Final Project Combined

Zev Rosen

12/12/2019

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(stringr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

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

```
library(knitr)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
library(DT)
library(cowplot)
```

```
##
## Attaching package: 'cowplot'
```

```
## The following object is masked from 'package:lubridate':
##
##     stamp
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(countrycode)
library(maps)
```

```
##
## Attaching package: 'maps'

## The following object is masked from 'package:purrr':
##
##     map
```

```
library(mapproj)
```

```
## Warning: package 'mapproj' was built under R version 4.0.5
```

Load in Datasets

```
players = read_csv("players.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   player_id = col_double(),
##   first_name = col_character(),
##   last_name = col_character(),
##   hand = col_character(),
##   birth_date = col_double(),
##   country_code = col_character()
## )
```

```
rankings = read_csv("rankings.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   ranking_date = col_double(),
```

```
##   ranking = col_double(),
##   player_id = col_double(),
##   ranking_points = col_double(),
##   tours = col_double()
## )
```

```
matches = read_csv("matches.csv")
```

```
## Warning: Missing column names filled in: 'X33' [33]
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   loser_entry = col_character(),
##   loser_hand = col_character(),
##   loser_ioc = col_character(),
##   loser_name = col_character(),
##   minutes = col_logical(),
##   round = col_character(),
##   score = col_character(),
##   surface = col_character(),
##   tourney_id = col_character(),
##   tourney_level = col_character(),
##   tourney_name = col_character(),
##   winner_entry = col_character(),
##   winner_hand = col_character(),
##   winner_ioc = col_character(),
##   winner_name = col_character(),
##   X33 = col_logical()
## )
## i Use `spec()` for the full column specifications.
```

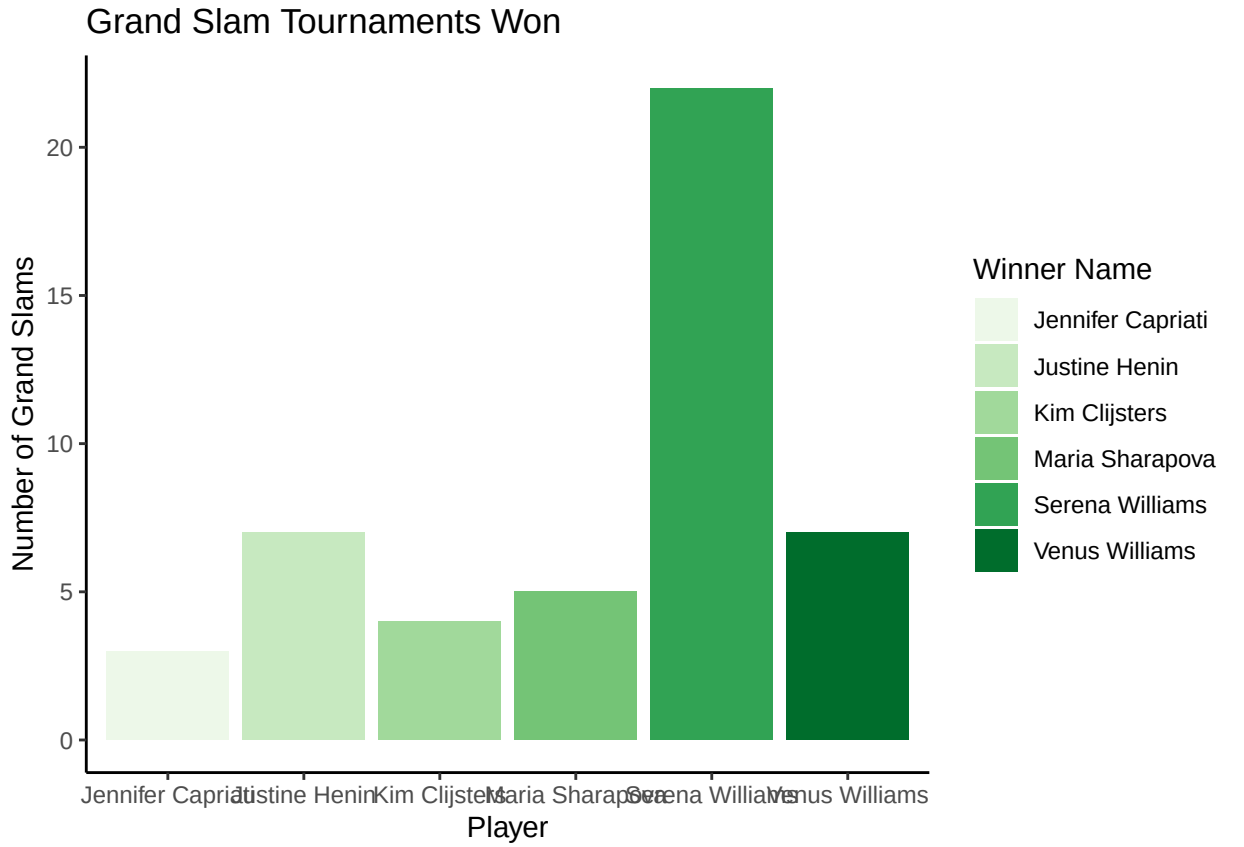
```
## Warning: 5008 parsing failures.
##   row          col           expected      actual          file
## 11907 loser_rank   a double           0           'matches.csv'
## 11907 tourney_date a double           Hard        'matches.csv'
## 11907 winner_ht    a double           R           'matches.csv'
## 11907 winner_rank  a double           Sania Mirza 'matches.csv'
## 11907 X33          1/0/T/F/TRUE/FALSE 2003        'matches.csv'
## ..... ............ .................. ........... ............
## See problems(...) for more details.
```

```
matches$tourney_name = str_replace(matches$tourney_name,"Us Open","US Open")
```

