21BDS0340

Abhinav Dinesh Srivatsa

Exploratory Data Analysis Lab

Experiment – IV

Code:

```
library(dplyr)
library(ggplot2)
library(lubridate)
setwd("/Users/abhi/College Work/Year 4 Semester 1 (Sem 7)/Exploratory Data Analysis Lab/Assignment 2")
data = read.csv("./DS2_Match.csv")
```

Output:

```
> library(dplyr)
> library(ggplot2)
> library(lubridate)
> setwd("/Users/abhi/College Work/Year 4 Semester 1 (Sem 7)/Exploratory Data Analysis Lab/Assignment 2")
> data = read.csv("./DS2_Match.csv")
```

Code:

```
# viewing data
View(data)
```

Output:

data

Filter

	Match_Id	Match_Date	Team_Name_Id	Opponent_Team_Id	Season_Id	Venue_Name	Toss_Winner_Id	Toss_Decision	IS_Superover	IS_Result	Is_DuckWorthLewis	Win_Type
1	335987	18-Apr-08	2	1	1	M Chinnaswamy Stadium	2	field	0	1	0	by runs
2	335988	19-Apr-08	4	3	1	Punjab Cricket Association Stadium, Mohali	3	bat	0	1	0	by runs
3	335989	19-Apr-08	6	5	1	Feroz Shah Kotla	5	bat	0	1	0	by wickets
4	335990	20-Apr-08	7	2	1	Wankhede Stadium	7	bat	0	1	0	by wickets
5	335991	20-Apr-08	1	8	1	Eden Gardens	8	bat	0	1	0	by wickets
6	335992	21-Apr-08	5	4	1	Sawai Mansingh Stadium	4	bat	0	1	0	by wickets
7	335993	22-Apr-08	8	6	1	Rajiv Gandhi International Stadium, Uppal	8	bat	0	1	0	by wickets
8	335994	23-Apr-08	3	7	1	MA Chidambaram Stadium, Chepauk	7	field	0	1	0	by runs
9	335995	24-Apr-08	8	5	1	Rajiv Gandhi International Stadium, Uppal	5	field	0	1	0	by wickets
10	335996	25-Apr-08	4	7	1	Punjab Cricket Association Stadium, Mohali	7	field	0	1	0	by runs
11	335997	26-Apr-08	2	5	1	M Chinnaswamy Stadium	5	field	0	1	0	by wickets
12	335998	26-Apr-08	3	1	1	MA Chidambaram Stadium, Chepauk	1	bat	0	1	0	by wickets
13	335999	27-Apr-08	7	8	1	Dr DY Patil Sports Academy	8	field	0	1	0	by wickets
14	336000	27-Apr-08	4	6	1	Punjab Cricket Association Stadium, Mohali	6	bat	0	1	0	by wickets
15	336001	28-Apr-08	2	3	1	M Chinnaswamy Stadium	3	bat	0	1	0	by runs
16	336002	29-Apr-08	1	7	1	Eden Gardens	1	bat	0	1	0	by wickets
17	336003	30-Apr-08	6	2	1	Feroz Shah Kotla	2	field	0	1	0	by runs
18	336004	01-May-08	8	4	1	Rajiv Gandhi International Stadium, Uppal	4	field	0	1	0	by wickets
19	336005	01-May-08	5	1	1	Sawai Mansingh Stadium	5	bat	0	1	0	by runs
20	336006	02-May-08	3	6	1	MA Chidambaram Stadium, Chepauk	3	bat	0	1	0	by wickets
21	336007	25-May-08	8	2	1	Rajiv Gandhi International Stadium, Uppal	8	bat	0	1	0	by wickets
22	336008	03-May-08	4	1	1	Punjab Cricket Association Stadium, Mohali	4	bat	0	1	0	by runs
23	336009	04-May-08	7	6	1	Dr DY Patil Sports Academy	6	field	0	1	0	by runs
24	336010	04-May-08	5	3	1	Sawai Mansingh Stadium	3	bat	0	1	0	by wickets
25	336011	05-May-08	2	4	1	M Chinnaswamy Stadium	4	field	0	1	0	by wickets
26	336012	06-May-08	3	8	1	MA Chidambaram Stadium, Chepauk	8	field	0	1	0	by wickets
27	336013	07-May-08	7	5	1	Dr DY Patil Sports Academy	7	field	0	1	0	by wickets
28	336014	08-May-08	6	3	1	Feroz Shah Kotla	3	field	0	1	0	by wickets
29	336015	08-May-08	1	2	1	Eden Gardens	1	bat	0	1	0	by runs
30	336016	09-May-08	5	8	1	Sawai Mansingh Stadium	5	field	0	1	0	by wickets
31	336017	28-May-08	2	7	1	M Chinnaswamy Stadium	7	field	0	1	0	by wickets
32	336018	10-May-08	3	4	1	MA Chidambaram Stadium, Chepauk	4	field	0	1	0	by runs
33	336019	11-May-08	8	1	1	Rajiv Gandhi International Stadium, Uppal	1	bat	0	1	0	by runs
34	336020	11-May-08	5	6	1	Sawai Mansingh Stadium	5	field	0	1	0	by wickets
35	336021	12-May-08	4	2	1	Punjab Cricket Association Stadium, Mohali	2	bat	0	1	0	by wickets
36	336022	13-May-08	1	6	1	Eden Gardens	1	bat	0	1	0	by runs
37	336023	14-May-08	7	3	1	Wankhede Stadium	7	field	0	1	0	by wickets

