Project 1 - Part 1

PART 1 – INTRODUCTION TO DATA SCIENCES AND ALGORITHMS

1. **What do you understand about data science?**

From my understanding, Data Science is an umbrella interdisciplinary field of knowledge that spans the branches of Computer Science (Machine Learning, Deep Learning, Data Management, Visualization), Mathematics and Statistics (Linear Algebra, Time Series Analysis), and Domain Knowledge (Business, Finance, Medicine, Economics). It is the tool or knowledge that utilizes these three components and creates meaningful insights to help steer the intended domain in a better direction using Information and Technology. However, there is a focus on utilizing large datasets, which are growing in complexity due to the nature of exponentially accessible data (Peng & Parker, 2022)

Therefore, Data Scientists who introduce Data Science into specific domains often work as data architects, data engineers, or data analysts. They represent the shift that emphasizes practical applications and data transformation into actionable knowledge that Data Science brings into various domains (Gibert et al., 2018). Examples of use cases are utilizing deep learning to help diagnose breast cancer based on available data of ultrasound images, helping identify potential customers for companies using targeted ads, and streaming services using targeted media for better viewership.

References

1. Gibert, K., Izquierdo, J., Sànchez–Marrè, M., Hamilton, S., Rodríguez-Roda, I., & Holmes, G. (2018). Which method to use? an assessment of data mining methods in environmental data science. Environmental Modelling & Software, 110, 3-27. https://doi.org/10.1016/j.envsoft.2018.09.021
2. Peng, R. and Parker, H. (2022). Perspective on data science. Annual Review of Statistics and Its Application, 9(1), 1-20. https://doi.org/10.1146/annurev-statistics-040220-013917

2. **Assuming you are a department manager, and would like to investigate the customer's preference on three types of your company's products. Hence, give a bit introduction about your company and what types of products you want to investigate. Next, explain the active roles of data scientist for this task.**

As a department manager in Company A in charge of selling properties from 3 housing areas: A, B, and C, I want to increase my sales to achieve my sales KPI. Thus, I want to employ data scientists to help me analyze my company's millions of available sales data to understand better and get a targeted response to improve my sales.

Data scientists play various roles in transforming questions into actionable insights. The first role begins at the top with a data architect. These professionals ensure the quality of the millions of data available and utilize tools to enhance data availability more efficiently, such as using SQL servers or MongoDB. Being able to retrieve and analyze data without requiring excessive computing power is crucial for effective data processing. Quality insights are derived from quality data; therefore, if the initial steps are not optimized, the resulting outputs could be inaccurate or useless.

Data scientists also play the role of analysts. Once quality data can be assured, the next step is to perform deep analysis. Data Scientists use statistical methods, often much more sophisticated than superficial descriptive statistics, such as predicting potential customers for targeted ads based on previous customers using machine learning and deep learning. Thus, a deeper understanding of the customer behaviour pattern will ultimately reflect better predictions and targeted ad responses.

Lastly, at the end of the step, the insights are communicated in bite-size information relevant to stakeholders. Data Scientists often act as the bridge between understanding data insights and actions. Using visualization tools such as PowerBI and Tableau, interactive charts and dashboards can be tailored to

relevant parties such as the CEO, CFO, and CRO. Often, business decisions like this would include monetary involvement. Thus, proper convincing is required, which can be aided by the tools used by Data Scientists.

3. Based on notes week two of "Basic of Algorithms", find/create one problem. You may refer to example of "Direction of numbered NYC streets algorithms" or "Class average algorithms".
    a. *State the problem, input, processing and output*
    b. *From the problem specified in part a) above, create three popular program design tool (flowcharts, pseudocode and hierarchy charts). For flowchart, please include as many symbols as possible.*

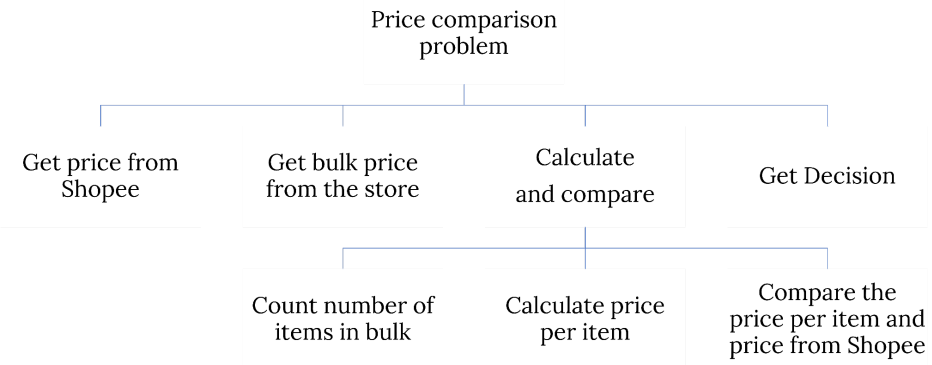
    Note: your solution must be unique.

Problem	:	Is buying in bulk cheaper from the store than buying from Shopee?
Input	:	Price in bulk from the store
Processing	:	Find the price of bulk item; count the total number of items in the bulk; calculate the price per item in the bulk
Output	:	Decision to buy

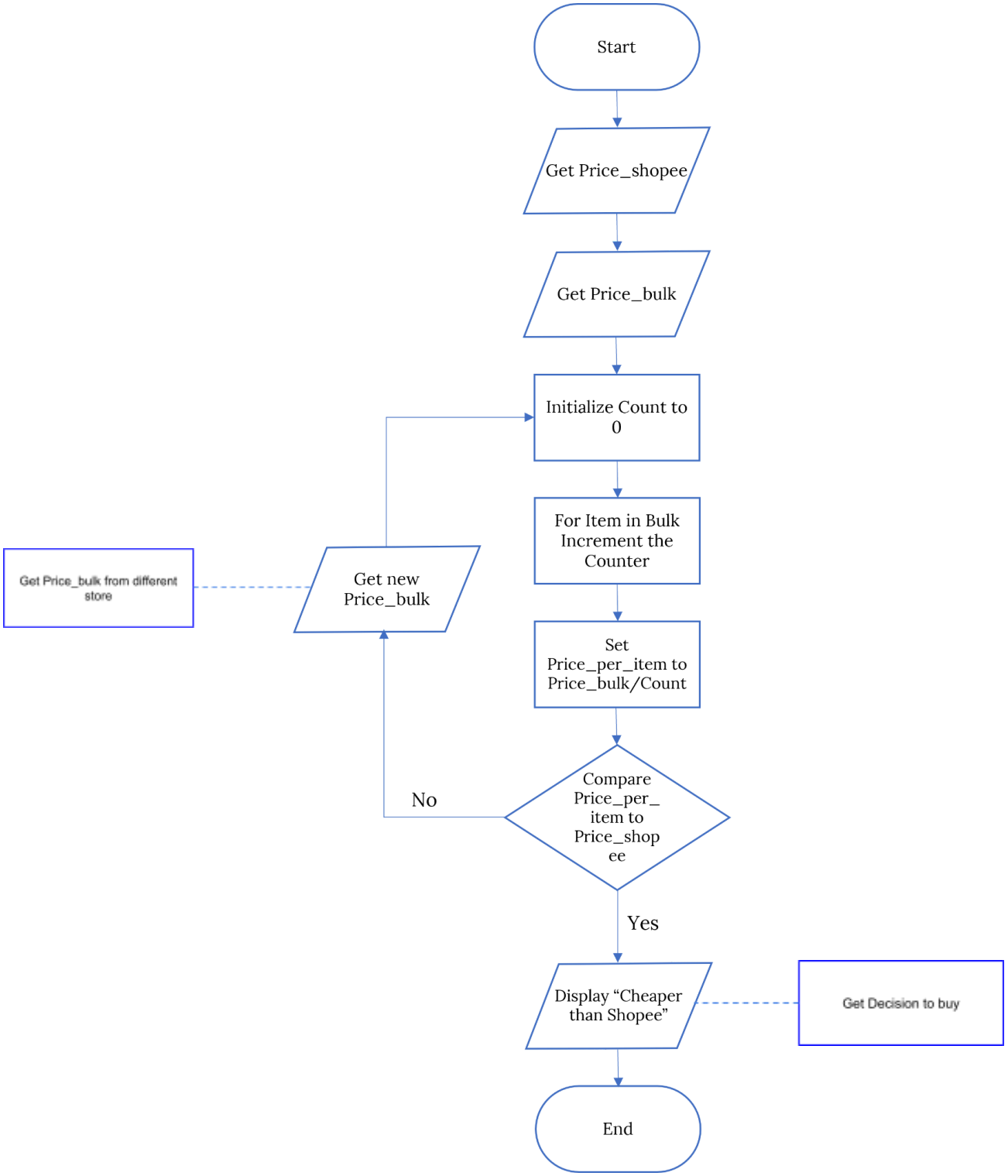
Pseudocode

Calculate the price per item in bulk.
Get Price_shopee from the price of the item from Shopee
Get Price_bulk from the price of bulk items from the shop
Initialize Count to 0
For Item in Bulk
    Increment the Counter
Set Price_per_item to Price_bulk/Count
Compare Price_per_item to Price_shopee
Display "Cheaper than Shopee"

Hierarchy Chart



Flowcharts


Start
Get Price_shopee
Get Price_bulk
Initialize Count to 0
For Item in Bulk Increment the Counter
Set Price_per_item to Price_bulk/Count
Compare Price_per_ item to Price_shop ee
No
Get new Price_bulk
Get Price_bulk from different store
Yes
Display "Cheaper than Shopee"
Get Decision to buy
End

4. **Give two examples of current data technology and its explanation.**

Two prominent examples of current data technology are data management and data utilization. Data technology has evolved significantly since the early 1900s and even the early 2000s. The growth of hardware has been paralleled by rapid software development to support the expanding technological landscape.

The first example of current data technology is storage. In the early days of digitalization, technology faced significant limitations primarily due to inadequate storage capacity. Humanity has progressed from using a mere 8KB diskettes to advanced data storage solutions that can hold hundreds of terabytes. This expansion has introduced concepts like "*data warehouses*" and "*data lakes*," which emphasize the vast amounts of data we can now manage. By utilizing various data types and database management techniques, we can store not only string data but also multiple file types, forming structured, semi-structured, and unstructured data. The increased storage capacity allows for a wider variety of data to be maintained and analyzed. This advancement is not just confined to physical storage; the growth of the network industry has led to the rise of cloud storage, which enhances connectivity among users and protects local data from corruption and loss. This development has further fueled the expansion of related sectors within the data industry.

As storage capabilities grow, new database management systems have emerged. Efficiently querying and parsing vast amounts of data has become the next challenge in data technology. Today, numerous database management systems cater to various data types, such as Apache HBASE for wide-column storage systems, Neo4j and OrientDB for graph-based systems, and MongoDB for document-based systems. These systems often extend the syntax of existing languages such as SQL, NoSQL, Python, C++, and Scalar. They are designed to be versatile and user-friendly, offering greater flexibility and easier navigation of large databases compared to older systems, which often require higher expertise and are less accessible.

Apache Spark has also experienced significant growth as a key program in this era. The ability to access cloud storage and perform parallel querying—which was once a purely academic concept—has now become essential across multiple industries. Parallel querying refers to breaking down large data sets into smaller chunks and processing them simultaneously across

multiple devices. This approach alleviates the workload on individual machines and speeds up querying.

The exponential growth of storage capacity and querying capabilities has empowered end-users to manipulate data more easily, leading to new insights and innovations. This progress has paved the way for the era of artificial intelligence (AI), fostering the development of advanced models like ChatGPT, Claude, and Google Gemini. These systems, powered by complex neural networks, have transformed various industries by addressing sophisticated challenges in natural language processing, computer vision, and more. However, a significant portion of the world's data remains untapped or inaccessible, underscoring the need for improved data governance and ethical frameworks. As AI continues to develop, it will be crucial to address challenges like data bias, privacy concerns, and responsible use to fully unlock its potential and ensure that it enhances human capabilities in a responsible manner.