2022-1 빅데이터 분석 하계 경진대회 보고서

팀장 통계학전공 5526355 손명균

팀원 통계학전공 5526463 이호정

목 차

I. 데이터 탐색 및 전처리 과정

1.1 데이터 설명 및 선정 이유

- 제공받은 데이터(이하 데이터1)의 경우, 점포라인(https://www.jumpoline.com/)에서 2018년부터 2022년까지 거래된 데이터로, 37,400개의 행과 8개의 열로 이루어진 파일이다.
- 상점 매매에 대한 권리금과 지역별 전출·전입 인구 간 영향이 있는지 파악하기 위하여 KOSIS 국가통계포털(https://kosis.kr)에서 데이터(이하 데이터2)를 수집하였으며, 1,084개의 행과 55개의 열로 이루어져 있다.
- 권리금과 지역별 대학교의 개수 간 영향이 있는지 파악하기 위하여 KOSIS 국가통계포털(https://kosis.kr)에서 데이터(이하 데이터3)를 수집하였으며, 153개의 행과 4개의 열로 이루어져 있다.
- 지역별 지하철과 상점 매매에 대한 권리금과 거리 간 영향이 있는지 파악하기 위하여 산림 빅데이터 거래소(https://www.bigdata-forest.kr)에서 데이터(이하 데이터4)를 수집였으며, 1,015개의 행과 40개의 열로 이루어져 있다.
- Covid-19의 확산으로 인하여 상점 매매에 어떤 영향을 미치는가에 대하여 알아보기 위하여 코로나바이러스감염증-19(http://ncov.mohw.go.kr)에서 데이터(이하 데이터5)를 수집 하였다.

