The product that we are developing in an Auto ML solution (automated machine learning). So, data cleaning, data manipulation is big part of the process. User can input any dataset and we have to detect what's date, what's character etc. Since no data is perfect, a very simple data cleaning code won't be able to read most of the variables and will remove those variables by the time it comes to predictive modelling. Consider this and then try to seek answers to the tasks given.

Please use google colab, and if any package you feel is missing (that maybe elsewhere), use alternate package.

In case of any questions, You can take basic assumptions if you want to, we want to see how innovative you think. We want to see your thinking abilities and how much big you can think from this given information.

We are going to look into how deep you think and how robust the code is.

Only send us a google colab link and for the theoretical part, put it in comments in the colab itself. There is a google response link. fill your name, email and google colab link there.

1) Write a function in python that inputs a dataframe and identify which columns have date in them. Using these date columns make new columns which are difference between these columns taking 2 at a time. (for instance if there is date1, date2, date3 columns, output should be like date1-date2, date2-date3, date1-date3). **For this problem only, print out data in the colab.**

Thing to consider

- Date column might have some invalid entries in them
- Date can be of different format throughout the column
- Code should be efficient and fast
- Code should be well commented and easy to interpret
- Use google Colab
- Code should be robust enough to run on any dataset
- Make a dummy dataset by yourself.

2) Write a function in python that take dataframe as input and drop columns having Pearson correlation more than 0.85

Thing to consider

- Code should drop least amount of variable as possible. (this is an important point)
- Code should be efficient and fast
- Code should be well commented and easy to interpret
- Use google Colab
- Code should be robust enough to run on any dataset
- Make a dummy dataset by yourself or pass any publicly available dataset to test out your logic

Hint: There is no restriction on copying code from the internet, but remember that most of the code found over the internet,

- Works on near perfect data which is impossible in the real world
- Is not equipped to work on every dataset which is central to our business model
- Have multiple flaws

We have specifically designed each question to see your thinking ability.