Reflective Summary: Property Price Prediction Using PySpark

In this project, I explored how to predict property prices using Linear Regression models in PySpark. The main goal was to test how well different combinations of property-related features could predict price, and to compare how the models performed using metrics like R^2, MSE, and RMSE. Here's a summary of my approach and what I learned.

I carefully selected five features from the dataset that I believed would have the strongest impact on property prices:

Square_Footage,Num_Bedrooms,Num_Bathrooms,Year_Built,Lot_Size

These features were chosen based on general real estate knowledge—larger, newer properties with more rooms typically cost more. I used these in different combinations to build multiple models, which helped me see how each feature contributed to the overall prediction accuracy.

Comparing the Models

I built three different models:One with just two basic features, Another with four features And a final model using all five features

As expected, the model with all five features performed the best, achieving the highest R^2 score. This showed me that including more relevant information helped the model make better predictions. However, I also learned that adding features blindly can sometimes cause overfitting or noise, so it's important to be selective.

Challenges and How I Solved Them

One issue I ran into was missing or messy data. Some rows had null values, which caused errors in the pipeline. I solved this by cleaning the dataset using. dropna() before training the models. Another challenge was understanding how to connect all the pieces in a PySpark ML pipeline—like assembling features, scaling them, and applying the regression model. Once I understood the order and purpose of each step, it became much easier to manage and test multiple models.

What I Learned

This exercise taught me the importance of thoughtful feature selection and the benefits of building reusable pipelines in PySpark. I also gained experience using evaluation metrics to compare models fairly. One key takeaway was how even small changes in the feature set could lead to noticeable differences in model performance. Overall, this project helped me strengthen my practical skills in machine learning and data preparation, which will be valuable in future projects.