Ly,Cuong,Assignment-1

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:

```
library(faraway)
data("worldcup")
```

==========

Question 1

==========

```
dim(worldcup)             # There are 595 observations with 7 variables
```

```
## [1] 595   7
```

```
names(worldcup)           # 7 attributes in this dataset are Team, Position, Time, Shots,
```

```
## [1] "Team"     "Position" "Time"     "Shots"    "Passes"   "Tackles"  "Saves"
```

```
class(worldcup)           # Data Frame
```

```
## [1] "data.frame"
```

```
sapply(worldcup, class)    # Numerical variables are Time, Shots, Passes, Tackles, Saves
```

```
##      Team  Position      Time     Shots    Passes   Tackles     Saves
##  "factor"  "factor" "integer" "integer" "integer" "integer" "integer"
```

```
                           # Categorical variables are Team and Position
```

==========

Question 2

==========

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# a. The mean of saves that all players made
mean(worldcup$Saves)    # The mean of saves is 0.6672269
```

```
## [1] 0.6672269
```

```
# b. Number of players in each position
count(worldcup,c('Position'))    # Defender: 188   Forward: 143    Goalkeeper: 36
```

```
##   c("Position")   n
## 1      Position 595
```

```
# c. The median number of shots that all plays made
median(worldcup$Shots)    # The median numbers of shots is 1
```

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

```
# d. The interquartile range (IQR) of passes that players made
IQR(worldcup$Passes)    # The interquartile range of passes is 86.5
```

```
## [1] 86.5
```

==========

Question 3

==========

```r
# a. Number of forwards on each team
table(worldcup$Team,worldcup$Position)[,2]
```

```
##      Algeria    Argentina    Australia       Brazil     Cameroon        Chile
##            3            6            3            4            5            5
##      Denmark      England       France      Germany        Ghana       Greece
##            3            4            5            6            5            5
##     Honduras        Italy  Ivory Coast        Japan       Mexico  Netherlands
##            5            6            5            4            5            5
##  New Zealand      Nigeria  North Korea     Paraguay     Portugal       Serbia
##            4            6            4            5            4            3
##     Slovakia     Slovenia South Africa  South Korea        Spain  Switzerland
##            5            5            3            3            3            4
##          USA      Uruguay
##            5            5
```

```r
# Team had the most shots in total among all its forwards

arrange(summarize(group_by(filter(worldcup,Position=='Forward'),Team),total_shots = sum(Shots)),desc(to
```

```
## # A tibble: 32 x 2
##    Team        total_shots
##    <fct>             <int>
##  1 Uruguay              46
##  2 Argentina            45
##  3 Germany              41
##  4 Netherlands          34
##  5 Spain                33
##  6 Ghana                32
##  7 Portugal             28
##  8 Paraguay             25
##  9 Brazil               23
## 10 USA                  21
## # ... with 22 more rows
```

```r
# Team had the most shots in total among all its forwards is Uruguay
```

```r
# b. Team(s) had the defender with the most tackles

most_tackle <- max(tapply(worldcup$Tackles,worldcup$Team,max)) # The most tackles is 34
worldcup$Team[worldcup$Tackles==most_tackle]
```

```
## [1] Uruguay
## 32 Levels: Algeria Argentina Australia Brazil Cameroon Chile ... Uruguay
```

```r
# Uruguay is the team that had defender  with the most tackles
```

```r
# c. Player played the longest time in the field? Which team did he come from?
#    How long did he play in the field?
```

```r
max_time <- max(aggregate(worldcup$Time ~ worldcup$Team, max, data = worldcup)[,2]) # The max time is 5

worldcup[which(worldcup$Time==max_time),]  # Use which to determine the index
```

```
##                 Team   Position Time Shots Passes Tackles Saves
## Arevalo Rios Uruguay Midfielder  570     5    195      21     0
## Maxi Pereira Uruguay Midfielder  570     5    182      15     0
## Muslera      Uruguay Goalkeeper  570     0     75       0    16
```

