PDS Extended Project

GUIDED

LIST OF CONTENT

PROBLEM 1

WHOLESALE CUSTOMER EXPERIENCE

DATA DESCRIPTION & It's DICTIONARY

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 channels (Hotel, Retail).

DATA OVERVIEW:

The data set contains 440 entries with the following columns.

1. Buyer/Spender: IDs of customers
2. Channel: Sales channel (hotel/retail)
3. Region: Region of the distributor (Lisbon, Oporto, others)
4. Fresh: Spending on fresh vegetables
5. Milk: Spending on milk
6. Grocery: Spending on grocery
7. Frozen: Spending on frozen food
8. Detergents_paper: Spending on detergents and toilet paper
9. Delicatessen: Spending on instant food

DATA CLEANING AND PREPARATION:

- There are no missing values in the data set.
- There are no duplicate rows in the data set.
- The buyer/ spender column has been dropped as it is an ID and not relevant for the analysis.

EXPLORATORY DATA ANALYSIS (EDA)

Based on the EDA performed on the data set, exploring all categorical variables and observations on their frequency.

UNIVARIATE ANALYSIS

- **Region vs Number of wholesale distributors**

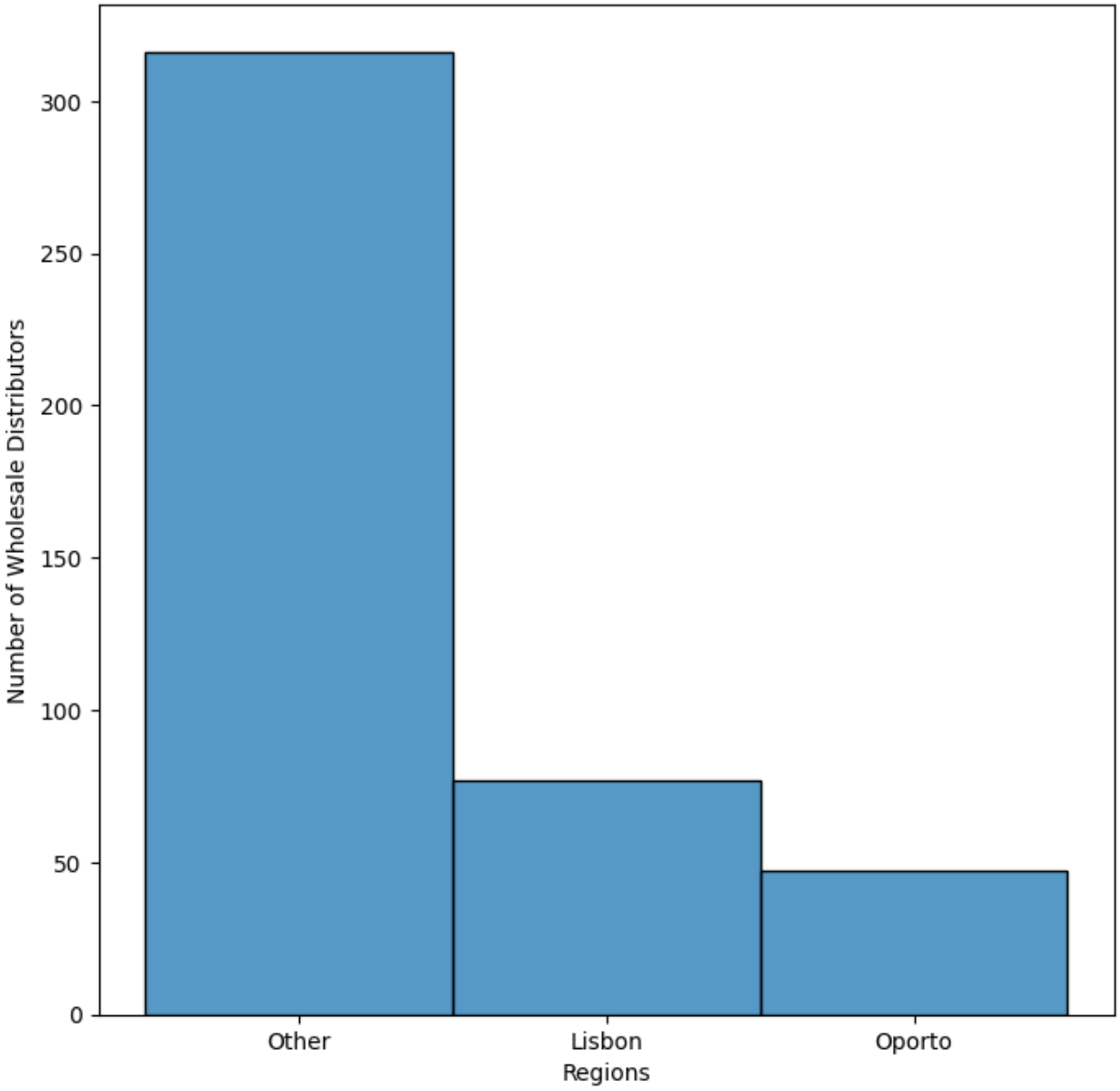

FIGURE 1

- From the above histogram we can see that the least number of wholesale distributors is in Oporto which is around 50 in count
- Lisbon has around 75 wholesale distributors.

- Whereas the maximum number of wholesale distributors are in the ‘other’ region.

- **Channel vs Number of wholesale distributors**

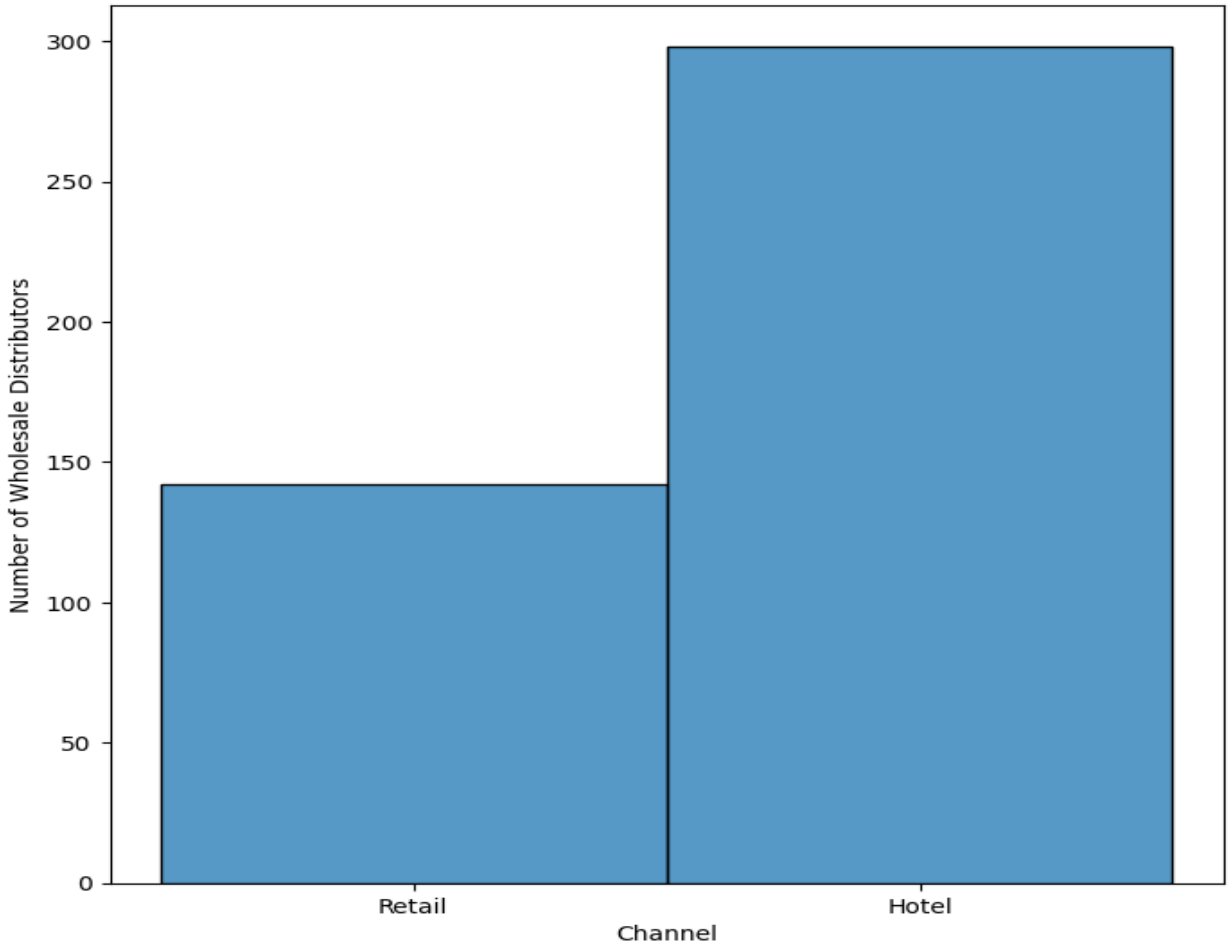

FIGURE 2

- The number of wholesale distributors in the retail channel is around 145 in number.
- Whereas the hotel has around 300 wholesale distributors.

- This shows that a vast number of wholesale distributors come from the hotel channel.

- **Distribution of spending across all categories**

The histogram provides visual representation of the distribution of the annual spending of fresh, milk, grocery, frozen, detergent_paper and delicatessen.

