GitHub Project Link (for Python Code):
https://github.com/laulpee/Soccer-Data-Analytics-Personal-Project

Google Sheets Link (for Spreadsheet Data & Analysis):
https://docs.google.com/spreadsheets/d/1FhX07fm0EIE_zFtxyGBLt_znADnvKLROqXku5KTBmBM/edit?usp=sharing

Soccer Data Analysis for Goals & Shooting Statistics

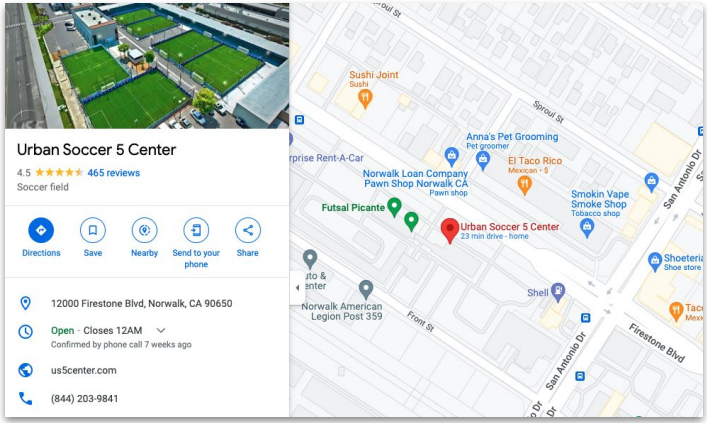
Personal Project
May 2022

Performed by: Paul Lee
Email: paul3pjl@gmail.com
LinkedIn: https://www.linkedin.com/in/paullee-1/

Table of Contents

Background

- In September 2021, my friends and I started a soccer group that plays every Tuesday at Urban Soccer (US) 5 Center in Norwalk, CA.
- Each game is 1.5 hours long with a 5 minute half time break. Typically, each game consists of 10-12 players, resulting in 5 or 6 players on each team.
- In December 2021, I discovered that the US 5 Center provides free game video footage (shown below) through their wide-angle cameras on each field. Ever since then, I have utilized the footage to collect data of the games' shot events.
- Initially, I collected data related to the scoreline, the players who scored and the players who assisted each goal. Currently, I am adding onto the data collection metrics by also collecting shot result types, shot location and body type used to take the shot by each player. The types of metrics I can collect continue to grow as the possibilities for in-game metrics are endless.
- Currently, with the 7 games that I have collected all goal & shot metrics data for, I processed, compiled, and analyzed the overall data using Google Sheets & Python to produce insightful data results detailed by the listed objectives below.
- Special thanks to US 5 Center for providing the field space and in-game video footage to make all of this possible.



Project Objectives

- Process and compile all in-game data from the games that I collected all goal & shot metrics for to produce the following analysis results:
  1. Shooting Heat Map of goals scored, shots attempted and goal percentage respective to the shot location on the field using Google Sheets and Python
  2. Pie Charts and Tables displaying shot results for each body type considered for shooting (left foot, rIght foot, head, overall) and comparing the specified player's goals, shots and goal percentage statistics with the group (league) average or game-wide average statistics using Google Sheets and Python
  3. Scatter Plot displaying every player's goal percentages for close-range and far-range goals using Python. This plot ultimately helps players better understand their preferred shooting spots and how clinical their shots are from close-range to far-range compared to other players in the league.
  4. Scatter Plot displaying every player's goals during the 1st-Half and 2nd-Half periods of the game using Python. This plot ultimately shows the goal impact that each player brings to each half of the game and which players have an earlier or later attacking impact on the games played.

Project Timeline

Data Collection

Data Compiling

Prelim. Analysis in Google Sheets

Analysis in Python

Data Visualization

Data Collection Overview

- The data collection process consisted of watching game video footage (6-7 15-minute video sections) provided by the US 5 Center venue for each 1.5-hour long game played each Tuesday. As the types of metrics are unlimited in the sport of soccer, only the current metrics of interest were recorded using Google Sheets.

- For the purposes of the project objectives, the following data metrics collected for this Project are shown in the table on the right (portion of the Collection Data Set for the Feb. 17th game) and the details for each metric are explained below:

  - **Date : Calendar Date of the Game**
  - **V# and Vid T : Video # and Video Time of the Game Video Footage**
    - Each 1.5-hour long game has 6-7 videos with a duration of 15 minutes per video.
  - **Player : Name of Player who took the Corresponding Shot**
  - **G : Goal Result**
    - Y = Goal
    - N = No Goal
  - **T : Result of the Shot**
    - G = Goal
    - Y = On Target (but No Goal)
    - B = Blocked (by a Player that isn't the Goalkeeper)
    - N = Off Target
  - **Z : Field Location of the Shot**
    - The field is discretized into 24 zones to determine the location of the shot. Refer to the next slide for further details.
  - **F : Body Type Used for Shot**
    - L = Left Foot
    - R = Right Foot
    - H = Head

Date	V#	Vid T	Player	G	T	Z	F
2/17/2022	1	0:11:30	Sung	N	N	10	R
2/17/2022	1	0:11:53	Stipe	N	B	14	L
2/17/2022	1	0:13:24	Stipe	N	Y	24	R
2/17/2022	1	0:13:30	Paul	N	Y	18	R
2/17/2022	1	0:13:53	Ryo	N	Y	19	R
2/17/2022	1	0:14:02	Joseph	N	Y	17	R
2/17/2022	1	0:14:04	Luis	Y	G	19	R
2/17/2022	1	0:14:39	Sung	N	B	18	L
2/17/2022	2	0:00:19	Sung	N	B	16	R
2/17/2022	2	0:00:36	Stipe	N	Y	18	R
2/17/2022	2	0:01:39	Stipe	N	N	14	R
2/17/2022	2	0:02:00	Luis	N	B	19	R
2/17/2022	2	0:02:44	Sung	N	Y	19	R

Data Collection

Determining the Field Location of a Shot

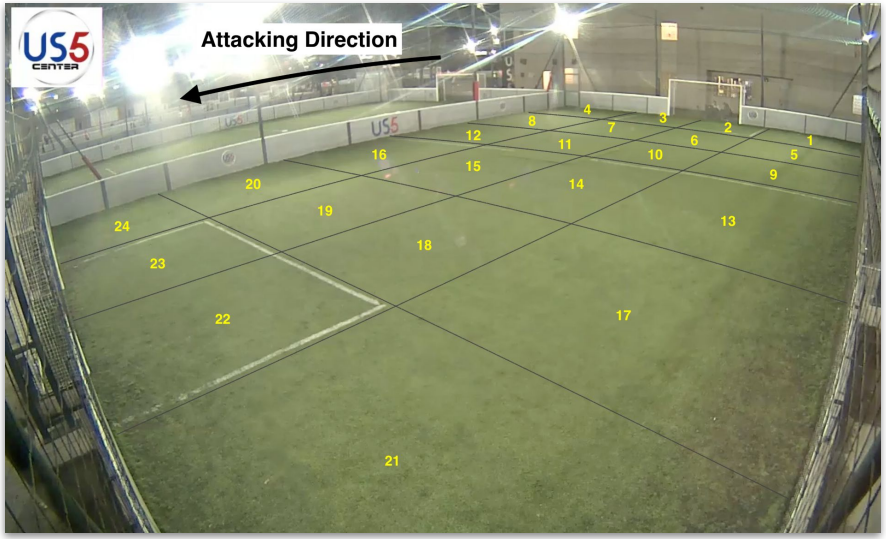
- The location of each shot was determined using the field location map shown below. The location map is also mapped onto an image of the actual field that the players play on.

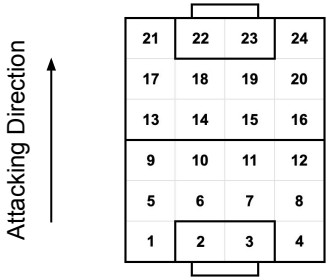
- Due to the lack of GPS monitoring systems, I discretized the field into 24 location zones to ultimately determine the location of each shot in the game.

- There are 3 important notes to mention:
    1. There is a level of inaccuracy for shots that are taken on the borders of location zones and also for shots that are taken on the other side of the field, which looks smaller as shown in the left image due to the wide-angle camera.

    2. Regardless of which area of the field each team plays on, the location map and zones are designed so that the higher-numbered zones are closer to the goal that the respective player is trying to score on. In simpler terms, field locations 1 - 12 consist of the corresponding player's own half while field locations 13 - 24 consist of the opposition's half where most of the corresponding player's shots will be taken.

    3. There is a rule established for each game, which states that a player can only shoot with their head in the small box (locations 22 & 23). Therefore shots in these zones are rare.





