NAVARI FAMILY
CENTER for
DIGITAL
SCHOLARSHIP

HANDS-ON R

October 4, 2018

James Ng, PhD
james.ng@nd.edu

UNIVERSITY OF
NOTRE DAME
Hesburgh Libraries

Tutorials

1. Exploring World Cup data
2. Plotting graphs using ggplot2
3. Factors
4. Repetition using for loop, apply family, function

TUTORIAL 1
Exploring World Cup data

UNIVERSITY OF NOTRE DAME
Hesburgh Libraries
NAVARI FAMILY
CENTER for DIGITAL SCHOLARSHIP

Tutorial 1: World Cup data

```
install.packages('faraway')


library(faraway)


dat <- data.frame(worldcup)


head(dat)

str(dat)

summary(dat)
```

TUTORIAL 1: World Cup (cont'd)

1. Retrieve the value in row 17, column 3.
2. Retrieve the first five columns for the first six rows.
3. Retrieve values by row and column names.
4. Retrieve all column values for row Alonso.
5. Retrieve all row values for column Team.
6. Extract the row names and store them in a new column Player.
7. Reorder column Player to the furthest left.
8. What's the max number of shots taken on each team?
9. Which player took the most shots on each team?

TUTORIAL 2
Plotting graphs with ggplot2

UNIVERSITY OF NOTRE DAME
Hesburgh Libraries
NAVARI FAMILY
CENTER for DIGITAL SCHOLARSHIP

TUTORIAL 2: plotting

1. Start a new R Script.
2. Clear objects in environment, set working directory

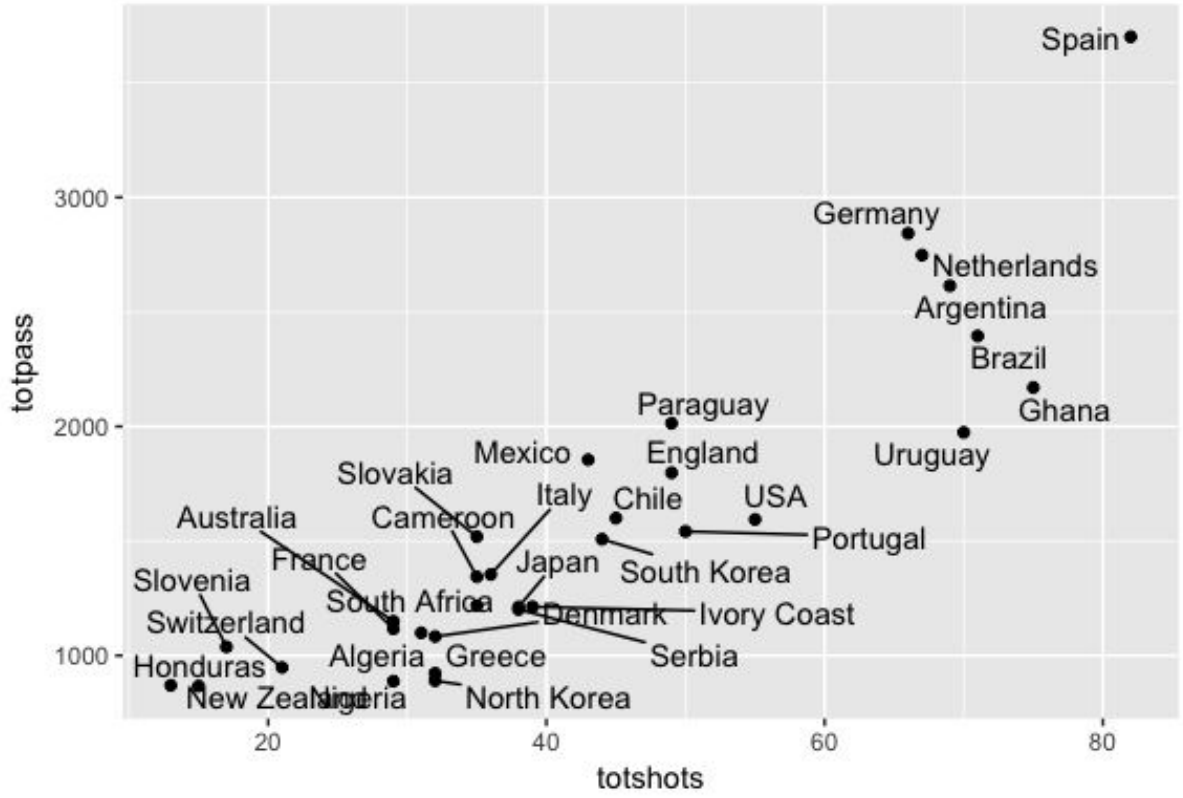
   rm(list=ls())