Grand Slam Leaders Barplot

This graph shows a barplot of the players with the most grand slam titles. Serena and Venus Williams have the most over this time period.

```
slam_matches <- filter(matches, tourney_level =="G")
slam_matches <- filter(slam_matches, round =="F")
summary_slam_matches <- summarize(group_by(slam_matches, winner_name), count=n())
summary_slam_matches <- rename(summary_slam_matches, number_of_grand_slams = count)
summary_slam_matches_2 <- subset(summary_slam_matches, (number_of_grand_slams >=3))
#barplot of most grand slams
grandslam_barplot <- ggplot(summary_slam_matches_2)+geom_bar(aes(x=winner_name, y=number_of_grand_slams
  labs(x="Player",y="Number of Grand Slams",
       title="Grand Slam Tournaments Won")+theme_classic()
grandslam_barplot + scale_fill_brewer(name= "Winner Name", palette = "Greens")
```




```
players_rankings = inner_join(players,rankings)
```

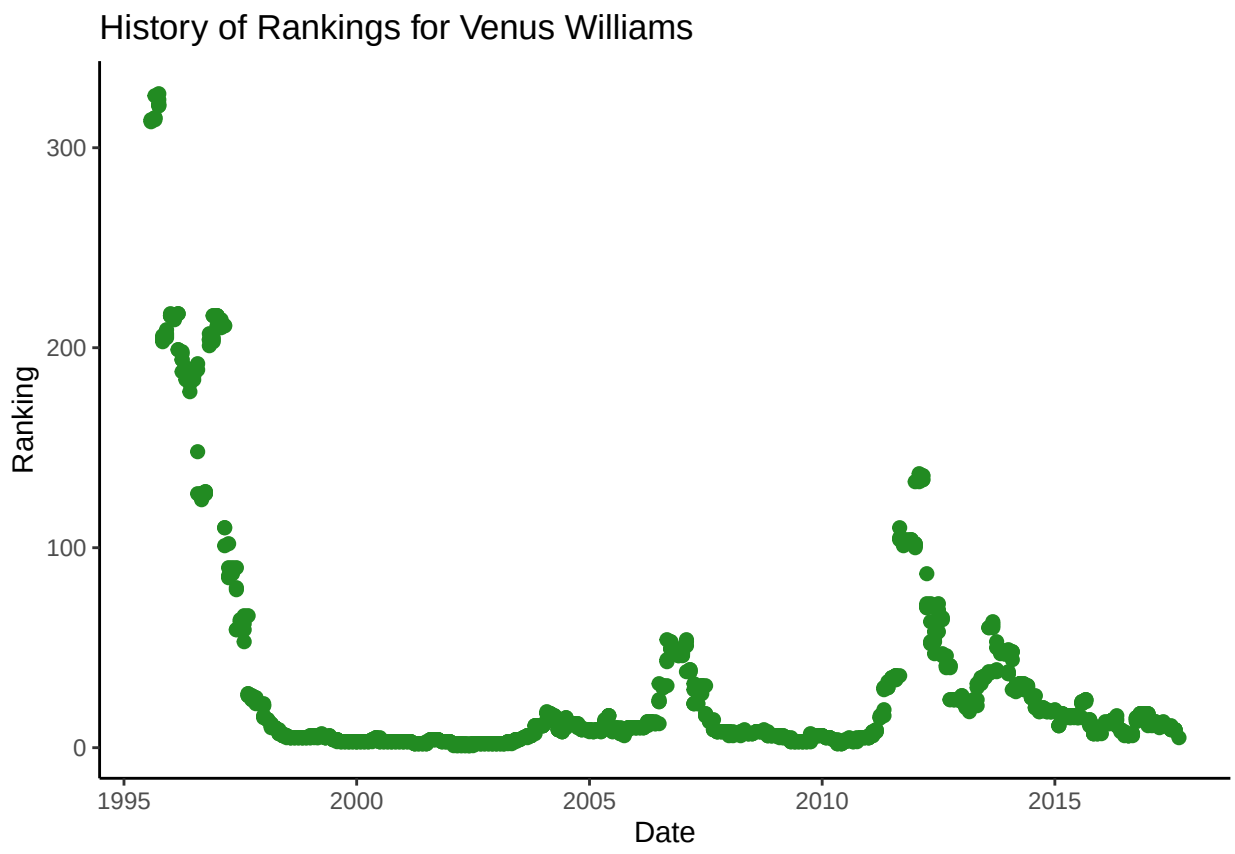
```
## Joining, by = "player_id"
```

```
players_rankings$year = str_sub(players_rankings$ranking_date,1,4)
players_rankings$month = str_sub(players_rankings$ranking_date,5,6)
```

Venus Williams Ranking Graph

This is a graph showing Venus Williams' rankings from her time coming on tour until 2017. We see Williams get to the top 10 in the rankings in the late 1990's. Aside from a downturn in 2012 and 2013, she has stayed near the top of the rankings since.

```
venus_rankings <- filter(players_rankings, first_name=="Venus", last_name =="Williams")
venus_rankings <- mutate(venus_rankings, YearMonth=make_date(year = year, month = month))
#graph
venus <- ggplot(venus_rankings,aes(x=YearMonth,y=ranking, group=1))+geom_point(size=2, color="forestgre
venus
```

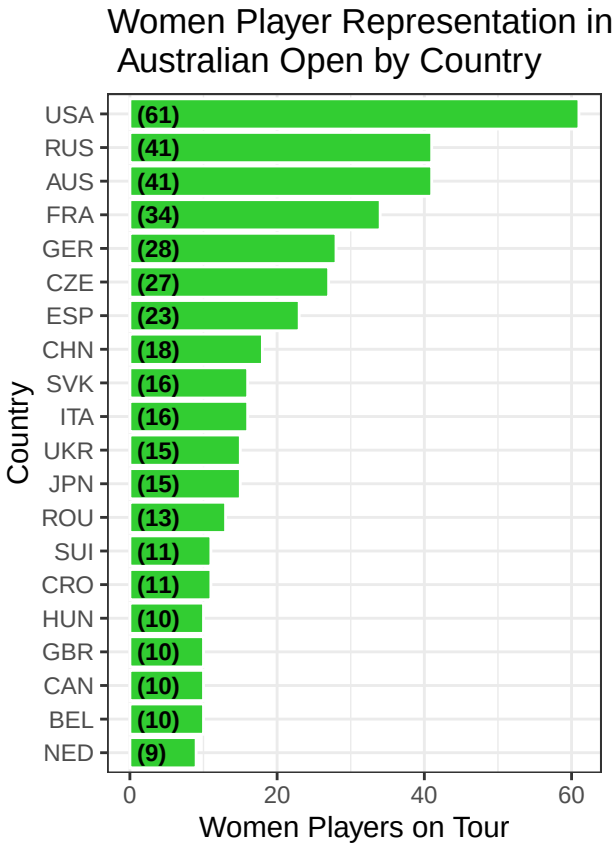


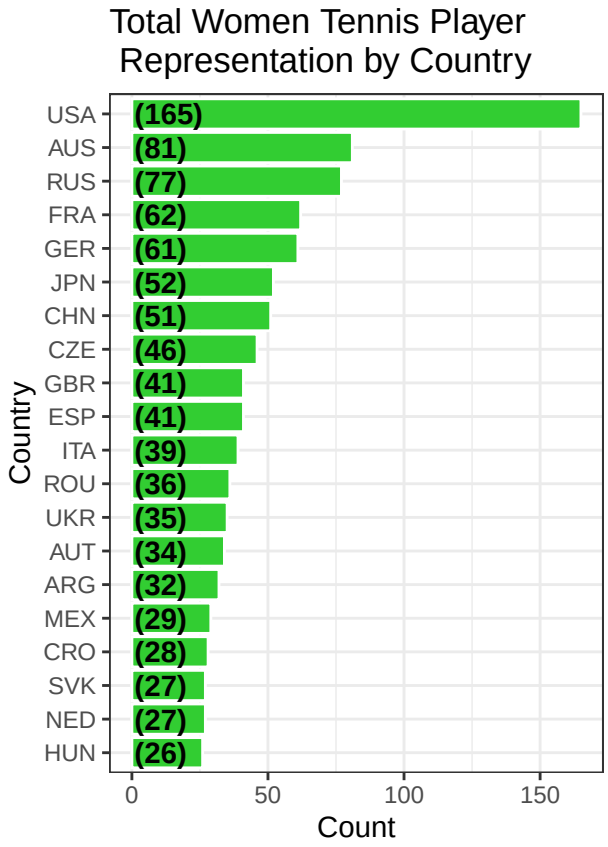
Representation by Country in Australian Open

These barplots depict the numbers of players who have represented their country from 2000-2017. The original graph, recreated on the right, displays the total number of players by country. The graph on the left displays the total number of players by country in the Australian Open. We compared the two to see whether demographics change depending on the tournament: for example, if Australian players are more represented in the Australian Open than in general.

```
matches_australia <- filter(matches, tourney_name == "Australian Open")
matches_australia_winner <- select(matches_australia, winner_name, winner_ioc)
matches_australia_winner <- rename(matches_australia_winner, name=winner_name, ioc=winner_ioc)
matches_australia_loser <- select(matches_australia, loser_name, loser_ioc)
matches_australia_loser <- rename(matches_australia_loser, name=loser_name, ioc=loser_ioc)
matches_country <- unique(rbind(matches_australia_winner, matches_australia_loser))
matches_country <- group_by(matches_country, ioc)
matches_country <- summarise(matches_country, Count=n())
matches_country <- arrange(matches_country, desc(Count))
matches_country <- ungroup(matches_country)
```

```
matches_country <- mutate(matches_country, ioc=reorder(ioc, Count))
matches_country <- head(matches_country, 20)

 australia <- ggplot(matches_country, aes(x = ioc,y = Count)) +
 geom_bar(stat='identity',colour="white", fill = "limegreen") +
 geom_text(aes(x = ioc, y = 1, label = paste0("(",Count,")",sep="")),
 hjust=0, vjust=.5, size = 3, colour = 'black',
 fontface = 'bold') +
 labs(x = 'Country',
 y = 'Women Players on Tour',
 title = 'Women Player Representation in \n Australian Open by Country') +
 coord_flip() +
 theme_bw()

matches_country_winner <- select(matches, winner_name, winner_ioc)
matches_country_winner <- rename(matches_country_winner, name=winner_name, ioc=winner_ioc)
matches_country_loser <- select(matches, loser_name, loser_ioc)
matches_country_loser <- rename(matches_country_loser, name=loser_name, ioc=loser_ioc)
matches_country <- unique(rbind(matches_country_winner, matches_country_loser))
matches_country <- group_by(matches_country, ioc)
matches_country <- summarise(matches_country, Count=n())
matches_country <- arrange(matches_country, desc(Count))
matches_country <- ungroup(matches_country)
matches_country <- mutate(matches_country, ioc=reorder(ioc, Count))
matches_country <- head(matches_country, 20)

 total <- ggplot(matches_country, aes(x = ioc,y = Count)) +
 geom_bar(stat='identity',colour="white", fill = "limegreen") +
 geom_text(aes(x = ioc, y = 1, label = paste0("(",Count,")",sep="")),
 hjust=0, vjust=.5, size = 4, colour = 'black',
 fontface = 'bold') +
 labs(x = 'Country',
 y = 'Count',
 title = 'Total Women Tennis Player \n Representation by Country') +
 coord_flip() +
 theme_bw()

grid.arrange(australia,total, nrow=1)
```





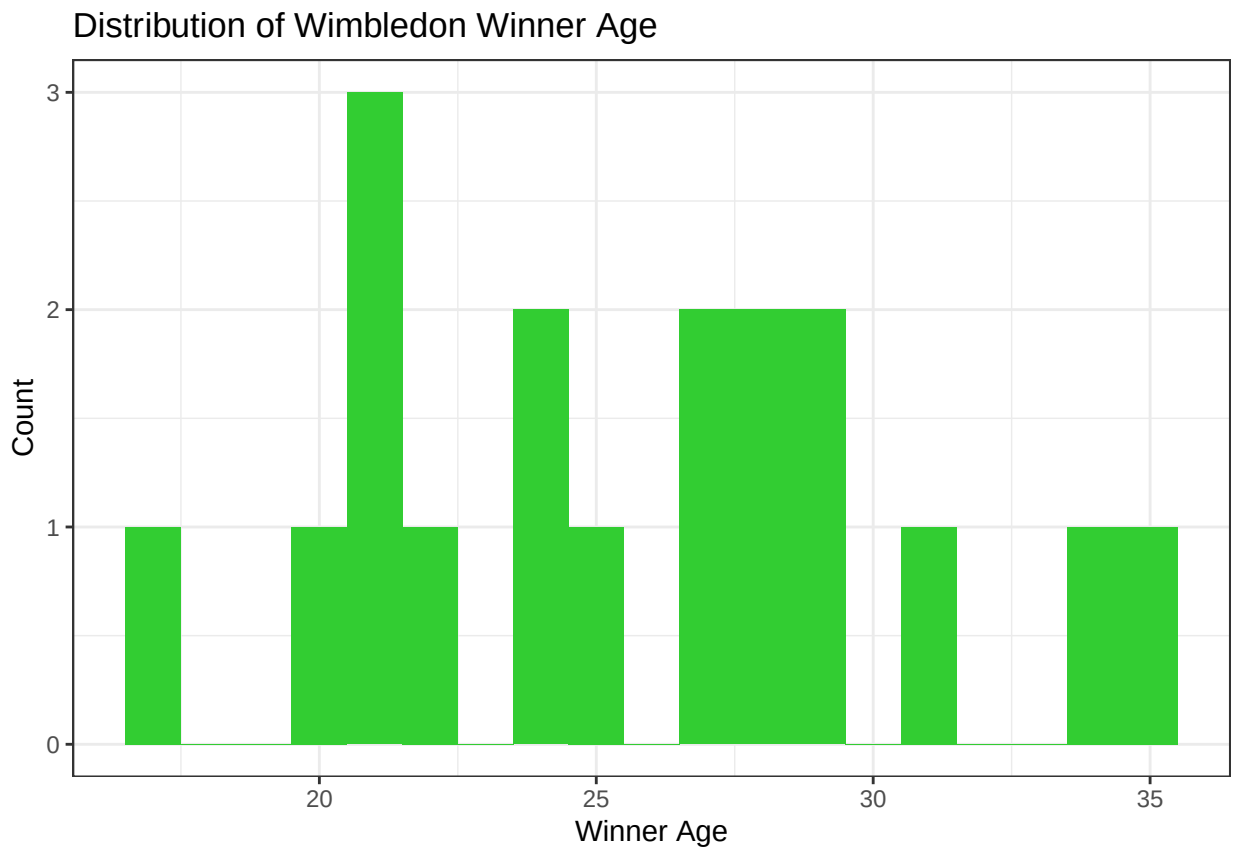
Distribution of Age Graph for Wimbledon

This barplot compares the age distributions of Wimbledon winners to losers. The graph indicates that there is greater variety in winners' ages than in losers' ages. Losers also tend to be younger, potentially indicating that more experienced players fare better.

```
wimbledon_matches <- filter(matches, tourney_name == "Wimbledon")
wimbledon_finals <- filter(wimbledon_matches, round == "F")
wimbledon_winners <- mutate(wimbledon_finals, agediff = loser_age - winner_age)

summary(wimbledon_winners$winner_age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.17   21.40   25.99   25.64   28.55   34.75
```

