MLB Data and Analysis

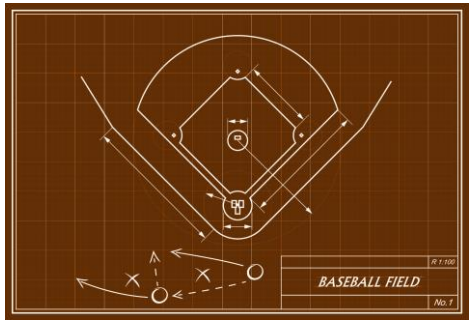
By - Manaswi Mishra ,Aalok Patel, Meghna Diwan, Abhishek Yadav

Agenda

- Analytics in Baseball
- Research Objective
- Methodology
- Key Findings
- Recommendations
- Tableau Dashboard

Baseball and Analytics

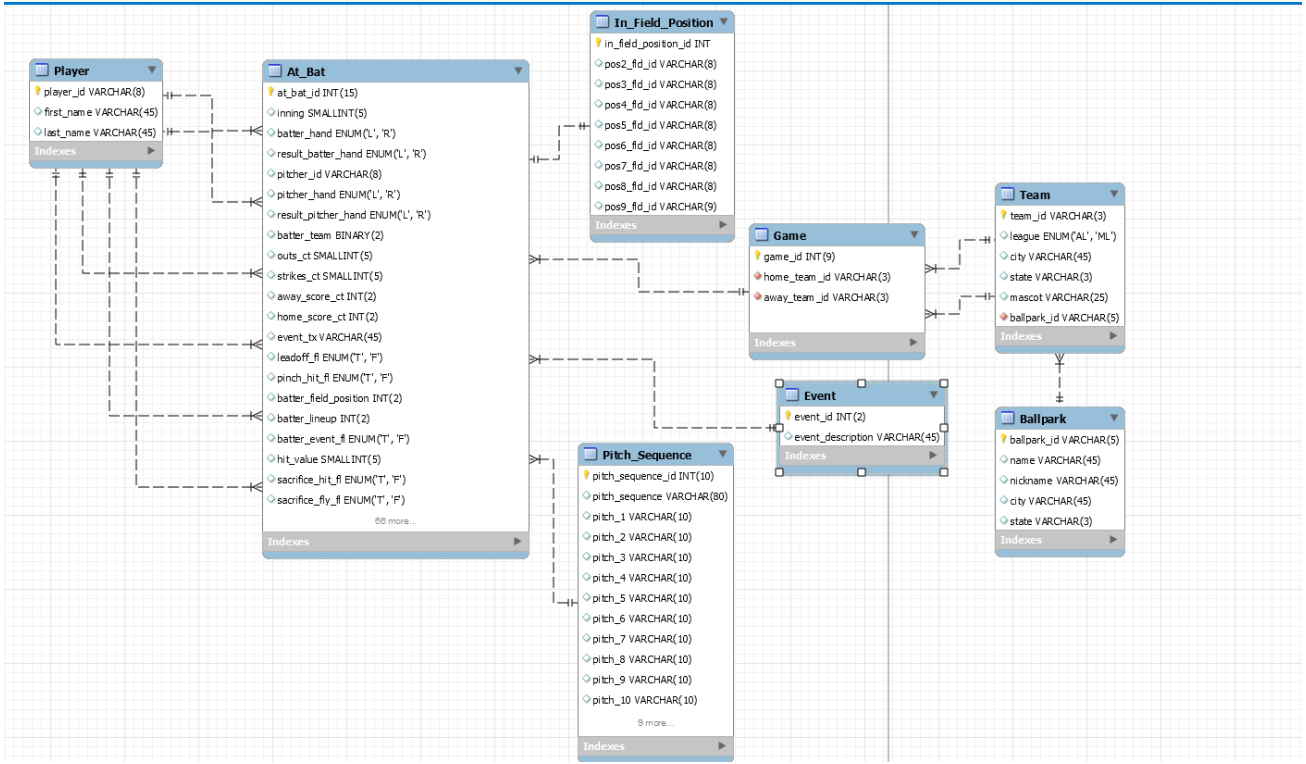
- Major League Baseball is one of the pinnacle sports which uses Data Analytics for calculated gambles during an ongoing game as well as player transactions
- First used extensively in the early 1990s by the Oakland Athletics, analytics is commonly used throughout the league.



Research Objective

- **Data to be used**
  - Raw at-bat data- This represents the data collected during the game in its rawest form
- **Data utilization**
  - Leverage this raw data to recreate advanced metrics, we can minimize the size of the datasets needed during these analyses.
- **Findings + Insights**
  - Identifying competitive metrics and trends within games, between players, and between teams
- **Easy to read dashboards** for non-technical end-users

Methodology



Methodology - ETL

Extraction - .csv

Transformation

Loading

- Received from popular compiler of at-bat data
- Limited to 2010 - 2018 and to all games played by the St. Louis Cardinals
- 114,176 total rows and 132 columns + 5 Reference Tables - pitcher and batter demographics
  - Pitch sequence
  - Outfield composition
  - Contextual knowledge within at-bat, within inning, within game
  - Team information
  - Outcome of at-bat

Methodology - ETL

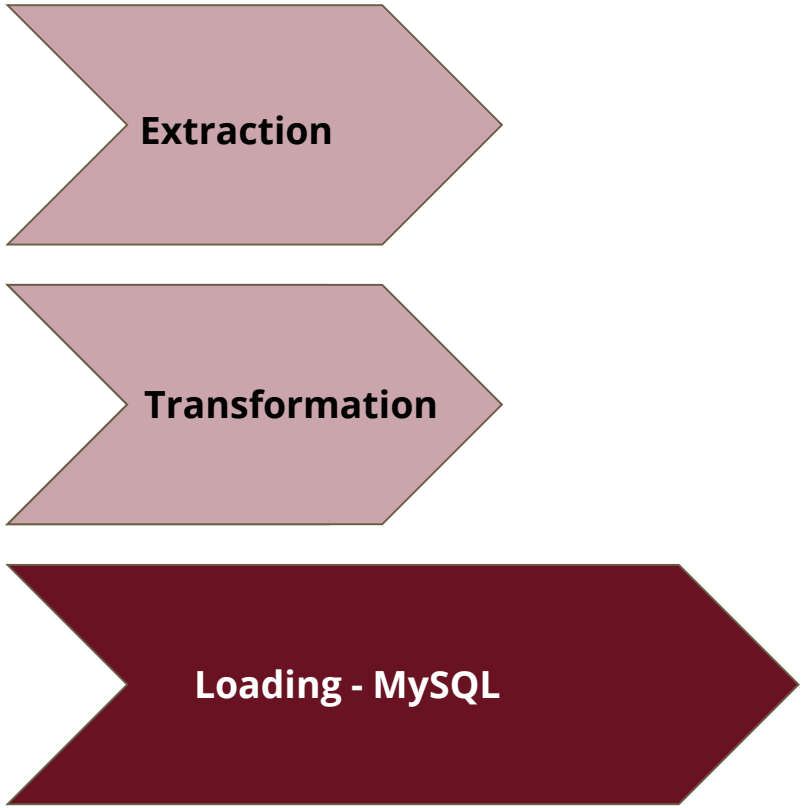
Extraction

Transformation - GCP

Loading

- **Standardized the team id** for franchise that changed their name during this period of time (FLO to MIA)
- **Dropped irrelevant columns** that were unnecessary for the analysis
- **Created alternative pitch sequence string** that ignored symbols indicating events unnecessary for this analysis
- **Split pitch sequence** (stored as string) into separate columns for each individual pitch
- **Created date field**
- **Converted** some variables to categorical flags for easy filtering
- **Standardized T/F** to True/False across all columns

