|  |  |
| --- | --- |
| DIGICROME ACADEMY  PROJECT REPORT | **Problem Statement**  Predicting the selling price of residential properties is a complex task influenced by numerous factors such as location, size, number of rooms, and year built. Real estate buyers, sellers, and developers often rely on human intuition or past trends to estimate property value, which can be inaccurate or biased.  Mr. Aniket Gajanan Tayade  Jr. Data Scientist |

**1. Introduction to the Project**

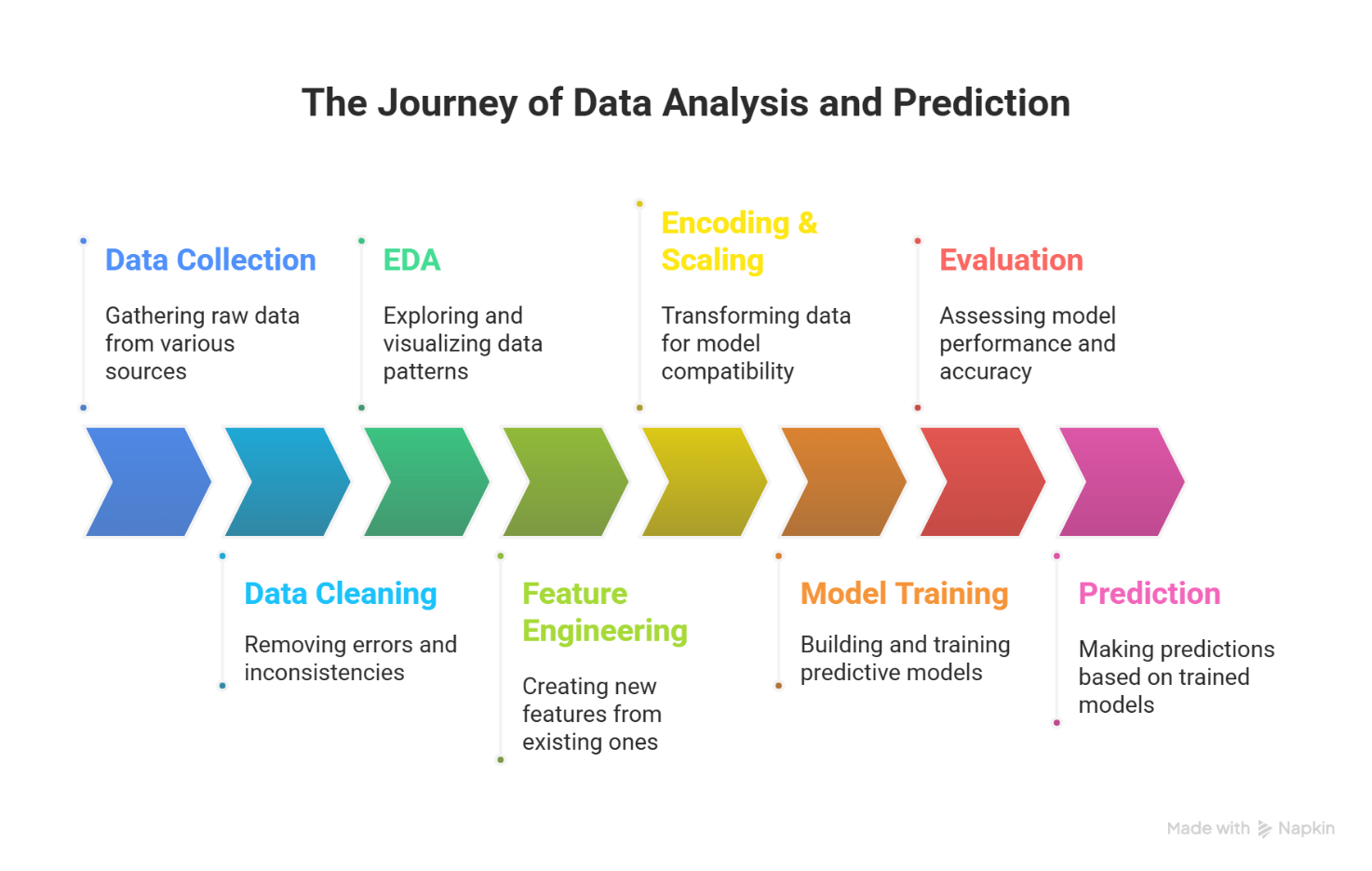
This project is about predicting house prices using machine learning. Normally, people guess prices based on experience or past trends. But that can be wrong. So, we used real data and machine learning to make better predictions. This can help buyers, sellers, and property developers make smarter decisions.

