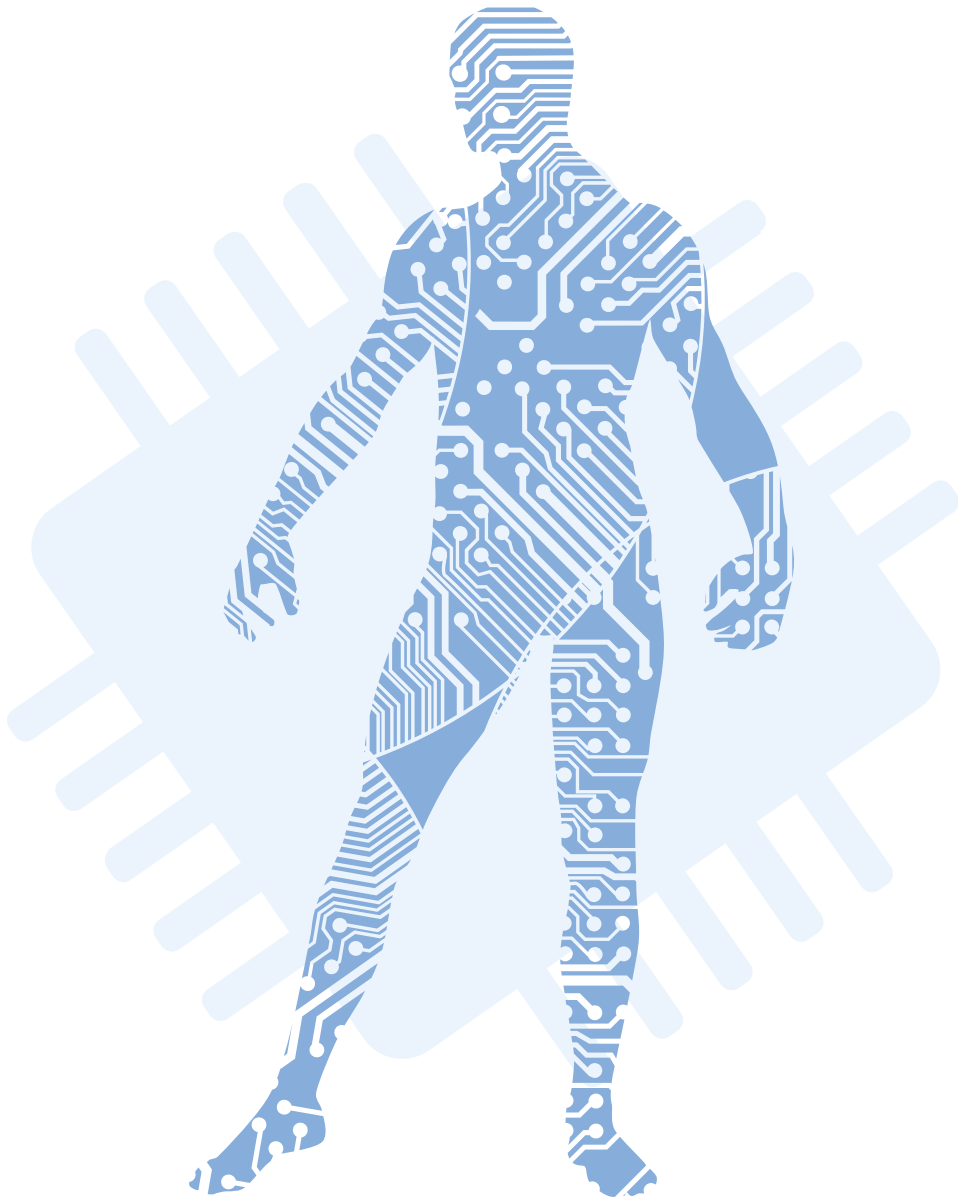
드럭스토어 수익예측

Team 약팔이

안우진 박태건 권도윤 정희주 한승희

Contents



데이터 소개

train.csv

obs: 949194

features

Store
DayofWeek
Date
Sales
Customer
Open
Promo
StateHoliday
SchoolHoliday

store.csv

obs: 1115

features

Store
StoreType
Assortment
CompetitionDistance
CompetitionOpenSinceMonth
CompetitionOpenSinceYear
Promo2
Promo2SinceWeek
Promo2SinceYear
PromoInterval

test.csv

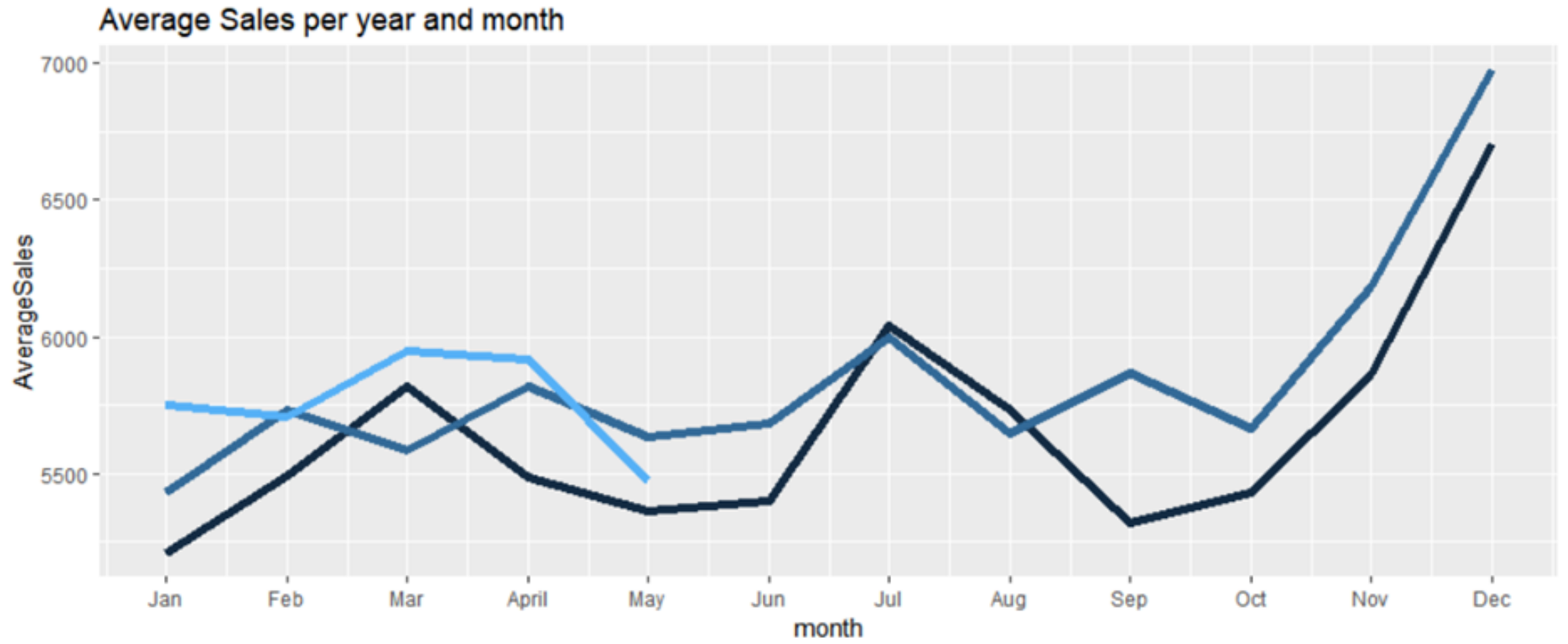
obs: 46830

features

ID
Store
DayofWeek
Date
Open
Promo
StateHoliday
SchoolHoliday

시각화

연월별 수익분포


Average Sales per year and month
7000
6500
6000
5500
AverageSales
Jan
Feb
Mar
April
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
month

