MAX-503 Assignment 3

```
setwd("C:/Users/Hema/Desktop/Market Research/Mark Research - Assignment 3")
getwd()
```

```
## [1] "C:/Users/Hema/Desktop/Market Research/Mark Research - Assignment 3"
```

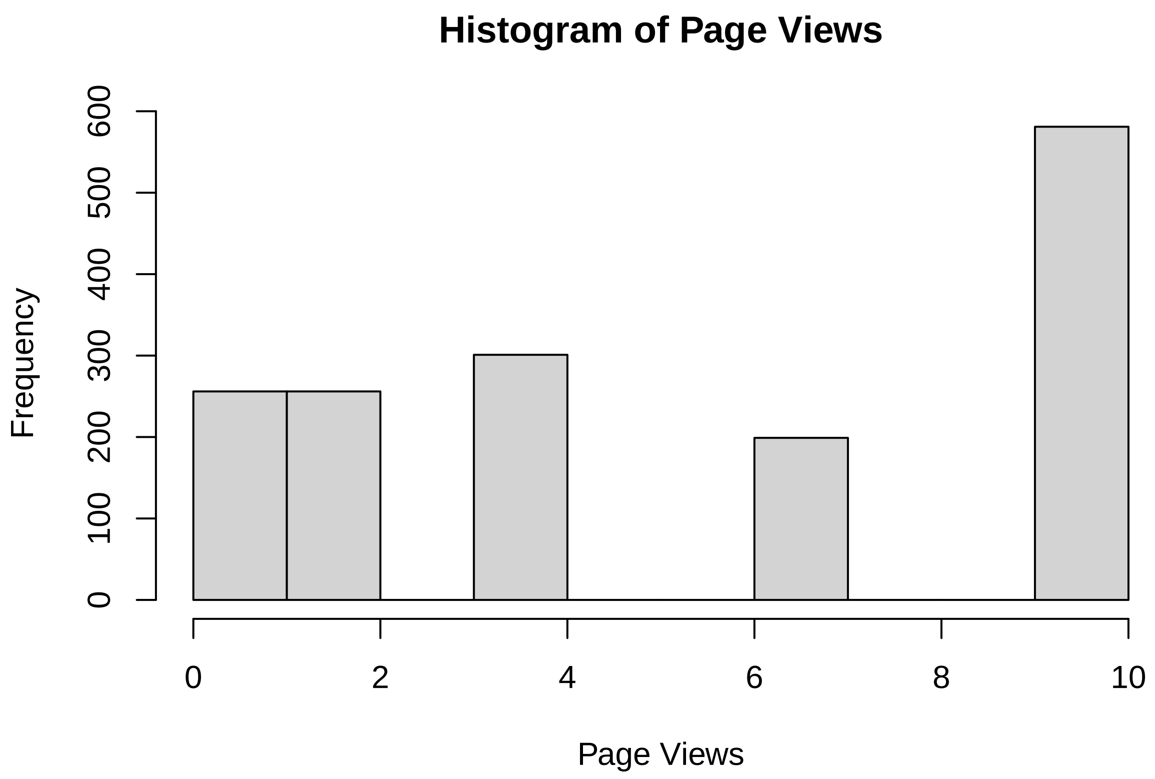
```
ecomm.df <- read.csv("ecommerce-data.csv")
str(ecomm.df)
```

```
## 'data.frame':    1593 obs. of  45 variables:
##  $ dateTime                   : chr  "7/25/2014 14:10" "7/25/2014 15:01" "7/25/2014 15:15" "7/25/2014
##  $ country                    : chr  "United States" "United States" "United States" "United States"
##  $ city                       : chr  "Monroe" "Ambler" "Beaumont" "Cedar City" ...
##  $ region                     : chr  "LA" "PA" "TX" "UT" ...
##  $ screenRed                  : chr  "1280x1024" "1280x800" "768x1024" "1280x960" ...
##  $ surveyType                 : chr  "At Exit" "At Exit" "At Exit" "At Exit" ...
##  $ purposeProductInfo         : chr  "Products" "" "" "Products" ...
##  $ purposeBuyFromSite         : chr  "" "Buy from this site" "" "" ...
##  $ purposeComparePricing      : chr  "" "Compare pricing" "Compare pricing" "" ...
##  $ purposeInfoAndResources    : chr  "Resources" "" "" "" ...
##  $ purposeInfoOnOrder         : chr  "" "" "" "" ...
##  $ purposeOther               : chr  "" "" "" "" ...
##  $ taskFindWhatLookingFor     : chr  "" "" "" "Most or all of it" ...
##  $ concernShippingCost        : chr  "" "" "" "Shipping costs" ...
##  $ concernDeliverySpeed       : chr  "" "" "" "" ...
##  $ concernWarranties          : chr  "" "" "" "" ...
##  $ concernEaseToReturnProduct : chr  "Ease of returning (if I am not satisfied with product)" "" ""
##  $ concernProductSafety       : chr  "" "" "" "" ...
##  $ concernRightForMyChild     : chr  "" "" "" "" ...
##  $ concernProductQuality      : chr  "Product durability/quality" "" "" "" ...
