Hypot2: Popularity by Country

Hypothesis: AOE2 is most popularly played in the US

```r
library(ggplot2)
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
```

```r
library(tidyr)
aoe2 <- read.csv("../Data/aoe2_leaderboard_sample.csv")
```

Let's first start off by seeing the countries with the most amount of players

```r
aoe2 %>%
  group_by(country) %>%
  summarise(count = n()) %>%
  subset(count > 100) %>%
  arrange(desc(count))
```

```
## # A tibble: 18 x 2
##    country count
##    <chr>   <int>
##  1 DE        888
##  2 US        711
##  3 FR        634
##  4 <NA>      417
##  5 AR        403
##  6 GB        284
##  7 ES        245
##  8 TR        245
##  9 AU        216
## 10 BR        190
## 11 CA        177
## 12 NL        148
## 13 CL        143
## 14 MX        141
## 15 CN        132
## 16 CH        116
## 17 TW        105
## 18 AT        101
```

The most significant gap in the data seems to be between Argentina (AR) and Great Britain (GB) where the count falls from 403 to 284. Past Great Britain all the countries lie in the 100 to 300 range.

```
pop_countries_count <- aoe2 %>%
  group_by(country) %>%
  summarise(count = n()) %>%
  subset(count > 400) %>%
  arrange(desc(count))

pop_countries_count
```

```
## # A tibble: 5 x 2
##   country count
##   <chr>   <int>
## 1 DE        888
## 2 US        711
## 3 FR        634
## 4 <NA>      417
## 5 AR        403
```

Let's select these countries in our popular countries vector and define the dataset "pop_countries" as players only from these countries.

```
pop_countries_v <- pop_countries_count$country
pop_countries_v <- pop_countries_v[!is.na(pop_countries_v)]
pop_countries_v
```

```
## [1] "DE" "US" "FR" "AR"
```

```
pop_countries <- aoe2[aoe2$country %in% pop_countries_v, ]
```

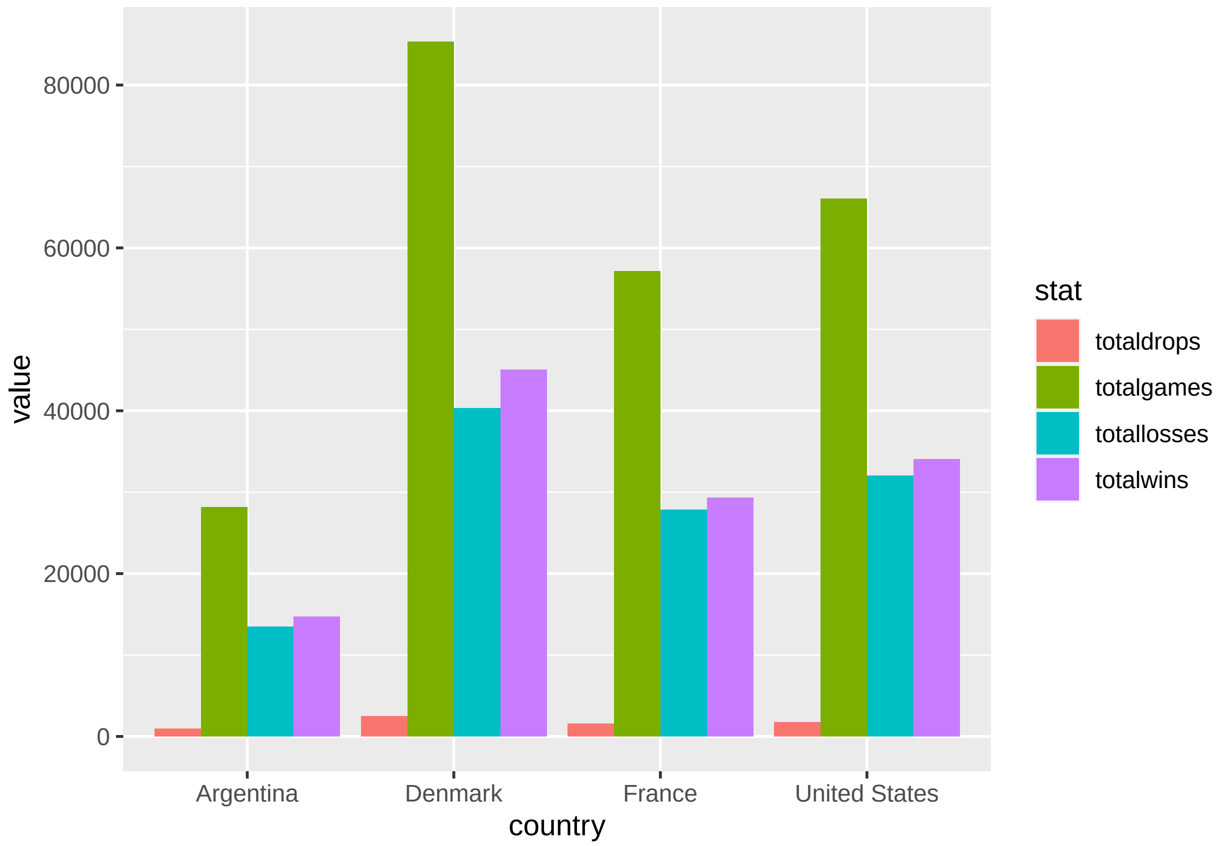
In order to test popularity, lets summarise each country's data at each stat

```
pop_countries_totals <- pop_countries %>%
  group_by(country) %>%
  summarise(totalgames = sum(games), totalwins = sum(wins), totallosses = sum(losses), totaldrops = sum
  arrange(desc(totalgames))

pop_countries_totals
```

```
## # A tibble: 4 x 5
##   country totalgames totalwins totallosses totaldrops
##   <chr>        <int>     <int>       <int>      <int>
## 1 DE           85329     45049       40280       2490
## 2 US           66039     34033       32006       1746
## 3 FR           57159     29295       27864       1581
## 4 AR           28167     14705       13462        922
```

