mlb-visualization

March 4, 2025

1 Final Project: Visualizing MLB Batting Performance Trends

1.1 Recap of Data, Goals, and Tasks

You can work with the interactive notebook at https://mlb-dashboard-app-d27477e94865.herokuapp.com/voila/render/mlb-visualization.ipynb

Note book can be found here: https://github.com/gareytwin1/mlb-dashboard

Dataset Overview

- **Source:** Kaggle (2023 MLB Player Stats)
- **Focus:** Batting statistics for the 2023 MLB season
- **Key Attributes Used:** Home Runs (HR), Batting Average (AVG), Slugging Percentage (SLG), On-Base Percentage (OBP), Team names

Project Goals

1. Analyze the relationship between Batting Average (AVG) and Home Runs (HR)
2. Compare power-focused teams (HR, SLG) vs. contact-focused teams (AVG, OBP)
3. Identify the top home-run-hitting teams in 2023

Planned Visualizations

- **Scatter Plot:** HR vs. AVG (color-coded by team)
- **Violin Plot:** Distribution of HR, SLG, AVG, and OBP by team
- **Bar Chart:** Total HR per team (sorted in descending order)

1.2 Import Libraries and Load Data

[1]:
```
import pandas as pd
import altair as alt

import warnings
warnings.filterwarnings('ignore')

# Load the data with specified encoding and delimiter
```

```
mlb_data = pd.read_csv('2023 MLB Player Stats - Batting.csv',␣
 ↪encoding='ISO-8859-1', delimiter=';')
```

[2]:
```
mlb_data.columns
```

[2]:
```
Index(['Rk', 'Name', 'Age', 'Tm', 'Lg', 'G', 'PA', 'AB', 'R', 'H', '2B', '3B',
       'HR', 'RBI', 'SB', 'CS', 'BB', 'SO', 'BA', 'OBP', 'SLG', 'OPS', 'OPS+',
       'TB', 'GDP', 'HBP', 'SH', 'SF', 'IBB'],
      dtype='object')
```

1.3 Visualization Implementation

Scatter Plot: Shows the relationship between Home Runs (HR) and Batting Average (AVG), colored by team. This helps analyze whether power hitters maintain high averages.

[6]:
```
# Convert categorical columns to strings (Altair prefers this)
mlb_data["Tm"] = mlb_data["Tm"].astype(str)

# Dropdown filter with "HOU" pre-selected
team_dropdown = alt.binding_select(options=mlb_data["Tm"].unique().tolist(),␣
 ↪name="Select Team: ")
team_selection = alt.selection_single(fields=["Tm"], bind=team_dropdown,␣
 ↪name="Team")

# Scatter Plot: Home Runs vs. Batting Average
scatter_plot = alt.Chart(mlb_data).mark_circle(size=80).encode(
    x=alt.X("HR:Q", title="Home Runs"),
    y=alt.Y("BA:Q", title="Batting Average"),
    color=alt.Color("Tm:N", legend=None),
    tooltip=["Name:N", "Tm:N", "HR:Q", "BA:Q"]
).add_selection(
    team_selection
).transform_filter(
    team_selection
).interactive().properties(
    title="Home Runs vs Batting Average (Interactive)",
    width=600,
    height=400
)

# Add regression line to the scatter plot
scatter_plot = scatter_plot + scatter_plot.transform_regression("HR", "BA").
 ↪mark_line(color="red")


# Display the scatter plot
```

```
scatter_plot.show()
```

```
alt.LayerChart(…)
```

Violin Plot: Displays the distribution of Home Runs across different teams, helping compare power-focused and contact-focused teams.

[9]:
```
import plotly.express as px

# Initial plot with HR
fig = px.violin(mlb_data, x="Tm", y="HR", box=True, points=False,␣
 ↪title="Distribution by Team",
                labels={"Tm": "Team", "HR": "Home Runs"}, color="Tm")

# Show the plot
fig.show()
```

Bar Chart: Ranks MLB teams by total Home Runs, making it easy to identify the most power-heavy teams.

[8]:
```
# Using Altair, create a bar chart of the total home runs by team
# Aggregate total home runs by team
hr_by_team = mlb_data.groupby('Tm')['HR'].sum().reset_index()

# Sort teams by total home runs in descending order
hr_by_team = hr_by_team.sort_values(by='HR', ascending=False)

# Create a bar chart
bar_chart = alt.Chart(hr_by_team).mark_bar().encode(
    x=alt.X('Tm:N', sort='-y', title='Team'),
    y=alt.Y('HR:Q', title='Total Home Runs'),
    color=alt.Color('HR:Q', scale=alt.Scale(scheme='blues'), legend=None),
    tooltip=['Tm:N', 'HR:Q']
).properties(
    title='Total Home Runs by Team',
    width=600,
    height=400
).interactive()

bar_chart
```

[8]:
```
alt.Chart(…)
```

1.4 Summary of Key Design Elements

Scatter Plot Enhancements:

- Color-coded by team for quick identification.
- No legend clutter. User picks filters team by interacting with dropdown menu.

