Novel Generation of Purely Synthetic Educational Dataset Effective For Research Applications

A Preprint

Austin Nicolas *
Department of Biology
Grinnell College
Grinnell, IA 50122
nicolasa@grinnell.edu

December 13, 2024

Abstract

The research applications of currently available education datasets are limited by ethical concerns due to privacy loss revealing sensitive information about the students in the datasets. Current educational research uses machine learning to augment and analyze the currently available datasets. In parallel, in silico research is developing purely synthetic datasets generated based on empirical datasets. When carefully handled, these datasets can effectively limits privacy loss concerns. A purely synthetic dataset was developed for use in educational research by sampling from known distributions and creating novel mappings (Nicolas & Sakib, 2024). However, no testing was done to compare the dataset to the original sampling distributions. This work examines the dataset to see if data generation aligned with sampling distributions and if weighted mappings effectively correlated features. We found that the generative function mostly sampled and mapped effectively to create significant correlations and distributions. Thus, the purely synthetic dataset mostly matches empirical data when such correlations and distributions were encoded in the generation function and the dataset, with some changes, can be used for educational research cases without privacy loss.

***K*eywords** synthetic data generation · educational research

1 Introduction

In the past 20 years, educational research has begun using machine learning and artificial intelligence for a wide variety of purposes. The most common uses are for computer assisted instruction, virtual reality, tutoring systems, and augmented reality, which combined make up over 70% of recent educational research in artificial intelligence and machine learning (Guan et al., 2020). However some educators warn that data driven approaches cannot replace educator experience and intuition, raising concerns about overreliance on artificial intelligence and machine learning outputs for determining educational policy and practices (Mandinach & Schildkamp, 2021). These works almost always use empirical data which can leak private information about students despite data privatization practices.

Synthetic datasets have been uplifted as an alternative to empirical data for data augmentation purposes and to limit privacy loss. However, synthetic datasets are not automatically private, with privacy leakage risks of the data used to generate the synthetic dataset (Jordon et al., 2022). Thus, purely synthetic datasets are not based on specific empirical datasets have arisen as a solution to privacy loss concerns. Recent in silico work has used purely synthetic data in a wide variety of ways, including improving machine learning algorithm

*The author's previous work that generated the dataset analyzed in this work can be found here https://github.com/austineamonn/SummerResearch2024

training speed (Gonsior et al., 2021), restoring resolution to degraded low resolution images (Wang et al., 2021), and annotating medical images (Liu et al., 2024).

One recent study used machine learning to predict future courses that university students would be interested in based on previous classes taken (Shao et al., 2021). More recently, a similar study was conducted that used a purely synthetic dataset based on known distributions and novel feature mappings rather than real world data that could be traced back to real students (Nicolas & Sakib, 2024). Nicolas and Sakib developed a novel technique to generate the purely synthetic educational dataset for educational research purposes. Nicolas and Sakib focused on the privacy loss implications under several data privatization methods rather than the accuracy of the purely synthetic dataset. This work takes the dataset and examines if the generating function worked as the authors had intended.

We hypothesize that the data generation function from Nicolas & Sakib (2024) functioned as intended. To support this hypothesis, we test whether observed feature distribution fits the sampled empirical distributions and whether features are correlated with eachother. Our first prediction is that dataset feature distributions will match the empirical sampling distribution. Next, we predict that the explicitly mapped features will be correlated. We also predict that the converse of the second prediction will be true. In other words, all features that were not explicitly mapped together will not be correlated.

Hypothesis	The data generation function worked as intended.
Prediction 1	Dataset feature distributions will match the empirical sampling distribution.
Prediction 2	Explicitly mapped features will be correlated.
Converse of Prediction 2	Non explicitly mapped features will not be correlated.

Table 1: Hypothesis and Predictions.

2 Methods

First, initial data preparation was done on the two datasets and then functions were developed to streamline the analysis process. Finally, Demographics, Majors, and Learning Styles were analyzed to see if the observed distribution matched the expected distribution and if features were correlated with one another. Here is a quick outline of this section:

1. Datasets (2.1)
2. Data Preparation (2.2)
3. Demographics (2.3)
4. Major (2.4)
5. Learning Style (2.5)
6. Complex Correlations (2.6)

Before we explain the methodology of the paper we quickly overview the datasets involved in the work.

2.1 Datasets

There were two datasets used in this analysis, both of which were generated in a paper by Nicolas & Sakib (2024). The *Purely Synthetic Education Dataset* was created by a novel generating function while the *Majors Dataset* was created by webscrapping data from "College Majors Explorer" (n.d.).

2.1.1 Purely Synthetic Education Dataset

The *Purely Synthetic Education Dataset* was generated by sampling from known distributions or randomly selecting when distributions were unknown. Additionally, novel mappings were built between different columns to reflect known connections (Nicolas & Sakib, 2024). The flowchart (Figure ref{fig:data_generation}) shows the process used for generating each observation in the dataset.

Table 2 from Nicolas & Sakib (2024) showcases the different features of the dataset. In this work, **Ethnoracial Group**, **Gender**, **International Student Status**, **Socioeconomic Status**, **GPA**, **Learning Style(s)**, **Student Semester**, **Major(s)**, **Previous Course Types**, and **Previous Course Subjects**, were analyzed.

Table 3 shows the first row of the dataset. The row was transposed and the long string list columns were concatenated to prevent the row from running of the page.

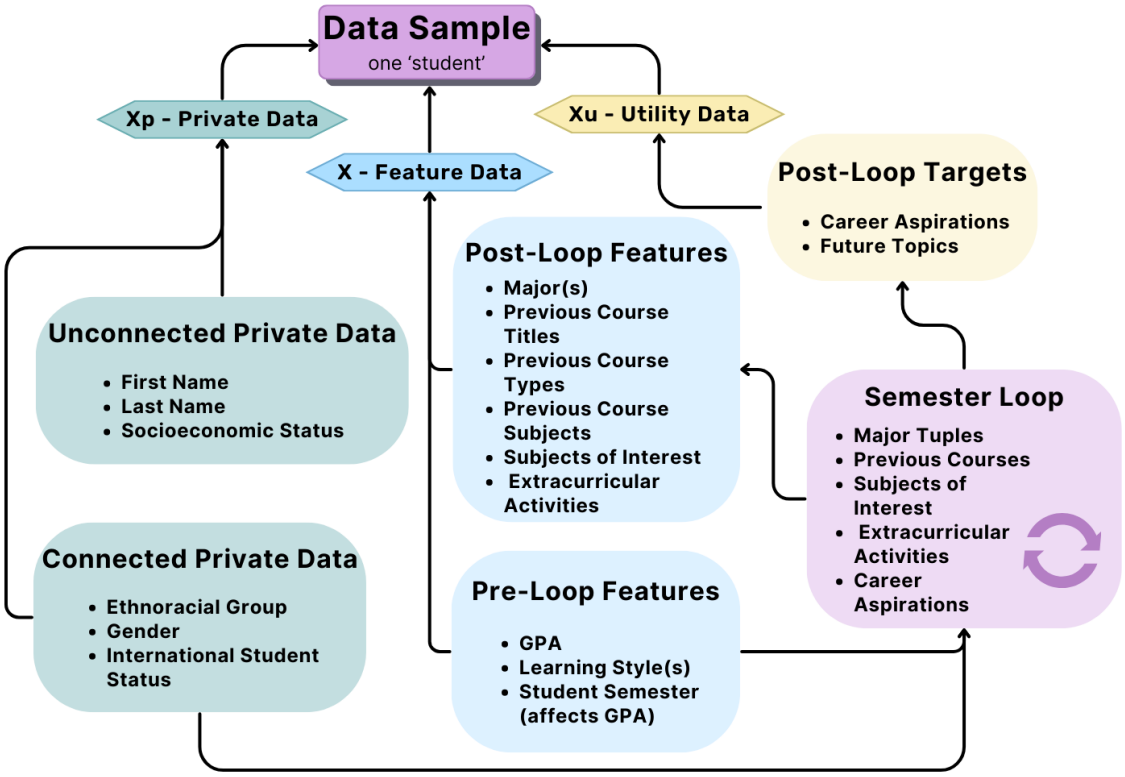

Figure 1: This data generation flowchart was taken from (Nicolas & Sakib, 2024).

Features	Data Category	Data Type	Number of Options
First Name	X_p	String	32,952
Last Name	X_p	String	151,671
Ethnoracial Group	X_p	String	7
Gender	X_p	String	3
International Student Status	X_p	String	2
Socioeconomic Status	X_p	String	5
GPA	X	Double	201
Learning Style(s)	X	List of Strings	4
Student Semester	X	Integer	15
Major(s)	X	List of Strings	173
Previous Course Titles	X	List of Strings	3,476
Previous Course Types	X	List of Strings	21
Previous Course Subjects	X	List of Strings	236
Subjects of Interest	X	List of Strings	141
Extracurricular Activities	X	List of Strings	304
Career Aspirations	X_u	List of Strings	149
Future Topics	X_u	List of Strings	226

Table 2: Information about Data Features.

Column Name	**First Entry**
First name	Raycen
Last Name	Tallini
Ethnoracial Group	European American or white
Gender	Female
International Student Status	International
Socioeconomic Status	Middle Income
Learning Style	['Read/Write']
GPA	3.86
Student Semester	10
Major	['Teacher Education: Multiple Levels']
Previous Courses	['Intensive Elementary Spanish', 'General Chemistry Lab II', ..., 'Advanced Study Abroad']
Course Types	[['Laboratory-Discussion', 'Lecture', 'Practice'], ['Lecture-Discussion', 'Online'],..., ['Discussion/Recitation', 'Lecture']]
Course Subjects	['ACES', 'EPOL', 'GWS',..., 'ART']
Subjects of Interest	['History', 'Philosophy',..., 'Engineering']
Extracurricular Activities	['Education Society', 'Engineering Society', ..., 'Electrical Engineering Club']
Career Aspirations	['Postsecondary Teacher', 'Other Teacher and Instructor', ..., 'Material Engineer']
Future Topics	['Leadership Studies', 'Sociology', ..., 'Human Development and Family Studies']

Table 3: The first row of the dataset, transposed. Long list features concatenated.

2.1.2 Majors Dataset

The Majors Dataset was stored as a JSON file so it was imported using the JSONlite package (Ooms et al., 2024). Table 4 shows the features of the *Majors Dataset* from Nicolas & Sakib (2024), who based it on a data webscrapped from "College Majors Explorer" (n.d.).

Data Column	Data Type	Data From
Major	String	Majors List
National Percent Female	float	Represents a percentage so between 0 and 1
Popularity Ranking	Integer	Between 1 and 173
Division	Factor	One of 6 Divisions
Top 5 Careers	List of Strings	Careers List

Table 4: Majors Dataset

The *Majors Dataset* contains the following major divisions: Business, Education, Health, Liberal Arts, Social Sciences, and STEM.

2.2 Data Preparation

The preprocessing pipelines for the features analyzed in this work are summarized in Table 5.

Renaming columns and converting to factors was done through Tidyverse (Wickham, 2023). Since the lists of strings in the *Purely Synthetic Education Dataset* were stored as strings, they had to be stripped and split into multiple columns. This was done for **Learning Styles(s)**, **Major(s)**, **Previous Course Types**, and **Previous Course Subjects** but the same process could be applied to the other list of string features. Then each split column was put into its own copy of the *Purely Synthetic Education Dataset* and they were recombined rowwise.

2.2.1 Dictionary

The Dict package (Asai, 2020) was used to create a dictionary with feature names as the keys and the column name for the feature as the values. The key and value for **Ethnoracial Group** are shown below:

Features	Rename Column	Convert to a Factor	Split Lists and Recombine
Ethnoracial Group	YES	YES	NO
Gender	YES	YES	NO
International Student Status	YES	YES	NO
Socioeconomic Status	YES	YES	NO
GPA	YES	NO	NO
Learning Style(s)	YES	NO	YES
Student Semester	YES	NO	NO
Major(s)	YES	NO	YES
Previous Course Types	YES	NO	YES
Previous Course Subjects	YES	NO	YES

Table 5: Data Preparation Pipeline.

Key: "Ethnoracial Group", Value: "ethnoracial_group"

2.3 Demographics

The demographic section looks at the following 6 features:

1. Ethnoracial Group
2. Gender
3. International Student Status
4. Socioeconomic Status
5. GPA
6. Student Semester

The string demographics (**Ethnoracial Group**, **Gender**, **International Student Status**, and **Socioeconomic Status**) were sampled based on empirical data distributions. The numerical demographics (**GPA** and **Student Semester**) were sampled from a uniform distribution. All the demographics were sampled independently of one another (Nicolas & Sakib, 2024).

2.3.1 Statistical Testing

The Chi Squared Goodness of Fit Test was run on **Ethnoracial Group**, **Gender**, **International Student Status**, and **Socioeconomic Status** comparing them with their known distributions. Known distributions were taken from Nicolas & Sakib (2024).

The Chi Squared Test of Independence was run between **Ethnoracial Group**, **Gender**, **International Student Status**, and **Socioeconomic Status**.

2.3.2 Graphing

For each combination of **Ethnoracial Group**, **Gender**, **International Student Status**, and **Socioeconomic Status** a bar chart comparing the proportion of one feature across the options for the other feature was created using ggplot2 (Wickham, 2023) and RColorBrewer (Neuwirth, 2022).

For **GPA** and **Student Semester** the number of students in a given semester or with a certain GPA were graphed with a horizontal line representing the expected uniform distribution of the features. This was done using ggplot2 (Wickham, 2023).

2.4 Major

The major function from Nicolas & Sakib (2024) determined a student's major based on a weighting by gender and major popularity. For a major m, let P_m be the popularity ranking of the major and let F_m be the percentage of female identifying students in the major. Since there are 173 majors we know that $P_m \in [1, 173] \cap \mathbb{N}$.

$$W_m = 1 - \frac{P_m}{173} + \begin{cases} 1 - \frac{F_m}{100} & \text{if student identifies as Male} \\ \frac{F_m}{100} & \text{otherwise} \end{cases}$$

2.4.1 Joining the Majors Dataset

A left join was used to combine the *Majors Dataset* with the *Purely Synthetic Education Dataset* where the **Major(s)** feature was split and recombined.

2.4.2 Statistical Testing

The Chi Squared Goodness of Fit Test was run on the popularity of each major vs the expected popularity ranking. The observed popularity was calculated by taking the number of students in each major (splitting double majors) and ranking them. The results were graphed in a bar graph using ggplot2 (Wickham, 2023) and viridis (Garnier et al., 2024).

The Chi Squared Test of Independence was run with **Majors(s)** against **Ethnoracial Group**, **Gender**, **International Student Status**, and **Socioeconomic Status**.

2.4.3 Gender By Major

Since **Gender** was the only feature that was explicitly called in the generating function for picking **Major(s)** then the Chi Squared Goodness of Fit test was run to compare the expected percent female of each major from the *Majors Dataset* vs what was observed in the *Purely Synthetic Education Dataset*. Then the dataset was filtered using Wickham (2023) to get the top ten majors with the highest percentage of male, female, and gender minority (female and nonbinary) students. Nonbinary students were never filtered alone since the sample size was too small. Then the results were graphed in a bar graph using ggplot2 (Wickham, 2023) and RColorBrewer (Neuwirth, 2022).

2.5 Learning Style

The learning style function from Nicolas & Sakib (2024) determined a student's learning style(s) based on randomly sampling from a weighted distribution based on the empirical distribution of the learning styles. There is also a 10% chance of a student getting a second learning style, though if it is a duplicate it is ignored.

2.5.1 Statistical Testing

The Chi Squared Goodness of Fit Test was run on the popularity of each learning style vs the expected distribution.

The Chi Squared Test of Independence was run with **Learning Style(s)** against **Ethnoracial Group**, **Gender**, **International Student Status**, and **Socioeconomic Status**.

2.6 Complex Correlations

The `pick_classes` function from Nicolas & Sakib (2024) determined a student's classes based on a list of classes with each class weighted as follows:

Class matches Student's	**Weighting**
Learning Style(s)	2
Major(s)	5
Nothing	1

Table 6: Class Weightings.

Classes are based on a student's **Majors(s)** and **Learning Style(s)**. So, a student with a math major would be more likely to take a math course. From each class the course subject was taken (ex: MAT for a math class). Additionally, a student with a Read/Write learning style would be more likely to take courses that focus on reading and writing. From each class the course type was taken (ex: Lecture for a math class), with each course type being associated with one or more learning styles. Thus, we want to determine if there are correlations between these variables.

Take the joined dataset that was already split by **Major(s)** and split it by **Course Subjects**. Then recombine using the `rbind` function from base R. Then the Chi Squared Test Test of Independence is run to determine if **Major(s)** and **Course Subjects** are correlated.

Take the dataset that was already split by **Learning Style(s)** and split it by **Course Types**. Then recombine using the `rbind` function from base R. Then the Chi Squared Test Test of Independence is run to determine if **Learning Style(s)** and **Course Types** are correlated.

3 Results

1. Demographics (3.1)
2. Major (3.2)
3. Learning Style (3.3)
4. Complex Correlations (3.4)

3.1 Demographics

First, we look at how well the demographic feature distribution fits the empirical data from which they were sampled. Then, we examine if the demographics are correlated together or if they are independent. Finally, we graph the numerical features (**GPA** and **Student Semester**) and compare them to the uniform distribution from which they were sampled.

3.1.1 Distribution Fit

Table 7 shows how well each of the feature distributions matched the empirical distribution. None of them were significant.

Demographic	**P-Value**
Ethnoracial Group	1
Gender	1
International Student Status	0.997
Socioeconomic Status	1

Table 7: Chi Squared Goodness of Fit Test. *p<0.05

3.1.2 Demographic Correlations

Table 8 shows how well each of the features correlate with each other. None of them were significant. Though **Socioeconomic Status** and **International Student Status** were close to significant at p=0.08951.

	Gender	International Student Status	Socioeconomic Status
Ethnoracial Group	0.8605	0.5301	0.5707
Gender		0.1464	0.1799
International Student Status			0.08951

Table 8: Chi Squared Test of Independence between Demographics. *p<0.05

The correlation graph (Figure 2) shows the correlation between ethnoracial group and gender and we see how similar the bars look.

The correlation graph (Figure 3) shows the correlation between ethnoracial group and international student status and we see how similar the bars look.

The correlation graph (Figure 2) shows the correlation between ethnoracial group and socioeconomic status and we see how similar the bars look.

The correlation graph (Figure 5) shows the correlation between gender and international student status and we see how similar the bars look.

The correlation graph (Figure 6) shows the correlation between gender and socioeconomic status and we see how similar the bars look.

The correlation graph (Figure 7) shows the correlation between international student status and socioeconomic status and we see how similar the bars look.

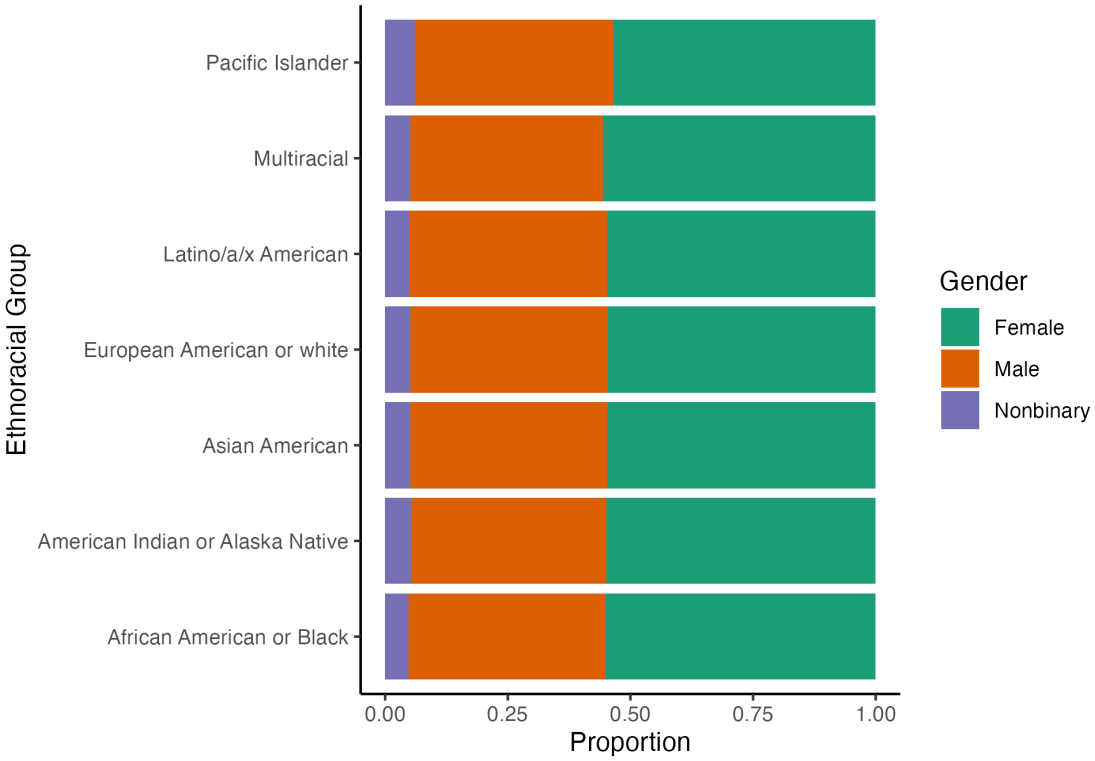

Figure 2: Correlation graph between ethnoracial group and gender.

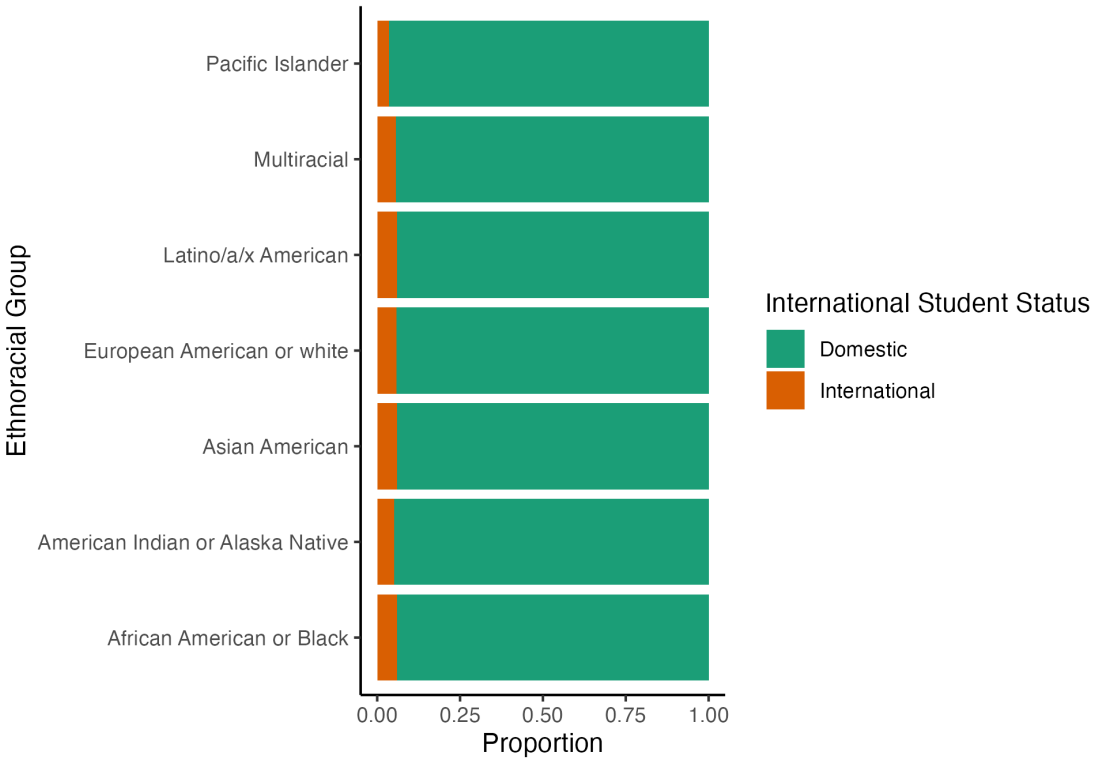

Figure 3: Correlation graph between ethnoracial group and international student status.

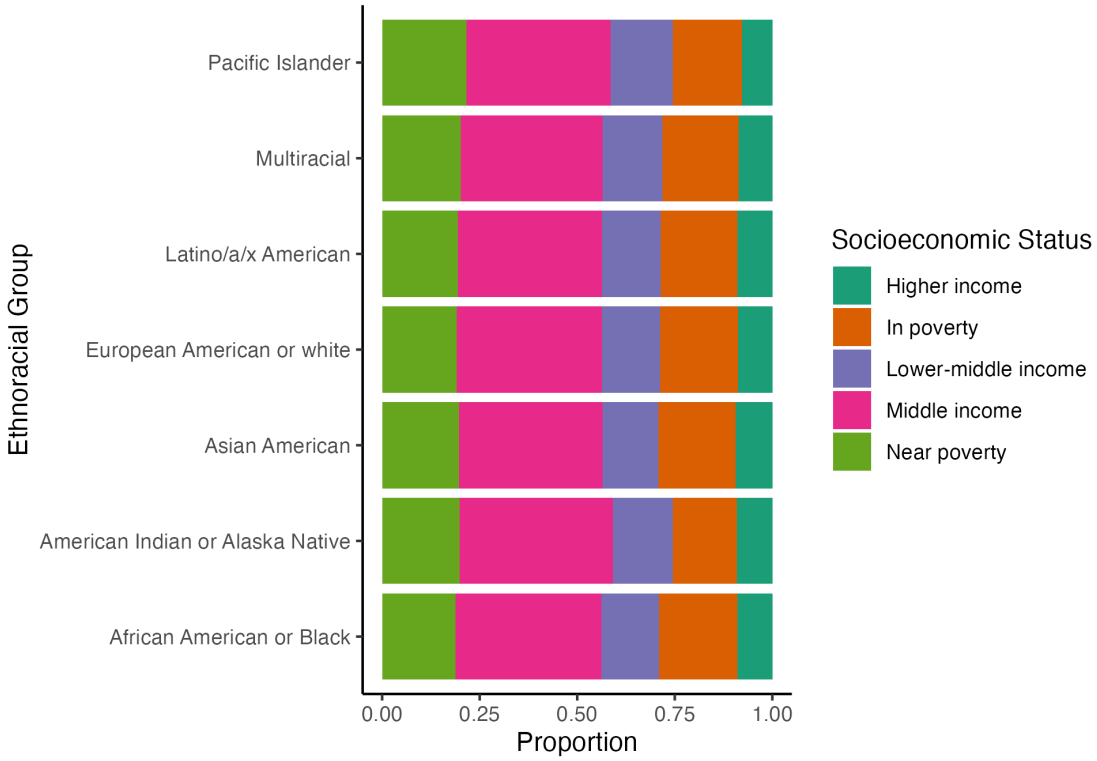

Figure 4: Correlation graph between ethnoracial group and socioeconomic status.

