[2020 혁신성장 청년인재 집중양성 사업]

프로젝트 기반 데이터 과학자 양성과정

빅데이터 분석

- 4주차 -

#데이터의 이해 #시각화

© RedInfo inc.

A table of Contents

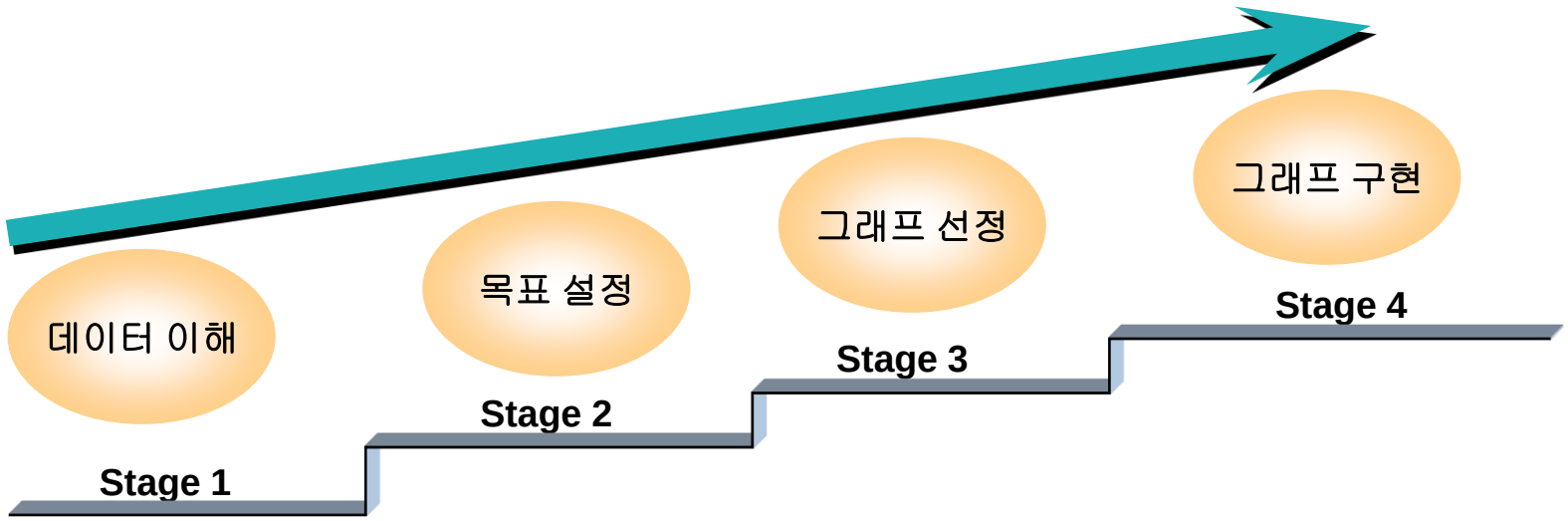
Part 1

시각화 개요

시각화 개요

시각화 목적

- 자료의 내재된 정보를 효과적인 그림으로 표현하는 게 목표
- 가공되지 않은 원천데이터로부터 정보를 추출하여 가시적으로 표현



시각화 개요

시각화 단계

- (1단계) 데이터 이해 : 데이터의 유형과 수집 기간, 그리고 데이터 내용 파악
- (2단계) 목표 설정 : 무엇을 알고 싶은지?
- (3단계) 그래프 선정 : 어떤 그래프가 좋을까?
- (4단계) 그래프 구현 : 핵심적인 의미를 담기 위한 옵션 선택과 그래프 구현

