
**MODULE CODE AND NAME:**

DA3304 – APPLIED PROGRAMMING FOR DATA ANALYTICS

**TITLE OF ASSIGNMENT:**

Assignment 1: *Data Visualizations*

**NAME:**

Muhammad Hidayat Bin Mohd Yusof (20FTT2910)

**SUBMISSION DATE:**

14th March 2022

**MODULE LECTURER:**

Norfarrah Muhd Masdi

**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

Data visualization refers to the graphical representation of data (tableau, 2022). It is the portrayal of data in the most simplified way possible which in the form of shapes and lines. Despite being simplified, it still can show new insight into the data that would have been discernable otherwise. This is useful for businesses who can use it to make better informed decisions.

Hence, this report intends to analyze a dataset regarding supermarket sales called ‘Supermarket_Sales.csv’. Its significance to the business will be examined through its contents. Then, the data will be prepared for the creation of meaningful visualizations where all of which would also be entirely examined.

Hence, by the end of this report you will have:

- Understood the industry of the provided dataset and the business benefits it gives.
- Understood the contents of the dataset specifically its attributes, the dependent and independent variables.
- Understood what steps are needed to prepare the data and the importance of each one.
- Learned all the different visualizations that can be inferred from the data as well as their strengths and weaknesses.

# 2. Business Understanding

To reiterate, the name of the dataset is ‘Supermarket_Sales.csv’ and it contains sales data collected from 3 different supermarkets in Myanmar. Supermarkets fall under the Retail Industry which concerns with the selling of goods to consumers for their own consumption.

### 2.1. Business Benefits

The main benefit this dataset brings to the businesses involved is it provides a method to record sale transactions that have occurred for analyzation. The dataset can be used to identify which products are popular in each supermarket as well as the characteristics of its purchasers. That information can then be used to make targeted advertisings or promotions to increase sales. It also provides an overview of the status of each supermarket such as their total gross incomes and their most popular payment method that can be used to rank them in terms of profits. Additionally, the general shopping experience among customers can be gauged to determine whether the supermarket is performing well or not.

### 2.2. Business Objectives

Supermarkets obtain profits by buying goods in bulk from suppliers and selling them at a higher price to consumers. Yet, the pricing must not be too expensive, or it will potentially make the customer reluctant to buy. Hence, the objective is to sell as many goods as possible while keeping each product at a reasonable price to encourage purchasing. A small portion of the customer's payment of goods is taken as tax by the supermarkets as another source of income.

## 3. Data Understanding

### 3.1. Description of Dataset

| Attributes | Description | Variable Type |
|---|---|---|
| Invoice ID | It is the identification number for the sales invoice | Independent |

| | | |
|---|---|---|
| Branch | The branch of the supercenter | Dependent |
| City | The city the supercenter was in | Dependent |
| Customer type | The type of the customer, which is either Members (ones that use a member card) or Normal (ones who do not) | Dependent |
| Gender | The gender of the customer | Dependent |
| Product line | The category the item purchased falls in | Dependent |
| Unit price | The price of each item in dollars ($) | Dependent |
| Quantity | The number of items purchased | Dependent |
| Tax 5% | The cost for the 5% customer tax fee | Dependent |
| Total | Total price with tax included | Dependent |
| Date | Date of purchase | Dependent |
| Time | Time of purchase | Dependent |
| Payment | The amount paid by customer | Dependent |
| cogs | Costs of goods sold | Dependent |
| gross margin percentage | Percentage of gross margin. | Dependent |
| gross income | Amount profited from item’s sale. | Dependent |

| | | |
|---|---|---|
| Rating | Customer stratification rating on their overall shopping experience (On a scale of 1 to 10) | Dependent |

### 3.2. Suggestions

A few new attributes that record the first, second, third and so on, items bought by the customer. They can be used to perform market basket analysis where the items that are likely to be bought together can be predicted. This information will then be used to arrange the placement aisles/goods in a way to encourage further customer purchases. In turn, increasing profit to be made.

Another suggestion, attributes that record personal information such as phone number, address and name so as to be able to contact them to collect more detailed feedback.

## 4. Data Preparation

### 4.1. Dummy Variables

In order to conduct market basket analysis, the product line attribute would have to be broken up into dummy variables. This process involves separating each entity into separate attributes, filled with either 0 or 1. 0 means the customer did not purchase the item while 1 means it was purchased. Since a customer can only buy a single type of item per transaction, only a single attribute will have a value of 1 while the rest has 0.

Moving on, the pandas method ‘get dummies’ is used as can be seen in the source code. As stated in the paragraph before, the product line attribute’s values are split into separate attributes namely, Health and beauty', 'Electronic accessories', 'Home and lifestyle', 'Sports and travel', 'Food and beverages', and 'Fashion accessories' with a prefix ‘Product line=’ so they can still be associated with the original column. This data is stored in a variable called ‘Dummies’. Then, a new dataframe Is created with only the first four columns of the original dataset called ‘with_dummies. The attributes in the ‘Dummies’ variable is then joined with the ‘with_dummies’ through the ‘join’ method followed by the rest of the columns from the original dataset. The result is a new dataframe with dummy variables.

## 4.2. Discretization and Binning

The ‘Rating’ attribute contains individual ratings of customers of their shopping experience from a scale of 1 to 10. Despite being simple to understand, it is not enough as it is to clearly show whether a provided rating is good or bad. Hence, to remedy this issue, the ‘Rating’ attribute should be placed into separate bins of different levels of ratings. The different categories are ‘Poor’, ’Below Average’, ‘Average’, ‘Good’ and ‘Great’. Respectively, their intervals are 0,2,4,6,8,10.

The process for the discretization and binning of the ‘Rating’ attribute requires the use of the pd.cut () method to place each occurrence in one of those bins. The data to be discretized is the ‘Rating’ attribute thus, the data parameter will have it as a series. Inside the label parameter will be a variable called ‘rating_groups’ which contains the categories stated earlier and of course the bin parameter will contain a list of intervals forming the bins. The result will then be stored as a new series called ‘new_ratings’ with the ‘Rating’ column renamed to ‘Old Rating’. After that the series will be joined with the previously created dataframe ‘with dummies’ forming a new dataframe called ‘NewR_df’.

## 4.3. Regex on time

The next technique involves the ‘Time’ attribute and the use of regular expressions. Time is much different to normal numbers whereas each digit can go up to number 9 but when digits are adjacent with each other they are limited, specifically up to 24 for the hour side and up to 60 for the minute side. Hence, this attribute cannot be treated as numeric despite its use of numbers. For that reason, it shall be treated as a string for the sake of progress.

The ’Time’ attribute contains the exact times of the transactions between the customers and counter of the supermarket. However, despite being easily legible on its own, the attribute is a chore to read if the intention is to just find the time of day a transaction occurred. Would it not be simpler if there was an attribute that displays the time of day it occurred? Therefore, the purpose of regex in this part is to read the time of each transaction and determine whether it is in the morning, afternoon, or evening.

Before regex can be used the patterns that are to be searched must be created. In this case, it is the time span for each phase of the day which are 00:00-11:59 for the morning, 12:00-18:59 for the afternoon and 19:00-23:59 for the evening. Each pattern is then compiled in their own individual variables namely regex1, regex2 and regex 3. Since a new attribute must be created especially one whose values are dictated by another attribute in the same row, then a special python built-in function called .map() is the best choice for the job. Within it, is another smaller function called lambda x which enables each ‘Time’ value to be used as an input and replaced with another value as an output based on a provided expression (Sharma, 2020). The expressions to be passed are ‘regex.sub’ which will check each value for matches with the pattern and if there is, will be replaced with the specified value. Due to the fact ‘regex.sub’ can only

accept a single pattern at a time, three separate mappings must be made where all three will look for matches with a specific phase and replace those matches with the correct time of day. Once all values have been replaced, the result is an entire attribute consisting of values from one of the three categories which are ‘Evening’, ’Morning’ and ‘Afternoon’. This new attribute will be saved as a new series called ‘Time Of Day’ and be joined with the previously created dataframe ‘NewR_df’ to form a new dataframe called ‘DF_With_TOD’.

# 5. Data Visualization

## 5.1. Exploration results

**Pairplot of supermarket_sales(prepared)**

Before data visualization can begin, it is good practice to create a pairplot for the whole dataset. A pairplot assists in finding relationships between variables by plotting pairwise relationships across all combinations of attributes (SL, 2019). It provides an overview of the dataset which will help in planning further visualizations. It is created with the seaborn method ‘sns.pairplot’ and the parameter “diag_kind=’kde’” is passed.

After analysis, it can be seen that ‘Total’, ’Cogs’ and ‘Tax 5%’ variables show a strong positive correlation with each other, but it is most likely caused by each variable being involve in each other’s calculation therefore, these correlations should just be ignored. Aside from that, all other combinations show uninteresting results. However, the nominal attributes are omitted from this visualization so there is still more to find.

**Different ratings across all cities and branches**

This visualization features two counts plots with two different x-axis values. The first is the ‘City’ attribute containing city names ‘Yangon’, ’Naypyitaw’ and Mandalay. The second is the ‘Branch’ attribute with branches a, b, and c. Both plots are grouped by the ‘New Rating’ attribute which means they will both show the counts of customers that have provided different ratings of their shopping experience across all cities and branches.

