Homework 0

Tina Zhang

2019/4/14

Question 1

```
a <- 1:5
print(a)
```

```
## [1] 1 2 3 4 5
```

```
Mindy <- 12
print(Mindy)
```

```
## [1] 12
```

```
c <- matrix(1:6, nrow = 2, ncol = 3, byrow=TRUE)
print(c)
```

```
##      [,1] [,2] [,3]
## [1,]    1    2    3
## [2,]    4    5    6
```

```
d <- matrix(1:6, nrow = 2, ncol = 3)
print(d)
```

```
##      [,1] [,2] [,3]
## [1,]    1    3    5
## [2,]    2    4    6
```

```
e <- matrix(1, nrow = 10, ncol = 10)
f <- c ("THIS", "IS", "A", "VECTOR")
g <- function(a, b, c) {
  sum <- a+b+c
  return(sum)
}

h <- function(numb) {
  if(numb > 10){
    ret <- "No"
  }
  else{
    ret <- "Yes"
  }
  return(ret)
}
set.seed(235)
```

```r
g <- rnorm(1000, 10, 1)
y <- rnorm(1000, 5, 0.5)
x <- 1:1000
for (i in 1:1000)
{
  x[i] <- mean(sample(g,10, TRUE))
}
fit <- lm(y ~ x)
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.53376 -0.33352  0.02432  0.32661  1.25128
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.47333    0.49714   11.01   <2e-16 ***
## x           -0.04568    0.04963   -0.92    0.358
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.491 on 998 degrees of freedom
## Multiple R-squared:  0.0008481,  Adjusted R-squared:  -0.000153
## F-statistic: 0.8472 on 1 and 998 DF,  p-value: 0.3576
```

The results show the estimated regression equation to be y = 5.47333 -0.04568x. The R-squared is 0.0008481 and the adjusted R-squared is -0.000153, which is very low. The intercept is significantly different from 0, but the coefficient is not. Overall this suggests x's explanatory power of y is very low.

Question 2a,b

```r
library("descr")
library("ggplot2")

setwd("C:/Users/tzwhi/Desktop/Northwestern/311-2/R")
pums <- read.csv("pums_chicago.csv")
print(dim(pums))
```

```
## [1] 50000   204
```

There are 204 variables in the dataset

Question 2(c)

```r
mai <- mean(pums$PINCP, na.rm=TRUE)
print(mai)
```

```
## [1] 38247.62
```

The mean annual income is approximately $38,247.62

Question 2(d)

```r
pums$PINCP_LOG <- log(pums$PINCP)
```

NaNs were produced because some entries of PINCP have values of 0 or NA

Question 2 e-j

```r
# Assumes that GED/alternative credential doesn't count as post-high-school education
grad_dummy <- c()
for(i in 1:5000){
  grad_dummy[i] <- "no grad"
  if(!is.na(pums$SCHL[i]) & pums$SCHL[i] > 17){
    grad_dummy[i] <- "grad"
  }
}
pums$SERIALNO <- NULL
write.csv(pums,"new dataset for part (g).csv")
under16 <- subset(pums, is.na(ESR))
employed <- subset(pums,ESR == 1 | ESR == 2)
unemployed <- subset(pums, ESR == 3)
inarmedforce <- subset(pums, ESR == 4 | ESR == 5)
notlaborforce <- subset(pums, ESR == 6)
#Note: I made the assumption that "Armed forces, at work" is not included in the employed category
employed_af <- rbind(employed, inarmedforce)
employed_af <- subset(employed_af, select = c("AGEP", "RAC1P", "PINCP_LOG"))
```

Question 2k(i)

mean: 34.84, median: 30, 80th percentile: 45

```r
summary(pums$JWMNP)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00   20.00   30.00   34.84   45.00  149.00   27668
```

```r
print(quantile(pums$JWMNP,.8, na.rm = TRUE))
```

```
## 80%
##  45
```

Question 2k(ii)

The correlation is -0.04205232

```
cor(pums$JWMNP, pums$WAGP, use = "complete.obs")
```

