R (퀴즈-통계)

> **A. 아래의 url의 데이터는 2018년 전국의 소매점에서 쌀과 찹쌀의 소비자가격을 각각 조사한 자료이다.**
> 여기서 ID는 소매점 구분번호이다.\
> http://datamining.dongguk.ac.kr/data/rice_price_survey.csv

1. 위 자료를 읽어 read.csv 파일을 이용해 R데이터프레임 `rice_price` 로 저장하라.

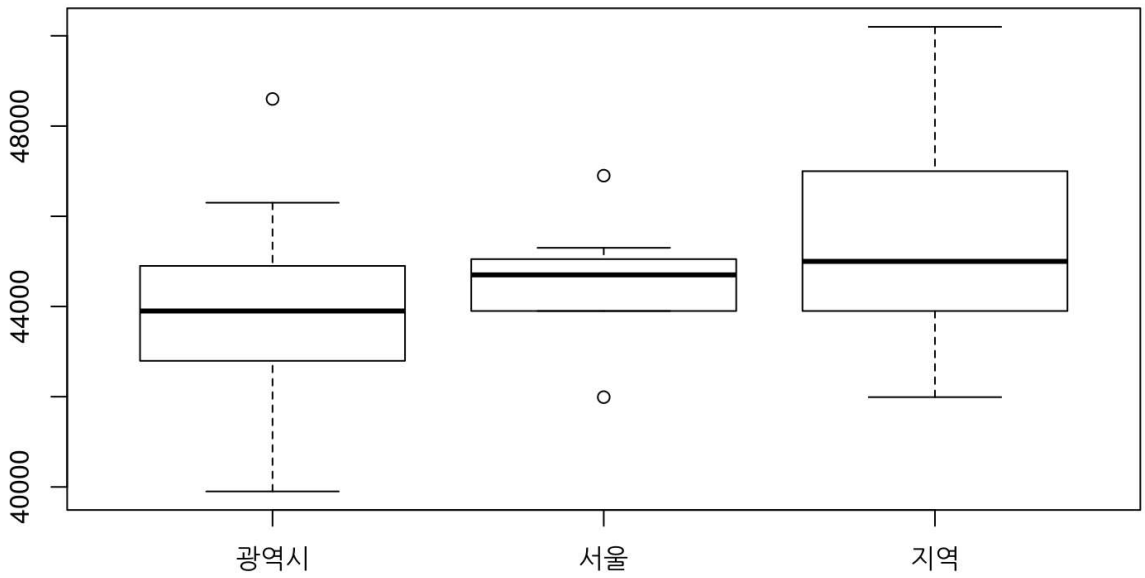
```
##   ID 품목 당일가격 전일가격 조사지역명
## 1  1   쌀    44600    44600       서울
## 2  2   쌀    46900    46900       서울
## 3  3   쌀    45300    45300       서울
## 4  4   쌀    44800    44800       서울
## 5  5   쌀    43900    43900       서울
## 6  6   쌀    41990    41990       서울
## 7  7   쌀    43900    43900       서울
## 8  8   쌀    44800    44800       서울
```

2. 자료를 품목별로 2개로 나누어라. 즉, `쌀` 은 `rice_price_1` , `찹쌀` `rice_price_2` .

```
##   ID 품목 당일가격 전일가격 조사지역명
## 1  1   쌀    44600    44600       서울
## 2  2   쌀    46900    46900       서울
## 3  3   쌀    45300    45300       서울
## 4  4   쌀    44800    44800       서울
## 5  5   쌀    43900    43900       서울
## 6  6   쌀    41990    41990       서울

##    ID 품목 당일가격 전일가격 조사지역명
## 46  1 찹쌀    45300    45300       서울
## 47  2 찹쌀    47400    47400       서울
## 48  3 찹쌀    46000    46000       서울
## 49  4 찹쌀    46800    46800       서울
## 50  5 찹쌀    44900    44900       서울
## 51  6 찹쌀    46990    46990       서울
```

3. `rice_price_1` 데이터에서 지역구분 별 `쌀` 의 당일가격을 상자그림으로 나타내어라.



