JEJU BIG DATA COMPETITION

Team Name : 감귤

Team Members : 최연웅, 김보경, 김태호, 부대권

Jeju

JTP 제주테크노파크 JEJU TECHNOPARK

DACON DATA TO VALUE

CONTENT

1. CONCEPT
2. 클러스터링 기준

1. RESULTS
2. CONCLUSION

STEP #1
DATA PREPROCESSING & EDA

STEP #2
MODEL BUILDING & EVALUATION

STEP #3
RESULT & CONCLUSION

1. PSUEDO-CODE
2. non_COV 계산법
3. COV_ratio 계산법
4. 각 그룹별 non_COV 및 COV_ratio

CONTENT

1

STEP #1 DATA PREPROCESSING & EDA

1. CONCEPT
2. 클러스터링 기준

STEP #1-1 CONCEPT

이번 대회의 목적은 2019년 1월부터 2020년 4월까지의 카드매출 데이터를 가지고 2020년 7월의 업종, 지역별 총 매출을 예상하는 것이다. Public 점수는 3월까지의 데이터만으로 예측한 4월의 매출을 이용하여 측정하였다. 하지만, 2020년 4월과 7월은 코로나19의 영향력도 다르고, 코로나19가 없는 상황이라고 하더라도 여러 업종별 매출이 많이 다르게 나타난다. 따라서 우리 팀은 4월 데이터를 이용한 Validation결과를 매우 제한적으로 사용하였다.

또한, 여러 가지 머신러닝 기법을 사용하기보다는 그룹들을 특정한 기준으로 클러스터링하여 그 클러스터마다 독립적인 방법으로 7월의 매출을 예측하는 것이 더 효과적일 것이라 판단하였다. 따라서 우리는 그룹의 편차, 변동성, 선형성, 계절성을 기준으로 클러스터링하여 각 그룹마다 다른 방법으로 7월의 매출을 예측하였다.

7월의 매출을 결정하는 요소는 크게 코로나19가 없는 상황(가상)에서의 7월 매출과 2020년 7월의 코로나19의 영향력으로 나누어 볼 수 있다. 우리는 그룹별로 코로나19가 없는 상황의 7월 매출, 7월의 코로나19의 영향력을 독립적으로 구하여 이를 곱하는 방식으로 그룹별 7월 매출을 결정하였다.

STEP #1-1 CONCEPT

코로나로 인한
매출 변화율

$$COV\_07 = non\_COV\_07 \times COV\_ratio$$

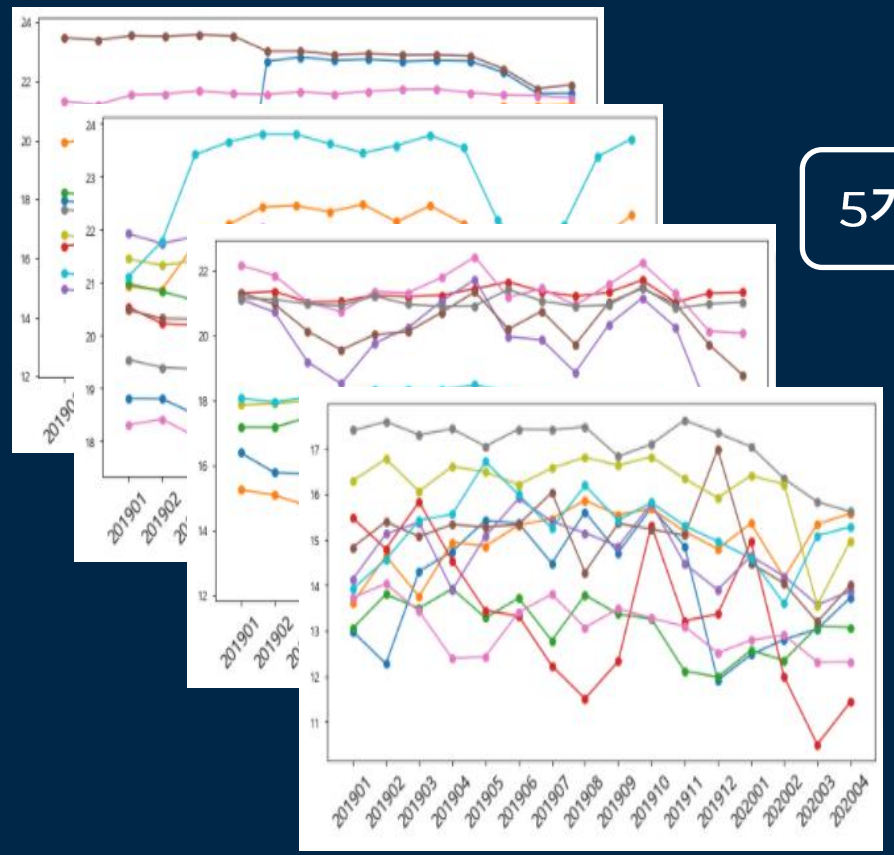
7월 총 매출

코로나가 없을 때
예상되는 7월 매출

STEP #1-1 CONCEPT

COV_07 = non_COV_07 × COV_ratio

650개의 그룹



5가지 판단 기준

8개의 CASE

STEP #1-2 클러스터링 기준

1. 결측치 유무

결측값 존재?

그룹: 시도 및 업종
(예: 제주, 한식 음식점업)

동일한 그룹 내 2019년 1월 ~ 2020년 2월 중 하나라도 매출이 집계되지 않는 기간이 존재할 경우 결측 그룹으로 판단한다.

STEP #1-2 클러스터링 기준

2. std (표준편차)

std = 그룹 내 월별 AMT의 표준편차

그룹: 시도 및 업종
(예: 제주, 한식 음식점업)

std: 그룹 내 월별 AMT의 표준편차
std가 작을 경우 2019년 ~ 2020년 1월의 평균으로 non_cov_07을 예측한다.

std가 크면 해당 그룹의 변동성이 크므로 다른 방법으로 non_cov_07을 예측한다.

STEP #1-2 클러스터링 기준

3. p 변수(변동성)

$$p = \frac{\text{그룹 내 이용 고객이 3명 이하인 카테고리 수}}{\text{그룹 내 카테고리 수}}$$

그룹 : 지역 및 업종
카테고리: 데이터상 해당 그룹에 해당하는 row 수

CSTMR_CNT(이용 고객 수)이 3 미만인 카테고리는 없음. (비식별 조치로 삭제됨) CSTMR_CNT가 낮은 카테고리는 없어질 가능성이 있다. 또한 이용 고객 수가 매우 적기 때문에 불규칙하게 변동하는 경우가 많다.

새롭게 정의된 변수 p를 통해서 **그룹의 변동성**을 판단한다.

p가 크다면,
1. **대부분의 row에서 이용 고객이 3명 이하**
2. **시기별 데이터 변동이 커질 것이다.**

STEP #1-2 클러스터링 기준

4. R2_score (선형성)

R2_score= 그룹 내에서 선형회귀 분석 후 계산한 r2_score

그룹: 시도 및 업종
(예: 제주, 한식 음식점업)

동일한 그룹 내에서 시간을 설명변수, 총 AMT를 반응 변수로 하는 선형회귀 모델을 훈련한다.
훈련결과 r2_score에 따라서 해당 그룹의 시간에 대한 선형성 유무를 판단한다.

STEP #1-2 클러스터링 기준

5. RMSE, diff_per(**계절성**)

그룹: 시도 및 업종
(예: 제주, 한식 음식점업)

시간에 따라 AMT총액의 변화를 살펴보면 두 번의 극댓값과 한 번의 극솟값을 갖는 **4차 다항함수** 꼴을 보이는 그룹들이 존재한다. 이러한 그룹들은 **여름에는 매출이 증가하고 겨울에는 매출이 감소하는 "계절성"**을 갖는다고 판단하였다.

$$\text{RMSE} = \text{4차 다항함수 학습 후 측정한 RMSE 오차}$$

$$\text{diff\_per} = \frac{abs(\text{19년1월} - \text{20년1월})}{\text{최대월 매출} - \text{최소월 매출}} \times 100$$

- RMSE
  4차 다항함수 학습 후 측정한 RMSE 값이다. RMSE가 작은 그룹들을 계절성의 충분조건으로 보았다.
- diff_per
  계절성을 갖는 그룹이라면, 19년 1월과 20년 1월의 매출이 비슷해야 한다. 따라서 그룹내 최대, 최소 매출 차이에 비하여 19년 1월과 20년 1월 매출 차이가 작은 것을 계절성의 충분조건으로 보았다.

2 STEP #2 MODEL BUILDING & EVALUATION

1. PSUEDO-CODE
2. non_COV 계산법
3. COV_ratio 계산법
4. 각 그룹별 non_COV 및 COV_ratio

STEP #2 PSUEDO-CODE

- **결측치**
  2019.01~2020.01에 결측 데이터의 존재 유무
- **표준편차**
  std가 0.15보다 클 때, 분산이 큰 경우로 판단
- **선형성**
  r2 score가 0.5보다 클 경우, 선형성 존재
- **변동성p**
  p가 0.5보다 클 때, 변동성이 큰 경우로 판단
- **계절성**
  작년 1월과 올해 1월의 매출액 차이와 그룹의 최대 최소값의 차이의 비율이 30% 보다 작고,
  4차 함수의 RMSE가 0.3보다 작을 경우 계절성 존재

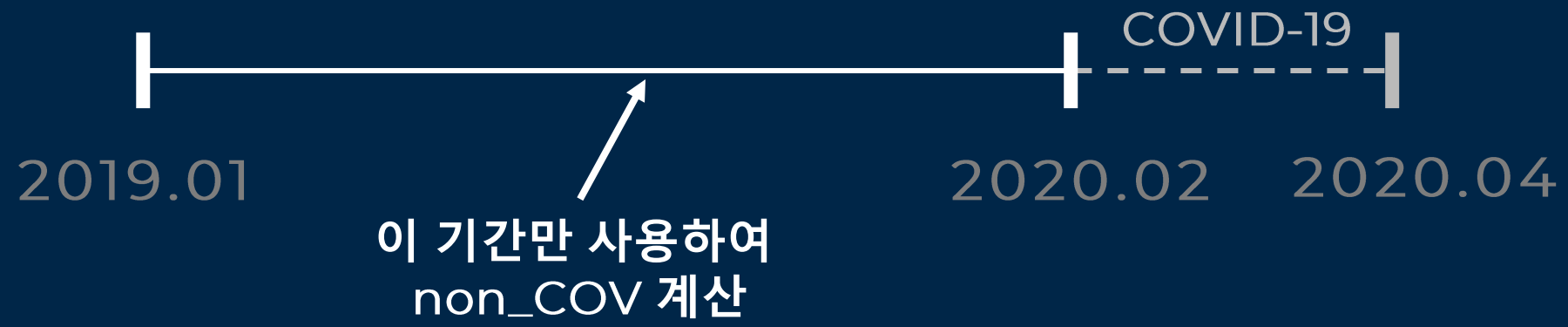
```
if (19년 데이터 + 20년 1월) 전부 존재:
    (19년 데이터 + 20년 1월)만 가지고 std를 계산
    if std > 0.15: # error_group
        if p < 0.5:  ## 변동성이 적다.

            if 변화가 선형적이다: .........1)
                cov_ratio_07 = (1 + 3 * mean(20Y02M, 20Y03M, 20Y04M) / mean(19Y07M, 19Y08M, 19Y09M)) / 4
                cov_ratio_04 = mean(20Y02M, 20Y03M) / mean(19Y07M, 19Y08M, 19Y09M)

                non_cov_07 = mean(19Y11M, 19Y12M, 20Y01M)
                non_cov_04 = mean(19Y11M, 19Y12M, 20Y01M)

                예측 7월 = non_cov_07 * cov_ratio_07
                예측 4월 = non_cov_04 * cov_ratio_04

            else 변화가 비선형적이다:

                if 계절성을 갖는다.: .........2)
                    # 다항회귀 모델 예측후 rmse 가 0.3보다 작고, diff_per가 30% 미만일 경우
                    cov_ratio_07 = (1 + 3 * 20Y04M/19Y04M) / 4
                    cov_ratio_04 = 20Y04M/19Y04M

                    non_cov_07 = mean(19Y06M, 19Y07M, 19Y08M)
                    non_cov_04 = mean(19Y03M, 19Y04M, 19Y05M)

                    예측 7월 = non_cov_07 * cov_ratio_07
                    예측 4월 = non_cov_04 * cov_ratio_04

                elif 특별한 규칙이 없다: .........3)
                    # 다항회귀 모델 예측후 rmse 가 0.3보다 크거나 같고, diff_per가 30% 이상일 경우
                    cov_ratio_07 = (1 + 4 * mean(20Y02M, 20Y03M, 20Y04M) / mean(19Y02M, 19Y03M, 19Y04M)) / 5
                    cov_ratio_04 = mean(20Y02M, 20Y03M) / mean(19Y02M, 19Y03M)

                    non_cov_07 = mean(19Y05M, 19Y06M, 19Y07M, 19Y08M)
                    non_cov_04 = mean(19Y02M, 19Y03M, 19Y04M, 19Y05M)

                    예측 7월 = non_cov_07 * cov_ratio_07
                    예측 4월 = non_cov_04 * cov_ratio_04
```

