Great Learning

SMDM

PROJECT REPORT

HARI HARAN

12th March, 2023

CONTENTS	PAGE

`

Great Learning

PROBLEM 1 :

Wholesale Customers Analysis (Download Data) 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.

Performing EDA for the given data:

```
<class 'pandas.core.frame.DataFrame'>
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)
memory usage: 31.1+ KB
```

```
Buyer/Spender       0
Channel             0
Region              0
Fresh               0
Milk                0
Grocery             0
Frozen              0
Detergents_Paper    0
Delicatessen        0
dtype: int64
```

- No missing values found or Null values found

Table - 1.1 Data Info and Checking for null values.

The above data shows there are 2 data types ,i.e., 7 - integer types and 2-object types. We can ignore the Buyer/Spender numerical value for further analysis as it won't have any effect on the current data. Also there are no missing values.

```
Other     316
Lisbon     77
Oporto     47
Name: Region, dtype: int64
```

```
Hotel     298
Retail    142
Name: Channel, dtype: int64
```

Further Analysis shows the following -

- Channel consist of Hotels and Retail Business channels on which
  - Hotel = 298
  - Retail = 142

- Region consist of namely Lisbon, Oporto and Other areas in which the business is done.

- The data consists of 440 channels through which the wholesale distributors conduct its business in three different regions Lisbon, Oporto and Other by supplying to the Hotel and Retail channels.

	count	mean	std	min	25%	50%	75%	max
Buyer/Spender	440.0	220.500	127.161	1.0	110.75	220.5	330.25	440.0
Fresh	440.0	12000.298	12647.329	3.0	3127.75	8504.0	16933.75	112151.0
Milk	440.0	5796.266	7380.377	55.0	1533.00	3627.0	7190.25	73498.0
Grocery	440.0	7951.277	9503.163	3.0	2153.00	4755.5	10655.75	92780.0
Frozen	440.0	3071.932	4854.673	25.0	742.25	1526.0	3554.25	60869.0
Detergents_Paper	440.0	2881.493	4767.854	3.0	256.75	816.5	3922.00	40827.0
Delicatessen	440.0	1524.870	2820.106	3.0	408.25	965.5	1820.25	47943.0

Table - 1.2 Data Description

```
Buyer/Spender          0.000000
Fresh                  2.561323
Milk                   4.053755
Grocery                3.587429
Frozen                 5.907986
Detergents_Paper       3.631851
Delicatessen          11.151586
dtype: float64
```

Table - 1.3 - Skewness in the data

From the above analysis we can conclude the following -

- Maximum values are higher than the median shows presence of outliers.
- Consumers buy **Fresh products** and grocery ***more*** compare to other products.
- **Delicatessens** are the ***least*** purchased item.
- Data given is skewed (ignoring the number of Buyer/Spender, as it won't change) and for the other columns we can visualize through Boxplot method.

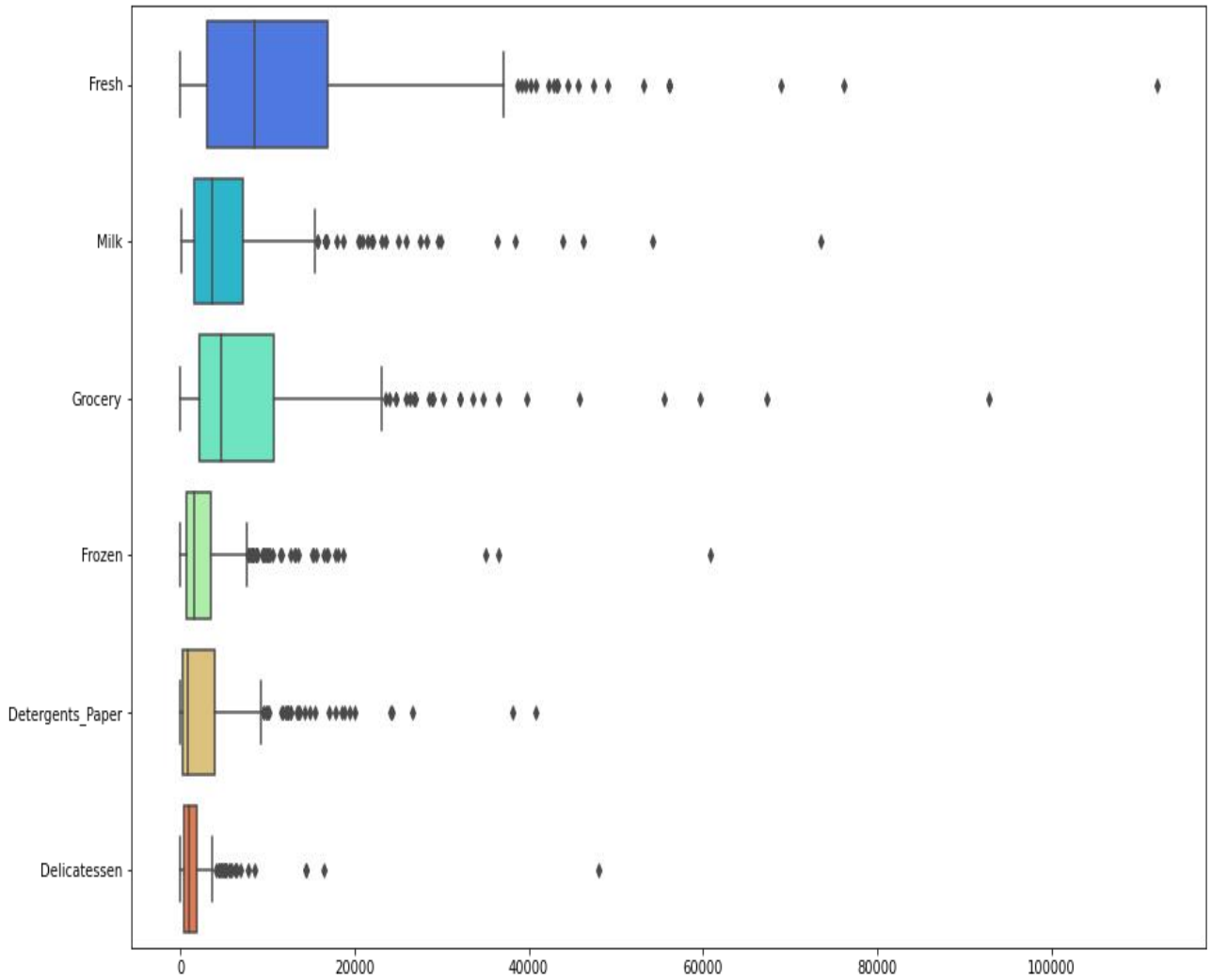


Fig.1.1 - Boxplot showing ouliers in spending on different items.

- The above plot shows presence of outliers which means and is ***right skewed.***
- This also shows that there are some buyer/spenders who are spending more than others.

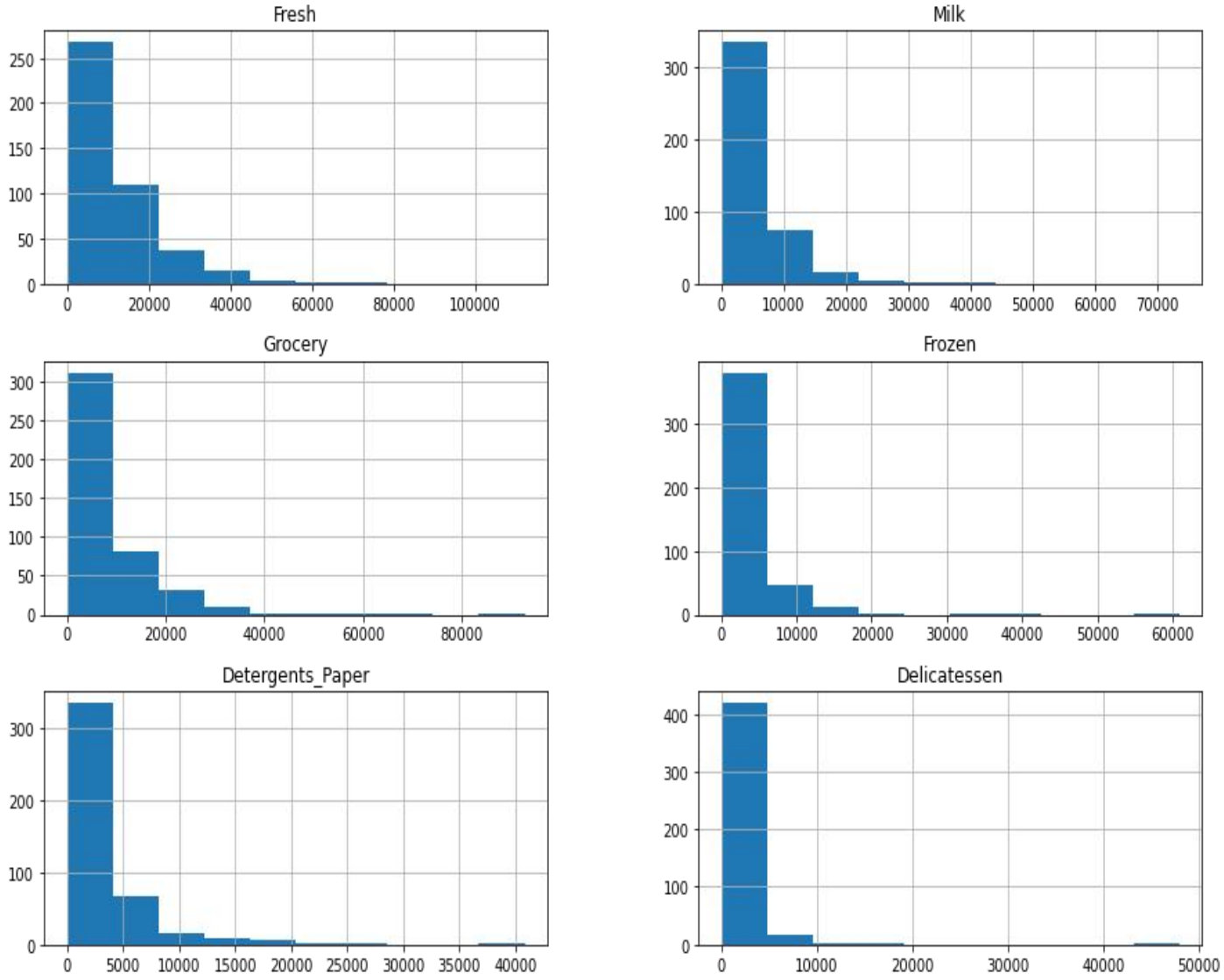


Fig.1.2 - Histogram showing spending on different items.

1.1.2 Which Region and which Channel spent the most?
1.1.3 Which Region and which Channel spent the least?

1. To understand this we need to calculate the total spending of all each channel and region
2. adding total expense column to the data frame = "Total_exp"

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

Tabel 1.3 - Showing 'Total_exp' Column

Using groupby and sum() method,

```
Channel
Hotel     7999569
Retail    6619931
```

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

In the above results obtained from analysis we have:

- The highest spending **Channel** is **Hotel** and the lowest is **Retail**.
- The highest spending **Region** is **Other** and the lowest is **Oporto**.

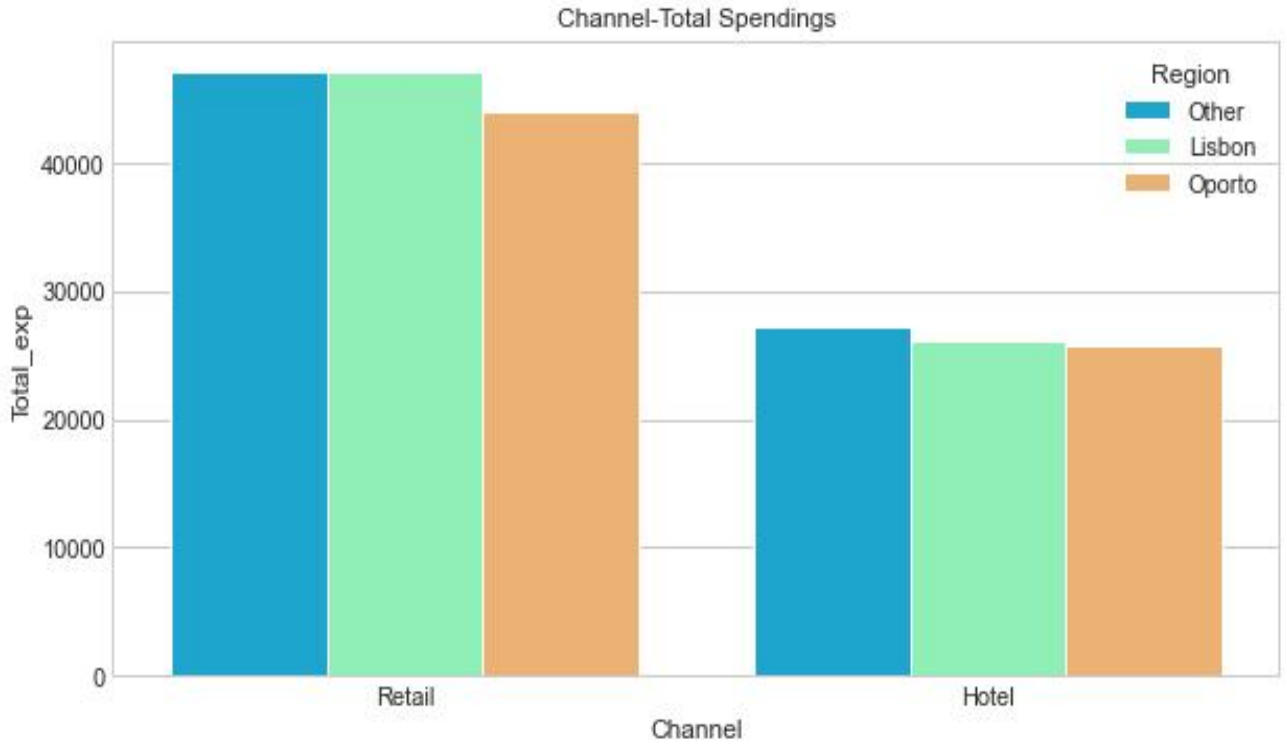


Fig1.3 - Channel wise spending in three Regions

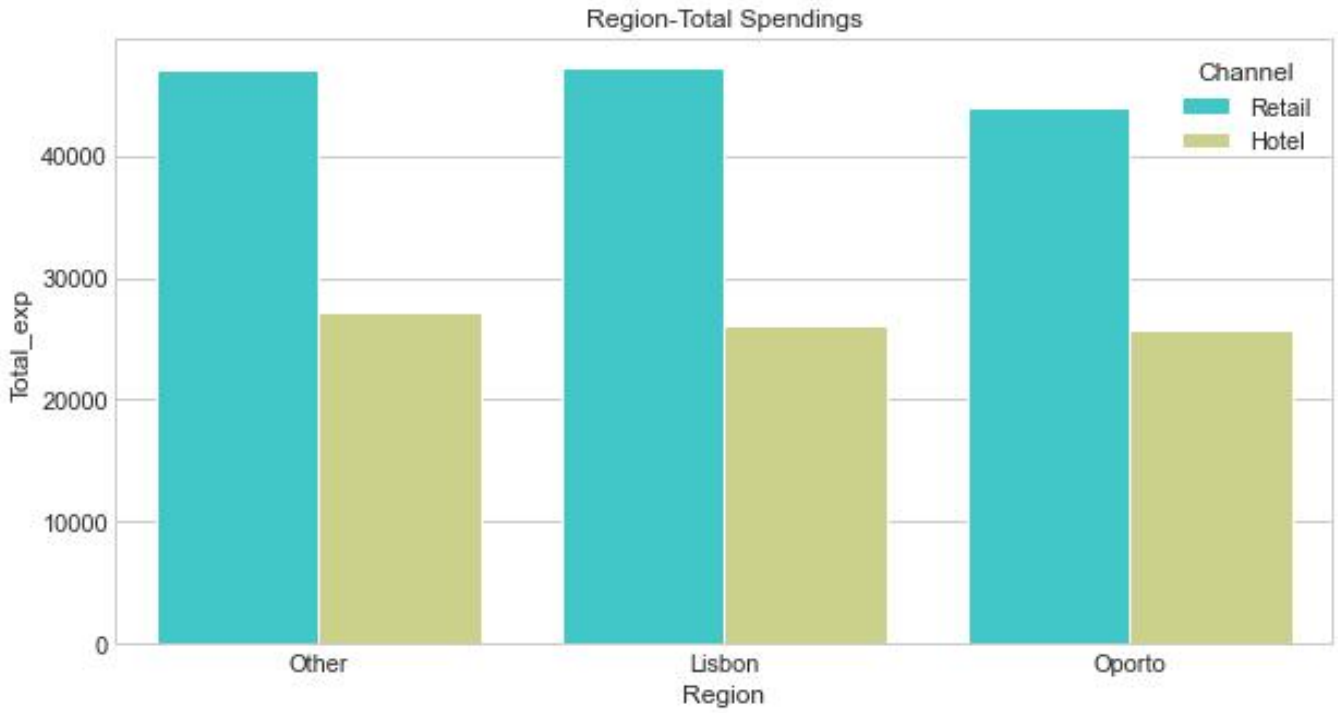


Fig1.4 - Region wise spending of Channels

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.

