SDS 322E: Project 1 Report

```
# load tidyverse and necessary datasets
library(tidyverse)
rankssingle <- read_tsv("WCA_export_RanksSingle_333.tsv.bz2")
results <- read_tsv("WCA_export_Results_333.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")
persons <- read_tsv("WCA_export_Persons.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.

```
## create a new object 'active_speedcuber_count' that joins the 'results' and 'competitions' datasets, filter years between 2022 and 2024 and filter the unique ids that attended at least two competitions. Count the number of rows in 'active_speedcuber_count' to find the number of active speedcubers

active_speedcuber_count <-
results |>
  rename(id = competitionId) |>
  left_join(competitions, by = 'id') |>
  filter((year >= 2022) & (year <= 2024)) |>
  group_by(personName) |>
  filter(n_distinct(id) >= 2) |>
  summarize()

nrow(active_speedcuber_count)
```

```
## [1] 39438
```

There are 39,438 active (3x3x3) speedcubers registered with the WCA.

World Records

1. Who holds the current world record single? On what date was this record set?

```
## join the 'results' and 'competitions' datasets, filter out invalid best times and
world record, and arrange best to find the minimum best time in world record single a
nd the corresponding date of record set

results |>
  rename(id = competitionId) |>
  left_join(competitions, by = "id") |>
  filter((best != -1) & (regionalSingleRecord == "WR")) |>
  select(best, personName, year, month, day) |>
  group_by(best) |>
  arrange(best) |>
  head(1)
```

```
## # A tibble: 1 × 5
## # Groups:   best [1]
##    best personName  year month   day
##   <dbl> <chr>      <dbl> <dbl> <dbl>
## 1   313 Max Park    2023     6    11
```

Max Park currently holds the current world record single, and this record was set on June 11th, 2023.

2. Who *previously* held the world record single? On what date was this previous record set?

```
## join the 'results' and 'competitions' datasets, filter out invalid best times and
world record, and arrange best to find the second best time in world record single an
d the corresponding date of record set

results |>
  rename(id = competitionId) |>
  left_join(competitions, by = "id") |>
  filter((best != -1) & (regionalSingleRecord == "WR")) |>
  select(best, personName, year, month, day) |>
  group_by(best) |>
  arrange(best) |>
  head(2)
```

```
## # A tibble: 2 × 5
## # Groups:   best [2]
##    best personName           year month   day
##   <dbl> <chr>               <dbl> <dbl> <dbl>
## 1   313 Max Park             2023     6    11
## 2   347 Yusheng Du (杜宇生)  2018    11    24
```

Yusheng Du previously held the world record single, and this record was set on November 24th, 2018.

Regional Rankings

1. Who is the top ranked male speedcuber (for single best time) in Australia?

```
## join 'persons' and 'rankssingle' datasets, filter by male and Australia and arrang
e countryRank to find the top ranked male speedcuber in Australia

persons |>
  rename(personId = id) |>
  left_join(rankssingle, by = "personId") |>
  select(name, countryId, gender, countryRank) |>
  group_by(countryRank) |>
  filter(countryId == "Australia") |>
  filter(gender == "m") |>
  filter(countryRank != 0) |>
  arrange(countryRank)
```

```
## # A tibble: 8,154 × 4
## # Groups:   countryRank [4,103]
##    name              countryId gender countryRank
##    <chr>             <chr>     <chr>        <dbl>
##  1 Jode Brewster     Australia m                1
##  2 Feliks Zemdegs    Australia m                2
##  3 Riley Dexter      Australia m                3
##  4 Phoenix Patterson Australia m                4
##  5 Charlie Eggins    Australia m                5
##  6 Jayden McNeill    Australia m                6
##  7 Toby Seufert      Australia m                7
##  8 Sora Sato         Australia m                8
##  9 Tomoya Firman     Australia m                9
## 10 Bryan Eng         Australia m               10
## # ℹ 8,144 more rows
```

The top ranked male speedcuber in Australia is Jode Brewster.

2. Who is the top ranked female speedcuber (for single best time) in Europe?

```
## join 'persons', 'rankssingle', and 'continents' datasets, filter by female and Eur
ope and arrange continentRank to find the top ranked female speedcuber in Europe

continents <- countries |>
  rename(countryId = id)

persons |>
  rename(personId = id) |>
  left_join(rankssingle, by = "personId") |>
  left_join(continents, by = "countryId")|>
  select(name.x,continentRank, gender,continentId) |>
  group_by(continentId, continentRank) |>
  filter(gender == "f") |>
  filter(continentId == "_Europe") |>
  arrange(continentRank)
```

```
## # A tibble: 5,157 × 4
## # Groups:   continentId, continentRank [3,603]
##    name.x                 continentRank gender continentId
##    <chr>                          <dbl> <chr>  <chr>
##  1 Magdalena Pabisz                   4 f      _Europe
##  2 Juliette Sébastien                10 f      _Europe
##  3 Tamar Dolenjishvili               90 f      _Europe
##  4 Celine Tran                      192 f      _Europe
##  5 Nino Zguladze                    261 f      _Europe
##  6 Sofia Saletnich                  393 f      _Europe
##  7 Katie Moughan                    403 f      _Europe
##  8 Irina Drobitjko                  438 f      _Europe
##  9 Ilona Ansel                      487 f      _Europe
## 10 Marisa Revaliente Ruiz           489 f      _Europe
## # ℹ 5,147 more rows
```

The top ranked female speedcuber (for single best time) in Europe is Magdalena Pabisz.

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?

```
## create an object 'new_comp' to have matching column name as 'results' dataset.
## join 'results', 'rankssingle', and 'new_comp' datasets to create a new object 'top
_10' to get only the top 10 speedcuber in the world. Make a new column for dates and
all the values, and find the number of dates that are less than the first date for ea
ch speedcuber and find the average

new_comp <-
  competitions |>
  rename(competitionId = id)

top_10 <- results |>
  left_join(rankssingle, by = "personId") |>
  left_join(new_comp, by = "competitionId") |>
  group_by(worldRank) |>
  arrange(worldRank) |>
  filter(worldRank <= 10)

top_10 |>
  mutate(date = as.Date(paste(year, month, day, sep = "-"))) |>
  pivot_longer(cols = c(value1, value2, value3, value4, value5),
               names_to = 'value',
               values_to = 'time') |>
  group_by(personName) |>
  mutate(min_date = min(date[time < 500])) |>
  summarize(n_solves = sum(date < min_date)) |>
  summarize(avg_solves = mean(n_solves))
```

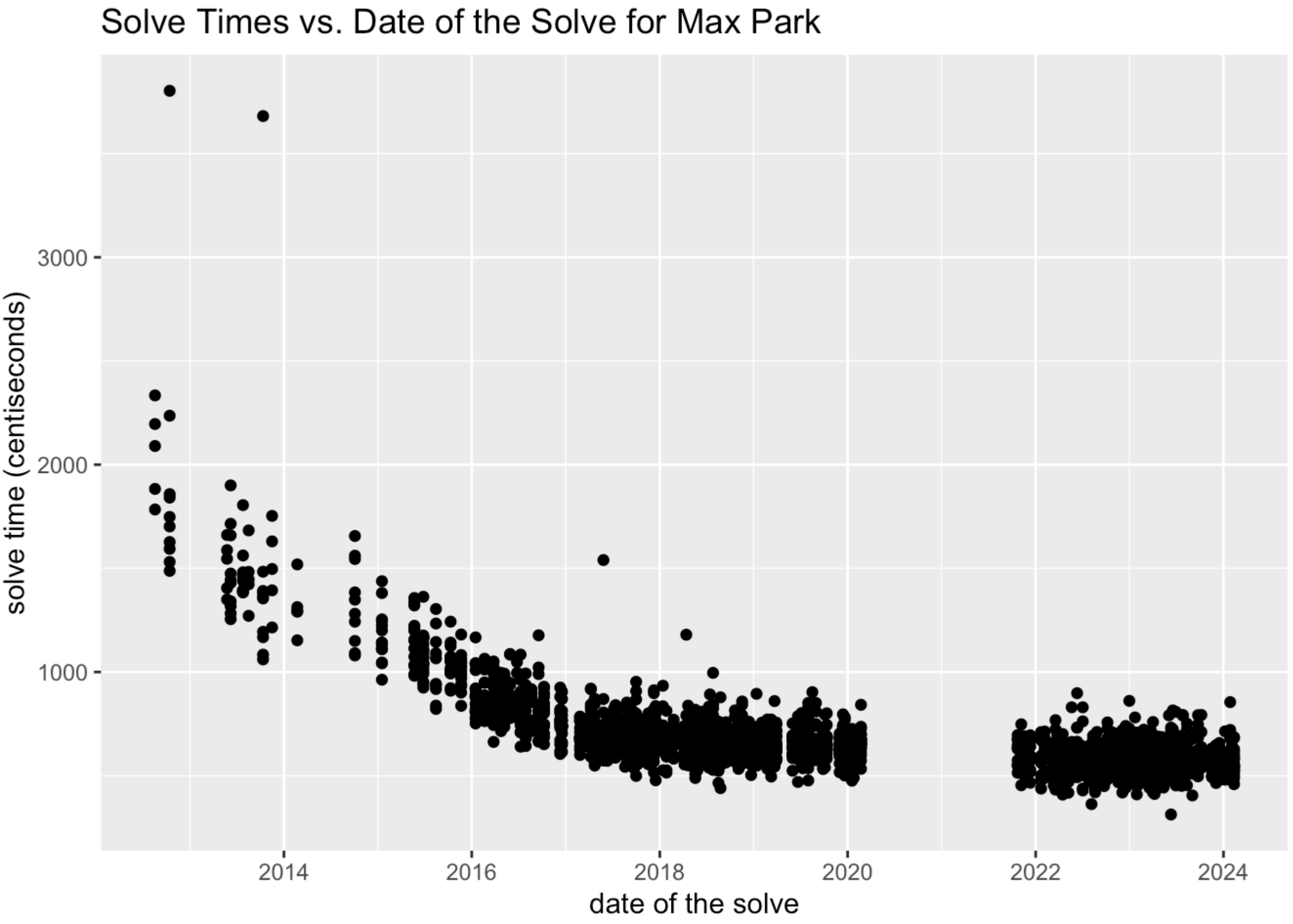
```
## # A tibble: 1 × 1
##   avg_solves
##        <dbl>
## 1       124.
```

On average, they have to do around 124 solves before achieving a sub-5 second solve.

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

```
## filter for Max Park and create a scatterplot of his solve times vs. date of solve

top_10 |>
  mutate(date = as.Date(paste(year, month, day, sep = "-"))) |>
  pivot_longer(cols = value1:value5,
               names_to = 'value',
               values_to = 'time') |>
  filter(time != -1) |>
  filter(personName == "Max Park") |>
  ggplot(aes(x = date, y = time)) +
  geom_point() +
  labs(x = "date of the solve",
       y = "solve time (centiseconds)",
       title = "Solve Times vs. Date of the Solve for Max Park")
```



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.

```
## create a new object 'below_top10' that joins 'results', 'rankssingle', and 'new_co
mp' (created from previous questions) datasets to find worldRank below 10,000.

below_top10 <- results |>
  inner_join(rankssingle, by = "personId") |>
  left_join(new_comp, by = "competitionId") |>
  group_by(worldRank) |>
  arrange(worldRank) |>
  filter(worldRank > 10000)

## group_by person name and filter out NA and invalid values for best time, use lag()
to find best time of previous year and create a new variable 'improvement' to find th
e difference between current best time and best time of previous year. Find the avera
ge and total improvement over the years for each person and arrange them in descendin
g order to find the top 5 speedcubers who improved the most over time

below_top10 |>
  group_by(personName) |>
  filter(!is.na(best.x) & !(best.x == -1)) |>
  group_by(personName, personId) |>
  arrange(year) |>
  mutate(improvement = best.x - lag(best.x)) |>
  filter(!is.na(improvement)) |>
  summarize(avg_improvement = format(mean(improvement, na.rm = TRUE), scientific = FA
LSE),
            total_improvement = sum(improvement, na.rm = TRUE)) |>
  filter(avg_improvement < 0, total_improvement < 0) |>
  arrange(desc(avg_improvement), desc(total_improvement)) |>
  head(5)
```

```
## # A tibble: 5 × 4
## # Groups:   personName [5]
##   personName             personId   avg_improvement total_improvement
##   <chr>                  <chr>      <chr>                       <dbl>
## 1 James Marshall         2022MARS07 -999.5                      -1999
## 2 Lukas Otto             2018OTTO02 -999.25                     -3997
## 3 Biel Ubia van Spaandonk 2018SPAA01 -999                        -999
## 4 Ethan Basco            2023BASC01 -999                         -999
## 5 Hugo Karlsson          2019KARL03 -999                         -999
```

Five “up-and-coming” speedcubers and their WCA ID

1. James Marshall (2022MARS07)
2. Lukas Otto (2018OTTO02)

3. Biel Ubia van Spaandonk (2018SPAA01)
4. Ethan Basco (2023BASC01)
5. Hugo Karlsson (2019KARL03)

Justification:

I calculated the speedcubers' improvements in their best time over the years. Since the calculation was the current best single time minus the best single time of the previous year, having a negative average improvement indicates that the speedcubers have gotten faster (shorter time) over the years. Therefore, the five people with the largest negative times, on average 9.99 seconds faster over the years, are up-and-coming speedcubers we should keep an eye on in the future.

Additional Questions

Question 1

State the question here: Which country has the most top 10,000 speedcubers?

State your expectation here: I expect Korea or Japan to have the most speedcubers who are in the top 10,000 speedcubers worldwide for best single time.

```
## join 'persons' and 'rankssingle' datasets, filter the top 10,000 speedcuber, group by country, and arrange and count the number of speedcubers in each country
persons |>
  rename(personId = id) |>
  left_join(rankssingle, by = "personId") |>
  filter(worldRank <= 10000) |>
  group_by(countryId) |>
  summarize(n = n()) |>
  arrange(desc(n))
```

```
## # A tibble: 111 × 2
##    countryId          n
##    <chr>          <int>
##  1 USA             2076
##  2 China           1027
##  3 India            394
##  4 Canada           376
##  5 Poland           358
##  6 Australia        348
##  7 Philippines      334
##  8 United Kingdom   254
##  9 Vietnam          226
## 10 France           207
## # ℹ 101 more rows
```

The USA has the most top 10,000 speedcubers with 2,076 speedcubers from the USA.

Question 2

State the question here: What are the top 5 competitions with the highest number of participants?

State your expectation here: I expect the top 5 competitions with the highest number of participants to be from the UK Opens, Stevenage competitions that occur during summer (June, July, August), and the Shri Ram Cubing Challenges.

```
## join 'results' and 'competitions' datatsets, group by competition name, and arrang
e and find the number of unique person IDs in each competition

results |>
  left_join(competitions, by = c(competitionId = "id")) |>
  group_by(name) |>
  summarize(n_participants = n_distinct(personId)) |>
  arrange(desc(n_participants)) |>
  head(5)
```

```
## # A tibble: 5 × 2
##   name                                   n_participants
##   <chr>                                           <int>
## 1 Rubik's WCA World Championship 2023             1142
## 2 CubingUSA Nationals 2023                        1014
## 3 World Rubik's Cube Championship 2017             913
## 4 Asian Championship 2016                          895
## 5 China Championship 2019                          843
```

Rubik's WCA World Championship 2023, CubingUSA Nationals 2023, World Rubik's Cube Championship 2017, Asian Championship 2016, and China Championship 2019 were the top 5 competitions that had the highest number of participants with 1,142 people, 1,014 people, 913 people, 895 people, and 843 people, respectively.

Discussion

From the data, it is clear that a lot of effort and time was needed to stand out among the world's best speedcubers. This is a very competitive competition, and slight differences in centiseconds can make a great impact on records and rankings.

Reflection on additional questions

For my first additional question, I expected Korea or Japan to have the most speedcubers in the top 10,000 speedcubers worldwide for the best single time; however, the result showed that the USA has the most speedcubers in the top 10,000. This may be because there are millions more people in the US than Korea or Japan, resulting in higher proportions of the top 10,000 speedcubers. It also may be that people in the US are more interested in speedcubing. I would want to find the number of participants in WCA 3x3x3 competitions from each country who are not in the top 10,000 to see if the result supports the idea that generally, more people in the US participate in speedcubing compared to Korea, Japan, or other countries.

For my second additional question, my expectations of the top 5 competitions with the most participants were inconsistent with the results. The discrepancy between my expectations and the results may be due to the fact that the competitions that I expected were more centered around Europe and South Asia, which was less spread out worldwide and diverse than the actual results (Korea, USA, France, China). In relation to my first additional question, Korea, the USA, France, and China cover the majority of the top 10,000 speedcubers, which makes me want to explore if there is a correlation between the number of top 10,000 speedcubers from a particular country and the attendance of competitions hosted by that country.

Reflection on challenges

Navigating the datasets was challenging as it required a deep understanding of the information in each dataset and its relationships. Determining the crucial steps and columns necessary for answering the questions was also challenging. From this process, I learned the importance of a clear step-by-step approach, and it enhanced my ability to organize my thoughts and problem-solving skills. Throughout the process, I discovered multiple ways of approaching the same problem and could find shorter methods to reach the same result.

Acknowledgements

Bose guided me during the project day lab, and my lab group members assisted me with clarification questions. I also had assistance from many of my classmates via GroupMe.