Stats Assignment 2

2024-10-11

Comparing the Salary Distribution Across Divisions

Setting up the environment and separating the data based on the divisions:

```
library("ISLR")
```

```
## Warning: package 'ISLR' was built under R version 4.3.3
```

```
library("tidyverse")
```

```
## Warning: package 'tidyverse' was built under R version 4.3.3

## Warning: package 'ggplot2' was built under R version 4.3.3

## Warning: package 'tidyr' was built under R version 4.3.3

## Warning: package 'readr' was built under R version 4.3.3

## Warning: package 'purrr' was built under R version 4.3.3

## Warning: package 'dplyr' was built under R version 4.3.3

## Warning: package 'stringr' was built under R version 4.3.3

## Warning: package 'forcats' was built under R version 4.3.3

## Warning: package 'lubridate' was built under R version 4.3.3

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library('moderndive')
```

```
## Warning: package 'moderndive' was built under R version 4.3.3
```

```r
library('skimr')
```

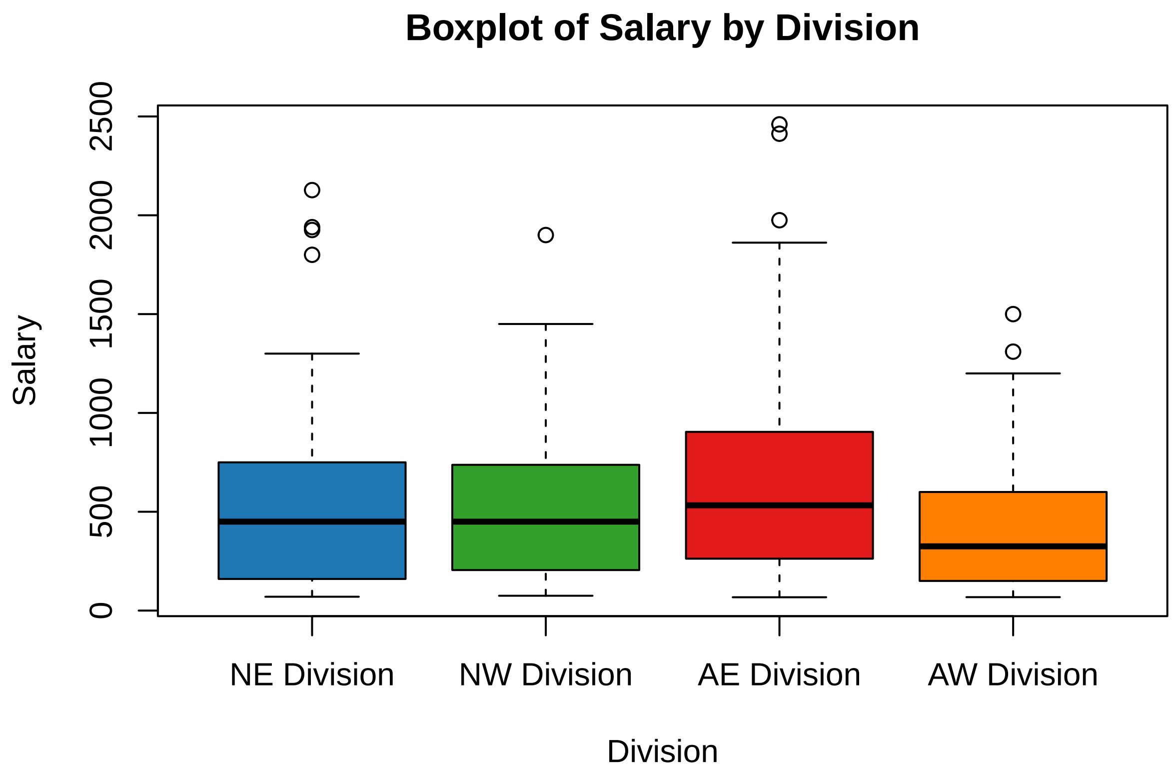
```
## Warning: package 'skimr' was built under R version 4.3.3
```

```r
library('lars')
```

```
## Loaded lars 1.3
```

```r
data <- Hitters


colors = c('red', 'blue', 'yellow', 'purple', 'green')
colors_vector <- c("#1f78b4", "#33a02c", "#e31a1c", "#ff7f00", "#6a3d9a",
                   "#b15928", "#a6cee3", "#b2df8a", "#fb9a99", "#fdbf6f",
                   "#cab2d6", "#ffff99", "#b15928", "#8dd3c7", "#bebada")


noNAHitter <- na.omit(data)
hittersN <- noNAHitter[noNAHitter$League == 'N',]
hittersA <- noNAHitter[noNAHitter$League == 'A',]

hittersNE <- hittersN[hittersN$Division == 'E',]
hittersNW <- hittersN[hittersN$Division == 'W',]
hittersAE <- hittersA[hittersA$Division == 'E',]
hittersAW <- hittersA[hittersA$Division == 'W',]
```

```r
# Boxplot of salaries with division labels
boxplot(hittersNE$Salary, hittersNW$Salary,
        hittersAE$Salary, hittersAW$Salary,
        col = colors_vector,
        names = c("NE Division", "NW Division", "AE Division", "AW Division"),  # Adding labels
        main = "Boxplot of Salary by Division",
        xlab = "Division",
        ylab = "Salary")
```

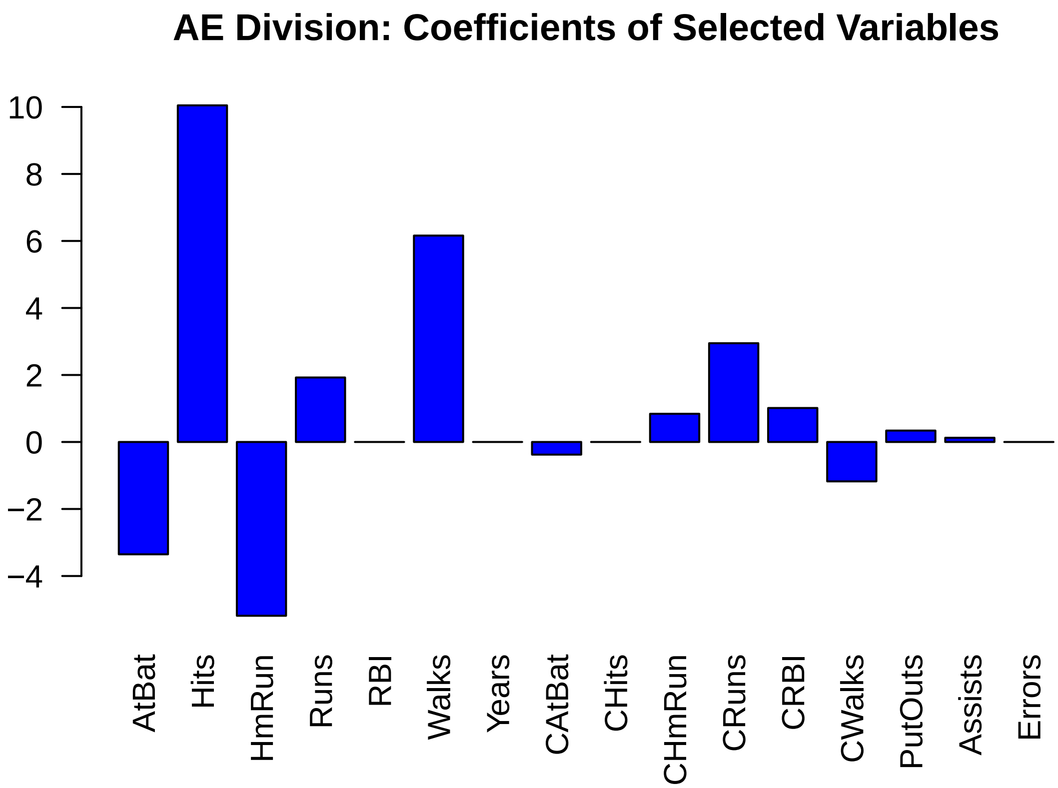


The median salaries seem to be around the same level for all the divisions, with the it being slightly higher for the AE division with a larger spread. AW seems to a slightly lower median and less spread as well, but overall the median salaries seem to be around the same level.

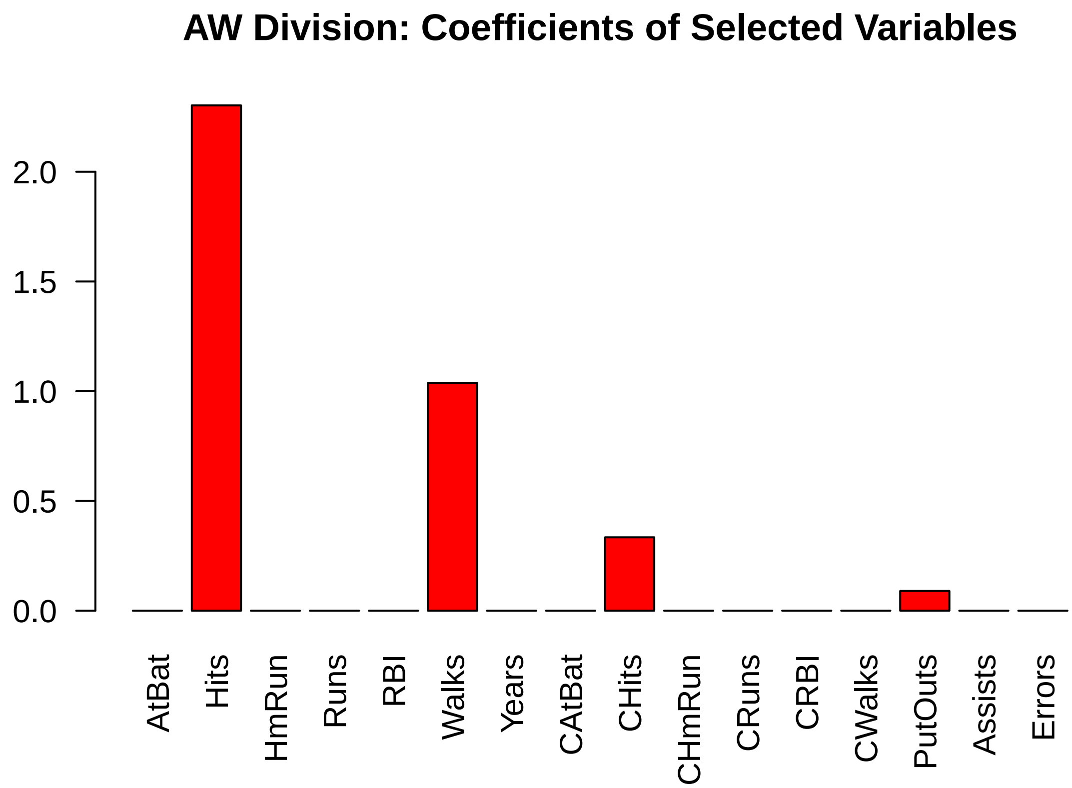
linear regression with LARS using Cp

Creating lars models for all the divisions using the numerical variables to find the significant predictors for salary for each division:

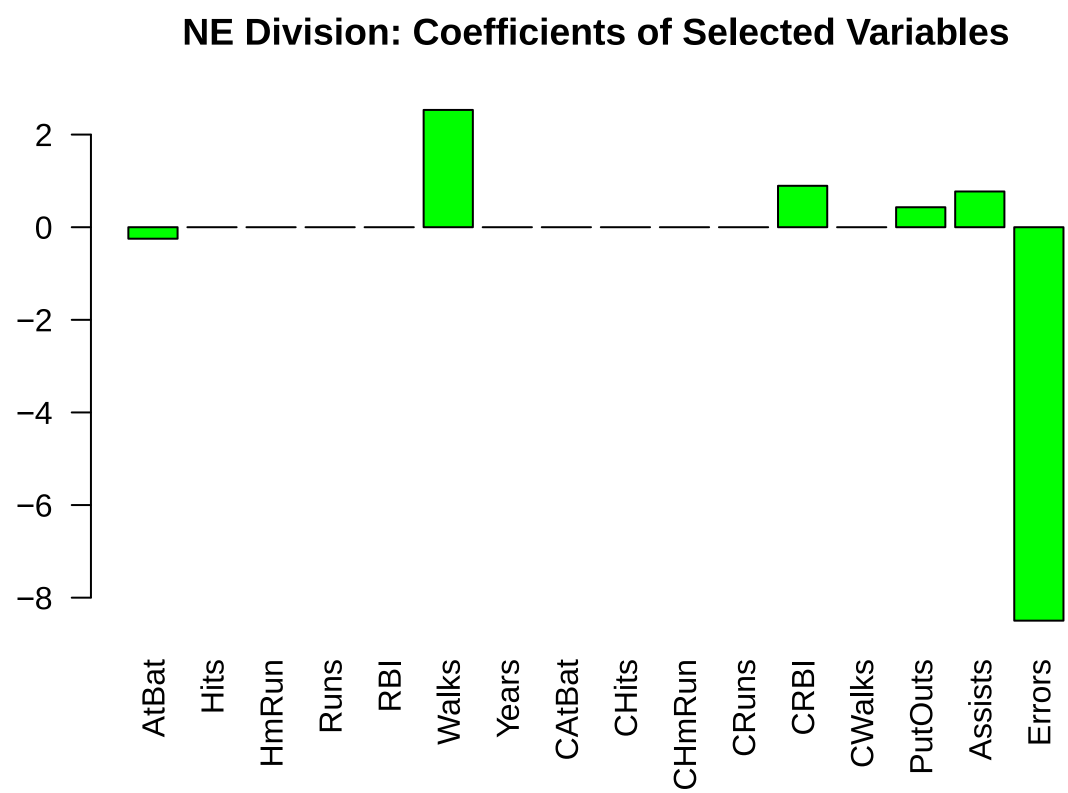
```
#seperating the numerical variables to do linear regression
numerical_cols <- c("AtBat", "Hits", "HmRun", "Runs", "RBI", "Walks",
                    "Years", "CAtBat", "CHits", "CHmRun", "CRuns",
                    "CRBI", "CWalks", "PutOuts", "Assists", "Errors")


# For AE Division
XAE <- hittersAE[, -c(20, 19, 14, 15)]  # Remove non-numeric columns
YAE <- hittersAE$Salary
larsAE <- lars(as.matrix(XAE), YAE)
chosen_index_AE <- which.min(larsAE$Cp)  # Choose model with smallest Cp
selected_AE <- larsAE$beta[chosen_index_AE,]

# Barplot for AE Division
barplot(selected_AE, main = "AE Division: Coefficients of Selected Variables",
        col = "blue", las = 2)
```



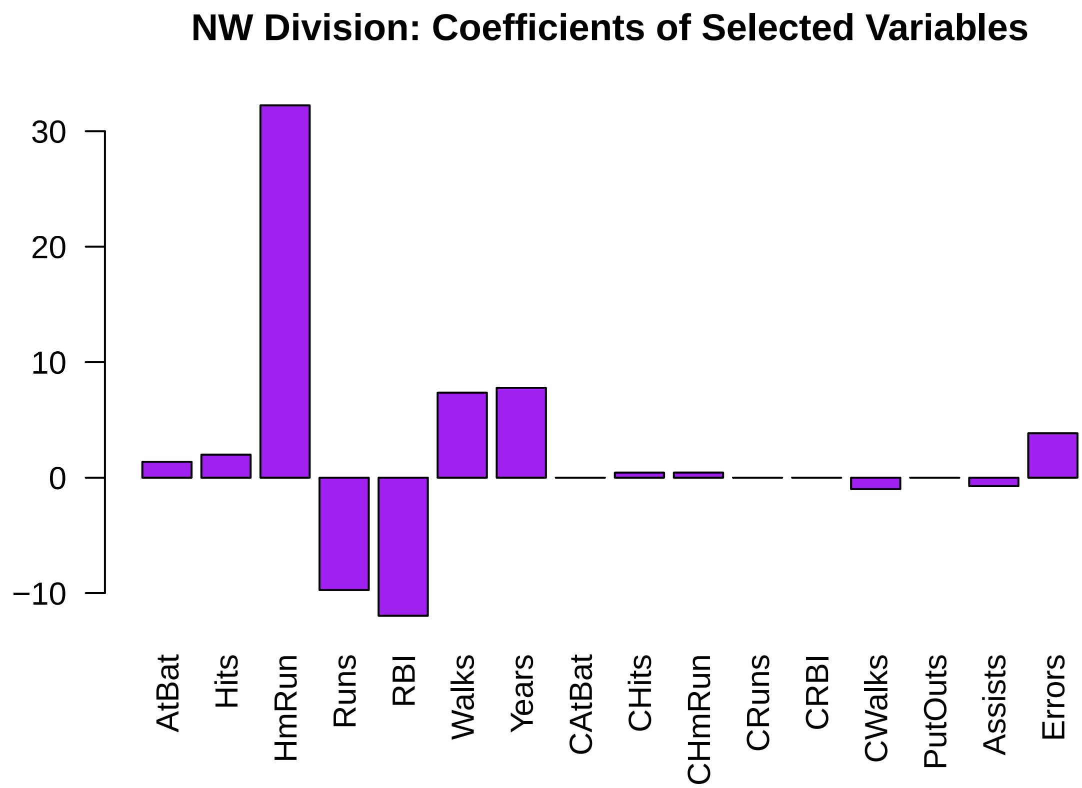

```r
# For AW Division
XAW <- hittersAW[, -c(20, 19, 14, 15)]  # Remove non-numeric columns
YAW <- hittersAW$Salary
larsAW <- lars(as.matrix(XAW), YAW)
chosen_index_AW <- which.min(larsAW$Cp)  # Choose model with smallest Cp
selected_AW <- larsAW$beta[chosen_index_AW,]

# Barplot for AW Division
barplot(selected_AW, main = "AW Division: Coefficients of Selected Variables",
        col = "red", las = 2)
```




```r
# For NE Division
XNE <- hittersNE[, -c(20, 19, 14, 15)]  # Remove non-numeric columns
YNE <- hittersNE$Salary
larsNE <- lars(as.matrix(XNE), YNE)
chosen_index_NE <- which.min(larsNE$Cp)  # Choose model with smallest Cp
selected_NE <- larsNE$beta[chosen_index_NE,]

# Barplot for NE Division
barplot(selected_NE, main = "NE Division: Coefficients of Selected Variables",
        col = "green", las = 2)
```



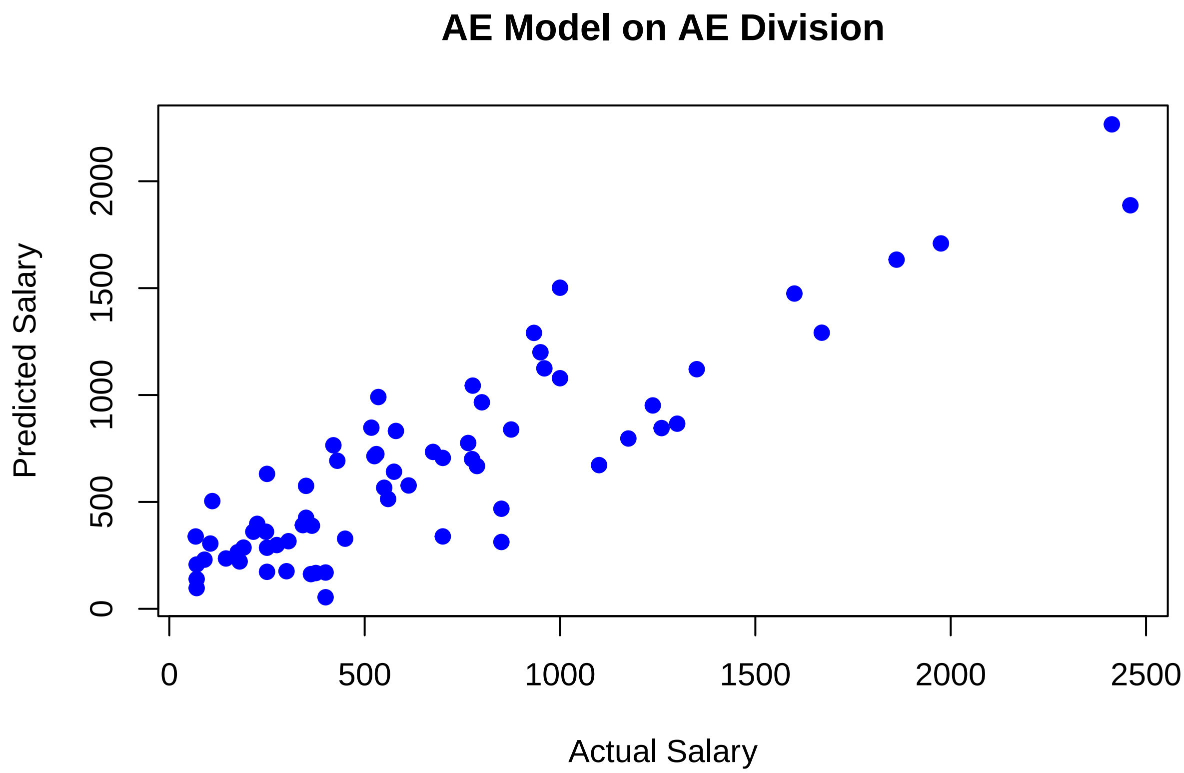

```r
# For NW Division
XNW <- hittersNW[, -c(20, 19, 14, 15)]  # Remove non-numeric columns
YNW <- hittersNW$Salary
larsNW <- lars(as.matrix(XNW), YNW)
chosen_index_NW <- which.min(larsNW$Cp)  # Choose model with smallest Cp
selected_NW <- larsNW$beta[chosen_index_NW,]

# Barplot for NW Division
barplot(selected_NW, main = "NW Division: Coefficients of Selected Variables",
        col = "purple", las = 2)
```



Plotting the predictors and their impact on the salary, it is evident the a variety of things affect the player salary differently when it comes to the AE division, with the Hits playing the strongest role, followed by the walks. Unlike the AE division, the AW division's salaries seems to have a small amount of correlation with the Hits, walks and PutOuts, while all other predictors seem to have 0 effect on the salary. The NE division's salaries seem to be strongly influenced negatively by the Errors, followed by some effect from the walks. The NW division's salaries seem to be strongly influenced by the Home runs, with significant negative influence from the Runs and RBI, followed by some positive influence based on the Walks and Years.

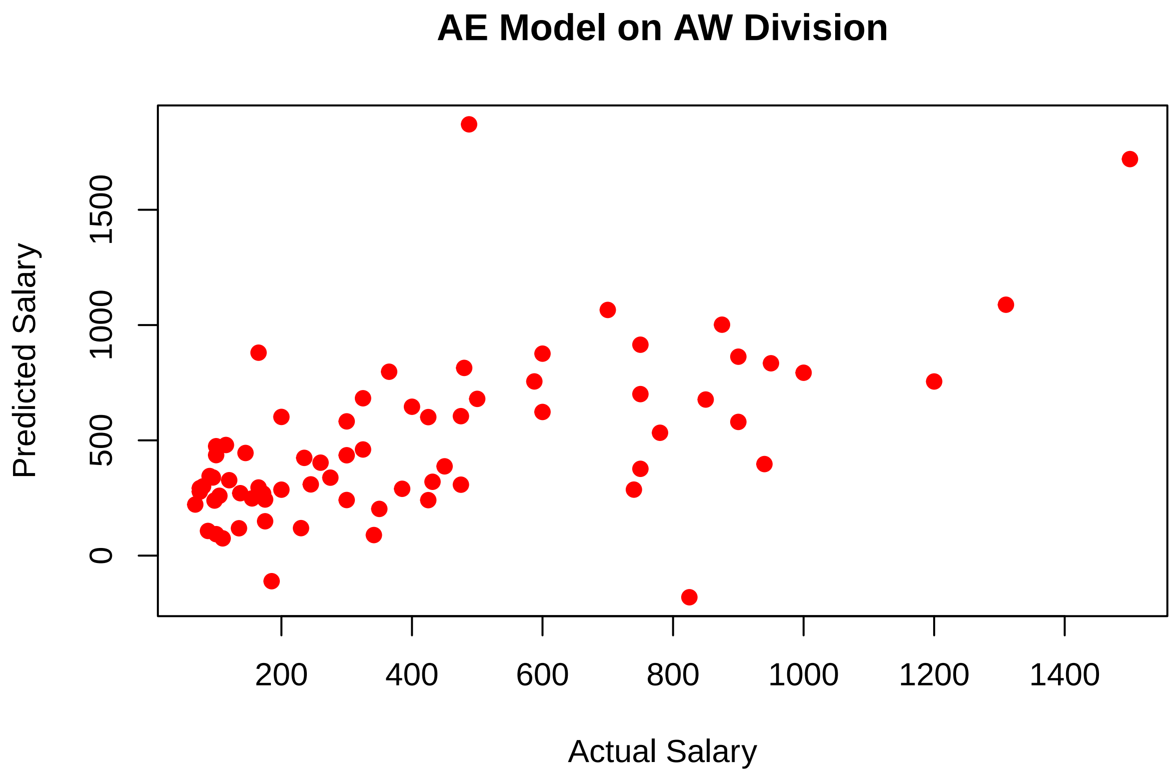
Applying the models of all the divisions onto all the divisions. Generating 16 scatterplots:

```r
# For AE Model:
predicted_salaries_AE_AE <- predict(larsAE, as.matrix(XAE), s = chosen_index_AE, type = "fit")$fit  # A
predicted_salaries_AE_AW <- predict(larsAE, as.matrix(XAW), s = chosen_index_AE, type = "fit")$fit  # A
predicted_salaries_AE_NE <- predict(larsAE, as.matrix(XNE), s = chosen_index_AE, type = "fit")$fit  # A
predicted_salaries_AE_NW <- predict(larsAE, as.matrix(XNW), s = chosen_index_AE, type = "fit")$fit  # A

# Scatterplot for AE model on AE division
plot(hittersAE$Salary, predicted_salaries_AE_AE,
     main = "AE Model on AE Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "blue", pch = 19)
```

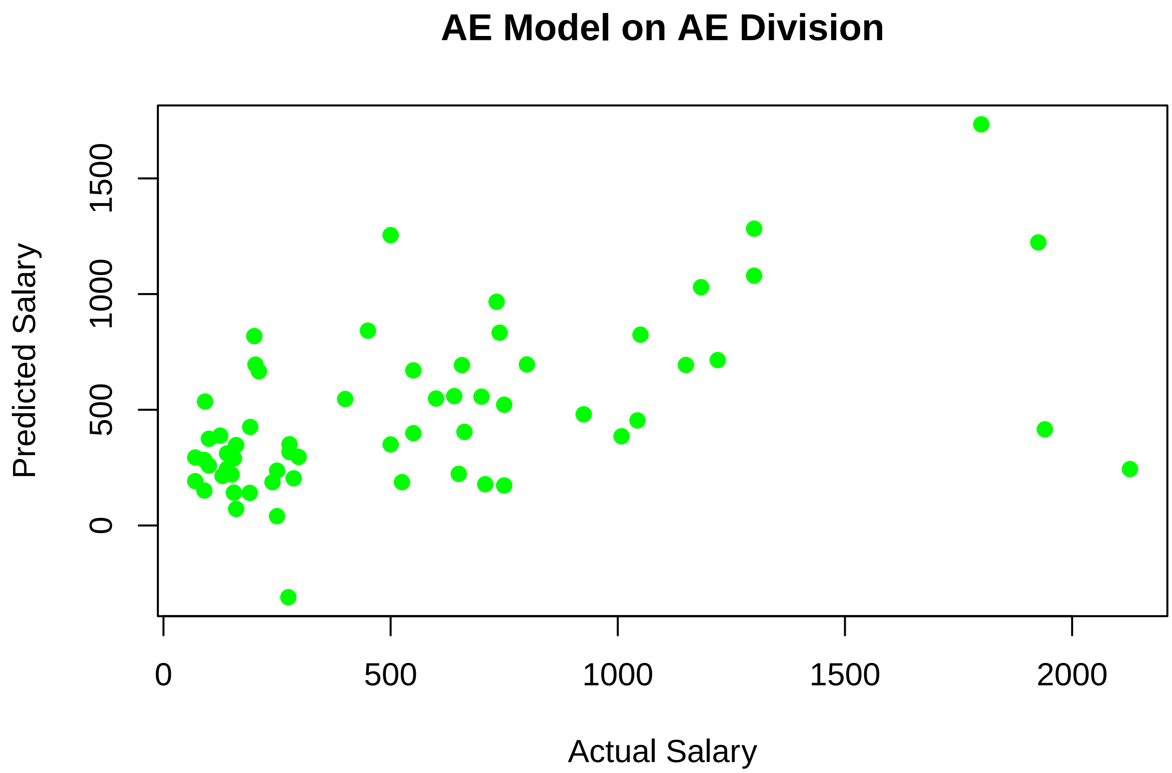


For AE Model:

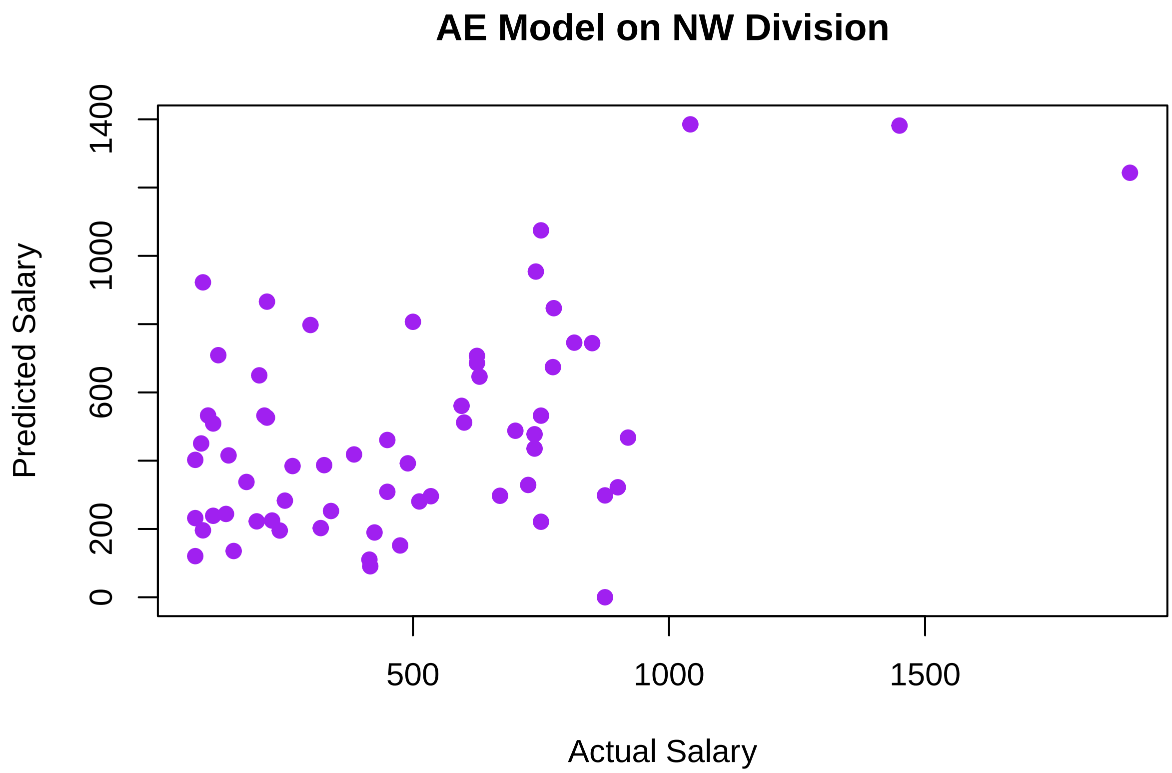
```
# Scatterplot for AE model on AW division
plot(hittersAW$Salary, predicted_salaries_AE_AW,
     main = "AE Model on AW Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "red", pch = 19)
```




```r
# Scatterplot for AE model on NE division
plot(hittersNE$Salary, predicted_salaries_AE_NE,
     main = "AE Model on AE Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "green", pch = 19)
```



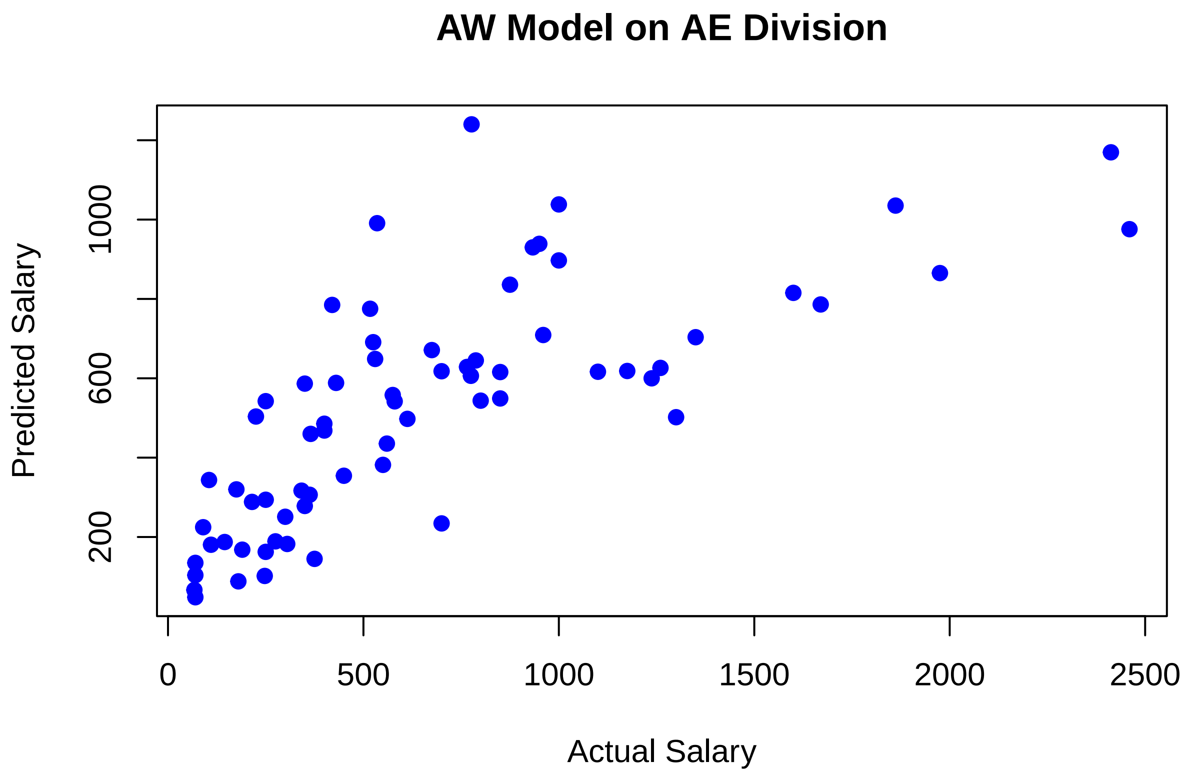

```r
# Scatterplot for AE model on AW division
plot(hittersNW$Salary, predicted_salaries_AE_NW,
     main = "AE Model on NW Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "purple", pch = 19)
```



It is quite evident when looking at the scatterplots that the model for the AE divisions seems to work best when applied to the data from the AE division. When applied to the AW division, there does seem to be some resemblance of a pattern, but it is definitely not a significant enough to establish a clear relationship. This makes sense, since if look at the bar graphs for the AE and AW divisions' predictors, we find that that both the divisions share the same top 2 predictors, Hits and Walks. However, the model does not work for AW because the significance of these 2 predictors differs a lot between AE and AW, and AE has many other significant predictors that play a role.

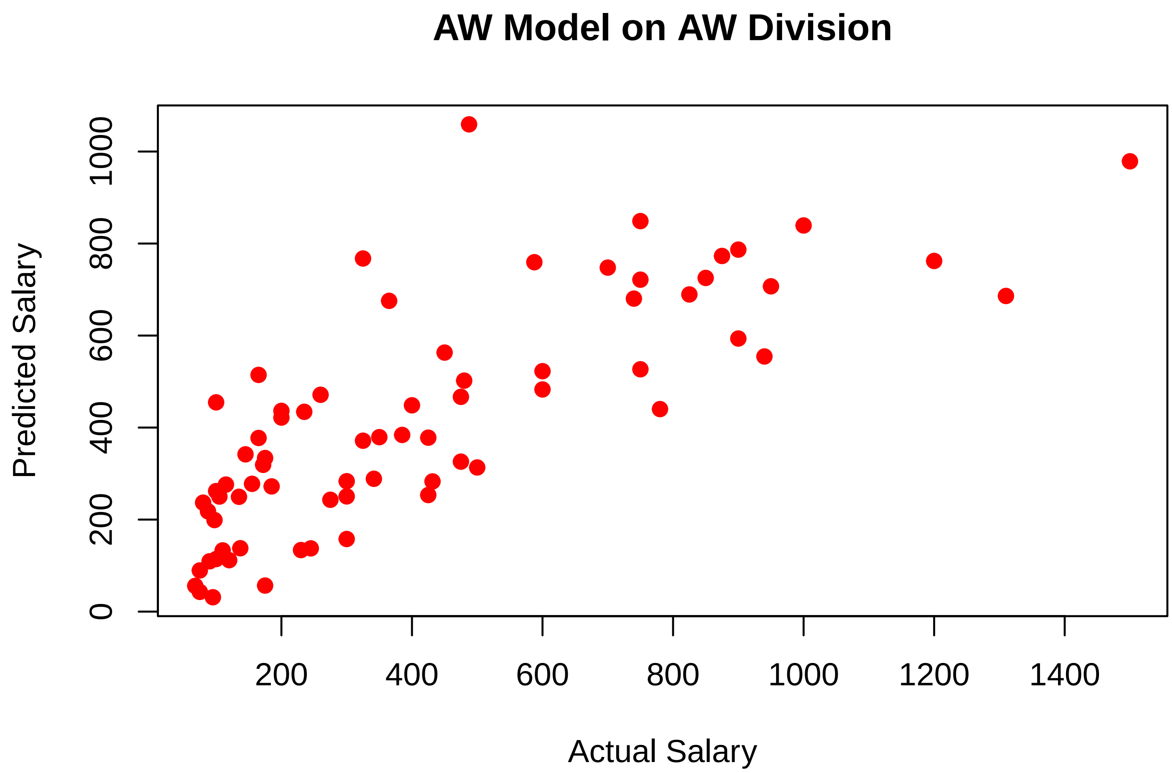
When it comes to applying the AE model onto NE and NW, it seems to be completely worthless

```
# For AW Model:
predicted_salaries_AW_AE <- predict(larsAW, as.matrix(XAE), s = chosen_index_AW, type = "fit")$fit  # A
predicted_salaries_AW_AW <- predict(larsAW, as.matrix(XAW), s = chosen_index_AW, type = "fit")$fit  # A
predicted_salaries_AW_NE <- predict(larsAW, as.matrix(XNE), s = chosen_index_AW, type = "fit")$fit  # A
predicted_salaries_AW_NW <- predict(larsAW, as.matrix(XNW), s = chosen_index_AW, type = "fit")$fit  # A

# Scatterplot for AW model on AE division
plot(hittersAE$Salary, predicted_salaries_AW_AE,
     main = "AW Model on AE Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "blue", pch = 19)
```

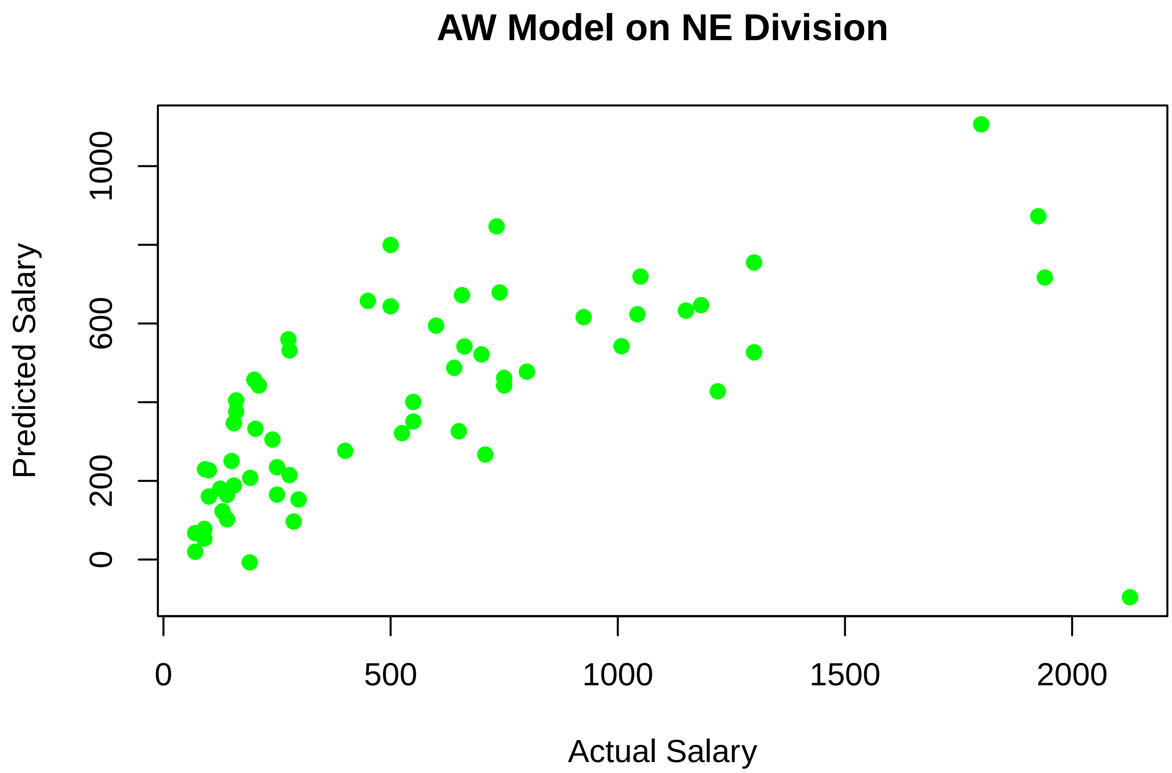


For AW Model:

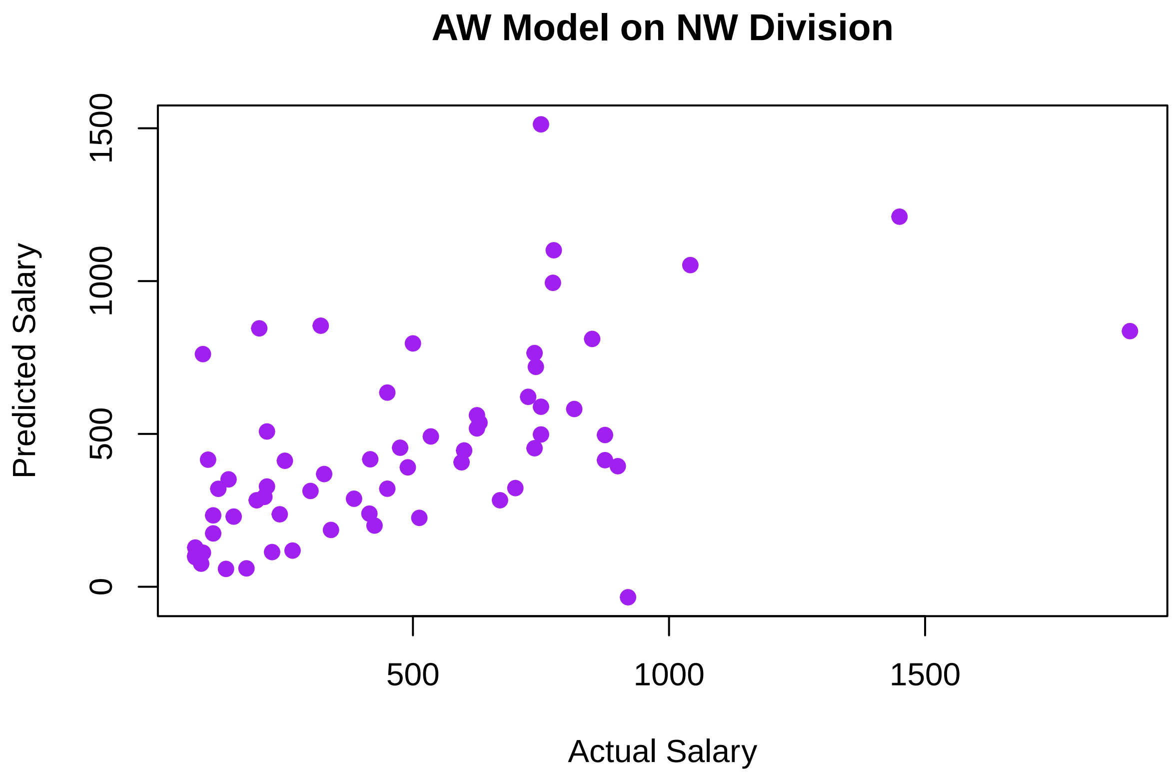
```
# Scatterplot for AW model on AW division
plot(hittersAW$Salary, predicted_salaries_AW_AW,
     main = "AW Model on AW Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "red", pch = 19)
```




```r
# Scatterplot for AW model on NE division
plot(hittersNE$Salary, predicted_salaries_AW_NE,
     main = "AW Model on NE Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "green", pch = 19)
```



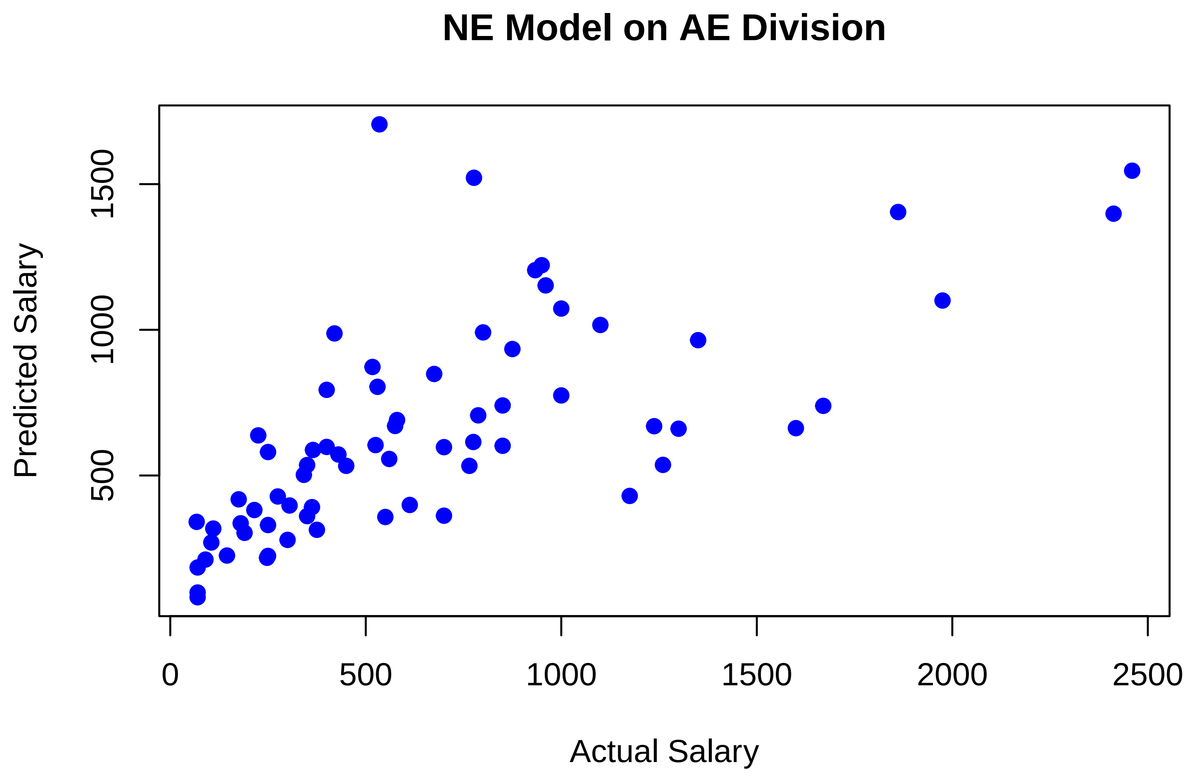

```r
# Scatterplot for AW model on NW division
plot(hittersNW$Salary, predicted_salaries_AW_NW,
     main = "AW Model on NW Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "purple", pch = 19)
```



The AW model does function somewhat well for its own division, producing somewhat of a linear pattern although not as good as the one that the AE model could with the AE division since it spreads outward into the high salary ranges. It may be due to the fact the the AW model's predictors have much lower significance as compared to the AE model's predictors. A similar shape can also be seen when trying to use the AW model on the AE division, which does not work well similar to when we tried applying the AE model onto the AW division.

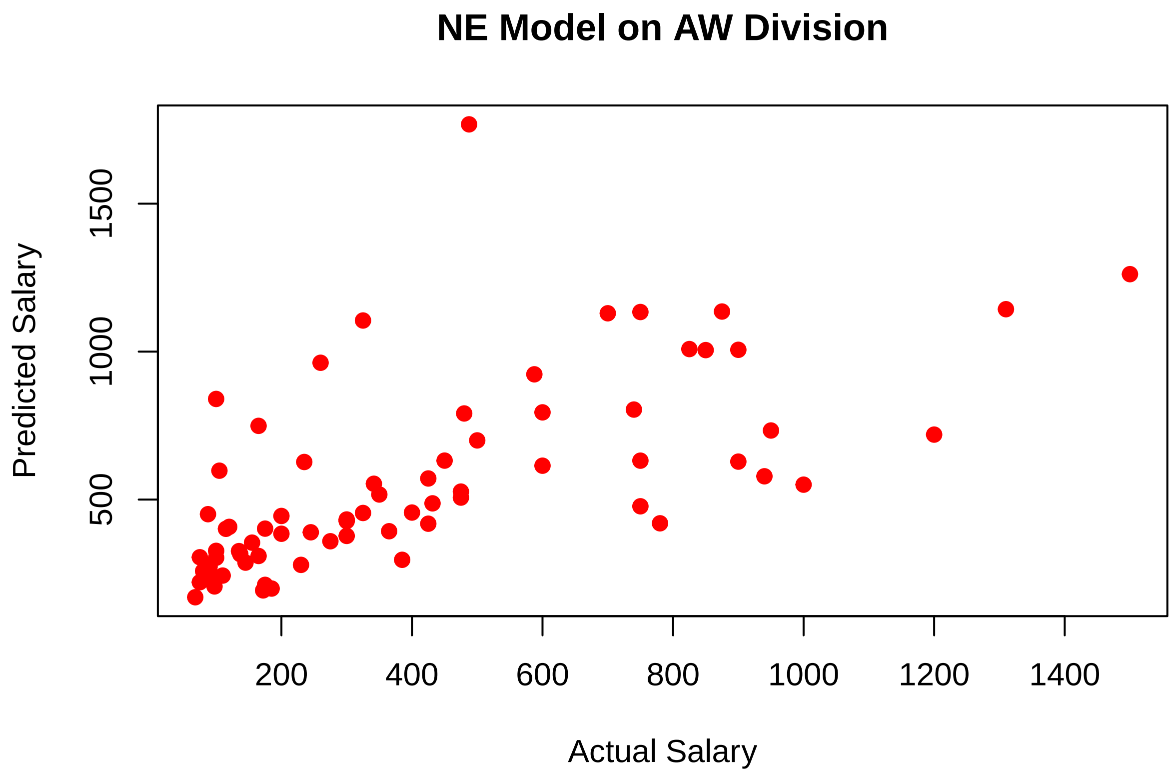
The AW model does not work well with the NE and NW divisions as well.

```
# For NE Model:
predicted_salaries_NE_AE <- predict(larsNE, as.matrix(XAE), s = chosen_index_NE, type = "fit")$fit  # N
predicted_salaries_NE_AW <- predict(larsNE, as.matrix(XAW), s = chosen_index_NE, type = "fit")$fit  # N
predicted_salaries_NE_NE <- predict(larsNE, as.matrix(XNE), s = chosen_index_NE, type = "fit")$fit  # N
predicted_salaries_NE_NW <- predict(larsNE, as.matrix(XNW), s = chosen_index_NE, type = "fit")$fit  # N

# Scatterplot for NE model on AE division
plot(hittersAE$Salary, predicted_salaries_NE_AE,
     main = "NE Model on AE Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "blue", pch = 19)
```

