ASSIGNMENT 1

DATA WRANGLING 1

Peform the following operations using Python on any open source dataset :

1. Import all the required Python Libraries
2. Locate an open source data
3. Load the dataset into pandas data frame
4. Data preprocessing
5. Data Formatting and Data Normalization
6. Turn categorical variables into quantitative variables in Python

> Importing Required Libraries, Loading the dataset

In [1]:

```python
import pandas as pd
import numpy as np
```

In [2]:

```python
df = pd.read_csv("titanic1.csv")
```

In [3]:

```
df
```

Out[3]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2 3101282
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450
...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C 6607
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376

891 rows × 12 columns

Data Preprocessing

In [4]:

```
# first 5 rows
df.head()
```

Out[4]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450

In [5]:

```python
# last 5 rows
df.tail()
```

Out[5]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	1
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	3
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	2
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	3
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	

In [6]:

```python
# checks total size(rows*columns)
df.size
```

Out[6]:

```
10692
```

In [7]:

```python
# checks dimensions of dataframe
df.shape
```

Out[7]:

```
(891, 12)
```

In [8]:

```python
# checks the columns present
df.columns
```

Out[8]:

```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibS
p',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
      dtype='object')
```

In [9]:

```
# checks datatypes of each column
df.dtypes
```

Out[9]:

```
PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
Embarked        object
dtype: object
```

In [10]:

```
# prints information about the dataframe
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null    int64  
 3   Name         891 non-null    object 
 4   Sex          891 non-null    object 
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64  
 7   Parch        891 non-null    int64  
 8   Ticket       891 non-null    object 
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object 
 11  Embarked     889 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

In [11]:

```
# checks initial statistics
df.describe()
```

Out[11]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512

In [12]:

```
# checks for missing values
df.isnull().sum()
```

Out[12]:

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

Since there are null values present, we will replace or fill them.

> Data Formatting and Normalization

We can handle missing values by:

1. replace() : method by which we can replace the null values with our own values.

In [13]:

```
df['Cabin'] =df['Cabin'].replace(to_replace = np.nan,value = "unknown")
```

```
In [14]: df
```

Out[14]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticke
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 2117
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 1759
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2 310128
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	11380
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	37345
...	...	...	...	...	...	...	...	...	..
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	21153
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	11205
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C 660
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	11136
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	37037

891 rows × 12 columns

2. interpolate() : performs linear interpolation(by default,can be changed) to replcae null values.

In [17]:

```
df['Age'] = df['Age'].interpolate()
```

In [18]:

```
df
```

Out[18]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450
...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	22.5	1	2	W./C. 6607
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376

891 rows × 12 columns

3. fillna() : method by which we can perform forward fill(ffill) or backward fill(bfill) to handle missing values.

In [19]:

```
df['Embarked'].fillna(method='ffill',inplace=True)
```

```
In [20]: df
```

Out[20]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450
...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	22.5	1	2	W./C. 6607
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376

891 rows × 12 columns

In [21]:

```
df.isnull().sum()
```

Out[21]:

```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Cabin          0
Embarked       0
dtype: int64
```

We can drop all the rows having null values at once, just by using dropna().

> Data Transformation

In [22]:

```
#changing datatypes
df.dtypes
```

Out[22]:

```
PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
Embarked        object
dtype: object
```

In [23]:

```
df['Age'] = df['Age'].astype('int64')
```

In [24]:

```
df.dtypes
```

Out[24]:

```
PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age              int64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
Embarked        object
dtype: object
```

> Turning categorical variables into quantitative variables

There are multiple ways to convert categorical variables into quantitative variables :

1. Dummy variables
2. One Hot Encoding
3. Label Encoding

We will use dummy variables here.

In [26]:

```
quantitative_data = pd.get_dummies(df.Embarked,prefix='Embarked')
```

In [27]:

```
quantitative_data
```

Out[27]:

```
     Embarked_C  Embarked_Q  Embarked_S
  0           0           0           1
  1           1           0           0
  2           0           0           1
  3           0           0           1
  4           0           0           1
...         ...         ...         ...
886           0           0           1
887           0           0           1
888           0           0           1
889           1           0           0
890           0           1           0

891 rows × 3 columns
```

In [28]:

```
df = df.join(quantitative_data)
```

In [29]:
```
df
```

Out[29]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
0	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38	1	0	PC 17599
2	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803
4	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450
...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27	0	0	211536
887	888	1	1	Graham, Miss. Margaret Edith	female	19	0	0	112053
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	22	1	2	W./C. 6607
889	890	1	1	Behr, Mr. Karl Howell	male	26	0	0	111369
890	891	0	3	Dooley, Mr. Patrick	male	32	0	0	370376

891 rows × 15 columns

In [30]:
```
df.drop(['Embarked'],axis=1,inplace=True)
```

In [31]:

```
df
```

Out[31]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
0	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38	1	0	PC 17599
2	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803
4	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450
...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27	0	0	211536
887	888	1	1	Graham, Miss. Margaret Edith	female	19	0	0	112053
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	22	1	2	W./C. 6607
889	890	1	1	Behr, Mr. Karl Howell	male	26	0	0	111369
890	891	0	3	Dooley, Mr. Patrick	male	32	0	0	370376

891 rows × 14 columns

In [32]:

```
quantitative_sex = pd.get_dummies(df.Sex,prefix='Sex')
```

In [33]:

```
quantitative_sex
```

Out[33]:

	Sex_female	Sex_male
0	0	1
1	1	0
2	1	0
3	1	0
4	0	1
...	...	...
886	0	1
887	1	0
888	1	0
889	0	1
890	0	1

891 rows × 2 columns

In [34]:

```
df = df.join(quantitative_sex)
```

In [35]:

```
df
```

Out[35]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
0	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38	1	0	PC 17599
2	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803
4	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450
...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27	0	0	211536
887	888	1	1	Graham, Miss. Margaret Edith	female	19	0	0	112053
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	22	1	2	W./C. 6607
889	890	1	1	Behr, Mr. Karl Howell	male	26	0	0	111369
890	891	0	3	Dooley, Mr. Patrick	male	32	0	0	370376

891 rows × 16 columns

In [36]:

```
df.drop(['Sex'],axis=1,inplace=True)
```

```
In [37]:  df
```

Out[37]:

	PassengerId	Survived	Pclass	Name	Age	SibSp	Parch	Ticket	Far
0	1	0	3	Braund, Mr. Owen Harris	22	1	0	A/5 21171	7.250
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	38	1	0	PC 17599	71.283
2	3	1	3	Heikkinen, Miss. Laina	26	0	0	STON/O2. 3101282	7.925
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	35	1	0	113803	53.100
4	5	0	3	Allen, Mr. William Henry	35	0	0	373450	8.050
...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	27	0	0	211536	13.000
887	888	1	1	Graham, Miss. Margaret Edith	19	0	0	112053	30.000
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	22	1	2	W./C. 6607	23.450
889	890	1	1	Behr, Mr. Karl Howell	26	0	0	111369	30.000
890	891	0	3	Dooley, Mr. Patrick	32	0	0	370376	7.750

891 rows × 15 columns