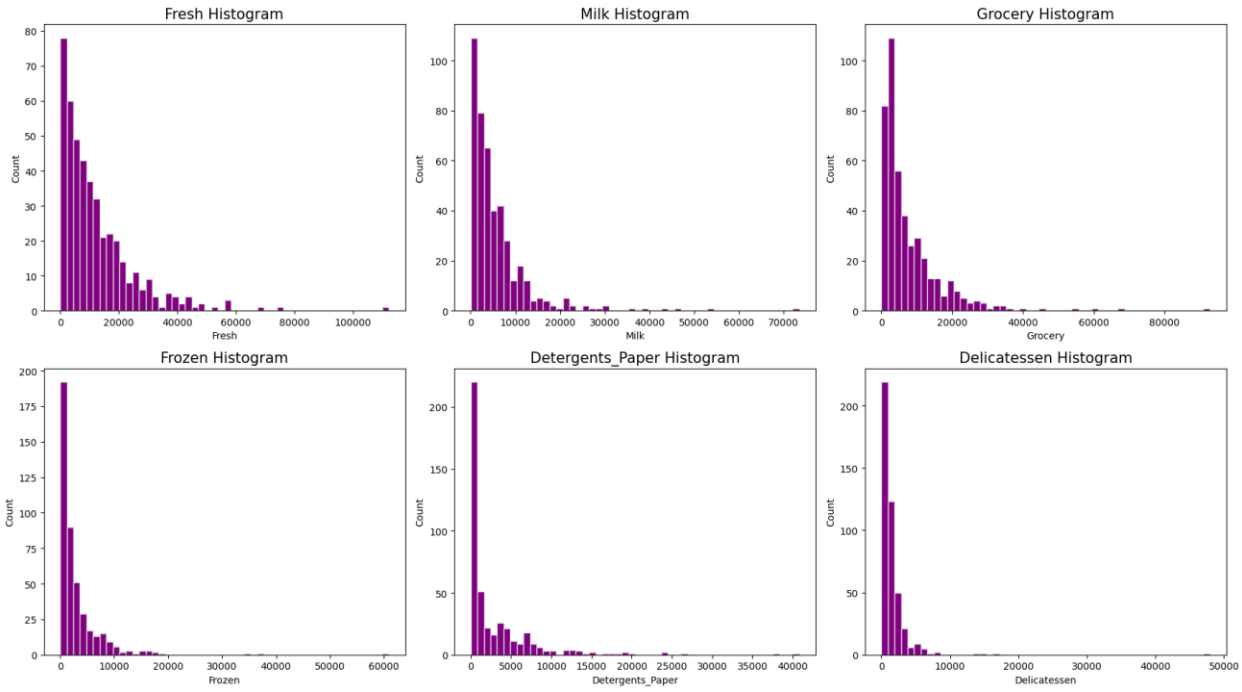

FIGURE 3

a) Fresh:

- The spending on fresh products is highly right skewed.

- Most retailers spend a small amount on fresh products with few spending higher amounts.
- Many customers might be small to medium sized retailers.

b) Milk:

- The spending on milk products is moderately right skewed.
- There is a concentration of retailers with lower spending.
- A good number of retailers spend moderately to highly on milk products.

c) Grocery:

- The spending on the grocery appears to have a binomial distribution.
- There are 2 groups of retailers: one with low spending and another with moderate to high spending on groceries.

d) Frozen:

- The spending on frozen products is rightly skewed.
- Most retailers spend less on frozen products with a few exceptions spending more.

e) Detergent_Paper:

- The plot shows a strong right skewness.
- Mostly retailers have low spending on detergent_paper.
- A small number of retailers prefer spending a large amount.

f) Delicatessen:

- The histogram shows the plot is right skewed.
- There is a significant number of retailers with low spending.
- But also, a good number of moderate to higher spenders.

- **Boxplot analysis**

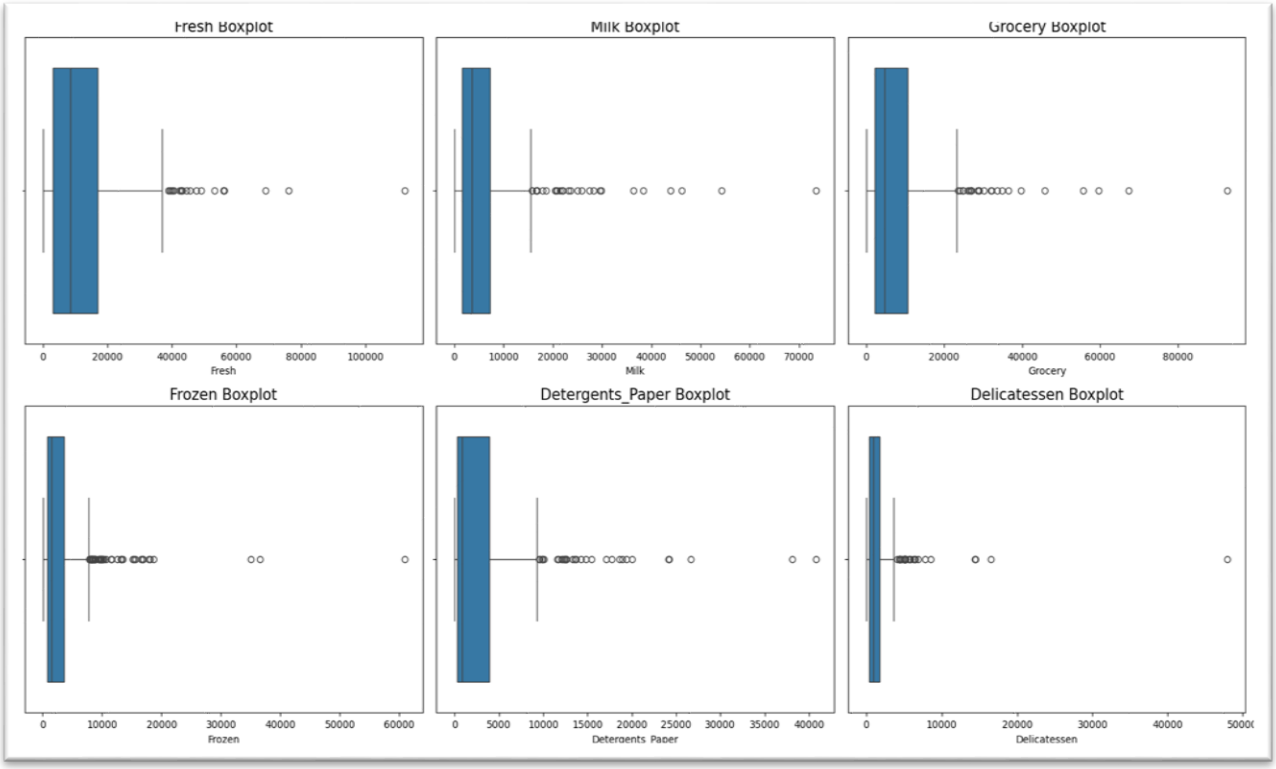

FIGURE 4

a) Fresh:

- The spending on fresh products has a higher Inter Quartile Range (IQR)
- Several outliers are present.
- Outliers indicate that some retailers spend significantly more than the rest.

b) Milk:

- The spending on milk products shows a high IQR with several outliers.
- The significant spending shows a diverse purchasing behavior among the retailers.

c) Grocery:

- Spending on groceries has a moderate IQR with numerous outliers.
- There is mixed spending nature i.e. some spending much than the rest.

d) Frozen:

- The spending on frozen products shows a lower IQR with significant outliers.

- Most retailers spend within a lower range with few spending much higher amounts.

e) Detergent_Paper:

- The spending on Detergent_paper has a lower IQR with some outliers.
- There are few high spenders.

f) Delicatessen:

- The spending on delicatessen has a moderate IQR.
- Many outliers are present.
- While most spend within a moderate range, there are several who spend significantly more.

- **Boxplot analysis: <u>Total spending by region</u>**

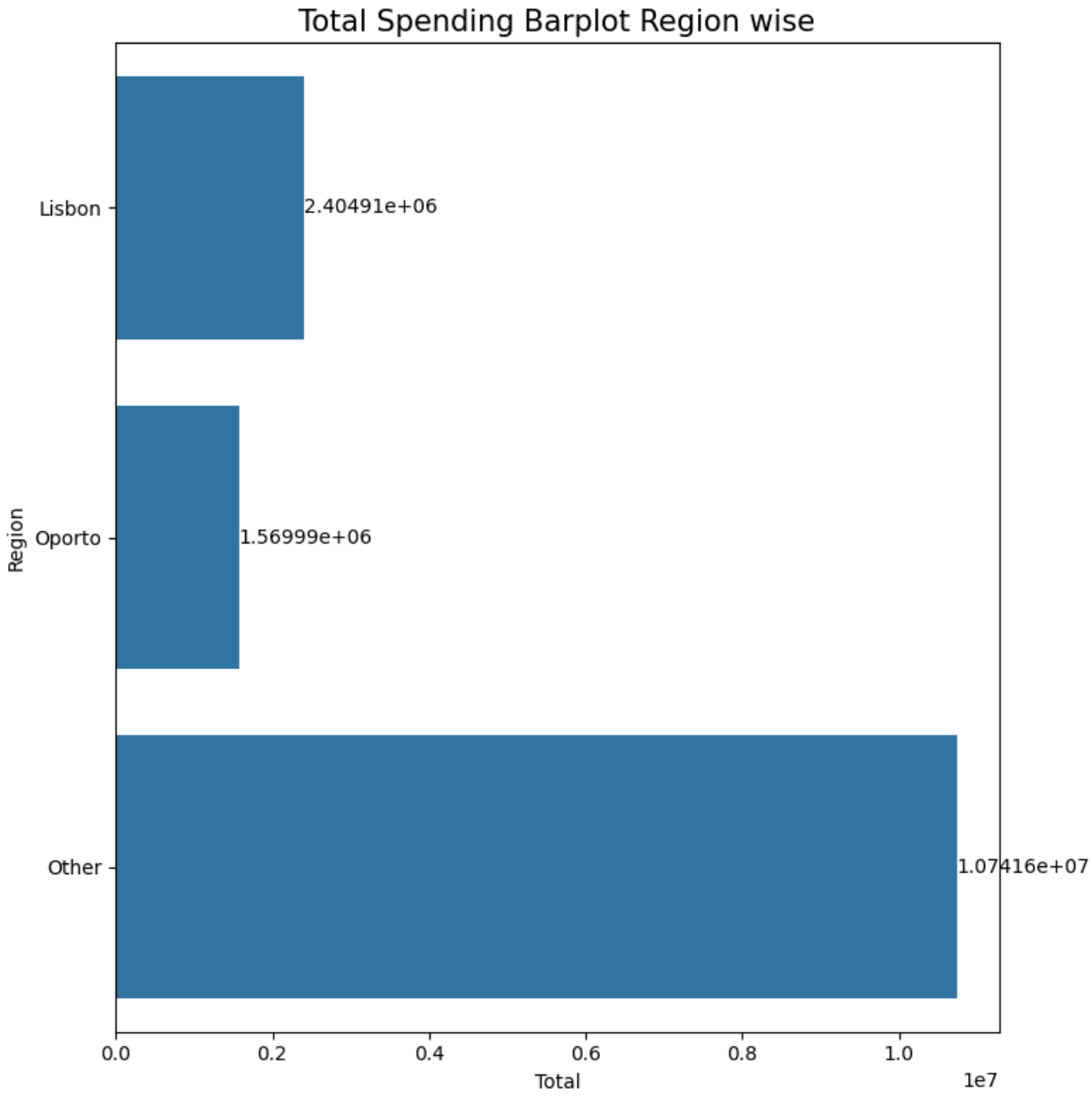

FIGURE 5

1. Lisbon:

- Lisbon shows the highest spending among all the regions.
- This shows that Lisbon is a major market for the wholesale distributor.

- Retailers in Lisbon contribute to the overall sales.

2. Oporto:

- Oporto shows moderate total spending
- Oporto has lower total spending as compared to Lisbon.

3. Other:

- The ‘other’ category shows the lowest total spending.
- These regions contribute the least to overall sales, showing a smaller market presence or low spending capacity.

Bar Plot Analysis: <u>Total spending by Channel:</u>

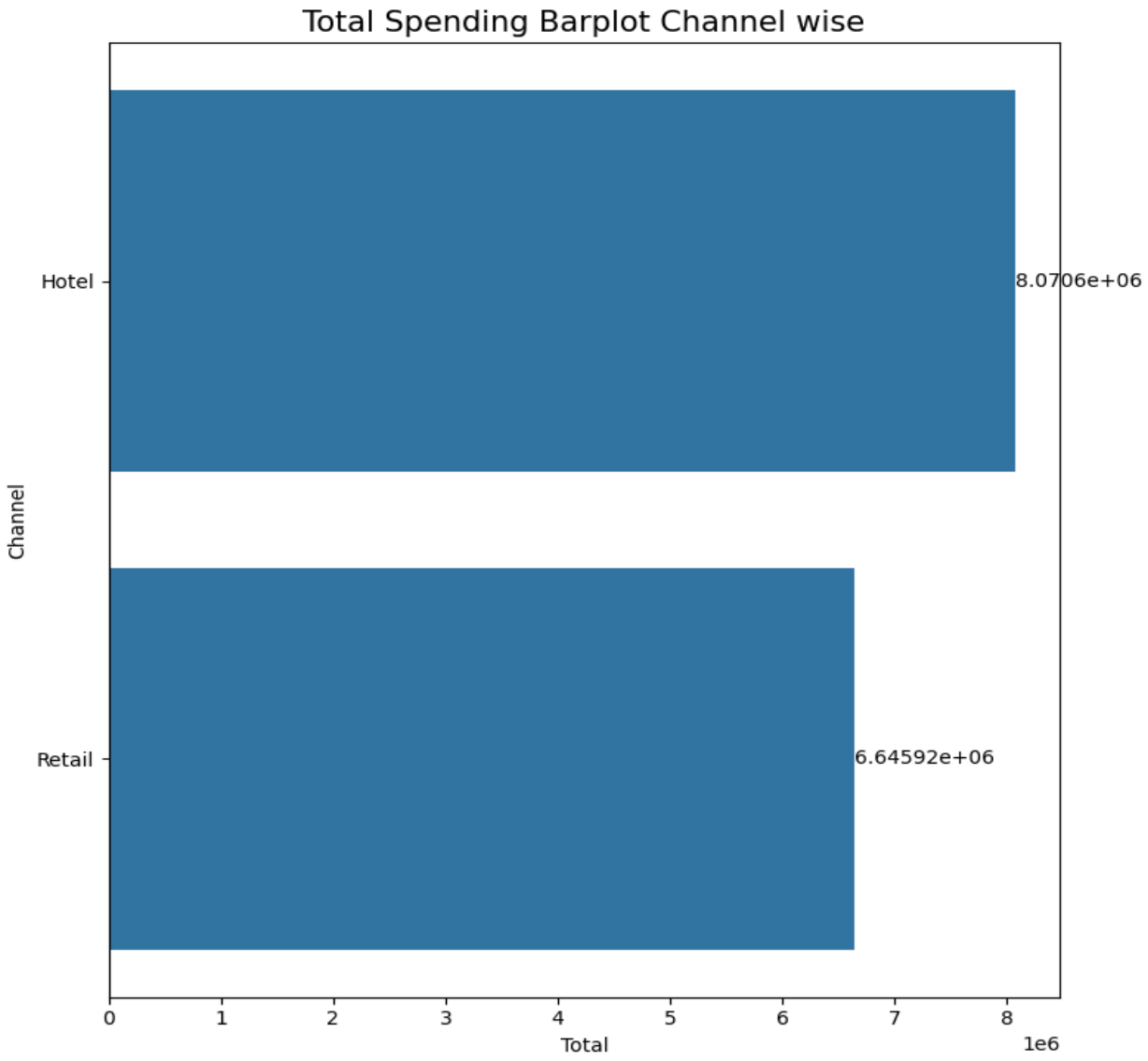

FIGURE 6

1. Retail Channel:

- The retail channel shows higher total spending as compared to the hotel channel.
- This shows that the retail channel is a major contributor to the overall sales of the wholesale distributor.
- Retailers tend to spend more.

2. Hotel Channel:

- The hotel channel shows the lower total spending compared to the retail channel.
- Hotels contribute less to the overall sales, due to the small purchase volumes.

Box Plot Analysis: Total spending by Region & Channel

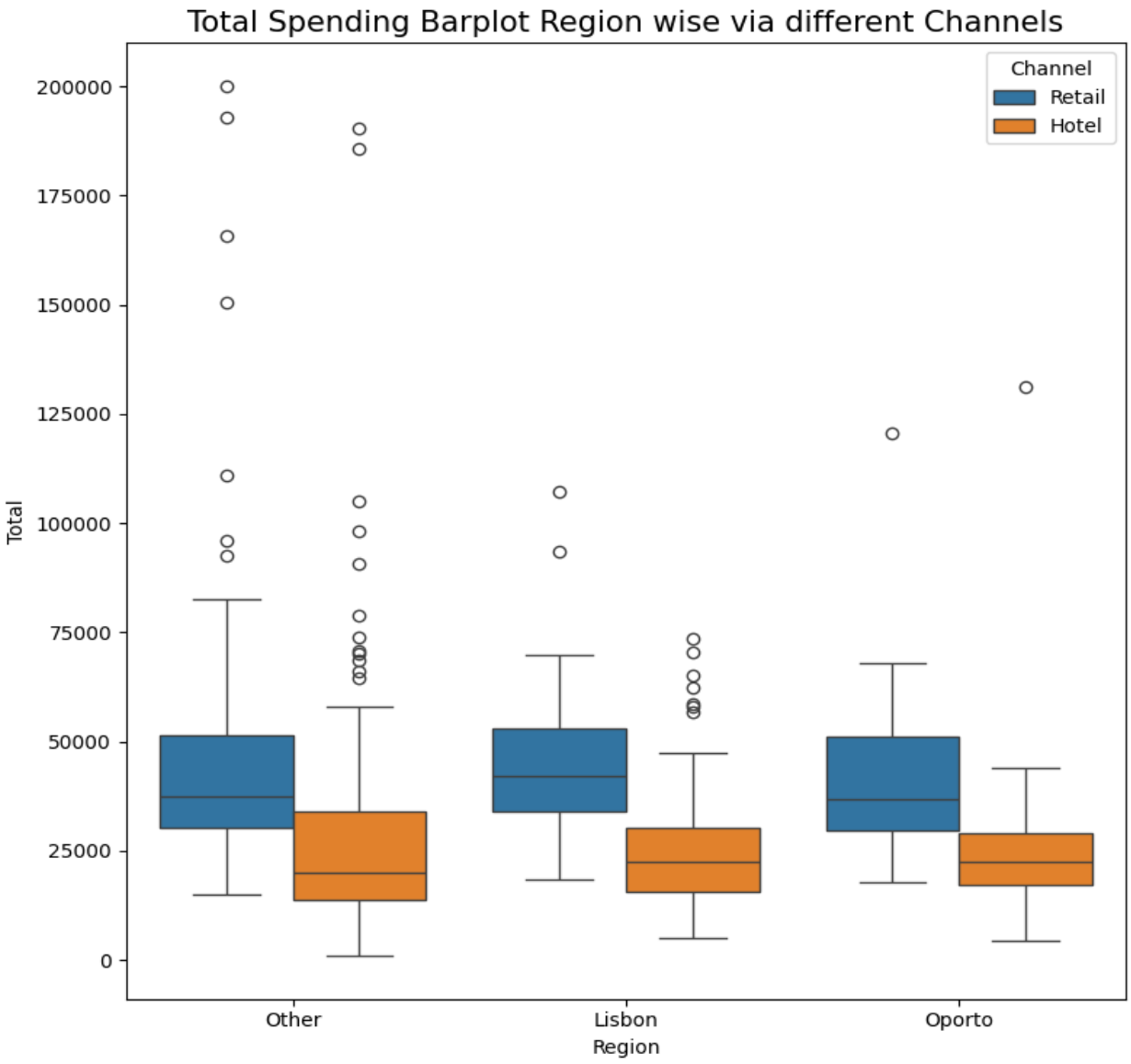

FIGURE 7

1. Lisbon:

- The total spending on the retail channel is higher than on the hotel channel.
- The distribution shows higher median spending and wider range.
- The total spending in hotel channel is lower.
- The distribution is concentrated with low median spending.

2. Oporto:

- The total spending on the retail channel is higher than on the hotel channel.
- The spending distribution is slightly narrower showing consistent spenders.
- The total spending in hotel channel is lower.
- It shows less variation indicating uniform spenders.

3. Other:

- The retail channel shows higher total spending.
- The distribution shows relatively consistent spending patterns.
- The spending in the hotel channel is lowest among all the regions.
- The distribution is concentrated on showing uniform spenders.

