STATISTICAL METHODS FOR DECISON MAKING

REPORT

PREPARED BY RAJEESH J

APRIL 2023

PROBLEM 1 ANALYSIS OF WHOLESALE CUSTOMERS

PREPARED BY:

RAJEESH J

PROBLEM STATEMENT:

A wholesale distributor operating in different regions of Portugal has information on the annual spending of several items in their stores across different regions and channels. The data consists of 440 large retailers' annual spending on 6 different varieties of products in 3 different regions (Lisbon, Oporto, Other) and across different sales channels (Hotel, Retail).

1.1.1 Use methods of descriptive statistics to summarize data

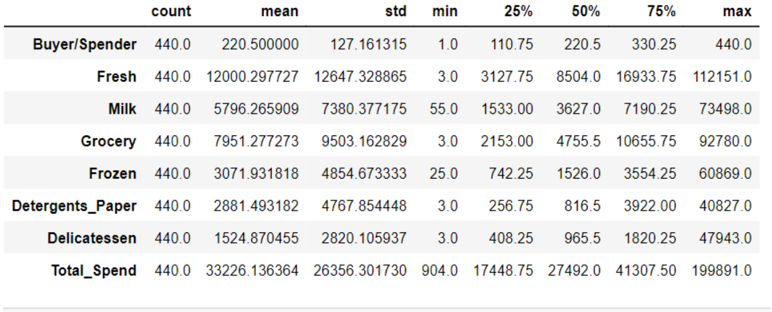
METHODS:

1) Using Describe, Shape, info from Numpy and Pandas library of Python
2) Using Univariate, Bivariate analysis, Multivariate analysis (Exploratory Data Analysis)

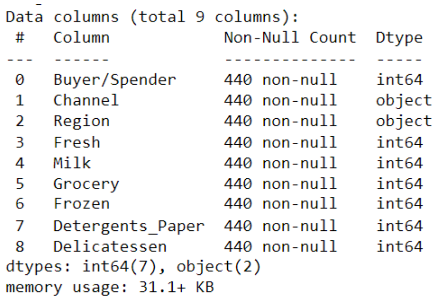
EXECUTIVE COMMENTARY:

1) There are no Null values, Missing Values. There are 7 Numerical variables and 2 categorical variables in 440 rows
2) The Total Spending for a buyer was calculated by summation of spending across, all 6 categories

3) Range of the spending varies from 904 to 199891 with mean of 33226.



	count	mean	std	min	25%	50%	75%	max
Buyer/Spender	440.0	220.500000	127.161315	1.0	110.75	220.5	330.25	440.0
Fresh	440.0	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0
Milk	440.0	5796.265909	7380.377175	55.0	1533.00	3627.0	7190.25	73498.0
Grocery	440.0	7951.277273	9503.162829	3.0	2153.00	4755.5	10655.75	92780.0
Frozen	440.0	3071.931818	4854.673333	25.0	742.25	1526.0	3554.25	60869.0
Detergents_Paper	440.0	2881.493182	4767.854448	3.0	256.75	816.5	3922.00	40827.0
Delicatessen	440.0	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	47943.0
Total_Spend	440.0	33226.136364	26356.301730	904.0	17448.75	27492.0	41307.50	199891.0



```
Data columns (total 9 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   Buyer/Spender     440 non-null    int64 
 1   Channel           440 non-null    object
 2   Region            440 non-null    object
 3   Fresh             440 non-null    int64 
 4   Milk              440 non-null    int64 
 5   Grocery           440 non-null    int64 
 6   Frozen            440 non-null    int64 
 7   Detergents_Paper  440 non-null    int64 
 8   Delicatessen      440 non-null    int64 
dtypes: int64(7), object(2)
memory usage: 31.1+ KB
```

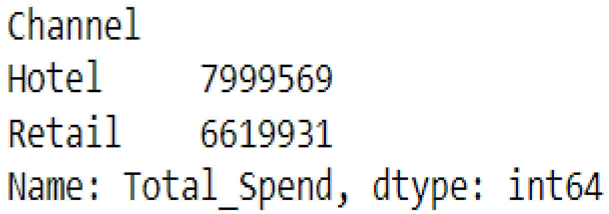
1.1.2 Which Region and which Channel spent the most?

METHODS:

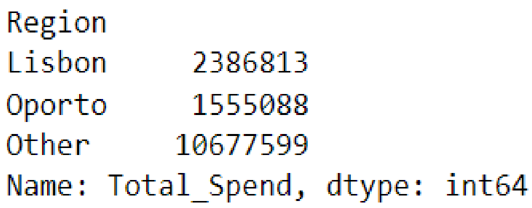
To identify max. spending, we have already calculated total spending for individual IDs. Now we are grouping it based on Region and Channel & Finding the total spends.

EXECUTIVE COMMENTARY:

1) "Other" is the region that has spent the most with 10677599 followed by Lisbon with a Total spend of 2386813.
2) Hotel is the channel have spent the most 7999569



```
Channel
Hotel      7999569
Retail     6619931
Name: Total_Spend, dtype: int64
```



```
Region
Lisbon      2386813
Oporto      1555088
Other      10677599
Name: Total_Spend, dtype: int64
```

1.1.3 Which Region and which Channel spent the least?

METHODS:

To identify max. spending, we have already calculated total spending for individual IDs by summing up spends on individual product categoies. Now we are grouping it based on Region and Channel & Finding the total spends.

EXECUTIVE COMMENTARY:

1) "Oporto" is the region that has spent the least with 1555088
2) Retail is the channel have spent the least 6619931

1.2 There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer.

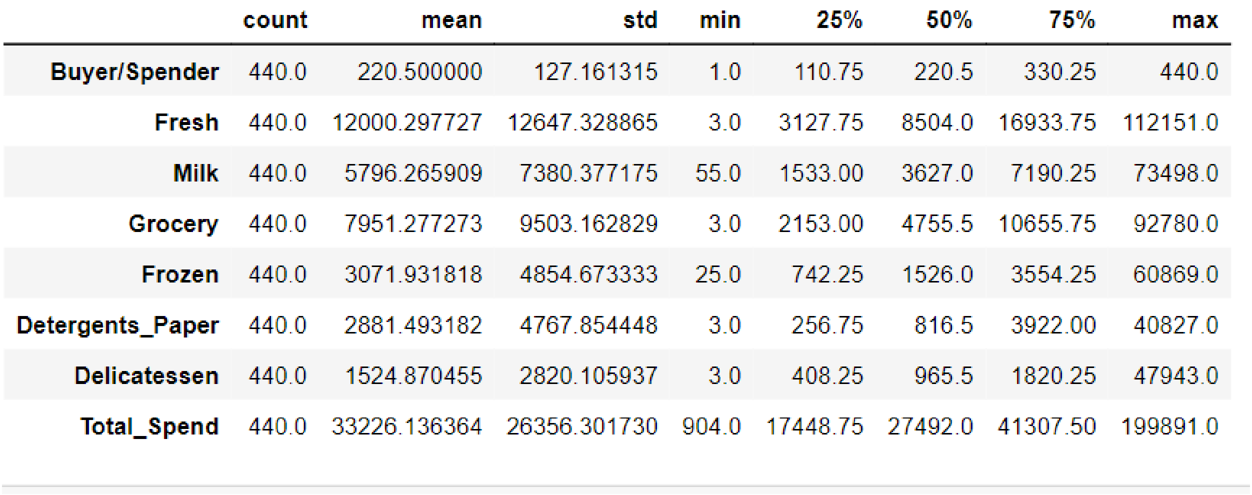
METHODS:

To understand we drew all the descriptive statistical data for all the variables from that we understood underlying descriptive patterns.

