A1_FGS

Frederick Strathmann

5/22/2017

Code chunk #1: Setup, data import and inspection of the data

```
## 1.A. load needed libraries
library(psych)
library(rmarkdown)
library(scatterplot3d)

## set directories and constants
dir <- '~/R/IS_6482/A1'
fileName <- 'bank_full.csv'

## import data as characters first
dat <- read.csv(gettextf('%s/%s', dir, fileName), colClasses = 'character', stringsAsFactors = FALSE)

## 1.B. Overall structure
str(dat)
```

```
## 'data.frame':    2451 obs. of  13 variables:
##  $ deposit  : chr  "yes" "yes" "yes" "yes" ...
##  $ age      : chr  "42" "33" "36" "56" ...
##  $ job      : chr  "admin" "services" "management" "technician" ...
##  $ education: chr  "secondary" "secondary" "tertiary" "secondary" ...
##  $ default  : chr  "no" "no" "no" "no" ...
##  $ housing  : chr  "yes" "yes" "yes" "yes" ...
##  $ loan     : chr  "yes" "no" "no" "no" ...
##  $ contact  : chr  "telephone" "telephone" "telephone" "unknown" ...
##  $ month    : chr  "oct" "oct" "oct" "oct" ...
##  $ duration : chr  "519" "144" "140" "518" ...
##  $ campaign : chr  "1" "1" "1" "1" ...
##  $ pdays    : chr  "166" "91" "143" "147" ...
##  $ poutcome : chr  "other" "failure" "failure" "success" ...
```

```
## summary showing the mean and the five-number statistics indicating the spread of each column's values
summary(dat)
```

```
##    deposit             age                job           
##  Length:2451        Length:2451        Length:2451       
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##   education            default             housing        
##  Length:2451        Length:2451        Length:2451       
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##      loan              contact              month          
##  Length:2451        Length:2451        Length:2451       
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##    duration            campaign             pdays          
##  Length:2451        Length:2451        Length:2451       
```

```
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##    poutcome
##  Length:2451
##  Class :character
##  Mode  :character
```

```
## 1.C. convert strings to factors
colStrings <- colnames(dat)[sapply(dat[1,], function(x){grepl('[a-zA-z]', x)})]
dat[,colStrings] <- lapply(dat[,colStrings], as.factor)

## overall structure - after factoring
str(dat)
```

```
## 'data.frame':    2451 obs. of  13 variables:
##  $ deposit  : Factor w/ 2 levels "no","yes": 2 2 2 2 1 1 1 1 1 1 ...
##  $ age      : chr  "42" "33" "36" "56" ...
##  $ job      : Factor w/ 12 levels "admin","bluecollar",..: 1 8 5 10 2 11 5 10 5 5 ...
##  $ education: Factor w/ 4 levels "primary","secondary",..: 2 2 3 2 2 2 3 2 3 3 ...
##  $ default  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ housing  : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 1 ...
##  $ loan     : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 2 1 ...
##  $ contact  : Factor w/ 3 levels "cellular","telephone",..: 2 2 2 3 2 2 2 1 1 1 ...
##  $ month    : Factor w/ 12 levels "apr","aug","dec",..: 11 11 11 11 11 10 10 10 10 10 ...
##  $ duration : chr  "519" "144" "140" "518" ...
##  $ campaign : chr  "1" "1" "1" "1" ...
##  $ pdays    : chr  "166" "91" "143" "147" ...
##  $ poutcome : Factor w/ 4 levels "failure","other",..: 2 1 1 3 2 1 1 1 1 2 ...
```

```
## summary showing the mean and the five-number statistics indicating the spread of each column's values
summary(dat)
```

```
##  deposit        age                 job          education
##  no :1863   Length:2451        management:532   primary  : 293
##  yes: 588   Class :character   bluecollar:488   secondary:1256
##             Mode  :character   technician:375   tertiary : 808
##                                admin     :318   unknown  :  94
##                                services  :231
##                                retired   :145
##                                (Other)   :362
##  default    housing     loan          contact         month
##  no :2434   no : 937   no :2123   cellular :2241   may    :763
##  yes:  17   yes:1514   yes: 328   telephone: 188   apr    :341
##                                   unknown  :  22   nov    :335
##                                                    feb    :261
##                                                    aug    :157
##                                                    jan    :136
##                                                    (Other):458
##    duration           campaign           pdays             poutcome
##  Length:2451        Length:2451        Length:2451        failure:1457
##  Class :character   Class :character   Class :character   other  : 547
##  Mode  :character   Mode  :character   Mode  :character   success: 446
##                                                           unknown:   1
##
##
##
```

```r
## 1.D. Retrieve, save and show number of rows and columns
rows <- nrow(dat)
cols <- ncol(dat)
cat(gettextf('Number of rows in data: %.0f', rows))
```

```
## Number of rows in data: 2451
```

```r
cat(gettextf('Number of columns in data: %.0f', cols))
```

```
## Number of columns in data: 13
```

```r
## 1.E. Show head and tail
## first 10 instances
head(dat, n=10)
```

```
##    deposit age        job education default housing loan   contact month
## 1      yes  42      admin secondary      no     yes  yes telephone   oct
## 2      yes  33   services secondary      no     yes   no telephone   oct
## 3      yes  36 management  tertiary      no     yes   no telephone   oct
## 4      yes  56 technician secondary      no     yes   no   unknown   oct
## 5       no  44 bluecollar secondary      no     yes   no telephone   oct
## 6       no  33 unemployed secondary      no     yes   no telephone   nov
## 7       no  30 management  tertiary      no     yes   no telephone   nov
## 8       no  51 technician secondary      no     yes   no  cellular   nov
## 9       no  44 management  tertiary      no     yes  yes  cellular   nov
## 10      no  38 management  tertiary      no      no   no  cellular   nov
##    duration campaign pdays poutcome
## 1       519        1   166    other
## 2       144        1    91  failure
## 3       140        1   143  failure
## 4       518        1   147  success
## 5       119        1    89    other
## 6       175        1   174  failure
## 7        86        1   174  failure
## 8        79        1   129  failure
## 9        58        1   188  failure
## 10      146        1   104    other
```

