WALMART STORE SALES FORECASTING

회귀분석을 활용한
월마트 주간 판매량 예측

STORE SALES FORECASTING

INDEX

분석 목적

STORE SALES FORECASTING

분석 목적

💡 주제 선정 배경

대규모 소매점 데이터에 대한 팀원들의 높은 관심

🎯 목표

1. 스토어별 주간 판매량을 회귀 예측하는 모델 구축
2. 모델을 활용해 주간 보고서용 대시보드 구현하기

📦 기대효과

재고 관리 및 마케팅 전략 수립 시 필요한 인사이트를 제공

데이터 소개

STORE SALES FORECASTING

테이블정보

Table 1 | Stores

45 (rows) X 3 (columns)

변수명	변수 설명	변수 타입	비고
Store	지점 번호	Int	1~45
Type	매장 타입	Str	A, B, C
Size	매장 면적	Int	-

Table 2 | Train

421570 (rows) X 5 (columns)
2010-02-05 ~ 2012-10-26 (995 일간)

변수명	변수 설명	변수 타입	비고
Store	지점 번호	Int	1~45
Dept	부서 번호	Int	1~99 (81개)
Date	날짜	TimeSeries	'yyyy-mm-dd' 형태
Weekly_Sales	주간 매출	Float	-
IsHoliday	공휴일 여부	Boolean	-

Table 3 | features

8190 (rows) X 12 (columns)
2010-02-05 ~ 2013-07-26 (1268 일간)

변수명	변수 설명	변수 타입	비고
Store	지점 번호	Int	1~45
Date	날짜	TimeSeries	'yyyy-mm-dd' 형태
Temperature	주간 평균 기온	Float	경제지표
Fuel_Price	주간 평균 유가	Float	경제지표
MarkDown	홍보 마크다운	Float	1~5번 컬럼까지 존재 (결측치 존재)
CPI	주간 소비자물가지수	Float	경제지표
Unemployment	주간 실업률	Float	-
IsHoliday	공휴일 여부	Boolean	-

STORE SALES FORECASTING

테이블 스키마

별첨자료 1 참고

Store

Store	Int
Type	Str
Size	Int

데이터 측정 범위

2010-02-05 ~ 2012-10-26 (995 일간)

train

Store	Int
Dept	Int
Date	Date
Weekly_Sales	Float
IsHoliday	Boolean

데이터 측정 범위

2010-02-05 ~ 2013-07-26 (1268 일간)

feature

Store	int
Date	date
Temperature	float
Fuel_Price	float
MarkDown1~5	float
CPI	float
Unemployment	float
IsHoliday	boolean

데이터 탐색 및 전처리

STORE SALES FORECASTING

사전 전처리

상관관계 분석 위해 Type, IsHoliday 컬럼을 라벨 인코딩

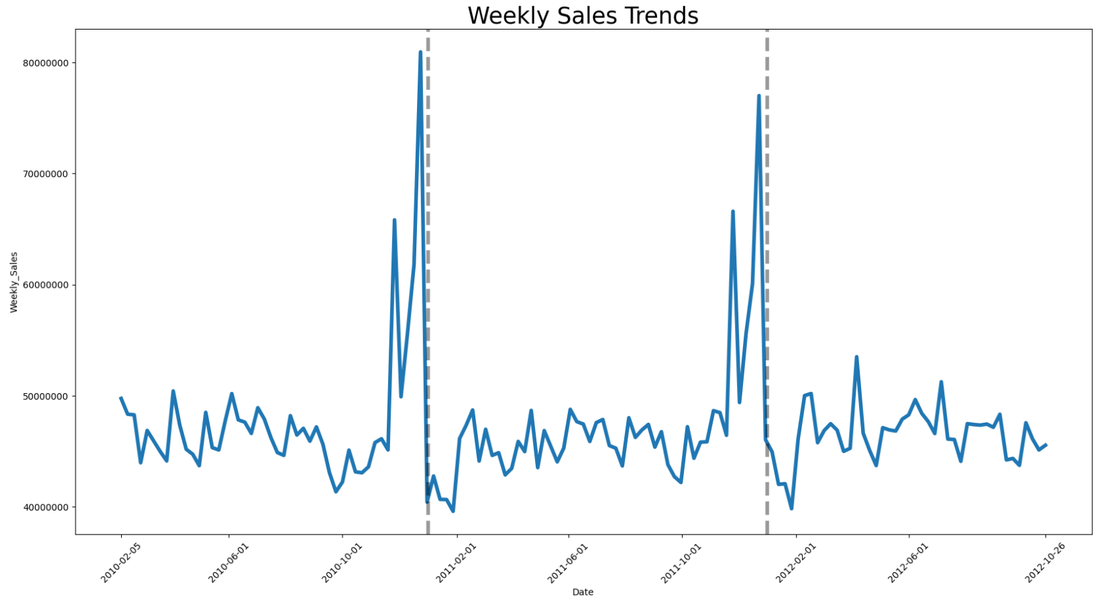
Store	Dept	Date	Weekly_Sales	IsHoliday	Type	Size	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment
1	1	2010-02-05	24924.50	False	A	151315	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106
1	2	2010-02-05	50605.27	False	A	151315	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106
1	3	2010-02-05	13740.12	False	A	151315	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106
1	4	2010-02-05	39954.04	False	A	151315	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106
1	5	2010-02-05	32229.38	False	A	151315	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106

