Datathon: Integrated Postsecondary Education Data System (IPEDS)

Presented by Correlation One

Problem Statement

Welcome to the 2022 Datathon! This document explains the topic of the Datathon, important details about the datasets you'll be using, and guidance on how to submit your results.

Background

Postsecondary educational institutions, private or public, are a critical component in producing educated workforces within the United States. In fact, 1.3% of total US GDP is spent each year on US postsecondary education alone. Furthermore, postsecondary education within the US has been on the rise for the past 12 years, with the percentage of U.S. adults 25-34-year-olds with a postsecondary degree increasing by 10% between 2010 and 2020, now reported to be 7% higher than the OECD average of 45%. It is without doubt that postsecondary educational institutions will be valuable organizations in educating our workforce over the ensuing decade, and thus dissecting them using data is critical to assess their strong and weak points.

Organizations like the National Center for Education Statistics (NCES) provide open access data for decision-makers & academics to analyze data surveyed on our postsecondary education systems. Systems for accessing this data offered by NCES include IPEDS, the Integrated Postsecondary Education Data System.

Today, Correlation One staff have prepared a rich and expansive compilation of longitudinal data from IPEDS providing you all with the chance to dive deep into the rich data landscape of our postsecondary educational institutions.

Your Task

You are asked to pose your own question and answer it using the available datasets, as well as any supplementary datasets that you find to aid your analysis. Both the creativity of your question, and the quality of your analysis are of paramount importance. **You need not be comprehensive; depth of insight is more important than breadth of question poses**.

Submissions may be predictive, using machine learning to classify or predict patterns. Submissions may also be illuminating by way of data visualization or sound statistical inference.

Consider exploring one of the sample questions below, or creating your own variation. Creativity in formulating your own question is encouraged; **however, it should not be at the expense of analytical depth, precision, and rigor, which are far more important**.

Sample Question 1: What roles do academic libraries play in the success of a postsecondary institution?

Sample Question 2: What variables on a postsecondary institution infer lower or higher ethnic, age, gender or veteran diversity?

Sample Question 3: Which variables of institutional programs tend to be most predictive of post graduation job attainment?

Datasets

The provided datasets are stored in the "Datathon Materials" folder on Google Drive. Your team need only use the data / datasets that are relevant to your chosen question / topic.

The full data consists of 26 longitudinal datasets that holistically cover almost all surveyed information on US postsecondary education institutions between 2015-2021. Tables are broken down into higher level categories (e.g. Institutional Finances vs Graduation Rates) containing various tables that stratify the categories by different features.

12-Month Enrollment

EFFY_2015-2021_data.csv –12-month headcount broken down by race/ethnicity, gender and level of student.

EFIA_2015-2021_data.csv – 12-month instructional activity data.

Academic Library

AL_2015-2020_data.csv – Information on Academic Libraries per fiscal year.

Admissions and Test Scores

ADM_2015-2021_data.csv – Admission considerations, applications, admissions, enrollees and test scores.

Completions

C_A_2015-2021_data.csv – Awards/degrees conferred by program (6-digit CIP code), award level, first/second major, race/ethnicity, and gender.

C_B_2015-2021_data.csv – Number of students receiving awards/degrees broken down by race/ethnicity and gender.

C_C_2015-2021_data.csv – Number of students receiving awards/degrees, by award level, gender, race/ethnicity and further stratified by age categories.

CDEP_2015-2021_data.csv – Number of programs offered and number of programs offered via distance education, by award level.

Employees by Assigned Position

EAP_2015-2020_data.csv – Number of staff by occupational category, faculty and tenure status.

Fall Enrollment

EFA_2015-2020_data.csv – Race/ethnicity, gender, attendance status, and level of student.

EFA_DIST_2015-2020_data.csv – Distance education status and level of student.

EFB_2015-2020_data.csv – Age category, gender, attendance status, and level of student.

EFC_2015-2020_data.csv – Residence and migration of first-time freshman.

EFD_2015-2020_data.csv – Total entering class, retention rates, and student-to-faculty ratio.

Graduation Rates

GR_2015-2021_data.csv – Graduation rate data of individuals that took 150% of normal time to complete for cohorts of 2 & 4 year institutions.

GR_2015-2021_data.csv – Graduation rate data, 150% of normal time to complete (less-than-2-year institutions).

Institutional Characteristics

HD_2015-2021_data.csv – Directory information for IPEDS

IC_2015-2021_data.csv – Educational offerings, organization, services and athletic associations

IC_AY_2015-2021_data.csv – Student charges for academic year programs

IC_PY_2015-2021_data.csv – Student charges by program (vocational programs)

Institutional Finances

F_F1A_1415-1920_data.csv – Public institutions - Government Accounting Standard Boards Information (GASB 34/35) fiscal years.

F_F2_1415-1920_data.csv – Private not-for-profit institutions or Public institutions using FASB fiscal years.

F_F3_1415-1920_data.csv – Private for-profit institutions fiscal years.

Outcome Measures

OM_2015-2021_data.csv – Award and enrollment data at four, six and eight years of entering degree/certificate-seeking undergraduate cohorts from degree-granting institutions, by Pell status

Student Financial Aid

SFA_1415-2021_data.csv – Student financial aid and net price data.

