# CASE STUDY 009

# [R]

# Wikipedia Page Views

## Difficulty Level: 3 of 3

*Disclaimer: SuperDataScience has no affiliation with Wikipedia. The scenario is made up for educational purposes.*

You are a Data Analyst working for Wikipedia. You have been given a large, dirty data set detailing the page views for a single hour. The head of the Analytics Team would like you to write a script to clean the data and create some visualisations.

1. Optional very difficult question: Load the following files and append them to create one data file. See the hints page for some help to do this in a very powerful way
    a. wikipedia_data.csv, aa.csv, abw.csv, ab.csv, aaw.csv
2. If you completed question 1 go to question 3, otherwise load the file wikipedia_data_full.csv
3. Complete the following data cleaning steps
    a. Name the columns as follows "source", "page_name", "views", "size"
    b. Format the variables
    c. Remove duplicate rows
    d. Remove rows with NA in column "views"
    e. Add a column called "language" (see hints page for details)
4. Which language has the largest number of view? (please create a histogram of the top 10 languages)
5. Which page, excluding Main Pages, has the highest number of views?
6. Which is the largest page? (please create a histogram of the top 10 pages)

Good luck! The organisation depends on you.

*Difficulty note: this is a difficult assignment. Do not be surprised that there will be lots of nuances we have not covered off in the courses. But just like in the Real Life – there will be things training has not prepared you for and you will need to do research to find how to solve the problems at hand. If you get stuck, check the clues file.*