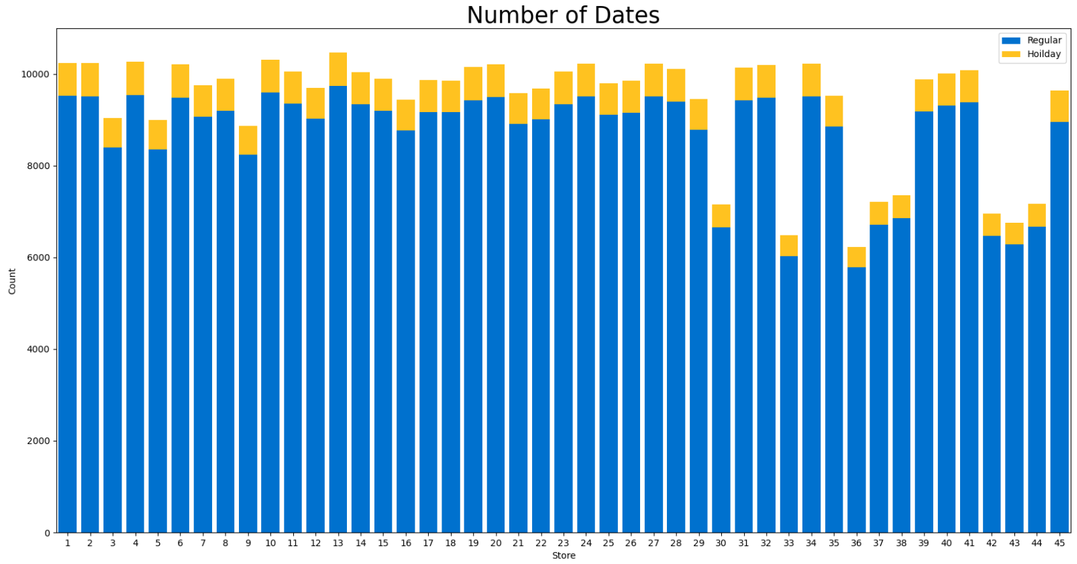
Store	Dept	Date	Year	Month	Day	Weekly_Sales	IsHoliday	Type	Size	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment
1	1	2010-02-05	2010	2	5	24924.50	0	0	151315	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106
1	2	2010-02-05	2010	2	5	50605.27	0	0	151315	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106
1	3	2010-02-05	2010	2	5	13740.12	0	0	151315	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106
1	4	2010-02-05	2010	2	5	39954.04	0	0	151315	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106
1	5	2010-02-05	2010	2	5	32229.38	0	0	151315	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106

Date (YYYY-MM-DD) 컬럼을 Year, Month, Day로 분리
컬럼수 11개 ▶ 19개

STORE SALES FORECASTING

데이터 탐색





Weekly Sales는 계절성이 있음

Date를 기준으로 Weekly Sales의 합계를 확인 한 결과 계절성이 확인됨.

각 store별 수집된 날짜 구간은 상이하나 Holiday의 비율은 비슷

각 Store별 데이터가 수집된 날짜들이 상이하다는 사실을 확인,

그러나 Holiday의 비율 7% 로 비슷함.

STORE SALES FORECASTING

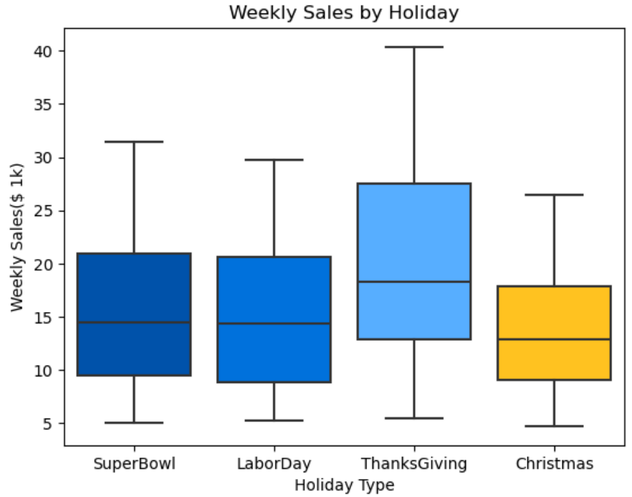
데이터 탐색

다음과 같이 각 월의 특정 주간은 Holiday주간 (IsHoliday = T/F)이다.

슈퍼볼 (2월), 노동절 (9월), 추수감사절(11월), 크리스마스(12월)

Holiday 주간의 매출은 평균적으로 일반적인 경우보다 높다.

⁉ 어떤 Holiday 주간의 Weekly Sales가 높을까?



- 평균적으로 Weekly Sales가 높은 주간은 **Thanksgiving** 기간이다.
- 평균 Weekly Sales가 가장 낮은 주간은 **Christmas** 기간이다.

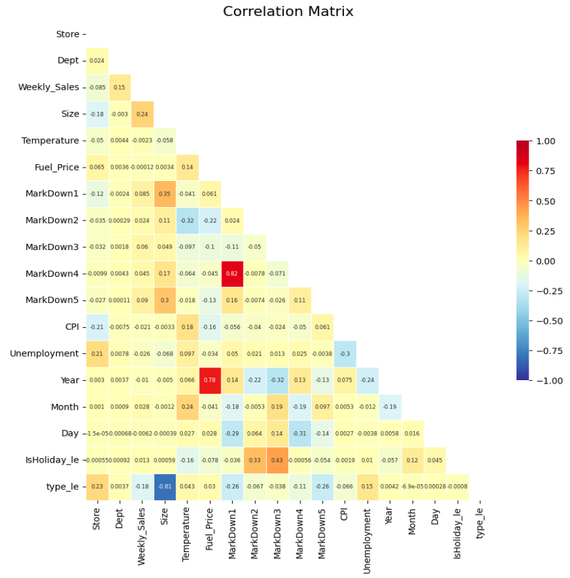
STORE SALES FORECASTING

상관관계 및 초기변수 탐색

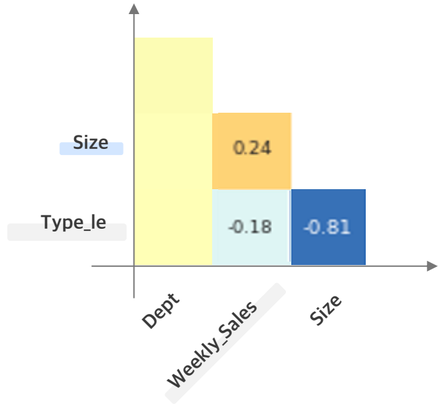
별첨자료 2 참고

초기변수

Store Dept Year Month Day Size Weekly Sales
IsHoliday_le Fuel_Price Temperature CPI Unemployment



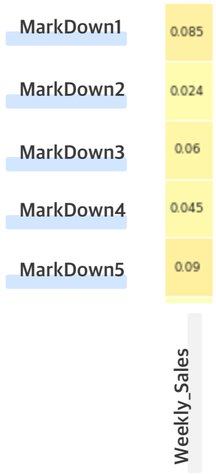
상관관계 | SIZE & TYPE

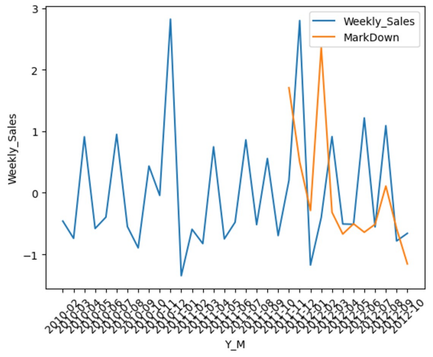


두 컬럼이 높은 관계성을 가진 것을 확인.

SIZE와 TYPE 중 Weekly Sales와 관계성이 더 높은 Size만 선택

상관관계 | MARKDOWN





마크다운 결측치 : 전체의 64%

컬럼 정보 부족 & 다량의 결측치

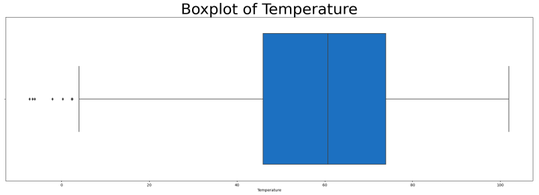
MarkDown1~5 의 데이터 수집이 짧은 기간 진행되어
유의미한 분석이 불가능 할 것으로 판단

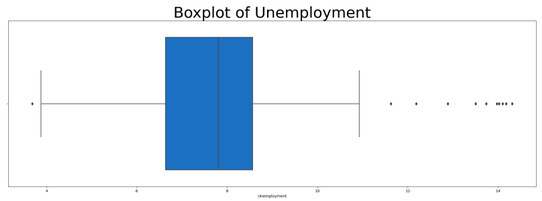
STORE SALES FORECASTING

데이터 표준화

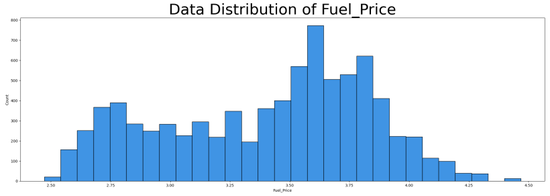
데이터 표준화

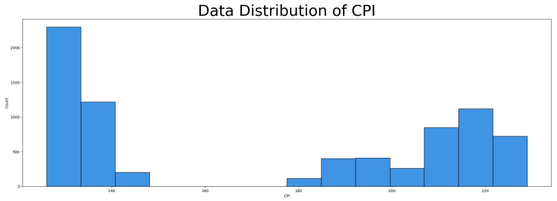
이상치 데이터





다중 분포형태 데이터





이상치 데이터에 취약

1. MinMaxScaler

2. StandardScaler

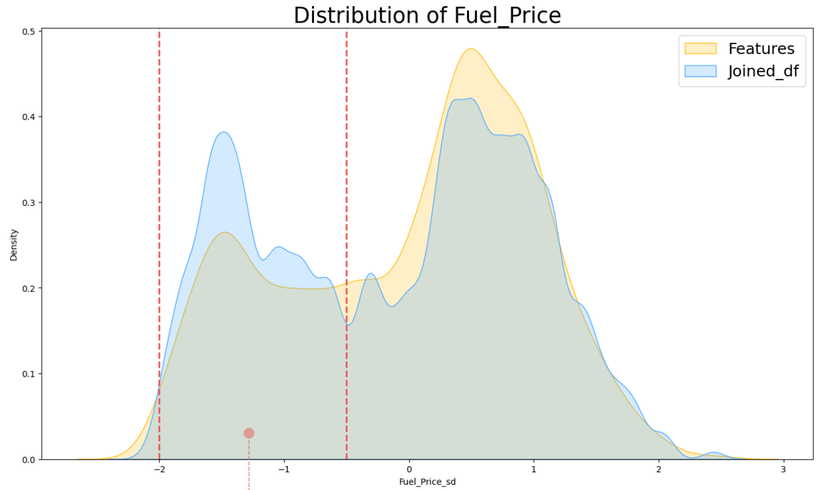
‘Temperature’, ‘Unemployment’

‘Size’ ‘Fuel_Price, ‘CPI’

컬럼명에 ‘_sd’ 추가

단, 적용 시 Join 후 테이블이 아닌

기존 ‘Store’와 ‘Feature’ 테이블에서 Fitting 후 스케일링 적용



‘Feature’ 테이블의 분포와, 조인된 테이블의 분포가 다른 것을 확인

원 분포를 활용하기 위해 ‘Store’와 ‘Feature’ 테이블에서 Fitting 후 스케일링 적용

STORE SALES FORECASTING

데이터 인코딩

별첨자료 3 참고

범주형 인코딩

Year

IsHoliday (Boolean)

Store

Dept

LabelEncoding 적용

두 가지 이상의 클래스를 가진 변수

인코딩 방식	One-Hot Encoding	기존 형식 유지 (1-45/1-99)
Feature 개수	135개	11개

One-Hot Encoding시, Feature의 개수 영향으로
성능 저하 예상 → **기존 데이터 형식 유지**

	Store	Dept	Year_le	Month	Day	IsHoliday_le	Size_sd	Temperature_sd	Fuel_Price_sd	CPI_sd	Unemployment_sd
0	1	1	0	2	5	0	0.333175	-0.912661	-1.933624	0.972312	0.148726
1	1	2	0	2	5	0	0.333175	-0.912661	-1.933624	0.972312	0.148726
2	1	3	0	2	5	0	0.333175	-0.912661	-1.933624	0.972312	0.148726
3	1	4	0	2	5	0	0.333175	-0.912661	-1.933624	0.972312	0.148726
4	1	5	0	2	5	0	0.333175	-0.912661	-1.933624	0.972312	0.148726
...	...	...	...	...	...	...	...	...	...	...	...
421565	45	93	2	10	26	0	-0.191193	-0.027102	1.103633	0.499502	0.447586
421566	45	94	2	10	26	0	-0.191193	-0.027102	1.103633	0.499502	0.447586
421567	45	95	2	10	26	0	-0.191193	-0.027102	1.103633	0.499502	0.447586
421568	45	97	2	10	26	0	-0.191193	-0.027102	1.103633	0.499502	0.447586
421569	45	98	2	10	26	0	-0.191193	-0.027102	1.103633	0.499502	0.447586

421570 rows × 11 columns

최종 선택된 컬럼

모델 학습

STORE SALES FORECASTING

모델학습 세팅

🍀 Random State = 73 고정

Train - Test 분리

Train(70%), Test(30%)

IsHoliday 기준 **층화추출(Stratify)***

평가 지표

본 프로젝트에서는 **RMSE, R^2, WMAE***

세 가지 평가지표를 사용합니다.

WMAE

$$WMAE = \frac{1}{\sum \omega_i} \sum_{i=1}^{n} \omega_i |y_i - \hat{y}_i|$$

- n : 행의 수
- $\hat{y}_i$: 예측 가격
- y_i : 실제 가격
- ω_i : 가중치

Holiday 주간인 경우 가중치로 5를 적용하고, 아닌 경우 1으로 한다.

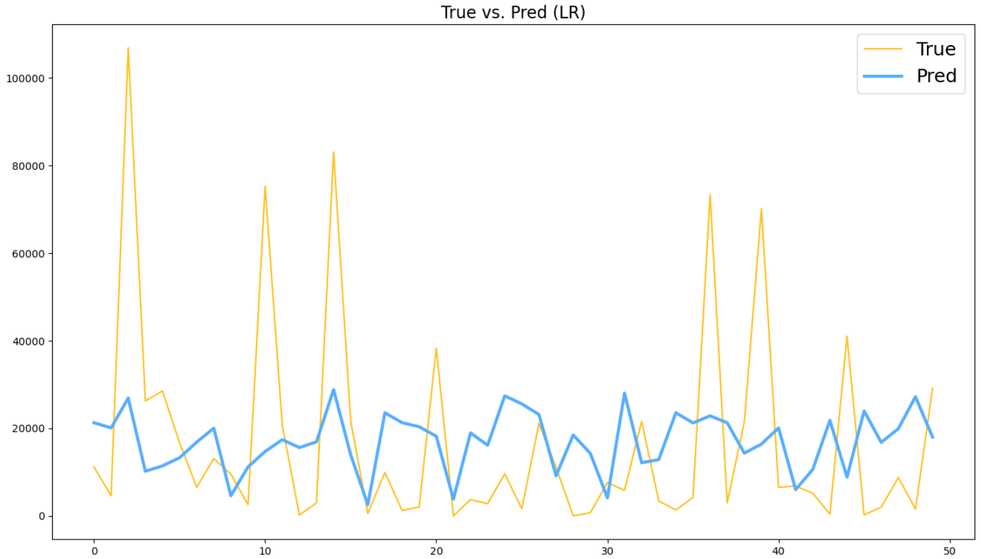
* **층화추출**

각 클래스의 비율을 맞추어 샘플을 선택함으로써, 효과적인 클래스별 분석과 모델 학습을 가능하게 하는 샘플 추출법

STORE SALES FORECASTING

선형 회귀 모델

별첨자료 4 참고



성능이 매우 낮음

WMAE	RMSE	R-Squared
14793.8502	21663.8066	0.0845

사용한 변수 간 비선형 관계가 강하게 보임

Store , Dept , Year , Month , Day , IsHoliday ,
Size , Temperature , Fuel_Price , CPI , Unemployment

학습에 사용한 모든 독립변수들과 종속변수 간
피어슨 상관계수가 절댓값 0.3 미만으로 **선형 관계가 약함**

Ramsay RESET Test 결과, 독립변수들과 종속변수 간 관계가
비선형성을 강하게 띄고 있음을 확인 가능

→ 선형 기반 회귀 모델은 비선형 관계 포착에 한계가 있음.

STORE SALES FORECASTING

트리기반앙상블모델

따라서 비선형 관계를 잘 포착하는

트리기반 **앙상블 모델*** 사용

RandomForest, Gradient Boosting, XGBoost

앙상블 모델*

여러 개의 개별 모델을 조합하여 최적의 모델로 일반화하는 방법.

의사결정나무의 **과대적합 문제를 해결함.**

보팅(voting), 배깅(bagging), 부스팅(boosting), 스태킹(stacking)이 있음.

Ver 1

선택 변수 Store, Dept, Year_le, Day, IsHoliday_le, Size_sd, Temperature_sd, Fuel_Prie_sd, CPI_sd, Unemployment_sd

정규화 일부 변수(Store, Dpet, Month, Day) 제외
Sklearn.preprocessing.Standard Scaler 활용

인코딩 Year, IsHoliday 라벨 인코딩

	R^2	RMSE	WMAE
GBM	0.7407	11529.9645	7143.3501
RF	0.9747	3604.8407	1602.6524
XgBoost	0.9393	5576.5045	3264.2371

STORE SALES FORECASTING

모델 학습 타임라인

STORE SALES FORECASTING

모델 변수 선택

💡 가설 1

Week가 Day보다 데이터의 시간성을 잘 반영할 것이다.

실제로는 2010년 10월 30일 부터 2010년 2월 5일까지의 **주간 판매량 데이터**

Store	Dept	Date	Weekly_Sales	Year	Month	Day	Week
1	1	2010-02-05	24924.50	2010	2	5	5
1	2	2010-02-05	50605.27	2010	2	5	5
1	3	2010-02-05	13740.12	2010	2	5	5
1	4	2010-02-05	39954.04	2010	2	5	5
1	5	2010-02-05	32229.38	2010	2	5	5
...	...	...	...	...	...	...	...

Week : 일년 중 해당 주간이 몇 번째 주간인지 숫자로 표기(1~52)
e.g.해당 주의 경우 2010년의 5번째 주차

Ver 2

선택 변수 Store , Dept , Year_le , Week , IsHoliday_le , Size_sd , Temperature_sd , Fuel_Prie_sd , CPI_sd , Unemployment_sd

	R^2	RMSE	WMAE
GBM	0.744 ▲	11456.2127 ▼	7086.2212 ▼
RF	0.9739 ▼	3657.4015 ▲	1657.7968 ▲
XgBoost	0.9446 ▲	5328.8998 ▼	3109.6522 ▼

Ver1 대비 증감(■ 성능 개선, ■ 성능 저하)

랜덤 포레스트 모델을 제외하고 나머지 모델에서 성능 개선

STORE SALES FORECASTING

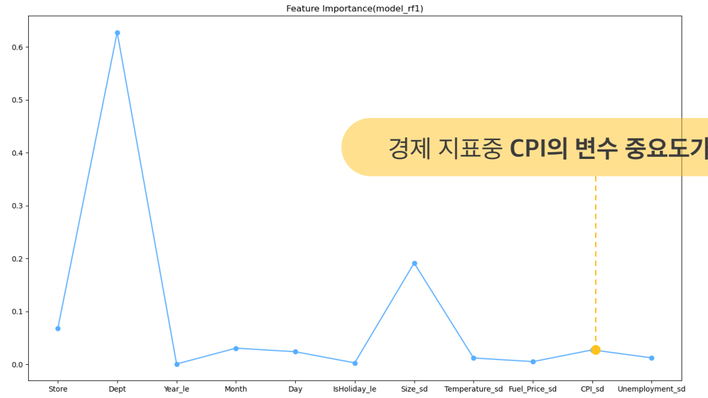
모델 변수 선택

💡 가설 2

경제 지표를 대표하는 **CPI만 사용**하면 **모델의 성능이 개선**될 것이다.

