2024학년도 2학기 중간과제물(온라인 제출용)

교과목명	: 빅데이터의이해와활용
학번	: 202034-153746
성명	: 이동열
연락처	: 010-5264-5565

1-1

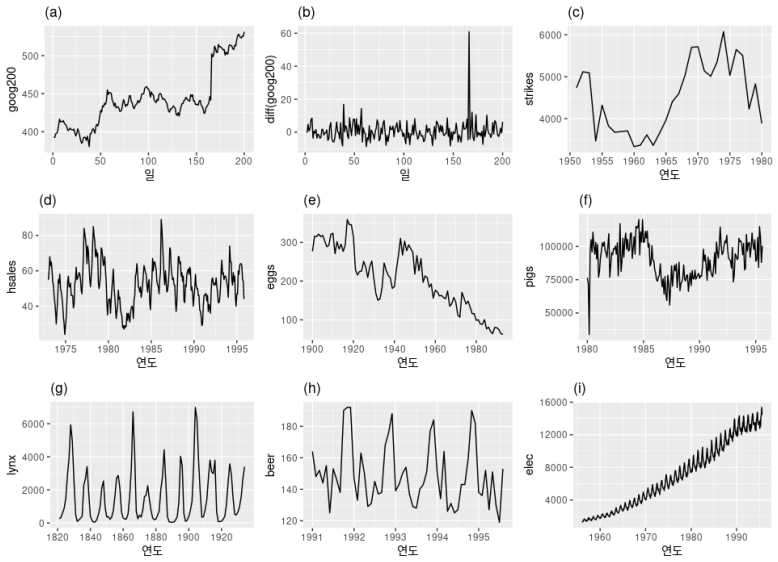
시계열 데이터를 분석하기 위한 기법으로 평활화, 차분, 변수변환이 있다. 평활화는 변동 짧은 주기를 가진 변동 요인을 제거해 단기적인 흐름을 확인할 수 있게하는 방법으로 중심화이동평균, 가중이동평균, 후방이동평균, 이중이동평균 등이 존재하며 이 중에서 중심화이동평균 평활화 기법이 가장 많이 이용된다. 중심화이동평균 평활화 기법을 사용한 시계열은 기존 시계열과 시차구조가 동일하며 이동평균항 수를 늘릴 경우 장기적인 흐름도 파악할 수 있다는 장점이 있다. 하지만 시계열의 처음과 마지막 자료가 없어 이동 평균값을 구하지 못해 장기적인 이동평균을 구하는데 제한이 있다는 단점이 있다.

연도	판매량 (GWh)	5-MA
1989	2354.34	
1990	2379.71	
1991	2318.52	2381.53
1992	2468.99	2424.56
1993	2386.09	2463.76
1994	2569.47	2552.60
1995	2575.72	2627.70
1996	2762.72	2750.62
1997	2844.50	2858.35
1998	3000.70	3014.70
1999	3108.10	3077.30
2000	3357.50	3144.52
2001	3075.70	3188.70
2002	3180.60	3202.32
2003	3221.60	3216.94
2004	3176.20	3307.30
2005	3430.60	3398.75
2006	3527.48	3485.43
2007	3637.89	
2008	3655.00	

출처 : https://otexts.com/fppkr/moving-averages.html#moving-averages

이에 대한 예시로 1989 ~ 2008년 동안의 호주 남부 주거용 전기 판매량 표가 있다. 위 표에서 5-MA라고 적힌 부분은 차수가 5인 이동평균을 의미하는데 맨 마지막 열은 두 번째 열의 5년치 판매량의 평균을 의미한다. 예를 들어 1991년 행의 마지막 열 값은 1989년부터 1993년까지의 사용량 평균 값이다. 보는 것처럼 표의 처음과 마지막 부분은 5년치 데이터가 없기 때문에 비어있는 것을 확인할 수 있다.

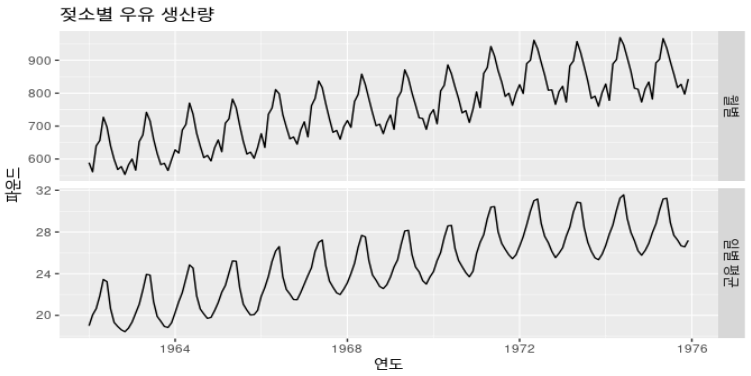
이동평균 평활화가 단기적인 변동을 제거하는데 주로 이용되었다면 차분은 장기적인 변동을 제거하는데 유용하게 사용될 수 있는 기법이다. 차분은 시계열의 현재 시점에서 과저의 인접한 시점의 자료를 차감하는 기법이다. 이렇게 인접한 시점의 자료를 차감하면 시계열이 증감을 확인할 수 있다. 차분 기법을 사용하면 추세 변동을 제거할 수 있지만 단기적인 변동이 증폭되기 때문에 신호가 정확하지 않을 수 있다.