{width="800"}

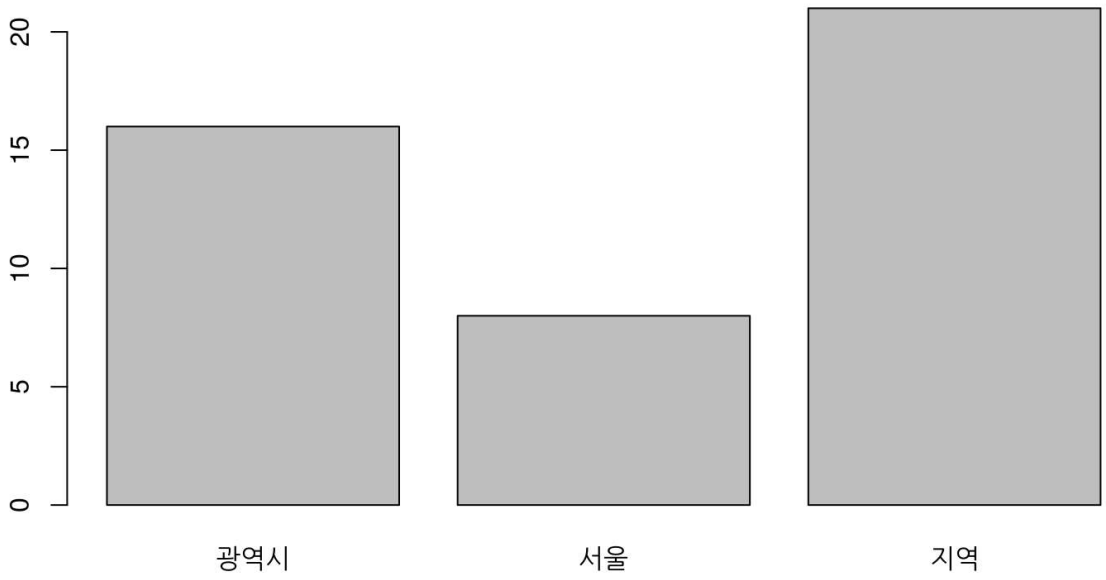
4. `rice_price_1` 데이터에서 지역구분 별 `쌀` 의 `당일가격` 평균과 중앙값, 표준편차를 구하라.

```
##   조사지역명 당일가격.평균 당일가격.중앙값 당일가격.표준편차
## 1     광역시      43973.125        43900.000         1980.719
## 2       서울      44523.750        44700.000         1394.150
## 3       지역      45188.095        45000.000         2548.701
```

5. `rice_price_1` 데이터에서 `조사지역명` 별 조사대상 소매점의 수를 구하라.

```
##
## 광역시   서울   지역
##     16      8     21
```

6. `조사지역명` 별 조사대상 소매점의 수를 막대그래프로 나타내어라.



{width="800"}

7. `rice_price_1` 데이터에서 `조사지역명` 을 `서울+광역시` 와 `지역` 두 개로 나누어 새로운 변수 `조사지역명2` 에 추가하여라.

```
##    ID 품목 당일가격 전일가격 조사지역명 조사지역명2
## 1   1  쌀    44600    44600       서울 서울+광역시
## 2   2  쌀    46900    46900       서울 서울+광역시
## 3   3  쌀    45300    45300       서울 서울+광역시
## 4   4  쌀    44800    44800       서울 서울+광역시
## 5   5  쌀    43900    43900       서울 서울+광역시
## 6   6  쌀    41990    41990       서울 서울+광역시
## 7   7  쌀    43900    43900       서울 서울+광역시
## 8   8  쌀    44800    44800       서울 서울+광역시
## 9   9  쌀    45000    45000     광역시 서울+광역시
## 10 10  쌀    43900    43900     광역시 서울+광역시
```

8. `rice_price_1` 데이터에서 `조사지역명2` (`서울+광역시` 와 `지역`)간의 쌀의 소비자 `당일가격` 에 대한 모분산이 서로 다른지를 가설검정하는 R코드를 작성하고, 실행결과를 해석하라.

9. 위의 결과를 바탕으로 `rice_price_1` 데이터에서 `조사지역명2` (`서울+광역시` 와 `지역`)간의 쌀의 소비자 `당일가격` 이 평균적으로 같은지를 가설검정하는 R코드를 작성하고, 실행결과를 해석하라.

10. `rice_price_1` 데이터에서 쌀의 `당일가격` 이 `전일가격` 보다 평균적으로 높은지를 가설검정하는 R코드를 작성하고, 실행결과를 해석하라.

> **B. 아래의 url의 데이터는 어느 은행과 거래하는 고객의 정보이다.**
> 각 컬럼은 `나이`, `학력수준`, `소득` 으로 구성되어 있다.\
> http://datamining.dongguk.ac.kr/data/finance.csv

1. 위 자료를 읽어 R데이터프레임 `customer_info` 로 저장하라.

```
##    나이 학력수준      소득
## 1    54       14 66814.195
## 2    40       12 42144.338
## 3    35       14 25697.767
## 4    55       12 35976.874
## 5    40       12 39060.606
## 6    82       12 13362.839
## 7    26       16 61674.641
## 8    50       14 53451.356
## 9    71       12 16446.571
## 10   70        6  9867.943
```

2. 위 자료에서 `나이` 를 아래와 같은 간격으로 범주화하여 `customer_info` 의 `age_group` 변수로 저장하여라(단 `factor` 형 변수로 저장할 것).

나이구간 `edu_group`

나이 \< 40 20~30대
40[\(\le\)]{.math .inline} 나이 \< 60 40~50대
[\(\ge\)]{.math .inline} 60 60대이상

```
##    나이 학력수준      소득 age_group
## 1    54       14 66814.195   40~50대
## 2    40       12 42144.338   20~30대
## 3    35       14 25697.767   20~30대
## 4    55       12 35976.874   40~50대
## 5    40       12 39060.606   20~30대
## 6    82       12 13362.839  60대이상
## 7    26       16 61674.641   20~30대
## 8    50       14 53451.356   40~50대
## 9    71       12 16446.571  60대이상
## 10   70        6  9867.943  60대이상
```

3. 위 자료에서 `학력수준` 를 다음과 같이 세개의 구간으로 나누어 범주화하고 `customer_info` 의 `edu_group` 변수로 저장하여라(단 `factor` 형 변수로 저장할 것).

학력수준구간 `edu_group`

학력수준 \< 12 고졸이하
12[\(\le\)]{.math .inline} 학력수준 \< 15 대졸
학력수준 [\(\ge\)]{.math .inline} 12 대학원이상

```
##     나이 학력수준       소득 age_group   edu_group
## 1    54       14 66814.195   40~50대        대졸
## 2    40       12 42144.338   20~30대    고졸이하
## 3    35       14 25697.767   20~30대        대졸
## 4    55       12 35976.874   40~50대    고졸이하
## 5    40       12 39060.606   20~30대    고졸이하
## 6    82       12 13362.839  60대이상    고졸이하
## 7    26       16 61674.641   20~30대  대학원이상
## 8    50       14 53451.356   40~50대        대졸
## 9    71       12 16446.571  60대이상    고졸이하
## 10   70        6  9867.943  60대이상    고졸이하
```

4. 위 자료에서 `소득` 를 다음과 같이 네개의 구간으로 나누어 범주화하고 `customer_info` 의 `income_group` 변수로 저장하여라(단 `factor` 형 변수로 저장할 것, 아래 표에서 [\(Q_1, Q_2, Q_3\)]{.math .inline}는 `소득` 의 제1, 제2, 제3사분위수임).

소득구간 `income_group`

소득[\(\<Q_1\)]{.math .inline} 매우낮음
[\(Q_1\le\)]{.math .inline}소득[\(\<Q_2\)]{.math .inline} 낮음
[\(Q_2\le\)]{.math .inline}소득[\(\<Q_3\)]{.math .inline} 높음
소득 [\(\ge Q_3\)]{.math .inline} 매우높음

```
##     나이 학력수준       소득 age_group   edu_group income_group
## 1    54       14 66814.195   40~50대        대졸      매우높음
## 2    40       12 42144.338   20~30대    고졸이하          높음
## 3    35       14 25697.767   20~30대        대졸          낮음
## 4    55       12 35976.874   40~50대    고졸이하          낮음
## 5    40       12 39060.606   20~30대    고졸이하          높음
## 6    82       12 13362.839  60대이상    고졸이하      매우낮음
## 7    26       16 61674.641   20~30대  대학원이상      매우높음
## 8    50       14 53451.356   40~50대        대졸          높음
## 9    71       12 16446.571  60대이상    고졸이하      매우낮음
## 10   70        6  9867.943  60대이상    고졸이하      매우낮음
```

