Python을 활용한 통계 분석 및 웹 서비스 구현

Store sales

Time Series Forecasting

Team Attention

유수빈, 장수경, 김범모, 오성준, 배상일

연세 IT 미래 교육원

CONTENTS

01

프로젝트 개요

대회의 목표

시계열 알고리즘을 사용하여 에콰도르 식료품 소매업체 매출 예측

잠재적 영향

매출에 영향을 주는 주 요소 파악 및 매출 증가

모델 평가

Root Mean Squared Logarithmic Error

<출처: https://www.kaggle.com/competitions/store-sales-time-series-forecasting>

대회의 평가 지표 : *Root Mean Squared Logarithmic Error*

RMSLE는 다음과 같이 계산합니다.

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

RMSLE

예측 값과 실제 값의 로그 차이를 측정하여, 평균 제곱 오차(MSE)와 비슷한 방식으로 계산됩니다.

RMSLE를 사용하면 예측 값과 실제 값의 오차를 측정하며 이를 통해 모델의 성능을 평가할 수 있습니다.

- n: 총 인스턴스의 수
- $\hat{y}_i$: 인스턴스 i에 대한 **예측된 타겟 값**
- y_i: 인스턴스 i에 대한 **실제 타겟**
- log: 자연 로그

출처 : https://ahnjg.tistory.com/90

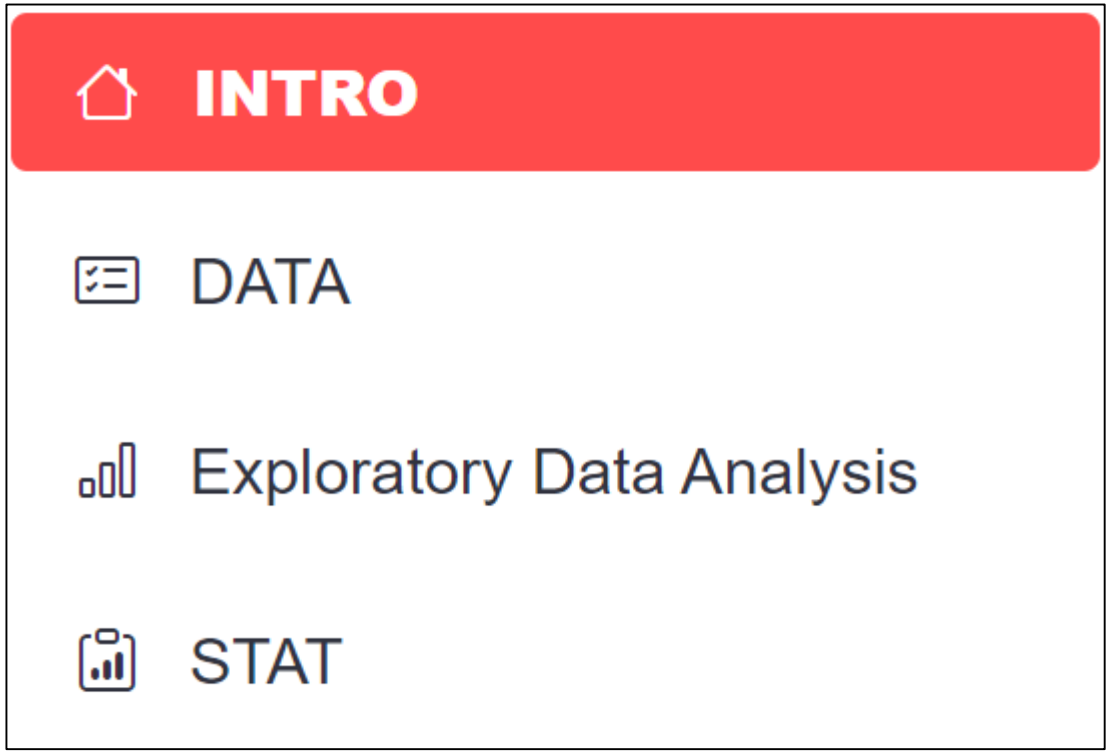
◆ Dataset Description

이 대회에선 에콰도르에 위치하고 있는 Favorita 상점에서 판매되는 수천 개의 제품군의 매출을 예측할 수 있습니다.

train.csv	상점 번호, 제품군, 프로모션 및 목표 매출로 구성된 시계열 기능으로 구성된 데이터
test.csv	학습 데이터와 동일한 테스트 데이터
store.csv	상점 메타데이터
oil.csv	일일 유가
holidays_events.csv	휴일 및 이벤트 데이터

<출처: https://www.kaggle.com/competitions/store-sales-time-series-forecasting/data>

대회를 진행한 내용을 웹 서비스로 구현하는 작업을 진행



01 대회 정보

02 데이터셋 정보

03 탐색적 자료 분석

04 통계 분석

김범모

- 시계열 예측 및 평가

오성준

- EDA
- 대시보드 웹 서비스 구현

유수빈

- 데이터 전처리
- 머신러닝 모델 평가

배상일

- 통계 분석
- 대시보드

장수경

- EDA
- 머신러닝 모델링

02

수행 절차 및 방법

언어 및 도구

언어 :

Python 3.10.10

도구 :

