**|목차|**

**Ⅰ. 자료 설명…………………………………………………………………………3**

**Ⅱ. 데이터 시각화….……………………………………………………………..4**

1. 히트맵…………………………………………………………………………………………………………………………….4
2. 막대그래프……………………………………………………………………………………………………………………..4
3. 박스 플롯…………….……………………………………………………………………………………………………….….5
4. 원그래프………………………………………………………………………………………………………………………….8

**Ⅲ. 결론………………………………………………………………………………….9**

**Ⅳ. 부록…………………………………………………………………………………10**

**|자료 설명|**

본 보고서에는 학과와 입학전형에 따른 영어, 수학 시험 결과 정보가 담겨있는 데이터 셋을 사용했다. 이 데이터 셋은 424개의 데이터와 7개의 특징들을 가지고 있다. 다음은 R 프로그램의 summary 함수를 사용하여 수치적 정보를 요약한 것이다.

‘텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

(학과와 입학전형에 따른 점수 데이터 summary함수 출력 결과 정리)

**텍스트, 스크린샷, 직사각형, 사각형이(가) 표시된 사진

자동 생성된 설명**위 데이터 셋을 가지고 자료의 경향, 성격을 여러 그래프를 통해 파악해볼 것이다. 그래프와 설명의 형식으로 작성되어 있으며 데이터 시각화 부분에서는 짧은 해석을, 결론부분에서 시각화의 결과를 전체적으로 분석한다. 아래는 네가지 점수별 상관관계를 보여주는 히트맵이다.

(영어실력시험, 수학실력시험, 미적1평점, 미적2 평점 특성들에 관한 Heatmap)

**|데이터 시각화|**

**텍스트, 스크린샷, 직사각형, 사각형이(가) 표시된 사진

자동 생성된 설명<히트맵>**

(영어실력시험, 수학실력시험, 미적1평점, 미적2 평점 특성들에 관한 Heatmap)

**<막대 그래프>**

텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명

(정시, 수시 인원수 막대그래프)

**[설명]**

- 4개의 시험 점수에 관한 상관관계를 보기 위해 히트맵 출력.

**-** 미적1 평점과 미적 2 평점 사이의 상관관계가 높은 것으로 보아 둘중 하나의 점수가 높다면 다른 하나의 점수 또한 높을 것이라 예상 가능하다.

**[설명]**

- 정시와 수시 인원 수를 확인하기 위해 출력

- 수시 학생이 정시학생 보다 두배 많다.

**<막대그래프>**

텍스트, 스크린샷, 도표, 디자인이(가) 표시된 사진

자동 생성된 설명

(학과별 인원수 막대그래프)

**<박스플롯>**

텍스트, 도표, 스크린샷, 평면도이(가) 표시된 사진

자동 생성된 설명

(수시와 정시의 시험 점수 분포 박스 플롯)

**[설명]**

- 학과별 인원 수를 확인 하기 위해 출력

- D와C학과의 학생이 타과생보다 적게는 두배 많게는 세배 많다.

**[설명]**

- 입학시기별 시험점수 경향을 비교하기 위해 박스플롯 출력

- 정시로 입학한 사람과 수시로 입학한 사람들 중 대체로 정시로 입학한 사람의 성취도가 높게 나온다.

**< 박스 플롯\_2-1>**

도표, 스크린샷, 다채로움, 텍스트이(가) 표시된 사진

자동 생성된 설명

(학과별 영어실력시험 점수 박스플롯)

**< 박스 플롯\_2-2>**

도표, 텍스트, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명

(학과별 수학실력시험 점수 박스플롯)

**[설명]**

- 학과별 영어실력시험 점수의 경향을 보기 위해 출력

- A학과의 성취도가 제일 낮으며 E학과의 성취도가 가장 높다.

**[설명]**

- 학과별 수학실력시험 점수의 경향을 보기 위해 출력

- 대체로 비슷한 수준의 성취를 보이고 있다.

**<박스 플롯\_2-3>**

도표, 사각형, 다채로움, 스크린샷이(가) 표시된 사진

자동 생성된 설명

(학과별 미적1 평점 박스플롯)

**<박스 플롯\_2-4>**

도표, 다채로움, 사각형, 스크린샷이(가) 표시된 사진

자동 생성된 설명

(학과별 미적2평점 박스플롯)

**[설명]**

- 학과별 미적 1 평점경향을 보기 위해 출력

- 대체로 비슷한 수준의 성취를 보이고 있으나 B학과가 조금 낮은 성취를 기록함

**[설명]**

- 학과별 미적 2 평점경향을 보기 위해 출력

- C, D, E 학과가 나머지 두 학과보다 높은 성취를 기록함

**<박스 플롯\_2-5>**도표, 텍스트, 다채로움, 라인이(가) 표시된 사진

자동 생성된 설명

(학과별 점수 박스플롯)

**<원그래프>**

텍스트, 도표, 원, 스크린샷이(가) 표시된 사진

자동 생성된 설명

(학과별 정시, 수시 인원 비율)

**[설명]**

- 학과별 시험 점수 경향을 한눈에 보고 비교하기 위해 박스플롯 출력

- 대체로 C, D, E 학과의 성취도가 나머지 두 학과 보다 높게 나타난다.

