Yue Su & Dan Sun

Assignment 2

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

- Generate Descriptive Statistics

Used R language to generate descriptive statistic table.

```
> Rowname = c("spend", "apret", "top10", "rejr", "tstsc", "pacc", "strat", "salar")
> Columnname = c("Min.","1stQuartile","Median","Mean","3rdQuartile","Max.","StdDev","Count")
> spend = c(as.vector(summary(Retention$spend)), sd(Retention$spend), nrow(Retention))
> apret = c(as.vector(summary(Retention$apret)), sd(Retention$apret), nrow(Retention))
> top10 = c(as.vector(summary(Retention$top10)), sd(Retention$top10), nrow(Retention))
> rejr = c(as.vector(summary(Retention$rejr)), sd(Retention$rejr), nrow(Retention))
> tstsc = c(as.vector(summary(Retention$tstsc)), sd(Retention$tstsc), nrow(Retention))
> pacc = c(as.vector(summary(Retention$pacc)), sd(Retention$pacc), nrow(Retention))
> strat = c(as.vector(summary(Retention$strat)), sd(Retention$strat), nrow(Retention))
> salar = c(as.vector(summary(Retention$salar)), sd(Retention$salar), nrow(Retention))
>
> Summary = matrix(c(spend, apret, top10, rejr, tstsc, pacc, strat, salar), nrow = 8, ncol = 8, byrow = TRUE, dimnames = list(Rowname, Columnname))
```

Table shown below is the result.

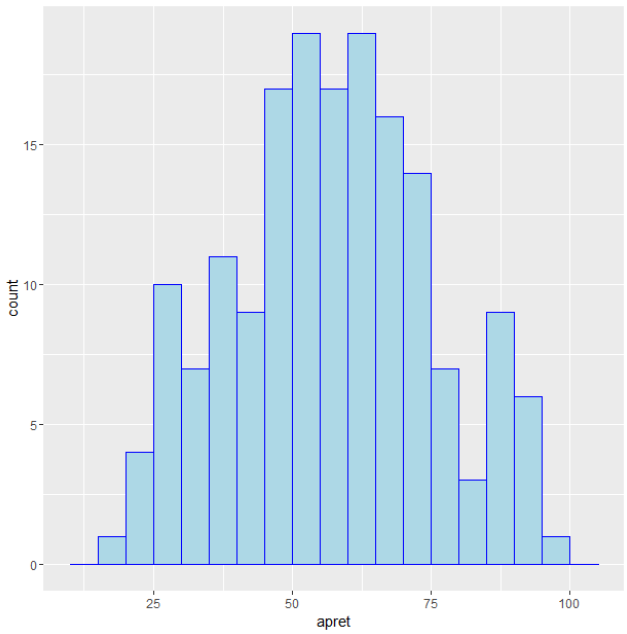
Data: Summary

	row.names	Min.	1stQuartile	Median	Mean	3rdQuartile	Max.	StdDev	Count
1	spend	4125.000	7372.00	9265.00	10970.00	12840.00	35860.00	5500.065580	170
2	apret	18.750	45.37	55.71	56.72	68.69	95.25	18.077097	170
3	top10	8.000	22.00	30.00	38.46	49.50	98.00	23.406393	170
4	rejr	0.000	19.17	27.39	30.65	36.81	84.07	17.098104	170
5	tstsc	48.120	61.11	64.78	66.16	70.45	87.50	6.975306	170
6	pacc	8.964	33.90	40.85	43.17	51.77	76.25	13.105195	170
7	strat	7.200	13.40	16.00	16.09	18.58	29.20	4.006503	170
8	salar	38640.000	54650.00	61150.00	61360.00	67100.00	87900.00	9802.786457	170

- Plot Histograms

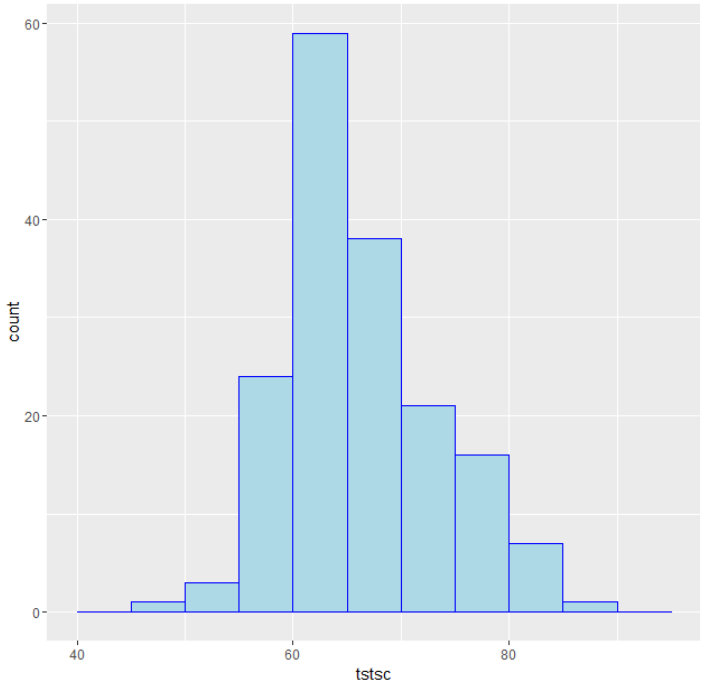
apret:

ggplot(Retention,aes(apret))+geom_histogram(binwidth=5, fill="LightBlue", colour="Blue")



