Women's vs. Men's Tennis

Is the WTA really more 'inconsistent' than the ATP?

Group E: Julie Ye, Chris Meng, Alan Zhou

December 14, 2021

Contents

Introduction: Emma Raducanu, Anomaly or Continuing a Trend?

Entering the 2021 US Open, Emma Raducanu was an 18-year-old teenager who had just finished her British A-level exams (Schifano, 2021). Stymied by the COVID-19 pandemic, Raducanu entered the tournament ranked a lowly #150 with zero WTA[1]-level match wins under her belt, needing to navigate the qualifying[2] draw in order to merely participate in the US Open (WTA, 2021). Yet, two weeks later, Raducanu was the last woman standing as champion of one of the biggest tournaments in the tennis world without dropping a single set (WTA, 2021). In other words, she dominated the opponents in front of her with extremely limited prior professional experience. To top it off, her opponent in the final was another teenager ranked a lowly #73, who also caused a string of stunning upsets over top players on her way to the final.

On the other hand, the men's final of the US open was contested between the top 2 seeds[3] (Loop, 2021). Novak Djokovic, the #1 seed, was looking to win the fourth and final Grand Slam[4] of the calendar year after winning the previous three (Loop, 2021). Daniil Medvedev, the #2 seed, caused what was considered to be a "major upset" relative to men's tennis when he defeated Djokovic in straight sets (Loop, 2021).

Since we began following tennis in the early 2010s, the contrast in the outcomes of the US Open between the WTA and ATP[5] players was just one example of a larger theme of the mainstream perception that the women's tour is more inconsistent than the men's tour. In this report, we explored the question: Is the WTA *really* more inconsistent than the ATP? We wanted to break down this idea, see how much data supports this perception, and learn more about the specific differences that might influence this perception.

Loading Data

On the state of tennis data, Jeff Sackmann, a pioneer of widely accessible tennis data and owner/creator of the datasets that we used in this analysis, stated the following: "So, first off, the sorry state of tennis data. There's a lot of cooks in the kitchen, and there aren't even any plates. There are lots of people out there who are collecting some data but not really any who are giving it to anyone."[6] In response to this issue that he identified, Sackmann created a public Github repository for ATP and WTA tennis rankings, results, and stats (Sackmann, 2021).

In order to address our research question about differences in consistency between the ATP and WTA tours, we extracted data from the repositories `tennis_wta` and `tennis_atp` from Sackmann's GitHub and combined the datasets as such:

```r
data_URLs <- paste("https://raw.githubusercontent.com/JeffSackmann/tennis_wta/master/wta_matches_",
                   1968:2021,
                   ".csv",
                   sep = "")
wtaresults <- read_csv(data_URLs)

data_URLs2 <- paste("https://raw.githubusercontent.com/JeffSackmann/tennis_atp/master/atp_matches_",
                    1968:2021,
                    ".csv",
                    sep = "")
atpresults <- read_csv(data_URLs2)
```

[1] The WTA is the Women's Tennis Association, the organization for professional women's tennis.

[2] If Emma's ranking was high enough, she would be able to enter the tournament directly.

[3] The players with the 2 highest ranks entering the tournament are seeded #1 and #2.

[4] The biggest tournaments in tennis, namely, the Australian Open, French Open, Wimbledon, and the US Open.

[5] The ATP is the Association of Tennis Professionals, the organization for professional men's tennis

[6] SSAC15: CA - First Service: The Advent of Actionable Tennis Analytics

```r
# adding column to prepare to combine datasets
wtaresults <- wtaresults %>%
  mutate(tour = "WTA")

atpresults <- atpresults %>%
  mutate(tour = "ATP")

# moving tour column to front for ease
wtaresults <- wtaresults[,c(50,1:49)]
atpresults <- atpresults[,c(50,1:49)]

# combining the datasets
tennis_results <- rbind(wtaresults, atpresults)

# making date objects from date using lubridate
tennis_results <- tennis_results %>%
  mutate(tourney_date = ymd(tourney_date)) %>%
  mutate(year = year(tourney_date)) %>%
  mutate(month = month(tourney_date))

# reorganizing date columns together
tennis_results <- tennis_results[,c(1:7,51:52,8:50)]
```

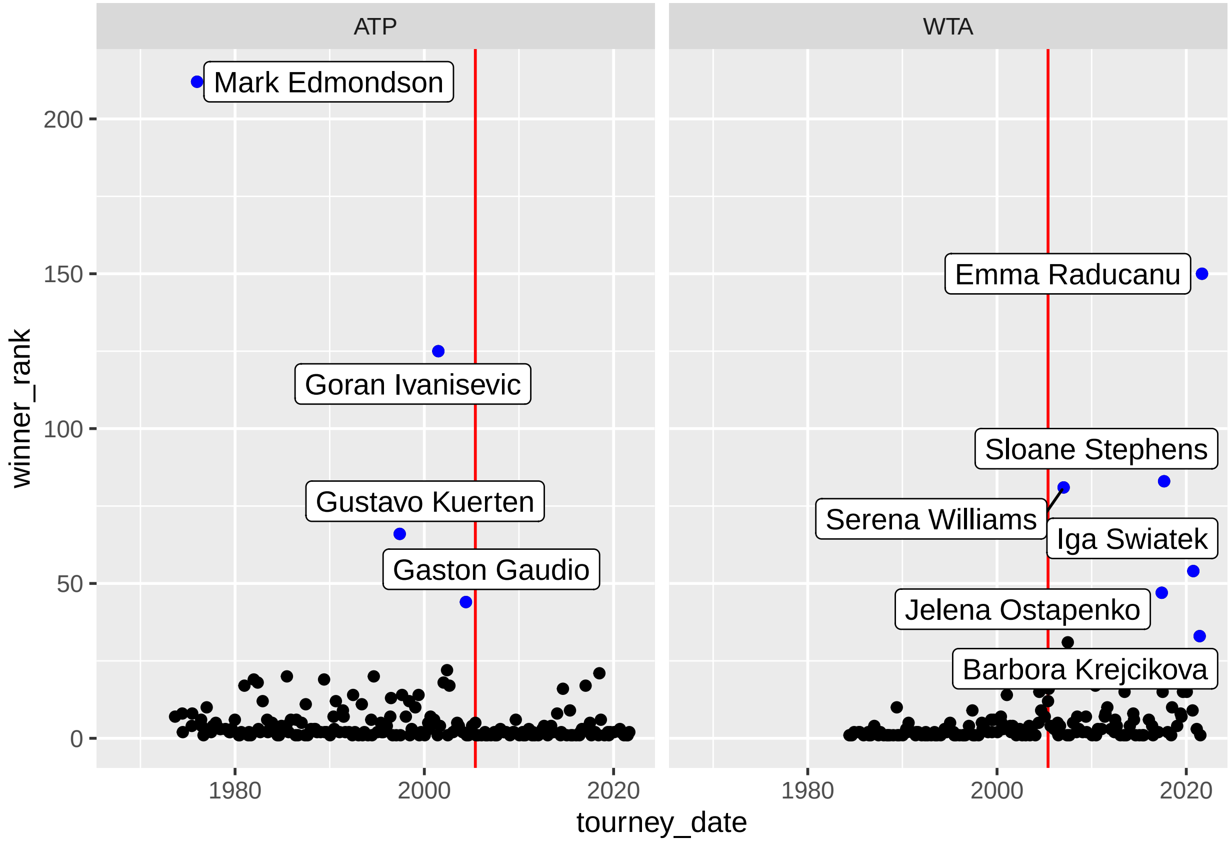
Though not included in this knitted document, we also did lots of cleaning to obtain the dataset that we wanted. We had to fix entry errors and accidental duplications, filter out results that were exhibition, junior, or challenger level matches to focus on the main professional tour, as well as mutate new variables that we needed. In addition, since each observational unit was a single tennis match, in order to compare statistics across matches (such as how well a particular player performs over all their matches), we also had to do additional wrangling to produce the desired aggregate results.

Results

To address our research question of whether or not the WTA tour is more inconsistent than the ATP tour, we first had to figure out how to measure this concept of "(in)consistency." In the end, we decided on what we believe to be three of the most public-facing measures of consistency: 1. the performance of top players at the biggest tournaments (the Grand Slam tournaments), 2. playing style and how this compares across the tours, and 3. the number of unique players who achieve top ranks per year.

Top Player Performance (Permutation Test)

```r
rank_point_results <- tennis_results %>%
  filter(round == "F", tourney_level == "G")

highlight_rank_point <- tennis_results %>%
  filter(round == "F", tourney_level == "G", winner_rank > 32)

rank_point_results %>%
  ggplot(aes(x = tourney_date, y = winner_rank)) +
  geom_vline(xintercept = rank_point_results$tourney_date[150], # split at 2005
             color = "red") +
  geom_point() +
  geom_point(data = highlight_rank_point, color = "blue") +
  facet_grid(~ tour) +
  geom_label_repel(data = highlight_rank_point, aes(label = winner_name))
```

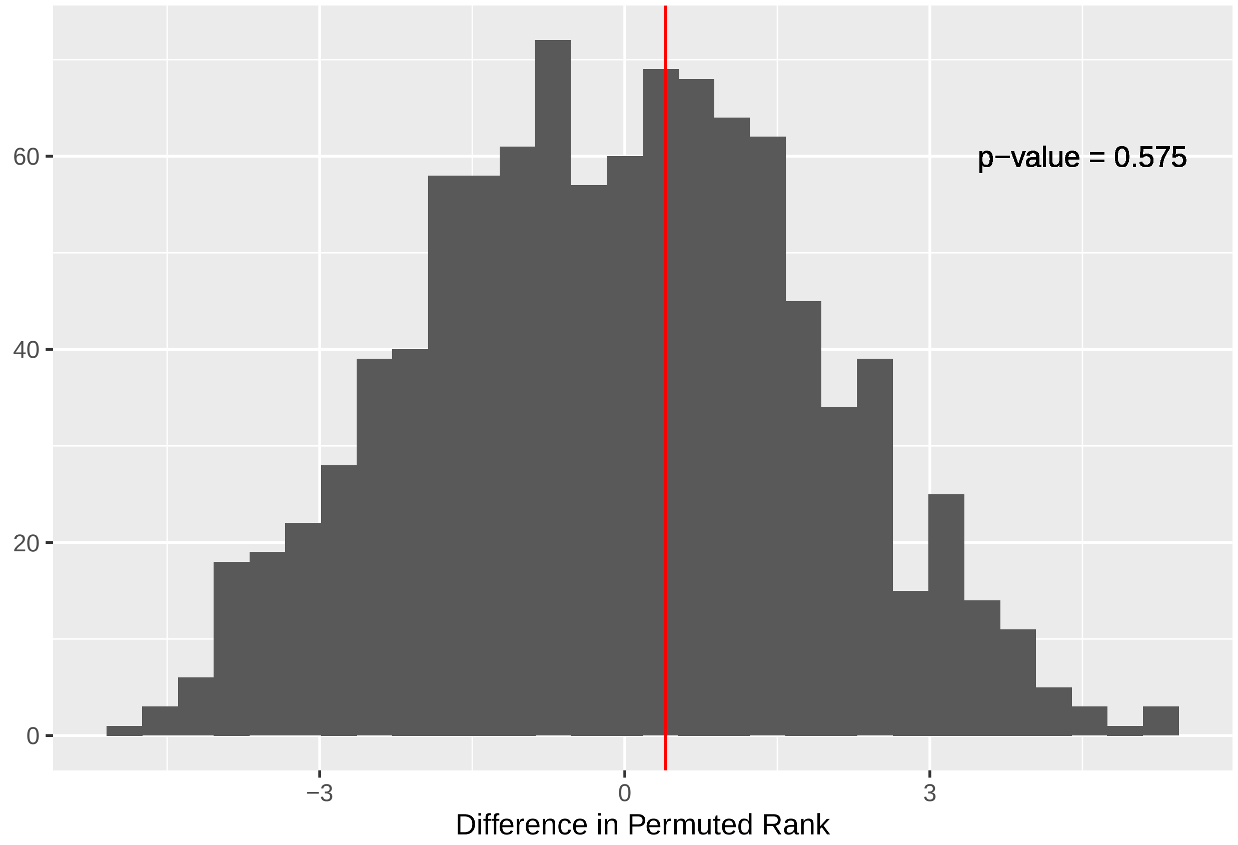


After filtering through our dataset to highlight the winners of the Grand Slam tournaments whose ranks were greater than 32 as our outliers, we made this graph to visualize the rank of Grand Slam winners over time, split by the two tours. Most of the winners, as expected, are very highly ranked, which posits them lower along the y-axis. We take note that all of the men's outliers occurred before 2005 (marked by the red lines), while all of the women's outliers occurred after 2005.

We decided to perform a permutation test by permuting the ranks of Grand Slam winners across the tours. Our null hypothesis, $H_0 : \mu_{ATP} = \mu_{WTA}$, is that there is no difference in average winner rank across the tours, so our observation shouldn't be too extreme if we permute the ranks across everyone and simulate a distribution of the differences between the tours:

```
diff_rank_func <- function(.x){
  rank_point_results %>%
  filter(!is.na(winner_rank)) %>%
  mutate(permrank = sample(winner_rank, replace = FALSE)) %>%
  group_by(tour) %>%
  summarize(avg_permrank = mean(permrank),
            avg_rank = mean(winner_rank)) %>%
  summarize(diff_permrank = diff(avg_permrank),
            diff_rank = diff(avg_rank))
  }
```

```
set.seed(47)
perm_diff_rank <- map_df(1:1000, diff_rank_func)
perm_diff_rank %>%
  ggplot() +
  geom_histogram(aes(x = diff_permrank)) +
  geom_vline(aes(xintercept = diff_rank), color = "red") +
  geom_text(aes(x = 4.5, y = 60, label = "p-value = 0.575")) +
  xlab("Difference in Permuted Rank") +
  theme(axis.title.y = element_blank())
```



```
perm_diff_rank %>%
  summarize(pval = sum(diff_rank > diff_permrank) / 1000)
```

```
## # A tibble: 1 x 1
##    pval
##   <dbl>
## 1 0.589
```

And it turns out that we get a big p-value of 0.575, which means we fail to reject the null hypothesis; there is no significant difference in average player rank of Grand Slam winners across the ATP and WTA tours.

Player Style Comparison (Clustering)

Next, we looked at the playing style/player type across the WTA and ATP tours. We performed k-means clustering in order to explore the different groups of players. We were inspired by an ATP cluster analysis that was performed on the same dataset using the results from 2011-2021 (Austin, 2021). We followed the steps there in order to replicate the results for both the ATP and WTA. For concision purposes, we only provided the code for the WTA cluster analysis; the ATP cluster analysis follows the exact same code but with the data filtered to include only ATP matches.

First, we had to replace data entries coded as 0 for number of service games (`w_SvGms` and `l_SvGms`) and length of the match in minutes (`minutes`) with NAs.[7] Realistically, these variables could almost never return a value of 0, unless one player retired from the match within one minute of the start. So, we applied a function that replaces 0 with NAs (Tierney et al., 2021).

```
tennis_results <- tennis_results %>%
  replace_with_na(replace = list(w_SvGms = 0, l_SvGms = 0, minutes = 0))
```

Next, we selected matches from 2011 onward and created the statistics that the original cluster analysis used as input variables in the unsupervised learning.

```
playerstyle_WTA_cluster <- tennis_results %>%
  filter(year >= 2011, tour == "WTA", !is.na(w_1stIn), !is.na(w_svpt), !is.na(l_1stIn), !is.na(l_svpt),
  mutate(w_1stsvpct = w_1stIn/w_svpt,
         l_1stsvpct = l_1stIn/l_svpt,
         w_svpctWon = (w_1stWon+w_2ndWon)/w_svpt,
         l_svpctWon = (l_1stWon+l_2ndWon)/l_svpt,
         w_1stsvWon = w_1stWon/w_1stIn,
         l_1stsvWon = l_1stWon/l_1stIn,
         w_2ndsvWon = w_2ndWon/(w_svpt-w_1stIn),
         l_2ndsvWon = l_2ndWon/(l_svpt-l_1stIn),
         w_acepct = w_ace/w_svpt,
         l_acepct = l_ace/l_svpt,
         w_dfpct = w_df/w_svpt,
         l_dfpct = l_df/l_svpt,
         w_ptspersvgame = w_svpt/w_SvGms,
         l_ptspersvgame = l_svpt/l_SvGms,
         w_bpSavepct = w_bpSaved/w_bpFaced,
         l_bpSavepct = l_bpSaved/l_bpFaced,
         w_bppersvgame = w_bpFaced/w_SvGms,
         l_bppersvgame = l_bpFaced/l_SvGms,
         w_pctptWon = (w_1stWon+w_2ndWon+(l_svpt-l_1stWon-l_2ndWon))/(w_svpt+l_svpt),
         l_pctptWon = (l_1stWon+l_2ndWon+(w_svpt-w_1stWon-w_2ndWon))/(w_svpt+l_svpt),
         w_1stretWon = (l_1stIn-l_1stWon)/l_1stIn,
         l_1stretWon = (w_1stIn-w_1stWon)/w_1stIn,
         w_2ndretWon = (l_svpt-l_1stIn-l_2ndWon)/(l_svpt-l_1stIn),
         l_2ndretWon = (w_svpt-w_1stIn-w_2ndWon)/(w_svpt-w_1stIn),
         w_retpctWon = 1-(l_1stWon+l_2ndWon)/l_svpt,
         l_retpctWon = 1-(w_1stWon+w_2ndWon)/w_svpt,
         w_ptsperretgame = l_svpt/l_SvGms,
         l_ptsperretgame = w_svpt/w_SvGms,
         w_bpConvpct = 1-l_bpSaved/l_bpFaced,
         l_bpConvpct = 1-w_bpSaved/w_bpFaced,
         w_bpperretgame = l_bpFaced/l_SvGms,
         l_bpperretgame = w_bpFaced/w_SvGms,
         w_retace = l_ace/l_svpt,
         l_retace = w_ace/w_svpt,
```

[7] Feel free to explore our glossary for a more in-depth explanation of these variables.

```
        w_retdf = l_df/l_svpt,
        l_retdf = w_df/w_svpt,
        ptspermin = (w_svpt+l_svpt)/minutes) %>%
  select(1:30, 49:90)
```

We then created two rows for each match, one for the winner and one for the loser, in order to properly aggregate match statistics for every player across this time span. We renamed each of the columns so that they do not specify whether the statistics apply to the winner and loser; the two datasets can now be run through `rbind`, which combines the datasets together, to create the master dataset for match statistics.

```
cluster_w <- playerstyle_WTA_cluster %>%
  select(tour, tourney_id, tourney_name, surface, tourney_date, year, month, winner_id, winner_name, wi
  mutate(result = 1)

colnames(cluster_w) <- c("tour", "tourney_id", "tourney_name", "surface", "tourney_date", "year", "mont

cluster_l <- playerstyle_WTA_cluster %>%
  select(tour, tourney_id, tourney_name, surface, tourney_date, year, month, loser_id, loser_name, lose
  mutate(result = 0)

colnames(cluster_l) <- c("tour", "tourney_id", "tourney_name", "surface", "tourney_date", "year", "mont

final_cluster_data <- rbind(cluster_w,cluster_l)
```

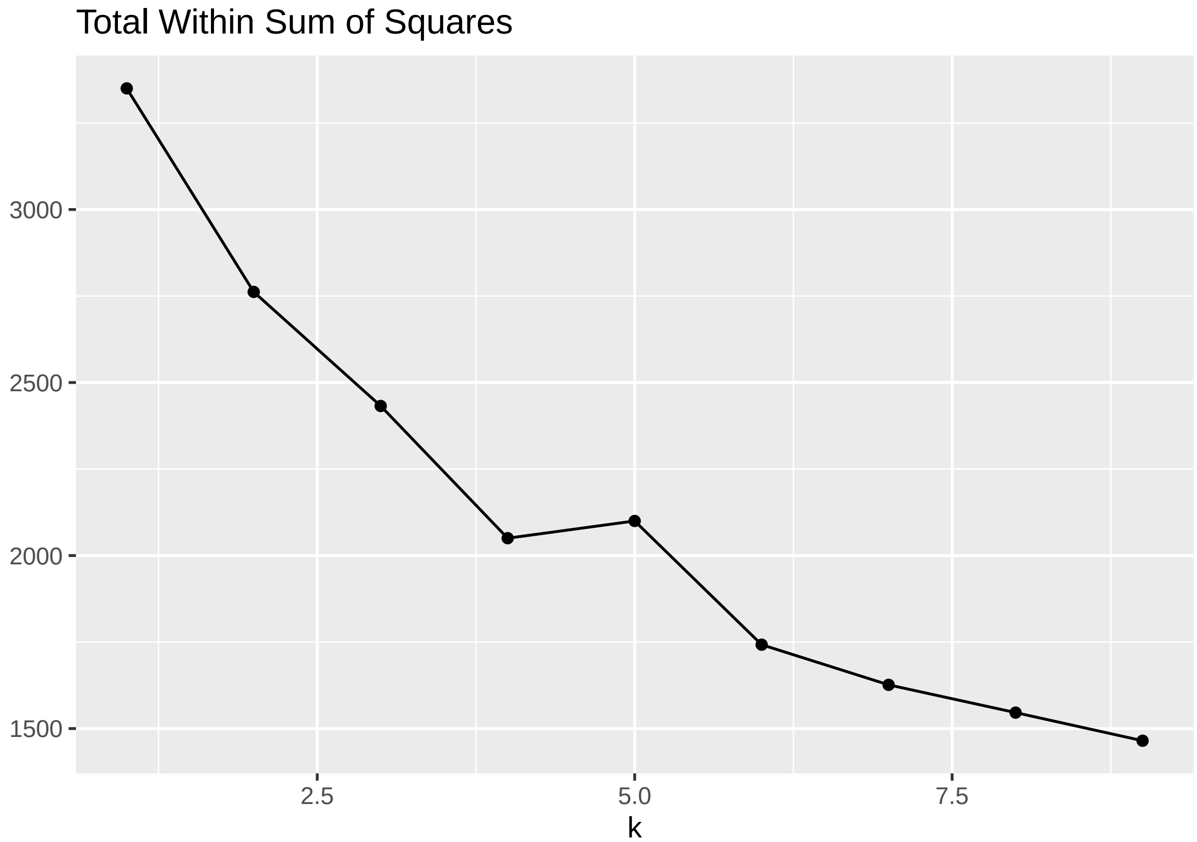
Match statistics are then calculated for every unique player in the dataset. As many individual matches did not log every match statistic, we had to specify our calculations to ignore the cells that were coded as NA. Finally, we also had to wrangle three new variables about the percentage of tournaments that they play on each surface (i.e. clay, grass, and hard), for a total of 25 variables heading into our analysis.

```
WTA_player_stats <- final_cluster_data %>%
  group_by(id, name) %>%
  summarize(height = mean(height),
            age = max(age),
            win_perc = mean(result),
            perc_points_won = mean(pctptWon, na.rm = TRUE),
            "1st_serv_perc" = mean(`1stsvpct`, na.rm = TRUE),
            "1st_win" = mean(`1stsvWon`, na.rm = TRUE),
            ace_perc = mean(acepct, na.rm = TRUE),
            df_perc = mean(dfpct, na.rm = TRUE),
            "2nd_win" = mean(`2ndsvWon`, na.rm = TRUE),
            svc_perc_win = mean(svpctWon, na.rm = TRUE),
            points_per_svc_game = mean(ptspersvgame, na.rm = TRUE),
            break_point_save_perc = mean(bpSavepct, na.rm = TRUE),
            bp_per_game = mean(bppersvgame, na.rm = TRUE),
            return_1st_win = mean(`1stretWon`, na.rm = TRUE),
            return_ace_perc = mean(retace, na.rm = TRUE),
            return_df_perc = mean(retdf, na.rm = TRUE),
            return_2nd_win = mean(`2ndretWon`, na.rm = TRUE),
            return_perc_win = mean(retpctWon, na.rm = TRUE),
            points_per_return_game = mean(ptsperretgame, na.rm = TRUE),
            bp_convert_perc = mean(bpConvpct, na.rm = TRUE),
            return_bp_per_game = mean(bpperretgame, na.rm = TRUE),
            points_per_minute = mean(ptspermin, na.rm = TRUE))
```

```r
surface_stats <- final_cluster_data %>%
  group_by(id, name, surface) %>%
  summarize(count = n()) %>%
  mutate(freq = count / sum(count)) %>%
  pivot_wider(id_cols = c(id, name), names_from = surface, values_from = freq) %>%
  summarize(clay_perc = Clay, grass_perc = Grass, hard_perc = Hard) %>%
  select(-2)

final_WTA <- cbind(surface_stats, WTA_player_stats) %>%
  select(-1) %>%
  rename(id = id...5) %>%
  select(4:7, 1:3, 8:27)
```

Finally, we performed k-means cluster analysis following example code (Hardin, 2021). After dropping all players in our dataset who still had values of NA within at least one of their match statistics (meaning there were no matches for that player in the original dataset that collected that match statistic), we were left with 135 WTA players and 297 ATP players. First, we examined an elbow plot in order to determine the best number of clusters (k-value) for our analysis. Considering that we are performing a replication, we also kept in mind the fact that the original analysis used $k = 4$.

```r
final_WTA_km <- final_WTA %>%
  drop_na() %>%
  select(height:points_per_minute) %>%
  mutate(across(height:points_per_minute, scale))

set.seed(13)
final_WTA_kclusts <-
  tibble(k = 1:9) %>%
  mutate(final_WTA_kclust = map(k, ~kmeans(final_WTA_km, .x)),
    glanced = map(final_WTA_kclust, glance),
    tidied = map(final_WTA_kclust, tidy),
    augmented = map(final_WTA_kclust, augment, final_WTA_km)
  )

clusters <-
  final_WTA_kclusts %>%
  unnest(cols = c(tidied))

assignments <-
  final_WTA_kclusts %>%
  unnest(cols = c(augmented))

clusterings <-
  final_WTA_kclusts %>%
  unnest(cols = c(glanced))

clusterings %>%
  ggplot(aes(x = k, y = tot.withinss)) +
  geom_line() +
  geom_point() + ylab("") +
  ggtitle("Total Within Sum of Squares")
```



From the elbow plot, we can see that $k = 4$ is also the choice that we should make for the WTA cluster analysis. After assigning a cluster label to each observation in the dataset, we summarized the means of each of the 25 variables in order to get a sense of how each cluster differentiates from each other (Cheng, 2020). We then created a bar graph for each variable, where each of the 4 clusters on the WTA and ATP tours (for a total of 8 clusters) could be compared. Keep in mind that tennis is a game of small percentages, so a 1-2% difference in, for example, percentage of first serve points won, is extremely significant The clusters are described below:

1. **Top Players** — The group of players that win the most matches and have some of the highest match statistics (i.e. serve and return) across the board. (e.g. Naomi Osaka, Serena Williams)
2. **Tier 2** — The group of players below the top players who are solid tour mainstays but aren't quite as good as the top players.
3. **Between Tours** — The group of players who play a significant number of matches at both the challenger[8] and main tour levels OR have retired at some point during the time span of the dataset used for analysis (2011-2021). They have won less matches than the other clusters.
4. Strong Servers (ATP)/Strong Returners (WTA)

- **Strong Servers** (ATP) — The group of ATP players who are taller and have extremely strong service statistics, even relative to the top players, and comparably weak return statistics.
- **Strong Returners** (WTA) — The group of WTA players who are shorter and have extremely strong return statistics, even relative to the top players, and comparably weak serve statistics[9]

[8] The level of professional tournaments below the main tour, organized by a different tennis organization: the International Tennis Federation (ITF).

[9] These results generally hold when we run the cluster analysis multiple times.

There are several other interesting trends that you can explore at our Shiny app[10].

The main takeaway from our cluster analysis is the exploratory difference that we see between the fourth cluster of players on the ATP and WTA tour: the strong servers (ATP) and the strong returners (WTA). The difference in strength on serve for men and strength on return for women extends beyond just that cluster; across every cluster and almost every single serve statistic (e.g. percentage of first serve points won, ace percentage, percentage of 2nd serve points won, percentage of service points won, and percentage of break points saved),[11] the ATP is uniformly stronger on serve than the WTA. Conversely, the WTA is uniformly stronger on return than the ATP.

However, the game of tennis has a serving bias. Much like baseball where the pitch is the only play in the game that a player has complete control over, the serve is the only shot in tennis that a player has complete control over. So, how well a player can control their service games is often a marker of the quality of a tennis match. Since serve and return are directly connected, it's difficult to disentangle whether, for example, the ATP has stronger serve stats because they have weaker return stats and vice versa. This serving bias could influence the perception of female players as more inconsistent, less stable, and lower quality, simply because they have worse serving statistics than male players.

