

「2020 빅콘테스트」 데이터 분석 계획서

* 해당란에 ☒ 표시

참가분야	<input checked="" type="checkbox"/> 혁신아이디어분야	<input type="checkbox"/> 데이터분석분야
세부분야	<input type="checkbox"/> 퓨처스리그	<input type="checkbox"/> 챔피언리그 *데이터분석 분야에 한함
개인/팀여부	<input type="checkbox"/> 개인 <input checked="" type="checkbox"/> 팀(구성원 3명)	개인/팀명 Moim
대표ID	leewon33373@gmail.com	

※ 5장 내외로 목차는 준수하여 자유롭게 작성

분석 주제명	코로나 이후 산업분석을 통한 상권지수 개발
분석 배경	<p>코로나 이후 우리들은 언택트 시대를 맞이하면서 대부분의 생활들을 온라인에서 해결하고 있다. 이러한 사회적 변화는 사람들의 소비에도 영향을 미쳤으며 유례없는 경제 불황을 야기하였다. 그럼에도 불구하고 어떠한 형태로든 이루어지고 있는 사람들의 '소비 활동'에 집중하여 이와 밀접한 관계가 있는 물류, 유통, 신한카드 등의 데이터를 분석해보기로 하였다. 이를 통해 감염병으로 인해 바뀐 소비의 흐름과 사람들의 패턴을 파악하여 코로나 이전과 이후에 어떠한 차이가 있었는지를 파악해볼 것이다. 나아가 구체적으로 어떠한 이유로 상권의 매출에 변화가 생겼는지 그리고 앞으로의 바이러스가 다가올 경우 이 지역의 산업 전망이 어떻게 될 것인지를 예측함으로써 다음 바이러스 시대에 여러 산업들이 경제 불황을 대비하고 이를 예방할 수 있는 가이드라인을 제공하고자 한다.</p>
분석 내용 요약	<p>전반적으로 소비가 줄었지만, 유통관련 소비에서 차지하는 항목들의 비율은 감염병 상황과 관련없이 유사한것으로 나타났다. 외출이 자제되는 상황속에서 물류량은 폭증하였고 선호하는 소비 품목에는 변화가 나타났다. 유통인구는 전반적으로 크게 감소하였고 밤과 새벽에 더 큰폭으로 감소하였다. 감염상황이 동일함에도 불구하고 동네별로 매출 하락율은 상이하였고 산업 대비 '의복', '문화취미', '숙박' 산업이 크게 감소. 반면 '가구', '전기' 사업은 증가한 것으로 나타났다.</p>
분석방법 및 계획	<p>▶ 분석 방향</p> <p>우리는 코로나19 전후 지역 별 상권분석을 통해 매출이 증가하였는지, 감소하였는지, 해당 산업이 배달로 전환하여 매출을 꺾을 수 있는지 등을 총망라해 산업군에 따른 활성화, 성장성, 안정성 등을 나타내는 경제안정지수를 도출해낼 것이다. 그리하여 전염병 발생 시 지역경제를 위해 어느 정도의 경제적인 면역력을 가지고 있는지 수치화하고 예측하여 제공할 것이다.</p> <p>▶ 분석에 활용되는 추가데이터</p> <p>- 코로나 데이터</p>

지역 별 코로나19 확진자 데이터 수집

- 소상공인시장진흥공단_상가(상권)정보_대구, 서울데이터

2020년 6월 30일자의 대구, 서울 지역 업종 대분류, 소분류, 업소명, 도로명 주소, 위도, 경도 등 데이터 존재

▶ 분석 프로세스

1. 데이터 수집 및 전처리

수집한 데이터를 우리가 원하는 형태에 맞게 가공 및 선별

2. 상권 내 데이터(매출액, 물류 전환률, 유통 등) 분석

3. 경제안정지수 및 지표 개발

4. 지수 예측 모델 구축

5. 모델 평가

6. 지도 시각화

▶ 분석에 적용·활용할 통계·분석 기법, 방법론

1. 경제안정지수 및 지표 개발에 적용할 기법들

- 연관규칙학습 (Association Rule Learning)

지역별로 어느 산업이 함께 소비되었는지를 '연관 규칙'을 통해 분석함으로써 해당 지역의 산업 소비 경향과 코로나 이전 또는 이후의 매출 변화를 알아 내 지수에 영향을 미칠 요인들을 파악한다.