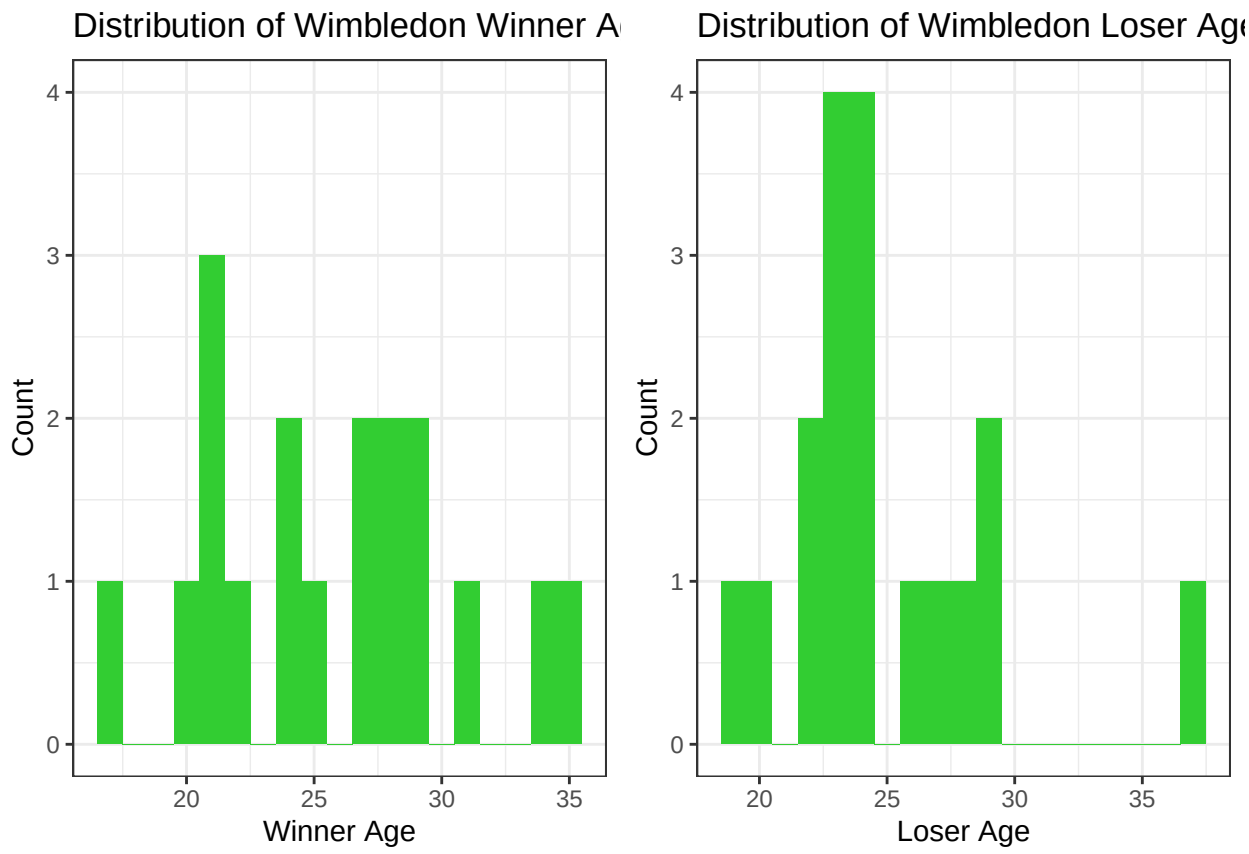
```
winners <- ggplot(wimbledon_winners, aes(x = winner_age)) +
 geom_histogram(binwidth = 1,fill = "limegreen") +
 labs(x= 'Winner Age',y = 'Count', title = paste("Distribution of", 'Wimbledon Winner Age ')) +
 theme_bw()
winners
```




```
wimbledon_losers <- mutate(wimbledon_finals, agediff = winner_age - loser_age)

summary(wimbledon_losers$loser_age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.07   22.73   23.90   24.83   26.50   37.04
```

```
losers <- ggplot(wimbledon_losers, aes(x = loser_age)) +
 geom_histogram(binwidth = 1,fill = "limegreen") +
 labs(x= 'Loser Age',y = 'Count', title = paste("Distribution of", 'Wimbledon Loser Age ')) +
 theme_bw()

grid.arrange(winners+ylim(0,4),losers,nrow=1)
```



Upsets

Here, we are finding how many upsets occured in grand slams and see that Wimbledon, due to unique seeding and surface, has the most upsets as a whole.

