HW1_problem6

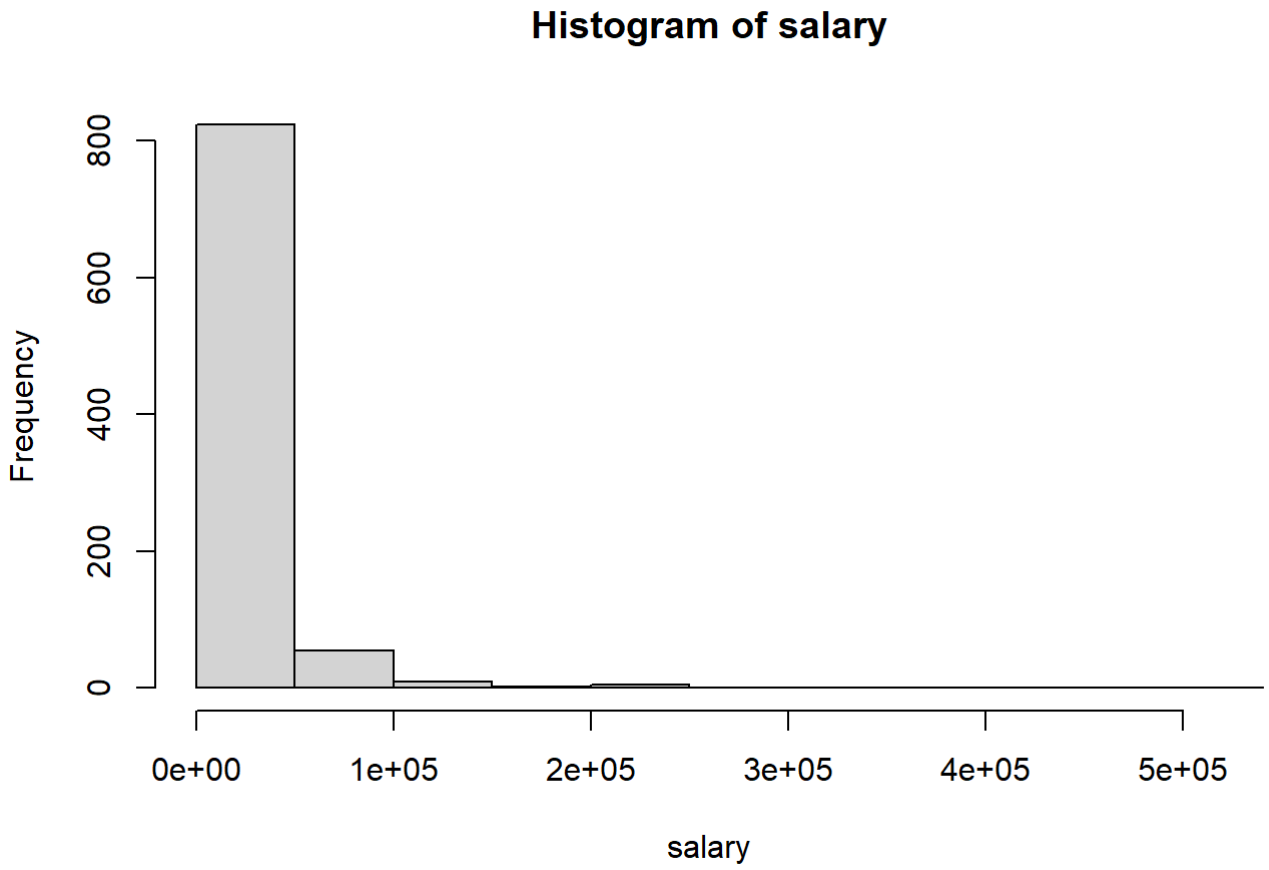
TA Sunmin Lee

2022-09-15

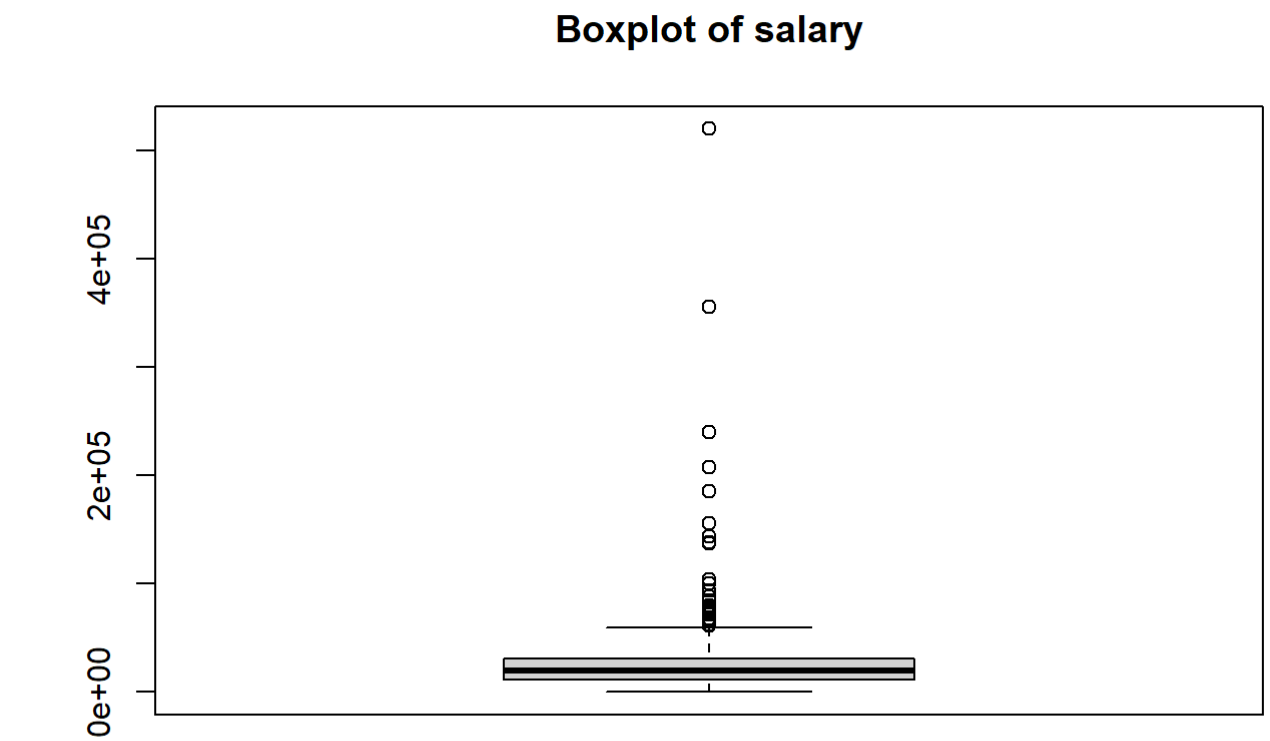
(a)

By comparing the following boxplots and histograms, we can check the distribution of *salary* is heavily skewed to the right while *lsalary* is unimodal and relatively symmetric.