##  $ concernProductEffectiveness: chr  "Product effectiveness/will it work" "" "" "" ...
##  $ concernOther               : chr  "" "" "" "" ...
##  $ concernNone                : chr  "" "" "" "" ...
##  $ intentWasPlanningToBuy     : chr  "" "Yes" "" "" ...
##  $ profile                    : chr  "Parent" "Parent" "Parent" "Person with [condition A]" ...
##  $ whenSiteUsed               : chr  "In the past week" "In the past year" "Never. This is my first
##  $ purchasedBefore            : chr  "Yes, once" "Yes, once" "" "" ...
##  $ purchasedWhen              : chr  "In the past month" "In the past year" "" "" ...
##  $ productKnewWhatWanted      : chr  "Yes" "Yes" "Yes" "Somewhat" ...
##  $ productSiteHasWhatWanted   : chr  "" "" "" "Yes, several of them" ...
##  $ purchaseExpectInNextMonth  : int  5 3 3 3 5 3 5 NA 5 4 ...
##  $ siteFirstHeardAbout        : chr  "In the past year" "More than 1 year ago" "Just now, from the we
##  $ age                        : chr  "25-34" "35-44" "35-44" "25-34" ...
##  $ gender                     : chr  "Female" "Female" "Female" "Female" ...
##  $ behavNumVisits             : int  13 3 2 1 1 1 4 1 2 2 ...
##  $ behavReferral              : chr  "Direct" "Unbranded Search" "Unbranded Search" "Unbranded Searc
##  $ behavPageviews           : chr  "4 to 6" "1" "10+" "10+" ...
##  $ behavHomePage            : int  1 0 0 0 0 1 0 1 1 1 ...
##  $ behavDetailProdA         : int  1 0 0 1 0 1 1 0 1 1 ...
##  $ behavDetailProdB         : int  0 0 0 1 0 1 1 1 1 0 ...
##  $ behavDetailProdC         : int  0 0 0 0 0 0 1 0 1 0 ...
##  $ behavAnySolution         : int  0 0 1 1 0 0 1 0 1 0 ...
##  $ behavAnySale             : int  0 0 1 0 0 0 1 0 1 1 ...
##  $ behavCart                : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ behavConversion          : int  0 0 0 0 0 0 0 0 0 0 ...
```

5. Plot a histogram of the newly added integer estimate of page views (pageviewInt).

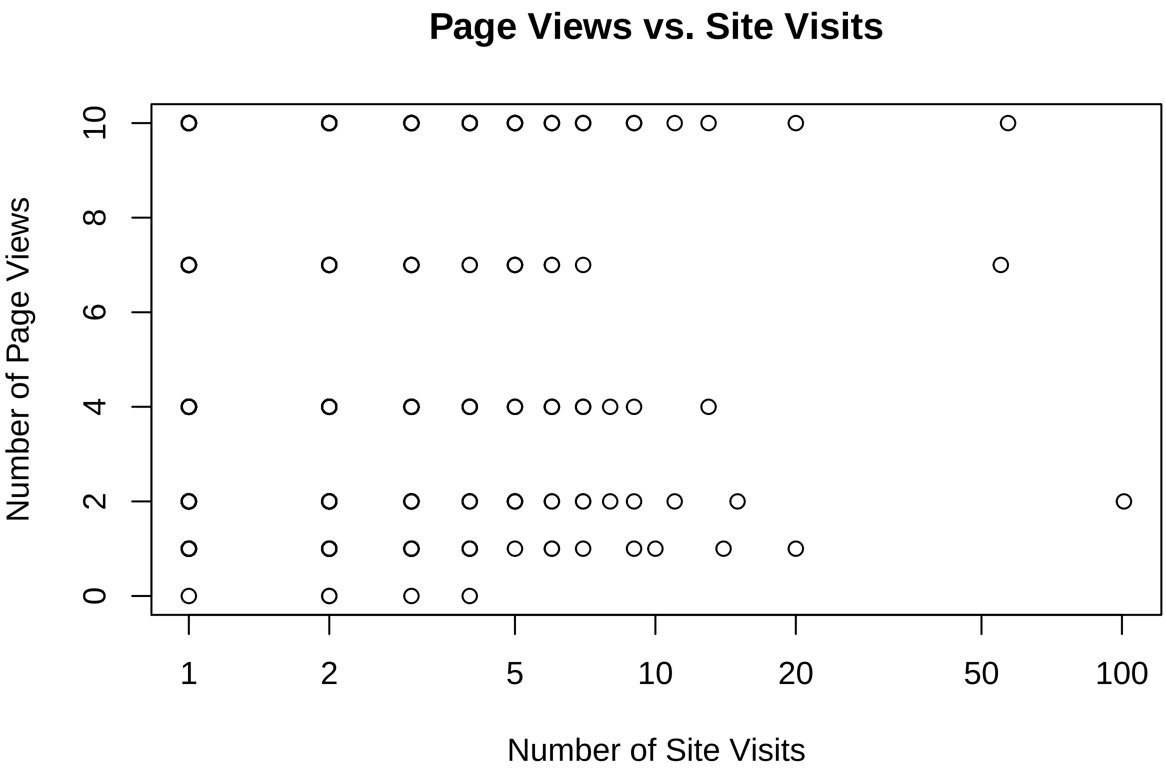
```
#Adding new integer variable, pageviewInt
pageViewInt <- rep(NA, length(ecomm.df$behavPageviews))
pageViewInt[ecomm.df$behavPageviews=="0"] <- 0
pageViewInt[ecomm.df$behavPageviews=="1"] <- 1
pageViewInt[ecomm.df$behavPageviews=="2 to 3"] <- 2
pageViewInt[ecomm.df$behavPageviews=="4 to 6"] <- 4
pageViewInt[ecomm.df$behavPageviews=="7 to 9"] <- 7
pageViewInt[ecomm.df$behavPageviews=="10+"] <- 10
ecomm.df$pageViewInt <- pageViewInt

hist(ecomm.df$pageViewInt, main = "Histogram of Page Views",
     xlab = "Page Views")
```



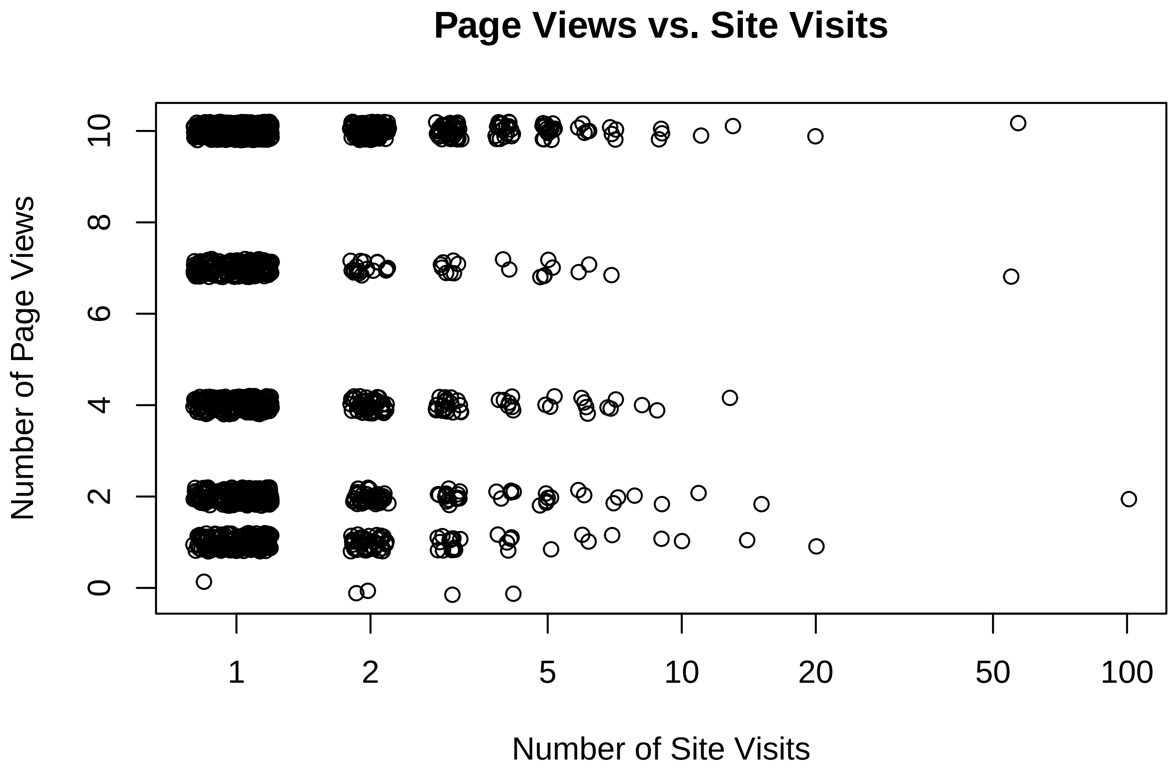
6. For a first exploration, make a scatterplot for the integer estimate of page views vs. the number of site visits. It's better to have the number of visits on a log scale.

```
plot(ecomm.df$pageViewInt ~ ecomm.df$behavNumVisits,
     main = "Page Views vs. Site Visits",
     ylab = "Number of Page Views", xlab = "Number of Site Visits",
     log = "x")
```



