Business Report for
Wholesale Customer Analysis
and
Education Post 12th Standard Analysis

Name: Aishwariya Hariharan
PGP-DSBA Online September’ 23
Date: 28/11/2023

Contents

Sl. no.	Topic	Page no.
	Wholesale customer analysis	
1.	Business context	3
2.	Objective	3
3.	Data overview	3
4.	Univariate analysis	6
5.	Multivariate analysis	9
6.	Actionable insights and recommendation	14
	Education post 12th standard analysis	
1.	Objective	16
2.	Data overview	16
3.	Univariate analysis	19
4.	Multivariate analysis	22
5.	Actionable insights and recommendation	23

Wholesale Customer Analysis

Business Context

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 varieties of products in 3 different regions (Lisbon, Oporto, Other) and across two sales channels - Hotel and Retail.

Objective

The distributor wants to analyze the data to get a fair idea about the demand of different buyers which will help them enhance their customer experience. We need to analyze the data to answer questions that will help the company improve its business.

Data Overview

- A sample of the dataset.
    - ★ The following table shows the top 5 rows of the dataset. It gives us an overview of what the dataset, comprising of transactions, looks like.

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
0	1	Retail	Other	12669	9656	7561	214	2674	1338
1	2	Retail	Other	7057	9810	9568	1762	3293	1776
2	3	Retail	Other	6353	8808	7684	2405	3516	7844
3	4	Hotel	Other	13265	1196	4221	6404	507	1788
4	5	Retail	Other	22615	5410	7198	3915	1777	5185

- Data Dictionary

  - ★ Buyer/Spender- IDs of customers
  - ★ Channel - Mode of sales (hotel or retail)
  - ★ Region- Region of the distributor
  - ★ Fresh- Spending on fresh vegetables
  - ★ Milk- Spending on milk
  - ★ Grocery- Spending on grocery
  - ★ Frozen- Spending on frozen food
  - ★ Detergents_Paper- Spending on detergents and toilet paper
  - ★ Delicatessen- Spending on instant food

- Data information

  - ★ The dataset has 440 rows (number of entries) and 9 columns (which are the variables)
  - ★ Channel and region are categorical variables and the rest of them are numerical variables (continuous).
  - ★ Buyer/Spender variable represents a unique row number for each transaction detail.
  - ★ There are two channels through which the sales are happening - hotel and retail.
  - ★ The distributor primarily operates in 3 different regions of Portugal - Lisbon, Oporto, and other regions of Portugal.
  - ★ The dataset has customers' transaction information for 6 different categories of items - fresh products, milk, grocery, frozen food, detergents, and instant food.
  - ★ There are no duplicate or null values in the dataset.

```
RangeIndex: 440 entries, 0 to 439
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)
```

- Statistical summary of the dataset

  ★ The statistical summary shows the central tendencies - mean, standard deviation, and the 5-point summary (min, 25%, 50% (i.e. the median), 75%, and max)

  ★ We can see that the maximum average spending is on groceries with a mean of 7951.

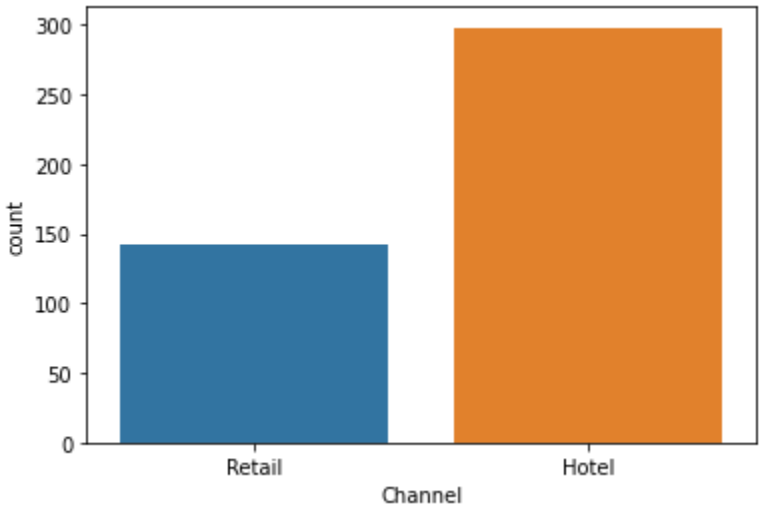
	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
mean	220.500000	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	127.161315	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937
min	1.000000	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	110.750000	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000
50%	220.500000	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000
75%	330.250000	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000
max	440.000000	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000

Exploratory Data Analysis

Univariate Analysis

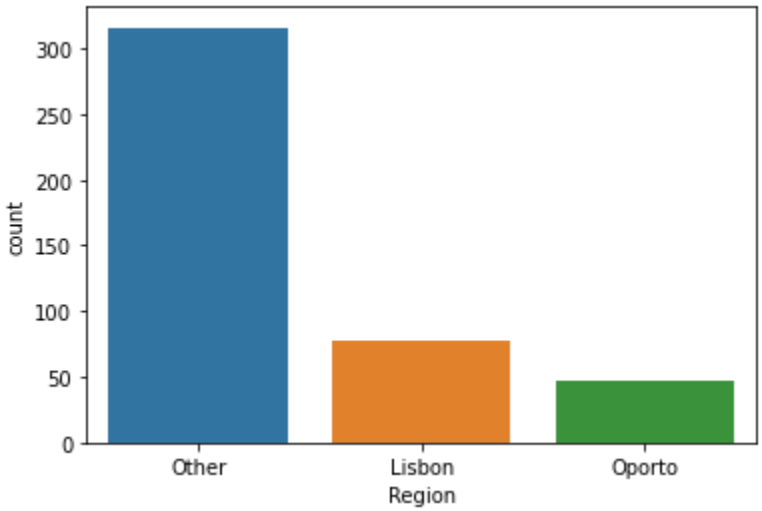
- Observation of channel distribution

  ★ The number of buyers/spenders through hotels is almost double the number of buyers/spenders through the retail channel.



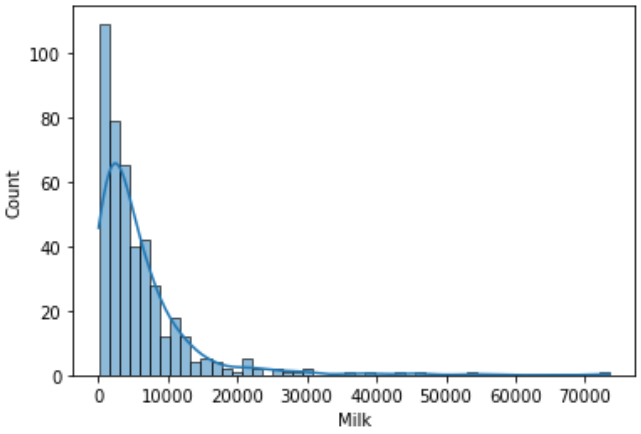
- Observation of sales across regions

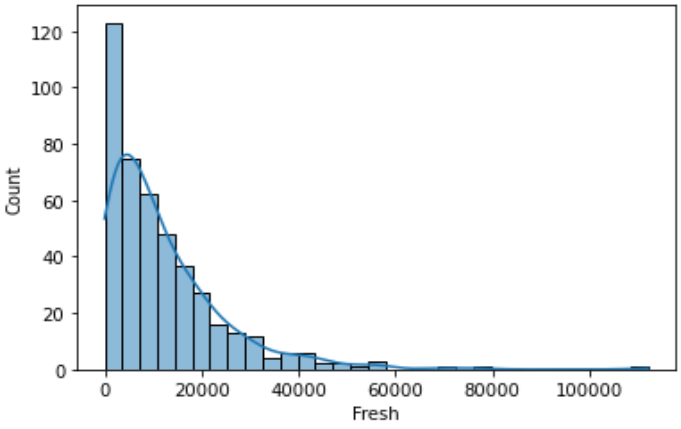
  ★ The majority of buyers/spenders are distributed across other regions of Portugal.

