**BUSINESS REPORT**

**Statistical Methods for Decision Making**

**Prepared by: Dhruv Dosad**



**Agenda**

|  |  |
| --- | --- |
| **Content** | **Page No** |
| **Problem I** | **1-8** |
| 1.1.1 Use methods of descriptive statistics to summarize data | 2 |
| 1.1.2 Which Region and which Channel spent the most? | 3 |
| 1.1.3 Which Region and which Channel spent the least? | 3 |
| 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. | 4 |
| 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? | 5 |
| 1.4 Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments. | 6 |
| 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. | 7 |
| **Problem II:** | **8-16** |
| 2.1 Perform Exploratory Data Analysis [Univariate, Bivariate, and Multivariate analysis to be performed]. What insight do you draw from the EDA? | 8 |

## 

## 

## 

## 

## **List of Figures**

* Fig 1.1  Total Spending across Channels
* Fig 1.2  Avg Spending across Channels
* Fig 1.3 Total Spending across Region
* Fig 1.4  Avg Spending across Region
* Fig 1.5  Total Spending across Channels & Regions
* Fig 1.6  Varieties of items across Region & Channel
* Fig 1.7  Outliers identification using boxplot
* Fig 1.8  Outliers for each variable across Channels & Regions
* Fig 2.1  Univariate Analysis using histogram
* Fig 2.2  Interaction between Application vs Accept
* Fig 2.3 Interaction between Application vs Enroll
* Fig 2.4 Interaction between Application vs Grad.Rate
* Fig 2.5 Top 10 Names with highest Application
* Fig 2.6 Top 10 Names with Highest Accept
* Fig 2.7 Top 10 Names with highest full time Undergraduates
* Fig 2.8 Top 10 Names with highest PHD ratio
* Fig 2.9 Pairplot
* Fig 2.10 Correlation Heatmap
* Fig 2.11 Outlier check using boxplot

## **List of Tables**

* Table 1.1 Sample of Dataset
* Table 1.2 Information about Dataset
* Table 1.3 Null values
* Table 1.4 Statistical summary of the dataset
* Table 1.5 Summary of the Spendings across Channels & Regions
* Table 1.6 Descriptive measure of variability
* Table 2.1 Sample of Dataset
* Table 2.2 Information about Dataset
* Table 2.3 Null values
* Table 2.4 Statistical summary of the dataset

**Problem Statement 1:**

A wholesale distributor operating in different regions of Portugal has information on 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 channel (Hotel, Retail).

**Data dictionary:-**

|  |  |  |
| --- | --- | --- |
| **Sr No** | **Column Name** | **Description** |
| 1 | Buyer/Spender | Buyer/Spender's id |
| 2 | Channel | Sales channel- Hotel or Retail |
| 3 | Region | Region of sales- Lisbon, Oporto, Other |
| 4 | Fresh | Variety of products |
| 5 | Milk | Variety of products |
| 6 | Grocery | Variety of products |
| 7 | Frozen | Variety of products |
| 8 | Detergents\_Paper | Variety of products |
| 9 | Delicatessen | Variety of products |

**Dimensions of the dataset**

* The dataset have 440 rows & 9 column

**First & last 5 rows of the dataset**

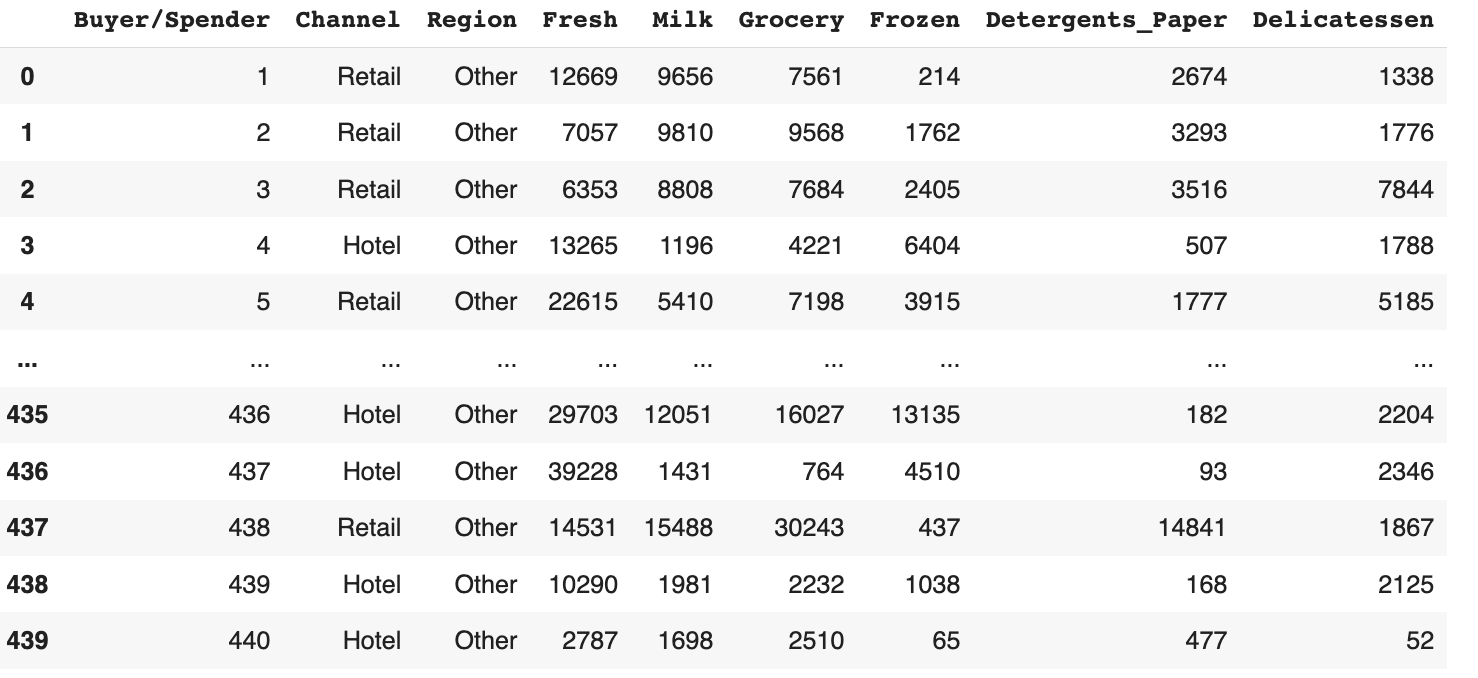


Table 1.1 Sample of Dataset

**Information about the dataset & datatype**

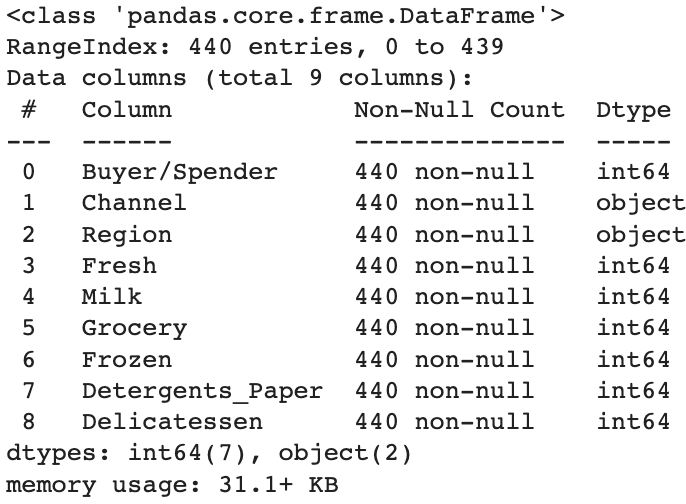
****

Table 1.2 Information about Dataset

* There are 7 Integer/Numeric and 2 Object data types
* All the data types are not null
* There seems to be no missing values

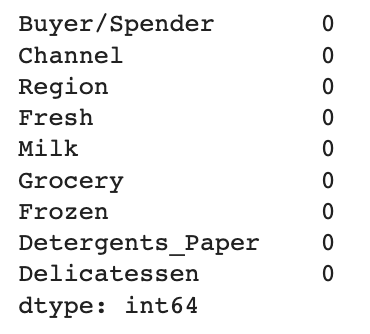


Table 1.3 Null values

**1.1.1 Use methods of descriptive statistics to summarize data.**

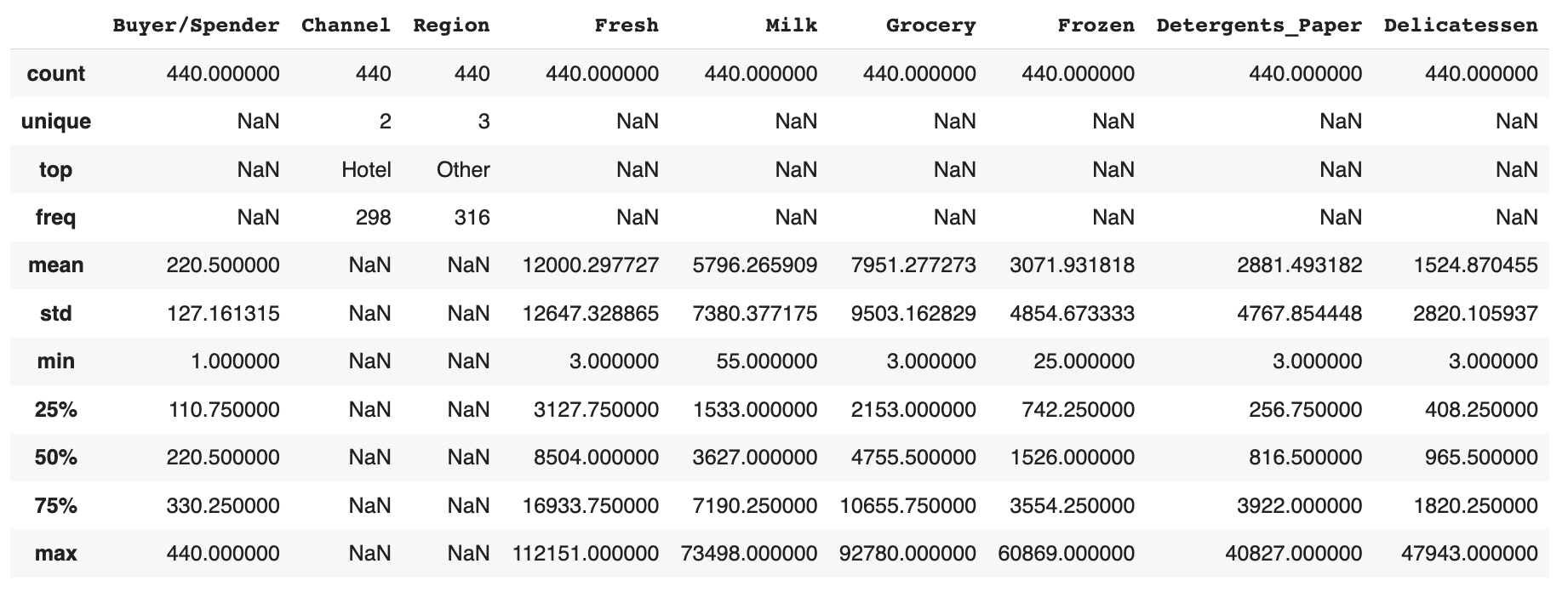
****

Table 1.4 Statistical summary of the dataset

* The dataset contains **440** entries representing buyers or spenders.
* The **majority** of buyers are associated with the **hotel** **channel**.
* **Most buyers** come from the "**Other**" region.
* Average spending varies across different product categories.
* **Fresh** products have an average spending of approximately **12,000** units.
* **Milk** and **grocery** items have average spendings of around **5,800** and **7,950** units, respectively.
* The dataset shows a wide range of values with some extreme values in certain categories.

**1.1.2 Which Region and which Channel spent the most?**

**1.1.3 Which Region and which Channel spent the least?**

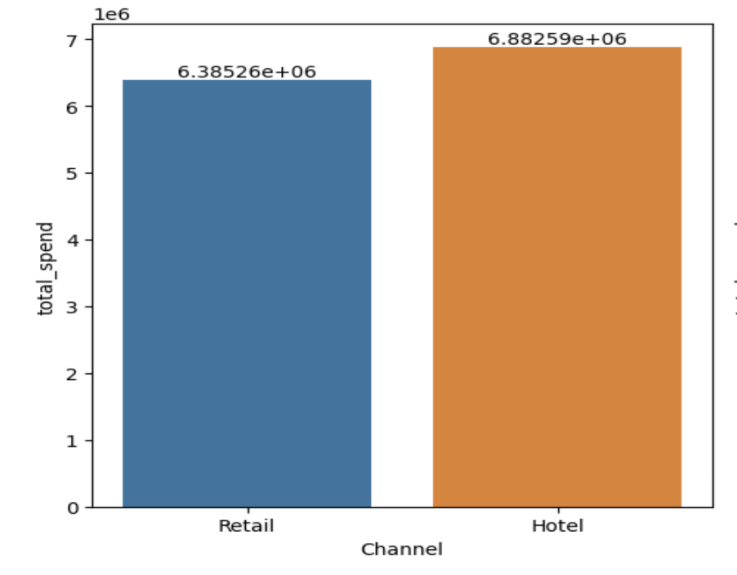
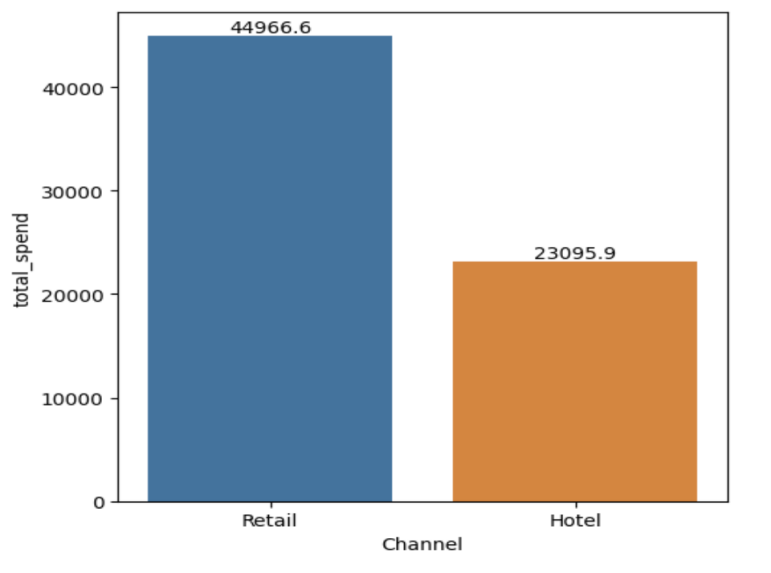
 

Fig 1.1 Total Spending across Channels Fig 1.2 Avg Spending across Channels

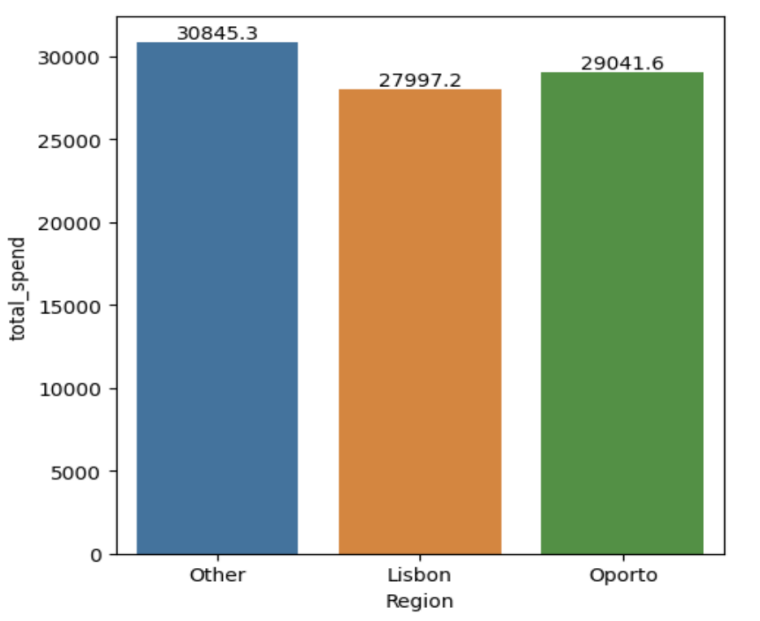
 

Fig 1.3 Total Spending across Region Fig 1.4 Avg Spending across Region

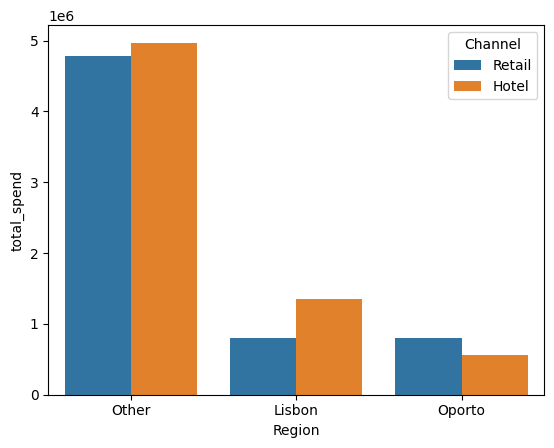


Fig 1.5 Total Spending across Channels & Regions

**Among the two Channels,**

* The **Hotel** channel had the **highest** total spending, amounting to **6,882,590.**
* On the other hand, the **Retail** channel had a relatively **lower** total spending of **6,385,260.**
* However, when we check **average spent**, **retailer** spends **more** than hotel
* **Hotels**, in general, **spend more than retailers**. However, in the **Oporto** region, **retailers spend more than hotels.**

