



Favorita Store Sales



Use machine learning to predict grocery sales

미니 프로젝트 :
Python을 활용한 통계 분석 및 웹서비스 구현

1. 프로젝트 개요

- 대회 소개
- 평가지표
- 데이터 소개
- 스트리밍 대시보드 소개

2. 프로젝트 팀 구성 및 역할

- 팀 구성 및 역할

3. 프로젝트 수행 절차 및 방법

- 프로세스 소개
- 개발환경 소개

4. 통계 분석 & 머신러닝

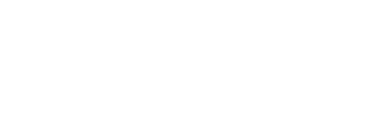
- 시계열 데이터 분석 방법
- 데이터 전처리 과정
- 탐색적 자료 분석
- 모델 선택 및 구현
- 결과 평가 및 검증

5. 대시 보드 웹 서비스 구현

- 대시보드 시각화 및 해석
- 대시보드 구현 영상

6. 자체 평가

- 자체 평가 및 문제점
- 개선점과 발전 가능성





Project 1 개요

대회소개 | 평가 지표 | 데이터소개 | 대시보드 소개



캐글 대회



대회 목표

- ✓ 시계열 알고리즘 사용
- ✓ 매출 변화 추이 분석
- ✓ 수요 예측 판매 모델링

기대 효과
최적의 모델을 찾아 매출 이익 극대화

RMSLE는 회귀 모델에서 예측 값과 실제 값의 평균 차이를 측정한 것으로 예측 모델의 정확도를 나타냅니다.

$$\text{RMSLE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\log(p_i + 1) - \log(y_i + 1))^2}$$



1 데이터 소개

개요

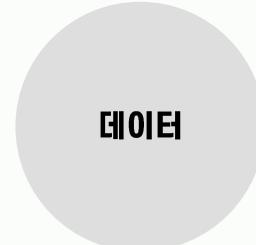
팀 구성 및 역할 | 수행 절차 및 방법 | 통계 분석 | 자체 평가

파일 설명 및 데이터 필드 정보

kaggle

Store Sales Data

에콰도르 전역 4년간 매출 데이터



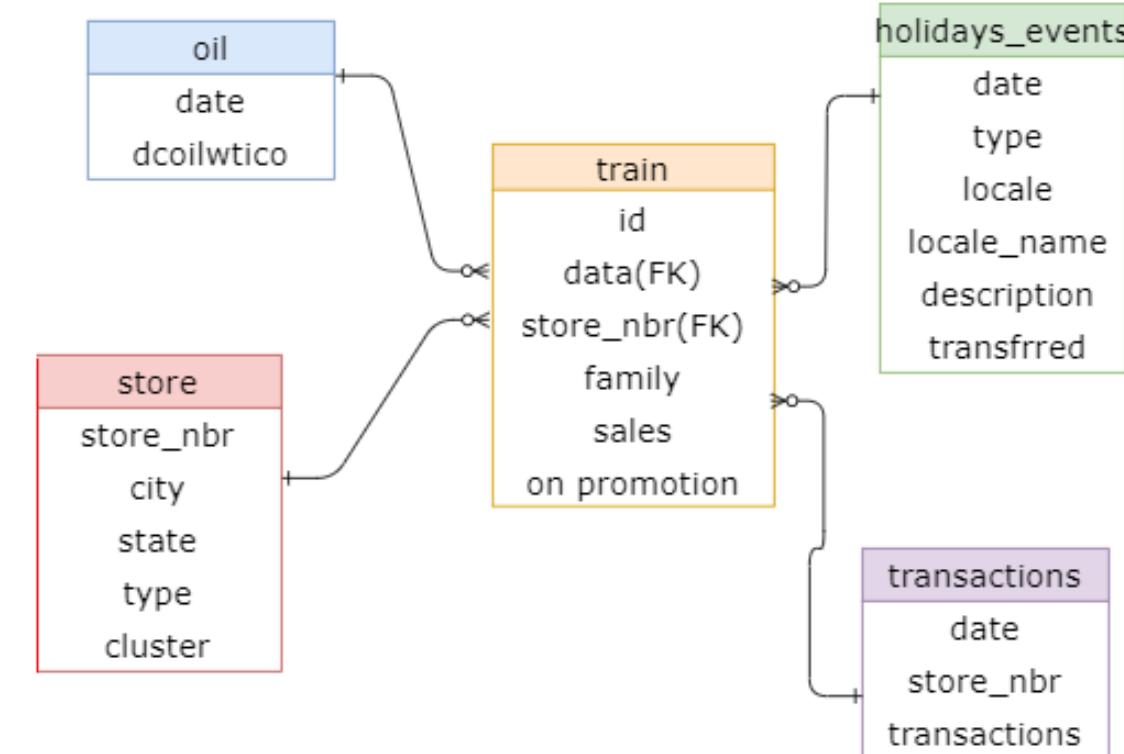
train

stores

oil

holidays

transactions



1 스트림릿 대시보드 소개

개요

팀 구성 및 역할 | 수행 절차 및 방법 | 통계 분석 | 자체 평가

웹 서비스 구현

Main Menu

Home

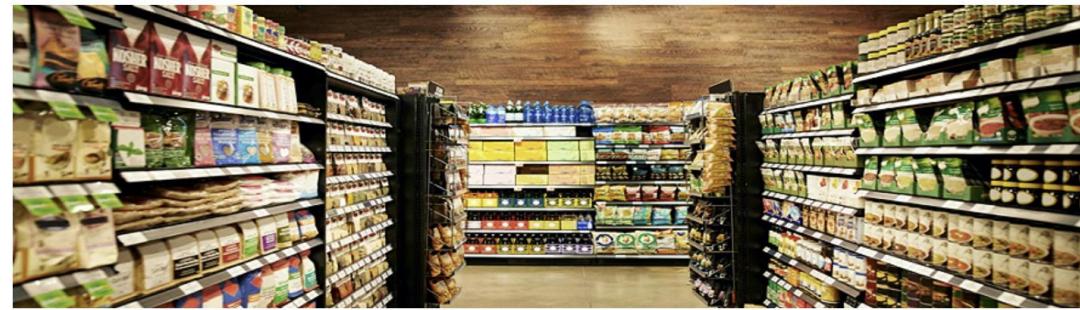
Description

Data

EDA

STAT

ML



☞ Store Sales - Time Series Forecasting

| Member | Skills | GitHub & Blog |
|-----------------|--------------------------|---|
| Jeong-An Choi | Analysis & Planning | GitHub : https://github.com/CHOIJEONGAN & TISTORY : https://choi-the-programmer.tistory.com/ |
| Jae-Myoung Choi | Analysis & Preprocessing | GitHub : https://github.com/ChoiJMS2m & TISTORY : https://james-choi88.tistory.com |
| Yong-Seok Kwon | Analysis & Dashboard | GitHub : https://github.com/MaestroYongseok & : https://blog.naver.com/maestrokwon78 |
| Yong-Jun Yoon | Dashboard & PPT | GitHub : https://github.com/yunjy1998 & : https://blog.naver.com/yunjy1998 |
| Geon-Yong Lee | Research & Dashboard | GitHub : https://github.com/leek1111 & : https://blog.naver.com/lgy2233 |

스트림릿(Streamlit)이란?

파이썬을 사용해 데이터를 활용해 빠르게
프로토타입 형태의 웹 앱을 구현할 수 있는 도구



미니 프로젝트 :

Python을 활용한 통계 분석 및 웹서비스 구현



Project 2 팀 구성 및 역할

2 프로젝트 팀 구성 및 역할

개요

팀 구성 및 역할

수행 절차 및 방법 | 수행 결과 | 자체 평가

202팀



최정안 팀장
분석/기획



권용석
발표/대시보드



윤용준
대시보드/ppt



이건웅
대시보드/영어번역

미니 프로젝트 :

Python을 활용한 통계 분석 및 웹서비스 구현



Project 3 수행 절차 및 방법

프로세스 소개 | 개발환경 소개 | 프로젝트 일정

3 프로세스 수행 절차

개요 | 팀 구성 및 역할 | **수행절차 방법** | 통계 분석 | 자체 평가

| 4/24 ~ 5/17 | 4월 4주차 | 5월 1주차 | 5월 2주차 | 5월 3주차 | 5월 16일 | 5월 17일 |
|-------------|----------|------------|--------|--------|--------|---------|
| 주제 분석 | 분석 방향 설정 | | | | | |
| 코드 분석 | 코드 필사 | 코드 분석, 시각화 | | 모델 평가 | | |
| 대시보드 | | 대시보드 제작 | | | 배포 | |
| PP T | | | PPT 제작 | | 최종 점검 | 발표 및 제출 |
| 발표 준비 | | | | | 발표 준비 | |

개발 환경 및 라이브러리 활용

파이썬 기반으로 데이터를 분석하고 머신러닝 코드를 분석하여 시각화를 진행하였고
통계분석 결과를 스트림릿을 이용하여 웹 서비스를 구현하였습니다.



Pandas



Pycharm



Python3.9



Streamlit



Colab



Project 4 통계&머신 러닝

시계열 데이터 분석 방법 | 데이터전처리 |
탐색적 자료분석 | 모델 선택 및 구현 | 결과 평가 및 검증

4 시계열 데이터 분석 방법

개요 | 팀 구성 및 역할 | 수행 절차 방법 | 통계 분석 | 자체 평가

시계열 데이터란?

시계열 데이터 정의

시계열 데이터란 시간적으로 연속된 데이터를 의미함
예) 주식 가격, 기온, 판매량 등

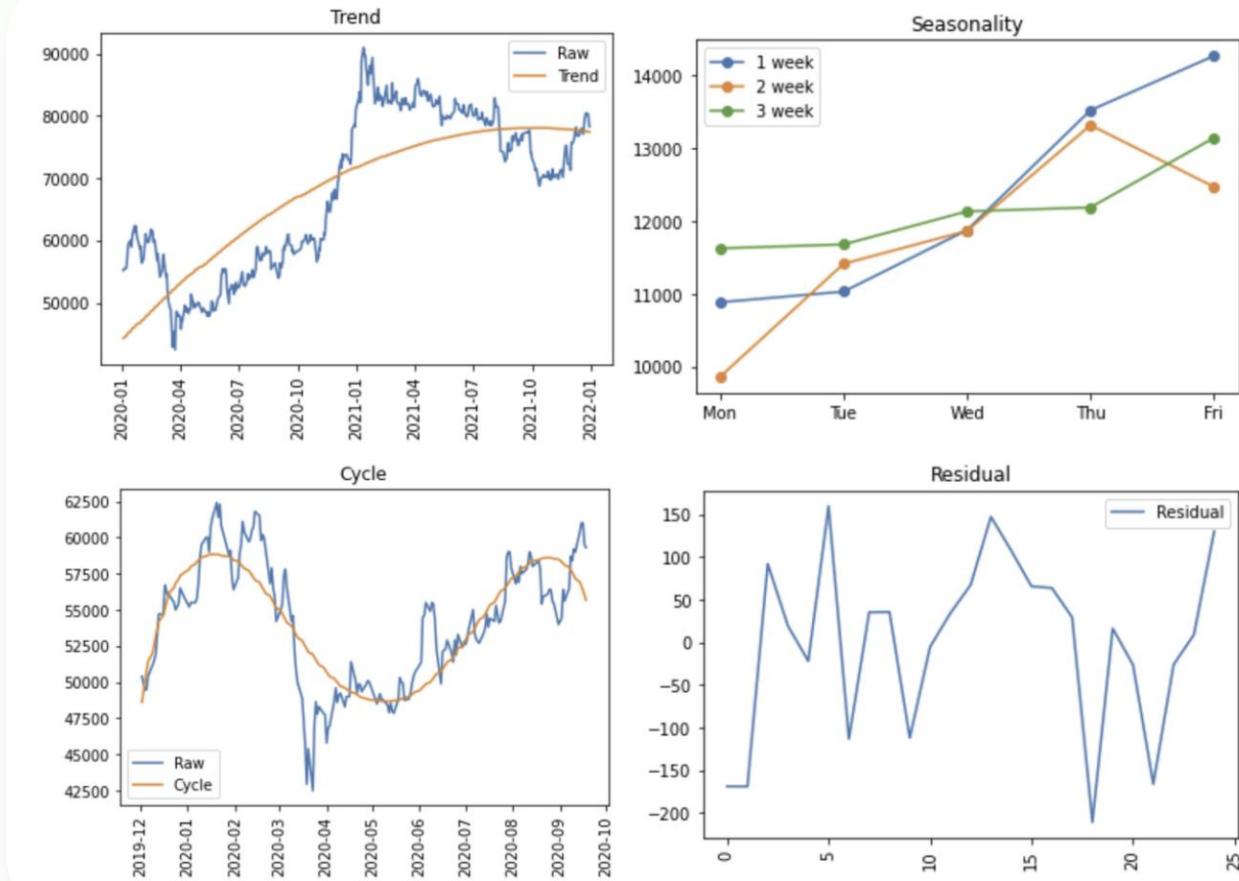
시계열 데이터 패턴

- ✓ Trend(추세)
- ✓ Seasonal component(계절성)
- ✓ Cyclical component(주기)
- ✓ Random component(불규칙)

시계열 데이터 분석 방법

- 이동평균 & 지수평활법
- ARIMA
- Prophet
- LSTM

과거 동향을 파악, 미래 수요 패턴을 예측하여
재고 최적화, 마케팅 전략 등
중요한 비즈니스 결정을 지원함



ADF(Augmented Dickey-Fuller)란?

ADF 정상성 검정

시계열 데이터의 정상성을 검정하는 통계적인 방법

정상성은 데이터가 시간에 따라 평균과 분산이 일정한 상태를 의미함

ADF 검정 과정

- ✓ 귀무가설과 대안가설 설정
- ✓ ADF 통계량 계산
- ✓ 임계값과 비교
- ✓ 정상성 여부 판단

ADF 검정을 통해 정상성이 확인된 데이터는
정확한 예측 모델을 구축해서
비즈니스 결정을 내리는데 도움을 줌

Augmented Dickey-Fuller Test Results for Oil Prices



H_0 : Series is non-stationary, or series has a unit root.

| index | result |
|-----------------------------|---------|
| Test Statistic | -1.9807 |
| p-value | 0.2951 |
| #lags used | 0 |
| number of observations used | 503 |
| critical value (1%) | -3.4434 |
| critical value (5%) | -2.8673 |
| critical value (10%) | -2.5698 |

사용한 라이브러리 소개



Numpy

- 과학적인 계산을 위한 다차원 배열과 함수를 제공하는 라이브러리

Pandas

- 데이터 조작과 분석을 위한 효과적인 데이터 구조와 도구를 제공하는 라이브러리

Matplotlib

데이터 시각화를 위한 다양한 그래프와 플롯을 생성하는 라이브러리

Statsmodels

통계 모델링과 추정을 위한 다양한 통계적 도구를 제공하는 라이브러리

Scikit-learn

머신러닝 알고리즘과 모델을 구현하고 평가하기 위한 라이브러리

Plotly

상호작용적인 시각화를 위한 다양한 그래프와 차트를 제공하는 라이브러리

탐색적 자료 분석 단계

1  데이터 수집 및 이해

2  데이터 정리 및 전처리

3  변수 분석

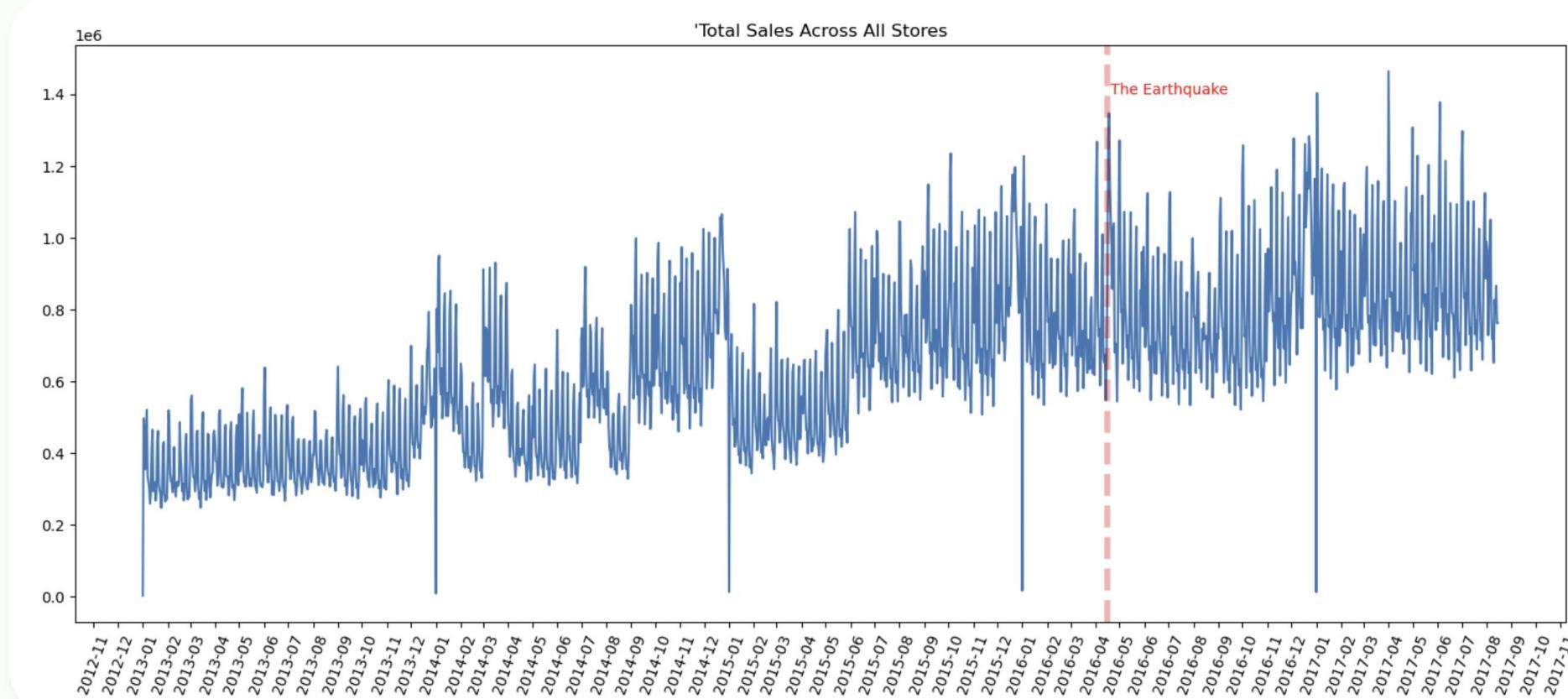
4  변수 간 관계 분석

5  시각화 및 탐색

6  결과 및 결론 도출



총 매출



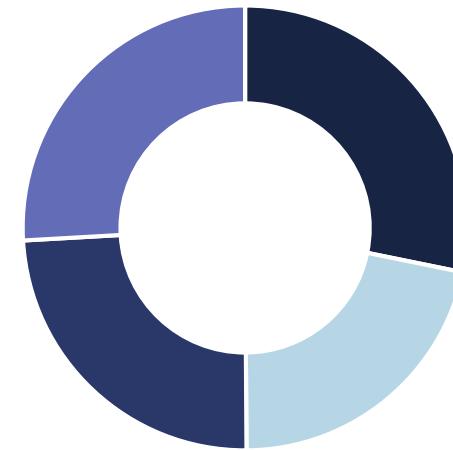
시계열 데이터에서 총매출은
비즈니스 전반적인 매출 성과를 파악하는 데에 중요한 역할

