# 2004-NPcrestHori_B-W

# School of InfoComm Technology

**Data Exploration & Analysis Assignment**

Diploma in DS

April 2022 Semester

**ASSIGNMENT 2**

(40% of DEA Module)

11th July 2022 – 7th August 2022

**Submission Deadline:**

**Presentation: 7th August 2022 (Sunday), 11:59PM**

**Report and files: 7th August 2022 (Sunday), 11:59PM**

|  |  |  |
| --- | --- | --- |
| **Tutorial Group** | **:** | **T01 / T02 / T03** |
| **Student Name** | **:** |  |
| **Student Number** | **:** |  |

**Penalty for late submission:**

10% of the marks will be deducted every calendar day after the deadline.

**NO** submission will be accepted after 14th August 2022 (Sunday), 11:59PM.

Contents

[*1.* *Abstract/Overview* 3](#_Toc110736730)

[*2.* *Creating Dashboard to Answer Business Questions* 4](#_Toc110736731)

[*3.* *Further Analysis* 10](#_Toc110736732)

[*4.* *Summary* 15](#_Toc110736733)

Assignment Presentation Link:

<https://connectnpedu.sharepoint.com/:v:/s/Test599/EdY-lsH8MMtMpCNmMyFfJFwBKPXhdIIWVEcqT9u4aoFCBA?e=5fsLB1>

# Abstract/Overview

Having been given the FoodFresh dataset, this report will cover the process of understanding and exploring the data as well as providing recommendations based on the findings and insights to improve the business’ performance. This report is primarily split into two sections, the creation of a dashboard to address business questions and further analysis using advanced analytical techniques.

The steps taken to prepare the data for exploration is included in both sections. For the section regarding the dashboard, the rationale behind the business questions and how each visualization within the dashboard is thoroughly explained. In Further Analysis, the dataset will be explored using univariate, bivariate, multivariate and data mining with a regression model to uncover more findings that were not discovered from the dashboard.

# Creating Dashboard to Answer Business Questions

To identify patterns and trends, the dashboard should be able to answer business questions and highlight key performance indicators. The questions that the aforementioned dashboard would answer with their rationale are as follows:

1. **Which period has the most Sales?** – To identify if there is any seasonality in sales as this will help to optimize management of inventories, staffing and timing of appropriate marketing campaigns.
2. **Who are the Top 10 High Spending customers?** – Grouping “High Spending” consumers to help identify the most appropriate target for FoodFresh products and thus, to efficiently and effectively use marketing activities to engage them and to keep them as FoodFresh customers.
3. **What are the Top & Bottom 5 Selling items/brands?** – It is generally a good business practice for a company to pay more attention to its high selling products, to make some effort to sustain those low selling products and decide if it is still sustainable/profitable for the business to continue low selling products.
4. **Which branch produces the most and least revenue?** – Similar to question 3, keeping an account for each branch’s sustainability/profitability is important in determining its value and deciding on the attention and budget the branch would need to deliver more business and see a profit.
5. **How are registered customers distributed across each region?** – Cross-referencing this to question 4, the distribution of customers in each region might have a direct correlation to the regional branch’s revenue. This also leads to determining both the region and branch’s value to the company and whether more or less accommodation is required.
6. **Which age group of customers generate the most sales and revenue?** – Determining the value of each age group allows FoodFresh to effectively gear their marketing strategies to their more valuable age groups to better engage and keep them as FoodFresh Customers
7. **How many products are under each brand?** – Linking this to question 3, the number of products in a brand may have a direct correlation to a brand’s performance and keeping track would help to optimize the number of products FoodFresh to introduce or remove from the brand for sustainability.

Before creating any visualizations for the dashboard, data preparation steps are taken such as creating relationships between datasets, creating new columns and formatting them:

1. Load FoodFresh excel file into PowerBi
2. Go to Model tab to edit data relationships
3. Drag CustomerID from Cust to Trans to create relationship
4. Went to Power Query Editor to edit dataset
5. Added a new column, ‘Revenue’, under TRANS table with formula, ‘= [Quantity]\*[Price]’
6. Added a conditional column, ‘Age groups’, under CUST table for age ranges
7. Closed and applied Power Query Editor
8. Changed ‘Revenue’ column datatype to be Decimal Number
9. Changed format of ‘Time’ column to (Long Time) as previous format type included the incorrect date
10. Changed format of currency related columns to include $ signs
11. Two measures, ‘Average Revenue per Month’ and ‘Number of Transactions per Month’, is created to account for missing data in November and December of 2021

Chart, line chart

Description automatically generated

Figure 1. Average Revenue and Transactions per Month

Figure 1 is an Area Chart consisting of the average revenue and number of transactions per month in the y-axis and each corresponding month in the x-axis. An additional dotted line is drawn for the maximum revenue that is reached in the chart.