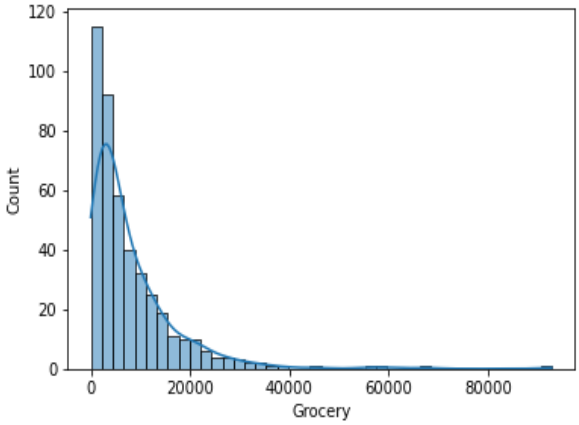


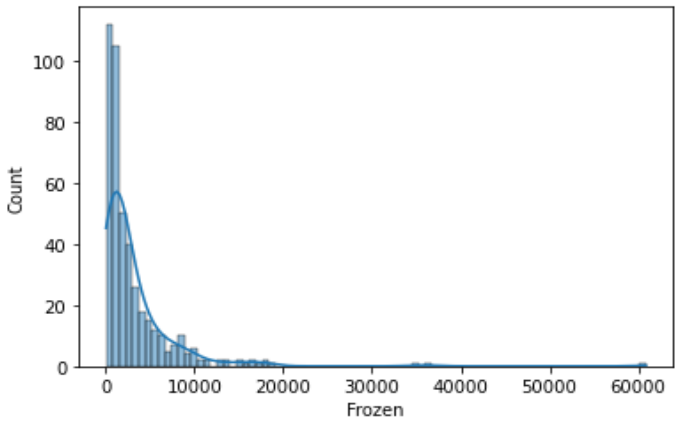
- Observation of spending across categories of items

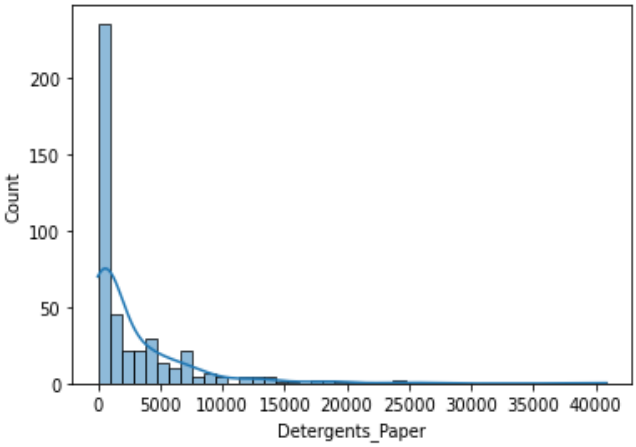
  ★ Most purchased category - fresh food
  ★ Least purchased category - instant food (Delicatessen)
  ★ Purchases across all categories are right-skewed

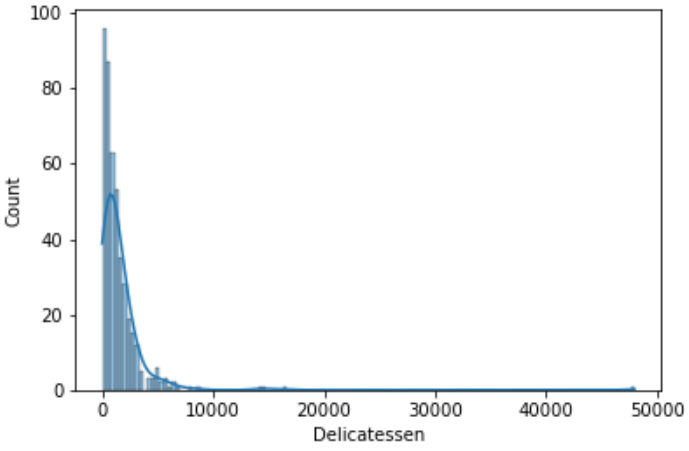






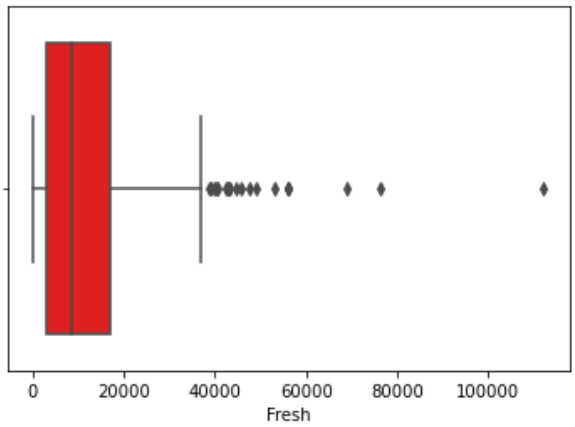


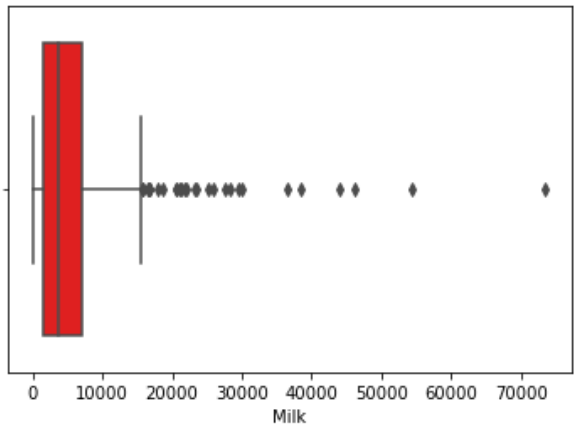


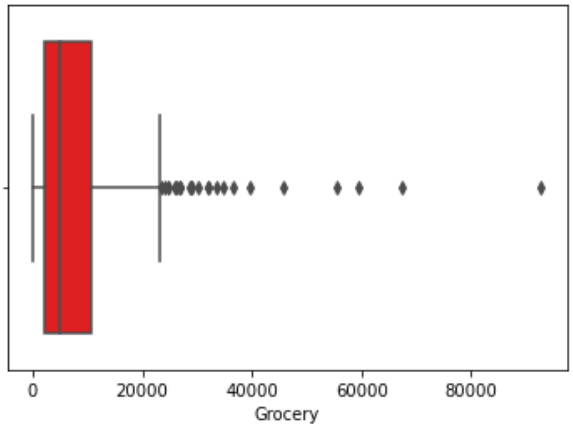


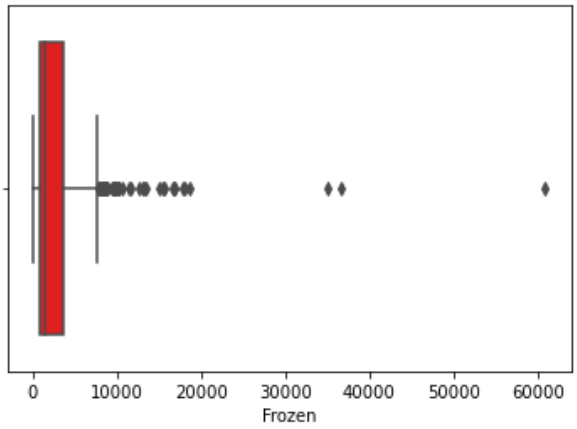
- Outliers

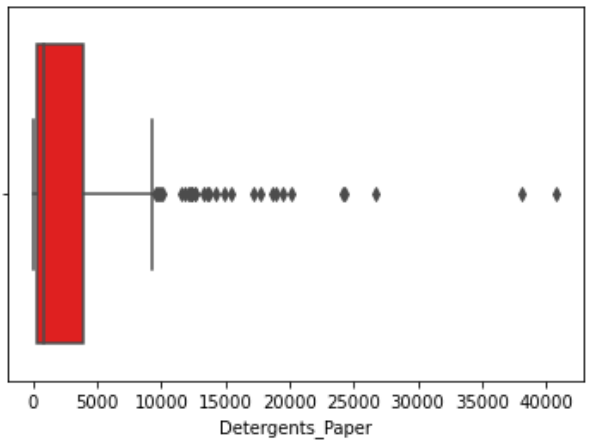
  ★ All categories have outliers and they can be treated using the z-score or box plot method.
  ★ However, since this project is restricted to exploratory data analysis, we refrain from treating the outliers.