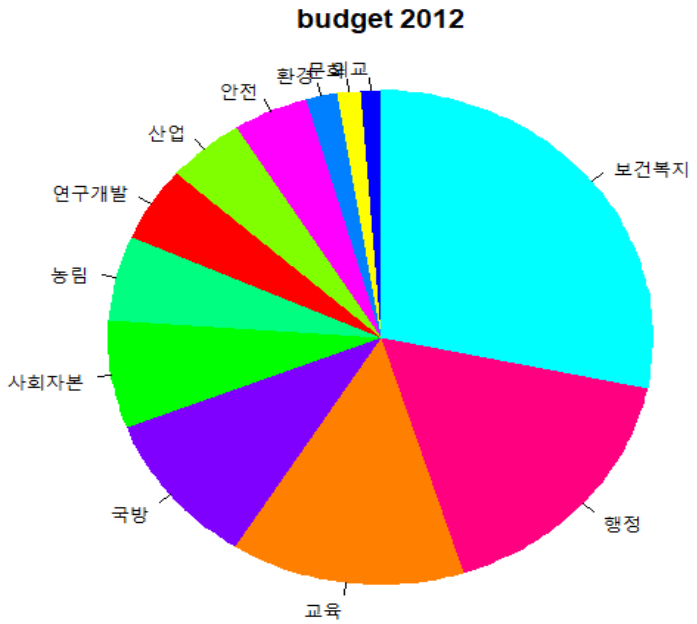
Part 2

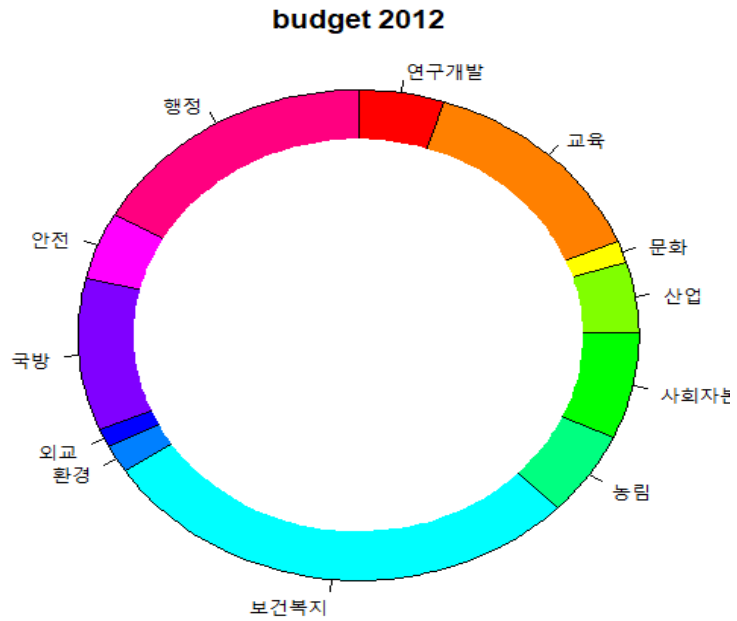
기본 그래프

파이 차트(Pie chart)

파이 차트 개요

- 전체에서 각 항목이 차지하는 비율을 파악하기 위한 그래프
- 데이터의 통계적인 정보를 그림의 형태로 나타내어 분포의 구성을 상대적으로 비교하는 데 유용
- 재학생의 연령별 구성, 선거에서 후보별 득표수 등

