Nathaniel Hemans / Niecia Say / Cobi Millier

Studypool clients | Benardo avancena / helen merhout

SCHOOL INDEX DOCUMENTATION

TAG Group Data Science Capstone Project

[1 Project Description 4](#_Toc165662133)

[2 TAG Team Members 4](#_Toc165662134)

[2.1 Clients - STUDYPOOL 5](#_Toc165662135)

[3 Client Presentation 6](#_Toc165662136)

[4 Team Plan 7](#_Toc165662137)

[4.1 First Iteration 7](#_Toc165662138)

[4.2 Second Iteration 7](#_Toc165662139)

[4.3 Third Iteration 7](#_Toc165662140)

[5 GitHub Repository & Colab Notebook 8](#_Toc165662141)

[5.1 GitHub Picture and Link 8](#_Toc165662142)

[5.2 Colab Notebook Picture and Link 9](#_Toc165662143)

[6 Web Crawling Methods & Procedures 10](#_Toc165662144)

[6.1 Extraction of Schools’ Summaries 10](#_Toc165662145)

[6.2 Extraction of Schools’ Logos 10](#_Toc165662146)

[7 Data Analysis Methods 11](#_Toc165662147)

[8 Features Implemented 11](#_Toc165662148)

[9 Technologies Used 12](#_Toc165662149)

[9.1 Data Modeler Technology 12](#_Toc165662150)

[9.2 Data Analyzer Technology 12](#_Toc165662151)

[9.3 Data Visualizer Technology 12](#_Toc165662152)

[10 Collections 13](#_Toc165662153)

[10.1 Dataset Reference 13](#_Toc165662154)

[10.2 Website Template Reference 13](#_Toc165662155)

[10.3 Early Website Design 13](#_Toc165662156)

[11 Live Demo of Project 14](#_Toc165662157)

[12 Known Issues 14](#_Toc165662158)

[13 TODO Tasks 15](#_Toc165662159)

[13.1 Tasks Remaining 15](#_Toc165662160)

[13.2 Future Features 15](#_Toc165662161)

[14 Burndown and Velocity Charts 16](#_Toc165662162)

[14.1 First Iteration Burndown Chart 16](#_Toc165662163)

[14.2 Second Iteration Burndown Chart 16](#_Toc165662164)

[14.3 Third Iteration Burndown Chart 17](#_Toc165662165)

[14.4 Velocity Chart 17](#_Toc165662166)

[15 Project Flow Charts 18](#_Toc165662167)

[15.1 Iteration 1 18](#_Toc165662168)

[15.2 Iteration 2 18](#_Toc165662169)

[15.3 Iteration 3 19](#_Toc165662170)

# Project Description

The objective of this project is to develop a comprehensive database encompassing colleges, universities, and online schools across the United States. The database aims to assist students in their research endeavors by providing information on various aspects of each educational institution. Notably, the system will incorporate filtering functionalities, enhancing the efficiency of student searches, and facilitating access to relevant data.

# TAG Team Members

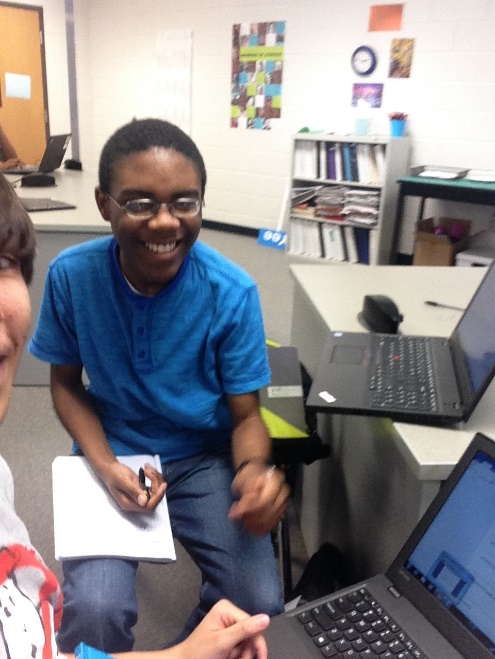
Niecia Say – Data Modeler & Team Manager



Cobi Miller – Data Analysis & Client Liaison



Nathaniel Hemans – Visualizer & Project Documentation



## Clients - STUDYPOOL

Benardo Avancena – Client, StudyPool COO

Helen Merhout – Client, StudyPool Full Stack Engineer

# Client Presentation

StudyPool is a company dedicated to serving the academic needs of students across a wide range of topics and courses. They provide a range of services aimed at making students successful and deepening their learning experience overall. These services would entail 24/7 Study Help, a database comprised of notes and guides made by other students, and personal video tutoring.

Their 24/7 Study Help is there for students to ask questions answered by verified tutors with expertise in the subject. For the *‘Notebank’,* students are paid for their book synopsis, study guides, essays that are readily accessible by other students. Lastly, StudyPool’s video tutoring is a one-on-one service where, after a student provides the necessary information, they can be paired with a specialized tutor that best suits the student’s needs.

# Team Plan

## First Iteration

* Met with the clients for the first time to understand the functional requirements and tasks of the project.
* Wrote a comprehensive list of functional requirements as per what the clients requested.
* Set up Jira so that tasks can be correctly defined, categorized, and tracked.
* Collected a series of datasets which were then cleaned and filtered out for analysis and database implementation.
* Established plans for the kind of statistics to be done and the technology for statistical analysis.
* Created a mock website that demonstrates the user experience for searching schools and its information.

## Second Iteration

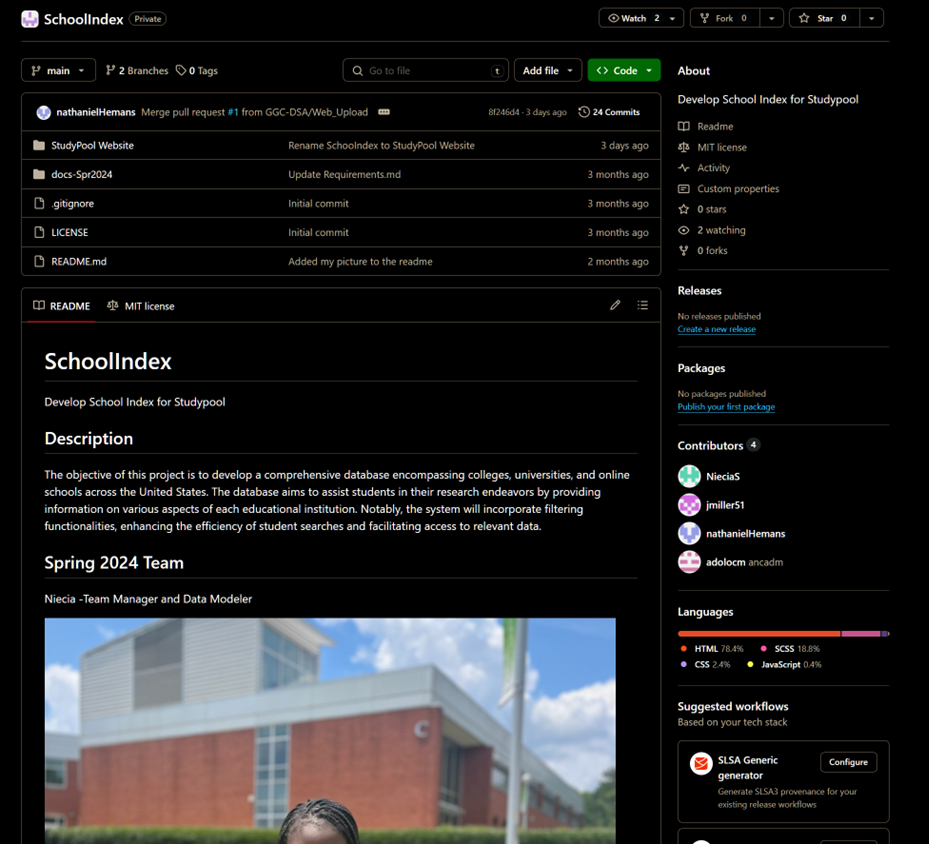
* Created a process for automating the extraction process of ETL.
* Created a data file for the logos of colleges.
* Set up a database utilizing MySQL as the basis of the project where data will be stored as required by the client.
* A second iteration of the website was created that supports code manipulation with HTML and CSS files.