7. There are only a few values of X and Y in the previous plot. Adjust the plot to visualize more clearly the frequencies occurring at each point on the plot.

```
plot(jitter(ecomm.df$pageViewInt) ~ jitter(ecomm.df$behavNumVisits),
     main = "Page Views vs. Site Visits",
     ylab = "Number of Page Views", xlab = "Number of Site Visits", log = "x")
```



8. What is the Pearson's r correlation coefficient between number of visits and the integer estimate of page views?

```
cor(ecomm.df$pageViewInt, ecomm.df$behavNumVisits)
```

```
## [1] 0.005626593
```

The correlation between number of visits and the integer estimate of page views is 0.005626593

What is the correlation if you use log of visits instead?

```
cor(ecomm.df$pageViewInt, log(ecomm.df$behavNumVisits))
```

```
## [1] 0.04003549
```

The correlation between log of number of visits and the integer estimate of page views is 0.04003549

9. Is the correlation from the previous exercise statistically significant?

```
cor.test(ecomm.df$pageViewInt, ecomm.df$behavNumVisits)
```

```
##
##  Pearson's product-moment correlation
```

```
##
## data:  ecomm.df$pageViewInt and ecomm.df$behavNumVisits
## t = 0.22443, df = 1591, p-value = 0.8224
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.04349882  0.05472487
## sample estimates:
##         cor
## 0.005626593
```

As the p-value = 0.8224, it is more than the significance level (0.05). This means that we fail to reject the null hypothesis. Therefore, this indicates that the correlation is not statistically significant.

```
cor.test(ecomm.df$pageViewInt, log(ecomm.df$behavNumVisits))
```

```
##
##  Pearson’s product-moment correlation
##
## data:  ecomm.df$pageViewInt and log(ecomm.df$behavNumVisits)
## t = 1.5982, df = 1591, p-value = 0.1102
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.009095793  0.088973938
## sample estimates:
##        cor
## 0.04003549
```

As the p-value = 0.1102, it is more than the significance level (0.05). This means that we fail to reject the null hypothesis. Therefore, this indicates that the correlation is not statistically significant.

For the remaining exercises, we use the Salaries data from the car package.

```
library(car)
```

```
## Loading required package: carData
```

