Reviewing 2023 MLB Data

A Data Analytics Presentation by Matthew Byrne

About this Project

Purpose of this Project

Baseball has become more and more analytical over the years, so why not use some recently acquired data analytics tools to analyze 2023 MLB data. No matter whether you're a die-hard baseball fan or a complete beginner, the insights and takeaways from this project will help you better understand player and team performance and potentially help you find your new favorite team or player.

Link to DataSet

Link to GitHub Repository

Link to Tableau Dashboard

Tools Used

Excel, PostgreSQL, pandas, seaborn, Matplotlib, tableau

Project Workflow

I'll be working through every stage of the Data Analytics process to conduct my analysis and gather my takeaways for this project.

Data Gathering & Storage

I visited Kaggle to find a dataset that includes 2023 MLB performance data.

I created an ERD using QuickDB to design my database schema and then uploaded my data into PostgreSQL

Data Cleaning

I imported my data into a Jupyter Notebook and used Pandas & ETL to clean the data.

Cleaning included checking for nulls, renaming columns, etc.

Data Analysis

I imported my cleaned data into PostgreSQL and began performing Exploratory Data Analysis techniques to explore my dataset.

Data Visualization

Using Jupyter Notebook I used Matplotlib, Seaborn, and Plotly Express to create my charts.

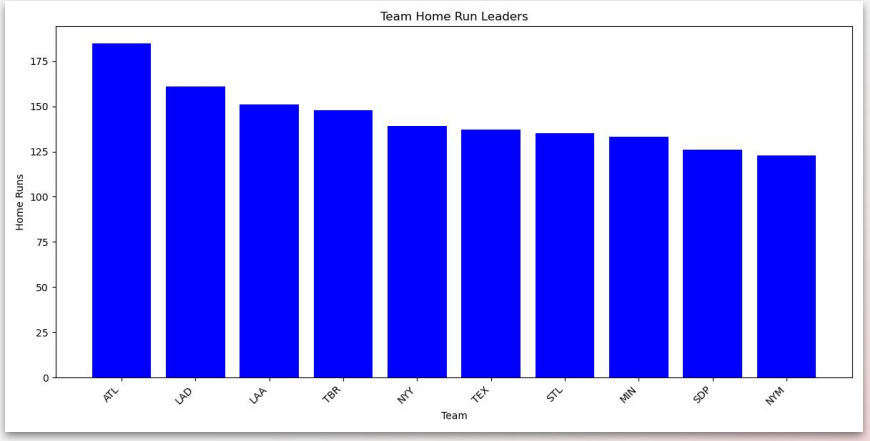
Separately, I created a Tableau dashboard to visualize the data.

Hitting Analysis

Home Run Leaders

Teams & Players who Hit the Most Home Runs

A home run is one of the most exciting plays in baseball. These charts show you who you should keep an eye on to hit the most home runs in future seasons.

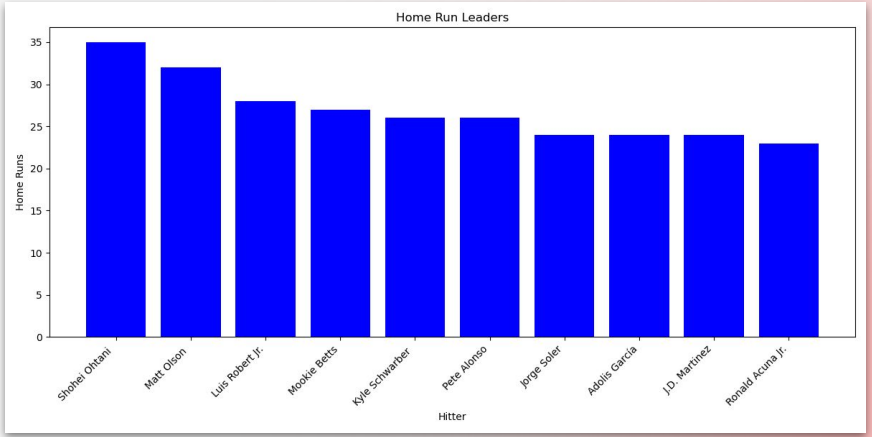



```sql
-- Which players had the most total home runs?

SELECT DISTINCT "Name", "HR" as "Most Home Runs"
FROM "HITTING"
ORDER BY "HR" DESC
LIMIT 10;
```

Data Output | Messages | Notifications

	Name character varying	Most Home Runs integer
1	Shohei Ohtani	35
2	Matt Olson	32
3	Luis Robert Jr.	28
4	Mookie Betts	27
5	Kyle Schwarber	26
6	Pete Alonso	26
7	Jorge Soler	24
8	Adolis García	24
9	J.D. Martinez	24
10	Rafael Devers	23

```python
# Create bar chart using Matplotlib

players = HRLeaders['Name']
HRs = HRLeaders['HR']

# Create a bar chart
plt.figure(figsize=(12, 6))
plt.bar(players, HRs, color='blue')

# Add labels and title
plt.xlabel('Hitter')
plt.ylabel('Home Runs')
plt.title('Home Run Leaders')

# Rotate x-axis labels for better readability
plt.xticks(rotation=45, ha='right')

# Show plot
plt.tight_layout()
plt.savefig('HRLeaders-Players')
plt.show()
```



RBI Leaders

Who Lead the League in RBIs?

American League

```python
# Which players had the most RBIs in each league?
# Starting with the American League (AL)

AL_hitters = hitting_df[hitting_df['League'] == 'AL']

sorted_by_RBI = AL_hitters.sort_values(by='RBI', ascending=False)

# Step 3: Select the top 10 rows and keep only the desired columns
top_10_RBI = sorted_by_RBI.head(10)[['Name', 'RBI']]

# Display the result

top_10_RBI
```

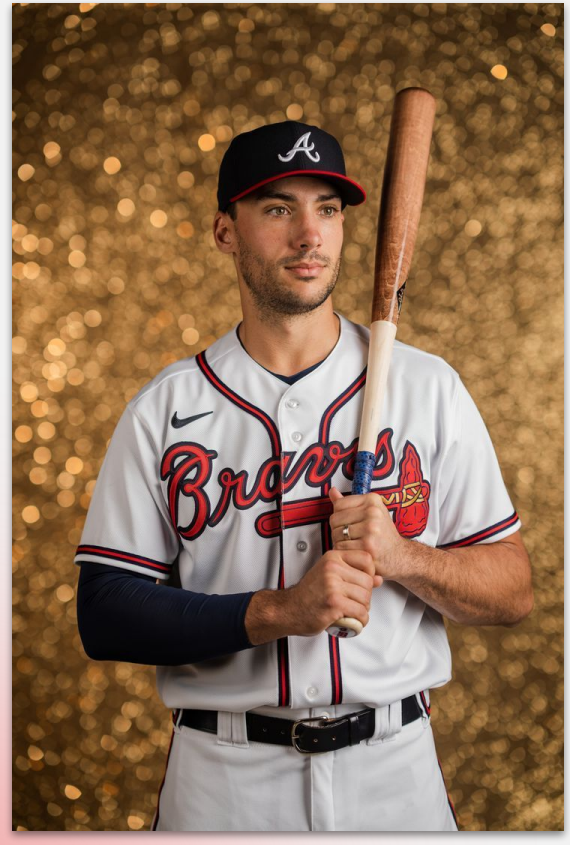
```sql
-- Which players had the most RBIs in each league?

SELECT
    (SELECT "Name" FROM "HITTING" WHERE "League" = 'AL' ORDER BY "RBI" DESC LIMIT 1) AS AL_RBI_Leader,
    (SELECT "RBI" FROM "HITTING" WHERE "League" = 'AL' ORDER BY "RBI" DESC LIMIT 1) AS Amount,
    (SELECT "Name" FROM "HITTING" WHERE "League" = 'NL' ORDER BY "RBI" DESC LIMIT 1) AS NL_RBI_Leader,
    (SELECT "RBI" FROM "HITTING" WHERE "League" = 'NL' ORDER BY "RBI" DESC LIMIT 1) AS Amount;
```

Data Output | Messages | Notifications

	al_rbi_leader character varying	amount integer	nl_rbi_leader character varying	amount integer
1	Adolis García	80	Matt Olson	80

	Name	RBI
215	Adolis García	80
447	Shohei Ohtani	76
433	Josh Naylor	76
157	Rafael Devers	73
634	Kyle Tucker	68
267	Jonah Heim	67
636	Justin Turner	64
77	Alex Bregman	63
33	Randy Arozarena	62
244	Vladimir Guerrero Jr.	62

RBI Leaders

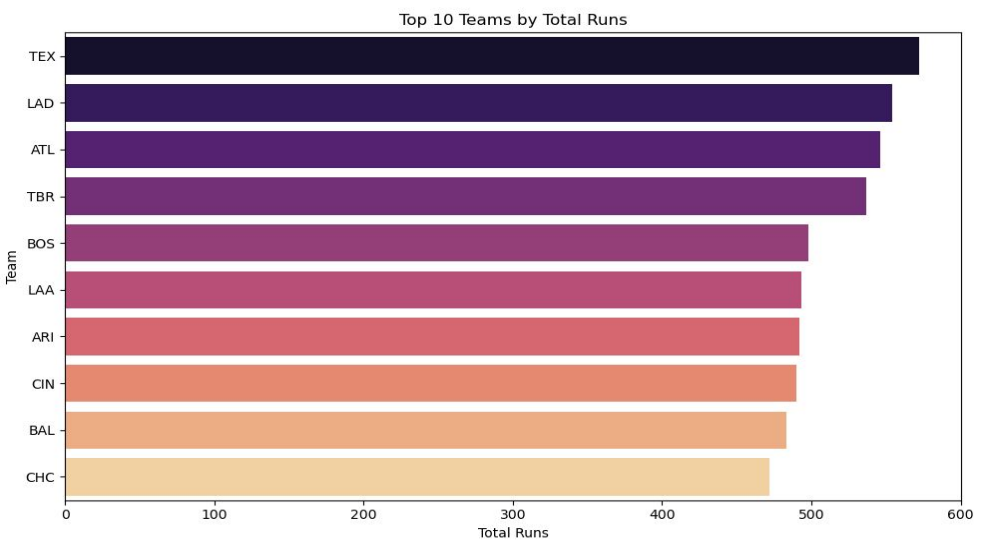


Who Lead the League in RBIs?

National League

```
# Now let's do the National League

NL_hitters = hitting_df[hitting_df['League'] == 'NL']

sorted_by_RBI = NL_hitters.sort_values(by='RBI', ascending=False)
top_10_RBI = sorted_by_RBI.head(10)[['Name', 'RBI']]

# Display the result
top_10_RBI
```

	Name	RBI
450	Matt Olson	80
31	Nolan Arenado	76
369	J.D. Martinez	73
207	Freddie Freeman	70
67	Mookie Betts	67
11	Ozzie Albies	66
663	Christian Walker	63
18	Pete Alonso	63
235	Nolan Gorman	62
342	Francisco Lindor	61



Top Scoring Offenses

Which Teams Scored the Most Runs?

High profile offenses are a nightmare for pitchers to face. Use this data to identify which teams might be creating the most stress for opposing pitching staffs or pick your team to watch this year if offense is your thing.

```sql
-- Which teams scored the most runs?

SELECT "Team", SUM("R") as "Teams with the Most Runs Scored"
FROM "HITTING"
GROUP BY "Team"
ORDER BY SUM("R") DESC
LIMIT 10;
```

Data Output | Messages | Notifications

	Team (character varying)	Teams with the Most Runs Scored (bigint)
1	TEX	1144
2	LAD	1108
3	ATL	1092
4	TBR	1074
5	BOS	996
6	LAA	986
7	ARI	984
8	CIN	980
9	BAL	966
10	CHC	944

```python
# Create a horizontal bar chart using Seaborn

plt.figure(figsize=(10, 6))  # Adjust the figure size as needed
sns.barplot(x='R', y='Team', data=top_10, palette='magma')

# Add labels and title
plt.xlabel('Total Runs')
plt.ylabel('Team')
plt.title('Top 10 Teams by Total Runs')

# Show plot
plt.tight_layout()
plt.savefig('RunLeaders-Teams')
plt.show()
```

Stolen Base Leaders

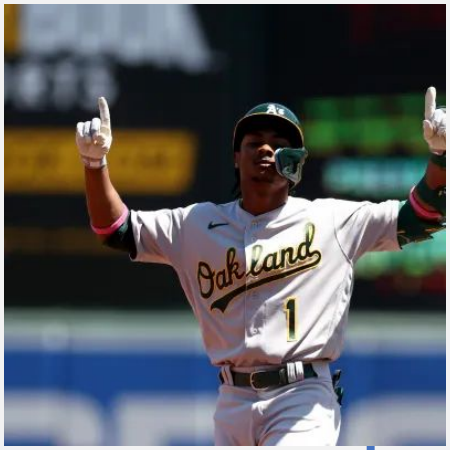
Who Stole the Most Bases?

Successful steal attempts are an absolute game changer during an inning, increasing any offense's chances of scoring. These results show us which players to watch on the base paths.

```
1  -- Which players stole the most bases?
2
3  SELECT DISTINCT "Name", "Team", "SB"
4  FROM "HITTING"
5  ORDER BY "SB" DESC
6  LIMIT 10;
7
```

Data Output | Messages | Notifications

	Name character varying	Team character varying	SB integer
1	Ronald Acuna Jr.	ATL	45
2	Esteury Ruiz	OAK	43
3	Corbin Carroll	ARI	29
4	Wander Franco	TBR	28
5	Bobby Witt	KCR	27
6	Starling Marte	NYM	24
7	Jorge Mateo	BAL	23
8	Willi Castro	MIN	22
9	Jake McCarthy	ARI	22
10	Julio Rodríguez	SEA	22



Rally Enders

Who Grounded Into the Most Double Plays?

There's nothing that ruins a good rally more than a double play. Be cautious with these teams and players who ground into double plays often.


```sql
-- Who Grounded into the most double plays?

SELECT DISTINCT "Name", "GDP" as "Double Plays"
FROM "HITTING"
ORDER BY "GDP" DESC
LIMIT 10;
```

Data Output | Messages | Notifications

	Name character varying	Double Plays integer
1	Carlos Correa	19
2	Ty France	15
3	Vladimir Guerrero Jr.	14
4	Austin Riley	14
5	Luis Arraez	13
6	George Springer	13
7	Jorge Soler	13
8	Alec Bohm	13
9	Hunter Renfroe	13
10	Manny Machado	13

```sql
-- Which teams grounded into the most double plays?

SELECT "Team", SUM("GDP") as "Double Plays"
FROM "HITTING"
GROUP BY "Team"
ORDER BY "Double Plays" DESC
LIMIT 10;
```

Data Output | Messages | Notifications

	Team character varying	Double Plays bigint
1	MIA	210
2	LAA	166
3	TOR	164
4	HOU	160
5	COL	160
6	ATL	158
7	STL	158
8	MIL	158
9	WSN	154
10	SDP	148

Pitching Analysis

Workhorse Starting Pitchers

With pitching injuries on the rise, we've seen a decrease in total innings pitched by starting pitchers. However, these 3 starters have no issues taking on a heavy workload.

```sql
-- Which pitchers pitched the most innings?

SELECT DISTINCT "Name", "IP"
FROM "PITCHING"
ORDER BY "IP" DESC
LIMIT 10;
```

Data Output | Messages | Notifications

	Name character varying	IP numeric
1	Logan Webb	134.1
2	Zac Gallen	130.1
3	Gerrit Cole	129.1
4	Miles Mikolas	126.2
5	Sandy Alcantara	126.1
6	Aaron Nola	126.1
7	Nathan Eovaldi	123.2
8	Mitch Keller	123.0
9	Marcus Stroman	122.1
10	Framber Valdez	122.1

Logan Webb

Logan Webb lead the entire MLB with 134.1 IP during this time span - 4 innings more than anyone else.