**2. Objectives of the Project**

* Clean and prepare property data
* Understand data using visual analysis (EDA)
* Create new useful features from existing ones
* Train and test different machine learning models
* Choose the best model (we used Gradient Boosting in the end)
* Make predictions and save the model for future use

**3. Flow Chart of the Project**

Data Collection → Data Cleaning → EDA → Feature Engineering → Encoding & Scaling → Model Training → Evaluation → Prediction



**4. Python Libraries Used**

* Pandas, NumPy → For data handling
* Matplotlib, Seaborn → For graphs and charts
* Scikit-learn → For model training and evaluation
* XGBoost → For advanced boosting model (used in testing)
* Joblib → For saving and loading the model

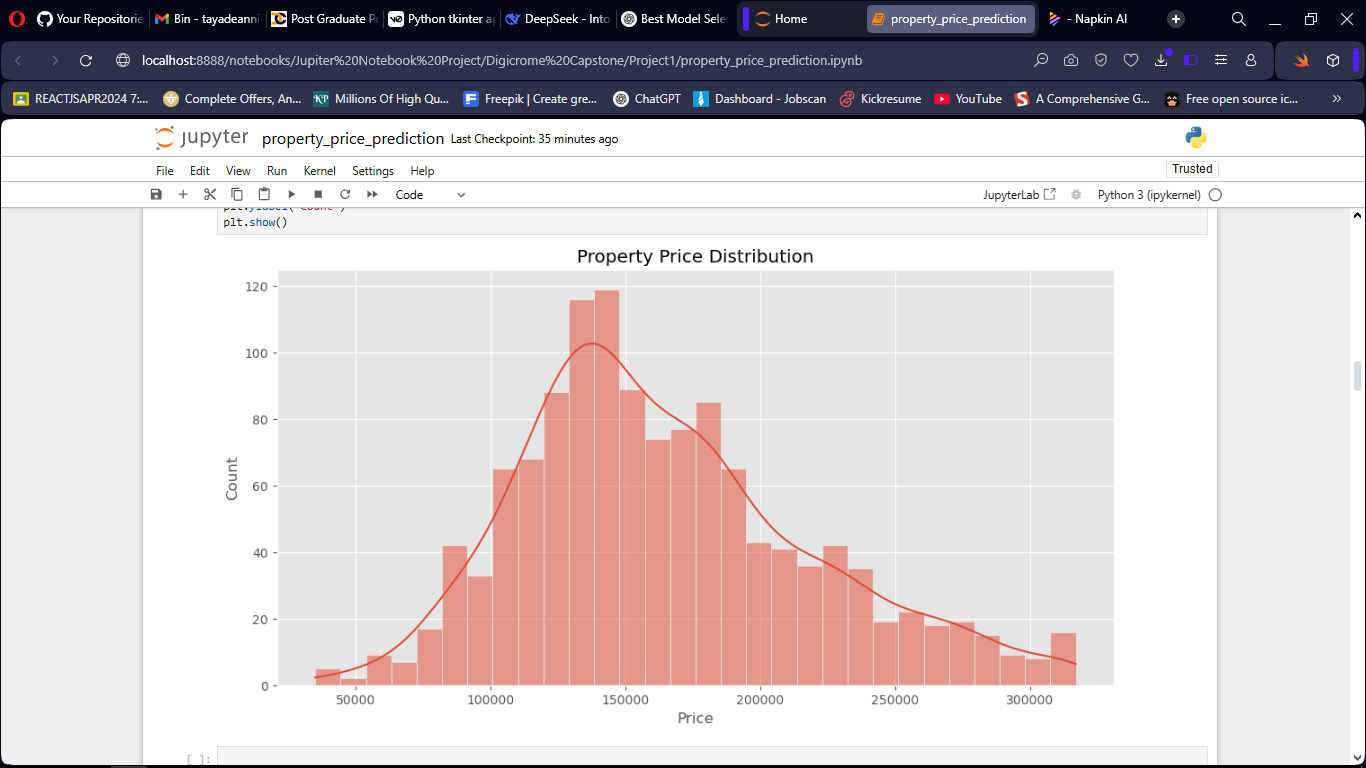
**5. What the Code Does**

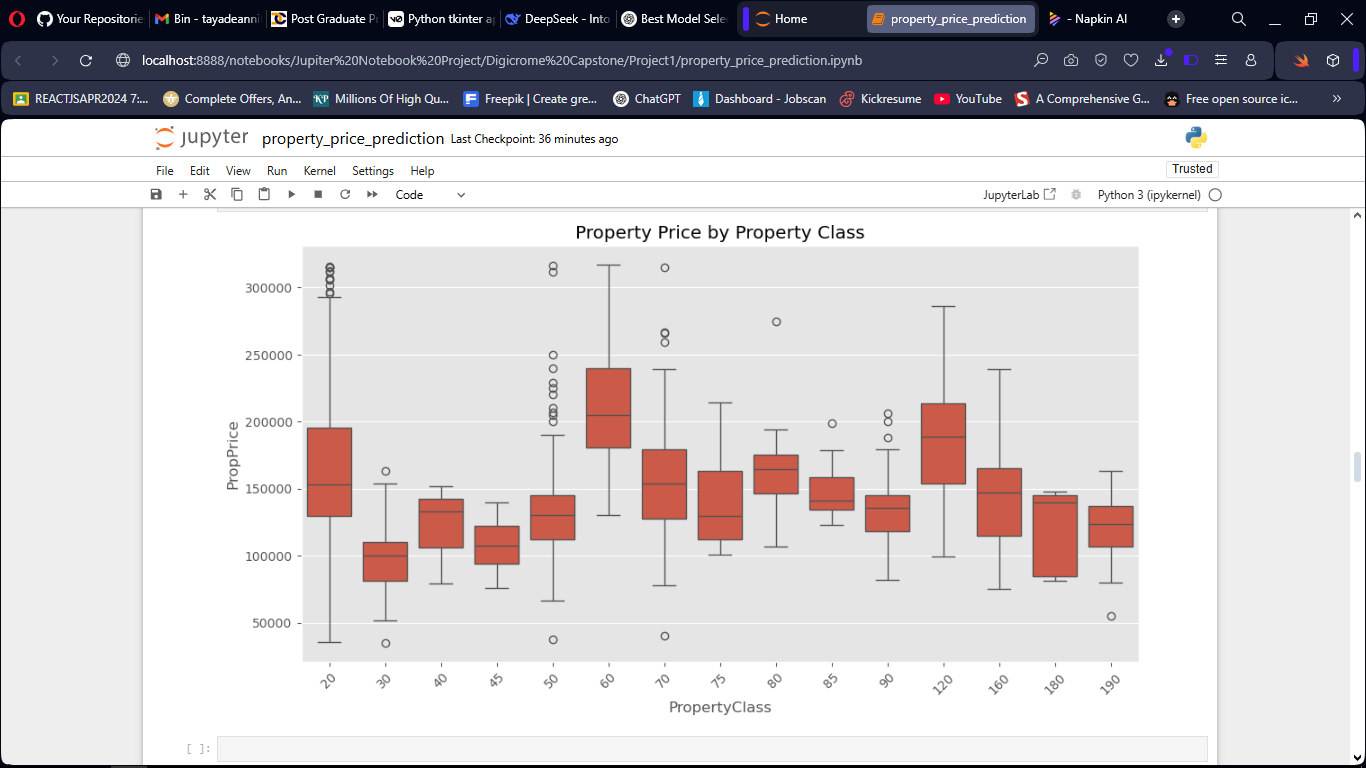
* Loads the dataset
* Fills missing values and removes outliers
* Creates new features like `TotalSF`, `Age`, etc.
* Applies log transformation to fix skewed price data
* Prepares data using pipelines (scaling & encoding)
* Trains Gradient Boosting model with hyperparameter tuning
* Saves the final model
* Makes predictions using new property data

