SMDM Project Report

Contents

List of Figures

List of Tables

Problem 1: Wholesale Customers Analysis

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 Use methods of descriptive statistics to summarize data. Which Region and which Channel spent the most? Which Region and which Channel spent the least?

Data Description:

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 channel (Hotel, Retail).

Domain:

Retail

Data Summary:-

Summary statistics of data is as shown below :-

	count	mean	Std	min	25%	50%	75%	max
Buyer/Spender	440.0	220.50	127.16	1.0	110.75	220.5	330.25	440.0
Fresh	440.0	12000.30	12647.33	3.0	3127.75	8504.0	16933.75	112151.0
Milk	440.0	5796.27	7380.38	55.0	1533.00	3627.0	7190.25	73498.0
Grocery	440.0	7951.28	9503.16	3.0	2153.00	4755.5	10655.75	92780.0
Frozen	440.0	3071.93	4854.67	25.0	742.25	1526.0	3554.25	60869.0
Detergents_Paper	440.0	2881.49	4767.85	3.0	256.75	816.5	3922.00	40827.0
Delicatessen	440.0	1524.87	2820.11	3.0	408.25	965.5	1820.25	47943.0

Table 1: Wholesale Data Set Summary

Findings from the summary statistics are:-

- First column which is just a reference of Buyer/Spender no which can be ignored for analysis.
- Maximum values of all the column attributes are high as compared to the median value. Hence there seems to be many outliers in this data.
- On checking the median values (50%), it appears that retailers spend more on Fresh products and grocery products as compared to others.

- 75% of 440 retailers spend only 1820 or less annually on Delicatessen. So annual spend of Delicatessen appears to be least among all.

REGION WISE SPENDING:

Pivot Table showing Region wise spending :-

Region	Buyer/Spender	Delicatessen	Detergents_Paper	Fresh	Frozen	Grocery	Milk	Total_Spend
Other	64026	512110	890410	3960577	930492	2495251	1888759	10677599
Lisbon	18095	104327	204136	854833	231026	570037	422454	2386813
Oporto	14899	54506	173311	464721	190132	433274	239144	1555088

Table 2: Region wise spending

We can see from the above Pivot Table that 'Other' region is the highest spender. It seems like the obvious one because as seen in our EDA, 75% data is coming from Other category. Further we can see that lowest spenders are in 'Opporto' region.

Bar Plot of Total Spend Vs Region

Bar Plot showing Region wise spending pattern for Total Spend is as shown below :

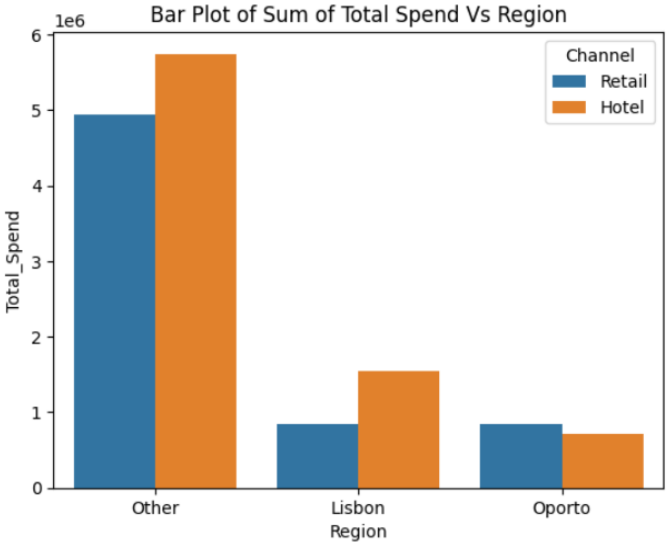


Figure 1: Bar plot of Total Spend vs Region

CHANNEL WISE SPENDING:

Pivot Table showing Channel wise spending :-

Channel	Buyer/Spender	Delicatessen	Detergents_Paper	Fresh	Frozen	Grocery	Milk	Total_Spend
Hotel	71034	421955	235587	4015717	1116979	1180717	1028614	7999569
Retail	25986	248988	1032270	1264414	234671	2317845	1521743	6619931

Table 3: Channel wise spending

We can see that spending is highest under 'Hotel' channel as compared to 'Retail' channel. Visually it can be seen under the Bar plot and Donut Charts. These charts are shown below.

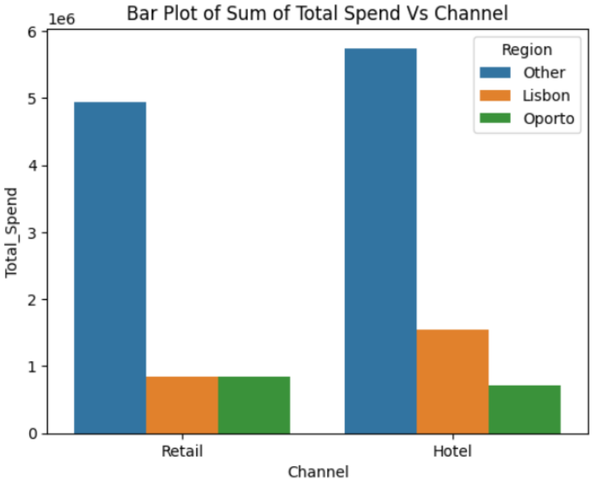
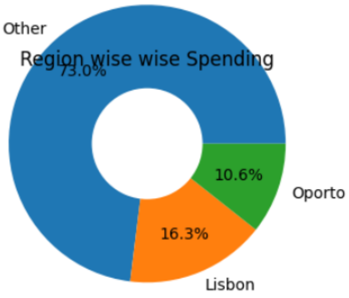


Figure 2: Bar plot of Sum Total Spend vs Channel

Donut Chart showing region wise and channel wise spending

Donut showing spending pattern region wise and channel wise is shown below:-



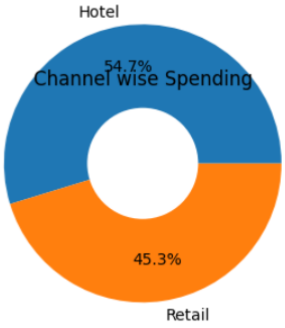


Figure 3: Donut chart of Region wise spending

Figure 4: Donut chart of Channel wise spending

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.

Now for checking the behaviour of varieties across region/channel, I have compared the 5 point summary along with Coefficient of Variation (CV) and Skewness of each variety across 3 regions and 2 channels. Visually I have created a boxplot and swarmplot charts to see the distribution pattern across region and channels. Coding details are present in the notebook file attached along with this report.

Varieties across Regions

Fresh Variety behaviour across all three regions

5 points summary, Coefficient of Variation and Skewness.

	Fresh_Other	Fresh_Oporto	Fresh_Lisbon
Count	316	47	77
Mean	12533.47	9887.68	11101.73
Std	13389.21	8387.9	11557.44
Min	3	3	18
25%	3350.75	2751.5	2806
50%	8752.5	8090	7363
75%	17406.5	14925.5	15218
Max	112151	32717	56083
CV	1.07	0.84	1.03
Skew	2.62	0.98	2.01

Table 4: Summary of Fresh Variety across all three regions

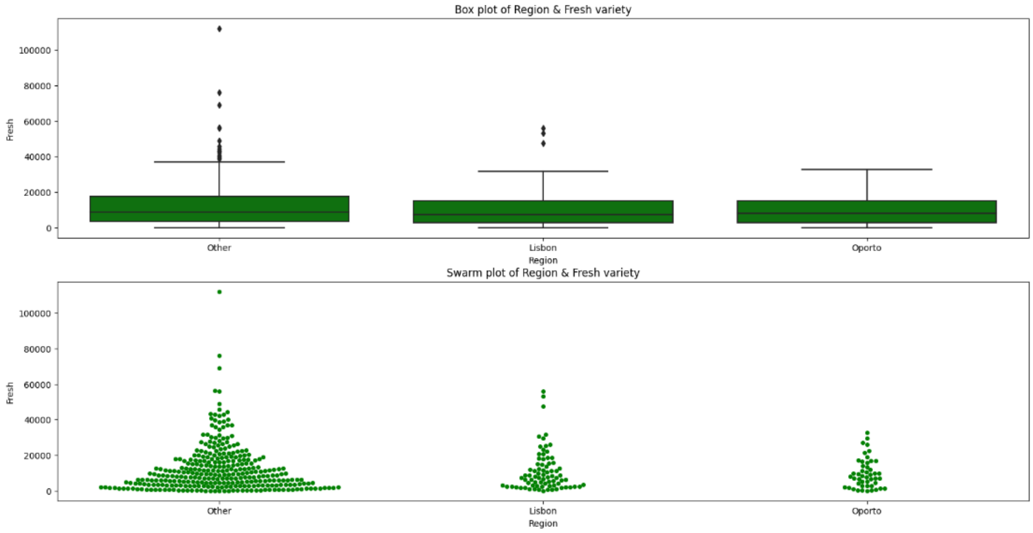


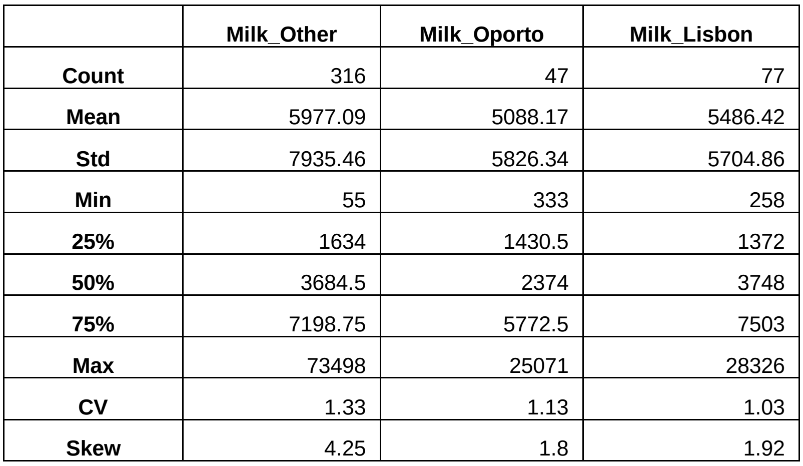
Figure 5: Box and Swarm plot of Fresh Variety across all three regions

Key points as seen in the summary statistics and swarm plot/box plot,

- Data contains more retailers in Other region as compared to Lisbon and Oporto.
- Except Oporto region, 'Other' and 'Lisbon' region Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
- Maximum annual spending in 'Other' region is very high as compared to other regions.
- Annual median spend of Other region is slightly higher(8752) than of Lisbon(7363) and Oporto region(8090)
- Although mean and median value of 'Other' region is highest but its volatility is also high i.e. it is the most inconsistent region for Fresh Variety
- Spread of data looks similar across all regions with distribution being right/positive skewed and 75% of retailers spending less than 17.5K annually across all three regions.
- Footfall is more for buyers under 'Other' Region and majority of buyers(75%) are spending less than around 8K across all three regions.

Milk Variety behaviour across all three regions

5 points summary , Coefficient of Variation and Skewness



	Milk_Other	Milk_Oporto	Milk_Lisbon
Count	316	47	77
Mean	5977.09	5088.17	5486.42
Std	7935.46	5826.34	5704.86
Min	55	333	258
25%	1634	1430.5	1372
50%	3684.5	2374	3748
75%	7198.75	5772.5	7503
Max	73498	25071	28326
CV	1.33	1.13	1.03
Skew	4.25	1.8	1.92

Table 5: Summary of Milk Variety across all three regions

Box Plot and Swarm Plot

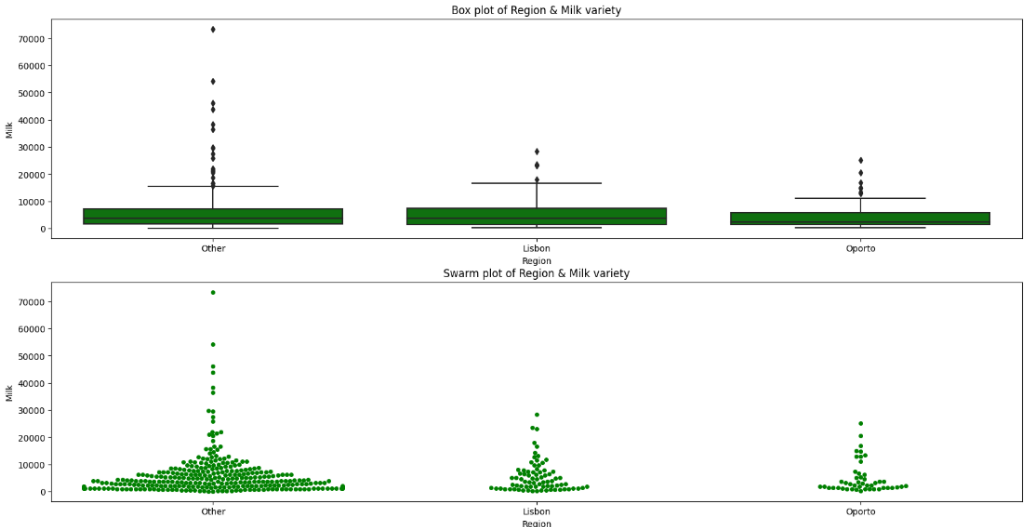


Figure 6: Box and Swarm plot of Milk Variety across all three regions

Key Points as seen in the summary statistics and swarm plot/box plot,

- Data contains more retailers in Other region as compared to Lisbon and Oporto.
- All three regions Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
- Maximum annual spending in 'Other' region is very high as compared to Lisbon/Oporto regions.
- Minimum annual spending in 'Other' region is low as compared to other Lisbon/Oporto regions.
- Annual median spend of Oporto region is slightly lower(2374) than of Lisbon(3748) and Other region(3684)
- Volatility of other region is highest among all i.e. it is the most inconsistent region for Milk Variety
- Spread of data looks similar across all regions with distribution being right/positive skewed and 75% of retailers spending less than 7.5K annually across all three regions.
- Footfall is more for buyers under 'Other' Region and majority of buyers(75%) are spending less than around 3K across all three regions.

Grocery Variety behaviour across all three regions

5 points summary , Coefficient of Variation and Skewness

	Grocery_Other	Grocery_Oporto	Grocery_Lisbon
Count	316	47	77
Mean	7896.36	9218.6	7403.08
Std	9537.29	10842.75	8496.29
Min	3	1330	489
25%	2141.5	2792.5	2046
50%	4732	6114	3838
75%	10559.75	11758.5	9490
Max	92780	67298	39694
CV	1.21	1.16	1.14
Skew	3.84	3.64	2.02

Table 6: Summary of Grocery Variety across all three regions

Box Plot and Swarm Plot

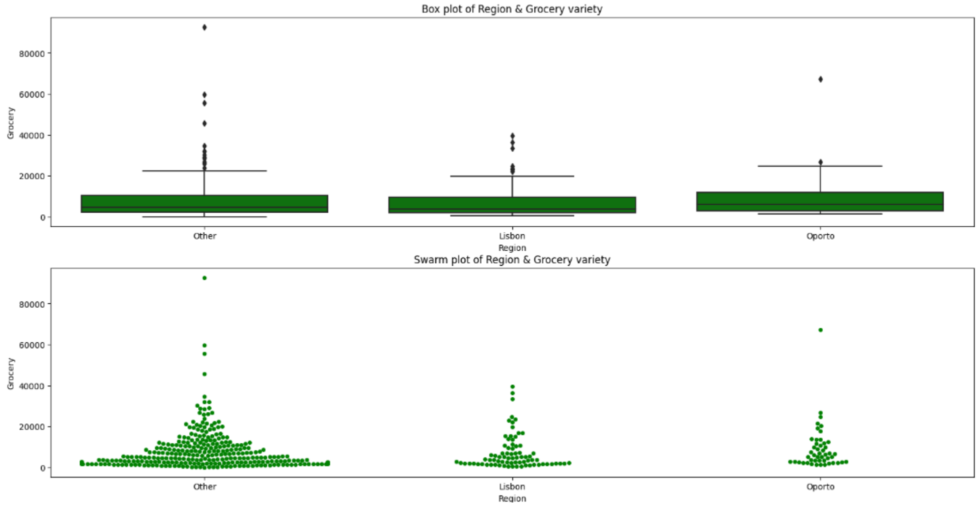


Figure 7: Box and Swarm plot of Grocery Variety across all three regions

Key Points as seen in the summary statistics and swarm plot/box plot,

- Data contains more retailers in Other region as compared to Lisbon and Oporto.
- All three regions Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
- Maximum annual spending in 'Other' region is very high as compared to Lisbon/Oporto regions.
- Minimum annual spending in 'Other' region is low as compared to other Lisbon/Oporto regions.
- Annual median spend of Oporto region is highest(6114) as compared to Lisbon(3838) and Other region(4732)
- Volatility of other region is highest among all i.e. it is the most inconsistent region.

- Spread of data looks similar across all regions with distribution being right/positive skewed and 75% of retailers spending less than 11.7K annually across all three regions.

- Footfall is more for buyers under 'Other' Region and majority of buyers(75%) are spending less than around 10-11K across all three regions.

Frozen Variety behaviour across all three regions

5 points summary , Coefficient of Variation and Skewness

	Frozen_Other	Frozen_Oporto	Frozen_Lisbon
Count	316	47	77
Mean	2944.59	4045.36	3000.34
Std	4260.13	9151.78	3092.14
Min	25	131	61
25%	664.75	811.5	950
50%	1498	1455	1801
75%	3354.75	3272	4324
Max	36534	60869	18711
CV	1.44	2.24	1.02
Skew	3.96	5.49	2.33

Table 7: Summary of Frozen Variety across all three regions

Box Plot and Swarm Plot

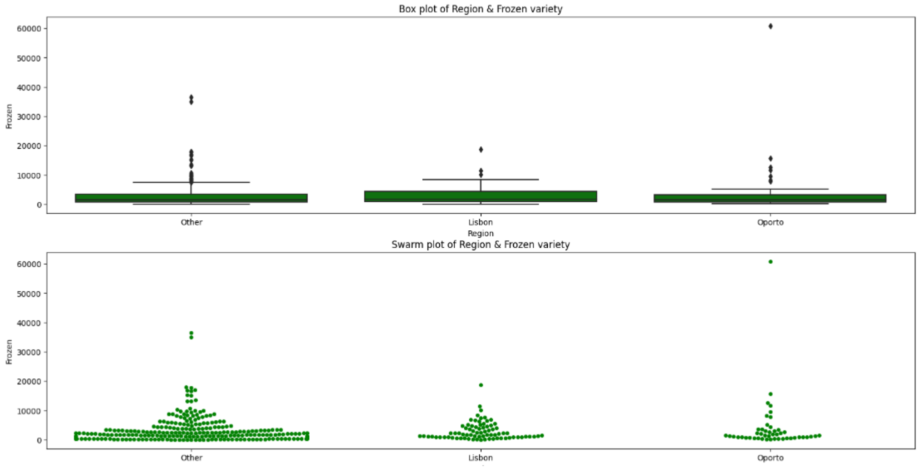


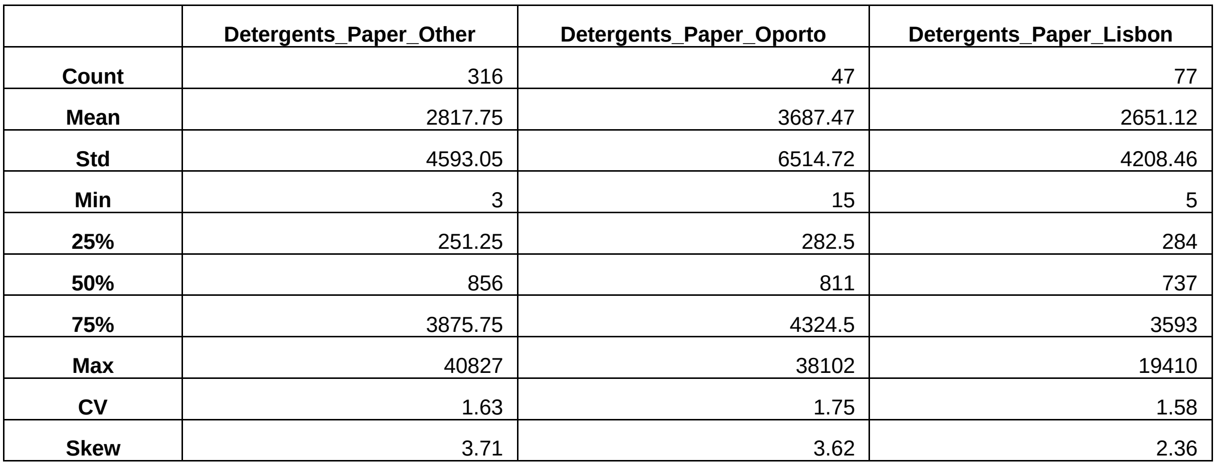
Figure 8: Box and Swarm plot of Frozen Variety across all three regions

Key Points as seen in the summary statistics and swarm plot/box plot,

- Data contains more retailers in Other region as compared to Lisbon and Oporto.
- All three regions Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
- Maximum annual spending in 'Oporto' region is very high as compared to Lisbon/Other regions.
- Minimum annual spending in 'Other' region is low as compared to other Lisbon/Oporto regions.
- Annual median spend of Lisbon region is highest(1801) as compared to Oporto(1455) and Other region(1498)
- Volatility of 'Oporto' region is highest among all i.e. it is the most inconsistent region.
- Spread of data looks similar across all regions with distribution being right/positive skewed and 75% of retailers spending less than 4.3K annually across all three regions.
- Footfall is more for buyers under 'Other' Region and majority of buyers(75%) are spending less than around 3-4K across all three regions.

Detergents_Paper Variety behaviour across all three regions

5 points summary , Coefficient of Variation and Skewness



	Detergents_Paper_Other	Detergents_Paper_Oporto	Detergents_Paper_Lisbon
Count	316	47	77
Mean	2817.75	3687.47	2651.12
Std	4593.05	6514.72	4208.46
Min	3	15	5
25%	251.25	282.5	284
50%	856	811	737
75%	3875.75	4324.5	3593
Max	40827	38102	19410
CV	1.63	1.75	1.58
Skew	3.71	3.62	2.36

Table 8: Summary of Fresh Detergents_Paper Variety across all three regions

Box Plot and Swarm Plot

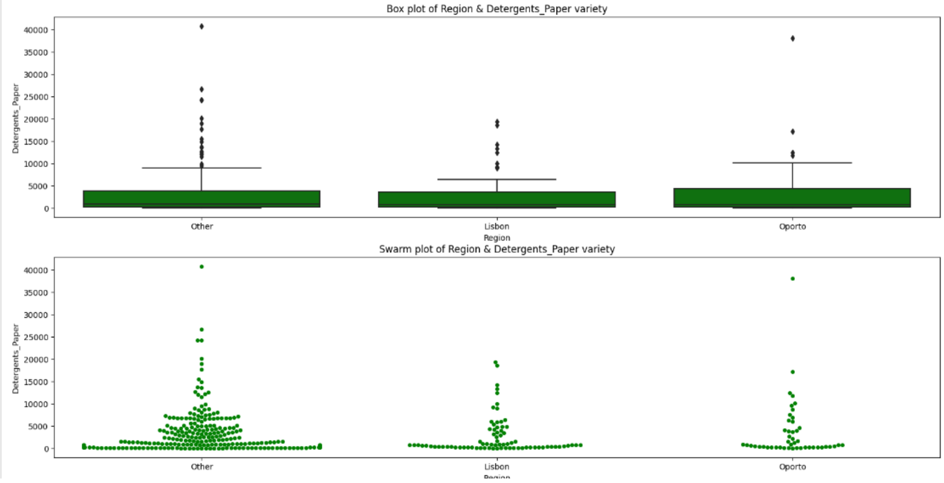


Figure 9: Box and Swarm plot of Detergents_Paper Variety across all three regions

Key Points as seen in the summary statistics and swarm plot/box plot,

- Data contains more retailers in Other region as compared to Lisbon and Oporto.
- All three regions Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
- Maximum annual spending in 'Other' region is very high as compared to Lisbon/Oporto regions.
- Minimum annual spending in 'Other' region is low as compared to other Lisbon/Oporto regions.
- Annual median spend of 'Other' region is highest(856) as compared to Oporto(811) and Lisbon region(737)
- Volatility of 'Oporto' region is highest among all i.e. it is the most inconsistent region.
- Spread of data looks similar across all regions with distribution being right/positive skewed and 75% of retailers spending less than 4.3K annually across all three regions.
- Footfall is more for buyers under 'Other' Region and majority of buyers(75%) are spending less than 3-4K annually across all three regions.

Delicatessen Variety behaviour across all three regions

5 points summary , Coefficient of Variation and Skewness

	Delicatessen_Other	Delicatessen_Oporto	Delicatessen_Lisbon
count	316	47	77
mean	1620.6	1159.7	1354.9
Std	3232.58	1050.74	1345.42
Min	3	51	7
25%	402	540.5	548
50%	994	898	806
75%	1832.75	1538.5	1775
Max	47943	5609	6854
CV	1.99	0.9	0.99
Skew	10.21	2.15	2.05

Table 9: Summary of Delicatessen Variety across all three regions

Box Plot and Swarm Plot

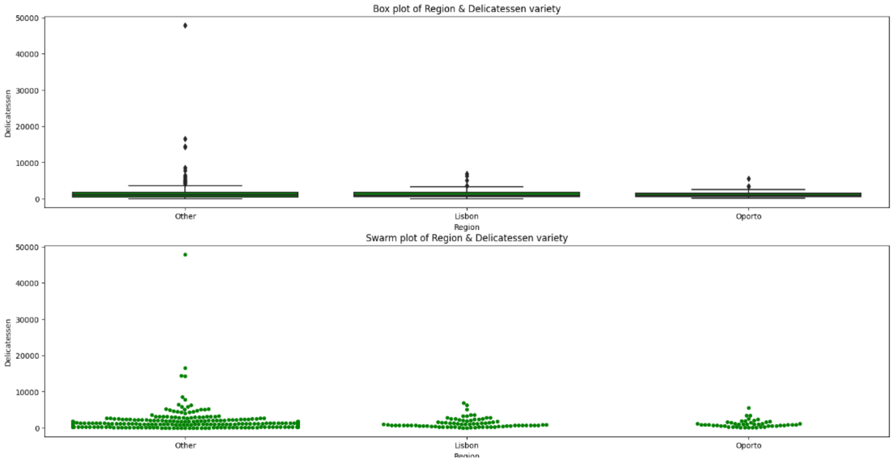


Figure 10: Box and Swarm plot of Delicatessen Variety across all three regions

Key Points as seen in the summary statistics and swarm plot/box plot,

- Data contains more retailers in Other region as compared to Lisbon and Oporto.
- All three regions Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
- Maximum annual spending in 'Other' region is very high as compared to Lisbon/Oporto regions.
- Minimum annual spending in 'Other' region is low as compared to other Lisbon/Oporto regions.
- Annual median spend of 'Other' region is highest(994) as compared to Oporto(898) and Lisbon region(806)
- Volatility of 'Other' region is highest among all i.e. it is the most inconsistent region.
- Spread of data looks similar across all regions with distribution being right/positive skewed and 75% of retailers spending less than 1.83K annually across all three regions.

- Footfall is more for buyers under 'Other' Region and majority of buyers(75%) are spending less than around 1.8K across all three regions.

Varieties across Channels

Fresh Variety behaviour across all 2 channels

5 points summary , Coefficient of Variation and Skewness.

