
비즈니스모델링1

복지 데이터 분석



제출일	2020-06-27	전공	경영학과
과목	비즈니스모델링1	학번	60180998
담당교수	김 일주 교수님	이름	이 서현

목 차

제 1 장 서론	3
제 2 장 저소득 가구의 가구 형태	4
제 1 절 저소득 가구 파악	4
제 2 절 최저임금 인상의 정당성	5
제 3 장 흡연과 수면의 관계 분석	6
제 1 절 흡연자와 비흡연자 비교 분석	6
제 2 절 결론	13
제 4 장 최종 학력과 임금과의 관계 분석	14
제 1 절 저학력자와 고학력자 비교 분석	14
제 2 절 직업 만족도 분석	19
제 3 절 결론	22

제 1 장 서론

본 연구는 한국복지패널조사(Koweps) 자료를 이용하여 3가지의 주제를 선정하여 조사를 시도하였다.

첫번째로 소득에 따른 가구 구분 변수(h10_hc)와, 가구의 형태 변수(h1001_110)를 이용하여 저소득 가구의 가구 형태를 파악하고 최저임금 상승의 정당성을 보이기 위한 결과 자료로 연구했다.

두번째로 현대사회에 주된 이슈에는 불면증이 있다. 따라서 불면증의 원인 중 흡연과 관련이 있을지 연구하게 되었다. 담배와 불면증과 관련된 학위 논문은 RISS 학술연구정보서비스에서 조사해봤지만 담배와 수면에 관한 논문은 발견하지 못했다. 담배와 수면의 관계를 파악하기 위해 현재 흡연 여부 변수(p1005_3aq5)와 하루 평균 흡연량(개비) 변수(p1005_3aq6)와 잠을 설치는 정도(p1005_13)을 사용했다.

세번째로 최종 학력과 임금과의 상관 관계에 대해서 연구했다. 학생들의 주 관심사는 대학진학일 것이다. 많은 초중고학생들이 대학에 진학하기 위해 12년동안 공부를 진행하고 있다. 최종 학력에 따른 임금이 어떻게 변화되는지 한국복지패널조사 자료를 이용하여 연구했다. 사용한 변수로는 일을 한 달의 월 평균 임금(단위: 만원)(p1002_8aq1), 최종학력 변수(p1007_3aq1)와 직업 만족도 변수(p1003_9)를 사용했다.

코드

```
library(foreign)
library(dplyr)
library(ggplot2)

originalData <- read.spss("Koweps_hpc10_2015_beta1.sav", to.data.frame = T)
```

본 연구에서는 foreign, dplyr, ggplot2 패키지를 이용했다. foreign 패키지를 이용하여 sav 파일을 originalData 변수에 담았다.

제 2 장 저소득 가구의 가구 형태

제 1 절 저소득 가구 파악

코드

```
Data1 <- Data1 %>%
  select(h10_hc, h1001_110)

Data1 <- Data1 %>%
  rename('classification_income' = h10_hc) %>%
  rename('form' = h1001_110)

Data1$form <- ifelse(Data1$form == 5, NA, Data1$form)

Result1 <- Data1 %>%
  filter(!is.na(form) & classification_income == 2) %>%
  group_by(classification_income , form) %>%
  summarise(count = n())

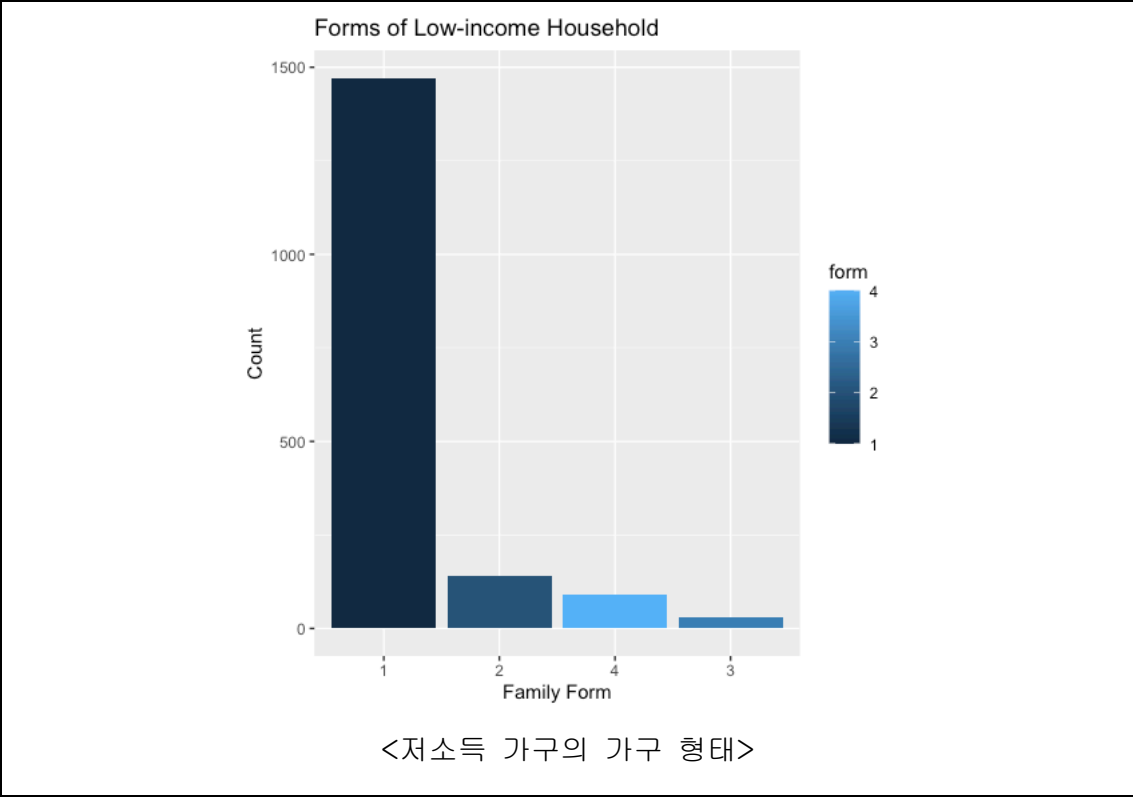
# 저소득 가구의 형태 그래프
ggplot(Result1, aes(reorder(form, -count), count, fill = form)) +
  geom_col() +
  xlab('Family Form') + ylab('Count') +
  ggtitle('Forms of Low-income Household')
```

코드 설명

select 함수를 사용하여 필요한 변수만을 원본 데이터에서 가져 왔다. rename 함수를 사용하여 변수들의 이름을 변경해주었다. h1001_110의 데이터는 1.단독, 2.모자, 3.부자, 4. 조손 가구 or 소년/소녀 가장, 5. 기타로 구성되어 있다. 이 연구에서는 5번 기타 자료가 가구의 형태가 다양해 연구 결과의 정확성을 저해시켜 이를 결측치(Missing Data)로 변환해주어 연구에서 제외했다.

