Notes

Passwords

Home directory: 10023
Username: claraslateliu
Pass: Heyangel02
Token: Use duo

CSV File Commands

Head: give me the headers of this csv file
Tail: last 10 things in file
History: history of all commands just done

Linux Codes

Ls: list all files
Mv: move files
Cd: Change to particular directory
. name is the same name
Rm: remove (use carefully)
Pwd: where you are (if you get lost)
Scp: Secure copy
Cp: Copy file
Asterix: Moves it to where you want

Research Question

- Women's soccer growth in the NWSL, WSL and World Cup
- Women's soccer growth in the united states, england and internationally

June 3

Possible Topics

- Women’s Soccer and Women’s sports

- Mental health and viewership
- Looking for data sets

Womens Sports Viewership

https://www.nielsen.com/insights/2023/womens-sports-viewership-on-the-rise/
https://www.shu.edu/news/poll-on-women-s-sports-show-us-more-we-ll-watch-more.html

https://www.gwi.com/hubfs/Downloads/Women's_Sports.pdf?utm_medium=email&_hsenc=p2ANqtz--YyuZ3tmdS8sOW1Sbuna8S2265gaiQD9tzYaDLrPCKrRioXtF_8T0_S0EnFF5gmQiKBH_zQ2uNIvJW6qxefdR5GFfYDQ&_hsmi=290172601&utm_content=290172601&utm_source=hs_automation

https://www.rainforgrowth.com/pdf/rain-media-impact-report-44/

https://womensmediacenter.com/news-features/six-signs-of-progress-for-womens-sports

https://www.pbs.org/newshour/show/how-womens-sports-are-breaking-through-and-scoring-big-wins-with-mainstream-audiences

https://journals.library.brocku.ca/index.php/jess/article/view/4552/3292

https://business.yougov.com/content/47309-the-growing-interest-in-womens-sport-across-the-world

https://www.forbes.com/sites/lindseyedarvin/2023/10/31/media-coverage-for-womens-sports-has-nearly-tripled-in-five-years-according-to-new-research/?sh=44785295ebbc

Concussion Male and Female Sports

https://fslab.org/datasets/02_ConcussionsInMaleAndFemaleCollegeAthletes.html
Raw:
https://raw.githubusercontent.com/fslaborg/datasets/main/data/ConcussionsInMaleAndFemaleCollegeAthletes_adapted.tsv

General Statistics for Women's Soccer

https://fbref.com/en/squads/df9a10a1/Portland-Thorns-FC-Stats#all_stats_standard

https://public.tableau.com/app/profile/sarah.rodgers/viz/FIFAWomensWorldCup_15612491877330/FIFAWomensWorldCup

NWSLR

https://github.com/adror1/nwslR/blob/master/nwslR.Rproj
https://github.com/nwslR/nwslR/blob/main/R/util.R

NCAA Databases

https://www.ncaa.org/sports/2017/9/5/ncaa-research-interactive-databases.aspx

June 4

Taac Analysis Portal New System Applications

1) DCV: Job script writes connection information to a file
2) Jupyter Notebook: Most common, using to prototype and work something out
3) R Studio: What we will be using
4) VNC: Do not use, same as DCV\

p3YuQDiY2T0ABtGXzWolMc9XRoy-nKy4NIFqD_Ih0FY

Once you are in RStudio….. $Work is unique to me, and $Stockyard is also unique to me… once on frontera can see specifically in these file systems

RStudio:

Scripts: $Home
Data: $Work or $Stockyard

Linux Commands: https://logit.io/blog/post/linux-command-cheat-sheet/

June 5

Moving file/dataset from laptop into the terminal

- Download data set into documents and folder (can get free dataset on Kaggle)
- Click on document and press command i (this shows you were the file is on your computer)
- In the terminal: cd, ls
- cd Documents/ then ls
- It should list all folders in documents
- Cd FOLDER NAME (or file name) then ls
- Type in: scp FILENAME.csv claraslateliu@frontera.tacc.utexas.edu: /work2/10023/claraslateliu/.
- MAKE sure that there is not a space between the .edu: and /work
- Remember that 10023 is my Home directory number
- It will ask you to log in using user, password, token
- When logged in, should list filename.csv
- Then type $STOCKYARD
- Then it will upload to stockyard/the terminal and you are good

THEN:

- To move into frontera terminal: cd $STOCKYARD, ls$STOCKYARD and file names that we moved from laptop to terminal to

Moving file/dataset from online to terminal

- Find data set online
- Download (get the link to copy)
- In the terminal put: wget
- Will copy the link into code

Statista File Links

https://www.statista.com/statistics/1275906/engagement-women-s-sporting-events-worldwide/

https://www.statista.com/statistics/272800/average-number-of-spectatators-at-the-fifa-womens-world-cup/

https://www.statista.com/statistics/1008207/uefa-womens-euro-cumulative-live-attendance/

https://www.statista.com/statistics/1269811/womens-sport-engagement/

https://datasetsearch.research.google.com/search?src=0&query=women%20sports%20viewership%20and%20ticket%20sales&docid=L2cvMTF2MHExZ25sMA%3D%3D

https://datasetsearch.research.google.com/search?src=0&query=women%20sports%20viewership%20&docid=L2cvMTF2ZGZ4cGx3bQ%3D%3D

https://datasetsearch.research.google.com/search?src=0&query=women%20sports%20viewership%20&docid=L2cvMTFya2Z4NDNwZA%3D%3D

https://datasetsearch.research.google.com/search?src=0&query=women%20sports%20viewership%20&docid=L2cvMTF2MDd6bmtqdw%3D%3D

https://www.statista.com/statistics/1386733/womens-world-cup-total-tv-viewership/

https://www.statista.com/statistics/1008207/uefa-womens-euro-cumulative-live-attendance/

https://www.statista.com/statistics/1236737/wnba-average-attendance/

Kaggle

https://www.kaggle.com/datasets/mgibert/womens-football

RSTUDIO

Log through the TAAC Portal and use user: claraslateliu and the password copy from the TAAC portal, changes everytime

