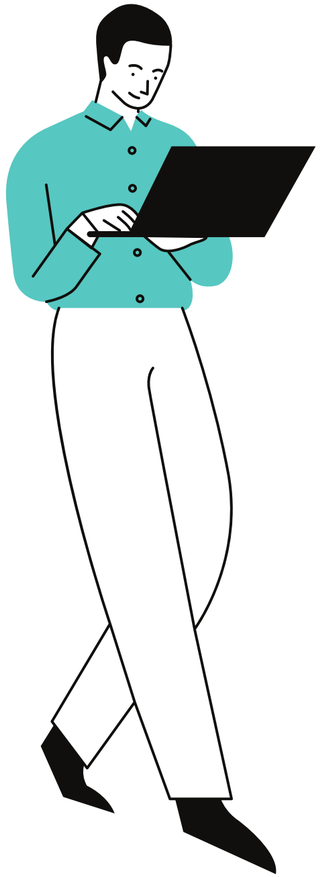
R 을 이용한
데이터분석

CU 일반탄산음료 판매 분석

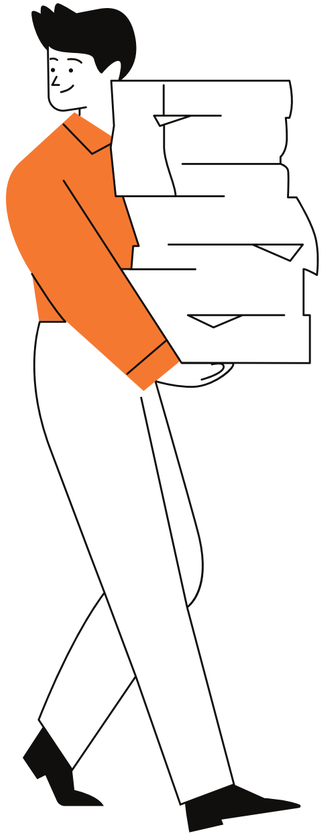
서울 남부 기술 교육원 - 2조

2023년 4월 24일

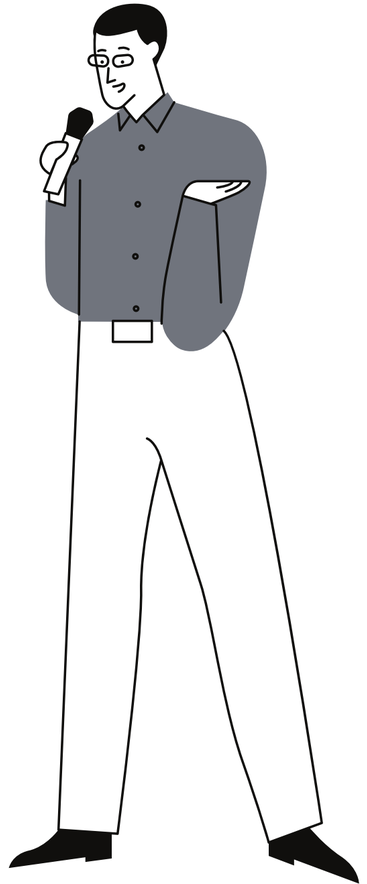
김진성
데이터분석 메인



김성진
스토리텔링-PPT

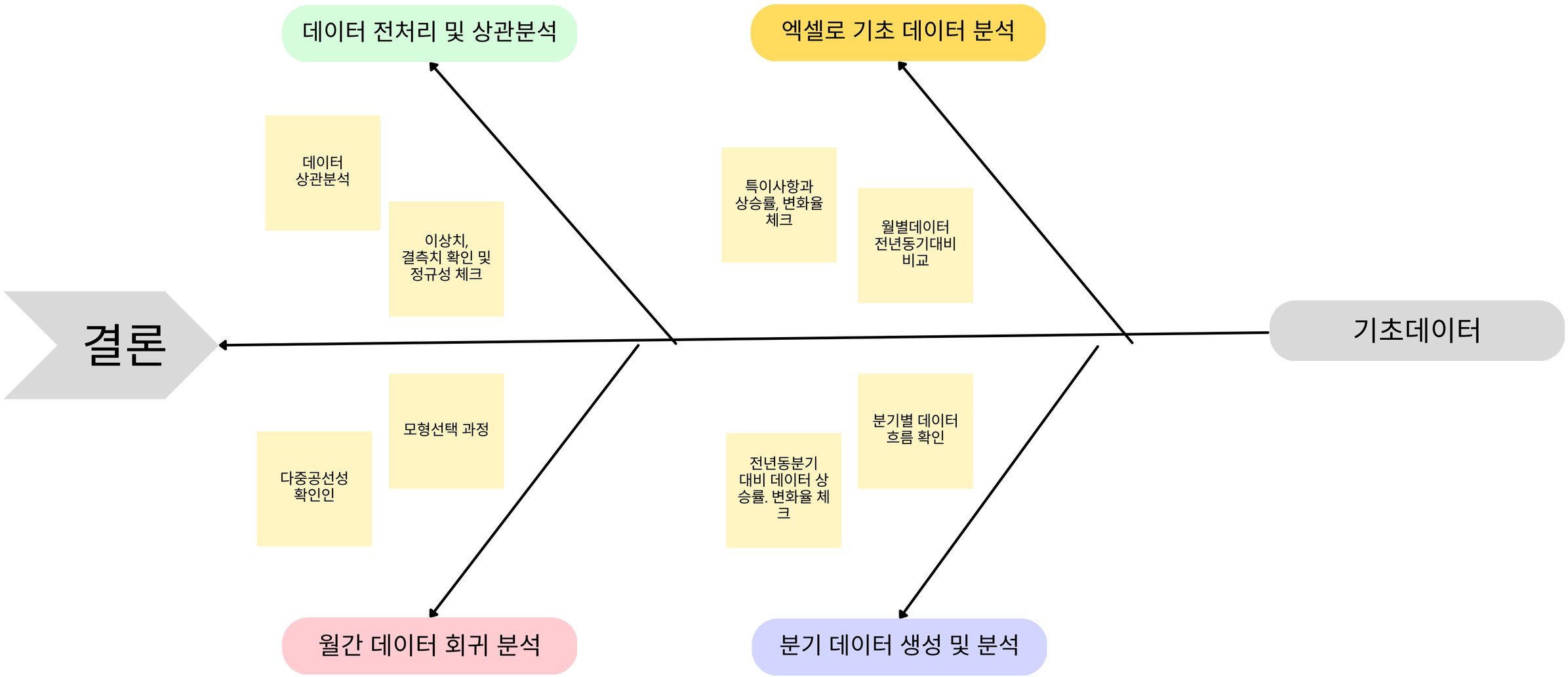


김호성
발표, 데이터 스토리텔링



그럼 시작할게요!

2009년~2013년 CU 일반탄산음료 카테고리 데이터 분석


데이터 전처리 및 상관분석
엑셀로 기초 데이터 분석
데이터
상관분석
이상치,
결측치 확인 및
정규성 체크
특이사항과
상승률, 변화율
체크
월별데이터
전년동기대비
비교
결론
기초데이터
모형선택 과정
다중공선성
확인인
분기별 데이터
흐름 확인
전년동분기
대비 데이터 상
승률. 변화율 체
크
월간 데이터 회귀 분석
분기 데이터 생성 및 분석

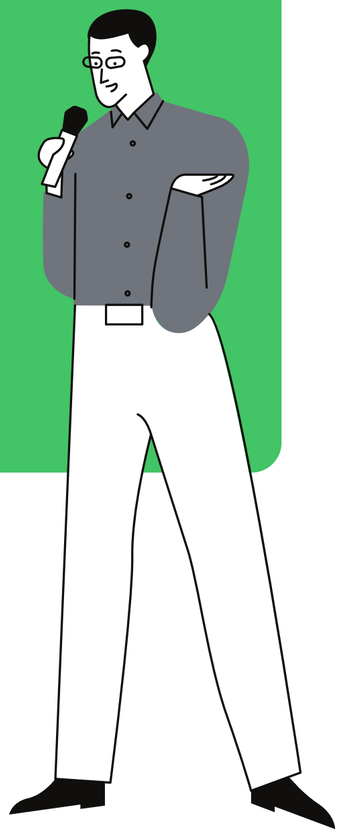
1. 데이터 기초

1. 데이터 기초분석

2009년 1월부터 2013년 12월 까지의 CU 편의점 **일반 탄산음료 판매량 데이터**와
해당 기간내의 **일반 탄산음료 평균 가격, 월간 최대기온, 월간전체판매일(매장수X월오픈일),**
비온날, 휴일 데이터를 기초 데이터로 분석을 진행 하였습니다.