Result1 변수에 filter함수를 이용하여 결측치를 제거해주고, 저소득 가구만 가져오도록 하여 저소득 가구의 가구 형태가 어떻게 되는지 그래프로 나타냈다.

결과물



그래프를 보면 알 수 있듯이, 1번 단독에 해당하는 가구의 형태가 압도적으로 나타났다.

제 2 절 최저임금 인상의 정당성

정부는 2020년까지 최저임금을 1만원으로 인상하기로 결정했다. 2019년 최저임금은 시급 8,350원으로 전년대비 10.9% 인상되었으며, 2020년 최저임금은 시급 8,590원으로 전년대비 2.9% 인상되었다. 최저임금 인상은 찬반이 나뉘고 있다. 최저임금 인상에 반대 측의 입장은 빈곤은 가구의 문제이기 때문에 최저임금 인상이 아닌 다른 수단으로 풀어야 한다고 주장한다. 하지만 한국복지패널조사 데이터에 따르면 빈곤은 단독 가구가 압도적으로 많은 것을 알 수 있다.

현대사회는 1인 단독 가구가 급증하고 있다. 단독 가구의 경우에는 최저임금이

유일한 소득원이다. 즉 최저임금이 인상이 되어야, 저소득 가구의 다수를 차지하고 있는 단독 가구의 생계 수준이 향상되는 것이다.

저소득 가구의 생계 수준 향상을 위해 최저임금의 인상은 정당하다. 하지만 최저임금의 인상이 소기업, 자영업자에게 부담을 주는 것은 사실이다. 따라서 정부는 국가재정으로 소기업 자영업자들에게 한시적으로 임금 보조하는 방안 등을 제시하여 최저임금 인상 관련 논란과 문제점들을 개선할 필요가 있다.

제 3 장 흡연과 수면의 관계 분석

제 1 절 흡연자와 비흡연자 비교 분석

코드

```
Data2 <- originalData

Data2 <- Data2 %>%
  select(p1005_3aq5, p1005_3aq6, p1005_13) %>%
  filter(p1005_3aq5 %in% c(1,2) & p1005_13 %in% c(1,2,3,4)) %>%
  rename(avg_daily_smoking = p1005_3aq6,
         sleeplessness = p1005_13,
         smoking = p1005_3aq5)

Data2$smoking <- ifelse(Data2$smoking == 1, 'yes', 'no')
table(Data2$smoking)
```

코드 설명

select함수와 filter함수를 이용해 결측치를 제거하고 흡연자와 비흡연자 그리고 잠을 설치는 정도의 변수를 가져왔다.

smoking 변수에는 ifelse함수를 이용하여 1이면 흡연자로 yes를 그렇지 않은 경우에는 비흡연자로 no로 데이터 값을 변환해주었다.

결과

no	yes
10728	2110

table함수를 통해 비흡연자와 흡연자의 수를 살펴보았다. 그 결과 비흡연자 수가 10,728명으로 흡연자의 수에 비해 5배 가량 많은 것을 확인할 수 있다.

코드

```
# 흡연자 Data
smoker_data <- Data2 %>%
  filter(smoking == 'yes')
non_smoker_data$sleeplessness <- ifelse(non_smoker_data$sleeplessness ==
1, 'A', ifelse(non_smoker_data$sleeplessness == 2, 'B',
               ifelse(non_smoker_data$sleeplessness == 3, 'C', 'D'))))
# 비흡연자 Sampling
non_smoker <- Data2 %>%
  filter(smoking == 'no')

non_smoker_row <- sample(rownames(non_smoker),
                        dim(non_smoker)[1]*(2110/10728))
non_smoker_data <- non_smoker[non_smoker_row,]

non_smoker_data$sleeplessness <- ifelse(non_smoker_data$sleeplessness ==
1, 'A', ifelse(non_smoker_data$sleeplessness == 2, 'B',
               ifelse(non_smoker_data$sleeplessness == 3, 'C', 'D'))))
# p1005_13 : 잠을 설치는 정도
# 1. 극히 드물다.(일주일에 1일 미만)
# 2. 가끔 있었다.(일주일에 1~2일간)
# 3. 종종 있었다.(일주일에 3~4일간)
# 4. 대부분 그랬다.(일주일에 5일 이상)
```

코드 설명

흡연자와 비흡연자 데이터 수를 맞춰주기 위해 흡연자 비율(2,110명)에 맞춰 비흡연자 중에서 Sampling을 진행했으며, ifelse함수를 통해 잠을 설치는 정도에 따라 A~D의 등급으로 구분해주었다.

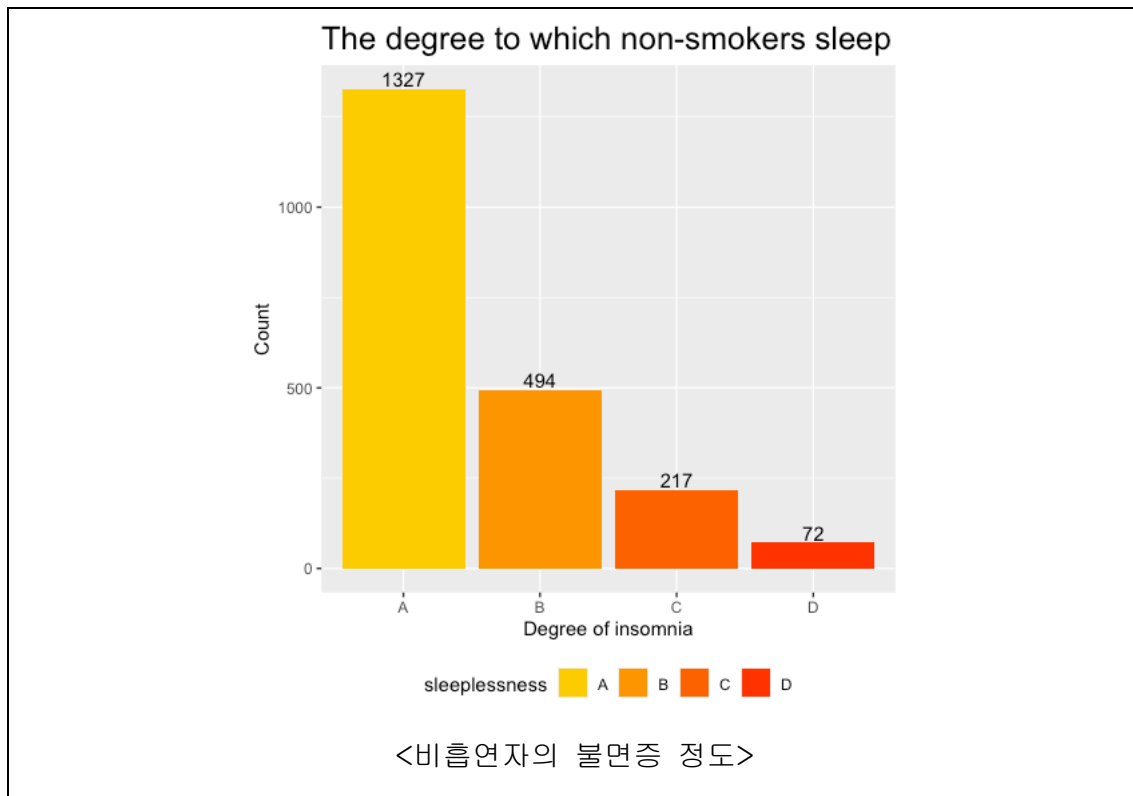
코드

```
# 비흡연자 불면증 정도
non_smoker_data %>%
  group_by(sleeplessness) %>%
  summarise(Count = n()) %>%
  ggplot(aes(x = sleeplessness, y = Count, fill = sleeplessness, label = Count))
+
  geom_col() +
  geom_text(vjust = -0.2, color = 'black') +
  ggtitle('The degree to which non-smokers sleep') +
  xlab('Degree of insomnia') + ylab('Count') +
  scale_fill_manual(values = c( "#FFCC00", "#FF9900",
                                "#FF6600", "#FF3300")) +
  theme(plot.title = element_text(size=18), legend.position = 'bottom')
```