	Fresh_Hotel	Fresh_Retail
Count	298	142
Mean	13475.56	8904.32
Std	13831.69	8987.71
Min	3	18
25%	4070.25	2347.75
50%	9581.5	5993.5
75%	18274.75	12229.75
Max	112151	44466
CV	1.02	1.01
Skew	2.51	1.59

Table 10: Summary of Fresh Variety across all channels

Box Plot and Swarm Plot

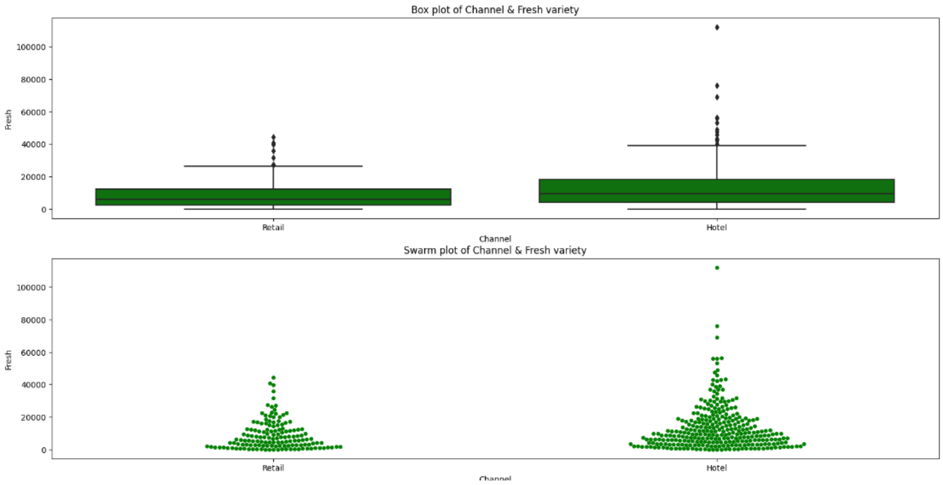


Figure 11: Box and Swarm plot of Fresh Variety across all channels

Key points as seen in the summary statistics and swarm plot/box plot,

- Data contains more retailers in Hotel channel as compared to Retail channel
- Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
- Maximum annual spending in 'Hotel' Channel is high as compared to Retail Channel
- Min annual spending in 'Hotel' Channel is low as compared to Retail Channel
- Annual median spend of Hotel channel is higher(9581) than of Retail channel(5993)
- Volatility of both the channels is similar.

- Spread of data looks similar across both channels with distribution being right/positive skewed and 75% of retailers spending less than 18K annually on Hotel channel and 12.2K on Retail channel respectively.
- Footfall is more for buyers of Fresh variety under Hotel channel and buyers are spending much more money on Milk variety under Hotel channel.

Milk Variety behaviour across all channels

5 points summary , Coefficient of Variation and Skewness

	Milk_Hotel	Milk_Retail
Count	298	142
Mean	3451.72	10716.5
Std	4352.17	9679.63
Min	55	928
25%	1164.5	5938
50%	2157	7812
75%	4029.5	12162.75
Max	43950	73498
CV	1.26	0.9
Skew	4.66	3.41

Table 11: Summary of Milk Variety across all Channels

Box Plot and Swarm Plot

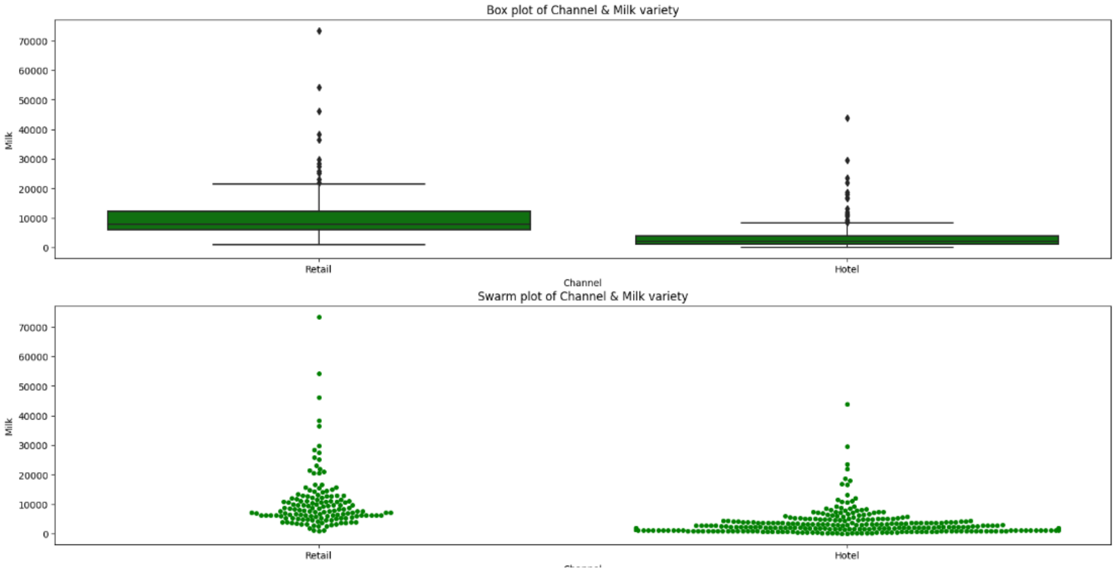


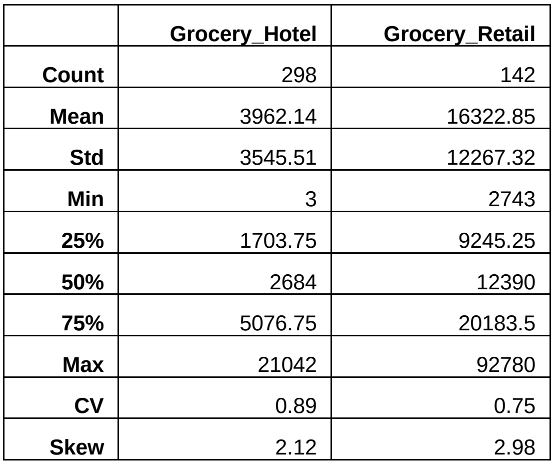
Figure 12: Box and Swarm plot of Milk Variety across all channels

Key Points as seen in the summary statistics and swarm plot/box plot,

- Data contains more retailers in Hotel channel as compared to Retail channel
- Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
- Maximum annual spending in 'Retail' Channel is high as compared to Hotel Channel
- Min annual spending in 'Hotel' Channel is low as compared to Retail Channel
- Annual median spend of Hotel channel(2157) is very low as compared to Retail channel(7812)
- Volatility of Hotel channel is higher than of Retail channel.
- Spread of data looks similar across both channels with distribution being right/positive skewed and 75% of retailers spending less than 4K annually on Hotel channel and 12K on Retail channel respectively.
- Although footfall is more for buyers of Milk variety under Hotel channel but buyers are spending more money on Milk variety under Retail channel.

Grocery Variety behaviour across all channels

5 points summary , Coefficient of Variation and Skewness



	Grocery_Hotel	Grocery_Retail
Count	298	142
Mean	3962.14	16322.85
Std	3545.51	12267.32
Min	3	2743
25%	1703.75	9245.25
50%	2684	12390
75%	5076.75	20183.5
Max	21042	92780
CV	0.89	0.75
Skew	2.12	2.98

Table 12: Summary of Grocery Variety across all Channels

Box Plot and Swarm Plot

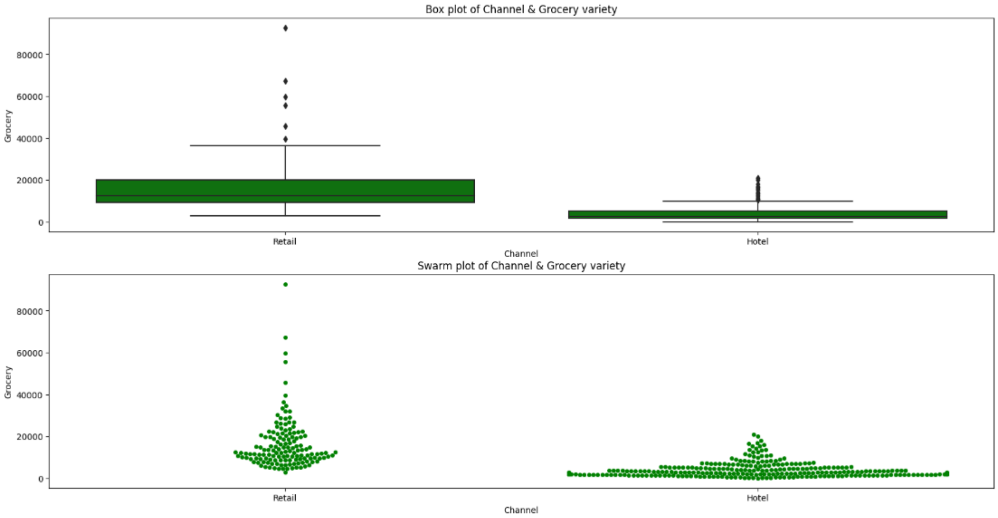


Figure 13: Box and Swarm plot of Grocery Variety across all channels

Key Points as seen in the summary statistics and swarm plot/box plot,

- Data contains more retailers in Hotel channel as compared to Retail channel
- Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
- Maximum annual spending in 'Retail' Channel is high as compared to Hotel Channel
- Min annual spending in 'Hotel' Channel is low as compared to Retail Channel
- Annual median spend of Hotel channel(2684) is very low as compared to Retail channel(12390)
- Volatility of Hotel channel is higher than of Retail channel.
- Spread of data looks similar across both channels with distribution being right/positive skewed and 75% of retailers spending less than 5K annually on Hotel channel and 20K on Retail channel respectively.
- Although Footfall is more for buyers of Grocery variety under Hotel channel but the buyers are spending more money on Grocery variety under Retail channel.

Frozen Variety behaviour across all channels

5 points summary , Coefficient of Variation and Skewness

	Frozen_Hotel	Frozen_Retail
Count	298	142
Mean	3748.25	1652.61
Std	5643.91	1812.8
Min	25	33
25%	830	534.25
50%	2057.5	1081
75%	4558.75	2146.75
Max	60869	11559
CV	1.5	1.09
Skew	5.21	2.53

Table 13: Summary of Frozen Variety across all Channels

Box Plot and Swarm Plot

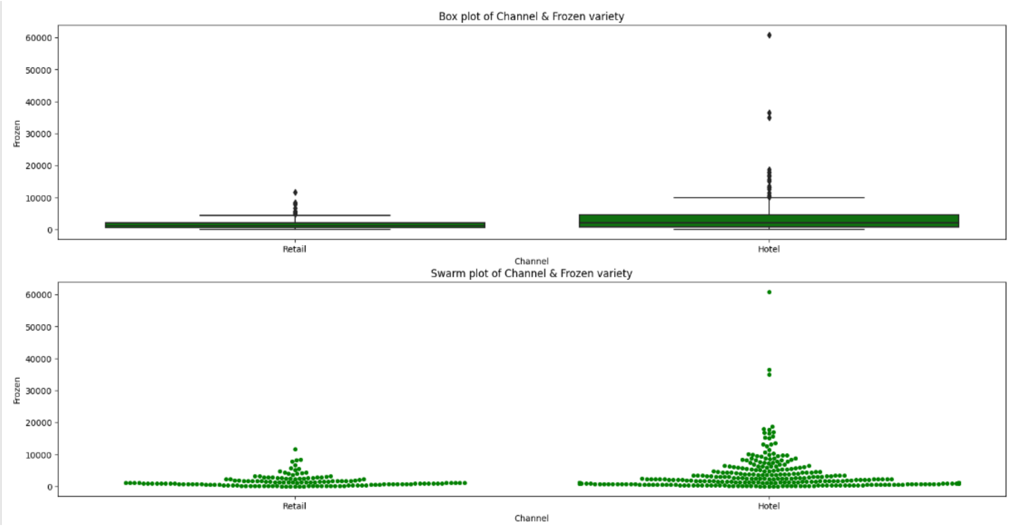


Figure 14: Box and Swarm plot of Frozen Variety across all channels

Key Points as seen in the summary statistics and swarm plot/box plot,

- Data contains more retailers in Hotel channel as compared to Retail channel
- Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
- Maximum annual spending in 'Hotel' Channel is high as compared to Retail Channel
- Min annual spending in 'Hotel' Channel is similar compared to Retail Channel
- Annual median spend of Hotel channel(2057) is high as compared to Retail channel(1081)
- Volatility of Hotel channel is higher than of Retail channel.
- Spread of data looks similar across both channels with distribution being right/positive skewed and 75% of retailers spending less than 4.5K annually on Hotel channel and 2.1K on Retail channel respectively.
- The Footfall is more for buyers under Hotel channel and the buyers are spending more money on Frozen variety under Hotel channel as compared to Retail channel

Detergents_Paper Variety behaviour across all channels

5 points summary , Coefficient of Variation and Skewness

	Detergents_Paper_Hotel	Detergents_Paper_Retail
Count	298	142
Mean	790.56	7269.51
Std	1104.09	6291.09
Min	3	332
25%	183.25	3683.5
50%	385.5	5614.5
75%	899.5	8662.5
Max	6907	40827
CV	1.39	0.86
Skew	2.86	2.61

Table 14: Summary of Detergents_Paper Variety across all Channels

Box Plot and Swarm Plot

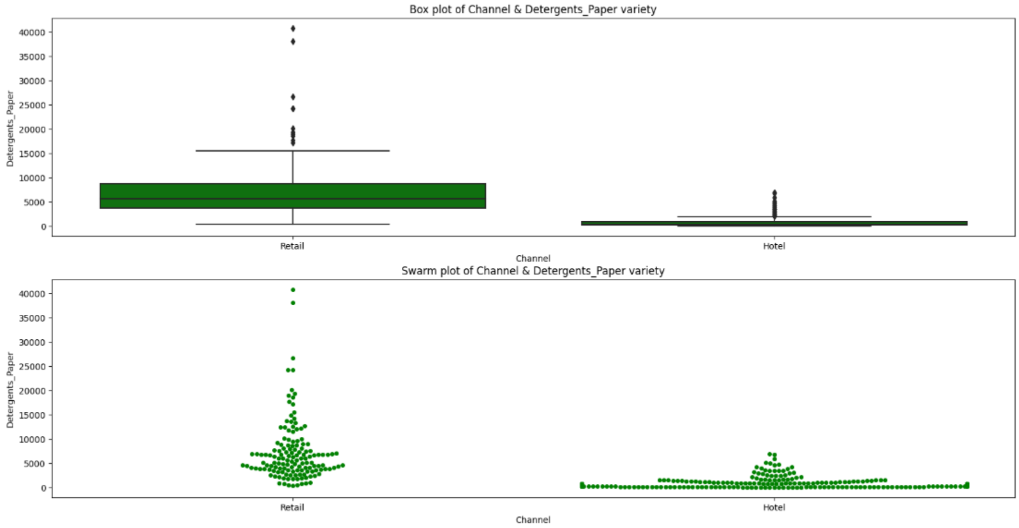


Figure 15: Box and Swarm plot of Detergents_Paper Variety across all channels

Key Points as seen in the summary statistics and swarm plot/box plot,

- Data contains more retailers in Hotel channel as compared to Retail channel
- Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
- Maximum annual spending in 'Retail' Channel is high as compared to Hotel Channel
- Min annual spending in 'Hotel' Channel is low as compared to Retail Channel
- Annual median spend of Hotel channel(385) is very low as compared to Retail channel(5614)
- Volatility of Hotel channel is higher than of Retail channel.