## Third Iteration

* Continuation of automated data extraction using web crawling.
* Created a data file for the descriptions of colleges.
* Ported data within the MySQL database and sent it for sharing to StudyPool.
* Created finalized visualizations using Tableau.
* Researched process for implementing website to the database with PHP.
* Added more functionalities to the website.

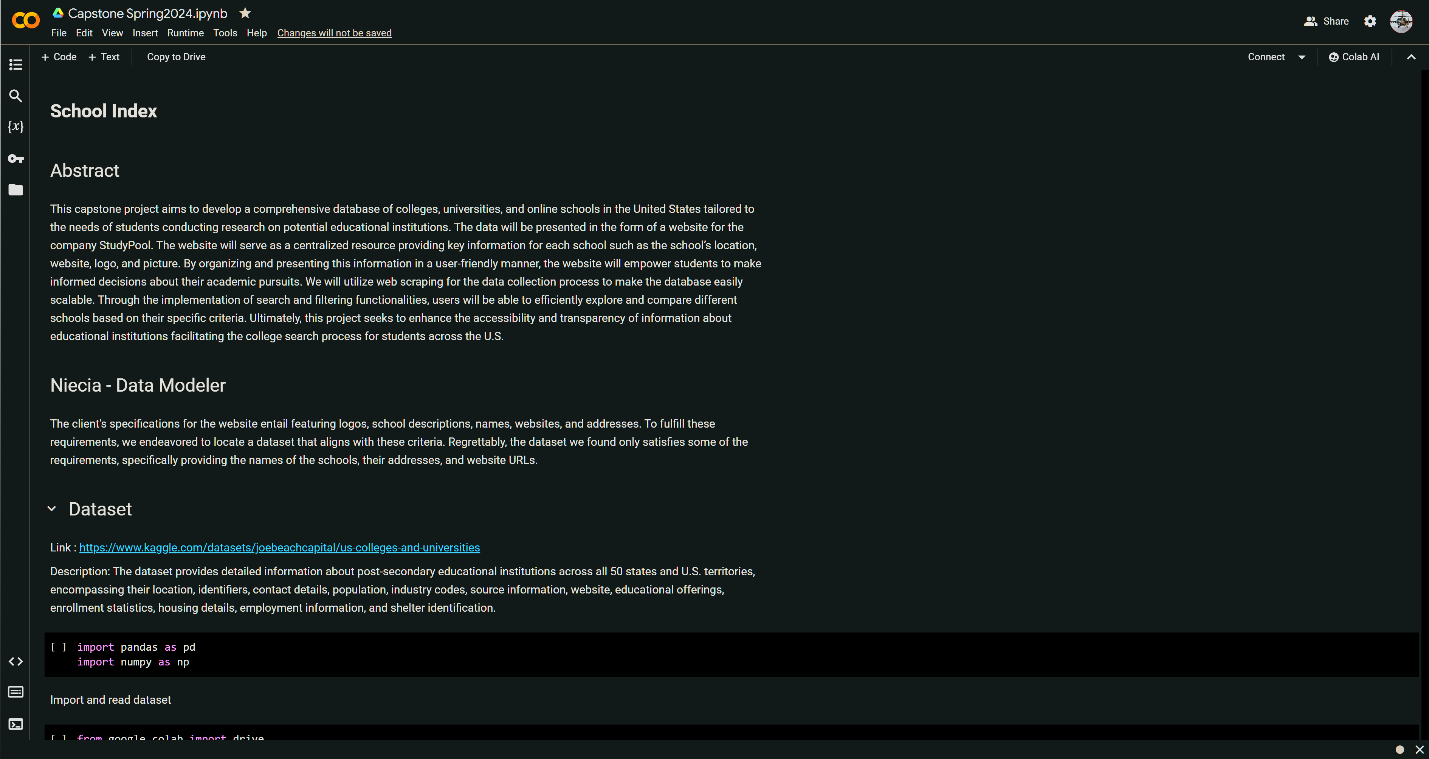
# GitHub Repository & Colab Notebook

## GitHub Picture and Link



<https://github.com/GGC-DSA/SchoolIndex?tab=readme-ov-file>

## Colab Notebook Picture and Link



<https://colab.research.google.com/drive/1g_mjqjC1Nx2u5h3dB9Y6kTNLTih8utay?usp=sharing>

# Web Crawling Methods & Procedures

## Extraction of Schools’ Summaries

The code uses the Wikipedia library to access Wikipedia's vast repository of articles and retrieve summaries for a list of schools. The Wikipedia library acts as an interface to interact with Wikipedia's content programmatically, providing functions to access article text, summaries, images, and more. To use this code, you need to ensure that you have the necessary libraries installed, including csv and Wikipedia. Prepare a CSV file containing a list of school names, ensuring that it is formatted correctly and named "GA\_College\_Name.csv". Then, simply run the script. It will read the CSV file, iterate through each school name, retrieve its summary from Wikipedia using the ‘get\_school\_summary’ function, and store the summaries in a new CSV file named "school\_summaries.csv". Once the script finishes execution, you can find the collected summaries saved in the output CSV file. With this script, you can efficiently gather comprehensive summaries of schools from Wikipedia.

## Extraction of Schools’ Logos

The code utilizes the SerpApi library to perform an image search for "colleges in Georgia logos" on Google Images. It retrieves image results using pagination to ensure comprehensive coverage of the search. To use the script, ensure that you have the necessary libraries installed, including csv, serpapi, and dotenv. Set up a .env file containing your SerpApi key. Then, run the script, and it will automatically perform the image search, retrieve results, and save them into the specified CSV file.

# Data Analysis Methods

As this project focused more on data modeling than data analysis, limited data analysis was performed on the data set.

* Filtered the data set down to one hundred colleges in the state of Georgia.
* All the columns in the original data set were combed through, and the ones that may have utility in the future were extracted. The remaining columns were discarded.
* Filtered the data set to remove outlier values for the total enrollment column – excluded colleges with a null value (-999) or with more than 10,000 enrolled students.
* Experimented with linear regression machine learning through Python.
  + This was not successful because the regression was done on enrollment vs. latitude. The latitude alone does not convey meaningful information – longitude is needed for coordinate data to be meaningful.
* Main visualization initially created with Matplotlib plots remaining colleges in the data set on a longitude vs. latitude graph.
* Visualization developed in Tableau to show the colleges on a map of Georgia and colored the dots to reflect the number of students enrolled in each college.

# Features Implemented

* Implemented a web scraper to extract the schools’ logos and summaries
* Developed a data visualization showing schools in the data set on a map of Georgia with a color gradient showing how many total students are enrolled in the data set.
* Developed a MySQL database to connect with the website.
* Added a primary key and two unique constraints to columns in the MySQL database.

# Technologies Used

## Data Modeler Technology

* Google Colab
* SerpApi
* Python
* Excel
* Wikipedia API

## Data Analyzer Technology

* Excel
* Python Notebook
* Matplotlib Python library
* Tableau
* MySQL

## Data Visualizer Technology

* Figma UI/UX website design platform: <https://www.figma.com>
* HTML
* CSS

# Collections

## Dataset Reference

* Title: Comprehensive Dataset of U.S. Colleges and Universities
* Source: <https://www.kaggle.com/datasets/joebeachcapital/us-colleges-and-universities>
* Description: Provides extensive information on post-secondary educational institutions across all 50 states and U.S. territories. It includes details such as location, identifiers, contact information, population, industry codes, source references, website URLs, educational programs, enrollment statistics, housing facilities, employment data, and shelter identification.