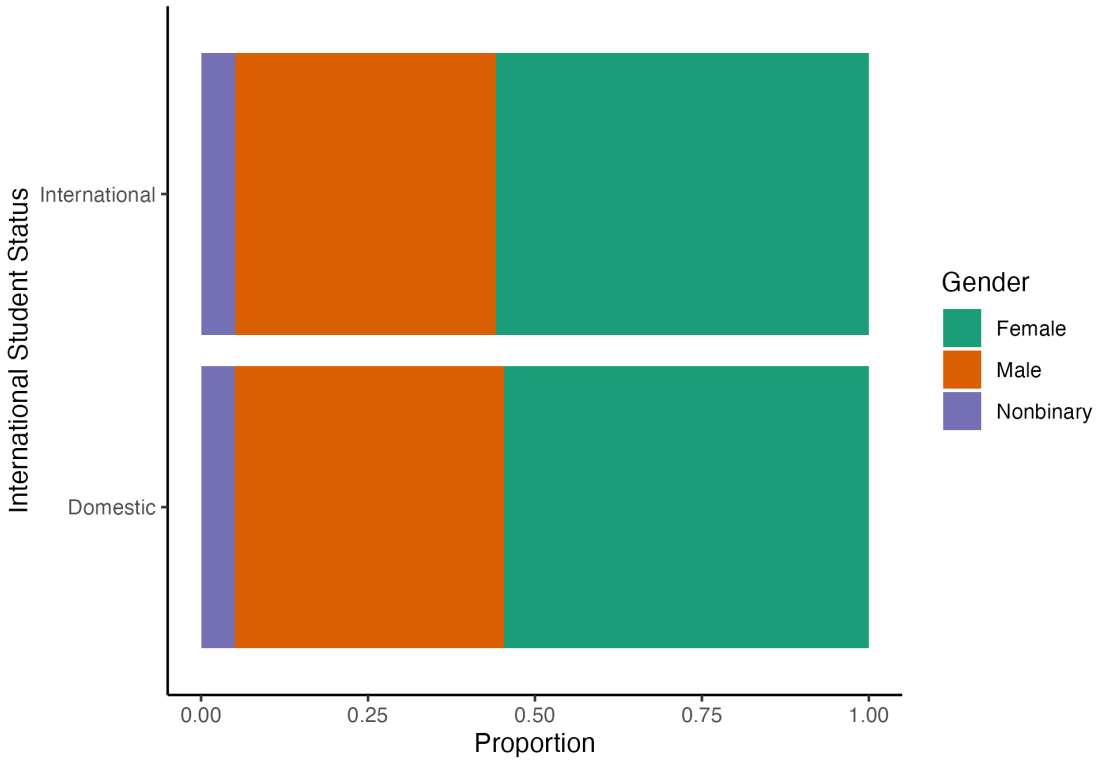

Figure 5: Correlation graph between gender and international student status.

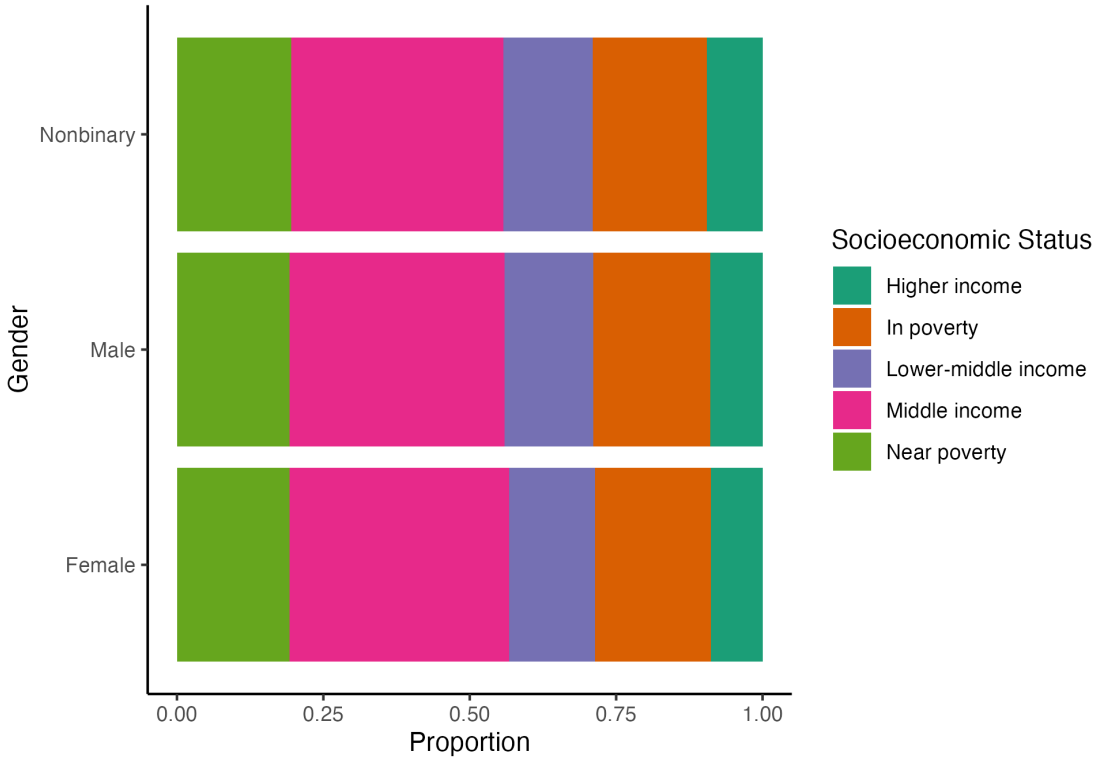

Figure 6: Correlation graph between gender and socioeconomic status.

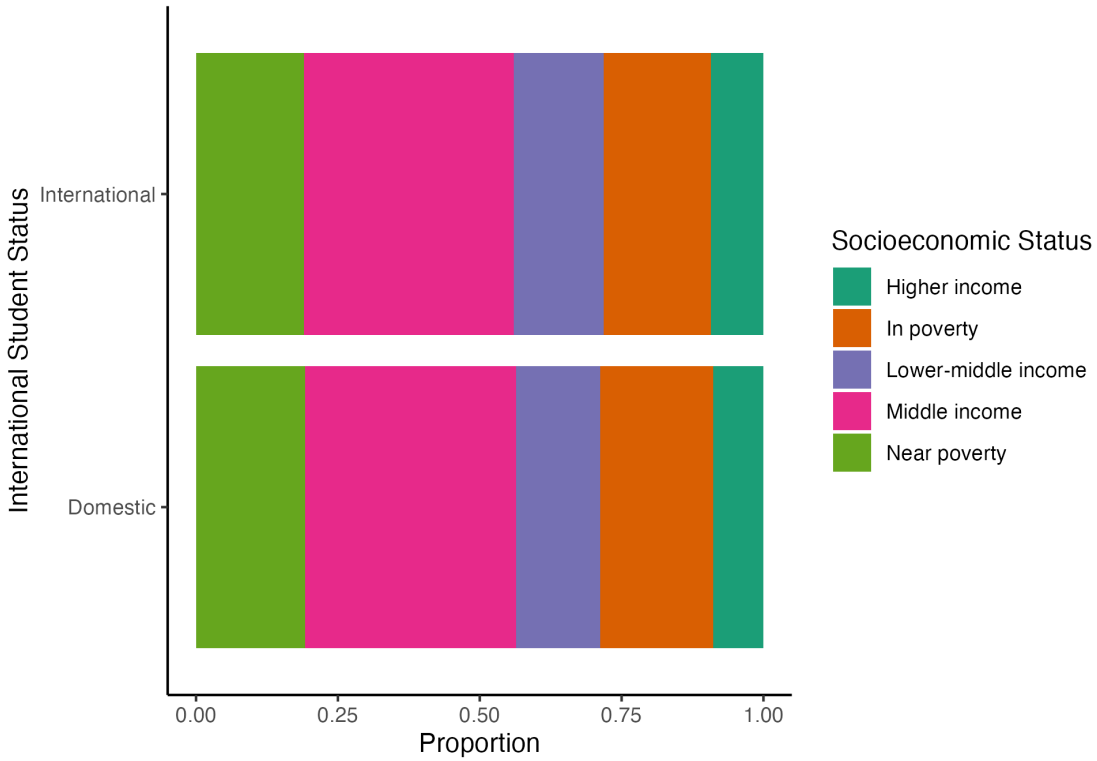

Figure 7: Correlation graph between international student status and socioeconomic status.

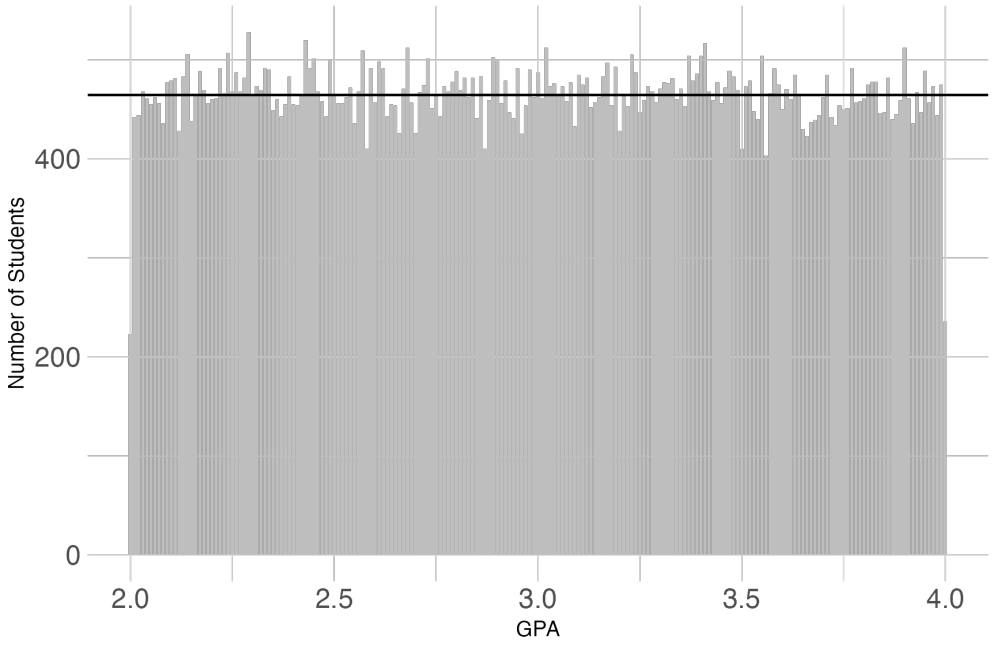

Figure 8: Number of students with different GPAs. The horizontal gray line (y=464.4527) represents the expected values based on a uniform distribution.

3.1.3 Numerical Graphs

The graphs were generated using ggplot2 (Wickham, 2023) and use a theme from hrbrthemes (Rudis, 2024).

The gpa distribution graph (Figure 8) shows the number of students with each GPA and we see how similar the distribution is to the horizontal line which represents a uniform distribution.

The student semester distribution graph (Figure 9) shows the number of students in each semester and we see how similar the distribution is to the horizontal line which represents a uniform distribution.

3.2 Major

First, we discuss the popularity of each major. Then, we cover the correlations between **Major(s)** and different demographics. Finally, we graph various distributions of major choice depending on gender identity.

3.2.1 Major Popularity

Figure 10 shows the top ten most popular majors in the *Purely Synthetic Educational Dataset*.

Figure 11 shows the majors comparing their observed and expected popularity. The Chi Squared Goodness of Fit Test between the expected and observed major popularity gave a p-value of $< 2.2e^{-16}$.

3.2.2 Major Correlations

Table 9 shows the correlations between **Major(s)** and several demographics. Only **Gender** is not independent of the **Major(s)** features.

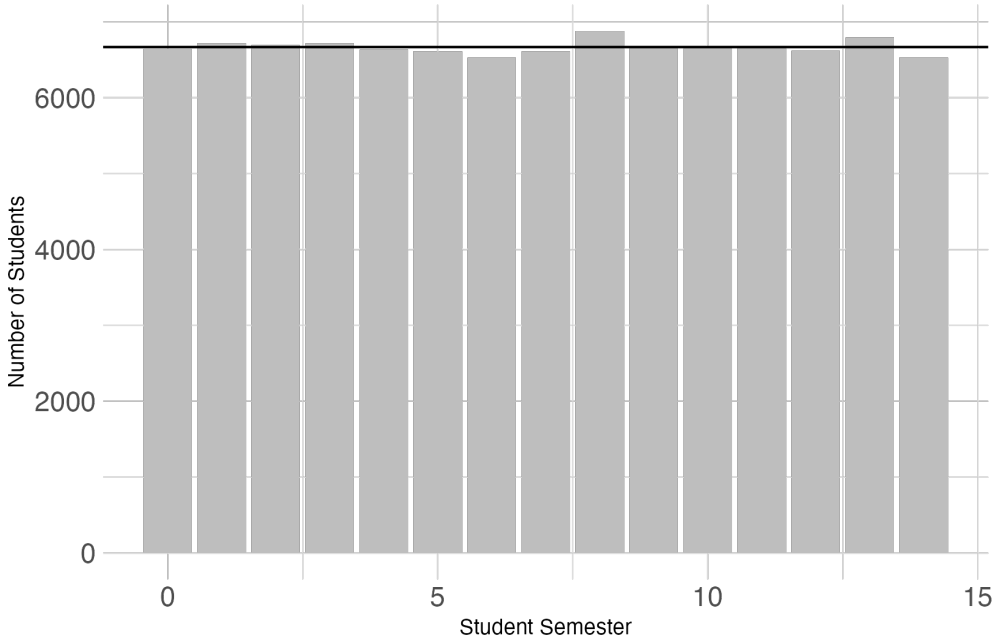

Figure 9: Number of students in each semester. The horizontal gray line (y=6666.667) represents the expected values based on a uniform distribution.

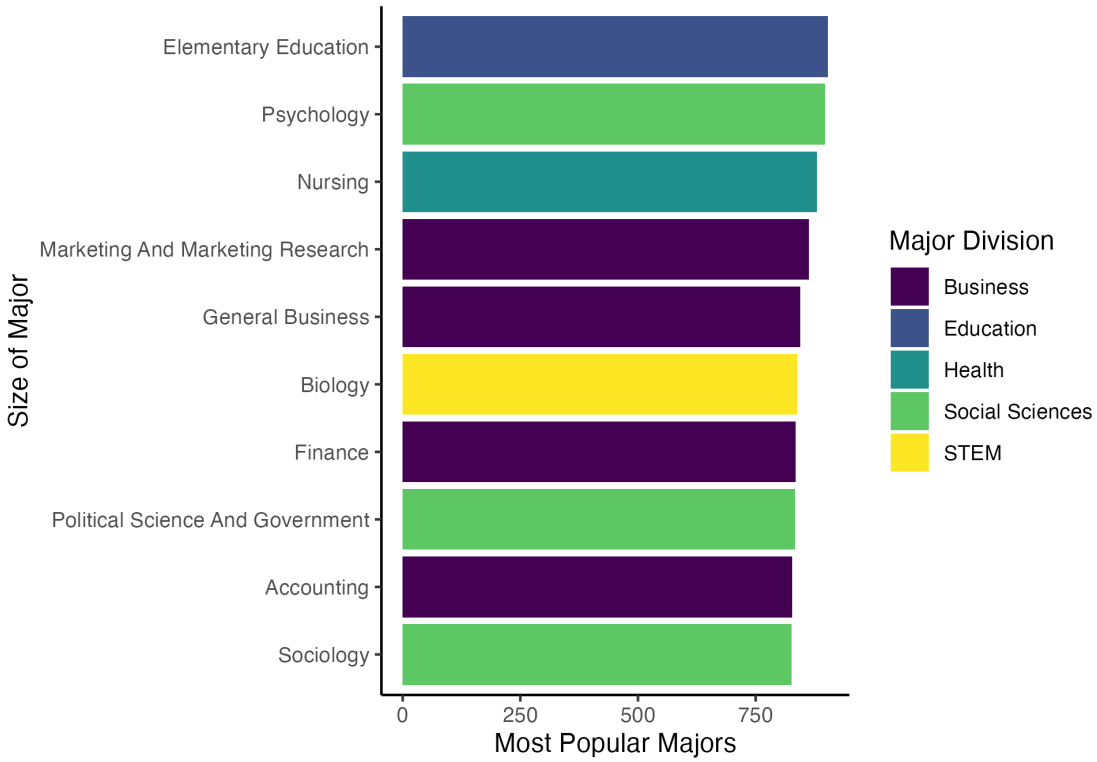

Figure 10: Top 10 majors by popularity organized by division.

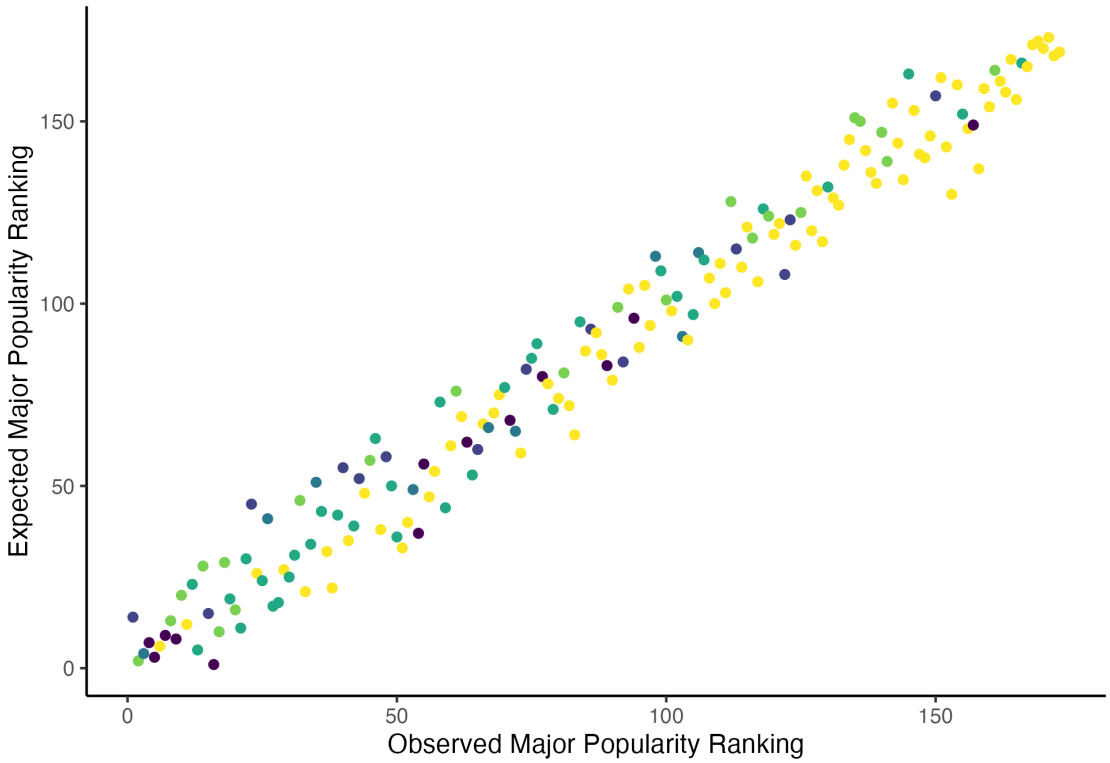

Figure 11: Observed vs expected major popularity organized by division.

Demographic	Major Correlation	Division Correlation
Ethnoracial Group	0.8966	0.1593
Gender	$< 2.2e^{-16}$*	$< 2.2e^{-16}$*
International Student Status	0.7619	0.2205
Socioeconomic Status	0.7008	0.3414

Table 9: Chi Squared Test of Independence between Major and the demographics. *p<0.05

3.2.3 Gender by Major

Figure 12 shows the expected and observed percent of female identifying students in each major. The observed (dark gray line) and expected (light gray line) linear regression are visually dissimilar. The Chi Squared Goodness of Fit Test between the expected and observed percent of female identifying students gave a p-value of 1.

Figure 12 shows the expected and observed percent of female identifying students in each major. This graph was interactive in the HTML, but that does not work in pdf and so the circles mostly cover each other since the size represents the number of students in each major.

Figure 14 shows the top ten majors that are dominated by male identifying students. All these majors are in the STEM division.

Figure 15 shows the top ten majors that are dominated by female identifying students. The majors are from a variety of divisions.

Figure 16 shows the top ten majors that are dominated by gender minority (female and nonbinary) students. The majors are from a variety of divisions.

No graph was made of nonbinary students, since there are not enough nonbinary students for any major to be majority nonbinary.

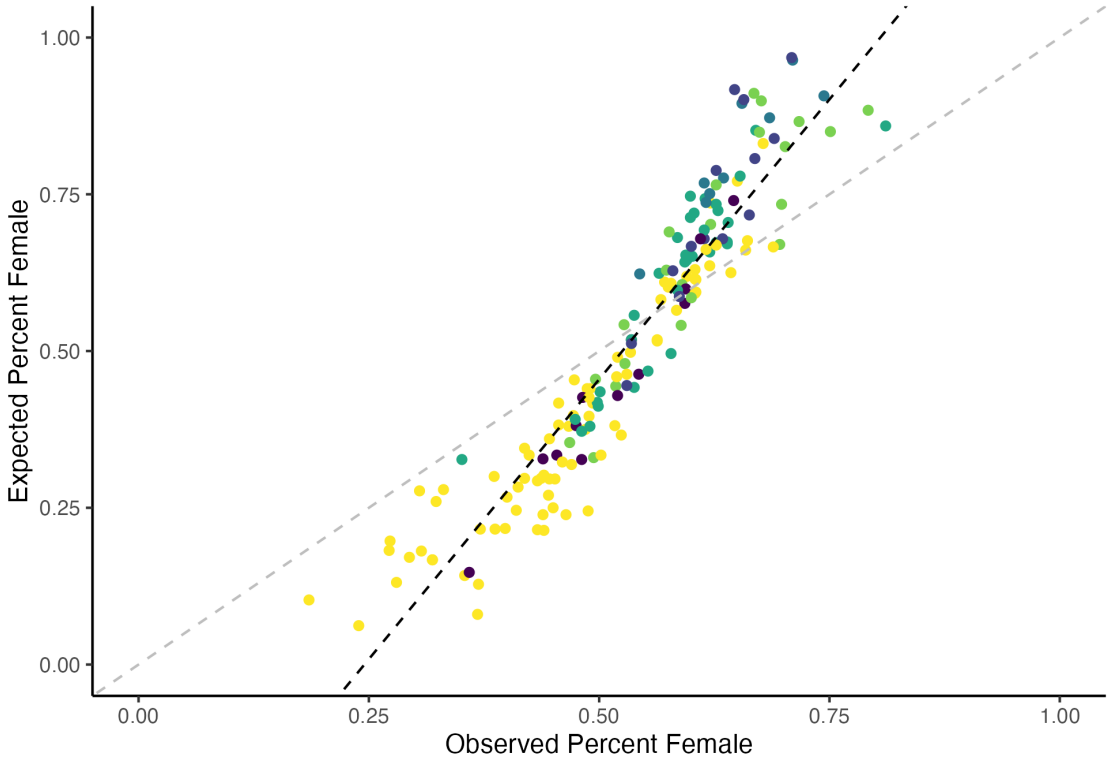

Figure 12: Expected vs observed percent female per major, colored by division. The dark gray line (y=1.7855x-0.4378) is the linear regression. The light gray line (y=x) is the expected linear regression, since expected and observed.

3.3 Learning Style

Table 10 shows the correlations between **Learning Style(s)** and the demographic features. None of them are significant, so they are all independent.

Demographic	**Learning Style Correlation**
Ethnoracial Group	0.4939
Gender	0.98
International Student Status	0.6115
Socioeconomic Status	0.6034

Table 10: Chi Squared Test of Independence between Learning Style and the demographics. *p<0.05

3.4 Complex Correlations

Table 11 shows the correlations between **Major(s)** and **Previous Course Subjects** as well as **Learning Style(s)** and **Previous Course Types**. The test between **Major(s)** and **Previous Course Subjects** was significant and so they are correlated. However, the test between **Learning Style(s)** and **Previous Course Types** was not significant, so they are independent.