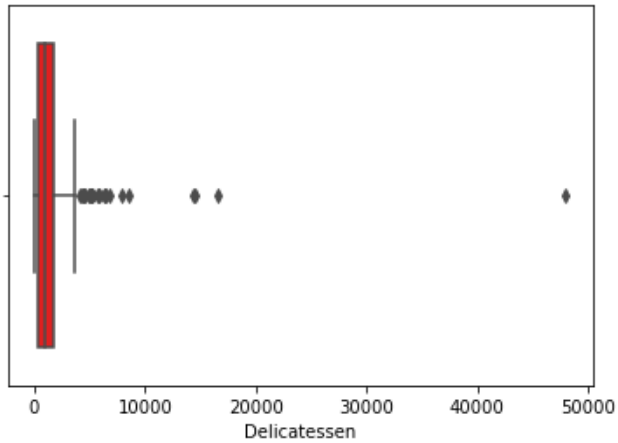












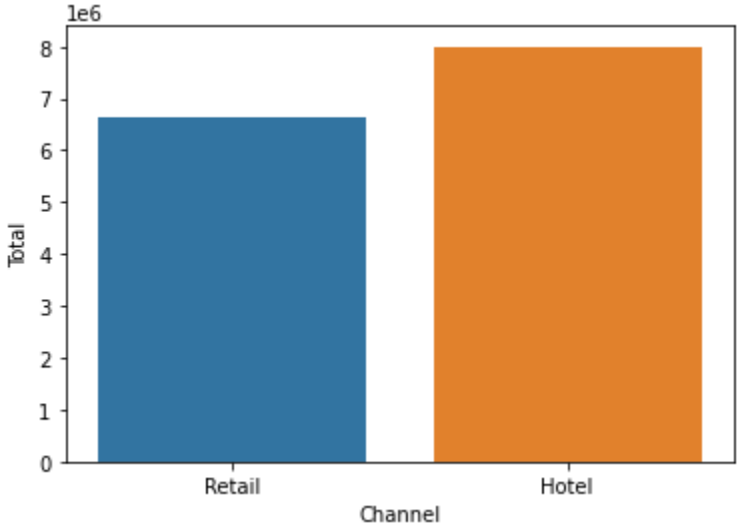
Multivariate Analysis

- The total spending by each customer is shown in the last column of the dataset.

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
0	1	Retail	Other	12669	9656	7561	214	2674	1338	34112
1	2	Retail	Other	7057	9810	9568	1762	3293	1776	33266
2	3	Retail	Other	6353	8808	7684	2405	3516	7844	36610
3	4	Hotel	Other	13265	1196	4221	6404	507	1788	27381
4	5	Retail	Other	22615	5410	7198	3915	1777	5185	46100

- Total spending across channels

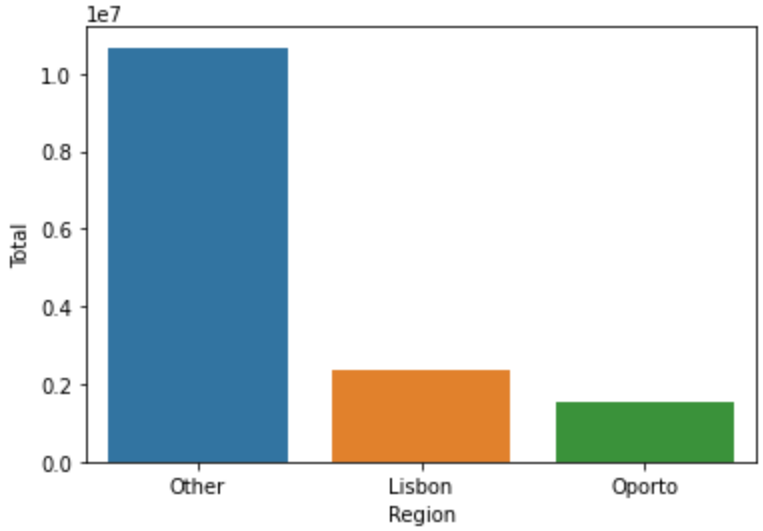
  ★ Total spending by hotels is 7999569, which is higher than the retail mode.
  ★ Total spending by retail customers is 6619931



- Total spending across regions

```
Region
Lisbon     2386813
Oporto     1555088
Other     10677599
```

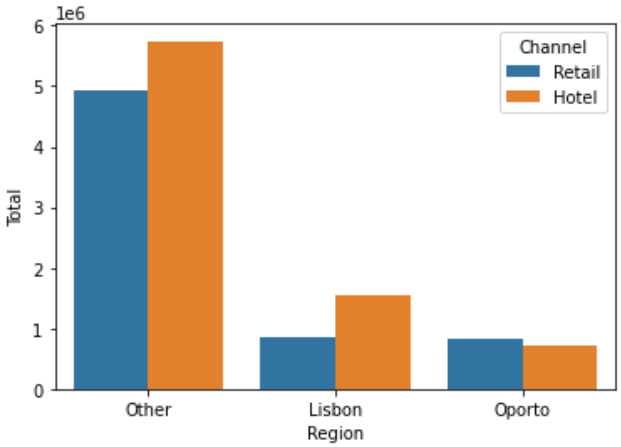
  ★ Most of the revenue is generated from other regions.
  ★ The least revenue comes from the Oporto region.

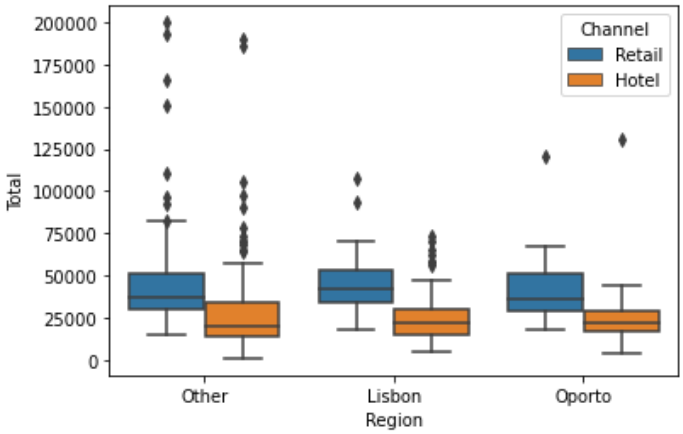


- Total spending across regions via different channels

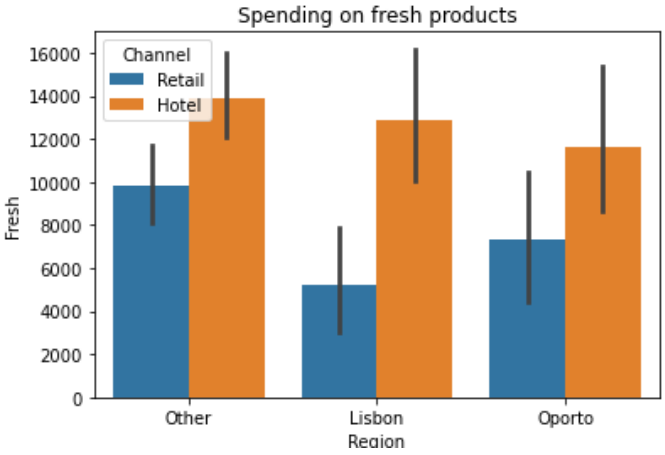
  ★ The total spending by hotel customers of the distributors is higher than retailers across different regions, except for Oporto where the difference is not very high. This could be because of the outlier values.

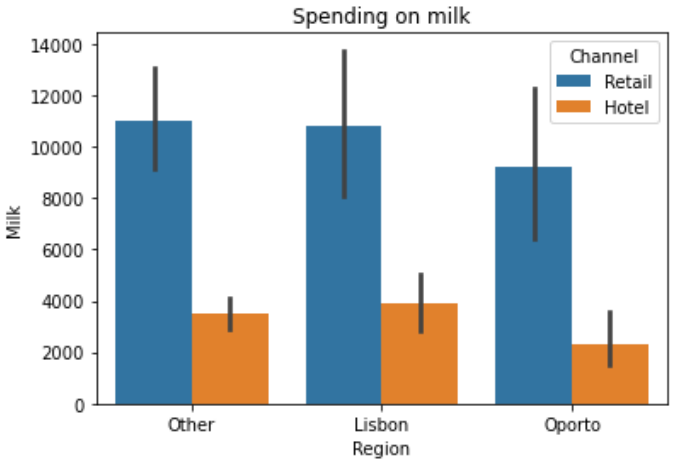
  ★ The average spending of retail customers is higher than that of hotel customers across regions.

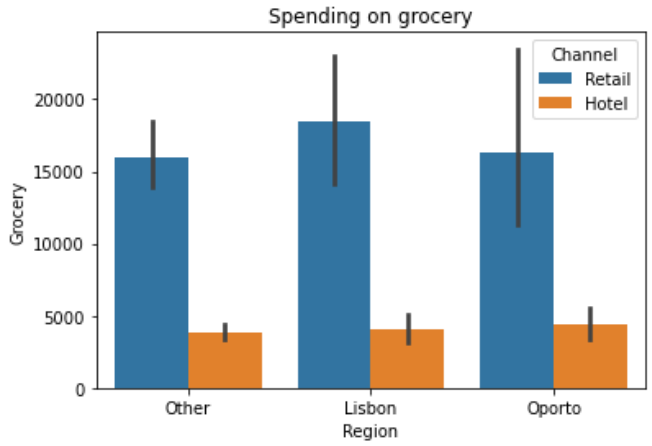


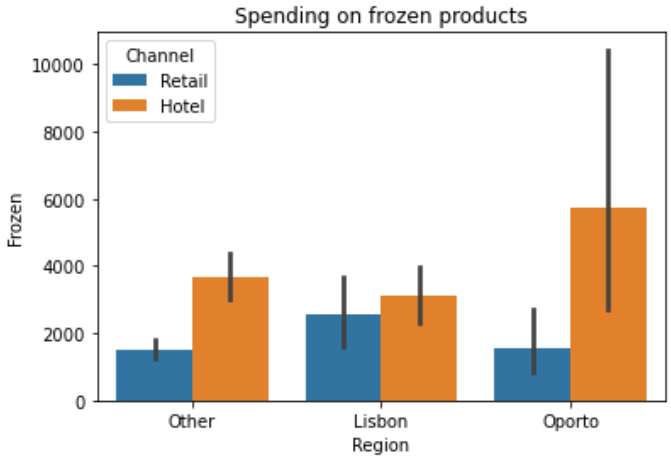


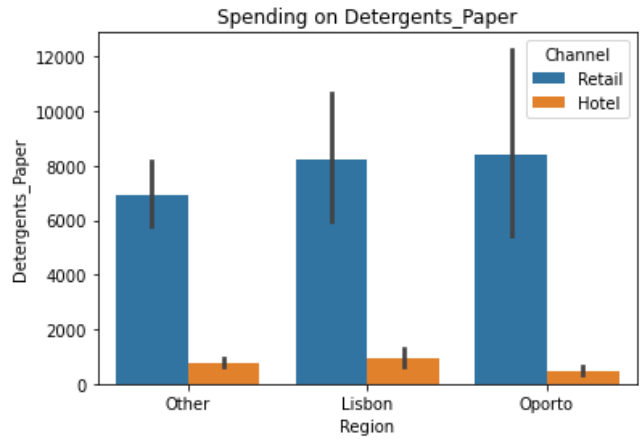
- Spending on each category across regions and channels.
  - ★ As we saw previously, expenditure on fresh products is the highest. Hotels in other regions of Portugal have spent the highest among others.
  - ★ On milk, grocery, detergents & paper, and instant food, retail customers spend way higher than hotel customers across all regions.
  - ★ Oporto region consumes more frozen food than other regions.

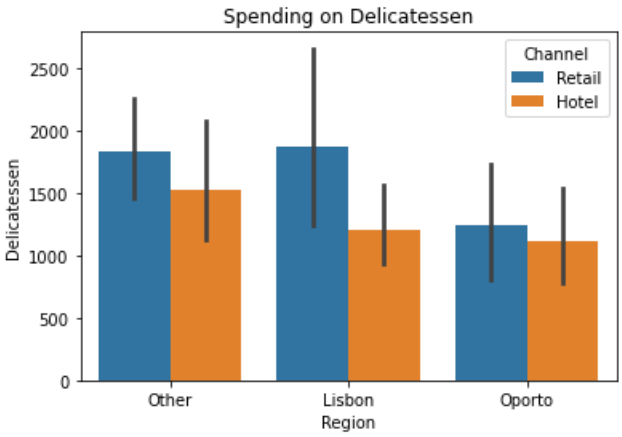












- Statistical summary - variable wise

    ★ The standard deviation varies considerably, across a huge range, for all variables.
    ★ So the variables do not show similar behaviour.
    ★ These summaries help us understand better the following parameters for each variable:
        ○ The total number of buyers/spenders category/variable-wise
        ○ The highest and lowest minimum amount spent, and particularly on the kind of products it was spent

For hotel

	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
count	298.000000	298.000000	298.000000	298.000000	298.000000	298.000000	298.000000	298.000000
mean	238.369128	13475.560403	3451.724832	3962.137584	3748.251678	790.560403	1415.956376	26844.191275
std	120.910343	13831.687502	4352.165571	3545.513391	5643.912500	1104.093673	3147.426922	22164.839073
min	4.000000	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000	904.000000
25%	137.250000	4070.250000	1164.500000	1703.750000	830.000000	183.250000	379.000000	13859.250000
50%	241.500000	9581.500000	2157.000000	2684.000000	2057.500000	385.500000	821.000000	21254.500000
75%	344.500000	18274.750000	4029.500000	5076.750000	4558.750000	899.500000	1548.000000	32113.750000
max	440.000000	112151.000000	43950.000000	21042.000000	60869.000000	6907.000000	47943.000000	190169.000000

For retail

	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
count	142.000000	142.000000	142.000000	142.000000	142.000000	142.000000	142.000000	142.000000
mean	183.000000	8904.323944	10716.500000	16322.852113	1652.612676	7269.507042	1753.436620	46619.232394
std	132.136132	8987.714750	9679.631351	12267.318094	1812.803662	6291.089697	1953.797047	29346.866491
min	1.000000	18.000000	928.000000	2743.000000	33.000000	332.000000	3.000000	14993.000000
25%	61.250000	2347.750000	5938.000000	9245.250000	534.250000	3683.500000	566.750000	30147.250000
50%	166.500000	5993.500000	7812.000000	12390.000000	1081.000000	5614.500000	1350.000000	37139.000000
75%	303.750000	12229.750000	12162.750000	20183.500000	2146.750000	8662.500000	2156.000000	51650.500000
max	438.000000	44466.000000	73498.000000	92780.000000	11559.000000	40827.000000	16523.000000	199891.000000