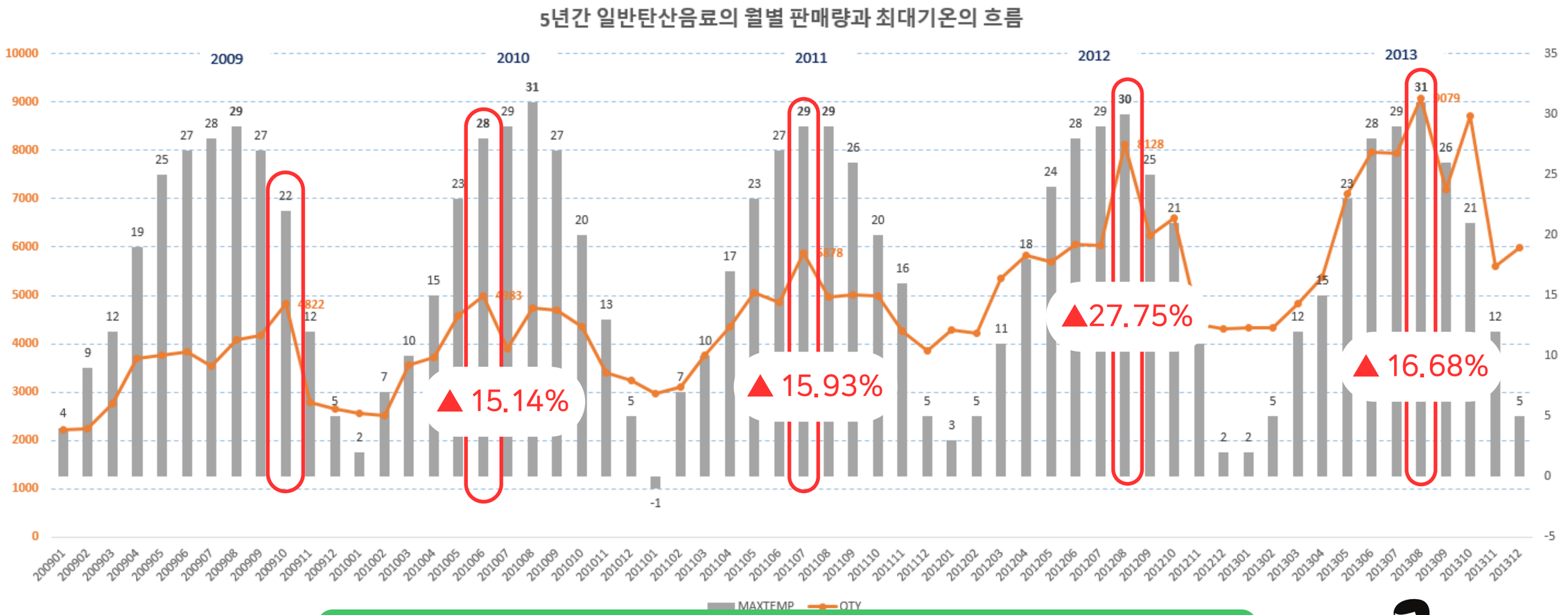
먼저 엑셀 데이터를 기반으로 기초 시계열 단순 분석을 진행하고,
월간 데이터에서 분기로 데이터 확장을 진행 했습니다.

해당 기초 데이터 를 상관분석과 회귀분석을 순서대로
진행 후 최종 분석 결과를 도출 하였습니다.



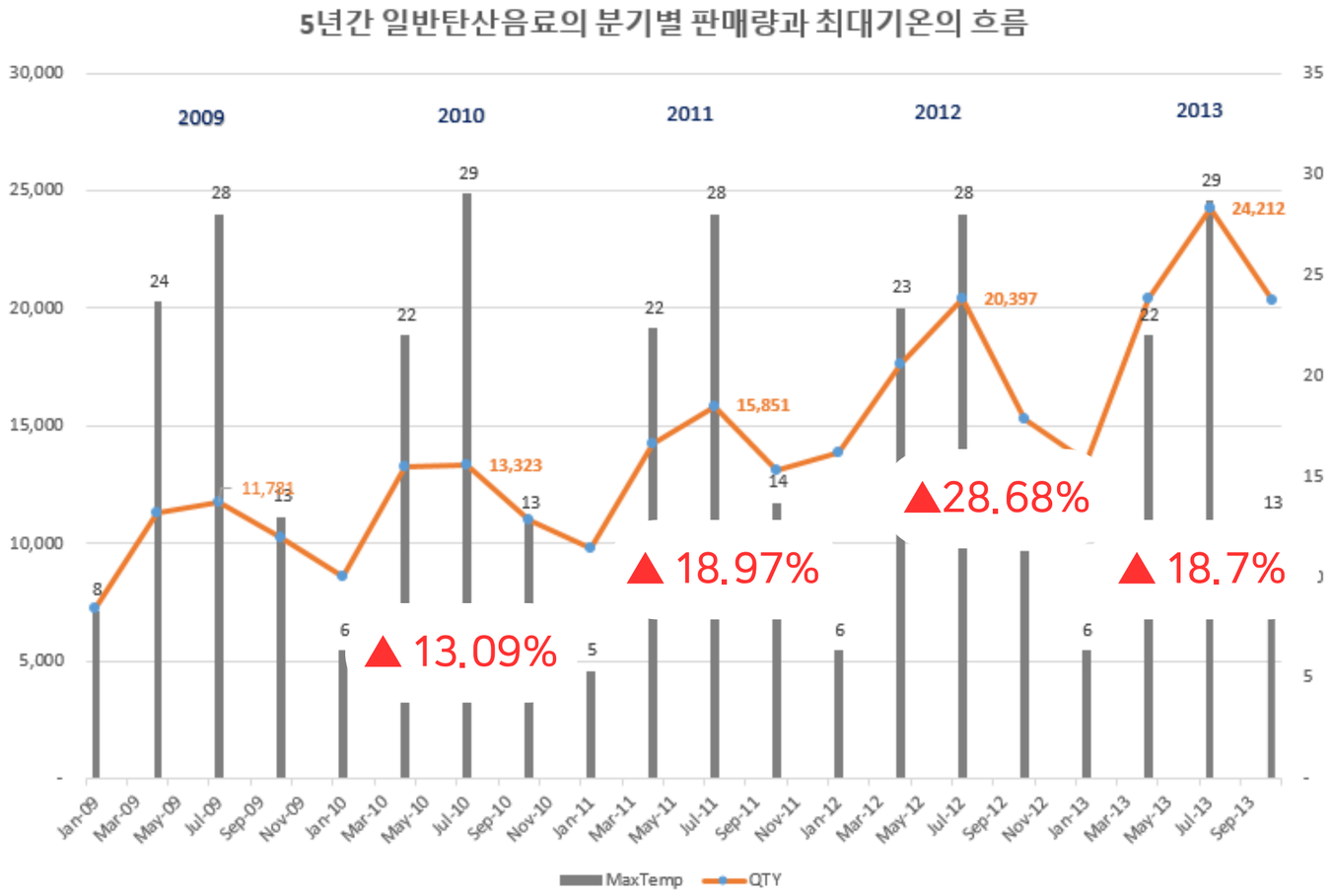
1. 데이터 기초분석

▲ 기간내 매년 평균 18.87% 판매량 상승



월별 판매량과 월 최대기온의 흐름을 그래프화 해본 결과
2009년의 10월 최대기온 22도 일때 최대 판매량이 발생했던것을 제외하고
2010,2011,2012,2013년 4년간 최대기온이 28도 이상이 되었을때
최대 판매량이 발생한것으로 보여지며, 매년 전년대비 최대 판매량을 15% 이상 갱신하는 실적 상승이 있습니다. (2012년은 전년 동월 대비 27.7% 이상 판매량 상승)