CORRELATION BETWEEN THE DIFFERENT ITEMS IN TERM OF SPENDING

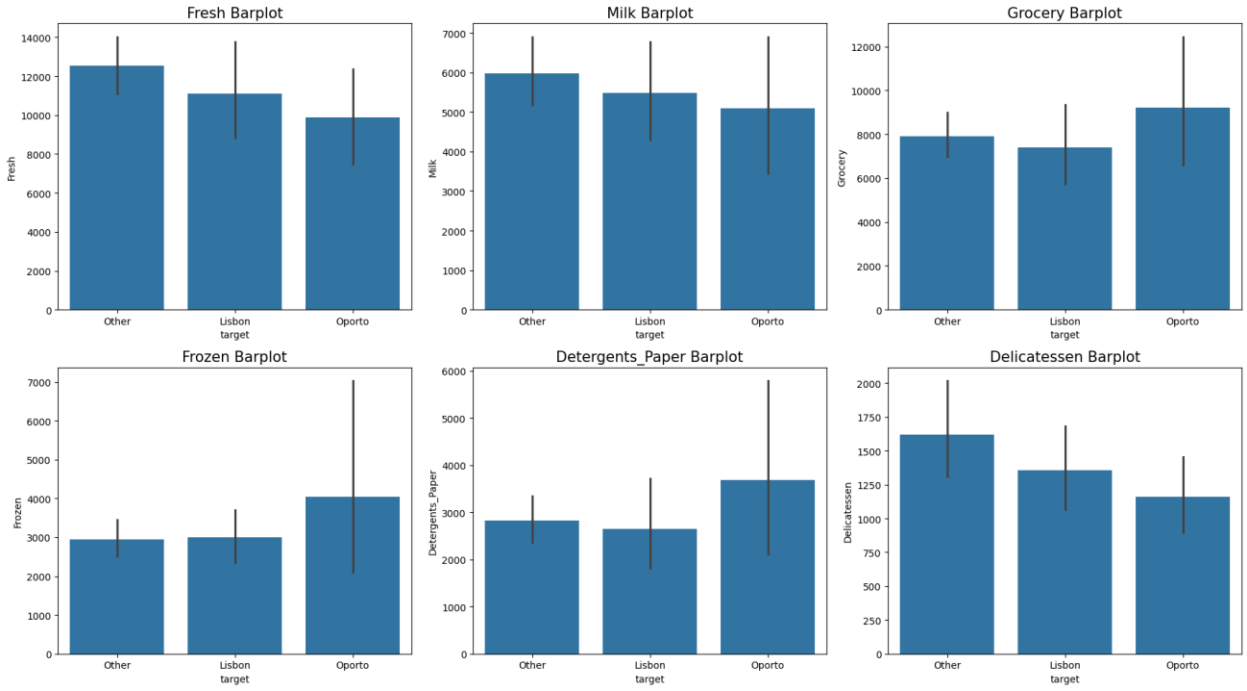

FIGURE 8

Insights based on the Bar plot:

1. Fresh Products:

- ‘Others’ has the highest average spending followed by Lisbon and then Oporto customers.
- There is variation in spending among ‘Other’ customers showing different purchasing behavior within this group.

2. Milk:

- ‘Other’ customers spend the most on milk with Lisbon target and Oporto customers followed.
- There is variation in milk spending among ‘Other’ showing wide range of milk consumptions.

3. Grocery:

- Lisbon customers spend the least on groceries. Oporto customers have the highest average spending followed by ‘Other’ customers.
- The variation in spending is uniform across all customer groups showing consistent purchasing behavior.

4. Frozen Products:

- Oporto customers spend the most on frozen products followed by Lisbon and ‘Other’.
- The variation in spending on frozen products among Oporto shows diverse needs.

5. Detergent_Paper:

- Oporto spends the highest on Detergent_Paper followed by ‘Other’ customers. Least is spent by Lisbon customers.
- The variation in spending by Oporto customers shows the stocking behavior.

6. Delicattesen:

- ‘Other’ customers spend the highest on delicattesen followed by Lisbon and then Oporto customers.
- The variation in spending by ‘Other’ shows a wide range of purchasing habits.

Correlation between different item varieties in terms of spending

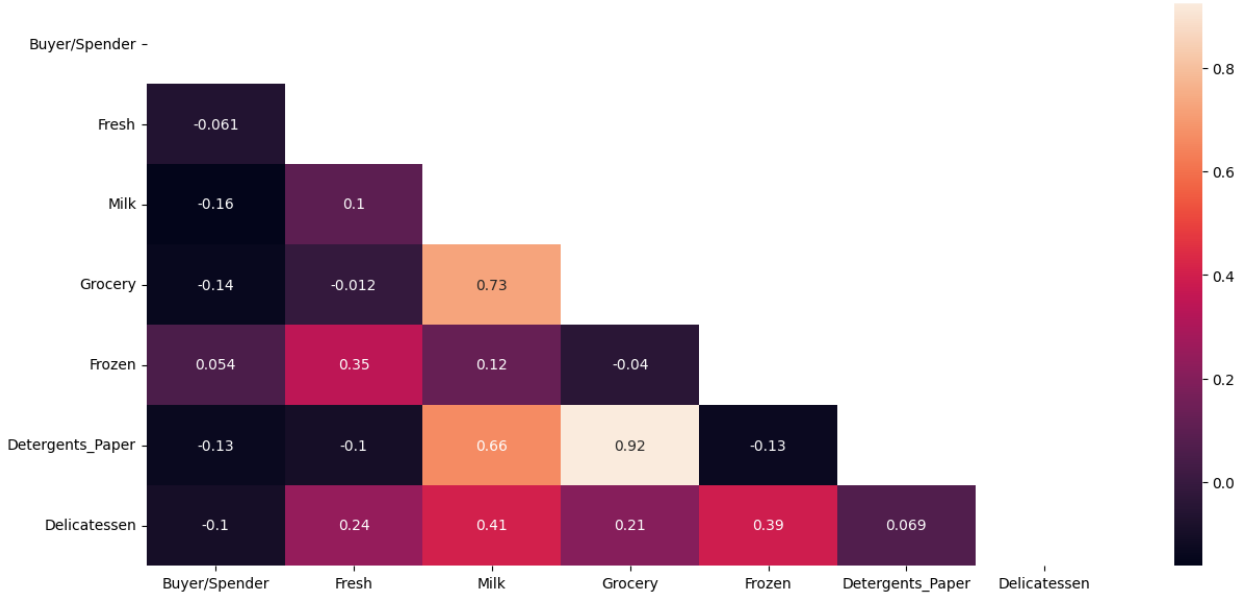

FIGURE 9

Insights on the above heatmap:

- **Strong positive correlation:** Milk & grocery, detergent_paper & milk and grocery & detergent_paper shows a strong positive correlation suggesting that the customers who buy one of the following items tend to spend more on the other product.
- **Moderate positive correlation:** Delicatessen & milk, delicatessen & grocery, frozen & fresh, delicatessen & frozen shows a moderate positive correlation. Customers who spend more on one of these items tend to spend more on the other product too.
- **Weak or no correlation:** The buyer/spender with other variables correlation is low with the highest being –0.16 with milk showing no strong linear relationship with any of these categories.
- Negative Correlation: There is no strong correlation as such in the heatmap. The correlation between milk & buyer/ spender (-0.16) and Detergent_paper (-0.10) is negative and tends to be weak.

CONCLUSION AND BUSINESS RECOMMENDATIONS:

Based on the Exploratory Data Analysis performed, here is the conclusion and business recommendations for overall business growth.

Conclusions:

1. **Spending pattern by product categories:**

- The distribution analysis shows significant variations in spending across different categories.
- Fresh, milk, grocery categories have higher mean spending compared to frozen, detergent_paper and delicatessen.

2. **Regional analysis:**

- Spending on fresh products is higher in Lisbon as compared to other regions.
- Oporto shows higher spending on the milk & grocery category.
- The other region has varied spending across all categories indicating diverse demand patterns.

3. **Channel analysis:**

- Hotels tend to spend more on fresh & delicatessen products, likely due to the need for high quality and fresh ingredients.
- Retail channels have higher spending on groceries, milk and detergents_paper which is important everyday items for customers.

4. **Correlation Insights:**

- There is a high correlation between spending on detegents_paper & grocery showing that the customers who buy more grocery items are more likely to buy more dtergents_paper.
- Another important correlation includes milk & groceries showing that these items are purchased more frequently together.

Business Recommendation:

1. **Enhance channel specific promotions:**

- Hotels: Create offers/ discounts for fresh & delicatessen items. Collaborate with high end hotels to provide customized solutions that meet their specific needs.
- Retail: Focus on value driven promotions for everyday essentials like milk, groceries and detergents_paper. Implement cross promotional strategies that increase bulk buying and regular purchases.

2. Tailor marketing strategies by region:

- Oporto: Bundle offers that combine milk & grocery products with complementary items like frozen food can be effective and strengthen marketing efforts for milk & grocery categories.
- Lisbon: Focus on promoting fresh products as there is already demand. Make loyalty programs too.
- Other regions: Flexible promotional strategies that can cater to different demands in different areas. Conduct futher analysis to identify local preference

3. Supply chain optimization:

- Implement dynamic supply chain management system that can quickly adapt to changing demands and ensure timely replacement of high demand products.
- Align inventory management strategies with regional/ channel specific to reduce waste.

4. Influence correlation insights for cross selling:

- Create bunch offers that combine highly correlated products such as detergents_paper & grocery. This can increase transactional values and improve customer satisfaction.

5. **Enhanced customer experience:**

- Use data driven insights to personalize customer interactions and offer tailored recommendations that improve the overall shopping experience.

PROBLEM – 2

EDUCATION - POST 12TH STANDARD

DATA DESCRIPTION AND ITS DICTIONARY:

- Names: Names of various university 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 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 Ph.D.’s
- 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 CLEANING AND PREPARATION:

- The data contains a total of 18 columns and 777 rows.
- There are no missing values in the data set.
- There are 17 numerical data types and 1 categorical feature.

EXPLORATORY DATA ANALYSIS (EDA)

UNIVARIATE ANALYSIS

Insights based on the histogram & Boxplot:

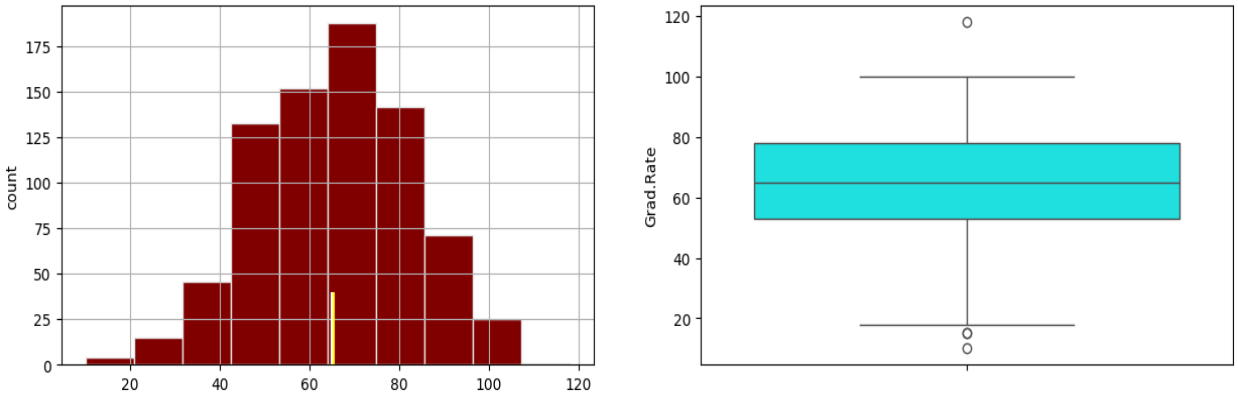

FIGURE 10

- The histogram above shows a roughly normal distribution of graduation rates. Most of the institutions have graduation rate between 30% to 80%
- The mean graduation rate appears to be around 65% which indicates that the average institution has a graduation rate around this value.
- Few of the institutes have low rates around 20% and as high as 100%
- The mode of histogram is around 60% - 70 % which shows many institutions fall within this range.
- The median graduation rate is approx. 65%
- The IQR lies around 55% to 75%
- There are few outliers present in the boxplot, lower outliers below 40% shows institutions with lower graduation rates.
- Upper outliers above 100% indicate institutions with an unusually high graduation rate.

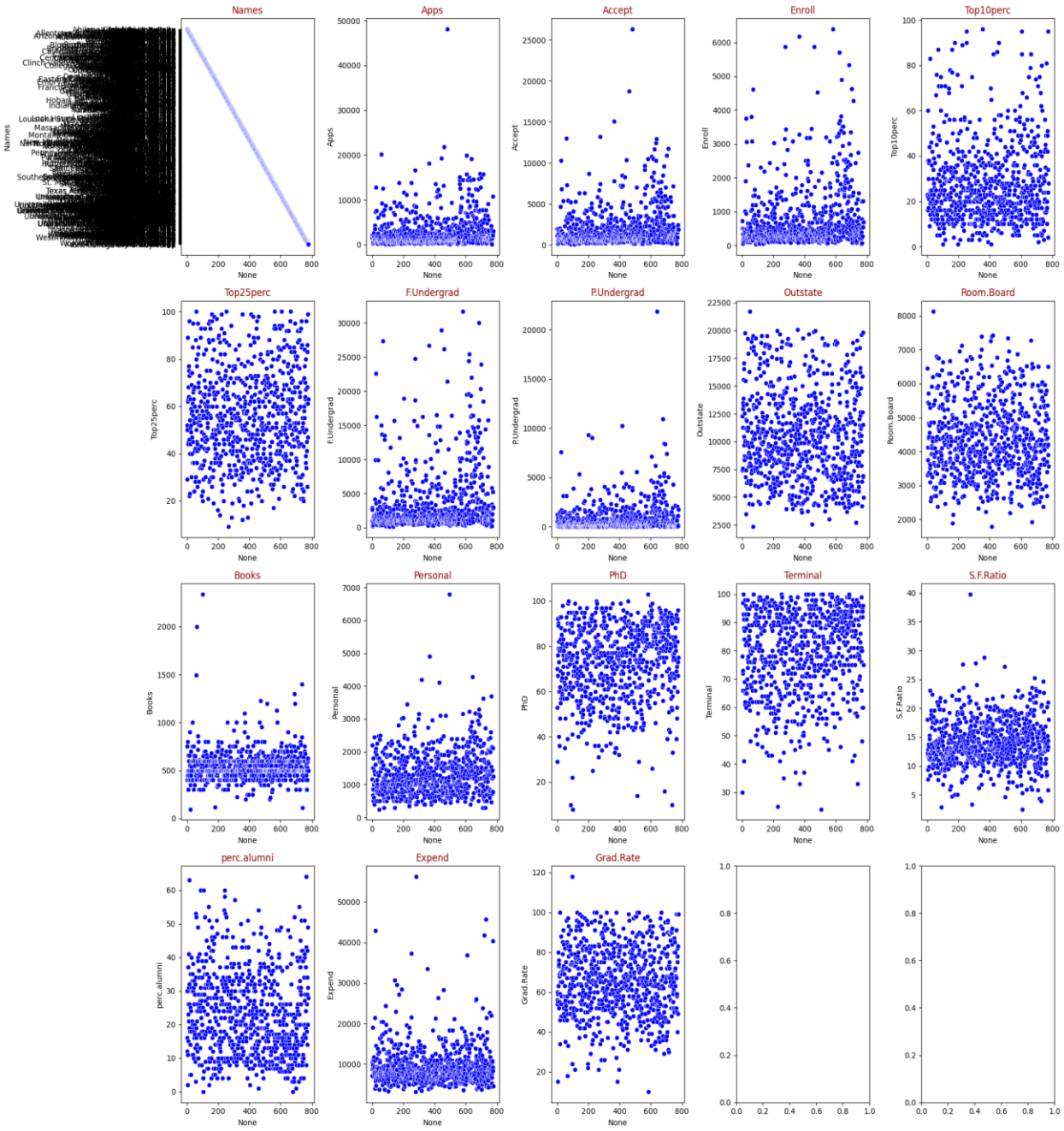

FIGURE 11

Insights based on the scatterplot:

- There is a variation in percentage of new students from the top 10% and 25%

- Outliers with very high values suggest that only a few institutions receive more applications.
- There is a wide range of applications, acceptances & enrollment among institutions.
- The plot shows that full time undergraduates are more prevalent than part time undergraduates.
- Some institutions have a higher number of full-time students.
- There is no outlier with higher cost.
- The student faculty ratio varies, some institutions have a lower ratio, indicating smaller class sizes.
- Alumni donation rates show that some of the institutes have high engaging alumni though the majority have moderate to low engagement.
- Expenditure on students varies, showing differences in resource uses.
- Graduation rate also varies a lot showing some institutions achieving high graduation rates.

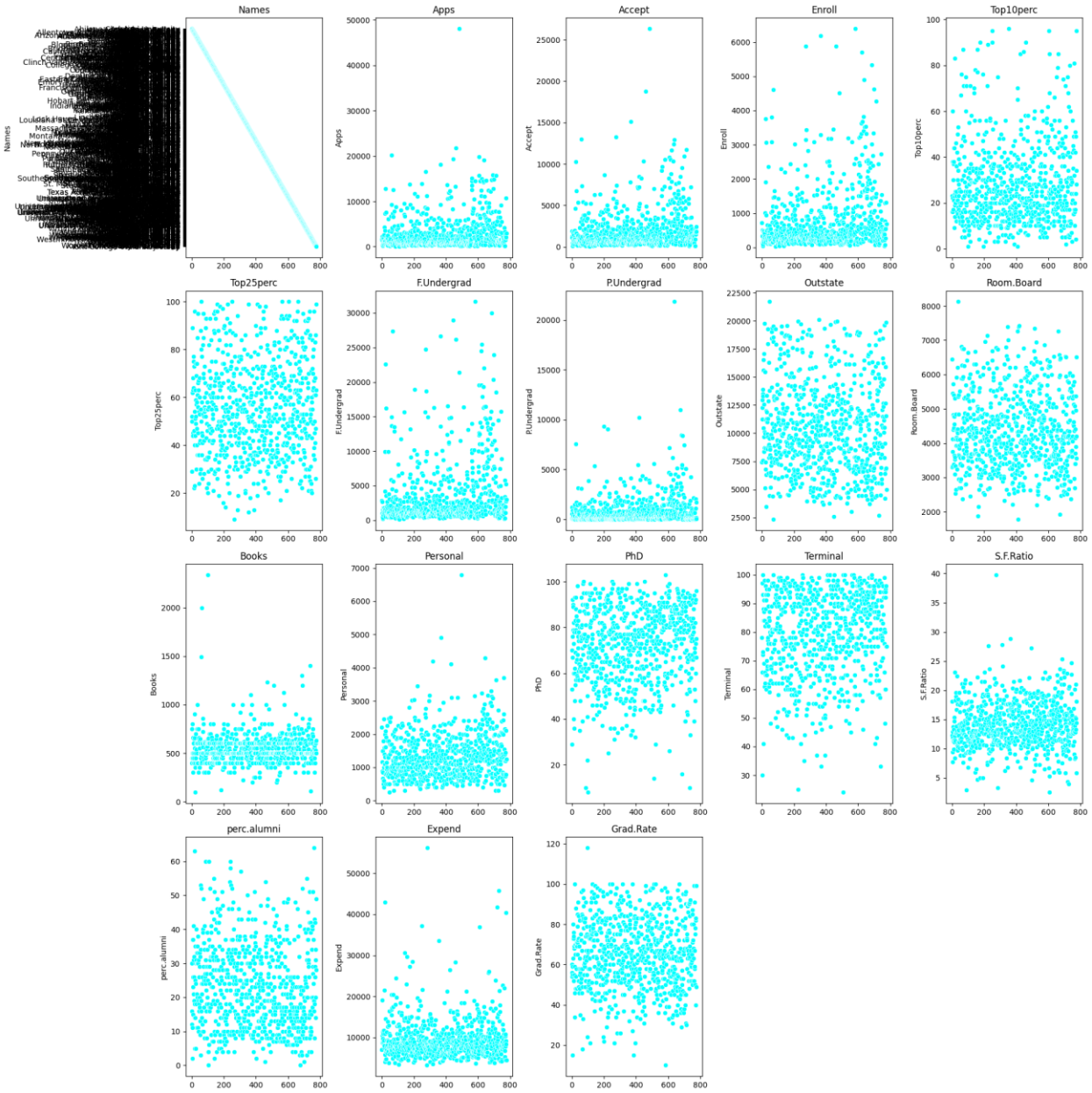

FIGURE 12

Insights based on the scatterplot:

- Many institutes have high qualified faculty which correlate with better educational outcome.
- Several outliers in the plot in applications, enrollment, cost, expenditures suggest that few institutions stand out.

- Alumni engagement varies, showing different levels of support and satisfaction.

Scatterplot: Apps vs Names

FIGURE 13

Insights:

- The scatterplot shows a wide range of apps across different institutions.
- Correlation heatmap shows no correlation due to categorical variable.

Scatterplot: Accept vs Apps

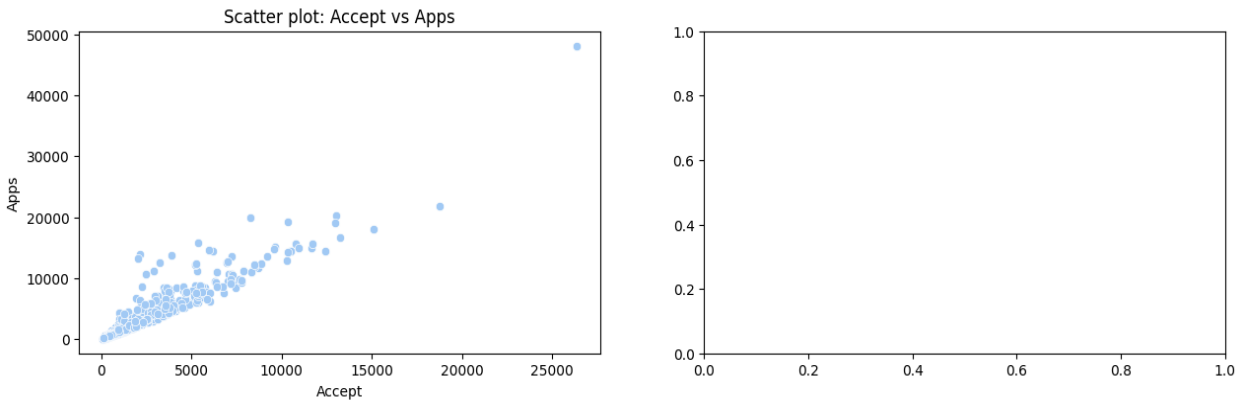

FIGURE 14

Insights:

- As institutions receiving more applications tend to accept more students showing positive correlations.
- The heatmap shows a strong positive relationship.

Scatterplot: Enroll vs Accept

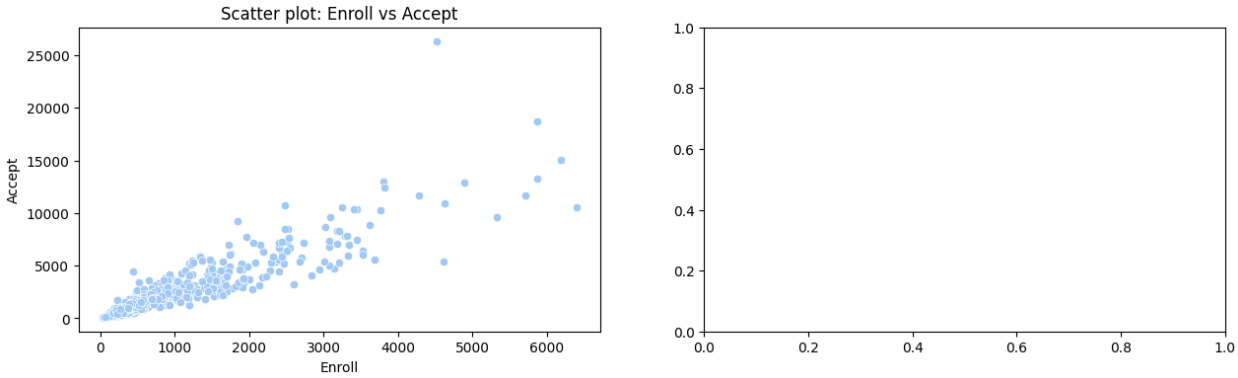

FIGURE 15

Insights:

- As more accepted students lead to higher enrollment showing a positive correlation.
- The heatmap shows a strong positive correlation.

Scatterplot: Top 10perc vs Enroll

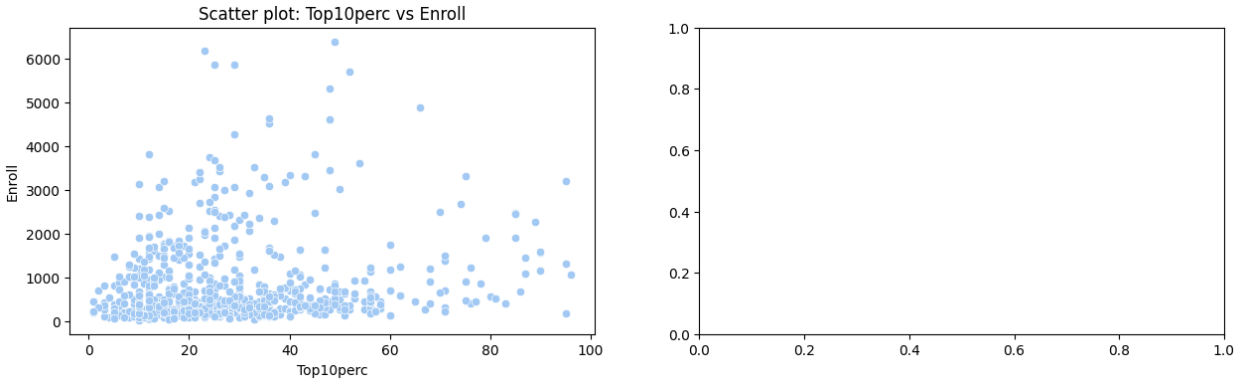

FIGURE 16

Insights:

- There is nothing clear in the scatterplot.
- Correlation heatmap shows weak correlation.

Scatterplot: Top 25perc vs Top 10perc

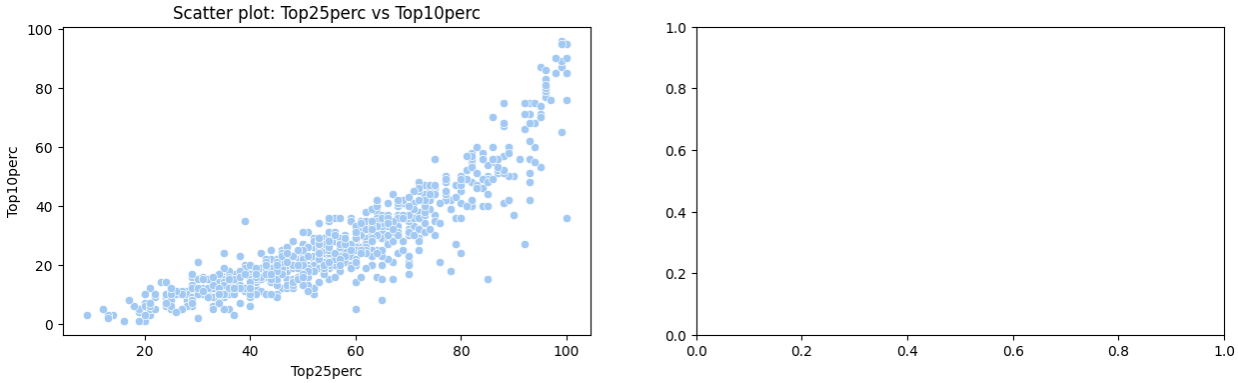

FIGURE 17

Insights:

- As institutions with a higher percentage of 10% students also tend to have higher percentage of 25% students.
- Correlation heatmap shows positive correlation.

Scatterplot: F. Undergrad vs Top 25perc

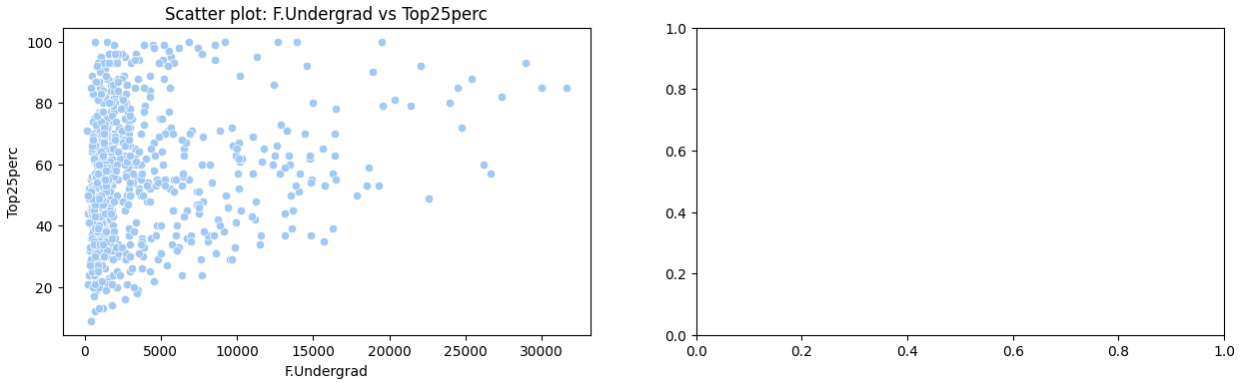

FIGURE 18

Insights:

- Nothing clear in the scatterplot.
- Weak correlation in the heatmap

Scatterplot: P. Undergrad vs F. Undergrad

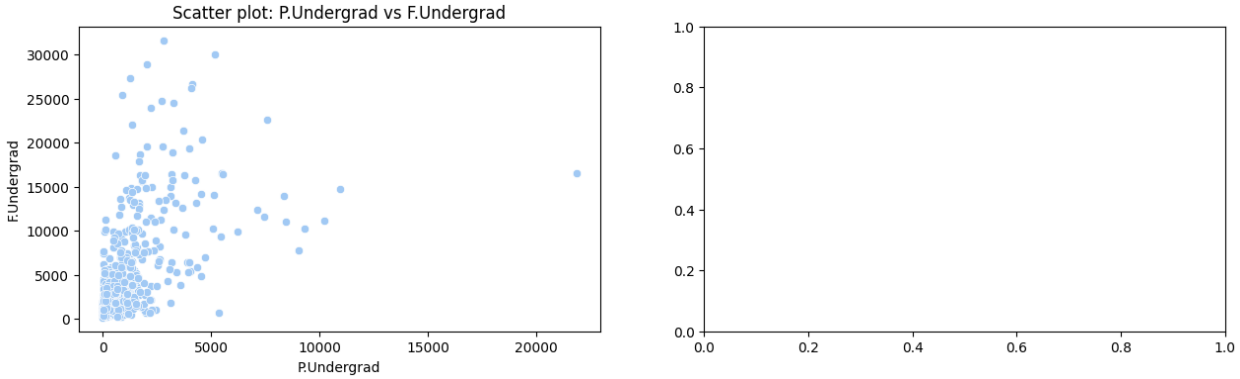

FIGURE 19

Insights:

- The number of part time undergraduates does not strongly relate to full time undergraduates.
- Correlation heatmap shows weak correlation.

Scatterplot: Outstate vs P. Undergrad

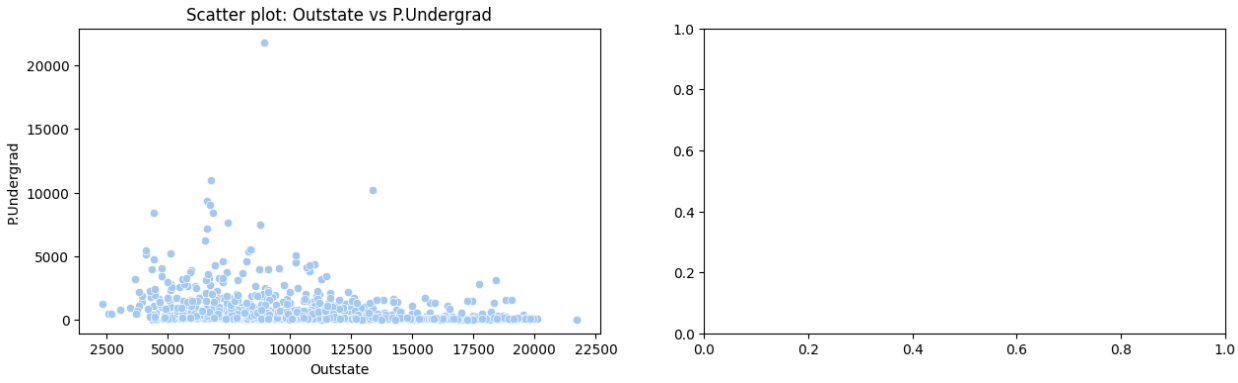

FIGURE 20

Insights:

- No clear trend in the scatterplot.

- Correlation heatmap shows weak relationship.

Scatterplot: Room board vs Outstate

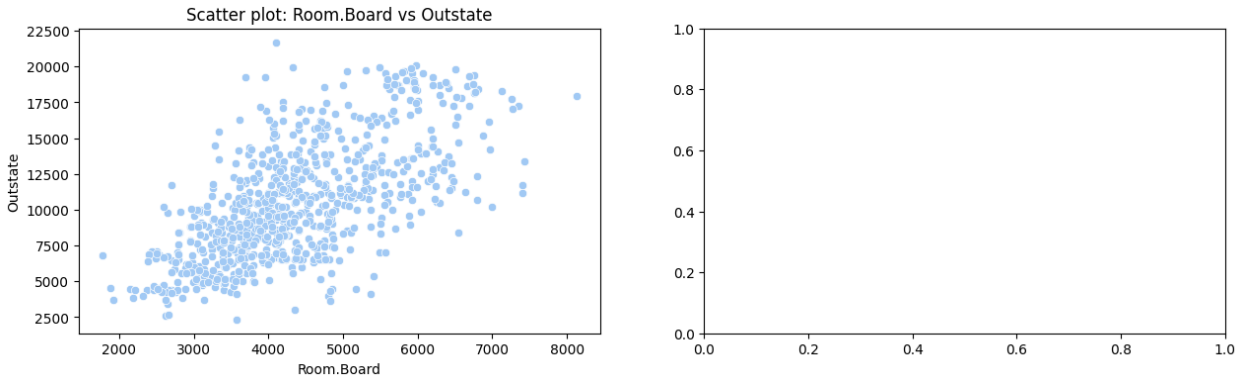

FIGURE 21

Insights:

- No clear trend.
- According to the heat map weak correlation.

Scatterplot: Books vs Room board

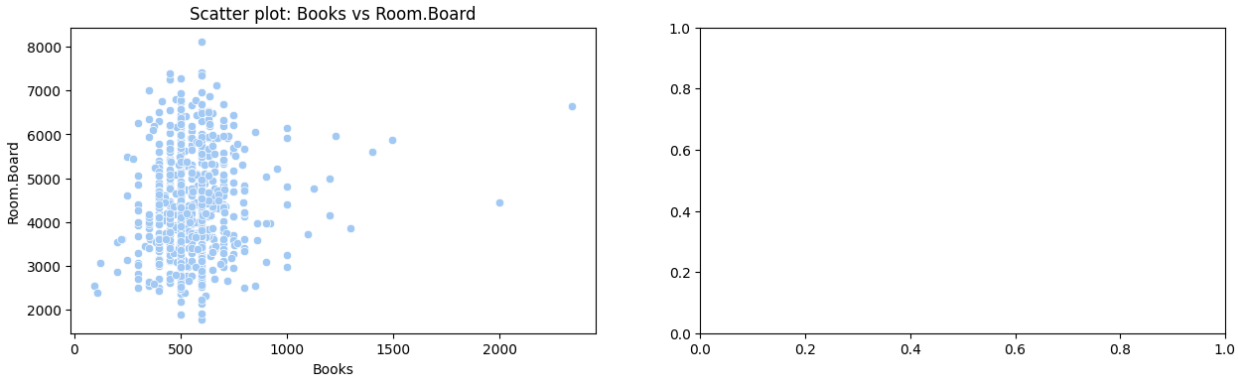

FIGURE 22

Insights:

- Nothing clear in the boxplot.
- Weak correlation in the heatmap.

Scatterplot: Personal vs Books

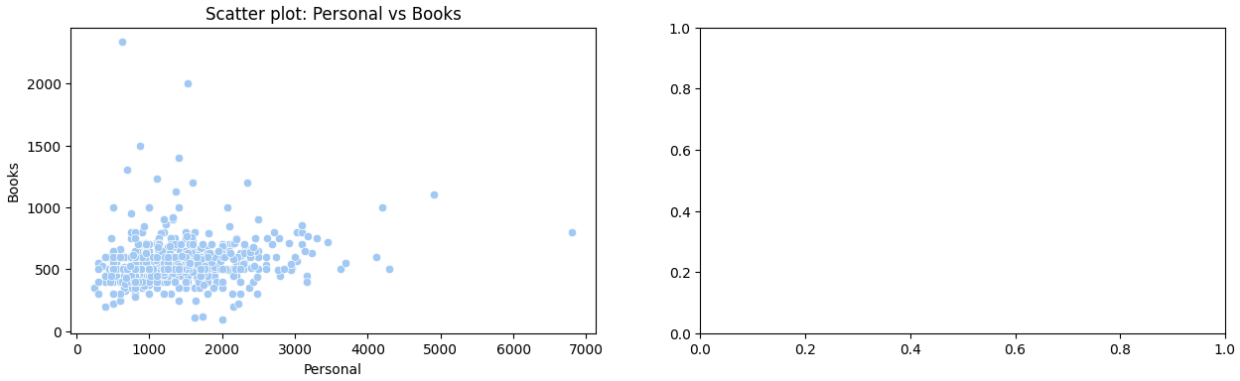

FIGURE 23

Insights:

- No clear trend.
- Weak correlation in the heatmap.

Scatterplot: PhD vs Personal

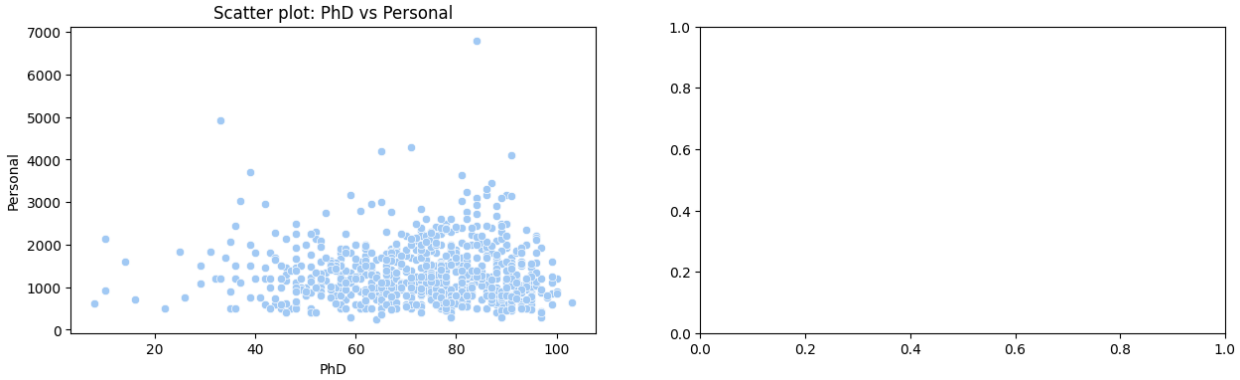

FIGURE 24

Insights:

- Nothing clear in the scatterplot.

- Weak correlation in the heatmap.

Scatterplot: Terminal vs PhD

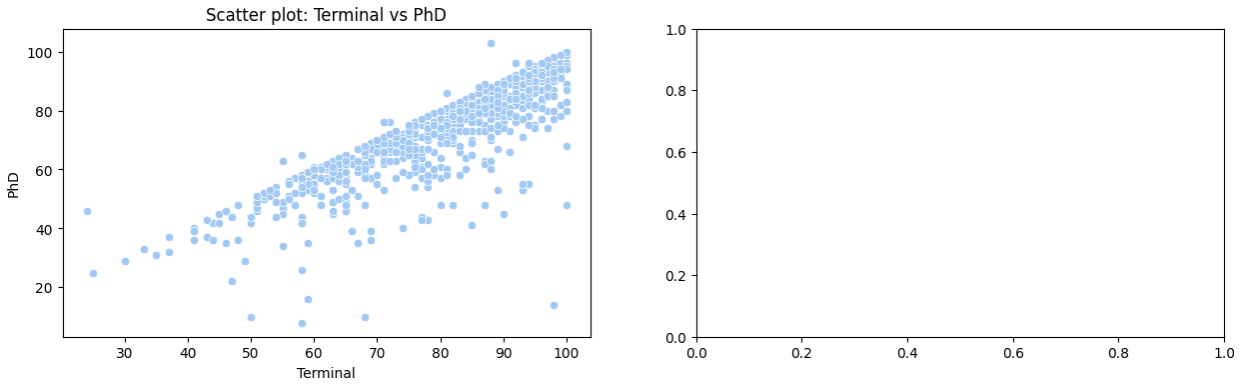

FIGURE 25

Insights:

- As institutions with more PhD qualified faculty also tend to have higher terminal degree qualification.
- Heatmap shows strong positive correlation.

Scatterplot: S.F Ratio vs Terminal

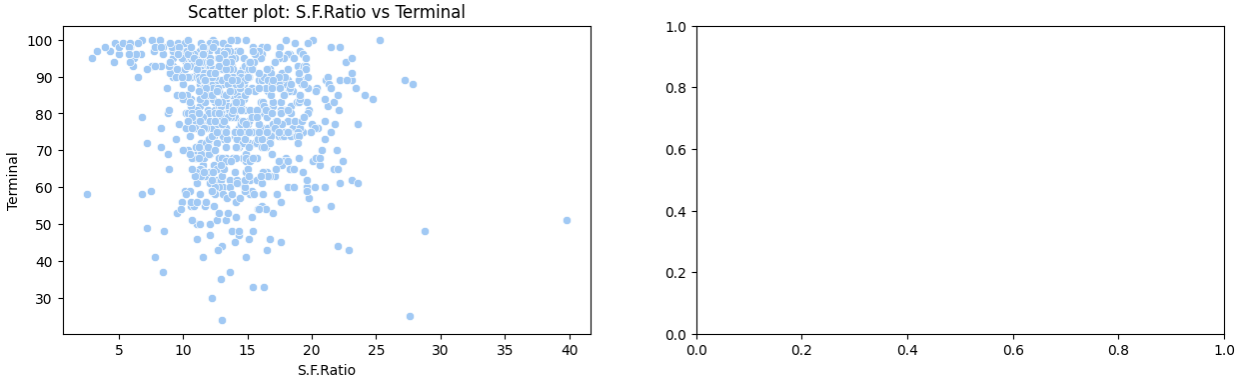

FIGURE 26

Insights:

- No clear trend can be seen.
- Weak correlation in the heatmap.

Scatterplot: Perc.alumni vs S.F Ratio

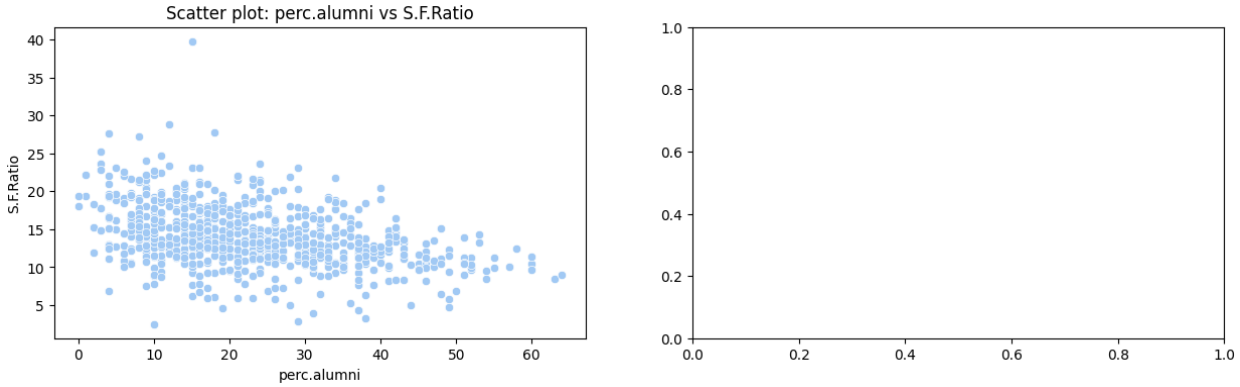

FIGURE 27

Insights:

- No clear trend.
- Weak correlation in the heatmap.

Scatterplot: Expend vs Perc.alumni

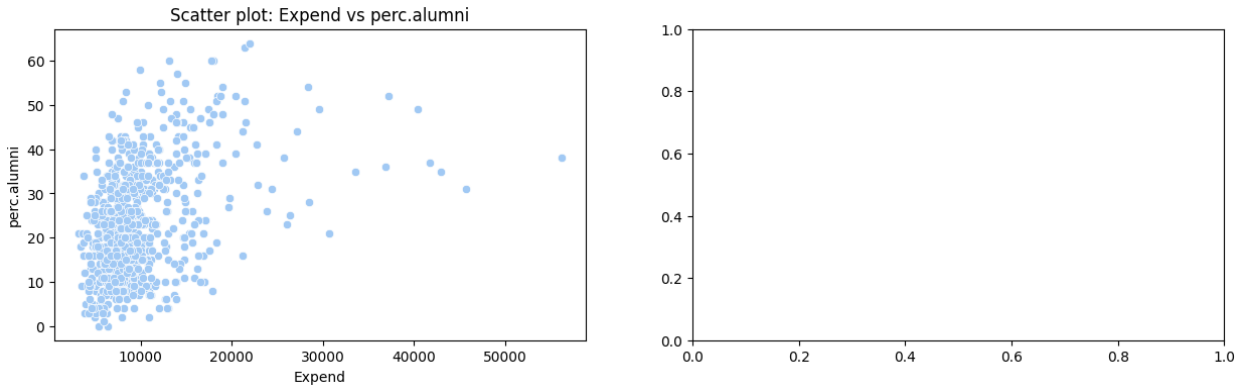

FIGURE 28

Insights:

- Nothing clear in the boxplot.
- Weak correlation can be seen in the heatmap.

Scatterplot: Grad.Rate vs Expend

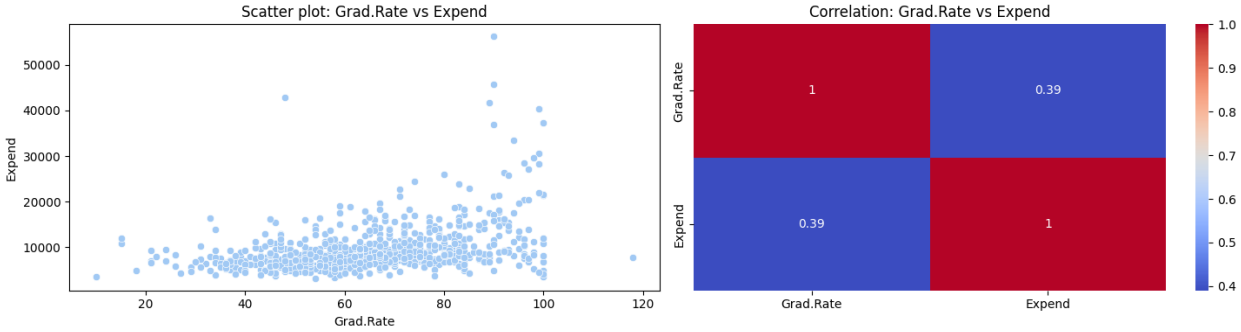

FIGURE 29

Insights:

- No clear trend in the scatterplot.
- Weak correlation in the heatmap.

CORRELATION BETWEEN COLUMNS:

- Applications and accept has a very high positive correlations suggesting that colleges receiving more applications have higher chances to accept more students.
- Accept and enroll also shows significantly strong positive correlation showing higher acceptance rates lead to higher enrollment numbers.
- Institutions with a higher percentage of faculty with terminal degrees also tend to have more PhD holding faculty.

- Higher expenditure and alumni donation are linked to better student faculty ratio. The higher percentage of top performing students and out of state students are linked with higher costs.

BUSINESS RECOMMENDATIONS AND CONCLUSION:

Conclusion:

- ◊ The acceptance rate is highly correlated with the number of applications and enrollment, suggesting that larger institutions with higher application volumes tend to have higher acceptance and enrollment numbers.
- ◊ Institutions receiving more applications tend to enroll more students.
- ◊ The percentage of students from top 10% & top 25% is positively correlated showing institutions attracting high
- ◊ achieving students perform well in the other.
- ◊ Correlation of Room. Board & Outstate shows that institutes with higher outstate tuitions have higher room and board cost.
- ◊ Positive correlation of Expend & Grad. Rate shows that investment in resources and faculty may lead to better student outcomes.
- ◊ The negative correlation of faculty/ student ratio suggests that institutions with more faculty per student tend to invest more in faculty and resources.

- ◊ The positive correlation between perc. Alumni, grad. Rate & expend suggests that institutions with high investment in students are more likely to have engaged alumni.

Business Recommendations:

- ◊ Colleges with low graduation rates should invest in students support services to improve student retention and success.
- ◊ Institutions should ensure financial aid programs to help manage high costs of room and other expenses making education more accessible to more students.
- ◊ Institutions with high applications but low acceptance rates should consider strategies to maintain selection, improving their academic reputation.
- ◊ Hiring qualified faculty and staff for the enhancement of quality of education leads to graduation rates and satisfaction.