**PHX #2 - SABR Qualifying Scenario #2 (Data)**

*The intent of this qualifying scenario is to help determine which individuals will make up the teams for the* [*Diamond Dollars Case Competition*](https://sabr.org/analytics/case) *at the* [*SABR Analytics Conference*](https://sabr.org/analytics) *in Phoenix in March. This case study must be done individually.*

*You may use any* ***inanimate*** *source of data, materials, computers, software, references, websites, books, etc. You may not discuss the case study with anyone other than Mr. Benesh. Be sure to credit all sources used.*

Upload your final analysis as a PDF and all supporting documentation (including any files – Excel, code in Notepad, etc.) in Schoology by Tues 17 Dec @ 4 PM.

My assigned team is \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

For this qualifying scenario, you will be judged on:

* The reasonableness and defensibility of your models as well as their validity.
* The logic included in your code (pseudo-code) or Excel calculations.
* The ability to properly create the files requested then pull the requested data from the files.

**Part I – Sabermetric Model**

1. For last season (2024) for the entire league (all 30 teams – ie, 30 rows of data), using team-level data from [StatCast from Baseball Savant](https://baseballsavant.mlb.com/league) create a model estimating the variable wOBA (in the ‘Statcast Hitting’ table) using the variables in the ‘Plate Discipline’ and ‘Batted Ball Profile’ tables. You CANNOT use variables from the ‘Statcast Hitting’ table.
2. Next, apply your model to the **2023** season (change year in upper right dropdown) to estimate wOBA. Calculate the absolute value of the difference between the actual values in 2021 and your estimated values for each team. Take the average of those absolute value differences to determine the [Mean Absolute Error](https://en.wikipedia.org/wiki/Mean_absolute_error).

**Part II – Simulation**

In baseball, a pitcher must record 27 outs in a row to throw a perfect game. In this Monte Carlo Simulation, we will be simulating the first 27 batters who come to the plate. If each and every batter makes an out (doesn’t get on base), our simulated pitcher will have thrown a perfect game. Otherwise, at least one batter will have reached base thus no perfect game, and we will ignore the rest of the game. Using On-Base Percentage (OBP – i.e., not an out!) of 0.317, simulate the outcome of 27 batters. Total up the number of batters who get on base. A perfect game would have a total of zero batters reach base. This represents one simulated game.

Run the simulation the 2,430 times (this is the total number of games in an MLB regular season).

Run the simulation several times. How frequently did you get a perfect game?

**Part III – Data Management A**

1. Create a master file for the AL (Lg=’AL’) and a master file for the NL (Lg=’NL’) by joining the files for the provided 10 years in the ‘Data Management’ folder.
2. Using the master files separately, create the following for **both** leagues (AL and NL) separately:
   1. Top 10 player/year combinations ranked by WAR (higher is better) for the decade (2007-2016) for each league
   2. Top 10 player/year combinations (Age > 30) by WAR (higher is better) for the decade (2007-2016) for each league
   3. Sum of WAR by the entire year (ie, total WAR for 2007), then rank the 10 years by WAR (higher is better)

**Part IV – Data Management B**

1. For your assigned team, create a master file by joining the files for the provided 10 years in the ‘Data Management’ folder for your team only.
2. Next, create:
   1. Top player WAR (higher is better) for each year of the decade (2007-2016) for your team
   2. Top 10 player/year combinations (AB > 400) by OPS (higher is better) for the decade