Data Stucture Assignment 3

Jing Pang, Tian Xue, Chuqian Ma, Jiaxiang Leng

In the beginning, we load our data into R for further analysis.

```
retention <- read.delim("~/R-workspace/Retention.txt")
```

To be understanding the data pattern, we use summary command code to gather our results. It will show all the detail information, including mean, median, maximum and minimum of every column in the "retention" table.

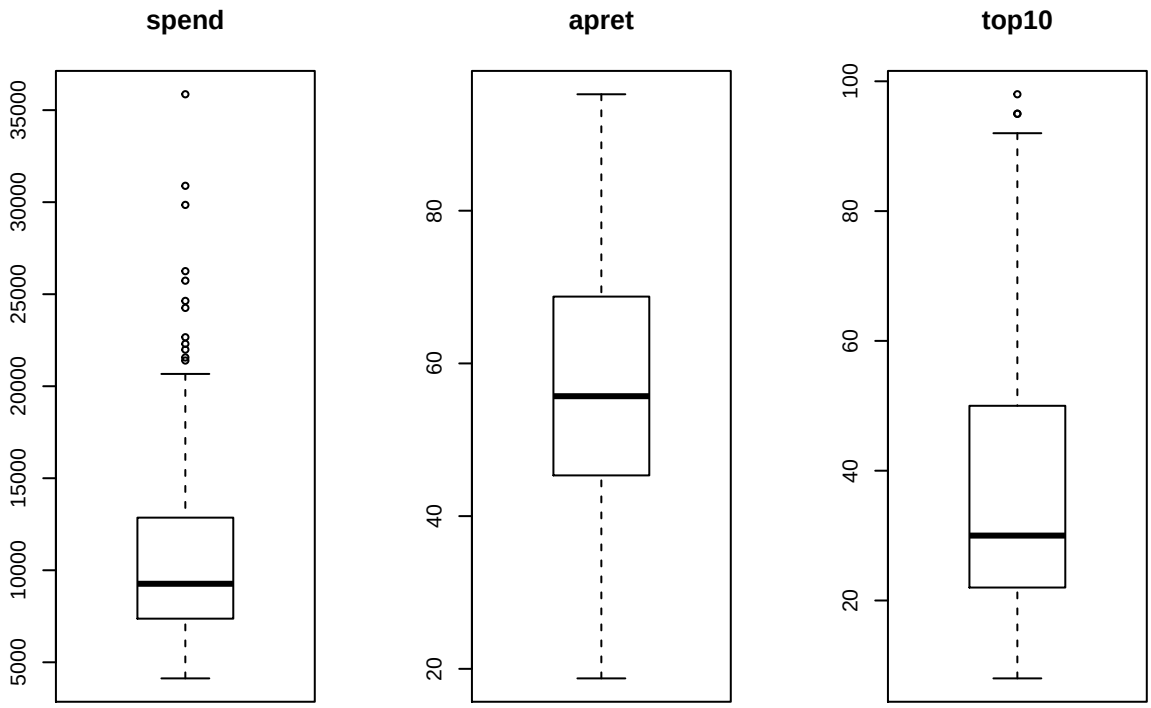
```
summary(retention)
```

```
##      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
```

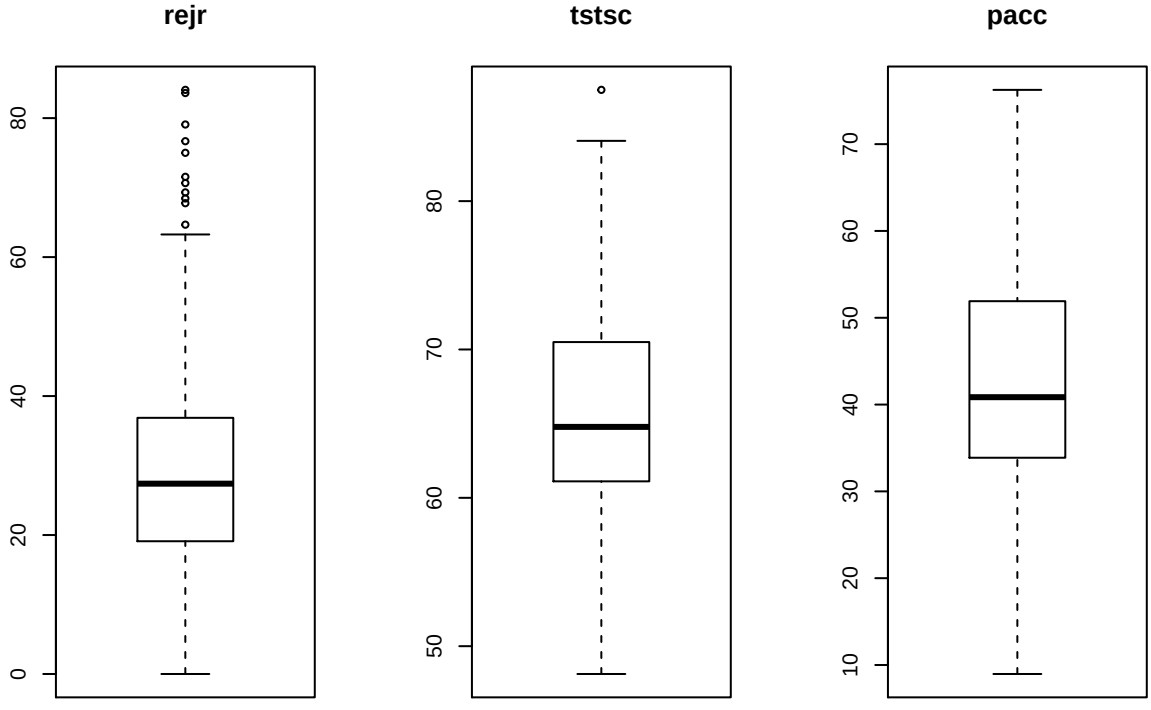
We could use boxplot to identify our outliers for each column easily. To achieve some basic statistics of each attribute, we use boxplot and histogram to reflect the distribution of them.

Observing the boxplots, each column of data have some outliers, especially in spend and "rejr" columns. This kind of problem requires more data to justify the data accuracy.

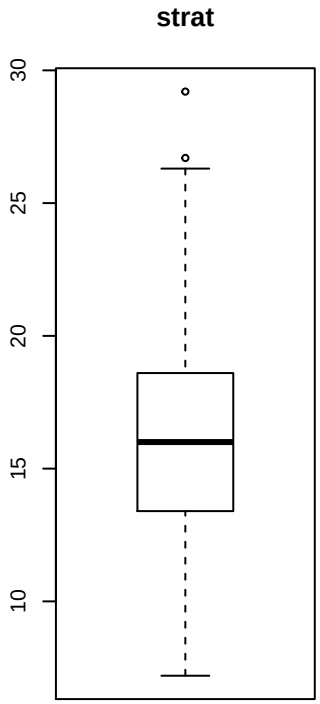
```
par(mfrow=c(1, 3))
boxplot(retention$spend, main="spend")
boxplot(retention$apret, main="apret")
boxplot(retention$top10, main="top10")
```




```
boxplot(retention$rejr, main="rejr")
boxplot(retention$tstsc, main="tstsc")
boxplot(retention$pacc, main="pacc")
```

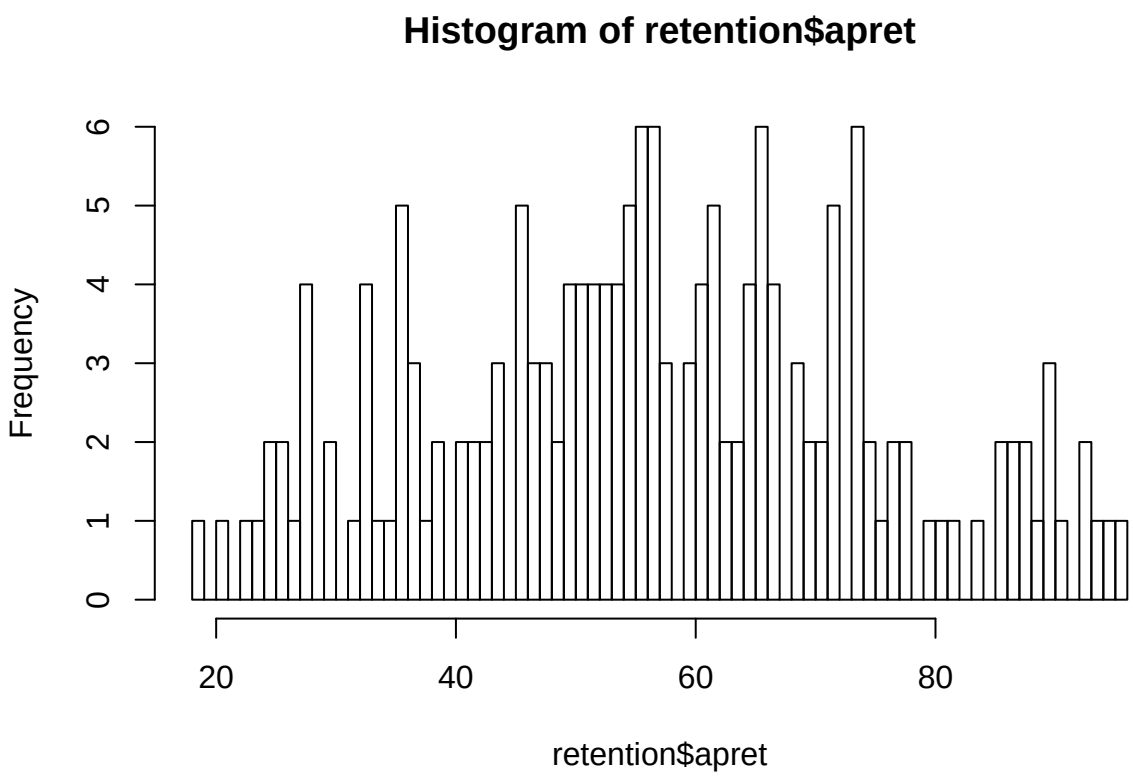



```
boxplot(retention$strat,main="strat")
```

