Data Analysis Documentation

Data Cleaning and Preprocessing

Missing data:

Missing data can be of forms like

* **NaN** which is typically formed by use of pandas
* **Null/None** which is typically formed by use of sql databases
* **“” empty string** in csv files
* **Special numbers like -999,9999** in old datasets
* **Spaces /empty** in some datasets

**Types of missing data :**

* MCAR: Missing Completely At Random
* MAR: Missing At Random(Missingness is depends on observed data not on unobserved data )
* MNAR :Missing Not At random (Missingness is depends on unobserved data not on observed data )

**Techniques to handle missing data :**

* Deletion : removing corresponding rows where null values occurred .
* Imputation : replacing null values with estimated values

🡪 **Mean/Median/Mode :** replacing null values with mean/median/ mode which are corresponding to particular column , **But it arises bias when null values spread over dataset randomly**

**🡪 KNN( k- nearest neighbors )**

**🡪Statistical model**

List of Methods to Handle Missing Values in a Dataset

* Deleting the Missing Values
* Imputing the Missing Values
* Imputing the Missing Values for Categorical Features
* Imputing the Missing Values using Sci-kit Learn Library
* Using “Missingness” as a Feature

**Using deletion method :**

df = train\_df.dropna(axis=0)

axis refers to rows(0) or columns (1)

**Using imputing method:**

1. Replacing with some arbitrary value:

Train\_df.fillna(0)

2)replacing with respective column mean :

train\_df.fillna(train\_df.mean())

for particular column : train\_df[‘column’].fillna(train\_df[‘column’].mean())

3)Replacing with mode value :

df.fillna(df.mode())

if having more than one mode value

modes=df.mode()  
 df.fillna(modes.iloc[0])

4) Replacing with medan:

df.fillna(df.median())

5) Replacing with the next value – backward fill

# Backward-Fill

df.fillna(method=‘bfill')

6) Interpolate

df.interpolate()

Missing values can also be imputed using interpolation. Pandas’ interpolate method can be used to replace the missing values with different interpolation methods like ‘polynomial,’ ‘linear,’ and ‘quadratic.’ The default method is ‘linear.’

**Handing categorical data :**

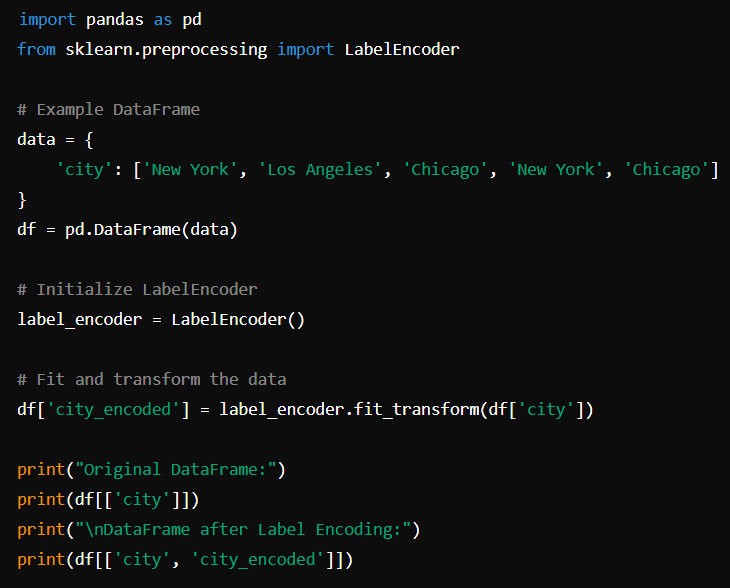
Types of categorical data :

1. Nominal data : having no inheriting relation (eg: female ,male )
2. Ordinal data : having relation like ranking(eg: high,low,medium)

Techniques to convert categorical data to numerical :

1)Label encoding :

Where it will assign unique integer for every category within single column



2)One -hot encoding :

It creates separate columns for every category within each column

