ڤوليتيكنيك برونئ
POLITEKNIK BRUNEI

MODULE CODE AND NAME:

DA3307 – Data Visualization

TITLE OF ASSIGNMENT:

Assessment One: Data preparation and Visualization

NAME:

Muhammad Hidayat Bin Mohd Yusof (20FTT2910)

SUBMISSION DATE:

22nd March 2022

MODULE LECTURER:

Ak Md Waqiuddin Bin Pg Amiruddin

NAME OF INSTITUTE:

School of Information, Communication and Technology, Polytechnic Brunei

PROGRAMME:

Level 5 Diploma in Data Analytics

Session:

2022 – 2023

Table of Contents

1. Introduction 3
2. Dataset Description 4
3. Data preparation 4
   3.1. Correcting inconsistent categorical values 4
   3.2. Missing Values 5
   3.3. Outliers 7
4. Data Visualization 8
   4.1. Count of each ProductType 8
   4.2. Average MRP of each ProductType 9
   4.3. FatContent in each establishment 10
   4.4. Average ProductVisibility of each ProductType 11
   4.5. Change in OutputSales against ProductVisibility scatterplot 12
   4.6. Total sales achieved by each OutletType 13
5. Chart benefits and limitations 13
6. Dataset limitations and recommendations 14
7. Conclusion 14

References 15

1. Introduction

1.1. Background

Data visualization is the act of representing data in the form of graphs. It attempts to convey information through the use of visual aids such as shapes, lines and labels which are much simpler to understand than the data in its original form. This makes it very friendly towards the average human who would find it more understandable than the walls of numbers of a dataset. Additionally, visualizing data could potentially reveal interesting details and hidden relationships within the data. These can then be utilized by the business to increase profits or to make improvements.

1.2. Purpose

In essence, data visualization provides many benefits to any willing to make use of it but in what ways can it be done? This is the question that this report will answer through the display of numerous interesting visualizations of a chosen dataset.

The dataset in question contains sales data of different products from 10 different Big Mart supermarket stores across multiple cities. Supermarkets are considered part of the retail industry which is concerned with the selling of goods to consumers, not for profit but for their own consumption (Farfan, 2020).

Hence, by the end of this report the reader will understand:

- The dataset chosen for this report
- How the dataset has been prepared
- What visualizations can be extracted
- The benefits and limitations of the charts
- Recommendations for improvements can be made

2. Dataset Description

Variable Name	Description	Type
ProductID	The unique ID number of the product.	Categorical
Weight	The weight of the product.	Numerical
FatContent	The amount of fat content contained within the product.	Numerical
ProductVisibility	The percentage of display area allocated for this product out of all products on display.	Numerical
ProductType	The category of product.	Categorical
MRP	The maximum retail price of the product.	Numerical
OutletID	The unique ID number of the store the product is sold in.	Categorical
EstablishmentYear	The year of establishment of the outlet.	Categorical
OutletSize	The size of the store.	Categorical
LocationType	The type of city the store is located in.	Categorical
OutletType	What type of outlet it is either a grocery store or supermarket.	Categorical
OutletSales	The sales of the product in that particular store.	Numerical

3. Data preparation

3.1. Correcting inconsistent categorical values

```
[9]
#Correcting inconsistent naming of FatContent nominal attribute
df.replace(to_replace={'reg':'Regular','LF':'Low Fat','low fat':'Low Fat'},inplace=True)
df['FatContent'].value_counts()

Low Fat    5517
Regular    3006
Name: FatContent, dtype: int64
```

Figure 1 .replace method to standardize the categories

After obtaining all the unique values of the FatContent attribute, it can be seen that there are multiple categorical values that are the same but are spelt different namely ‘reg’ for ‘Regular’, ‘LF’ and ‘low fat’ instead of ‘Low Fat’. This issue may confuse readers in thinking there are for separate FatContent unique when in actuality there are only two. To fix this problem, ‘. replace’ method is used to replace those values with their correct spelling, standardizing the values.

3.2. Missing Values

```
[10]
df.isnull().sum()

ProductID              0
Weight              1463
FatContent             0
ProductVisibility      0
ProductType            0
MRP                    0
OutletID               0
EstablishmentYear      0
OutletSize          2410
LocationType           0
OutletType             0
OutletSales            0
dtype: int64
```

Figure 2 Missing values are present in both Weight and OutletSize attributes

```
#Missing values in 'Weight' attribute is repalced with its mean and missing values in
# 'OutletSize' is replaced with 'Medium' which is the columns mode
df.fillna({'Weight':df['Weight'].mean(),'OutletSize':'Medium'},inplace=True)
df.isnull().sum()
```

```
ProductID            0
Weight               0
FatContent           0
ProductVisibility    0
ProductType          0
MRP                  0
OutletID             0
EstablishmentYear    0
OutletSize           0
LocationType         0
OutletType           0
OutletSales          0
dtype: int64
```

Figure 3 There are no missing values anymore after they are all imputated

It is found the dataframe contains numerous missing values through '.isnull()' method. Missing Values in the dataset can introduce bias within the data making it much more inaccurate and reduce its statistical power. The most common procedure to treat Missing Values is to delete records they occur in but since more than 20% of the data contains a missing value it is not the wisest thing to do. Instead, they should be imputed with an estimated value such as the mean, mode or median of their respective attributes.

The missing values resides in both the 'Weight' and 'OutletSize' attributes. So, the '.fillna' method will be applied to these two attributes. 'Weight' attribute is a numerical attribute therefore, missing values within it will be filled with the mean of the attribute. 'OutletSize' is a categorical attribute so mode must be used instead of mean. Once both actions have been carried out, it can be seen that there are no more missing values after 'isnull()' method is run again.

3.3. Outliers

```
numeric=df.select_dtypes('number')
numeric=numeric.drop(axis=1,columns='EstablishmentYear')
numeric
```

	Weight	ProductVisibility	MRP	OutletSales
0	9.300	0.016047	249.8092	3735.1380
1	5.920	0.019278	48.2692	443.4228
2	17.500	0.016760	141.6180	2097.2700
3	19.200	0.000000	182.0950	732.3800
4	8.930	0.000000	53.8614	994.7052
...	...	...	...	...
8518	6.865	0.056783	214.5218	2778.3834
8519	8.380	0.046982	108.1570	549.2850

Figure 4 A dataframe containing only numeric attributes

```
q1=numeric.quantile(q=0.25)
q3=numeric.quantile(q=0.75)
IQR=q3-q1
lowerbound=q1-IQR*1.5
upperbound=q3+IQR*1.5

not_outlier=numeric[~((numeric<lowerbound)|(numeric>upperbound)).any(axis=1)]

no_outler_df=df.loc[not_outlier.index]
no_outler_df
```

Figure 5 Code for keeping only non-outlier rows

Outliers can increase variance within the data, causing a skew in its distribution which can cause misleading interpretations of the data. In order to prevent that from occurring, they must be removed.

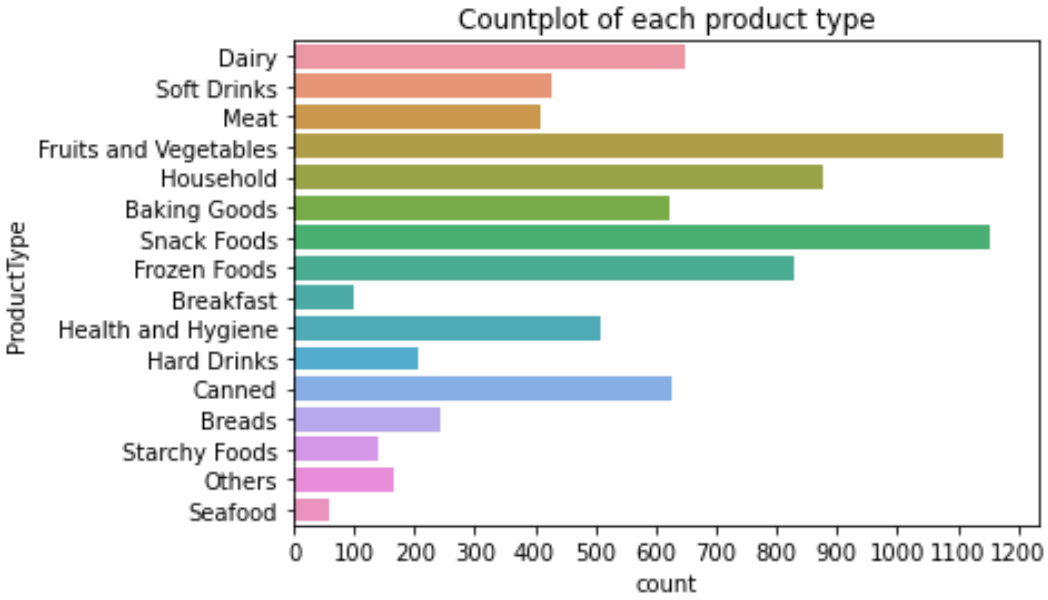
Outliers are values that are above the upper quartile, Q3 and below the lower quartile, Q1 by 1.5 times the interquartile range (IQR). Since outliers can only exist in numerical variables, a new dataframe is made consisting of only numerical variables from the dataset. The quantiles of those variables are then

found with the '. quantile' method. The results are then use to calculate the upper and lower bounds of each variable.

Next, a new dataframe called 'not_outlier' is made. It contains only rows that are not considered outliers. The index of that dataframe is then used to filter for non-outlier rows in the original dataframe but this time all attributes are present. The result is saved in a variable called 'no_outlier_df'.

4. Data Visualization

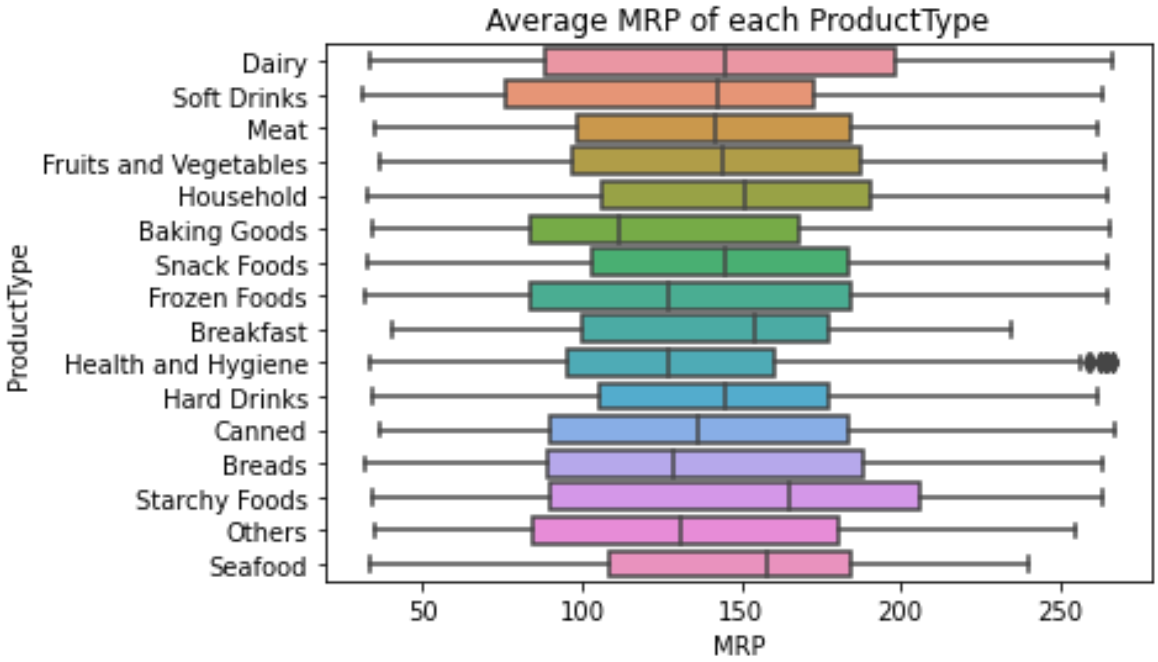
4.1. Count of each ProductType



This is a visualization of the number of each ProductType sold in the whole dataset. There are 16 different types of Products that can be sold. Among all of them 'Fruits and Vegetables' have been sold the most with 'Snack Foods' coming to a close second. This must be due to how cheap they are and how they can be purchased in large quantities. Furthermore, both have their own appeals. Fruits and Vegetables are considered healthy foods, bringing many health benefits. Snack Foods are tasty despite being bad if consumed regularly. Both product types are also very compact and portable, allowing them to be brought anywhere for any occasion.

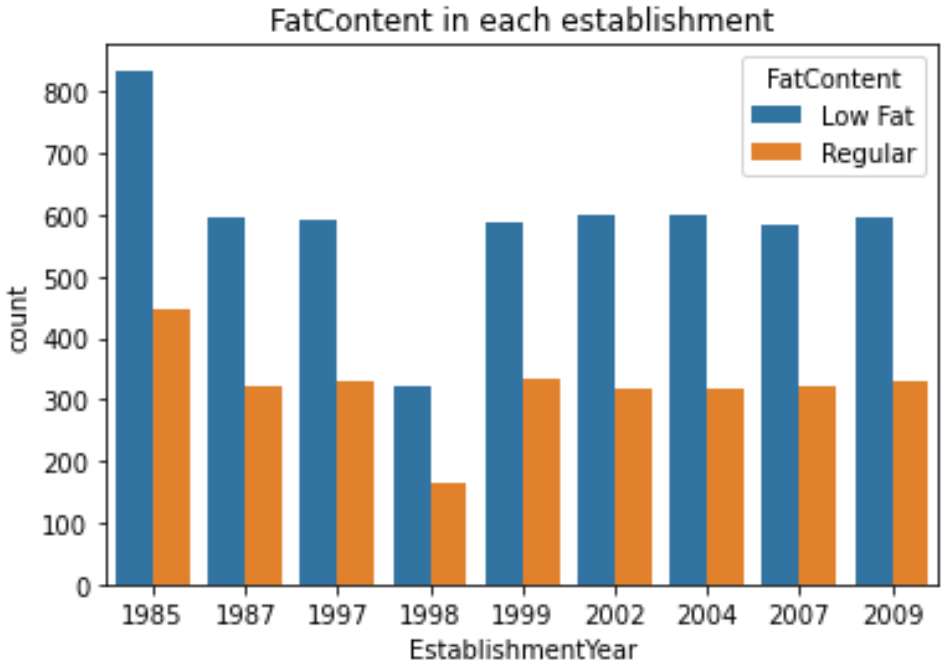
Besides that, the least sold product type is Seafood. This is probably due to the cuisine culture of the country not being fish based. In addition to that, fish in fish dishes require much preparation in the form of de-scaling and gutting the fish.

4.2. Average MRP of each ProductType



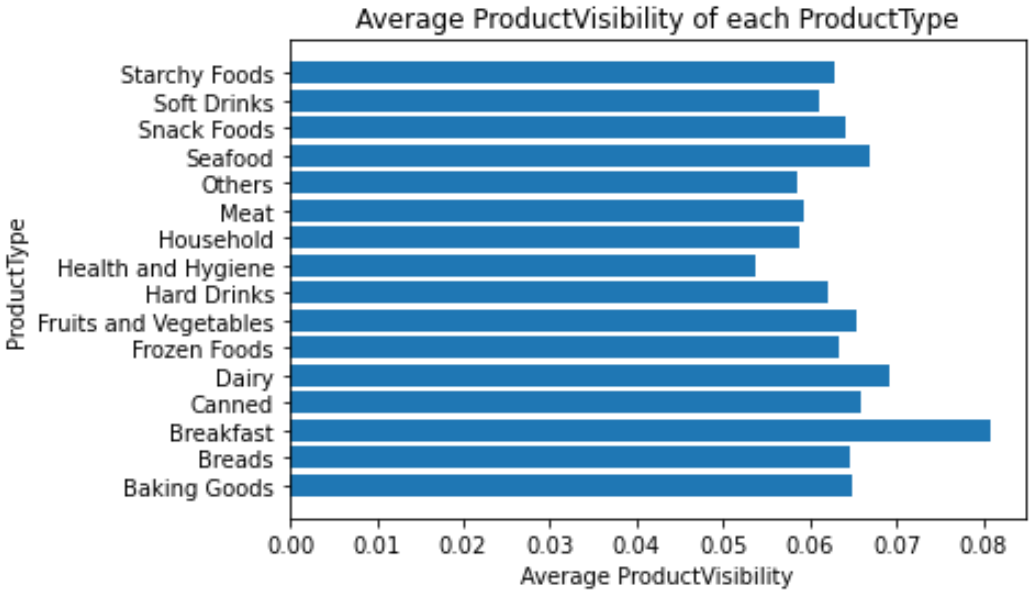
This visualization shows boxplots of MRP for each ProductType where MRP is on the x-axis while ProductType is on the y-axis. This box-whisker graph shows the distribution of max retail price of each product type. After examining the figure, it can be seen that the average(q2) MRP of 'Starchy foods' is the highest among others. This may be due to it being in low supply hence, an increase in retail price (Kramer, 2021) . Another reason could be due to the cost of purchasing it from suppliers which is reflected in its retail price.