**Among the Region,**

* The **Other** region has the **highest** total spending among the three, amounting to **9,747,107.**
* The **lowest** is for the **Oporto** region having a total spending of **1,364,956.**
* On an **average** all regions spend almost same **Lisbon** being the **least**

**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.**

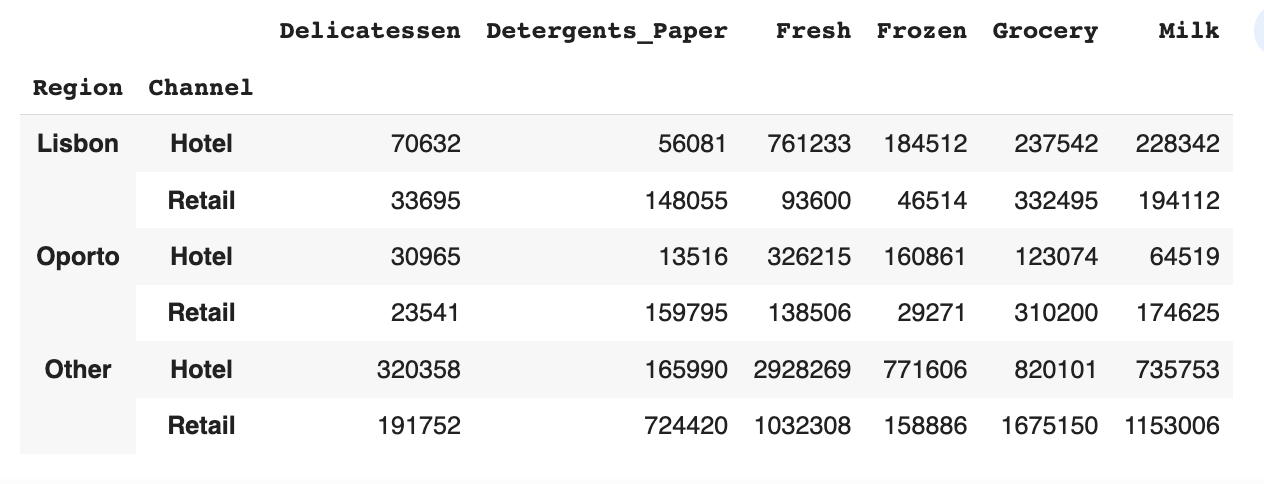
****

Table 1.5 Summary of the Spendings across Channels & Regions

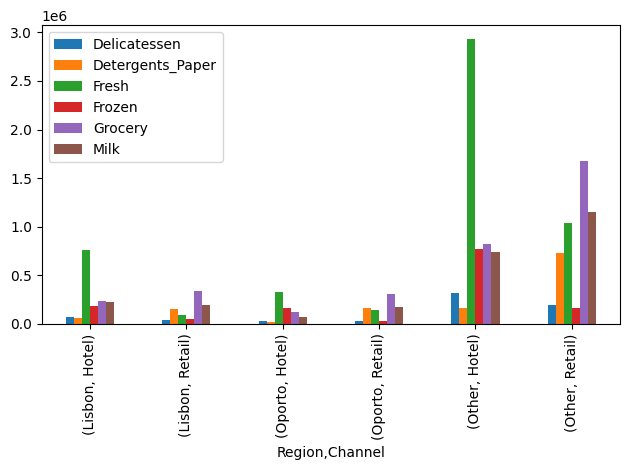


Fig 1.6 Varieties of items across Region & Channel

* **Fresh** category has **high** spending across all regions and channels.
* **Retailers** in **Lisbon and Oporto** spend **more on Grocery** items compared to **hotels**.
* **Hotels** in the **Other region** have significantly **higher spending** across all categories.
* **Retailers** have **higher spending** on **Detergents\_Paper** compared to **hotels**.
* **Hotels** in the **Other region** spend significantly **more on Delicatessen**.

**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?**

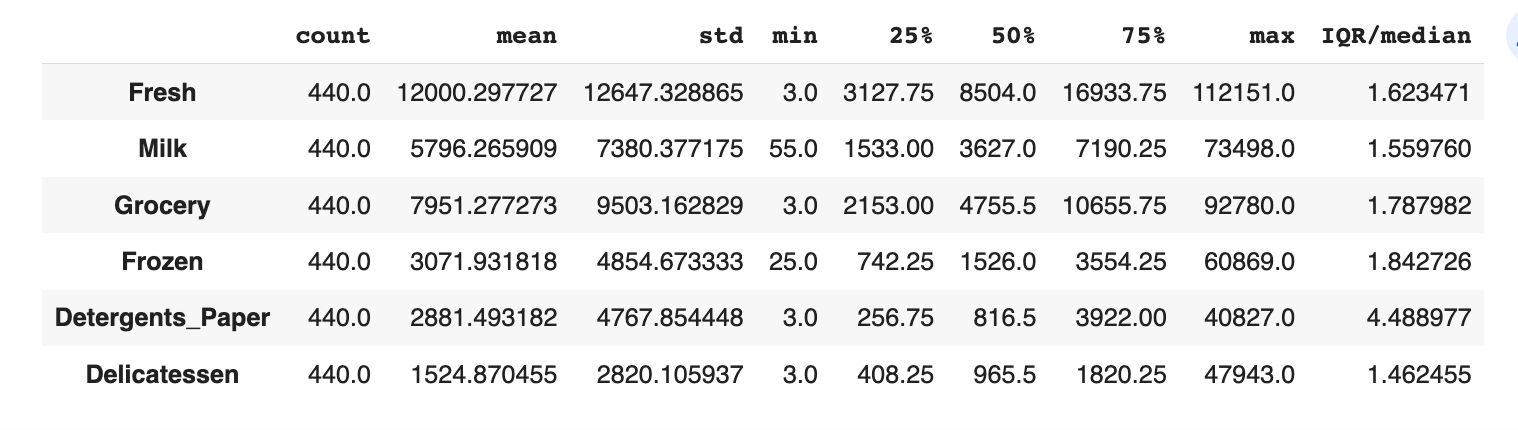
****

Table 1.6 Descriptive measure of variability

* The item **Detergents\_Paper** exhibits the **highest** level of inconsistency, with a relatively high IQR/median value of **4.488977.**
* **Delicatessen** and **Milk** show the **least** inconsistency, with lower IQR/median values of **1.462455** and **1.559760** respectively.
* **Higher** IQR/median values indicate **greater variability** and **inconsistency** in the data, while lower values suggest more consistency and less spread.

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

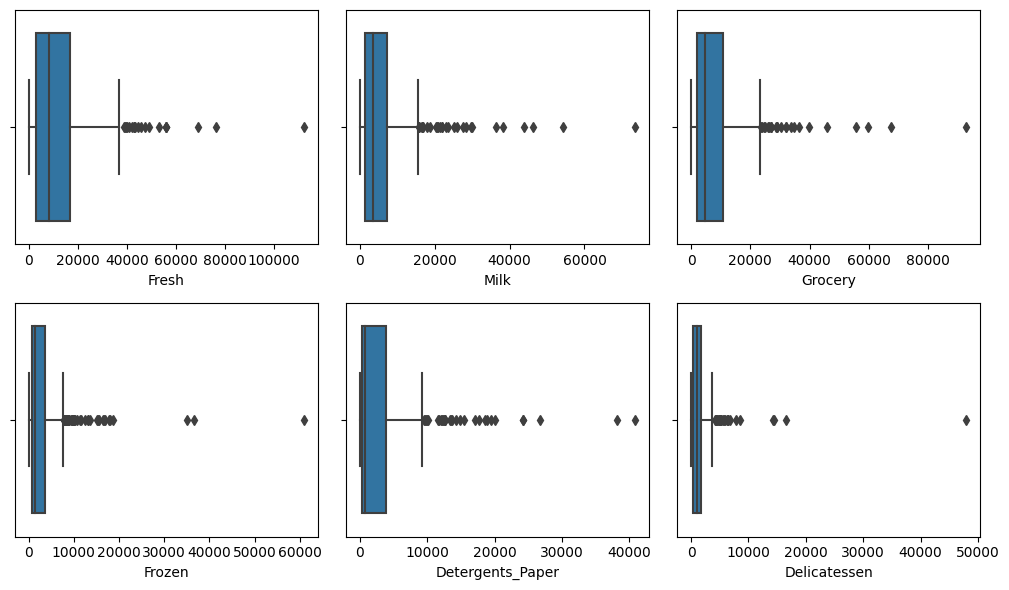
****

Fig 1.7 Outliers identification using boxplot

* From above, we can see the presence of outliers across all varieties.

