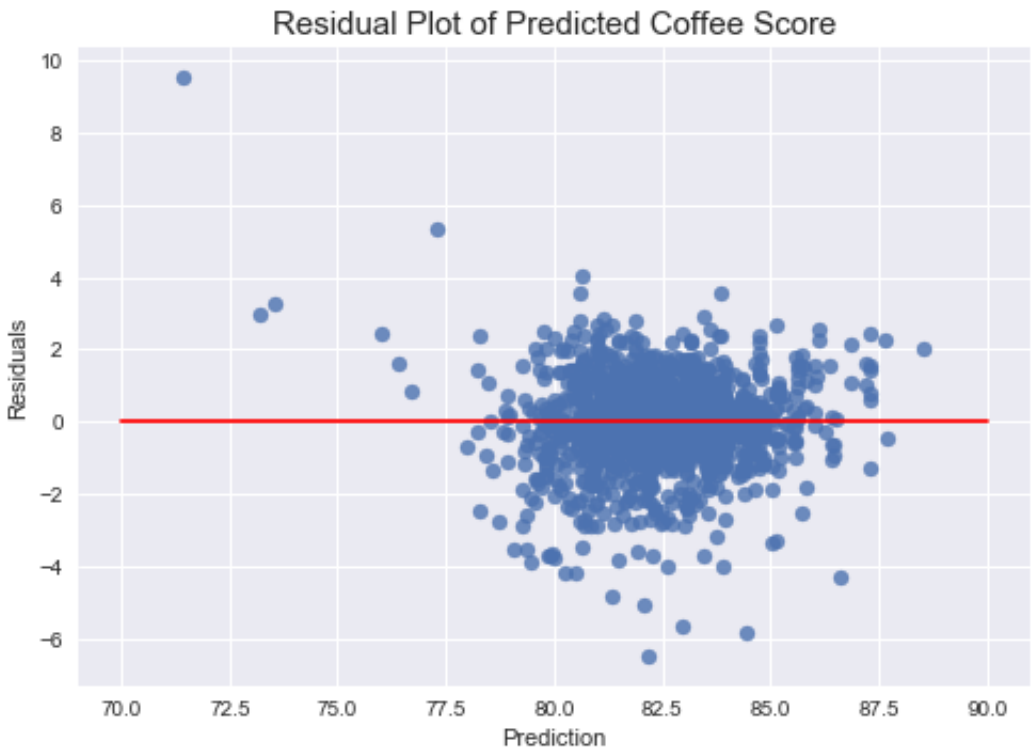
Predict Coffee Quality Without Having to Taste It

The goal of this project is to predict the coffee quality score from the Coffee Institute without having to taste the coffee. The scores are in the range of 0 - 100, modeled using attributes obtained from the Coffee Institute . If time allows, I will incorporate data on the total killigram exported by each country from the International Coffee Organization, which I can use to calculate the market share of each producer.

Attributes Obtained from the Coffee Institute:

Categorical Variables	Quantitative Variables
Country	Coffee Quality Score (Target Variable)
Farm	Bag weight
Mill	Number of bags
Company	Altitude
Region	Acidity
Producer	Sweetness
Owner	Moisture
In_country_partner	Quakers
Variety	Defects Category 1
Harvest_year	Defects Category 2
Grading_date	
Processing_method	
Color	

I will be modeling using Ordinary Least Squares for this project, incorporating cross validation, dummy variables, potentially regularization, and treatment of outliers and missing values. On first pass to test the viability of this project, I passed the quantitative variables to predict Total Coffe Score, my results show the following residual plot. It appears the lower Coffee Scores have a large residual, signaling that I need to pay attention to the outliers in the lower end.



Questions

Any areas of concern? Suggestions for handling outliers and missing values?