

Exercise 2 - Data management

1. Input data

- Download and install the packages "ellipse", "reshape2", "ggplot2", "dygraphs", "dplyr"
- Load the "Teams_overall2020.csv" dataset

2. Data preview

- Provide a description of the data frame including a preview of data, the number of rows and columns and the classes of data.

3. Data subsetting

- Create a new dataframe containing only the following fields:
"Teams", "Seasons", "W", "L"
- Create a new dataframe containing only the "W" and "L" made by a chosen Team.
- Create a new dataframe containing only the "PTS", "PTS.ALL" and "W." for the Season 2004-5

4. Data grouping

- Rank teams by points scored in the Season 2004/2005
- Provide a TIME-SERIES plot of the number of Teams per Season over the sample 1980-2018
- Provide a TIME-SERIES plot of the average number of points scored per season over the sample 1980-2018

5. Graph and data Analysis

- With the aid of a QQ-plot compare the distribution of the average number of points per game scored over the sample 1980-2018 with the normal distribution
- Compare the correlation among three points percentages, two point percentages and total wins over the sample 1980-1999 and 2000-2018
- Provide a TIME-SERIES plot of the average number of points scored per season over the sample 1980-2018

6. Model Simulation

From 10,000 simulations each derive a probability chart of teams of winning %s of (0.2, 0.5 and 0.8) of achieving win streaks of lengths (3w, 5w, and 7w) in a regular NBA Season.