Zac Gallen

Zac Gallen joins his fellow NL West competitor, Logan Webb, and comes in 2nd place with 130.1 IP

Gerrit Cole

New York Yankees ace Gerrit Cole fell just 1 inning short from topping Gallen & comes in at 3rd place with 129.1 IP

```sql
SELECT
    DISTINCT "Name",
    "ERA",
    "IP"  as "Innings Pitched"
FROM "PITCHING"
WHERE "IP" > 100
ORDER BY "ERA"
LIMIT 10;
```

Data Output | Messages | Notifications

	Name character varying	ERA numeric	Innings Pitched numeric
1	Blake Snell	2.67	108.0
2	Nathan Eovaldi	2.69	123.2
3	Gerrit Cole	2.78	129.1
4	Shane McClanahan	2.89	106.0
5	Framber Valdez	2.94	122.1
6	Justin Steele	2.95	103.2
7	Luis Castillo	3.04	118.1
8	Marcus Stroman	3.09	122.1
9	Jordan Montgomery	3.14	109.0

```sql
SELECT
    "Team",
    ROUND(AVG("ERA"), 2) AS "Avg Team ERA"
FROM "PITCHING"
GROUP BY "Team"
ORDER BY "Avg Team ERA"
LIMIT 10;
```

Data Output | Messages | Notifications

	Team character varying	Avg Team ERA numeric
1	CLE	3.69
2	ATL	3.89
3	NYY	4.11
4	TOR	4.13
5	MIN	4.24
6	NYM	4.53
7	HOU	4.87
8	PIT	4.91
9	SEA	5.00

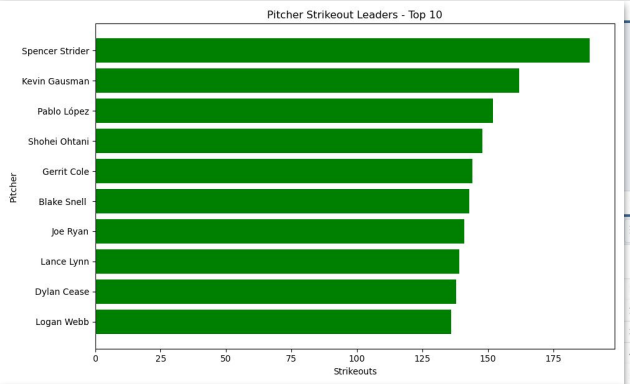
Earned Run Average Leaders

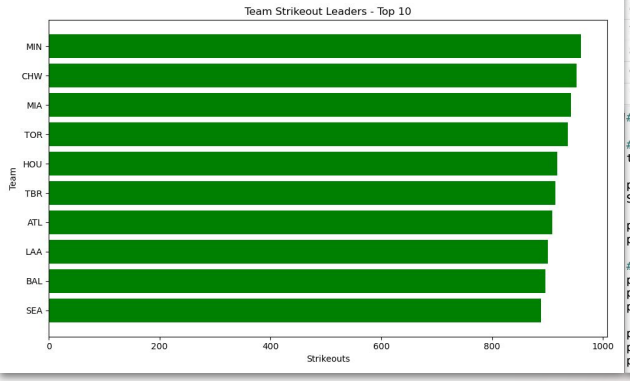
Teams & Players that Were the Best at Stifling Opposing Offenses

For years Earned Run Average (ERA) has been the top stat to analyze pitcher performance - The pitchers with the lowest ERA are viewed as the best pitchers in the league, proving they can prevent the opposing team from scoring runs.

Blake Snell lead the way with a 2.67 ERA and eventually won the 2023 NL Cy Young Award given to the league's best pitcher.

The Cleveland Guardians further established themselves as a pitching factory, posting a 3.69 ERA, which is the best ERA out of any MLB team in this time frame.





Query | Query History

```sql
-- Which pitchers struck out the most hitters?

SELECT
    DISTINCT "Name",
    "SO"
FROM "PITCHING"
ORDER BY "SO" DESC
LIMIT 10;
```

Data Output | Messages | Notifications

	Name character varying	SO integer
1	Spencer Strider	189
2	Kevin Gausman	162
3	Pablo López	152
4	Shohei Ohtani	148
5	Gerrit Cole	144
6	Blake Snell	143
7	Joe Ryan	141
8	Lance Lynn	139
9	Dylan Cease	138
10	Logan Webb	136

```python
# Create bar chart using Matplotlib

# Sort the DataFrame by Strikeouts in descending order
top10_Ks = top10_Ks.sort_values(by='SO', ascending=True)

players = top10_Ks['Name']
SO = top10_Ks['SO']

plt.figure(figsize=(10, 6))
plt.barh(players, SO, color='green')

# Add labels and title
plt.xlabel('Strikeouts')
plt.ylabel('Pitcher')
plt.title('MLB Strikeout Leaders - Top 10')

plt.tight_layout()
plt.savefig('MLB Strikeout Leaders.png')
plt.show()
```

Strikeout Leaders

Teams & Players Who K'd the Most Batters

A strikeout is one of the most dominant outcomes in baseball - Let's take a look at which teams and pitchers were the most dominant.

- Spencer Strider struck out 27 more batters than any other pitcher in the league.
- The Miami Marlins finished with the 3rd most strikeouts despite not having a single pitcher finish in the top 10.
- The Minnesota Twins, who finished with the most strikeouts from any team, were the only team two have two pitchers finish in the top 10: Pablo Lopez & Joe Ryan.

Saves Leaders

The closer has become both one of the most exciting and pivotal positions on an MLB team. These are the closers who ‘shut the door’ most frequently in this time span.

Giants’ closer Camilo Doval lead the way with 30 saves - one more than Reds’ closer Alexis Diaz. Guardian’s flamethrower Emmanuel Clase & Orioles’ closer Felix Bautista tied for third place with 27 total saves.


```sql
-- Which pitchers led the league in saves?

SELECT
    DISTINCT "Name",
    "SV"
FROM "PITCHING"
ORDER BY "SV" DESC
LIMIT 10;
```

Data Output | Messages | Notifications

	Name character varying	SV integer
1	Camilo Doval	30
2	Alexis Díaz	29
3	Emmanuel Clase	27
4	Félix Bautista	27
5	Jordan Romano	26
6	Devin Williams	25
7	Josh Hader	24
8	Ryan Pressly	23
9	Carlos Estévez	22
10	Kenley Jansen	21

Pitchers Who Struggled

Teams & Individual Pitchers that Allowed the Most Home Runs

Hitters get excited when they face these pitchers… Here are the teams and individual pitchers who gave up the most home runs in this time frame.

Lance Lynn and Chase Anderson both allowed a league leading 28 home runs to opposing hitters.

The Oakland Athletics as a staff combined to allow the most home runs out of all MLB teams (142). The Chicago White Sox & Washington Nationals were close behind, allowing a total of 139 home runs.


```sql
-- Which pitchers allowed the most home runs? Minimum 100 innings pitched
-- Dividing by 2 here since the amounts were somehow being doubled.

SELECT
    DISTINCT "Name",
    SUM("HR")/2 as "HRs Allowed"
FROM "PITCHING"
GROUP BY "Name"
HAVING SUM("IP") > 100
ORDER BY "HRs Allowed" DESC
LIMIT 10;
```

Data Output | Messages | Notifications

	Name character varying	HRs Allowed bigint
1	Chase Anderson	28
2	Lance Lynn	28
3	JP Sears	24
4	Austin Gomber	22
5	Dean Kremer	22
6	Max Scherzer	22
7	Tyler Wells	22
8	Yusei Kikuchi	22
9	Aaron Nola	21
10	Jordan Lyles	20

```sql
-- Dividing by 2 here since the amounts were somehow being doubled.

SELECT
    DISTINCT "Team",
    SUM("HR")/2 as "HRs Allowed"
FROM "PITCHING"
GROUP BY "Team"
ORDER BY "HRs Allowed" DESC
LIMIT 10;
```

Data Output | Messages | Notifications

	Team character varying	HRs Allowed bigint
1	OAK	142
2	CHW	139
3	WSN	139
4	COL	138
5	CIN	134
6	BOS	128
7	NYM	128
8	TOR	126
9	LAA	121

THANK YOU!