Microsoft Excel

Flourish

Visual Studio Code

PyCharm

라이브러리

Pandas 1.5.3

NumPy 1.24.3

Math

Seabron 0.11.2

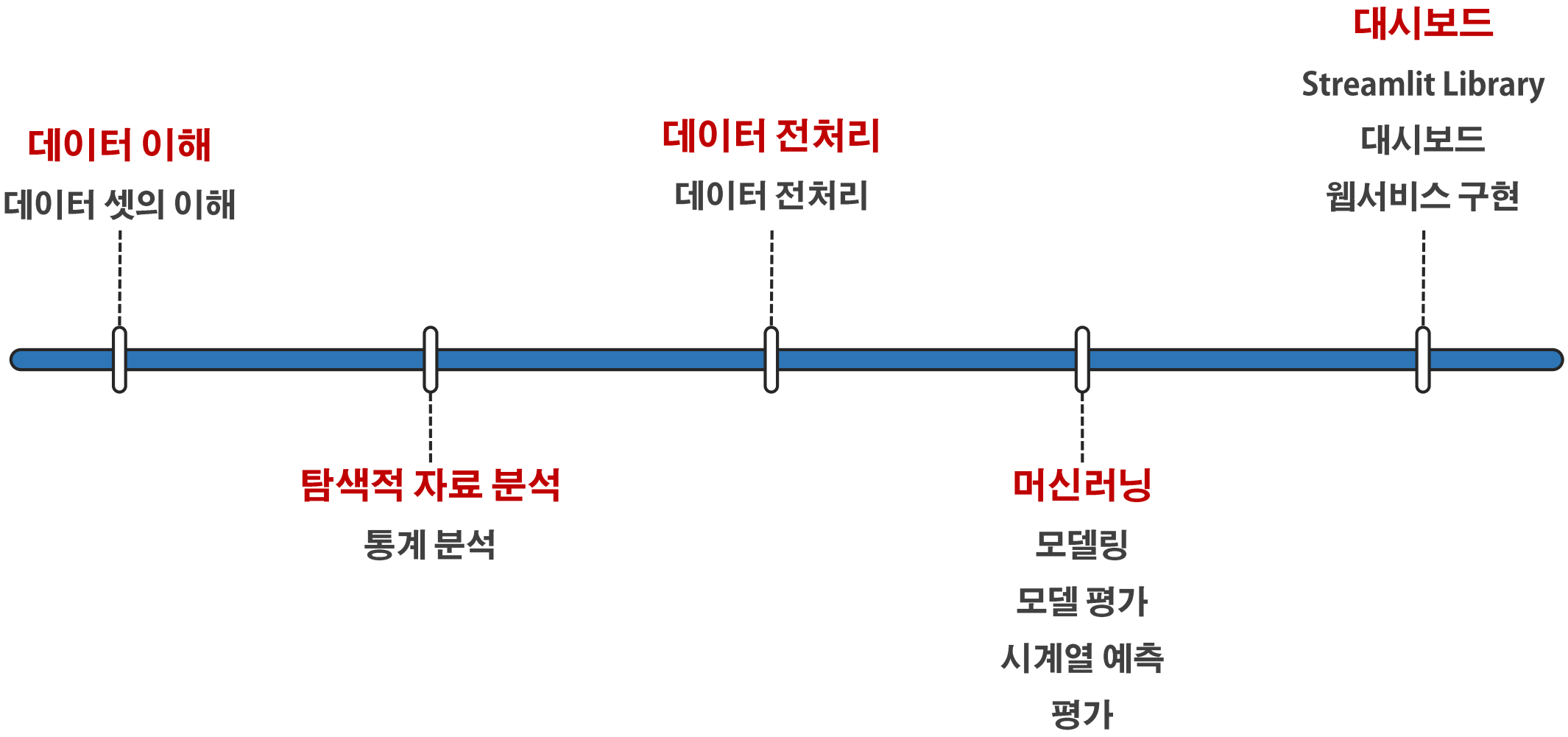
Streamlit 1.22.0

Matplotlib 3.7.1

Plotly 5.14.1

Scikit-learn 1.2.2

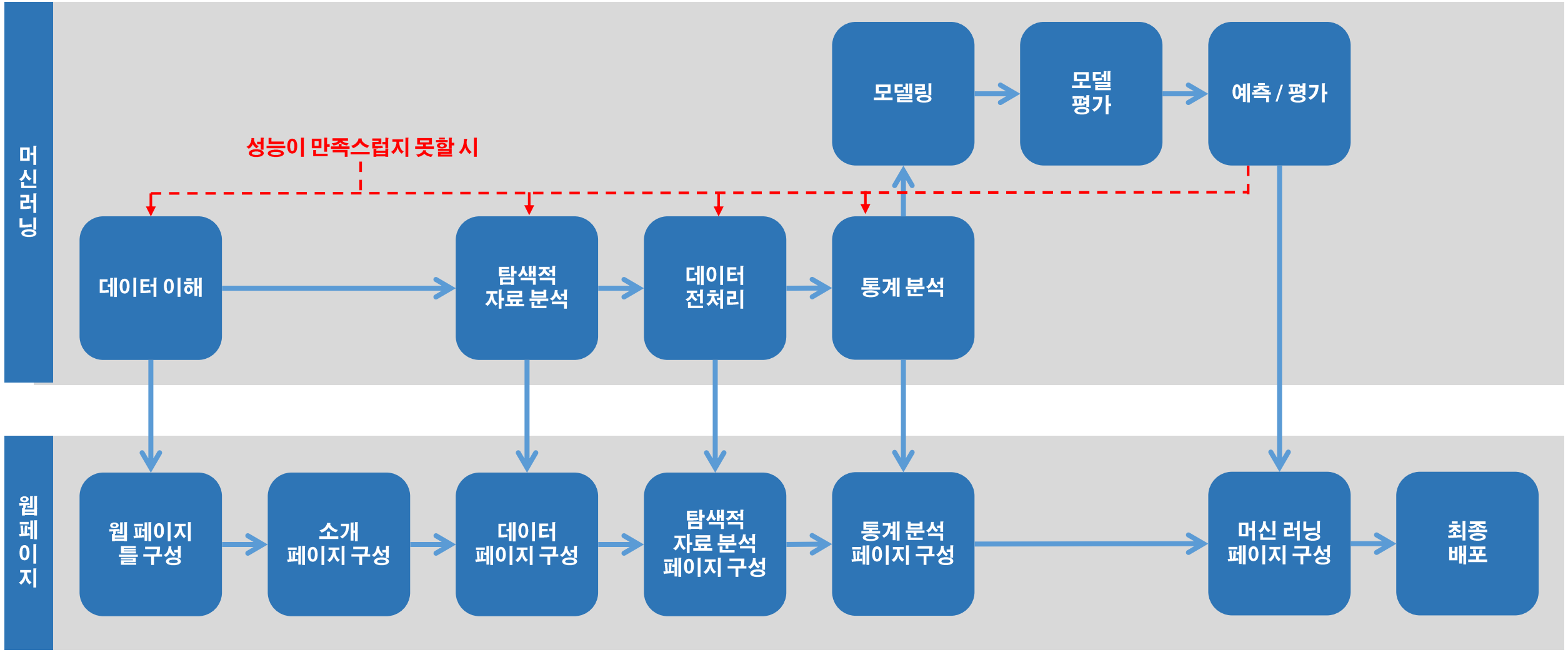
Scipy 1.10.1


데이터 이해
데이터 셋의 이해
탐색적 자료 분석
통계 분석
데이터 전처리
데이터 전처리
머신러닝
모델링
모델 평가
시계열 예측
평가
대시보드
Streamlit Library
대시보드
웹서비스 구현

프로젝트 수행 기간(5월)

구분	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. 데이터 이해																	
2. 데이터 전처리																	
3. 1차 대시보드 시각화																	
4. 통계 분석																	
5. 머신러닝																	
6. 대시보드 웹 서비스 구현																	
7. 최종 수정																	

개발 환경	수행 절차	수행 기간	순서도



03

통계 / 머신러닝

EDA

- 매출 패턴 파악
- 거래량 패턴 파악
- 매출과 거래량의 상관 관계 파악
- 매출 추세, 계절성 패턴 파악
- 유가 정보에 따른 매출 패턴 파악
- 휴일 정보에 따른 평균 매출 비교

데이터 전처리

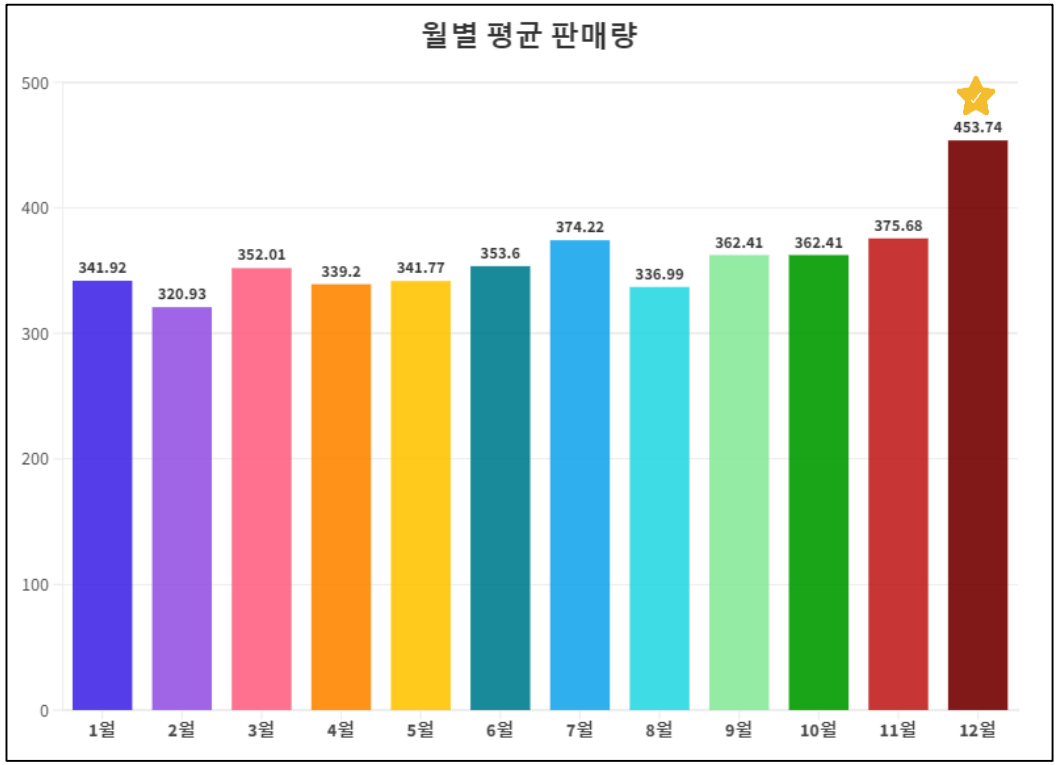
- 이상치 제거
- 결측값 전처리
- 더미 전처리
- 지연값 전처리
- 추세 전처리
- 계절성 전처리

