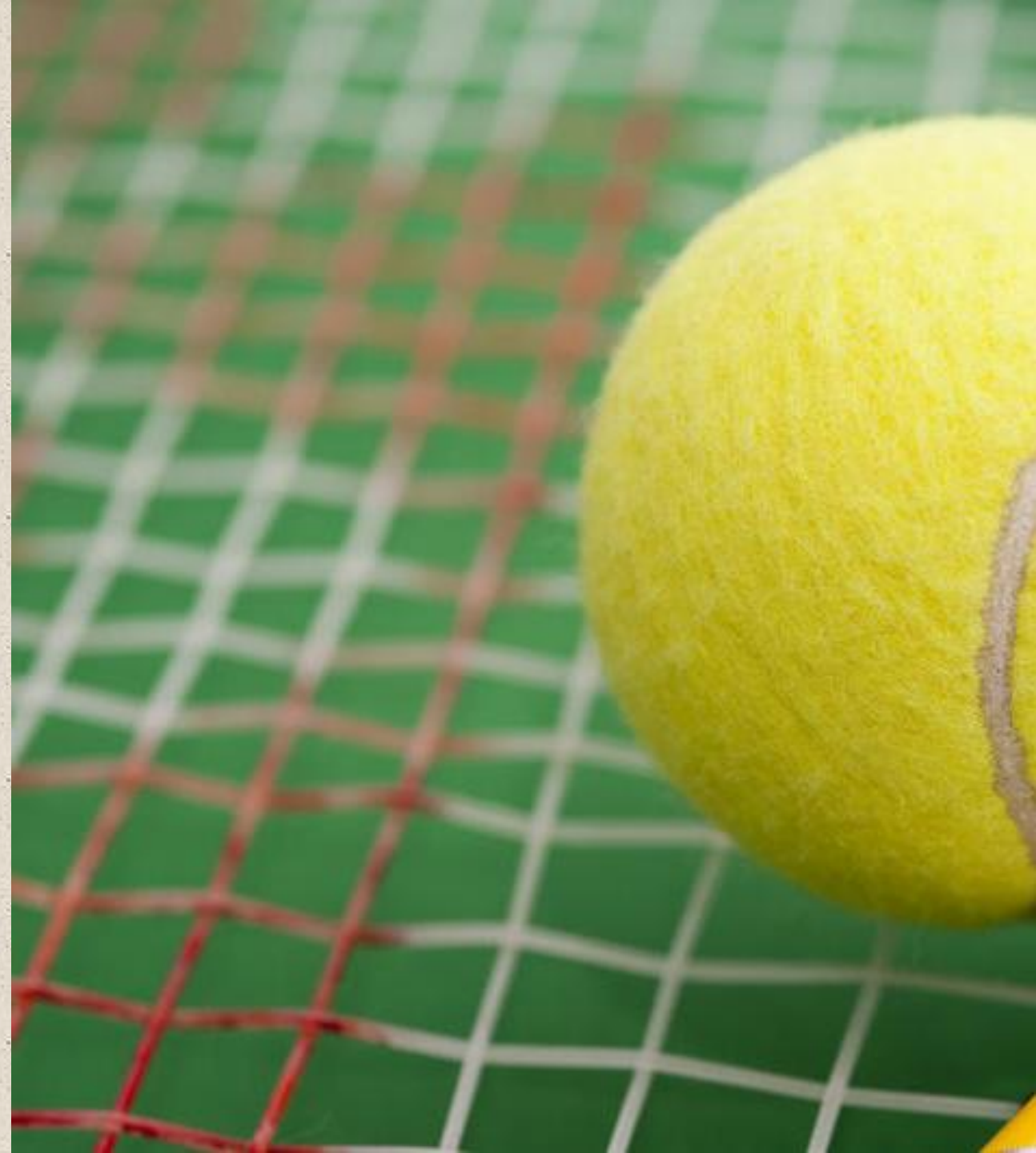
Mahanthi Bukkapatnam
Rebecca Johnson
Danny Kramer
Ram Yellepeddi

WHO IS THE G.O.A.T?

Question: Who is the greatest male professional tennis player in the Open Era?

- Analyze 46 years of ATP (Association of Tennis Professionals) data
- Create an index to compare players
- This question is hotly debated in the media and among tennis fans



THE SITUATION

It's too close to call…

- Roger Federer: most Grand Slam titles and weeks ranked at #1
- Jimmy Connors: most overall titles
- Rafael Nadal: currently ranked #1 and highest career winning percentage overall
- Novak Djokovic: only player in history to hold all four Grand Slams on three different surfaces at the same time
- Someone else?

A BRIEF HISTORY

Men's Professional Tennis

- The "Open Era" began in 1968 when Grand Slam tournaments agreed to allow professional players to compete with amateurs [1]
- Since 1972 the professional tournaments have been part of various tour circuits [2]
  - Grand Prix, World Championships, ATP Tour
- Australian Open was played on grass until 1988 [3]
- US Open has been played on 3 surfaces [4]
  - grass until 1974
  - clay from 1975 until 1977
  - hard courts from 1978 until today
- Player rankings: determined over the past 52 weeks by adding up ranking points from tour events [5] :
  - 4 Grand Slams
  - 8 Master's 1000s
  - 4 best ATP 500s
  - 2 best ATP 250s
  - ATP Finals if player qualifies

Sources:

1 https://en.wikipedia.org › wiki › History_of_tennis

2 https://en.wikipedia.org › wiki › Association_of_Tennis_Professionals

3 https://en.wikipedia.org › wiki › Australian_Open

4 https://en.wikipedia.org › wiki › Tennis_court

5 https://en.wikipedia.org › wiki › ATP_Rankings

- **Contains Jeff Sackmann's master ATP player file, historical rankings, results, and match stats**

- **Includes tour-level main draw matches:**
  - **159,812 observations of 50 variables**

- **The different tournament levels are represented by single letters: 'G' = Grand Slams, 'M' = Masters 1000s, 'A' = other tour-level events**

- **We used Wikipedia for # of weeks ranked #1 data due to gaps in years for Jeff Sackmann's rankings data**

THE DATA

1. **ATP Tennis Results and Stats on GitHub**
2. **Wikipedia: One table: List of ATP #1 Ranked Singles Players**

PREPARING THE DATA

Programming in R

- Read in data
  - Read in Jeff Sackmann csv files: 1973:2019
  - Wikipedia Rankings: extracted table
- Column Changes
  - deleted 35, renamed 2, added 6 and defined classes
- Used ifelse ( ) to define index variables by Round and Level or by Surface
- Removed NA's by using filter( )

ANALYSIS

Programming in R

- Created new dataframes for key variables
- Added a new weight column for each dataframe
- Assigned a value for each weight column
- Created df_index and df_our_ranks for our final calculation

ANALYSIS

Create Player Indices

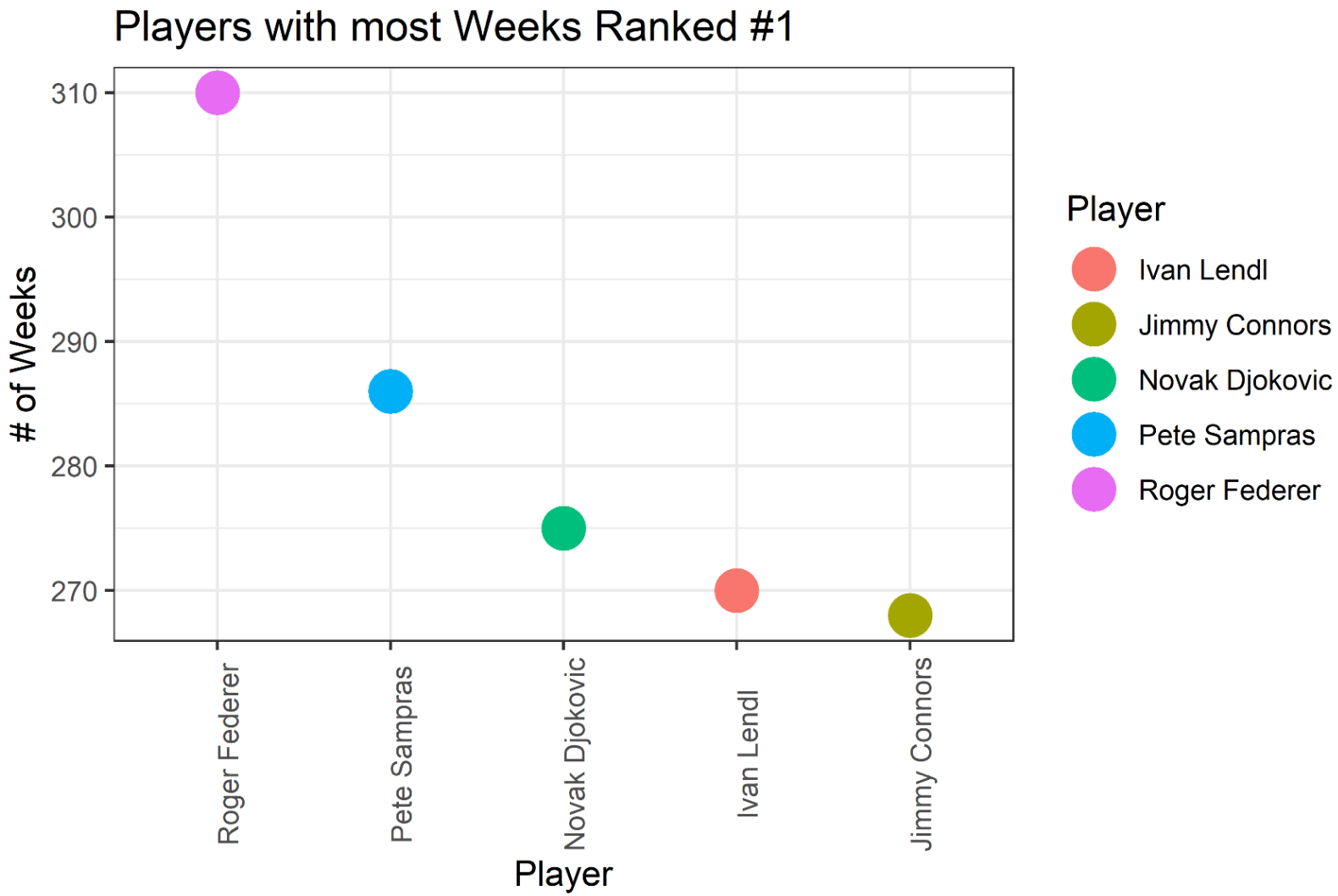
Developed a method to compare players across the Open Era

- Assigned a weight to each of the following
- Variables:
  - Number of
    - Grand Slam Titles
    - Grand Slam Finals
    - Grand Slam Semifinals
    - All Tournament Titles
    - All Tournament Finals
    - All Tournament Semifinals
    - Weeks Ranked at #1 for ATP Tour
    - Career Match Wins on Clay Courts
    - Career Match Wins on Grass Courts
    - Career Match Wins on Hard Courts

	A	B	C	D	E	F
1	**Variable**	**# of Obs**	**Our Weight**	**Multiplier**	**New our weight**	**Final Weight**
2	GrandSlamTitles	184	0.3	1000	300	1.6304
3	GrandSlamFinals	184	0.18	1000	180	0.9783
4	GrandSlamSemis	368	0.108	1000	108	0.2935
5	All Tourney Titles	3478	0.1125	1000	112.5	0.0323
6	All Tourney Finals	3478	0.0675	1000	67.5	0.0194
7	All Tourney Semis	6769	0.0405	1000	40.5	0.0060
8	Weeks Ranked #1	2410	0.047875	1000	47.875	0.0199
9	Match wins Clay	64860	0.047875	1000	47.875	0.0007
10	Match wins Hard	57538	0.047875	1000	47.875	0.0008
11	Match Wins Grass	16762	0.047875	1000	47.875	0.0029
12						
13			**Final weight = New our Weight / # of OBS**			

