[중간고사]빅데이터마이닝 패턴탐색

전공: 데이터테크전공 학번: 20173204 성명: 곽명빈

※ 1~3번은 연결된 문제입니다. 4번은 1번~3번 중, 1개의 문항을 대신할 수 있습니다. 1번에서 3번 중, 어떤 문제를 대신할지 정확히 표기하고 4개의 문항 중, 3개를 선택하여 답안지에 표기하세요. 단, 포기한 문제가 맞고, 선택한 4번이 틀릴 경우, 선택한 4번으로 대체하여 총점됩니다. 실습관련하여 5~7번 문제도 같은 답안지에 pdf 파일로 제출하길 바랍니다!

1. 중국 칭다오 도시의 HL-대형마트에 새로 부임한 최 우헌 매장관리 팀장은 20대의 최연소 메니저로서, 최근 10년간 부진한 매출을 높이기 위해 임용되었다. 대표이사와의 계약 체결은 직전년도 영업 대비 200% 매출증가와 증가분의 10%를 성과급으로 받는다는 조건으로 연봉체결을 진행하였다(직전년도 1월1일~12월 31일까지 영업수익: 20만엔, 1년간 인건비: 18만엔, 유틸리티 제반 등의 지출비용: 15만엔).

최우헌 매니저는 가장 먼저 매장의 구분된 품목의 데이터를 살펴보았다. 식품관, 전자제품관, 의류관, 유아용품관, 주류관, 가구관으로 나뉘어져 있었다. 가장 적자를 이루는 아래의 표에서 분류된 2개 관의 매출수익 현황이다. 하루평균 마트의 거래 건수는 200개이다.

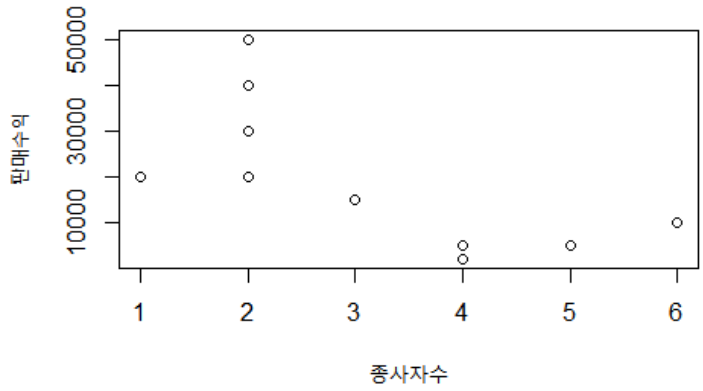
(종사자 인건비 : 1년 평균 5천엔 적용)

구분	종사자수	상품종류(품목)	판매수익(엔)	재고 비율
식품관	3	곡류	1.5 만엔	80%
	6	생선류	1.0 만엔	60%
	5	건어물류	0.5 만엔	50%
	2	고기류	4.0 만엔	10%
	2	채소류	3.0 만엔	10%
	1	면류	2.0 만엔	0%
주류/음료관	2	고량주	2.0 만엔	40%
	2	맥주	5.0 만엔	10%
	4	와인	0.5 만엔	90%
	4	주스류	0.2 만엔	80%
	2	탄산음료류	4.0 만엔	20%

1-1) 본인이 최우헌 매니저라면, 위의 가장 낮은 영업실적을 지닌 현황표를 보고, 매출수익 증가를 위한 전략을 기획하기 위하여 어떤 문제제기를 하고, 이에 해결을 위하여 접근하는 방법을 제시하시오.

문제를 해결하기 위해 Excel로 데이터를 만들어 문제점을 조사해보았다. 시각화는 R프로그래밍을 이용하였다.

1) 종사자수와 판매수익이 비례하지 않는다.

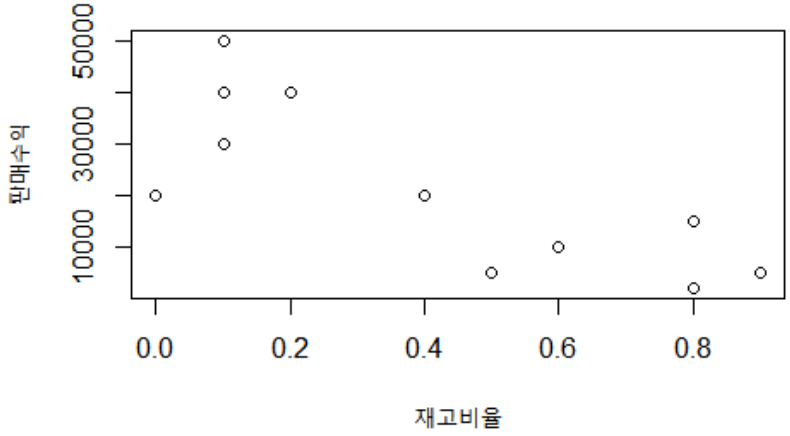


R로 데이터의 분포를 확인해 보았을때 종사자수가 높다고 판매수익이 높은게 아님을 확인하였고 상관계수 또한 확인해보았다.

```
> cor(df$종사자수, df$판매수익)
[1] -0.6830668
```

음의 상관관계를 보인다.

2) 재고비율과 판매수익이 비례하지 않는다.



재고비율과 판매수익의 분포또한 확인하였을때 보통 재고가 적을수록 판매수익이 높다는 사실을 확인 할 수 있었다.

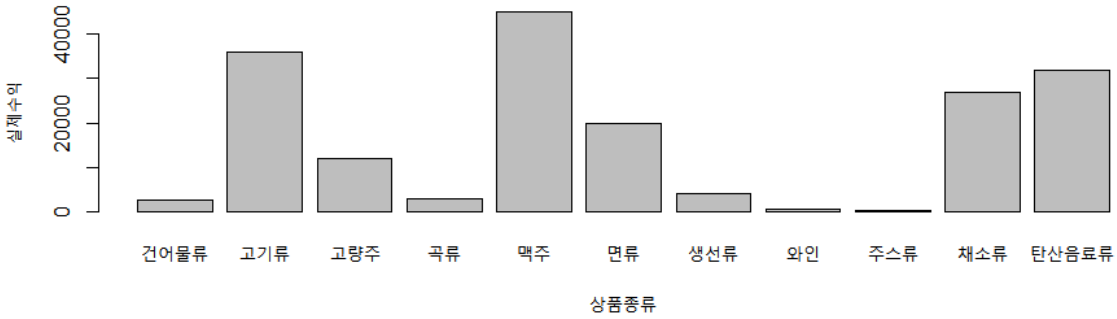
```
> cor(df$재고비율, df$판매수익)
[1] -0.7850379
```

음의 상관관계를 보인다

3) 수익성

위 분포를 바탕으로 수익성이라는 열을 만들었다. * 수익성 = 수익 X (1-재고비율)

수익성이란 재고량과 상관없이 수익만으로 어떤 품목이 좋을까(인기가 있을까)를 판단할 수 있는 근거로 사용한다.



재고량을 제외한 수익성이 좋은 품목들을 막대그래프를 통해 확인할 수 있다.

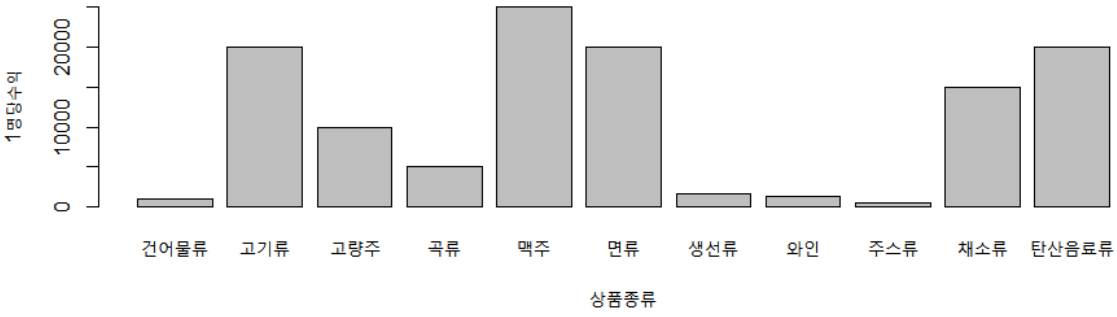
살펴본 결과 맥주, 고기, 탄산, 채소, 면 순으로 수익성이 좋다고 할 수 있으며,

기존의 판매수익의 순서와 크게 다르지 않아 위 5가지의 품목의 수익성이 좋다고 판단 하였다.

4) 종사자 1인당 수익

종사자 1명당 수익을 살펴보았다. 종사자 1명당 수익 = 판매수익 / 종사자수

기존의 수익성이 높은 결과와 크게 다르지 않다는 것을 확인하였다.



```
> summary(df)
    종사자수    상품종류            판매수익         재고비율          인건비          수익성
 Min.   :1   Length:11          Min.   : 2000   Min.   :0.0000   Min.   : 5000   Min.   :  400
 1st Qu.:2   Class :character   1st Qu.: 7500   1st Qu.:0.1000   1st Qu.:10000   1st Qu.: 2750
 Median :2   Mode  :character   Median :20000   Median :0.4000   Median :10000   Median :12000
 Mean   :3                      Mean   :21545   Mean   :0.4091   Mean   :15000   Mean   :16582
 3rd Qu.:4                      3rd Qu.:35000   3rd Qu.:0.7000   3rd Qu.:20000   3rd Qu.:29500
 Max.   :6                      Max.   :50000   Max.   :0.9000   Max.   :30000   Max.   :45000
```

위의 결과를 바탕으로 종사자수와 판매수익은 비례하지 않는다 는 사실을 확인 하였다.

재고량을 제외한 수익성이 좋은 품목이 판매수익이 높은 품목과 비슷하다는 점 또한 확인할 수 있었 다.

재고가 많은 품목의 수익성이 낮고, 종사자 수가 많다는 사실을 데이터를 통해 확인할 수 있었다.

결국 매출수익 증가를 위해서는 기존의 인기 없는 품목의 종사자의수를 줄이고, 수익성이 좋지 못한 품목의 매입량을 줄여야만 한다.

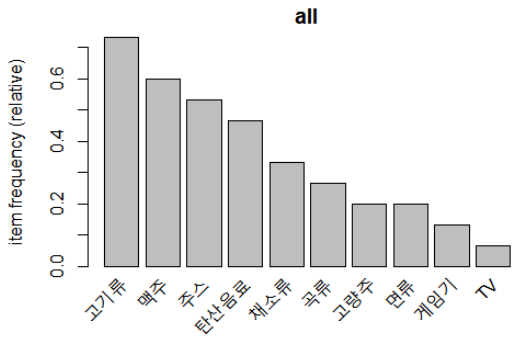
수익성의 평균 값인 16582원을 기준으로하여 높은 품목의 종사자수를 늘리고 낮은 품목의 종사자 수와 매입량을 조절해야 할 것이다.

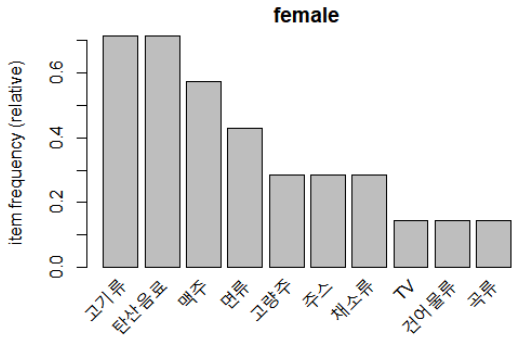
2. 365일 중 어느 날 매장을 방문하여 구매한 거래 장부를 살펴보니, 다음과 같다.

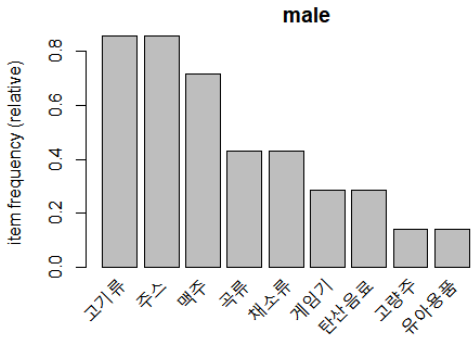
거 래	성 별	출생년도	품 목
1	F	1982	면류, 고기류, 생필품, 주스
2	F	1969	곡류, 고량주, 채소류, 생선류, 탄산음료, TV
3	F	1958	면류, 고량주, 채소류, 고기류, 탄산음료, 휴지
4	M	1959	곡류, 맥주, 채소류, 고기류, 주스, 유아용품
5	M	2002	맥주, 고기류, 주스, 탄산음료
6	M	2001	맥주, 탄산음료, 주스, 고기류, 게임기
7	F	1999	탄산음료, 맥주, 건어물류, 면류
8	M	2001	게임기, 고기류, 맥주, 주스
9	M	2004	곡류, 고기류, 주스, 채소류
10	F	2006	맥주, 고기류
11	F	1989	맥주, 탄산음료, 주스, 고기류
12	F	2002	고기류, 탄산음료, 맥주
13	M	1997	고기류, 곡류, 채소류, 고량주
14	M	1996	맥주, 주스

2-1) 상품판매의 특징과 고객의 특징을 분류하여 탐색하시오

성별에 따른 구매품목의 차이





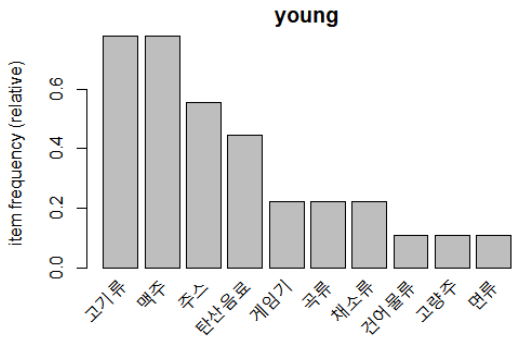


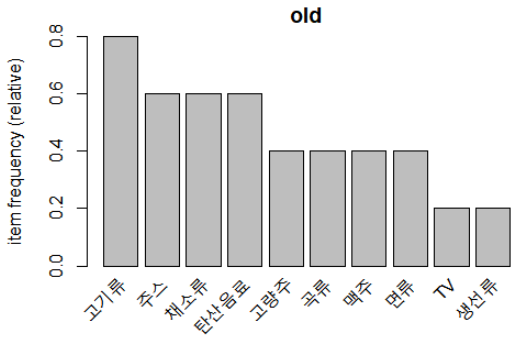
남성과 여성의 상품 구매 빈도수를 살펴보면

남성은 고기, 탄산, 맥주, 면류를 많이 사는것으로 나타났고,

여성은 고기, 주스, 맥주, 곡류를 많이 사는것으로 나타났다.

나이에 따른 구매품목의 차이





나이에 따른 구매품목의 빈도수도 살펴 보았다. (나이는 평균값인 1990을 기준으로 크면 young, 낮으면 old에 해당)

어린 그룹에서는 고기, 맥주, 주스, 탄산을 많이 사는것으로 나타났고,

늙은 그룹에서는 고기, 주스, 채소, 탄산을 많이 사는것으로 나타났다.

위의 두 결과를 바탕으로 나이, 성별에 상관없이 고기+마실것을 많이 사는것으로 파악된다.

남성은 고기와 탄산을 선호하고 여성은 고기와 주스를 선호하는 것을 알 수 있고, 맥주는 두 성별 모두에게 인기가 있다.

어린그룹과 늙은그룹의 차이는 나이가 많을수록 맥주의 빈도가 줄어들며 채소류가 증가한다는 사실을 그래프를 통해 파악 가능하다.

2-2) 매출증진을 위한 전략 계획을 설명하고, 그 이유를 제시하시오.

위 그래프를 통해 고기의 빈도수가 항상 높다는 사실을 알고있다. 그렇기 때문에 고기류에 초점을 맞춘 매출증진 전략을 사용해야 한다.

