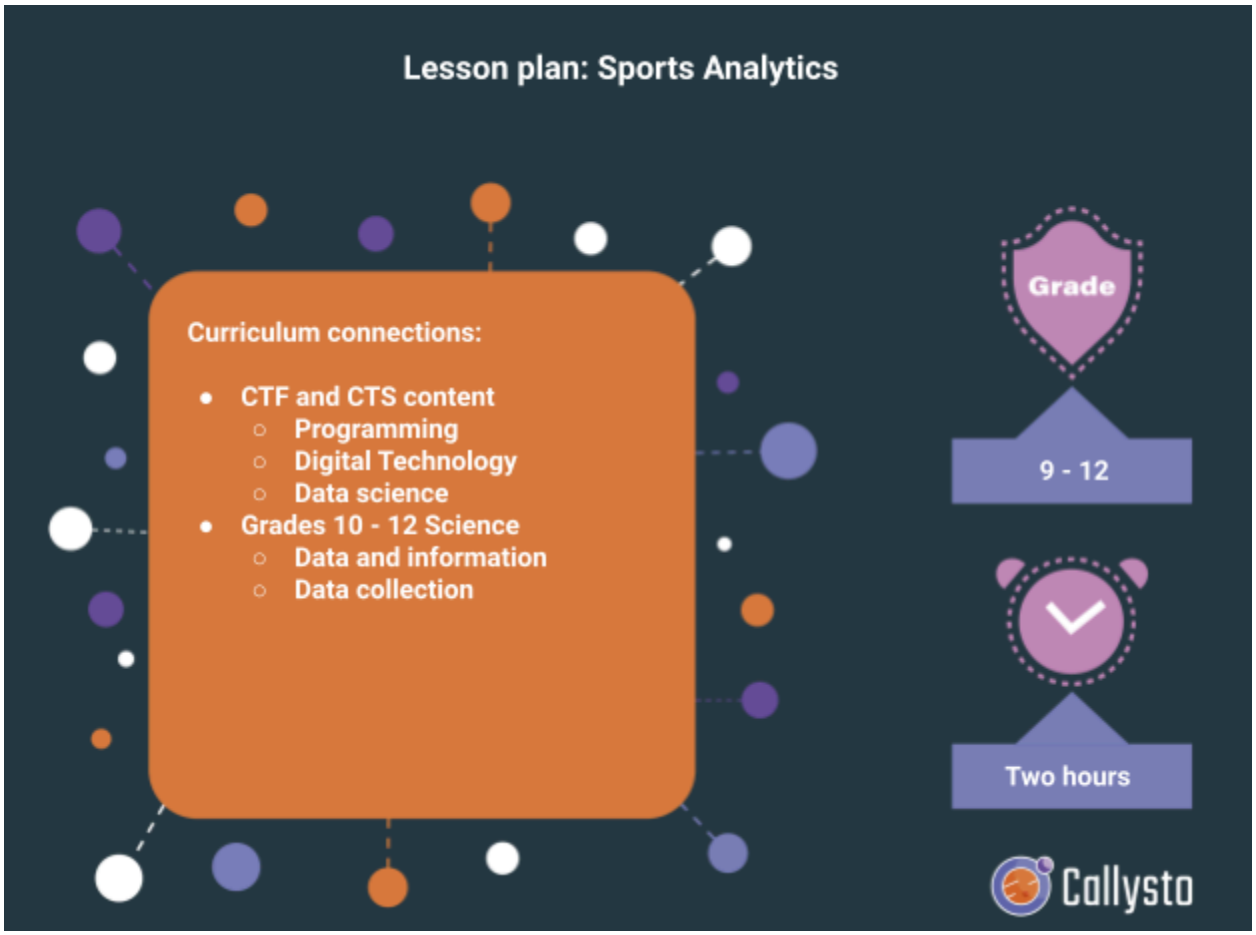

Webscraping

Planned by: Michael Lamoureux

Date: March 2024

Duration: Two hours

Introduction

This is a free resource for teachers and students and is part of the Callysto project, a federally-funded initiative to foster computational thinking and data literacy in Canadian Grade 5-12 classrooms.

Sports Analytics is the process of collecting data about a sport, the games played in the sport, and the players involved in that sport, then making informed decisions based on an analysis of that data. This can be done for fun and entertainment as well as for making important financial decisions about a team and its players. The movie Moneyball, with Brad Pitt and Jonah Hill, showed how sports analytics was able to help create a winning team in baseball.

Many professional sports leagues and their teams gather detailed information about games and players and post much of this data on publicly available websites. We can use computer code to gather all the information and then do an analysis on that data. Often, we preform data visualizations to help understand this data. Code libraries are available to help make this process easy for beginner programmers.

In these notebooks, students will learn how to use Python software libraries to gather data from online web pages and put the information into a Pandas dataframe, suitable for statistical analysis. The notebooks focus on finding interesting sports data and displaying it in an informative graph or chart. We also explore the idea of regression and best-fit lines to observe correlations between types of data.

The notebooks here cover soccer and baseball analytics as a starting point for students. Students are also encouraged to explore a variety of other data sources and web scraping methods.

Grade level and audience

Grades 9 - 12

Learning outcomes/curriculum connections

- This lesson connects CTF and CTS outcomes by blending digital technology, programming, and hands-on data collection and analysis. It provides students with a well-rounded learning experience, merging theory with practical skills to prepare them for the changing tech field.

- Grade 10-12 Science curriculum
    - Data and information
    - Data collection
    - Statistical analysis

Required materials

Required materials

1. A charged computer.
2. Access to the internet.
3. An installed internet browser, preferably Google Chrome.
4. A Google or an Outlook email account.

Students should know how to log in to the Callysto Hub as well as run a notebook prior to interacting with it. Teachers, to get started with Callysto notebooks and running material on the Callysto Hub, see our Starter Kit.

In-class activities

Activity 1 : Baseball analytics

Time: 60 minutes

Link to notebooks:

- baseball-part1.ipynb
- baseball-part2.ipynb
- baseball-part3.ipynb

In this activity, students will learn how to use Python libraries to gather baseball data from both a spreadsheet and online websites, then save in a dataframe. The data includes performance of baseball players and how well they hit the ball. Part 1 examines where the player tends to hit a ball – in left or right field. This is called a spray chart.Part 2 examines the strike zone and where the batter is able to make a good hit. Part 3 looks at correlations between the player's ability to produce runs, and stats like batting average.

Students are encouraged to use the code to access other interesting statistics about Major League players.

The notebooks continue with more analysis in Activity 2, focused on soccer. .

Activity 2 : Soccer Analytics

Time: 60 minutes

Link to notebook:

- soccer-partI.ipynb
- soccer-partII.ipynb
- soccer-partIII.ipynb

Students continue the work on collecting data from webpages, using additional Python tools for gathering the data with a focus on soccer data. These tools let the students download data and store them in a data frame for further statistical analysis.

Part I examines the types of play on the soccer field for a sample game, with a visualization of the playing field. Part II looks at scoring techniques for games played in the UEFA Champions League website. Part III examines the relationship between ball possession and scoring rates.

Students are encouraged to use the code to access other interesting statistics about professional soccer..

Reflections

- *What parts of this project do you think worked really well and why?*
- *Did you face any difficulties during the project, and how did you manage to solve them?*
- *If you had the chance to do the project again, what things would you want to change or make better, and what would you keep the same? Explain why you feel that way.*
- *Can you see any ways the things you learned in this project could be useful in your other schoolwork, lessons, or things you've experienced*

Next steps

For more information, you can check out our YouTube videos, online courses, or callysto.ca for learning modules, tutorials, lesson plans, exercises and events.

Contact

If you encounter any issues or have any suggestions, please get in touch with us at contact@callysto.ca or twitter.com/callysto_canada.