STEP #2 PSUEDO-CODE

```
elif p >= 0.5:  ## 변동성이 크다.

    if 변화가 선형적이다: .........4)
        cov_ratio_07 = (1 + 3 * mean(20Y02M, 20Y03M, 20Y04M) / mean(19Y07M, 19Y08M, 19Y09M)) / 4
        cov_ratio_04 = mean(20Y02M, 20Y03M) / mean(19Y07M, 19Y08M, 19Y09M)

        non_cov_07 = mean(19Y05M, 19Y06M, 19Y07M, 19Y08M)
        non_cov_04 = mean(19Y05M, 19Y06M, 19Y07M, 19Y08M)

        예측 7월 = non_cov_07 * cov_ratio_07
        예측 4월 = non_cov_04 * cov_ratio_04

    else 변화가 비선형적이다:
        if 계절성을 갖는다.: .........5)
            # (다항회귀 모델 예측후 rmse 가 0.3보다 작다) and (diff_per가 30% 미만이다)
            cov_ratio_07 = (1 + 4 * mean(20Y02M, 20Y03M, 20Y04M) / mean(19Y02M, 19Y03M, 19Y04M)) / 5
            cov_ratio_04 = mean(20Y02M, 20Y03M) / mean(19Y02M, 19Y03M)

            non_cov_07 = mean(19Y05M, 19Y06M, 19Y07M, 19Y08M, 19Y09M)
            non_cov_04 = mean(19Y02M, 19Y03M, 19Y04M, 19Y05M, 19Y06M)

            예측 7월 = non_cov_07 * cov_ratio_07
            예측 4월 = non_cov_04 * cov_ratio_04

        elif 특별한 규칙이 없다: .........6)
            # (다항회귀 모델 예측후 rmse 가 0.3보다 크다) or (diff_per가 30% 이상이다)
            cov_ratio_07 = (1 + 4 * mean(20Y02M, 20Y03M, 20Y04M) / mean(19Y02M, 19Y03M, 19Y04M)) / 5
            cov_ratio_04 = mean(20Y02M, 20Y03M) / mean(19Y02M, 19Y03M)

            non_cov_07 = ((19Y05M, 19Y06M, 19Y07M, 19Y08M) * 2 + 나머지 기간 AMT) / (전체 기간의 길이 + 4)
            non_cov_04 = ((19Y02M, 19Y03M, 19Y04M, 19Y05M) * 2 + 나머지 기간 AMT) / (전체 기간의 길이 + 4)

            예측 7월 = non_cov_07 * cov_ratio_07
            예측 4월 = non_cov_04 * cov_ratio_04
```

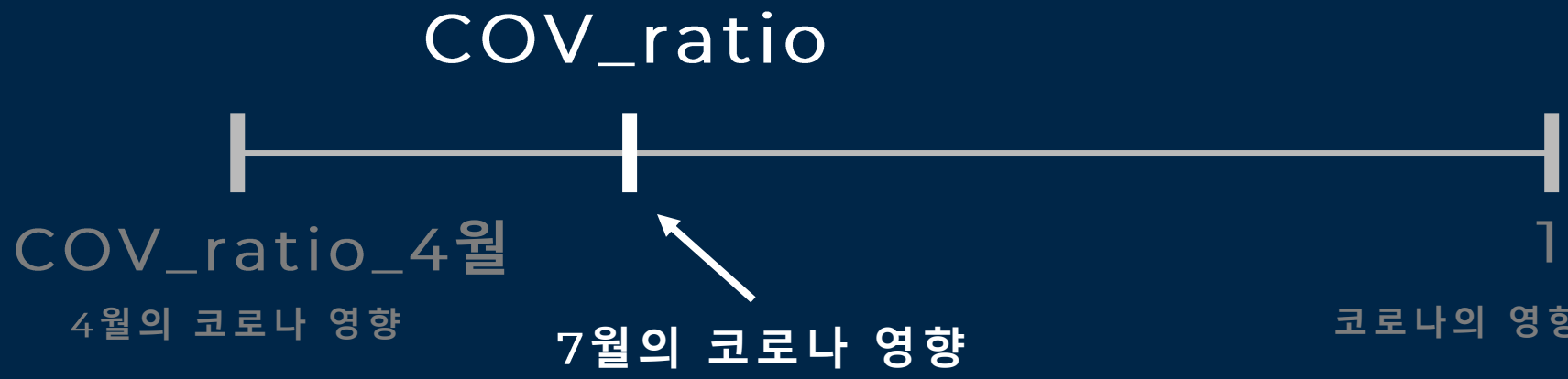
```
        else std <= 0.15:
            cov_ratio_07 = (1 + 3 * 20Y04M/19Y04M) / 4
            cov_ratio_04 = 20Y03M/19Y03M

            non_cov_07 = mean(19Y06M, 19Y07M, 19Y08M)
            non_cov_04 = mean(19Y03M, 19Y04M, 19Y06M)

            전체 평균 * cov_ratio .........7)

    else 19년 데이터가 전부 존재하지 않는다: # uncomplete_group
        # cov_ratio_07 구하기
        if 20년 4월 데이터가 존재한다:
            cov_ratio_07 = 20Y04M / 전체평균
        elif 20년 3월 데이터가 존재한다:
            cov_ratio_07 = 0.7 * 20Y03M / 전체평균
        else:
            cov_ratio_07 = 0.4

        # cov_ratio_04 구하기
        if 20년 3월 데이터가 존재한다:
            cov_ratio_04 = 20Y03M / 전체평균
        elif 20년 3월 데이터가 존재한다:
            cov_ratio_04 = 20Y02M / 전체평균
        else:
            cov_ratio_04 = 0.4

        # 결측을 0으로 채워준다.
        # 20년 1월까지 순차적으로 가중치를 크게 부여한다.
        for i, month in enumerate([20Y01M, 19Y12M, 19Y11M, ..., 19Y01M]):
            총 AMT += AMT[month] * 0.9^(i+1)
        가중치 평균 AMT = 총 AMT / (0.9 + 0.9^2 + 0.9^3 + ...)

        non_cov_07 = 가중치 평균 AMT
        non_cov_04 = 가중치 평균 AMT
```

