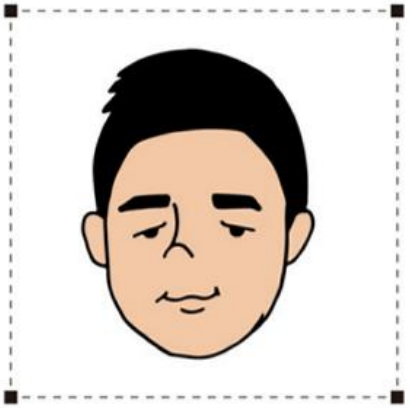
히든:그레이스

실전! 데이터 분석 R

-일단 함 해보자! 실전 데이터 분석 첫걸음-

Profile



김영우

㈜히든그레이스 데이터분석팀장

데이터분석 교육(데이터분석방법론, R, SPSS, AMOS)

연구방법론 및 통계분석 대학원 출강(2013~현재)

데이터 저널리즘 언론매체 '데이터 저널(datajournal.kr)' 대표

오마이뉴스 시민기자 - '데이터로 보는 뉴스' 연재

연애심리 연구 집단 '모태솔로연구소' 소장

한국데이터베이스진흥원 빅데이터 분석 전문가 과정 수료

1.카이검정

1단계.준비하기

패키지 설치 & 로드

```
# 패키지 설치
install.packages("dplyr")    # 데이터 정제
install.packages("ggplot2") # 그래프 만들기

# 패키지 로드
library(dplyr)
library(ggplot2)

# 데이터 불러오기
raw_01 <- read.csv(file="01_chi_t.csv", header = T)

# 복사본 생성
df_a <- raw_01
```

2단계.전처리

```r
# 데이터 탐색
head(df_a)
```

```
##   id sex alc snack prefer1 prefer2 prefer3 m_prefer
## 1  1   1   2     1       4       5       4     4.33
## 2  2   1   2     2       3       3       1     2.33
## 3  3   1   2     2       5       4       5     4.67
## 4  4   1   1     1       4       3       4     3.67
## 5  5   1   2     1       4       4       3     3.67
## 6  6   1   2     2       1       5       5     3.67
```

```r
str(df_a)
```

```
## 'data.frame':    100 obs. of  8 variables:
##  $ id      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ sex     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ alc     : int  2 2 2 1 2 2 1 1 2 1 ...
##  $ snack   : int  1 2 2 1 1 2 2 1 2 1 ...
##  $ prefer1 : int  4 3 5 4 4 1 5 4 3 5 ...
##  $ prefer2 : int  5 3 4 3 4 5 5 5 4 5 ...
##  $ prefer3 : int  4 1 5 4 3 5 3 3 5 3 ...
##  $ m_prefer: num  4.33 2.33 4.67 3.67 3.67 3.67 4.33 4 4 4.33 ...
```

```r
summary(df_a)
```

```
##        id              sex             alc           snack      
##  Min.   :  1.00   Min.   :1.00   Min.   :1.00   Min.   :1.0  
##  1st Qu.: 25.75   1st Qu.:1.00   1st Qu.:1.00   1st Qu.:1.0  
##  Median : 50.50   Median :1.00   Median :2.00   Median :1.0  
##  Mean   : 50.50   Mean   :1.49   Mean   :1.55   Mean   :1.3  
##  3rd Qu.: 75.25   3rd Qu.:2.00   3rd Qu.:2.00   3rd Qu.:2.0  
##  Max.   :100.00   Max.   :2.00   Max.   :2.00   Max.   :2.0  
##     prefer1         prefer2         prefer3         m_prefer   
##  Min.   :1.00   Min.   :1.00   Min.   :1.00   Min.   :1.33  
##  1st Qu.:2.00   1st Qu.:2.00   1st Qu.:2.00   1st Qu.:2.33  
##  Median :3.00   Median :3.00   Median :3.00   Median :3.33  
##  Mean   :3.28   Mean   :3.23   Mean   :3.33   Mean   :3.28  
##  3rd Qu.:4.00   3rd Qu.:4.00   3rd Qu.:4.00   3rd Qu.:4.33  
##  Max.   :5.00   Max.   :5.00   Max.   :5.00   Max.   :5.00
```

```r
# 값 변경
df_a$sex <- ifelse(df_a$sex == 1, "M", "F")
table(df_a$sex)
```

```
##
##  F  M
## 49 51
```

```r
df_a$alc <- ifelse(df_a$alc == 1, "Beer", "Soju")
table(df_a$alc)
```

```
##
## Beer Soju
##   45   55
```