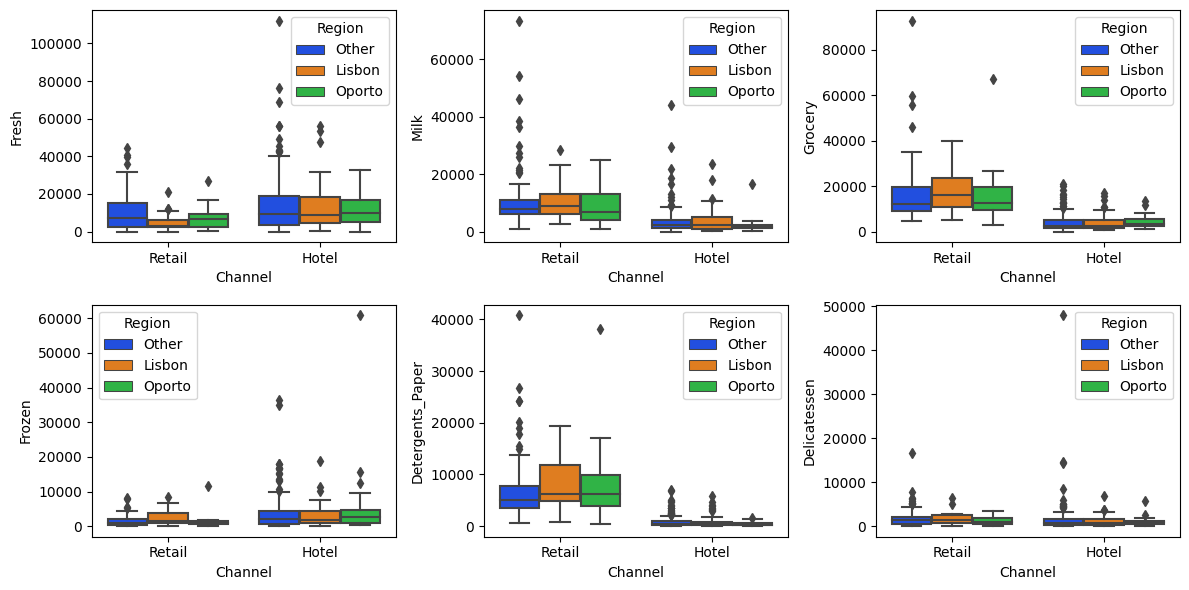


Fig 1.7 Outliers for each variable across Channels & Regions

**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.**

1. **Retailers spend more than hotels**. This means that we can increase our revenue by targeting more retailers. We can do this by expanding our reach to new retailers or by increasing our marketing efforts to existing retailers.
2. **There are customers in other regions**. Currently, we are only focused on Oporto and Lisbon. However, there are customers in other regions as well. We should focus on expanding our reach to these regions.
3. **Hotels buy fresh food, retailers buy grocery & milk**. This means that we can maximize our sales by focusing on these items in each channel. For example, we can offer discounts on fresh food to hotels or we can offer loyalty programs for grocery & milk purchases to retailers.
4. **Frozen foods, delicatessen & detergents\_paper are least popular**. This means that we may not be able to sell these items as well in the current channels. We can look for other channels or industries that have demand for these items. For example, we could sell frozen foods to restaurants or delicatessen items to grocery stores.

Problem Statement 2:

The dataset [Education - Post 12th Standard.csv](https://olympus.mygreatlearning.com/courses/87091/files/7283009/download?wrap=1) contains information on various colleges. The data dictionary of the 'Education - Post 12th Standard.csv' can be found in the following file: [Data Dictionary.xlsx.](https://olympus.mygreatlearning.com/courses/87091/files/7283010/download?wrap=1)

**Data Dictionary**

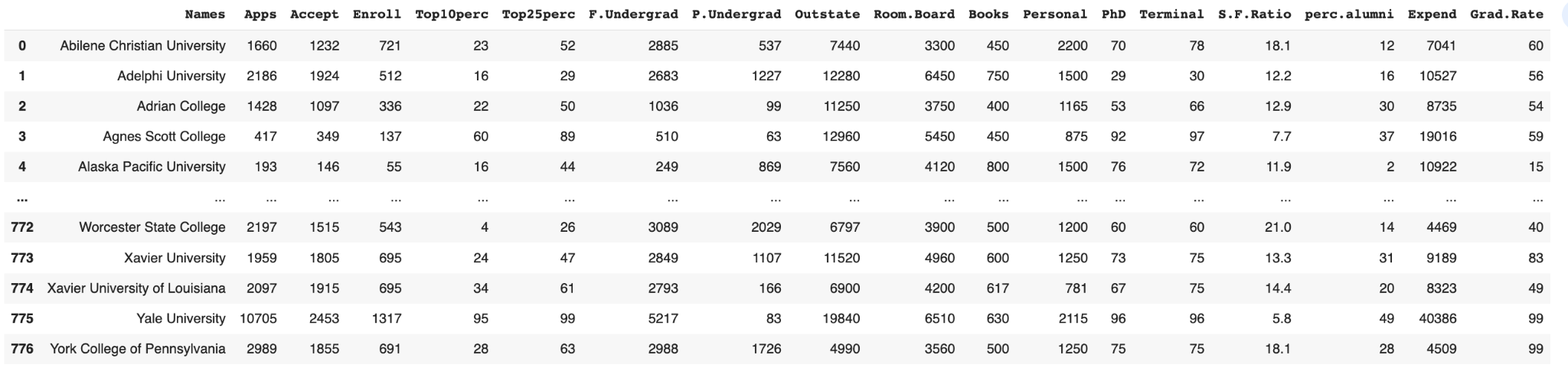
1. Names: Names of various university and colleges
2. Apps: Number of applications received
3. Accept: Number of applications accepted
4. Enroll: Number of new students enrolled
5. Top10perc: Percentage of new students from top 10% of Higher Secondary class
6. Top25perc: Percentage of new students from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

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

**Dimensions of the dataset**

* The dataset have 777 rows & 18 column

**First & last 5 rows of the dataset**

Table 2.1 Sample of Dataset

**Information about the dataset & datatype**

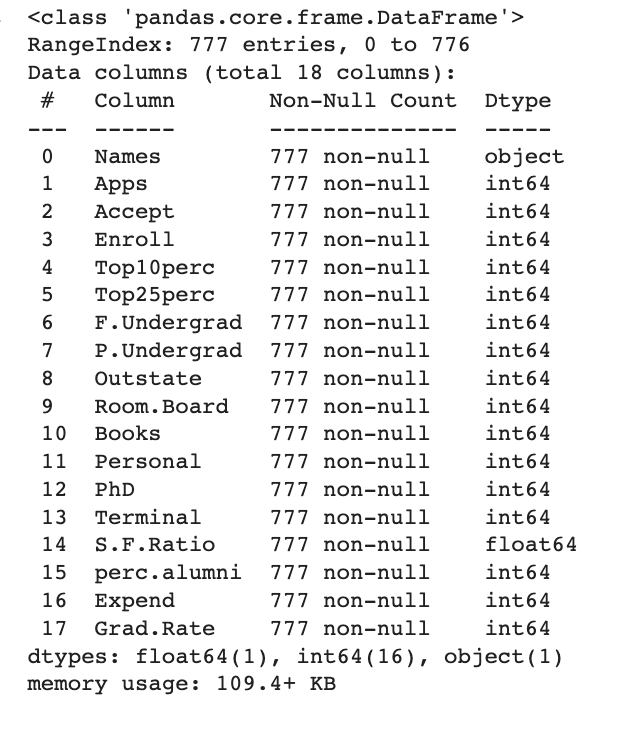
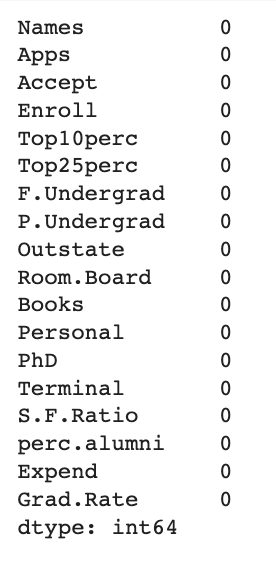
**** 

Table 2.2 Information about Dataset Table 2.3 Null values

* The dataset provides insights into applications, acceptance, enrollment, student performance, expenses, faculty qualifications, and more.
* The dataset contains information on **777** educational institutions.
* It consists of **18 columns** representing various aspects of these institutions.
* The data is complete, with **no missing** values.
* There are **no duplicate values**

**Descriptive statistics to summarize data**

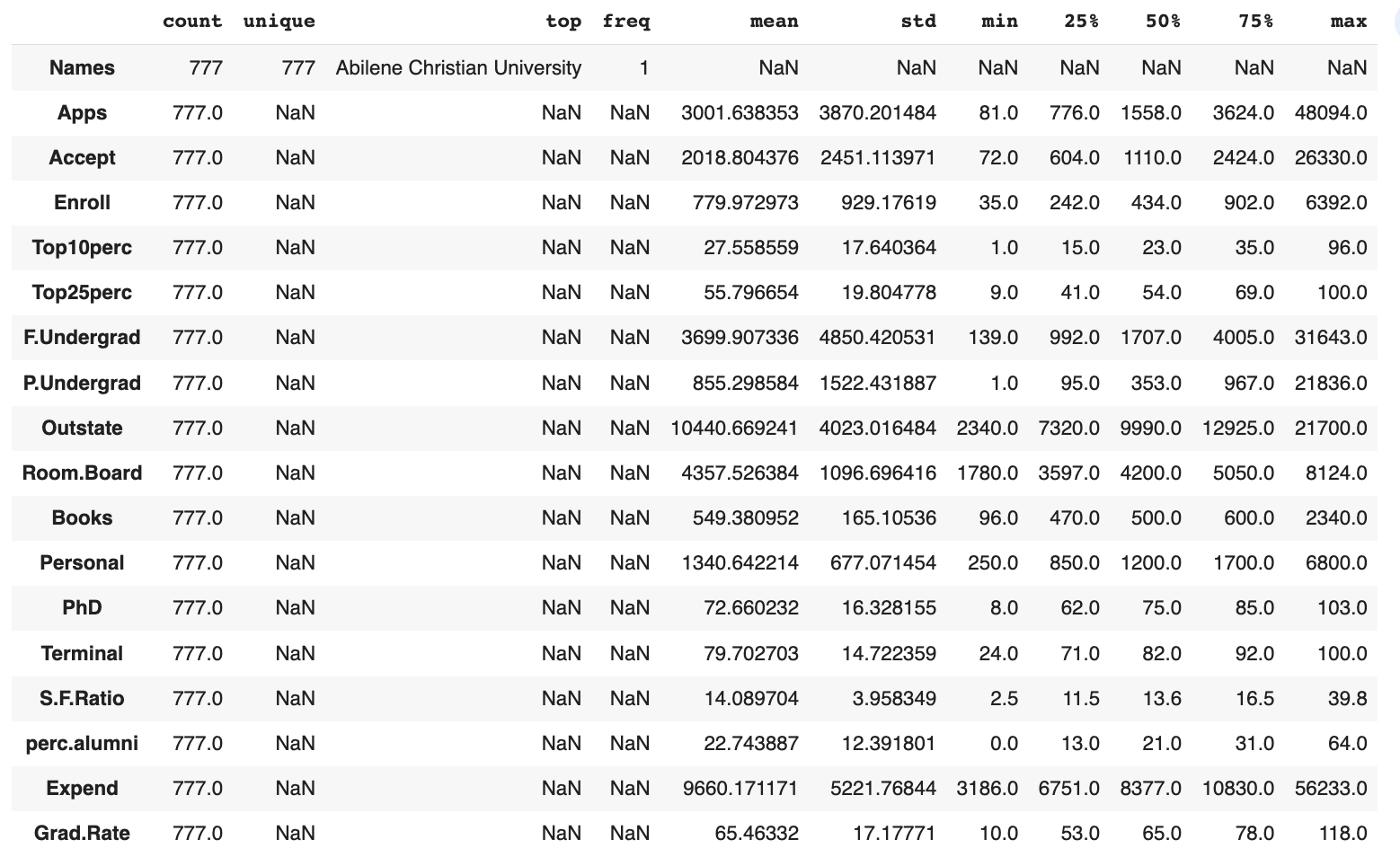
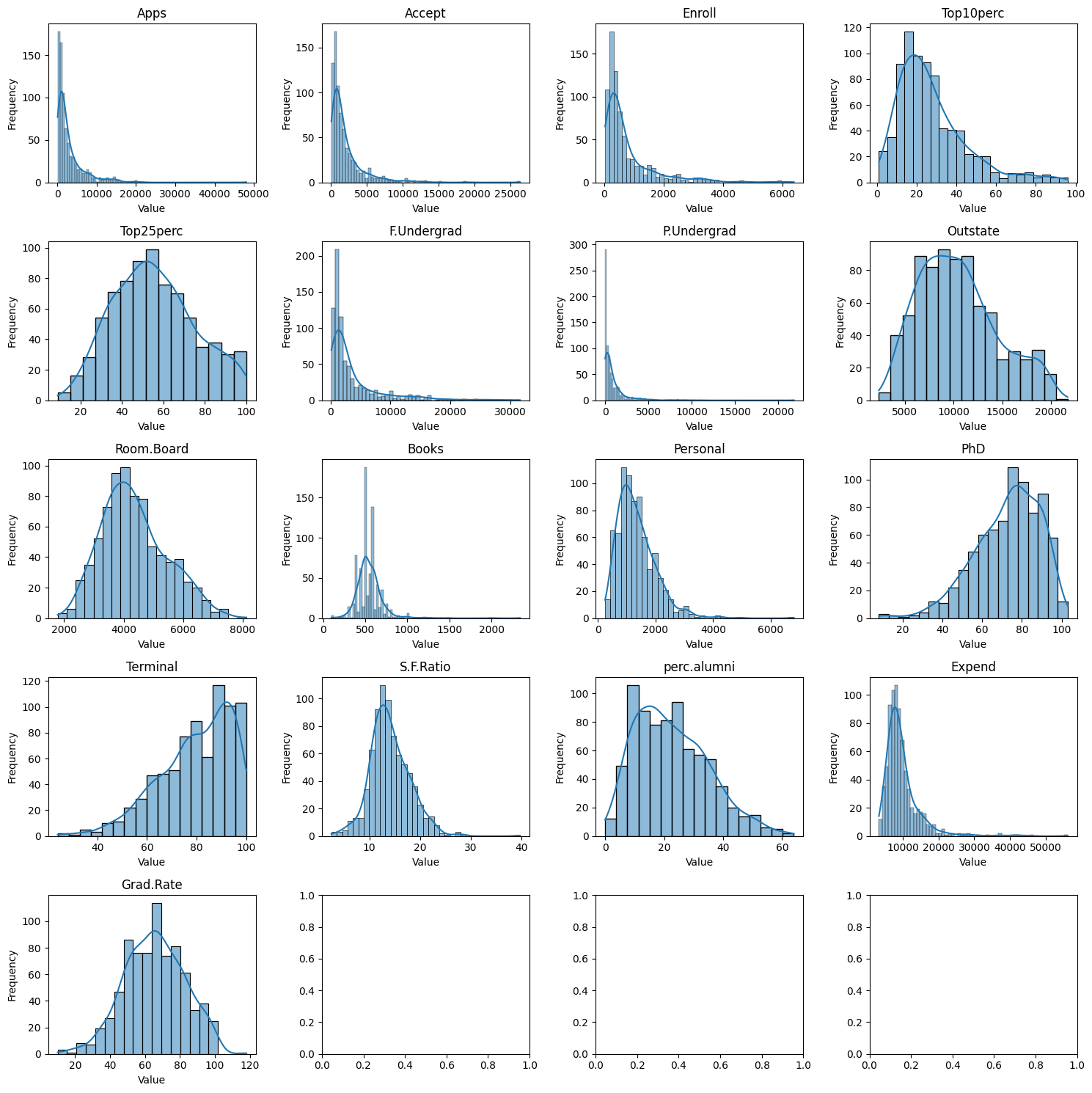
****

Table 2.4 Statistical summary of the dataset

* The dataset contains information on **777** educational institutions, providing a comprehensive view of the higher education landscape.
* The **average** number of **applications** received by institutions is **around** **3001**, indicating a competitive admissions process.
* The **mean** student-to-faculty ratio is approximately **14**, suggesting a **relatively balanced** **student-teacher interaction** in these institutions.
* The **average graduation rate** across the institutions is **approximately 65%,** reflecting the percentage of students who **successfully complete their degree programs**.
* The **average instructional expenditure** per student is **around $9660**, indicating the financial resources allocated to supporting education and academic programs in these institutions.

**Univariate Analysis**

Fig 2.1 Univariate Analysis using histogram

**Basis the above univariate Analysis:**

* Most of the columns such as top25perc, top10perc, outstate, Room.Board, Books, Personal tend to relatively distribute **normally** with **skewness** in some.
* Variables such as Apps, Accept, Enroll etc are highly skewed towards left while Terminal, PHD etc are skewed towards right.

**Bivariate Analysis**

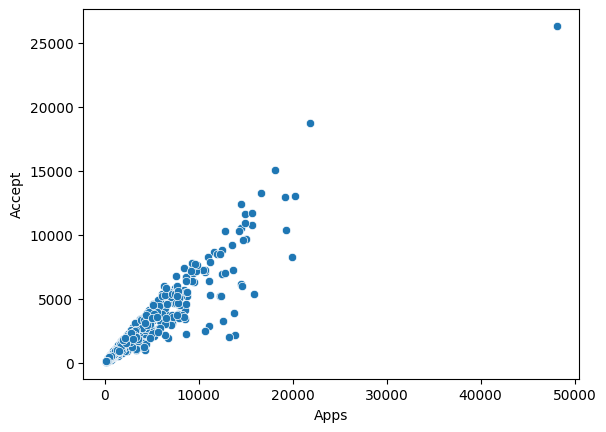
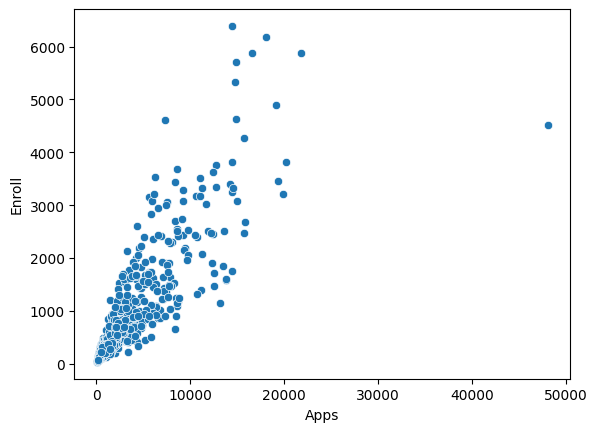
** **

Fig 2.2 Interaction between Application vs Accept Fig 2.3 Interaction between Application vs Enroll

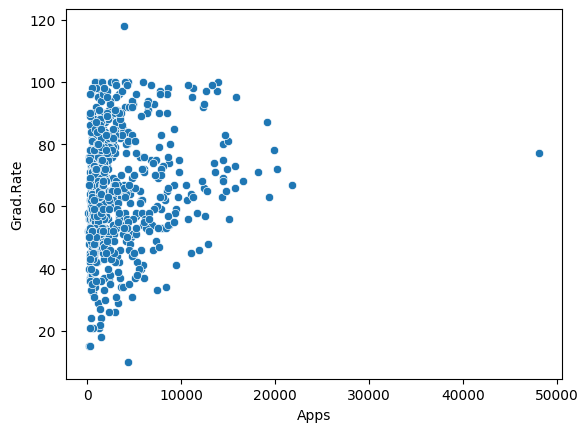
****

Fig 2.4 Interaction between Application vs Grad.Rate

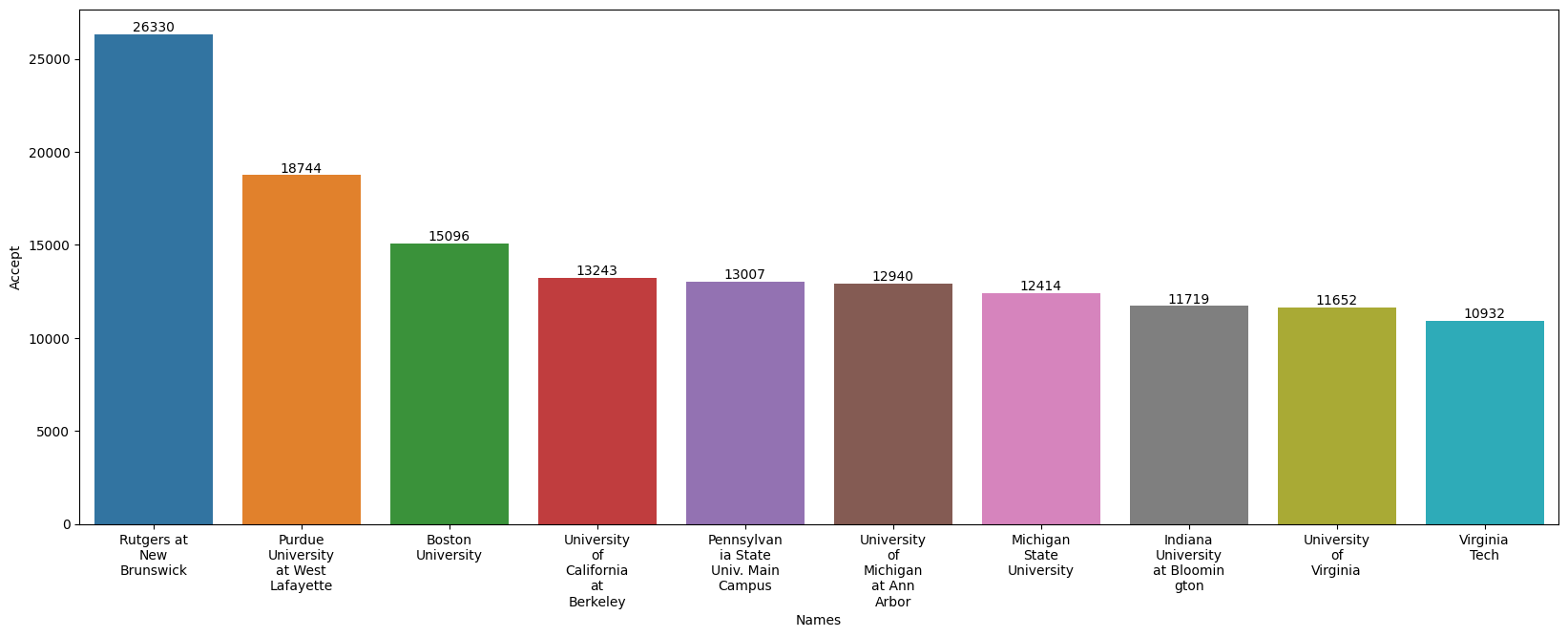
****

Fig 2.5 Top 10 Names with highest Application

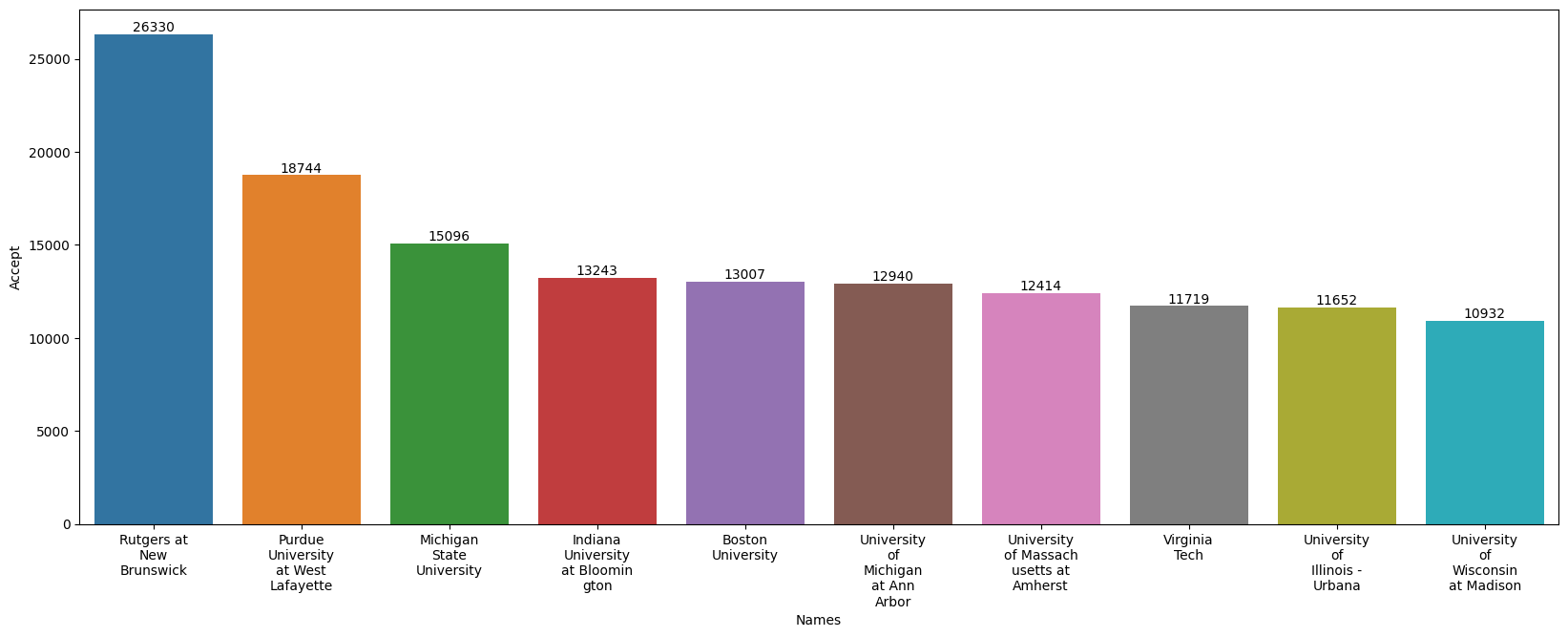


Fig 2.6 Top 10 Names with Highest Accept

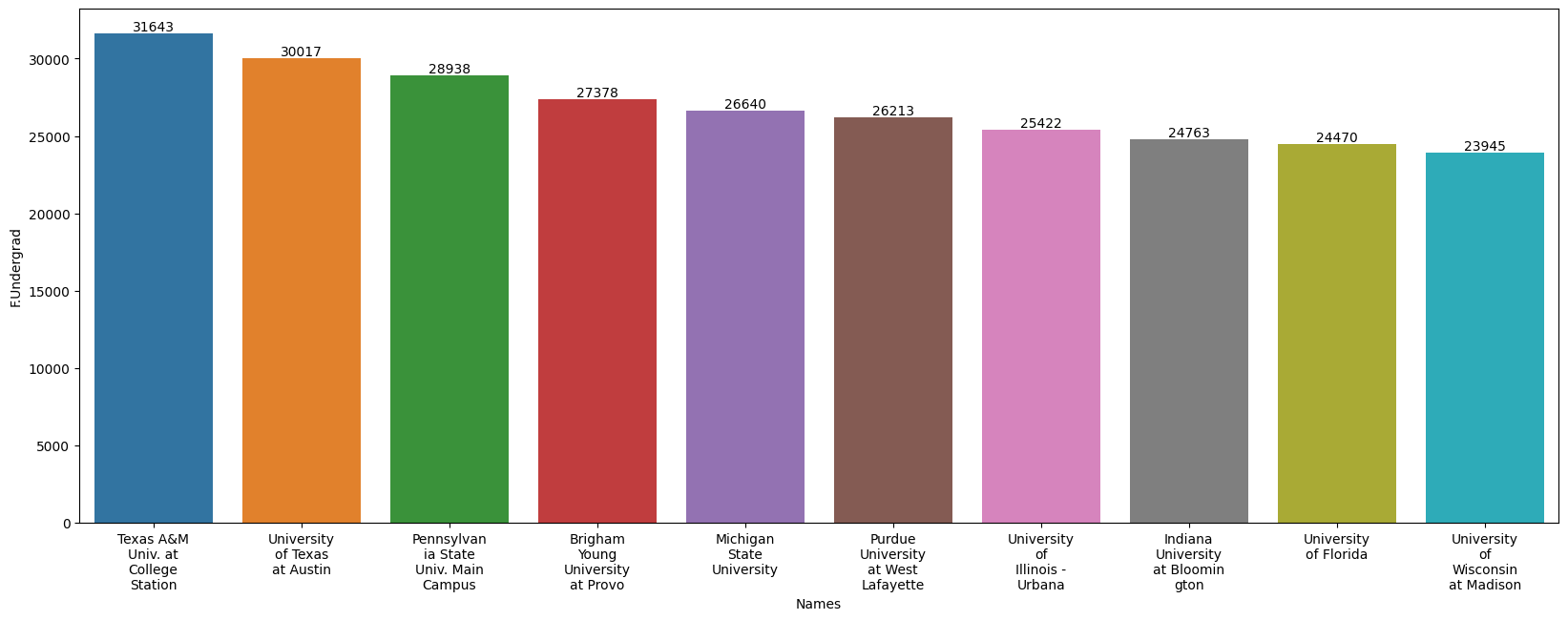


Fig 2.7 Top 10 Names with highest full time Undergraduates

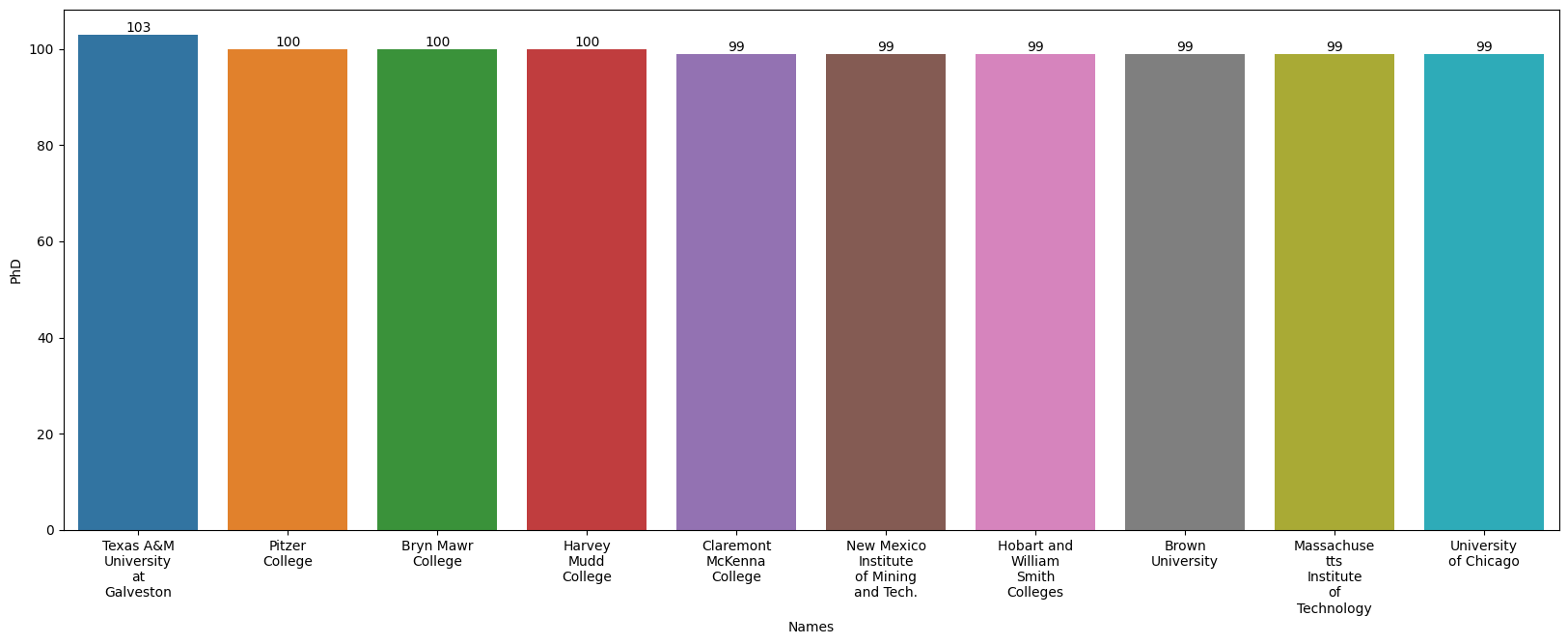
****

Fig 2.8 Top 10 Names with highest PHD ratio

**Multivariate Analysis**

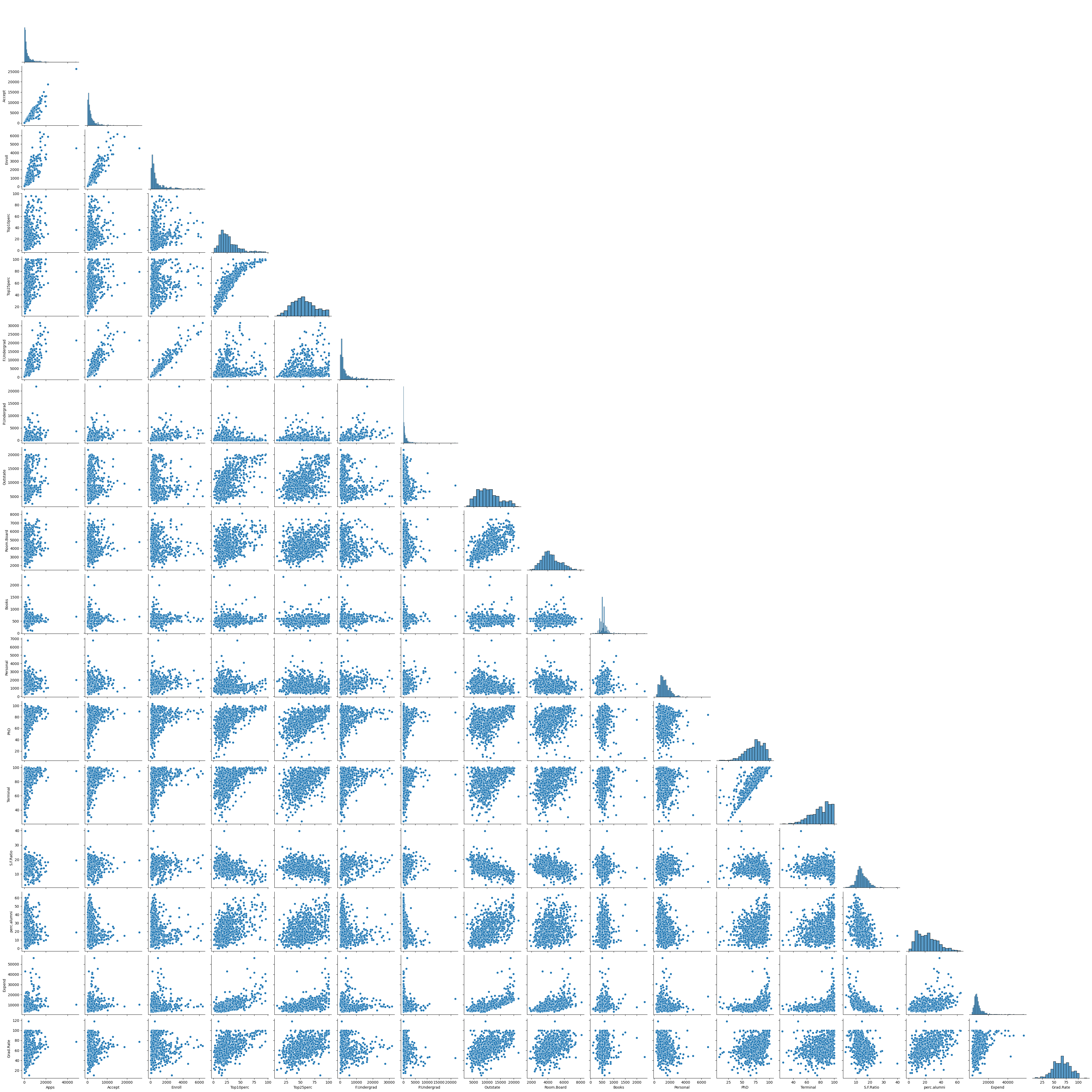


Fig 2.9 Pairplot

**Correlation Heatmap**

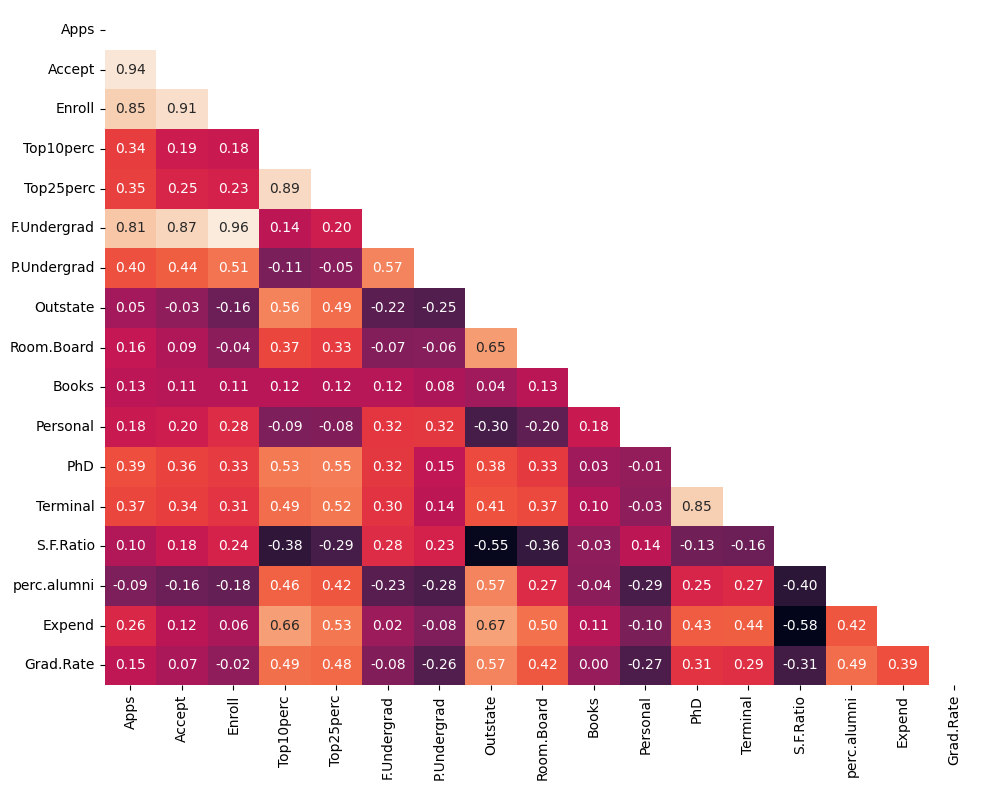
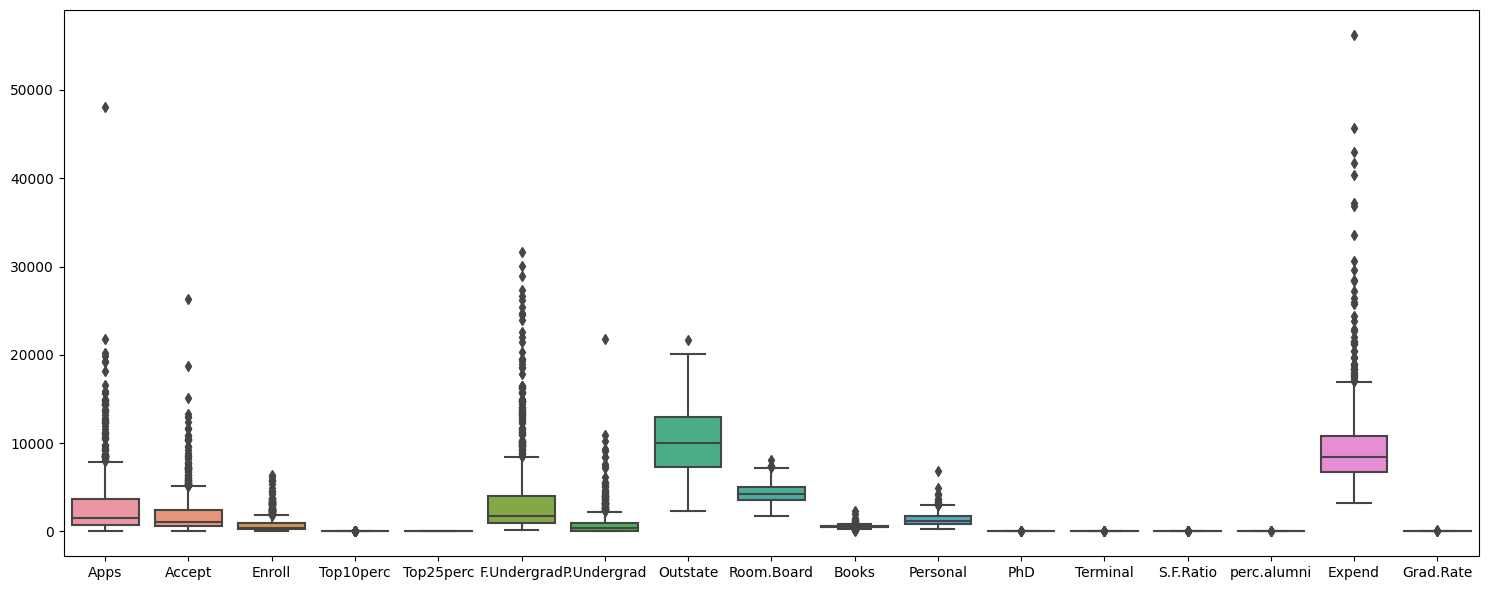
****

Fig 2.10 Correlation Heatmap

**Basis the Multivariate Analysis above, we can infer that:-**

* There is a **strong positive correlation** between the **number of applications, acceptances, and enrollments.**
* The presence of t**op-performing students** is associated with **higher** values in various variables.
* **Higher student-to-faculty ratios** are linked to **lower out-of-state tuition**, **room** and **board expenses, expenditure per student, and graduation rates**.
* Institutions with a **higher percentage** of **faculty** holding **PhD** and **terminal degrees** tend to have a **strong positive correlation**.
* **Higher instructional expenditures** per student are associated with **increased out-of-state tuition, room and board expenses, spending on books, personal expenses, faculty with advanced degrees, and graduation rates.**



**Outlier Check Using Boxplot**

Fig 2.11 Outlier check using Boxplot

* Outliers are present in all variables except for **Top25perc**
* The outliers can be treated as per the business use case by different methods such as removal, imputing the outliers with the mean, median, or mode of the variable, Winsorization etc basis the data.

**Overall Insights:**

* The dataset provides insights into **applications, acceptance, enrollment, student performance, expenses, and faculty qualifications** among **777** **educational institutions**.
* The dataset is **complete**, with **no missing** values and **no duplicate** values.
* The **average** **number of applications** received by institutions is around **3001**, indicating a competitive admissions process.
* The **mean student-to-faculty ratio** is approximately **14**, suggesting a balanced student-teacher interaction.
* The **average graduation rate** across institutions is around **65%,** reflecting successful degree program completion.
* The **average instructional expenditure** per student is approximately **$9660**, indicating the allocation of financial resources to support education.
* Most variables follow a **relatively normal distribution** with some skewness.
* The analysis shows a **strong positive correlation between applications, acceptances, and enrollments.**
* **Higher student-to-faculty ratios** are linked to lower tuition, room and board expenses, expenditure per student, and graduation rates.
* Institutions with a **higher percentage** of **faculty holding PhD** and **terminal** degrees have a strong positive correlation.
* **Higher instructional expenditures** per student are associated with **increased tuition, room and board expenses, spending on books, personal expenses, faculty with advanced degrees, and graduation rates.**
* **Outliers are present** in all variables **except for Top25per**.

**End of Report**