#problem 1.1

CPS1988 %>%

filter(parttime == 'yes') %>%

select('wage','experience','education')

#problem 1.2

mean\_Wage <- mean(CPS1988$wage)

CPS1988 %>%

filter(wage > mean\_Wage) %>%

select(experience) %>%

summarise(total = log(experience)) %>%

summarise(mean = mean(total))

#problem1.3

#problem2.1

multiTweeters <- tweets %>%

group\_by(tweet, user\_id) %>%

summarise(count = n()) %>%

arrange(desc(count)) %>%

filter(count > 1) %>%

ungroup() %>%

select(user\_id)

#problem 2.2

lateTweeters <- tweets %>%

group\_by(tweet, user\_id) %>%

arrange(desc(date,time)) %>%

select(user\_id, date, time, tweet) %>%

distinct()

#problem 2.3

tweets %>%

filter(date == '2016-04-16' & time == '12:44:00') %>%

#not sure afterwards

#problem2.4

#Not sure how to solve this

#problem3.1

#they produce the same result because finding the mean of the doubled wage and doubling the mean wage is the same.

#problem 3.2

# Finding the mean of the log wage and finding the log value of the wage mean would be different. Kind of hard to explain why but I get it.

#problem3.3

#Would be the same because it is inner join. Inner join gets rid of the NA rows.

#problem 3.4

#The results would be different because left\_join means the other dataframe is inserted into an another dataframe which can have totally different cells in the row.

#problem 4

#This was a very interesting problem. Even though I couldn't figure out and I am very weak in statistics, but tried my best.

#P value is a measurement of the probability that an observed difference can occur under the null hypothesis.

mean(results$trmt) - mean(results$ctrl)

t.test(results$trmt, results$ctrl)

ctrl <- rnorm(25, mean(results$ctrl), sd(results$ctrl))

ates <- rep(0, 10000)

ates

for (i in 1:10000) {

ates[i] <- sum(ctrl)/25

}

hist(ates, breaks = 10)