Fresh - Region & Channel wise Analysis :

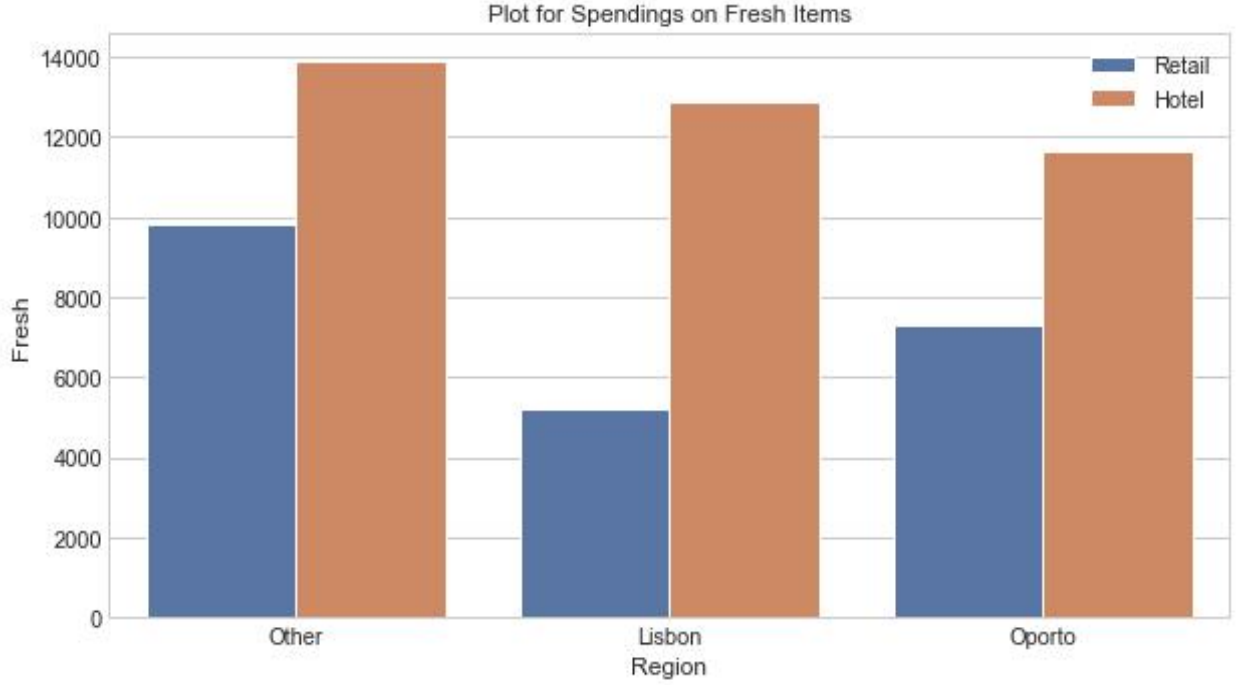


Fig1.5 - Region-wise spending on Fresh items by Channels

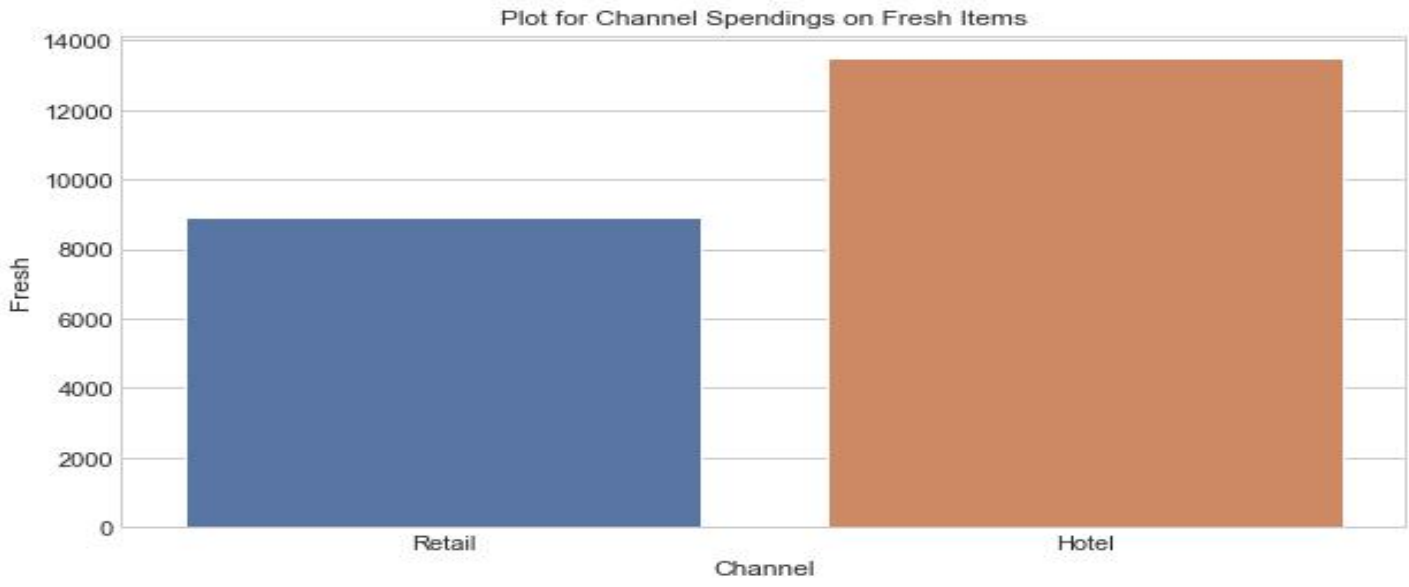


Fig1.6 - Channel-wise spending on Fresh items

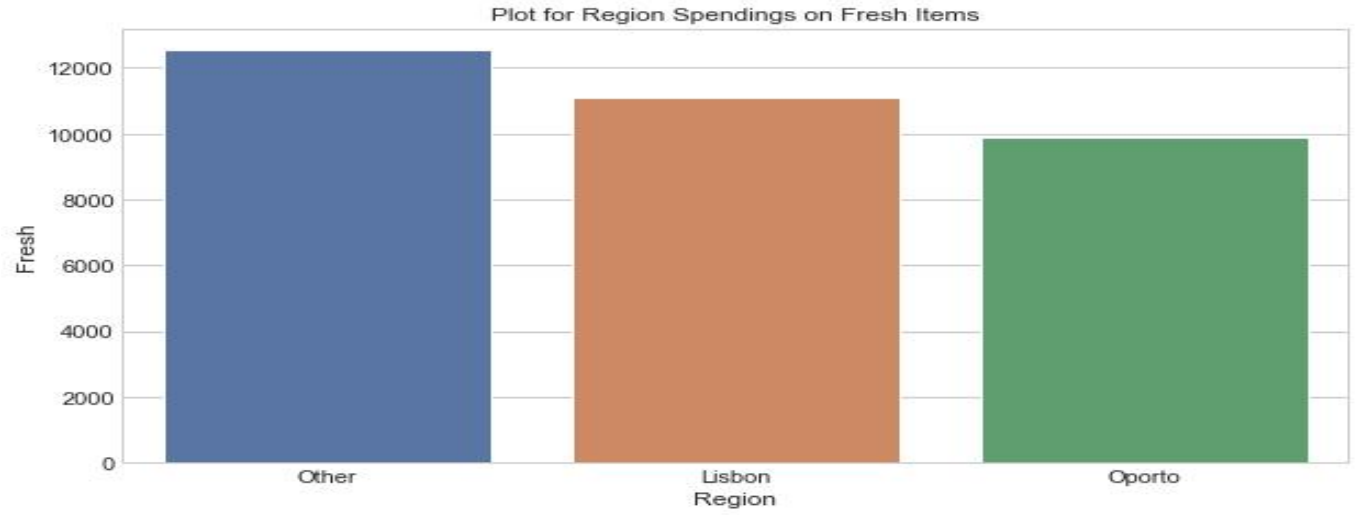


Fig1.7 - Region-wise spending on Fresh items.

- Above graphs show, ***Hotel Channel*** is consuming highest amount of Fresh items in ***Other Region.***
- Average spending on Fresh products is the highest, which is around 12000.30 by 'Hotel' Channel in 'Other' Region.
- This gives the inference that Hotels prefer fresher items for their guests and hence the highest consumption.

Milk - Region & Channel wise Analysis :

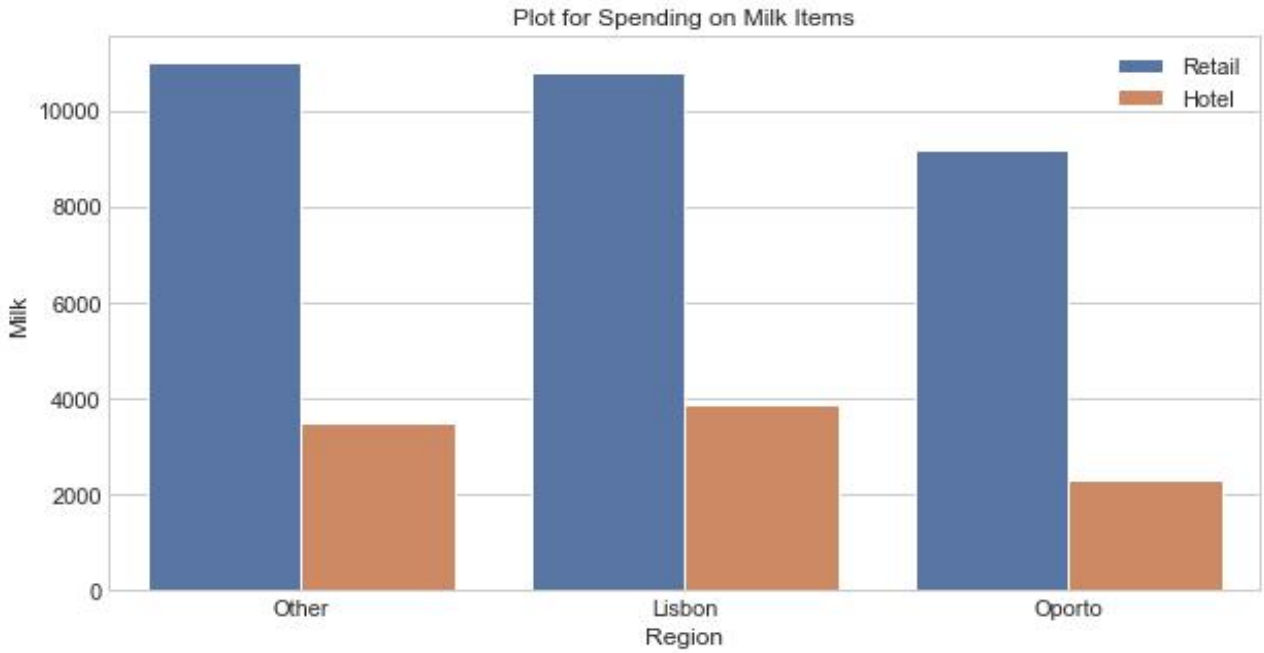


Fig1.8 - Region-wise spending on Milk items by Channels

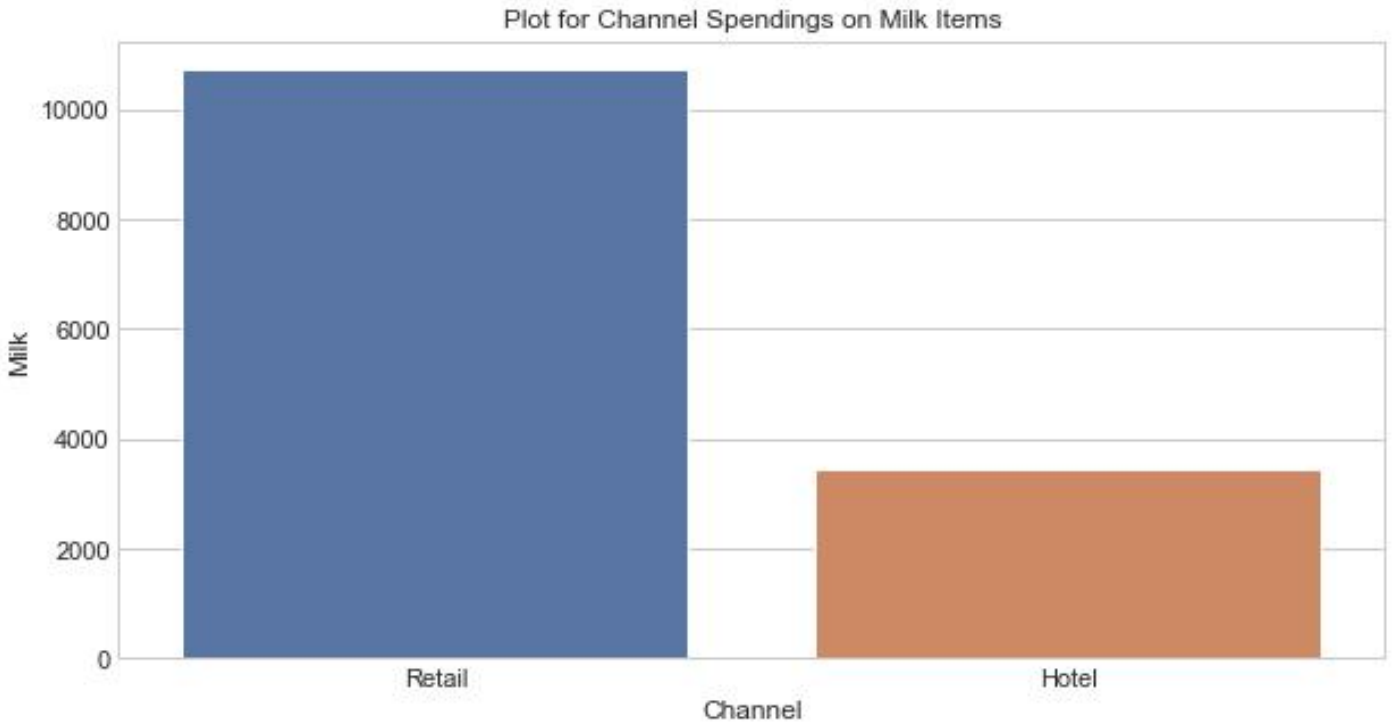


Fig1.9 - Channel-wise spending on Milk items.

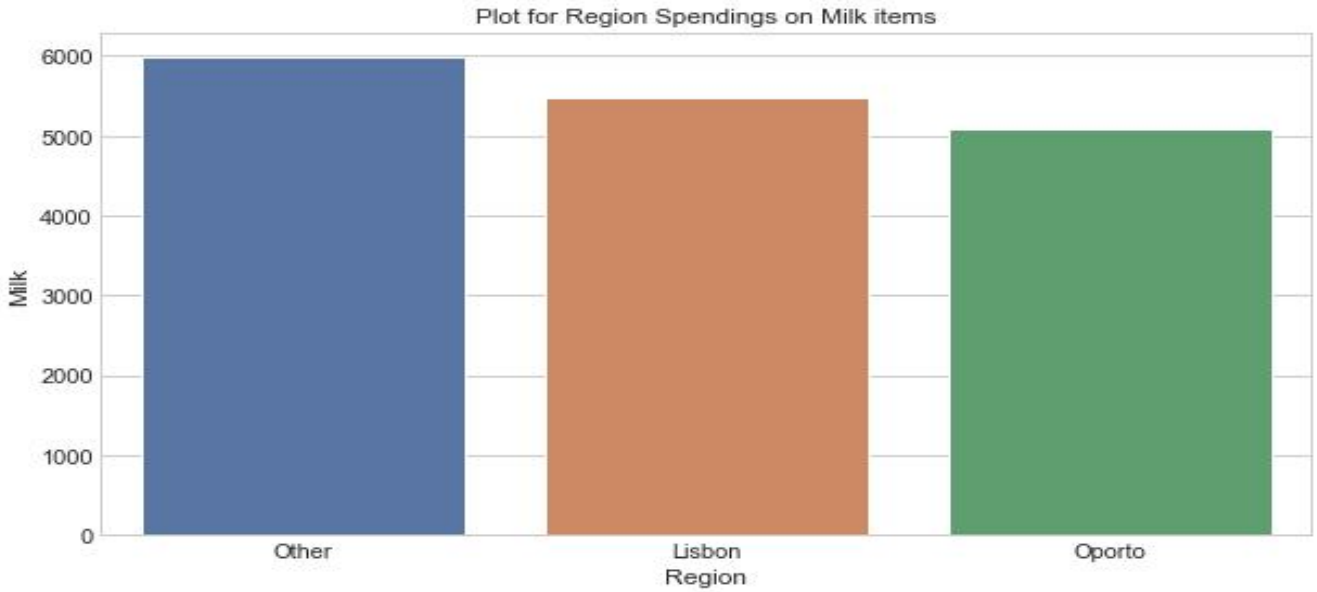


Fig1.10 - Region-wise spending on Milk items.

- Above graphs show, ***Retail Channel*** is consuming highest amount of Milk items in ***Other Region.***
- Average spending on Milk items is 5796.30 by 'Retail' Channel in 'Other' Region.
- This gives the inference that Retail Channel have a best business for Milk and milk products .
- And also tell the nature of the customers,i.e.,consumers lives locally like family or nearby residents who are likely their regular customers.

Grocery - Region & Channel wise Analysis :

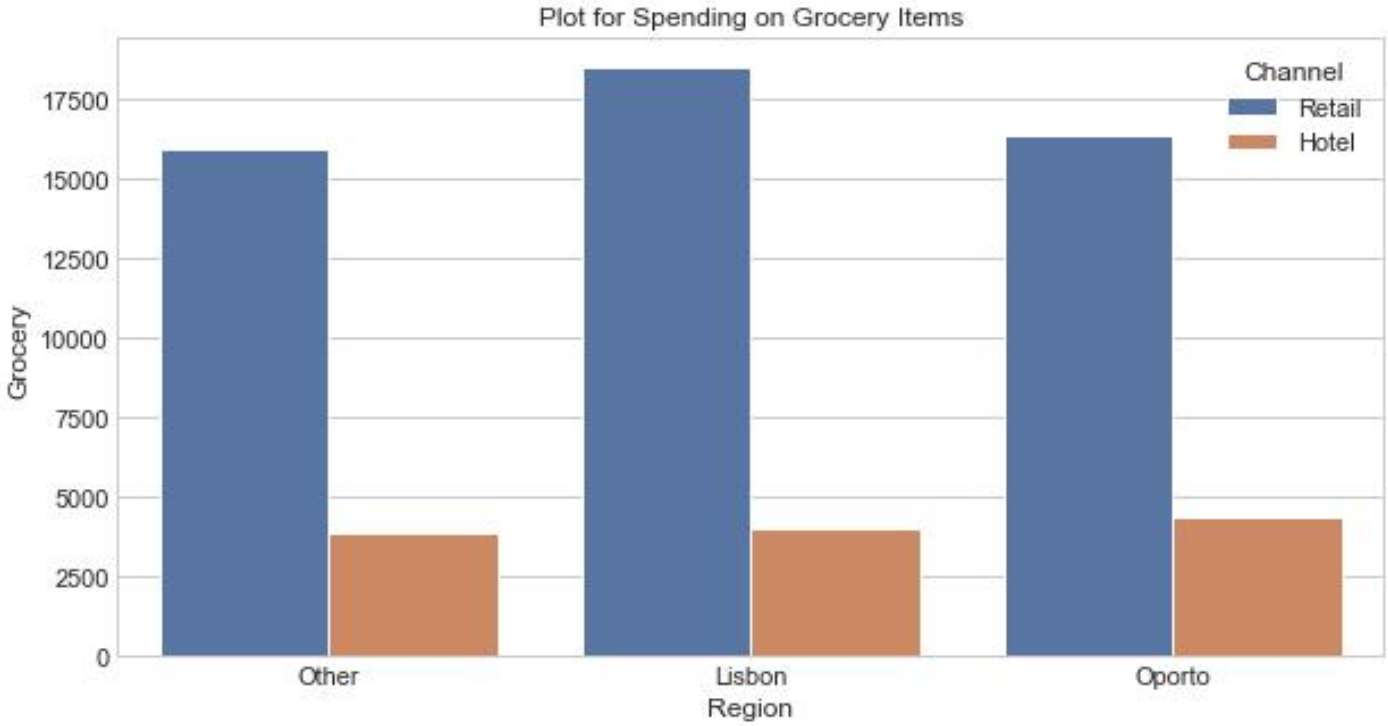


Fig1.10 - Region-wise spending on Grocery items by Channels

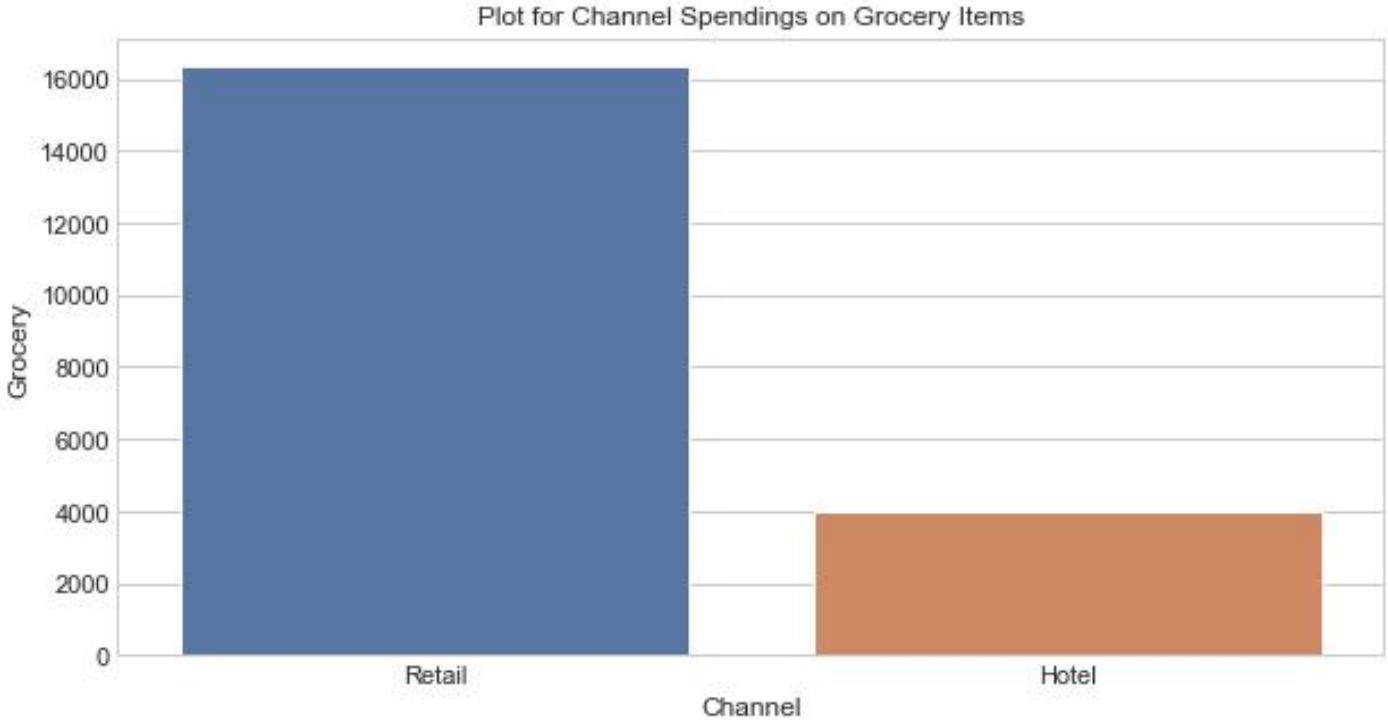


Fig1.11 - Channel-wise spending on Grocery items.

- ✧ Above graphs show, ***Retail Channel*** is consuming highest amount of Grocery items in ***Oporto Region.***
- ✧ Average spendings on 'Grocery' item is 7951.28 by 'Retail' Channel in 'Oporto' Region

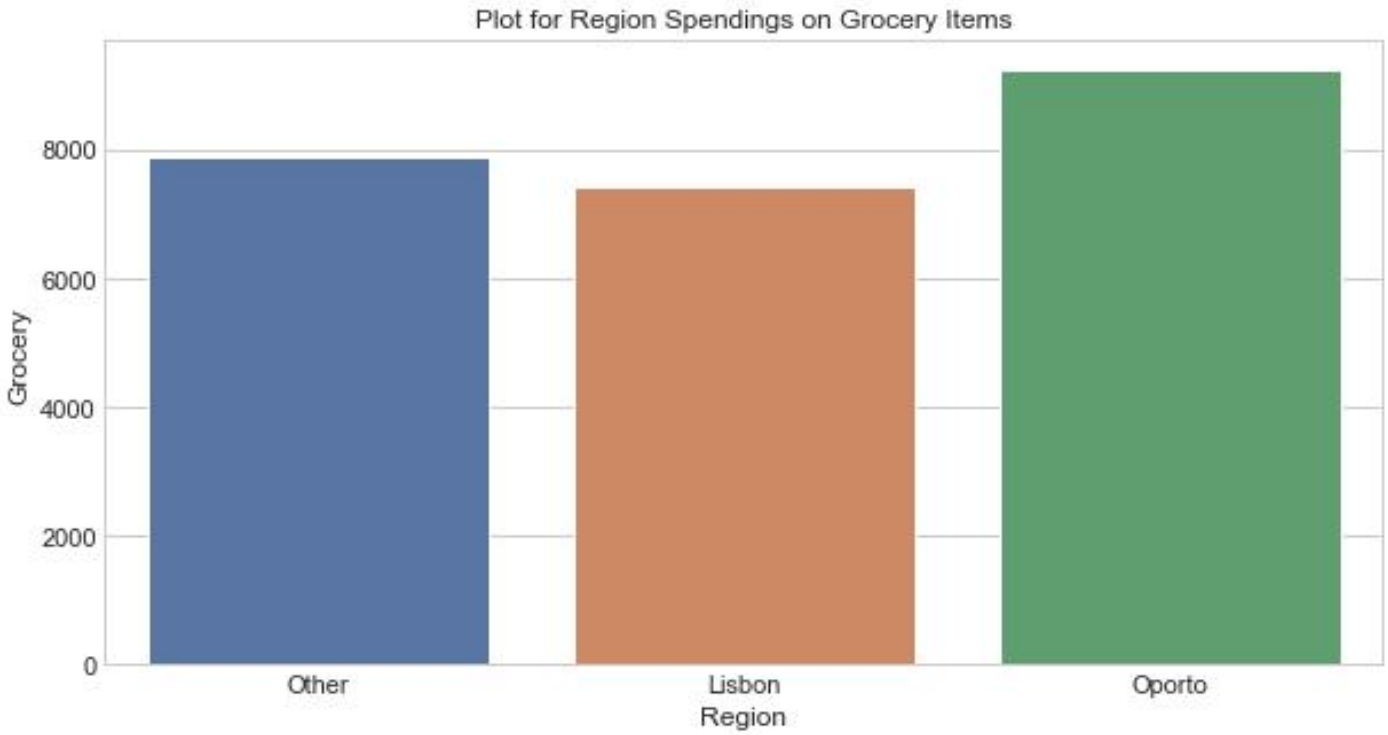


Fig1.12 -Region-wise spending on Grocery items.

- ✧ This gives the inference that Retail Channel have a good business for Grocery items.
- ✧ And also tell the nature of the customers,i.e.,consumers lives locally like family or nearby residents who are likely their regular customers.