코드 설명

ggplot2 패키지를 이용하여 비흡연자 불면증 정도를 막대 그래프로 나타냈다.

결과



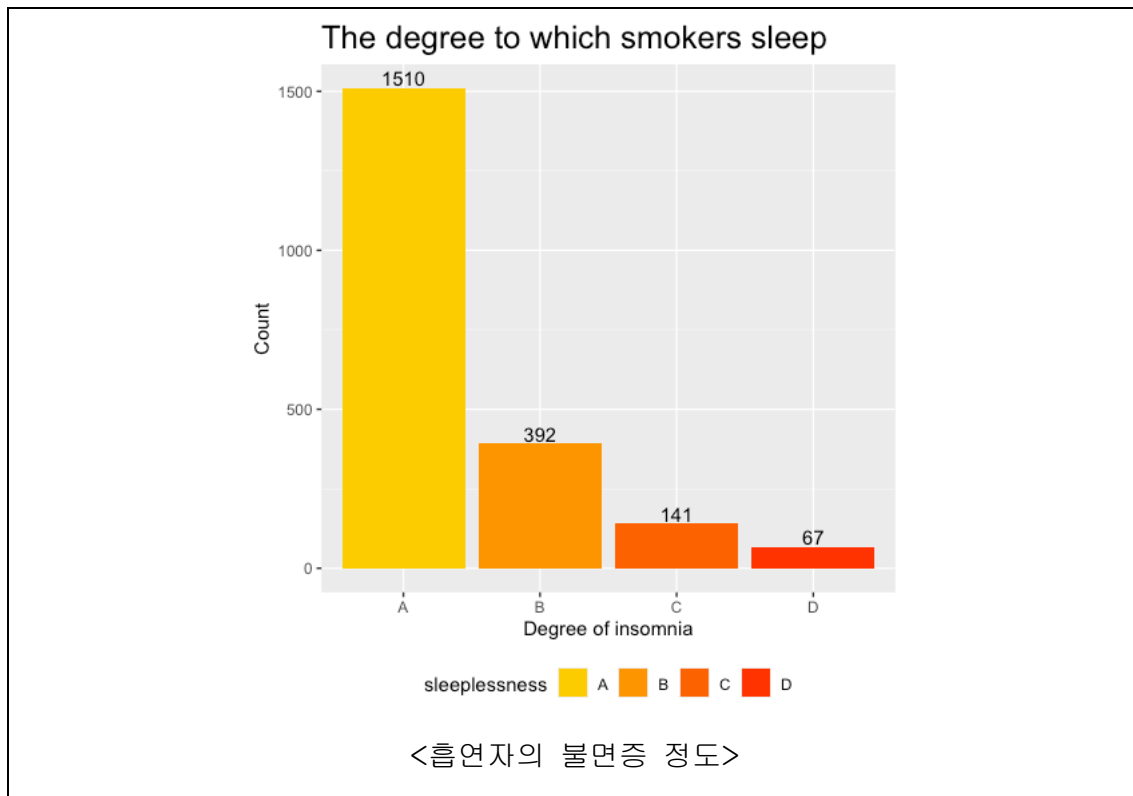
비흡연자의 불면증 정도는 A(극히 드물다.(일주일에 1일 미만))등급 1327명, B(가끔 있었다.(일주일에 1~2일간))등급 494명, C(종종 있었다.(일주일에 3~4일간))등급 217명, D(대부분 그랬다.(일주일에 5일 이상)) 72명으로 나타났다.

코드

```
# 흡연자 불면증 정도
smoker_data %>%
  group_by(sleeplessness) %>%
  summarise(Count = n()) %>%
  ggplot(aes(x = sleeplessness, y = Count, fill = sleeplessness, label = Count))
+
  geom_col() +
  geom_text(vjust = -0.2, color = 'black') +
  ggtitle('The degree to which smokers sleep') +
  xlab('Degree of insomnia') + ylab('Count') +
```

```
scale_fill_manual(values = c( "#FFCC00", "#FF9900",
                              "#FF6600", "#FF3300")) +
theme(plot.title = element_text(size=18), legend.position = 'bottom')
```

결과



흡연자의 경우 A등급 1510명, B등급 392명, C등급 141명, D등급 67명으로 나타났다.

