Final Project Report – Grizzly Paths

Project Description

For our Data Science and Analytics Capstone project, our team worked with client, GGC alum, Hugh Smith led by Dr. Anca Doloc-Mihu to create a functional website for IT students to attain valuable skills based on the real-world job market. Grizzly Paths is a website where students can enter skills they want to attain and match them with corresponding courses that teach them. They may also view skills that are in-demand, web scraped from popular job sites such as Glassdoor, LinkedIn, and Indeed, in their respective domains to help ensure their marketability in their dream job fields.

Team Members' Introductions & Group Photo

Team Members:

Krishan Bhalsod, Lucas Leon, Nikhita, Michelle Webb

Group Photo:

https://github.com/GGC-DSA/itskills/blob/main/media/Spr%20'25/teamphoto2.jpg

Client Presentation

Client: Hugh Smith

Hugh is a current DSA professional and GGC alum. He started the project in Summer 2021. His feedback and experience provided valuable insight and ultimately drove the project.

Team Plan

Our team used Jira for sprint planning and task management throughout the project. Communication with each other and our client was maintained through Discord, Microsoft Outlook, and Zoom. During each sprint, we individually brainstormed and created tasks based on our assigned roles, collaborating closely to review and provide feedback as tasks were completed. Since our project was a continuation of an ongoing effort that began in 2021 under Hugh Smith, we first analyzed the progress and data collected during earlier iterations. Building on that foundation, we scraped additional data from major job posting sites, then cleaned, analyzed, and merged it into a comprehensive master dataset. Lucas and Krishan collaborated to develop an interactive chatbot for user integration on the site, complemented by visualizations created by Nikhita.

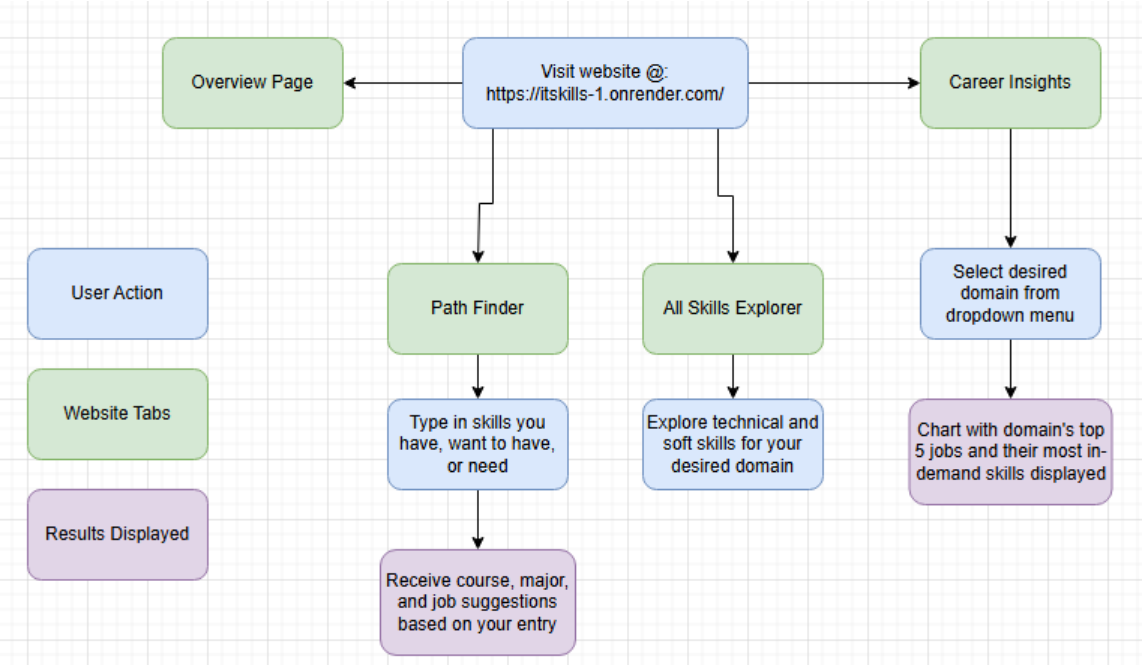
Roles

Name	Role
Michelle Webb	Project Manager & Data Analyst
Nikhita	Visualizer & Project Documenter
Krishan Bhalsod	Data Analyzer & Modeler
Lucas Leon	Visualizer & Client Liaison

Technologies Used

- **Frontend:** HTML/CSS, Vue.js
- **Backend:** Python
- **Data Analysis Tools:** Pandas, Plotly, Prince, MCA, K-Means Clustering
- **Notebooks:** Google Colab, PowerBi
- **Project Management:** Jira
- **Deployment:** Render
- **Version Control:** Git & GitHub