Writing in R:

#: Comment, can write whatever and will not code
Command Return: Run code
Highlight and Press Run (green arrow): Run code

Getwd(): get work directory

Imputing Dataset from terminal to R:

```
#Set Working Directory
setwd("/work2/10023/claraslateliu")

#Get the current working directory
current_directory <- getwd()

#Print the current working directory
print(current_directory)

#Read data
attendance_dataset <-read.csv("AttendanceDataset.csv")
```

June 6

Learning R:

- Think in native language, what are my questions and what is my data? Load data into R. Thinking in our native language will translate to coding language through code
- Load data
- Verify data that just loaded…. Is data the same as data loaded into R?
- After verified… can move on

Loading packages

- It is crucial to load packages because packages are what contains the different codes to decipher the data
- library(tidyverse): packages are inside of the library
- After we load our library which has our packages, we can start to subset the data

- #Load Packages
- library(tidyverse)

Subsetting data by column

- Type: #Subsetting Data By Columns

- This sets the heading for what you are about to do, remember hashtag before anything will not run that… just for taking notes purposes.
- After, underneath, specify which columns you want to take out specifically. For example I only want to take out “Team” and “Total”. So I would write: onlyTeamAndTotalData <- select(attendance_dataset, Team, Total)
- Put the arrow after the two columns you want and then in brackets the file name, column you want, another column you want)
- Done. The code should run and it will show up in your environment and you can now click on the new data set you just generated with those specific columns.

Subsetting data by Rows

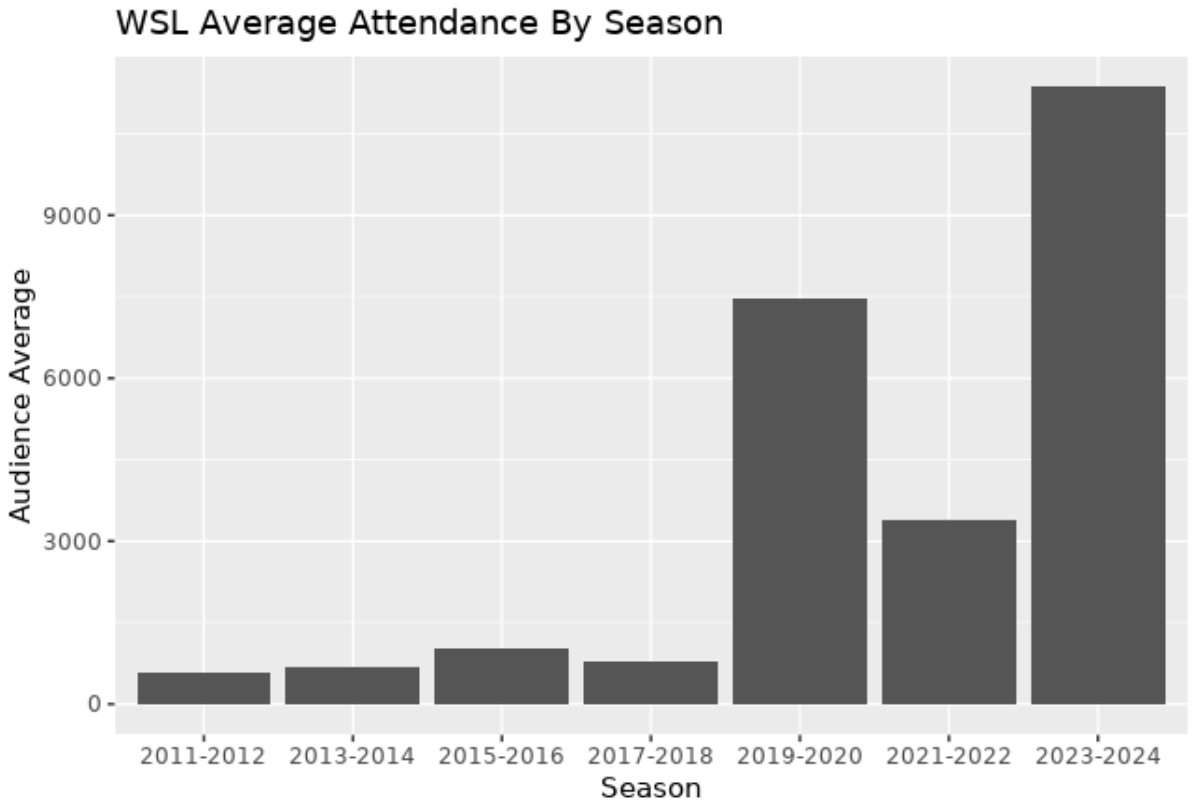
- Type #Subsetting Data By Rows
- With the hashtag, we create a title (will not code)
- Like columns, you are going to write down specifically which ROWS you want to take out. For example, I only want to take out the row USA.
- Then type: onlyUSA <- filter(attendance_dataset, Country == "USA")
- USA is in the column country so using == means that we only want USA in from that country column.
- If I want to do more than one row, I can. For example if I want USA and Spain I will write: onlyUSAandSpain <- filter(attendance_dataset, Country == "USA"|Country == "Spain"
- We can separate the two rows by using a |. For example: Country == "USA"|Country == "Spain
- It will show up in your environment and a new data set has been created with desired rows only.

