Assignment 2B

Does random forest generalize better across divisions than linear models for the process of salary selection? Why might it be reasonable to think this given the way trees work?

Yes, random forests predict salary selection across divisions better than linear models.

A basic tree model partitions the data into nodes by creating simple rules to define a threshold on which the data can be split. This approach in turn, picks up on nonlinear trends, and in this case, different divisions may have different relationships. It also creates the downside of overfitting the data because trees are sensitive to all data points since they can be split until each node contains a single point. This can occur in multiple divisions in this dataset.

Generating a random forest means that we are generating trees repeatedly on the data until a random number of trees, or a "forest," is established. The final prediction is the average overall decision trees in the forest.

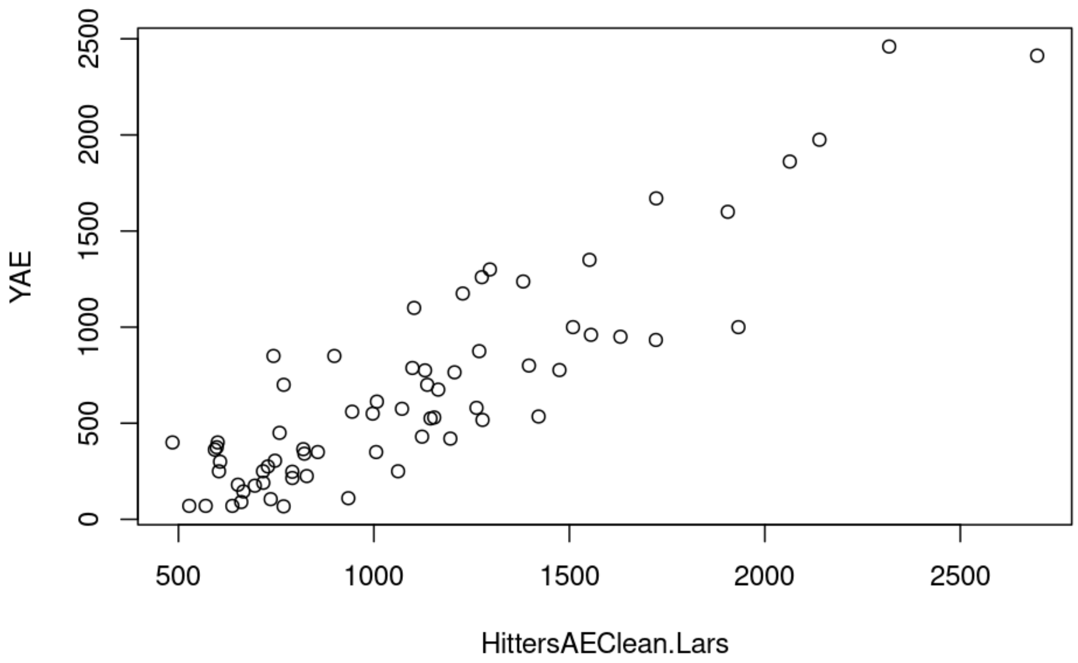
Use the plot and correlation of the cross predictions for Lars vs salary and random forest vs salary to answer this question.

The cross-prediction correlations support this:

- The random forest predictions correlate reasonably well across both divisions (0.67 for AE, 0.72 for NW).
- The lasso predictions correlate well on the training AE division (0.89) but much worse on the NW division (0.51).

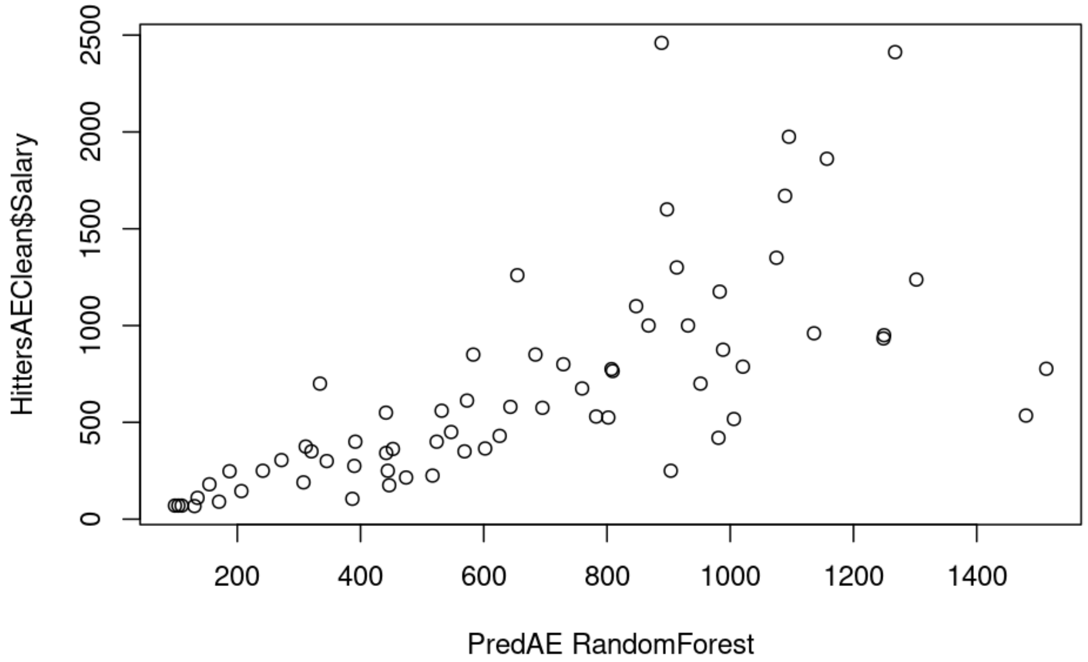
So, the random forest does seem to generalize better across divisions than the linear lasso model in this salary prediction task. Examining the cross prediction correlations lets us quantitatively assess how each model transfers.

After an analysis of the plots generated in all divisions via LARS and Random Forest, the derived conclusion supports the accuracy of the random forest models compared to the Lars models in predicting salary selection. This was determined by the correlation values found for each pair of plots (LARS vs Random Forest).



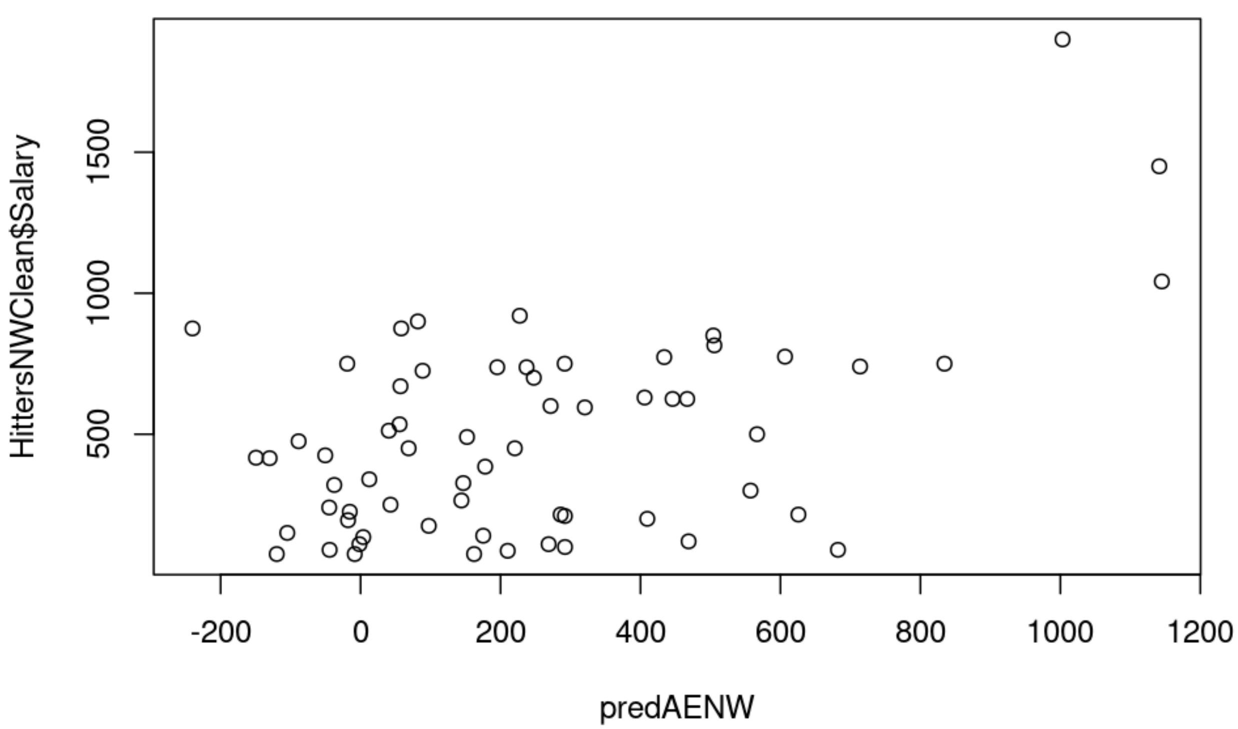
Correlation value: 0.8890144

Since the correlation value is so close to 1, there is a strong correlation between the data for League A and Division E and salary. There are no clear outliers that go against the trend. This plot utilizes the LARS regression model, which displays the relationship well.



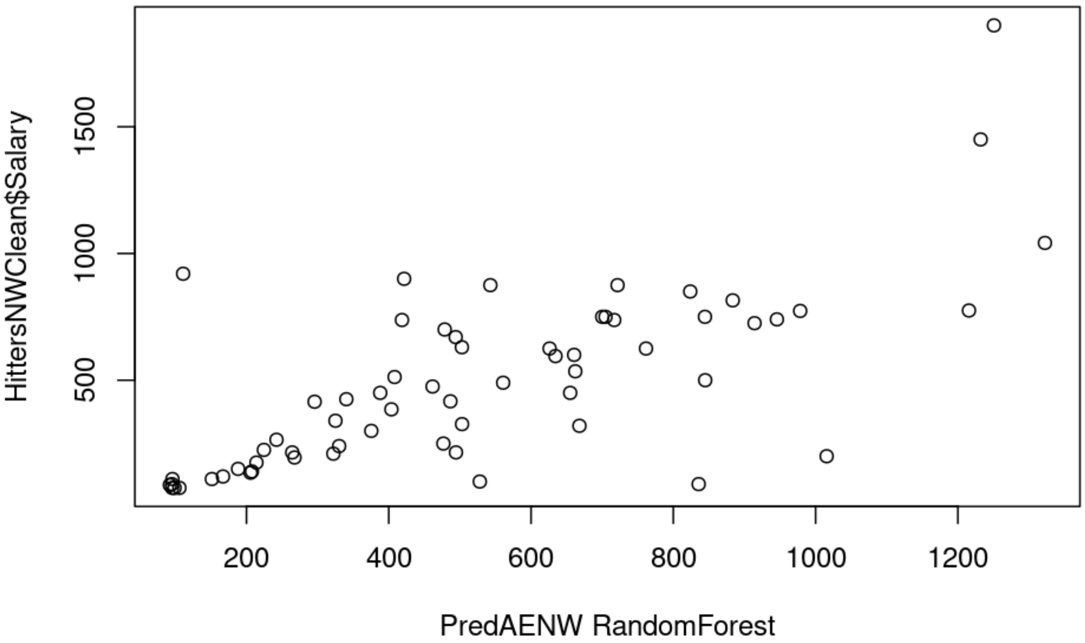
Correlation value: 0.6900489

This scatter plot utilizes the random forest technique to create a prediction of the relationship between the data representing the performance of the hitters for League A and Division E versus their corresponding salary. There is a lower correlation value using the Random Forest technique than there was using the LARs method. In this case, the random forest method yielded a lower and worse correlation value, however, this is just an anomaly to the method as even the Random Forest method could rarely pick an unrepresentative sample.



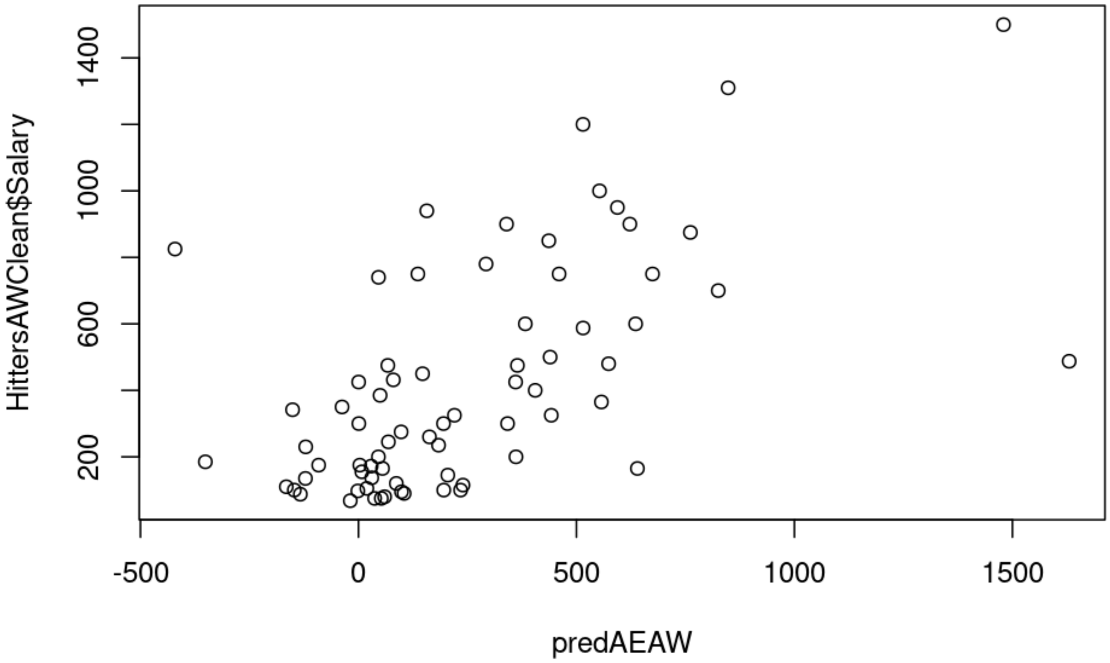
Correlation value: 0.506991

Once again, utilizing the LARs regression method, we found the correlation between League A, Division E, League N, Division W, and the NW salary to be quite low, which goes against the expected trend of a strong correlation. All salaries stay within a certain range of 0 to 1000 regardless of the predicted data values.



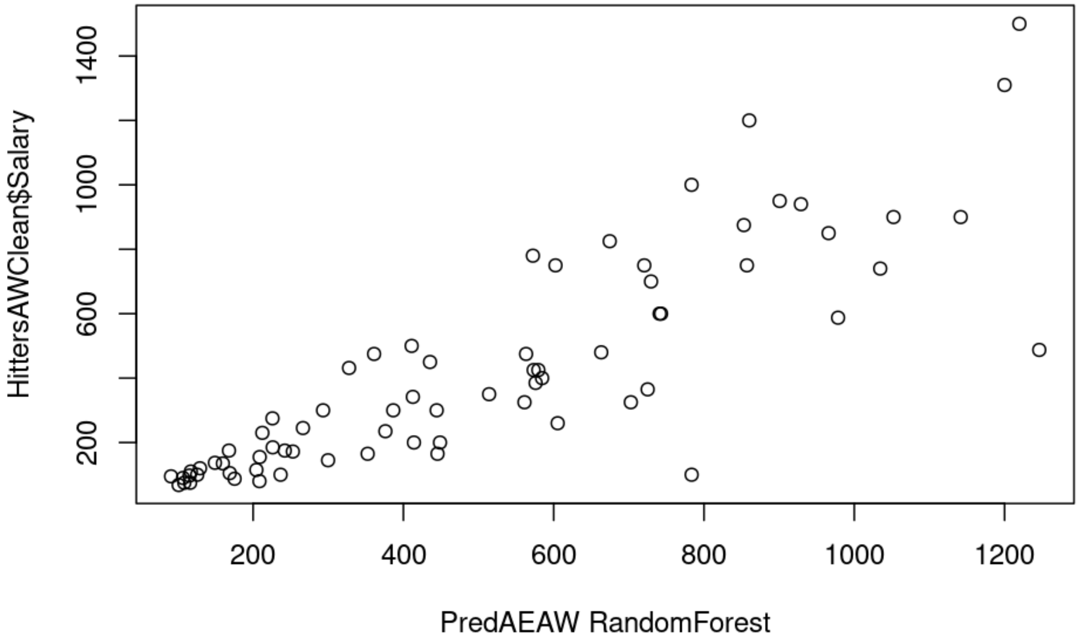
Correlation value: 0.7174673

Analyzing the same relationship stated above, it is clear that the Random Forest regression method is more accurate than Lars because the correlation value for random forest is higher than when Lars was used. The regression scatter plot represents a positive correlation between the data and salary. There is an outlier at about 80, 800, and about 1050, which makes the correlation moderately strong.



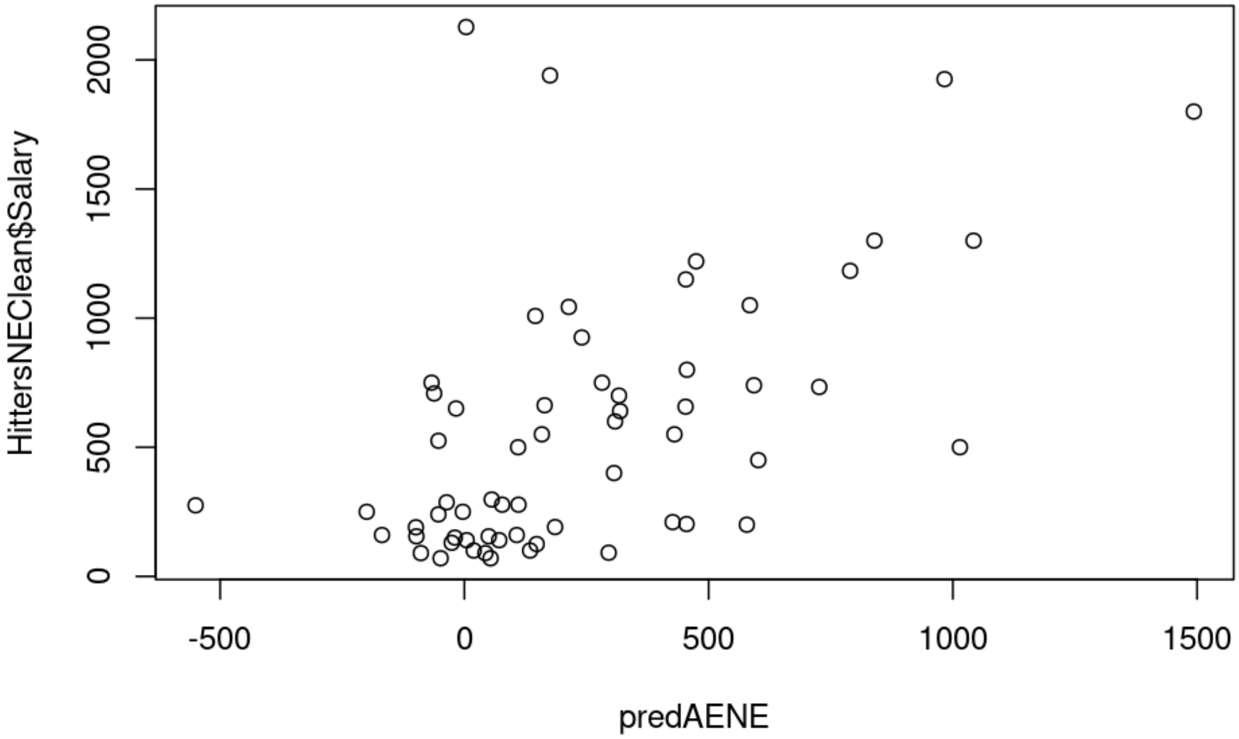
Correlation value: 0.5973853

We analyzed the relationship between the data for League A Division E & League A Division W against the League A Division W salary. The relationship is moderately positively correlated, indicated by the correlation value 0.5973853. There is clustering around -100 to 250, which might have contributed to the low correlation value.



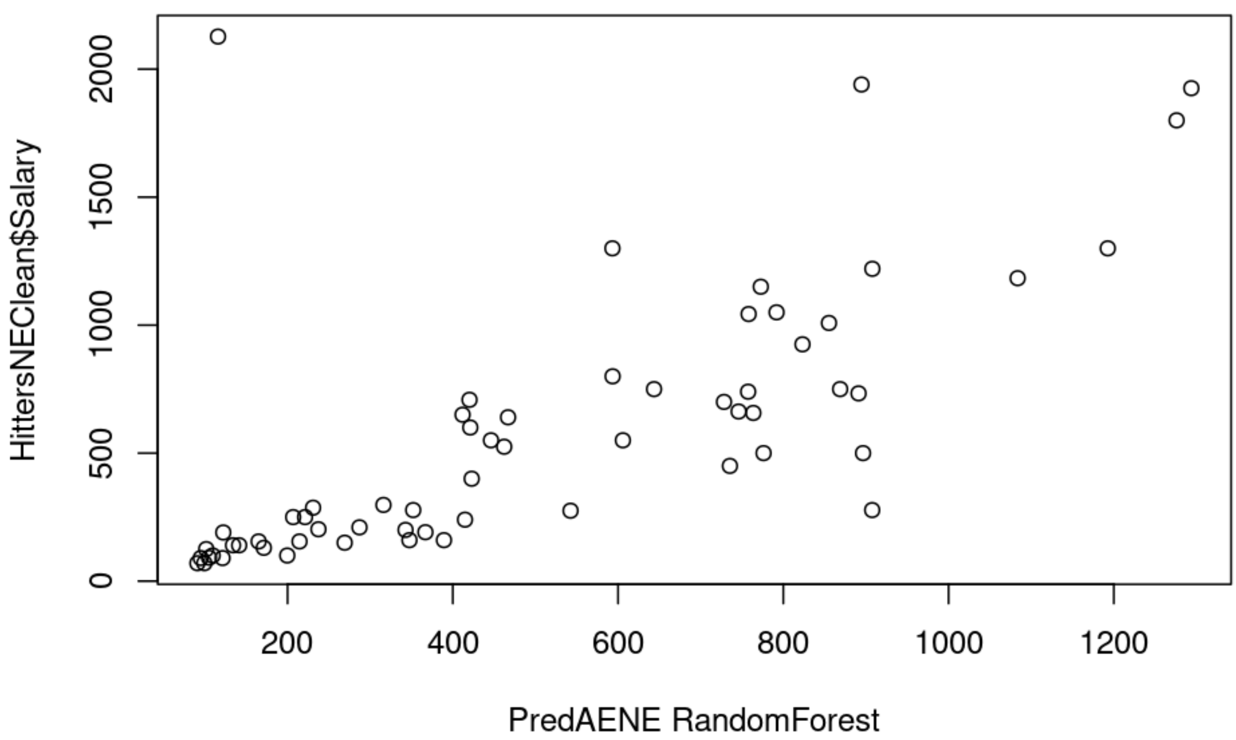
Correlation value: 0.847302

The correlation value of the Random Forest plot is higher than the Lars plot (0.847302>0.5973853), allowing us to conclude that the Random Forest plot was a better predictor of the data and the relationship for the League A Division W salary versus League A Division E & League A Division W.