1. 데이터 기초분석

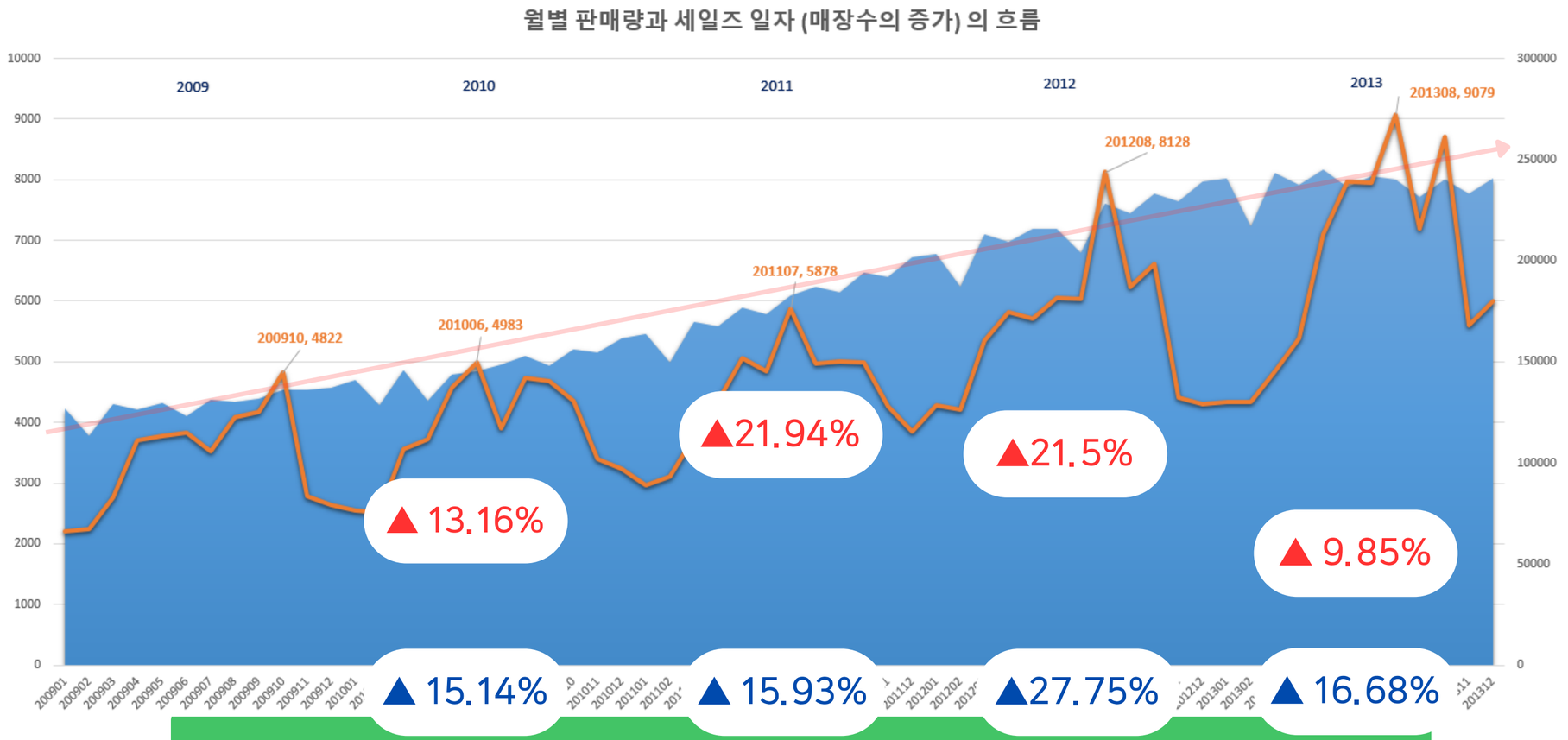


전년동 분기 대비
최대 28.68%
평균 19.86%
최저 13.09%
판매량의
상승발생

기초데이터 (판매량과 최대기온) **분기 단위**로 보면
3분기에 최대 온도가 가장 높고, 해당 분기 판매량이 매년 최대치를 갱신하는것으로 보여집니다.
또한 분기 판매량은 전년 동 분기 대비 최소 13% ~ 최대 28.68% 상승하고 있습니다.

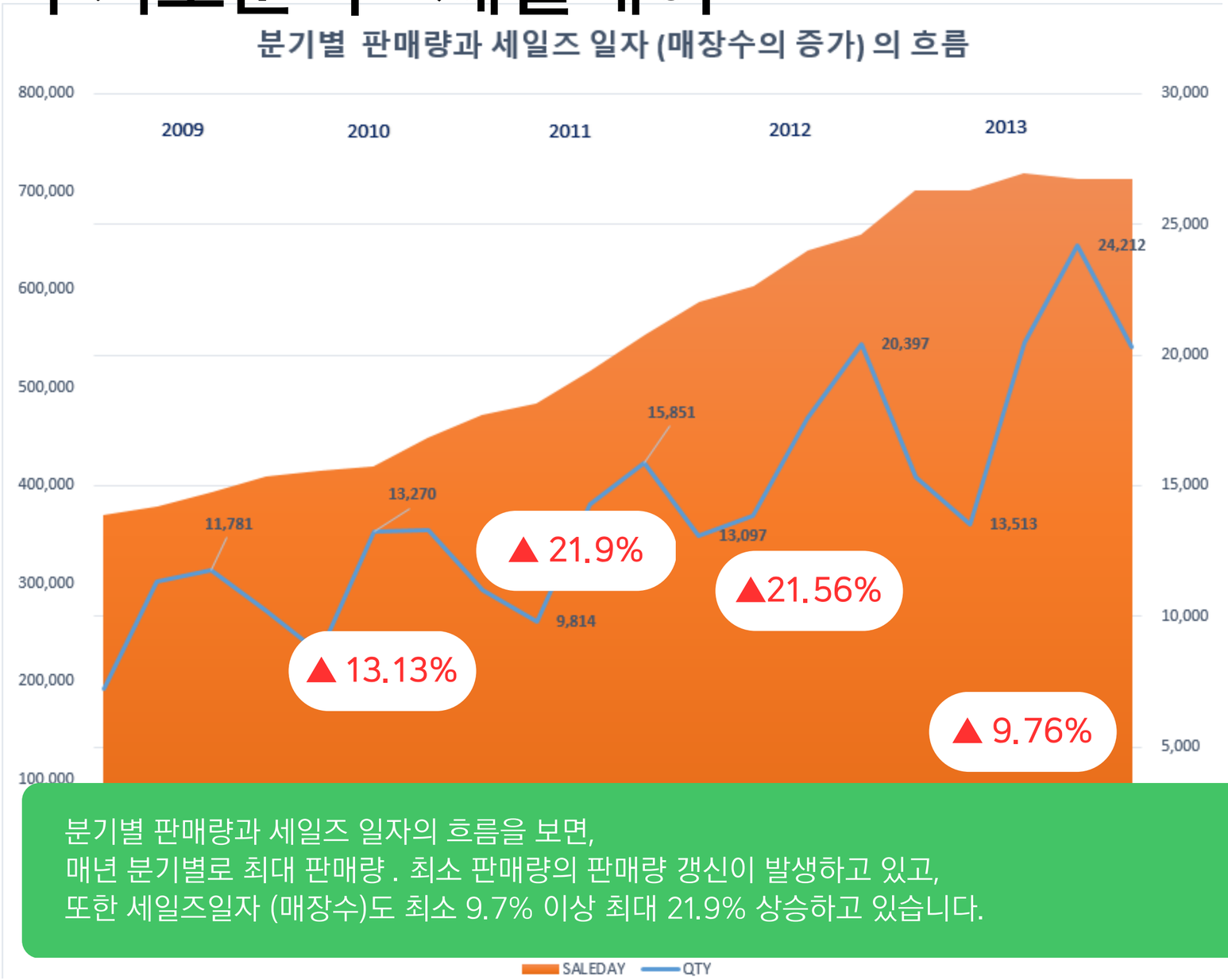
1. 데이터 기초분석-세일데이

▲ 기간내 매년 평균 16.63% 세일데이 상승



월별 판매량과 세일즈 일자를 시계열로
확인해보면 전체 세일즈 일자가 증가하면서 (매장수의 증가)
판매량의 최대/평균/최소치 모두 상승하는 흐름이 보입니다. 다만 2013년의 경우 세일즈일자의 상승은 월 평균 9.85% 로 전기 대비 낮은 상승률을 보였으나, 판매량은 월평균 16.6% 상승하여 선전하였다고 보여집니다.

1. 데이터 기초분석 -세일데이



분기별 판매량과 세일즈 일자의 흐름을 보면,
매년 분기별로 최대 판매량 . 최소 판매량의 판매량 갱신이 발생하고 있고,
또한 세일즈일자 (매장수)도 최소 9.7% 이상 최대 21.9% 상승하고 있습니다.

