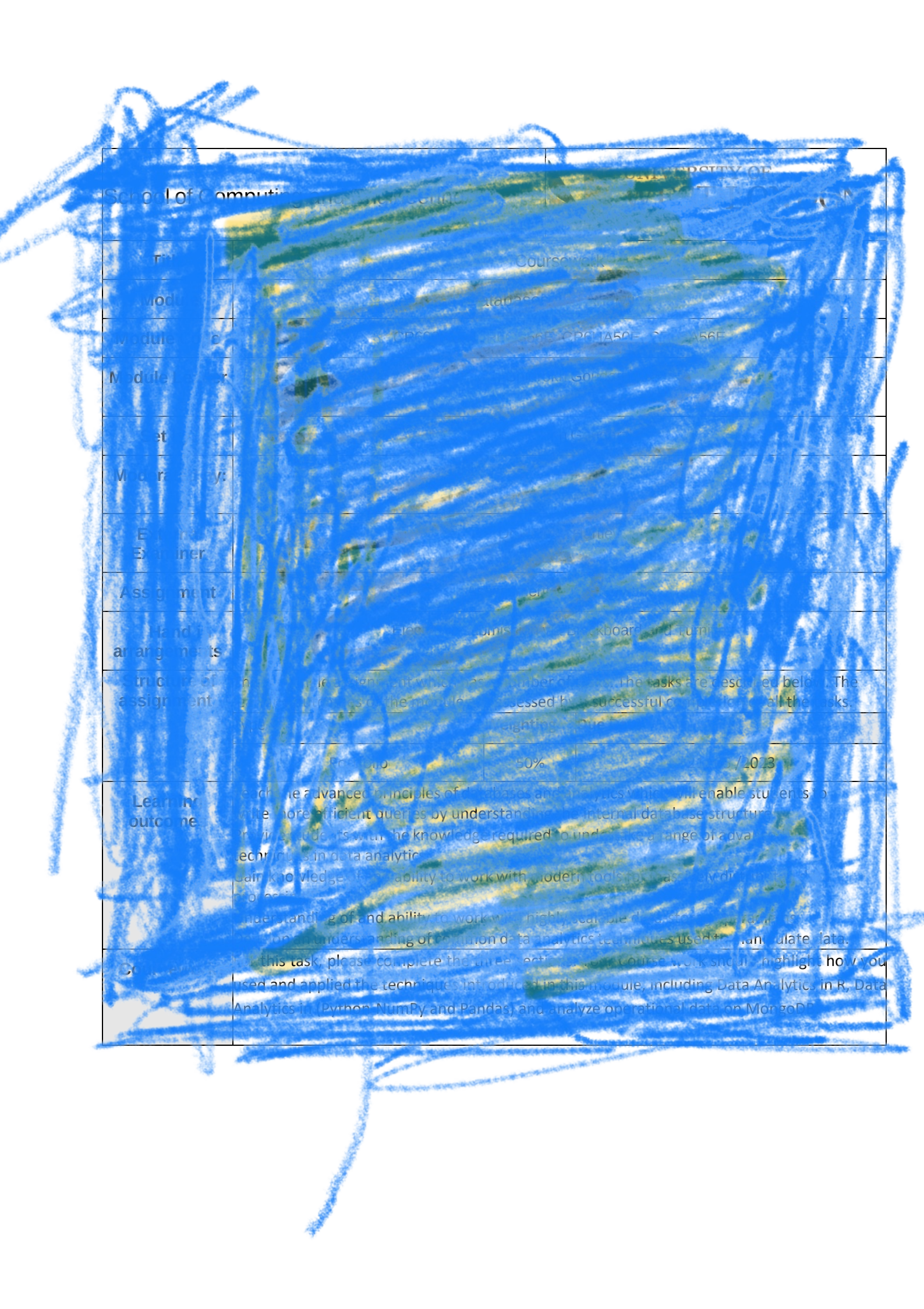

Module Assessment No.: Assessment brief:

Assessment Marking Criteria:

Criteria	Issues	Mark	Marking breakdown
Section 1 - Data analysis visualization in R- Analyzing the Energy Efficiency of Buildings (15%) **(Learning Outcome - SQL functions for data manipulation and aggregation/ Utilize R packages for analytics and data Analyzing the Energy Efficiency of Buildings.)**	As a data analyst, Importing the dataset to GitHub and Google Colab. (1 Mark) Performing SQL operations such as select, insert, update, and delete. (2 Marks) Using mathematical and aggregate functions for data manipulation. (3 Marks) Executing SQL queries in R. (3 Marks) Data manipulation and transformation using built-in functions and packages in R. (3 Marks) Creating informative and visually appealing plots using R. (3 Marks)	15	A dataset "Energy_dataset.csv" is provided for data analytics. There are some questions that should be executed on provided dataset in Google Colab. **15 marks**
Section 2 - Telecom Case Study (15%) **(Learning Outcome – Pandas and NumPy Exercise for Data Analysis) Telecom Case Study**	Importing the "telecom_dataset.csv" data and combining all data files into one consolidated dataframe. (2 Marks) Analyzing the telecom dataset using NumPy and Pandas libraries. (10 Marks) Creating plots for each pair of numerical features in the input dataframe. (3 Marks)	15	A dataset "telecom_dataset.csv" is provided for data analytics. There are some questions that should be executed on provided dataset in Google Colab. **15 marks**
Section 3 - (Learning Outcome – Analyse operational data on MongoDB) (20%)	Implementing data insertion on MongoDB Atlas. (2 Marks) Designing NoSQL schema for the company, including collections such as customers, pastOrders, products, ratings, suppliers, dailyInventoryRecord, partners, and partnerHistory. (5 Marks) Performing CRUD operations (find/insert/delete/retrieve/update) on MongoDB database. (5 Marks) Using Apache Spark in data analytics. (5 Marks) Creating plots for each pair of numerical features in the input dataframe. (3 Marks)	20	There are some questions that should be executed on MongoDB compass/atlas. **20 marks**
Total		50	

Online submission on Blackboard. You will find a link to the Assignment from the **Assessments area** of the Blackboard course menu.

Section 1 - Analyzing the Energy Efficiency of Buildings (15%) (Learning Outcome - Executing SQL Queries in R- Data Analytics in R)

Energy Efficiency of Buildings

Description:

Analyzing the Energy Efficiency of Buildings

The dataset includes features such as surface/wall/roof area, glass area, glass area distribution and orientation of 768 simulated building shapes. Based on this data, it is aimed to estimate the heating and cooling load of the building. In this age where environmental awareness is very important, it can be a super consultancy initiative to reduce our carbon footprint, as we still haven't switched to renewable energy sources.

Data Set Information:

We perform energy analysis using 12 different building shapes. The buildings differ with respect to the glazing area, the glazing area distribution, and the orientation, amongst other parameters. We simulate various settings as functions of the afore-mentioned characteristics to obtain 768 building shapes. The dataset comprises 768 samples and 8 features, aiming to analyse the energy consumption patterns in buildings and to predict two real valued responses.

Attribute Information:
The dataset contains eight attributes (or features, denoted by X1…X8) and two responses (or outcomes, denoted by Y1 and Y2). The aim is to use the eight features to analyze the dataset and predict each of the two responses.

X1: Relative Compactness, X2: Surface Area, X3: Wall Area, X4: Roof Area, X5: Overall Height, X6: Orientation, X7: Glazing Area, X8: Glazing Area Distribution, Y1: Heating Load, Y2: Cooling Load. A dataset "Energy_dataset.csv" is provided for data analytics.

Further relevant assumptions can be made about the aspects not specified above if required.