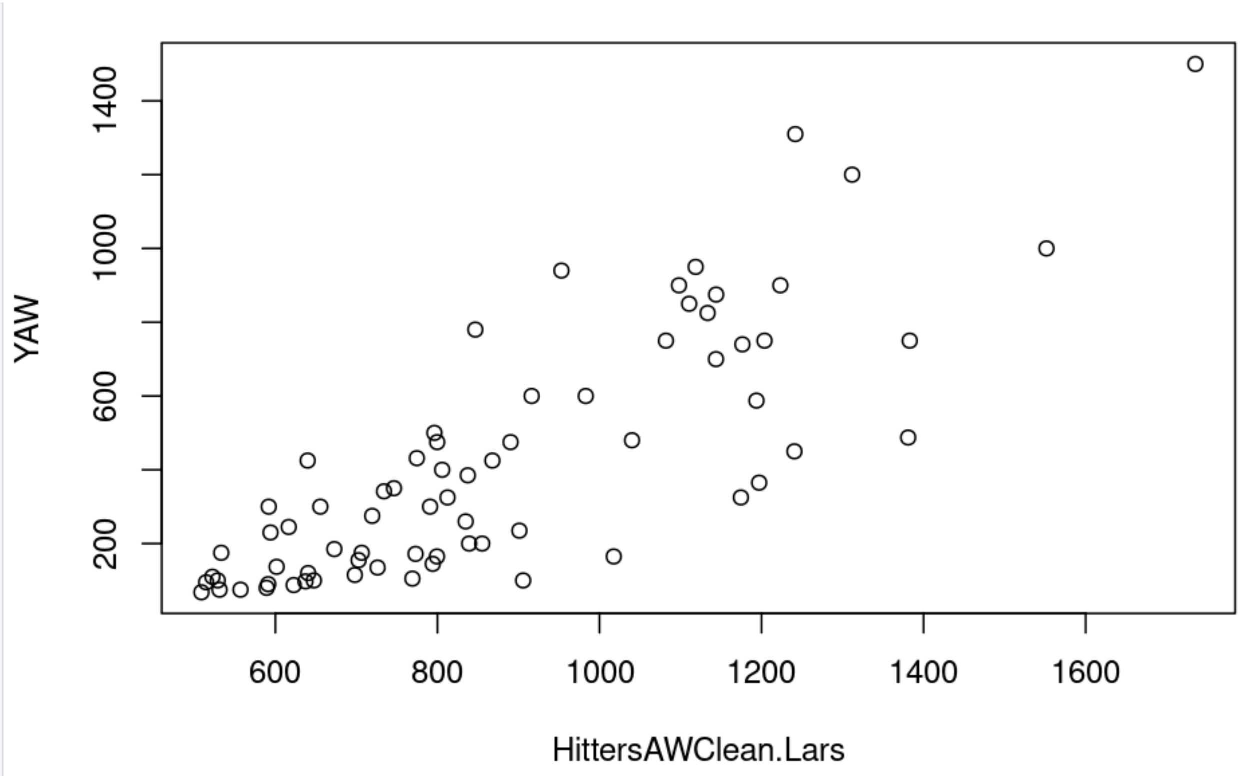
<u>Correlation value: 0.5493544</u>

This plot uses the Lars model to show the correlation between the salaries in comparison to the divisions AE and NE. We can see that the majority of the data is condensed between the x-values values of approximately - 200 to 500 and the y-values of less than 1000. While we cannot claim that there is a strong, positive linear relationship between the two variables, we can say that there is a slight positive trend that can be seen from the graph. The correlation value is 0.5493544, which shows that the plot is somewhat of a decent representation of the correlation between the two variables.



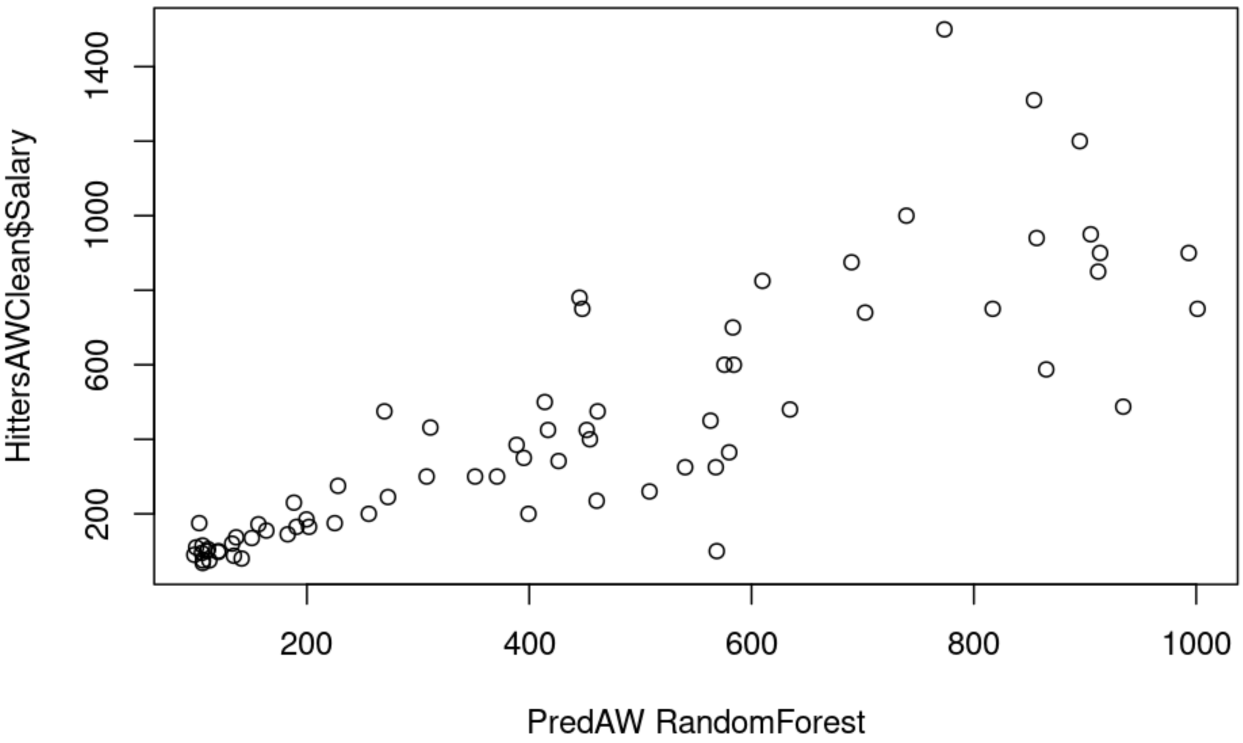
Correlation value: 0.7092483

This plot uses the Random Forest model to show the correlation between the salaries in comparison to the divisions AE and NE. We can see that a part of the data is condensed between the x-values of approximately 0 to 200 and the y-values of less than 1000. We can also observe that the values are on a rise as the x-values rise, thus signifying that there is a positive, and linear relationship between the two variables, though it may not be the strongest. The correlation value is 0.7092483, which shows that the plot is a decent representation of the correlation between the two variables. Also, this plot has a higher correlation value than the Lars model, thus showing that this plot is a better representation of the data. This helps emphasize the claim that Random Forest is a better predictor of the trends than Lars.



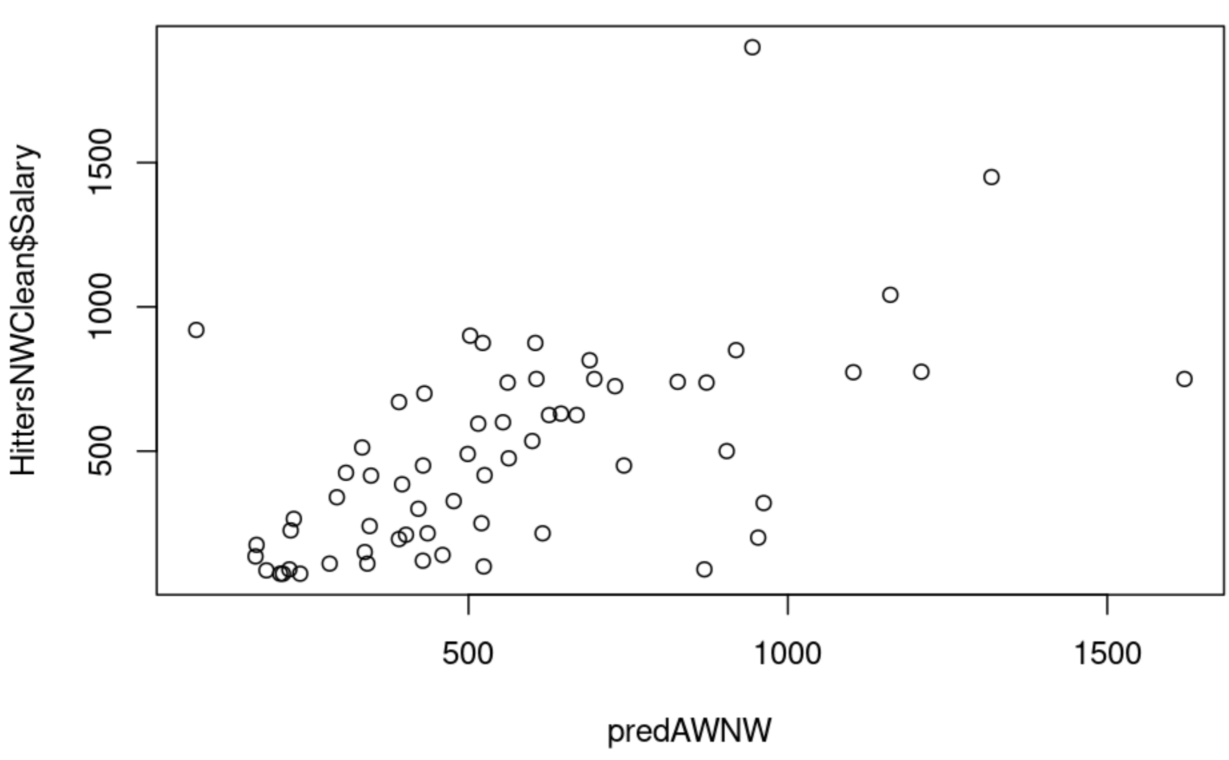
<u>Correlation value: 0.817936</u>

This plot uses the Lars model to show the correlation between the salaries in comparison to the division AW. We can see that the majority of the data is condensed between the x-values of approximately 0 to 800 and the y-values of less than 1000. While we cannot claim that there is a strong, positive linear relationship between the two variables, we can say that there is a slight positive trend that can be seen from the graph. The correlation value is 0.817936, which shows that the plot is a good representation of the correlation between the two variables.



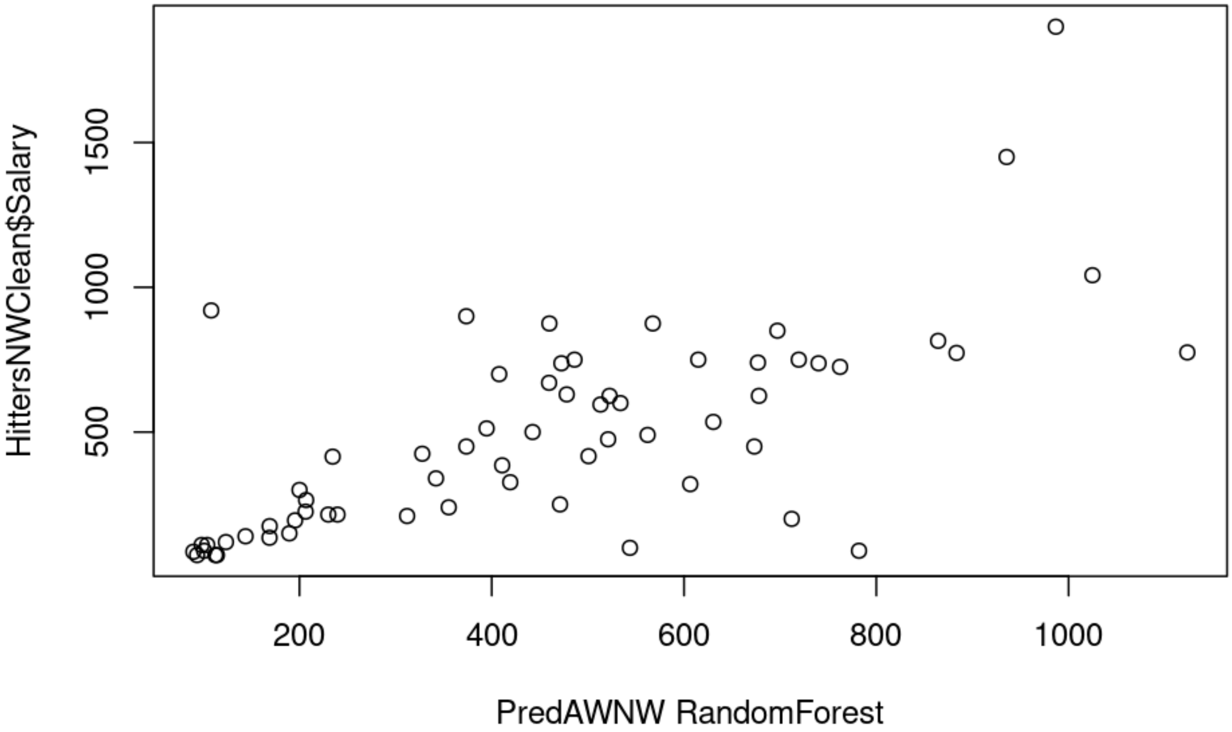
Correlation value: 0.844974

This plot uses the Random Forest model to show the correlation between the salaries in comparison to the division AW. We can see that a part of the data condensed between the x-values of approximately 0 to 200 and the y-values of less than 600. We can also observe that the values have somewhat of a positive trend, though it may not be the strongest. The correlation value is 0.844974, which shows that the plot is a good representation of the correlation between the two variables. Also, this plot has a higher correlation value than the Lars model, thus showing that this plot is a better representation of the data. This helps emphasize the claim that Random Forest is a better predictor of the trends than Lars.