In fact, there is even a somewhat derisive term, "break[12] fest," that describes a tennis match riddled with breaks of serve (i.e. the players' inability to win their service games). For example, in writing about the clash between Sebastian Korda and Karen Khachanov at the 2021 Wimbledon Championships, tennis writer Steve Tignor labeled the 5th set of the match as a "break fest" (2021). He wrote with the expectation that holding serve was what one *expects* from a (men's) tennis match, and the "surprising" change in the 5th set was due to nerves.

The takeaway here is that women's tennis is mocked for their inability to hold their service games, but men's tennis lacks the same level of disdain for their inability to win their return games. Exploring why this double standard exists is a ripe area for future research. Does the root of the serving bias come from the fact that men had access to the game for far longer? Does the foundation of a game have an unwitting bias towards men, who are typically taller and therefore, more likely to have stronger service games? One might think that how the early norms of the games were established would continue to have influence on the present perception of the game, even as women's tennis gains an increasingly equal share of the pie.

Rankings (Permutation Test)

Finally, we looked at the number of unique players who achieved top rankings per year, namely those who hit rank 1 or 2. Fortunately, there were also datasets in the same repository that contained the ranking information that we needed, so we were able to use `read_csv()` to extract these data:

```r
rank_URLs <- c("https://raw.githubusercontent.com/JeffSackmann/tennis_wta/master/wta_rankings_90s.csv",
               "https://raw.githubusercontent.com/JeffSackmann/tennis_wta/master/wta_rankings_00s.csv",
               "https://raw.githubusercontent.com/JeffSackmann/tennis_wta/master/wta_rankings_10s.csv",
               "https://raw.githubusercontent.com/JeffSackmann/tennis_wta/master/wta_rankings_20s.csv",
               "https://raw.githubusercontent.com/JeffSackmann/tennis_wta/master/wta_rankings_current.c
wtarank <- read_csv(rank_URLs)
```

```r
# 20s file is missing first row with variable names so loaded separately
atp_rankings_90s <- read_csv("https://raw.githubusercontent.com/JeffSackmann/tennis_atp/master/atp_rank
atp_rankings_00s <- read_csv("https://raw.githubusercontent.com/JeffSackmann/tennis_atp/master/atp_rank
atp_rankings_10s <- read_csv("https://raw.githubusercontent.com/JeffSackmann/tennis_atp/master/atp_rank
```

[10] A work in progress, our learnings and process for the Shiny app are described later in the section "Going Beyond Math 154." The link is here (though it takes a while to load): https://alanyiyu.shinyapps.io/shiny/

[11] See comparable variables in glossary for more in-depth explanations of these variables.

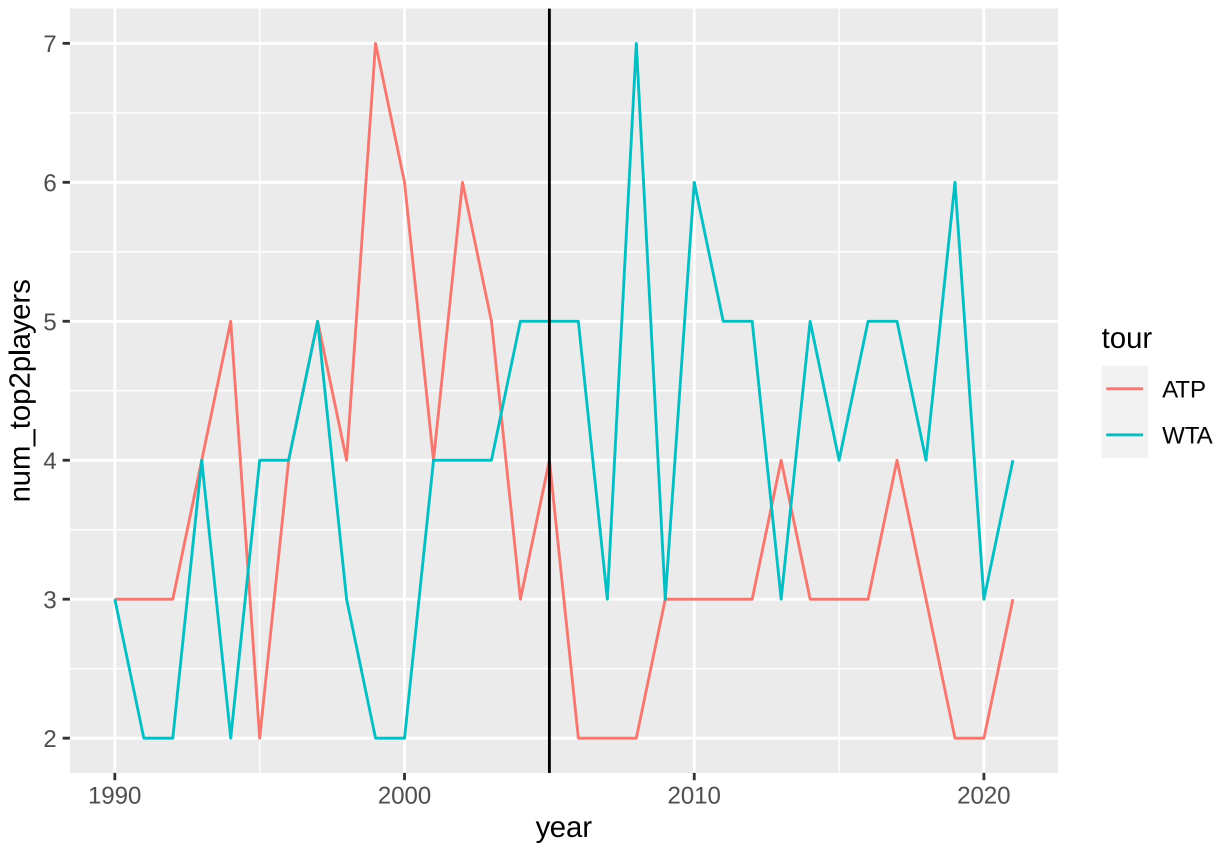
[12] A break in tennis is when the server loses their service game (or the returner wins their return game/opponent's service game.

```
atp_rankings_20s <- read_csv("https://raw.githubusercontent.com/JeffSackmann/tennis_atp/master/atp_ranki
    col_names = FALSE)
colnames(atp_rankings_20s) <- c("ranking_date", "rank", "player", "points")
atp_rankings_current <- read_csv("https://raw.githubusercontent.com/JeffSackmann/tennis_atp/master/atp_r
atprank <- rbind(atp_rankings_90s,
                 atp_rankings_00s,
                 atp_rankings_10s,
                 atp_rankings_20s,
                 atp_rankings_current)
```

```
library(tidyverse)
# adding column to prepare to combine datasets
wtarank <- wtarank %>%
  mutate(tour = "WTA")

atprank <- atprank %>%
  mutate(tours = NA, tour = "ATP")

# moving tour column to front for ease
wtarank <- wtarank[,c(6,5,1:4)]
atprank <- atprank[,c(5,6,1:4)]

# combining the datasets
tennis_rankings <- rbind(wtarank, atprank)

# making date objects from date
library(lubridate)
tennis_rankings <- tennis_rankings %>%
  mutate(ranking_date = ymd(ranking_date)) %>%
  mutate(year = year(ranking_date)) %>%
  mutate(month = month(ranking_date)) %>%
  mutate(week = week(ranking_date))

# reorganizing date columns together
tennis_rankings <- tennis_rankings[,c(1:3,7:9,4:6)]
```

