

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** | **:** | **T02** |
| **Student Name** | **:** | Ong Sze Hwee Alec |
| **Student Number** | **:** | S10222628K |

**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 and Overview 3](#_Toc110797183)

[1.1 Report Objective 3](#_Toc110797184)

[1.2 Data Preparation and Cleaning 3](#_Toc110797185)

[2. Creating Dashboard to answer business questions 4](#_Toc110797186)

[2.1 Which period has the most Sales? (Time analysis) 5](#_Toc110797187)

[2.1.1 Visual Explanations 5](#_Toc110797188)

[2.1.2 Rationale and Conclusion 7](#_Toc110797189)

[2.2 Who are the Top 10 High Spending Customers? (Customer analysis) 8](#_Toc110797190)

[2.2.1 Visual Explanations 8](#_Toc110797191)

[2.2.2 Rationale and Conclusion 10](#_Toc110797192)

[2.3 What are the Top & Bottom 5 Selling items/brands? (Product analysis) 10](#_Toc110797193)

[2.3.1 Visual Explanations 10](#_Toc110797194)

[2.3.2 Rationale and Conclusion 12](#_Toc110797195)

[2.4 Which branch and area are the most/least popular? (Location analysis) 12](#_Toc110797196)

[2.4.1 Visual Explanations 12](#_Toc110797197)

[2.4.2 Rationale and Conclusion 14](#_Toc110797198)

[2.5 What are the trends for customers demographic? (Customer analysis) 14](#_Toc110797199)

[2.5.1 Visual Explanations 14](#_Toc110797200)

[2.5.2 Rationale and Conclusion 15](#_Toc110797201)

[2.6 What is the highest revenue product category (Transaction analysis) 15](#_Toc110797202)

[2.6.1 Visual Explanations 15](#_Toc110797203)

[2.6.2 Rationale and Conclusion 15](#_Toc110797204)

[3.Further analysis 16](#_Toc110797205)

[3.1 Univariate analysis 16](#_Toc110797206)

[3.1.1 Product Code 16](#_Toc110797207)

[3.1.2 Quantity 17](#_Toc110797208)

[3.1.4 BranchCode 18](#_Toc110797209)

[3.1.5 CustomerID 19](#_Toc110797210)

[3.1.6 Gender 19](#_Toc110797211)

[3.1.7 Region 20](#_Toc110797212)

[3.1.8 Age 20](#_Toc110797213)

[3.1.9 Product Category 21](#_Toc110797214)

[3.2 Bivariate analysis 22](#_Toc110797215)

[3.2.1 Price Vs ProductCode 22](#_Toc110797216)

[3.2.2 Price vs Quantity Sold 23](#_Toc110797217)

[3.2.3 Gender vs Product Categories 24](#_Toc110797218)

[3.3 Multivariate analysis 25](#_Toc110797219)

[3.3.1 Correlation Matrix for Transaction Table 25](#_Toc110797220)

[3.4 Data Mining 26](#_Toc110797221)

[3.4.1 Predictions for the missing months 26](#_Toc110797222)

[4. Summary and Recommendations 27](#_Toc110797223)

[4.1 Keeping and Removal of Product Categories and Products 27](#_Toc110797224)

[4.2 Expanding their branches towards the southern region 27](#_Toc110797225)

[4.3 Promotional campaign for non-members 27](#_Toc110797226)

[4.4 Having marketing sales event during February 28](#_Toc110797227)

[4.5 Focus more on their leading brands 28](#_Toc110797228)

# 1. Abstract and Overview

## 1.1 Report Objective

This report aims to address business questions by using the “FoodFresh Data.xlsx” given. Detailed yet easy to comprehend visuals will be created to form a dashboard to carry out the explanations of the business questions and to assist the company to see critical trends and data evaluations to achieve business objectives that would highly benefit the well-being and success of FoodFresh.

