Homework2

QIang Li QIL45@pitt.edu

Team Member: Qiang Li, Qiao Zhao

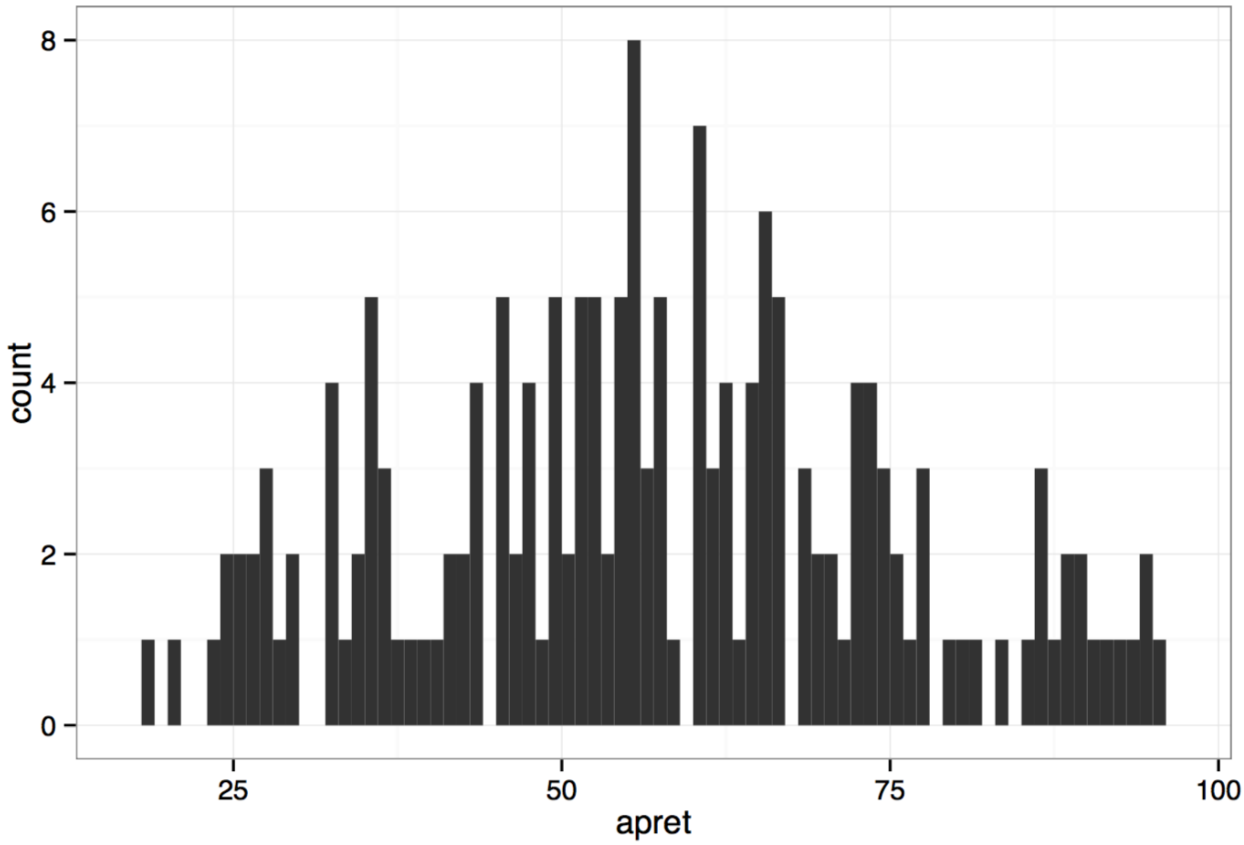
load the data and packages

```
library(ggplot2)
data
=read.table("http://chirayukong.github.io/infsci2725/resources/lecture4/Retention.txt",
header = T)
summary(data)
```