Timeline:

Scheduled to be completed by the Week 15: (21/ 05 /2023).

Deliverables Section 1:

Complete the following steps and takes a screenshot of the output for each table and query in Google Colab.

Importing the Dataset: (2 Mark)
How to import the dataset to GitHub?
Can you explain how to import a dataset to GitHub?

How to import the dataset in Google Colab? (3 Mark)
Can you demonstrate how to select, insert, update, and delete records from tables using SQL statements?
How to use mathematical expressions/ aggregate functions/ arithmetic functions to query and manipulate data?

Executing SQL Queries in R: (4 Marks)
How do you execute SQL queries in R?
Can you provide three examples of SQL queries executed within R?

Data Manipulation and Transformation in R: (3 Marks)
How to create informative and visually appealing plots using the R package for data visualization
Explain how to perform data manipulation and transformation tasks using built-in functions and packages in R.
Can you demonstrate three examples of data manipulation tasks using R functions and packages?

Data Visualization in R: (3 Marks)
How do you create informative and visually appealing plots using the R package for data visualization?
Provide three examples of different types of plots that can be created using R and explain their significance in data analysis.

Section 2 - Telecom Case Study (15%) (Learning Outcome – Pandas and NumPy Exercise for Data Analysis)

Telecom Case Study

Description:

This case study is about a telecom company. Data analytics can help this company to make smarter decisions that lead to higher productivity and more efficient operations. One of the key parameters that should be checked for this company is churn rate. Churn, often known as "churn rate," is the measurement of the number of subscription customers who have cancelled their subscription over a set period of time.

Generally, churn rate, sometimes known as attrition rate in business, is the rate at which customers stop doing business with a company over a given period of time. Churn rate is the percentage of subscribers to a service that discontinue their subscription to that service in each period. In order for a company to expand its clientele, its growth rate (i.e. its number of new customers) must exceed its churn rate.

The first way to reduce the churn rate and consequently improve business retention is to make the most effective use of a telco's internal data. All this data allows telecom companies to know customers better–to the point where they can collect them into precise and consistent segments. In this case study, with some variables we need to show whether a particular customer will switch to another telecom provider or not. In telecom terminology, this is referred to as churning and not churning, respectively. Some improvements will make possible in this company through your data analytics. A dataset "telecom_dataset.csv" is provided for data analytics. You need to complete each step in the following order.

Deliverables Section 2:

Complete the following steps and take a screenshot of the output for each step.

Importing and Combining Data: (2 Marks)
How to import the "telecom_dataset.csv" data and combining all data files into one consolidated dataframe?
Can you outline the steps to import a CSV file into a Pandas DataFrame?
How would you ensure data integrity when combining multiple data files into a single DataFrame?

Could you demonstrate how to verify that the combined DataFrame contains all the data from the individual files without any loss or duplication?

Analyzing Data with NumPy and Pandas: (10 Marks)

How do you analyze“telecom_dataset.csv” data in NumPy and Pandas libraries?

Explain how you would handle missing data during data analysis using Pandas.

Can you calculate and provide examples of commonly used statistics in data analysis, such as mean, median, standard deviation, minimum, maximum, and quantiles, using Pandas and NumPy?

How would you interpret these statistics in the context of your analysis?

Creating Plots for Numerical Features: (3 Marks)

How do you create some plots for each pair of numerical features in the input datafram?

Provide examples of data visualization techniques commonly used in exploratory data analysis and demonstrate how to create them using libraries like Matplotlib or Seaborn.

How would you decide which type of plot is most appropriate for visualizing the relationship between two numerical features?

Could you demonstrate how to create pairwise relationship plots for numerical features in a dataset, such as scatter plots, pair plots, and correlation matrices, using Python libraries?

Section 3 - (Learning Outcome – Analyze operational data on MongoDB) (20%)

Description:

Data Modelling

The process of designing and implementing the database are guided based on the 4 steps:

(i) Gather requirements.
(ii) Understand relationships between entities.
(iii) Identify the data structure.
(iv) Apply design patterns.

