MVA_Assignment_4

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

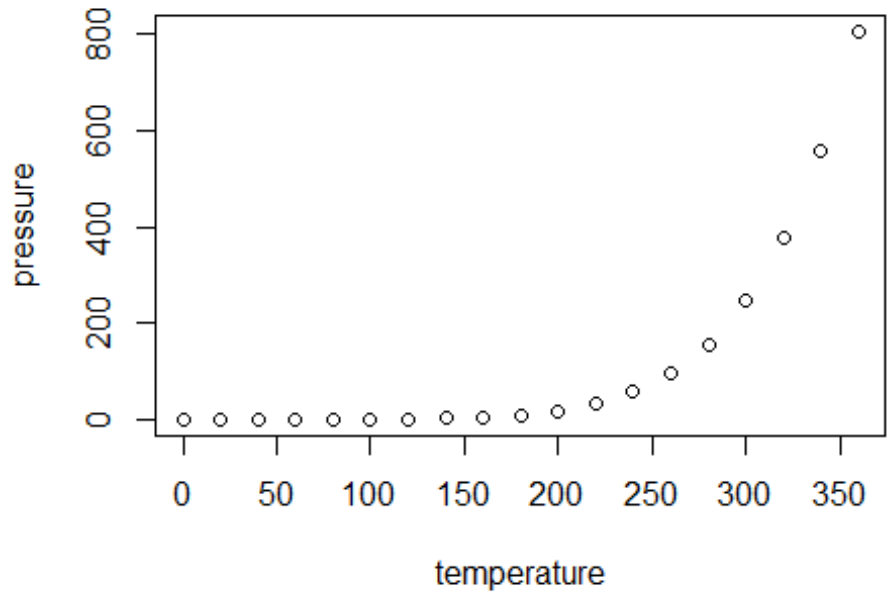
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

```
library(data.table)
library(tidyverse)  # data manipulation

## -- Attaching packages --------------------------------------- tidyvers
e 1.3.0 --

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.4
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ------------------------------------------ tidyverse_conf
licts() --
## x dplyr::between()   masks data.table::between()
## x dplyr::filter()    masks stats::filter()
## x dplyr::first()     masks data.table::first()
## x dplyr::lag()       masks stats::lag()
## x dplyr::last()      masks data.table::last()
## x purrr::transpose() masks data.table::transpose()

library(data.table) # fast file reading
library(gridExtra)  # arranging ggplot in grid

##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine

library(rmarkdown)
library(tinytex)
library(latexpdf)
library(latex2exp)

bank <- read.csv("C:/Users/Shamali/Desktop/Rutgers Spring/multivariat/project
/bank-marketing-dataset/bank.csv")

bank1<- bank[ ,c(1,5,6,10,12,13,14,15,17)]

bank1_pca <- prcomp(bank1[,-1],scale=FALSE)

bank1_pca

## Standard deviations (1, .., p=8):
## [1] 3225.4233974  347.0547106  108.7037873    8.4012846    2.6798358
## [6]    1.9739796    0.4320433    0.1213404
##
## Rotation (n x k) = (8 x 8):
##                    PC1           PC2           PC3           PC4
PC5
## default  -2.301152e-06 -2.901819e-06  3.977508e-05  0.0002243681  1.103771
e-03
## balance   9.999968e-01 -2.437063e-03  6.111252e-04 -0.0000315371  1.087512
e-05
## day       2.731106e-05 -4.495770e-04  6.082131e-03  0.9988831188 -4.640695
e-02
## duration  2.442855e-03  9.999500e-01 -9.652725e-03  0.0005246749  3.339633
e-04
## campaign -1.173648e-05 -3.211287e-04  2.602661e-03  0.0464153238  9.987384
e-01
## pdays     5.875094e-04 -9.656965e-03 -9.998741e-01  0.0062555015  2.228881
e-03
## previous  2.189243e-05 -1.912152e-04 -1.067605e-02 -0.0058038794  9.171479
e-03
## deposit   1.257854e-05  6.470195e-04 -7.541632e-04 -0.0022337784 -1.665930
e-02
##                    PC6           PC7           PC8
## default   1.127990e-03  7.758188e-03  9.999686e-01
## balance   1.633325e-05  1.011121e-05  2.168037e-06
## day      -6.317666e-03 -1.333395e-03 -1.556720e-04
## duration -7.468342e-05  6.493507e-04 -2.148716e-06
## campaign  8.514529e-03 -1.693552e-02 -9.911438e-04
## pdays     1.067406e-02  4.750723e-04  2.015439e-05
## previous -9.996699e-01  2.067405e-02  9.588572e-04
## deposit  -2.079730e-02 -9.996115e-01  7.797799e-03
```

```r
summary(bank1_pca)
```