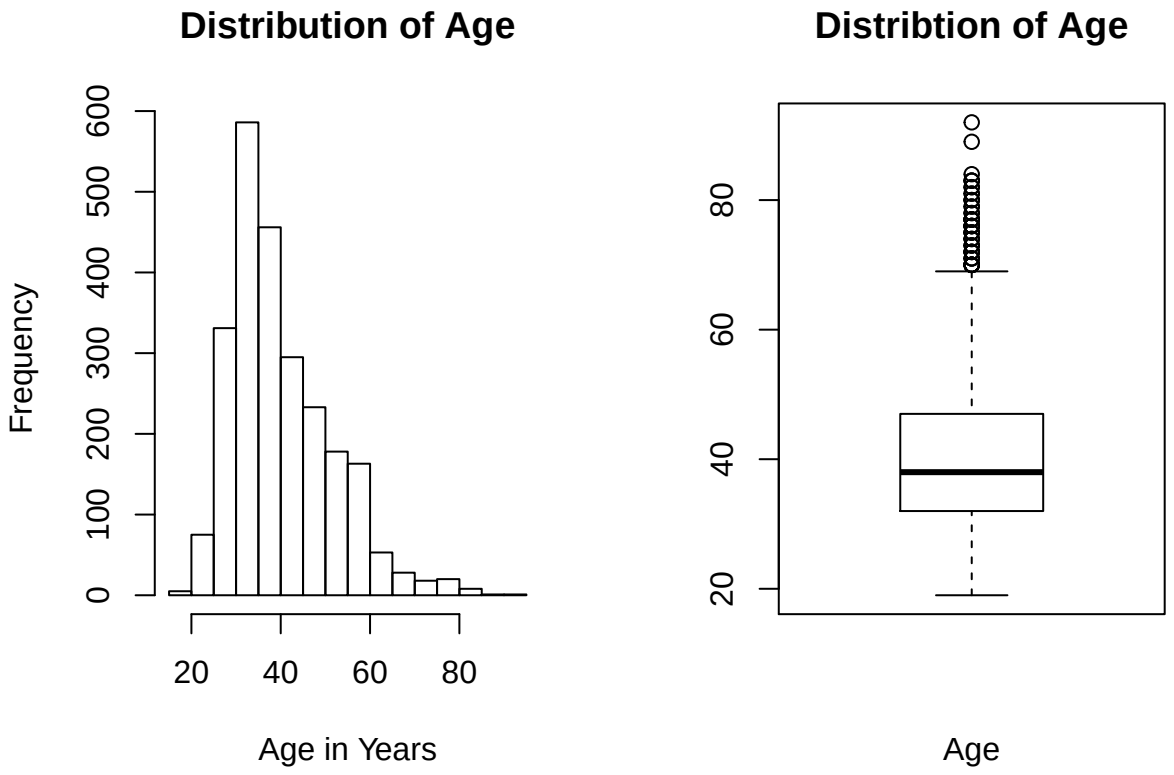
```r
## last 10 instances
tail(dat, n=10)
```

```
##      deposit age        job education default housing loan   contact month
## 2442     yes  62    retired  tertiary      no     yes   no  cellular   nov
## 2443      no  19    student   primary      no      no   no telephone   nov
## 2444      no  30 technician  tertiary      no      no   no  cellular   nov
## 2445      no  36      admin  tertiary      no      no   no  cellular   nov
## 2446     yes  36      admin secondary      no     yes   no  cellular   nov
## 2447     yes  34 bluecollar secondary      no     yes   no  cellular   nov
## 2448      no  34 technician secondary      no      no   no  cellular   nov
## 2449      no  66    retired secondary      no      no   no  cellular   nov
## 2450     yes  68    retired secondary      no      no   no  cellular   nov
## 2451     yes  72    retired secondary      no      no   no  cellular   nov
##      duration campaign pdays poutcome
## 2442      404        1    57  success
## 2443       98        2   110    other
## 2444      134        1    92  success
```