4.3. FatContent in each establishment



This visualization shows a countplot for FatContent in each establishment year. From 1985 to 1987, it can be seen that the number of LowFat products have dropped significantly. The cause may be attributed to less demand for that type of product. As to why? It has drop in demand. One hypothesis is that the prices for LowFat products have risen during that period, leading to less sales. Another is they have fallen out of fashion also leading to less sales. In 1998, sales dropped even further back rebounds back in the next year.

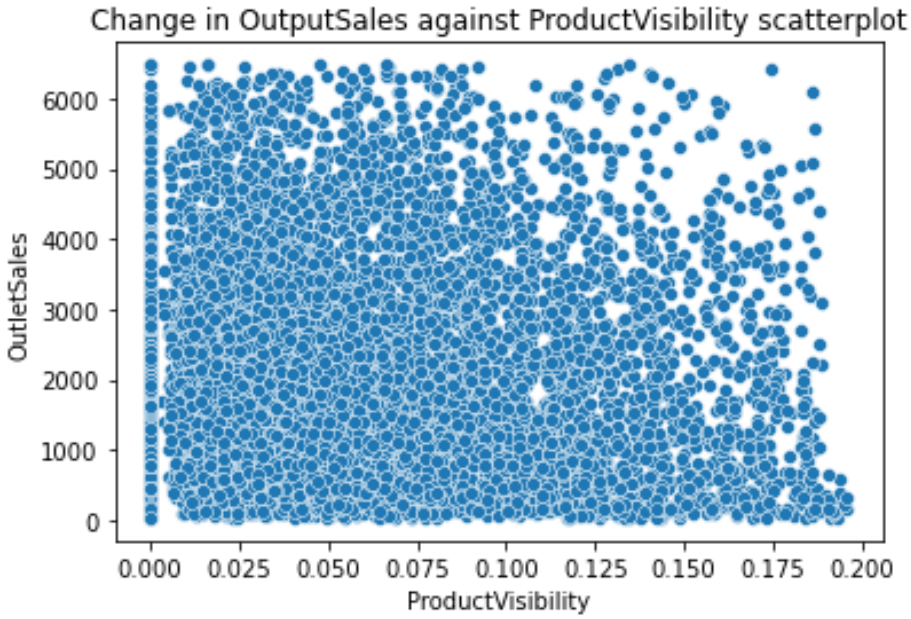
4.4. Average ProductVisibility of each ProductType



This horizontal bar chart plots average ProductVisibility in the x-axis against ProductType in the y-axis. The aim of this visualization is to give an idea how much exposure each product receives on average.

The graph shows that breakfast has higher product visibility on average than all other types. This may be due to the large amounts of space that breakfast products take like cereal boxes so a big portion of the store's display area must be set aside for them.

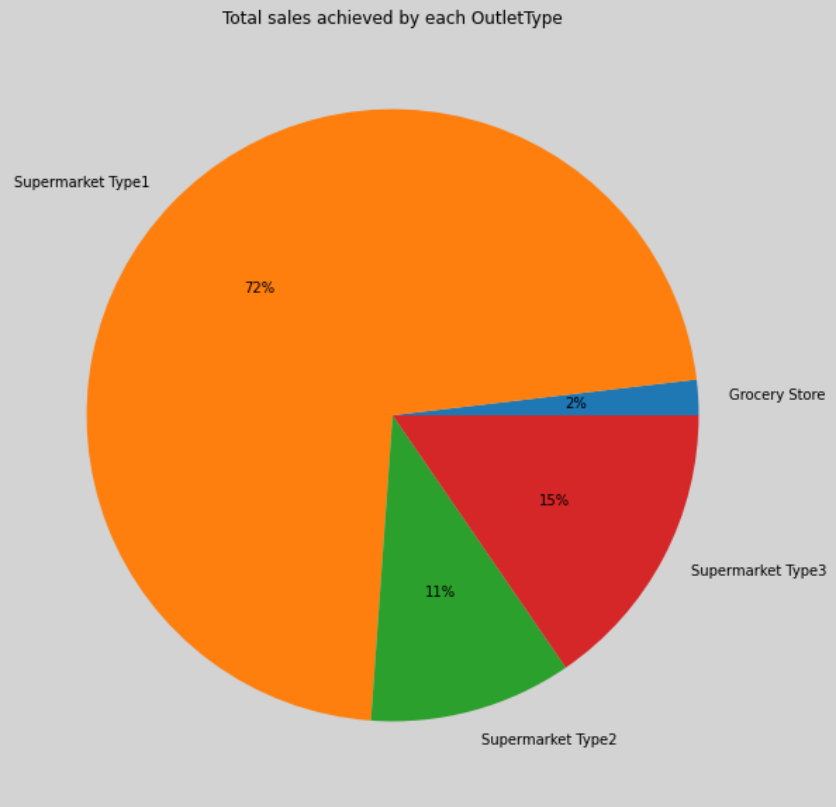
4.5. Change in OutputSales against ProductVisibility scatterplot



This is a scatterplot which plots ProductVisibility on the x-axis against OutputSales on the y-axis. Based on the graph, it shows that as product visibility increases the less points there are. Likely due to less products are sold at high values of product visibility.

The points are also distributed in a way which shows a negative correlation. Hence, the higher the product visibility the lower the outlet sales.

4.6. Total sales achieved by each OutletType



This last visualization is a pie chart of Total OutletSales by all OutletTypes. It shows the total sales achieve by each OutletType. According to the chart, Supermarket Type 1 has made the most revenue in sales. This type must have some qualities that all other types do not have. Perhaps type 1 has better prices for it products so people would prefer to purchase products from there.

5. Chart benefits and limitations

The charts allow thousands of rows to be condensed into useful visualizations. Many of which reveal interesting details that would not have been noticed from the original dataset. This makes them the perfect tools for presenting information to a large audience. However, the accuracy of the charts solely depends on the data itself. Regardless of how they are presented if the data is flawed so will the visualizations. To add more fuel to the fire, the stories made for each chart still have a chance of being completely wrong because they are more or less just assumptions as they are made without any in-depth knowledge of the organization the dataset originates from.

6. Dataset limitations and recommendations

Market basket analysis is a method of understanding associations between items purchased together (Li, 2017). The information gathered can then be used to predict what items the customer is most likely going to buy after the previous item. The product placement or layout of the supermarket can be modified based on those predictions so profit can be maximized. Nevertheless, each item that the customer has purchased together would have to be recorded which this dataset lacks. A column listing each item bought should next time be made for market basket analysis to be conducted.

In the same vein, the characteristics of the customers should also be recorded to understand their purchasing habits. Once understood, companies can then do targeted marketing that are in line with their habits to further boost the chances of the customers spending more. Characteristics such as gender, occupation, age and more can all contribute to the business.

7. Conclusion

Data visualization will bring many benefits to businesses willing to makes use of it. It has revealed many useful details which can assist in making informed decisions for the business. Yet, it is still limited by the data itself since it is just a different way of showing it so extra care must be exercised during the data collection and preparation stage. The dataset also lacks additional attributes that would assist in the business immensely so there are changes that have to be made next time.

References

Farfan, B. (2020, June 30). *What Is Retail?* Retrieved from The Balance Small Business: https://www.thebalancesmb.com/what-is-retail-2892238#:~:text=Retail%20is%20a%20very%20broad,and%20consumption%20by%20the%20purchaser.

Kramer, L. (2021, April 29). *How Does the Law of Supply and Demand Affect Prices?* Retrieved from Investopedia: https://www.investopedia.com/ask/answers/033115/how-does-law-supply-and-demand-affect-prices.asp#:~:text=There%20is%20an%20inverse%20relationship,quantity%20of%20goods%20and%20services.

Li, S. (2017, September 25). *A Gentle Introduction on Market Basket Analysis — Association Rules*. Retrieved from Towardsdatascience: https://towardsdatascience.com/a-gentle-introduction-on-market-basket-analysis-association-rules-fa4b986a40ce