코드

```
# 흡연자 비흡연자 그래프 같이 보기
smoker_nonsmoker_data <- bind_rows(smoker_data, non_smoker_data)

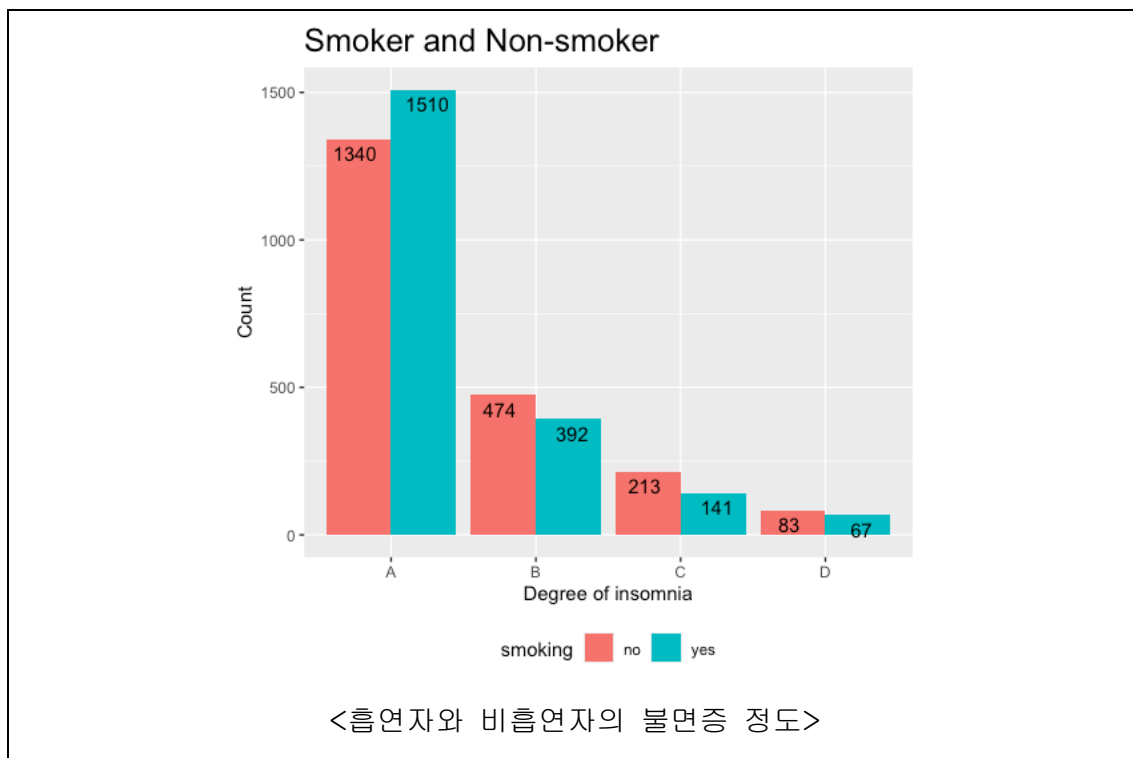
smoker_nonsmoker_data %>%
  group_by(sleeplessness, smoking) %>%
  summarise(count = n()) %>%
  ggplot(aes(sleeplessness, count, fill = smoking, label = count)) +
```

```

geom_col(position = 'dodge') +
geom_text(vjust = 1.5, color = 'black') +
ggtitle('Smoker and Non-smoker') +
xlab('Degree of insomnia') + ylab('Count') +
theme(plot.title = element_text(size=18), legend.position = 'bottom')

```

결과



흡연자와 비흡연자의 불면증 정도를 비교했을 때 흡연자가 불면증이 적은 것으로 나타났다. 비흡연자의 Sampling을 다수 진행해보았으나 대부분 위와 비슷한 결과가 나타났다.

코드

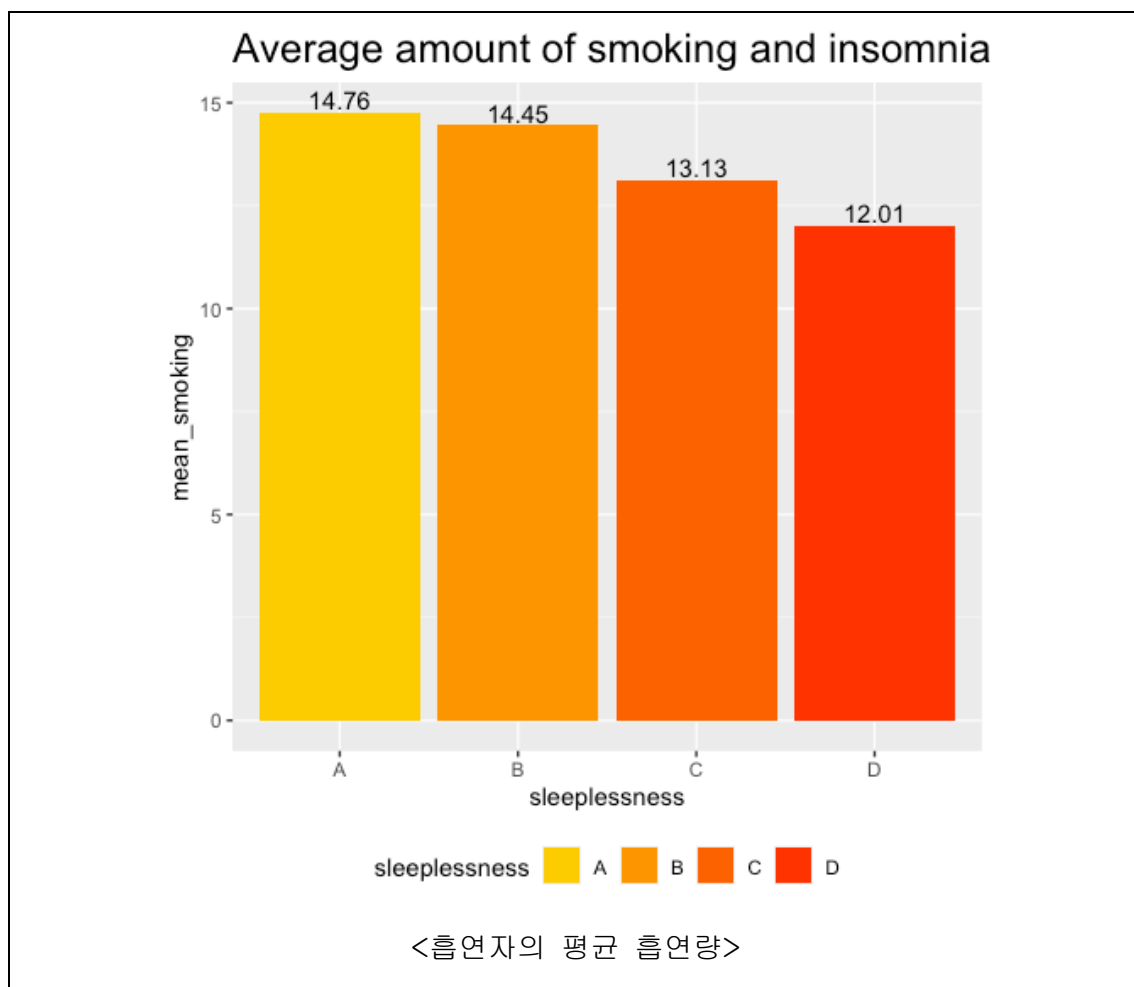
```

smoker_data %>%
  group_by(sleeplessness) %>%
  summarise(mean_smoking = round(mean(avg_daily_smoking), 2)) %>%
  ggplot(aes(sleeplessness, mean_smoking, fill = sleeplessness, label =

```

```
mean_smoking)) +
  geom_col() +
  geom_text(vjust = -0.2, color = 'black') +
  ggtitle('Average amount of smoking and insomnia') +
  scale_fill_manual(values = c( "#FFCC00", "#FF9900",
                                "#FF6600", "#FF3300")) +
  theme(plot.title = element_text(size=18), legend.position = 'bottom')
```

결과



흡연자의 평균 흡연량(개비)에 따른 불면증 정도를 막대 그래프로 나타냈다. A등급은 14,76개비로 가장 많은 흡연량을 가졌으며, B등급은 14,45개비로 두 번째로 많은 흡연량을 보여준다.

