Machine learning assignment 1

This report documents my work on Assignment 1: predicting house sale prices using the Ames Housing dataset. I outline the end-to-end process—data cleaning, visualization, and feature engineering—followed by training and evaluating three regression approaches (Linear Regression, Polynomial Features + Linear Regression, and an ensemble model). I then answer the required questions about which features matter, why, what the visualizations reveal, how I ensured an unbiased test set, possible data biases, and how results could be improved without changing the model.

Which features are important for house price prediction?
The most important drivers of SalePrice in my results are overall quality and home size. Specifically, OverallQual is the strongest single predictor, followed by GrLivArea. After these come TotalBsmtSF (basement size), GarageCars ,FullBath ,YearBuilt and YearRemod/Add, plus LotArea and GarageArea with smaller but consistent effects. These priorities are the same whether you look at simple correlations/plots or the feature importances from the tree models.

How did you decide they're important?

I judged feature importance by combining exploratory analysis and model-based evidence. First, I computed a correlation matrix with SalePrice and inspected scatter/box plots (e.g., quality vs. price, living area vs. price) to confirm strong, monotonic relationships and spot outliers. Then I trained RandomForest and GradientBoosting on the cleaned, one-hot-encoded data (with ID columns removed) and used their feature importances to rank predictors. Both approaches consistently elevated OverallQual, GrLivArea, TotalBsmtSF, GarageCars, FullBath, YearBuilt/YearRemodAdd, plus LotArea/GarageArea. The agreement between correlations/plots and tree-based importances—and their stability across runs—was my basis for deciding these features are most important.What did I learn from visualization?

I learned that salesprice is right skewed. Also these is a clear upward trends on higher quaty and larger living/basement/garage area that brings to higher price. More bathroom and garage capacity associate with higher prices. Also the diagnostic for the best model GBR, shows under prediction on very expensive homes and larger residuas at higher prices.

Was there a need to create an unbiased test set? If yes, how?
Yes. I created an unbiased test set with train_test_split(test_size=0.2, random_state=42), holding out 20% that the models never saw during training. To avoid leakage, preprocessing

(imputation/encoding) should be fit on the training data and then applied to the test set only, and the fixed random_state makes the split reproducible.

Any biases in the data fed to the model?

Yes. The data show several sources of bias: (1) Right-skewed prices—a few luxury homes stretch the tail and can distort errors. (2) Imbalanced geography/time—some neighborhoods and build years are over-represented, so the model may favor common neighborhoods/periods and underfit rare ones. (3) Imputation bias—filling missing values with the median or "None" can wash out rare patterns. (4) Potential leakage from identifiers (e.g., PID/Order) was removed to avoid inflating performance. (5) Heteroscedasticity—residuals grow with price, so high-end homes have larger errors.

How to get better results without changing the model?

Improve results by cleaning and enriching the data, not swapping algorithms: train linear/polynomial on log1p(SalePrice) and invert with expm1 to stabilize variance; cap/winsorize outliers (very large living area/price); add simple features like *TotalBaths* and *HouseAge* (or age since remodel); log-transform skewed numerics (e.g., lot/basement areas) and standardize for linear/polynomial so scales don't dominate; and fit imputation/encoding on the training set only before applying to test to avoid leakage.

Group work

I worked alone. I divided tasks into data preparation/visualization and modeling/evaluation, reviewed the code and figures, nd finalized on final feature choices and conclusions.