EXECUTIVE COMMENTARY:

1) From the descriptive summary we understood two categorical variables are Region and channel. Which are Lisbon, Oporto, Other & Hotel, Retail respectively.
2) Max transactions (2/3rd of transactions) happened in Hotels channel and (3/4th of all transactions) in Other Region
3) Fresh – It has a spends varying from 3 to 112151, with an average of 12000 per transaction dispersed with a range of 112148 and an IQR of 13806
4) Milk - It has a spends varying from 55 to 73498, with an average of 5797 per transaction dispersed with a range of 73443 and an IQR of 5657

5) Grocery - It has a spends varying from 3 to 92780, with an average of 7952 per transaction dispersed with a range of 92777 and an IQR of 8502
6) Frozen - It has a spends varying from 25 to 60869, with an average of 3071 per transaction dispersed with a range of 60844 and an IQR of 2812
7) Detergents Paper - It has a spends varying from 3 to 40827, with an average of 2881 per transaction dispersed with a range of 40824 and an IQR of 3665
8) Delicatessen - It has a spends varying from 3 to 47943, with an average of 1524 per transaction dispersed with a range of 60844 and an IQR of 1412
   Ref [4]



	count	mean	std	min	25%	50%	75%	max
Buyer/Spender	440.0	220.500000	127.161315	1.0	110.75	220.5	330.25	440.0
Fresh	440.0	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0
Milk	440.0	5796.265909	7380.377175	55.0	1533.00	3627.0	7190.25	73498.0
Grocery	440.0	7951.277273	9503.162829	3.0	2153.00	4755.5	10655.75	92780.0
Frozen	440.0	3071.931818	4854.673333	25.0	742.25	1526.0	3554.25	60869.0
Detergents_Paper	440.0	2881.493182	4767.854448	3.0	256.75	816.5	3922.00	40827.0
Delicatessen	440.0	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	47943.0
Total_Spend	440.0	33226.136364	26356.301730	904.0	17448.75	27492.0	41307.50	199891.0

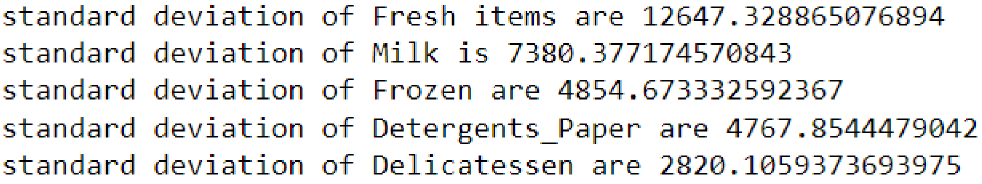
1.3 On the basis of a descriptive measure of variability, which item shows the most inconsistent behavior? Which items show the least inconsistent behavior?

Method:

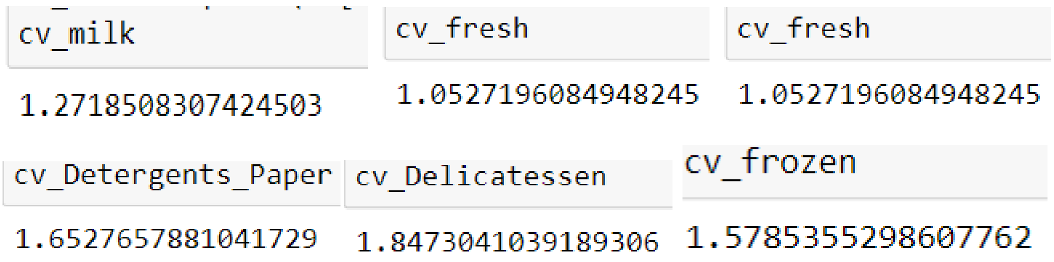
To understand the measure of variability, we studied Standard deviation and Covariance

Executive Commentary:

1) By studying Standard Deviation which is the measure of variability of any datapoint of a variable from its mean, Fresh has highest standard deviation of 12747 and Delicatessen has a standard deviation of 2820. If we take Standard deviation as a measure, Fresh is most inconsistent and delicatessen is least inconsistent.
2) By Studying Covariance which is also an indicator for measuring consistency of data, fresh has Covariance of 1.05 and Delicatessen has covariance of 1.85. If we take Covariance as a measure, Fresh is least inconsistent and Delicatessen has a most inconsistent.



```
standard deviation of Fresh items are 12647.328865076894
standard deviation of Milk is 7380.377174570843
standard deviation of Frozen are 4854.673332592367
standard deviation of Detergents_Paper are 4767.8544479042
standard deviation of Delicatessen are 2820.1059373693975
```



```
cv_milk
1.2718508307424503

cv_fresh
1.0527196084948245

cv_fresh
1.0527196084948245

cv_Detergents_Paper
1.6527657881041729

cv_Delicatessen
1.8473041039189306

cv_frozen
1.5785355298607762
```

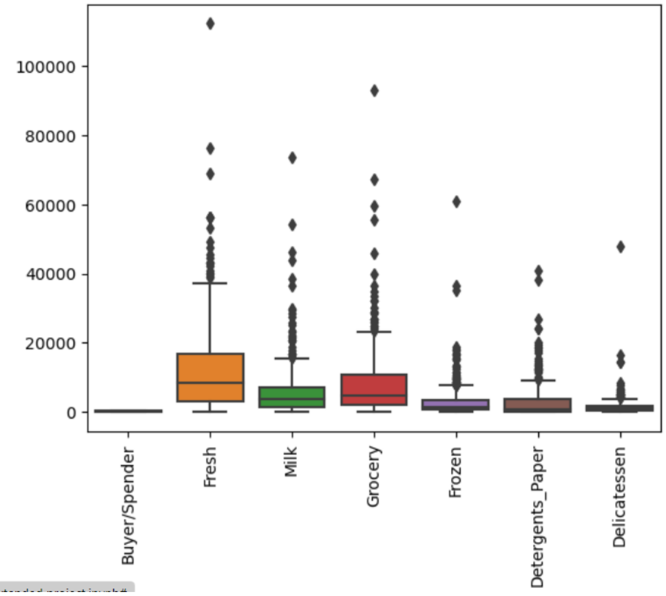
1.4 **Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.**

METHODS:

We used Box plots to understand and identify the outliers in each variable.

EXECUTIVE COMMENTARY:

Yes. There are outliers in the data as visualized with box plot across all the products



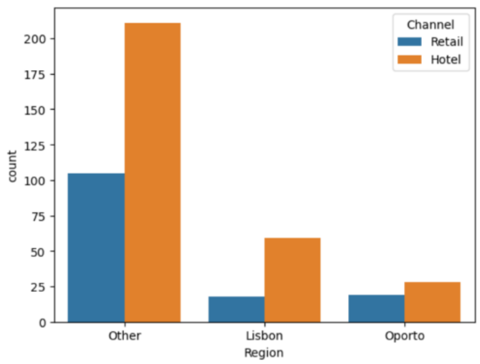
1.5 On the basis of your analysis, what are your recommendations for the business? How can your analysis help the business to solve its problem? Answer from the business perspective.