	CPI	Fuel_Price	Unemployment
CPI	1.000000	-0.164210	-0.299953
Fuel_Price	-0.164210	1.000000	-0.033853
Unemployment	-0.299953	-0.033853	1.000000

다른 지표 대비 **경제 지표 간의 상관성이 높다.**



경제 지표중 **CPI의 변수 중요도가 가장 높음**

Ver 3

선택 변수 Store , Dept , Year_le , Week , Day , IsHoliday_le , Size_sd , Temperature_sd , CPI_sd

	R^2	RMSE	WMAE
GBM	0.7434 ▲	11468.6906 ▼	7091.2188 ▼
RF	0.976 ▲	3507.1593 ▼	1579.366 ▼
XgBoost	0.9449 ▲	5316.2909 ▼	3106.3117 ▼

Ver1 대비 증감(■ 성능 개선, ■ 성능 저하)

모든 모델에서 성능 개선

STORE SALES FORECASTING

모델 변수 선택

앞의 가설을 데이터의 시간성을 잘 반영하는 **Week를 Month와 Day 대신 사용**,
경제지표를 대표하는 CPI만 선택해 변수를 선정함.

Ver 4

선택 변수 Store , Dept , Year_le , Week , IsHoliday_le , Size_sd , Temperature_sd , CPI_sd

	R^2	RMSE	WMAE
GBM	0.7448 ▲	11438.6736 ▼	7115.2914 ▼
RF	0.975 ▲	3578.0588 ▼	1634.702 ▲
XgBoost	0.95 ▲	5061.3442 ▼	3000.1364 ▼

Ver1과 32.05 차이

Ver1 대비 증감(■ 성능 개선, ■ 성능 저하)

전반적으로 모든 모델에서 성능 개선

STORE SALES FORECASTING

모델 변수 선택

💡 가설 3

경제 지표 중 **CPI만으로는 주간 판매량을 예측하기 어려울 것이다.**

Ver 5

유가(Fuel_Price) 컬럼 추가

선택 변수 Store , Dept , Year_le , Week , IsHoliday_le , Size_sd , Temperature_sd , CPI_sd , Fuel_Prie_sd

	R^2	RMSE	WMAE
GBM	0.7448 -	11438.6736 -	7115.2914 -
RF	0.9744 ▼	3619.3998 ▲	1651.4686 ▲
XgBoost	0.9456 ▼	5279.5781 ▲	3089.7855 ▲

Ver4 대비 증감(■ 성능 개선, ■ 성능 저하)

모든 모델에서 성능이 개선되지 않았다.

Ver 6

실업률(Unemployment) 컬럼 추가

선택 변수 Store , Dept , Year_le , Week , IsHoliday_le , Size_sd , Temperature_sd , CPI_sd , Unemployment_sd

	R^2	RMSE	WMAE
GBM	0.744 ▼	11456.2127 ▲	7086.2212 ▼
RF	0.9743 ▼	3627.1725 ▲	1643.6511 ▲
XgBoost	0.9475 ▼	5188.031 ▲	3057.5611 ▲

Ver4 대비 증감(■ 성능 개선, ■ 성능 저하)

모든 모델에서 성능이 개선되지 않았다.

STORE SALES FORECASTING

모델 변수 선택

💡 가설 4

변수 **Year를 라벨 인코딩**한 것은 모델의 **성능에 영향을 주지 않을 것이다.**

Store	Dept	Date	Year_le
1	1	2010-02-05	0
1	2	2010-02-05	0
1	3	2010-02-05	0
1	4	2010-02-05	0
1	5	2010-02-05	0
...	...	...	...
45	93	2012-10-26	2
45	94	2012-10-26	2
45	95	2012-10-26	2
45	97	2012-10-26	2
45	98	2012-10-26	2

0 → 2010
1 → 2011
2 → 2012

Store	Dept	Date	Year
1	1	2010-02-05	2010
1	2	2010-02-05	2010
1	3	2010-02-05	2010
1	4	2010-02-05	2010
1	5	2010-02-05	2010
...	...	...	...
45	93	2012-10-26	2012
45	94	2012-10-26	2012
45	95	2012-10-26	2012
45	97	2012-10-26	2012
45	98	2012-10-26	2012

Ver 7

선택변수 Store , Dept , Year , Week , IsHoliday_le , Size_sd , Temperature_sd , CPI_sd

	R^2		RMSE		WMAE	
GBM	0.7448	-	11438.6736	-	7115.2914	-
RF	0.975	-	3578.0588	-	1634.702	-
XgBoost	0.95	-	5061.3442	-	3000.1364	-

Ver4 대비 증감(■ 성능 개선, ■ 성능 저하)

모든 모델의 성능이 변하지 않았다.
고로, **연도(Year) 변수는 라벨링 하지 않아도 된다.**

STORE SALES FORECASTING

모델 최적화

하이퍼 파라미터 조정

앞의 가설을 데이터의 시간성을 잘 반영하는 **Week를 Month와 Day 대신 사용**,
경제지표를 대표하는 CPI만 선택해 변수를 선정,
변수 **Year의 라벨 인코딩이 불필요하다고 판단해** 인코딩 하지 않음.

| 랜덤 포레스트 & 랜덤서치

	R^2	RMSE	WMAE
rf7	**0.9750**	**3578.06**	**1643.70**
rf8	0.9309	5951.90	3127.57
rf9	0.9644	4274.06	1907.84

| 세팅 파라미터

rf 8 params

n_estimators=300, min_sample_split= 8, min_sample_leaf= 12, max_depth= 12

rf 9 params

n_estimators=300, min_sample_split= 8, min_sample_leaf= 12

!

rf8단계에서 랜덤서치 결과인 최적화한 파라미터 사용시
전반적으로 **성능이 크게 저하되었다.**

rf9단계에서 **max depth만 'None'으로 설정하여**
무한대로 늘린 경우, 부분적으로 성능이 회복되었지만
처음보다 과소적합이 진행되었다.