시각화

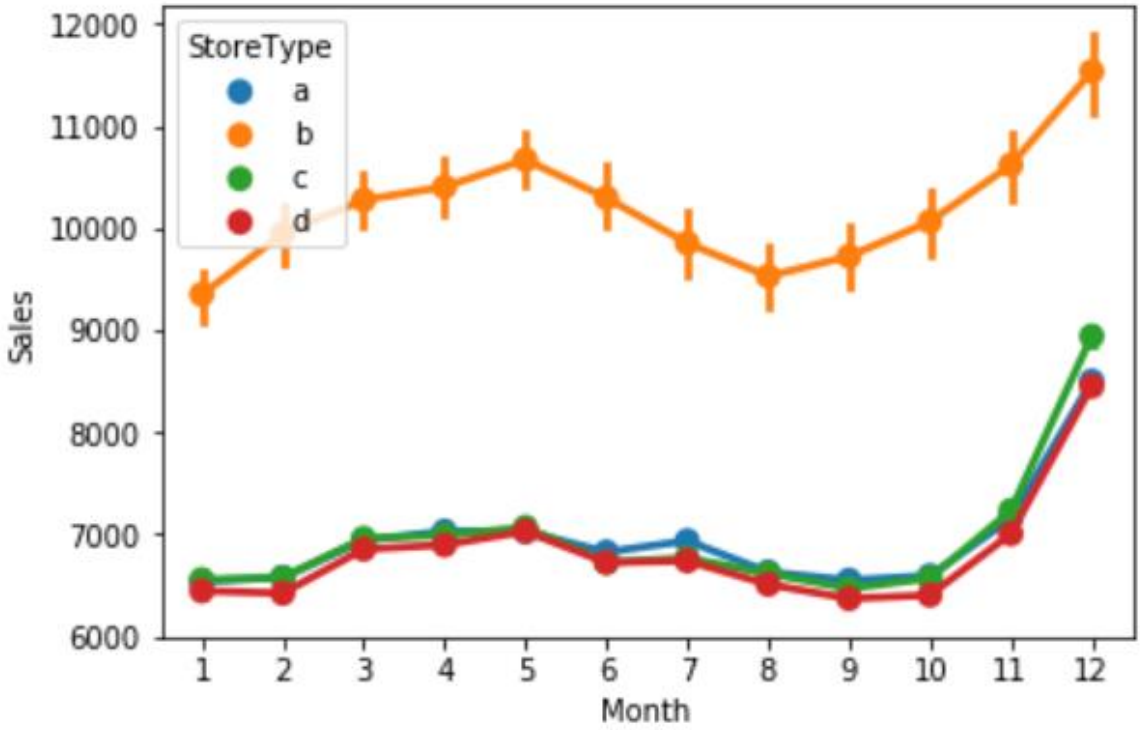
연월별 수익분포

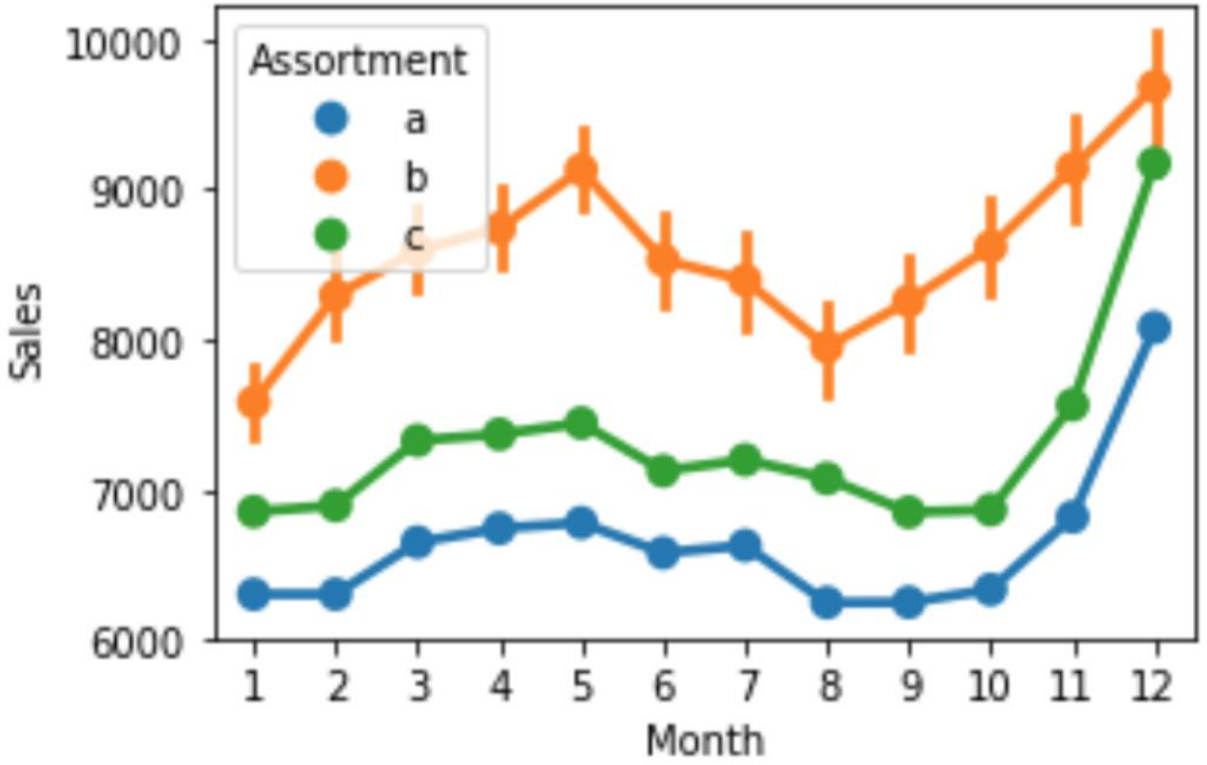
Insight 1.
10 ~ 12월 수익이 상대적으로 높음

Insight 2.
3년간 평균 수익분포 대체적으로 비슷함

시각화

