Q1

1. What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

- The optimal value of alpha in Ridge and Lasso regression models depends on the dataset and the problem being addressed. Typically, the best alpha value is determined through hyperparameter tuning using techniques like cross-validation, grid search, or information criteria, where different alpha values are tested, and the one that gives the best performance (often evaluated using metrics like mean squared error or R-squared) on unseen data is chosen.
- Changes with Double the Value of Alpha:
  - Ridge Regression: Doubling the alpha in Ridge regression further shrinks the coefficients toward zero, increasing the regularization effect. This could potentially lead to underfitting and lower predictive accuracy.
  - Lasso Regression: For Lasso regression, doubling the alpha intensifies the sparsity effect, pushing more coefficients to exactly zero. This leads to a model with fewer features. However, it might also reduce model accuracy as important features might be discarded.
- Most Important Predictor Variables after Change
  - Ridge Regression: After implementing the change, in Ridge regression, the most important predictor variables will likely be those with non-zero coefficients or larger absolute coefficients, indicating higher impact on model predictions despite increased regularization.
  - Lasso Regression: For Lasso regression, post-change, the most important predictor variables would be the ones with non-zero coefficients, as those are the features that remain in the model.

specific to this dataset

These were the best features for Lasso

GrLivArea
LotArea OverallQual_9

YearBuilt
TotalBsmtSF OverallQual_8
BsmtFinSF1
OverallCond_3
Neighborhood_Crawfor
Neighborhood_StoneBr

after doubling the alpha it became

['MSSubClass_180', 'Exterior1st_AsphShn', 'Foundation_Wood', 'Foundation_Stone', 'Foundation_Slab', 'Foundation_CBlock', 'BsmtQual_None', 'OverallQual_10', 'OverallQual_6', 'OverallQual_2']

For Ridge the top features were

['Exterior2nd_HdBoard', 'GarageType_Detchd', 'MSSubClass_70', 'Neighborhood_CollgCr', 'BsmtFinType1_BLQ', 'BsmtFinType1_Unf', 'Exterior1st_HdBoard', 'MSSubClass_40', 'MSSubClass_80', 'Exterior1st_AsphShn']

after doubling alpha it became

['Exterior1st_WdShing', 'Neighborhood_CollgCr', 'Exterior2nd_VinylSd', 'GarageType_Detchd', 'Exterior2nd_HdBoard', 'MSSubClass_80', 'FireplaceQu_TA', 'TotRmsAbvGrd_6', 'BsmtFinType1_BLQ',

```
#     'Foundation_Stone']
```

Q2

1. You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Choosing Based on Optimal Lambda If the optimal lambda for Lasso regression significantly reduces the number of features while maintaining or improving predictive performance, and if the feature selection aligns with domain knowledge or expectations about the dataset, then Lasso regression is a good choice. However, if interpretability of the model is crucial and I believe most features might contribute to the outcome, I might choose Ridge regression or a smaller lambda for Lasso.

Ultimately, the choice between Ridge and Lasso regression, or the optimal lambda value, is a balance between model complexity, interpretability, and performance. It's often beneficial to compare the performance of both models (using metrics like mean squared error, R-squared, etc.) and consider the implications of their feature selection capabilities before making a final decision.

Q3

1. After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

In one of my lasso models the top 5 features were

- GrLivArea
- LotArea
- TotalBsmtSF
- YearBuilt
- OverallQual

Best features dropping top 5 features are

- 1stFlrSF
- 2ndFlrSF
- BsmtFinSF1
- Neighborhood_StoneBr
- OverallCond

Q4

1. How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Ensuring a model is robust and generalizable involves several key strategies:

- Data Quality and Diversity:
    - High-Quality Data: Train the model on accurate and representative data, avoiding biases, outliers, and missing values.
    - Diverse Data Representation: Augment the dataset using techniques like noise injection and transformations to improve its representativeness and generalize better to unseen data.
- Model Selection and Training Techniques:
    - Regularization: Employ techniques like L1 and L2 regularization to control model complexity, prevent overfitting, and enhance robustness to noise.

    - Cross-Validation: Validate the model on separate training and validation splits to avoid overfitting and determine the most generalizable model configuration.
    - Ensemble Methods: Combine predictions from multiple models to create a stronger, more robust model that generalizes better.
    - Evaluation and Analysis:
        - Metrics Beyond Accuracy: Use metrics like precision, recall, and F1-score to evaluate model performance across different scenarios and distributions, rather than relying solely on accuracy.
        - Error Analysis: Analyze the types of errors the model makes on unseen data to understand limitations and improve generalizability.
        - Interpretability: Choose interpretable models or techniques to understand model predictions, identify biases, and limitations.
- Implications for Model Accuracy:
    - Trade-offs: Prioritizing robustness and generalizability often involves trade-offs with accuracy on the training data.
    - Overfitting Avoidance: Models optimized solely for training data might not perform well on new data, while robust models tend to capture underlying patterns and perform better on unseen data.
    - Complexity vs. Interpretability: Highly accurate models might be complex and less interpretable. Simplifying models enhances generalizability but might reduce raw accuracy.
- Why it Matters:
    - Real-World Applicability: Models need to perform well on new, unseen data to be useful in real-world scenarios.
    - Trust and Reliability: Robust, generalizable models are more trustworthy and reliable for decision-making in diverse situations.
- Balancing accuracy with robustness and generalizability is crucial in model development. While raw accuracy is desirable, ensuring the model can perform well across various scenarios and datasets enhances its reliability and applicability in real-world settings.