Comparison	**Correlation**
Major vs Previous Course Subjects	$< 2.2e^{-16}*$
Learning Style vs Previous Course Types	0.9967

Table 11: Chi Squared Test of Independence. *p<0.05

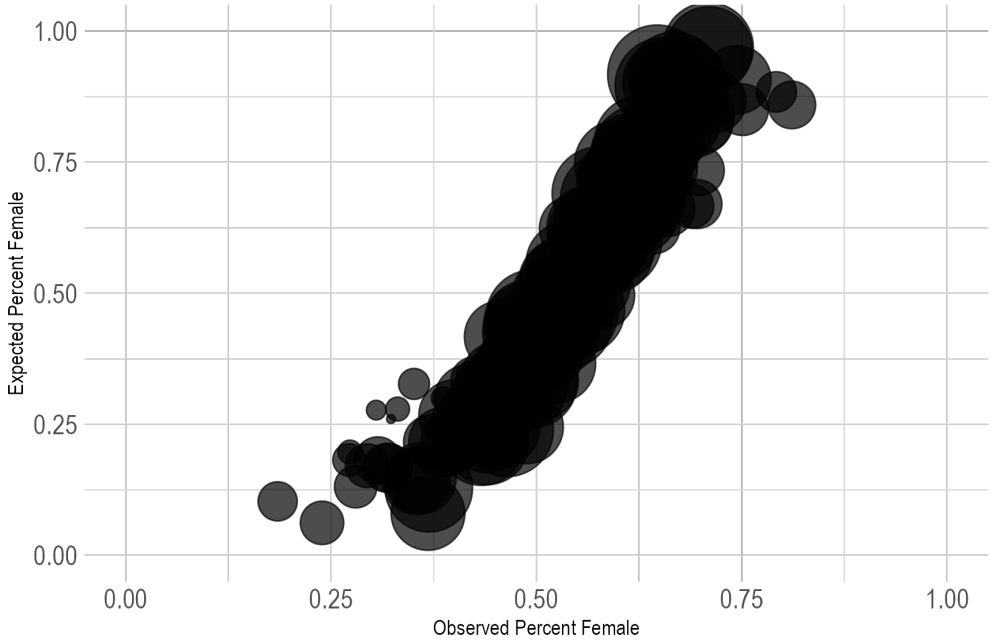

Figure 13: Observed vs expected percent female by major. Circles are the number of majors. In the HTML this was colored by Major Division and an interactive graph, but that does not work with a pdf output.

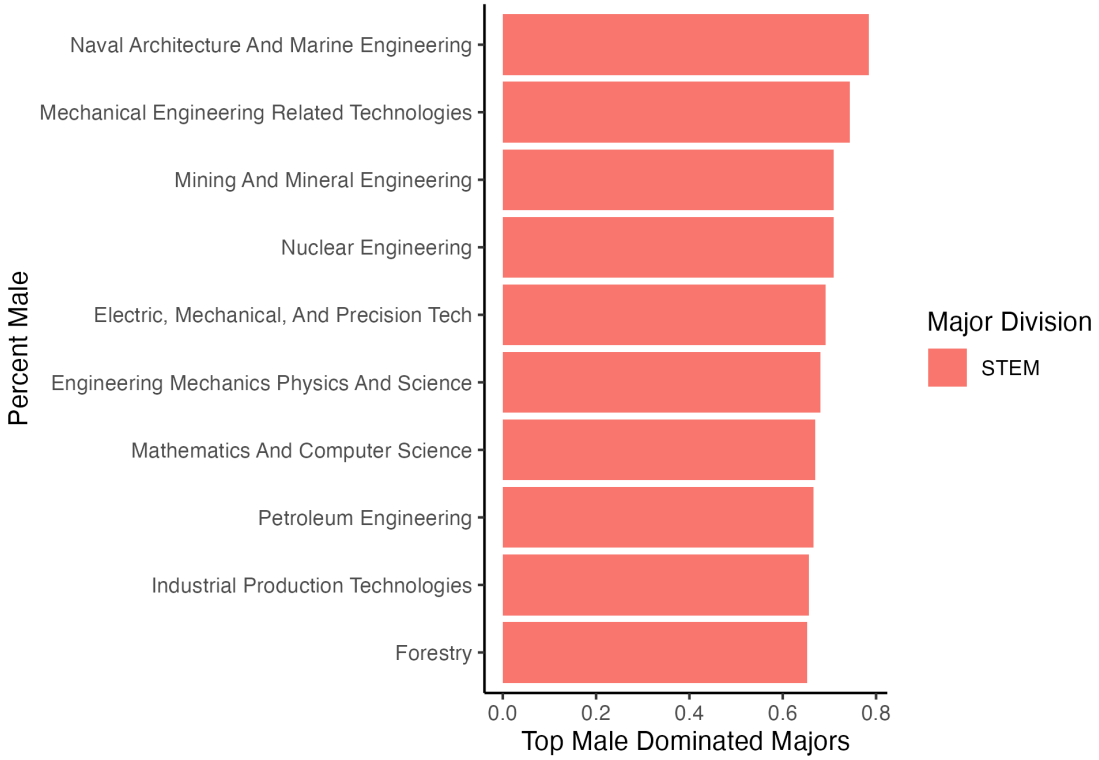

Figure 14: Top ten majors with the highest percentage of male students. Colored by division.

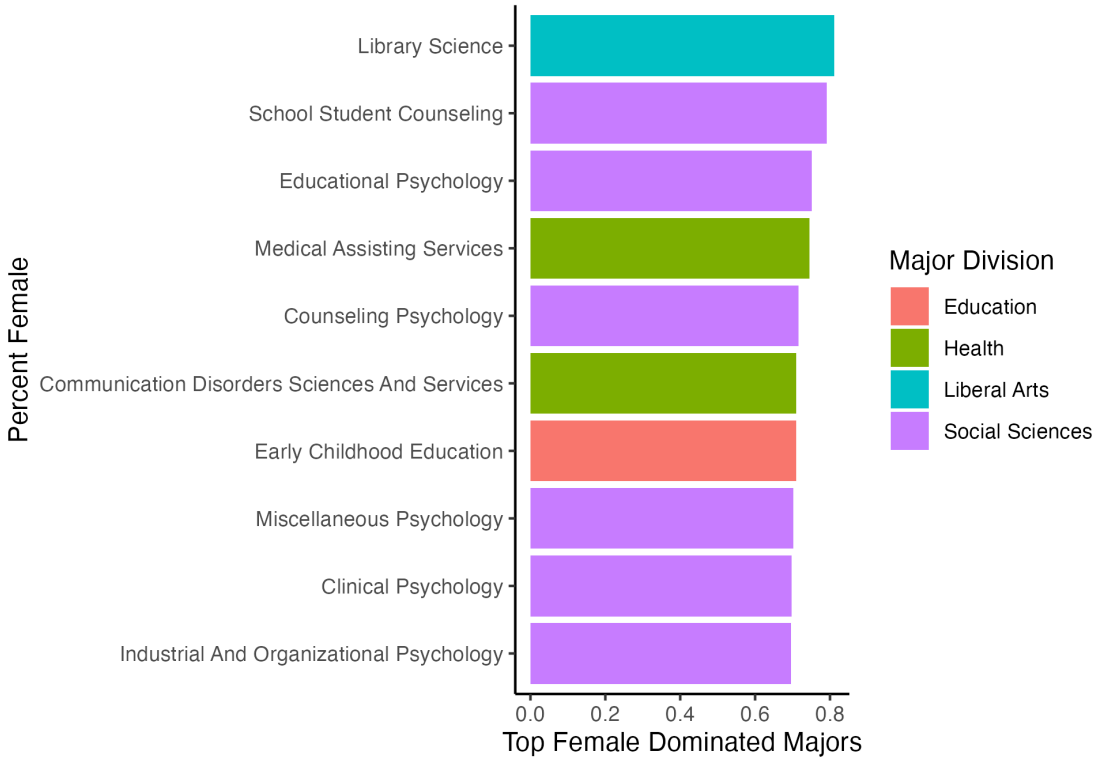

Figure 15: Top ten majors with the highest percentage of female students. Colored by division.

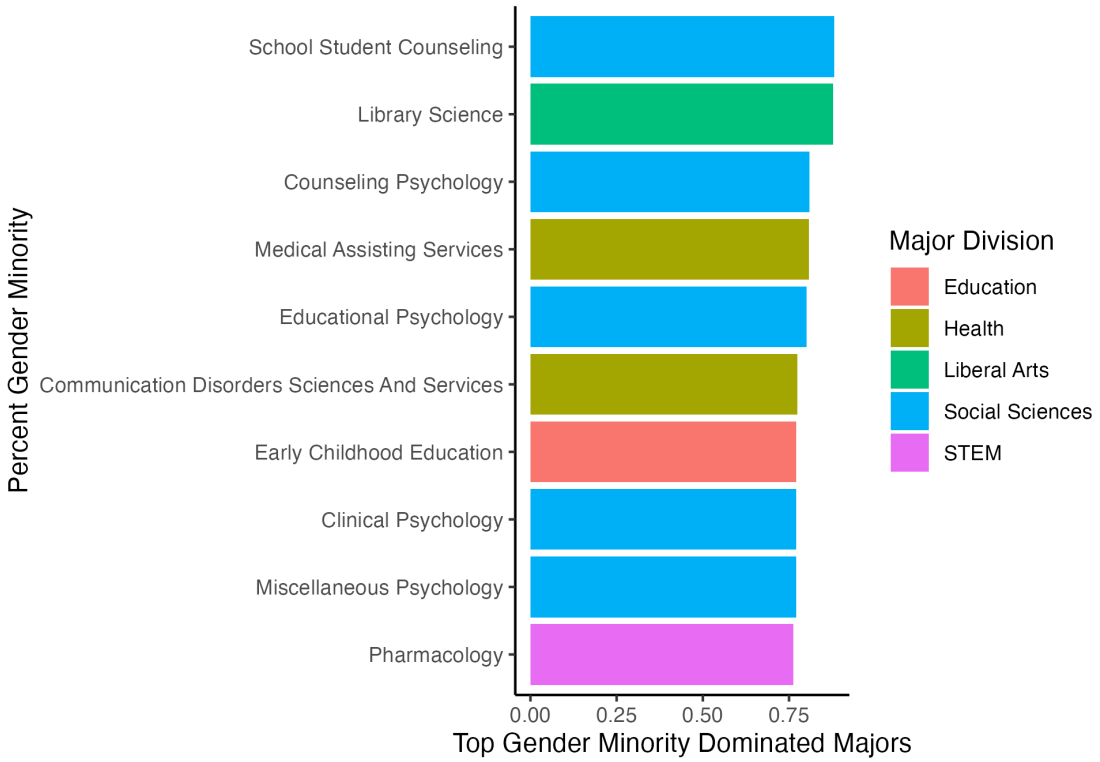

Figure 16: Top ten majors with the highest percentage of male students. Colored by division.

