a) variables included in my model

1. **'Unit SF' (Total Livable Area):** It's highly impact on housing prices, as evident by the high importance through PCA and Random Forest analysis. It was subjected to standard scaling for normalization of the data and ensuring even influence across the model.
2. **'Lot SF' (Lot Square Footage)**: Chosen, as size of the lot is supposed to have great influence on the market value. Thereafter, it was also standardized to ensure it did not over-dominance the model, having confirmed it did not correlate strongly with other similar features (corr <= 0.8).
3. **'Miles' (to downtown)**: This field captures the geographic advantage, which is a key determinant of property desirability and pricing. Standard scaling has been applied to keep uniform influence on predictive outcomes.
4. **HOA Dues/Mo (Homeowners Association Dues per Month)**: Included as part of running the property costs and affordability of the buyer due to its relevance but standardized to assess its effect always in relation with the other cost-dependent features.

Why did I choose those features in my model?

First, I extracted the information first from the form column and converting the categorical feature to numerical using one-hot encoding and target encoding, I further use correlation filtering out some highly related features in case of multicollinearity and use PCA and RFE to select the top 15 important features, lastly using random forest feature importance to cross-validate. those feature are best for the model

b) Explain what functional form you have assumed. (10 points)

The predictive model for house prices employs a stacking ensemble method, enhancing accuracy by integrating predictions from multiple base models with a meta-model. The ensemble comprises:

1. Base Models: Random Forest and Gradient Boosting handle complex, non-linear relationships using multiple decision trees and sequential error correction, respectively. These models effectively capture intricate patterns from variables like 'Unit SF', 'Lot SF', 'Miles', and 'HOA Dues/Mo', without assuming a specific functional form.
2. Meta-model: Simple Linear Regression acts as the meta-model, utilizing a linear functional form. It processes inputs from the base models and combines them via a weighted linear equation.

This stacking approach, combining complex non-linear data analysis with straightforward linear synthesis, the functional form for this model is:

$$Price = \beta_0 + \beta_1 \times Random\ Forest + \beta_2 \times Gradient\ Boosting + \epsilon$$

Specifically in my best stocking mode's l functional form is:

$$Price = -10564 + 0.48 \times Random\ Forest + 0.54 \times Gradient\ Boosting + \epsilon$$

However, this is not straightforward because the stacking model does not learn a direct

c) Comment on the explanatory power of the model. (10 points)

The model's performance in predicting house prices is effectively summarized by its Mean Absolute Error (MAE) of $24,768.28 and an R^2 score of 0.73. The MAE indicates that the model's predictions typically deviate by about $24,768 from actual house prices, offering stakeholders a clear measure of prediction accuracy. An R^2 of 0.73 signifies that the model explains approximately 73% of the variance in house prices, demonstrating strong explanatory power within the context of real-world data.

d) Comment on the signs and significance of the model's explanatory variables. (20 points)

In assessing the predictive model for house prices, we focus on the signs and statistical significance of the explanatory variables' coefficients:

1. **Unit SF**: Coefficient: 44,460; Sign: Positive;Statistical Significance: Highly significant ($p < 0.001$)
    - Interpretation: A larger unit size correlates with higher prices, aligning with market expectations and confirming its strong predictive value.
2. **Lot SF:**Coefficient: 7,551; Sign: Positive;Statistical Significance: Not significant ($p = 0.173$)
    - Interpretation: Despite a positive sign indicating that larger lots might fetch higher prices, the lack of statistical significance suggests an unreliable relationship in the data.
3. **Miles:** Coefficient: -14,100; Sign: Negative; Statistical Significance: Significant ($p = 0.023$)
    - Interpretation: Distance from downtown negatively impacts house prices, with statistical significance confirming the importance of location in property valuation.
4. **HOA Dues/Mo:** Coefficient: 20,210; Sign: Positive; Statistical Significance: Highly significant ($p = 0.001$)
    - Interpretation: Higher HOA dues are linked to higher house prices, likely reflecting better amenities or services that add value.

The model's explanatory variables generally align with real estate principles: larger properties and better amenities increase value, and proximity to key locations enhances appeal. The significant coefficients for 'Unit SF' and 'HOA Dues/Mo' validate their strong influence on price, while the non-significance of 'Lot SF' invites further analysis to understand its variable impact. This comprehensive evaluation aids both sellers in pricing strategies and buyers in making informed decisions.

e) Estimate the price of a base unit

Using our stacking model with features including 'Unit SF', 'Miles', 'HOA Dues/Mo', 'Beds', 'Pool', 'Full Baths', and 'Half Baths', the estimated price for a base unit in The Lakes townhome subdivision is $662,027.75. This prediction reflects the unit’s specific attributes and location.

f) Using the base unit, determine the marginal amenity value for an additional full bathroom (vs the half bath)

In the analysis using our stacking model, we calculated the marginal amenity values for specific features in a base unit within The Lakes townhome subdivision:

1. Additional Full Bathroom: The introduction of an additional full bathroom (as opposed to a half bath) adds a marginal value of $850.76 to the unit. This value reflects the increased functionality and appeal that an extra full bathroom provides.
2. Proximity to Downtown Woodstock: Each unit closer to Downtown Woodstock sees an increase in value by $2,408.00. This increment underscores the desirability and convenience associated with proximity to central urban amenities.

g) Graph and comment on the price profile of such a unit at different distances from downtown Woodstock. (20 points)

1. **Initial Drop in Price**: There is a sharp decrease in the predicted price as the distance from downtown increases from 0 to 0.25 miles, suggesting a high value placed on properties within a quarter-mile radius of the downtown area.
2. **Gradual Decline**: Between 0.25 miles and approximately 1 mile, the price continues to decline, albeit at a slower rate, indicating that while the proximity to downtown is still valued, the premium for being very close to downtown decreases.
3. **Plateauing of Prices**: After around 1 mile, the price profile flattens out, indicating that the marginal value of being closer to downtown diminishes after this point, and other factors may become more influential in determining the property value.
4. **Stabilization:** Beyond 1.5 miles, the predicted prices stabilize, suggesting that the distance from downtown does not significantly affect property values beyond this point within the context of this model.