Contents

[Overview 2](#_Toc134563457)

[Background 2](#_Toc134563458)

[About the Dataset 2](#_Toc134563459)

[Objective 2](#_Toc134563460)

[Dashboard Creation Methods and Rationale 3](#_Toc134563461)

[Dashboards Creation and Rationale 3](#_Toc134563462)

[Which period has the most Sales? 3](#_Toc134563463)

[Who are the Top 10 High Spending customers? 5](#_Toc134563464)

[What are the Top & Bottom 5 Selling items/brands? 7](#_Toc134563465)

[Where are the members located? 12](#_Toc134563466)

[What is the most & least popular product category? 15](#_Toc134563467)

[Further Analysis 19](#_Toc134563468)

[Univariate Analysis 19](#_Toc134563469)

[Member Age 19](#_Toc134563470)

[Customer Order Sales 20](#_Toc134563471)

[Bivariate 21](#_Toc134563472)

[Product Category vs Product 21](#_Toc134563473)

[Continuous variables correlation 22](#_Toc134563474)

[Product vs Sales 23](#_Toc134563475)

[Multivariate 24](#_Toc134563476)

[Member vs Non-member behavioural patterns 24](#_Toc134563477)

[Data Mining 25](#_Toc134563478)

[Detrending 25](#_Toc134563479)

[Auto Regression 26](#_Toc134563480)

[Auto Regression Integrated Moving Average (ARIMA) 26](#_Toc134563481)

[Summary 27](#_Toc134563482)

[Insights 27](#_Toc134563483)

[Sales 27](#_Toc134563484)

[Customers 27](#_Toc134563485)

[Product 28](#_Toc134563486)

[Miscellaneous Findings 28](#_Toc134563487)

# Overview

## Background

The FreshFood, the company of this dataset, was founded in 2020, and have a total of 6 brands under them. FreshFood has started a analytics team, of which I am under, to use the past 2 years of data from 2020 and 2021 to generate insights and assist management in making effective business decisions.

## About the Dataset

The dataset is clean, without nulls or erroneous data and contains ample data, with 30,000 transactional data points. It contains data from January 2020 to October of 2021. The dataset contains several tables contain information about its branches, brands, registered customers, products, and all transactions taken place.

For the year 2021, it will be important to note the lack of data from November and December. This will cause the months November and December to have half the data as other months, making appear as if those months did poorly when viewing all data by month. For this report, *all data will be from the year 2020 unless otherwise stated or as seen from the year filter from the dashboards*.

## Objective

The goal of this assignment to identify and answer key business questions such that it is possible to see a clear key performance indicator. By getting an actionable metric to focus on, management will be able to plan accordingly to target these metrics most efficiently. To achieve this, dashboards will be made to make data more digestible and viewable for management and further analysis will be used to identify patterns and trends from the data.

# Dashboard Creation Methods and Rationale

To achieve the objective of identifying key patterns and trends from the dataset we must first identify questions the business would want to know and its rationale. Then, breakdown these business questions into exploratory questions, which can be directly answered with visualisations made from data. The first 3 business questions and their rationale are provided.

## Dashboards Creation and Rationale

### Which period has the most Sales?

* To identify if there is any seasonality in sales as this will help to optimise management of inventories, staffing, and timing of appropriate marketing campaigns.
* Exploratory Questions

1. Do weekends have the higher sales?
2. Which month has the highest sales?
3. How does sales change according to yearly quarter?
4. Does the month date have an effect?
5. What time of day has the highest sales?
6. Does the morning make have a higher sale than afternoon?

Chart

Description automatically generated

The above dashboard is a collection of visualisations for sales. There is a filter to switch between the 2 years.

A picture containing text

Description automatically generated A picture containing text, clipart

Description automatically generated A picture containing text

Description automatically generated

There are several cards to provide immediate information for the viewer. The count of sales and average sale amount gives clues to the viewer that a small change in the average sale would significantly change the total sales.

Chart, line chart

Description automatically generated

On the right is, rather counter-intuitively, a line graph of Quantity of product sold by month. This is so that and change in quantity of product sold will be reflected closely in the bar chart in the very middle of the dashboard. This allows a direct comparison of goods sold to sales amount

The purpose of the top third of the dashboard is to provide context to the viewer for the rest of the visualisations.