For Oporto

	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
count	47.000000	47.000000	47.000000	47.000000	47.000000	47.000000	47.000000	47.000000
mean	317.000000	9887.680851	5088.170213	9218.595745	4045.361702	3687.468085	1159.702128	33086.978723
std	13.711309	8387.899211	5826.343145	10842.745314	9151.784954	6514.717668	1050.739841	24234.507325
min	294.000000	3.000000	333.000000	1330.000000	131.000000	15.000000	51.000000	4129.000000
25%	305.500000	2751.500000	1430.500000	2792.500000	811.500000	282.500000	540.500000	20611.500000
50%	317.000000	8090.000000	2374.000000	6114.000000	1455.000000	811.000000	898.000000	26953.000000
75%	328.500000	14925.500000	5772.500000	11758.500000	3272.000000	4324.500000	1538.500000	36158.500000
max	340.000000	32717.000000	25071.000000	67298.000000	60869.000000	38102.000000	5609.000000	130877.000000

For Lisbon

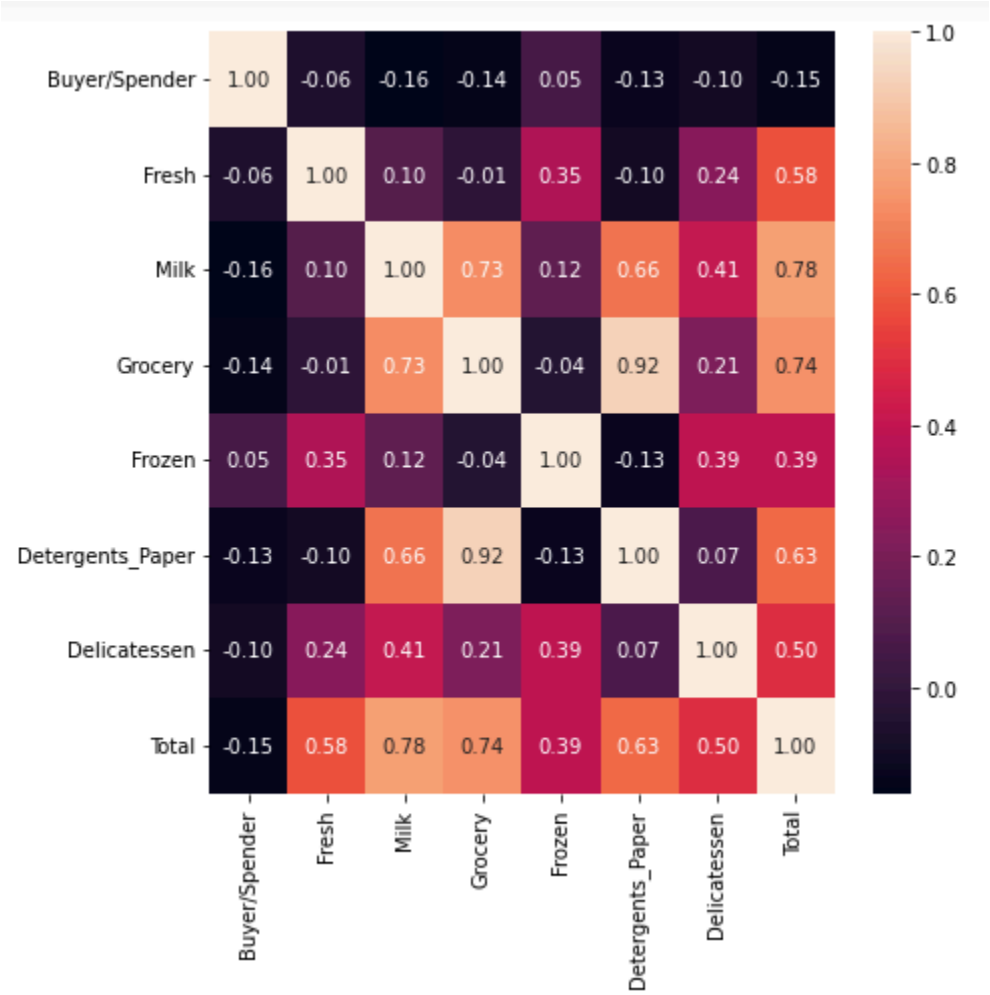
	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
count	77.000000	77.000000	77.000000	77.000000	77.000000	77.000000	77.000000	77.000000
mean	235.000000	11101.727273	5486.415584	7403.077922	3000.337662	2651.116883	1354.896104	30997.571429
std	22.371857	11557.438575	5704.856079	8496.287728	3092.143894	4208.462708	1345.423340	20321.813773
min	197.000000	18.000000	258.000000	489.000000	61.000000	5.000000	7.000000	4925.000000
25%	216.000000	2806.000000	1372.000000	2046.000000	950.000000	284.000000	548.000000	17184.000000
50%	235.000000	7363.000000	3748.000000	3838.000000	1801.000000	737.000000	806.000000	25385.000000
75%	254.000000	15218.000000	7503.000000	9490.000000	4324.000000	3593.000000	1775.000000	38699.000000
max	273.000000	56083.000000	28326.000000	39694.000000	18711.000000	19410.000000	6854.000000	107155.000000

For other regions of Portugal combined

	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
count	316.000000	316.000000	316.000000	316.000000	316.000000	316.000000	316.000000	316.000000
mean	202.613924	12533.471519	5977.085443	7896.363924	2944.594937	2817.753165	1620.601266	33789.870253
std	143.615303	13389.213115	7935.463443	9537.287778	4260.126243	4593.051613	3232.581660	27949.337752
min	1.000000	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000	904.000000
25%	79.750000	3350.750000	1634.000000	2141.500000	664.750000	251.250000	402.000000	17209.250000
50%	158.500000	8752.500000	3684.500000	4732.000000	1498.000000	856.000000	994.000000	28029.000000
75%	361.250000	17406.500000	7198.750000	10559.750000	3354.750000	3875.750000	1832.750000	42492.250000
max	440.000000	112151.000000	73498.000000	92780.000000	36534.000000	40827.000000	47943.000000	199891.000000

- Correlation between different numerical variables

    ★ The highest correlation (0.92) is between groceries and detergents. I.e. if a buyer purchases groceries, then they are very likely to purchase detergents as well and vice-versa.
    ★ There's also a good correlation between groceries and milk.



Actionable Insights and Recommendations

1. Hotel spends more than Retail across all regions. Identifying this spending pattern, offers and marketing strategies can be tailored to reach out to more Hotels in and around the region.
2. The average spending of Retail is greater than Hotels across all regions, and retail spends significantly more on milk, groceries, and detergents/paper compared to Hotels. The minimum amount spent by Retail on these products is also quite high. The

strategies used for these products in these regions for Retail can be mirrored for more regions in the Retail channel

3. Revenue generation from Oporto is very low. Marketing strategies, offers, partnerships, etc., can be explored to increase sales in Oporto.
4. The skewed distribution of spending across all categories implies that customers prefer spending over a certain price range of products, i.e., on specific products. This knowledge can help manage inventory and time the offers and promotions.
5. Frozen food sales are higher in Hotels than in Retail, and Oporto has a greater customer base for frozen food than any other region. Understanding why this is the case can help increase frozen food sales in other or similar regions.
6. Tailored offers, for Hotels or Retail, membership options, loyalty programs, etc. can help increase and retain the customer base.
7. Create marketing campaigns and promotions based on the insights provided by the data for different regions, channels of sales, and categories of products.
8. Insights from this data can be utilised to implement sales strategies in regions similar to Portugal demographically.
9. More focus on increasing revenue generation from the Oporto region can be done by increasing brand awareness, offers, etc.

