FinalProject

Liam McFall, Cam Farrugia

5/4/2020

# What is the goal of the visualization?

In the round of 16 stage of the 2014 world cup, Did the home teams score more goals or did the away teams?

## Data Collection

How was the data collected? Is there any sampling bias that needs accounting for?

## Data Loading

worldcup<-read.csv("WorldCupMatches.csv", stringsAsFactors = FALSE)

## Data Manipulation

# Should any variables be factors/categorical?

The varibles are in the correct types already.

# Are there any types of variables

summary(worldcup)

## Year Datetime Stage Stadium   
## Min. :1930 Length:836 Length:836 Length:836   
## 1st Qu.:1970 Class :character Class :character Class :character   
## Median :1990 Mode :character Mode :character Mode :character   
## Mean :1985   
## 3rd Qu.:2002   
## Max. :2014   
##   
## City Home.Team.Name Home.Team.Goals Away.Team.Goals  
## Length:836 Length:836 Min. : 0.000 Min. :0.000   
## Class :character Class :character 1st Qu.: 1.000 1st Qu.:0.000   
## Mode :character Mode :character Median : 2.000 Median :1.000   
## Mean : 1.824 Mean :1.022   
## 3rd Qu.: 3.000 3rd Qu.:2.000   
## Max. :10.000 Max. :7.000   
##   
## Away.Team.Name Win.conditions Attendance   
## Length:836 Length:836 Min. : 2000   
## Class :character Class :character 1st Qu.: 29800   
## Mode :character Mode :character Median : 41000   
## Mean : 44859   
## 3rd Qu.: 61009   
## Max. :173850   
## NA's :1   
## Half.time.Home.Goals Half.time.Away.Goals Referee   
## Min. :0.0000 Min. :0.000 Length:836   
## 1st Qu.:0.0000 1st Qu.:0.000 Class :character   
## Median :0.0000 Median :0.000 Mode :character   
## Mean :0.7189 Mean :0.427   
## 3rd Qu.:1.0000 3rd Qu.:1.000   
## Max. :6.0000 Max. :5.000   
##   
## Assistant.1 Assistant.2 RoundID   
## Length:836 Length:836 Min. : 201   
## Class :character Class :character 1st Qu.: 262   
## Mode :character Mode :character Median : 337   
## Mean :10860927   
## 3rd Qu.: 249722   
## Max. :97410600   
##   
## MatchID Home.Team.Initials Away.Team.Initials  
## Min. : 25 Length:836 Length:836   
## 1st Qu.: 1184 Class :character Class :character   
## Median : 2114 Mode :character Mode :character   
## Mean : 56775774   
## 3rd Qu.: 43950047   
## Max. :300186515   
##

# Do NA’s need to be removed/omitted?

Removing NA’s would be dangerous in this example because only the Win Conditions column has data

# If removing NA’s do we lose data from other variables that we need?

worldcupNA<-na.omit(worldcup)  
summary(worldcupNA)

## Year Datetime Stage Stadium   
## Min. :1930 Length:835 Length:835 Length:835   
## 1st Qu.:1970 Class :character Class :character Class :character   
## Median :1990 Mode :character Mode :character Mode :character   
## Mean :1985   
## 3rd Qu.:2002   
## Max. :2014   
## City Home.Team.Name Home.Team.Goals Away.Team.Goals  
## Length:835 Length:835 Min. : 0.000 Min. :0.000   
## Class :character Class :character 1st Qu.: 1.000 1st Qu.:0.000   
## Mode :character Mode :character Median : 2.000 Median :1.000   
## Mean : 1.824 Mean :1.022   
## 3rd Qu.: 3.000 3rd Qu.:2.000   
## Max. :10.000 Max. :7.000   
## Away.Team.Name Win.conditions Attendance   
## Length:835 Length:835 Min. : 2000   
## Class :character Class :character 1st Qu.: 29800   
## Mode :character Mode :character Median : 41000   
## Mean : 44859   
## 3rd Qu.: 61009   
## Max. :173850   
## Half.time.Home.Goals Half.time.Away.Goals Referee   
## Min. :0.0000 Min. :0.0000 Length:835   
## 1st Qu.:0.0000 1st Qu.:0.0000 Class :character   
## Median :0.0000 Median :0.0000 Mode :character   
## Mean :0.7198 Mean :0.4275   
## 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :6.0000 Max. :5.0000   
## Assistant.1 Assistant.2 RoundID   
## Length:835 Length:835 Min. : 201   
## Class :character Class :character 1st Qu.: 262   
## Mode :character Mode :character Median : 337   
## Mean :10873628   
## 3rd Qu.: 249722   
## Max. :97410600   
## MatchID Home.Team.Initials Away.Team.Initials  
## Min. : 25 Length:835 Length:835   
## 1st Qu.: 1182 Class :character Class :character   
## Median : 2098 Mode :character Mode :character   
## Mean : 56484265   
## 3rd Qu.: 43950046   
## Max. :300186515

#We would only lose one data entry which isn't too bad, but it doesn't seem necessary to do so.

# Is imputation a better alternative in this case to account for NA’s?

The only reason Win Condition would have data in it is if the game went to overtime, so the imputation would be hard to code for this example because the data is a string in that column.

# Does data need to be aggregated prior to visualization?

Not for this example

# Does data need to be filtered?

To answer our question, yes!

world2=worldcup %>%   
 filter(Year == "2014", Stage=="Round of 16")

# Do multiple data sets need to be combined?

Not for this example

# Manipulation libraries  
library(dplyr)  
library(lubridate)

##   
## Attaching package: 'lubridate'

## The following object is masked from 'package:cowplot':  
##   
## stamp

## The following object is masked from 'package:base':  
##   
## date

# Check for NA  
#sum(is.na(data))  
# Remove NA's if you don't lose important data  
#data <- na.omit(data)  
  
# Check summary of data to see the current format each variable is in  
#summary(data)  
#str(data)  
  
# Change any data to the needed new format  
#data$date <- date(data$date)  
#data$exColumn <- factor(data$exColumn)  
  
# Other data manipulations  
# Dplyr library has tons of useful data manipulation tools and functions  
# https://rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf  
#data <- data %>%   
# filter(col2 = "string", col3 < num) %>%   
# left\_join(data2)

## Type of visualization

# What type of data are we using?

Strings and Integers

# Are we looking at trends or timeseries data?

Trends for this example

# Are we showing a relationship between variables?

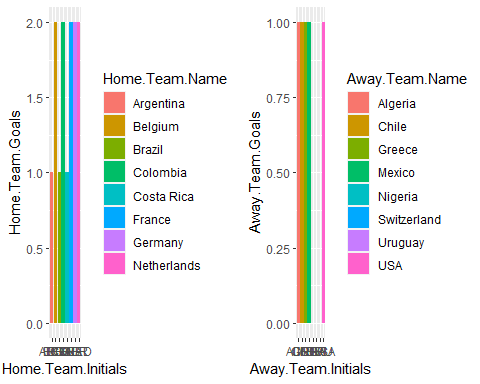
Yes, between home/away teams and goals

# Are we trying to show geographic data?

Not for this example

# R has base plotting capabilities, but ggplot has a lot more options  
library(ggplot2)  
  
# ggplot objects for a variety of types of charts  
#geom\_point()  
#geom\_line()  
#geom\_histogram()  
#geom\_tile()  
#geom\_bar()  
# add coords\_polar to geom\_bar to make it a pie chart  
#geom\_density()  
#geom\_boxplot()

home<-ggplot(world2, aes(x=Home.Team.Initials, y=Home.Team.Goals, fill=Home.Team.Name))+geom\_col()  
  
away<-ggplot(world2, aes(x=Away.Team.Initials, y=Away.Team.Goals, fill=Away.Team.Name))+geom\_col()  
plot\_grid(home, away)



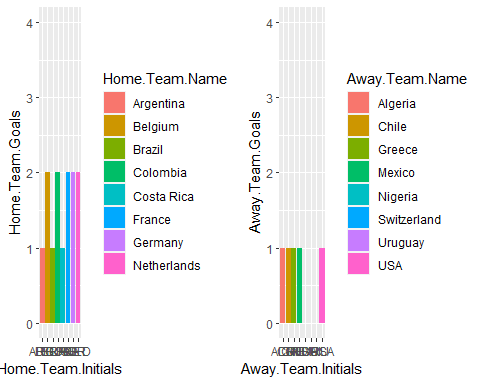
There is a lot wrong with this picture though. The graphs ranges are different. The home teams graph scales up to 2 whereas the aways team only scales to 1. The color coding leads to assumptions that teams like Argentia played Algeria and so forth, but that was not the actual case. You can call this bias, or you can call this poor design choices, regardless one should not settle for this and look into how to make it better.

## Scale and Axis choices

# Should any of the scales be nonlinear?

Not for this one, but they do need to be of equal length.

home<-ggplot(world2, aes(x=Home.Team.Initials, y=Home.Team.Goals, fill=Home.Team.Name))+geom\_col()+coord\_cartesian(ylim = c(0, 4))  
away<-ggplot(world2, aes(x=Away.Team.Initials, y=Away.Team.Goals, fill=Away.Team.Name))+geom\_col()+coord\_cartesian(ylim = c(0, 4))  
plot\_grid(home, away)

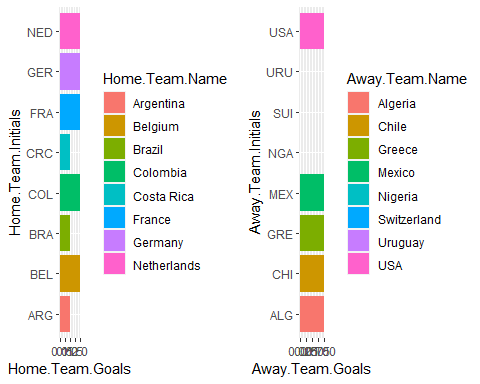


# Should scales be reversed?

home<-ggplot(world2, aes(x=Home.Team.Initials, y=Home.Team.Goals, fill=Home.Team.Name))+geom\_col()+coord\_flip()  
away<-ggplot(world2, aes(x=Away.Team.Initials, y=Away.Team.Goals, fill=Away.Team.Name))+geom\_col()+coord\_cartesian(xlim = c(0, 4))+coord\_flip()

## Coordinate system already present. Adding new coordinate system, which will replace the existing one.

plot\_grid(home, away)



Probably not for this one, I would say the picture doesn’t look as good. The only thing more visible are the team initals.

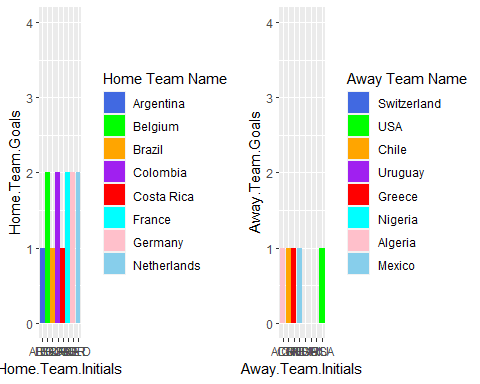
# Are tick marks needed?

Not for this example

# What other theme choices should be considered?

The color scheme to match the teams that played each other.

home<-ggplot(world2, aes(x=Home.Team.Initials, y=Home.Team.Goals, fill=Home.Team.Name))+geom\_col()+coord\_cartesian(ylim = c(0, 4))+scale\_fill\_manual(  
 values=c("royalblue", "green", "orange", "purple","red", "cyan","pink", "skyblue"), limits=c("Argentina", "Belgium", "Brazil","Colombia", "Costa Rica", "France", "Germany","Netherlands"), breaks=c("Argentina", "Belgium", "Brazil","Colombia", "Costa Rica", "France", "Germany","Netherlands"), name="Home Team Name",labels=c("Argentina", "Belgium", "Brazil","Colombia", "Costa Rica", "France", "Germany","Netherlands")  
)  
  
