General instructions:

You're provided with 2 problems with corresponding datasets. Some tasks have been provided in the problem sets. Your aim is to accomplish those tasks and perform any other data analysis that you think could provide useful insights into the datasets.

Google colab should be used for the tasks. You're free to use any open source python libraries in the process.

Your scores would be based on:

- Correctness of outcome for the given tasks
- Efficiency of the code performing the tasks
- Clarity of plots/ conclusions drawn from the datasets
- Attempt both the problems

Total duration of the test: 3 hours

Problem 1.

Given a dataset of DNA strings (each DNA sequence is a 40 character string made of characters A,C,T and G).
a) Perform analysis of co-occurrence of kmers (A kmer is a substring of length k). Show distribution of co occurrence of kmers of upto length 3 in form of heatmap. Kmers are said to co-occur when one kmer is immediately followed by another.
(Eg: In ACTTGA, for k=2, [AC,TT], [CT, TG], [TT, GA] are co occurring)
b) Create a heatmap of the occurrences of kmers (for k=2) at each position. The color should indicate the percentage of times a particular kmer occurs at a given position.
c) Any other analysis of your choice of the kmers in the given dataset

Problem 2.

Given a csv containing multiple user's activity information in the form of a time series. Each row contains information about a single activity session performed by a user. Fields in the data indicate:

1. User id: Unique id per user
2. Activity: One of walking/ running/ meditation/ aerobics/ pilates
3. Start time of the activity
4. End time of the activity

One day should be demarcated between 04:00:00 am on a given date to 3:59:59 am the next day.

a) Given user id and a particular date, generate a summary of the day’s activity
b) Given user id and month/ year, generate average monthly stats for a user for a given month (most frequently performed activity and the activity performed for the longest duration)
c) From the user data, find the users that should be given notifications to engage them more on the platform (Explain your reasoning for the same)
d) Some summary statistics for the manager of the platform to observe user activity.