**Graduating Midwest Liberal Arts Colleges:**

**The influence of student characteristics on graduation rates**

Sneha Verma

Department of Economics, Accounting, and Management

ECON 493: Senior Honors Project

Professor Bird

17 May 2022

**Data Appendix**

**Purpose of Research:**

The goal of this research is to quantify the relationships between race, gender, and Pell grant status with graduation rate of students at private, liberal arts colleges in the Midwest.

**Variables:**

1. Race Dataset
2. college

Chart, bar chart

Description automatically generatedThis is a categorical variable that identifies the college of an individual observation. It has no unit of measurement. It is a created variable to allow for the comparison of graduation rates among colleges and there are no missing values.

|  |  |
| --- | --- |
| **College** | **Frequency** |
| Carleton College | 126 |
| Grinnell College | 126 |
| Lawrence University | 126 |
| Macalester College | 90 |

The reason that the frequency of Macalester College is less than those of the other colleges is because there are fewer categories of race in Macalester College.

Macalester College is missing the categories of White and Non-resident Alien students due to which there are fewer instances of observations from Macalester College.

1. race

This is a categorical variable that identifies the race of an individual observation. It has no unit of measurement. This variable was extracted from the individual datasets of each college. It was wrangled to ensure that the factors of race are consistent among all the colleges. There are no missing values.

|  |  |
| --- | --- |
| **Race** | **Frequency** |
| Asian | 72 |
| Black/African American | 72 |
| Hispanic/Latino | 72 |
| Native Hawaiian/Pacific Islander | 72 |
| Non-resident Alien | 54 |
| Two or more races | 72 |
| White | 54 |

Chart, bar chart

Description automatically generated

As mentioned before, the reason for the lower frequencies of White and Non-resident Alien students is that Macalester College did not provide graduation rates for these groups of students. Since the data was wrangled to ensure that the categories of race are the same across all colleges, the frequencies of all race categories are the same (except that of White and Non-resident alien students, as explained above).

1. year

This is a numerical/quantitative variable based on the years that each data observation was collected or recorded. The variable was extracted from the individual datasets of each college. The years are consistent for each college regardless of if there is an observation; this happened by pivoting the dataset into long form during the Chart, histogram

Description automatically generatedprocessing step of the research.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Min. | 1st Quart. | Median | Mean | 3rd Quart. | Max. | Std. Dev |
| 2000 | 2004 | 2008 | 2008 | 2013 | 2017 | 5.19 |

Since this is a quantitative variable, the summary statistics describe the variable. The histogram shows the frequencies of various groups of years. Each year has a frequency of 26 and the histogram groups years by 5. The groups are as follows - 2000-2005, 2006-2010, 2011-2015, 2016-2017.; due to this grouping, the distribution of the histogram is not uniform.

1. grad\_rate

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Min. | 1st Quart. | Median | Mean | 3rd Quart. | Max. | Missing Values | Std. Dev |
| 0 | 0.69 | 0.818 | 0.76 | 0.875 | 1 | 195 | 0.4 |

This is a numerical/quantitative variable that describes the graduation rate of students from a specific college and a particular race during a certain year. The unit of measurement for graduation rates is percentage. This value was extracted from the individual datasets of each college. There are 195 missing values in this variable. This is because of several reasons. First, there are graduation rates that were not reported for certain groups of students for certain years that were converted into missing values. Second, while pivoting the appended dataset, missing values were created in response to

Chart, histogram

Description automatically generatedeach college reporting graduation rates over different time periods.

While the high number of missing values may be shocking, they are expected because of the data manipulation and wrangling steps taken during the processing file. The histogram is extremely left-skewed. The median and mean show that graduation rates on average are higher than 50% which is indicative of high graduation rates across students of all races and colleges. However, the standard deviation is 40% which is high and indicates high fluctuations in graduation rates in this dataset.

1. Sex dataset
2. college

Chart, bar chart

Description automatically generatedThis is a categorical variable that identifies the college of an individual observation. It has no unit of measurement. It is a created variable to allow for the comparison of graduation rates among colleges and there are no missing values.

|  |  |
| --- | --- |
| **College** | **Frequency** |
| Carleton College | 36 |
| Grinnell College | 36 |
| Lawrence University | 36 |

The frequencies of all colleges are the same because of the steps taken during the processing stage of the research, i.e., after pivoting the instances of each college became the same. By appending and pivoting the datasets, the same number of instances were created for each college.

1. sex

Chart

Description automatically generatedThis is a categorical variable that identifies the sex of an individual observation. It has no unit of measurement. This variable was extracted from the individual datasets of each college. It was wrangled to ensure that the factors of sex are consistent among all the colleges. There are no missing values.

|  |  |
| --- | --- |
| **Sex** | **Frequency** |
| Female | 54 |
| Male | 54 |

Similar to the frequencies of colleges, the frequencies of each sex category are the same because of the steps taken during the processing stage to ensure that the graduation rates for each gender are reported for each college.

