# PRACTICAL NO 1

Exp1: Extract the data from database using python and demonstrate various data pre-processing techniques for a given dataset

In [4]:
```
%matplotlib inline
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import sklearn
```

Create a random dataset

In [55]:
```
np.random.seed(42)
data = {
    'A': np.random.randn(5),
    'B': [1, 2, np.nan, 4, 5],
    'C': ['foo', 'bar', 'baz', 'qux', 'quux'],
    'D': [True, False, True, False, True]
}
```

Create a DataFrame from the dictionary

In [56]:
```
df = pd.DataFrame(data)
```

Display the original dataset

In [57]:
```
print("Original Dataset:")
print(df)
```

```
Original Dataset:
          A    B     C      D
0  0.496714  1.0   foo   True
1 -0.138264  2.0   bar  False
2  0.647689  NaN   baz   True
3  1.523030  4.0   qux  False
4 -0.234153  5.0  quux   True
```

Export data to a CSV file

In [58]:
```
df.to_csv('random_dataset.csv', index=False)
```

# Simulate extracting data from a database

Read data from the CSV file

In [59]:
```
df_from_csv = pd.read_csv('random_dataset.csv')
```

Display the extracted dataset

In [60]:
```
print("\nDataset Extracted from CSV:")
print(df_from_csv)
```

```
Dataset Extracted from CSV:
          A    B     C      D
0  0.496714  1.0   foo   True
1 -0.138264  2.0   bar  False
2  0.647689  NaN   baz   True
3  1.523030  4.0   qux  False
4 -0.234153  5.0  quux   True
```

# Data Pre-processing Techniques

1. Check for missing values

In [61]:
```
print("Isnull:\n", df_from_csv.isnull())
```

```
Isnull:
        A      B      C      D
0  False  False  False  False
1  False  False  False  False
2  False   True  False  False
3  False  False  False  False
4  False  False  False  False
```

1. Check for non-missing values

In [62]:
```
print("\nNotnull:\n", df_from_csv.notnull())
```

```
Notnull:
       A      B     C     D
0  True   True  True  True
1  True   True  True  True
2  True  False  True  True
3  True   True  True  True
4  True   True  True  True
```

In [63]:
```
print("\nNotnull:\n", df_from_csv.notnull())
```

```
Notnull:
       A      B     C     D
0  True   True  True  True
1  True   True  True  True
2  True  False  True  True
3  True   True  True  True
4  True   True  True  True
```

1. Drop rows with missing values

In [64]:
```
df_dropna = df_from_csv.dropna()
print("\nDropna:\n", df_dropna)
```

```
Dropna:
          A    B     C      D
0  0.496714  1.0   foo   True
1 -0.138264  2.0   bar  False
3  1.523030  4.0   qux  False
4 -0.234153  5.0  quux   True
```

1. Fill missing values with a specific value

In [65]:
```
df_fillna = df_from_csv.fillna(0)
print("\nFillna:\n", df_fillna)
```

```
Fillna:
          A    B     C      D
0  0.496714  1.0   foo   True
1 -0.138264  2.0   bar  False
2  0.647689  0.0   baz   True
3  1.523030  4.0   qux  False
4 -0.234153  5.0  quux   True
```

1. Replace values with another value

In [66]:
```
df_replace = df_from_csv.replace({'baz': 'replaced_value'})
print("\nReplace:\n", df_replace)
```

```
Replace:
          A    B              C      D
0  0.496714  1.0            foo   True
1 -0.138264  2.0            bar  False
2  0.647689  NaN  replaced_value   True
3  1.523030  4.0            qux  False
4 -0.234153  5.0           quux   True
```

1. Interpolate missing values

In [67]:
```
df_interpolate = df_from_csv.interpolate()
print("\nInterpolate:\n", df_interpolate)
```

```
Interpolate:
          A    B     C      D
0  0.496714  1.0   foo   True
1 -0.138264  2.0   bar  False
2  0.647689  3.0   baz   True
3  1.523030  4.0   qux  False
4 -0.234153  5.0  quux   True
```

1. Creating a bool series for NaN values

In [68]:
```
bool_series = df_from_csv.isna()
print("\nBool Series for NaN Values:\n", bool_series)
```

```
Bool Series for NaN Values:
       A      B      C      D
0  False  False  False  False
1  False  False  False  False
2  False   True  False  False
3  False  False  False  False
4  False  False  False  False
```

1. Filtering data based on a condition

In [69]:
```
filtered_data = df_from_csv[df_from_csv['B'] > 2]
print("\nFiltered Data:\n", filtered_data)
```

```
Filtered Data:
          A    B     C      D
3  1.523030  4.0   qux  False
4 -0.234153  5.0  quux   True
```

1. Creating a DataFrame using a dictionary

In [70]:
```
new_data = {'A': [1.0, 2.0, 3.0], 'B': [4, 5, 6]}
new_df = pd.DataFrame(new_data)
print("\nNew DataFrame from Dictionary:\n", new_df)
```

```
New DataFrame from Dictionary:
      A  B
0  1.0  4
1  2.0  5
2  3.0  6
```

1. Using notnull() function

In [71]:
```
not_null_values = df_from_csv.notnull()
print("\nUsing notnull() function:\n", not_null_values)
```

```
Using notnull() function:
       A      B     C     D
0   True   True  True  True
1   True   True  True  True
2   True  False  True  True
3   True   True  True  True
4   True   True  True  True
```

1. Filling a missing value

In [72]:
```
df_fill_specific_value = df_from_csv['B'].fillna(-1)
print("\nFilling a Missing Value:\n", df_fill_specific_value)
```

```
Filling a Missing Value:
 0    1.0
1    2.0
2   -1.0
3    4.0
4    5.0
Name: B, dtype: float64
```