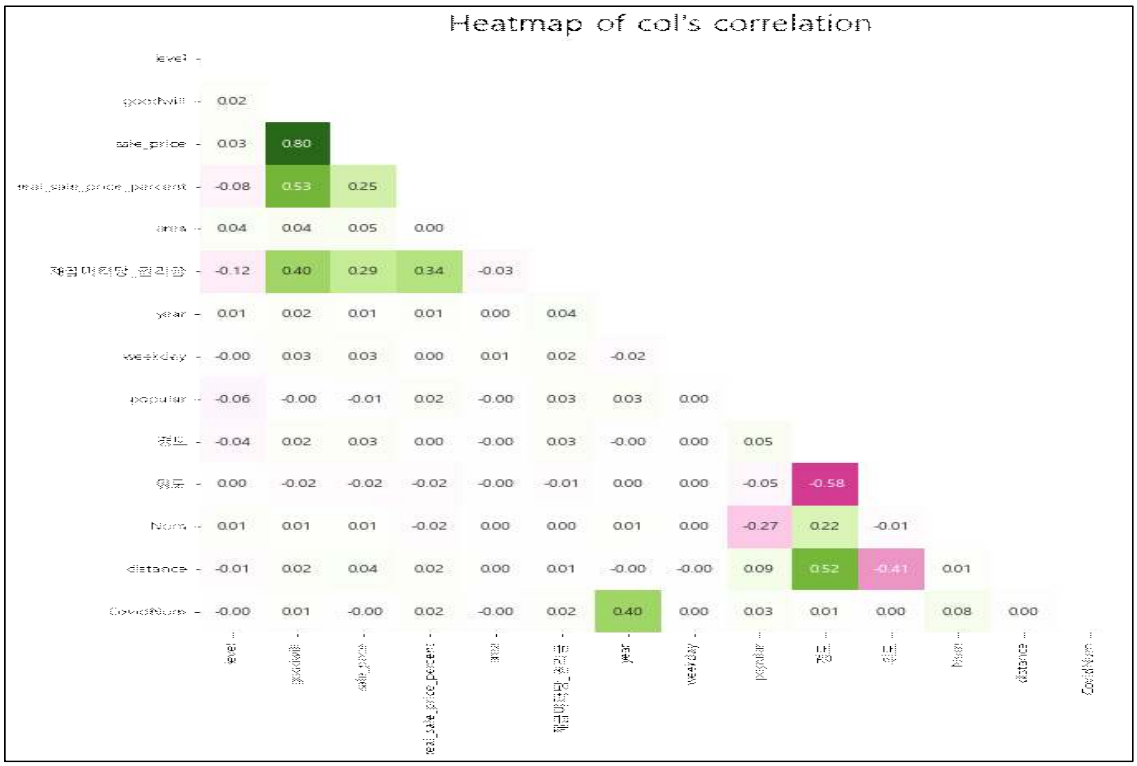
1.2 데이터 전처리

- 데이터1의 'store_type' 열에 결측치 존재하는 데이터 1개 삭제하였다.
- 데이터1의 'store_type' 열에서 점포라인에서 제공하는 소분류 기준을 대분류 기준으로 바꾸었다.
- 데이터1의 'area' 열에서 단위를 뜻하는 m^2을 삭제한 뒤, 제곱미터당 권리금을 소수점 2자리까지 나타내었다.
- 데이터1의 'address'열에서 스페이스를 기준으로 잘라, 지역 값을 '지역', 시군구 값을 '시' 열로 생성해 주었다. 그 후, '지역', '시'를 합친 '지역_시' 열을 생성하였다.
- 데이터1의 '지역' 열을 광역시와 특별시로 분류해 주었다.
- 데이터1의 모든 열이 같을 경우, 삭제하였다.
- 데이터1의 이상치 층을 직접 추적하여 해당하는 값으로 변경해주었다.
- 데이터1의 'level' 열에서 지하층수는 층수 앞에 -표시를 해주었고, 단위를 나타내는 '층'을 삭제하였다.
- 데이터2의 '행정구역(시군구)별' 열에서 월별, 지역별 (전체 전입 - 전체 전출)에 해당하는 열을 생성하였다.
- 데이터1의 'address'열을 Geo-Coding을 활용하여 좌표를 생성한 뒤, 데이터4의

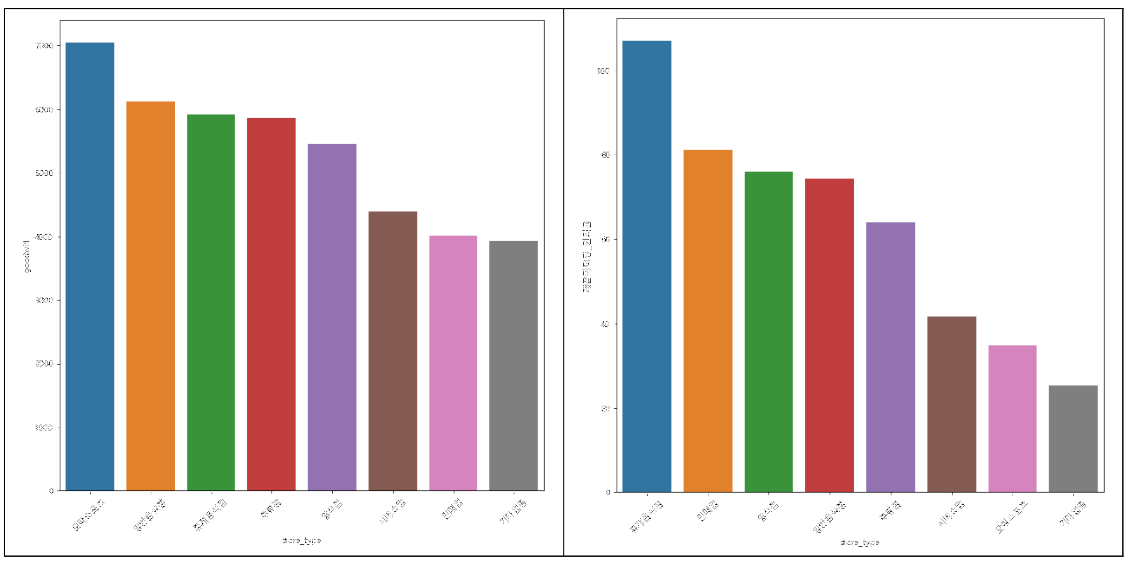
좌표와 최소 거리를 계산하여 'distance'열을 생성하였다.

- 데이터1,2,3,4,5를 모두 합친다. 이때, 만약 Covid-19가 발생하기 전일 경우, 0으로 입력하였다.

1.3 데이터 탐색 및 시각화

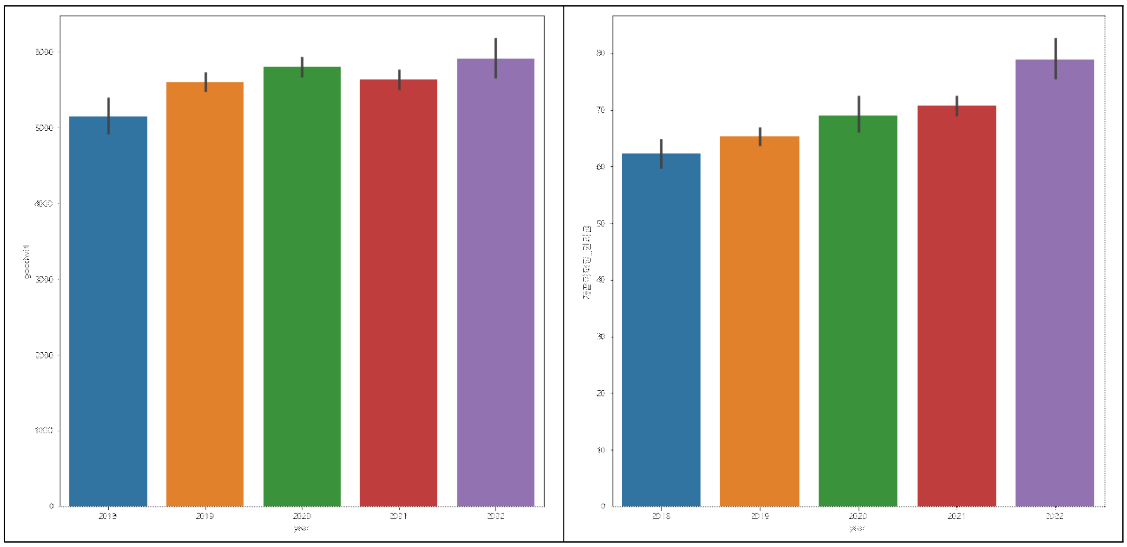


- 수치형 변수별 상관계수를 확인한 결과, 'goodwill'과 'sale_price'간의 상관계수가 0.80로 높다고 볼 수 있다.

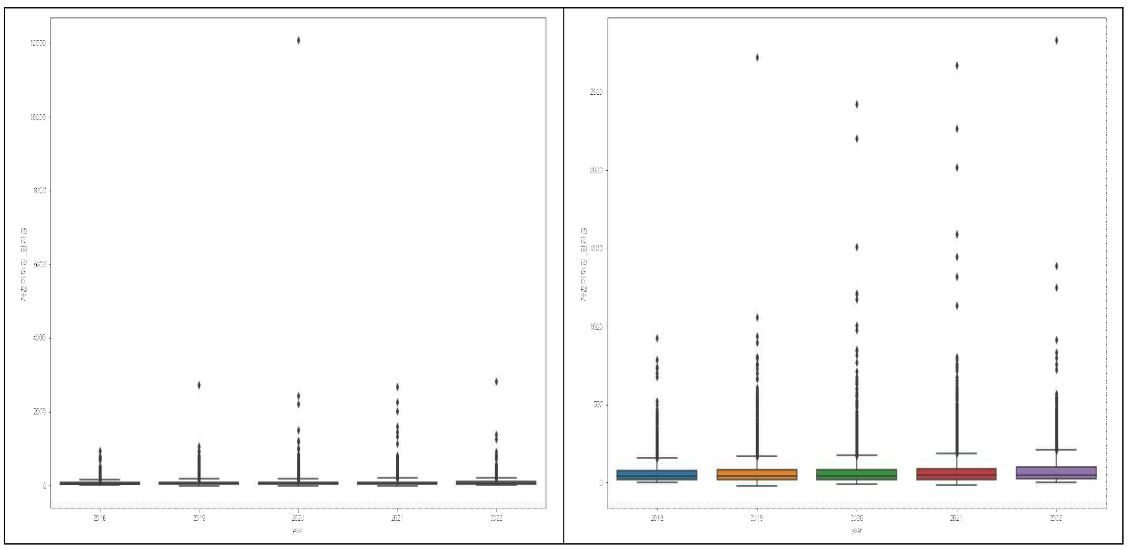


- 'store_type'에 따른 'goodwill'과 '제곱미터당_권리금'을 나타낸 막대그래프이다.

- ‘goodwill’ 확인 결과, 상위 3개 업종이 오락 스포츠, 일반음식점, 휴게음식점인 것에 반해, ‘제곱미터당_권리금’ 확인 결과, 상위 3개 업종이 휴게음식점, 판매업, 일식점으로 나타났다.
- 이를 통해 오락 스포츠의 경우, 제곱미터당_권리금이 낮음에도 ‘goodwill’자체가 높은 이유로, ‘area’가 넓기 때문이라고 예측할 수 있다.



- ‘year’에 따른 ‘goodwill’과 ‘제곱미터당_권리금’을 나타낸 막대그래프이다.
- ‘goodwill’ 확인 결과, 제곱미터당_권리금은 연도가 지날수록 상승하는 것에 반해, 2020년보다 2021년의 ‘goodwill’의 값이 더 감소하는 것을 알 수 있다.
- 이를 통해, 2020년에서 2021년 사이 ‘goodwill’에 영향을 줄 수 있는 어떠한 일이 있었다고 예상할 수 있다.

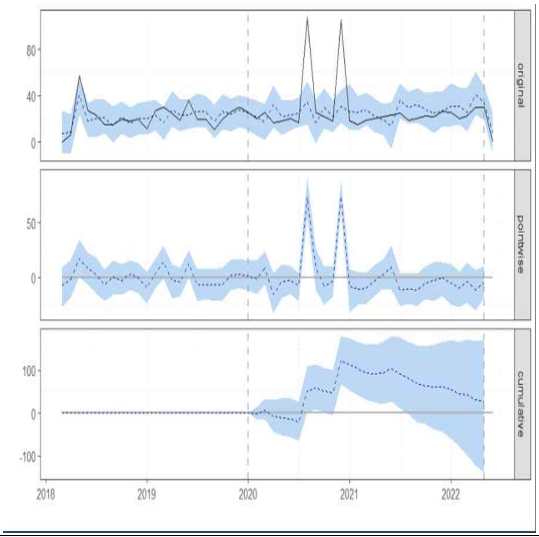


- ‘year’에 따른 ‘제곱미터당_권리금’을 나타낸 상자그림이다.
- ‘제곱미터당_권리금’에서 2020년에 12000이 넘는 값을 제거 후, 상자그림을 다시 그렸을 때, 상자그림의 IQR이 더 뚜렷해지는 것을 알 수 있다.

II. 분석과정

2.1 과제1 코로나의 영향 파악

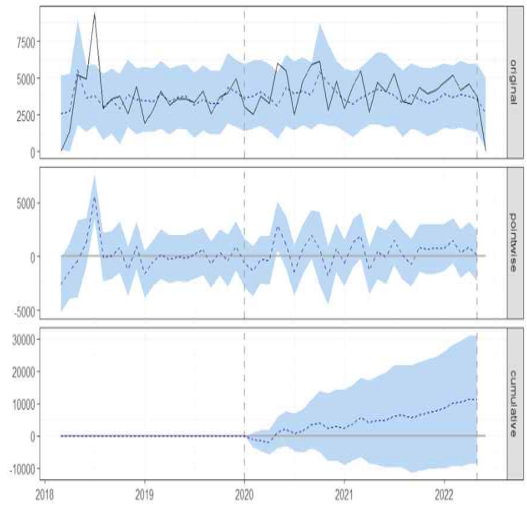
업종별 제곱미터당 권리금




```
Posterior inference {CausalImpact}

                          Average        Cumulative
Actual                    28             778
Prediction (s.d.)         27 (2.8)       751 (78.9)
95% CI                    [22, 33]       [610, 917]

Absolute effect (s.d.)    0.98 (2.8)     27.52 (78.9)
95% CI                    [-4.9, 6]      [-138.3, 169]

Relative effect (s.d.)    3.7% (11%)     3.7% (11%)
95% CI                    [-18%, 22%]    [-18%, 22%]

Posterior tail-area probability p:   0.348
Posterior prob. of a causal effect:  65%

For more details, type: summary(impact, "report")
```

- 코로나 이후 2022년 6월까지의 업종별 제곱미터당 권리금의 평균값은 약 28만원 이다. 코로나가 발생하지 않았다면 평균 27만원 으로 예측된다. 업종별 제곱미터당 권리금 예측의 95% 구간은 [22, 33]이다. 관측된 반응에서의 예측을 빼면 코로나가 제곱미터당 권리금에 미치는 인과 관계의 추정치가 산출되는데, 이 각 시점에서의 개입효과 는 9,800원이며, 95% 신뢰구간은 [-4.9, 6]이다. 이를 통해 코로나는 제곱미터당 권리금 9,800원 증가시키는 효과가 있었다고 할 수 있다.
- 코로나 이후 기간의 제곱미터당 권리금의 전체 값은 778만원이다. 코로나가 발생하지 않았다면 751만원으로 예측된다. 해당 예측의 95% 신뢰구간은 [610, 917]이다. 상대적으로 제곱미터당 권리금이 3.7% 증가하는 것으로 나타나고, 95% 신뢰구간은 [-18%, 22%]이다.
- 이는 코로나 이후 전체적으로 볼 때 업종별 제곱미터 당 권리금에 긍정적인 영향을 미친 것으로 볼 수 있지만, 해당 예측의 $p-value$ = 0.348이고, 코로나로 인하여 제곱미터당 권리금에 영향을 준 확률은 65%이다. 이는 통계적으로 유의하다고 할 수 없다. 따라서 코로나로 인해 업종별 제곱미터당 권리금의 가격이 바뀌었다고 보기에는 어렵다.

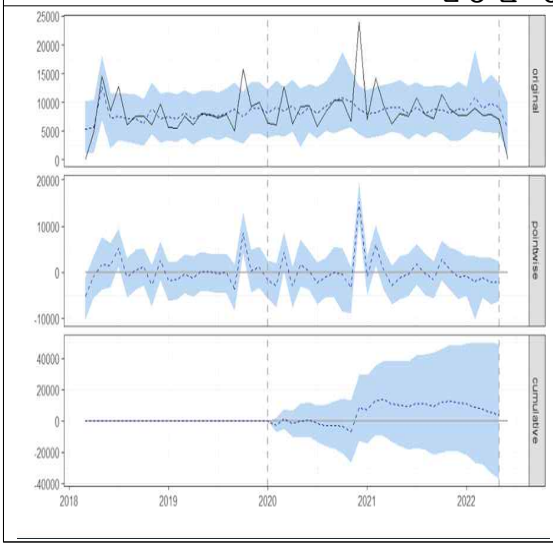
업종별 평균 권리금




```
Posterior inference {CausalImpact}

                          Average        Cumulative
Actual                    4255           119153
Prediction (s.d.)         3849 (363)     107780 (10178)
95% CI                    [3137, 4573]   [87837, 128039]

Absolute effect (s.d.)    406 (363)      11373 (10178)
95% CI                    [-317, 1118]   [-8886, 31316]

Relative effect (s.d.)    11% (9.4%)     11% (9.4%)
95% CI                    [-8.2%, 29%]   [-8.2%, 29%]

Posterior tail-area probability p:   0.11426
Posterior prob. of a causal effect:  89%

For more details, type: summary(impact, "report")
```

- 코로나 이후 2022년 6월까지의 업종별 권리금의 평균값은 약 4,255만원이다. 코로나가 발생하지 않았다면 평균 3,849만원으로 예측된다. 코로나이후 업종별 관리금 예측의 95% 구간은 [3137, 4573]이다. 관측된 반응에서의 예측을 빼면 코로나가 관리금에 미치는 인과관계의 추정치가 산출된다. 이 각 시점에서의 개입효과 는 406만원이고, 95% 신뢰구간은 [-317, 1118]이다. 이를 통해 코로나는 권리금 406만원 증가시키는 효과가 있었다고 할 수 있다.
- 코로나 이후 기간의 권리금의 전체 값은 11억 9,153만원이다. 코로나가 발생하지 않았다면 10억 7,780만원으로 예측된다. 해당 예측의 95% 신뢰구간은 [87837, 128039]이다. 상대적으로 권리금이 11% 증가하는 것으로 나타나며, 95% 신뢰구간은 [-8.2%, 29%]이다.
- 이는 코로나 이후 전체적으로 볼 때 업종별 권리금에 긍정적인 영향을 미친 것으로 볼 수 있지만, 해당 예측의 $p-value$ = 0.11426이고, 코로나로 인하여 업종별 권리금에 영향을 줄 확률은 89%이다. 이는 통계적으로 유의하다고 할 수 없다. 따라서 코로나로 인해 업종별 권리금의 가격이 바뀌었다고 보기에는 어렵다.

