Fast campus × upstage

Upstage AI Lab

ML 프로젝트 3조
: Predict Grocery Sales with Machine Learning

24.04.17

www.fastcampus.co.kr

목차

01

팀원 소개

Upstage AI Lab

모든일을 즐기면서 최선을 다하자!!

조용중

Interested in

- 창업
- Micro SaaS Business
- ML/DL 을 사업 아이템에 적용하기

Introduction

- 기계공학 석사
- 다수 국책 사업 Management
- C, DB, System Architecture

Role

- 전반적인 것을 따로 또 같이 함
: Preprocessing(EDA), Modeling(Features Eng., ML), Post processing (Streamlit)

In Upstage AI Lab

- 오랜기간 멀어졌던 분야에서의 예전 감각을 다시 살

이윤재: the present is the only thing that has no end

Intereste in

- 아직 배우는 중입니다

Introduction

- 경영학
- SI, 금융업
- Web(js), 하둡, 텍스트마이닝(R), RecoSys

Role

- EDA, feature engineering, 모델 생성
- Streamlit

In Upstage AI Lab

- kaggle 데이터 분석을 좀 더 도전해보고 싶어요.
- 남은 딥러닝 공부도 열심히 하겠습니다.

Upstage AI Lab

내가 하고 싶은 일을 하면서 열심히 하자!

Intereste in

- 스마트팜
- 기업분석

Role

- 데이터 전처리, 전반적인 일 보조

Introduction

- 경영학
- 프롬프트 엔지니어 공부

In Upstage AI Lab

- 머신러닝을 공부해 기업분석, 스마트팜에 적용할 수 있으면 좋겠습니다.
- 딥러닝, 머신러닝에 대한 지식 획득

02

프로젝트 주제 소개

Upstage AI Lab

프로젝트 소개

: ML 프로젝트

주제

Grocery Data와 머신러닝 기법을 활용하여 매장 판매량 예측

목표

목표

에콰도르의 가게,제품, 지진 등의
데이터들을 가공, 분석하여 각 요소 간의
상관관계를 분석하고, 머신러닝 알고리즘을
통해 앞으로의 판매량을 예측한다.

개요

소개 및 배경 설명

Kaggle의 에콰도르 데이터들을 활용하여
앞으로의 판매량을 예측하는 머신러닝 모델 개발

기간

2024. 06. 03 ~ 2024. 06.17

Upstage AI Lab

프로젝트 진행 방법

: ML 프로젝트

팀원 소개 : 조용중, 이윤재, 이승민
스크럼 진행 횟수 및 일정:
매일 1회 1시간 씩 정기적인 미팅을 하며 진행 내용을 공유하였습니다.
프로젝트 진행 장소 : 실시간 비대면 (ZOOM)
프로젝트 진행 방법 : 자율 학습법 선택하여 진행
프로젝트 진행 시 생긴 문제점 :
모델 적용 train셋 범위, test 셋 모델예측 과정에서 질문
문제 해결 방법 : 강사님의 도움 및 가이드라인의 도움을 통해 해결

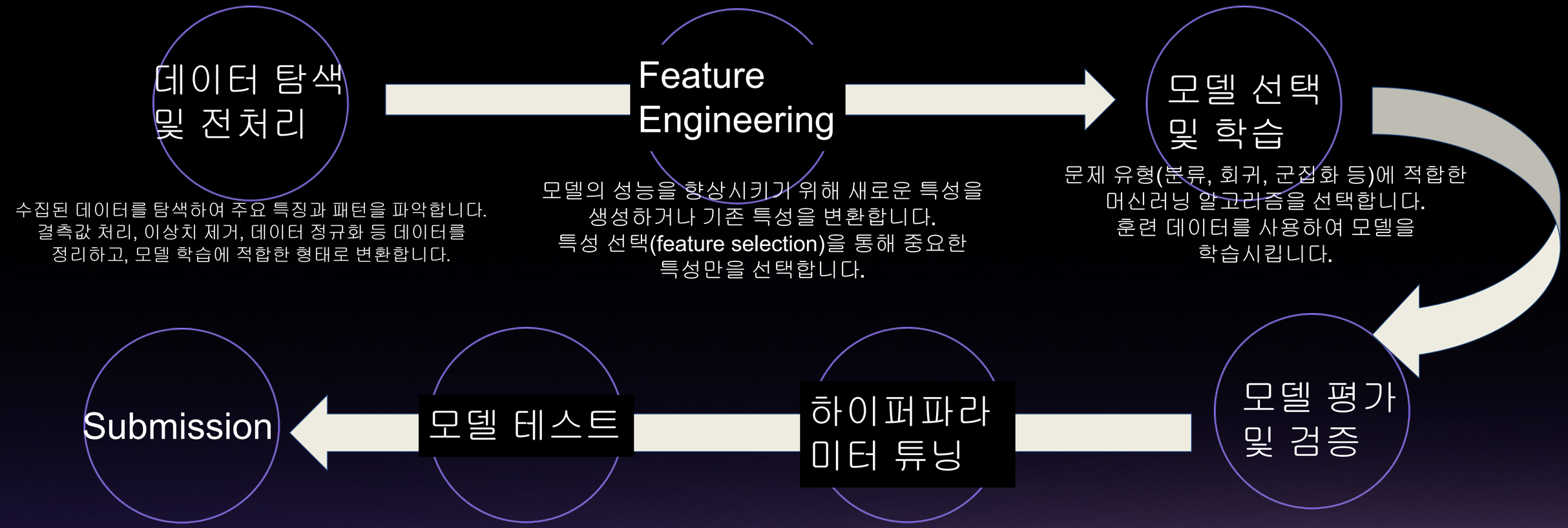
03

이론 및 데이터 설명

Upstage AI Lab

이론 및 개념 설명

: ML 프로젝트



Upstage AI Lab

데이터 설명

: Data Structures

train.csv: *특정 매장에서 특정 날짜에 판매된 제품군의 판매 데이터를 포함하는 훈련 데이터.*
'store_nbr'은 판매된 매장을, 'family'는 판매된 제품 유형을, 'sales'는 제품군의 총 판매량을, 'onpromotion'은 해당 날짜에 프로모션된 제품 수를 나타냅니다.

stores.csv: *도시, 주, 유형, 클러스터를 포함한 매장 메타데이터*

holidays_events.csv: *휴일 및 이벤트와 메타데이터. 정부에 의해 다른 날짜로 이동된 휴일은 'transferred' 컬럼에서 확인할 수 있습니다. 'Bridge' 유형의 날은 휴일을 연장하기 위해 추가된 날이며, 'Work Day' 유형의 날은 통상적인 근무일이 아닌 날을 의미합니다.*

