Final Description
James Bischoff, Jeffrey Tsang, Joaquin Amante

Overview

A. **A description of the data**. Report where you got the data. Describe the variables. If you had to reformat the data or filter it in any way, provide enough details that someone could repeat your results. If you combined multiple datasets, specify how you integrated them. Mention any additional data that you used, such as shape files for maps. Editing is important! You are not required to use every part of the dataset. Selectively choosing a subset can improve usability. Describe any criteria you used for data selection. (10 pts)

B. **A description of the mapping from data to visual elements**. Describe the scales you used, such as position, color, or shape. Mention any transformations you performed, such as log scales. (10 pts)

C. **The story**. What does your visualization tell us? What was surprising about it? (5 pts)

A) Description of Dataset

We utilized the MLB Statcast data for hardest-hit baseballs based on Exit Velocity from http://m.mlb.com/statcast/leaderboard#exit-velo. We used all data of the hardest-hit balls based on exit velocity from 2015 to 2017 during the regular seasons. The MLB Statcast dataset tabulates the following set of parameters for each data point: Batter, Exit Velocity, Distance, Launch Angle, Hit Type (i.e. Outcome of At-Bat), Pitcher, Pitch Speed, and Date. We used the following subset of these parameters for our visualization: Exit Velocity, Distance, Launch angle and Outcome of At-Bat. This data was copied from the MLB statcast website into an Excel sheet and ultimately converted to a .CSV file so that we could parse the data for the visualization.

The exit velocity is the velocity of the baseball as it comes off the bat, immediately after a batter makes contact. The launch angle is a vertical angle from the ground at which the ball leaves a player's bat after being struck. The distance is the linear distance from home plate the ball travels after being hit, and the outcome of the pitch is whether it results in a homerun or not.

Using launch angle and exit velocity, we aimed to calculate the distance predicted by kinematic equations of motion. Simplifying and combining the kinematic equations for the vertical and horizontal directions, we arrive at the following expression for distance:

$$x(t) = v * \mathrm{Cos}[\theta] * \frac{\left(-v * \mathrm{Sin}[\theta] - \sqrt{(v * \mathrm{Sin}[\theta])^2 - 4 * g * h}\right)}{2 * g} \qquad \text{(Eq. 1)}$$

where x = predicted distance from home plate, v = exit velocity, θ = launch angle, h = initial height at $x(t = 0)$, and g = gravitational acceleration. For each batted ball resulting in a home run, we perform this calculation, ultimately comparing the range of predicted distances to the range of actual distances. As provided, the data set uses imperial units (i.e. miles per hour, feet, angles in degrees) for all values. Before substituting values into the kinematic equation, all quantities were converted to SI units (i.e. meters per second, meters, angles in radians) for simplicity. After processing, the calculated values were converted to back to the appropriate imperial unit before aesthetic mapping.

A second dataset describes the kinematic-predicted minimum angle to achieve the shortest recorded home run distance (378 ft). Using Wolfram Mathematica, we iterate over the span of recorded exit velocities, back-solving for the launch angle required to attain x = 378 ft at each increment in exit velocity. Collectively, these points define the minimum launch angle threshold needed to hit a home run according to the kinematic model. This second dataset contains tuples of the form (α, v), where α = the calculated minimum angle and v = the corresponding exit velocity. These points were compiled into a .CSV file and loaded alongside the MLB statcast data. We ultimately use this data to compare the minimum recorded launch angle for a home run to the kinematically derived prediction.

Additionally, two external SVG files were imported into the visualization: one for a baseball batter silhouette, and one for an illustration of a baseball. These can be found at the following addresses:

- http://www.supercoloring.com/sites/default/files/silhouettes/2015/07/baseball-batter-black-silhouette.svg
- https://upload.wikimedia.org/wikipedia/commons/9/92/Baseball.svg

B) Description of Visual Mapping

We used a baseball batter silhouette for both visualizations to help guide the reader. In each visualization, the trajectory of the baseball emanates from the baseball in the figure, emphasizing the context of our dataset.

First Visualization - Predicted vs. actual home run distances

The first visualization presents a comparison of the span of kinematic-projected home run distances to the span of recorded home run distances. For our kinematic model projections, utilizing the data from MLB statcast pasted in the CSV, we substitute launch angle and exit velocity into Eq.1 to calculate the predicted distance of the home run balls. The actual distances and kinematic distances are plotted together on a linear scale, which occupies the right-hand side of the screen.

We found the min and max home run distance calculated from our kinematic model and plotted that on the linear scale in blue. In contrast, the actual range of home run distances is plotted in red on the same scale. By showing the span of distances in this plot, we hope to convey the disparity between the predictions and the actual distance data.

We include parabolic arcs to simulate the trajectory of the baseball, which serves to contextualize the home run distance data, and to emphasize the difference between the predictions and actual data. We used a dashed outline and a low-opacity fill for the trajectories of the kinematic predictions to make them distinct from those of the actual data, which are depicted as red parabolic arcs.

Second Visualization - Exit velocity of batted balls and their launch angles

The second visualization shows the hardest-hit batted balls plotted according to their exit velocities and launch angles. We plotted the home run balls as red circles and all other balls that did not result home runs as grey circles. Plotting batted balls using this binary color scale helps to emphasize the cut-off point between home runs and non-home runs. This visualization uses a polar plot to emphasize the spatial nature of our data set. The angular axis corresponds to Launch Angle, and the radial axis corresponds to Exit Velocities. Each axis employs a linear mapping between data values and position along the axis.

On this coordinate system, we placed a red line to denote the minimum recorded angle that resulted in a home run. The angle from the data was 13.5 degrees.

Using the same kinematic model from above, we then solved for the predicted launch angles needed to hit a home run. We assumed that the minimum required distance was 378 feet, in accordance with minimum home run distance from the actual data set. We numerically solved for the launch angle needed to hit the ball 378 feet as a function of exit velocity of the ball. Thus, for a range of exit velocities, we plotted a line to show the necessary angles needed to hit the ball 378 feet at a specific velocity. This was a dashed line colored in blue to maintain continuity with how we visualize distance projections.

C) The Story

Our first visualization compares the home run distance of the hardest baseballs hit with the predictions from our kinematic model. As the illustration makes evident, the actual data and the predictions are very different. This suggests that there are more forces in play than just gravity that may result in farther home runs than predicted. This point is reiterated in the second visualization, comparing the minimum launch angle needed to hit a home run ball with our kinematic model prediction. The model predicts a greater launch angle needed for a home run than the actual launch angle.

We were surprised to find out that our predictions were not close to the real data. We assumed that since drag is not accounted for, our predictions for home run distance would go beyond the data. In fact, the opposite was true, which we believe is caused by other forces in play like lift on the ball, backspin, wind speed and air pressure. These factors may decrease the launch angle needed and may result in farther home run distances. The kinematic models fails to account for these variables and thus does not predict these homerun trajectories as well.

Overall, our visualization shows the fastest balls hit that traveled between 114 to 120 mph require a launch angle between 13.5 and 30 degrees in order to result in a homerun. In addition, launch angle is more important than exit velocity when hitting home runs. Balls that did not result in home runs reached higher exit velocities. Generally, as the exit velocity increases for homerun balls, the launch angle needed to hit the ball decreases. In addition, our kinematic model underestimates home run distances based on same exit velocities and launch angles because it fails to account for other forces such as lift, backspin, wind speed and air pressure. It is likely that lift is the most important force that we have not accounted for, since the inclusion of lift would skew the projected distances upward. This would narrow the gap between projected distances and actual distances.