The first plot shows that among all cities there were no ‘poor’ ratings but there exist some ‘Below Average’ ratings in each city. The highest of the 3 is the city of ‘Mandalay’ but yet, it also boasts the highest number of good ratings. Nevertheless, they cannot be ignored and as a result further investigation should be made into the cause. All 3 three cities consist of ratings of average or more while Naypyitaw has the highest number of great ratings. Perhaps the supermarkets there are doing something right so whatever is being done there can be implemented to the rest.

**Payment type popularity among customer types**

This visualization shows a histogram where the x-axis is the ‘gross income’ attribute, and the y-axis is the count. There are 10 bins with equal intervals of 10. Each are colored based the number of values from the ‘Time Of Day’ attribute that are present inside each bin. Furthermore, the distribution of each value are represented by a line across all bins. It is the effect of enabling Kernel Density estimate (KDE) through the method’s arguments.

It can be quickly recognized that much of the income is made in the afternoon. The other two are seen

**Average & Total gross income against time of day**

This visualization contains two horizontal bar plots. The First plots Average gross income in the x-axis against Time of Day in the y-axis. The second shares the same y-axis but its x-axis is the Total gross income.

After examining the first plot, it is apparent the evening makes the most income on average while the other two phases follow close behind. Yet, when looking at the second plot, surprisingly the afternoon has the most income in total despite not making the most on average.

**Average cogs against customer type**

This next visualization is also a horizontal bar-plot. This time the Average cogs is plotted against Customer type. This one reveal that the average cost of goods sold to customers who are members are higher than normal customers. This is a problem as it would make becoming a member less appealing which is detrimental to the supermarket.

**Boxplots of unit prices in each city**

As stated in its title, this a boxplot that plots Unit Price attribute in the x-axis against the city attribute in the y-axis. An interesting observation from this visualization is that the median of Naypyitaw is more to the right than the others.

This indicates that customers there generally have bought more expensive items than customers of other stores.

**Count of product line between both genders**

This is a count plot made with the cat plot method as it allows the separation into multiple plots based on a categorical variable. This categorical variable is specified in the ‘col=’ parameter of the method. On the x-axis is the count and on the y-axis is the ‘Product line’ attribute. In the ‘col=’ parameter, the attribute ‘Gender’ has been specified therefore, there will be two different plots with different conditions. One is for Males; the other is for females. The goal of this visualization is to determine what product lines are popular in both genders. In this case, among females the most purchased product line is ‘Fashion accessories’ while among males it is ‘Health and beauty’.

**Payment type use in different total values**

This is a distplot of the ‘Total’ attribute, where each bin is colored to show the most used payment type. A displot is similar to a histogram where it plots only a single variable but what differentiates it is the addition of a line which graphically depicts the data’s distribution (JournalDev, 2022). This visualization will reveal what payment type is favored when paying for goods of a certain price. From what can be seen on the figure, ’Cash’, ‘credit card’ and especially wallet is used many times at low total prices but as it rises (move to the right), their use gets lower due to less people paying for such amounts. ‘Credit card’ shows a less dramatic reduction in use as it is most likely preferred for making large payments.

**Change of Total against Old rating regplot**

Comparable to a distplot, a regplot is just a variation of a familiar visualization chart. This time it is a scatterplot but with an addition of a line to show the change in direction of the overall data. This regplot plots ‘Total’ in the x-axis against ‘Old Rating’ in the y-axis. The graph does not show any interesting relationships between the two variables. Although, datapoints appear much less as ‘Total’ increases, its cause can be easily deduced as people are less likely to make expensive purchases which means they won’t leave a rating.

**Count of different product lines in each city**

The last visualization produced is a count plot of the product line attribute (x-axis) but is split into three separate graphs representing each city. Hence, each of the three count plots will show all types and the number of product lines bought from supermarkets in a particular city. Starting with Yangon, the most frequently purchased item comes from the home and lifestyle product line. In Naypyitaw, it is the food and beverages product line. Finally, in Mandalay it is the sports and travel product line.

## 5.2. Important visualizations

Among all visualizations the four most important ones to the business are:

## I. Different ratings across all cities

Customer feedback is important to any business. If a customer gives a poor rating due to a bad shopping experience, they may never come back to shop there ever again which means a loss of potential income. Furthermore, the issues that caused them to give a poor rating would still be unsolved leading to further poor reviews from customers impacted by them and more loss of potential income. Hence, it is imperative to have a method to obtain feedback so management can be aware of the existence of an issue in order to immediately act.