출처 :https://otexts.com/fppkr/stationarity.html

위 표에서 (a)는 구글의 주식가격 추세 그래프다. 이를 차분하게 되면 구글 주식 가격의 일일 변동 그래프가 되는데 이가 (b) 그래프이다. 이처럼 차분은 전체적인 변동성을 확인할 수 있게 해준다.

변수변환은 데이터의 예측이 힘들 때 일부 변수를 변환하거나 조정해 예측하기 쉽게 만드는 기법이다. 변수변환 방식을 사용하면 과거 데이터 패턴을 일관성 있게 만들어서 패턴을 단순하게 만들 수 있다. 이 기법은 간단한 변환으로 큰 효과가 있는 경우가 많지만 기존 척도로 예측값을 얻기 위해서는 다시 역으로 변환을 수행해야 한다.



출처 : https://otexts.com/fppkr/transformations.html

위 그래프는 우유 생산량을 나타내는 그래프이다. 월별 그래프의 경우 매 월마다 날짜 개수가 다르기 때문에 월마다 생산량에 차이가 발생할 수 있다. 이때 월 평균 대신 일 평균으로 변수를 조정해 날짜 개수로 인한 변동성을 제거할 수 있다.

1-2

A/B 검증은 추천 알고리즘이 사용되었을 때와 사용되지 않았을 때의 매출을 비교하기 위한 통계적인 방법론이다. 추천시스템의 성능 지표는 예측된 추천의 정확도나 추천된 상품을 구매했는지 여부 등이 있는데 이러한 지표는 기존 알고리즘과 신규 알고리즘을 비교할 때 주로 사용된다. 하지만 이러한 추천시스템의 성능 지표로 매출 증감분을 확인하기에는 힘든 부분이 있다. 추천시스템의 주요 성능 지표인 추천된 상품 구매여부를 보면, 이 지표는 상품을 구매 생각이 없는 사용자에게 상품을 추천해 구매를 유도할 수도 있지만 상품을 구매할 생각이 있는 사용자에게 상품을 추천해 구매를 유도할 수도 있다. 전자의 경우 구매여부(성능)가 매출과 관련이 있지만 후자의 경우 추천과 상관없이 이미 확정된

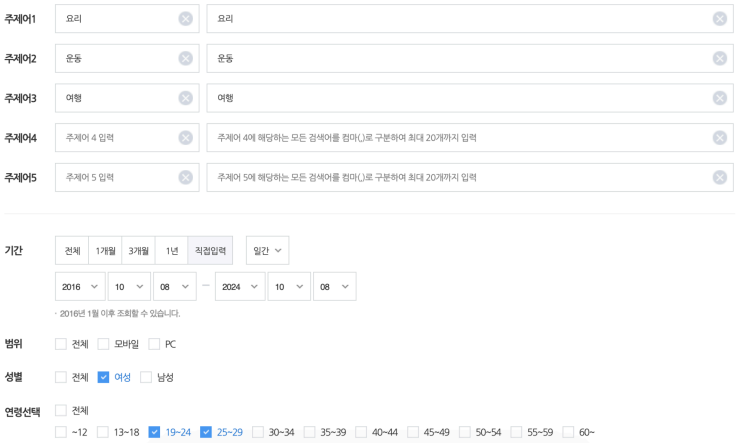
매출이므로 구매여부와 매출이 관련이 없다. 때문에 추천시스템의 성능을 매출로 확인하려면 A/B 검증을 수행해야 한다.

A/B 검증을 수행하려면 먼저 고객을 두 그룹으로 나눈다. 그룹을 A와 B로 나누었을 때 A 그룹에는 추천시스템을 사용해 상품을 추천하고 B 그룹에는 추천 시스템을 사용하지 않는다. 이렇게 일정기간이 지난 후 두 그룹의 매출을 비교하면 된다. 만약에 A 그룹이 B 그룹보다 매출액이 크다면 이는 추천시스템의 성능이 우수하다고 판단할 수 있다.

