Data Analysis Questions

1. (Linear Regression) When predicting when the year a particular Programming Language was released, which predictors (Active Repositories, Total Pushes, Total Pushes, Pushes Per Repository, New Forks Per Repository, Opened Issues Per Repository, and New Watchers Per Repository) improves the accuracy of the model’s prediction when excluded from the model? What does this suggest about the nature of the Programming Languages used in GitHub repositories?
2. (Clustering) When considering the number of Opened Issues Per Repository and the number of New Watchers Per Repository, what clusters emerge? What can be said about those clusters? What percentage of those clusters contain Programming Languages that were released prior to 1993?
3. (Feature Reduction) Apply LASSO to the Linear Regression model created in Question 1 and based on the variables that LASSO determines to have a coefficient, create another Linear Regression model. How did this effect the original model? What does this tell you about the GitHut data regarding Programming Languages used in GitHub repositories?