STORE SALES FORECASTING

모델 최적화

하이퍼 파라미터 조정

| Xgb & 그리드 서치

	R^2	RMSE	WMAE
Xgb7(M : 6 / N : 100)	0.9500	5061.34	3000.14
Xgb8	0.9850	2774.57	1432.85
Xgb9(M : 10 / N : 1,000)	**0.9862**	**2660.32**	**1385.19**

| 세팅 파라미터

Xgb 8 params

colsample_bytree=0.9, learning_rate=0.3, max_depth=10, min_child_weight=5, n_estimators=630

Xgb 9 params

colsample_bytree =0.9, learning_rate =0.3, max_depth =10, min_child_weight =5,

n_estimators=1000 , lambda=10, alpha=2

Xgb 8에서 **그리드 서치 결과 반영**,

max_depth & n_estimators의 영향으로 **비약적인 성능 상승**

원래 Xgb모델의 기본값 max depth = 6 ▶ 10

Xgb 9에서 성능 상승을 위해 **n_estimators를 키우고,**

과적합 방지 목적으로 **alpha, lambda를 통해 L1, L2 규제**

STORE SALES FORECASTING

모델 최적화

하이퍼 파라미터 조정

| 랜덤 포레스트

	R^2	RMSE	WMAE
rf7	0.9750	3578.06	1643.70

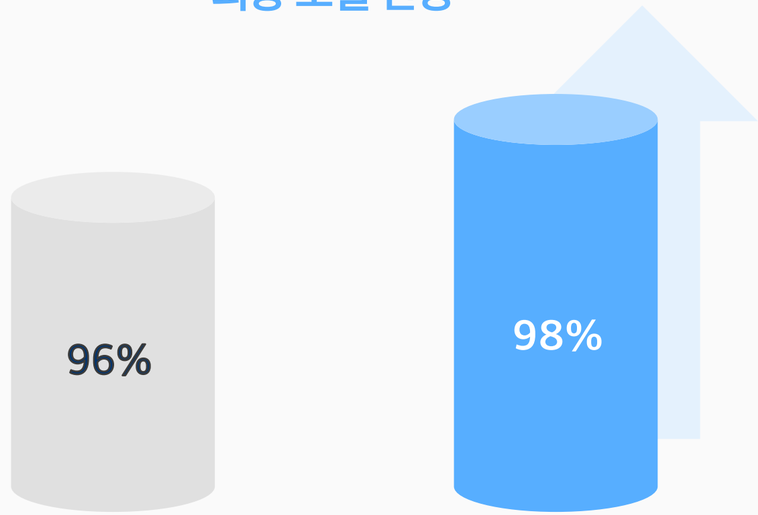
| Xgb

	R^2	RMSE	WMAE
Xgb9(M : 10 / N : 1,000)	**0.9862**	**2660.32**	**1385.19**

| XGB9 RF7 의 성능차이

	R^2	RMSE	WMAE
Xgb9 RF7 의 성능차이	0.0218	1613.74	522.65

최종 모델 선정



Random Forest XGBoost

TRAIN 295,100 (개)

TEST 126,500 (개)

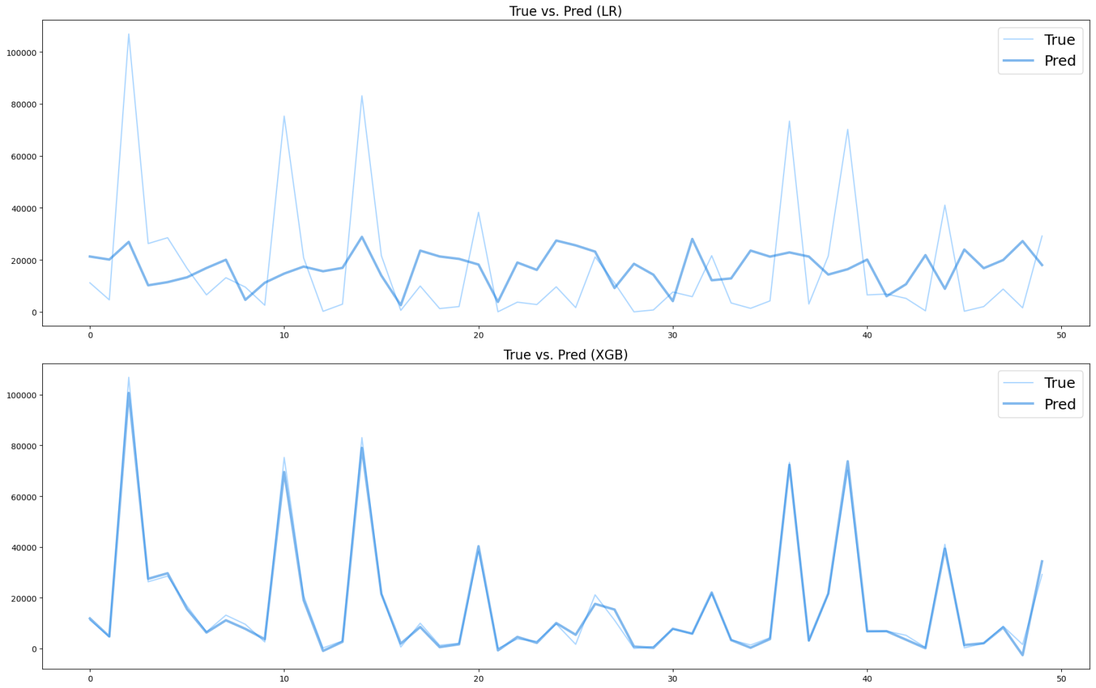
rf7 학습시간 | 95.15 & 예측 3.8 (초)

Xgb9 학습시간 | 5.6 & 예측 0.36 (초)

STORE SALES FORECASTING

모델 최적화

최종 모델과의 비교



	Linear Regression	XgBoost (ver9)	
R^2	0.0845	0.9862	▲ 91.43%
RMSE	21663.8066	2660.32	▼ 87.72%
WMAE	14793.8502	1385.19	▼ 90.64%

초기 선형회귀모델과 비교했을 때, 하이퍼파라미터 세팅을 마친

최종모델(Xgboost)의 예측 정확도가 높다.