제 2 절 결론

처음 연구를 시작했을 때 예상한 것은 흡연량이 많을 수록 숙면에 어려울 것을 예상하고 연구를 시작했다. 하지만 연구 내용에 따르면 흡연자가 비흡연자보다 수면의 질이 좋은 것으로 나타났으며, 흡연량이 많을 수록 수면의 질이 더 좋은 것으로 나타났다. 본고의 추정 결과는 담배가 수면에 미치는 영향이 없는 것으로 나타난다. 하지만 주의할 것은 수면의 질을 높이기 위해 흡연을 하는 것은 절대 좋지 않은 행동이다. 흡연이 인체에 미치는 유해성은 많은 연구 결과로 입증되었다. 따라서 이 문제에 대한 실증적 분석은 우연성이 어떻게 흡연과 수면의 질에 영향을 미치는지에 대한 우리의 이해를 깊게 해줄 수 있다.

본 연구는 흡연자 2,110명, 비흡연자 2,110명을 대상으로 조사를 했기 때문에 더 많은 데이터를 가지고 연구를 진행한다면, 다른 결과가 나올 수 있다.

제 4 장 최종 학력과 임금과의 관계 분석

제 1 절 저학력자와 고학력자 비교 분석

코드

```
# p1002_8aq1 : 일한달의 월 평균 임금(만원)
# p1007_3aq1 : 최종 학력
#   1. 중학교 졸업 이하
#   2. 고등학교 중퇴, 졸업
#   3. 전문대학 재학, 중퇴, 졸업
#   4. 대학교(4년제) 재학, 중퇴, 졸업
#   5. 대학원 이상
# p1003_9 : 직업 만족도
#   1. 매우 불만족 2. 대체로 만족 3. 그저그렇다. 4. 대체로 만족. 5. 매우만족

Data3 <- originalData

Data3_Income <- Data3 %>%
  rename(avgMonthIncome = p1002_8aq1,
         finalEdu= p1007_3aq1) %>%
  filter(!is.na(finalEdu), !is.na(avgMonthIncome)) %>%
  select(avgMonthIncome, finalEdu) %>%
  group_by(finalEdu) %>%
  summarise(avgIncome = round(mean(avgMonthIncome), 2),
            maxIncome = max(avgMonthIncome),
            minIncome = min(avgMonthIncome))

Data3_Income$finalEdu <- ifelse(Data3_Income$finalEdu == 5, 'Very High',
                               ifelse(Data3_Income$finalEdu == 4, 'High',
                                       ifelse(Data3_Income$finalEdu == 3, 'Middle',
```

```
ifelse(Data3_Income$finalEdu == 2, 'Low', 'Very Low'))))

View(Data3_Income)
```

코드 설명

filter함수를 이용해 finalEdu와 avgMonthIncome의 결측치를 제거해 주었다. 위 코드를 통해 변수에 최종 학력에 따른 한 달간의 평균 임금과 최고 임금, 최저 임금 데이터를 담았다.

결과

	finalEdu	avgIncome	maxIncome	minIncome
1	Very Low	65.00	70.0	60
2	Low	221.30	1000.0	42
3	Middle	200.25	335.4	110
4	High	262.09	845.6	30
5	Very High	252.21	428.0	145

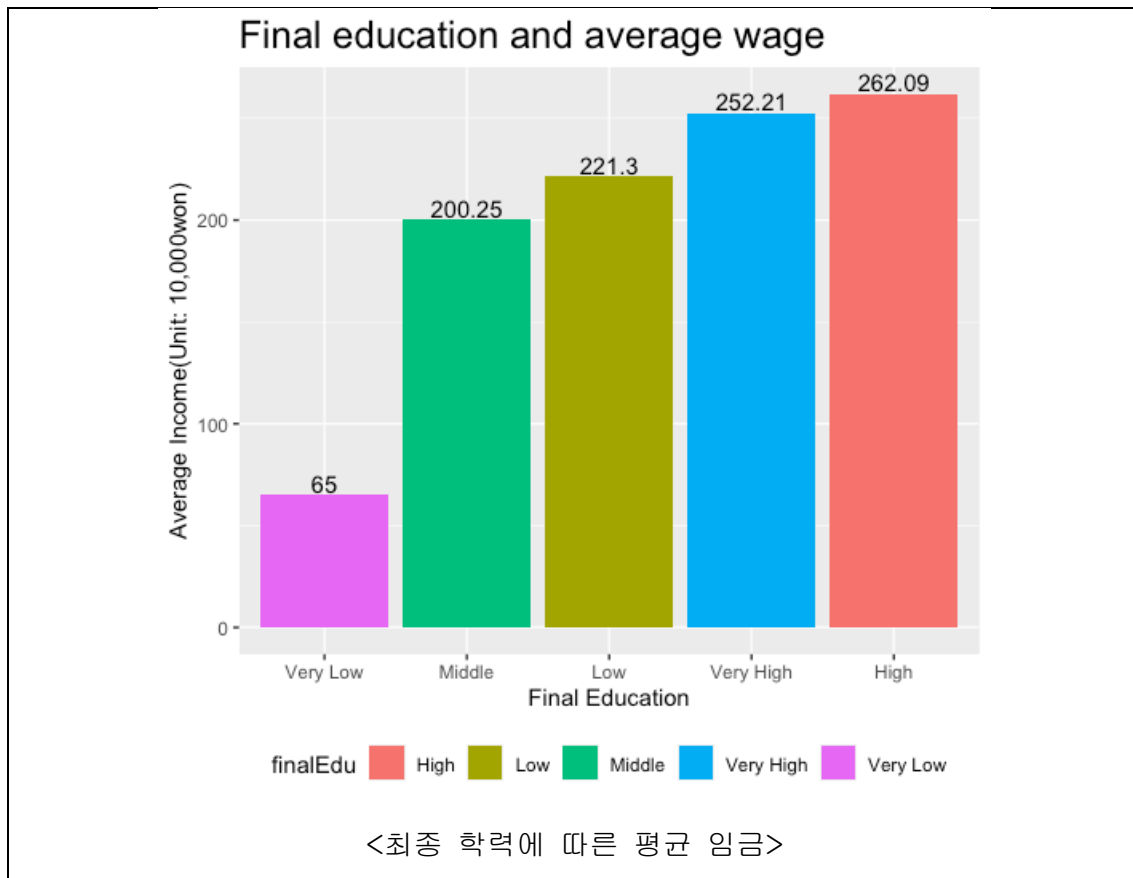
<Data3_Income의 Table>

최종 학력별로 평균 임금, 최고 많은 임금, 최고 적은 임금의 데이터가 담겼다.

코드

```
# 최종학력과 평균 임금
Data3_Income %>%
  ggplot(aes(reorder(finalEdu, avgIncome), avgIncome, label = avgIncome)) +
  geom_col(aes(fill = finalEdu)) +
  geom_text(vjust = -0.2, color = 'black') +
  ggtitle('Final education and average wage') +
  xlab('Final Education') + ylab('Average Income(Unit: 10,000won)') +
  theme(plot.title = element_text(size=18), legend.position = 'bottom')
```

결과



임금이 가장 높은 학력은 High(대학교(4년제) 재학, 중퇴, 졸업)로 2,620,000원으로 가장 높으며 다음으로 높은 학력은 Very High(대학원 이상) 2,522,100원으로 높다. 세번째로 높은 평균 임금은 Low(고등학교 중퇴, 졸업) 2,213,000원이며 네번째로 높은 평균 임금은 Middle(전문대학 재학, 중퇴, 졸업) 2,002,500원 이다. 그리고 가장 적은 평균 임금을 받는 최종 학력은 Very Low(중학교 졸업 이하) 650,000원으로 나타났다.

