Supervised Learning Capstone

World Migration

Contents

- Data
- Explorations and Correlations
- Models
    - Linear Regression
        - Ridge
        - Lasso
    - Support Vector Regression
    - Random Forest
    - Gradient Boosting
    - KNN Regression

Data

Source

Dataset found at https://www.kaggle.com/fernandol/countries-of-the-world

As taken from 2013 US government data gathered in the World Factbook found at https://www.cia.gov/library/publications/the-world-factbook/docs/faqs.html

Each observation is a record of an individual country featuring diverse attributes

Data

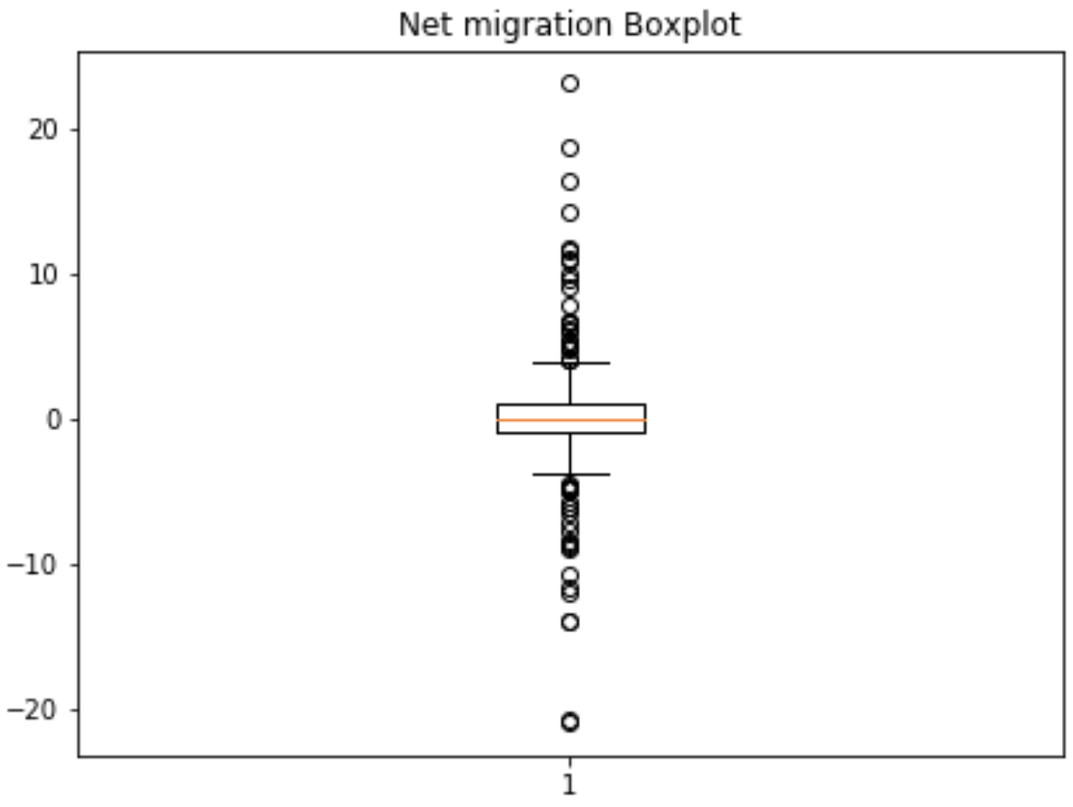
Columns:

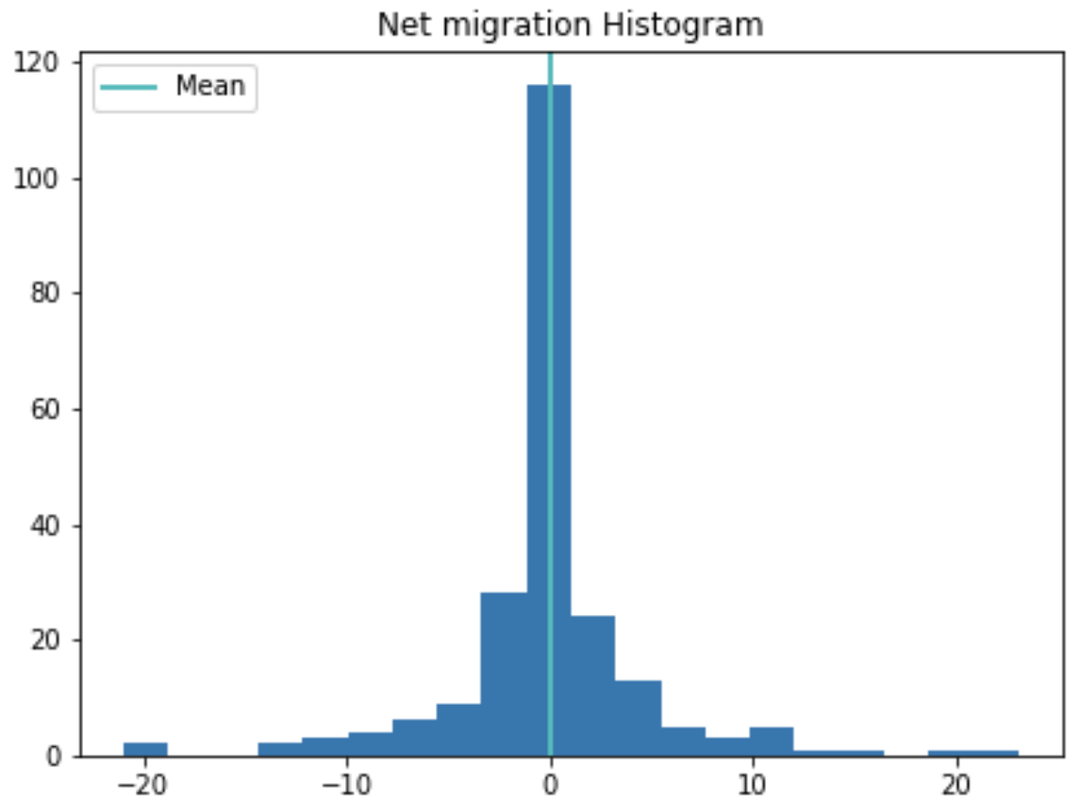
- ‘Country'
- ‘Region'
- ‘Population'
- 'Area (sq. mi.)’
- 'Pop. Density (per sq. mi.)’
- 'Coastline (coast/ area ratio)’
- 'Net migration’
- 'Infant mortality (per 1000 births)’
- 'GDP ($ per capita)’
- 'Literacy (%)’
- 'Phones (per 1000)’
- 'Arable (%)’
- 'Crops (%)’
- 'Other (%)’
- ‘Climate'
- ‘Birthrate'
- ‘Deathrate'
- ‘Agriculture'
- ‘Industry'
- 'Service'

Net Migration

Normally distributed variable with heavy tails and mean of 0.

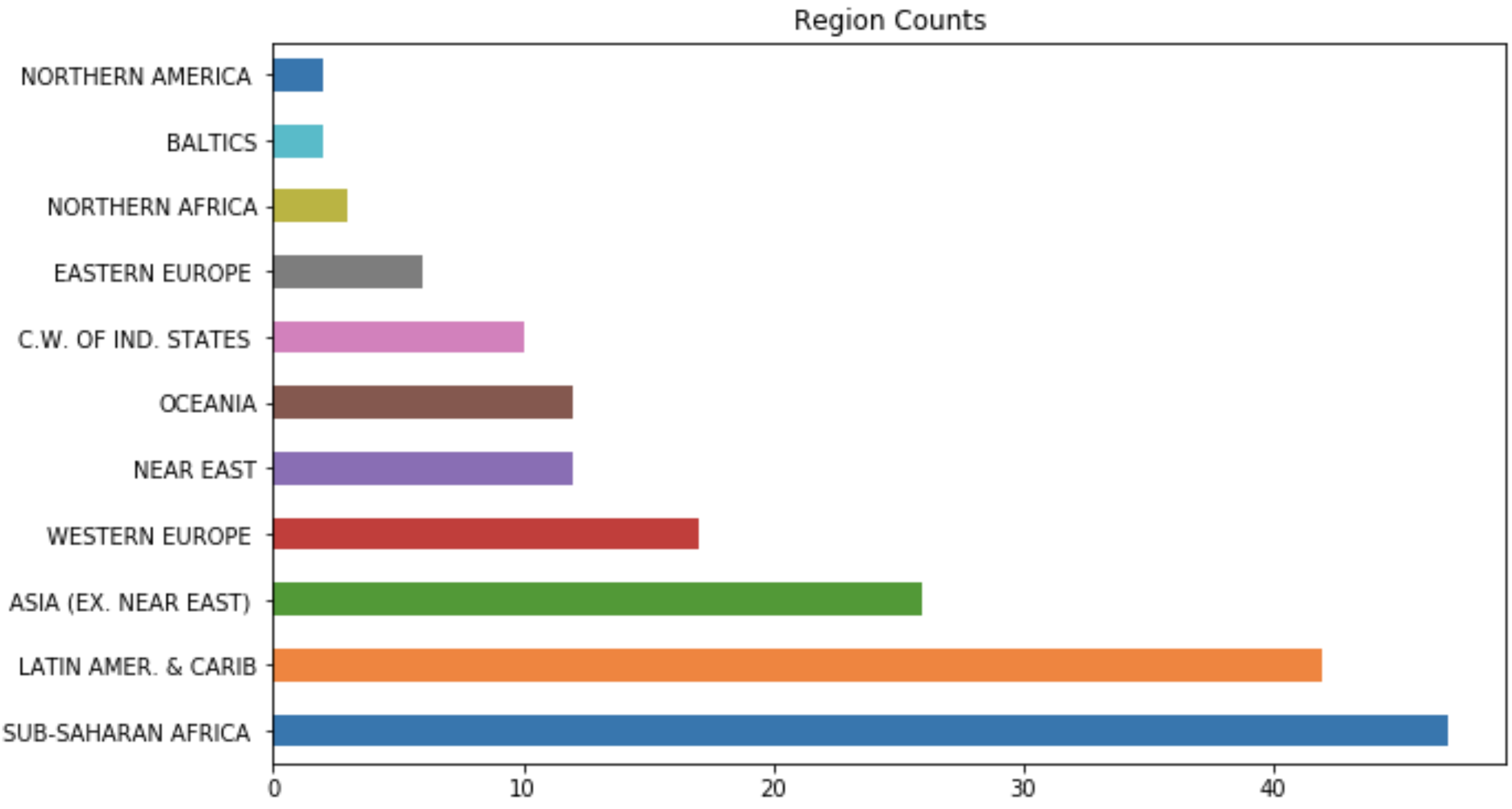
Several notable outliers on both high and low ends, maxing out at +20 and -20 from Afghanistan and Micronesia, respectively.



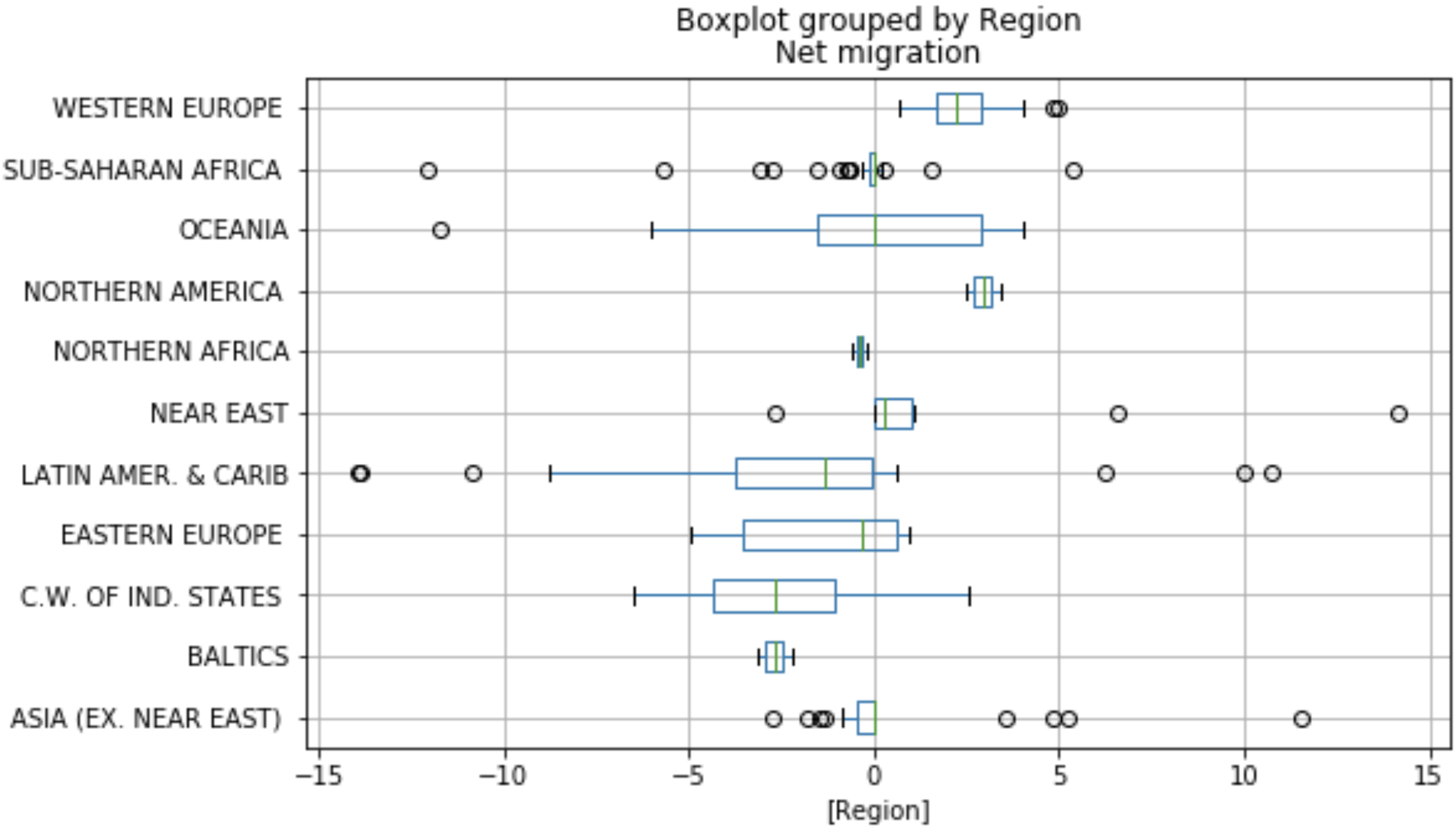


Region

Highest number of countries in Sub-Saharan Africa and Latin America & Caribbean, of the 179 total countries.



