Exercise 3 : Perform ETL on the air pollution CSV data

Step 1 : Read provided csv File which contains pm2.5 values

In [1]:
```
import pandas as pd
df = pd.read_csv('30201130PM25.csv',header = None)
df.columns = ['pm25']
```

In [2]:
```
df.head()
```

Out[2]:

	pm25
0	4.0
1	4.0
2	3.0
3	2.0
4	3.0

In [3]:
```
df.shape
```

Out[3]:
```
(8176, 1)
```

Step 2 : Check whether missing values are present or not

In [4]:
```
# check whether 9999(which is considered as missing value) is in the datafr

9999 in df.values
```

Out[4]:
```
True
```

Step 3 : Replace missing values in the dataframe with NaN

In [8]:
```
import numpy as np

df_Imp = df.copy()
df_Imp['pm25'] = df_Imp['pm25'].replace([9999],np.nan)
```

In [9]:
```
df_Imp.isnull().sum()
```

Out[9]:
```
pm25    101
dtype: int64
```

In [10]:
```
df_Imp
```

Out[10]:

	pm25
0	4.0
1	4.0
2	3.0
3	2.0
4	3.0
...	...
8171	22.0
8172	24.0
8173	27.0
8174	23.0
8175	24.0

8176 rows × 1 columns

we have missing values in our data

Step 4 : Perform various Imputation Techniques on the data

Basic Imputation Techniques

4.1 Replace missing values(NaN) with zero/any constant value

In [21]:
```
imputed_df1 = df_Imp.copy()
imputed_df1  = imputed_df1.fillna(0)
```

In [22]:
```
imputed_df1.isnull().sum()
```

Out[22]:
```
pm25    0
dtype: int64
```

4.2 Applying Mean Imputation technique to replace missing values

In [17]:
```
#Impute the values using scikit-learn SimpleImpute Class

from sklearn.impute import SimpleImputer

dfMeanImp = df_Imp.copy()
print(dfMeanImp.isna().sum())
imp_mean = SimpleImputer(strategy='mean') #for median imputation replace 'm
imp_mean.fit(dfMeanImp)
imputed_df2 = pd.DataFrame(imp_mean.transform(dfMeanImp))
```

```
pm25    101
dtype: int64
```

In [19]:
```
imputed_df2.isnull().sum()
```

Out[19]:
```
0    0
dtype: int64
```

4.3 Applying Most Frequent Imputation technique to replace missing values

In [20]:
```
#Impute the values using scikit-learn SimpleImpute Class

from sklearn.impute import SimpleImputer

dfMeanImp = df_Imp.copy()
print(dfMeanImp.isna().sum())
imp_mean = SimpleImputer( strategy='most_frequent')
imp_mean.fit(dfMeanImp)
imputed_df3 = pd.DataFrame(imp_mean.transform(dfMeanImp))
imputed_df3.head()
```

```
pm25    101
dtype: int64
```

Out[20]:

	0
0	4.0
1	4.0
2	3.0
3	2.0
4	3.0