- Spread of data looks similar across both channels with distribution being right/positive skewed and 75% of retailers spending less than 0.9K annually on Hotel channel and 8.6K on Retail channel respectively.
- Although Footfall is more for buyers under Hotel channel but the buyers are spending more money on Detergent_Paper variety under Retail channel.

Delicatessen Variety behaviour across all channels

5 points summary , Coefficient of Variation and Skewness

	Delicatessen_Hotel	Delicatessen_Retail
Count	298	142
Mean	1415.96	1753.44
Std	3147.43	1953.8
Min	3	3
25%	379	566.75
50%	821	1350
75%	1548	2156
Max	47943	16523
CV	2.22	1.11
Skew	11.52	3.77

Table 15: Summary of Delicatessen Variety across all Channels

Box Plot and Swarm Plot

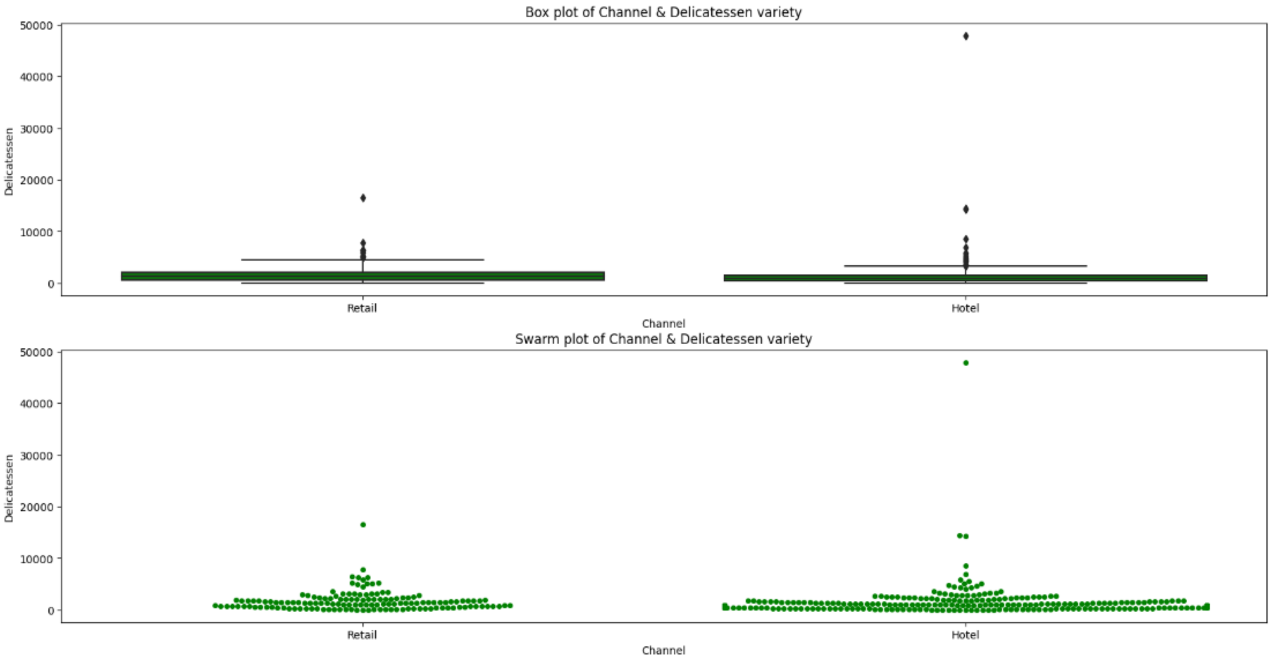


Figure 16: Box and Swarm plot of Delicatessen Variety across all channels

Key Points as seen in the summary statistics and swarm plot/box plot,

- Data contains more retailers in Hotel channel as compared to Retail channel
- Data contains outliers as seen in box plot. Hence we are using median values for comparison instead of mean.
- Maximum annual spending in 'Hotel' Channel is high as compared to Retail Channel
- Min annual spending in 'Hotel' Channel is similar to Retail Channel
- Annual median spend of Hotel channel(821) is very low as compared to Retail channel(1350)
- Volatility of Hotel channel is higher than of Retail channel.
- Spread of data looks similar across both channels with distribution being right/positive skewed and 75% of retailers spending less than 1.5K annually on Hotel channel and 2.1K on Retail channel respectively.
- Although Footfall is more for buyers under Hotel channel but the buyers are spending more money on Delicateessen variety under Retail channel.

Conclusions

Spending pattern of all 6 items across region appears to be similar

Spending pattern of 6 items across channels gives us following points

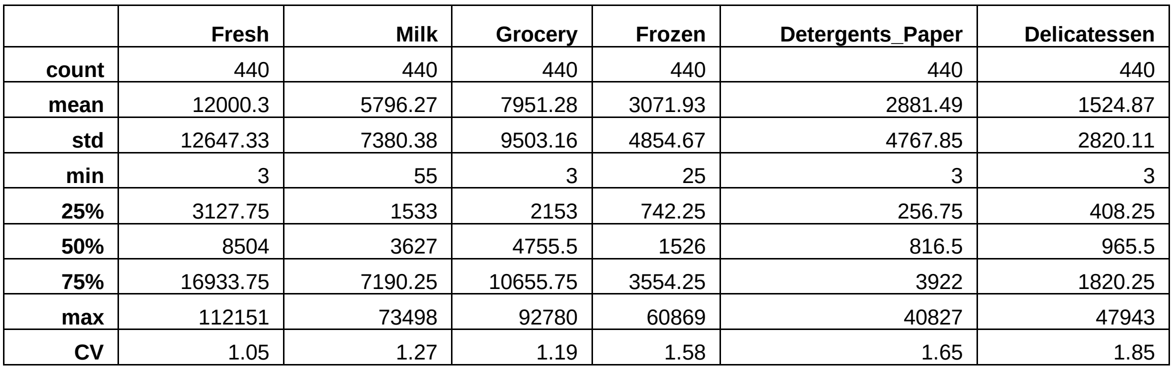
- Fresh and Frozen variety is purchased more via Hotel Channel (annual median around 1.5 to 2 times) as compared to Retail channel

- Milk, Grocery is purchased more via Retail Channel(annual median around 4 to 5 times) as compared to Hotel Channel

- Detergents_Paper is purchased very less via Hotel Channel as compared to Retail Chan nel (annual median around 14 times)

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

For checking the most inconsistent behaviour, I have used CV (Coefficient of Variation) parameter for comparison. Code which is calculating the CV of all the 6 items can be seen in the notebook file attached. I have attached here is the summary report showing CV, mean & SD



	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	440	440	440	440	440	440
mean	12000.3	5796.27	7951.28	3071.93	2881.49	1524.87
std	12647.33	7380.38	9503.16	4854.67	4767.85	2820.11
min	3	55	3	25	3	3
25%	3127.75	1533	2153	742.25	256.75	408.25
50%	8504	3627	4755.5	1526	816.5	965.5
75%	16933.75	7190.25	10655.75	3554.25	3922	1820.25
max	112151	73498	92780	60869	40827	47943
CV	1.05	1.27	1.19	1.58	1.65	1.85

Table 16: summary of Groceries after adding Coefficient of Variation

As seen above from the CV values, item which is showing most inconsistent behaviour is 'Delicatessen' and item which is showing lowest inconsistent behaviour is 'Fresh'

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

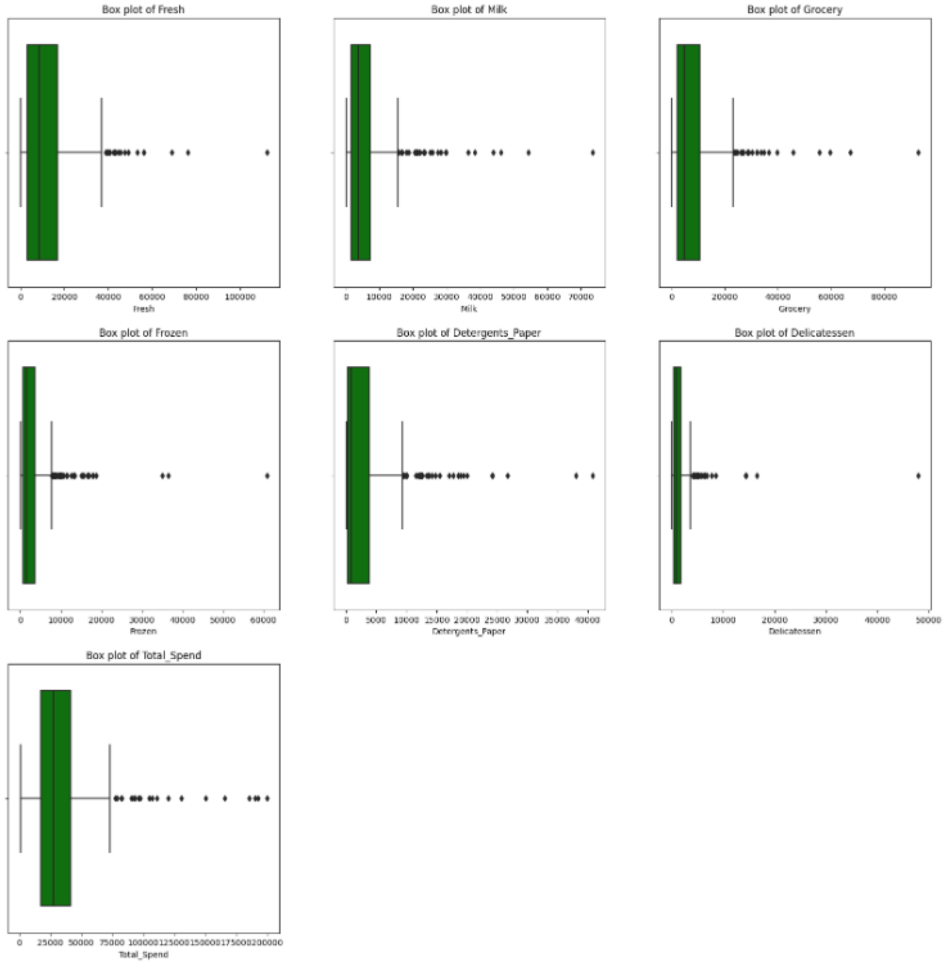


Figure 17: Box plot of numerical variables

As evident from above box plots, all attributes have outliers. All outliers are on maximum side i.e. there are few retailers which are spending much more than the majority of the retailers. Data appears to be right skewed.

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

I have created a Pivot table on region and channel, and following table is the output

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

Table 17: Summary of data across all region and channel

- In this table, we could see that Delicatessen total spent is very less as compared to other products. We could easily spot the highest selling varieties i.e. Fresh, Grocery and Milk.
- Then I created another Pivot table based on Channel and aggregated on median (not used mean as data contains outlier).

	Buyer/Spender	Delicatessen	Detergents_Paper	Fresh	Frozen	Grocery	Milk	Total_Spend
Channel								
Hotel	241.5	821	385.5	9581.5	2057.5	2684	2157	21254.5
Retail	166.5	1350	5614.5	5993.5	1081	12390	7812	37139

Table 18: Pivot table based on Channel using aggrerator

- In this table, we could see that some varieties are purchased more via Hotel channel (Fresh & Frozen) and some are purchased more via Retail channel (Milk & Grocery). I was not able to see any noticeable pattern in region wise pivot table.
- Then on checking the correlation we could see that there is very strong correlation between Grocery and Detergent_Paper , Grocery and Milk. Also there exist a minor negative correlation between Frozen and Detergent_Paper , Detergent_paper & Fresh.