## Website Template Reference

* Provides already established websites packaged with proper CSS and HTML files that can be edited and changed
* This is where the bases of the website used in this project came from
* <https://www.free-css.com/free-css-templates/page294/hirevac>

## Early Website Design

* This was an early idea of the website based on the functional requirements of the client
* Done without the use of any HTML or CSS code using a UI/UX design platform called Figma
* Link to Figma: <https://www.figma.com/file/NcFuqP4PtsDglRWvYZkAui/StudyPool-School-Index?type=design&node-id=0%3A1&mode=design&t=4VWaIBPC1b2ynkr3-1>

# Live Demo of Project

* Link of Demo of the project: <https://youtu.be/PRgV5qmOW-o?si=UOPrtdF9YKvSRovf>

# Known Issues

* After scraping the schools’ summaries, some summaries were not found for certain schools. This issue arose either because the school could not be on Wikipedia or due to a Disambiguation Error. Such errors occur when school names are ambiguous, resulting in multiple potential matches on Wikipedia, making it challenging for the API to automatically determine the correct page.
* In the logos data file, all serpapi links have expired and need to be re-scraped. This should only affect one school.
* The logos data file contains schools that are not in the current school data set, and there are schools in the school data set without a logo in the logo data set.

# TODO Tasks

## Tasks Remaining

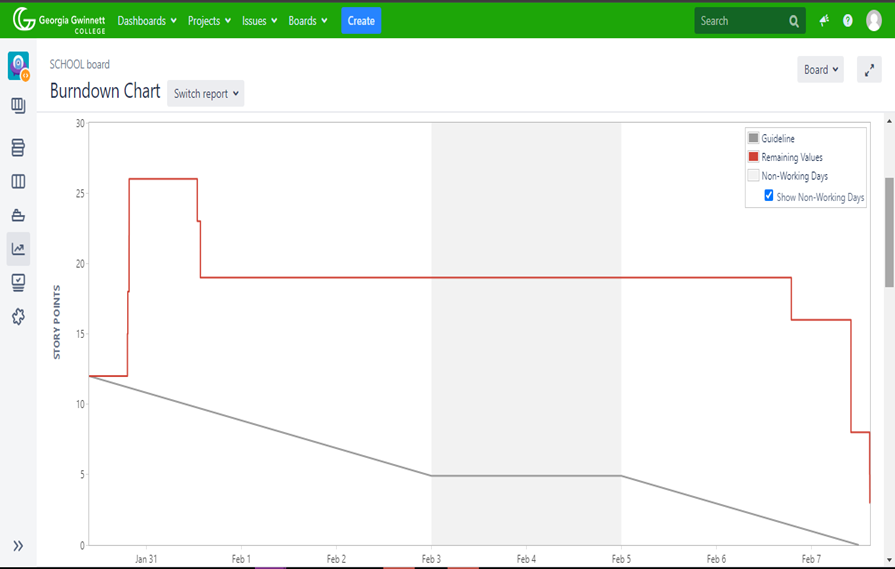
* Set up local server to host website to allow for testing of client’s API.
* Finish cleaning the logos data file and clean the descriptions file.
* Merge cleaned logo/description files into the SQL database.
* Send the client the output of a mysqldump Powershell command.
* Finish connecting the MySQL database to the website.
* Extract logos and summaries from the remaining schools in Georgia and across the US in the dataset.
* Find a solution to address schools without logos and summaries.
* Address disambiguation errors caused by ambiguous school names.

