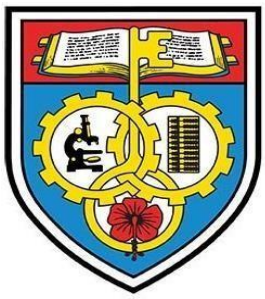


TARC

TUNKU ABDUL RAHMAN
UNIVERSITY COLLEGE

COURSE NAME:	**Introduction to Data Science**	
COURSE CODE:	**AACS1573**	
ASSIGNMENT TITLE:	**Data Science Application**	
SEMESTER:	**202301**	
SUBMISSION DEADLINE:	**Week 14 Monday before 12pm**	
TOTAL MARK AWARDED:	**Task Report Total:**	**80%**
	Presentation Total:	**20%**
	Assignment Total:	**100%**
WEIGHTAGE TO FINAL MARK:	**60%**	
STUDENT NAME and Student ID	**1. YONG VIN SEN (22WMD03343)** **2. CHUA TENG HUI (22WMD02924)** **3. WONG WAI KIN (22WMD02893)** **4. SIOW JUN JIE (22WMD02900)**	

AACS1573 Introduction to Data Science

LEARNING OUTCOMES

Upon completion, students are expected to achieve the followings: -

CLO2:	**Interpret data into actionable insights that are found through data science process (C3, PLO5)**
CLO3:	**Perform Exploratory Data Analysis (EDA) to analyse the main characteristics of datasets. (P3, PLO3)**

GUIDELINES TO STUDENTS

Students are required to observe the followings: -

1. Students will form into groups of **3 - 4** members per group. Every team member is expected to contribute and participate actively in the entire process of completing the assignment. Sharing of ideas and assistance in the completion of assignments among members is required.
2. Upon completion of each assignment task, students are required to prepare an assignment task report.
3. All references and citations shall use the Harvard Referencing Style.
4. The assignment should be typed using 1.5 spaces between lines in 12 Times New Roman point-font, not more than 5 pages.
5. Plagiarism is strictly prohibited. Marks are awarded for your own (original) analysis. Therefore, use the time and information to build a well-constructed report.
6. Check carefully the submission date and the instructions given with the assignment. For late submission, there will be a reduction of absolute marks from the mark's score submitted:
    i. Late 1 to 3 days after the deadline of submission: minus 10 marks;
    ii. Late 4 to 7 days after the deadline of submission: minus 20 marks;
    iii. Late more than 7 days after the deadline of submission: 0 marks
7. The presentation will be on Week 14 (Duration at least 5 to 10 minutes per student depending on the lecturer's allocation).
8. Presentation requirements including communication skills, visual aids and personal grooming are the criteria for obtaining the marks allocated.
9. Assignment format:
    a) Front Cover
    b) Assignment Marking Criteria
    c) Table of content
    d) Content following the Name of the assigned tasks and the corresponding task number.
    e) References and Appendix after each assignment task report.

ASSIGNMENT

Introduction

Data Science has a wide variety of applications. It is used in several fields ranging from health, education to transportation, and manufacturing. Various industries are using Data Science to boost their production, make smarter decisions, and develop innovative products that are tailored for customer needs.

Example of data science case studies from the following fields but not limited to:

- Manufacturing
- Pharmaceutical Industries
- Biotech
- Education
- Business

Task

- You are required to choose one of the above data science case study fields and describe clearly the data science application you have chosen.
- Your documentation should consist of the following:
    1. Introduction (include definition and significance of data science)
    2. Case study: Name and description (field and application)
    3. Data Science Process for the selected case study
    4. Results and Discussion
    5. Conclusion with the advantages and disadvantages of the selected case study
- Present your assignment during week 12 (practical class).

END OF ASSIGNMENT

MARKING CRITERIA

Your written assignment and presentation will be assessed against the following criteria.

ASSIGNMENT RUBRICS (CLO3)

No.	Criteria	Evaluation/Marks						Mark Achieved
		0 – Nothing Presented/ Plagiarized	**2 – Developing**	**4 – Approaching Expectation**	**6– Complete**	**8 – Excellent**	**10 – Beyond Expectation**	
1.	**Logic and Organisation**	Does not develop ideas cogently, uneven and ineffective overall organisation, unclear introduction and conclusion	Still developing ideas cogently, uneven and ineffective overall organisation, unclear introduction and conclusion	Develops unified and coherent ideas within paragraphs with generally adequate transitions; clear overall organisation relating most ideas together with good introduction and conclusion	Develops ideas cogently organises them logically with paragraphs and connects them with effective transitions Clear and specific introduction and conclusion	Strongly Developed ideas cogently organise them logically with paragraphs and connect them with effective transitions Clear and specific introduction and conclusion n	High innovative idea cogently organised and logically sequenced with a clear and specific conclusion.	
2.	**Conceptual Understanding**	Does not respond using course content	Respond using appropriate and sufficient course content	Responds using appropriate and sufficient course content	Respond clearly and effectively using appropriate and sufficient course content and outside sources	Good Response in using course content and resources in explaining your understanding of the topic	Highly effective response in using course content and sources to explain the understanding	
3.	**Evidence**	Does not present data	Presents accurately some of the necessary data	Presents clearly and accurately most of the necessary data	Presents clearly and accurately all of the necessary data	Good Sequence and Logical thinking behind Presenting data	Most suitable content explaining the end to end points and with clarity	

4.	**Relevance of Content**	Irrelevant content	Partially relevant content is mentioned	Good relevant content is mentioned	Right appropriate Content is mentioned	Right content with clear logic content is mentioned	Highly effective content explaining the clear logic behind the topic	

No.	**Criteria**	**Evaluation/Marks**						**Mark Achieved**
		0 – Nothing Presented/ Plagiarized	**2 – Developing**	**4 – Approaching Expectation**	**6– Complete**	**8 – Excellent**	**10 – Beyond Expectation**	
5.	**In-text citation and Referencing**	No in-text citation and no References	There are in-text citation or References	There are in-text citations and References but the citation and references were not matching. References were not using the appropriate format.	There are in-text citations and References. The citation and reference were tallying. References were not using the appropriate format.	There are in-text citations and References. The citation and reference were tallying. References were using an appropriate format.	There are in-text citations, References, Bibliography. The citation and reference were tallying. References were using the appropriate format.	
6.	**Use of Language**	Imprecise or inappropriate choice of words	Express thoughts marginally	Appropriate choice of words	Uses rich choice of words and imaginative language	Good mix of words expressing the logic of thoughts clearly	Strong Effective use of Language expressing ideas clearly	
7.	**Conclusion**	Does not draw conclusion or inference	Draws valid conclusions or inferences	Draw valid conclusions or inferences supported by content	Draws clear and valid conclusions or inferences supported by content	Good inferences by relevant content representation and inference	Effective conclusion based on right content representation and inference	
8.	**Reading/ Research**	No research has been done	Non-relevant research	Relevant research	Relevant research displayed in the content	Relevant research displayed in the content with appropriate examples and evidence	Relevant research displayed in the content with appropriate	

							examples and evidence and recommendation.	
							TOTAL MARKS (80%)	

PRESENTATION RUBRICS (CLO2)

No	Evaluation Categories	0 - No submission	1 – Weak /Unsatisfactory	2 – Developing /Needs Improvement	3 - Satisfactory	4 - Very Good	5 – Exceptional /Excellent	Mark Achieved
1.	**INTRODUCTION** How well did the speaker set the scene	Did not appear/attend the presentation	Lack of interest and enthusiasm. The presented topic was not introduced and explained. Individuals were not introduced.	Shows a bit of interest and enthusiasm. The presented topic was introduced but vague and irrelevant. Introducing self but not the members.	Seems interested, but could be more informed on the introduction. All members were addressed, including self.	Interested and enthusiastic about the presentation. The introduction has been laid out properly and to the point. All members were addressed, including self.	Eager about the presentation. Exceptional introduction which includes the summary. All members were addressed, including self.	
2.	**CONTENTS** Was the objective identified? Was the presentation adapted to a wide range of audiences? Was it well organized?		No clear statement offered. Scope too broad or too narrow; lacks depth; AND uses too much technical language/ jargon. No clear information sequence; very difficult to follow.	Incomplete or unfocused. Scope too broad or too narrow OR lacks depth OR uses too much technical language/ jargon. Evidence of some organization but not in an optimal order; difficult to follow.	Reasonably clear. Reasonable scope and depth; lapse into detail that may not be accessible to the audience. Ideas presented in logical sequence; reasonably easy to follow.	Clear and concise. Good scope & depth without losing the audience in technical detail; a good learning experience. Presented in logical & interesting ways; easy to follow but not oversimplified.	Clear, concise. Engaging, and thought-provoking. Exceptional scope & depth; a true learning experience; exceeds expectations. Exceptional organization because the topic is complex.	

No	Evaluation Categories	0 - No submission	1 – Weak /Unsatisfactory	2 – Developing /Needs Improvement	3 - Satisfactory	4 - Very Good	5 – Exceptional /Excellent	Mark Achieved
3.	**EVIDENCE** Did the speaker demonstrate that actual work was carried out independently? Was careful thought put into the work		No appropriate evidence was presented to support the presentations central claims. None of the descriptions contains an explanation. Does not display a clear grasp of the subject being discussed.	Some evidence is present, but is either insufficient or not supportive of the main claims. Somewhat display some explanations but not clearly explain the significance of the central claims.	Evidence used to support the central claims is well chosen with some degree of detail. Some of the descriptions contain explanations that clearly explain the significance of the central claims. Displays a reasonable grasp of the subject. The opinion was clearly expressed.	Evidence well-chosen & detailed; the connection between argument & evidence is clear; opposing evidence considered.	Well-chosen, detailed, rich; highly compelling; opposing evidence considered and refuted. Each description contains an explanation that clearly explains the significance of the central claims. Displays excellent understanding of the subject. Opinion strongly expressed and supported.	
4.	**CONCLUSIONS CONFIDENCE (reading materials)** How well did the author conclude, summarize and recommend?		No apparent conclusions; no discussion of implications.	Conclusions are restatements of previous statements.	Brings closure with some synthesis but does not address implications.	Synthesizes the work; brings closure; allude to broader implications.	Synthesizes; brings closure; conveys real implications; suggest new perspectives.	
							TOTAL MARKS (20%)	

Table of Content

1.0 Introduction

1.1 Introduction to Data Science

Data science is a branch of study that combines subject-matter expertise, programming prowess, and understanding of math and statistics to derive practical insights from data. Data scientists build Artificial Intelligence (AI) systems that carry out operations that ordinarily call for human intellect by applying machine learning algorithms to numbers, text, pictures, video, audio, and other sorts of data. The insights produced by these technologies may then be transformed into real commercial value by analysts and business users (Jonathan, n.d.).

1.2 Significance of Data Science

Data Science plays a crucial role in collecting thousands of datasets that are useful for our analysis. Data science combines tools, methods, and technology to enable us to understand the pattern of datasets that may have either structured data or unstructured data by identifying and analyzing the data from many related sources. It also allows researchers to draw insights from the data that has been analysed.

Therefore, data science can also make predictions from the data and know what is the trend in the future. Data scientists can also draw conclusions from the data. Hence, data science is useful in many industries such as e-commerce, manufacturing, banking, healthcare, transport, finance and so on.

2.0 Case Study

Name: Supermart Grocery Sales

Field: Sales

Application: Business owners will face many adversities when they are managing their supermarts. On the other hand, customer's choices will vary their quantity of buying products in the supermart based on some situations too. Therefore, there are many aspects that will affect supermart grocery sales. Hence, our purpose of this analysis is to find out the factors that strongly affected the sales of supermart grocery.

Within this case study, we will analyse the customers' region, supermarts' located city, profit margin of the supermarts' city, category and subcategory of goods, customers' order date and order's discount. These are the main factors that might affect the supermart sales. All of these factors will be analyzed and recorded in a data table and it will be carried out as the proof that these factors might affect the data of the sales order.

3.0 Data Science Process

3.1 Data Preparation

Data preparation is the first step in the data science process, which involves reading the data from myriad sources and preparing it for analysis. This process involves reading the data that is compatible with the analysis tools being used, such as HDFS. This step also involves converting the data into the same format type like JSON, CSV, or Excel may be used depending on the form of the data.

The second step in data preparation is cleansing the data to ensure its accuracy, consistency, and quality. This can involve removing corrupt or problematic data, filling in missing values, and removing stop words or special characters. It may also involve normalizing the data to improve consistency and reduce redundancy.

3.2 Data Exploration

In this process, we tried to analyze and get some useful data from the dataset which are related to our title. In addition, we can analyze various factors from the dataset such as the customers' region, supermarts' located city, profit margin of the supermarts' city and so on. After that, we can thoroughly know and find out which factors will affect the supermart grocery sales. Through this process, we are able to find out the similarities, differences and outliers in order to identify the relationships between the factors and the sales. Therefore, some useful data and hidden potential information in the dataset can be figured out so that we will not get influenced by irrelevant data.

3.3 Data Representation

After data exploration, data representation is our next step. Data representation is a technique for transforming and describing the data. This process allows us to transform raw data into a dataset which can optimize computer memory and storage for a faster execution. For example, we transform and integrate the data into the appropriate data types like integer values should store in integer data type. Therefore, data representation can help us to interpret the data more easily and accurately and it ensures that the data is clearer and prepares for the upcoming step in our analysis.

3.4 Data Discovery

Data discovery is a crucial step in the data science process that involves conducting initial data analysis to comprehend the data, its characteristics, and potential uses. It includes collecting and evaluating data from different sources, using visualizations to identify patterns and correlations, and performing statistical analysis to test hypotheses. Through data discovery, we can extract insights from the Supermart Grocery Sales Retail Analytics Dataset such as the top-selling products and customer purchasing behaviors, which can aid in developing strategies to optimize sales and increase profitability. Furthermore, data discovery allows us to filter out less significant correlations based on statistical measures and discard less meaningful relationships, such as the relationship between the order discount and customer's region.

3.5 Learning from Data

Last but not least, learning from data is vital in this data science process. It requires a certain knowledge and creativity in the examination of a dataset through the statistical methods and machine learning systems. Through this process, we can learn types of useful data from the datasets. After going through those data, we can totally understand what the datasets are trying to interpret. For example, we are able to identify what factors will affect the sales after learning the datasets about the supermart grocery sales.

4.0 Result and Discussion

4.1 Customers' Region

By referring to the pie chart in Appendix 1, we are able to acknowledge the sales order of supermart grocery sales by customers' region in the state of Tamil Nadu, India. The state is divided into 5 regions that are included in the dataset including Western region, Eastern region, Central region, Southern region and Northern region. By observing the pie chart, we know that the Western region has the highest proportion of customers (32.05%) followed by the Eastern region that occupies 28.50% of the customers. The Central region and the Southern region also accumulate 23.24% and 16.20% customers respectively. Lastly, the Northern region has the smallest percentage of customers among all regions which is 0.01%. In summary, Western region customers shop more frequently than customers of other regions.

4.2 Sales Order Based on City

Based on the horizontal bar chart in Appendix 2, it shows all of the cities in the state of Tamil Nadu, India such as Kanyakumuri, Tirunelveli, Bodi and so on. The most sales order placed from the supermarts' city is Kanyakumari, which has 459 sales orders. Tirunelveli has the second most sales order placed from our supermarts' city after Kanyakumari, which has 446 sales orders. Then, the third most sales order placed from the supermarts' city is Bodi which has 442 sales orders. The least sales order placed from the supermarts' city is Trichy which has only 357 sales orders. In conclusion, the top 3 cities which have the most sales orders are Kanyakumari, Tirunelveli and Bodi.

4.3 Profit Margins Based on City

Based on the bar chart in Appendix 3, it shows the profit margin based on city. The most and second most profitable city is Karur and Bodi which have 26.36% and 26.03% of profit margins respectively. The least and second least profitable cities are Tenkasi and Namakkal which have 24.27% and 24.31% of profit margins respectively. In short, the most profitable city is Karur and the least profitable city is Tenkasi.

4.4 Category of Goods

According to the pie chart in Appendix 4 that represents the number of sales based on category, there are 7 main categories of goods that are sold in the supermart. The category that obtained the most sales is snacks, which is 15.15% from the total sales followed by eggs, meat & fish that have the proportion of 14.91%. The top third and fourth most sold items by category are fruits & veggies and bakeries which is 14.19% and 14.14% respectively. The fifth and sixth categories are beverages and food grains that have a similar proportion which accumulate 14.01% and 13.99% of total goods respectively. The last category is oil & masala which only includes 13.62% from total goods. In short, snacks is the most popular category and conversely oil & masala is the least popular category in the supermart.

4.5 Sub-Category of Goods

From the pie chart of the number of sales based on sub-category in Appendix 4, we can see that two of the most popular subcategories are health drinks and soft drinks which belong to beverages category that have the proportion of 7.19% and 6.81% of total goods respectively. On the other hand, the two least popular subcategories are dals & pulses and rice which belong to food grains category that have the percentage of 3.43% and 3.30% of total goods respectively. In short, we know that customers usually come to the supermarts to buy health drinks and soft drinks.

4.6 Order Date Based on Year

According to the pie chart in Appendix 5, we can determine the sales order quantities made by customers at the supermarts in the Indian state of Tamil Nadu, which is rising from 2015 to 2018. The pie chart plainly shows the lowest percentage of sales order quantities made by customers at the supermarts in 2015, which was 19.94%. In contrast, the largest percentage of sales order quantities made by customers at the supermarts in 2018 was 33.14%. Based on the pie chart, we can see that obviously the sales order of the supermarts keep increasing by years.

4.7 Discount

From the vertical bar chart in Appendix 6, the bar chart represents the relationship between sales order and the discount amount. When the discount rate reaches 17%, the total sales order only has 336, which is the least. On the other hand, 25% of the discount will have the most total sales order which is 438. In a nutshell, customers will mostly make sales orders when the supermarts are having a 25% discount.

5.0 Conclusions

In conclusion, we are able to conclude from our case study that various factors such as the region and city in which customers originate, the category and subcategory of the goods, will influence the sales of supermarts. However, there are several benefits and drawbacks to the factors that were selected for this case study.

One of the advantages of the case study is we are able to identify which category and subcategory of goods are the most bought by customers based on the dataset we have analysed. Therefore, we are able to predict which category and subcategory of goods should be restocked with priority to prevent out of stock problems so that when customers visit our supermart, they can buy the goods they want without running into any issues, especially a lack of stock. By doing this, it can increase the customer's satisfaction and retention towards the supermart. Hence, the sales revenue of the supermart will increase indirectly as well.

Moreover, by analyzing the sales order based on city, we are able to identify which cities have less customers compared to cities that have more customers. This is significant because the supermart will spend more time and money to do advertising and marketing in the area that has fewer customers in order to increase the supermart's popularity within the area. As a result, the customers of the supermart and the supermart's sales in the city might increase exponentially in the future.

However, we also find out some irrelevant factors after analyzing the dataset. For example, we cannot define the discount as a factor that affects the sales order. A high discount rate does not necessarily lead to an increase in sales orders. It is possible that

customers will only purchase the products based on their specific needs rather than being influenced by the discount rate. Thus, we can know that the correlation between discount rate and sales orders is relatively weak. To sum up, the analysis reveals that the discount rates do not significantly affect sales or profits, indicating that the customers are not heavily influenced by discounts when making their purchasing decisions.

Last but not least, we are able to conclude that having a large number of sales orders does not indicate that it will bring high profit margin to our supermart by comparing the results between appendix 2 and appendix 3. This is because the customers may buy different subcategories of products and different sub-categories have different profit margins (one product stands for one sales order) according to the bar chart in appendix 7. For instance, we only earn 23.82% of profit margin when we sell one spice. However, when we sell one noodle, we can have 26.34% of profit margin. Therefore, the profit margin of different cities is not closely related to the quantity of sales orders made by customers.

6.0 References

1. Jonathan, B. (n.d.). Introduction to Data Science. [online] Available at: https://fti.unai.edu/wp-content/uploads/2020/02/Introduction-to-Data-Science.pdf.

2. Embibe Exams. (2021). Data Representation: How Data is Represented with Examples - Embibe. [online] Available at: https://www.embibe.com/exams/data-representation/.

3. Blue-pencil.ca. (2018). Data Cleansing: What Is It and Why Is it Important? [online] Available at: https://www.blue-pencil.ca/data-cleansing-what-is-it-and-why-is-it-important/.