This answers ‘Which period has the most Sales?’, by showing each period’s revenue and number of transactions. From the chart, we can see that December has the highest average revenue of $933.49K as well as number of transactions, and February the lowest. There is also a direct correlation between the number of transactions and revenue generated. Thus, December is the period that has the most sales and should have the most accommodation in terms of marketing campaigns, inventory and staffing.

Chart, table

Description automatically generated

Figure 2. Top 10 highest spending customers

Figure 2 is a Bar Chart of the names of the top ten customers’ total revenue that they generated for the company. They are sorted in descending order and the amount that they spent can be seen within each bar.

This answers ‘Who are the Top 10 High Spending customers?’, by showing the ten customers who generated the most revenue for FoodFresh from their purchases. The customer who spent the most is Koh Eng Mei with $8.4K followed closely by Chong Kuan Choon with $8.2K. Targeting FoodFresh’s marketing campaigns towards these ten customers would help keep them as returning customers.

Graphical user interface

Description automatically generated with medium confidence

Figure 3. Top 5 selling products by quantity (with revenue)

Figure 3 is a Bar Chart with the names of the top five products based on their quantity sold ranked in descending order with the bars representing their total generated revenue. Hovering over a bar will show the total quantity the product has sold, the exact revenue and the average price of one product. Since the quantities of the top five products are very similar, displaying the bars of each of their revenues is more meaningful than having same-sized bars.

This answers ‘What are the Top & Bottom 5 Selling items/brands?’, by showing the top five products sold. The highest selling product in terms of quantity is Matcha Garden followed by Dark Chocolate Souffle, Coffee Indulgence, Strawberry Shortcake and Peanut Butter Addiction. Despite being the most popular product, Matcha Garden does not generate the most revenue for FoodFresh. A business decision would be to increase the average price of this product which can be seen by hovering over the bar which would then lead to more revenue.

Chart, bar chart

Description automatically generated

Figure 4. Bottom 5 selling products by quantity (with revenue)

Similar to Figure 3, Figure 4 is Bar Chart with the names of the bottom five products based on their quantity sold. The features and reasonings of this Figure are the same however this time, it shows the lowest five.

This answers the question ‘What are the Top & Bottom 5 Selling items/brands?’, by showing the bottom five products sold. The least popular product in terms of quantity is the Brisket King Beef Noodle followed by Little Hanoi Noodle Bar, Miso Japanese, Chinatown Beef Noodle and Penyet & BBQ. Out of the five, Brisket King Beef Noodle also generated the least revenue. From this, discontinuing this product is a good course of action due to its low profitability.

Chart, line chart

Description automatically generated

Figure 5. Total Revenue and Number of Products under each Brand

Figure 5 is a Line and Column Chart consisting of each brand’s total revenue and number of products. Each bar represents the total revenue of the respective brand, the line represents the brand’s total number of products. Hovering over the bar and line shows the exact revenue and number of products of the brand.

This answers ‘What are the Top & Bottom 5 Selling items/brands?’ and ‘How many products are under each brand?’, by showing the total revenue generated by each of the six brands as well as the number of products that they have. CakeFresh is the most popular brand in terms of revenue and KopiFresh the least. CakeFresh also has the least number of products at 12 and FoodFresh the most at 28. Despite having the highest number of products, FoodFresh is the second least popular. A possible decision taken from this would be to reduce the number of products under FoodFresh due to under-performance and increase it in CakeFresh to further boost its revenue.

A picture containing chart

Description automatically generated

Figure 6. Total Revenue from each Branch Location

Figure 6 is a Map featuring each branch and their total revenue. Each bubble on the map represents a branch placed at it’s location latitude and longitude and the size of the bubble is indicative of the total revenue generated and the colours of the bubble is region-based. Hovering over a bubble displays the branch location’s name, exact total revenue and region that it is in.

This answers ‘Which branch produces the most and least revenue?’, by showing the total revenue generated by each branch determined by the bubble’s size. Hovering over the largest bubble, the Lim Chu Kang branch in the west produces the most revenue at $965K. For the smallest bubble, the Choa Chu Kang branch also in the west produces the least revenue at $942K. From this, we can tell that the Lim Chu Kang is the most and Choa Chu Kang the least valuable branch in terms of revenue generated. However, this is not by much as both their performance are relatively close and thus, no action is needed as they are performing extremely well but if necessary, shutting down the Choa Chu Kang branch would lead to an overall more efficient business.

