## Lahman Baseball Database Exercise

- this data has been made available [online](http://www.seanlahman.com/baseball-archive/statistics/) by Sean Lahman

- A data dictionary is included with the files for this project.

\*\*Directions:\*\*

\* Within your repository, create a directory named "scripts" which will hold your scripts.

\* Create a new branch for each question. Name these branches using the question number and your initials. For example, Johnny Cash would name his branch for question 3 as question\_3\_jc. On each branch, you'll create a script to hold your answer. By the end of the project, each student should have ten scripts, one for each of the ten initial questions.

\* On Saturday morning, you should be prepared to discuss your answers to each question. Each group will review the ten initial questions and merge answers to the main branch. Merging can only take place after the group has reviewed and agreed that the script is correct. After the review session, each group should have a script for each question on the main branch. Each student needs to merge at least two of their answers to the main branch.

\* Each student should choose one of the open-ended questions or come up with their own and use the database to investigate. At the end of class on Thursday, July 6, each student will give a 5 minute presentation of their findings. For the open-ended portion, pull data from the database, import it into Excel, and use Excel to create visuals.

\*\*Initial Questions\*\*

1. What range of years for baseball games played does the provided database cover?

**SELECT MIN(yearid) AS earliest\_year, MAX(yearid) AS latest\_year**

**FROM appearances**

**--1871 to 2016**

2. Find the name and height of the shortest player in the database. How many games did he play in? What is the name of the team for which he played?

**--2. Find the name and height of the shortest player in the database. How many games did he play in? What is the name of the team for which he played?**

**--Find the name and height of the shortest player in the database**

**SELECT namefirst||' '||namelast AS player\_name, height**

**FROM people**

**ORDER BY height**

**--Eddie Gaedel at 43 units**

**--How many games did he play in?**

**SELECT p.namefirst||' '||p.namelast AS player\_name, a.G\_all AS total\_games\_played**

**FROM people AS p**

**INNER JOIN appearances AS a**

**ON p.playerid = a.playerid**

**WHERE p.namelast = 'Gaedel'**

**-- 1 game**

**--What is the name of the team for which he played?**

**SELECT DISTINCT p.namefirst||' '||p.namelast AS player\_name, t.franchID AS franchise, t.name**

**FROM people AS p**

**INNER JOIN appearances AS a**

**ON p.playerid = a.playerid**

**INNER JOIN teams AS t**

**ON a.teamid = t.teamid**

**WHERE p.namelast = 'Gaedel'**

**--SLA/BAL**

3. Find all players in the database who played at Vanderbilt University. Create a list showing each player’s first and last names as well as the total salary they earned in the major leagues. Sort this list in descending order by the total salary earned. Which Vanderbilt player earned the most money in the majors?

**--How Todd approached it along with some edits from team**

**--Run the code to see the remaining Vanderbilt players**

**SELECT p.namefirst||' '||p.namelast AS player\_name, sc.schoolname AS university\_attended, SUM(sa.salary)::numeric::money AS total\_salary\_earned**

**FROM people AS p**

**INNER JOIN collegeplaying AS cp**

**ON p.playerid = cp.playerid**

**INNER JOIN schools AS sc**

**ON cp.schoolid = sc.schoolid**

**INNER JOIN salaries AS sa**

**ON sa.playerid = p.playerid**

**WHERE sc.schoolname LIKE '%Vanderbilt%'**

**GROUP BY p.playerid, p.namefirst, p.namelast, sc.schoolname**

**ORDER BY total\_salary\_earned DESC**

**--Jeremy's approach with editing from team**

**--Run the code to see the remaining Vanderbilt players**

**WITH vandy\_players AS (SELECT playerid**

**FROM collegeplaying**

**WHERE schoolid= 'vandy' )**

**SELECT namefirst, namelast, SUM(CAST(s.salary AS numeric))::MONEY AS total\_salary**

**FROM people**

**INNER JOIN salaries AS s**

**USING(playerid)**

**INNER JOIN vandy\_players**

**USING (playerid)**

**GROUP BY namefirst, namelast**

**ORDER BY total\_salary DESC**

**--"David Price" "Vanderbilt University" "$245,553,888.00"**

4. Using the fielding table, group players into three groups based on their position: label players with position OF as "Outfield", those with position "SS", "1B", "2B", and "3B" as "Infield", and those with position "P" or "C" as "Battery". Determine the number of putouts made by each of these three groups in 2016.