   setwd('/Users/jng2/Dropbox/Work/Library/CDS/R-RStudio/hands-on')
3. Load 'tidyverse' (may have to install first)

   library(tidyverse)
4. Reload 'worldcup' data

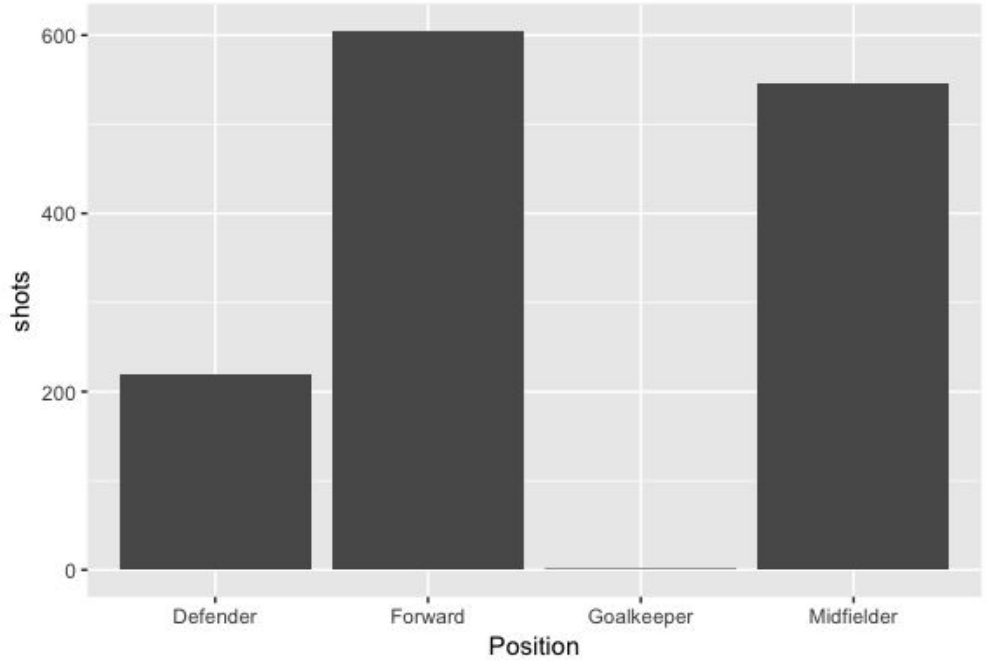
TUTORIAL 2: plotting (cont'd)

Total passes vs total shots



TUTORIAL 2: plotting (cont'd)