3단계.분석

1.분할표 만들기

(1)빈도 분할표

```
tab_alc <- table(df_a$alc, df_a$sex)
tab_alc
```

```
##
##          F  M
##   Beer 29 16
##   Soju 20 35
```

(2)비율 분할표 - 열 기준

```
proptab_alc <- prop.table(tab_alc, 2)
proptab_alc
```

```
##
##                  F         M
##   Beer 0.5918367 0.3137255
##   Soju 0.4081633 0.6862745
```

```r
# 백분율 표기
proptab_alc <- prop.table(tab_alc, 2)*100
proptab_alc
```

```
##
##                F        M
##   Beer 59.18367 31.37255
##   Soju 40.81633 68.62745
```

```r
# 소숫점 둘째자리까지 표기
proptab_alc <- round(prop.table(tab_alc, 2)*100, 2)
proptab_alc
```

```
##
##              F     M
##   Beer 59.18 31.37
##   Soju 40.82 68.63
```

(3)파일로 저장

```
# 빈도표, 비율표 합치기
tab.cross <- cbind(tab_alc, proptab_alc)
tab.cross
```

```
##        F  M     F     M
## Beer 29 16 59.18 31.37
## Soju 20 35 40.82 68.63
```

```
#csv 로 저장
write.csv(tab.cross, "output_chi.csv")
```

2.카이검정

```
chisq.test(df_a$sex, df_a$alc, correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df_a$sex and df_a$alc
## X-squared = 7.8096, df = 1, p-value = 0.005197
```

```r
chisq.test(df_a$sex, df_a$alc)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  df_a$sex and df_a$alc
## X-squared = 6.7263, df = 1, p-value = 0.0095
```

[참고] default - 2x2 경우 예이츠 수정지표

4단계.시각화

(1)비율표 -> 데이터프레임 변환

```
proptab_alc
```

```
##
##              F     M
##   Beer 59.18 31.37
##   Soju 40.82 68.63
```

```
df_proptab <- as.data.frame(proptab_alc)
df_proptab
```

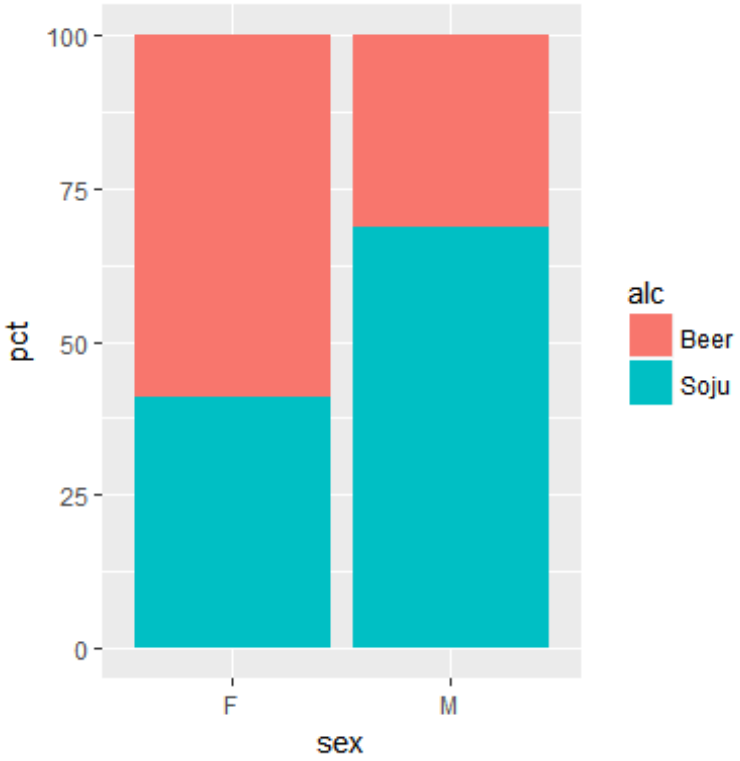
```
##   Var1 Var2  Freq
## 1 Beer    F 59.18
## 2 Soju    F 40.82
## 3 Beer    M 31.37
## 4 Soju    M 68.63
```

```
# 변수명 변경
df_proptab <- rename(df_proptab,
                     alc = Var1,
                     sex = Var2,
                     pct = Freq)

df_proptab
```

```
##    alc sex   pct
## 1 Beer   F 59.18
## 2 Soju   F 40.82
## 3 Beer   M 31.37
## 4 Soju   M 68.63
```