Research Questions

- Is women's soccer growing worldwide?
- Women's soccer growth domestically and internationally
- Women's soccer growth world cup and club
- Did audience numbers increase throughout the women's world cup 2015-2024?
- Did audience numbers grow for the NWSL

Bar Graphs

3



June 10

Github In RStudio

- https://nsf-all-spice-alliance.github.io/SDG-Analytics-in-R/rmarkdowns/installation_guide.html#Create_a_Github_Account
- TOKEN/PASS: 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
- Create a github account

- In Rstudio, run a new job. Look down and there is consol, terminal and job. We are going to be typing in the terminal:

- Type: whichgit
- Then paste with our own username: `git config --global user.name "Jane Doe"`
- Then paste with our own email: `git config --global user.email janedoe@example.com`
- Paste: `git config --list --global`
- Then our username and email should run and show up

- Then going to install package: `install.packages("usethis")`
- `install.packages("usethis")`
- This should bring you to github, create a token name and set it to no expiration. Then generate token and copy the numbers/letters.

- Back to the console, paste: `gitcreds::gitcreds_set()`
- Will prompt for token, paste copied numbers/letters.
- Should say done

Terminal: setup
Console: writing code and installing packages
Environment: Names
Bottom right: all folders saved

June 14

TO DO:

Bar chart for world cup
Line graph WSL
Line graph NWSL

Creating Visuals Using R

- We need to download ggplot, a package that will help us create the visuals
- We can code that by typing: install.packages("ggplot2") or library(ggplot2)

- Before we begin creating our visuals we must load our package, go to our library and import ggplot.

Creating Line Graphs Using R

```
#Creating Visuals: Line Graph WSL Average Attendnace every second year 2011-2024
WSLdata <- data.frame(
  season = c("2011-2012", "2013-2014", "2017-2018", "2019-2020", "2021-2022", "2023-2024"),
  av_att = c(239, 671, 722, 3066, 1589, 6988)
)

ggplot(WSLdata, aes(x = season, y= av_att)) +
  geom_line(aes(group = 1), color = "navy") +
  geom_point()+
  labs(
    title = "WSL Average Game Attendance",
    subtitle = "Seasons 2011-2024",
    x = "Season",
    y = "Average Attendance"
  )+
  theme_light()
```

Explaining:

This R code generates a line graph using the `ggplot2` package to visualize the average game attendance in the Women's Super League (WSL) for selected seasons from 2011 to 2024, with some additional formatting:

Creating Data Frame:
r

Copy code

```
WSLdata <- data.frame(
  season = c("2011-2012", "2013-2014", "2017-2018",
"2019-2020", "2021-2022", "2023-2024"),
  av_att = c(239, 671, 722, 3066, 1589, 6988)
)
```

1.
    - `WSLdata` is a data frame that contains two columns:
        - `season`: Represents the seasons from 2011-2012 to 2023-2024.
        - `av_att`: Represents the corresponding average attendance numbers for each season.

Creating the Plot:
r
Copy code

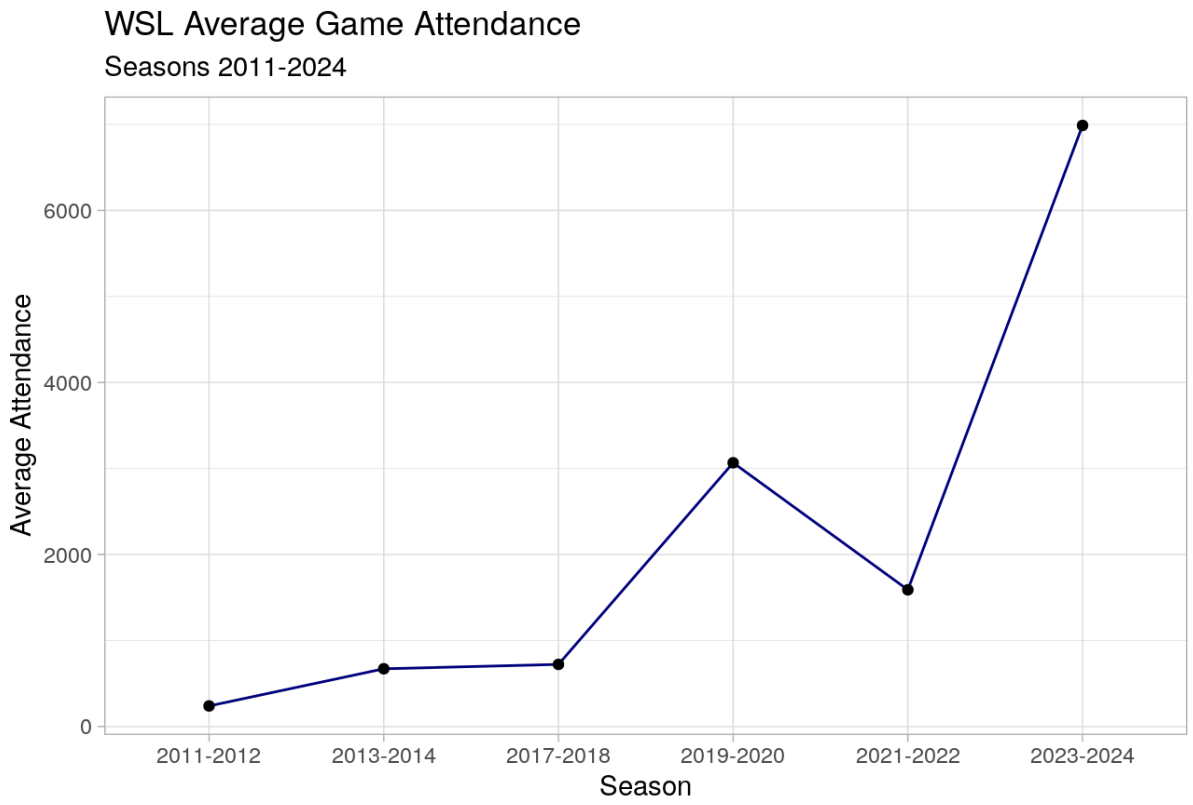
```
ggplot(WSLdata, aes(x = season, y = av_att)) +
  geom_line(aes(group = 1), color = "navy") +
  geom_point() +
  labs(
    title = "WSL Average Game Attendance",
    subtitle = "Seasons 2011-2024",
    x = "Season",
    y = "Average Attendance"
  ) +
  theme_light()
```

2.
    - `ggplot(WSLdata, aes(x = season, y = av_att))`: Initializes a ggplot object with `WSLdata` as the data frame and specifies `season` as the x-axis variable and `av_att` as the y-axis variable.
    - `geom_line(aes(group = 1), color = "navy")`: Adds a line plot (`geom_line`) to the plot. `aes(group = 1)` ensures that all points are connected by a single line. The `color = "navy"` specifies the color of the line to navy blue.
    - `geom_point()`: Adds points (`geom_point`) to the plot, which represent the actual data points for each season.
    - `labs(title = "WSL Average Game Attendance", subtitle = "Seasons 2011-2024", x = "Season", y = "Average`

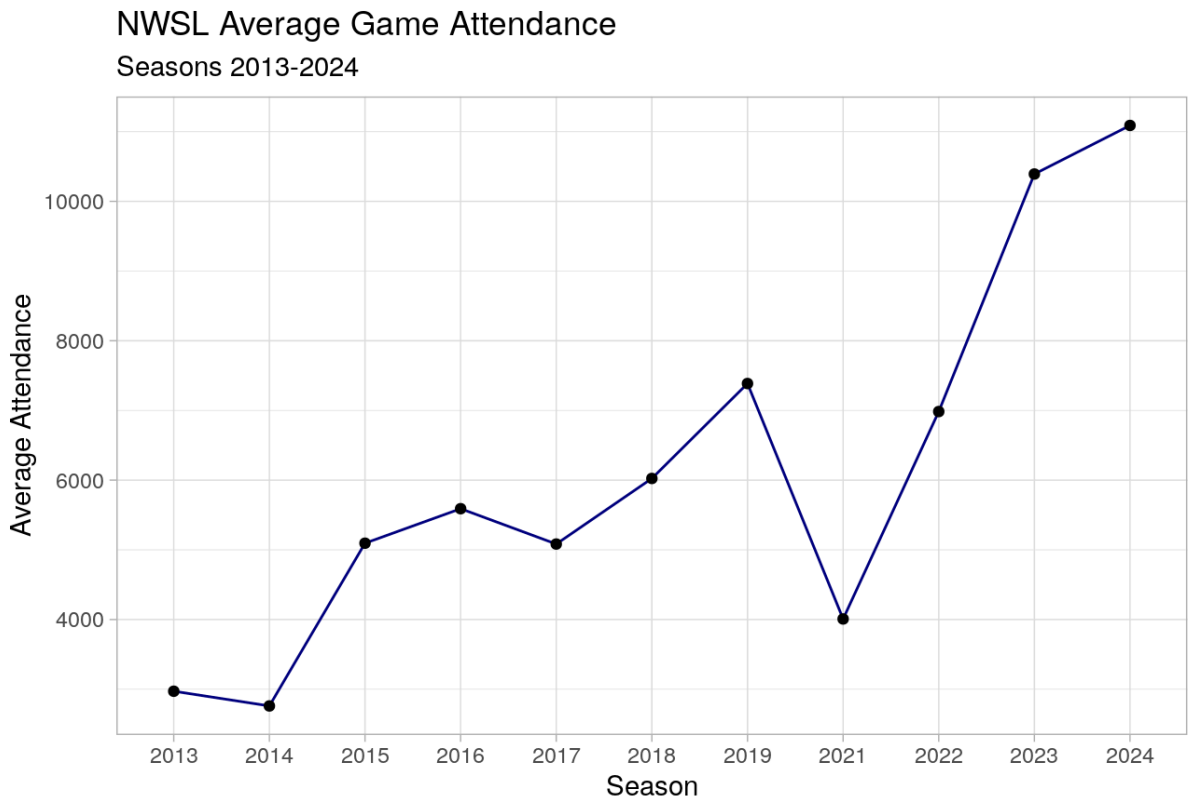
`Attendance")`: Sets the plot labels - the title of the plot is "WSL Average Game Attendance", with a subtitle indicating the range of seasons plotted, and labels for the x-axis ("Season") and y-axis ("Average Attendance").

- `theme_light()`: Applies a light theme to the plot, which affects the background, grid lines, and other visual elements to give a consistent appearance.

Therefore, when you run this code, it will produce a line graph showing the average game attendance in the Women's Super League for the specified seasons (2011-2012 to 2023-2024). Points will be marked for each season, and these points will be connected by a line to visualise the trend in attendance over time. The plot will have a title, subtitle, and labelled axes, with a light theme applied for aesthetic consistency.



I repeated the same steps with the NWSL Data and got this graph:



Creating Bar Graphs Using R

```
#Creating Visuals" Bar Graph for WorldCup
WorldCupData <- data.frame(
year = c("2011", "2015", "2019", "2023"),
av_att = c(26398, 21628, 21756, 30911)
)

ggplot(WorldCupData, aes(x = year, y = av_att)) +
  geom_bar(stat = "identity", fill= "navy") +
  geom_text(aes(label= av_att), vjust = -0.5, color = "black", size = 3.3) +
```

```
  labs(
    title = "Women's World Cup Average Attendance",
    x = "Year",
    y = "Average Attendance"
  ) +
  theme_light()
```

Explaining

Code Explanation

r

Copy code

```r
# Step 1: Create the Data Frame
WorldCupData <- data.frame(
  year = c("2011", "2015", "2019", "2023"),
  av_att = c(26398, 21628, 21756, 30911)
)
```

Here, a data frame named `WorldCupData` is created with two columns:

- `year`: A vector of years in which the World Cups were held.
- `av_att`: A vector of average attendances for those years.

r

Copy code

```r
# Step 2: Load the ggplot2 Package
library(ggplot2)
```

This line loads the `ggplot2` package, which is required to create the plot. This line should be included at the top of your script if `ggplot2` is not already loaded.

r

Copy code

```
# Step 3: Create the Bar Graph

ggplot(WorldCupData, aes(x = year, y = av_att)) +

  geom_bar(stat = "identity", fill= "navy") +

  geom_text(aes(label= av_att), vjust = -0.5, color = "black",
size = 3.3) +

  labs(

    title = "Women's World Cup Average Attendance",

    x = "Year",

    y = "Average Attendance"

  ) +

  theme_light()
```

Detailed Breakdown:

1. **`ggplot(WorldCupData, aes(x = year, y = av_att))`**:
   - Initializes the `ggplot` object with `WorldCupData` as the data source.
   - `aes(x = year, y = av_att)` maps the `year` column to the x-axis and the `av_att` column to the y-axis.
2. **`geom_bar(stat = "identity", fill= "navy")`**:
   - Adds a bar graph layer to the plot.
   - `stat = "identity"` means that the heights of the bars represent the actual values in the `av_att` column.
   - `fill = "navy"` sets the color of the bars to navy blue.
3. **`geom_text(aes(label = av_att), vjust = -0.5, color = "black", size = 3.3)`**:
   - Adds text labels to the bars.

- `aes(label = av_att)` sets the text labels to be the values of the `av_att` column.
- `vjust = -0.5` vertically adjusts the position of the labels so they appear just above the bars.
- `color = "black"` sets the color of the text labels to black.
- `size = 3.3` sets the size of the text labels.

4. **`labs(title = "Women's World Cup Average Attendance", x = "Year", y = "Average Attendance")`**:
   - Adds labels and a title to the plot.
   - `title` sets the main title of the plot.
   - `x` sets the label for the x-axis.
   - `y` sets the label for the y-axis.
5. **`theme_light()`**:
   - Applies the `theme_light()` theme to the plot, which gives a light, clean background.

