데이터 통합
3차 보고서

2015122053 이재환
2014131026 한인욱
2017122035 홍장미

<목차>

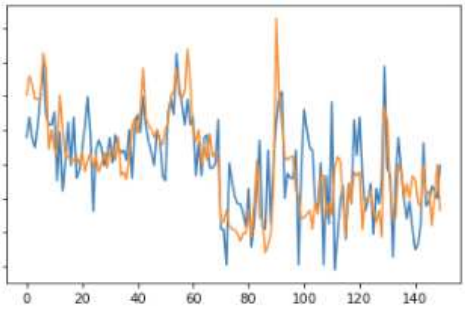
1. 코멘트 반영 & Data Integration

1) 네이버, 구글 검색량 결측치

구글 검색량은 전국과 대전 데이터가 있는 대신에 일요일 값만 있는 주별 데이터였고, 네이버 검색량은 전국 데이터인 대신에 일별 데이터였다. 결과적으로 우리는 브랜드에 대한 관심도를 반영하는 변수로서 네이버 검색량을 채택하였다. 이에 대한 근거와 과정은 다음과 같다.

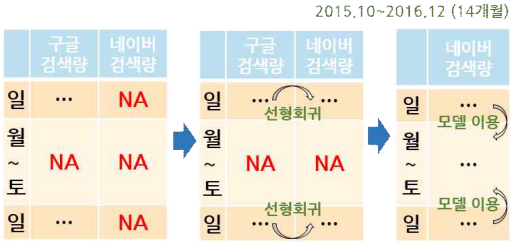
①전국 데이터 → 대전 데이터

구글 검색량의 시도표(그림1)를 보면, 전국의 검색 경향과 대전의 검색 경향이 유사하다. 이는 전국 검색량 데이터를 사용해도 대전의 트렌드를 어느 정도 알 수 있음을 시사한다. 또한 네이버 검색량과 구글 검색량은 높은 상관관계(r=0.63)를 가졌다. 따라서 전국 데이터인 네이버 검색량을 사용해도 괜찮겠다고 판단하였다.



▲ (그림1) 구글 검색량(전국, 대전)

②NA Imputation



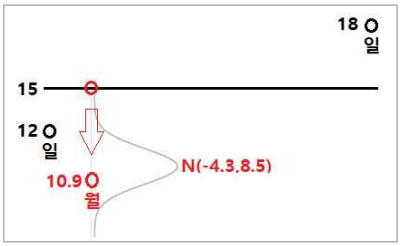
▲ (그림2) imputation 개요

```
lm(formula = 네이버검색량 ~ 구글검색량, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-6.6025 -2.1322 -0.4989  1.7955 10.6913

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.46616    0.63198   14.98   <2e-16 ***
구글검색량   0.14302    0.01146   12.48   <2e-16 ***
```

▲(그림3) 회귀분석 결과



▲(그림4) 모델 가정

네이버 검색량 변수는 전체 5년 중 앞 14개월이 결측치이다. 이를 imputation 하기 위해, 깊은 관련이 있었던 구글 검색량으로 네이버 검색량의 일요일 값을 선형회귀를 이용해 채워 넣었다(그림3). 그리고 일요일이 채워진 네이버 검색량 변수 내에서, 아래와 같은 두 가지를 관찰 가능했다.

① 각 요일의 검색량은 인접한 두 일요일의 검색량 평균의 영향을 받는다.

② 요일별로 다른 검색량을 가진다.

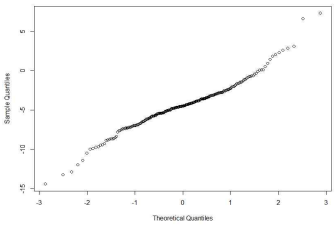
	Mean	Variance
월	-4.286531	8.475655
화	-4.094882	8.684276
수	-4.310949	7.747393
목	-4.457831	6.466559
금	-1.157331	6.845821
토	1.246372	7.177955

▲(그림5)요일별 평균,분산

이에 따라 월~토 의 결측값을 채우기 위한 모델을

$$Y_{A\text{요일 검색량}} = \frac{(Y_{\text{전주일요일}} + Y_{\text{다음주일요일}})}{2} + \epsilon_{A\text{요일}}\ ,\ (\epsilon_{A\text{요일}} \sim N(\mu_A, \sigma_A))$$

로 결정하였다. 즉, '인접한 두 일요일 사이 A요일의 검색량은 두 일요일 값의 평균에 A요일의 error를 더한 것' 이라는 모델이다(그림4). 이때 요일별 error는 전체 네이버 검색량 데이터에서 요일별로 전, 후 일요일 평균에 대비한 편차의 mean과 variance를 계산해(그림5) μ_A, σ_A 를 추정하였다. 그리고 요일별 error가 Normal인지 검정하기 위해 네이버 검색량의 요일별 에러를 qqplot으로 그려보았을 때 normal을 따름을 알 수 있었다(그림6). 이 모델을 근거로, 일요일 사이의 평일(NA)을 채워 넣었다.



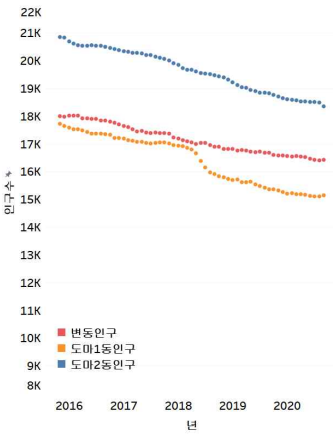
▲(그림6) 목요일 qqplot

2) 인구수

우리가 얻은 데이터는 월별 데이터 인구수였다. 이를 일별로 채우기 위해서 등차수열을 생각했고, 그와 더불어 구조를 알아보고자 분산을 추정해 보았다. 별도 데이터에서 대전시가 제공한 전입, 전출자 비율을 보면 약 15:16이었다. 만일 각 인구수의 변화가 전입:전출 비율대로 변화한다고 가정한다면, 인구수가 1 줄어드는 것은 15명이 대전으로 오고, 16명이 대전을 나갔다고 생각할 수 있

다. 우리는 전입, 전출이 위의 비율대로 이루어진다고 가정했다. X_i=개인의 인구이동이라 하면 전입인 경우 X_i=1, 전출인 경우는 X_i=-1이 된다. 그에 따라 $P(X_i=1)=\frac{15}{31}$, $P(X_i=-1)=\frac{16}{31}$, $E(X_i)=-\frac{1}{31}$, $V(X_i)=E(X_i^2)-E(X_i)^2\approx 1$ 즉 약 1이 된다. 즉 일별 인구이동이 N 명이면 $\sum_{i=1}^{N}X_i$ 값이 일별 인구수의 변화가 된다.

이때 하루에 평균적으로 10명의 인구가 감소했다면 N=310의 인구이동이 있었다고 생각할 수 있다, 이에 따라 일별 인구수 변화의 분산은 각 개인이 독립이라고 가정하면 $V(\sum_{i=1}^{310}X_i)\approx$약 310이다. 즉 일별 분산이 인구 변화 추세에 비해 매우 크다고 할 수 있다. 우리는 인구의 전체적인 경향을 보고 싶으므로 분산을 반영하기보다는 등차수열로 채우기로 하였다.



▲(그림7) 월별 인구수 그래프

3) 대전시 승하차 인원

대전시 지하철 승하차 인원은 전체 날짜 중 2015년 10월부터 2015년 12월 31까지가 결측치였다. 그렇기에 이 결측치를 CIA모델을 가정하여 채워 넣기로 했다. 때문에 관련 데이터를 모색하던 도중 2015년 10월 ~ 2016년 6월까지 대전시 일별 버스 승하차 인원 데이터를 확보했다. 주문수(Y), 버스승하차 인원(X), 지하철 승하차(Z), 나머지 관련 변수(W)이라고 가정한다면, 지하철 승하차인원과 버스 승하차인원 사이에는 높은 correlation(dependency)가 관측됨을 확인하였다. 즉 $f(Y|X,Z,W)\simeq f(Y|Z,W)$ 일 것이다. 이에 따라 지하철 승하차 인원만 사용하고, 버스 승하차 인원은 지하철 승하차 인원의 NA를 채우는 데에만 사용하였다.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	9.281e+04	4.593e+03	20.209	< 2e-16	***
버스승하차인원	9.698e-02	1.306e-02	7.423	4.96e-12	***
주말여부	-2.726e+04	1.728e+03	-15.777	< 2e-16	***
방학여부	-3.269e+03	1.597e+03	-2.048	0.0421	*
강우여부	-3.985e+03	1.655e+03	-2.408	0.0171	*
공휴일	-3.270e+04	3.421e+03	-9.558	< 2e-16	***

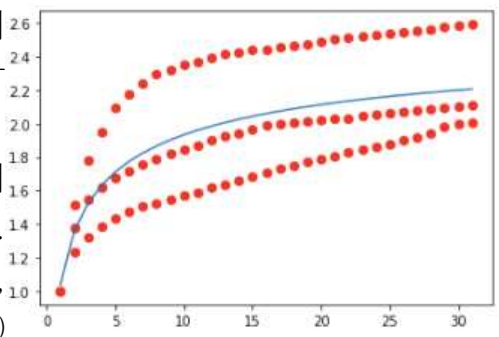
▲ (그림8) 지하철 승하차인원을 예측하기 위해 사용된 관련 변수들

NA를 채우기 위해 하나의 데이터를 일별로 두 개로 분리하여 나눈 뒤(2015.10~2015.12 , 2016.1~2016.6) CIA approach를 사용하여 $\hat{Z}\sim X,W$ 를 regression의 예측값으로 채워 넣었다. 실제 관련 변수들을 넣어 모델을 fitting시켜 AIC stepwise selection을 한 결과 그림 8과 같은 계수들을 최종적으로 사용하여 지하철 탑승 인원의 결측값을 채워 넣었다.

4) 유튜브 조회수 결측치

유튜브 조회수 변수는 신메뉴 출시의 영향을 알아보기 위해 출시일에 근접한 신메뉴 리뷰 유튜브 영상들의 누적 조회수 합산값을 붙여넣은 것이다. 즉 출시일 이외의 날들은 결측치이다. 리뷰 유튜브 영상들의 경우 일별 조회수 데이터를 유튜브가 제공하지 않아서, 우리가 직접 모델을 만들어 채워 넣어야 했다.

누적 조회수를 일별로 분배하기 위한 모델을 고안하기 위해, 일별 조회수가 존재하는 동영상 조회수 데이터를 수집하였다.[1] 그 이후 동영상마다 조회수 scale이 다르므로, 첫날의 조회수가 1이 되게 scaling 하였다. 그리고 누적 조회수 추이에 대해 근사하는 2차 로그식을 Least Square Method를 이용하여 구하였다. 그 모델은 $-0.04\log(x)^2+0.49\log(x)+1.04$ = x일 이후의 누적 조회



▲(그림9) 예시 3개 동영상의 조회수와 우리의 모델식(파란색)

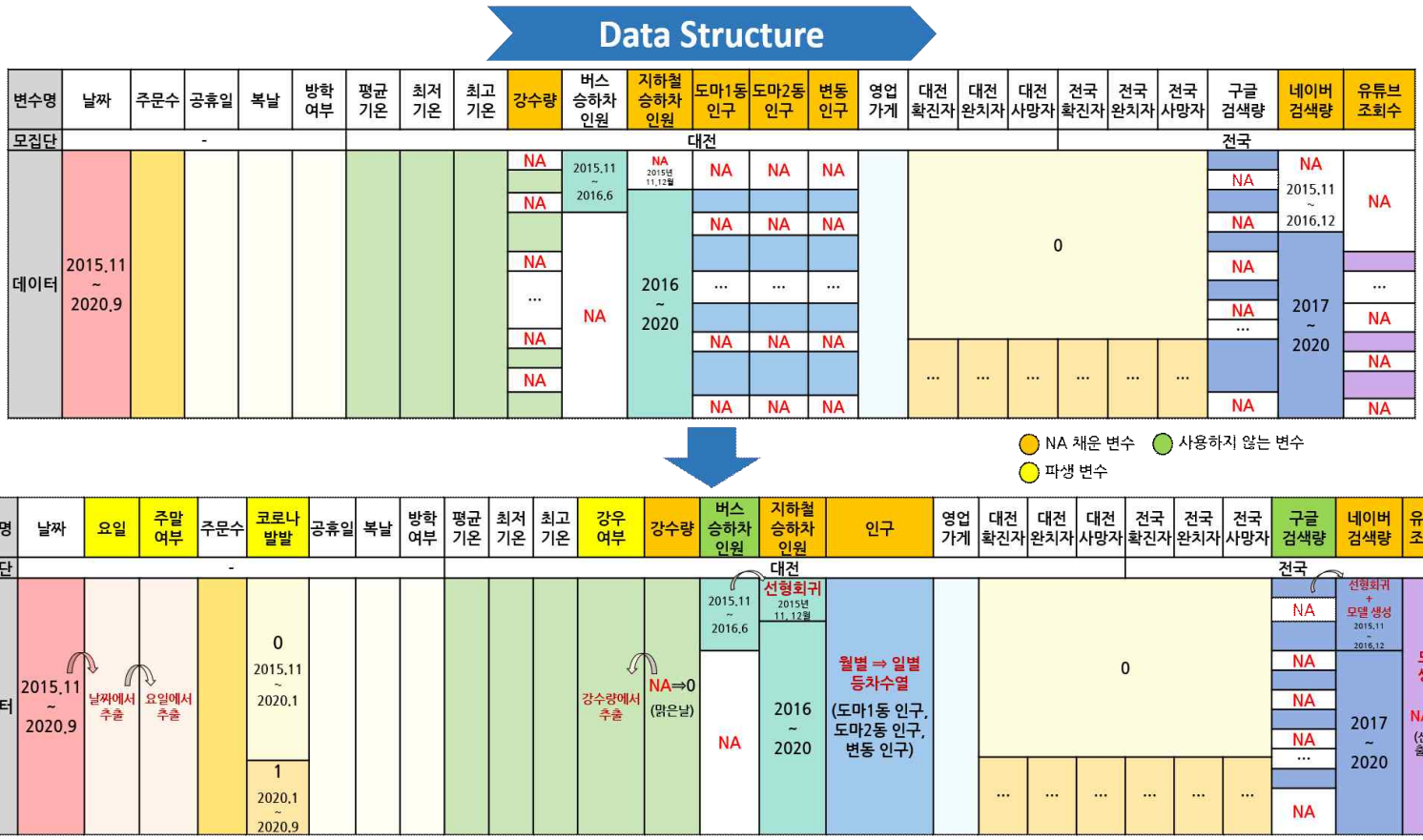
1) 유튜브는 일별 조회수를 제공하지 않기 때문에 다음과 같은 별도의 사이트를 참고했다: http://stats.informational.xyz/index.php

수 가 된다. 이 모델을 이용해 유튜브 총 조회수를 이 근사식의 증가 비율에 맞게 일별로 분배하였고, 30일 이후 나머지 값은 신메뉴의 영향이 없으리라 판단해 0으로 채웠다.