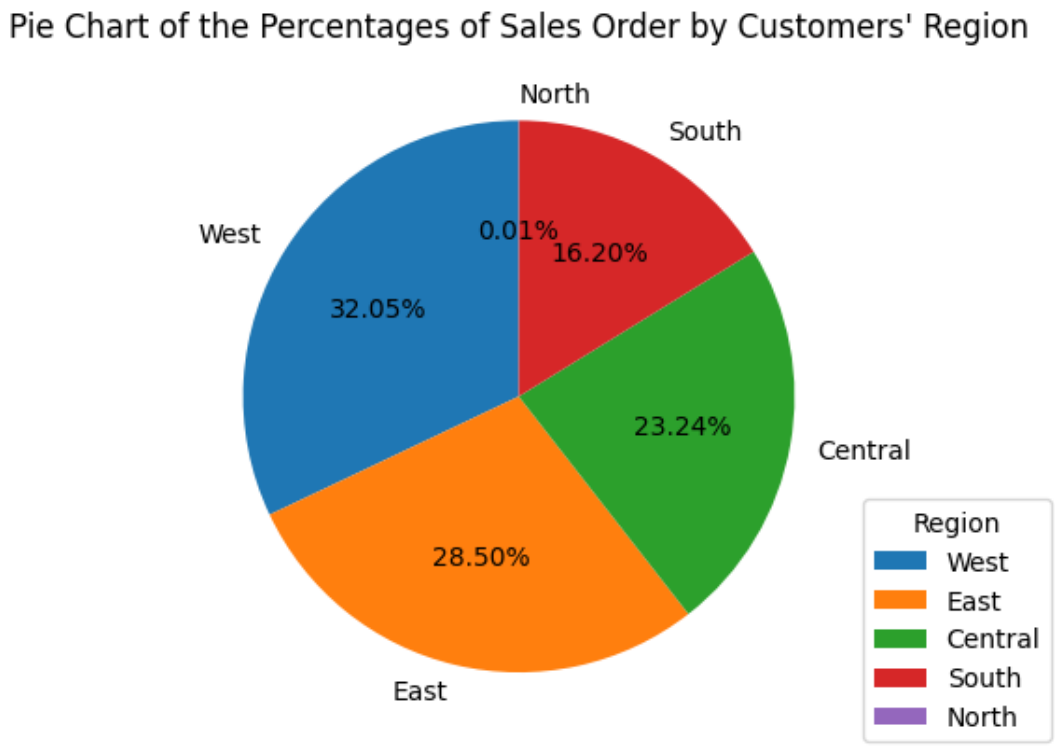
4. Anon, (2021). Matplotlib Pie Chart Tutorial - Python Guides. [online] Available at: https://pythonguides.com/matplotlib-pie-chart/#:~:text=Matplotlib%20pie%20chart%20autopct%20position [Accessed 13 May 2023].

5. www.leadfuze.com. (2021). Sales Data Analysis: 9 Ways to Help You Easily Make Revenue : LeadFuze. [online] Available at: https://www.leadfuze.com/sales-data-analysis/ [Accessed 13 May 2023].

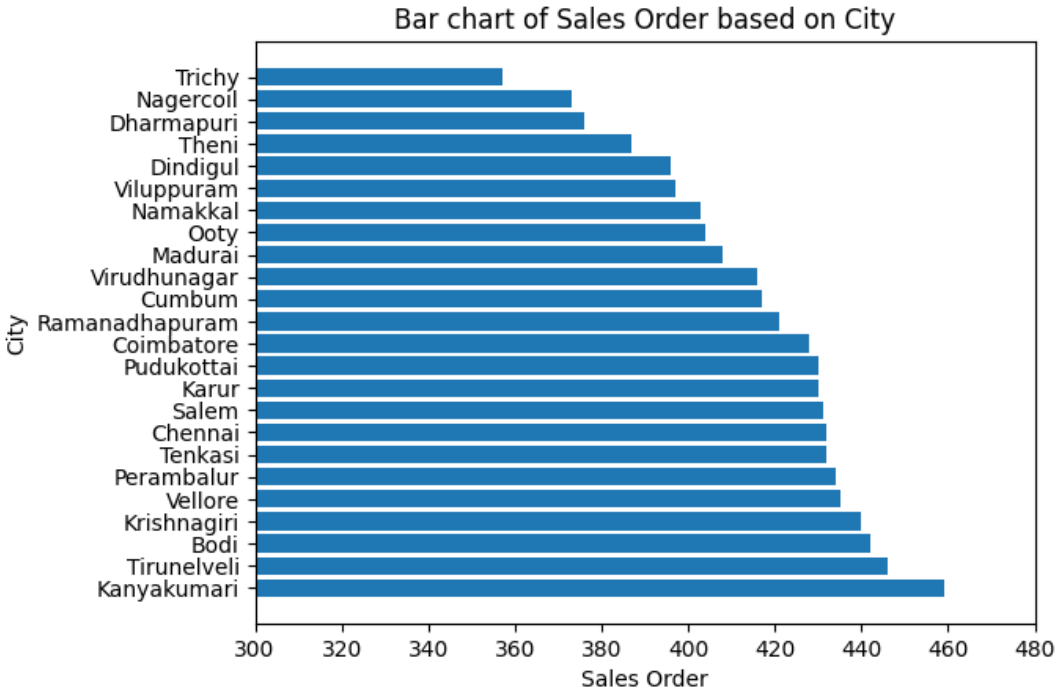
6. www.kaggle.com. (n.d.). Supermart Grocery Sales - Retail Analytics Dataset. [online] Available at: https://www.kaggle.com/datasets/mohamedharris/supermart-grocery-sales-retail-analytics-dataset.

7.0 Appendix

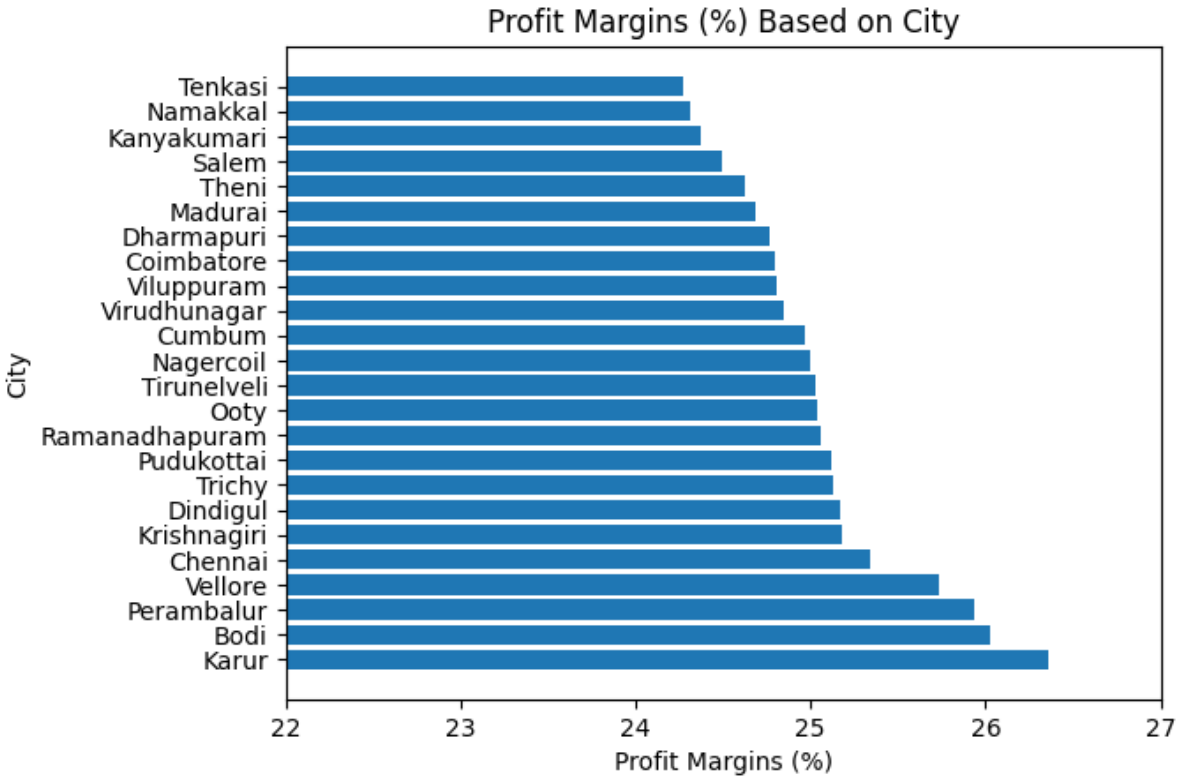
Appendix 1



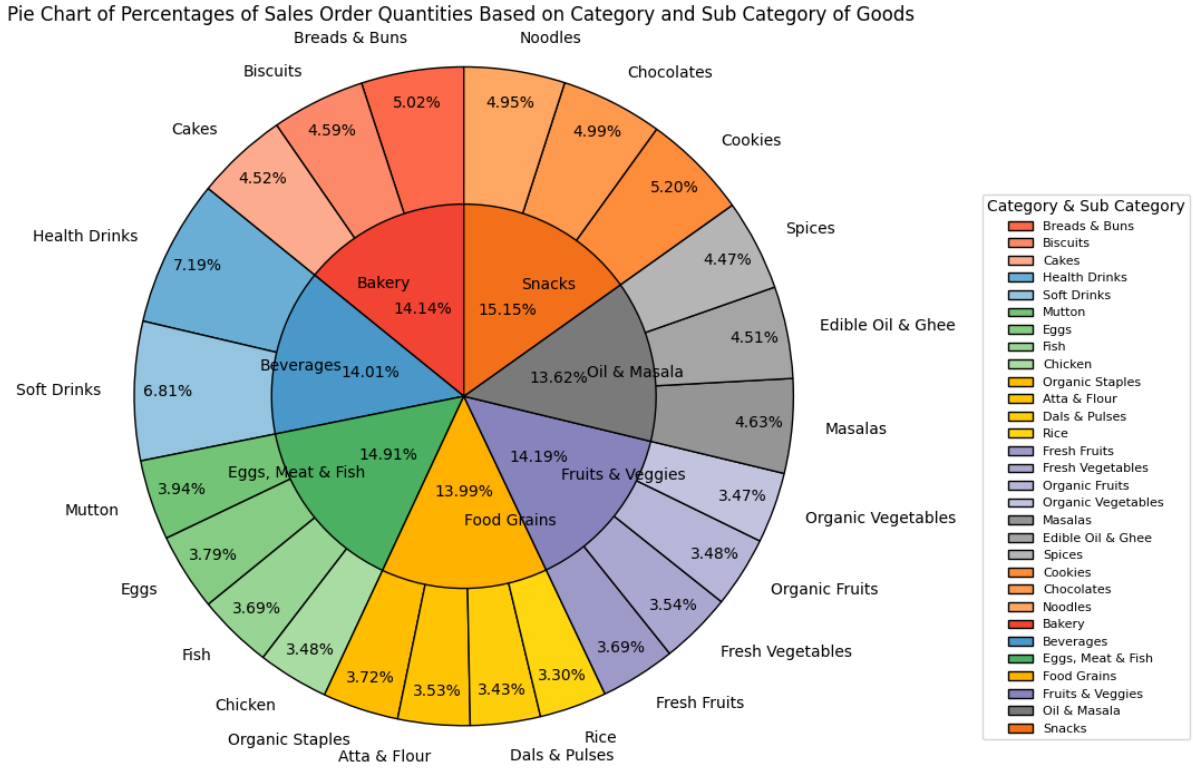
Appendix 2



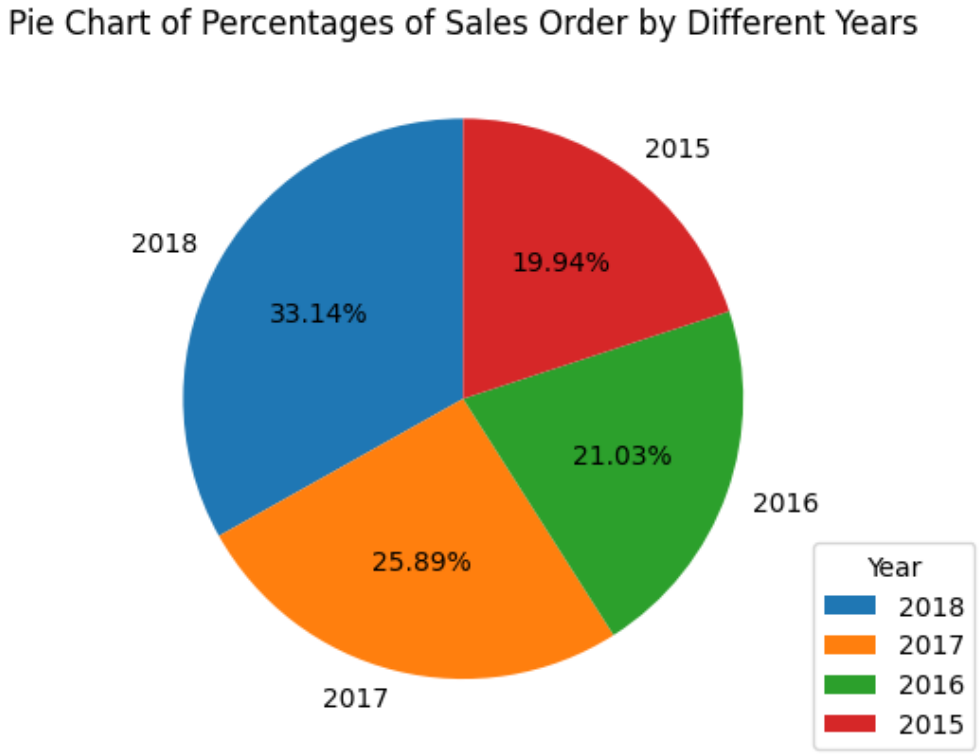
Appendix 3



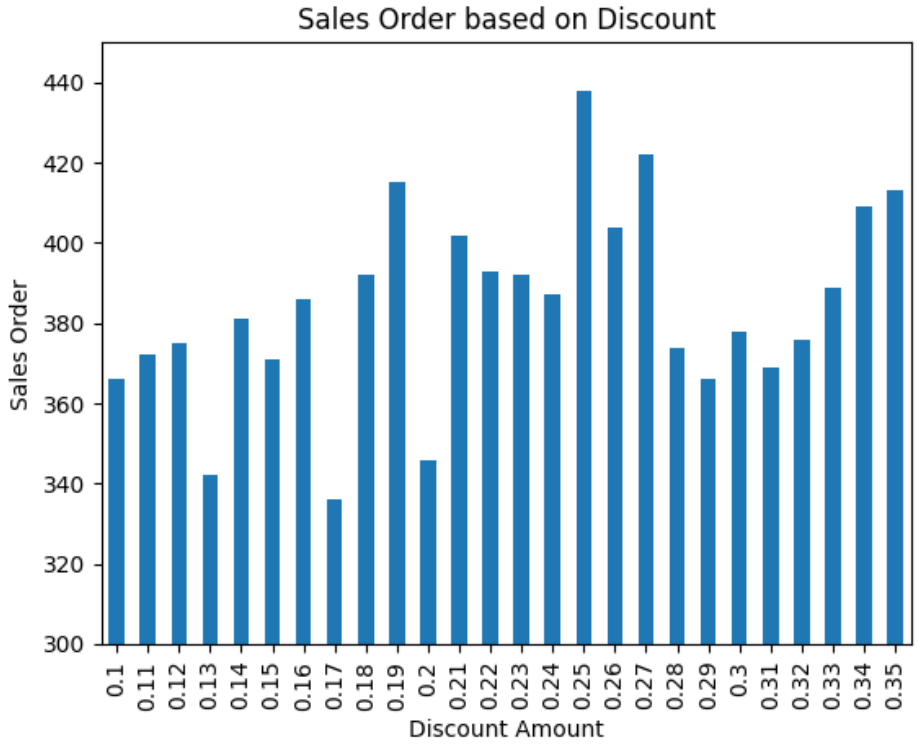
