House Pricing in Egypt - DOCUMENTATION

1. **Dataset Description:**

   This dataset has been collected from users listings on **olx.com**, the dataset was last updated in **2022**, having approximately **27,000 rows** in total and **11 columns** listed as follows:

   1. **Type:** Type of property.
   2. **Price:** Price of property.
   3. **Bedrooms:** Total number of bedrooms.
   4. **Bathrooms:** Total number of bathrooms.
   5. **Area:** Area of the property *(measured by m^2)*.
   6. **Furnished:** Indicates whether the property is furnished or not.
   7. **Level:** The floor the property is.
   8. **Compound:** ***(Removed)***
   9. **Payment_Option:** Indicates available payment options (Cash, installments, etc..)
   10. **Delivery_Date:** ***(Removed)***
   11. **Delivery_Term:** ***(Removed)***
   12. **City:** The city where the property is located.

2. **Missing Data:**

   a. **Price:** 2
   b. **Bedrooms:** 203
   c. **Bathrooms:** 171
   d. **Area:** 471

   Total: Approx. 850

3. **Data Preprocessing / Cleaning Procedures:**

A. Data Cleaning:

1. "**Compound**", "**Delivery_Term**" and "**Delivery_Date**" columns were dropped.
2. Renamed cities that appeared **10 times** or less to "**Other**".
3. Dropped rows where "**Price**" is unknown.
4. Dropped "**Studio**" and "**Penthouse**" because they're uncommon, therefore irrelevant.
5. Dropped levels above 10th floor.
6. Dropped unknown values in the "**Level**" column in "**Chalet**".
7. Changed all unknown levels to "**-1**"
8. Changed all unknown values in the "**Furnished**" column with "**No**"

B. Data Preprocessing:

1. Worked on data consistency *(Eg. Dropped rows where **Area / Bedrooms < 7**)*.
2. Created "**Price per area**" column to remove outliers, using mean -+ std..
3. Removed outliers using Z-Score ***(Normalization).***

4. Label encoder, Ordinal encoder and Replace to transform string to numerical values.

5. Usage of "**Log(Price)**" instead of "**Price**" to solve the **Skewness** problem.

4. Shapes:

X-Train: (11836, 8)
X-Test: (5073, 8)
Y-Train: (11836,)
Y-Test: (5073,)

5. Used Algorithms:

	Linear Regression	KNN Regression	Polynomial
Accuracy (Train Set)	70.82249056120516	0.8645354018651873	75.58011120912857
Accuracy (Test Set)	71.21386202003794	0.8055429670358469	76.12201449381041
MSE	0.20294970573761895	0.13639029163266328	0.1694061206298791
MAE	0.36280646436842406	0.2680564721195458	0.3334219093157052
R-Squared	0.7094837375092709	0.8055429670358469	0.7575003480316774