Then, we made a graphic to show how many different players make it to the top 2 ranks from 1990 to today.

```
tennis_rankings %>%
  group_by(tour, year) %>%
  filter(rank <= 2) %>%
  summarize(num_top2players = n_distinct(player)) %>%
  ggplot(aes(x = year, y = num_top2players)) +
  geom_line(aes(color = tour)) +
  geom_vline(xintercept = 2005)
```



This statistic has been fluctuating more for the WTA since 2005, while it fluctuated more for the ATP before 2005. This is very similar to the trends that we saw before in the graph of the player rank of Grand Slam winners, as well as in the graph of ranking dominance. We can calculate the difference in the average number of players who reach the top 2 before and after 2005 by tour:

```
tennis_rankings %>%
  group_by(tour, year) %>%
  filter(rank <= 2, year <= 2005) %>%
  summarize(num_top2players = n_distinct(player)) %>%
  summarize(mean_top2players = mean(num_top2players))
```

```
## # A tibble: 2 x 2
##   tour  mean_top2players
##   <chr>            <dbl>
## 1 ATP               4.25
## 2 WTA               3.44
```

```
tennis_rankings %>%
  group_by(tour, year) %>%
  filter(rank <= 2, year > 2005) %>%
  summarize(num_top2players = n_distinct(player)) %>%
  summarize(mean_top2players = mean(num_top2players))
```

```
## # A tibble: 2 x 2
##   tour  mean_top2players
##   <chr>            <dbl>
## 1 ATP               2.81
## 2 WTA               4.56
```

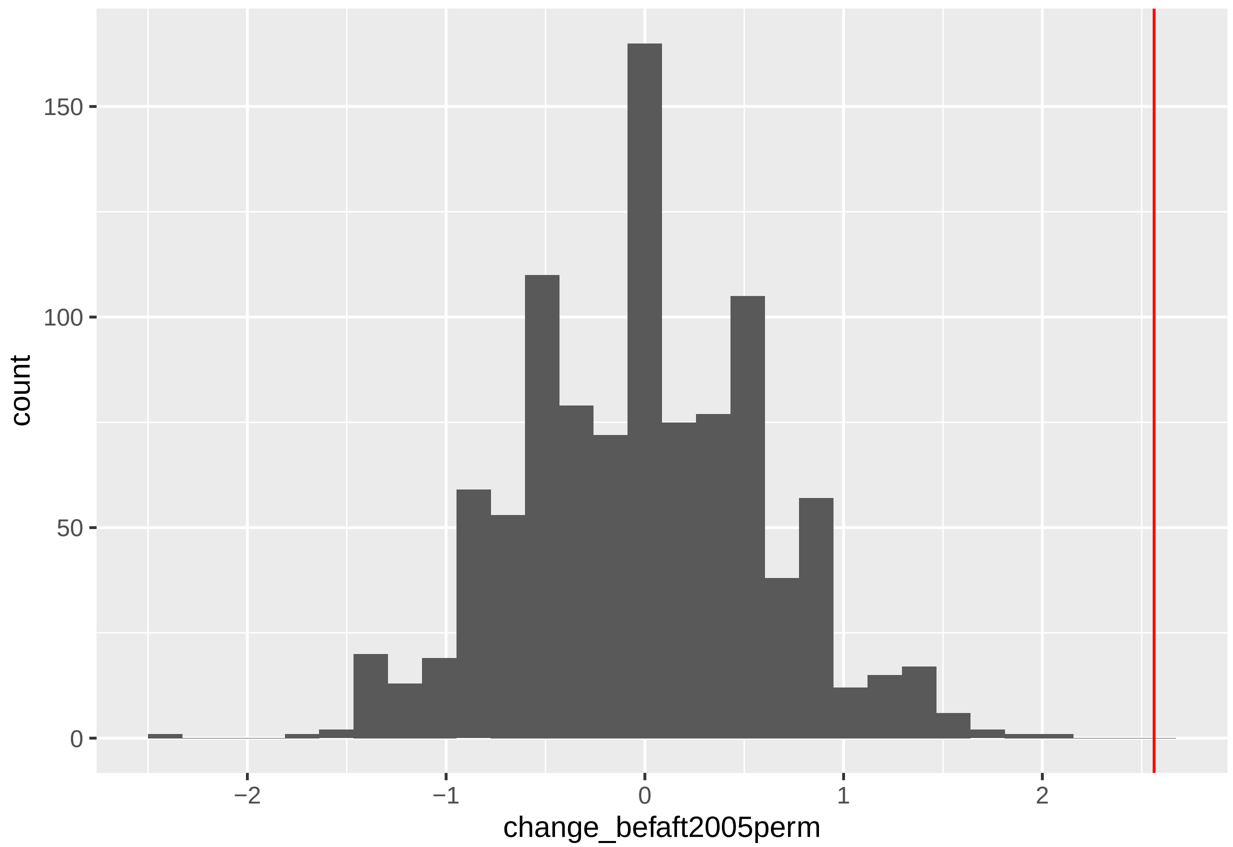
```
differences <- matrix(c(3.4375, 4.5625, 1.1250,
                        4.2500, 2.8125, -1.4375,
                        -0.8125, 1.7500, 2.5625),
                      ncol=3,byrow=TRUE)
colnames(differences) <- c("Before 2005","After 2005","Difference")
rownames(differences) <- c("ATP","WTA","Difference")
differences <- as.table(differences)
differences
```

```
##             Before 2005 After 2005 Difference
## ATP              3.4375     4.5625     1.1250
## WTA              4.2500     2.8125    -1.4375
## Difference      -0.8125     1.7500     2.5625
```

Before 2005, there was a $3.4375 - 4.25 = -0.8125$ difference in the tours, meaning on average, there were 0.8125 less players who reached the top 2 ranks in the WTA than in the ATP. However, after 2005, this difference became $4.5625 - 2.8125 = 1.75$, meaning that on average, the WTA had 1.75 players more than the ATP who reached the top 2 ranks. The observed change in differences before and after 2005 between the two tours is 2.5625 players, which means that the average number of unique top 2 players per year has increased in the WTA, while that of the ATP has decreased.

To wrap it up, we performed another permutation test:

```
top2byyear <- tennis_rankings %>%
  group_by(tour, year) %>%
  filter(rank <= 2)

top2_rank_func <- function(.x){
  top2byyear %>%
    summarize(num_top2players = n_distinct(player)) %>%
    mutate(bef_2005 = ifelse(year <= 2005, 1, 0)) %>%
    group_by(tour) %>%
    mutate(num_top2perm = sample(num_top2players, replace = FALSE)) %>%
    group_by(tour, bef_2005) %>%
    summarize(avg_bef2005 = mean(num_top2players),
              avg_bef2005_perm = mean(num_top2perm)) %>%
    group_by(bef_2005) %>%
    summarize(diff_bef2005 = diff(avg_bef2005),
              diff_bef2005perm = diff(avg_bef2005_perm)) %>%
    summarize(change_befaft2005 = diff_bef2005[1]-diff_bef2005[2],
              change_befaft2005perm = diff_bef2005perm[1]-diff_bef2005perm[2])
}
set.seed(47)
perm_diff_top2 <- map_df(1:1000, top2_rank_func)

perm_diff_top2 %>%
  ggplot() +
  geom_histogram(aes(x = change_befaft2005perm)) +
  geom_vline(aes(xintercept = change_befaft2005), color = "red")
```



```
perm_diff_top2 %>%
  summarize(pval = 1-sum(abs(change_befaft2005) > change_befaft2005perm) / 1000)
```

```
## # A tibble: 1 x 1
##    pval
##   <dbl>
## 1     0
```

By creating a simulated null distribution (where $H_0 : \mu_{difference<2005} = \mu_{difference>2005}$) through permuting the number of unique players who reach the top 2 ranks across players in the same tour, we found that the p-value was 0, meaning that our observed change in difference of 2.5625 players was significant and that we reject our null hypothesis.

