Project 2: Data Querying & Cleaning (SQL/Power BI)

Overview

* Making queries with SQL
* Basics, control flow and data structures in Python
* Useful Python libraries: numpy, matplotlib, and pandas
* Basic R
* R data structures: vectors, matrices, lists, and data frames
* Graphics using R

In this project I’m going to be applying Data Querying and Cleaning skills in to explore baseball statistics. In the process I’ll practice a **vital**data science skill – the ability to research new tools and read documentation.

Concepts covered:

* Reading and applying technical documentation
* Querying SQL data in R/Python
* Cleaning and preparing data
* Visualizing data in R/Python

The Dataset

The dataset I’ll be using is [Lahman’s Baseball Database](http://www.seanlahman.com/baseball-archive/statistics/" \t "_blank). This dataset contains a variety of baseball statistics gathered from 1871-2018, including both individual and team statistics.

More information can be found at the supplied link. We’ll be downloading and using the “2018 – SQL version” file. You will need to read the [documentation](http://www.seanlahman.com/files/database/readme2017.txt) to find out more about the various tables and statistics available in the database.