Complete Code

r

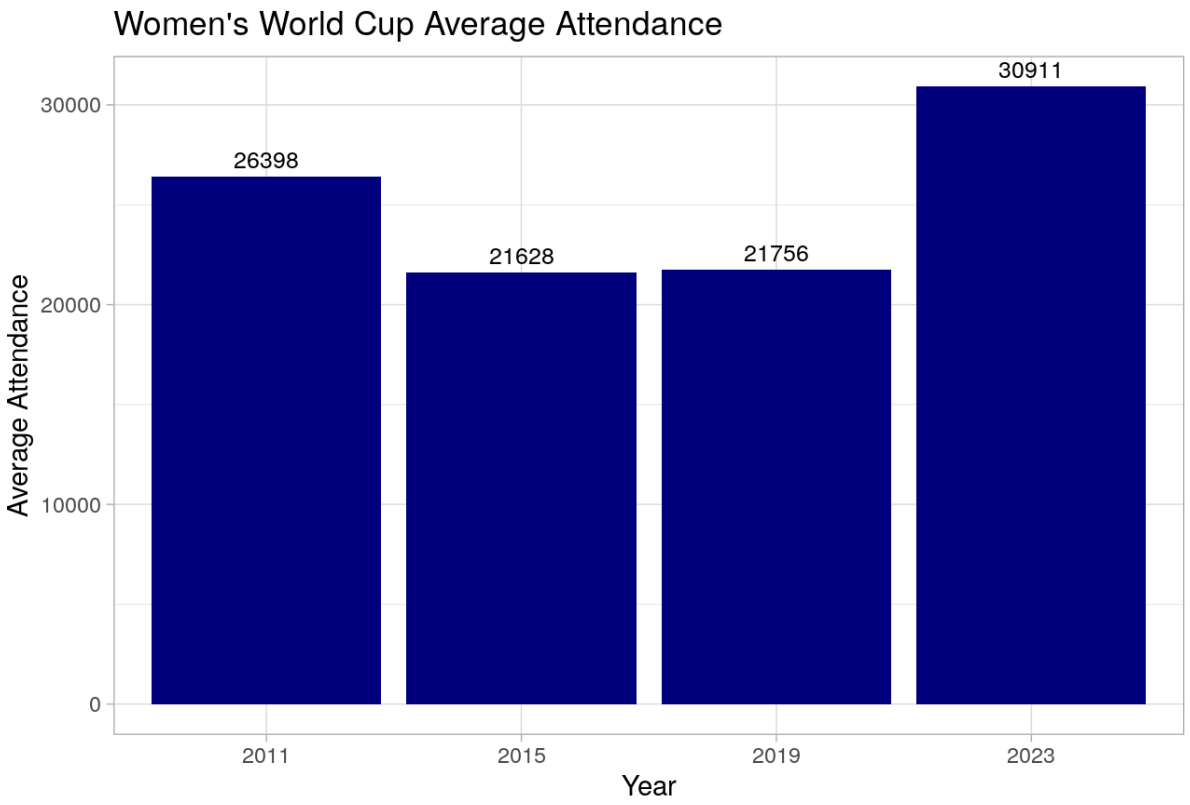
Copy code

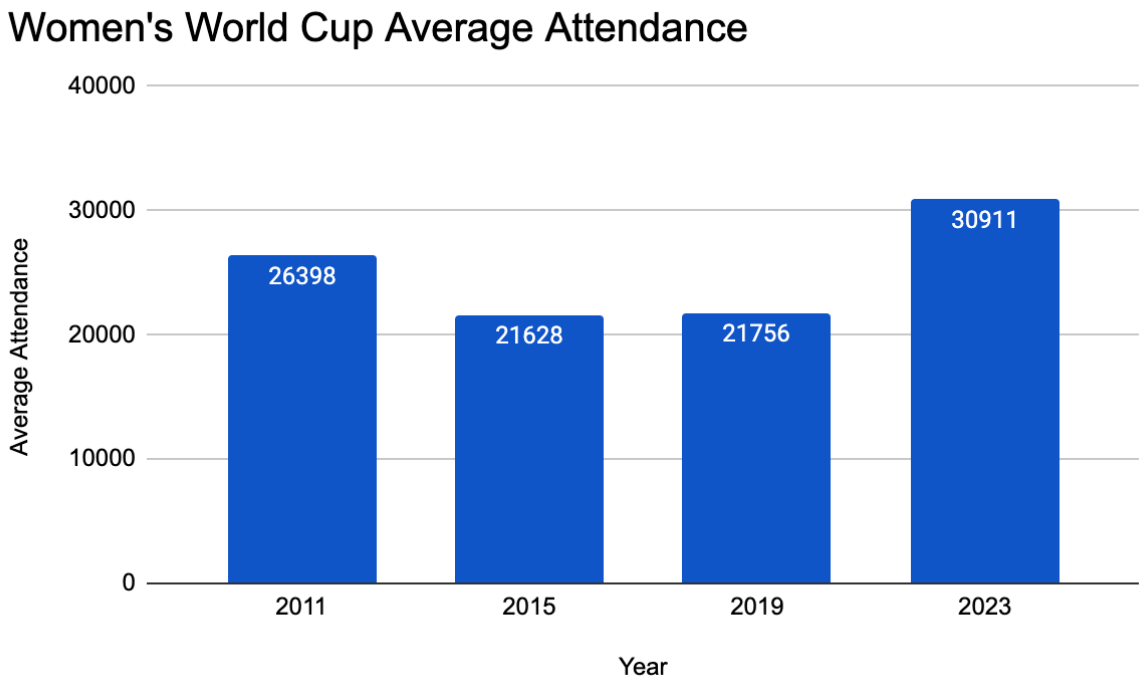
```r
# Load the ggplot2 package

library(ggplot2)


# Creating Visuals: Bar Graph for WorldCup

WorldCupData <- data.frame(

  year = c("2011", "2015", "2019", "2023"),

  av_att = c(26398, 21628, 21756, 30911)

)


ggplot(WorldCupData, aes(x = year, y = av_att)) +

  geom_bar(stat = "identity", fill= "navy") +
```

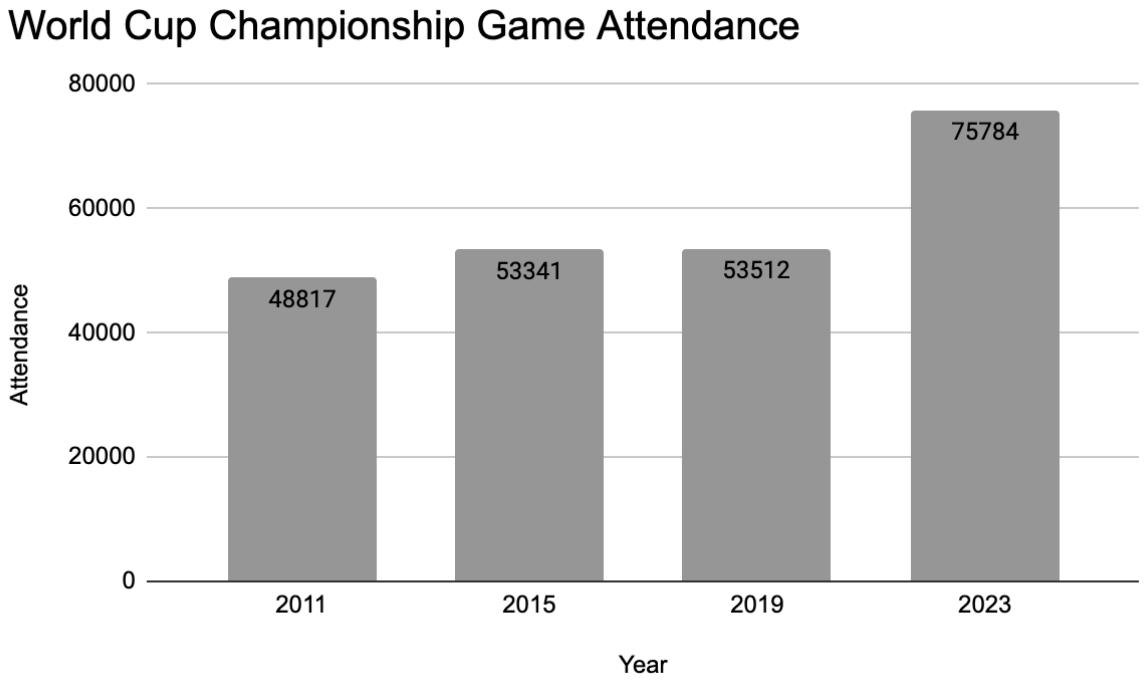
```
  geom_text(aes(label= av_att), vjust = -0.5, color = "black",
size = 3.3) +

  labs(

    title = "Women's World Cup Average Attendance",

    x = "Year",

    y = "Average Attendance"

  ) +

  theme_light()
```

