**Understand The Data**

**By**

**Israel Nolazco**

**Section 1.**

* Which CSV file/table would you use to determine the total number of players?

In the zip file downloaded from <http://www.seanlahman.com/baseball-archive/statistics/> shows two folders. A Core folder and an Upstream folder. The best one to work with is the core folder given that this folder has multiple .csv files breaking down specific aspect of the overall data collected. Therefore, finding out the total number of players can be easily found by looking at the People.csv. This dataset shows a total of 19,879 rows/people in this file.

* Which CSV file/tablewould you use to determine a player such as Derek Jeter's  salary for the year 2010?

In order to answer this question, there are two spreadsheet needed to look at. One of them is the People.csv file and the other is the Salaries.csv. The Salaries file provides the salary information and assigns a playerID. The playerID can then be decrypted by using the People.csv.

* Which CSV file/tablewould you use to determine the player's date of birth and country of birth?

The People.csv provides the date of birth and country of birth for all the players.

* Which CSV file/tablewould you use to determine whether the player was inducted into the Hall of Fame?

A combination of the HallOfFame.csv and People.csv files. Just like the Salaries; a playerID is assigned for those players that have been inducted in the hall of fame.

* Which CSV file/table would you use to determine the name of the team a player played in, in the year 2000?

In order to get the name of a team player from a team in 2000 there three files that need to be access. The first one is obviously teams file. The file arranges the name of the team with an ID and organizes it by year. The second one is the Salaries file; with the team ID at hand we can cross it with the salary. Lastly, the playerID from the Salaries can be double cross with the People.csv file and finally get the name of the player.

* Which CSV file/table would you use to determine the number of home runs scored by a player such as Derek Jeter in 2010?

A combination of the Pitching.csv and People.csv files are only needed to find this information. The Pitching.csv file provides the number of homeruns a player (PlayerID) has made in its career. Then the People.csv would decrypt the PlayerID.

* Which CSV file/table and which column would you use to check if the player is still alive?

The People.csv and specifically the Birth Year and Death Year are the only one required. For quality purposes you could run a SQL query with returns that show if the Death Year is anything greater than 1 then show results and another query on top of that that would also look at Birth Year say if the Birth Year is two hundred less than current year and the Death Year still empty. Then the results will show fields that would probably need to be updated.

**Section 2.**

1. What is the total number of baseball players?

SELECT count(\*) FROM People

1. How many players were born in the year 1960 and earlier?

SELECT count(\*) FROM People

WHERE BirthYear < = 1960

1. How many players were born in the USA?

SELECT count(\*) FROM People

Where BithCountry = USA

1. How many players were born outside the USA?

SELECT count(\*) FROM People

Where BithCountry < > USA

1. Display the number of players born in each year starting from 1960 thru 2000. For example, the output should show: 1980 4 ( where 4 is the number of players born in 1980)

SELECT BirthYear, count (\*) from People

WHERE BirthYear >= 1960 and Birth Year < =2000

GROUP by BirthYear

Order by BirthYear

1. How many players and managers were inducted into the Hall of Fame?

SELECT inducted, category count(\*) FROM HallOfFam

Where inducted = Y

Where category = Player AND Managers

1. Provide a list of all players for any team and from any year. For example, print the list of players who played for Chicago Cubs in 2000.

SELECT Appearances.yearID, Appearances.teamID, Appearances.playerID, Teams.name, People.nameGiven

INNER JOIN Teams ON Appearances.TeamID = Teams.teamID

INNER JOIN People ON Appearances.playerID = People.playerID