Let's graph it to see how it looks.

```
test <- gather(pop_countries_totals, country)
colnames(test) <- c("stat", "value")
test$country <- rep(c("Denmark", "United States", "France", "Argentina"), 4)

ggplot(test, aes(fill = stat, y = value, x = country)) +
  geom_bar(position="dodge", stat="identity")
```



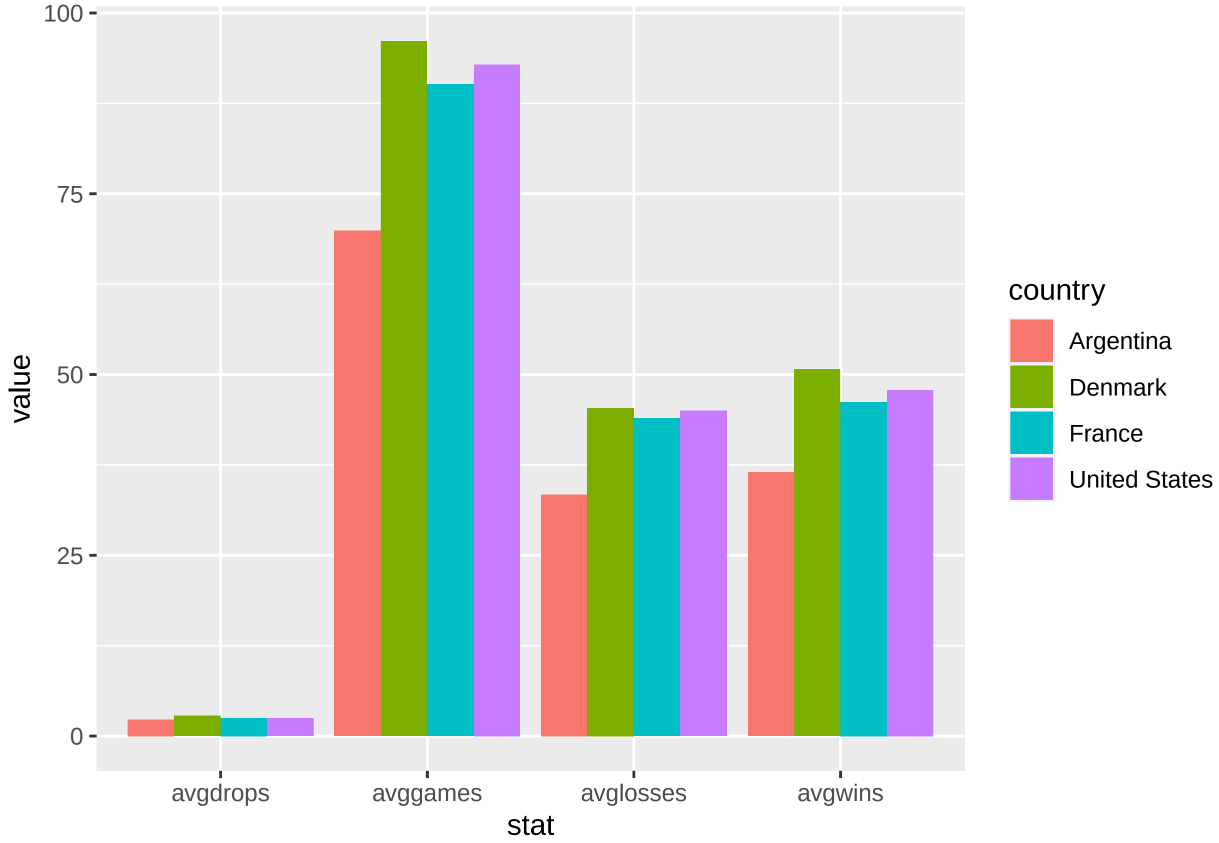
While this graph does show Denmark's dominance in total games, wins, losses, and drops, the data has a lot more to show, and looking to average games might show more about the game's culture/competitiveness in a country.

```
pop_countries_averages <- pop_countries %>%
  group_by(country) %>%
  summarise(avggames = mean(games), avgwins = mean(wins), avglosses = mean(losses), avgdrops = mean(drop
  arrange(desc(avgwins))

pop_countries_averages
```

```
## # A tibble: 4 x 5
##   country avggames avgwins avglosses avgdrops
##   <chr>      <dbl>   <dbl>     <dbl>    <dbl>
## 1 DE          96.1    50.7      45.4     2.80
## 2 US          92.9    47.9      45.0     2.46
## 3 FR          90.2     46.2      43.9     2.49
## 4 AR          69.9     36.5      33.4     2.29
```