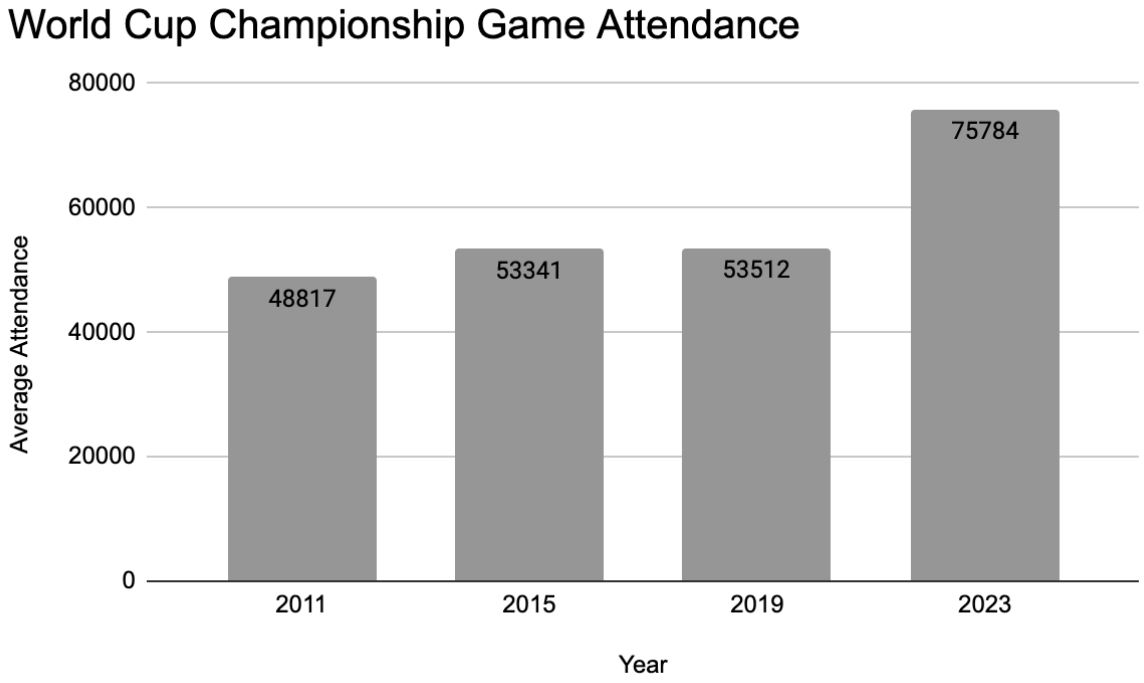
This code will produce a bar graph showing the average attendance for the Women's World Cup in the years 2011, 2015, 2019, and 2023, with each bar labeled with its respective attendance value.