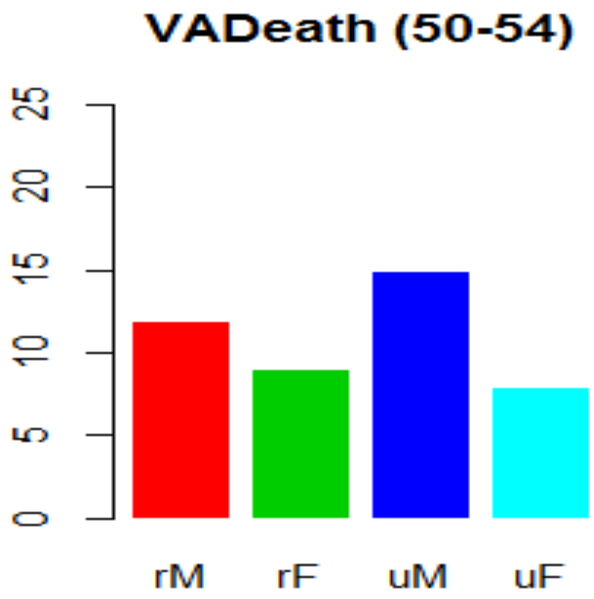


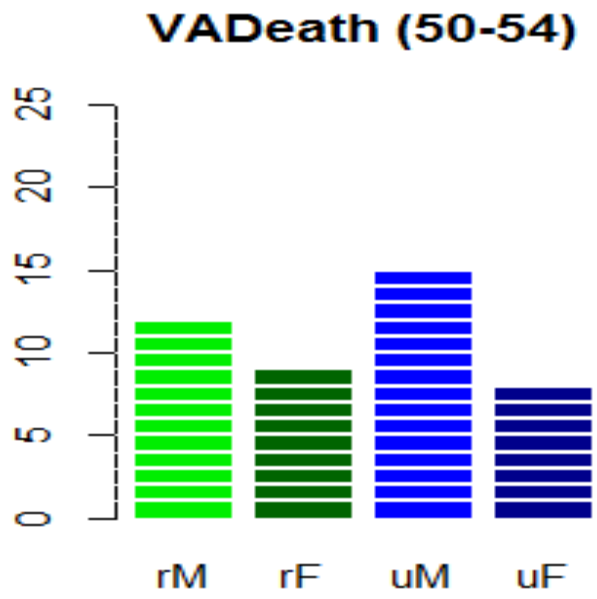


막대 그래프(bar plot)