Now lets graph it.

```
pop_countries_averages_long <- gather(pop_countries_averages, country)
colnames(pop_countries_averages_long) <- c("stat", "value")
pop_countries_averages_long$country <- rep(c("Denmark", "United States", "France", "Argentina"), 4)

ggplot(pop_countries_averages_long, aes(fill = country, y = value, x = stat)) +
  geom_bar(position="dodge", stat="identity")
```

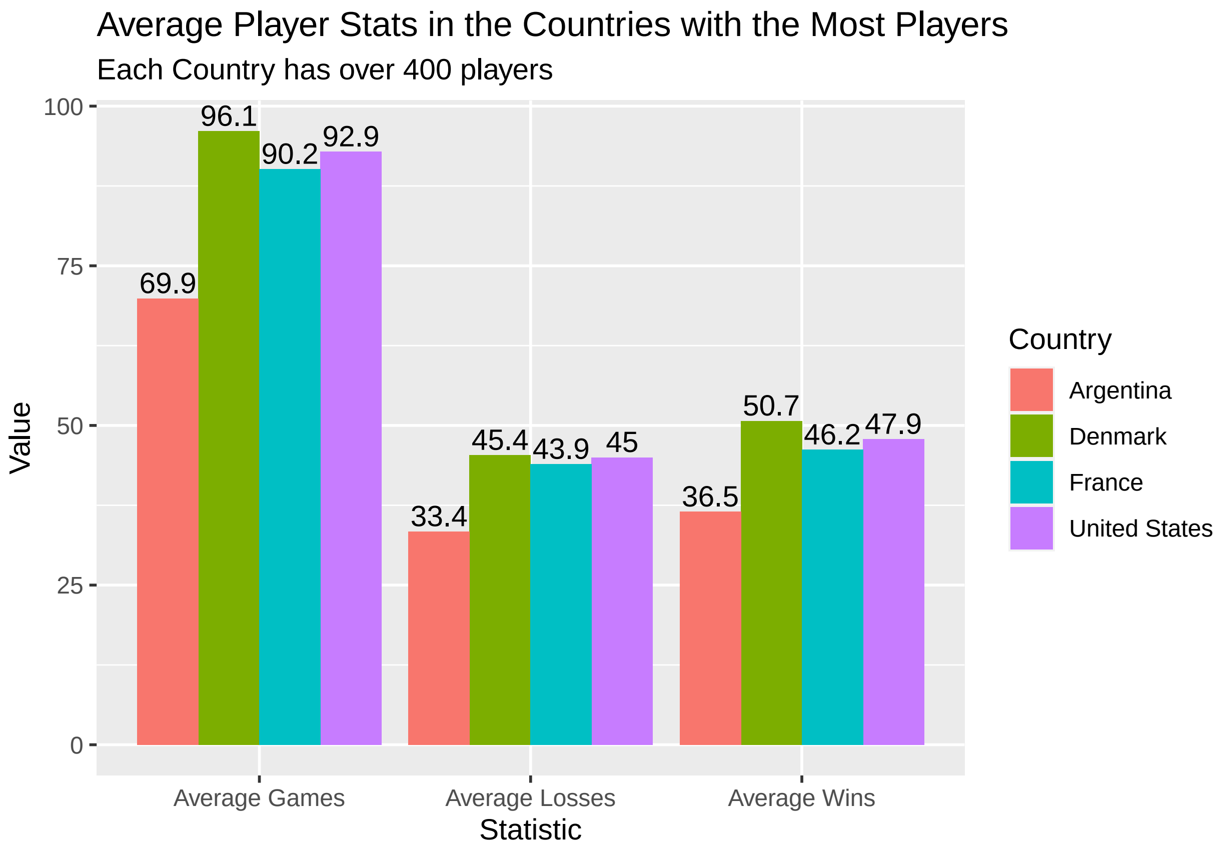


The average drops statistic seems to be pretty insignificant here, so lets take that out and finalize our graph!

```
pop_countries_averages$avgdrops <- NULL
pop_countries_averages
```

```
## # A tibble: 4 x 4
##   country avggames avgwins avglosses
##   <chr>      <dbl>   <dbl>     <dbl>
## 1 DE          96.1    50.7      45.4
## 2 US          92.9    47.9      45.0
## 3 FR          90.2    46.2      43.9
## 4 AR          69.9    36.5      33.4
```

```r
pop_countries_averages_long <- gather(pop_countries_averages, country)
colnames(pop_countries_averages_long) <- c("stat", "value")
pop_countries_averages_long$country <- rep(c("Denmark", "United States", "France", "Argentina"), 3)
pop_countries_averages_long$stat <- c(rep("Average Games", 4), rep("Average Wins", 4),
                                      rep("Average Losses", 4))


ggplot(pop_countries_averages_long, aes(fill = country, y = value, x = stat)) +
