**Task 3 – Conclusion and Finalising Best Fit Model**

1. **Comparision and conclusion of Residuals distribution.**

|  |  |
| --- | --- |
| 1. **Model1 – Without SUB\_GRADE – Multiple Linear Regression** | 1. **Model1 – Without ADDR\_STATE – Multiple Linear Regression** |
| 1. **Model2 – Without SUB\_GRADE – Gradient Boost Tree Regression** | 1. **Model3 – Without SUB\_GRADE – Multiple Linear Regression – Polynomial Degree = 2** |
| 1. **Model4 – Model 3 With PARAM GRID SEARCH – 12 variations** | 1. **Comparision of R² (coefficient of determination).** 2. Model1 – Without SUB\_GRADE – Multiple Linear Regression - **94.78 %** 3. Model1 – Without ADDR\_STATE – Multiple Linear Regression - **99.43 %** 4. Model2 – Without SUB\_GRADE – Gradient Boost Tree Regression - **94.86 %** 5. Model3 – Without SUB\_GRADE – Multiple Linear Regression – Polynomial Degree = 2 – **94.82 %** 6. Model4 – Model 3 With PARAM GRID SEARCH – 12 variations – **94.65 %** |

* 1. Rug chart distribution shows better model fit if ‘addr\_state’ is not considered.
  2. Rug chart shows big variation between different states.
  3. R² value shows Model1 wihout ‘addr\_state’ feature to be the best.
  4. GBT regressor happens to be the best without ‘SUB\_GRADE’ feature.
  5. Also the features ‘annual\_inc’ and ‘loan\_amnt’ were transformed with LOG(ln) functions which makes the model prediction accuracy much better.