Final Visualizations

Dataset:

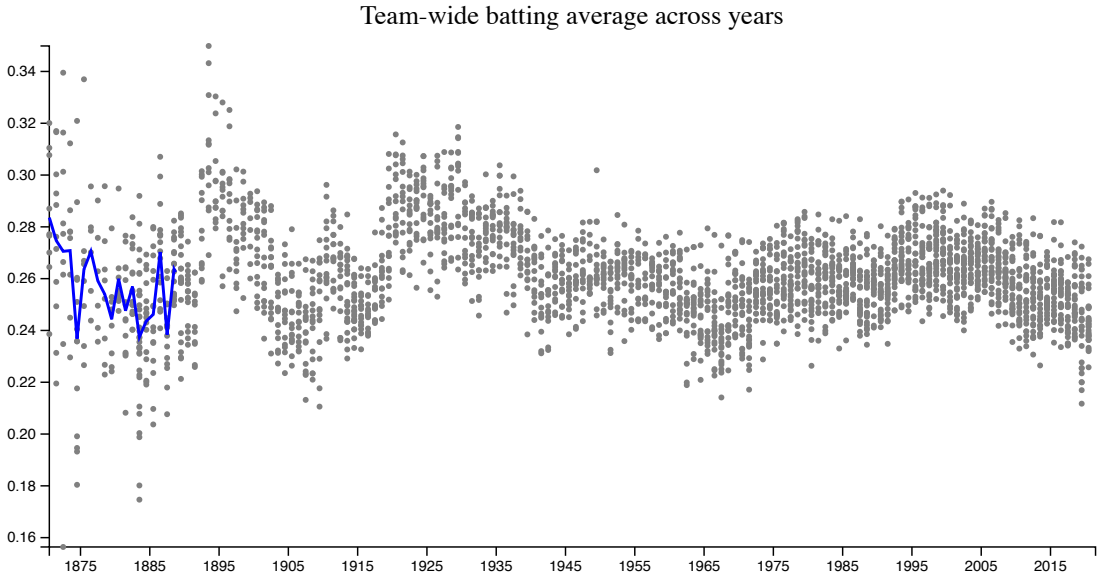
Data visualization is very important for the sport of baseball. Statistics have been kept about every pitch, hit, out, win, much more since the mid 1800s. This makes data analysis very fun to to because there is so much data in the baseball realm. For this project I have chosen a dataset called Teams.csv from the Lahman database that includes data of every year from every MLB team starting in 1871. The dataset was pretty clean, except I needed to get information about the state that each MLB park was located in as well as the longitude and latitude for that state. I did this by merging the Teams.csv dataset with the Parks.csv dataset from seamheads.com/ballparks. The following visualizations come from the merged dataset called teams1.csv.

My visualizations answer questions about how various baseball stats change from year to year and within a team.

Visualization 1:

This first visualization answers the following question: How have team batting averages changed over time?
Batting average is calculated by dividing number of hits(H) by number of at bats (AB). I created a scatterplot with years on the x-axis and calculated batting average on the y-axis. The blue line that is animated across the scatterplot is the average batting average for each year across all teams. You can drag your mouse across a point and the team name will display in a tooltip. This allows you to see which teams have the highest of lowest batting average for a given year. I have chosen grey for the scatterpoints because we are just looking for a general pattern of points and because we want to be able to see the animated line easily. I chose blue for the animated line because it is a color that is easy to see with a grey background. This graph uses point marks and bertical and horizontal spatial position channels.

Plot:



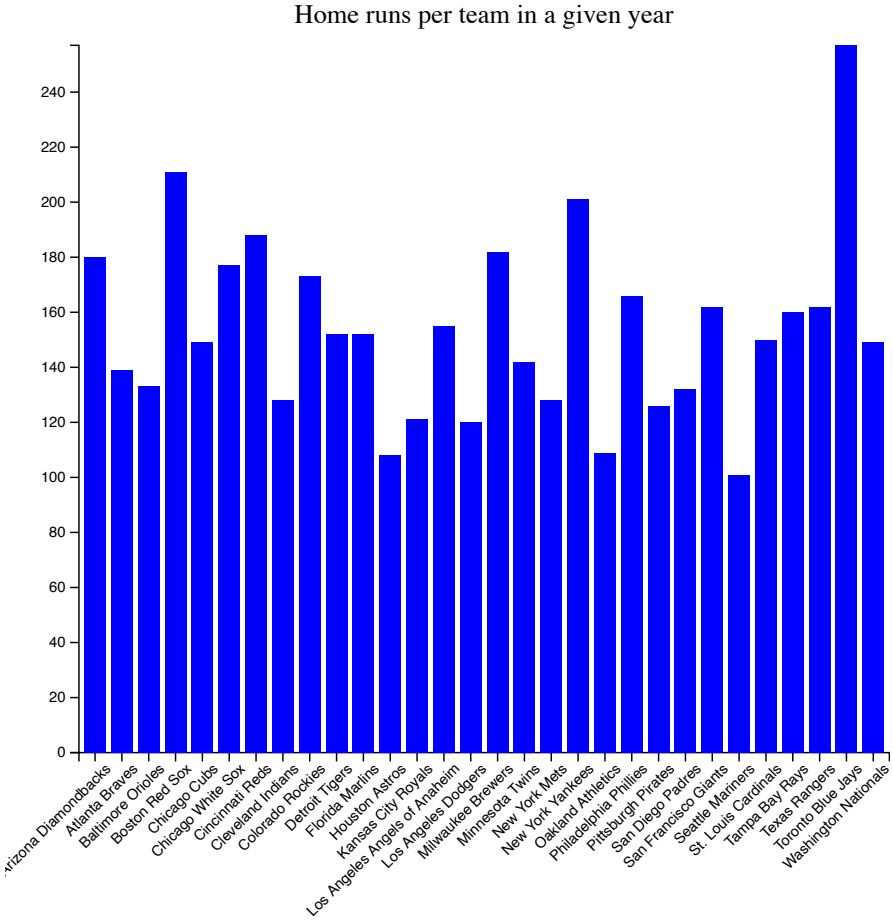
We can see from the plot that over time, batting average across teams has tightened up a lot and now typically ranges between 0.20 and 0.28 compared to earlier years where the range is much higher and the scatter points as well as the average line appear to have a sinusoidal shape.

Visualization 2:

This first visualization answers the following question: Which teams have the most home runs in a given year?
I have created a bar chart with buttons where you can select a year to see a barchart that contains the number of home runs in a given year for each team. I have chosen the color blue for the bar chart because it is easy to see and there is no other colors displayed in the graph. This graph uses a line mark and a spatial position channel.

○ 2000 ○ 2001 ○ 2002 ○ 2003 ○ 2004 ○ 2005 ○ 2006 ○ 2007 ○ 2008 ○ 2009 ◉ 2010

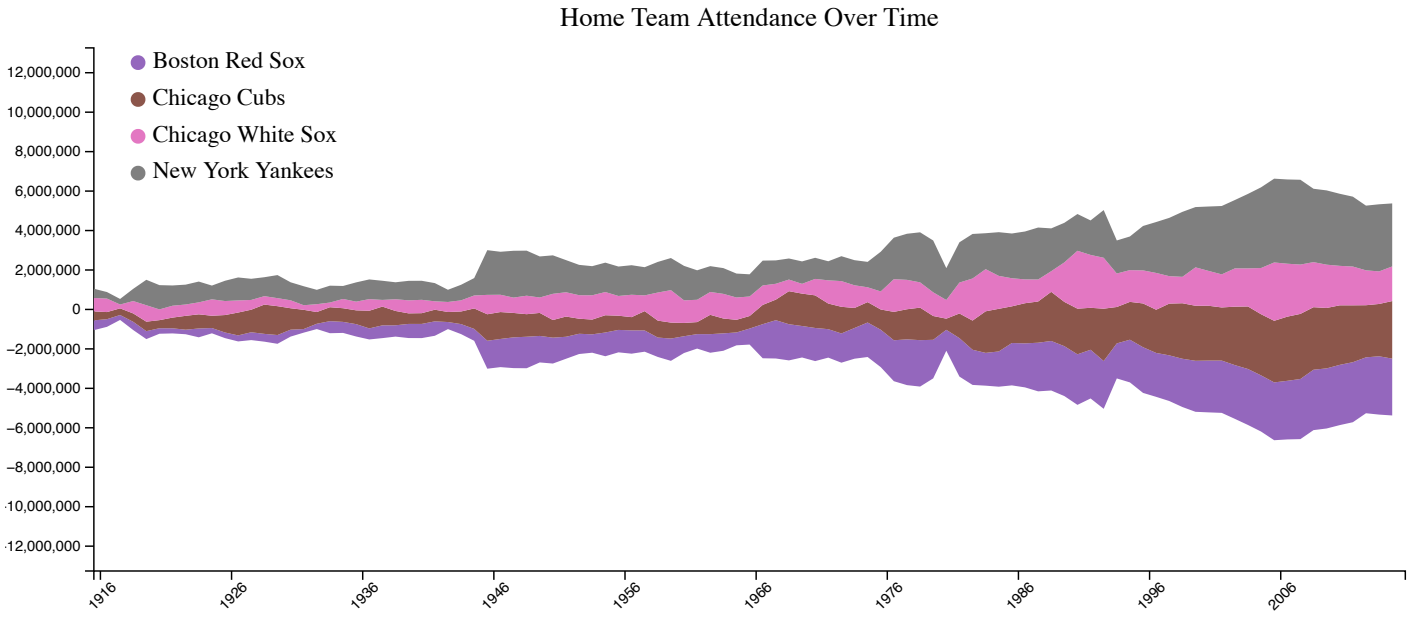
Plot:



Visualization 3:

This visualization answers the question: How has home team fan attendance changed over time?
This steamgraph displays the home team attendance for 4 of the teams that have been around the longest (Boston Red Sox, Chicago Cubs, Chicago White Sox, New York Yankees). We can see that over time, attendance has increased for all teams. There were dips in the early 1980s as well as around 1995. I have selected the following colors using different colors from schemeCategory10 because the coloring is for categorical data and there are no red and greens since they cannot be differentiated by colorblind people. This steamgraph have line marks, color channgels and horizontal and vertical position channels.

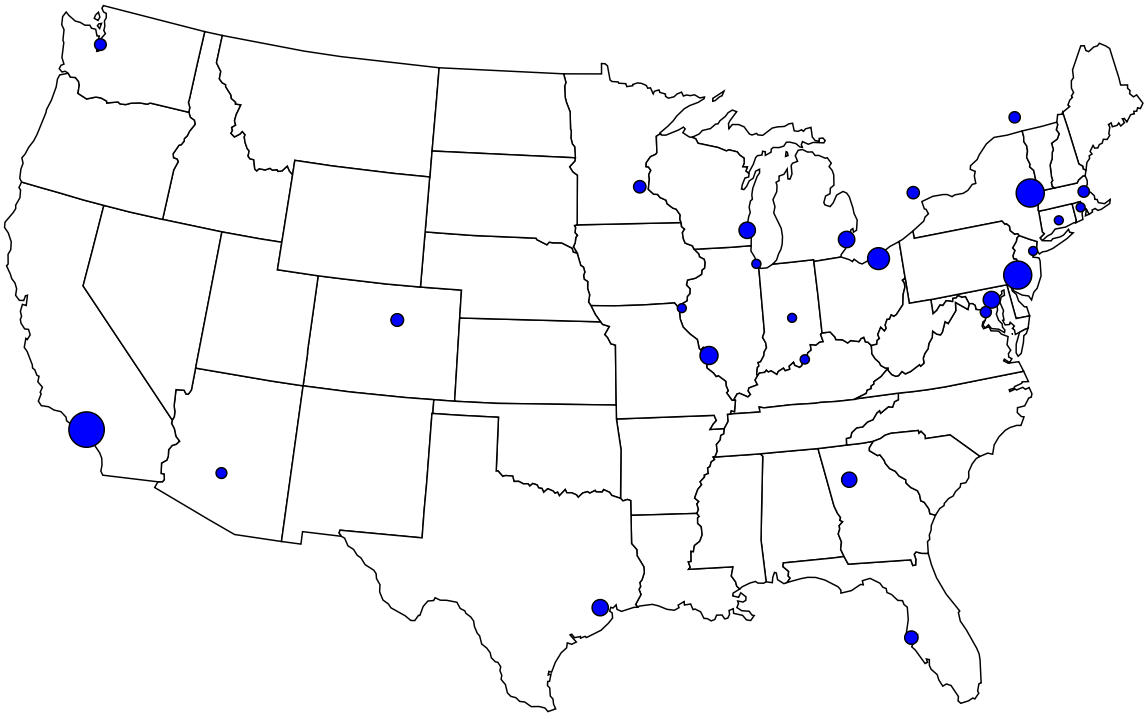
Plot:



Visualization 4:

This graph answers the question: Where in the United States are the most home runs hit?
How far a ball travels is dependent on air density and weather and thus it is easier to hit home runs in certian parks. This graph has points on a map with the radius representing the amound of home runs hit. A larger radius corresponds to more home runs. I chose blue as the color of points because it is easy to view and contrasts well with the white background map. This graph has area and point marks, size channgels and vertical and horizontal position channels.

Plot:



Visualization 5:

This graph answers the question: What is the distribution of wins per year for the San Francisco Giants?
This visualization is a box plot of the number of wins per year for the San Francisco Giants. There are line marks with vertical and horizontal position channels.

Plot: