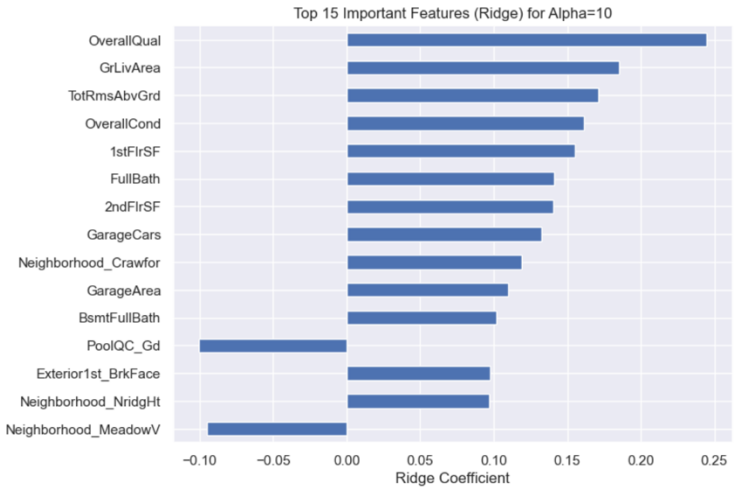
Question 1

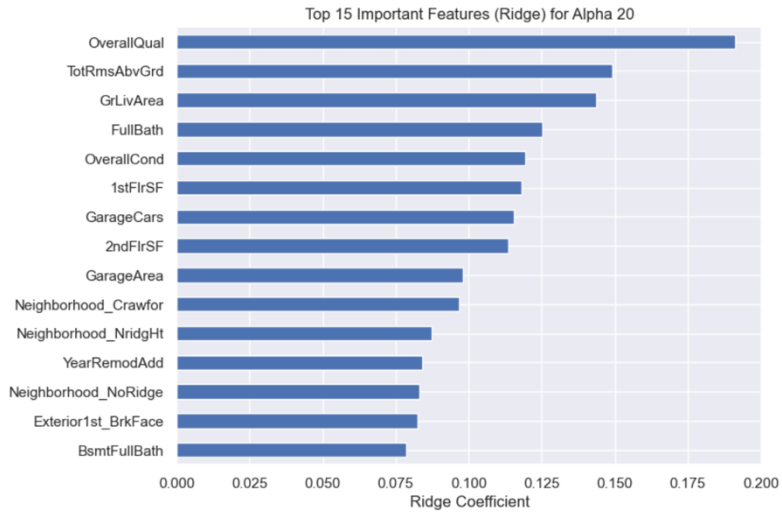
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:

For Ridge Alpha value chosen in 10. With Ridge regression at alpha =10, the top 10 most important variables were OverallQual, GrLivArea, TotRmsAbvGrd, OverallCond, 1stFlrSF, FullBath, 2ndFlrSF, GarageCars, Neighborhood_Crawfor, and GarageArea. These features highlight that house price is primarily driven by overall construction quality, total living area, number of rooms, physical condition, floor space distribution, bathroom count, garage capacity, and location. After doubling the regularization strength to alpha =20, the top 10 variables shifted slightly to OverallQual, TotRmsAbvGrd, GrLivArea, FullBath, OverallCond, 1stFlrSF, GarageCars, 2ndFlrSF, GarageArea, and Neighborhood_Crawfor. While the ranking of some variables changed, the core predictors remained the same, indicating strong model stability.

Increasing the value of alpha made the model reduce the strength of less important features. However, the main and most important features did not change. The top predictors were still related to house quality, house size, how usable the space is, and the desirability of the neighborhood.





For Lasso: Alpha choosen is 0.001

When the alpha value is doubled in Lasso regression, the strength of regularization increases, which makes the model more selective. As a result, more coefficients are shrunk to zero, reducing the number of selected features from 87 to 62. This leads to a simpler and more interpretable model, but with a small trade-off in predictive accuracy. Doubling the Alpha value leads to underfitting

After increasing the Lasso alpha, the most important predictors of house prices are GrLivArea, OverallQual, GarageCars, TotRmsAbvGrd, OverallCond, BsmtFullBath, FullBath, Neighborhood_NridgHt, CentralAir_Y, and BsmtExposure_Gd, indicating that living area, quality, amenities, and location remain the key drivers of pricing.

With a lower alpha, the Lasso model is less strict and retains more features, including weaker predictors such as PoolQC_Gd and multiple neighborhood indicators. When the alpha is increased, these weaker features are removed, and the model focuses only on the strongest and most stable predictors like living area, overall quality, garage capacity, and key location and comfort features.

```
Number of selected features for Alpha 0.001: 87
GrLivArea              0.925578
OverallQual            0.530168
PoolQC_Gd              0.423338
OverallCond            0.248555
GarageCars             0.208587
TotRmsAbvGrd           0.148321
BsmtFullBath           0.146818
Neighborhood_Crawfor   0.115125
Neighborhood_NridgHt   0.111582
FullBath               0.102319
dtype: float64
```

```
Number of selected features for Alpha 0.002: 62
GrLivArea               0.712142
OverallQual             0.607388
GarageCars              0.256380
TotRmsAbvGrd            0.160997
OverallCond             0.154729
BsmtFullBath            0.108231
FullBath                0.098801
Neighborhood_NridgHt    0.093377
CentralAir_Y            0.089080
BsmtExposure_Gd         0.087313
dtype: float64
```

Question 2

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?

We observe that both Ridge and Lasso models show very similar performance on the test data. Ridge performs marginally better, with a slightly higher test R^2 and lower RMSE, but the difference compared to Lasso is very small and practically negligible.

Despite this, Lasso is chosen as the final model because it provides the additional benefit of feature selection. While maintaining comparable prediction accuracy, Lasso reduced the original feature set from over 258 variables to ~90 important features, making the model simpler and easier to interpret. This helps in identifying the key predictors that influence house prices and supports better business decision-making.

```
10
Ridge Train R2: 0.9127731647815557
Ridge Test R2: 0.8744827545479026
Ridge RMSE: 0.13744649630277958
```

```
Alpha: 0.001
Lasso Train R2: 0.9088427486286539
Lasso Test R2: 0.8736070972240357
Lasso RMSE: 0.13792510323469795
```

Question 3:

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?

Answer:

For alpha value of 0.001 and using Lasso Regularization -

Earlier the most important factors were : This table shows the Variable and their coefficients

GrLivArea	0.925578
OverallQual	0.530168
PoolQC_Gd	0.423338
OverallCond	0.248555
GarageCars	0.208587

After removal of above variables most important variable using lasso are: (This table shows the Variable and their coefficients)

1stFlrSF	1.023815
PoolQC_Gd	0.606218
2ndFlrSF	0.444642
GarageArea	0.256442
Neighborhood_NridgHt	0.139319

Question 4:

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

A model is robust and generalizable when it performs well not only on training data but also on new, unseen data. This can be achieved by using techniques like train–test split, cross-validation, and regularization, which help prevent the model from overfitting.

The impact on accuracy is that training accuracy may reduce slightly, but test accuracy becomes more reliable. This happens because the model focuses on learning real patterns instead of memorizing the training data, leading to better performance in real-world situations.

_____EOF_____