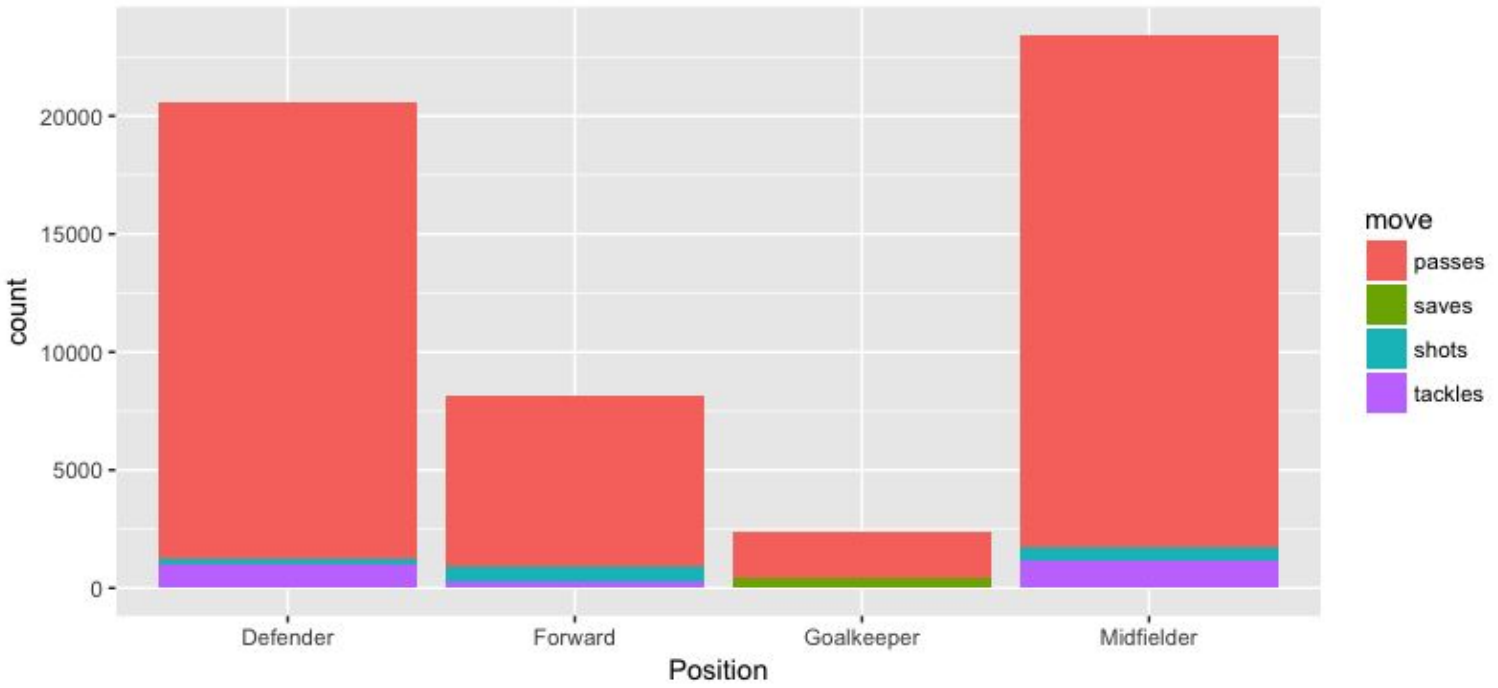
Number of shots by position




```
# A tibble: 4 x 5
  Position   passes shots tackles saves
  <fct>       <int> <int>   <int> <int>
1 Defender    19297   219    1027     0
2 Forward      7268   605     289     0
3 Goalkeeper   2003     1       1   397
4 Midfielder  21722   546    1177     0
```

TUTORIAL 2: plotting (cont'd)

