Exploring Differences Between East Coast Colleges in the United States

Andras Horvath & Ting-Jung Lee

STAT 5375 Project Report

1 Background

When choosing a university, individuals often consider a variety of factors that reflect their unique preferences, values, and goals. The decision-making process varies significantly from person to person. The dataset chosen for this project is from the StatCrunch website and is titled "Sample College Data." It can be seen directly through the following link:

https://www.statcrunch.com/app/index.html?dataid=3133754. The dataset concentrated on colleges and universities located on the East Coast of the United States (Northeastern and Southern United States), encompassing both categorical and quantitative variables. A detailed understanding of the variables is outlined in Table 1.

Variable	Description
University	The colleges and universities included in our dataset are located in Delaware, the District of Columbia, Maryland, Pennsylvania, Virginia, or West Virginia in 2011.
State	The dataset covers the states of Delaware, the District of Columbia, Maryland, Pennsylvania, Virginia, and West Virginia in the United States.
Location	The positioning of universities or colleges encompasses urban, suburban, town, or rural settings.
Public/Private	The classification of universities or colleges falls into the categories of public or private institutions.
Admissions Rate	The admissions rate represents the percentage of applicants admitted to a particular college or university in 2011.
SAT Reading (75%)	The 75th percentile SAT Reading Scores represent the score that is higher than 75% of individuals who took the SAT Reading section.
SAT Math (75%)	SAT Math (75%) denotes the score that exceeds the scores of 75% of test-takers in the SAT Math section.
Tuition & Fees	Tuition and fees refer to the costs associated with enrollment and education at a college or university. Tuition typically covers the academic instruction provided by the university or college, while fees include additional charges for services or facilities.

Average Financial Aid	Average financial aid is the mean amount that the university assists a student in covering college expenses.
Enrollment	The enrollment figure for a college or university represents the number of students who were registered in 2011.
Undergraduate Enrollment	The undergraduate enrollment figure for a college or university denotes the number of students at the undergraduate level who were registered in 2011.
Retention Rate	The retention rate signifies the percentage of first-time or first-year undergraduate students at a university or college who continue at the same institution in the subsequent year.
Student-Teacher Ratio	The student-teacher ratio is a measure that indicates the average number of students for each teacher in a college or university. It is calculated by dividing the total number of students enrolled by the total number of teachers or instructors.
Graduation Rate (5-year)	The graduation rate (5-year) represents the percentage of students who complete their degree within a five-year period at a college or university.

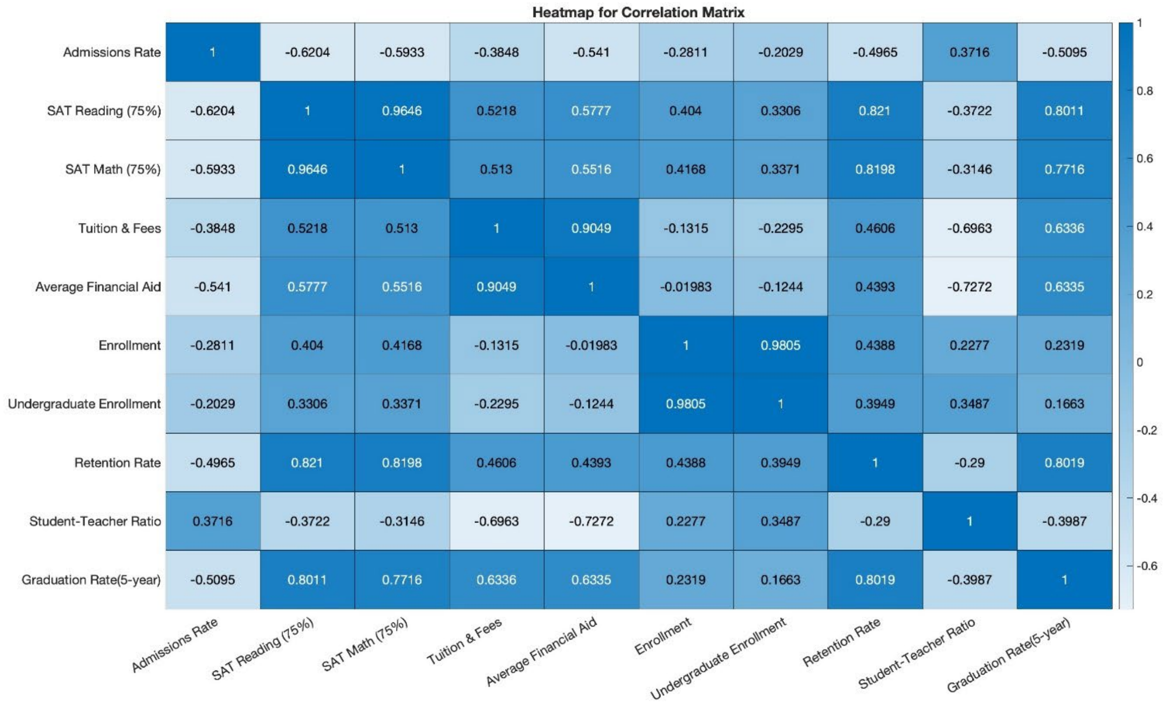
Table 1: Description of Variables in the Dataset