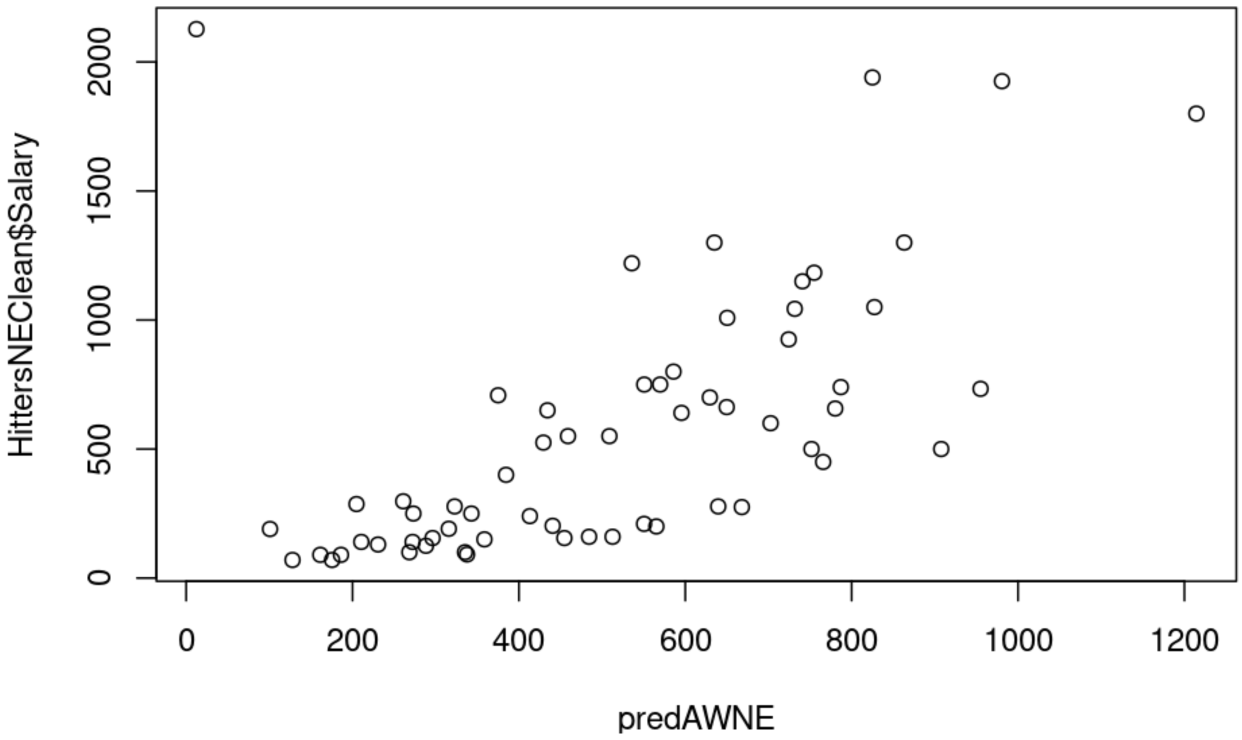
Correlation value: 0.5721472

This plot uses the Lars model to show the correlation between the salary of the NW Division and the divisions' AW and NW. We can see that there aren't any major patterns or trends in the plot. The data values of the group mostly seem pretty scattered, but nonetheless, they are mostly condensed around the x-values of 200 - 600 and the y-values of less than 1000. The correlation value is 0.5721472, which shows that the plot is somewhat of a decent representation of the correlation between the two variables.



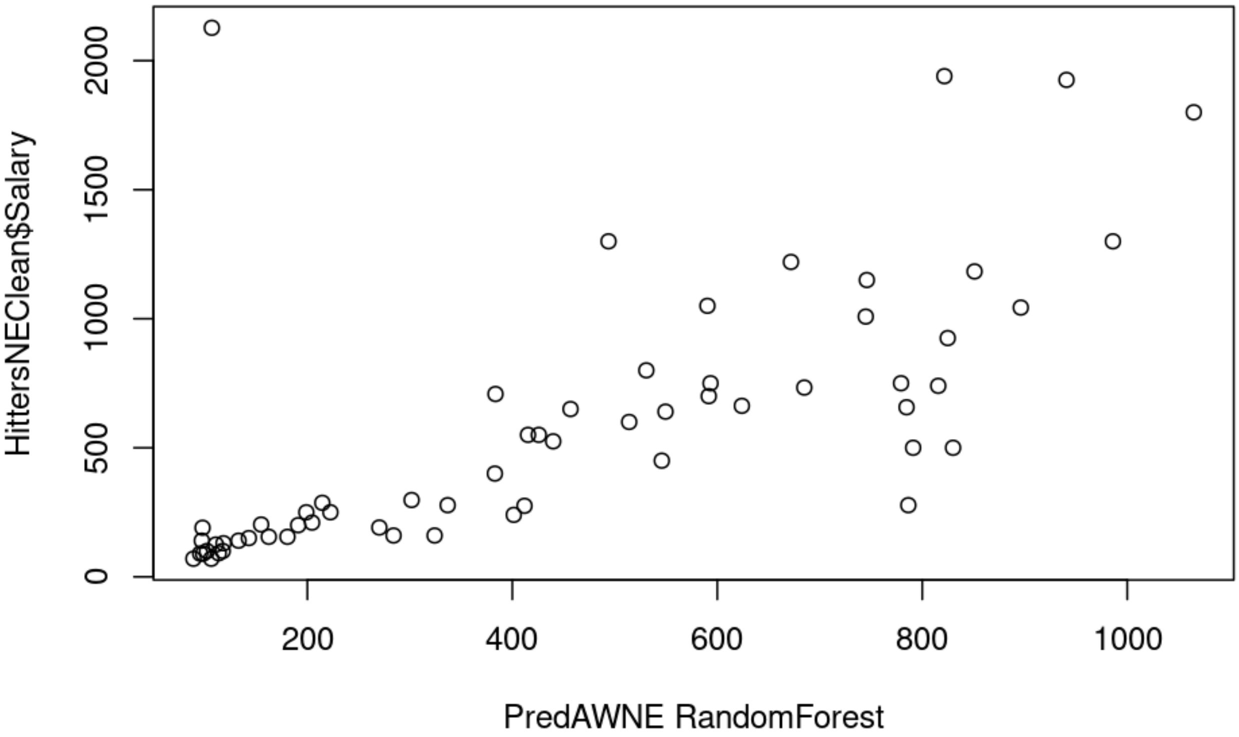
Correlation value: 0.6958524

This plot uses the Random Forest model to show the correlation between the salary of the NW division in comparison to the 0. We can see that a part of the data is condensed between the x-values values of approximately 0 to 200 and the y-values of less than 600. We can also observe that the values have somewhat of a positive trend, though it may not be the strongest. The correlation value is 0.6958524, which shows that the plot is a decent representation of the correlation between the two variables. Also, this plot has a higher correlation value than the Lars model, thus showing that this plot is a better representation of the data. This helps emphasize the claim that Random Forest is a better predictor of the trends than Lars.



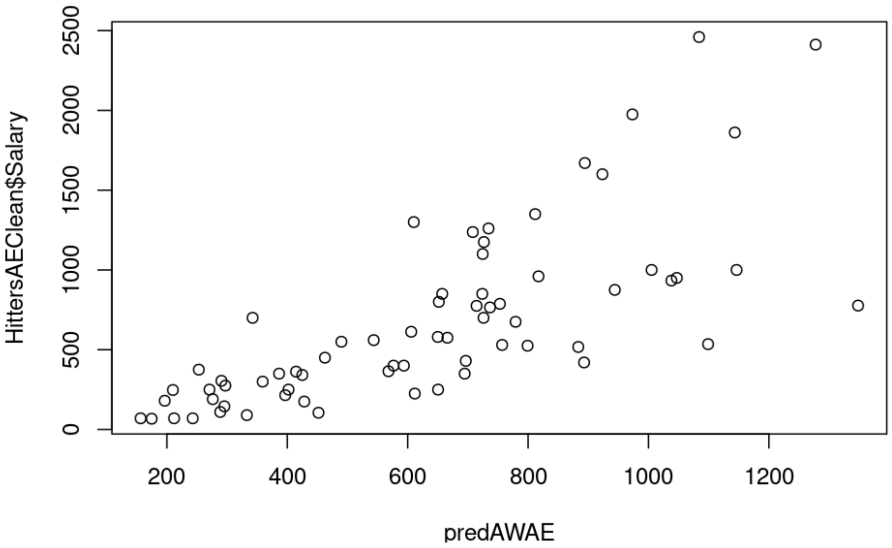
Correlation value: 0.5792761

This plot uses the Random Forest model to showcase the correlation between the salary of the NE Division and the divisions' AW and NE. The correlation between the salaries is 0.5792761, which is almost 50/50, based on how strong the correlation is between the two divisions. There is a moderate positive correlation, and more points lie between 200 and 400 on the x-axis. The correlation is not strong enough to make precise predictions about one division's salaries based solely on the other division's salaries. Additional factors and variables may also be influencing the relationship between these two divisions' salaries. Also, this plot has a higher correlation value than the Lars model, which shows that this plot is a better representation of the data.



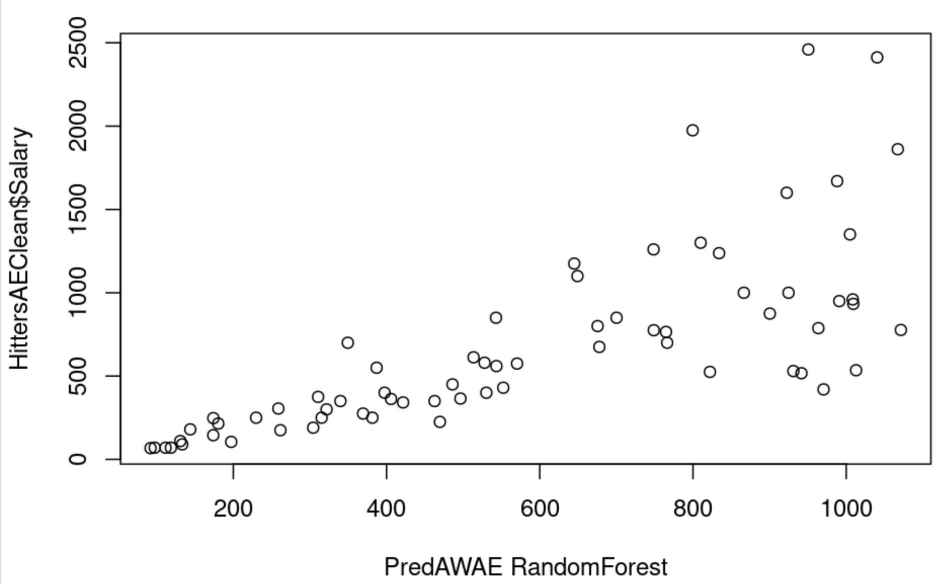
Correlation value: 0.6930962

This plot uses the Random Forest model to show the correlation between the salary of the AW division in comparison of the AW salary. The correlation is 0.6930962, which is a decent correlation to say that it has a decently strong relation to all the data points in the model. Also, this plot has a higher correlation value than the Lars model, which implies that this plot is a better representation of the data. This further proves that Random Forest is a better predictor of the trends than Lars.



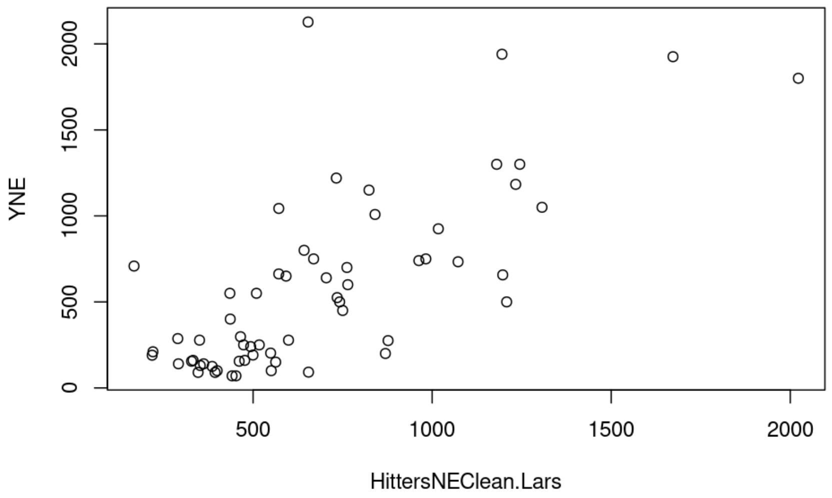
Correlation value: 0.7368015

This plot is generated for the AWAE division using the LARS linear model. The correlation is less than that determined by the random forest model. We see a linear relationship between salary and prediction, and the points have a greater spread as the x-y values increase. We also do not see clustering in the lower values as strongly as with Random Forest.



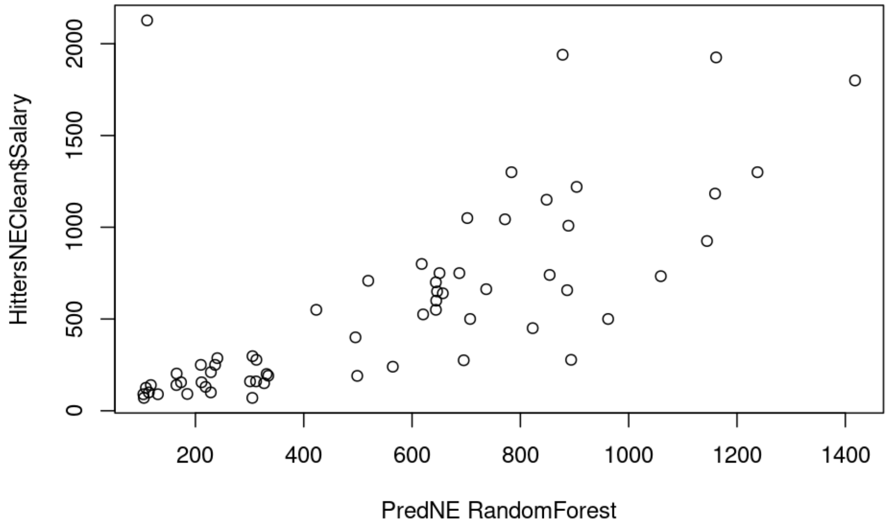
Correlation value: 0.7442501

Here, the correlation value derived from the plot generated by the Random Forest for the AWAE division is larger than the correlation from the linear LARS model. We see a linear relationship generated with a clustering at the lower x-y values, with a greater variance as both x and y values increase.



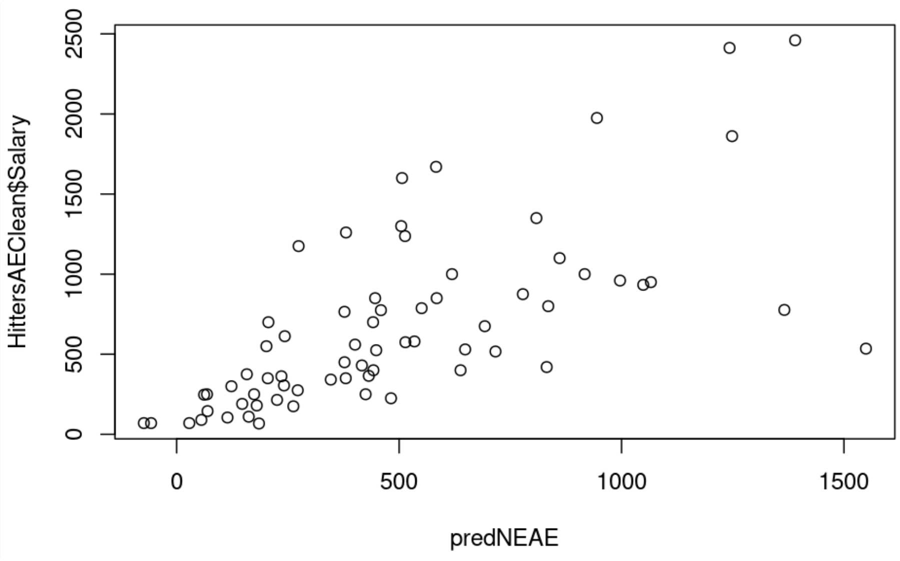
Correlation value: 0.7247419

This plot is generated for the NE division using the LARS linear model. It has a higher correlation value than the corresponding Random Forest model. There is no great linear relationship determined by the produced plot. We see a strong cluster at the lower values (especially sub 500) and outliers spanning all x-values that relate to high predicted salaries.



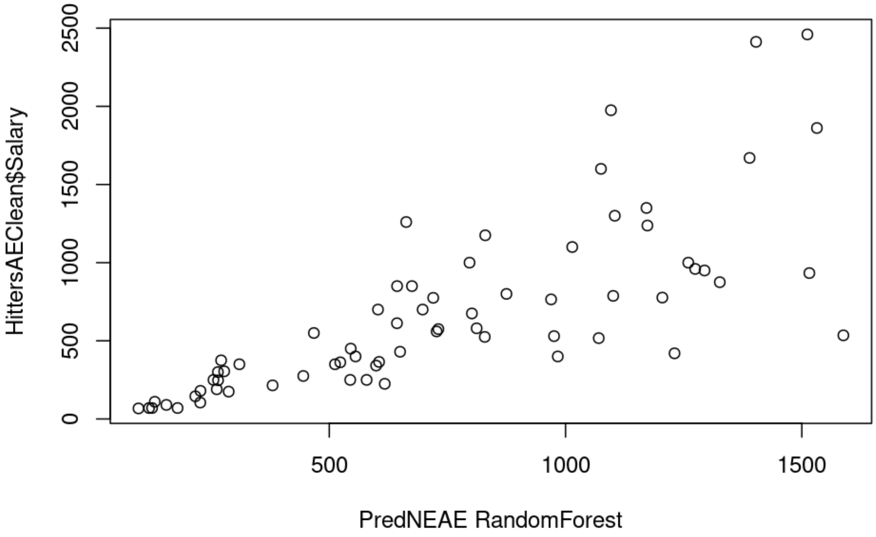
Correlation value: 0.684643

This plot is produced by the Random Forest prediction model, and it has a lower correlation value compared to the LARS model. We see similar features in this plot, including a sub-500 cluster and similar outliers, but there is a stronger linear correlation exhibited in this plot. The outliers are at lower predictions (lower than 1400).



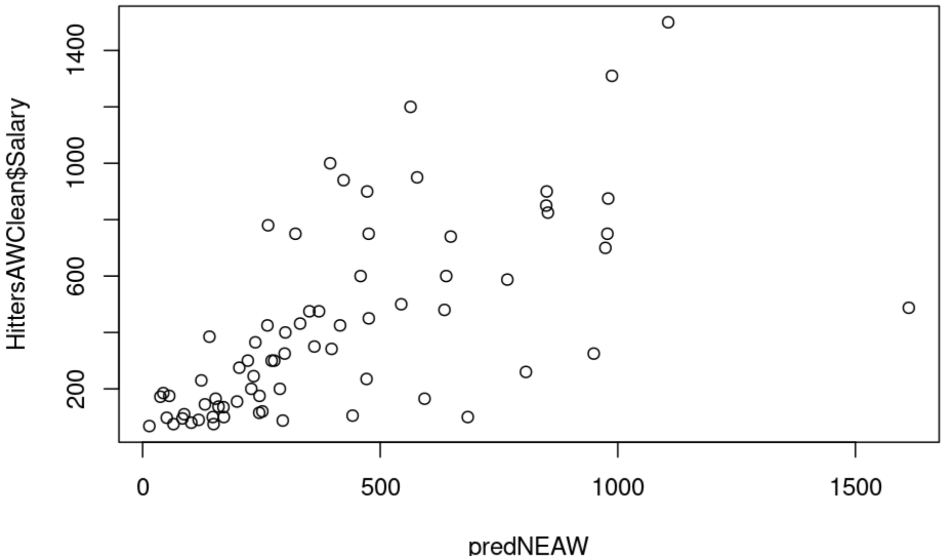
Correlation value: 0.6803984

This is the plot for the NEAE division using the linear LARS prediction model. We see a lower correlation value than there is for the random forest model. There is somewhat of a linear trend, although it is not distinct, and most of the clustering occurs about the lower-to-mid values.



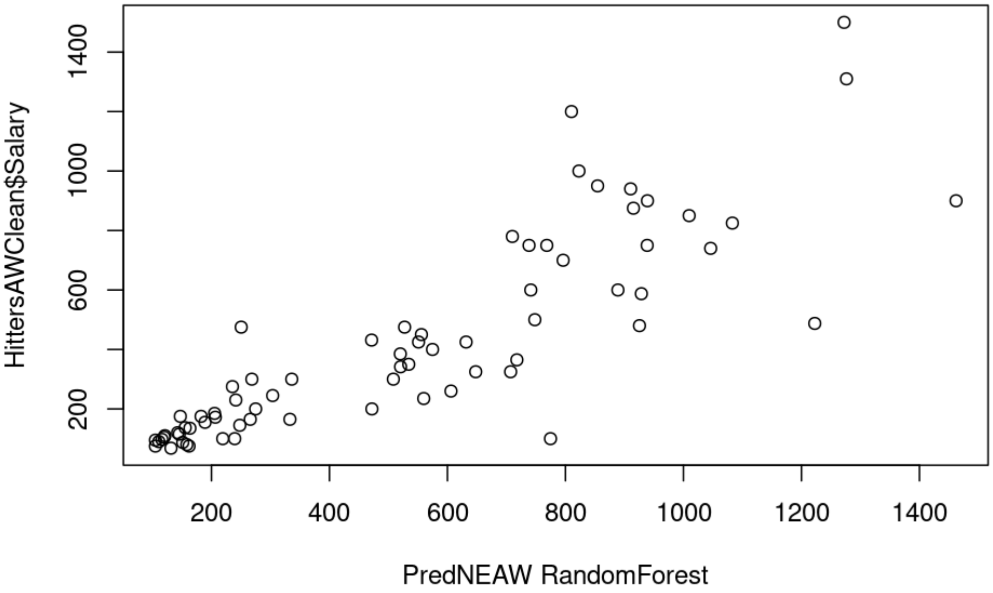
Correlation value: 0.7625007

The plot was generated for the NEAE division via the Random Forest model, which has a higher correlation value than the corresponding LARS model. There is a stronger linear relationship exhibited in this plot, with clustering at lower values, and a higher spread at values above x=1000.



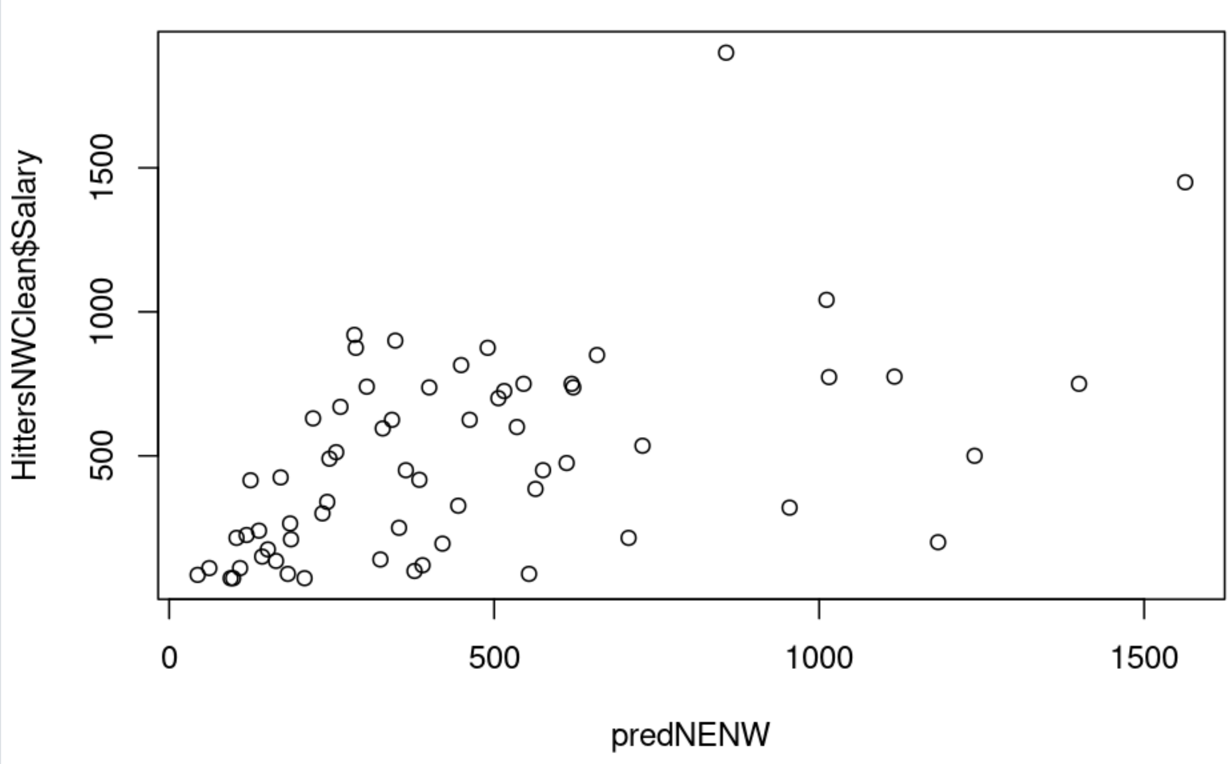
Correlation value: 0.6390477

This is the plot for the NEAW division via the Random Forest model, which has a correlation of 0.6390477, higher than the LARS model. This moderately strong and positive correlation value shows that as the salaries in one division increase, the salaries in the other division tend to increase as well, but not in a perfectly linear manner. There is a cluster of points between 0 and 500 on the x-axis and a single outlier on a point past the x-value of 1500.



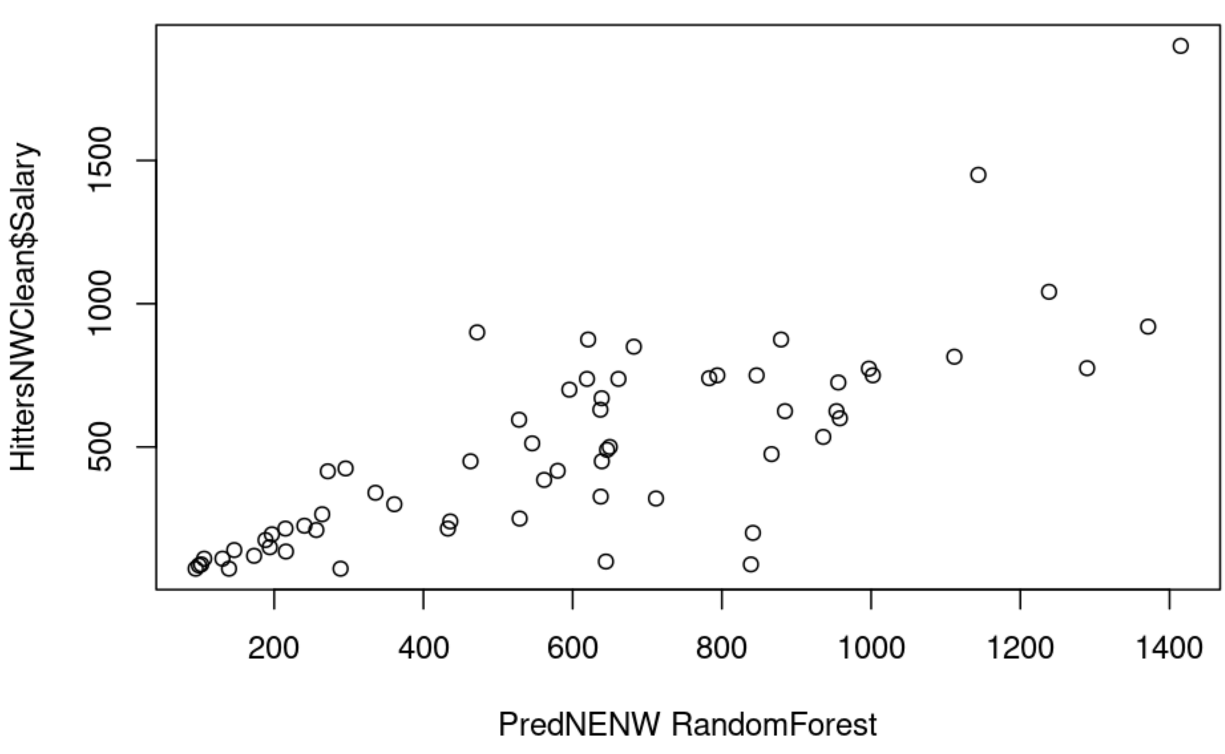
Correlation value: 0.8522953

This is the plot for the NEAW division via the Random Forest model. The standout feature of this particular plot is the remarkably high correlation coefficient of 0.8522953. A correlation coefficient ranging from 0 to 1 indicates a positive correlation, with values closer to 1 signifying a stronger positive relationship. In this case, the correlation coefficient of 0.8522953 suggests an exceptionally strong positive correlation between the salaries of employees in the two divisions being compared. There is a cluster of points right below the 200 points on the x-axis and y-axis. This clustering suggests that a considerable portion of employees or positions in both divisions have salaries within a similar range. This plot has a higher correlation value than the Lars model, showing that this plot better represents the data.



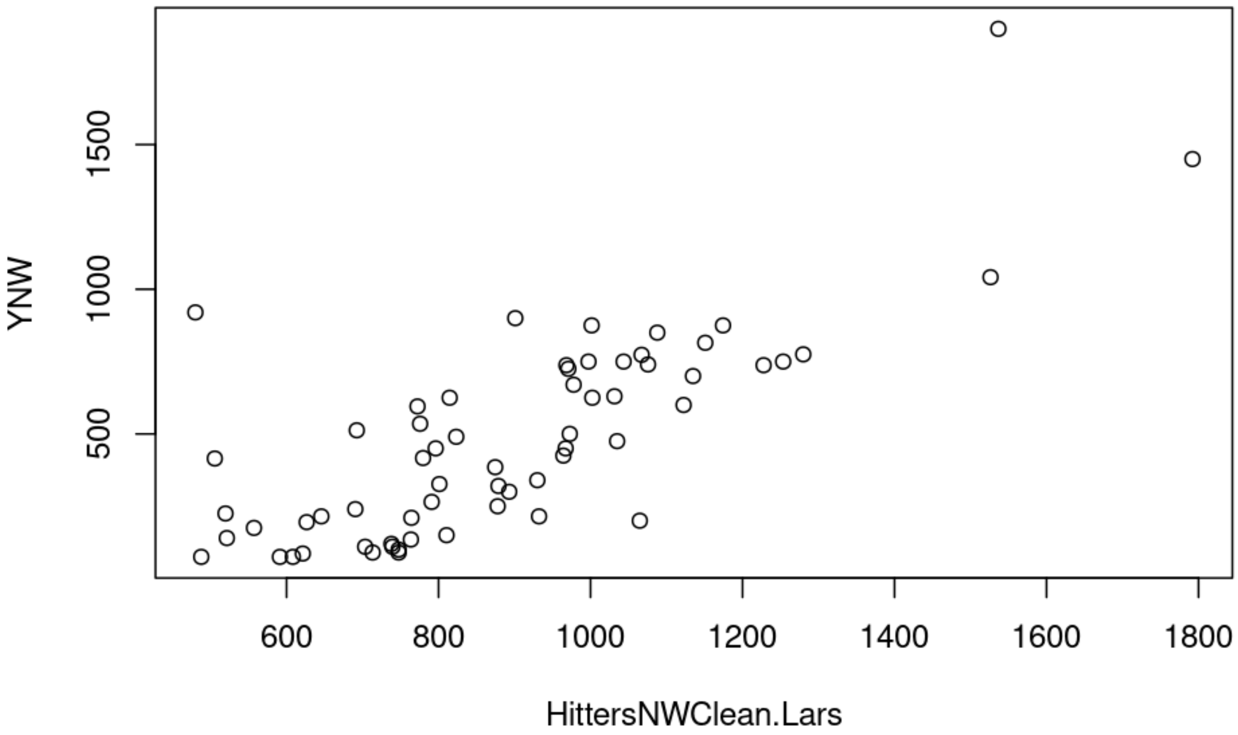
<u>Correlation value: 0.5272795</u>

This is the plot for the NENW division, and the correlation value is 0.5272795, which suggests a moderate positive correlation between the salaries of employees in the NENW division and the other divisions being compared. A moderately positive correlation in this case implies that as the salaries in one division increase, there is a tendency for the salaries in the other division also to increase, but not in a perfectly linear or proportional manner. This relationship is not as strong as a high positive correlation, but it is still significant and indicative of a connection between the salary structures of these two divisions.