After reading the instructions, the following tasks in the database need to be designed. The list of tasks and the assumptions are explained in Table 1. The data structure and entity relationships (ii and iii) should be addressed and built based on the most important functionalities identified for the operation of the business.

Tasks:

Complete the following takes a screenshot of the output for each task in MongoDB (Atlas).

Task	Subtask	Sub Subtask
1. Designing NoSQL Schema for Company	1.1 Designing collections in Schema	customers
		pastOrders
		products
		ratings
		suppliers
		dailyInventoryRecord
		partners
		partnerHistory
2. Implementation on MongoDB	2.1 Data Insertion	customers
		pastOrders
		products
		ratings
		suppliers
		dailyInventoryRecord
		partners
		partnerHistory
	2.1 Querying	Query (1-20)

Based on the instructions, the focus of the business is to deliver all products as soon as possible.
The collections (in JSON format) are presented here. Also, all collections files are uploaded in Blackboard. These collections should be used and imported in the designed database in MongoDB (Atlas/Compass).

1. **Customers:**
    - **_id:** unique identifier
    - **Name:** customer name
    - **Gender:** gender of the customer
    - **Age:** age of the customer
    - **Phone number:** contact number of the customer
    - **Addresses**
        - **_id:** unique identifier
        - **House:** house name/number
        - **Street:** street name/number
        - **City:** city
        - **Post code:** postcode of the address
            - **Location**
                - **Coordinates:** geospatial location based on longitude and latitude
            - Current orders
                - **_id:** unique identifier
                - **Date:** date when order was placed
                - **Order status:** Can be one of 6, 1. Added to Cart, 2. Paid, 3. Ongoing 4. Dispatched, 5. Delivered, 6. Delayed
                    - **Order details**
                        - **Total cost:** total cost of the order
                        - **partner _id:** unique identifier of the partner making the delivery
                        - **Shipping_id:** unique identifier of the shipping address
                        - **Supplier_id:** unique identifier of the warehouse supplying the product
                    - **Recommended products**
                        - **Product_id:** unique identifier of the product being recommended
                        - **Avg_rating:** average rating of the recommended product

2. **pastOrders:**
    - **_id:** unique identifier
    - **Order_date:** date order was placed
    - **Customer_id:**
    - **Order_details**
        - **Product_id:** unique product id
        - **Quantity:** quantity ordered
        - **Cost:** unit cost of the particular product (£'s)
    - **Total_cost:** total cost, unit cost x quantity
    - **Partner_id:** unique identifier of the partner (delivery drivers)
    - **Shipping_id:** the address ID to which the order shipped to
    - **Supplier_id:** the supplier unique id
    - **Order_status:** Can be one of 6, 1. Added to Cart, 2. Paid, 3. Ongoing 4. Dispatched, 5. Delivered, 6. Delayed.

3. **products:**
    - **_id:** unique identifier
    - **Product_name:** name of the product
    - **Short_desc:** short description of the product
    - **Dimensions**
        - **Length:** in cm
        - **Width:** in cm
        - **Height:** in cm
    - **Quantity_per_unit:** kilograms or litres (based on the product)
    - **Avg_ratings:** the average rating of the product (based off all customer ratings)
    - **Std_price:** standard price of the product
    - **Supp_price:** price that the suppliers charges
    - **Category:** product type
    - **Fresh**
        - **Category:** either one of bakery, drinks or fruits & vegetables
        - **Best_before:** number of days until product expires
        - **Country_of_origin:** what country product is from
    - **Books**
        - **Author_name:** name of the author
        - **Publisher:** publisher name
        - **Year_publication:** year it was published
        - **ISBN:** 13-digit code which identifies a specific edition of a book title from a publisher
    - **CDs**
        - **Artist_name:** full name of the artist
        - **No_of_tracks:** number of tracks
        - **Total_play_time:** total play time in minutes
        - **Publisher:** publisher of cd