## Future Features

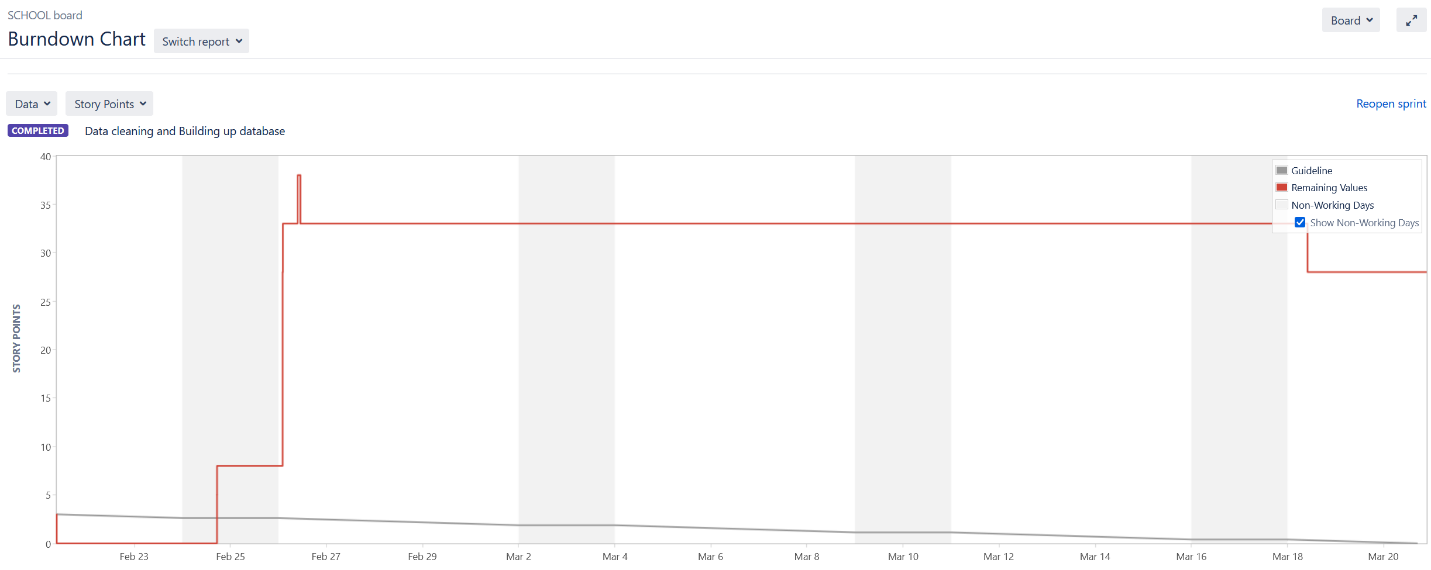
* Mobile-friendly website interface.
* Re-styling of website to match Studypool's website theming.
* Expansion of data set beyond Georgia colleges.