1. 데이터 기초분석- 결론 페이지

18.8%

평균 판매량 상승

월간데이터를 기준으로 매년 평균 18.87% 판매량 상승

28.6%

분기 최대 판매량 상승

매년 최대 판매가 이루어지는 3분기를 기준으로 최대 2012년 3분기에 전년 동분기 대비 28.68% 판매량 상승

16.6%

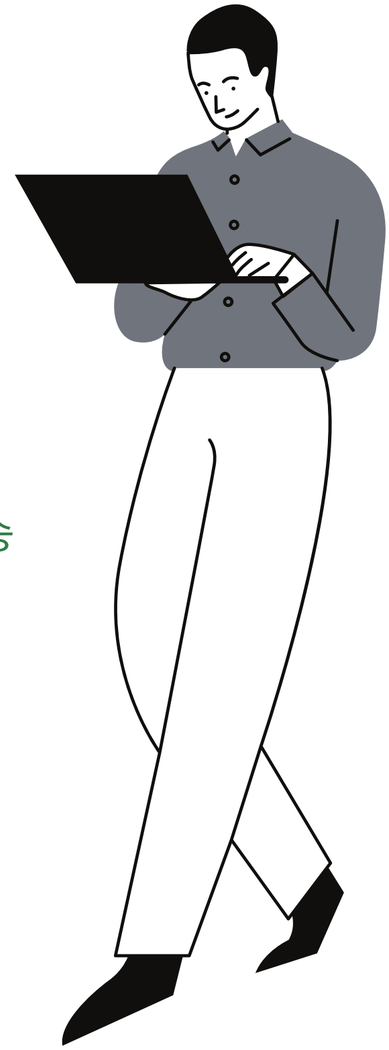
월간 세일즈데이 상승

매년 평균 16.63% 세일즈데이가 증가하고 있습니다. 특히 2011년의 세일즈 데이 증가폭이 가장 큽니다.

21.9%

분기 최대 세일즈데이 상승

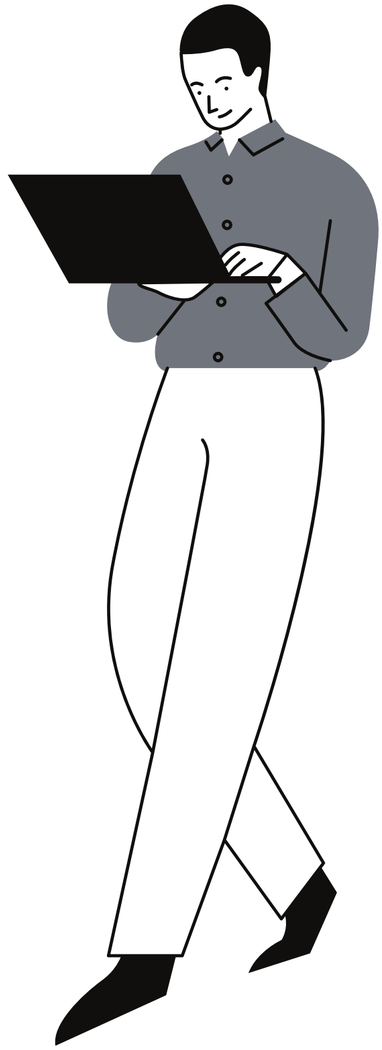
2011년 3분기에 전년동분기 대비 21.9%의 세일즈데이 증가분 발생



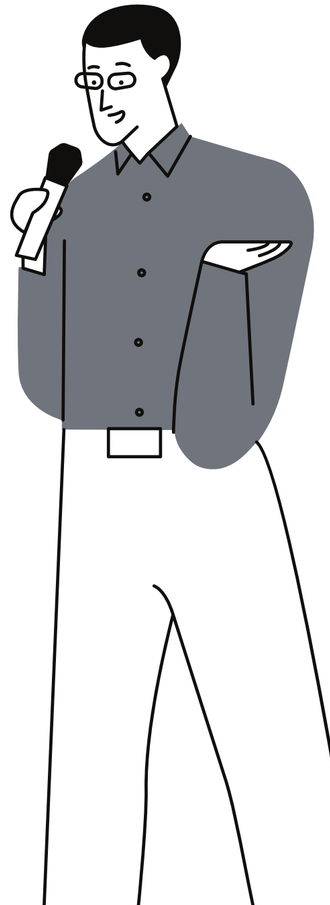
1. 질문셋

2009년을 제외한 모든 연도에서 3분기에 최대 판매실적이 보임
최대온도는 월 판매량 최대치가 발생하는 절대적인 요인인가?

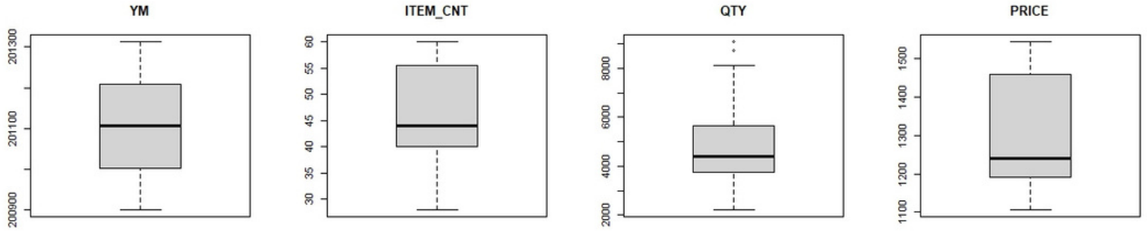
판매량의 최대값이 지속적으로 상승하는것과
가장 관련이 높은 독립 변수는 무엇인가?

