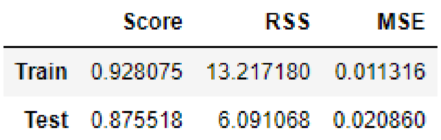
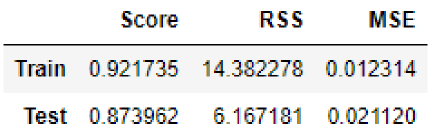
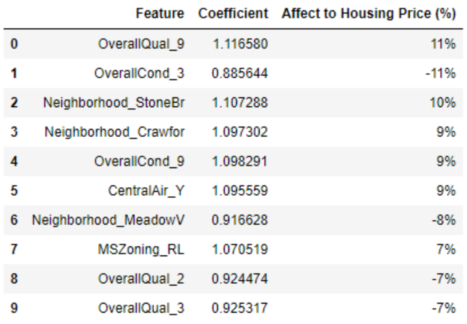
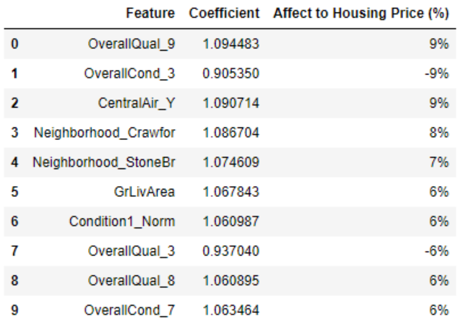
Assignment-based Subjective Questions

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 alpha value for **Ridge Regression** is 8 while the optimal value for **Lasso Regression** is 0.001

If we double the alpha value for both Ridge and Lasso

- ***Alpha value of 16 for Ridge***

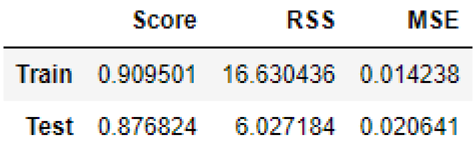
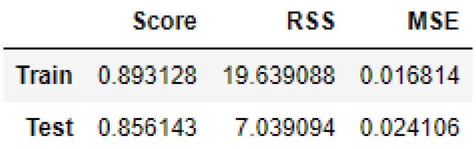
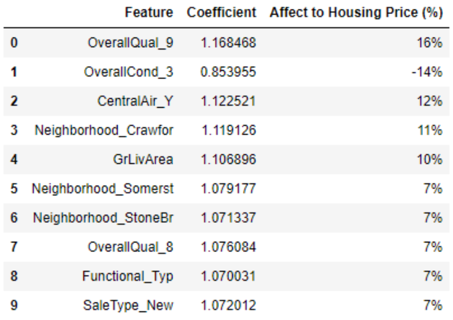
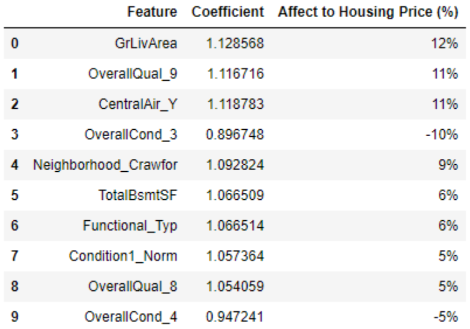
<table>
<tr><th></th><th>Before</th><th>After</th></tr>
<tr><td><b>Model Score</b></td><td><table><tr><th></th><th>Score</th><th>RSS</th><th>MSE</th></tr><tr><td>Train</td><td>0.928075</td><td>13.217180</td><td>0.011316</td></tr><tr><td>Test</td><td>0.875518</td><td>6.091068</td><td>0.020860</td></tr></table></td><td><table><tr><th></th><th>Score</th><th>RSS</th><th>MSE</th></tr><tr><td>Train</td><td>0.921735</td><td>14.382278</td><td>0.012314</td></tr><tr><td>Test</td><td>0.873962</td><td>6.167181</td><td>0.021120</td></tr></table></td></tr>
<tr><td><b>Most important features</b></td><td><table><tr><th></th><th>Feature</th><th>Coefficient</th><th>Affect to Housing Price (%)</th></tr><tr><td>0</td><td>OverallQual_9</td><td>1.116580</td><td>11%</td></tr><tr><td>1</td><td>OverallCond_3</td><td>0.885644</td><td>-11%</td></tr><tr><td>2</td><td>Neighborhood_StoneBr</td><td>1.107288</td><td>10%</td></tr><tr><td>3</td><td>Neighborhood_Crawfor</td><td>1.097302</td><td>9%</td></tr><tr><td>4</td><td>OverallCond_9</td><td>1.098291</td><td>9%</td></tr><tr><td>5</td><td>CentralAir_Y</td><td>1.095559</td><td>9%</td></tr><tr><td>6</td><td>Neighborhood_MeadowV</td><td>0.916628</td><td>-8%</td></tr><tr><td>7</td><td>MSZoning_RL</td><td>1.070519</td><td>7%</td></tr><tr><td>8</td><td>OverallQual_2</td><td>0.924474</td><td>-7%</td></tr><tr><td>9</td><td>OverallQual_3</td><td>0.925317</td><td>-7%</td></tr></table></td><td><table><tr><th></th><th>Feature</th><th>Coefficient</th><th>Affect to Housing Price (%)</th></tr><tr><td>0</td><td>OverallQual_9</td><td>1.094483</td><td>9%</td></tr><tr><td>1</td><td>OverallCond_3</td><td>0.905350</td><td>-9%</td></tr><tr><td>2</td><td>CentralAir_Y</td><td>1.090714</td><td>9%</td></tr><tr><td>3</td><td>Neighborhood_Crawfor</td><td>1.086704</td><td>8%</td></tr><tr><td>4</td><td>Neighborhood_StoneBr</td><td>1.074609</td><td>7%</td></tr><tr><td>5</td><td>GrLivArea</td><td>1.067843</td><td>6%</td></tr><tr><td>6</td><td>Condition1_Norm</td><td>1.060987</td><td>6%</td></tr><tr><td>7</td><td>OverallQual_3</td><td>0.937040</td><td>-6%</td></tr><tr><td>8</td><td>OverallQual_8</td><td>1.060895</td><td>6%</td></tr><tr><td>9</td><td>OverallCond_7</td><td>1.063464</td><td>6%</td></tr></table></td></tr>
</table>

