Capstone Project (Outlier Detection and Removal)
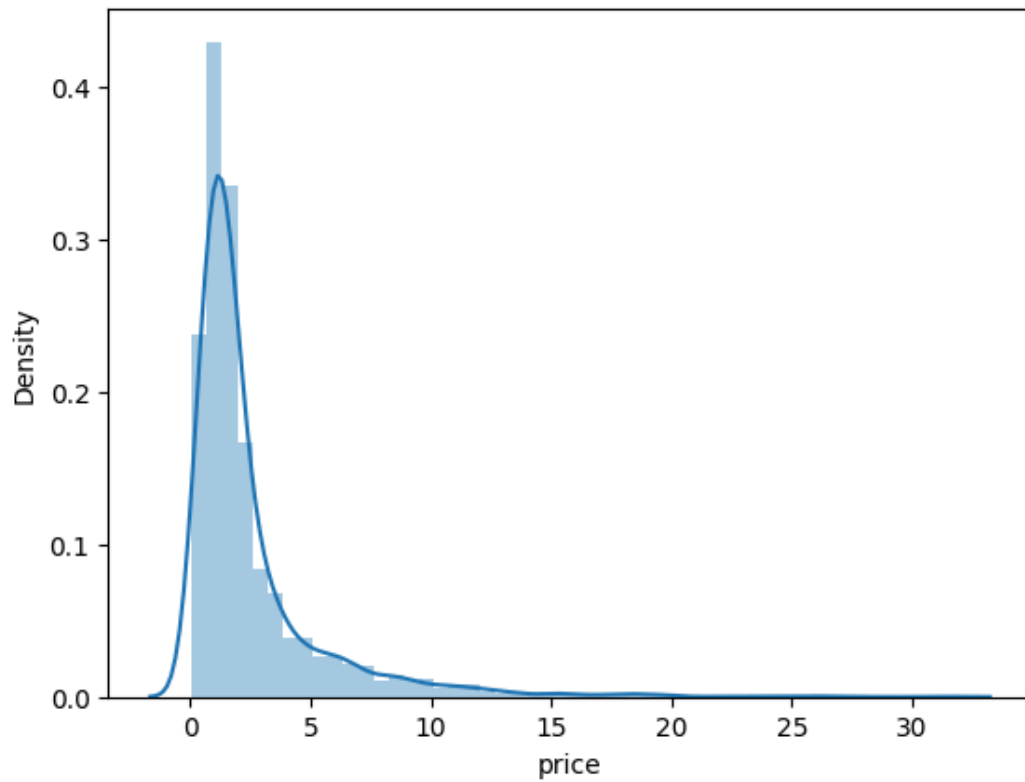
```
df = pd.read_csv('gurgaon_properties_cleaned_v2.csv').drop_duplicates()
```

```
df.shape

Output: (3677, 23)
```

```
# outliers on the basis of price column
sns.distplot(df['price']);
```
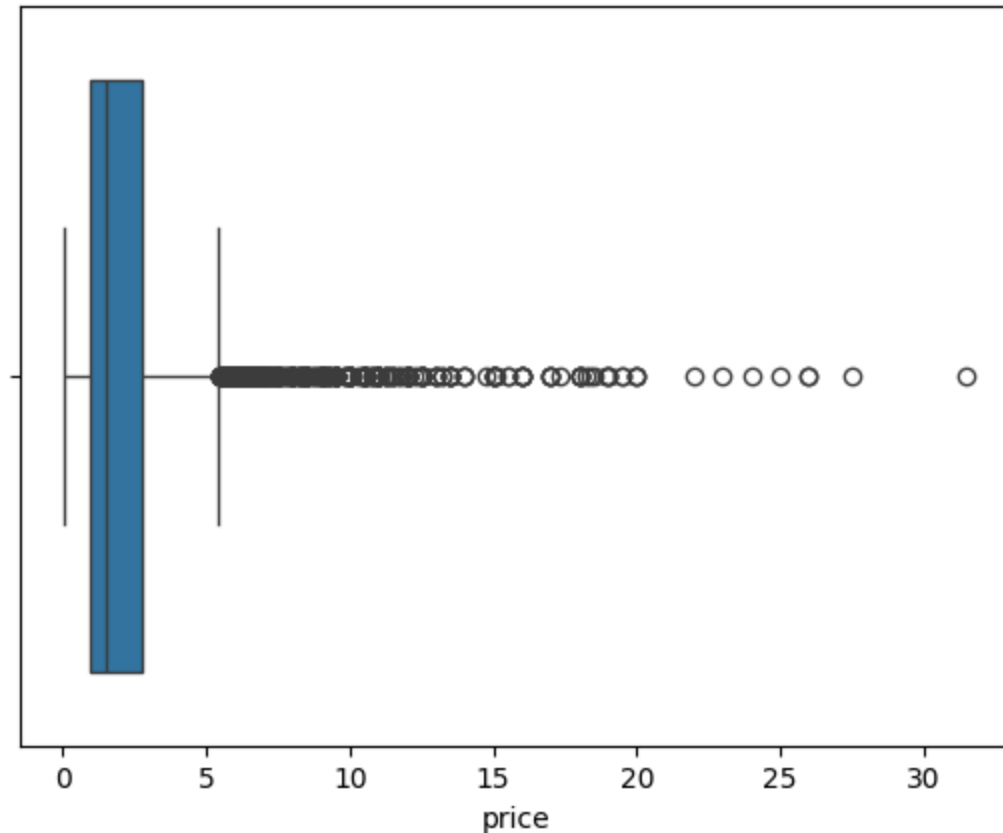
> 💡 `distplot` will be deprecated.
> Use:
> `displot` or `histplot`

```
sns.boxplot(x=df['price'])
```

Calculate the IQR for the 'price' column

```python
# Calculate the IQR for the 'price' column
Q1 = df['price'].quantile(0.25)
Q3 = df['price'].quantile(0.75)
IQR = Q3 - Q1

# Define bounds for outliers
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Identify outliers
outliers = df[(df['price'] < lower_bound) | (df['price'] > upper_bound)]

# Displaying the number of outliers and some statistics
num_outliers = outliers.shape[0]
```

```
outliers_price_stats = outliers['price'].describe()

num_outliers, outliers_price_stats
```

```
(425,
 count    425.000000
 mean       9.235624
 std        4.065259
 min        5.460000
 25%        6.460000
 50%        8.000000
 75%       10.750000
 max       31.500000
 Name: price, dtype: float64)
```

- 425 outliers

```
outliers.sort_values('price',ascending=False).head(20)
```

	property_type	society	sector	price	price_per_sqft	area	areaWithType	bedRoom	b
586	house	arjun marg/ sector- 26 phase- 1/ golf course road	sector 26	31.50	35000.0	9000.0	Plot area 1000(836.13 sq.m.)	7	
744	house	independent	sector 43	27.50	24366.0	11286.0	Plot area 1254(1048.5 sq.m.)	6	
260	house	independent	sector 26	26.00	82540.0	3150.0	Plot area 402(336.12 sq.m.)Built Up area: 400 ...	16	
1597	house	dlf city plots	sector 26	26.00	57206.0	4545.0	Plot area 505(422.24 sq.m.)	6	
374	house	suncity township	sector 54	25.00	31111.0	8036.0	Plot area 1000(836.13 sq.m.)	4	
1012	house	emaar the palm springs	sector 54	24.00	600000.0	400.0	Plot area 400(37.16 sq.m.)	5	
298	house	independent	sector 110	23.00	25556.0	9000.0	Plot area 1000(836.13 sq.m.)	4	

Did this for all columns.