**[설명]**

- 학과별 정시와 수시 인원 비율을 확인하기 위해 출력

- 정시인원의 비율이 높으면 시험 성취도가 높을 것이라는 가설을 가지고 확인하기 위해 비율 출력했는데 대부분 30~35퍼센트의 정시 인원을 가지고 있음

-예상과 달리 평균 성취가 낮은 A학과의 정시생 비율이 높게 나옴

**|결론|**

본 과제에서 학과와 입학전형에 따른 영어, 수학 시험 결과 정보가 담겨있는 데이터 셋으로 여러가지 정보를 시각화 해보았다. 단 학번에 대한 데이터는 학과별로 고르게 분포해 있는 것이 아닌 특정 범위에 따라 학과가 나뉘어 있어서 의미 없는 정보라 판단해 분석에 사용하지 않았다. 진행 방향은 ‘공부를 잘 하는 학생들은 어떤 특징에 속해 있는지’ 에 중점을 두고 분석해보았다. 먼저 히트맵을 통해 영어실력 시험, 수학실력시험, 미적1, 미적2 특징들 간의 상관성을 확인하여 미적1과2사이의 상관성이 높다는 것을 확인했다. 수학실력실험 점수 또한 미적1, 미적2의 점수와 높은 상관성을 기록할 것이라 예상 했으나 낮은 상관성을 보여주었다. 다음으로 데이터의 정시와 수시인원 수를 을 확인하기 위해 막대그래프를 출력했다. 정시인원과 수시 인원사이의 점수 관계를 확인하고자 박스 플롯을 출력했고 대체로 정시입학생의 성취도가 높다는 것을 알 수 있었다. 그리고 특정 학과에 속해 있으면 공부를 잘 한다라는 가설을 세우고 학과별 점수 박스플롯을 출력했다. 그 결과 성취도가 대부분 비슷하지만 그나마 E 학과 학생들이 공부를 잘 한다는 정보를 얻을 수 있었다. 그 이유가 정시생 비율이 높기 때문일 것이라는 예상을 가지고 학과별 정시, 수시생 비율을 원그래프로 출력해보았으나 정시생 비율이 모든 학과에서 비슷하게 나옴으로 위 추측은 틀렸다 라는 것을 확인 할 수 있었다.

결론적으로 수시생보다 정시생들의 시험 점수가 높고 E 학과에 속한 학생들이 공부를 잘한다. 그러나 학과와 정시생 사이에는 상관성이 없다는 결론을 도출했다.

**|부록|**

**- 사용한 R스크립트**

library(openxlsx)

library(ggplot2)

library(gridExtra)

library(gplots)

library(dplyr)

library(reshape2)

# 엑셀 데이터 가져오기

data <- read.xlsx("C:/기말고사.xlsx")

summary(data)

data\_df <- as.data.frame(data)

# 히트맵 생성

data\_without\_some <- data\_df |>

select(-학번, -학과, -입학시기)

correlation\_matrix <- cor(data\_without\_some)

cor\_df <- melt(correlation\_matrix)

heatmap\_plot <- ggplot(data = cor\_df, aes(x = Var1, y = Var2, fill = value,label = round(value, 2))) +

geom\_tile() +

geom\_text(color = "black", size = 3) +

scale\_fill\_gradient(low = "white", high = "red") +

labs(title = "Heatmap of Features", x = "Features", y = "Features") +

theme\_minimal() +

theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust = 1))

print(heatmap\_plot)

# "정시"와 "수시" 인원 수 비교 막대그래프프

frequency <- table(data\_df$입학시기)

frequency\_df <- as.data.frame(frequency)

names(frequency\_df) <- c("입학시기", "빈도")

ggplot(frequency\_df, aes(x = 입학시기, y = 빈도)) +

geom\_bar(stat = "identity", fill = "skyblue", width = 0.5) +

geom\_text(aes(label = 빈도), vjust = -0.5, size = 3.5, color = "black") +

labs(x = "입학시기", y = "인원 수", title = "정시 vs 수시 인원 수") +

theme\_minimal()