***Insights*:**

- Train score decreases from 0.928 to 0.922, Test score decreases from 0.876 to 0.874
- Train RSS increases from 13.217 to 14.382, Test RSS increases from 6.091 to 6.167
- Train MSE increases from 0.011 to 0.012, Test MSE in increases from 0.020 to 0.021
- Most important features after the change:
  - OverallQual_9
  - OverallCond_3
  - CentralAir_Y
  - Neighborhood_Crawfor
  - Neighborhood_StoneBr
  - GrLivArea
  - Condition1_Norm
  - OverallQual_3

    - OverallQual_8
    - OverallCond_7

- ***Alpha value of 0.002 for Lasso***

<table>
<tr><th></th><th>Before</th><th>After</th></tr>
<tr><td><b>Model Score</b></td><td>
<table>
<tr><th></th><th>Score</th><th>RSS</th><th>MSE</th></tr>
<tr><td>Train</td><td>0.909501</td><td>16.630436</td><td>0.014238</td></tr>
<tr><td>Test</td><td>0.876824</td><td>6.027184</td><td>0.020641</td></tr>
</table></td><td>
<table>
<tr><th></th><th>Score</th><th>RSS</th><th>MSE</th></tr>
<tr><td>Train</td><td>0.893128</td><td>19.639088</td><td>0.016814</td></tr>
<tr><td>Test</td><td>0.856143</td><td>7.039094</td><td>0.024106</td></tr>
</table></td></tr>
<tr><td><b>Most important features</b></td><td>
<table>
<tr><th></th><th>Feature</th><th>Coefficient</th><th>Affect to Housing Price (%)</th></tr>
<tr><td>0</td><td>OverallQual_9</td><td>1.168468</td><td>16%</td></tr>
<tr><td>1</td><td>OverallCond_3</td><td>0.853955</td><td>-14%</td></tr>
<tr><td>2</td><td>CentralAir_Y</td><td>1.122521</td><td>12%</td></tr>
<tr><td>3</td><td>Neighborhood_Crawfor</td><td>1.119126</td><td>11%</td></tr>
<tr><td>4</td><td>GrLivArea</td><td>1.106896</td><td>10%</td></tr>
<tr><td>5</td><td>Neighborhood_Somerst</td><td>1.079177</td><td>7%</td></tr>
<tr><td>6</td><td>Neighborhood_StoneBr</td><td>1.071337</td><td>7%</td></tr>
<tr><td>7</td><td>OverallQual_8</td><td>1.076084</td><td>7%</td></tr>
<tr><td>8</td><td>Functional_Typ</td><td>1.070031</td><td>7%</td></tr>
<tr><td>9</td><td>SaleType_New</td><td>1.072012</td><td>7%</td></tr>
</table></td><td>
<table>
<tr><th></th><th>Feature</th><th>Coefficient</th><th>Affect to Housing Price (%)</th></tr>
<tr><td>0</td><td>GrLivArea</td><td>1.128568</td><td>12%</td></tr>
<tr><td>1</td><td>OverallQual_9</td><td>1.116716</td><td>11%</td></tr>
<tr><td>2</td><td>CentralAir_Y</td><td>1.118783</td><td>11%</td></tr>
<tr><td>3</td><td>OverallCond_3</td><td>0.896748</td><td>-10%</td></tr>
<tr><td>4</td><td>Neighborhood_Crawfor</td><td>1.092824</td><td>9%</td></tr>
<tr><td>5</td><td>TotalBsmtSF</td><td>1.066509</td><td>6%</td></tr>
<tr><td>6</td><td>Functional_Typ</td><td>1.066514</td><td>6%</td></tr>
<tr><td>7</td><td>Condition1_Norm</td><td>1.057364</td><td>5%</td></tr>
<tr><td>8</td><td>OverallQual_8</td><td>1.054059</td><td>5%</td></tr>
<tr><td>9</td><td>OverallCond_4</td><td>0.947241</td><td>-5%</td></tr>
</table></td></tr>
<tr><td><b>Feature Eliminated</b></td><td>191</td><td>221</td></tr>
</table>

***Insights*:**

- Train score decreases from 0.91 to 0.893, Test score decreases from 0.877 to 0.856
- Train RSS increases from 16.630 to 19.639, Test RSS increases from 6.027 to 7.039
- Train MSE increases from 0.014 to 0.017, Test MSE in increases from 0.021 to 0.024
- Number of eliminated features increases from 191 to 221
- Most important features after the change:
    - GrLivArea
    - OverallQual_9
    - CentralAir_Y
    - OverallCond_3
    - Neighborhood_Crawfor
    - TotalBsmtSF
    - Functional_Typ
    - Condition1_Norm
    - OverallQual_8
    - OverallCond_4

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?**

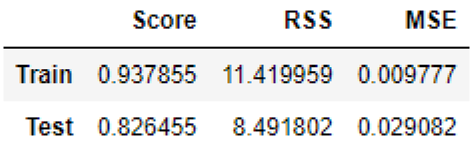
   The optimal value of lambda for **Ridge** is 8, and for **Lasso** is 0.001

   Both models score is good, but the business goal is to find the most important features so feature selection should be performed. Also, feature elimination helps making the model simple and robust. So, **we should choose Lasso in this scenario to eliminate less important features**

3. **After building the model, you realized 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?**

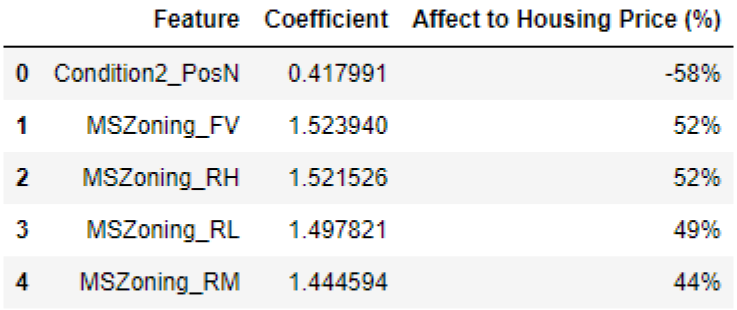
Top 5 variables to be removed: **OverallQual_9, CentralAir_Y, Neighborhood_Crawfor, GrLivArea, Neighborhood_StoneBr**

After using cross validation, optimal alpha value for Lasso Model is now **0.0001**



	Score	RSS	MSE
Train	0.937855	11.419959	0.009777
Test	0.826455	8.491802	0.029082

Top 5 variables and its effect to housing price per 1 unit are:



	Feature	Coefficient	Affect to Housing Price (%)
0	Condition2_PosN	0.417991	-58%
1	MSZoning_FV	1.523940	52%
2	MSZoning_RH	1.521526	52%
3	MSZoning_RL	1.497821	49%
4	MSZoning_RM	1.444594	44%

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?**

   To make sure a model is robust and generalizable, we should:

   - The robustness and generalization of a model can be achieved by removing the impact of outliers. For example, regression-based models are usually affected by outliers, while tree-based models are not
   - Switching from mean squared error to mean absolute difference also helps reducing the impact of outliers.
   - Keeping the model as simple as possible to avoid overfitting, using Lasso instead of Ridge to make use of its feature selection.
   - Transforming the data to reduce skewness, for example, use a log transform for a skewed distribution of data.
   - Removing outliers in training data, those outliers could badly affect the accuracy of the model

   There are trade-offs between accuracy and robustness, we need to keep balance between them like the balance between bias-variance trade-offs, keep the model as simple as possible but don’t make it underfitting.