Data Wrangling, I

Perform the following operations using Python on any open source dataset (e.g., data.csv)

1. Import all the required Python Libraries.
2. Locate an open source data from the web (e.g., https://www.kaggle.com). Provide a clear description of the data and its source (i.e., URL of the web site).
3. Load the Dataset into pandas dataframe.
4. Data Preprocessing: check for missing values in the data using pandas isnull(), describe() function to get some initial statistics. Provide variable descriptions. Types of variables etc. Check the dimensions of the data frame.
5. Data Formatting and Data Normalization: Summarize the types of variables by checking the data types (i.e., character, numeric, integer, factor, and logical) of the variables in the data set. If variables are not in the correct data type, apply proper type conversions.
6. Turn categorical variables into quantitative variables in Python. In addition to the codes and outputs, explain every operation that you do in the above steps and explain everything that you do to import/read/scrape the data set.

1. Import all the required Python Libraries

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

2. Load dataset

```
df = pd.read_csv("datasets/ass1/tested.csv")
df
```

```
     PassengerId  Survived  Pclass  \
0            892         0       3
1            893         1       3
2            894         0       2
3            895         0       3
4            896         1       3
..           ...       ...     ...
413         1305         0       3
414         1306         1       1
415         1307         0       3
416         1308         0       3
417         1309         0       3
```

```
                                                Name     Sex   Age  SibSp
Parch  \
0                                  Kelly, Mr. James    male  34.5      0
0
1                  Wilkes, Mrs. James (Ellen Needs)  female  47.0      1
0
2                         Myles, Mr. Thomas Francis    male  62.0      0
0
3                                  Wirz, Mr. Albert    male  27.0      0
0
4    Hirvonen, Mrs. Alexander (Helga E Lindqvist)  female  22.0      1
1
..                                             ...     ...   ...    ...
...
413                            Spector, Mr. Woolf    male   NaN      0
0
414                  Oliva y Ocana, Dona. Fermina  female  39.0      0
0
415                  Saether, Mr. Simon Sivertsen    male  38.5      0
0
416                           Ware, Mr. Frederick    male   NaN      0
0
417                     Peter, Master. Michael J    male   NaN      1
1

                  Ticket      Fare Cabin Embarked
0                 330911    7.8292   NaN        Q
1                 363272    7.0000   NaN        S
2                 240276    9.6875   NaN        Q
3                 315154    8.6625   NaN        S
4                3101298   12.2875   NaN        S
..                   ...       ...   ...      ...
413           A.5. 3236    8.0500   NaN        S
414            PC 17758  108.9000  C105        C
415  SOTON/O.Q. 3101262    7.2500   NaN        S
416              359309    8.0500   NaN        S
417                2668   22.3583   NaN        C

[418 rows x 12 columns]
```

Link for the dataset = https://www.kaggle.com/datasets/brendan45774/test-file/data

Description : This dataset is a titanic dataset imported of from kaggle website This is great for making charts to help you visualize. This also will help you know who died or survived.

The dataset contains 12 columns:

- PassengerId - ID of Passenger
- Survived - If that particular passenger survived or not
- Pclass - Describes the class of passengers

- Name - Name of the passenger
- Sex - Gender of passenger
- Age - Age of Passenger
- Sibsp - Number of Siblings/spouses aboard
- Parch - Number of parents/children aboard
- Ticket - Ticket id
- Fare - Price of the Ticket
- Cabin - Cabin number the passenger was in
- Embarked - Where the passenger boarded from

4. Data Preprocessing

```
df.isna().sum()
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age             86
SibSp            0
Parch            0
Ticket           0
Fare             1
Cabin          327
Embarked         0
dtype: int64
```

Datase has 86 missing values in Age column and 327 in Cabin column

```
df.describe()
```

```
       PassengerId    Survived      Pclass         Age       SibSp  \
count   418.000000  418.000000  418.000000  332.000000  418.000000
mean   1100.500000    0.363636    2.265550   30.272590    0.447368
std     120.810458    0.481622    0.841838   14.181209    0.896760
min     892.000000    0.000000    1.000000    0.170000    0.000000
25%     996.250000    0.000000    1.000000   21.000000    0.000000
50%    1100.500000    0.000000    3.000000   27.000000    0.000000
75%    1204.750000    1.000000    3.000000   39.000000    1.000000
max    1309.000000    1.000000    3.000000   76.000000    8.000000

            Parch        Fare
count  418.000000  417.000000
mean     0.392344   35.627188
std      0.981429   55.907576
min      0.000000    0.000000
```

```
25%      0.000000    7.895800
50%      0.000000   14.454200
75%      0.000000   31.500000
max      9.000000  512.329200
```

Variable Description

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  418 non-null    int64  
 1   Survived     418 non-null    int64  
 2   Pclass       418 non-null    int64  
 3   Name         418 non-null    object 
 4   Sex          418 non-null    object 
 5   Age          332 non-null    float64
 6   SibSp        418 non-null    int64  
 7   Parch        418 non-null    int64  
 8   Ticket       418 non-null    object 
 9   Fare         417 non-null    float64
 10  Cabin        91 non-null     object 
 11  Embarked     418 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 39.3+ KB
```

Dimensions of dataframe

```
df.shape
```

```
(418, 12)
```

Data Formatting and Data Normalization

Type Conversions

```
df['Survived'] = df['Survived'].astype(bool)
df['Survived'].head()
```

```
0    False
1     True
```

```
2    False
3    False
4     True
Name: Survived, dtype: bool
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  418 non-null    int64  
 1   Survived     418 non-null    bool   
 2   Pclass       418 non-null    int64  
 3   Name         418 non-null    object 
 4   Sex          418 non-null    object 
 5   Age          332 non-null    float64
 6   SibSp        418 non-null    int64  
 7   Parch        418 non-null    int64  
 8   Ticket       418 non-null    object 
 9   Fare         417 non-null    float64
 10  Cabin        91 non-null     object 
 11  Embarked     418 non-null    object 
dtypes: bool(1), float64(2), int64(4), object(5)
memory usage: 36.5+ KB
```

```
df['Sex'] = df['Sex'].apply(lambda x: 1 if x=='male' else 0)
df['Sex']
```

```
0      1
1      0
2      1
3      1
4      0
      ..
413    1
414    0
415    1
416    1
417    1
Name: Sex, Length: 418, dtype: int64
```

```
df.isna().sum()
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age             86
```

```
SibSp            0
Parch            0
Ticket           0
Fare             1
Cabin          327
Embarked         0
dtype: int64
```

```
df.ffill(inplace = True)
```

```
df.isna().sum()
```

```
PassengerId     0
Survived        0
Pclass          0
Name            0
Sex             0
Age             0
SibSp           0
Parch           0
Ticket          0
Fare            0
Cabin          12
Embarked        0
dtype: int64
```

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

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  418 non-null    int64  
 1   Survived     418 non-null    bool   
 2   Pclass       418 non-null    int64  
 3   Name         418 non-null    object 
 4   Sex          418 non-null    int64  
 5   Age          418 non-null    int32  
 6   SibSp        418 non-null    int64  
 7   Parch        418 non-null    int64  
 8   Ticket       418 non-null    object 
 9   Fare         418 non-null    float64
 10  Cabin        406 non-null    object 
 11  Embarked     418 non-null    object 
dtypes: bool(1), float64(1), int32(1), int64(5), object(4)
memory usage: 34.8+ KB
```