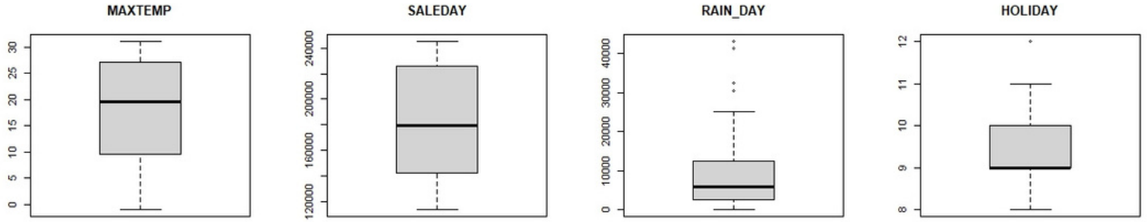


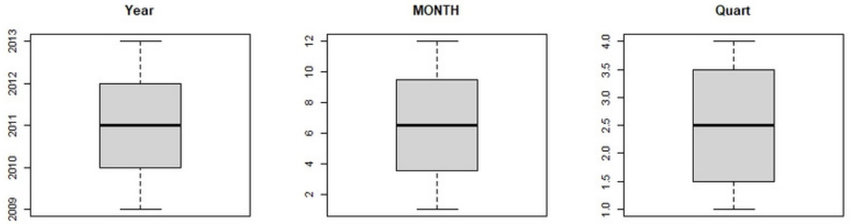
2. 데이터 전처리 및 상관분석



2. 데이터 전처리



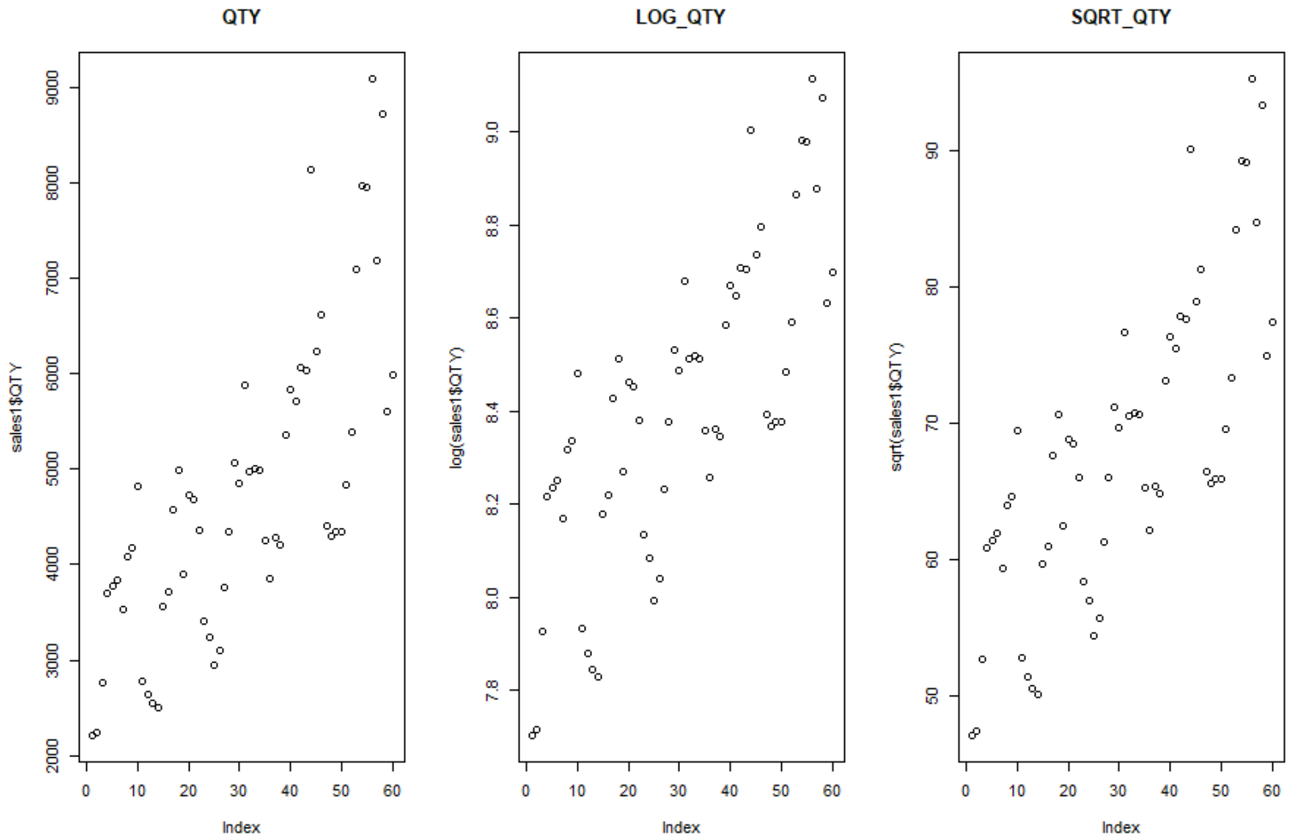





```
> table(is.na(sales1))

FALSE
  720
```

우선, 데이터를 분석하기 앞서 모집단의 이상치와 결측치는 없는 지 확인해 보았습니다.
QTY와 RAIN_DAY, HOLIDAY에서 이상치가 발견되었으나, 각 데이터의 범주에서 크게 벗어나지 않아 데이터 변환하지 않았습니다. 결측치는 없는 것으로 확인되었습니다.

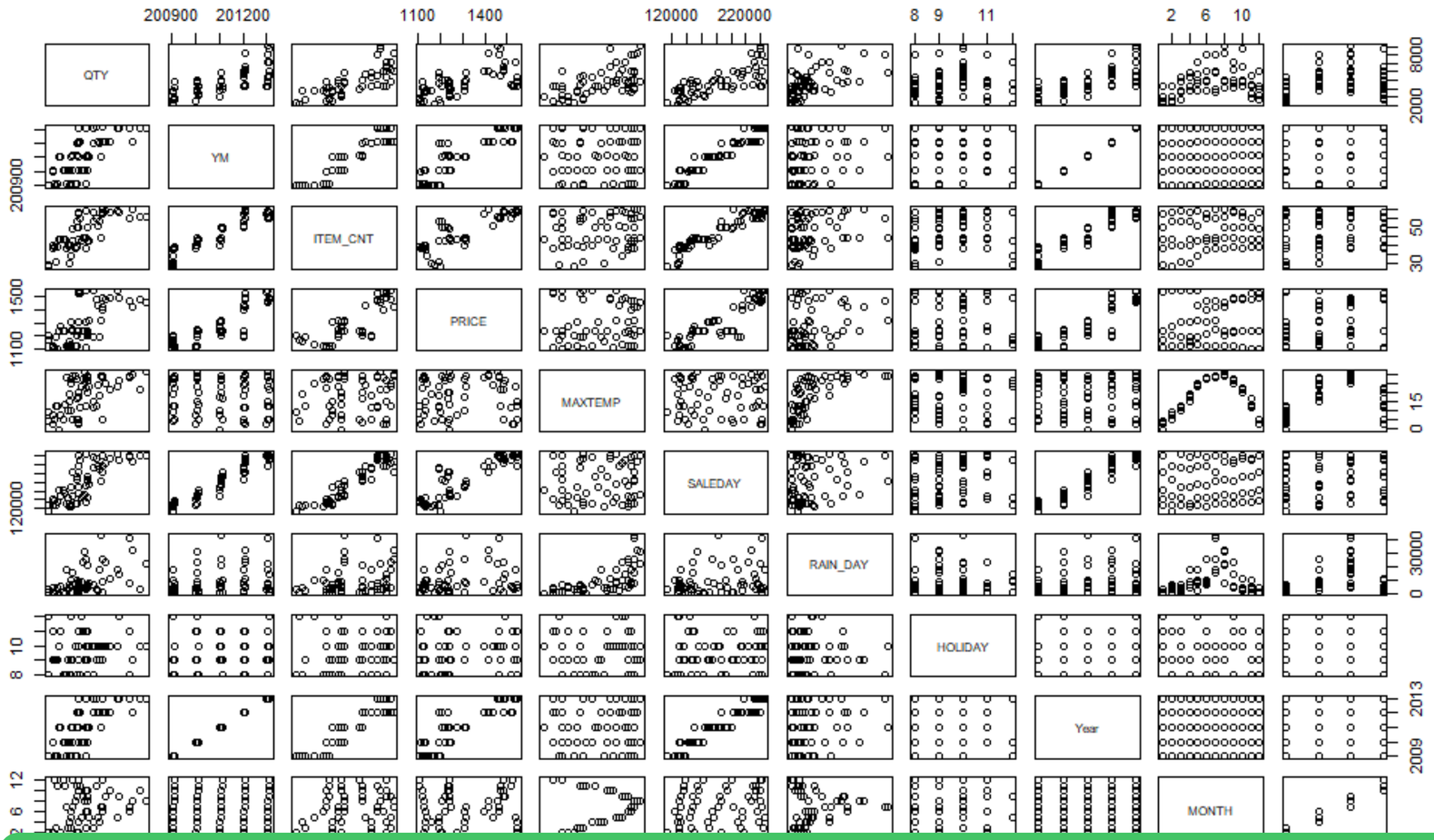
2. 데이터 전처리




```
> shapiro.test(sales1$QTY)

        Shapiro-Wilk normality test

data:  sales1$QTY
W = 0.94651, p-value = 0.01071

> shapiro.test(log(sales1$QTY))

        Shapiro-Wilk normality test

data:  log(sales1$QTY)
W = 0.98488, p-value = 0.6638

> shapiro.test(sqrt(sales1$QTY))

        Shapiro-Wilk normality test

data:  sqrt(sales1$QTY)
W = 0.97596, p-value = 0.2822
```

종속변수가 될 QTY데이터의 정규성을 확인해 보았습니다.
QTY 데이터는 기본적으로 정규성을 갖으나 **LOG로 데이터 변환** 했을 때, 정규성이 더욱 뚜렷해지는 것을 확인할 수 있습니다. 따라서, 종속변수는 LOG로 데이터 변환된 QTY로 사용했습니다.

2. 데이터 상관분석



상관분석은 변수간의 관계를 이해하기 위한 목적에 있으며,
변수 사이의 연관정도와 긍부정 관계를 식별하고, 또한 회귀모델에 사용하기에 적합한 변수식별에 도움이 됩니다.

2. 데이터 상관분석

	QTY	YM	ITEM_CNT	PRICE	MAXTEMP	SALEDAY	RAIN_DAY	HOLIDAY	Year	MONTH	Quart
QTY	1.0000	0.7269	0.7315	0.6547	0.5608	0.7564	0.5204	0.1605	0.7205	0.2704	0.2809
YM	0.7269	1.0000	0.9266	0.8780	-0.0214	0.9642	0.1958	0.1478	0.9997	0.0244	0.0237
ITEM_CNT	0.7315	0.9266	1.0000	0.8235	0.0917	0.9445	0.2624	0.1069	0.9212	0.2333	0.2233
PRICE	0.6547	0.8780	0.8235	1.0000	0.0451	0.8893	0.2146	0.1287	0.8752	0.1225	0.1161
MAXTEMP	0.5608	-0.0214	0.0917	0.0451	1.0000	0.0262	0.6749	-0.0007	-0.0280	0.2725	0.2784
SALEDAY	0.7564	0.9642	0.9445	0.8893	0.0262	1.0000	0.2135	0.1509	0.9599	0.1854	0.1813
RAIN_DAY	0.5204	0.1958	0.2624	0.2146	0.6749	0.2135	1.0000	-0.0514	0.1929	0.1206	0.1538
HOLIDAY	0.1605	0.1478	0.1069	0.1287	-0.0007	0.1509	-0.0514	1.0000	0.1503	-0.1026	-0.1141
Year	0.7205	0.9997	0.9212	0.8752	-0.0280	0.9599	0.1929	0.1503	1.0000	0.0000	0.0000
MONTH	0.2704	0.0244	0.2333	0.1225	0.2725	0.1854	0.1206	-0.1026	0.0000	1.0000	0.9716
Quart	0.2809	0.0237	0.2233	0.1161	0.2784	0.1813	0.1538	-0.1141	0.0000	0.9716	1.0000

회귀분석은 하나이상의 독립 변수로 종속 변수의 값을 예측하는 모델을 구축하는 기술입니다.
다른 변수를 통제하면서 종속변수와 독립변수 관계를 이해하는데 사용됩니다.