## 1.2 Data Preparation and Cleaning

The data given was relatively clean and well put together with most details correctly given. Each spreadsheet is also named appropriately to match its content and all columns are well organized to be used.

However, there is lack of 2021 November and December Data for transactions, making the data somewhat incomplete but critical analysis can still be done. Furthermore, we can also see that there is no “South” region for any customers or regions and the company could potentially expand towards that region.

I chose to add the revenue column with the formulae of (“Quantity” \* “Price”) as the prices given are per unit. Moreover, I have also formatted the values in “Price” and “Revenue” to be currency, following 2 decimals as they represent monetary values.

It is also to note that the “0” in the CustomerID would serve as an indicator for non-members who has a transaction with the company.

In Power BI I have also created a relationship between the Trans and Customer tables using “CustomerID” column as this would allow me to extract out the names of the customers in visuals.

Furthermore, I have also added a month column indicating the months in a shorter abbreviation so to keep my months slicer short.

# 2. Creating Dashboard to answer business questions

Chart, Excel

Description automatically generated

## 2.1 Which period has the most Sales? (Time analysis)

Chart

Description automatically generated

### 2.1.1 Visual Explanations

Starting out the dashboard (On the top left) would be the period visuals where it shows different degrees of time such as yearly, quarterly, monthly to daily. This allows me to identify any trends in terms of time with relation to sales (Counting transaction ID).

Chart, waterfall chart

Description automatically generated

Firstly, using the clustered bar chart is better when showing only 2 bars, it is also easier to tell which year is performing better when stacked next to each other. Hence this year visual simply shows how there is more sales (in millions) in 2020 compared to 2021, from 0.35m to 0.30m.

Line chart

Description automatically generated with low confidence

Secondly, by using sparkline, it is concise and pleasing to the eyes in showing the Sales by Date from left to right, showing that nearing the end of the month, 29th onwards, it sees a significant decrease in sales.

Although it might be due to that not all the months contains 30th or 31st as well as there being missing data for 2021 November and December. We would still have to take into consideration of the dip on 29th nearing the end of the month.

Diagram, line chart

Description automatically generated

Lastly, the sales and quarter visual shows that there is a significant decrease in sales during the final quarter of the year and a dip during the month of February. However, the dip in November and December seems to be caused by the lack of data entries in those months during the year of 2021. Hence by only checking 2020, February would still be the worst performing month and the company should put efforts in generating sales for that month.

### 2.1.2 Rationale and Conclusion

This business question would aim to target viewing the sales data from a period perspective where we can identify when the best timing for company is to inject promotions and marking campaigns to drive up sales (transactions count) for those low performing time such as February or at the end of months.

The rationale for my visuals is to showcase all the levels from Year to Days while keeping it concise and easy to understand with logo-themed colors and bright indicators for important data. Furthermore, by knowing when sales are high and stable, they can manage inventory and optimally spend on storage only on months they know that are performing well to reduce cost.

## 2.2 Who are the Top 10 High Spending Customers? (Customer analysis)

Graphical user interface, table

Description automatically generated

### 2.2.1 Visual Explanations

Following the time visuals, on the right side we would have the customers analysis visuals to tackle the business question. I have showcased important data such as members and non-members to know how much revenue is generated between both parties as well as a bar graph for Top 10 spending customers which would be the amount of revenue they generated for the company.

A picture containing graphical user interface

Description automatically generated

Firstly, it is important to show that customers includes both Members and Non-Members. Moreover, these cards are simple but effective in highlighting values therefore we can see that there is a significant amount of revenue being generated by non-members and that the company should still put in efforts in catering to non-members.

However, it also shows that the company is successful it in member acquisition and making them beneficial for the company as members still generated more revenue than non-members.

Table

Description automatically generated

Secondly, the Top customers would be best to only cater towards members as it is not possible to track non-members without knowing who they are. Therefore, in answering the business question of top 10 customers, we will only be seeing top 10 members that generated the highest revenue for the company. This chart would show the top spenders and the company can further improve their services or keep these customers incentivized to continue their support.