1 Data Collection
2 Data Compiling
3 Prelim. Analysis in Google Sheets
4 Analysis in Python
5 Data Visualization

Project Timeline

Data Collection

Data Compiling

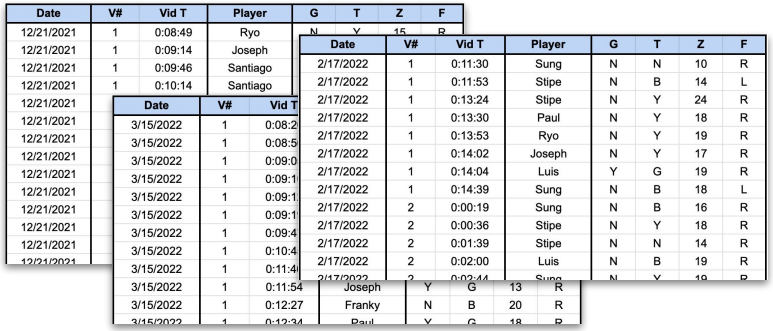
Prelim. Analysis in Google Sheets

Analysis in Python

Data Visualization

Data Compiling Overview

- As of now, I have collected full data sets for the metrics of interest for a total of 7 games. For this current edition of the Project, I have used the datasets for the 7 games to analyze goal & shooting metrics.
- Because the data sets for each game are collected in its own respective Google Sheets file, I compiled each data set into 1 main dataset so the analysis in Google Sheets and Python can be as efficient as possible, preventing having to call upon multiple Google Sheets files/tables.
- The steps and functions used in the Data Compiling process are as shown:



Date	V#	Vid T	Player	G	T	Z	F
12/21/2021	1	0:08:49	Ryo				
12/21/2021	1	0:09:14	Joseph				
12/21/2021	1	0:09:46	Santiago				
12/21/2021	1	0:10:14	Santiago				
2/17/2022	1	0:11:30	Sung	N	N	10	R
2/17/2022	1	0:11:53	Stipe	N	B	14	L
2/17/2022	1	0:13:24	Stipe	N	Y	24	R
2/17/2022	1	0:13:30	Paul	N	Y	18	R
2/17/2022	1	0:13:53	Ryo	N	Y	19	R
2/17/2022	1	0:14:02	Joseph	N	Y	17	R
2/17/2022	1	0:14:04	Luis	Y	G	19	R
2/17/2022	1	0:14:39	Sung	N	B	18	L
2/17/2022	2	0:00:19	Sung	N	B	16	R
2/17/2022	2	0:00:36	Stipe	N	Y	18	R
2/17/2022	2	0:01:39	Stipe	N	N	14	R
2/17/2022	2	0:02:00	Luis	N	B	19	R
3/15/2022	1	0:08:2					
3/15/2022	1	0:08:5					
3/15/2022	1	0:09:0					
3/15/2022	1	0:09:1					
3/15/2022	1	0:09:1					
3/15/2022	1	0:09:1					
3/15/2022	1	0:09:4					
3/15/2022	1	0:10:4					
3/15/2022	1	0:11:4					
3/15/2022	1	0:11:54	Joseph	Y	G	13	R
3/15/2022	1	0:12:27	Franky	N	B	20	R

Step 1: For the games that have complete collection datasets for shooting metrics, the link for each game's Google Sheets dataset file is added into the table in a separate Google Sheets file where the separate datasets will be added into 1 final dataset for analysis.

Note: As shown in table below, there are 7 total games with complete collection considered for analysis.

Game ID	Google Sheets Link	Date
13	https://docs.google.com/spreadsheets/d/1z0x5VqnNgb1jCjaH24GxgATFe0xC5mzOHgyoFxz5VzY/edit#gid=246451811	2021.12.21
18	https://docs.google.com/spreadsheets/d/18vilGzU6MbkYCXueMcT_0700c8izeHsWEocjGqdikw0/edit#gid=246451811	2022.02.01
19	https://docs.google.com/spreadsheets/d/11R_rkMTcSk7pNia6CTpFp0oV5JdKOKDBNuiunjoVCWo/edit#gid=246451811	2022.02.08
20	https://docs.google.com/spreadsheets/d/19gQUr2RcQd-fA5p_fwOB4ykrBtUwtwU_5W4nbc2VUN0/edit#gid=246451811	2022.02.15
21	https://docs.google.com/spreadsheets/d/1AhXmGDM_WR4_zAj2OtDi4U92WTmWPXKmWfyLBJXZ20w/edit#gid=246451811	2022.02.17
23	https://docs.google.com/spreadsheets/d/1SeU_JwxxNHKBgmvGDdMQPp5lj4HzgYWlE6a5-Lg7_2o/edit#gid=246451811	2022.03.01
25	https://docs.google.com/spreadsheets/d/1ZmNTQrS5SiLuucCo_1LZf__ydjAa0gbYx0F6J1arxlQ/edit#gid=246451811	2022.03.15

Step 2: The links table is used to import each game's shooting metrics dataset and compile all datasets using the IMPORTRANGE() function into the overall dataset shown above. The dataset contains 1365 rows, indicating that 1365 shot attempts over 7 games were analyzed in this project.

1	Date	V#	Vid T	Player	G	T	Z	F
2	2021.12.21	1	0:08:49	Ryo	N	Y	15	R
3	2021.12.21	1	0:09:14	Joseph	Y	G	18	R
4	2021.12.21	1	0:09:46	Santiago	Y	G	19	L
232	2022.02.01	2	0:04:04	Luis	N	Y	17	R
1029	2022.03.01	4	0:02:33	Santiago	N	N	14	L
1172	2022.03.15	2	0:06:50	Danny	N	B	20	R
1173	2022.03.15	2	0:07:05	Luis	Y	G	17	R
1365	2022.03.15	8	0:03:48	Ricky	Y	G	19	L

Step 3: With the compiled dataset, each column of data was cleaned and verified using the UNIQUE() function in Google Sheets.

For example, UNIQUE() was used for the Player column to verify that there weren't any mis-typed names inputted during data collection.

1 Data Collection
2 Data Compiling
3 Prelim. Analysis in Google Sheets
4 Analysis in Python
5 Data Visualization

Project Timeline

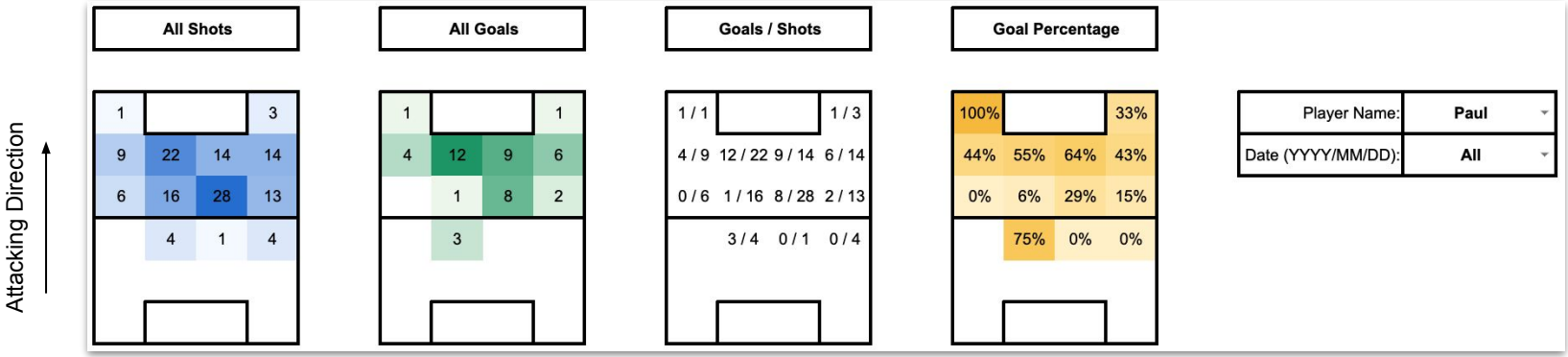
Data Collection

Data Compiling

Prelim. Analysis in Google Sheets

Analysis in Python

Data Insights & Conclusion



This visualization is from the "Prelim. Shooting Heat Map Display" sheet in the Google Sheets spreadsheet link on the title page.

Preliminary Analysis in Google Sheets

Shooting Heat Map Overview

The first preliminary analysis was performed using Google Sheets. With the main dataset of shooting metrics now available for analysis, the first analysis was performed to create a simple dashboard that displays 4 heat maps related to shooting metrics of the selected player and game date. If "All" is selected for Date, then shooting statistics for every game played by the selected player are shown.

The four heat maps display the following: shot attempts, goals, goals over shots ratio and goal percentage for every shot that was taken in each of the 24 discretized field locations.

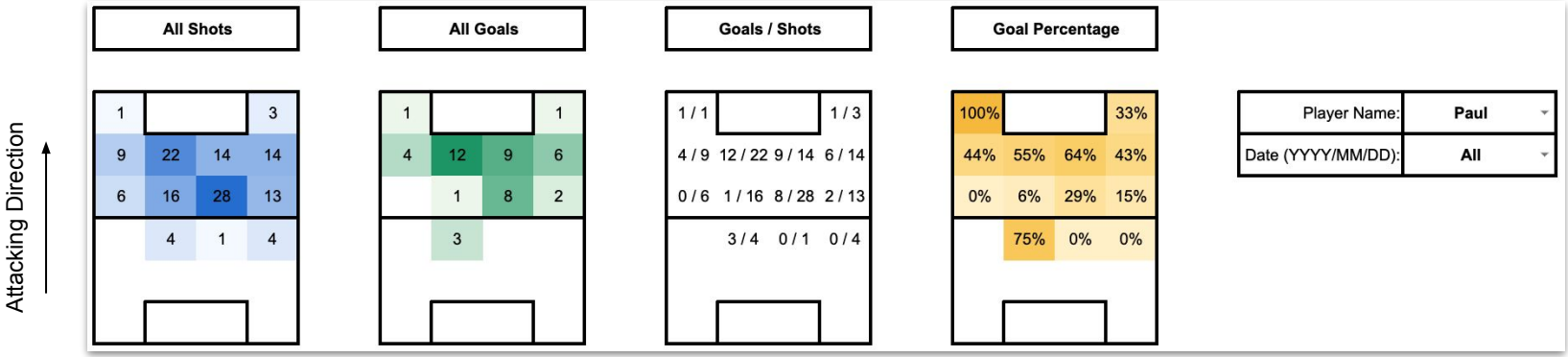
Step 1: Extracted a distinct list from main dataset of all the players that have played and the dates of the 7 games played by using the UNIQUE() and FILTER() function.

Step 2: For "All Shots" and "All Goals" heat maps, each cell used the QUERY() function to query the Goal Result (G) column of the dataset with the following criteria:

- The corresponding field location zone # based on the map in slide 6
- Selected player name
- Selected game date (or All)
- For "All Goals", the shots with "Y" in the Goal Result column were queried

Step 3: For the "Goals/Shots" heat map, the CONCATENATE() function was used to combine the Goals & Shot Attempts for each location into 1 string for each zone.

Step 4: For the "Goal Percentage" heat map, simple division of the Goals by Shot Attempts was performed to determine the goal % from each location.



This visualization is from the "Prelim. Shooting Heat Map Display" sheet in the Google Sheets spreadsheet link on the title page.

Preliminary Analysis in Google Sheets

Shooting Heat Map: Advantages and Disadvantages of Heat Map Analysis in Google Sheets

Advantages:

- Because the heat map dashboard was created in the same spreadsheet as the main dataset, performing the heat map analysis in Google Sheets did not require a lot of time and any type of coding knowledge. As explained in the previous slide, only a couple of built-in Google Sheets functions in addition to conditional formatting was needed to perform the analysis and visualization.
- The short amount of time required to perform the analysis and create the dashboard can provide an easily-accessible, updated heat map dashboard to players who want to access any player's shooting stats for any selected game.

Disadvantages:

- Although 1365 rows of data may be easily handled by the spreadsheet for the analysis, the process time for updating the dashboard when the selected player or game date is changed by the user takes longer than a second. Once more data is collected and the main dataset grows, the process time will get slower and the dashboard will become less user-friendly.
- The dashboard can be difficult to understand as there are four heat maps with different color spectrums for each visualization. For example, the top-left location zone displays a 100% goal percentage, which may show that I am efficient and clinical from that location. However, when looking at the "All Shots" heat map, I took only 1 shot in that location and happened to score. Although it could've been a well executed goal, the 100% goal percentage is unreliable compared to goal percentages where a lot of shots were attempted.

To address these disadvantages, a heat map analysis was performed using Python, which will be explained later in this document.

Date (or All): All	Left Foot			Right Foot			Head			Overall		
	Goals	Attempts	Goal %	Goals	Attempts	Goal %	Goals	Attempts	Goal %	Goals	Attempts	Goal %
Ryo	1	4	25.00%	26	100	26.00%	0	2	0.00%	27	106	25.47%
Joseph	0	2	0.00%	21	161	13.04%	2	6	33.33%	23	169	13.61%
Santiago	52	132	39.39%	6	12	50.00%	0	1	0.00%	58	145	40.00%
Ozzie	0	5	0.00%	0	0		0	0		0	5	0.00%
Ricky	20	54	37.04%	11	28	39.29%	0	1	0.00%	31	83	37.35%
Danny	3	12	25.00%	26	152	17.11%	1	1	100.00%	30	165	18.18%
Appa	1	8	12.50%	11	70	15.71%	0	0		12	78	15.38%
KyungSoo	0	3	0.00%	1	20	5.00%	0	0		1	23	4.35%
Paul	3	15	20.00%	44	120	36.67%	0	0		47	135	34.81%
Fredy	0	0		2	4	50.00%	0	0		2	4	50.00%
KChung	0	3	0.00%	4	35	11.43%	0	0		4	38	10.53%
Daniel	0	0		1	13	7.69%	0	0		1	13	7.69%
Mikey	1	3	33.33%	3	22	13.64%	0	0		4	25	16.00%
Luis	18	68	26.47%	16	72	22.22%	1	1	100.00%	35	141	24.82%
Isaac	0	0		0	6	0.00%	0	0		0	6	0.00%
Sung	3	8	37.50%	12	46	26.09%	1	2	50.00%	16	56	28.57%
David	0	0		0	1	0.00%	0	0		0	1	0.00%
Stipe	1	5	20.00%	9	30	30.00%	0	0		10	35	28.57%
Chris	6	29	20.69%	2	11	18.18%	0	0		8	40	20.00%
Franky	3	14	21.43%	2	36	5.56%	0	0		5	50	10.00%
KCamal	1	5	20.00%	5	12	41.67%	0	0		6	17	35.29%
Oscar	1	7	14.29%	5	21	23.81%	0	1	0.00%	6	29	20.69%
League Total:	114	377		207	972		5	15		326	1364	
League AVG:	5.18	17.14	19.59%	9.41	44.18	21.58%	0.23	0.68	35.42%	14.82	62.00	20.06%

Table: League-wide Shooting Statistics for Every Player on Selected Game Date (or All Games)

This visualization is from the "Shooting Statistics Table" sheet in the Google Sheets spreadsheet link on the title page.

Player:	Danny	Shot Results						Shot Result Percentages				
Date (or All):	2022.02.08	Goal	On Target	Blocked	Off Target	Total		Goal	On Target	Blocked	Off Target	Total
		G	Y	B	N			G	Y	B	N	
Left Foot	L	0	0	0	1	1		0.00%	0.00%	0.00%	100.00%	100.00%
Right Foot	R	9	8	3	5	25		36.00%	32.00%	12.00%	20.00%	100.00%
Head	H	0	0	0	0	0		#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
	Total	9	8	3	6	26						
	Percentages	34.62%	30.77%	11.54%	23.08%	100.00%						

Table: Detailed Shooting Statistics for Selected Player on Selected Game Date (or All Games)

This visualization is from the "Shooting Statistics Table" sheet in the Google Sheets spreadsheet link on the title page.

Preliminary Analysis in Google Sheets

Shooting Statistics Tables

In addition to the Shooting Heat Map dashboard, shooting statistics analysis was performed using the main dataset in Google Sheets. It's important to note that the tables can display statistics for all games or a specific game selected by the user. For the bottom-left table, a player can be selected by the user to display the respective player's statistics. The 2 tables are explained in detail below:

- **League-Wide Shooting Statistics Table:**
  - As shown in the top-left table, shooting statistics such as goals, attempts and goal % for each body type used for shooting (left foot, right foot, head) were calculated for every player that played any of the 7 games covered in this project.
  - Aside from mathematical functions, the main function used to perform this analysis was the COUNTIFS() function in Google Sheets, in which different criteria based on body type, player name, game date, goal result, and shot attempts were used to extract the appropriate data from the dataset.

- **Detailed Shooting Statistics Table for Specific Player:**
  - As shown in the bottom-left table, more detailed shooting statistics with a focus on shot results were calculated for each player. The result of each shot taken by the selected player is categorized with the body type used.
  - Like the analysis for the top-left table, the COUNTIFS() function was used to perform the analysis, in which criteria specific to the shot result type and body type were used to extract the appropriate data from the dataset.

- Although the tables provide a detailed perspective of each player's shooting performances, the tables are statistics-heavy and can be less user-friendly to players who are interested in straight-forward data visualizations. Therefore, the same analysis was performed in Python to transform these data results into a dashboard consisted of pie-charts and smaller tables.

Project Timeline

Analysis in Python

Shooting Heat Map - Analysis Steps

As mentioned in slide 11, a Shooting Heat Map analysis was performed using Python to address the disadvantages of the Heat Map visualization in Google Sheets and to ultimately create a Shooting Heat Map that is more insightful and easier to understand.

As shown in the scatter plot on the right, the Shooting Heat Map designed in Python is a scatter plot with hidden axes that represents the field and its discretized 24 location zones. The upward arrow indicates the attacking direction of play. A more detailed explanation of the Heat Map analysis result/visualization will be on the next slide.

Like the Heat Map dashboard in Google Sheets, the Heat Map visualization was coded in Python to take the Player Name and Game Date (or All Games) inputs from the user. With this feature, a shooting heat map for any player for any game can be displayed.

The main analysis steps in the Python code are summarized in this slide. For the full, detailed Python code, please refer to the GitHub link below to find the "shotHeatMap_statsDashboard.py" code used to create the Shooting Heat Map visualization.