Frozen - Region & Channel wise Analysis :

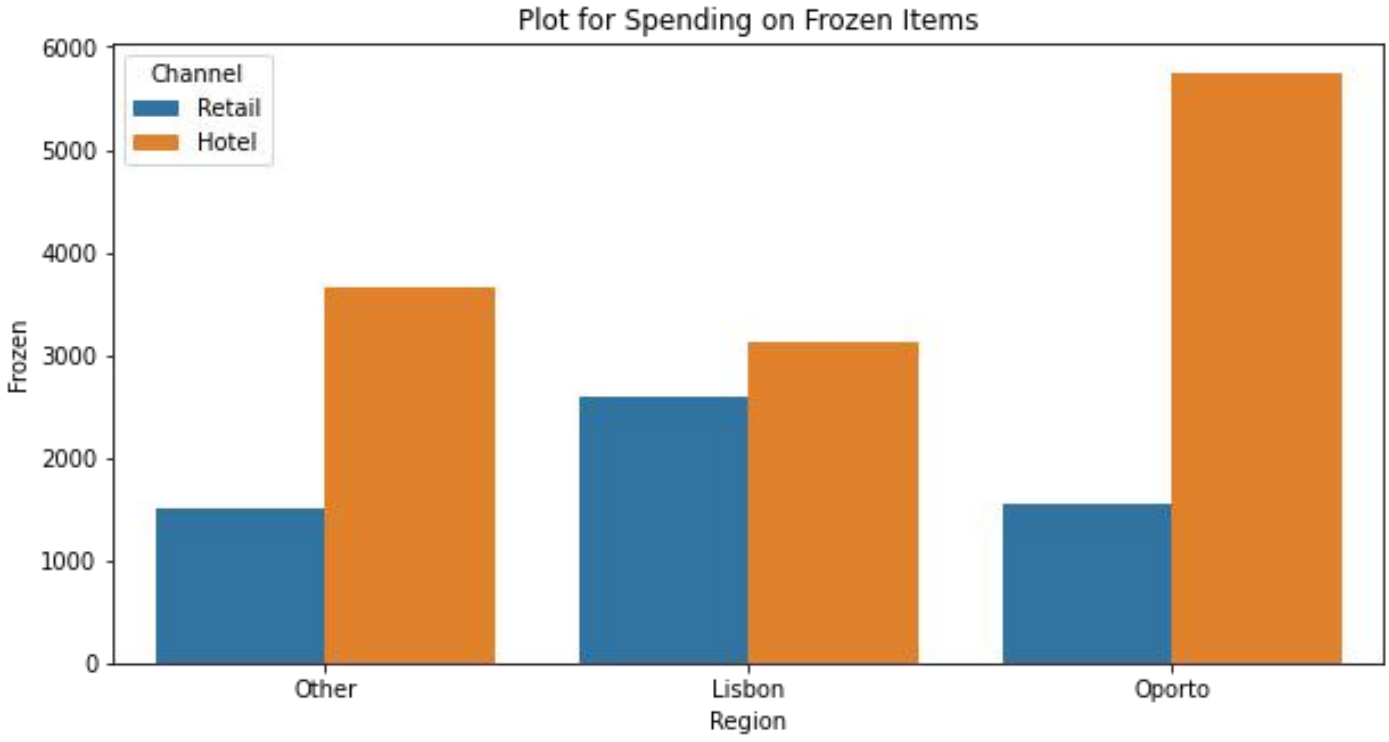


Fig1.13 - Region-wise spending on Frozen items by Channels

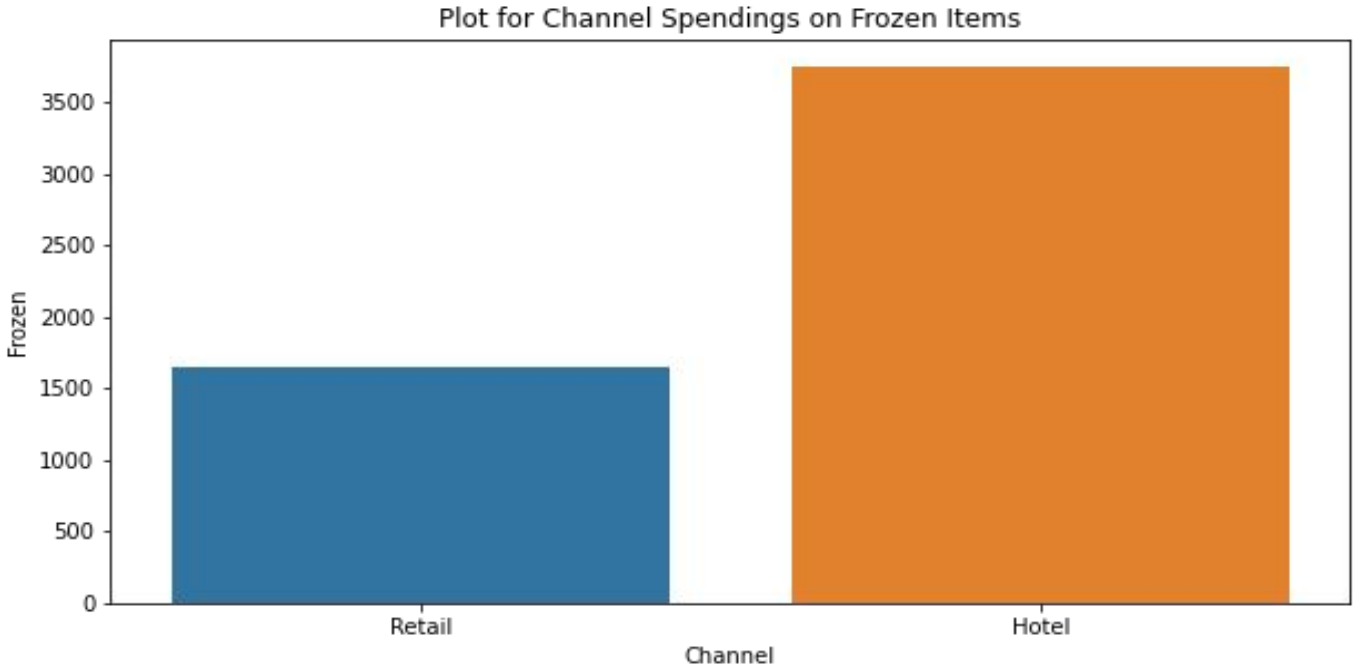


Fig1.14 - Channel-wise spending on Frozen items

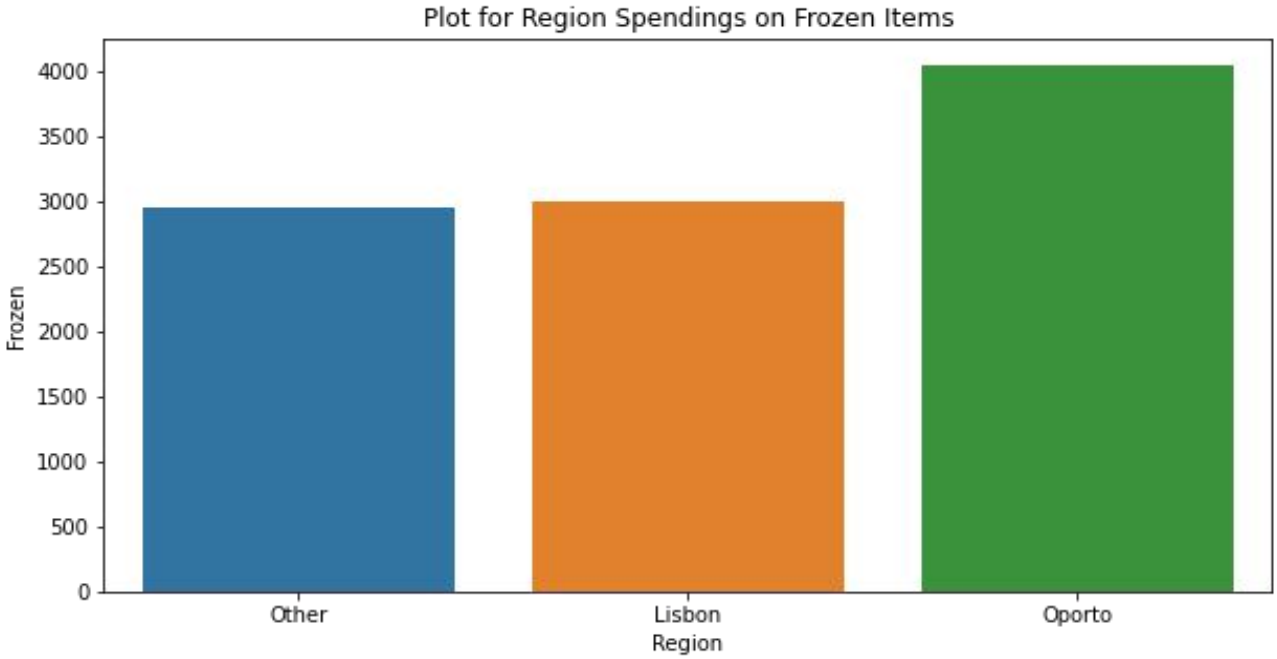


Fig1.14 - Region-wise spending on Frozen items

- ✧ Above graphs show, ***Hotel Channel*** is consuming highest amount of Frozen items in ***Oporto Region.***
- ✧ Average spending on 'Frozen' item is 3071.93 by 'Hotel' Channel in 'Oporto' Region.
- ✧ This gives the inference that Hotel Channel have a good business for Frozen items.

Detergents_Paper - Region & Channel wise Analysis :

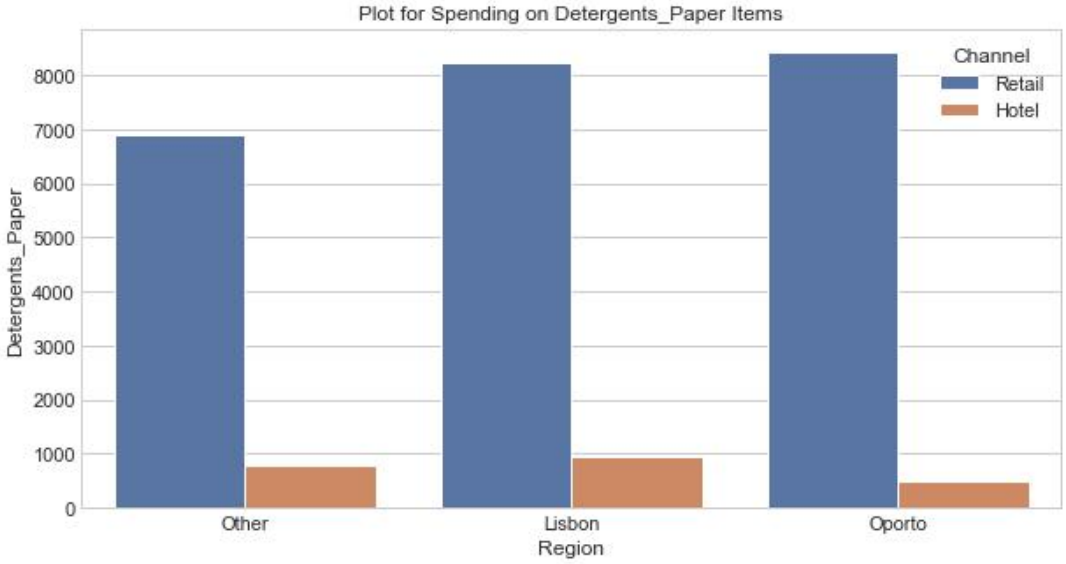


Fig1.15 - Region-wise spending on Detergents_Paper items by Channels

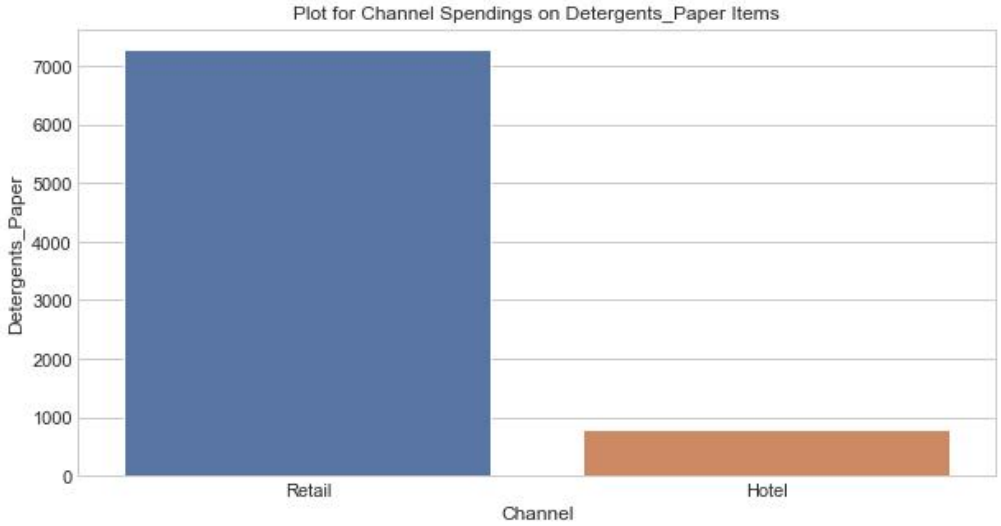


Fig1.16 - Channel-wise spending on Detergents_Paper items

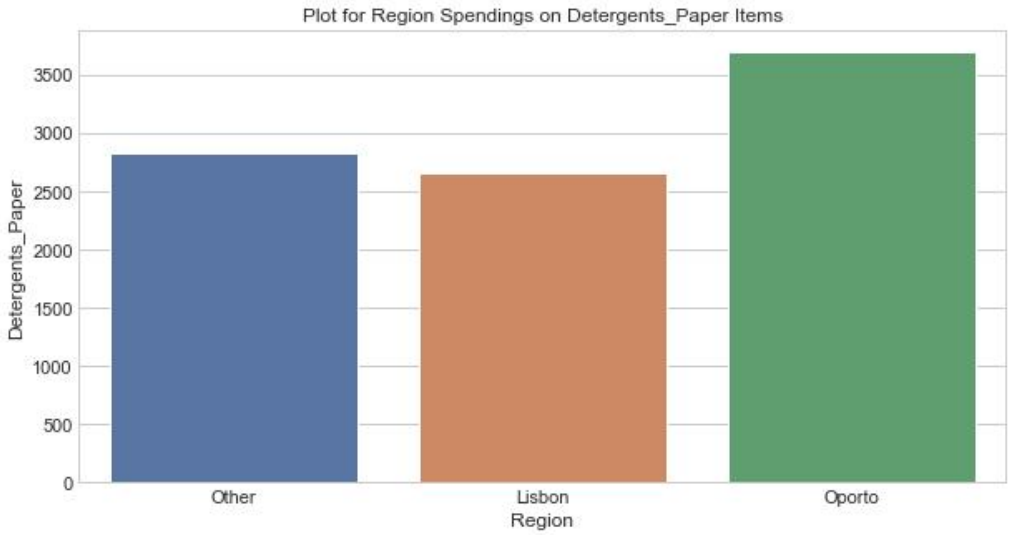


Fig1.17 - Region-wise spending on Detergents_Paper items

- Above graphs show, ***Retail Channel*** is consuming highest amount of Detergents_Paper items in ***Oporto Region.***
- Average spending on 'Detergents_Paper' item is 2881.49 by 'Hotel' Channel in 'Oporto' Region.
- This gives the inference that Retail Channel have a normal business for Detergents_Paper items compared to other items.

Delicatessen - Region & Channel wise Analysis :

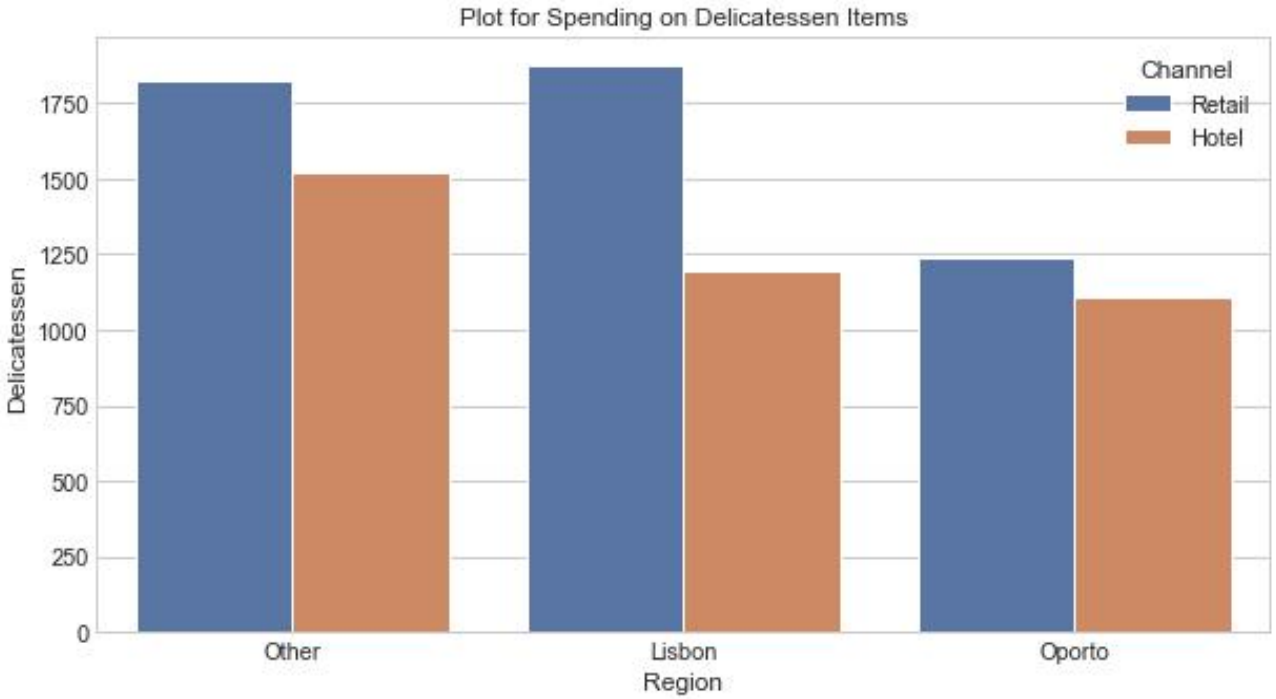


Fig1.18 - Region-wise spending on Delicatessen items by Channels

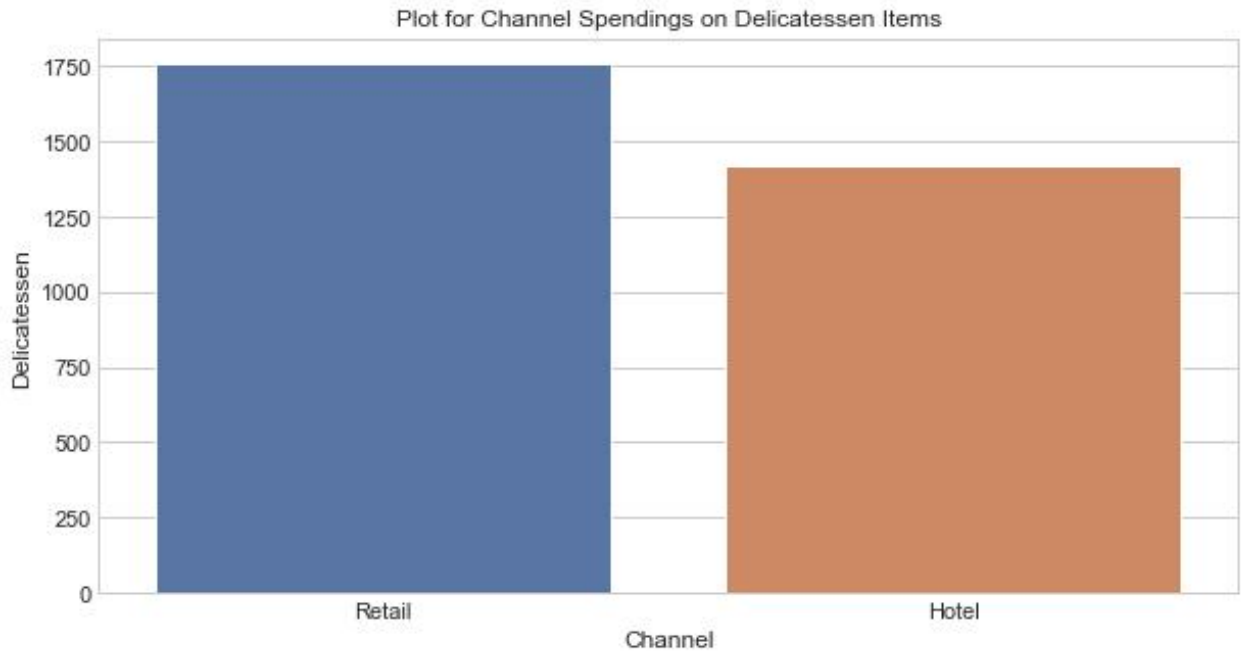


Fig1.19- Channel-wise spending on Delicatessen items

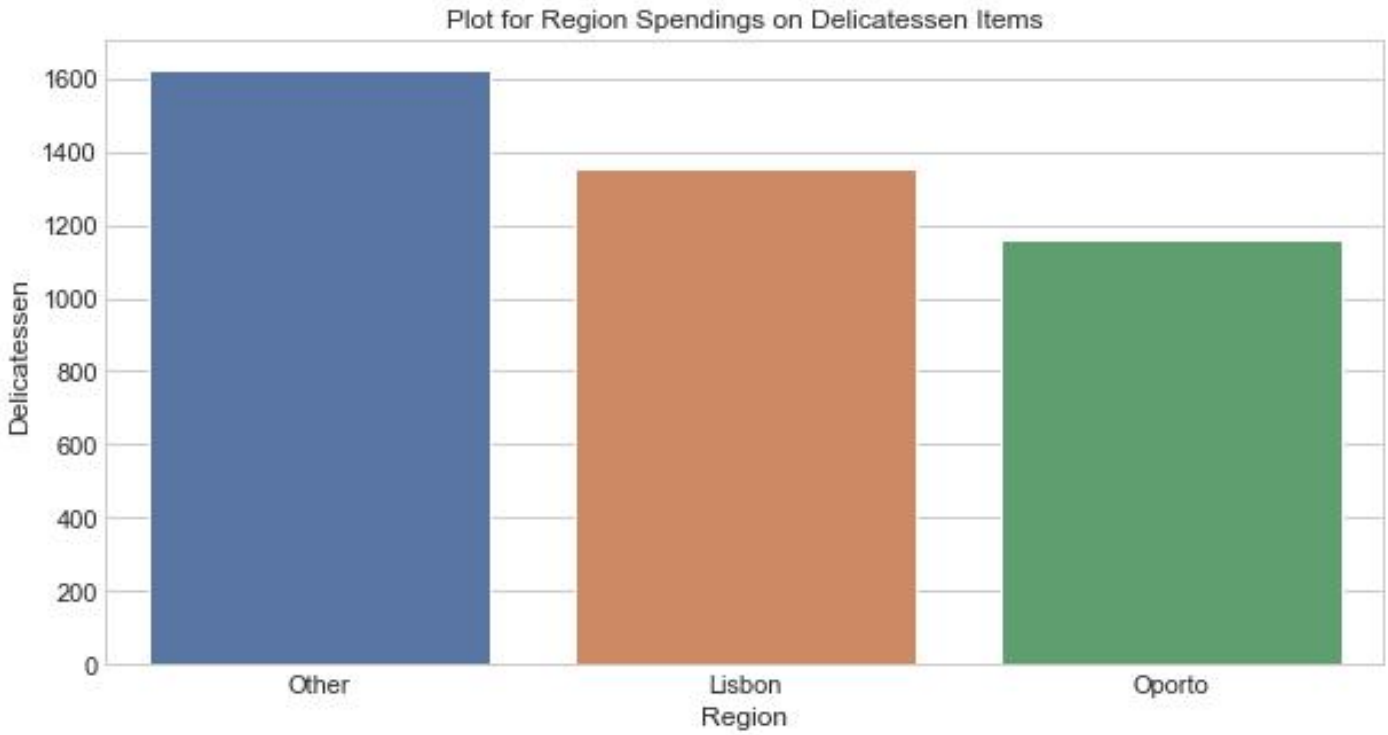


Fig1.20- Region-wise spending on Delicatessen items

- Above graphs show, ***Retail Channel*** is consuming highest amount of Delicatessen items in ***Other Region.***
- Average spending on 'Delicatessen' item is 1524.87 by 'Hotel' Channel in 'Oporto' Region.
- This gives the inference that Retail Channel have the least business for Delicatessen items compared to all the other items.
- The business needs to work on its advertisement in least consumed region through their distribution channels to increase the business volume.

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?

	count	mean	std	min	25%	50%	75%	max
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

Table 1.4-Data Description with Mean Values of all Items

The average consumption of fresh items is highest and the least is Delicatessen.

Calculating the variance we get,

Fresh	1.599549e+08
Milk	5.446997e+07
Grocery	9.031010e+07
Frozen	2.356785e+07
Detergents_Paper	2.273244e+07
Delicatessen	7.952997e+06
dtype: float64	

Fresh	1.05
Milk	1.27
Grocery	1.19
Frozen	1.58
Detergents_Paper	1.65
Delicatessen	1.85
dtype: float64	

Tabel 1.5 - Variance and Co-variance between items

From the above data, we can conclude that 'Fresh' items are most consistent in nature, whereas Delicatessen have the least consistency.

1.4 Are there any outliers in the data?

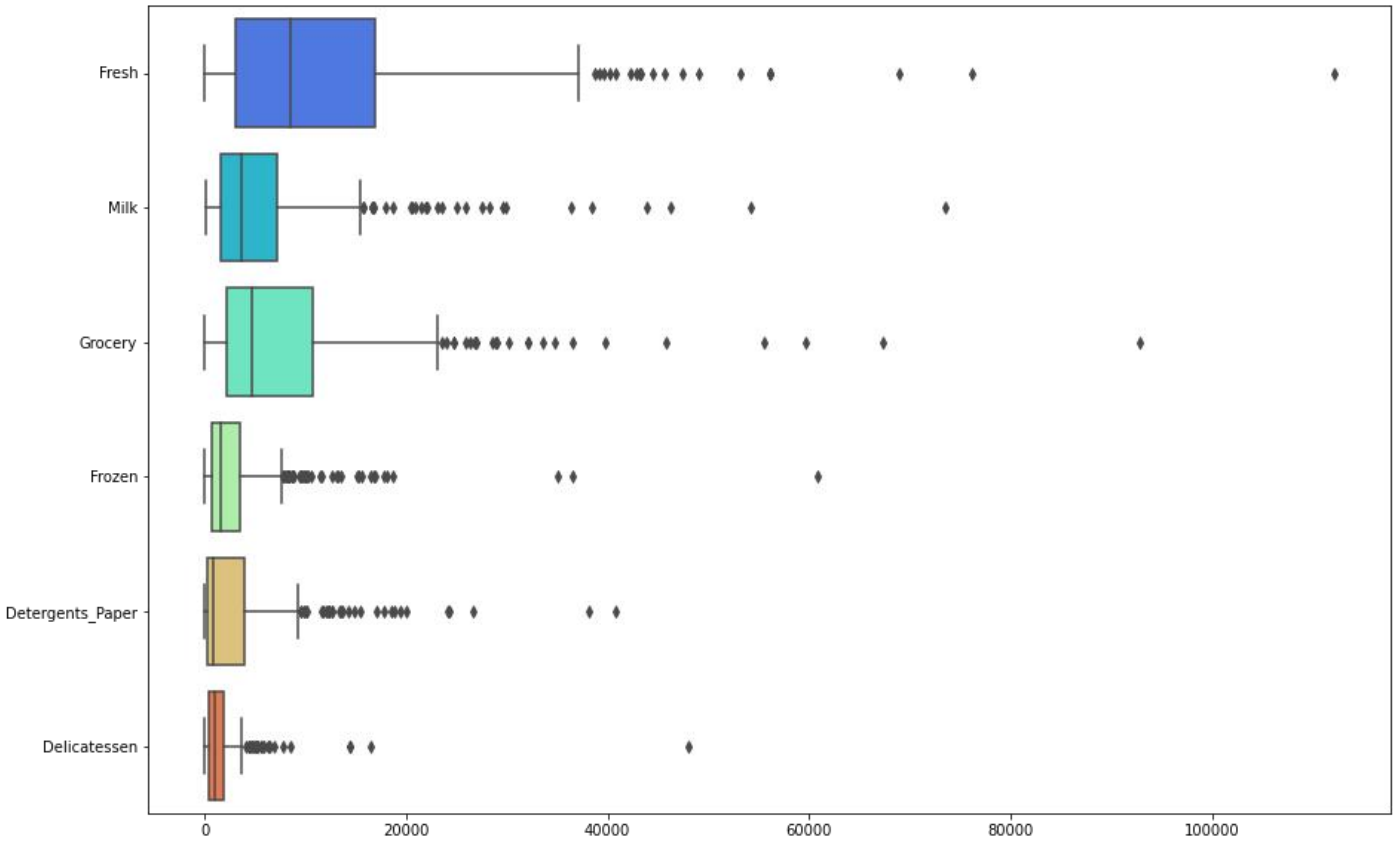


Fig.1.21- Boxplot for Checking outliers

The above plot shows outliers are present in the data.

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.

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Fresh	1.000000	0.100510	-0.011854	0.345881	-0.101953	0.244690
Milk	0.100510	1.000000	0.728335	0.123994	0.661816	0.406368
Grocery	-0.011854	0.728335	1.000000	-0.040193	0.924641	0.205497
Frozen	0.345881	0.123994	-0.040193	1.000000	-0.131525	0.390947
Detergents_Paper	-0.101953	0.661816	0.924641	-0.131525	1.000000	0.069291
Delicatessen	0.244690	0.406368	0.205497	0.390947	0.069291	1.000000

Table 1.5 Correlation Values

Fig.1.22- Correlation Heat Map

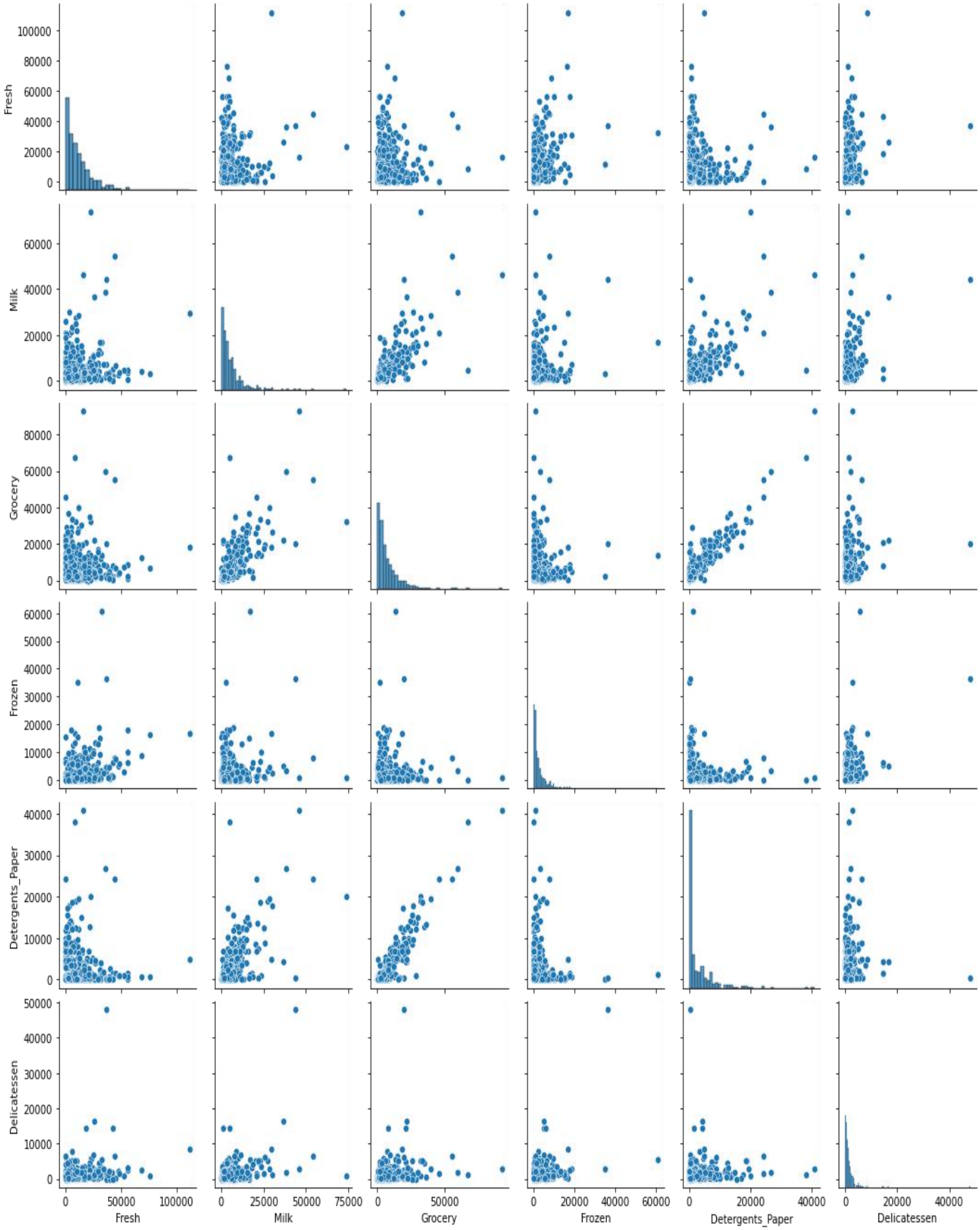


Fig.1.23- Pair Plot - Relation between different items.

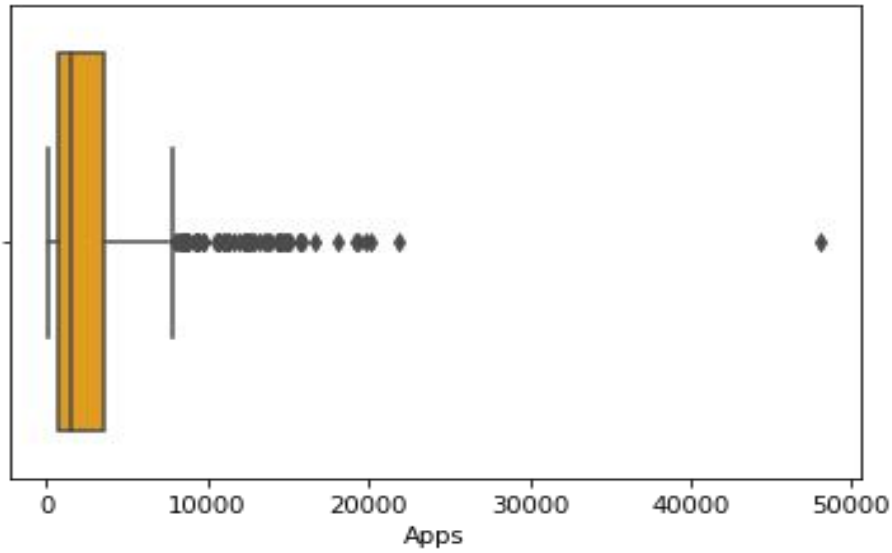
Business Inferences and Suggestions :

- The data suggests that most of the consumers/clients are in the 'Other' region and hence the business needs to advertise or plan more ad-campaigns in 'Lisbon' and 'Oporto' region.
- As seen in the data, consumers are spending much highest on Fresh items, Grocery and Milk. The business needs to maintain the ensure the availability of these items at all times.
- Delicatessen is the least purchased item. Business may need to reduce the price or invest into the marketing of these items to increase the profit from this item.
- Fresh and Frozen are purchased more via Hotel Channel as compared to Retail channel, as it is nature of Hotel business to give the best possible service and fresh.
- Milk, Grocery is purchased more via Retail Channel as compared to Hotel Channel.
- Business needs to focus on the promotion of Frozen, Detergent_Paper and Delicatessen, since they are lowest purchased items.
- Milk has positive correlation with all the products. It also signifies it's purchase may increase with other products.
- Similarly, Fresh items has negative correlation with Grocery and Detergent_Paper.
- All items with positive correlation will help increase the business volume and negative correlation items will affect the business adversely.

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

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



Fg.2.1 Boxplot for apps

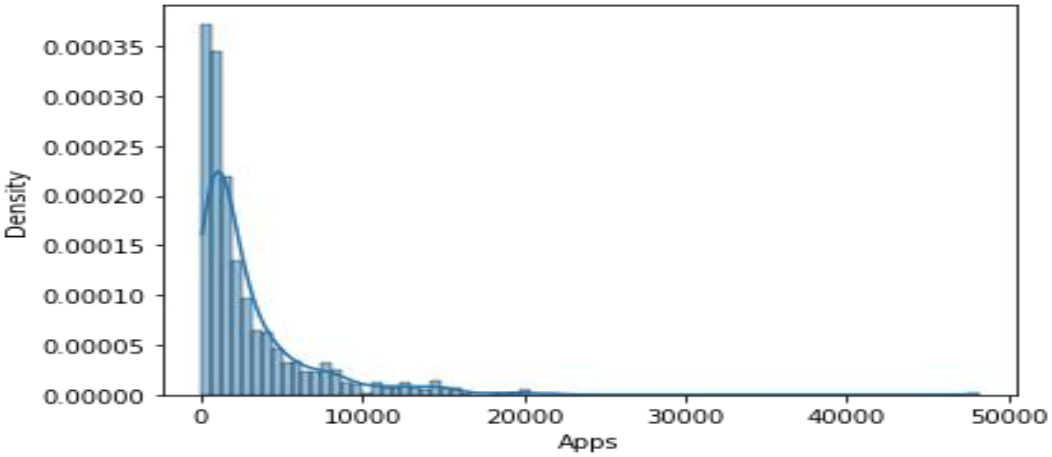


Fig.2.2 Distribution for Apps

The above plot have outliers, the applicants are 75% and above. The distribution curve is not normal and right skewed.

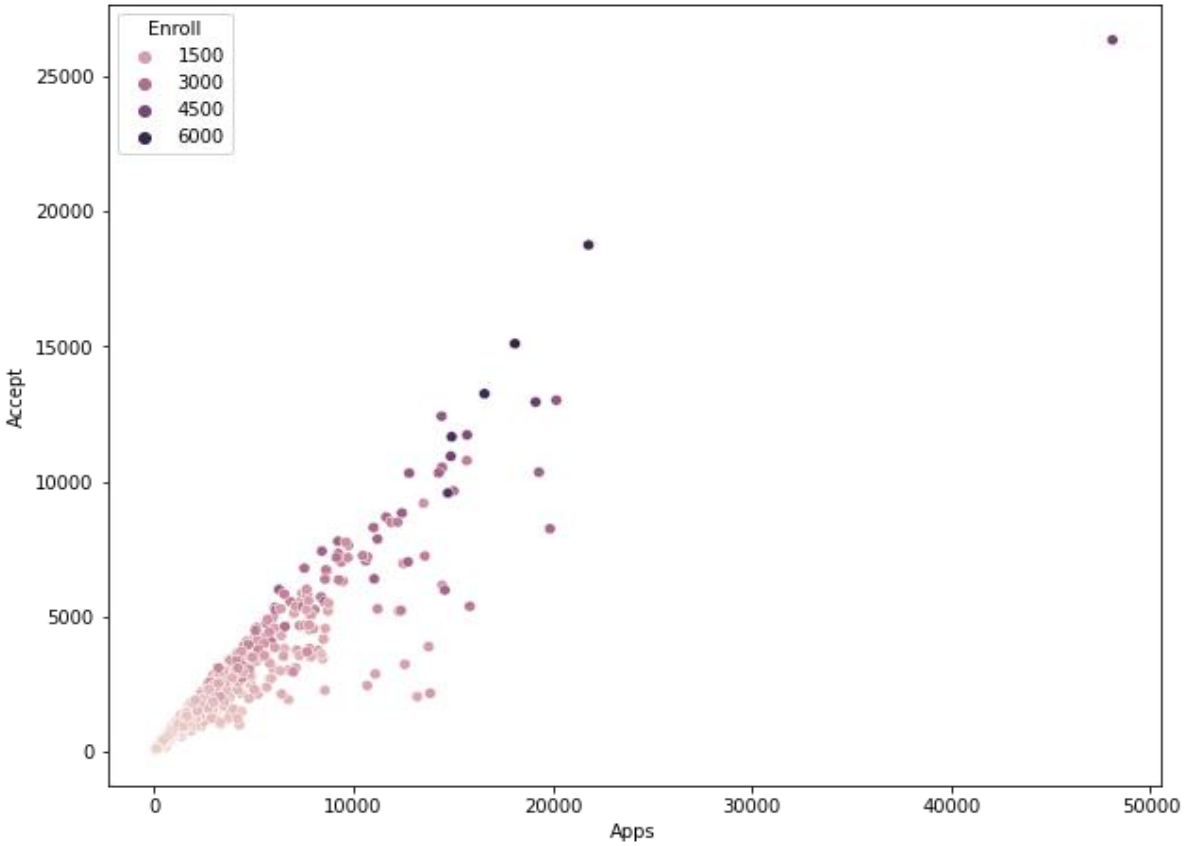


Fig.2.3. Scatter plot for Apps Vs to Accept with Enrol

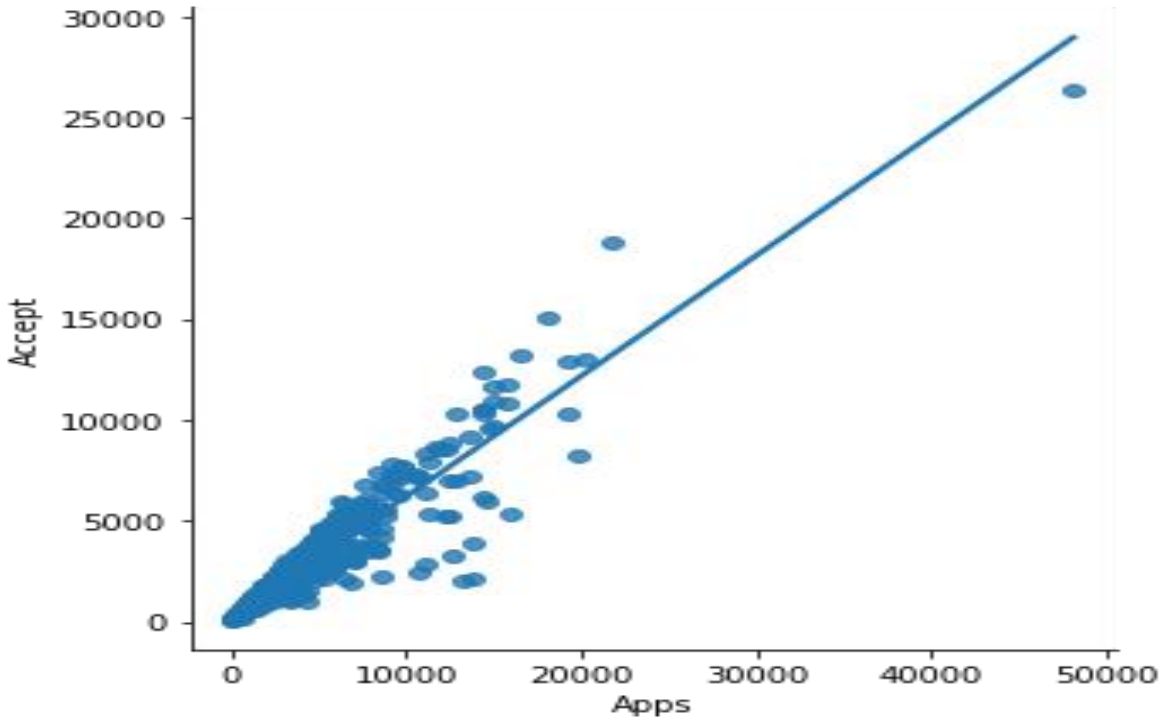


Fig.2.4. Scatter plot for Apps Vs Accept

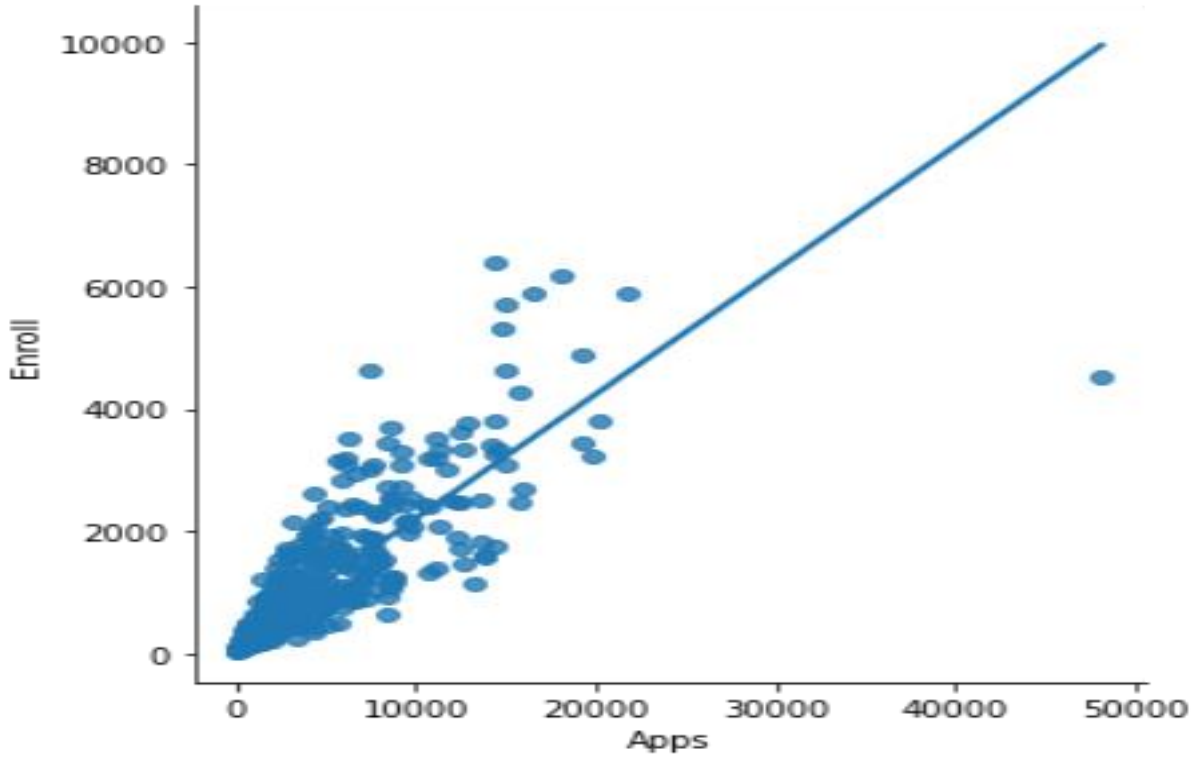


Fig.2.4. Scatter plot for Apps Vs Enroll

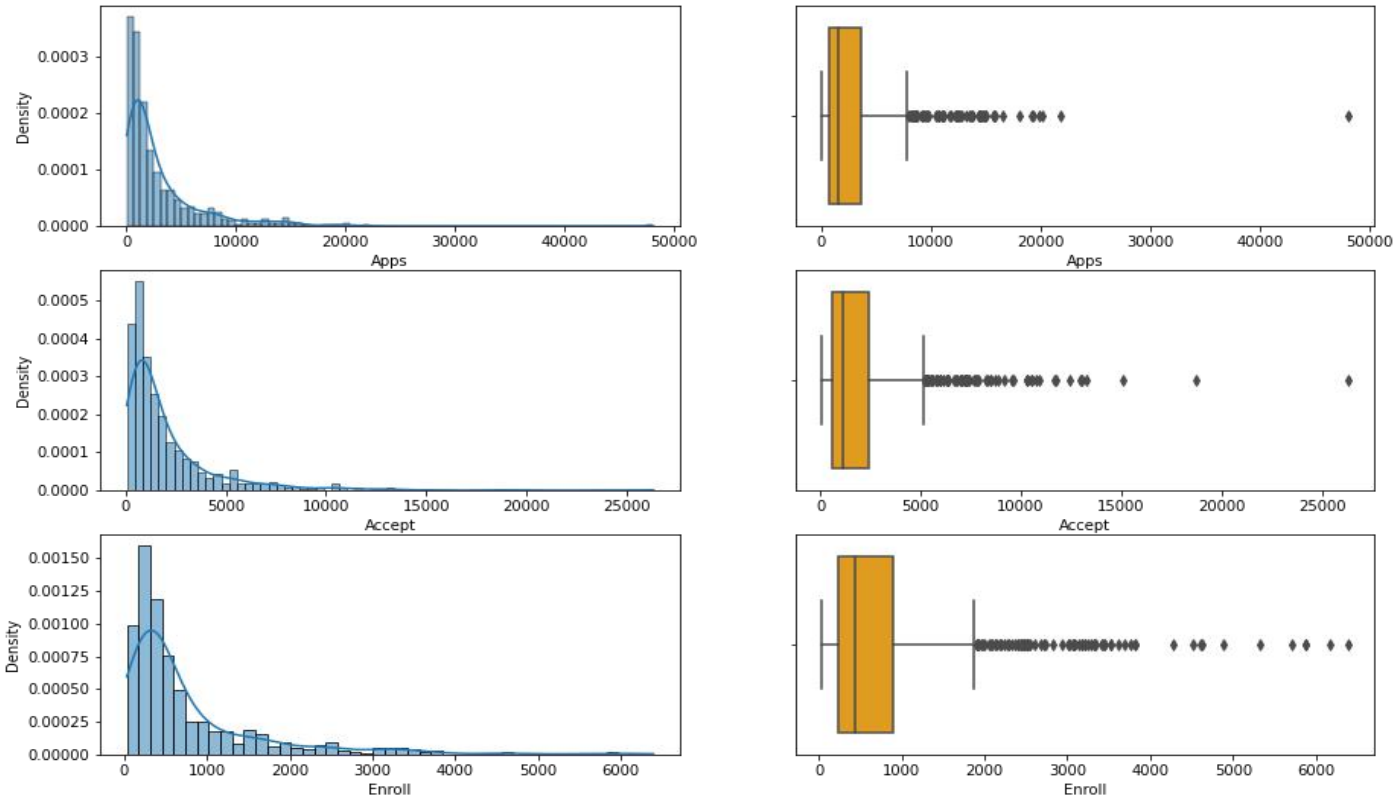


Fig. 2.5. Comparison of Distribution and Box Plot between Application, Accepted and Enrollment.

The above plot have outliers, most data lies in 75% and above. The distribution is curver is not normal.

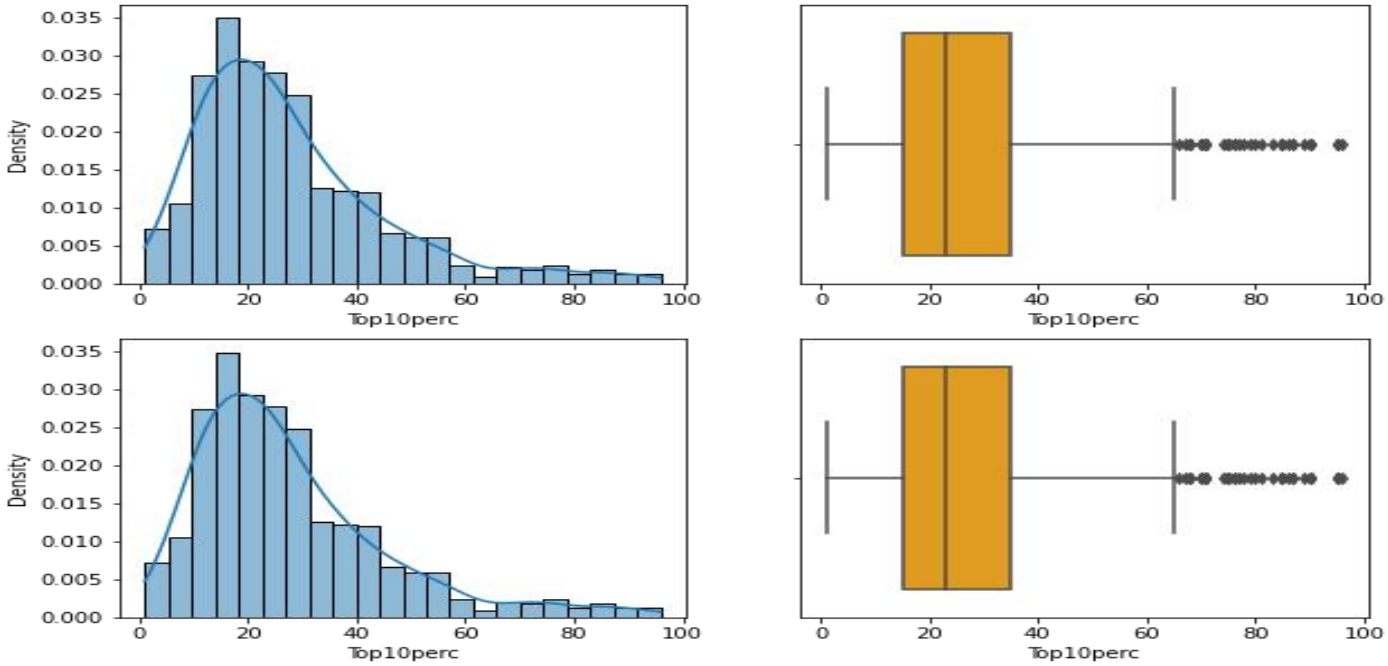


Fig. 2.6. Comparison of Distribution and Box Plot between Top 10% and 25% Students

The distribution is near normal but still skewed.

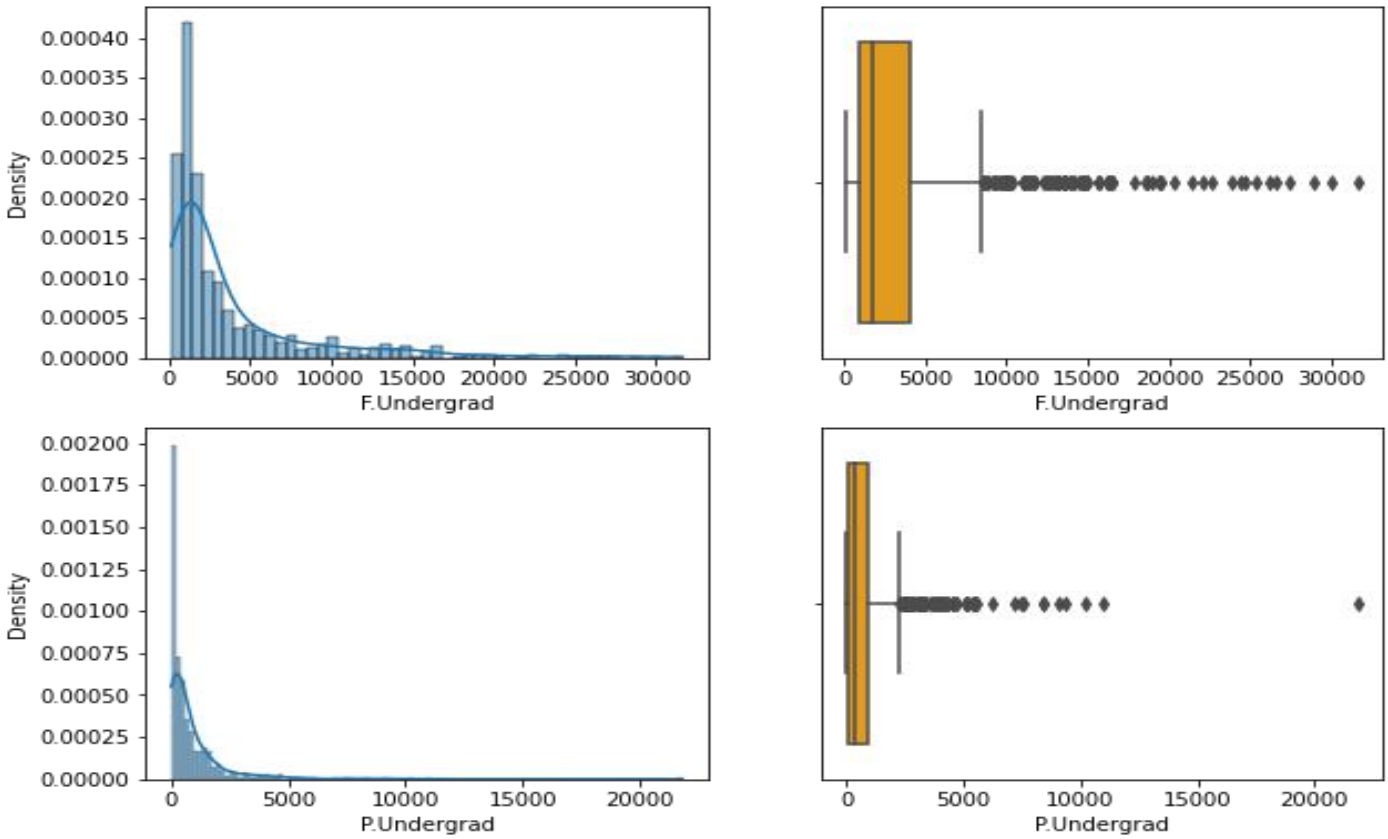


Fig. 2.6. Comparison of Distribution and Box Plot between F. Undergrad and P. undergrad.

The above plot have outliers and there is no normal distribution and highly skewed.

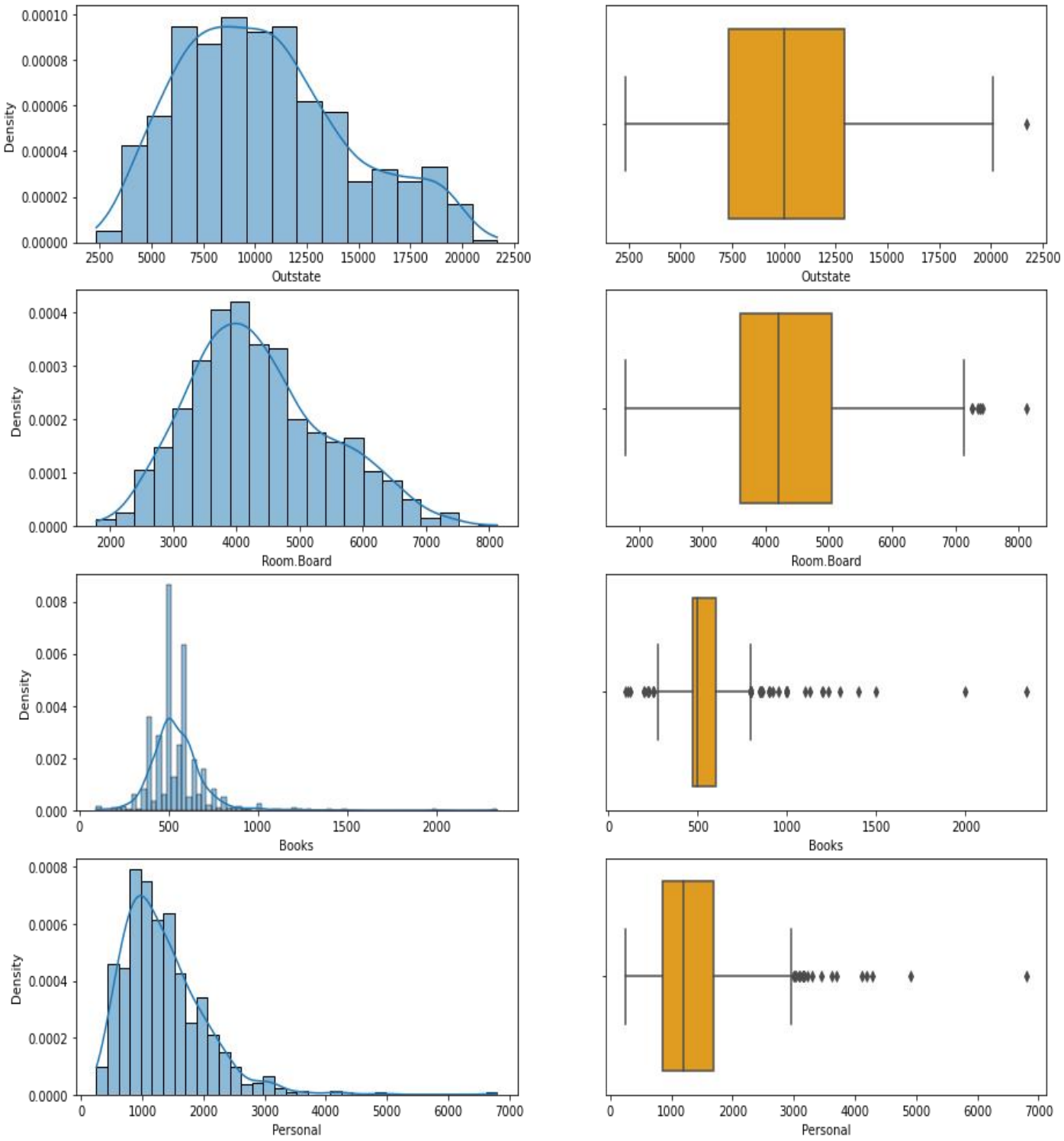


Fig. 2.6. Comparison of Distribution and Box Plot of all the expenses of students.

This distribution seems somewhat normally with less outliers.