away<-ggplot(world2, aes(x=Away.Team.Initials, y=Away.Team.Goals, fill=Away.Team.Name))+geom\_col()+coord\_cartesian(ylim = c(0, 4))+scale\_fill\_manual(  
 values=c("royalblue", "green", "orange", "purple","red", "cyan","pink", "skyblue"), limits=c("Switzerland", "USA", "Chile","Uruguay", "Greece", "Nigeria", "Algeria","Mexico"), breaks=c("Switzerland", "USA", "Chile","Uruguay", "Greece", "Nigeria", "Algeria","Mexico"), name="Away Team Name",labels=c("Switzerland", "USA", "Chile","Uruguay", "Greece", "Nigeria", "Algeria","Mexico")  
)  
plot\_grid(home, away)

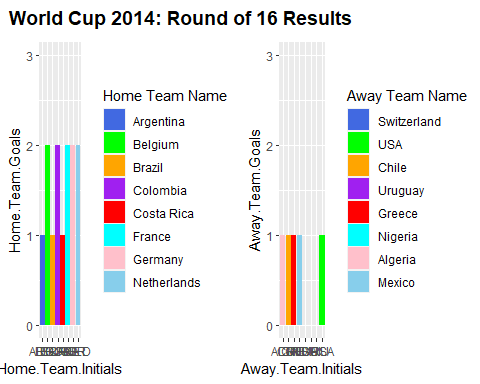


Now the colors from each graph correspond and if you match up the colors, those are the teams that played each other.

# ggplot has a bunch of x and y scale options to choose from  
#scale\_x\_log10()  
#scale\_y\_log10()  
#scale\_x\_binned()  
#scale\_x\_reverse()  
#scale\_y\_binned()  
#scale\_y\_reverse()  
#coord\_flip()  
#coord\_polar()  
#expand\_limits()

## Final Visualization

home<-ggplot(world2, aes(x=Home.Team.Initials, y=Home.Team.Goals, fill=Home.Team.Name))+geom\_col()+coord\_cartesian(ylim = c(0, 3))+scale\_fill\_manual(  
 values=c("royalblue", "green", "orange", "purple","red", "cyan","pink", "skyblue"), limits=c("Argentina", "Belgium", "Brazil","Colombia", "Costa Rica", "France", "Germany","Netherlands"), breaks=c("Argentina", "Belgium", "Brazil","Colombia", "Costa Rica", "France", "Germany","Netherlands"), name="Home Team Name",labels=c("Argentina", "Belgium", "Brazil","Colombia", "Costa Rica", "France", "Germany","Netherlands")  
)  
  
away<-ggplot(world2, aes(x=Away.Team.Initials, y=Away.Team.Goals, fill=Away.Team.Name))+geom\_col()+coord\_cartesian(ylim = c(0, 3))+scale\_fill\_manual(  
 values=c("royalblue", "green", "orange", "purple","red", "cyan","pink", "skyblue"), limits=c("Switzerland", "USA", "Chile","Uruguay", "Greece", "Nigeria", "Algeria","Mexico"), breaks=c("Switzerland", "USA", "Chile","Uruguay", "Greece", "Nigeria", "Algeria","Mexico"), name="Away Team Name",labels=c("Switzerland", "USA", "Chile","Uruguay", "Greece", "Nigeria", "Algeria","Mexico")  
)  
plot\_row<-plot\_grid(home, away)  
  
title <- ggdraw() +   
 draw\_label(  
 "World Cup 2014: Round of 16 Results",  
 fontface = 'bold',  
 x = 0,  
 hjust = 0  
 ) +  
 theme(  
 # add margin on the left of the drawing canvas,  
 # so title is aligned with left edge of first plot  
 plot.margin = margin(0, 0, 0, 7)  
 )  
plot\_grid(  
 title, plot\_row,  
 ncol = 1,  
 # rel\_heights values control vertical title margins  
 rel\_heights = c(0.1, 1)  
)



There are always many ways to do things or answer questions. This way just one example showing a process of trying to make the visualization as clear as possible to the reader.