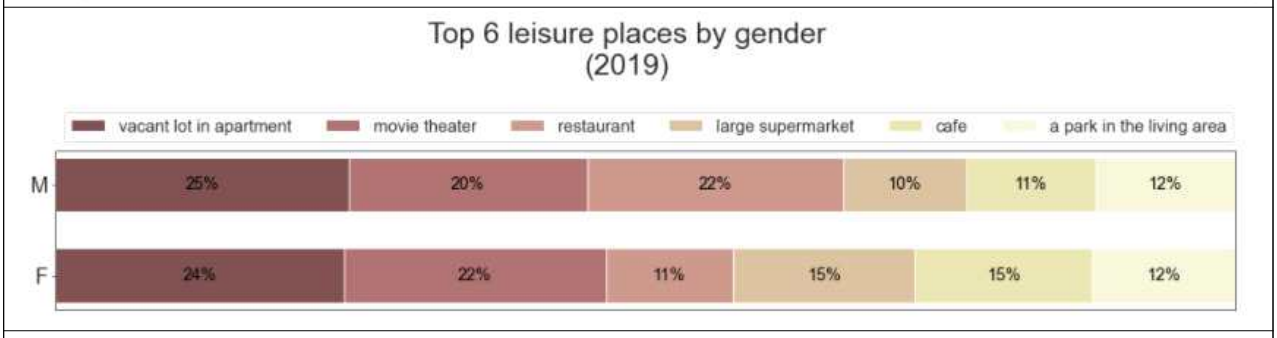


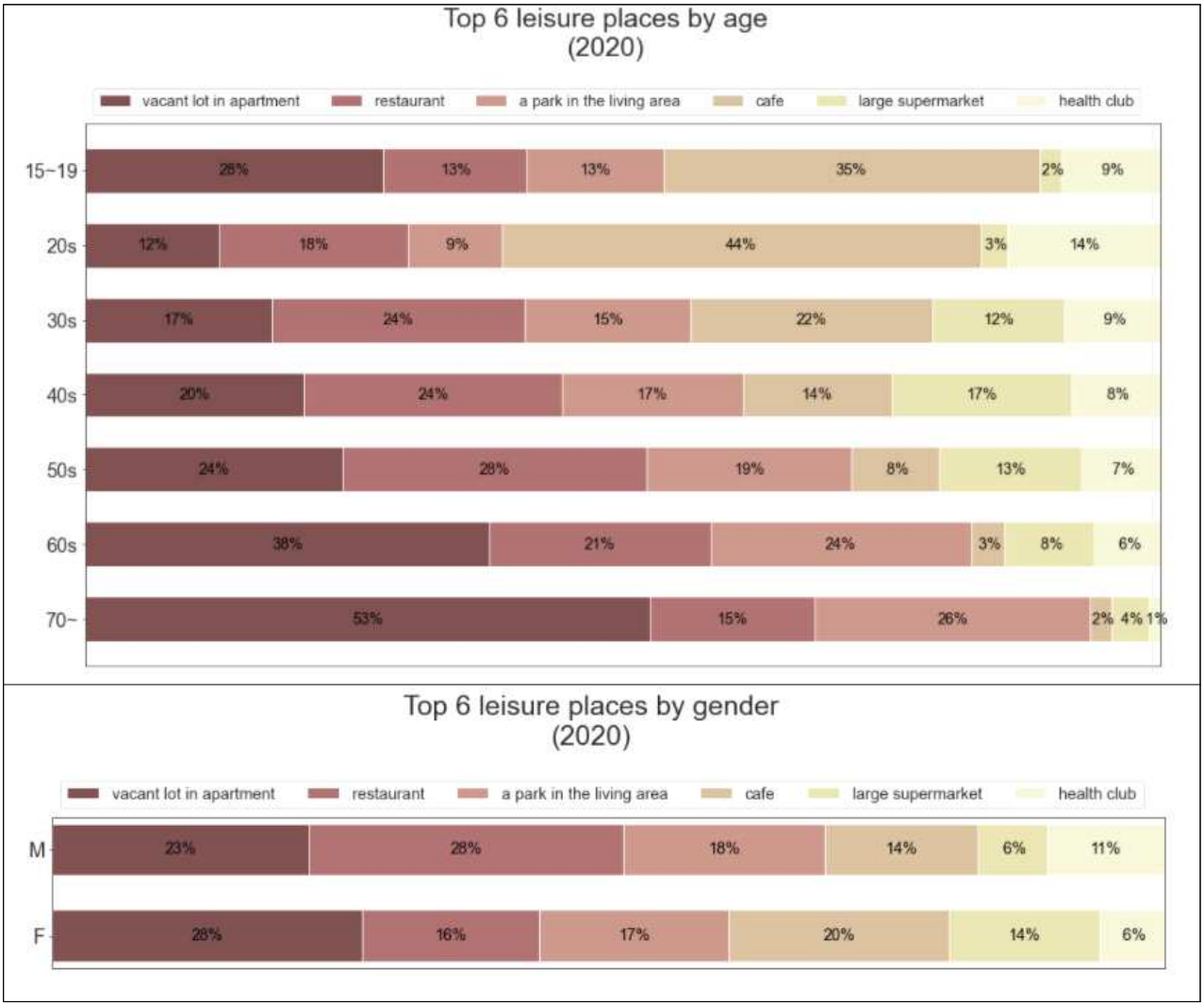




7. 여가 생활의 변화에도 궁금증이 생겨서 데이터를 찾아 보았다. 2019년에는 영화관을 1순위로 뽑은 사람이 많은 비중을 차지하였지만 2020년에는 영화관을 거의 가지 않는 것으로 보이지만 카페를 많이 가는 사람의 비중은 오히려 늘어났다. 영화관 대신 헬스 클럽이 어느 정도의 비중을 차지하게 되었다.







4. 보완점

1. covid19가 다양한 분야에서 어떤 영향을 끼쳤는지 알아보기 위해 자료를 수집하려고 했지만 원하는 시기의 특정 물품의 매출 데이터와 같이 상업적인 정보로 활용할 수 있는 자료 등 가치를 창출할 수 있는 데이터의 경우에는 찾기 어려웠고, 찾는다고 하더라도 무료로 자료를 배포하는 경우가 드물었다. 그에 따라 기존에 공적인 용도로 또는 무료로 공유된 한정된 자료들 속에서 목표와 연관 지어 생각해볼 수 있는 자료를 선택해 분석했다는 점에서 새롭게 분석할 만한 부분이 적었다는 것이 아쉬웠다.

2. covid19에 관한 데이터가 전국을 대상으로 하는 데이터여서 직접 조사하는 데 무리가 있어 기존에 배포된 데이터를 이용해 분석하는데 수집한 자료 내에서 결측치가 많이 존재하였고 이를 추측할 수 있는 수단이 없어 이를 제거한 채 분석을 했다는 점에서 설명력이 많이 부족했다고 생각한다.