Flow Chart of the Project



Data Collections

- **Bright Data for Webscraping:** https://brightdata.com/
- **Final Cleaned/Merged Dataset:** https://github.com/GGC-DSA/itskills/blob/main/docs-Spring2025/final_files/merged_jobs_cleaned%20(6).csv

---

Methods Used to Clean the Datasets

- Removed null values from key columns like job_title, skills
- Filtered out senior/executive-level roles using keyword lists
- Combined titles such as developer + software developer for consistency
- Extracted keywords from job_description

Methods Used to Analyze Data and Results

Techniques Used:

- Multiple Correspondence Analysis (MCA)

- K-Means Clustering
- Skill frequency analysis – value counts
- Interactive visualizations using Plotly and Matplotlib

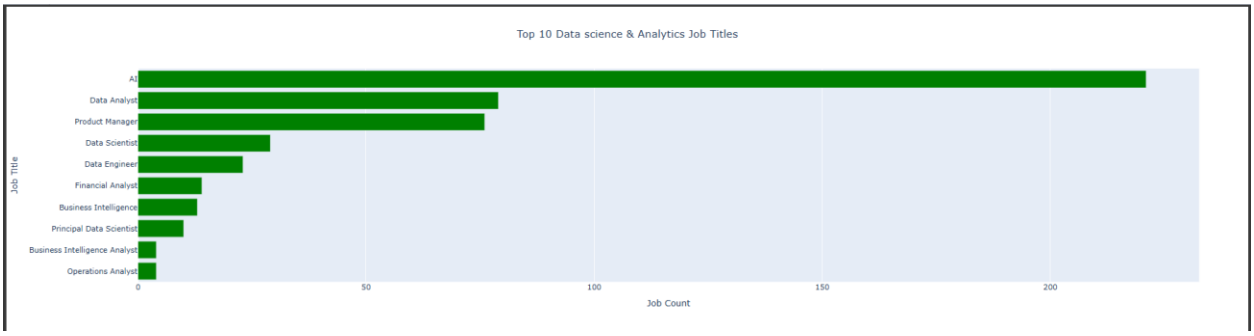
Results:

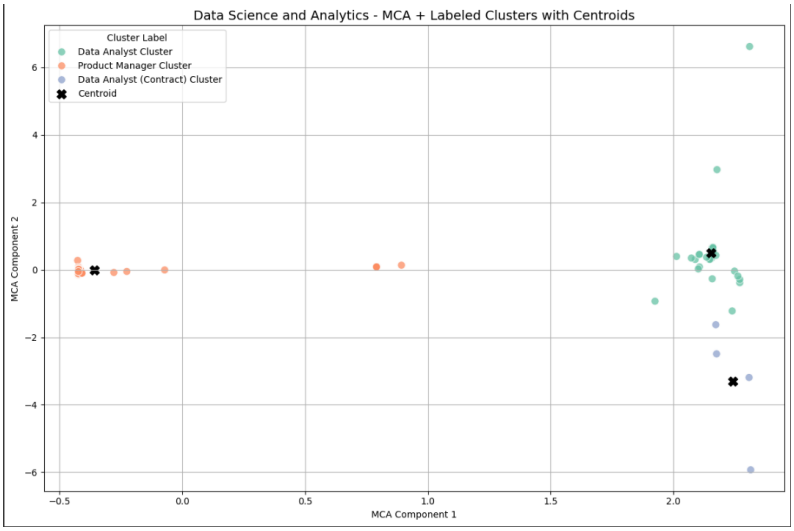
Insert summary of insights — e.g.,

- Entry-level job clusters primarily require Python, SQL, and soft skills.
- Data Science clusters were more skill-dense than Software Development.
- Clear gap between course offerings and required job skills identified.