막대 그래프 개요

- 여러 목적으로 아주 흔히 쓰이는 통계 그래프
- 어느 항목의 막대가 제일 긴지 보여줌
- 매체별 선호도, 재학생 연령별 분포 등

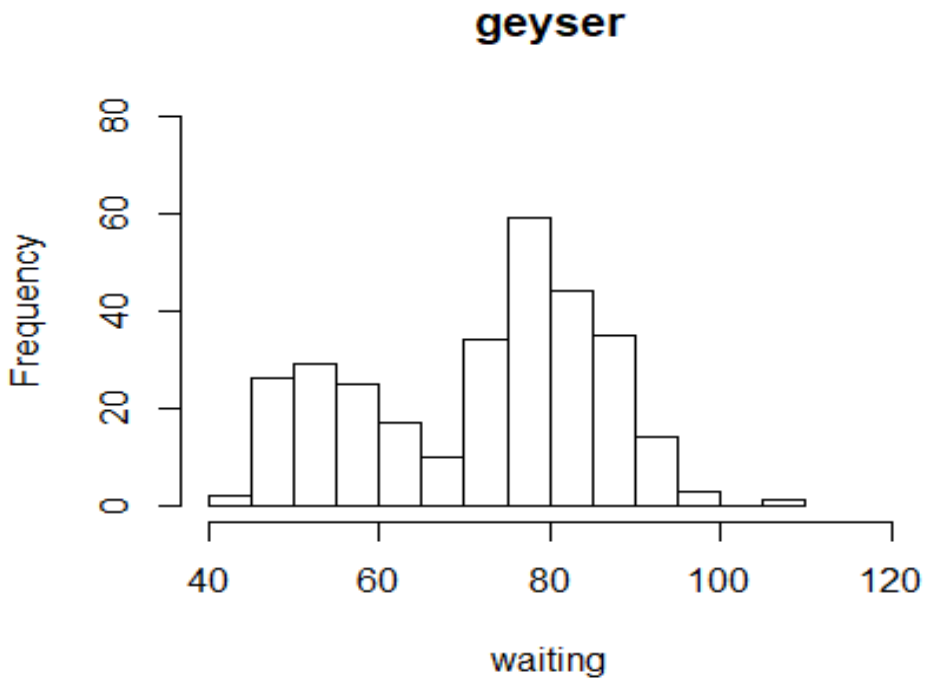


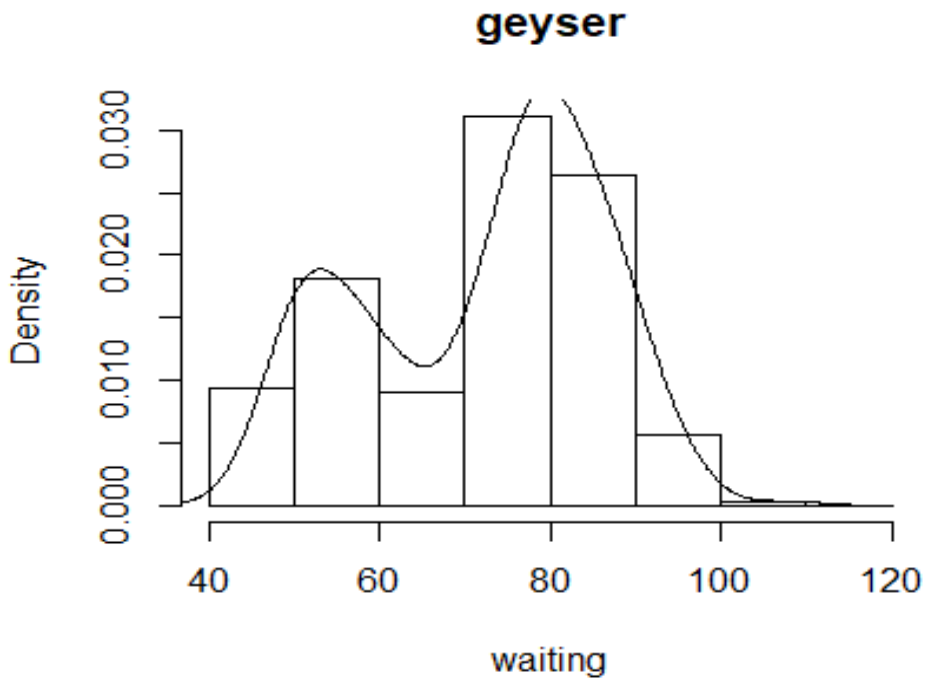


히스토그램(histogram)

히스토그램 개요

- 연속형 데이터의 구간별 도수를 상대적인 막대의 길이로 나타낸 그래프
- 분포를 파악하기 유리함

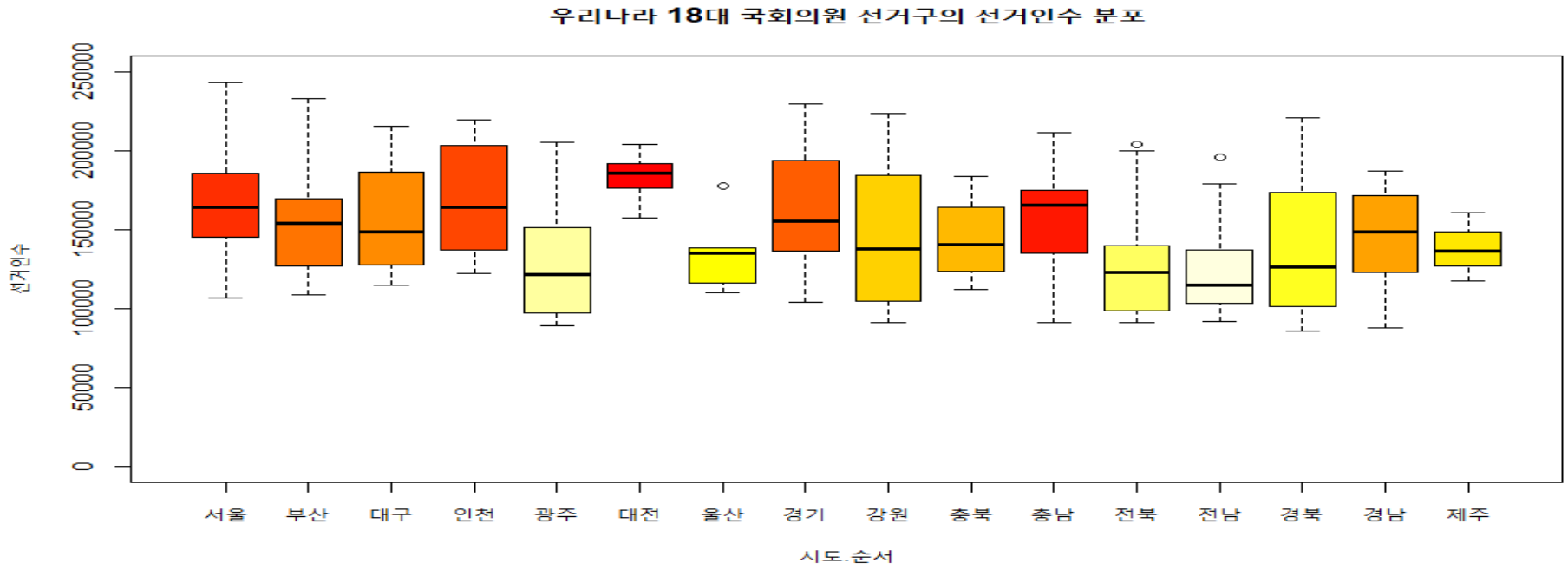




상자그림(Box plot)

상자그림 개요

- 사분위수와 중앙값으로 상자를 만들고 최대/최소에 선을 연결한 자료의 퍼짐 정도를 보고 분포를 파악함
- 상자의 중앙선은 중앙값(median), 상자의 위/아래 모서리는 3분위/1분위, 상자에 연결된 줄의 양끝은 최대/최소



줄기 잎 그림(Stem & Leaf plot)

줄기 잎 그림 개요

- 수치로 된 자료를 줄기와 잎으로 분류하여 자료의 분포를 파악

```
> stem(data, scale=1)

  The decimal point is 1 digit(s) to the right of the |

  12 | 133678
  13 | 4666779
  14 | 011589
  15 | 099
```

```
> stem(data, scale=2)

  The decimal point is 1 digit(s) to the right of the |

  12 | 133
  12 | 678
  13 | 4
  13 | 666779
  14 | 011
  14 | 589
  15 | 0
  15 | 99
```

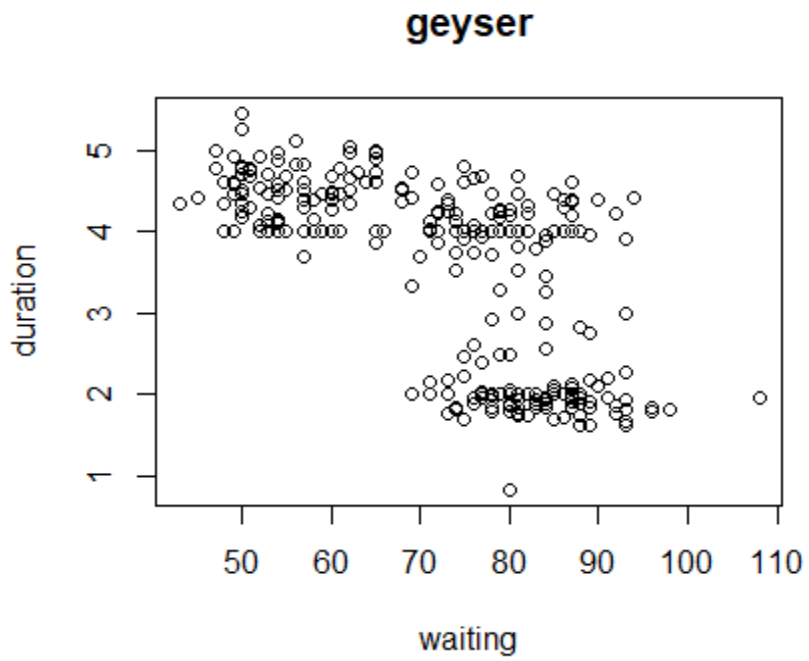
Part 3

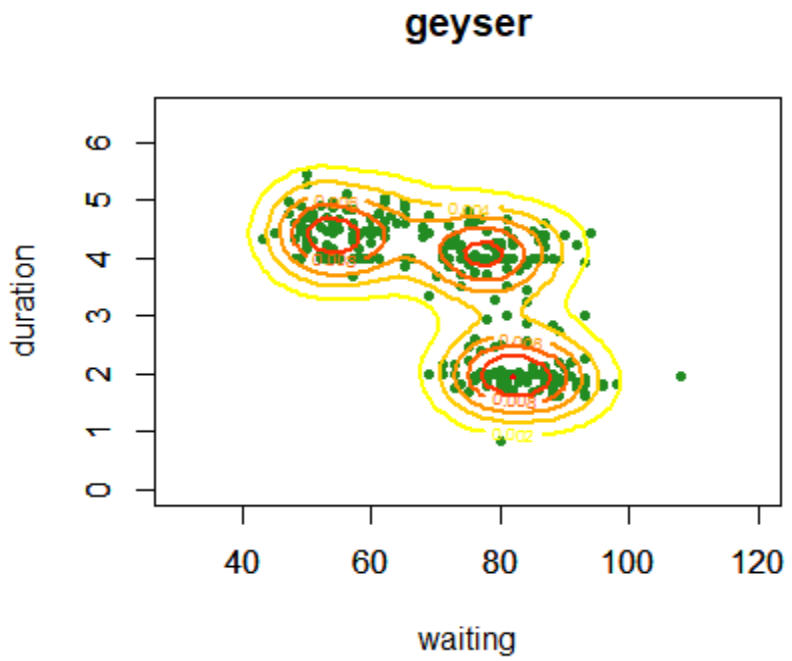
이변량 그래프

산점도(Scatter plot)

