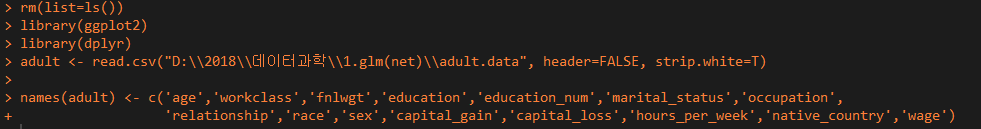
+ggplot2

Install.packages(‘ggplot2’)

Library(ggplot2)

#ggplot2설치가 안되면 install.packages("ggplot2", repos="https://cran.rstudio.com/")이 코드 실행



#adult데이터 불러오기

1. wage ~ age

Wage는 범주형, age는 연속형 변수이므로 적절한 시각화 방법은 boxplot

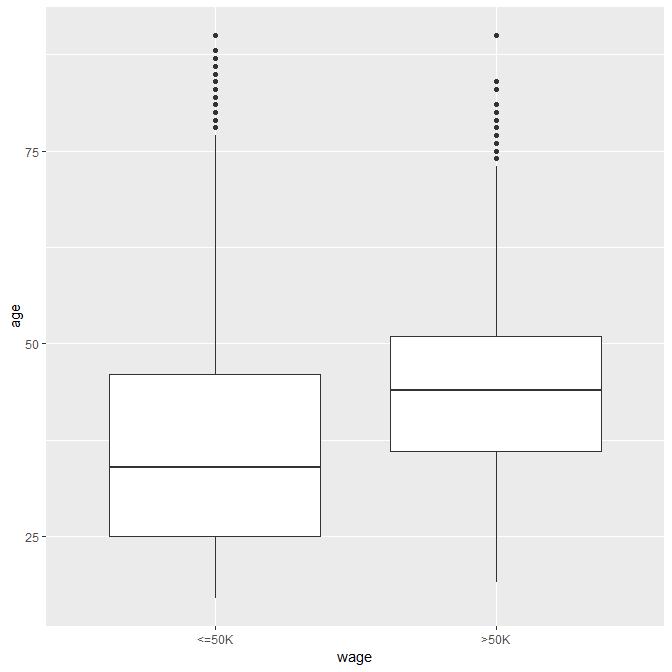


%>%(파이프연산자)를 통해 adult데이터를 사용하고, ggplot함수를 사용하겠다는 의미

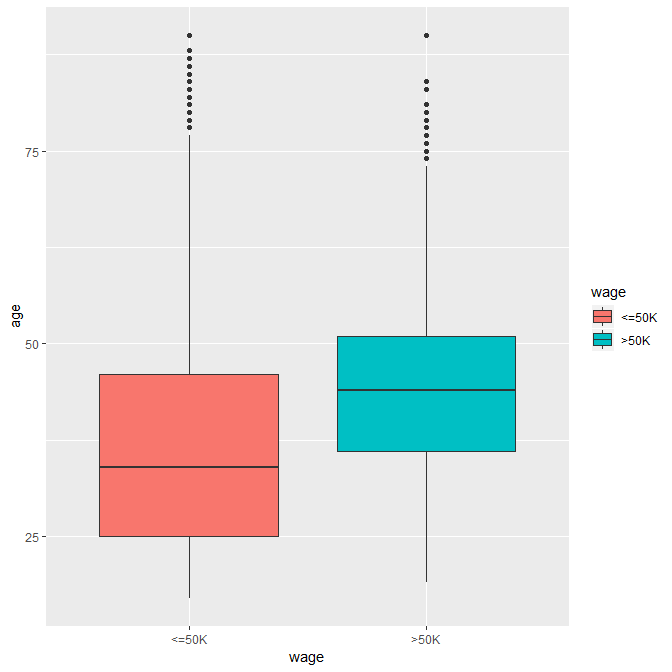
aes(aesthetic)옵션은 ‘심미적인’ 이라는 의미로 ggplot함수를 이용해 그림을 그릴 때,

x, y축에 어떤 변수를 사용할 건지, “fill=”옵션을 이용해서 어떤 변수에 대해 색을 칠할 것인지,

“group=”라는 변수를 통해 어떤 변수를 묶어서 뒤에 “geom\_func”에 해당하는 그림을 그룹별로 그려줄 것인지 정해줄 수 있다.