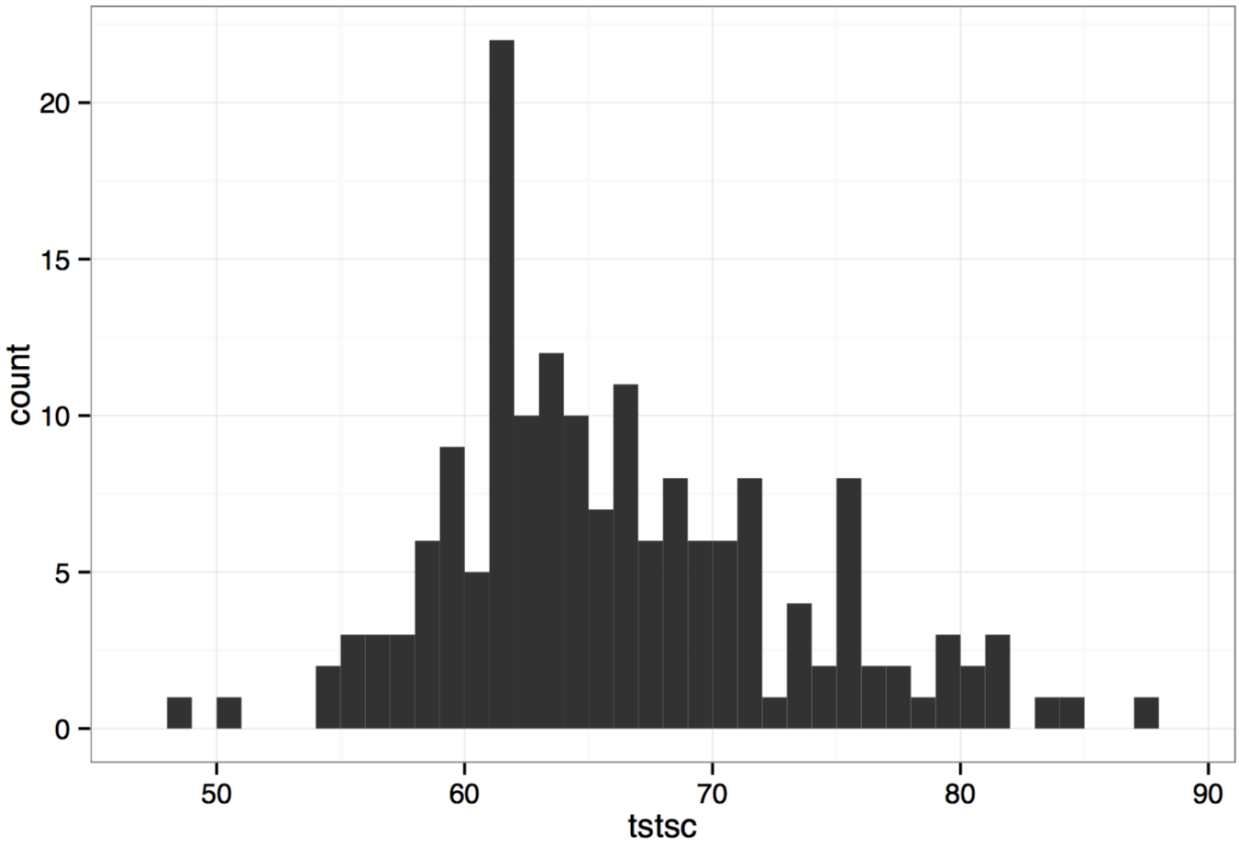
```
##      spend           apret           top10           rejr
##  Min.   : 4125   Min.   :18.75   Min.   : 8.00   Min.   : 0.00
##  1st Qu.: 7372   1st Qu.:45.37   1st Qu.:22.00   1st Qu.:19.17
##  Median : 9265   Median :55.71   Median :30.00   Median :27.39
##  Mean   :10975   Mean   :56.72   Mean   :38.46   Mean   :30.65
##  3rd Qu.:12838   3rd Qu.:68.69   3rd Qu.:49.50   3rd Qu.:36.81
##  Max.   :35863   Max.   :95.25   Max.   :98.00   Max.   :84.07
##      tstsc            pacc            strat           salar
##  Min.   :48.12   Min.   : 8.964   Min.   : 7.20   Min.   :38640
##  1st Qu.:61.11   1st Qu.:33.904   1st Qu.:13.40   1st Qu.:54650
##  Median :64.78   Median :40.850   Median :16.00   Median :61150
##  Mean   :66.16   Mean   :43.173   Mean   :16.09   Mean   :61358
##  3rd Qu.:70.45   3rd Qu.:51.773   3rd Qu.:18.57   3rd Qu.:67100
##  Max.   :87.50   Max.   :76.253   Max.   :29.20   Max.   :87900
```

plot histograms for the following three columns: apret, tstsc, and salar.