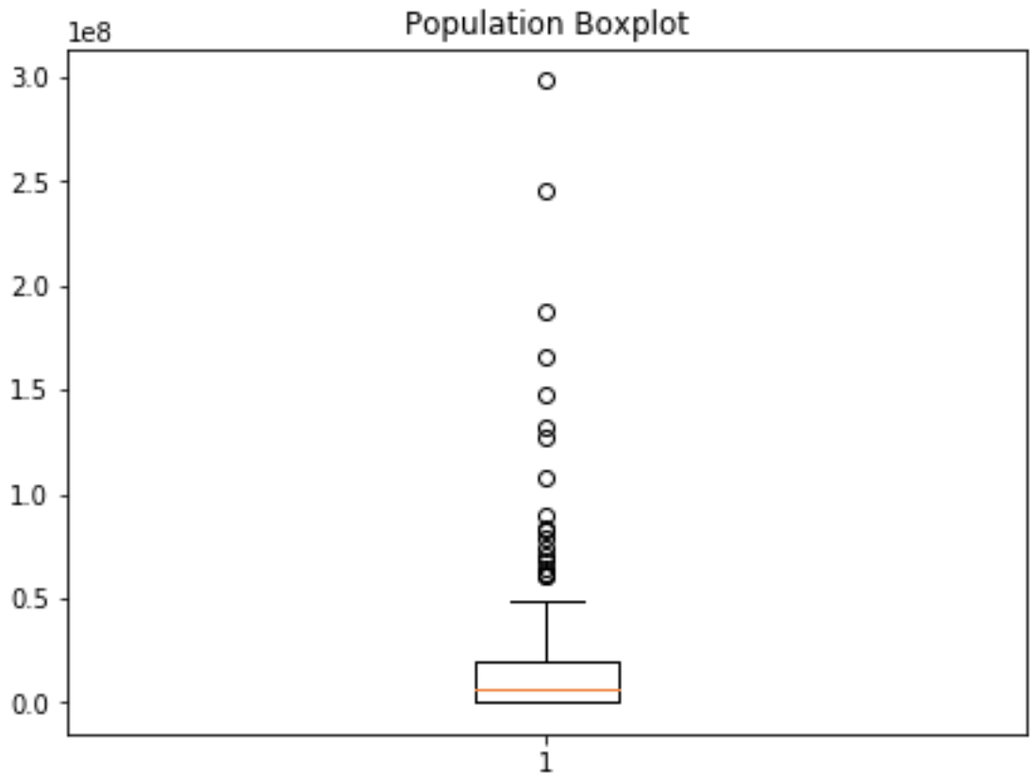
Net Migration by Region

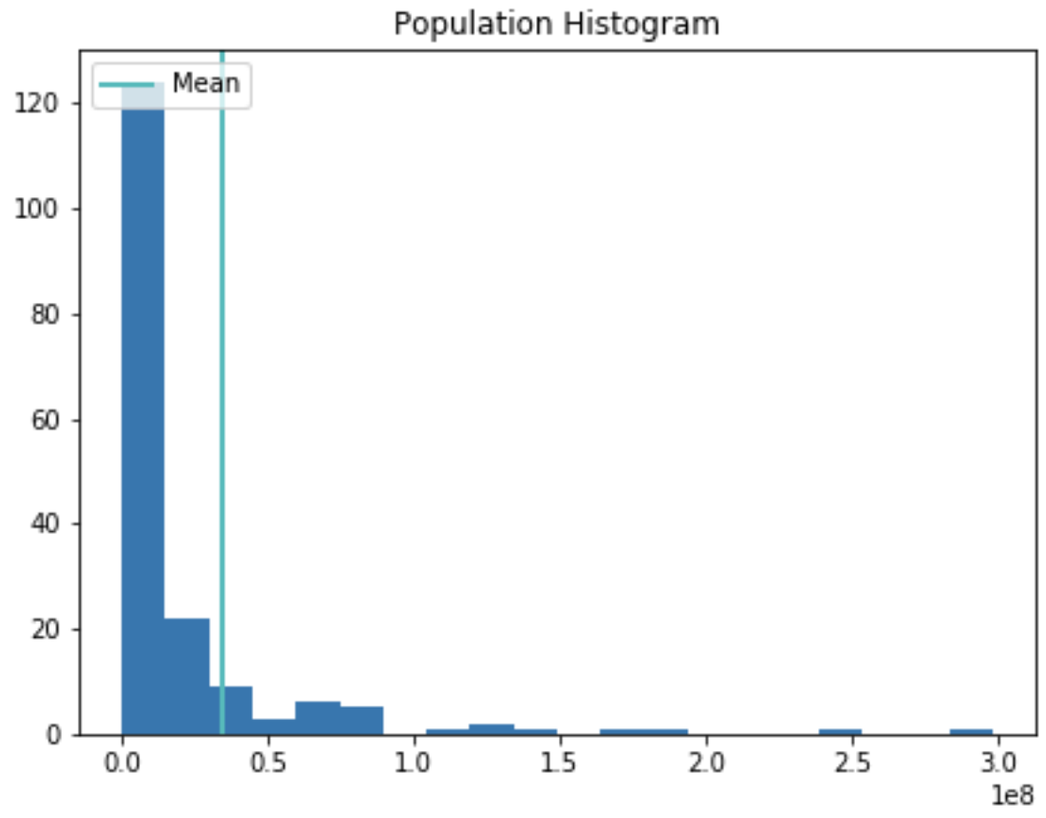


Outliers with net migration beyond +/- 15 have been removed from the visual (4 countries).

Population

Heavily right skewed variable, with most countries having low population, well below the mean of 34 million.

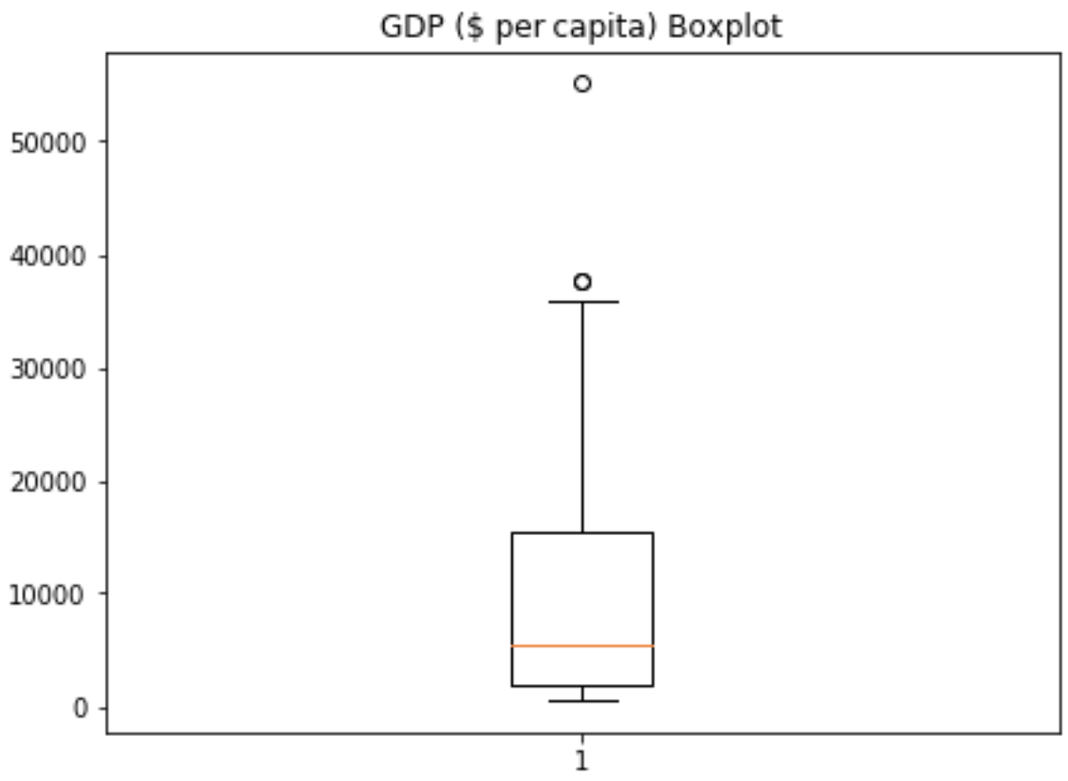


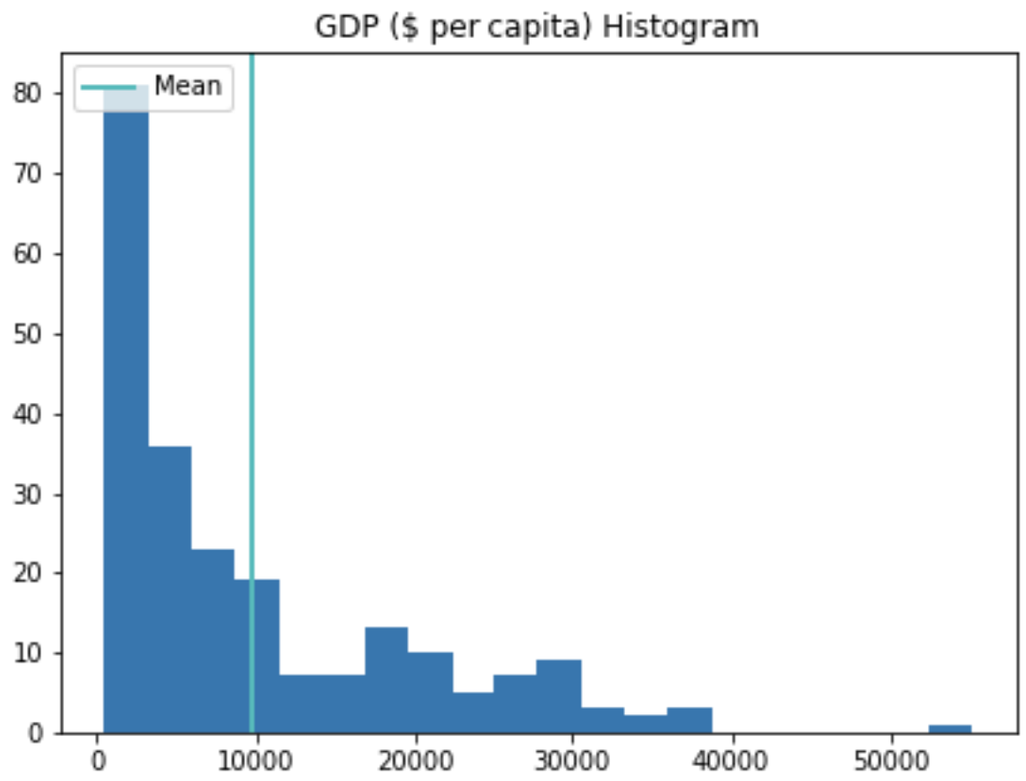


Notable outliers of China and India, each with populations of over one billion, are not included in the image

GDP ($ per capita)

Heavily right skewed, most nations have very low GDP, well below the mean of $9126.

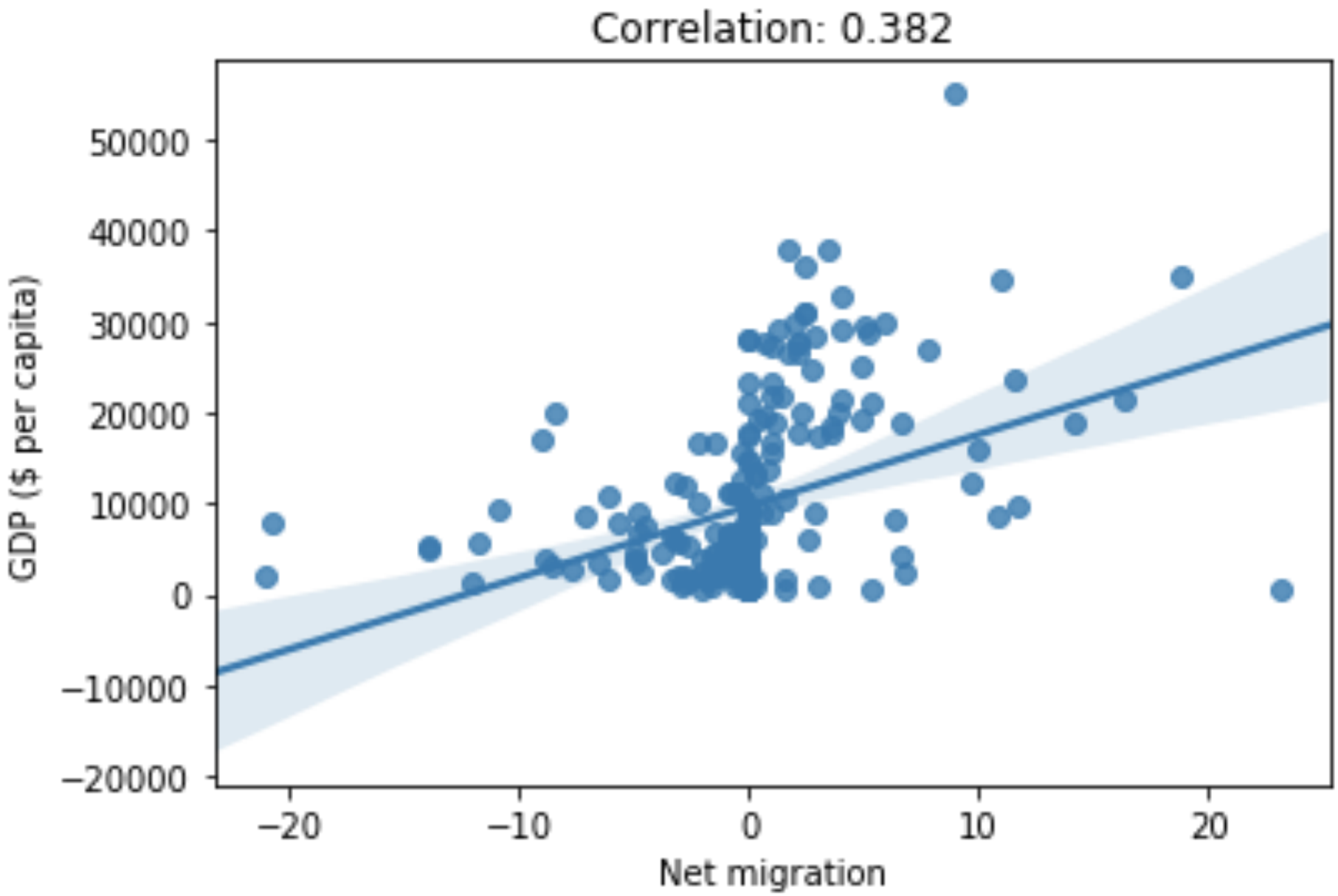




Notable outliers with very high GDP are Luxembourg, Norway and United States and are not included in the image.