5. Find the average number of strikeouts per game by decade since 1920. Round the numbers you report to 2 decimal places. Do the same for home runs per game. Do you see any trends?

**--From Abram:**

**SELECT(yearid / 10 \* 10) || 's' AS decade, ROUND(AVG(so)::numeric, 2) AS average\_strikeouts\_per\_game**

**FROM batting**

**WHERE yearid >= 1920**

**GROUP BY decade**

**ORDER BY decade**

**--Also from Abram:**

**--Find the average number of homeruns per game by decade since 1920. Round the numbers you report to 2 decimal places. Do you see any trends?**

**SELECT(yearid / 10 \* 10) || 's' AS decade, ROUND(AVG(hr)::numeric, 2) AS homeruns\_per\_game**

**FROM batting**

**WHERE yearid >= 1920**

**GROUP BY decade**

**ORDER BY decade**

**--Combined**

**SELECT(yearid / 10 \* 10) || 's' AS decade, ROUND(AVG(hr)::numeric, 2) AS homeruns\_per\_game, ROUND(AVG(so)::numeric, 2) AS average\_strikeouts\_per\_game**

**FROM batting**

**WHERE yearid >= 1920**

**GROUP BY decade**

**ORDER BY decade**

**--Approach by Jeremy after some extra discussion**

**SELECT**

**CASE WHEN yearid BETWEEN 1920 AND 1929 THEN '1920s'**

**WHEN yearid BETWEEN 1930 AND 1939 THEN '1930s'**

**WHEN yearid BETWEEN 1940 AND 1949 THEN '1940s'**

**WHEN yearid BETWEEN 1950 AND 1959 THEN '1950s'**

**WHEN yearid BETWEEN 1960 AND 1969 THEN '1960s'**

**WHEN yearid BETWEEN 1970 AND 1979 THEN '1970s'**

**WHEN yearid BETWEEN 1980 AND 1989 THEN '1980s'**

**WHEN yearid BETWEEN 1990 AND 1999 THEN '1990s'**

**WHEN yearid BETWEEN 2000 AND 2009 THEN '2000s'**

**ELSE '2010s' END AS decade,**

**ROUND(AVG(so),2) AS avg\_strikeouts\_per\_game,**

**ROUND(AVG(hr),2) AS avg\_homeruns\_per\_game**

**FROM batting**

**GROUP BY decade**

6. Find the player who had the most success stealing bases in 2016, where \_\_success\_\_ is measured as the percentage of stolen base attempts which are successful. (A stolen base attempt results either in a stolen base or being caught stealing.) Consider only players who attempted \_at least\_ 20 stolen bases.

**--From Mel after a few team edits**

**SELECT CONCAT((namefirst),' ',(namelast)) AS player\_name, ROUND((SUM(sb)::numeric/NULLIF(SUM(sb + cs),0))\*100,2)||'%' AS success\_rate**

**FROM batting**

**JOIN people**

**USING (playerid)**

**WHERE sb+cs >=20**

**AND yearID = 2016**

**GROUP BY player\_name**

**ORDER BY success\_rate DESC**

**--From Abram with team edits**

**SELECT p.namefirst||' '||p.namelast AS player\_name, SUM(b.sb+b.cs) AS total\_attempts, ROUND((SUM(sb)::numeric / NULLIF(SUM(sb+cs), 0)) \* 100, 2)||'%' AS success\_rate, yearid AS season**

**FROM batting AS b**

**JOIN people AS p**

**USING (playerid)**

**WHERE yearID = 2016**

**AND sb + cs >= 20**

**GROUP BY b.sb,p.namefirst,p.namelast,b.yearid**

**ORDER BY success\_rate DESC;**

**--From Jeremy after a few team edits**

**SELECT**

**SUM(sb+cs) AS total\_attempts,**

**ROUND(SUM(CAST(sb AS NUMERIC))/SUM(sb+cs)\*100,2)||'%' AS sucess\_rate,**

**namefirst|| ' ' ||namelast AS player\_name**

**FROM batting**

**JOIN people**

**USING (playerid)**

**WHERE sb+cs >=20**

**AND yearid=2016**

**GROUP BY namefirst, namelast**

**ORDER BY sucess\_rate DESC**

7. From 1970 – 2016, what is the largest number of wins for a team that did not win the world series? What is the smallest number of wins for a team that did win the world series? Doing this will probably result in an unusually small number of wins for a world series champion – determine why this is the case. Then redo your query, excluding the problem year. How often from 1970 – 2016 was it the case that a team with the most wins also won the world series? What percentage of the time?

**--From 1970 – 2016, what is the largest number of wins for a team that did not win the world series?**

**SELECT teamid AS team, yearid AS season, MAX(w) AS total\_wins**

**--CASE WHEN wswin = 'Y' AND w = 'MAX(wins)' THEN 1**

**--ELSE 0 END AS count\_years**

**FROM teams**

**WHERE wswin = 'N'**

**AND yearid BETWEEN 1970 AND 2016**

**GROUP BY yearid, teamid**

**ORDER BY total\_wins DESC**

**LIMIT 1**

**--SEA Mariners in 2001 won 116 games but lost the WS**

**--What is the smallest number of wins for a team that did win the world series?**

**SELECT name AS team\_name, SUM(w) AS number\_of\_wins, yearid AS season**

**FROM teams**

**WHERE yearid <= 2016**

**AND yearid >= 1970**

**AND wswin = 'Y'**

**AND yearid NOT IN (1981) --this doesn't change the result-------**

**GROUP BY name, yearid**

**ORDER BY number\_of\_wins ASC**

**LIMIT 1;**

**--83 wins for the St Louis Cards in 2006**

**--Doing this will probably result in an unusually small number of wins for a world series champion – determine why this is the case.**

**SELECT name AS team\_name, SUM(w) AS wins, yearid AS season**

**FROM teams**

**WHERE yearid <= 2016**

**AND yearid >= 1970**

**AND wswin = 'Y'**

**GROUP BY name, yearid**

**ORDER BY wins ASC;**

**--Above pulls yearid's for wins and we can see low wins per year 1981; That season there was a strike and the season was shortened**

**--how often did the team with the most wins win world series?**

**--team with most wins that has won WS**

**SELECT name AS team\_name, SUM(w) AS wins, yearid AS season**

**FROM teams**

**WHERE yearid <= 2016**

**AND yearid >= 1970**

**AND wswin = 'Y'**

**AND yearid NOT IN (1981)**

**GROUP BY name, yearid**

**ORDER BY yearid DESC**

**-- WITH full\_batting AS (-- SELECT-- playerid,-- SUM(sb) AS sb,-- SUM(cs) AS cs-- FROM batting-- WHERE yearid = 2016-- GROUP BY playerid-- )-- SELECT-- namefirst || ' ' || namelast AS full\_name,-- sb, -- sb + cs AS attempts,-- sb \* 100.0 / (sb + cs) AS sb\_pct-- FROM full\_batting-- INNER JOIN people-- USING(playerid)-- WHERE sb + cs >= 20-- ORDER BY sb\_pct DESC;**

8. Using the attendance figures from the homegames table, find the teams and parks which had the top 5 average attendance per game in 2016 (where average attendance is defined as total attendance divided by number of games). Only consider parks where there were at least 10 games played. Report the park name, team name, and average attendance. Repeat for the lowest 5 average attendance.