10. How do you load the Salaries data from the car package? (Hint: review the data() function)

```
data("Salaries")
```

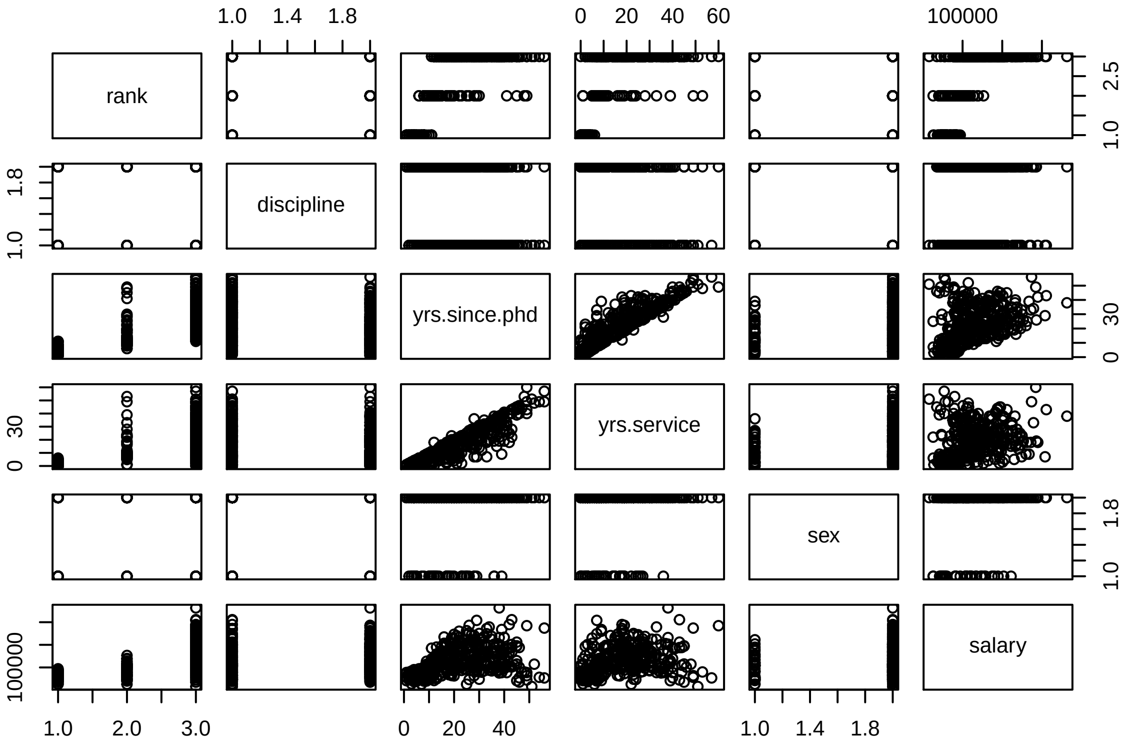
Within R itself, how can you find out more detail about the Salaries data set?

```
?Salaries
```

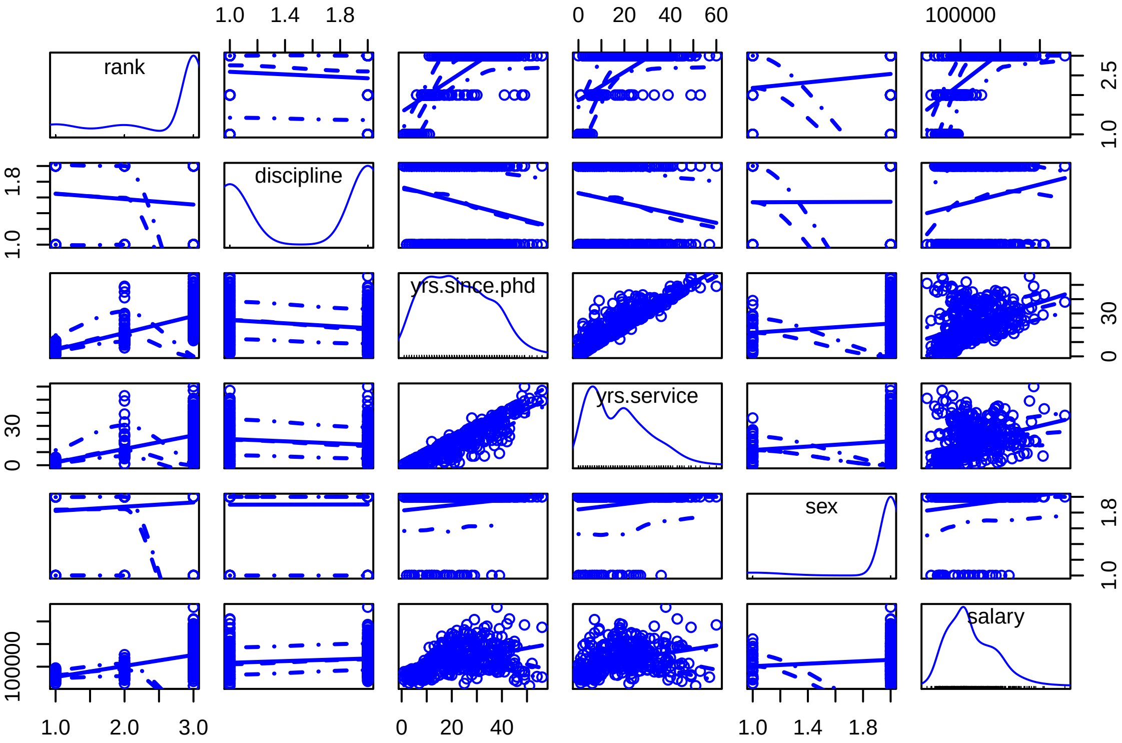
```
## starting httpd help server ... done
```