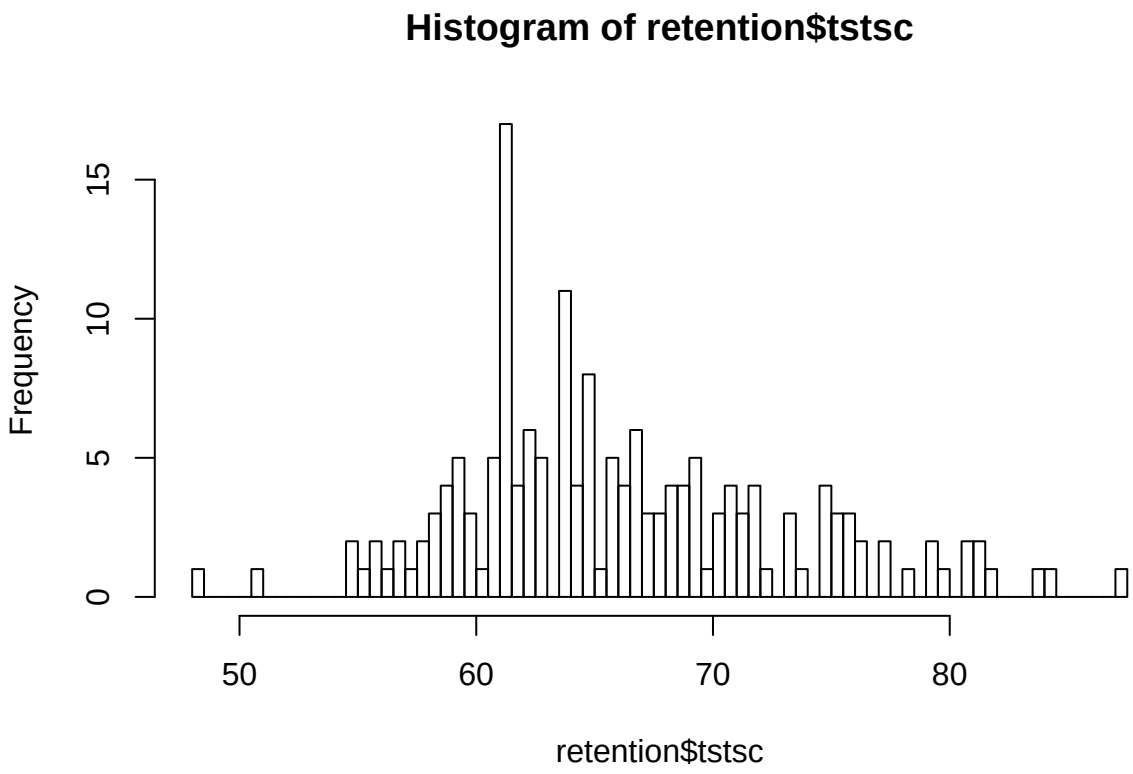


Then, we will look closer, consider to analysis three specific columns: apret, tstsc, and salar in histogram plot.Then, we will look closer, consider to study three particular columns: apret, tstsc, and salar in the histogram plot.

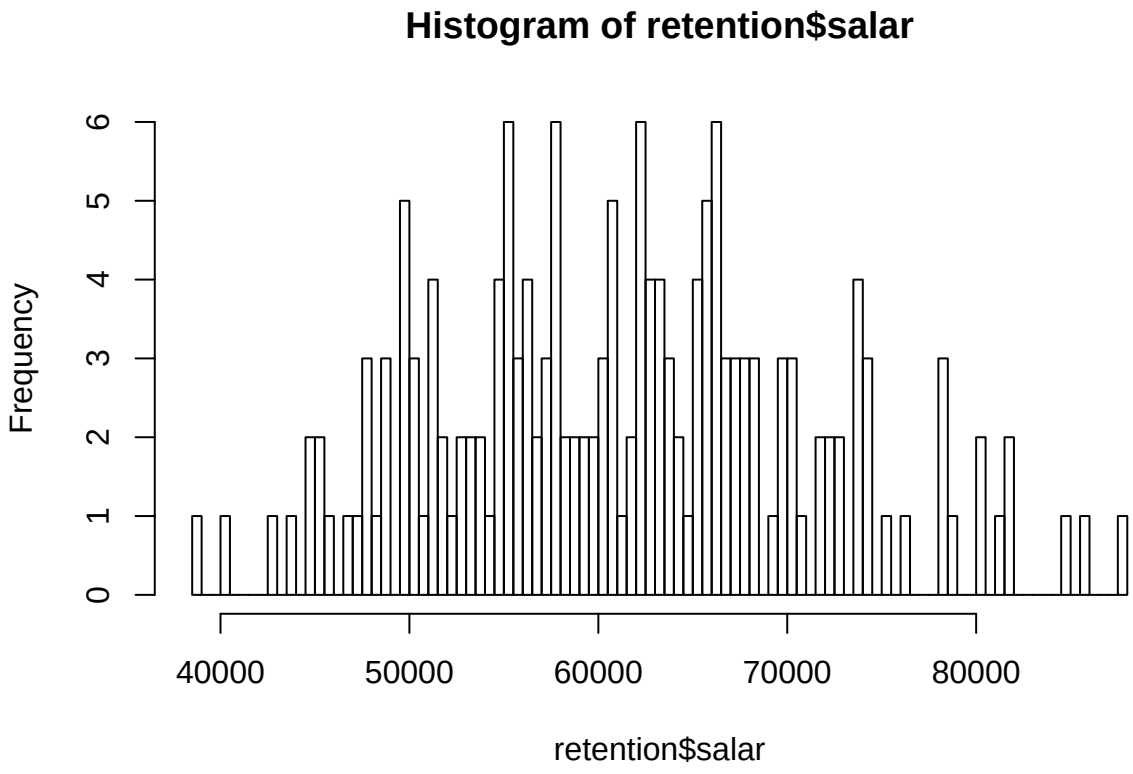
```
hist(retention$apret, 100)
```




```
hist(retention$tstsc, 100)
```