상관분석 후 회귀분석을 수행함으로써 종속변수와 상관관계가 높은 적절한 변수를 선택하여 보다 정확한 모형을 구축하려고 상관분석을 먼저 수행했습니다.

3. 데이터 회귀분석

3. 데이터 회귀분석

```
> model=lm(log(QTY)~YM+ITEM_CNT+PRICE+MAXTEMP+SALEDAY+RAIN_DAY+HOLIDAY+Year+MONTH+Quart,data=sa
> summary(model)

Call:
lm(formula = log(QTY) ~ YM + ITEM_CNT + PRICE + MAXTEMP + SALEDAY +
    RAIN_DAY + HOLIDAY + Year + MONTH + Quart, data = sales1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.18723 -0.06945  0.00605  0.04911  0.22779

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.016e+02  8.731e+01  -3.455  0.00113 **
YM          -1.826e-05  1.548e-02  -0.001  0.99906
ITEM_CNT    -5.708e-03  4.671e-03  -1.222  0.22745
PRICE       -5.312e-04  1.867e-04  -2.845  0.00642 **
MAXTEMP      2.106e-02  1.806e-03  11.665 7.01e-16 ***
SALEDAY      3.541e-06  1.485e-06   2.384  0.02095 *
RAIN_DAY    -1.670e-06  1.719e-06  -0.971  0.33611
HOLIDAY      8.437e-03  1.056e-02   0.799  0.42790
Year         1.559e-01  1.538e+00   0.101  0.91967
MONTH               NA         NA      NA       NA
Quart        3.592e-02  4.620e-02   0.778  0.44052
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09262 on 50 degrees of freedom
Multiple R-squared:  0.9341,    Adjusted R-squared:  0.9222
F-statistic: 78.75 on 9 and 50 DF,  p-value: < 2.2e-16
```

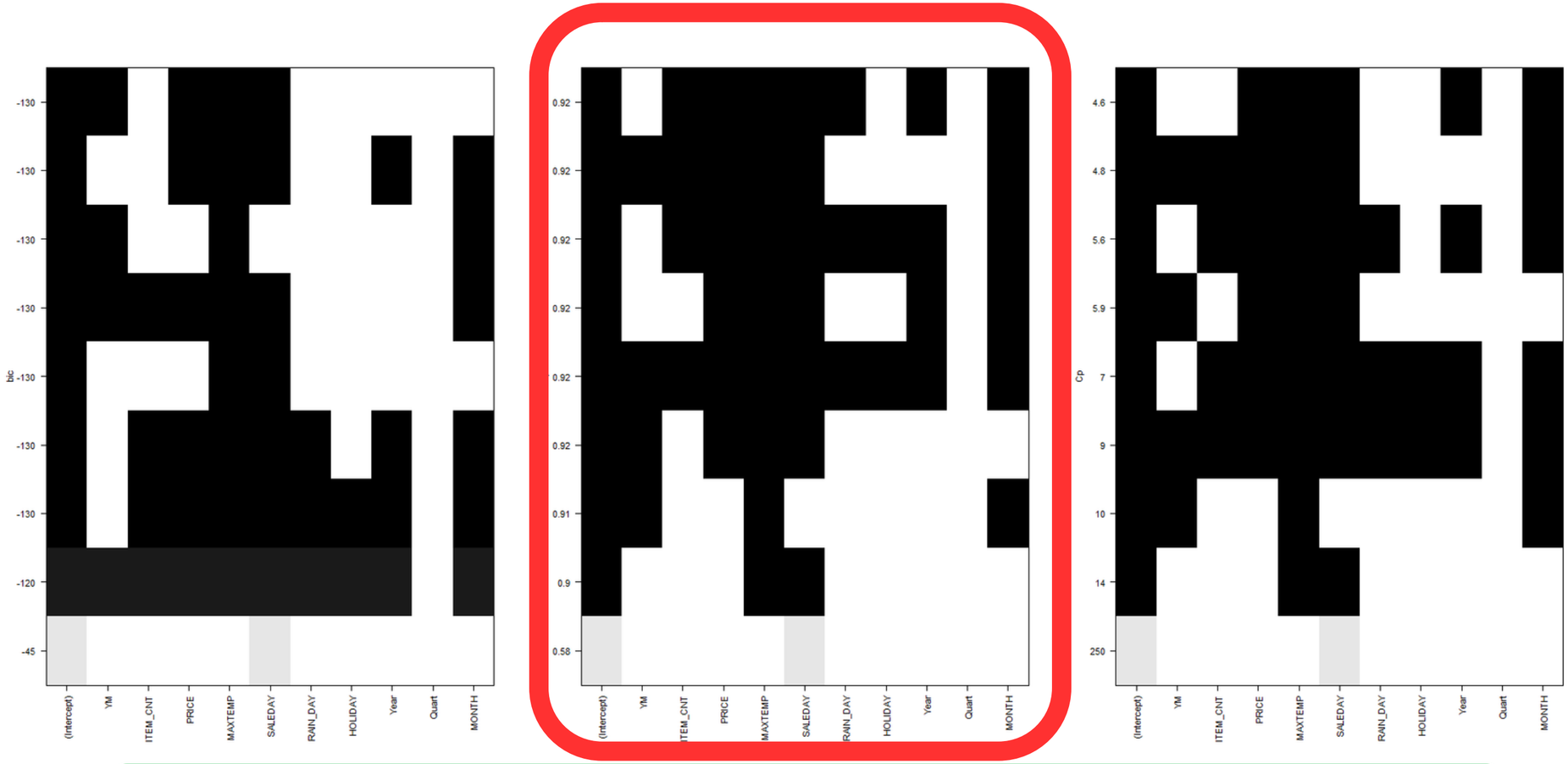
QTY를 종속변수로 정하고, 범주형 변수인 카테코리를 제외한 모든 컬럼을 독립변수로 잡았습니다. 선형 회귀모델 코드를 구성했습니다. model이라는 이름으로 구성했습니다. model을 summary하여 얻은 결과는 다음과 같습니다.

3. 데이터 회귀분석

```
install.packages("leaps")
library(leaps)
leaps=regsubsets(log(QTY)~YM+ITEM_CNT+PRICE+MAXTEMP+SALEDAY+RAIN_DAY+HOLIDAY+
Year+MONTH+Quart,data=sales1,nvmax=10)
summary(leaps)
plot(leaps)
plot(leaps,scale="adjr2")
plot(leaps,scale="Cp")
```

bic, adjr2, Cp 3가지로 leaps 모델을 구성했습니다. 3가지로 분석했을 때 공통적으로 포함되는 핵심 독립변수 PRICE, MAXTEMP,SALEDAY가 있습니다. 그중 y축 범위가 0~1로, 분석이 용이하다고 판단된 adjr2를 중심으로 이후 분석을 진행했습니다. (Adjusted R squared Value 가 적합성을 가장 잘 나타내는 모델이라고 판단하여 선택하였습니다)

3. 데이터 회귀분석



bic, adjr2, Cp 3가지로 leaps 모델을 구성했습니다. 3가지로 분석했을 때 공통적으로 포함되는 핵심 독립변수 PRICE, MAXTEMP,SALEDAY가 있습니다. 그중 y축 범위가 0~1로, 분석이 용이하다고 판단된 adjr2를 중심으로 이후 분석을 진행했습니다. (Adjusted R squared Value 가 적합성을 가장 잘 나타내는 모델이라고 판단하여 선택하였습니다)

3. 데이터 회귀분석

summary.out=summary(leaps)
which.max(summary.out$adjr2)
summary.out$which[7,]

```
> summary.out$which[7,]
(Intercept)          YM     ITEM_CNT       PRICE     MAXTEMP     SALEDAY    RAIN_DAY
       TRUE       FALSE        TRUE        TRUE        TRUE        TRUE        TRUE
    HOLIDAY        Year       MONTH       Quart
      FALSE        TRUE       FALSE        TRUE
```

adjr2 leaps 분석결과 YM, HOLIDAY, MONTH가 FALSE가 나왔습니다. 회귀분석 정확도를 높이기 위해 FALSE 가 나온 값들을 제외하기로 결정했습니다.