코드

```
# 최종학력과 최고 임금
Data3_Income %>%
  ggplot(aes(reorder(finalEdu, maxIncome), maxIncome, label = maxIncome)) +
  geom_col(aes(fill = finalEdu)) +
```

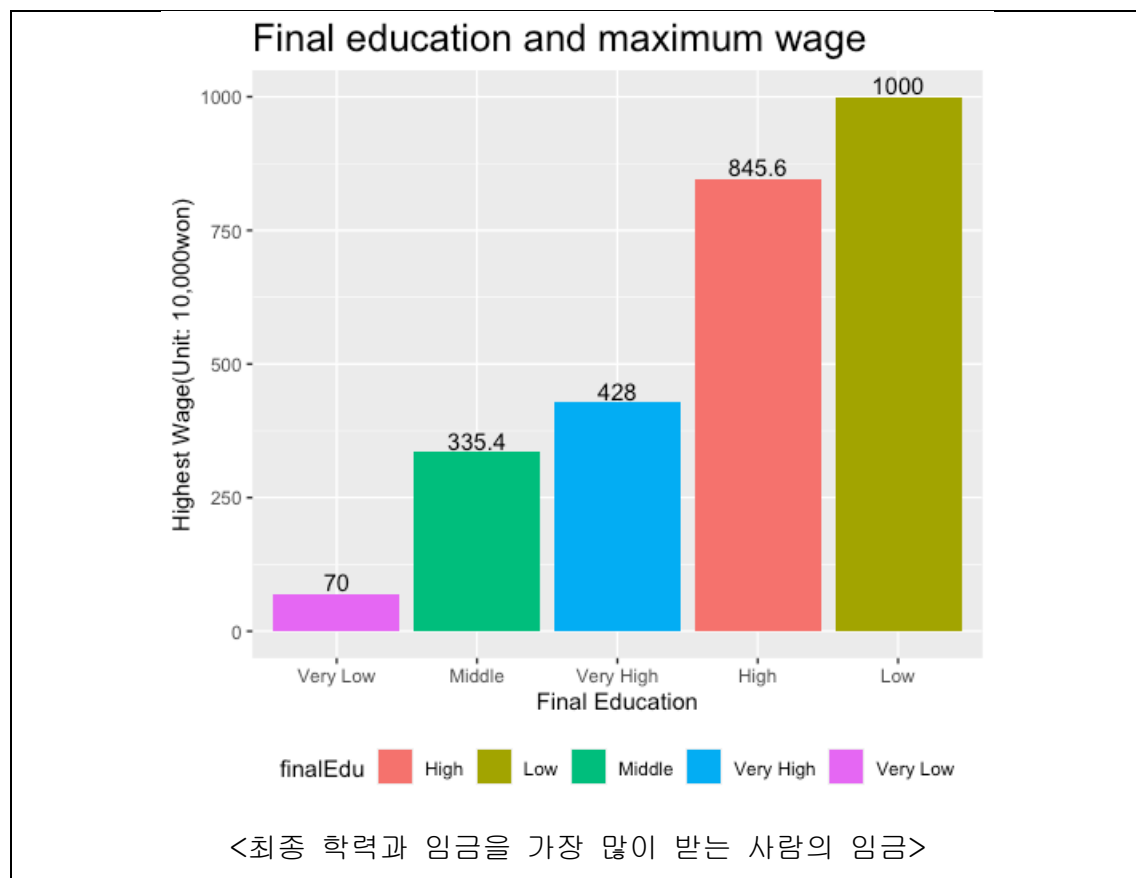


```
geom_text(vjust = -0.2, color = 'black') +
ggtitle('Final education and maximum wage') +
xlab('Final Education') + ylab('Highest Wage(Unit: 10,000won)') +
theme(plot.title = element_text(size=18), legend.position = 'bottom')
```

코드 설명

최종 학력과 그에 따른 가장 높은 임금을 그래프로 나타냈다.

결과

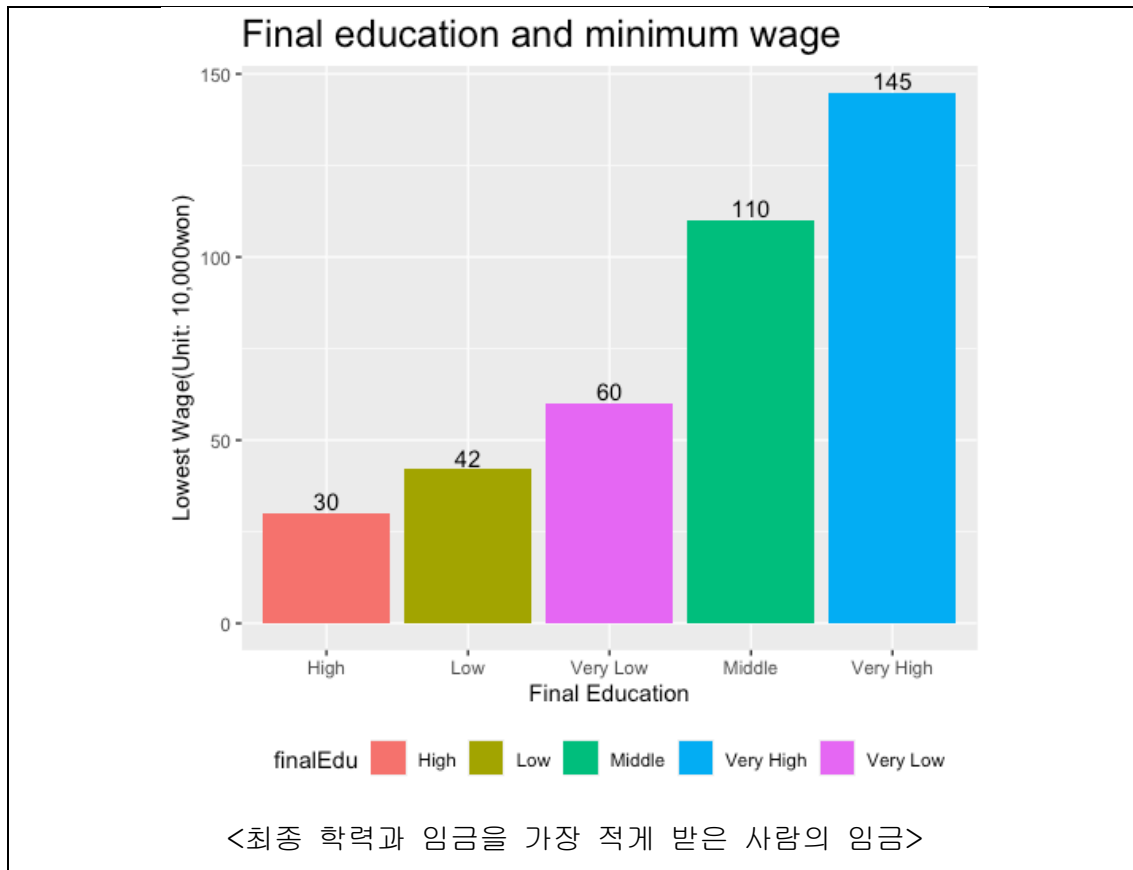


가장 높은 임금을 받은 학력은 Low로 10,000,000원으로 나타났으며, 그 다음으로 높은 임금을 받은 학력은 High로 8,456,000원, 그 다음으로 Very High로 4,280,000원 그 다음으로 Middle이 3,350,000원 그리고 마지막으로 Very Low가 700,000원을 받는 것으로 나타났다.

코드

```
# 최종학과와 최저 임금
Data3_Income %>%
  ggplot(aes(reorder(finalEdu, minIncome), minIncome, label = minIncome)) +
  geom_col(aes(fill = finalEdu)) +
  geom_text(vjust = -0.2, color = 'black') +
  ggtitle('Final education and minimum wage') +
  xlab('Final Education') + ylab('Lowest Wage(Unit: 10,000won)') +
  theme(plot.title = element_text(size=18), legend.position = 'bottom')
```

결과



가장 적은 임금을 받은 학력은 High가 300,000원 그 다음으로 Low가 420,000원 그 다음으로 Very Low가 600,000원 그 다음으로 Middle이 1,100,000원으로 나타났다. 그리고 Very High가 1,450,000원으로 가장 적게 받은 사람의 임금 중에서 가장 높은 임금을 받았다.

제 2 절 직업 만족도 분석

코드

```
job_Satisfaction <- Data3 %>%
  select(p1007_3aq1, p1003_9) %>%
  rename(finalEdu = p1007_3aq1,
         Satisfaction = p1003_9) %>%
  filter(!is.na(finalEdu) & !is.na(Satisfaction)) %>%
  group_by(finalEdu, Satisfaction) %>%
  summarise(Count = n())

View(job_Satisfaction)

table(job_Satisfaction$Satisfaction)

job_Satisfaction$finalEdu <- ifelse(job_Satisfaction$finalEdu == 5, 'Very High',
                                   ifelse(job_Satisfaction$finalEdu == 4, 'High',
                                           ifelse(job_Satisfaction$finalEdu == 3, 'Middle',
                                                  ifelse(job_Satisfaction$finalEdu == 2, 'Low', 'Very Low'))))

job_Satisfaction$Satisfaction <- ifelse(job_Satisfaction$Satisfaction == 5, 'A',
                                       ifelse(job_Satisfaction$Satisfaction == 4, 'B',
                                               ifelse(job_Satisfaction$Satisfaction == 3, 'C',
                                                      ifelse(job_Satisfaction$Satisfaction == 2, 'E', 'F'))))

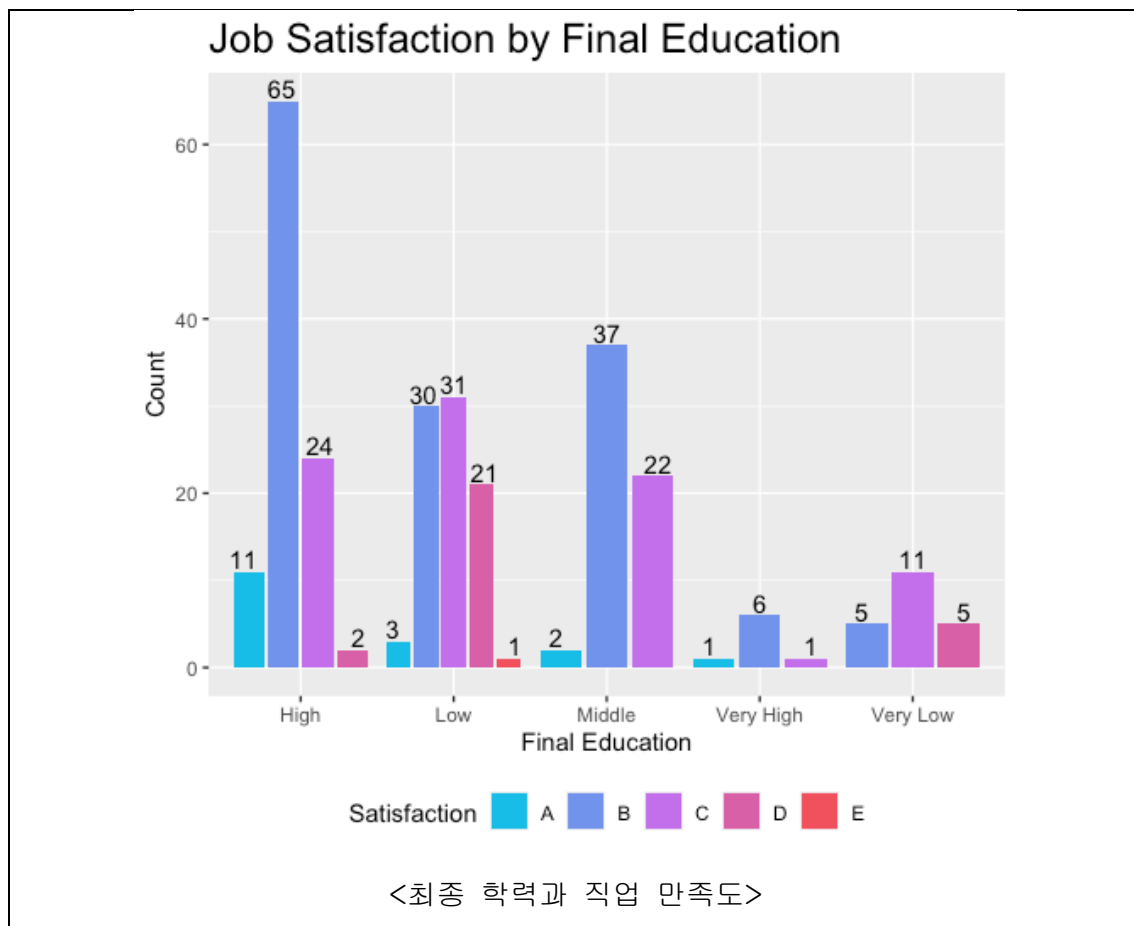
job_Satisfaction %>%
  ggplot(aes(finalEdu, Count, fill = Satisfaction, label = Count)) +
  geom_col(position = 'dodge2') +
  geom_text(vjust = -0.2, color = 'black', position = position_dodge(width = 1))
+
  xlab('Final Education') + ylab('Count') +
```

```
ggtitle('Job Satisfaction by Final Education') +
  theme(plot.title = element_text(size=18), legend.position = 'bottom') +
  scale_fill_manual(values = c( "#18bfe9", "#7595eb", "#c371ed",
                                "#db62a9", "#f4505e"))
```

코드 설명

ifelse 함수를 이용하여 최종 학력 변수의 1~5번 데이터 값을 Very Low, Low, Middle, High, Very High의 데이터 값으로 변환해 주었으며, 직업 만족도 변수의 1~5의 Categorical 변수를 D~A의 데이터 값으로 변환해 주었다.