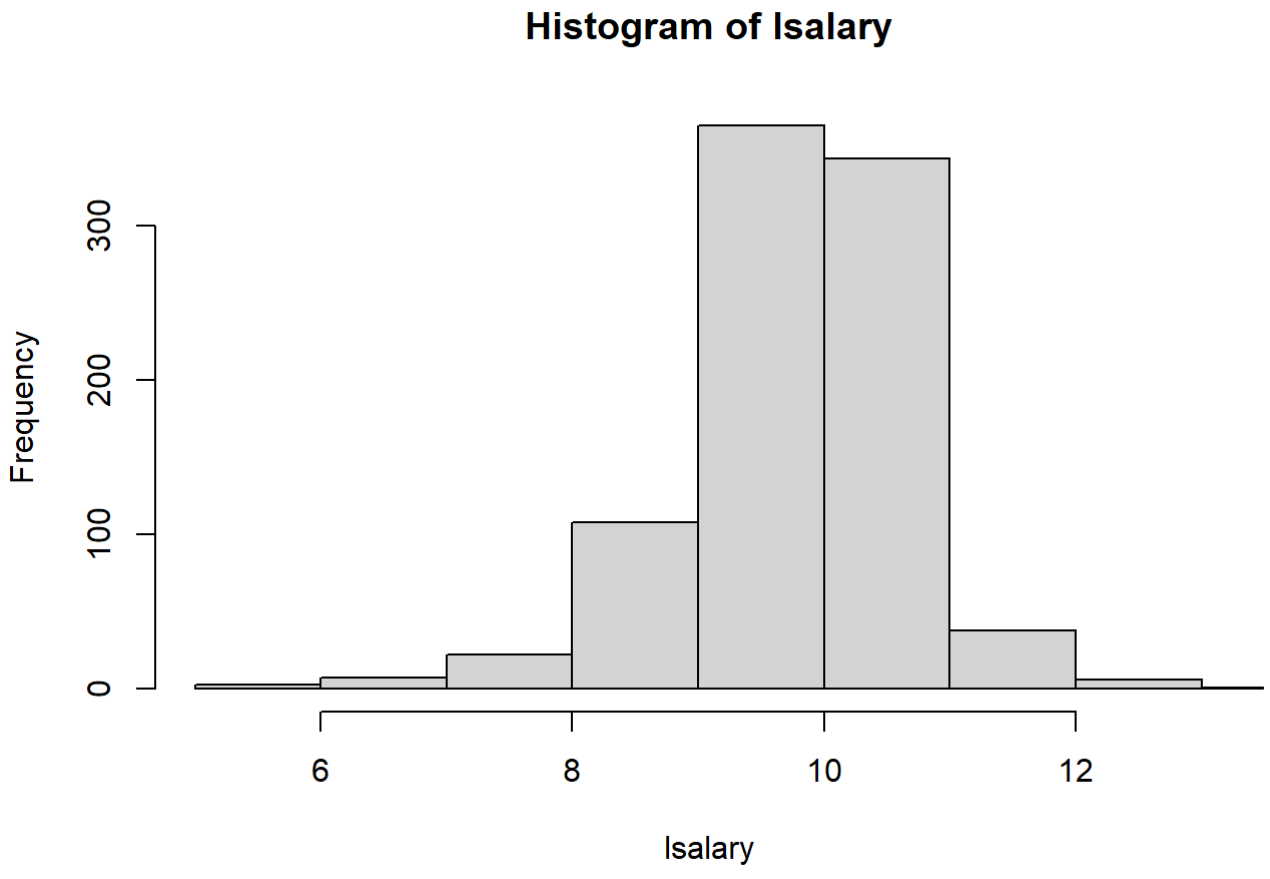
```
GA <- data.frame(read.csv('gallup.csv', header=TRUE))
hist(GA$salary, main='Histogram of salary', xlab='salary', xlim=c(min(GA$salary), max(GA$salary)))
```