GitHub Link for Project:

https://github.com/laulpee/Soccer-Data-Analytics-Personal-Project

Corresponding Python Code File:

shotHeatMap_statsDashboard.py

(user defined function: heatmap)

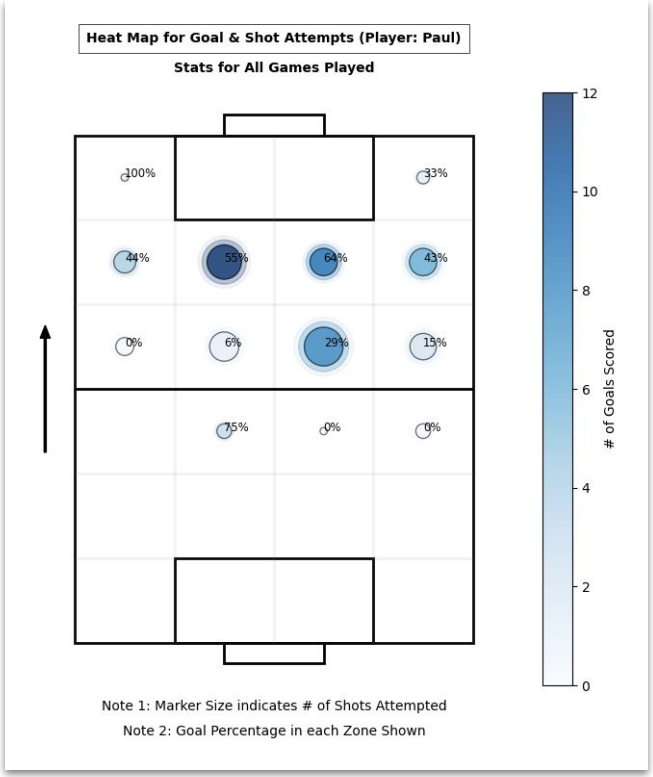
Step 1: Import Python libraries (pandas, matplotlib, numpy, math)

Step 2: Read CSV file of main dataset exported from Google Sheets, and CSV file with (x,y) coordinates of field location zones

Step 3: Ask user to select Player and Game Date (or All Games)

Step 4: Perform analysis using the query and count functions to calculate total goals, shot attempts and goal percentage for each field location zone of selected player and game

Step 5: Plot scatter plot with field lines to represent the field and plot gray grid to represent the 24 field location zones



Analysis in Python

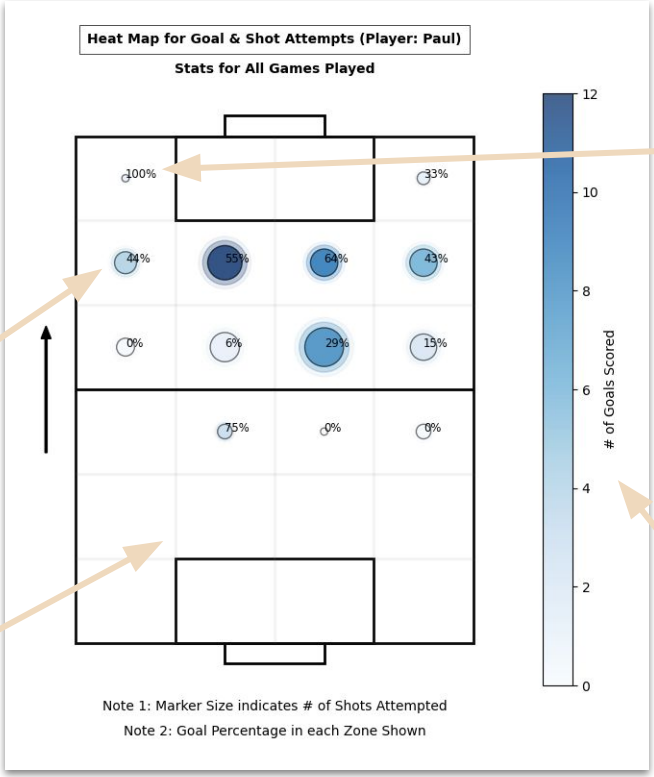
Shooting Heat Map - Visualization

GitHub Link for Project:
https://github.com/laulpee/Soccer-Data-Analytics-Personal-Project

Corresponding Python Code File:
shotHeatMap_statsDashboard.py
(user defined function: heatmap)

Marker Size:
As stated in Note 1, the marker size is displaying the # of shots attempted in each location. With this heat map visualization, I can conclude that majority of my shots are attempted more in the middle of the field and less in the wide areas.

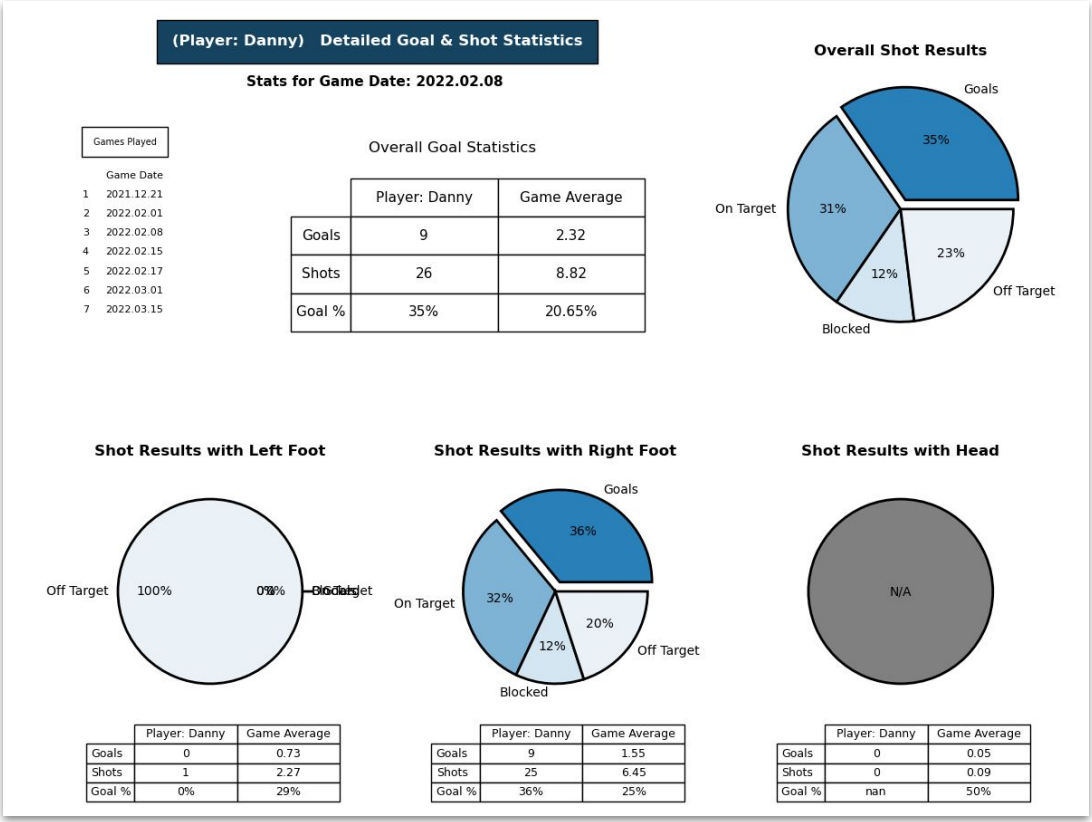
Grid Lines:
The gray grid lines are plotted to show the 24 field location zones which were established in the data collection process due to the lack of precise GPS tracking data. The location zones were essential in creating this Heat Map visualization.



Goal Percentage:
The goal percentage at each location that a shot has been attempted is labelled on the marker. Combined with the marker size and color, this eliminates the potential misunderstanding mentioned in slide 11, in which a 100% goal percentage doesn't necessarily represent a positive result. As shown here, the 100% goal percentage in this location is paired with a small, light-colored marker, which indicates that the high goal percentage is not reliable compared to the locations where more shots were attempted.

Marker Color:
As stated in the color bar legend, the color of each marker represents the # of goals scored in the respective location. The locations with a darker color indicate that more goals were scored in that respective location. As shown on the Heat Map, I scored the most goals outside of the penalty box.

1 Data Collection
2 Data Compiling
3 Prelim. Analysis in Google Sheets
4 Analysis in Python
5 Data Visualization



Analysis in Python

Shooting Statistics Pie Charts & Tables Overview

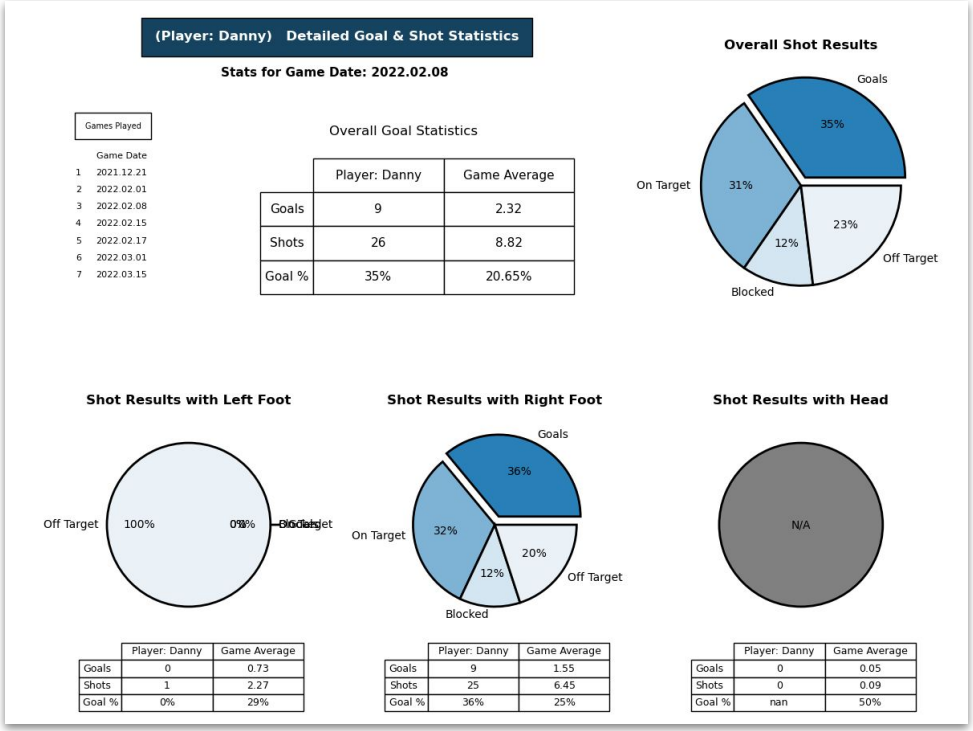
As mentioned in slide 12, data analysis was performed using Python to create a dashboard with pie charts & tables that displays detailed shooting statistics for better readability and understanding than tables with numbers only as shown in the tables in Google Sheets. Because the analysis is the same as the analysis performed in Google Sheets, the shooting statistics in the dashboard can be confirmed and verified with the statistics table created using Google Sheets.

With the pie charts supported by the small tables in the dashboard, the shooting statistics for goal, shot attempts and goal percentage for each body type (or overall) are clearly portrayed for the user to make insightful conclusions of the selected player's shooting performance.

The analysis steps and features of the dashboard visualization will be discussed in the following slides.

GitHub Link for Project:
https://github.com/laulpee/Soccer-Data-Analytics-Personal-Project

Corresponding Python Code File:
shotHeatMap_statsDashboard.py
(user defined functions: goalLeagueGameAverages, goalPlayerStats, pieChart)



The main analysis steps in the Python code are summarized in this slide. For the full, detailed Python code, please refer to the GitHub link to find the “shotHeatMap_statsDashboard.py” code used to create the Shooting Statistics dashboard visualization.

Analysis in Python

Shooting Statistics Pie Charts & Tables Analysis Steps

GitHub Link for Project:
https://github.com/laulpee/Soccer-Data-Analytics-Personal-Project

Step 1: Import Python libraries (pandas, matplotlib, numpy, math)

Step 2: Read CSV file of main dataset in Google Sheets

Step 3: Ask user to select Player and Game Date (or All Games)

Step 4: Perform analysis using the query, count, sum and mean functions to calculate total goals, shot attempts and goal percentage for each body type for ENTIRE league or for ENTIRE selected game. (user defined function: goalLeagueGameAverages)

Step 5: Perform same analysis as step 4 but for selected player. (user defined function: goalPlayerStats)

Step 6: Compile League-wide (or game-wide) shooting statistics with the selected player’s statistics for each shot body type into matrices for table display purposes. (user defined function: pieChart)

Step 7: Create a figure with subplots for the dashboard containing pie charts and tables for shooting statistics involving the left foot, right foot, head and overall. (user defined function: pieChart)

1 Data Collection
2 Data Compiling
3 Prelim. Analysis in Google Sheets
4 Analysis in Python
5 Data Visualization

Analysis in Python

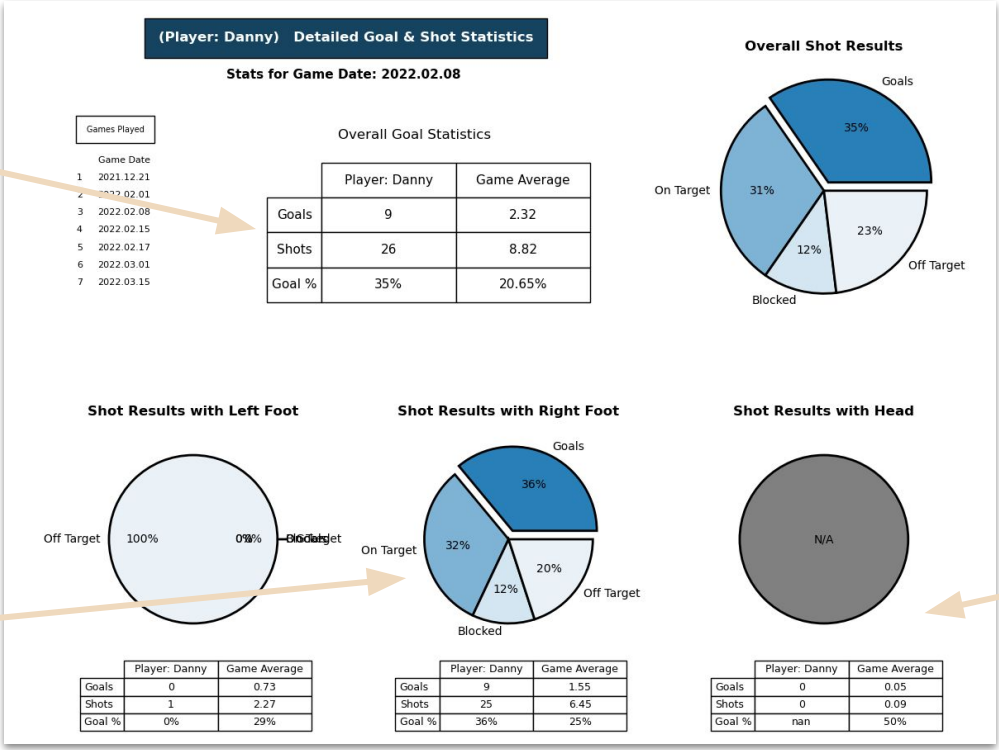
Shooting Statistics Pie Charts & Tables Visualization

Statistics Tables:

For each shot body type and overall statistics, a table displays the player's goals, shots and goal percentage metrics compared to the game average (or league-wide average if all games are selected by the user) shooting statistics.

Pie Chart Breakdown:

For each shot body type and overall statistics, a pie chart shows the shot result types (goal, on target, blocked and off target) and its percentages. It is important to note that the darker colored areas indicate the better result of the shot. In simpler terms, the best-to-worst rank of shot results are in the following: goals, on target, blocked (by a player other than the goalkeeper) and off target.



Statistics for 0 Attempted Shots:

For cases where a shot was not attempted overall or by one of the body types, the pie chart is displayed with a gray filler with N/A text. It is important to note that the goal percentage for cases like this are not set to 0% because 0 shots were attempted. Therefore, the goal percentage was displayed as the non-value nan in Python.

GitHub Link for Project:

https://github.com/laulpee/Soccer-Data-Analytics-Personal-Project

Corresponding Python Code File:

shotHeatMap_statsDashboard.py

(user defined functions: goalLeagueGameAverages, goalPlayerStats, pieChart)

The shooting statistics displayed in the dashboard are verified to be the same as the statistics calculated in Google Sheets, shown in slide 12.

1 Data Collection
2 Data Compiling
3 Prelim. Analysis in Google Sheets
4 Analysis in Python
5 Data Visualization

Analysis in Python

Close-Range vs Far-Range Goal % Case Study for League

Overview

After analysis was performed for each player's shooting statistics, an analysis was performed for a league-wide case study on players' goal percentages for shots taken from close-range and far-range. As shown in the field location map on the right, close-range is defined as shots attempted in locations 17 - 24 while far-range is defined as shots attempted in locations 1 - 16.

Through this analysis of the main dataset, a scatter plot of every player's goal % from the 2 range criteria is created. With this visualization, insightful conclusions can be made to better understand players' preferred shooting spots and how clinical each player's shooting performance is from close-range and far-range compared to other players or the leagues average.

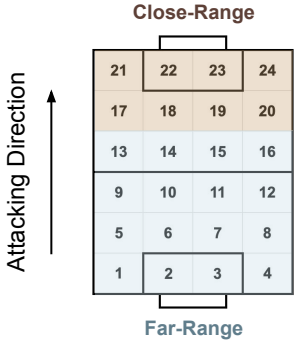
Important note: as shown in the scatter plot on the right, names of every player who has played in the 7 games covered in this project are not displayed in the plot. Because there are players with low number of shot attempts, only the players with shot attempts over the 25th percentile of the league's shot attempts taken are considered for displaying on the plot.

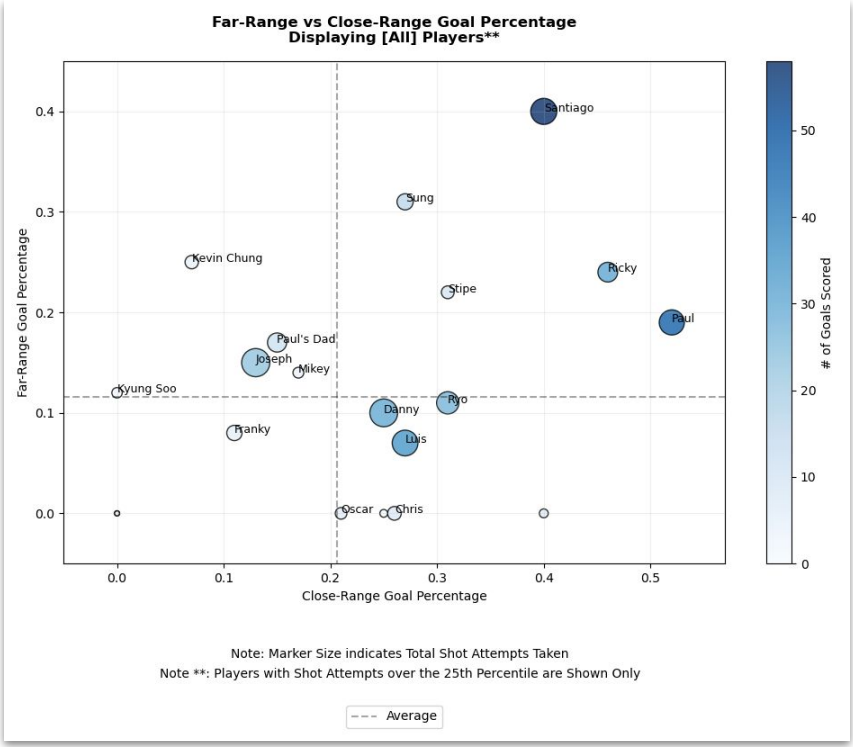
GitHub Link for Project:

https://github.com/laulpee/Soccer-Data-Analytics-Personal-Project

Corresponding Python Code File:

closeGoal_farGoal.py





1 Data Collection
2 Data Compiling
3 Prelim. Analysis in Google Sheets
4 Analysis in Python
5 Data Visualization

Analysis in Python

Close-Range vs Far-Range Goal % Case Study for League

Analysis Steps

The main analysis steps in the Python code are summarized in this slide. For the full, detailed Python code, please refer to the GitHub link to find the “closeGoal_farGoal.py” code used to create the scatter plot visualization.