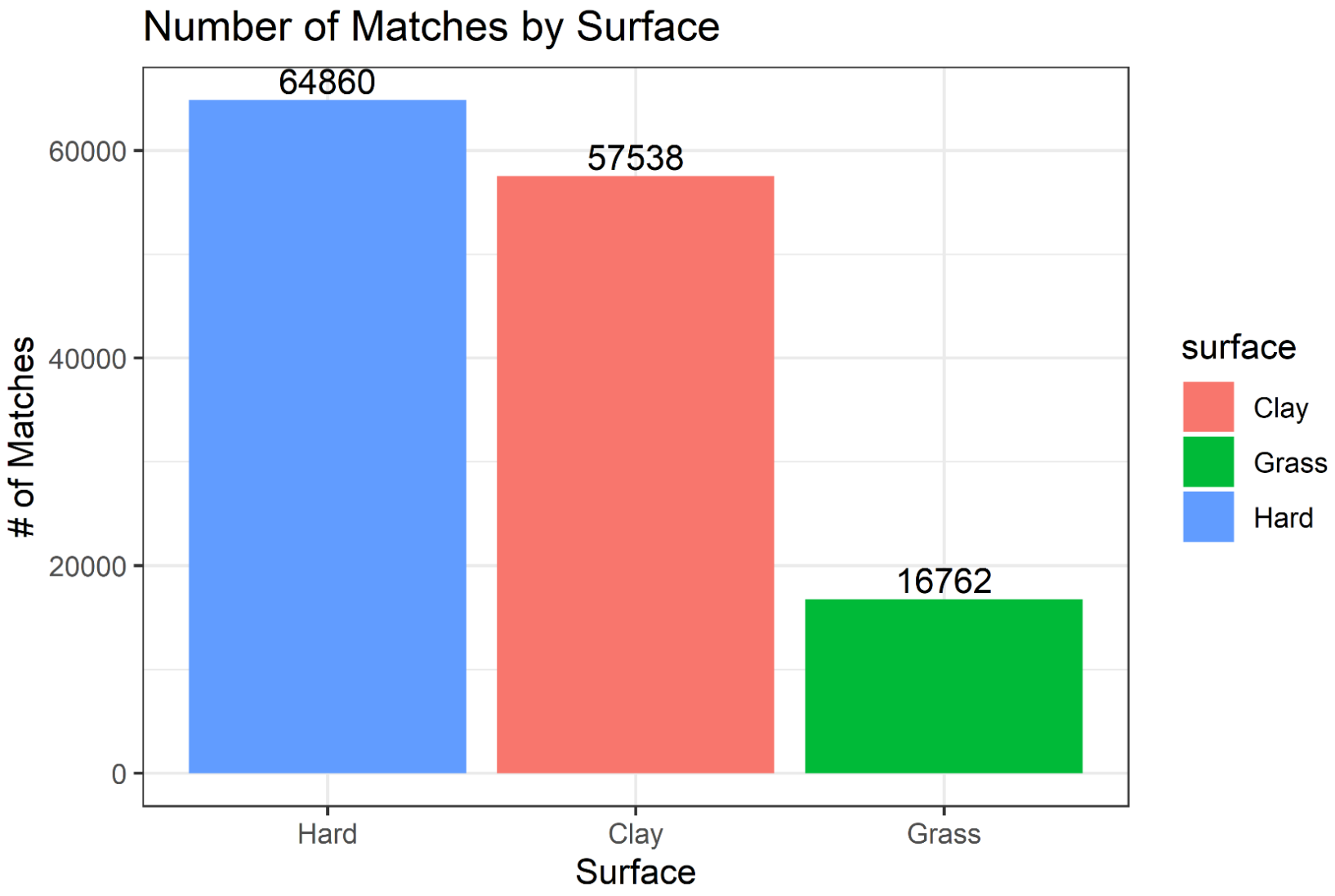
ANALYSIS

Variable Weights



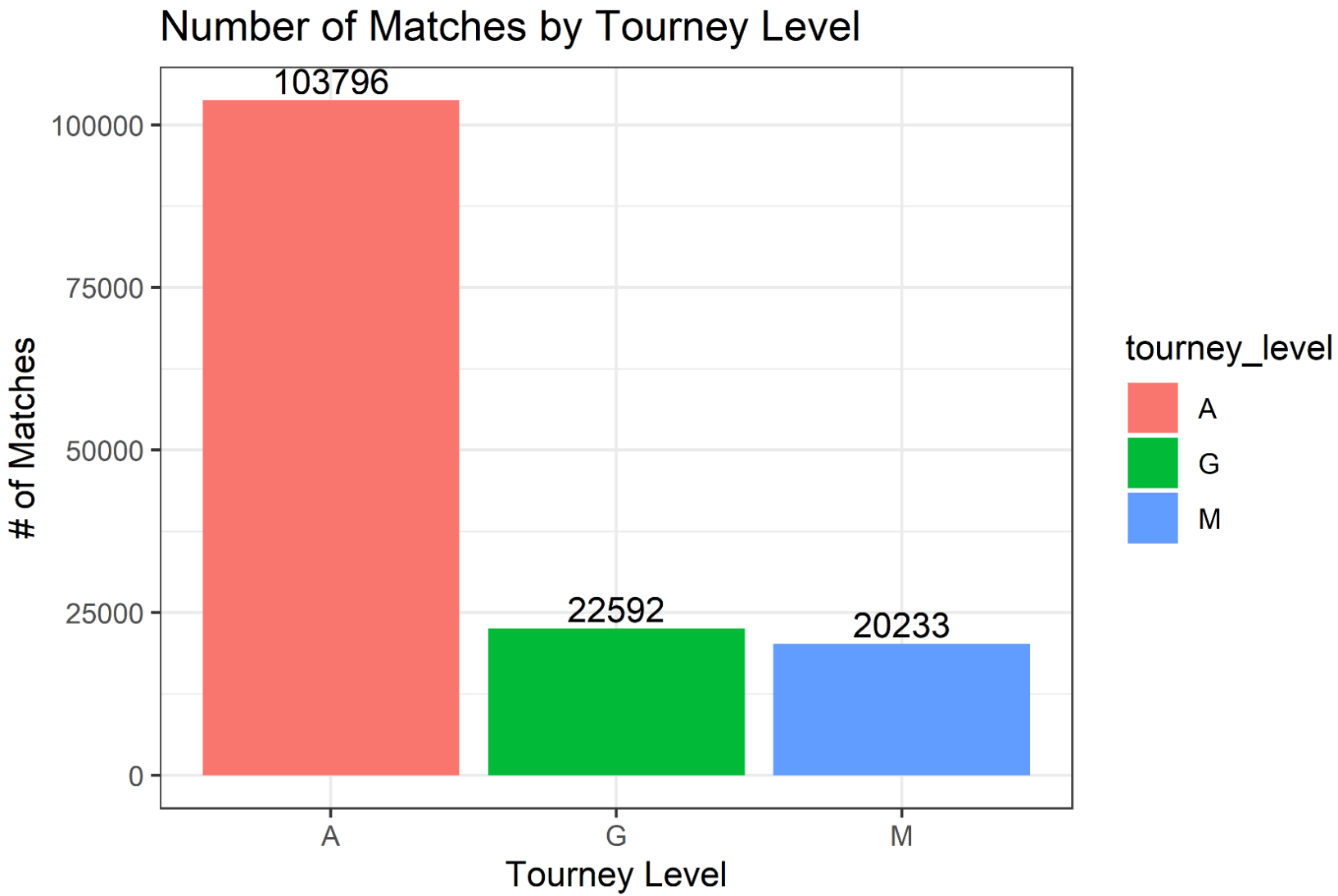
ANALYSIS

Exploring the Data



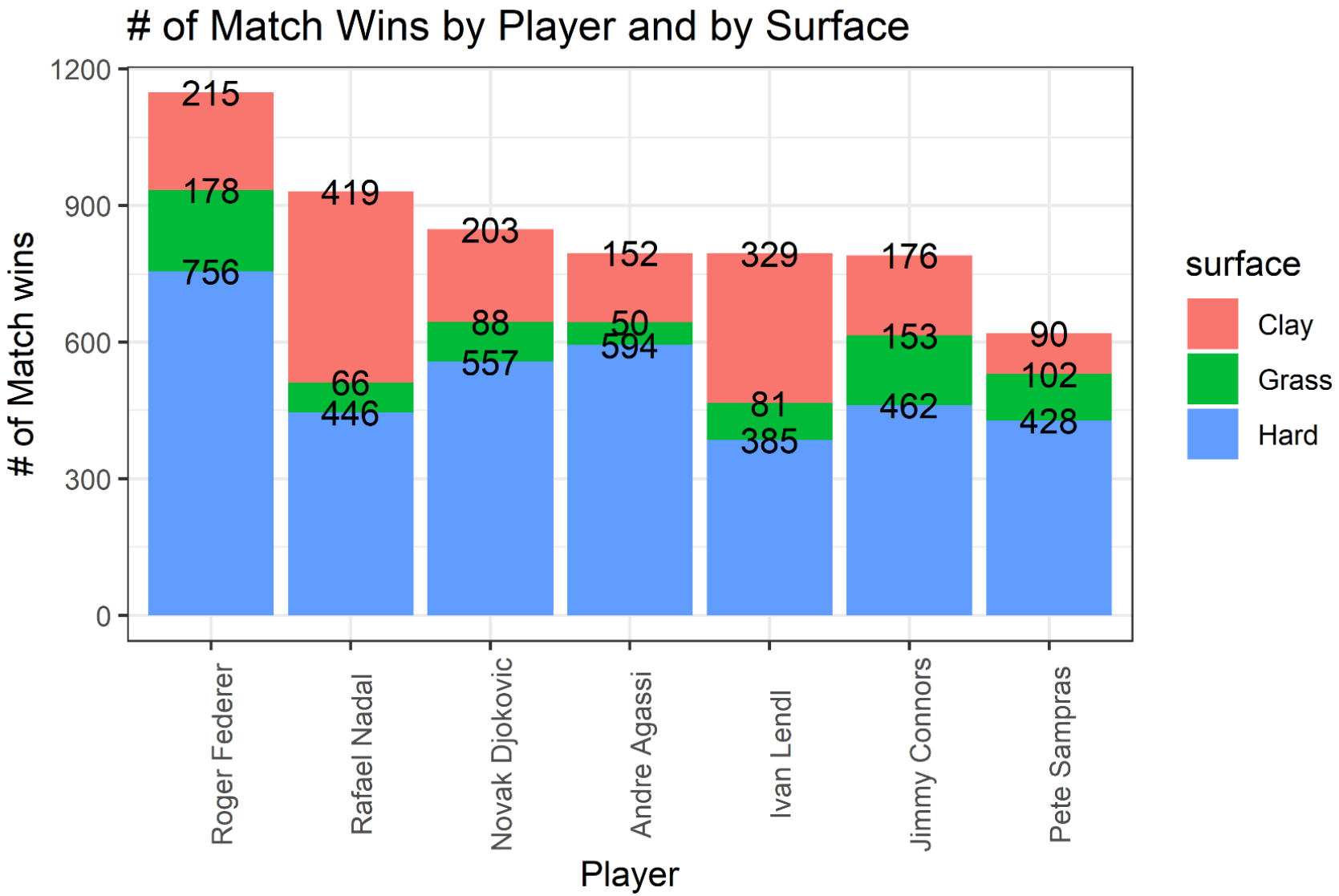
ANALYSIS

Exploring the Data



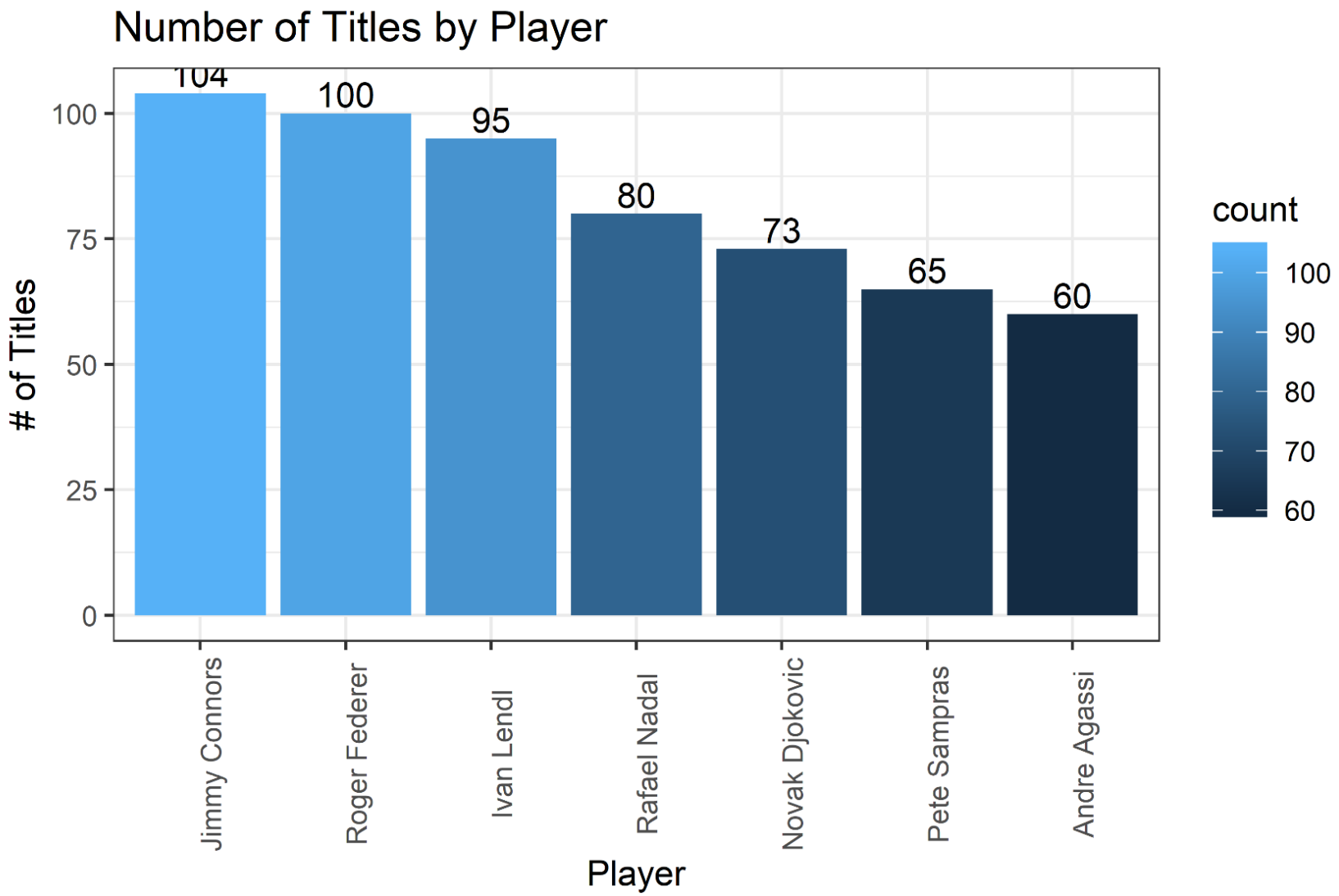
ANALYSIS