oil.csv: *훈련 및 테스트 데이터 기간 동안의 일일 오일 가격. (에콰도르는 오일 가격 변동에 매우 취약한 오일 의존 국가입니다.)*

test.csv: *훈련 데이터와 동일한 특성을 가진 테스트 데이터. 이 파일의 날짜에 대한 판매량을 예측합니다.*

sample_submission.csv: *올바른 형식의 샘플 제출 파일..*

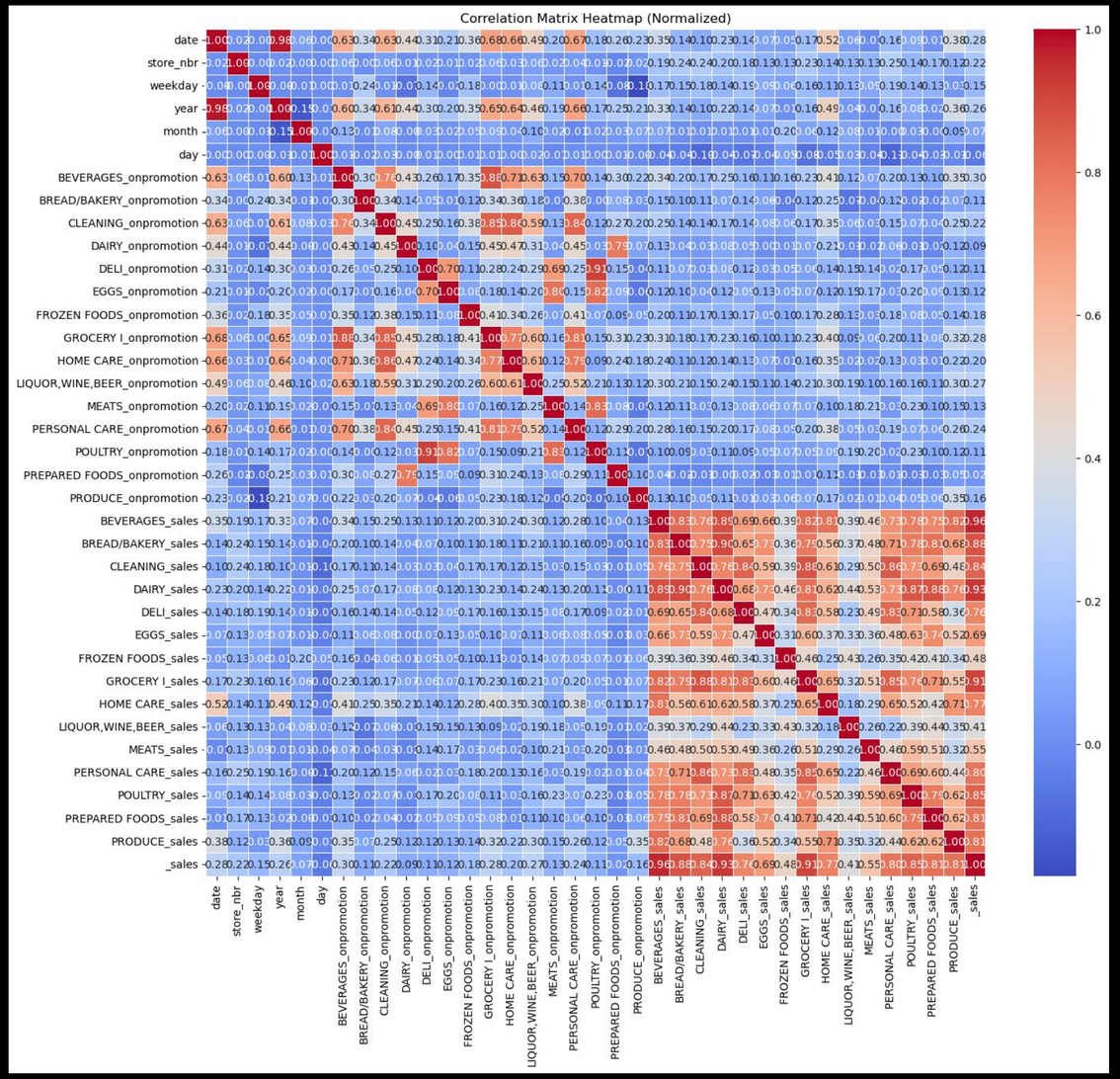
Upstage AI Lab

데이터 설명 : EDA

: train.csv

분석

train 데이터 안의 날짜별 상품군과 프로모션 상품 수에 따른 상관 계수 분석



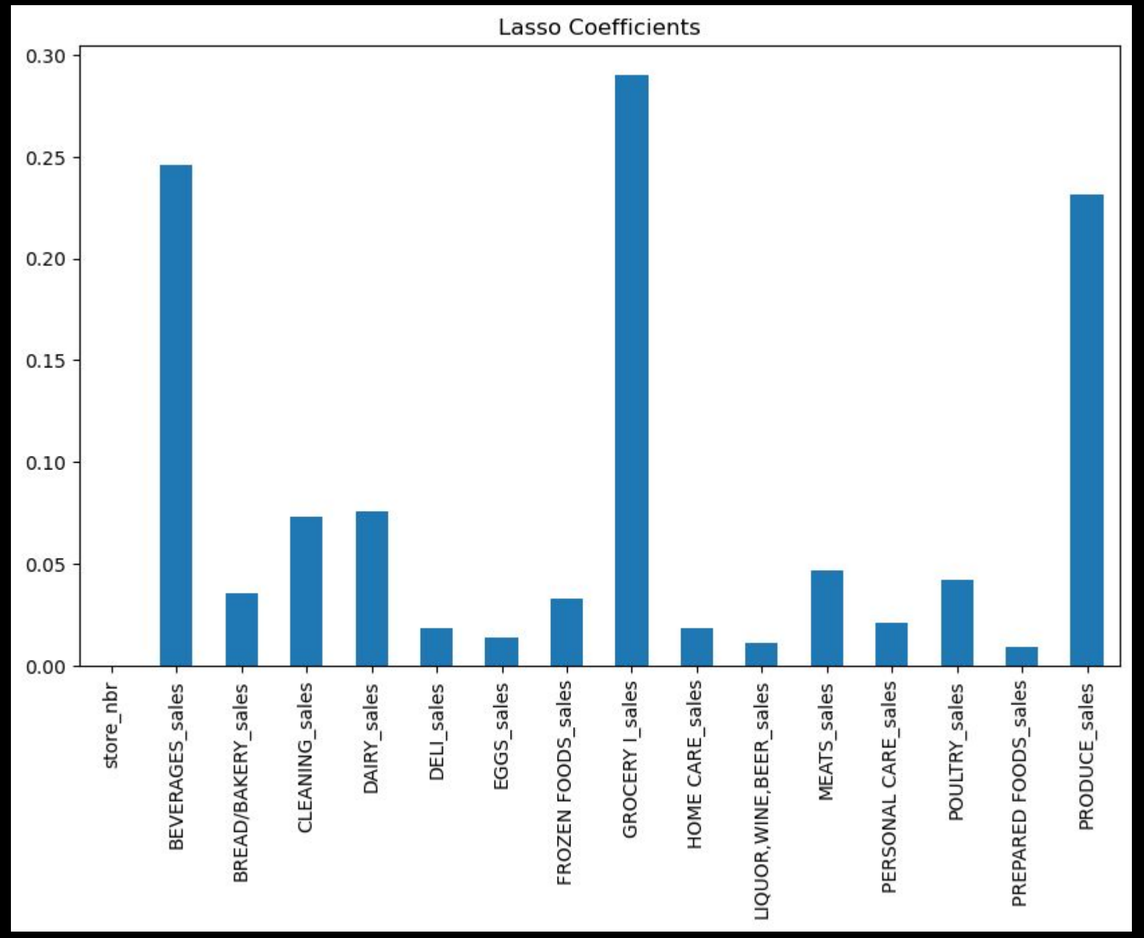
Upstage AI Lab

데이터 설명 : EDA

: train.csv

분석

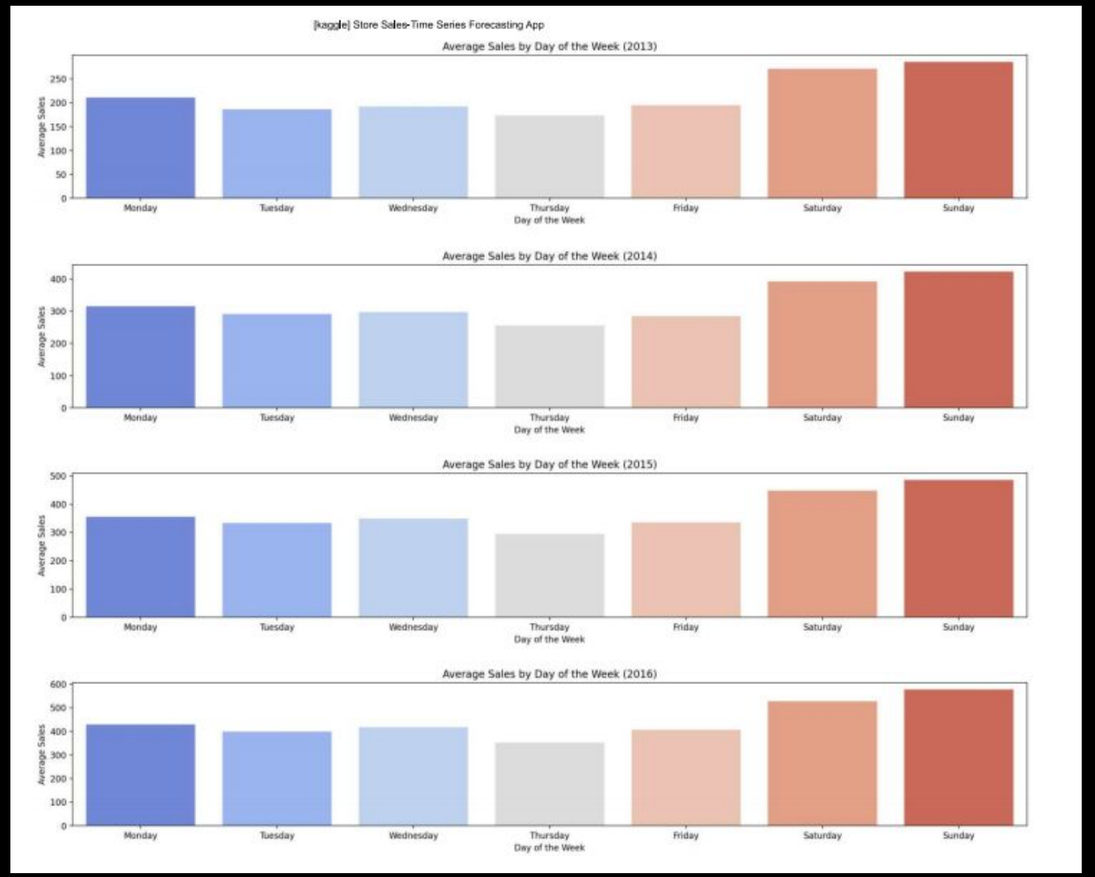
train 상품군 별 Sales와의 Lasso coefficients



데이터 설명 : EDA

: Days of week → 상점마다 요일별 뚜렷한 매출 trend를 보임

요일별 가장 많은 판매량을 .나타내는 그래프.
2013~2017년까지의 각 요일별 평균 판매량을 나타냄



Upstage AI Lab

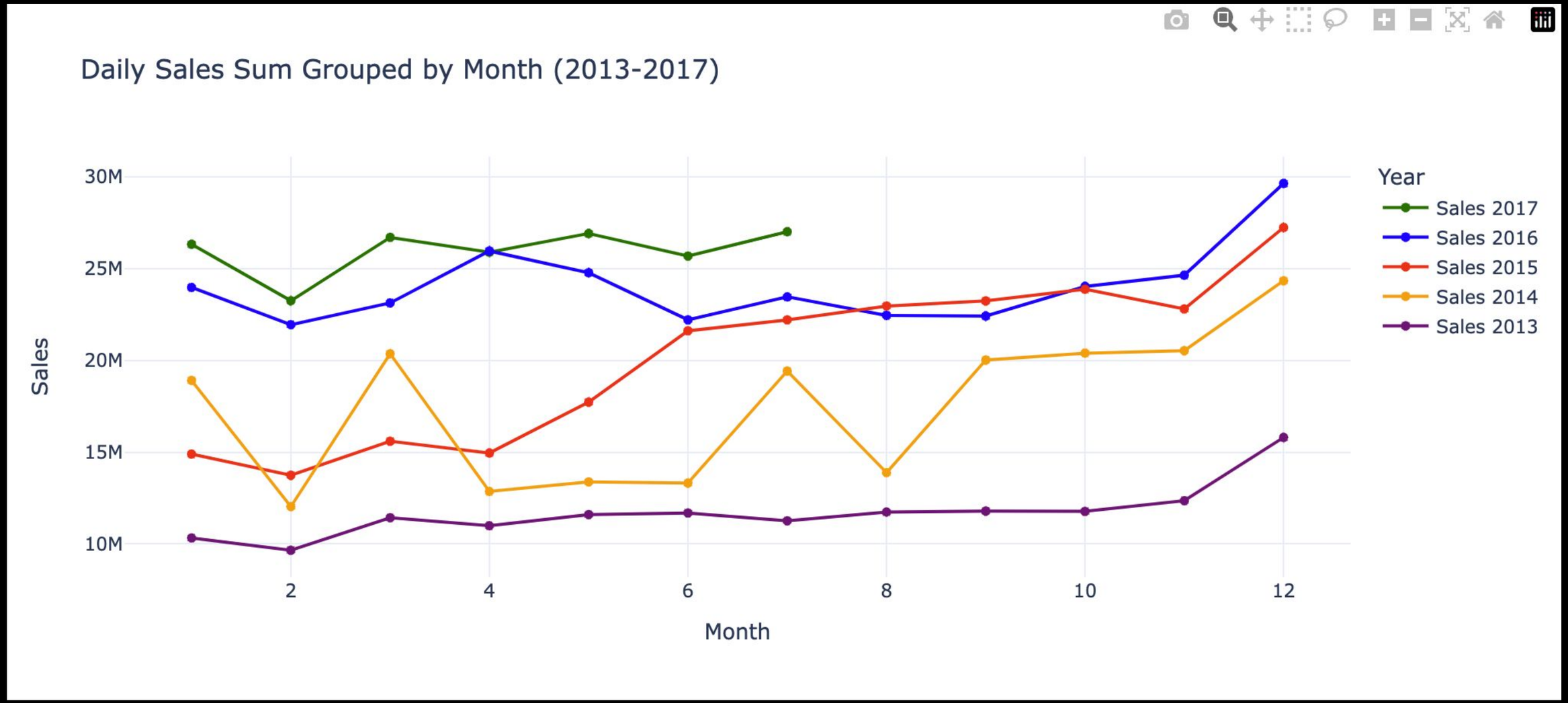
데이터 설명 : EDA

: 월별 transactions

Upstage AI Lab

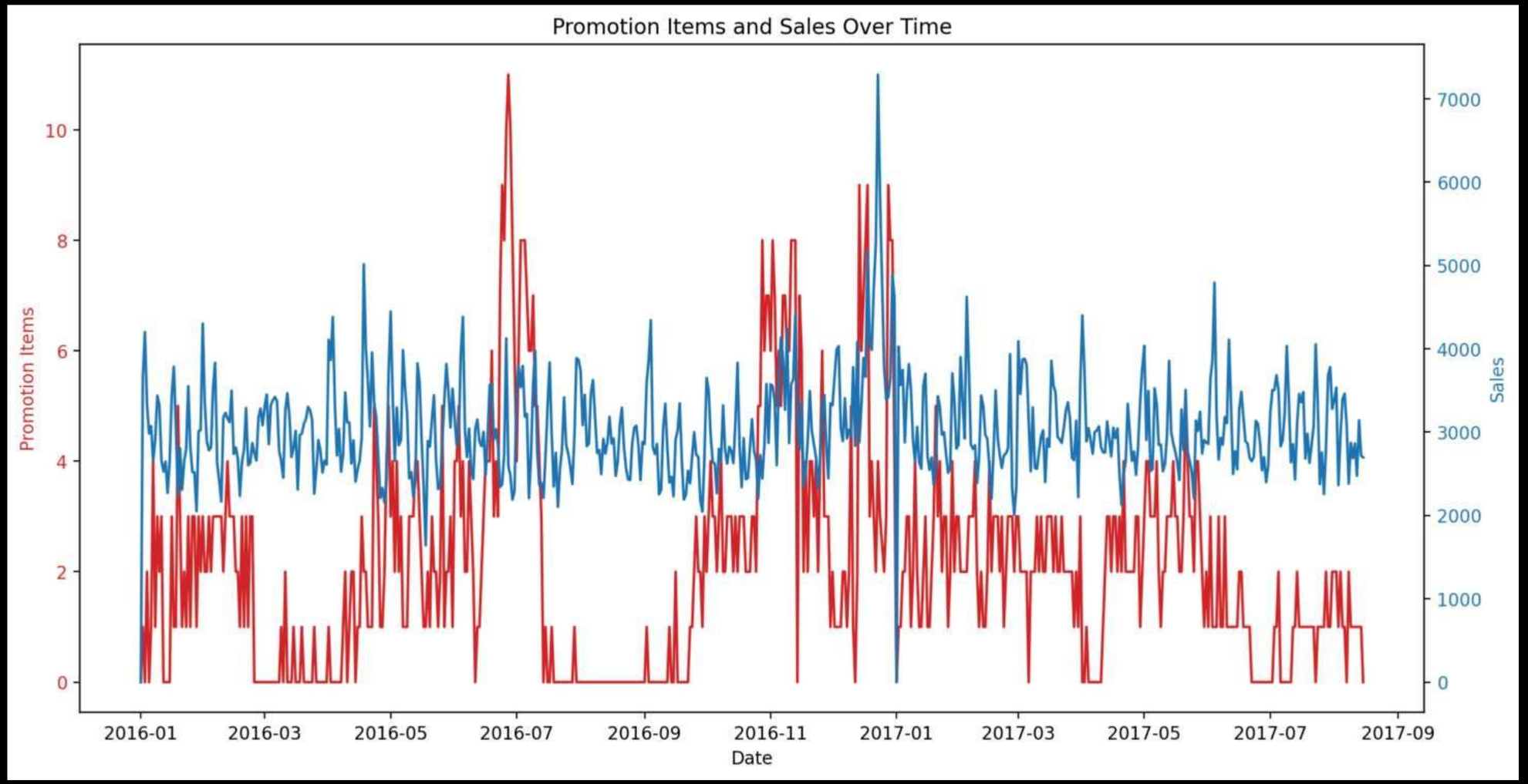
데이터 설명 : EDA

: 월별 transactions을 연도별로 비교 → oil 영향이 있는 것으로 판단



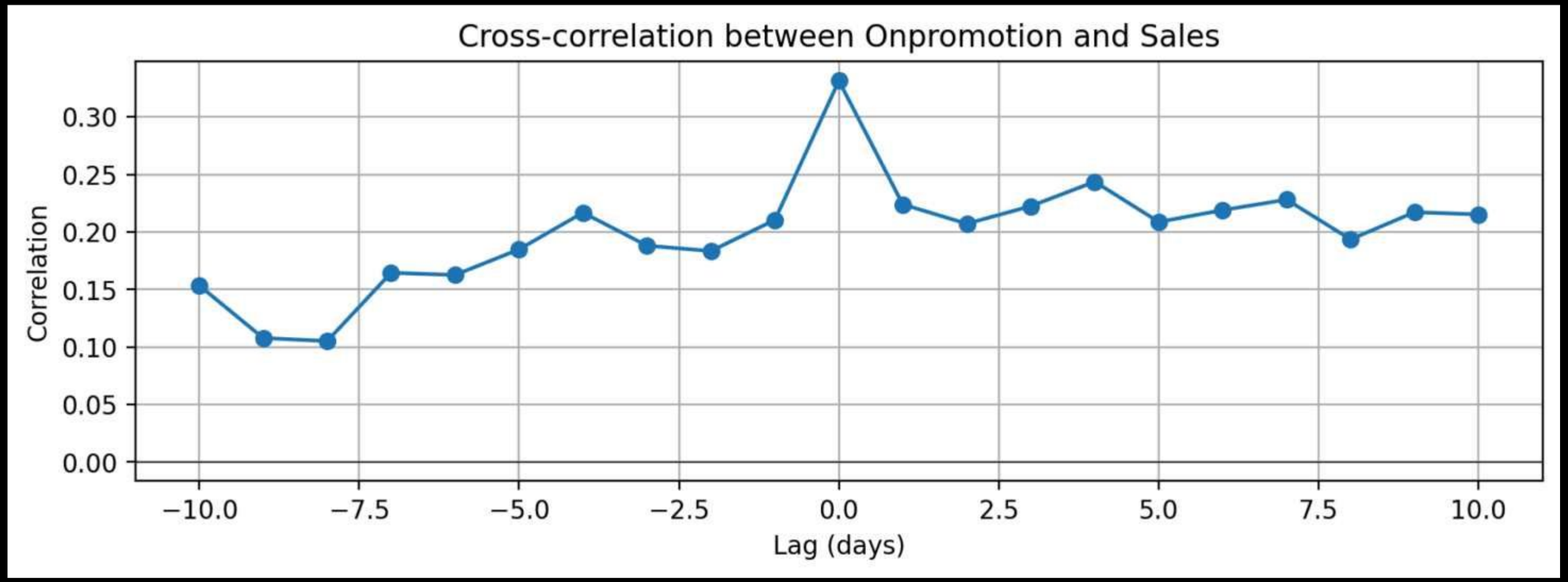
데이터 설명 : EDA

: Promotion Items and Sales Over Time → onpromotion 에 sales가 즉각적 반응을 나타내지는 않음



데이터 설명 : EDA

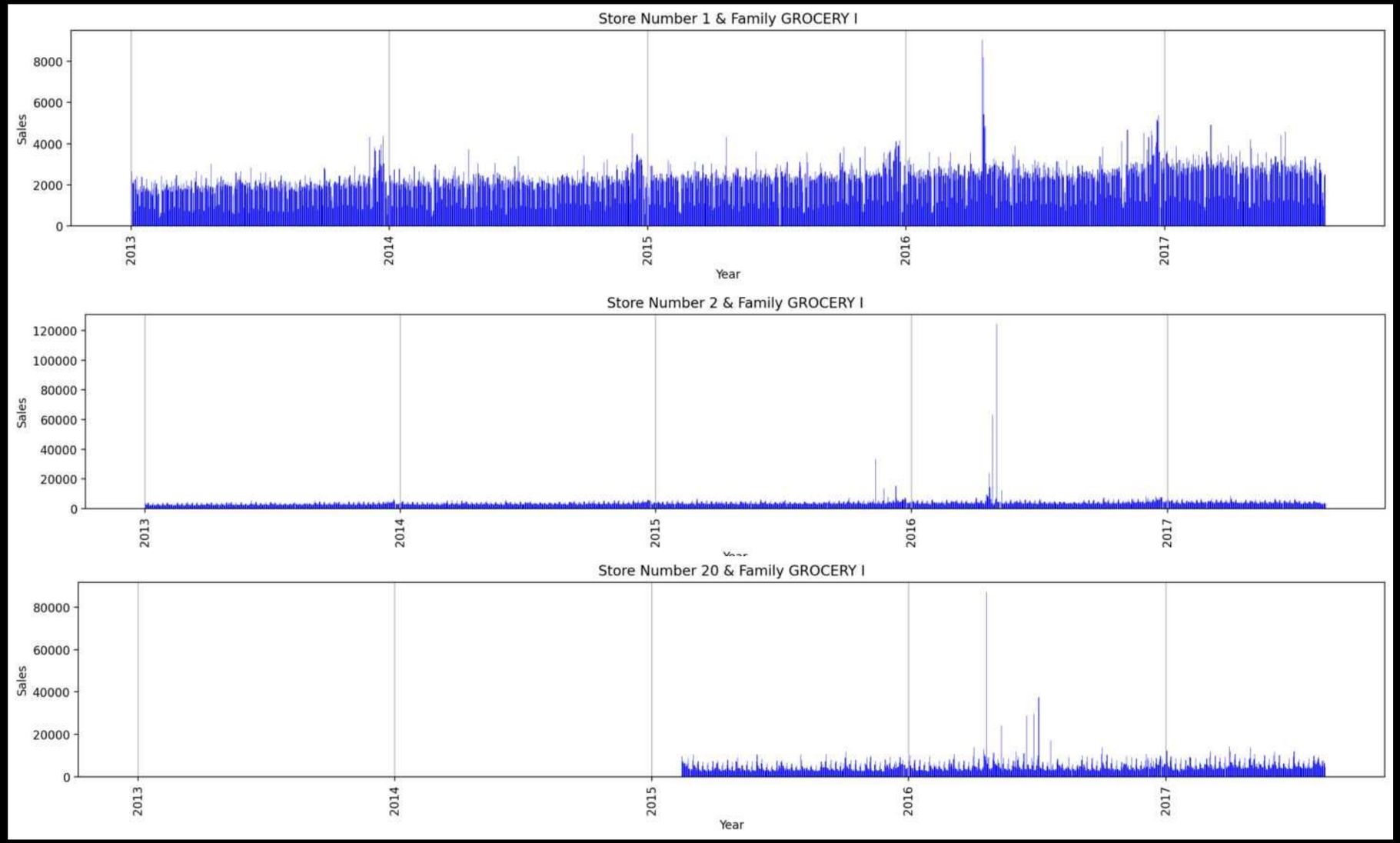
: Sales와 프로모션 lagging 의 상관 관계 분석 → promotion +1day 에 효과가 있는 것으로 판단



Upstage AI Lab

데이터 설명 : EDA

: 상점 별 상품군의 연도별 판매 데이터 분석 → family, store_nbr 마다 다른 sales trend



데이터 설명 : EDA

: 결측치 확인 → transactions가 없는데 sales가 있는 경우 존재하여 결측치 메꿈

0	2013-01-01 00:00:00	5	0
1	2013-01-02 00:00:00	5	1,903
2	2013-01-03 00:00:00	5	1,740
3	2013-01-04 00:00:00	5	1,642
4	2013-01-05 00:00:00	5	1,643
5	2013-01-06 00:00:00	5	1,754
6	2013-01-07 00:00:00	5	1,577
7	2013-01-08 00:00:00	5	1,504
8	2013-01-09 00:00:00	5	1,513
9	2013-01-10 00:00:00	5	1,449

