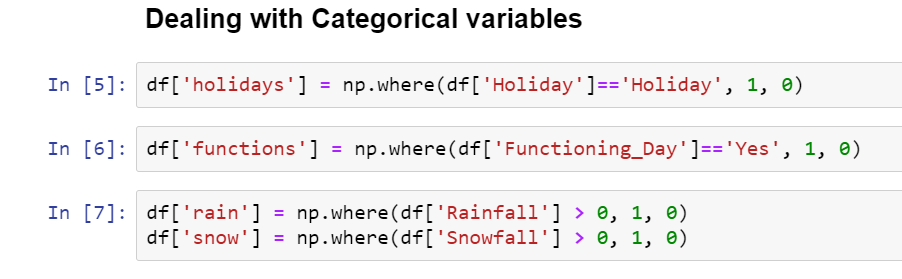
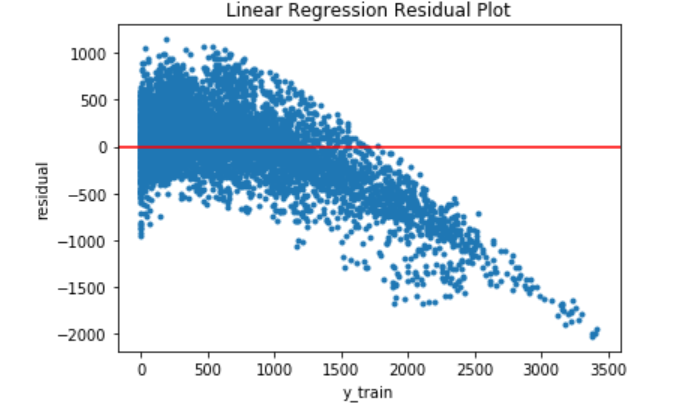
# In Depth Analysis

The data set consists of the count of hourly rental bike and its corresponding weather and datetime information. Hence, the first model that was used to give it a shot is linear regression model. When using linear regression model, it is necessary to transform the categorical variables into a set of true/false (binary) dummy variables. In addition to the “Holiday” and “functioning\_day” features, “Rainfall” and “snowfall” features are also transformed to binary variables for modeling.

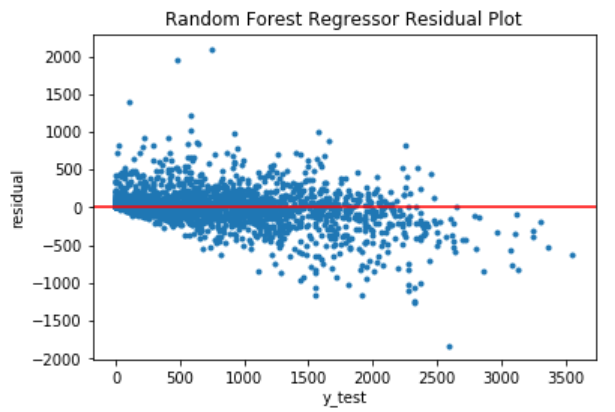


Linear Regression

Rented bike count is the used as the target variable, while all other columns become the features to the model. The training result for linear regression is not that optimal. The r-squared is showing 0.549, with some of the features, such as Visibility and the log of visibility, not showing significance in the model. Then, I took out those features with less t-values and re-train the model. The R-squared did not improve. The graph below is the residual plot on the training dataset. The mean absolute error is 321.93. In conclusion, linear regression model does not perform well for this problem.

Random Forest Model

The random forest regressor model, on the other hand, performs very well on predicting the rented bike count. With the default hyperparameters, the model is already scoring very high, with 0.97 on the training dataset, and 0.867 on the testing dataset. The mean absolute error is 139, comparing to the one from linear regression, the random forest is doing so much better. The graph below is the residual plot of the random forest model. The error is range from -2000 to 2000 in this case.



In order to make the model perform better, the hyperparameters tuning is necessary. GridSearch cross validation is used in the tuning. The new parameters are improving the mean absolute error from 139 to 134.