## II. Average cogs against customer type.

Memberships provides supermarkets with customer data whenever the membership cards are used. Data such as the customer's name, address, purchases, and more are collected. These are used to conduct targeted marketing based on the customer's purchasing habits or for analytical purposes. Their use often come with benefits such as discounts, loyalty points and more. This is to entice normal customers to obtain membership for its benefits in exchange for their data.

However, this visualization has shown that members are paying more on average than normal customers. This detail breaks the perception that obtaining membership places oneself in a better position compared to without it and would discourage them from obtaining it. Moreover, current members may even withdraw their membership if this information is leaked to the public

Therefore, without this visualization such an alarming detail would have gone unnoticed, and no action will be taken against it.

### III. Count of different product in each city

The benefits this visualization provides is the ability to predict customers reception towards a product. This is powerful because it allows product launches to be set up for success by releasing them in supermarkets with right people. For example, a new product, a new type of bicycle is to be launched so rather that releasing it on any random supermarket, would it not be more fruitful to release it on a supermarket who's most popular sold product is from the 'Sports and travel' product line. The insights obtained from this visualization could also be used to entice suppliers to sell their products in the one of business's supermarkets.

### IV. Average & Total gross income against time of day

This visualization has shown that the majority of the income was made is the afternoon which is unsurprisingly known as the busiest time for the supermarket. People usually work in the morning and get off of work in the afternoon where they would make a short shopping run before heading back home. Even though this is good for business, it could introduce a few other issues. Firstly, it would create more congestion at the supermarket whether it be vehicular or human, increasing the difficulty in finding parking spots, increasing waiting time at counters, and creating the perfect environment for the spread of disease. Secondly, with so many people purchasing goods in such a short span of time, a supply problem may occur. Therefore, with the assistance of this visualization that busy times like that can be anticipated and prepared for.

# 6. Recommendations

The visualizations have provided much unseen insights into the dataset. Displayed in a simple yet detailed manner that is quickly produced through a few lines of code. Additionally, they can be displayed in however way that is desired as they are very customizable. Nevertheless, they still have a few weaknesses. For one, no matter how good it is displayed, the data will always have the possibility as being false. Visualizations can never change the accuracy of the data hence if it was flawed in the beginning so will the visualization be. In addition, the assumptions made from visualizations are usually made without knowing the full detail of how the business operates or not made from a place of experience. Hence, as more information is released, there is good chance of those preconceptions to be completely wrong.

Since data visualizations, require basic knowledge in coding. It is to the industries' best interest to train talents in its use so they can utilize it as a tool for the quick sharing of information. The widespread use of data visualization in the retail industry will speed up progress made and help the industry reach new heights.

Furthermore, all people within the industry should also learn the practice of data cleaning. This is to combat inaccuracy in visualizations further boosting their effectiveness.

# 7. Conclusions

To summarize, data visualization is a method that can reveal new facets of a dataset. That can then be used to assist businesses to make better informed decisions. The 'supermarket_sales.csv' dataset contained large reserves of information regarding the retail industry, but as good practice still had to be

prepared. Many new attributes were created without the importation of foreign data and new options of visualization were opened.

The dataset was thoroughly explored, and numerous graphs were made to show interesting findings within the data. Some had very important significance to the industry/business, concerning with its future so was explained completely. Those few visualizations could bring fortune and improvements. However, despite all its strengths there still exists some weaknesses. But, through the close cooperation of everyone within the industry, these weaknesses can and will eventually be beaten.

Through the help of data visualization, the business objectives can be achieved as it can ensure prices are kept in check and do not rise up too high. Also, it can assist in identifying the target audience of products for targeted marketing maximizing profits. Customer feedback will always be analyzed so any problem amongst customers that may arise can immediately dealt with.

# References

JournalDev. (2022, 3 14). *Seaborn Distplot: A Comprehensive Guide*. Retrieved from journaldev: https://www.journaldev.com/39993/seaborn-distplot#:~:text=A%20Distplot%20or%20distribution%20plot,with%20different%20variations%20in%20it.

Sharma, A. (2020, March 12). *What are Lambda Functions? A Quick Guide to Lambda Functions in Python*. Retrieved from analyticsvidhya: https://www.analyticsvidhya.com/blog/2020/03/what-are-lambda-functions-in-python/

SL, S. (2019, September 29). *PAIRPLOT VISUALIZATION*. Retrieved from medium: https://medium.com/analytics-vidhya/pairplot-visualization-16325cd725e6#:~:text=Pairplot%20visualizes%20given%20data%20to,attractive%20and%20informative%20statistical%20graphics.

tableau. (2022, March 14). *What Is Data Visualization? Definition, Examples, And Learning Resources*. Retrieved from tableau: https://www.tableau.com/learn/articles/data-visualization