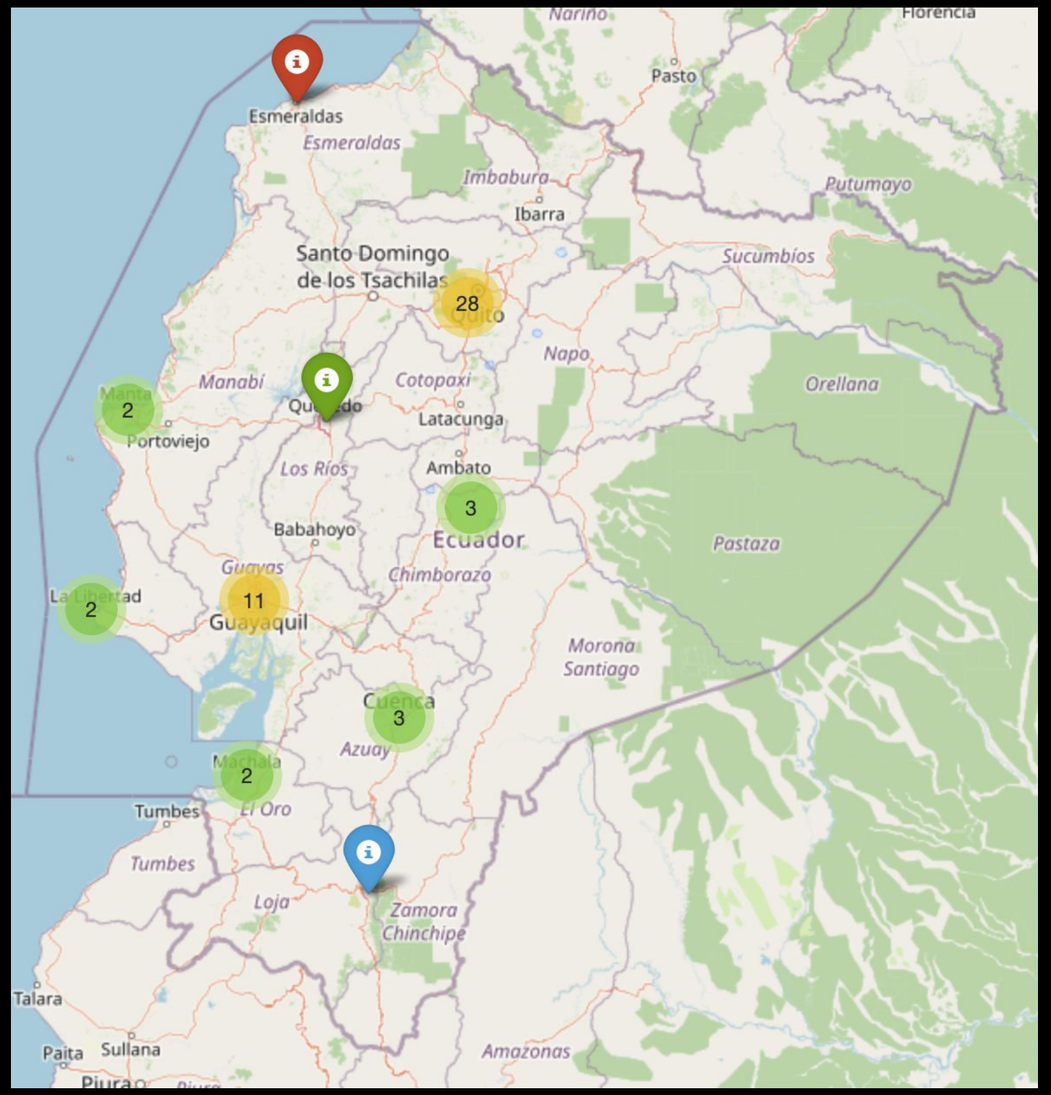
Upstage AI Lab

데이터 설명 : EDA

: 지진여파가 큰 2주치 데이터를 기존 sales와 transactions와 비레하여 대체함

분석

GEOPandas를 활용한 상점들의 위치 표시



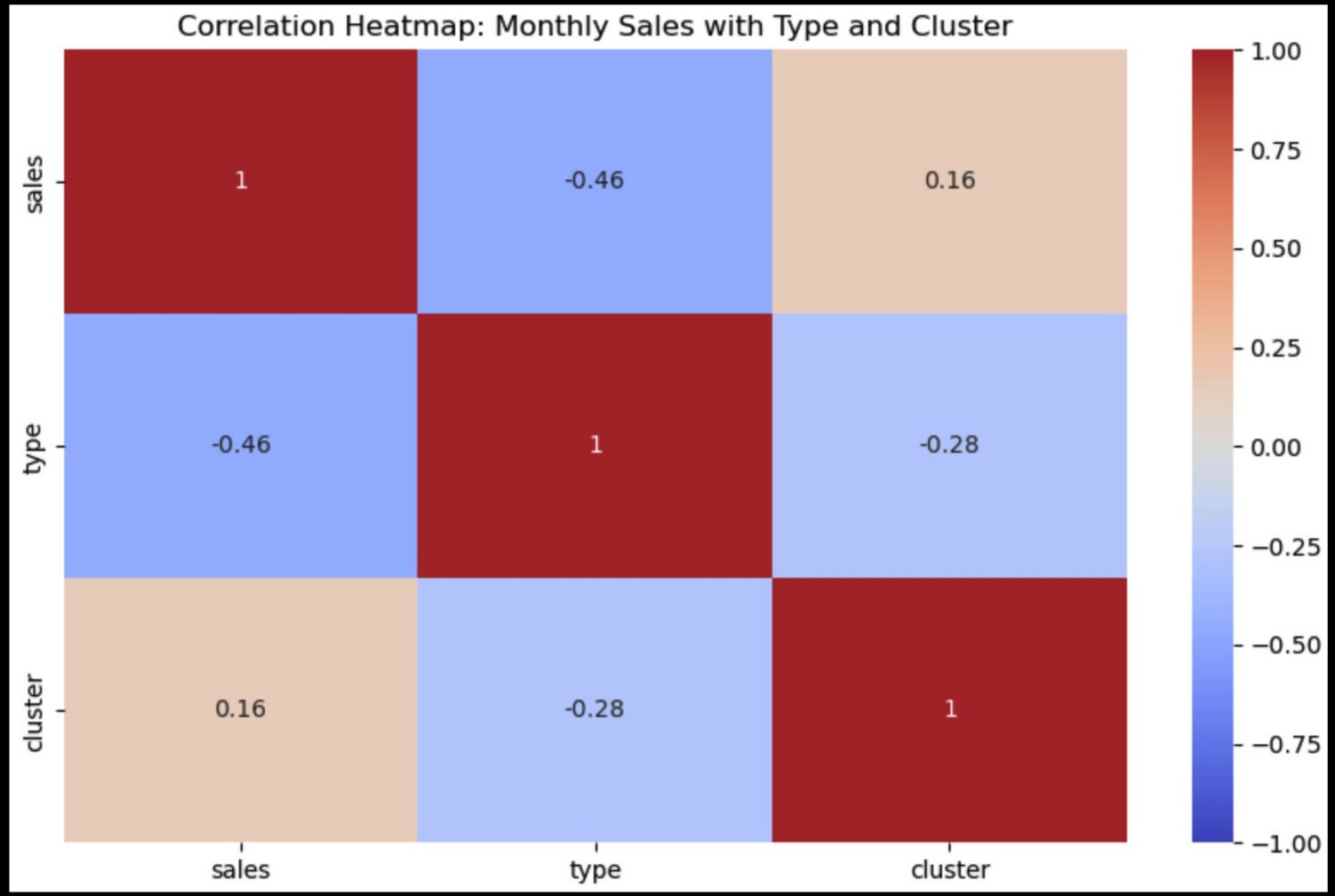
Upstage AI Lab

데이터 설명 : EDA

: cluster의 경우 큰 의미가 없는 것으로 판단함

분석

상점별 cluster 와 type 간의 히트맵(상관계수)