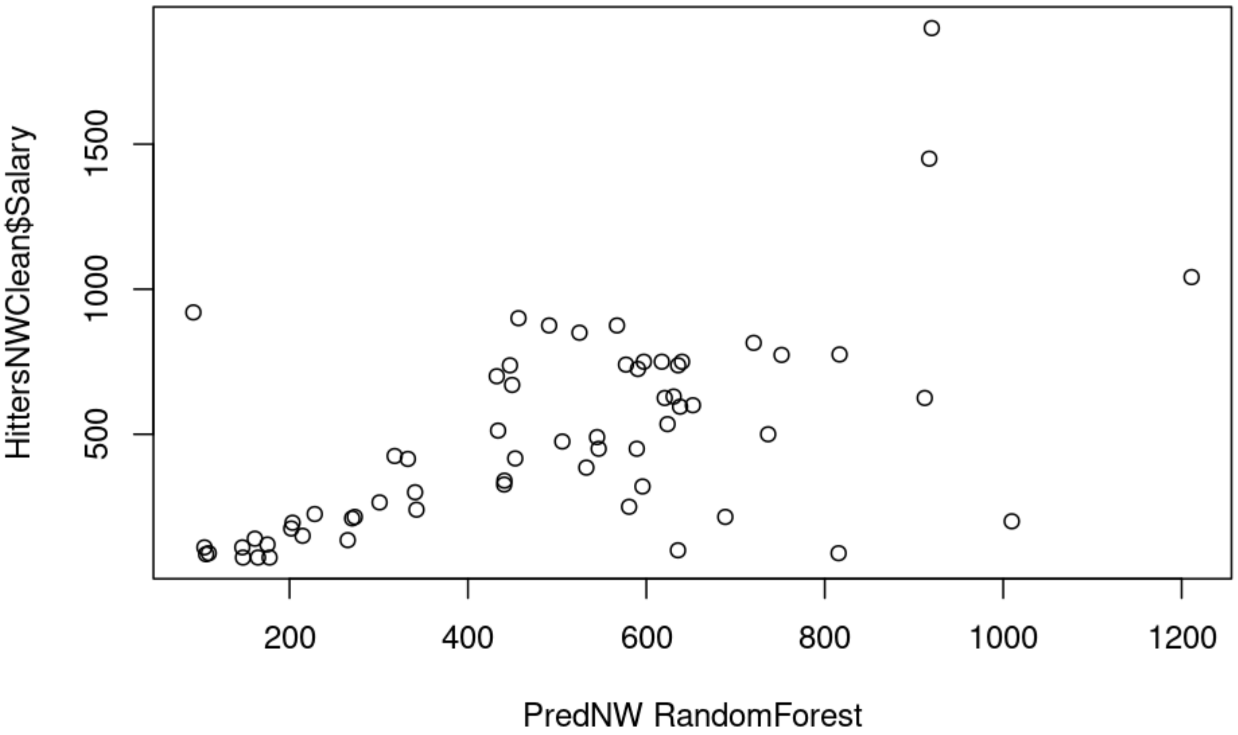
Correlation value: 0.7856406

This is the plot for the NENW divisions via the RandomForest model. The correlation value of this plot is 0.7856406, which indicates a strong positive correlation between the salaries of employees in the two NENW divisions being compared. There's a cluster of points below the 200 mark on the x-axis and below the 500 mark on the y-axis, which could suggest that a significant portion of employees or positions in both NENW divisions have salaries within these lower ranges. This plot has a higher correlation value than the Lars model, and that shows that this plot is a better representation of the data.



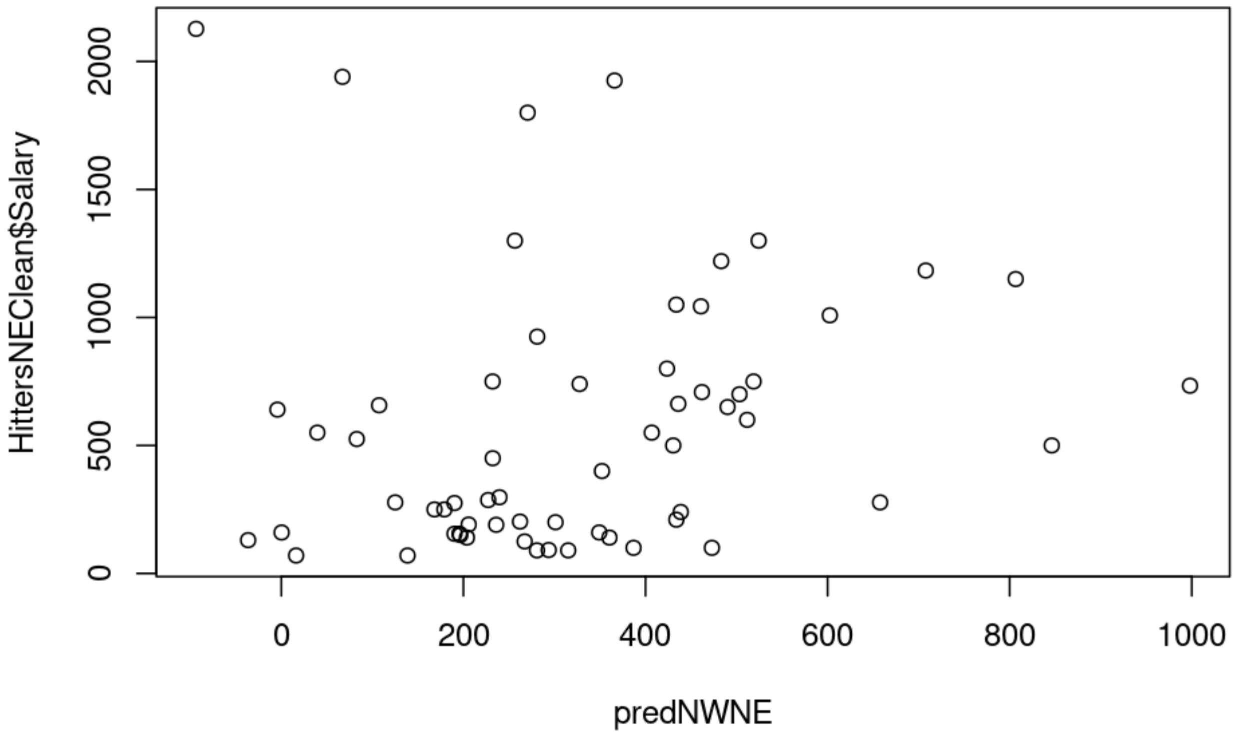
Correlation value: 0.774747

In this plot we can see the usage of the Lars model to demonstrate the correlation between the salaries in comparison to the NW division. A lot of the data points are clustered in the 600 to 100 range, and we see a strong positive correlation. It is clear that the relation appears to be pretty linear, and as seen by the correlation value of 0.774747, it is clear that the plot is a pretty good indication of the relationship between the salaries in comparison to the NW division.



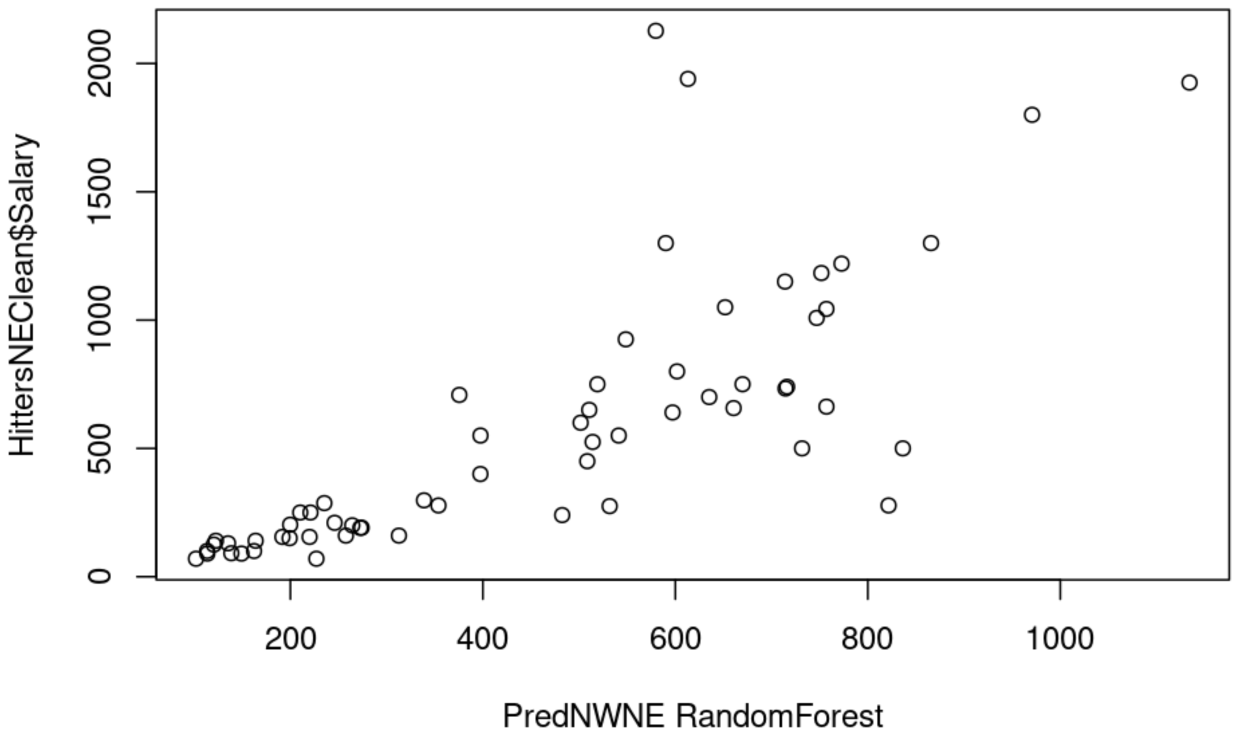
Correlation value: 0.598386

We can see that the random forest regression is less accurate than the Lars model for the NW division because the correlation value for the regression model is less than the Lars model by 0.20. The random forest model seems less clearly correlated than the Lars model, and though it looks like there is a positive relationship because as the x-axis values increase, so does the y-value. However, this positive correlation is weak/moderate from the x values from 600-1200 as they are distributed farther apart and more randomly apart. We see outliers at 1200 and 900.



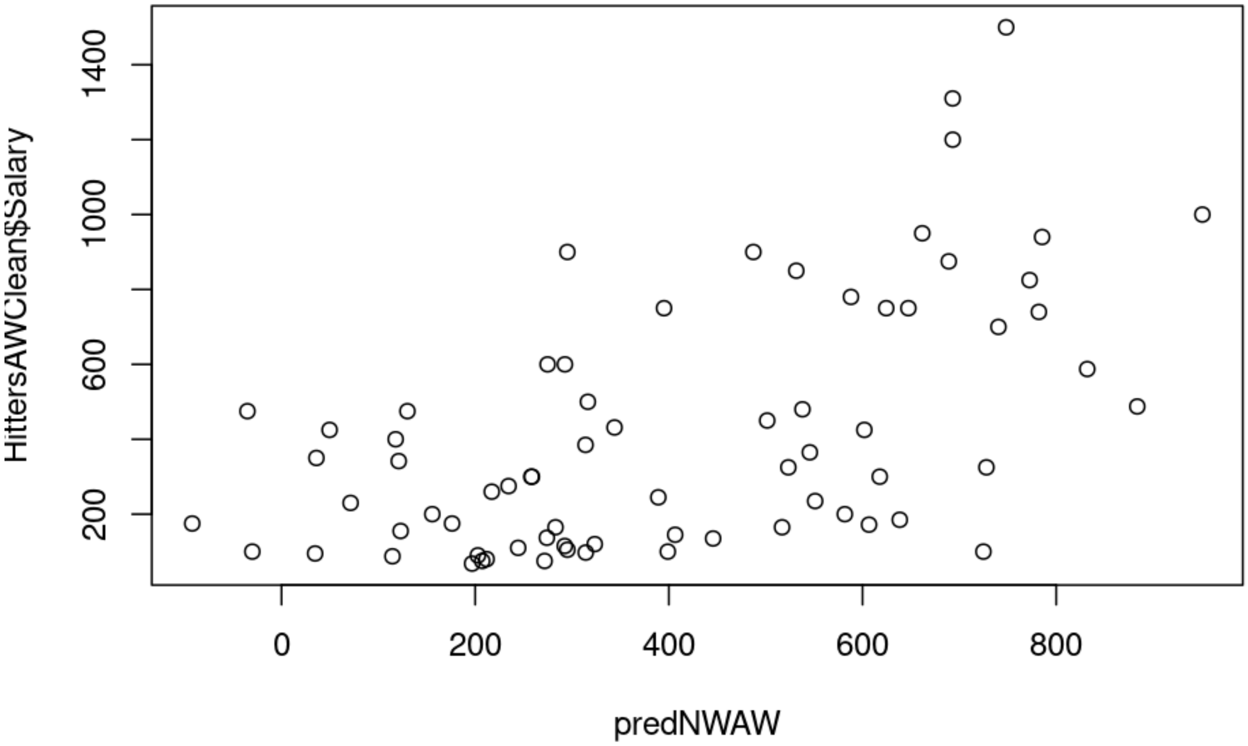
<u>Correlation value: 0.1454497</u>

Here, we can see the relationship between the League N division W and League N and Division E against the League N division E. We see that the correlation is very, very small, as it is 0.1454497. This indicates that there is no significant correlation, and it is a very weak relationship. The points generated show a very random relationship between salary and prediction. The distribution is widely spread across, and we see that there appear to be some clear outliers as the one at 2000 (y value) and 1000 (x value). Compared to the random forest, the correlation is much weaker, almost a 6th of the correlation shown for that of the random forest model.



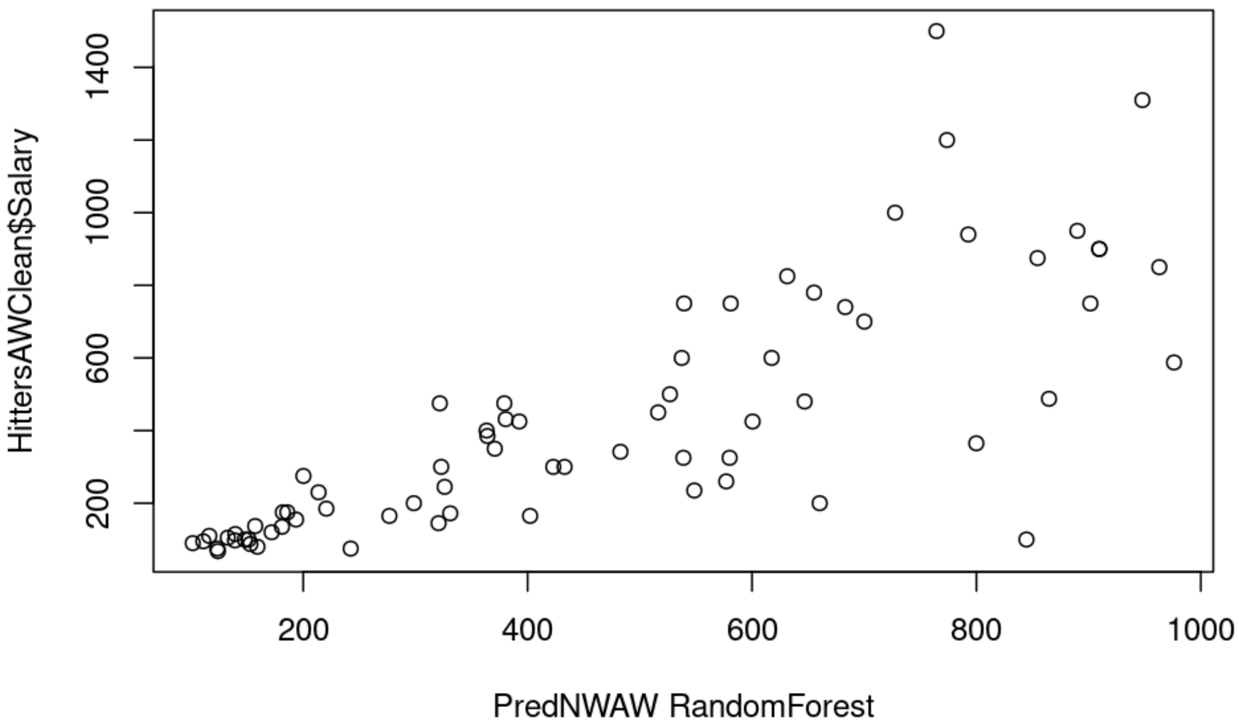
Correlation value: 0.7728121

This is the plot for the NWNE divisions via the RandomForest model, and the correlation value is 0.7728121, indicating a strong positive correlation between the salaries of employees in the two NWNE divisions being compared. This correlation value suggests that as the salaries in one division increase, there is a corresponding tendency for the salaries in the other division to increase as well. There is a cluster of points around the 200 mark on the x-axis and below the 500 mark on the y-axis. This plot has a higher correlation value than the Lars model, showing that this plot better represents the data.



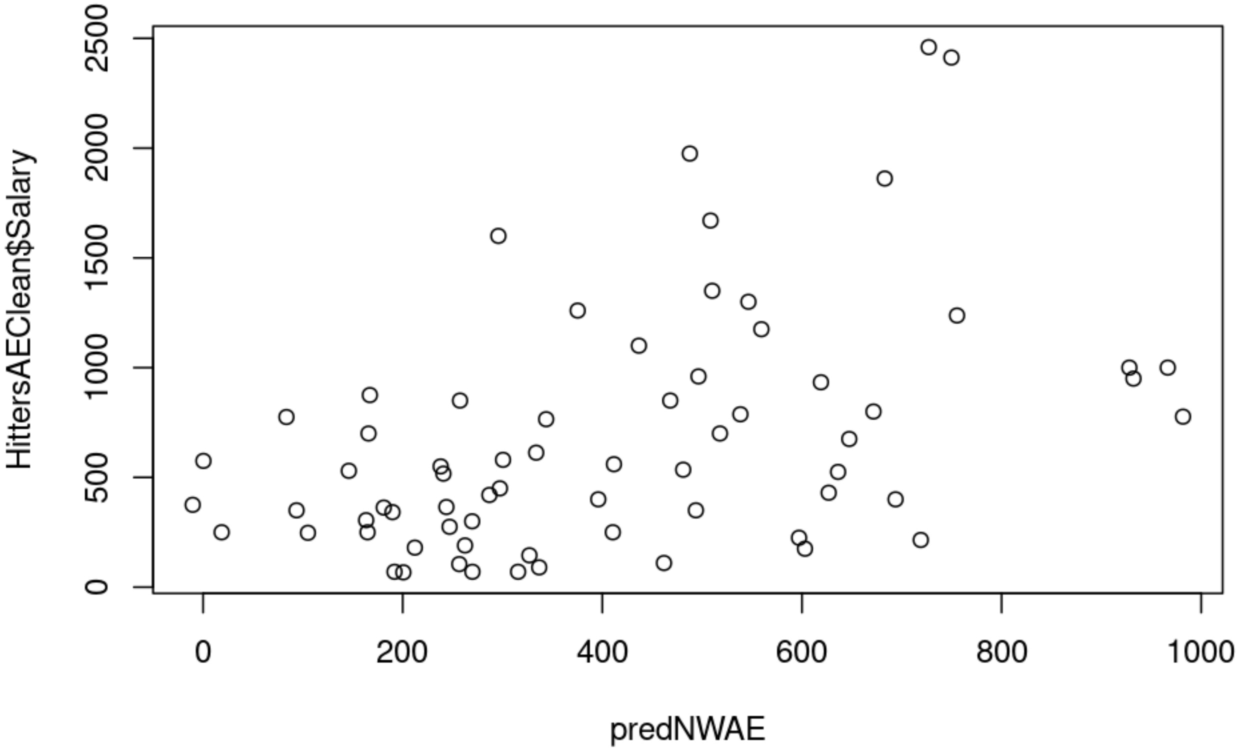
Correlation value: 0.5636847

This is the plot for the NWAW divisions, and the correlation between the data points is 0.5636847, which shows that there is a moderately positive correlation between the salaries. A correlation value of 0.5636847 suggests that as the salaries in one division increase, there is a tendency for the salaries in the other division to increase as well, but not in a perfectly linear manner. The plot reveals a very small cluster of points concentrated just above the 200 mark on the x-axis, implying that a portion of employees or positions in these divisions have salaries within a similar range, potentially representing entry-level or lower-tier roles.



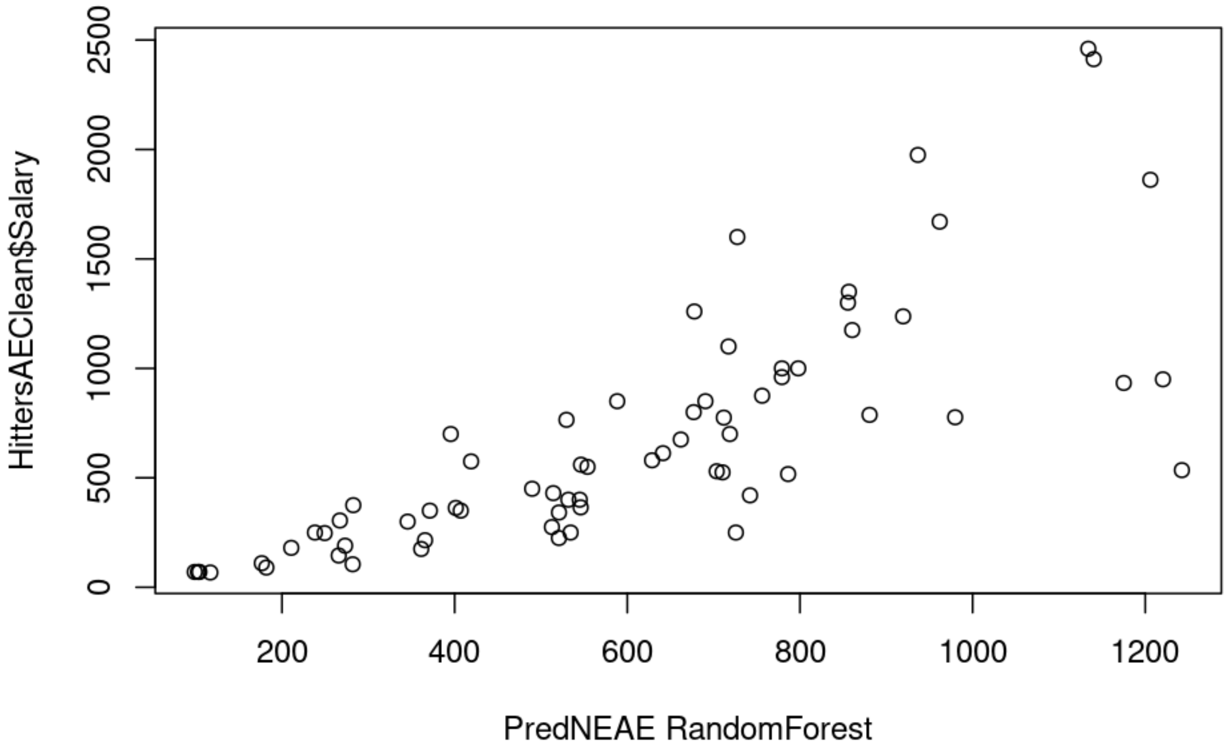
Correlation value: 0.7921005

Here, we can see that the correlation value derived from the plot that is generated from the Random Forest model for the NWAW division is 0.7921005. We can see a stronger, positive correlation in the model above compared to other division league comparisons. This correlation value indicates that as the salaries in one division increase, then there is a salary increase in other divisions as well. The only major clustering that occurs in this plot is between the 0-200 mark on the x-axis. After that, the rest of the data points are widely spread and there appears to be an outlier of the 700 mark on the x-axis.



Correlation value: 0.4554133

In this plot, we see the relationship between League N Division W and League A and Division E against League A Division E. The correlation is 0.4554133, indicating a somewhat moderate to weak correlation between the salary of employees in the NWAE division and the other divisions that it is compared against. This somewhat moderate relationship essentially implies that as the salaries for one division increase, it is likely that the salaries of the other divisions will increase alongside it. Compared to other divisions like NENW, this division has a smaller correlation value, meaning that though the salary of one division may increase, it will not be as directly proportional to the competing division increases. However, 0.4554133 indicates that it is still somewhat positive and significant.



Correlation value: 0.7731608

This plot showcases the relationship between League N Division E and League A Division E against League A Division E via the Random Forest model. The correlation value for this graph is 0.7731608, which indicates that there is a strong and positive relationship between the data points. In the context of the plot, this would mean that there is a strong positive relationship between the NEAE division's salaries and the other divisions whose salaries it is benign compared to. There are a couple of outliers near higher prediction values, which leads to a lower correlation value. Overall, the correlation value of Random Forest is higher than the Lars model, indicating that Random Forest was a better predictor of the data.