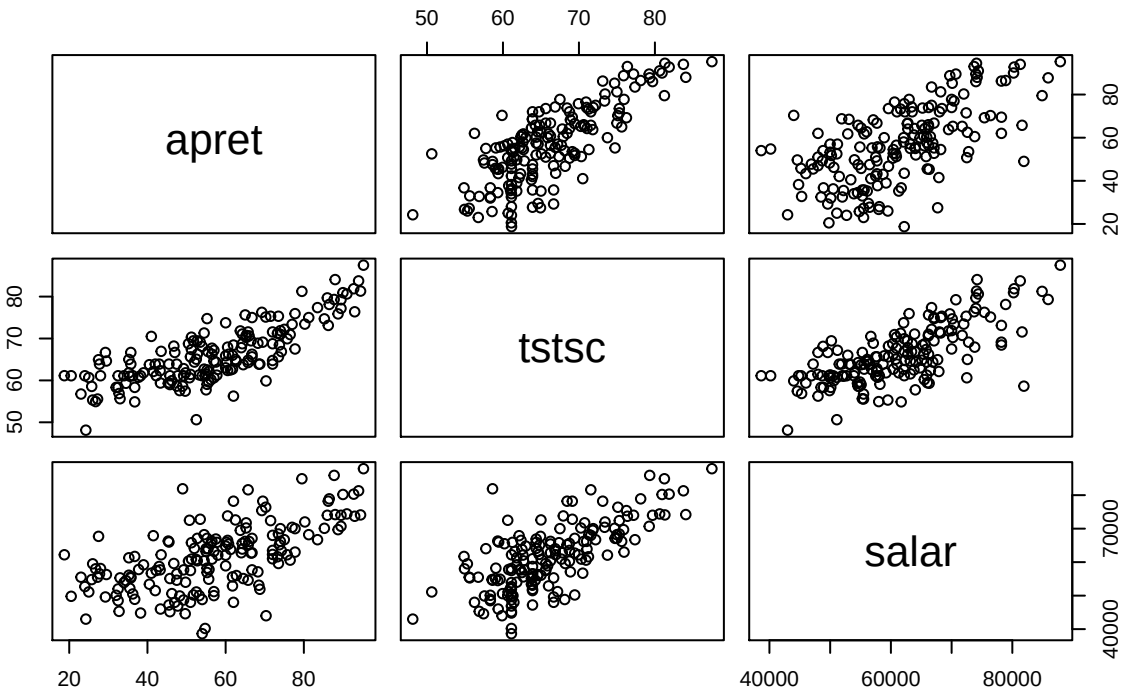



```
hist(retention$salar, 100)
```



To estimate the relationship in this three characters, we draw a dot plot graph in between each two of them.

```
pairs(~apret+tstsc+salar,data=retention, main="Cross Relationship Of Characters")
```

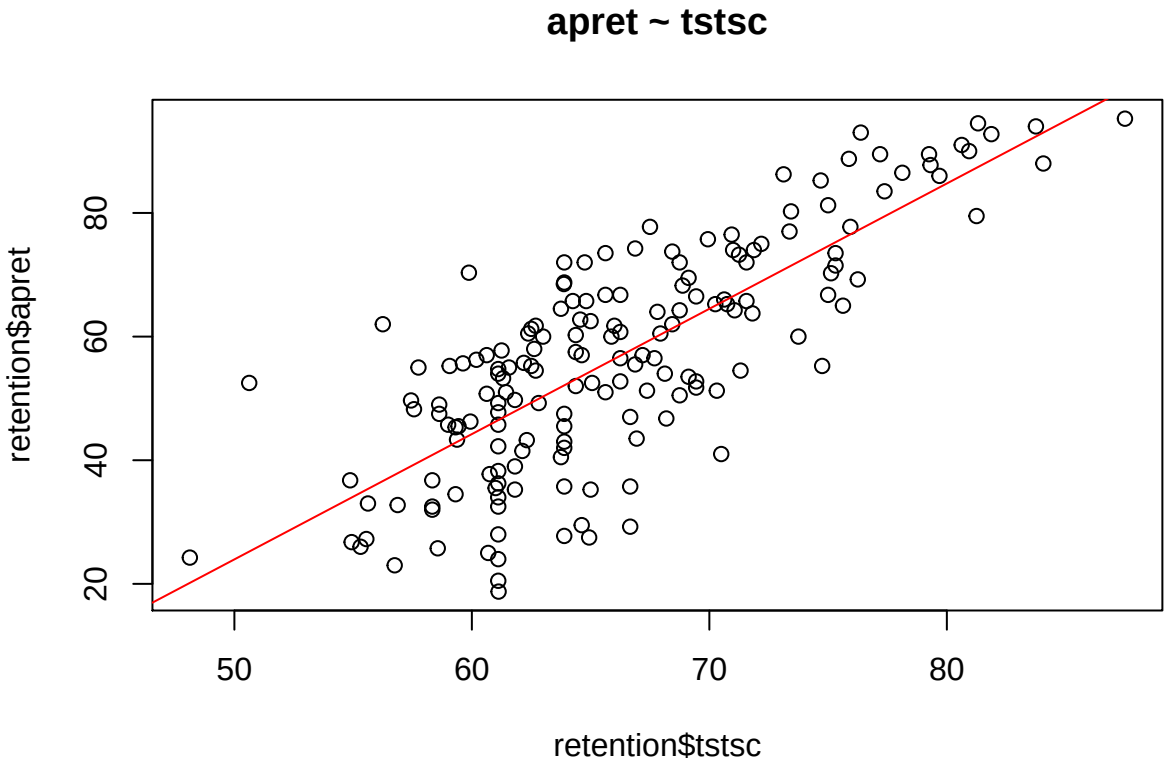


For further study, we plot a dot graph of "apret" character constructing separately with "tstsc" and "salar" characters. A linear relationship observes from figures. So we consider modifying a linear regression to predict the future outcomes.

