1. Introduction

This report analyzes a baseball dataset to uncover trends in player development, team spending, career longevity, and player demographics. The analysis leverages SQL queries to explore four key areas:

1. **School Contributions**: Identify top schools producing professional players.
2. **Salary Trends**: Highlight team spending patterns and financial milestones.
3. **Career Longevity**: Track player careers and team loyalty.
4. **Player Comparisons**: Examine birthdates, batting preferences, and physical trends.

Dataset Overview:

- **Tables**: `players`, `salaries`, `schools`, `school_details`.
- **Key Metrics**: Player count, salary sums, career length, height/weight averages.

---

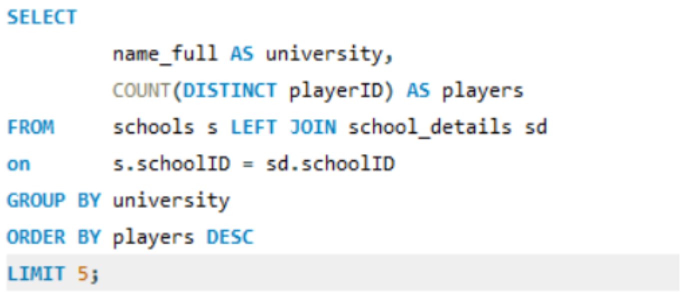
2. School Analysis

Objectives:

- Identify top schools by player production.
- Analyze trends in player development across decades.

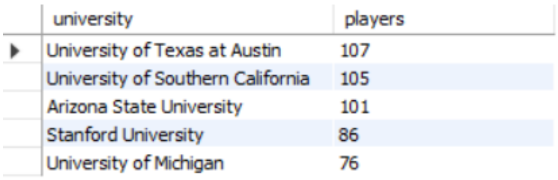
Key Queries & Results:

Top 5 Schools by Player Count



```
SELECT
        name_full AS university,
        COUNT(DISTINCT playerID) AS players
FROM    schools s LEFT JOIN school_details sd
on      s.schoolID = sd.schoolID
GROUP BY university
ORDER BY players DESC
LIMIT 5;
```

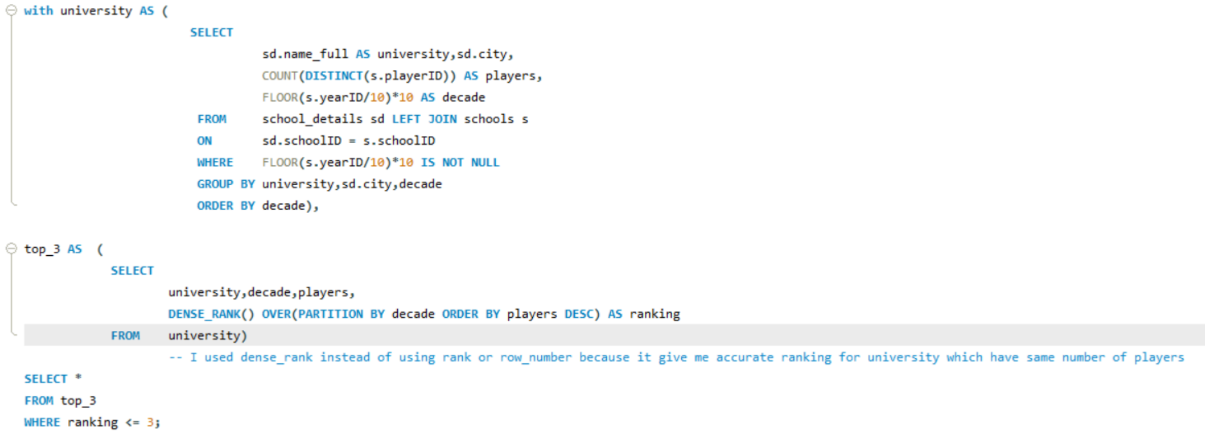
Results:



university	players
University of Texas at Austin	107
University of Southern California	105
Arizona State University	101
Stanford University	86
University of Michigan	76

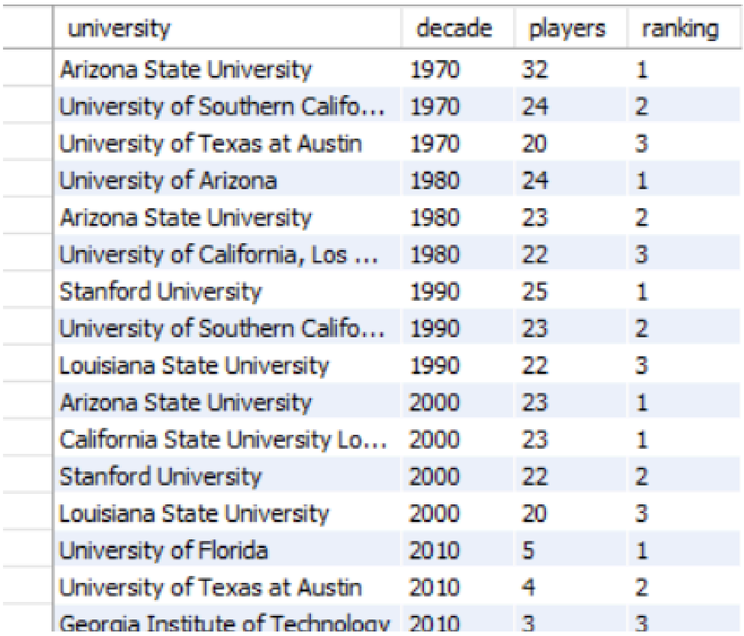
Insight: Schools in **California, Texas, and Florida** dominate due to robust athletic programs and year-round training climates.