4 탐색적 자료 분석

개요 | 팀 구성 및 역할 | 수행 절차 방법 | 통계 분석 | 자체 평가

평균 매출 분석

분기별 평균 매출 분석

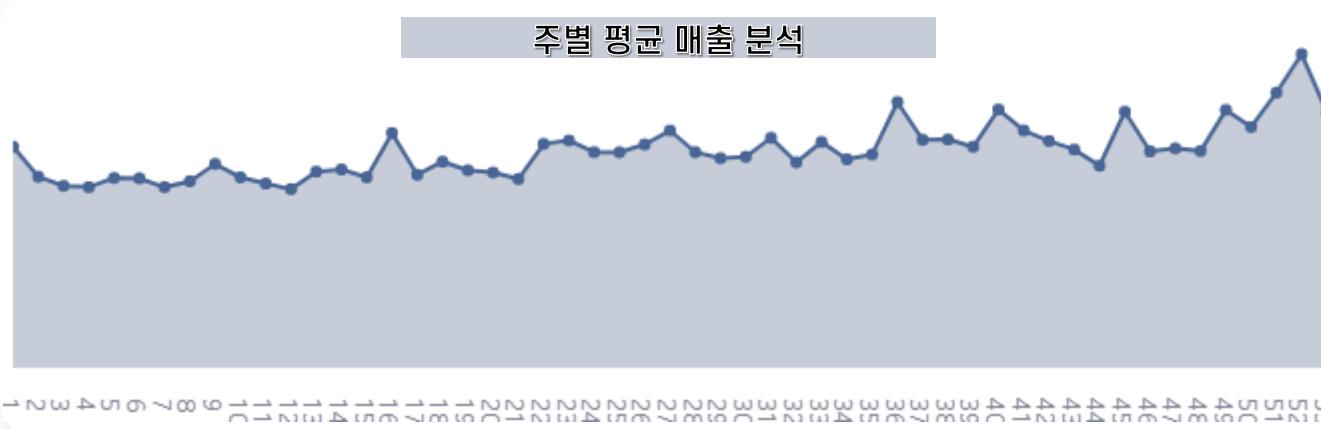


1분기 21.7%
2분기 24.2%
3분기 25.9%
4분기 28.2%

분기별 평균 매출 비교

- 가장 적은 매출 : 첫번째 분기
- 가장 높은 매출 : 마지막 분기

주별 평균 매출 분석



월별 평균 매출 분석



4 탐색적 자료 분석

개요 | 팀 구성 및 역할 | 수행 절차 방법 통계 분석 자체 평가

유가분석

Oil price trend during time



시간에 따른 석유 가격 추이 분석

1단계 : 2013년 1월 ~ 2014년 7월 => 안정된 추세

2단계 : 2014년 7월 ~ 2015년 1월 => 하락세

3단계 : 2015년 1월 ~ 2017년 7월 => 안정된 추세

결측치 처리

시계열 데이터의 패턴을 유지하기 위해 Backward fill 방식으로 처리

결측치 이후에 있는 첫번째 유효한 값으로 결측치를 대체하는 방법

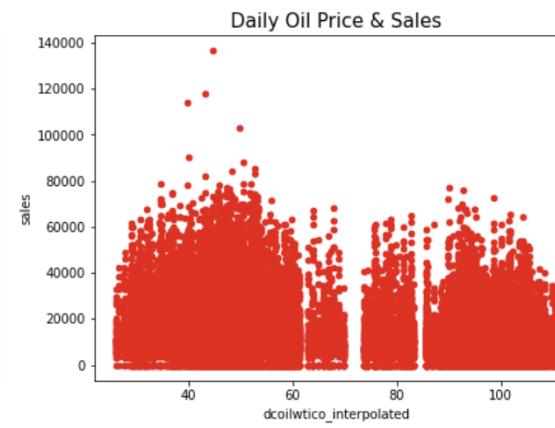
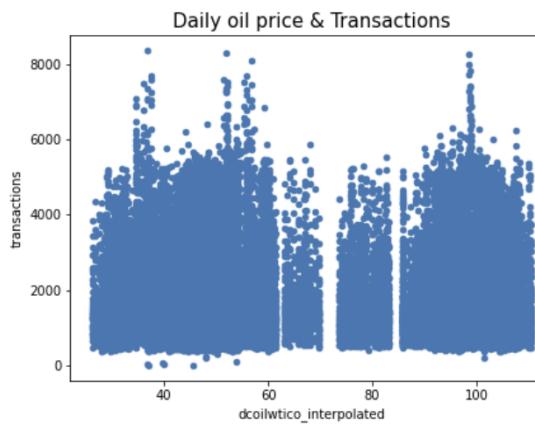
일일 유가와 매출 간 상관관계 분석

상관관계는 강하지 않지만, 매출이 음수

→ 일일 유가가 높을 때 예과도르 경제가 좋지 않을 것으로 예상

제품 가격 ↑ 판매량 ↓

→ 제품의 가격과 판매량은 음의 관계임을 알 수 있음



미니 프로젝트 :

Python을 활용한 통계 분석 및 웹서비스 구현

4 탐색적 자료 분석

개요 | 팀 구성 및 역할 | 수행 절차 방법 | 통계 분석 | 자체 평가

매장별 지진 영향 분석

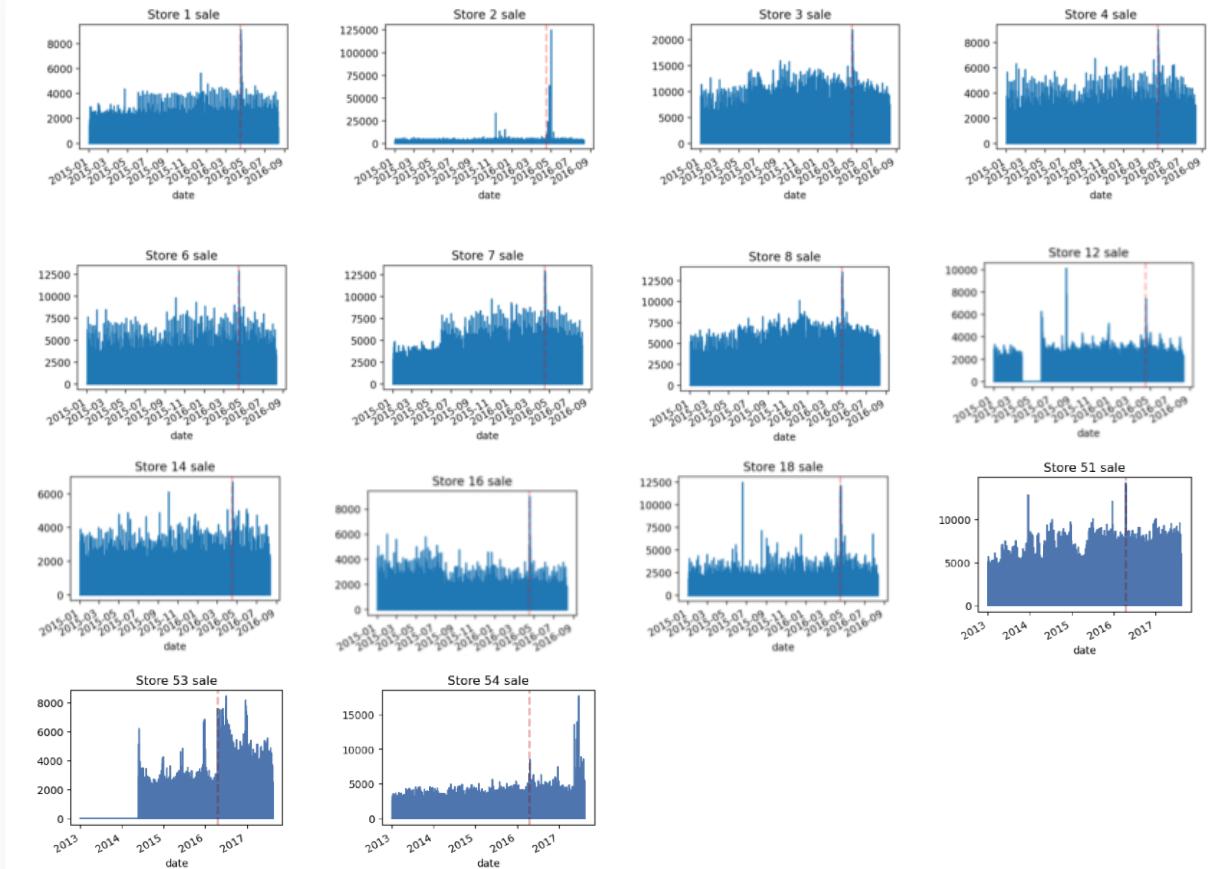
☞ 매출 영향 지역

- 1 Quito ➔ 1, 2, 3, 4, 6, 7, 8번 매장
- 2 Guayaquil ➔ 12, 14, 16, 18번 매장
- 3 Manta ➔ 51, 53, 54번 매장

☞ 지진 후 매출이 급등한 매장의 원인으로 추측 가능한 것

- ✓ 구호활동으로 인한 매출 상승

지진 이벤트는 특정 기간 동안의 판매량에 급격한 변동을 일으킬 수 있어 데이터에 이상치로 나타날 수 있음.
이상치는 평균적인 패턴과 다른 동작을 보여주므로,
분석 시에는 이를 고려하여 데이터를 해석해야 함



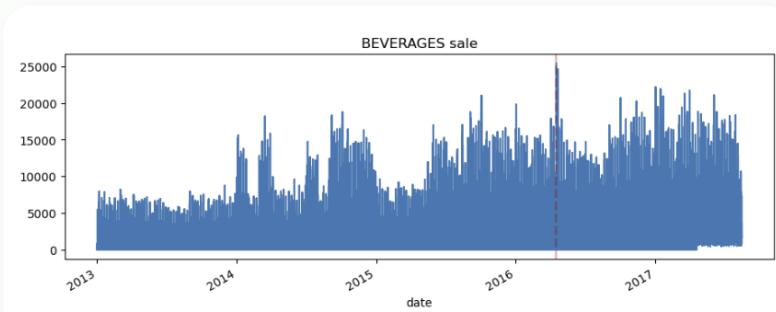
4 탐색적 자료 분석

개요 | 팀 구성 및 역할 | 수행 절차 방법 | 통계 분석 | 자체 평가

제품군 별 지진 영향 분석

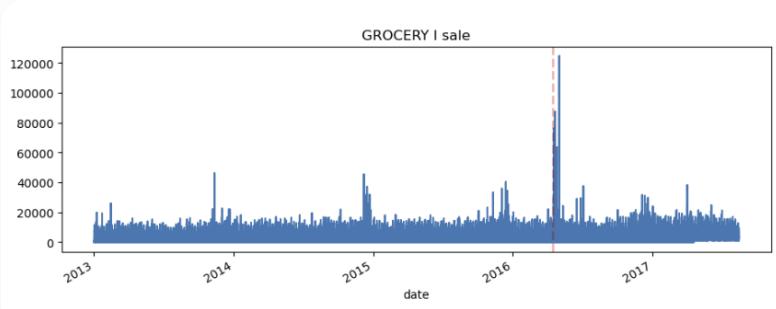
제품군 별 판매 패턴

- ✓ 냉동식품과 학용품, 사무용품은 계절성이 뚜렷함

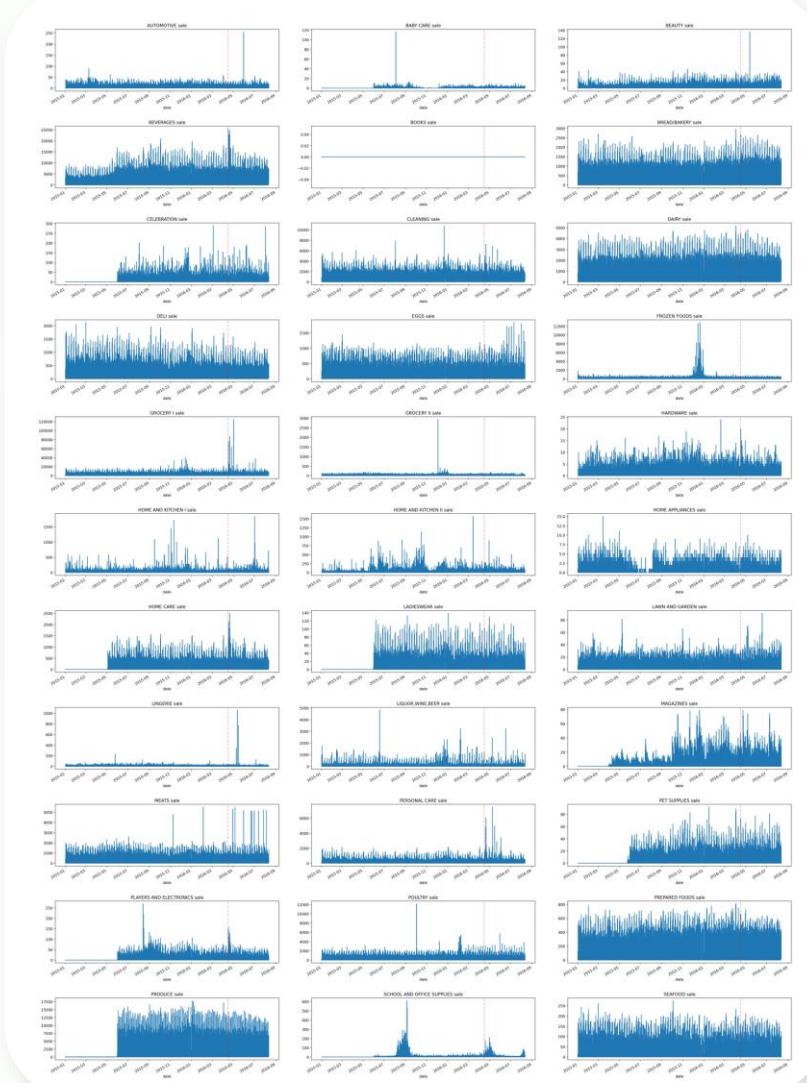
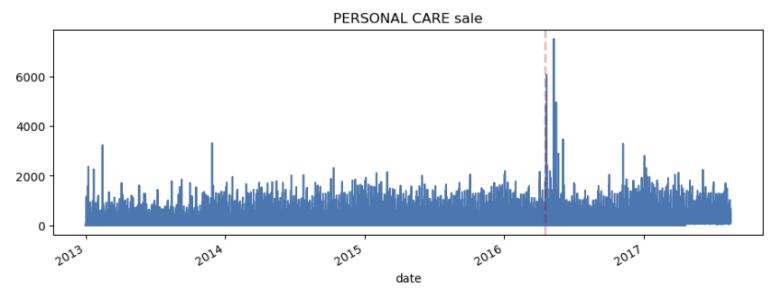


지진 이후 매출 급상승한 제품군

- 1 Beverage (음료)
- 2 Grocery (식료품)
- 3 Personal Care (개인용품)



지진 이후 음료, 식료품, 가정용품의
매출 급상승은 구호물품으로
사용되었을 가능성이 높음



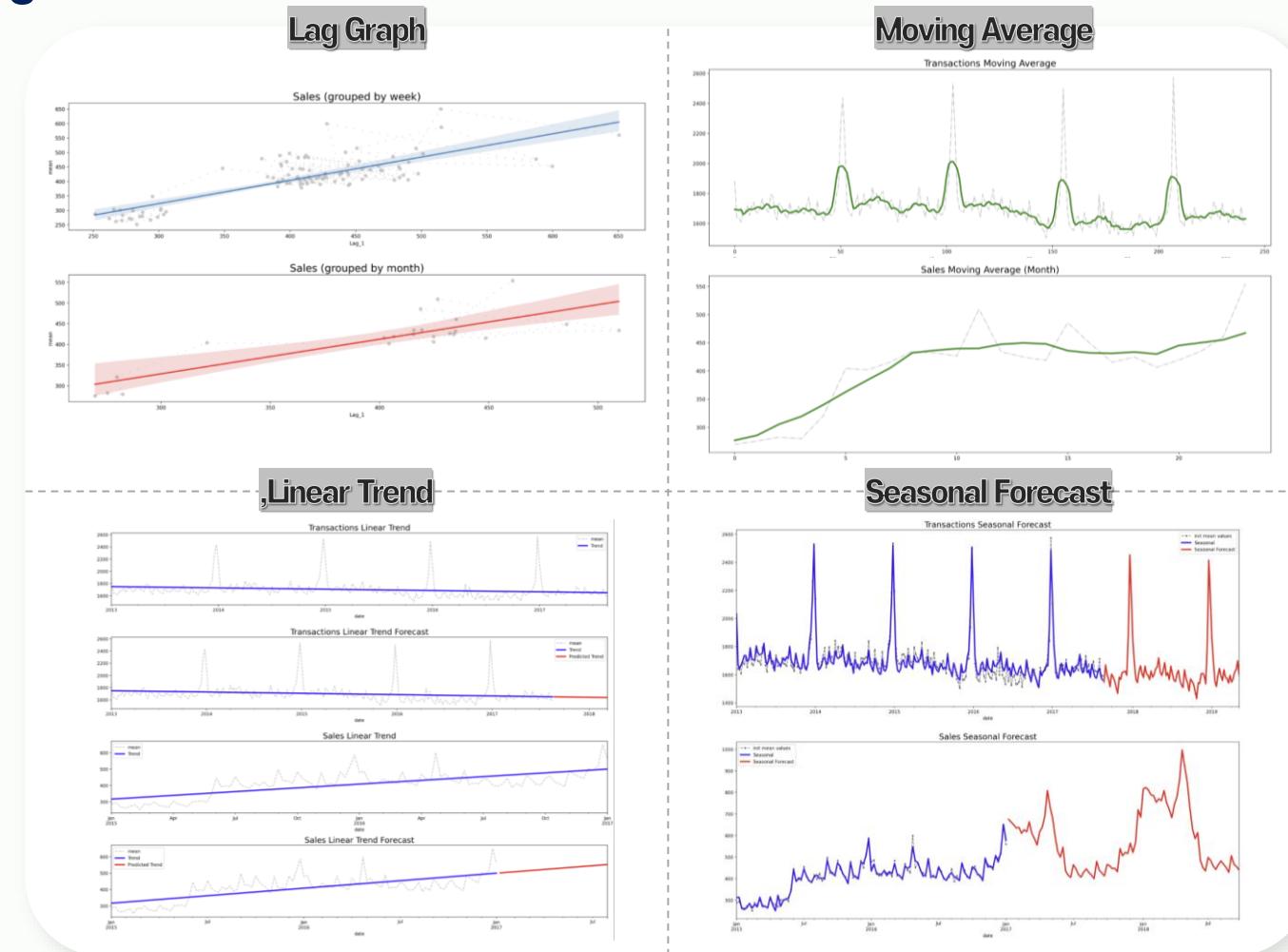
ARIMA(Autoregressive Integrated Moving Average) 모델이란?