SFAV_1415-2021_data.csv – Military Servicemembers and Veteran's awarded benefits.

Additional Datasets

Participants are welcome to scour the internet for their own custom datasets to supplement their analysis. All additional data used should be public and reputable. Additionally, any supplementary datasets should not exceed 1 GB unzipped (consult Correlation One's R&D team if you believe your idea is worthy of an exception).

Other Materials

We will provide you with the schema for each of the data tables in another packet.

Submissions: Content

Submissions should have two components:

1. Report – this should have two main sections:
    a. Non-Technical Executive Summary – What is the question that your team set out to answer? What were your key findings, and what are their significance? You must communicate your insights clearly – summary statistics and visualizations are encouraged to help explain your thought process
    b. Technical Exposition – What was your methodology / approach towards answering the questions? Describe your data manipulation and exploration

process, as well as your analytical and/or modeling steps. Again, the use of visualizations is highly encouraged when appropriate.

2. Code – please include all relevant code that was used to generate your results. **Although your code will not be graded, you <u>MUST</u> include it, otherwise your entire submission will be discarded.**

Additional information (e.g. roadblocks encountered, caveats, future research areas, and unsuccessful analysis pathways) may be placed in an appendix.

Judges will be evaluating your technical report without your team there to explain it; therefore, **your submission must "speak for itself".** Please ensure that your main findings are clear and that any visualizations are functionally labeled.

<u>Submissions: Evaluations</u>

The competition will have multiple rounds of evaluation. Your Report will judged as follows:

- **Technical Executive Summary**
  - *Insightfulness of Conclusions.* What is the question that your team set out to answer, and how did you choose that question? Are your conclusions precise and nuanced, as opposed to over-generalizations?
- **Technical Expositions**
  - *Wrangling & Cleaning Process.* Did you conduct proper quality control and handle common error types? How did you transform the datasets to better use them together? What sorts of feature engineering did you perform? Please describe your process in detail within your Report.
  - *Investigative Depth.* How did you conduct your exploratory data analysis (EDA) process? What other hypothesis tests and ad-hoc studies did you perform, and how did you interpret the results of those tests and analyses? What patterns did you notice, and how did you use these to make subsequent decisions?
  - *Analytics & Modeling Rigor.* What assumptions and choices did you make, and how did you justify them? How did you perform feature selection? If you build models, how did you analyze their performance, and what shortcomings do they exhibit? If you constructed visualizations and/or conducted statistical tests, what was the motivation behind the particular models you build, and what did you tell you?

<u>Submissions: Formats</u>

Reports can be produced using any tool you prefer (Python Notebook, Shiny Application, Microsoft Office, etc.); however, **your report <u>MUST</u> be in a universally accessible and readable format (HTML, PDF, PPT, Web link)**. It must not require dedicated software to open. For example, if your report is a Python Notebook, it should be exported to HTML. If you create a Shiny App, it should be published at an accessible Web link.

Please include the source file used to generate your report. For example, if you submit a PDF with math-type, equations, or symbols please include your LaTeX source file.

Code should be submitted in a single zipped collection of files separate from your report.

Your team will be sent a Google Form at the beginning of the competition; you will use this form to submit and upload your content. **Submissions MUST be received by Sunday, December 4th 11:59pm ET. Any submissions received after that time will NOT be evaluated by the judges.**

Tips and Recommendations

Since this is a virtual event and you will have almost a week to work on your submissions, you should start thinking early about what problem you want to solve. The outcome of this Datathon, and your overall success, will largely be a product of how well you planned prior to the event, and the insightfulness of the problem that you chose to solve.

For data engineering, exploration, and modeling, we highly recommend that you install Jupyter Notebook: http://jupyter.org/install.html. Jupyter Notebook is an interactive, real-time development environment that eliminates many pain points of the standard "terminal & text editor" environment, and is compatible with both Python and R.

We also recommend that your team stick to tools and techniques that you have previously used. Learning new skills is certainly valuable, but it can consume a large portion of your available time, leaving less time for completing the task at hand.

We've compiled 3 additional commonalities of successful teams and 3 pitfalls that successful teams will actively avoid. Of course, these may not apply to every team, so we recommend that you and your team apply any tips accordingly.

Tips for Success	Try to Avoid
1. Focus on hypothesis testing when brainstorming your research question	**1.** Do not try to exhaust all different models you know just to yield an ideal cross validation accuracy
2. Spend at least 3 hours on your report to ensure strong communications through both visualizations and writing	**2.** Do not violate assumptions of statistical models. Sometimes, specific models require specific features so it is best to make sure those conditions are sufficiently met
3. Engage in proper causal analysis. Just because your model passes standard cross-validation checks it does not demonstrate (or even suggest) causality	**3.** Do not pick research statements and blindly stick to it trying to get it to work. Often times, further data exploration will show that it's not true or worthwhile

Ask for Help

Correlation One's R&D team is here to help. Let us know about your struggles as early on as you can and we may be able to offer advice on how to best move forward.