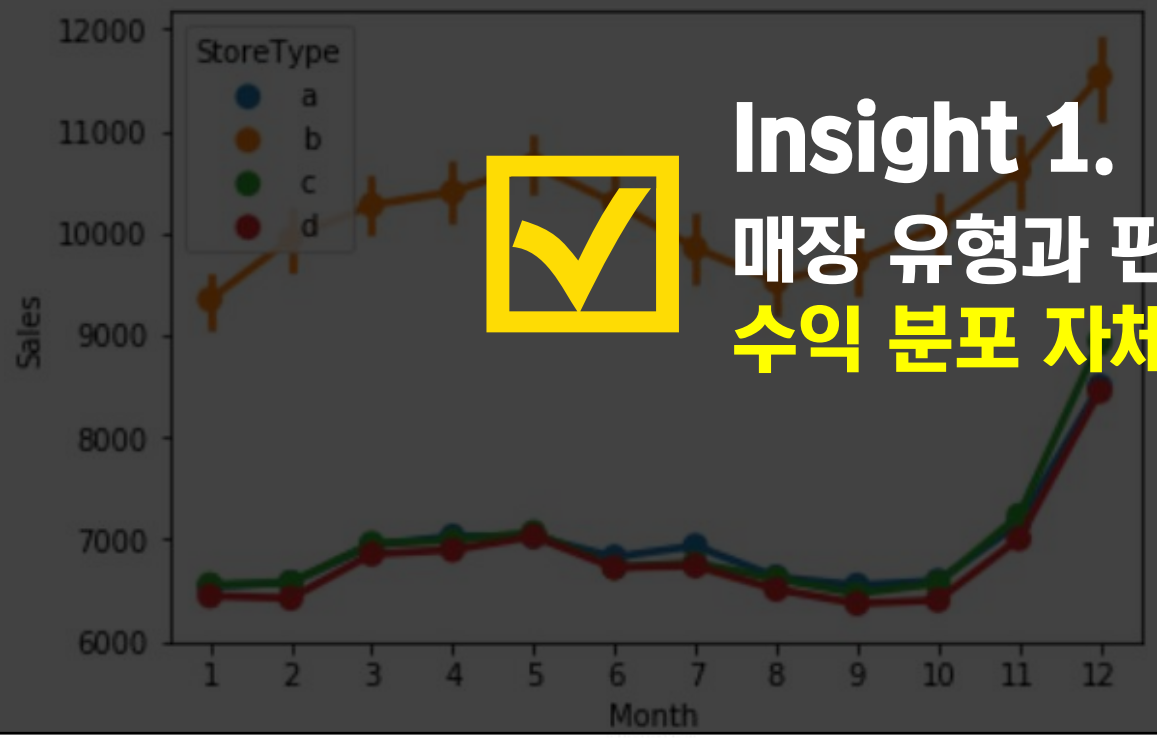
매장 유형 및 판매상품 별 수익분포

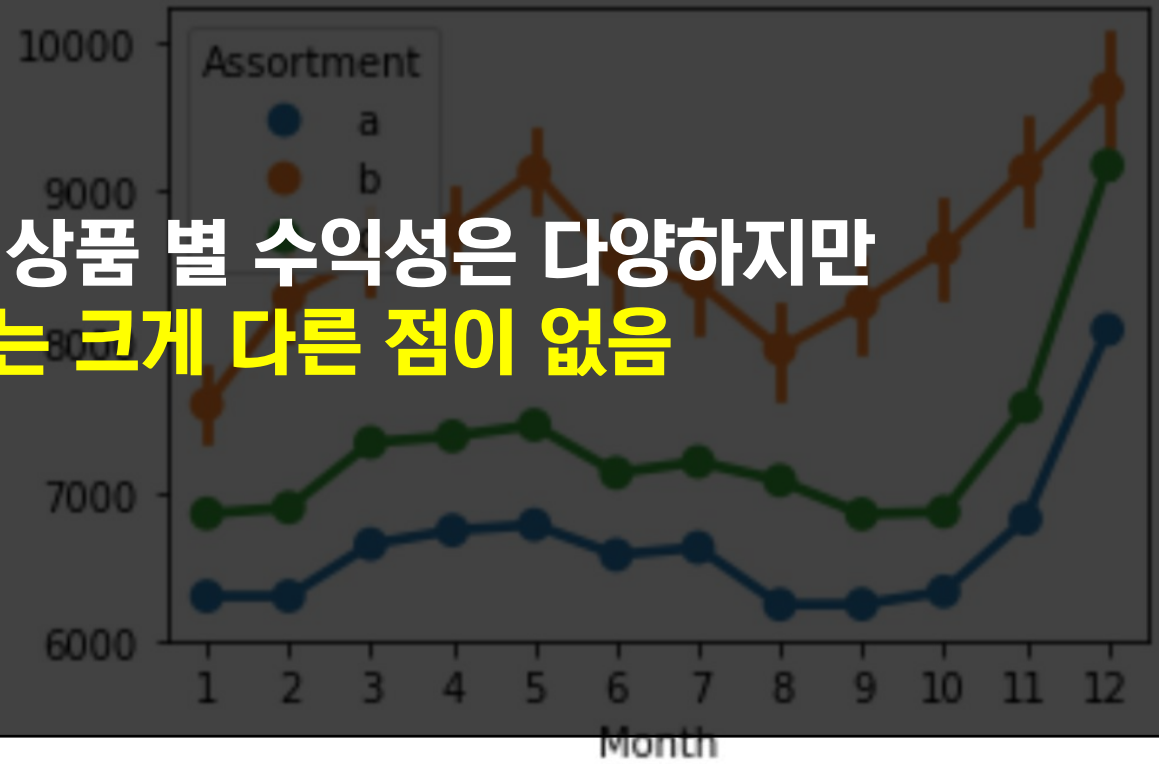




시각화

매장 유형 및 판매상품 별 수익분포





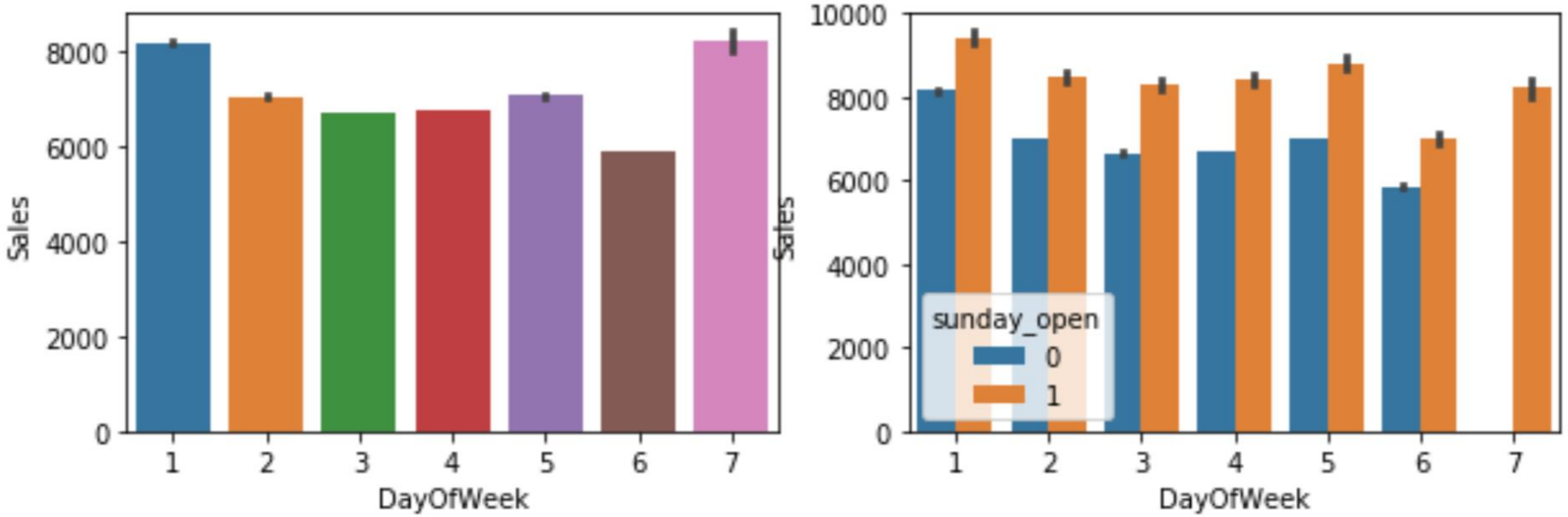
Insight 1.