```
## Importance of components:
##                             PC1       PC2       PC3     PC4  PC5   PC6
PC7
## Standard deviation     3225.4234 347.05471 108.70379 8.40128 2.68 1.974 0.
432
## Proportion of Variance    0.9874   0.01143   0.00112 0.00001 0.00 0.000 0.
000
## Cumulative Proportion     0.9874   0.99887   0.99999 1.00000 1.00 1.000 1.
000
##                            PC8
## Standard deviation     0.1213
## Proportion of Variance 0.0000
## Cumulative Proportion  1.0000
```

```r
(eigen_bank <- bank1_pca$sdev^2)
```

```
## [1] 1.040336e+07 1.204470e+05 1.181651e+04 7.058158e+01 7.181520e+00
## [6] 3.896595e+00 1.866614e-01 1.472349e-02
```

```r
names(eigen_bank) <- paste("PC",1:8,sep="")
eigen_bank
```

```
##          PC1          PC2          PC3          PC4          PC5
PC6
## 1.040336e+07 1.204470e+05 1.181651e+04 7.058158e+01 7.181520e+00 3.896595e
+00
##          PC7          PC8
## 1.866614e-01 1.472349e-02
```

```r
sumlambdas <- sum(eigen_bank)
sumlambdas
```

```
## [1] 10535701
```

```r
propvar <- eigen_bank/sumlambdas
propvar
```

```
##          PC1          PC2          PC3          PC4          PC5
PC6
## 9.874384e-01 1.143227e-02 1.121569e-03 6.699277e-06 6.816366e-07 3.698468e
-07
##          PC7          PC8
## 1.771704e-08 1.397486e-09
```

```r
cumvar_bank <- cumsum(propvar)
cumvar_bank
```

```
##       PC1       PC2       PC3       PC4       PC5       PC6       PC7
PC8
```

```
## 0.9874384 0.9988707 0.9999922 0.9999989 0.9999996 1.0000000 1.0000000 1.00
00000

matlambdas <- rbind(eigen_bank,propvar,cumvar_bank)

rownames(matlambdas) <- c("Eigenvalues","Prop. variance","Cum. prop. variance
")

round(matlambdas,4)

##                              PC1          PC2        PC3     PC4    PC5
PC6
## Eigenvalues         1.040336e+07 120446.9722 11816.5134 70.5816 7.1815 3.8
966
## Prop. variance      9.874000e-01      0.0114     0.0011  0.0000 0.0000 0.0
000
## Cum. prop. variance 9.874000e-01      0.9989     1.0000  1.0000 1.0000 1.0
000
##                         PC7    PC8
## Eigenvalues         0.1867 0.0147
## Prop. variance      0.0000 0.0000
## Cum. prop. variance 1.0000 1.0000

summary(bank1_pca)

