Project Summary: EDA & Data Cleaning for House Prices

This project involved data cleaning and exploratory data analysis (EDA) on a real estate dataset to prepare it for predictive modeling. The dataset contained multiple attributes, such as house price, lot size, square footage, and building age, which required processing before use in machine learning models. The primary tasks included handling missing values, detecting and removing outliers, transforming data, and generating insights using visualizations. After cleaning, the dataset became structured, consistent, and ready for feature engineering and model development.

Key Actions & Technologies Used:

- Utilized Python & Pandas to construct a structured DataFrame for data processing.
- Handled missing values using median imputation for numerical features and zero-filling for categorical attributes.
- Applied IQR-based outlier detection to remove extreme values affecting data distribution.
- Transformed skewed data using log normalization for better model performance.
- Generated key visualizations such as histograms, boxplots, scatterplots, and heatmaps to analyze trends.
- Conducted correlation analysis to identify relationships between numerical variables.
- Exported a cleaned dataset for further feature engineering and machine learning modeling.

References

1) Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques.* Elsevier.
2) Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
3) Hawkins, D. M. (1980). *Identification of Outliers.* Springer.
4) García, S., Luengo, J., & Herrera, F. (2015). *Data Preprocessing in Data Mining.* Springer.
5) Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow.* O'Reilly Media.
6) Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research.
7) Seabold, S., & Perktold, J. (2010). *Statsmodels: Econometric and Statistical Modeling with Python.*