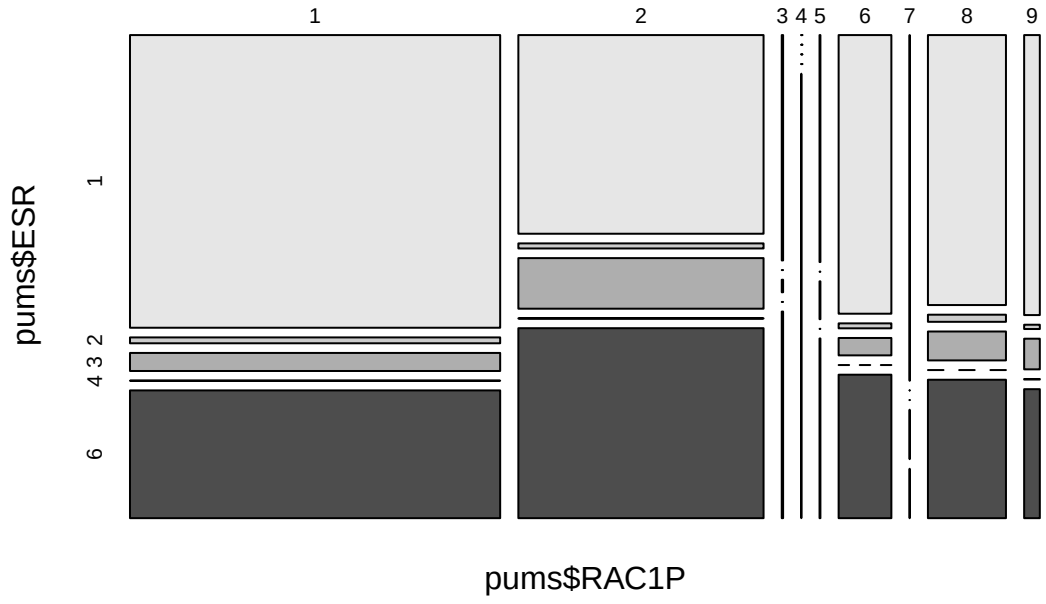
```
## [1] -0.04205232
```

Question 2k(iii)-(vii)

```
pdf("ivplot.pdf")
plot(pums$AGEP, pums$PINCP_LOG, main = "Graph for (iii): Log Income vs. Age",
                xlab = "Age (Years)", ylab = "Log Income", col = "#2E9FDF")
dev.off()
```

```
## pdf
##   2
```

```
crosstab(pums$ESR, pums$RAC1P)
```



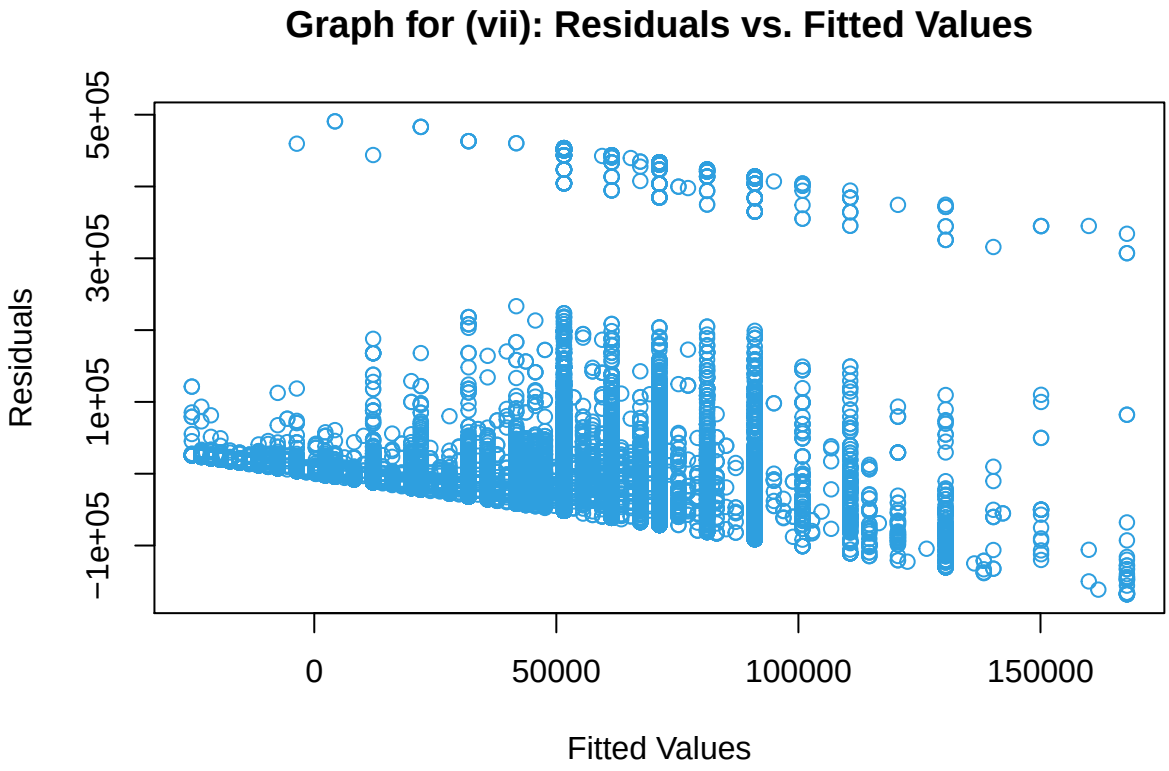

```
##    Cell Contents
## |-------------------------|
```

```
## |                   Count |
## |-------------------------|
##
## ========================================================================
##             pums$RAC1P
## pums$ESR        1       2    3   4    5      6   7      8     9   Total
## ------------------------------------------------------------------------
## 1           1.287e+04    5786   36   0   24   1746   7   2502   521   2.349e+04
## ------------------------------------------------------------------------
## 2             258     147    0   0    0     31   0     66     8     510
## ------------------------------------------------------------------------
## 3             794    1473    2   0    4    109   0    268    57    2707
## ------------------------------------------------------------------------
## 4               4       5    0   0    0      0   1      0     1       11
## ------------------------------------------------------------------------
## 6            5618    5533   33   2   19    899   1   1283   240   1.363e+04
## ------------------------------------------------------------------------
## Total       1.954e+04   1.294e+04   71   2   47   2785   9   4119   827   4.035e+04
## ========================================================================
```

```
Qkvi <- lm(WAGP~WKHP, pums)
summary(Qkvi)
```

```
##
## Call:
## lm(formula = WAGP ~ WKHP, data = pums)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -167856  -27577  -11577    9491  490723
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -27256.47    1253.63  -21.74   <2e-16 ***
## WKHP          1970.83      30.97   63.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61490 on 26206 degrees of freedom
##   (23792 observations deleted due to missingness)
## Multiple R-squared:  0.1339, Adjusted R-squared:  0.1338
## F-statistic:  4050 on 1 and 26206 DF,  p-value: < 2.2e-16
```

```
plot(Qkvi$fitted.values, Qkvi$residuals,
     main = "Graph for (vii): Residuals vs. Fitted Values",
     xlab = "Fitted Values", ylab = "Residuals", col = "#2E9FDF")
```



If a linear model were specified correctly, then the residuals should appear randomly distributed around 0, and their values (or variance) shouldn't be correlated with the fitted value. After all, the residuals give us an estimate of the error term in the regression equation.

However, in this graph, while the variance of the residuals seems relatively constant, the residuals seem to exhibit a linear relation with the fitted values, and they definitely don't look random. This suggests that the regression model may not have been specified correctly, perhaps due to omitted variable bias, or perhaps because the relationship between WAGP and WKHP is nonlinear.