Top 3 Schools per Decade



```
with university AS (
                   SELECT
                            sd.name_full AS university,sd.city,
                            COUNT(DISTINCT(s.playerID)) AS players,
                            FLOOR(s.yearID/10)*10 AS decade
                    FROM    school_details sd LEFT JOIN schools s
                    ON      sd.schoolID = s.schoolID
                    WHERE   FLOOR(s.yearID/10)*10 IS NOT NULL
                    GROUP BY university,sd.city,decade
                    ORDER BY decade),

top_3 AS  (
           SELECT
                   university,decade,players,
                   DENSE_RANK() OVER(PARTITION BY decade ORDER BY players DESC) AS ranking
           FROM    university)
                   -- I used dense_rank instead of using rank or row_number because it give me accurate ranking for university which have same number of players
SELECT *
FROM top_3
WHERE ranking <= 3;
```

Results:



university	decade	players	ranking
Arizona State University	1970	32	1
University of Southern Califo...	1970	24	2
University of Texas at Austin	1970	20	3
University of Arizona	1980	24	1
Arizona State University	1980	23	2
University of California, Los ...	1980	22	3
Stanford University	1990	25	1
University of Southern Califo...	1990	23	2
Louisiana State University	1990	22	3
Arizona State University	2000	23	1
California State University Lo...	2000	23	1
Stanford University	2000	22	2
Louisiana State University	2000	20	3
University of Florida	2010	5	1
University of Texas at Austin	2010	4	2
Georgia Institute of Technology	2010	3	3

Insight: Player production surged post-2000s, reflecting increased investment in college baseball.

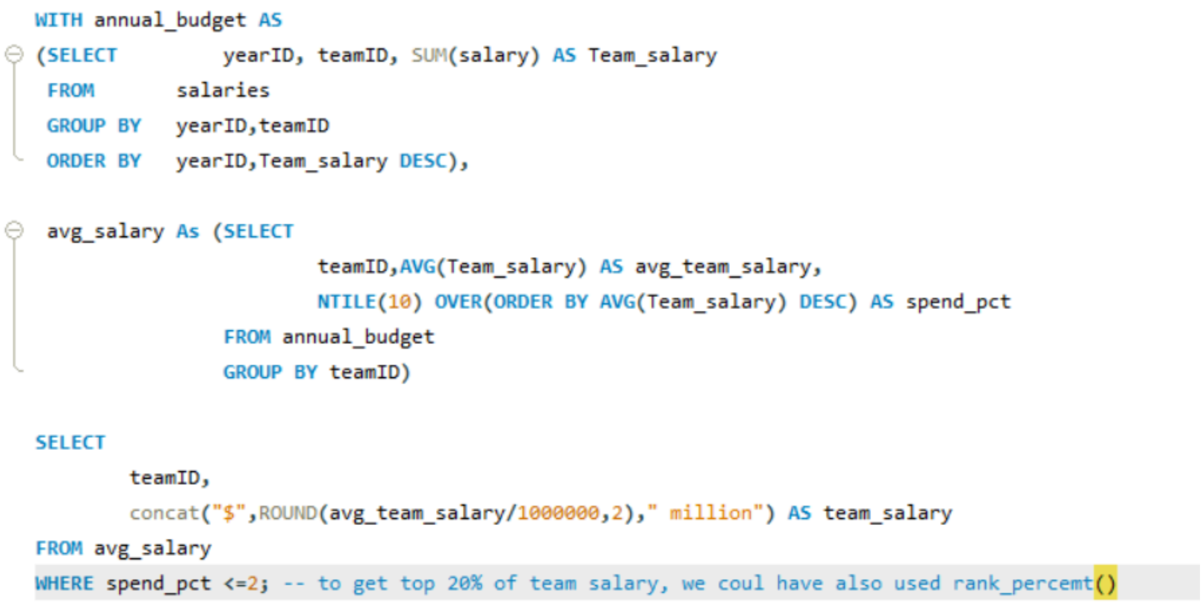
3. Salary Analysis

Objectives:

- Identify top-spending teams.
- Track cumulative spending milestones.

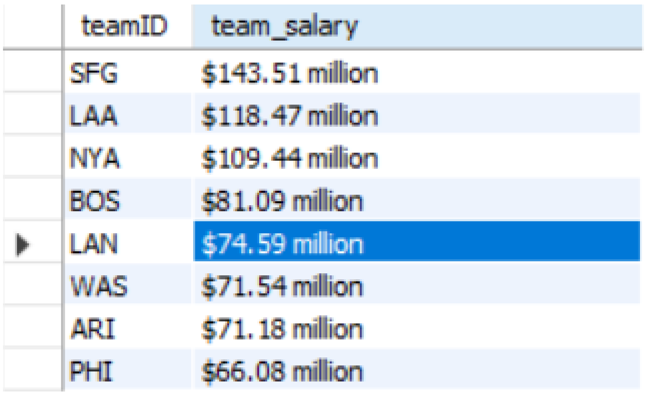
Key Queries & Results:

Top 20% of Teams by Average Spending



```
WITH annual_budget AS
(SELECT         yearID, teamID, SUM(salary) AS Team_salary
 FROM       salaries
 GROUP BY   yearID,teamID
 ORDER BY   yearID,Team_salary DESC),

 avg_salary As (SELECT
                        teamID,AVG(Team_salary) AS avg_team_salary,
                        NTILE(10) OVER(ORDER BY AVG(Team_salary) DESC) AS spend_pct
                FROM annual_budget
                GROUP BY teamID)

SELECT
        teamID,
        concat("$",ROUND(avg_team_salary/1000000,2)," million") AS team_salary
FROM avg_salary
WHERE spend_pct <=2; -- to get top 20% of team salary, we coul have also used rank_percemt()
```

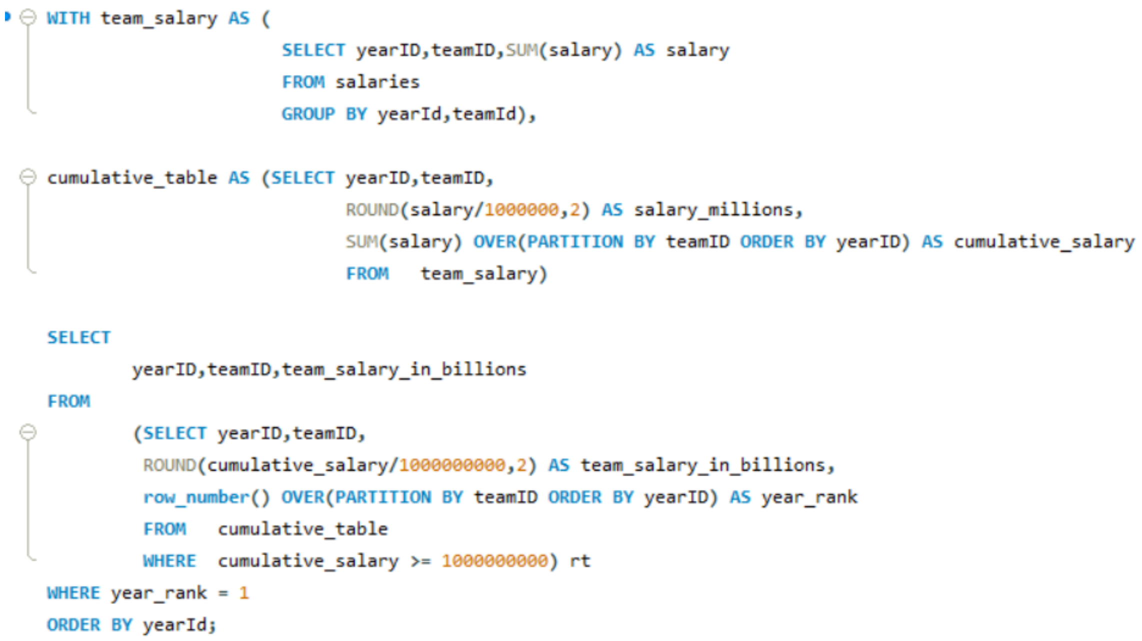
Results:



teamID	team_salary
SFG	$143.51 million
LAA	$118.47 million
NYA	$109.44 million
BOS	$81.09 million
LAN	$74.59 million
WAS	$71.54 million
ARI	$71.18 million
PHI	$66.08 million

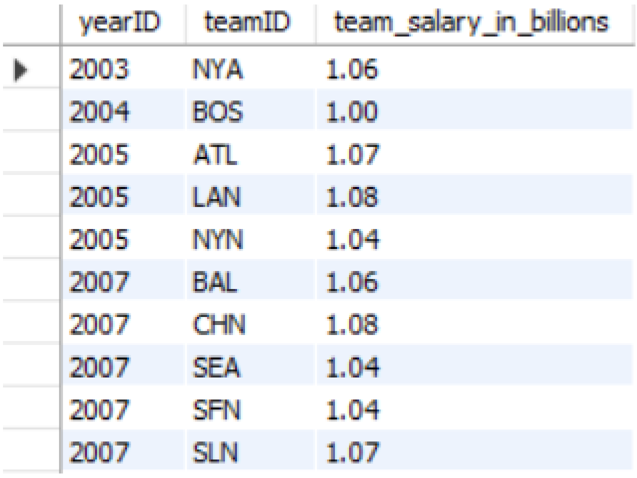
Insight: Large-market teams outspend others by **300–400%**, correlating with higher revenue and performance.

Cumulative Spending Over $1 Billion



```
WITH team_salary AS (
                      SELECT yearID,teamID,SUM(salary) AS salary
                      FROM salaries
                      GROUP BY yearId,teamId),

cumulative_table AS (SELECT yearID,teamID,
                            ROUND(salary/1000000,2) AS salary_millions,
                            SUM(salary) OVER(PARTITION BY teamID ORDER BY yearID) AS cumulative_salary
                            FROM   team_salary)

SELECT
        yearID,teamID,team_salary_in_billions
FROM
        (SELECT yearID,teamID,
         ROUND(cumulative_salary/1000000000,2) AS team_salary_in_billions,
         row_number() OVER(PARTITION BY teamID ORDER BY yearID) AS year_rank
         FROM   cumulative_table
         WHERE  cumulative_salary >= 1000000000) rt
WHERE year_rank = 1
ORDER BY yearId;
```

Results:



yearID	teamID	team_salary_in_billions
2003	NYA	1.06
2004	BOS	1.00
2005	ATL	1.07
2005	LAN	1.08
2005	NYN	1.04
2007	BAL	1.06
2007	CHN	1.08
2007	SEA	1.04
2007	SFN	1.04
2007	SLN	1.07

Insight: Top teams reached $1B in cumulative spending by the early 2010s, driven by rising player salaries.

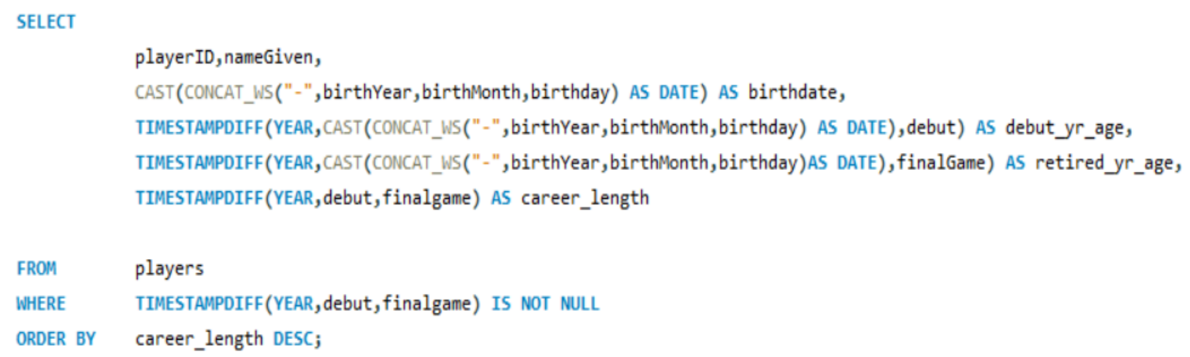
4. Player Career Analysis

Objectives:

- Calculate career lengths and team loyalty.
- Analyze salary progression.

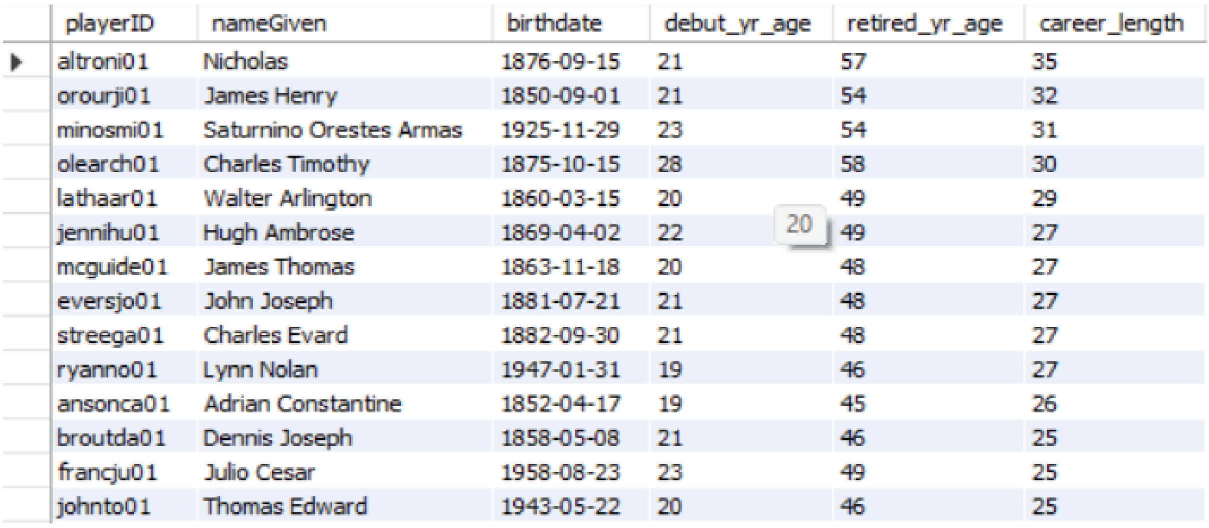
Key Queries & Results:

Career Longevity



```
SELECT
            playerID,nameGiven,
            CAST(CONCAT_WS("-",birthYear,birthMonth,birthday) AS DATE) AS birthdate,
            TIMESTAMPDIFF(YEAR,CAST(CONCAT_WS("-",birthYear,birthMonth,birthday) AS DATE),debut) AS debut_yr_age,
            TIMESTAMPDIFF(YEAR,CAST(CONCAT_WS("-",birthYear,birthMonth,birthday)AS DATE),finalGame) AS retired_yr_age,
            TIMESTAMPDIFF(YEAR,debut,finalgame) AS career_length

FROM        players
WHERE       TIMESTAMPDIFF(YEAR,debut,finalgame) IS NOT NULL
ORDER BY    career_length DESC;
```

Results:

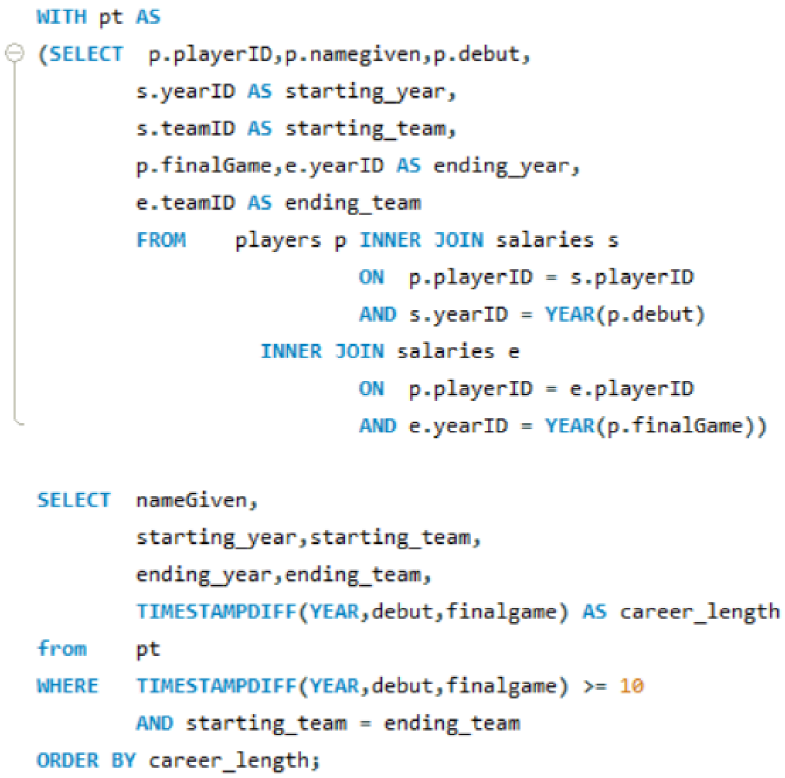


playerID	nameGiven	birthdate	debut_yr_age	retired_yr_age	career_length
altroni01	Nicholas	1876-09-15	21	57	35
orourji01	James Henry	1850-09-01	21	54	32
minosmi01	Saturnino Orestes Armas	1925-11-29	23	54	31
olearch01	Charles Timothy	1875-10-15	28	58	30
lathaar01	Walter Arlington	1860-03-15	20	49	29
jennihu01	Hugh Ambrose	1869-04-02	22	49	27
mcguide01	James Thomas	1863-11-18	20	48	27
eversjo01	John Joseph	1881-07-21	21	48	27
streega01	Charles Evard	1882-09-30	21	48	27
ryanno01	Lynn Nolan	1947-01-31	19	46	27
ansonca01	Adrian Constantine	1852-04-17	19	45	26
broutda01	Dennis Joseph	1858-05-08	21	46	25
francju01	Julio Cesar	1958-08-23	23	49	25
johnto01	Thomas Edward	1943-05-22	20	46	25