- 회귀분석 (Regression Analysis)

신한카드 데이터와 물류데이터에서 구한 구매 금액 또는 구매 건수에 대해 GS 유통데이터의 매출지수와 연관시켜 변수 사이의 관계를 회귀분석을 통해 찾아낼 수 있다. 즉, 매출에 대해 영향을 미치는 변수들을 찾고 우리의 지수 개발에 도움이 되는 변수들을 추론하기 위해 회귀분석을 활용한다.

- 소셜네트워크 분석 (Social Network Analysis)

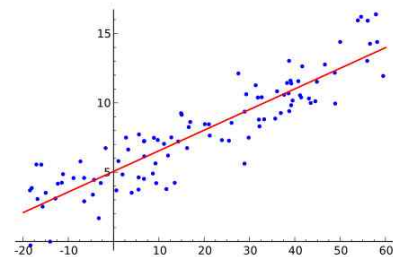
지역별 산업 매출과 유동인구 변화에 SNS 또는 블로그, 뉴스 등이 어떤 영향을 끼쳤는지 알아 보기 위해 소셜 네트워크 분석을 활용할 것이다. 분석 결과를 우리의 지표에 활용할 것이다.

Find *associations* or *co-occurrence* between entities

– (ex) What items are commonly purchased together?



An association rule: {Diapers} → {Beer}
("customers who buy diapers tend also to buy beer")



2. 지수 예측 모델에 적용할 기법들

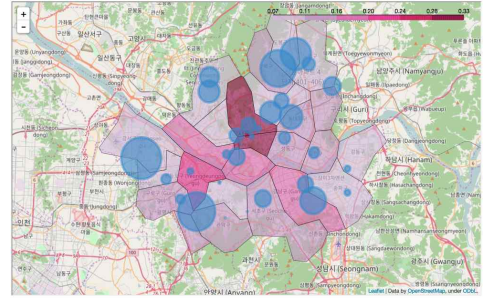
- XGBoost 및 앙상블 기법

머신러닝의 앙상블 모델 중 하나인 xgboost를 활용해 개발한 경제안정지수를 예측한다. 과적합 방지가 가능한 규제가 포함되어 있으며, CART 기반으로 분류와 회귀가 둘 다 가능하다. 경제안정지수를 구간화하여 예측해낼 수도, 점수 자체를 예측해낼 수도 있다.

▶ 분석 결과에 대한 시각화 방법

① Folium Map

지역별로 각 산업의
흥망성쇠를 알아보기 위하여
먼저 '지도'라는 시각적 도구
를 사용한다.

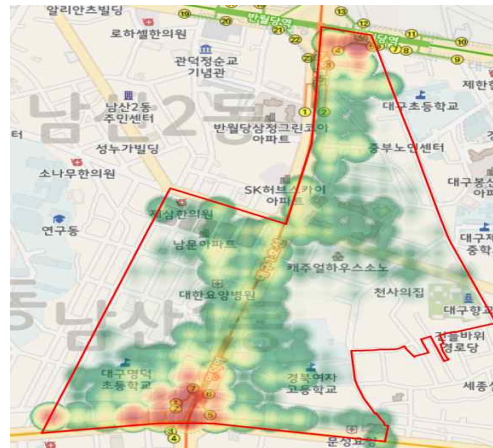


② Three dimensional Rectangle



각 지역마다 어떤 사업의
성장했고 퇴화했는지
색깔과 막대를 이용하여
한 눈에 알아볼 수 있게
Three dimensional
Rectangle 시각화 방법을
사용할 것이다.

③plotly 패키지를 활용한 인터랙티브 시각화



기본적인 그래프 시각화들과 항목별
로 클릭하면 움직이는 그래프를 만들
것이다. 또한 해당 시각화 기법을 활
용해 인구 밀집도에 관련한 시각화를
구현할 것이다.

분석결과 활용
및 시사점

분석결과는 모두에게 활용될 수 있게 공개하지만 주 타겟층은 자영업자들이다. 각 산업군과 상권, 지역 등이 감염병 상황으로부터 얼마나 강한 경제적인 면역력을 갖는지 수치화하여 제공함으로써 지역상권과 일반 시민들에게 전염병에 대비한 경제적인 가이드라인을 제공할 수 있을 것으로 기대한다. 더 넓은 범위의 데이터를 통해 범위를 넓혀간다면 국가단위의 경제적 어려움 해소에 도움을 줄 수 있을 것이다.

우리가 제공하는 경제안정지수가 앞으로 다가올 뉴노멀 시대에 입점 시 고려해볼 만한 새로운 지표 중 하나가 될 수 있을 것을 기대한다.

※ 제출자료는 최종 제품작 평가시 활용될 수 있음