Data Engineer Preliminary Test

Please complete this test in your GitHub and make sure the repository is set to public or if you want it private, you can provide us the access to your GitHub repository by inviting us (username: **kulina-data**) to your repository (here's the reference to invite other users to your repo). The set of requirements for this preliminary test can be downloaded from here. Please note that your submission should follow our submission directory structure illustrated in the next two sections.

Requirement Directory Structure

The directory structure of requirement folder can be described as follows:

- Database and SQL
  - Insertion.sql
  - ERD.png
- Statistics
  - dataset.csv
- Data Visualization
  - dataset.csv
- Machine Learning
  - train.csv
  - validation.csv
  - test.csv

Submission Directory Structure

You submission on your GitHub repository should be in following structure:

- Programming
  - tools.txt
  - rotate_box.[file format supported by the tool you use]
- Database and SQL
  - tools.txt
  - create_db.[file format supported by the tool you use]
  - delivery_history.[file format supported by the tool you use]
- Statistics
  - tools.txt
  - prob_dist.[file format supported by the tool you use]
  - business_recommendation.txt
- Data Visualization
  - tools.txt
  - cohort_retention.png
  - cohort_retention.[file format supported by the tool you use]
  - business_recommendation.txt
- Machine Learning

- tools.txt
- recommender.[file format supported by the tool you use]
- prediction.csv
- interpretation.txt

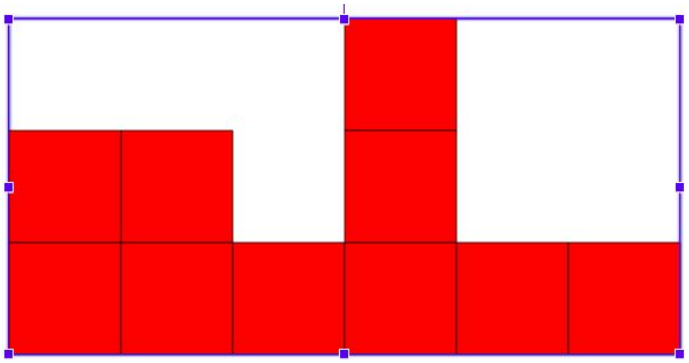
Programming (Max Point: 15)

Instruction

Complete this programming task using Python, R, Scala, Julia or any programming language you want. Libraries and frameworks are allowed to complete this task. Please provide programming languages, libraries and frameworks you use to complete the task, including its version.

Question Statement

Kulina has a box that contains food packages, the size of the box is $l \times 2 \times h$ where l and h are the length and the height of the box. The size of each food package is $2 \times 2 \times 2$. The box is represented as an array A and the food packages are represented as the value of each element of the array A. For example the food packages are arranged in a box with the following arrangement.



It can be represented as an array $A = [2, 2, 1, 3, 1, 1]$. Without loss of generality it can be assumed that the height of the box is equal to the maximum value of element in the array and the length of the box is equal to the length of the array. During delivery, the box is rotated 90 degrees clockwise k times so that the arrangement of food packages inside the box is changing due to gravitational force that pulls the food packages down. Please write a program that prints out the new arrangement of the array A after k times 90 degrees clockwise rotation.

Input Format

N : an integer represent the length of the array A, $0 \leq N < 2^{32}$
A_i : integers represent the element of array A, $0 \leq A_i < 2^{32}$
k : an integer representing the number of 90 degrees clockwise rotation, $0 \leq k < 2^{32}$

Output Format

Print array A'

Input Sample

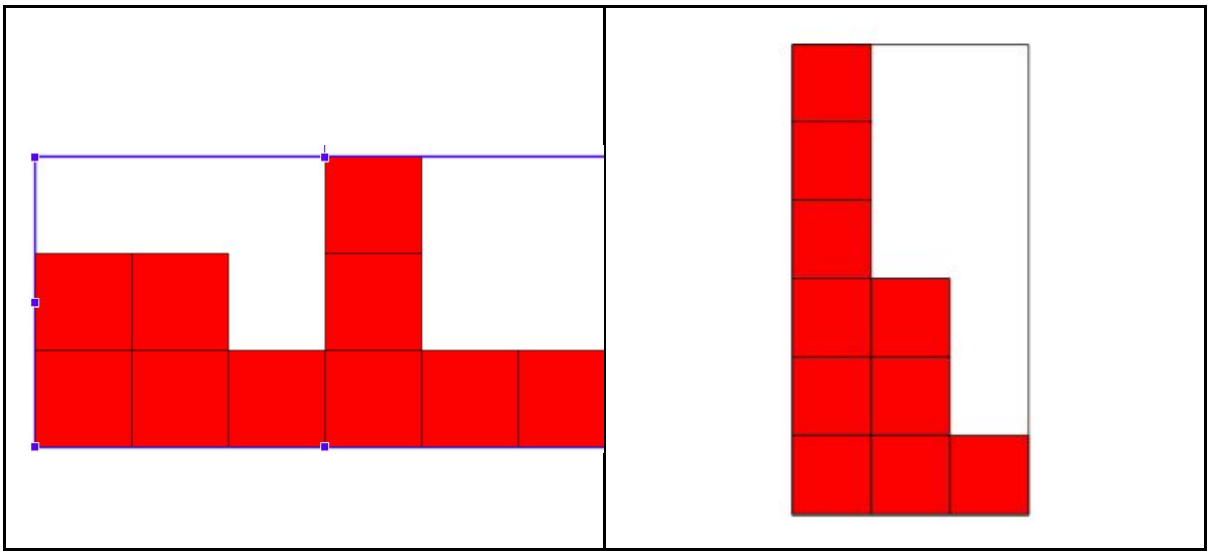
6
2 2 1 3 1 1
1

Output Sample

6 3 1

Explanation

If we rotate the box 90 degrees clockwise, the box will transform into following arrangement.



The arrangement on the right can be represented as an array [6,3,1]

Scoring

The scoring of this programming task is considered by following factors (ranked):

1) Correctness (Max: 7 Points)
2) Efficiency (Max: 5 Points)
3) Code Quality (Max: 2 Points)
4) Logic Flow (Max: 1 Point)

Database and SQL (Max Point: 30)

Instruction

Complete this database and SQL task using MySQL, PostgreSQL, SQLite or any relational database management system. Please provide RDBMS you use to complete the task including its version.

Database Design (Max Point: 10)

Take a look at the ERD Diagram provided in the Database and SQL folder.

Question Statement

Please write an SQL query to create a database that contains those tables with similar field names and relation. You are expected to determine the data type of each field which has a best suit with the field name. Save your query in database_design.sql.

Clue: You may also take a look at database contents file which include insertion query and look at the content of each field.

Scoring

The scoring of this database design task is considered by following factors (ranked):

1) Correctness of Data Relationship (Max: 5 Points)
2) Data type choices (Max: 3 Points)
3) Query Quality (Max: 2 Points)

SQL (Max Point: 20)

Please run the insertion query provided in insertion.sql file. Please take note that this insertion query has only tested on MySQL with the correct database design, if you have any problem on running this query, you can edit the insertion query depending on your needs, but it's okay if you choose to not running this query as this SQL test only test your Select query. The only problem you'll get is that you can't check whether your query runs smoothly or not which will affect the scoring.

Question Statement

Please write an SQL query to show the details of successful delivery history for each user in September 2019. The details include user id, username, user email, user phone number, delivery_date, delivered product name, product category (separated by commas), quantity and delivery address sorted by delivery date and followed again by user id. Based on that sorting value, you need to provide a 'total' column containing progressive sum of the quantity.

Note: if there is an error on the data, please contact us to provide revision, while you wait you can just ignore the error by creating the complete query without test it or you can just do the other question.

Scoring

The scoring of this SQL task is considered by following factors (ranked):

1) Correctness (Max: 10 Points)
2) Efficiency (Max: 7 Points)
3) Query Quality (Max: 3 Points)

Statistics (Max Point: 10)

Instruction

Complete this statistics task using Python, R, Scala, Julia or any programming language you want. Libraries and frameworks are allowed to complete this task. Microsoft Excel, Google Sheet, SPSS or any statistical software are allowed to complete this task. Please provide programming language, libraries, frameworks and software you use to complete the task including its version.

Question Statement

Using the available dataset, Please determine the probability distribution of every user total quantity they have purchased including your reason on picking that probability distribution. Please give some business recommendations to Kulina based on your findings.

Scoring

1) Correctness (Max: 6 Points)
2) Business Recommendation Quality (Max: 4 Point)

Data Visualization (Point: 15)

Instruction

Complete this data visualization task using D3.js, Plotly, Matplotlib or any data visualization tools. Tableau, Google Data Studio, Microstrategy or any GUI/WYSIWYG data visualization tools are also allowed to complete this task. Please provide data visualization tools you use to complete this task including its version.

Question Statement

Using the provided dataset, please create **cohort retention rate graph based on customer acquisition month with monthly period**. Acquisition month is defined as the first order month of the user. Please determine the metric you use, give some analysis and business recommendations to Kulina based on the visualization.

Please note that you are always able to create your own definition for the visualization as long as it's relevant. You can also search on any sources for the definition of the visualization but please don't forget to define it first and put the citation on your solution in order to make the same perspective on the visualization.

Scoring

1) Understandability of the Visualization (Max: 5 Points)
2) Definition Clarity (Max: 4 Points)
3) Metric Chosen (Max: 3 Points)
4) Analysis and Business Recommendation Quality (Max: 3 Points)

Machine Learning (Point: 30)

Instruction

Complete this machine learning task using Python, R, Scala, Julia or any programming language you want. Libraries and frameworks are allowed to complete this task. WEKA, KNIME, RapidMiner or any GUI Machine Learning tools are also allowed to complete this task. Please provide programming languages, libraries, frameworks and tools you use to complete the task, including its version.

Question Statement

Using the provided datasets, please create a machine learning model that can predict the value of rating based on the review text. You can also use meal_id as an additional attribute. Please provide an interpretation of your model and an analysis on the data.

Scoring

1) Accuracy from test data (Max: 15 Points)
2) Interpretation and Analysis (Max: 10 Points)
3) Efficiency (Max: 5 Points)