  geom_bar(position="dodge", stat="identity") +
  ggtitle("Average Player Stats in the Countries with the Most Players",
          subtitle = "Each Country has over 400 players") +
  xlab("Statistic") +
  ylab("Value") +
  labs(fill = "Country") +
  geom_text(aes(label=round(value, 1)), position=position_dodge(width=0.9), vjust=-0.25)
```



Interesting Insights: 1. Denmark still remains at the top with the best average player stats. 2. In terms of percentage, France and US would be tied for 2nd. 3. The average US player, who plays 3.2 less games on average then the average Denmark player, only loses an average of 0.4 less games, providing a generally significant difference for us to disprove our hypothesis

For fun, lets make the colors of the bars match the country they are representing.

```r
gdURL <- "http://www.stat.ubc.ca/~jenny/notOcto/STAT545A/examples/gapminder/data/gapminderCountryColors
countryColors <- read.delim(file = gdURL, as.is = 3) # protect color
str(countryColors)
```

```
## 'data.frame':    142 obs. of  3 variables:
##  $ continent: Factor w/ 5 levels "Africa","Americas",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ country  : Factor w/ 142 levels "Afghanistan",..: 95 39 43 28 118 121 127 69 86 3 ...
##  $ color    : chr  "#7F3B08" "#833D07" "#873F07" "#8B4107" ...
```