Violin Plot Enhancements:

- Inner quartile marks to show the distribution clearly.
- Interactive rollover to show values of violin distribution

Bar Chart Enhancements:

- Sorted in descending order for easy interpretation.
- Blue gradient (Blues_r) to emphasize higher values.
- Grid lines for readability.

1.5 Evaluation Approach

Participants:

- A mix of baseball coaches, fans, and analysts.
- If experts are unavailable, friends, family or classmates with an interest in baseball analytics will be recruited.

Evaluation Methods:

1. User Surveys & Questionnaires:
   - Rate clarity, effectiveness, and ease of interpretation.
   - Example question: "On a scale of 1-5, how easy was it to interpret the scatter plot?"
2. Task-Based Testing:
   - Participants will complete tasks such as:
     - "Identify the team with the most home runs."
     - "Determine whether high-HR players generally have high AVG."
   - For deeper insights, combine open-ended and multiple-choice questions: *"Which team shows the highest variability in home run distribution?"*
     a) Yankees

     b) Dodgers

     c) Braves

     d) Astros
3. Expert Feedback:
   - Baseball coaches will review statistical validity and practical use cases.

Success Criteria:

- 80% of users should find the visualizations intuitive.
- 70% of users should correctly identify key insights.
- 50% of users should interact with filters or annotations (if an interactive dashboard is built).

1.6 Summary of Key Elements of the Design and Justification

The visualizations in this project were designed to provide clear insights into MLB batting statistics, specifically focusing on HR (Home Runs), SLG (Slugging Percentage), BA (Batting Average), and OBP (On-Base Percentage).

The key visualizations include:

1. **Scatter Plot (HR vs. BA)**
   - **Why?** This explores the relationship between a player's home run power and their batting average.
   - **Justification:** Helps determine if power hitters sacrifice contact ability for home runs.
   - **Enhancements:** Added dropdown to filter by team. Also, added tooltips to display player stats on rollover.
2. **Violin Plot for HR distributions by Team**
   - **Why?** This visualization helps display the distribution of player statistics within each team, showing variability and density.
   - **Justification:** It allows analysts and fans to compare how teams emphasize power-hitting vs. contact-hitting and spot outliers.
   - **Enhancements:** interactivity and rollover was implement to show values to users.
3. **Bar Chart (Total HR per Team)**
   - **Why?** This visualization helps in identifying which teams rely on power-hitting for their offensive success.
   - **Justification:** A simple, intuitive visualization for team home run totals.
   - **Enhancements:** Sorted in descending order, with the top teams represented in a darker blue hue.

1.7 Final Evaluation Approach

Procedure

1. **Participants**
   - 2 MLB fans
   - 1 ex-baseball coach
2. **Evaluation Methods**
   - **Task-Based Testing**: Participants were asked to answer specific questions using the visualizations:
     - *Which team has the widest HR distribution?*
     - *Do power hitters also have a high batting average?*
     - *Which teams rely the most on home runs for scoring?*
   - **Observations & Feedback**: Their ease of use, struggles, and insights were recorded.
   - **Post-Evaluation Discussion**: A brief discussion about which visualizations were most useful and what needed improvement. Used scoring criteria to evaluate success criteria.

Results

- **Violin Plot**: Initially hard to understand for all participants. However, after a quick training session explaining violin plots and distributions, they were able to extract insights.

- **Scatter Plot:** Found slightly misleading because it included pitchers who rarely bat. This skewed the distribution.
- **Bar Chart**: Most intuitive and easiest to understand. All participants quickly grasped insights and were surprised that some of the top teams were not the league leaders in win totals. So home runs doesnt necessarily correlate to win totals.

1.8 Synthesis of Findings and Future Refinements

What Worked Well?

- **Bar Chart** is highly intuitive, great for team-level insights.
- **Dropdown Filtering** allowed flexibility in exploring different statistics.
- **Violin Plot (after training)** helped users identify variability and outliers.
- **Participants were able to correctly answer questions** once they understood the charts.

Areas for Improvement and Future Refinements **Scatter Plot Skewed by Pitchers**
- **Issue:** Pitchers with very few at-bats (AB) affected overall distribution.
- **Future Fix:** Filter out pitchers who have fewer than a certain number of ABs.

Violin Plot Complexity
- **Issue:** Hard for non-experts to understand initially.
- **Future Fix:** Provide a brief guide or tooltip explanation about violin plots.

Enhance Interactivity
- **Issue:** Some participants wanted to see individual player names more easily.
- **Future Fix:** Add hover tooltips with more detailed stats (e.g., OPS, RBI).

1.9 Conclusion

This evaluation provided valuable insights into which visualizations were most effective and which needed refinement.
The bar chart was the most accessible, while the violin plot required some user training but was useful afterward.
Future improvements will focus on clarifying complex plots, improving scatter plot data filtering, and enhancing interactivity.