# Burndown and Velocity Charts

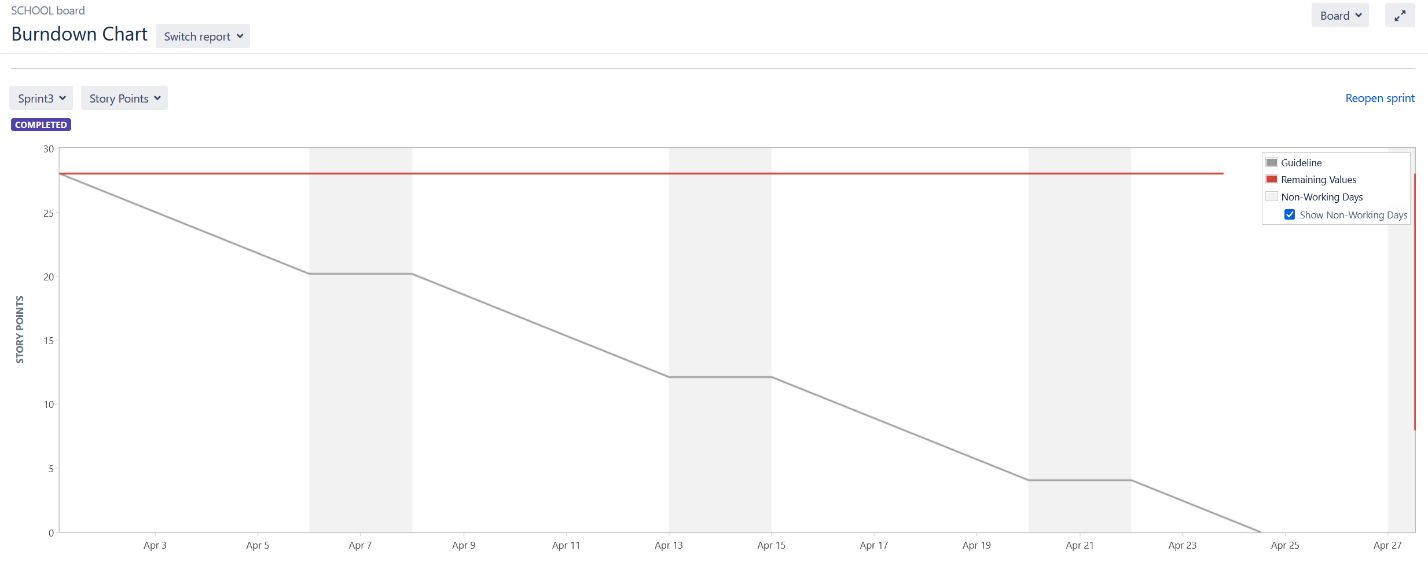
## First Iteration Burndown Chart



## Second Iteration Burndown Chart



## Third Iteration Burndown Chart