STEP #2 non_COV 계산법



- non_COV: 코로나의 영향을 고려하지 않은 예상 7월 매출액
- 2020년 2월부터 코로나가 매출에 직접적으로 영향을 끼친다고 판단
- 따라서, 2020년 1월까지의 데이터만을 이용하여 non_COV 예측

STEP #2 코로나 영향 변수 계산법



- 7월은 4월보다 코로나의 직접적인 영향이 적다고 판단
- COV_ratio는 1과 COV_ratio_4월의 a:b 내분점을 사용
- 4월 데이터가 포함된 COV_ratio_4월을 사용
  (단, validation에 사용된 COV_ratio_4월에는 4월 데이터 미포함)

$$COV\_ratio = \frac{COV\_ratio\_4월(+4월\ 데이터) + a + \frac{20년\ 2\sim4월\ 평균}{19년\ 7\sim9월\ 평균} \times b}{a+b}$$

STEP #2

CASE 1. 분산이 크고 변동성이 적은 경우

CASE 1.1 선형적일 때

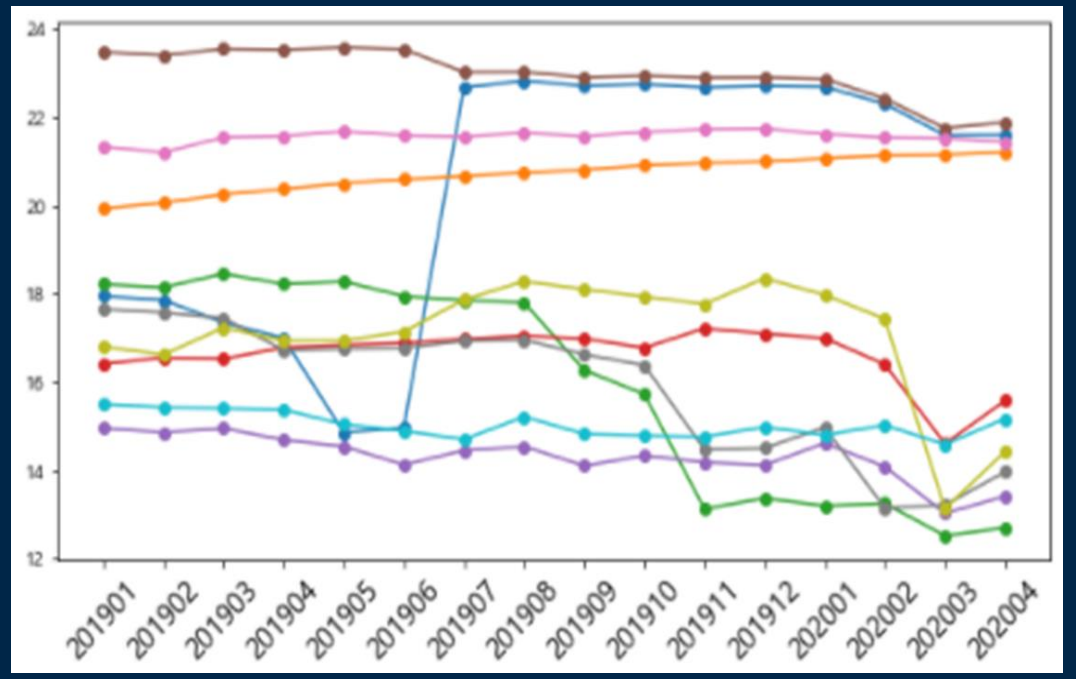
non_COV

19년 11월
19년 12월
20년 1월

평균

COV_ratio

