

DATA WAREHOUSING  
FINAL PROJECT  
BASEBALL MLB  
PLAYERS & TEAMS  
PROFILE



GROUP 11:  
SANA A WEHSE  
RAGHAV KHURANA  
SRINIDHI ACHARLA  
SAI PRAMOD  
RAHUL BANKEY

ISM6208 | Data Warehousing | June 19, 2023

UNIVERSITY OF SOUTH FLORIDA

## Contents

<b><u>EXECUTIVE SUMMARY .....</u></b>	<b><u>2</u></b>
<b><u>PROBLEM STATEMENT .....</u></b>	<b><u>3</u></b>
<b><u>LITERATURE REVIEW, DATA COLLECTION, AND PREPARATION .....</u></b>	<b><u>3</u></b>
<b><u>METHODOLOGY AND DATABASE DESIGN .....</u></b>	<b><u>4</u></b>
<b><u>EXPLORATORY DATA ANALYSIS .....</u></b>	<b><u>6</u></b>
<b><u>REPORTING, MODELING AND STORYTELLING.....</u></b>	<b><u>12</u></b>
<b><u>CONCLUSIONS .....</u></b>	<b><u>17</u></b>
<b><u>REFERENCES .....</u></b>	<b><u>18</u></b>

## Executive Summary

Our research project aims to provide valuable insights into Major League Baseball (MLB) teams by analyzing their composition, players' profile, and factors influencing their performance. By examining historical data and current trends, we will identify patterns and factors that contribute to team success on the field. This knowledge will help teams strategize and make informed decisions to enhance their competitive edge and attract better players.

Additionally, we will explore player demographics, primary positions, and performance metrics to uncover correlations between player attributes and team success. This information will assist teams in player recruitment, contract negotiations, and overall team composition. We will also assess the impact of external factors such as average team salary, hometown weather conditions, and fans' attendance on win percentages. By understanding these relationships, teams can adapt strategies, allocate resources effectively, and engage with their fan base more meaningfully.

The outcomes of our research will benefit MLB teams, fans, and stakeholders, enabling data-driven decision-making, deeper fan engagement, and optimized investments in the baseball industry.



## Problem Statement

Baseball history takes us back to the 18<sup>th</sup> century, which makes it the oldest major league sport in the United States and Canada. The sport's popularity exploded in the 20<sup>th</sup> century as it became one of America's national pastimes, attracting big business and millions of dollars in profits as people continue to attend games, and bring their children to watch events or support their progenies as they learn to play the sport.

This research will focus on analyzing the various Major League Baseball (MLB) teams' composition, the profile of their players, and evaluate some of the factors that may impact their winnings. Specific research questions will address:

- **What are MLB teams' performance, players' profiles, and fans attendance?**
- **Is there a correlation between the percentage of games won and the team's average salary?**
- **Is there a correlation between the percentage of games won and the weather in the team's hometown?**
- **Is there a correlation between the percentage of games won and fans' attendance?**

## Literature Review, Data Collection, and Preparation

According to the United States Census Bureau of each Major League team's metropolitan statistical areas (MSA) study, published in 2021, baseball includes 30 Major League teams. Twenty MSA teams in the United States and one in Toronto, Canada. Within the US, four MSAs have 2 teams each, namely Chicago, New York, Los Angeles, and San Francisco-Oakland. All other MSAs have only one team. This research used various publicly available data sources from sports websites and for the 2022 year. This data was retrieved from various websites, using the "get data" function within Power BI. Data sources used included:

The Spotrtract website - [MLB 2022 Payroll Tracker | Spotrac](#), provided information about players' salary for each MLB Team, which were aggregated from individual player's payroll salary and included base salary, incentives, and prorated signing bonus.

Home team weather information was procured from the publicly available - Current results weather and science facts website- [Average Annual Temperatures for Large US Cities - Current Results](#), and looked at the average annual temperature for large US cities, which was joined to team's home town city. This information was derived from the National Centers for Environmental Information (NOAA).

MLB team attendance was sourced from ESPN website - [2022 MLB Attendance - Major League Baseball – ESPN](#), which provided data for fans turnout at home games which take place in a players' home team field versus away or road games played at their competition's home town location.

A different ESPN website -

[https://www.espn.com/mlb/stats/rosters/\\_/sort/average\\_age/order/true](https://www.espn.com/mlb/stats/rosters/_/sort/average_age/order/true), provided MLB roster analysis information related to the number of players who are right handed versus left handed or switchers. These latter have the ability to bat with either arm. Similar information about pitchers was provided as well as players' average height, weight, and age for each team.