Correlation: Net Migration, GDP

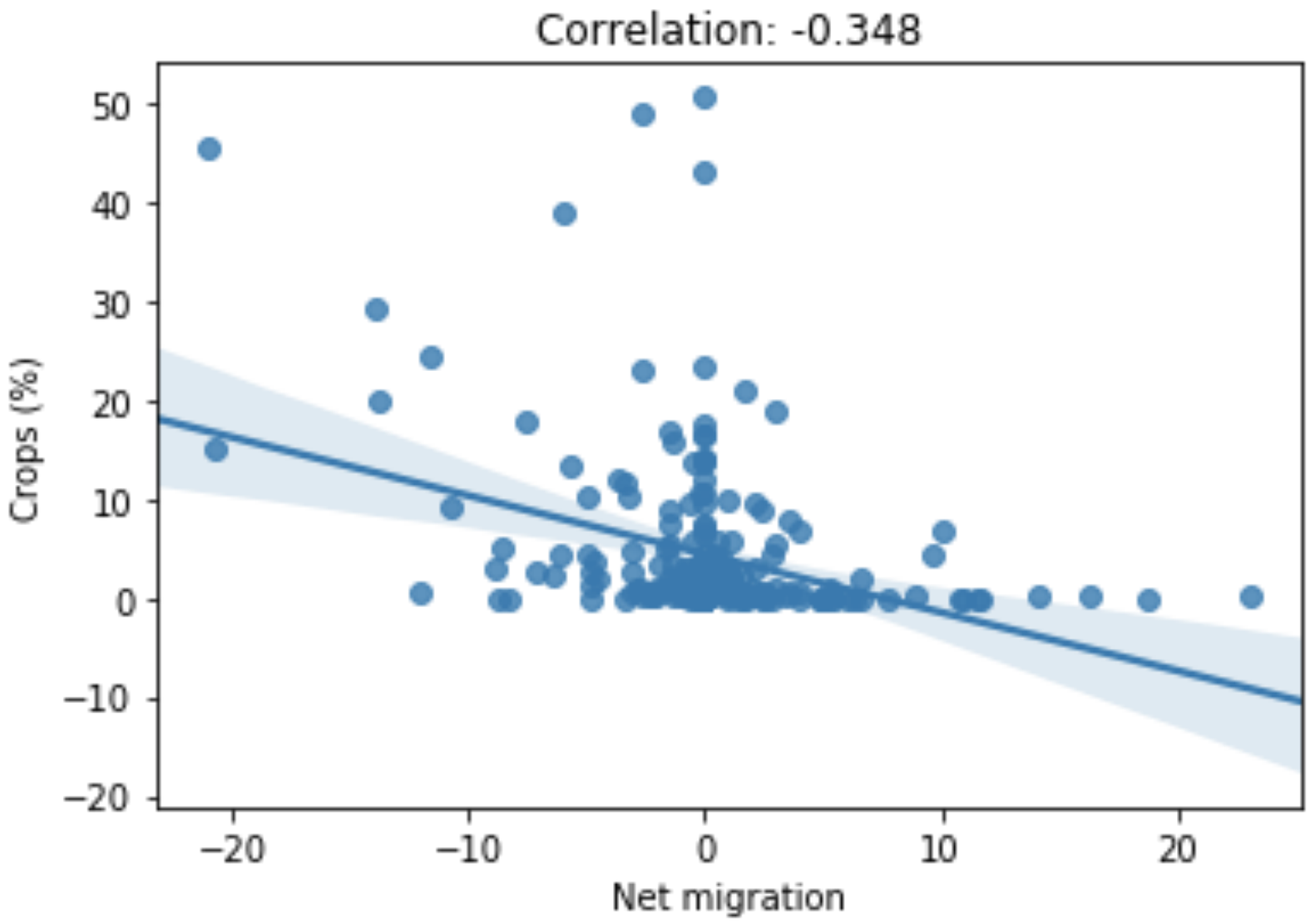
Positive correlation
Increased GDPs show increased net migration



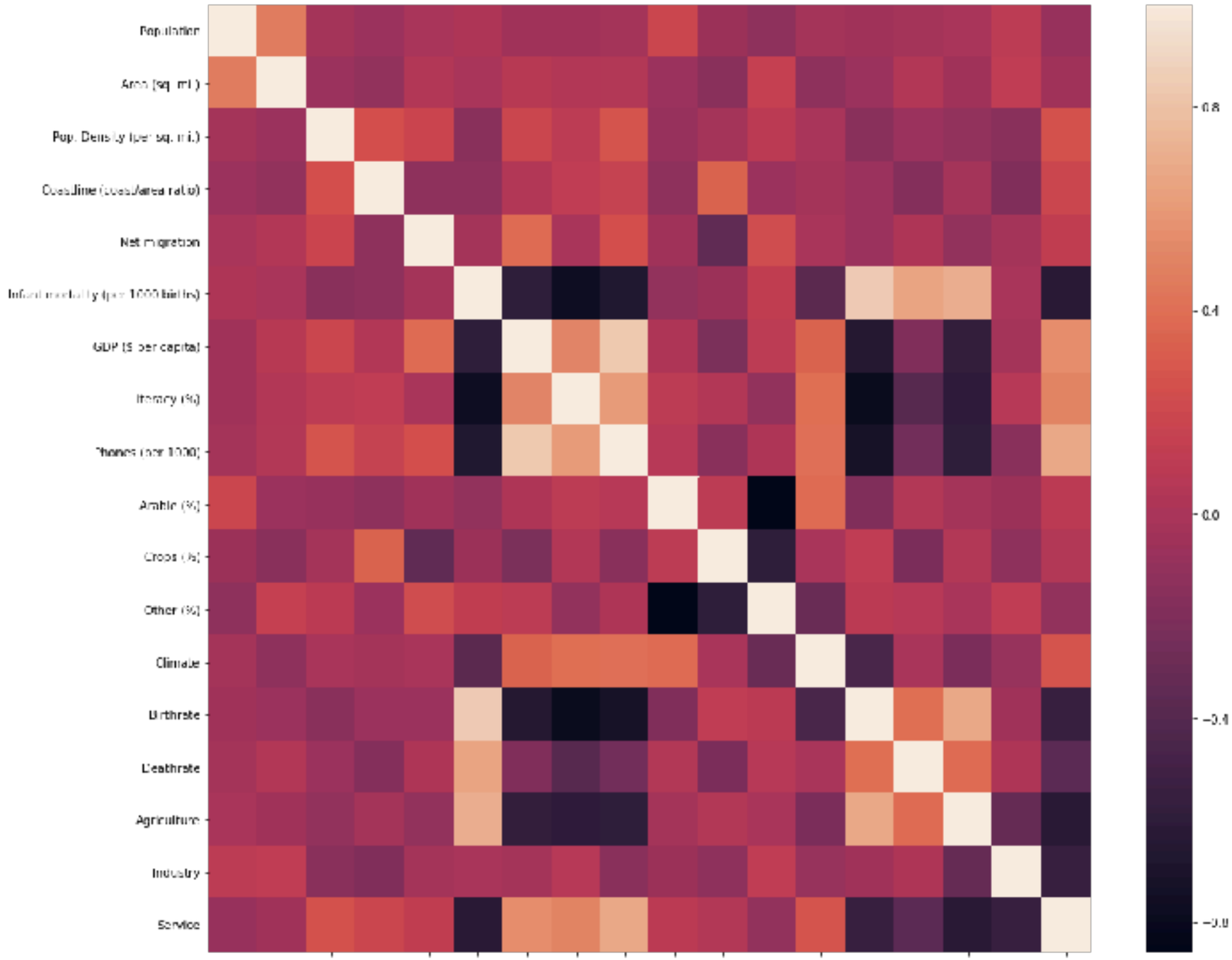
Correlation: Net Migration, Crop Land

Negative correlation

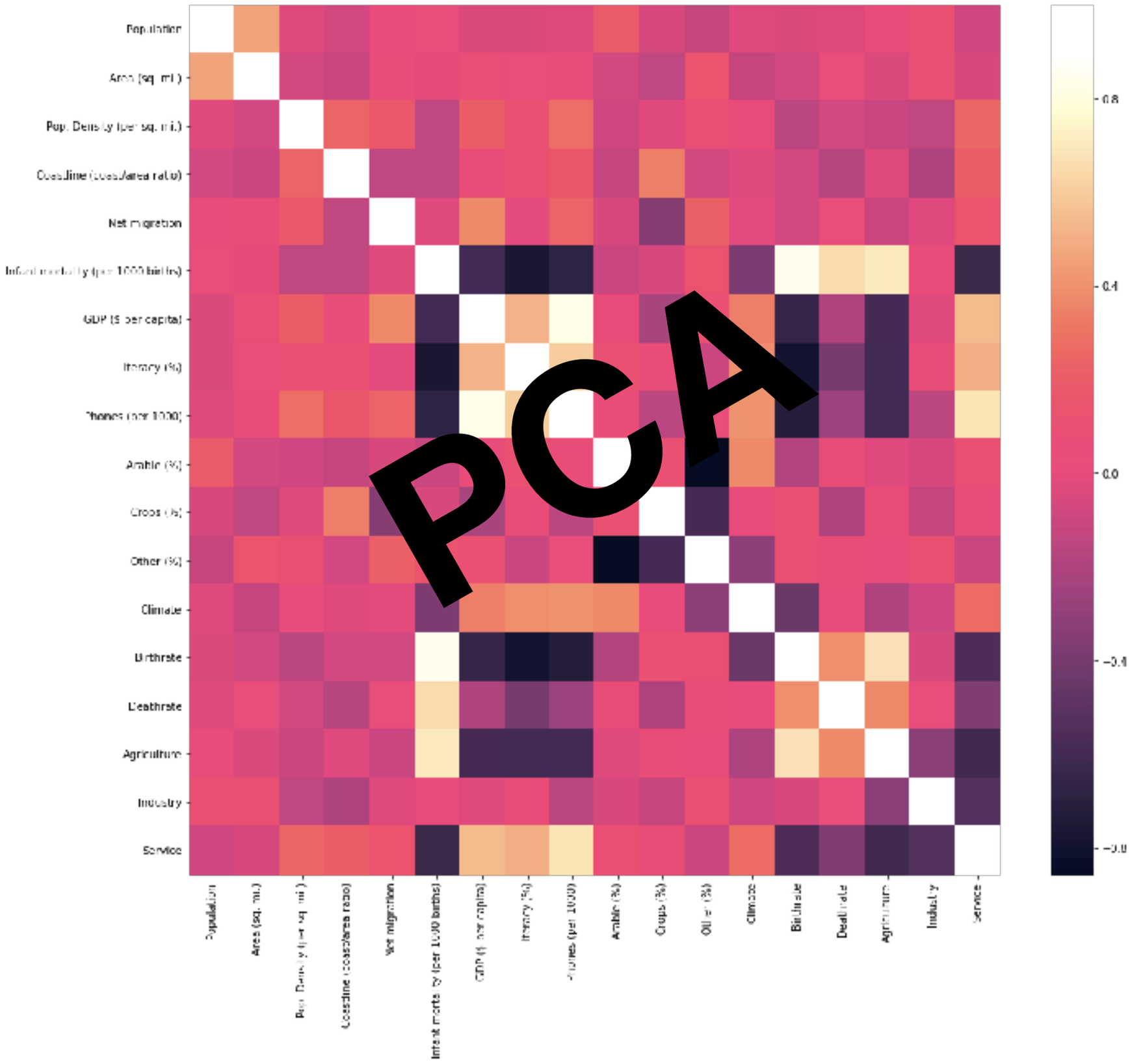
Countries using less land for crops, see increasing net migration



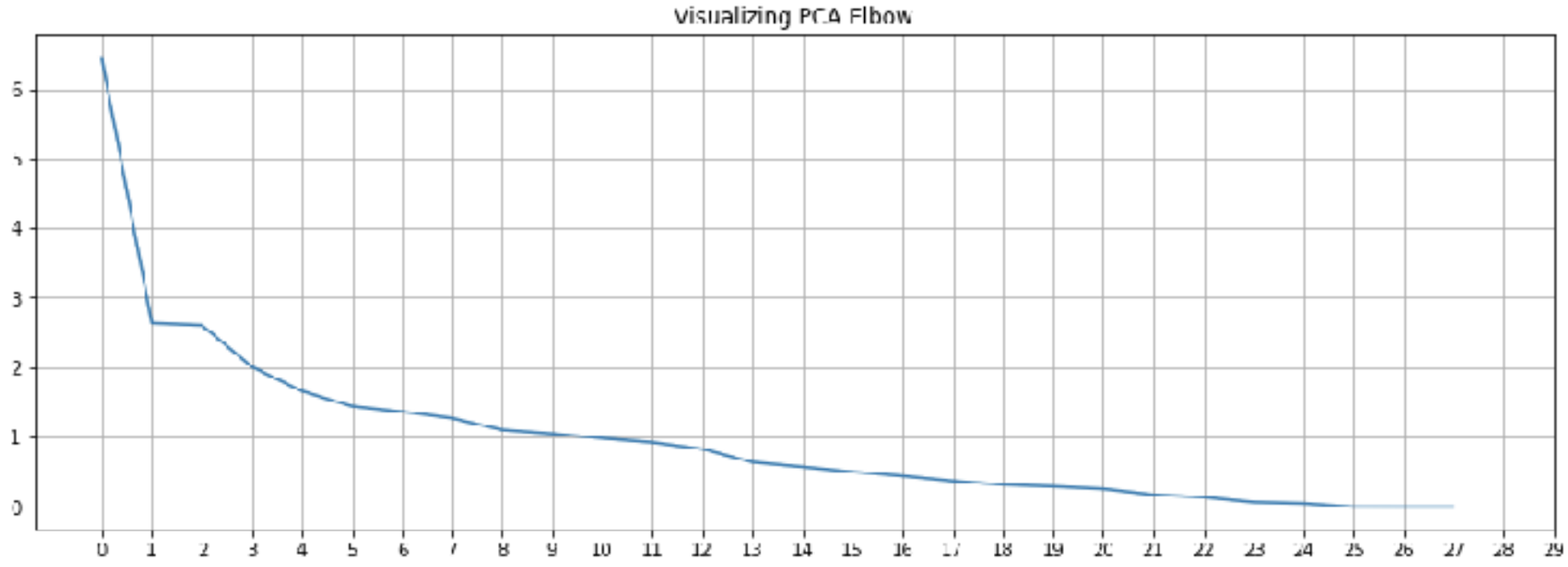
Correlation Matrix



Correlation Matrix



Principal component analysis



Elbow plot shows a sharp bend from 1 and 2 components. Two components will be used to fit models from PCA.

This reduction in features will also reduce the time to fit each model.

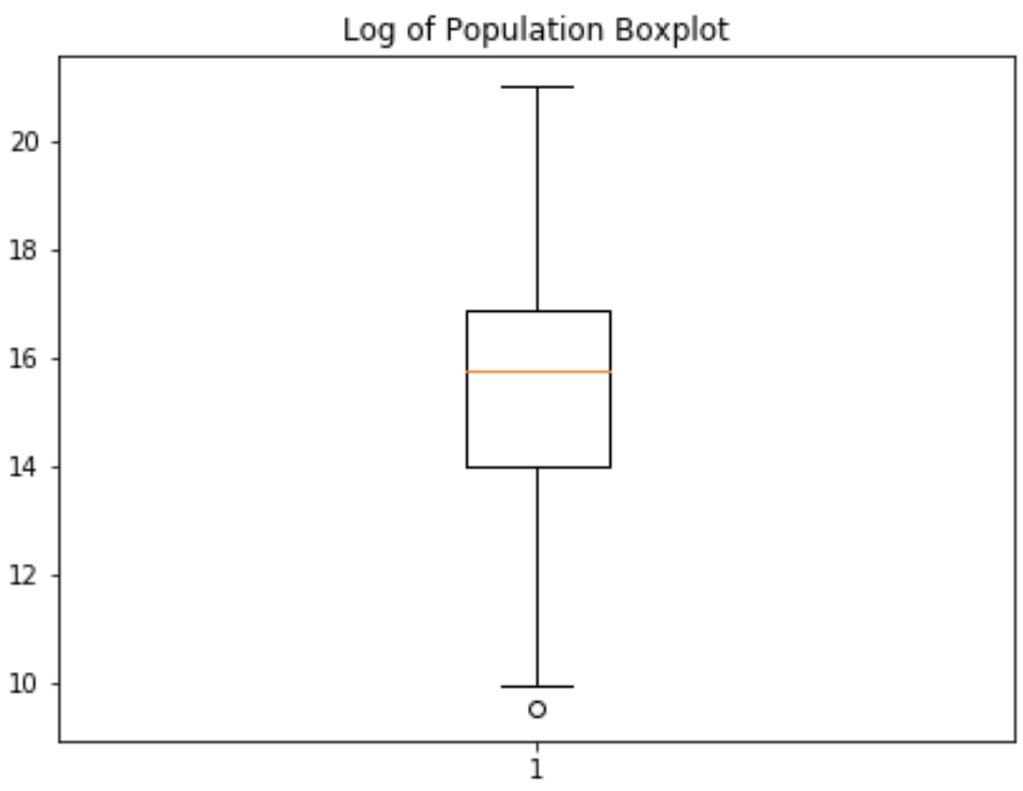
Feature Selection and Transformation

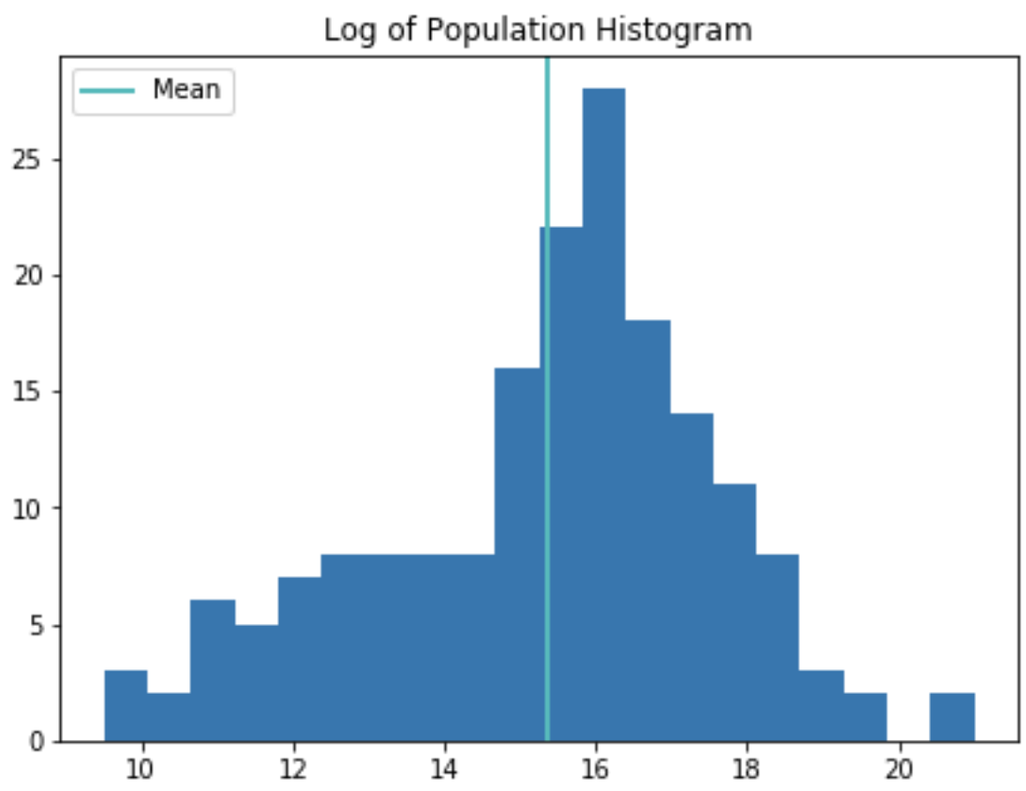
- Transforming Population with logarithm

- Random Forest feature selection

Log of Population

Transforming the population variable with a numpy logarithm, the variable becomes normally distributed, with a mean just over 15.





Random Forest Feature Selection

Lowest Scored Features

	features	score
21	Region_LATIN AMER. & CARIB	0.007326
10	Climate	0.004836
19	Region_C.W. OF IND. STATES	0.003754
17	Region_ASIA (EX NEAR EAST)	0.002199
22	Reg ion_NEAR EAST	0.000692
18	R egion_BALTICS	0.000631
27	Region_WE STERN EUROPE	0.000435
26	Region_SUB-SA HARAN AFRICA	0.000279
20	Region_EA STERN EUROPE	0.000189
25	R egion_OCEANIA	0.000162
23	Region_NO RTHERN AFRICA	0.000015
24	Region_NORT HERN AMERICA	0

Random Forest's built in feature scoring shows low significance in regions. It is worth testing models without those region features.

Models

Regression

Target variable: Net Migration

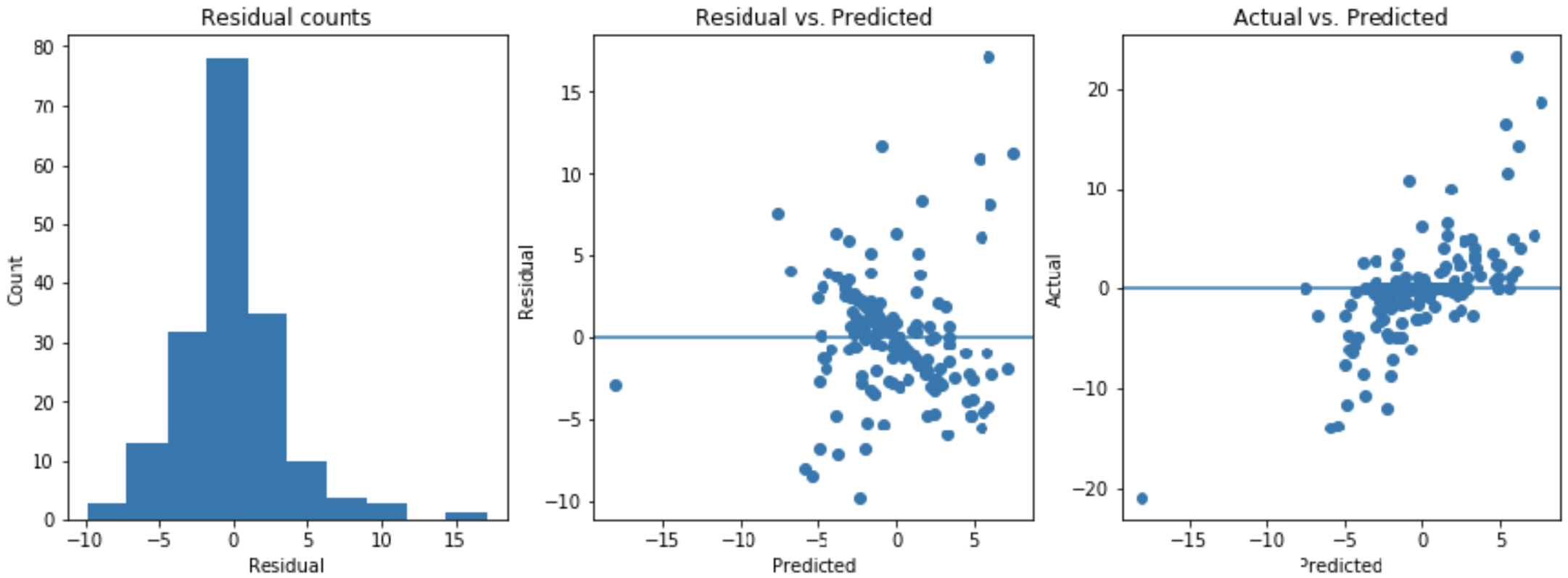
Feature variables: Numerical variables of dataset tested with and without added columns for numerical representation of categorical variable of region, or using two components from PCA

Linear Regression

- Using Ordinary Least Squares as the cost function
- Ridge regularization to shrink parameter estimates
- Lasso regularization to force small parameter estimates to zero, effectively dropping features

Regression Results

R-squared: 45%

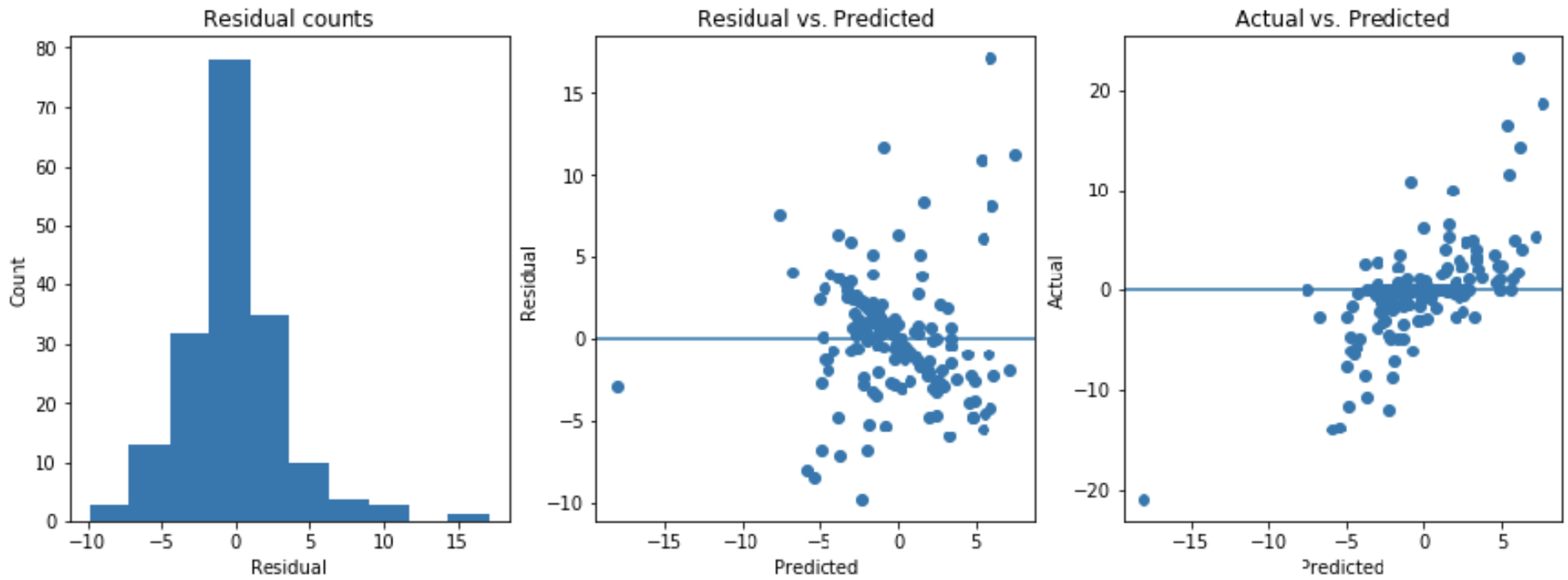


Using basic linear regression, the residual range of ~25 is rather large, with a fairly small R-squared.

This is not a very good fit for the data.

Regression Results without Region Features

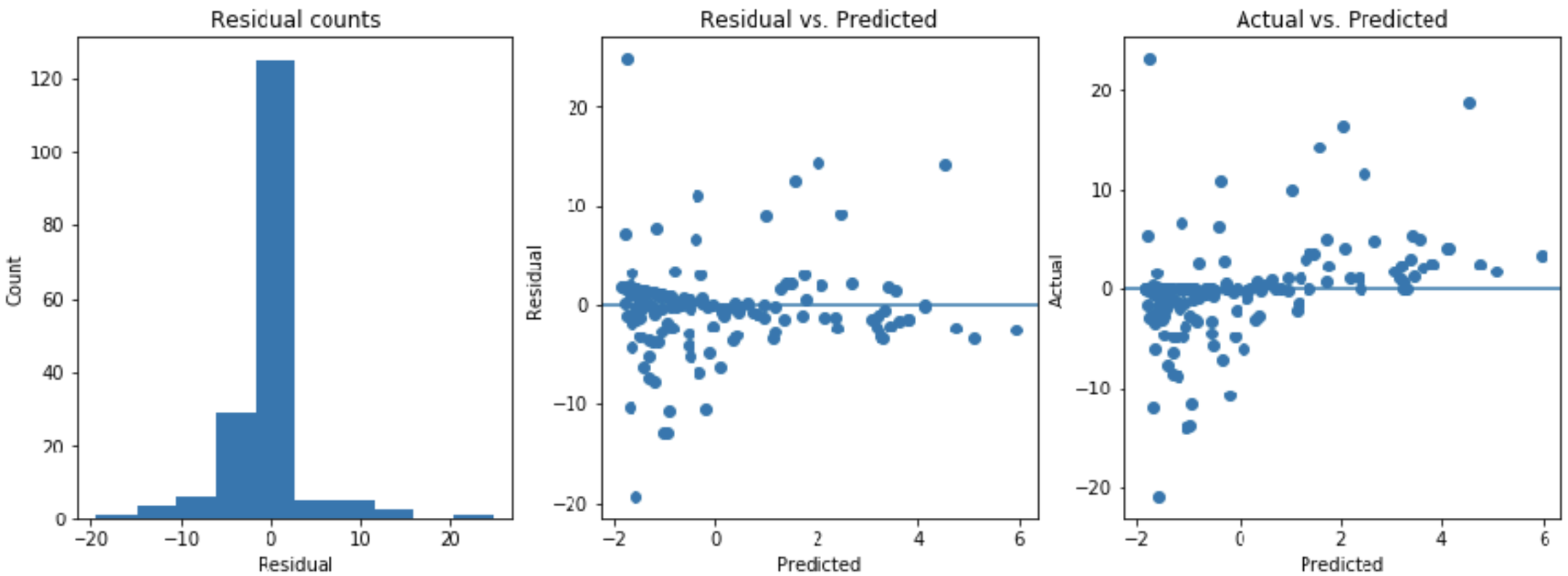
R-squared: 45%



As indicated in the Random Forest feature grading, there is no loss in accuracy when removing the region variables from the model. There is a very small reduction in time (0.004 seconds), not enough to be a factor.

Regression Results with PCA

R-squared: 14%

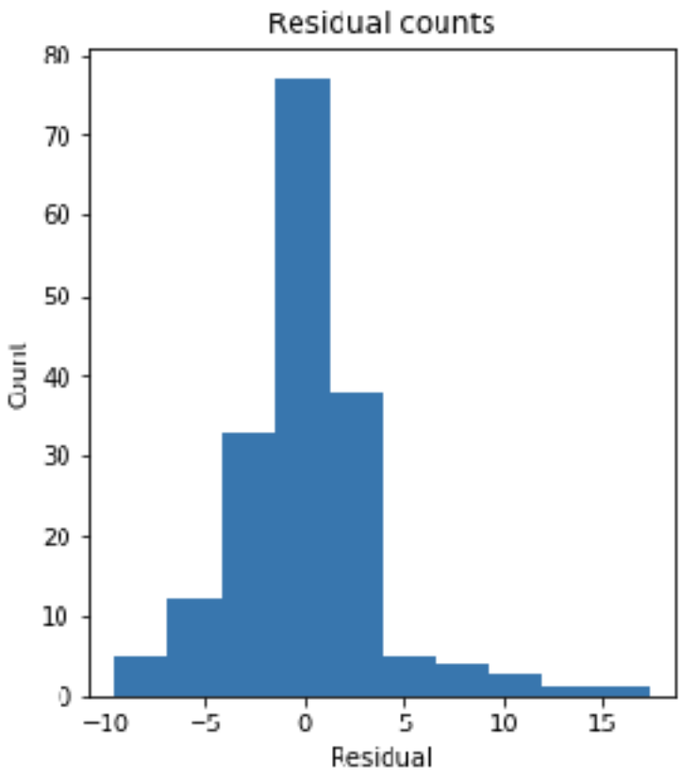


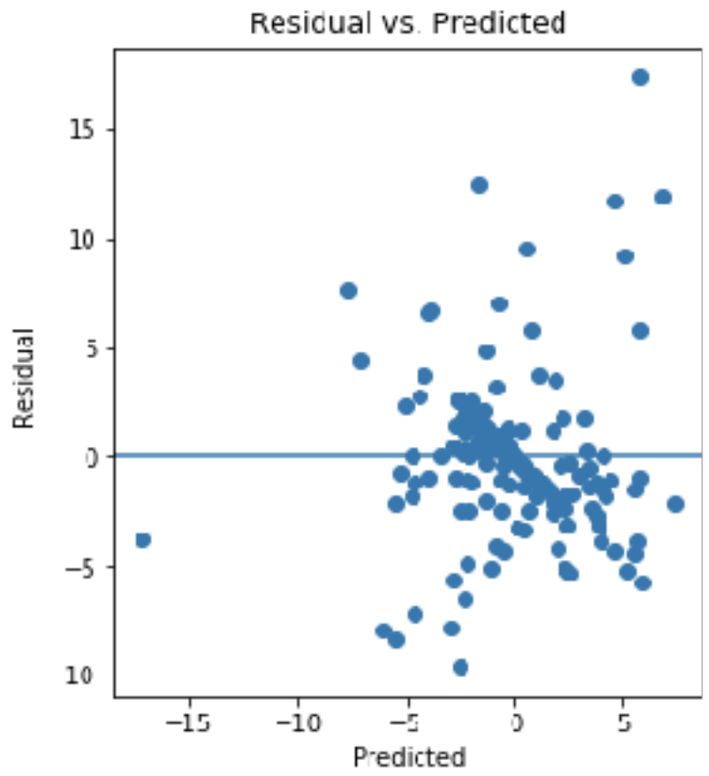
Using basic linear regression with PCA, the residual range increases to ~40 and the R-squared shrinks to 14%. The only improved metric is time, with a savings of 0.1 seconds.

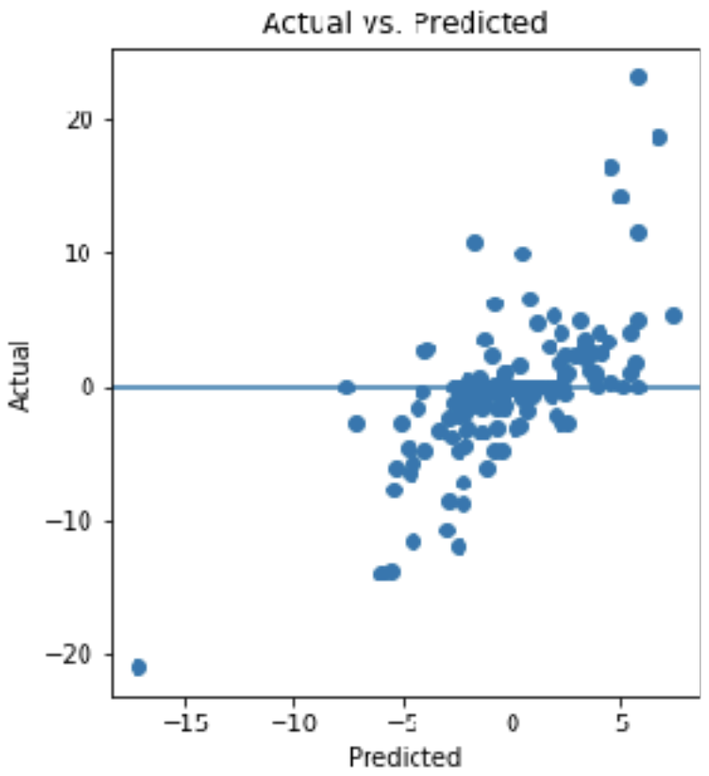
This is bad fit for the data.

Ridge Results

R-squared: 43%





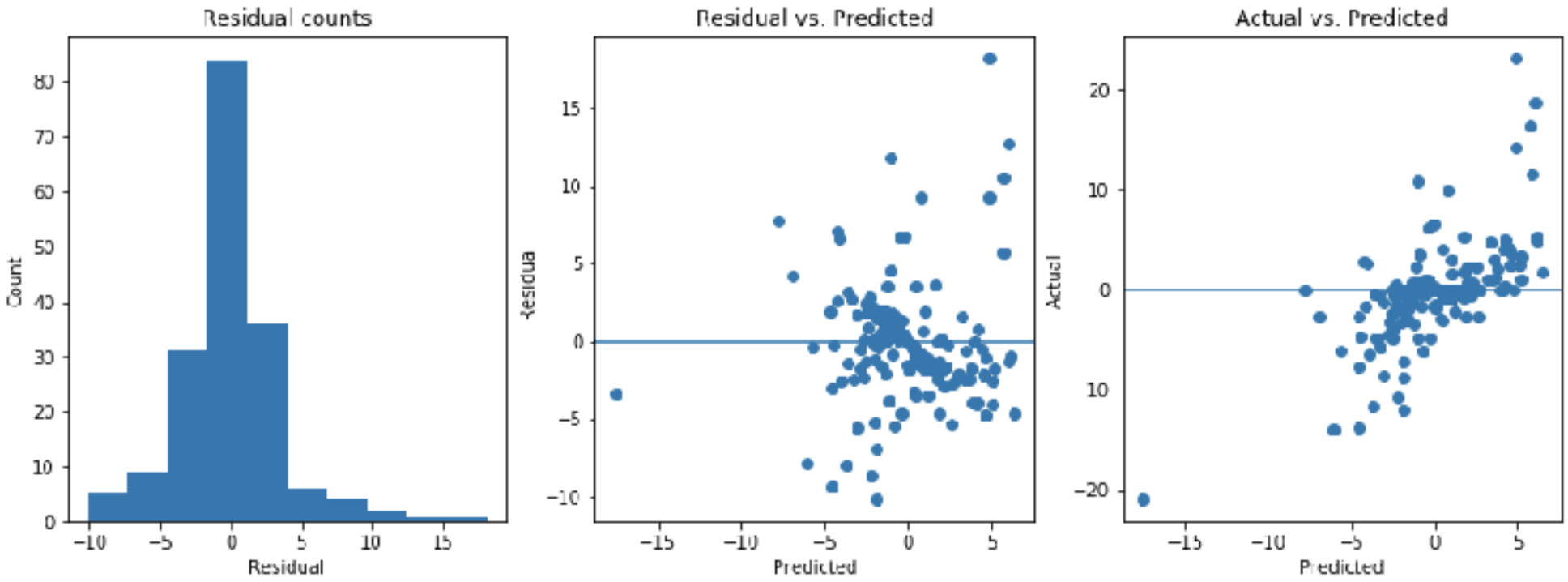


Coefficients are already too small for Ridge regularization to have an effect - results are almost identical. This is true with PCA as well as both sets of features (with and without regions).

This is not a good fit for the data.