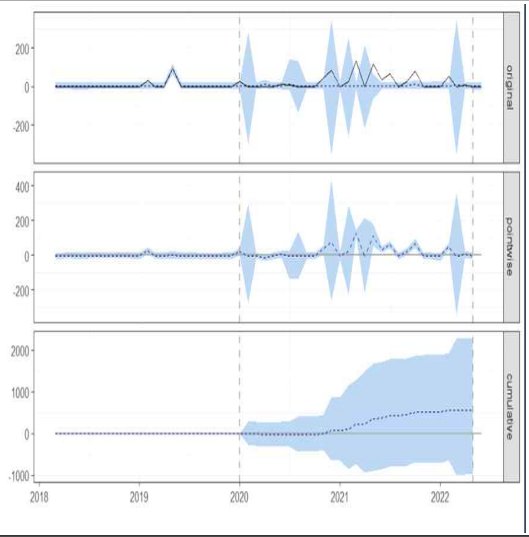
업종별 평균 매매가




```
Posterior inference {CausalImpact}

                          Average        Cumulative
Actual                    9101           254841
Prediction (s.d.)         8979 (762)     251426 (21345)
95% CI                    [7352, 10433]  [205844, 292135]

Absolute effect (s.d.)    122 (762)      3415 (21345)
95% CI                    [-1332, 1750]  [-37294, 48997]

Relative effect (s.d.)    1.4% (8.5%)    1.4% (8.5%)
95% CI                    [-15%, 19%]    [-15%, 19%]

Posterior tail-area probability p:   0.43207
Posterior prob. of a causal effect:  57%

For more details, type: summary(impact, "report")
```

- 코로나 이후 2022년 6월까지의 업종별 매매가의 평균값은 약 9,101만원이다. 코로나가 발생하지 않았다면 평균 8,979만원으로 예측된다. 코로나 이후의 매매가 평균 예측의 95% 구간은 [7352, 10433]이다. 관측된 반응에서의 예측을 빼면 코로나가 매매가에 미치는 인과관계의 추정치가 산출된다. 이 각 시점에서의 개입효과 는 122만원이고, 95% 신뢰구간은 [-1332, 1750]이다. 이를 통해 코로나는 평균 매매가 122만원 증가시키는 효과가 있었다고 할 수 있다.
- 코로나 이후 기간의 매매가의 전체 값은 25억 4,841만원이다. 코로나가 발생하지 않았다면 25억 1,426만원으로 예측된다. 해당 예측의 95% 신뢰구간은 [205844, 292135]이다. 상대적으로 매매가가 1.4% 증가하는 것으로 나타나며, 95% 신뢰구간은 [-15%, 19%]이다.
- 이는 코로나 이후 전체적으로 볼 때 업종별 매매가에 긍정적인 영향을 미친 것으로 볼 수 있지만, 해당 예측의 $p-value$ = 0.43207, 코로나로 인하여 매매가에 영향을 준 확률은 57%이다. 이는 통계적으로 유의하다고 할 수 없다. 따라서 코로나로 인해 매매가의 가격이 바뀌었다고 보기에는 어렵다.

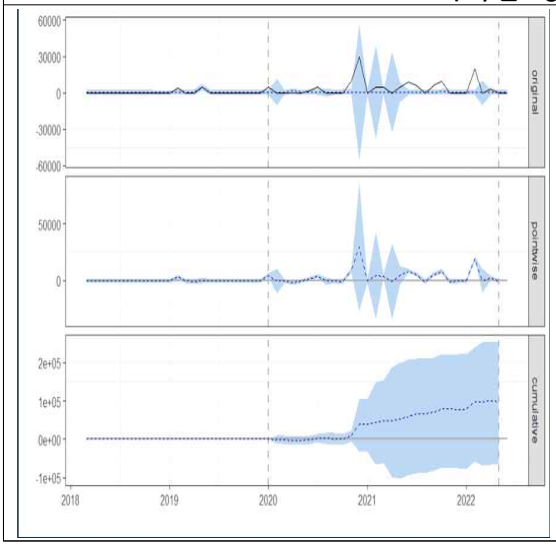
지역별 제곱미터당 권리금




```
Posterior inference {CausalImpact}

                          Average          Cumulative
Actual                    25               693
Prediction (s.d.)         4.5 (22)         126.5 (614)
95% CI                    [-57, 60]        [-1601, 1676]

Absolute effect (s.d.)    20 (22)          566 (614)
95% CI                    [-35, 82]        [-983, 2294]

Relative effect (s.d.)    448% (486%)      448% (486%)
95% CI                    [-777%, 1814%]   [-777%, 1814%]

Posterior tail-area probability p:   0.06809
Posterior prob. of a causal effect:  93%

For more details, type: summary(impact, "report")
```

