The data below shows the first six entries of the professional Japanese sumo wrestling data set that is utilized for this activity. Containing all records from 1957 to mid 2023, the results of individual wrestlers from the elite Makuuchi Banzuke division are displayed for each Basho (tournament). Several other important variables, explained in the data description, also accompany each wrestlers’ appearances in the data set. The full data (sumo\_since\_1957.csv) is available on the GitHub repository associated with this module.

**A picture containing text, font, screenshot, white

Description automatically generated**

1. a. Subset the data to contain wrestlers from Mongolia only.

sumo%>%

filter(Shusshin == "Mongolia")

b. Using one line of code, subset the full data to contain wrestlers from Mongolia or Tokyo only.

sumo%>%

filter(Shusshin == "Mongolia" | Shusshin == "Tokyo")

1. c. Using one line of code, subset the full data to contain wrestlers from Mongolia who train at Kataonami.

sumo%>%

filter(Shusshin == "Mongolia" & Heya == "Kataonami")

1. a. Extract rows 25 through 85 from the full data set.

sumo%>%

slice(25:85)

b. Identify the rows with the single highest and lowest recorded weights of wrestlers. Show only the top result for both lines.

sumo%>%

slice\_max(weight\_kg, with\_ties = FALSE)

sumo%>%

slice\_min(weight\_kg, with\_ties = FALSE)

1. Create a new variable that displays the wrestlers’ win percentage for each appearance in a Basho.

sumo%>%

mutate(win\_percentage = (wins/(wins+losses+ties))\*100)

1. Calculate how many times each wrestler appears in the data set.

sumo%>%

group\_by(Rikishi)%>%

summarise(total\_appearances = n())

1. Calculate the mean and standard deviation of weight for each wrestler. What information do these calculations provide and why is it important?

sumo%>%

group\_by(Rikishi)%>%

summarise(mean weight = mean(weight\_kg)

sd\_weight = sd(weight\_kg)

The calculation for mean shows the average weight for each wrestler over their carreer

which could be related to their career performance. The standard deviation shows how much their weight varied throughout their career. In tandem, we could see trends of target weights and weight variation.

1. Calculate the total losses for each wrestler and order the resulting data from least to most.

sumo%>%

group\_by(Rikishi)%>%

summarise(total\_losses = sum(losses))%>%

arrange(total\_losses)

1. Calculate the total wins for each wrestler and order the resulting data from most to least.

sumo%>%

group\_by(Rikishi)%>%

summarise(total\_wins = sum(wins))%>%

arrange(desc(total\_wins))

1. Calculate the overall proportion of wins for each wrestler over their career and arrange from highest to lowest. Why might this be a particularly important way to view the data?

sumo%>%

group\_by(Rikishi)%>%

summarise(total\_win\_percentage = sum(wins)/(sum(wins+losses+ties)))%>%

arrange(desc(total\_win\_percentage))

or

sumo%>%

group\_by(Rikishi)%>%

summarise(total\_win\_percentage = mean(wins)/(mean(wins+losses+ties)))%>%

arrange(desc(total\_win\_percentage))

It is important because we can see the wrestlers who have the best performance throughout history, not just per each event.