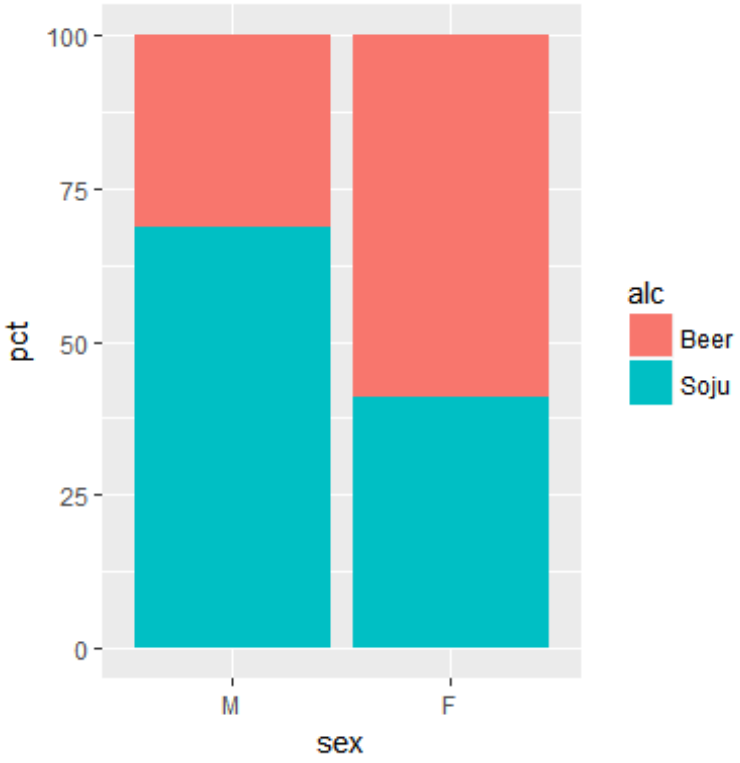
(2)그래프 생성

```
ggplot(data = df_proptab, aes(x = sex, y = pct, fill = alc)) +
  geom_col()
```



그래프 Tip - x축 순서 지정

```
ggplot(data = df_proptab, aes(x = sex, y = pct, fill = alc)) +
  geom_col() +
  scale_x_discrete(limit = c("M", "F"))
```



카이검정 복습!

1단계.준비하기

```
# 패키지 설치 & 로드
library(dplyr)
library(ggplot2)

# 데이터 로드
raw_01 <- read.csv(file="01_chi_t.csv", header = T)

# 복사본 생성
df_a <- raw_01
```

2단계.전처리

```r
# 데이터 탐색
str(df_a)
summary(df_a)
head(df_a)

# 값 변경
df_a$sex <- ifelse(df_a$sex == 1, "M", "F")
df_a$alc <- ifelse(df_a$alc == 1, "Beer", "Soju")
```

3단계.분석

```
# 1.분할표 만들기


#(1)빈도 분할표
tab_alc <- table(df_a$alc, df_a$sex)
tab_alc
```

```
##
##          F  M
##   Beer 29 16
##   Soju 20 35
```

```
#(2)비율 분할표
proptab_alc <- round(prop.table(tab_alc, 2)*100, 2)
proptab_alc
```

```
##
##             F     M
##   Beer 59.18 31.37
##   Soju 40.82 68.63
```

```r
#(3)파일로 저장

# 빈도표, 비율표 합치기
tab.cross <- cbind(tab_alc, proptab_alc)
tab.cross
```

```
##        F  M     F     M
## Beer 29 16 59.18 31.37
## Soju 20 35 40.82 68.63
```

```r
# csv 로 저장
write.csv(tab.cross, "output_crosstab.csv")
```

```r
# 2.카이검정
chisq.test(df_a$sex, df_a$alc, correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df_a$sex and df_a$alc
## X-squared = 7.8096, df = 1, p-value = 0.005197
```

4단계.시각화

```r
#(1)Data Frame 으로 변환
df_proptab <- as.data.frame(proptab_alc)

# 변수명 변경
df_proptab <- rename(df_proptab,
                     alc = Var1,
                     sex = Var2,
                     pct = Freq)

#(2)그래프 생성
ggplot(data = df_proptab, aes(x = sex, y = pct, fill = alc)) +
  geom_col() +
  scale_x_discrete(limit = c("M", "F"))
```

2.t-test

1.기술통계표 만들기

```
tab_t <- df_a %>%
  group_by(sex) %>%
  summarise(n = n(),
            m = round(mean(m_prefer), 2),
            sd = round(sd(m_prefer), 2))
tab_t
```

```
## # A tibble: 2 × 4
##     sex     n     m    sd
##   <chr> <int> <dbl> <dbl>
## 1     F    49  2.51  0.76
## 2     M    51  4.02  0.72
```

```
# 표 저장
write.csv(tab_t, "output_ttest.csv")
```

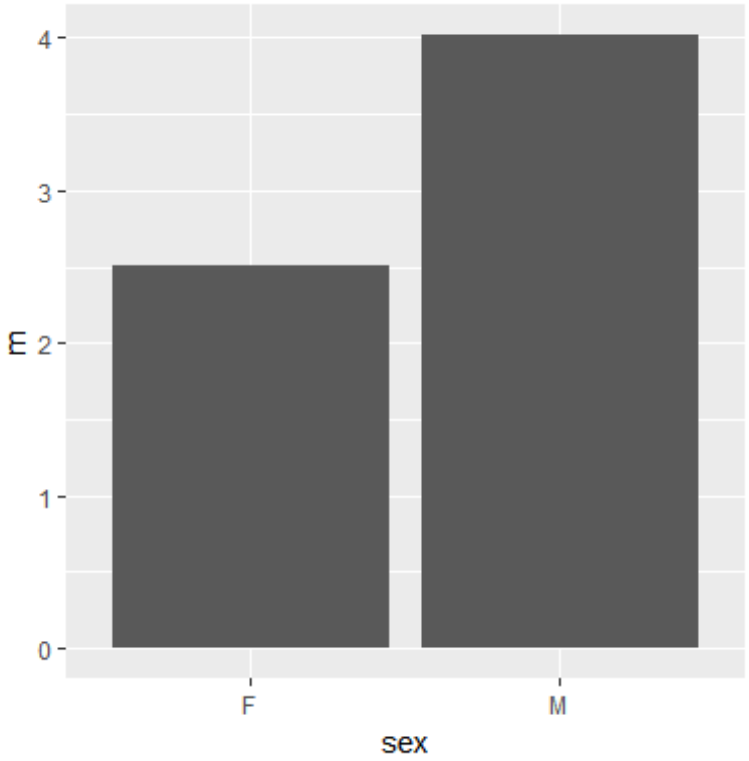
2.t-test

```r
t.test(data = df_a, m_prefer ~ sex, var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  m_prefer by sex
## t = -10.173, df = 98, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.803629 -1.214818
## sample estimates:
## mean in group F mean in group M
##        2.509796        4.019020
```

3.시각화

```
ggplot(data = tab_t, aes(x = sex, y = m)) + geom_col()
```



3.회귀분석

1. 준비하기

```
# 데이로 로드
raw_02 <-read.csv(file="02_reg.csv", header = T)

# 복사본 생성
df_b <- raw_02
```

```r
# 데이터 탐색
str(df_b)
```

```
## 'data.frame':    133 obs. of  3 variables:
##  $ id   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ skill: num  2.71 1.57 2.39 3.33 1.57 ...
##  $ sat  : num  3.18 1.23 2.05 1.3 1.74 ...
```

```r
summary(df_b)
```

```
##        id          skill            sat       
##  Min.   :  1   Min.   :1.157   Min.   :1.022  
##  1st Qu.: 34   1st Qu.:1.857   1st Qu.:1.907  
##  Median : 67   Median :2.443   Median :2.614  
##  Mean   : 67   Mean   :2.580   Mean   :2.598  
##  3rd Qu.:100   3rd Qu.:2.857   3rd Qu.:3.068  
##  Max.   :133   Max.   :5.000   Max.   :5.000
```

```r
head(df_b)
```

```
##   id    skill      sat
## 1  1 2.714286 3.181818
## 2  2 1.571429 1.231818
## 3  3 2.390000 2.045455
## 4  4 3.330000 1.295455
## 5  5 1.571429 1.735648
## 6  6 1.714286 1.045455
```

2.회귀분석

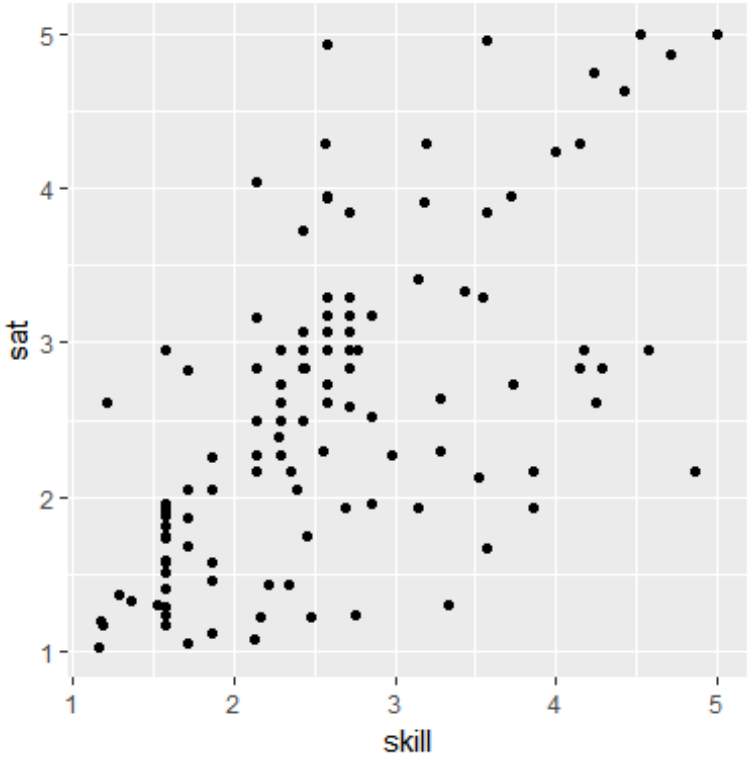
```
out_reg <- lm(data = df_b, sat ~ skill)  # 분석
summary(out_reg)                         # 분석결과 출력
```

```
##
## Call:
## lm(formula = sat ~ skill, data = df_b)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.03714 -0.51331  0.02348  0.48926  2.33349
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.78790    0.20973   3.757 0.000258 ***
## skill        0.70172    0.07701   9.113 1.19e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7758 on 131 degrees of freedom
## Multiple R-squared:  0.388,  Adjusted R-squared:  0.3833
## F-statistic: 83.04 on 1 and 131 DF,  p-value: 1.194e-15
```

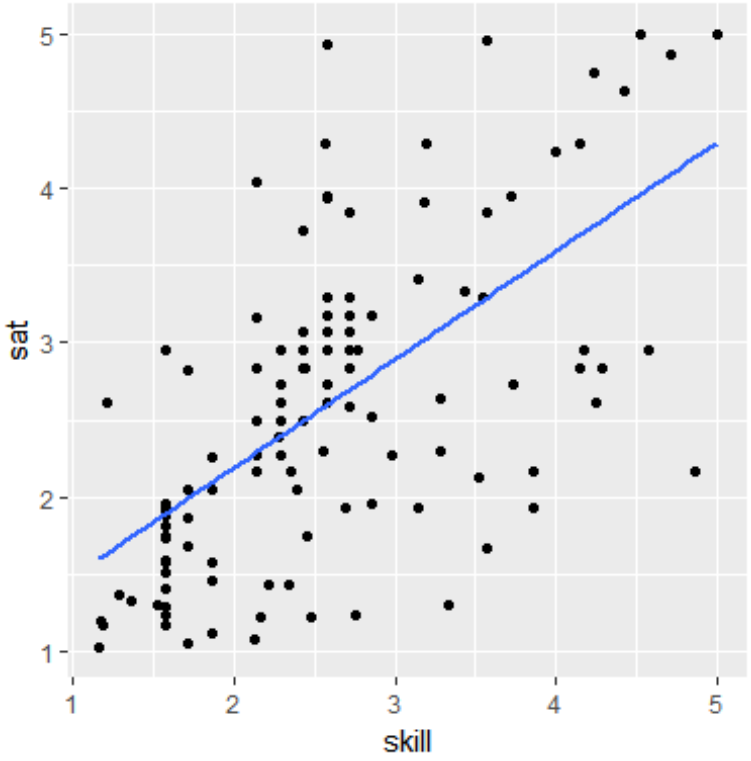
```r
# 회귀분석표 html 출력
install.packages("ztable") # 회귀분석표 정리용
library(ztable)

# html 출력
ztable(out_reg)
```

3.시각화

```r
# 산점도 만들기
ggplot(data = df_b, aes(x = skill, y = sat)) + geom_point()
```



```r
# 회귀선 추가
ggplot(data = df_b, aes(x = skill, y = sat)) +
  geom_point() +
  geom_smooth(method = lm, se = F)
```



실전 분석 프로젝트!

