MLB Player Comparison

This SQL portfolio project showcases a comprehensive analysis of Major League Baseball (MLB) player data from the Sean Lahman Baseball Database, covering the period from 1971 to 2014.

The project was developed in MySQL Workbench, leveraging structured queries and advanced SQL techniques to extract meaningful insights from complex relational data.

The database includes four key tables:

- players – demographic and career information of MLB athletes
- salaries – yearly salary details by player and team
- schools – player-school affiliations
- school details – information on educational institutions

The project is organized into four core analytical sections, each addressing specific business and analytical questions commonly relevant to sports analytics, talent scouting, and historical performance evaluations:

Part I: School Analysis

- Identified the number of schools that produced MLB players by decade
- Ranked top schools by player output overall and per decade
- Merged institutional data to enhance insight quality

1. In each decade, how many schools were there that produced players?

decade	num_schools
1860	2
1870	14
1880	34
1890	89
1900	148
1910	178
1920	196
1930	162
1940	142
1950	176
1960	301
1970	427
1980	473
1990	494
2000	372
2010	57

2. What are the names of the top 5 schools that produced the most players?

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

3. For each decade, what were the names of the top 3 schools that produced the most players?

decade	name_full	num_players
2010	University of Florida	5
2010	University of Texas at Austin	4
2010	University of South Carolina	3
2010	Georgia Institute of Technology	3
2000	California State University Long Beach	23
2000	Arizona State University	23
2000	Stanford University	22
2000	Louisiana State University	20
1990	Stanford University	25
1990	University of Southern California	23
1990	Louisiana State University	22
1980	University of Arizona	24
1980	Arizona State University	23
1980	University of California, Los Angeles	22
1970	Arizona State University	32
1970	University of Southern California	24
1970	University of Texas at Austin	20
1960	Arizona State University	18
1960	University of Southern California	17
1960	University of Michigan	14
1950	University of Southern California	12
1950	Michigan State University	9
1950	University of Texas at Austin	7
1940	University of Southern California	9
1940	University of Illinois at Urbana-Cham…	8
1940	University of Texas at Austin	7
1930	Duke University	14
1930	University of Texas at Austin	11

decade	name_full	num_players
1930	College of the Holy Cross	11
1930	Wake Forest University	9
1930	University of Alabama	9
1920	University of Alabama	19
1920	College of the Holy Cross	15
1920	University of Texas at Austin	12
1910	College of the Holy Cross	11
1910	St. Mary's College of California	11
1910	University of Arkansas	9
1910	Brown University	9
1910	University of Notre Dame	9
1910	University of Pennsylvania	7
1910	Washington and Lee University	7
1910	Santa Clara University	7
1910	Princeton University	7
1910	University of Michigan	7
1910	University of Alabama	7
1900	University of Notre Dame	16
1900	Manhattan College	14
1900	College of the Holy Cross	14
1900	Georgetown University	12
1900	Fordham University	12
1890	College of the Holy Cross	13
1890	Brown University	13
1890	University of Pennsylvania	9
1890	Georgetown University	6
1880	Yale University	6
1880	Brown University	5
1880	Cornell University	3
1880	College of the Holy Cross	3

1880	University of Michigan	3
1870	Yale University	3
1870	Brown University	3
1870	Washington and Lee University	1
1870	Villanova University	1
1870	Union College	1
1870	St. Mary's College of California	1
1870	Seton Hall University	1
1870	Princeton University	1
1870	Pennsylvania State University	1
1870	University of Michigan	1
1870	Manhattan College	1
1870	Harvard University	1
1870	Fordham University	1
1870	Dartmouth College	1
1860	Villanova University	1
1860	Fordham University	1

Part I: School Analysis

- Calculated and ranked team spending across all years
- Tracked cumulative salary investments and identified milestone years when teams exceeded $1 billion in total compensation
- Applied percentile filtering for comparative team budget assessments

1. Return the top 20% of teams in terms of average annual spending

teamID	avg_spend_millions
ML4	18.0
MON	20.4
CAL	22.7
PIT	33.2
FLO	35.5
KCA	39.2
SDN	39.2
OAK	39.6