2 Methodology

The goal of this study was to perform a principal component analysis on a dataset of East Coast United States Colleges from 2011 which contains 197 observations across 13 variables in order to visualize and determine if there were substantial differences between the institutions.

To subset our data into exploratory and confirmatory datasets a stratified sample was taken from the original dataset. As there are 197 observations we wanted to allocate no more than 1/3 of the observations to the exploratory dataset, meaning that 65 observations would be allocated to the exploratory dataset. The strata were determined by using the location categorical variable. This was chosen for a number of reasons. First, if the stratum were determined using the state categorical variable, then the exploratory dataset would have at most 1 observation from Delaware, as the original data contained only 3 observations from Delaware. This would not give much insight in the plots, and Delaware was not the only state with very few observations. If we instead used the Private vs Public institution variable then it is plausible that some of the states and location types will be underrepresented in the exploratory dataset. Therefore, we decided to use the location to determine the strata that would lead to the fewest potential problems. Of course, other choices for the subsetting is possible. The location variable has four categories: city, suburb, town, and rural setting. In the original dataset 71 schools were in cities, 63 in suburbs, 46 in towns, and 17 in rural areas. As a result, 23 observations from cities, 21 observations from suburbs, 15 observations from towns, and 6 observations from rural areas were randomly sampled to create the exploratory dataset. The remaining 132 observations would compose the confirmatory dataset.

A heatmap of the correlation matrix of the numeric variables in the exploratory dataset is shown in Figure 1.



	Admissions Rate	SAT Reading (75%)	SAT Math (75%)	Tuition & Fees	Average Financial Aid	Enrollment	Undergraduate Enrollment	Retention Rate	Student-Teacher Ratio	Graduation Rate(5-year)
Admissions Rate	1	-0.6204	-0.5933	-0.3848	-0.541	-0.2811	-0.2029	-0.4965	0.3716	-0.5095
SAT Reading (75%)	-0.6204	1	0.9646	0.5218	0.5777	0.404	0.3306	0.821	-0.3722	0.8011
SAT Math (75%)	-0.5933	0.9646	1	0.513	0.5516	0.4168	0.3371	0.8198	-0.3146	0.7716
Tuition & Fees	-0.3848	0.5218	0.513	1	0.9049	-0.1315	-0.2295	0.4606	-0.6963	0.6336
Average Financial Aid	-0.541	0.5777	0.5516	0.9049	1	-0.01983	-0.1244	0.4393	-0.7272	0.6335
Enrollment	-0.2811	0.404	0.4168	-0.1315	-0.01983	1	0.9805	0.4388	0.2277	0.2319
Undergraduate Enrollment	-0.2029	0.3306	0.3371	-0.2295	-0.1244	0.9805	1	0.3949	0.3487	0.1663
Retention Rate	-0.4965	0.821	0.8198	0.4606	0.4393	0.4388	0.3949	1	-0.29	0.8019
Student-Teacher Ratio	0.3716	-0.3722	-0.3146	-0.6963	-0.7272	0.2277	0.3487	-0.29	1	-0.3987
Graduation Rate(5-year)	-0.5095	0.8011	0.7716	0.6336	0.6335	0.2319	0.1663	0.8019	-0.3987	1

Figure 1: Heatmap of Correlations Between Quantitative Variables

In the figure, it is evident that several variables display high correlations. A closer look at the strongest correlations can be found in Table 2. This observation has led us to utilize the principal component analysis (PCA) as a methodological approach. In high-dimensional data, PCA works to simplify complexity while retaining trends, patterns, and enhancing our understanding of the data's underlying structure. By employing PCA, we aim to explore deeper into the convoluted relationships among these variables and effectively reduce dimensionality. This reduction not only provides a clearer representation through data visualization but also ensures that the new variables (y_i) are uncorrelated, thereby enhancing the interpretability and independence of the transformed dataset.