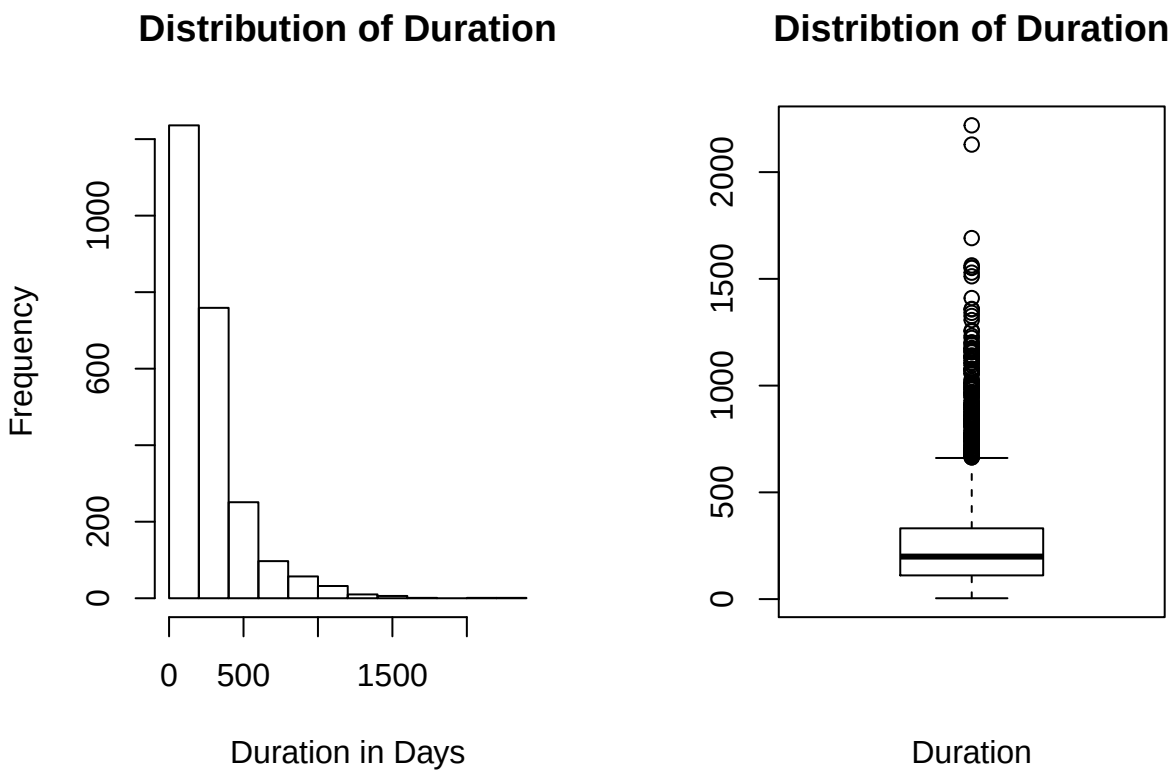
```
## 2445      118        4   104  failure
## 2446      482        1   374  success
## 2447      413        1    92  success
## 2448      319        1   100  failure
## 2449      414        2    27  failure
## 2450      212        1   187  success
## 2451     1127        5   184  success
```

Code Chunk #2: Exploration of variables of numeric data type

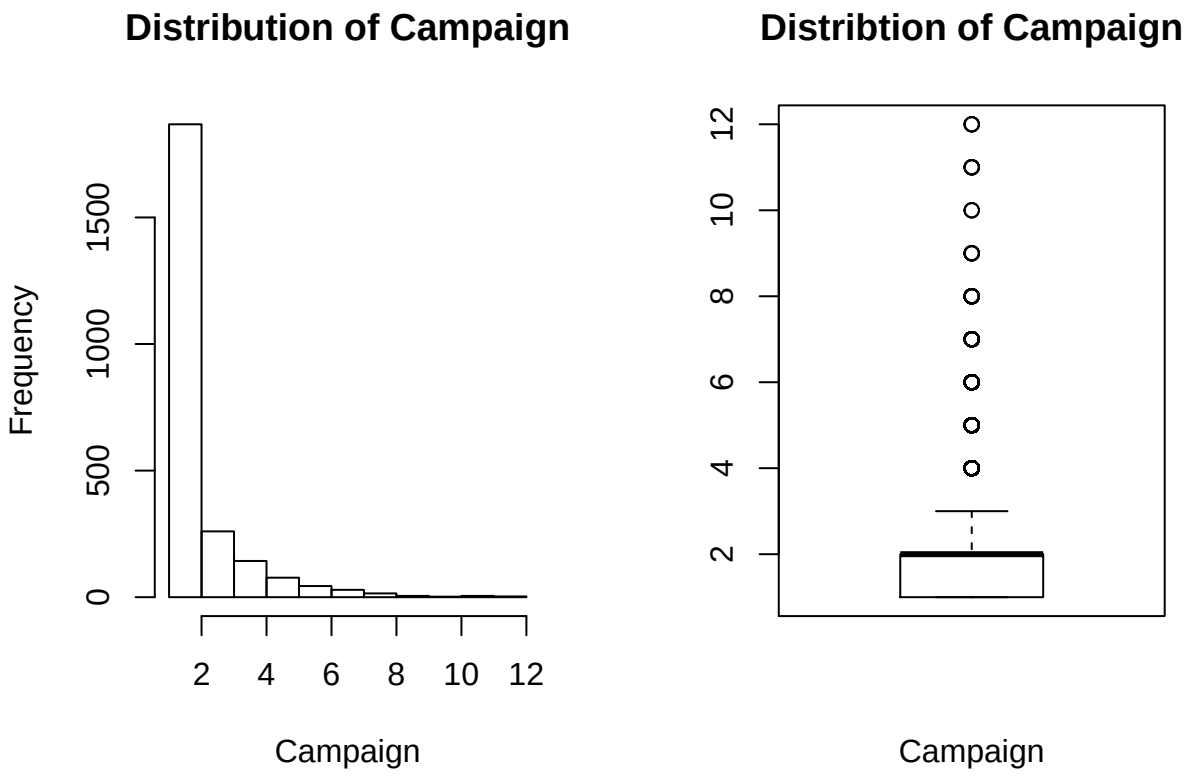
```
## 2.A. Histogram and Boxplot for each numeric variable
## generate numeric data first
colNumeric <- colnames(dat)[sapply(dat[1,], function(x){!grepl('[a-zA-z]', x)})]
dat[,colNumeric] <- lapply(dat[,colNumeric], as.numeric)

## histogram and boxplot for Age
par(mfrow = c(1,2))
with(dat, hist(age, main = 'Distribution of Age', xlab = 'Age in Years', ylab = 'Frequency'))
with(dat, boxplot(age, main = 'Distribtion of Age', xlab = 'Age'))
```



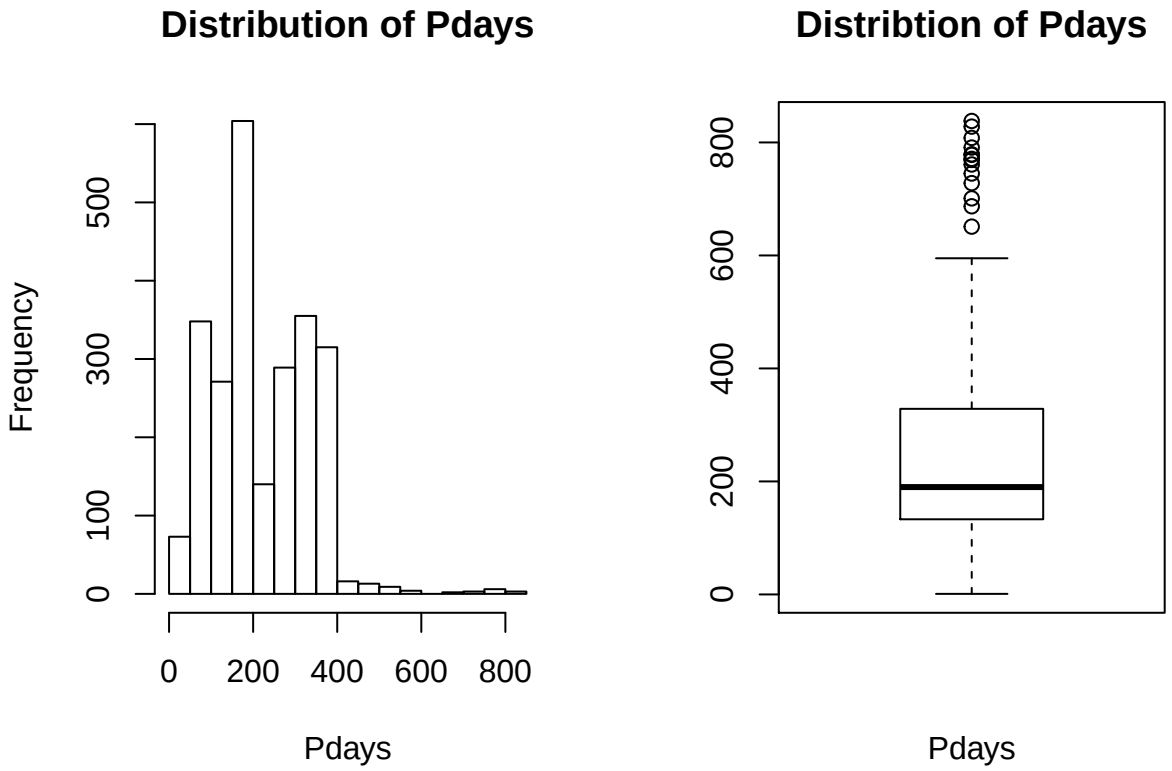

```
## histogram and boxplot for Duration
par(mfrow = c(1,2))
with(dat, hist(duration, main = 'Distribution of Duration', xlab = 'Duration in Days', ylab = 'Frequency
with(dat, boxplot(duration, main = 'Distribtion of Duration', xlab = 'Duration'))
```




```
## histogram and boxplot for Campaign
par(mfrow = c(1,2))
with(dat, hist(campaign, main = 'Distribution of Campaign', xlab = 'Campaign', ylab = 'Frequency'))
with(dat, boxplot(campaign, main = 'Distribtion of Campaign', xlab = 'Campaign'))
```



```r
## histogram and boxplot for Pdays
par(mfrow = c(1,2))
with(dat, hist(pdays, main = 'Distribution of Pdays', xlab = 'Pdays', ylab = 'Frequency'))
with(dat, boxplot(pdays, main = 'Distribtion of Pdays', xlab = 'Pdays'))
```




```r
## 2.B.i. Non normalized summary statistics (mean, variance, sd, quantiles, deciles) for duration, campa
## put evertying into a dataframe for ease

## initial stats - quantiles
quantDur <- as.data.frame(quantile(dat$duration), optional = TRUE)
quantCamp <- as.data.frame(quantile(dat$campaign), optional = TRUE)
quantpdays <- as.data.frame(quantile(dat$pdays), optional = TRUE)

## initial stats - deciles
decDur <- as.data.frame(quantile(dat$duration, seq(from = 0, to = 1, by = 0.10)), optional = TRUE)
decCamp <- as.data.frame(quantile(dat$campaign, seq(from = 0, to = 1, by = 0.10)), optional = TRUE)
decpdays <- as.data.frame(quantile(dat$pdays, seq(from = 0, to = 1, by = 0.10)), optional = TRUE)

## summary table - non normalized!!
sumStats <- with(dat, data.frame('Parameter' = c('Duration', 'Campaign', 'Pdays'),
                        'Mean' = c(mean(duration), mean(campaign), mean(pdays)),
                        'Variance' = c(var(duration), var(campaign), var(pdays)),
                        'Standard.Deviation' = c(sd(duration), sd(campaign), sd(pdays))
))

sumStatsQuant <- with(dat, data.frame('Quartile' = rownames(quantDur),
                                      'Raw.duration' = quantDur[,1],
                                      'Raw.campaign' = quantCamp[,1],
                                      'Raw.pdays' = quantpdays[,1]
```

```
))

sumStatsDec <- with(dat, data.frame('Decile' = rownames(decDur),
                                    'Raw.duration' = decDur[,1],
                                    'Raw.campaign' = decCamp[,1],
                                    'Raw.pdays' = decpdays[,1]
))

cat('Summary stats for NON NORMALIZED data')
```

```
## Summary stats for NON NORMALIZED data
```

```
sumStats
```

```
##   Parameter       Mean     Variance Standard.Deviation
## 1  Duration 265.539780 56609.030560         237.926523
## 2  Campaign   2.040392     2.360409           1.536362
## 3     Pdays 223.604243 12966.003313         113.868360
```

```
cat('Non-Normalized Quantiles')
```

```
## Non-Normalized Quantiles
```

```
sumStatsQuant
```

```
##   Quartile Raw.duration Raw.campaign Raw.pdays
## 1       0%          4.0            1       1.0
## 2      25%        111.0            1     133.0
## 3      50%        199.0            2     190.0
## 4      75%        331.5            2     328.5
## 5     100%       2219.0           12     838.0
```

```
cat('Non-Normalized Deciles')
```

```
## Non-Normalized Deciles
```

```
sumStatsDec
```

```
##    Decile Raw.duration Raw.campaign Raw.pdays
## 1      0%            4            1         1
## 2     10%           63            1        92
## 3     20%           98            1       110
## 4     30%          128            1       158
## 5     40%          160            1       181
## 6     50%          199            2       190
## 7     60%          243            2       257
## 8     70%          293            2       299
## 9     80%          383            3       343
## 10    90%          549            4       361
## 11   100%         2219           12       838
```

```
## 2.B.ii. Normalized summary statistics (mean, variance, sd, quantiles, deciles) for duration, campaign
## NORMALIZE duration, campaign and pdays
dat$durationNORM <- with(dat, round((duration - min(duration))/(max(duration) - min(duration)), digits =
dat$campaignNORM <- with(dat, round((campaign - min(campaign))/(max(campaign) - min(campaign)), digits =
dat$pdaysNORM <- with(dat, round((pdays - min(pdays))/(max(pdays) - min(pdays)), digits = 1))

## normalized stats - quantiles
quantDurN <- as.data.frame(quantile(dat$durationNORM), optional = TRUE)
```

```r
quantCampN <- as.data.frame(quantile(dat$campaignNORM),optional = TRUE)
quantpdaysN <- as.data.frame(quantile(dat$pdaysNORM), optional = TRUE)

## normalized stats - deciles
decDurN <- as.data.frame(quantile(dat$durationNORM, seq(from = 0, to = 1, by = 0.10)), optional = TRUE)
decCampN <- as.data.frame(quantile(dat$campaignNORM, seq(from = 0, to = 1, by = 0.10)), optional = TRUE)
decpdaysN <- as.data.frame(quantile(dat$pdaysNORM, seq(from = 0, to = 1, by = 0.10)), optional = TRUE)

## summary table - Normalized!!
sumStatsNORM <- with(dat, data.frame('Parameter' = c('Duration', 'Campaign', 'Pdays'),
                        'Mean' = c(mean(durationNORM), mean(campaignNORM), mean(pdaysNORM)),
                        'Variance' = c(var(durationNORM), var(campaignNORM), var(pdaysNORM)),
                        'Standard.Deviation' = c(sd(durationNORM), sd(campaignNORM), sd(pdaysNORM))
))

sumStatsNormQuant <- with(dat, data.frame('Quartile' = rownames(quantDurN),
                                          'NORM.duration' = quantDurN[,1],
                                          'NORM.campaign' = quantCampN[,1],
                                          'NORM.pdays' = quantpdaysN[,1]
))

sumStatsNormDec <- with(dat, data.frame('Decile' = rownames(decDurN),
                                        'NORM.duration' = decDurN[,1],
                                        'NORM.campaign' = decCampN[,1],
                                        'NORM.pdays' = decpdaysN[,1]
))

cat('Summary stats for NORMALIZED data')
```

```
## Summary stats for NORMALIZED data
```

```r
sumStatsNORM
```

```
##   Parameter      Mean   Variance Standard.Deviation
## 1  Duration 0.1161159 0.01288711          0.1135214
## 2  Campaign 0.1016320 0.02091162          0.1446085
## 3     Pdays 0.2577723 0.01862834          0.1364857
```

```r
cat('Normalized Quantiles')
```

```
## Normalized Quantiles
```

```r
sumStatsNormQuant
```

```
##   Quartile NORM.duration NORM.campaign NORM.pdays
## 1       0%           0.0           0.0        0.0
## 2      25%           0.0           0.0        0.2
## 3      50%           0.1           0.1        0.2
## 4      75%           0.1           0.1        0.4
## 5     100%           1.0           1.0        1.0
```

```r
cat('Normalized Deciles')
```

```
## Normalized Deciles
```

```r
sumStatsNormDec
```

```
##    Decile NORM.duration NORM.campaign NORM.pdays
```

```
## 1      0%           0.0           0.0        0.0
## 2     10%           0.0           0.0        0.1
## 3     20%           0.0           0.0        0.1
## 4     30%           0.1           0.0        0.2
## 5     40%           0.1           0.0        0.2
## 6     50%           0.1           0.1        0.2
## 7     60%           0.1           0.1        0.3
## 8     70%           0.1           0.1        0.4
## 9     80%           0.2           0.2        0.4
## 10    90%           0.2           0.3        0.4
## 11   100%           1.0           1.0        1.0
```