3. 데이터 회귀분석

```
> summary(model2)

Call:
lm(formula = log(QTY) ~ ITEM_CNT + PRICE + MAXTEMP + SALEDAY +
    RAIN_DAY + Year + Quart, data = sales1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.191174 -0.064337  0.009035  0.043871  0.234262

Coefficients:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -2.980e+02  8.366e+01  -3.562 0.000798 ***
ITEM_CNT    -5.947e-03  4.508e-03  -1.319 0.192864
PRICE       -5.354e-04  1.837e-04  -2.915 0.005237 **
MAXTEMP      2.124e-02  1.759e-03  12.076  < 2e-16 ***
SALEDAY      3.710e-06  1.444e-06   2.569 0.013097 *
RAIN_DAY    -1.833e-06  1.651e-06  -1.110 0.271947
Year         1.523e-01  4.173e-02   3.649 0.000611 ***
Quart        3.400e-02  1.490e-02   2.283 0.026575 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0914 on 52 degrees of freedom
Multiple R-squared:  0.9333,    Adjusted R-squared:  0.9243
F-statistic: 103.9 on 7 and 52 DF,  p-value: < 2.2e-16
```

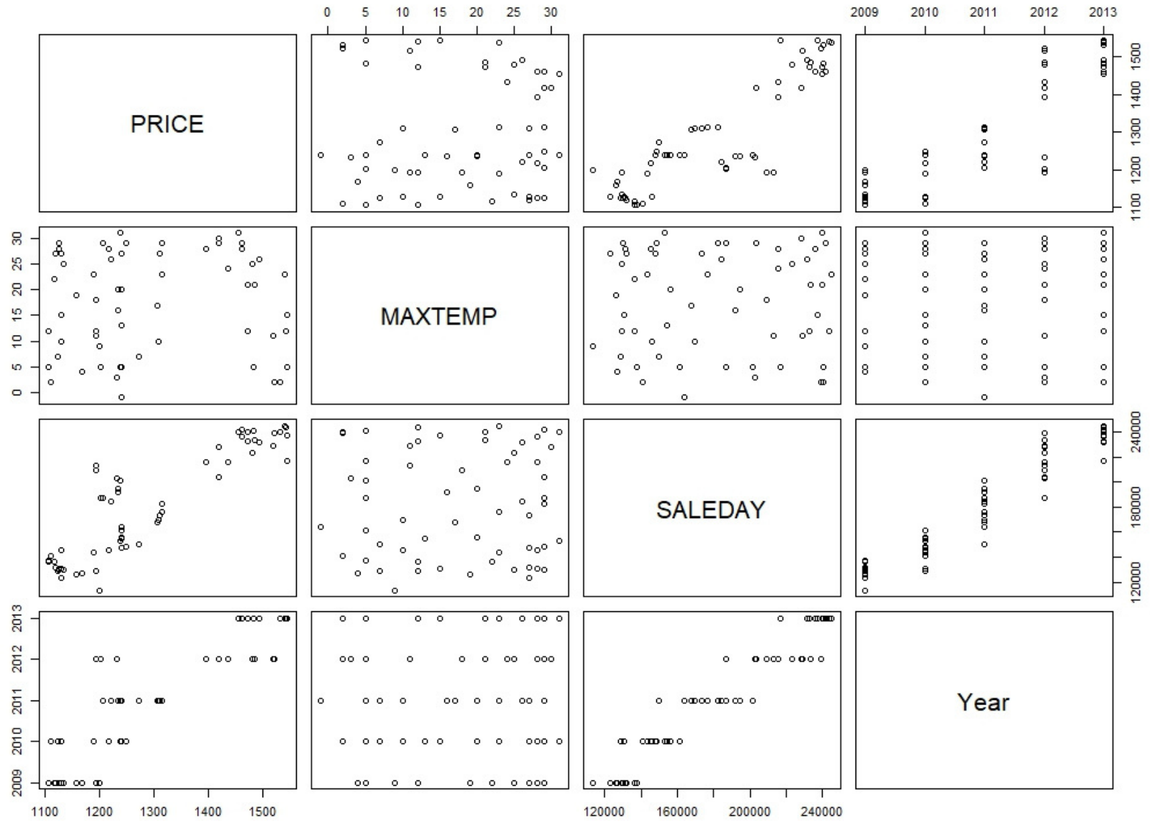
TRUE가 나온 독립변수로 model2 회귀모델을 만들었고, 결과는 다음과 같습니다.
이후, P-value(Pr(>|t|))값이 높게 나온 ITEM_CNT와 RAIN_DAY를 독립변수에서 제외시켰습니다.

3. 데이터 회귀분석

```
> summary(model3)

Call:
lm(formula = log(QTY) ~ PRICE + MAXTEMP + SALEDAY + Year + Quart,
    data = sales1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.201002 -0.059898  0.001393  0.048785  0.247986

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.511e+02  7.856e+01  -3.196  0.00233 **
PRICE       -4.959e-04  1.829e-04  -2.711  0.00897 **
MAXTEMP      1.956e-02  1.290e-03  15.165  < 2e-16 ***
SALEDAY      3.102e-06  1.378e-06   2.251  0.02850 *
Year         1.289e-01  3.918e-02   3.289  0.00177 **
Quart        2.849e-02  1.430e-02   1.992  0.05143 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09238 on 54 degrees of freedom
Multiple R-squared:  0.9292,    Adjusted R-squared:  0.9226
F-statistic: 141.7 on 5 and 54 DF,  p-value: < 2.2e-16
```

3번째 모델입니다. 이번엔 Quart의 p-value값이 낮게 측정되었습니다.
Quart도 독립변수에서 제외시켜줍니다.

3. 데이터 회귀분석



	PRICE	MAXTEMP	SALEDAY	Year
PRICE	1.0000	0.0451	0.8893	0.8752
MAXTEMP	0.0451	1.0000	0.0262	-0.0280
SALEDAY	0.8893	0.0262	1.0000	0.9599
Year	0.8752	-0.0280	0.9599	1.0000

```
pairs(sales1[,c("PRICE","MAXTEMP","SALEDAY","Year")])
cor(sales1[,c("PRICE","MAXTEMP","SALEDAY","Year")])
```

독립변수끼리의 상관관계, 다중공선성을 확인해봤습니다.
첫번째로 PRICE와 SALEDAY의 상관관계가 높은 것을 확인할 수 있습니다.
두번째로 독립변수 YEAR가 PRICE, SALEDAY와 높은 상관성을 보이고 있어 회귀모델의 정확성을 위해 YEAR도 독립변수에서 제외시켜줍니다.

3. 데이터 회귀분석

```
> summary(model5)

Call:
lm(formula = log(QTY) ~ PRICE + MAXTEMP + SALEDAY, data = sales
1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.199136 -0.064018 -0.000996  0.055346  0.249051

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.312e+00  1.546e-01  47.305  < 2e-16 ***
PRICE       -4.130e-04  1.935e-04  -2.134   0.0372 *
MAXTEMP      1.943e-02  1.330e-03  14.615  < 2e-16 ***
SALEDAY      7.135e-06  6.657e-07  10.717  3.5e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09939 on 56 degrees of freedom
Multiple R-squared:  0.915,     Adjusted R-squared:  0.9105
F-statistic:   201 on 3 and 56 DF,  p-value: < 2.2e-16
```

5번째 회귀모델입니다.
앞서, PRICE와 SALEDAY의 상관성이 높게 나왔기 때문에, p-value값이 낮게 나온 PRICE를 마지막으로 독립변수에서 제외시켜줍니다.

