



# **SUMMER BOOT CAMP PROJECT 2024**

By: Akhtar Raza

INDEX

INDEX

Sr No.	Topic	Page No.
1.	Cover Page	1
2.	Index	2
3.	List of Tables	3
4.	List of Figures	3
5.	Problem Statement/ Objective	4
6.	Data Dictionary	4
7.	Basic EDA	5-15
8.	Problem-1	16-18



# Data Dictionary:

- Names: Name of the university
- Apps: Number of applications received
- Accept: Number of applications accepted
- Enroll: Number of students enrolled
- Top10perc: Percentage of students from top 10% of their high school class
- Top25perc: Percentage of students from top 25% of their high school class
- F.Undergrad: Number of full-time undergraduates
- P.Undergrad: Number of part-time undergraduates
- Outstate: Out-of-state tuition
- Room.Board: Room and board cost
- Books: Estimated cost of books
- Personal: Estimated personal expenses
- PhD: Percentage of faculty with a PhD
- Terminal: Percentage of faculty with terminal degrees
- S.F.Ratio: Student-faculty ratio
- perc.alumni: Percentage of alumni who donate
- Expend: Instructional expenditure per student
- Grad.Rate: Graduation rate

# LIST OF FIGURES

- Distribution of Applications, Acceptances, and Enrollments:
  - Histogram or bar chart showing the number of applications, acceptances, and enrollments across universities.
- Top 10% and Top 25% High School Class Distribution:
  - Bar chart or pie chart displaying the percentage of students from the top 10% and top 25% of their high school classes.
- Undergraduate Population Distribution:
  - Bar chart showing the number of full-time and part-time undergraduates
- Out-of-State Tuition Costs:
  - Box plot or histogram illustrating the distribution of out-of-state tuition fees.
- Room and Board Costs:
  - Histogram or bar chart showing the distribution of room and board costs.
- Estimated Costs (Books and Personal Expenses):
  - Bar chart comparing estimated costs of books and personal expenses.
- Faculty Qualifications:
  - Bar chart showing the percentage of faculty with PhDs and terminal degrees.
- Student-Faculty Ratio:
  - Histogram or bar chart illustrating the distribution of student-faculty ratios.
- Alumni Donations:
  - Bar chart displaying the percentage of alumni who donate to their alma mater.
- Instructional Expenditure per Student:
  - Histogram or box plot showing the distribution of instructional expenditures per student.
- Graduation Rates:
  - Bar chart or histogram showing the distribution of graduation rates across universities.

## LIST OF TABLES

- Names
- Apps
- Accept
- Enroll
- Top10perc
- Top25perc
- F.Undergrad
- P.Undergrad
- Outstate
- Room.Board
- Books
- Personal
- PhD
- Terminal
- S.F.Ratio
- perc.alumni
- Expend
- Grad.Rate

# Overall Insights

- Identify the factors (applications, acceptance rate, enrollment, academic excellence, costs, faculty qualifications, student/faculty ratio, alumni donations, expenditures) most strongly associated with higher graduation rates.
- Provide recommendations for colleges to improve their graduation rates based on the data analysis.

# Methodology

## Data Cleaning:

- Handle missing values by imputing with the median.
- Remove or correct anomalies and outliers.
- Convert non-numeric columns to appropriate numeric types.
- Exploratory Data Analysis:
  - Calculate summary statistics for the relevant features.
  - Visualize data distributions and relationships using bar charts, histograms, scatter plots, and box plots.
- Correlation Analysis:
  - Compute the Pearson correlation coefficients between various numeric features.
  - Identify strong correlations with graduation rates and other key metrics.

## Statistical Analysis:

- Perform detailed analysis on key metrics such as acceptance rates, enrollment rates, academic excellence indicators, cost factors, and faculty qualifications.
- Draw insights and conclusions based on the statistical analysis and visualizations.

## Recommendations:

- Based on the insights gained from the analysis, provide actionable recommendations to educational institutions for improving their graduation rates and overall performance.

## Problem Statement

- The objective of this data science project is to analyze various factors that influence the performance and characteristics of colleges and universities, focusing on their impact on graduation rates. By examining different aspects such as applications, enrollment, academic excellence, student demographics, costs, faculty qualifications, student-faculty interaction, and alumni engagement, the goal is to identify key determinants of higher graduation rates and provide actionable recommendations for educational institutions.

## Objectives

### Application and Enrollment Analysis



- Determine the average number of applications received by colleges.
- Calculate the average acceptance rate across all colleges.
- Calculate the average enrollment rate (number of students enrolled divided by the number of applications accepted).
- Identify the college with the highest number of applications received.

### Academic Excellence

- Determine the average percentage of new students from the top 10% of their higher secondary class across all colleges.
- Determine the average percentage of new students from the top 25% of their higher secondary class.
- Investigate the correlation between the percentage of students from the top 10% and the top 25% of their higher secondary class.

### Student Demographics

- Calculate the average number of full-time undergraduate students per college.
- Calculate the average number of part-time undergraduate students per college.
- Identify the college with the highest number of out-of-state students.



## Cost and Spending

- Determine the average cost of room and board across all colleges.
- Calculate the average estimated book cost for a student.
- Calculate the average estimated personal spending for a student.
- Analyze how instructional expenditure per student varies across colleges.

## Faculty Qualifications

- Determine the average percentage of faculty with Ph.D.s across all colleges.
- Determine the average percentage of faculty with terminal degrees.
- Investigate the correlation between the percentage of faculty with Ph.D.s and the graduation rate.

## Student-Faculty Interaction

- Calculate the average student/faculty ratio across all colleges.
- Identify the college with the lowest student/faculty ratio.
- Investigate the correlation between the student/faculty ratio and the graduation rate.

## Alumni Engagement

- Determine the average percentage of alumni who donate across all colleges.
- Investigate the correlation between the percentage of alumni who donate and the graduation rate.

## Graduation Rates

- Determine the average graduation rate across all colleges.
- Identify the college with the highest graduation rate.
- Investigate the correlation between instructional expenditure per student and the graduation rate.



## Importing the necessary Libraries

+ 1 cell hidden

## Loading the Dataset

+ 1 cell hidden

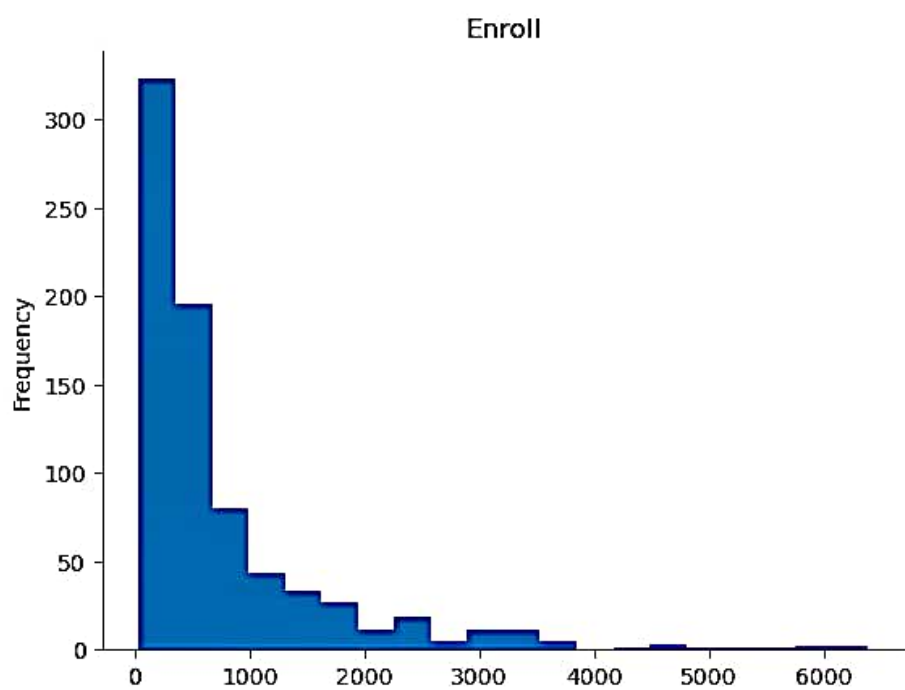
## Exploratory Dataset Analysis And Data Cleaning

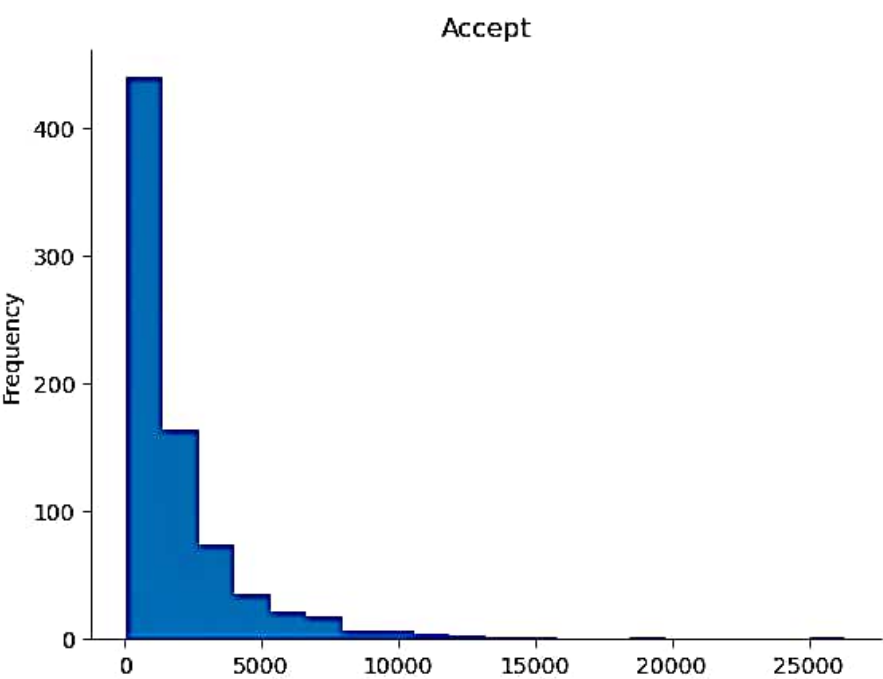
First Five Rows	<div>📄 ⬆ ⬇ ⬇ ⬆ 🗑</div>
-----------------	------------------------

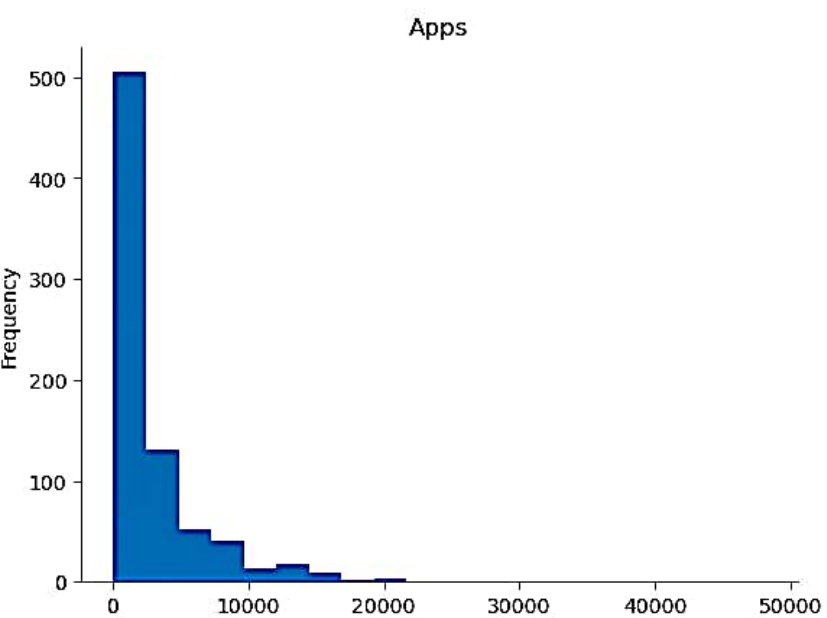
Names	Abilene Christian University	Adelphi University	Adrian College	Agnes Scott College	Alaska Pacific University
Apps	1660.0	2186.0	1428.0	417.0	193.0
Accept	1232	1924	1097	349	146
Enroll	721.0	512.0	336.0	NaN	55.0
Top10perc	23.0	16.0	22.0	60.0	16.0
Top25perc	52	29	50	89	44
F.Undergrad	2885	2683	1036	510	249
P.Undergrad	537	1227	99	63	869
Outstate	7440	12280	11250	12960	7560
Room.Board	3300	6450	3750	5450	4120
Books	450	750	400	450	800
Personal	2200.0	1500.0	1165.0	875.0	1500.0
PhD	70	29	53	92	76
Terminal	78	30	66	97	72
S.F.Ratio	18.1	?	12.9	7.7	11.9
perc.alumni	12	16	30	37	2
Expend	7041	10527	8735	19016	10922
Grad.Rate	60	56	54	59	15

