SUMMER BOOTCAMP PROJECT 2024

BY PULLE THIRUPATHI

2022513603

Education Post - 12th Data Analysis

List of Tables

List of Figures

Problem Statement

The objective of this analysis is to gain insights into the characteristics of colleges and answer key questions related to the educational landscape. By understanding the data, we aim to inform strategies for improving the quality of education and enhancing the overall college experience. The analysis will provide valuable insights and recommendations for stakeholders in the education sector.

Data Dictionary

- Names: Names of various university and colleges
- Apps: Number of applications received
- Accept: Number of applications accepted
- Enroll: Number of new students enrolled
- Top10perc: Percentage of new students from top 10% of Higher Secondary class
- Top25perc: Percentage of new students from top 25% of Higher Secondary class
- F.Undergrad: Number of full-time undergraduate students
- P.Undergrad: Number of part-time undergraduate students
- Outstate: Number of students for whom the particular college or university is Out-of-state tuition
- Room.Board: Cost of Room and board
- Books: Estimated book costs for a student
- Personal: Estimated personal spending for a student
- PhD: Percentage of faculties with Ph.D.'s
- Terminal: Percentage of faculties with terminal degree
- S.F.Ratio: Student/faculty ratio
- perc.alumni: Percentage of alumni who donate
- Expend: The Instructional expenditure per student
- Grad.Rate: Graduation rate

Basic Steps

1- Display the top 5 rows.

2- Display the last 5 rows.

3- Check the shape of dataset.

4- Check the datatypes of each feature.

5- Check the Statistical summary.

6- Check the null values.

7- Check the duplicate values.

8- Check the anomalies or wrong entries.

9- Check the outliers and their authenticity.

10- Do the necessary data cleaning steps like dropping duplicates, unnecessary columns, null value imputation, outliers treatment etc.

The top 5 rows are as displayed:

	0	1	2	3	4
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

Observations

In the column; Enroll, there is an entry given as NaN which is unique in the head dataset.

In the column; S.F.Ratio, there is an entry given as ? which seems to be a wrong entry in the head dataset.

The last 5 rows are as displayed:

	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

The shape of the dataset:

```
(777, 18)
```

Observations:

There are 777 rows and 18 columns in this dataset.

Datatypes of each feature:

```
<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         775 non-null    float64
 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    777 non-null    object 
 15  perc.alumni  777 non-null    int64  
 16  Expend       777 non-null    int64  
 17  Grad.Rate    777 non-null    int64  
dtypes: float64(4), int64(12), object(2)
memory usage: 109.4+ KB
```

Observations:

Number of strucrures of integer datatypes = 12

Number of strucrures of float datatypes = 4

Number of strucrures of object datatypes = 2

Statistical Data:

	count	mean	std	min	25%	50%	75%	max
Apps	775.0	3007.592258	3873.414660	81.0	778.0	1561.0	3635.0	48094.0
Accept	777.0	2018.804376	2451.113971	72.0	604.0	1110.0	2424.0	26330.0
Enroll	775.0	780.961290	930.077779	35.0	242.5	434.0	902.5	6392.0
Top10perc	773.0	27.620957	17.645470	1.0	15.0	23.0	35.0	96.0
Top25perc	777.0	55.796654	19.804778	9.0	41.0	54.0	69.0	100.0
F.Undergrad	777.0	3699.907336	4850.420531	139.0	992.0	1707.0	4005.0	31643.0
P.Undergrad	777.0	855.298584	1522.431887	1.0	95.0	353.0	967.0	21836.0
Outstate	777.0	10440.669241	4023.016484	2340.0	7320.0	9990.0	12925.0	21700.0
Room.Board	777.0	4357.526384	1096.696416	1780.0	3597.0	4200.0	5050.0	8124.0
Books	777.0	547.875161	167.426237	0.0	465.0	500.0	600.0	2340.0
Personal	774.0	1601.507752	7369.594038	50.0	855.0	1200.0	1687.5	205500.0
PhD	777.0	72.660232	16.328155	8.0	62.0	75.0	85.0	103.0
Terminal	777.0	79.702703	14.722359	24.0	71.0	82.0	92.0	100.0
perc.alumni	777.0	22.743887	12.391801	0.0	13.0	21.0	31.0	64.0
Expend	777.0	9660.171171	5221.768440	3186.0	6751.0	8377.0	10830.0	56233.0
Grad.Rate	777.0	65.463320	17.177710	10.0	53.0	65.0	78.0	118.0

Observations:

This whole data consists of a lot of outliers which needs to be removed, for a more qualitative analysis.

The null values are:

```
Names          0
Apps           2
Accept         0
Enroll         2
Top10perc      4
Top25perc      0
F.Undergrad    0
P.Undergrad    0
Outstate       0
Room.Board     0
Books          0
Personal       3
PhD            0
Terminal       0
S.F.Ratio      0
perc.alumni    0
Expend         0
Grad.Rate      0
dtype: int64
```

Observations:

There are 2 null values in column Apps, 2 null values in column Enroll, 4 null values in column Top25perc, 3 null values in column Personal.

Filling the null values:

Using 'fillna()' function

```
Names          0
Apps           0
Accept         0
Enroll         0
Top10perc      0
Top25perc      0
F.Undergrad    0
P.Undergrad    0
Outstate       0
Room.Board     0
Books          0
Personal       0
PhD            0
Terminal       0
S.F.Ratio      0
perc.alumni    0
Expend         0
Grad.Rate      0
dtype: int64
```

Duplicate Values:

```
0      False
1      False
2      False
3      False
4      False
       ...
772    False
773    False
774    False
775    False
776    False
Length: 777, dtype: bool
```

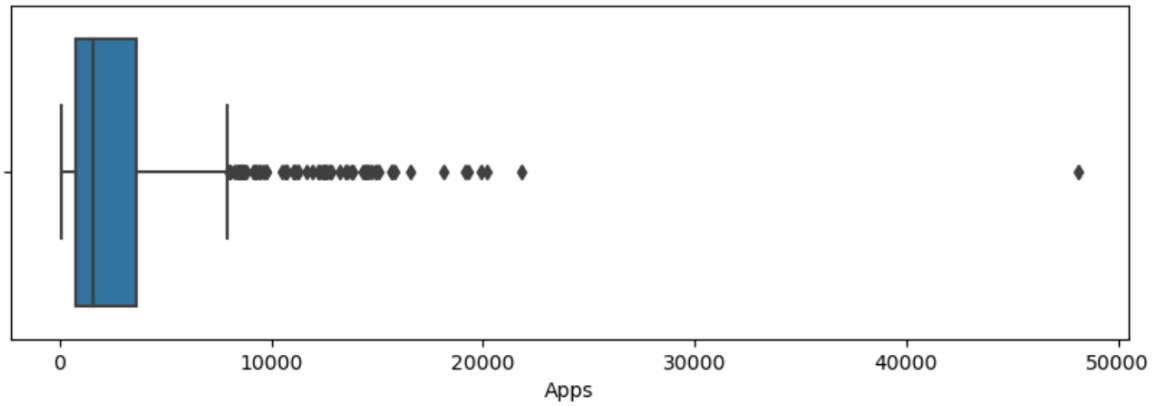
Number of duplicate values:

```
0
```

Observations

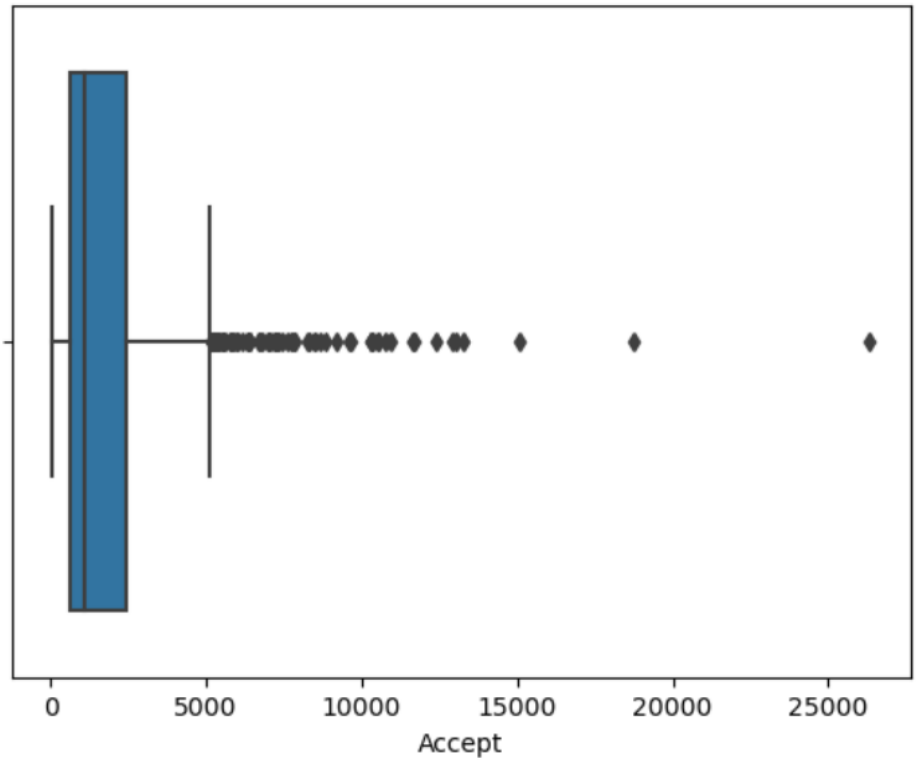
There are 0 duplicate values in the dataset.

Outliers and Their Authencity



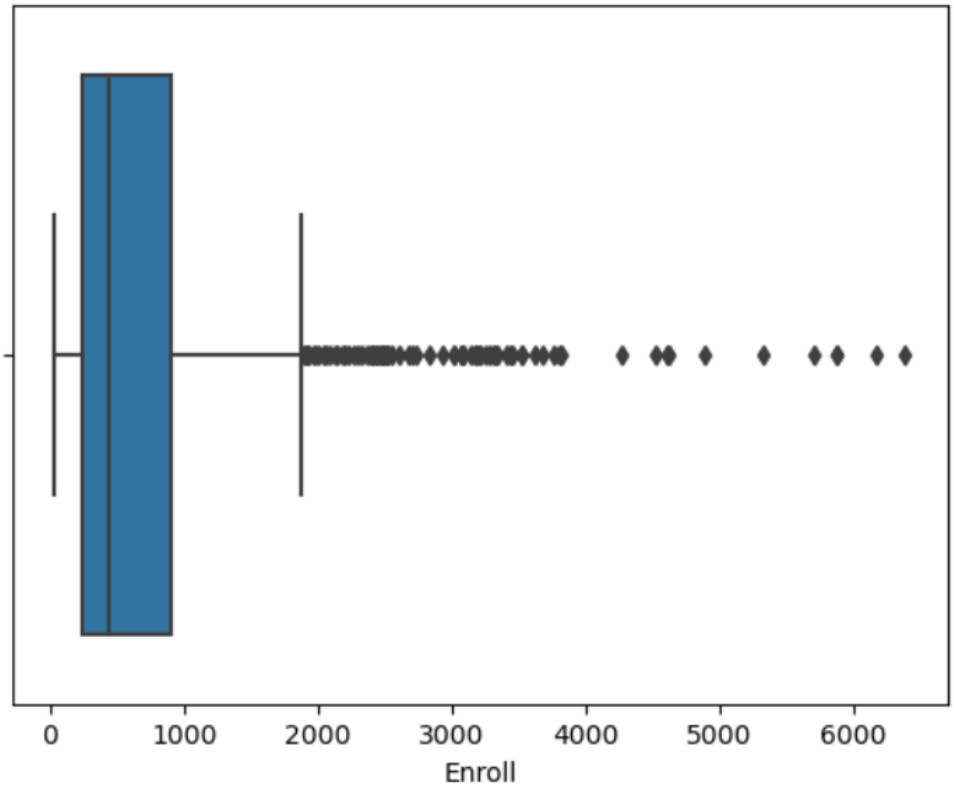
Observation:

In column Apps,the outlier with the highest value is 48094.



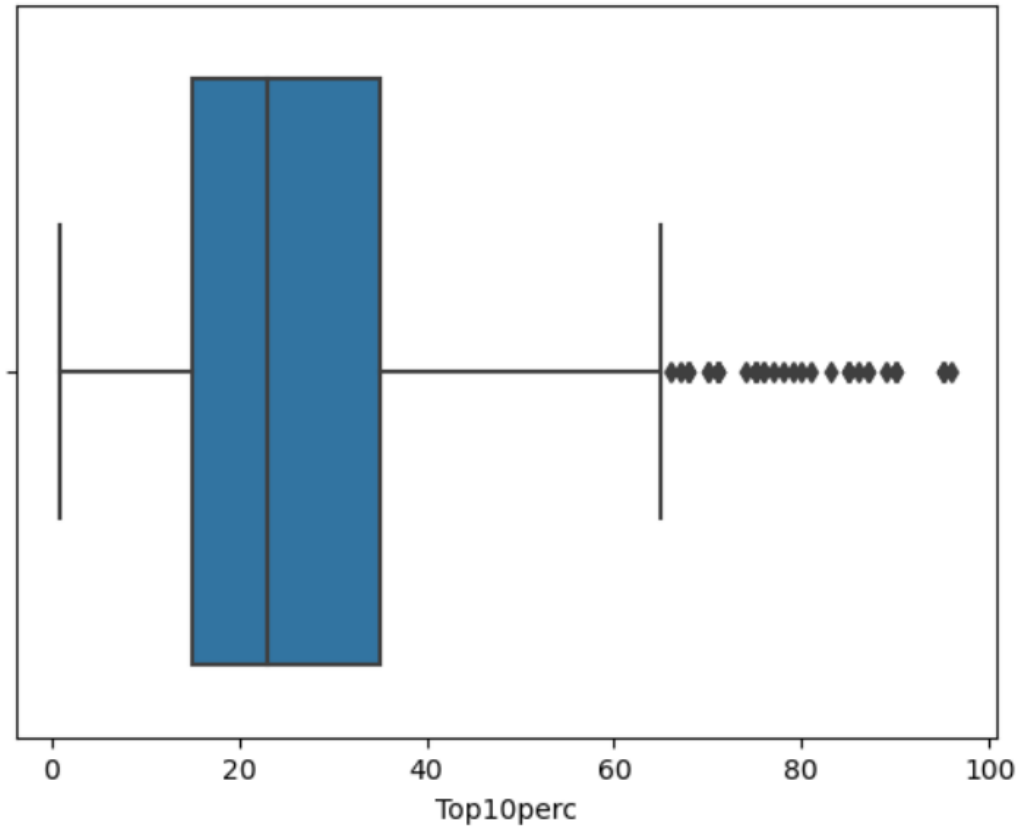
Observations:

In column Accept, the outlier with the highest value is 26330.



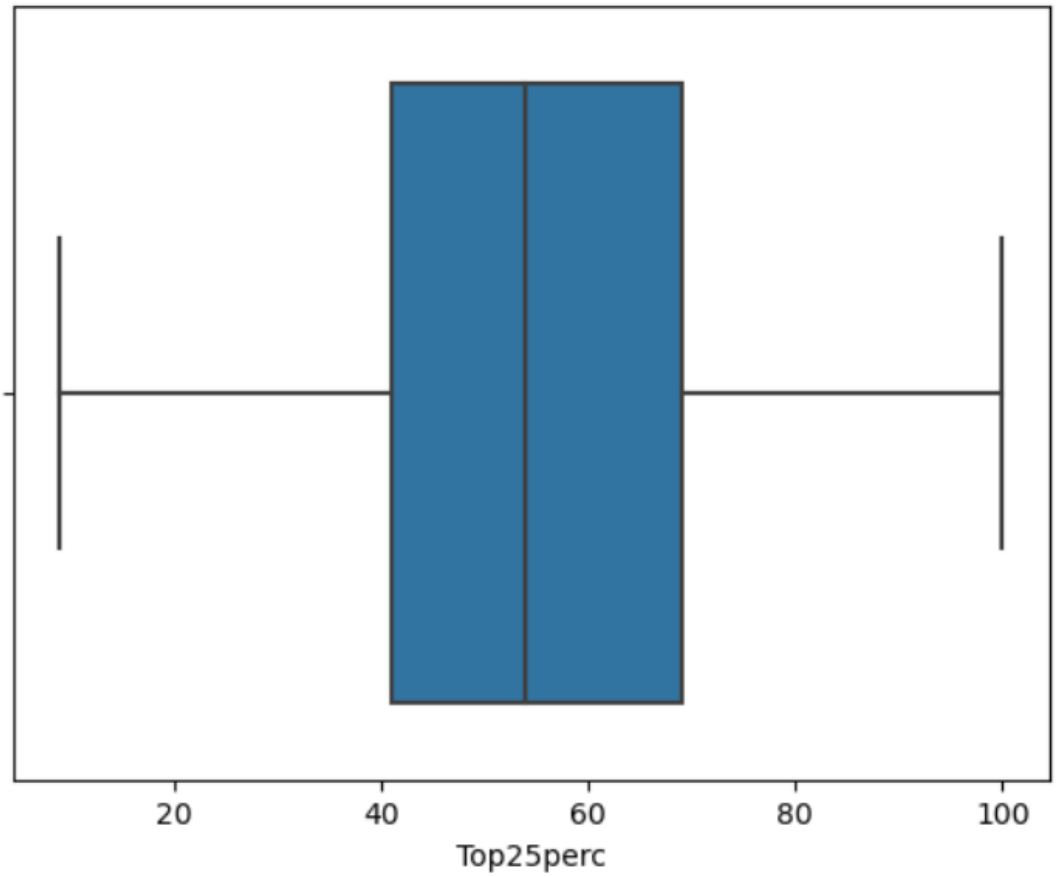
Observations:

In column Enroll, the outlier with the highest value is 6392.



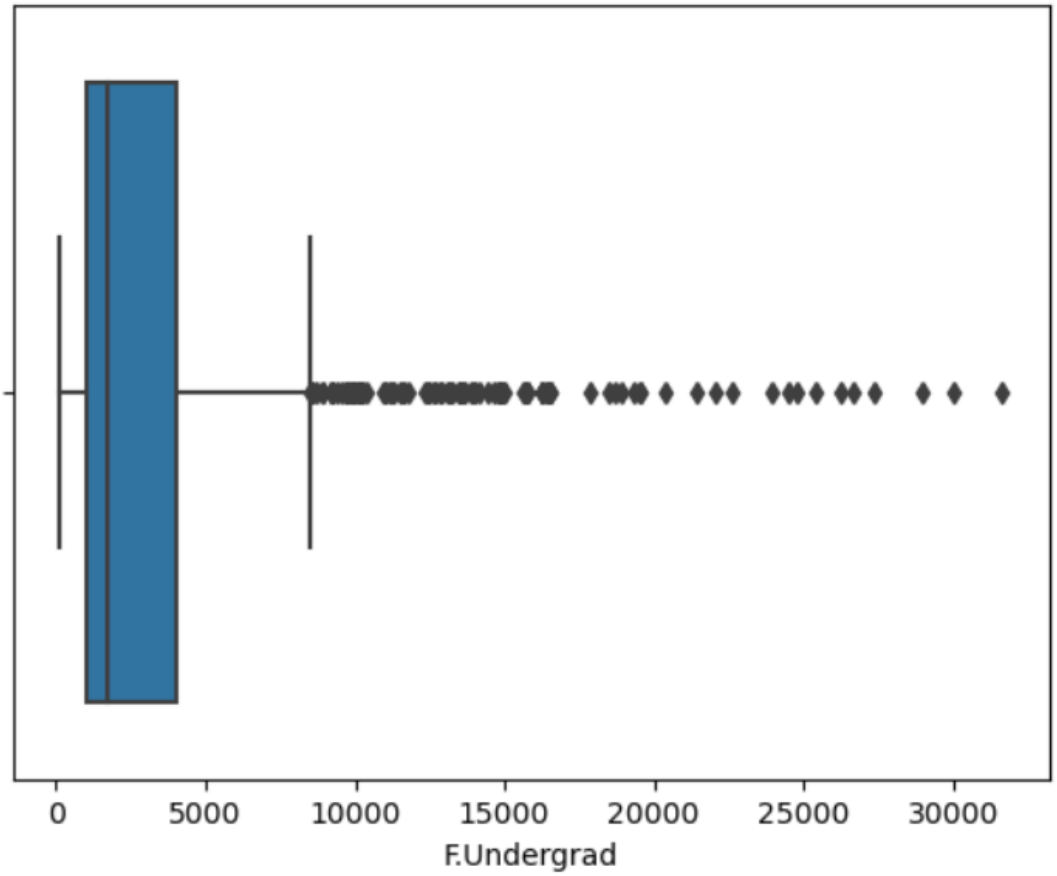
Observations:

In column Top10perc, the outlier with the highest value is 96.



