Lannet Technologies

The product that we are developing in an Aoto 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.

It's up to you to solve any number of questions within time frame, 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. https://forms.gle/et27noLnBfHHJLxu8 - fill your name, email and google colab link here.

Task 1.1 Write a function in python that identify which columns have date in them

Task 1.2 Using these date columns make new columns which are difference between these columns taking 2 at a time. (difference of days between dates). For instance - Data set contains 4 date columns which are start date, end date, DOB and Date of promotion. Then you will form 6 new columns containing difference of these date by taking 2 date at a time. One of the new column would be DOB - end date. But the data can contain any number of date columns (dataset has n number of date columns)

Task 1.3 Drop all the original columns containing the date and just keep the newly computed columns

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
- Make a dummy dataset by yourself to test out your logic
- Code should be well commented and easy to interpret
- Use google Colab
- Code should be robust enough to run on any dataset
- To test out the logic we will pass a random dataframe into your function

Task 2.1 Write a function in python that drop columns having Pearson correlation more than 0.85

Thing to consider

- Code should drop least amount of variable as possible
- Code should be efficient and fast
- Make a dummy correlation matrix by yourself to test out your logic
- Code should be well commented and easy to interpret
- Use google Colab
- Code should be robust enough to run on any dataset
- To test out the logic we will pass a random dataframe into your function

Task 3.1 Write an explanation about how operations like imputation , feature selection, normalization etc. changes across the training and the testing data. For example, do you do imputation separately for training and testing set…?

Task 4.1 How to you speed up a python code?

Task 5.1 Write a python function that extract zip code from an address column.

Thing to consider

- Code should be able to identify which columns contains address
- An address columns may not contain zipcode in some records
- These are US zipcode (5 digits)
- US zipcode can be of 9 digits but we want only the first 5 digits
    - https://smartystreets.com/articles/zip-4-code
- Code should be efficient and fast
- Make a dummy dataset by yourself to test out your logic
- Code should be well commented and easy to interpret
- Use google Colab
- Code should be robust enough to run on any dataset
- To test out the logic we will pass a random dataframe into your function

Task 6.1 Write a python function to address typos in a column

Thing to consider

- Code should identify which columns have categorical data and do operations on that columns
- Make a dummy dataset by yourself to test out your logic
- Code should be well commented and easy to interpret
- Use google Colab
- Code should be robust enough to run on any dataset

- To test out the logic we will pass a random dataframe into your function