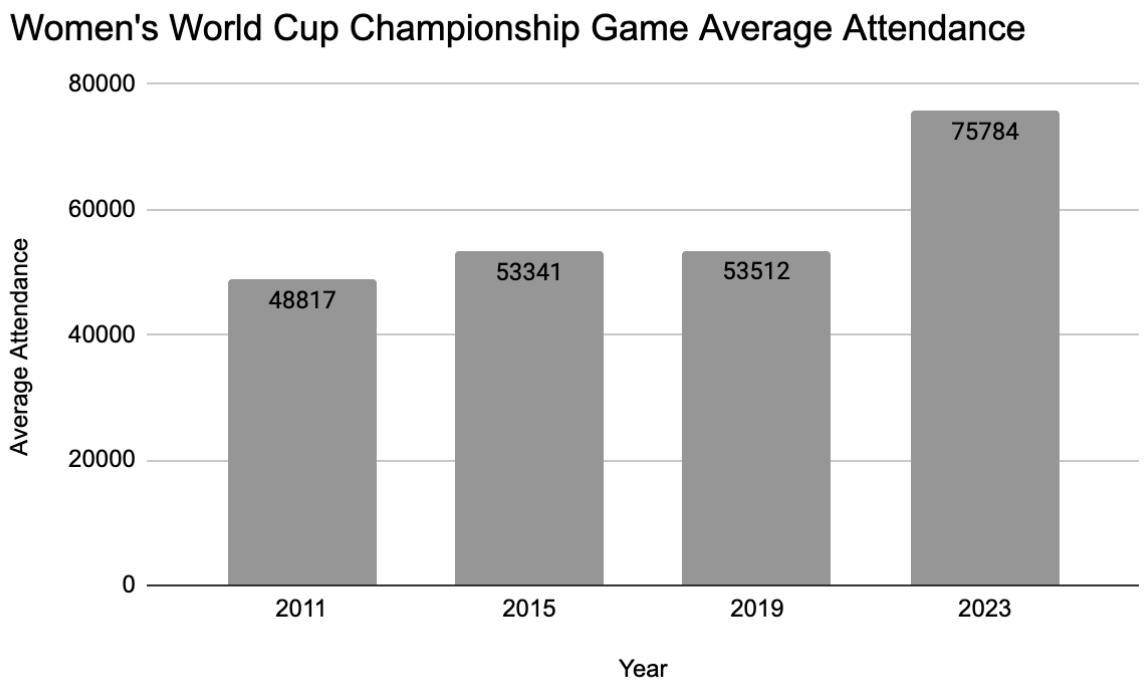
4o

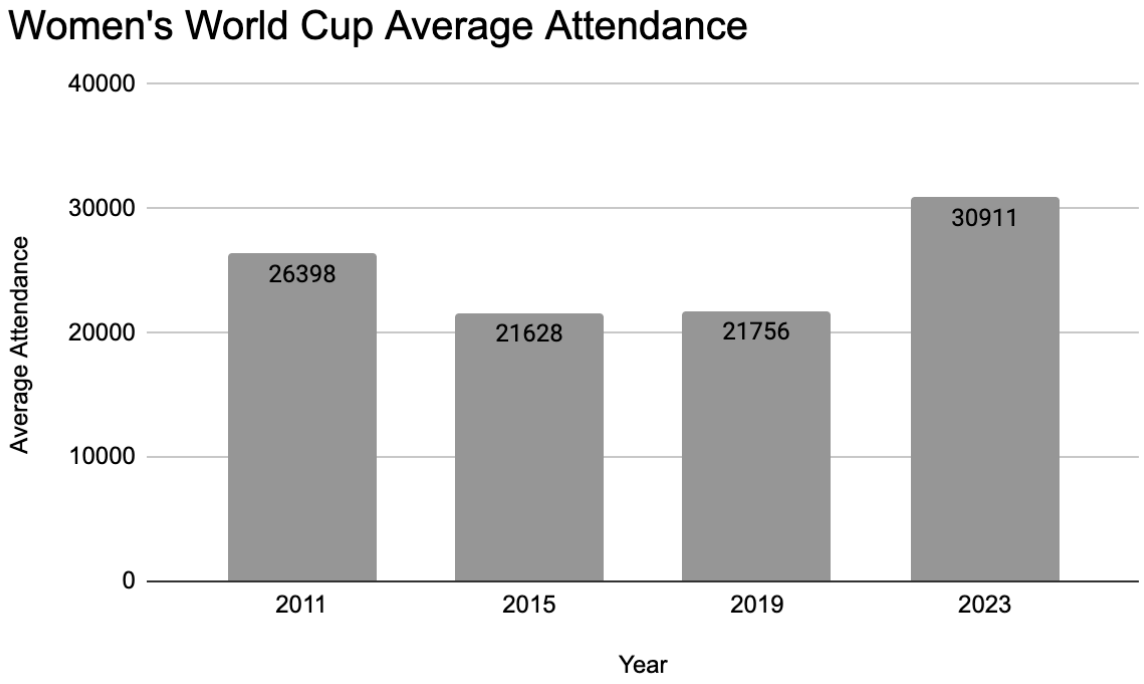

Women's World Cup Average Attendance
30000
20000
10000
0
Average Attendance
26398
21628
21756
30911
2011
2015
2019
2023
Year


Women's World Cup Average Attendance
40000
30000
20000
10000
0
Average Attendance
26398
21628
21756
30911
2011
2015
2019
2023
Year


World Cup Championship Game Attendance
80000
60000
40000
20000
0
Attendance
48817
53341
53512
75784
2011
2015
2019
2023
Year


World Cup Championship Game Attendance
80000
60000
40000
20000
0
Average Attendance
48817
53341
53512
75784
2011
2015
2019
2023
Year


Women's World Cup Championship Game Average Attendance
80000
60000
40000
20000
0
Average Attendance
48817
53341
53512
75784
2011
2015
2019
2023
Year



June 17

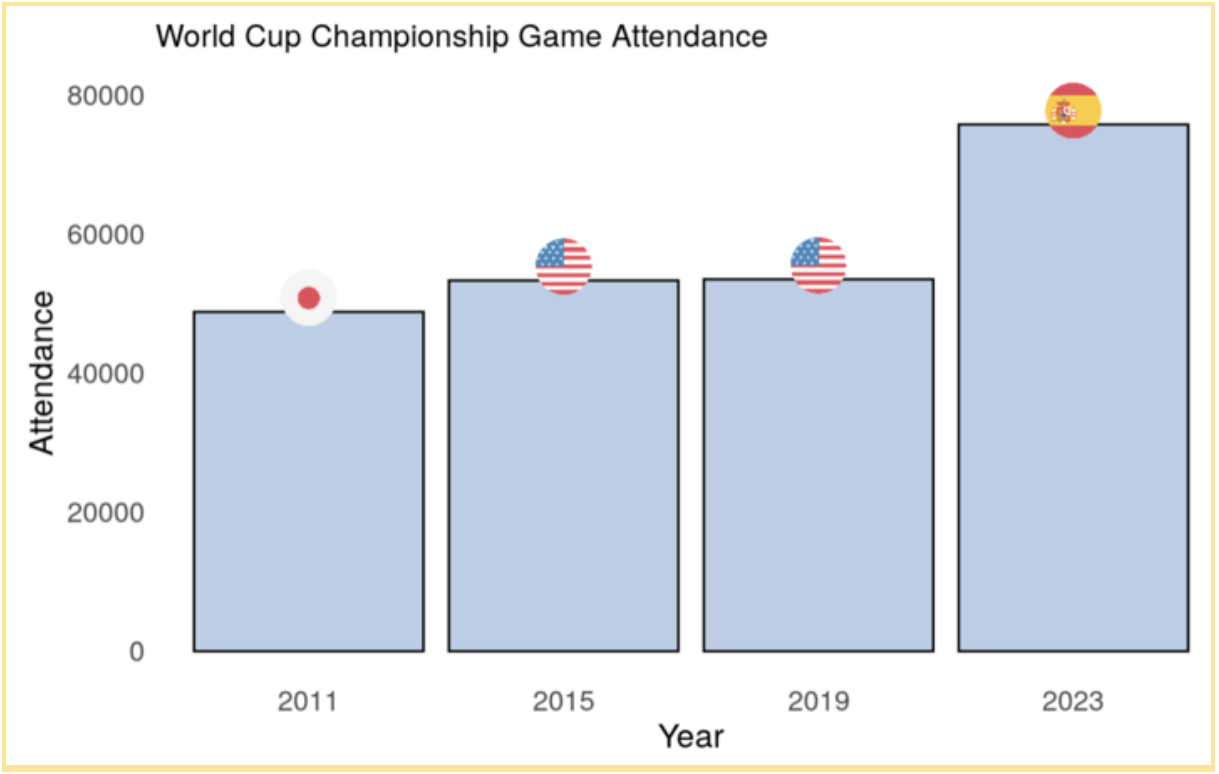
- Subset data- NWSL, WSL and Worldcup from Large Dataset, write this code and subsetted data before visuals

- Find code to make calculations of average of multiple numbers (instead of how we did it in google sheets- for method purpose)

Poster

Articles: Women's Soccer Growth/Attendance Growth

World Cup Championship Game Attendance Graph


World Cup Championship Game Attendance
80000
60000
40000
20000
0
Attendance
2011
2015
2019
2023
Year

Cham

Nsf

Tacc

Cifal

Unitar

Innovation

Aim

Epsor

Air force