Statcast Power Hitters Analysis

Executive Summary

This project applies data wrangling, exploratory analysis, and visualization techniques to *Major League Baseball's (MLB) Statcast* data, focusing on the performance of Aaron Judge and Giancarlo Stanton — two of the league's most dominant power hitters. Using granular pitch-by-pitch data from 2015–2017, the report uncovers similarities and differences in their home run production, pitch zone performance, and physical profiles. These insights demonstrate actionable analytical skills prized in sports analytics roles and showcase real-world data acumen for recruiters and data science professionals alike.

1. Business Context

- **Industry**: Sports Analytics / Major League Baseball
- **Challenge**: Teams and fans increasingly seek quantitative scouting—understanding what makes elite sluggers successful (or vulnerable) using high-resolution Statcast tracking data.
- **Objective**: Compare and contrast Aaron Judge and Giancarlo Stanton, illuminating how two towering sluggers generate (and differ in) home run production, pitch selection, and zone dominance.

2. Data Summary

Dataset	Description
judge.csv	All Statcast pitches to Aaron Judge (2015–17)
stanton.csv	All Statcast pitches to Giancarlo Stanton (2015–17)

Key Variables:

- `pitch_type`: Type of pitch thrown (e.g., fastball, slider)
- `zone`: Pitch location within strike zone (1–14)
- `events`: Result (e.g., home_run)
- `release_speed`, `launch_angle`, `launch_speed`: Ball physics metrics

- `description`/`des`: Human-readable outcome
- `game_date`, `game_year`: Temporal markers

3. Methodology

- **Data Loading & Inspection**: Imported Judge's and Stanton's Statcast CSVs using pandas; inspected tail of data for schema validation.
- **Feature Engineering**: Defined custom functions to map Statcast's complex strike zone numbering to `(x, y)` zone coordinates for effective visualization.
- **Exploratory Data Analysis (EDA)**:
  - Examined pitch types faced and home run outcomes.
  - Assessed strike zone coverage and "hot zones" for each hitter.
  - Compared physical attributes and home run tendencies.
- **Visualization**: Plotted spatial heatmaps and bar charts to show distinctive slugging patterns, powered by Matplotlib and Seaborn.

4. Key Findings

- **Physical Comparison**:
  - Aaron Judge: 6'7", 282lb — one of MLB's largest ever.
  - Both Judge and Stanton outpaced league peers in home runs (Judge: 52 in 2017; Stanton: 59 in 2017), indicating rare, game-changing power.
- **Home Run Analysis**:
  - Both hitters display "hot zones" — specific strike zones where they launch the most homers.
  - Stanton and Judge led MLB in home runs, with substantial gaps over the next closest hitter (45).
- **Pitch Zone Performance**:
  - Custom `assign_x_coord` and `assign_y_coord` functions facilitated mapping the 9-box strike zone for easily interpretable visualizations.

    - Analysis revealed that both batters exploit similar pitch locations, but heatmaps show subtle differences in where each thrives or struggles.
- **Statcast Value**: The project highlights the depth of insights available from granular, pitch-level Statcast data, showcasing modern sports analytics in action.

5. Business & Technical Impact

Business Value:

- **For Teams**: Refines scouting and pitching decisions—knowing hitter strengths/weaknesses by zone and pitch type.
- **For Recruiters**: Demonstrates ability to turn raw sports data into actionable insights, mirroring analytics-driven roles in sports franchises or data-driven companies.

Technical Value:

- Mastery in wrangling large, complex CSV data; strong EDA and visualization.
- Custom mapping and zone heatmap techniques.
- Concise storytelling and communication of technical findings for both analysts and non-technical stakeholders.

6. Recommendations & Next Steps

- Perform year-over-year trend analysis to determine if "hot" and "cold" zones change as pitchers adapt.
- Expand project to additional hitters or pitchers for league-wide benchmarking.
- Integrate Statcast batted ball and sprint speed metrics to explore multidimensional talent (beyond home runs).

7. Conclusion

This project demonstrates advanced data wrangling, EDA, and visualization skills using a celebrated public sports dataset. By clarifying both physical and performance differences between two elite sluggers, it bridges the gap between raw data and actionable, impactful insight. Such capabilities are valued in sports analytics, data science, and any data-driven decision-making context.