데이터를 보면 모든 그룹이 고기 + 고기와 함께 먹는음식으로 빈도수가 높다는 사실을 확인할 수 있고 연관규칙을 통해 고기와 함께 묶이는 것들의 지지도 신뢰도 향상도 를 살펴보면 다음과 같다. (지지도 0.2 이상, 신뢰도 0.25 이상, 고기를 포함)

	lhs		rhs	support	confidence	coverage	lift	count
[1]	{}	=>	{고기류}	0.7333333	0.7333333	1.0000000	1.0000000	11
[2]	{곡류}	=>	{고기류}	0.2000000	0.7500000	0.2666667	1.0227273	3
[3]	{고기류}	=>	{곡류}	0.2000000	0.2727273	0.7333333	1.0227273	3
[4]	{채소류}	=>	{고기류}	0.2666667	0.8000000	0.3333333	1.0909091	4
[5]	{고기류}	=>	{채소류}	0.2666667	0.3636364	0.7333333	1.0909091	4
[6]	{탄산음료}	=>	{고기류}	0.3333333	0.7142857	0.4666667	0.9740260	5
[7]	{고기류}	=>	{탄산음료}	0.3333333	0.4545455	0.7333333	0.9740260	5
[8]	{주스}	=>	{고기류}	0.4666667	0.8750000	0.5333333	1.1931818	7
[9]	{고기류}	=>	{주스}	0.4666667	0.6363636	0.7333333	1.1931818	7
[10]	{맥주}	=>	{고기류}	0.4666667	0.7777778	0.6000000	1.0606061	7
[11]	{고기류}	=>	{맥주}	0.4666667	0.6363636	0.7333333	1.0606061	7
[12]	{곡류,채소류}	=>	{고기류}	0.2000000	0.7500000	0.2666667	1.0227273	3
[13]	{고기류,곡류}	=>	{채소류}	0.2000000	1.0000000	0.2000000	3.0000000	3
[14]	{고기류,채소류}	=>	{곡류}	0.2000000	0.7500000	0.2666667	2.8125000	3
[15]	{주스,탄산음료}	=>	{고기류}	0.2000000	1.0000000	0.2000000	1.3636364	3
[16]	{고기류,탄산음료}	=>	{주스}	0.2000000	0.6000000	0.3333333	1.1250000	3
[17]	{고기류,주스}	=>	{탄산음료}	0.2000000	0.4285714	0.4666667	0.9183673	3
[18]	{맥주,탄산음료}	=>	{고기류}	0.2666667	0.8000000	0.3333333	1.0909091	4
[19]	{고기류,탄산음료}	=>	{맥주}	0.2666667	0.8000000	0.3333333	1.3333333	4
[20]	{고기류,맥주}	=>	{탄산음료}	0.2666667	0.5714286	0.4666667	1.2244898	4
[21]	{맥주,주스}	=>	{고기류}	0.3333333	0.8333333	0.4000000	1.1363636	5
[22]	{고기류,주스}	=>	{맥주}	0.3333333	0.7142857	0.4666667	1.1904762	5
[23]	{고기류,맥주}	=>	{주스}	0.3333333	0.7142857	0.4666667	1.3392857	5
[24]	{맥주,주스,탄산음료}	=>	{고기류}	0.2000000	1.0000000	0.2000000	1.3636364	3
[25]	{고기류,주스,탄산음료}	=>	{맥주}	0.2000000	1.0000000	0.2000000	1.6666667	3
[26]	{고기류,맥주,탄산음료}	=>	{주스}	0.2000000	0.7500000	0.2666667	1.4062500	3
[27]	{고기류,맥주,주스}	=>	{탄산음료}	0.2000000	0.6000000	0.3333333	1.2857143	3

고기와 묶이는것 대부분의 신뢰도가 높게 나타난다. 예상밖으로 [1]~[11]에서 고기 → 고기와 함께 먹는 음식 의 신뢰도 보다 고기와 함께 먹는음식 → 고기의 신뢰도가 높게 나타났다. 이는

고기 → 주스, 고기→맥주의 신뢰도가 높게 나타나고 향상도 또한 1보다 크기 때문에 매출증진을 위해서는 고기를 파는 매장 근처에 식음료 매장을 배치하면 좋을 것이다.

채소나 곡류를 사는사람 대부분은 고기류를 사기 때문에 위치 선정시 고기 코너보다 채소나 곡류 코너를 먼저 배치하여 이벤트를 열어곡류와 채소를 사게한다면 고기를 살 가능성이 높

다고 판단된다. 고기류 곡류류 ⇒ 채소류 와 고기류 채소류 ⇒ 곡류의 향상도가 3과 2.8로 굉장히 높기 때문에 실제 효용가치가 높다고 할 수 있다.

*나머지 품목은 지지도가 낮아 대표성을 확보하기 힘들기 때문에 배제하였다.

2-3) 고객과의 관계 전략을 위하여, 매출증진을 위한 어떤 이벤트를 어느 기간에 진행할 계획인지와 그 근거를 제시하시오.

```
[42] {고기류}          => {채소류}   0.2666667 0.3636364  0.7333333 1.0909091  4
[43] {}                 => {채소류}   0.3333333 0.3333333  1.0000000 1.0000000  5
[44] {고기류}          => {곡류}     0.2000000 0.2727273  0.7333333 1.0227273  3
```

고기류의 빈도는 높지만 채소나 곡류의 빈도는 높다고 할 수 없다. 곡류나 채소 → 고기의 신뢰도는 높지만 반대의 신뢰도는 낮기 때문에 고기를 사면 채소나, 곡물을 할인이나 덤으로 주는 방식으로 매출증진을 유도 해야한다.

```
[4]  {곡류}          => {채소류} 0.2666667 1.0000000  0.2666667 3.000000 4
[5]  {채소류}        => {곡류}   0.2666667 0.8000000  0.3333333 3.000000 4
```

채소류를 사는 사람 대부분은 곡물을 사기 때문에 고기 이벤트 한번으로 곡류와 채소류 두 품목의 매출 증진을 유도할 수 있다고 판단된다. (신뢰도도 높고 향상도도 높음)

3. 매출전략이 성공하여 최우헌 매니저는 1년 2만엔 연봉 + 성과급 4만엔을 받아서, 본인의 연봉 2배의 초과 급여 받았다는 소문이 인근 도시인 베이징 HL-마트 점장에게 알려졌다. 베이징 본사 점장은 최우헌 매니저가 빅데이터 전문가 자격증 ADsP를 소지하였다는 것을 확인한 후에, 본격적으로 본점의 매출증가를 위한 컨설팅을 부탁하였다. 관건은, 매장진열을 어떻게 하는 것이 좋을까에 있다. 예를 들어, 베이징은 해안과 멀기 때문에 수산물은 마트에서 아무리 잘 관리를 한다 해도 신선도를 유지하기 어렵다는 점을 알고 있기 때문에 고기류의 판매전략을 높혀서 전체 매출을 증가시키고 싶은 점에 목적이 있었다. 만일 베이징 구매고객의 구매품목이 위의 2)에서 제시한 항목과 같다면, 베이징 점장이 강조하고 싶은 고기류를 어느 품목과 연계하여 진열하면 좋을지를 최우헌 빅데이터 전문가에게 컨설팅을 했다고 한다.

3-1) 연관규칙 A→B의 의미를 해석하시오.

A와B의 지지도 $A \rightarrow B = \frac{A와B를포함한거래수}{전체거래수}$

전체 거래수 = 전체 표본공간의 확률

A와B를 포함한 거래수 = P(AnB)

결국 P(AnB) → A와B의 교집합의 확률

지지도가 낮으면 대표성이 떨어져 제외시킬때 사용하면 좋음

A와B의 신뢰도$A \rightarrow B = \frac{A와B를포함한거래수}{A를포함한거래수}$

A조건하에서 B가 발생할 확률 (조건부 확률)

신뢰도가 높지만 지지도가 낮으면 표본의 대표성이 떨어져 해석할 때 주의해야함

A와B의 향상도$A \rightarrow B = \frac{A \rightarrow B의신뢰도}{B를포함한거래비율} = \frac{P(B|A)}{P(B)}$

A와B가 독립이면 향상도가 1이 나옴

A→B B→A 의 향상도는 같다

향상도가 1보다 크면 의미가 있다고 해석할 수 있음

★ 연관규칙을 할때, 세개의 지표를 고려하고, 표본의 대표성을 고려해야함 →(해석시 각 연관 규칙별 정의을 고려해야함)

3-2) 만일 베이징 점장이 부탁한 고기류 판매를 증진하기 위해 연관된 품목을 제시하려면, 맥주, 고량주, 곡류, 면류, 채소류 중에 어떤 품목을 제안하겠는가? 그 근거와 이유는?

고기류의 판매전략을 높인다고 하였으니 2번에 서술한 내용과 비슷하다.