$\boldsymbol{|r_{ij}| > 0.7}$	**Variables**
$r_{6,7} = 0.9805$	Enrollment vs. Undergraduate Enrollment
$r_{2,3} = 0.9646$	SAT Reading (75%) vs. SAT Math (75%)
$r_{4,5} = 0.9049$	Tuition & Fees vs. Average Financial Aid
$r_{2,8} = 0.8210$	SAT Reading (75%) vs. Retention Rate
$r_{3,8} = 0.8198$	SAT Math (75%) vs. Retention Rate
$r_{8,10} = 0.8019$	Retention Rate vs. Graduation Rate (5-year)

$r_{2,10} = 0.8011$	SAT Reading (75%) vs. Graduation Rate (5-year)
$r_{3,10} = 0.7716$	SAT Math (75%) vs. Graduation Rate (5-year)
$r_{5,9} = -0.7272$	Average Financial Aid vs. Student-Teacher Ratio

Table 2: Strongest Correlations Between Quantitative Variables (|r| > 0.7)

There exist two distinct approaches to conducting principal component analysis. The first method involves utilizing a covariance matrix in the computations, while the latter approach begins by standardizing the data and subsequently employs a correlation matrix. The choice between these methods depends on the character of the variables involved. The first approach is preferred when the variables exhibit similar scales, whereas the second approach is often more suitable when the variables vary significantly in scale. In our specific case, the quantitative variables, such as the admission rate, SAT scores, and tuition fees present considerable differences in scales. Consequently, we have opted to standardize our data and employ the correlation matrix for the principal component analysis. This decision leads us to the second approach.

A scree plot is a valuable visual tool for determining the optimal number of principal components. On the y-axis, the eigenvalues, represented as $\widehat{\lambda_\iota}$, are arranged in descending order while the x-axis displays their respective numbers. The goal is to identify where a clear elbow or bend in the scree plot occurs, and to use this to determine the necessary number of components. As long as all the points after the bend, for some i, are small and have consistent magnitude then we will take the number of necessary principal components to be equal to the value for i.

In the following scree plot, the most pronounced elbow occurs at $i = 3$. Beyond $\hat{\lambda}_3$, the subsequent eigenvalues consistently exhibit diminished magnitudes. This observation leads to the inference that three principal components effectively capture and summarize the total sample variance.

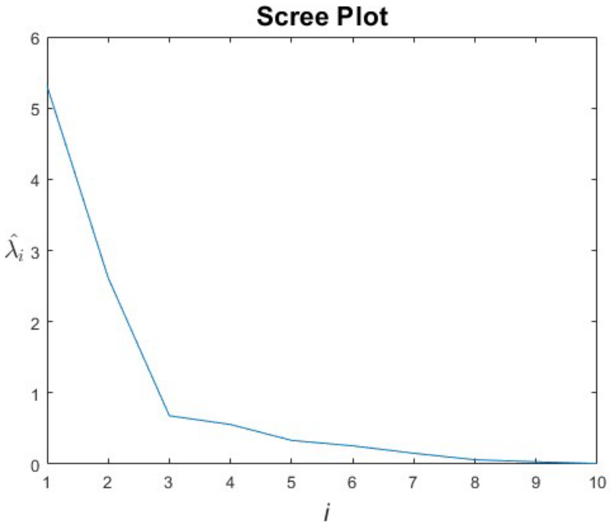


Figure 2: Scree Plot

Upon calculating the cumulative proportions of variation explained the first three principal components account for 86.03% of the total sample variance. Utilizing principal component analysis, this significant reduction in dimensionality to three components ensures a comprehensive dataset representation while alleviating the correlation among variables. The outcome signifies an effective consolidation of information, enhancing interpretability and analytical clarity in our analysis.

The first 3 principal components are as follows:

$$\hat{y}_1 = -0.3063 * z_1 + 0.4008 * z_2 + 0.3927 * z_3 + 0.3173 * z_4 + 0.3392 * z_5 + 0.1553 * z_6 + 0.1135 * z_7 + 0.3681 * z_8 - 0.2408 * z_9 + 0.3810 * z_{10}$$

$$\hat{y}_2 = -0.0268 * z_1 + 0.1045 * z_2 + 0.1211 * z_3 - 0.3344 * z_4 - 0.2876 * z_5 + 0.5318 * z_6 + 0.5645 * z_7 + 0.1563 * z_8 + 0.3914 * z_9 - 0.0104 * z_{10}$$

$$\hat{y}_3 = 0.4964 * z_1 + 0.2085 * z_2 + 0.2463 * z_3 - 0.0733 * z_4 - 0.3141 * z_5 - 0.3393 * z_6 - 0.2708 * z_7 + 0.3525 * z_8 + 0.3289 * z_9 + 0.3529 * z_{10}$$

Where $z_1, z_2, \dots, z_{10}$, are the standardized version of the variables found in Table 3.