Chart, bar chart

Description automatically generated

Looking at the middle portion of the dashboard, graphs to show sales plotted against day of week, month, and yearly quarter. These are plotted by sum of sales, so as to show if there is any overall seasonality or preference to a particular period. This answers exploratory questions 1-3, showing quarterly sales change, change from weekends to weekdays, and change in sales by month.

Graphical user interface, text, application

Description automatically generated

Further below, day of month, time of day and a bar graph for morning vs afternoon is plotted. These are plotted to the average sale amount, as the sum of sales would not show the variability in sales on such a small-time scale. Time of day is grouped in 10minute bins to increase readability. As the cards suggest on top, these small variations in average sale have large impacts to total sales. This answers exploratory questions 4-6, showing change in sale by month, peaks and drops throughout a day, and difference between morning and afternoon sales.

This can be used to pinpoint at which period has the most and least sales, so that the business can optimise maintenance, inventory stock, and staffing to ensure that the most profitable times are kept well serviced, and downtime is planned accordingly.

### Who are the Top 10 High Spending customers?

* Grouping “High Spending” consumers to help identify the most appropriate target for FreshFood products and thus, to efficiently and effectively use marketing activities to engage them and to keep them as FreshFood customers.
* Exploratory Questions

1. Who are the top spending customers?
2. How often do these customers make orders?
3. How many items do these customers order at a time?
4. Where are these customers?
5. What is the age demographic of these customers?

Chart

Description automatically generated

The above dashboard is a collection of visualisations for the top 10 members. There is a filter to switch between the years.

Text

Description automatically generated with low confidence

There are 2 static cards for the average total sales for the top 10 members against the total average. This is to show the degree of difference between the top customers against the average.

Graphical user interface, text, application

Description automatically generated

The 2 pie charts show the % of sales that come from members as a whole, then how much of sales come from the top 10 members. This is to illustrate that whilst the top 10 members make a large portion of total member sales on their own, standing at almost 0.7% from 10 members, it is a small subset against all customers.

These cards and pie charts serve as context and caveats to the viewer so that they interpret the importance of the top 10 members.

Chart

Description automatically generated

Looking at the middle portion of the dashboard, graphs to show top 10 members total sales, order frequency and order category. This answers exploratory questions 1-3, showing the names of the top members and their sales contribution, how often they have ordered, and what kind of products they order.

Chart

Description automatically generated

Further below, age and location of top 10 members can be seen. This answers exploratory questions 4-5, showing the general age group of members, and where they are located.

These visualisations can be used to find out exactly how much their best customers are worth in terms of sales, how often they will buy goods, and what type of goods they order. This can be used to determine which products to keep in order to retain these high value customers. Moreover, knowing their age group and region, it can provide clues to the rest of the customer base as to how and where to market towards.

### 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.
* Exploratory Questions
  1. Which product has the highest/lowest sales?
  2. What is the most/least sold product?
  3. Is there seasonality affecting these sales and quantity figures?
  4. Which brand generated the most/least sales?
  5. Does the product a brand sells affect sales significantly?

To answer the above questions, 3 dashboards will be used, separated by product sales, product quantity, and brand. Only the product dashboards will be filterable by brand and all dashboards will be filterable by year.

#### Product Sales

Graphical user interface

Description automatically generated with low confidence

The above dashboard shows the top and bottom 5 products by sales.



There are 4 cards show the best and worst performing product in a single brand along with lifetime sales of a product and its average. This is best comparable to the rest of the visualisations when no year filter is selected.

Chart

Description automatically generated

The bar graphs show the top and bottom 5 products and their total sale amount. Below, the stacked area graph shows sales of the product by month. This answers exploratory question 1, showing the best and worst performing products, and if there is any particular seasonality behind their sales.