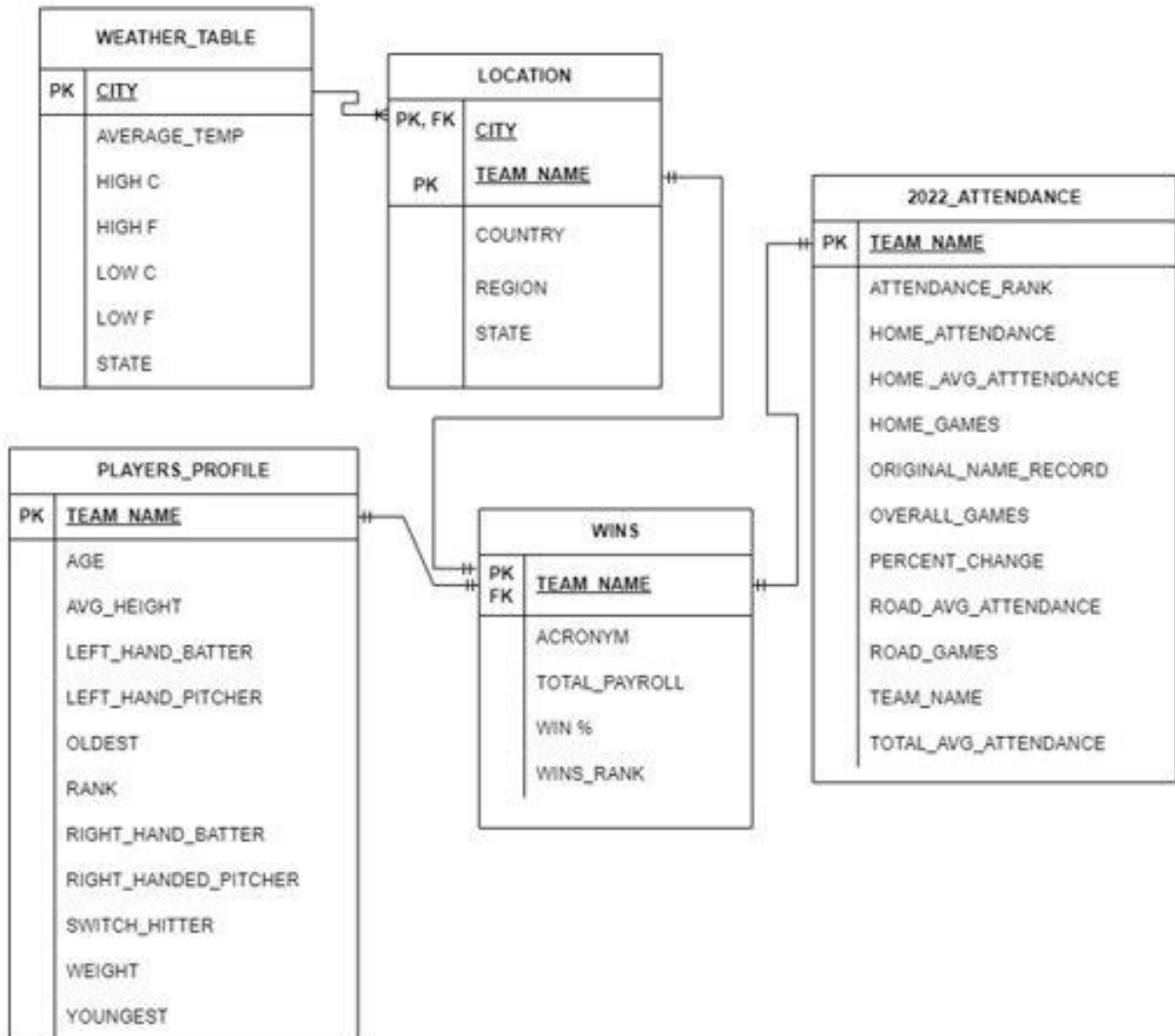
Our hometown MLB team, the Tampa Bay Rays' roster information for each player was derived from the USA Today website, <https://usatoday.sportsdirectinc.com/baseball/mlb-teams.aspx?page=/data/mlb/teams/rosters/roster2960.html>, and included each player's main position, number, height, weight, date of birth, and salary.

## Methodology and Database Design

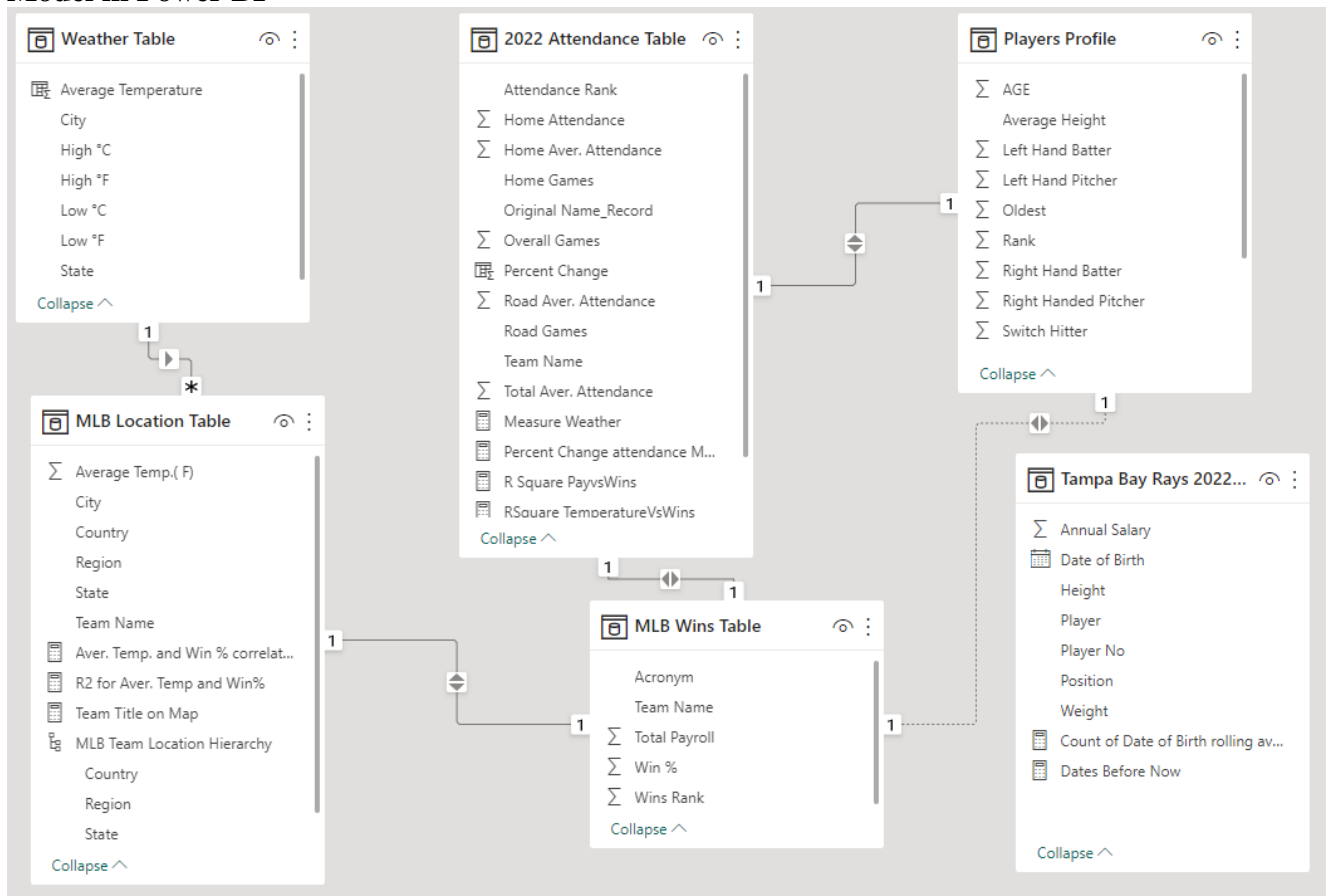
We extracted the data from various websites using Power BI "get data" functionality. Data was cleaned-up using Power BI query editor for the most part and excel. For instance, players' height was converted from feet and inches to inches only, teams name was edited to allow us to join tables by team name. We had to export cities temperature information to excel and manually cleaned-up cities' names and location, as well as added temperatures for 2 cities that were missing from the dataset. Column headers were relabeled. The first row of data was converted to column headers, and extraneous data from the source website were deleted where needed.