## Importance of components:
##                              PC1       PC2       PC3     PC4  PC5   PC6
PC7
## Standard deviation     3225.4234 347.05471 108.70379 8.40128 2.68 1.974 0.
432
## Proportion of Variance    0.9874   0.01143   0.00112 0.00001 0.00 0.000 0.
000
## Cumulative Proportion     0.9874   0.99887   0.99999 1.00000 1.00 1.000 1.
000
##                            PC8
## Standard deviation     0.1213
## Proportion of Variance 0.0000
## Cumulative Proportion  1.0000

bank1_pca$rotation

##                    PC1           PC2           PC3           PC4
PC5
## default  -2.301152e-06 -2.901819e-06  3.977508e-05  0.0002243681  1.103771
e-03
## balance   9.999968e-01 -2.437063e-03  6.111252e-04 -0.0000315371  1.087512
e-05
## day       2.731106e-05 -4.495770e-04  6.082131e-03  0.9988831188 -4.640695
e-02
## duration  2.442855e-03  9.999500e-01 -9.652725e-03  0.0005246749  3.339633
e-04
```

```
## campaign -1.173648e-05 -3.211287e-04  2.602661e-03  0.0464153238  9.987384
e-01
## pdays     5.875094e-04 -9.656965e-03 -9.998741e-01  0.0062555015  2.228881
e-03
## previous  2.189243e-05 -1.912152e-04 -1.067605e-02 -0.0058038794  9.171479
e-03
## deposit   1.257854e-05  6.470195e-04 -7.541632e-04 -0.0022337784 -1.665930
e-02
##                    PC6           PC7           PC8
## default   1.127990e-03  7.758188e-03  9.999686e-01
## balance   1.633325e-05  1.011121e-05  2.168037e-06
## day      -6.317666e-03 -1.333395e-03 -1.556720e-04
## duration -7.468342e-05  6.493507e-04 -2.148716e-06
## campaign  8.514529e-03 -1.693552e-02 -9.911438e-04
## pdays     1.067406e-02  4.750723e-04  2.015439e-05
## previous -9.996699e-01  2.067405e-02  9.588572e-04
## deposit  -2.079730e-02 -9.996115e-01  7.797799e-03
```

```
# Sample scores stored in bank_pca$x
head(bank1_pca$x)
```

```
##             PC1       PC2      PC3       PC4         PC5       PC6
PC7
## [1,]   816.0646  668.4989 46.29391 -10.71400 -0.91235277 0.2804991 -0.0850
8411
## [2,] -1480.8899 1099.0781 40.78714 -10.41854 -0.79540938 0.2112248  0.1676
5440
## [3,]  -256.0843 1018.0966 42.28868 -10.49809 -0.80813650 0.2370584  0.1293
9127
## [4,]   947.9331  205.1979 50.84440 -10.96112 -1.06553139 0.3172499 -0.3843
8871
## [5,] -1343.8300  304.7787 48.53895 -10.79310 -0.06032618 0.2813083 -0.3634
6016
## [6,] -1528.1006  194.2326 49.49795 -10.84553 -0.09939713 0.2865929 -0.4373
9855
##                PC8
## [1,] -0.009320344
## [2,] -0.015215698
## [3,] -0.012392252
## [4,] -0.008037139
## [5,] -0.014199403
## [6,] -0.014359815
```

```
# Identifying the scores by their deposit status
deposit_pca <- cbind(data.frame(bank1),bank1_pca$x)
head(deposit_pca)
```

```
##   age default balance day duration campaign pdays previous deposit
PC1
## 1  59       0    2343   5     1042        1    -1        0       1   816.0
646
```

```
## 2  56       0      45   5     1467        1    -1        0       1 -1480.8
899
## 3  41       0    1270   5     1389        1    -1        0       1  -256.0
843