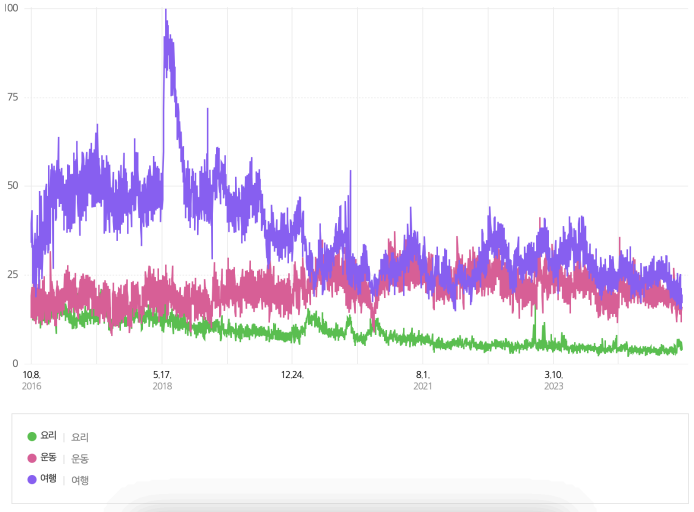
이러한 검증 방법은 절대적인 성능을 비교하기 때문에 효율적이면서도 과학적이다. 하지만 추천시스템의 성능이 우수할 경우 검증을 수행하는 기간동안 추천시스템이 적용되지 않은 사용자에 대한 매출 감소가 발생하기 때문에 기업은 이에 대한 비용을 감수해야 한다는 단점이 있다.

2-1

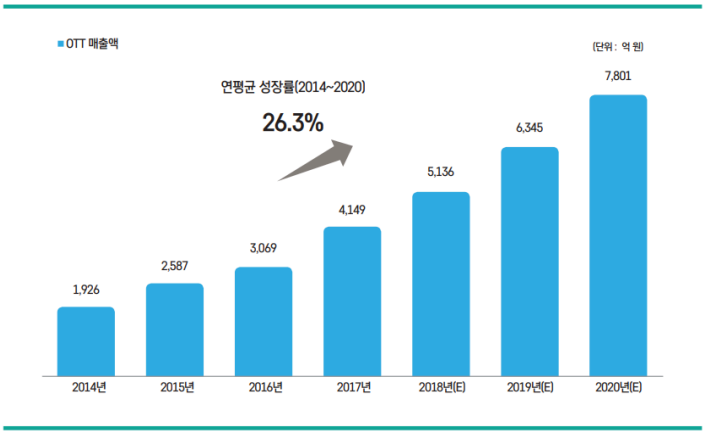
먼저 취미생활에 대한 키워드로 요리, 운동, 여행을 선택했다.



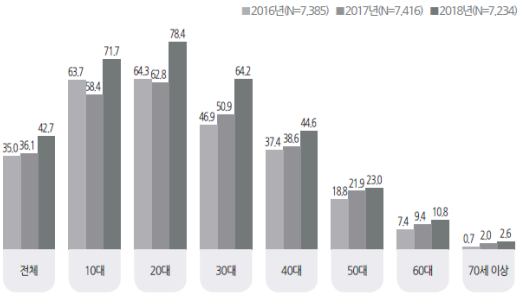
각 키워드 별로 지표를 따로 확인할 예정이므로 주제어를 각 키워드별로 설정한다. 기간은 최대한 많은 데이터를 확인하기 위해 2016년부터 지금까지로 설정하고 성별은 여성, 연령대는 20대로 설정한다. 이렇게 설정하고 데이터를 조회하면 다음과 같다.



전체적으로 여행에 대한 검색이 가장 많은데, 2019년 하반기부터 코로나로 인해 여행에 대한 검색량이 크게 줄어들었다. 하지만 여행 제한이 풀린 이후로도 검색량은 감소하는 것을 볼 수 있다.