Conclusions

Bringing it back to the 2021 US Open, Raducanu and Medvedev seem like fitting winners considering the trends that we saw in our data. For the past 10-15 years, the WTA has indeed been "less consistent" according to our analysis of Grand Slam winner rank and the average number of unique top 2 players in a calendar year. Raducanu was one outlier in what is an emerging trend of unseeded winners of Grand Slams on the women's side in the past few years; it would not even be that surprising if we had another "surprise" Grand Slam winner in the next year or two.

Meanwhile, the recent decade of "consistency" as seen in the ATP can be attributed to the longevity of the current generation of top players, namely the Big 4 (Nadal, Federer, Djokovic, and Murray). Whether this is the natural cycling of new and old players or a more permanent trend lasting into the future, the Big 4 is in fact nearing retirement age; maybe the ATP will follow suit and show more unpredictable results once the last of the current Big 4 retires, and perhaps Medvedev is a sign of the changing of guards in men's tennis.

Limitations and Future Directions

As the owner and maintainer of the datasets we used, Jeff Sackman says, the world of tennis statistics is mostly untapped—there are many people collecting data in some capacity and a dearth of people working with that data in sophisticated ways. When we approached the problem of inconsistency and the public perception of the difference between women's and men's tennis, we cast a wide net of exploratory data analysis. However, the analysis of consistency we continued with was fairly narrow in scope, relying heavily on ranking data as a key variable. This certainly does not capture the entire metric of consistency. Indeed, part of the limitations of our inquiry was the difficulty in fairly capturing, particularly quantitatively, the idea of consistency across a tour. Future analyses could consider measures such as the percentage of tournaments that top players win or make to the final round in — of the tournaments they play in. We also suggest exploring more sophisticated measures of consistency that aggregate data across all players, not just those at the top.

As previously mentioned, there is a double standard between the belittling of women's players for not holding their service games and a lack of criticism for men's weakness in returning. This idea exists in a broader area of further research we can glibly refer to as "Why are sports fans so sexist? And in what ways does that sexism manifest?" It would also be interesting to see if these perceptions and differences apply to the doubles'[13] tour. Oftentimes, most of the attention goes to singles' matches, even though doubles contain more "exciting" forays to the net, quick reflexes, and trick shots. Similarly, it would have been interesting to look at the challenger circuit, or the level below the main tour, where most of the players on the professional WTA and ATP circuits compete.

Finally, the dataset is humongous—we could have easily performed many more analyses. Other areas of research include looking at upsets, or lower-ranked players defeating higher-ranked players, the number of tournament wins by different levels of competitions, and win percentage. It would have also been interesting to back our claim that the mainstream perception is in fact that women's tennis is more inconsistent than men's tennis. Here, we could have performed a sentiment analysis of headlines, social media, or a different source of data where people are discussing the results of the tours.

Ethical Considerations

After running through several exploratory data analyses and technical simulations comparing the WTA and ATP tour, we noticed that the women's dataset had more missing values and less observations, meaning that less matches were played or at least recorded on the women's tour than the men's tour. It's important to note that we did not dive into the foundational history of the WTA tour and how the Original 9 fought to establish a professional women's tour after decades of growing inequality in prize money and opportunity (WTA, 2019). This disparity in prize money still exists today, and the WTA tour is still fighting the remnants of a history in the sport that has traditionally marginalized and sidelined women from headlining the same tournaments and stadiums as the ATP tour. Moving forward, we should consider how to increase the quality and quantity of data collection for women's tennis to match or exceed the level of men's tennis.

We are also very aware that any attempt to quantify differences between women's and men's sports will unavoidably be interpreted in different ways by people who hold different predispositions and established views on the subject. It seems like most people find offensive plays more riveting, which means that they would likely gravitate towards favoring strong service games rather than acknowledging the skill demanded by playing on the returning side. However, this kind of perspective unfortunately downplays the achievements and skill level of female tennis players who have outstanding return performance.

[13] A doubles match consists of two players playing on each side of the net as a team.

Going Beyond Math 154

When we began considering how to go about presenting our conclusions, we kept coming back to the narrative thread that had motivated our exploration in the first place. It felt like we should represent this in our end product, so we decided (after playing around with a few different potential Shiny layouts) to make a scrollytell — the goal being to capture the “story” through the use of a scrolling visualization. This required a few steps. First, we had to learn the basics of how to put together a Shiny app. Then, we had to figure out which of our graphics to include on the app and learn how to integrate the interactive element of updating graphs using scrolling input. Here, we leaned on a few key packages: `scrollytell` and `plotly`. Since we wanted to provide a tool to understand relationships in our cluster analysis, we created a reactive visualization that compared differences (of any given variable in the cluster analysis) between the WTA and ATP for each of the four clusters, utilizing a dropdown menu to select the desired variable. We also used interactivity to allow viewers to hover points in our graph for more details — the idea being to identify which points, particularly outliers, belong to which players. Along the way, we also built some cool animations using `gganimate`, which we showed in our presentation. Learning these tools was a lot of fun (and useful!), and we look forward to making more cool visualizations in the future.

Citations

Agence France-Presse (AFP). "Emma Raducanu reaches U.S. Open 2021 final." *iNews*, https://inews.co.uk/sport/tennis/emma-raducanu-us-open-final-maria-sakkari-new-york-leylah-fernandez-1192444

Austin, D. (2021). "ATP tennis cluster analysis." *towards data science*, https://towardsdatascience.com/atp-tennis-cluster-analysis-91bbcce61595

Chang et al. (2021). "shiny: Web Application Framework for R. R package version 1.7.1." https://CRAN.R-project.org/package=shiny

Cheng, J. (2020). "calculate the mean for each column of a matrix in R." *stack overflow*, https://stackoverflow.com/questions/21807987/calculate-the-mean-for-each-column-of-a-matrix-in-r

Loop, N. (2021). "US Open tennis 2021: Men's final winner, score and Twitter reaction." *Bleacher Report.* https://bleacherreport.com/articles/2948351-us-open-tennis-2021-mens-final-winner-score-and-twitter-reaction

Hardin, J. (2021). "R k-means example." *Computational Statistics*, http://st47s.com/Math154/Notes/unsup.html#r-k-means-example

Rothschild, C. (2019). "How to scrollytell in R." https://www.connorrothschild.com/post/automation-scrollytell/

Sackmann, J. (2021). "ATP tennis rankings, results, and stats." *GitHub.* https://github.com/JeffSackmann/tennis_atp

Sackmann, J. (2021). "WTA tennis rankings, results, and stats." *GitHub.* https://github.com/JeffSackmann/tennis_wta

Schifano, I. (2021). "Meet Emma Raducanu, the 18-year-old tennis star from London who just won the US Open." *The Tab,* https://thetab.com/uk/2021/09/13/who-is-emma-raducanu-age-instagram-net-worth-university-222598

Sievert, C. (2020). "Interactive Web-Based Data Visualization with R, plotly, and shiny." Chapman and Hall/CRC Florida.