Methodology - ETL



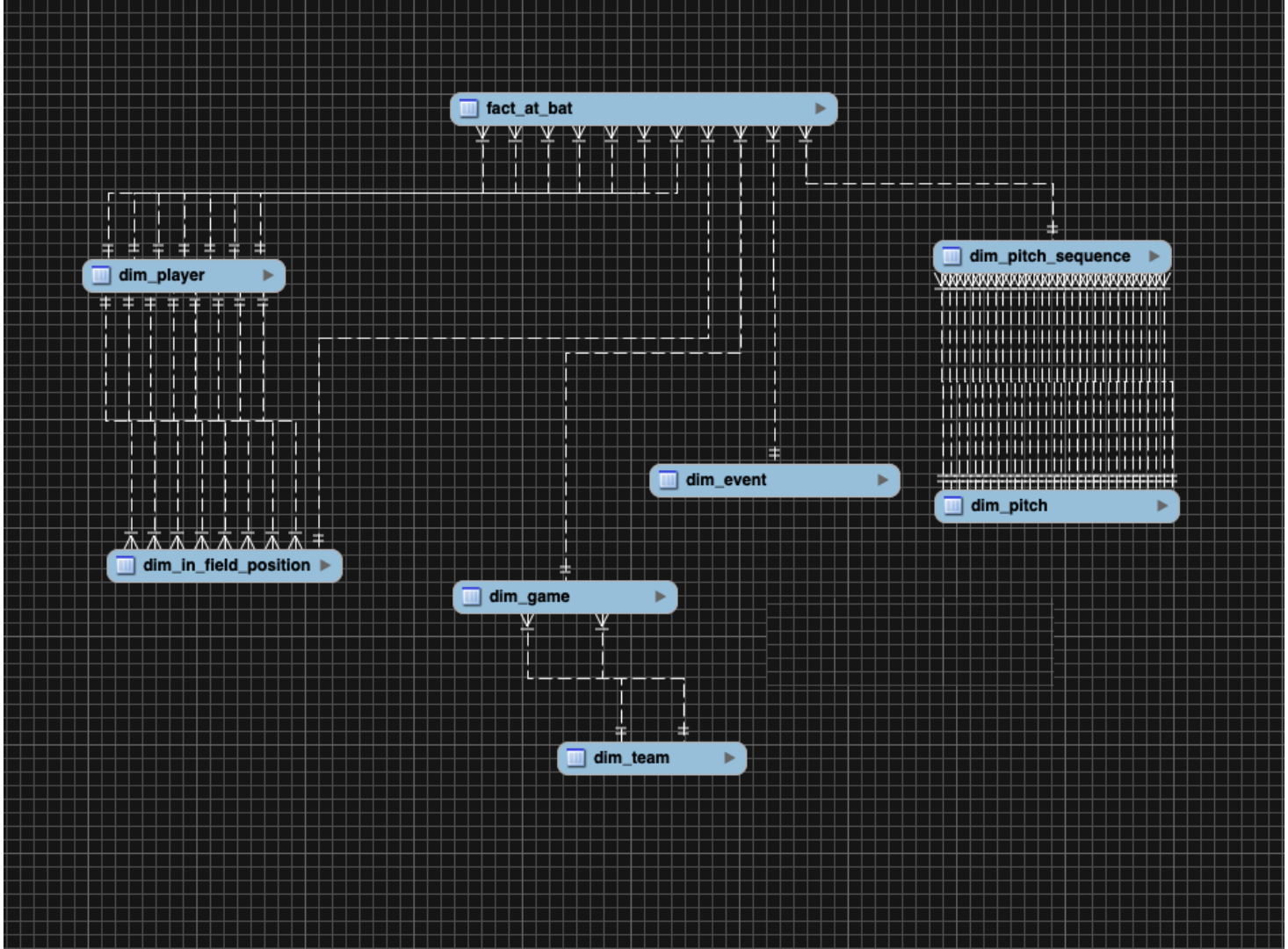
Snowflake Dimensional Database

1. Created a dummy database for the raw data
2. Identified Primary and Foreign Keys
3. Built 9 dim tables using DDL and DML

Why?

Normalized data for the following reasons :

1. Avoid Partial Functional Dependencies (Delete Anomaly, Update Anomaly and Insert Anomaly)
2. Avoid Transitive Functional Dependencies


fact_at_bat
dim_player
dim_pitch_sequence
dim_event
dim_pitch
dim_in_field_position
dim_game
dim_team

Key Findings

Pythagorean Winning %

What?

Key metric to determine if the team under or over performed in the season based on the run scored and allowed at home or away in a season.

Result?

For the years 2016, 2017 and 2018 the St. Louis Cardinals were on par on what their expected season wins/losses. They did not make it to the playoffs in that time period, failed to make any big trades/signings and performed as per prediction.

Year	RunsScored	RunsAllowed	Win	Loss	PythogoreanWinPercentage	PredictedWin	PredictedLoss
2016	771	700	81	68	0.543	80.907	68.093
2017	753	695	80	72	0.536	81.472	70.528
2018	742	680	78	68	0.539	78.694	67.306

First-Pitch Strike %

What?

Calculated first pitch strike percentage and the resulting Strikeout to Walk ratio when the first pitch is a strike.

Result?

The hitting percentage is less for pitchers who throw more first pitch strikes. In 2018, Cardinals pitcher Miles Mikolas lead the MLB in First Pitch Strikes which resulted in:

1. Leading all MLB pitchers in First Pitch Strike Percentage - 66.25%.
2. Highest Strikeout - Walk ratio in all MLB - 5.4
3. Being the only pitcher to pitch a shutout game.