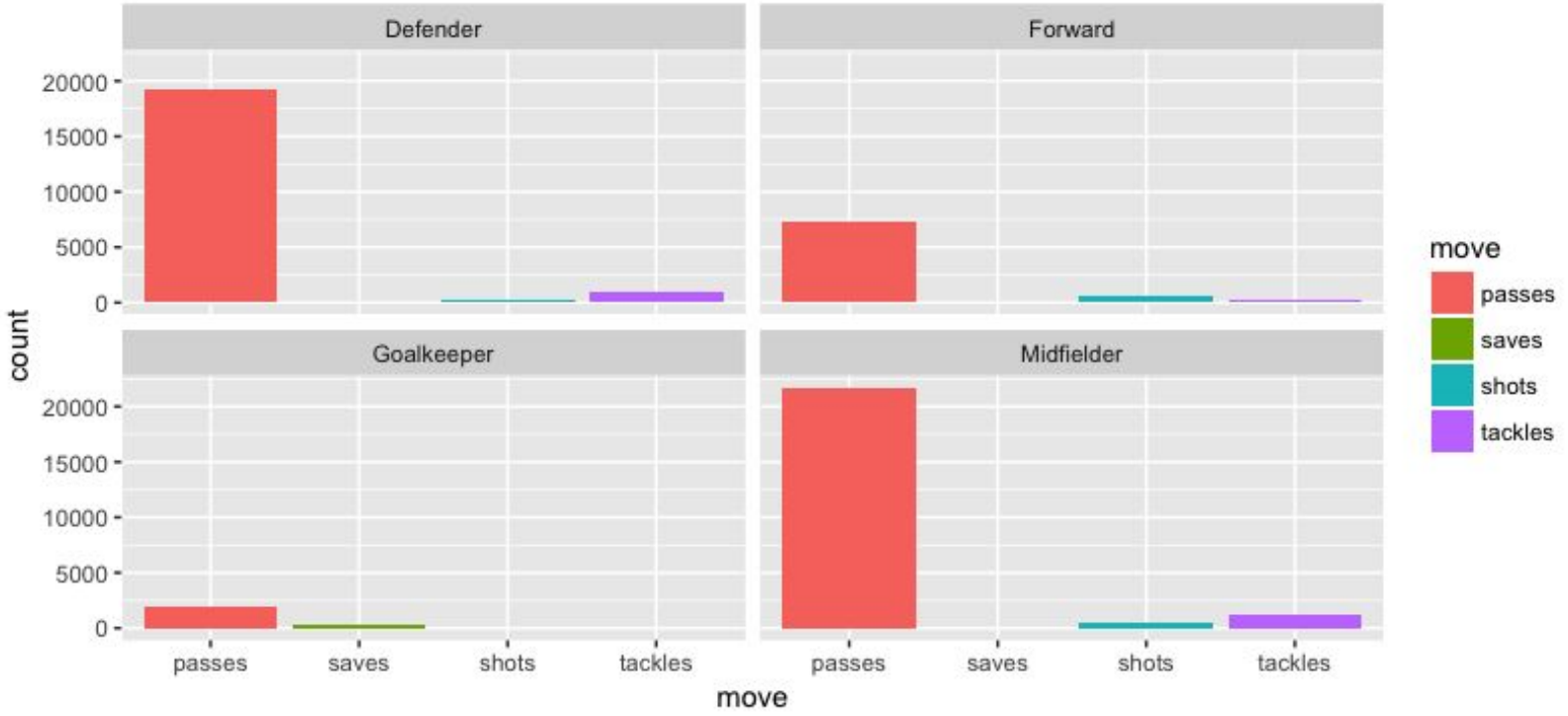
Moves (shots, passes, tackles, saves) by position




```
# A tibble: 16 x 3
   Position   move    count
   <fct>      <chr>   <int>
 1 Defender   passes  19297
 2 Forward    passes   7268
 3 Goalkeeper passes   2003
 4 Midfielder passes  21722
 5 Defender   shots     219
 6 Forward    shots     605
 7 Goalkeeper shots       1
 8 Midfielder shots     546
 9 Defender   tackles  1027
10 Forward    tackles   289
11 Goalkeeper tackles     1
12 Midfielder tackles  1177
13 Defender   saves       0
14 Forward    saves       0
15 Goalkeeper saves     397
16 Midfielder saves       0
```

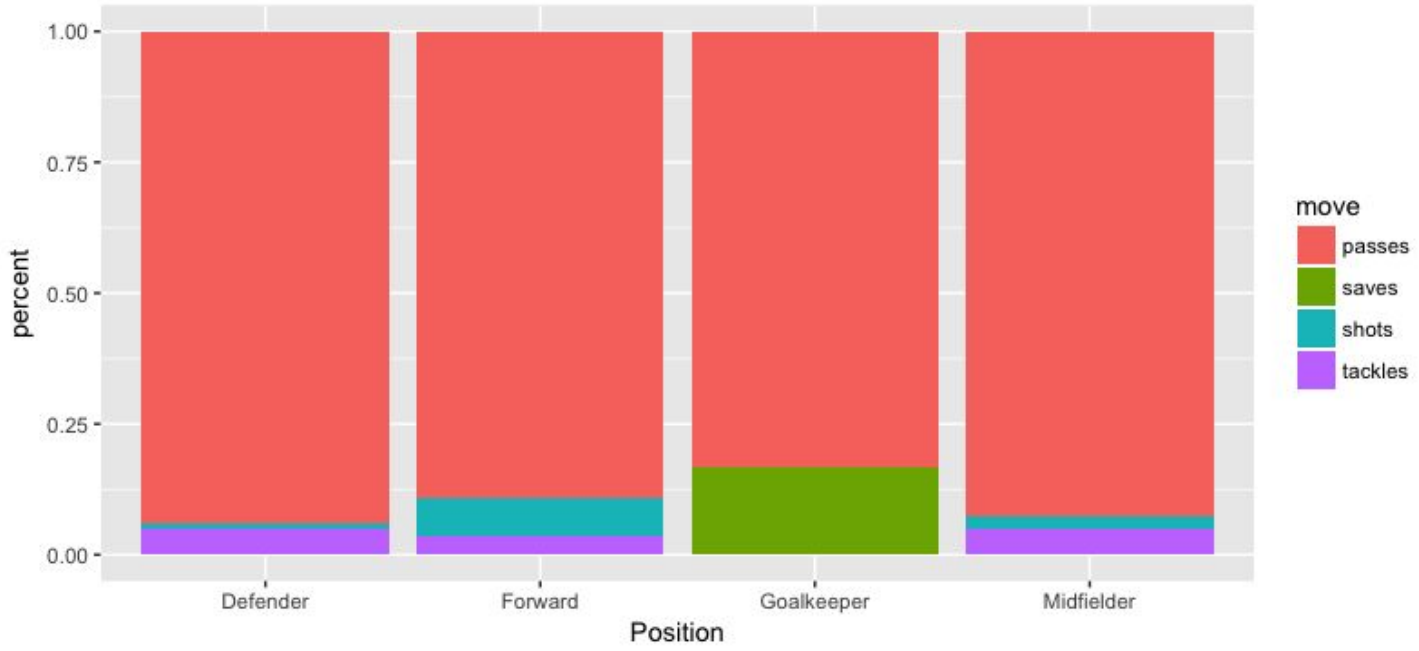
TUTORIAL 2: plotting (cont'd)

