An Analysis of Baseball Salary and Player Effectiveness

These pages are part of rdb4@illinois.edu / rdbisch@gmail.com project for CS498 Data Visualization. Here is the accompanying narrative visualization this report describes.

Messaging

The intent of my narrative visualization is to have the reader understand something about baseball salaries. The primary conclusions are that they are growing much faster than inflation, that player-performance is definitely not driving this, and that baseball's salary structure makes weird things happen to rookie's data.

Narrative Structure

The structure this story uses most closely resembles an interactive slide show. While there are no slides, the information is presented linearly in a long scrolling webpage, and each scene allows manipulation along the way. The text copy and scenes start off tied together with initial parameters that are connected--for example the text talks about the Cubs and White Sox, and sure enough that's what the first chart shows, for example. The user is encouraged to try the charts out and find supporting or counter evidence.

Visual Structure

There are a few different elements employed to ensure continuity and understanding from the reader.

- Consistent colors are used to represent team selections. These colors were also chosen to be friendly to the color blind.
- Even though each chart type and style changes from scene, at most only one-dimension is changing. It starts with (inflation x time) and then go to (performance x time) and then go to (performance-per-dollar x time), and finally (performance-per-dollars x player-year).
- Where possible and practical, transition animations were used to make clearer what is changing.
- Presumably baseball fans are consuming this, and common baseball iconography is used to represent team association.

Scenes

There are four main scenes displayed in this narrative visualization:

- The Consumer Product Index (CPI) vs. Aggregate Baseball Salaries.

  This scenes sets the stage for the entire visualization by establishing how much faster than inflation these sports players salaries are. The labels (e.g. baseline, CPI, etc.) are cleverly set to match the slope of the line-chart, creating a powerful illustration. The user can mouse-over the chart and see the annotation below change further highlighting the discrepancy of inflation vs. player salary.

- Player performance (per salary) over time.

  This is used to establish evidence that player's performance is not rising consummate with their income. The data points are plotted, and then two trend lines representing the series are overlaid-on-top. Most of the teams have very flat trends which is in stark contrast to the slope we saw with salaries above.

- Best-Player performance per salary over time.

  This is used to test the conjecture that averaging a team's performance may not be fair and it washed out any signal The textual style is similar to the previous scene, except that now we are displaying actual player results instead of a team's result. Player initials are used instead of filled circles to highlight this change and to also provide the reader with a visual cue that some players are represented over multiple years. Further there is data-on-demand available by clicking on an individual player's initials.

- Finally, Performance by Tenure.

  This is used to highlight the effect of baseball's salary contracts they have with the players. It is very similar to the 2nd scene in the hopes that it sparks some familiarity--there are data points and trend lines, and visually the two charts are identical.

The order displayed is very intentional and linear. They don't make much sense to viewed in any other way. The use of formatting is used to accentuate scene breaks

Annotations

A specific template of annotation was not used for this project, rather each of the charts has its own kind of annotation.

- The CPI chart annotation is presented as a dynamic table just under the plot and shows in numbers how large the gap is between salary and CPI. The annotation changes with a mouse move. This illustrates to the user that the tabular information is tied to the graphic.
- The performance over time annotation is a simple block of text pointing to one of the trend lines. The arrow here is dynamic following any new regression driven by a user selecting a different team parameter. This is so that the annotation still makes sense when the user selects a different region.
- The best-player annotation is a simple highlighted region with some text. It is not dynamic as it represents an annotation about the whole data. There is additional detail-on-demand annotation and changes when the user picks different players. This was chosen so they can verify the assumptions in the text and annotation, and allows them to try to find their favorite player.
- The performance by player tenure does not have any annotations.

Another annotative device used is the regression trend line shown in both charts 2 and 4. Both use a simple trend line over time to aid the user in understanding the data.

Parameters

Other than the first scene, the last three scenes all use two parameters to allow users to compare and contrast their favorite and rival teams. These are not shared across charts because it is unlikely all three scenes would ever by displayed simultaneously. In retrospect, the first chart could have also utilized these parameters to show how team salary changed compared to one another.

Some of the charts also use pre-selections as parameters. That is to say, for the sake of the narrative story, the code is preselecting points so that their data on demand is showing when user scrolls there.

Triggers

For the CPI scene, any mouseover will instantly change the highlighted time that the annotation is drawn on, and also will update the annotated table below the chart. There is no indication that this is possible to the user, but also it is not an important element. If they don't discover it, it will do very little to diminish their understanding.

In the aggregate player scene, there are two drop-down dialog boxes that are pre-selected. It is an obvious UI element that users are used to using and they will instantly know its purpose. When the selection of each box is changed, the corresponding data is reloaded into the chart and transitions animated.

In the player-detail scene, there are the same triggers as the previous scene. There are also the inclusion of triggers caused by mouse-clicks. This is non-obvious to the user. The author preselected two interesting teams and two interesting players that made an compelling story. It is hoped that by drawing a light line connecting the annotation to the selected player, the possibility of interaction with this scene becomes apparent. When users click on a set of initials, the details for the selected team will change. In the case where two teams are right on top of each other, both teams will change their selection.

Finally the triggers on the last scene are the same as tne 2nd scene. HTML Select elements that when changed load new data for that chart and animate the transitions.

Conclusion

All the code and history are contained in the github here. Of note:

- There was some significant data preprocessing done outside of the webpage. This was done in both SQLite and Python.
- The JavaScript and HTML code are structured so that the narrative visualization itself is in "report.html" and calls out to the various charts, "cpi_chart" and "ops_chart1", "ops_chart2", and so on.
- Some more data processing was done inside of "parseBaseball.js". While less than ideal it was faster to make minor modifications to data handling here rather than continually swap tools.
- No external libraries were used. MLB Icons and Wrigley field graphics are external and sources quoted either in the text or in the repository.

This was an interesting project, and while happy with the overall result, there is definitely room for improvement. The analysis could dig deeper into other plausible causes and justifications of player salaries. Pitchers could have been contemplated as well. The visualization could allow the user to brush/sweep on the time dimension. A deep dive on a specific player (like Mike Trout) would have been an interesting addition as well. Maybe next time!