결과



직업 만족도를 그래프로 나타낸 모습이다. 데이터 양의 불균형이 있으므로 비율로 그래프를 만들어 확인할 필요가 있다.

코드

```
job_Satisfaction <- job_Satisfaction %>%
  group_by(finalEdu) %>%
  mutate(Total_People = sum(Count))

job_Satisfaction <- job_Satisfaction %>%
  mutate(Rate = round(Count/Total_People, 2))

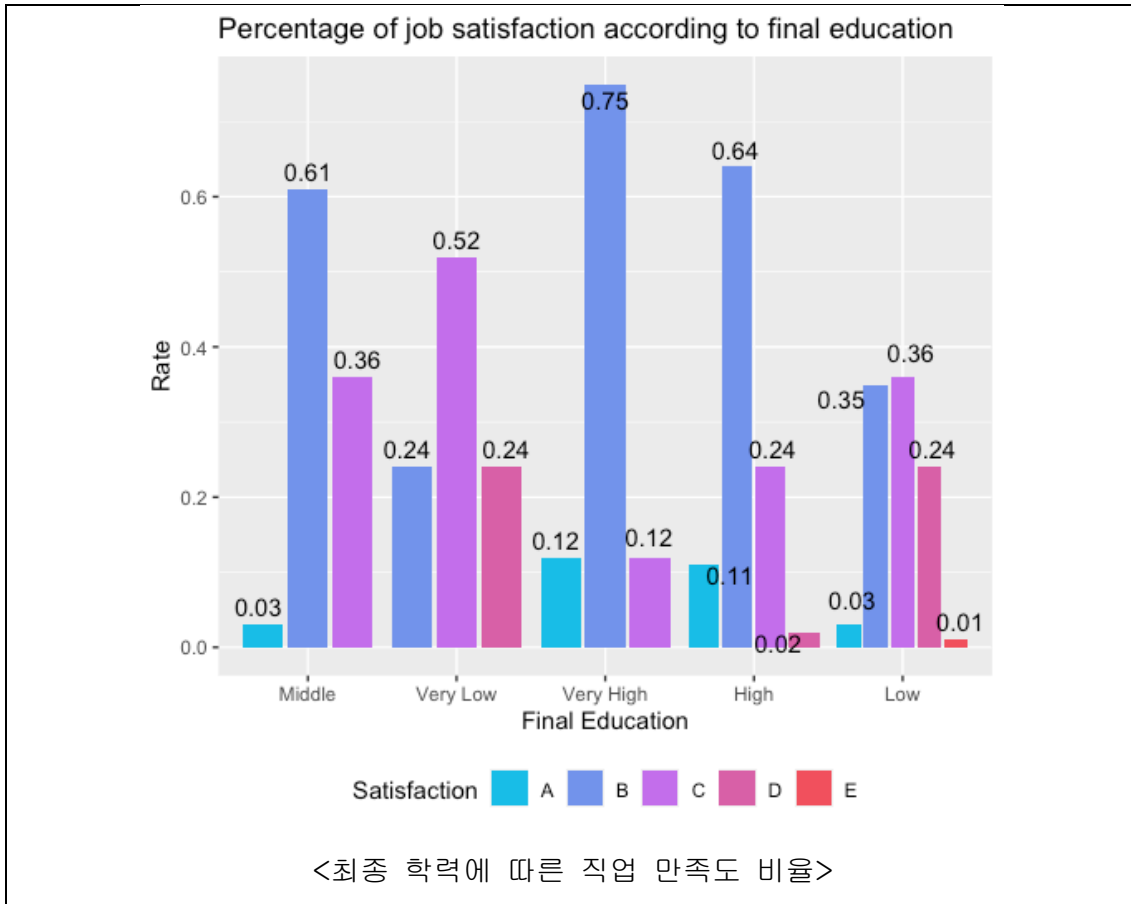
job_Satisfaction

# 만족도 비율 보기
install.packages('ggrepel')
library('ggrepel')

job_Satisfaction %>%
  ggplot(aes(reorder(finalEdu, -Rate), Rate, fill = Satisfaction, label = Rate)) +
  geom_col(position = 'dodge2') +
  geom_text_repel(hjust = 0.5, vjust = -0.5, position = position_dodge(width =
1)) +
  xlab('Final Education') + ylab('Rate') +
  ggtitle('Percentage of job satisfaction according to final education') +
  theme(plot.title = element_text(size=13), legend.position = 'bottom') +
  scale_fill_manual(values = c( "#18bfe9", "#7595eb", "#c371ed",
                                "#db62a9", "#f4505e"))
```

같은 학력에 해당하는 전체 사람 수를 구한 후 mutate함수를 이용하여 Total_People변수에 전체 사람 수의 데이터를 넣었다. 그리고 같은 학력의 전체 사람수와 각 만족도에 해당하는 사람 수를 나누어주어 비율을 나타내는 Rate 변수를 추가해주었다. ggplot의 geom_text함수를 사용하면 label이 겹치거나 삭제되는 문제가 발생한다. 따라서 ggrepel패키지를 설치하여 label이 겹치거나 삭제되는 문제를 해결해주었다.

결과



제 3 절 결론

<최종 학력에 따른 평균 임금> 그래프에서 알 수 있 듯이 최종 학력이 높을 수록 평균 임금이 높은 것을 확인할 수 있다. 하지만 가장 많은 임금을 받는 학력은 Low(고등학교 중퇴, 졸업)가 나왔다. <최종 학력과 임금을 가장 적게 받은 사람의 임금>의 그래프를 보면 High가 300,000만원으로 높은 학력에도 불구하고 매우 적은 임금을 받고 있음을 확인할 수 있다.

<최종 학력에 따른 직업 만족도 비율> 그래프를 보면 Very Low의 해당하는 학력은 만족도가 A가 없다. Very High의 학력을 보면 D와 E의 만족도가 없으며, A와 B의 비율이 0.87로 가장 높은 직업 만족도를 가지고 있다. High의 경우에는 A와 B의 비율이 0.75로 두번째로 높은 직업 만족도를 가지고 있다.

이로 인해 알 수 있는 것은, 학력이 높아질 수록 평균 임금이 높아지는 것은 맞

으나, 임금은 자신의 직업에 따라 최종 학력과 상관 없이 높은 임금을 받을 수도 낮은 임금을 받을 수도 있다는 것이다. 또한 직업 만족도의 경우도 자신의 직업에 따라 만족도가 달라지지만, 최종 학력이 고학력일수록 평균적으로 직업 만족도는 높아지고, 저학력일수록 평균적으로 직업 만족도는 낮아지는 것을 확인할 수 있다.