데이터 통합
2차 보고서

2015122053 이재환
2014131026 한인욱
2017122035 홍장미

<목차>

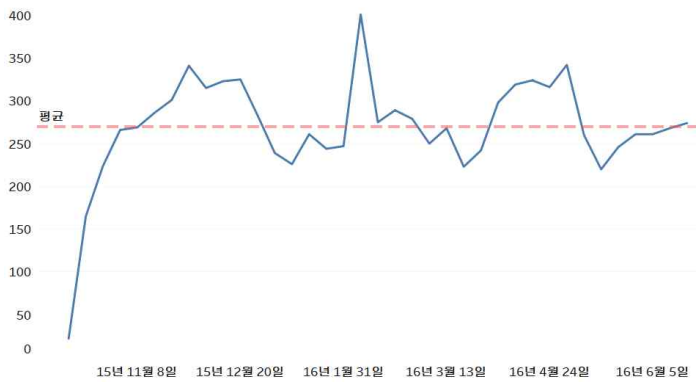
1. 코멘트 반영

지난달 제출했던 1차 보고서의 코멘트에 대해 다음과 같이 반영하였다.

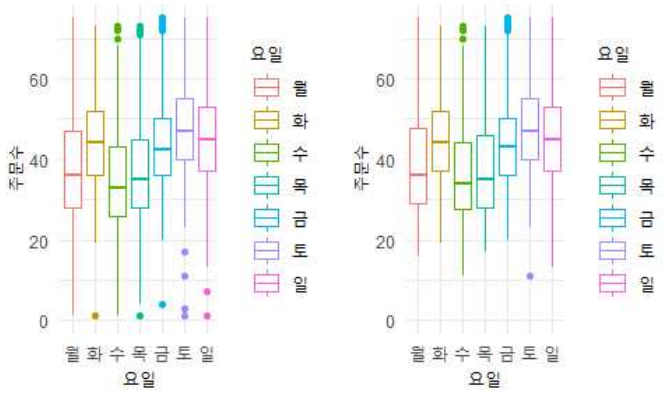
1. 요일별 주문수 데이터의 분산이 크다.

⇒ 분산을 안정화하기 위해 가게가 개점하고 주문수가 안정화될 때까지의 데이터를 제거하였다. [그림 1]에서 우리는 주별 주문수가 전체적인 평균에 도달한 2015년 11월 8일을 안정화에 접어든 시점으로 보았고, 우리는 가게가 안정된 이후의 분석을 원하므로 그 이후부터의 데이터를 사용하기로 하였다.

또한 주문수 10개 이하인 날을 제거하였다. 이는 주문수를 기록하는 포스기의 오류 혹은 일정 시간 휴업으로 인한 것인데, 이는 데이터의 질을 떨어뜨리므로 제거하였다. 기존의 전체 데이터와 초기 1개월, 이상치 제거 이후의 데이터의 박스플롯을 비교해보면 아래와 같이 분산이 감소한 것을 볼 수 있다. ([그림2] 참조)



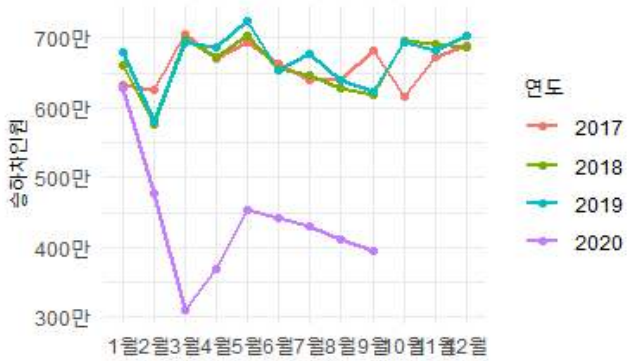
[그림 1] 주문수 시계열 그래프(주별), 점선: 총평균(주별)



[그림 2] 전체 데이터의 요일별 주문수 박스플롯(좌측)과 초기 1개월+이상치 제거 데이터의 박스플롯(우측)

2. 코로나와 관련하여 유동성(mobility)에 대한 변수가 추가되면 좋겠다.

⇒ 대전광역시 일별 지하철 승하차인원 데이터를 추가하였다. 승하차인원을 역별로 모두 합산하여 그날의 유동인구를 나타내는 데이터로 생각하였다. 그림 3에서 2017~2019년의 지하철 승하차인원은 대체로 비슷하다가 코로나가 발발한 2020년 1월 이후는 인원이 절반 미만 수준으로 감소한다. 때문에 유동성(mobility)의 감소를 잘 드러내는 변수라고 판단하여 해당 변수를 데이터셋에 추가하였다.



[그림 3] 대전광역시 일별 지하철 승하차인원

3. 개인정보 보호와 관련하여 비식별화의 필요성이 있을 수 있다.

데이터를 검토한 결과 개인정보가 포함되지 않았으므로 이슈가 발생하지 않았다. 대한민국 정부 산하 관계부처에서 만든 자료[1]에 따르면, 개인정보 비식별화 조치가 필요한 경우는 크게 식별자(개인 혹은 개인과 관련한 사물에 고유하게 부여된 값 혹은 이름) 혹은 속성자(개인과 관련된 정보로서 다른 정보와 쉽게 결합할 경우 특정 개인을 알아볼 수 있는 정보)일 때이다. 우리 조의 "고객수" 변수는 두 가지 케이스에 모두 해당이 되지 않기 때문에 비식별화 없이 그대로 분석을 진행하였다.

1) 개인정보 비식별 조치 가이드라인(비식별 조치 기준 및 지원 관리체계 안내)(2016)-대한민국 정부

2.데이터 수집

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	날짜	요일	주문수	평균기온	최저기온	최고기온	강수량mm	구글검색량	네이버검색량	확진자	사망자	완치자	지하철승하차인원	공휴일	복날	방학여부	도마1동인구	도마2동인구	변동인구	영업가게	유튜브조회수	전국확진자	전국완치자	전국사망자	코로나발발
2	2015-11-08	일	64	13.9	13.4	14.5	15.5	26	NA	0	0	0	NA	0	0	0	17755	20916	18026	45	NA	0	0	0	0
3	2015-11-09	월	31	13.1	11.5	15.2	1.7	27.428571	NA	0	0	0	NA	0	0	0	17753	20913	18025	45	NA	0	0	0	0
4	2015-11-10	화	19	11	8	14.5	NA	28.857143	NA	0	0	0	NA	0	0	0	17751	20910	18024	45	NA	0	0	0	0
5	2015-11-11	수	28	11.9	6.8	18.3	NA	30.285714	NA	0	0	0	NA	0	0	0	17749	20907	18023	45	NA	0	0	0	0
6	2015-11-12	목	24	13.6	8.8	19.4	NA	31.714286	NA	0	0	0	NA	0	0	0	17747	20904	18022	45	NA	0	0	0	0
7	2015-11-13	금	50	12.2	11.1	13.6	30.7	33.142857	NA	0	0	0	NA	0	0	0	17745	20901	18021	45	NA	0	0	0	0
8	2015-11-14	토	53	13.5	12.2	15.7	3.7	34.571429	NA	0	0	0	NA	0	0	0	17743	20898	18020	45	NA	0	0	0	0
9	2015-11-15	일	54	13.5	9.8	17	NA	36	NA	0	0	0	NA	0	0	0	17741	20895	18019	45	NA	0	0	0	0
10	2015-11-16	월	34	11.7	8.4	14.2	26.3	35.142857	NA	0	0	0	NA	0	0	0	17739	20892	18018	45	NA	0	0	0	0
11	2015-11-17	화	37	13.7	12.1	15.1	1.5	34.285714	NA	0	0	0	NA	0	0	0	17737	20889	18017	45	NA	0	0	0	0
12	2015-11-18	수	33	12.1	10.2	13.5	1.9	33.428571	NA	0	0	0	NA	0	0	0	17735	20886	18016	45	NA	0	0	0	0
13	2015-11-19	목	32	10.3	9.5	11.3	1.8	32.571429	NA	0	0	0	NA	0	0	0	17733	20883	18015	45	NA	0	0	0	0
14	2015-11-20	금	47	11.2	8.7	15.1	NA	31.714286	NA	0	0	0	NA	0	0	0	17731	20880	18014	45	NA	0	0	0	0
15	2015-11-21	토	49	10	7.6	13	NA	30.857143	NA	0	0	0	NA	0	0	0	17729	20877	18013	45	NA	0	0	0	0