Appendix 4



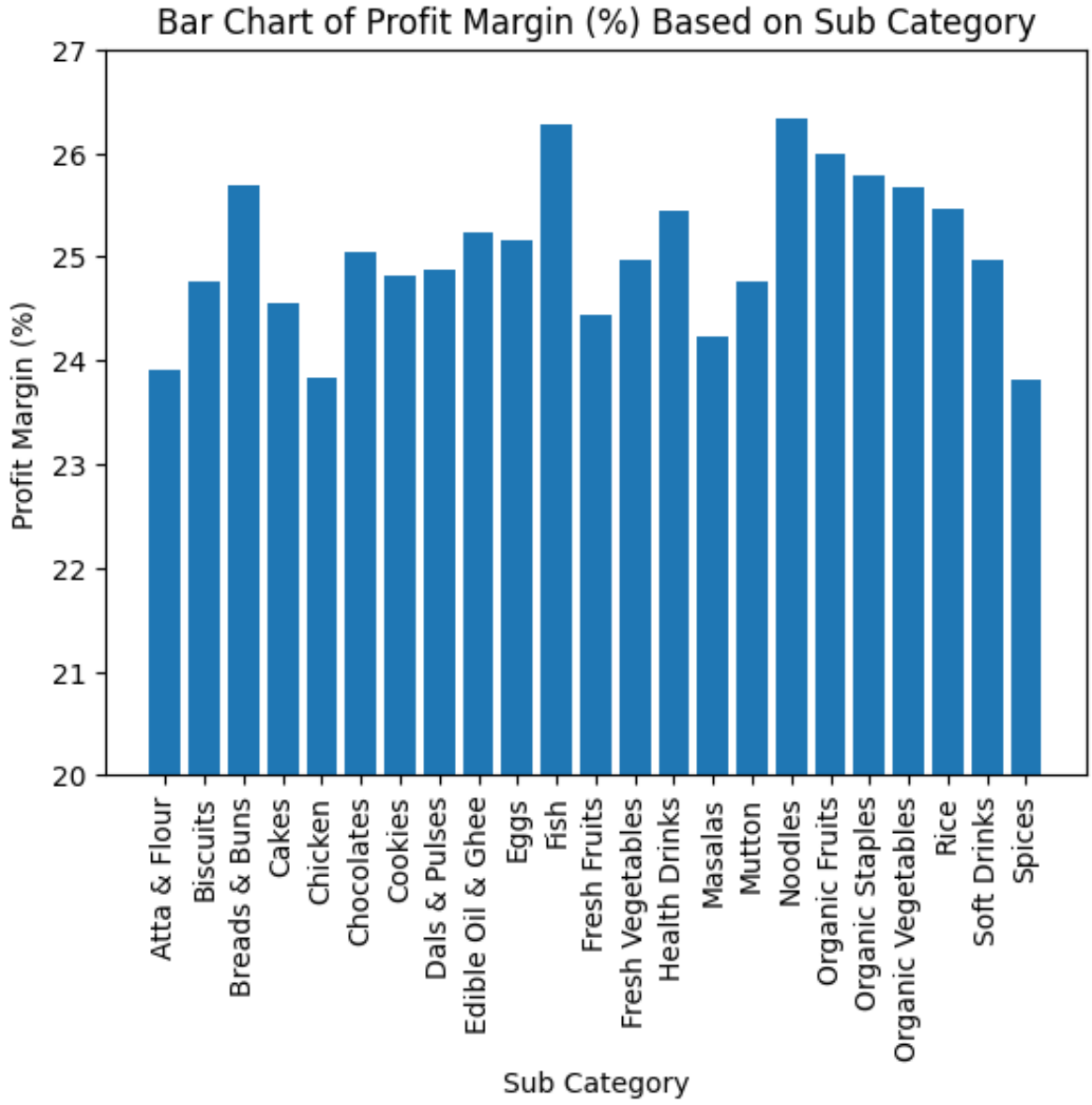
Appendix 5



Appendix 6



Appendix 7


Bar Chart of Profit Margin (%) Based on Sub Category
Profit Margin (%)
27
26
25
24
23
22
21
20
Atta & Flour
Biscuits
Breads & Buns
Cakes
Chicken
Chocolates
Cookies
Dals & Pulses
Edible Oil & Ghee
Eggs
Fish
Fresh Fruits
Fresh Vegetables
Health Drinks
Masalas
Mutton
Noodles
Organic Fruits
Organic Staples
Organic Vegetables
Rice
Soft Drinks
Spices
Sub Category