Heatmap and pairplot graphs are shown below :-

Heatmap

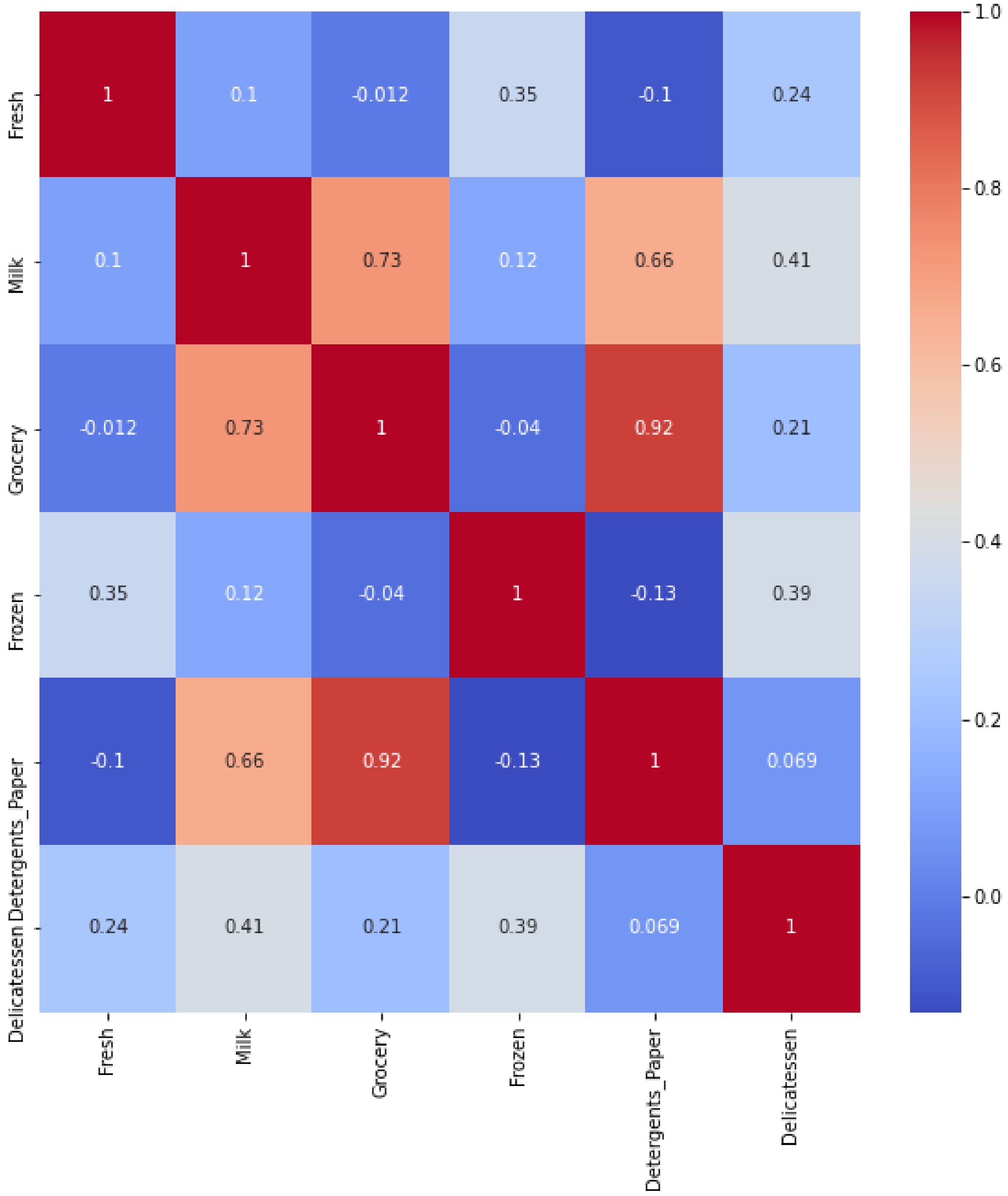


Figure 18: Heatmap of numerical variables

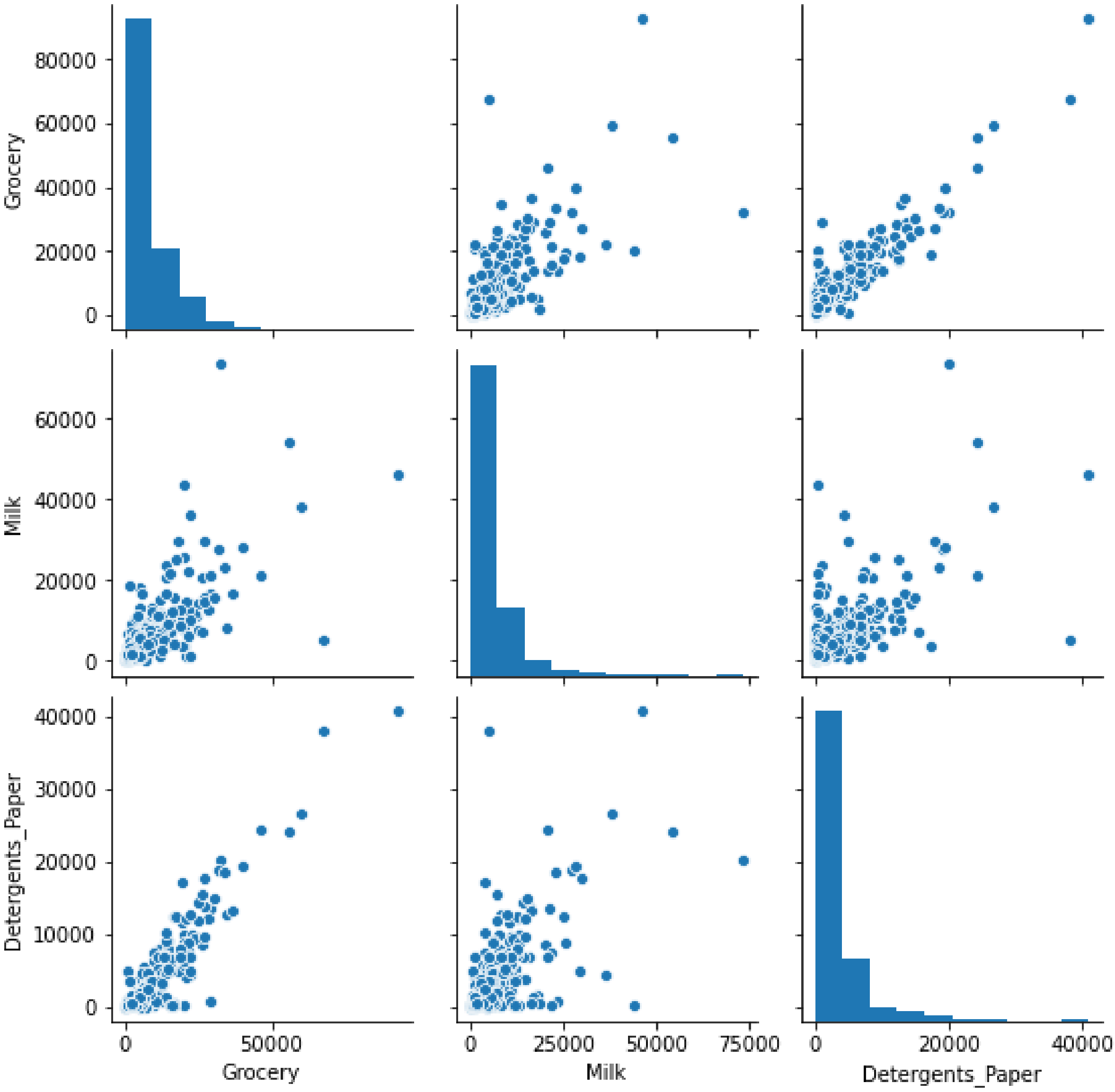


Figure 19: Pair plot of Grocery, Milk and Detergents_Paper

- Pairplot showing strong positive correlation

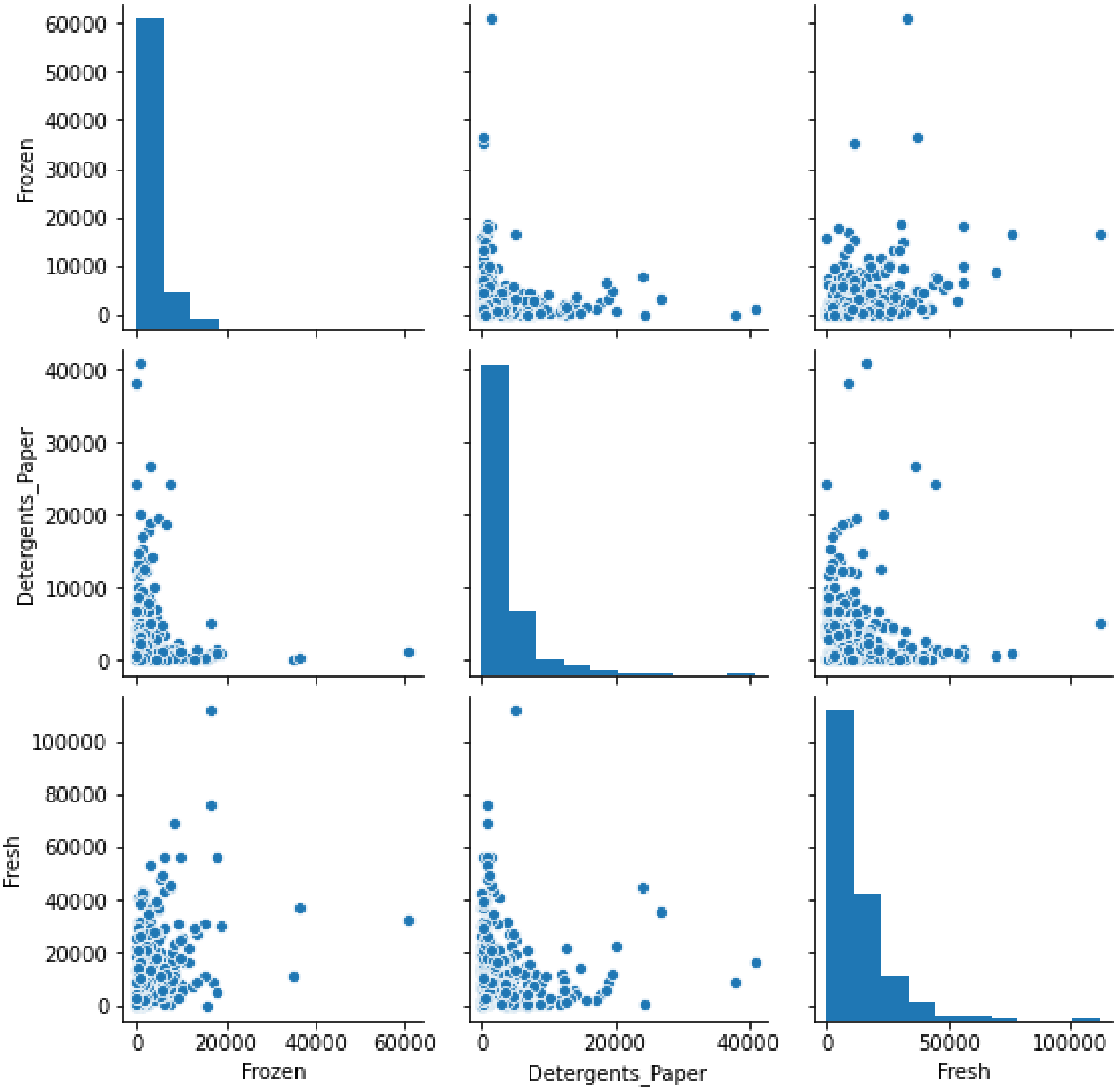


Figure 20: Pair plot of Frozen, Detergents_Paper & Fresh

- Pairplot showing negative correlation

Thus after analysis , my recommendation for business is :-

- As seen in summary and Pivot tables, Buyers are spending much more on Fresh Products, Grocery and Milk varieties. It means Fresh Products, Grocery and Milk varieties are high selling items and business should ensure proper supply of these varieties.

- As seen in summary, 75% of buyers are spending less than 4k annually on Frozen and Detergents Paper variety. Delicateessen variety is the least consumed among all the 6 varieties with around 2K annual spend by 75% of buyers.

- Delicateessen variety share is only 4.5% of total. May be price of these items could be very high causing low sales. So Business either need to drop this variety or perhaps needs put in more sales promotion/marketing efforts to boost its sale.
- Fresh and Frozen variety is purchased more via Hotel Channel (annual median around 1.5 to 2 times) as compared to Retail channel

- Milk, Grocery is purchased more via Retail Channel (annual median around 4 to 5 times) as compared to Hotel Channel

- Detergents_Paper is purchased very less via Hotel Channel as compared to Retail Channel (annual median around 14 times)

- Thus Business should focus on sale of more Fresh and Frozen varieties under Hotel Channel and more Milk, Grocery and Detergent_Paper varieties under Retail Channel.

- Since there is a strong correlation between Grocery & Milk, Grocery & Detergent_Paper business could give combo offers for these varieties like BOGO scheme etc for sale promotion.

- Also clubbing of negative correlated varieties like Detergent_Paper & Frozen, Detergent_Paper & Fresh for any combo pack scheme should be avoided by Business.

Problem 2:

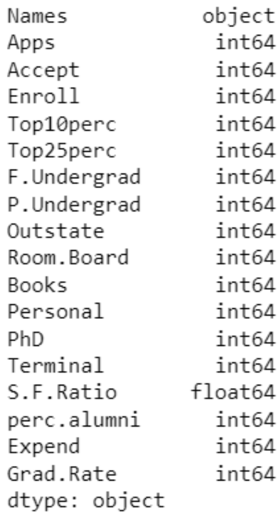
The dataset Education - Post 12th Standard.csv contains information on various colleges. You are expected to do a Principal Component Analysis for this case study according to the instructions given. The data dictionary of the 'Education - Post 12th Standard.csv' can be found in the following file: Data Dictionary.xlsx.

- **Perform Exploratory Data Analysis [Univariate, Bivariate, and Multivariate analysis to be performed]. What insight do you draw from the EDA?**

Inference of the Data set

Shape:
We have 777 rows and 18 columns in our Data set.