**--From Jeremy with help from team**

**--Find the teams and parks which had the top 5 average attendance per game in 2016. Report the park name, team name, and average attendance.**

**SELECT park\_name, team, SUM(attendance)/SUM(games) AS avg\_attendance**

**FROM homegames**

**INNER JOIN parks**

**USING (park)**

**WHERE year=2016**

**AND games >= 10**

**GROUP BY team, park\_name**

**ORDER BY avg\_attendance DESC**

**LIMIT 5**

**--Repeat for the lowest 5 average attendance. Report the park name, team name, and average attendance.**

**SELECT park\_name, team, SUM(attendance)/SUM(games) AS avg\_attendance**

**FROM homegames**

**INNER JOIN parks**

**USING (park)**

**WHERE year=2016**

**AND games >= 10**

**GROUP BY team, park\_name**

**ORDER BY avg\_attendance**

**LIMIT 5**

9. Which managers have won the TSN Manager of the Year award in both the National League (NL) and the American League (AL)? Give their full name and the teams that they were managing when they won the award.

**--original attempt Todd made with team help**

**SELECT p.namefirst||' '||p.namelast AS manager\_name, a.awardid AS award\_name, a.yearid AS year, a.lgid AS league\_name, m.teamid AS city**

**FROM awardsmanagers AS a**

**INNER JOIN people AS p**

**ON a.playerid = p.playerid**

**INNER JOIN managershalf AS m**

**ON p.playerid = m.playerid**

**WHERE a.awardid = 'TSN Manager of the Year'**

**AND a.lgid IN ('AL','NL')**

**GROUP BY a.playerid, a.awardid, a.yearid, a.lgid, p.namefirst, p.namelast, m.teamid**

**--Jeremy's work with team help as well as Dibran's help**

**WITH both\_leagues AS (**

**SELECT playerid --yearid, COUNT(DISTINCT(lgid)) AS counts**

**FROM awardsmanagers**

**WHERE awardid = 'TSN Manager of the Year'**

**AND lgid IN ('AL','NL')**

**GROUP BY playerid**

**HAVING COUNT(DISTINCT(lgid)) = 2**

**ORDER BY COUNT(DISTINCT(lgid)) DESC**

**)**

**,**

**award\_year AS (**

**SELECT DISTINCT am.yearid, am.playerid**

**FROM both\_leagues AS bl**

**INNER JOIN awardsmanagers AS am**

**USING (playerid)**

**--WHERE playerid IN ('johnsda02','leylaji99')**

**--AND awardid = 'TSN Manager of the Year'**

**)**

**SELECT namefirst|| ' ' || namelast AS manager\_name, m.teamid AS team, m.yearid AS season, m.lgid AS league**

**FROM people AS p**

**INNER JOIN award\_year AS ay**

**USING (playerid)**

**INNER JOIN managers AS m**

**ON ay.playerid = m.playerid AND ay.yearid = m.yearid**

10. Find all players who hit their career highest number of home runs in 2016. Consider only players who have played in the league for at least 10 years, and who hit at least one home run in 2016. Report the players' first and last names and the number of home runs they hit in 2016.

\*\*Open-ended questions\*\*

11. Is there any correlation between number of wins and team salary? Use data from 2000 and later to answer this question. As you do this analysis, keep in mind that salaries across the whole league tend to increase together, so you may want to look on a year-by-year basis.

12. In this question, you will explore the connection between number of wins and attendance.

\* Does there appear to be any correlation between attendance at home games and number of wins? </li>

\* Do teams that win the world series see a boost in attendance the following year? What about teams that made the playoffs? Making the playoffs means either being a division winner or a wild card winner.

13. It is thought that since left-handed pitchers are more rare, causing batters to face them less often, that they are more effective. Investigate this claim and present evidence to either support or dispute this claim. First, determine just how rare left-handed pitchers are compared with right-handed pitchers. Are left-handed pitchers more likely to win the Cy Young Award? Are they more likely to make it into the hall of fame?