SMT Data Challenge Notes

Overall Goal:

- We want to make a “story” of baseball based on the position data given
    - Could be something like focusing on 2 players
- How is “story” significant in the grand scheme of things? (i.e. how relevant/applicable is it to real world scenarios, and how important is the impact)
- How does our data analysis of the “story” impact that story?

Goal for 7/5 Meeting:

- Understand our objective / what to build
- Understand the dataset given (check metadata info)
    - Understand baseball terminology and concepts
    - https://www.youtube.com/watch?v=roWV4BFVlEg&ab_channel=SportsExplained
    - Watch Moneyball 2011 (movie, if interested)
    - Who’s on first: https://youtu.be/nZ5vspsNS1g (pretty funny, I get it now)
    - Learn R

Ideas:

- Stolen bases: interactive between catcher throwing ball to 2B or SS (DOING THIS)
    - **Whether it's important for the catcher to reduce time between catching and release or throwing the ball harder**
    - https://www.youtube.com/watch?v=xPLdza7DOfw
    - https://www.youtube.com/watch?v=Sorz3YzAsU4&ab_channel=TheOnDeckCircle (a good birds eye view visual of how a base is stolen)

- Success probability of lefty vs righty (OUT)
    - Double play (rightly more prone to double play?)
        - Ball hits ground

- Tagging up (if runner starts the run when the outfielder catches the ball) (OUT)
    - How outfielders can affect the tagging up (from 3rd to home plate)
    - https://www.youtube.com/watch?v=yuP5WGUjTa8
    - Very important to end game
- Task:
    - [Anonymous] - Learn more about baseball, and start cleaning data (filtering)

List of Possible Edge Cases:

- Ball bounces in front of home plate before catcher gathers the ball (good)
    - Ball bouncing is part of event sequence (pitch, bounce, 2-catch, 2-throw, 4-catch)
    - Separate from a bunt (pitch, hit, bounce, 2-catch, 2-throw, 4-catch)
- Ball could be recovered by by 2nd baseman or shortstop
    - Sequence ends in 4-catch or 6-catch

- Ball is overthrown into the outfield
    - Sequence would include being gathered by 8 (center fielder) and then most likely thrown to 5 (third baseman) after runner attempts to steal third base
- Pitch is thrown to backstop (not considered)
    - Ball will definitely bounce before gathered by catcher but he will gather it further from second base and attempt a longer throw
    - Pop up time is less as he will most likely pick up the ball with bare hand
- Ball bounces on its way to second base (possibly multiple times)
    - Sequence is (pitch, 2-catch, 2-throw, ball bounce, 4-catch)
- Stolen base attempt to 3rd base
    - Sequence (pitch, 2-catch, 2-throw, 5-catch)
- **Sequence we're looking for:** 1-Pitch -> possible bounce -> 2-catch -> 2-throw -> possible bounce -> 4 or 6 -catch
- Game events table (player position, event code): (1,1), (2,2), (2,3), (4,2), (0,5)
    - (1,1): pitcher pitches ball
    - (2,2): catcher catches the pitch
    - (2,3): catcher throws the ball to second base
    - (4,2): second base gets the ball from catcher
    - OR (6,2): where shortstop gets the ball from catcher
    - (0, 5): end of play
- What happens if 1B tries to steal 2B and at the same time when the opposing team is distracted getting 1B, then 3B steals to Home? How to account for this.

July 12th Meeting Notes:

- Once we filter all of the bases, we need to find the following
    - Distance and time between catcher throwing the ball and second baseman receiving the ball (find player_pos table to get distance of catcher and 2B or SS, then get the time difference between (2,3) and ((4,2) or (6,2)), finally calculate velocity) TODO: do for (6,2) event code and for all of the player_pos
    - Time between catcher catching the ball and releasing the ball (time difference between (2,2) and (2,3)) **[Anonymous] work on this**
    - Connect to other tables to determine if the runners were safe or out as a result of stolen base attempt **[Anonymous] working on this**
- For player_pos and ball_pos tables, group by game and play_id then combine with timestamp (**left_join(player_pos, by = c("game_str", "timestamp")**)
- We could possibly look into where the runner is located using player_pos table when the ball leaves the pitchers hand

Could do further analysis, where we can compare teams and perhaps one pitcher is better than the others (based on speed of ball throw, etc.)

July 26th Meeting Notes:

- Figuring out whether it is worth it for the catcher to throw the ball to beat the runner going to second base or to focus on another possible runner

    - We can do this by looking at the x y coordinates of the runner when the catcher catches the ball (2,2) and the distance he is from the base he is trying to get to
    - Check whether baserunner was successful in stealing 2B **(Part of [Anonymous]'s task)**
- Another thing: **[Anonymous] can work on this part**
    - How far a baserunner (individually) should be from the 1B when stealing 2B
        - Checking distance between 1B and baserunner when catcher catches ball and compare that with successful stolen base attempts
        - Scatter plot of stolen base attempts vs distance from 1B (success in green, failure in red)
- Ask **[Anonymous]**: **[Anonymous]**
    - Why player codes > 14
    - Why 0 0 0 for baserunners in game-info even though baserunner pos is moving in that stolen base play

August 2nd Meeting Notes:

- Graph 1 and 2 is main analysis (can do multiple different types of graphs to show these)
- Once ball is in SS or 2B hands, then its an independent variable outside of our project
- Since catcher doesn't affect once ball is in SS/2B hands, and player doesn't care either
- Combine SS and 2B, not compare the two (does not affect)
- **[Anonymous]** Notes:
    - the data aren't as good in the beginning
    - operator failed to mark someone on base
    - point out these issues in report
    - can assume that they made it based on the player position
    - How far is a lead off is safe ? minimum lead they can get away with
    - Who should the catcher throw to? 2nd baseman or shortstop
    - look at the individual catchers and see if someone is better or more successful?
    - look at player position for each "nothing" (0,0,0)
    - can send her rows that only have one pairing (or any abnormalities/issues)
- Why didn't we do other base attempts? Not enough data
- In edge cases where speed is slow (<20mps), remove since it doesn't affect our values

Response from [Anonymous] email about Report:

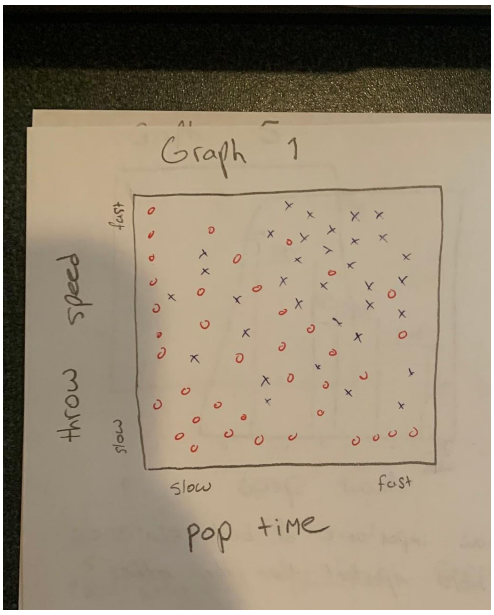
> "The paper should be a PDF, maximum 2000 words. That word count is the text itself, and does not include figure captions, references, appendices etc. There's no strict section format, though I discussed ways to organize things during the most recent Q&A (link below).

Current Analysis Ideas:

Graph 1: Scatterplot displaying each play's combination of pop time and throw speed

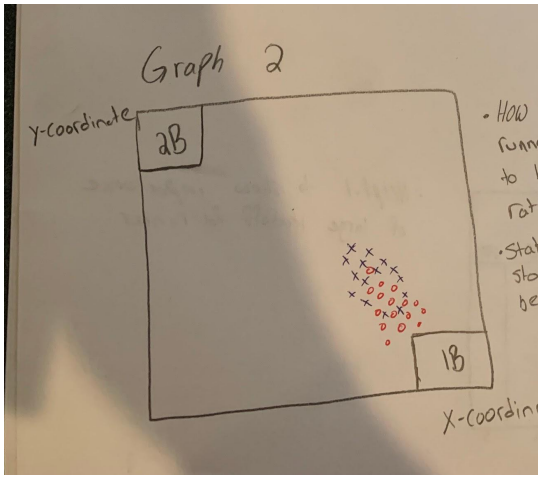
- Catcher successes as red O's, Catcher failures as blue X's
- What should the target ratio be for catchers? What ratio results in a 50% success rate for catchers (catcher avg should be around 33%)

- To achieve this desired ratio, should a catcher improve their pop time or throw speed (which is easier to change without significantly affecting the other)
- Is there a maximum pop time or minimum velocity that must be respected at all times in order to have a 10% success rate? What are they? (Find certain threshold)
- Possible box and whisker?