매장 유형과 판매 상품 별 수익성은 다양하지만

수익 분포 자체에는 크게 다른 점이 없음

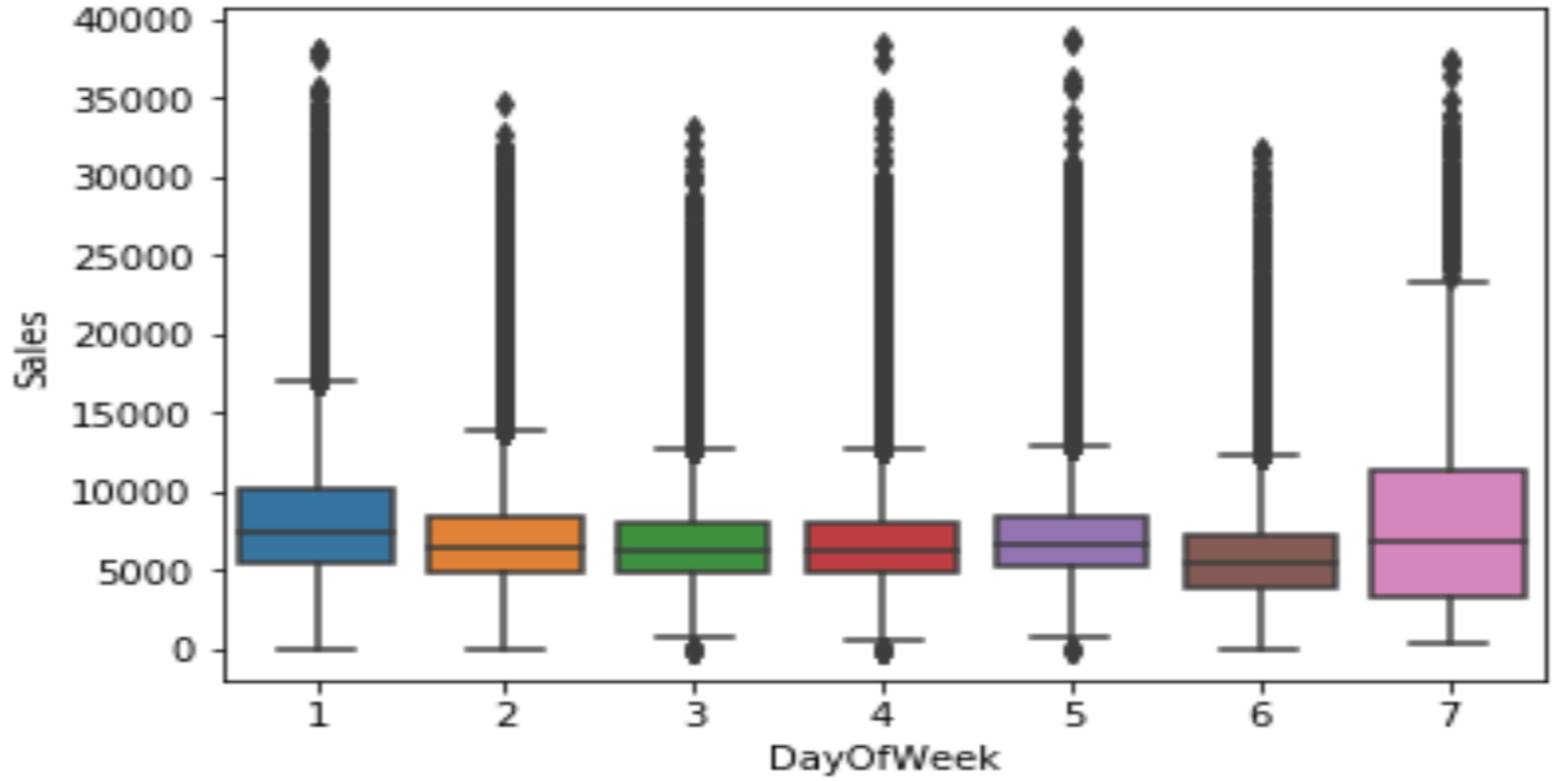
시각화

공휴일 이외의 정기 휴일에 따른 수익분포



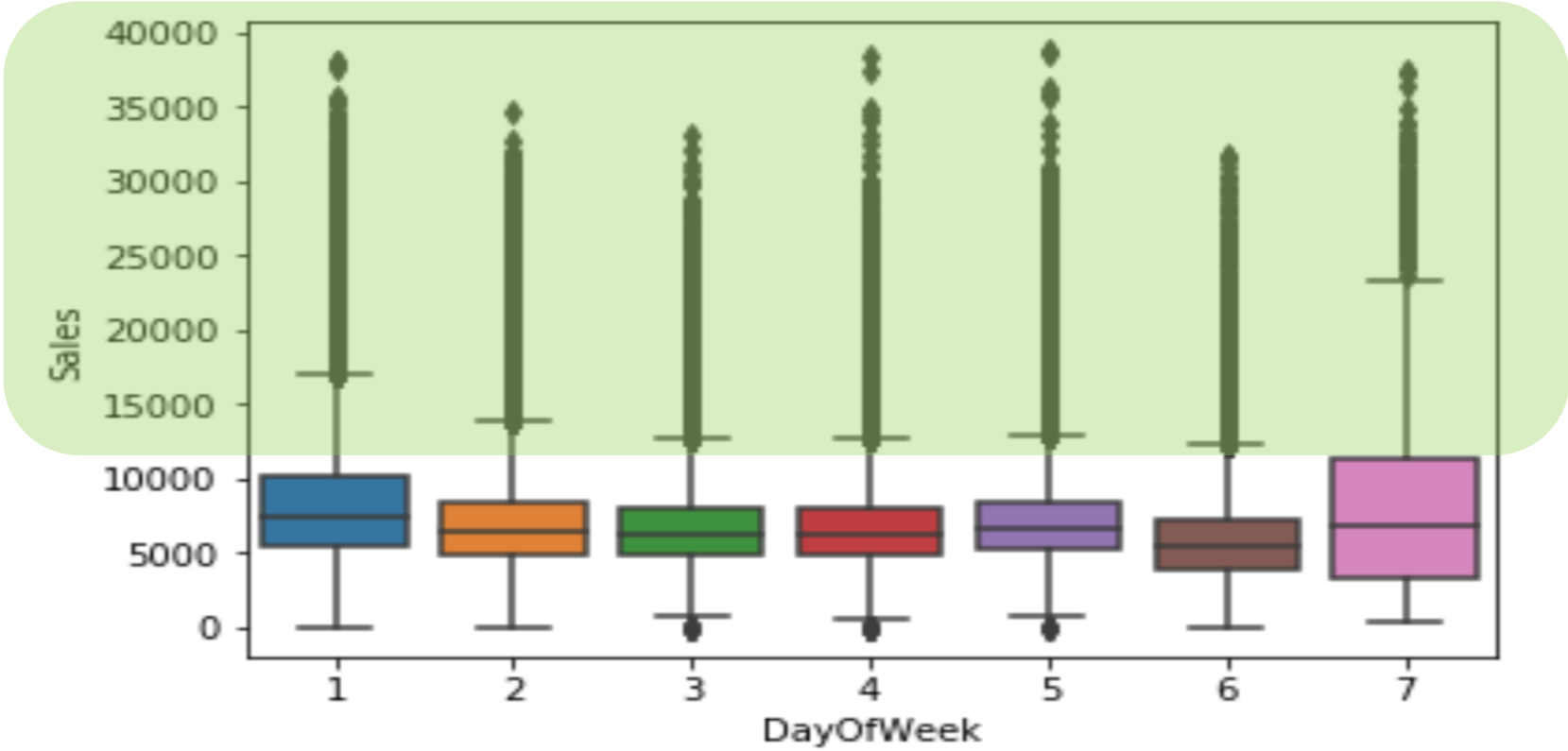
시각화

요일에 따른 수익분포



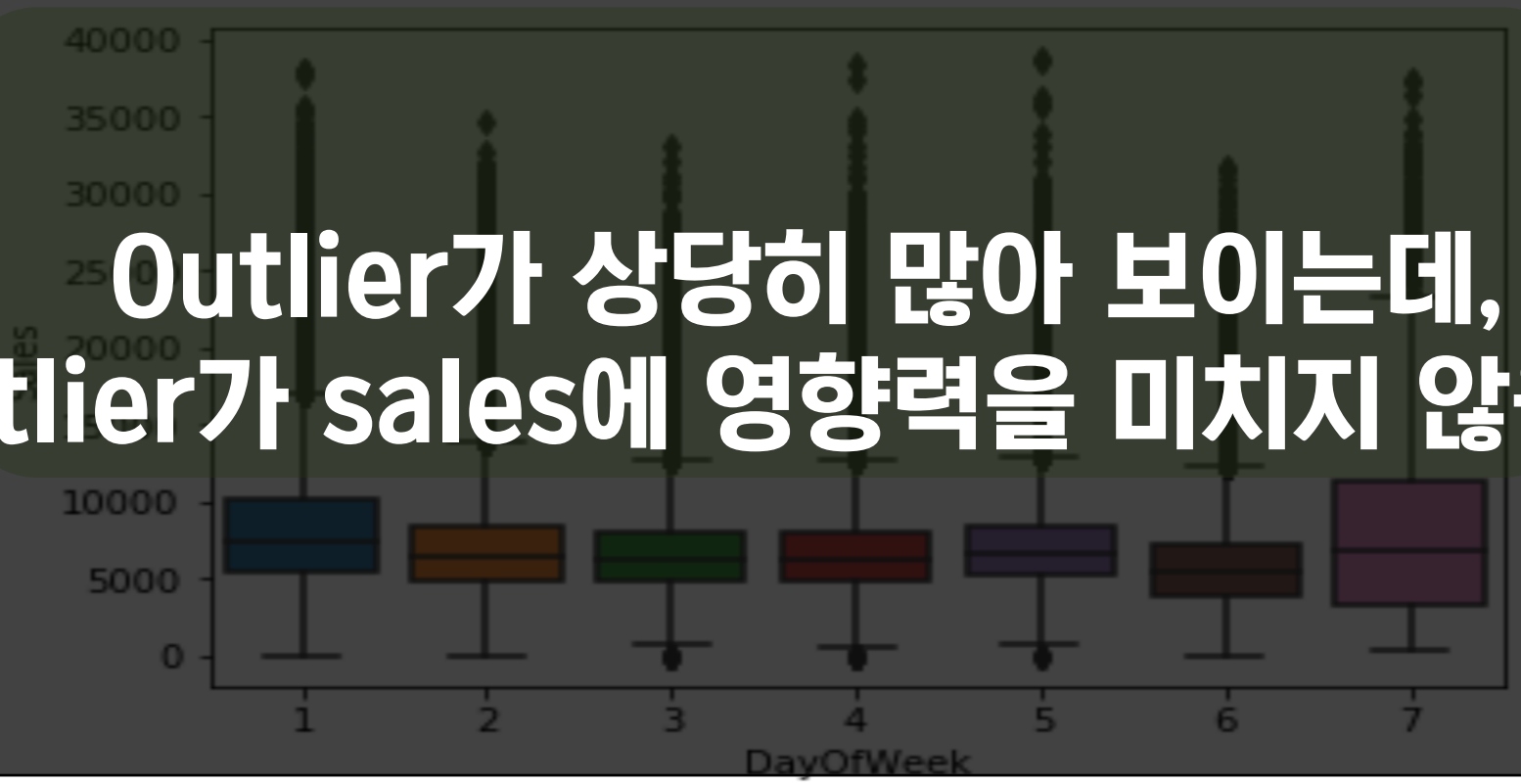
시각화

요일에 따른 수익분포



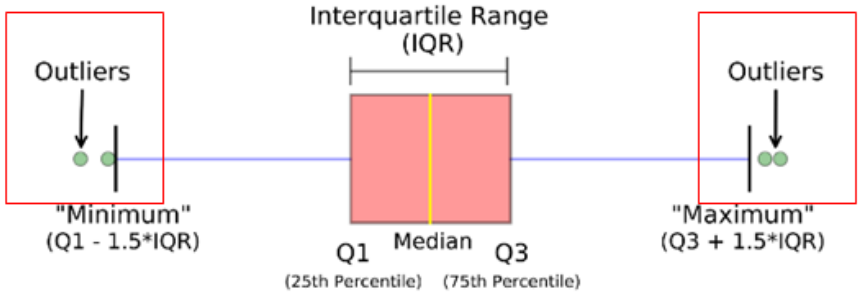
시각화

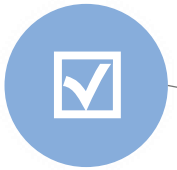
요일에 따른 수익분포

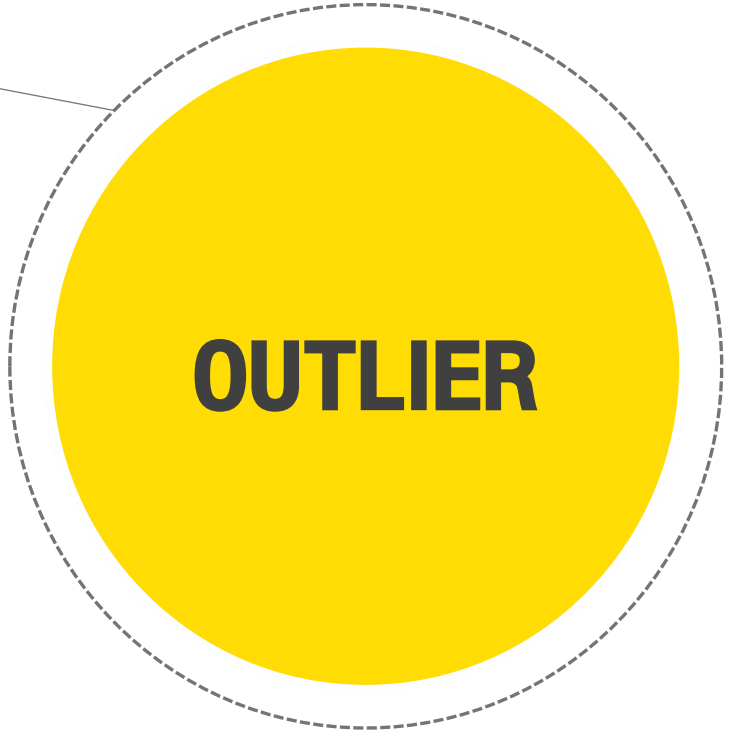


Outlier가 상당히 많아 보이는데,
Outlier가 sales에 영향력을 미치지 않을까?

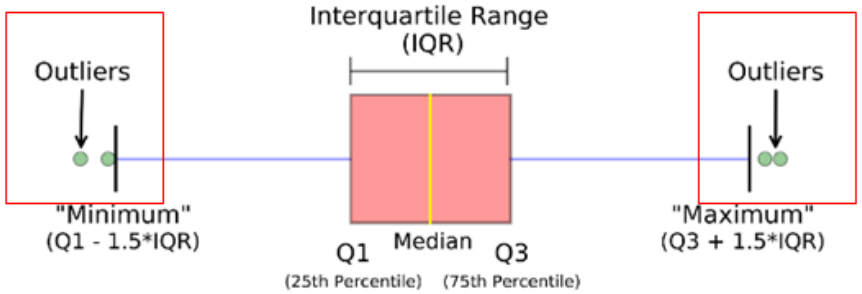