Looking at his efficiency, the club awarded him a **4 year $68 million** extension.

pitcher_id	TotalPitches	FirstPitchStrikes	FirstPitchStrikePercentage	HitPercentageAfterFirstPitchStrike	Strikeout	Walk	StrikeToWalkRatio
mikom001	809	536	0.6625	0.1656	146	25	5.8400
flahj002	628	347	0.5525	0.1099	182	56	3.2500
weavl001	614	341	0.5554	0.1547	121	52	2.3269
martc006	527	299	0.5674	0.1328	117	56	2.0893
gantj002	494	288	0.5830	0.1194	95	54	1.7593

What really matters to win a game?

St. Louis Cardinals

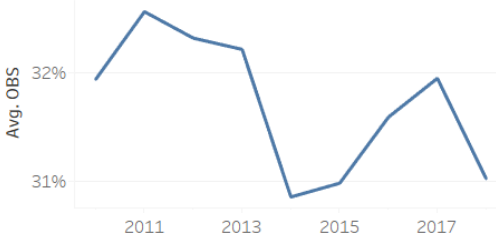
Comparison of key Base Ball metrics

Batting Average

Batter T..	2010	2011	2012	2013	2014	2015	2016	2017	2018
SLN	0.250	0.259	0.257	0.259	0.241	0.240	0.246	0.242	0.237

Year of Date

On Base Percentage



Home Runs

Year of D..	SLN
2010	149
2011	152
2012	156
2013	120
2014	101
2015	128
2016	214
2017	192
2018	199

Batter Team ID

Batter Team ID: SLN

Year of Date: (All)

Runs Conceded

Year of D..	SLN
2010	830
2011	954
2012	855
2013	863
2014	735
2015	722
2016	987
2017	863
2018	930

Runs Scored

Year of D..	SLN
2010	530
2011	480
2012	564
2013	502
2014	449
2015	410
2016	483
2017	520
2018	460

Run Batted In

SLN
6,035

Total Hits

Year of D..	SLN
2010	2,169
2011	2,233
2012	2,268
2013	2,151
2014	1,926
2015	2,076
2016	2,346
2017	2,228
2018	2,167

Strikeouts

Year of D..	SLN
2010	1,001
2011	930
2012	1,149
2013	1,071
2014	1,104
2015	1,230
2016	1,251
2017	1,297
2018	1,337

What really matter to win a game?

What?

Hit for Average, Get on Base and Strike Out Less. Home Runs Don’t MATTER as much.

Result?

- 2011 and 2013 Cardinals went to the **World Series**. They had high batting average, high on base percentage, and less strikeouts.
- On the contrary, during 2016, 2017 and 2018 seasons the cardinals hit more homeruns but their batting average and on base percentage fell and strikeouts increased. This resulted, in the team **not** make it to the playoffs all 3 years.
- Baseball is moving towards hitting for power, exit velocity and launch angle but the trick is still to get on base.

Recommendations

Suggested Use Cases

Player Development

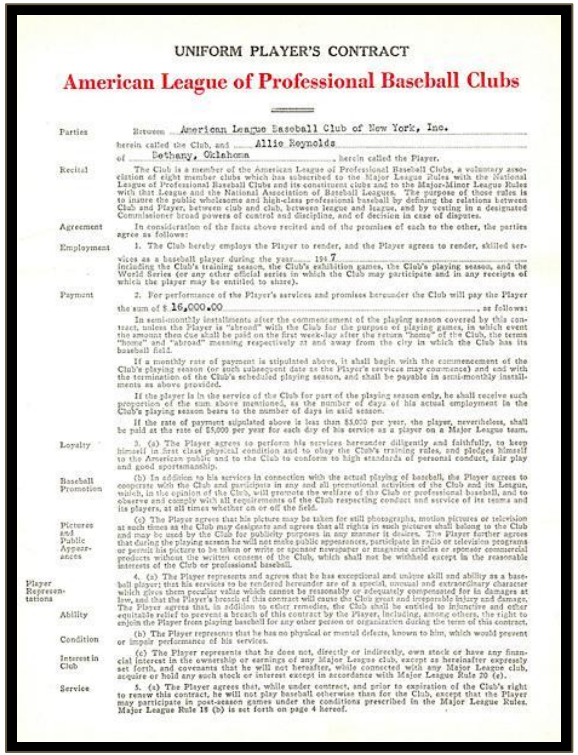
- Track player performance during season to identify developing players or replacing them

Player Contracts

- Mid-Season contract negotiations favor the team
- Identifying a target for retention earlier is favorable for a franchise as they can potentially secure the player for less

In-game management

- Once informed about performance in specific game situations, the manager can exploit the matchup between the pitcher and batter



UNIFORM PLAYER'S CONTRACT

American League of Professional Baseball Clubs

Parties — Between American League Baseball Club of New York, Inc. herein called the Club, and Allie Reynolds of Bethany, Oklahoma, herein called the Player.