### 2.2.2 Rationale and Conclusion

The business question would help the company to see who the top spenders and they are can look up in their database to see their spending habits as well as their region to effectively market more promotions in those areas to attract more members and to appropriately target the right type of people with their promotions. After seeing the visuals, we can see that none of the top members are from the “North” region and that the company should exercise marketing strategies in areas that are not as popular.

## 2.3 What are the Top & Bottom 5 Selling items/brands? (Product analysis)

Chart, diagram

Description automatically generated

### 2.3.1 Visual Explanations

These visualizations show the Top best and bottom worst selling items in terms of quantity sold as it is most appropriate to see how popular a product is regardless of its price so we know which item can sell the most or least. Furthermore, the brands can also be ranked in terms of how much of their products are being sold to know which brand the company should work on.

Table

Description automatically generated

Firstly, this bar chart would show how well the top sellers are compared to the bottom sellers which would allow the company to easily see what their best and worst selling products for the year are 2020 and 2021. Furthermore, it will also retroactively change whenever a certain period is chosen from the slicers. As such, the company can know what items are being sold the most seasonally in terms of quarters or even months and would know what to stock up or promote during those periods.

Diagram

Description automatically generated

Secondly, this area graph shows how the brands are ranked (highest from the right). Although, they are very close in values, all hovering around 315k to 330k in terms of quantity sold. This graph would show in deeper details how well each brand is doing in a more focused scale.

### 2.3.2 Rationale and Conclusion

The rationale behind the business question is to allow the company to know which items or brand they should focus on selling while also considering if they should stop supplying their poor performing products. With this, we can see that Noodles Fresh is performing relatively the best while products such as “Matcha Garden”, “Dark Chocolate Souffle” and “Coffee Indulgence” under the cake category are their top selling items. Hence, they should pay more attention to Cake Fresh as they have potential to be the leading brand for the company.

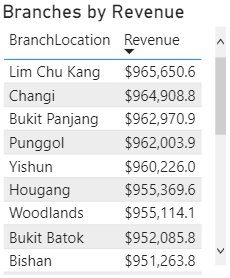
## 2.4 Which branch and area are the most/least popular? (Location analysis)

Chart, bubble chart

Description automatically generated

### 2.4.1 Visual Explanations

Location analysis is crucial for a company to know which areas they should focus their attention on to increase their revenue while also knowing what the best areas are to model after. My visuals effectively show the branches that are performing the best in descending order while also showing a map to visualize where in Singapore are those branches located and how they are performing in terms of sales.



Firstly, in order to distinguish each branch on its right as the labels overlap on the map, I have first used a table chart to clearly show the names of the branches. The table provides the name of the location and how much revenue it is generating from highest to lowest so the company can see their top and lowest performers.

Chart, bubble chart

Description automatically generated

Secondly, the map plot has bubbles that are ordered with color density, with the darker and bigger circles indicating better performance while lighter and smaller circle indicating otherwise. I have also made sure that the circles colors chosen is themed with the overall dashboard yet still being easily distinguished from each other.

### 2.4.2 Rationale and Conclusion

From these visuals, we can see that top performers comes from the North, Northeast and Centre branches of Singapore and there is a lack of branches in the Southern area and good performers in the west side. Furthermore, we can also see that the top performers generated around 965k in revenue while the lowest are not too far being 940k. Therefore, the company should investigate how the top performers are set up and promote marketing strategies in regions or branches that are not performing as well.

## 2.5 What are the trends for customers demographic? (Customer analysis)

Chart, pie chart

Description automatically generated

### 2.5.1 Visual Explanations

For my customer analysis visuals, they are kept to being simple and concise in delivering important aspects about the demographics of the members. As such, they are themed around the color of the logo easily and shows Gender, Age and Region respectively. Allowing the company to know who their targeting audiences are and where they are from.

