Overview

You’ve covered a lot of ground in the past there weeks, learning about

* 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 your second project you’re going to be applying these skills in to explore baseball statistics. In the process you’ll practice a **vital**data science skill – the ability to research new tools and read documentation. You may choose to do this project in either R or Python.

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 we’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.

Requirements

If you’re unfamiliar with baseball, you can find a quick summary of the objectives of the game [here](https://www.baseball-rules.com/basicbb.htm), though you don’t need to understand much about baseball to complete this project. You need to write a Python or R script that does the following:

* Connects to an SQL database file and queries for all players who have played at least 50 games and are still active.  Use the “finalGame” field from the “People” table to determine if a player is active. Retrieve weight, throws, bats, throws, all birth-related and all name-related columns from the “People” table and retrieve all columns from the “Batting” table.
* Converts this data into either an R data frame or a pandas data frame.
* Adds a calculated column with the player’s age and a calculated column with each player’s first and last name concatenated.
* Once the calculated columns are added, drops the other columns related to birth date and name.
* Deletes any rows with missing values
* Answers the following questions:  
  + Which active player had the most runs batted in (“RBI” from the Batting table) from 2015-2018?
  + How many double plays did Albert Pujols ground into (“GIDP” from Batting table) in 2016?
* Creates the following plots:  
  + A histogram of triples (3B) per year.
  + Create a scatter plot relating triples (3B) and steals (SB).
* Comes up at least three additional questions about the data and answers them. At least one should be a question about the relationship between two variables, e.g., triples and steals, as above.

To meet these requirements you will need to research additional packages and/or functionality from existing packages:

* Connecting to an SQL database file and performing SQL queries. For Python, use the popular package *sqlite3* and for R use *RSQLite*.
* Manipulating data, in this case for adding and removing columns, rearranging and grouping, etc. In Python, the *pandas* package should provide the necessary functionality; read the documentation to find out more. For R, the *dplyr* package is the most popular choice.

For ease of grading, please use only the above packages (along with any packages used in the preceding courses).

Note: while manipulating data, be careful about player “stints”. Baseball players sometimes play for multiple teams in the same year (due to trades); such players will have multiple rows for such a year, one for each “stint” with a team.

Submission

When you have completed the project, email your instructor following deliverables:

1. A single Python or R script. Excel workbook and/or Power BI report (attached to the email).
2. A document (Word file, PDF, or plain text) containing:

* + Details on how your project met the above
  + A description of the additional questions your project answered
  + Ideas for future improvement of your project (minimum of 3)