ETL_Data_Pipeline

September 13, 2024

```
import pandas as pd
[39]: df = pd.read_csv("tested.csv")
      df
[39]:
                                     Pclass
            PassengerId
                          Survived
      0
                    892
                                           3
      1
                    893
                                  1
                                           3
      2
                    894
                                  0
                                           2
      3
                    895
                                  0
                                           3
      4
                                           3
                    896
                                  1
                                           3
                                  0
      413
                    1305
      414
                    1306
                                  1
                                           1
      415
                    1307
                                  0
                                           3
                                  0
                                           3
      416
                    1308
                                           3
      417
                    1309
                                  0
                                                        Name
                                                                              SibSp
                                                                                      Parch
                                                                  Sex
                                                                         Age
      0
                                                                                   0
                                           Kelly, Mr. James
                                                                 male
                                                                        34.5
                                                                                           0
      1
                                                                        47.0
                                                                                           0
                         Wilkes, Mrs. James (Ellen Needs)
                                                               female
                                                                                   1
      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
                                                                                   0
                                                                                           0
                                         Spector, Mr. Woolf
                                                                 male
                                                                         NaN
      414
                              Oliva y Ocana, Dona. Fermina
                                                                                   0
                                                                                           0
                                                               female
                                                                        39.0
      415
                              Saether, Mr. Simon Sivertsen
                                                                                           0
                                                                 male
                                                                        38.5
                                                                                   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
                                                          S
                                              NaN
      2
                         240276
                                    9.6875
                                                          Q
                                              NaN
      3
                         315154
                                    8.6625
                                              NaN
                                                          S
                                                          S
      4
                        3101298
                                   12.2875
                                              NaN
```

```
413
                     A.5. 3236
                                  8.0500
                                            NaN
                                                       S
      414
                      PC 17758
                                           C105
                                                        С
                                108.9000
      415 SOTON/O.Q. 3101262
                                                        S
                                  7.2500
                                            NaN
      416
                                                        S
                        359309
                                  8.0500
                                            NaN
      417
                          2668
                                 22.3583
                                            NaN
                                                        C
      [418 rows x 12 columns]
[40]: df.info()
     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 418 entries, 0 to 417
     Data columns (total 12 columns):
          Column
                        Non-Null Count
                                         Dtype
          PassengerId 418 non-null
      0
                                         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
                        332 non-null
                                         float64
          Age
      6
          SibSp
                        418 non-null
                                         int64
      7
          Parch
                        418 non-null
                                         int64
          Ticket
                        418 non-null
                                         object
          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
[41]: # age value replace with according pclass...
      df['Age'] = df.groupby('Pclass')['Age'].transform(lambda x: x.fillna(x.mean()))
[42]: df
           PassengerId
                         Survived
                                   Pclass
      0
                   892
                                0
                                         3
      1
                   893
                                1
                                         3
      2
                   894
                                0
                                         2
      3
                   895
                                0
                                         3
      4
                   896
                                         3
                                1
                   •••
                                         3
      413
                   1305
                                0
      414
                   1306
                                         1
                                1
      415
                   1307
                                0
                                         3
      416
                   1308
                                0
                                         3
```

[]:

[42]:

```
Name
                                                                 Sex
                                                                             Age
                                                                                   SibSp \
                                          Kelly, Mr. James
      0
                                                                male
                                                                      34.500000
      1
                         Wilkes, Mrs. James (Ellen Needs)
                                                              female
                                                                      47.000000
                                                                                       1
      2
                                Myles, Mr. Thomas Francis
                                                                male
                                                                      62.000000
                                                                                       0
      3
                                          Wirz, Mr. Albert
                                                                male
                                                                      27.000000
                                                                                       0
      4
           Hirvonen, Mrs. Alexander (Helga E Lindqvist)
                                                              female
                                                                      22.000000
                                                                                       1
      . .
      413
                                        Spector, Mr. Woolf
                                                                       24.027945
                                                                male
                                                                                       0
                             Oliva y Ocana, Dona. Fermina
      414
                                                              female
                                                                      39.000000
                                                                                       0
      415
                             Saether, Mr. Simon Sivertsen
                                                                male
                                                                      38.500000
                                                                                       0
      416
                                       Ware, Mr. Frederick
                                                                male
                                                                      24.027945
                                                                                       0
      417
                                 Peter, Master. Michael J
                                                                male
                                                                      24.027945
                                                                                       1
           Parch
                                Ticket
                                             Fare Cabin Embarked
      0
                                330911
                                           7.8292
                                                     NaN
                0
                                                                 Q
                0
                                           7.0000
                                                                 S
      1
                                363272
                                                     NaN
      2
                0
                                240276
                                           9.6875
                                                     NaN
                                                                 Q
      3
                0
                                315154
                                           8.6625
                                                     NaN
                                                                 S
                                                                 S
      4
                1
                               3101298
                                          12.2875
                                                     NaN
      . .
      413
                0
                             A.5. 3236
                                           8.0500
                                                                 S
                                                     NaN
                                                    C105
                                                                 C
      414
                0
                              PC 17758
                                         108.9000
      415
                   SOTON/O.Q. 3101262
                                           7.2500
                                                     NaN
                                                                 S
      416
                0
                                359309
                                           8.0500
                                                     NaN
                                                                 S
                                   2668
      417
                1
                                          22.3583
                                                     NaN
      [418 rows x 12 columns]
[43]: # Drop Irrelevant Columns:
      df.drop(['Name','Cabin','Ticket'], axis=1, inplace = True)
[44]:
     df
[44]:
            PassengerId
                          Survived
                                    Pclass
                                                 Sex
                                                                  SibSp
                                                                          Parch
                                                                                      Fare
                                                             Age
      0
                    892
                                 0
                                          3
                                               male
                                                      34.500000
                                                                       0
                                                                              0
                                                                                    7.8292
                    893
      1
                                 1
                                          3
                                             female
                                                      47.000000
                                                                       1
                                                                              0
                                                                                    7.0000
      2
                    894
                                 0
                                          2
                                               male
                                                      62.000000
                                                                       0
                                                                              0
                                                                                    9.6875
      3
                    895
                                 0
                                          3
                                                                       0
                                                                              0
                                               male
                                                      27.000000
                                                                                    8.6625
      4
                    896
                                 1
                                             female
                                                      22.000000
                                                                                   12.2875
      . .
                    •••
      413
                   1305
                                 0
                                          3
                                               male 24.027945
                                                                       0
                                                                              0
                                                                                    8.0500
      414
                                             female
                                                      39.000000
                                                                      0
                                                                                  108.9000
                   1306
                                 1
                                          1
                                 0
                                          3
      415
                   1307
                                               male
                                                      38.500000
                                                                      0
                                                                              0
                                                                                    7.2500
      416
                   1308
                                 0
                                          3
                                               male
                                                      24.027945
                                                                       0
                                                                              0
                                                                                    8.0500
      417
                   1309
                                 0
                                          3
                                               male
                                                      24.027945
                                                                              1
                                                                                   22.3583
                                                                       1
```

417

1309

0

3

```
Embarked
0
            S
1
            Q
2
            S
3
4
            S
            S
413
414
            С
415
            S
            S
416
            С
417
```

[418 rows x 9 columns]

```
[45]: # Convert Data Types:Convert the Sex column into numerical data for easier

analysis.

df['Sex'] = df['Sex'].map({'male':1 , 'female':0})
```

[46]: df

[46]:	${\tt PassengerId}$	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	\
0	892	0	3	1	34.500000	0	0	7.8292	
1	893	1	3	0	47.000000	1	0	7.0000	
2	894	0	2	1	62.000000	0	0	9.6875	
3	895	0	3	1	27.000000	0	0	8.6625	
4	896	1	3	0	22.000000	1	1	12.2875	
	•••	•••		•••		•••			
413	1305	0	3	1	24.027945	0	0	8.0500	
414	1306	1	1	0	39.000000	0	0	108.9000	
415	1307	0	3	1	38.500000	0	0	7.2500	
416	1308	0	3	1	24.027945	0	0	8.0500	
417	1309	0	3	1	24.027945	1	1	22.3583	

Embarked

0	Q
1	S
2	Q
3	S
4	S
	•••
413	S
414	C
415	S
416	S

```
417 C
```

Pclass

Sex

0

0

[418 rows x 9 columns]

```
[49]: # Feature Engineering: Create new features based on existing ones (e.g., Age_
       \hookrightarrow group).
      df['Age_group'] = pd.cut(df['Age'], bins=[0, 18, 60, 100], labels=["Child", __

¬"Adult", "Senior"])
[50]: df
[50]:
            PassengerId
                          Survived Pclass
                                              Sex
                                                               SibSp
                                                                       Parch
                                                                                   Fare \
                                                          Age
                    892
                                  0
                                                   34.500000
                                                                    0
                                                                                 7.8292
      0
                                                1
      1
                    893
                                  1
                                          3
                                                   47.000000
                                                                    1
                                                                            0
                                                                                 7.0000
                                          2
      2
                    894
                                  0
                                                                    0
                                                1
                                                   62.000000
                                                                            0
                                                                                 9.6875
                                  0
                                          3
      3
                    895
                                                   27.000000
                                                                    0
                                                                                 8.6625
                                                1
      4
                    896
                                          3
                                                   22.000000
                                                                    1
                                                                            1
                                                                                12.2875
                                                1 24.027945
      413
                   1305
                                  0
                                          3
                                                                    0
                                                                           0
                                                                                 8.0500
      414
                                          1
                                                   39.000000
                                                                    0
                                                                           0
                                                                               108.9000
                   1306
                                  1
                                                0
      415
                                  0
                                          3
                                                1
                                                   38.500000
                                                                    0
                                                                            0
                                                                                 7.2500
                   1307
      416
                   1308
                                  0
                                          3
                                                   24.027945
                                                                    0
                                                                            0
                                                                                 8.0500
                                  0
                                          3
      417
                   1309
                                                   24.027945
                                                                    1
                                                                                22.3583
          Embarked Age_group
      0
                  Q
                         Adult
      1
                  S
                         Adult
      2
                  Q
                        Senior
      3
                  S
                         Adult
                  S
      4
                         Adult
      . .
      413
                  S
                         Adult
      414
                         Adult
                  С
                         Adult
      415
                  S
      416
                  S
                         Adult
      417
                  С
                         Adult
      [418 rows x 10 columns]
[51]: df.isnull().sum()
[51]: PassengerId
                       0
      Survived
                       0
```

```
SibSp
                      0
      Parch
                      0
      Fare
      Embarked
                      0
      Age_group
                      0
      dtype: int64
[52]: # Correct way to fill missing values in the 'Fare' column
      df['Fare'] = df['Fare'].fillna(df['Fare'].mean())
[53]: df
[53]:
           PassengerId Survived Pclass
                                                        Age SibSp
                                                                     Parch
                                                                                 Fare \
                                            Sex
                    892
                                 0
                                         3
                                               1
                                                  34.500000
                                                                  0
                                                                          0
                                                                               7.8292
      1
                    893
                                 1
                                         3
                                                  47.000000
                                                                  1
                                                                          0
                                                                               7.0000
                    894
                                 0
                                         2
      2
                                               1
                                                  62.000000
                                                                  0
                                                                          0
                                                                               9.6875
      3
                    895
                                 0
                                         3
                                               1
                                                  27.000000
                                                                  0
                                                                          0
                                                                               8.6625
                    896
                                          3
                                                  22.000000
      4
                                 1
                                                                  1
                                                                              12.2875
                                          3
                                                                          0
      413
                   1305
                                 0
                                                  24.027945
                                                                  0
                                                                               8.0500
                                                                             108.9000
      414
                   1306
                                 1
                                         1
                                               0
                                                  39.000000
                                                                  0
      415
                   1307
                                 0
                                         3
                                                  38.500000
                                                                  0
                                                                               7.2500
                                               1
      416
                                 0
                                         3
                                                  24.027945
                                                                  0
                                                                          0
                                                                               8.0500
                   1308
                                               1
      417
                   1309
                                 0
                                         3
                                                  24.027945
                                                                  1
                                                                          1
                                                                              22.3583
          Embarked Age_group
                        Adult
      0
                  Q
                  S
      1
                        Adult
      2
                  Q
                       Senior
      3
                  S
                        Adult
                  S
      4
                        Adult
      . .
      413
                  S
                        Adult
      414
                  С
                        Adult
      415
                  S
                        Adult
                  S
                        Adult
      416
      417
                  С
                        Adult
      [418 rows x 10 columns]
[55]: df.isnull().sum()
[55]: PassengerId
                      0
      Survived
                      0
      Pclass
                      0
      Sex
                      0
```

Age

0

```
Age
                      0
      SibSp
                      0
                      0
      Parch
      Fare
                      0
      Embarked
                      0
      Age_group
                      0
      dtype: int64
[56]: df.dtypes
[56]: PassengerId
                         int64
      Survived
                         int64
      Pclass
                         int64
      Sex
                         int64
                       float64
      Age
      SibSp
                         int64
      Parch
                         int64
      Fare
                       float64
      Embarked
                        object
      Age_group
                      category
      dtype: object
[57]: import mysql.connector
      from sqlalchemy import create_engine
[58]: engine = create_engine('mysql+pymysql://root:rohit9828@localhost/
       ⇔titanic_transformed')
      df.to_sql('titanic_transformed', con=engine, if_exists='replace', index=False)
      print("Data loaded successfully!")
     Data loaded successfully!
[37]: df
                                                                    Parch
[37]:
                         Survived Pclass
           PassengerId
                                            Sex
                                                        Age
                                                             SibSp
                                                                                Fare \
      0
                    892
                                 0
                                         3
                                                  34.500000
                                                                         0
                                                                              7.8292
                                              1
                                                                  0
                                         3
      1
                    893
                                 1
                                                 47.000000
                                                                  1
                                                                         0
                                                                              7.0000
                    894
                                 0
                                         2
      2
                                                 62.000000
                                                                  0
                                                                         0
                                                                              9.6875
      3
                    895
                                 0
                                         3
                                                  27.000000
                                                                  0
                                                                         0
                                                                              8.6625
      4
                    896
                                         3
                                                 22.000000
                                                                             12.2875
                                 1
                                              0
                                                                  1
                                                                         1
      . .
      413
                   1305
                                0
                                         3
                                                 24.027945
                                                                  0
                                                                         0
                                                                              8.0500
                                              1
                                                                            108.9000
      414
                   1306
                                         1
                                                 39.000000
                                                                 0
                                                                         0
                                 1
                                              0
                                         3
      415
                   1307
                                 0
                                              1
                                                 38.500000
                                                                 0
                                                                              7.2500
      416
                   1308
                                 0
                                         3
                                              1
                                                  24.027945
                                                                  0
                                                                         0
                                                                              8.0500
      417
                   1309
                                                 24.027945
                                                                  1
                                                                             22.3583
```

${\tt Embarked}$ Q S 0 1 Q S 2 3 S 4 413 S 414 S 415 416 С 417

[418 rows x 9 columns]

[]: