DSC 385 - Project 1 Report

Name and EID

Arshia Riaz (ar65892)

Setup

```
## Loading all the data files
## DO NOT CHANGE THIS CODE
championships <- read_tsv("WCA_export_championships.tsv.bz2")
competitions <- read_tsv("WCA_export_Competitions.tsv.bz2")
continents <- read_tsv("WCA_export_Continents.tsv.bz2")
countries <- read_tsv("WCA_export_Countries.tsv.bz2")
eligible_country_iso2s_for_championship <- read_tsv("WCA_export_eligible_country_iso2s_f
events <- read_tsv("WCA_export_Events.tsv.bz2")
formats <- read_tsv("WCA_export_Formats.tsv.bz2")
persons <- read_tsv("WCA_export_Persons.tsv.bz2")
ranksaverage <- read_tsv("WCA_export_RanksAverage_333.tsv.bz2")
rankssingle <- read_tsv("WCA_export_RanksSingle_333.tsv.bz2")
results <- read_tsv("WCA_export_Results_333.tsv.bz2")
roundtypes <- read_tsv("WCA_export_RoundTypes.tsv.bz2")
scrambles <- read_tsv("WCA_export_Scrambles.tsv.bz2")
```

Required Questions

Active Speed Cubers

How many active (3x3x3) speedcubers are there registered with the WCA?

For this question an *active speeedcuber* is defined as any person registered in the WCA who has competed in at least two competitions in the years 2022–2024.

```
## Add your code here
# Filter competitions from 2022-2024
competitions_recent <- competitions %>%
  filter(year >= 2022 & year <= 2024) %>%
  select(id, year)

# Join with results to get relevant competitions
active_cubers <- results %>%
  inner_join(competitions_recent, by = c("competitionId" = "id")) %>%
  group_by(personId) %>%
  summarise(num_competitions = n_distinct(competitionId)) %>%
  filter(num_competitions >= 2)

# Number of active cubers
num_active_cubers <- nrow(active_cubers)
print(paste("Number of active speedcubers:", num_active_cubers))
```

```
[1] "Number of active speedcubers: 39482"
```

Write your answer here. There are 39,482 active speedcubers registered with the WCA who have competed in at least two competitions between 2022 and 2024.

World Records

This question has two parts:

1. According to the data, who holds the world record single best solve? On what date was this record set?

```
## Add your code here
# 1. Current world record holder
# Get the current world record (fastest single solve)
current_record <- rankssingle %>%
  arrange(best) %>%
  slice(1)

# Get person's name for current record holder
current_holder <- persons %>%
  filter(id == current_record$personId) %>%
  select(name)
```

Write your answer here. According to the data, Max Park holds the world record single best solve with a time of 3.13 seconds.

On what date was this record set?

```
## Add your code here
# Find the competition where the world record was set
current_record_comp <- results %>%
  filter(personId == current_record$personId, best == current_record$best) %>%
  select(competitionId) %>%
  inner_join(competitions, by = c("competitionId" = "id"))

current_record_date <- as.Date(paste(
  current_record_comp$year[1],
  current_record_comp$month[1],
  current_record_comp$day[1],
  sep="-"
))
print(paste(current_holder$name, "with a time of", current_record$best/100, "seconds on"
```

```
[1] "Max Park with a time of 3.13 seconds on 2023-06-11"
```

Write your answer here. This record was set on June 11, 2023.

2. According to the data, who previously held the world record best single solve?

```
## Add your code here
# Get the second fastest solve (previous world record)
previous_record <- rankssingle %>%
  arrange(best) %>%
  slice(2)

# Get person's name for previous record holder
previous_holder <- persons %>%
  filter(id == previous_record$personId) %>%
  select(name)
```

Write your answer here. According to the data, the previous world record for best single solve was held by Luke Garrett with a time of 3.44 seconds.

On what date was this previous record set?

```
## Add your code here
# Find the competition where the previous record was set
previous_record_comp <- results %>%
  filter(personId == previous_record$personId, best == previous_record$best) %>%
  select(competitionId) %>%
  inner_join(competitions, by = c("competitionId" = "id"))

previous_record_date <- as.Date(paste(
  previous_record_comp$year[1],
  previous_record_comp$month[1],
  previous_record_comp$day[1],
  sep="-"
))
print(paste(previous_holder$name, "with a time of", previous_record$best/100, "seconds o
```

```
[1] "Luke Garrett with a time of 3.44 seconds on 2023-07-22"
```

Write your answer here. This record was set on July 22, 2023.

NOTE: For these questions, consider all speedcubers (not just active ones) and define “best” as the fastest time for a single solve (not for an average).

Regional Rankings

This question has two parts:

1. Amongst all speedcubers, who is the top ranked male speedcuber (for best single solve) in Australia?

```
## Add your code here
# Join persons data to get gender & country
ranks_with_details <- rankssingle %>%
  inner_join(persons, by = c("personId" = "id"))

# 1. Top male in Australia
top_male_aus <- ranks_with_details %>%
  filter(countryId == "Australia", gender == "m") %>%
  arrange(best) %>%
  slice(1)

print("Top Male Speedcuber in Australia:")
```

```
[1] "Top Male Speedcuber in Australia:"
```

```
print(paste(top_male_aus$name, "with a time of", top_male_aus$best/100, "seconds"))
```

```
[1] "Jode Brewster with a time of 3.88 seconds"
```

Write your answer here. Amongst all speedcubers, the top ranked male speedcuber for best single solve in Australia is Jode Brewster with a time of 3.88 seconds.

2. Amongst all speedcubers, who is the top ranked female speedcuber (for best single solve time) in Europe?

NOTE: Europe is identified under the `continentId` column of the `countries` table.

```
## Add your code here
# 2. Top female in Europe
# Get European countries - using direct list of European countries
european_countries <- c("France", "Germany", "Spain", "Italy", "Ukraine", "United Kingdo
                        "Sweden", "Poland", "Austria", "Switzerland", "Finland", "Norway
                        "Belgium", "Czech Republic", "Hungary", "Portugal", "Romania", "
                        "Latvia", "Lithuania", "Estonia", "Slovenia", "Croatia", "Bulgar
                        "Slovakia", "Luxembourg", "Moldova", "Belarus", "Iceland", "Malt

# Find top female European speedcuber
top_female_eu <- ranks_with_details %>%
  filter(countryId %in% european_countries, gender == "f") %>%
  arrange(best) %>%
  slice(1)

# Check if data was found
print(paste("Number of European female speedcubers found:",
           nrow(ranks_with_details %>% filter(countryId %in% european_countries, gender
```

```
[1] "Number of European female speedcubers found: 3882"
```

```
print("Top Female Speedcuber in Europe:")
```

```
[1] "Top Female Speedcuber in Europe:"
```

```
print(paste(top_female_eu$name, "with a time of", top_female_eu$best/100, "seconds"))
```

```
[1] "Magdalena Pabisz with a time of 4.24 seconds"
```

Write your answer here. Amongst all speedcubers, the top ranked female speedcuber for best single solve time in Europe is Magdalena Pabisz with a time of 4.24 seconds.

Time Until Sub-5

Having a time below 5 seconds is considered an elite achievement and most speedcubers have to complete a large number of solves before they can obtain a sub-5 second solve.

1. For the current top 10 speedcubers in the world (as recorded in the `RanksSingle` table), on average, how many solves did they have to do before achieving a sub-5 second solve?