결론

STORE SALES FORECASTING

프로젝트 요약

최종 모델 소개

R-squared
모델의 설명력
98.62%

WMAE
공휴일을 반영한 오차(MAE)
1385.118

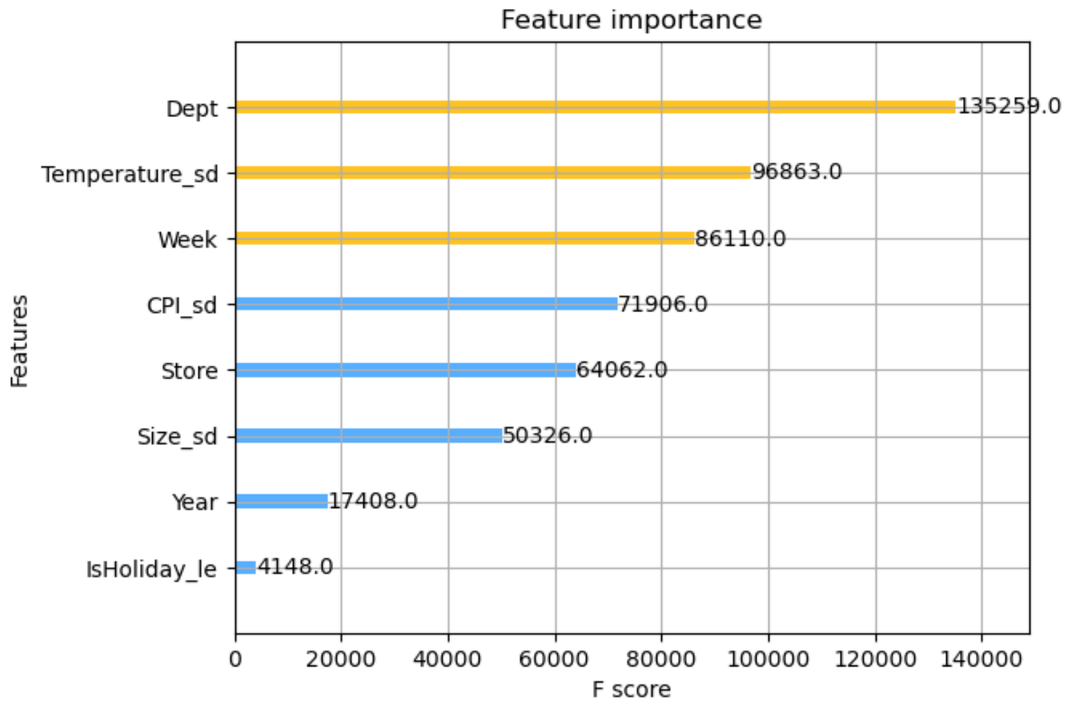
Ver 9

선택 변수 Store , Dept , Year , Week , IsHoliday , Size , Temperature , CPI

정규화 일부 변수(Store, Dept) 제외
Sklearn.preprocessing.Standard Scaler 활용

인코딩 IsHoliday 라벨 인코딩

최종 모델의 변수 중요도 (F-Score)



STORE SALES FORECASTING

대시보드

최종 모델을 기반으로

월마트의 n번 매장의 주간 보고서를 목적으로 하는

대시보드를 다음과 같이 생성하였다.

Walmart Store No.45

주간 매출 보고서

- 2011년도 36주차 결산 -

노동절 주간

Holiday 기간 표시

연누계 (천달러 단위)	Best Dept	Worst Dept
$ 27,093.68	95	18

주간 총 매출 (천달러 단위)	지난 주 총 매출	작년 동기 총 매출
$ 746.13	$ 726.48 +3%	$ 721.46 +3%

부서 분석

1

주간 총 매출	지난 주 총 매출	작년 동기 총 매출
$ 11,402.12	$ 11,142.98 +2%	$ 11,730.76 -3%

매장 전체 금주 예측치	1번 부서 금주 예측치
$ 745,815.94	$ 11,402.12

Graph Table

SQl 쿼리 조회

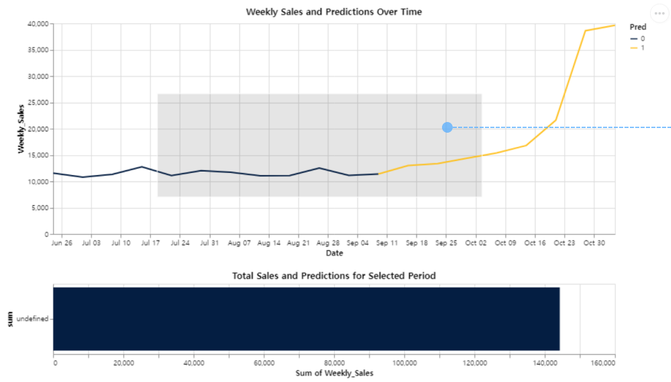
Graph Table

20 rows 9 columns 180 cells Run SQL Query Export

```
SELECT * FROM $table
```

Run Query Reset Data

	Date	Week	IsHoliday	Weekly_Sales	MarkDown1	MarkDown2
1	2011-06-24	25	false	11570.03		
2	2011-07-01	26	false	10796.27		
3	2011-07-08	27	false	11346.06		
4	2011-07-15	28	false	12773.75		
5	2011-07-22	29	false	11090.63		
6	2011-07-29	30	false	12019.61		
7	2011-08-05	31	false	11745.45		
8	2011-08-12	32	false	11060.92		
9	2011-08-19	33	false	11096.27		
10	2011-08-26	34	false	12543.74		



주간 매출 예측 그래프 (반응형)

