Pandas : Analysis & Manipulation

1. info()
2. describe()
3. head(), tail()
4. columns
5. dtypes
6. sort_values()
7. set_index()
8. value_counts()
9. groupby()
10. pivot_table()
11. cross_tab()
12. Ass()

Data Preprocessing

1. Fixing Rows/Columns
   - Delete
   - Add
   - Rename
   - Datatype
2. Handling Duplicates
3. Missing Values
4. Outliers

emp	Role	Salary

np.where( ) .

np.where(df['Salary'] > 10,000 , 'A' , 'B')

↳ np.where(✓ , ,)

① Handling Duplicates ⟶ identify ⟶ drop ⟶ ② drop_duplicates( )

↓

① duplicated().sum() : total no. of duplicates

df[df.duplicated()] : Display duplicate records

③ Handling missing values → Identify → filling / dropping

① Identify : isnull(), isna().sum()

② Deleting : drop_na() ✓ [when % is above 30%]

③ Filling : fillna( )
- (imputation)
- value → mean / median / mode / value

* Interpolation → filling by using prediction → interpolate( )

Data normal - Mean
Abnormal - Median (Outliers)
Repititive - Mode

outlier → extreme high
extreme low

Age : [18, 50], 2, 3, -18 → X
200, 150, 90 → X