## Last Five Rows

	772	773	774	775	776
Names	Worcester State College	Xavier University	Xavier University of Louisiana	Yale University	York College of Pennsylvania
Apps	2197.0	1959.0	2097.0	10705.0	2989.0
Accept	1515	1805	1915	2453	1855
Enroll	543.0	695.0	695.0	1317.0	691.0
Top10perc	4.0	24.0	34.0	95.0	28.0
Top25perc	26	47	61	99	63
F.Undergrad	3089	2849	2793	5217	2988
P.Undergrad	2029	1107	166	83	1726
Outstate	6797	11520	6900	19840	4990
Room.Board	3900	4960	4200	6510	3560
Books	500	600	617	630	500
Personal	1200.0	1250.0	781.0	2115.0	1250.0
PhD	60	73	67	96	75
Terminal	60	75	75	96	75
S.F.Ratio	21	13.3	14.4	5.8	18.1
perc.alumni	14	31	20	49	28
Expend	4469	9189	8323	40386	4509
Grad.Rate	40	83	49	99	99







```
Names      object
Apps       float64
Accept     int64
Enroll     float64
Top10perc  float64
Top25perc  int64
F.Undergrad int64
P.Undergrad int64
Outstate   int64
Room.Board int64
Books      int64
Personal   float64
PhD        int64
Terminal   int64
S.F.Ratio  object
perc.alumni int64
Expend     int64
Grad.Rate  int64
dtype: object
```

## These Are columns Name And Data Types

## Observations

- Features are having null values
- The data types of "S.F.Ratio" should be float but it is appearing as object, so we need to check that.
- The Apps column shows the number of applications received by each university. It is a float type but it should be as whole number(integer)
- Top10perc and Top25perc show the percentage of students who were in the top 10% and top 25% of their high school class, respectively. Top10perc is a float, while Top25perc is an integer. So we have to check that.



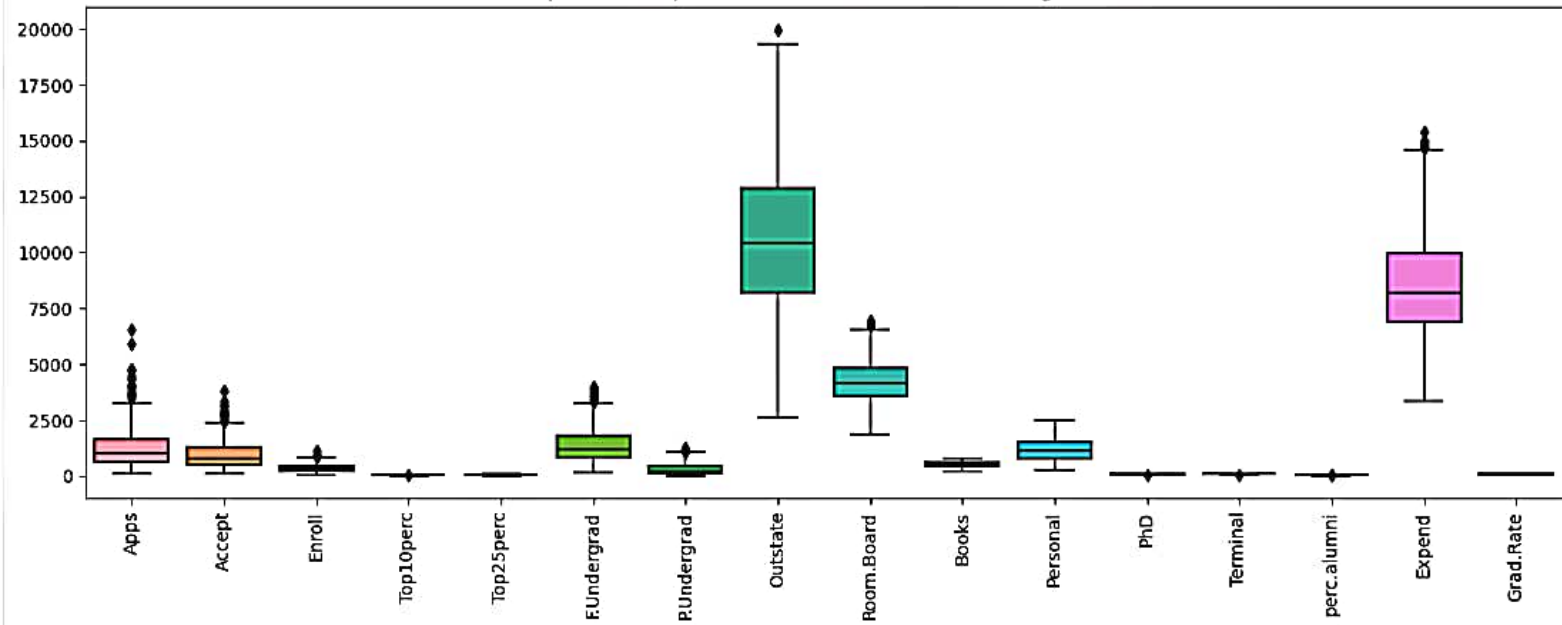
```
] : Names      object
Apps      float64
Accept     int64
Enroll     float64
Top10perc  float64
Top25perc  int64
F.Undergrad  int64
P.Undergrad  int64
Outstate   int64
Room.Board  int64
Books      int64
Personal   float64
PhD        int64
Terminal    int64
S.F.Ratio   object
perc.alumni  int64
Expend      int64
Grad.Rate   int64
dtype: object
```