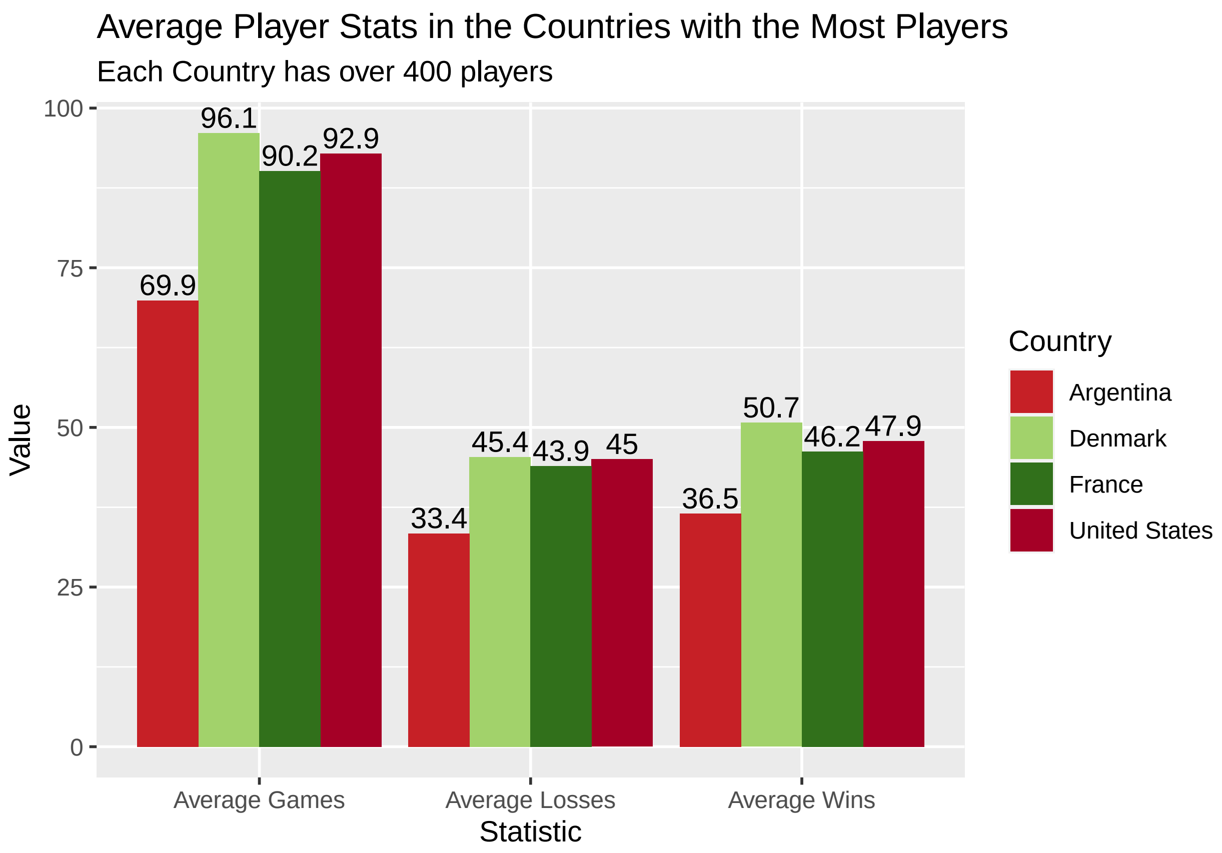
```r
head(countryColors)
```

```
##   continent          country   color
## 1    Africa          Nigeria #7F3B08
## 2    Africa            Egypt #833D07
## 3    Africa         Ethiopia #873F07
## 4    Africa Congo, Dem. Rep. #8B4107
## 5    Africa     South Africa #8F4407
## 6    Africa            Sudan #934607
```

```r
jColors <- countryColors$color
names(jColors) <- countryColors$country
head(jColors)
```

```
##          Nigeria            Egypt         Ethiopia Congo, Dem. Rep.
##        "#7F3B08"        "#833D07"        "#873F07"        "#8B4107"
##     South Africa            Sudan
##        "#8F4407"        "#934607"
```

```r
ggplot(pop_countries_averages_long, aes(fill = country, y = value, x = stat)) +
  geom_bar(position="dodge", stat="identity") +
  ggtitle("Average Player Stats in the Countries with the Most Players",
          subtitle = "Each Country has over 400 players") +
  xlab("Statistic") +
  ylab("Value") +
  labs(fill = "Country") +
  geom_text(aes(label=round(value, 1)), position=position_dodge(width=0.9), vjust=-0.25) +
  scale_fill_manual(values = jColors)
```



Unfortunately, the colors are not very distinct, so we will use the first full graphic.

Test Code / Scratch:

```
# aoe2 %>%
#   group_by(country) %>%
#   summarise(avggames = mean(games), avgwins = mean(wins), avglosses = mean(losses), avgdrops = mean(d
#   arrange(desc(avggames))
#
#
# aoe2 %>%
#   group_by(country) %>%
#   summarise(count = n()) %>%
#   arrange(desc(count))
```