(지지도 0.1 이상, 신뢰도 0.25 이상, A → 고기)

```
> inspect(rules)
     lhs                    rhs      support   confidence coverage  lift      count
[1]  {}                  => {고기류} 0.7333333 0.7333333  1.0000000 1.0000000 11
[2]  {게임기}            => {고기류} 0.1333333 1.0000000  0.1333333 1.3636364  2
[3]  {면류}              => {고기류} 0.1333333 0.6666667  0.2000000 0.9090909  2
[4]  {고량주}            => {고기류} 0.1333333 0.6666667  0.2000000 0.9090909  2
[5]  {곡류}              => {고기류} 0.2000000 0.7500000  0.2666667 1.0227273  3
[6]  {채소류}            => {고기류} 0.2666667 0.8000000  0.3333333 1.0909091  4
[7]  {탄산음료}          => {고기류} 0.3333333 0.7142857  0.4666667 0.9740260  5
[8]  {주스}              => {고기류} 0.4666667 0.8750000  0.5333333 1.1931818  7
[9]  {맥주}              => {고기류} 0.4666667 0.7777778  0.6000000 1.0606061  7
[10] {게임기,주스}       => {고기류} 0.1333333 1.0000000  0.1333333 1.3636364  2
[11] {게임기,맥주}       => {고기류} 0.1333333 1.0000000  0.1333333 1.3636364  2
[12] {고량주,채소류}     => {고기류} 0.1333333 0.6666667  0.2000000 0.9090909  2
[13] {곡류,채소류}       => {고기류} 0.2000000 0.7500000  0.2666667 1.0227273  3
[14] {곡류,주스}         => {고기류} 0.1333333 1.0000000  0.1333333 1.3636364  2
[15] {주스,채소류}       => {고기류} 0.1333333 1.0000000  0.1333333 1.3636364  2
[16] {주스,탄산음료}     => {고기류} 0.2000000 1.0000000  0.2000000 1.3636364  3
[17] {맥주,탄산음료}     => {고기류} 0.2666667 0.8000000  0.3333333 1.0909091  4
[18] {맥주,주스}         => {고기류} 0.3333333 0.8333333  0.4000000 1.1363636  5
[19] {게임기,맥주,주스}  => {고기류} 0.1333333 1.0000000  0.1333333 1.3636364  2
[20] {곡류,주스,채소류}  => {고기류} 0.1333333 1.0000000  0.1333333 1.3636364  2
[21] {맥주,주스,탄산음료} => {고기류} 0.2000000 1.0000000  0.2000000 1.3636364  3
```

```
[4]  {곡류}              => {채소류} 0.2666667 1.0000000  0.2666667 3.000000 4
[5]  {채소류}            => {곡류}   0.2666667 0.8000000  0.3333333 3.000000 4
```

고기류 판매를 증진하기 위해서는 A → 고기의 연관 규칙을 봐야한다. (어떤것을 사는 사람이 고기류를 사는가)

고기 → B의 연관규칙 대부분은 신뢰도와 향상도가 높기 때문에 고기 매출의 증가와 관련이 없다고 생각한다.

따라서 A → 고기의 연관규칙에서 찾아야하는데, 지지도가 0.1 미만인 것은 아무리 신뢰도가 높아도 지지도가 낮아 표본의 대표성이 부족하기 때문에 제외 시켰다.

연관규칙을 살펴보면 주스 → 고기류와 맥주 → 고기 의 신뢰도가 높지만 반대의 경우의 신뢰도도 높기 때문에 이미 충분하다고 생각한다. 하지만 채소류와 곡류는 채소류 → 고기, 곡류 → 고기의 신뢰도는 높지만 반대의 경우 낮기 때문에 채소와 곡류를 고기류 판매 증진을 위한 연관품목으로 제안할 것이다. 또한 채소와 곡류간의 신뢰도와 향상도가 높기 때문에, 둘중 하나만 고기와 연관품목으로 설정해도 고기+채소+곡류+식음료의 매출이 한번에 오를 것으로 판단 된다.

3-3) 만일 연관규칙 고기류 → 주스의 지지도가 높다면, 주스가 고기류의 매출을 증가시키는 후보 품목으로 채택될 수 있는가? 없다면, 왜 그런지, 있다면, 왜 그런지 이유를 쓰시오.

```
[9]  {고기류} => {주스}      0.4666667 0.6363636  0.7333333 1.193182 7
```

고기류와 주스간의 지지도는 0.47로 높기 때문에 주스가 고기류의 매출을 증가시키는 후보 품목으로 채택 될 수 있다. 지지도가 낮다면 대표성이 떨어져 매출증가 후보품목에 포함시킬 수 없지만 높다면 매출을 증가시키는 후보로 채택 될 수 있다. 지지도의 경우 A → B의 지지도와 B → A의 지지도가 같기 때문에 증가시킨다고 할 수있다.

전체 고객의 47%가 고기류를 살때 주스를 사고 고기를 사는 고객 63%가 주스를 사기 때문에 충분히 주스가 고기류의 매출을 증가 시킨다고 판단 할 수 있다. 향상도 또한 1.19로 1보다 크기 때문에 이벤트나 계절성을 탄다고 할 수 없다.

빅데이터마이닝:패턴탐색

데이터테크전공 20173204 곽명빈

2021 10 19

```
library(knitr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tidyr)
```

5. 2주차 실습데이터인 "REM_고객예측_의사결정_데이터"를 이용하여 지역1 :
서울특별시의 행을 뽑아 새벽시간대의 평균 지출비용(가격*구입갯수)을 구하시오.

1. 중복행을 제거하여 진행
2. 지출비용 = 구입갯수 * 가격
3. 시간대 하루를 4그룹으로 나누어서 진행
   (새벽 : 0시~6시, 아침 : 6시~12시,
    오후 : 12시~18시, 저녁 : 18~24시)
4. 지역1 칼럼에서 서울특별시 행만을 뽑아서 진행

```
# 데이터 읽기
df <- read.csv('REM_고객예측_의사결정_데이터_중간고사.csv')

# 1) 중복행을 제거하여 진행
dat <- unique(df)
str(dat)
```

```
## 'data.frame':    812 obs. of  10 variables:
##  $ 고객코드: int  1293 1302 1813 3573 3573 3573 3714 3714 4804 4804 ...
##  $ 구입날짜: chr  "2010-01-26" "2009-11-18" "2010-01-17" "2010-03-16" ...
##  $ 구입시간: int  1259 2219 2212 1651 1908 1908 1347 1433 1456 1617 ...
##  $ 구입갯수: int  1 1 1 1 1 1 1 1 10 5 ...
##  $ 가격    : int  27500 32000 24500 39000 48000 15000 23900 23900 19800 19800 ...
##  $ 나이    : int  44 39 38 40 40 40 39 39 40 40 ...
##  $ 성별    : chr  "남성" "여성" "남성" "남성" ...
##  $ 지역1   : chr  "전북" "경기도" "경기" "서울" ...
##  $ 지역2   : chr  "군산시" "김포시" "성남시" "강남구" ...
##  $ 상품분류: chr  "컴퓨터/주변기기/게임" "화장품/이미용" "남성의류" "컴퓨터/주변기기/게임"
...
```

```
# 2) 지출비용 = 구입갯수 * 가격

dat['지출비용'] = dat$구입갯수*dat$가격
str(dat)
```

```
## 'data.frame':    812 obs. of  11 variables:
##  $ 고객코드: int  1293 1302 1813 3573 3573 3573 3714 3714 4804 4804 ...
##  $ 구입날짜: chr  "2010-01-26" "2009-11-18" "2010-01-17" "2010-03-16" ...
##  $ 구입시간: int  1259 2219 2212 1651 1908 1908 1347 1433 1456 1617 ...
##  $ 구입갯수: int  1 1 1 1 1 1 1 1 10 5 ...
##  $ 가격    : int  27500 32000 24500 39000 48000 15000 23900 23900 19800 19800 ...
##  $ 나이    : int  44 39 38 40 40 40 39 39 40 40 ...
##  $ 성별    : chr  "남성" "여성" "남성" "남성" ...
##  $ 지역1   : chr  "전북" "경기도" "경기" "서울" ...
##  $ 지역2   : chr  "군산시" "김포시" "성남시" "강남구" ...
##  $ 상품분류: chr  "컴퓨터/주변기기/게임" "화장품/이미용" "남성의류" "컴퓨터/주변기기/게임"
...
##  $ 지출비용: int  27500 32000 24500 39000 48000 15000 23900 23900 198000 99000 ...
```

```
# 3) 시간대를 하루 4그룹으로 나누어서 진행
dat$group_time <- ifelse(dat$구입시간 <= 600, "새벽",
                         ifelse(dat$구입시간 <= 1200, "아침",
                                ifelse(dat$구입시간 <= 1800, "오후", "저녁")))

# 4) 지역 1 칼럼에서 서울특별시 행만을 뽑아서 진행

seoul <- filter(dat, 지역1 == '서울특별시')

# 5) 새벽시간대의  평균 지출 비용

summary(seoul[seoul$group_time == '새벽',])  # summary 이용
```

```
##     고객코드         구입날짜            구입시간         구입갯수
##  Min.   : 16771   Length:28          Min.   :  3.00   Min.   :1.000
##  1st Qu.: 36423   Class :character   1st Qu.: 16.25   1st Qu.:1.000
##  Median : 92371   Mode  :character   Median : 83.50   Median :1.000
##  Mean   : 86647                      Mean   :122.36   Mean   :1.036
##  3rd Qu.:125595                      3rd Qu.:217.25   3rd Qu.:1.000
##  Max.   :145105                      Max.   :457.00   Max.   :2.000
##       가격             나이           성별              지역1
##  Min.   :  9900   Min.   :28.00   Length:28          Length:28
##  1st Qu.: 24975   1st Qu.:33.75   Class :character   Class :character
##  Median : 35650   Median :35.00   Mode  :character   Mode  :character
##  Mean   : 47221   Mean   :35.29
##  3rd Qu.: 43250   3rd Qu.:36.50
##  Max.   :230400   Max.   :41.00
##     지역2             상품분류            지출비용       group_time
##  Length:28          Length:28          Min.   :  9900   Length:28
##  Class :character   Class :character   1st Qu.: 24975   Class :character
##  Mode  :character   Mode  :character   Median : 35650   Mode  :character
##                                        Mean   : 47575
##                                        3rd Qu.: 43250
##                                        Max.   :230400
```

```
aggregate(지출비용 ~ group_time ,seoul, mean) # aggregate
```

```
##   group_time 지출비용
## 1       새벽 47575.00
## 2       아침 84100.00
## 3       오후 49818.12
## 4       저녁 54357.23
```

```
## 서울지역 새벽시간대의 평균 지출비용 = 47575
```

6. R의 내장데이터 “iris” 데이터를 이용하여 상관관계가 가장 낮은 두 칼럼을 뽑아 상관계수를 구하고, 이상치를 제거한 후에 상관계수를 다시 구하여라.

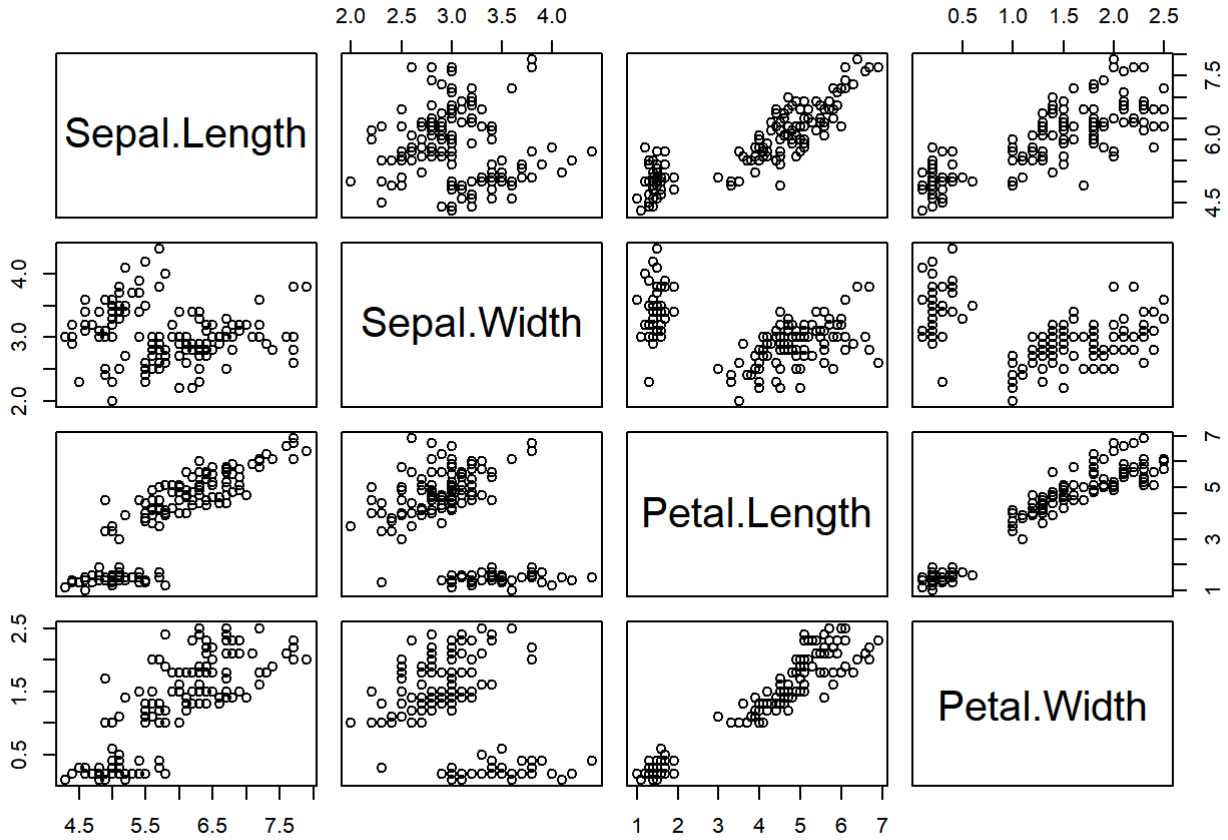
```
library(corrplot)
```

```
## corrplot 0.90 loaded
```

