# **Data Wrangling**

• Importing Libraries

```
In [ ]:
          import numpy as np
          import pandas as pd
          import seaborn as sns
In [ ]:
          kashti=sns.load_dataset("Titanic")
          ks1=kashti
          ks2=kashti
In [ ]:
          kashti.head()
Out[ ]:
            survived pclass
                                                                                         adult_male
                                          sibsp
                                                parch
                                                          fare
                                                               embarked
                                                                          class
                                                                                   who
                                                                                                    deck
                                     age
         0
                   0
                          3
                               male
                                     22.0
                                              1
                                                        7.2500
                                                                          Third
                                                                                   man
                                                                                               True
                                                                                                     NaN
                             female
                                    38.0
                                                       71.2833
                                                                                              False
                                                                                                       C
                                                                           First woman
         2
                             female
                                    26.0
                                              0
                                                    0
                                                        7.9250
                                                                          Third
                                                                                              False
                                                                                woman
                                                                                                     NaN
         3
                             female
                                     35.0
                                              1
                                                       53.1000
                                                                                              False
                                                                                                       C
                                                                           First
                                                                                woman
         4
                   0
                          3
                               male 35.0
                                              0
                                                    0
                                                        8.0500
                                                                          Third
                                                                                               True
                                                                                                    NaN
                                                                                   man
In [ ]:
          (kashti['age']+12).head(10)
               34.0
Out[]:
               50.0
         2
               38.0
         3
               47.0
         4
               47.0
         5
                NaN
         6
               66.0
         7
               14.0
         8
               39.0
         9
               26.0
         Name: age, dtype: float64
In [ ]:
          kashti.isnull().sum()
```

```
survived
                           0
Out[]:
         pclass
                           0
         sex
                           0
         age
                         177
         sibsp
                           0
         parch
                           0
         fare
                           0
         embarked
                           2
         class
                           0
         who
                           0
         adult_male
                           0
         deck
                         688
         embark_town
                           2
         alive
                           0
         alone
                           0
         dtype: int64
In [ ]:
          kashti.shape
         (891, 15)
Out[]:
In [ ]:
          # dropna method
          kashti.dropna(subset=['deck'],axis=0,inplace=True)
          kashti.isnull().sum()
                          0
         survived
Out[]:
                          0
         pclass
                          0
         sex
         age
                         19
                          0
         sibsp
         parch
                          0
         fare
                          0
         embarked
                          2
                          0
         class
                          0
         who
                          0
         adult_male
                          0
         deck
         embark_town
                          2
         alive
                          0
         alone
                          0
         dtype: int64
In [ ]:
          kashti.shape
         (203, 15)
Out[ ]:
In [ ]:
          # to remove null value from whole dataset
          kashti=kashti.dropna()
          kashti.isnull().sum()
```

```
survived
                         0
Out[]:
                         0
         pclass
         sex
                         0
                         0
         age
                        0
         sibsp
         parch
                         0
         fare
                         0
         embarked
                         0
         class
                         0
         who
                         0
         adult_male
                        0
         deck
                         0
                         0
         embark_town
         alive
                         0
         alone
         dtype: int64
In [ ]:
          kashti.shape
         (182, 15)
Out[]:
        Replace the missing value with the average of that column
In [ ]:
```

```
# finding the average(mean)
         mean=ks1['age'].mean()
         mean
         35.77945652173913
Out[ ]:
In [ ]:
         ks1['age']=ks1['age'].replace(np.nan,mean)
In [ ]:
         # replacing nan with mean
         ks1.isnull().sum()
        survived
                        0
Out[]:
        pclass
                        0
                        0
        sex
                        0
        age
        sibsp
                        0
                        0
        parch
        fare
                        0
                        2
        embarked
        class
                        0
        who
                        0
        adult_male
        deck
                        0
        embark_town
                        2
                        0
        alive
                        0
        alone
        dtype: int64
```

