PM566 Final Project Report

Sylwia Lipior

LA Galaxy U17 2022-2023 Season Game Data

Introduction

Catapult devices are GPS trackers worn by athletes. Wearable-based tracking technologies are used throughout sport to support performance monitoring. In addition to GPS capability, these devices contain inertial sensors comprising of an accelerometer (to measure acceleration forces), a gyroscope (to measure rotation), and a magnetometer (to measure body orientation). Inertial sensors collect data in three axes, or directions, allowing sensitive 'maps' of athlete movements and actions to be created. For Catapult's website, they claim: "The combination of the wearable tracking device and the inertial sensors creates a powerful athlete monitoring tool that ensures that key performance decisions are always supported with objective data." The sports performance department at LA Galaxy uses Catapult data to make decisions about performance readiness, rehabilitation, and training prescription.

This data is Catapult data collected over the course of the U17 2022-2023 season. I will specifically be look at U17 game data for that season. As a student in the USC Sports Science program and an intern at the LA Galaxy Sports Performance Department, I've had the opportunity to assist with collecting this data since January 2023. This data is typically visualized using either Catapult's Cloud where they offer many widgets to visualize data, or an internal athlete management system. LA Galaxy has been developing an athlete management system using Microsoft Azure. They export the data from Catapult and import it to Azure and have customized many different dashboards to visualize data. For this project, I decided to export CSVs directly from Catapult and try to wrangle the data myself.

Catapult data is collected at every training session and game. The players wear the devices on vests produced by Catapults and the GPS units are stored in a secure pouch on the back of the vest. During training or games, a member of the sport performance department will have an iPad which has the Vector app created by Catapult. The Vector app allows the user to input information about the training session or game, and it produces a live view of the Catapult data per player. The user can start and stop "Periods" based on training drills and which players are participating. After training, all the units are collected, put into a dock,

and uploaded to a computer. This data is then available in the Cloud and can be exported to the athlete management system for further visualization.

When thinking about this data, my research question became: does fatigue affect player's physical performance in soccer matches? More specifically, are players less physically productive when they are tired? To answer this question, I looked at the data at a few levels. To start off, I look at the difference in player's maximum velocities in the first half of games vs the second half. Then, I look at a string of five games in seven games that the team played in difficult conditions in the MLS Next Tournament, which was played in June of this year.

Methods

Preparing the data frame

When you export bulk CSVs from Catapult, you get observations for every player involved in the session for 1699 variables. A lot of that data is a little redundant, but I wrote a function to subset the data with only around 34 variables of interest to make it more manageable. The CSVs don't have the activity name easily accessible, so I wrote a function to extract the names from the names of the CSV files. I then wrote a for loop to read in all the data (~57 CSVs which corresponds to data from 57 games), making a new variable for the date of the session, and a new variable for the activity name. Finally, I de-identified the data since the data contains player names.