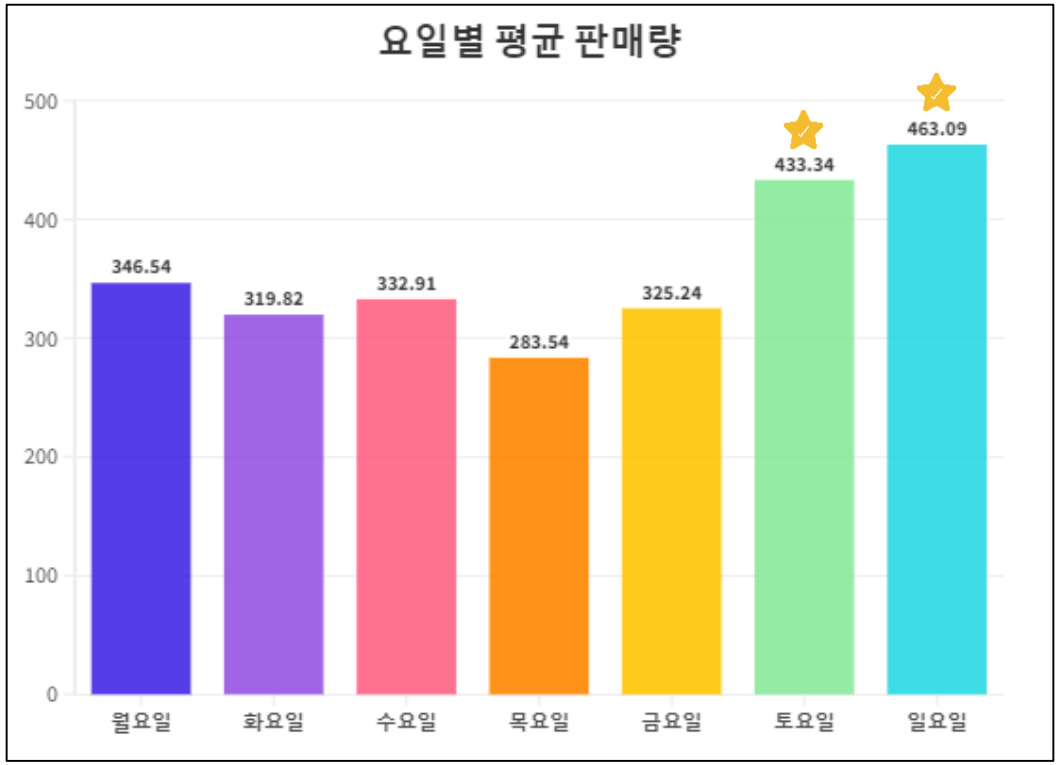
모델링

- 선형 회귀 분석 및 시각화
- 릿지 회귀 분석 및 시각화

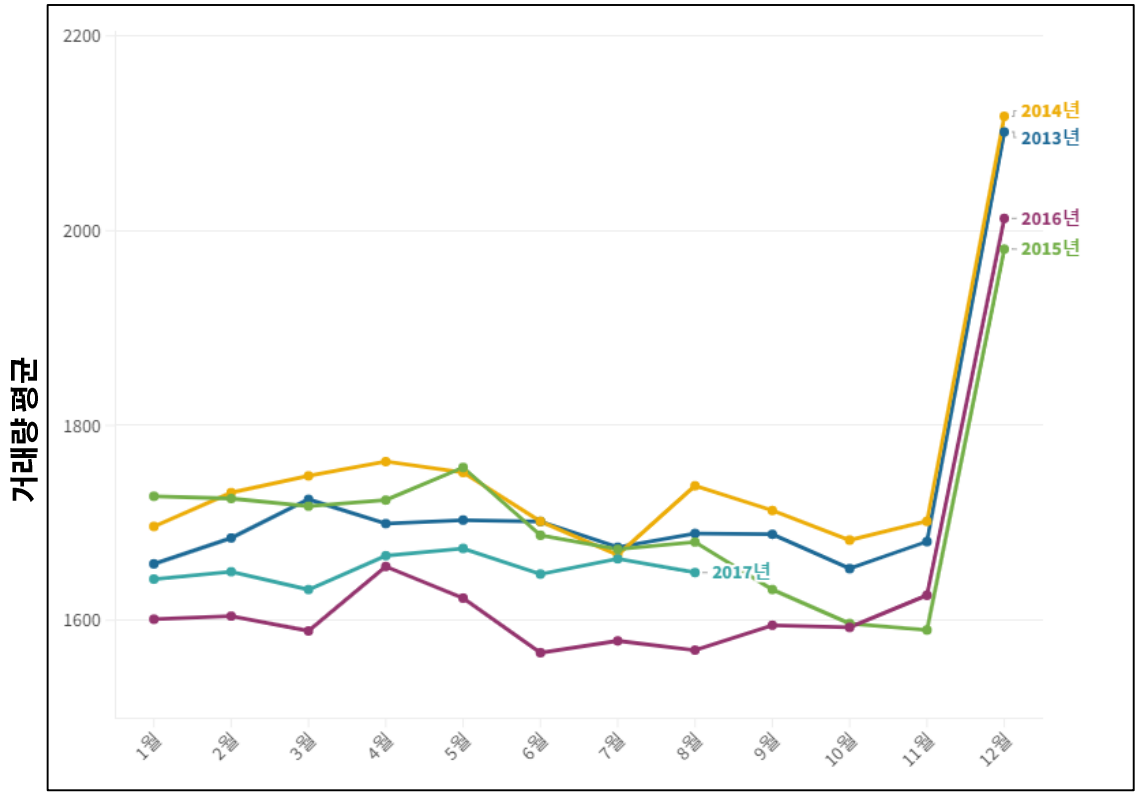
예측 / 평가

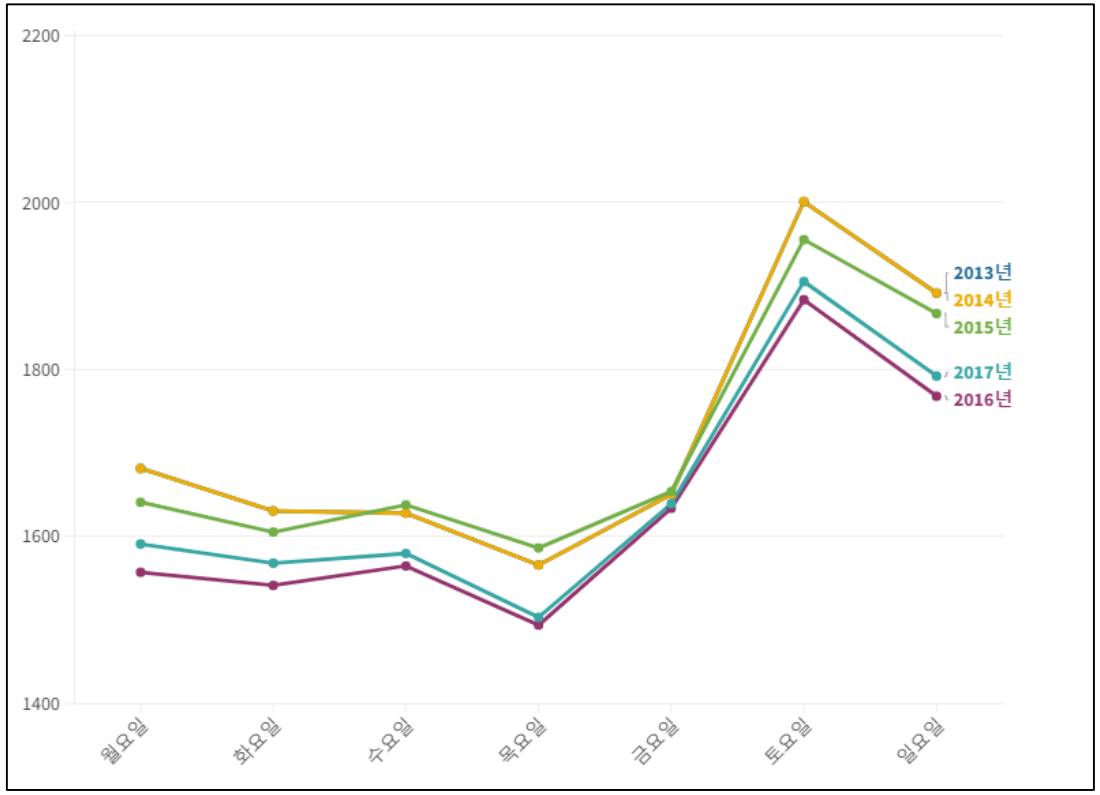
- 매장별 제품별 매출 예측 및 성능 평가



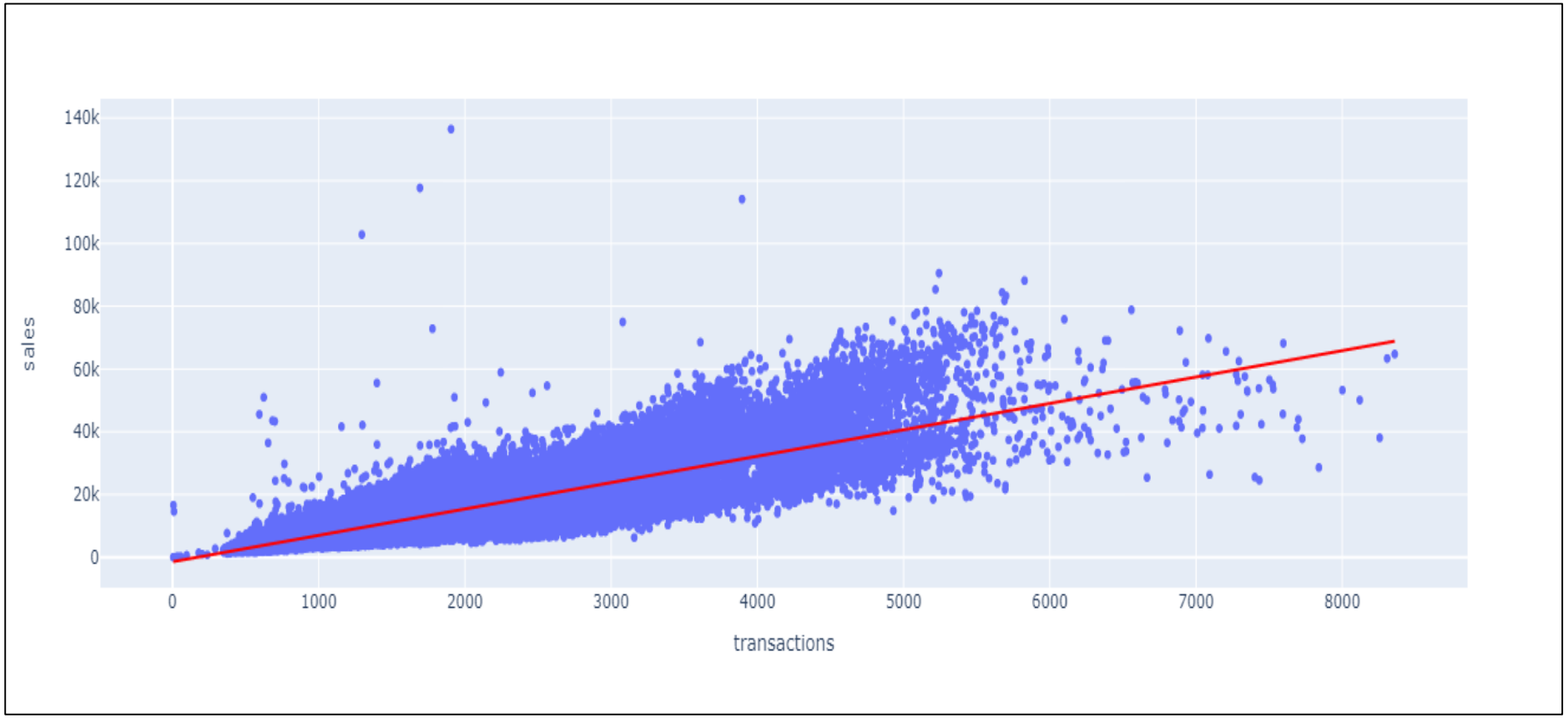


월별, 요일 별 평균 판매량을 살펴 봤을 때 연말과 주말에 매출이 높은 것을 확인 할 수 있었음.





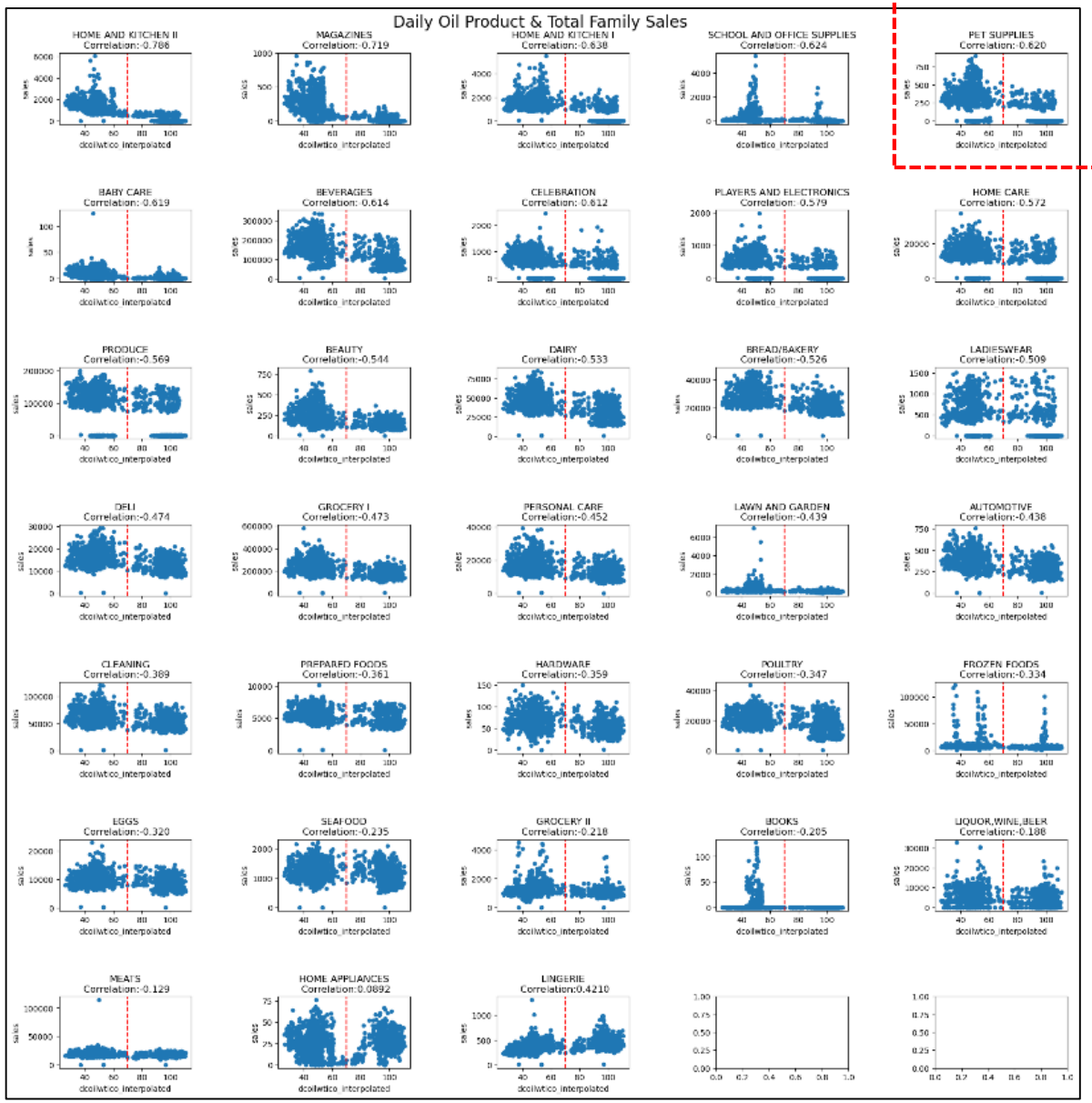
거래량도 마찬가지로, 연말과 주말이 거래량이 높음.

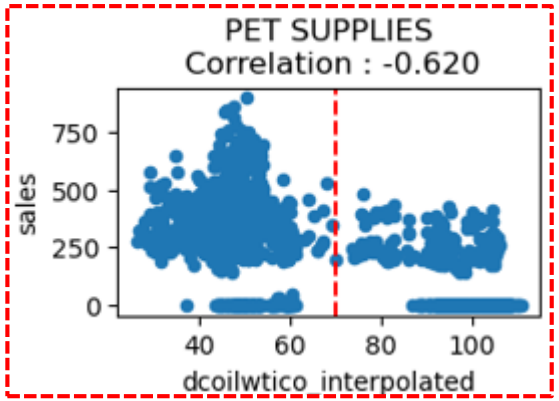


매출(Sales)과 거래량(Transaction)의 산점도가 회귀선과 비슷한 양상을 띄며, 관계가 유의미하다고 판단됨.

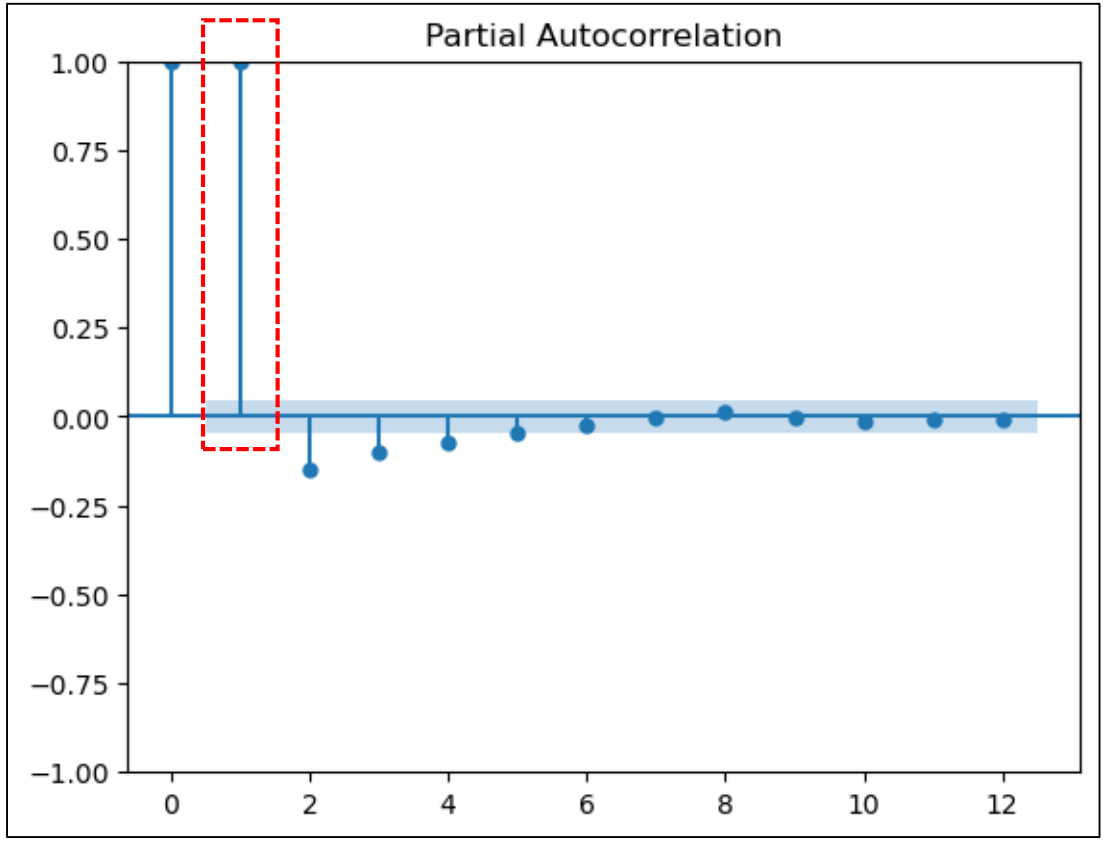
매출 패턴에 대해서 **관측값**(Observed), **추세**(Trend), **계절성**(Seasonal), **잔차**(Residuals)에 대해 패턴 파악.