2. For each team, show the cumulative sum of spending over the years

teamID	yearID	cumulative_sum_millions
ANA	1997	31.1
ANA	1998	72.4
ANA	1999	127.8
ANA	2000	179.3
ANA	2001	226.8
ANA	2002	288.5
ANA	2003	367.6
ANA	2004	468.1
ARI	1998	32.3
ARI	1999	101.1
ARI	2000	182.1
ARI	2001	267.2
ARI	2002	370.0
ARI	2003	450.6
ARI	2004	520.4
ARI	2005	582.7
ARI	2006	642.4
ARI	2007	694.5
ARI	2008	760.7
ARI	2009	833.8
ARI	2010	894.5
ARI	2011	948.2
ARI	2012	1022.0
ARI	2013	1112.1
ARI	2014	1210.0
ATL	1985	14.8
ATL	1986	31.9
ATL	1987	48.5
ATL	1988	61.2
ATL	1989	72.3

3. Return the first year that each team's cumulative spending surpassed 1 billion

teamID	yearID	cumulative_sum_billions
ARI	2012	1.02
ATL	2005	1.07
BAL	2007	1.06
BOS	2004	1.00
CHA	2008	1.07
CHN	2007	1.08
CIN	2010	1.06
CLE	2009	1.06
COL	2011	1.05
DET	2009	1.11
HOU	2008	1.03
KCA	2012	1.02
LAA	2013	1.06
LAN	2005	1.08
MIL	2014	1.05
MIN	2011	1.02
NYA	2003	1.06
NYN	2005	1.04
OAK	2012	1.05
PHI	2008	1.03
SDN	2012	1.04
SEA	2007	1.04
SFN	2007	1.04
SLN	2007	1.07
TEX	2007	1.04
TOR	2008	1.05

Part III: Player Career Analysis

- Assessed player ages at debut and retirement, and computed career lengths
- Tracked starting and ending teams per player
- Highlighted players with decade-long careers who remained loyal to a single franchise

1. For each player, calculate their age at their first game, their last game, and their career length (all in years). Sort from longest career to shortest career.

nameGiven	starting_age	ending_age	career_length
Nicholas	21	57	35
James Henry	21	54	32
Saturnino Orestes Armas	23	54	31
Charles Timothy	28	58	30
Walter Arlington	20	49	29
James Thomas	20	48	27
Lynn Nolan	19	46	27
Charles Evard	21	48	27
Hugh Ambrose	22	49	27
John Joseph	21	48	27
Adrian Constantine	19	45	26
Dennis Joseph	21	46	25
Jamie	23	49	25
Julio Cesar	23	49	25
Thomas Edward	20	46	25
William J.	21	45	24
Rickey Nelson Henley	20	44	24
John Picus	25	50	24
John Bernard	20	44	24
Early	19	43	24
Michael Thomas	18	42	24
Louis Norman	22	46	24
Jesse Russell	21	46	24
Walter James Vincent	20	43	23

2. What team did each player play on for their starting and ending years?

nameGiven	starting_year	starting_team	ending_year	starting_team
Kirk Edward	1985	CAL	1996	CHA
Mariano	1985	LAN	1997	NYA
Teodoro Valenzuela	1985	ML4	1994	ML4
Roger Alan	1985	NYN	1996	BAL
Timothy Dean	1985	OAK	1990	CIN
Vincent Maurice	1985	SLN	1997	DET
Paul Andre	1986	ATL	1999	CLE
Edward R.	1986	ATL	1988	ATL
Robert Clifford	1986	ATL	1986	ATL
Charles Edward	1986	CAL	2002	CLE
Wallace Keith	1986	CAL	2001	ANA
Roberto Martin An…	1986	CHA	2001	SLN
Ronald Joseph	1986	CHA	1997	CHA
Joel Jacob	1986	CHA	1987	CHA
Robert Thomas	1986	CHA	1994	SEA
Joseph Michael	1986	CHN	1986	CHN
Jamie	1986	CHN	2012	COL
Kalvoski	1986	CIN	1992	LAN
Tracy Donald	1986	CIN	1991	SEA
Barry Louis	1986	CIN	2004	CIN
Kurt Andrew	1986	CIN	1996	TEX
Andrew Neal	1986	CLE	1995	CAL
Scott Alan	1986	CLE	1998	TEX
James Cory	1986	CLE	1994	LAN

3. How many players started and ended on the same team and also played for over a decade?

nameGiven	starting_year	starting_team	ending_year	starting_team
Ronald Joseph	1986	CHA	1997	CHA
Barry Louis	1986	CIN	2004	CIN
Thomas Michael	1987	ATL	2008	ATL
Ellis Rena	1987	BOS	2004	BOS
Thomas Alan	1987	SLN	1998	SLN
George Kenneth	1989	SEA	2010	SEA
Samuel Peralta	1989	TEX	2007	TEX
David Michael	1990	PHI	2002	PHI
Raymond Lewis	1990	SLN	2004	SLN
Bernabe	1991	NYA	2006	NYA
Patrick George	1991	TOR	2004	TOR
Larry Wayne	1993	ATL	2012	ATL
Brad William	1995	MIN	2006	MIN
Andrew Eugene	1995	NYA	2013	NYA
Mariano	1995	NYA	2013	NYA
Richard Santo	1995	SFN	2009	SFN
Todd Lynn	1997	COL	2013	COL
Kerry Lee	1998	CHN	2012	CHN
Chase Cameron	2003	PHI	2014	PHI

Part IV: Player Comparison Analysis

- Matched players with shared birthdays
- Analyzed batting hand distribution per team
- Explored historical trends in debut age, height, and weight, with decade-over-decade comparisons

1. Which players have the same birthday?

b_day	players
1980-01-03	Bradley Keith
1980-01-10	Matthew Stephen
1980-01-12	Robert Edward
1980-01-15	Jeffrey Darrin, Matthew Thomas
1980-01-16	Brooks Litchfield, Jose Alberto
1980-01-17	Thomas Joseph, Michael Gregory
1980-01-20	Franklyn Miguel, Luis
1980-01-25	Phillip Matthew
1980-01-26	Brandon Edward, Antonio Miguel
1980-02-01	Hector R.
1980-02-03	Jared Michael
1980-02-04	Stephen John, Douglas
1980-02-07	Brad Martin
1980-02-10	Cesar David
1980-02-11	Matthew Raymond
1980-02-12	Adam James
1980-02-13	Drew Daniel
1980-02-15	Donald Thomas
1980-02-18	Walter Ernest
1980-02-20	Ryan David
1980-02-22	Ramon A.
1980-02-26	Gary Wayne
1980-02-27	John Duane
1980-03-01	James Micah
1980-03-04	John Joseph
1980-03-07	Scott Michael
1980-03-11	Christopher Allen, Richard Joseph, Daniel Cooley
1980-03-13	Byron Earl
1980-03-15	Freddie Lee
1980-03-25	Neal James
1980-03-31	Chien-Ming
1980-04-03	Justin Barnett

2. Create a summary table that shows for each team, what percent of players bat right, left and both

teamID	bats_right	bats_left	bats_both
ATL	62.0	30.6	7.4
BAL	64.4	27.0	8.6
BOS	63.1	27.8	8.8
CAL	61.9	28.8	9.4
CHA	60.2	31.8	8.0
CHN	59.4	32.1	8.5
CIN	60.4	29.9	9.6
CLE	61.2	29.6	9.2
DET	62.0	27.4	10.5
HOU	59.3	28.6	12.0
KCA	63.7	27.7	8.6
LAN	62.2	28.2	9.3
MIN	63.0	24.1	12.9
ML4	60.0	30.0	10.0
MON	61.9	26.5	11.5
NYA	63.5	27.1	9.4
NYN	59.6	29.6	10.8
OAK	60.6	29.3	10.1

PHI	62.6	29.1	8.2
PIT	65.9	27.5	6.6
SDN	63.2	27.6	9.1
SEA	61.0	28.5	10.5
SFN	63.2	26.0	10.8
SLN	63.0	27.2	9.7
TEX	63.7	25.2	11.1
TOR	65.4	25.4	9.3
COL	64.7	26.5	8.5
FLO	63.7	27.0	9.4
ANA	66.4	24.5	9.1
ARI	61.1	29.5	9.0
MIL	64.6	28.8	6.6
TBA	63.2	28.6	8.1
LAA	68.6	20.7	10.7
WAS	61.3	28.2	10.4
MIA	63.3	28.3	6.7
NYM	66.7	29.2	4.2
SFG	55.6	25.9	18.5

3. How have the average height and weight at a debut game changed over the years, and what's the decade-over-decade difference?

decade	height_diff	weight_diff
1870	NULL	NULL
1880	0.7423	5.8693
1890	0.4023	1.3236
1900	0.5436	3.7460
1910	0.2519	-2.2125
1920	0.1276	1.2309
1930	0.7343	5.7174
1940	0.4079	3.5361
1950	0.4140	2.0629
1960	0.4139	1.4574
1970	0.1921	0.1835
1980	0.2722	1.6483
1990	0.1460	6.1865
2000	0.1893	11.9966
2010	-0.0746	1.4347