출처 : 방송통신위원회 방송시장 경쟁상황 평가 ; 메조미디어(2019), '2019 OTT 서비스 트렌드 리포트'에서 재인용

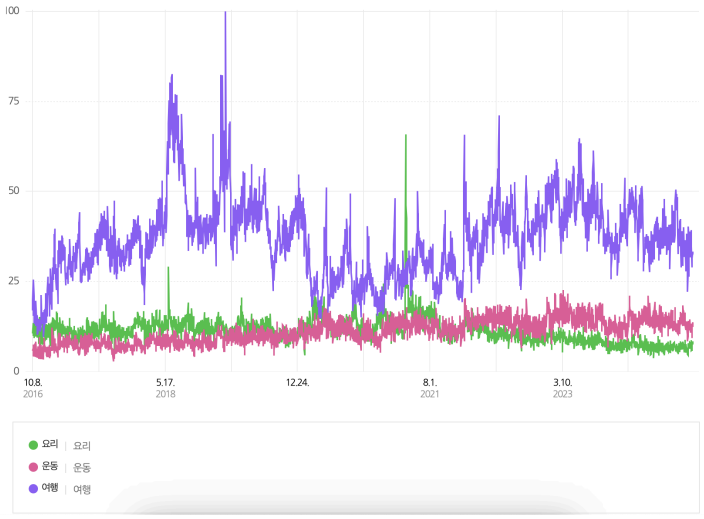


출처 : 방송통신위원회

위 자료를 보면 OTT 서비스가 매년 평균 20% 이상씩 성장하고 있으며 20대에서 큰폭으로 이용률이 증가하고 있는 것을 볼 수 있다. OTT와 같이 집에서 스마트폰으로 즐길 수 있는 취미가 늘어나면서 여행에 대한 관심도가 줄어든 것으로 볼 수 있다.

출처 : 프레시지

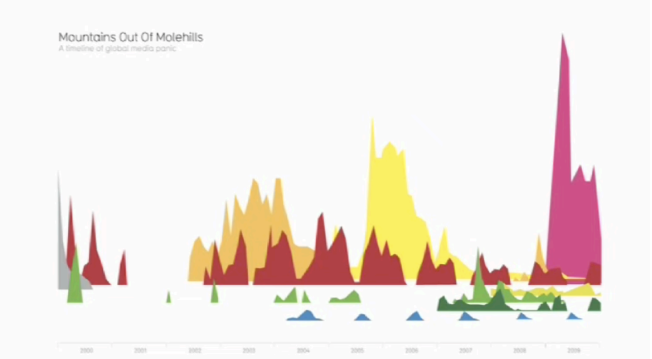
요리의 경우 시기와 상관없이 꾸준하게 줄어들고 있는 것을 볼 수 있는데 위 그래프를 보면 밀키트 시장 급격하게 성장하고 있고 배달 음식이 보편화되면서 요리에 대한 관심도가 자연스럽게 줄어든 것으로 볼 수 있다.



40대 또한 코로나 전에 여행에 대한 검색량이 많았지만 20대와 다른 점은 코로나 이후로도 꾸준하게 여행에 대한 검색량이 꾸준하게 유지되고 있다는 점이다. 위에 OTT 그래프를 보면 20대에 비해 이용률이 높지 않고 낮은 폭으로 이용률이 증가하는 것을 볼 수 있으며 가정을 꾸리고 있는 40대 특성상 야외활동에 관심이 많을 수 밖에 없기 때문에 여행에 대한 검색량이 높을 수 밖에 없다.

요리 또한 코로나때 잠깐 요리에 대한 검색량이 증가하였으나 코로나 이후로 점차 감소하는 것을 볼 수 있다. 밀키트 및 배달음식 이용률이 20대보다 많은 것을 고려했을 때 감소하는 이유 또한 동일한 것으로 볼 수 있다.

3-1



영상 3분 26초 부분을 보면 Mountains Out Of Molehills라는 것을 사용해 연도별로 특정 독감이 얼마나 발생했는지 등을 비교하고 있다. 이 시각화 방법은 연도에 따라 서로 다른 종류의 발생빈도 등을 효과적으로 비교할 수 있기 때문에 연도별, 국가별로 기후위기로 인해 발생한 피해액을 비교해본다면 기후위기가 실제로 어떻게 영향을 미치고 있는지 효과적으로 파악할 수 있을 것이다.

먼저 x축에는 연도를 그대로 두고 y축에는 자연재해로 인한 피해액을 표시한다. z축에는 각 국가들을 나열해 연도마다 각 국가들이 자연재해로 인해 얼마나 많은 비용을 지불하고 있는지 시각화할 수 있다. 이때 z축에 나열할 국가는 기후위기에 취약한 국가와 기후위기에 영향을 많이 받지 않는 국가를 적절하게 선정해 배치한다.

이렇게 시각화했을 때, 시간이 지날수록 기후위기에 취약한 국가의 비용은 증가하면서 기후위기에 영향을 받지 않는 국가의 비용은 크게 변하지 않는다면 기후위기가 발생하고 있다는 것을 증명할 수 있게 된다.

[참고 문헌]

- 빅데이터의 이해와 활용 | 이긍희, 함유근, 김용대, 이준환, 원중호 공저
- The beauty of data visualization | David McCandless Youtube
- Forecasting: Principles and Practice