Exploring the Data



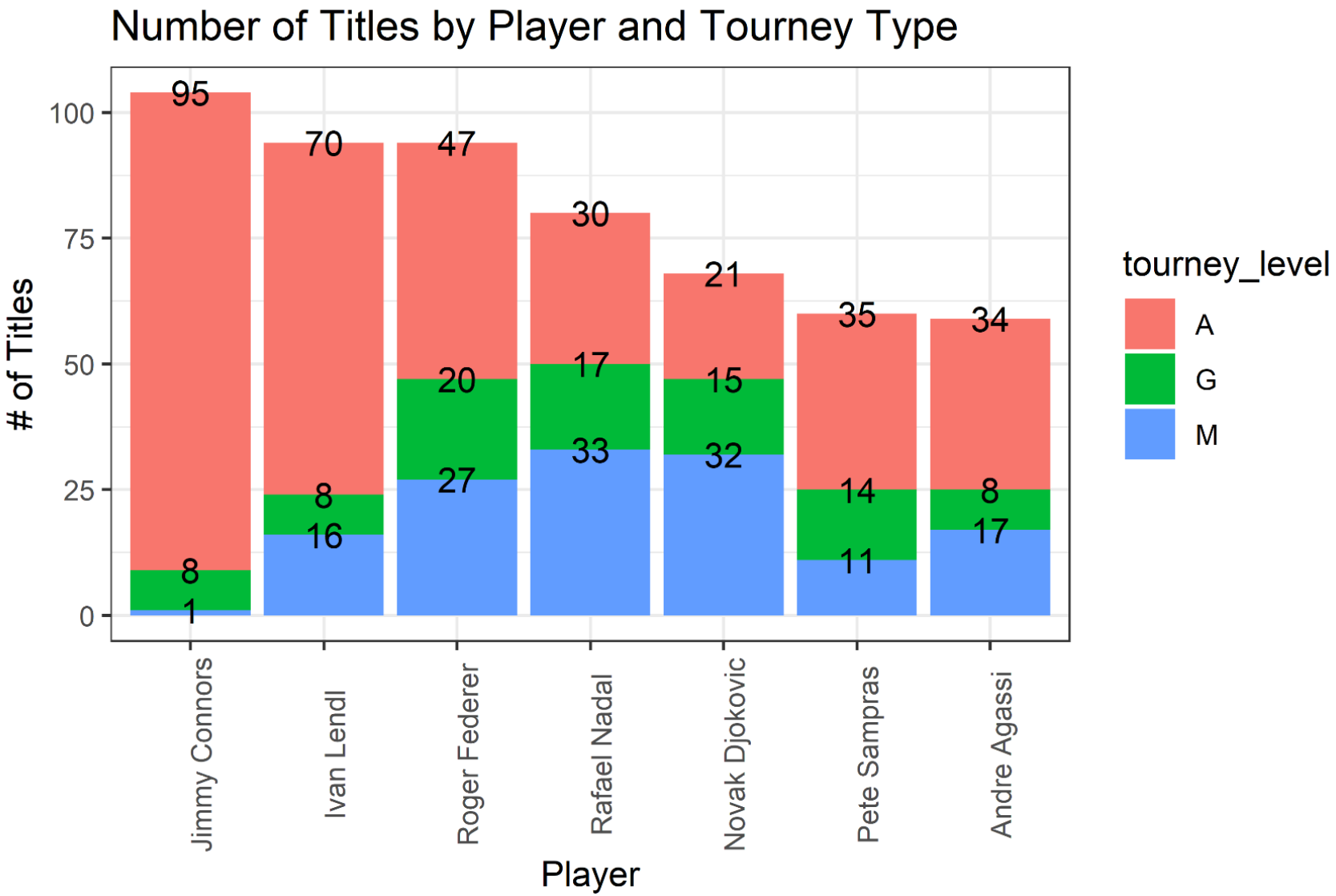
ANALYSIS

Exploring the Data



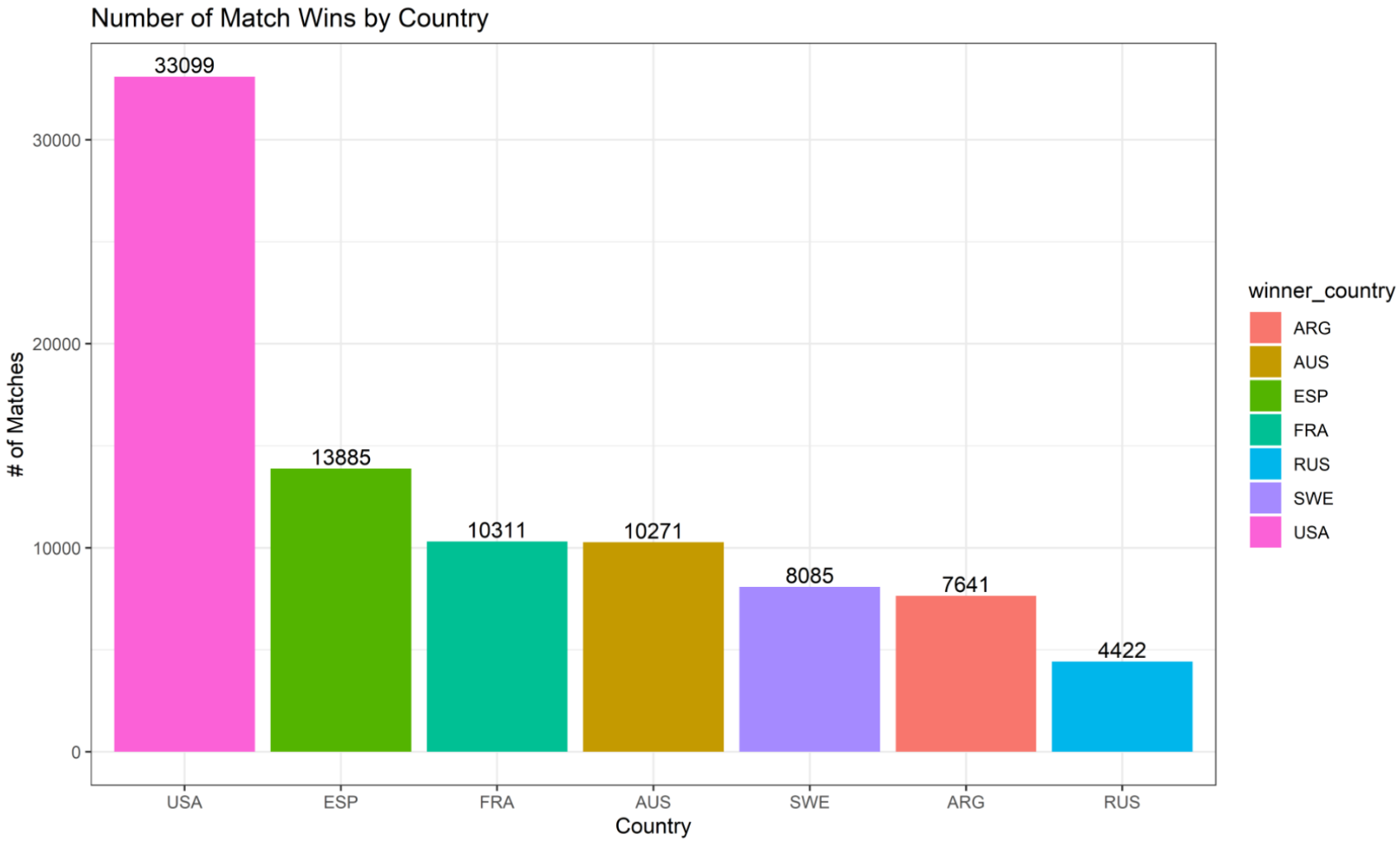
ANALYSIS

Exploring the Data



ANALYSIS

Exploring the Data



ANALYSIS

Exploring the Data

Challenges for our project

- **Sourcing complete tennis data**
  - Professional men's tennis experienced significant changes and leadership between 1973-2019
  - There are some tour-level matches with missing stats because ATP doesn't have them
  - We spent a lot of time figuring out the best data source
  - We weren't able to include 2019 March – October data
- **Downloading/merging multiple csv files**
  - variables with differing factor levels caused warnings when loading our data and during our analysis
- **Determining weights for player indices**
- **Reordering the plot x-axis results**

OVERCOMING OBSTACLES

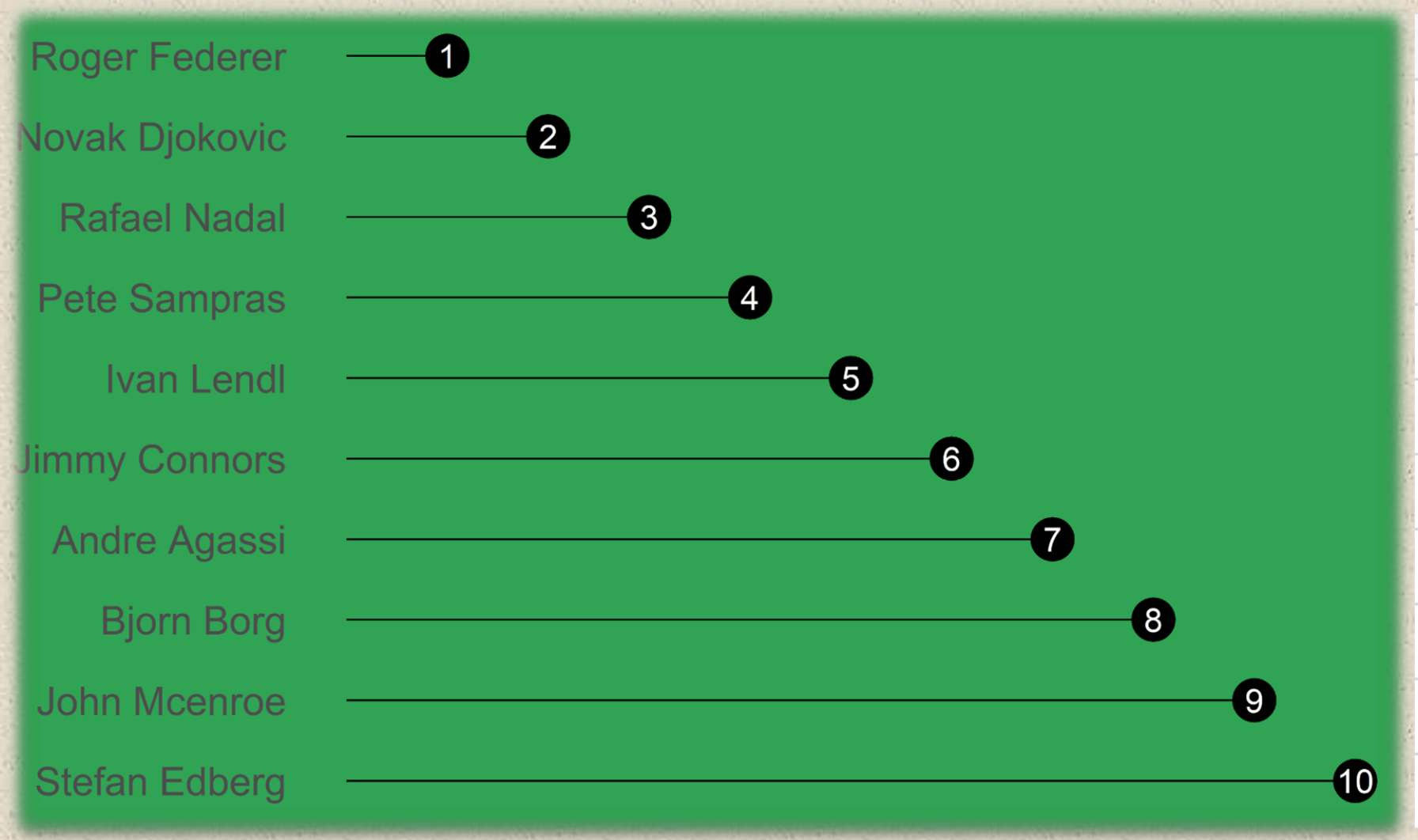
Our Solutions

- **Spent time searching for the best data source**
  - **We found the best data source available for free**
  - **We used a separate data set for # Weeks Ranked #1**
  - **We added data from 2019 for our final predictions**
- **We brought the factor variables into R as characters and converted them to factors**
- **We based our initial weights on the same percentages as Tour level events**
- **We created a spreadsheet to determine weights based on total observations for each variable**
- **Spent time exploring ggplot capabilities**

AND THE G.O.A.T IS.....

Roger Federer

TOP TEN PLAYERS FROM OUR INDEX



PlayerName	total	Rank
Roger Federer	56.9784	1
Novak Djokovic	44.8372	2
Rafael Nadal	44.4853	3
Pete Sampras	36.6153	4
Ivan Lendl	36.3440	5
Jimmy Connors	35.0324	6
Andre Agassi	28.0836	7
Bjorn Borg	25.7403	8
John Mcenroe	21.4285	9
Stefan Edberg	21.2756	10

ONE LAST THOUGHT

Predictions

Additional Analysis in R

How many Grand Slams will the top 3 players win over next 5 years?

- **Jeff Sackmann has a formula with points for:**
  - **Last two years of Grand Slam Titles, Finals and Semis**
  - **Player age over/under 27 years**
- **Federer, Nadal, and Djokovic stand out**
  - **Won all 8 Grand Slams in 2018 and 2019**
  - **Won 54 of last 64 Grand Slams (beginning in 2004)**
- **We created a data file for top 3 current players:**
  - **2018 and 2019 Grand Slam results and birthdays**

R FUNCTION

Predictions

- Initialized player data vectors and entered data
- Created points vector and calculated points for each player
- Wrote a function to sum the contents of each players' points vector:

```
Myfunc <- function(vector) {
    total <- 0
    for(i in vector) {
        total <-  i + total
    }
     total
}
```

- Calculated predicted number of Grand Slams:

DjokovicPredSlams <- (myfunc(DjokovicPoints))/100

Analysis Results:

- **Nadal will win 3 more for a total of 22 Grand Slams**
- **Federer will win 0 more for a total of 20 Grand Slams**
- **Djokovic will win 3 more for a total of 19 Grand Slams**

PREDICTIONS