Chart, pie chart

Description automatically generated

Figure 7. Registered Customers in each Region

Figure 7 is a Pie Chart showing the number of customers in each region. Each slice in the pie chart represents a region and the size of which representing the number of registered customers in that region. Additionally, the colour representing each region matches the regions in figure 6 for easier comparison.

This answers ‘How are registered customers distributed across each region?’, by showing the distribution of customers for each region. West has the highest number of customers making up nearly 29% of FoodFresh’s total registered users and Central is the least at 13%. With this in mind, marketing strategies can be targeted in the West as that is where most of FoodFresh’s customers are. Referencing the map in figure 6, we can also place a new branch in the Central which may increase the number of customers in that region as there are currently only three branches in that area.

Chart

Description automatically generated

Figure 8. Total revenue and transactions by age group

Figure 8 is a Line and Column Chart with the total revenue and number of transactions generated by each age group. The ages of each customer are binned into six different columns with their bars representing the total revenue and the line representing the number of transactions. Hovering over the bar or line displays the exact revenue and number of transactions for that age group.

This answers ‘Which age group of customers generate the most sales and revenue?’, by showing the total revenue coming from each age group as well as number of tran.sactions. The age group ’75 and above’ produces the most sales and revenues followed by 30-45. Additionally, there is a positive correlation between the number of transactions and revenue generated. From this, FoodFresh would know to target their marketing activities towards these two age groups as they generate the most revenue and number of transactions.