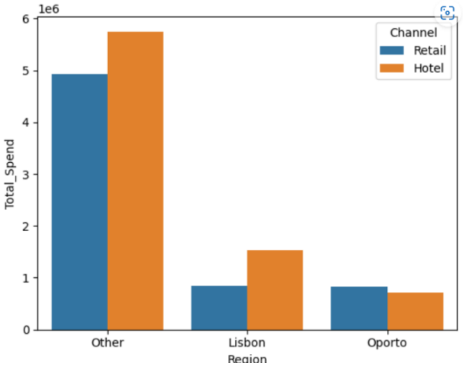
EXECUTIVE COMMENTARY:

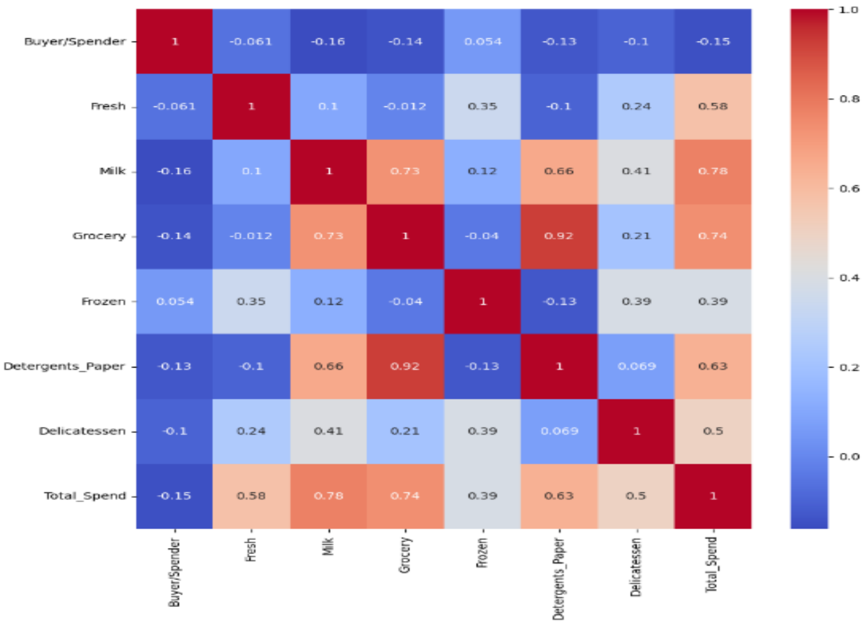
1) Customers spend most on grocery, Milk, Fresh and comparing to other category products. Supply to be made seamless for these product categories
2) Delicatessen category is the least spent category. The low sales may be due to price or due to low demand. If it is due low demand, company can drop the product, if it is incase of price company can consider repricing.
3) Spending pattern across region seems more or less similar however, if we see the pattern in channels,

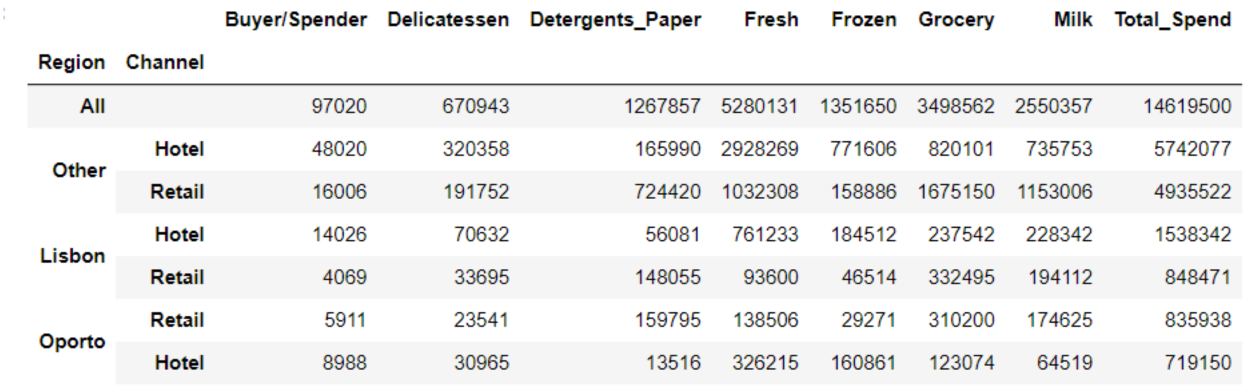
   Frozen and Fresh products are sold most in Hotel in comparison to retail

   Milk and Grocery category is sold most in Retail
4) There is strong correlation between Grocery and Mlk, Detergent paper and Grocery, Business can use this data to use it in merchandising and promotion schemes like Buy one Get one type of offers

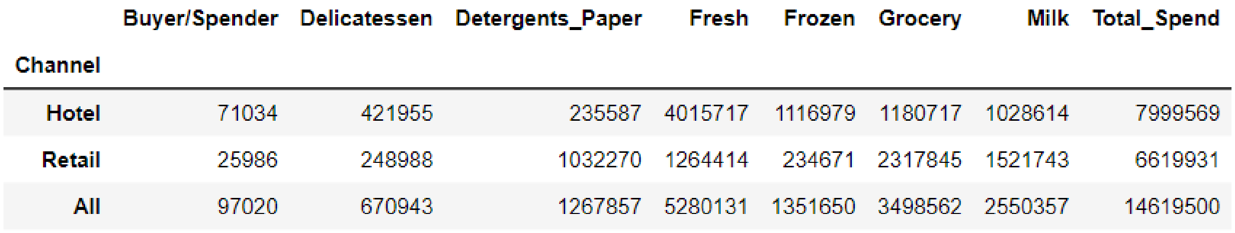






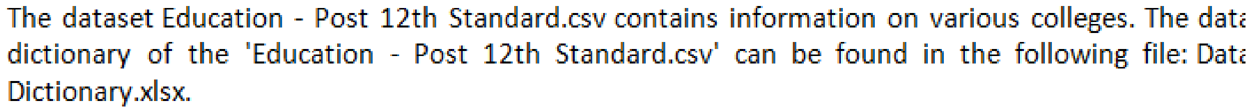


		Buyer/Spender	Delicatessen	Detergents_Paper	Fresh	Frozen	Grocery	Milk	Total_Spend
Region	**Channel**								
All		97020	670943	1267857	5280131	1351650	3498562	2550357	14619500
Other	**Hotel**	48020	320358	165990	2928269	771606	820101	735753	5742077
	Retail	16006	191752	724420	1032308	158886	1675150	1153006	4935522
Lisbon	**Hotel**	14026	70632	56081	761233	184512	237542	228342	1538342
	Retail	4069	33695	148055	93600	46514	332495	194112	848471
Oporto	**Retail**	5911	23541	159795	138506	29271	310200	174625	835938
	Hotel	8988	30965	13516	326215	160861	123074	64519	719150



	Buyer/Spender	Delicatessen	Detergents_Paper	Fresh	Frozen	Grocery	Milk	Total_Spend
Channel								
Hotel	71034	421955	235587	4015717	1116979	1180717	1028614	7999569
Retail	25986	248988	1032270	1264414	234671	2317845	1521743	6619931
All	97020	670943	1267857	5280131	1351650	3498562	2550357	14619500

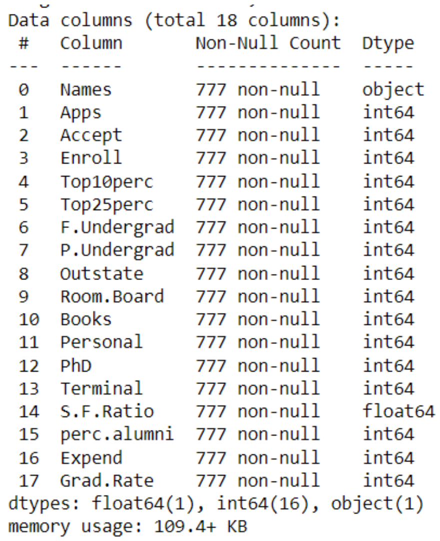
PROBLEM 2


The dataset Education - Post 12th Standard.csv contains information on various colleges. The datа dictionary of the 'Education - Post 12th Standard.csv' can be found in the following file: Data Dictionary.xlsx.

