**The facts which are related to baseball and stored in Lahman’s Baseball Database would be analyzed in the given report**

**Mahamadou Mahi Diallo**

**July 21, 2024**

**Introduction**

Baseball has many statistical figures that help to assess the players’ efficiency and the teams’ prospects. To underline these observations, we employ Lahman’s Baseball Database that incorporates the statistics from 1871 to 2018. The tasks of this undertaking include writing queries and mining baseball databases to determine trends and responses to certain sets of questions about player and team performances. With the help of Python, and other useful libraries like pandas, matplotlib, and sqlite3, it would be possible to explore data to solve numerous research questions and visualize outcomes properly. In this report, the author outlines actions that were deemed necessary to translate into connecting with an SQL database, manipulating data and creating a visualization in order to answer questions regarding the performance of the different players.

**Methodology**

**1. Connecting to SQL Database**

First, we established our linkage to the SQLite database that Engages Lahman’s Baseball Database using the package of Python’s sqlite3. This dependency was made with the help of the database file Lahman2018. sqlite

**2. Querying Data**

We wrote a SQL code that would help extract data from two tables namely the People and Batting tables. The query was carried out based on the criteria of players who have played at least 50 games and are active (as deduced by the NULL value in the final game field). Thus, all data from the People table containing weight, throws, total bats, birth-related records, and name details, as well as all entries in the Batting table, have been selected.

**3. Data Conversion and Manipulation**

Most of the data was cleaned before the retrieved data was loaded into a panda DataFrame for further data manipulation. The following transformations were performed: The following transformations were performed:

* **Age Calculation**: Aging of the player was calculated as the current year subtracting the birth year of the player and using the formula 1 if the birthday has occurred in the current year otherwise 0.
* **Name Concatenation**: Furthermore, a new column was generated through assembling the player’s first and last names.
* **Column Dropping**: The operations that have been performed include deleting the columns relative to the birth date and name.
* **Handling Missing Values**: Such observation led to the deletion of any row that contained a missing value so as to guarantee the correct analysis was carried out.

**4. Answering Research Questions**

With the cleaned data, we addressed specific research questions: With the cleaned data, we addressed specific research questions:

* **Player with Most RBI from 2015-2018**: Established which player from the active players was most instrumental in scoring the total number of Run Batted In (RBI) within the stipulated timeframe.
* **Albert Pujols’ Double Plays in 2016**: Determined the number of double plays Albert Pujols grounded into in the year 2016.

**5. Data Visualization**

Using matplotlib, we created the following plots: Using matplotlib, we created the following plots:

* **Histogram of Triples (3B) per Year**: Also displayed the first three columns of the spreadsheet where it indicated triples by year.
* **Scatter Plot of Triples vs. Steals (SB)**: Analyzed the correlation between triples and stolen bases among the players.

**6. Additional Analysis**

We formulated and answered additional questions to gain further insights: We formulated and answered additional questions to gain further insights:

* **Average Weight of Players by Year**: Sought a way of analyzing how the weight of the players has changed over the years and drew a conclusion as to the average weight per year.
* **Average Age of Players by Year**: Analyzed shifts in the years of experience or players ‘mean’ age as the company unfolded.
* **Correlation Between Triples and RBI**: Assigned the relationship between triples and RBI.

**Conclusion**

The findings of this study of Lahman’s Baseball Database give an extensive outline of the players’ and team stats. By making the connection to the SQL database, selecting applicable data, and then doing further complex manipulations and visualization of manipulations, we have obtained a broad view of the player’s performance statistics trend. Using the results, it is possible to not only answer several questions of the study but also further develop research hypotheses and ideas. Future upgrades could be the inclusion of other statistical methods beyond basic regression, addition of features such as interactive graphics, and broader analysis in relation to more performance indicators and players’ characteristics. In general, this project proved the high relevance of data science competencies in data analysis and questioning.