GitHub Link for Project:

https://github.com/laulpee/Soccer-Data-Analytics-Personal-Project

Step 1: Import Python libraries (pandas, matplotlib, numpy, math)

Step 2: Ask user to select which player names to display on scatter plot (All, Top #, or Bottom #)

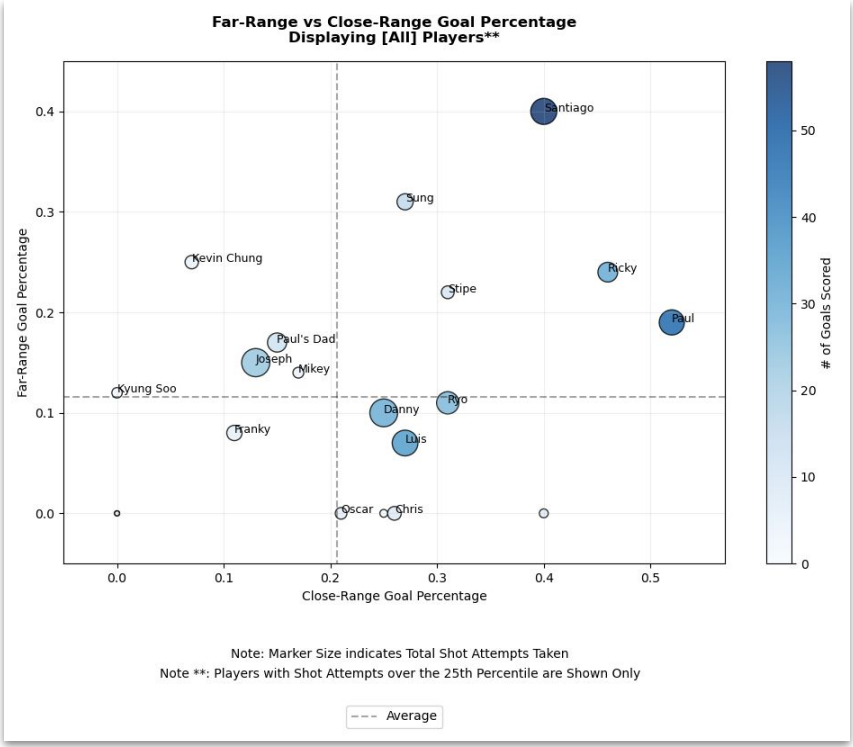
Step 3: Read CSV file of main dataset in Google Sheets

Step 4: Perform analysis using the sum function to calculate goals, shot attempts, and goal % for each player from close-range and far-range.

Step 5: Calculate total goal % L2-norm of each player’s close-range and far-range goal % using numpy’s sqrt and square functions. The norm is later used to determine the rankings of each player’s normalized goal %.

Step 6: Calculate the 25th percentile of shot attempts taken within the entire league. This will exclude players who took less shots than the 25th percentile.

Step 7: Plot scatter plot with close range goal % on the x-axis and far-range goal % on the y-axis. Plot lines indicating the league average for each axis.



1 Data Collection
2 Data Compiling
3 Prelim. Analysis in Google Sheets
4 Analysis in Python
5 Data Visualization

Analysis in Python

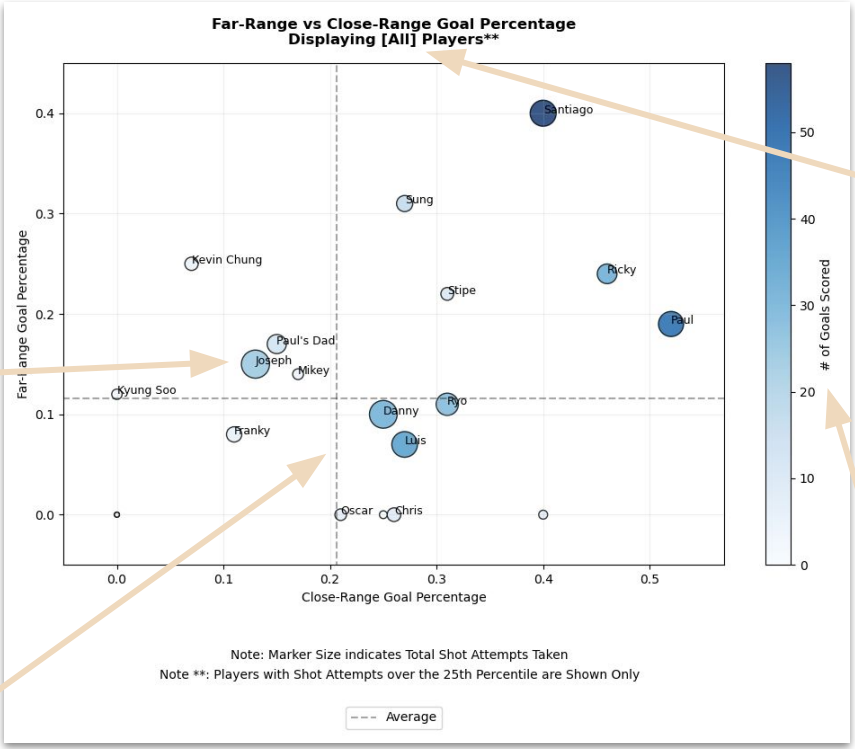
Close-Range vs Far-Range Goal % Case Study for League Visualization

GitHub Link for Project:
https://github.com/laulpee/Soccer-Data-Analytics-Personal-Project

Corresponding Python Code File:
closeGoal_farGoal.py

Marker Size:
As stated in the first note, the marker size is displaying the total # of shots attempted. The larger markers indicate that the respective player took more shot attempts then a player with a smaller marker.

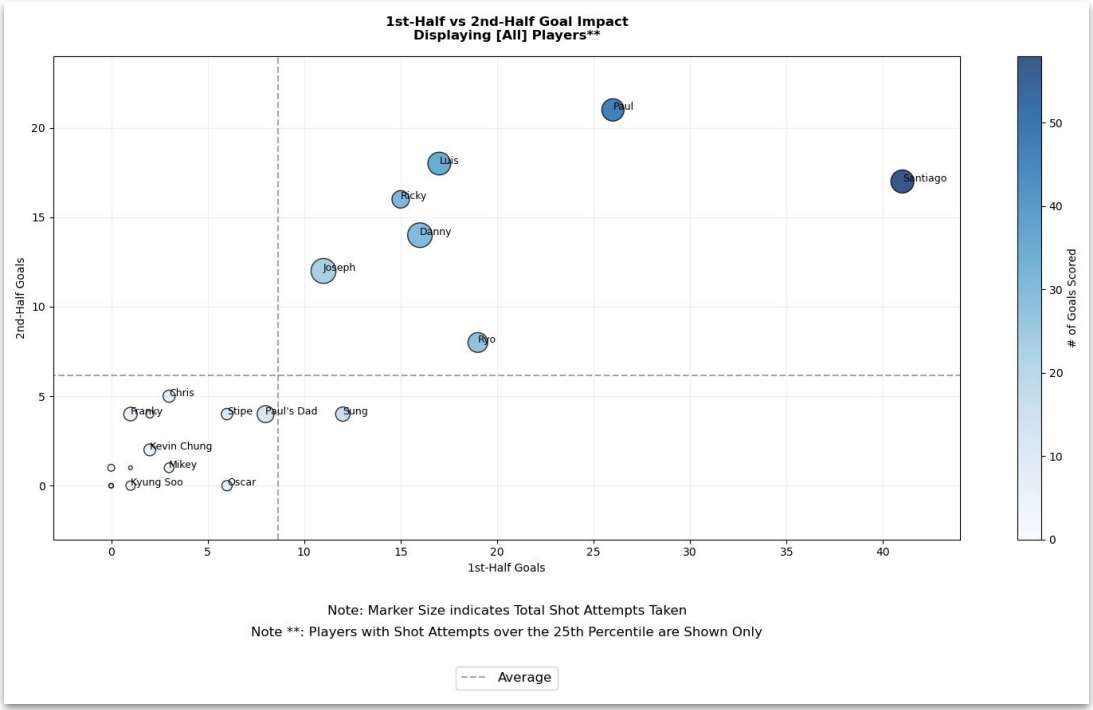
League Average:
The gray dashed lines are plotted to show the league average for close-range and far-range goal %. This feature can help make conclusions on which players are underperforming or outperforming in the league.



Rankings Using L2-Norms:
The Python code allows the user to choose to display all player names, top ranked player names or bottom ranked player names on the scatter plot based on the L2-norm of each player's close-range and far-range %. In the scatter plot shown, it displays all players names; however, if "Top 1" is inputted into the code, the top player with the highest goal % norm is Santiago.

Marker Color:
As stated in the color bar legend, the color of each marker represents the # of goals scored by the corresponding player. It's interesting to note that most players with a higher # of goals scored are performing higher than the league average in close-range goal %.

1 Data Collection
2 Data Compiling
3 Prelim. Analysis in Google Sheets
4 Analysis in Python
5 Data Visualization



1st-Half vs 2nd-Half Goal Impact Case Study for League

Overview

In addition to the close-range vs far-range goal % case study, an analysis for a league-wide case study on players' goal impact in the first half and second half of each game was performed.

Through this analysis of the main dataset, a scatter plot of every player's goal count from each half of a game is created. With this visualization, insightful conclusions can be made to better understand the goal impact that each player brings to each half of the game and which players have an earlier or later impact in the games played.

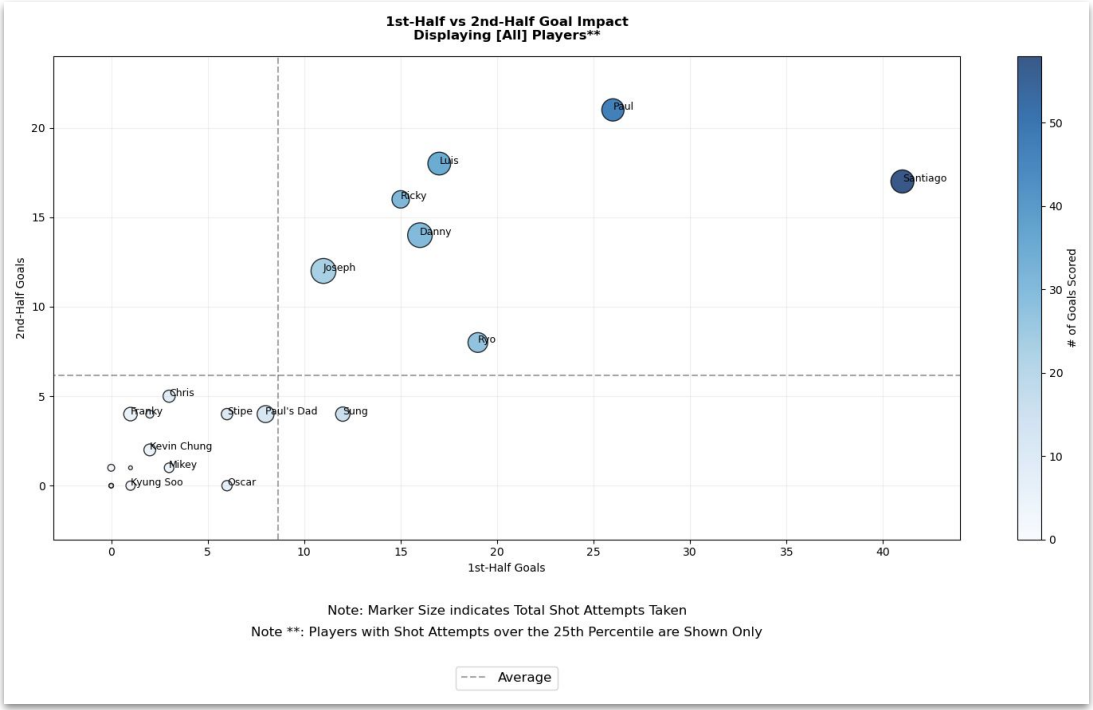
Important note: as shown in the scatter plot on the left, names of every player who has played in the 7 games covered in this project are not displayed in the plot. Because there are players with low number of shot attempts, only the players with shot attempts over the 25th percentile of the league's shot attempts taken are considered for displaying on the plot.

GitHub Link for Project:
https://github.com/laulpee/Soccer-Data-Analytics-Personal-Project

Corresponding Python Code File:
firstHalfGoals_2ndHalfGoals.py

1st-Half vs 2nd-Half Goal Impact Case Study for League

Analysis Steps



The main analysis steps in the Python code are summarized in this slide. For the full, detailed Python code, please refer to the GitHub link to find the "firstHalfGoals_2ndHalfGoals.py" code used to create the scatter plot.

GitHub Link for Project:
https://github.com/laulpee/Soccer-Data-Analytics-Personal-Project

Step 1: Import Python libraries (pandas, matplotlib, numpy, math)

Step 2: Ask user to select which player names to display on scatter plot (All, Top #, or Bottom #)

Step 3: Read CSV file of main dataset in Google Sheets

Step 4: Calculate average of every video section # in dataset to determine the half time mark relative to the video section # of the game footage.

Step 5: Perform analysis using the sum function to calculate goals, shot attempts, and goal % for each player in the 1st-half and 2nd-half.

Step 6: Determine rankings by calculating total goal L2-norm of each player's 1st-Half and 2nd-Half goals using numpy's sqrt and square functions.

Step 7: Calculate the 25th percentile of shot attempts taken within the entire league. This will exclude players who took less shots than the 25th percentile of the league.

Step 8: Plot scatter plot and plot lines indicating the league average for each axis.

Analysis in Python

1st-Half vs 2nd-Half Goal Impact Case Study for League Visualization

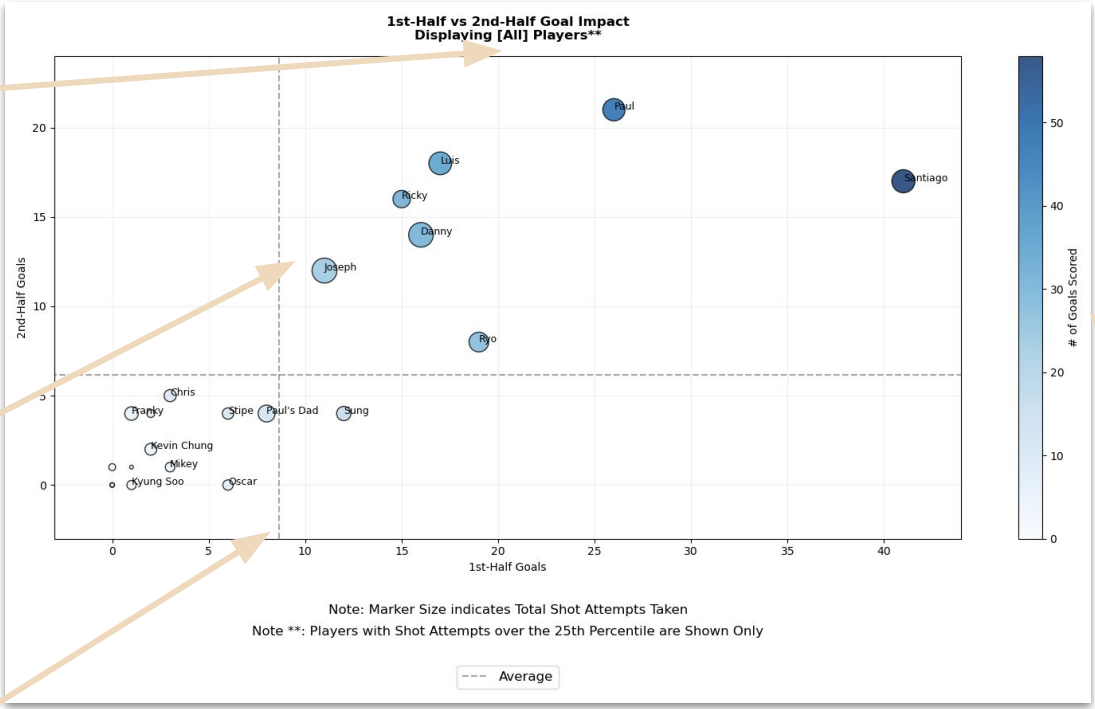
GitHub Link for Project:
https://github.com/laulpee/Soccer-Data-Analytics-Personal-Project

Corresponding Python Code File:
firstHalfGoals_2ndHalfGoals.py

Rankings Using L2-Norms:
The Python code allows the user to choose to display all player names, top ranked player names or bottom ranked player names on the scatter plot based on the L2-norm of each player's 1st-Half and 2nd-Half goals. In the scatter plot shown, it displays all player names; however, if "Top 1" is inputted into the code, the top player with the highest goal count norm is Santiago.

Marker Size:
As stated in the first note, the marker size is displaying the # of shots attempted. The larger markers indicate that the respective player took more shot attempts then a player with a smaller marker.

League Average:
The gray dashed lines are plotted to show the league average for 1st-Half and 2nd-Half goals. This feature can help make conclusions on which players are underperforming or outperforming in the league.



Marker Color:
As stated in the color bar legend, the color of each marker represents the # of goals scored by the corresponding player.

1 Data Collection
2 Data Compiling
3 Prelim. Analysis in Google Sheets
4 Analysis in Python
5 Data Visualization

Project Timeline

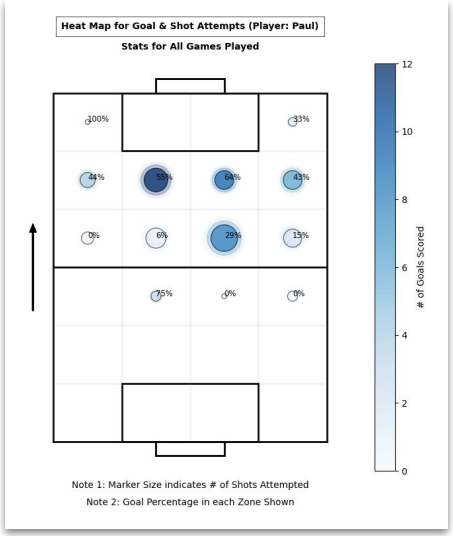
Data Visualization

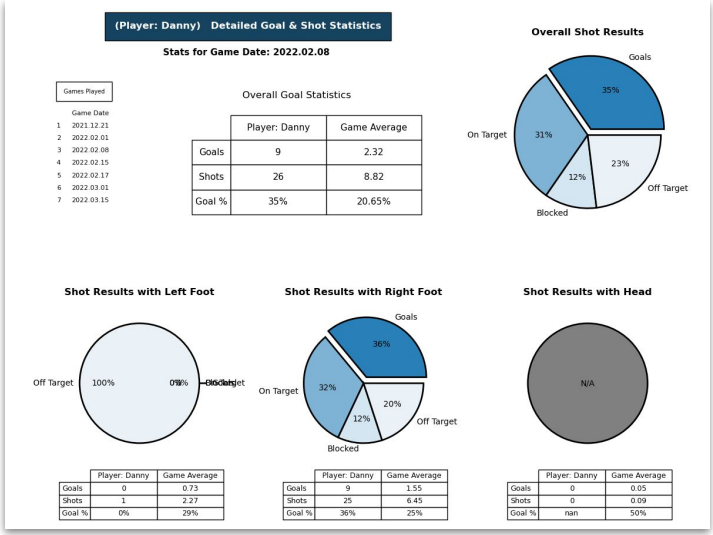
Insight from Python Analysis Results

Shooting Heat Map:

The Shooting Heat Map visualization provides an in-depth summary of a player's highly preferred shot locations, efficient goal-scoring locations and favorite spots for scoring goals. One of the most important takeaways from this visualization is determining if the player's highly-preferred shot locations correspond with a high goal percentage and goal count. This will help players improve their shot selection and overall shooting performance.

In the example shown in the heat map, I (as the player) can conclude that locations with large and light-colored marker sizes paired with low goal % are areas I should avoid shooting from.





Shooting Statistics Pie Charts & Tables Dashboard:

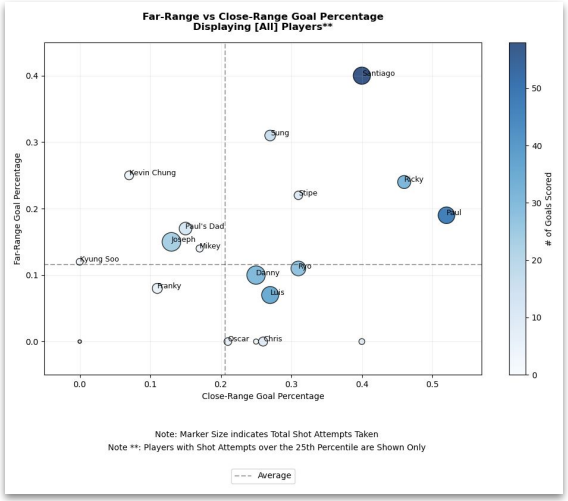
The dashboard provides a detailed overview of a player's shooting statistics of 3 different body types used for shooting (or overall statistics). With the main goal of scoring goals and taking shots that are On Target, players can refer to this dashboard to learn more about their shooting performance specifically related to accuracy. In addition, this dashboard can be used for pre-game scouting to determine which players prefer which leg to shoot with.

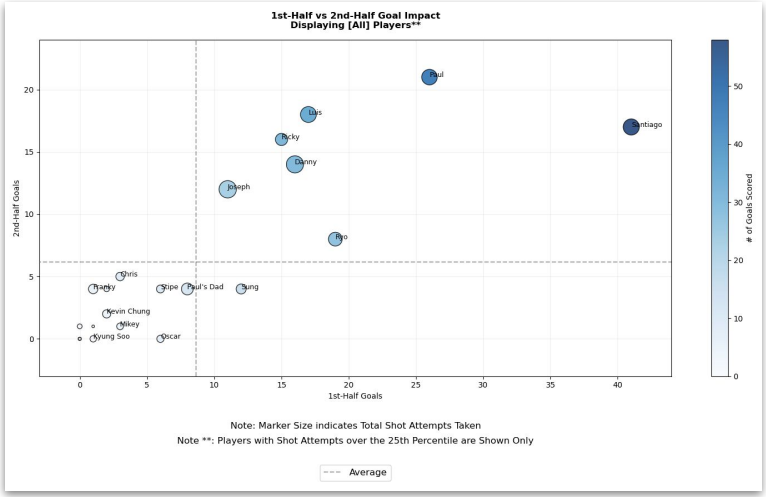
It is also important to note that the dashboard provides comparisons between the player's shooting performance with the league or game average.

1 Data Collection
2 Data Compiling
3 Prelim. Analysis in Google Sheets
4 Analysis in Python
5 Data Visualization

Data Visualization

Insight from Python Analysis Results





Close-Range vs Far-Range Goal Percentage Scatter Plot:

The close-range vs far-range goal % case study provides a detailed overview of the league's shooting performances related to the shot's distance from the goal. For pre-game scouting purposes, the plot can help players make conclusions on how to defend against the opposing player's shots. For example, a player playing against me (Paul) can make a decision to play tighter defense when I am close to the goal compared to when I am far from the goal based on my 2x higher close-range goal percentage (~50%) than my far-range goal percentage (~20%).

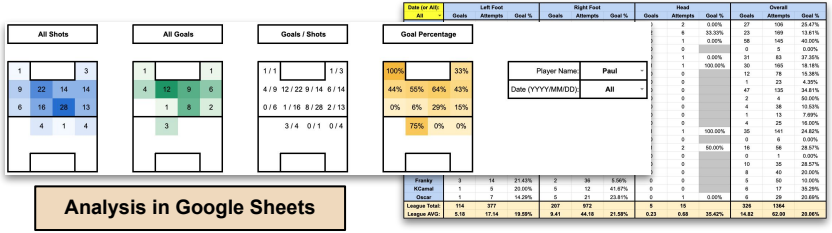
1st-Half vs 2nd-Half Goal Impact Scatter Plot:

The 1st-Half vs 2nd-Half goal impact case study provides an in-depth league-wide summary of players' attacking impact relative to the time of the game. As expected, majority of the players have a 1-to-1 ratio of 1st-Half to 2nd-Half goals. It's interesting to note that the player (Santiago) with the most goals primarily makes a goal impact earlier in a game than in the 2nd-half. With this plot, further research can be conducted on why certain players are more active in making an attacking impact in a specific half of a game.

Data Collection

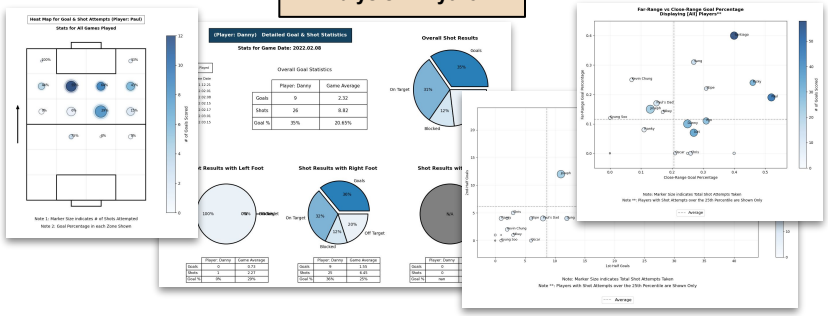
Data Compiling

	Date	V#	Vid T	Player	G	T	Z	F
2	2021.12.21	1	0:08:49	Ryo	N	Y	15	R
3	2021.12.21	1	0:09:14	Joseph	Y	G	18	R
4	2021.12.21	1	0:09:46	Santiago	Y	G	19	L
232	2022.02.01	2	0:04:04	Luis	N	Y	17	R
1029	2022.03.01	4	0:02:33	Santiago	N	N	14	L
1172	2022.03.15	2	0:06:50	Danny	N	B	20	R
1173	2022.03.15	2	0:07:05	Luis	Y	G	17	R
1365	2022.03.15	8	0:03:48	Ricky	Y	G	19	L



Analysis in Google Sheets

Analysis in Python



Concluding Thoughts

- From watching game video footage to collecting and compiling data and to running data analyses in Google Sheets and Python, the objectives I have set for this project have been achieved.
- The story told by the granular data set of shots over 7 games played by players in the league provide insightful results on various shooting metrics and shooting performances of each player.
- The visualization of the analysis results provide a user-friendly and in-depth dashboard for players to use for reasons such as self-improvement, pre-game scouting and overall statistics review for shooting performances.
- As the creator and analyst of this project, I have learned so much on how to process a mid-sized dataset in Google Sheets and how to write detailed analysis codes in Python through my own research. It has been absolutely amazing to see this project start from scratch with just watching game video footage to ultimately building a full dataset and insightful visualizations of shooting statistics for weekly recreational soccer games my friends and I play. I look forward to expanding this project with future action steps and diving deeper into implementing more advanced data analytics methodology into the project.

Future Action Steps

- With more games being played, I plan on collecting more data to increase the integrity and sample size of the main dataset.
- As data is collected for more games, the main dataset will continue to grow over 1365 rows of shot data. Once the dataset gets larger, I aim to implement SQL software programs to process the large dataset with more efficient and powerful SQL queries.
- For more dynamic visualizations, I hope to learn and utilize data visualization software, such as Tableau and Power BI, to build dashboards with more control and smoother design workflow.
- As of the current version of this project, the only types of in-game data metric that are being collected are shots and goals. Going forward, I'm setting a goal to collect data related to other events of the game, such as goalkeeper saves, passes, dribble attempts and defensive positioning. One of my dreams for this project is to compile a diverse dataset so that I can build an expected goal (xG) model for each player and each game.

Thank you for your time spent on viewing this project documentation! If you have any questions, comments or feedback, you can reach out at any of the following links and I'll get back to you as soon as I can.


paul3pjl@gmail.com


https://www.linkedin.com/in/paullee-1/


https://twitter.com/_laulpee


https://github.com/laulpee