"한국인의 삶의 질, 원인 분석 프로젝트" - 한국복지패널 데이터

분석1. 성별에 따른 소득

1.준비하기

```
# foreign 패키지 설치 & 로드
install.packages("foreign")
library(foreign)

# 복지패널데이터 로드
raw_03 <- read.spss("data_spss_Koweps2014.sav", to.data.frame=T)

# 복사본 생성
df_c <- raw_03
```

```r
# 데이터 검토
dim(df_c)
str(df_c, list.len=10)
summary(df_c)
head(df_c)
View(df_c)
```

2.독립변수, 종속변수 검토 및 정제

변수명 변경

```
df_c <- rename(df_c,
               sex = h0901_4,      # 성별
               income = h09_din)   # 소득
```

변수 검토 - 성별

```r
class(df_c$sex)
```

```
## [1] "numeric"
```

```r
summary(df_c$sex)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   1.309   2.000   2.000
```

```r
table(df_c$sex)
```

```
##
##    1    2
## 4873 2175
```

```r
# 이상치 결측처리 - 모름/무응답=9
df_c$sex <- ifelse(df_c$sex == 9, NA, df_c$sex)
table(is.na(df_c$sex))
```

```
##
## FALSE
##  7048
```

```r
# sex 항목 이름 부여
df_c$sex <- ifelse(df_c$sex == 1, "M", "F")
table(df_c$sex)
```

```
##
##    F    M
## 2175 4873
```

변수 검토 - 소득

```
class(df_c$income)
```

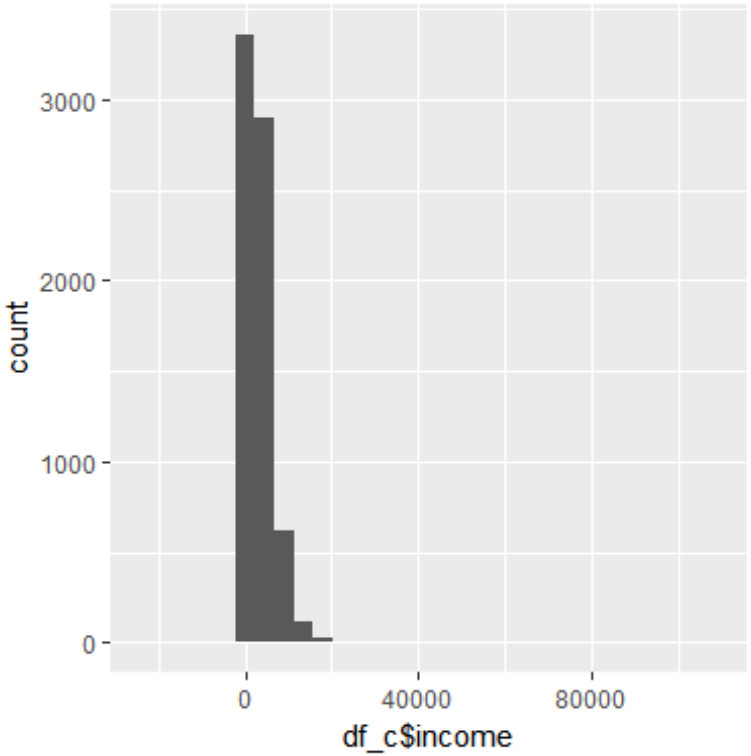
```
## [1] "numeric"
```

```
summary(df_c$income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -20520    1108    2404    3336    4642  108900
```

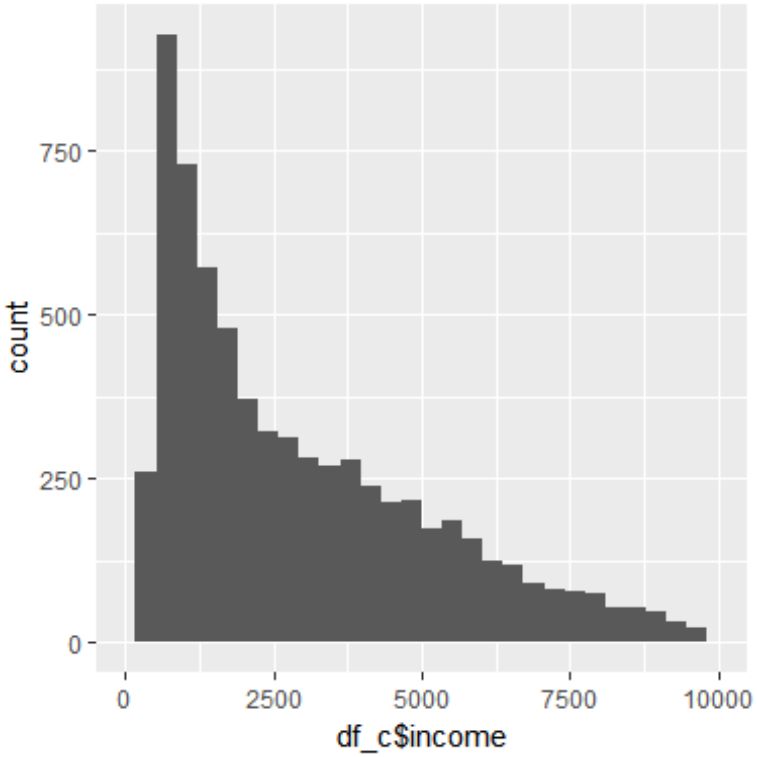
```
qplot(df_c$income)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
qplot(df_c$income) + xlim(0, 10000)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```r
# 소득 결측치 확인
table(is.na(df_c$income))
```

```
##
## FALSE
##  7048
```

