Name - Sajjad ali dhuniya

USN - 21BTRCD064

Department - CSE (Data Science)

In [1]:
```
import pandas as pd
```

In [2]:
```
# Load the dataset
house_df = pd.read_excel('my_house.xlsx')
```

Using dropna()

In [3]:
```
house_df = house_df.dropna()
```

Fill missing values with mean, median or mode

In [4]:
```
median = house_df['Number of windows'].median()
house_df['Number of windows'] = house_df['Number of windows'].fillna(median)
```

In [5]:
```
house_df
```

Out[5]:

	Room	Size (in square feet)	Number of windows	Flooring type
0	Kitchen	150	2.0	Tile
1	Living room	250	3.0	Carpet
3	Bedroom 2	175	2.0	Carpet
4	Bathroom	75	1.0	Tile

Encoding Techniques

Replace categorical values with numerical values We can replace the categorical values in the 'Flooring type' column with numerical values using the replace method:

In [7]:
```
house_df['Flooring type'] = house_df['Flooring type'].replace({'Tile': 0, 'C
```

In [8]:
```
house_df
```

Out[8]:

	Room	Size (in square feet)	Number of windows	Flooring type
0	Kitchen	150	2.0	0
1	Living room	250	3.0	1
3	Bedroom 2	175	2.0	1
4	Bathroom	75	1.0	0

One-hot encoding

We can use one-hot encoding to convert the 'Flooring type' column into binary columns for each category using the get_dummies method:

In [9]:
```
house_df = pd.get_dummies(house_df, columns=['Flooring type'])
```

In [10]:
```
house_df
```

Out[10]:

	Room	Size (in square feet)	Number of windows	Flooring type_0	Flooring type_1
0	Kitchen	150	2.0	1	0
1	Living room	250	3.0	0	1
3	Bedroom 2	175	2.0	0	1
4	Bathroom	75	1.0	1	0

Sklearn-ordinalencoder

We can use the OrdinalEncoder from scikit-learn to encode the 'Flooring type' column with numerical values:

In [13]:
```
from sklearn.preprocessing import OrdinalEncoder
```

In [16]:
```
encoder = OrdinalEncoder()
house_df['Flooring type'] = encoder.fit_transform(house_df[['Flooring type_0
```

In [17]:
```
house_df
```

Out[17]:

	Room	Size (in square feet)	Number of windows	Flooring type_0	Flooring type_1	Flooring type
0	Kitchen	150	2.0	1	0	1.0
1	Living room	250	3.0	0	1	0.0
3	Bedroom 2	175	2.0	0	1	0.0
4	Bathroom	75	1.0	1	0	1.0

Write short description of encoding & its methods.

In machine learning, encoding is the process of transforming categorical data into numerical data that can be used by algorithms to build predictive models. This is necessary because many machine learning algorithms can only handle numerical data, and cannot directly work with categorical data.

There are several methods of encoding categorical data, including:

1) Label Encoding: This method assigns each unique value in a categorical column with a unique integer. For example, if a column has values 'red', 'green', and 'blue', these might be encoded as 0, 1, and 2. This method is suitable for ordinal categorical data where there is a natural ordering between the categories.

2) One-Hot Encoding: This method creates a new binary column for each unique value in a categorical column. If a row has a certain value for a categorical column, the corresponding binary column will have a value of 1, and all other binary columns will have a value of 0. This method is suitable for nominal categorical data where there is no natural ordering between the categories.

3) Binary Encoding: This method converts each unique value in a categorical column to a binary code. For example, if a column has values 'red', 'green', and 'blue', these might be encoded as 00, 01, and 10. This method is suitable for categorical data with many unique values.

4) Target Encoding: This method replaces each unique value in a categorical column with the mean of the target variable for that value. This method is suitable for categorical data where the target variable is continuous.

5) Frequency Encoding: This method replaces each unique value in a categorical column with its frequency in the dataset. This method is suitable for categorical data where the frequency of the value is informative.

6) Ordinal Encoding: This method assigns each unique value in a categorical column with a numerical value based on its rank or order. This method is suitable for categorical data where there is a natural ordering between the categories, but the categories cannot be assumed to have an equal distance between them.

The choice of encoding method depends on the nature of the categorical data and the specific requirements of the machine learning algorithm being used.