Data Type:



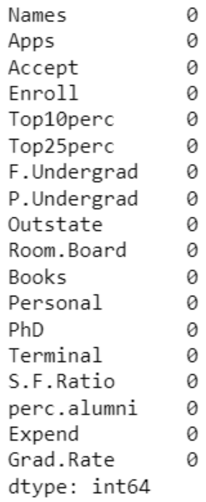
```
Names           object
Apps             int64
Accept           int64
Enroll           int64
Top10perc        int64
Top25perc        int64
F.Undergrad      int64
P.Undergrad      int64
Outstate         int64
Room.Board       int64
Books            int64
Personal         int64
PhD              int64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

- All the columns seem to be integer or float values except Names.
- The Names column alone is a categorical value.

Duplicates

There are no duplicates found in our data set.

Null Check



```
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
```

- There is no Null data present in our data set.

Uni Variate Analysis:

Apps:

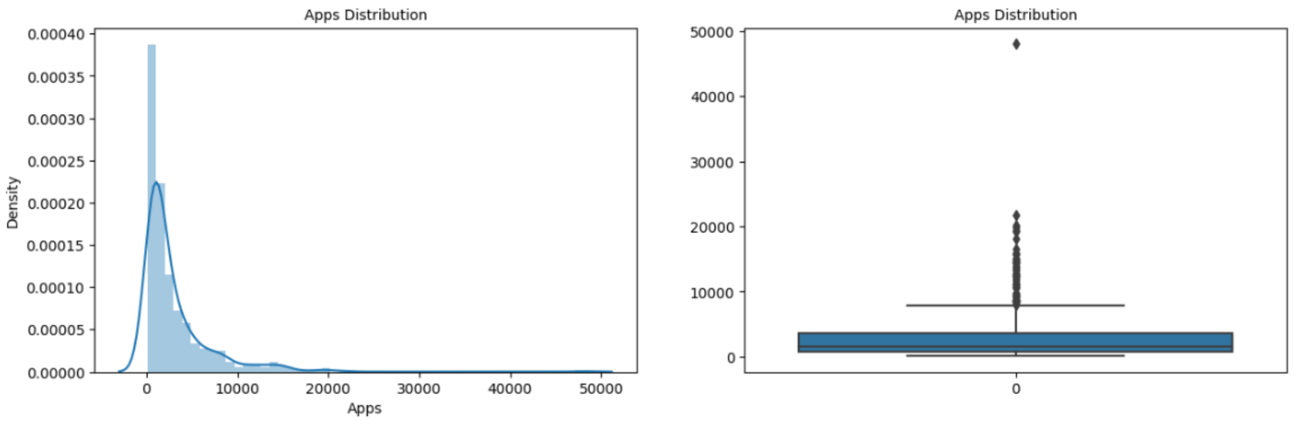


Figure 21: Dist and Box plot of Apps

- The Box plot of Apps variable seems to have outliers, the distribution of the data is skewed.
- We could also understand that each collage or university offers application in the range 3000 to 5000.
- The max applications seem to be around 50000.

Accept:

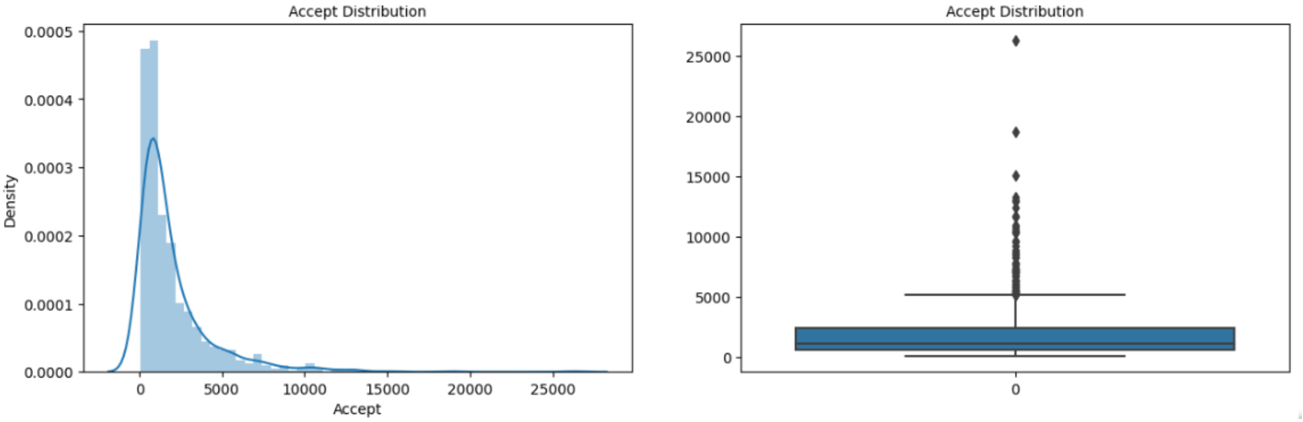


Figure 22: Dist and Box plot of Accept

- The accept variable seems to have outliers.
- The dist plot shows us most applications accepted from each university are in the range from 70 to 1500.
- The accept variable seems to be positively skewed.

Enrol

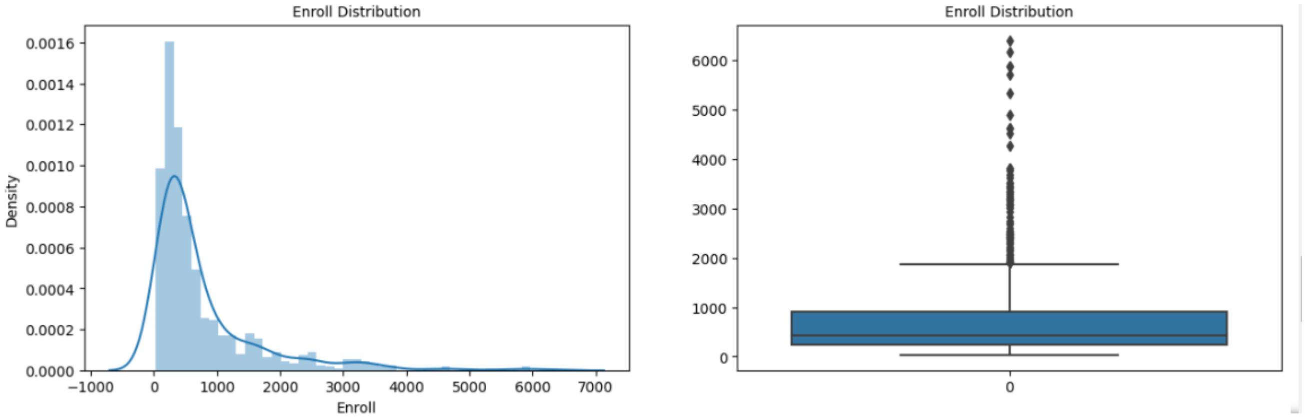


Figure 23: Dist and Box plot of Enroll

- The Box plot of the enroll variable also has outliers.
- The distribution of the data is positively skewed.
- From the dist plot we can understand most of the colleges have enrolled students in the range of 200 to 500 students.

Top 10 Percent:

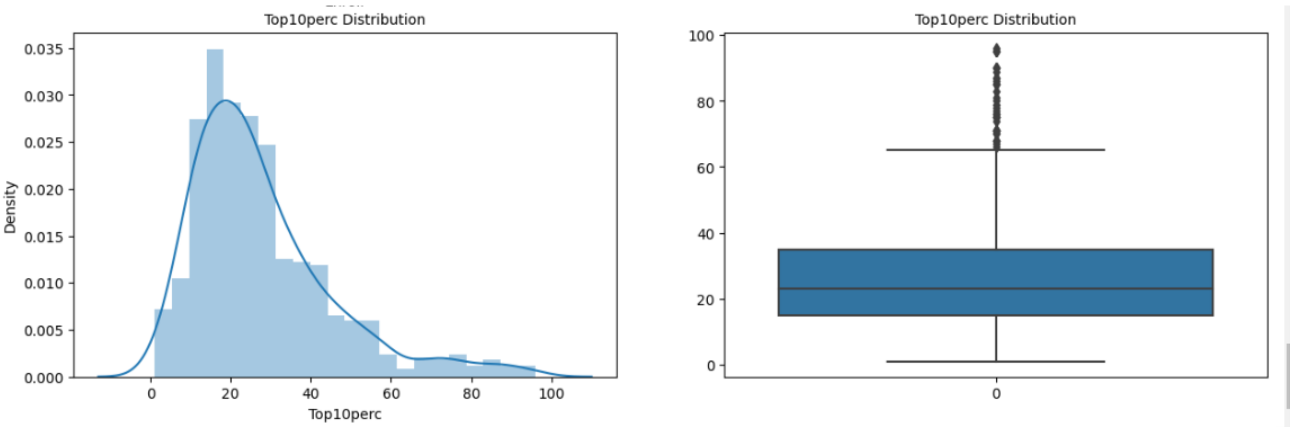


Figure 24: Dist and Box plot of Top 10 Percent

- The box plot of the students from top 10 percentage of higher secondary class seems to have outliers.
- The distribution seems to be positively skewed.
- There is good amount of intake about 30 to 50 students from top 10 percentage of higher secondary class.

Top 25 Percent:

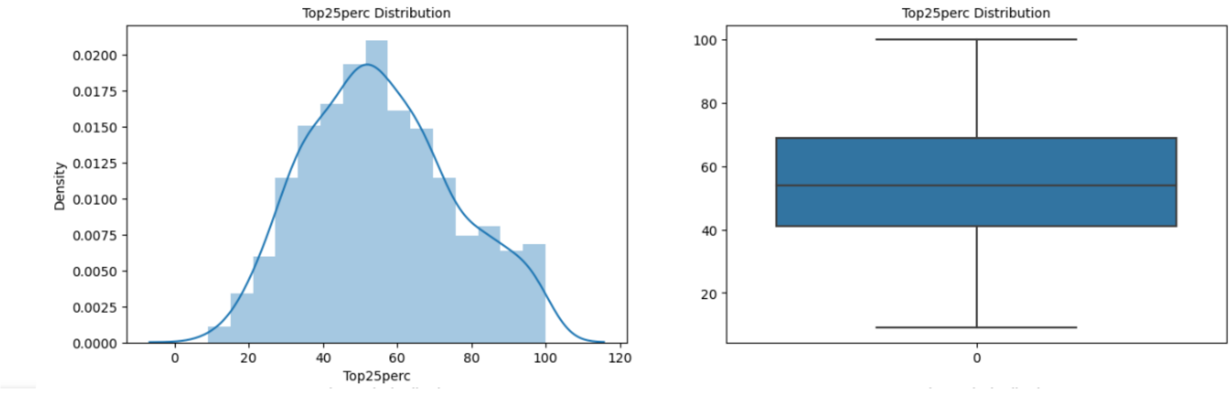


Figure 25: Dist and Box plot of Top 25 Percent

- The box plot for the top 25% has no outliers.
- The distribution is almost normally distributed.
- Majority of the students are from top 25% of higher secondary class.

Full Time Undergraduate:

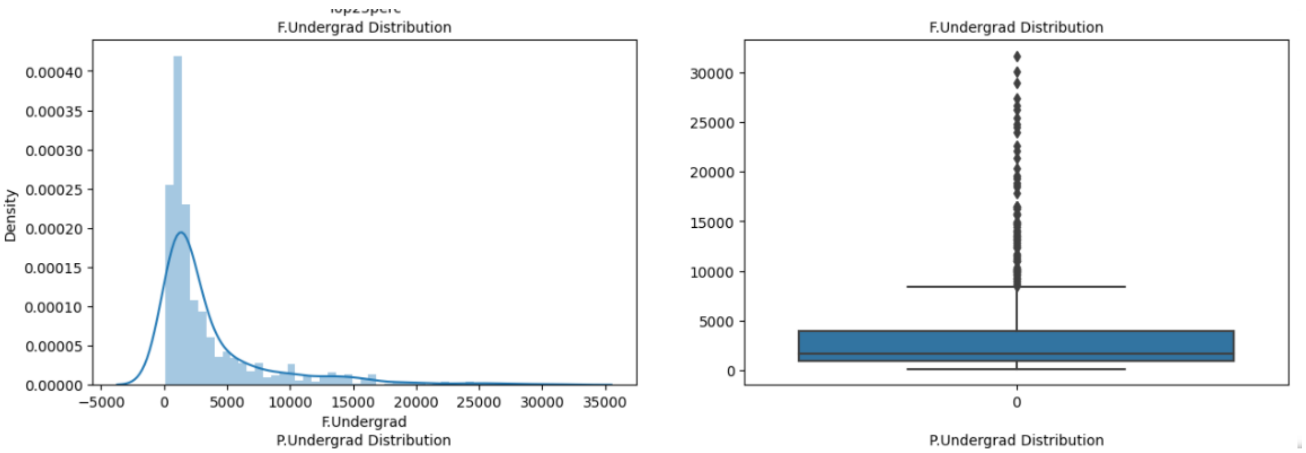


Figure 26: Dist and Box plot of Full Time Undergraduate

- The box plot of full time graduate has outliers.
- The distribution of the data is positively skewed.
- In the range about 3000 to 5000 they are full time graduates studying all the university.

Part Time Graduate

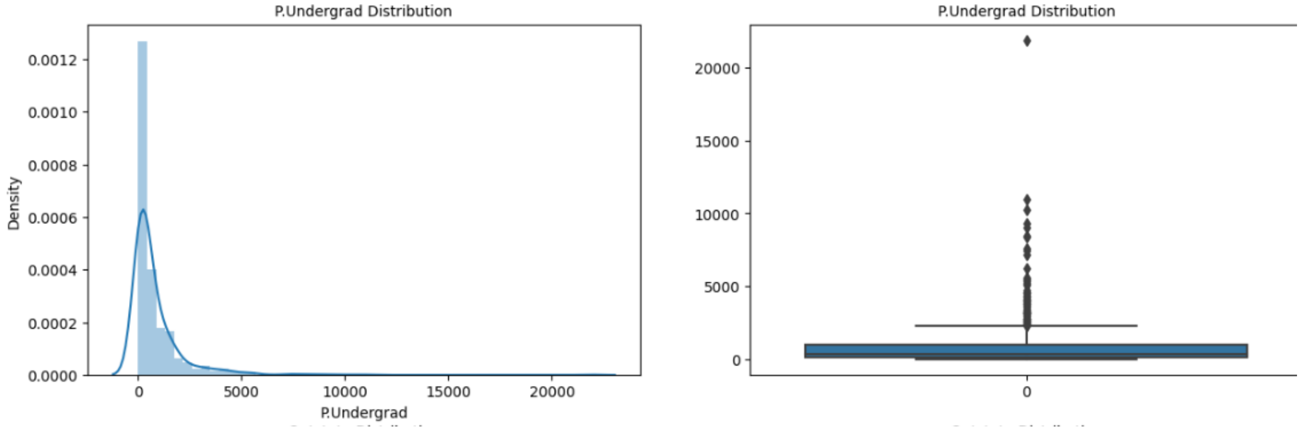


Figure 27: Dist and Box plot of Part Time Graduate

- The box plot of part time graduate has outliers.
- The distribution of the data is positively skewed.
- In the range about 1000 to 3000 they are part time graduates studying all the university.

OutState:

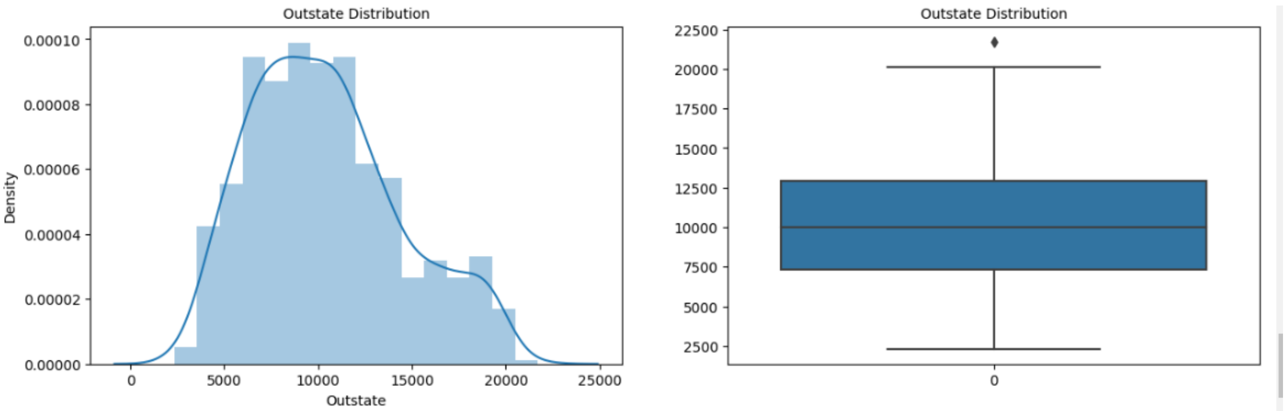


Figure 28: Dist and Box plot of Outstate

- The Box plot has only one outlier.
- The distribution is almost normally distributed.

ROOM BOARD

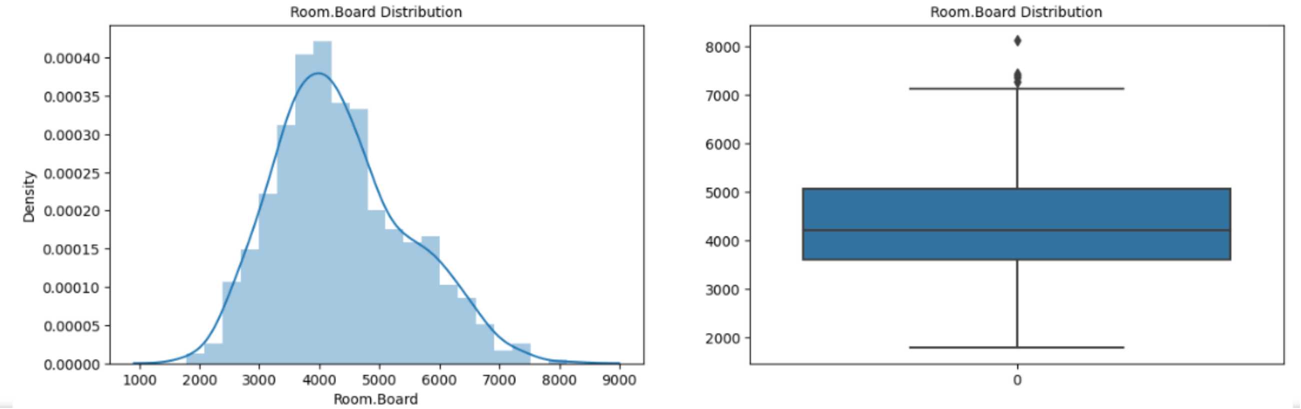


Figure 29: Dist and Box plot of Room Board

- The Room board has few outliers.
- The distribution is normally distributed.

BOOKS

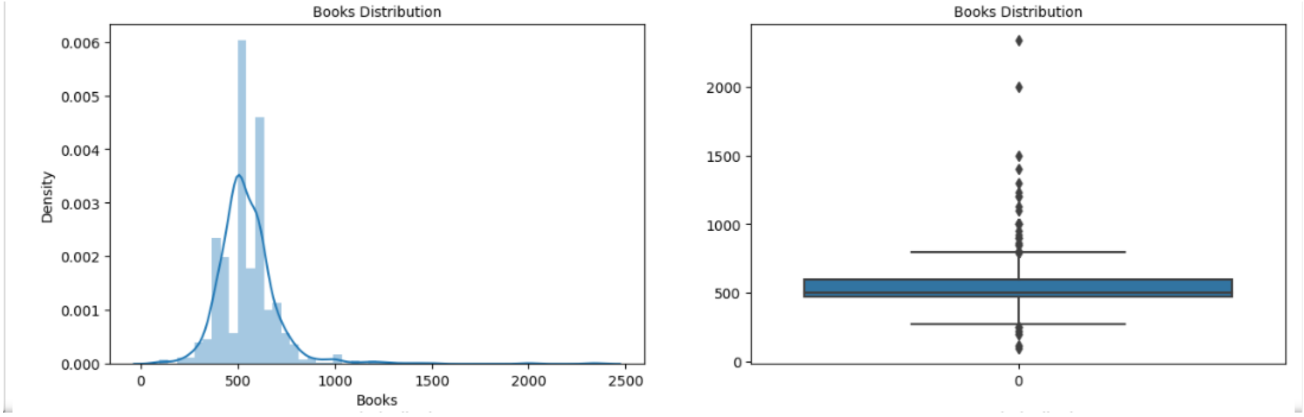


Figure 30: Dist and Box plot of Books

- The box plot of books has outliers.
- The distribution seems to be bimodal. The cost of books per student seems to be in the range of 100 to 500.

PERSONAL

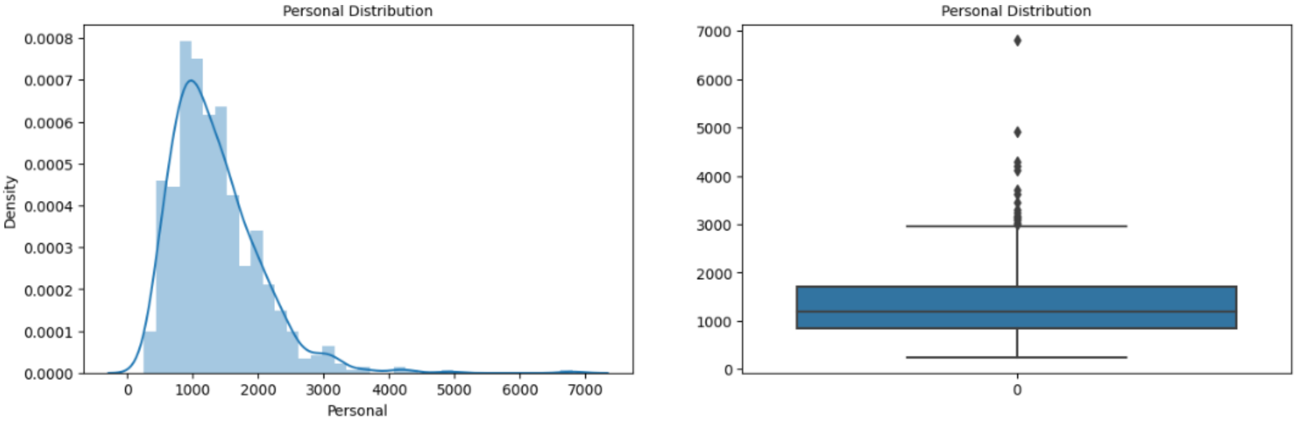


Figure 31: Dist and Box plot of Personal

- The box plot of personal expense has outliers.
- Some student’s personal expense is way bigger than the rest of the students.
- The distribution seems to be positively skewed.

PhD

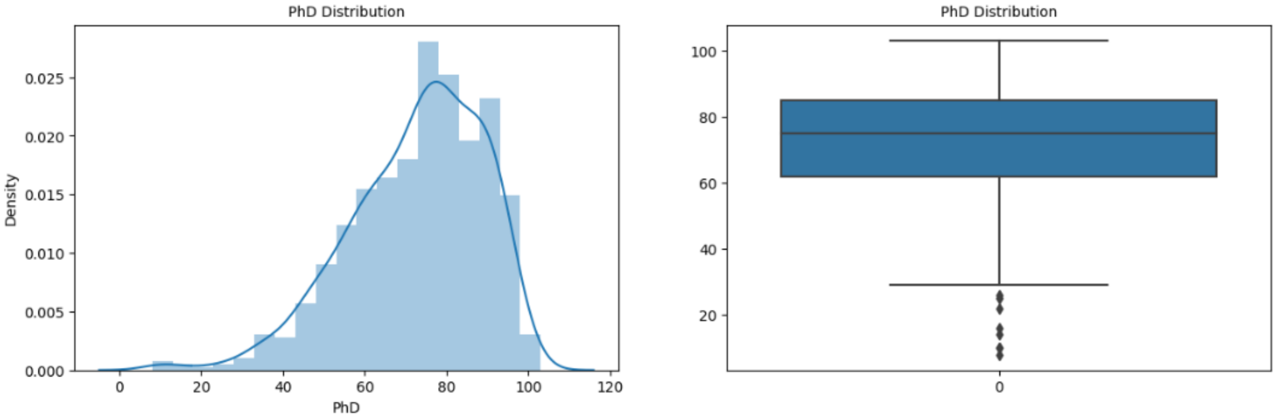


Figure 32: Dist and Box plot of PHD

- The box plot of PHD has outliers.
- The distribution seems to be negatively skewed.

TERMINAL

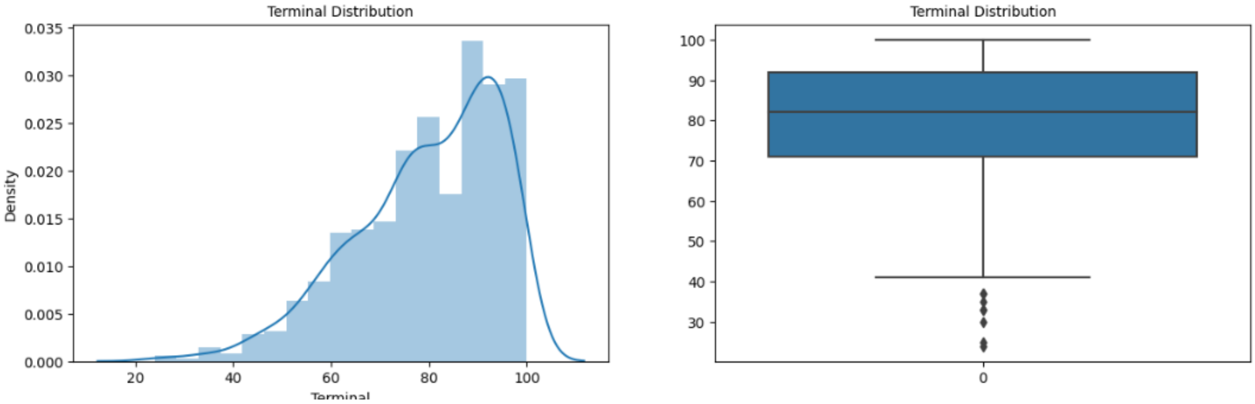


Figure 33: Dist and Box plot of TERMINAL

- The box plot of terminal seems to have outliers in the dataset.
- The distribution also seems to be negatively skewed.

SF RATIO