3.분석

(1)기술통계표 만들기

```
income_sex <- df_c %>%
  group_by(sex) %>%
  summarise(n = n(),
            m = round(mean(income), 2),
            sd = round(sd(income), 2))
income_sex
```

```
## # A tibble: 2 × 4
##     sex     n       m      sd
##   <chr> <int>   <dbl>   <dbl>
## 1     F  2175 1581.25 2103.14
## 2     M  4873 4118.90 3870.89
```

```
# 표 저장
write.csv(income_sex, "output_income_sex.csv")
```

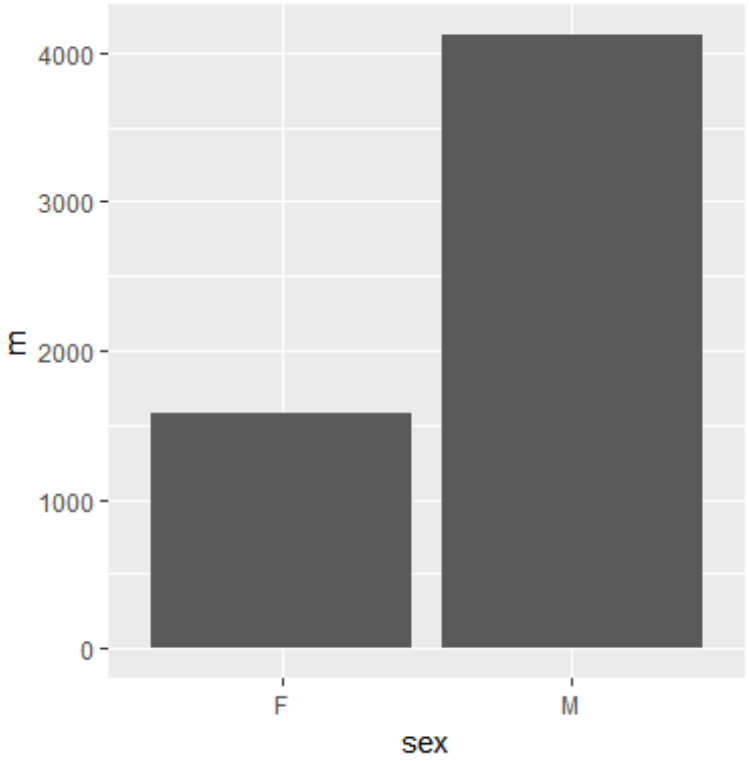
(2)t-test

```
t.test(data = df_c, income ~ sex, var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  income by sex
## t = -28.738, df = 7046, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2710.746 -2364.551
## sample estimates:
## mean in group F mean in group M
##        1581.255        4118.903
```

4.시각화

```
ggplot(data = income_sex, aes(x = sex, y = m)) + geom_col()
```



분석2. 30~40대의 성별에 따른 소득차이

1.준비하기

- 패키지 설치 & 로드
- 데이터 불러오기

2.데이터 추출

30~40대 추출

```
# 변수명 변경
df_c <- rename(df_c,
               birth = h0901_5)

# 나이 변수 생성
df_c$age <- 2014 - df_c$birth + 1
summary(df_c$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.00   49.00   63.00   62.01   75.00   97.00
```

```
# 30~40 대 추출
df_3040 <- df_c %>%
  filter(age >= 30 & age <= 49)

# 데이터 확인
summary(df_3040$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   30.00   38.00   42.00   41.48   46.00   49.00
```