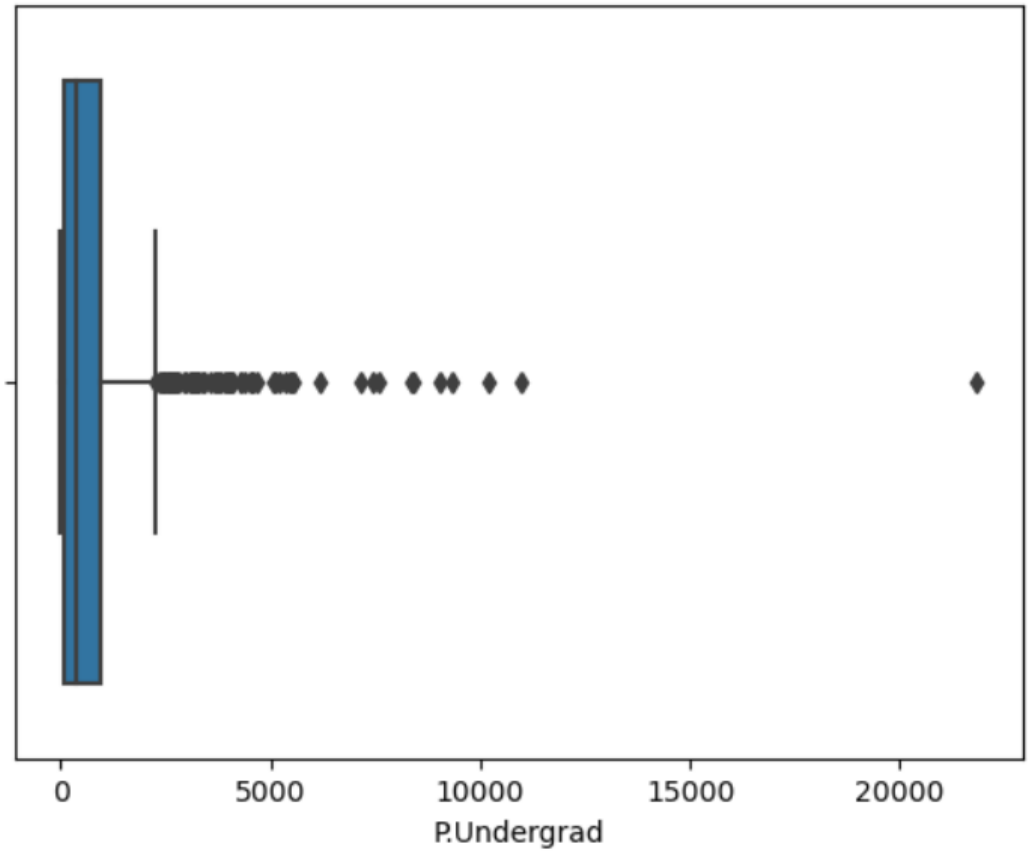
Observations:

In column Top25perc, there are no outliers present.



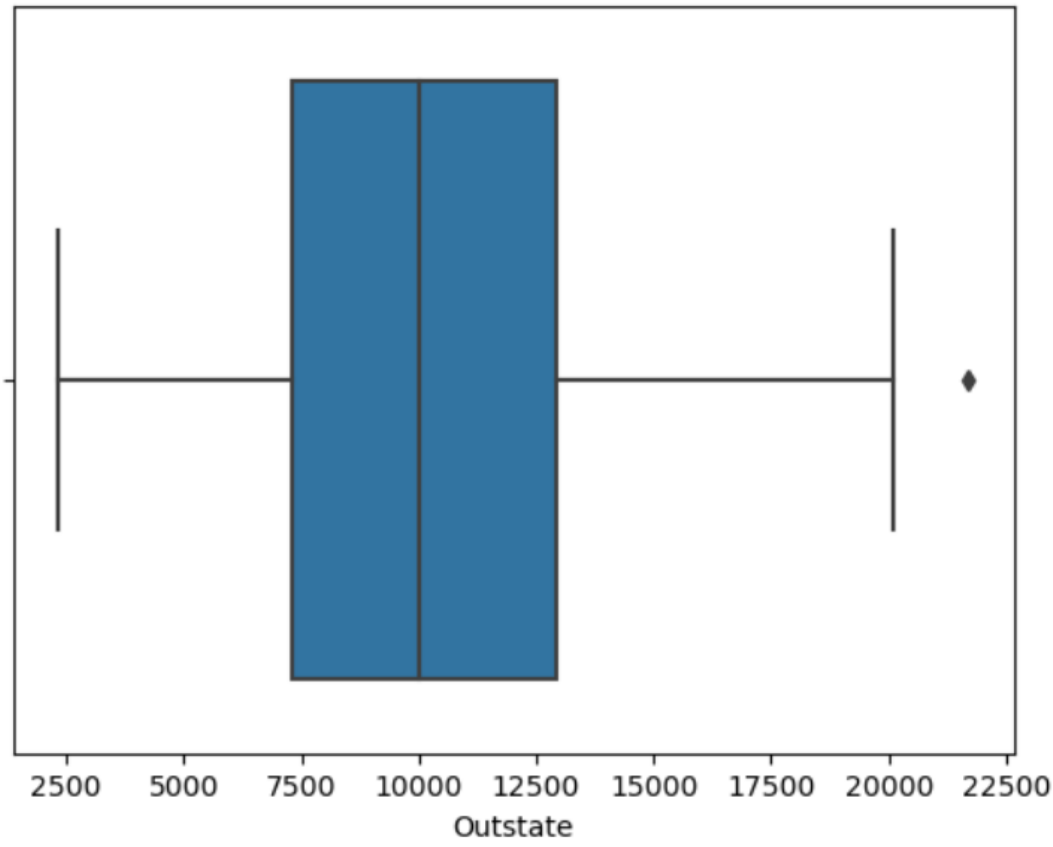
Observations:

In column F.Undergrad, the outlier with the highest value is 31643.



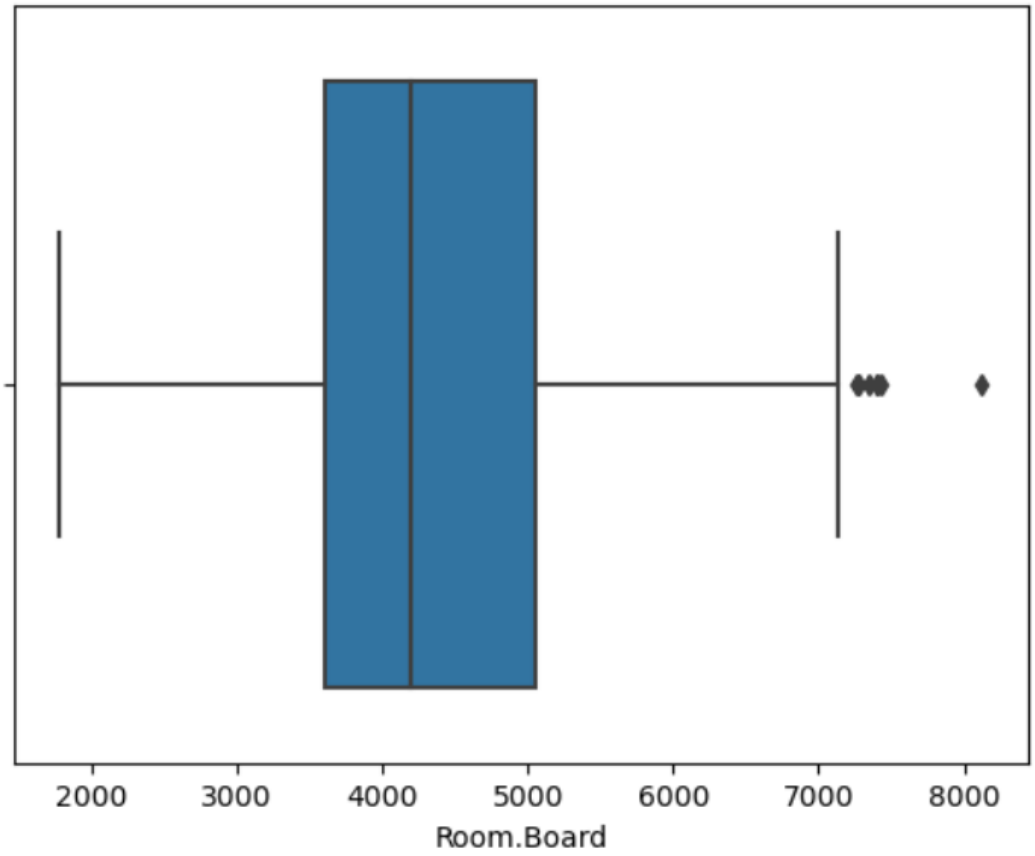
Observations:

In column P.Undergrad, the outlier with the highest value is 21836.