Graph 2: Scatterplot displaying x and y coordinates of runners in regard to first and second base

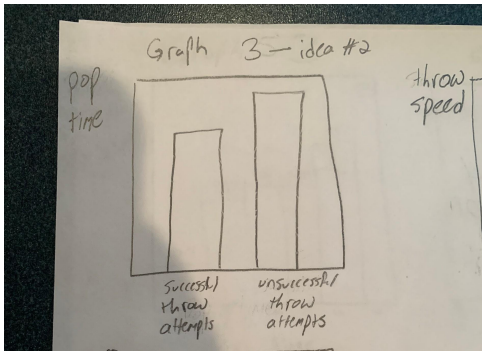
- Runner successes as blue X's and runner failures as red O's
- What leadoff distance does the runner need to have to have a 75% success rate? (runner avg is 67%)
- Is there a minimum leadoff distance for a runner to even consider running? (50% success rate) (a threshold to achieve a certain success rate)
- Are the x and y coordinates of any importance in relation to one another or is intuition correct in that you would prefer to be on the direct path to second base?

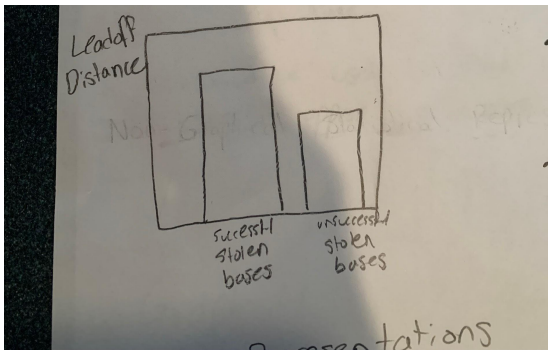


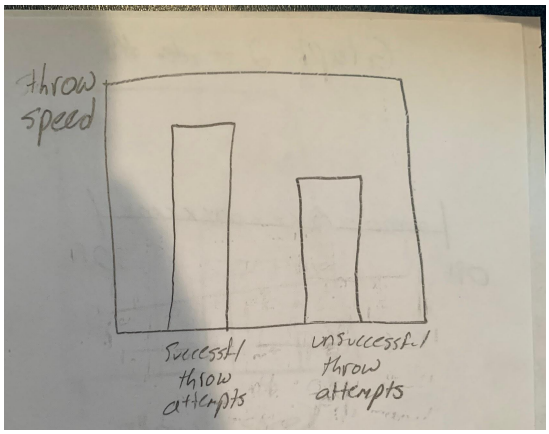
[insert 1D version] [could also use a heat map?]

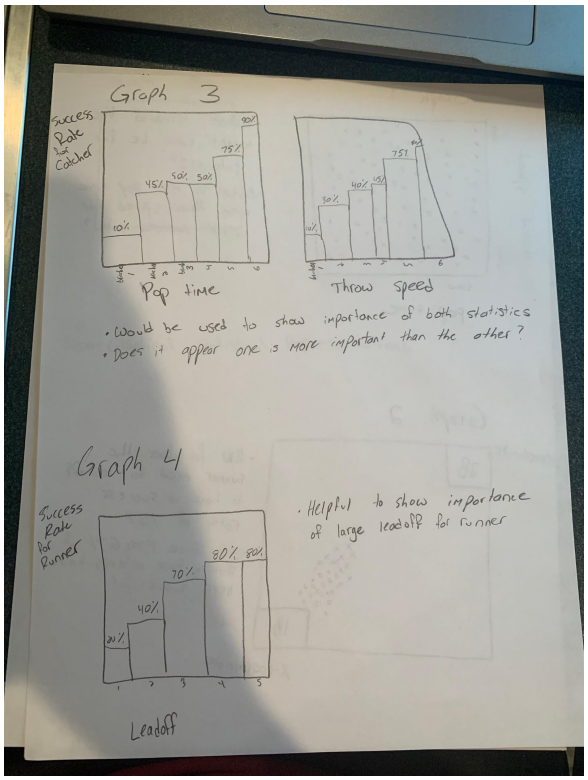
Graph 3: Group of 3 bar charts each displaying avg pop time, throw speed, and leadoff distance for successful and unsuccessful attempts

- Why are we looking at pop time, throw speed, and leadoff distance? What makes these statistics more important than others?
- Are these statistics really indicative of whether a runner will be safe or out?
- This graph will be used earlier on in the report to validate that our methodology is sound









August 9th Meeting Notes:
-Analysis Time!
-figured out what to do for report
-went over our work this week
-Divvy up Tasks:
-[Anonymous]: Report
-First Draft of Introduction
-Methodology: What variables to target, what statistics to calculate/use, and what tables to use
-Methodology: What play sequence to use and find

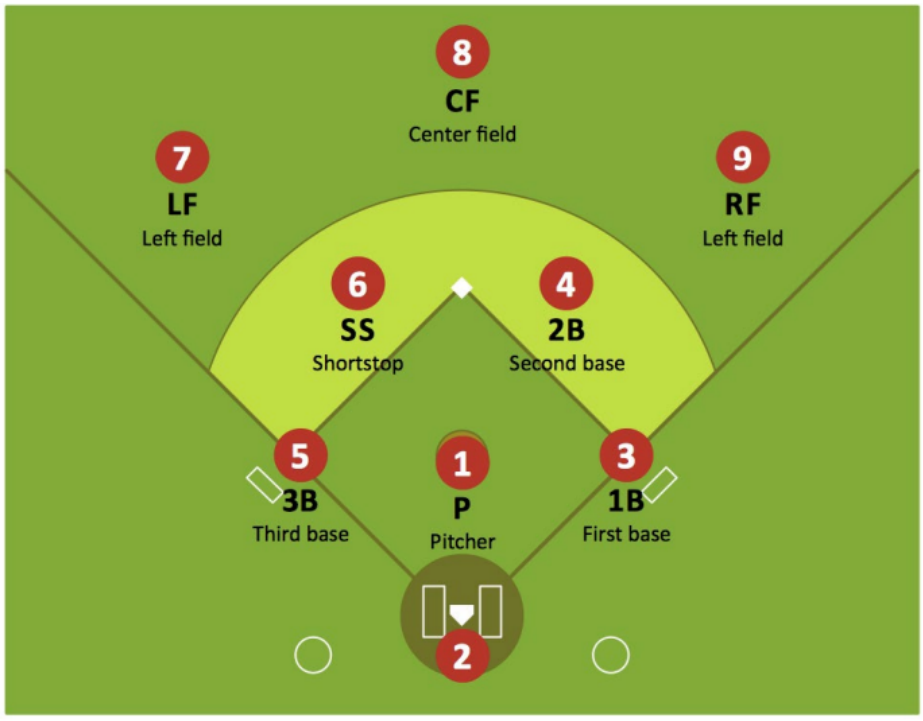
-[Anonymous]: Report
    -Methodology: What play sequence to use and find
    -Methodology: Things we considered and things we didn't consider/factor in and why (e.g. ball bouncing before reaching 2B or SS)
    -Analysis
        -Make graph 3
-[Anonymous]: Analysis
    -Finish filtering for lead off distance
    -Start making graphs 1 and 2 (graph 2 priority)
    -Report:
        -Thought process on what tech stack to use (R over Python)
        -How we filtered data ([Anonymous] - technical/implementation wise)
    -Code:
        -Cleaning up code, add comments, make repo
        -Upload notes? (Make it more navigable)
-[Anonymous]: Information / Report
    -Ask [Anonymous]: field dimensions (what is the coordinates of all the bases), is it a good idea to upload notes
    -Finish up filtering for stolen base successes/failures
        -[Anonymous] can help w/ this (maybe call next week)
    -Review / comment on current report work

Next week proposal: Meet for 2 hours from 7-9PM EST for work session
    -End of Wednesday Goal: First draft for Intro and Methodology, 60% graphs done, start Analysis part

-Better resolution to save graphs/plots

```
ggsave("test.tiff", runner_xy_plot, width=10, height=10, units="cm", dpi=700)
```

Reference:



1	pitch
2	ball acquired
3	throw (ball-in-play)
4	ball hit into play
5	end of play
6	pickoff throw
7	ball acquired - unknown field position
8	throw (ball-in-play) - unknown field position
9	ball deflection
10	ball deflection off of wall
11	home run
16	ball bounce

Dimensions (from [Anonymous] email):

- Home Plate (I'm assuming 18" rather than 17")
  - Back tip: (0, 0)
  - Front: (-0.75, 1.5) -> (0.75, 1.5)
  - Center: (0, 0.75)
- First Base (bases are 15" on a side)
  - Center: (62.76, 63.64) +/- 0.885
  - Right Corner: (63.64, 63.64)
  - Left Corner: (61.87, 63.64)
  - Upper Corner: (62.76, 64.53)
  - Lower Corner: (62.76, 62.86)
- Second Base (note that the center of second base lines up with the outer corners of first- and third base.)
  - Center: (0, 127.28) +/- 0.885
  - Right Corner: (0.89, 128.28)
  - Left Corner: (-0.89, 128.28)
  - Upper Corner: (0, 129.17)
  - Lower Corner: (0, 127.40)
- Third Base
  - Center: (-62.76, 63.64) +/- 0.885
  - Right Corner: (-61.87, 63.64)
  - Left Corner: (-63.64, 63.64)
  - Upper Corner: (-62.76, 64.53)
  - Lower Corner: (-62.76, 62.86)