DATASET INSIGHTS

Univariate Analysis

We understand the data from post 12th standard data. It has 777 rows and 18 columns [1]. It has more integer and float variables and only one categorical variable.


```
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       777 non-null    int64  
 4   Top10perc    777 non-null    int64  
 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     777 non-null    int64  
 12  PhD          777 non-null    int64  
 13  Terminal     777 non-null    int64  
 14  S.F.Ratio    777 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(1), int64(16), object(1)
memory usage: 109.4+ KB
```

Data Preprocessing

As we are checking we did not find any missing values or duplicates or null value

Exploratory Analysis

Univariate Analysis:

It helps us understand the pattern of the distribution of the data and detecting outliers if any. We'll check one by one

Applications: Apps

The boxplot seems to have outliers and skewed in right. It is between range of 3000 to 5000. Total number of people applications seems to highest at 50000.

Accept

This variable also has outliers as plotted from boxplot and it is positively skewed. And it is between range between70 to 1500 and highest accepts are in the zone of 25000

Enroll

This variable also has outliers as plotted from box plot and distplot and it is positively skewed and majority of colleges have enrolled from 200 to 500

Top 10 Percentage

The boxplot of top10 percentage of higher secondary class also have outliers. It is also positively skewed. There are almost 10 to 50 students from top 10 percentage on higher secondary class

Top 25 percentage

The boxplot of top 25 percentage of higher secondary class. It does not seems to have any outliers and it is normally distributed

Full time undergraduate

The Full time undergraduate students seems to have almost 3000 to 5000 students, it is positively skewed as visualized by using box plot and also have outliers

Part Time undergraduate

There are 1000 to 3000 part time undergraduate students from 1000 to 3000, it is also positively skewed with outliers

Outstate

This distribution is almost normal distribution and has only one outlier, most of the outstate students are in the range from the range of 7500 to 12000 students

Room Board

This distribution is also normally distributed with a few outliers

Books

This follows a bi modal distribution with a range from 500 to 600 and there are lot of outliers

PHD

The distribution is negatively skewed with a range of 60 to 80 and it has outliers

Terminal

The distribution is negatively skewed with a range of 70 to 90 and it has outliers

Student faculty ratio

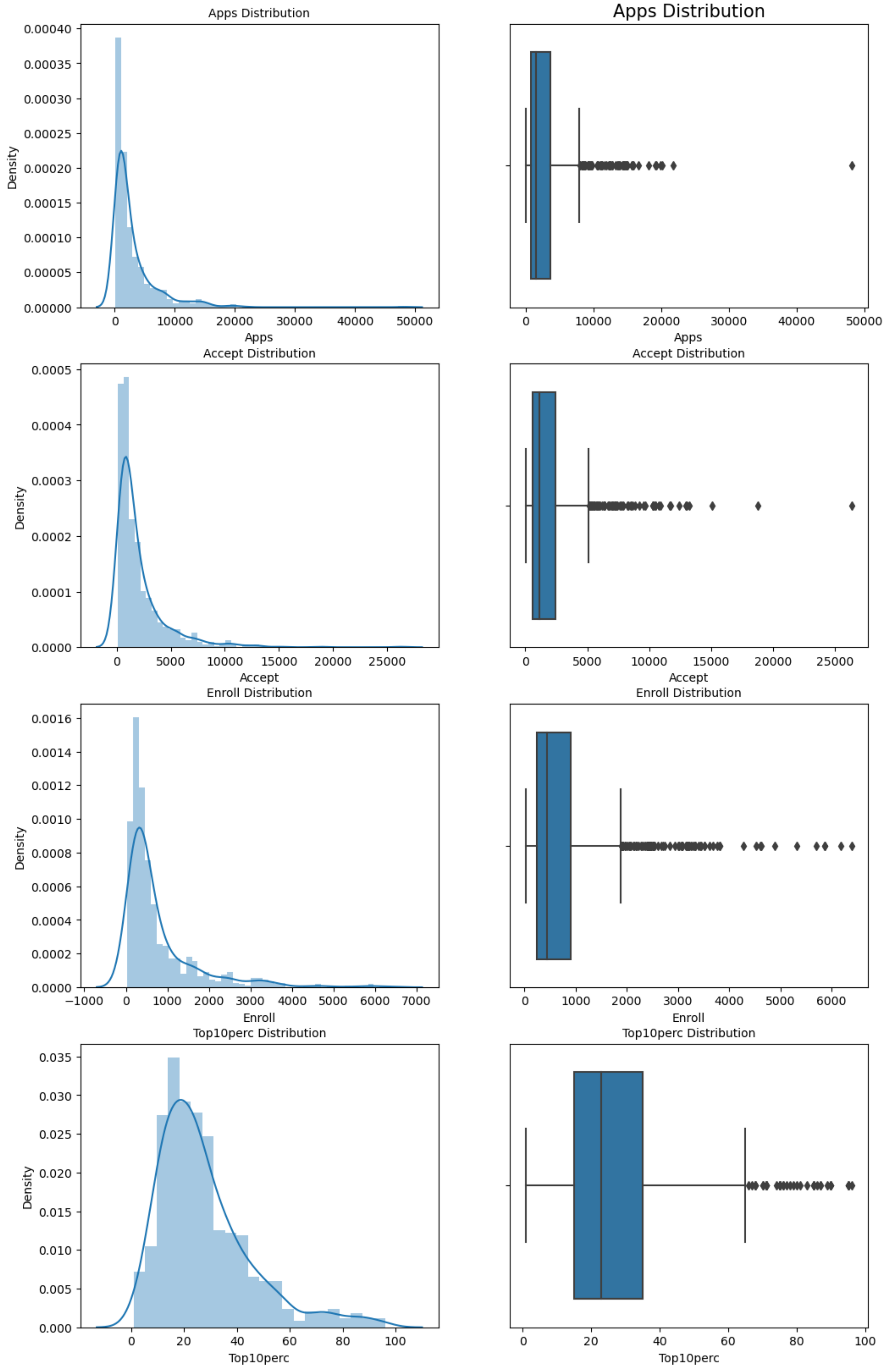
This distribution is also almost normally distributed with a range of 10 to 17

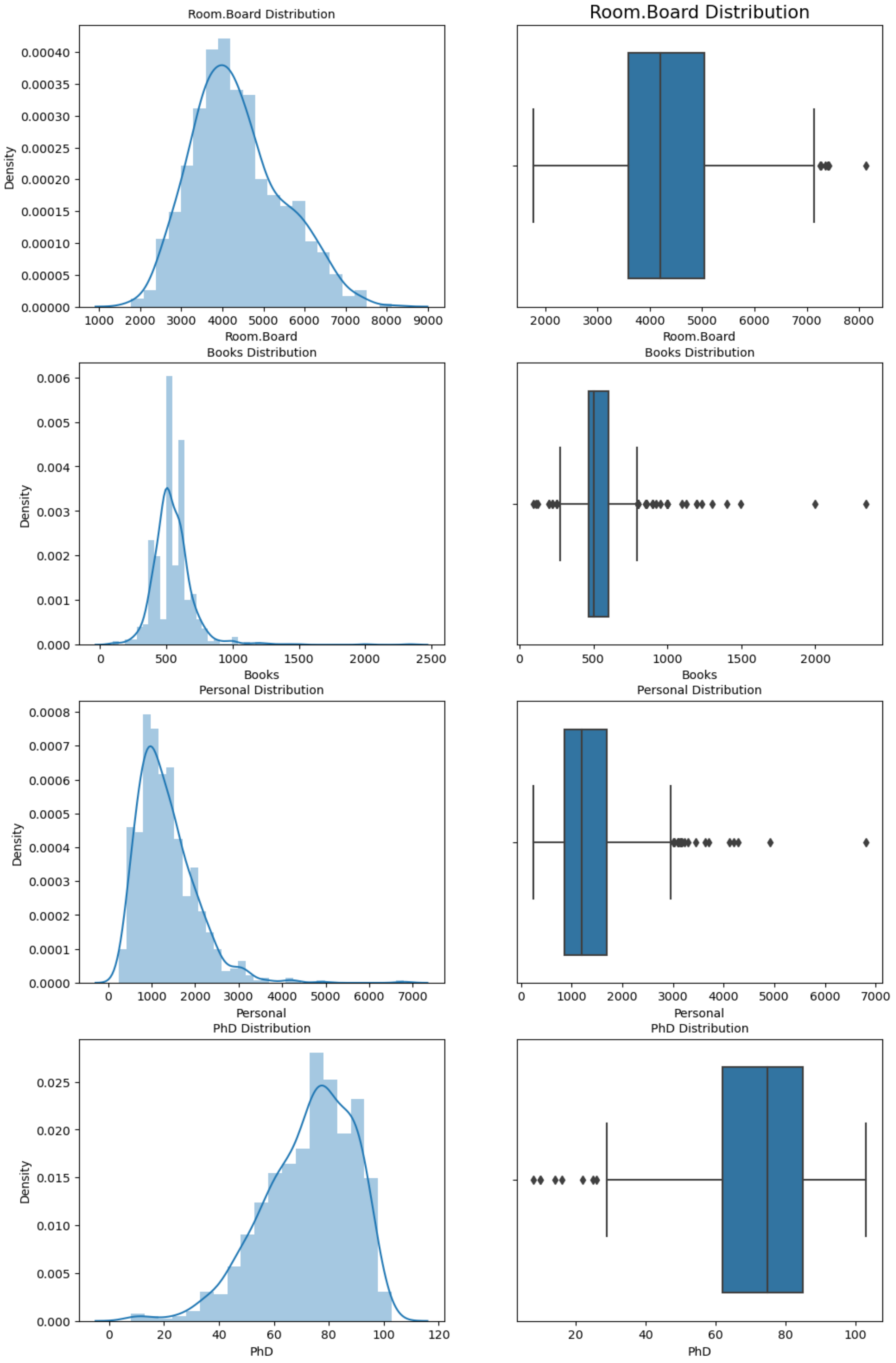
Percent Alumni

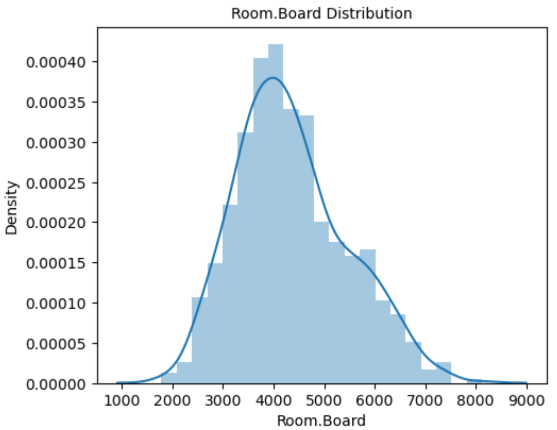
The distribution of % alumni is also normally distributed and has outliers

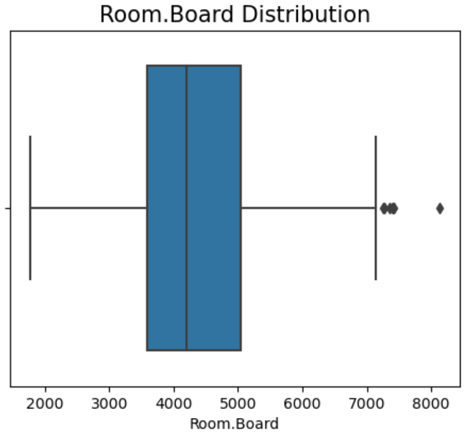
Expenditure

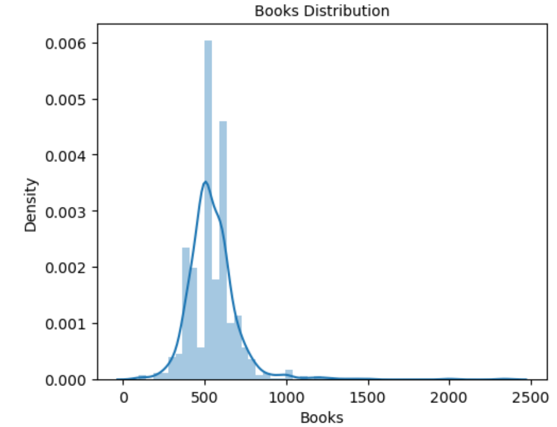
The expenditure is skewed positively and it also carries lot of outliers

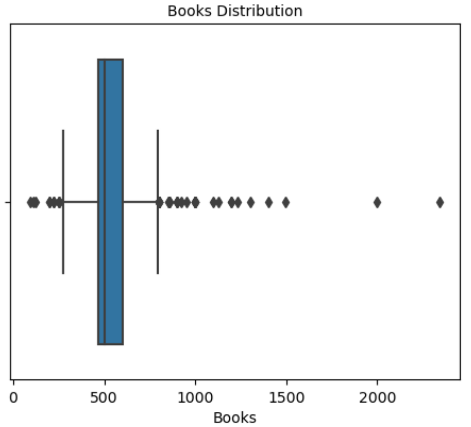

Apps Distribution
Density
0.00040
0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
0
10000
20000
30000
40000
50000
Apps
Apps Distribution
0
10000
20000
30000
40000
50000
Apps
Accept Distribution
Density
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
0
5000
10000
15000
20000
25000
Accept
Accept Distribution
0
5000
10000
15000
20000
25000
Accept
Enroll Distribution
Density
0.0016
0.0014
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
−1000
0
1000
2000
3000
4000
5000
6000
7000
Enroll
Enroll Distribution
0
1000
2000
3000
4000
5000
6000
Enroll
Top10perc Distribution
Density
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100
Top10perc
Top10perc Distribution
0
20
40
60
80
100
Top10perc

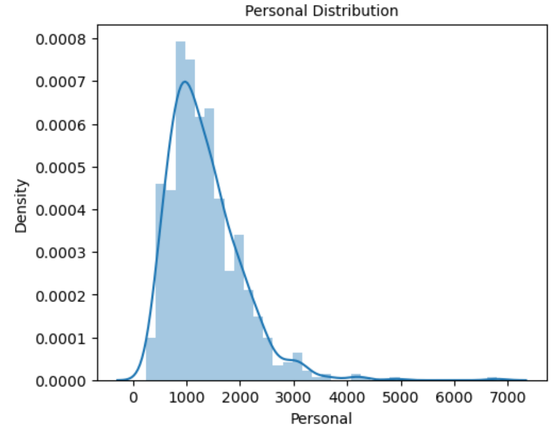

Room.Board Distribution
Density
0.00000
0.00005
0.00010
0.00015
0.00020
0.00025
0.00030
0.00035
0.00040
1000
2000
3000
4000
5000
6000
7000
8000
9000
Room.Board
Room.Board Distribution
2000
3000
4000
5000
6000
7000
8000
Room.Board
Books Distribution
Density
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0
500
1000
1500
2000
2500
Books
Books Distribution
0
500
1000
1500
2000
Books
Personal Distribution
Density
0.0000
0.0001
0.0002
0.0003
0.0004
0.0005
0.0006
0.0007
0.0008
0
1000
2000
3000
4000
5000
6000
7000
Personal
Personal Distribution
0
1000
2000
3000
4000
5000
6000
7000
Personal
PhD Distribution
Density
0.000
0.005
0.010
0.015
0.020
0.025
0
20
40
60
80
100
120
PhD
PhD Distribution
20
40
60
80
100
PhD

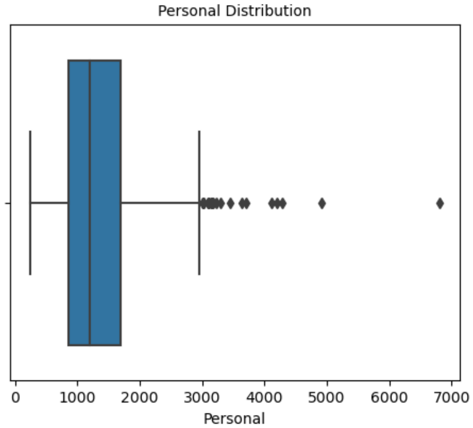

Room.Board Distribution
Density
0.00040
0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
1000
2000
3000
4000
5000
6000
7000
8000
9000
Room.Board

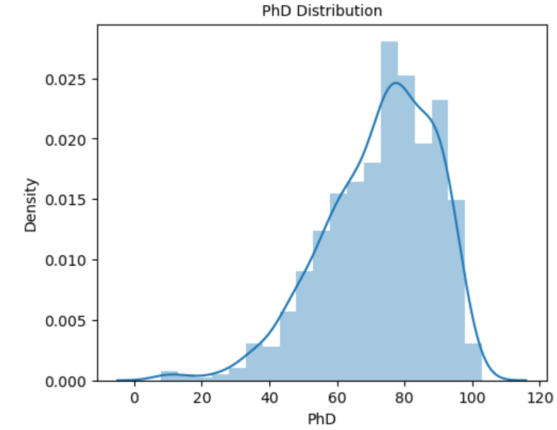

Room.Board Distribution
2000
3000
4000
5000
6000
7000
8000
Room.Board

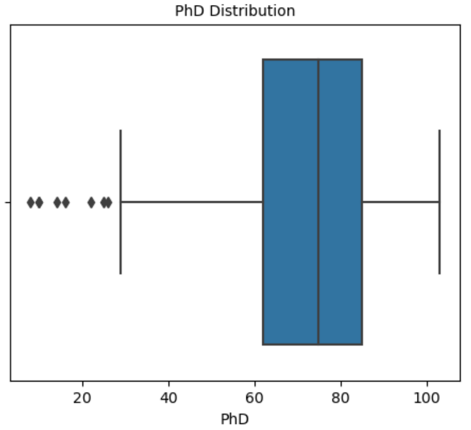

Books Distribution
Density
0.006
0.005
0.004
0.003
0.002
0.001
0.000
0
500
1000
1500
2000
2500
Books

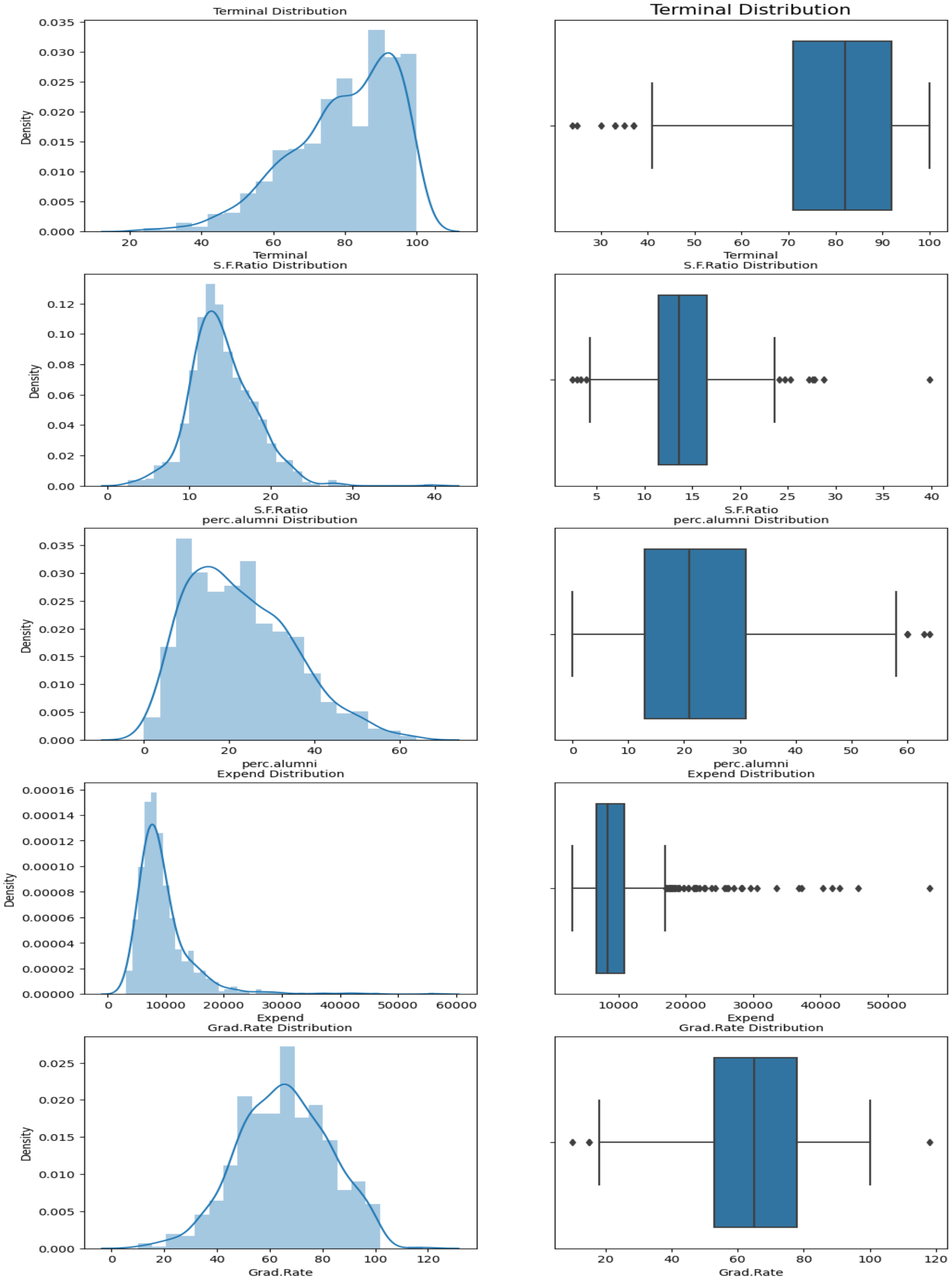

Books Distribution
0
500
1000
1500
2000
Books


Personal Distribution
Density
0.0008
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
0
1000
2000
3000
4000
5000
6000
7000
Personal


Personal Distribution
0
1000
2000
3000
4000
5000
6000
7000
Personal


PhD Distribution
Density
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100
120
PhD


PhD Distribution
20
40
60
80
100
PhD

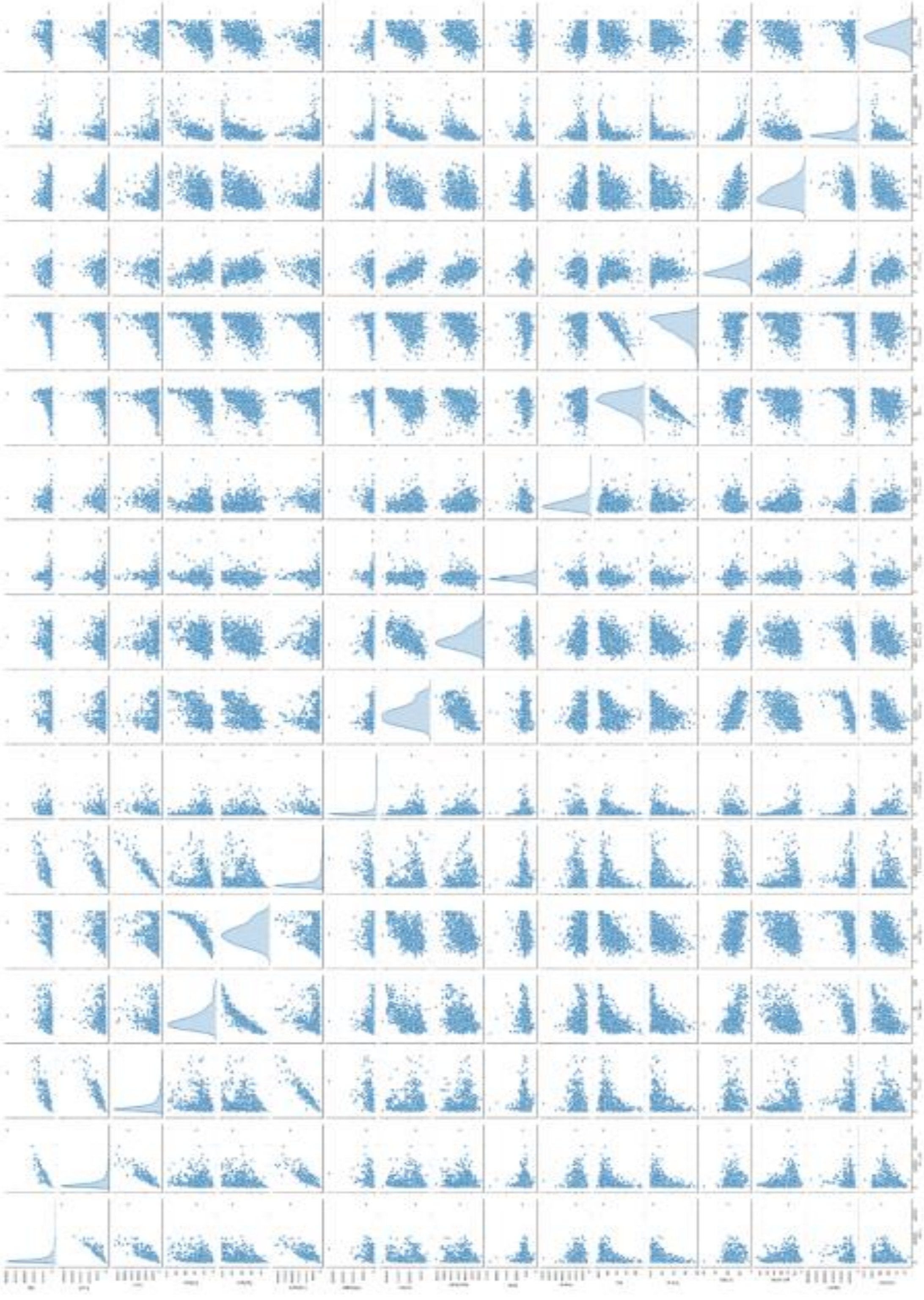


Bivariate-Multivariate Analysis:

The numerical variables are analyzed with correlation heatmap, variation was studied with Standard deviation and the relationship was studied with a pairplot

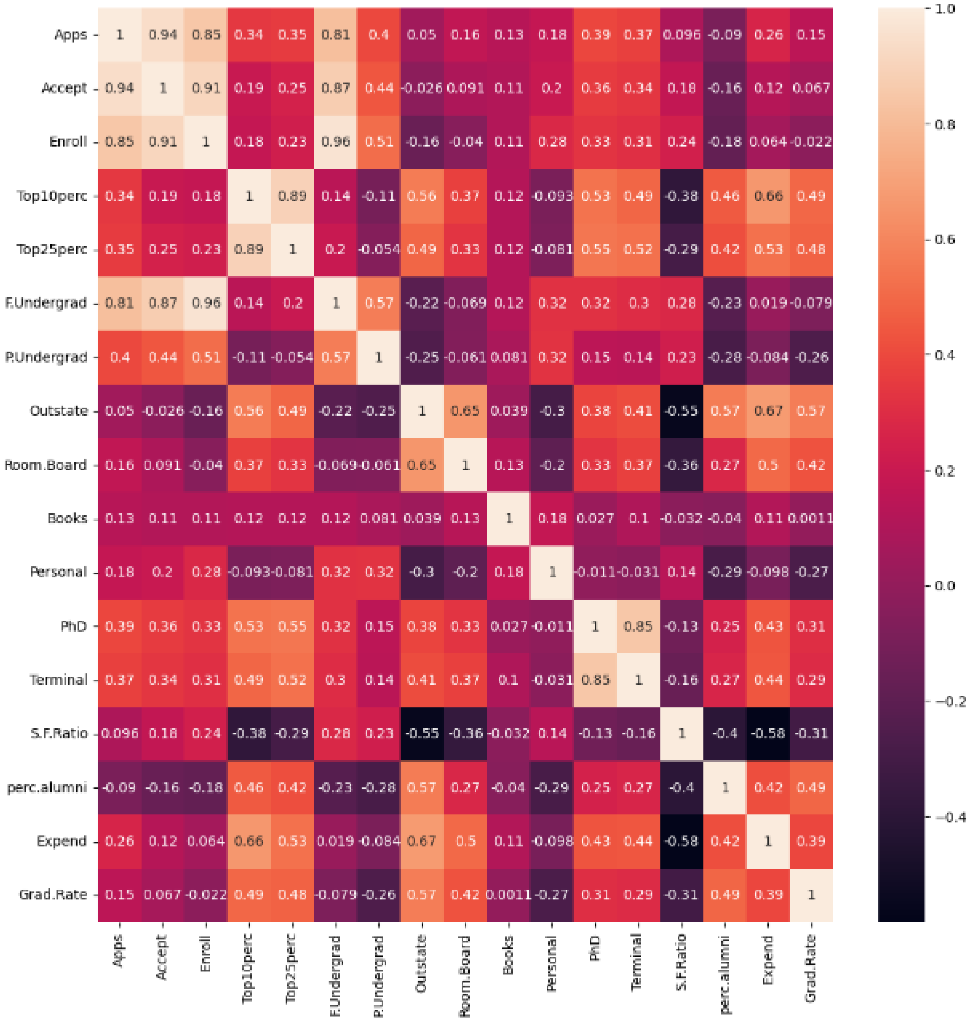
Pairplot

With Pairplot we are trying to find the relationship amongst variables across the dataset. We found some significant linear relationship



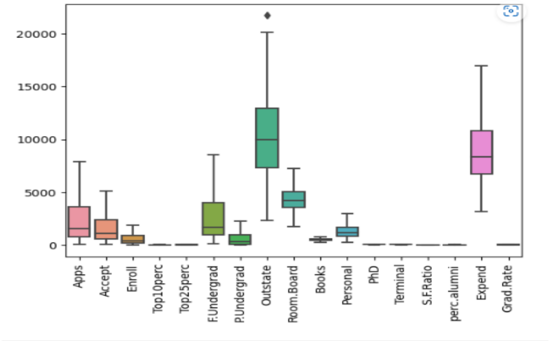
Correlation Heatmap

The Correlation coefficient was calculated amongst variables and results are discussed in commentary



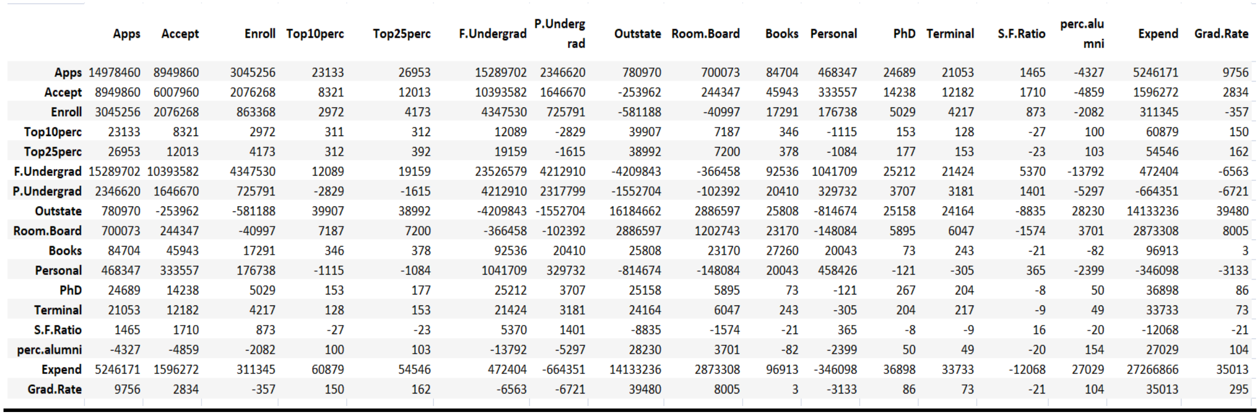
Outlier treatment:

The outliers were removed by using capping the upper range and lower range, calculated by using IQR



Covariance Matrix

As a part of Bivariate analysis, we calculated Covariance also represented in form of matrix



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Apps	14978460	8949860	3045256	23133	26953	15289702	2346620	780970	700073	84704	468347	24689	21053	1465	-4327	5246171	9756
Accept	8949860	6007960	2076268	8321	12013	10393582	1646670	-253962	244347	45943	333557	14238	12182	1710	-4859	1596272	2834
Enroll	3045256	2076268	863368	2972	4173	4347530	725791	-581188	-40997	17291	176738	5029	4217	873	-2082	311345	-357
Top10perc	23133	8321	2972	311	312	12089	-2829	39907	7187	346	-1115	153	128	-27	100	60879	150
Top25perc	26953	12013	4173	312	392	19159	-1615	38992	7200	378	-1084	177	153	-23	103	54546	162
F.Undergrad	15289702	10393582	4347530	12089	19159	23526579	4212910	-4209843	-366458	92536	1041709	25212	21424	5370	-13792	472404	-6563
P.Undergrad	2346620	1646670	725791	-2829	-1615	4212910	2317799	-1552704	-102392	20410	329732	3707	3181	1401	-5297	-664351	-6721
Outstate	780970	-253962	-581188	39907	38992	-4209843	-1552704	16184662	2886597	25808	-814674	25158	24164	-8835	28230	14133236	39480
Room.Board	700073	244347	-40997	7187	7200	-366458	-102392	2886597	1202743	23170	-148084	5895	6047	-1574	3701	2873308	8005
Books	84704	45943	17291	346	378	92536	20410	25808	23170	27260	20043	73	243	-21	-82	96913	3
Personal	468347	333557	176738	-1115	-1084	1041709	329732	-814674	-148084	20043	458426	-121	-305	365	-2399	-346098	-3133
PhD	24689	14238	5029	153	177	25212	3707	25158	5895	73	-121	267	204	-8	50	36898	86
Terminal	21053	12182	4217	128	153	21424	3181	24164	6047	243	-305	204	217	-9	49	33733	73
S.F.Ratio	1465	1710	873	-27	-23	5370	1401	-8835	-1574	-21	365	-8	-9	16	-20	-12068	-21
perc.alumni	-4327	-4859	-2082	100	103	-13792	-5297	28230	3701	-82	-2399	50	49	-20	154	27029	104
Expend	5246171	1596272	311345	60879	54546	472404	-664351	14133236	2873308	96913	-346098	36898	33733	-12068	27029	27266866	35013
Grad.Rate	9756	2834	-357	150	162	-6563	-6721	39480	8005	3	-3133	86	73	-21	104	35013	295

EXECUTIVE COMMENTARY

We can recognize the apps variable is positively correlated with accepted, students enrolled and full time graduates.

Henceforth this map gives the pattern on when student whoever is submitting the application, it is accepted and he will be enrolled as fulltime graduate.

We can see inverse relationship between apps and percentage of alumni. This gives us an insight that not all the students are part of alumni of their respective or university.

Max expenditure happens from outstate students, they also board many rooms.

Similarly good positive correlations are seen students of top 10 and top 25% of higher secondary class and the graduation rate. Most of PHD holders are having Terminal positions

These variables also have underlying commonality amongst themselves and can be grouped. With no. of numerical variables being high, it is recommended to group instead of having Top 10 percent, Top 25 percent can be grouped as , Full time undergraduate, part time undergraduate can be grouped as Undergraduate and so forth

High variance is found in expenditure, Outstate students, Full time undergraduates