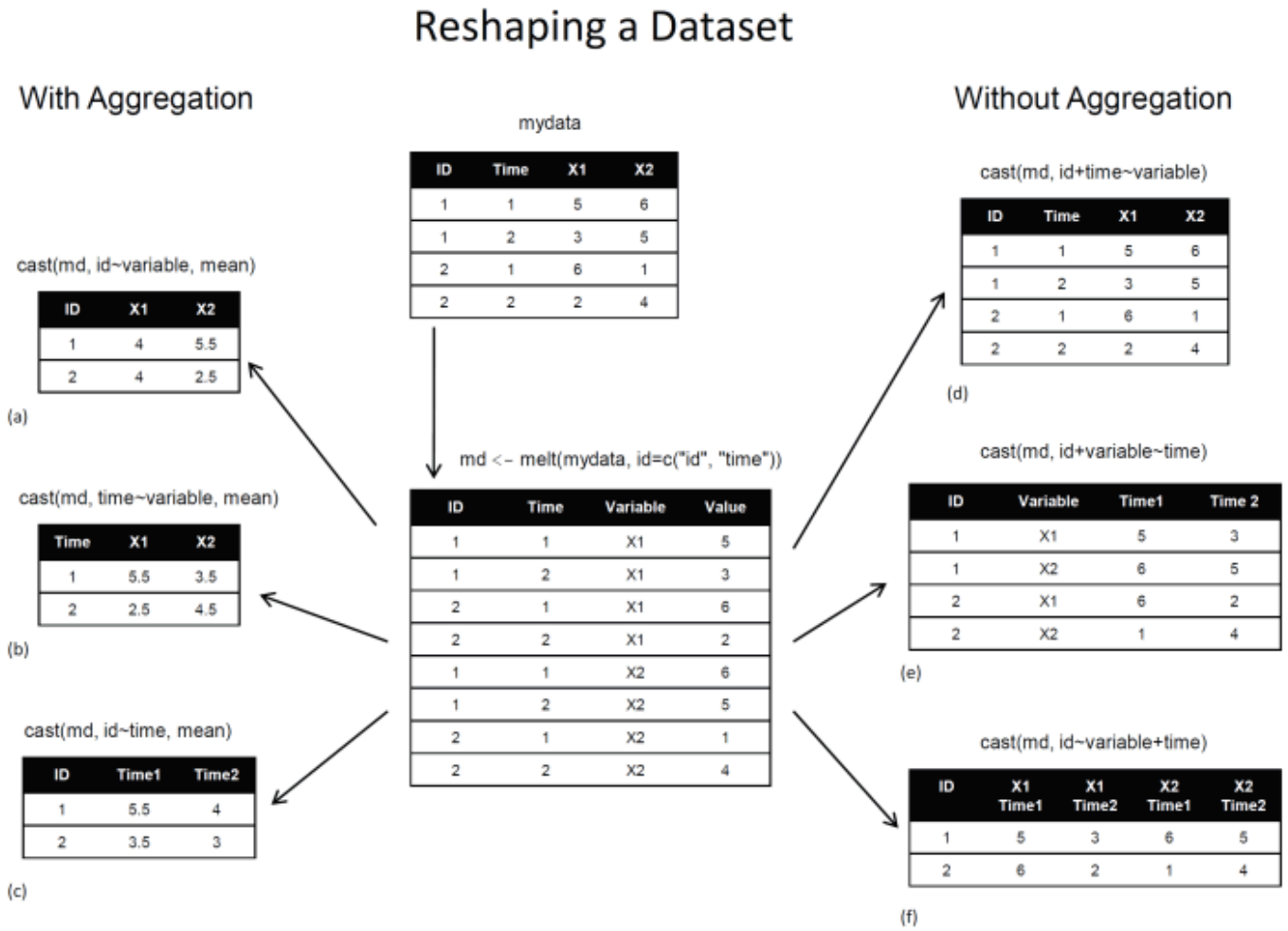
LAB 4-1

Data Manipulation

- MASS package의 Cars93 데이터를 읽고, 다음 문제를 다양한 방법으로 풀어 보세요.
  - data(Cars93, package="MASS")
  - 1) 현대(Hyundai) 자동차만 추출해 보세요.
  - 2) 현대 자동차를 높은 가격 순으로 정렬해보세요. (price 필드 기준)
  - 3) 현대 자동차 중 가장 비싼 자동차를 추출해 보세요.
  - 4) 가장 비싼 현대 자동차보다 비싼 다른 브랜드 자동차를 추출해 보세요. (가격순 정렬 / 제조사, 모델, 타입, 가격, 도시연비 MPG.city 정보만 추출)
  - 5) 4기통 자동차 중 도시연비(MPG.city)가 가장 좋은 차는? (제조사, 모델, 타입, 가격, 도시연비 정보만 추출)
  - 6) Cars93.num<-Cars93[,sapply(Cars93,is.numeric)]
  - 7) Cars93.fac<-Cars93[,sapply(Cars93,is.factor)]

Data Manipulation: reshape package

- **melt()**: restructuring where multiple categorical columns are 'melted' into unique rows
- cast(): Cast a molten data frame into the reshaped or aggregated form you want



Data Manipulation: reshape package

- Question
  - tips data에 관하여 다음 질문에 답해 보세요.
  - data(tips,package="reshape")
  - 1.어느 요일에 팁을 가장 많이 받았나요?
  - 2.요일별 평균적으로 받은 팁은 얼마인가요?
  - 3.남성고객과 여성 고객중 어느쪽이 팁에 후한편인가요?
  - 4.흡연 여부에 따른 성별 고객의 팁은?

Data Manipulation: sqldf Package

- With the sqldf package, you can use SQL to handle the data frame
  - library(sqldf); data(mtcars)
  - newdf <-sqldf("select * from mtcars where mpg>15 order by mpg",row.names=TRUE)
- SQL (Structured Query Language)
  - SELECT *attributes*
  - FROM *databases*
  - WHERE *conditions*
  - GROUP BY *attributes*
  - HAVING *conditions*
  - ORDER BY *attributes (DESC | ASC)*
- Question
  - mtcars 데이터에서 실린더가 4기통 또는 6기통 자동차의 Gear 별로 평균 mpg, disp를 구하시오.
    (SQL에서 평균은 avg 함수 이용)

Data Manipulation: dplyr Package

- **filter** – It filters the data based on a condition
- **select** – It is used to select columns of interest from a data set
- **arrange** – It is used to arrange data set values on ascending or descending order
- **mutate** – It is used to create new variables from existing variables
- **summarise** (with **group_by**) – It is used to perform analysis by commonly used operations such as min, max, mean count etc
- **%>%**: then

Data Manipulation: dplyr Package

- # filter (행추출) use *tips* data,
  - data(tips)str(tips)
  - tips %>% filter(sex=="Female",smoker=="Yes") %>% head
  - tips %>% filter(sex=="Male") %>% summarise(count=n())
- # select (열추출)
  - tips %>% select(starts_with("s")) %>% head
  - tips %>% select(contains("a")) %>% head
- # arrange(정렬)
  - tips %>% arrange(day,time) %>% head
  - tips %>% arrange(desc(total_bill)) %>% head

Data Manipulation: dplyr Package

- # group_by(묶기), summarise(요약)
  - tips %>% group_by(day) %>% summarise(sum.tip=sum(tip),mean.tip=mean(tip))
- # mutate(변수추가),transmute
  - tips %>% mutate(tip.bill=tip/total_bill) %>% head
  - tips %>% transmute(tip.bill=tip/total_bill) %>% head