3. 데이터 회귀분석

```
> summary(model6)

Call:
lm(formula = log(QTY) ~ MAXTEMP + SALEDAY, data = sales1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.187309 -0.073353 -0.005191  0.060218  0.250036

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.009e+00  6.266e-02  111.86   <2e-16 ***
MAXTEMP     1.930e-02  1.369e-03   14.10   <2e-16 ***
SALEDAY     5.871e-06  3.137e-07   18.72   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1024 on 57 degrees of freedom
Multiple R-squared:  0.9081,    Adjusted R-squared:  0.9049
F-statistic: 281.6 on 2 and 57 DF,  p-value: < 2.2e-16
```

최종 회귀모델입니다.
회귀분석 결과 MAXTEMP, SALEDAY는 모두 QTY와 유의미한 관계가 있다는 것을 확인할 수 있습니다.

3. 데이터 회귀분석

최종회귀식

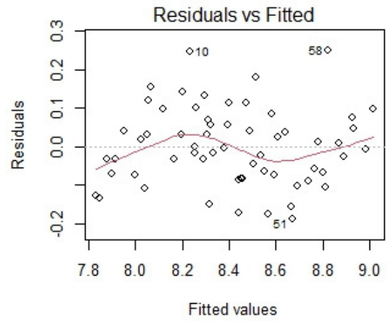
log(QTY) = 7.009 + 0.01930 * MAXTEMP + 0.000005871 * SALEDAY

```
> summary(model6)

Call:
lm(formula = log(QTY) ~ MAXTEMP + SALEDAY, data = sales1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.187309 -0.073353 -0.005191  0.060218  0.250036

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.009e+00  6.266e-02  111.86   <2e-16 ***
MAXTEMP     1.930e-02  1.369e-03   14.10   <2e-16 ***
SALEDAY     5.871e-06  3.137e-07   18.72   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1024 on 57 degrees of freedom
Multiple R-squared:  0.9081,    Adjusted R-squared:  0.9049
F-statistic: 281.6 on 2 and 57 DF,  p-value: < 2.2e-16
```

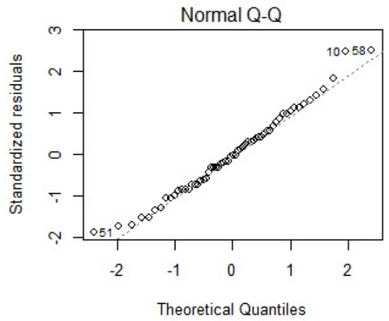
MAXTEMP와 SALEDAY는 종속변수 QTY와 양의 상관관계를 가지고,
R-squared은 0.9081로 QTY 변동의 90.81%를 설명할 수 있습니다.
F-통계량은 281.6이며, P-value 값은(p-value: < 2.2e-16) 매우 작은 값으로, 해당 회귀모델이 유의하다는 것을 알 수 있습니다.

3. 데이터 회귀분석

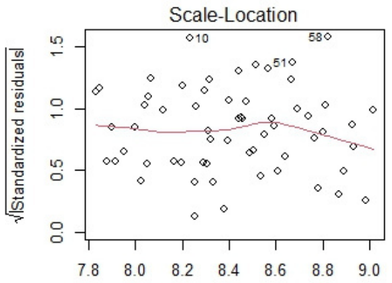
최종회귀모델의 잔차분석과 오차분석

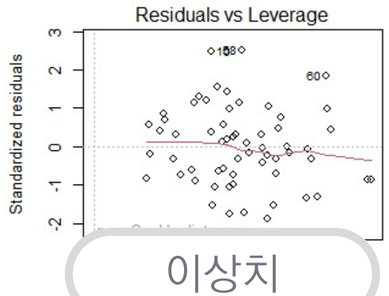


등분산성



정규성





이상치

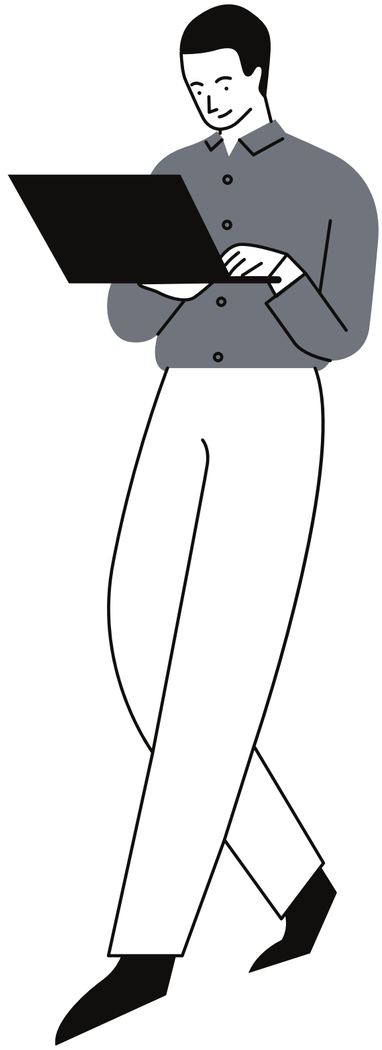
```
> #잔차들의 정규성 검정
> res= residuals(model6)
> shapiro.test(res)

        Shapiro-Wilk normality test

data:  res
W = 0.98392, p-value = 0.6142

> #잔차들의 독립성 검정
> dwtest(model6)

        Durbin-Watson test

data:  model6
DW = 1.4421, p-value = 0.006951
alternative hypothesis: true autocorrelation is greater than 0

> k=mean((sales1$QTY-exp(model6$fitted.values))^2)
> sqrt(k)
[1] 507.413
> mean(sales1$QTY)
[1] 4757
```

최종회귀모델의 잔차분석입니다.
그래프와 코드를 보면 잔차가 정규성을 갖는 것을 확인 할 수 있고,
오차는 507.413으로 실제값이 4757인 것에 비해 낮으므로 해당 회귀모델이 유의하다는 것을 알 수 있습니다.

3.데이터 회귀분석 - 결론

MAXTEMP와 SALEDAY는 종속변수 QTY와 양의 상관관계를 가지고,
R-squared은 0.9081로 QTY 변동의 90.81%를 설명할 수 있습니다.
F-통계량은 281.6이며, P-value 값은(p-value: < 2.2e-16) 매우 작은 값으로, 해당 회귀모델이 유의하다는 것을 알 수 있습니다.

따라서, 데이터 기초분석에서 확인한 바와 같이 일반 탄산음료의 매출량은 그달의 최대기온과 영업일수에 가장 영향을 받는 것으로 확인됐습니다.

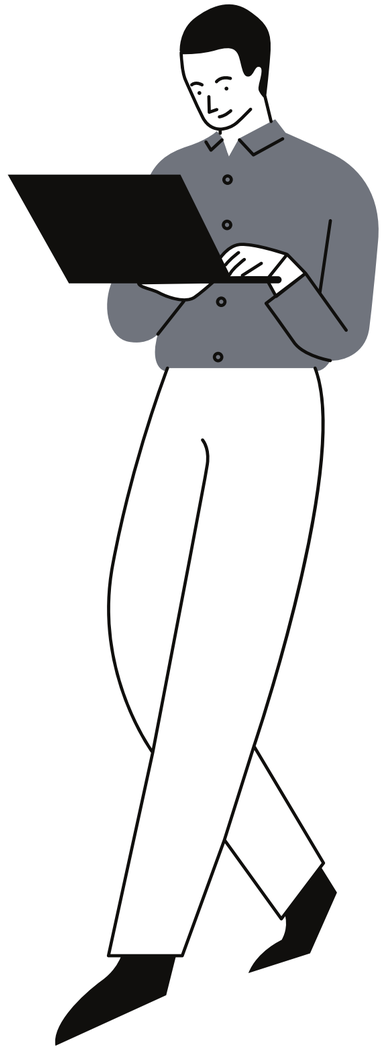


4.요약 및 결론

4.요약 및 결론

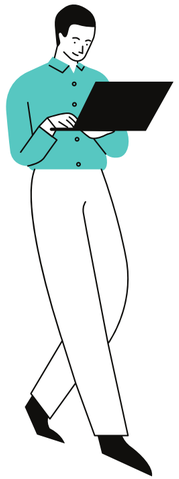
월별 일반 탄산 음료의 판매량에 가장 영향을 끼치는것은
최대 온도의 변화이다.

전체 판매량의 상승분에 가장 영향을 끼치는 것은
세일데이 (매장의 증가) 이다.

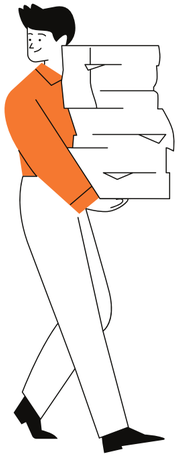


감사합니다.

김진성
데이터분석 메인



김성진
스토리텔링-PPT



김호성
발표, 데이터 스토리텔링