3.독립변수, 종속변수 검토 및 정제 - 성별, 소득

- 앞에서 완료

4.분석

(1)기술통계표 만들기

```
tab_3040 <- df_3040 %>%
  group_by(sex) %>%
  summarise(n = n(),
            m = round(mean(income), 2),
            sd = round(sd(income), 2))

tab_3040
```

```
## # A tibble: 2 × 4
##     sex     n       m      sd
##   <chr> <int>   <dbl>   <dbl>
## 1     F   202 3223.85 5153.03
## 2     M  1550 5472.34 4518.54
```

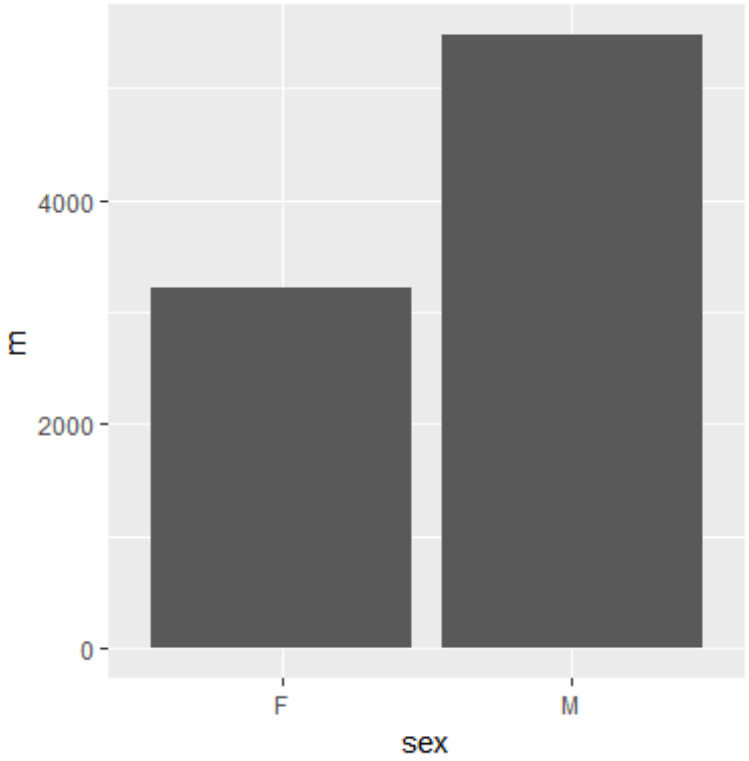
(2)t-test

```
t.test(data = df_3040, income ~ sex, var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  income by sex
## t = -6.5403, df = 1750, p-value = 8.036e-11
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2922.773 -1574.208
## sample estimates:
## mean in group F mean in group M
##        3223.847        5472.337
```

5.시각화

```
ggplot(data = tab_3040, aes(x = sex, y = m)) + geom_col()
```



‘통계, 히든:그레이스에서 답을 찾다

데이터분석강의

공공데이터분석

논문통계분석

논문통계강의

빅데이터분석

HOON

HOON

세상의 히든카드를 Joker로 만들다

서비스 소개

데이터 분석

예측분석 모델링

데이터분석 교육

논문통계

세상의 히든카드를 Joker로 만들다

Contact

전화 010-5461-7445

메일 stats7445@gmail.com

김영우, Kim Young Woo
데이터분석팀장
www.hidden-analysis.co.kr
010.5461.7445
stats7445@gmail.com

김성은, Kim Sung Eun
대표이사
www.hidden-analysis.co.kr
010.3558.8121
admin@hiddenjgrace.com

세상의 히든카드를 Joker로 만들다

찾아오는 길

대중교통 이용시

지하철

강남역 신분당선 하차 후 강남역 5번 출구, 도보 1분거리

강남역 하차

간선버스 140, 400, 402, 407, 420, 440, 441, 462, 470, 471, 541, 542
광역버스 M4403, M6427, M7412, M7426, 9404, 9408

강남역 도씨에빛 II 하차

직행버스 1311, 5300, 5300-1, 6800
광역버스 9500, 9501, 9802

찾는 방법

(STEP 1) 파리바게뜨와 아리따움 화장품 가게를 지난다.
(STEP 2) CU편의점과 GUGUS 옆 모모안경점 사이에 도씨에빛 2 건물 출입구를 찾는다.
(STEP 3) 엘레베이터 18층을 누른다.
(STEP 4) 오른쪽 끝에 있는 1801호 (유리문)으로 온다.