완성된 데이터셋의 변수 리스트는 [날짜, 요일, 주문수, 평균기온, 최저기온, 최고기온, 강수량mm, 구글검색량, 네이버검색량, 확진자, 사망자, 완치자, 지하철승하차인원, 공휴일, 복날, 방학여부, 도마1동인구, 도마2동인구, 변동인구, 영업가게, 유튜브조회수, 전국확진자, 전국완치자, 전국사망자, 코로나발발]이다. 밑줄친 변수는 여부를 나타내는 0,1로 이루어진 변수이고, 노랗게 칠한 변수들은 1차 보고서의 계획과는 다르게 새로 추가된 변수이다.

- **구글검색량**과 **네이버검색량**은, 각각 구글 트렌드와 네이버 트렌드의 'bhc' 키워드 지수이다.
- **영업가게**는 해당 일자에 주변에서 영업 중인 치킨집 수이다.
- **지하철승하차인원**은 1. 코멘트 반영에서 언급한 유동성 관련 변수로, 대전시 일별 지하철 모든 역의 승·하차인원을 총합산한 것이다. 2017년 1월 1일 데이터부터 수집에 성공하였다.
- **유튜브 조회수**는 신메뉴의 출시로 인해 주문수가 받는 영향을 반영하기 위해 넣은 변수이다. 신메뉴를 리뷰한 유튜브 영상 중에서 출시일과 근접한 영상을 추린 뒤 그중 조회수 상위 5개의 합을 구한 것이다. 합산된 조회수를 신메뉴의 출시일자에 맞추어 데이터로 추가해주었다.

3.발생한 이슈 및 처리방안 모색

1. 인구수, 구글검색량은 각각 월별, 주별 데이터로 다른 형식이었다.
⇒ 인구수와 검색량은 결측치를 등차수열을 이루게 채워넣었다. K-NN이나 기타 방법들을 사용하지 않은 이유는, 인구수나 검색량 변수는 그 추세가 크게 변하지 않는 변수들이라고 판단하였기 때문이다. 두 데이터는 sample 끼리가 독립적인 데이터가 아니고 전, 후 데이터의 영향을 받는 시계열 데이터이기 때문에 그 전과 후의 값을 고려하여 선형으로 채워넣었다.

구글 검색량	구글 검색량
24	24
NA	27
NA	30
NA	33
NA	36
NA	39
NA	42
45	45

[예시]

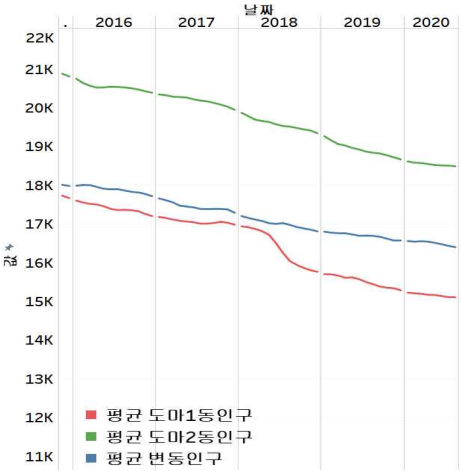
2. 지하철 승하차 인원의 결측치 처리
⇒ 지하철의 승하차의 경우 2015년 10월부터 2016년 12월까지는 모두 NA 처리가 되어있다. 이에 알맞은 결측치를 채우기 위한 모델을 모색하는 것은 추후에 결정할 수 있을 것이다.

3. 유튜브 조회수 처리
⇒ 현재는 신메뉴가 출시된 날들에만 유튜브 조회수 합산값을 붙여넣은 상태이다. 하지만 신메뉴의 출시에 대한 인기는 날이 갈수록 효과가 떨어지는 구조이다. 이를 반영하기 위해 인기가 언제까지 지속되고, 어떤 함수를 그리며 떨어지는지를 조사하고, 그 함수에 따라 유튜브 조회수를 뒤 날짜에 대해 분배할 생각이다.