- 코로나 이후 2022년 6월까지의 지역별 제곱미터당 권리금 평균값은 약 25만원이다. 코로나가 발생하지 않았다면 평균 4.5만원으로 예측된다. 코로나 이후의 제곱미터당 권리금 평균 예측의 95% 구간은 [-57, 60]이다. 관측된 반응에서의 예측을 빼면 코로나가 제곱미터당 권리금에 미치는 인과 관계의 추정치가 산출된다. 이 각 시점에서의 개입효과는 20만원이고, 95% 신뢰구간은 [-35, 82]이다. 이를 통해 코로나는 제곱미터당 권리금 20만원 증가시키는 효과가 있었다고 할 수 있다.
- 코로나 이후 기간의 제곱미터당 권리금의 전체 값은 693만원이다. 코로나가 발생하지 않았다면 126.5만원으로 예측된다. 해당 예측의 95% 신뢰구간은 [-1601, 1676]이다. 상대적으로 제곱미터당 권리금이 448% 증가하는 것으로 나타나고, 95% 신뢰구간은 [-777%, 1814%]이다.
- 이는 코로나 이후 전체적으로 볼 때 지역별 제곱미터당 권리금에 긍정적인 영향을 미친 것으로 볼 수 있고, 해당 예측의 $p-value$ = 0.06809, 코로나로 인하여 지역별 제곱미터당 권리금에 영향을 준 확률은 93%이다. 유의수준 10% 내에서 통계적으로 유의하다. 따라서 코로나로 인해 지역별 제곱미터당 권리금은 영향을 받았다고 할 수 있다.

지역별 평균 권리금




```
Posterior inference {CausalImpact}

                          Average          Cumulative
Actual                    4075             114100
Prediction (s.d.)         530 (2281)       14847 (63865)
95% CI                    [-4987, 6447]    [-139633, 180505]

Absolute effect (s.d.)    3545 (2281)      99253 (63865)
95% CI                    [-2372, 9062]    [-66405, 253733]

Relative effect (s.d.)    669% (430%)      669% (430%)
95% CI                    [-447%, 1709%]   [-447%, 1709%]

Posterior tail-area probability p:   0.04714
Posterior prob. of a causal effect:  95.286%

For more details, type: summary(impact, "report")
```

- 코로나 이후 2022년 6월까지의 지역별 권리금 평균값은 약 4,075만원이다. 코로나가 발생하지 않았다면 평균 530만원으로 예측된다. 코로나 이후의 권리금 예측의 95% 구간은 [-4987, 6447]이다. 관측된 반응에서의 예측을 빼면 코로나가 권리금에 미치는 인과 관계의 추정치가 산출된다. 이 각 시점에서의 개입효과 는 3,545만원이고, 95% 신뢰구간은 [-2372, 9062]이다. 이를 통해 코로나는 권리금 3,545만원을 증가시키는 효과가 있었다고 할 수 있다.
- 코로나 이후 기간의 권리금의 전체 값은 11억 4,100만원이다. 코로나가 발생하지 않았다면 1억 4,847만원으로 예측된다. 해당 예측의 95% 신뢰구간은 [-139633, 180505]이다. 상대적으로 권리금이 669% 증가하는 것으로 나타나고, 95% 신뢰구간은 [-447%, 1709%]이다.
- 이는 코로나 이후 전체적으로 볼 때 지역별 권리금에 긍정적인 영향을 미친 것으로 볼 수 있고, 해당 예측의 $p-value$ = 0.04714, 코로나로 인하여 지역별 권리금에 영향을 준 확률은 95.286%이다. 유의수준 5%내에서 통계적으로 유의하다. 따라서 코로나로 인해 지역별 권리금은 영향을 받았다고 할 수 있다.

2.2 과제2 권리금 결정 요인

1. 데이터준비

```
데이터의 형태 :  (35287, 21)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 35287 entries, 0 to 35286
Data columns (total 21 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   address                   35287 non-null  object
 1   level                     35287 non-null  int64
 2   goodwill                  35287 non-null  int64
 3   sale_price                35287 non-null  int64
 4   contract_date             35287 non-null  object
 5   real_sale_price_percent   35287 non-null  float64
 6   store_type                35287 non-null  object
 7   area                      35287 non-null  float64
 8   제곱미터당_권리금                 35287 non-null  float64
 9   year_month                35287 non-null  object
 10  year                      35287 non-null  int64
 11  weekday                   35287 non-null  int64
 12  지역                        35287 non-null  object
 13  시                         35287 non-null  object
 14  지역_시                       35287 non-null  object
 15  popular                   35287 non-null  int64
 16  경도                        35287 non-null  float64
 17  위도                        35287 non-null  float64
 18  Num                       35287 non-null  int64
 19  distance                  35287 non-null  float64
 20  CovidNum                  35287 non-null  int64
dtypes: float64(6), int64(8), object(7)
```

- 권리금 결정 요인을 파악하기 위하여 데이터를 파악한 뒤, ‘store_type’을 one-hot encoding을 시행한다.
- 수치를 가진 데이터들의 범위를 일치화 시켜주기 위하여 'level', 'goodwill', '제곱미터당_권리금', 'popular', 'Num', 'distance', 'CovidNum' 열을 MinMax 정규화를 시행해준다.
- 분석에 필요 없는 변수라고 판단한 ‘address', 'sale_price', 'contract_date', 'real_sale_price_percent', 'area', 'weekday', '지역', 'year_month', 'year', '시', '지역_시', '경도', '위도' 열을 삭제한다.
- 예측 모델을 생성하기 위하여, train : test = 70:30으로 나누어 준다.

2. 검증 결과

- 사용한 예측 모델은 Linear Regression, Ridge, XGBoost 모델이며, 각 모델에 대한 RMSE와 R^2값은 <표 1>과 같다.

모델 / 평가지표	Linear Regression	Ridge	XGBoost
RMSE	0.0298	0.0298	0.0273
R^2	0.2939	0.2935	0.4062

<표 1> 모델 및 평가지표

III. 분석결과

Linear Regression을 통하여 goodwill을 예측한 결과, 추정된 회귀식은 다음과 같다.

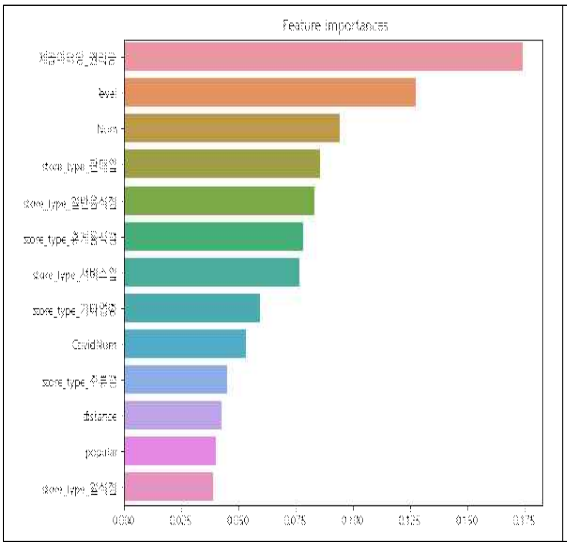
goodwill = 0.0457*level +0.6131*제곱미터당_권리금 -0.005*popular +0.0004*Num
+0.0118*distance +0.0002*CovidNum -0.0168*store_type_기타업종
-0.0163*store_type_서비스업 -0.0114*store_type_일반음식점
-0.015*store_type_일식점 -0.0111*store_type_주류점
-0.0247*store_type_판매업 -0.0195*store_type_휴게음식점

Ridge를 통하여 goodwill을 예측한 결과, 추정된 회귀식은 다음과 같다.

goodwill = 0.044*level +0.5807*제곱미터당_권리금 -0.0048*popular +0.0004*Num
+0.0117*distance +0.0003*CovidNum -0.0168*store_type_기타업종
-0.0162*store_type_서비스업 -0.011*store_type_일반음식점
-0.0145*store_type_일식점 -0.0108*store_type_주류점
-0.0243*store_type_판매업 -0.0187*store_type_휴게음식점

Linear Regression, Ridge 회귀식 모두 '제곱미터당_권리금' 열의 가중치가 가장 크다는 것을 알수 있다. 현재 회귀식의 경우, goodwill 또한 정규화가 되어있다. 따라서, 반정규화를 시켜주어야한다. 예를 들어, goodwill= 0.2일 경우, 0.2*(max-min)+min을 해주어야 한다. 본 문제의 경우 31809.6으로 나타난다.

XGBoost의 경우 트리구조 모델로, 변수 중요도를 확인한 결과 '제곱미터당_권리금', 'level', 'Num' 순으로 나타나는 것을 확인할 수 있다.



	Feature importances
제곱미터당_권리금	0.1744
level	0.1277
Num	0.0942
store_type_판매업	0.0857
store_type_일반음식점	0.0833
store_type_휴게음식점	0.0784
store_type_서비스업	0.0769
store_type_기타업종	0.0594
CovidNum	0.0533
store_type_주류점	0.0452
distance	0.0426
popular	0.0400
store_type_일식점	0.0389