11. Using the Salaries data, create scatterplot matrix plots using two different plotting functions. Which do you prefer and why?

```
pairs(formula = ~ rank + discipline + yrs.since.phd + yrs.service + sex + salary, data = Salaries)
```



```
scatterplotMatrix(formula = ~ rank + discipline + yrs.since.phd + yrs.service + sex + salary, data = Sal
```



Scatterplot Matrix is better because it details both the correlation and distribution of variables.

12. Which are the numeric variables in the Salaries data set?

```
str(Salaries)
```

```
## 'data.frame':    397 obs. of  6 variables:
##  $ rank         : Factor w/ 3 levels "AsstProf","AssocProf",..: 3 3 1 3 3 2 3 3 3 3 ...
##  $ discipline   : Factor w/ 2 levels "A","B": 2 2 2 2 2 2 2 2 2 2 ...
##  $ yrs.since.phd: int  19 20 4 45 40 6 30 45 21 18 ...
##  $ yrs.service  : int  18 16 3 39 41 6 23 45 20 18 ...
##  $ sex          : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 1 ...
##  $ salary       : int  139750 173200 79750 115000 141500 97000 175000 147765 119250 129000 ...
```

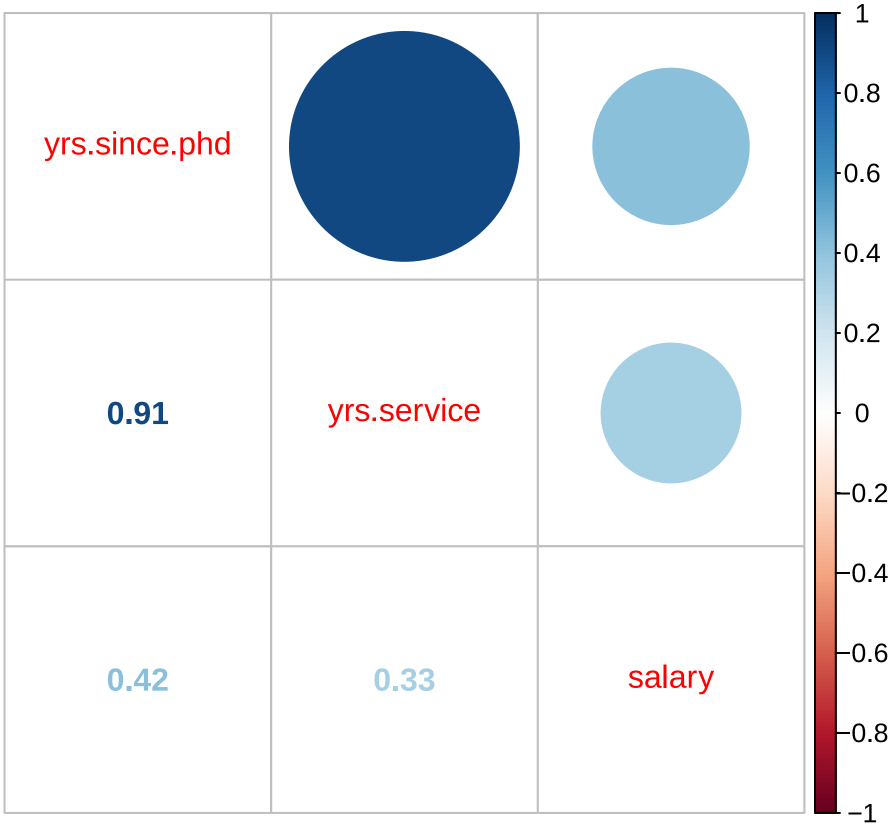
Numeric variables are yrs.since.phd, yrs.service, & salary.

Create a correlation plot for them, with correlation coefficients in one area of the plot. Which two variables are most closely related?

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
corrplot.mixed(cor(Salaries[,c(3,4,6)]))
```



The two variables that are the most closely related are yrs.since.phd & yrs.service