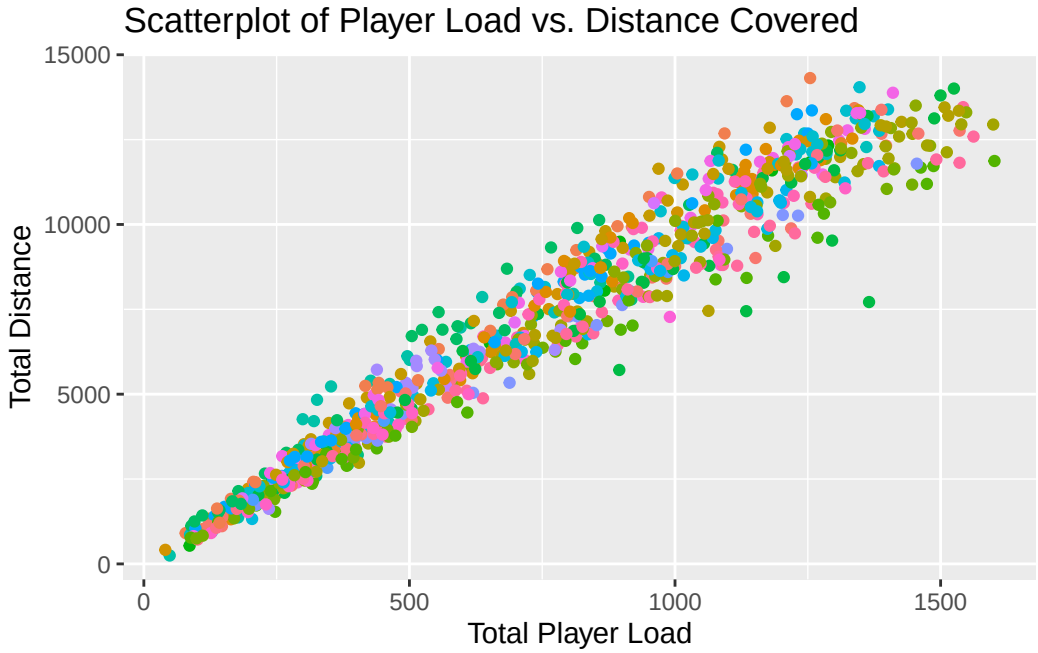
Initial visualization of the data

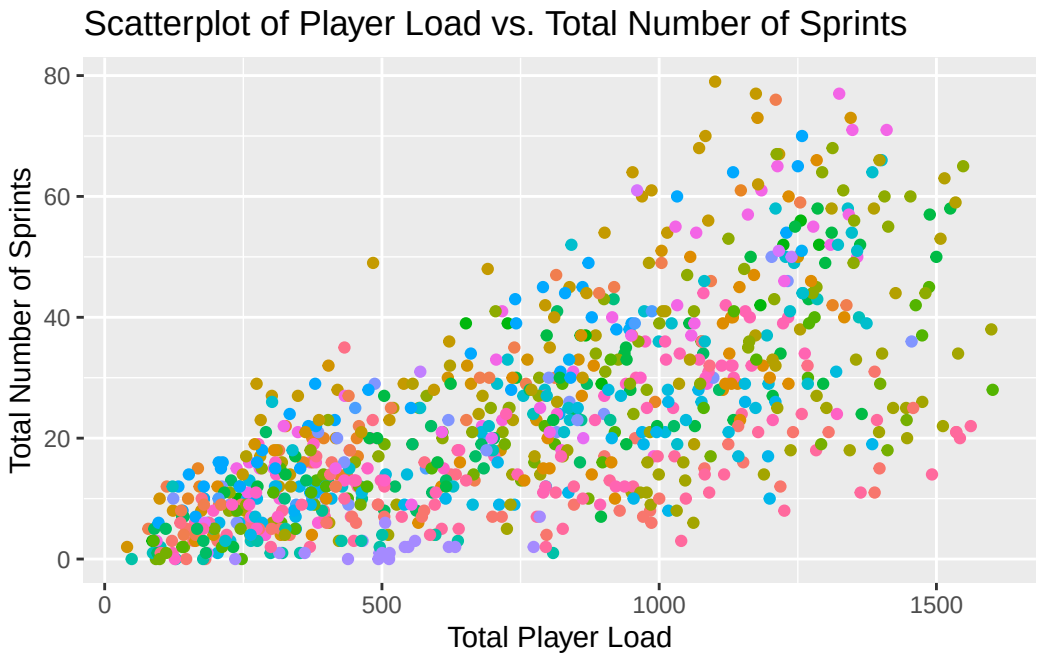
Next, I wanted to make sure the data looks how I would expect it to. Since the session names are inputted by staff, there is some room for error, and I wanted to make sure I have only game data here. To accomplish this, I decided to plot maximum velocity for each activity by month.

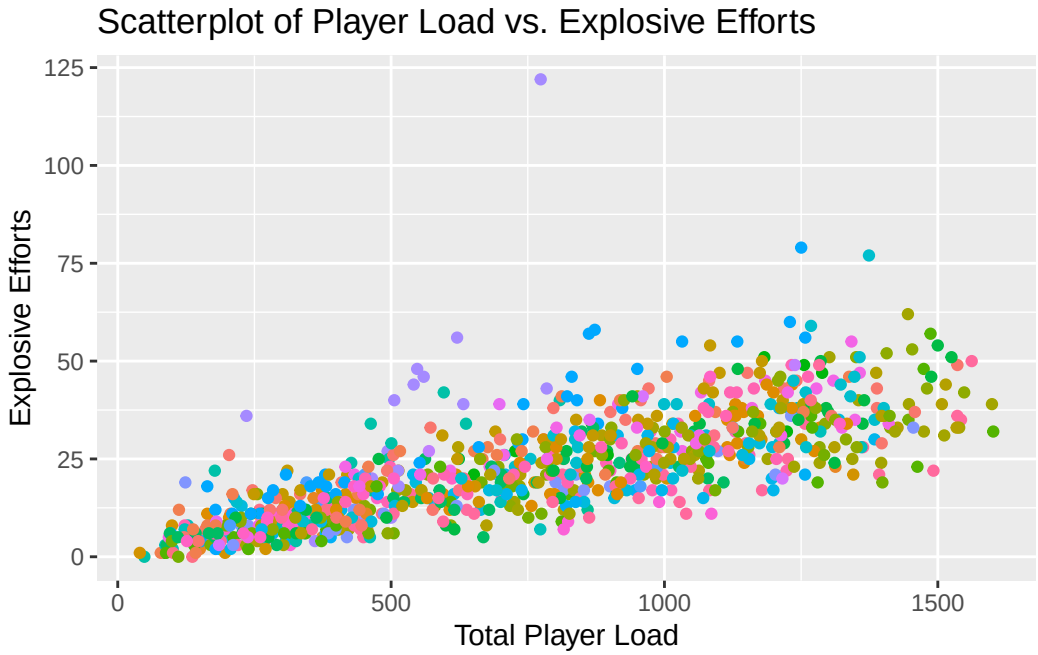
From looking at the initial box plots, I figured out that there is some data in the set that doesn't belong. Specifically, "U17 GD vs RSL" seems to have an issue with the GPS data since the maximum velocities are so low, so I decided to remove that data. In addition, the maximum velocities for "U17 GD Pre Season Day 2" are much lower than expected. Once I looked at the period names, I realized that this is data from a training session that was mislabeled as a game, so I removed it from the data set. I included the code to produce these box plots, but decided not to render here, because I did not use them for further analysis.

Player Load Scatter Plots

For some more initial visualization, I decided to look at the relationship between "Player Load", which is defined by Catapult as "the sum of the accelerations across all axes of the internal tri-axial accelerometer during movement", and a few other physical metrics. Specifically, I looked at scatterplot of Player Load vs Total Distance Covered, Total Number of Sprints, Explosive Efforts, and Total High Intensity Bouts.