- 추세(Trend) : 장기적인 변동 패턴을 나타내며 여기서는 **시계열에 따라 증가함**을 보임
- 계절성(Seasonal) : 주기적으로 반복되는 패턴이나 변동을 나타내며 여기서는 **연말에 매출이 증가함**을 보임
- 잔차(Residuals) : 추세와 계절성을 제거한 나머지 데이터

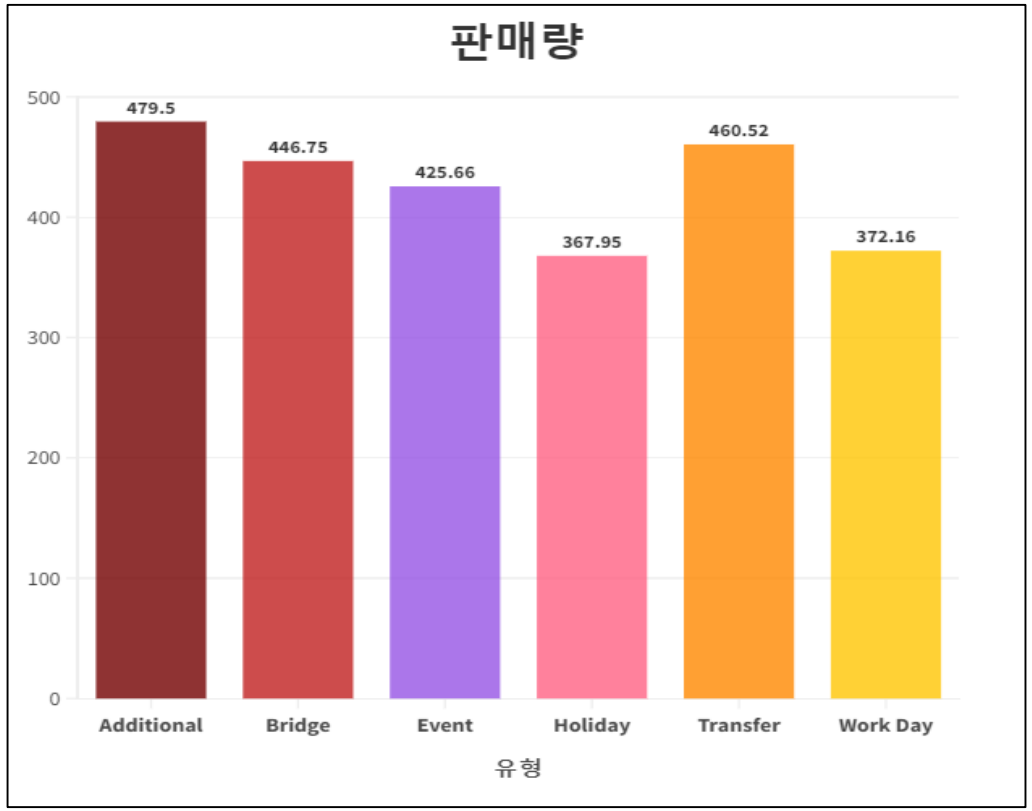




유가 데이터를 사전 **전처리하여 결측값은 보간으로 처리**한 후 매출과의 상관관계를 분석한 결과 특별한 상관관계는 나오지 않았지만 **유가 70을 기준으로 매출이 상이하게 다른 제품군이 있다는 것을 발견했음.**



유가에 대해서 지연에 따라 인접한 관측치와 그에 인접한 관측치 간의 자기 상관 관계를 보면 **Lag = 1 이 강한 상관 관계를 가짐**을 볼 수 있음.



앞에 휴무는 매출에 영향을 미칠 수 있다는 결과를 바탕으로 휴무일 또는 이벤트에 대해서 판매량을 조회해 보았을 때, **휴무일 또는 이벤트의 판매량 평균의 차이가 크지 않다고 판단**하고 휴무일 또는 이벤트에 대해 전처리와 더미 처리를 하였음.

이상치 제거

예) 매출이 0 인 경우 매장이 오픈 전이거나 값의 누락으로 인해 0으로 처리했을 가능성을 염두해 이상치 처리를 한다.

결측값 전처리

예) 유가에 대한 정보에는 결측값이 존재함으로 결측값에 대해 보간 처리하였다.

더미 전처리

예) 휴일은 매출에 영향을 준다고 할 때, 요일 정보와 휴무, 이벤트 정보 등, 평균 매출보다 영향을 더 준다고 판단되는 것에 대해서 더미 처리를 한다.

지연값 전처리

예) 판매량 및 유가 데이터와 같은 시계열 데이터에 대해 데이터의 패턴과 예측 성능 개선을 위한 이전 값과의 자기상관성을 파악한 지연값을 준다.

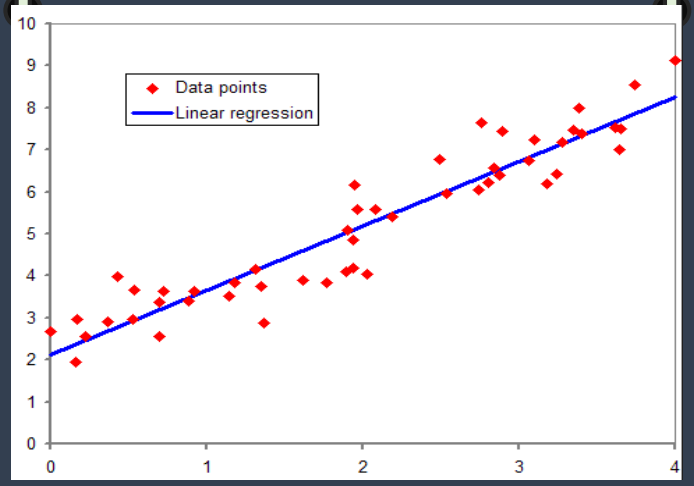
추세 전처리

예) 판매량 및 유가 데이터와 같은 시계열 데이터에 대해 데이터의 장기적인 변화를 모델에 반영하고, 추세의 성분을 추출하거나 예측 할 수 있게 한다.

계절성 전처리

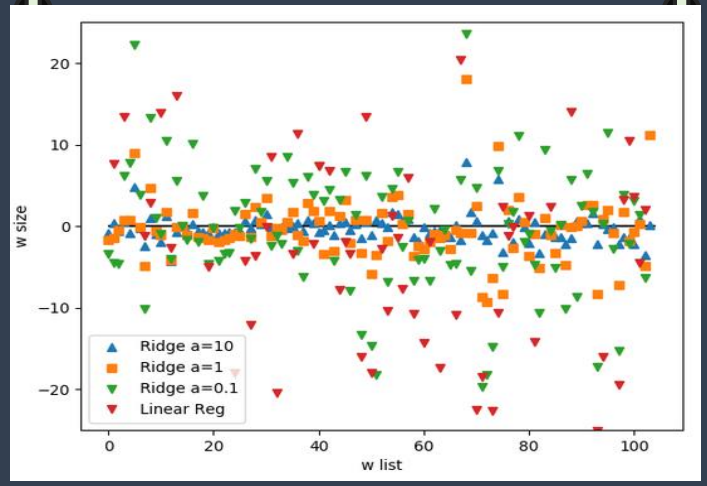
예) 판매량 및 유가 데이터와 같은 시계열 데이터에 대해 주기성을 모델에 반영하고, 계절성 패턴을 분석하고 예측하게 한다.

선형회귀(Linear Regression)



- 종속 변수와 한 개 이상의 독립 변수 간의 **선형 관계를 모델링하는 방법** 중 하나
- 선형 회귀는 한 개 이상의 독립 변수 x와 종속 변수 y의 선형 관계를 모델링

릿지회귀(Ridge Regression)



- 릿지 회귀는 선형 회귀에서 L2 규제를 추가한 방법
- 선형 회귀 모델에서는 데이터에 대한 잔차의 합을 최소화하는 파라미터 값을 찾아내는데, 릿지 회귀에서는 **추가적으로 모델 파라미터의 크기를 제한하기 위해 L2 규제를 사용**

선형 회귀(Linear Regression)

선형 회귀의 정의

선형 회귀(Linear Regression)는 널리 사용되는 대표적인 회귀 알고리즘이다. 선형 회귀는 종속 변수 y와 하나 이상의 독립 변수 x와의 선형 상관관계를 모델링하는 기법이다. 만약 독립 변수 x가 1개라면 단순 선형 회귀라고 하고, 2개 이상이면 다중 선형 회귀라고 한다.