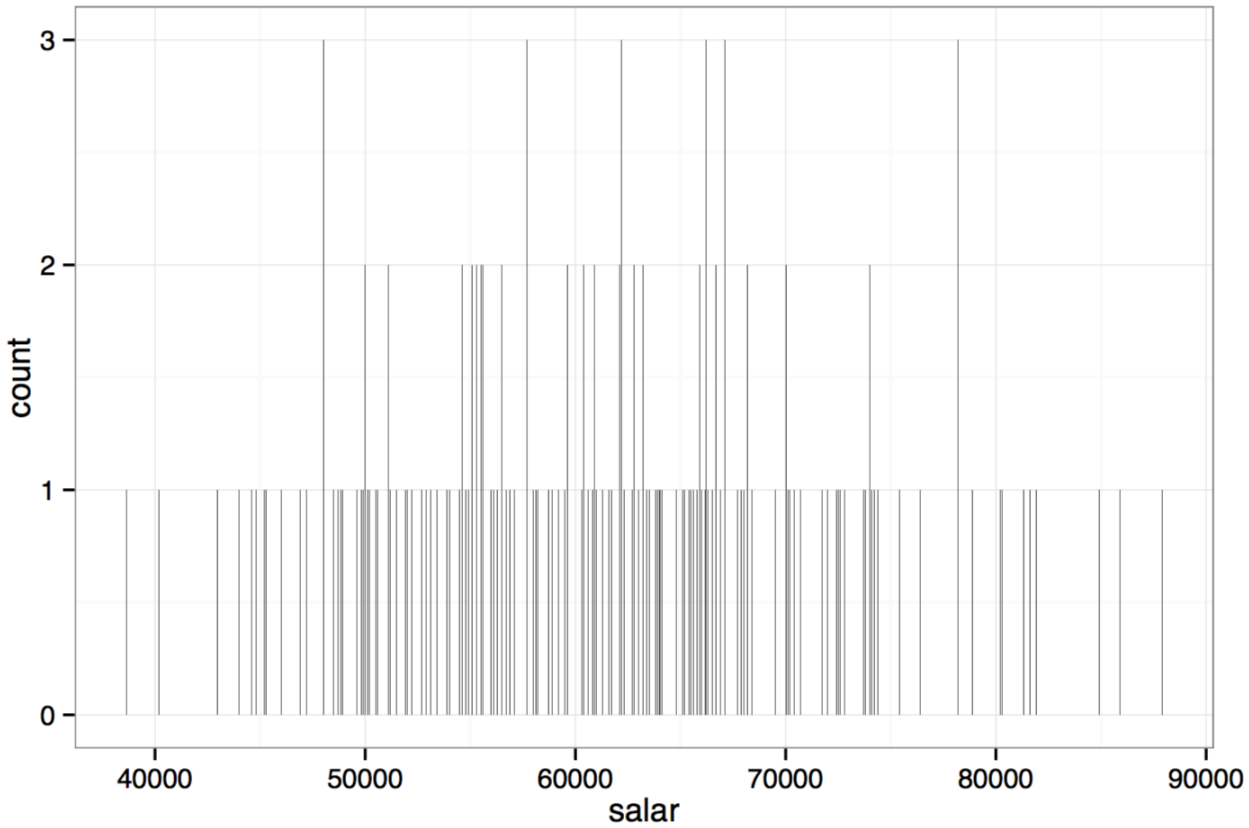
```
theme_set(theme_bw())
ggplot(data , aes(x = apret)) + geom_histogram(binwidth=1)
```



ggplot(data , **aes**(x = tstsc)) + **geom_histogram**(binwidth=1)