Education Post 12th Standard

Objective

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 Overview

- A sample of the dataset.
    - ★ The following table shows the top 5 rows of the dataset. It gives us an overview of what the dataset, comprising of transactions, looks like.

	Names	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.a
0	Abilene Christian University	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	
1	Adelphi University	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	
2	Adrian College	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	
3	Agnes Scott College	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	
4	Alaska Pacific University	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	

- Data dictionary
    - ★ Names: Names of various universities 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 PhDs
- ★ 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

- Data information

  - ★ The dataset has 777 rows (number of entries) and 18 columns (which are the variables).
  - ★ Only Names is a categorical variable, the rest of the variables are numerical (continuous).
  - ★ There are no duplicate or null values in the dataset.

```
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       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
```

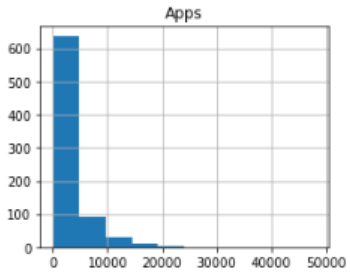
- Statistical summary of the dataset

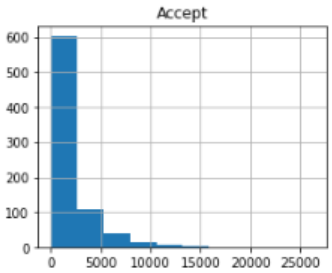
    ★ The statistical summary shows the central tendencies - mean, standard deviation, and the 5-point summary (min, 25%, 50% (i.e. the median), 75%, and max)

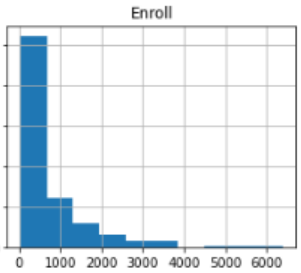
	count	mean	std	min	25%	50%	75%	max
Apps	777.0	3002.0	3870.0	81.0	776.0	1558.0	3624.0	48094.0
Accept	777.0	2019.0	2451.0	72.0	604.0	1110.0	2424.0	26330.0
Enroll	777.0	780.0	929.0	35.0	242.0	434.0	902.0	6392.0
Top10perc	777.0	28.0	18.0	1.0	15.0	23.0	35.0	96.0
Top25perc	777.0	56.0	20.0	9.0	41.0	54.0	69.0	100.0
F.Undergrad	777.0	3700.0	4850.0	139.0	992.0	1707.0	4005.0	31643.0
P.Undergrad	777.0	855.0	1522.0	1.0	95.0	353.0	967.0	21836.0
Outstate	777.0	10441.0	4023.0	2340.0	7320.0	9990.0	12925.0	21700.0
Room.Board	777.0	4358.0	1097.0	1780.0	3597.0	4200.0	5050.0	8124.0
Books	777.0	549.0	165.0	96.0	470.0	500.0	600.0	2340.0
Personal	777.0	1341.0	677.0	250.0	850.0	1200.0	1700.0	6800.0
PhD	777.0	73.0	16.0	8.0	62.0	75.0	85.0	103.0
Terminal	777.0	80.0	15.0	24.0	71.0	82.0	92.0	100.0
S.F.Ratio	777.0	14.0	4.0	2.5	11.5	13.6	16.5	39.8
perc.alumni	777.0	23.0	12.0	0.0	13.0	21.0	31.0	64.0
Expend	777.0	9660.0	5222.0	3186.0	6751.0	8377.0	10830.0	56233.0
Grad.Rate	777.0	65.0	17.0	10.0	53.0	65.0	78.0	118.0