시각화


Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
Q1
Median
Q3
(25th Percentile)
(75th Percentile)
"Maximum"
(Q3 + 1.5*IQR)




OUTLIER

시각화



고려해야 하는 이유

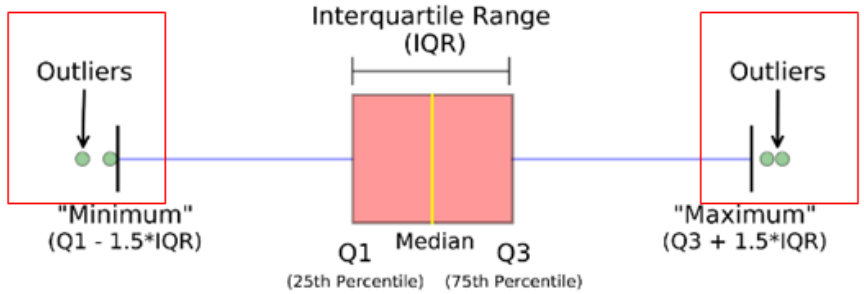
무시할 수 없을 정도로 큰 비중 차지

Outlier의 개수: 16167개

(전체 obs의 2.06%)

OUTLIER

시각화



고려해야 하는 이유

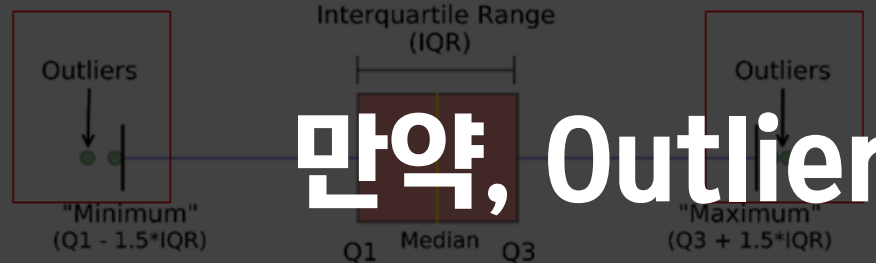
무시할 수 없을 정도로 큰 비중 차지
Outlier의 개수: 16167개
(전체 obs의 2.06%)

OUTLIER분석으로 인한 기대 효과

Outlier에 영향을 주는 요인들을 찾고 그 영향력을 분석하여 모델의 정확성을 높일 수 있지 않을까?

