Equitable College Planner

Designed not to rank colleges by prestige, but to help students find institutions where they can thrive academically, financially, and personally.

Adam Mahmoud
Peter Dickson

UC Berkeley Datathon
16 November 2025

Agenda

01. **Background**

02 **Project Overview**

03. **Website Walkthrough**

04 **Methodology**

05 **Data Critique**

Background

College decisions are confusing: students face thousands of institutions and hundreds of variables (earnings, debt, net price, demographics, graduation outcomes).

High school counselors serve over 400+ students each → high-quality personalized guidance is rare.

Low-income students face higher barriers:

- High net price after aid
- Higher debt burdens
- Higher childcare costs for student parents
- Limited guidance resources

Existing tools (IPEDS, College Scorecard) are not personalized and often overwhelming.

Project Overview

To Address these Problems, We Built

- An interactive College Recommendation Platform powered by the Track 2 datasets.

- Provides personalized, data-driven, and equity-focused college recommendations.

- Designed to help students understand cost, debt, ROI, equity outcomes, and institutional fit.

How It Works

- User enters preferences: state, residency, income, degree level, MSI preference, locality, enrollment size, competitiveness, and faculty ratio.

- Backend pipeline filters, merges, and scores institutions based on user-chosen weights.

- Logistic similarity function + ROI model generate final rankings.

Website Walkthrough

[College Match Explorer]

Methodology

Data Sources

- *College Results dataset*: earnings, graduation rates, admissions, retention, faculty ratio, etc.

- *Affordability Gap dataset*: net price, state minimum wage, childcare-adjusted work hours to close cost gap, MSI flags, institutional type, locality.

- Cleaned, normalized, and merged via DuckDB for speed.

Similarity-Based Matching

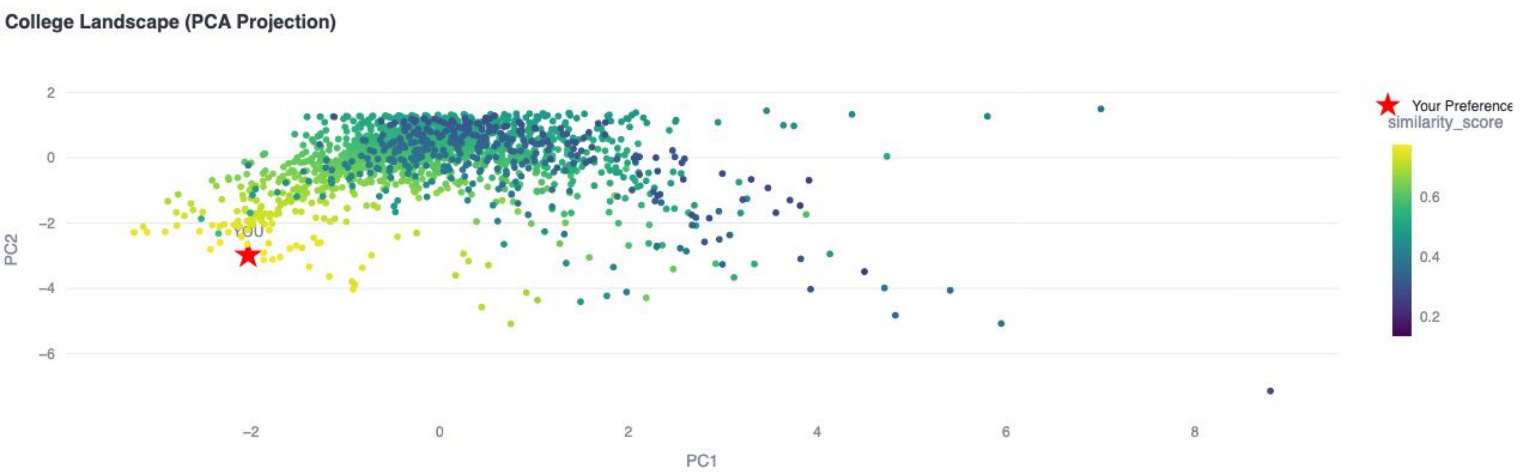
- **Filters schools first by accessibility — degree availability, residency rules, and family income bracket**
- **Uses student-defined soft preferences (sector, campus setting, enrollment size, student-faculty ratio, acceptance rate, MSI status)**
- **Numeric preferences scored using normalized distance to student's target values**
- **Categorical preferences scored using structured similarity matrices (not binary matches)**
- **MSI handling designed to support identity and belonging while preserving choice**
- **Final similarity score generated using a logistic scaling function — ranks relative fit, not "best school"**

ROI-Based Financial Model

- Computes cost based on residency: in-state, out-of-state, or net price fallback
- Estimates completion time using graduation and retention rates (bounded to avoid extreme distortions)
- Calculates expected earnings using federal median wage data + adjustment for institutional selectivity
- Applies work-burden penalty to reflect hidden affordability barriers for lower-income students
- Final ROI standardized so values are comparable across institutions
- Designed to promote financial sustainability and transparency, not prestige-based outcomes

Principal Component Analysis Graph

- Converts key quantitative attributes (enrollment size, admit rate, student-faculty ratio) into a 2D representation
- Standardizes values first so no metric dominates due to scale differences
- Student preferences are projected into the same space — visualizing personal alignment with the landscape



Data Critique

Limited Columns

- As a team we limited ourselves to only choosing a small number of columns that we believed were important to calculate similarity scores and ROI
- Drawbacks of this include potentially untested and missed information that could have bettered enabled us to give users more accurate and equitable results.
- Due to time constraints we were unable to look at whether students who have dependents could have potentially different ROIs at certain colleges than regular students
- Next Steps: Spend more time looking at more features and deciding what to include in our cleaned dataset.

High Amounts of Null Values in Features

- These datasets contained columns with high percentages of the values being null.
- Factored in the amount of non-null data in decision to keep columns
- This potentially also affected ROI scores as colleges lacked the proper data for the ROI to be calculated.
- Solution: keep as many colleges for people to choose from and leave -99.9 placeholder for Null ROI values, then order by similarity instead

Thank you