FIT5147 Visualisation Project Report

Nicholas Pennell, 30338913

Introduction

The idea behind my visualization is to crate a dashboard to show of three findings from my exploration report;

- the change of what pitch types are thrown from game to game and year to year,
- the location of where each pitch is thrown,
- And Where and what happens to each pitch (the outcome).

As well adding some extra information, and visualization features threw the use of filtering the data and having links to videos.

The visualization was design for people with a knowledge of baseball looking to find more visual information about their favorited player all in one place.

Design

To start out the design procedure all ideas for plots and features of each plot where written down, (idea sheet). This yield four main sections.

The time series plot, a plot where the y-axis shows the percentage of what pitches where thrown on the corresponding date, the x-axis. The idea was to have a button to switch between how the data is group, by year or game. As well as add a way to show key events and important dates.

The next was the location plots. A way to show where each pitch is thrown to the batter. Ideas include a scatter plot, heat map, or both with a way to switch. Any of these options would be overlaid with a strike zone and a home plate to help show the ground. The plot would also have a way to link to a video the pitch, and the data be filtered to look a certain scenario.

The plot to show the outcome of a pitch yielded two ideas. A bar chart to show the count of outcomes or a scatter plot that overlays a baseball field to show where the outcome occurs. Either option would share the data filter options of the location plot.

The last idea was to add a bio page, or a way to share information about the player and/or the visualization.

The first design (sheet 2) was a simple dashboard, it included the time series plot form the idea sheet with a date slider below. A bar chart to count how many pitches where thrown and what type in the same time frame as the time series. With the location and outcome plots below that. The filter options would be on the left side of the page and effect the location and outcome plot.

The second design (sheet 3), was like first but aggraded as a shiny site, adding the bio page while splitting each plot into its own tab to give more space and detail to each plot.

The third design (sheet 4) was a d3 article/slide show, like the second design plots where broken up into their own page. As it would be an article style design each page would have more text and have and any linked videos embedded into the page rather than a link to an external source.

The finale design (sheet 5), is a combination of the other sheets. It takes the dashboard idea and uses that as one tab, takes the bio page shiny site idea as uses that as another tab. This idea was selected as I felt it displayed the data in a way that would be easy to follow and use, not overloading the user with information and be achievable to accomplish in the given time frame.

Implementation

For the implementation stage more data was sourced from several places:

- A list of player id's numbers and names, from baseball prospectus,
- Shape information of baseball field's to plot, from GitHub user bdilday,
- Json data was manually scraped from baseballsavant.mlb.com to get video links of pitches,
- A career information table was scraped from baseballsavant.mlb.com as well. It was saved to keep load times of the app shorter.
- I then made a event data file with key events, from information gathered from Baseball Reference and found matching videos on YouTube

The choice was made to use shiny to create the app, due to having more knowledge and a higher comfort level with it rather than other options. The plots are made using the plotly package due to its added interactive features over ggplot2. A full list of used packages can be found in the references.

The implementation of the bio page was straight forward, the bio section on the left side of the screen is just a head shot, ever MLB player has one and the url is the same just with a different id number, (which was in the original data frame). The table on the bio page was scraped as said above, this table was selected as it has a all the key simple stats. The table is then printed using kableExtra package due, to its html settings as well as it visually appealing look. The Tm column has its elements coloured to match the team they represent, to draw a user's attention to that attribute. The top 3 values are in each other column is also highlighted for the same reason, this time with lighter colour to be less grabbing. The * symbol was also used to show all start years. The foot note at the bottom of the table is added to explain this to the user.

The dashboard page was a much more time-consuming process with many ideas not coming to fruition, and many compromises. The time series plot at the top of the page, starts in the game-by-game view, with each line representing a pitch type. A point can be hover on to show more information about said point. The dots above the lines represent career events, and have more information when hovered on, the brighter pink dots can be clicked on, if done a new window opens to video of the event.

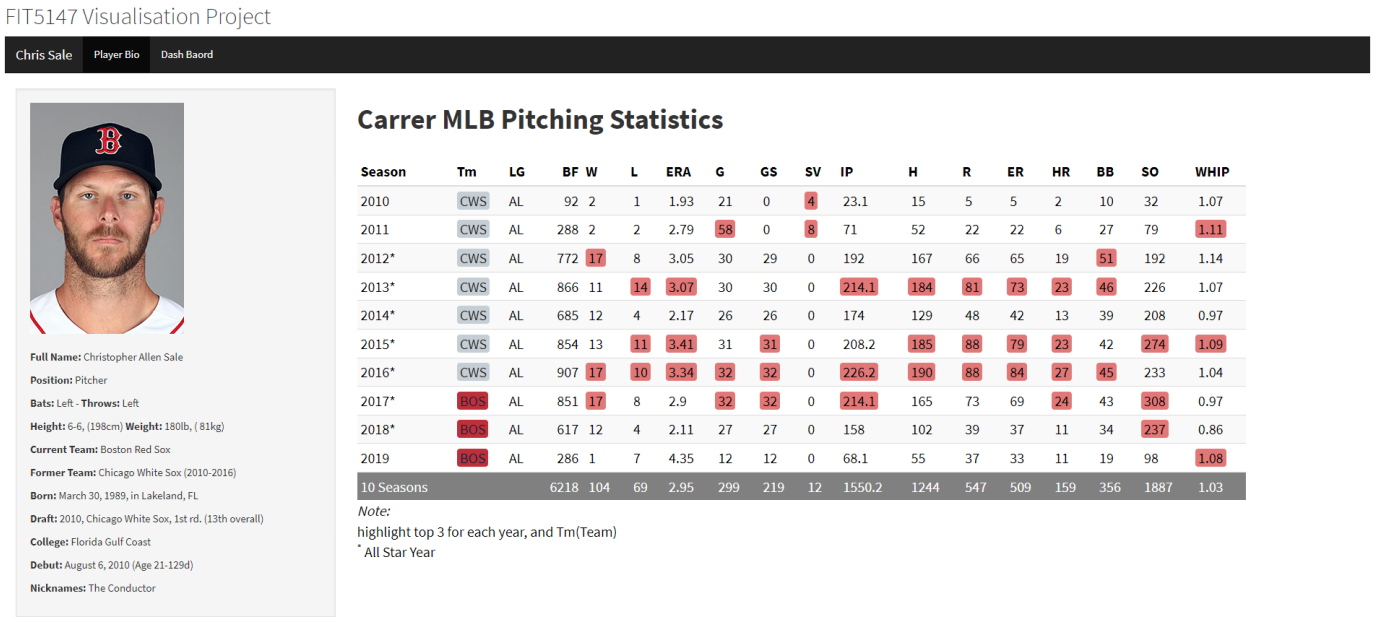
The time series plot at the top of the page, starts in the game-by-game view, with each line representing a pitch type. A point can be hover on to show more information about said point. The dots above the lines represent career events, and have more information when hovered on, the brighter pink dots can be clicked on, if done a new window opens to video of the event.

The location and hit spray chart are located below the time series. They share a row, so the user can view all plots at one time. The location plot shares colours based on the pitch type with the time series and is also faceted by pitch type to allow the user to a better view. The strike zone (the square) is the average size of an MLB strike zone and is added to show reference as well as the home plate is added to show where the ground is. The option to switch between a heat map and a scatter plot was scraped once I had to program the facet feature for Plotly myself. Some points (2019 only) have point that are clickable and will open a window to a video and some information on baseballsavant.mlb.com.

The spray chart has the same click feature as the location plot, but it over lays a shape of the baseball filed using the data from Git Hub user bdilday. A major downside with this plot is that the legend was removed due to some scenarios lead to the legend being to large and blocking the plot itself. The hover information does say what the outcome is though. The text around the baseball filed shape is the distance in feet, all measurements are in imperial as that's the standard in the for baseball and the MLB, to the edge of the filed that is plotted. The filter setting on the left side has the option I felt where the most important there are many more that could be added, the list was cut short to keep it all in one page. The date option effects all three options, while the other options just effect the location and it spray chart. The starting filter setting where set as such to show off all features but to also keep load time down.

User guide

The bio page is straight forward and has new user interaction. To get to the dashboard a user must just click on the dashboard tab.