#### Product Inventory

Graphical user interface, application, table, Excel

Description automatically generated

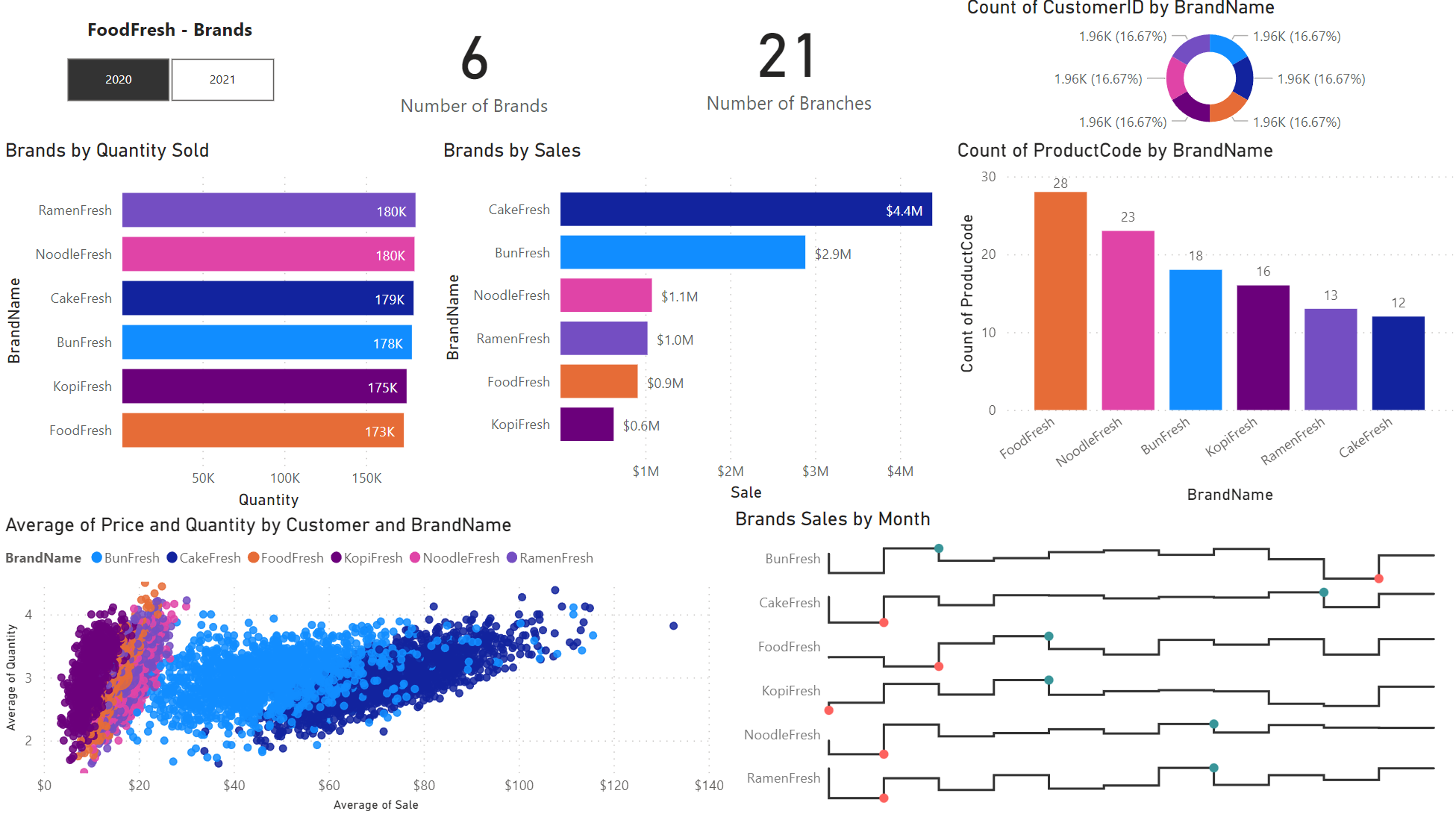
The above dashboard shows the top and bottom 5 products by quantity, or inventory, sold.

There are 4 cards showing inventory sold, the average amount sold per order coupled with the most and least sold products. The bar graphs and stacked area graphs are the exact same as the above sales dashboard but for quantity.

The above 2 dashboards answer exploratory questions 1-3, showing the top and bottom products for both quantity and sales, alongside if there were any unusual changes throughout the year that affected these metrics.

These dashboards can be used to identify which product is doing best so as to further investigate that product to better understand the underlying reasons so that it can be applied to other products, bolstering sales. It can also be used to decide if the cost of inventory management for lower sale items are worth keeping or to cut the product out entirely.

#### Brands



The above dashboard shows various metrics by brand.

Graphical user interface

Description automatically generated with medium confidence

The cards show general information, and a pie chart shows distribution of data between the brands by customer count.

Graphical user interface, application

Description automatically generated

The middle section of the dashboard shows brand by inventory sold, sales and number of unique products it sells. This answers exploratory questions 4, showing a brand’s sales. It also shows the relationship between inventory moved against total sales and how many unique products it requires to sell.

Chart, scatter chart

Description automatically generated

The bottom portion of the dashboard shows brand by a scatterplot of average sale and inventory moved, and sales by time. This answers exploratory questions 3 and 5, showing that a brand’s sales change as months pass, in addition to the degree of change in sales as quantity increases.

This dashboard can be used to determine which and when a brand is doing well or poorly and its relationship to the product it sells. Businesses always aim to reduce complexity, meaning a low number of unique products, with high sales, regardless of quantity. This dashboard allows for that identification and decide which brands need a change in product line-up

### Where are the members located?

* It is important to for a business to know where their general members are located and distributed so as to be able to pinpoint where their customers are and are not so as to market and advertise activities to the widest potential customer base.
* Exploratory Questions
  1. Which regional branch has the most members?
  2. Where are my members located?
  3. How does sales change based on a member’s region?
  4. What is the age demographic of my members?
  5. Which region has the most members?

Graphical user interface, application

Description automatically generated

The above dashboard shows visualisations based on all registered members.

Graphical user interface

Description automatically generated with medium confidence

The cards show a comparison of transactions made by members vs non-members

Text

Description automatically generated with medium confidence

The 2 pie charts show the proportion of all transactions made by members vs non-members. Since non-members will not have personal details, visualisations in this dashboard only apply to members. It also shows gender proportion of said members.

Chart, bubble chart

Description automatically generated

The map shows branches, coloured by region, and sized on number of members. This answers exploratory question 1-2, showing how many members each branch serves by bubble size and the branch proximity to each other.

Chart, scatter chart

Description automatically generated

The scatter plot shows average price and quantity by members, coloured by region. This answers exploratory question 3, showing if there is a specific change in spending habits across regions.

Graphical user interface, application, Teams

Description automatically generated

The stacked bar chart and horizontal bar chart show age demographic and count of members by region. This answers exploratory question 4-5, showing if a region’s age demographic and count of members.

This dashboard can be used to pinpoint where members are, how they purchase goods, their demographic coupled with their distribution. The business can then use appropriate marketing tools to reach their members, according to their distribution and spending habits. It can also identify and market branches which are underutilized to its associated member base.

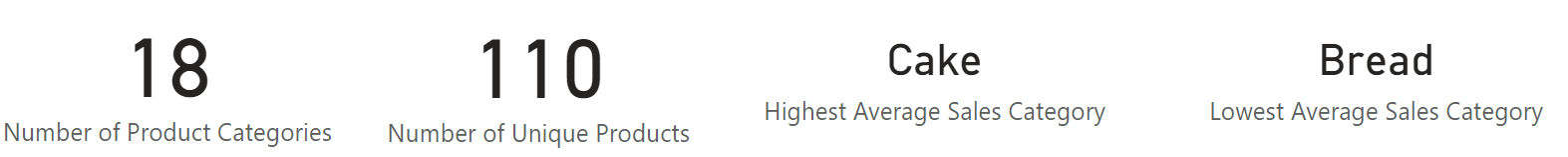
### What is the most & least popular product category?

* It is important to identify which category of products are performing the best and worst so as to investigate further into why the products in a category are performing in such a way. If the entire category is performing a certain way, new products in that category can be expected to perform similarly.
* Exploratory Questions
  1. Which category has the highest sales?
  2. How many products does this category have?
  3. How many orders does the category have?
  4. How many units of inventory has this category sold?

Chart

Description automatically generated

The above dashboard shows various metrics by product category.



The cards show general information and categories of interest to look into first.

Table

Description automatically generated with medium confidence

The left-middle section shows product category by sales, number of unique products, and number of orders. This answers exploratory question 1-3, showing the aforementioned attributes for the viewer to easily make comparisons between categories.

Chart, bar chart, histogram

Description automatically generated

The vertical bar chart shows product category by units sold, which answers exploratory question 4, showing the number of units sold.

Chart, scatter chart

Description automatically generated

An additional scatterplot organising product category by average quantity, price, and sales by size grants additional context to differentiating categories.

This dashboard can be used to make comparisons between entire product line-ups against the statistics. If a product category is performing a certain way, therein lies a certain aspect that customers are looking for or actively avoiding. Moreover, products with low inventory sold and unique products can be identified and determined if the number of products for the category should be reduced.

# Further Analysis

## Univariate Analysis

### Member Age

Chart, bar chart, treemap chart

Description automatically generated

The dashboard is a collection of visualizations showing age of members

From the 2 cards, both the mean and median of members can be seen to be very similar, this would indicate that the distribution of members age is close or is normally distributed.