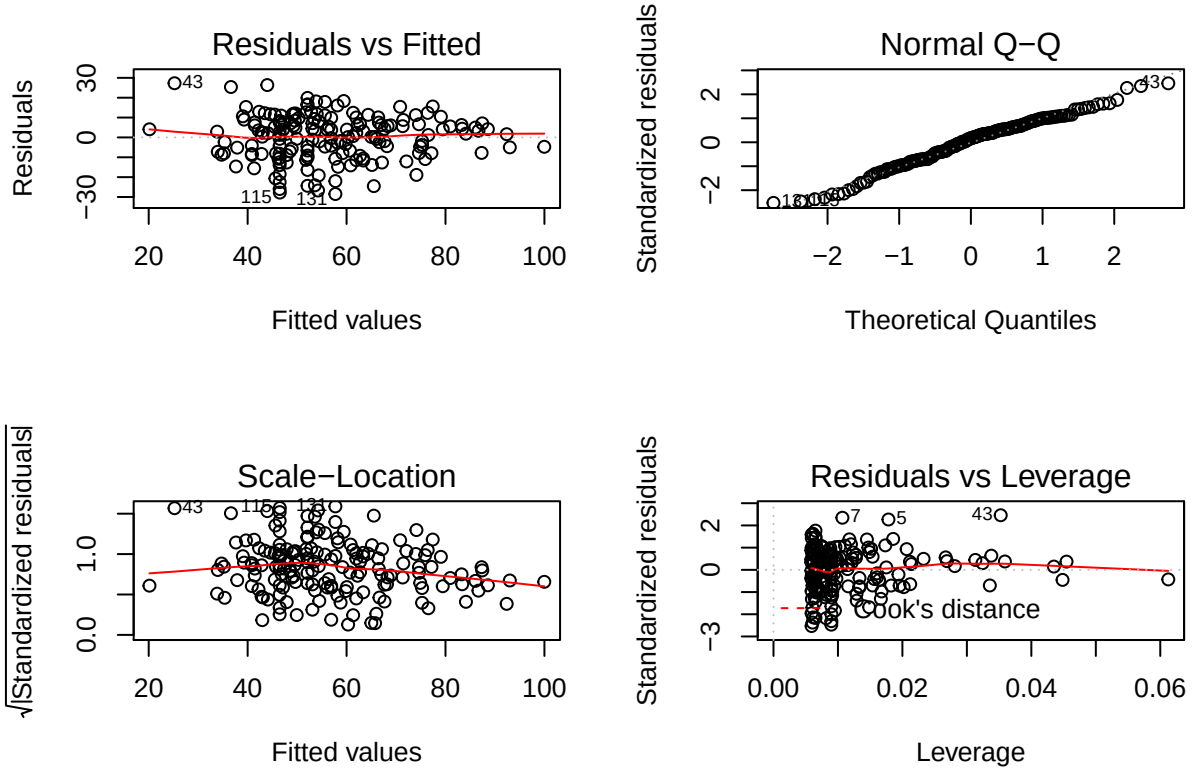
```
plot(y=retention$apret, x=retention$tstsc, main="apret ~ tstsc")
lm1 <- lm(apret ~ tstsc, data=retention)
summary(lm1)
```

```
##
## Call:
## lm(formula = apret ~ tstsc, data = retention)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.490  -7.957   1.857   7.552  27.278
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -77.3999     8.2878  -9.339   <2e-16 ***
## tstsc         2.0271     0.1246  16.272   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.3 on 168 degrees of freedom
## Multiple R-squared:  0.6118, Adjusted R-squared:  0.6095
## F-statistic: 264.8 on 1 and 168 DF,  p-value: < 2.2e-16
```

```
abline(lm1,col="red")
```




```
par(mfrow=c(2,2))
plot(lm1)
```

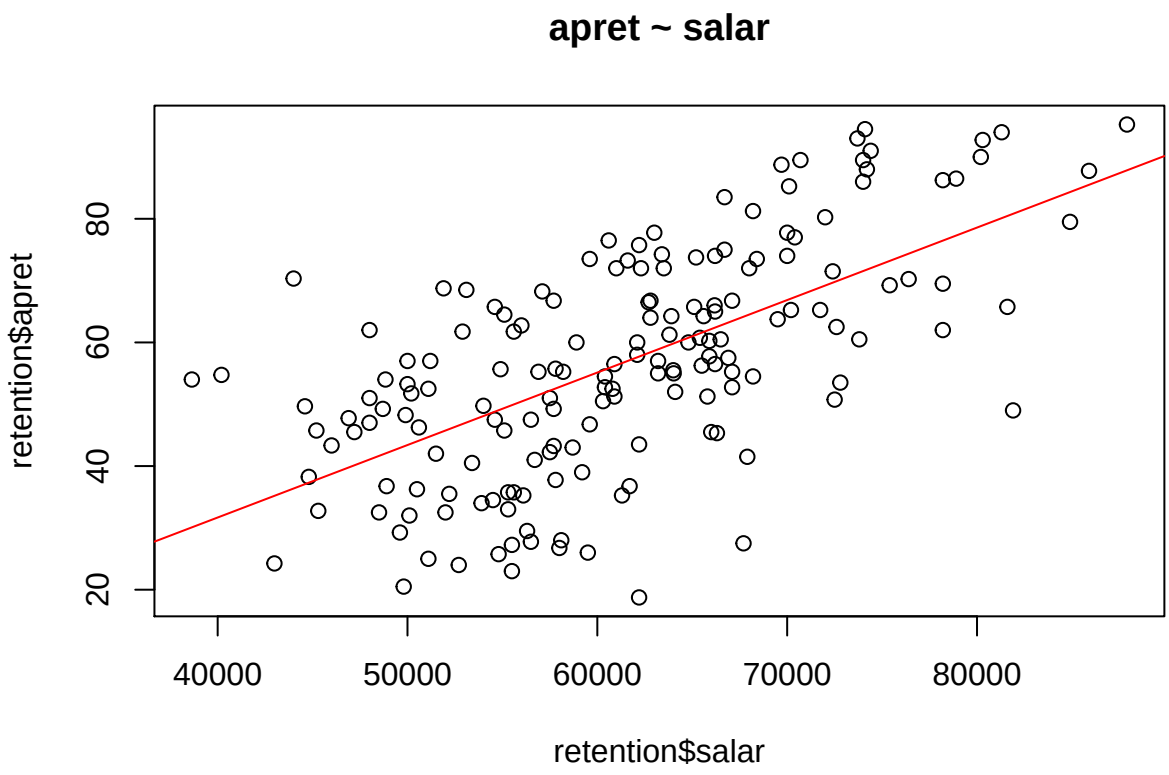



```
plot(y=retention$apret, x=retention$salar, main="apret ~ salar")
lm2 <- lm(apret ~ salar, data=retention)
```

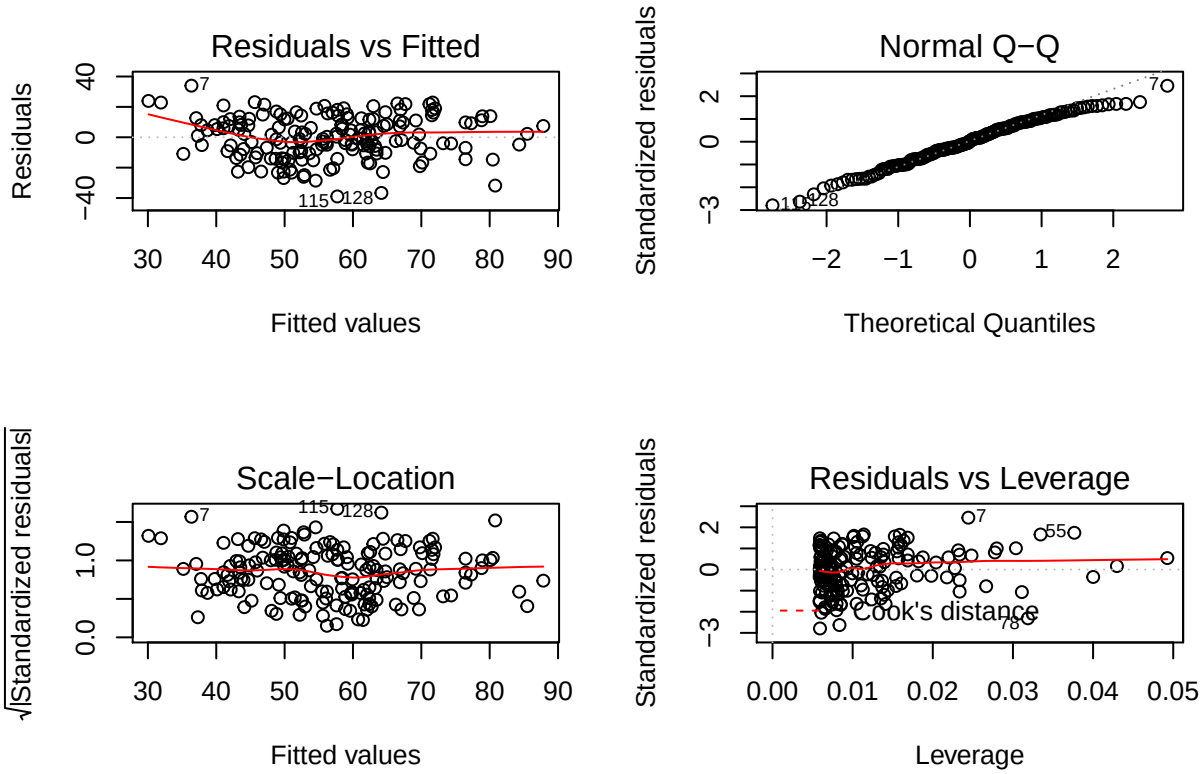
```r
summary(lm2)
```

```
##
## Call:
## lm(formula = apret ~ salar, data = retention)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.959 -10.170   0.362  11.151  33.965
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.522e+01  6.823e+00  -2.231    0.027 *
## salar        1.173e-03  1.098e-04  10.678   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.99 on 168 degrees of freedom
## Multiple R-squared:  0.4043, Adjusted R-squared:  0.4008
## F-statistic:   114 on 1 and 168 DF,  p-value: < 2.2e-16
```

```r
abline(lm2,col="red")
```




```r
par(mfrow=c(2,2))
plot(lm2)
```



Performing "apret" character on both "tstsc" and "salar" characters displays how the linear relationship occurs on all three characters. The predictable regression function happening to be a plane supports our assumption at the beginning.

```
library(scatterplot3d)
```

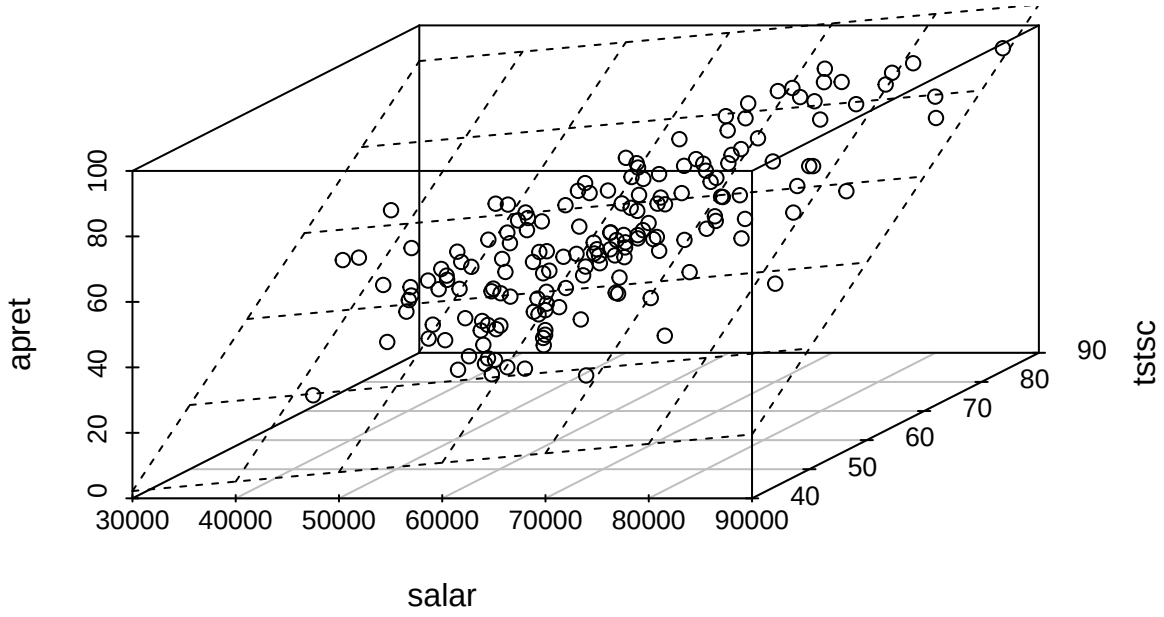
```
## Warning: package 'scatterplot3d' was built under R version 3.4.4
```

```
attach(retention)
s3dplot<- scatterplot3d(salar,tstsc,apret)
lm3 <- lm(apret~salar+tstsc)
summary(lm3)
```

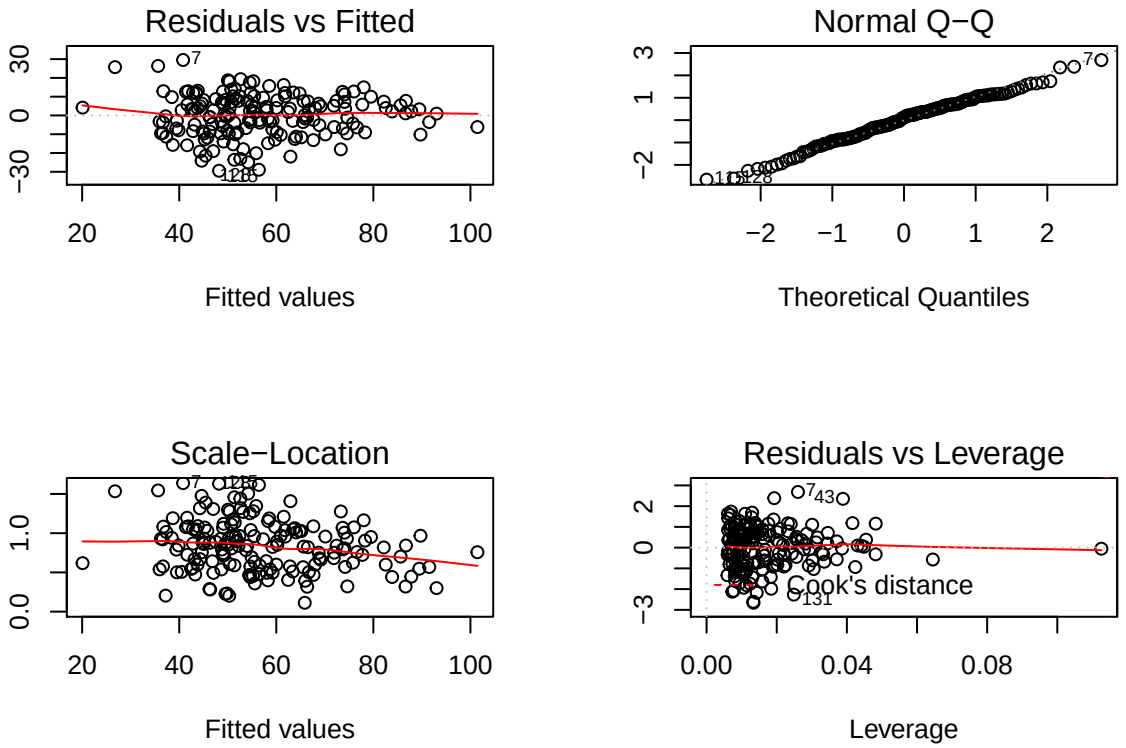
```
##
## Call:
## lm(formula = apret ~ salar + tstsc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.458  -7.915   1.270   7.777  29.538
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.591e+01  8.210e+00  -9.246   <2e-16 ***
## salar        2.880e-04  1.253e-04   2.298   0.0228 *
## tstsc        1.738e+00  1.761e-01   9.868   <2e-16 ***
## ---
## 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
```

```
s3dplot$plane3d(lm3)
```




```
par(mfrow=c(2,2))
plot(lm3)
```



We also can test the relationship through another method Anova.

```
fit1 <- lm(apret ~ tstsc + salar, data=retention)
fit2 <- lm(apret ~ tstsc, data=retention)
fit3 <- lm(apret ~ salar, data=retention)
```

```
anova(fit1, fit2, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: apret ~ tstsc + salar
## Model 2: apret ~ tstsc
## Model 3: apret ~ salar
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1    167 20781
## 2    168 21438 -1    -657.3 5.2826 0.02278 *
## 3    168 32898  0  -11459.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Also, we could use the correlation function to find other characters linear correlation. In this part, we test correction in two different methods Spearman and Pearson. If the value is closer to zero, the relationship of the linear correlation between them is weaker, which means they are more independent on each other.

```
cor(retention, method="spearman")
```

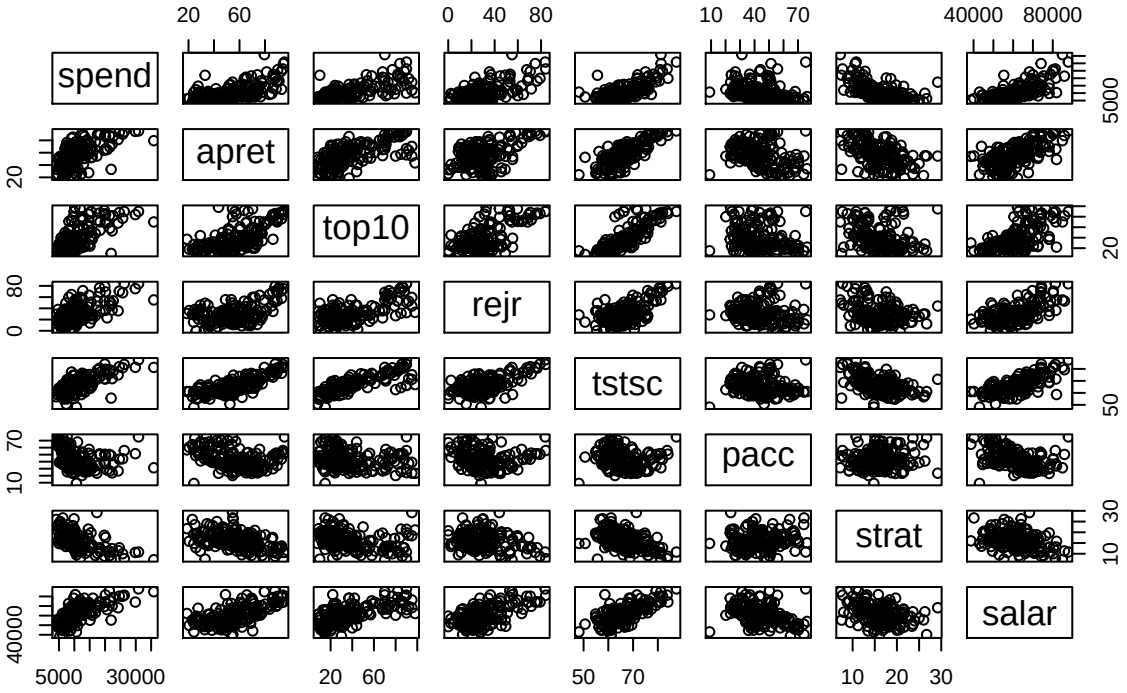
```
##            spend      apret      top10       rejr      tstsc       pacc
## spend  1.0000000  0.5859626  0.6280528  0.5237807  0.6584567 -0.4038082
## apret  0.5859626  1.0000000  0.6369981  0.3520061  0.7496700 -0.3469449
## top10  0.6280528  0.6369981  1.0000000  0.4507527  0.8222439 -0.2446597
## rejr   0.5237807  0.3520061  0.4507527  1.0000000  0.4695680 -0.1591321
## tstsc  0.6584567  0.7496700  0.8222439  0.4695680  1.0000000 -0.2823545
## pacc  -0.4038082 -0.3469449 -0.2446597 -0.1591321 -0.2823545  1.0000000
## strat -0.5894018 -0.4618781 -0.3213480 -0.2142416 -0.4697187  0.1383233
## salar  0.7321529  0.6387381  0.6135785  0.5500964  0.6936353 -0.4295698
##            strat      salar
## spend -0.5894018  0.7321529
## apret -0.4618781  0.6387381
## top10 -0.3213480  0.6135785
## rejr  -0.2142416  0.5500964
## tstsc -0.4697187  0.6936353
## pacc   0.1383233 -0.4295698
## strat  1.0000000 -0.3272242
## salar -0.3272242  1.0000000
```

```
cor(retention, method="pearson")
```

```
##            spend      apret      top10        rejr      tstsc        pacc
## spend  1.0000000  0.6012312  0.6756556  0.63354382  0.7149101 -0.23673000
## apret  0.6012312  1.0000000  0.6424645  0.51495797  0.7821831 -0.30283389
## top10  0.6756556  0.6424645  1.0000000  0.64316348  0.7988074 -0.20750524
## rejr   0.6335438  0.5149580  0.6431635  1.00000000  0.6286011 -0.07152073
## tstsc  0.7149101  0.7821831  0.7988074  0.62860107  1.0000000 -0.16422305
## pacc  -0.2367300 -0.3028339 -0.2075052 -0.07152073 -0.1642230  1.00000000
## strat -0.5617553 -0.4583114 -0.2478568 -0.28361659 -0.4652263  0.13185837
## salar  0.7118376  0.6358517  0.6376482  0.60677651  0.7154715 -0.37524020
##            strat      salar
## spend -0.5617553  0.7118376
## apret -0.4583114  0.6358517
## top10 -0.2478568  0.6376482
## rejr  -0.2836166  0.6067765
## tstsc -0.4652263  0.7154715
```

```
## pacc   0.1318584 -0.3752402
## strat  1.0000000 -0.3476728
## salar -0.3476728  1.0000000
```

```
pairs(~spend+apret+top10+rejr+tstsc+pacc+strat+salar,data=retention,
   main="")
```



In the end, we try to use GeNIe digging more statistical understanding from the dataset. This program helps us locate a Bayesian network between each character. The relationship diagram shows on the following.