Data Analysis Introduction
Final Project
ATP Tennis Tournaments Analysis

Introduction:

The Association of Tennis Professionals (ATP) was formed in September 1972 to protect the interests of male professional tennis players.

Research Objectives:

In this study, we expect to analyse and answer below set of questions:

- How many ATP tournaments in each year? and which year has most ATP tournaments?
- How many of them indoor, how many of them outdoor tournaments?
- What kind of surface usually had been played on in the tournaments?
- Who is in the top 10 most successful players in the ATP history?
- Which player has most won between 4 grand slam tournaments? (aus-open, roland garros, wimbledon, us open)
- How is the competition between the big four (Nadal, Federer, Djokovic, Murray) in the grand slams?
- How is the Big Four's performances in different surfaces?
- How is the competition between Rafael Nadal and Novak Djokovic in the clay surface?
- How is the competition between Rafael Nadal and Novak Djokovic in the hard surface?

Dataset Information:

atp-world-tour-tennis-data_zip:This dataset contains tennis data from the ATP World Tour website. The data is updated annually in October. The data contains ATP tournaments, match scores, match stats, rankings and players overview. The latest available data is for 2017.

Data actively used for this project | **Tournaments.csv**: Only includes ATP tournaments-- winner names.

Data source: https://datahub.io/sports-data/atp-world-tour-tennis-data/r/0.html

Tools and Packages Used for This Analysis

R packages were used to visualize the plots in this document.

library(tidyverse)
library(plyr)
library(dplyr)
library(ggplot2)

Loading the data

We need to start with reading the csv file.

```
atp_tennis <- read.csv("tournaments.csv",  header=TRUE, stringsAsFactors=FALSE, na.strings=c(""))
dim(atp_tennis)
```

```
## [1] 4114   28
```

Data Pre-Processing

Before we start analysis, need to make sure that data is consistent, and clean.

There are some columns that not related the purpose of our analysis. For example doubles data, or the tournament and player IDs, url suffixes etc. We can start cleaning our data by removing these.

```
tennis <- atp_tennis[, !(colnames(atp_tennis) %in% c("tourney_order", "tourney_name", "tourney_id", "tourney_date
s", "tourney_month", "tourney_day", "tourney_doubles_draw", "tourney_fin_commit", "tourney_url_suffix", "singles_
winner_url", "singles_winner_player_slug", "singles_winner_player_id", "doubles_winner_1_name", "doubles_winner_1
_url", "doubles_winner_1_player_slug", "doubles_winner_1_player_id", "doubles_winner_2_name", "doubles_winner_2_u
rl", "doubles_winner_2_player_slug", "doubles_winner_2_player_id", "tourney_year_id"))]

tennis <- na.omit(tennis)

dim(tennis)
```

```
## [1] 3984    7
```

Now it’s time to make the colnames more consistent and readable.

```
colnames(tennis) <- c("year", "name", "location", "singles_draw", "conditions", "surface", "winner_name" )
colnames(tennis)
```

```
## [1] "year"          "name"          "location"      "singles_draw"
## [5] "conditions"    "surface"       "winner_name"
```

Let’s check the part of the data and see how it looks like now.

```
head(tennis)
```

```
##   year      name              location singles_draw conditions surface
## 1 1877 wimbledon  London, Great Britain           32    Outdoor   Grass
## 2 1878 wimbledon  London, Great Britain           64    Outdoor   Grass
## 3 1879 wimbledon  London, Great Britain           64    Outdoor   Grass
## 4 1880 wimbledon  London, Great Britain           64    Outdoor   Grass
## 5 1881 wimbledon  London, Great Britain           64    Outdoor   Grass
## 6 1881   us-open Newport, United States           32    Outdoor   Grass
##       winner_name
## 1    Spencer Gore
## 2     Frank Hadow
## 3    John Hartley
## 4    John Hartley
## 5 William Renshaw
## 6    William Glyn
```

For the future analysis, setting some variables as factors needed.

```
tennis$name <- as.factor(tennis$name)
tennis$location <- as.factor(tennis$locatio)
tennis$conditions <- as.factor(tennis$conditions)
tennis$surface <- as.factor(tennis$surface)
tennis$winner_name <- as.factor(tennis$winner_name)
```