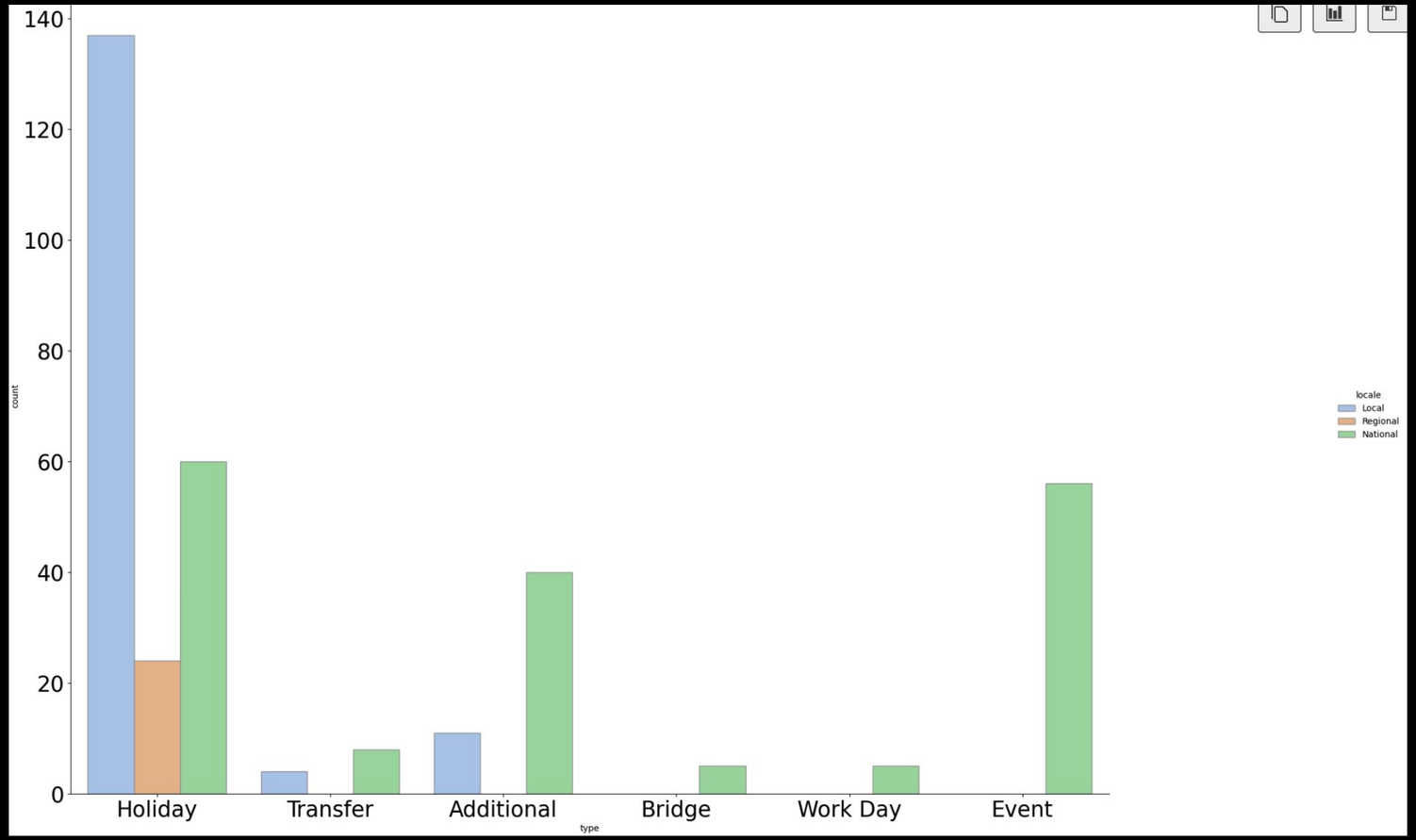
Upstage AI Lab

데이터 설명 : EDA

: holiday를 분할하여 살펴봄

분석

휴일의 Locale 별 구성



Upstage AI Lab

전처리 및 모델 학습(1)

:modeling_for_submission.ipynb

```
for index, row in target_temp.iterrows():
    matching_source_temp = source_temp[
        (source_temp['store_nbr'] == row['store_nbr']) &
        (source_temp['family'] == row['family']) &
        (source_temp['day_of_week'] == row['day_of_week'])
    ]

    if not matching_source_temp.empty:
        source_row = matching_source_temp.iloc[0]

        # 트랜잭션 비례로 sales 계산
        transaction_ratio = row['transactions'] / source_row['transactions']
        adjusted_sales = source_row['sales'] * transaction_ratio
        new_sales.append(adjusted_sales)
    else:
        new_sales.append(row['sales'])  # 매칭되는 소스 데이터가 없을 경우 원래 sales 값을 사용
```

데이터 대체

목적: *특정 기간의 매출 데이터를 대체하여 일관된 분석을 할 수 있도록 함.*
데이터 필터링: *target_temp와 source_temp는 각각 대체하려는 기간과 기준이 되는 기간의 데이터셋입니다.*
매칭 기준: *store_nbr, family, day_of_week가 같은 행을 찾아서 대체.*
트랜잭션 비례 대체: *트랜잭션 비율에 따라 매출을 조정합니다.*
결과: *대체된 매출 데이터를 new_sales 리스트에 저장하고, 이를 target_temp의 매출 데이터로 업데이트합니다.*

특성 엔지니어링

목적: 매출 데이터의 통계적 특징을 추출하여 모델 성능을 향상시킴.
7일과 14일의 이동 창을 사용하여 특성을 계산.
특성 종류:
slope: 기울기, 추세를 나타냄.
std: 표준편차, 데이터의 변동성을 나타냄.
mean: 평균값.
skew: 왜도, 데이터의 비대칭 정도.
kurt: 첨도, 데이터의 뾰족한 정도.
min, max: 최소값과 최대값.
결과: 각 창 크기별로 다양한 통계적 특성을 계산하여 데이터프레임에 추가합니다.

```
# get_slope 함수 정의
def get_slope(array):
    y = np.array(array)
    x = np.arange(len(y))
    slope, intercept, r_value, p_value, std_err = linregress(x, y)
    return slope


# get_rolling 함수 정의
def get_rolling(df):
    df['slope7'] = df['sales'].rolling(7).apply(get_slope, raw=True)
    df['std7'] = df['sales'].rolling(7).std(raw=True)
    df['mean7'] = df['sales'].rolling(7).mean(raw=True)
    df['skew7'] = df['sales'].rolling(7).skew()
    df['kurt7'] = df['sales'].rolling(7).kurt()
    df['min7'] = df['sales'].rolling(7).min()
    df['max7'] = df['sales'].rolling(7).max()

    df['slope14'] = df['sales'].rolling(14).apply(get_slope, raw=True)
    df['std14'] = df['sales'].rolling(14).std(raw=True)
    df['mean14'] = df['sales'].rolling(14).mean(raw=True)
    df['skew14'] = df['sales'].rolling(14).skew()
```

Upstage AI Lab

전처리 및 모델 학습(1)

:modeling_for_submission.ipynb

```
, 'mean7', 'skew7', 'kurt7', 'min7', 'max7', 's

_size=0.2, random_state=42)
```

모델 학습

*특성 선택: **X**에는 예측에 사용될 다양한 특성들이 포함되고, **y**는 타겟인 매출 데이터입니다.*
*데이터 분할: 학습 데이터와 테스트 데이터로 **80:20** 비율로 분할.*
*모델 선택: 랜덤 포레스트 회귀 모델(**RandomForestRegressor**) 사용.*
*모델 학습: **X_train**과 **y_train**을 사용하여 모델을 학습시킵니다*

예측 및 결과 저장

목적: 학습된 모델을 사용하여 테스트 데이터의 매출을 예측.
*예측 수행: **deepAR** 데이터셋에 대해 예측을 수행하여 **test_pred**에 저장.*
*결과 저장: 예측 결과를 **deepAR_origin** 데이터프레임에 추가하고, 모든 예측 결과를 **final_predictions**로 결합하여 **CSV** 파일로 저장.*

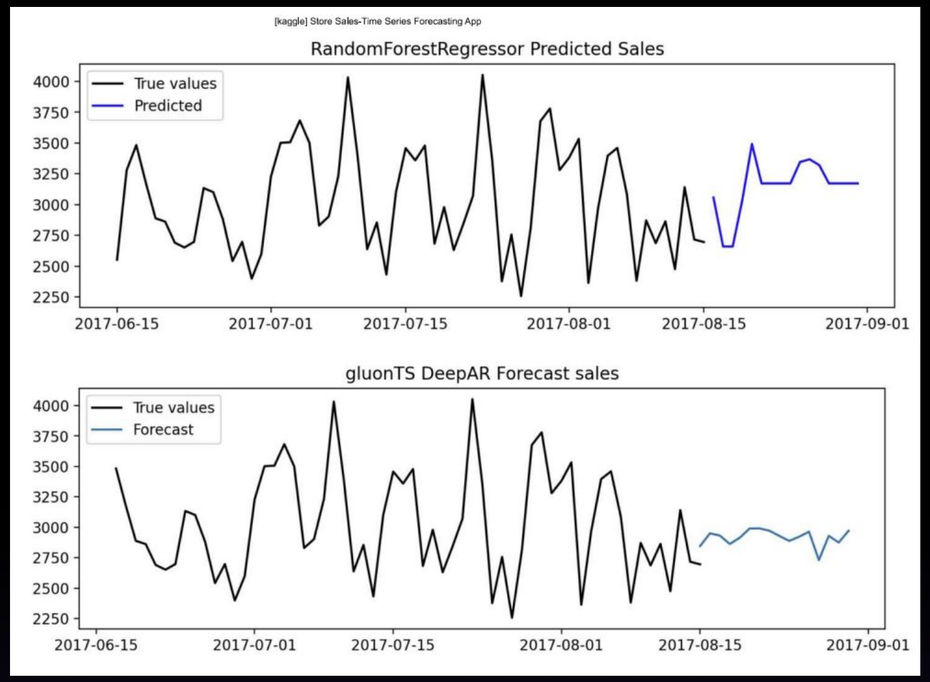
```
dataframe
test_pred


in)
############ end store: {store}, family: {family}##

ingle dataframe
ictions)
ctions_{store}.csv', index=False)
```