6. categorical to quantitative

```
df = pd.get_dummies(df,columns=['Embarked','Sex','Pclass'])
df
```

```
     PassengerId  Survived
Name  Age  \
0            892     False                                Kelly, Mr.
James   34
1            893      True              Wilkes, Mrs. James (Ellen
Needs)   47
2            894     False                     Myles, Mr. Thomas
Francis   62
3            895     False                                Wirz, Mr.
Albert   27
4            896      True  Hirvonen, Mrs. Alexander (Helga E
Lindqvist)   22
..           ...       ...                                        .
..  ...
413         1305     False                             Spector, Mr.
Woolf   28
414         1306      True                   Oliva y Ocana, Dona.
Fermina   39
415         1307     False                   Saether, Mr. Simon
Sivertsen   38
416         1308     False                            Ware, Mr.
Frederick   38
417         1309     False                       Peter, Master. Michael
J   38

     SibSp  Parch              Ticket      Fare Cabin  Embarked_C
Embarked_Q  \
0        0      0              330911    7.8292   NaN       False
True
1        1      0              363272    7.0000   NaN       False
False
2        0      0              240276    9.6875   NaN       False
True
3        0      0              315154    8.6625   NaN       False
False
4        1      1             3101298   12.2875   NaN       False
False
..     ...    ...                 ...       ...   ...         ...
...
413      0      0          A.5. 3236    8.0500   C78       False
False
414      0      0           PC 17758  108.9000  C105        True
False
415      0      0  SOTON/O.Q. 3101262    7.2500  C105       False
False
```

```
416      0      0              359309    8.0500  C105       False
False
417      1      1                2668   22.3583  C105        True
False

     Embarked_S  Sex_0  Sex_1  Pclass_1  Pclass_2  Pclass_3
0         False  False   True     False     False      True
1          True   True  False     False     False      True
2         False  False   True     False      True     False
3          True  False   True     False     False      True
4          True   True  False     False     False      True
..          ...    ...    ...       ...       ...       ...
413        True  False   True     False     False      True
414       False   True  False      True     False     False
415        True  False   True     False     False      True
416        True  False   True     False     False      True
417       False  False   True     False     False      True

[418 rows x 17 columns]
```

```
df.drop(columns=['Embarked_S','Sex_1','Pclass_3'],inplace = True)
```

```
df.Cabin.bfill(inplace = True)
```

```
df
```

```
     PassengerId  Survived
Name  Age  \
0            892     False                                Kelly, Mr.
James   34
1            893      True              Wilkes, Mrs. James (Ellen
Needs)   47
2            894     False                     Myles, Mr. Thomas
Francis   62
3            895     False                                Wirz, Mr.
Albert   27
4            896      True  Hirvonen, Mrs. Alexander (Helga E
Lindqvist)   22
..           ...       ...                                        .
..  ...
413         1305     False                             Spector, Mr.
Woolf   28
414         1306      True                   Oliva y Ocana, Dona.
Fermina   39
415         1307     False                   Saether, Mr. Simon
Sivertsen   38
416         1308     False                            Ware, Mr.
Frederick   38
417         1309     False                       Peter, Master. Michael
J   38
```

```
     SibSp  Parch              Ticket      Fare Cabin  Embarked_C
Embarked_Q  \
0        0      0              330911    7.8292   B45       False
True
1        1      0              363272    7.0000   B45       False
False
2        0      0              240276    9.6875   B45       False
True
3        0      0              315154    8.6625   B45       False
False
4        1      1             3101298   12.2875   B45       False
False
..     ...    ...                 ...       ...   ...         ...
...
413      0      0           A.5. 3236    8.0500   C78       False
False
414      0      0            PC 17758  108.9000  C105        True
False
415      0      0  SOTON/O.Q. 3101262    7.2500  C105       False
False
416      0      0              359309    8.0500  C105       False
False
417      1      1                2668   22.3583  C105        True
False

     Sex_0  Pclass_1  Pclass_2
0    False     False     False
1     True     False     False
2    False     False      True
3    False     False     False
4     True     False     False
..     ...       ...       ...
413  False     False     False
414   True      True     False
415  False     False     False
416  False     False     False
417  False     False     False

[418 rows x 14 columns]
```