5) 강수량

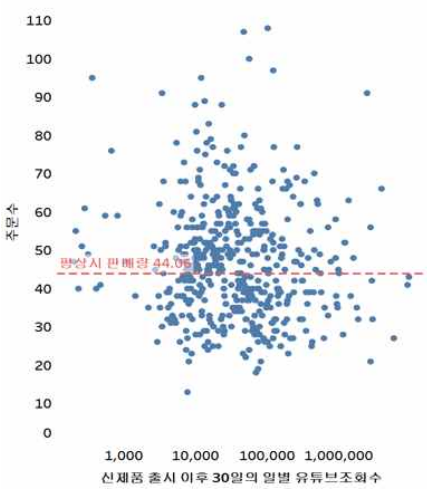
원본 데이터에서 강수량은 0과 NA값이 혼재하여 있었다. 기상청에서 강수량을 측정하는 AWS 장비는 NA는 비가 아예 오지 않은 경우, 0은 매우 미량의 비가 온 경우로 기록한다. 때문에 NA값을 단순히 0으로 채워 넣으면 원본 데이터를 훼손할 수도 있기에 새로운 column(강수여부)를 추가하여 0은 비가 오지 않은 경우(원본 데이터의 NA), 1은 비가 온 경우(원본 데이터의 0이상의 값)로 정의하여 원본 데이터의 정보를 보존하였다.

2. 데이터 최종 형태



※1~2 차 보고서의 결과와 더불어 3차 보고서 때의 과정을 모두 합하면 위와 같은 데이터 형태가 나온다.

3.EDA

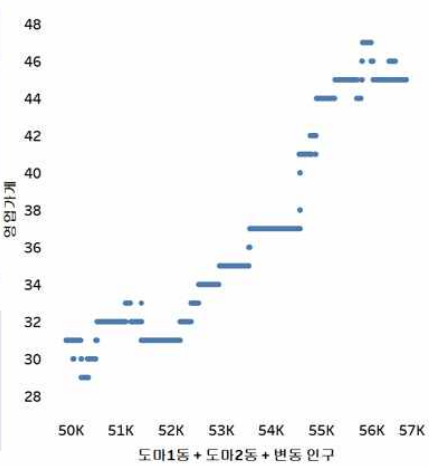


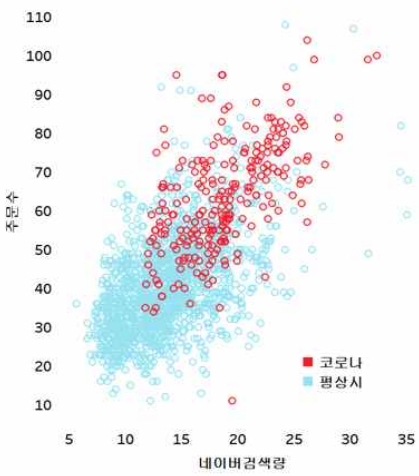
유튜브 조회수(신제품 출시)와 주문수

- 조회수가 신제품별로 크게 차이남을 알 수 있다.
- 신제품이 출시되었을 때와, 주문수의 관계는 크게 없어보인다.
- 주된 소비자층이 기존 메뉴를 더 선호하거나, 신제품에 큰 관심이 없어보인다.

배달지역 인구수와 영업가게수

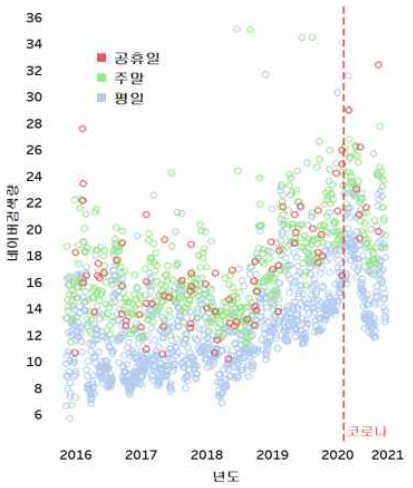
- 인구와 영업가게가 선형관계가 있어 보인다.
- 치킨집 1개마다 운영 가능한 최소 인구수가 존재해, 이를 넘으면 이익이 나지않아 폐업하는 것으로 보인다.