These Are columns Name And Data Types

## Observations:

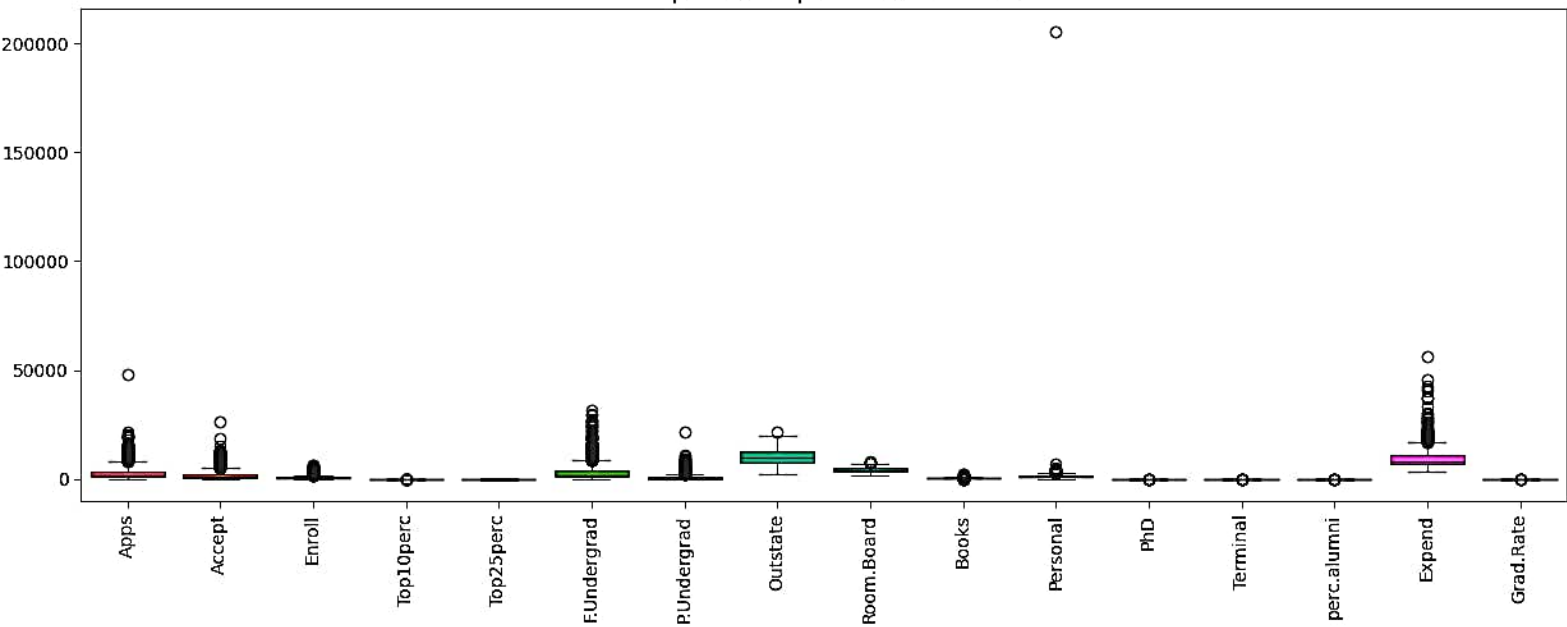
- The number of Rows in our dataset is 777.
- The number of Columns in our dataset is 18.

Boxplots of all quantitative columns after removing outliers



First check for outliers

Boxplots of all quantitative columns

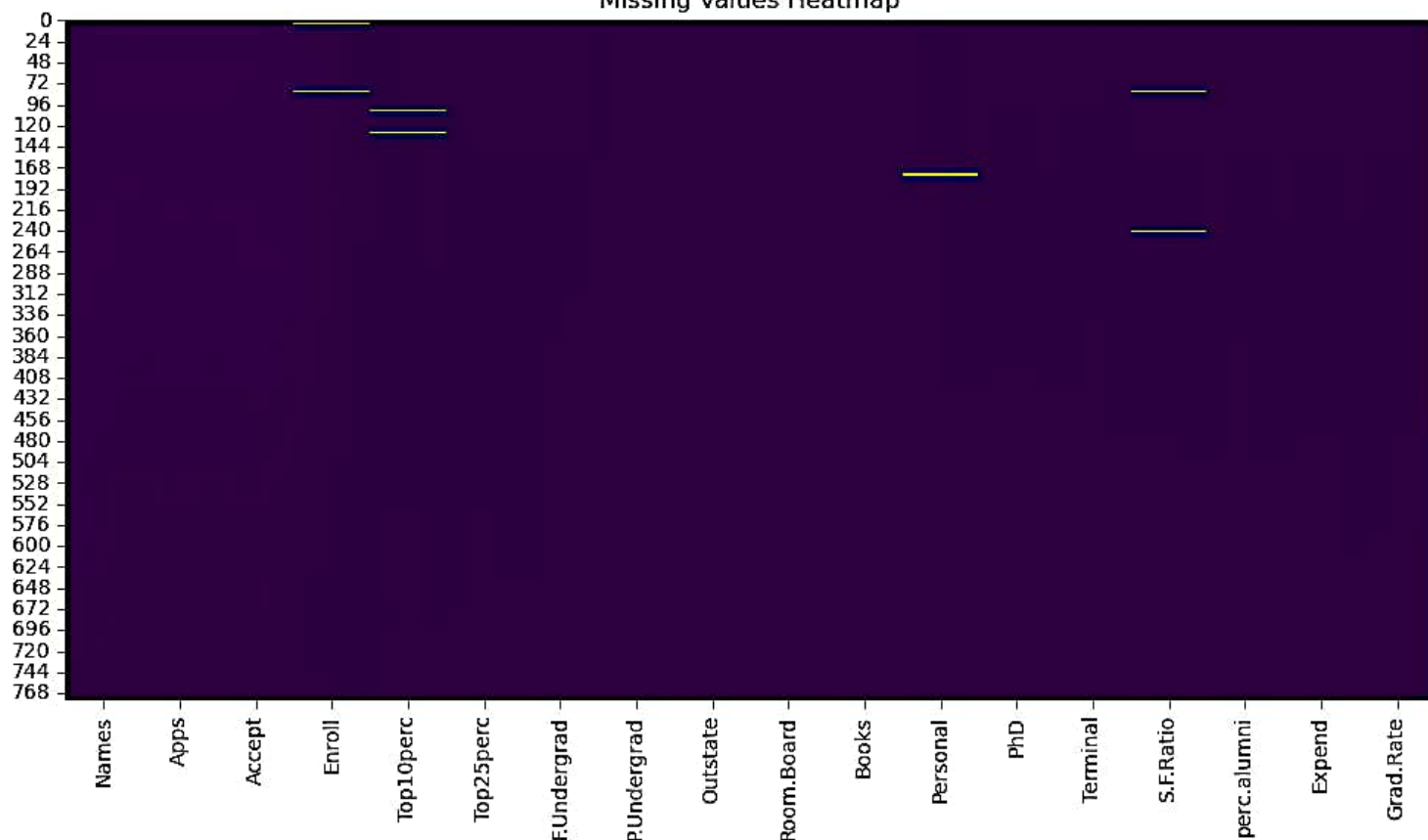


```
# Check for percentage wise missing values in columns  
data.isnull().sum()/len(data)*100
```

```
Names          0.000000  
Apps           0.000000  
Accept         0.000000  
Enroll         0.257400  
Top10perc      0.514801  
Top25perc      0.000000  
F.Undergrad    0.000000  
P.Undergrad    0.000000  
Outstate       0.000000  
Room.Board     0.000000  
Books          0.000000  
Personal       0.386100  
PhD            0.000000  
Terminal       0.000000  
S.F.Ratio      0.386100  
perc.alumni    0.000000  
Expend         0.000000  
Grad.Rate      0.000000  
dtype: float64
```

Check for percentage wise missing values in columns

Missing Values Heatmap





	Names	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alum
1	Adelphi University	2186	1924	512.0	16.0	29	2683	1227	12280	6450	750	1500.0	29	30	NaN	
3	Agnes Scott College	417	349	NaN	60.0	89	510	63	12960	5450	450	875.0	92	97	7.7	
81	Campbell University	2087	1339	NaN	20.0	54	3191	1204	7550	2790	600	500.0	77	77	NaN	
102	Central Connecticut State University	4158	2532	902.0	NaN	24	6394	3881	5962	4444	500	985.0	69	73	16.7	
103	Central Missouri State University	4681	4101	1436.0	NaN	35	8094	1596	4620	3288	300	2250.0	69	80	19.7	
128	College of Notre Dame	344	264	97.0	NaN	42	500	331	12600	5520	630	2250.0	77	80	10.4	
129	College of Notre Dame of Maryland	457	356	177.0	NaN	61	667	1983	11180	5620	600	700.0	64	64	11.5	
166	Dillard University	1998	1376	651.0	41.0	88	1539	45	6700	3650	500	NaN	52	52	14.1	
175	Earlham College	1358	1006	274.0	35.0	63	1028	13	15036	4056	600	NaN	90	94	10.6	
177	East Tennessee State University	3330	2730	1303.0	15.0	36	6706	2640	5800	3000	600	NaN	73	75	14.0	
241	Gwynedd Mercy College	380	237	104.0	30.0	56	716	1108	11000	5550	500	500.0	36	41	NaN	

Filled NaN in place missing in S.F.Ratio

```
<bound method Series.sum of 0      False
2      False
5      False
6      False
7      False
...
769    False
770    False
771    False
773    False
774    False
Length: 448, dtype: bool>
```