Our data model included 6 tables. Four tables had one-to-one relationship using Team name as a join, and 2 tables used city as a join in a one-to-many relationship. One table for the Tampa Bay Rays was analyzed as a standalone table. Here is a view of our data model.

## Star Schema (ER Diagram)



## Model in Power BI



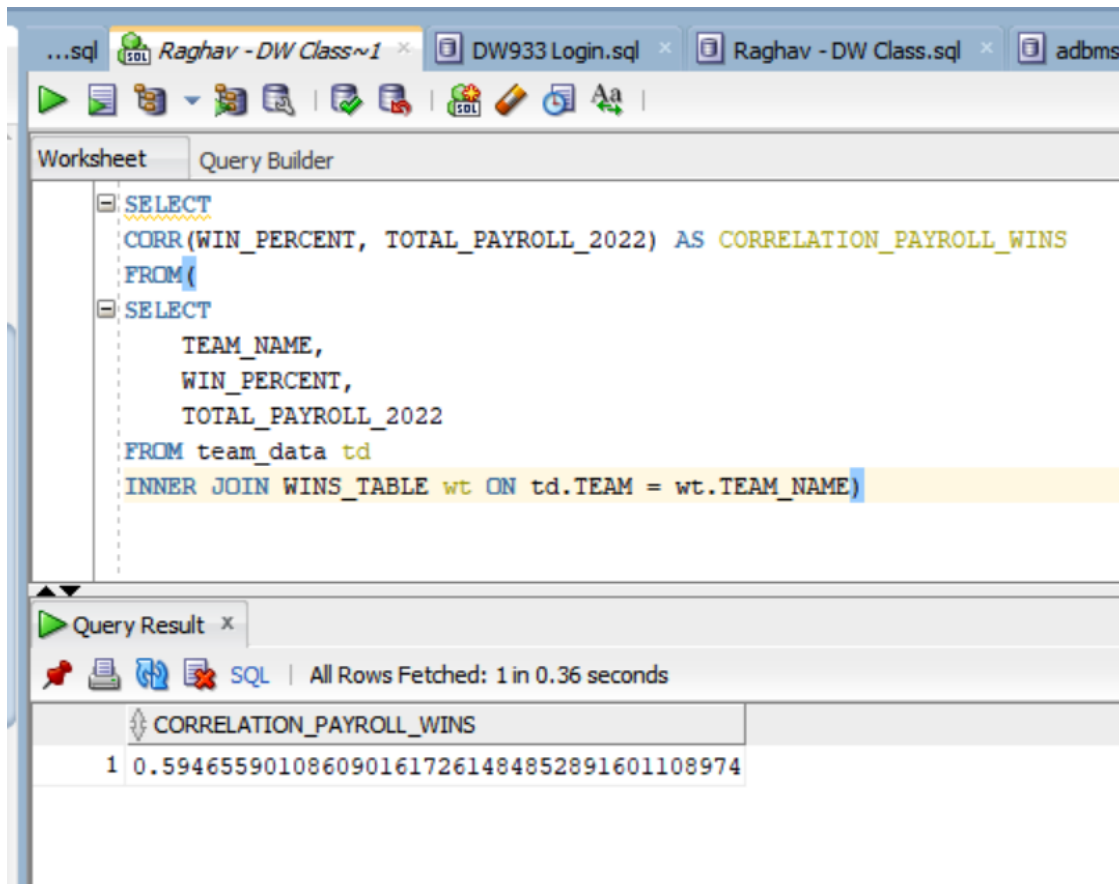
## Exploratory Data Analysis

Several correlation analyses were performed to identify factors that may impact MLB teams' percentage of games won such as salary, temperature of their respective hometown, and fans attendance. A deeper dive into one of the MLB teams, the Tampa Bay Rays, was performed to look at individual players' profile and characteristics. This analysis leveraged the use of Power BI Quick Measure and custom Measures functionality. Standard out of the box Power BI visualizations were used for the most part, such as a stacked bar, table matrix, card, line chart, pie chart, dual y-axis, a bubble chart, dynamic slicer and title. Quick measures were used to calculate correlation coefficients, and custom measures were leveraged to display the R-square and other calculations. A custom visual certified by Microsoft, Play-Axis, was used to show an animated trend line with a trail. The Power BI map displayed the location of each team, leveraging a hierarchy with drill down from Country to region, state, and city.

We run some queries which helped us get a good understanding of the database structure before moving on to visualize the outputs we had seen and been inspired by.

