ETL_Data_Pipeline

September 13, 2024

[38]:
```
import pandas as pd
```

[39]:
```
df = pd.read_csv("tested.csv")
df
```

[39]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	0	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893	1	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
2	894	0	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
3	895	0	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
4	896	1	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S
..	…	…	…	…	…	…	…	…	…	…	…	…
413	1305	0	3	Spector, Mr. Woolf	male	NaN	0	0				
414	1306	1	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0				
415	1307	0	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0				
416	1308	0	3	Ware, Mr. Frederick	male	NaN	0	0				
417	1309	0	3	Peter, Master. Michael J	male	NaN	1	1				

```
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]
```

[40]:
```
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
```

[ ]:
```

```

[41]:
```
# age value replace with  according pclass...
df['Age'] = df.groupby('Pclass')['Age'].transform(lambda x: x.fillna(x.mean()))
```

[42]:
```
df
```

[42]:
```
     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  \
0                                    Kelly, Mr. James    male   34.500000      0
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    male   24.027945      0
414                  Oliva y Ocana, Dona. Fermina  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        0              330911    7.8292   NaN        Q
1        0              363272    7.0000   NaN        S
2        0              240276    9.6875   NaN        Q
3        0              315154    8.6625   NaN        S
4        1             3101298   12.2875   NaN        S
..     …                   …       …   …      …
413      0           A.5. 3236    8.0500   NaN        S
414      0            PC 17758  108.9000  C105        C
415      0  SOTON/O.Q. 3101262    7.2500   NaN        S
416      0              359309    8.0500   NaN        S
417      1                2668   22.3583   NaN        C

[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         Age  SibSp  Parch      Fare  \
0            892         0       3    male   34.500000      0      0    7.8292
1            893         1       3  female   47.000000      1      0    7.0000
2            894         0       2    male   62.000000      0      0    9.6875
3            895         0       3    male   27.000000      0      0    8.6625
4            896         1       3  female   22.000000      1      1   12.2875
..           …       …     …      …        …    …      …       …
413         1305         0       3    male   24.027945      0      0    8.0500
414         1306         1       1  female   39.000000      0      0  108.9000
415         1307         0       3    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      1   22.3583
```

```
    Embarked
0          Q
1          S
2          Q
3          S
4          S
..       …
413        S
414        C
415        S
416        S
417        C

[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]:
```
     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

[418 rows x 9 columns]
```

[49]:
```
# Feature Engineering:Create new features based on existing ones (e.g., Age group).


df['Age_group'] = pd.cut(df['Age'], bins=[0, 18, 60, 100], labels=["Child", "Adult", "Senior"])
```

[50]:
```
df
```

[50]:
```
     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 Age_group
0          Q     Adult
1          S     Adult
2          Q    Senior
3          S     Adult
4          S     Adult
..       …       …
413        S     Adult
414        C     Adult
415        S     Adult
416        S     Adult
417        C     Adult

[418 rows x 10 columns]
```

[51]:
```
df.isnull().sum()
```

[51]:
```
PassengerId    0
Survived       0
Pclass         0
Sex            0
```

```
Age            0
SibSp          0
Parch          0
Fare           1
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  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 Age_group
0          Q     Adult
1          S     Adult
2          Q    Senior
3          S     Adult
4          S     Adult
..       …       …
413        S     Adult
414        C     Adult
415        S     Adult
416        S     Adult
417        C     Adult

[418 rows x 10 columns]
```

[55]:
```
df.isnull().sum()
```

[55]:
```
PassengerId    0
Survived       0
Pclass         0
Sex            0
```

```
Age            0
SibSp          0
Parch          0
Fare           0
Embarked       0
Age_group      0
dtype: int64
```

[56]:
```
df.dtypes
```

[56]:
```
PassengerId       int64
Survived          int64
Pclass            int64
Sex               int64
Age             float64
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
```

[37]:

	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

[418 rows x 9 columns]
```

[ ]: