Housing Prices Case Study Subjective Questions

1. What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose to 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 lambda** values and **score metrics** for regularized models are as follows

Lasso Regression with alpha equal to 0.001

	R-squared	Adj. R-squared	RSS	MSE	RMSE
Train	0.897	0.876	16.599	0.016	0.128
Test	0.880	0.857	8.626	0.020	0.140

Ridge Regression with alpha equal to 6.0

	R-squared	Adj. R-squared	RSS	MSE	RMSE
Train	0.898	0.890	16.441	0.016	0.127
Test	0.878	0.869	8.802	0.020	0.142

If we choose to **double the value of alpha**, we will have the following changes.

Lasso Regression with alpha equal to 0.002

	R-squared	Adj. R-squared	RSS	MSE	RMSE
Train	0.881	0.858	19.031	0.019	0.137
Test	0.877	0.853	8.875	0.020	0.142

Ridge Regression with alpha equal to 12.0

	R-squared	Adj. R-squared	RSS	MSE	RMSE
Train	0.893	0.886	17.099	0.017	0.129
Test	0.878	0.869	8.790	0.020	0.142

In both updated models we can see a slight drop in the R-squared and Adjusted R-squared values, but values for RSS and RMSE increase, which aligns with the theory that the higher the value of alpha, the greater the error since the model is underfitting. If we keep on increasing alpha, we'll have more errors.

Now, we'll **compare the most important predictor variables** (top 10) before and after doubling the alpha value.

Lasso Regression: before in the left, after in the right.

	Feature	Coefficient
0	Constant	11.934
1	Neighborhood_Crawfor	0.118
2	Neighborhood_NridgHt	0.108
3	GrLivArea	0.102
4	OverallQual	0.100
5	Neighborhood_Somerst	0.090
6	Neighborhood_ClearCr	0.076
7	BsmtExposure_NA	-0.077
8	MSSubClass_160	-0.069
9	HouseAge	-0.068
10	Neighborhood_NoRidge	0.062

	Feature	Coefficient
0	Constant	11.947
1	OverallQual	0.113
2	GrLivArea	0.102
3	Neighborhood_Crawfor	0.082
4	Neighborhood_NridgHt	0.067
5	HouseAge	-0.066
6	OverallCond	0.051
7	MSSubClass_20	0.049
8	Neighborhood_Somerst	0.049
9	BsmtExposure_Gd	0.046
10	MSSubClass_160	-0.048

We have different top 10 columns and common ones in different order of affection to the *SalePrice*. *OverallQual* and *GrLivArea* affect the house price the most. *OverallCond*, *MSSubClass_20*, and *BsmtExposure_Gd* are the new columns.

The coefficients are smaller, which aligns with the theory that the greater the value of alpha the more the coefficients are pushed towards zero (and will become zero).

Ridge Regression: before in the left, after in the right.

	Feature	Coefficient
0	Constant	11.982
1	OverallQual	0.128
2	MSSubClass_80	0.108
3	MSSubClass_20	0.101
4	MSSubClass_30	0.099
5	Neighborhood_MeadowV	0.094
6	WoodDeckSF	0.094
7	MSSubClass_75	0.092
8	MSZoning_FV	0.085
9	BsmtFinSF1	-0.093
10	LotFrontage	0.084

	Feature	Coefficient
0	Constant	12.006
1	OverallQual	0.128
2	MSSubClass_80	0.093
3	LotFrontage	0.092
4	Neighborhood_MeadowV	0.089
5	MSSubClass_30	0.087
6	MSSubClass_20	0.084
7	BsmtFinSF1	-0.081
8	MSSubClass_75	0.072
9	MasVnrArea	-0.077
10	MSZoning_FV	0.070

We have common columns in different order of affection to the *SalePrice* and a new one. *MasVnrArea* is the new column. The coefficients are smaller, which aligns with the theory that the greater the value of alpha the more the coefficients are pushed towards zero (never becoming zero).

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?

   Both models perform good in terms of score metrics, but **Lasso Regression has slightly better scores**. As the **interpretability** of the model is also important, we need to **compare both models' top selected features.**

   We can see that even if during RFE for Ridge Regression the method chose many columns that Lasso Regression also selected, the **Ridge model gave more importance to variables that do not fully represent a house.**
   MSSubClass is a categorical variable that seems to have values that are manually assigned and derived from the other features, and the column is weakly related with *SalePrice*. In future improvements, the possibility of removing this column should be considered.

   Features selected by Lasso Regression model include variables we saw **are highly correlated with *SalePrice*** (from Numerical Analysis) and **can be logically and easily explained in terms of business**.

   In summary:
   - **The Lasso Regression model performed better** than Ridge Regression in terms of scoring metrics and feature selection.
   - As **Lasso Regression performs feature selection,** we do not need to analyze all variables to remove some, so it is easier to implement.

   From the above analysis and comparisons, **we will choose to apply Lasso Regression for modeling the house prices.**

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?

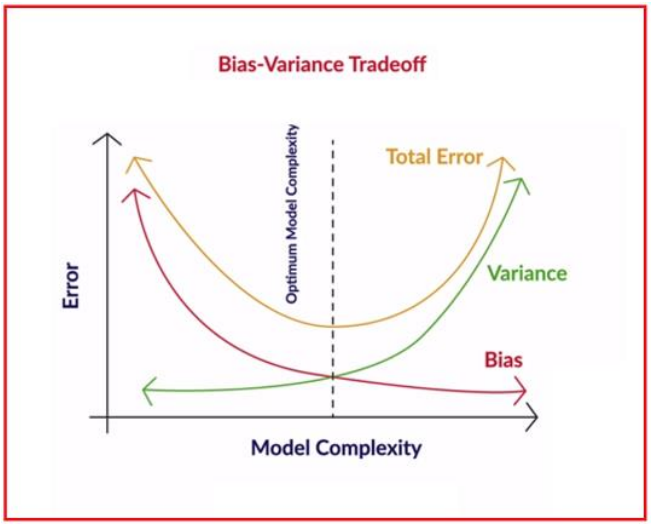
   The five most important predictor variables in Lasso model are the following:
   - *Neighborhood_Crawfor*
   - *Neighborhood_NridgHt*
   - *GrLivArea*
   - *OverallQual*
   - *Neighborhood_Somerst*

After creating another model excluding the above variables, for which we performed a search again to find optimal alpha and turned out to be the same as before (0.001), **the five most important predictor variables in Lasso model are** the following:

- *MSSubClass_50*
- *Foundation_Slab*
- *BsmtFinSF1*
- *1stFlrSF*
- *Neighborhood_ClearCr*

4. 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?

We can understand the implications of a model complexity by looking at the **following Bias and Variance tradeoff** graph.



We want to reduce both bias and variance because the expected total error of a model is the sum of the errors in bias and variance, so the optimal point is represented at the middle of the graph, where variance and bias intercept.

In the **extremes** of the relationship between bias and variance we have low accuracy:

- With high bias and low variance, the model is underfitting.
- With low bias and high variance, the model is overfitting.

Then, **a model should be as simple as possible**. When we talk about *simplicity*, we mean that the model should have **low bias and low variance** and should have enough features to describe the target variable in general terms (**as least features as possible**).

Then, a simpler model is more robust and generalisable. When achieving these characteristics, the model performs equally well on both training and test data, i.e., it has good accuracy on both sides.