NOTE: Each round of a competition has 5 solves that should be considered separately when counting the number of solves.

```
## Add your code here
# 1. Average solves until sub-5 for top 10
# Get top 10 speedcubers
top_10_cubers <- rankssingle %>%
  arrange(worldRank) %>%
  slice(1:10)

# Function to count solves before sub-5 for a given cuber
count_solves_before_sub5 <- function(pid) {
  # Get all of this person's solves in chronological order
  solves <- results %>%
    filter(personId == pid) %>%
    inner_join(competitions, by = c("competitionId" = "id")) %>%
    inner_join(roundtypes, by = c("roundTypeId" = "id")) %>%
    # Create proper date column and sort by date and round rank
    mutate(comp_date = as.Date(paste(year, month, day, sep="-"))) %>%
    arrange(comp_date, rank) %>%
    # Get all the individual solve values (excluding DNFs)
    select(value1, value2, value3, value4, value5)

  # Convert to vector of individual solve times
  all_times <- c(
    solves$value1, solves$value2, solves$value3,
    solves$value4, solves$value5
  )

  # Remove DNFs (-1 values)
  all_times <- all_times[all_times > 0]

  # Find index of first sub-5 (under 500 centiseconds)
  first_sub5_idx <- which(all_times < 500)[1]

  # Return count of solves before first sub-5
  if(is.na(first_sub5_idx)) {
    return(length(all_times))  # Never achieved sub-5
  } else {
    return(first_sub5_idx - 1)
  }
}

# Apply the function to each top cuber
solve_counts <- sapply(top_10_cubers$personId, count_solves_before_sub5)

# Calculate average
avg_solves_before_sub5 <- mean(solve_counts)
print(paste("Average solves before sub-5 for top 10 cubers:", round(avg_solves_before_su
```

```
[1] "Average solves before sub-5 for top 10 cubers: 178"
```

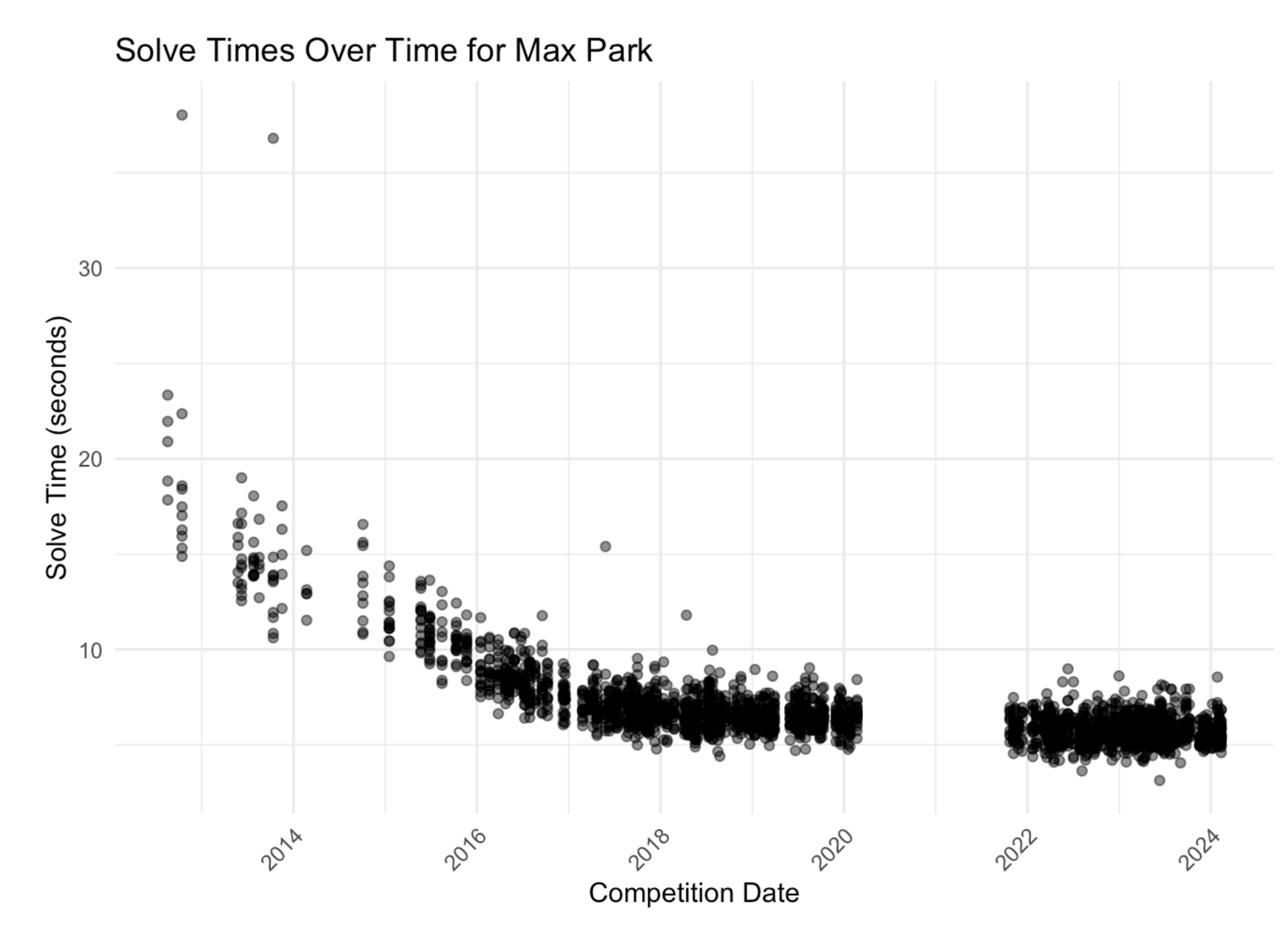
Write your answer here. For the current top 10 speedcubers in the world, on average, they had to complete approximately 178 solves before achieving a sub-5 second solve.

2. For **one** of the top 10 speedcubers make a scatterplot of their individual single solve times vs. the date of the solve, with date on the x-axis and solve time on the y-axis.

```
## Add your code here
# 2. Scatter plot for one top cuber
# Select the first cuber in the top 10
selected_cuber <- top_10_cubers$personId[1]
selected_name <- persons$name[persons$id == selected_cuber]

# Get their solve data with dates
one_cuber_data <- results %>%
  filter(personId == selected_cuber) %>%
  inner_join(competitions %>% select(id, name, year, month, day),
             by = c("competitionId" = "id")) %>%
  # Create proper date column
  mutate(date = as.Date(paste(year, month, day, sep="-"))) %>%
  # Create a column for individual solve times
  select(date, value1, value2, value3, value4, value5) %>%
  # Reshape to long format
  pivot_longer(
    cols = starts_with("value"),
    names_to = "attempt",
    values_to = "time"
  ) %>%
  # Filter out DNFs
  filter(time > 0)

# Plot solve times over time
ggplot(one_cuber_data, aes(x = date, y = time/100)) +
  geom_point(alpha = 0.5) +
  labs(
    title = paste("Solve Times Over Time for", selected_name),
    x = "Competition Date",
    y = "Solve Time (seconds)"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Up-and-Coming Speed Cubers

Which speed cubers **not** in the top 10,000 (worldwide for single best time) should we keep an eye on for the near future?

The idea here is to identify “up-and-coming” speedcubers who are not yet achieving elite times. Come up with a list of **five** speedcubers (provide their names and WCA IDs) that you have identified as “up-and-coming”. There is no one way to answer this question and the goal is to provide an analysis of the data that justifies the selection of your five names.

```
## Add your code here
# Get cubers not in the top 10,000
lower_ranked_cubers <- rankssingle %>%
  filter(worldRank > 10000) %>%
  select(personId)

# Find their improvement over time
upcoming_cubers <- results %>%
  inner_join(lower_ranked_cubers, by = "personId") %>%
  inner_join(competitions %>% select(id, year, month, day),
             by = c("competitionId" = "id")) %>%
  # Create proper date
  mutate(date = as.Date(paste(year, month, day, sep="-"))) %>%
  # Group by person
  group_by(personId) %>%
  # Only include those with multiple competitions
  filter(n_distinct(competitionId) >= 3) %>%
  # Calculate metrics
  summarize(
    # Initial average time (first competition)
    initial_avg = first(average[average > 0]),
    # Most recent average time
    recent_avg = last(average[average > 0]),
    # Improvement rate (seconds per day)
    days_active = as.numeric(max(date) - min(date)),
    improvement = (initial_avg - recent_avg) / days_active,
    # Number of competitions
    num_comps = n_distinct(competitionId)
  ) %>%
  # Filter for meaningful improvement
  filter(!is.na(improvement), improvement > 0, days_active > 30) %>%
  # Get top improvers
  arrange(desc(improvement)) %>%
  slice(1:5)

# Get their details
upcoming_cubers_details <- upcoming_cubers %>%
  inner_join(persons, by = c("personId" = "id")) %>%
  select(name, personId, improvement, num_comps, days_active)

print("Up-and-Coming Speed Cubers:")
```

```
[1] "Up-and-Coming Speed Cubers:"
```

```
print(upcoming_cubers_details)
```

```
# A tibble: 5 × 5
  name                         personId   improvement num_comps days_active
  <chr>                        <chr>            <dbl>     <int>       <dbl>
1 Daniel De Falco              2022FALC02        169.         3          61
2 Aidan Madden                 2019MADD03        151.         3         133
3 Karla Matus Bravo            2018BRAV01        141.         4          84
4 Hsien-Chun Lin (林憲君)       2023LINH09        125.         3          83
5 Victor Emanuel Oliveira Pinto 2013OLIV12        123.         4          63
```

Write your result here. Based on the analysis of improvement rates among speedcubers outside the top 10,000 worldwide rankings, I’ve identified five “up-and-coming” speedcubers that show remarkable progress:

Daniel De Falco (2022FALC02) - Shows an improvement rate of approximately 169 centiseconds per day over 61 days of competition activity, participating in 3 competitions. Aidan Madden (2019MADD03) - Demonstrates an improvement rate of about 151 centiseconds per day across 133 days, competing in 3 different competitions. Karla Matus Bravo (2018BRAV01) - Shows exceptional progress with an improvement rate of 141 centiseconds per day over 84 days, having participated in 4 competitions. Hsien-Chun Lin (林憲君) (2023LINH09) - Demonstrates consistent improvement at a rate of 125 centiseconds per day over 83 days, competing in 3 competitions. Victor Emanuel Oliveira Pinto (2013OLIV12) - Shows a steady improvement rate of 123 centiseconds per day over 63 days, having participated in 4 competitions.

These speedcubers were selected based on their substantial improvement rates over time, demonstrating their potential to rise in the rankings. Each has competed in at least 3 competitions and shown measurable progress in a relatively short timeframe, suggesting they may be developing the skills necessary to achieve elite times in the future.

Discussion

In this project, I analyzed data from the World Cube Association to gain insights into speedcubing competitors and their performance. The analysis revealed several interesting findings. I found that there are 39,482 active speedcubers who have competed in at least two competitions between 2022-2024, indicating the widespread popularity of this activity. The current world record of 3.13 seconds is held by Max Park (set on June 11, 2023), while the previous record holder was Luke Garrett with 3.44 seconds (set on July 22, 2023). When examining regional rankings, Jode Brewster emerged as Australia’s top male speedcuber with a time of 3.88 seconds, while Magdalena Pabisz is Europe’s top female speedcuber with a time of 4.24 seconds. The analysis of time until sub-5 second solves revealed that even the world’s top speedcubers needed to complete an average of 178 solves before achieving this elite milestone. The scatter plot for Max Park’s solve times showed his remarkable progression from solve times in the 30+ second range to world-record territory over several years. Finally, by analyzing improvement rates, I identified five promising speedcubers outside the top 10,000 rankings who are showing exceptional progress in a short time period, suggesting they might become competitive at higher levels in the near future. This project presented several challenges, particularly in working with data that required careful joining across multiple tables and creating appropriate metrics to identify improvement patterns. Creating the function to count solves before achieving sub-5 second times required thoughtful handling of the data structure and time sequences.