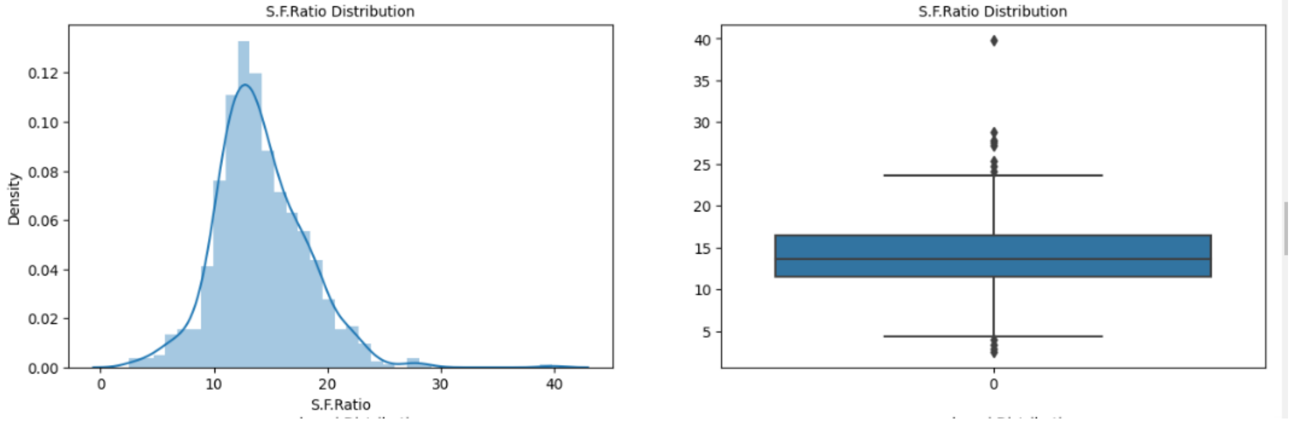


Figure 34: Dist and Box plot of SF Ratio

- The SF ration variable also has outliers in the dataset.
- The distribution is almost normally distributed.
- The student faculty ratio is almost same in all the university and colleges.

PERCENT ALUMNI

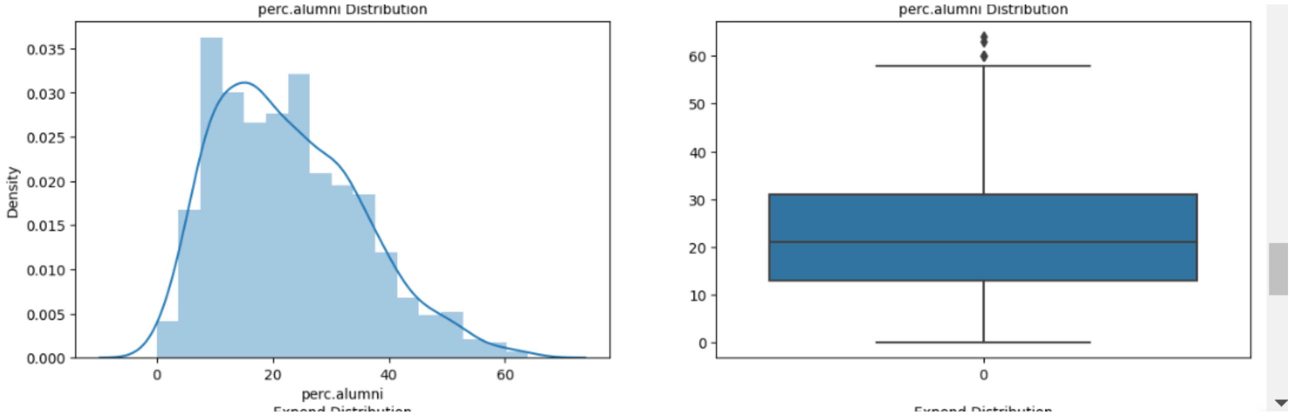


Figure 35: Dist and Box plot of Percentage of Alumni

- The percentage of alumni box plot seems to have outliers in the dataset.
- The distribution is almost normally distributed.

EXPENDITURE

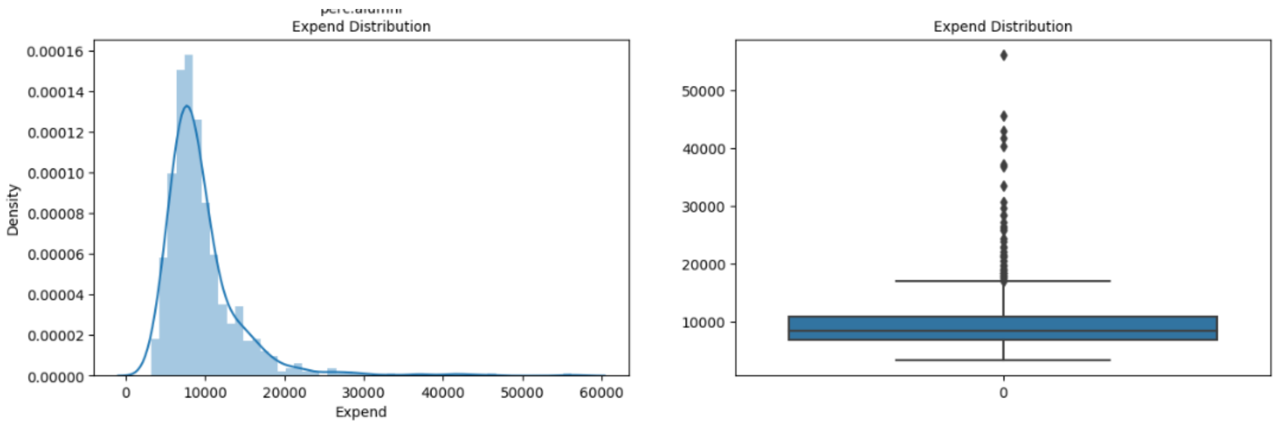


Figure 36: Dist and Box plot of Expenditure

- The expenditure variable also has outliers in the dataset.
- The distribution of the expenditure is positively skewed.

GRAD RATE

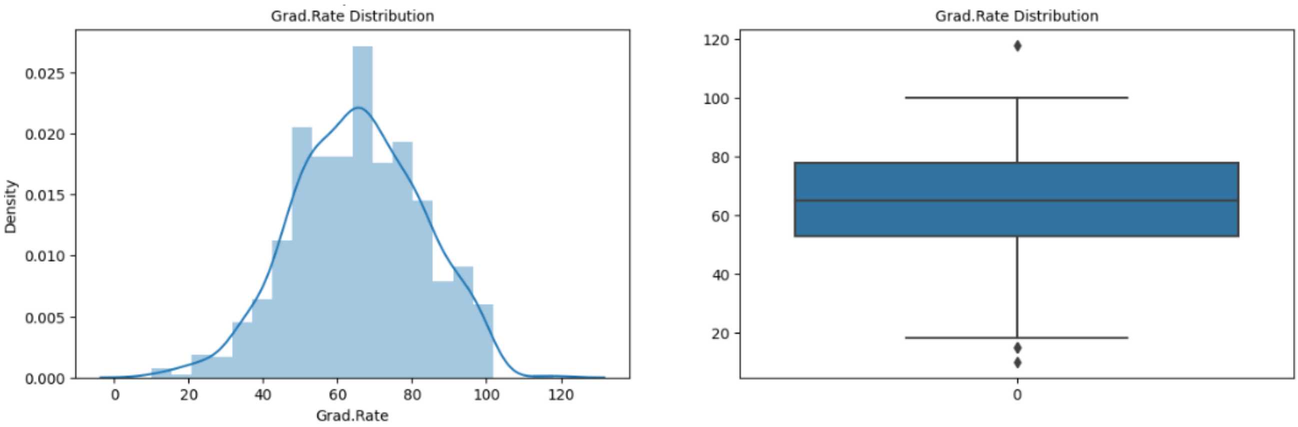


Figure 37: Dist and Box plot of Grad Rage

- The graduation rate among the students in all the university above 60%.
- The boxplot of the graduation rate has outliers in the data set.
- The distribution is normally distributed.

Multivariate Analysis:

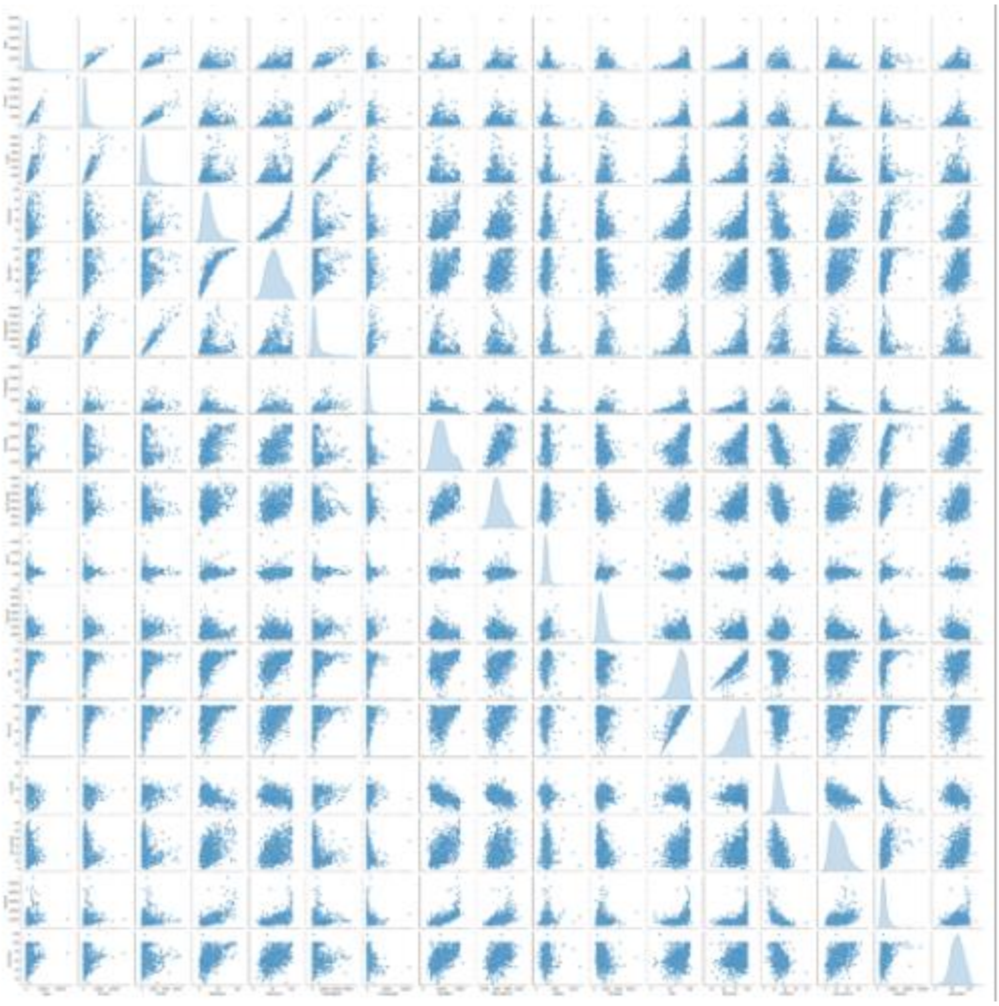


Figure 38: Pair plot of numerical variables

- The pair plot helps us to understand the relationship between all the numerical values in the data set.
- On comparing all the variables with eat other we could understand the patterns or trends in the dataset.

Heat Map:

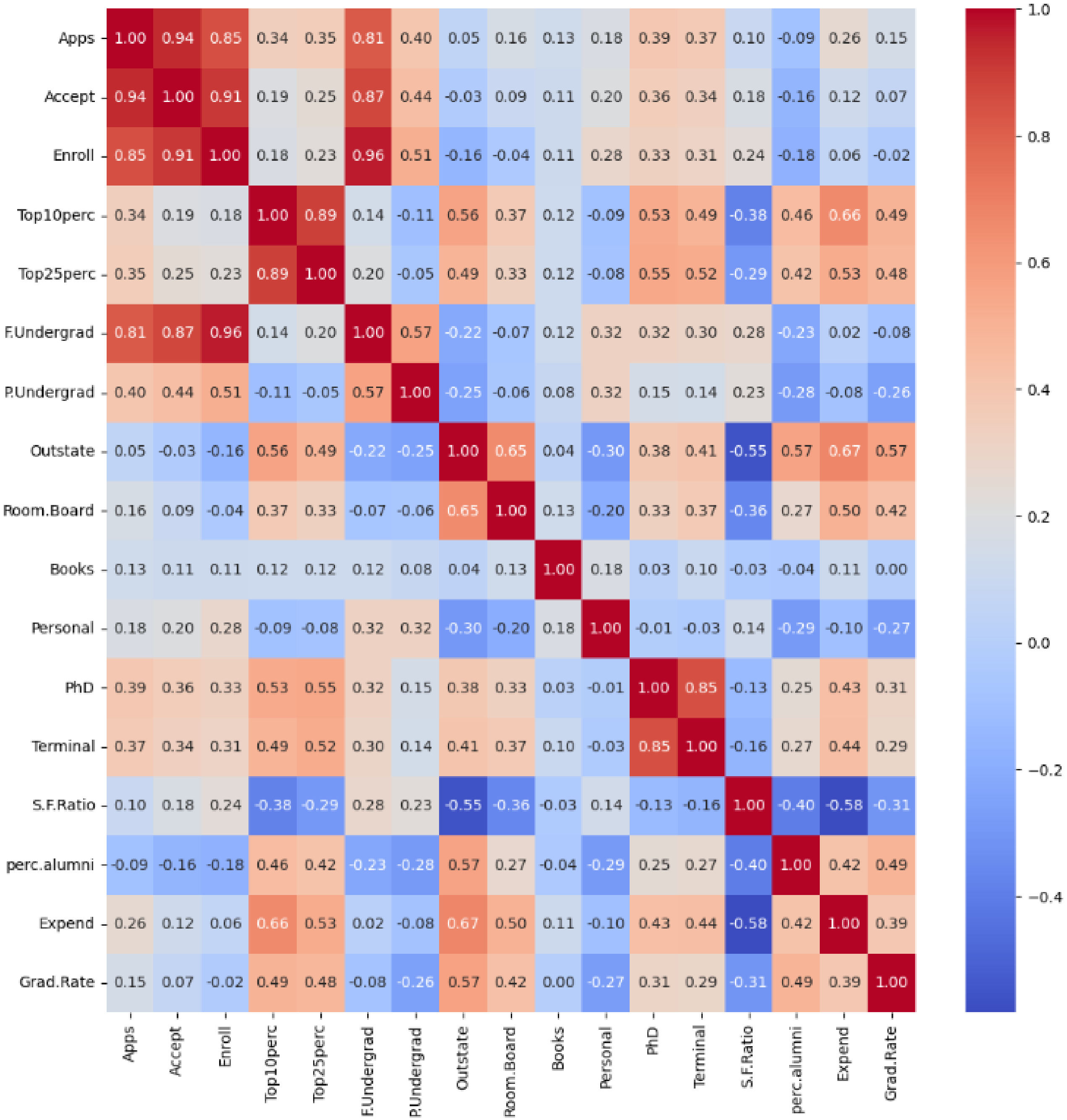


Figure 39: Heat map numerical variables

- This Heat map gives us the correlation between two numerical values.
- We could understand the application variable is highly positively correlated with application accepted, students enrolled and full-time graduates.
- So this relationship gives insights on when student submits the application it is accepted and student is enrolled as fulltime graduate.
- We can find negative correlation between application and percentage of alumni. This indicates us not all students are part of alumni of their college or university.
- The Application with top 10, 25 of higher secondary class, outstate, room board, books, personal, PhD, terminal, S.F ratio, expenditure and Graduation ratio are positively correlated.