Once the dashboard is selected the user can change setting on the left side of the page. The date filter range takes dates from the start of 2010 to the start of 2020. The drop down box filters based on the outcome of a pitch. Any number of the tick boxes can be selected at the one time.

A user may click on a bright pink dot on the time series to open a video of the event, or a point on either the location plot or the spray chart if the point has an accoupling link. Points on any of the plots can be hovered over to show more information.

Conclusion

For me this is I would have liked to spend more time on and make many changes to. There are a few things that I'm proud of, the chunk of code that facets the location plot for example, or the extra json data used to link videos. Many ideas where scraped dew to a lack of time to get them working are ways to fit it all in. A few changes or things I would like to add give more time are, more filter options for the data, a way to better link all the plots and have them interact with each other. Several extra java script functions, better use of space on the page and the ability to embed the linked videos in the page as well as an option to select a view more players. Some of this could be achieved quickly in r, but to achieve the results I would like more time needs to be spent learning and implementing with D3 and JavaScript.

references

- baseball prospectus
- baseballsavant.mlb.com
- Baseball Reference
- YouTube
- mlb.com

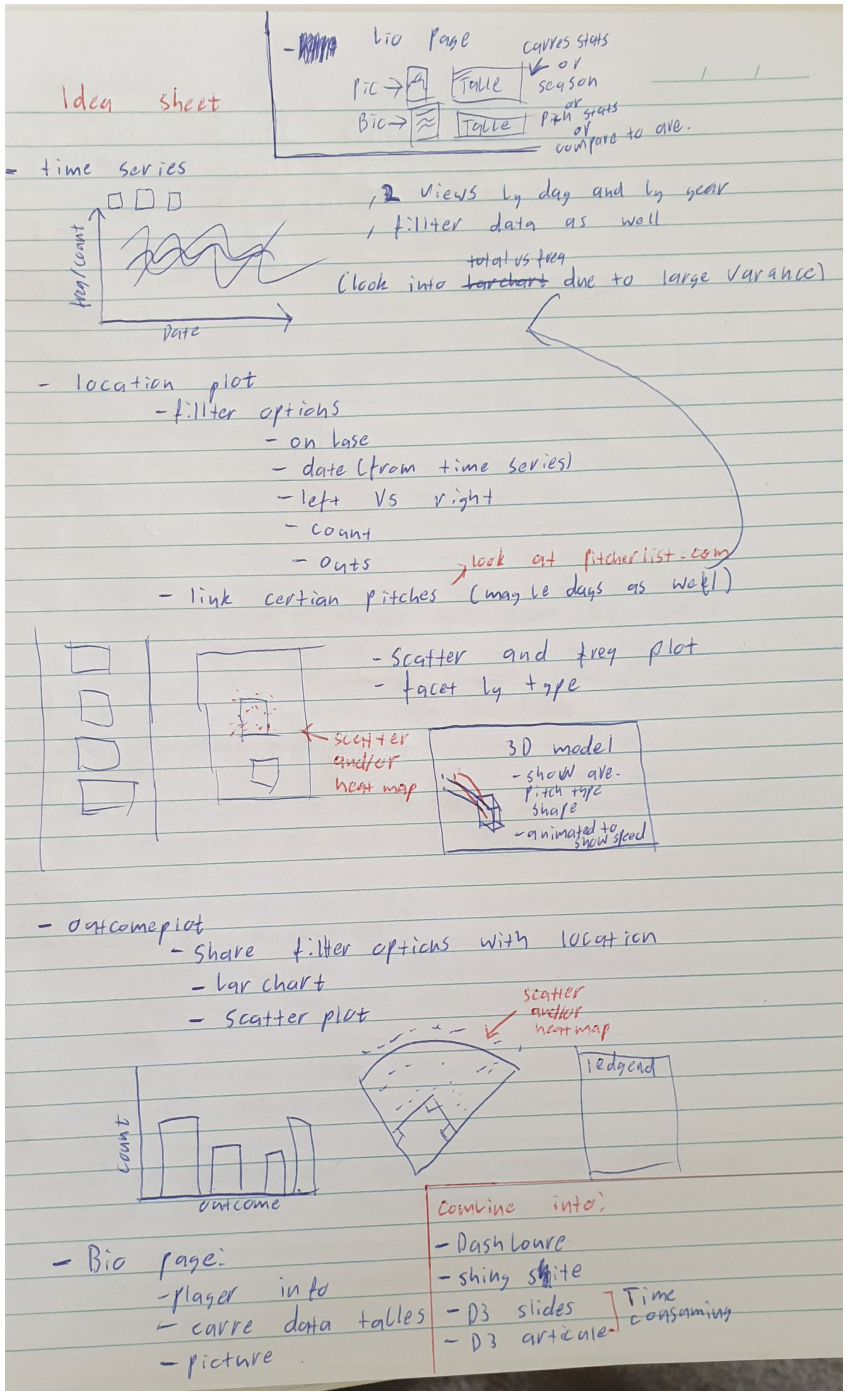
Packge list:

- shiny
- htmlwidgets, html and javascript
- shinythemes, shiny themes
- readr, read in files
- jsonlite, read json
- dplyr, data wrangling
- tidyr, data wrangling
- data.table, data wrangling
- lubridate, data wrangling
- XML, scraping
- RCurl, scraping
- rlist, scraping
- plotly, plot
- knitr, html table
- kableExtra, html table
- RColorBrewer, colour pallet

Appendix

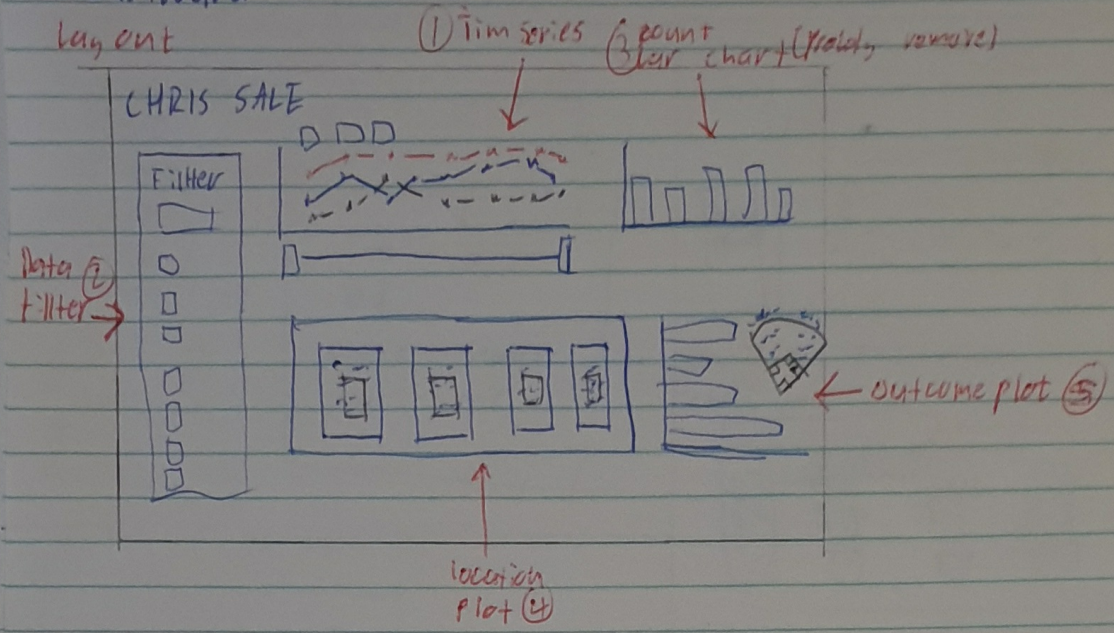
note: my five sheet method was not done 100% like the examples, as i found this way best for me to explore my ideas.

Sheet 1


Idea sheet
- Bio Page
Pic →
Table
carries stats or season
Bio →
Table
Pitch stats or compare to ave.
- time series
freq/count
Date
2 views by day and by year
filter data as well
(look into total vs freq ~~bar chart~~ due to large variance)
- location plot
- filter options
- on base
- date (from time series)
- left Vs right
- count
- outs
look at pitcherlist.com
- link certian pitches (may be days as well)
- Scatter and freq plot
- facet by type
scatter ~~and~~ or heat map
3D model
- show ave. pitch type shape
- animated to show speed
- outcomeplot
- Share filter options with location
- bar chart
- scatter plot
scatter ~~and~~ or heat map
legend
count
outcome
- Bio page:
- player info
- carre data tables
- picture
Combine into:
- Dash Loure
- shiny ~~s~~ite
- D3 slides
- D3 article
Time consuming

Dashboard

Layout



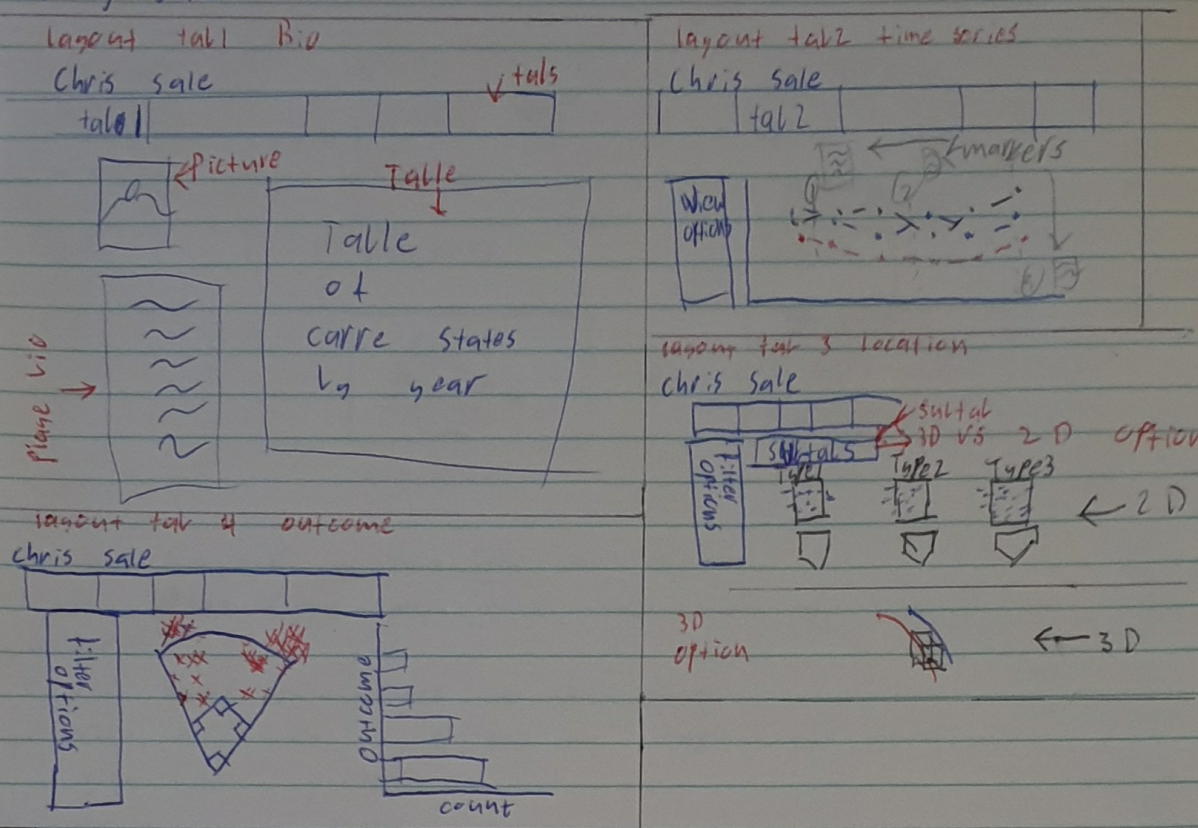
Operations

1. Time series
   - has buttons above to change ~~view~~, from by year to by date, (maybe by month/quarter)
   - possible option to change between total pitches and %
   - slider under time series use to filter data for location and outcome plots
2. Data filter
   - mostly check boxes in sections, sections being, outs(0,1,2), count(0-0, 0-1, 1-0, ...)
   - drop down select for hand, left, right, both
3. Count bar chart
   - bar chart to show total pitche by year
4. ~~R/S~~. Location plot, ~~outcome plot~~
   - ~~same~~ scatter plot over a frequency plot(low, alpha) to show pitch location, with some point linking to a gif of the pitch
5. Outcome plot
   - same outcome plot from repot, but with a baseball diamond showing where hits are located.