별첨자료 1

데이터 기술통계량

	Store	Dept	Date	Weekly_Sales	Type	Size	Temperature	Fuel_Price	MarkDown					CPI	Unemployment	IsHoliday
									1	2	3	4	5			
count	421570	421570	421570	421570	421570	421570	421570	421570	150681	111248	137091	134967	151432	421570	421570	421570
unique	-	-	143	-	3	-	-	-	-	-	-	-	-	-	-	2
top	-	-	2011.12.23	-	A	-	-	-	-	-	-	-	-	-	-	FALSE
freq	-	-	3027	-	215478	-	-	-	-	-	-	-	-	-	-	391909
mean	22.2005	44.2603	-	15981.2581	-	136727.9157	60.0901	3.3610	7246.4202	3334.6286	1439.4214	3383.1683	4628.9751	171.2019	7.9603	-
std	12.7853	30.4921	-	22711.1835	-	60980.5833	18.4479	0.4585	8291.2213	9475.3573	9623.0783	6292.3840	5962.8875	39.1593	1.8633	-
min	1	1	-	-4988.9400	-	34875	-2.06	2.472	0.27	-265.76	-29.1	0.22	135.16	126.064	3.879	-
25%	11	18	-	2079.6500	-	93638	46.68	2.933	2240.27	41.6	5.08	504.22	1878.44	132.0227	6.891	-
50%	22	37	-	7612.0300	-	140167	62.09	3.452	5347.45	192	24.6	1481.31	3359.45	182.3188	7.866	-
75%	33	74	-	20205.8525	-	202505	74.28	3.738	9210.9	1926.94	103.99	3595.04	5563.8	212.4170	8.572	-
max	45	99	-	693099.3600	-	219622	100.14	4.468	88646.76	104519.54	141630.61	67474.85	108519.28	227.2328	14.313	-

별첨자료2

마크다운 컬럼을 포함한 모델 검증

마크다운 컬럼의 정보가 존재하지 않지만,
마케팅과 연관된 매출데이터라고 판단하여 결측치가 없는 기간을 추출해 예측모델을 구축하였다.

| 사용 데이터

MarkDown이 기록된 이후 데이터
Train : 106,002개 / Test : 45,430개

| 사용 컬럼

Xgb
Store, Dept, Year, Week, IsHoliday_le, Size_sd, Temperature_sd, CPI_sd

Xgb_s
XGB + Markdown_sum

Xgb_m
XGB + Markdown1~5

| 각 XGB 모델별 평가지표

	R^2	RMSE	WMAE
Xgb	**0.9507**	**5115.6780**	**3125.6781**
Xgb_s	0.9466 ▼	5323.0853 ▲	3215.9792 ▲
Xgb_m	0.9437 ▼	5469.6044 ▲	3316.4209 ▲

Xgb_s | MarkDown 컬럼들을 **연산(합, 평균)**하여 학습에 사용한 모델의 성능이 기존 모델보다 전반적으로 하락하였다.

Xgb_m | MarkDown 컬럼들을 **모두 포함**하여 학습에 사용한 결과, 연산하여 사용한 경우보다 더 성능이 떨어졌다.

별첨자료3

원핫 인코딩 모델 검증

Store, Dept의 경우 각 매점, 부서별 번호이기 때문에, 범주형 데이터이다.
범주형의 경우 원핫 인코딩을 하는 것이 일반적이다.

프로젝트에서는 각 데이터 범위가 1~45, 1~99 이기 때문에, 컬럼 수가 비대해지는 것을 막기 위해
별도의 인코딩 없이 그대로 진행하였지만, 참고를 위해 **one-hot 인코딩으로 모델을 확인한 A/B Test** 의 성능을 기록하였다.

| 사용 컬럼

Xgb
Store, Dept, Year, Week, IsHoliday_le, Size_sd, Temperature_sd, CPI_sd

Xgb_oh
Store1~45, Dept1~99, Year, Week, IsHoliday_le, Size_sd, Temperature_sd, CPI_sd

| 각 XGB 모델별 평가지표

	R^2	RMSE	WMAE
Xgb	**0.9500**	**5061.3442**	**3000.1364**
Xgb_oh	0.9356 ▼	5746.3508 ▲	3402.8074 ▲

Xgb_oh | 원핫 인코딩을 진행한 컬럼들을 학습에 사용한 모델의 성능이 기존 모델보다 전반적으로 하락하였다.

별첨자료4

선형 회귀가 적합하지 않은 이유

1. Ramsay RESET Test

$Model: y = \beta_0 + \beta_1 x_1$

$RESET\ Model: y = \beta_0 + \beta_1 x_1 + \delta_1(\hat{\beta}_0 + \hat{\beta}_1 x_1)^2$ (다차항 삽입)

다차항의 회귀계수 δ_1에 대한 유의성 판단

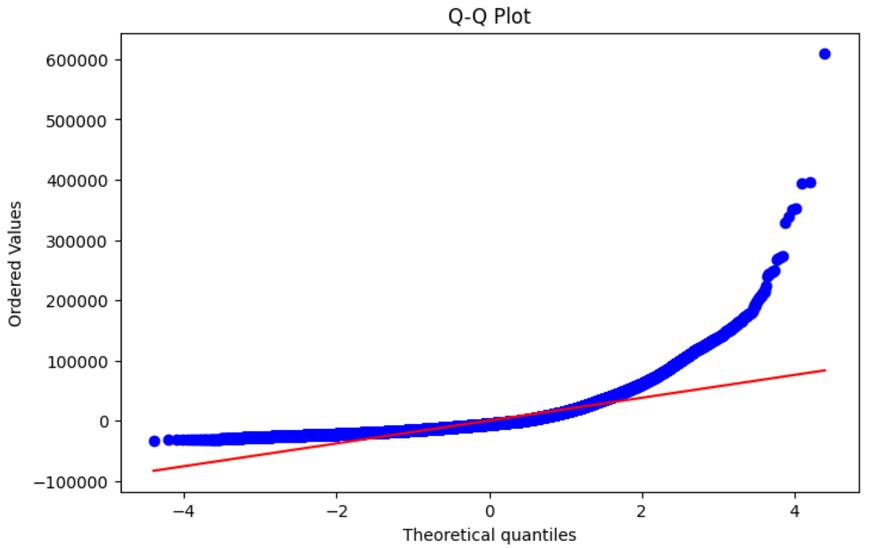
귀무가설 | 독립변수들과 종속변수는 비선형 관계를 띈다. (회귀계수 δ_1가 유의)

대립가설 | 독립변수들과 종속변수는 선형 관계를 띈다. (회귀계수 δ_1가 유의하지 않음)

F 통계량	P-Value
6516.895	< 0.05

귀무가설 채택, 독립변수들과 종속변수는 비선형관계

2. Q-Q Plot



대부분의 잔차가 직선 밖에 위치, **오차가 정규성을 만족하지 못함**

해당 데이터는 선형 모델로 설명 하기 힘듦