2ND DATA SCIENCE PROJECT – DATA QUERYING AND CLEANING

INTRODUCTION TO DATA

The dataset we were using was Lahman's Baseball Database. This dataset contains a variety of baseball statistics gathered from 1871-2018, including both individual and team statistics.

DETAILS ON HOW MY PROJECT MET THE REQUIREMENTS

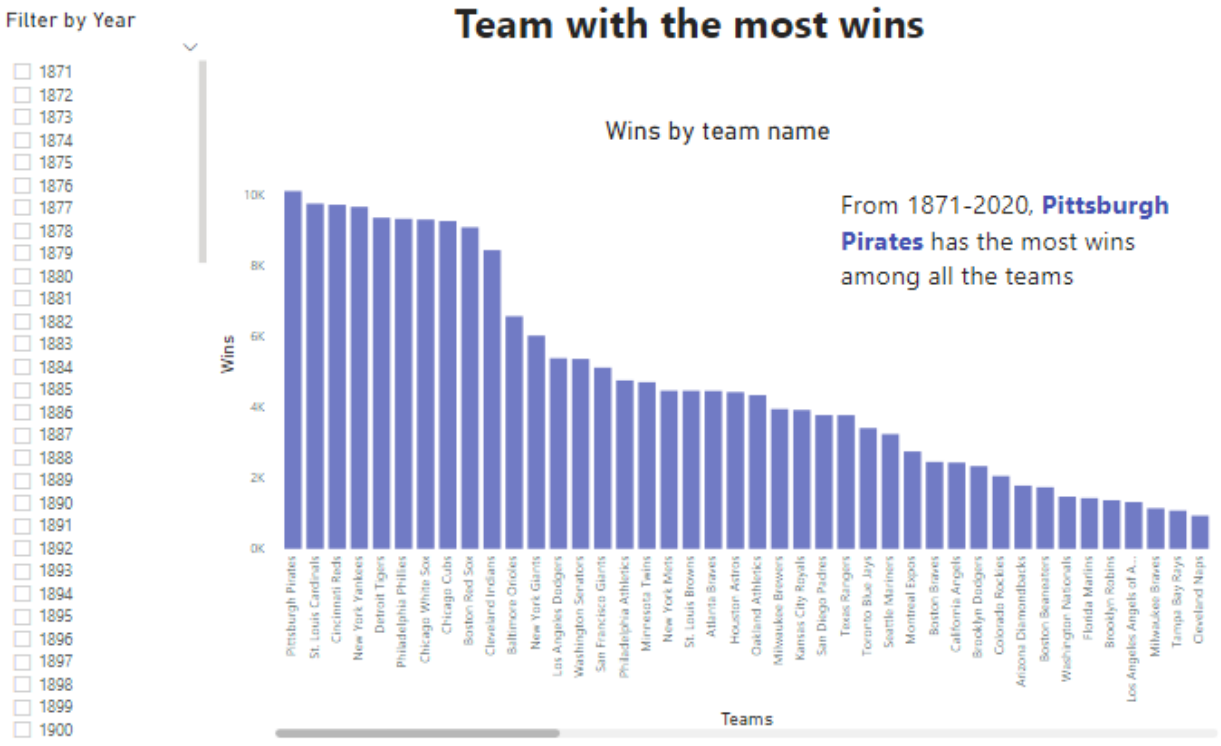
- Used **PyCharm** IDE for python
- Separate CSV files were given for the separate tables of database. Downloaded **DB browser for SQLite** and created a database containing all the tables in it
- Installed **SQLite** on PyCharm, imported it and connected the database
- Wrote a query and joined 3 tables to fetch required columns from the database. Used WHERE keyword in that query to retrieve specific rows from the tables. The three conditions were:
    - Age of the player should be less than or equals to 45 (We were told to put 40 here but needed Albert Pujols data and he is 41 years old)
    - Death year should be null (To determine active players)
    - Who has played at least 50 games (It was required)
- Imported **pandas** and used **read_sql_query** function to convert the data into a data frame
- Created calculated column for the player's "**Fullname**" (First name + Last name)
- Created calculated column "**Age**" (2021 – Birth year)
- Used pandas **drop** function to drop name related and birth related columns which were not of any use anymore
- Used pandas **dropna** function to delete any rows with missing values
- Dropped duplicated with pandas **drop_duplicates** function
- For the **player with most batted runs** from 2015 to 2018:
    - Retrieved rows from 2015 to 2018
    - Sorted values for better understanding
    - Used RBI.idxmax() to get the maximum value from RBI column
    - Printed the player's name with his most batted runs
- For **double plays of Albert Pujols** in 2016
    - Retrieved rows to get Albert Pujols records from the year 2016
    - Printed his name with his GIDP (Double plays)
- For **histogram of triples per year**
    - Sorted the yearID column with unique values
    - Used for loop to get 3B (triples) for each year

- Used plot function with kind=hist for plotting histogram
- For a **scatter plot relating triples (3B) and steals (SB)**
  - Used sum function for all of the playerIDs to get sum of the 3Bs and SBs of each player
  - Used scatter function to plot scatter chart

THREE ADDITIONAL QUESTIONS ABOUT THE DATA

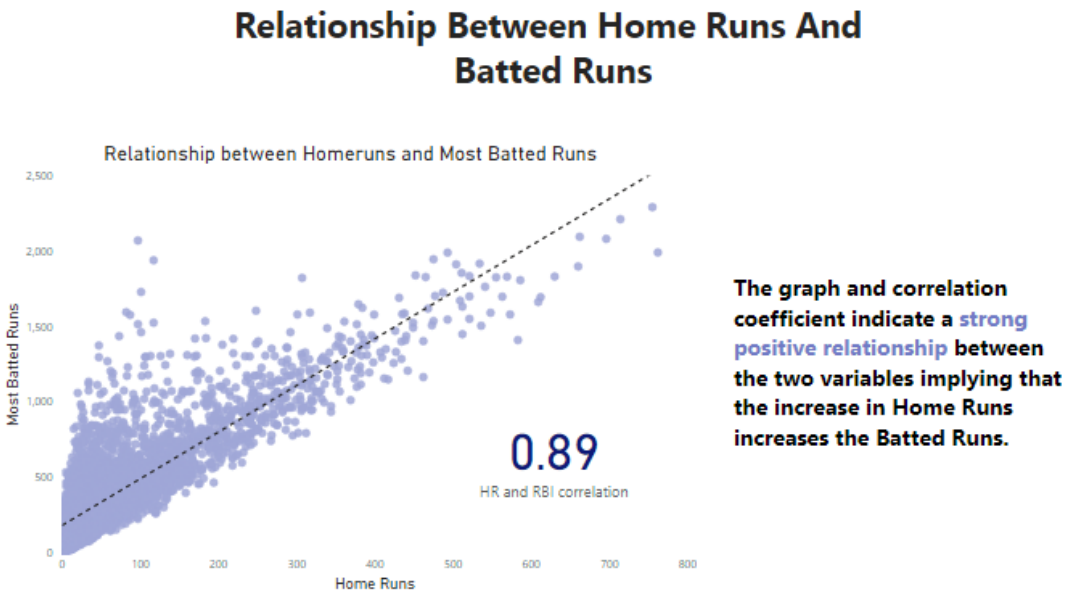
Used power BI to answer three additional questions with visualizations

1. Team with the most wins



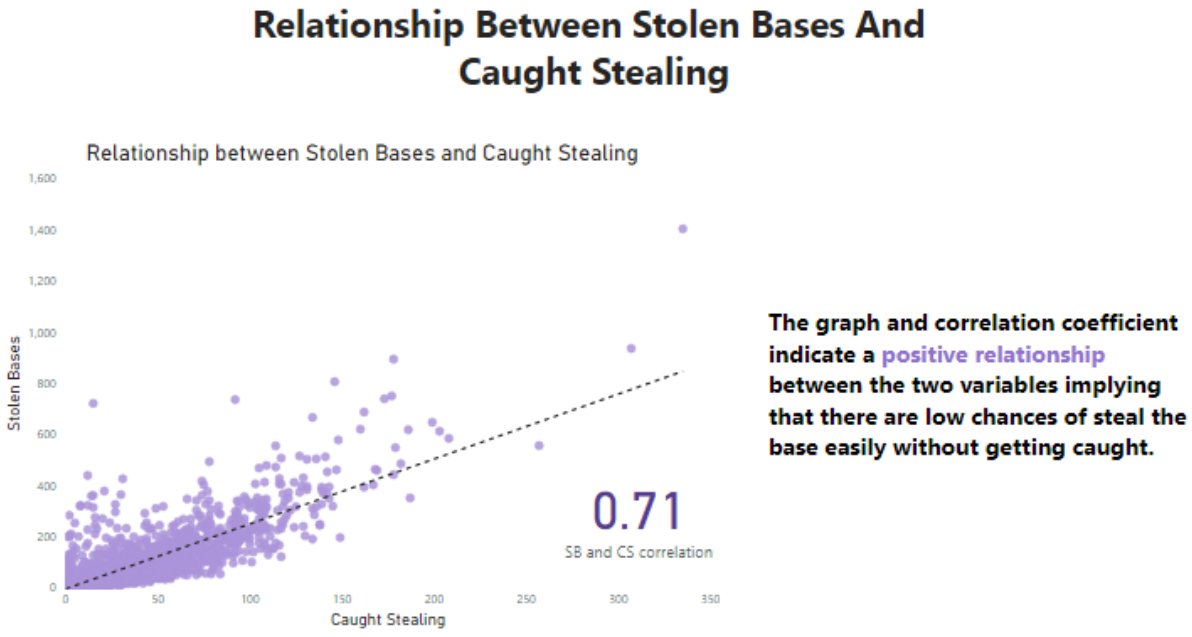
Above diagram shows that the team Pittsburgh Pirates has the most wins from 1871-2020. Added a slicer too to filter by year

2. Relationship between Homeruns and batted runs



As the diagram shows we have a strong positive relationship between player's homeruns and batted runs. The more home runs, the more batted runs.

3. Relationship between Stoles bases and Caught Stealing



We have a positive relationship between the player's stolen bases and caught stealing.