However, that theory that the age is normally distributed however is quickly proven wrong when looking at the rest of the dashboard. The pie chart shows that the overall distribution of age groups is relatively similar throughout, with not one age group being overrepresented.

The bar chart shows that members aged less than 20 and between 80-90 are the largest age demographic, but not by much. This is reinforced by the tree map, showing that there is no single major age group.

This dashboard reveals that though the mean and median age is similar and would suggest a normal distribution, it is instead because the distribution of ages throughout all members is equal. Moreover, this shows that it is actually the very extremes of ages that make up the highest percentage of members, and ages between 20-80 having about the same distribution throughout. This means that there is no particular age demographic to target during marketing, as even the largest difference is minute.

### Customer Order Sales

Chart

Description automatically generated

The dashboard is a collection of visualizations showing the sales amount by order.

From the 2 cards, both the mean and median of a sale. The mean of sale is ~$30, whereas the median of sales amount is ~$16. This means that the true standard amount a sale can be expected to be is ~$16, but due to a high number of large sales, the mean has been pushed to double that of the median.

This is further shown via the pie chart, showing that about a third of total sales comes from orders over $100.

The bar chart and tree map both show the number of orders, grouped by their sales amount per order. It can be clearly seen that more than half of all transactions are less than $10 and between $10-$30. Despite this, the smaller amount of high value sales, particularly orders above $100, still make a majority of total sales.

This dashboard reveals that the average sales amount is severely weighted by high sales transactions as seen by the almost doubling of the median. Despite the fact that overwhelmingly most sales make below $30 per order, a small number of high value orders can outweigh and almost match the total sales earned. This means that the total sales of high volume, low value sales, is comparable to a small number of high value sales, raising the importance of such transaction.

## Bivariate

### Product Category vs Product

Chart, treemap chart

Description automatically generated

The dashboard is a collection of visualizations showing the product category against number of products.

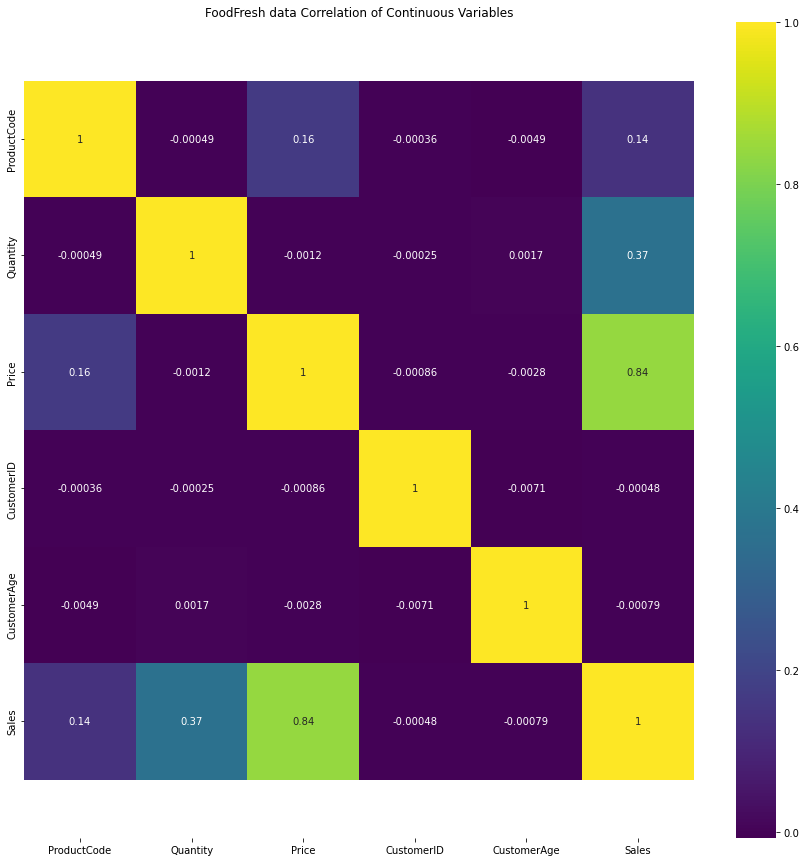
From the 3 cards, the total number of products, categories and the theorical average number of products in each category.

The pie chart is the first clue into the distribution of products in a category. It can be seen that the top 3 product categories make up more than half of all products. This means that 3 of 18 product categories make up more than half of all products.

The bar chart and tree map both show the product category against count of products, reinforcing what was seen in the pie chart. It can be seen from the bar chart that bottom few product categories only have one product in them. From the tree map, the excluding the top 4 product categories, the remain 14 categories make up only slightly more than a quarter of all products.

This dashboard reveals that there is a huge disparity between number of products in a product category. There are categories with only one product, and on the other extreme, 25 products. This means that the administrative and logistical overhead of maintaining these small product categories still exist, without fully capitalizing on the productive value of them.

### Continuous variables correlation



The above visualisation is a heatmap showing correlation where the lighter the colour, the more positively correlated they are, and vice versa for negative correlation.

It can be seen that the majority of all variables have little to no correlation to each other and there are no negative correlations.

As expected, Sale has a positive relationship with both quantity as price. However, product code has a relatively weak but not insignificant correlation to both sales and price. This would suggest that as product code increases, so would sales and price. Further analysis would be required to find the cause.

### Product vs Sales

Graphical user interface, chart, scatter chart

Description automatically generated

The above dashboard is simply 1 visualisation and tables to view data from said visualisation.

The scatterplot shows product code by sales, coloured by product category. In interpretation, product code is simply the ID of a product and should have no bearing on its sales. Exactly for this reason, it is particularly interesting that for the products under the ‘cake’ category, there are 2 distinct clusters of produce codes which do well, alongside a cluster that performed similarly to other categories.

Extracting these datapoints to a table, it is possible to see the product names and separate the products for ‘cake’ to those which sold similarly to other products, and those which sold much more.

From the table, it becomes clear all the cakes that did well have a very long name, depicting that it is very complex. This is compared to cakes that did comparatively poorly, which can be seen to have short, one-word names.

This means that cakes with longer names, thereby are more complex, have a significantly higher chance of generating more sales than shorter, plain cakes. Moreover, if these product codes were generated by when they were created, it would be possible to trace back the time in which they were created, allow for further investigation on the how the cake was thought of.

## Multivariate

### Member vs Non-member behavioural patterns

Chart, bar chart

Description automatically generated

The above bar graph is shows transactions sorted by product category, brand, region, and average price with quantity, coloured by if the customer was a register member or not. Green is for members and red is for non-members.

Using the graph, it is possible to compare the difference of how members and non-members interact with the various metrics. The height of the bar graph shows the count of that category. What is important is not the specific height, but the shape of the graph comparing members against non-members.

The above picture of the graph does not show all categories and is only a snippet of all the bar graphs. However, this is not important as the observation to be made here is the shape of the bar graphs between red and green.

As can be seen, the shape of the graph between non-members and members, is remarkably similar. What this means that regardless of whether customers are members, where they are, brand preference, and purchasing habits are the exact same. From this, we can infer that largely, if there are no discrepancies in the data, the behaviour of a members will reflect non-members. This is a vital insight as it means that data such as age of members can be expected to reflect non-members.

## Data Mining

### Detrending

The above 2 line graphs show sales by month. The second line graph is detrended to remove any underlying increasing or decreasing trend that may exist from the initial data. This is to be able to visualise the month by month change within the data.   
 It can be seen that for both years, in February, there is a huge decrease in sales, followed by a peak in March, then a repeating dip and rise until June. This is consistent for both years. This suggest that there is a particular consistent and predictable factor that happens every February which results in a massive decrease in sales.

### Auto Regression

The above line graph shows current sales, in blue, and predicted sales in orange. It can be seen that the predicted sale is very flat as compared to the actual sales. This is most certainly due to the high variability of the actual sales, but consistent overall average.

All that can be drawn from the auto regression line graph is that whilst day to day variability of sales may change, the overall sales can be expected to not change.

### Auto Regression Integrated Moving Average (ARIMA)

The above line graph shows current sales, in blue, and predicted sales in orange. It can be seen that the predicted sale and actual sale mirror each other almost perfectly. This suggests that the moving average, or within a small time period, sales can be expected to stay similar to the above graph.

# Summary

All in all, owing to the large dataset available, many insights have been obtained by exploring and visualising the data. This section will re-establish the insights obtained throughout the report into 5 main categories focusing on, Sales, Customers, and Products.

## Insights

### Sales

It can be seen that the overall sales changes very little throughout the year. The only exception is the drop February and November. These 2 months should be further investigated identify the factors behind the drop. Otherwise, it can be expected that the total sales up to quarterly year can be expected to be the same as the last, with little variation.

It is observed that throughout the day, sales do not have any particular drops or peaks. This is contrary to the norm, where food related sales should peak around mealtimes, and drop during off-peak hours. This means that the products being sold at FoodFresh are not generally sought out during mealtimes but evenly throughout the day. This has the advantage of having a slow but incredibly steady number of sales on an hourly basis, ensuring that off-peak hours are not wasted staff time.

It should be noted that there is little growth between months, quarters or even years. With only 2 years’ worth of data, it is impossible to definitively say that business growth is stagnating, however the possibility of sales stagnation is not insignificant, especially when taking into consideration predictions from data mining, showing a similar story of non-growth.

These observations can be seen from ‘[Which period has the most Sales?](#_Which_period_has)’

Overall, the FoodFresh can expect sales to have no major drop in the future, however the same would be true to increasing sales. Should the status quo remain, FoodFresh will not grow at any significant pace. The business must make a introduce changes to the way it operates if it wants to see any significant growth, even if at the expense of the unchanging sales.

### Customers

First, it is vital to consider that both register members and non-members have similar if not almost identical spending habits, opting to spend the same amount, buy the number of products, be in similar proportions according to region, and prefer similar product categories. This is imperative knowledge as it means that any insights generated from member information can also be applied to non-members. (‘[Member vs Non-member behavioural patterns](#_Member_vs_Non-member)’)

Another thing to take note of, is the change in age distribution of the top 10 customers. In 2020, the age distribution is spread evenly throughout, however in 2021, this has shifted to the extreme ends of the age brackets. This means that the business should at least expect for a year-on-year change of its most valued customers and should always make adaptations to ensure each high paying customer stays as long as possible. (‘[Who are the Top 10 High Spending customers?](#_Who_are_the)’)

The region with least customers, Central, also has the least number of branches, the opposite is true for West, with the most customers but also the most branches. Looking at the positioning of Central region branches, there is a substantial lack of coverage for this region, especially considering this region is Singapore’s Central Business District (CBD) wherein the wealthiest consumers are located. (‘[Where are the members located?](#_Where_are_the)’)

Overall, FoodFresh’s customer base has an incredibly even and distributed age demographic, though its most valued customers are shifting towards the extremes of the age brackets. FoodFresh enjoys a peculiar parity between behaviour of its non-member and member customers, meaning that the business can expect the same response when introducing new products or branches, regardless of member status.

### Product

FoodFresh has 18 unique product categories, however the top 3 categories make up more than 50% of all sales. There are 10 product categories with only 1-2 products within them. It then begs to question the reasoning behind maintaining the logistical and administrative overhead to manage these low volume product categories. It is highly advisable to cut or grow these product categories to justify their upfront costs.

The best performing food category is cake, comprising of more than a quarter of all sales alone. Moreover, due to its high price, the inventory required to ensure these sales continue is comparatively lower to other products. This means that FoodFresh pays less for transportation whilst generating sales.

These observations can be seen from ‘[Product Category vs Product](#_Product_Category_vs)’

Overall, FoodFresh should review their best and worst products, both individually and by category, to determine if it is worth keeping them to potentially grow them in the future, or to cut costs now and redirect resources elsewhere. It should also continue to heavily invest in bolstering its best products under the cake category, as it is by far in a way their best products.

### Miscellaneous Findings

The majority of transactions that take place have a sale value of $0-$30. This is supported by the median sale of $16. However, the average sale price of all transactions is $30. This is about a 100% difference in the median and mean sale. This would indicate that a small number of high value transactions is severely skewing the average. In ‘[Customer Order Sales](#_Customer_Order_Sales)’ dashboard, it can be seen that transactions $100> make up a disproportionate amount of the total sales comparative to the number of such transactions. FoodFresh should aim to encourage these types of transactions as a priority.

There is an observed inverse relationship in a brand’s number of unique products sold and total sales. Seen from ‘[Brands](#_Brands)’ dashboard, it can be seen that the brand FoodFresh, has the highest number of unique products with 28, 5 more than the second highest, yet ranks nearly bottom in terms of sales. This is in comparison to the brand ‘CakeFresh’ which has half the number of products and the least overall, and ranking top of sales by brand. This would indicate that for the brand FoodFresh, it should reduce the number of unique products it sells, to focus heavily into only a few products.