Pandas, get\_dummies, and feature engineering was utilized to get a better fit model due to our dataset size. States and years were transformed into categorical data. Testing and training data was preprocessed through scaling to normalize the data within a smaller range. We ran our analysis our residual plot. Prior to that, we decided to run linear regressions to see the correlation between income, bachelor’s degree and post-secondary degree. Actual and predicted data was plotted on the same scatterplot.

Chart, scatter chart

Description automatically generated

This plot above shows the actual and predicted data follows the same pattern, showing a strong correlation between postsecondary and income.

Chart, scatter chart

Description automatically generated

This plot above shows the actual and predicted data follows the same pattern, showing a strong correlation and a strong trend between Bachelors and income.

Chart, scatter chart

Description automatically generated

The R2 of .80 indicates that the model explains about 80% of the variability of the data around the mean. The mean scored error, which shows the averaged squared difference between the predicted and actual values was .09, close to the desired 0 range. Overall, the probabilities of a positive correlation between bachelors percentage, post secondary and per capita income combined are relatively high