## 4  55       0    2476   5      579        1    -1        0       1   947.9
331
## 5  54       0     184   5      673        2    -1        0       1 -1343.8
300
## 6  42       0       0   5      562        2    -1        0       1 -1528.1
006
##         PC2      PC3       PC4         PC5       PC6         PC7
PC8
## 1  668.4989 46.29391 -10.71400 -0.91235277 0.2804991 -0.08508411 -0.009320
344
## 2 1099.0781 40.78714 -10.41854 -0.79540938 0.2112248  0.16765440 -0.015215
698
## 3 1018.0966 42.28868 -10.49809 -0.80813650 0.2370584  0.12939127 -0.012392
252
## 4  205.1979 50.84440 -10.96112 -1.06553139 0.3172499 -0.38438871 -0.008037
139
## 5  304.7787 48.53895 -10.79310 -0.06032618 0.2813083 -0.36346016 -0.014199
403
## 6  194.2326 49.49795 -10.84553 -0.09939713 0.2865929 -0.43739855 -0.014359
815
```

```
# Means of scores for all the PC's classified by Deposit status
tabmeansPC <- aggregate(deposit_pca[,2:6],by=list(Deposit=bank$deposit),mean)
tabmeansPC
```

```
##   Deposit     default  balance      day duration campaign
## 1       0 0.019751405 1280.227 16.10812 223.1303 2.839264
## 2       1 0.009831726 1804.268 15.15825 537.2946 2.141047
```

```
tabmeansPC <- tabmeansPC[rev(order(tabmeansPC$Deposit)),]
tabmeansPC
```

```
##   Deposit     default  balance      day duration campaign
## 2       1 0.009831726 1804.268 15.15825 537.2946 2.141047
## 1       0 0.019751405 1280.227 16.10812 223.1303 2.839264
```

```
tabfmeans <- t(tabmeansPC[,-1])
tabfmeans
```

```
##                      2           1
## default  9.831726e-03    0.0197514
## balance  1.804268e+03 1280.2271412
## day      1.515825e+01   16.1081219
## duration 5.372946e+02  223.1302571
## campaign 2.141047e+00    2.8392644
```

```r
colnames(tabfmeans) <- t(as.vector(tabmeansPC[1]))
tabfmeans
```

```
##                       1            0
## default  9.831726e-03    0.0197514
## balance  1.804268e+03 1280.2271412
## day      1.515825e+01   16.1081219
## duration 5.372946e+02  223.1302571
## campaign 2.141047e+00    2.8392644
```

```r
# Standard deviations of scores for all the PC's classified by Deposit status
tabsdsPC <- aggregate(deposit_pca[,2:6],by=list(Deposit=bank$deposit),sd)
tabfsds <- t(tabsdsPC[,-1])
colnames(tabfsds) <- t(as.vector(tabsdsPC[1]))
tabfsds
```

```
##                       0            1
## default     0.1391567 9.867575e-02
## balance  2933.4119343 3.501105e+03
## day         8.3220713 8.501875e+00
## duration  208.5775301 3.925253e+02
## campaign    3.2444741 1.921826e+00
```

```r
t.test(PC1~bank$deposit,data=deposit_pca)
```

```
##
##  Welch Two Sample t-test
##
## data:  PC1 by bank$deposit
## t = -8.5333, df = 10359, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -645.385 -404.267
## sample estimates:
## mean in group 0 mean in group 1
##       -248.6834        276.1425
```

```r
t.test(PC2~bank$deposit,data=deposit_pca)
```

```
##
##  Welch Two Sample t-test
##
## data:  PC2 by bank$deposit
## t = -51.649, df = 7853.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -324.4163 -300.6914
## sample estimates:
## mean in group 0 mean in group 1
##       -148.1004        164.4534
```

```r
t.test(PC3~bank$deposit,data=deposit_pca)
```

```
##
##  Welch Two Sample t-test
##
## data:  PC3 by bank$deposit
## t = 17.396, df = 10206, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  31.71374 39.76818
## sample estimates:
## mean in group 0 mean in group 1
##        16.93549       -18.80547
```

```
t.test(PC4~bank$deposit,data=deposit_pca)
```

```
##
##  Welch Two Sample t-test
##
## data:  PC4 by bank$deposit
## t = 3.9672, df = 10966, p-value = 7.317e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3199010 0.9447568
## sample estimates:
## mean in group 0 mean in group 1
##       0.2996226      -0.3327063
```

```
t.test(PC5~bank$deposit,data=deposit_pca)
```

```
##
##  Welch Two Sample t-test
##
## data:  PC5 by bank$deposit
## t = 9.7124, df = 9863.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3829862 0.5766678
## sample estimates:
## mean in group 0 mean in group 1
##       0.2273611      -0.2524659
```

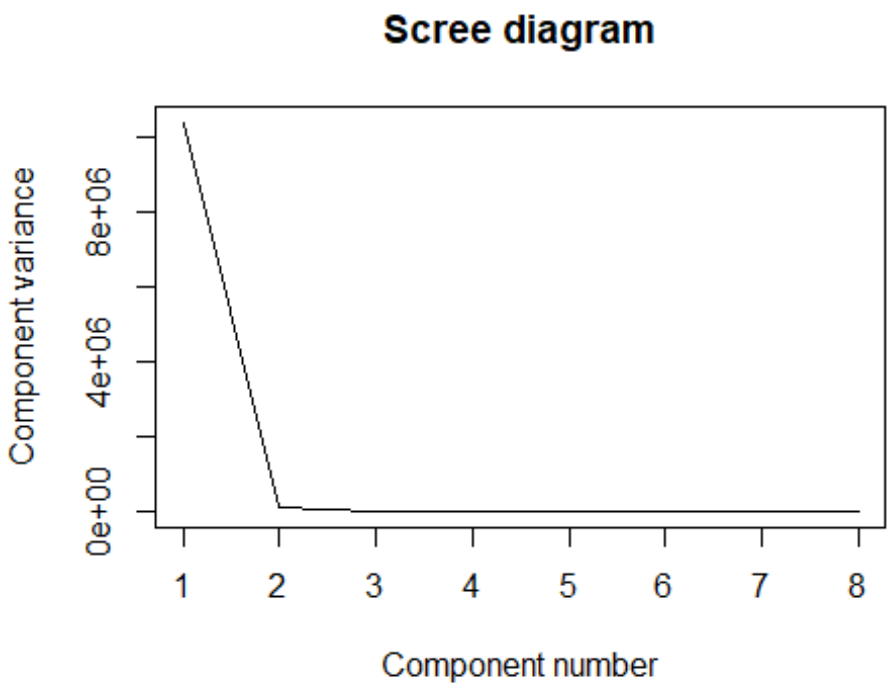
```
#F-test
var.test(PC1~bank$deposit,data=deposit_pca)
```

```
##
##  F test to compare two variances
##
## data:  PC1 by bank$deposit
## F = 0.70202, num df = 5872, denom df = 5288, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6660526 0.7398708
```

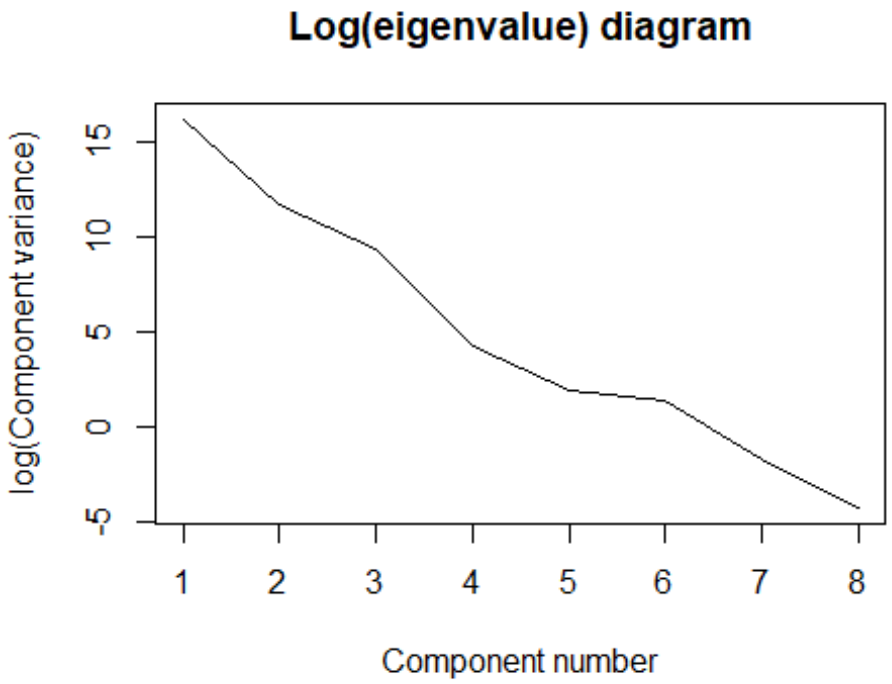
```
## sample estimates:
## ratio of variances
##          0.7020178

var.test(PC2~bank$deposit,data=deposit_pca)

##
##  F test to compare two variances
##
## data:  PC2 by bank$deposit
## F = 0.28137, num df = 5872, denom df = 5288, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2669565 0.2965431
## sample estimates:
## ratio of variances
##          0.2813715

var.test(PC3~bank$deposit,data=deposit_pca)

##
##  F test to compare two variances
##
## data:  PC3 by bank$deposit
## F = 0.66204, num df = 5872, denom df = 5288, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6281230 0.6977374
## sample estimates:
## ratio of variances
##            0.66204

var.test(PC4~bank$deposit,data=deposit_pca)

##
##  F test to compare two variances
##
## data:  PC4 by bank$deposit
## F = 0.9451, num df = 5872, denom df = 5288, p-value = 0.03509
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8966825 0.9960612
## sample estimates:
## ratio of variances
##          0.9451011

var.test(PC5~bank$deposit,data=deposit_pca)

##
##  F test to compare two variances
##
## data:  PC5 by bank$deposit
```

```
## F = 2.6818, num df = 5872, denom df = 5288, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  2.544403 2.826397
## sample estimates:
## ratio of variances
##           2.681794
```

```
plot(eigen_bank, xlab = "Component number", ylab = "Component variance", type = "l", main = "Scree diagram")
```




```
plot(log(eigen_bank), xlab = "Component number",ylab = "log(Component variance)", type="l",main = "Log(eigenvalue) diagram")
```


Log(eigenvalue) diagram
log(Component variance)
15
10
5
0
-5
1
2
3
4
5
6
7
8
Component number