To start off, the pie chart is used as there are only two variables (male and female) and the colors and percentage would easily show how they are segmented. Moving on, with the multi cards showing the average age of males and females, the company can tell that their audiences are mainly older adults. Lastly, the donut chart is used to not only distinguish itself from the pie chart but also to sustain more variables when showcasing each region.

### 2.5.2 Rationale and Conclusion

Although the visuals are simple, it helps in answering how the customers demographic are within the company. We can see that majority of their members comes from West, most are in their late 40s and have an equal amount of both genders. The rationale behind visuals is to provide an overview of the customers demographics within the dashboard while not overly describing details that are unnecessary.

## 2.6 What is the highest revenue product category (Transaction analysis)

Chart, treemap chart

Description automatically generated

### 2.6.1 Visual Explanations

A tree map is used to show how all the categories are segmented and the size of each tree section shows the magnitude of where most of the revenue comes from. The tree map is made by using revenue earned from each product category and the colors are chosen to fit the theme and have an ascending green tone from darker to lighter to show the contrast between high and low performing categories.

### 2.6.2 Rationale and Conclusion

We can see that more than half of the revenue generated are from Cake and they outshine other products even by multiple folds as compared to the 2nd and 3rd highest earning categories which are appetizers and noodles. Furthermore, the rationale behind using tree map is to only show significant categories such as “Cake”, “Appetizers”, “Noodle”, “Rice” and “Desserts” which would optimally show the company which categories are their leads.

# 3.Further analysis

## 3.1 Univariate analysis

### 3.1.1 Product Code

Chart, bar chart

Description automatically generated

Table

Description automatically generatedTable

Description automatically generated

By counting the “product code” variable and sorting them from lowest count to highest count. I can determine that Product 320 has the least number of transactions at 3594 while 403 have the highest at 9235 transactions. This shows us the various counts and ranking of all 109 products and tells us that there is a large gap in terms of the highest and lowest transactional products while also showing us that there is a total of 650974 transactions of products from Jan 2020 – Oct 2021.

### 3.1.2 Quantity

Chart

Description automatically generated

By counting the quantity sold and showing them in a bar chart, we can determine what is the most and least commonly sold quantity. It shows us that there are 650974 transactions in which there are 130642 counts of items being sold in pairs while the least popular bought quantity would be 3 at 129753 counts. Furthermore, we can also see that the maximum amount of item bought is capped at 5 and the range between items sold in 3s and 2s are very insignificant. Hence the company could just pay attention to the items that are being sold in 2s and what characteristics do they have.

#### 3.1.3 Price

Table, Excel

Description automatically generated

By using descriptive statistics data analysis, I was able to generate these data for the column “Price”. This shows many important data such as the average price of a product sold being around $10.24, the product with the highest frequency sold was priced at $7. The range between the cheapest and the most expensive product is $47.9 with the cheapest being only at $0.10 and most expensive at $48. We can also see that there is a count of 650974 transactions that took place with shows no empty values.

### 3.1.4 BranchCode

Graphical user interface, application, table

Description automatically generated

By plotting out the counts for each Branch in the “BranchCode” column, we can get a bar graph that is sorted from the lowest count branch to the highest. We can observe that branch “115” is has the least number of counts at 30579 which would mean that it has the lowest number of transactions hence it is performing the worst. Furthermore, it is also significantly lower than the rest of the branches with branch “109” performing the best with 31295 counts. This will allow us to see how each branch is performing relatively to one another and how the company should model after branch “109” when seeking improvement characteristics.’

### 3.1.5 CustomerID

Graphical user interface, application

Description automatically generated



Firstly, we can see that there is a total number of 290310 counts of “0” which indicates that out of 650974, there is 290310 transactions made by non-members. After analyzing with the chart, we can see that the least performing member would be “1000071” with a value of 131 while the highest would be “1000625” with 234 counts. Therefore, we can see that there is a relatively significant difference in terms of the number of transactions with members. Hence the company should target their marketing strategy towards members that are not performing well while keeping top performing members.