네이버검색량과 주문수

- 검색량과 주문수가 어느정도 양의 상관관계가 있어보인다.
- 코로나때에 검색량과 주문수 모두 늘어났다.
- 언택트 시대에 맞추어 사람들의 인터넷 사용량이 더 많아진 것으로 보인다.

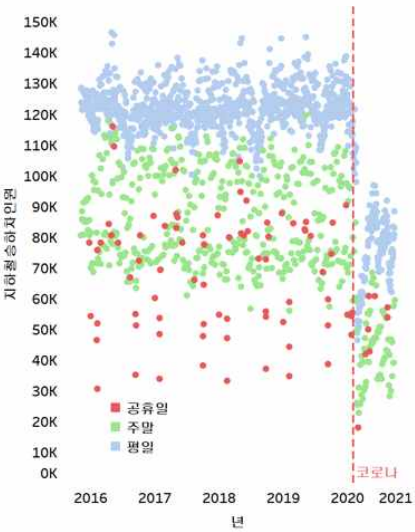


일별 네이버검색량의 분포

- 공휴일,주말이 검색량이 더 많았다.
- 2018~2020 년의 검색량이 증가함을 볼 수 있었는데, 이는 해당년도 사이의 브랜드 파워가 증가하였음을 나타낸다(실제로 브랜드 이미지 자료에서 10위권 -> 5위권으로 상승).

지하철승하차인원과 주문수의 변화

- 코로나 이후 지하철 승하차 인원은 급감했고, 주문량은 급증했다.
- 코로나로 인해 유동성이 감소하고, 이로 인해 배달 주문량이 급증했다고 볼 수 있다.



일별 지하철승하차인원의 분포

- 평일,주말,공휴일 순으로 승하차 인원이 많다.
- 공휴일은 열차의 휴무, 학교 및 회사의 휴무로 인해 제일 적어진 것으로 보인다.
- 코로나 이후의 승객수는 필수불가결한 일 때문에 이용하는 승객수라고 생각된다.
- 코로나 이전 - 이후의 승객수 = 자가용 이용자 혹은 놀러 다니는 사람들이라고 볼 수 있을 것이다.

4. 결론

EDA를 통해 최종 데이터에 관한 여러 가지 intuition을 얻을 수 있었다. 그럼에도 불구하고 해당 데이터를 분석 및 다른 목적으로 사용할 때 몇 가지 유의할 점들이 있을 것이다.

▷ 전국과 대전의 네이버 검색 경향이 비슷하긴 하나 분산이 다를 것이다.
▷ 인구수의 결측치를 채울 때 분산이 너무 커질 거라 생각해 단순 경향만 반영한 등차수열을 사용했는데, 분산을 고려하지 않았기 때문에 이 데이터를 이용하면 error가 있으므로 주의해야 할 것이다.
▷ 결측치를 Regression 으로 채울 때 CIA 가정을 사용했는데 이 가정이 틀릴 경우 NA를 채운 값은 error 가 클 것이다.
▷ 대전의 일별 코로나 확진자,사망자 수가 적어 데이터가 너무 Sparse 했다. 그리고 코로나는 인터넷, 뉴스 등을 통해 전국적으로 그 정보가 실시간으로 퍼져나가므로 전국 데이터와 같이 사용해도 된다고 판단하여서, 전국 데이터도 붙여넣었다. 이때 대전과 전국 코로나 데이터가 경향성이 다를 수 있으므로 유의해야 할 것이다.

위의 주의사항과 각 데이터의 특성을 염두에 두면서 통합된 데이터를 개인 프로젝트에 사용하여 개인별 보고서에 알맞은 분석 결과를 도출해야 할 것이다.