$$\frac{1+\frac{\text{20년 2~4월 평균}}{\text{19년 7~9월 평균}}\times 3}{4}$$



<13개>

STEP #2

CASE 1. 분산이 크고 변동성이 적은 경우

CASE 1.2 계절성이 있을 때

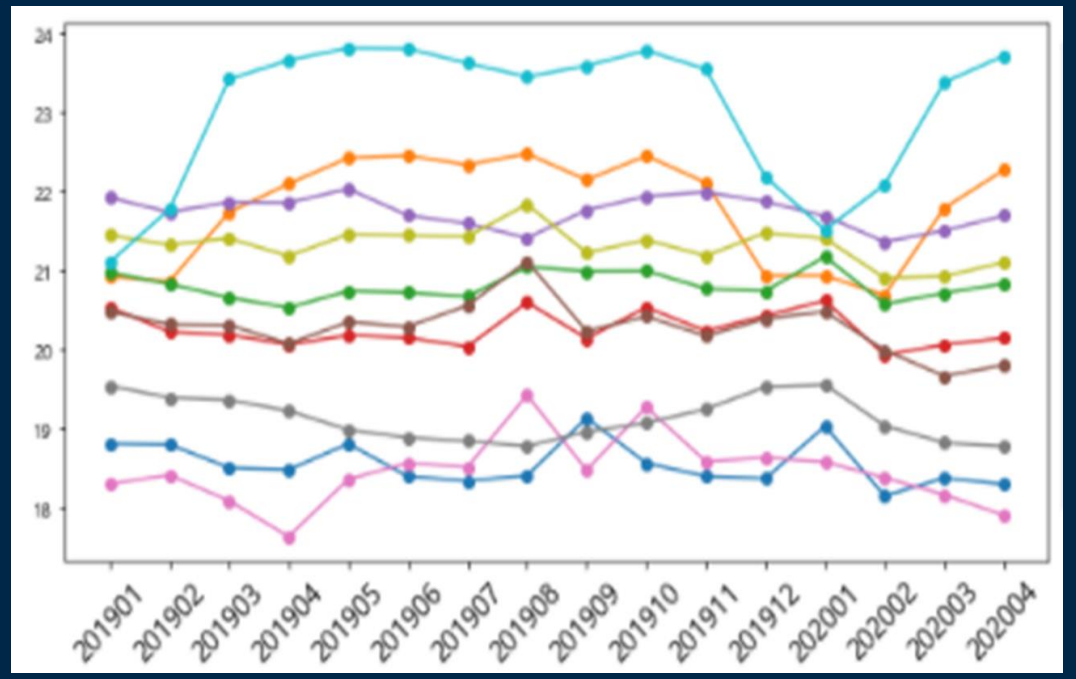
non_COV

19년 6월
19년 7월
19년 8월

평균

COV_ratio

$$\frac{1+\frac{\text{20년 4월 평균}}{\text{19년 4월 평균}}\times 4}{5}$$



<126개>

STEP #2

CASE 1. 분산이 크고 변동성이 적은 경우

CASE 1.3 불규칙할 때

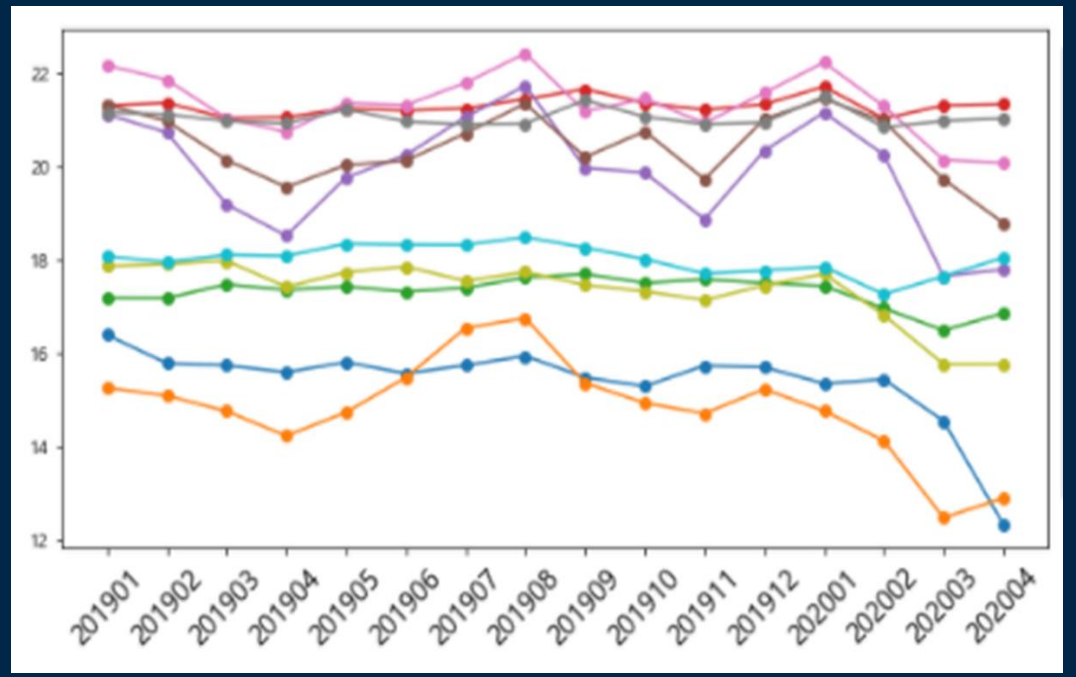
non_COV

19년 5월
19년 6월
19년 7월
19년 8월

평균

COV_ratio

$$\frac{1+\frac{\text{20년 2~4월 평균}}{\text{19년 2~4월 평균}}\times 4}{5}$$



<71개>

STEP #2

CASE 2. 분산이 크고 변동성이 큰 경우

CASE 2.1 선형적일 때

non_COV

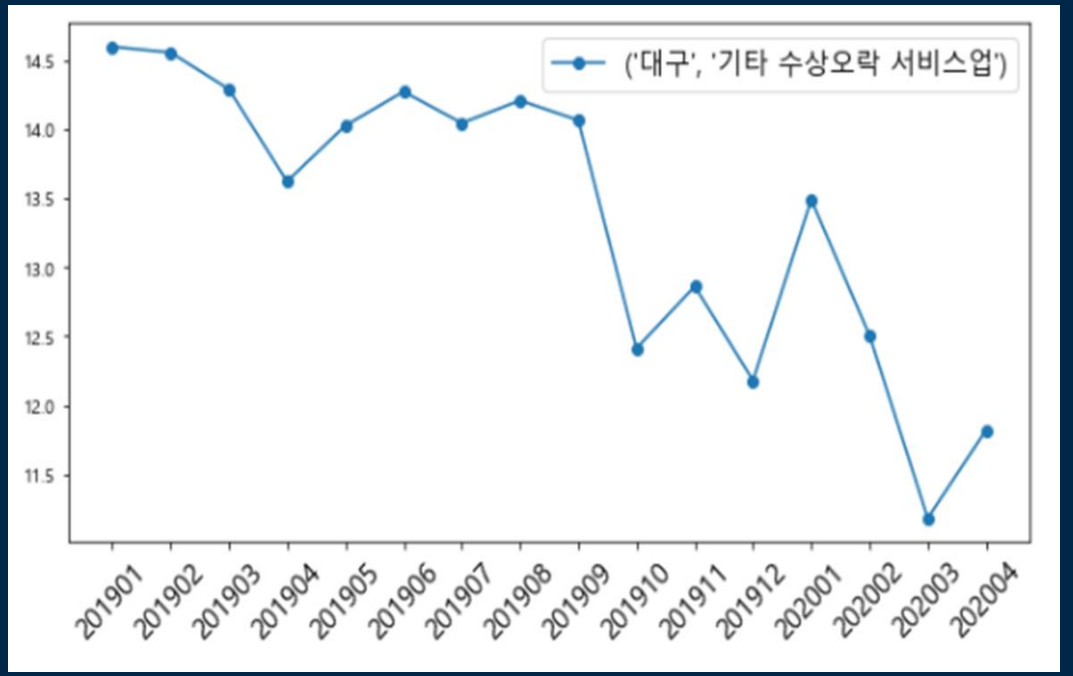
19년 5월
19년 6월
19년 7월
19년 8월

평균

COV_ratio

$$1+\frac{\text{20년 2~4월 평균}}{\text{19년 7~9월 평균}}\times 3$$

$$\frac{}{4}$$



<1개>

STEP #2

CASE 2. 분산이 크고 변동성이 큰 경우

CASE 2.2 계절성이 있을 때

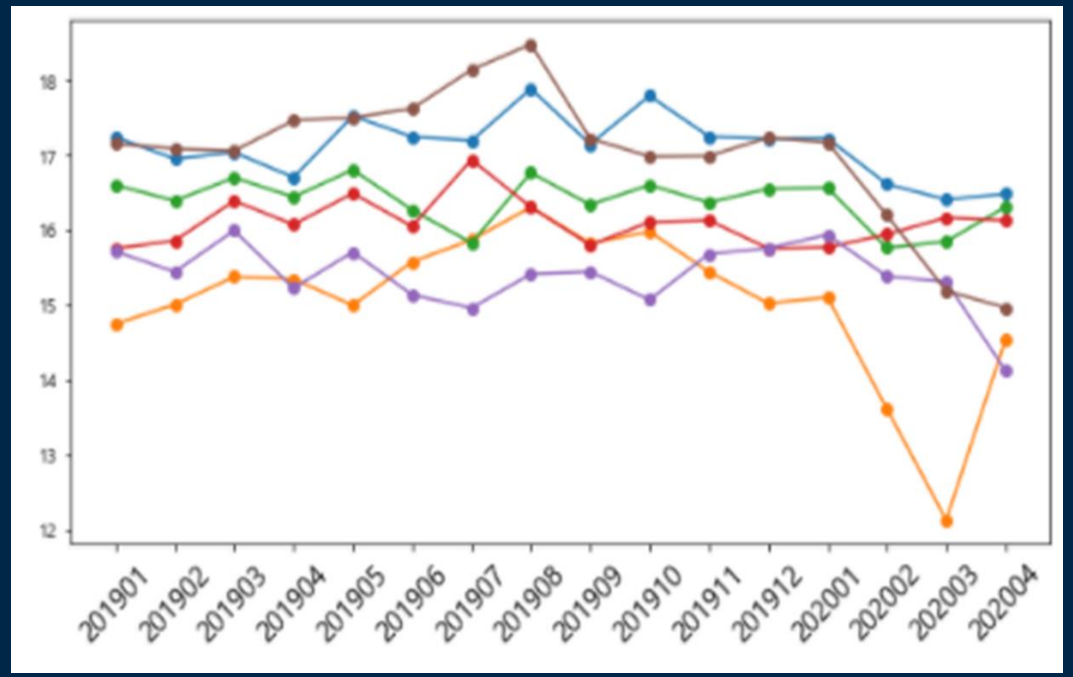
non_COV

19년 5월
19년 6월
19년 7월
19년 8월
19년 9월

평균

COV_ratio

$$\frac{1+\frac{\text{20년 2~4월 평균}}{\text{19년 2~4월 평균}}\times 4}{5}$$



<6개>

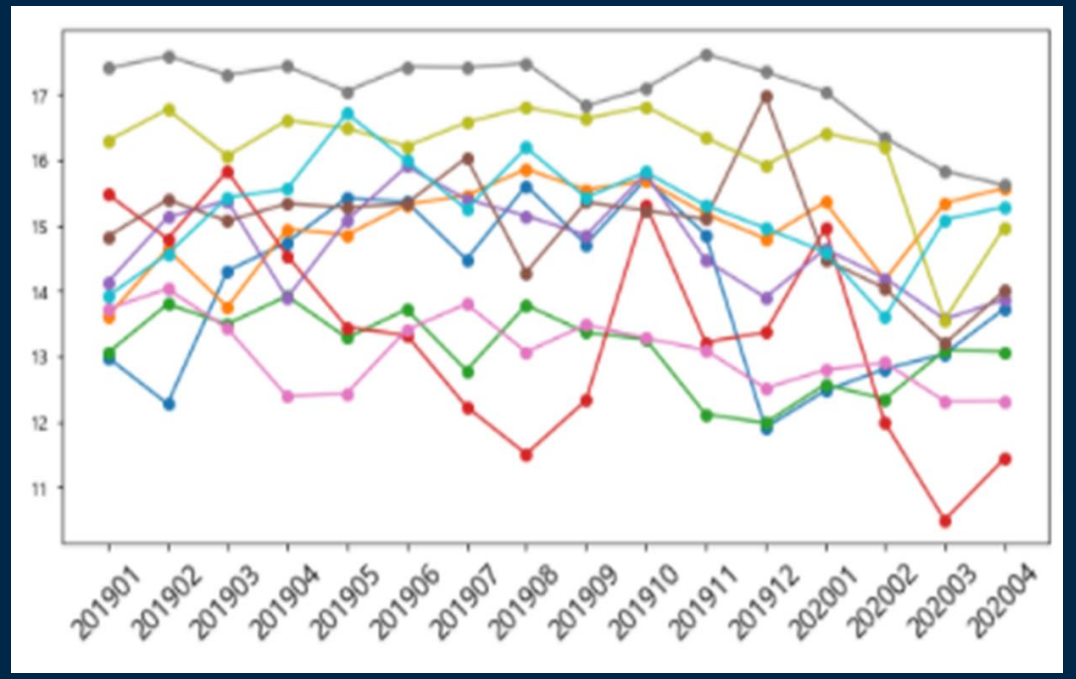
STEP #2

CASE 2. 분산이 크고 변동성이 큰 경우

CASE 2.3 불규칙할 때

$$non\_COV = \frac{(19년\ 5~8월) \times 2 + 나머지}{전체\ 기간의\ 길이 + 4}$$