## Checking Duplicate values

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 777 entries, 0 to 776
Data columns (total 18 columns):
#   Column          Non-Null Count  Dtype
---  -
0   Names           777 non-null   object
1   Apps            777 non-null   int64
2   Accept          777 non-null   int64
3   Enroll          775 non-null   float64
4   Top10perc       773 non-null   float64
5   Top25perc       777 non-null   int64
6   F.Undergrad     777 non-null   int64
7   P.Undergrad     777 non-null   int64
8   Outstate        777 non-null   int64
9   Room.Board      777 non-null   int64
10  Books           777 non-null   int64
11  Personal        774 non-null   float64
12  PhD             777 non-null   int64
13  Terminal        777 non-null   int64
14  S.F.Ratio       774 non-null   float64
15  perc.alumni     777 non-null   int64
16  Expend          777 non-null   int64
17  Grad.Rate       777 non-null   int64
dtypes: float64(4), int64(13), object(1)
memory usage: 109.4+ KB
```

## Converting S.F.Ratio into float

# Data Cleaning

## 1- Correcting Data types

## Observations

- Books: Minimum value of 0 is not realistic.
- Personal: Minimum value of 50 is too low.

books\_zero\_data = data[data['Books']==0]

print(books\_zero\_data)

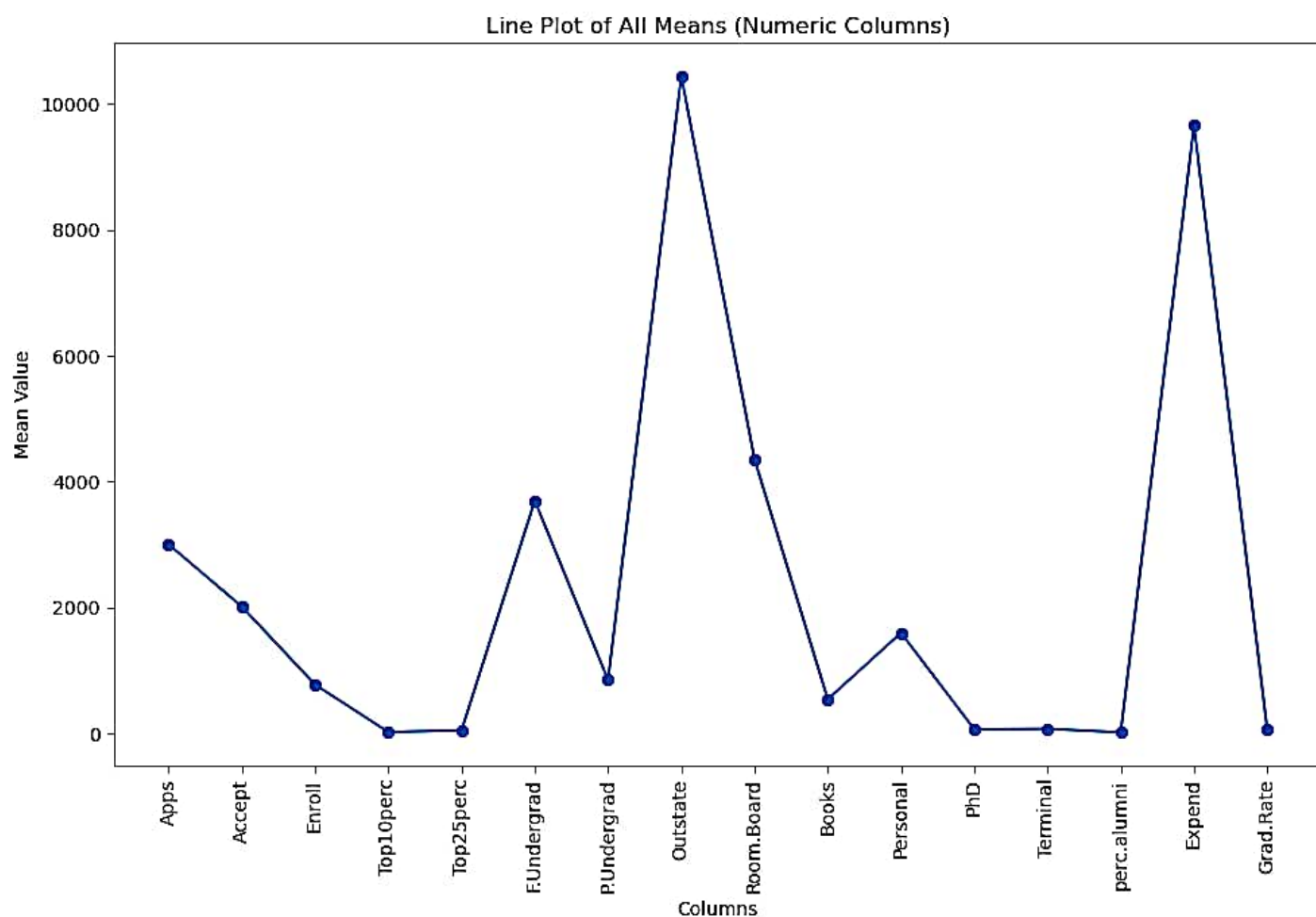
data[data['Personal']==50]

	Names	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alun
180	Eastern Connecticut State University	2172.0	1493	564.0	14.0	50	2766	1531	5962	4316	650	50.0	71	76	16.9	

# Observations

\* Books value is zero at Catawba College and Catawba College.

\* Concordia University Ca is also has zero value



Line plot of means

	count	mean	std	min	25%	50%	75%	max
<b>Apps</b>	448.0	1274.979911	955.728071	81.0	597.75	981.5	1660.00	6548.0
<b>Accept</b>	448.0	960.430804	644.840889	72.0	484.50	791.5	1266.50	3813.0
<b>Enroll</b>	448.0	352.294643	197.900304	46.0	199.50	307.0	464.25	1123.0
<b>Top10perc</b>	448.0	24.591518	12.525611	1.0	15.00	22.0	33.00	62.0
<b>Top25perc</b>	448.0	52.595982	17.402195	9.0	40.00	52.0	65.25	93.0
<b>F.Undergrad</b>	448.0	1422.533482	808.391506	139.0	828.00	1199.0	1819.25	3957.0
<b>P.Undergrad</b>	448.0	303.933036	281.912373	1.0	74.00	213.5	466.00	1235.0
<b>Outstate</b>	448.0	10535.060268	3466.235253	2580.0	8172.50	10435.0	12850.00	19964.0
<b>Room.Board</b>	448.0	4245.225446	952.350757	1880.0	3600.00	4140.0	4821.50	6950.0
<b>Books</b>	448.0	514.493304	98.069965	225.0	450.00	500.0	600.00	800.0
<b>Personal</b>	448.0	1157.843750	483.142030	250.0	800.00	1100.0	1500.00	2500.0
<b>PhD</b>	448.0	69.390625	15.138484	26.0	59.00	71.0	80.00	103.0
<b>Terminal</b>	448.0	76.553571	14.266771	33.0	67.00	78.0	89.00	100.0
<b>perc.alumni</b>	448.0	24.328125	11.301320	2.0	16.00	24.0	32.00	58.0
<b>Expend</b>	448.0	8596.354911	2481.720429	3365.0	6878.75	8192.5	9981.25	15365.0
<b>Grad.Rate</b>	448.0	65.883929	15.835562	21.0	55.00	67.0	78.00	100.0

## Statical Summary



	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
0	2200.0	70	78	18.1	12	7041	60
2	1165.0	53	66	12.9	30	8735	54
5	675.0	67	73	9.4	11	9727	55
6	1500.0	90	93	11.5	26	8861	63
7	850.0	89	100	13.7	37	11487	73
..	...	...	...	...	...	...	...
765	1550.0	69	81	13.9	8	7264	91
768	1400.0	48	48	8.5	26	8960	50
769	800.0	82	95	12.8	29	10414	78
770	1440.0	91	92	15.3	42	7875	75
774	781.0	67	75	14.4	20	8323	49