z_i	**Standardized Variable**
z_1	Admissions Rate
z_2	SAT Reading (75%)
z_3	SAT Math (75%)
z_4	Tuition & Fees
z_5	Average Financial Aid
z_6	Enrollment
z_7	Undergraduate Enrollment
z_8	Retention Rate
z_9	Student-Teacher Ratio
z_{10}	Graduation Rate (5-year)

Table 3: Standardized Variables z_i

We can make a few observations from the given equations. First, z_6 and z_7 contribute roughly ½ as much to the first principal component as the other standardized variables. Second, z_1 and z_{10} contribute very little to the second principal component. Also notice that the coefficient for z_2 and

z_3 as well as z_8 is a little larger than 0.1. We can use this to determine how much more the other linear combinations of standardized variables contribute to the second principal component. Compared to the aforementioned linear combination of standardized variables, z_4, z_5, z_9, and z_{10} contribute roughly 3 times as much, and z_6 and z_7 contribute roughly five times as much to the second principal component. Finally, z_4 contributes very little to the third principal component, while z_1 contributes a little less than twice as much as all the other standardized variables.

Finally, are there any trends in the data which would lead to tests of interest? Based on the Scree Plot in Figure 2, the first three principal components were used to create a 3-dimensional scatterplot in order to look for trends. The subsequent 2-dimensional scatterplots are shown below for their respective categorical variables.

First, we looked for trends in the data based on the state which the college was located in. Figure 3 shows the scatterplot of the 1st vs 3rd principal components, and Figure 4 shows the scatterplot of the 2nd vs 3rd principal components. Note that the scatterplot of the 1st vs 2nd principal components was omitted as it did not provide any additional information which the above scatterplots did not provide. In each of these figures the state was coded as a quantitative variable where 1 corresponds to Delaware, 2 corresponds to the District of Columbia, 3 corresponds to Maryland, 4 corresponds to Pennsylvania, 5 corresponds to Virginia, and 6 corresponds to West Virginia. Note that because Delaware has only 3 observations in the original dataset, the subsetting procedure did not actually catch any observations from Delaware.

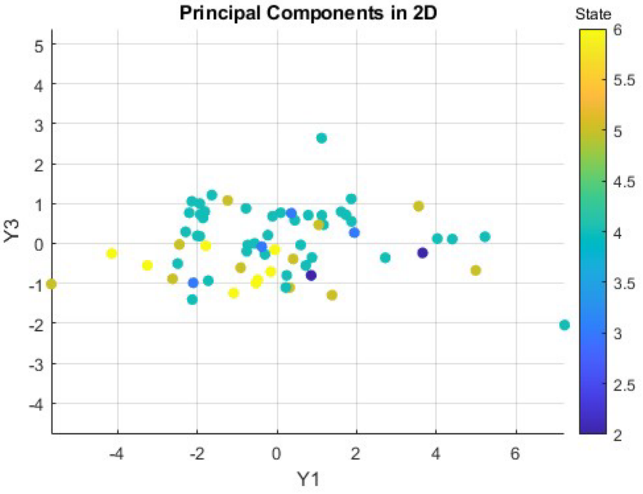


Figure 3: Scatterplot of 1st vs 3rd Principal Components for State

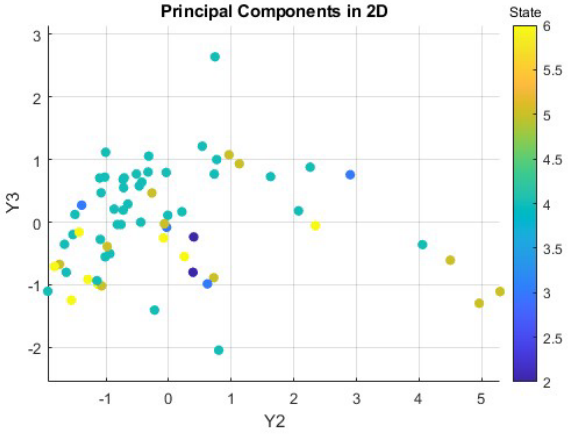


Figure 4: Scatterplot of 2nd vs 3rd Principal Components for State

Based on the figures, there does not appear to be any major differences between the means of the different states as there is no distinct clustering. Each group is just a part of a cloud of points with no major distinctions or separation from one another. However, there does appear to be a difference in the variances of certain states as some clouds of points are far more spread out than others. Pennsylvania and Virginia both appear to have a much larger variance than West Virginia, the District of Columbia, and Maryland. We thus have strong evidence that the state in which a college resides does make an impact on the covariance matrix structure among these different states.

Next, we looked for trends in the data based on the location type (city/suburb/town/rural) that the college was located in. Figure 5 shows the scatterplot of the 1st vs 2nd principal components, and Figure 6 shows the scatterplot of the 1st vs 3rd principal components. Note that the scatterplot of the 2nd vs 3rd principal components was omitted as it did not provide any additional information. In each of these figures the location type was coded as a quantitative variable of decreasing population size i.e., 1 corresponds to City, 2 corresponds to Suburb, 3 corresponds to Town, and 4 corresponds to Rural.

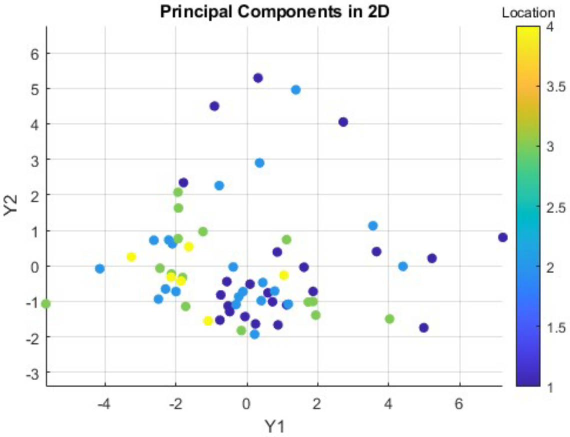


Figure 5: Scatterplot of 1st vs 2nd Principal Components for Location Type

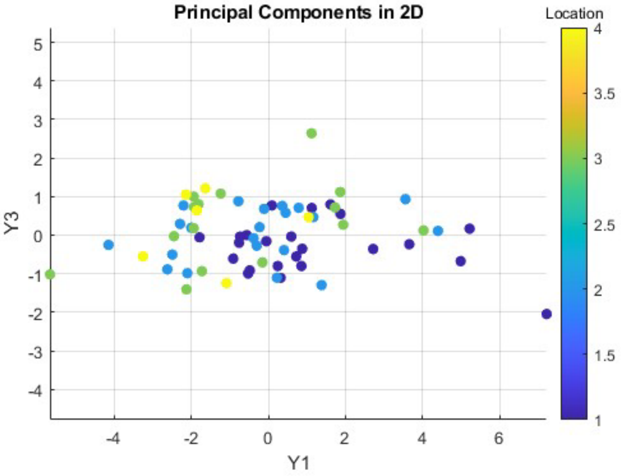


Figure 6: Scatterplot of 1st vs 3rd Principal Components for Location Type

Based on the figures, there does not appear to be any major differences between the means of the different location types. There is no distinct clustering taking place. However, there does appear to be a difference in the variances of the different location types. Both cities and suburbs appear to have a much larger variance than towns or rural areas. We thus have evidence that the location type in which a college resides does make an impact on the covariance matrix structure among these location types.

Finally, we looked for trends in the data based on whether the college was a Private or Public institution. Figure 7 shows the scatterplot of the 1st vs 2nd principal components, and Figure 8 shows the scatterplot of the 2nd vs 3rd principal components. Note that the scatterplot of the 1st vs 3rd principal components was omitted as it did not provide any additional information. In each of these figures Private vs Public was coded as a quantitative variable where 1 corresponds to a Private College, and 2 corresponds to a Public College.

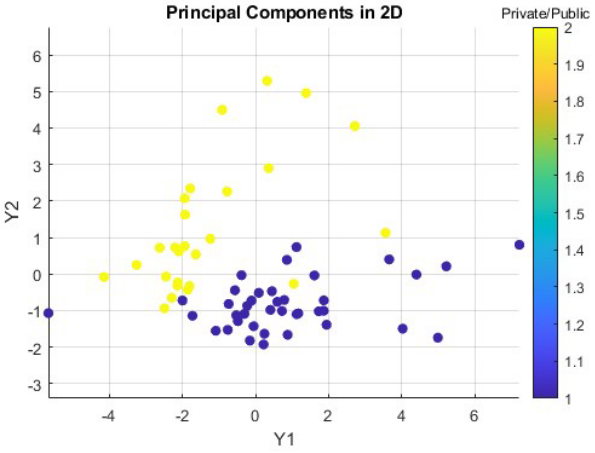


Figure 7: Scatterplot of 1^{st} vs 2^{nd} Principal Components for Private/Public College

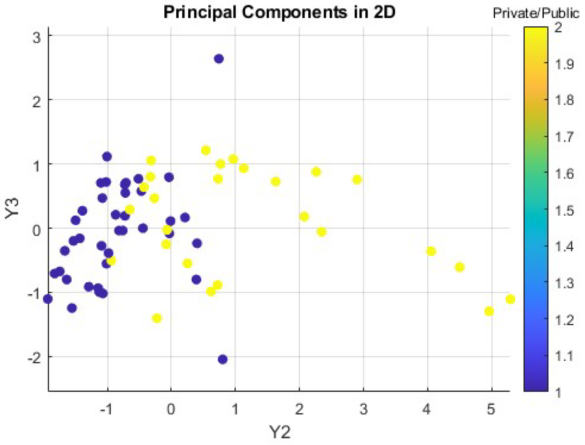


Figure 8: Scatterplot of 2^{nd} vs 3^{rd} Principal Components for Private/Public College

Based on the above two figures there does appear to be a sort of clustering taking place between the two groups. There is very little overlap between the two groups. This suggests that there probably is a difference between the means of the two categories of colleges. There also appears to be a difference in the variance of the two groups. For example, in Figures 8 while it appears that the cloud of points has the same variation along the axis Y3, this is not the case along the Y2 axis. It is apparent that the variation is higher for the Public Colleges along the Y2 axis. Therefore, we have evidence that the type of institution, whether Public or Private, does have an impact on the means and covariance structures.

3 Confirmatory Analysis

For this project, we intended to explore potential differences in the mean and covariance structure among the categorical variables, including locations, states, and the distinction between public and private universities or colleges. The final results of the exploratory analysis give rise to hypotheses to be tested using the confirmatory dataset. These tests are highlighted below. Assume all of the hypothesis tests are at significant level $\alpha = 0.05$.

We initiated the analysis by assessing the multivariate normality of the dataset. We opted for the application of Mardia's Test as our chosen method. The details of the hypothesis testing are outlined below.

$$\begin{cases} H_0: X \sim MVN \\ H_1: X \nsim MVN \end{cases}, \alpha = 0.05$$

First, we computed Mardia's test statistic for multivariate skewness (corrected for small samples) as 1.4133×10^3, and Mardia's test statistic for multivariate kurtosis as 22.4482. Subsequently, the associated p-values were determined to be approximately 0. Therefore, based on these findings, we rejected the null hypothesis (H_0) at a significance level of $\alpha = 0.05$. We can thus infer that the dataset does not follow a multivariate normal distribution ($X \nsim MVN$).

Following the exploratory analysis, we then determined that the type of institution, location, and the state in which a college resides in does have a discernible impact on the covariance matrix structure. Consequently, the hypotheses to be tested in the confirmatory analysis are as follows ($\alpha = 0.05$):

$$(1) \begin{cases} H_0: \Sigma_{Private} = \Sigma_{Public} \\ H_1: \Sigma_{Private} \neq \Sigma_{Public} \end{cases}$$

$$(2) \begin{cases} H_0: \Sigma_{City} = \Sigma_{Suburb} = \Sigma_{Town} = \Sigma_{Rural} \\ H_1: At\ least\ one\ of\ the\ covariance\ matrices\ differ \end{cases}$$

$$(3) \begin{cases} H_0: \Sigma_{Delaware} = \Sigma_{DC} = \Sigma_{Maryland} = \Sigma_{Pennsylvania} = \Sigma_{Virginia} = \Sigma_{West\ Virgina} \\ H_1: At\ least\ one\ of\ the\ covariance\ matrices\ differ \end{cases}$$

The selected method to evaluate the equality of covariance matrices is Box's Test with a χ^2 approximation. However, this test assumes that the data conform to multivariate normality, and our previous analysis indicated that our dataset departs from this assumption. Accordingly, we opt for an alternative approach, specifically comparing a Box's Test permutation critical value instead (assuming random permutations of the order of observations $B = 1000$ times). Under a significance level of $\alpha = 0.05$, we will reject the null hypothesis (H_0) if $C >$ the Box's Test permutation critical value. The test statistic will be provided below.

Let $u = \left[\sum_l \frac{1}{n_l - 1} - \frac{1}{\sum_l (n_l - 1)}\right]\left[\frac{2p^2+3p-1}{6(p+1)(g-1)}\right]$, where p is the number of variables and g is the number of groups. Then, the test Statistic C would be defined as the following,

$$\boldsymbol{C = (1-u)M = (1-u)\left\{\left[\sum_l (n_l - 1)\right]\ln|S_{pooled}| - \sum_l [(n_l - 1)\ln|S_l|]\right\}}$$

, where $S_{pooled} = \frac{1}{\sum_l (n_l - 1)}\{(n_1 - 1)S_1 + (n_2 - 1)S_2 + \cdots + (n_g - 1)S_g\}$,

S_l is the lth group sample variance matrix.