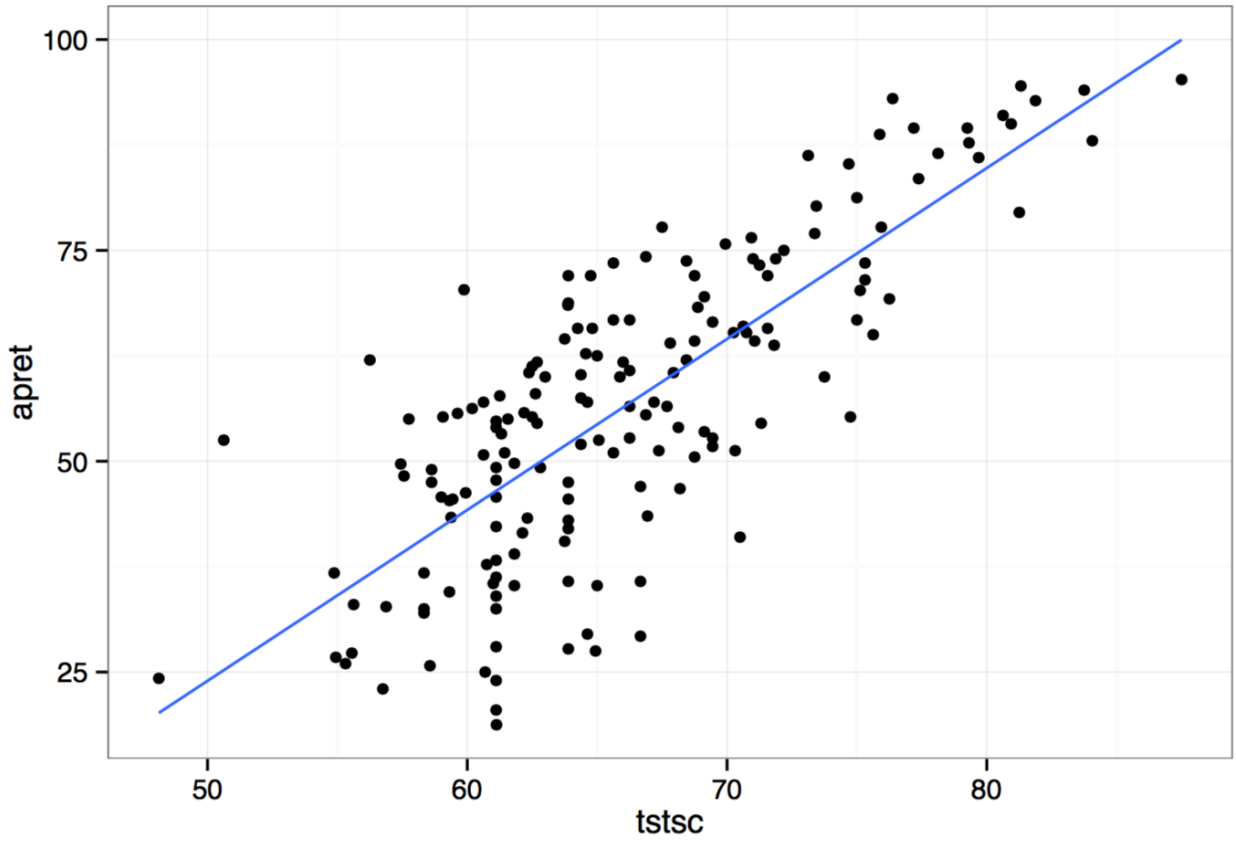

```
ggplot(data , aes(x = salar)) + geom_histogram(binwidth=1)
```