- **Phones**
    - **Brand:** brand of the phone
    - **Model:** model of the phone
    - **Colour:** phone colour
    - **Features:** list of features comma separated
- **Home_appliances**
    - **Colour:** home appliance colour
    - **Voltage:** in volts
    - **Style:** style of home appliance

4. **ratings:**
    - **_id**: unique identifier
    - **PublishDate:** date the rating is published
    - **Order_id:** unique order id
    - **Customer_id:** unique customer id
    - **Order_date:** date the order was placed
    - **Product_id;** unique product id
    - **Rating:** rating given to the product out of 5

5. **Suppliers:**
    - **_id**: unique identifier
    - **Name:** Supplier's name
    - **Address:** supplier's address
    - **City:** supplier's city
    - **Post_code:** supplier's postcode
    - **Location**
        - **Coordinates:** geospatial location based on longitude and latitude
    - **Realtime_inventory**
        - **Product_id:** unique product identifier
        - **Timestamp:** time when the product was supplied
        - **Quantity:** number of products

6. **dailyInventoryRecord**: inventory records for a product in a supplier in a day for past 10 years
    - **_id**: unique identifier
        - **Supplier_id**: supplier identifier
        - **Product_id**: product identifier
        - **Start_date**: the starting datetime of this document (00:00 of the day)
        - **End_date**: the ending datetime of this document (23:59 of the day)
    - **Supplier_location**
        - **Coordinates**: geospatial location based on longitude and latitude
    - **Inventory_data**: list of inventory record within the day
        - **Datetime**: datetime when the inventory quantity was recorded
        - **Inventory_quantity**: inventory quantity at the datetime

7. **partners:**
    - **_id:** Unique Identifier
    - **name:** partner's name
    - **Age:** partner's age
    - **Gender:** gender of the partner
    - **Phone:** phone number of the partner
    - **Email:** partner email
        - **Bank_account:**
            - **Account_name:** partner's name on bank account
            - **Account_number:** partner's bank account number
            - **Sort_code:** 6 digit number which identifies the bank
        - **Availability:**
            - **Is_active:** binary field. 1 indicates partner is active, 0 indicates not active
            - **On_delivery:** binary field. 1 indicates partner is on a delivery, 0 indicatesnot on a delivery
            - **Location:**
                - **type**
                - **coordinates**
    - **Deliveries_made:** total number of historical deliveries
    - **Avg_per_week:** average weekly deliveries historical
    - **Best_week:** best historical weekly average
    - **Number_of_week:** current number of deliveries in the week

8. **partnerHistory:**
    - **_id:** ObjectId
    - **Partner_id:** Unique partner ID
    - **Start_date:** starting datetime for the partners shift period
    - **End_date:** ending datetime for the partners shift period
    - **Orders**
        - **Order_id:** unique order identifier
        - **Timestamp**
        - **Order_details**
            - **Product_id:** unique product identifier
            - **Quantity:** quantity of a product

Deliverables Section 3:

Write the following queries and takes a screenshot of the output for each query.

Implementation on MongoDB Atlas (Data Insertion): (2 Marks)

How to do implementation on MongoDB atlas (Data Insertion)?

Provide two examples of inserting data into a MongoDB Atlas database.

Designing NoSQL Schema for Company: (5 Marks)

How to design a NoSQL schema for a company?

Provide examples of a NoSQL schema designed for a company.

Data Manipulation in MongoDB: (10 Marks)

How to find, insert, delete, retrieve, and update data from a MongoDB database?

Provide five examples of finding, inserting, deleting, retrieving, and updating data in MongoDB.

Creating Plots for Numerical Features: (3 Marks)

How do you create plots for each pair of numerical features in a dataframe?

Provide examples of creating three different plots for pairwise numerical features.