OUTLIER

시각화



만약, Outlier의 영향력이 있다면,

1. Mean으로 그린 Sales 그래프와
2. Median으로 그린 Sales 그래프는

차이를 보일 것이다

고려해야 하는 이유

무시할 수

Outlier의 개수: 16167개

(전체 ob

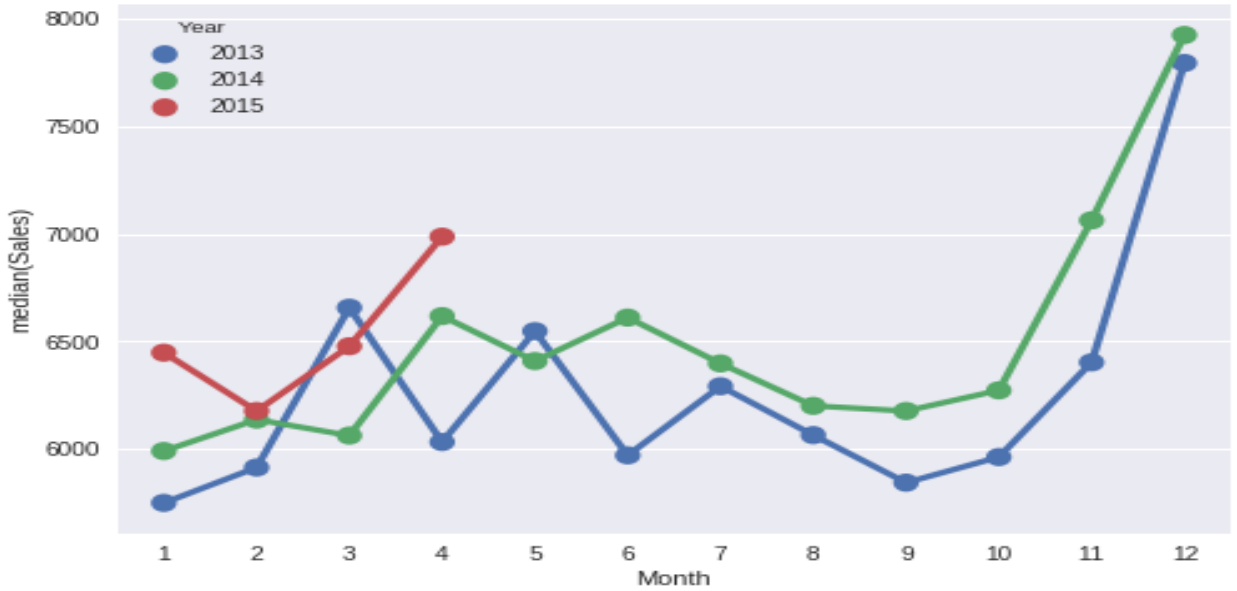
OUTLIER

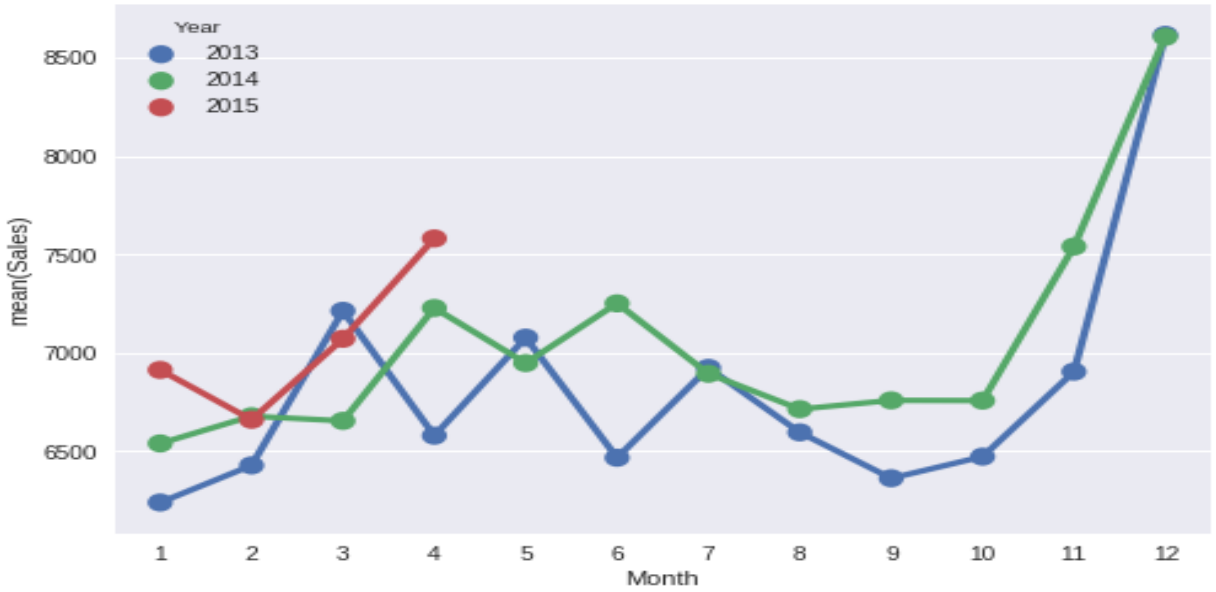
OUTLIER분석으로 인한 기대 효과

Outlier에 영향을 주는 요인들을 찾고 그 영향력을 분석하여 모델의 정확성을 높일 수 있지 않을까?

시각화

전체 매장의 연월 sales mean 분포 VS 전체 매장의 연월 sales median 분포

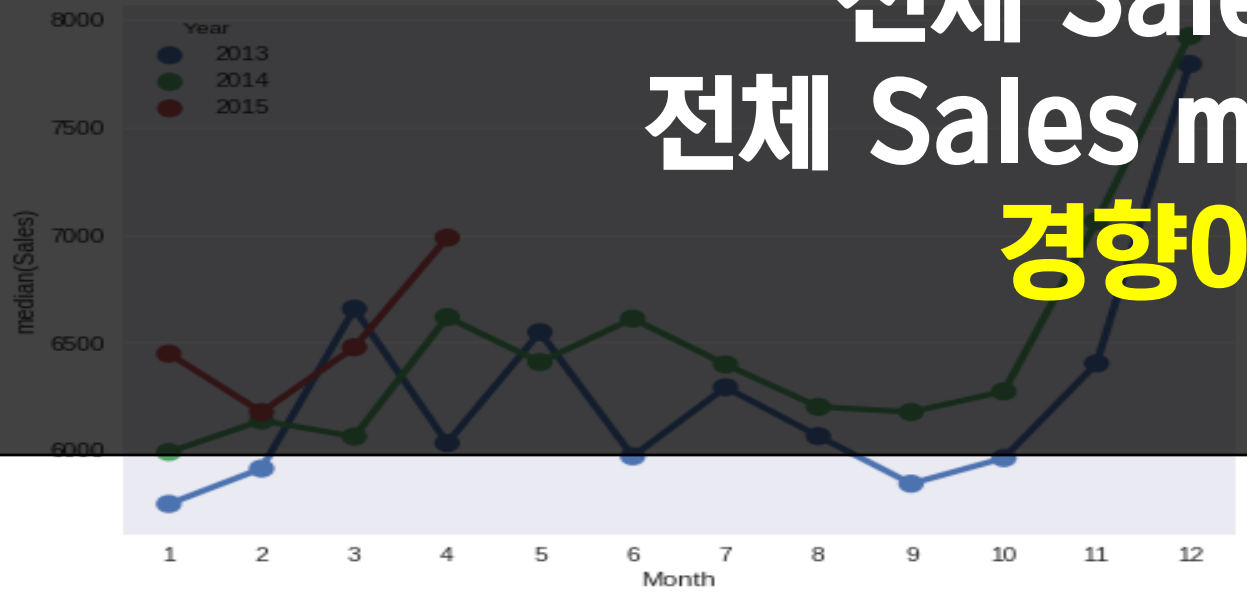




시각화

전체 매장의 연월 sales mean 분포 VS 전체 매장의 연월 sales median 분포

전체 Sales mean과
전체 Sales median 분포의
경향이 유사



전체 sales mean 과 전체 sales median 분포의 경향성이 유사하다

변수선정

train.csv

obs: 949194

features

Store
DayofWeek
Date
Sales
Customer
Open
Promo
StateHoliday
SchoolHoliday

store.csv

obs: 1115

features

Store
StoreType
Assortment
CompetitionDistance
CompetitionOpenSinceMonth
CompetitionOpenSinceYear
Promo2
Promo2SinceWeek
Promo2SinceYear
PromoInterval

test.csv

obs: 46830

features

ID
Store
DayofWeek
Date
Open
Promo
StateHoliday
SchoolHoliday

변수선정

train.csv

obs: 949194

features

Store
DayofWeek
Date
Sales
Customer
Open
Promo
StateHoliday
SchoolHoliday

store.csv

obs: 1115

features

Store
StoreType
Assortment
CompetitionDistance
CompetitionOpenSinceMonth
CompetitionOpenSinceYear
Promo2
Promo2SinceWeek
Promo2SinceYear
PromoInterval

test.csv

obs: 46830

features

ID
Store
DayofWeek
Date
Open
Promo
StateHoliday
SchoolHoliday

변수선정

TRAIN SET

train.csv

obs: 949194

features

Store
DayofWeek
Date
Sales
Customer
Open
Promo
StateHoliday
SchoolHoliday

store.csv

obs: 1115

features

Store
StoreType
Assortment
CompetitionDistance
CompetitionOpenSinceMonth
CompetitionOpenSinceYear
Promo2
Promo2SinceWeek
Promo2SinceYear
PromoInterval

변수선정

TEST SET

store.csv

obs: 1115

features

Store
StoreType
Assortment
CompetitionDistance
CompetitionOpenSinceMonth
CompetitionOpenSinceYear
Promo2
Promo2SinceWeek
Promo2SinceYear
PromoInterval

test.csv

obs: 46830

features

ID
Store
DayofWeek
Date
Open
Promo
StateHoliday
SchoolHoliday

변수선정

store.csv

obs: 1115

features

Store
StoreType
Assortment
CompetitionDistance
CompetitionOpenSinceMonth
CompetitionOpenSinceYear
Promo2
Promo2SinceWeek
Promo2SinceYear
PromoInterval

변수선정

store.csv

obs: 1115

☑ 변수 별 영향도에 따른 Binary 변수 고려

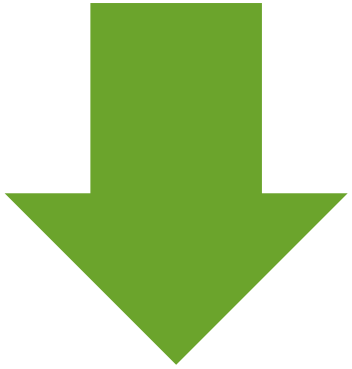
☑ Lag를 이용한 변수 고려

Store
StoreType
Assortment
CompetitionDistance
CompetitionOpenSinceMonth
CompetitionOpenSinceYear
Promo2
Promo2SinceWeek
Promo2SinceYear
PromoInterval

변수선정

☑ 변수 별 영향도에 따른 Binary 변수 고려

CompetitionOpenSinceMonth
CompetitionOpenSinceYear



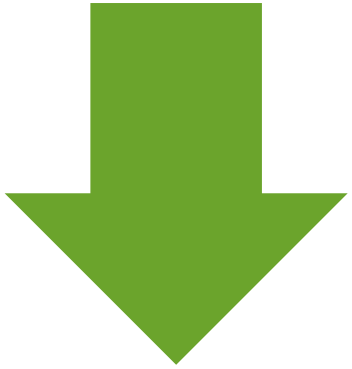
CompetitionOpen

CompetitionOpenSinceMonth	CompetitionOpenSinceYear
9	2008
11	2007
12	2006
9	2009
4	2015
12	2013
4	2013
10	2014
8	2000
9	2009
11	2011
NA	NA

변수선정

☑ 변수 별 영향도에 따른 Binary 변수 고려

CompetitionOpenSinceMonth
CompetitionOpenSinceYear



CompetitionOpen

Date
2013-01-01
2013-01-02
2013-01-03
2013-01-04
2013-01-05
2013-01-06
2013-01-07
2013-01-08

CompetitionOpenSinceMonth	CompetitionOpenSinceYear
9	2008
11	2007
12	2006
9	2009
4	2015
12	2013
4	2013
10	2014

변수선정

변수 별 영향도에 따른 Binary 변수 고려

CompetitionOpenSinceMonth
CompetitionOpenSinceYear

경쟁업체의 **개업정보**를 담은 변수 설정
개업 이전 = 0 / 개업 이후 = 1

Date	CompetitionOpenSinceMonth	CompetitionOpenSinceYear
2013-01-02	11	2007
2013-01-03	12	2006
2013-01-04	9	2009
2013-01-05	4	2015
2013-01-06	12	2013
2013-01-07	4	2013
2013-01-08	10	2014

CompetitionOpen

변수선정

☑ 변수 별 영향도에 따른 Binary 변수 고려

Promo2
Promo2SinceWeek
Promo2SinceYear
PromoInterval

Promo2Score

Promo2	Promo2SinceWeek	Promo2SinceYear	PromoInterval
0	NA	NA	
1	13	2010	Jan,Apr,Jul,Oct
1	14	2011	Jan,Apr,Jul,Oct
0	NA	NA	
0	NA	NA	
0	NA	NA	
0	NA	NA	
0	NA	NA	
0	NA	NA	
0	NA	NA	
1	1	2012	Jan,Apr,Jul,Oct
1	13	2010	Jan,Apr,Jul,Oct
1	45	2009	Feb,May,Aug,Nov
1	40	2011	Jan,Apr,Jul,Oct
1	14	2011	Jan,Apr,Jul,Oct
0	NA	NA	

변수선정

변수 별 영향도에 따른 Binary 변수 고려

Promo2
Promo2SinceWeek
Promo2SinceYear
PromoInterval

Promo2Score

Promo2	Promo2SinceWeek	Promo2SinceYear	PromoInterval
0	NA	NA	
1	13	2010	Jan,Apr,Jul,Oct
1	14	2011	Jan,Apr,Jul,Oct
0	NA	NA	

Promo2	Promo2SinceWeek	Promo2SinceYear	PromoInterval	Date
0	NA	NA		NA
1	13	2010	Jan,Apr,Jul,Oct	2010-04-01
1	14	2011	Jan,Apr,Jul,Oct	2011-04-01
0	NA	NA		NA

변수선정

변수 별 영향도에 따른 Binary 변수 고려

Promo2
Promo2SinceWeek
Promo2SinceYear
PromoInterval

Promo2Score

Promo2	Promo2SinceWeek	Promo2SinceYear	PromoInterval
0	NA	NA	
1	13	2010	Jan,Apr,Jul,Oct
1	14	2011	Jan,Apr,Jul,Oct
0	NA	NA	

Promo2	Promo2SinceWeek	Promo2SinceYear	PromoInterval	Date
0	NA	NA		NA
1	13	2010	Jan,Apr,Jul,Oct	2010-04-01
1	14	2011	Jan,Apr,Jul,Oct	2011-04-01
0	NA	NA		NA

변수선정

변수 별 영향도에 따른 Binary 변수 고려

Promo2
Promo2SinceWeek
Promo2SinceYear
PromoInterval

정기 프로모션을 진행여부를 담은 변수 설정
해당하지 않는 달 = 0
해당하는 달 = 1

Promo2	Promo2SinceWeek	Promo2SinceYear	PromoInterval
0	NA	NA	
1	14	2011	Jan,Apr,Jul,Oct
0	NA	NA	

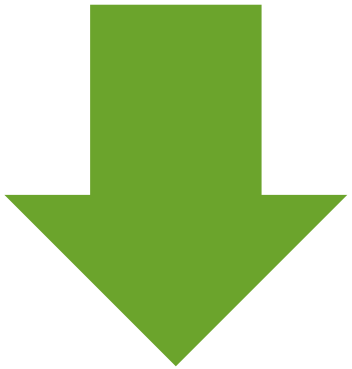
Promo2	Promo2SinceWeek	Promo2SinceYear	PromoInterval	Date
0	NA	NA		NA
1	13	2010	Jan,Apr,Jul,Oct	2010-04-01
1	14	2011	Jan,Apr,Jul,Oct	2011-04-01
0	NA	NA		NA

Promo2Score

변수선정

☑ Lag를 이용한 변수 고려

CompetitionOpenSinceMonth
CompetitionOpenSinceYear



CompetitionSince

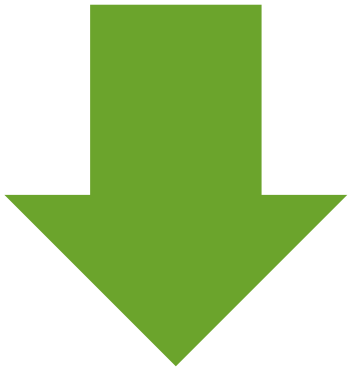
Date	CompetitionOpenSinceMonth	CompetitionOpenSinceYear
2013-01-01	9	2008
2013-01-02	11	2007
2013-01-03	12	2006
2013-01-04	9	2009
2013-01-05	4	2015
2013-01-06	12	2013
2013-01-07	4	2013
2013-01-08	10	2014

Promo2Since = (Year - Promo2SinceYear) * 12 +
(WeekOfYear - Promo2SinceWeek) / 4

변수선정

Lag를 이용한 변수 고려

CompetitionOpenSinceMonth
CompetitionOpenSinceYear



CompetitionSince

Date	CompetitionOpenSinceMonth	CompetitionOpenSinceYear	CompetitionSince
2013-01-01	9	2008	43.0
2013-01-02	11	2007	46.0
2013-01-03	12	2006	
2013-01-04	9	2009	0.0
2013-01-05	4	2015	53.0
2013-01-06	12	2013	
2013-01-07	4	2013	0.0
2013-01-08	10	2014	

변수선정

☑ Lag를 이용한 변수 고려

CompetitionOpenSinceMonth
CompetitionOpenSinceYear

☑ 경쟁 지속 기간

Date	CompetitionOpenSinceMonth	CompetitionOpenSinceYear	CompetitionSince
2013-01-01	9	2008	43.0
2013-01-02	11	2007	46.0
2013-01-03	12	2006	
2013-01-04	9	2009	0.0
2013-01-05	4	2015	
2013-01-06	12	2013	53.0
2013-01-07	4	2013	
2013-01-08	10	2014	0.0

CompetitionSince

변수선정

Lag를 이용한 변수 고려

Promo2
Promo2SinceWeek
Promo2SinceYear

Promo2Since

Promo2SinceWeek	Promo2SinceYear	PromoInterva	Year	Month	Day	WeekOfYear
1.0	2012.0	Jan,Apr,Jul,Oc	2015	4	19	16
0.0	0.0		2015	4	19	16
0.0	0.0		2015	4	19	16
0.0	0.0		2015	4	19	16
0.0	0.0		2015	4	19	16

CompetitionSince = (Year - CompetitionOpenSinceYear) * 12 + (Month - CompetitionOpenSinceMonth)

변수선정

Lag를 이용한 변수 고려

Promo2
Promo2SinceWeek
Promo2SinceYear

Promo2Since

Promo2SinceWeek	Promo2SinceYear	PromoInterva	Year	Month	Day	WeekOfYear	Promo2Since
1.0	2012.0	Jan,Apr,Jul,Oc	2015	4	19	16	39.75
0.0	0.0		2015	4	19	16	0.00
0.0	0.0		2015	4	19	16	0.00
0.0	0.0		2015	4	19	16	0.00
0.0	0.0		2015	4	19	16	0.00

변수선정

Lag를 이용한 변수 고려

Promo2
Promo2SinceWeek
Promo2SinceYear

프로모션 지속 기간

Promo2SinceWeek	Promo2SinceYear	PromoInterva	Year	Month	Day	WeekOfYear	Promo2Since
1.0	2012.0	Jan,Apr,Jul,Oc	2015	4	19	16	39.75
0.0	0.0		2015	4	19	16	0.00
0.0	0.0		2015	4	19	16	0.00
0.0	0.0		2015	4	19	16	0.00
0.0	0.0		2015	4	19	16	0.00

Promo2Since

변수선정

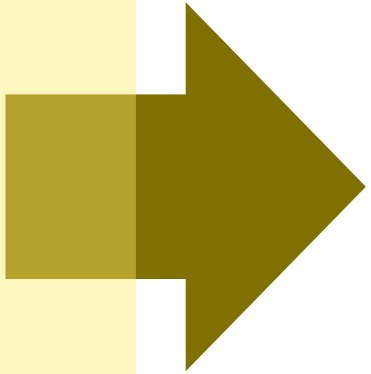
TRAIN SET

drugstore.df

obs: 949194

14 features + 1 target

Open	Promo
Store	Promo2Score
StoreType	Promo2Since
Assortment	DayOfWeek
CompetitionOpen	Date
ConpetitionSince	StateHoliday
CompetitionDistance	SchoolHoliday



Sales

데이터 통합

Store	DayOfWeek	Date	Sales	Open	Promo	StateHoliday	SchoolHoliday	StoreType	Assortment	CompetitionDistance	Promo2Score	CompetitionOpen
391	7	2015-03-29	0	0	0	0	0	1	1	460	0	1
391	1	2015-03-30	10276	1	1	0	1	1	1	460	0	1
391	2	2015-03-31	8418	1	1	0	1	1	1	460	0	1
391	3	2015-04-01	7758	1	1	0	1	1	1	460	0	1
391	4	2015-04-02	8616	1	1	0	1	1	1	460	0	1
391	5	2015-04-03	0	0	1	2	1	1	1	460	0	1
391	6	2015-04-04	3675	1	0	0	0	1	1	460	0	1
391	7	2015-04-05	0	0	0	0	0	1	1	460	0	1
391	1	2015-04-06	0	0	0	2	1	1	1	460	0	1
391	2	2015-04-07	5552	1	0	0	1	1	1	460	0	1
391	3	2015-04-08	4888	1	0	0	1	1	1	460	0	1
391	4	2015-04-09	4830	1	0	0	1	1	1	460	0	1
391	5	2015-04-10	4992	1	0	0	1	1	1	460	0	1
391	6	2015-04-11	2767	1	0	0	0	1	1	460	0	1
391	7	2015-04-12	0	0	0	0	0	1	1	460	0	1
391	1	2015-04-13	9509	1	1	0	0	1	1	460	0	1
391	2	2015-04-14	7228	1	1	0	0	1	1	460	0	1
391	3	2015-04-15	5604	1	1	0	0	1	1	460	0	1
391	4	2015-04-16	6309	1	1	0	0	1	1	460	0	1
391	5	2015-04-17	6567	1	1	0	0	1	1	460	0	1
391	6	2015-04-18	3100	1	0	0	0	1	1	460	0	1
391	7	2015-04-19	0	0	0	0	0	1	1	460	0	1
391	1	2015-04-20	4958	1	0	0	0	1	1			
391	2	2015-04-21	4592	1	0	0	0	1	1			
391	3	2015-04-22	4335	1	0	0	0	1	1			
391	4	2015-04-23	4576	1	0	0	0	1	1			

TRAIN SET

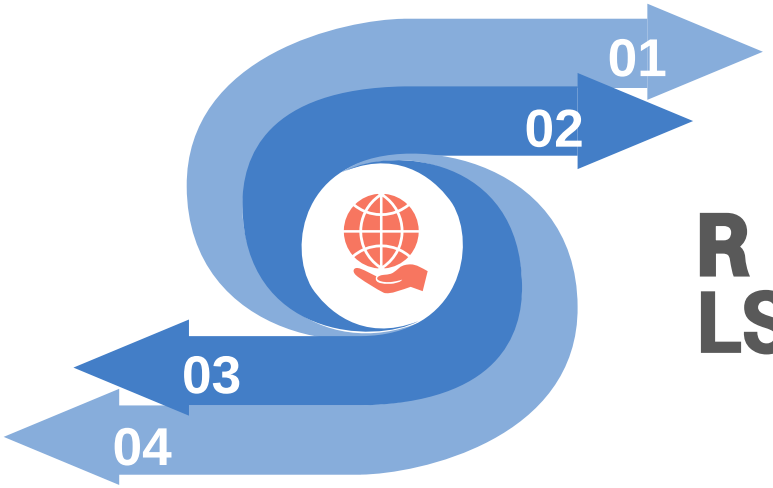
데이터 통합

Store	DayOfWeek	Date	Open	Promo	StateHoliday	SchoolHoliday	StoreType	Assortment	CompetitionDistance	Promo2Score	CompetitionOpen
17	2	2015-07-28	1	1	0	1	1	1	50	1	NA
17	3	2015-07-29	1	1	0	1	1	1	50	1	NA
17	4	2015-07-30	1	1	0	1	1	1	50	1	NA
17	5	2015-07-31	1	1	0	1	1	1	50	1	NA
18	6	2015-06-20	1	0	0	0	4	3	13840	0	NA
18	7	2015-06-21	0	0	0	0	4	3	13840	0	NA
18	1	2015-06-22	1	0	0	0	4	3	13840	0	NA
18	2	2015-06-23	1	0	0	0	4	3	13840	0	NA
18	3	2015-06-24	1	0	0	0	4	3	13840	0	NA
18	4	2015-06-25	1	0	0	0	4	3	13840	0	NA
18	5	2015-06-26	1	0	0	0	4	3	13840	0	NA
18	6	2015-06-27	1	0	0	0	4	3	13840	0	NA
18	7	2015-06-28	0	0	0	0	4	3	13840	0	NA
18	1	2015-06-29	1	1	0	0	4	3	13840	0	NA
18	2	2015-06-30	1	1	0	0	4	3	13840	0	NA
18	3	2015-07-01	1	1	0	0	4	3			
18	4	2015-07-02	1	1	0	0	4	3			

TEST SET

다음 주

Python 팀
XG Boost 사용하여 모델 제작



R 팀
LSTM 사용하여 모델 제작

새로운 변수(?)

상금 겟


감사합니다
모든 팀 다 수고하셨어요!