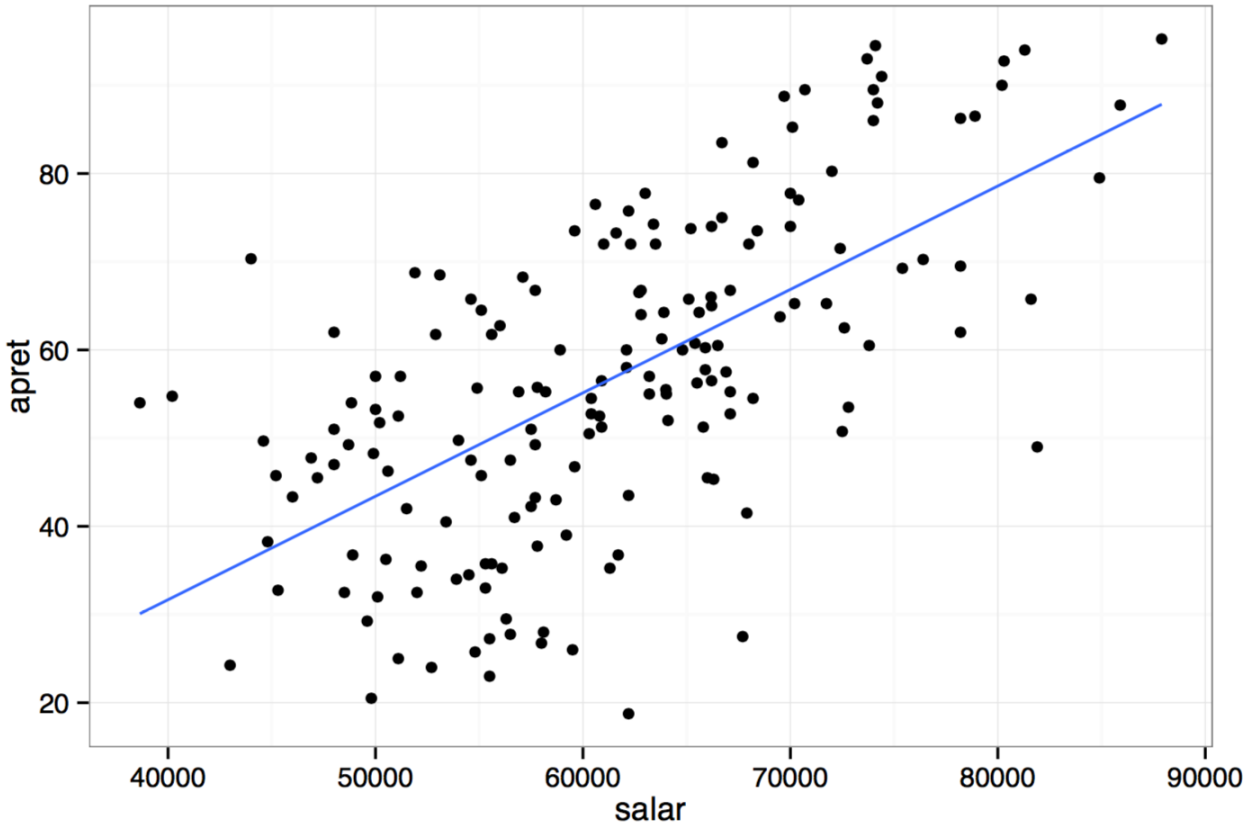
According to the histograms, it is easy to see that apret and tstsc have a normal distribution while the salar has a very sparse distribution

Then, perform linear regression of apret on tstsc and salar separately.

```
ggplot(data, aes(x = tstsc, y = apret)) + geom_point() +
geom_smooth(method=lm, # add linear regression line
se=FALSE)
```




```
ggplot(data, aes(x = salar, y = apret)) + geom_point() +
geom_smooth(method=lm, # add linear regression line
se=FALSE)
```



then, perform linear regression of apret on both tstsc and salar.

```
fit = lm(apret ~ tstsc+salar, data=data)
 summary(fit)
```

```
##
## Call:
## lm(formula = apret ~ tstsc + salar, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.458  -7.915   1.270   7.777  29.538
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.591e+01  8.210e+00  -9.246   <2e-16 ***
## tstsc        1.738e+00  1.761e-01   9.868   <2e-16 ***
## salar        2.880e-04  1.253e-04   2.298   0.0228 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.16 on 167 degrees of freedom
## Multiple R-squared:  0.6237, Adjusted R-squared:  0.6192
## F-statistic: 138.4 on 2 and 167 DF,  p-value: < 2.2e-16
```