[347 rows x 18 columns]

## Cleaned Data

	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	\
0	52	2885	537	7440	3300	450	
2	50	1036	99	11250	3750	400	
5	62	678	41	13500	3335	500	
6	45	416	230	13290	5720	500	
7	68	1594	32	13868	4826	450	
..	...	...	...	...	...	...	
765	34	1207	157	7820	3400	550	
768	41	282	22	9100	3700	500	
769	68	1980	144	15948	4404	400	
770	83	1059	34	12680	4150	605	
774	61	2793	166	6900	4200	617	

	Names	Apps	Accept	Enroll	Top10perc	\
0	Abilene Christian University	1660.0	1232	721.0	23.0	
2	Adrian College	1428.0	1097	336.0	22.0	
5	Albertson College	587.0	479	158.0	38.0	
6	Albertus Magnus College	353.0	340	103.0	17.0	
7	Albion College	1899.0	1720	489.0	37.0	
..	...	...	...	...	...	
765	Wingate College	1239.0	1017	383.0	10.0	
768	Wisconsin Lutheran College	152.0	128	75.0	17.0	
769	Wittenberg University	1979.0	1739	575.0	42.0	
770	Wofford College	1501.0	935	273.0	51.0	
774	Xavier University of Louisiana	2097.0	1915	695.0	34.0	

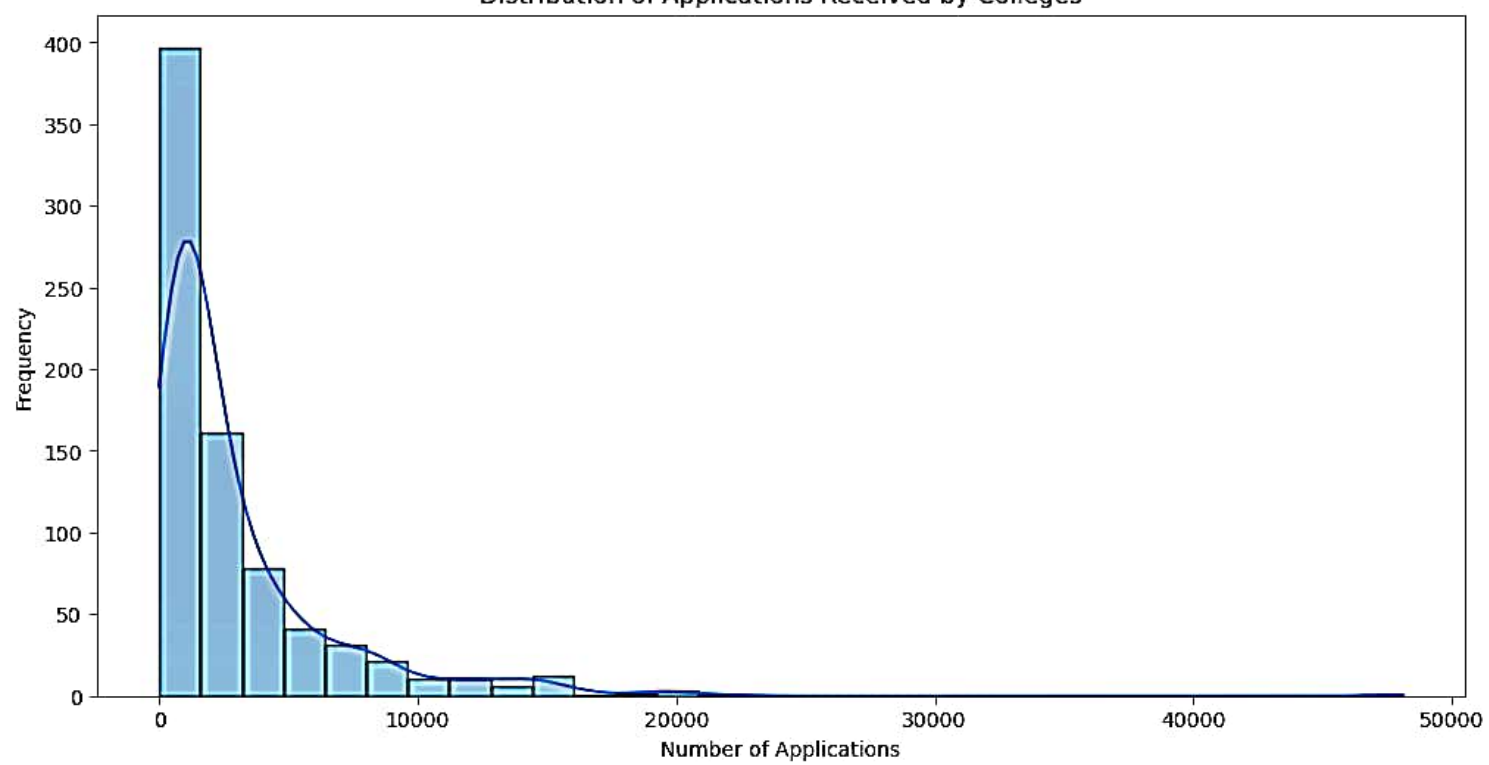
## Application and Enrollment Analysis

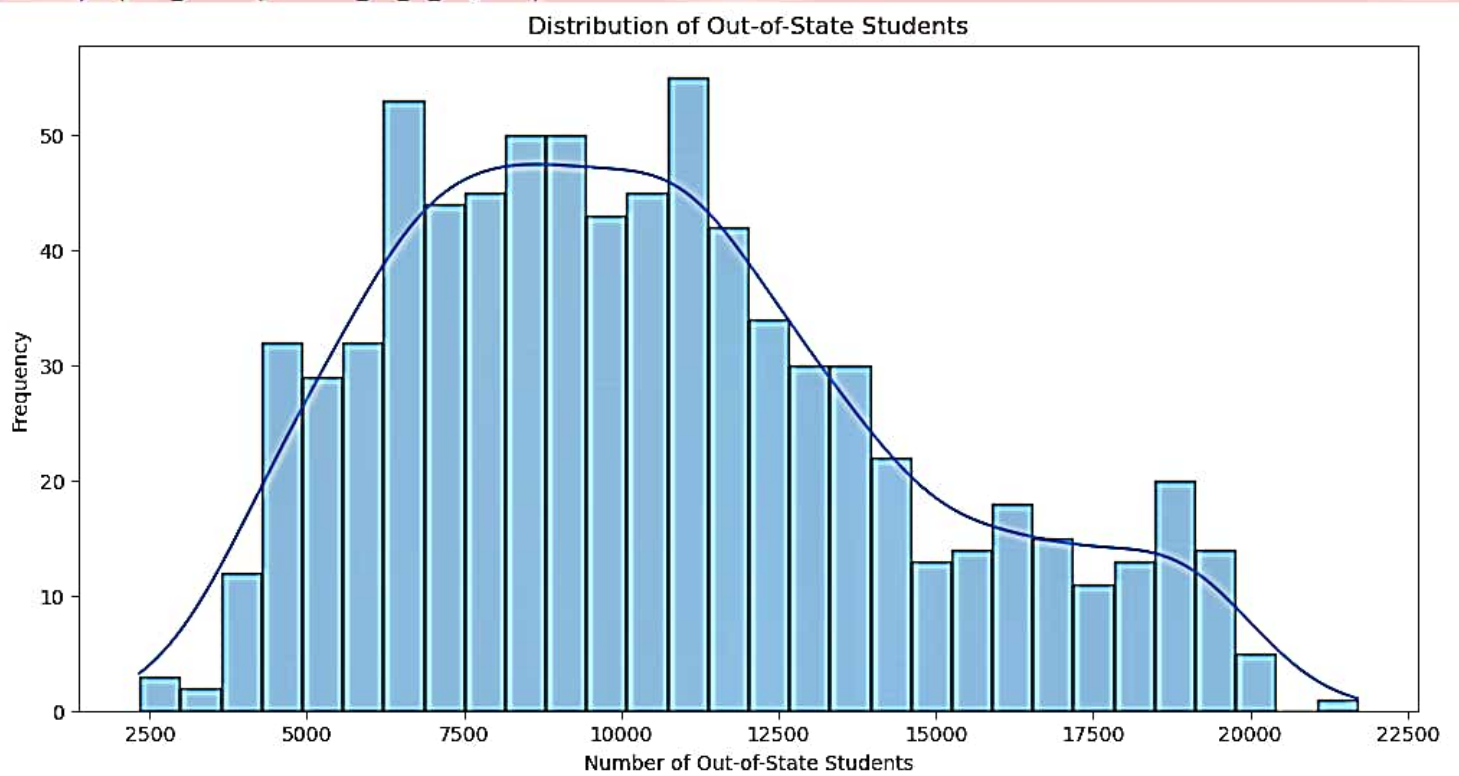
- What is the average number of applications received by colleges?
- What percentage of applications are accepted on average across all colleges?
- What is the average enrollment rate (number of students enrolled divided by number of applications accepted)?
- Which college has the highest number of applications received?

## Answers

- Average number of applications received: 1282.923250564334
- Average acceptance rate: inf
- Average enrollment rate: 40.6806479936538
- % College with the highest number of applications: Bucknell University

Distribution of Applications Received by Colleges





**College with the highest number of out-of-state students: Bennington College**

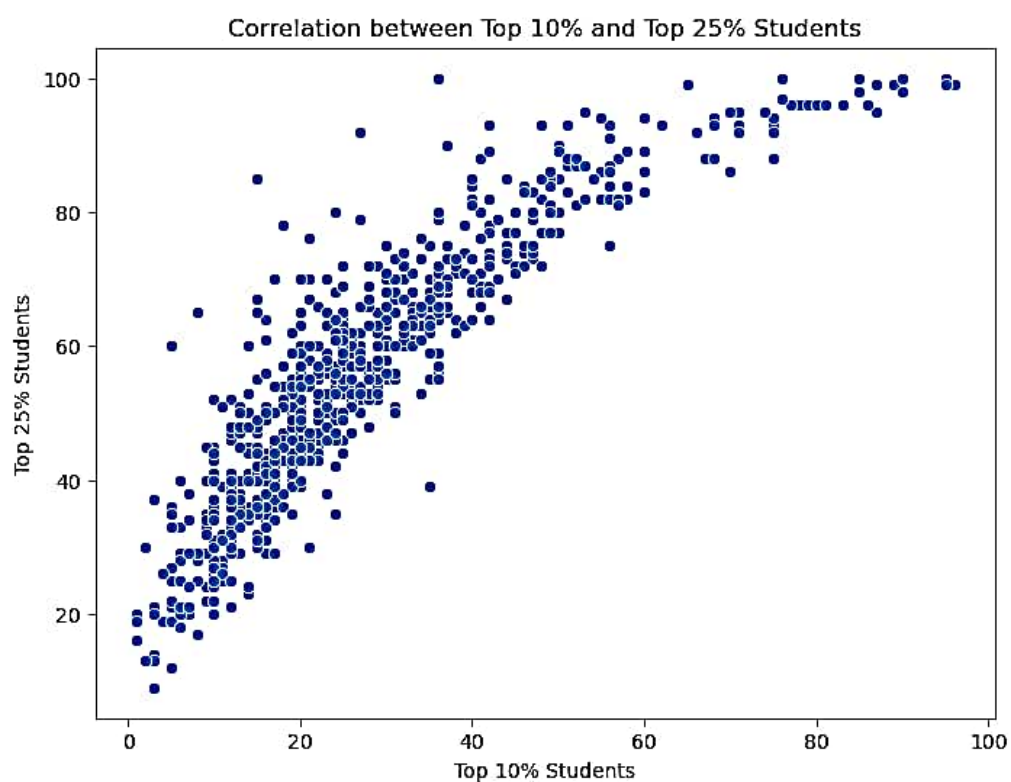
## Graduation Rates

- What is the average graduation rate across all colleges
- • Which college has the highest graduation rat
- • Is there a correlation between the instructional expenditure per student and the graduation rate?

## Answers

- Average graduation rate: 66.058690744921%
- College with the highest graduation rate: College of Mount St. Josep
- Correlation between instructional expenditure and graduation rate: 0.355992583405635





## 7 Alumni Engagement

- What is the average percentage of alumni who donate across all colleges
- Is there a correlation between the percentage of alumni who donate and the graduation rate?

## Answers

📄 ⬆ ⬇ ⬅ ➡ 🔍

- Average percentage of alumni who donate: 24.611738148984198%
- Correlation between alumni donations and graduation rate: 0.40513854471067384

## Student-Faculty Interaction







- What is the average student/faculty ratio across all colleges?
- Which college has the lowest student/faculty ratio?
- Is there a correlation between the student/faculty ratio and the graduation rate?

## Answers

- Average student/faculty ratio: 13.451241534988714
- College with the lowest student/faculty ratio: Sarah Lawrence Colleg
- Correlation between student/faculty ratio and graduation rate: -0.15579649410338467

## Faculty Qualifications

- What is the average percentage of faculties with Ph.D.s across all colleges
- • What is the average percentage of faculties with terminal degree
- • Is there a correlation between the percentage of faculties with Ph.D.s and the graduation rate?

## Answers

- Average percentage of faculties with Ph.D.s: 69.76749435665914%
- Average percentage of faculties with terminal degrees: 76.85778781038374
- Correlation between Ph.D.s and graduation rate: 0.33427167204722763

## Cost and Spending

- What is the average cost of room and board across all colleges
- • What is the average estimated book cost for a studen
- • What is the average estimated personal spending for a stude
- • How does the instructional expenditure per student vary across colleges?

### Answers

- Average cost of room and board: 4259.200902934537
- Average estimated book cost: 515.446952595936
- AvAverage estimated personal spending: 1159.2392776523
- 2 Average instructional expenditure per student: 8686.270880361173

## Student Demographics

- What is the average number of full-time undergraduate students per college
- • What is the average number of part-time undergraduate students per colleg
- • Which college has the highest number of out-of-state students?

## Answers ¶

- Average number of full-time undergraduate students: 1423.3205417607223
- Average number of part-time undergraduate students: 301.2889390519187
- College with the highest number of out-of-state students: Gettysburg Collegeg

# Academic Excellence

- What is the average percentage of new students from the top 10% of their higher secondary class across all colleges
- • What is the average percentage of new students from the top 25% of their higher secondary clas
- • Is there a correlation between the percentage of students from the top 10% and the top 25% of their higher secondary class?

# Answers

- Average percentage of new students from the top 10%: 24.851015801354404%
- Average percentage of new students from the top 25%: 53.00451467268623
- Correlation between top 10% and top 25%: 0.90192003495869\*

# Overall Insights

- Which factors (applications, acceptance rate, enrollment, academic excellence, costs, faculty qualifications, student/faculty ratio, alumni donations, expenditures) are most strongly associated with higher graduation rates?

# Factors strongly associated with higher graduation rates:

* Grad.Rate=	1.000000		
* Outstate=	0.495601		
* perc.alumni=	0.405139		
* Top25perc=	0.393640		
* Top10perc=	0.382983	Room.Board	0.35628
* Expend=	0.355993		
* Apps=	0.353564		
* PhD=	0.334272		
* Accept=	0.327748		
* Terminal=	0.323001		
* Enroll=	0.241371		
* F.Undergrad=	0.197217		
* Books=	0.024272		
* P.Undergrad=	0.005998		
* S.F.Ratio=	-0.155796		
* Personal=	-0.212902		
* Name: Grad.Rate, dtype: float64			

## Recommendation

Recommendations to improve graduation rates based on the data analysis:

1. Increase faculty qualifications, particularly the percentage of faculty with Ph.D.s and terminal degrees.
2. Improve student-faculty interaction by lowering the student/faculty ratio.
3. Encourage alumni engagement and donations, as there is a positive correlation with graduation rates.
4. Increase instructional expenditure per student to support academic and instructional resources.
5. Focus on attracting top students from higher secondary classes, particularly those in the top 10% and 25%.
6. Maintain a balanced acceptance rate to manage the quality of incoming students and their potential graduation outcomes.