### 3.1.6 Gender

Chart, pie chart

Description automatically generated

The fastest way to observe the “gender” column for members is to have a pie chart and have it show the difference in the count of each gender through the colored regions. We can see that the genders are quite evenly distributed with males being only having slightly higher by 19 which is insignificant. However, we can see that there is a total of 1957 members and can infer that the company membership promotion is effective at targeting both genders equally.

### 3.1.7 Region

Chart, pie chart

Description automatically generated

By plotting another pie chart for region, we can see how the regions are distributed in terms of their counts and percentage of the entire chart. We can see that West has the highest count of 565 and 29% while Central have the lowest number of members of 260 and being only 13%. Furthermore, we can also see that there is no missing data as the total is 1957 which is the same as the total number of members. This would show the company which region they should focus on, in order to increase members.

### 3.1.8 Age

Table

Description automatically generated

We can analysis the “Age” column with descriptive statistics data analysis. This shows how the “Age” column for members are distributed. Firstly, the average age is around 49, with the most occurring age being at 75 and the age ranges has a range of 83 years from 7 to 90. It also confirms that there is no missing value since the count is 1957. This tells the company the general age demographic of their members and allows them to plan promotions that would better suit different age groups.

### 3.1.9 Product Category

Chart

Description automatically generated

This plot depicts how each product category is ranked from the highest count to lowest. We can see that “Cake” are dominating the product category followed by noodles and appetizers. Furthermore, we can also see that there is a lot of variety for some categories such as “Indian”, “Vietnam”, “Asian”,” Others”,” Thai” and “Salad”, as they only have 1 product each. Therefore, the company should consider increasing in variety or remove the product if the category is not highly successful.

## 3.2 Bivariate analysis

### 3.2.1 Price Vs ProductCode

Chart, scatter chart

Description automatically generated

By plotting a scatter plot to view the various average prices of each product. We are able to see that most product are below $10 and they generally have code between 320 and 400. We can also see that the highest price for a product here is around $50 and can tell how much of a price difference it is from other products.

### 3.2.2 Price vs Quantity Sold

Chart, scatter chart

Description automatically generated

From this scatter plot we can see that there is no direct correlation that the higher the price of the product, more quantity is sold as there are products being sold in high quantity throughout all price range. We can also see some clusters forming such as the cheaper products below $10 having 5 different groups of quantity sold. Therefore, for the company to improve their revenue, they should see what the characteristics of the high performing products are and model the other products after them.

### 3.2.3 Gender vs Product Categories

Chart, bar chart

Description automatically generated

To compare how each gender is distributed for each type of product, I have used a clustered bar chart to and set the X-axis as the product category and the Y-Axis as the count of the customerID per transaction of each product while also further color filtering the total count of customerID into female and males.

With this visual, we can see that most of the categories are quite evenly distributed between male and female with males having slightly higher counts than female which can be due to males having a slightly higher member count. Therefore, we can see that the company products are good at targeting both genders equally.

## 3.3 Multivariate analysis

### 3.3.1 Correlation Matrix for Transaction Table

Application, table, Excel

Description automatically generated

With this matrix, we can see how each of the variables is related with one another. With 1 and -1 being having the highest correlation and 0 having no correlation. We can see that Date have a very high positive correlation coefficient with TransID at 0.999995, which we can determine that as TransID increases, the Dates would also increase in a highly similar frequency.

Moving on, we can see that price have a relatively high correlation coefficient with receipt number at -0.56457 but it is not close to 1 enough to form a negative relationship between those 2 variables as logically ReceiptNo should have no effects on the price of the product.

To end off, the rest of the variables does not seem to have correlation with one another as their correlation coefficient is close to 0.