**Is there any correlation between MLB team's wins percentage and team's average salary?**

```
SELECT  
CORR(WIN_PERCENT, TOTAL_PAYROLL_2022) AS CORRELATION_PAYROLL_WINS  
FROM(  
SELECT  
    TEAM_NAME,  
    WIN_PERCENT,  
    TOTAL_PAYROLL_2022  
FROM team_data td  
INNER JOIN WINS_TABLE wt ON td.TEAM = wt.TEAM_NAME)
```



The screenshot shows a SQL query editor with the following query entered:

```
SELECT  
CORR(WIN_PERCENT, TOTAL_PAYROLL_2022) AS CORRELATION_PAYROLL_WINS  
FROM(  
SELECT  
    TEAM_NAME,  
    WIN_PERCENT,  
    TOTAL_PAYROLL_2022  
FROM team_data td  
INNER JOIN WINS_TABLE wt ON td.TEAM = wt.TEAM_NAME)
```

The query result is displayed in a table below the editor:

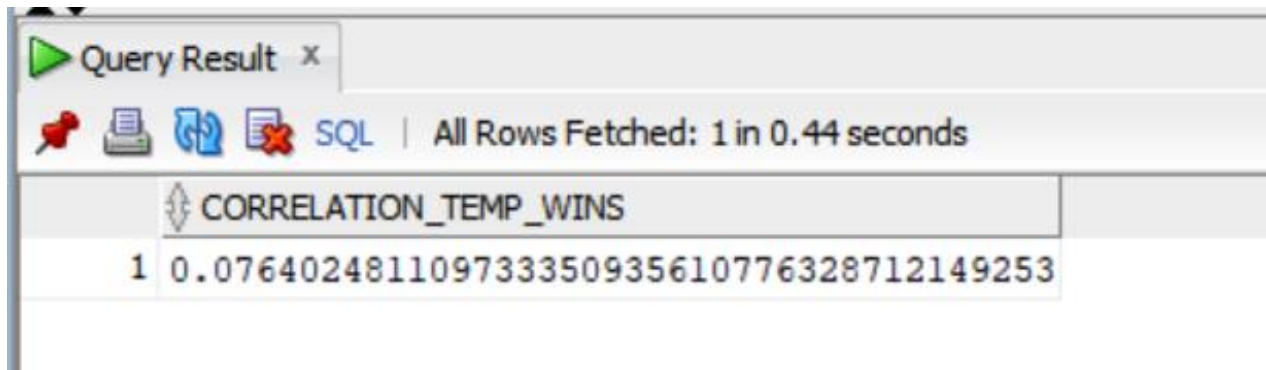
	CORRELATION_PAYROLL_WINS
1	0.5946559010860901617261484852891601108974

This shows that there is a medium correlation between a team's total salary and the percentage of games won.



**Is there a correlation between win percentage and high temperature on game day?**

```
SELECT
CORR(WIN_PERCENT, AVG_CITY_TEMP_F_HIGH) AS CORRELATION_TEMP_WINS
FROM(
SELECT
    TEAM_NAME,
    WIN_PERCENT,
    AVG_CITY_TEMP_F_HIGH
FROM team_data td
INNER JOIN WINS_TABLE wt ON td.TEAM = wt.TEAM_NAME);
```



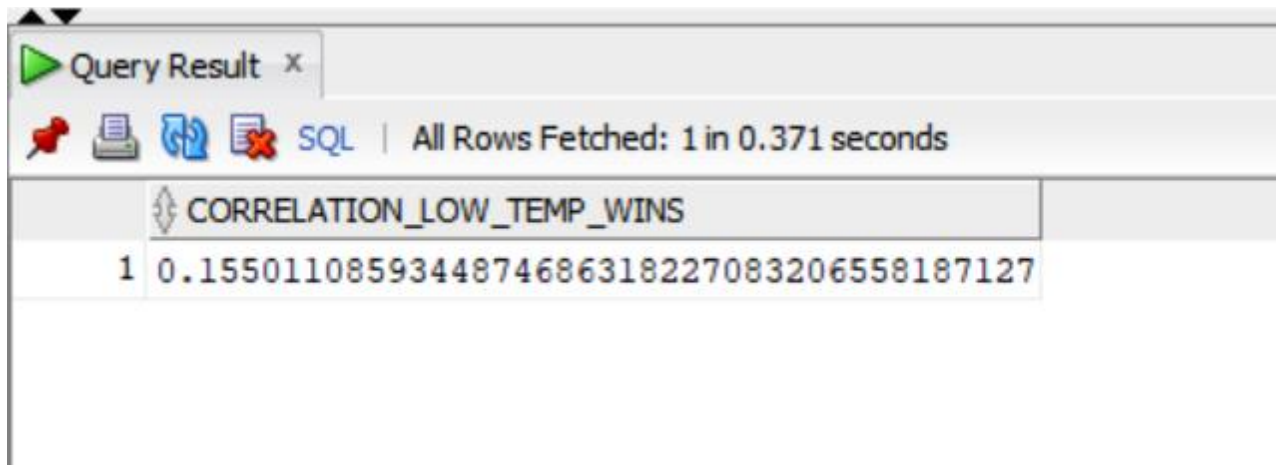
The screenshot shows a 'Query Result' window with a toolbar containing icons for a pin, print, refresh, and error, along with an 'SQL' label. The status bar indicates 'All Rows Fetched: 1 in 0.44 seconds'. The result table has one column, 'CORRELATION\_TEMP\_WINS', and one row with the value '0.076402481109733350935610776328712149253'.

CORRELATION_TEMP_WINS
0.076402481109733350935610776328712149253

There is almost no correlation between the percentage of games won and high temperature on game day.

**Is there a correlation between win percentage and low temperature on game day?**

```
SELECT
CORR(WIN_PERCENT, AVG_CITY_TEMP_F_LOW) AS
CORRELATION_LOW_TEMP_WINS
FROM(
SELECT
    TEAM_NAME,
    WIN_PERCENT,
    AVG_CITY_TEMP_F_LOW
FROM team_data td
INNER JOIN WINS_TABLE wt ON td.TEAM = wt.TEAM_NAME);
```



Query Result x

SQL | All Rows Fetched: 1 in 0.371 seconds

	CORRELATION_LOW_TEMP_WINS
1	0.1550110859344874686318227083206558187127

There is a low correlation between the percentage of games won and low temperature on game day.