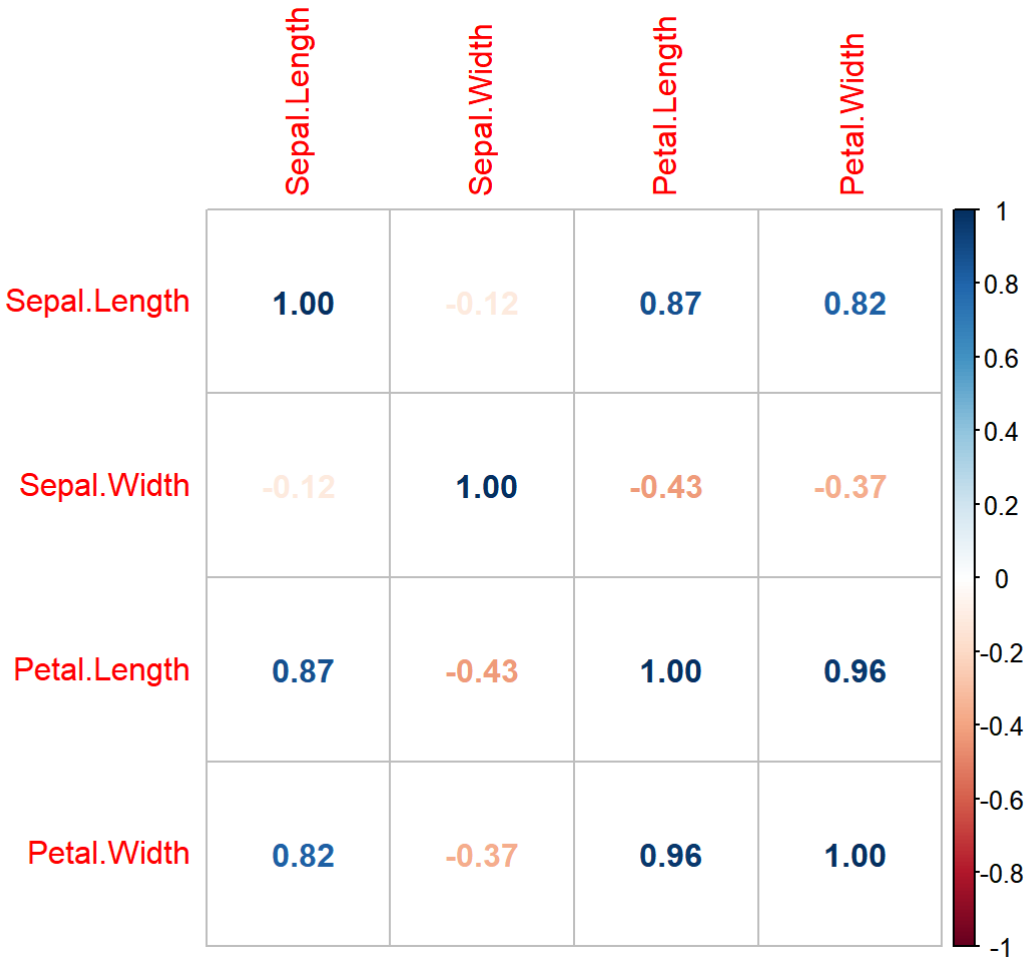
```
library(ggplot2)

# iris 데이터 불러오기

df <- iris

str(df)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

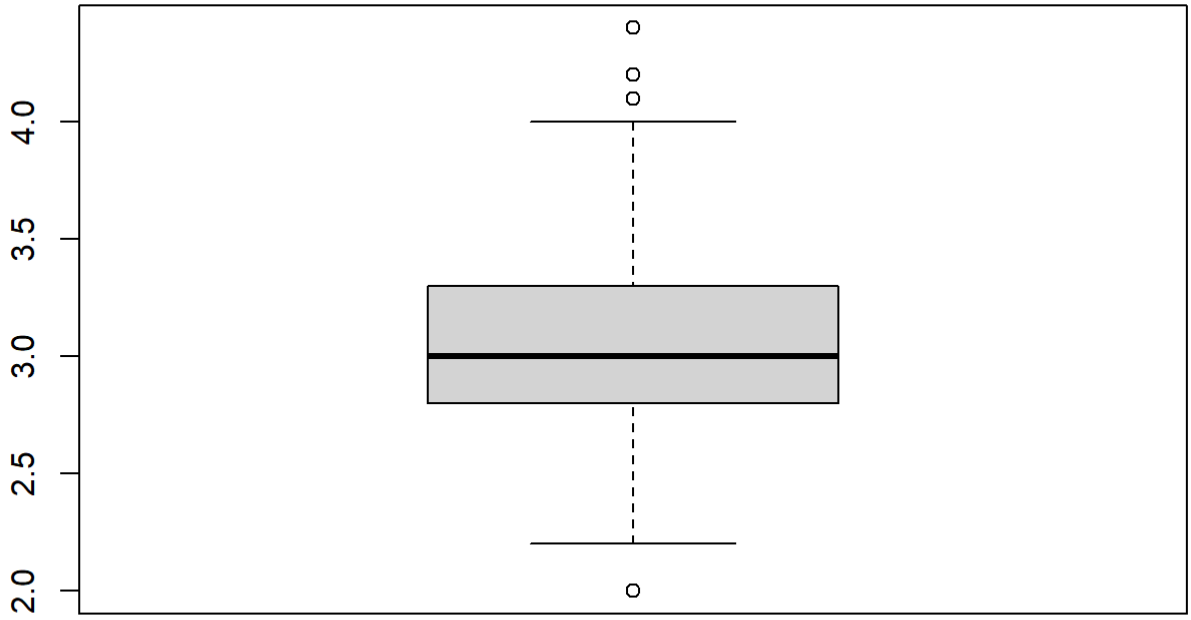
```
# 데이터 분포
plot(df[,1:4])
```



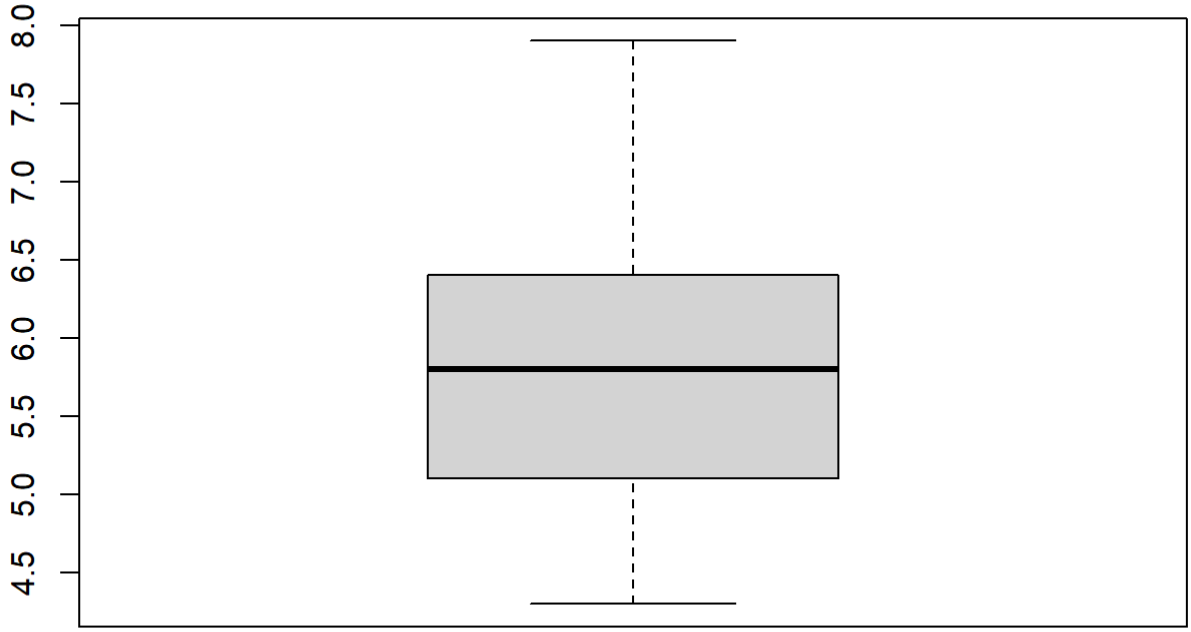

```
# 상관계수 행렬
iris_cor <- cor(df[,1:4])

corrplot(iris_cor, method="number")
```



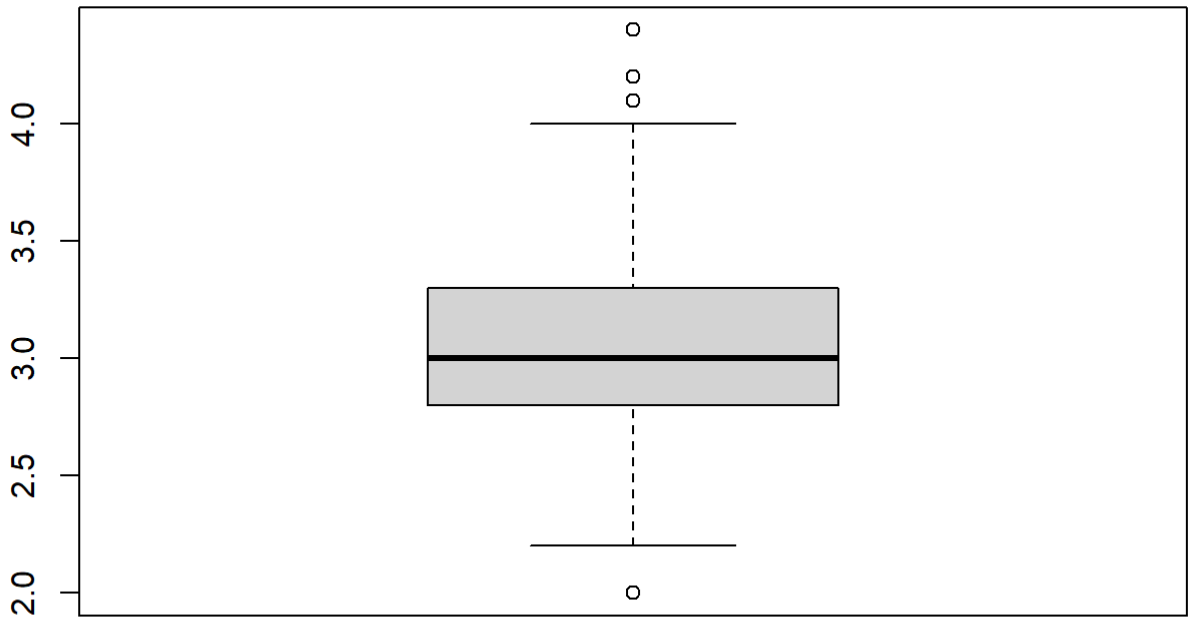

```
# Sepal.Width 와 Sepal.Length 가 -0.12로 제일 낮은 것을 확인

# 이상치 확인을 위해 Boxplot 이용
boxplot(df$Sepal.Width)
```




```
boxplot(df$Sepal.Length)
```

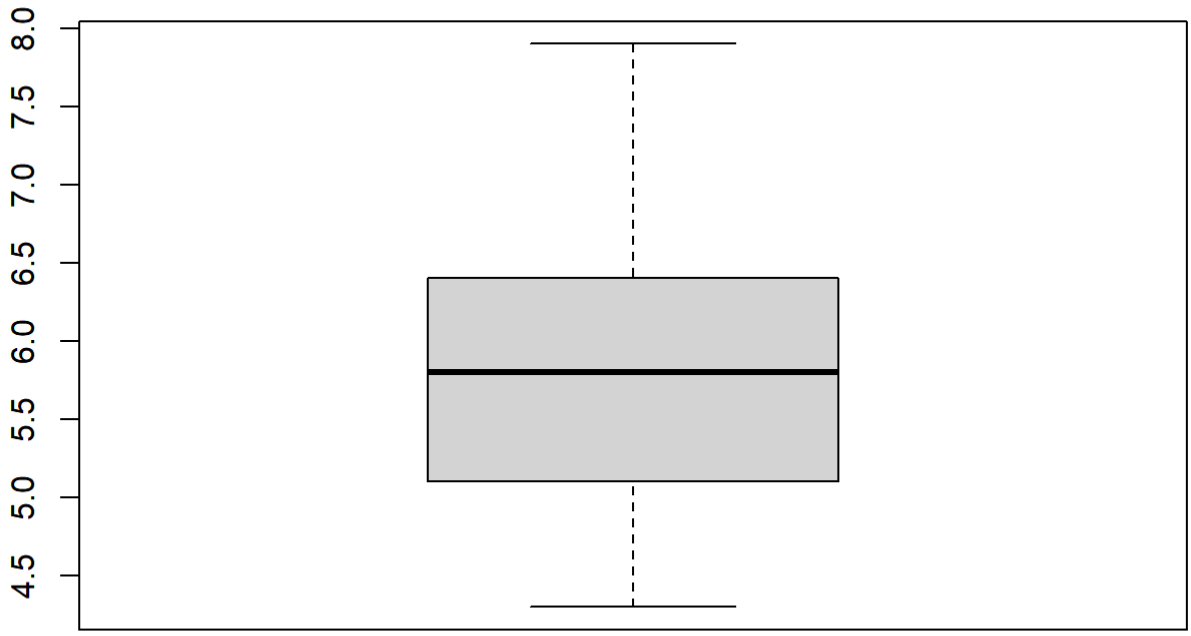


```
boxplot(df$Sepal.Width)$stat  # 확인결과 2.2보다 낮고 4보다 높으면 이상치로 판단
```




```
##      [,1]
## [1,]  2.2
## [2,]  2.8
## [3,]  3.0
## [4,]  3.3
## [5,]  4.0
```

```
boxplot(df$Sepal.Length)$stat # 확인결과 4.3보다 낮고 7.9보다 높으면 이상치로 판단 -> 없음
```




```
##      [,1]
## [1,]  4.3
## [2,]  5.1
## [3,]  5.8
## [4,]  6.4
## [5,]  7.9
```

```
df$Sepal.Width<- ifelse(df$Sepal.Width < 2.2 | df$Sepal.Width > 4, NA, df$Sepal.Width) # 이상치를 Na로 바꾼후 제거

# 4개의 이상치를 제거
sum(is.na(df))
```

```
## [1] 4
```

```
na <- df %>% filter(!is.na(df$Sepal.Width))

# 이상치 제거후 상관계수 -0.1231441
cor(na$Sepal.Length, na$Sepal.Width)
```

```
## [1] -0.1231441
```

7. R의 내장데이터 "Groceries" 데이터를 활용해서 {yogurt, margarine, waffles} -> {whole milk}의 지지도, 신뢰도, 향상도를 구하여라.

```
# 식료품데이터
library(arules)
```

```
## 필요한 패키지를 로딩중입니다: Matrix
```

```
##
## 다음의 패키지를 부착합니다: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
##
## 다음의 패키지를 부착합니다: 'arules'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following objects are masked from 'package:base':
##
##     abbreviate, write
```

```
data("Groceries")

dat <- Groceries

summary(dat)
```

```
## transactions as itemMatrix in sparse format with
##  9835 rows (elements/itemsets/transactions) and
##  169 columns (items) and a density of 0.02609146
##
## most frequent items:
##       whole milk other vegetables       rolls/buns             soda
##             2513             1903             1809             1715
##           yogurt          (Other)
##             1372            34055
##
## element (itemset/transaction) length distribution:
## sizes
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
## 2159 1643 1299 1005  855  645  545  438  350  246  182  117   78   77   55   46
##   17   18   19   20   21   22   23   24   26   27   28   29   32
##   29   14   14    9   11    4    6    1    1    1    1    3    1
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   4.409   6.000  32.000
##
## includes extended item information - examples:
##        labels  level2           level1
## 1 frankfurter sausage meat and sausage
## 2     sausage sausage meat and sausage
## 3  liver loaf sausage meat and sausage
```

```
inspect(dat[1:5])
```

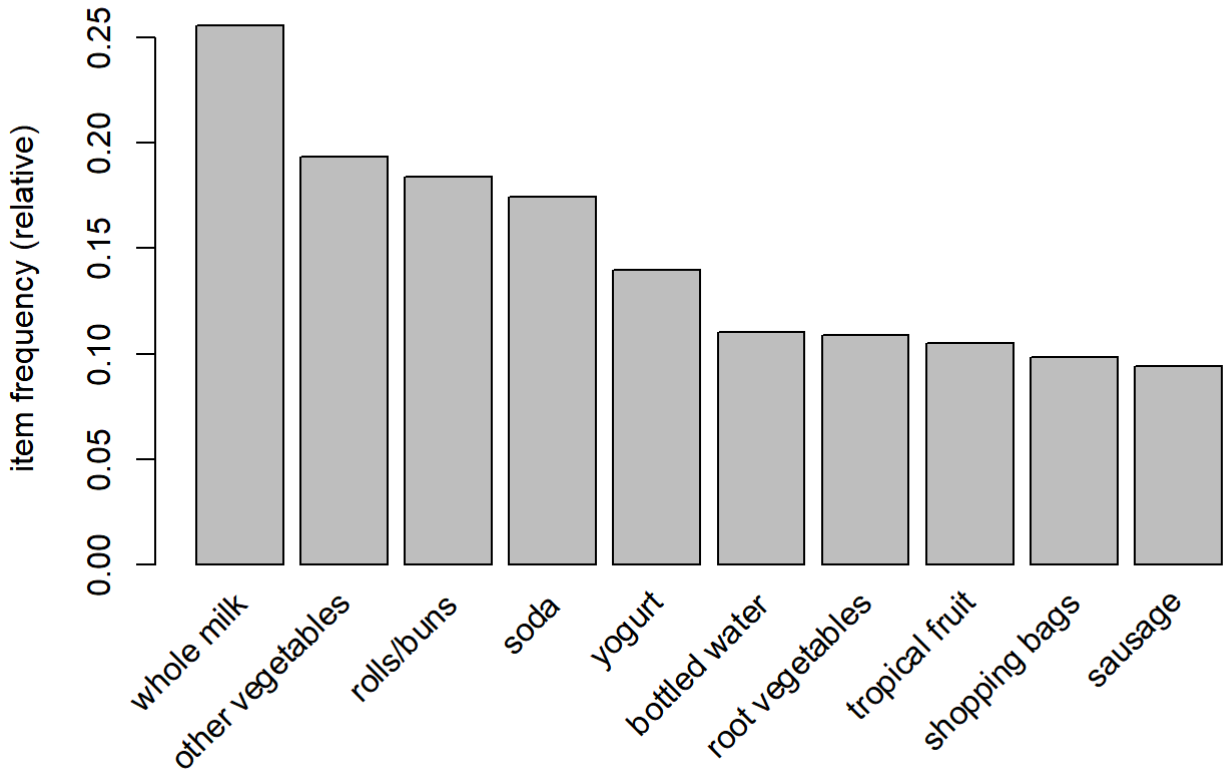
```
##     items
## [1] {citrus fruit,
##      semi-finished bread,
##      margarine,
##      ready soups}
## [2] {tropical fruit,
##      yogurt,
##      coffee}
## [3] {whole milk}
## [4] {pip fruit,
##      yogurt,
##      cream cheese ,
##      meat spreads}
## [5] {other vegetables,
##      whole milk,
##      condensed milk,
##      long life bakery product}
```