1) 단순 선형 회귀(Simple Linear Regression)

단순 선형 회귀는 $y=W_x+b$의 식으로 나타난다. 머신러닝에서는 독립 변수 x에 곱해지는 W값을 가중치, 상수항에 해당하는 b를 편향(bias)이라고 부른다. 따라서 단순 선형 회귀 모델을 훈련하는 것은 적절한 W와 b의 값을 찾는것이다.
(그래프의 형태는 직선으로 나타낸다.)

2) 다중 선형 회귀(Multiple Linear Regression)

다중 선형 회귀는$y = W1x1 + W2x2 + \cdots + w_n x_n+b$의 식으로 나타난다. 여러 독립 변수에 의해 영향을 받는 경우이다.
만약 2개의 독립 변수면 그래프는 평면으로 나타날 것이다.

<출처: 선형 회귀 - 위키백과, 우리 모두의 백과사전 (wikipedia.org)>

릿지 회귀(Ridge Regression)

릿지 회귀의 정의

릿지 회귀(Ridge Regression)는 독립 변수들이 강한 상관관계를 가지는 다중 회귀 모델에서 회귀 계수를 추정하는 방법입니다. 선형 회귀 모델에서 다중 공선성 문제로 인해 최소 제곱 추정치의 부정확함을 해결하기 위한 방법으로 개발되었습니다.

과적합된 다중 선형 회귀 모델은 단 하나의 특이값에도 회귀선의 기울기가 크게 변할 수 있다. 릿지 회귀는 어떤 값을 통해 이 기울기가 덜 민감하게 반응하게끔 만드는데, 이 값을 (lambda, λ)라고 한다.
릿지 회귀의 식은 아래와 같다.

$$\beta_{ridge}: argmin[\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 + \lambda \sum_{j=1}^{p} \beta_j^2]$$ (n : 샘플 수, p : 특성 수, λ: .튜닝 파라미터(패널티))

식의 앞부분은 다중 선형 회귀에서의 최소제곱법(OLS, Ordinary Least Square)과 동일하며 뒤쪽의 람다가 붙어 있는 부분이 기울기를 제어하는 패널티 부분이다.

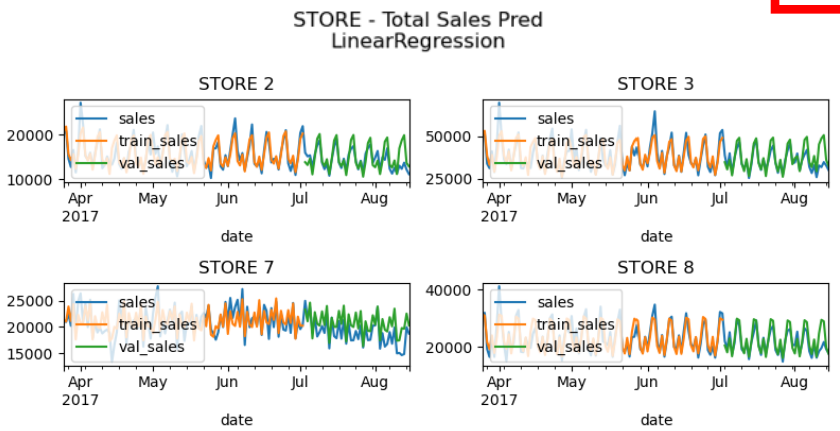
뒷부분을 자세히 보면 회귀계수 제곱의 합으로 표현되어 있는데, 이는 L2 Loss와 같다. 이런 이유로 릿지 회귀를 L2 정규화(L2 Regularization)라고도 하며 만약 람다가 0이면 위 식은 다중 선형 회귀와 동일하다.

반대로 람다가 커지면 커질수록 다중 회귀선의 기울기를 떨어뜨려 0으로 수렴하게 만든다. 이는 덜 중요한 특성의 개수를 줄이는 효과로도 볼 수 있다.

<출처: 능선 회귀 - 위키백과, 우리 모두의 백과사전 (wikipedia.org)>

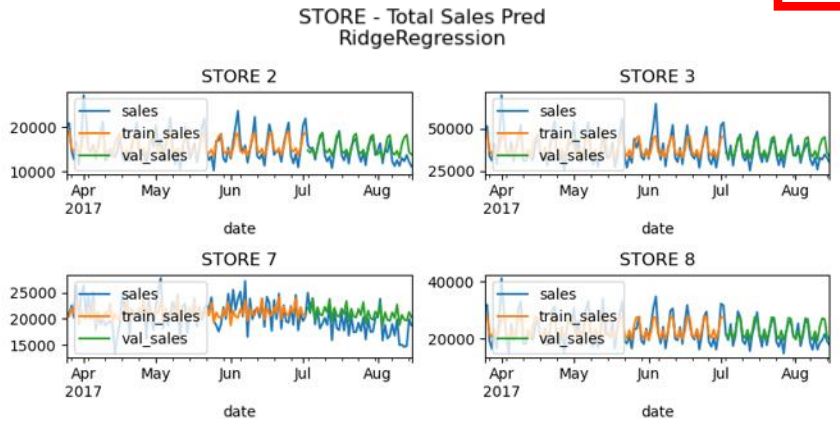
선형 회귀(Linear Regression)

RMSLE
0.44671

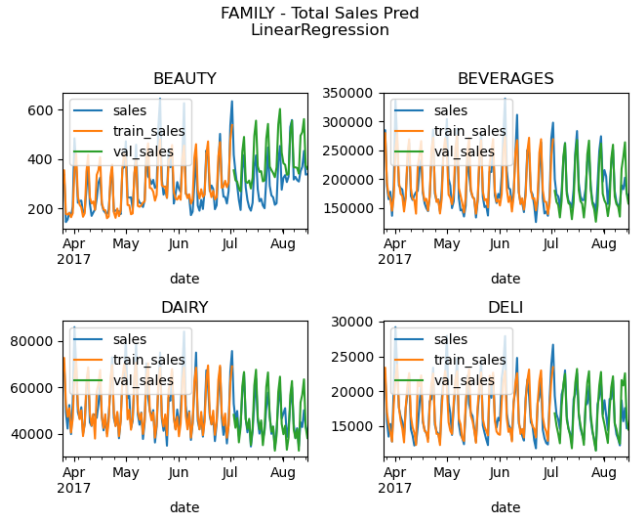


릿지 회귀(Ridge Regression)

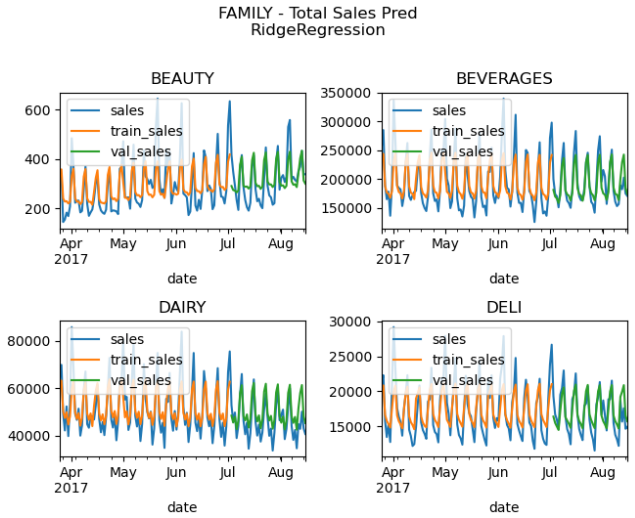
RMSLE
0.38648



매장별 예측



제품군별 예측



기존에 사용했던 선형 회귀 모델보다 릿지 회귀 모델이 더 나은 결과값을 출력하는 것을 확인할 수 있음.

04

대시보드

4-1 소개

4-2 데이터

4-3 탐색적 자료 분석

4-4 통계 분석

Main Menu

- INTRO
- DATA
- Exploratory Data Analysis
- STAT

메인 메뉴

Intro, Data, EDA, Stat
총 4개의 탭 중 원하는
페이지 선택 가능

소개

Intro의 3가지 탭 중 배경소개와 대회의
개요가 적혀있는 페이지

배경 소개 및 대회 개요

Intro의 3가지 탭 중 배경소개 및
대회의 개요가 적혀있는 페이지

Store Sales

소개 목표 분석 단계

대회 목표

- 이번 대회의 목표는 **시계열 예측** 을 사용하여 에콰도르에 본사를 두고 있는 대형 식료품 소매업체인 "Corporación Favorita"의 데이터를 분석하고 매장의 앞으로의 매출을 예측하는 것입니다.
- 구체적으로는 여러 Favorita 매장에서 판매되는 수많은 품목의 판매 단가를 보다 **정확하게 예측하는 모델을 구축하는 것이 최종 목표** 입니다.
- 날짜, 매장 및 품목 정보, 프로모션, 판매 단가로 구성된 비교적 접근성이 좋은 학습 데이터 셋을 통해 머신러닝 모델들을 연습할 수도 있습니다.

평가

- 이 대회의 평가 지표는 *Root Mean Squared Logarithmic Error* (평균 제곱근 오차)입니다.

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

- n n은 **총 인스턴스의 수** 입니다.
- $\hat{y}_i$ i는 인스턴스 i에 대한 **예측된 타겟 값** 입니다.
- y_i 는 인스턴스 i에 대한 **실제 타겟** 입니다.
- log 는 **자연 로그** 입니다.

대회 정보

More Detailed : Store Sales - Time Series Forecasting

목표, 분석 단계

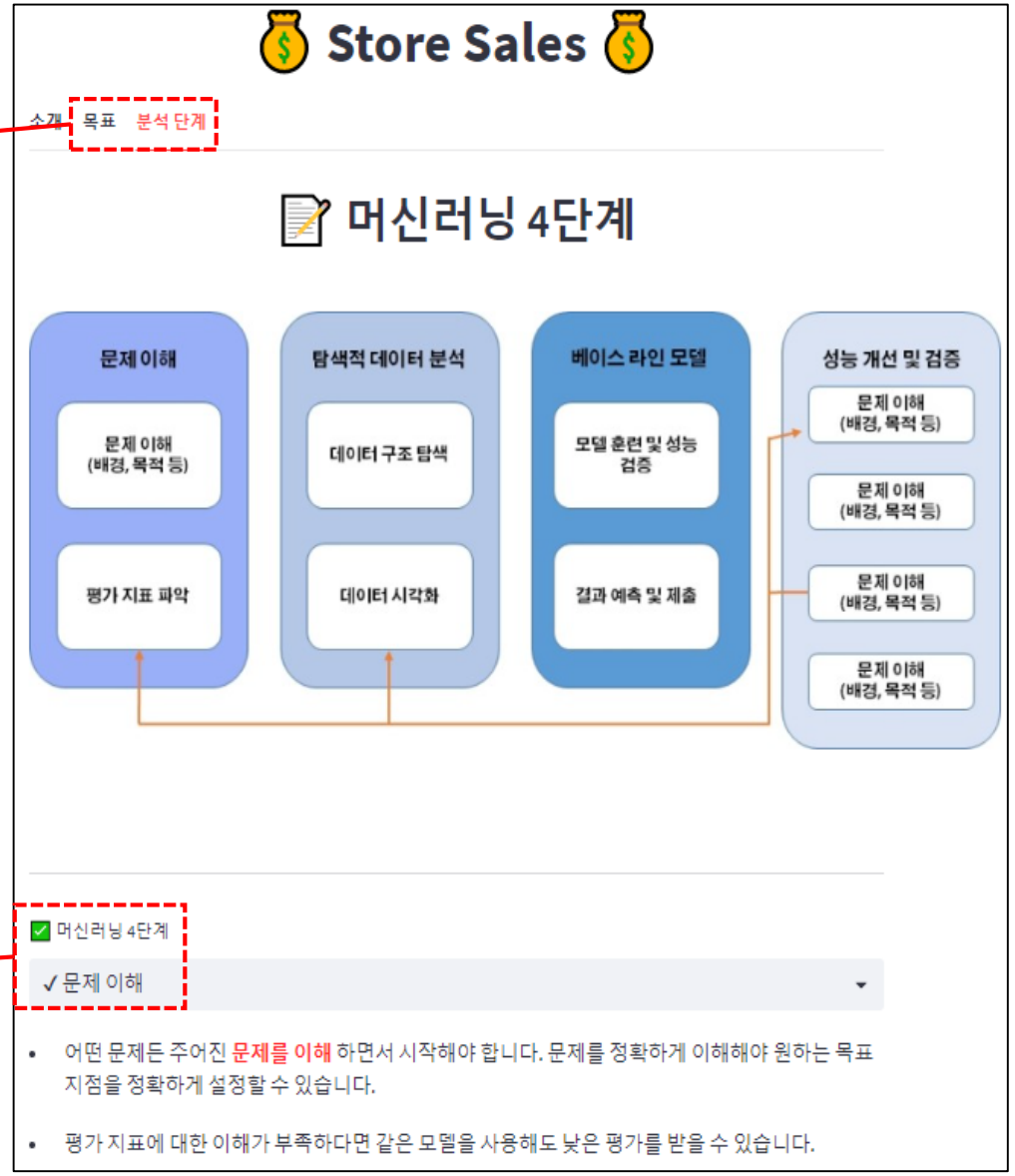
Intro의 3가지의 탭 중 대회의 최종 목표와 평가 지표. 그리고 대회의 정보를 찾아 볼 수 있는 페이지

대회 목표, 평가 대회 정보

대회의 최종 목표와 평가 지표 그리고 대회 정보에 대한 자세한 내용을 볼 수 있다.

머신러닝 4단계

Select box를 통해 머신러닝 4단계에 대한 자세한 정보를 볼 수 있다.



데이터 선택

Select box를 통해
데이터 6개에 대한
자세한 정보 확인 가능

데이터 간략 설명

Select box를 통해
고른 데이터의 간략한 설명

Store Sales

데이터 종류

Train

Train Data Description

- train Data 는 **상점 번호, 제품군, 프로모션 및 목표 매출** 로 구성된 시계열 데이터입니다.
- store_nbr 은 제품이 판매되는 **상점 번호** 를 나타냅니다.
- family 는 판매되는 **제품 유형** 을 나타냅니다.
- sales 는 특정 날짜에 특정 가게에서 **판매되는 제품군의 총 매출** 을 나타냅니다. (일부 제품은 소수점 단위로 판매될 수 있으므로 분수 값이 가능합니다.)
- onpromotion 은 특정 날짜에 상점에서 **프로모션 중인 제품군의 항목 수** 를 나타냅니다.

Describe

Describe : 데이터의 분포를 요약한 결과를 출력합니다. 이를 통해 데이터의 중심 경향성, 산포도 등을 쉽게 파악 가능

Data / Data Type

Data : 원본 데이터 확인 가능

Data Type : 데이터 유형 확인가능

Data

	id	date	store_nbr	family	sales	onpromotion	year	mon
0	1,297,296	2015-01-01	1	AUTOMOTIVE	0	0	2,015	
1	1,297,297	2015-01-01	1	BABY CARE	0	0	2,015	
2	1,297,298	2015-01-01	1	BEAUTY	0	0	2,015	
3	1,297,299	2015-01-01	1	BEVERAGES	0	0	2,015	
4	1,297,300	2015-01-01	1	BOOKS	0	0	2,015	
5	1,297,301	2015-01-01	1	BREAD/BAKERY	0	0	2,015	
6	1,297,302	2015-01-01	1	CELEBRATION	0	0	2,015	
7	1,297,303	2015-01-01	1	CLEANING	0	0	2,015	
8	1,297,304	2015-01-01	1	DAIRY	0	0	2,015	

Data Type

	0
id	int64
date	object
store_nbr	int64
family	object
sales	float64
onpromotion	int64
year	int64
month	int64
day	int64

Describe

	id	store_nbr	sales	onpromotion	year	mont
count	1,299,078	1,299,078	1,299,078	1,299,078	1,299,078	1,299
mean	1,946,834.5	27.5	407.6234	3.5424	2,015.5007	6.
std	375,011.6608	15.5858	1,201.05	14.7025	0.5	3.
min	1,297,296	1	0	0	2,015	
25%	1,622,065.25	14	1	0	2,015	
50%	1,946,834.5	27.5	17	0	2,016	
75%	2,271,603.75	41	237	1	2,016	
max	2,596,373	54	124,717	741	2,016	