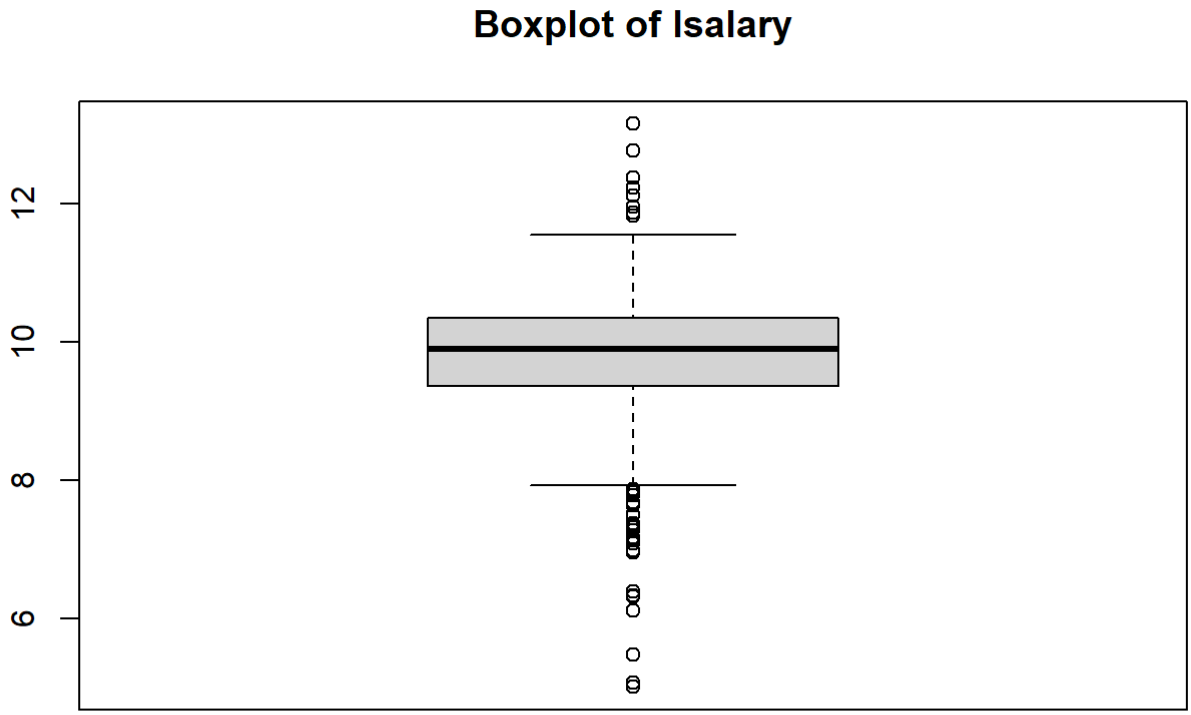

```
boxplot(GA$salary, main='Boxplot of salary')
```




```
hist(GA$lsalary,  main='Histogram of lsalary', xlab='lsalary', xlim=c(min(GA$lsalary), max(GA$lsalary)))
```



```
boxplot(GA$lsalary, main='Boxplot of lsalary')
```



(b)

The following results show each value in the problem:

```
#report Q1, Q2, mean and median
summary(GA$lsalary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.011   9.377   9.903   9.794  10.348  13.162
```

```
#report 35th percentile
cat("35th percentile=", quantile(GA$lsalary, 0.35), '₩n')
```

```
## 35th percentile= 9.654834
```

```
#report IQR
cat("IQR=", IQR(GA$lsalary))
```

```
## IQR= 0.9714982
```

(c)

```
L5 <- GA[GA$location==5,]
L6 <- GA[GA$location==6,]
L7 <- GA[GA$location==7,]
cat('Standard variation of location 5 is ', sd(L5$wage), 'and the other values can be obtained from the following data:₩n')
```

```
## Standard variation of location 5 is  10.25878 and the other values can be obtained from the following data:
```

```
summary(L5$wage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.27    5.00    8.33   10.56   12.50  100.00
```

```
cat('Standard variation of location 6 is ', sd(L6$wage), 'and the other values can be obtained from the following data:₩n')
```

```
## Standard variation of location 6 is  9.126596 and the other values can be obtained from the following data:
```

```
summary(L6$wage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.93    6.25    9.21   11.52   13.51   66.67
```

```
cat('Standard variation of location 7 is ', sd(L7$wage), 'and the other values can be obtained from the following data:₩n')
```

```
## Standard variation of location 7 is  14.75312 and the other values can be obtained from the following data:
```

```
summary(L7$wage)
```