$$COV\_ratio = \frac{1 + \frac{20년\ 2~4월\ 평균}{19년\ 2~4월\ 평균} \times 4}{5}$$



<14개>

STEP #2

CASE 3. 분산이 작은 경우

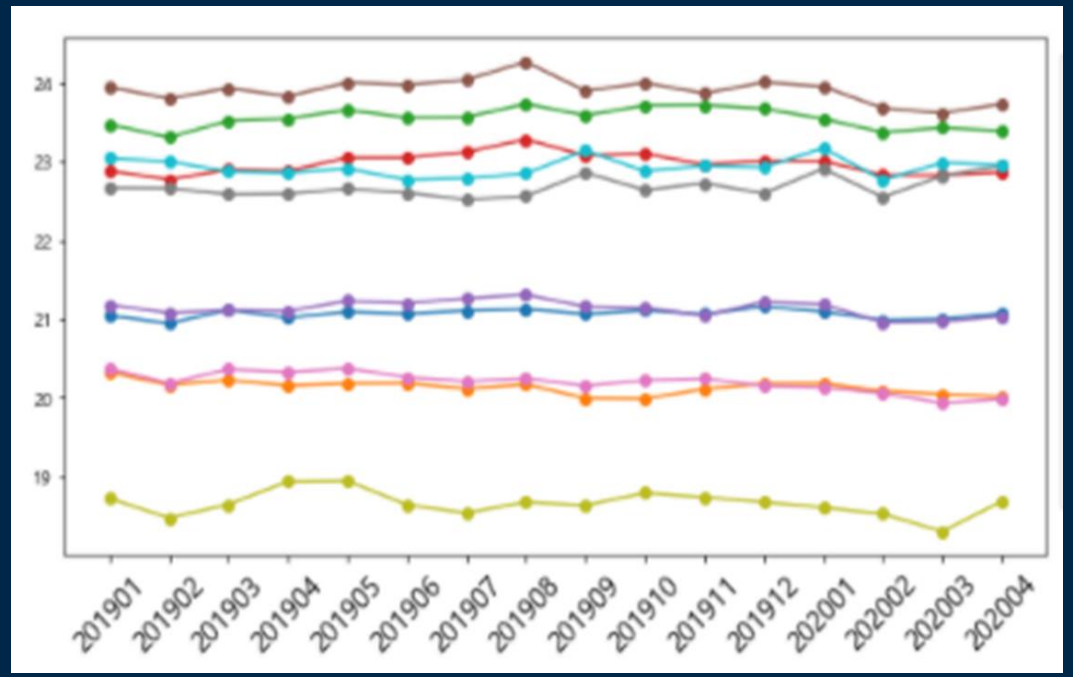
non_COV

19년 6월
19년 7월
19년 8월

평균

COV_ratio

$$\frac{1+\frac{\text{20년 4월 평균}}{\text{19년 4월 평균}}\times 4}{5}$$



<363개>

STEP #2

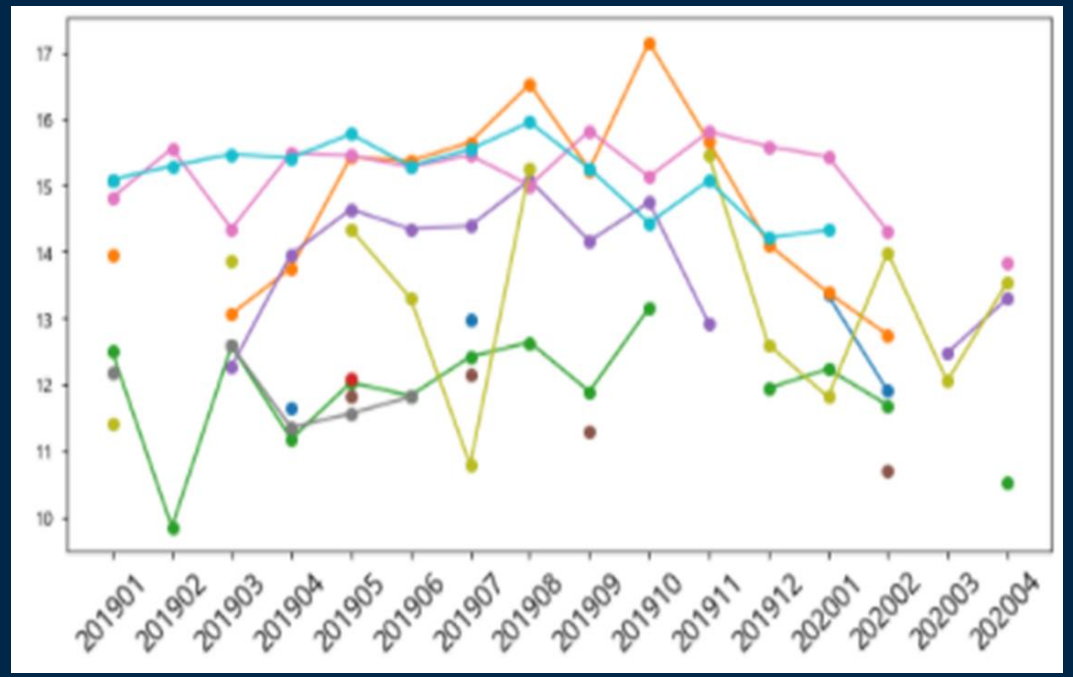
CASE 4. 결측 데이터 존재

non_COV

1. 결측 월 매출 0으로 채우기
2. 20년 1월에 가까울수록 매출에 가중치를 크게 하여 평균

COV_ratio

$\frac{\text{20년 4월}}{\text{전체 평균}}$	$\frac{\text{20년 3월}}{\text{전체 평균}} \times 0.7$	0.4
4월 데이터가 존재하는 경우	4월 데이터가 없지만 3월 데이터가 존재하는 경우	3, 4월 데이터가 모두 없는 경우