ARIMA 모델

시계열 데이터의 추세, 계절성, 그리고 이동평균을 고려하여 예측하는 모델로 Autoregressive는 자기회귀모형을, Moving Average는 이동평균모형을 의미함

ARIMA 모델을 위한 탐색적 분석 과정

- ✓ Lag Graph(지연 그래프) : 시계열 데이터의 현재 값과 이전 시점의 값 사이의 관계를 분석. 자기상관성을 파악하고 시계열 데이터가 이전 값에 얼마나 의존하는지 확인
- ✓ Moving Average(이동평균) : 데이터의 추세를 파악하고, 추세를 제거한 후의 데이터를 분석에 활용
- ✓ Linear Trend(선형 추세) : 데이터의 시간에 따른 증감을 분석하여 추세의 방향성과 기울기를 확인하고, 데이터의 선형적인 추세를 파악
- ✓ Seasonal Forecast : 계절성 패턴을 시각화하여 특정 기간의 데이터가 어떤 패턴을 따르는지 확인하고, 해당 패턴을 기반으로 계절성 예측을 수행



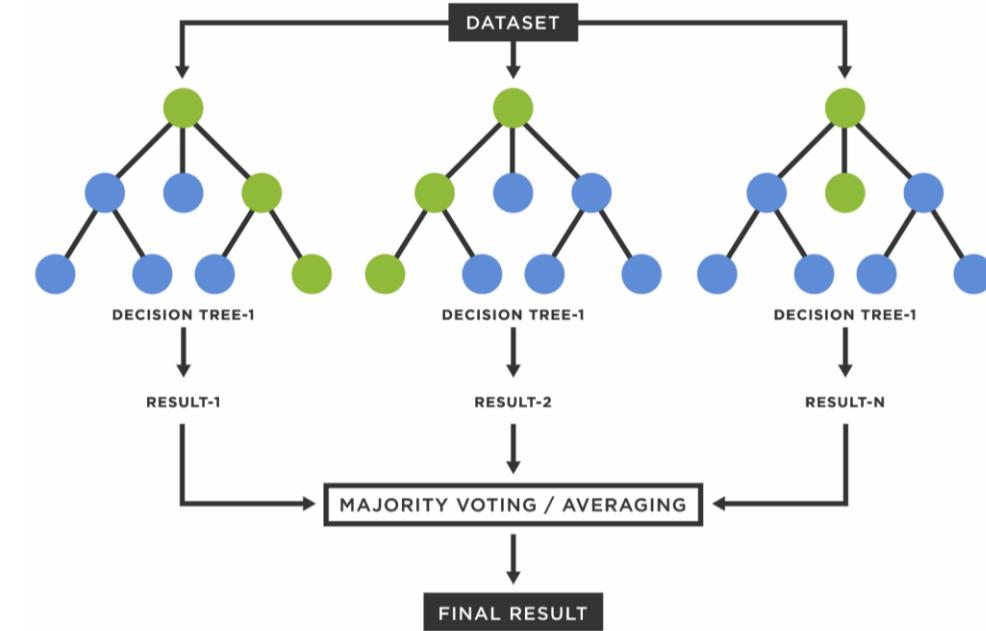
랜덤 포레스트 회귀(Random Forest Regression)란?

☞ 랜덤 포레스트 회귀 알고리즘 정의

다중 결정 트리를 활용하여 회귀 분석을 수행하는 방법
다양한 특성을 고려하여 예측 모델을 구축할 수 있음

☞ 비즈니스에서 랜덤 포레스트 알고리즘을 쓰는 이유

- ✓ 정확한 예측 : 신뢰도 있는 데이터로 비즈니스 결정에 도움
- ✓ 다변량 분석 : 여러 독립변수를 고려, 종속변수에 대한 복잡한 관계를 모델링
- ✓ 예측 변수 중요도 평가 : 중요한 변수에 집중하여 리소스를 효율적으로 할당
- ✓ 오버피팅 방지 : 데이터에 대해 신뢰할 수 있는 예측을 제공, 신뢰성 강화



랜덤 포레스트 회귀(Random Forest Regression) 대시보드 구현

랜덤 포레스트 회귀 모델링

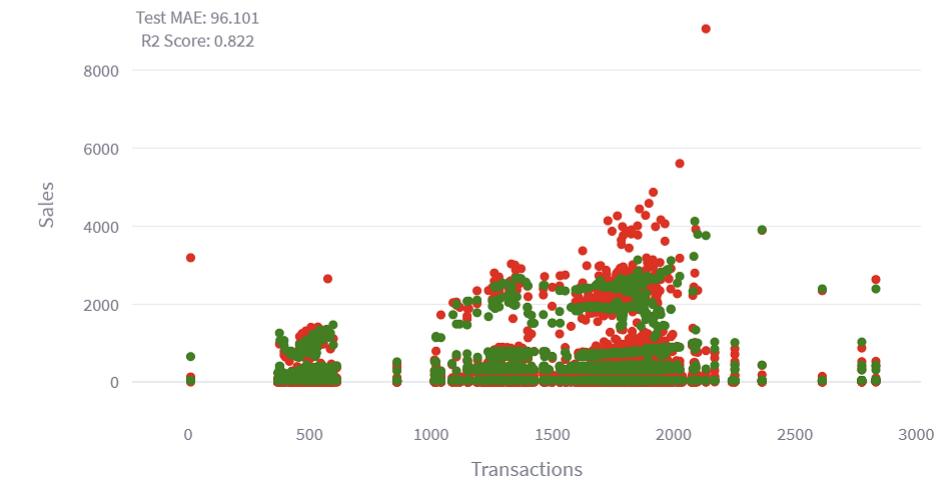
max depth와 minimum samples leaf는 랜덤 포레스트 모델의 성능에 영향을 줄 수 있는 중요한 하이퍼파라미터

랜덤 포레스트 회귀 모델링의 하이퍼파라미터 설명

- ✓ Max depth : 결정 트리의 최대 깊이. 결정 트리의 복잡성을 제어하는데 사용되며 값이 높을수록 트리가 더 복잡해지고 값이 낮을수록 트리가 더 단순해짐.
일반적으로 max depth는 5에서 10 사이의 값으로 설정
- ✓ Minimum samples leaf : 각 리프 노드에 필요한 최소 샘플 수. 트리의 일반화 성능을 제어하는 데 사용되며 값이 높을수록 트리가 더 일반화되고 값이 낮을수록 트리가 특이해짐
일반적으로 minimum samples leaf는 5에서 10 사이의 값으로 설정



Sales Prediction with RandomForestRegressor by Store Number



RMSLE(Root Mean Squared Log Error) 평가란?

👉 RMSLE 정의

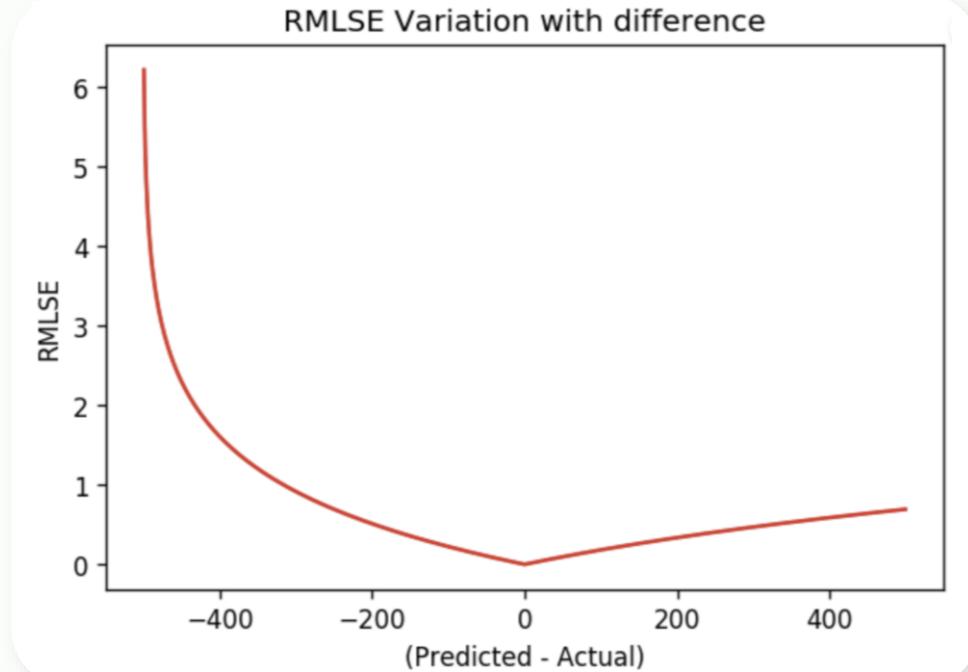
회귀 문제에서 예측 성능을 평가하는 지표
값이 작을수록 예측값과 실제값의 차이가 없다는 뜻

👉 RMSLE가 매출 예측 모델링에서 평가지표로 사용되는 이유

- ✓ 비율적인 오차 고려 : 로그 변환으로 예측 성능의 정확도가 올라감
- ✓ 오차의 대칭성 보정 : 오차를 일관성 있게 평가할 수 있도록 함
- ✓ 이상치 영향 완화 : 예측 성능 평가를 안정적으로 수행할 수 있도록 함

👉 RMSE(Root Mean Squared Error)와 비교

RMSE는 절대적인 예측 오차를 고려하는 지표
두 평가 지표를 함께 사용하면 예측 모델 성능을 더 정확하게 평가할 수 있음
비즈니스에서는 매출 예측 모델의 신뢰성과 정확성을 높이는 데 도움을 줌



평가 지표로 RMSLE를 사용하면 비율적인 오차를 고려, 오차의 대칭성을 보정, 이상치 영향을 완화할 수 있어 보다 정확하고 신뢰성 있는 매출 예측 가능



Project 5 대시보드 웹서비스 구현

대시보드 시각화 및 해석 | 대시보드 구현 영상

5 대시보드 시각화 및 해석

개요 | 팀 구성 및 역할 | 수행 절차 방법 | 통계 분석 | 자체 평가

대시보드 웹 서비스 구현

대시보드 측면 소개

측면 요소는 데이터를 조직화하고 시각화하여 사용자가 다양한 관점에서 이해하고 분석하는데 도움을 줌

측면 구성 요소

- 1 Data Overview (데이터 개요) : 데이터의 전반적인 정보와 특징을 보여주는 섹션
- 2 Variables (변수) : 다양한 변수들에 대한 세부 정보와 통계량을 제공하는 섹션
- 3 Interactions (상호작용) : 사용자 입력을 기반으로 데이터를 탐색하고 조작할 수 있는 기능을 제공하는 섹션
- 4 Correlations (상관관계) : 변수 간의 상관관계를 시각화하고 분석하는 섹션
- 5 Missing Value (결측치) : 누락된 값을 처리하고 보완하는 방법에 대한 섹션
- 6 Sample (샘플) : 데이터셋에서 임의의 샘플을 선택하여 보여줌

Data Preview

train stores oil transactions holidays_events

Train Data

Overview

Overview Alerts 8 Reproduction

Dataset statistics

| | |
|-------------------------------|----------|
| Number of variables | 9 |
| Number of observations | 104858 |
| Missing cells | 0 |
| Missing cells (%) | 0.0% |
| Duplicate rows | 0 |
| Duplicate rows (%) | 0.0% |
| Total size in memory | 12.0 MiB |
| Average record size in memory | 120.3 B |

Variable types

| | |
|-------------|---|
| Numeric | 6 |
| Categorical | 3 |

5 대시보드 시각화 및 해석

개요 | 팀 구성 및 역할 | 수행 절차 방법 통계 분석 자체 평가

데이터 분석 결과 요약

Overview

Alerts 8

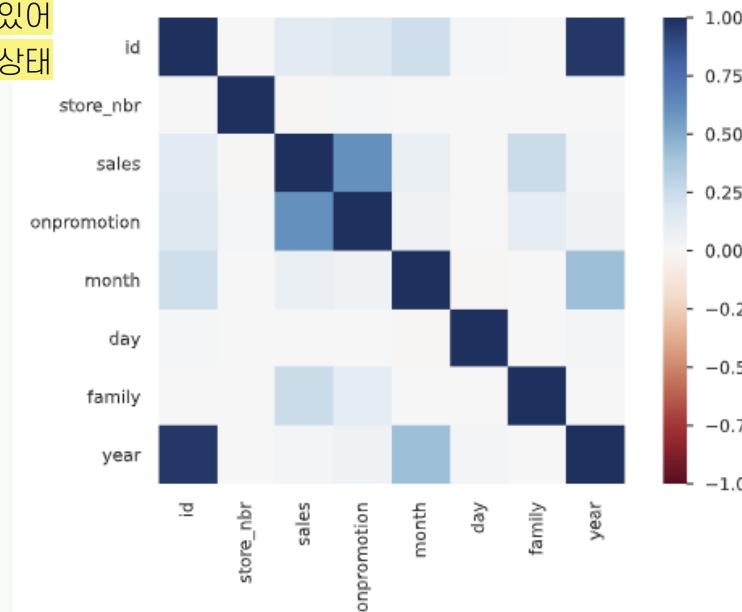
Reproduction

카디널리티가 높다 = 해당 변수에 대해
다양한 값을 가지고 있어
중복되는 값이 적거나 거의 없는 상태

Alerts

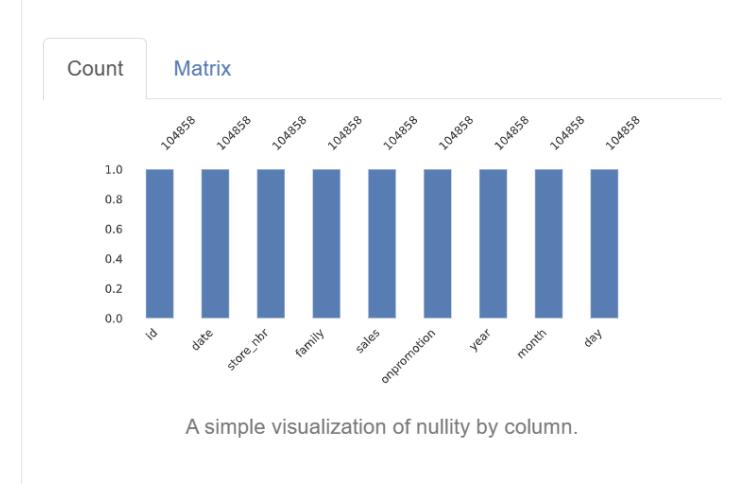
- `date` has a high cardinality: 589 distinct values High cardinality
- `id` is highly overall correlated with `year` High correlation
- `sales` is highly overall correlated with `onpromotion` High correlation
- `onpromotion` is highly overall correlated with `sales` High correlation
- `year` is highly overall correlated with `id` High correlation
- `id` has unique values Unique
- `sales` has 26046 (24.8%) zeros Zeros
- `onpromotion` has 78814 (75.2%) zeros Zeros

데이터 특성 알림



상관관계

Missing values



결측치

파이썬으로 전처리하기 전, 대시보드를 통해 데이터의 구성, 변수들의 분포, 결측치 여부 등을
쉽게 확인할 수 있어 데이터 전처리 과정에서 필요한 정보를 빠르게 파악 가능

미니 프로젝트 :

Python을 활용한 통계 분석 및 웹서비스 구현

두 평균의 비교

☞ 독립 표본 t-검정(Independent Samples t-test)

두 개의 독립된 표본 간의 평균 차이를 검정하는 통계적 방법

- ✓ **스튜던트 t-검정** : 두 개의 표본은 동일한 분산을 가진 정규 분포를 따른다
- ✓ **웰치 t-검정** : 두 개의 표본은 동일한 분산을 가진 정규 분포를 따르지 않는다

두개의 시계열 데이터 또는 그룹 간에 통계적으로 유의미한 차이가 있는지 확인

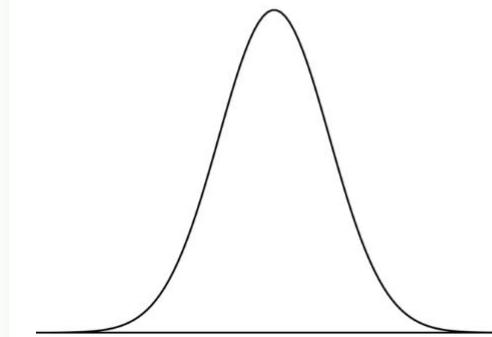
☞ 비즈니스 데이터 분석에서 스튜던트 t-검정이 쓰이는 이유

- ✓ 특정 기간의 대비로 두 그룹 간의 평균 차이를 분석, 그룹 간의 유의미한 차이를 확인할 수 있음
- ✓ 이를 통해 특정 전략, 제품, 서비스의 효과를 평가하거나 다양한 그룹 간의 성과를 비교할 수 있음

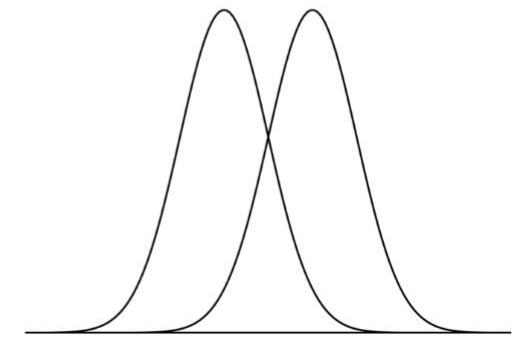
지진 같은 외부 요인이 데이터에 영향을 미친 경우,
독립 표본 t-검정을 사용하여 이 요인이 실제로 평균에
유의미한 변화를 가져왔는지 확인할 수 있음

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

귀무가설
null hypothesis



대립가설
alternative hypothesis



☞ 귀무가설 (Null Hypothesis)과 대립가설 (Alternative Hypothesis)

- H0 : 두 집단의 평균이 동일하다, 아무런 차이가 없다고 가정
- H1 : 두 집단의 평균이 서로 다르다, 평균값에 유의미한 차이가 있다고 주장

5 대시보드 시각화 및 해석

개요 | 팀 구성 및 역할 | 수행 절차 방법 통계 분석 자체 평가

두 평균의 비교

Select Column 1

BEVERAGES

Select Column 2

BREAD/BAKERY

Sample Size is Equal?

| | family | sales |
|-------------|--------|---------|
| BEVERAGES | 3,348 | 2,301.3 |
| BREAD/BAKER | 3,348 | 483.6 |

Independent Test

| | T | dof | alternative | p-val | CI95% | cohen-d | BF10 | power |
|--------|---------|-------|-------------|-------|----------------|---------|------|-------|
| T-test | 45.2252 | 6,694 | two-sided | 0 | 1738.92 1896.5 | 1.1054 | inf | 1 |

H_1 : The means for the two populations are not equal. 두 표본집단의 평균은 같지 않다

👉 독립 표본 t-검정의 결과

두 평균 간의 차이가 통계적으로 유의미하다고 보이므로 귀무가설을 기각하고 대립가설을 채택한다

데이터 정제 방법 결정

두 샘플 데이터가 같거나 연관성이 있는가?

YES

Paired T-Test

NO

두 샘플 데이터가 같은 사이즈인가?

YES

Equal Variance T-Test

NO

두 샘플 데이터의 분산이 동일한가?

YES

Unequal Variance T-Test

NO



Project 6 자체 평가

자체평가 및 문제점 | 개선점과 발전 가능성

6 자체 평가

개요 | 팀 구성 및 역할 | 수행 절차 및 방법 | 통계 분석 | 자체 평가

자체 평가 결과

👉 모델 평가

SMSLE 점수



submission.csv

Complete · 25s ago

2.15214

👉 데이터의 한계

⚠️ 제한된 데이터 수집 기간

⚠️ 제품군에 대한 개별적인 예측 모델링이 어려운 데이터

⚠️ 지진 이벤트가 이상치로 작용, 경제 지표, 시장 동향, 경쟁사의 활동 등 외부 데이터가 고려되지 않아 예측력이나 일반화 능력에 한계가 있음

👉 모델의 문제점

✗ 예측 정확도가 떨어짐

👉 개선해야 할 점과 발전 가능성

더 다양한 Feature 활용(기상정보나 경쟁업체의 정보 활용)

딥러닝 모델을 활용하면 더욱 정확한 예측 가능 (LSTM / GRU 등)

👉 프로젝트 평가

“ 도메인 지식에 기반한 피처 엔지니어링이 부족해 중요 피처들 간의 상호작용을 적절히 고려하지 못해 데이터의 품질이 떨어졌다. ”

“ 머신러닝에 대한 이해도가 떨어져서 모델의 복잡성을 고려하지 못하였다. ”

“ 모델의 성능을 위해 하이퍼파라미터 튜닝이 적절히 이루어지지 못하였다. ”



Project 7 부록

출처 및 참고자료

캐글 노트북 링크



캐글 대회 페이지

- [Store Sales - Time Series Forecasting](#)
- | Kaggle



참고한 캐글 노트북 페이지

- [Store Sales TS Forecasting - A Comprehensive Guide](#)
- [Time Series Forecasting Tutorial](#)
- [Store Sales Forecasting - Exploration](#)
- [First project : Store sales](#)
- [Following TS tutorial](#)

The screenshot shows the Kaggle interface. On the left, there's a sidebar with navigation links: Create, Home, Competitions, Datasets, Models, Code, Discussions, Learn, More, Your Work, Recently Viewed, and Recently Edited. The main area displays the competition details for 'Store Sales - Time Series Forecasting'. It features a search bar at the top right. Below the title 'Store Sales - Time Series Forecasting' and subtitle 'Use machine learning to predict grocery sales', it shows 'Kaggle · 727 teams · Ongoing'. There are tabs for Overview, Data, Code, Discussion, Leaderboard, and Rules. A large 'Submit Predictions' button is on the right. The 'Overview' section includes a 'Description' table with rows for 'Goal of the Competition' (Forecast store sales), 'Evaluation' (using time-series forecasting), and 'Frequently Asked Questions'. The 'Goal of the Competition' row contains text about forecasting store sales for Corporación Favorita using machine learning.



Guido van Rossum 



"Life is too short, You need Python."

17 May 2023

