DATA ANALYTICS

Assignment two

February 14, 2016

Assignment 2 : Fundamental concepts from statistics

- **Group members:**

- Yifan Zhao
- Nanxun Xie

- **Using System** : RStudio

- **Using Programming language** : R, MATLAB

preparation

STEP 1: Read Retention.txt to RStudio.

R code:

```
# read txt file "Retention" into a data frame
Retention <- read.table("Retention.txt",header = TRUE)
```

STEP 2: Analyze data by fundamental statistics

Pre-requested jars:

```
# package rJava, xlsx and xlsxjars are used to tramsform the data frame to xlsx file.
library(rJava)
library(xlsxjars)
library(xlsx)
# package psych contain the "describe" function to generate descriptive statistics
library(psych)
```

Question1:generate descriptive statistics and plot histograms for the following three columns: apret, tstsc, and salar.

- **descriptive statistics**

R code:

```
# create a vector "s" with three specific strings "apret","tstsc","salar" which we used to do some analytics
s <- c("apret","tstsc","salar")
# call describe function to generate descriptive statistics of three specific colunms, and then put the result into a data frame “x"
#Retention[s] stored three specific colunms of data given by vector"s"
x <- describe(Retention[s])
# tramsform data frame "x" to a xlsx file "statistics.xlsx"
write.xlsx(x,"statistics.xlsx",row.names = TRUE,col.names = TRUE)
```

the result shown below:

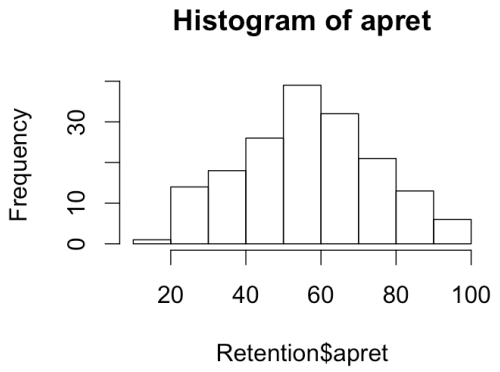
	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
apret	1	170	56.72108	18.077097	55.7085	56.42157	18.40944	18.750	95.25	76.500	0.08761814	-0.6018289	1.3864500
tstsc	2	170	66.16416	6.975306	64.7815	65.70476	5.93040	48.125	87.50	39.375	0.56314164	0.1185932	0.5349816
salar	3	170	61357.64706	9802.786457	61150.0000	61050.99265	9340.38000	38640.000	87900.00	49260.000	0.25334376	-0.2915902	751.8394005

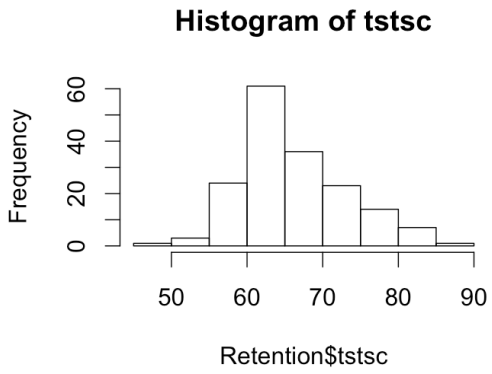
❖ **histograms**

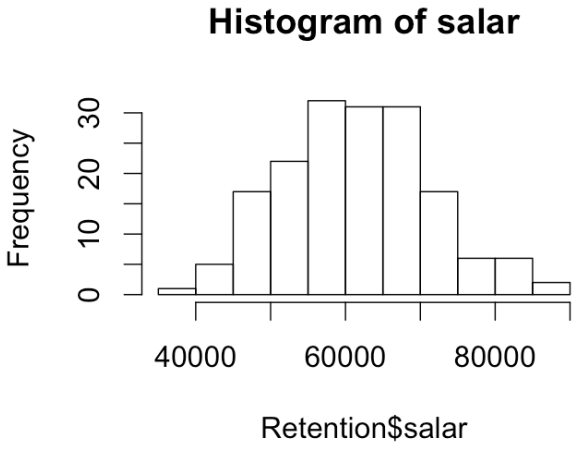
R code:

```
# generete plot histograms for the following three colunms: apret, tstsc, and salar
# call hist function to draw histograms for a given colunm of data.
# main argument set the head name of histograms
hist(Retention$apret, main="Histogram of apret")
hist(Retention$tstsc, main="Histogram of tstsc")
hist(Retention$salar, main="Histogram of salar”)
```

the result shown below:







Question2: perform linear regression of apret on tstsc and salar separately and then of apret on both tstsc and salar.

❖ **linear regression of apart on tstsc**

R code:

```
# perform linear regression of apret on tstsc
# call the plot function, using the "tstsc" column of countries for the horizontal axis, and the "apret" column for the vertical axis
plot(Retention$apret ~ Retention$tstsc)
# call lm function to fit linear model, given a response variable(apret),and a predictor variable(tstsc),and put the result into a vector Retention.reg
Retention.reg <- lm(Retention$apret ~ Retention$tstsc, data = Retention)
# call abline function to draw a line fitting the data on the plot
abline(Retention.reg , col = 2, lty = 2)
# call summary function to produce result summaries of the results of various model fitting functions.
summary(Retention.reg)
# call anova function to compute an analysis of variance table for the linear model fits.
anova(Retention.reg)
```

the result shown below:

- summary function

```
Call:
lm(formula = Retention$apret ~ Retention$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 ***
Retention$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
```

from the result, we can get the linear equation：

$$y = 2.0271x - 77.3999$$

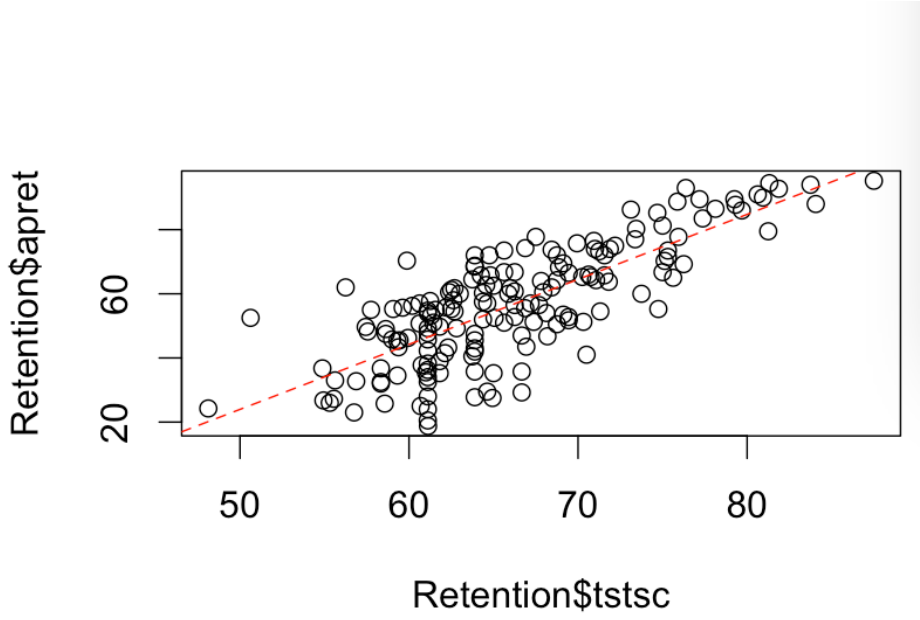
y: apret
x: tstsc

- anova function

```
Analysis of Variance Table

Response: Retention$apret
                 Df Sum Sq Mean Sq F value    Pr(>F)
Retention$tstsc   1  33788   33788  264.78 < 2.2e-16 ***
Residuals       168  21438     128
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- plot and fitting line



❖ **linear regression of apart on salar**

R code:

```
# perform linear regression of apret on salar
# call the plot function, using the "salar" column of countries for the horizontal axis, and the "apret" column for the vertical axis
plot(Retention$apret ~ Retention$salar)
# call lm function to fit linear model, given a response variable(apret),and a predictor variable(salar)
Retention.reg <- lm(Retention$apret ~ Retention$salar, data = Retention)
# call abline function to draw a line fitting the data on the plot
abline(Retention.reg, col = 2, lty = 2)
# call summary function to produce result summaries of the results of various model fitting functions.
summary(Retention.reg)
# call anova function to compute an analysis of variance table for the linear model fits.
anova(Retention.reg)
```

- summary function

```
Call:
lm(formula = Retention$apret ~ Retention$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 *
Retention$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
```

from the result, we can get the linear equation：

$y = 0.001173x - 15.22$

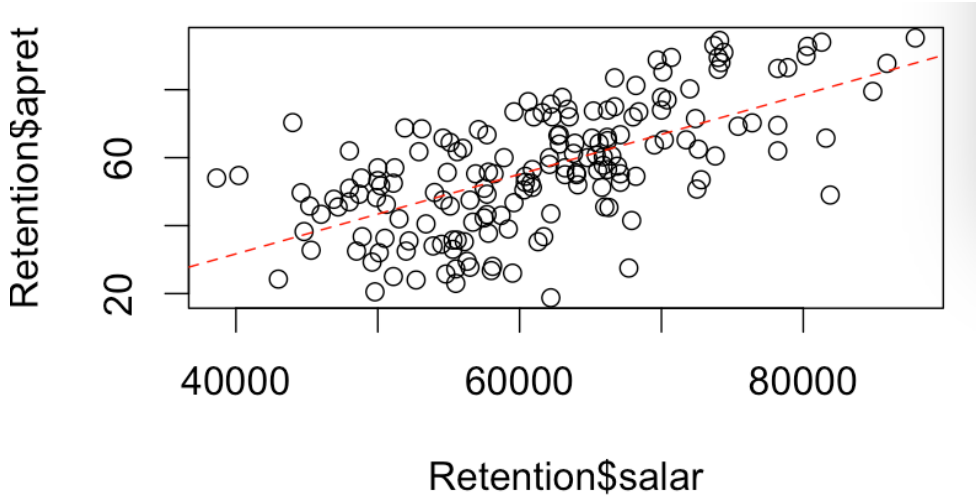
y: apret
x: salar

- anova function

```
Analysis of Variance Table

Response: Retention$apret
                 Df Sum Sq Mean Sq F value    Pr(>F)
Retention$salar   1  22328 22328.3  114.02 < 2.2e-16 ***
Residuals       168  32898   195.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- plot and fitting line



❖ **linear regression of apart on tstsc and salar**

R code:

```
# perform linear regression of apret on tstsc and salar
# call lm function to fit linear mode, given a response variable(apret),and a predictor variable(salar+tstsc)
Retention.reg <- lm(Retention$apret ~ Retention$salar + Retention$tstsc , data = Retention)
# call summary function to produce result summaries of the results of various model fitting functions.
summary(Retention.reg)
# call anova function to compute an analysis of variance table for the linear model fits.
anova(Retention.reg)
```

- summary function

```
Call:
lm(formula = Retention$apret ~ Retention$salar + Retention$tstsc,
    data = Retention)

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 ***
Retention$salar  2.880e-04  1.253e-04   2.298   0.0228 *
Retention$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
```

from the result, we can get the linear equation：

z＝ 0.000288y + 1.738x - 75.91

z: apret
y: salar
x: tstsc

- anova function

```
Analysis of Variance Table

Response: Retention$apret
                 Df Sum Sq Mean Sq F value    Pr(>F)
Retention$salar   1  22328 22328.3 179.436 < 2.2e-16 ***
Retention$tstsc   1  12117 12116.9  97.375 < 2.2e-16 ***
Residuals       167  20781   124.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- plot and fitting line(use MATLAB)