산점도 개요

- 이변량 연속형 자료를 2차원 평면에 넣은 그래프
- 가로와 세로의 비는 1:1
- 회귀적 관계

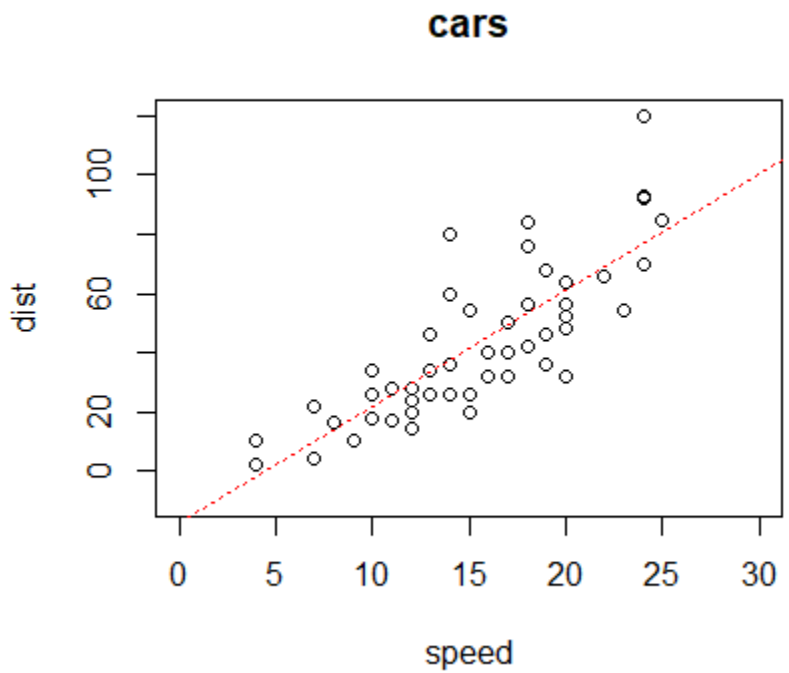


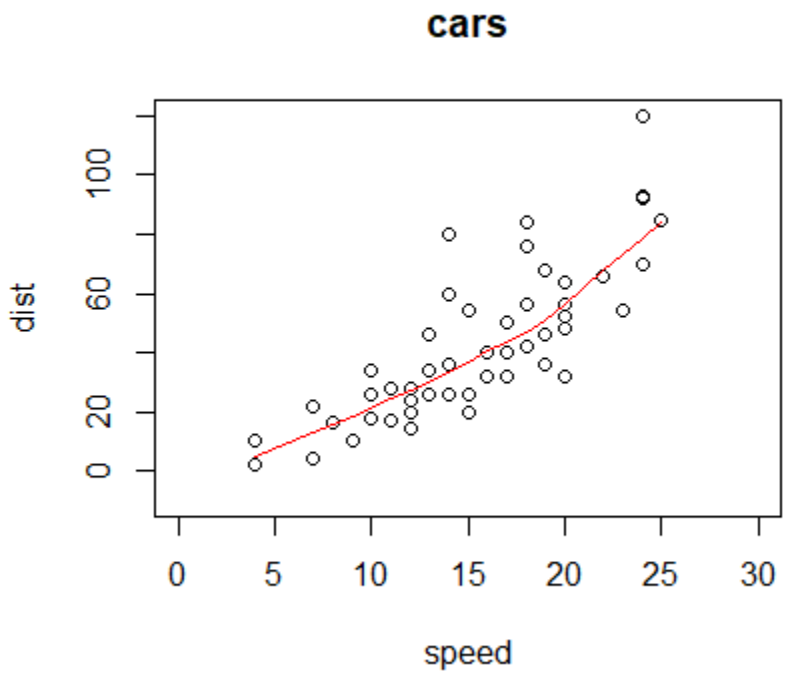


회귀모형

회귀모형 개요

- 설명변수와 반응변수가 각각 1개인 경우, 이변량 산점도로 회귀적인 관계를 시각화 할 수 있음

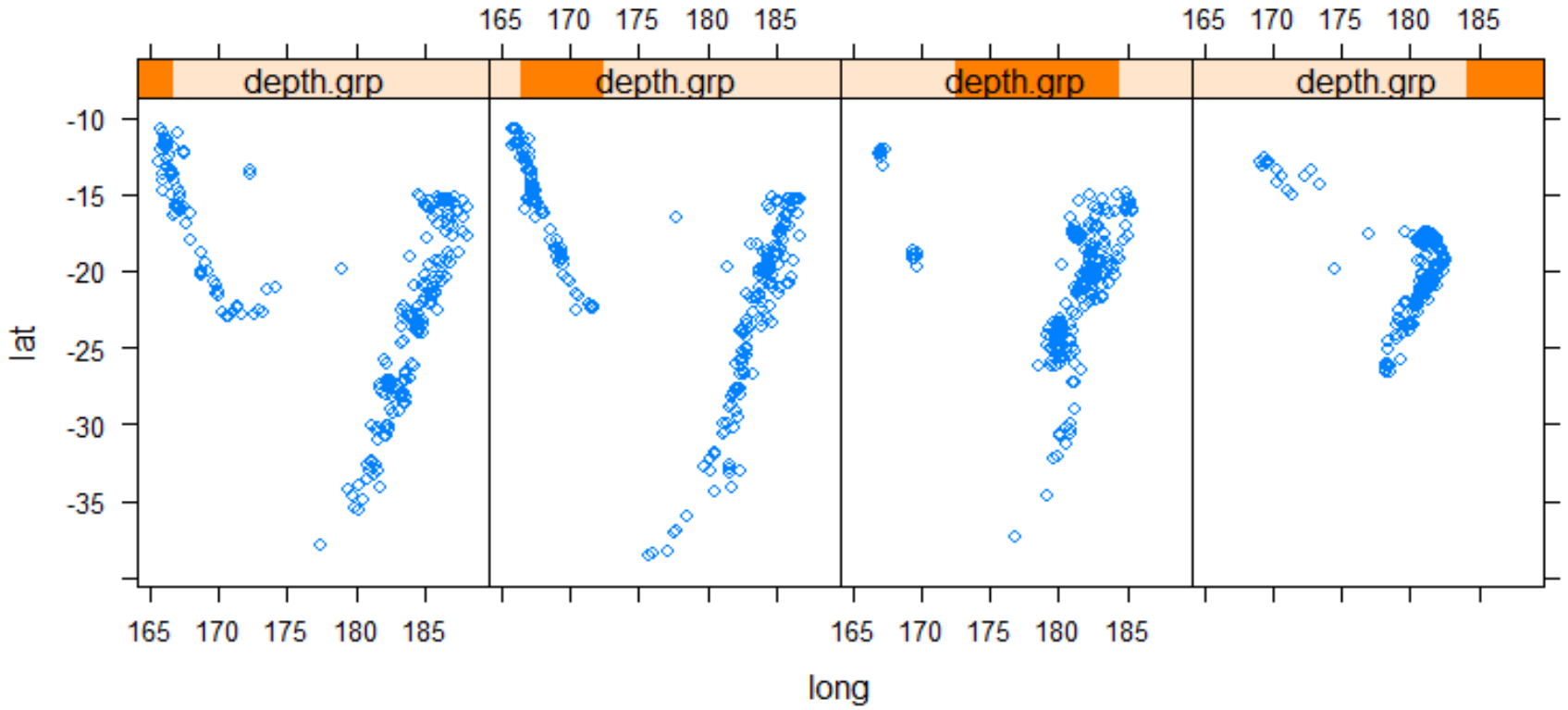




조건부 플랏(Conditioning plot)

조건부 플랏 개요

- 제 3의 변수의 수준에 따라 병렬되는 일련의 통계 그래프



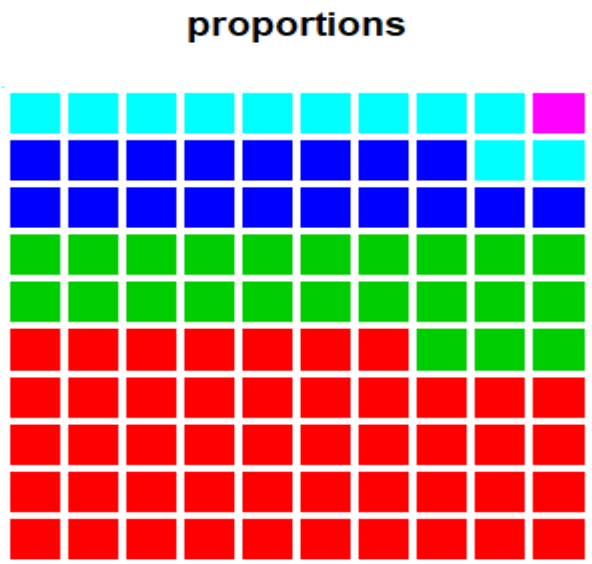
Part 4

유형별 그래프

사각 타일(Square tiles)

사각 타일

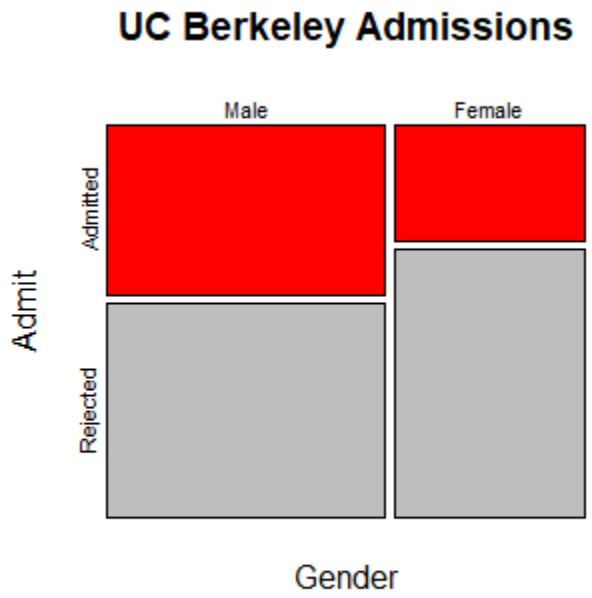
- 사각형 틀에 배열한 100개의 색 타일로 시각화
- 특정 속성의 구성 비율을 시각화

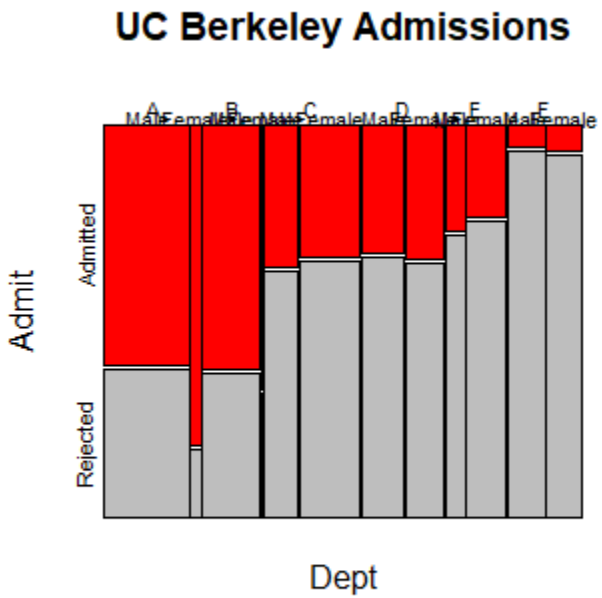


모자이크 그림(Mosaic plot)

모자이크 그림

- 2원 3원 교차표의 시각화
- 전체 정사각형 도형을 행 빈도에 비례하는 직사각 도형으로 나누고, 다시 각 도형을 행 내 열의 빈도에 해당하는 직사각도형으로 나눔





Part 5

ggplot 2

ggplot2 개요

ggplot2 문법

- 데이터를 이해하는 데 좋은 시각화 툴
- 문법 내에서 간단한 코드 추가/삭제가 가능

```
ggplot(...) + plot의 종류(eg. box plot, bar plot, scatter plot 등)<BR>
```

```
p + geom_point(color = 'red')
```

```
p + geom_point(aes(color = factor(gear)))
```

```
p + geom_line()
```

```
p + geom_line(aes(color = factor(gear)))
```

Part 6

Summary