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

**Ans:- pd.to\_datetime() funtion is used to identify if columns have date in them.**

**df.dtypes (to check the dataset types)**

**Task 1.2** Using these date columns make new columns which are difference between these columns taking 2 at a time

**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 does, feature selection, normalization etc. changes across the training and the testing data. For example, do you do imputation separately for training and testing set…?

**Ans:- Imputaion , sometime data contain missing values or NAN’s or blank space or missing values, so when we fill the missing values it impact on the rest of the dataset. Here Data Mining can be helpfull. Missing values can be ignore.**

**Feature Selection:- This function is very beneficial in terms of identifying the actual column or important column which has more use than other column it , It is also Known as Variable Selection. Best to use in model construction.**

**Normalization:- It is the technique that organizes table in manner that reduces redundancy and dependency of data.**

**Yes, normalization or imputation separately for training and testing set ,we should split before the pre-processing or imputation, it is easy to remember what we did in training set ,if test set also need pre-Processing or imputation, so that way we can do imputation in both sets.**

**Task 4.1** How to you speed up a python code?

**Ans:- Familiar with all the functions.**

**Libraries should be known for the perticular dataset so the functions run accurately.**

**Using Proper data structures and algorithms .**

**Use updated version of the python or platform where we typing code (Jupyter notebook), it can also affect on runtime for good performance.**

**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 l
  + [https://smartystreets.com/ar]ticles/zip-4-code](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

**Ans:- I didn’t work with extraction part with this type of description yet, still learning.**

**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