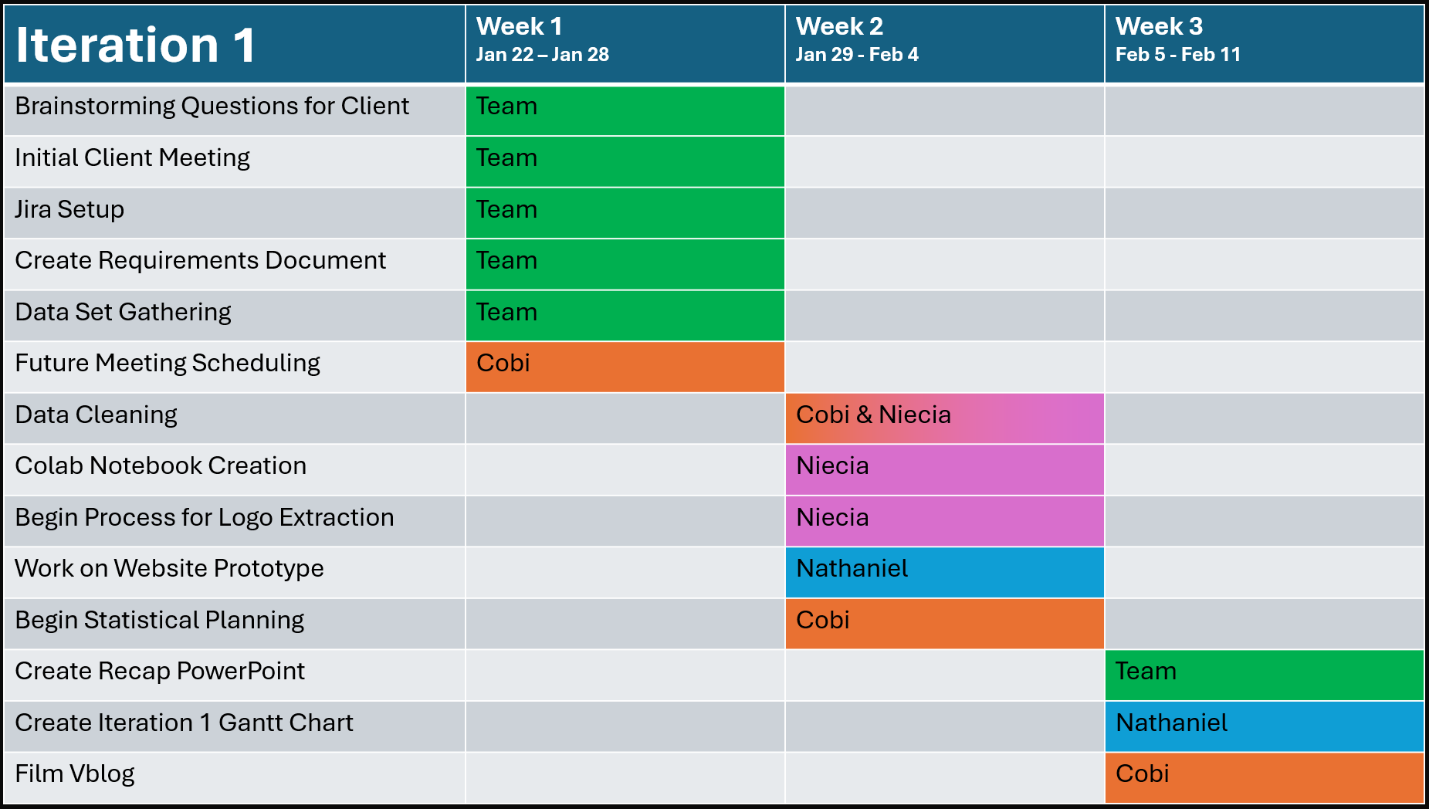


## Velocity Chart

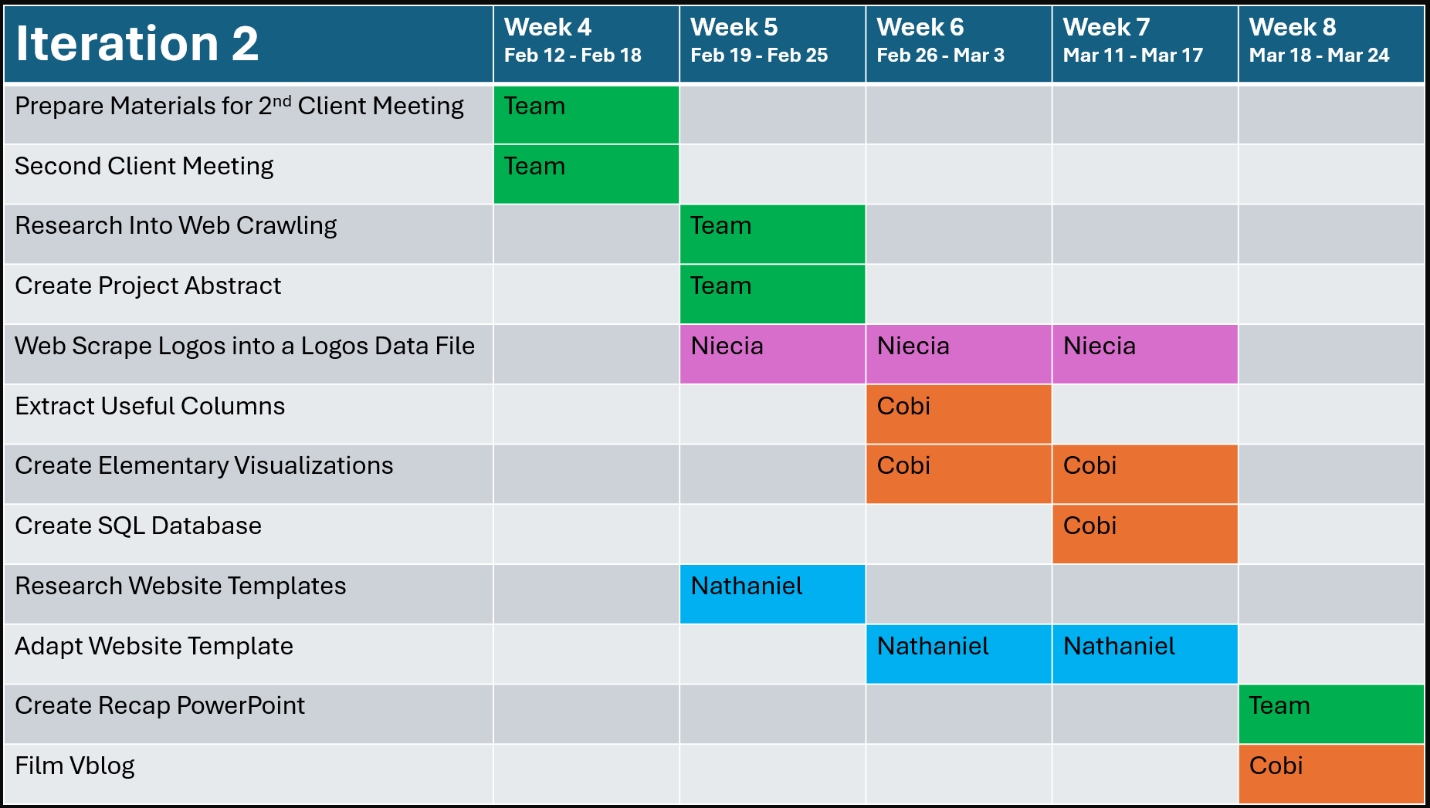


# Project Flow Charts

## Iteration 1



## Iteration 2



## Iteration 3