```r
# Arevalo Rios, Maxi Pereira, and Muslera played the longest time in the field.
# They all came from Uruguay and they played 570 minutes
```

```r
# d. The mean number of saves just among the goalkeepers
mean(worldcup[worldcup$Position == 'Goalkeeper', 'Saves'])
```

```
## [1] 11.02778
```

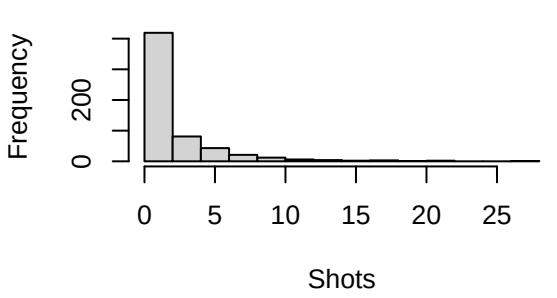
```r
# The mean number of saves just among the goalkeepers is 11.02778
```

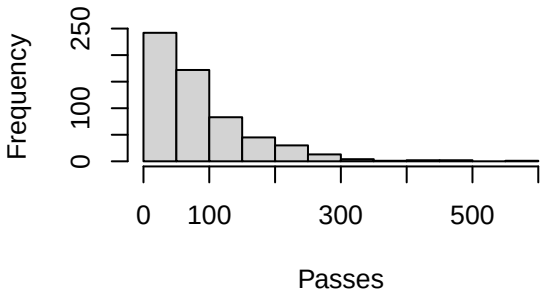
==========

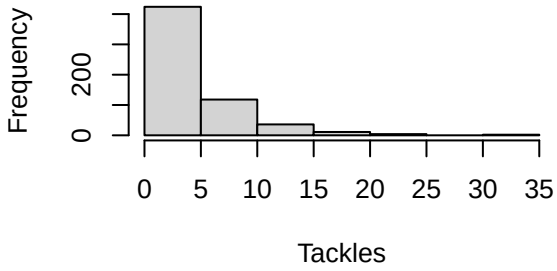
Question 4

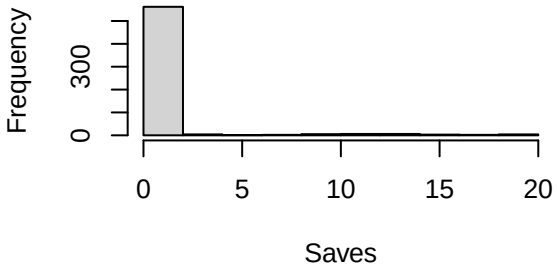
==========

```r
# a. Create a histogram
par(mfrow=c(2,2)) ##2 plots in each row, 2 plots in each column
for(i in 4:7) { hist(worldcup[,i], xlab=names(worldcup)[i], main=NULL) }
```



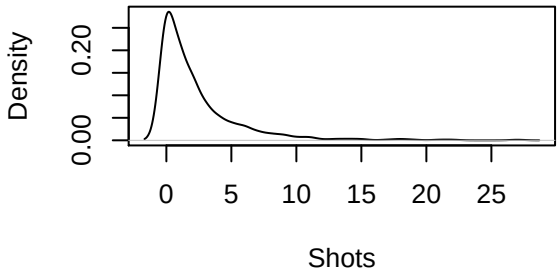


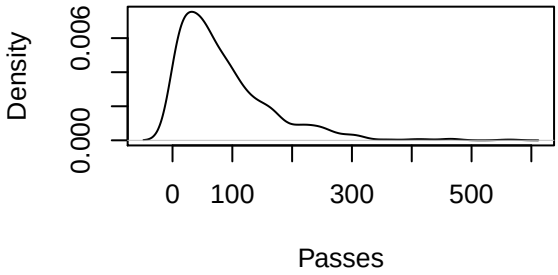


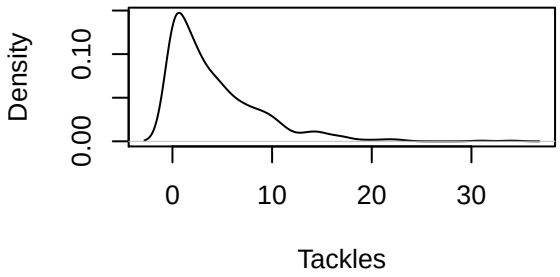


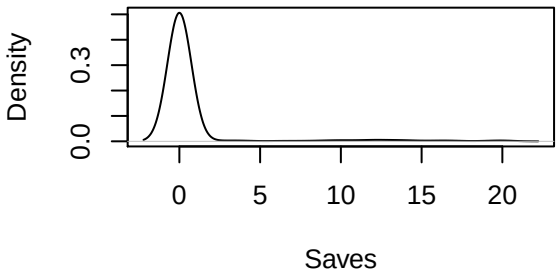

```
# We can see that four graphs are all skewed to the right
# A majority of players attempted less than 5 shots
# More than 90% of players passed less than 200
# Players mostly tackled less than 5 attempts
# 1 save is the majority
```

```
# b. Create a kernel density
par(mfrow=c(2,2))
for(i in 4:7) { plot(density(worldcup[,i]), main=NA, xlab=names(worldcup)[i]) }
```



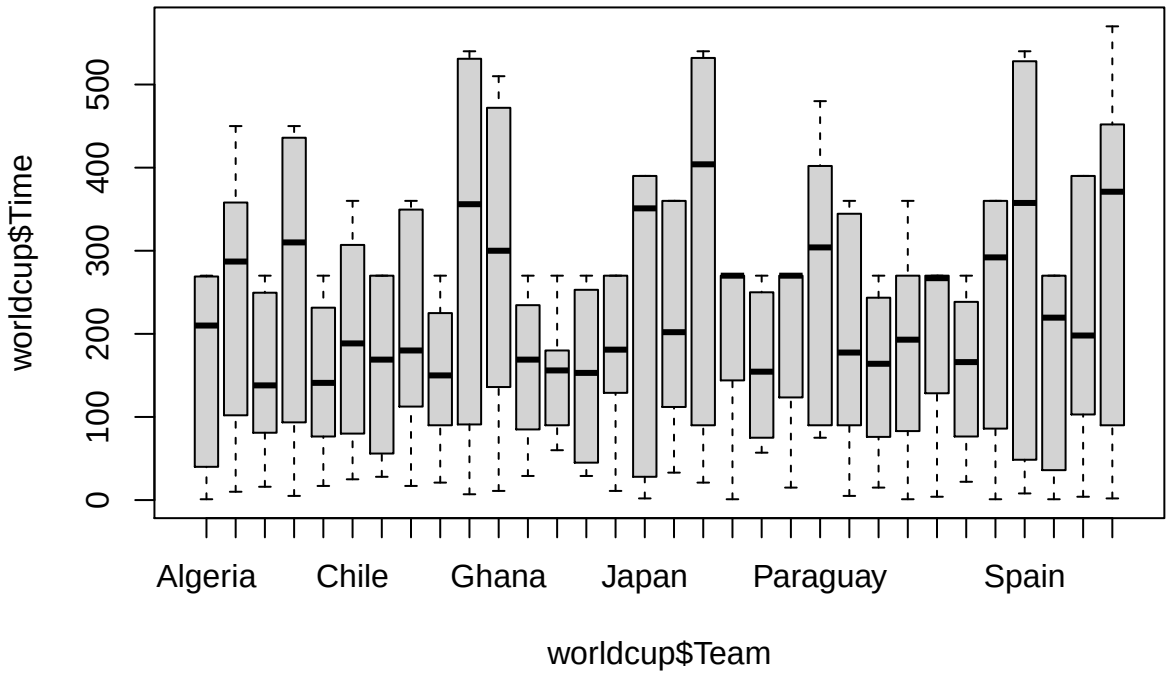







```
# Saves is the most difficult variable to distinguish the team's performance because the
# number of observations center heavily around 0. In contrast, although Shots, Passess and
# Tackles are also skewed to the right, there are some observations that lie in the right
# (upper levels)
```

```
# c.  A boxplot to show the distribution of playing time in field by each country
boxplot(worldcup$Time~worldcup$Team)
```




```
# The mean of playing time of more than half the teams is below 200
# 12 out of 32 teams have range of playing time more than 200
# The minimum value is fairly similar among teams. However, there's a big dissimilarity in
# the maximum values among teams
```

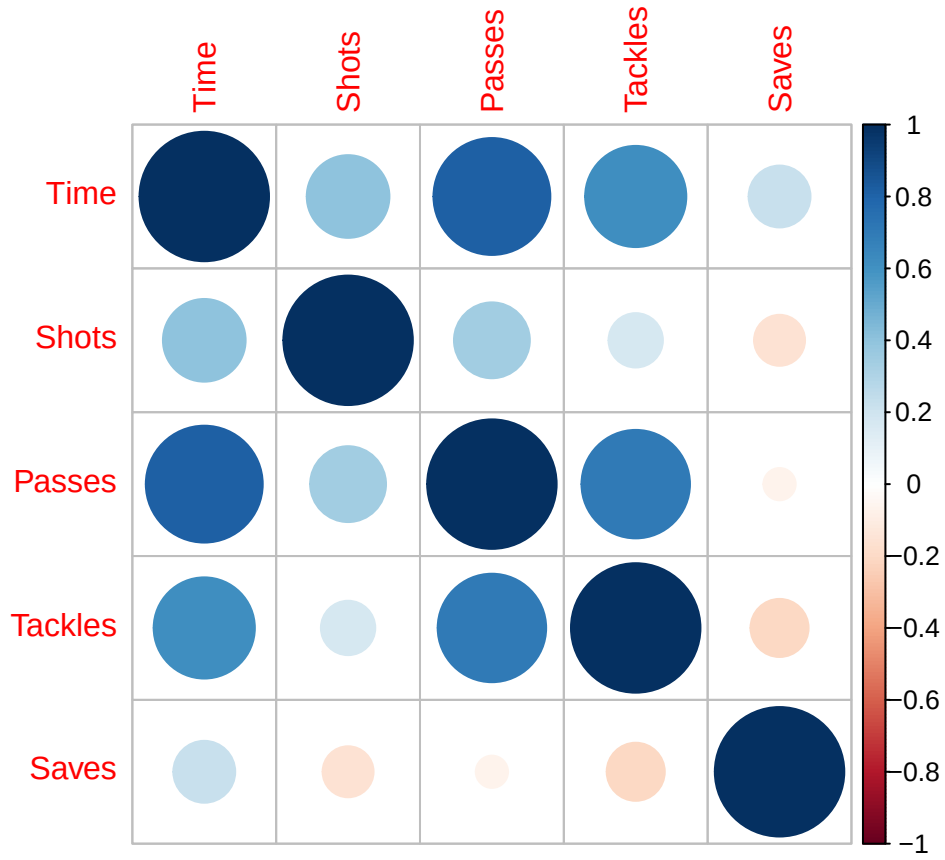
```
# d. Correlation among all the numeric variables. Plot the correlation matrix
cor(worldcup[,3:7])
```

```
##              Time      Shots      Passes    Tackles       Saves
## Time    1.0000000  0.4079231  0.81511932  0.6106735  0.22847723
## Shots   0.4079231  1.0000000  0.34316326  0.1762829 -0.15495828
## Passes  0.8151193  0.3431633  1.00000000  0.7020965 -0.06205701
## Tackles 0.6106735  0.1762829  0.70209651  1.0000000 -0.20118978
## Saves   0.2284772 -0.1549583 -0.06205701 -0.2011898  1.00000000
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

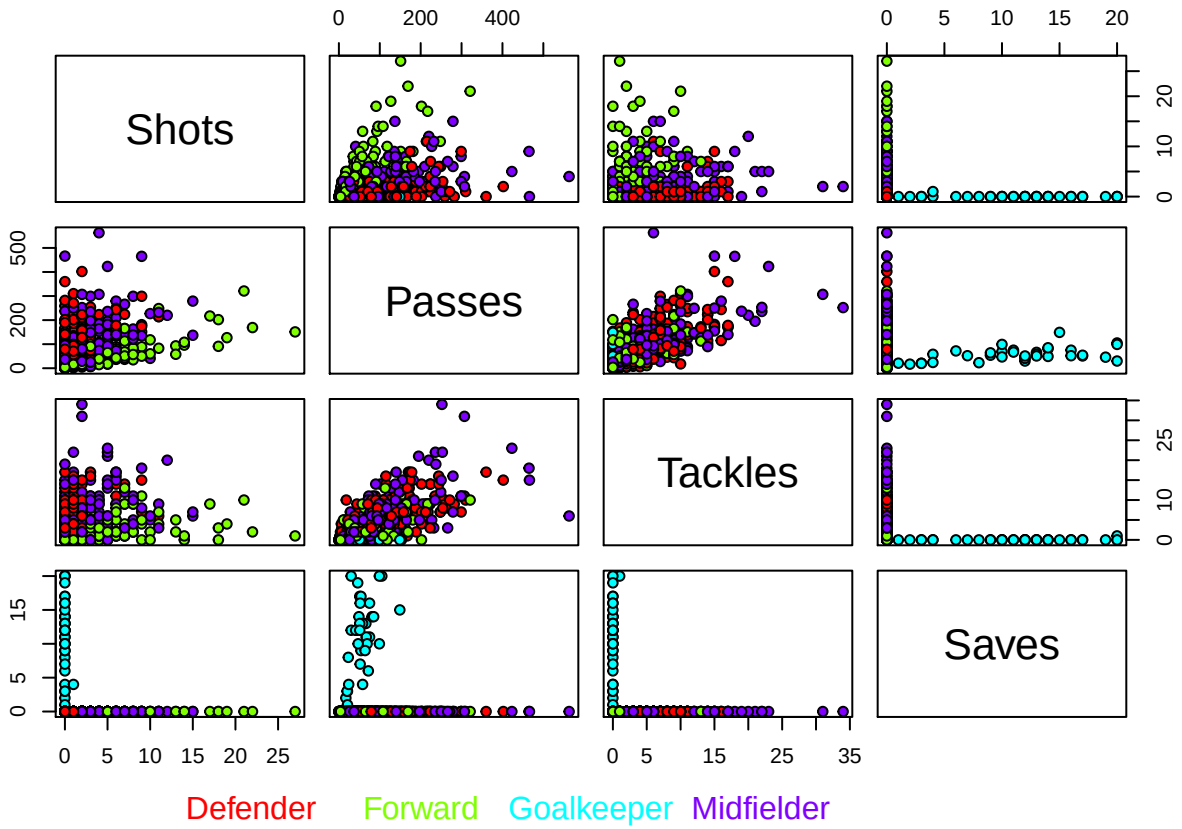
```
corrplot(cor(worldcup[,3:7]))
```




```
# Passes and Time have the highest correlation
# Tackles and Save have the lowest correlation
# Passes and Time also have high correlation
# Passes and Tackles have high correlation
# For Saves, except Time, it has a negative correlation with Shots, Passess, and Tackles
```

```
# e. A scatter plot to show the associations among shots, passes, tackles and saves by each position.

plot(worldcup[4:7], pch=21, bg=rainbow(4)[worldcup$Position], oma=c(4,3,3,3))
mtext(at=c(0.2,0.35,0.5,0.65), side=1, line=4, text=levels(worldcup$Position), col=rainbow(4))
```




```
# Midfielder has more passes but forward has the most shots
# Midfielder also has the most tackles
# Goalkeeper has the most saves
# Defender does tackles, but less than midfielder
# Midfielder is the busiest position on the field.
```