Question 2(l)

```
data("mtcars")
Qli <- lm(mpg~wt, mtcars)
summary(Qli)
```

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
## wt           -5.3445     0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

```
manualfit <- lm(mpg~wt, subset(mtcars, am == 1))
autofit <- lm(mpg~wt, subset(mtcars, am == 0))
summary(manualfit)
```

```
##
## Call:
## lm(formula = mpg ~ wt, data = subset(mtcars, am == 1))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4190 -1.4937 -1.2234  0.8228  6.0909
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   46.294      3.120  14.839 1.28e-08 ***
## wt            -9.084      1.257  -7.229 1.69e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.686 on 11 degrees of freedom
## Multiple R-squared:  0.8261, Adjusted R-squared:  0.8103
## F-statistic: 52.26 on 1 and 11 DF,  p-value: 1.688e-05
```

```
summary(autofit)
```

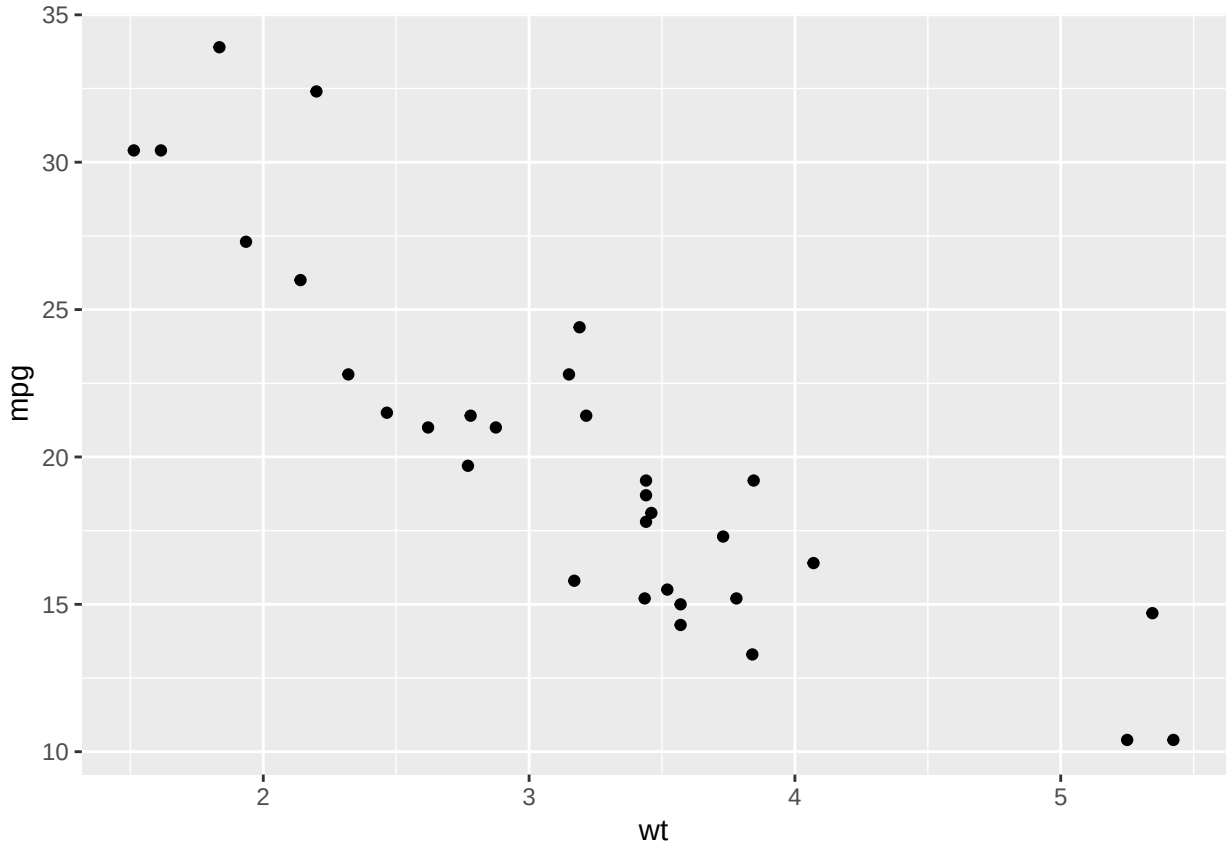
```
##
## Call:
## lm(formula = mpg ~ wt, data = subset(mtcars, am == 0))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6004 -1.5227 -0.2168  1.4816  5.0610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  31.4161     2.9467  10.661 6.01e-09 ***
## wt           -3.7859     0.7666  -4.939 0.000125 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.528 on 17 degrees of freedom
## Multiple R-squared:  0.5893, Adjusted R-squared:  0.5651
## F-statistic: 24.39 on 1 and 17 DF,  p-value: 0.0001246
```

```r
Qliii <- lm(mpg~log(hp), mtcars)
summary(Qliii)
```

```
##
## Call:
## lm(formula = mpg ~ log(hp), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9427 -1.7053 -0.4931  1.7194  8.6460
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   72.640      6.004  12.098 4.55e-13 ***
## log(hp)      -10.764      1.224  -8.792 8.39e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.239 on 30 degrees of freedom
## Multiple R-squared:  0.7204, Adjusted R-squared:  0.7111
## F-statistic:  77.3 on 1 and 30 DF,  p-value: 8.387e-10
```

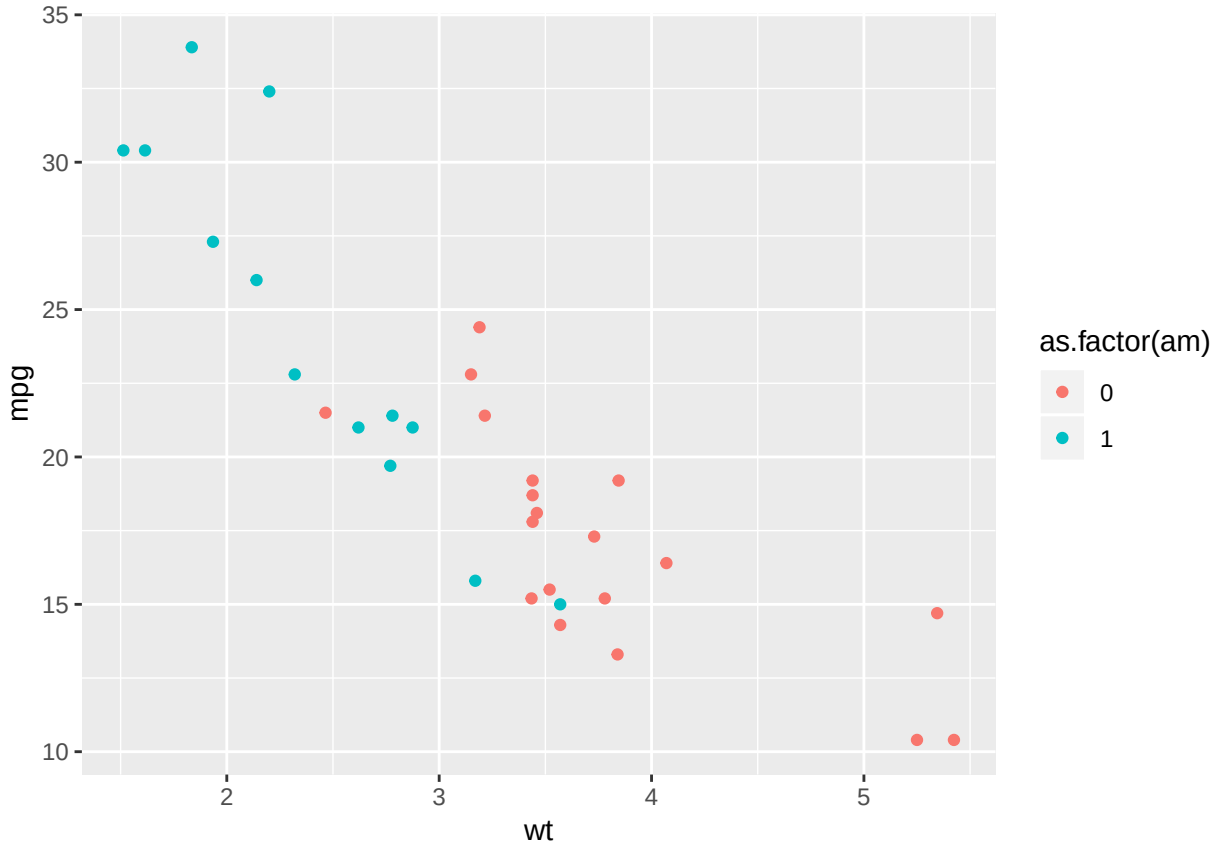
Question 2(m)i

```r
ggplot(mtcars)+geom_point(aes(wt, mpg))
```



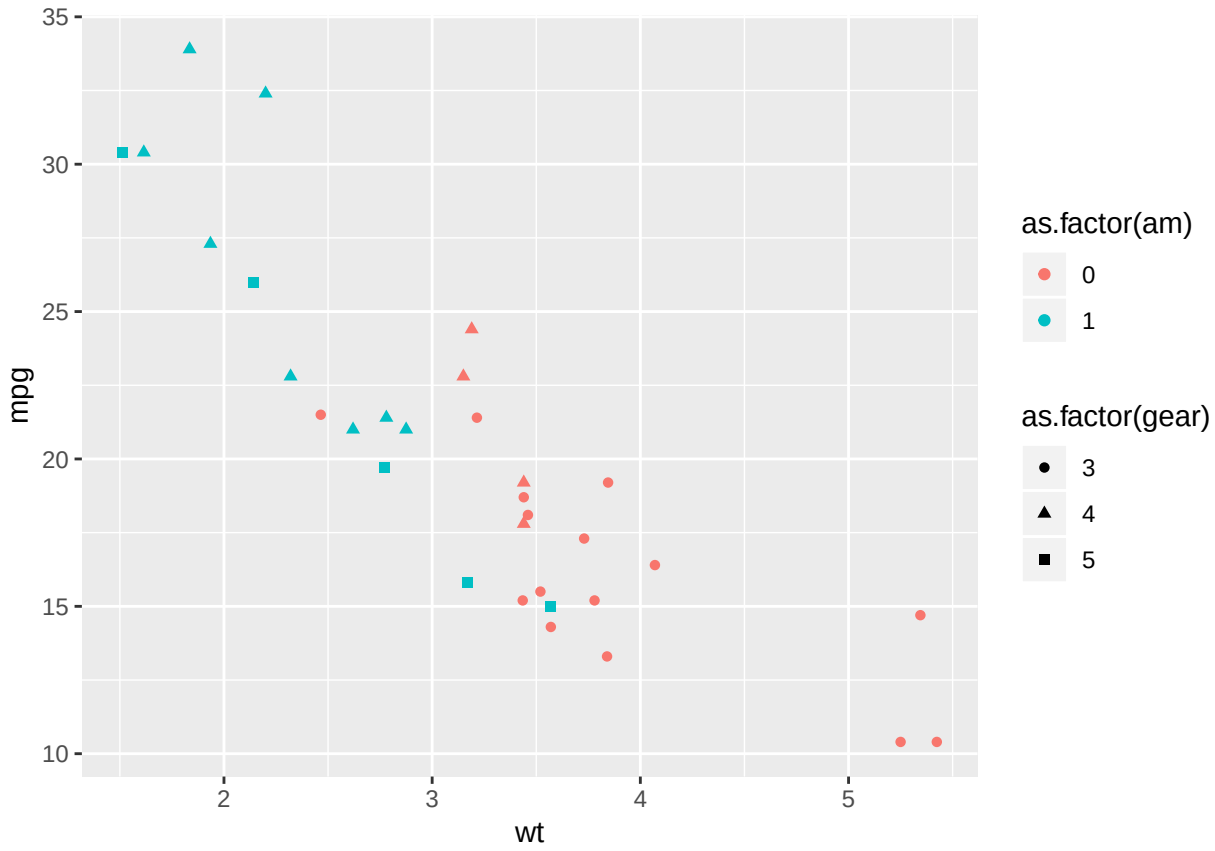
Question 2(m)ii

```
ggplot(mtcars)+
  geom_point(aes(wt, mpg, color = as.factor(am)))
```



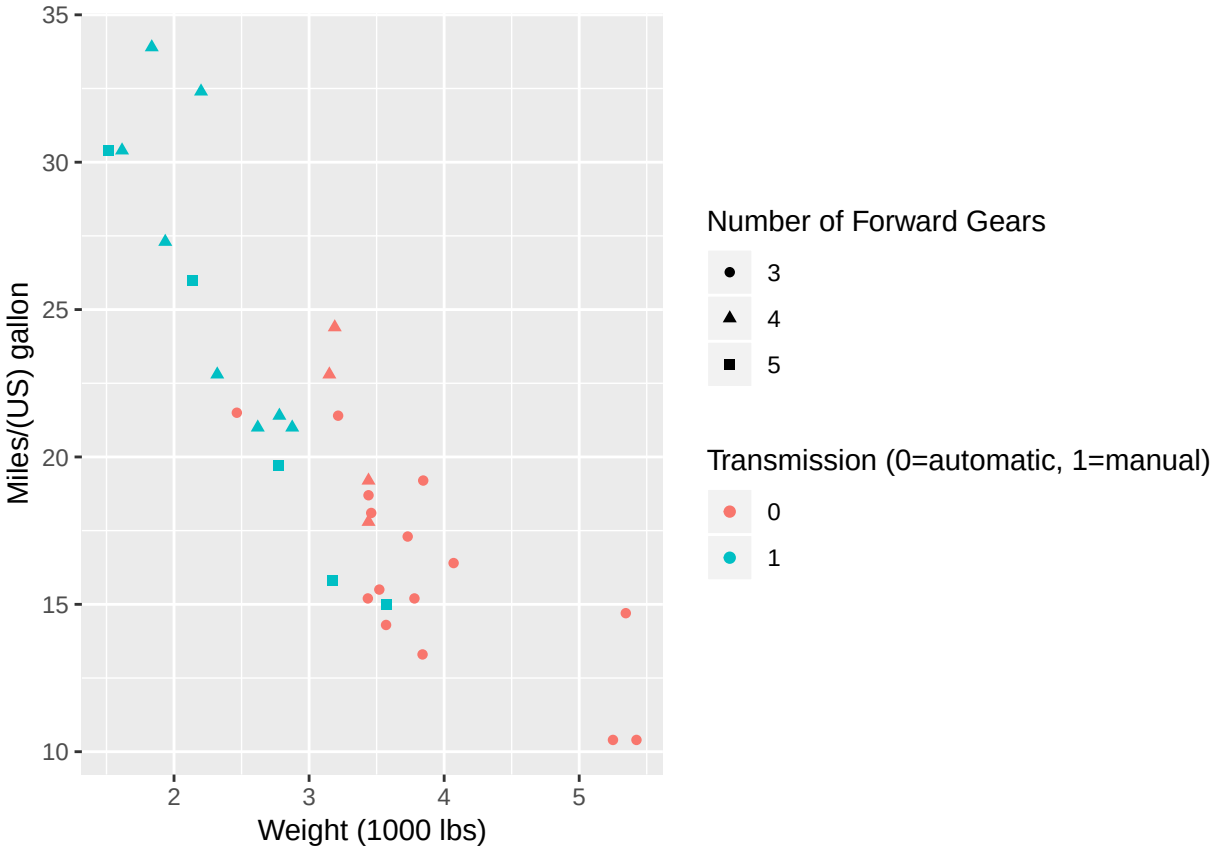
Question 2(m)iii

```
ggplot(mtcars)+
  geom_point(aes(wt, mpg, color = as.factor(am), shape = as.factor(gear)))
```



Question 2(m)iv

```
ggplot(mtcars)+
  geom_point(aes(wt, mpg, color = as.factor(am), shape = as.factor(gear))) +
  labs(x = "Weight (1000 lbs)", y = "Miles/(US) gallon", shape = "Number of Forward Gears", color = "Tra
```



Question 2(m)v

```
ggplot(mtcars)+
  geom_point(aes(wt, mpg, color = as.factor(am), shape = as.factor(gear))) +
  labs(x = "Weight (1000 lbs)", y = "Miles/(US) gallon", shape = "Number of Forward Gears", color = "Tra
  theme(panel.background = element_rect("lightblue"))
```


35
30
25
20
15
10
Miles/(US) gallon
2
3
4
5
Weight (1000 lbs)
Number of Forward Gears
3
4
5
Transmission (0=automatic, 1=manual)
0
1