Showing 1 to 37 of 577 entries, 19 total columns

Code:

```
# dimentions and names of columns
dim(data)
names(data)
```

Output:

```
> dim(data)
[1] 577  19
> names(data)
 [1] "Match_Id"            "Match_Date"          "Team_Name_Id"
"Opponent_Team_Id"
 [5] "Season_Id"           "Venue_Name"          "Toss_Winner_Id"
"Toss_Decision"
 [9] "IS_Superover"        "IS_Result"           "Is_DuckWorthLewis"   "Win_Type"
[13] "Won_By"              "Match_Winner_Id"     "Man_Of_The_Match_Id"
"First_Umpire_Id"
[17] "Second_Umpire_Id"    "City_Name"           "Host_Country"
```

Code:

```
# sorting data by win type
head(sort(data$Win_Type))
```

Output:

```
> head(sort(data$Win_Type))
[1] "by runs" "by runs" "by runs" "by runs" "by runs" "by runs"
```

Code:

```
# summary of data
summary(data)
```

Output:

```
> summary(data)
    Match_Id       Match_Date         Team_Name_Id    Opponent_Team_Id   Season_Id
 Min.   :335987   Length:577         Min.   : 1.000   Min.   : 1.000   Min.
:1.000
 1st Qu.:419140   Class :character   1st Qu.: 3.000   1st Qu.: 3.000   1st
Qu.:3.000
 Median :548353   Mode  :character   Median : 5.000   Median : 5.000   Median
:5.000
 Mean   :591636                      Mean   : 5.102   Mean   : 5.211   Mean
:5.029
 3rd Qu.:734004                      3rd Qu.: 7.000   3rd Qu.: 7.000   3rd
Qu.:7.000
 Max.   :981024                      Max.   :13.000   Max.   :13.000   Max.
:9.000

  Venue_Name        Toss_Winner_Id   Toss_Decision       IS_Superover
IS_Result
 Length:577         Min.   : 1.000   Length:577         Min.   :0.0000   Min.
:0.0000
```

```
 Class :character   1st Qu.: 3.000   Class :character   1st Qu.:0.0000   1st
Qu.:1.0000
 Mode  :character   Median : 5.000   Mode  :character   Median :0.0000   Median
:1.0000
                    Mean   : 5.192                      Mean   :0.0104   Mean
:0.9948
                    3rd Qu.: 7.000                      3rd Qu.:0.0000   3rd
Qu.:1.0000
                    Max.   :13.000                      Max.   :1.0000   Max.
:1.0000

 Is_DuckWorthLewis   Win_Type            Won_By          Match_Winner_Id
 Min.   :0.000     Length:577         Length:577         Min.   : 1.000
 1st Qu.:0.000     Class :character   Class :character   1st Qu.: 3.000
 Median :0.000     Mode  :character   Mode  :character   Median : 5.000
 Mean   :0.026                                           Mean   : 4.991
 3rd Qu.:0.000                                           3rd Qu.: 7.000
 Max.   :1.000                                           Max.   :13.000
                                                         NA's   :3
 Man_Of_The_Match_Id First_Umpire_Id Second_Umpire_Id  City_Name
Host_Country
 Min.   :  1.0       Min.   :470.0   Min.   :471.0    Length:577         Length:577
 1st Qu.: 40.0       1st Qu.:475.0   1st Qu.:488.0    Class :character   Class
:character
 Median :105.5       Median :482.0   Median :493.0    Mode  :character   Mode
:character
 Mean   :139.8       Mean   :484.1   Mean   :495.2
 3rd Qu.:209.5       3rd Qu.:493.0   3rd Qu.:500.0
 Max.   :460.0       Max.   :511.0   Max.   :521.0
 NA's   :3
```

Code:

```
# finding min and max of first umpire id
min(data$First_Umpire_Id)
max(data$First_Umpire_Id)
```

Output:

```
> min(data$First_Umpire_Id)
[1] 470
> max(data$First_Umpire_Id)
[1] 511
```

Code:

```
# finding mean and median of won by amount
data$Won_By = sapply(data$Won_By, function(x) {
  if (x == "NULL") {
    return(0)
  }
  x
})
data$Won_By = as.numeric(data$Won_By)
```

```
mean(data$Won_By)
median(data$Won_By)
```

Output:

```
> mean(data$Won_By)
[1] 17.07972
> median(data$Won_By)
[1] 8
```

Code:

```
# finding quantiles of won by
quantile(data$Won_By)
```

Output:

```
> quantile(data$Won_By)
  0%  25%  50%  75% 100%
   0    6    8   20  144
```

Code:

```
# checking NaN values (if cleanup is required)
sum(apply(data, 2, is.nan))
```

Output:

```
[1] 0
```

Code:

```
# check different host countries
levels(factor(data$Host_Country))
```