Recital — The Club is a member of the American League of Professional Baseball Clubs, a voluntary association of eight member clubs which has subscribed to the Major League Rules with the National League of Professional Baseball Clubs and its constituent clubs and to the Major-Minor League Rules with that League and the National Association of Baseball Leagues. The purpose of those rules is to insure the public wholesome and high-class professional baseball by defining the relations between Club and Player, between club and club, between league and league, and by vesting in a designated Commissioner broad powers of control and discipline, and of decision in case of disputes.

Agreement — In consideration of the facts above recited and of the promises of each to the other, the parties agree as follows:

Employment — 1. The Club hereby employs the Player to render, and the Player agrees to render, skilled services as a baseball player during the year 194 7, including the Club's training season, the Club's exhibition games, the Club's playing season, and the World Series (or any other official series in which the Club may participate and in any receipts of which the player may be entitled to share).

Payment — 2. For performance of the Player's services and promises hereunder the Club will pay the Player the sum of $ 16,000.00, as follows:

In semi-monthly installments after the commencement of the playing season covered by this contract, unless the Player is "abroad" with the Club for the purpose of playing games, in which event the amount then due shall be paid on the first week-day after the return "home" of the Club, the terms "home" and "abroad" meaning respectively at and away from the city in which the Club has its baseball field.

If a monthly rate of payment is stipulated above, it shall begin with the commencement of the Club's playing season (or such subsequent date as the Player's services may commence) and end with the termination of the Club's scheduled playing season, and shall be payable in semi-monthly installments as above provided.

If the player is in the service of the Club for part of the playing season only, he shall receive such proportion of the sum above mentioned, as the number of days of his actual employment in the Club's playing season bears to the number of days in said season.

If the rate of payment stipulated above is less than $5,000 per year, the player, nevertheless, shall be paid at the rate of $5,000 per year for each day of his service as a player on a Major League team.

Loyalty — 3. (a) The Player agrees to perform his services hereunder diligently and faithfully, to keep himself in first class physical condition and to obey the Club's training rules, and pledges himself to the American public and to the Club to conform to high standards of personal conduct, fair play and good sportsmanship.

Baseball Promotion — (b) In addition to his service in connection with the actual playing of baseball, the Player agrees to cooperate with the Club and participate in any and all promotional activities of the Club and its League, which, in the opinion of the Club, will promote the welfare of the Club or professional baseball, and to observe and comply with all requirements of the Club respecting conduct and service of its teams and its players, at all times whether on or off the field.

Pictures and Public Appearances — (c) The Player agrees that his picture may be taken for still photographs, motion pictures or television at such times as the Club may designate and agrees that all rights in such pictures shall belong to the Club and may be used by the Club for publicity purposes in any manner it desires. The Player further agrees that during the playing season he will not make public appearances, participate in radio or television programs or permit his picture to be taken or write or sponsor newspaper or magazine articles or sponsor commercial products without the written consent of the Club, which shall not be withheld except in the reasonable interests of the Club or professional baseball.

Player Representations — 4. (a) The Player represents and agrees that he has exceptional and unique skill and ability as a baseball player; that his services to be rendered hereunder are of a special, unusual and extraordinary character which gives them peculiar value which cannot be reasonably or adequately compensated for in damages at law, and that the Player's breach of this contract will cause the Club great and irreparable injury and damage. The Player agrees that, in addition to other remedies, the Club shall be entitled to injunctive and other equitable relief to prevent a breach of this contract by the Player, including, among others, the right to enjoin the Player from playing baseball for any other person or organization during the term of this contract.

Ability

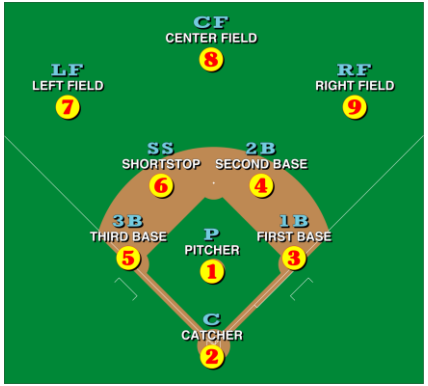
Condition — (b) The Player represents that he has no physical or mental defects, known to him, which would prevent or impede performance of his services.

Interest in Club — (c) The Player represents that he does not, directly or indirectly, own stock or have any financial interest in the ownership or earnings of any Major League club, except as hereinafter expressly set forth, and covenants that he will not hereafter, while connected with any Major League club, acquire or hold any such stock or interest except in accordance with Major League Rule 20 (e).

Service — 5. (a) The Player agrees that, while under contract, and prior to exercise of the Club's right to renew this contract, he will not play baseball otherwise than for the Club, except that the Player may participate in post-season games under the conditions prescribed in the Major League Rules. Major League Rule 18 (b) is set forth on page 4 hereof.

Corrective Measures

- Structure **pitch sequence** such that the full sequence can be utilized
- Use **defensive field positions** to help strategize fielding especially against specific teams
- Use **defensive errors** to identify common mistakes and players prone to errors



Scope for Improvement

- Scale up to include data regarding other teams
  - Can gather insights on opponents - both players and teams - for in-game, between game, and between season changes

- Expand to include in-game tracking like:
  - Speed/Acceleration
  - Launch Angle
  - Exit Velocity
  - Hard Hit Rate
  - Spin rate

Lessons Learned

- Potential of data analytics in baseball
- Real Life data is unstructured and dirty
  - Used SQL & DataPrep to clean and transform the data
- Team collaboration using Github, GCP, Google Drive, Tableau Online
- Leadership Skills
  - Relying upon team members strengths and domain expertise

Appendix

Data Sources

- https://www.retrosheet.org/tools.htm
  - Player IDs
  - Franchise IDs
- https://www.kaggle.com/gghatano/mlbplaybyplay2010s
  - Yearly at-bat dataset summarized in CSV files
  - Note: Ultimately, the MLB tracks and maintains the data generated each game.

Data/Tools Used

Player IDs
Franchise IDs
Event IDs
Yearly at-bat dataset

Excel
MySQL
Cloud SQL
Google Cloud Platform

Tableau

Pythagorean Win %

```sql
SELECT
    Wins.Year,
    Wins.RunsScored,
    Wins.RunsAllowed,
    Wins.Win,
    Wins.Loss,
    Wins.PythogoreanWinPercentage,
    (Wins.Win + Wins.Loss) * Wins.PythogoreanWinPercentage AS 'PredictedWin',
    (Wins.Win + Wins.Loss) * (1 - Wins.PythogoreanWinPercentage) AS 'PredictedLoss'
FROM
    (SELECT
        LEFT(runs.game_id, 4) AS 'Year',
            SUM(runs.RunScoredHome + runs.RunScoredAway) AS 'RunsScored',
            SUM(runs.RunAllowedHome + runs.RunAllowedAway) AS 'RunsAllowed',
            SUM(CASE
                WHEN runs.RunScoredHome > runs.RunAllowedHome THEN 1
                WHEN runs.RunScoredAway > runs.RunAllowedAway THEN 1
                ELSE 0
            END) AS 'Win',
            SUM(CASE
                WHEN runs.RunScoredHome < runs.RunAllowedHome THEN 1
                WHEN runs.RunScoredAway < runs.RunAllowedAway THEN 1
                ELSE 0
            END) AS 'Loss',
            ROUND(POWER(SUM(runs.RunScoredHome + runs.RunScoredAway), 1.8) / (POWER(SUM(runs.RunScoredHome + runs.RunScoredAway), 1.8) + POWER(SUM(runs.RunAllowedHome + runs.RunAllowedAway), 1.8)), 3) AS 'PythogoreanWinPercentage'
    FROM
        (SELECT
        game_id,
            CASE
                WHEN home_team_id = 'SLN' THEN home_score_ct
                ELSE 0
            END AS 'RunScoredHome',
            CASE
                WHEN home_team_id = 'SLN' THEN away_score_ct
                ELSE 0
            END AS 'RunAllowedHome',
            CASE
                WHEN home_team_id != 'SLN' THEN away_score_ct
                ELSE 0
            END AS 'RunScoredAway',
            CASE
                WHEN home_team_id != 'SLN' THEN home_score_ct
                ELSE 0
            END AS 'RunAllowedAway'
    FROM
        MLB.mlbdatatransformations
    WHERE
        game_end_fl = 'true'
    GROUP BY game_id) runs
    GROUP BY LEFT(runs.game_id, 4)) Wins
```

Hitting with RISP

```
SELECT
    d.batter_id,
    COUNT(*) AS 'AtBatWithRunnersInScoringPosition',
    SUM(CASE
        WHEN event_id IN ('20' , '21', '22', '23') THEN 1
        ELSE 0
    END) AS 'HitWithRunnersInScoringPosition',
    SUM(CASE
        WHEN event_id IN ('20' , '21', '22', '23') THEN 1
        ELSE 0
    END) / COUNT(*) AS 'BattingAverageWithRunnersInScoringPosition'
FROM
    (SELECT
        *
    FROM
        MLB.mlbdatatransformations
    WHERE
        LEFT(date, 4) = '2018'
            AND (base2_run_id != '' OR base3_run_id != '')) d
GROUP BY d.batter_id
ORDER BY AtBatWithRunnersInScoringPosition DESC
```

First Pitch Strikes to Strikeout-Walk Ratio

```sql
SELECT
    pitcher_id,
    COUNT(*) AS 'TotalPitches',
    SUM(CASE
        WHEN d.pitch_1 IN ('C' , 'F', 'L', 'M', 'O', 'R', 'S', 'T') THEN 1
        WHEN
            (d.pitch_1 = 'X'
                AND d.event_id IN ('2' , '18', '19'))
        THEN
            1
        ELSE 0
    END) AS 'FirstPitchStrikes',
    (SUM(CASE
        WHEN d.pitch_1 IN ('C' , 'F', 'L', 'M', 'O', 'R', 'S', 'T') THEN 1
        WHEN
            (d.pitch_1 = 'X'
                AND d.event_id IN ('2' , '18', '19'))
        THEN
            1
        ELSE 0
    END)) / COUNT(*) AS 'FirstPitchStrikePercentage',
    (SUM(CASE
        WHEN
            d.pitch_1 IN ('C' , 'F', 'L', 'M', 'O', 'R', 'S', 'T', 'X')
                AND d.event_id IN ('20' , '21', '22', '23')
        THEN
            1
        ELSE 0
    END)) / COUNT(*) AS 'HitPercentageAfterFirstPitchStrike',
    SUM(CASE
        WHEN EVENT_ID = '3' THEN 1
        ELSE 0
    END) AS `Strikeout`,
    SUM(CASE
        WHEN EVENT_ID = '14' THEN 1
        ELSE 0
    END) AS `Walk`,
    SUM(CASE
        WHEN EVENT_ID = '3' THEN 1
        ELSE 0
    END) / SUM(CASE
        WHEN EVENT_ID = '14' THEN 1
        ELSE 0
    END) AS `StrikeToWalkRatio`
FROM
    (SELECT
        *
    FROM
        MLB.mlbdatatransformations
    WHERE
        pitch_1 NOT IN ('N' , 'V', '')
            AND LEFT(game_id, 4) = '2018') d
GROUP BY pitcher_id
ORDER BY TotalPitches DESC
LIMIT 15
```