Code Chunk #3: Exploration of variables of factor data type

```
## 3.A Count value and % value for job, eduation, contact, poutcome

## count tables
jobTable <- with(dat, table(job))
eduTable <- with(dat, table(education))
contTable <- with(dat, table(contact))
poutTable <- with(dat, table(poutcome))

cat('Count Values')
```

```
## Count Values
```

```
cat('Job Table')
```

```
## Job Table
```

```
jobTable
```

```
## job
##        admin   bluecollar entrepreneur    housemaid   management
##          318          488           75           37          532
##      retired selfemployed     services      student   technician
##          145           92          231           84          375
##   unemployed      unknown
##           68            6
```

```
cat('Education Table')
```

```
## Education Table
```

```
eduTable
```

```
## education
##   primary secondary  tertiary   unknown
##       293      1256       808        94
```

```
cat('Contact Table')
```

```
## Contact Table
```

```
contTable
```

```
## contact
##  cellular telephone   unknown
##      2241       188        22
```

```r
cat('Outcome Table')
```

```
## Outcome Table
```

```r
poutTable
```

```
## poutcome
## failure   other success unknown
##    1457     547     446       1
```

```r
## proportion tables
jobTableP <- with(dat, round(prop.table(jobTable)*100, digits = 1))
eduTableP <- with(dat, round(prop.table(eduTable)*100), digits = 1)
contTableP <- with(dat, round(prop.table(contTable)*100), digits = 1)
poutTableP <- with(dat, round(prop.table(poutTable)*100), digits = 1)

cat('Proportion Tables as Percent')
```

```
## Proportion Tables as Percent
```

```r
cat('Job Table')
```

```
## Job Table
```

```r
jobTableP
```

```
## job
##        admin   bluecollar entrepreneur    housemaid   management
##         13.0         19.9          3.1          1.5         21.7
##      retired selfemployed     services      student   technician
##          5.9          3.8          9.4          3.4         15.3
##   unemployed      unknown
##          2.8          0.2
```

```r
cat('Education Table')
```

```
## Education Table
```

```r
eduTableP
```

```
## education
##   primary secondary  tertiary   unknown
##        12        51        33         4
```

```r
cat('Contact Table')
```

```
## Contact Table
```

```r
contTableP
```

```
## contact
##  cellular telephone   unknown
##        91         8         1
```

```r
cat('Outcome Table')
```

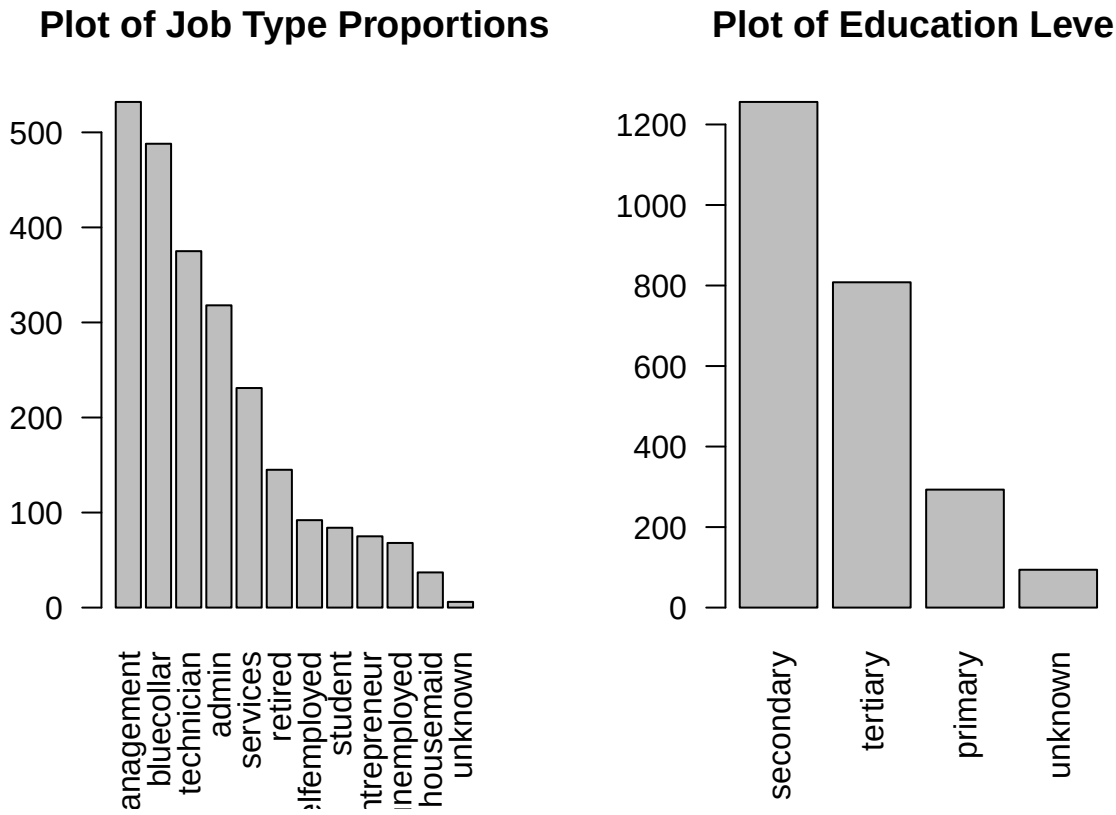
```
## Outcome Table
```

```r
poutTableP
```

```
## poutcome
## failure   other success unknown
```

```
##      59      22      18       0
```

```
## 3.B bar plot for Jobs and Education
par(mfrow = c(1,2))
barplot(sort(jobTable, decreasing=TRUE), main = 'Plot of Job Type Proportions', las=2)
barplot(sort(eduTable, decreasing=TRUE), main = 'Plot of Education Level', las=2)
```




```
## 3.C Retrieve and save number of levels for contact and poutcome
contL <- nlevels(dat$contact)
cat('Number of levels for contact: ', contL)
```

```
## Number of levels for contact:  3
```

```
poutL <- nlevels(dat$poutcome)
cat('Number of levels for poutcome: ', poutL)
```

```
## Number of levels for poutcome:  4
```

Code Chunk #4: Demonstration of relationships amongst multiple variables

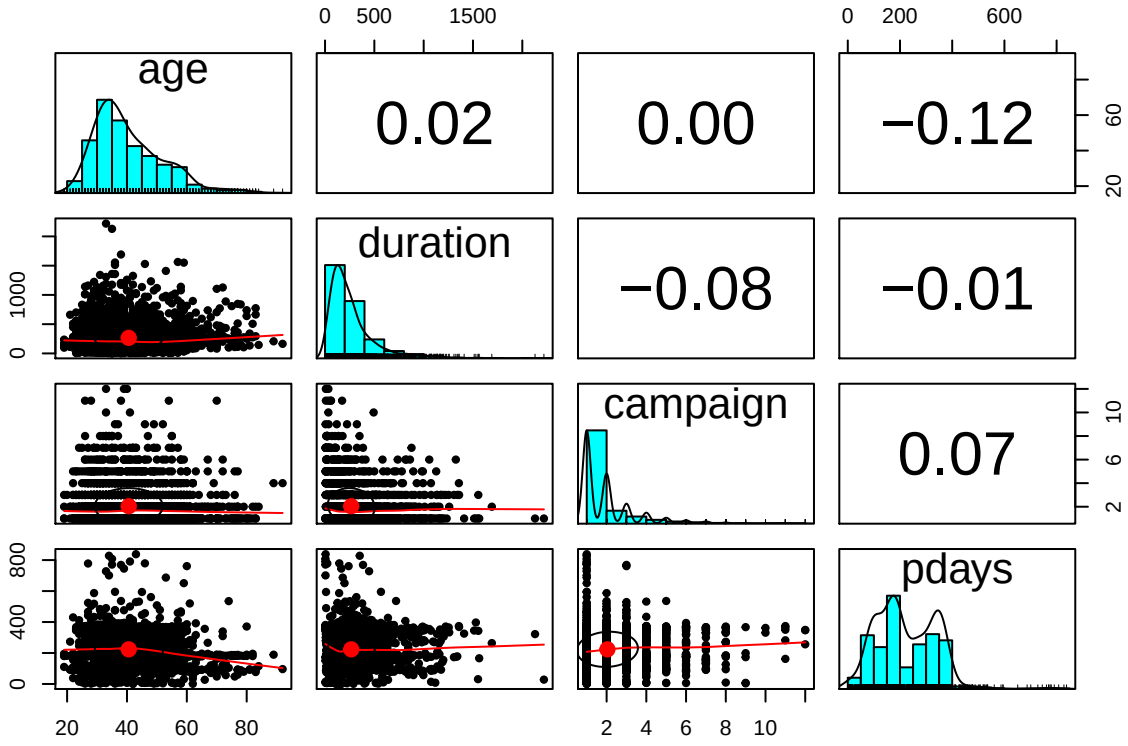
```
## 4.A. Correlations and pairwise graphs for all numeric variables
cat('Correlation coefficients')
```

```
## Correlation coefficients
```