## 3.4 Data Mining

### 3.4.1 Predictions for the missing months

Chart, line chart

Description automatically generated

Table

Description automatically generated

By using regression modelling predictions, I was able to predict the sales for the months of Nov 2021 and Dec 2021 to be 29.4K and 29.6K respectively by using a formular that uses the Intercept and Sale(T-1) coefficients and previous month values.

From this we can see that there are some high performing months such as Dec 2020 and Apr 2021. However, this prediction may not be the most accurate as it deviates from the original values (in blue) by a significant amount. Therefore, the prediction should only be use as a reference for how Nov2021 and Dec2021 could perform.

# 4. Summary and Recommendations

In summary, with the data given from Food fresh, I was able to create many visuals in analyzing important variables that would determine how well the company is doing and a dashboard that showcases many aspects of the company while also answering valuable business questions. By using the visuals and analysis that I have done so far, I am able to come up with some recommendations that the company should consider for them to improve and be more successful.

## 4.1 Keeping and Removal of Product Categories and Products

From the analysis we have done so far, we can see that there are some Product categories such as “Cake”, “Appetizers”, “Noodle”, “Rice” and “Desserts” as well as top products mostly being cakes, doing very well in generating sales and revenue for the company. This would tell the company to put in more promotional and cost-reduction methods such as bulk buying of ingredients to keep up the stock of these goods while also reducing the cost of making hence increasing their profit margins.

On the other hand, we can also see that there are products in products categories that the company should consider removing such as “Asian”,” Others”,” Salad”, “Vietnam” and “Indian” or improve it by merging it with food items in top categories for a new product to sustain the variety of that category.

## 4.2 Expanding their branches towards the southern region

From the map and seeing how regions are segmented, we can observe that there is a lacking number of branches in the southern region of Singapore. Furthermore, there is no “South” region when distinguishing where members are from. Hence the company should expand their branches towards the south to capture more customers as starting a branch in a place that does not have other branches nearby would also reduce competition between branches.

## 4.3 Promotional campaign for non-members

From the cards visuals showcasing the how much revenue is generated for Members and Non-Members; we can see that there is a sizeable amount of revenue coming from non-members. Therefore, apart from trying to convert non-members into members, the company should also host promotional campaigns such as lucky draws, vouchers, and online gift codes to target non-members. This is to promote loyalty for non-members; however, the company should also not forget to give additional benefits for their members so that it also incentives non-members to convert such as having higher discounted prices, more entries in lucky draws and exclusive member bundles for seasonal festivals.

## 4.4 Having marketing sales event during February

After predicting how sales will be throughout the year of 2021, we are able to see that February is still significantly lower in sales and that the company should counteract it with marketing strategies. One way of doing it would be having seasonal promotions during the period for festivals and celebrated days during the month. Additionally, the company is doing well in their Cakes product, it would be wise to come out with more discounts for them in February to incentives buying them for Valentine’s Day or birthdays.

Furthermore, as Chinese New Year often happen during the month of February, the company should also change their product aesthetic or categories to cater towards the Asian demographics that would be buying food related product during Chinese New Year such as changing their box art and bundling their products to sell it in higher quantity together as people usually buy more products to host guest during this season.

## 4.5 Focus more on their leading brands

By viewing the area graph in the dashboard that shows how each brand is doing while also comparing with the findings found for each product and product categories within the brands in the univariate and bivariate analysis. We can see that NoodlesFresh, RamenFresh and Cakefresh seems to be leading while FoodFresh itself has been relatively doing the worst. The company should consider merging of brands to reduce logistics and branding cost while also keeping products quality more uniformed for each of their brands.

Therefore, they should seek into optimizing their main-brand “FoodFresh” by advertising them on their NoodoleFresh/RamenFresh/CakeFresh products, such as indicating that they should also try out product from FoodFresh while also remembering to keep quality of food high as using their top brands reputation can be a double-edge sword if the quality of the lesser-known brand they are promoting is not as good.