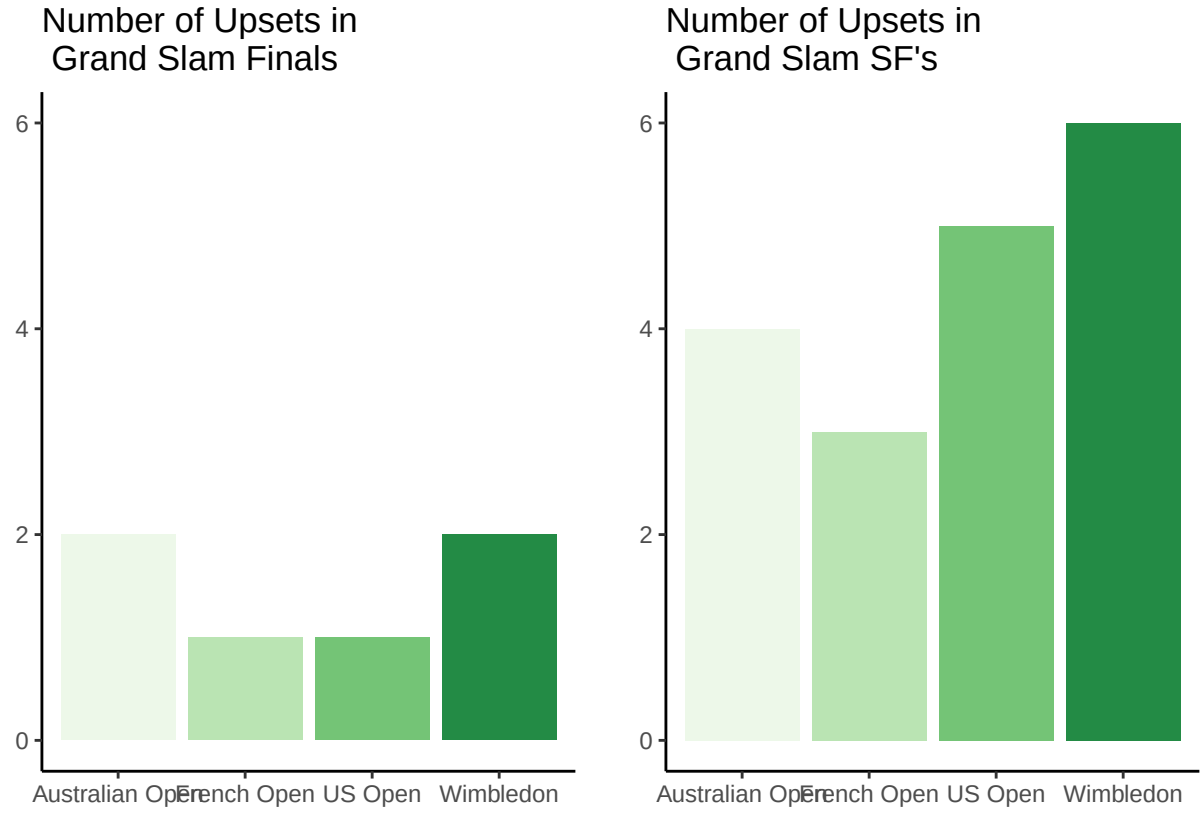
```
tournamentwinners <- mutate(matches, agediff = winner_age - loser_age,
                                rankingdiff = loser_rank - winner_rank)
ausopen = filter(tournamentwinners, tourney_name=="Australian Open")
french = filter(tournamentwinners, tourney_name=="French Open")
wimbledon = filter(tournamentwinners, tourney_name=="Wimbledon")
usopen = filter(tournamentwinners, tourney_name=="US Open")
grandslam= rbind(ausopen,french,wimbledon,usopen)
grandslamupsets = filter(grandslam, rankingdiff <= -10)

grandslamupsetsfinals = filter(grandslam, round == "F", rankingdiff <= -10)
```

Grand Slam Upsets Graph

We classify a game as an "upset" if the winner rank is greater than or equal to the loser rank by 10 slots. We plot the ranking difference between winners and losers for the grand slams in Finals and Semi-Finals.

```
g1 <- ggplot(grandslamupsetsfinals)+
  geom_bar(aes(x=tourney_name,fill=tourney_name,
               y = (..count..)))+
```

```
  scale_fill_brewer(palette="Greens",direction=1,
                    name="")+
  labs(x="",y="",
       title="Number of Upsets in \n Grand Slam Finals")+theme_classic()+theme(legend.position="none")+y

grandslamupsetsSF = filter(grandslam, round == "SF", rankingdiff <= -10)

g2 <- ggplot(grandslamupsetsSF)+
  geom_bar(aes(x=tourney_name,fill=tourney_name,
               y = (..count..)))+
  scale_fill_brewer(palette="Greens",direction=1,
                    name="")+
  labs(x="",y="",
       title="Number of Upsets in \n Grand Slam SF's")+theme_classic()+theme(legend.position="none")

grid.arrange(g1, g2, nrow=1)
```

