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:

Optimal alpha for Ridge regression : 6
Optimal alpha for Lasso regression : 0.001

Here are the changes observed when the alpha value was doubled:

- Slight change in the metrics can be observed for both Ridge and Lasso, but it is not that significant.
- Lasso regression has eliminated 83 features when alpha doubled. In optimal alpha case it eliminated 61 features
- When alpha was doubled, Ridge Regression displayed a slight better performance/metric compared to Lasso regression

Metric table with optimal alpha

	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.910915	0.894430
1	R2 Score (Test)	0.886188	0.888002
2	RSS (Train)	14.488083	17.169148
3	RSS (Test)	7.821143	7.696485
4	MSE (Train)	0.119474	0.130059
5	MSE (Test)	0.133934	0.132863

Metric table when Alpha is doubled

	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.904144	0.875296
1	R2 Score (Test)	0.887735	0.877729
2	RSS (Train)	15.589209	20.280862
3	RSS (Test)	7.714829	8.402480
4	MSE (Train)	0.123931	0.141355
5	MSE (Test)	0.133021	0.138823

Most important variables are: Overall quality, Overall condition. Neighbourhood

Lasso with alpha doubled:

```
OverallQual_9: 0.1835628118394943
OverallCond_3: -0.10031294687359829
Neighborhood_Crawfor: 0.09929945862763964
MSSubClass_30: -0.09336699992288719
OverallQual_8: 0.08865044622226378
OverallQual_3: -0.07350067830890708
OverallQual_4: -0.07024112776998369
Neighborhood_IDOTRR: -0.06390478059192428
OverallCond_4: -0.06158991020053291
TotRmsAbvGrd: 0.061347290923643966
1stFlrSF: 0.06114187789199146
Neighborhood_NridgHt: 0.0610674031936282
BsmtExposure_Gd: 0.060645004776857935
GarageArea: 0.05504329030889074
Functional_Typ: 0.05473436563855146
Condition1_Norm: 0.05339134143915787
FireplaceQu_NA: -0.0529210116154539
FullBath: 0.04868698691770389
ExterQual_TA: -0.048227993170429286
HalfBath: 0.044046340124347834
Neighborhood_Edwards: -0.04176397108561379
SaleCondition_Partial: 0.0399041934973562
MSSubClass_160: -0.03656580962799536
OverallCond_7: 0.03440655834174169
KitchenQual_TA: -0.03326714508220582
```

Ridge with alpha doubled:

```
OverallQual_9: 0.15667527956602548
OverallQual_3: -0.10941987983876665
OverallCond_3: -0.1030164868279577
Neighborhood_Crawfor: 0.10145438589670736
Neighborhood_NridgHt: 0.09889230641033685
Neighborhood_IDOTRR: -0.0979262688639369
Neighborhood_StoneBr: 0.0854669378809203
MSSubClass_30: -0.08392421082019144
OverallCond_9: 0.07821391255134857
Neighborhood_MeadowV: -0.07683924733471904
OverallQual_8: 0.07636081726333101
Neighborhood_NoRidge: 0.0749109778431321
SaleCondition_Partial: 0.07450673419221117
OverallQual_4: -0.06946432783604219
OverallCond_4: -0.06905939368339355
Exterior1st_BrkFace: 0.06866304432080259
1stFlrSF: 0.06379402630628289
BsmtExposure_Gd: 0.06339233460106945
LandContour_HLS: 0.06314003475363933
OverallCond_7: 0.06158843316513589
Alley_Pave: 0.059551900188136817
SaleCondition_Normal: 0.05898026403999413
OverallQual_10: 0.05701963423082702
MSSubClass_160: -0.05669532658546908
Functional_Typ: 0.0553362222570052
```

Lasso with optimal alpha:

```
OverallQual_9: 0.23180834994848692
OverallCond_3: -0.1433480687640675
OverallQual_3: -0.14192146261349692
Neighborhood_Crawfor: 0.12393599144839308
MSSubClass_30: -0.11040868766033969
OverallQual_8: 0.10529280165532796
Neighborhood_IDOTRR: -0.10457959049613784
Neighborhood_NridgHt: 0.09715702162632396
Neighborhood_MeadowV: -0.08124917770512712
Neighborhood_NoRidge: 0.07884107336229927
OverallQual_4: -0.07609443471993738
BsmtExposure_Gd: 0.07396109024321482
Neighborhood_StoneBr: 0.07178978409139304
SaleCondition_Partial: 0.07114892215785995
OverallCond_4: -0.0670686373441287
Exterior1st_BrkFace: 0.06604112997671949
Condition1_Norm: 0.05871878825428102
1stFlrSF: 0.05749622026580367
MSSubClass_160: -0.05726737424939389
OverallQual_10: 0.05624524243575949
TotRmsAbvGrd: 0.0556999329409701
Functional_Typ: 0.05558295418962677
SaleCondition_Normal: 0.05413032764623093
OverallCond_9: 0.05273538179512369
FireplaceQu_NA: -0.048856568504996346
```

Ridge with optimal alpha:

```
OverallQual_9: 0.19652987941242814
OverallQual_3: -0.1389828713348906
OverallCond_3: -0.1272977419260038
Neighborhood_IDOTRR: -0.1243659404938063
Neighborhood_Crawfor: 0.11743830188650753
Neighborhood_NridgHt: 0.11647677318184468
Neighborhood_StoneBr: 0.10889438411303765
Neighborhood_MeadowV: -0.10463432911846711
OverallCond_9: 0.10319255118004424
Neighborhood_NoRidge: 0.09612081933379585
OverallQual_10: 0.0899107701644299
OverallQual_8: 0.08973435998747965
SaleCondition_Partial: 0.08957497708997422
MSSubClass_30: -0.08823812169605044
Condition2_Norm: 0.08208493654477092
Exterior1st_BrkFace: 0.08140806141654759
Heating_GasW: 0.07759209196451808
LandContour_HLS: 0.07579295697574985
Alley_Pave: 0.07343001615149523
OverallQual_4: -0.07156783085346168
Functional_Maj2: -0.0714530352745558
OverallCond_4: -0.07095213678847591
HouseStyle_2.5Unf: 0.07087427727534425
HouseStyle_2.5Fin: 0.07007271943732513
SaleCondition_Normal: 0.06837050972154438
```

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?

Lasso and Ridge regression has almost same train and test accuracy (Observed from the metric table)

Lasso regression model is slightly better than Ridge model (metric on test data determines the same).Lasso regression has omitted 61 features.

Lasso regression shall be chosen to apply based on aforementioned observations.

	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.910915	0.894430
1	R2 Score (Test)	0.886188	0.888002
2	RSS (Train)	14.488083	17.169148
3	RSS (Test)	7.821143	7.696485
4	MSE (Train)	0.119474	0.130059
5	MSE (Test)	0.133934	0.132863

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?

A new model was created by excluding the 5 most important predictor variables. Here are the 5 most important predictors based on the new lasso model.

Following list includes the significant features that impact the house price in both positive and negative manner:

- Functional (Home functionality)
- LotShape (General shape of property)
- LandContour (Flatness of the property)
- BsmtCond (Evaluates the general condition of the basement)
- TotRmsAbvGrd (Total rooms above grade)

```
Functional_Typ: 0.11570012714622015
LotShape_IR3: -0.10372148037609494
LandContour_HLS: 0.09906664202977859
BsmtCond_Gd: 0.07651039337530996
TotRmsAbvGrd: 0.07556836505913686
SaleCondition_Partial: 0.07430255433370099
Exterior1st_BrkFace: 0.07259720717069743
1stFlrSF: 0.0715686361647663
Condition1_Norm: 0.06921850133030655
FireplaceQu_NA: -0.06515414626076281
BldgType_Twnhs: -0.06319676200857655
SaleCondition_Normal: 0.061146717509894415
BsmtCond_TA: 0.060625426985427784
ExterQual_TA: -0.058915749778067794
GarageArea: 0.05770866340767669
Alley_Pave: 0.05627448442329541
LotConfig_CulDSac: 0.04689475922592703
BsmtFinType1_GLQ: 0.043778776650482014
KitchenQual_TA: -0.042912724572459635
FullBath: 0.040387868407064154
YearRemodAdd: 0.03999149671866924
HalfBath: 0.03996320777839013
RoofStyle_Hip: 0.03647829365798285
HeatingQC_TA: -0.03644215333547513
CentralAir: 0.0341588586398057
```

Above is the snapshot of params and its coefficients for reference

Question 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 make sure model is robust and generalisable by following below steps:

- Data cleaning , outlier detection and removal
- Feature selection, feature transformation, creating new features
- Dividing dataset into train, validation, test data. If dataset is not big enough for dividing into validation then use cross validation techniques
- Check for imbalance data and use stratified sampling accordingly
- Check for underfitting and overfitting
- Use regularisation techniques to prevent overfitting
- Hyperparameter tuning
- Make use of appropriate metrics for different models for evaluation
- Make sure data leak is not happening i.e. model evaluation should happen only on test data (unseen data)

Implications of the above for the accuracy of the model:

- Robust and generalisable model maintains consistent performance
- It performs well on unseen data (no overfitting)
- Prediction will be accurate
- Helps in better decision making