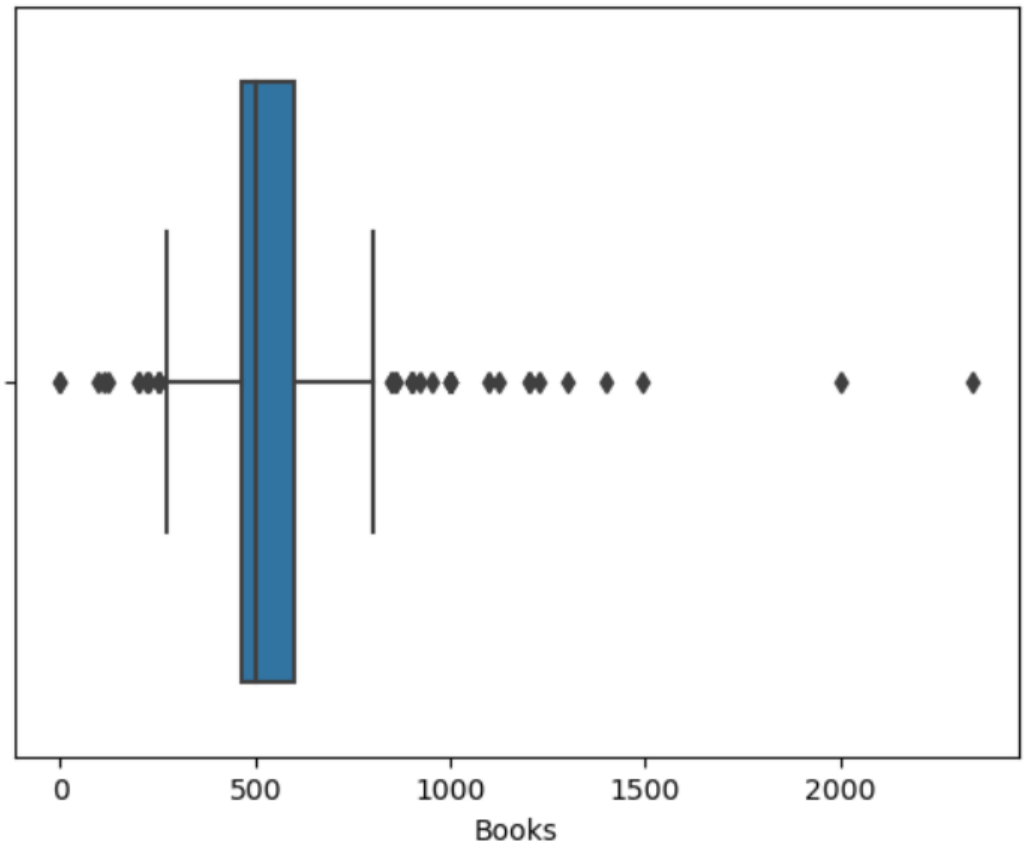
Observations:

In column Outstate, there is only one outlier which has value 21700.



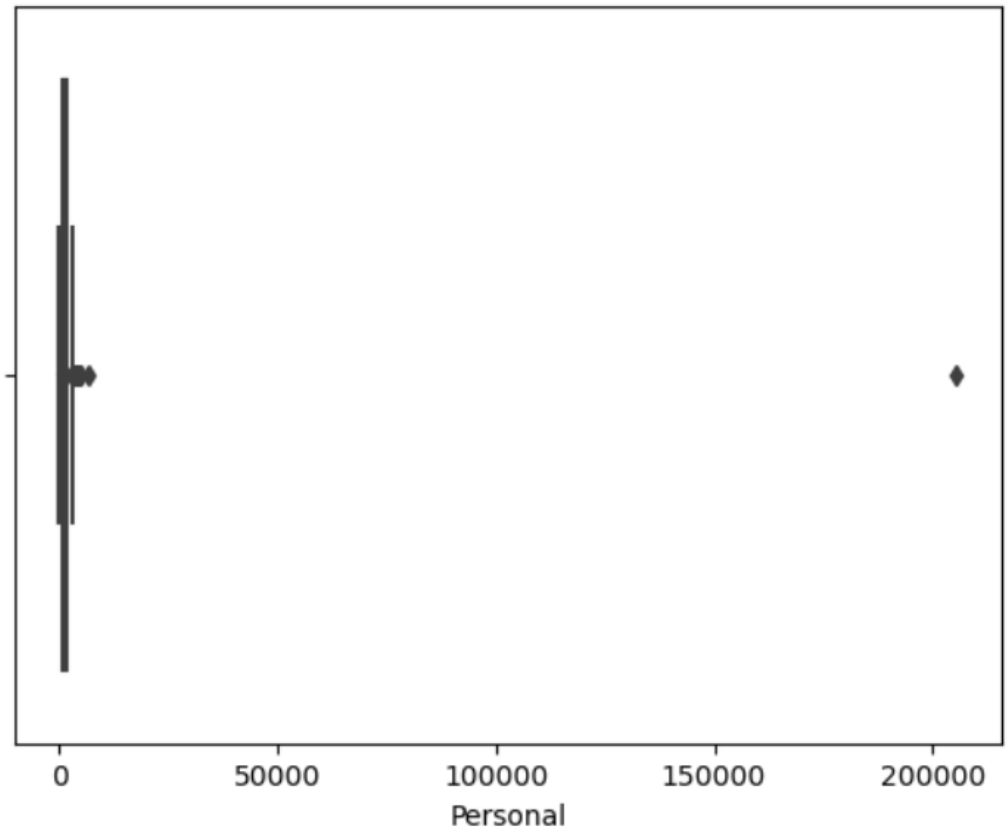
Observations:

In column Room.Board, the outlier with the highest value is 8124.



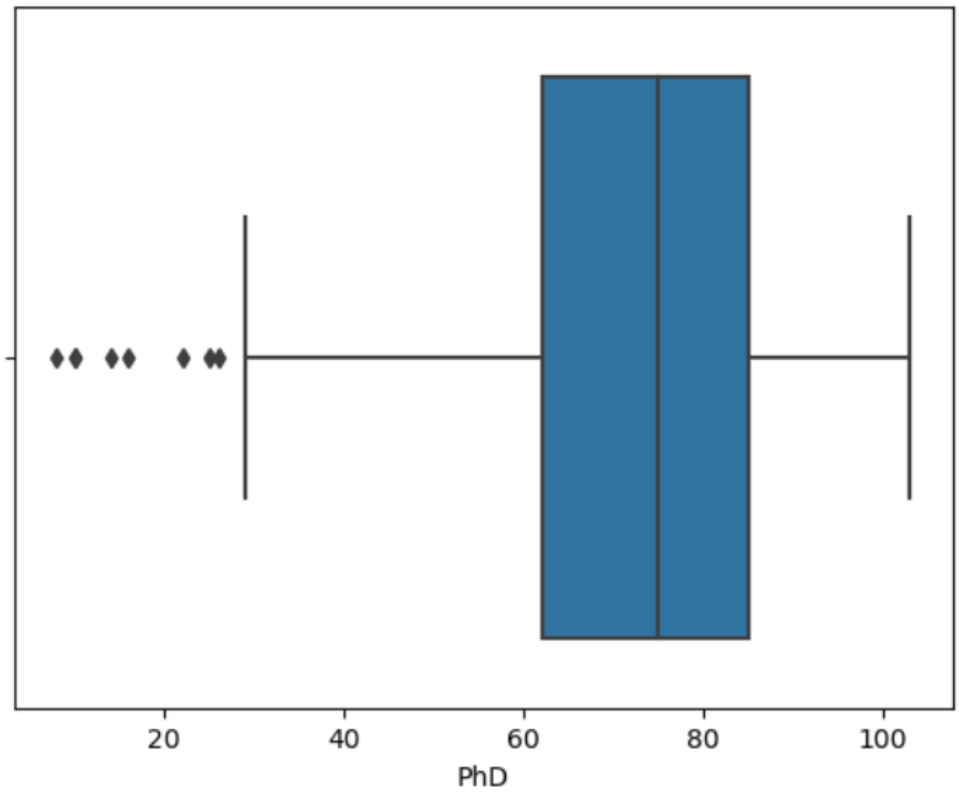
Observations:

In column Books, the outlier with the highest value is 2340.



