

Mini Project: Cohort 21: Affordable Housing Price Estimator

Date: 27.02.2026

Today, our group had a discussion regarding the progress made by the data team. The data team explained the steps they followed in data preprocessing and data cleaning. They walked us through how the raw dataset was handled, including removing missing values, correcting inconsistent data, handling outliers, and formatting the data properly for analysis.

We understood how important preprocessing is before building a model, as clean and structured data helps improve model accuracy and performance. The team also explained the tools and techniques they used during the preprocessing stage.

After understanding the cleaned and preprocessed dataset, we discussed the next phase of the project, which is model building. We planned to start selecting suitable machine learning algorithms and begin training the model using the prepared dataset. Roles and responsibilities for the model development phase were also briefly discussed.

Overall, the meeting was productive and helped all team members clearly understand the preprocessing workflow and the upcoming steps for model implementation.

Group Discussion: Day 2



Data Preprocessing:

In this phase, we prepared the Bengaluru Housing dataset for Machine Learning model building. First, we imported essential Python libraries such as Pandas and NumPy to handle data manipulation and numerical operations. The dataset was loaded using the `read_csv()` function and stored as a DataFrame.

During data exploration, we identified missing values and inconsistencies. The society column had more than 40% missing values, so it was removed to improve data quality. The total_sqft column contained range values (e.g., "2100-2850"), which were converted into single numeric values by calculating their average.

Missing values were handled carefully:

- Bath column – filled using median (to avoid impact of outliers)
- Balcony column – filled using mode (most frequent value)
- Size column – filled using mode (most common property size)

We also removed unrealistic outliers such as properties with more than 10 bathrooms or more than 10 BHK, as these were likely data entry errors.

Finally, the cleaned dataset was saved as a new CSV file for further steps like Exploratory Data Analysis and Model Building. After preprocessing, the dataset became clean, consistent, and ready for building the Affordable Housing Price Estimator model under SDG 11 – Sustainable Cities and Communities.

Code Snippets:

```
import pandas as pd
import numpy as np
✓ 0.0s

df = pd.read_csv("Bengaluru_House_Data_Final.csv")
print("Shape:", df.shape)
df.head()

print("Null Values:")
print(df.isnull().sum())

# Reload fresh dataset
import pandas as pd
import numpy as np

df = pd.read_csv("Bengaluru_House_Data_Final.csv")
print("Dataset loaded fresh!")
print("Shape:", df.shape)
print("Columns:", df.columns.tolist())
```

```
# 1. Drop society (41% missing)
df = df.drop('society', axis=1)

# 2. Drop null location
df = df.dropna(subset=['location'])
df['location'] = df['location'].astype(str).str.strip()

# 3. Fix total_sqft ranges like "2100-2850"
def convert_sqft(val):
    val = str(val).strip()
    if '-' in val:
        try:
            parts = val.split('-')
            return (float(parts[0]) + float(parts[1])) / 2
        except:
            return np.nan
    try:
        return float(val)
    except:
        return np.nan

df['total_sqft'] = df['total_sqft'].apply(convert_sqft)
df = df.dropna(subset=['total_sqft'])

# 4. Fill missing values
df['bath'] = df['bath'].fillna(df['bath'].median())
```

```

df['bath'] = df['bath'].fillna(df['bath'].median())
df['balcony'] = df['balcony'].fillna(df['balcony'].mode()[0])
df['size'] = df['size'].fillna(df['size'].mode()[0])

# 5. Remove bath outliers (>10)
df = df[df['bath'] <= 10]

# 6. Remove size outliers (>10 BHK)
def get_bhk(size):
    try:
        return int(str(size).split()[0])
    except:
        return np.nan

df['bhk'] = df['size'].apply(get_bhk)
df = df[df['bhk'] <= 10]
df = df.drop('bhk', axis=1)

print("Data Cleaning Done!")
print("Final Shape:", df.shape)

print("== SHAPE ==")
print(df.shape)

```

```

print("== SHAPE ==")
print(df.shape)

print("\n== COLUMNS ==")
print(df.columns.tolist())

print("\n== NULL VALUES ==")
print(df.isnull().sum())

print("\n== SAMPLE DATA ==")
df.head(10)

# Save cleaned dataset
df.to_csv("Bengaluru_House_Cleaned.csv", index=False)
print("Cleaned dataset saved as Bengaluru_House_Cleaned.csv")

```

Processed Dataset:

A	B	C	D	E	F	G	H	I	J	K	L
1	area_type	availability	location	size	total_sqft	bath	balcony	price	Distance_to_nearest_MRT_station_m		
2	Super built	19-Dec	Electronic	2 BHK	1056	2	1	39.07	1249.45		
3	Plot Area	Ready To N	Chikka Tir	4 BHK	2600	5	3	120	3901.43		
4	Built-up	AI	Ready To N	Uttarahalli	1440	2	3	62	1678.39		
5	Super built	Ready To N	Lingadhee	3 BHK	1521	3	1	95	3197.32		
6	Super built	Ready To N	Kothanur	2 BHK	1200	2	1	51	2312.04		
7	Super built	Ready To N	Whitefield	2 BHK	1170	2	1	38	987.19		
8	Super built	18-May	Old Airport	4 BHK	2732	4	2	204	869.7		
9	Super built	Ready To N	Rajaji Nagar	4 BHK	3300	4	2	600	733.09		
10	Super built	Ready To N	Marathahalli	3 BHK	1310	3	1	63.25	1521.34		
11	Plot Area	Ready To N	Gandhi Ba	6 BHK	1020	6	2	370	3416.15		
12	Super built	18-Feb	Whitefield	3 BHK	1800	2	2	70	824.7		
13	Plot Area	Ready To N	Whitefield	4 BHK	2785	5	3	295	1963.89		
14	Super built	Ready To N	7th Phase	2 BHK	1000	2	1	38	3664.89		
15	Built-up	AI	Ready To N	Gottigere	1100	2	2	40	1054.81		
16	Plot Area	Ready To N	Sarjapur	3 BHK	2250	3	2	148	1018.19		
17	Super built	Ready To N	Mysore Ro	2 BHK	1175	2	2	73.5	1020.09		
18	Super built	Ready To N	Bisuvanh	3 BHK	1180	3	2	48	2608.48		

A	B	C	D	E	F	G	H	I	J
10878	Plot Area	Ready To N	Mirumetta	2 BHK	1100	1	1	200	4385.73
10879	Super built	18-Dec	Hebbal	3 BHK	1255	3	2	77.68	830.97
10880	Super built	Ready To N	Thanisand	2 BHK	1140	2	1	59	824.34
10881	Super built	Ready To N	Hosakerel	3 BHK	2376	3	1	240	7356.21
10882	Super built	19-Mar	Thanisand	1 BHK	663	1	1	46	1594.01
10883	Super built	Ready To N	Victoria La	2 BHK	1276	2	0	146	5177.29
10884	Super built	Ready To N	Kanakpura	3 BHK	1450	3	3	60.9	1184.01
10885	Super built	Ready To N	Kodichikk	3 BHK	1400	3	3	66	5376.46
10886	Super built	17-Oct	Divya Unn	2 BHK	1339	2	1	53.56	7761.11
10887	Built-up	AI	18-Dec	Electronic	1220	2	1	35.23	869.76
10888	Built-up	AI	Ready To N	Malleshwara	2200	3	2	275	9503.81
10889	Built-up	AI	Ready To N	Harohalli	1305	2	2	52	4557.65
10890	Built-up	AI	Ready To N	Green Gle	1740	3	3	80	9997.71
10891	Super built	Ready To N	Anekal	2 BHK	680	1	1	21	4514.57
10892	Super built	Ready To N	Bannergha	2 BHK	1154	2	2	47	1725.86
10893	Built-up	AI	Ready To N	Munnekoll	1200	2	2	40	4563.89
10894	Super built	Ready To N	Hennur	2 BHK	1155	2	2	54.5	897.02
10895	Super built	Ready To N	Sarjapur	3 BHK	1525	2	3	63	1849.92
10896	Built-up	AI	Ready To N	Giri Nasar	600	1	1	125	9011.26