We commenced by examining whether there exists a difference in the covariance structure among various institution types, considering sample sizes $n_1 = 82$, $n_2 = 50$, the number of variables $p = 10$, and the number of groups $g = 2$. The hypothesis testing for the comparison of institution types between private and public is outlined as follows.

$$\begin{cases} H_0\colon \Sigma_{Private} = \Sigma_{Public} \\ H_1\colon \Sigma_{Private} \neq \Sigma_{Public} \end{cases}, \alpha = 0.05$$

We computed the test statistic to be $C = 347.8662$, and the Box's Test permutation critical value as 147.5236. Since $C = 347.8662$ exceeds 147.5236, the Box's Test permutation critical value, we reject the null hypothesis (H_0) at a significant level $\alpha = 0.05$. Therefore, we deduce that there is evidence indicating a difference in the covariance structures among different institution types, i.e., $\Sigma_{Private} \neq \Sigma_{Public}$.

Next, we proceeded to examine whether there existed a distinction in the covariance structures within the location types, specifically encompassing city, suburb, town, and rural area.

$$\begin{cases} H_0\colon \Sigma_{City} = \Sigma_{Suburb} = \Sigma_{Town} = \Sigma_{Rural} \\ H_1\colon At\ least\ one\ of\ the\ covariance\ matrices\ differ \end{cases}, \alpha = 0.05$$

We begin by assessing whether a significant difference exists between city and town, as they displayed larger variances compared to suburb and rural area. Consequently, the null hypothesis is listed as follows: H_0: $\Sigma_{City} = \Sigma_{Town}$ with sample sizes $n_1 = 48$, $n_2 = 31$, the number of variables $p = 10$, and groups $g = 2$. The test statistic, calculated as $C = 201.2750$, surpasses the Box's Test permutation critical value of 122.8188. Consequently, we reject the null hypothesis (H_0) at a significance level of $\alpha = 0.05$. Table 4 displays all pairs of null hypothesis combinations in location types. As we reject at least one of the comparisons, we have evidence that there is a distinction in the covariance structures among the different location types.

Null Hypothesis	Test Statistic (C)	Box's Test Permutation Value	Result
$\Sigma_{City} = \Sigma_{Suburb}$	162.9632	122.8510	Reject H_0
$\Sigma_{City} = \Sigma_{Town}$	201.2750	122.8188	Reject H_0
$\Sigma_{City} = \Sigma_{Rural}$	110.6009	130.2541	Fail to reject H_0
$\Sigma_{Suburb} = \Sigma_{Town}$	124.4155	126.0680	Fail to reject H_0
$\Sigma_{Suburb} = \Sigma_{Rural}$	93.1654	137.2264	Fail to reject H_0
$\Sigma_{Town} = \Sigma_{Rural}$	90.5961	121.5793	Fail to reject H_0

Table 4: All Combination Pairs of Null Hypotheses in Location Types

Finally, we discovered compelling evidence from the exploratory analysis indicating that the states of Delaware, the District of Columbia (DC), Maryland, Pennsylvania, Virginia, and West Virginia had a discernible difference in the covariance structures. Hence, the hypothesis to be investigated in the confirmatory analysis is stated below:

$$\begin{cases} H_0\colon \Sigma_{Delaware} = \Sigma_{DC} = \Sigma_{Maryland} = \Sigma_{Pennsylvania} = \Sigma_{Virginia} = \Sigma_{West\ Virgina} \\ H_1\colon At\ least\ one\ of\ the\ covariance\ matrices\ differ \end{cases}, \alpha = 0.05$$

Following that, we began with testing the two largest states in our dataset, Pennsylvania and Virginia. The hypothesis is stated as $H_0\colon \Sigma_{Pennsylvania} = \Sigma_{Virginia}$, with sample sizes $n_1 = 67$, $n_2 = 29$, the number of variables $p = 10$, and groups $g = 2$. The calculated test statistic $C = 118.3530$ is less than the Box's Test permutation critical value of 135.6276. Consequently, we fail to reject the null hypothesis. Subsequently, we examine the covariance structure between the largest state (Pennsylvania) and the third-largest state (Maryland), with sample sizes $n_1 = 67$, $n_2 = 18$, the number of variables $p = 10$, and groups $g = 2$. We computed the test statistic to be $C = 145.4437$, along with the corresponding Box's Test permutation critical value of 132.4595. Since $C = 145.4437$ is larger than 132.4595, the Box's Test permutation critical value, we reject the null hypothesis (H_0) at a significance level of $\alpha = 0.05$. Table 5 illustrates the combination pairs of null hypotheses among the different states provided that n > p. Due to the sample sizes of Delaware, DC, and West Virginia being $3, 5$, and 10, respectively, there will be rank deficiency problems since there are 10 quantitative variables. As such, other testing procedures would need to be used to handle these cases, among which could include decreasing the number of quantitative variables considered if it was determined that they could be removed. Based on the following table, we reject at least one of the comparisons, and therefore deduce that there is a noticeable difference in the covariance structures among the different states.

Null Hypothesis	**Test Statistic (C)**	**Box's Test Permutation Value**	**Result**
$\Sigma_{Maryland} = \Sigma_{Pennsylvania}$	145.4437	132.4595	Reject H_0
$\Sigma_{Maryland} = \Sigma_{Virginia}$	107.2746	117.6178	Fail to reject H_0
$\Sigma_{Pennsylvania} = \Sigma_{Virginia}$	118.3530	135.6276	Fail to reject H_0

Table 5: Pairs of Null Hypotheses Combinations in States

After thoroughly testing all covariance structures, the final step involves investigating whether there is a significant difference between the means of the different institution types. The hypothesis test is listed as follows:

$$\begin{cases} H_0: \mu_{Private} = \mu_{Public} \\ H_1: \mu_{Private} \neq \mu_{Public} \end{cases}, \alpha = 0.05$$

Given the outcomes of the previous tests, we know that our dataset does not follow a multivariate normal distribution, and the covariance structures differ among different institution types. Therefore, we must determine if we are in a large or small sample size setting. Here, the numbers $n_1 - p$ is 72 and $n_2 - p$ is 40, which we consider to be large given that $p = 10$. Therefore, the approach that we opted for comparing the mean vectors was to use the T_2^2- statistic, which is defined below,

$$T_2^2 = (\bar{x}_1 - \bar{x}_2 - \delta_0)' \left(\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2\right)^{-1} (\bar{x}_1 - \bar{x}_2 - \delta_0).$$

We will reject H_0 if $T_2^2 > \chi_p^2(\alpha)$ at significant level $\alpha = 0.05$ ($\delta_0 = 0$).

The computed test statistic $T_2^2 = 662.2719$ exceeds $18.3070 = \chi_{10}^2(0.05)$. Therefore, we reject H_0 at significant level $\alpha = 0.05$, and conclude that the means of different institution types differ (i.e., $\boldsymbol{\mu_{Private}} \neq \boldsymbol{\mu_{Public}}$).

For the final segment of the confirmatory analysis we will calculate the confidence intervals for the institution types in order to determine which marginal means led to the rejection of the null hypothesis. The chosen approach for computing the confidence intervals is the Bonferroni simultaneous confidence intervals for δ_i.

The $100(1 - \alpha)\%$ Bonferroni simultaneous confidence intervals for δ_i, where $i = 1, \ldots, p$ and p is the number of variables, is defined below:

$$(\bar{x}_{1i} \text{ - } \bar{x}_{2i}) \pm z\left(\frac{\alpha}{2p}\right)\sqrt{\frac{S_{1,ii}}{n_1} + \frac{S_{2,ii}}{n_2}}$$

In our specific case, with 10 quantitative variables and a significance level of $\alpha = 0.05$, the results are presented in Table 6.

Variable	Lower Confidence Limit	Upper Confidence Limit	Conclusion
Admissions Rate	-9.1740	-6.1470	Reject H_0
SAT Reading (75%)	33.6915	44.9934	Reject H_0
SAT Math (75%)	19.6987	31.3325	Reject H_0
Tuition & Fees	1.8956×10^4	1.9954×10^4	Reject H_0
Average Financial Aid	1.0369×10^4	1.1092×10^4	Reject H_0
Enrollment	-6.6366×10^3	-5.0503×10^3	Reject H_0
Undergraduate Enrollment	-6.0626×10^3	-4.8733×10^3	Reject H_0
Retention Rate	-0.6996	1.0781	Fail to reject H_0
Student-Teacher Ratio	-5.7125	-5.2241	Reject H_0
Graduation Rate (5-year)	5.0906	8.4665	Reject H_0

Table 6: Bonferroni Simultaneous Confidence Intervals for Public vs Private Mean Vector

4 Conclusions and Future Research

We found that the assumptions made during the exploratory analysis did in fact align with the results derived from the confirmatory analysis. We first concluded that the covariance matrix structures differed across states, locations, and institution types. Subsequently, in the case of institution types, the mean vectors were found to differ as well. This confirms that there are indeed multiple distinct factors that lead to a person’s decision to attend the college of their choice. We now propose a few different topics that could be explored in future research. Having established that there are differences in the covariance structures and mean vectors within the dataset, one topic worth exploring would be the application of Principal Component Analysis (PCA) to rank the colleges in this dataset. What is determined as “best” will be a subjective choice of the researcher based on a multitude of factors. Moreover, this dataset only comprises universities located along the eastern coast of the United States. Therefore, it would be interesting to expand the analysis conducted in this project to colleges in other regions of the United States which may provide insights into regional variations and enhance a more comprehensive understanding of the college institution system in the United States.