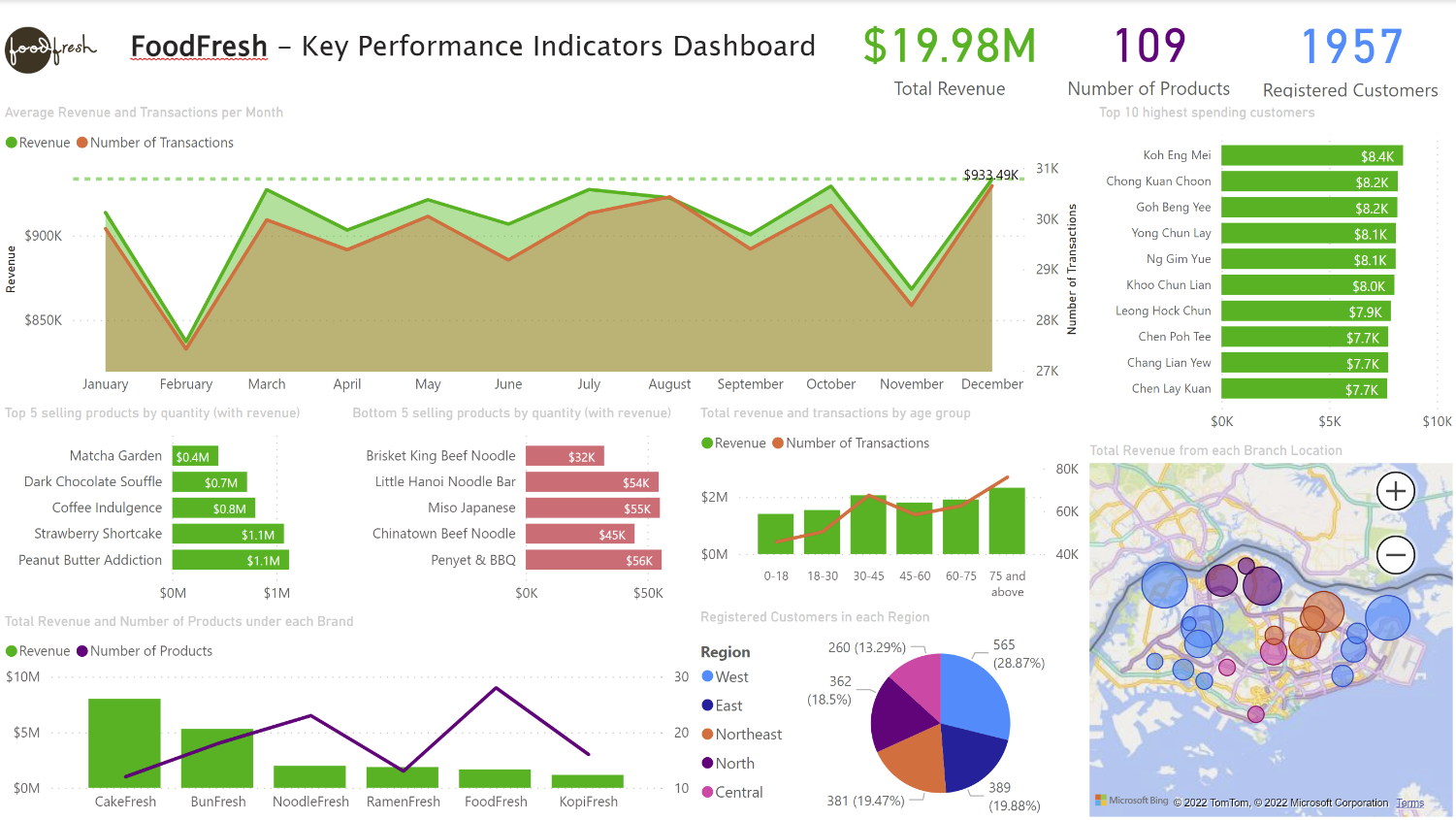


Figure 9. FoodFresh – Key Performance Indicators Dashboard

Figure 9 is the final dashboard after collating every visualization. A triadic colour scheme is used for the revenue (green), number of transactions (orange) and number of products (purple) with the exception of the ‘Bottom 5 selling products by quantity (with revenue)’ visualization which uses red to convey a negative message. Visualizations are grouped and sized based on comparison and importance. For instance, ‘Average Revenue and Transactions per Month’ is the largest as it holds the important information and Top and Bottom 5 selling products are placed next to each other for easy comparison. This dashboard can also be filtered down by clicking on any of the visualization’s data such as a branch location on the map or a brand and other visualizations would dynamically show only the data based on that filter. This would give a more focused look into specific areas of the data.

# Further Analysis

This section will dive deeper into the FoodFresh dataset by using advanced analytical techniques to discover any trends that could help improve the business. Univariate, bivariate and multivariate techniques will be used on the dataset as well as a data mining through a linear regression model to showcase these findings. Exploration will be done in Excel’s Pivot Charts and Jupyter Notebook.

Before being able to create any Pivot Charts within Excel, each sheet from the FoodFresh dataset would need to be merged based on their primary keys accordingly.

1. Create a new Excel file, ‘FoodFresh PivotTables.xlsx’
2. Get Data from Excel Workbook File, ‘FoodFresh Data.xlsx’
3. Tick the ‘Select multiple items’ and select the five Excel sheets to be merged
4. Click on Transform Data to edit the data
5. Merge Queries under Combine section
6. Select a table to merge with the Trans (Transactions) Table
7. Select the corresponding primary and foreign keys of the two tables and merge using a Left Outer
8. Repeat steps 6 and 7 for every table until the full dataset is arrived
9. Expand each column tables while unselecting the primary keys of the table
10. Close & Load

Similarly, the tables will also need to be joined and merged within Jupyter Notebook. Each sheet from the ‘FoodFresh Data.xlsx’ is exported as a csv file and stored in a new file called ‘data’. The following code is then used inside Jupyter Notebook to merge the files.

Text

Description automatically generated

An additional column is also created, ‘Revenue’, by multiplying the Quantity and Price columns.

Univariate exploration generally involves one feature per visualization and is used to find the distribution of the data. Each visualization that is used primarily focuses on one singular feature which are usually bar graphs, pie charts and box plots.

Chart, pie chart

Description automatically generated

Figure 10. Distribution of Gender

Figure 10 is a pie chart showing the distribution of gender based on the number of transactions generated by registered customers. The number of transactions made by males and females are almost equal with males making slightly more transactions in comparison to females. This means that the gender of the customer has little correlation to the number of transactions that they make and thus, would not need to be further investigated.

Chart, bar chart

Description automatically generated

Figure 11. Number of Transactions for each quantity

Figure 11 is a bar chart showing the distribution of transactions across each amount of quantity of product ordered. Out of all the transactions, the highest quantity of product ordered in one transaction is 5 and the minimum is 1. From this chart, we can tell that customers generally favours ordering in lower quantities of one and two and tend to avoid three. To boost sales, FoodFresh could implement a small order fee where ordering lower quantities would include an additional fee which would either increase higher quantity orders or generally boost revenue.

Chart, box and whisker chart

Description automatically generated

Figure 12. Distribution of Product Price

Figure 12 is a box plot of the prices of the products based on the transactions. Based on this, we know that the average price of all the products is $10, the median is around $6-$7 and the inner range for prices of most products are around $5-$10. Additionally, the prices of products can go as low as a couple of cents and almost as high as $20. There are also several outliers which reach upwards of $48. Through this, FoodFresh can easily determine if they should increase the price of their products to increase the average which would lead to more revenue.