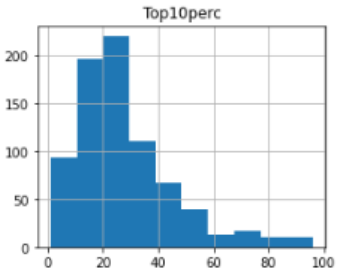
Exploratory Data Analysis

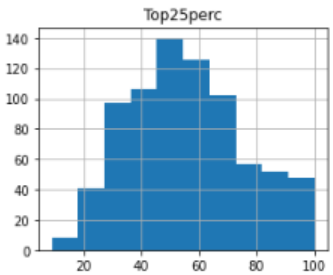
Univariate Analysis

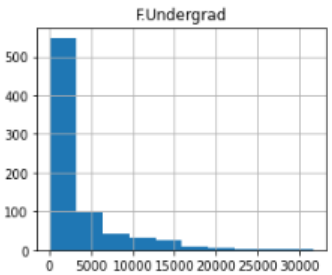


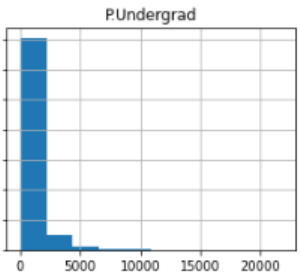


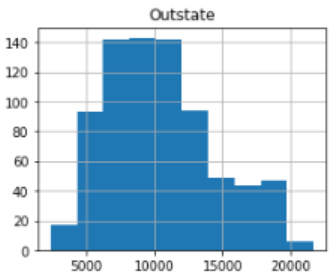


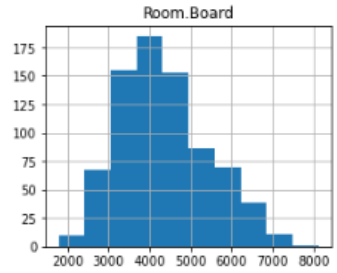


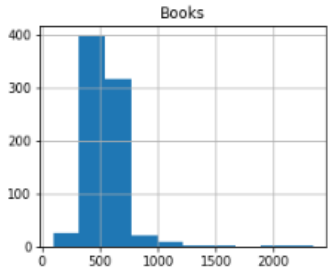


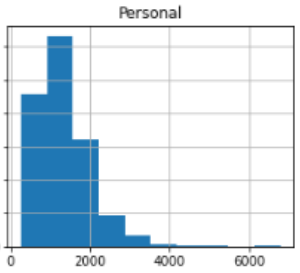


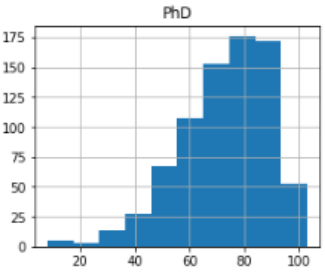


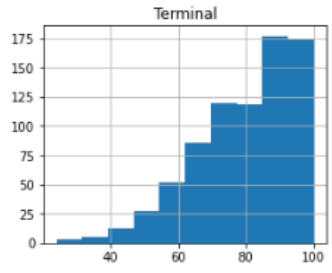


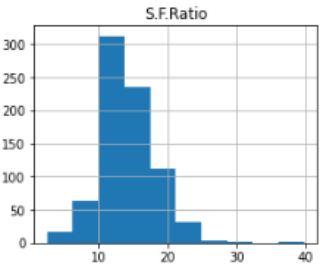


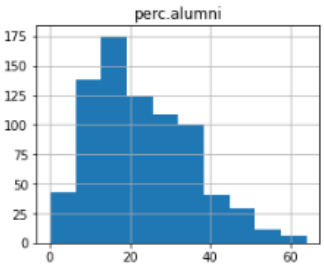


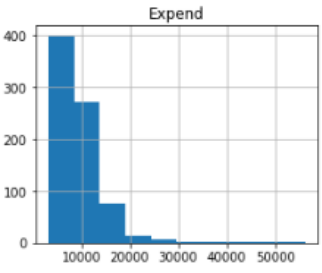


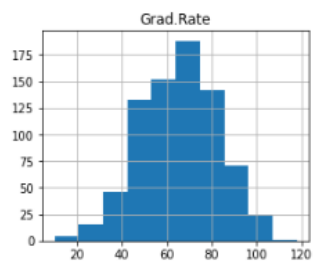






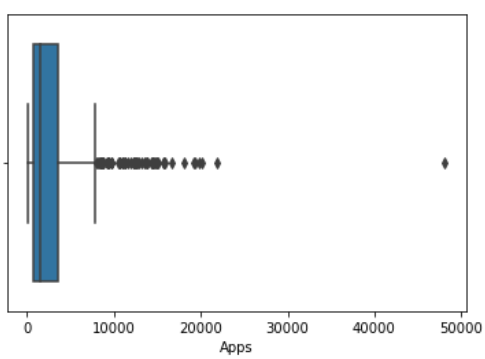


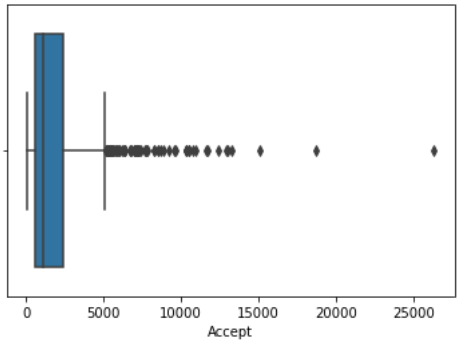


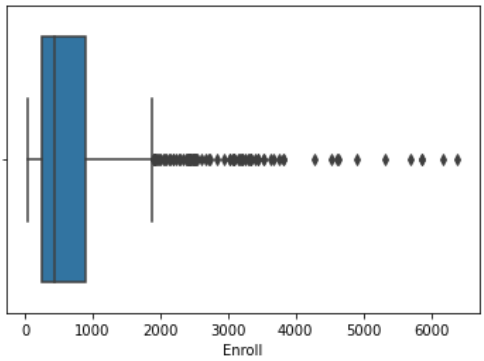


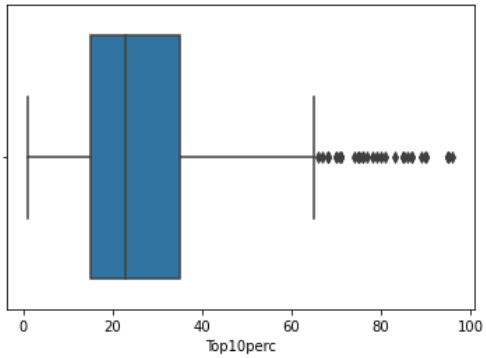
Observations:

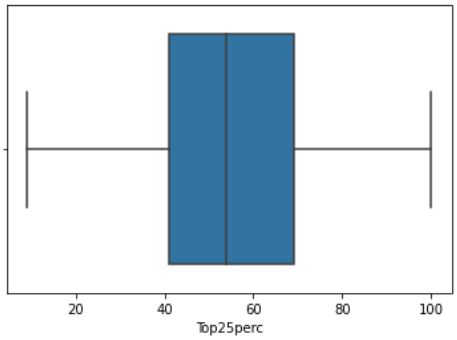
- ★ Apps, Accept, Enroll, Top10perc, F.Undergrad, P.Undergrad, Books, Personal, and Expend variables are highly skewed to the right.
- ★ Only PhD and Terminal are skewed to the left, that too moderately - i.e., the number of universities with a higher percentage of faculty members with PhD and Terminal Degrees is high.