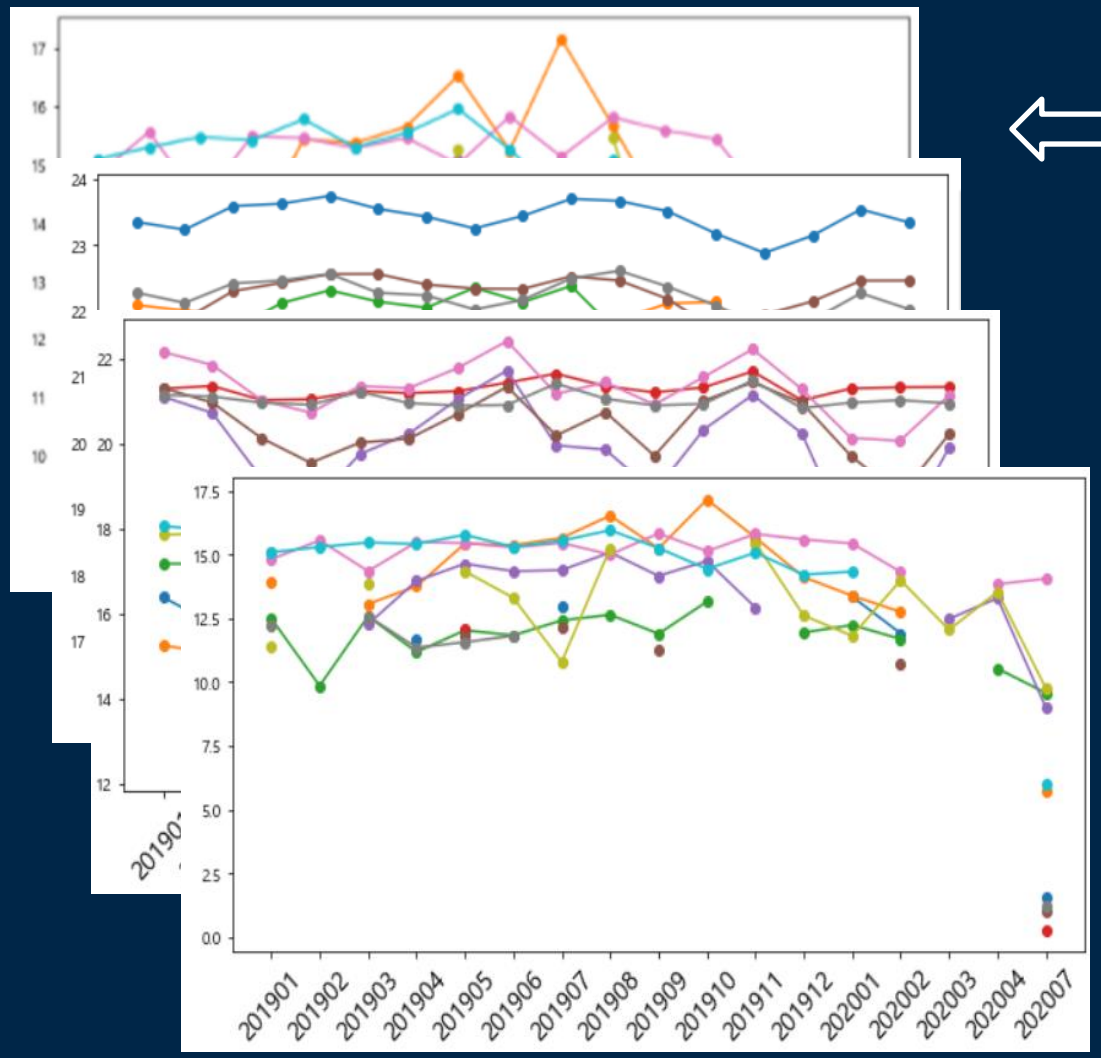


<56개>

3 STEP #3 RESULTS & CONCLUSION

1. RESULTS
2. CONCLUSION

STEP #3 RESULTS



⇦ **예측된 7월을 포함한 그래프**

< 예외 처리 >
데이터에 이상치나 결손이 있을 경우 예외 처리하였다.

- 경남 버스 운송업 : 2020년 4월 AMT 그대로 사용
- 충북 택시 운송업 : 2020년 4월 AMT 그대로 사용
- 충남 버스 운송업 : 2020년 4월 AMT 그대로 사용
- 제주 택시 운송업 : 2020년 4월 AMT 그대로 사용
- 제주 그외 기타 분류안된 오락관련 서비스업 : 2020년 4월 AMT 그대로 사용
- 부산 면세점 : {(2020년 3월 AMT) / 3} 사용

STEP #3 RESULTS

<예외 처리 근거>

- **경남 버스 운송업** : 하락 추세이지만, 코로나의 2,3,4월 대비 코로나가 급격하게 감소하지 않을 것이라 판단하여 2020년 4월 AMT 그대로 사용했다.
- **충북 택시 운송업** : 2020년 4월 AMT 그대로 사용
- **충남 버스 운송업** : 19년 9월 이전에는 데이터가 존재하지 않았으나, 그 뒤로는 일정 매출이 계속 기록되었다. 앞으로도 매출이 지속적이고, 코로나로 인한 영향도 크지 않다고 판단하여 2020년 4월 AMT 그대로 사용했다.
- **제주 택시 운송업** : 19년 10월 이전에는 데이터가 존재하지 않았으나, 그 뒤로는 일정 매출이 계속 기록되었다. 앞으로도 매출이 지속적이고, 코로나로 인한 영향도 크지 않다고 판단하여 2020년 4월 AMT 그대로 사용했다.
- **제주 그외 기타 분류안된 오락관련 서비스업** : 19년7월 이전에는 데이터가 존재하지 않았으나, 그 뒤로는 일정 매출이 계속 기록되었다. 앞으로도 매출이 지속적이고, 코로나로 인한 영향도 크지 않다고 판단하여 2020년 4월 AMT 그대로 사용했다.
- **부산 면세점** : 꾸준히 매출이 존재하던 부산 면세점업은 2020년 4월에만 데이터가 존재하지 않는다. 다른 관광, 항공 관련 산업들이 코로나19의 영향을 크게 받았다는 점을 고려하여 {(2020년 3월 AMT) / 3} 으로 7월을 예측하였다.

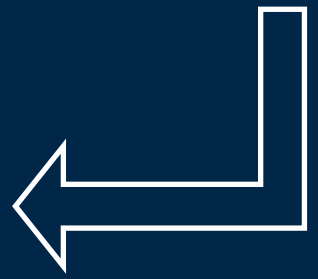
STEP #3 CONCLUSION

id	REG_YYMM	CARD_SIDO_NM	STD_CLSS_NM	AMT
0	202004	강원	건강보조식품 소매업	8.882399e+07
1	202004	강원	골프장 운영업	4.708347e+09
2	202004	강원	과실 및 채소 소매업	1.121029e+09
3	202004	강원	관광 민예품 및 선물용품 소매업	1.436078e+07
4	202004	강원	그외 기타 분류안된 오락관련 서비스업	0.000000e+00
...	...	...	...	...
1389	202007	충북	피자 햄버거 샌드위치 및 유사 음식점업	1.514481e+09
1390	202007	충북	한식 음식점업	2.091243e+10
1391	202007	충북	호텔업	1.488726e+07
1392	202007	충북	화장품 및 방향제 소매업	4.344208e+08
1393	202007	충북	휴양콘도 운영업	4.550172e+07

1394 rows × 4 columns

예측한 매출액의 exp()**를 취하여 "원" scale의 매출액** DataFrame**을 계산**

last_submission_exception.csv



STEP #3 CONCLUSION

일반적인 딥러닝, 머신러닝의 접근법은 적절한 validation 세트가 존재할 때 큰 위력을 발휘한다. 하지만, 이번 대회에서는 4월 데이터와 7월 데이터 사이의 큰 간극이 있어 적절한 validation을 할 수 없었다.

또한 비식별조치로 데이터가 삭제된 특수한 조건 하에서의 예측이라는 어려움이 있었다. 우리 팀은 적절한 데이터 관찰을 통해 feature를 생성하였고, 이는 타 팀보다 실제 데이터에 맞는 예측을 가능하게 하였다.

따라서 우리는 코로나19라는 특수한 상황속에서 업종을 클러스터링하는 새로운 접근법을 제시한다. 이는 통계적 이론과 머신러닝 기법, 그리고 특수 상황을 잘 반영할 수 있는 다양한 아이디어에서 출발한다.

THANK
YOU

JEJU BIG DATA COMPETITON

AI 알고리즘 활용 카드 사용량 예측

총상금 600만원

대회 기간 6월 22일 ~ 7월 31일

참가 방법 https://dacon.io

Jeju JTP 제주테크노파크

DACON