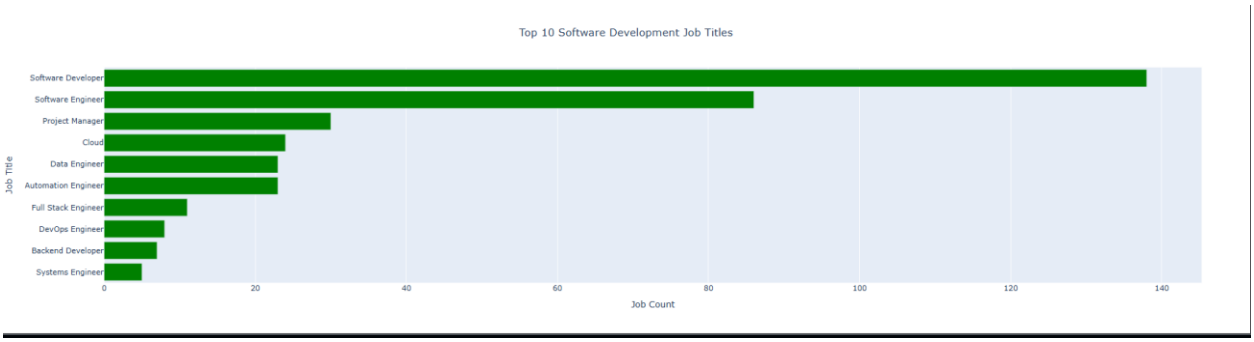
<u>Plots and Visuals:</u>

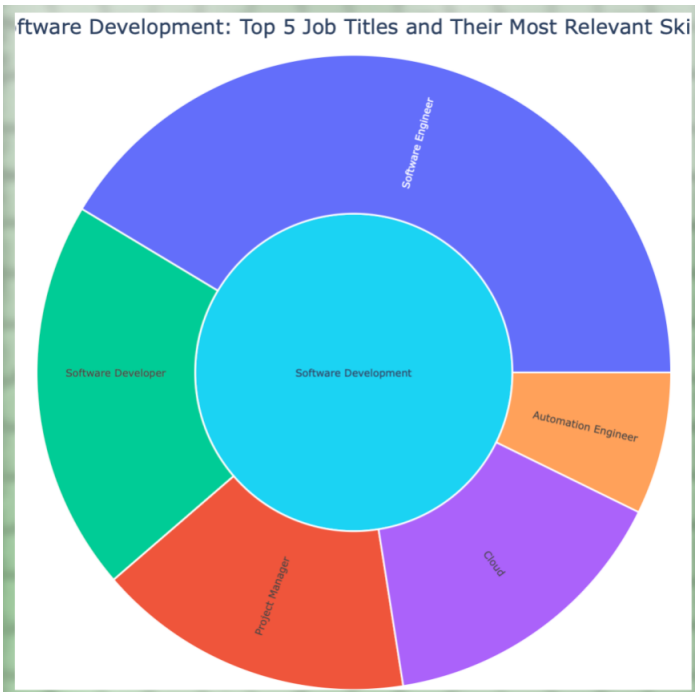
DSA:

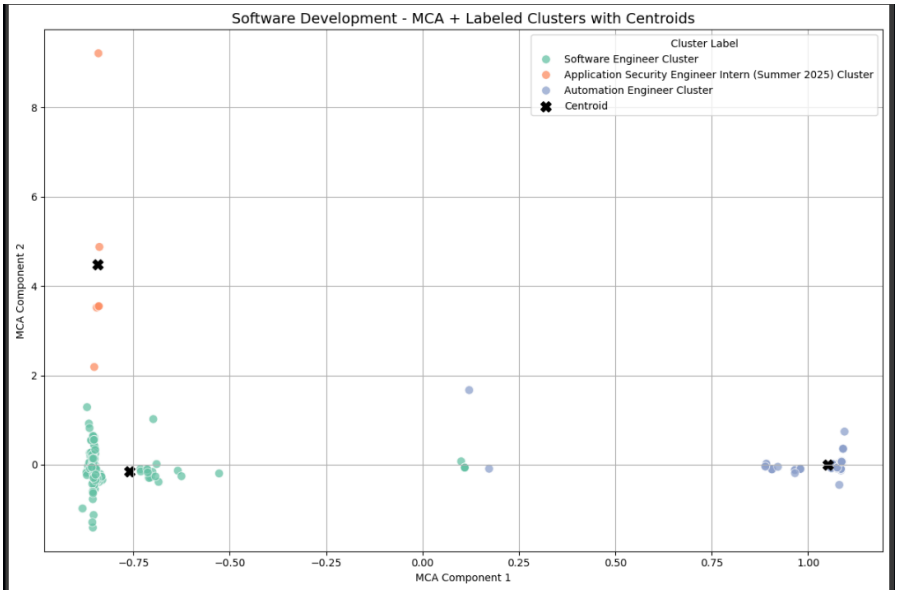




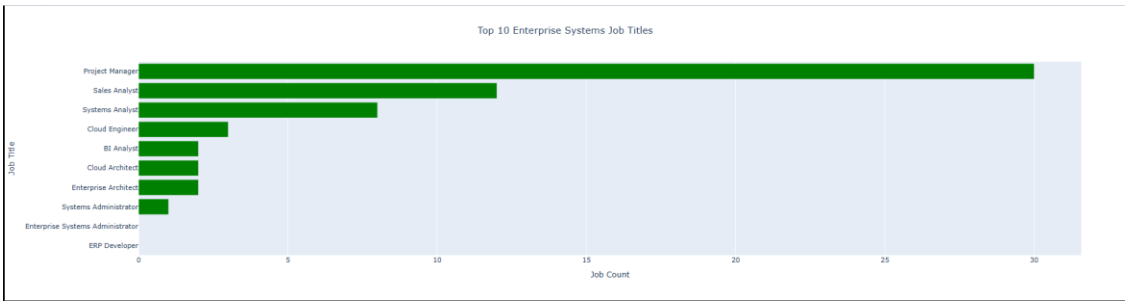
Software Development:

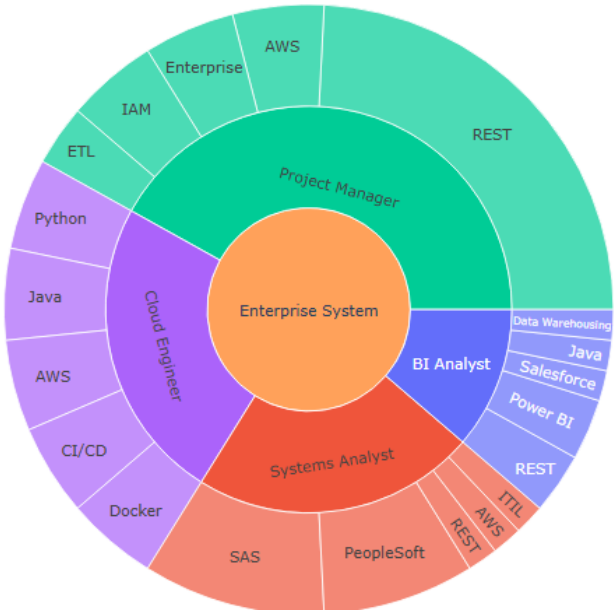


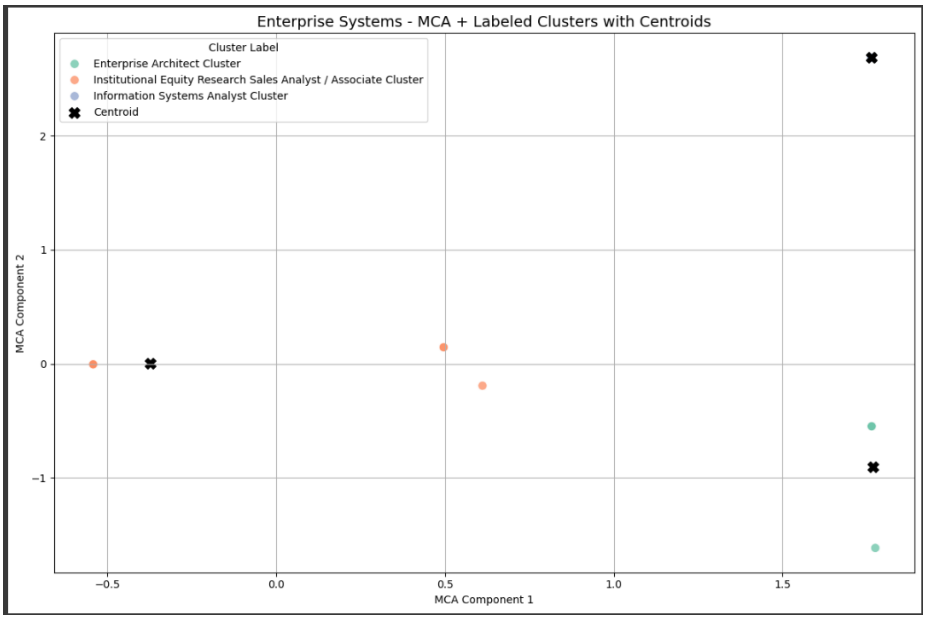




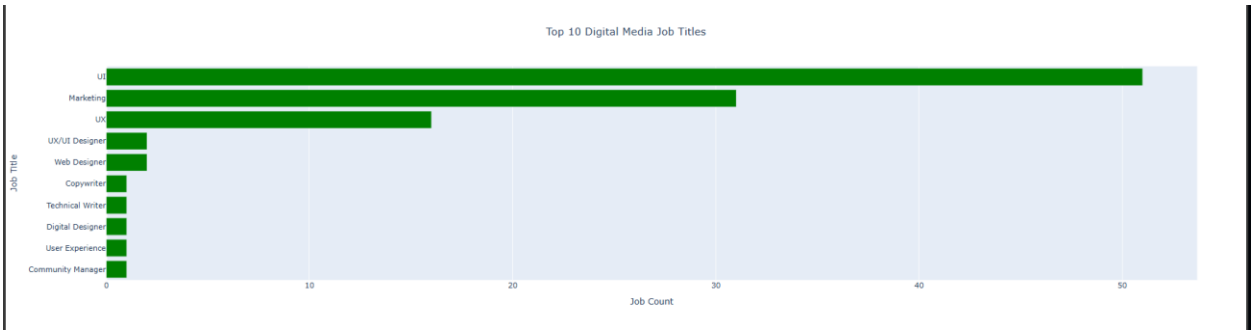
Enterprise Systems:

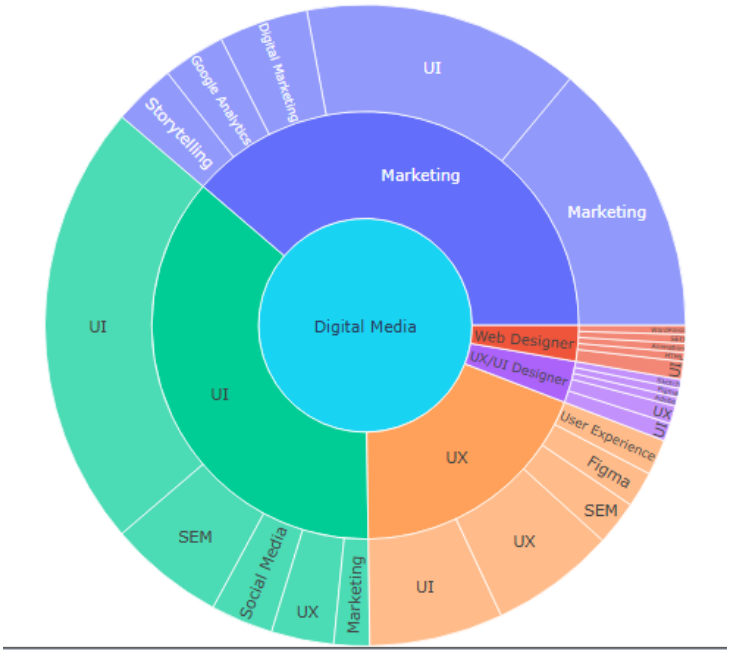


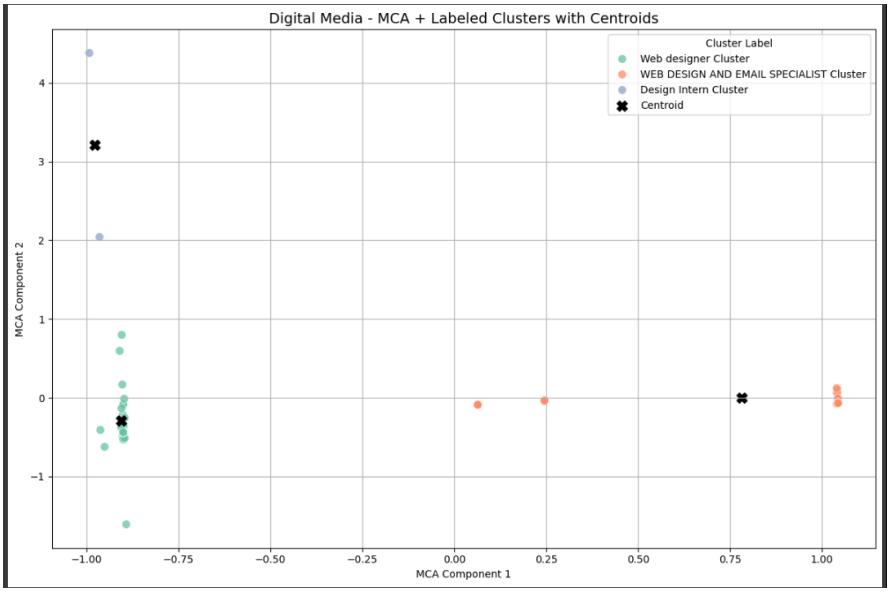




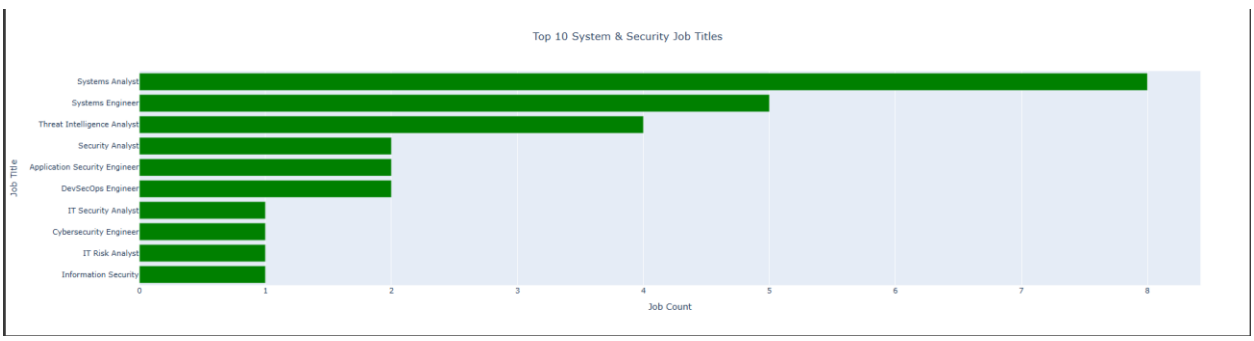
Digital Media:

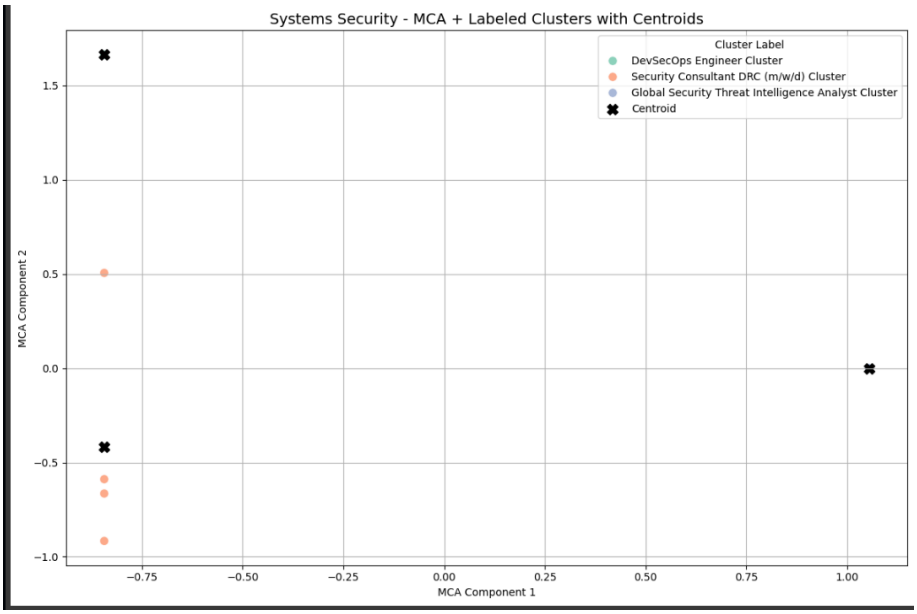






Systems and Security:





Summary of Each Iteration's Results

Iteration 1:

For iteration 1, our team webscraped data from Glassdoor, LinkedIn, and Indeed. We then began to do some general cleaning to ensure this data was valuable and would provide the insights we were looking for. We made some general graphs to see what the initial results looked like, then continued refining them in future iterations. We also created a skills survey to gather more data on Enterprise Systems courses based on the Fall '23 remaining scope and wrote our website requirements to determine what we needed to work toward in future iterations.

Iteration 2:

Iteration 2 was focused on analyzing the data and drilling down more insights. We integrated previous data with new data to create a more comprehensive master dataset, did more cleaning such as filtering out senior job titles and higher-level jobs. We created our first visualizations based on our own input to determine what it should look like, particularly for the sunburst chart that describes technical /soft skills. Our visualizer Nikhita also began designing an interactive dashboard with Power BI and Lucas and Krishan worked closely with one another to develop the logic for the suggestion chatbot integrated into the website.

Iteration 3:

The third iteration was all about finalizing the website design or tweaking the data (e.g. combining the developer job title with software developer title so they are the same) to

work toward deploying our website for student use! This process mostly involved making small design or logic changes for the website. Our data modeler, Krishan, designed the website and deployed it using Render.

GitHub Repo + Overview

GitHub Repository: https://github.com/GGC-DSA/itskills/tree/main

Overview:

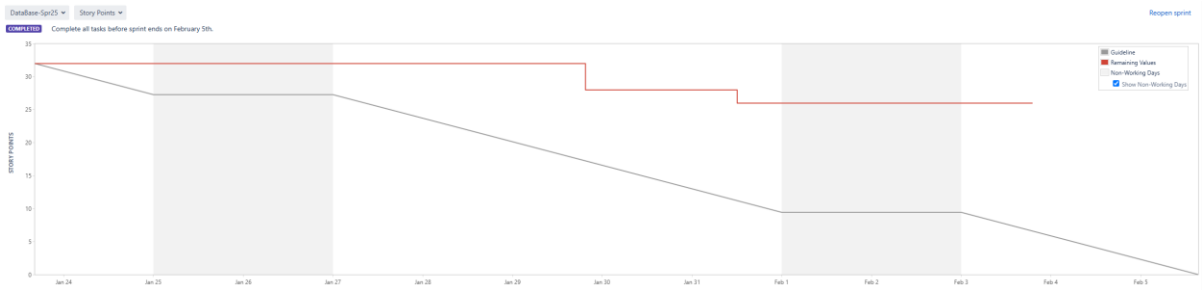
- Website linked under Spring 2025 on ReadMe
- Deployed using Render, linked on GitHub Repository
- ---

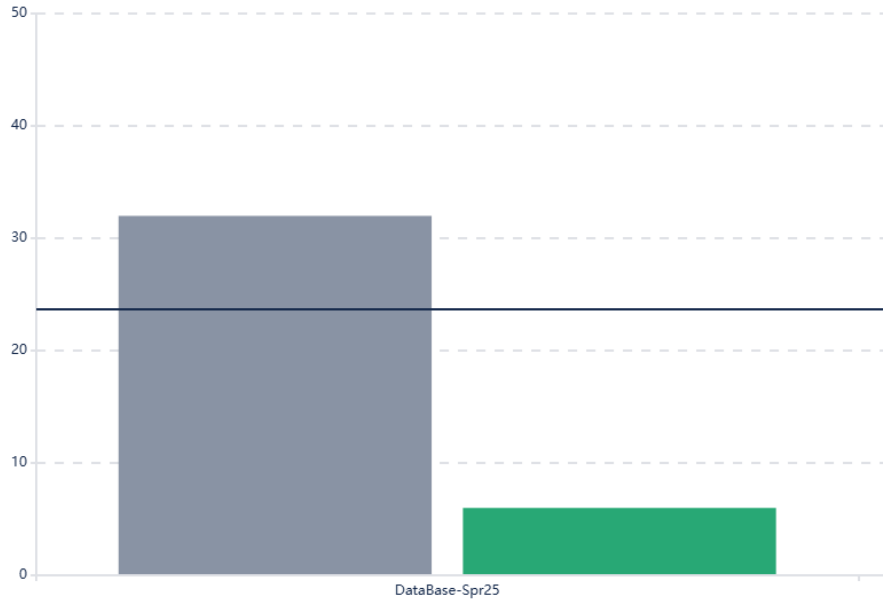
  All media, including results graphs for each major, MCA/K-Means, team photos and STaRS poster are in media/Spr ’25 folder
- Final Google Colab notebook located in code/Spr ‘25 folder

🌐 **Live Deployment Link:** https://itskills-1.onrender.com/

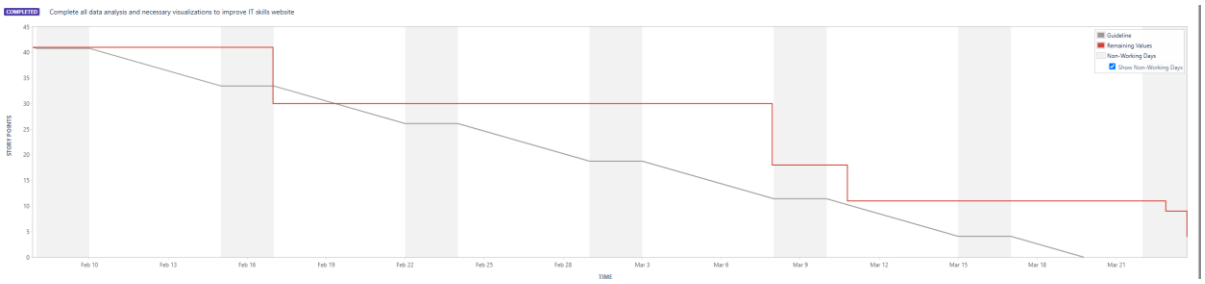
Jira Dashboard + Burndown/Velocity Charts

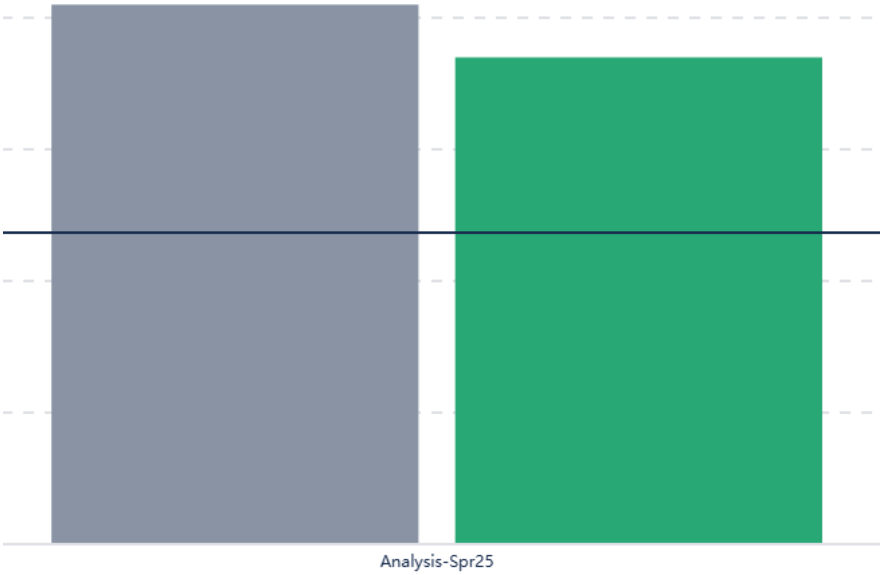
Sprint 1:



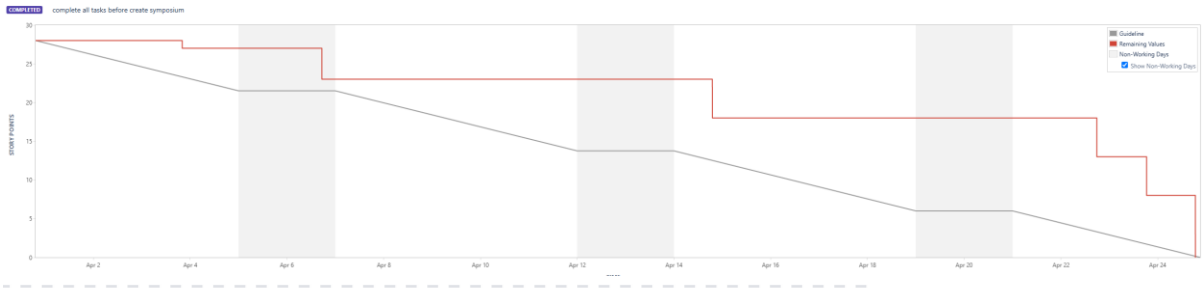


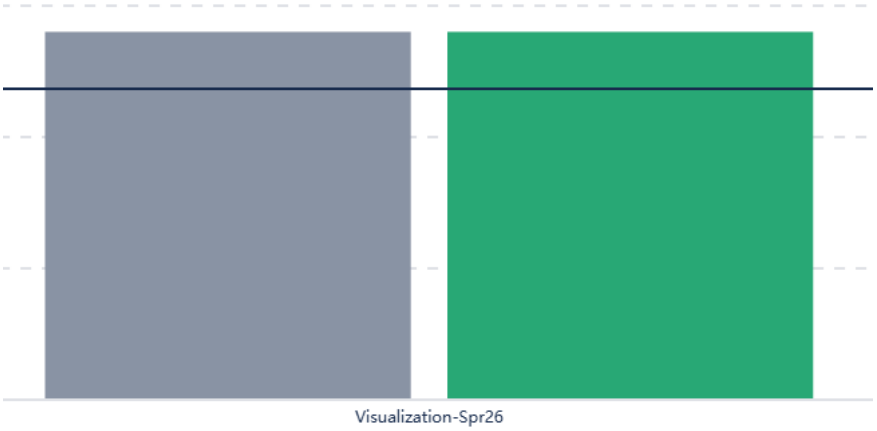
- Sprint 2


COMPLETED
Complete all data analysis and necessary visualizations to improve IT skills website
STORY POINTS
45
40
35
30
25
20
15
10
5
0
Feb 10
Feb 13
Feb 16
Feb 19
Feb 22
Feb 25
Feb 28
Mar 3
Mar 6
Mar 9
Mar 12
Mar 15
Mar 18
Mar 21
TIME
Guideline
Remaining Values
Non-Working Days
Show Non-Working Days


Analysis-Spr25

- Sprint 3:





Google Drive Demo Link/Vlog:

https://drive.google.com/file/d/1bj1rbmg5UvHdfhg2SQcOoPqxyHSE80DV/view?usp=drive_link

---

Features Implemented

- User input of desired course or skill
- Interactive charts showing job skill matches
- Dashboard with course/job skill stats

Known Issues

- Initial slow loading time upon first loading the site

- Static data, can become outdated for future use and should be regularly updated at least on an annual basis
- Full course skill list lacking

TODO Tasks (for Future Teams)

- Have a continuous influx of data with live job posting data
- Create a survey for IT faculty to gather which specific skills are taught for each class - Dr. Anca can distribute
- Include full range of courses available for domains
- Incorporate more data from other popular job sites or elsewhere
- Incorporating more interactive visuals as we could not implement our PowerBI dashboard into the website due to paywall
- Improve chatbot algorithm for better matching courses with job skills

Note: These tasks are also listed under Spring 2025 Remaining scope at the bottom of the ReadMe in the GitHub repository.