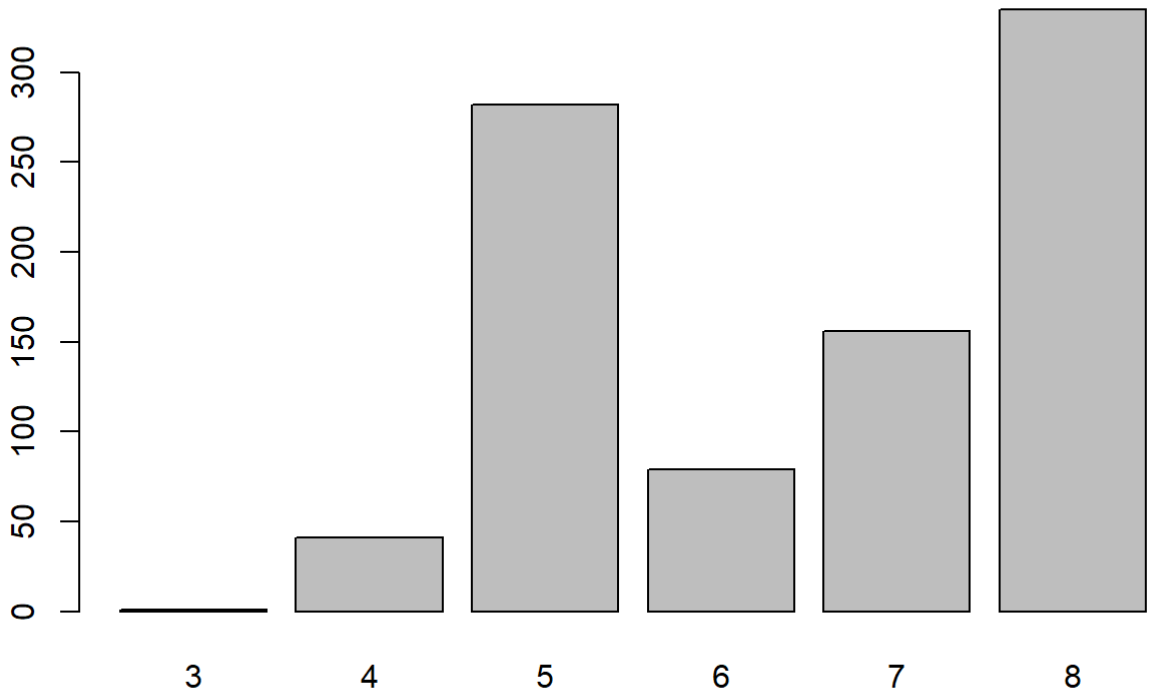
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.74    7.50   11.11   13.96   16.25  227.27
```

```
cat('Then, we can find the mean of location 7 (rest of count) is the highest with the value:', mean(L7$wage), '; Also, since the standard deviation of location 6 (Pittsburgh) is the lowest, location 6 has the lowest variance with the value:', var(L6$wage))
```

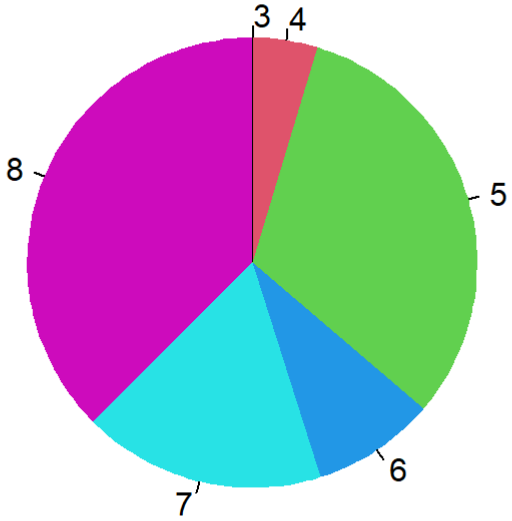
```
## Then, we can find the mean of location 7 (rest of count) is the highest with the value: 13.95737 ; Also, since the standard deviation of location 6 (Pittsburgh) is the lowest, location 6 has the lowest variance with the value: 83.29476
```

(d)

```
Tedu <- table(GA$educ)
barplot(Tedu)
```




```
pie(Tedu, clockwise=T, col = 1:6, lty=0)
```

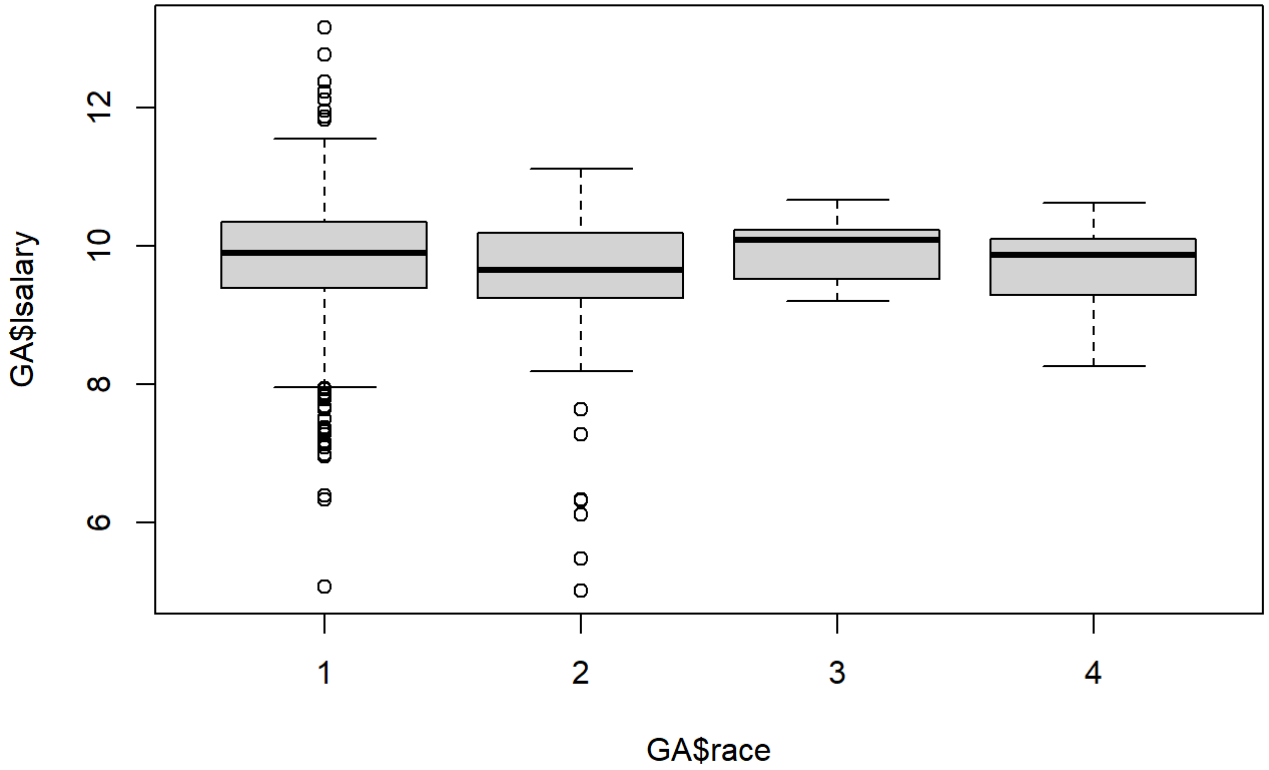



```
EDU <- data.frame(GA$educ)
x <- nrow(GA[GA$educ >= 5,])/nrow(GA)
cat(x*100, 'percent of the observations is at least high school graduate')
```

```
## 95.30201 percent of the observations is at least high school graduate
```

(e)

```
boxplot(GA$lsalary ~ GA$race)
```




```
Gwhite <- GA[GA$race == 1,]
Gblack <- GA[GA$race == 2,]
Gother <- GA[GA$race == 3,]
Ghispanic <- GA[GA$race == 4,]
summary(Gwhite$lsalary)
```

```
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  5.075  9.393  9.904  9.825 10.348 13.162
```

```
summary(Gblack$lsalary)
```

```
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  5.011  9.250  9.655  9.457 10.198 11.120
```

```
summary(Gother$lsalary)
```

```
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  9.200  9.524 10.086  9.926 10.237 10.674
```

```
summary(Ghispanic$lsalary)
```

```
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  8.269  9.309  9.876  9.685 10.089 10.636
```

```
cat('By comparing the mean of each race, we can obtain [other 〉 white 〉 hispanic 〉 black]₩n')
```

```
## By comparing the mean of each race, we can obtain [other 〉 white 〉 hispanic 〉 black]
```

```
cat('each variation of white, black, other and hispanic is', var(Gwhite$lsalary), var(Gblack$lsalary), var(Gother$lsalary), var(Ghispanic$lsalary), ',respectively₩n')
```

```
## each variation of white, black, other and hispanic is 0.7753157 1.511399 0.3005829 0.4172616 ,respectively
```

According to the box plot, we can find the following:
Distribution shape: they look roughly symmetric except for race 3 (other).
Centers: The centers are similar. Other has the highest and black has the lowest median values.
Variations:Black 〉 White 〉 Hispanic 〉 Other
Outliers: White has many outliers in both tails and black has several in the lower tail. Hispanic and other do not have any outliers.

(f)

```
RE <- data.frame(GA$educ, GA$rate)
table(RE)
```

```
##       GA.rate
## GA.educ  0  1  2
##      3  0  0  1
##      4 11  8 22
##      5 44 111 127
##      6 15 22 42
##      7 22 59 75
##      8 25 171 139
```

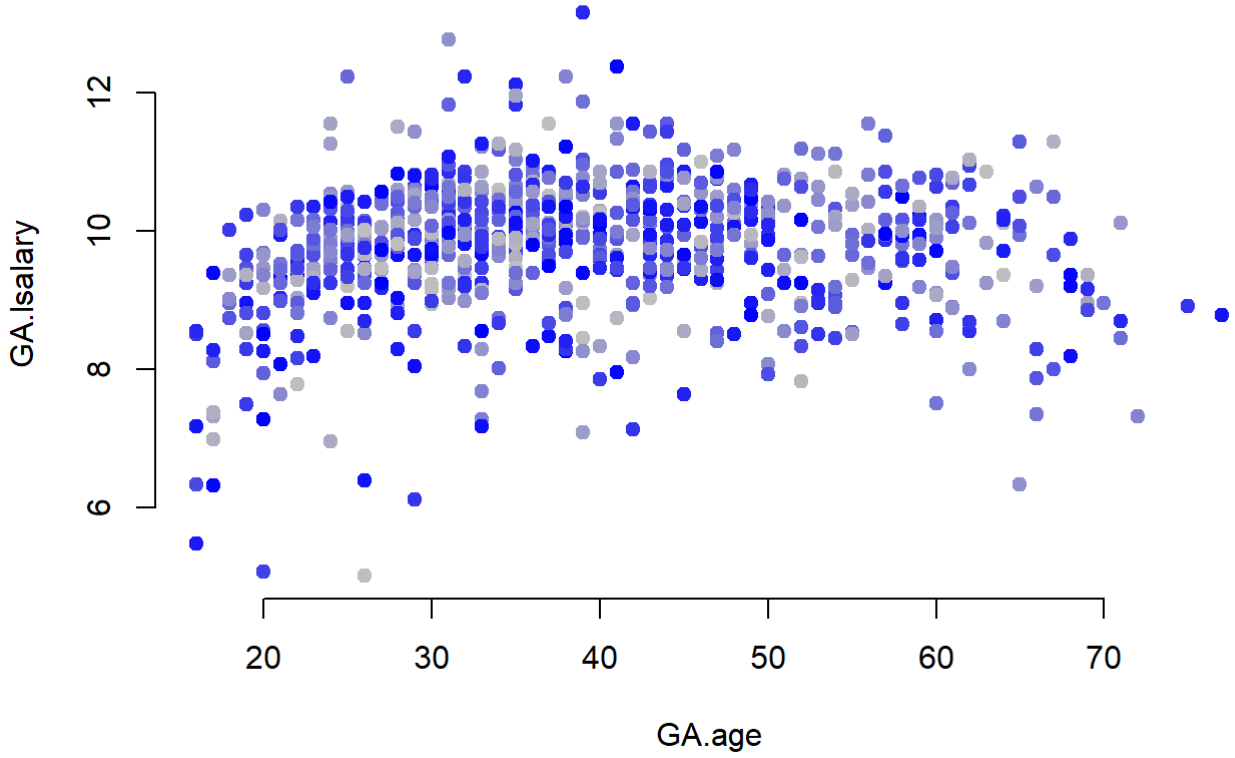
```
cat('we can find that among the excellent rate group, the percentage of high school incomplete or high school graduate is', (22+127)/nrow(RE[GA$rate==2,])*100, 'percent.')
```

```
## we can find that among the excellent rate group, the percentage of high school incomplete or high school graduate is 36.69951 percent.
```

(g)

There does not seem to be any strong relationship between the two variables except for a slight increase in early ages and a slight decrease in late ages.

```
LA <- data.frame(GA$age, GA$lsalary)
pal <- colorRampPalette(c("blue", "gray"))
plot(LA, col = pal(30), pch = 19, bty = "n")
```


GA.lsalary
6
8
10
12
20
30
40
50
60
70
GA.age