Tierney et al. (2021). "naniar: Data Structures, Summaries, and Visualisations for Missing Data." *R package version 0.6.1.* https://CRAN.R-project.org/package=naniar

Tignor, S. (2021). "Break fest at Wimbledon: Karen Khachanov wins a doozy of a fifth set over Sebastian Korda." *tennis.* https://www.tennis.com/news/articles/break-fest-at-wimbledon-karen-khachanov-wins-a-doozy-of-a-fifth-set-over-sebasti

Wickham et al. (2019). "Welcome to the tidyverse." *Journal of Open Source Software, 4*(43), 1686, https://doi.org/10.21105/joss.01686

WTA. (2021). "Emma Raducanu — Matches." https://www.wtatennis.com/players/328366/emma-raducanu/matches

WTA. (2021). "Looking back on the Original Nine." https://www.wtatennis.com/news/1451222/looking-back-on-the-original-nine

Glossary of Variables in the Dataset

1. tourney_level

- The level of the tournament.
  - O: Olympics
  - G: Grand Slam, the four biggest tournaments of the season (Australian, French, and US Opens and Wimbledon)
  - D: Davis Cup or Fed Cup, which is country vs. country competition
  - W: all other WTA tournaments not classified into another category, up until 2000 when the tier system was introduced
  - J: juniors tournaments (for 18 under players, not professional tour) -> very few, plan to delete as well
  - E: Exhibition, an unofficial tennis match
  - F: Year-End Finals tournament with the best players, currently the format is where the top 8 players compete in Round Robin style
  - T1-5: A classification system formerly used by the WTA, Tier 1 being the biggest tournaments and Tier 5 the smallest
  - CC: The challenger circuit, the level of competition below the main WTA and ATP tours. (Only includes a tiny fraction of matches from certain years and only WTA… plan to delete these observations.)
  - I: A WTA "international" level tournament (new WTA classification system in 2009), the lowest level of main tour competition (I believe it's equivalent to Tier 4-5)
  - PM: A WTA "Premier Mandatory" level tournament, the highest level of competition just below the grand slams (equivalent to ATP Masters, former Tier 1)
  - P: A WTA "Premier" tournament, the second highest level of competition below Premier Mandatories (former Tier 2-3)
  - A: All other ATP tournaments not classified into Masters?
  - M: Masters tournaments, only an ATP classification, highest level below Grand Slams (like Premier Mandatories)

2. winner_entry and loser_entry

- How a player enters the tournament, if not directly through high ranking.
  - `ALT`: Alternative, where a player enters the tournament that has no qualifiers, usually after another competitor's withdrawal
  - `IP`: Special case for the Olympics, where the highest rank player from an underrepresented country qualifies for the Olympics
  - `LL`: Lucky loser, where a player loses in the qualifying round but then enters the main draw, usually after the withdrawal of another competitor due to injury, illness, etc.
  - `PR`: Protected ranking, where a player uses their protected ranking from the first few months of their injury or pregnancy leave to enter tournaments when making a comeback
  - `Q`: Qualifier, where a player won the qualifying match
  - `SE`: Special exempt, where a player played well in a preceding tournament that would have overlapped with the dates of the current tournament; given a special exempt to enter into the main draw without playing in the qualifying matches
  - `SR`: See `PR`
  - `WC`: Wild card, where a player is invited to participate in an event that they wouldn't have able to qualify for under normal circumstances (e.g. young players with high potential, players returning from injuries, seasoned players who enter late)

variable	class	description
tour	character	`ATP` (Association of Tennis Professionals) or `WTA` (Women's Tennis Association)
tourney_id	character	Tournament unique ID
tourney_name	character	Tournament name (e.g. US Open, Wimbledon)
surface	character	Type of court (e.g. 'Clay', 'Grass, or 'Hard')
draw_size	numeric	Total number of players in the tournament
tourney_level	character	See **Variables/Glossary** #1
tourney_date	Date	Date of tournament in the format `YYYY-MM-DD`, often the Sunday or Monday of the tournament week
year	numeric	Year of tournament
month	numeric	Month of tournament
match_num	numeric	Match number
winner_id	numeric	Player ID of match winner
winner_seed	numeric	Seed of match winner
winner_entry	character	See **Variables/Glossary** #2
winner_name	character	Name of winner
winner_hand	character	Dominant hand of winner (e.g. 'R' for right, 'L' for left, 'U' for unknown)
winner_ht	numeric	Height of winner in centimeters
winner_ioc	character	Three-letter country abbreviation of winner
winner_age	numeric	Age of winner in years
loser_id	numeric	Player ID of match loser
loser_seed	numeric	Seed of match loser
loser_entry	character	See **Variables/Glossary** #2
loser_name	character	Name of loser
loser_hand	character	Dominant hand of loser (e.g. 'R' for right, 'L' for left, 'U' for unknown)
loser_ht	numeric	Height of loser in centimeters
loser_ioc	character	Three-letter country abbreviation of loser
loser_age	numeric	Age of loser in years
score	character	Final scores of match
best_of	numeric	Number of sets for the match (e.g. '3', '5')
round	character	Round of tournament (e.g. `R16` means 16 players left, `QF` for quarterfinals, `SF` for semifinals, `F` for finals, `RR` for round robin, `BR` for bronze medal (third place))
minutes	numeric	Length of match in minutes
w_ace	numeric	Winner number of service aces (no-touch serves that win the point)
w_df	numeric	Winner number of double fault counts (missing both attempts of the serve)
w_svpt	numeric	Winner number of service points played (total points won or lost when serving)
w_1stIn	numeric	Winner number of first serves made (serves made on the first attempt)
w_1stWon	numeric	Winner number of first serve points won
w_2ndWon	numeric	Winner number of second serve points won
w_SvGms	numeric	Winner number of service games won (games in which they served)
w_bpSaved	numeric	Winner number of break points saved (when the opponent is one point away from winning your service game, and you end up taking the point)

variable	class	description
w__bpFaced	numeric	Winner number of break points faced (total number of occurrences where your opponent was one point away from winning your service game)
l__ace	numeric	Loser number of service aces
l__df	numeric	Loser number of double fault counts
l__svpt	numeric	Loser number of service points played (total points won or lost when serving)
l__1stIn	numeric	Loser number of first serves made
l__1stWon	numeric	Loser number of first serve points won
l__2ndWon	numeric	Loser number of second serve points won
l__SvGms	numeric	Loser number of service games won
l__bpSaved	numeric	Loser number of break points saved
l__bpFaced	numeric	Loser number of break points faced
winner__rank	numeric	Winner's most recent ATP or WTA rank as of the tournament date
winner__rank__points	numeric	Winner's most recent number of ranking points as of the tournament date
loser__rank	numeric	Loser's most recent ATP or WTA rank as of the tournament date
loser__rank__points	numeric	Loser's most recent number of ranking points as of the tournament date
tourney__winner	character	Final tournament winner