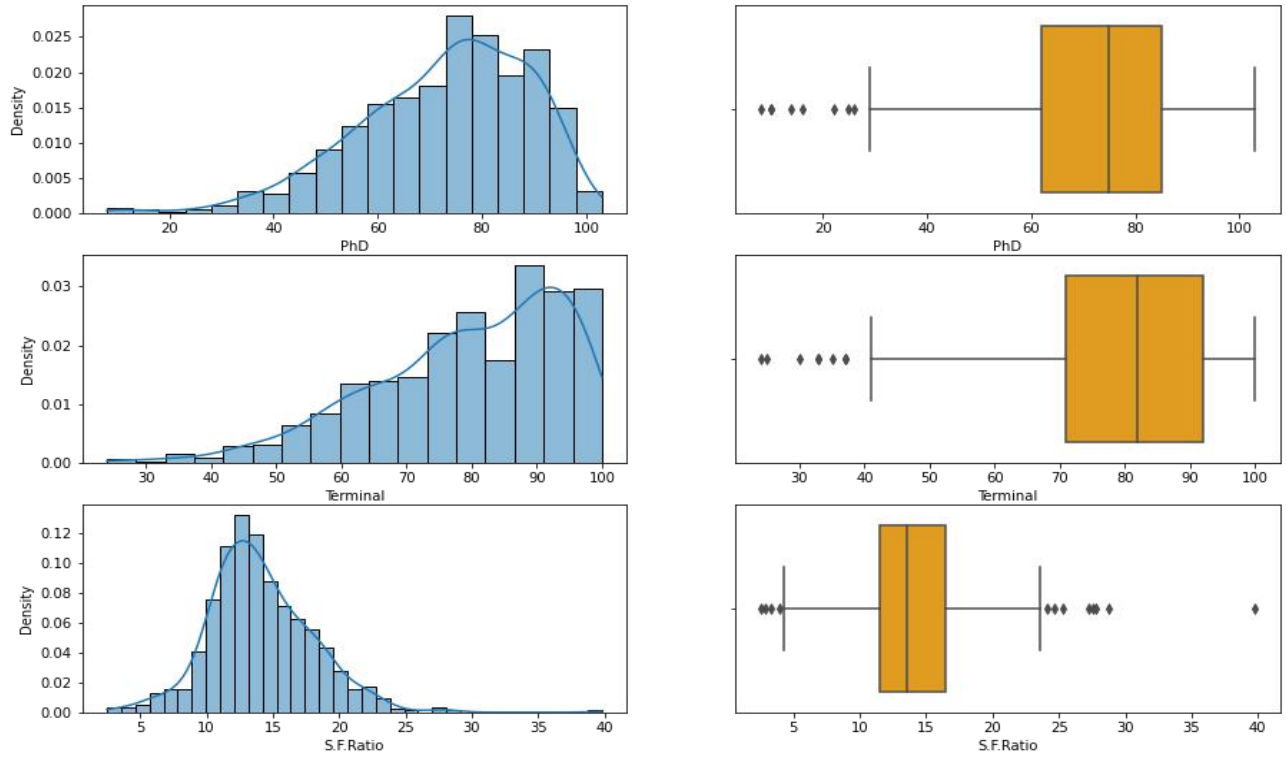


Fig. 2.7. Comparison of Distribution and Box Plot PhD, Terminal and S/F ratio.

PhD and Terminal is left skewed and box plots are left skewed.

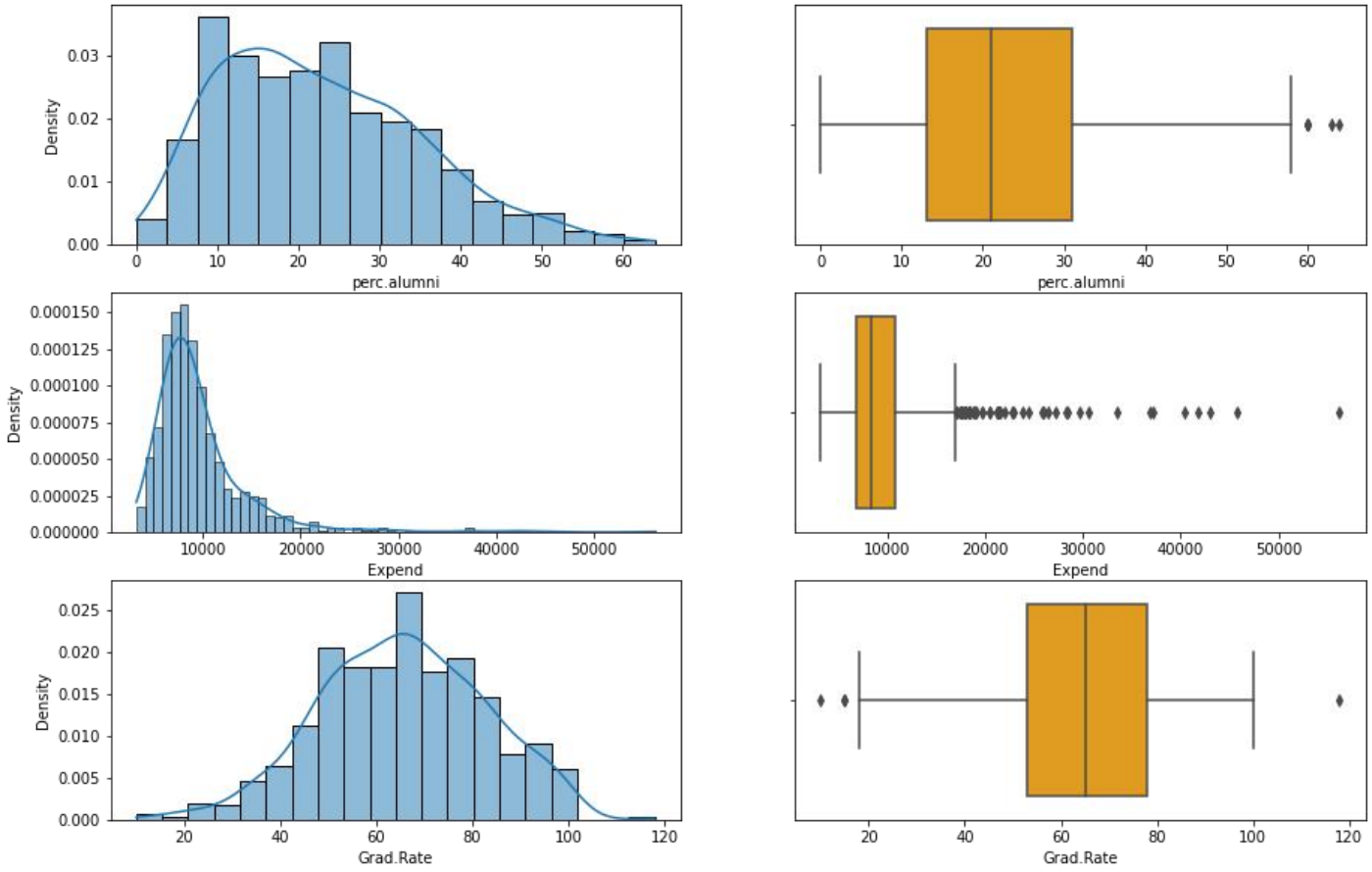


Fig. 2.8. Comparison of Distribution and Box Plot between Alumni, Expend and Grad Rate.

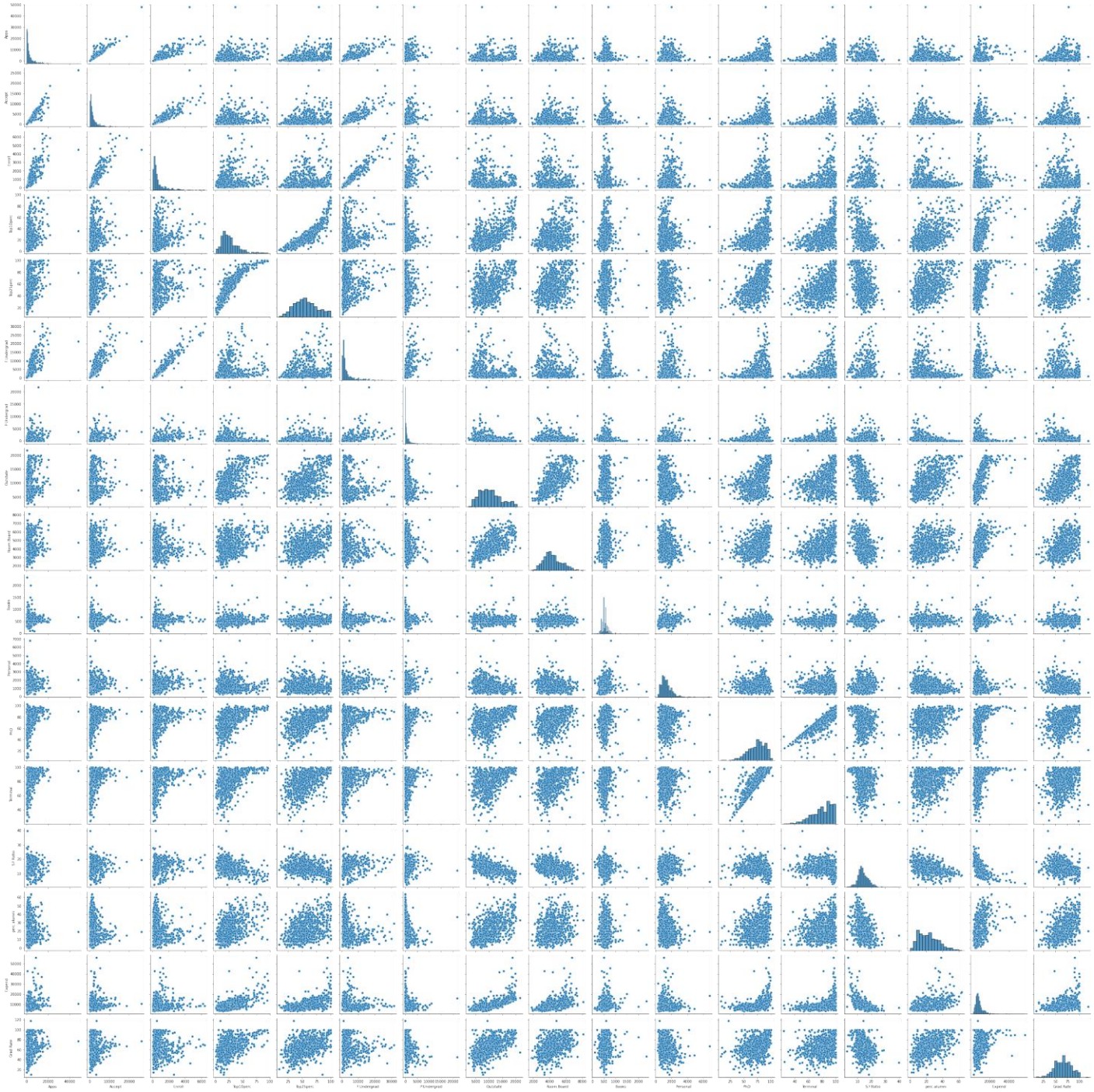


Fig. 2.9. Pair plot showing relation between all the data.

The above plot shows us relation between all the data and their effect on their one another.

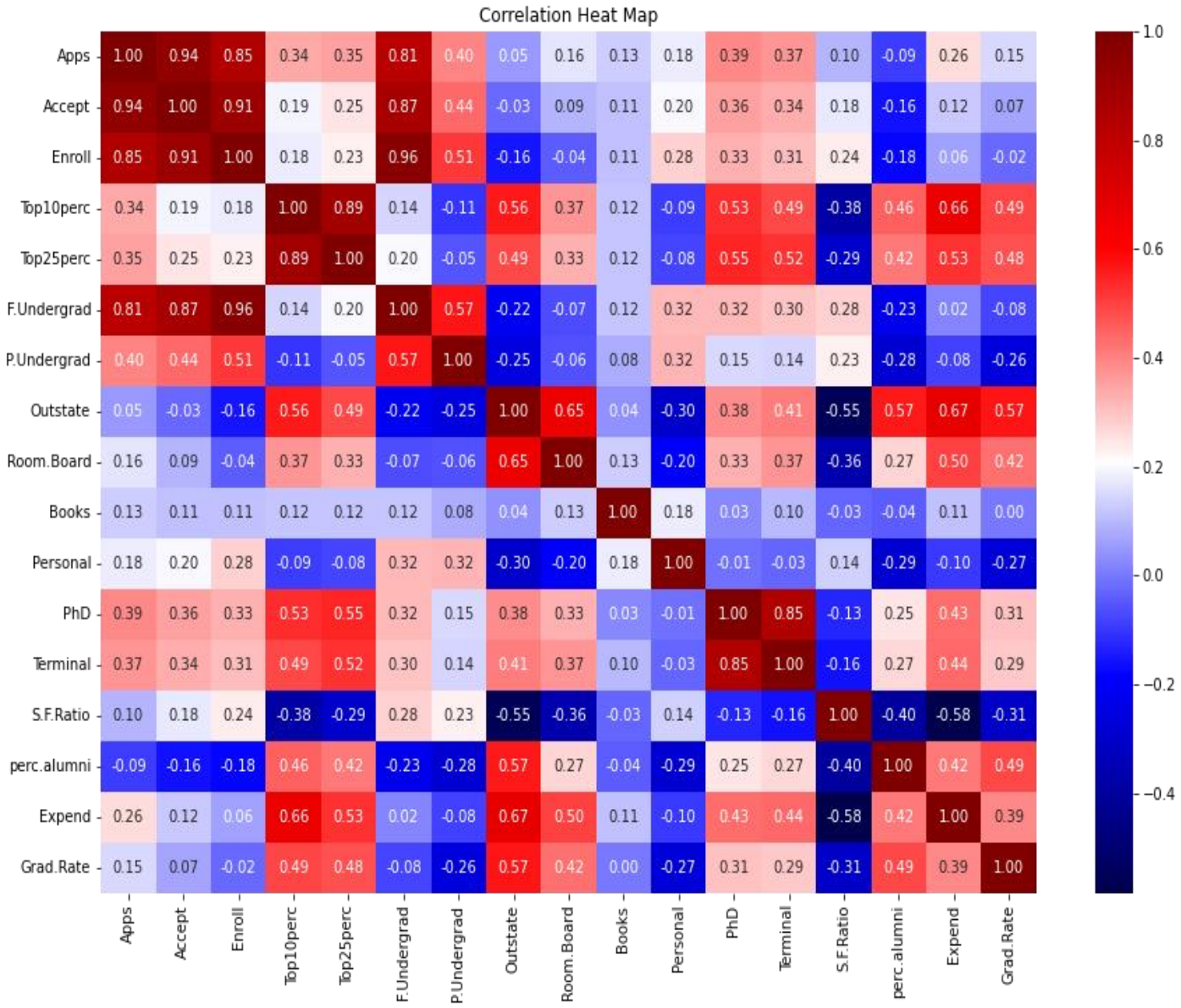


Fig. 2.10. Heat-Map Correlation Between data.

This map shows all the data and their positive or negative relation with each other.
Negative values indicates that they are likely related to each other like personal with Outstate. This means students who are from outstate are not able to spend for their personal requirements. They maybe paying higher tuition fees, as outstate and Room.Board are positively correlated and they are paying for their room and boarding.