Batter Record

```
SELECT
    d.batter_id,
    d.Single AS `1B`,
    d.Double AS `2B`,
    d.Triple AS `3B`,
    d.HomeRun AS `HomeRun`,
    d.RBI AS `RBI`,
    (d.Single + d.Double + d.Triple + d.HomeRun) / (d.AtBat - d.SacFly - d.SacHit) AS `BA`,
    (d.Single + d.Double + d.Triple + d.HomeRun + d.Walk + d.IntentionalWalk + d.HitByPitch) / (d.AtBat - d.SacFly - d.SacHit + d.Walk + d.IntentionalWalk + d.HitByPitch) AS `OBS`,
    (d.TotalBases) / (d.AtBat - d.SacFly - d.SacHit) AS `Slugging`,
    (d.Single + d.Double + d.Triple + d.HomeRun + d.Walk + d.IntentionalWalk + d.HitByPitch) / (d.AtBat - d.SacFly - d.SacHit + d.Walk + d.IntentionalWalk + d.HitByPitch) + (1 * d.Single + 2 * d.Double + 3 * d.Triple + 4 * d.HomeRun) / (d.AtBat - d.SacFly - d.SacHit) AS `OPS`,
    d.StrikeOut AS `StrikeOut`,
    d.TotalBases AS `TotalBases`,
    d.GroundedDoublePlay AS `GroundedDoublePlay`,
    d.HitByPitch AS `HitByPitch`,
    d.StrikeOut AS `StrikeOut`,
    d.SacFly AS `SacFly`,
    d.SacHit AS `SacHit`,
    d.FirstPitchSwingMissPercentage AS `FirstPitchSwingMissPercentage`,
    d.FirstPitchCalledStrikePercentage AS `FirstPitchCalledStrikePercentage`,
    d.FirstPitchFoulBallPercentage AS `FirstPitchFoulBallPercentage`,
    d.FirstPitchBallPercentage AS `FirstPitchBallPercentage`
FROM
    (SELECT
        BATTER_ID,
            SUM(CASE
                WHEN EVENT_ID NOT IN ('14' , '15', '16', '17') THEN 1
                ELSE 0
            END) AS `AtBat`,
            SUM(CASE
                WHEN EVENT_ID IN ('20' , '21', '22', '23', '2', '5', '15', '16', '14') THEN 1
                ELSE 0
            END) AS `PlateAppearance`,
            SUM(CASE
                WHEN EVENT_ID = '14' THEN 1
                ELSE 0
            END) AS `Walk`,
            SUM(CASE
                WHEN EVENT_ID = '15' THEN 1
                ELSE 0
            END) AS `IntentionalWalk`,
            SUM(CASE
                WHEN EVENT_ID = '16' THEN 1
                ELSE 0
            END) AS `HitByPitch`,
            SUM(CASE
                WHEN EVENT_ID = '2' THEN 1
                ELSE 0
            END) AS `StrikeOut`,
            SUM(CASE
                WHEN sacrifice_fly_fl = 'T' THEN 1
                ELSE 0
            END) AS `SacFly`,
            SUM(CASE
                WHEN sacrifice_hit_fl = 'T' THEN 1
                ELSE 0
            END) AS `SacHit`,
            SUM(CASE
                WHEN double_play_fl = 'true' THEN 1
                ELSE 0
            END) AS `GroundedDoublePlay`,
            SUM(CASE
                WHEN triple_play_fl = 'true' THEN 1
                ELSE 0
            END) AS `GroundedTriplePlay`,
            SUM(CASE
```

Batter Record Continued

```
        SUM(CASE
            WHEN EVENT_ID = '3' THEN 1
            ELSE 0
        END) AS `StrikeOut`,
        SUM(CASE
            WHEN sacrifice_fly_fl = 'T' THEN 1
            ELSE 0
        END) AS `SacFly`,
        SUM(CASE
            WHEN sacrifice_hit_fl = 'T' THEN 1
            ELSE 0
        END) AS `SacHit`,
        SUM(CASE
            WHEN double_play_fl = 'true' THEN 1
            ELSE 0
        END) AS `GroundedDoublePlay`,
        SUM(CASE
            WHEN triple_play_fl = 'true' THEN 1
            ELSE 0
        END) AS `GroundedTriplePlay`,
        SUM(CASE
            WHEN pitch_1 = 'S' THEN 1
            ELSE 0
        END) / (COUNT(pitch_1)) AS `FirstPitchSwingMissPercentage`,
        SUM(CASE
            WHEN pitch_1 = 'C' THEN 1
            ELSE 0
        END) / (COUNT(pitch_1)) AS `FirstPitchCalledStrikePercentage`,
        SUM(CASE
            WHEN pitch_1 = 'F' THEN 1
            ELSE 0
        END) / (COUNT(pitch_1)) AS `FirstPitchFoulBallPercentage`,
        SUM(CASE
            WHEN pitch_1 = 'B' THEN 1
            ELSE 0
        END) / (COUNT(pitch_1)) AS `FirstPitchBallPercentage`,
        SUM(CASE
            WHEN EVENT_ID = '20' THEN 1
            ELSE 0
        END) AS `Single`,
        SUM(CASE
            WHEN EVENT_ID = '21' THEN 1
            ELSE 0
        END) AS `Double`,
        SUM(CASE
            WHEN EVENT_ID = '22' THEN 1
            ELSE 0
        END) AS `Triple`,
        SUM(CASE
            WHEN EVENT_ID = '23' THEN 1
            ELSE 0
        END) AS `HomeRun`,
        SUM(rbi_ct) AS `RBI`,
        SUM(CASE
            WHEN EVENT_ID = '20' THEN 1
            WHEN EVENT_ID = '21' THEN 2
            WHEN EVENT_ID = '22' THEN 3
            WHEN EVENT_ID = '23' THEN 4
            ELSE 0
        END) AS `TotalBases`
FROM
    MLB.mlbdatatransformations
WHERE
    LEFT(GAME_ID, 4) = '2018'
GROUP BY BATTER_ID
ORDER BY HomeRun DESC) AS d
```