왼쪽의 그림은 x축에 wage(“>50K”, “<=50K)변수를 넣고,

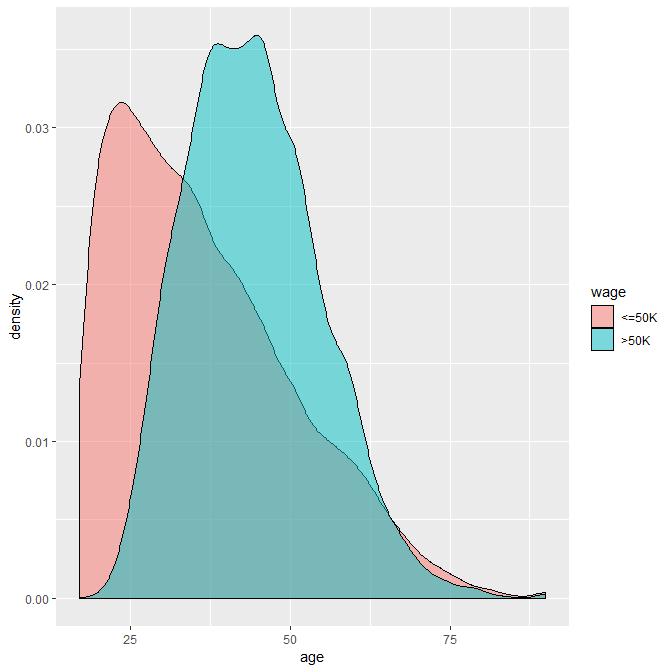
y축에 age를 boxplot으로 그린 그림.



“fill=wage”옵션을 통해 wage별로   
다른 색을 칠한 그림

ggplot함수로 도화지의 축을 설정하고, +뒤에 오는 geom\_function으로 어떤 그림을 도화지에 그릴 것인지 결정한다. geom=geometric으로써 어떤 기하학적인 그림을 보여줄 것인지를 의미함.



x축에는 age, fill=wage로 wage별로 그림을 그려라

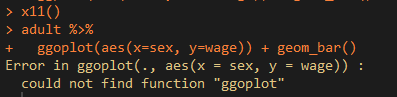
geom\_density로 확률분포를 그려 줌  
(여기서 말하는 확률분포는 정확한 확률분포가 아니라 histogram을 그린 후 각 막대기들을 적당히 선으로 연결한 것)

alpha=.5 옵션이 없으면 각 density그래프들이 너무 진하게 출력되기 때문에 그래프가 겹쳐서 보이지 않음.

alpha옵션으로 그림의 투명도를 조절

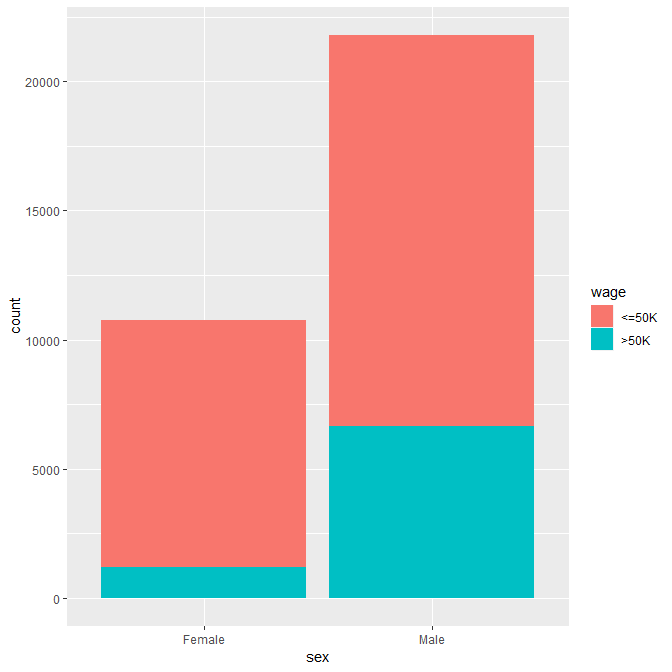
2. sex~wage

Sex, wage 모두 이산형 변수이므로 막대그래프 또는 히스토그램이 가장 적절한 시각화 방법.



위에서 했던 것처럼 aes(x=sex, y=wage)를 사용하면 에러가 뜬다.

왜냐하면 x축에는 sex(F, M) 2가지로 나타낼 수 있지만, y=wage를 하면 x=F일 때 wage의 위치를 나타낸다는 뜻인데 wage또한 이산형 변수이기 때문에 특정 위치를 표현할 수 없다.   
따라서 단순히 몇 개의 관측치가 있는지 count할 수밖에 없다.

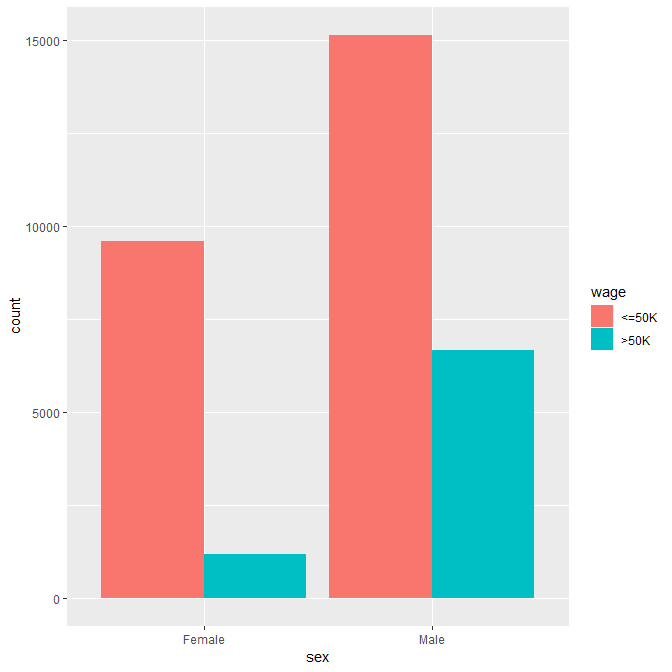


geom\_bar에 아무 옵션을 추가하지 않을 경우,

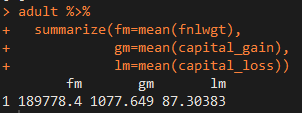
x값에 따라 wage들이 ‘>50K’와 ‘’<=50K”값들이

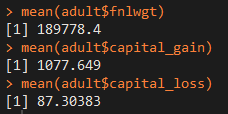
fill에 의해 서로 다른 색깔로 x=F, x=M위에 쌓이게 된다.



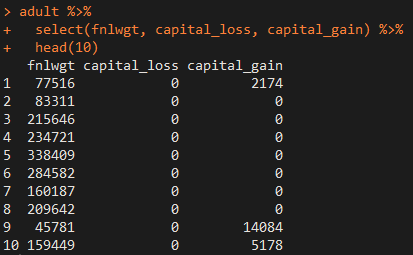
geom\_bar(position=’dodge’)옵션을 통해

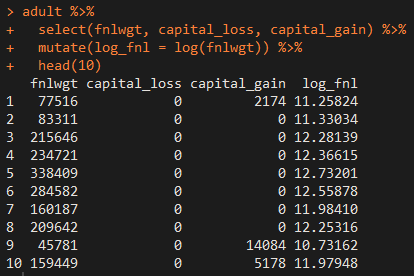
각 x값에 따라 wage들을 떼서 2개의 막대그래프를 생성해 준다.  
3. fnlwgt, capital\_loss, capital\_gain





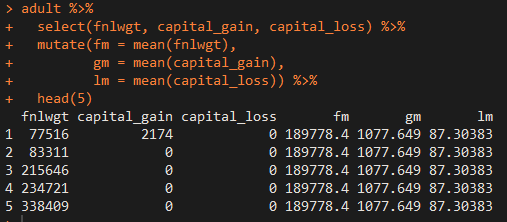
summarize함수로 fm, gm, lm이라는 새로운 변수명을 생성한 다음, 해당 변수명에 mean을 계산해서 할당해 줌.





mutate함수로 log\_fnl이라는 새로운 변수를 생성해줌.

Summarize는 요약된 기술 통계량을 출력해 주는 것이고, mutate는 새로운 변수를 생성해서   
각 obs들에 해당하는 새로운 변수 값을 열에 할당해 줌.



mutate함수를 써서 fm, gm, lm변수를 생성한 다음, mean을 해서 구할 수는 있지만,

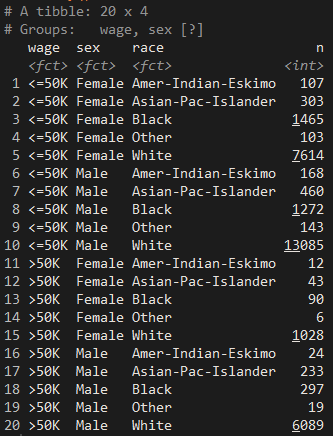
summarize와는 다르게 모든 평균값을 새로 만든 열에 모두 할당해주는 것이 차이점.

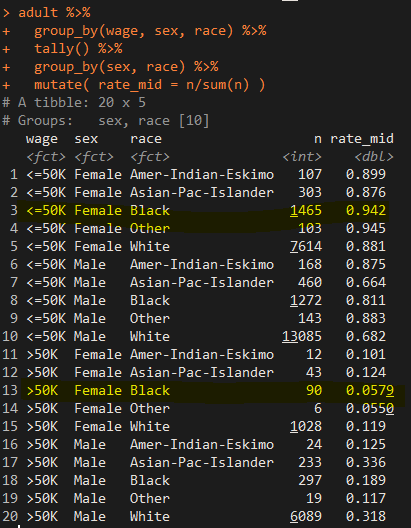
4. Sex, race, wage



adult데이터에서 wage, sex, race별로 그룹핑을 하고, 각 그룹에 대해서 count해라.

tally함수는 그룹별로 count해서 n이라는 변수에 값을 할당해 줌.

wage는 2개의 카테고리, sex도 2개,   
race는 5개의 범주를 가지므로   
총 2x2x5=20개의 항목에 대해 count해줌

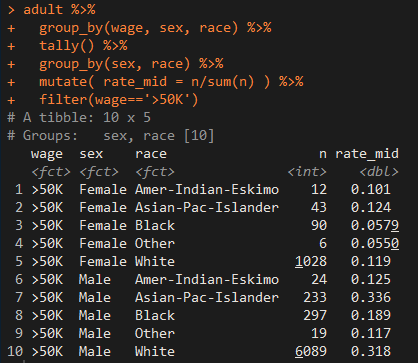


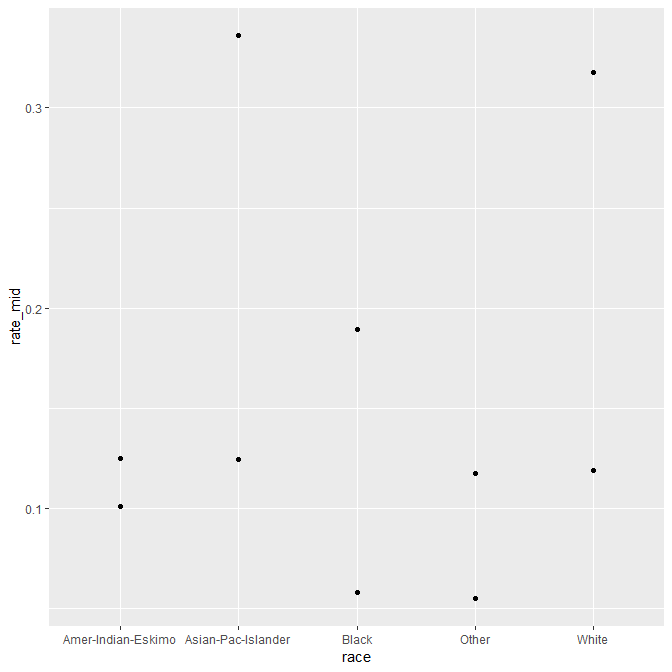
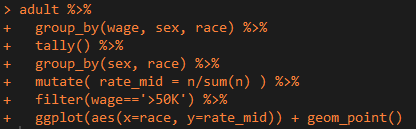
tally다음에 group\_by(sex, race)한 이유는 mutate에서 새로운 변수인 rate\_mid=n/sum(n)을   
실행하기 위해서임.

Wage는 새로 그룹핑을 하지 않았기 때문에 n/sum(n)을 실행하면   
sex=F, race=Black, wage=’>50K’와 sex=F, race=Black, wage=’<=50K’에 대해서 각각n을 계산하고 sum(n)은 sex=F, race=Black인 모든 관측치의 count값을 나타낸다.

따라서 n/sum(n)을 실행하게 되면, 동일 성별, 동일 인종에 대해 wage가 ‘>50K’인 비율, ‘<=50K’인 비율을 rate\_mid에 출력해 준다.

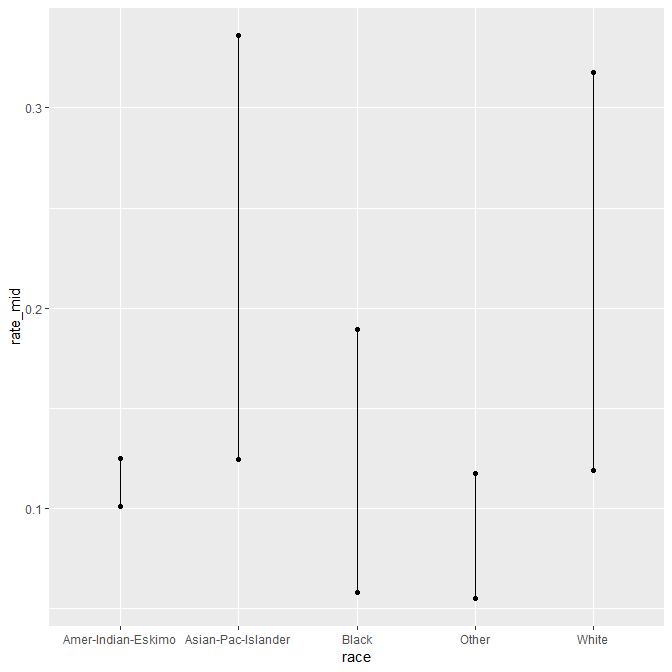
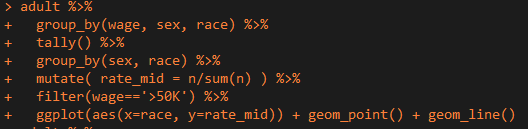
노란색으로 칠한 sex=F, race=Black들의 rate\_mid값을 합치면 1이됨.

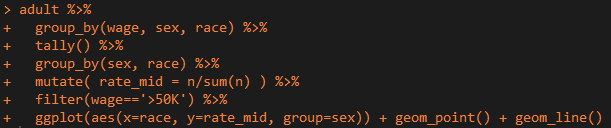
filter(조건) 함수로 조건에 부합하는 결과만 가져올 수 있음.

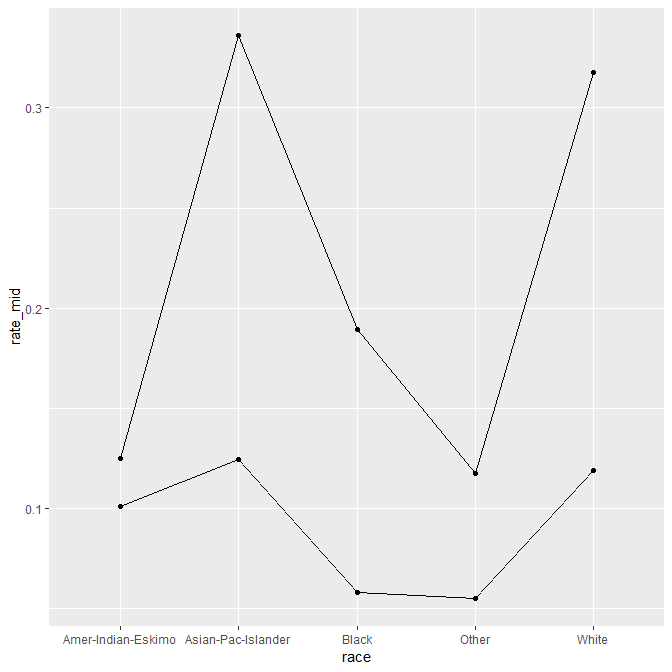
중산층의 비율이므로 wage==’>50K’인   
rate\_mid를 plotting하기 위해서   
filter(wage==’>50K’)를 사용  


ggplot함수에서 인종에 따른 rate\_mid의 위치를 점으로 나타낸 그림.

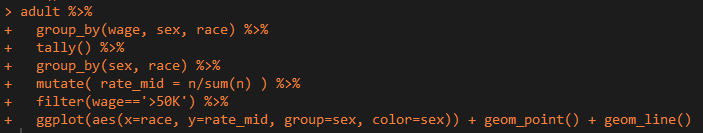
하지만 이 그림을 만든 이유는 인종과 성별에 따른 중산층의 비율의 추세를 보기 위함인데   
각 점들이 모두 검은색으로 나타나 있으므로 성별에 따른 정보를 확인하기 어려움.

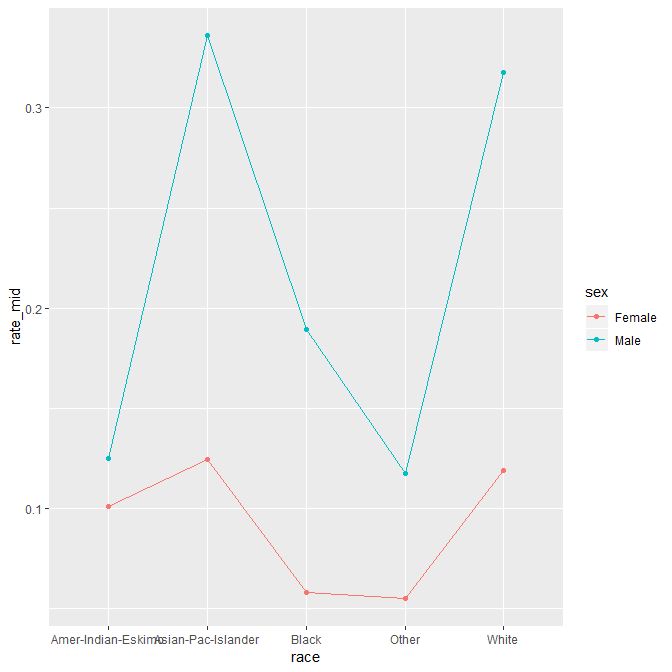


Geom\_line을 추가했을 때에도 동일한 x에 대해서 선을 긋기 때문에 잘못된 그래프임.



Group=sex 옵션을 통해 성별별로 점을 찍고, 선을 그어줌.

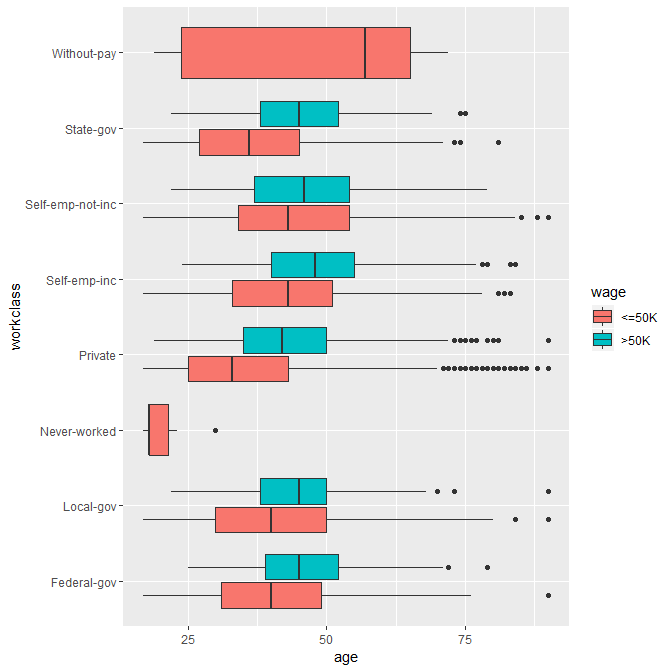




Color=sex옵션을 통해 성별에 따라 다른 색으로 그림을 그려줌.

(숙제)

1.



위의 그림처럼 만들고 코드 첨부해서 올리기.

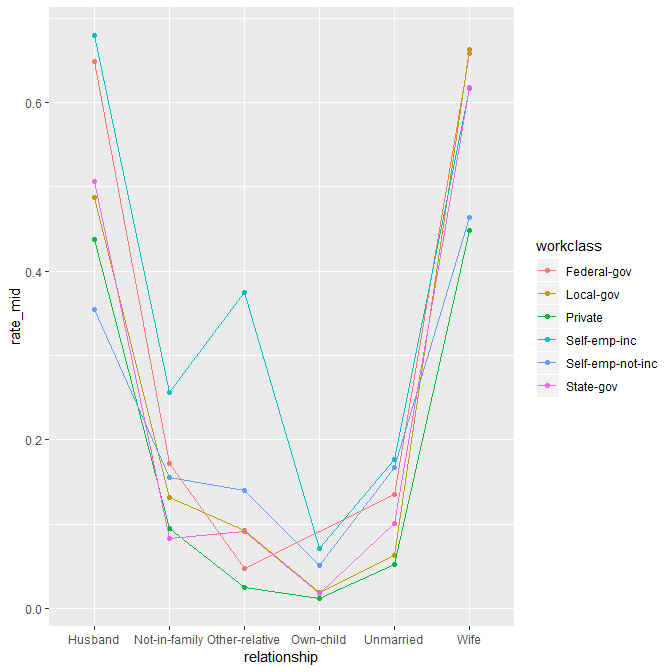
Hint

1. workclass가 ?인 관측치들을 먼저 제거.

2. wage별로 다른색깔

3. x축과 y축이 반전되어 있음.

2.



위 1번과 동일.

Hint

1.workclass==’?’인 관측치를 제거

2.x축은 relationship, y축은 rate\_mid, workclass별로 다른 색의 꺾은선 그래프