7-11-2024

Training Day – 36

Topic:* Data Cleaning

- Handled missing values and duplicates in a dataset.
- Example: Used fillna() to replace missing values with the column mean. visualization easier and identifying outliers easily.

What is data cleaning? Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled

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

# Load the dataset
df = pd.read_csv('titanic.csv')
df.head()
```

Output:

```
PassengerId    Survived    Pclass    Name    Sex    Age    SibSp    Parch    Ticket    Fare
Cabin    Embarked
0    1    0    3    Braund, Mr. Owen Harris    male    22.0    1    0    A/5 21171
7.2500    NaN    S
1    2    1    1    Cumings, Mrs. John Bradley (Florence Briggs Th...    female    38.0
1    0    PC 17599    71.2833    C85    C
2    3    1    3    Heikkinen, Miss. Laina    female    26.0    0    0    STON/O2. 3101282
7.9250    NaN    S
3    4    1    1    Futrelle, Mrs. Jacques Heath (Lily May Peel)    female    35.0    1
0    113803    53.1000    C123    S
4    5    0    3    Allen, Mr. William Henry    male    35.0    0    0    373450    8.0500
NaN    S
```

jupyter project-of-DA Last Checkpoint: 29 days ago

File Edit View Run Kernel Settings Help

Code

JupyterLab Python 3 (ipykernel)

[8]:
```
df.shape
```

[8]:
```
(36, 6)
```

[9]:
```
(117869/6122893)*100   #find active ration by dived by total
```

[9]:
```
1.9250540553297273
```

[10]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36 entries, 0 to 35
Data columns (total 6 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   State/UTs         36 non-null     object 
 1   Total Cases       36 non-null     int64  
 2   Active            36 non-null     int64  
 3   Deaths            36 non-null     int64  
 4   Active Ratio (%)  36 non-null     float64
 5   Death Ratio (%)   36 non-null     float64
dtypes: float64(2), int64(3), object(1)
memory usage: 1.8+ KB
```

[11]:
```
df.isnull().sum()
```

```
State/UTs           0
Total Cases         0
Active              0
Deaths              0
Active Ratio (%)    0
Death Ratio (%)     0
dtype: int64
```

[12]:
```
df.columns
```