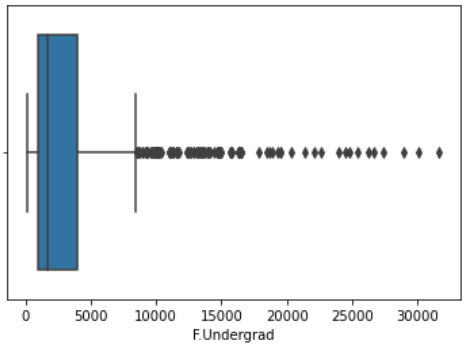


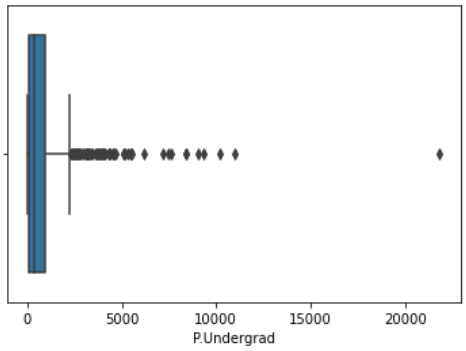


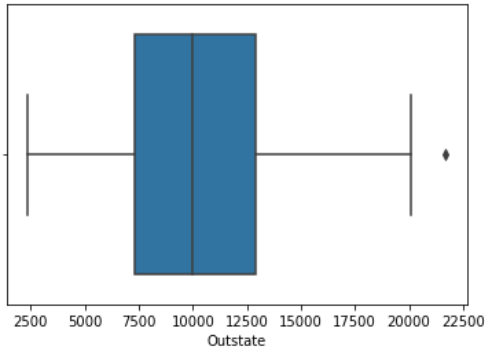


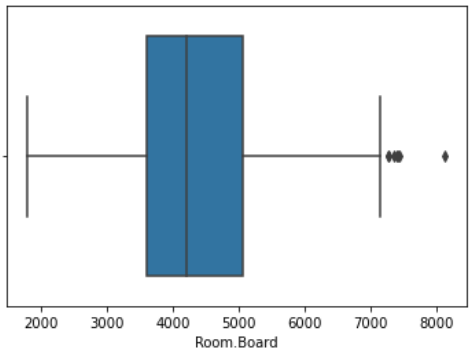


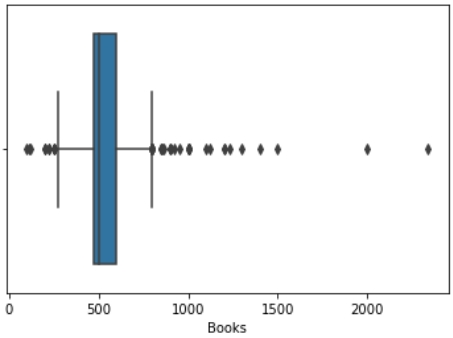


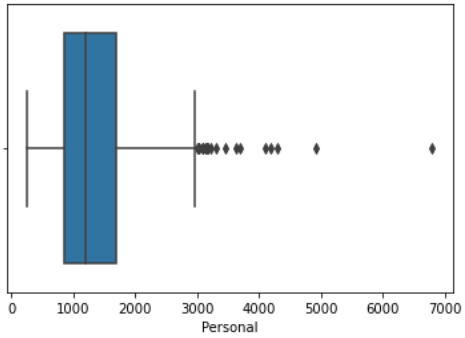


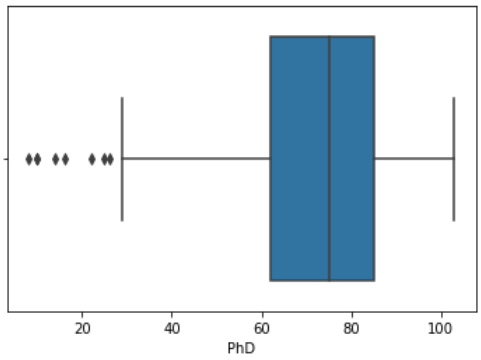


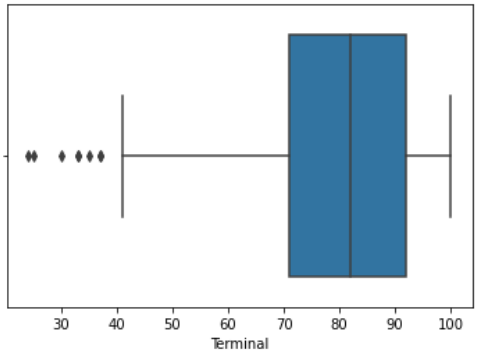


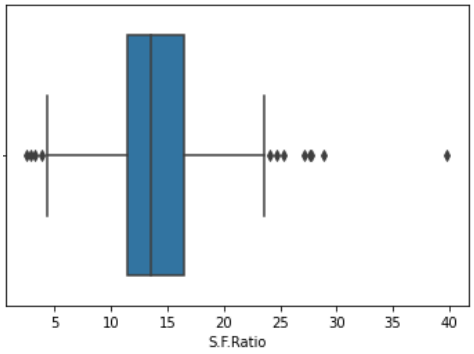


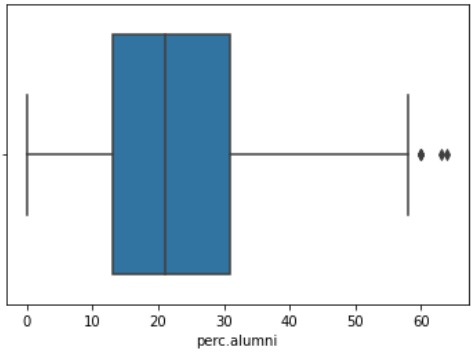


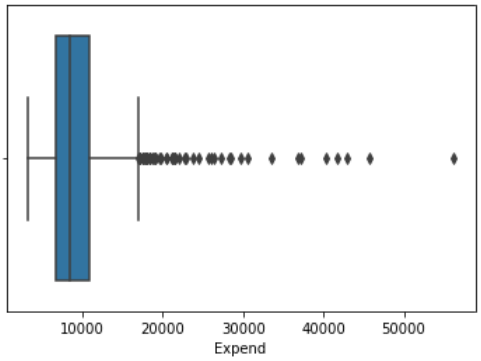


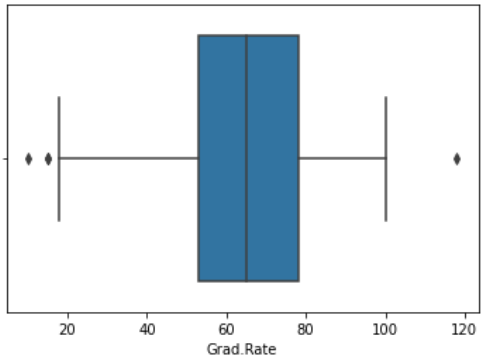










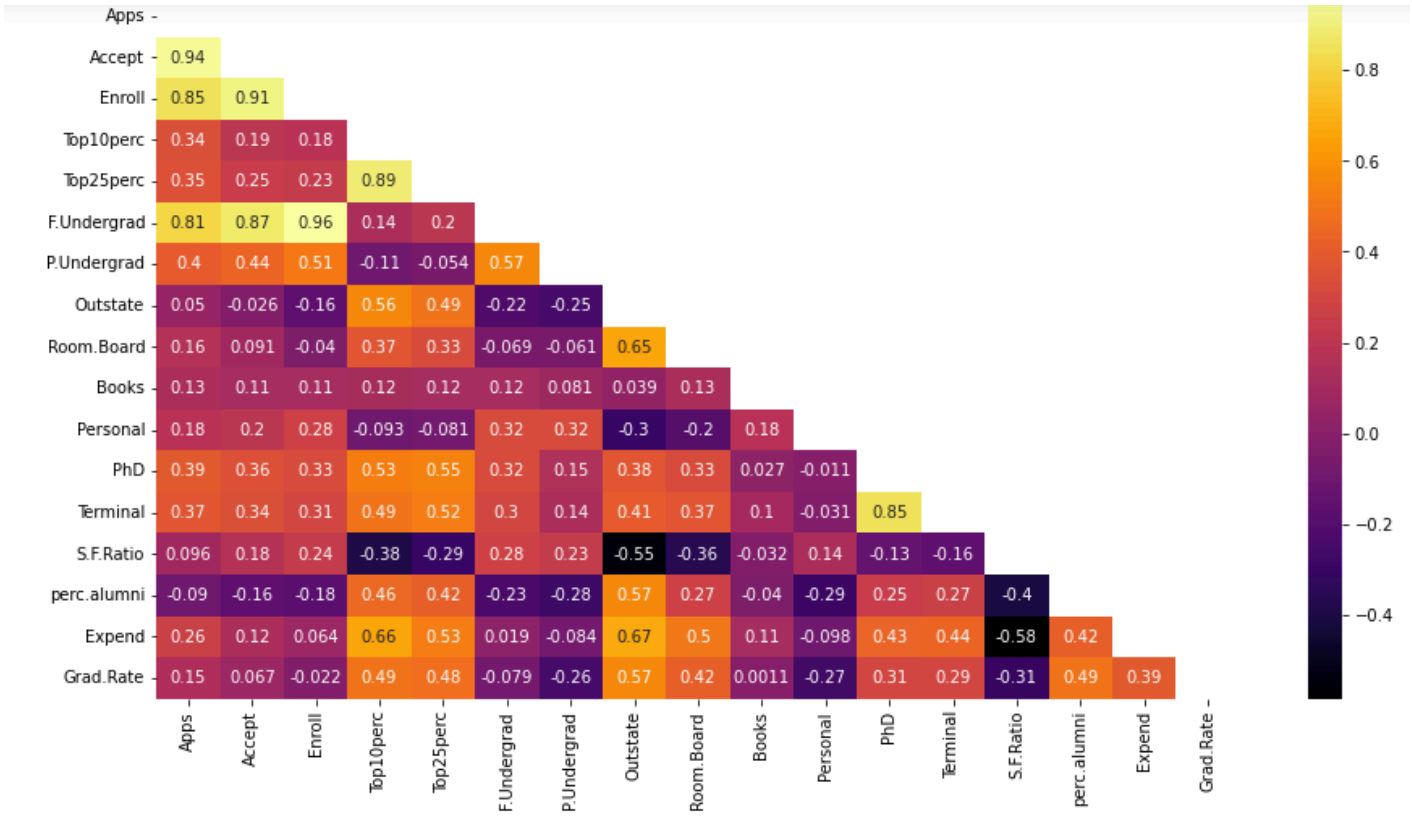


Observations:

★ All variables have outliers except the Top25perc variable.

Multivariate Analysis

Correlation between all variables.



Observations:

- ★ The highest correlation is 0.96, which is between F.Undergrad and Enroll variables. Implying, a large number of newly enrolled students belong to the full-time undergraduate category.
- ★ The correlation between Apps and Accept variables is 0.94. Implying, higher the number of applications, higher the acceptance of number of applications.
- ★ The correlation between the following variables is also quite high:
    - Enroll and Accept
    - Top25perc and Top10perc
    - Accept and F.Undergrad
    - Enroll and Apps
    - PhD and Terminal
    - Apps and F.Undergrad
    - Apps and Enroll

★ We can observe the first row from the bottom to see the effect of graduation rate on variables of interest such as perc.alumni, Terminal, PhD, Outstate, Top25perc, Top10perc.

Actionable Insights and Recommendations

1. The mean number of applications received is about 3000, however, the mean of acceptance is only around 2020. Understanding the reason – competency of students, application processing, criteria for filtering students, etc – can help improve the acceptance ratio.
2. The mean costs for students' rooms, books, and personal expenses are high and could cause potential financial distress to students. Interest-free loans, EMI options, scholarships, etc. can be provided to deserving students.
3. Accommodative measures/processes can be adopted to enable more part-time students to enroll.