### **Assignment 1**

```
In [ ]:
         kashti=sns.load_dataset("Titanic")
```

```
ks2=kashti
ks2.head()
```

```
Out[ ]:
                survived pclass
                                                                   embarked
                                                                              class
                                                                                             adult_male
                                   sex
                                        age sibsp
                                                    parch
                                                              fare
                                                                                       who
                                                                                                        deck
                      0
            0
                             3
                                  male
                                        22.0
                                                 1
                                                        0
                                                            7.2500
                                                                              Third
                                                                                       man
                                                                                                   True
                                                                                                         NaN
             1
                      1
                                female
                                        38.0
                                                 1
                                                           71.2833
                                                                           C
                                                                               First woman
                                                                                                   False
                                                                                                           C
                              1
             2
                      1
                              3
                                female
                                        26.0
                                                 0
                                                        0
                                                            7.9250
                                                                           S
                                                                              Third
                                                                                    woman
                                                                                                   False
                                                                                                         NaN
             3
                                 female
                                        35.0
                                                 1
                                                           53.1000
                                                                           S
                                                                               First woman
                                                                                                   False
                                                                                                           C
             4
                      0
                              3
                                  male 35.0
                                                 0
                                                            8.0500
                                                                              Third
                                                                                       man
                                                                                                   True
                                                                                                         NaN
4
   In [ ]:
             ks2.isnull().sum()
            survived
                               0
   Out[ ]:
            pclass
                                0
            sex
                                0
            age
                             177
            sibsp
                               0
            parch
                                0
            fare
                                0
            embarked
                                2
            class
                                0
            who
                                0
            adult male
                                0
            deck
                             688
            embark_town
                                2
            alive
                                0
            alone
                                0
            dtype: int64
   In [ ]:
             mode=ks2['deck'].mode
             mode
            <bound method Series.mode of 0</pre>
                                                     NaN
   Out[]:
            1
                      C
            2
                    NaN
            3
                       C
            4
                    NaN
            886
                    NaN
            887
                       В
            888
                    NaN
            889
                       C
            890
                    NaN
            Name: deck, Length: 891, dtype: object>
   In [ ]:
             ks2['deck']=ks2['deck'].replace(np.nan,mode)
   In [ ]:
             ks2.isnull().sum()
```

```
survived
                           0
Out[ ]:
         pclass
                           0
         sex
                           0
                         177
         age
         sibsp
                           0
         parch
                           0
         fare
                           0
         embarked
                           2
         class
                           0
         who
                           0
         adult_male
                           0
         deck
                           0
                           2
         embark_town
         alive
                           0
         alone
         dtype: int64
In [ ]:
         ks2['embark_town'].head()
              Southampton
Out[]:
         1
                Cherbourg
         2
              Southampton
         3
              Southampton
              Southampton
         Name: embark_town, dtype: object
In [ ]:
         ks2['embarked'].head()
              S
Out[]:
              C
         2
              S
         3
              S
         Name: embarked, dtype: object
In [ ]:
         mod=ks2['embark_town'].mode
         mod
         <bound method Series.mode of 0</pre>
                                               Southampton
Out[]:
                  Cherbourg
         2
                Southampton
         3
                Southampton
         4
                Southampton
         886
                Southampton
         887
                Southampton
         888
                Southampton
         889
                  Cherbourg
         890
                 Queenstown
         Name: embark town, Length: 891, dtype: object>
In [ ]:
          ks2['embark_town']=ks2['embark_town'].replace(np.nan,mod)
In [ ]:
          ks2.isnull().sum()
```

```
survived
                           0
Out[]:
         pclass
                           0
         sex
                           0
                         177
         age
         sibsp
                           0
         parch
                           0
         fare
                           0
         embarked
                           2
         class
                           0
         who
                           0
         adult_male
                           0
         deck
                           0
         embark_town
                           0
         alive
                           0
         alone
         dtype: int64
In [ ]:
          mo=ks2['embarked'].mode
         <bound method Series.mode of 0</pre>
                                               S
Out[]:
                C
                S
         2
         3
                S
                S
         4
         886
                S
         887
                S
         888
                S
                C
         889
         890
                Q
         Name: embarked, Length: 891, dtype: object>
In [ ]:
          ks2['embarked']=ks2['embarked'].replace(np.nan,mo)
In [ ]:
          ks2.isnull().sum()
         survived
                           0
Out[]:
         pclass
                           0
                           0
         sex
         age
                         177
         sibsp
                           0
         parch
                           0
         fare
                           0
         embarked
                           0
         class
                           0
         who
                           0
         adult_male
                           0
         deck
                           0
         embark_town
                           0
         alive
                           0
         alone
                           0
         dtype: int64
In [ ]:
          mean=ks2['age'].mean()
          mean
```

```
29.69911764705882
Out[ ]:
In [ ]:
          ks2['age']=ks2['age'].replace(np.nan,mean)
In [ ]:
          ks2.isnull().sum()
         survived
                         0
Out[]:
         pclass
                         0
                         0
         sex
                         0
         age
         sibsp
                         0
         parch
                         0
         fare
                         0
         embarked
                         0
         class
                         0
                         0
         who
         adult_male
                         0
         deck
                         0
         embark_town
                         0
                         0
         alive
         alone
         dtype: int64
In [ ]:
          ks2.shape
         (891, 15)
Out[ ]:
```

