**Team**: Alice Vadney, Athena Vo, Avery Donmoyer, Naylor Stone

**The best research is organized around a question, which should be organized around prediction.**

Which features related to property characteristics, location, quality and condition, and market conditions have the most significant influence on predicting house prices using a random forest model, and how do these features impact the model's predictive accuracy?

---

**What is an observation in your study?**

An observation in our study is an individual house sale record and all of its attached attributes, including house square footage, year built, number of bedrooms, number of bathrooms, neighborhood, and more.

**Supervised or unsupervised learning? Classification or regression?**

Our model of random forests will be supervised learning because we're using training data to inform the model that will ultimately predict prices. We'll use our model to predict prices, which means our analysis will be regression because our predictive values are continuous. Some of our input data are categorical, and we will encode these values so they are suited for the model.

**What models or algorithms do you plan to use in your analysis? How?**

We plan to use the random forest method to conduct our analysis. Our data has already been split into training and test sets, and we will use that built-in organization while building our model. The model will be trained on the training dataset, and we'll average together a large number of trees to better capture data patterns and relationships.

**How will you know if your approach 'works'? What does success mean?**

We will know if our approach worked by a number of indicators, including the value of $R^2$. As the value of $R^2$ increases, becoming closer to 1.0, it is an indication that our model explains a substantial proportion of the data variance. For this project, a threshold of around 0.8 or higher might indicate strong predictive power, though the specific threshold will depend on what's typical in similar housing price prediction models.

**Team**: Alice Vadney, Athena Vo, Avery Donmoyer, Naylor Stone

Another indicator of model success is the RMSE value. A lower RMSE indicates that the model minimizes large prediction errors. Our model would be successful if it achieves an RMSE that is competitive with or lower than simpler models, making the model practically useful.

We will also look at performance across the individual decision trees to ensure that the bootstrapped data generated for the purpose of creating a random forest. If metrics such as $R^2$ and RMSE vary widely across the individual decision trees, it may indicate an unreliable model from the bootstrapped data. Lastly, residual analysis can be an important factor in ensuring that there is no bias in the model, by over or underestimating certain price ranges, and that errors are uniformly distributed. If residuals show patterns, it may indicate missing variables or areas where the model could improve.

**What are weaknesses that you anticipate being an issue? How will you deal with them if they come up? If your approach fails, what might you learn from this unfortunate outcome?**

One major weakness is the high dimensionality and multicollinearity within the dataset. The dataset contains a large number of features, some of which might be highly correlated with each other. This can increase the model's complexity, leading to overfitting, where the model performs well on the training data but poorly on unseen data. Random forests inherently provide feature importance scores. Thus, after training the model, we will analyze these scores to identify and potentially remove features that contribute minimally to the prediction.

Although random forests are less prone to overfitting compared to individual decision trees, they can still overfit the training data if the number of trees is too large or the trees are too deep, capturing noise instead of the underlying pattern, resulting in poor generalization to unseen data. To mitigate this, we will carefully tune the hyperparameters such as the number of trees (*n_estimators*), maximum depth of each tree (*max_depth*), and minimum samples per leaf (*min_samples_leaf*). Random forests also provide an OOB error estimate, which we will monitor to detect overfitting and help adjust the hyperparameters.

Several features in our dataset contain missing values. Features like lot frontage, garage year built, or basement quality have many missing values. Additionally, outliers such as extremely high house prices or unusual feature values (e.g. unusually large lots) can distort the model's performance. To mitigate this, we will drop features that have a high proportion of

missing values to help reduce the complexity of the model. We will also detect outliers in numerical features using box plots and z-scores, then cap extreme values to a reasonable percentile (e.g. 99th percentile) to reduce their impact on the model.

If our approach fails, that could indicate that our chosen model (i.e. random forest) is either too simple, missing critical patterns, or is too complex, which would highlight the need for a better model to fit our data. Failure could also indicate that our dataset is missing critical variables that significantly influence house prices, such as economic conditions, neighborhood-specific trends, or unrecorded features like recent renovations. It could also indicate that the housing market is more complex, influenced by non-linear interactions between features, than we previously thought.

**Feature Engineering: How will you prepare the data specifically for your analysis? For example, are there many variables that should be one-hot encoded? Do you have many correlated numeric variables, for which PCA might be a useful tool?**

To prepare the data for analysis, we will first handle the numerical variables using log transformations. Features such as *GrLivArea*, *LotArea*, *TotalBsmtSF*, and *SalePrice* will undergo log transformations to stabilize their variance and make relationships between features more linear.

The dataset has many numerical features that are highly correlated. For instance, *1stFlrSF* and *TotalBsmtSF* are correlated because a larger basement area usually corresponds to a larger first floor. PCA would combine these to represent the overall size of the lower area of the house. *GarageArea* and *GarageCars* both describe garage capacity. PCA would help reflect overall garage size. *TotRmsAbvGrd* and *GrLivArea* represent the total number of rooms above ground and the above-ground living area. PCA would help capture the essence of house size. *YearBuilt* and *YearRemodAdd*, while not strictly numerical, often show correlation. PCA would help capture the overall age and modernization level of the property.

We will detect outliers by visually inspecting features using box plots to identify values that lie beyond the interquartile range (IQR). We will also calculate z-scores to detect values that are several standard deviations away from the mean. For features such as *GrLivArea* and *SalePrice*, we will cap them beyond a certain threshold. In extreme cases, outliers that are clearly unrepresentative of the general market (e.g. exceptionally low *SalePrice*) may be removed.

**Team**: Alice Vadney, Athena Vo, Avery Donmoyer, Naylor Stone

**Results: How will you communicate or present your results? This might be a table of regression coefficients, a confusion matrix, or comparisons of metrics like $R^2$ and RMSE or accuracy and sensitivity/specificity. This is how you illustrate why your plan succeeded or explain why it failed.**

To evaluate the models and our results, we will compare the $R^2$ value of individual trees to the $R^2$ value of our forest, to analyze how variation across the models affects the quality of predictions. We will present our results with a table summarizing the key metrics, including $R^2$ and RMSE, across the different decision trees. This summary will provide a comprehensive view of the models' consistency and reliability in predicting housing prices.

We also plan to communicate our results visually, including a plot that displays the error between predicted house prices and actual house prices. This plot will illustrate how closely the model's predictions align with actual sale prices, highlighting areas where the model may overestimate or underestimate prices.