**Database Exploration Questions**

What range of years does the provided database cover?

1871 to 2016

--Use MIN to find earliest year / MAX for latest

SELECT MAX(f.yearid)

FROM people p

INNER JOIN pitching i

ON p.playerid = i.playerid

INNER JOIN batting b

ON p.playerid = b.playerid

INNER JOIN fielding f

ON p.playerid = f.playerid

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?

43 inches, Eddie Gaedel, Saint Louis Browns

SELECT height, namefirst, namelast, debut, finalgame, b.teamid

FROM people p

INNER JOIN batting b

ON p.playerid = b.playerid

WHERE height =

(SELECT MIN(height) FROM people)

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?

|  |  |  |  |
| --- | --- | --- | --- |
| Earnings | schoolname | namefirst | namelast |
| $245,553,888 | Vanderbilt University | David | Price |
| $62,045,112 | Vanderbilt University | Pedro | Alvarez |
| $21,500,000 | Vanderbilt University | Scott | Sanderson |
| $20,512,500 | Vanderbilt University | Mike | Minor |
| $16,867,500 | Vanderbilt University | Joey | Cora |
| $12,800,000 | Vanderbilt University | Mark | Prior |
| $12,183,000 | Vanderbilt University | Ryan | Flaherty |
| $7,920,000 | Vanderbilt University | Josh | Paul |
| $4,627,500 | Vanderbilt University | Sonny | Gray |
| $4,188,836 | Vanderbilt University | Mike | Baxter |
| $3,702,000 | Vanderbilt University | Jensen | Lewis |
| $3,180,000 | Vanderbilt University | Matt | Kata |
| $2,000,000 | Vanderbilt University | Nick | Christiani |
| $1,154,400 | Vanderbilt University | Jeremy | Sowers |
| $540,000 | Vanderbilt University | Scotti | Madison |

SELECT SUM(sa.salary) AS dough,

s.schoolname, p.namefirst, p.namelast

FROM schools s

INNER JOIN collegeplaying c

ON s.schoolid = c.schoolid

INNER JOIN people p

ON p.playerid = c.playerid

INNER JOIN salaries sa

ON p.playerid = sa.playerid

WHERE LOWER(s.schoolname) LIKE '%vanderbilt%'

GROUP BY namefirst, namelast, s.schoolname

ORDER BY dough DESC

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.

Battery – 41,424

Infield – 58,934

Outfield – 29,560

SELECT SUM(PO) AS putouts,

CASE WHEN pos in ('SS', '1B', '2B', '3B')

THEN 'Infield'

WHEN pos = 'OF'

THEN 'Outfield'

WHEN pos in ('P', 'C')

THEN 'Battery'

END AS positional\_family

FROM fielding

WHERE yearid = '2016'

GROUP BY positional\_family

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?

Both home runs and strikeouts are increasing.

|  |  |  |
| --- | --- | --- |
| Average HR per game | Average Strikeouts per game | Decade |
| 0.8 | 5.63 | 1920 |
| 1.09 | 6.63 | 1930 |
| 1.05 | 7.1 | 1940 |
| 1.68 | 8.8 | 1950 |
| 1.65 | 11.39 | 1960 |
| 1.49 | 10.3 | 1970 |
| 1.61 | 10.69 | 1980 |
| 1.91 | 12.27 | 1990 |
| 2.15 | 13.12 | 2000 |
| 1.97 | 15.04 | 2010 |
|  |  |  |

WITH hr\_per\_year AS

(SELECT yearid, SUM(HR) AS total\_hr, SUM(HRA)

FROM teams

WHERE yearid >= '1920'

GROUP BY yearid),

total\_games\_per\_year AS

(SELECT yearid, SUM(g) / 2 AS total\_games

FROM teams

WHERE yearid >= '1920'

GROUP BY yearid),

total\_so\_per\_year AS

(SELECT yearid, SUM(so) AS total\_strikeouts FROM teams

WHERE yearid >= '1920'

GROUP BY yearid)

SELECT DISTINCT ROUND(AVG(hr\_per\_game)

OVER(PARTITION BY decade), 2) AS avg\_hr\_per\_game,

ROUND(AVG(so\_per\_game)

OVER(PARTITION BY decade), 2) AS avg\_so\_per\_game,

decade

FROM (SELECT

ROUND(CAST(h.total\_hr / CAST(t.total\_games AS float) AS numeric), 2)

AS hr\_per\_game,

ROUND(CAST(s.total\_strikeouts / CAST(t.total\_games AS float) AS numeric), 2) AS so\_per\_game,

(10 \* DATE\_PART('decade', TO\_DATE(h.yearid::text, 'YYYY'))) AS decade

FROM hr\_per\_year h

INNER JOIN total\_games\_per\_year t

ON h.yearid = t.yearid

INNER JOIN total\_so\_per\_year s

ON h.yearid = s.yearid) AS foo

ORDER BY decade

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.

Chris Owings

SELECT 100 \* (sb / (sb + cs)::float) AS stolen\_base\_pct,

p.namefirst, p.namelast, sb, cs

FROM batting b

INNER JOIN people p

ON p.playerid = b.playerid

WHERE b.yearid = '2016' and sb + cs >= 20

ORDER BY stolen\_base\_pct

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?

Most Wins no Series - 2001 Mariners

Least Wins with Series - LA Dodgers

1981 and 1995 had strikes

Pct Series winner won most: 22.6%

SELECT yearid, name, /\*MIN\*/MAX(w)FROM teams

WHERE yearid >= 1970

AND yearid <> 1981

AND wswin = 'N' /\*'Y'\*/

GROUP BY yearid, name

ORDER BY /\*MIN\*/MAX(w) DESC;

WITH sub AS

(SELECT distinct yearid, MAX(w)

OVER (PARTITION BY yearid) AS top\_wins

FROM teams

WHERE yearid >= 1970 ORDER BY yearid)

SELECT

(SELECT COUNT(\*)

FROM teams t

INNER JOIN sub s

ON t.yearid = s.yearid

WHERE top\_wins = w

AND t.wswin = 'Y') /

(SELECT COUNT(\*)

FROM teams t

INNER JOIN sub s

ON t.yearid = s.yearid

WHERE top\_wins = w)::float

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.

|  |  |  |
| --- | --- | --- |
| team | park\_name | avg\_attend |
| Top |  |  |
| LAN | Dodger Stadium | 45719.9 |
| SLN | Busch Stadium III | 42524.6 |
| TOR | Rogers Centre | 41877.8 |
| SFN | AT&T Park | 41546.4 |
| CHN | Wrigley Field | 39906.4 |
| Bottom |  |  |
| TBA | Tropicana Field | 15878.6 |
| OAK | Oakland-Alameda County Coliseum | 18784.0 |
| CLE | Progressive Field | 19650.2 |
| MIA | Marlins Park | 21405.2 |
| CHA | U.S. Cellular Field | 21559.2 |

SELECT team, h.park, p.park\_name,

(attendance / games::numeric) AS avg\_attend

FROM homegames h

INNER JOIN parks p

ON h.park = p.park

WHERE year = '2016'

AND games >= 10

ORDER BY avg\_attend DESC --ORDER BY avg\_attend ASC

LIMIT 5

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.

Jim Leyland Detroit/Pittsburgh

Davey Johnson Baltimore/Washington

WITH al\_awards AS

(SELECT a.yearid AS al\_year, m.teamid AS al\_team, p.namefirst, p.namelast,

awardid AS al\_award, a.playerid, a.lgid

FROM managers m

INNER JOIN awardsmanagers a

ON m.playerid = a.playerid

AND m.yearid = a.yearid

AND m.lgid = a.lgid

INNER JOIN people p

ON p.playerid = m.playerid

WHERE awardid = 'TSN Manager of the Year'

AND a.lgid = 'AL'),

nl\_awards AS

(SELECT a.yearid AS nl\_year, m.teamid AS nl\_team, p.namefirst, p.namelast, awardid AS nl\_award,

a.playerid, a.lgid

FROM managers m

INNER JOIN awardsmanagers a

ON m.playerid = a.playerid

AND m.yearid = a.yearid

AND m.lgid = a.lgid

INNER JOIN people p

ON p.playerid = m.playerid

WHERE awardid = 'TSN Manager of the Year'

AND a.lgid = 'NL')

SELECT DISTINCT al\_award, nl\_award,

al\_year, nl\_year,

al\_team, nl\_team, a.namefirst, a.namelast

FROM al\_awards a

INNER JOIN nl\_awards n

ON a.playerid = n.playerid

**Presentation Topic:** chosen from several options

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?

Left-handers are:

28% of the general population (10 games or more)

33% of Cy Young Award winners

22% of Hall of Fame inductees

More elaboration in PowerPoint presentation ‘Lefties v Righties’.

WITH pitchers AS

(SELECT yearid, playerid, SUM(g) AS total\_games

FROM pitching

GROUP BY playerid, yearid

HAVING SUM(g) > 10)

SELECT ROUND(

(SELECT COUNT(\*) as lefties

FROM people p

INNER JOIN pitchers i

ON p.playerid = i.playerid

WHERE throws = 'L') /

(SELECT COUNT(\*) as throws\_r\_l

FROM people p

INNER JOIN pitchers i

ON p.playerid = i.playerid

WHERE throws in ('L', 'R'))

::numeric, 2) as pct;

WITH pitchers AS

(SELECT yearid, playerid, SUM(g) AS total\_games

FROM pitching

GROUP BY playerid, yearid

HAVING SUM(g) > 10)

SELECT ROUND(

(SELECT COUNT(\*)

FROM pitchers p

INNER JOIN awardsplayers a

ON p.playerid = a.playerid

AND p.yearid = a.yearid

INNER JOIN people e

ON p.playerid = e.playerid

WHERE throws = 'L'

AND awardid = 'Cy Young Award') /

(SELECT COUNT(\*)

FROM pitchers p

INNER JOIN awardsplayers a

ON p.playerid = a.playerid

AND p.yearid = a.yearid

INNER JOIN people e

ON p.playerid = e.playerid

WHERE awardid = 'Cy Young Award')

::numeric, 2) as pct;

WITH pitchers AS

(SELECT yearid, playerid, SUM(g) AS total\_games

FROM pitching

GROUP BY playerid, yearid

HAVING SUM(g) > 10)

SELECT ROUND(

(SELECT COUNT(\*)

FROM pitchers p

INNER JOIN halloffame a

ON p.playerid = a.playerid

INNER JOIN people e

ON p.playerid = e.playerid

WHERE throws = 'L'

AND inducted = 'Y') /

(SELECT COUNT(\*)

FROM pitchers p

INNER JOIN halloffame a

ON p.playerid = a.playerid

INNER JOIN people e

ON p.playerid = e.playerid

WHERE inducted = 'Y')::numeric, 2) as pct;