# MSDS 6306: Doing Data Science – Preparing Data

## Live session Unit 09 assignment

## Due: 1 hour before your 10th live session

## Submission

**ALL MATERIAL MUST BE KNITTED INTO A SINGLE, LEGIBLE, AND DOCUMENTED HTML DOCUMENT.** Use RMarkdown to create this file. Formatting can be basic, but it should be easily human-readable. Unless otherwise stated, please enable {r, echo=TRUE} so your code is visible.

**TIPS:** If you are having problems with scraping, go to the website to check if it is online. If it is, then take a look at the actual website to verify how it is structured. Feel free to use View Source to narrow down good nodes to use. You are welcome to use any R libraries for this assignment. Off the top of my head, good ones to use might be rvest, dplyr, tidyr, ggplot2, reshape2, or stringr. You don't need to include install.packages for your final code, but you *need* library() for each.

## Questions

1. **Harry Potter Cast (50%)**

    **a.** In the IMDB, there are listings of full cast members for movies. Navigate to http://www.imdb.com/title/tt1201607/fullcredits?ref_=tt_ql_1. Feel free to View Source to get a good idea of what the page looks like in code.

    **b.** Scrape the page with any R package that makes things easy for you. Of particular interest is the table of the Cast in order of crediting. Please scrape this table (you might have to fish it out of several of the tables from the page) and make it a data.frame() of the Cast in your R environment

    **c.** Clean up the table

    - It should not have blank observations or rows, a row that should be column names, or just '…'
    - It should have intuitive column names (ideally 2 to start – Actor and Character)
    - In the film, Mr. Warwick plays two characters, which makes his row look a little weird. Please replace his character column with just "Griphook / Professor Filius Flitwick" to make it look better.
    - One row might result in "Rest of cast listed alphabetically" – remove this observation.

    **d.** Split the Actor's name into two columns: FirstName and Surname. Keep in mind that some actors/actresses have middle names as well. Please make sure that the middle names are in the FirstName column, in addition to the first name (example: given the

Actor Frank Jeffrey Stevenson, the FirstName column would say "Frank Jeffrey.")

**e.** Present the first 10 rows of the data.frame() – It should have only FirstName, Surname, and Character columns.

**2. SportsBall (50%)**

**a.** On the ESPN website, there are statistics of each NBA player. Navigate to the San Antonio Spurs current statistics (likely http://www.espn.com/nba/team/stats/_/name/sa/san-antonio-spurs). You are interested in the **Shooting Statistics** table.

**b.** Scrape the page with any R package that makes things easy for you. There are a few tables on the page, so make sure you are targeting specifically the Shooting Statistics table.

**c.** Clean up the table (You might get some warnings if you're working with tibbles)

- You'll want to create an R data.frame() with one observation for each player. Make sure that you do not accidentally include blank rows, a row of column names, or the Totals row in the table as observations.
- The column PLAYER has two variables of interest in it: the player's name and their position, denoted by 1-2 letters after their name. Split the cells into two columns, one with Name and the other Position.
- Check the data type of all columns. Convert relevant columns to numeric. Check the data type of all columns again to confirm that they have changed!

**d.** Create a colorful bar chart that shows the *Field Goals Percentage Per Game* for each person. It will be graded on the following criteria.

- Informative Title, centered
- Relevant x and y axis labels (not simply variables names!)
- Human-readable axes with no overlap (you might have to flip x and y to fix that). **Note:** You do not have to convert the decimal to a percentage.
- Color the columns by the team member's position (so, all PF's should have the same color, etc.)

**Reminder**

To complete this assignment, please submit **one** RMarkdown and matching HTML file at least one hour before your live session. Please submit all files at the same time; only one submission is granted.

Good luck!