MATHS 7107
Data Taming
Week 8 - Revision

Shenal Dedduwakumara

School of Mathematical Sciences, University of Adelaide

Lad the data

1. Load the libraries.
2. Read the data.

Taming data

3. Extract the country code from the player variable and create a new variable called Country. *hint*: You can use "$\backslash\backslash((.+?)\backslash\backslash)$" regular expression.
4. Remove the country code from the players name.
5. Extract only the number from the HS (Highest score) variable?
6. Change variable names as follow

```
%>% head()
```

► 100 as Centuries
► 50 as Fifties
► 0 as Zeros

7. Recode the relevant variables as factors and integers
8. Recode the factors using fct_recode() such that there are no typographical errors in the Country variable.
9. Create a new variable for the number of years the player has played cricket.
10. Save the dataset

Descriptive Statistics

11. Who scored highest test score? and from which country?
12. Compare the batting averages from each country
13. Create a bar graph for the proportion of centuries from each country?

14. Build up a model to predict the average score by using the following predicters.

- Mat
- NO
- HS
- Centuries
- Fifties
- Zeros
- Country
- Years

15. Check the assumptions
16. Predict the batting average of a player with the following statistics?

- ▶ 85 matches (Mat)
- ▶ 17 Notouts (NO)
- ▶ 27 Centuries
- ▶ 36 Fifties
- ▶ 12 Years

17. Can we use this model to predict the batting average of any player?