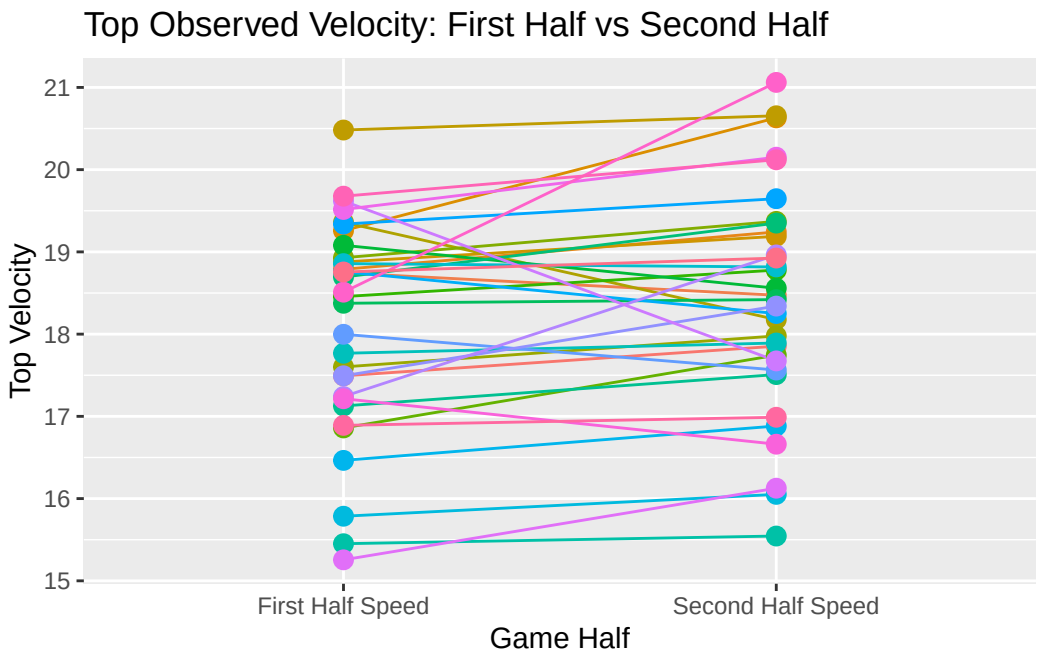


After looking at the plots, as expected there is a strong positive correlation between Player Load and all the other physical metrics. In other words, as a player covers more distance, or performs more sprints, explosive efforts, or high intensity bouts, their player load is expected to be higher.

Player Name	Half	Times in Top 3	Player Name	Half	Times in Top 3
Player_14	First Half	22	Player_14	Second Half	19
Player_17	First Half	17	Player_4	Second Half	14
Player_3	First Half	14	Player_12	Second Half	12
Player_4	First Half	11	Player_17	Second Half	12
Player_7	First Half	11	Player_7	Second Half	12
Player_19	First Half	9	Player_19	Second Half	11
Player_12	First Half	8	Player_21	Second Half	8
Player_15	First Half	7	Player_26	Second Half	7
Player_10	First Half	6	Player_3	Second Half	7
Player_20	First Half	6	Player_11	Second Half	6

Maximum Velocity Analysis

I thought it would be interesting to analyze maximum velocity from the dataset. Unfortunately, the data is a little difficult, because there are various Period Names that could signify either First Half or Second Half game data. Therefore, I used mutate to add a variable called "Period.Name.Halves" to denote which observations are from the first half vs the second half. I also removed Goal Keeper data, since their data looks very different from field players due to the nature of their position. I added a threshold of 10mph for speed, to make sure that I am actually using game data, and not some other potentially mislabeled data. I then visualized the maximum velocity data in a few different ways.



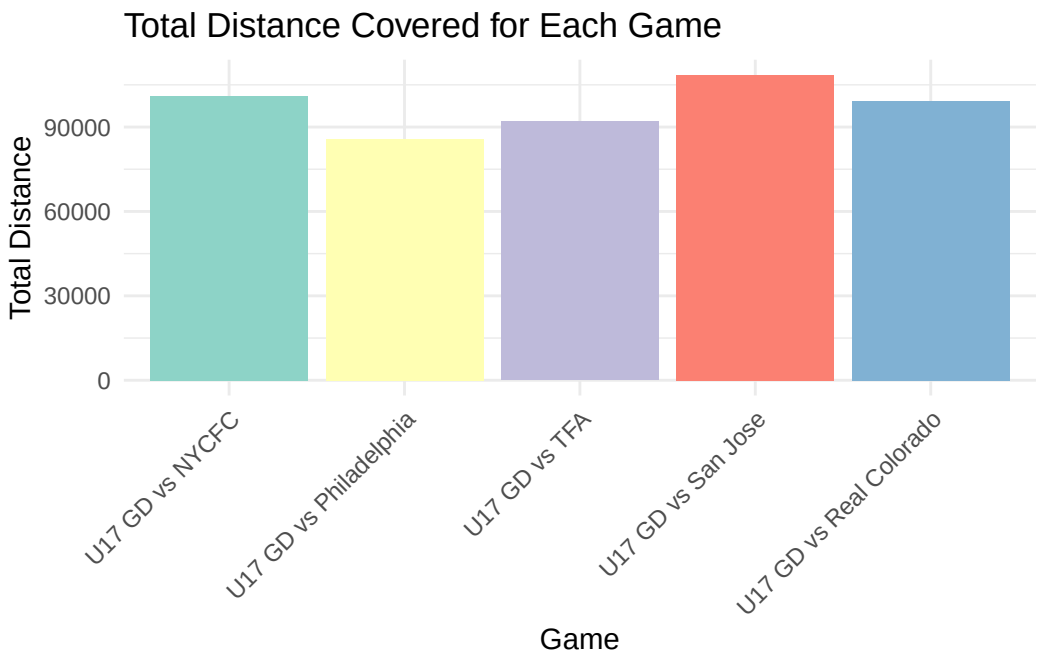
Activity Name	Game Date	Total Distance	Total High Intensity Bouts	Total Player Load	Total Explosive Efforts	Total Sprints	Total High Speed Distance	Total Very High Speed Distance	Total Sprinting Distance	Total Supra Velocity Distance
U17 GD vs NYCFC	2023-06-18	100962.69	468	9808.981	279	311	3741.93	691.55	244.39	35.18
U17 GD vs Philadelphia	2023-06-19	85746.10	438	8493.206	239	216	2843.61	651.58	216.41	60.81
U17 GD vs TFA	2023-06-21	91976.23	432	8732.986	255	269	3608.76	466.27	120.83	19.96
U17 GD vs San Jose	2023-06-23	108491.91	508	10766.929	283	304	3877.79	564.90	231.75	25.15
U17 GD vs Real Colorado	2023-06-25	99200.09	436	9329.604	236	230	2911.11	465.49	182.11	45.42

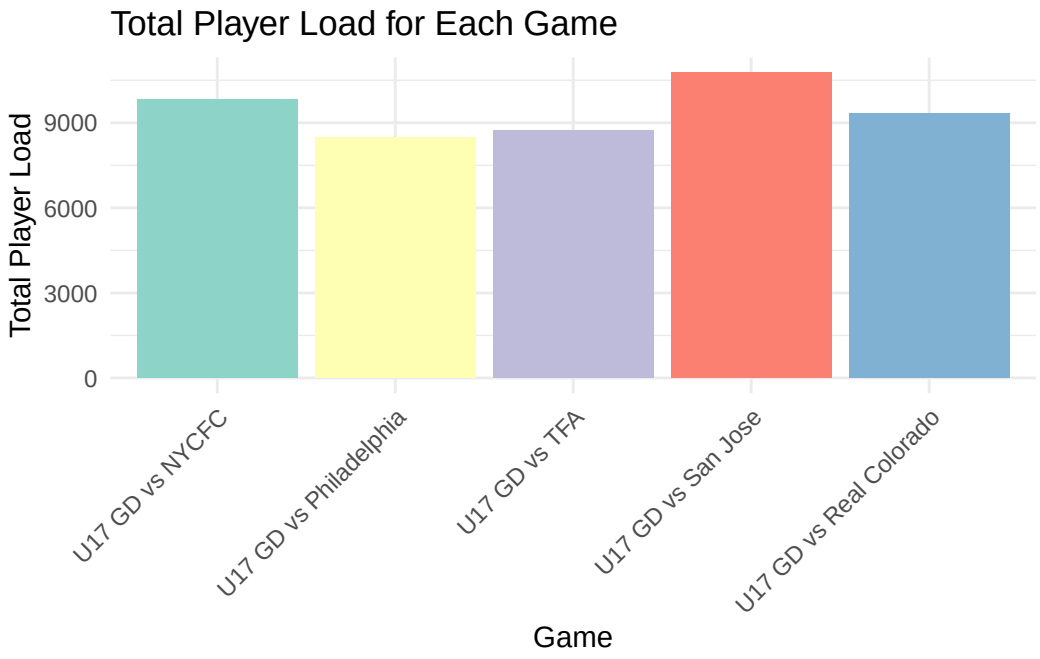
I was interested in looking at speed on the individual player level. I took the frequency of players that appeared in the top 3 max speed values per activity. "Player_14" is clearly the fastest player, since he appears in the top 3 the most times out of anyone. I made a stacked box plot looking at the five players that appeared in the top 3 for maximum velocity the most number of times. Four of the players appear in the top 3 the most times for both the first half and second half, but player 3 appeared in the top 3 the third most times for first half and player 12 appeared in the top 3 the third most times for the second half. That tells me that player 12 might be a second half substitute often while player 3 gets subbed off in the second half.

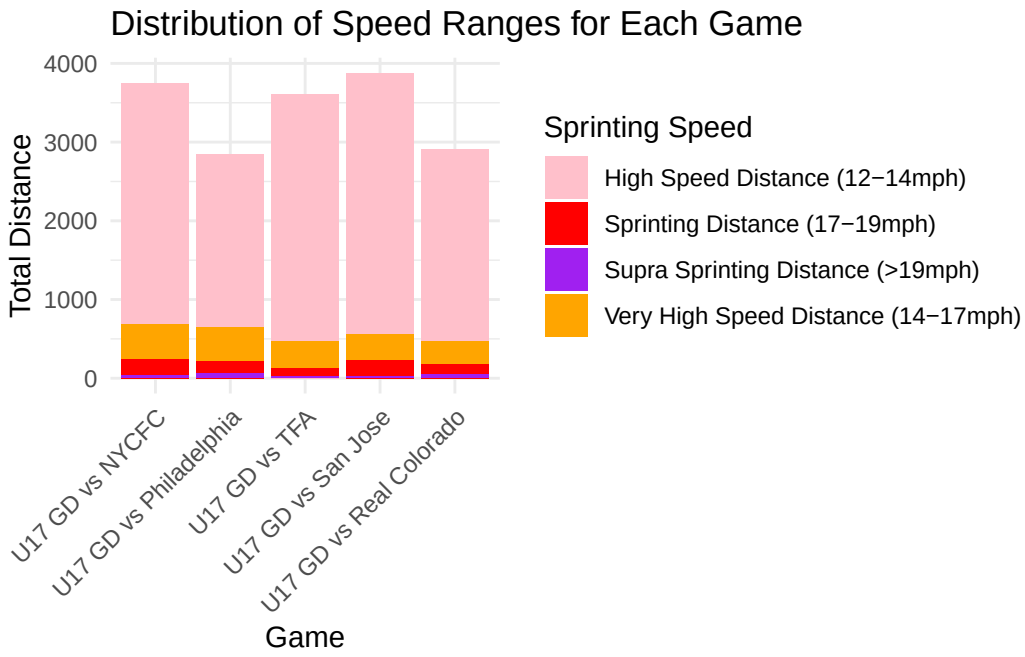
I then made a spaghetti plot looking at the top recorded velocity for the first half vs second half for each player to see if there is a difference in performance between halves. Interestingly, the trend seems to be that players achieve higher maximum velocities in the second half of games. That makes sense, because player's might be experience fatigue and might make errors where they have to achieve very high speeds to deal with counter attacks.

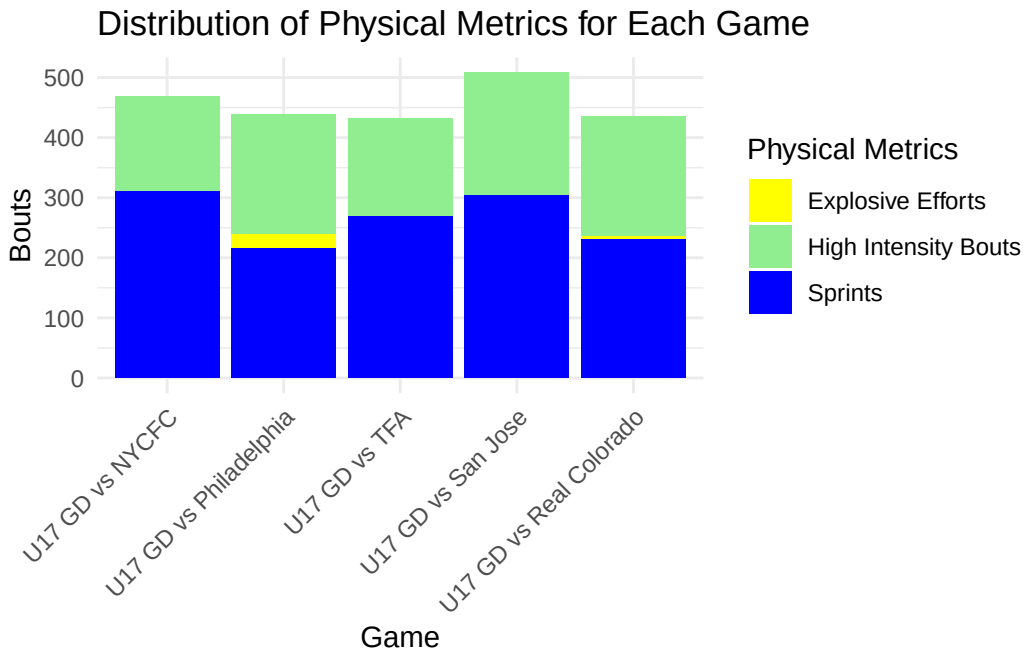
Extract Games from MLS Next Tournament (June 2023)

In June of 2023, the LA Galaxy U17 team won the MLS Next Tournament. In order to hoist the trophy, they played 5 matches in 7 days in very difficult and humid conditions in Dallas, Texas. To perform an analysis on this data, I had to subset the master data frame to extract the data for these matches. Then, I calculated the match totals for the following physical metrics: total distance, total high intensity bouts, total player load, total explosive efforts, total sprints, total high speed distance (12-14 mph), total very high speed distance (14-17 mph), total sprinting distance (17-19 mph), and total supra velocity distance (>19 mph). I found this data following comments from the midterm. I created a table summarizing this data, and made multiple bar charts.


Total Distance Covered for Each Game
Total Distance
90000
60000
30000
0
U17 GD vs NYCFC
U17 GD vs Philadelphia
U17 GD vs TFA
U17 GD vs San Jose
U17 GD vs Real Colorado
Game


Total Player Load for Each Game
Total Player Load
9000
6000
3000
0
U17 GD vs NYCFC
U17 GD vs Philadelphia
U17 GD vs TFA
U17 GD vs San Jose
U17 GD vs Real Colorado
Game





When looking at total distance covered, there doesn't seem to be a decrease as the matches progress. Interestingly, the first match had the most sprints, and almost the highest high speed running distance. This is expected since this was the first match of the tournament. There was a noticeable dip in the second match, but this was played the day after the first match.

There seemed to be some affects of fatigue in this match. The high speed running distance was the lowest of the tournament. The players also completed the fewest number of sprints in this match. There is also a dip in the number of sprints in the last match, which seems to be an affect of accumulated fatigue. Overall, all of the physical metrics looked at here are have comparable values. When looking at the distribution of speed ranges for each match, there was a noticeable decrease in high speed running in the second match (the one with the least rest) and the last match, again probably due to accumulated fatigue. The distribution of sprinting speeds seems to be fairly similar between matches. As expected, players cover the most distance at high speed, and a small amount of distance at "very" high speeds and sprinting speed. When considering the supra velocity distance, it was interesting that the last match had the second highest distance for this category. Even though you would expect the players to be very fatigued at this point, they still managed to hit some very high speeds.

Conclusion

From my analysis, I was able to get a better understanding of how fatigue affects physical metrics measured by Catapult devices. In my analysis of maximum velocity on the player level, I found that fatigue within a single match does not seem to affect whether players will hit a high maximum velocity. The average maximum velocity for the team is slightly higher in the first half and in the second half, which is expected. Intra-player differences in maximum velocity are very low between the first half and the second half. Furthermore, in my analysis of the MLS Next Tournament, I found that there was not a large impact of accumulated fatigue. The distribution of different sprinting speed distances was similar between the games (i.e. player's did not seem to be sprinting less as the tournament progressed). That's must mean that players are good at recovering and have good fitness. The matches that seemed the most affected by fatigue were the second match and the last match. As discussed above, that is expected because the second match was played the day after the first match. All of the other matches had at least one day's rest in between. The last match presumably had lower values due to accumulated fatigue. In conclusion, fatigue definitely has an effect on player's physical performance, but this needs to be analyzed further, and it varies on a case-by-case basis. High performing athletes seem to be very good at recovering quickly and minimizing the effects of fatigue.