Lasso Results

R-squared: 42%

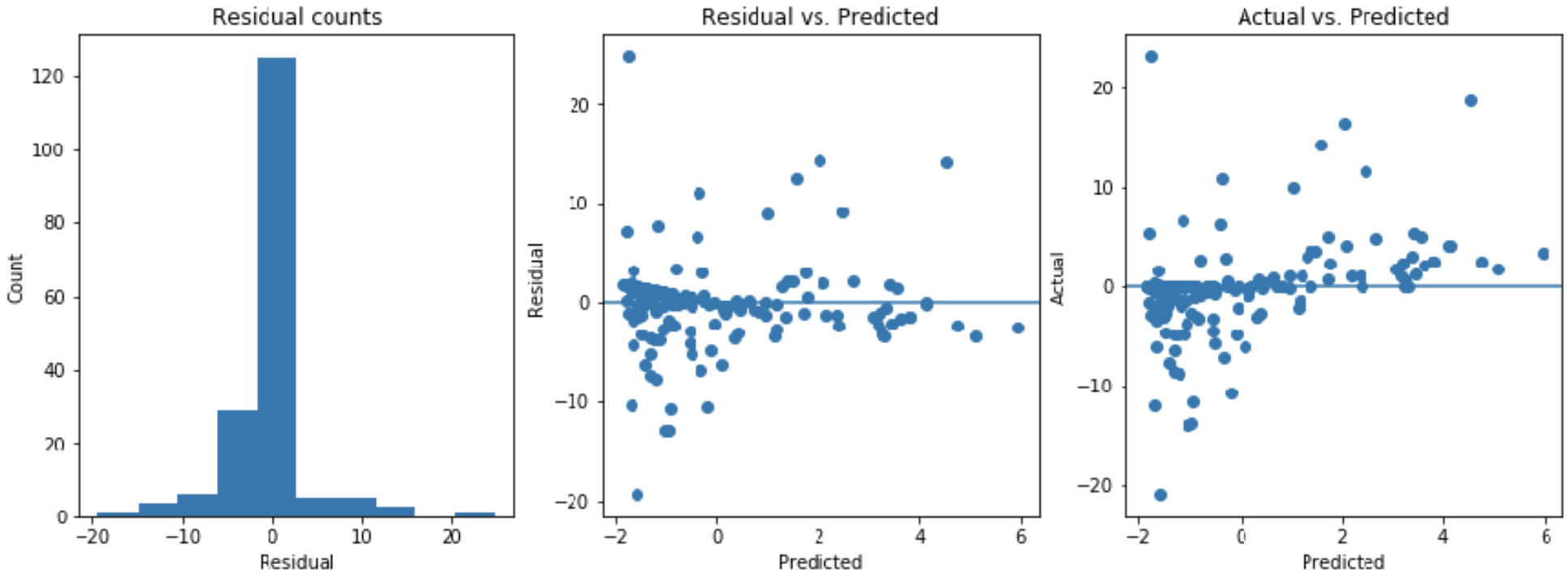


Lasso regularization is optimized with fit-intercept and lambda of 0.04. Most residuals are minimized, but the range is still at ~25. The R-squared again similar to regression without regularization.

Lasso is not a good fit for the data.

Lasso Results with PCA

R-squared: 14%

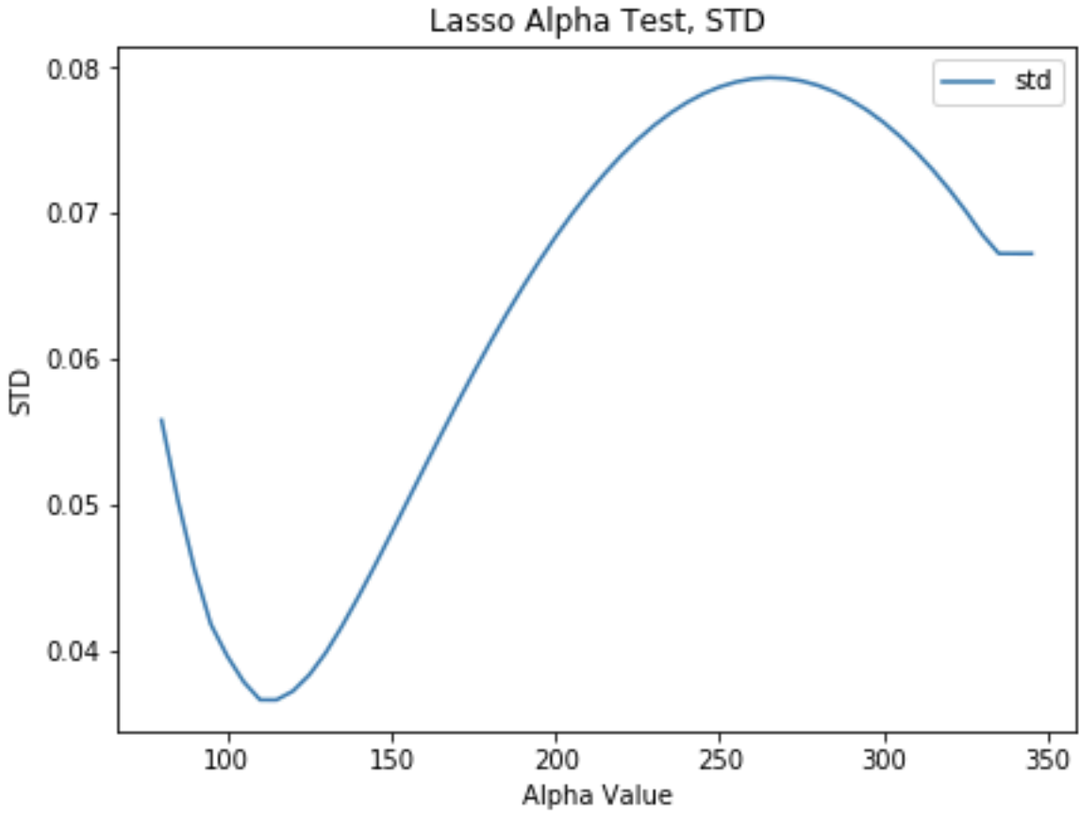


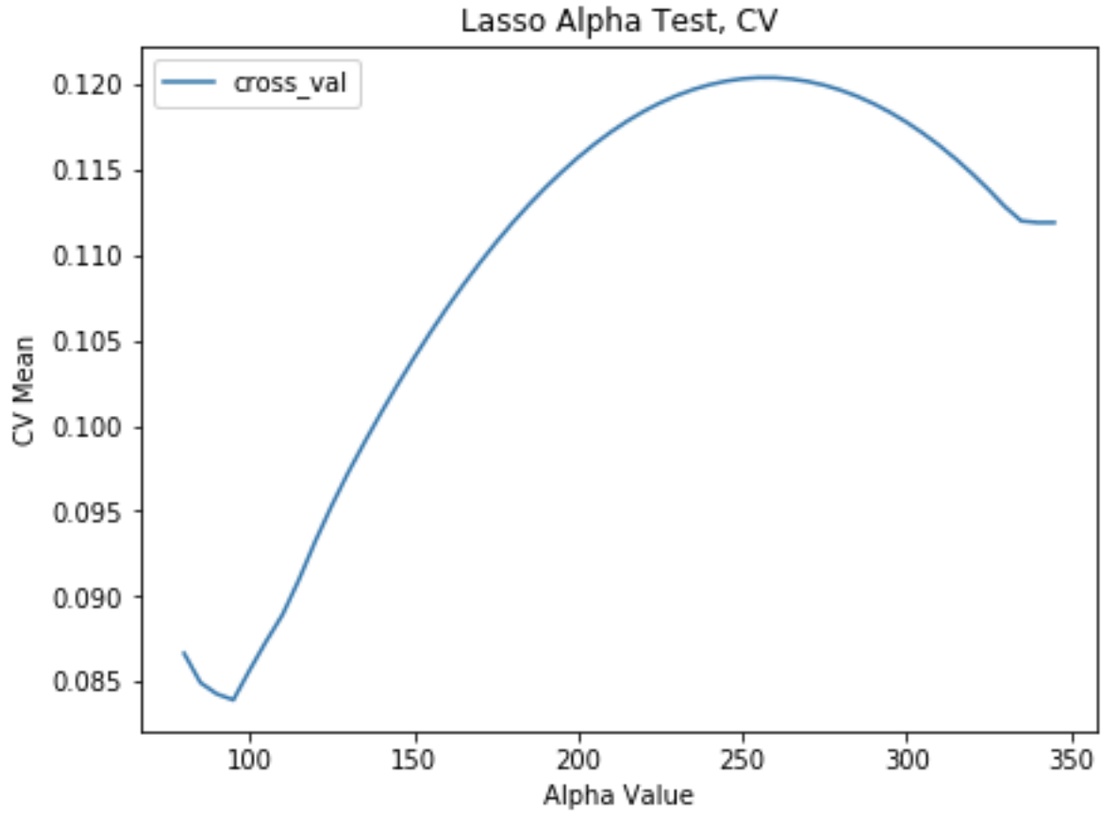
Lasso regularization on PCA components is optimized with fit-intercept and lambda of 0.09. The result is a low R-squares and large residual range of over 40.

Lasso is a very poor fit for the PCA data.

Tuning Lasso

When tuning the settings for Lasso, there is a tradeoff between high accuracy and low standard deviation.





Grid Search CV

```
# set parameters for tuning with grid search cv
parameters = {'C': [0.1, 1, 10],
    'epsilon': [0.01, 0.1, 1]}
svr = SVR()
svr_cv = GridSearchCV(svr, parameters)
```

Model tuning using GridSearchCV from SciKitLearn

- Setting range of parameters to be tested
- Cross validation scores are compared to identify the best combination for the model.

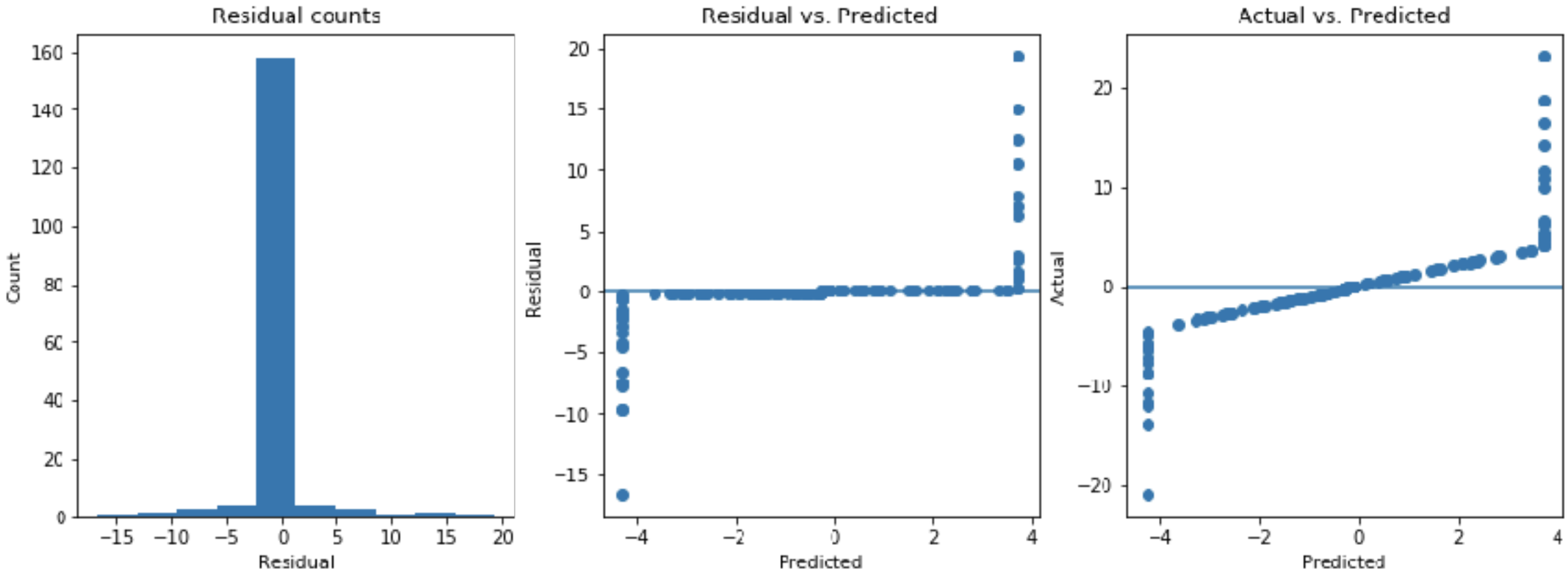
Support Vector Machine

Tuning the penalty for values outside the margin (C) and setting the margin size (epsilon) to set the sensitivity of the model

Optimized with C set to 0.01 and epsilon to 0.1

SVR Results

R-squared: 58%



The accuracy is very good in predicting mid-range values, but fails on the extremes, resulting in a large residual range. This is true with features as well as with PCA (when R-squared is 57%).

Support Vector Machines Regression is not a good fit for the data.

Decision Tree

The nature of decision trees leads to inconsistent performance, as each tree is potentially built differently with every fit.