4 Discussion

In this section we discuss each of the results sections and consider challenges and obstacles. Then we return to the hypothesis and the predictions and finish with the future work section.

1. Demographics (4.1)
2. Major (4.2)
3. Learning Style (4.3)
4. Obstacles and Challenges (4.4)
5. Hypothesis and Predictions (4.5)
6. Future Work (4.6)

4.1 Demographics

The extremely high p-values from the Chi Squared Goodness of Fit Test show that **Ethnoracial Group**, **Gender**, **International Student Status**, and **Socioeconomic Status** all closely math the empirical sampling distribution taken from Nicolas & Sakib (2024). Visually, we see that **GPA** and **Student Semester** match the uniform distribution from which they were sampled.

The insignificant p-values from the Chi Squared Test of Independence show that **Ethnoracial Group**, **Gender**, **International Student Status**, and **Socioeconomic Status** are independent and not correlated with eachother. Notably there is high variation in the p-values with **Ethnoracial Group** and **Gender** having a very high p-value (0.8605), while **International Student Status**, and **Socioeconomic Status** have an almost significant p-value (0.08951).

4.2 Major

The extremely high p-value from the Chi Squared Goodness of Fit Test show that gender distribution among **Major(s)** closely matches the empirical data used in the weighting function. However, the extremely low p-value from the Chi Squared Goodness of Fit Test show that major popularity within **Major(s)** does not match the empirical data used in the weighting function. Thus, the weighting function generated by Nicolas & Sakib (2024) does not properly weight the majors to reflect the empirical data.

The insignificant p-values from the Chi Squared Test of Independence show that **Ethnoracial Group**, **International Student Status**, and **Socioeconomic Status** are independent from and not correlated with **Major(s)**. This was expected since none of these variables are explicitly mapped to **Major(s)**. The very low p-values from the Chi Squared Test of Independence show that **Gender** and **Previous Course Subjects** are correlated with **Major(s)**. This was expected since **Gender** is explicitly mapped to **Major(s)** which is explicitly mapped to **Previous Courses**.

Looking at the top ten majors dominated by different gender identities, we see that all the male dominated majors are STEM while the female or gender minority dominated majors come from a variety of divisions. Looking at Figure 12 we see that for majors with few female students more female students were observed than expected. At the same time, for majors with more female students less female students were observed than expected. Since some randomization was used in the majors function (Nicolas & Sakib, 2024), perhaps this had a normalizing effect which is why only majors at the extremities (very high or very low percentage of female students) were pushed closer to being even balanced between genders (around 50% female).

4.3 Learning Style

The extremely high p-value from the Chi Squared Goodness of Fit Test show that **Learning Style(s)** all closely math the empirical sampling distribution taken from Nicolas & Sakib (2024). The insignificant p-values from the Chi Squared Test of Independence show that **Ethnoracial Group**, **Gender**, **International Student Status**, **Socioeconomic Status**, and **Previous Course Types** are independent and not correlated with **Learning Style(s)**. This was unexpected since **Learning Style(s)** was explicitly mapped to **Previous Course Types**.

4.4 Obstacles and Challenges

Since the list features in the *Purely Synthetic Educational Dataset* were imported as strings, once they were split and recombined the dataset became very large (>1,000,000 rows). This limited the tables and graphs

that could be made with these features. This is why only the list features **Major(s)** and **Learning Style(s)** which have at most 2 entries were examined closely as they never made the dataset larger than 200,000 rows. Another difficulty was how many of the graphics were made to be interactive and while this worked when knitted to HTML, knitting to pdf prevented interactive graphics. Thus, the graphs shown here are slightly altered versions of the interactive HTML graphics. Finally, the extremely large size of the dataset (100,000 rows and 17 columns) meant that not all features could be examined within the time frame.

4.5 Hypothesis and Predictions

Now we return to our predictions and hypothesis (Table 1). All of the features matched their empirical sampling distributions except for major popularity. Thus, the first prediction is mostly supported overall. For the second prediction, only three explicit mappings were measured and two of three showed correlations between the two features. The converse of the second prediction was true with all of the non explicitly mapped features being independent and not correlated with each other.

Thus, the three predictions overall do support the hypothesis that the data generation function from Nicolas & Sakib (2024) works as intended. The nuance comes from the major weighting function, and also the independence of **Learning Style(s)** and **Previous Course Types** which were explicitly mapped.

4.6 Future Work

Repeating the same testing on the remaining variables in the dataset can further support the hypothesis and the notion that the dataset is usable for educational research. Additionally, future work should focus on attempting to fix the parts of the generating function that did not work as expected. Once, they are changed, a new dataset could be generated and then the analysis in this work could be repeated on the newly generated dataset. This work only showed that certain features were correlated, but a more advanced statistical analysis is needed to show the strength of the correlation.

Once this work is done, and the dataset has been shown to truly represent the empirical data, then the dataset can be used for educational research. Since the dataset is purely synthetic it is not connected to any real people, and thus poses no privacy risk. Previous educational research (Guan et al., 2020; Shao et al., 2021) can be redone with the purely synthetic educational dataset to further test the effectiveness of the dataset at replacing empirical data or synthetic data that is based in empirical data.

References

Asai, S. (2020). *Dict: R6 Based Key-Value Dictionary Implementation*.

College Majors Explorer. (n.d.). In *Big Economics*. https://bigeconomics.org/college-majors-explorer/.

Garnier, S., Ross, N., Rudis, B., Sciaini, M., Camargo, A. P., & Scherer, C. (2024). *Viridis: Colorblind-Friendly Color Maps for R*.

Gonsior, J., Thiele, M., & Lehner, W. (2021). *ImitAL: Learning Active Learning Strategies from Synthetic Data* (arXiv:2108.07670). arXiv. `https://doi.org/10.48550/arXiv.2108.07670`

Guan, C., Mou, J., & Jiang, Z. (2020). Artificial intelligence innovation in education: A twenty-year data-driven historical analysis. *International Journal of Innovation Studies*, *4*(4), 134–147. `https://doi.org/10.1016/j.ijis.2020.09.001`

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., & Weller, A. (2022). *Synthetic Data – what, why and how?* (arXiv:2205.03257). arXiv. `https://doi.org/10.48550/arXiv.2205.03257`

Liu, C., Wan, Z., Wang, H., Chen, Y., Qaiser, T., Jin, C., Yousefi, F., Burlutskiy, N., & Arcucci, R. (2024). *Can Medical Vision-Language Pre-training Succeed with Purely Synthetic Data?* (arXiv:2410.13523). arXiv. `https://doi.org/10.48550/arXiv.2410.13523`

Mandinach, E. B., & Schildkamp, K. (2021). Misconceptions about data-based decision making in education: An exploration of the literature. *Studies in Educational Evaluation*, *69*, 100842. `https://doi.org/10.1016/j.stueduc.2020.100842`

Neuwirth, E. (2022). *RColorBrewer: ColorBrewer Palettes*.

Nicolas, A., & Sakib, S. K. (2024). *PipelineEDU: Creating High-Quality Synthetic Data for Educational Research and Applications*.

Ooms, J., Lang, D. T., & Hilaiel, L. (2024). *Jsonlite: A Simple and Robust JSON Parser and Generator for R*.

Rudis, B. (2024). *Hrbrthemes: Additional Themes, Theme Components and Utilities for 'Ggplot2'*.

Shao, E., Guo, S., & Pardos, Z. A. (2021). Degree Planning with PLAN-BERT: Multi-Semester Recommendation Using Future Courses of Interest. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(17), 14920–14929. https://doi.org/10.1609/aaai.v35i17.17751

Wang, X., Xie, L., Dong, C., & Shan, Y. (2021). Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data. *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 1905–1914. https://doi.org/10.1109/ICCVW54120.2021.00217

Wickham, H. (2023). *Tidyverse: Easily Install and Load the 'Tidyverse'*.