데이터 선택

Select box를 통해
Train, Transaction, Oil
각 3가지 데이터에 대한
시각화 및 설명 확인 가능

세부 선택

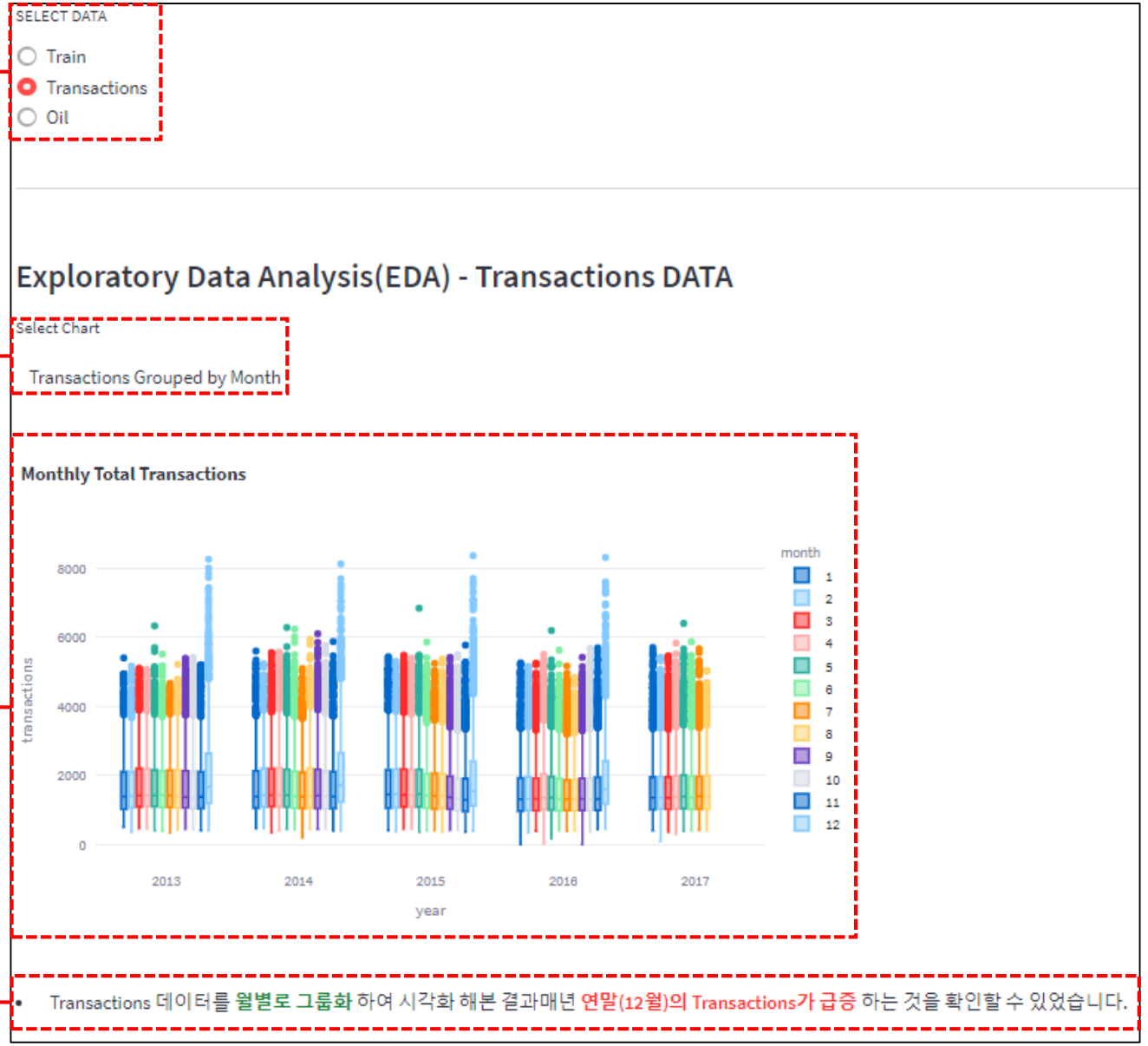
Select box를 통해
선택한 데이터를 세부적으로
나눠서 시각화한 데이터

시각화

세부 주제에 대한 시각화

데이터 선택

Select box를 통해
데이터 6개에 대한
자세한 정보 확인 가능



데이터 선택

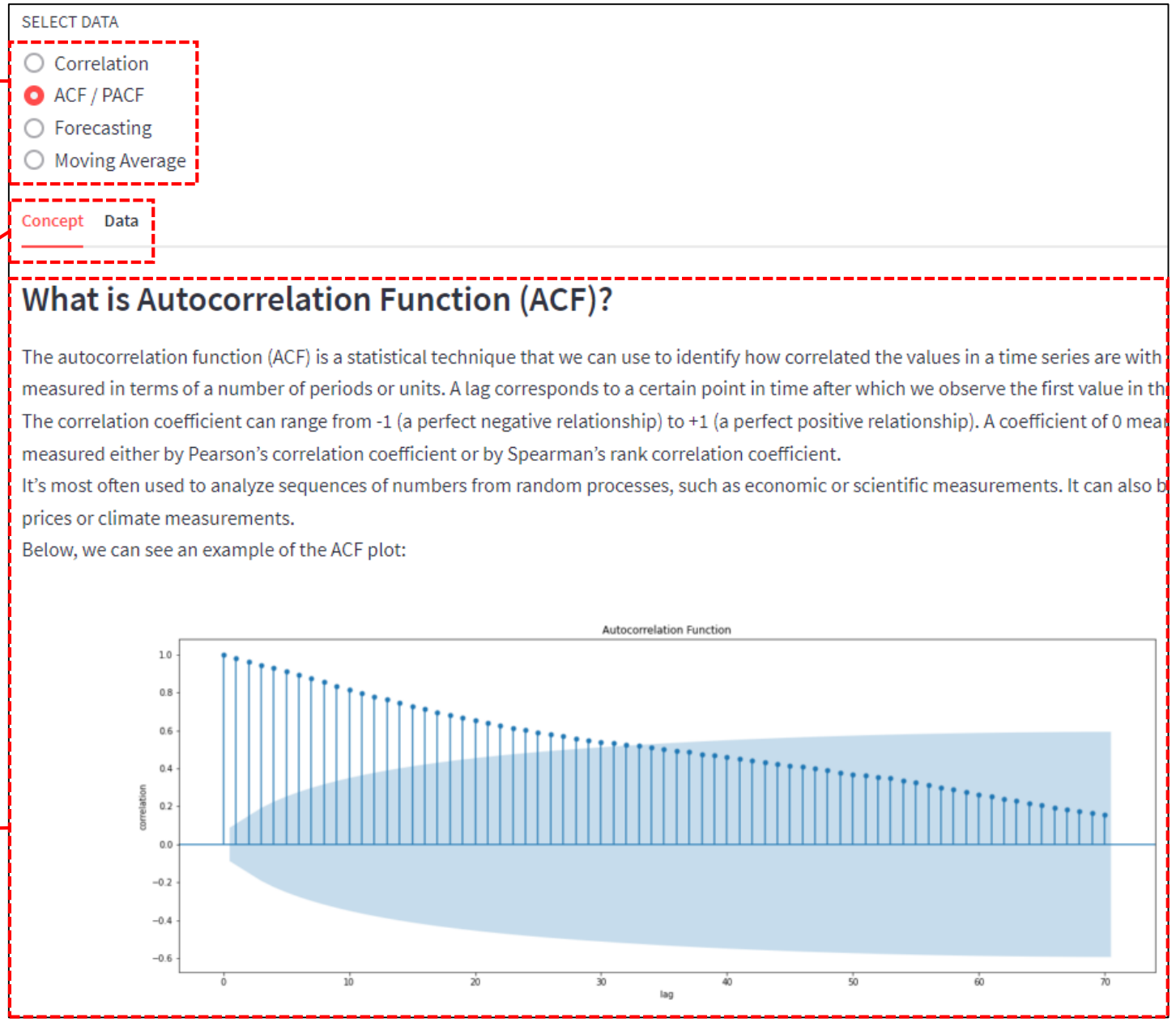
Radio Button을 통해
Correlation, ACF / PACF,
Forecasting, Moving Average
각 4가지 개념에 대한
시각화 및 내용 확인 가능

세부 선택

Concept / Data 탭을 통해
통계 분석에 대한 내용 확인

내용 / 시각화

분석에 사용한 개념들에 대한
내용과 Data 탭에서의
시각화



05

자체 평가

5-1 한계점

5-2 발전 가능성

한계점 및 발전 가능성

? 가장 기본적인 모델인 선형 회귀 모델을 사용하였을 때 RMSLE(오차율)가 약 0.44정도 나와서 만족스러운 결과를 얻지 못하였습니다.

✓ 그래서 더 적은 오차율을 찾기 위해 정보 수집 및 코드 분석을 하던 중 Ridge(릿지) 모델 사용에 대한 코드를 익히게 되었고 RMSLE 약 0.38로 더 나은 결과값을 도출할 수 있었습니다.

? ARIMA 모델, Prophet 모델 등을 적용해보지 못했으며, LSTM 모델의 경우 모델링 과정에서 잘못하여 너무 높은 값이 나오는 결과를 받았습니다.

✓ 더 많은 모델을 공부하는 계기가 되었고, 여러 모델을 응용하는 공부를 진행하다 보면 더 나은 결과값을 구현할 수 있을 것으로 생각됩니다.

? 이번 대시보드 웹 서비스 구현에서 Streamlit 라이브러리를 사용하여 구현하였으나 데이터 적재 문제로 인해 웹 서비스를 제작하는데 에러 사항이 있었습니다.

✓ 다음엔 다른 라이브러리를 사용하여 더 나은 웹 서비스를 구현해볼 수 있었으면 좋겠습니다.

PROPHET

ARIMA

TensorFlow

06

참고문헌

선형 회귀 : https://ko.wikipedia.org/wiki/%EC%84%A0%ED%98%95_%ED%9A%8C%EA%B7%80

릿지 회귀 : https://en.wikipedia.org/wiki/Ridge_regression

RMSLE : https://ahnjg.tistory.com/90

감사합니다