5. `age_group` 과 `income_group` 의 교차표를 작성하라.

```
##
##             매우낮음 낮음 높음 매우높음
##   20~30대         9   10    7       11
##   40~50대         6   11   12       11
##   60대이상        9    5    5        3
```

6. 위의 결과에서 `age_group` 이 `60대이상` 과 `40~50대` 의 소득(`income_group`)분포가 서로 다르다고 볼 수 있는지를 가설검정하려고 한다. 어떤 가설검정법을 사용해야 하는지를 밝히고, 가설검정을 위한 R코드를 작성 및 그 결과를 해석하라.

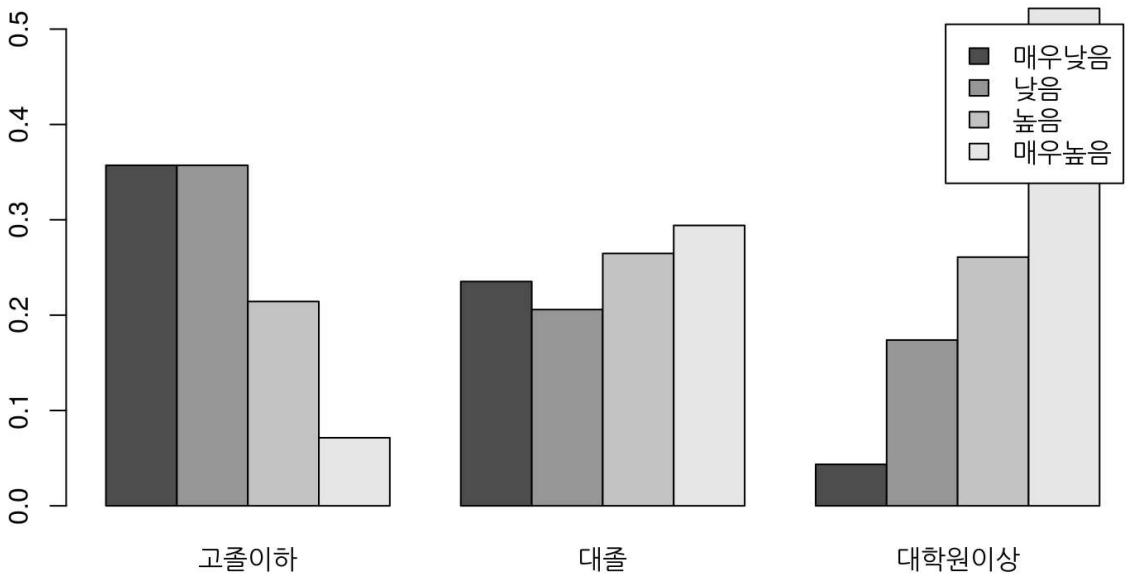
7. `edu_group` 과 소득(`income_group`)과의 교차표를 작성하라.

```
##
##             고졸이하 대졸 대학원이상
##   매우낮음       15    8          1
##   낮음           15    7          4
##   높음            9    9          6
##   매우높음        3   10         12
```

8. 위의 결과를 아래와 같이 각 컬럼의 합이 1이 되도록 비율로 변환하여라.

```
##
##               고졸이하       대졸 대학원이상
##   매우낮음 0.35714286 0.23529412 0.04347826
##   낮음     0.35714286 0.20588235 0.17391304
##   높음     0.21428571 0.26470588 0.26086957
##   매우높음 0.07142857 0.29411765 0.52173913
```

9. 위에서 계산한 비율을 아래와 같이 막대그래프로 나타내고, 결과를 해석하라.



{width="800"}

10. `edu_group` 과 소득(`income_group`)이 서로 연관성이 있는지를 가설검정하려고 한다. 어떤 가설검정법을 사용해야 하는지를 밝히고, 가설검정을 위한 R코드 작성 및 실행 결과를 해석하라.
:::