Now we can use summary function to see some insights from our pre-processed data.

```
summary(tennis)
```

```
##       year                   name                          location
##  Min.   :1877   us-open        : 137   London, Great Britain  : 104
##  1st Qu.:1978   wimbledon      : 131   London                 :  96
##  Median :1990   australian-open: 104   Paris                  :  81
##  Mean   :1988   roland-garros  :  87   Sydney                 :  62
##  3rd Qu.:2002   houston        :  72   New York, United States:  57
##  Max.   :2017   kitzbuhel      :  71   Tokyo                  :  56
##                 (Other)        :3382   (Other)                :3528
##   singles_draw      conditions     surface            winner_name
##  Min.   :  4.00   Indoor :1024   Carpet: 562   Jimmy Connors : 110
##  1st Qu.: 32.00   Outdoor:2960   Clay  :1347   Ivan Lendl    :  90
##  Median : 32.00                  Grass : 565   Roger Federer :  90
##  Mean   : 46.18                  Hard  :1510   John McEnroe  :  75
##  3rd Qu.: 56.00                                Rafael Nadal  :  75
##  Max.   :256.00                                Novak Djokovic:  67
##                                                (Other)       :3477
```

Using str function will give us some more insights.

```
str(tennis)
```

```
## 'data.frame':    3984 obs. of  7 variables:
##  $ year        : int  1877 1878 1879 1880 1881 1881 1882 1882 1883 1883 ...
##  $ name        : Factor w/ 274 levels "acapulco","adelaide",..: 270 270 270 270 270 259 270 259 270 259 ...
##  $ location    : Factor w/ 406 levels "Acapulco","Acapulco, Mexico",..: 200 200 200 200 200 266 200 266 200 26
6 ...
##  $ singles_draw: int  32 64 64 64 64 32 32 64 32 32 ...
##  $ conditions  : Factor w/ 2 levels "Indoor","Outdoor": 2 2 2 2 2 2 2 2 2 2 ...
##  $ surface     : Factor w/ 4 levels "Carpet","Clay",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ winner_name : Factor w/ 623 levels "Aaron Krickstein",..: 549 172 299 299 612 610 612 100 612 255 ...
##  - attr(*, "na.action")= 'omit' Named int  269 293 298 371 372 385 401 402 415 435 ...
##   ..- attr(*, "names")= chr  "269" "293" "298" "371" ...
```

Now we are good to go for our analysis. Considering our research objectives, we will be moving forward question by question.

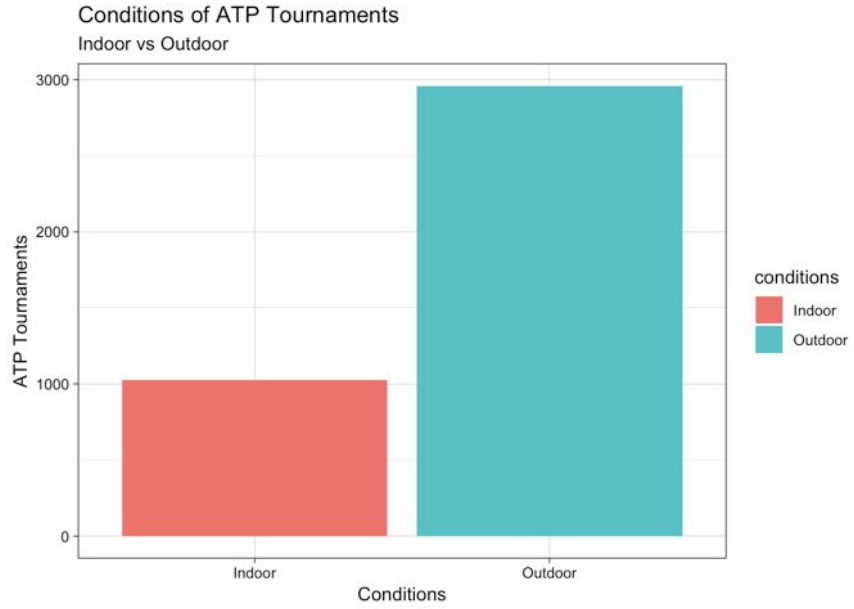
How many atp tournaments in each year? and which year has most ATP tournaments?

```
max_atp <- tennis %>%
  group_by(year) %>%
  summarize(number_of_tournaments = n()) %>%
  arrange(desc(number_of_tournaments))

head(max_atp)
```

```
## # A tibble: 6 x 2
##    year number_of_tournaments
##   <int>                 <int>
## 1  1982                   102
## 2  1976                   100
## 3  1977                   100
## 4  1980                    98
## 5  1973                    94
## 6  1974                    94
```

