Removing duplicate Columns

1. Import pandas library.

Scenario 1: Number of rows<50000 and number of columns<100

1. Read dataset using pandas methods such as read_csv, read_pickle,read_excel etc.
2. Transpose the dataset using the attribute “T” of a pandas dataframe. It will transform rows as columns and columns as rows. Now column names will become row index and row indexes will become column names.
3. Drop duplicated rows using method “drop_duplicates()” where key parameters to be passed are “keep=’first’ ” and “inplace=True”.

```
transposed=data.T
```

```
transposed.drop_duplicates(keep='first',inplace=True)
```

Scenario 2: Number of rows>50000 or number of columns >100

1. Write a custom python function/method which is based on comparison of series to find the set of duplicate columns.
2. Once you find the set of duplicate columns, drop them using **drop()** method. The key parameters to be passed are “columns= list of duplicated columns” and “inplace=True”.

```
def findDuplicateColumns(df):
    duplicateColumnNames = set()
    for x in range(df.shape[1]):
        col1 = df.iloc[:, x]
        for y in range(x + 1, df.shape[1]):
            col2 = df.iloc[:, y]
            if col1.equals(col2):
                duplicateColumnNames.add(df.columns.values[y])

    return list(duplicateColumnNames)
```

```
duplicated_columns=findDuplicateColumns(da)
```

```
data.drop(columns=duplicated_columns,inplace=True)
```