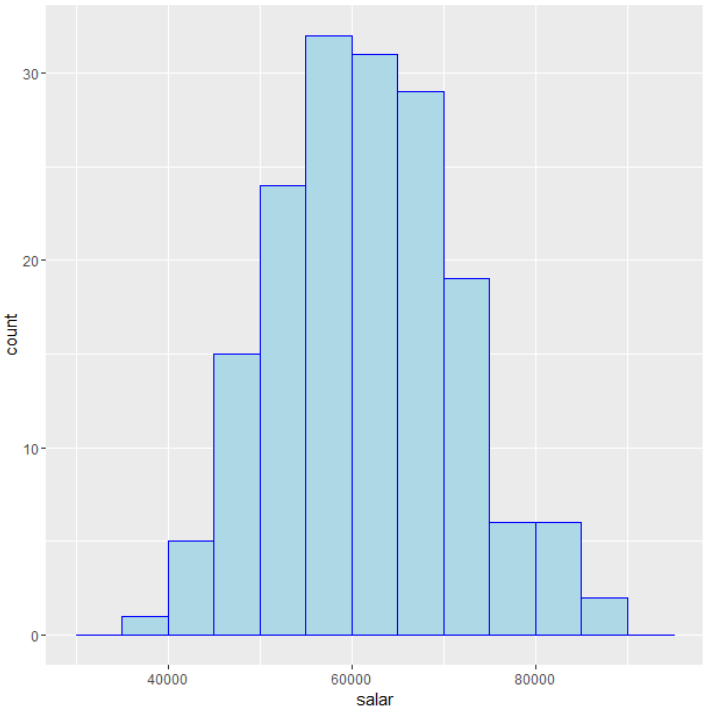
tstsc:

```
ggplot(Retention,aes(tstsc))+geom_histogram(binwidth=5, fill="LightBlue", colour="Blue")
```



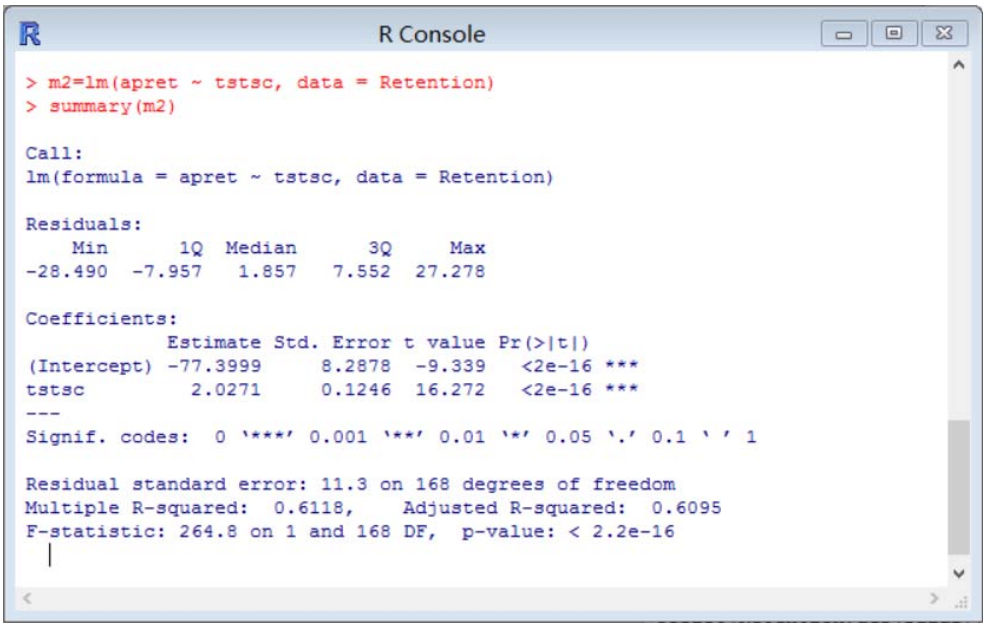
salar:

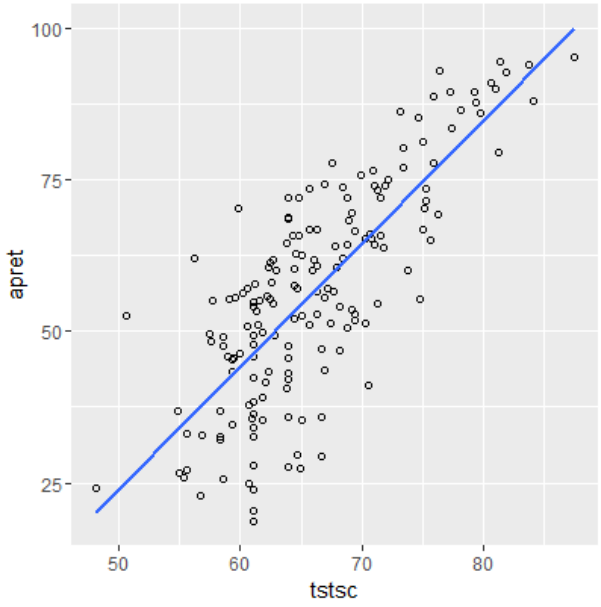
```
ggplot(Retention,aes(salar))+geom_histogram(binwidth=2000, fill="LightBlue",
colour="Blue")
```



2. Perform linear regression of apret on tstsc and salar separately and then of apret on both tstsc and salar.

- tstsc vs. apret



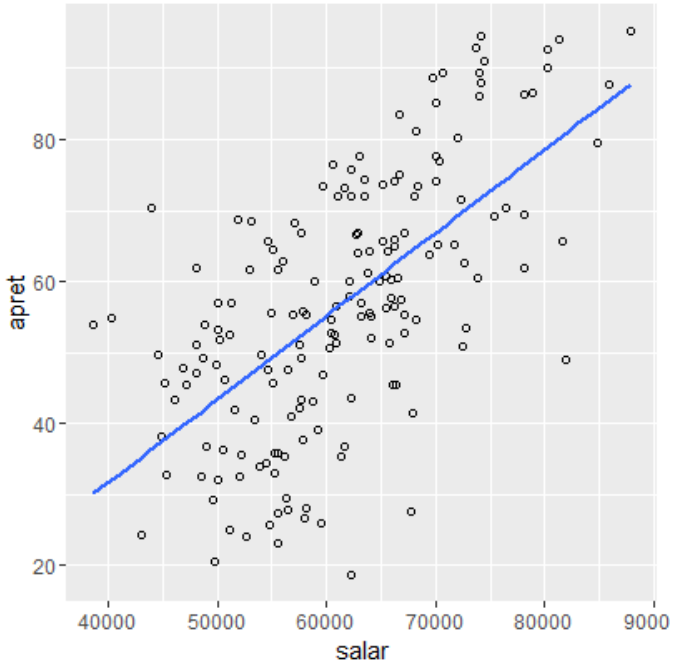


From the graph and summary table, we can conclude that variable tstsc and apret has a strong correlation. Tstsc is a significant factor for apret.

- salar vs. apret

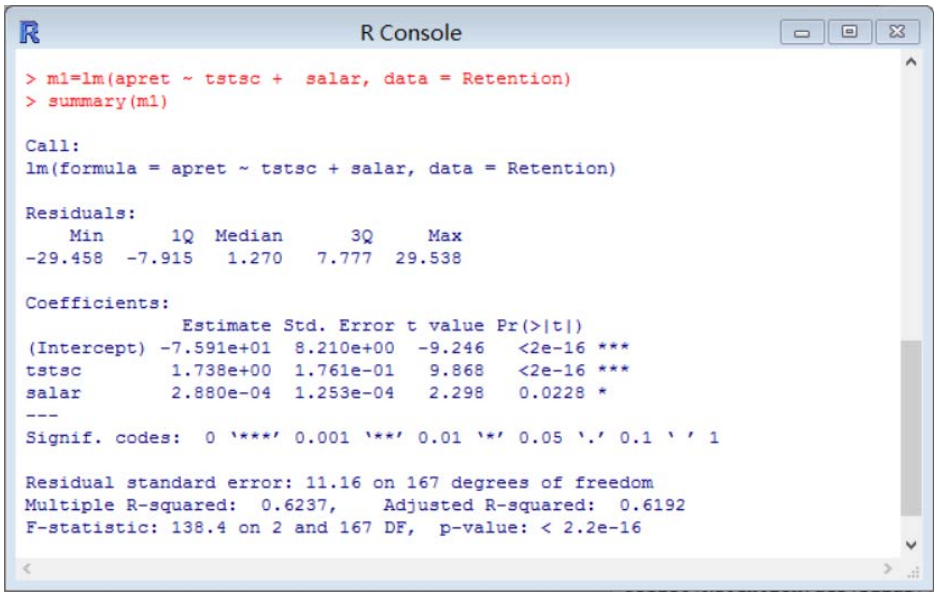
R Console

```
> m3=lm(apret ~ salar, data = Retention)
> summary(m3)

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
```



From the graph and summary table we can conclude that variable salar and apret has a loose correlation between each other. The points are distributed dispersedly and the R-square is 0.4043, smaller than the linear regression of apret on tstsc.

- tstsc, salar vs. apret



The summary table above indicates that linear regression of apret on both tstsc and salar is a good model. Based on the coefficient values and R-squared values, we can conclude that apret has a strong correlation together with both tstsc and salar.