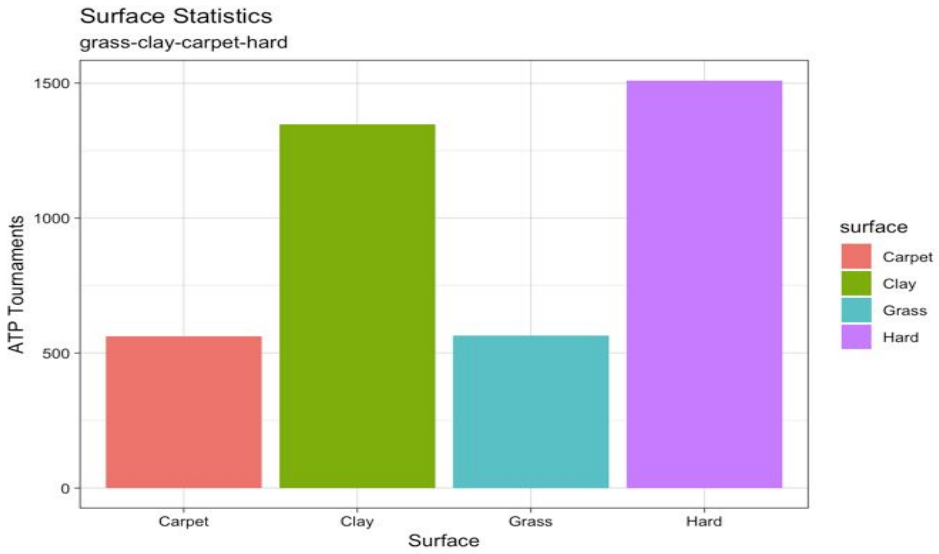
How many of them indoor, how many of them outdoor tournaments?

```
tennis %>%
  group_by(conditions) %>%
  ggplot() +
  stat_count(aes(conditions, fill = conditions)) + labs(title = "Conditions of ATP Tournaments", subtitle = "Indo
or vs Outdoor", y = "ATP Tournaments", x = "Conditions") + theme_linedraw()
```



What kind of surface usually had been played on in the tournaments?

```
tennis %>%
  group_by(surface) %>%
  ggplot() +
  stat_count(aes(surface, fill = surface)) + labs(title = "Surface Statistics", subtitle = "grass-clay-carpet-har
d", y = "ATP Tournaments", x = "Surface") + theme_linedraw()
```



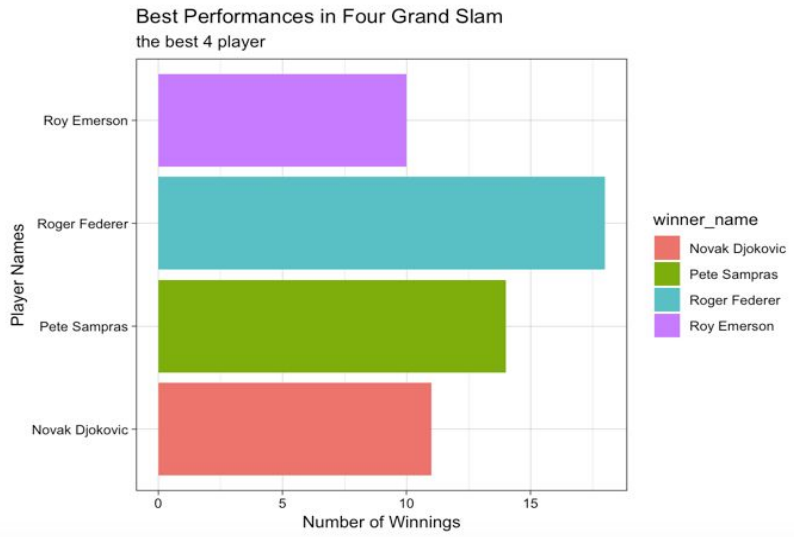
Who is in the top 10 most successful players in the ATP history.

```
top_10 <- tennis %>%
  group_by(winner_name) %>%
  summarize(number_of_wins = n()) %>%
  arrange(desc(number_of_wins)) %>%
  head(10)

top_10
```

```
## # A tibble: 10 x 2
##    winner_name     number_of_wins
##    <fct>                    <int>
##  1 Jimmy Connors              110
##  2 Ivan Lendl                  90
##  3 Roger Federer               90
##  4 John McEnroe                75
##  5 Rafael Nadal                75
##  6 Novak Djokovic              67
##  7 Bjorn Borg                  61
##  8 Guillermo Vilas             61
##  9 Andre Agassi                59
## 10 Pete Sampras                59
```

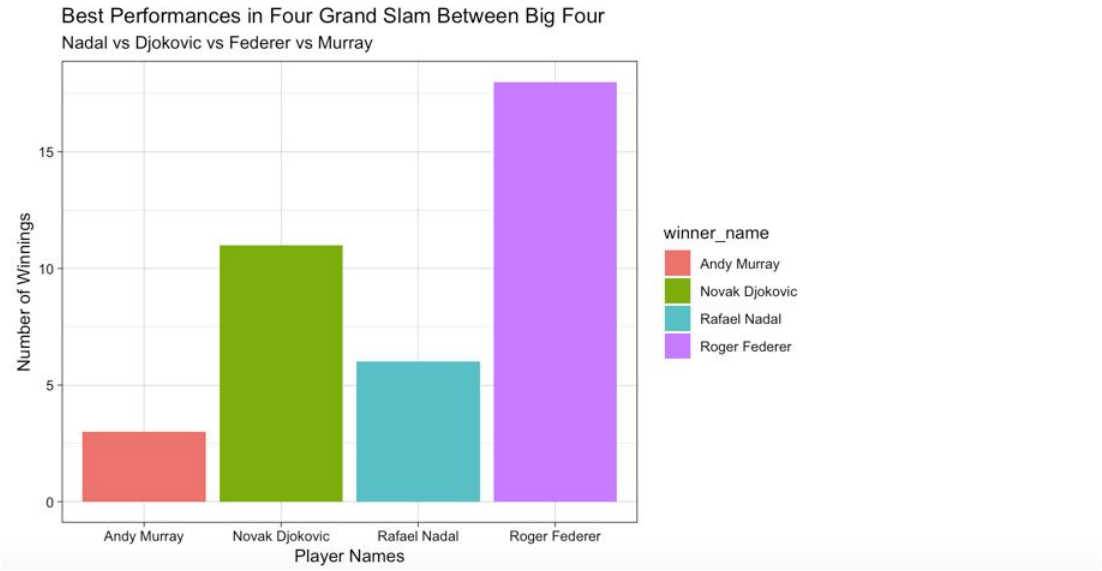
Which player has most won between 4 grand slam tournaments? (aus open, roland garros, wimbledon, us open)

```
grand_slams <- c("us-open", "wimbledon", "rolland-garros", "australian-open")

tennis %>%
  filter(name %in% grand_slams) %>%
  group_by(winner_name) %>%
  filter(n() >= 10) %>%
  ggplot() +
  geom_bar(aes(winner_name, fill = winner_name)) + coord_flip() + labs(title = "Best Performances in Four Grand S
lam", subtitle = "the best 4 player", y = "Number of Winnings", x = "Player Names") + theme_linedraw()
```