The best results came with the default settings:

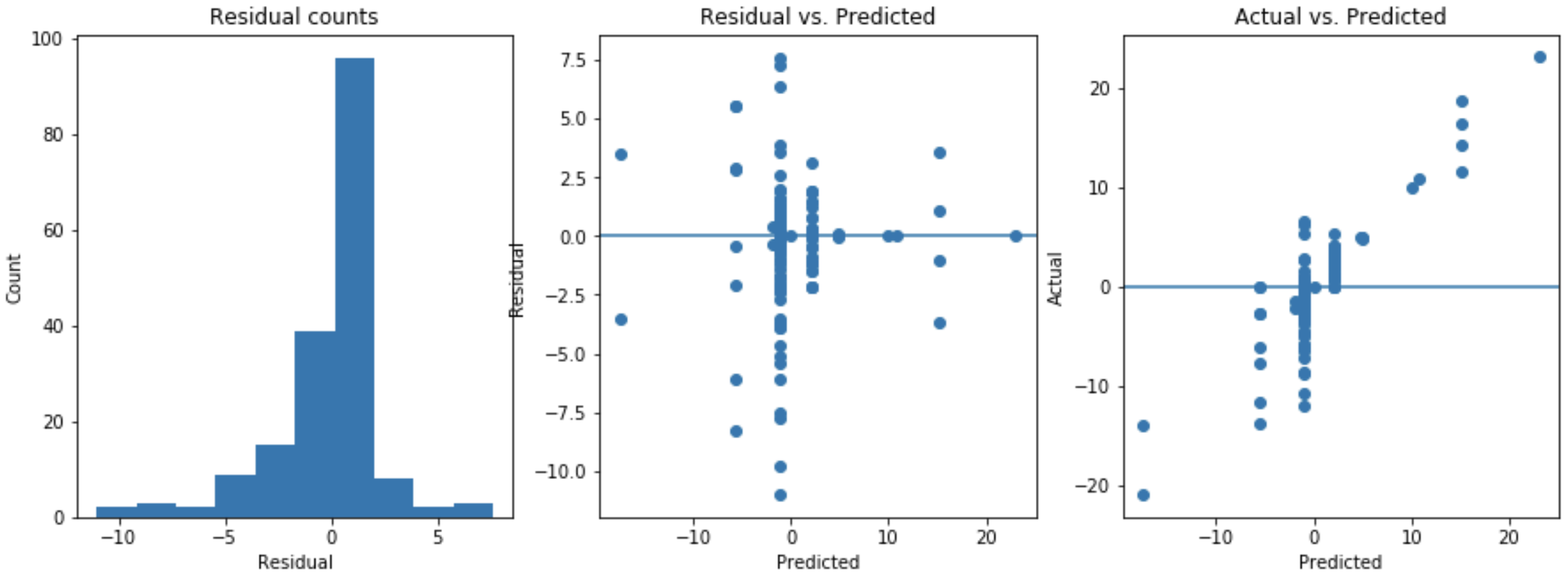
criterion is set to mse (mean squared error)

splitter is set to best

max depth is set 5

Decision Tree Results

R-squared: 79%

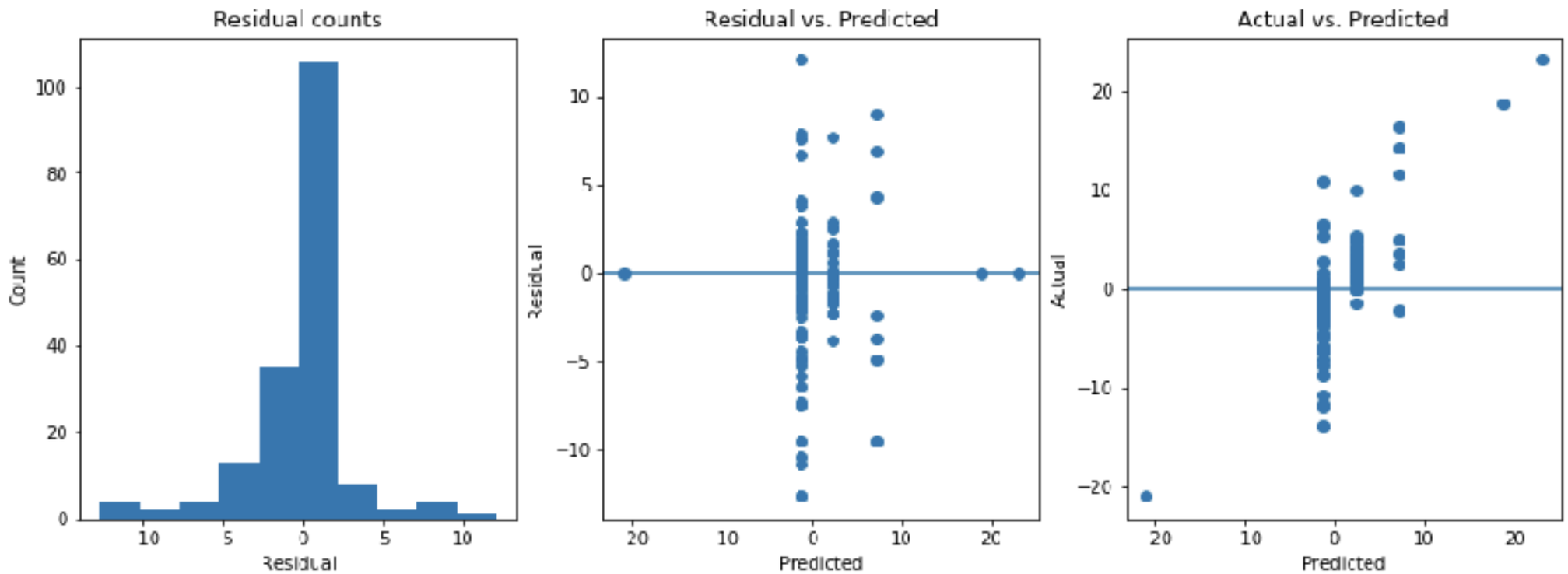


Using all features to fit the model, the residual range is almost 20.

This is a somewhat good fit for the data.

Decision Tree Results with regions removed

R-squared: 52%

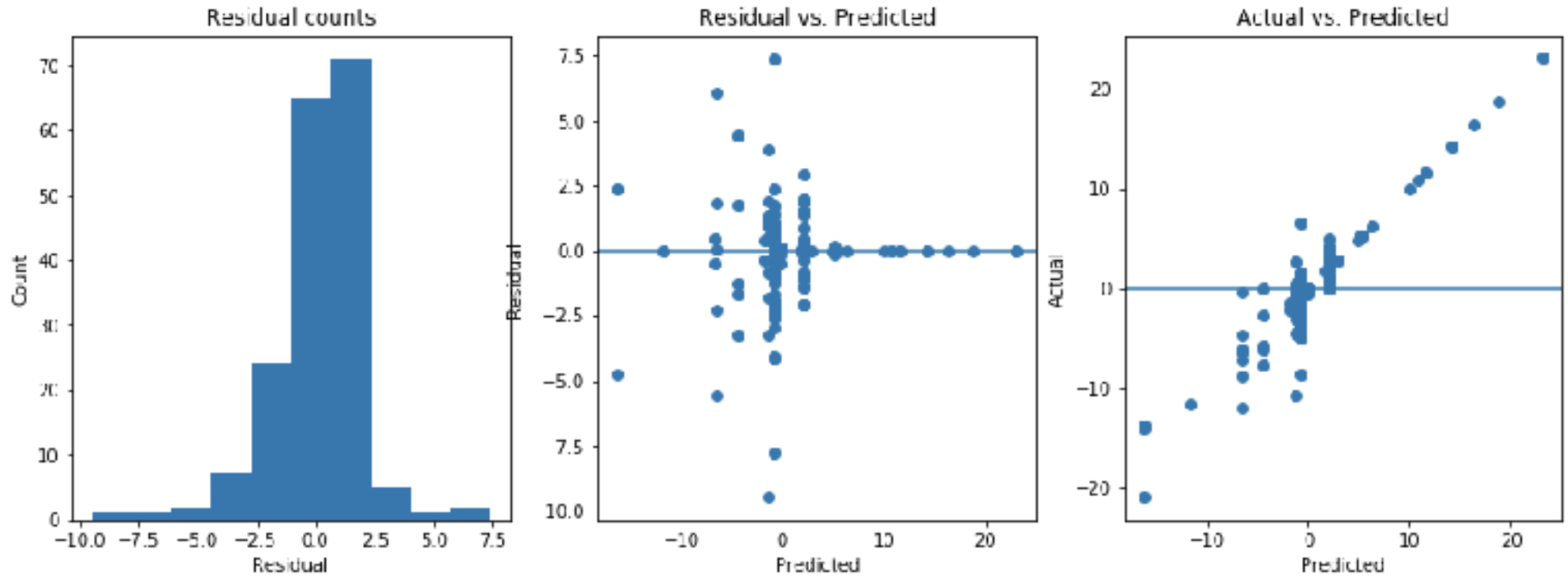


The removal of the Region features has an effect of the Decision Tree Model, reducing the R-squared by 27 points.

This is not a good fit for the data.

Decision Tree Results with PCA

R-squared: 84%



Using PCA to fit the model, the residual range is still less than 20, and the R-squared shows improvement over the fit with all features.

This is a better fit for the data.

Random Forest Regressor

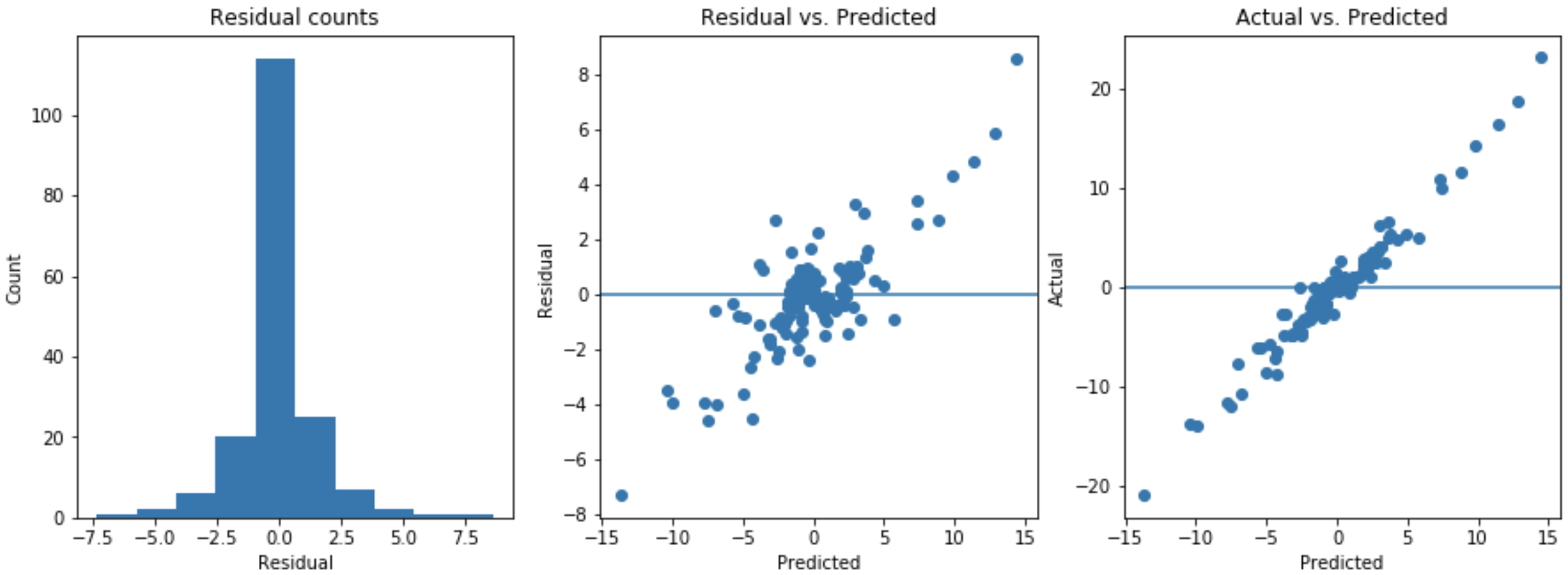
Ensemble model learns from multiple iterations of decision trees (weaker model) in random subspace.

When using all features, the model is tuned to 200 estimators, max depth of 10, 2 max features and default criterion of Max Squared Error (mse).

When using PCA, the model is tuned to 50 estimators, max depth of 10, and criterion of Max Absolute Error (mae).

Random Forest Results

R-squared: 88%

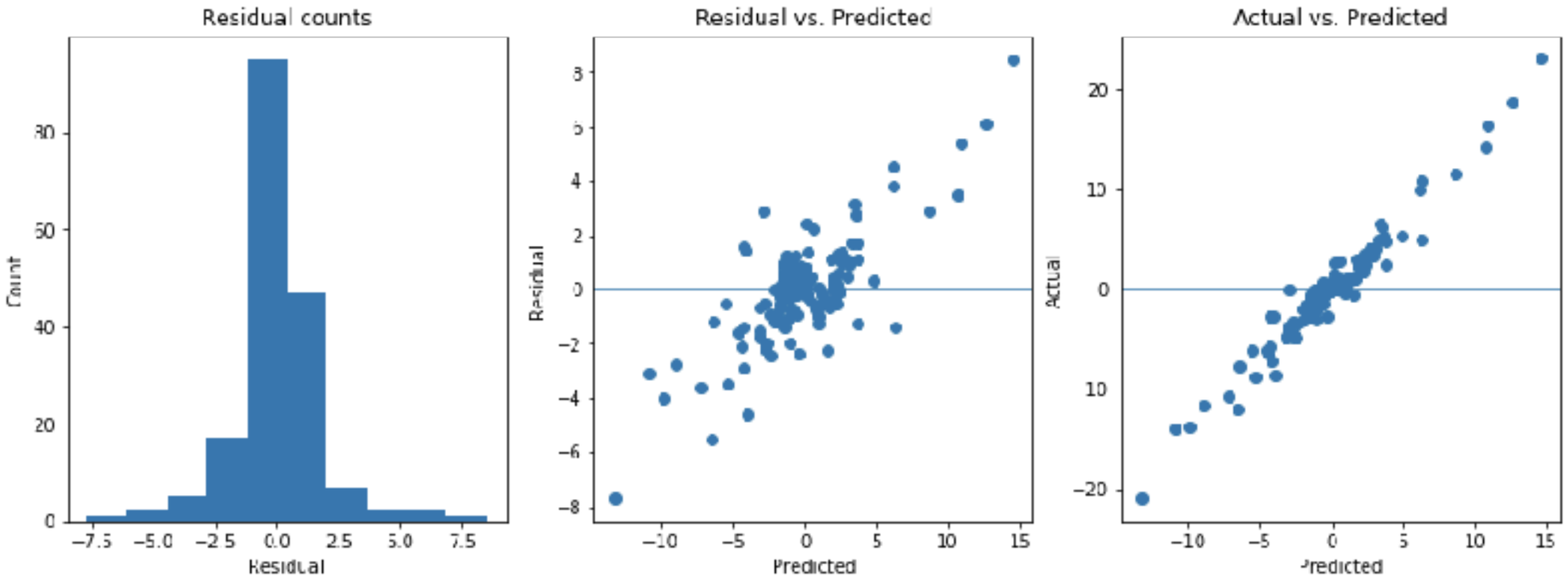


Good accuracy and residual range (~16) smaller than other models, although residuals increase in the extremes.