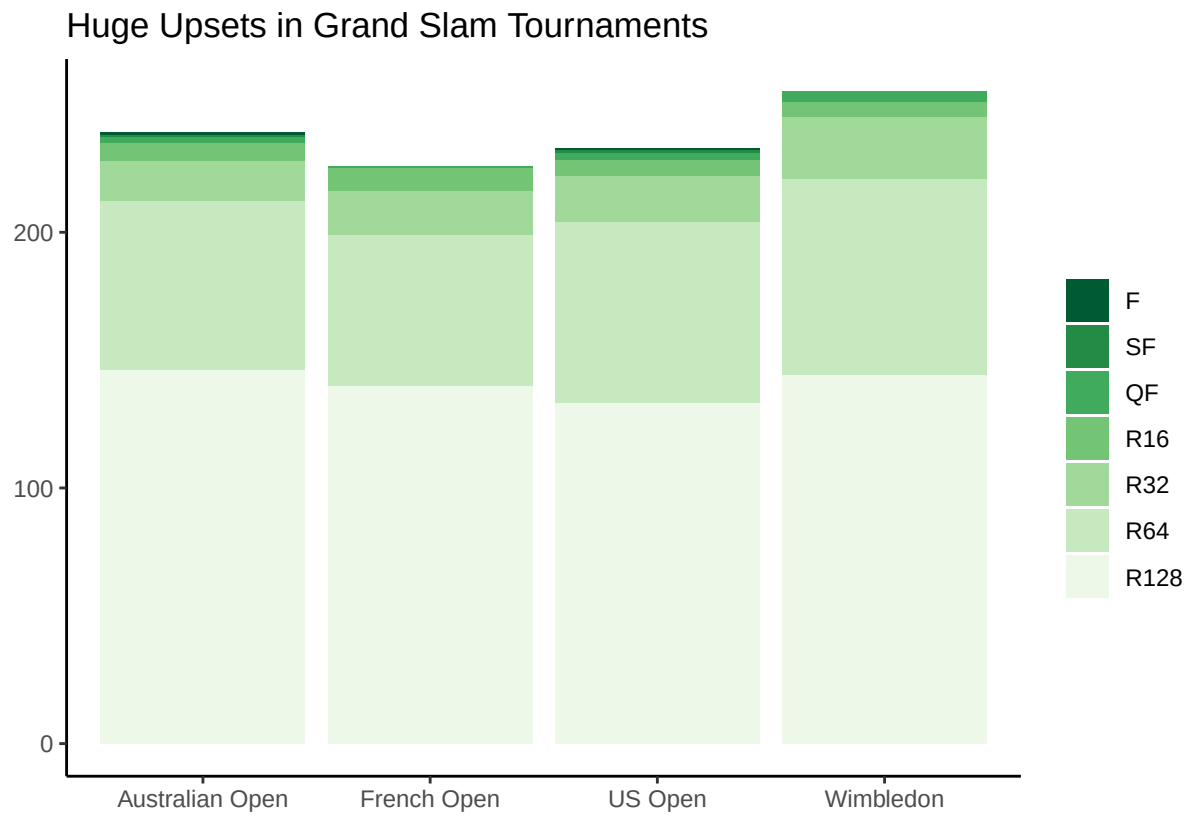


Grand Slam Big Upsets Graph

We are now calling upsets in Grand Slam tournaments of 50+ spots "huge upsets," and tracking them through every round.

```
grandslambigupsets <- filter(grandslam, rankingdiff <= -50)
grandslambigupsets <- mutate(grandslambigupsets, round = fct_relevel(round, "F", "SF", "QF", "R16","R32
```

```
ggplot(grandslambigupsets)+
  geom_bar(aes(x=tourney_name,fill=round,
               y = (..count..)))+
  scale_fill_brewer(palette="Greens",direction=-1,
                    name="")+
  labs(x="",y="",
       title="Huge Upsets in Grand Slam Tournaments")+theme_classic()
```



Chloropleth of Female Tennis Players by Country

This a chloropleth which displays female grand slam players by country over the last 17 years. A darker shade means more players from each country have been represented.

```
##CHLOROPLETH OF GRAND SLAM PLAYERS BY COUNTRY
#set match winners and losers variable
slam_matches <- filter (matches, tourney_level == "G")
winner_countries<- select(slam_matches, winner_name, winner_ioc)
winner_countries<-rename(winner_countries, name = winner_name, ioc = winner_ioc)

loser_countries<- select(slam_matches, loser_name, loser_ioc)
loser_countries <- rename(loser_countries, name = loser_name, ioc = loser_ioc)

#combine winners and loser
winners_losers<- unique(rbind(winner_countries,loser_countries))
```

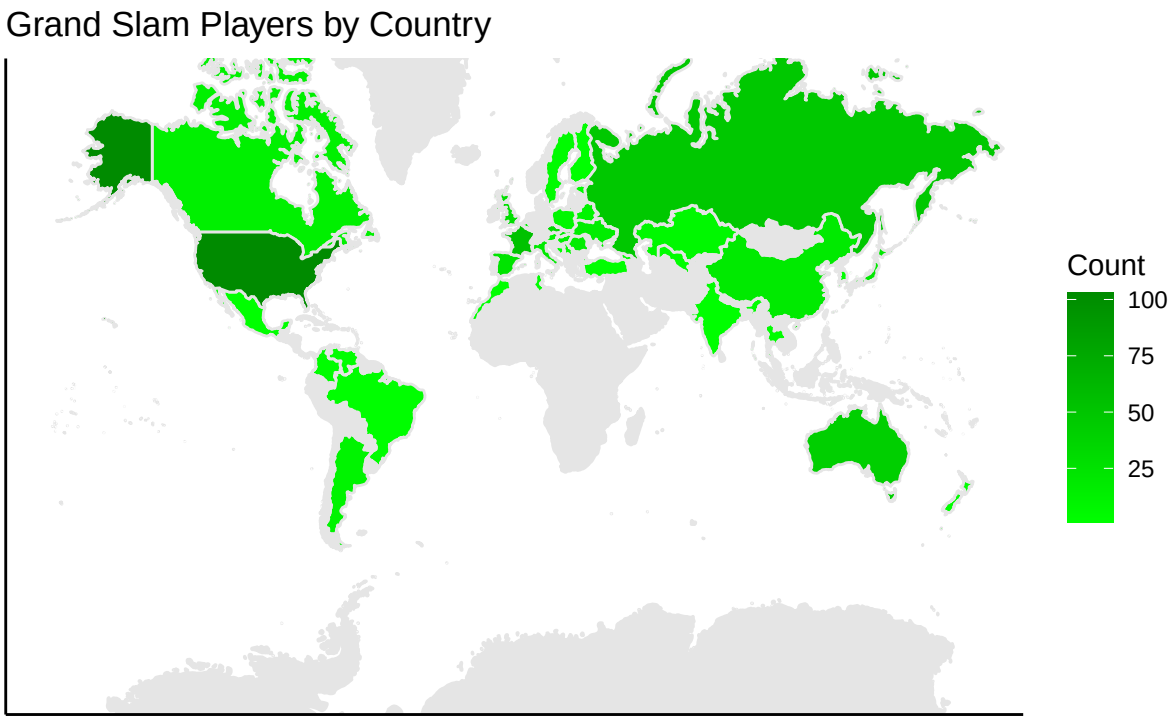
```
winners_losers <- summarise(group_by(winners_losers,ioc), count = n())

#load in map data and merge with player data
all_countries<-map_data("world")
names(all_countries)[5] <- "ioc"
#change country names in map data to abbreviation
all_countries[5] <- countrycode(all_countries[,5], 'country.name','iso3c')
```

```
## Warning in countrycode_convert(sourcevar = sourcevar, origin = origin, destination = dest, : Some val
```

```
#merge matches and map data
matches_country_map <- full_join(winners_losers, all_countries, by = "ioc")
matches_country_map
```

```
## # A tibble: 99,356 x 7
##    ioc   count  long   lat group order subregion
##    <chr> <int> <dbl> <dbl> <dbl> <int> <chr>
##  1 ARG      10 -64.5 -54.7    17  1148 Isla de los Estados
##  2 ARG      10 -64.4 -54.7    17  1149 Isla de los Estados
##  3 ARG      10 -64.2 -54.7    17  1150 Isla de los Estados
##  4 ARG      10 -64.1 -54.7    17  1151 Isla de los Estados
##  5 ARG      10 -64.1 -54.7    17  1152 Isla de los Estados
##  6 ARG      10 -64.0 -54.7    17  1153 Isla de los Estados
##  7 ARG      10 -63.9 -54.7    17  1154 Isla de los Estados
##  8 ARG      10 -63.8 -54.7    17  1155 Isla de los Estados
##  9 ARG      10 -63.8 -54.8    17  1156 Isla de los Estados
## 10 ARG      10 -64.0 -54.8    17  1157 Isla de los Estados
## # ... with 99,346 more rows
```

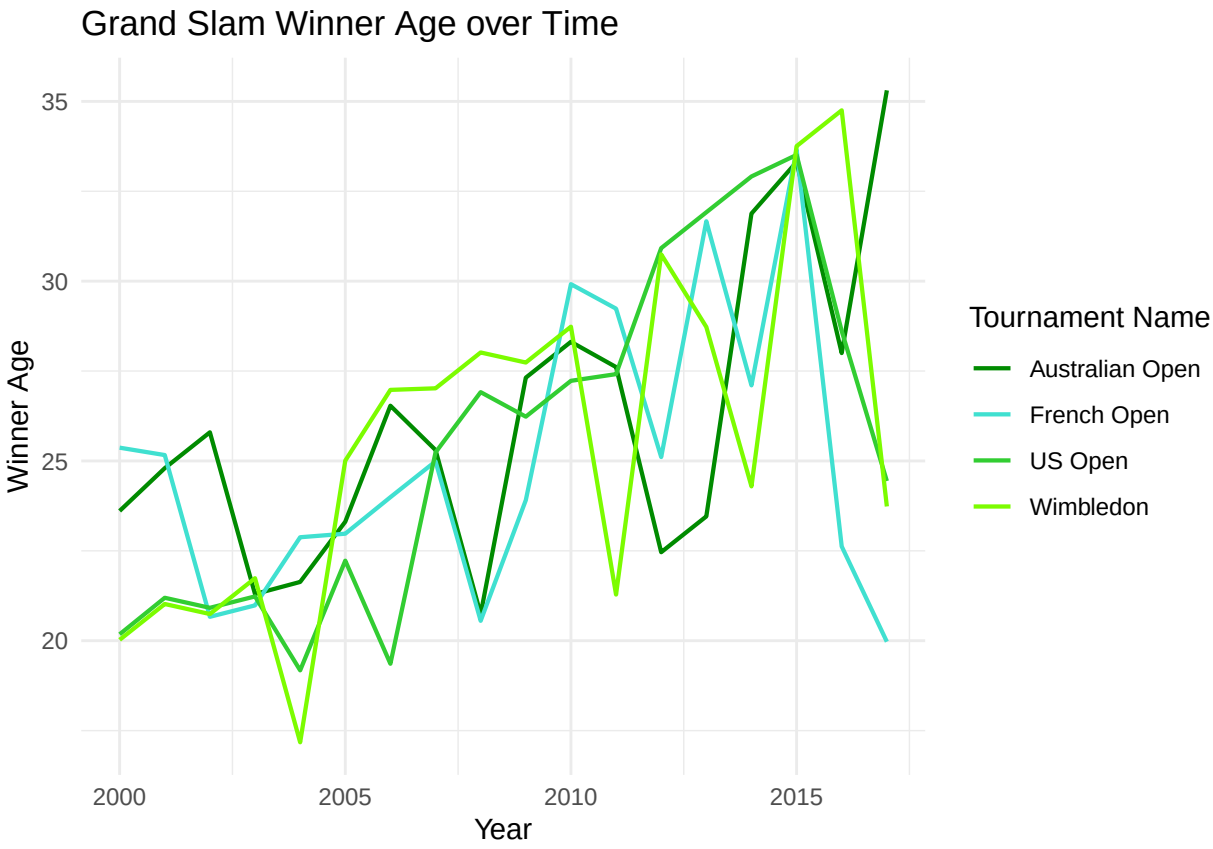
```
players_chloropleth <- ggplot(matches_country_map) + geom_polygon(aes_string(x="long", y = "lat", group
  labs(x = "", y = "", title = "Grand Slam Players by Country") + theme_classic() + theme(axis.ticks.y =
players_chloropleth
```



Age of Winners in Respective Grand Slams

This graph shows female grand slam winner age over time for the last 17 years. Each grand slam is represented by a different color line. Age tends to trend upwards over time for each grand slam though there are a high amount of fluctuations.

```
##LINEPLOT OF TOURNAMENT WINNER AGE OVER TIME

#get grand slam winners data for each slam
slam_matches <- filter(matches, tourney_level == "G")
slam_finals <- filter(slam_matches, round == "F")
slam_finals <- mutate(slam_finals, tourney_name = fct_recode(tourney_name,
                                                             "French Open" = "Roland Garros"))
g <- ggplot(slam_finals) + geom_line(size=.75,aes_string(x="year", y= "winner_age", group = "tourney_nar
g
```


Grand Slam Winner Age over Time
Winner Age
35
30
25
20
2000
2005
2010
2015
Year
Tournament Name
Australian Open
French Open
US Open
Wimbledon