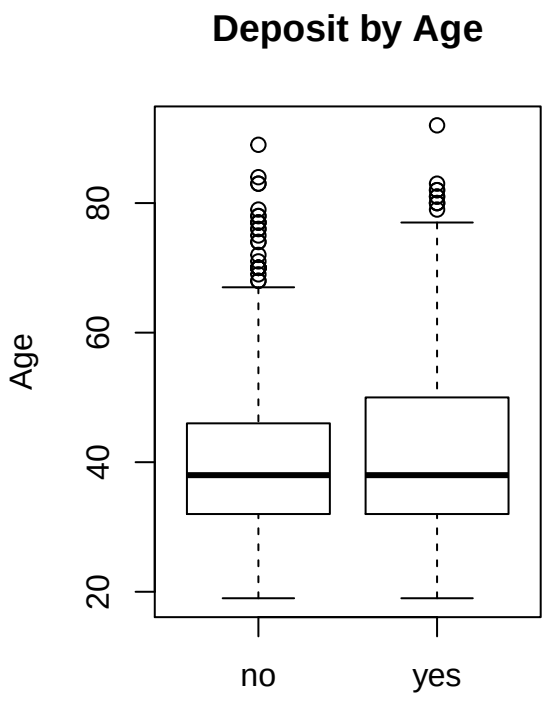
```
cor(dat[,colNumeric])
```

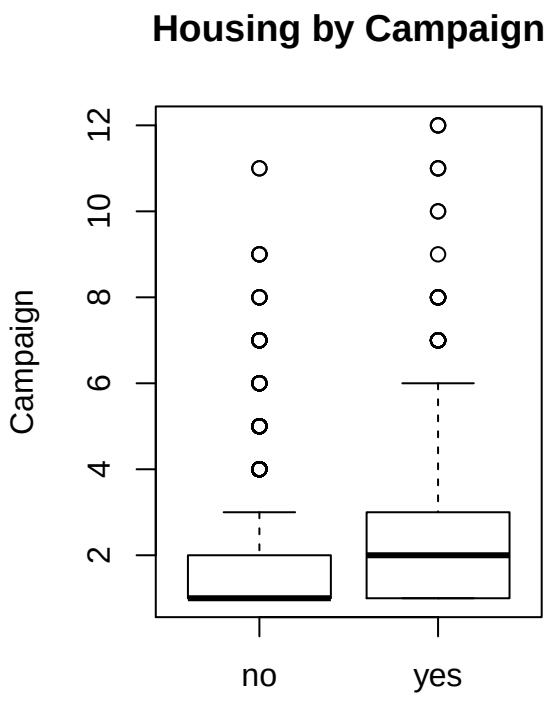
```
##                  age    duration     campaign       pdays
## age       1.000000000  0.02356400  0.002656224 -0.11759320
## duration  0.023564002  1.00000000 -0.079325956 -0.01278811
## campaign  0.002656224 -0.07932596  1.000000000  0.07311810
## pdays    -0.117593201 -0.01278811  0.073118097  1.00000000
```

```
## pairwise graphs
pairs.panels(dat[,colNumeric])
```




```
## 4.B.i Boxplot: age by deposit and campaign by housing
par(mfrow = c(1,2))
with(dat, boxplot(age~deposit, main = 'Deposit by Age', xlab = 'Deposit', ylab = 'Age'))
with(dat, boxplot(campaign~housing, main = 'Housing by Campaign', xlab = 'Housing', ylab = 'Campaign'))
```





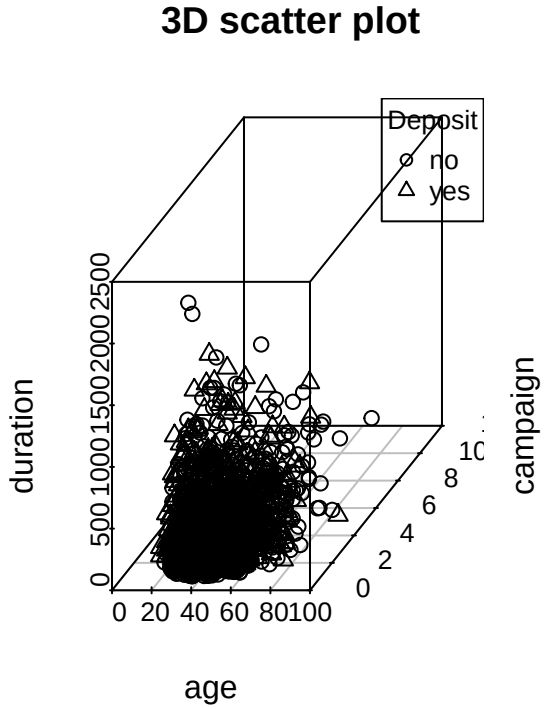

```
## 4.B.ii aggregated summary
aggregate(age~deposit, data = dat, summary)
```

```
##   deposit age.Min. age.1st Qu. age.Median age.Mean age.3rd Qu. age.Max.
## 1      no 19.00000    32.00000   38.00000 40.07837    46.00000 89.00000
## 2     yes 19.00000    32.00000   38.00000 42.24830    50.00000 92.00000
```

```
aggregate(campaign~housing, data = dat, summary)
```

```
##   housing campaign.Min. campaign.1st Qu. campaign.Median campaign.Mean
## 1      no      1.000000         1.000000        1.000000      1.925293
## 2     yes      1.000000         1.000000        2.000000      2.111625
##   campaign.3rd Qu. campaign.Max.
## 1         2.000000     11.000000
## 2         3.000000     12.000000
```

```
## 4.C 3d scatter plot
with(dat, scatterplot3d(age, campaign, duration, pch = as.numeric(deposit), main = "3D scatter plot", su
with(dat, legend('topright', legend = levels(deposit), cex = 0.8, pch = 1:2, title = 'Deposit'))
```



Age, Campaign, Duration