04

결과 및 인사이트 공유

실습 및 예시

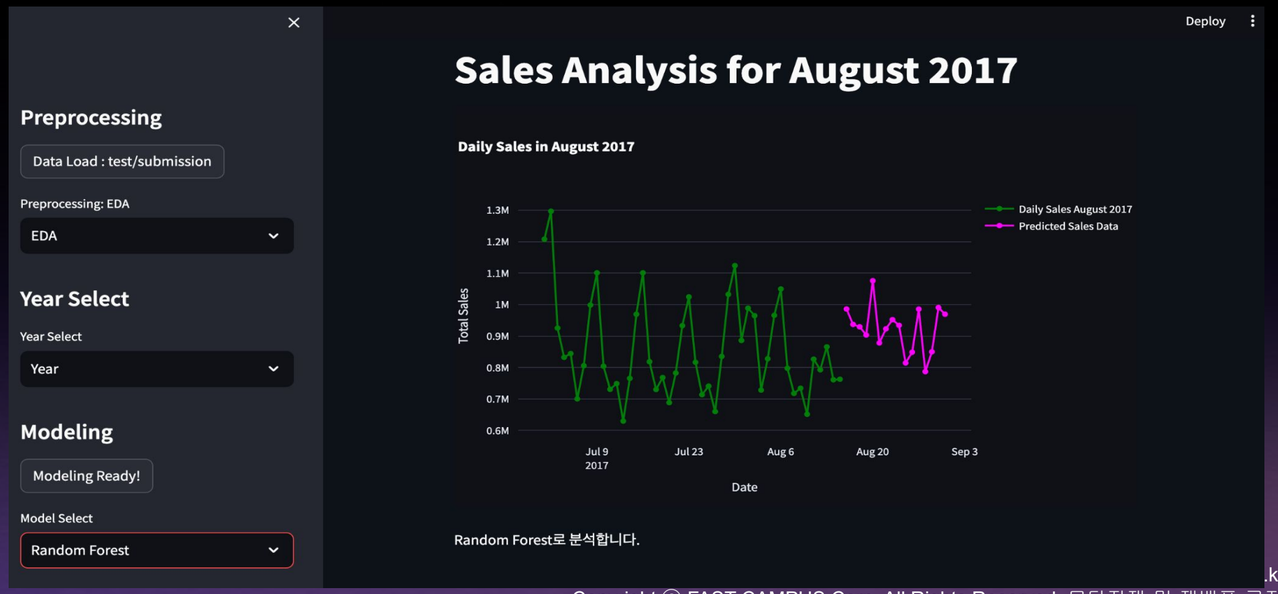
: ML 프로젝트



RandomForest 예측치

gluonTS DeepAR 예측치

스터디 진행 과정 및 학습 진행 상황에 대한 내용을 자유롭게 작성해주세요.
스터디 진행 과정 및 학습 진행 상황에 대한 내용을 자유롭게 작성해주세요.
스터디 진행 과정 및 학습 진행 상황에 대한 내용을 자유롭게 작성해주세요.



프로젝트 인사이트 공유

: ML 프로젝트

인사이트 1

***Colab GPU** 환경을 이용하려 했으나, 로컬과의 버전 차이와 환경 설정의 복잡성으로 인해 많은 문제가 발생했습니다.*
*특히 **gluonts** 모델에서는 해결하기 어려운 에러들이 발생하여 예측 모델링에 어려움을 겪었습니다.*
***Colab**에서 발생한 에러는 로컬에서는 재현되지 않는 문제들이었고, 결국 **GPU**를 활용하지 못했습니다.*

인사이트 2

*테스트 데이터셋에 중요한 **transactions** 컬럼이 없다는 사실을 프로젝트 마감 기한 직전에 알게 되었습니다.*
*이로 인해 **transactions** 값을 예측해야 했고, 빠르게 예측할 수 있는 모델을 사용하여 이 문제를 해결했습니다.*
***rolling** 변수 또한 **sales**를 기반으로 산출해야 하는데, 이를 예측하기 위해 **gluonts** 모델을 사용했습니다.*

인사이트 3

최종 예측 결과는 **RMSLE 0.47947**로, **1**등과 비교해 **26.94%**의 차이가 있었습니다.
이는 **transactions**와 **sales** 예측 부분을 더 고도화할 필요가 있음을 시사합니다.
앞으로 이 부분을 개선하여 예측 정확도를 높이는 방향으로 나아갈 계획입니다.

05

프로젝트 회고

프로젝트 진행 느낀점

: ML 프로젝트

Point 1

조용중

이유 : 초기에 우선 예측 데이터 구성을 확인하여 시작할것. 즉 프로젝트를 전체적으로 면밀하게 살피고 시작할것
향후 계획 : ML 알고리즘에 대한 이론 및 수학적 배경 지식을 좀더 쌓으면서 해야 겠다는 생각.

Point 1

이승민

이유 : 머신러닝에 대한 지식이 너무 부족해서 프로젝트에 많이 참여를 못했습니다.
향후 계획 : 나 혼자서 여러 머신러닝 프로젝트를 돌려가면서 프로젝트에 더 잘 참여할 수 있도록 하겠습니다.

Point 1

이윤재

이유 : 데이터셋에 대한 명확한 이해를 바탕으로 단계별 진행과정을 한 번 시도한 이후에 고도화하는 것이 좋겠다
향후 계획 : 다른 이들의 코드를 참고하며 다양한 feature 생성 방법을 알아두면 도움이 되겠다.

Upstage AI Lab

프로젝트 진행 소감

: ML 프로젝트

조용중

*처음 한 **ML** 프로젝트 답게 너무나 많은 시행착오를 했지만 결론적으로는 많은 것을 얻을수 있는 기회였다고 생각합니다. 다음 프로젝트는 좀 더 잘할수 있을 것이라는 자신감을 얻는 기회 이기도 합니다.*

이승민

*처음으로 진행한 **ML** 프로젝트가 너무나도 어려워서 제대로 참여를 하지 못해서 많이 아쉬웠습니다. 다음에 더 잘할 수 있도록 혼자서 여러 머신러닝 프로젝트를 돌리며 공부를 할 생각입니다.*

이윤재

*첫 캐글 **submission**을 완료할 수 있어서 만족스럽습니다. 모델을 좀 더 고도화하여 제출하면 더 좋은 결과를 얻을 수 있을 것이라고 생각합니다. 앞으로도 **forecsting** 문제 외에도 다양한 분야의 문제들을 다뤄볼 예정입니다.*

Life-Changing Education

감사합니다.