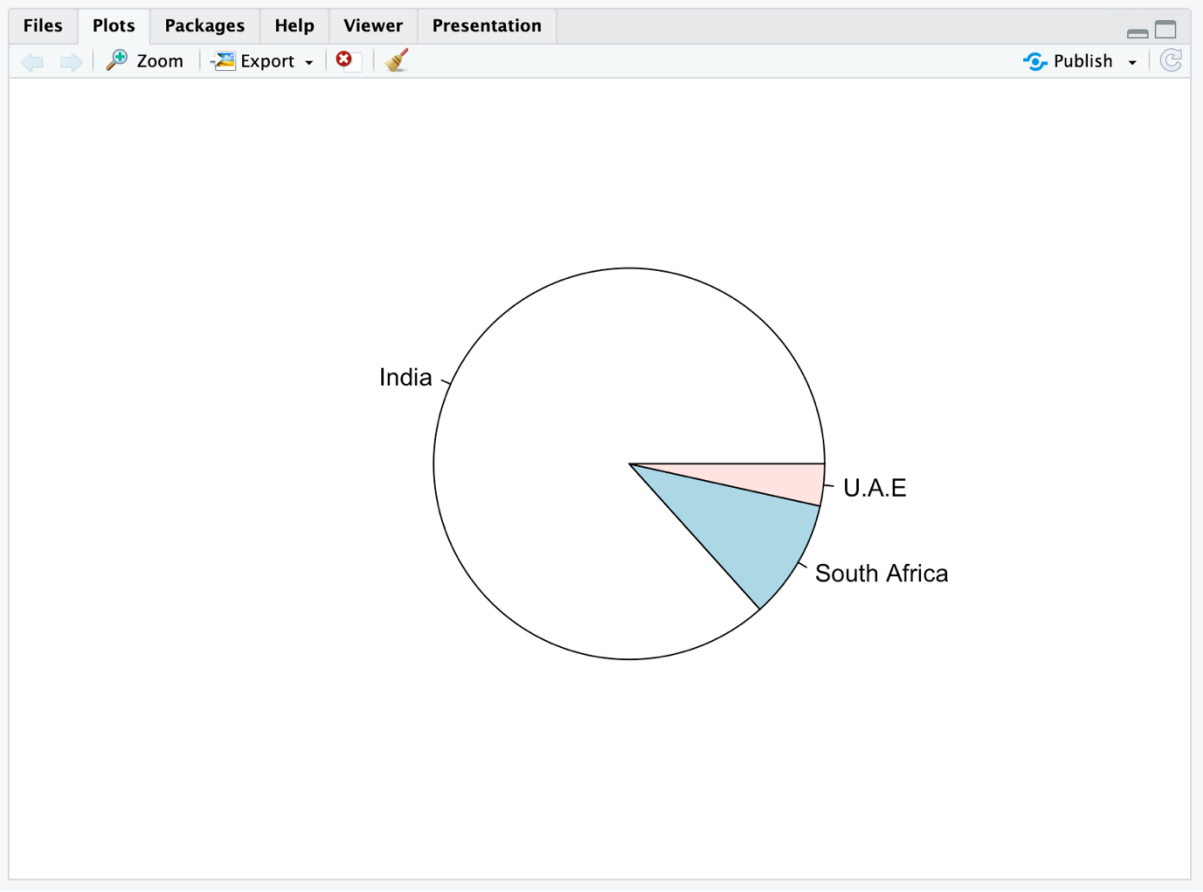
Output:

```
[1] "India"        "South Africa" "U.A.E"
```

Code:

```
# plotting by host country
country_counts = data %>%
  group_by(Host_Country) %>%
  summarise(count = length(Host_Country))
pie(country_counts$count, labels=country_counts$Host_Country)
```

Output:



Code:

```
# check different win conditions
levels(factor(data$Win_Type))
```

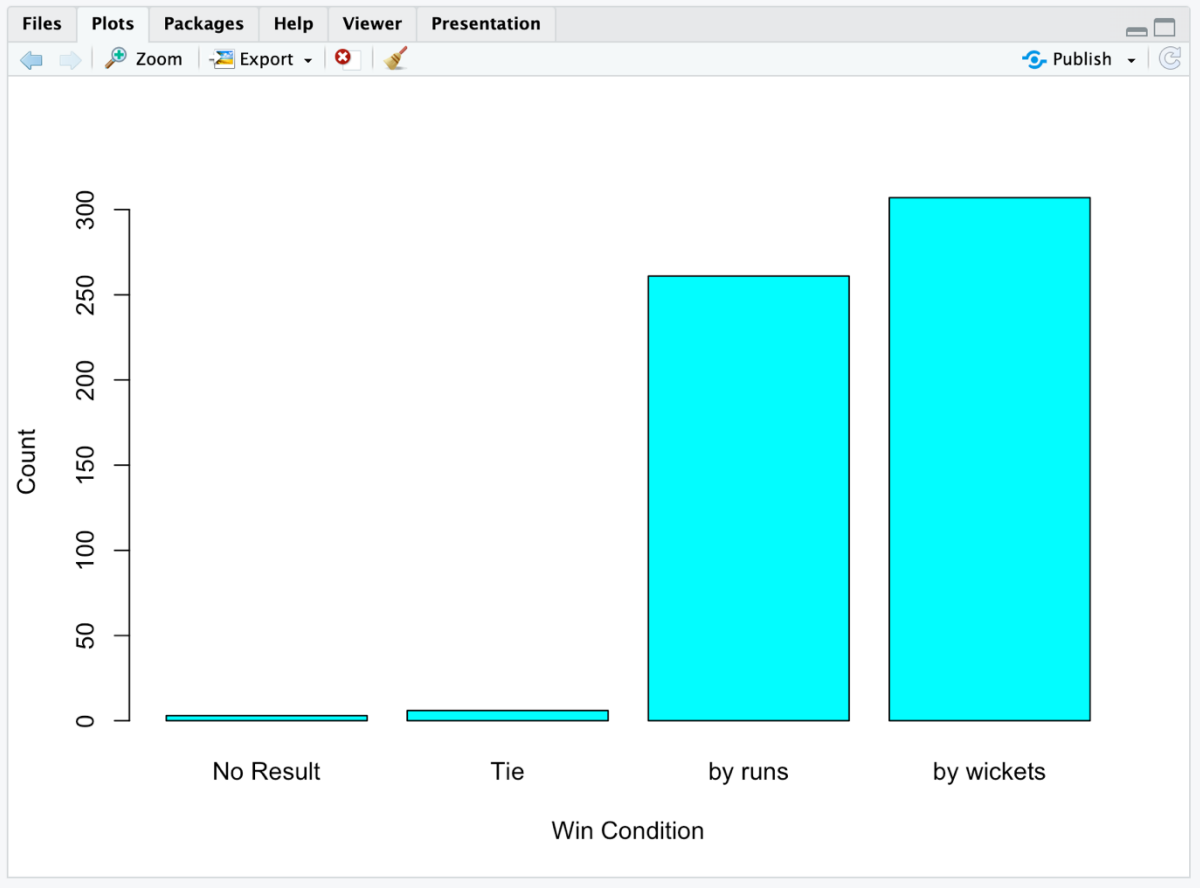
Output:

```
[1] "by runs"    "by wickets" "No Result"  "Tie"
```

Code:

```
# plotting by win condition
win_cond_count = data %>%
  group_by(Win_Type) %>%
  summarise(count = length(Win_Type))
barplot(win_cond_count$count, xlab="Win Condition", ylab="Count",
names.arg=win_cond_count$Win_Type, col="cyan")
```

Output:



Code:

```
# check different city names
levels(factor(data$City_Name))
```

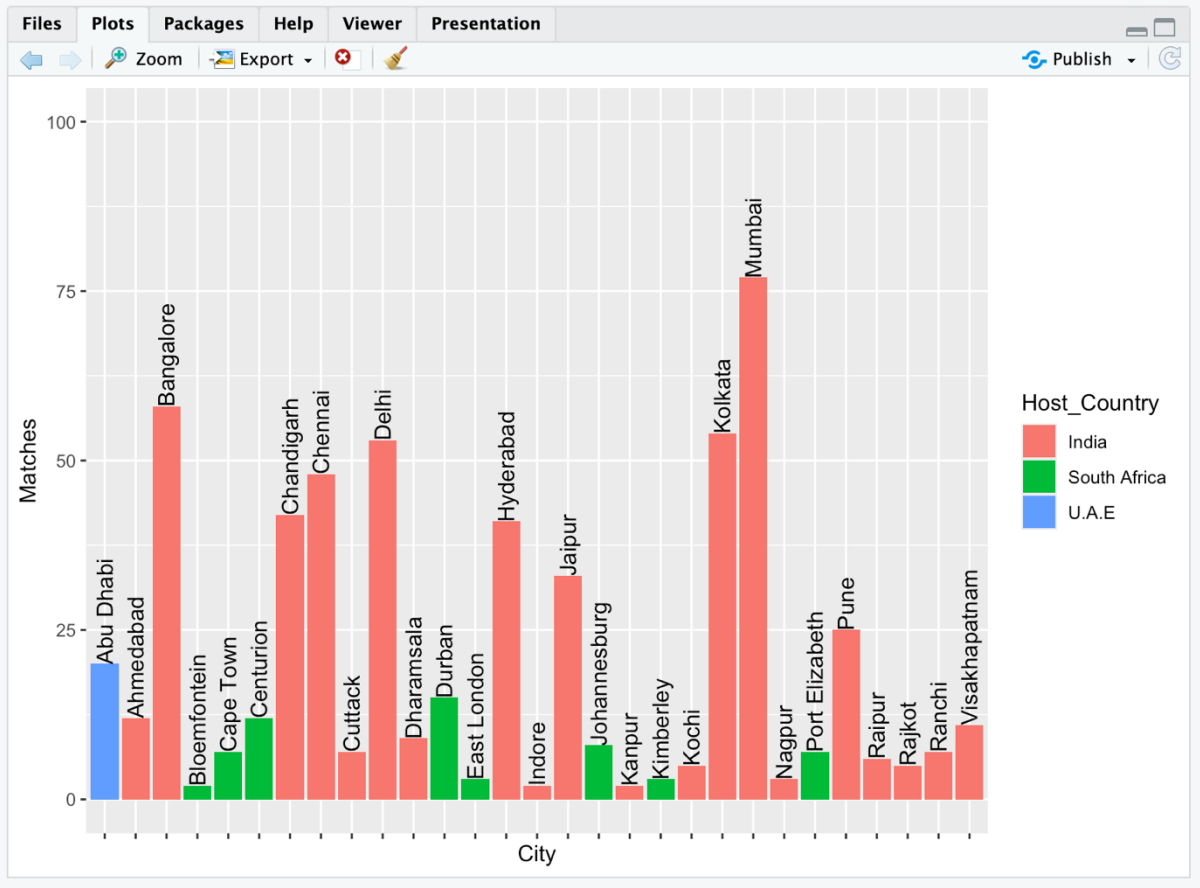
Output:

```
[1] "Abu Dhabi"      "Ahmedabad"      "Bangalore"      "Bloemfontein"   "Cape Town"
 [6] "Centurion"      "Chandigarh"     "Chennai"        "Cuttack"        "Delhi"
[11] "Dharamsala"     "Durban"         "East London"    "Hyderabad"      "Indore"
[16] "Jaipur"         "Johannesburg"   "Kanpur"         "Kimberley"      "Kochi"
[21] "Kolkata"        "Mumbai"         "Nagpur"         "Port Elizabeth" "Pune"
[26] "Raipur"         "Rajkot"         "Ranchi"         "Visakhapatnam"
```

Code:

```
# plotting by city name, color by host country
city = data %>%
  group_by(City_Name, Host_Country) %>%
  summarise(count = length(City_Name))
ggplot(city, aes(x=City_Name, y=count, fill=Host_Country)) +
  geom_bar(stat="identity") +
  geom_text(aes(label=City_Name), vjust=0.5, angle=90, hjust=0) +
scale_x_discrete(labels=NULL) +
  ylim(0, 100) +
  labs(x="City", y="Matches")
```

Output:



Code:

```
# casting match date column to date type
match_dates = data.frame(date=as.Date(data$Match_Date, format="%d-%b-%y"))
head(match_dates, 2)
```

Output:

```
        date
1 2008-04-18
2 2008-04-19
```

Code:

```
# finding the matches played per month
match_dates = match_dates %>%
  mutate(month=month(date)) %>%
  group_by(month) %>%
  summarise(count=length(month))
head(match_dates)
```

Output:

```
# A tibble: 4 × 2
  month count
  <dbl> <int>
1     3    29
2     4   261
3     5   285
```

```
4     6     2
```

Code:

```r
# plotting matches played by month
ggplot(match_dates, aes(x=month, y=count, fill=month)) +
  geom_bar(stat="identity") +
  geom_text(aes(label=month.name[month]), vjust=-0.5) +
  labs(x="Month", y="Matches Played")
```

Output: