

Practical Machine Learning and Data Exploration
First Trimester 2025

Project 1

Housing Dataset

ID : 243002097

Name : Ahmad Mostafa Abutalib

Data Preprocessing - Housing Dataset

Step 1: Import Libraries

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

Step 2: Load the Dataset

Load the Housing dataset

```
housing_df = pd.read_csv('Housing.csv')
```

Preview the first 5 rows

```
housing_df.head()
```

Step 3: Handling Missing Values

Get the count of missing values

```
missing_values = housing_df.isnull().sum()
```

Display only columns with missing values

```
print("Missing values in each column:")
```

```
print(missing_values[missing_values > 0])
```

```
Series([], dtype: int64)
```

Step 4: Handling Outliers (Visualizing Outliers)

Step 1: Import Libraries

```
import pandas as pd
```

```
import numpy as np
```

Step 2: Load the Housing dataset

```
housing_df = pd.read_csv('Housing.csv')
```

Step 3: Define a function to replace outliers with NaN using IQR method

```
def replace_outliers_with_nan(column):
```

```
    Q1 = column.quantile(0.25)
```

```
    Q3 = column.quantile(0.75)
```

```
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
return column.where((column >= lower_bound) & (column <=
upper_bound), np.nan)
```

Step 4: Apply the function to numerical columns

```
for col in housing_df.select_dtypes(include=[np.number]).columns:
    housing_df[col] = replace_outliers_with_nan(housing_df[col])
```