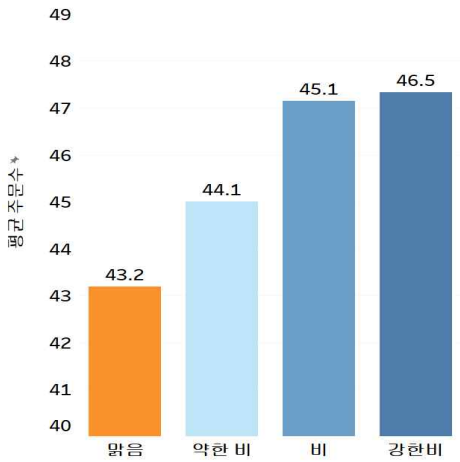
유튜브 조회수	유튜브 조회수
100	40
NA	25
NA	15
NA	10
NA	5
NA	3
NA	2

[예시]

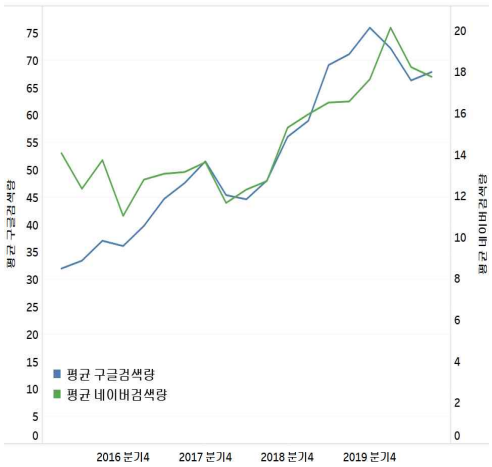
4.데이터 살펴보기



[그림 4] 각 동의 인구수 변화

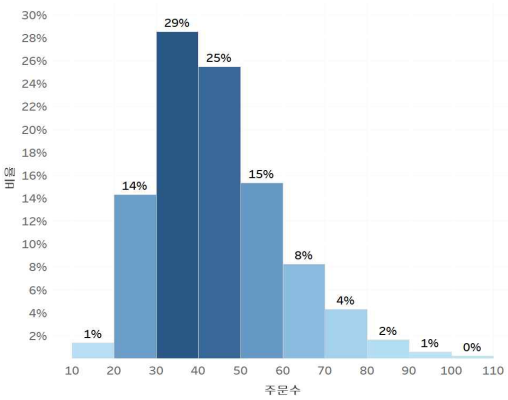


[그림 5] 강수량에 따른 평균 주문수

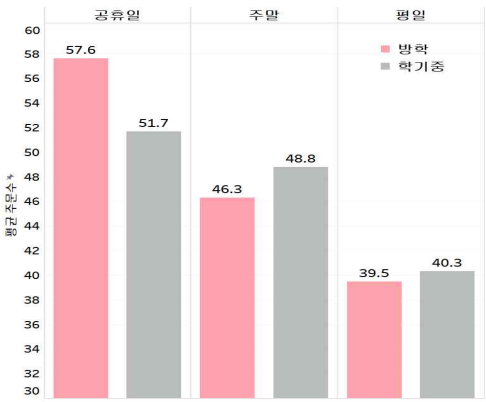


[그림 6] 구글검색량, 네이버검색량

- [그림 4] : 상권에 중요한 역할을 하는 인구수가 꾸준히 감소하고 있음을 볼 수 있다.
- [그림 5] : 강수량이 높을수록 주문수가 많음을 볼 수 있다.
- [그림 6] : 시간이 갈수록 검색량이 많아지고 인터넷의 영향이 커지고 있음을 알 수 있다.



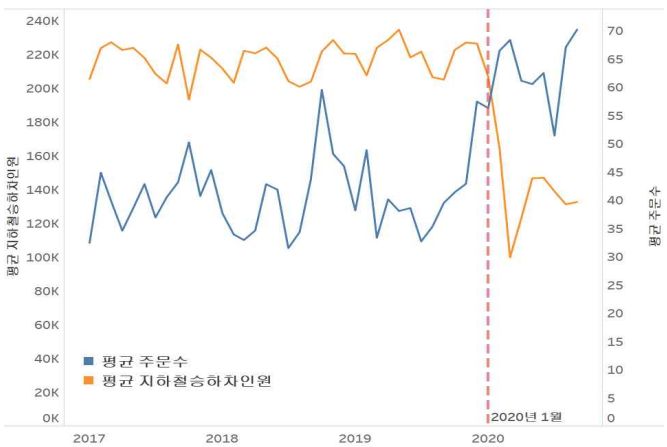
[그림 7] 주문수의 히스토그램



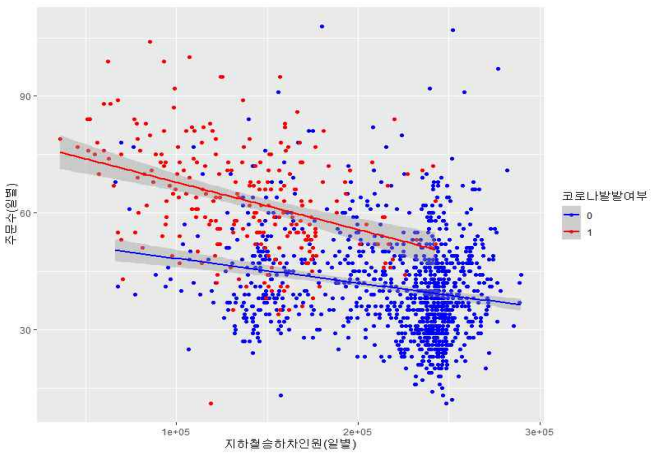
[그림 8] 공휴일/주말/평일에 따른 방학과 학기중의 평균 주문수

- [그림 7] : 대다수의 주문수가 30~50개이며, 분포의 형태가 right skewed 임을 볼 수 있다.
- [그림 8] : 공휴일>주말>평일 순으로 주문량이 많으며 공휴일과 주말에는 방학과 학기 중의 주문수가 차이가 나는 반면 평일에는 차이가 거의 없다.

5.간단한 분석



[그림9] 주문수와 지하철 승하차인원 (점선:코로나 발발)



[그림 10] 코로나에 따른 지하철 승하차인원과 주문수의 산점도

```
Call:
lm(formula = 주문수 ~ 지하철승하차인원, data = data_subway2)

Residuals:
    Min      1Q  Median      3Q     Max
-31.685  -9.229  -1.538   7.228  68.277

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.461e+01  2.035e+00  26.838  < 2e-16 ***
지하철승하차인원 -6.302e-05  9.189e-06  -6.858  1.2e-11 ***
```

[그림 11] 지하철 승하차인원과 주문수(코로나 이전)

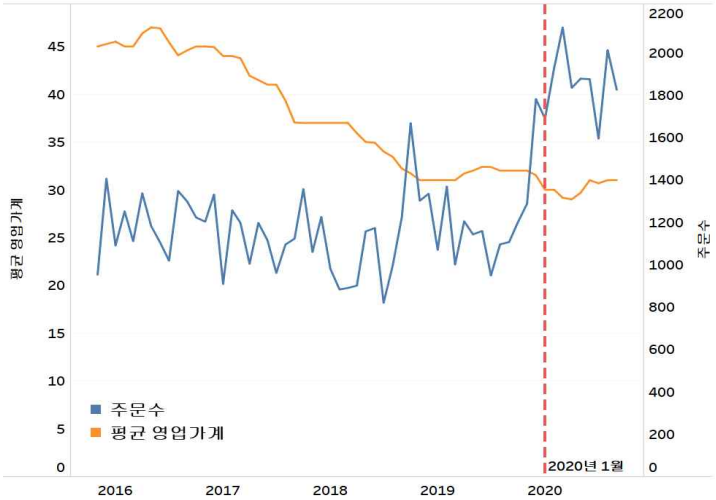
```
Call:
lm(formula = 주문수 ~ 지하철승하차인원, data = data_subway)

Residuals:
    Min      1Q  Median      3Q     Max
-54.576  -7.762  -0.908   8.754  34.368

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      7.997e+01  2.847e+00  28.086  < 2e-16 ***
지하철승하차인원 -1.212e-04  1.972e-05  -6.147 3.17e-09 ***
```

[그림 12] 지하철 승하차인원과 주문수(코로나 이후)

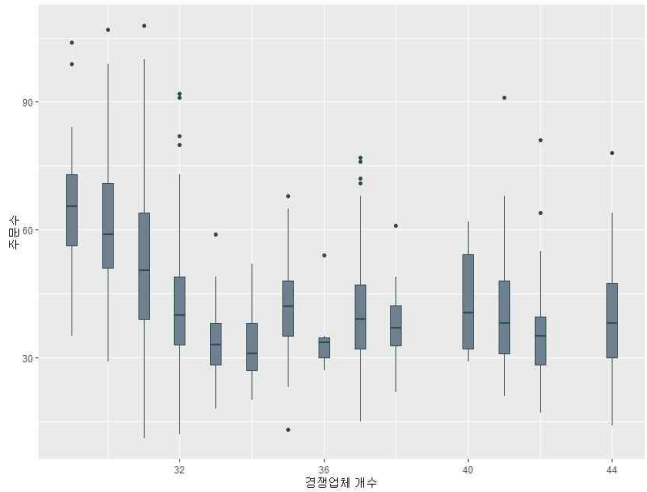
```
Call:
lm(formula = 고객수 ~ 영업가계, data = customer)

Residuals:
    Min      1Q  Median      3Q     Max
-37.659 -10.226  -1.226   8.574  59.341

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 72.42477    2.13059   33.99   <2e-16 ***
영업가계     -0.76664    0.05681  -13.49   <2e-16 ***
```

[그림 13] 경쟁업체 개수와 고객수

[그림 9]에서 볼 수 있듯이 코로나 발발 이후 평균 지하철 승하차인원은 급감하고 평균 주문수는 급격히 상승했다. 이는 유동성(mobility)이 감소함에 따라 배달업체의 매출이 증가할 것이라는 1차 보고서에서의 예측과 맞아떨어진다. [그림 10]을 통해 코로나 이전과 이후 모두 지하철 승하차인원이 감소함에 따라 주문량이 늘어남을 볼 수 있으며, [그림 11], [그림 12]에서 이 관계가 유의하다는 것이 확인 가능하다. 또한 코로나 발생 이전보다 이후의 영향력(계수의 절댓값)이 더 크다.



[그림 14] 평균 주문수 & 평균 영업 가게 수의 시계열 그래프



[그림 15] 경쟁업체 개수에 따른 주문수의 박스플롯

[그림 14]를 통해 시간이 흐름에 따라 주변 경쟁 영업체 수는 지속적으로 감소하고, 주문수는 변동이 있다가 2020년에 들어 크게 상승함을 볼 수 있다. 또한 경쟁업체 수가 줄어듦에 따라 전체적인 평균 주문수는 늘어나는 양상([그림 15])도 확인할 수 있고, 이 관계 또한 유의함을 [그림 13]을 통해 확인 가능하다.

6.결론 및 계획

첫 보고서에서 계획했던 데이터를 모두 통합하고, 시각화를 통해서 많은 정보를 얻을 수 있었다. 하지만 그 과정에서 결측치 문제가 지하철승하차인원, 네이버검색량 등의 변수에서 발생하였다. 그리고 이 결측치를 채우기 위해 각 변수에 맞는 적절한 방법들을 찾아야 할 것이다. 인구수나 구글 검색량은 전, 후 시점의 데이터에 영향을 크게 받고 트렌드가 존재해서 등차수열로 채울 수 있었지만, 지하철 승하차인원처럼 다른 특성을 가지는 데이터에 대해서는 아직 결측치를 채울 적절한 방법을 찾지 못한 상황이다. 이러한 변수들에 대해서는 수업에서 배운 ignorable missing mechanism을 활용한 모델이나 Nearest neighborhood imputation 등의 방법론들을 모색할 계획이다. 또한 앞에서도 언급했듯이 떨어지는 인기를 반영하며 유튜브 조회수를 알맞게 배분해야 할 것이다. 이렇게 발생한 이슈들을 처리하고, 개인 프로젝트를 진행하며 더 필요한 데이터는 원본 데이터에 통합해 나가면서 더욱 big(ger)한 데이터를 만들어 나갈 계획이다.