# "정시"와 "수시" 인원 수 비교 막대그래프프

frequency <- table(data\_df$학과)

frequency\_df <- as.data.frame(frequency)

names(frequency\_df) <- c("학과", "빈도")

ggplot(frequency\_df, aes(x = 학과, y = 빈도)) +

geom\_bar(stat = "identity", fill = "pink", width = 0.5) +

geom\_text(aes(label = 빈도), vjust = -0.5, size = 3.5, color = "black") +

labs(x = "학과", y = "인원 수", title = "학과별 인원 수") +

theme\_minimal()

# 입학시기별 시험 점수 박스플롯롯

data\_df$입학시기 <- factor(data\_df$입학시기)

# 박스플롯에"추가"라는 항목이 같이 출력되어서 그거 지워줌줌

data\_df <- data\_df[data\_df$입학시기 != "추가", ]

plot1 <- qplot(입학시기, 영어실력시험, data = data\_df, geom = "boxplot", fill = 입학시기) +

geom\_rug() +

labs(x = "입학시기", y = "영어실력시험") +

theme\_minimal()

plot2 <- qplot(입학시기, 수학실력시험, data = data\_df, geom = "boxplot", fill = 입학시기) +

geom\_rug() +

labs(x = "입학시기", y = "수학실력시험") +

theme\_minimal()

plot3 <- qplot(입학시기, 미적1평점, data = data\_df, geom = "boxplot", fill = 입학시기) +

geom\_rug() +

labs(x = "입학시기", y = "미적1평점") +

theme\_minimal()

plot4 <- qplot(입학시기, 미적2평점, data = data\_df, geom = "boxplot", fill = 입학시기) +

geom\_rug() +

labs(x = "입학시기", y = "미적2평점") +

theme\_minimal()

grid.arrange(plot1, plot2, plot3, plot4, ncol = 2)

# 학과별 영어시험점수 박스 플롯

qplot(학과, 영어실력시험, data = data\_df, geom = "boxplot", fill = 학과) +

geom\_rug() +

labs(x = "학과", y = "영어실력시험") +

theme\_minimal()

# 학과별 수학학시험점수 박스 플롯

qplot(학과, 수학실력시험, data = data\_df, geom = "boxplot", fill = 학과) +

geom\_rug() +

labs(x = "학과", y = "수학실력시험") +

theme\_minimal()

# 학과별 미적1평점 박스 플롯

qplot(학과, 미적1평점, data = data\_df, geom = "boxplot", fill = 학과) +

geom\_rug() +

labs(x = "학과", y = "미적1평점") +

theme\_minimal()

# 학과별 미적2평점 박스 플롯

qplot(학과, 미적2평점, data = data\_df, geom = "boxplot", fill = 학과) +

geom\_rug() +

labs(x = "학과", y = "미적2평점") +

theme\_minimal()

# 학과별 시험점수 박스플롯

data\_df$학과 <- factor(data\_df$학과)

plot1 <- qplot(학과, 영어실력시험, data = data\_df, geom = "boxplot", fill = 학과) +

geom\_rug() +

labs(x = "학과", y = "영어실력시험") +

theme\_minimal()

plot2 <- qplot(학과, 수학실력시험, data = data\_df, geom = "boxplot", fill = 학과) +

geom\_rug() +

labs(x = "학과", y = "수학실력시험") +

theme\_minimal()

plot3 <- qplot(학과, 미적1평점, data = data\_df, geom = "boxplot", fill = 학과) +

geom\_rug() +

labs(x = "학과", y = "미적1평점") +

theme\_minimal()

plot4 <- qplot(학과, 미적2평점, data = data\_df, geom = "boxplot", fill = 학과) +

geom\_rug() +

labs(x = "학과", y = "미적2평점") +

theme\_minimal()

grid.arrange(plot1, plot2, plot3, plot4, ncol = 2)

# 학과 정시, 수시 인원 비율율

summary\_df <- data\_df %>%

group\_by(학과, 입학시기) %>%

summarise(데이터\_개수 = n())

summary\_df <- summary\_df %>%

group\_by(학과) %>%

mutate(비율 = 데이터\_개수 / sum(데이터\_개수))

p <- ggplot(summary\_df, aes(x = "", y = 비율, fill = 입학시기)) +

geom\_bar(stat = "identity", width = 1) +

coord\_polar("y", start = 0) +

facet\_wrap(~학과) +

theme\_void() +

labs(title = "학과별 정시, 수시 인원 비율율")

p + geom\_text(aes(label = scales::percent(비율)), position = position\_stack(vjust = 0.5))

끝.