Discussion

- could get clutred and hard to manage space, -ve
- all in R, so less time consuming, +ve
- ~~hard~~ hard to explain to a user how to use

Shiny site



- ~~[redacted]~~ tab1 (Bio page)
  - picture of chris sale
  - a little bio about him, full name, age, college, Draft ~~order~~ no., etc
  - Stat summary table, ~~stats~~ highlight standout stats
  - maybe add radar chart to show some stats compared to ave.
- ~~[redacted]~~ tab 2 (time series)
  - time series, same options as in the dashboard, with some date markers for events of notes
- tab 3 (location)

subtab 1 - 2D: location plot same as in dashboard, with filter options one the left

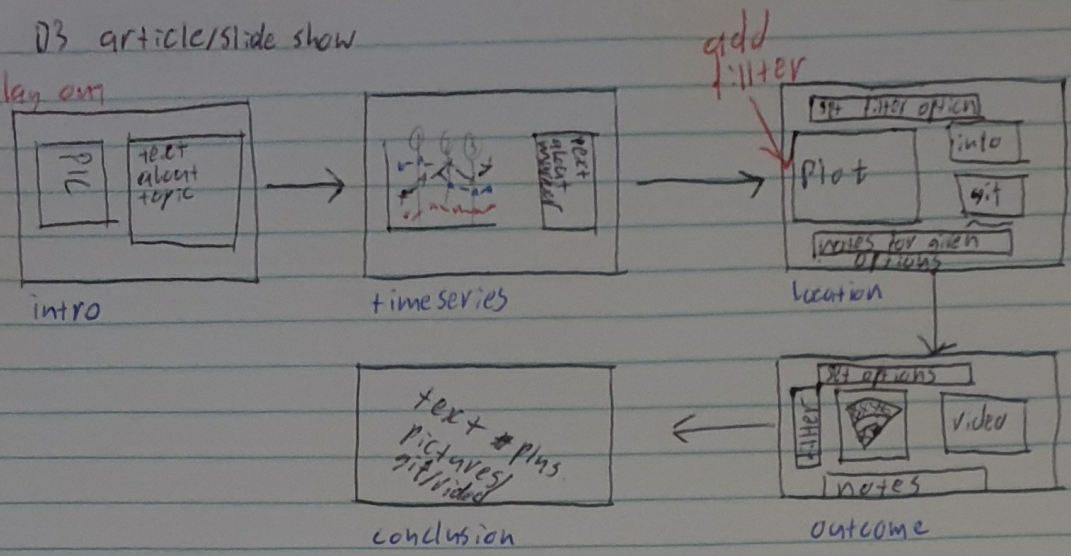
subtab 2 - 3D: a 3D version of the location plot

- tab 4 (outcome)
  - outcome plot, same as Dashboard, filter options on the left, like location

Discussion

- again all R, good for time, +ve
- hard to show one story, -ve
- filter options, might not be shared

D3 article/slide show



Intro
- would show and introduce who chris sale is
- intro the topic a bit about each slide and a quick link to each

timeseries
- the same ~~ter~~ timesers as the others
- when an event is clicked on explain the event and why it's important

location
- same location plot
- ~~when~~ filter settings to the left, also the same
- when a point is clicked on display info about it as well as a .gif of the pitch/pitch type
- Some predefined filter options, with notes on there importants

Outcome
- Similar set up to location, with filter, set options ~~and a video of~~
  - video of hit (prolaly just home runs)
- might put info if extra data can be sourced

Conclusion
- Summarise findings
- refrences
- glossary

Disscussion
- uses D3 so takes more time, -ve
- more info, +ve
- stronger narative, +ve
- better looking

- use plotly in R to create the time series and scatter plot, due to its slider, buttons to change view and link points to a url.
- outcome plot ggplot plus a package to map the shape of a baseball field, from MLBStadiums
- use shiny to host/create
- scrape summary table from MLB.com
- Glossary terms from, Baseball Reference
- Radar plot with ggplot
- makers info from MLB.com articles
- source any gif's from pitcherlist
- Viedos from MLB.com