Rather than exploring only one feature, Bivariate exploration involves two features and their relationship to one another. A bar chart featuring two variables and a correlation matrix for every numerical variable will be used to determine the relationships of certain features.

Chart, bar chart

Description automatically generated

Figure 13. Total number of transactions from each branch

Figure 13 is a bar chart of the total number of transactions from each corresponding branch location. From the graph, we can tell that the Changi, Bishan and Punggol branches sees the largest number of transactions and Choa Chu Kang the smallest. With this information, FoodFresh can accommodate the top three branches with management of inventory due to their high frequency of orders and potentially discontinue the Choa Chu Kang branch for its lower volume in sales.

Treemap chart

Description automatically generatedA picture containing chart

Description automatically generated

Figure 14. Correlation of features

Figure 14 is a correlation matrix showing the correlation between each numerical variable within the dataset. The closer the value is to 1, the stronger the correlation between two features. Unsurprisingly, the Price and Revenue has the strongest relationship 0.84 as Revenue is a derived from the multiplication of Price and Quantity. Besides that, the BranchCode and Price also has a good relationship of 0.5. This tells us that the price of a product is determined by the branch that it is in.

Multivariate analysis involves exploring three or more features simultaneously to explore their relationship. A simple bar chart with three features will be used to showcase this.

Chart, bar chart

Description automatically generated

Figure 15. Number of Transactions for each age group and gender

Figure 15 is a clustered column chart showcasing the total number of transactions for the registered customer’s age group based on their respective gender. From this we can see which gender has creates more transactions in specific age groups and would allow FoodFresh to home in on who they cater to. For example, the age group 7-21 has males make more transactions than females and alternatively, the age group 82-96 has more females making transactions than males. Essentially, when FoodFresh is determining who to accommodate, whether it be marketing strategies or sales campaigns, to customers in the age group 7-21, they should accommodate to the males and when it comes to the age group 82-96, they should accommodate to the females.

Data Mining involves identifying potentially useful patterns within the dataset. For this report, direct data mining will be done to attempt to infer a relationship based on the labeled training data to be used on the unlabeled data. Using Jupyter Notebook, a Linear Regression Model will be created using the dataset to predict the target variable based on its input variables. The target variable in this case will be the newly created feature, Revenue, which the regression model will attempt to predict.

Using the merged dataset earlier, I proceeded to drop columns ‘Quantity’ and ‘Price’ as they were involved in the creation of the target variable. A train and test split is then done with 70% of the data acting as the training data and 30% the testing data. Additionally, I used ordered ordinal encoding to turn the categorical features into numerical data for model training and a standard scaler to scale the data. Finally, I used a Linear Regression Model on the training data in order to predict the revenue.

Text, letter

Description automatically generated

Above is the Linear Regression Model’s Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and R-square (R2) scores which determines the performance and effectiveness of the model. For the MSE and RMSE, the lower the score, the better where 0 is the best possible score. For the R2 value which is the coefficient of determination, which measures how well the input data’s columns account for the target variable’s variance where 1 is the best score. From this, we can tell that the Linear Regression Model is relatively decent in predicting the revenue with all the current input variables.

Table

Description automatically generated

We can further investigate the importance of each feature to the revenue by taking the linear regression coefficient and comparing it with all the input variables. The Product’s Name is the most important variable in predicting the revenue at around 18 followed by the Brand at 7 and Product Category at 5.

# Summary

In conclusion, based on my findings from the dashboard in an attempt to answer business questions, FoodFresh should optimize their management for December as that is the period that has the most sales, target their marketing campaigns towards the top ten spending customers and the age groups 75 and above and 30-45, increase the average price of Matcha Garden, discontinue Brisket King Beef Noodle, increase the number of products under the CakeFresh brand and reduce them in FoodFresh and introduce a new branch in the central region.

With further analysis using univariate, bivariate, multivariate and data mining techniques, I have found that FoodFresh should the gender of the customer has little correlation to the number of orders they would make, customers tend to order in lower quantities which allows for the implementation of a small order fee, there is room for an increase in product price, there should be inventory accommodation for the Bishan, Changi and Punggol branch due to high volume of transactions, the price of the product is correlated to the branch that it is sold at, marketing campaigns should be targeted at males when focusing on the age group 7-21 and females when focusing on the age group 82-96 and finally, the product name and brand is the most important in predicting the revenue a transaction order would generate.