This is a good fit for the data.

Random Forest Results with regions removed

R-squared: 87%

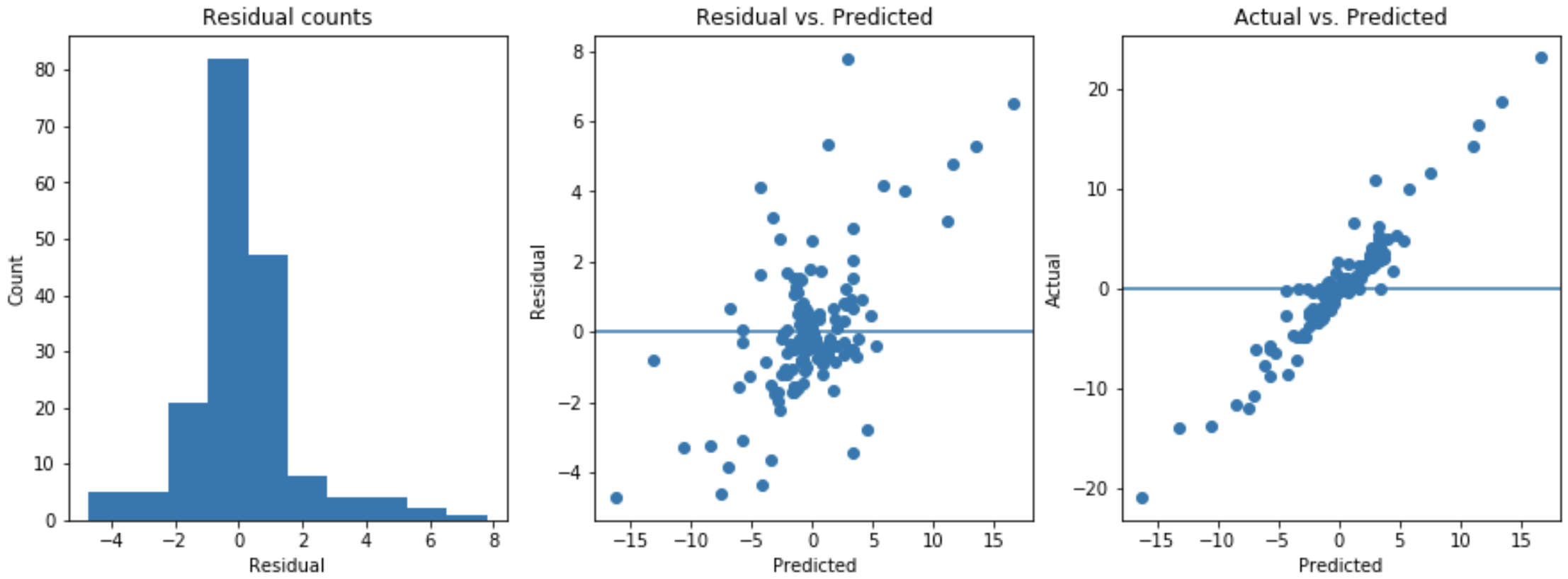


Random Forest is largely unaffected by the removal of the Region features. The R-squared dropped by 1 point and maintained the same residual range.

This is a good fit for the data.

Random Forest Results with PCA

R-squared: 87%



Using PCA with Random Forest decreased the residual range (~13), while keeping a high R-squared.

Random Forest with PCA is a good fit for the data.

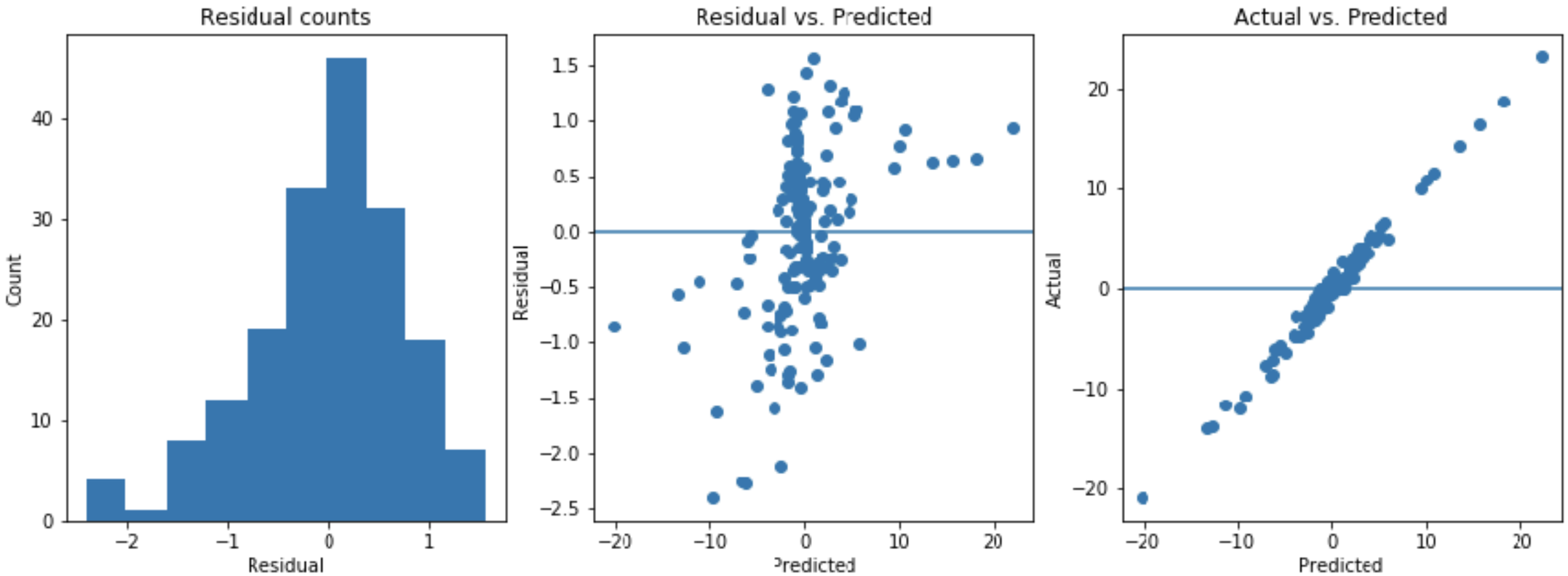
Gradient Boosting Regressor

Ensemble model using random forests to iteratively learn, minimizing exponential loss function (residuals).

Model is tuned to learning rate of 0.1, max depth of 3, 100 estimators and least squares loss function.

Gradient Boost Results

R-squared: 97%

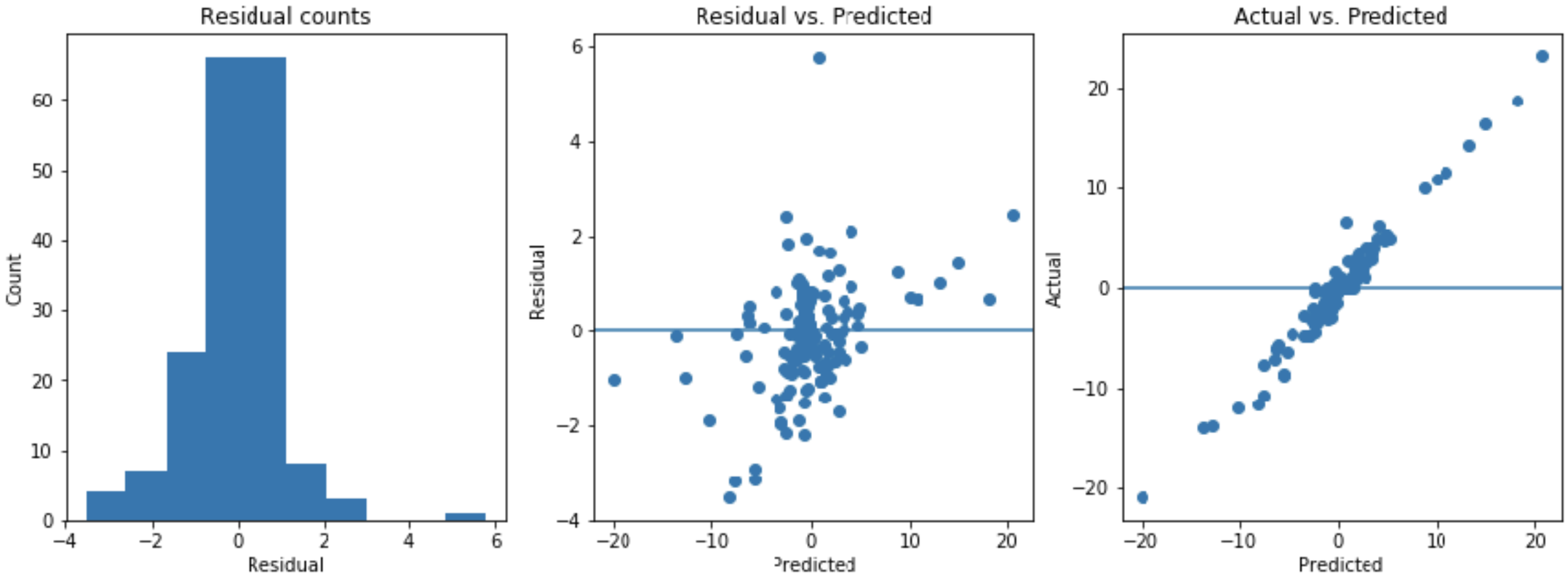


Excellent accuracy, significantly reducing residuals (range is just 4) with both sets of features.

This model is a very good fit for the data.

Gradient Boost Results with PCA

R-squared: 95%



Excellent accuracy is maintained with PCA (dropped just 2 points). Residuals are increased, with a range of 10.

Gradient Boost with PCA is a very good fit for the data.

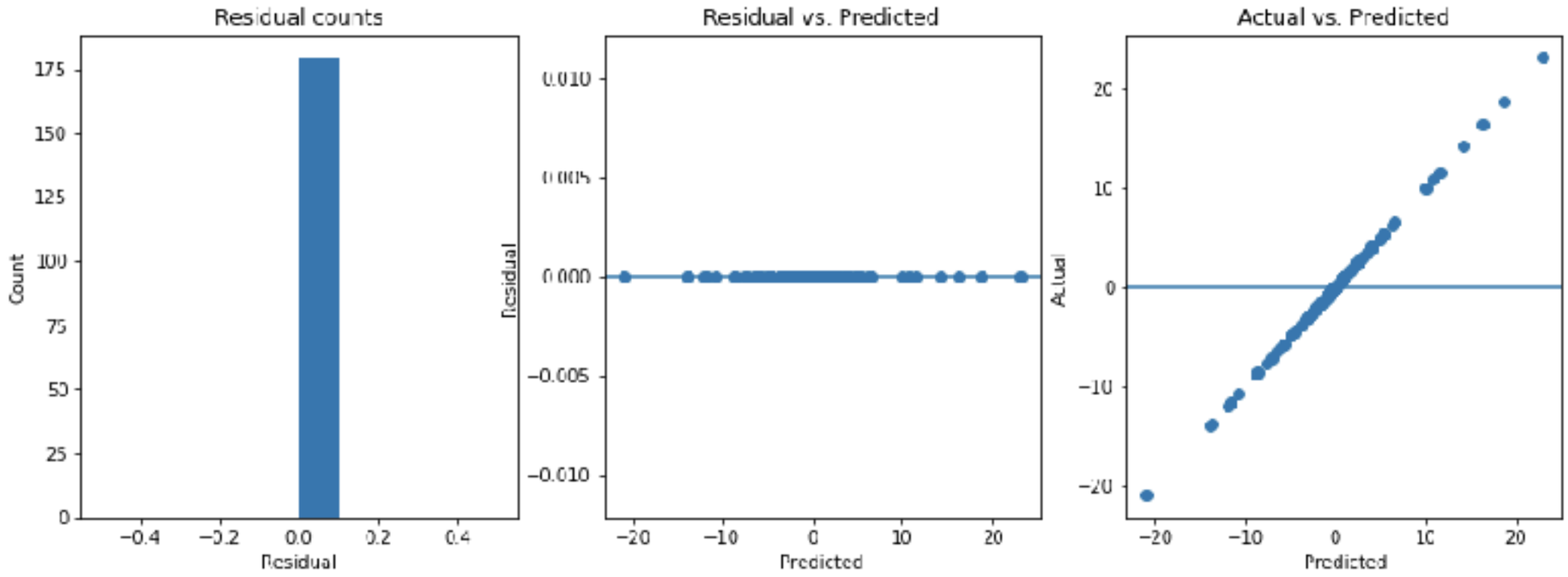
K- Nearest Neighbor

Using observations with similar feature values to predict target value. Optimized by number of observations (neighbors) considered and potential weighting by distance (deference given to obsevations with closest values).

This model is optimized when weighted by Distance, using 10 neighbors.

KNN Results

R-squared: 100%



Perfect accuracy with fit of both PCA and all features.

KNN provides the best fit for the data.

The Best Model

model	R-squared	Time to train	R-squared with PCA	Time to train with PCA
KNN	100%	0.028	100%	0.019
Gradient Boost	97%	0.325	95%	0.205
Random Forest	88%	1.703	87%	0.462
Decision Tree	79%	0.027	84%	0.016
SVM	58%	0.043	57%	0.021
Linear Regression	45%	0.027	22%	0.015
Ridge	44%	0.046	22%	0.024
Lasso	42%	0.026	14%	0.018