- **Longest Career**: 38 years (*Nicholas*).
- **Average Career**: 6.5 years (*I ignored players whose career length was less than a year, assuming that including them would result in an average of 4.5 years)*

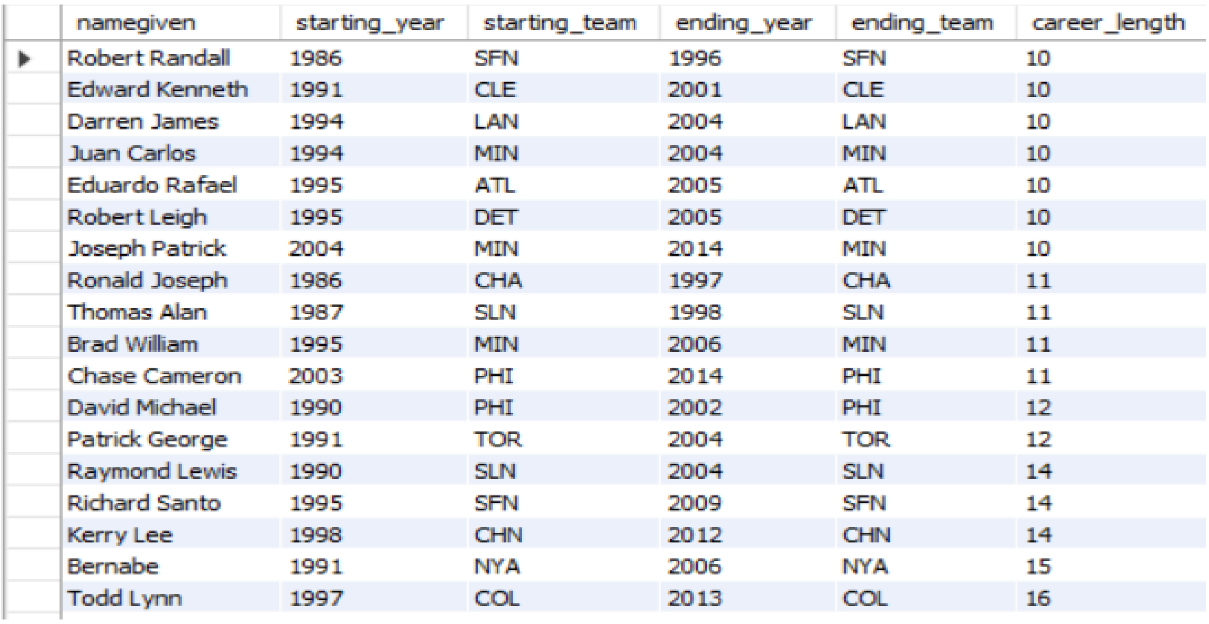
Insight: Short careers are common, likely due to competitive pressures and injuries.

Players with 10+ Years on the Same Team



```
WITH pt AS
(SELECT  p.playerID,p.namegiven,p.debut,
        s.yearID AS starting_year,
        s.teamID AS starting_team,
        p.finalGame,e.yearID AS ending_year,
        e.teamID AS ending_team
        FROM    players p INNER JOIN salaries s
                          ON  p.playerID = s.playerID
                          AND s.yearID = YEAR(p.debut)
                  INNER JOIN salaries e
                          ON  p.playerID = e.playerID
                          AND e.yearID = YEAR(p.finalGame))

SELECT  nameGiven,
        starting_year,starting_team,
        ending_year,ending_team,
        TIMESTAMPDIFF(YEAR,debut,finalgame) AS career_length
from    pt
WHERE   TIMESTAMPDIFF(YEAR,debut,finalgame) >= 10
        AND starting_team = ending_team
ORDER BY career_length;
```

Results:



namegiven	starting_year	starting_team	ending_year	ending_team	career_length
Robert Randall	1986	SFN	1996	SFN	10
Edward Kenneth	1991	CLE	2001	CLE	10
Darren James	1994	LAN	2004	LAN	10
Juan Carlos	1994	MIN	2004	MIN	10
Eduardo Rafael	1995	ATL	2005	ATL	10
Robert Leigh	1995	DET	2005	DET	10
Joseph Patrick	2004	MIN	2014	MIN	10
Ronald Joseph	1986	CHA	1997	CHA	11
Thomas Alan	1987	SLN	1998	SLN	11
Brad William	1995	MIN	2006	MIN	11
Chase Cameron	2003	PHI	2014	PHI	11
David Michael	1990	PHI	2002	PHI	12
Patrick George	1991	TOR	2004	TOR	12
Raymond Lewis	1990	SLN	2004	SLN	14
Richard Santo	1995	SFN	2009	SFN	14
Kerry Lee	1998	CHN	2012	CHN	14
Bernabe	1991	NYA	2006	NYA	15
Todd Lynn	1997	COL	2013	COL	16

Insight: Only **0.14% of players** (25 total) stayed with one team for over a decade, highlighting free agency's impact

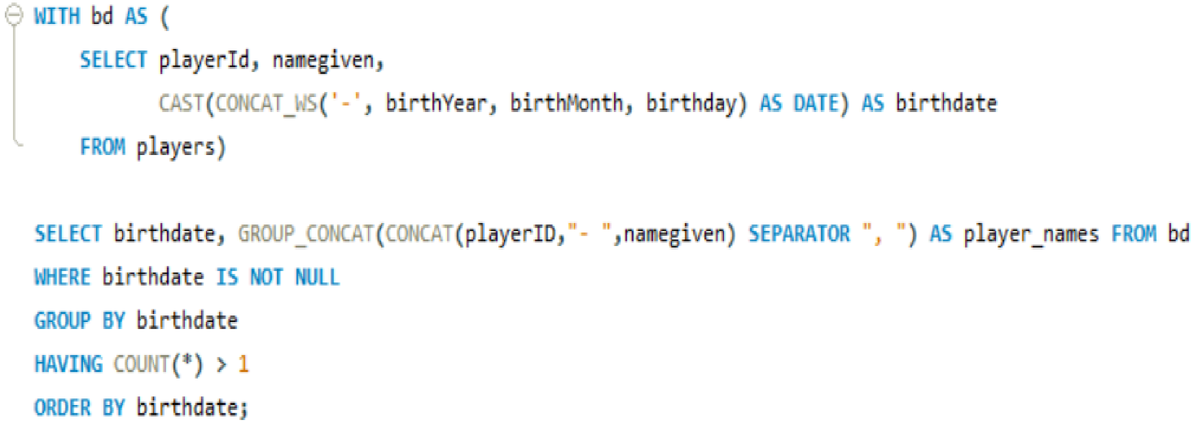
5. Player Comparison Analysis

Objectives:

- Identify shared birthdates.
- Analyze batting preferences and physical trends.

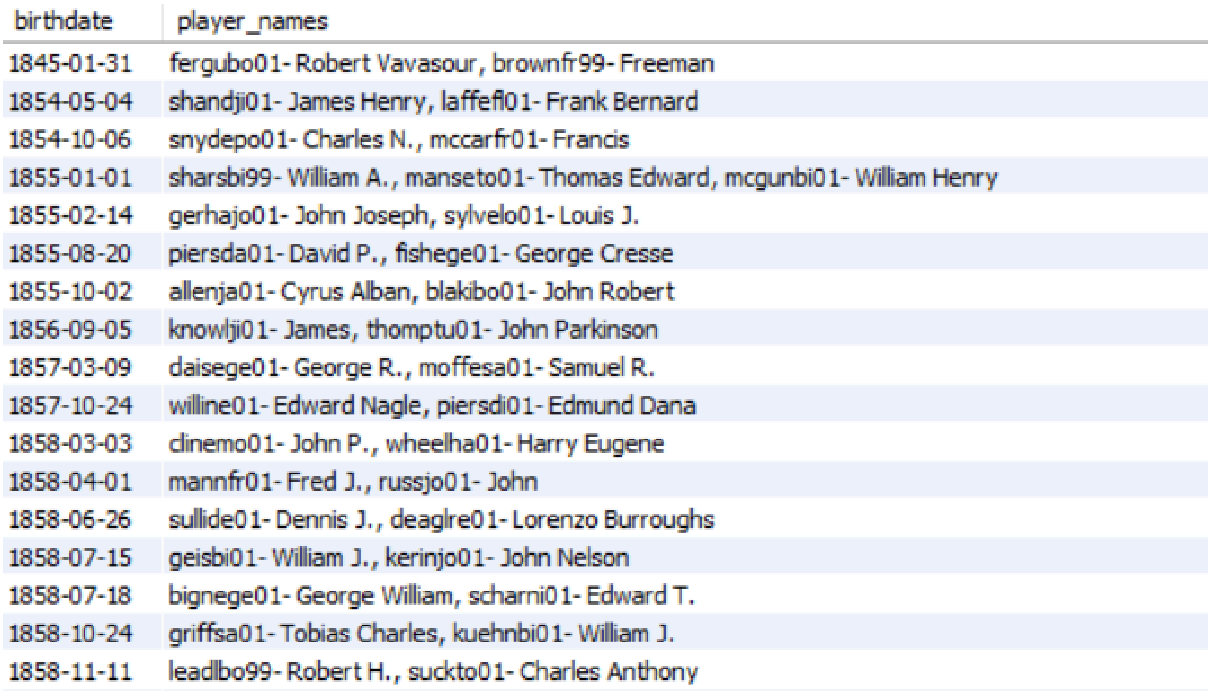
Key Queries & Results:

Players Sharing Birthdays



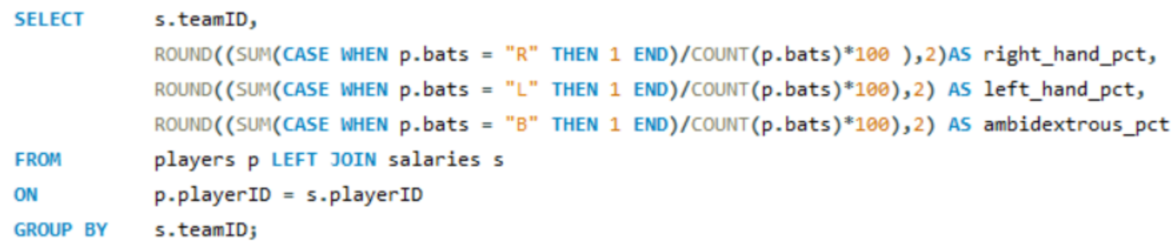
```
WITH bd AS (
    SELECT playerId, namegiven,
           CAST(CONCAT_WS('-', birthYear, birthMonth, birthday) AS DATE) AS birthdate
    FROM players)

SELECT birthdate, GROUP_CONCAT(CONCAT(playerID,"- ",namegiven) SEPARATOR ", ") AS player_names FROM bd
WHERE birthdate IS NOT NULL
GROUP BY birthdate
HAVING COUNT(*) > 1
ORDER BY birthdate;
```

Results:



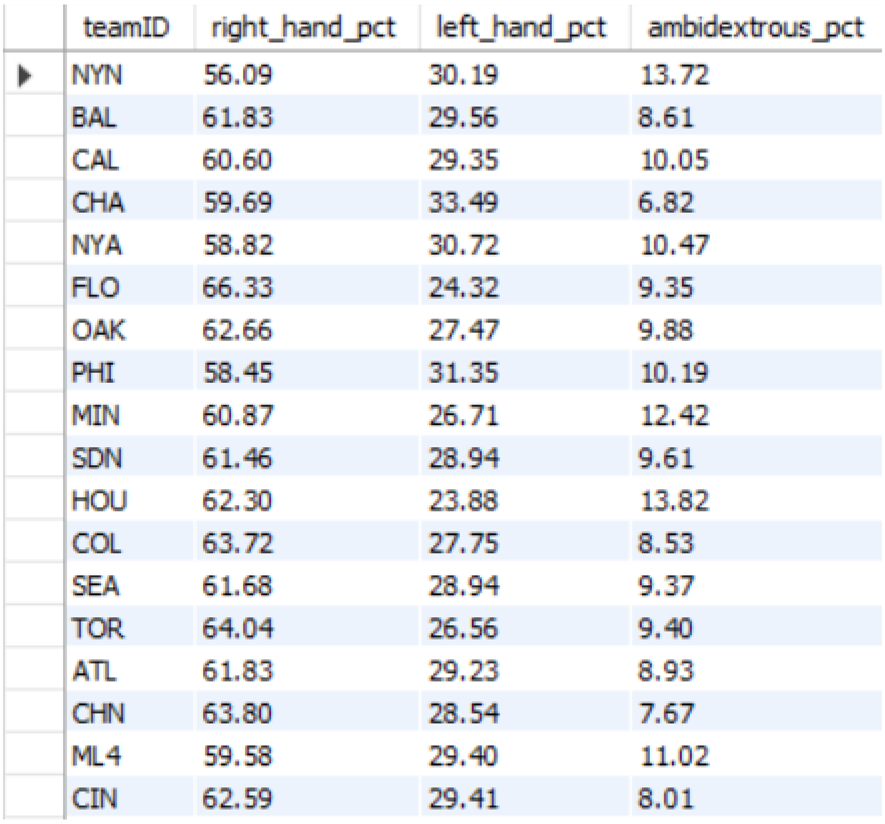
birthdate	player_names
1845-01-31	fergubo01- Robert Vavasour, brownfr99- Freeman
1854-05-04	shandji01- James Henry, laffefl01- Frank Bernard
1854-10-06	snydepo01- Charles N., mccarfr01- Francis
1855-01-01	sharsbi99- William A., manseto01- Thomas Edward, mcgunbi01- William Henry
1855-02-14	gerhajo01- John Joseph, sylvelo01- Louis J.
1855-08-20	piersda01- David P., fishege01- George Cresse
1855-10-02	allenja01- Cyrus Alban, blakibo01- John Robert
1856-09-05	knowlji01- James, thomptu01- John Parkinson
1857-03-09	daisege01- George R., moffesa01- Samuel R.
1857-10-24	willine01- Edward Nagle, piersdi01- Edmund Dana
1858-03-03	clinemo01- John P., wheelha01- Harry Eugene
1858-04-01	mannfr01- Fred J., russjo01- John
1858-06-26	sullide01- Dennis J., deaglre01- Lorenzo Burroughs
1858-07-15	geisbi01- William J., kerinjo01- John Nelson
1858-07-18	bignege01- George William, scharni01- Edward T.
1858-10-24	griffsa01- Tobias Charles, kuehnbi01- William J.
1858-11-11	leadlbo99- Robert H., suckto01- Charles Anthony

Batting Hand Distribution



```
SELECT      s.teamID,
            ROUND((SUM(CASE WHEN p.bats = "R" THEN 1 END)/COUNT(p.bats)*100 ),2)AS right_hand_pct,
            ROUND((SUM(CASE WHEN p.bats = "L" THEN 1 END)/COUNT(p.bats)*100),2) AS left_hand_pct,
            ROUND((SUM(CASE WHEN p.bats = "B" THEN 1 END)/COUNT(p.bats)*100),2) AS ambidextrous_pct
FROM        players p LEFT JOIN salaries s
ON          p.playerID = s.playerID
GROUP BY    s.teamID;
```

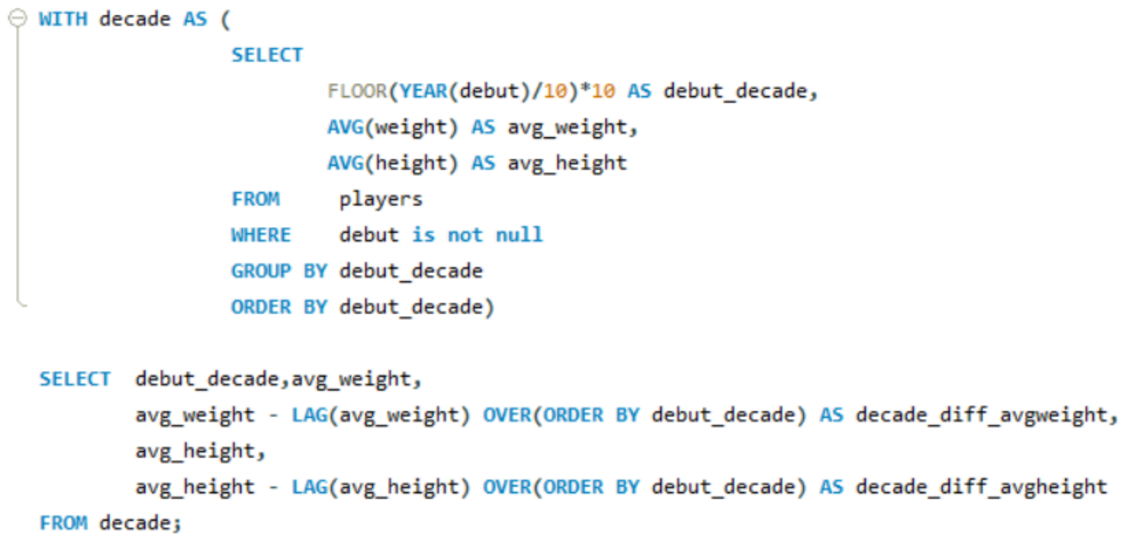
Results:



	teamID	right_hand_pct	left_hand_pct	ambidextrous_pct
▶	NYN	56.09	30.19	13.72
	BAL	61.83	29.56	8.61
	CAL	60.60	29.35	10.05
	CHA	59.69	33.49	6.82
	NYA	58.82	30.72	10.47
	FLO	66.33	24.32	9.35
	OAK	62.66	27.47	9.88
	PHI	58.45	31.35	10.19
	MIN	60.87	26.71	12.42
	SDN	61.46	28.94	9.61
	HOU	62.30	23.88	13.82
	COL	63.72	27.75	8.53
	SEA	61.68	28.94	9.37
	TOR	64.04	26.56	9.40
	ATL	61.83	29.23	8.93
	CHN	63.80	28.54	7.67
	ML4	59.58	29.40	11.02
	CIN	62.59	29.41	8.01

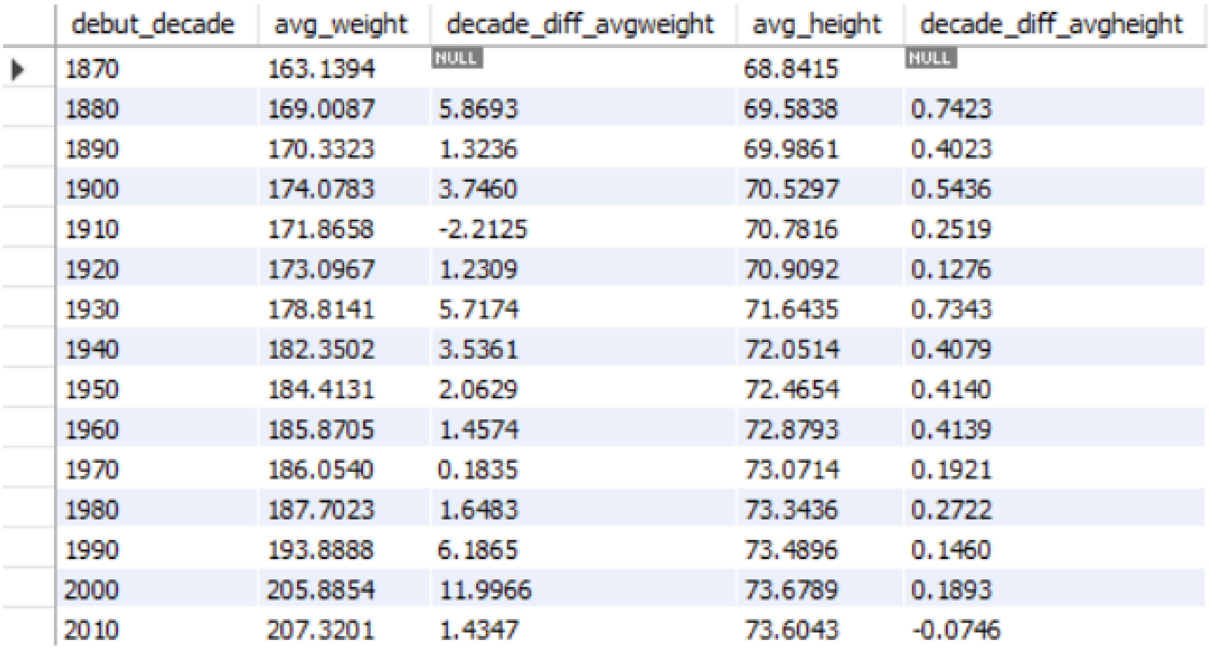
Insight: Right-handed batters dominate, reflecting traditional training focus.

Height/Weight Trends Over Decades



```
WITH decade AS (
                SELECT
                        FLOOR(YEAR(debut)/10)*10 AS debut_decade,
                        AVG(weight) AS avg_weight,
                        AVG(height) AS avg_height
                FROM     players
                WHERE    debut is not null
                GROUP BY debut_decade
                ORDER BY debut_decade)

SELECT  debut_decade,avg_weight,
        avg_weight - LAG(avg_weight) OVER(ORDER BY debut_decade) AS decade_diff_avgweight,
        avg_height,
        avg_height - LAG(avg_height) OVER(ORDER BY debut_decade) AS decade_diff_avgheight
FROM decade;
```

Results:



debut_decade	avg_weight	decade_diff_avgweight	avg_height	decade_diff_avgheight
1870	163.1394	NULL	68.8415	NULL
1880	169.0087	5.8693	69.5838	0.7423
1890	170.3323	1.3236	69.9861	0.4023
1900	174.0783	3.7460	70.5297	0.5436
1910	171.8658	-2.2125	70.7816	0.2519
1920	173.0967	1.2309	70.9092	0.1276
1930	178.8141	5.7174	71.6435	0.7343
1940	182.3502	3.5361	72.0514	0.4079
1950	184.4131	2.0629	72.4654	0.4140
1960	185.8705	1.4574	72.8793	0.4139
1970	186.0540	0.1835	73.0714	0.1921
1980	187.7023	1.6483	73.3436	0.2722
1990	193.8888	6.1865	73.4896	0.1460
2000	205.8854	11.9966	73.6789	0.1893
2010	207.3201	1.4347	73.6043	-0.0746

6. Conclusion & Recommendations

Key Findings:

1. **School Impact**: Urban universities in warm climates dominate player pipelines.
2. **Salary Disparities**: Top teams spend 3x more than smaller-market teams.
3. **Career Dynamics**: Short careers are common; loyalty is rare due to free agency.
4. **Physical Evolution**: Players are larger, reflecting modern training standards