# ftblData Read Me

Jonas Stillhard, November 2017
jonas.stillhard@wsl.ch

## Available Datasets

### Football Data

The available datasets are partially described in pres1.pdf. It might be worth to check out **footballsquads.co.uk**. They provide squad data for all world cups back to 1950, i did not find an API so far to get the datasets.

### World Bank data

World bank data can be downloaded via the **world banks online interface**. An API is available but it proved to be easier to just download the data as .csv from the website rather than getting it via the API.

### Other datasets

So far, i focused on World Bank datasets. Datasets such as the **World Penis dataset** need a registration, but might be worth having a look at.

### Datasets within repository

The following datasets are available within the repository in `'./1_data'`

- worldcup.db: Downloaded from **openfootball.db**.
- wc 2014 data: json from **Joe Kampschmidts** github repository. Not physically available in the repo but as part of the script `1_Get_Data.R`.
- wdi.RDS: Downloaded as csv from the **world banks data site**. Contains 500+ Indicators, of which only few are available for all countries. Dataset was subset to the countries that ever participated in a World Cup.
- Gender Stats: Contains 630 gender-related indicators such as *Women who believe a husband is justified in beating his wife when she refuses sex with him (%)*. Dataset was subset to the countries that ever participated in a World Cup.
- World Government Indicators (WGI): Contains government 36 related indicators. Dataset was subset to the countries that ever participated in a World Cup.
- The files events_teams.RDS, World_Cup_Doubled_up.RDS, worldcups.RDS, worldCupSummary.RDS and WorldCupSummaryAll.RDs are generated within one of the above mentioned scripts.

## Scripts

The Repository contains the following scripts and some other, deprecated scripts within the folder 4_Scripts/deprecated

### 1_Get_Data.R

The Script loads data from the sqLite database worldcups.db via package `sqLite` and does some data manipulation to the data loaded. In a second step, world cup 2014 data is loaded from **Joe Kampschmidts** git repo. The data is exported as .RDS.

**2_Create_Clean_Country_Codes.R**

Creates codes for matching with world bank datasets for (most) of the teams/countries. As for some countries (e.g., Soviet Union), no world bank datasets are available, they are removed in a latter script.

**3_PrepareFtblDataForCorrelation.R**

Creates three datasets: A dataset with summaries for every country for every world cup, a dataset with the summaries but also containing lines without any values for world cups that a team missed and a 'doubled up' dataset. That dataset might not be needed, it contains every game played on two lines, so every team appears in both columns (teamA and teamH etc.). Could be useful for analyses on game level.

**4_Add_World_Bank_Data.R**

Adds world bank data to the extended dataset (Teams appear even if they missed a world cup). Should be improved.