## **Data Formatting**

```
In [ ]:
         kashti=sns.load_dataset('titanic')
         kashti.dtypes
        survived
                           int64
Out[ ]:
                           int64
        pclass
                          object
        sex
        age
                         float64
                           int64
        sibsp
                           int64
        parch
        fare
                         float64
        embarked
                          object
        class
                        category
                          object
        who
        adult_male
                            bool
        deck
                        category
        embark_town
                          object
        alive
                          object
                            bool
        alone
        dtype: object
In [ ]:
         kashti['survived']=kashti['survived'].astype('float64')
         kashti.dtypes
```

```
survived
                         float64
Out[ ]:
        pclass
                           int64
        sex
                          object
                         float64
        age
                           int64
        sibsp
        parch
                           int64
        fare
                         float64
        embarked
                          object
        class
                        category
        who
                          object
        adult_male
                            bool
        deck
                        category
        embark_town
                          object
        alive
                          object
                            bool
        alone
        dtype: object
In [ ]:
          kashti['survived']=kashti['survived'].astype('int64')
         kashti.dtypes
        survived
                           int64
Out[]:
        pclass
                           int64
        sex
                          object
                         float64
        age
        sibsp
                           int64
        parch
                           int64
        fare
                         float64
        embarked
                          object
        class
                        category
        who
                          object
        adult male
                            bool
        deck
                        category
        embark_town
                          object
                          object
        alive
        alone
                            bool
        dtype: object
In [ ]:
          kashti['age']=kashti['age']*365
         kashti['age'].head()
               8030.0
Out[]:
              13870.0
        1
         2
               9490.0
              12775.0
        3
              12775.0
        Name: age, dtype: float64
In [ ]:
          kashti.rename(columns={'age':'age in days'},inplace=True)
         kashti.head()
```

Out[ ]:		survived	pclass	sex	age in days	sibsp	parch	fare	embarked	class	who	adult_male	d٤
	0	0	3	male	8030.0	1	0	7.2500	S	Third	man	True	Ν
	1	1	1	female	13870.0	1	0	71.2833	С	First	woman	False	
	2	1	3	female	9490.0	0	0	7.9250	S	Third	woman	False	Ν
	3	1	1	female	12775.0	1	0	53.1000	S	First	woman	False	
	4	0	3	male	12775.0	0	0	8.0500	S	Third	man	True	Ν
													•

## **Assignment 2**

```
In [ ]:
         kashti['fare'].round(1)
                 7.2
Out[]:
                71.3
                 7.9
                53.1
                 8.0
         886
                13.0
         887
                30.0
         888
                23.4
                30.0
         889
                 7.8
         890
         Name: fare, Length: 891, dtype: float64
```

#### **Data Normalization**

- Simple feature Scaling
  - x(new)=x(old)/x(max)
- Min-Max method
- Z-score(Standard Score -3 to +3)
- Log Transformation

```
In [ ]: kashti[['age in days','fare']].head()
```

```
Out[]: age in days fare

0 8030.0 7.2500

1 13870.0 71.2833

2 9490.0 7.9250

3 12775.0 53.1000

4 12775.0 8.0500
```

```
In [ ]:  # Simple Feature Scaling
```

```
kashti['fare']=kashti['fare']/kashti['fare'].max()
         kashti['fare'].head()
             0.014151
Out[ ]:
        1
             0.139136
             0.015469
        2
             0.103644
        3
             0.015713
        Name: fare, dtype: float64
In [ ]:
         # Min-Max method
         kashti['fare'].head()
             0.014151
Out[ ]:
             0.139136
        1
        2
             0.015469
             0.103644
        3
             0.015713
        Name: fare, dtype: float64
In [ ]:
         kashti['age in days']=kashti['age in days']/kashti['age in days'].max()
         kashti['age in days']=(kashti['age in days']-kashti['age in days'].min())/(kashti['age in days'].min())/
         kashti['age in days'].head()
             0.271174
Out[]:
             0.472229
        1
        2
             0.321438
             0.434531
        3
             0.434531
        Name: age in days, dtype: float64
In [ ]:
         kashti[['age in days','fare']].head()
Out[]:
           age in days
                         fare
        0
             0.271174 0.014151
        1
             0.472229 0.139136
        2
             3
             0.434531 0.103644
             0.434531 0.015713
In [ ]:
         i=sns.load_dataset('titanic')
         i.head(3)
           survived pclass
Out[ ]:
                            sex
                                age sibsp parch
                                                   fare embarked
                                                                  class
                                                                         who
                                                                              adult_male
                                                                                        deck
                                                                                         NaN
        0
                 0
                                22.0
                                                  7.2500
                                                                  Third
                       3
                           male
                                        1
                                              0
                                                                                    True
                                                                         man
        1
                                38.0
                                                71.2833
                                                                                           C
                 1
                         female
                                        1
                                              0
                                                               C
                                                                   First woman
                                                                                   False
                       1
        2
                 1
                         female
                                26.0
                                        0
                                              0
                                                  7.9250
                                                               S Third woman
                       3
                                                                                   False
                                                                                         NaN
```

```
In [ ]:
                          # Log Transformation
                          i['age']=np.log(i['age'])
                          i['age'].head(3)
                        C:\Users\786\AppData\Local\Programs\Python\Python39\lib\site-packages\pandas\core\arr
                        aylike.py:364: RuntimeWarning: divide by zero encountered in log
                              result = getattr(ufunc, method)(*inputs, **kwargs)
                         C: \Users \786 \App Data \Local \Programs \Python \Python 39 \Lib \site-packages \pandas \core \arresponse \Python 
                        aylike.py:364: RuntimeWarning: invalid value encountered in log
                              result = getattr(ufunc, method)(*inputs, **kwargs)
                                      1.128508
Out[]:
                                      1.291320
                                      1.181143
                        Name: age, dtype: float64
                      Bining
In [ ]:
                           import seaborn as sns
                           import pandas as pd
                           import numpy as np
                           k=sns.load dataset('titanic')
                          k.head(3)
Out[]:
                                survived pclass
                                                                                               age sibsp parch
                                                                                                                                                      fare
                                                                                                                                                                 embarked
                                                                                                                                                                                                class
                                                                                                                                                                                                                      who adult_male
                                                                                                                                                                                                                                                                 deck
                                                                                  sex
                        0
                                                 0
                                                                    3
                                                                                               22.0
                                                                                                                      1
                                                                                                                                                 7.2500
                                                                                                                                                                                               Third
                                                                               male
                                                                                                                                       0
                                                                                                                                                                                                                      man
                                                                                                                                                                                                                                                     True
                                                                                                                                                                                                                                                                   NaN
                         1
                                                  1
                                                                           female
                                                                                               38.0
                                                                                                                                             71.2833
                                                                                                                                                                                                                                                   False
                                                                                                                                                                                                                                                                          C
                                                                    1
                                                                                                                      1
                                                                                                                                                                                        C
                                                                                                                                                                                                  First woman
                         2
                                                  1
                                                                    3 female 26.0
                                                                                                                     0
                                                                                                                                       0
                                                                                                                                                 7.9250
                                                                                                                                                                                              Third woman
                                                                                                                                                                                                                                                   False
                                                                                                                                                                                                                                                                  NaN
In [ ]:
                          min(k['age'])
                        0.42
Out[ ]:
In [ ]:
                          max(k['age'])
                        80.0
Out[ ]:
In [ ]:
                           bins=np.linspace(min(k['age']),max(k['age']),4)
                           age_groups=['bachay','jawan','bhuray']
                           k['age']=pd.cut(k['age'],bins,labels=age_groups,include_lowest=True)
                           k['age']
```

```
bachay
Out[ ]:
         1
                   jawan
         2
                  bachay
                   jawan
         4
                   jawan
         886
                   jawan
         887
                  bachay
         888
                     NaN
         889
                  bachay
                   jawan
         890
         Name: age, Length: 891, dtype: category
         Categories (3, object): ['bachay' < 'jawan' < 'bhuray']</pre>
In [ ]:
          k.head()
Out[]:
             survived
                      pclass
                                              sibsp
                                                              fare
                                                                    embarked
                                                                               class
                                                                                        who
                                                                                              adult_male
                                 sex
                                        age
                                                    parch
                                                            7.2500
                                                                               Third
                           3
                                male
                                      bachay
                                                                                        man
                                                                                                    True
                                                                                                          Νá
                              female
                                       jawan
                                                           71.2833
                                                                                First woman
                                                                                                    False
          2
                              female
                                      bachay
                                                            7.9250
                                                                               Third
                                                                                     woman
                                                                                                    False
          3
                              female
                                       jawan
                                                           53.1000
                                                                                First woman
                                                                                                    False
                   0
                           3
                                male
                                       jawan
                                                 0
                                                            8.0500
                                                                               Third
                                                                                        man
                                                                                                    True
```

## **Converting Categories into dummies**

- 1. Ease to use for computation
- 2. Male Female(0,1)

```
In [ ]: pd.get_dummies(k['sex'])
```

]:		female	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
	]:	0 1 2 3 4  886 887 888 889	0 0 1 1 2 1 3 1 4 0 886 0 887 1 888 1 889 0

891 rows × 2 columns

In [ ]: k.head()

Out[ ]:		survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	de
	0	0	3	male	bachay	1	0	7.2500	S	Third	man	True	Νέ
	1	1	1	female	jawan	1	0	71.2833	С	First	woman	False	
	2	1	3	female	bachay	0	0	7.9250	S	Third	woman	False	Νέ
	3	1	1	female	jawan	1	0	53.1000	S	First	woman	False	
	4	0	3	male	jawan	0	0	8.0500	S	Third	man	True	Ná
4													

In [ ]: pd.get\_dummies(k['sex'],drop\_first=True,sparse=True)

```
Out[]:
                 male
             0
                    1
             1
                    0
             2
                    0
             3
                    0
             4
                    1
           886
                    1
           887
                    0
           888
                    0
           889
                    1
           890
                    1
```

891 rows × 1 columns

#### **Assignment 3**

```
In [ ]:
          k=sns.load_dataset('titanic')
          k.head(3)
Out[]:
             survived
                       pclass
                                      age sibsp
                                                 parch
                                                             fare
                                                                  embarked
                                                                              class
                                                                                       who
                                                                                            adult_male
                                                                                                        deck
                                 sex
          0
                    0
                           3
                                                           7.2500
                                                                              Third
                                male
                                      22.0
                                                       0
                                                                                       man
                                                                                                   True
                                                                                                         NaN
          1
                                      38.0
                                                1
                                                         71.2833
                                                                               First woman
                                                                                                            C
                           1
                              female
                                                       0
                                                                                                  False
          2
                                               0
                    1
                           3
                              female 26.0
                                                       0
                                                           7.9250
                                                                             Third
                                                                                    woman
                                                                                                  False
                                                                                                         NaN
In [ ]:
           dummies = pd.get dummies(k[['sex']], drop first=False)
           k = pd.concat([k.drop(['sex'],axis=1), dummies],axis=1)
In [ ]:
           k.head()
Out[]:
             survived
                       pclass
                              age
                                    sibsp
                                           parch
                                                     fare
                                                           embarked
                                                                      class
                                                                               who
                                                                                    adult_male
                                                                                                 deck
                                                                                                       embarl
          0
                    0
                              22.0
                                                                      Third
                                                   7.2500
                                                                               man
                                                                                           True
                                                                                                 NaN
                                                                                                        Southa
                                                                                                    C
                                                                                                          Che
          1
                              38.0
                                                  71.2833
                                                                       First
                                                                            woman
                                                                                           False
          2
                    1
                           3
                              26.0
                                               0
                                                   7.9250
                                                                      Third
                                                                            woman
                                                                                           False
                                                                                                 NaN
                                                                                                        Southa
          3
                    1
                           1
                              35.0
                                        1
                                               0
                                                  53.1000
                                                                   S
                                                                       First
                                                                            woman
                                                                                           False
                                                                                                    C
                                                                                                        Southa
                    0
                           3
                              35.0
                                        0
                                               0
                                                   8.0500
                                                                      Third
                                                                               man
                                                                                           True
                                                                                                 NaN
                                                                                                        Southa
```