Predicting house prices accurately is crucial for many people, including homebuyers, sellers, and real estate agents. It helps in making informed decisions. The goal of this project is to build a machine learning model that can predict the price of a house based on various features like location, size, number of rooms, and amenities.

**Dataset Description:**

We used a dataset from Kaggle that contains information about over 2,000 houses. The dataset includes features such as location, number of bedrooms, size in square feet, and the sale price.

**Initial Observations:**

* The dataset has 12 features and 2,150 records.
* Some columns have missing values, particularly 'Lot Frontage' and 'Garage Type'.
* The target variable (what we want to predict) is 'SalePrice'

**Key Insights from EDA:**

* There is a strong positive relationship between the size of the living area ('GrLivArea') and the sale price ('SalePrice').
* The neighborhood significantly affects house prices.
* Houses with more bathrooms and bedrooms tend to be priced higher.
* Some outliers in the sale price could skew the model's predictions

**Data Preprocessing:**

* **Handling Missing Values:** We filled missing values in 'Lot Frontage' with the median value and replaced missing 'Garage Type' entries with 'None'.
* **Encoding Categorical Variables:** We converted categorical features like 'Neighborhood' and 'House Style' into numerical values using one-hot encoding.
* **Feature Scaling:** We standardized numerical features like 'GrLivArea' and 'LotArea' to ensure they have a similar scale

**Models Built:**

* **Baseline Model:** Linear Regression
* **Advanced Models:** Random Forest, Gradient Boosting, and XGBoost

**Performance Metrics:**

* **Linear Regression:** RMSE (Root Mean Squared Error) of 42,000
* **Random Forest:** RMSE of 28,000
* **Gradient Boosting:** RMSE of 25,000
* **XGBoost:** RMSE of 24,000

**Addressing Overfitting:**

* We used cross-validation to ensure that the model's performance is reliable and not just tailored to the training data.
* We tuned hyperparameters using GridSearchCV to find the best settings for Random Forest and Gradient Boosting models

**Conclusion:**

Our best model, XGBoost, had the lowest RMSE of 24,000, indicating it predicts house prices quite accurately. This project shows that machine learning can be very effective in predicting house prices based on various features. For future improvements, we could add more detailed data or explore new features like proximity to schools or shopping centers.