**How do the average age of players vary across teams with high and low win percentages?  
Are teams with high win percentages mostly comprised of younger players?**

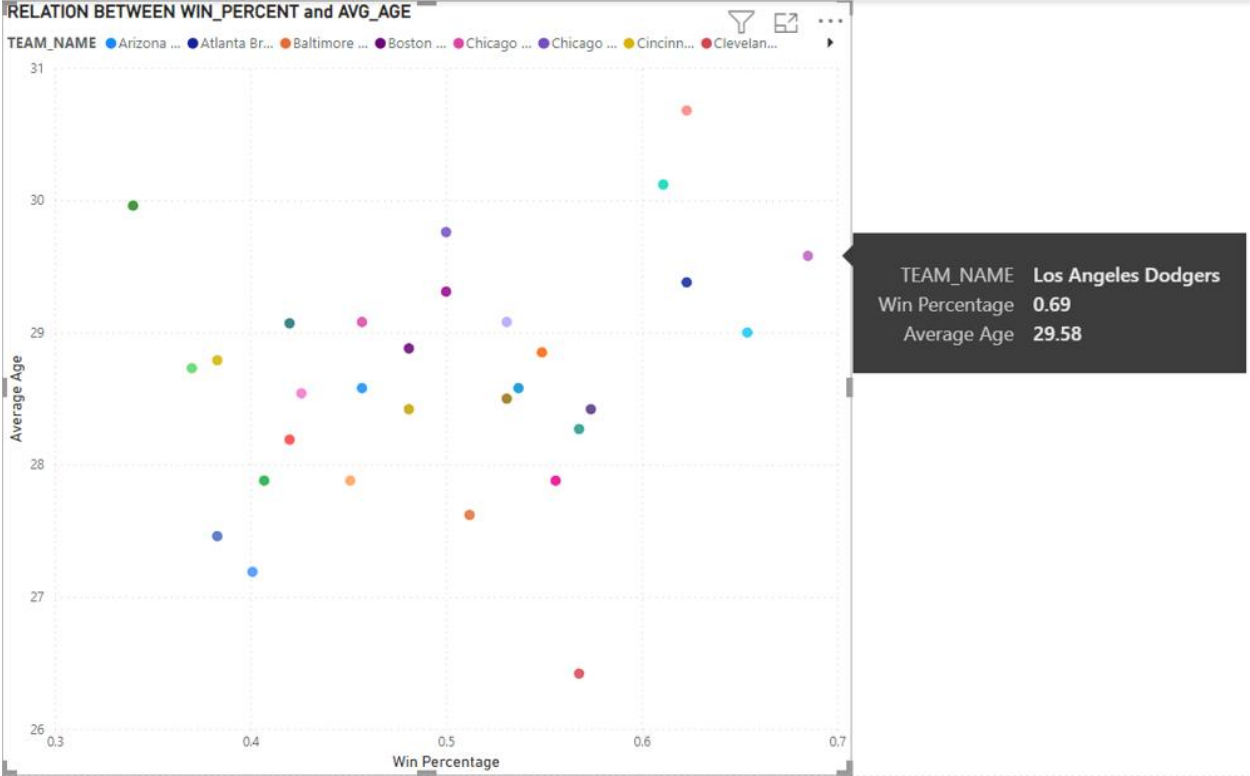
To help answer this question, we decided to display the team members' average age and the percentage games won.

```
SELECT
    TEAM_NAME,
    WIN_PERCENT,
    AVG_AGE
FROM team_data td
INNER JOIN WINS_TABLE wt ON td.TEAM = wt.TEAM_NAME
```

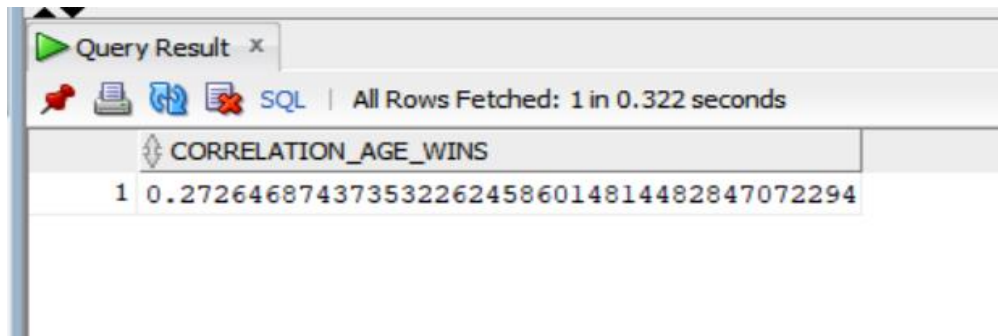


	TEAM_NAME	WIN_PERC...	AVG_AGE
1	Los Angeles Dodgers	0.685	29.58
2	Houston Astros	0.654	29
3	New York Mets	0.623	30.68
4	Atlanta Braves	0.623	29.38
5	New York Yankees	0.611	30.12
6	St. Louis Cardinals	0.574	28.42
7	Toronto Blue Jays	0.568	28.27
8	Cleveland Guardians	0.568	26.42
9	Seattle Mariners	0.556	27.88
10	San Diego Padres	0.549	28.85
11	Philadelphia Phillies	0.537	28.58
12	Milwaukee Brewers	0.531	29.08
13	Tampa Bay Rays	0.531	28.5
14	Baltimore Orioles	0.512	27.62
15	Chicago White Sox	0.5	29.76
16	San Francisco Giants	0.5	29.31
17	Minnesota Twins	0.481	28.42

Scatterplot for average age and win percentage:



**Is there a correlation between the Average Age of a team's players and Win percentage?**



Query Result x

SQL | All Rows Fetched: 1 in 0.322 seconds

	CORRELATION_AGE_WINS
1	0.2726468743735322624586014814482847072294

There is low correlation between a team's percentage of games won and the average age of a team's players.

These are some of the questions we asked ourselves before moving on to visualize the data.

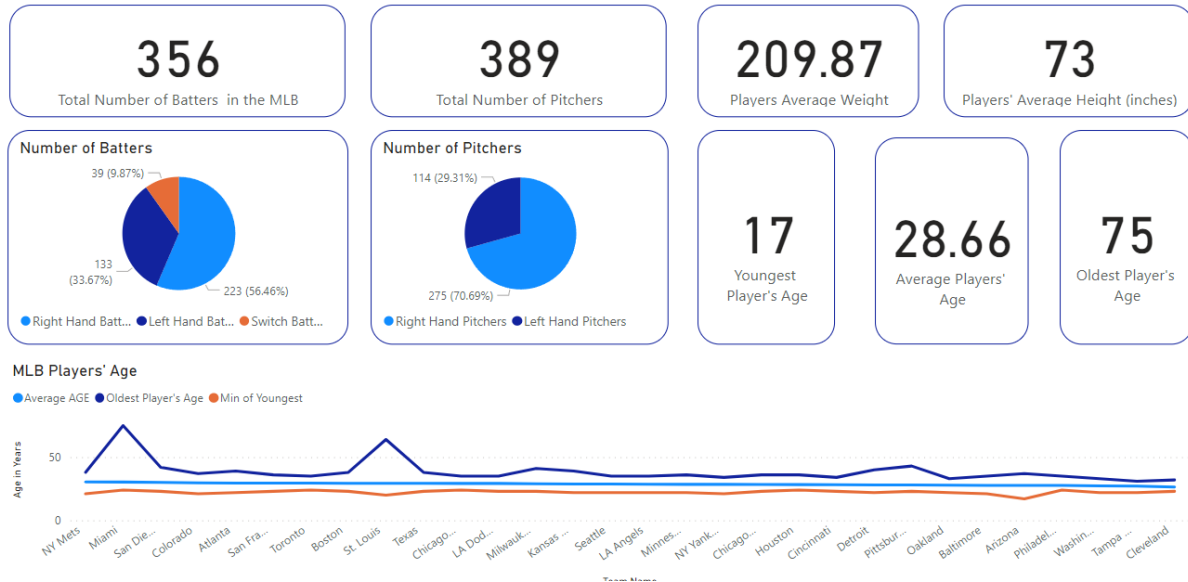
## Reporting, Modeling and Storytelling

This map shows the location of each MLB team's hometown location within the United States and Canada, along with average temperature, and percentage of wins. We can see a total of 30 teams.



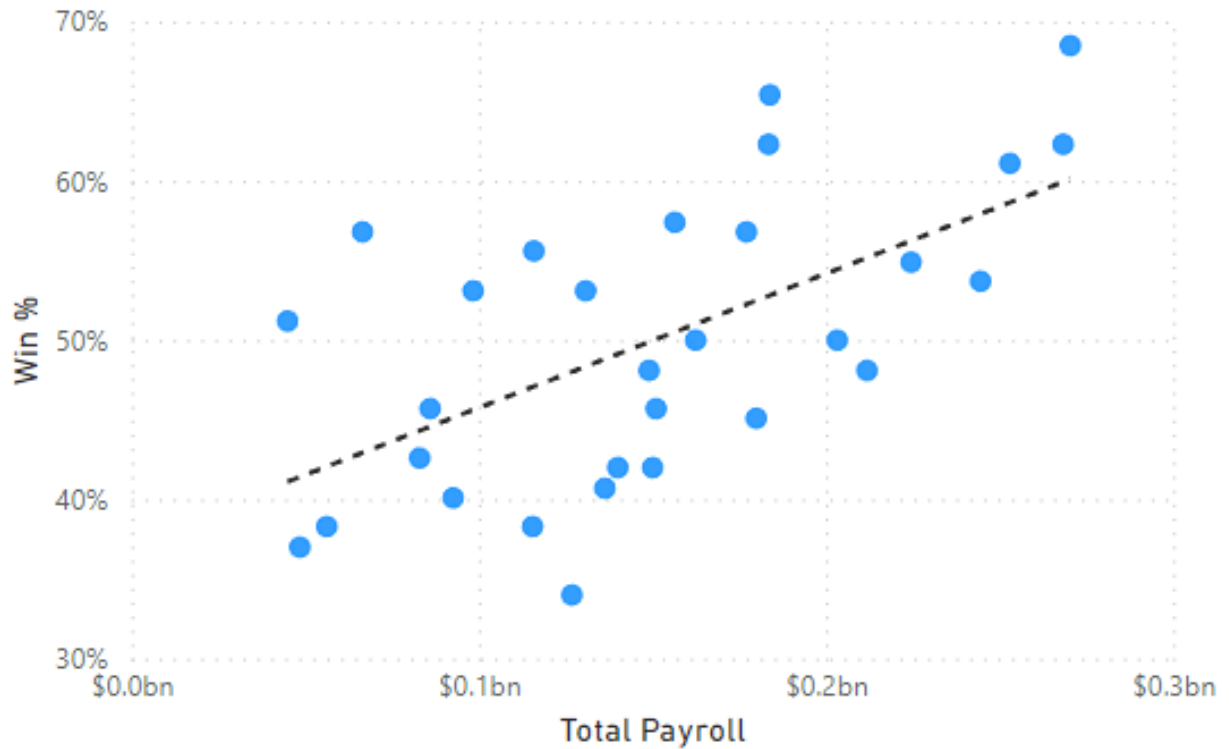
This dashboard provides aggregate information about the total number of batters and pitchers in the MLB, which respectively added-up to 356 and 389 in 2022. The average players' weight was 210 pounds, and average height was 73 inches or 6 feet and 1 inch. Most players are right-handed batters (56%), and right-handed pitchers (71%). The youngest player is 17 years old, part of the Diamond Back Arizona team and the oldest is 75 years old, part of the Miami Marlins. The youngest and oldest players are visibly outliers, which will be excluded in future analysis.

## MLB Players' Profile (2022)



Overall results show a medium to strong correlation between a team's percentage of wins versus salary with a correlation coefficient of .6 and a R2 of .36.

Total Payroll and Win %



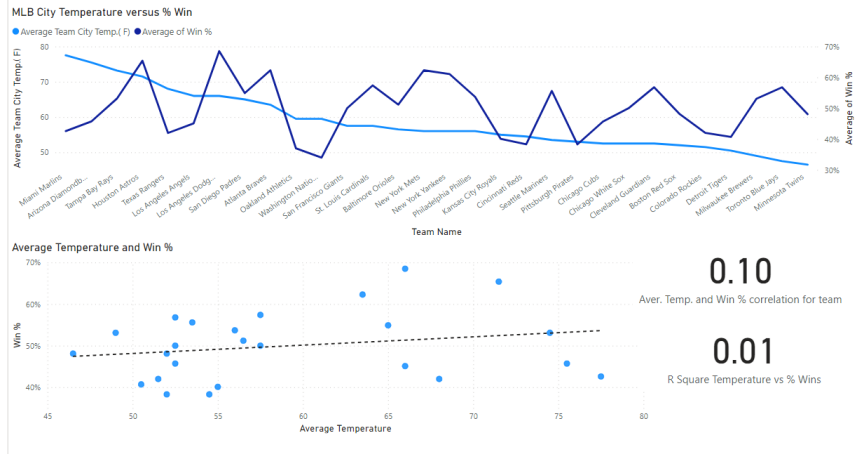
0.6

Team's Pay vs. Wins Correlation Coef

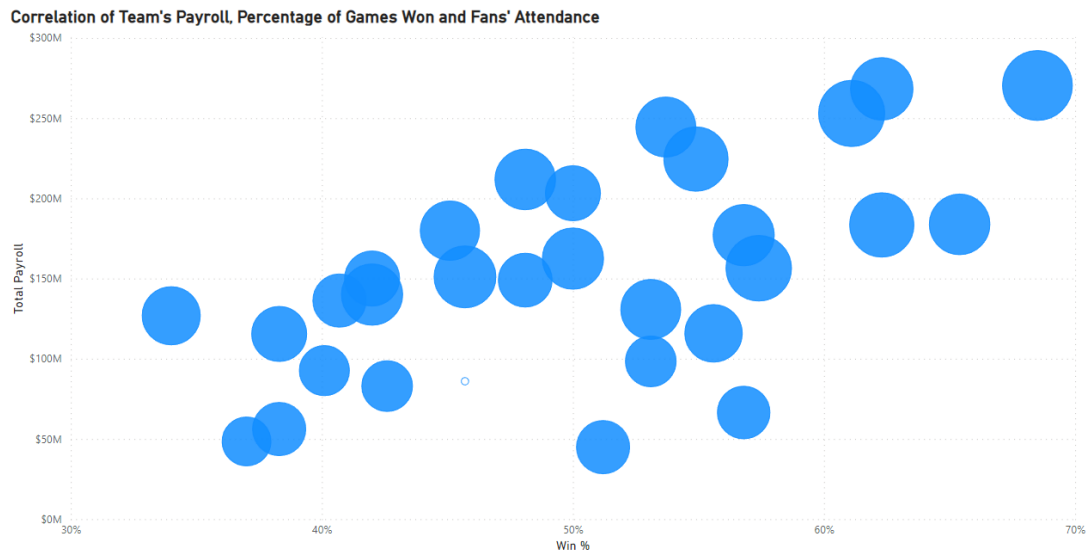
0.36

R Square PayvsWins

A similar analysis was performed between temperature in each team's home city and percentage of wins and showed a very a low correlation coefficient of .10. This can be explained by the fact that teams with hometowns averaging lower temperatures travel to practice in the warmer states, such as Florida.



A different correlation analysis using the bubble chart visual revealed a positive correlation between percentage of games won and total team's payroll. The size of the bubbles was used to introduce a third variable of fan attendance. For the most part, we can see that teams with larger bubbles in the top right quadrant of the visual tend to bring in the highest wins and salaries.

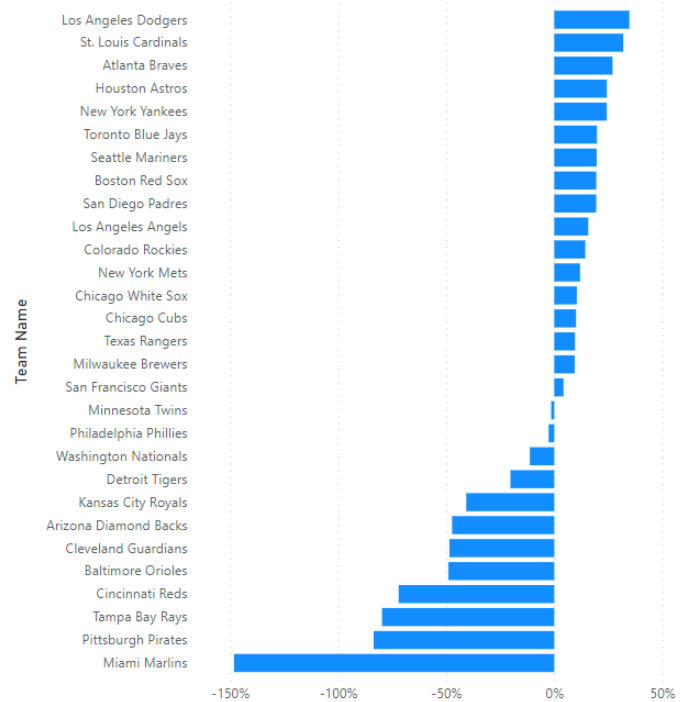


The visual below shows the teams that have a higher positive percentage of fans attending home games at the top versus away or road games. The teams listed at the bottom show a negative number with lower percentage attendance at home games versus away games.



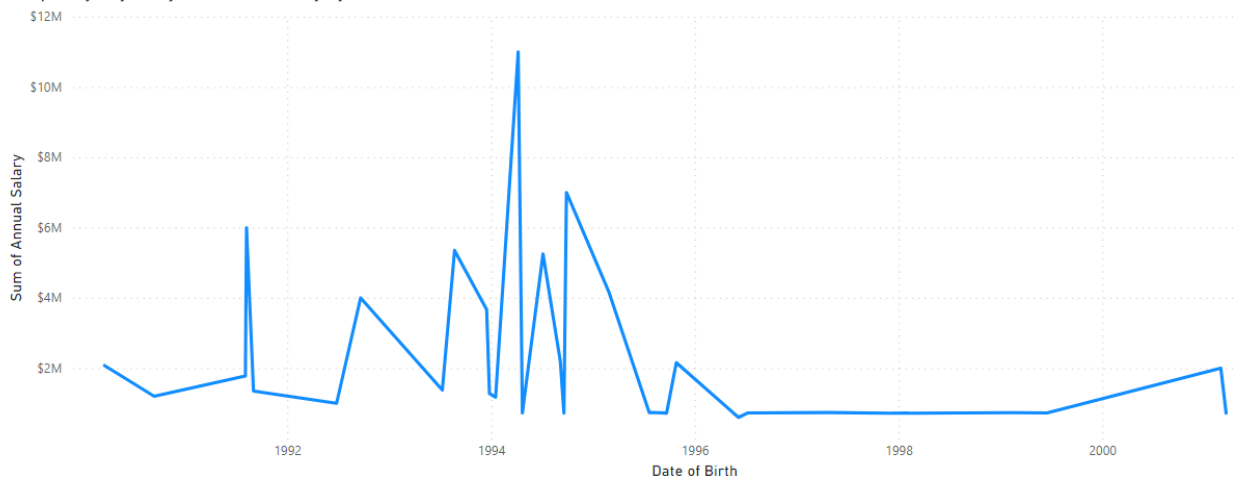
Team Name	Home Aver. Attendance	Road Aver. Attendance	Percent Change
Los Angeles Dodgers	47,671	31,104	34.75%
St. Louis Cardinals	40,994	27,895	31.95%
Atlanta Braves	38,641	28,216	26.98%
Houston Astros	33,197	25,112	24.35%
New York Yankees	40,207	30,418	24.35%
Toronto Blue Jays	32,763	26,280	19.79%
Seattle Mariners	28,590	22,973	19.65%
Boston Red Sox	32,408	26,106	19.45%
San Diego Padres	36,931	29,754	19.43%
Los Angeles Angels	30,339	25,561	15.75%
Colorado Rockies	32,467	27,801	14.37%
New York Mets	33,308	29,330	11.94%
Chicago White Sox	24,704	22,118	10.47%
Chicago Cubs	32,305	29,044	10.09%
Texas Rangers	24,831	22,443	9.62%
Milwaukee Brewers	30,155	27,290	9.50%
San Francisco Giants	30,650	29,337	4.28%
Minnesota Twins	22,514	22,827	-1.39%
Philadelphia Phillies	28,459	29,193	-2.58%
Washington Nationals	25,017	27,839	-11.28%
Detroit Tigers	19,634	23,612	-20.26%
Kansas City Royals	15,974	22,474	-40.69%
Arizona Diamond Backs	19,817	29,182	-47.26%
Cleveland Guardians	17,050	25,315	-48.48%

Home vs Road % change in Fan Attendance



A deeper dive into the Tampa Bay's player's age compared to salary indicates that 28 years old players (birth year equal to 1994) brought in the highest total salary of \$11Millions.

Tampa Bay Rays Player's Annual Salary by Date of Birth



## Conclusions

Overall, this research focused more on statistical information about MLB teams in aggregate and found that there is a positive correlation between salary and team performance.

For future research, we would like to look at individual players' profiles from different teams to help identify specific players' characteristics who are able to pull in the highest salaries. We will also evaluate the positions that attract the highest salaries and will explore players' tenure as we was puzzled with some players' age who we believe were outliers.

In retrospective, our research topic was inspired by our admiration for the American game of baseball, and our team's wish to learn more about the sport so that we can have more meaningful conversations with our fellow American friends who grew up with baseball while most of us on our team grew up around either Cricket or Soccer (football as we all refer to it outside of the United States). Our team's goal was achieved, and we are thrilled that we have developed our Oracle SQL and visualization development skills in the process.

## References

Average Annual Temperatures for Large US Cities - NOAA National Centers for Environmental Information (NCEI - [Climate Normals.](#)). [Average Annual Temperatures for Large US Cities - Current Results](#), accessed on April, 16, 2023

Baseball Is Back! All 30 Teams Play on Opening Day, March 30!  
<https://www.census.gov/content/dam/Census/library/visualizations/2023/comm/baseball-is-back.pdf>, accessed on April, 27, 2023.

Library of Congress Research Guides: Baseball. <https://guides.loc.gov/sports-industry/baseball>, accessed April 27, 2023.

MLB Team Payroll Tracker [MLB 2022 Payroll Tracker | Spotrac](#)

MLB Attendance Report – 2022 (Average tickets sold per home game) - [2022 MLB Attendance - Major League Baseball – ESPN](#) , accessed April 15, 2023.

Major League Roster [https://www.espn.com/mlb/stats/rosters/\\_/sort/average\\_age/order/true](https://www.espn.com/mlb/stats/rosters/_/sort/average_age/order/true) , accessed April 26, 2023.

Tampa Bay Roster - <https://usatoday.sportsdirectinc.com/baseball/mlb-teams.aspx?page=/data/mlb/teams/rosters/roster2960.html>, accessed April 17, 2023.