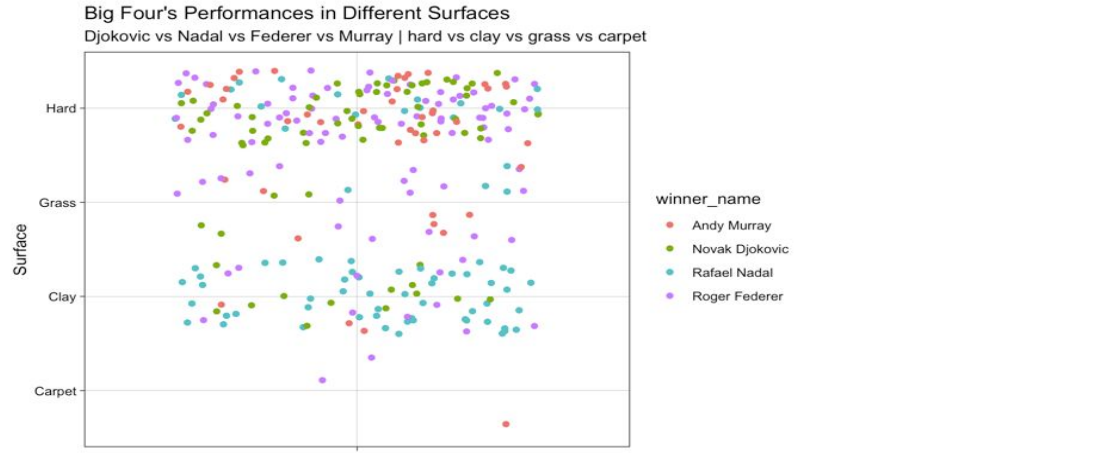
How is the competition between the big four (Nadal, Federer, Djocovic, Murray) in the grand slams?

```
big_four <- c("Novak Djokovic", "Rafael Nadal", "Roger Federer", "Andy Murray")

tennis %>%
  filter(name %in% grand_slams) %>%
  group_by(winner_name) %>%
  filter(winner_name %in% big_four) %>%
  ggplot() +
  geom_bar(aes(winner_name, fill = winner_name)) + labs(title = "Best Performances in Four Grand Slam Between Big
Four", subtitle = "Nadal vs Djokovic vs Federer vs Murray", y = "Number of Winnings", x = "Player Names") + theme
_linedraw()
```



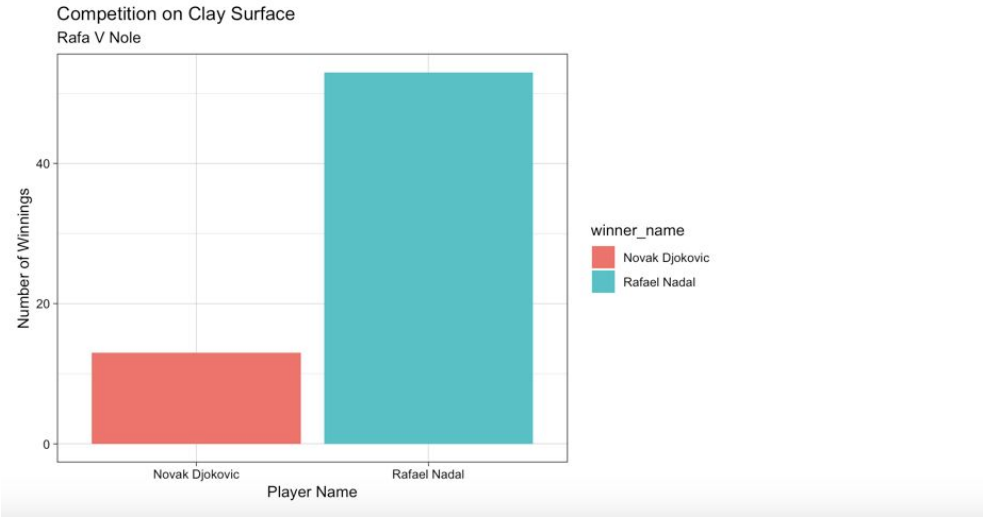
How is the Big Four's performances in different surfaces?

```
tennis %>%
  filter(winner_name %in% big_four) %>%
  group_by(surface) %>%
  ggplot(aes(x = "", y = surface)) +
  geom_jitter(aes(color = winner_name)) + labs(title = "Big Four's Performances in Different Surfaces", subtitle
= "Djokovic vs Nadal vs Federer vs Murray | hard vs clay vs grass vs carpet", x = "", y = "Surface") + theme_line
draw()
```



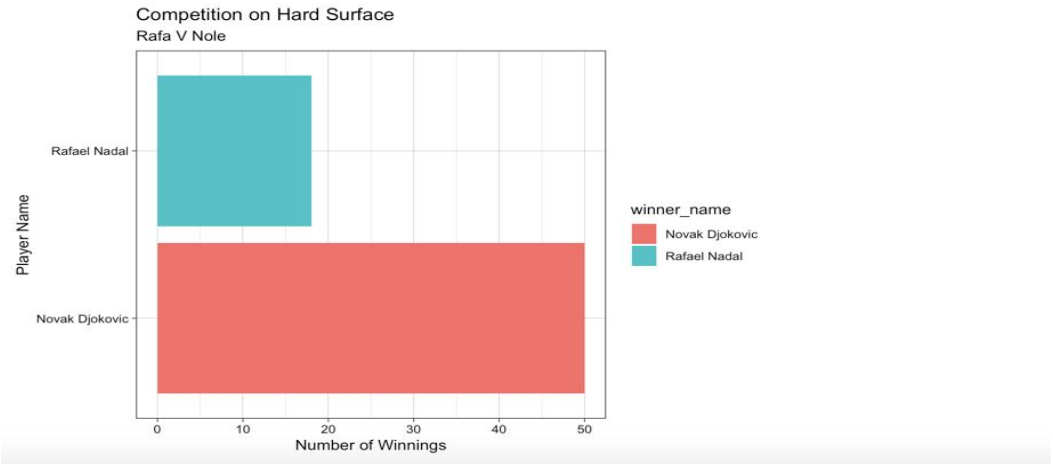
How is the competition between Rafael Nadal and Novak Djokovic in the clay surface?

```
rafa_nole <- c("Novak Djokovic", "Rafael Nadal")

tennis %>%
  filter(winner_name %in% rafa_nole) %>%
  group_by(winner_name) %>%
  filter(surface == "Clay") %>%
  ggplot() +
  geom_bar(aes(winner_name, fill = winner_name)) + labs(title = "Competition on Clay Surface", subtitle = "Rafa V
Nole", x = "Player Name", y = "Number of Winnings") + theme_linedraw()
```



How is the competition between Rafael Nadal and Novak Djokovic in the hard surface?

```
tennis %>%
  filter(winner_name %in% rafa_nole) %>%
  group_by(winner_name) %>%
  filter(surface == "Hard") %>%
  ggplot() +
  geom_bar(aes(winner_name, fill = winner_name)) + coord_flip() + labs(title = "Competition on Hard Surface", sub
title = "Rafa V Nole", x = "Player Name", y = "Number of Winnings") + theme_linedraw()
```



Conclusions

- 70’s, 80’s more tournaments was organized.
- 1982 is the year which more tournaments was organized in.
- After 1968 there is a big jump in the numbers of tournaments organized.
- In 1968 the "Open Era" began when major tournaments agreed to allow professional players to compete with amateurs.
- Most of the tournaments were organized in outdoor conditions.
- Indoors, the light is good, no wind, no sun, no temperature problems, no rain.
- That makes perfect conditions but also make the tournaments less exciting.
- Although grass courts are more traditional than other types of tennis courts, maintenance costs of grass courts are higher than those of hard courts and clay courts. Hard courts are most low maintenance of the three and hence most common across the world.
- Jimmy Connors is the most successful player in the ATP history with 110 winnings.
- Roger Federer is the most successful player within Grand Slam Tournaments.
- Between big four, Roger Federer is the most successful players within Grand Slams and Novak Djokovic is following him.
- In different surfaces, the big four has different performances in the ATP history. The most interesting insight was the data about Rafael Nadal and his success on the clay courts.
- Between Novak Djokovic and Rafael Nadal, Nadal is better than Djokovic on the clay, but Djokovic beats him on the hard courts.