Moves by position: faceting



TUTORIAL 2: plotting (cont'd)

Moves (%) by position




```
# A tibble: 16 x 4
# Groups:   Position [4]
   Position   move    count  percent
   <fct>      <chr>   <int>    <dbl>
 1 Defender   passes  19297 0.939
 2 Forward    passes   7268 0.890
 3 Goalkeeper passes   2003 0.834
 4 Midfielder passes  21722 0.927
 5 Defender   shots     219 0.0107
 6 Forward    shots     605 0.0741
 7 Goalkeeper shots       1 0.000416
 8 Midfielder shots     546 0.0233
 9 Defender   tackles  1027 0.0500
10 Forward    tackles   289 0.0354
11 Goalkeeper tackles     1 0.000416
12 Midfielder tackles  1177 0.0502
13 Defender   saves       0 0.
14 Forward    saves       0 0.
15 Goalkeeper saves     397 0.165
16 Midfielder saves       0 0.
```

TUTORIAL 2: plotting (cont'd)

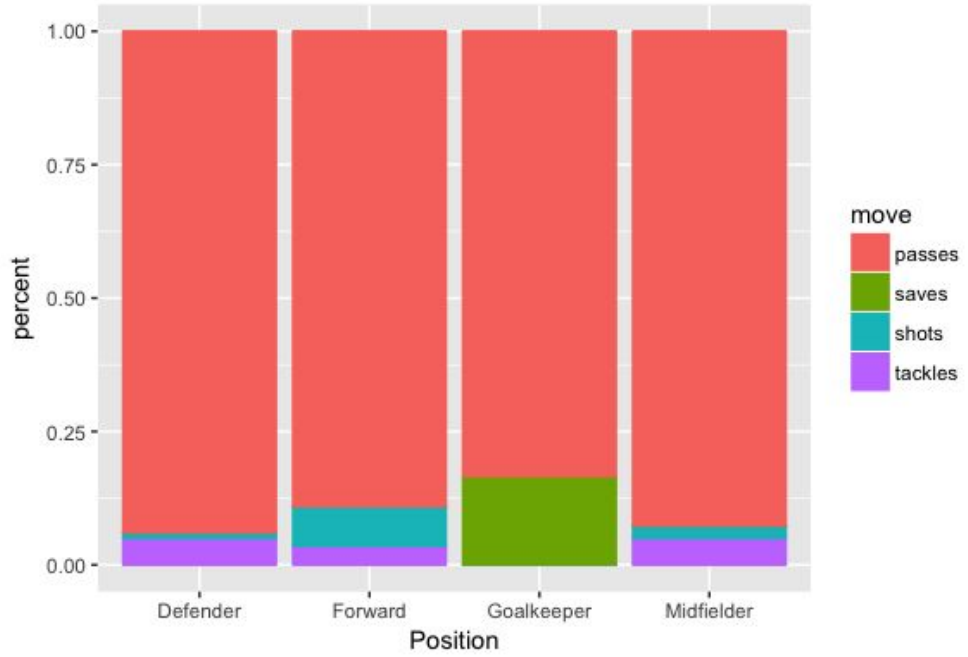
Moves (%) by position

```
# first compute percentage of each move taken per position

datplot3b <- datplot3 %>%
        group_by(Position) %>%
        mutate(percent = count/sum(count))


# then plot the graph

ggplot(datplot3b, aes(x=Position, y=percent, fill=move)) +
  geom_bar(stat="identity", position="fill")
```

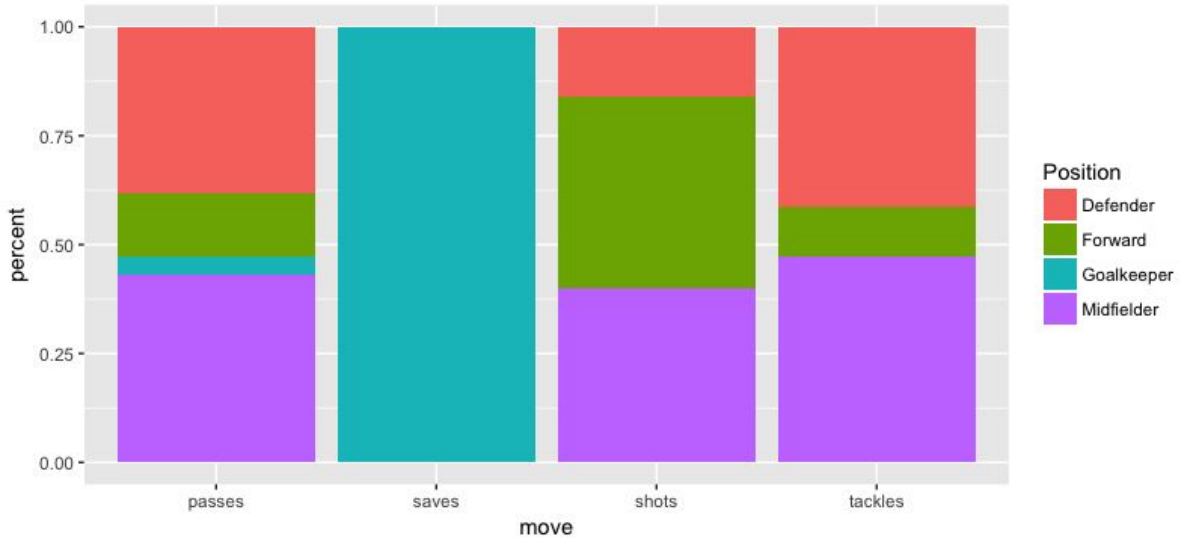


TUTORIAL 2: plotting (cont'd)

Positions (%) by move

```
datplot3c <- datplot3 %>%
        group_by(move) %>%
        mutate(percent = count/sum(count))


ggplot(datplot3c, aes(x=move, y=percent, fill=Position)) +
  geom_bar(stat="identity", position="fill")
```



TUTORIAL 3
Factors

UNIVERSITY OF NOTRE DAME
Hesburgh Libraries
NAVARI FAMILY
CENTER for DIGITAL SCHOLARSHIP

TUTORIAL 3: factors

1. Continuing using ‘worldcup’ data.
2. Team and Position are factor variables – R knows how to deal with factors properly

```
summary(dat$Team)

summary(dat$Position)

levels(dat$Position)

table(dat$Team, dat$Position)


mod <- lm(Shots ~ Position + Time + Passes + Tackles + Saves, data=dat)

summary(mod)
```

TUTORIAL 3: factors (cont'd)

1. Sometimes factor levels should be ordered
2. Create a factor variable categorizing amount of shooting: low, medium, high, extremely high
3. Order the levels of this factor variable such that low < medium < high < extremely high

TUTORIAL 4
Repetition using for loop, apply family, function

UNIVERSITY OF NOTRE DAME
Hesburgh Libraries
NAVARI FAMILY
CENTER for DIGITAL SCHOLARSHIP

TUTORIAL 3: repetition

1. Load ‘iris’ data
2. Compute the number of unique values in each column in iris using
   - For loop
   - apply
   - lapply
   - sapply

   This is what you should obtain:

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
35	23	43	22	3