1. year

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Min. | 1st Quart. | Median | Mean | 3rd Quart. | Max. | Std. Dev. |
| 2000 | 2004 | 2008 | 2008 | 2013 | 2017 | 5.21 |

This is a numerical/quantitative variable based on the years that each data observation was collected or recorded. The variable was extracted from the individual datasets of each college. The years are consistent for each college regardless of if there is an observation; this happened by pivoting the dataset into a long form during the processing step of the research.

Chart, histogram

Description automatically generatedThe histogram shows the frequencies of various groups of years. Each year has a frequency of 6 and the histogram groups the values by 5. The groups are as follows - 2000-2005, 2006-2010, 2011-2015, 2016-2017.; due to this grouping, the distribution of the histogram is not uniform.

1. grad\_rate

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Min. | 1st Quart. | Median | Mean | 3rd Quart. | Max. | Missing Values | Std. Dev |
| 0.53 | 0.68 | 0.842 | 0.787 | 0.884 | 0.949 | 28 | 0.363 |

This is a numerical/quantitative variable that describes the graduation rate of students from a specific college and a particular sex during a certain year. The unit of measurement for graduation rates is percentage. This value was extracted from the individual datasets of each college. There are 28 missing values in this variable. While pivoting the appended dataset, missing values were created because the years that each college reported graduation rates were not the same.

Chart, histogram

Description automatically generated

While the high number of missing values may be shocking, they are expected because of the data manipulation and wrangling steps taken during the processing file. The histogram appears to be slightly left-skewed. The median and mean show that graduation rates on average are higher than 50% which is indicative of high graduation rates across students of both races and all colleges. However, the standard deviation is 36% which is high and indicates high fluctuations in graduation rates in this dataset.

1. Pell dataset
2. college

Chart, bar chart

Description automatically generatedThis is a categorical variable that identifies the college of an individual observation. It has no unit of measurement. It is a created variable to allow for the comparison of graduation rates among colleges and there are no missing values.

|  |  |
| --- | --- |
| **College** | **Frequency** |
| Carleton College | 39 |
| Coe College | 39 |
| Grinnell College | 39 |
| Lawrence University | 39 |

The frequencies of all colleges are the same because of the steps taken during the processing stage of the research. By appending and pivoting the datasets, the same number of instances were created for each college.

1. pell\_status

This is a categorical variable that identifies the Pell status of an individual observation. It has no unit of measurement. This variable was extracted from the individual datasets of each college. It was wrangled to ensure that the factors of Pell status are consistent among all the colleges. There are no missing values.

A picture containing graphical user interface

Description automatically generated

|  |  |
| --- | --- |
| **Pell Status** | **Frequency** |
| Received Federal Pell Grant | 52 |
| Received Neither Pell Grant nor Stafford Loan | 52 |
| Received Subsidized Stafford Loan, No Pell Grant | 52 |

Similar to the frequencies of colleges, the frequencies of each category of Pell status are the same because of the steps taken during the processing stage to ensure that the graduation rates for each Pell status are reported for each college.

1. year

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Min. | 1st Quart. | Median | Mean | 3rd Quart. | Max. | Std. Dev. |
| 2005 | 2008 | 2011 | 2011 | 2014 | 2017 | 3.754 |

Chart, histogram

Description automatically generatedThis is a numerical/quantitative variable based on the years that each data observation was collected or recorded. The variable was extracted from the individual datasets of each college. The years are consistent for each college regardless of if there is an observation; this happened by pivoting the dataset into a long form during the processing step of the research.

Since this is a quantitative variable, the summary statistics describe the variable. The histogram shows the frequencies of various groups of years. Each year has a frequency of 12 and the histogram groups the values of year by 2. The last group is from 2016-2018, however, since the range of years in the dataset is 2005-2017, the frequency of the last column is 12 instead of 24.

1. grad\_rate

This is a numerical/quantitative variable that describes the graduation rate of students from a specific college and a particular Pell status during a certain year. The unit of measurement for graduation rates is percentage. This value was extracted from the individual datasets of each college. There are 51 missing values in this variable. While

Chart, histogram

Description automatically generatedpivoting the appended dataset, missing values were created because the years that each college reported graduation rates were not the same.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Min. | 1st Quart. | Median | Mean | 3rd Quart. | Max. | Missing Values | Std. Dev |
| 0.449 | 0.616 | 0.72 | 0.743 | 0.868 | 1 | 51 | 0.363 |

While the high number of missing values may be shocking, they are expected because of the data manipulation and wrangling steps taken during the processing file. The histogram is bi-nodal. The median and mean show that graduation rates on average are slightly higher than 50% which is indicative of high graduation rates across students of all Pell statuses and all colleges. However, the standard deviation is 36% which is high and indicates high fluctuations in graduation rates in this dataset.