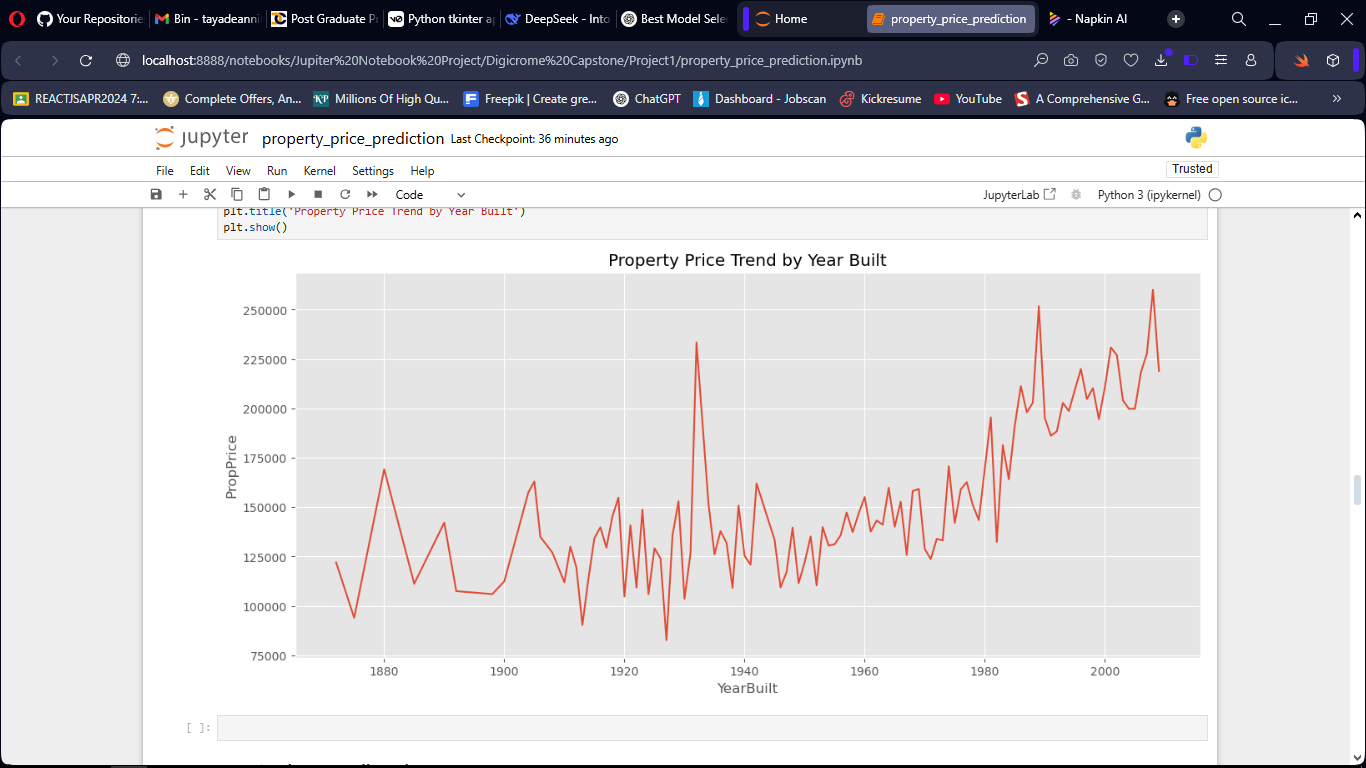
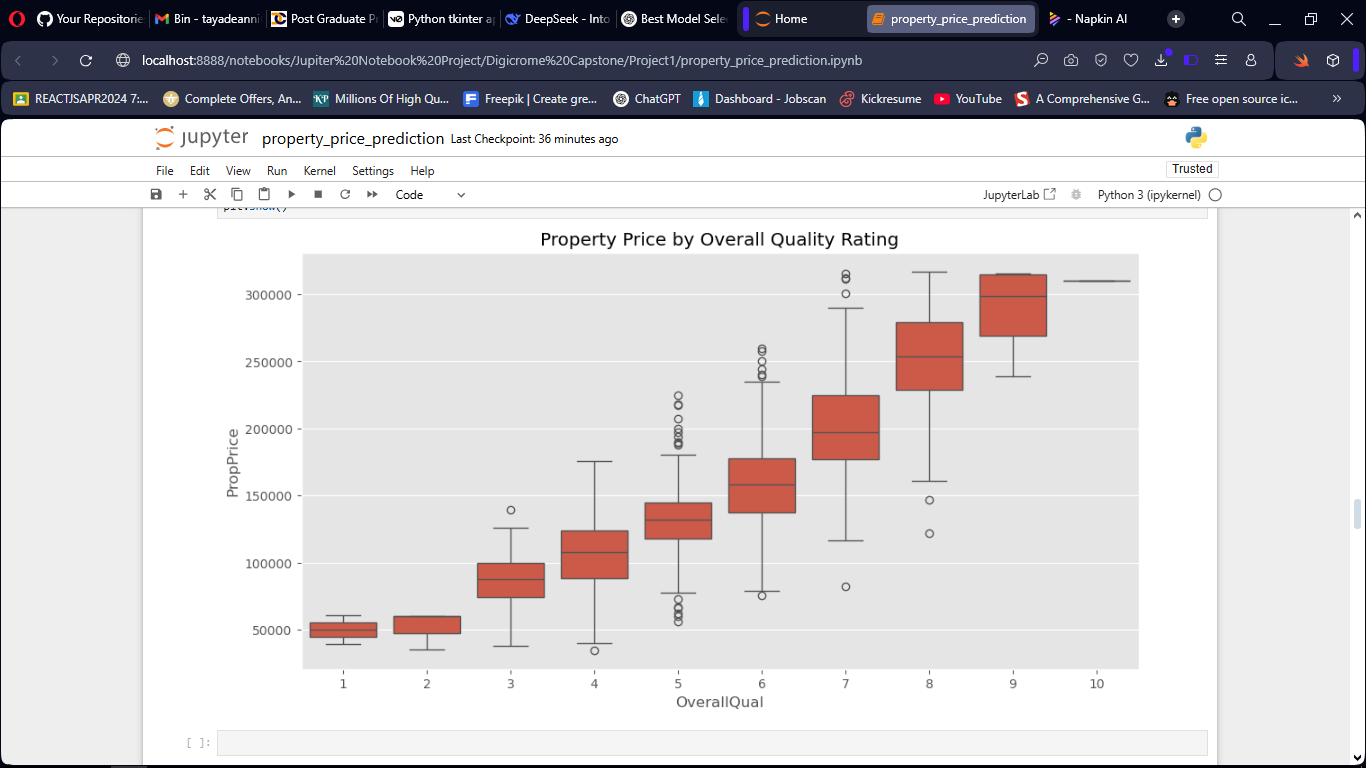
**6. Exploratory Data Analysis (EDA) Report**

From EDA, we found:

* Property price data was right-skewed (fixed using log)
* Bigger houses (more SF) usually had higher prices
* Some features like `OverallQual` had strong correlation with price
* Some outliers were removed from features like `GrLivArea` and `PropPrice`
* Created bar plots, scatter plots, heatmaps, etc. to visualize patterns







**7. Learning Outcomes**

We learned:

* How to clean and prepare data properly
* Importance of handling missing values and outliers
* Creating new features to improve model accuracy
* Using pipelines to preprocess data
* Comparing ML models and tuning them
* How to save and reuse models
* Writing a clean and complete ML workflow

**8. Conclusion**

This project built a machine learning model that can predict house prices accurately. We used real data and Gradient Boosting to get good results. Instead of guessing, this model gives more reliable estimates. In the future, we can improve it by adding more features and maybe even turning it into a web app.

**9. References**

*E-Books:*

[Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow](https://bcebakhtiyarpur.ac.in/wp-content/uploads/2020/03/file_5e748501c810d.pdf)

[Python for Data Analysis](https://nibmehub.com/opac-service/pdf/read/Python%20for%20Data%20Analysis%20_%20data%20wrangling%20with%20Pandas-%20NumPy-%20and%20IPython.pdf)

*Websites:*

[Scikit-learn](https://scikit-learn.org/)

[Pandas](https://pandas.pydata.org/)

[Matplotlib](https://matplotlib.org/)

[Towards Data Science Articles](https://towardsdatascience.com/explainable-generic-ml-pipeline-with-mlflow-2494ca1b3f96/)