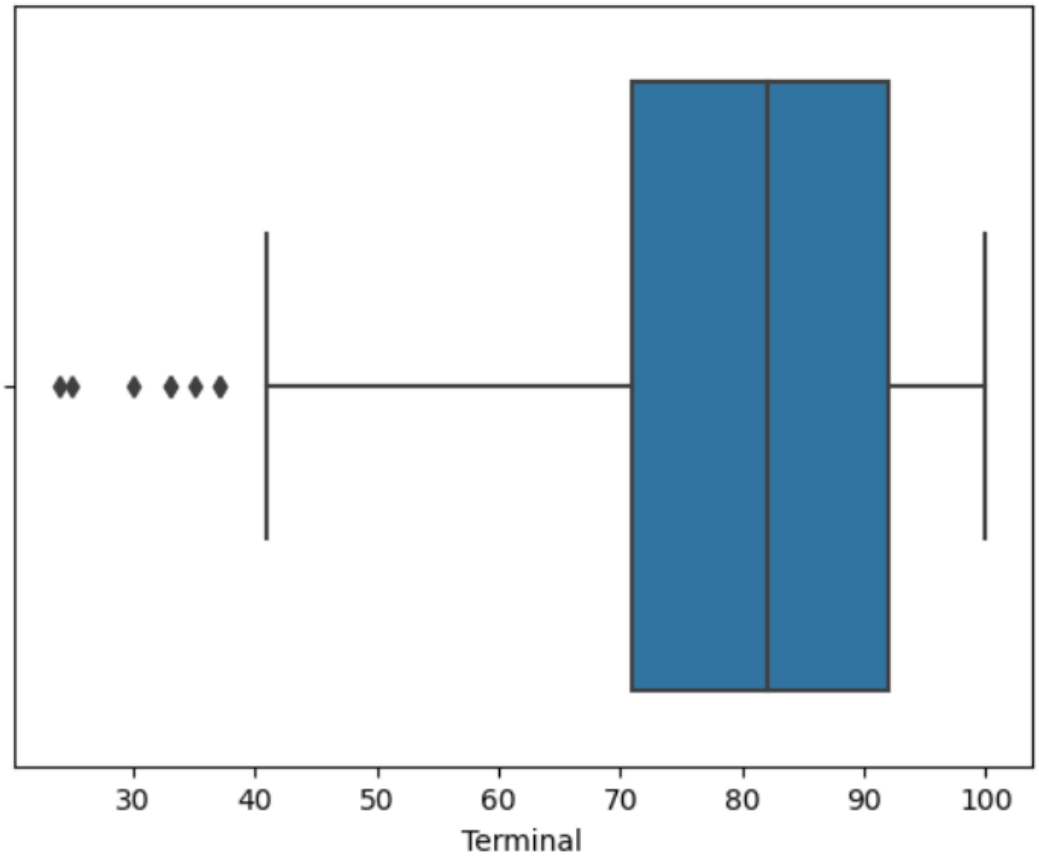
Observations:

In column Personal, the outlier with the highest value is 205500.



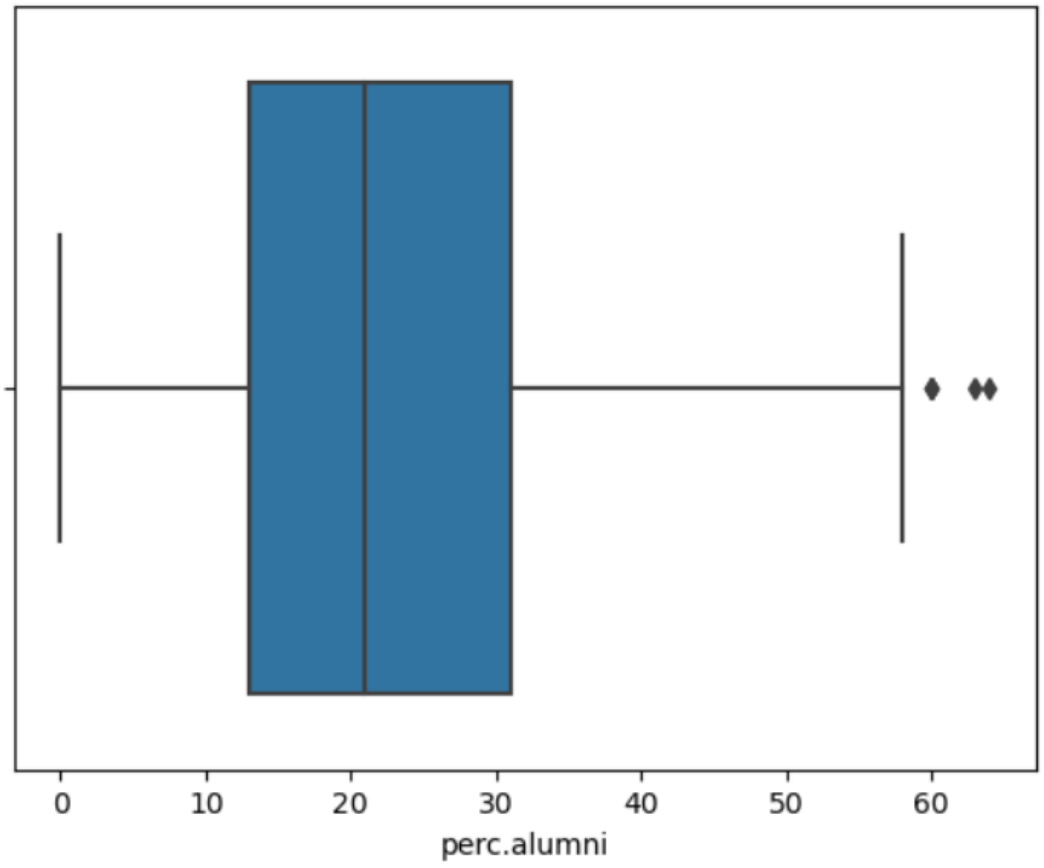
Observations:

In column PhD, the outlier with the lowest value is 8.



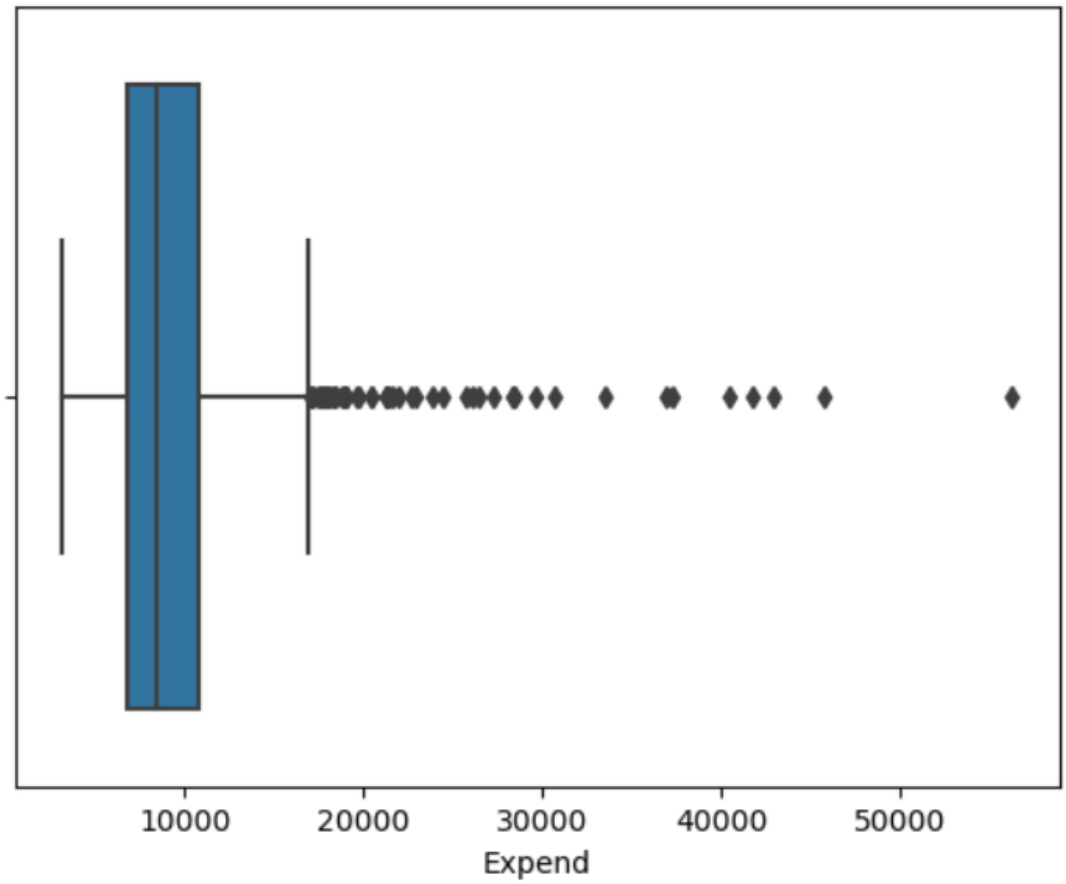
Observations:

In column Terminal, the outlier with the lowest value is 24.



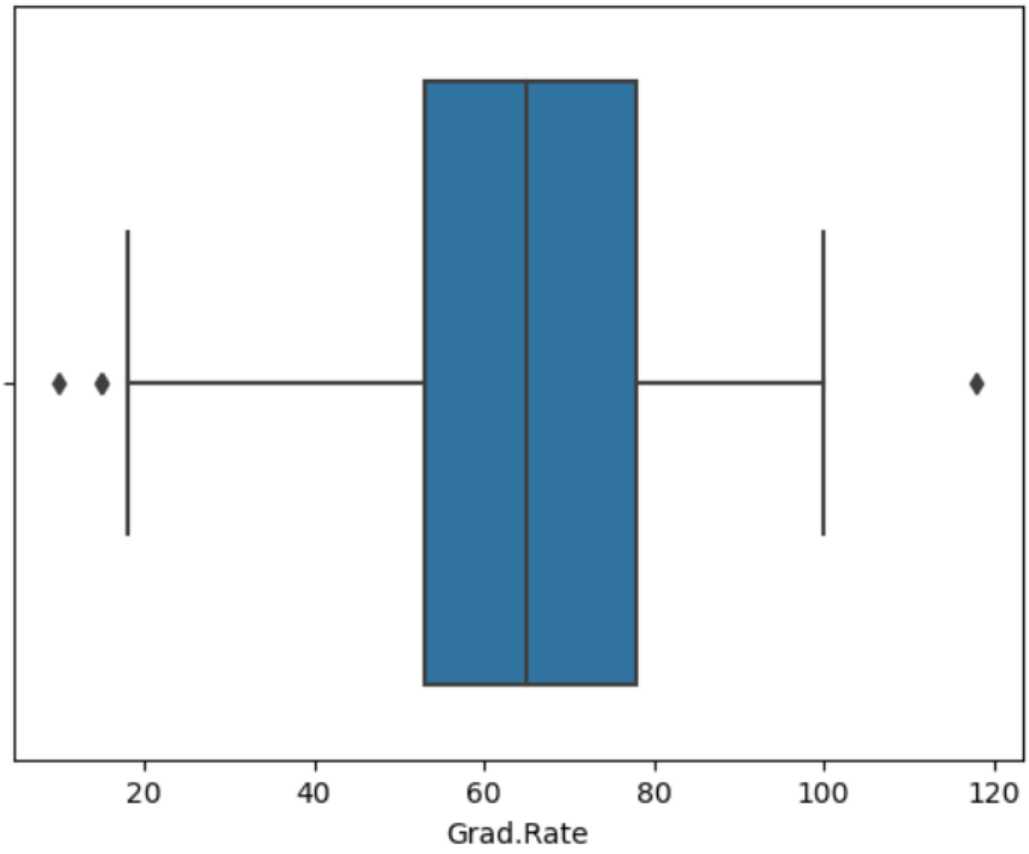
Observations:

In column perc.alumni, the outlier with the highest value is 64.



Observations:

In column Expend, the outlier with the highest value is 56233.



Observations:

In column Grad.Rate, the outlier with the highest value is 118.

Anomalies Or Wrong Entries

	1	81	241
Names	Adelphi University	Campbell University	Gwynedd Mercy College
Apps	2186.0	2087.0	380.0
Accept	1924	1339	237
Enroll	512.0	NaN	104.0
Top10perc	16.0	20.0	30.0
Top25perc	29	54	56
F.Undergrad	2683	3191	716
P.Undergrad	1227	1204	1108
Outstate	12280	7550	11000
Room.Board	6450	2790	5550
Books	750	600	500
Personal	1500.0	500.0	500.0
PhD	29	77	36
Terminal	30	77	41
S.F.Ratio	?	?	?
perc.alumni	16	34	22
Expend	10527	3739	7483
Grad.Rate	56	63	96

	0	1	2	3	4
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	434.0	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	13.6	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

Observations:

Hence the wronge entry - "?" has been replaced with null value

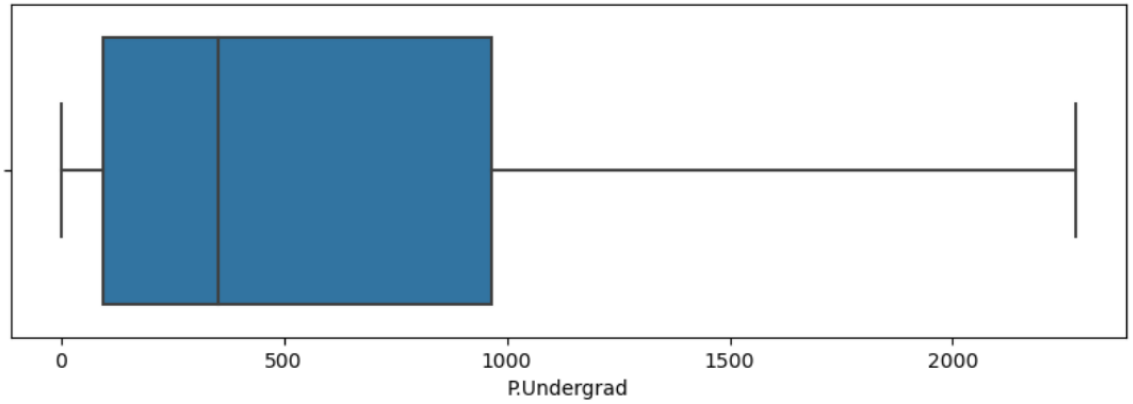
Removing Null Values

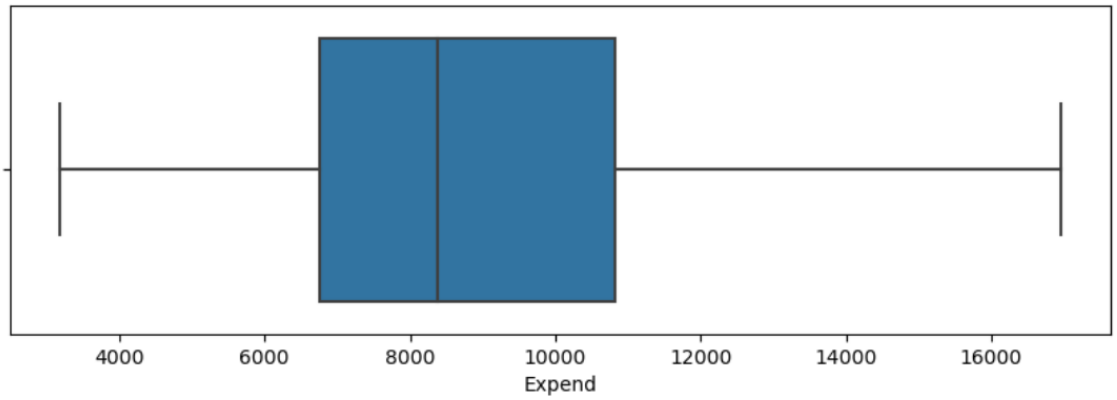
	0	1	2	3	4
Names	Abilene Christian University	Adelphi University	Adrian College	Agnes Scott College	Alaska Pacific University
Apps	1660.0	2186.0	1428.0	417.0	193.0
Accept	1232	1924	1097	349	146
Enroll	721.0	512.0	336.0	434.0	55.0
Top10perc	23.0	16.0	22.0	60.0	16.0
Top25perc	52	29	50	89	44
F.Undergrad	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	13.6	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

Observations:

There are no null values present in the dataset.

Removing Outliers





Observations:

There are no outliers present in the given dataset as can be seen through the examples P.Undergrad and Expend.

1. Application and Enrollment Analysis

• What is the average number of applications received by colleges?

```
The average number of applications received by colleges is 3007.59
```

• What percentage of applications are accepted on average across all colleges?

```
The average acceptance rate across all colleges is 74.66%.
```

• What is the average enrollment rate (number of students enrolled divided by number of applications accepted)?

```
The average enrollment rate across all colleges is 41.19%.
```

• Which college has the highest number of applications received?

```
The college with the highest number of applications received is Rutgers at New
Brunswick with 48094.0 applications.
```

2.Academic Excellence

• What is the average percentage of new students from the top 10% of their higher secondary class across all colleges?

```
The average percentage of new students from the top 10% of their higher seconda
ry class is 27.62%.
```

• What is the average percentage of new students from the top 25% of their higher secondary class?

The average percentage of new students from the top 25% of their higher s
econdary class is 55.80%.

• Is there a correlation between the percentage of students from the top 10% and the top 25% of their higher secondary class?

The correlation between the percentage of students from the top 10% and t
he top 25% of their higher secondary class is 0.89.

3. Student Demographics

• What is the average number of full-time undergraduate students per college?

The average number of full-time undergraduate students per college is 369
9.91.

• What is the average number of part-time undergraduate students per college?

The average number of part-time undergraduate students per college is 85
5.30.

• Which college has the highest number of out-of-state students?

The college with the highest number of out-of-state students is Benningto
n College with 21700 out-of-state students.

4. Cost and Spending

• What is the average cost of room and board across all colleges?

The average cost of room and board across all colleges is $4357.53.

• What is the average estimated book cost for a student?

The average estimated book cost for a student is $547.88.

• What is the average estimated personal spending for a student?

The average estimated personal spending for a student is $1601.51.

• How does the instructional expenditure per student vary across colleges?

```
The statistical summary of instructional expenditure per student across colleges is:
count      777.000000
mean      9660.171171
std       5221.768440
min       3186.000000
25%       6751.000000
50%       8377.000000
75%      10830.000000
max      56233.000000
Name: Expend, dtype: float64
```

5. Faculty Qualifications

• What is the average percentage of faculties with Ph.D.s across all colleges?

The average percentage of faculty with Ph.D.s across all colleges is 72.66%.

• What is the average percentage of faculties with terminal degrees?

The average percentage of faculty with terminal degrees across all colleges is 79.70%.

• Is there a correlation between the percentage of faculties with Ph.D.s and the graduation rate?

The correlation between the percentage of faculty with Ph.D.s and the graduation rate is 0.31.

6. Student-Faculty Interaction

• What is the average student/faculty ratio across all colleges?

```
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      float64
perc.alumni      int64
Expend           int64
Grad.Rate        int64
AcceptRate     float64
EnrollRate     float64
dtype: object
The average student/faculty ratio across all colleges is 14.09.
```

• Which college has the lowest student/faculty ratio?

```
The college with the lowest student/faculty ratio is Universi
ty of Charleston with a ratio of 2.50.
```

• Is there a correlation between the student/faculty ratio and the graduation rate?

```
The correlation between the student/faculty ratio and the gra
duation rate is -0.31.
```

7. Alumni Engagement

• What is the average percentage of alumni who donate across all colleges?

```
The average percentage of alumni who donate across all colleg
es is 22.74%.
```

• Is there a correlation between the percentage of alumni who donate and the graduation rate?

```
The correlation between the percentage of alumni who donate a
nd the graduation rate is 0.49.
```

8. Graduation Rates

• What is the average graduation rate across all colleges?

```
The average graduation rate across all colleges is 65.46%.
```

• Which college has the highest graduation rate?

```
The college with the highest graduation rate is Cazenovia Col
lege with a graduation rate of 118%.
```

• Is there a correlation between the instructional expenditure per student and the graduation rate?

```
The correlation between the instructional expenditure per stu
dent and the graduation rate is 0.39.
```

9. 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?

```
Correlation between Applications and graduation rate: 0.15
Correlation between Acceptance Rate and graduation rate: -0.29
Correlation between Enrollment Rate and graduation rate: -0.29
Correlation between Top 10% and graduation rate: 0.49
Correlation between Top 25% and graduation rate: 0.48
Correlation between Room and Board and graduation rate: 0.42
Correlation between Books and graduation rate: 0.01
Correlation between Personal Spending and graduation rate: 0.02
Correlation between PhD Faculty and graduation rate: 0.31
Correlation between Terminal Faculty and graduation rate: 0.29
Correlation between Student/Faculty Ratio and graduation rate: -0.31
Correlation between Alumni Donations and graduation rate: 0.49
Correlation between Instructional Expenditure and graduation rate: 0.39

The factor most strongly associated with higher graduation rates is Top 10% wit
h a correlation coefficient of 0.49.
```

• What recommendations can be made to colleges to improve their graduation rates based on the data analysis?

```
Recommendations:

 Support Financially Needy Students:
   - Offer financial aid packages that cover not just tuition but also living e
xpenses.
   - Provide affordable housing options and textbook rental programs to reduce
the financial burden on students.
```