```
# itemFrequency
itemFrequency(dat, type = "absolute") # 빈도수
```

```
##              frankfurter                   sausage               liver loaf
##                      580                       924                       50
##                      ham                      meat         finished products
##                      256                       254                       64
##           organic sausage                   chicken                   turkey
##                       22                       422                       80
##                     pork                      beef            hamburger meat
##                      567                       516                      327
##                     fish              citrus fruit            tropical fruit
##                       29                       814                     1032
##                pip fruit                    grapes                   berries
##                      744                       220                      327
##              nuts/prunes            root vegetables                   onions
##                       33                      1072                      305
##                    herbs           other vegetables packaged fruit/vegetables
##                      160                      1903                      128
##               whole milk                    butter                      curd
##                     2513                       545                      524
##                   dessert                butter milk                    yogurt
##                      365                       275                     1372
##         whipped/sour cream                  beverages                  UHT-milk
##                      705                       256                      329
##            condensed milk                     cream                soft cheese
##                      101                        13                      168
##             sliced cheese                hard cheese              cream cheese
##                      241                       241                      390
##           processed cheese              spread cheese                curd cheese
##                      163                       110                       50
##           specialty cheese                mayonnaise            salad dressing
##                       84                        90                        8
##                   tidbits          frozen vegetables              frozen fruits
##                       23                       473                       12
##              frozen meals                frozen fish            frozen chicken
##                      279                       115                        6
##                 ice cream             frozen dessert    frozen potato products
##                      246                       106                       83
##            domestic eggs                 rolls/buns               white bread
##                      624                      1809                      414
##              brown bread                    pastry            roll products
##                      638                       875                      101
##        semi-finished bread                  zwieback           potato products
##                      174                        68                       28
##                    flour                      salt                      rice
##                      171                       106                       75
##                    pasta                   vinegar                       oil
##                      148                        64                      276
##                 margarine              specialty fat                    sugar
##                      576                        36                      333
##           artif. sweetener                     honey                  mustard
##                       32                        15                      118
##                  ketchup                    spices                    soups
##                       42                        51                       67
##              ready soups      Instant food products                   sauces
##                       18                        79                       54
##                  cereals           organic products             baking powder
##                       56                        16                      174
##       preservation products             pudding powder          canned vegetables
```

```
##                         2                        23                       106
##              canned fruit         pickled vegetables       specialty vegetables
##                        32                       176                        17
##                       jam             sweet spreads              meat spreads
##                        53                        89                        42
##               canned fish                  dog food                  cat food
##                       148                        84                       229
##                  pet care                 baby food                    coffee
##                        93                         1                       571
##             instant coffee                       tea              cocoa drinks
##                        73                        38                        22
##              bottled water                      soda            misc. beverages
##                      1087                      1715                       279
##      fruit/vegetable juice                     syrup              bottled beer
##                       711                        32                       792
##               canned beer                    brandy                     whisky
##                       764                        41                         8
##                    liquor                       rum                    liqueur
##                       109                        44                         9
##         liquor (appetizer)               white wine            red/blush wine
##                        78                       187                       189
##                  prosecco            sparkling wine               salty snack
##                        20                        55                       372
##                   popcorn               nut snack            snack products
##                        71                        31                        30
##  long life bakery product                    waffles                  cake bar
##                       368                       378                       130
##               chewing gum                 chocolate         cooking chocolate
##                       207                       488                        25
##       specialty chocolate              specialty bar      chocolate marshmallow
##                       299                       269                        89
##                    candy             seasonal products                 detergent
##                       294                       140                       189
##                  softener               decalcifier              dish cleaner
##                        54                        15                       103
##           abrasive cleaner                   cleaner             toilet cleaner
##                        35                        50                         7
##           bathroom cleaner                hair spray                dental care
##                        27                        11                        57
##             male cosmetics           make up remover                  skin care
##                        45                         8                        35
##  female sanitary products            baby cosmetics                      soap
##                        60                         6                        26
##            rubbing alcohol          hygiene articles                   napkins
##                        10                       324                       515
##                   dishes                  cookware            kitchen utensil
##                       173                        27                         4
##           cling film/bags             kitchen towels     house keeping products
##                       112                        59                        82
##                  candles                light bulbs       sound storage medium
##                        88                        41                         1
##               newspapers                 photo/film                 pot plants
##                       785                        91                       170
##     flower soil/fertilizer           flower (seeds)              shopping bags
##                        19                       102                       969
##                     bags
##                         4
```

```
itemFrequencyPlot(dat, topN = 10)  # 상대도수 그래프
```




```
# 연관규칙 (지지도0.005 신뢰도0.25이상인 것만 추출)
rules <- apriori(dat, parameter = list(supp = 0.005, conf = 0.25))
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##        0.25    0.1    1 none FALSE            TRUE       5   0.005      1
##  maxlen target  ext
##      10  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 49
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
## sorting and recoding items ... [120 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 done [0.00s].
## writing ... [663 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
rules # 663개
```

```
## set of 663 rules
```

```
# "yogurt" ,"margarine", "waffles"가 포함된 연관규칙

rules_3 <- subset(rules, lhs %ain% c("yogurt" ,"margarine", "waffles"))

rules_3 # 0개
```

```
## set of 0 rules
```

```
# 연관규칙 (지지도0.0005 신뢰도0.25이상인 것만 추출)
rules <- apriori(dat, parameter = list(supp = 0.0005, conf = 0.25))
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##        0.25    0.1    1 none FALSE            TRUE       5   5e-04      1
##  maxlen target  ext
##      10  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 4
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[169 item(s), 9835 transaction(s)] done [0.01s].
## sorting and recoding items ... [164 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 5 6 7 done [0.05s].
## writing ... [95038 rule(s)] done [0.03s].
## creating S4 object  ... done [0.03s].
```

```
rules # 95038개
```

```
## set of 95038 rules
```

```
# "yogurt" ,"margarine", "waffles"가 포함된 연관규칙

rules_3 <- subset(rules, lhs %ain% c("yogurt" ,"margarine", "waffles"))

rules_3 # 12개
```

```
## set of 12 rules
```

```
inspect(rules_3) # 8번째에 있는것을 확인
```

```
##      lhs                          rhs                          support confidence     coverage
lift count
## [1]  {yogurt,
##       margarine,
##       waffles}                 => {chocolate marshmallow} 0.0006100661  0.4615385 0.0013218099
51.002593     6
## [2]  {yogurt,
##       margarine,
##       waffles}                 => {beef}                  0.0005083884  0.3846154 0.0013218099
7.330799     5
## [3]  {yogurt,
##       margarine,
##       waffles}                 => {frankfurter}           0.0005083884  0.3846154 0.0013218099
6.521883     5
## [4]  {yogurt,
##       margarine,
##       waffles}                 => {whipped/sour cream}    0.0005083884  0.3846154 0.0013218099
5.365521     5
## [5]  {yogurt,
##       margarine,
##       waffles}                 => {tropical fruit}        0.0007117438  0.5384615 0.0013218099
5.131559     7
## [6]  {yogurt,
##       margarine,
##       waffles}                 => {rolls/buns}            0.0006100661  0.4615385 0.0013218099
2.509249     6
## [7]  {yogurt,
##       margarine,
##       waffles}                 => {other vegetables}      0.0007117438  0.5384615 0.0013218099
2.782853     7
## [8]  {yogurt,
##       margarine,
##       waffles}                 => {whole milk}            0.0008134215  0.6153846 0.0013218099
2.408399     8
## [9]  {yogurt,
##       margarine,
##       waffles,
##       chocolate marshmallow} => {other vegetables}      0.0005083884  0.8333333 0.0006100661
4.306796     5
## [10] {other vegetables,
##       yogurt,
##       margarine,
##       waffles}                 => {chocolate marshmallow} 0.0005083884  0.7142857 0.0007117438
78.932584     5
## [11] {beef,
##       yogurt,
##       margarine,
##       waffles}                 => {whole milk}            0.0005083884  1.0000000 0.0005083884
3.913649     5
## [12] {whole milk,
##       yogurt,
##       margarine,
##       waffles}                 => {beef}                  0.0005083884  0.6250000 0.0008134215
11.912548     5
```

```
inspect(rules_3[8])
```

```
##     lhs                            rhs          support      confidence
## [1] {yogurt,margarine,waffles} => {whole milk} 0.0008134215 0.6153846
##     coverage   lift     count
## [1] 0.00132181 2.408399 8
```

```
#    lhs                            rhs          support   confidence coverage   lift     count
# {yogurt,margarine,waffles} => {whole milk} 0.0008134215 0.6153846  0.00132181 2.408399 8
# 지지도:0.0008134215 신뢰도:0.6153846 향상도:2.408399
```