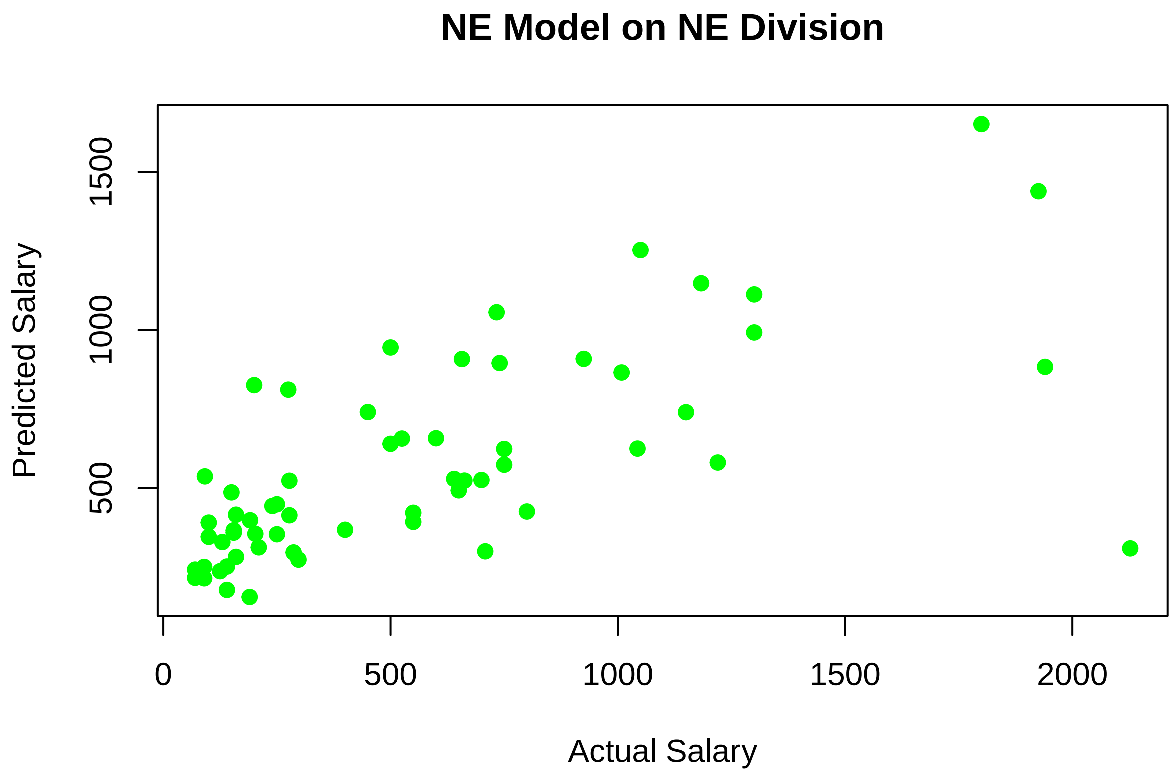


For NE Model:

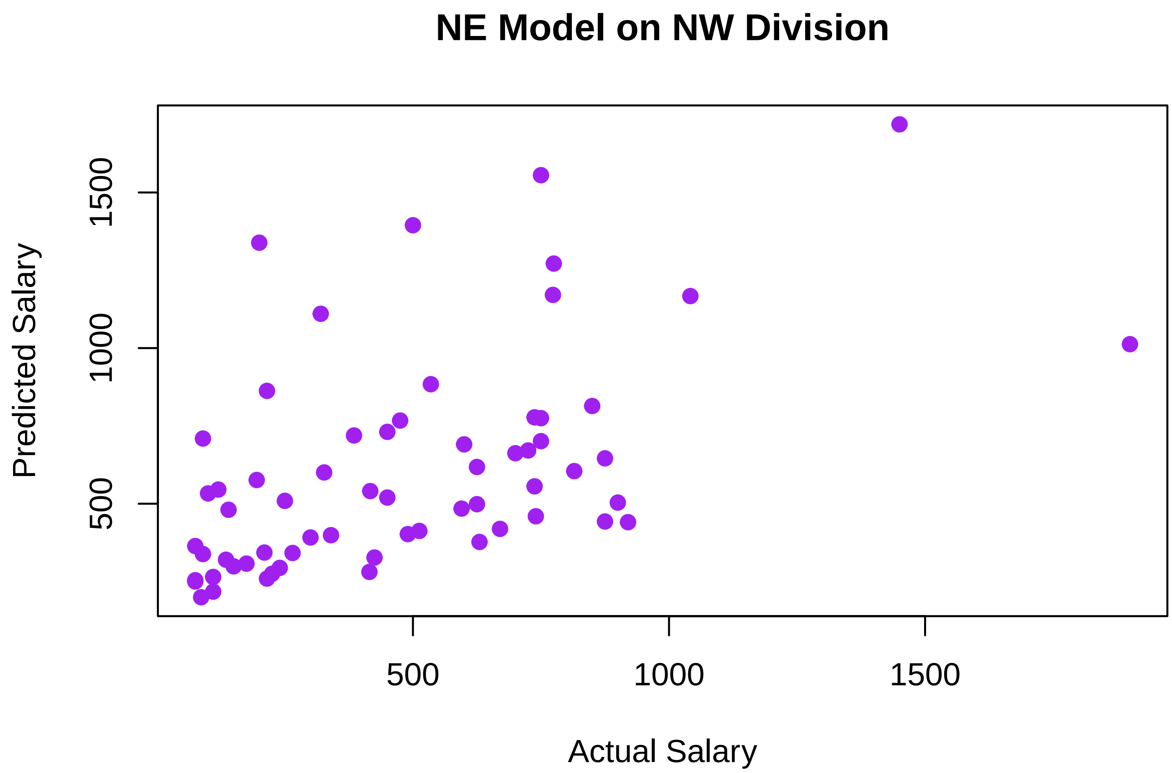
```
# Scatterplot for NE model on AW division
plot(hittersAW$Salary, predicted_salaries_NE_AW,
     main = "NE Model on AW Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "red", pch = 19)
```




```r
# Scatterplot for NE model on NE division
plot(hittersNE$Salary, predicted_salaries_NE_NE,
     main = "NE Model on NE Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "green", pch = 19)
```

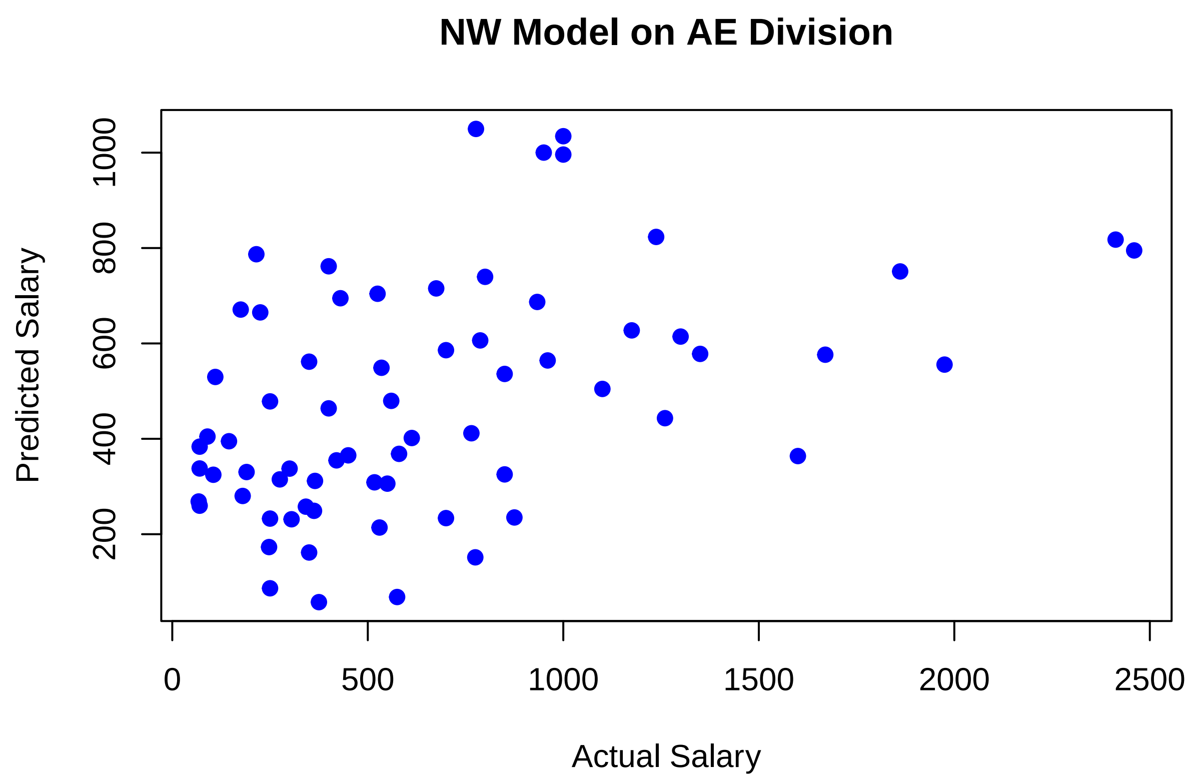



```r
# Scatterplot for NE model on NW division
plot(hittersNW$Salary, predicted_salaries_NE_NW,
     main = "NE Model on NW Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "purple", pch = 19)
```



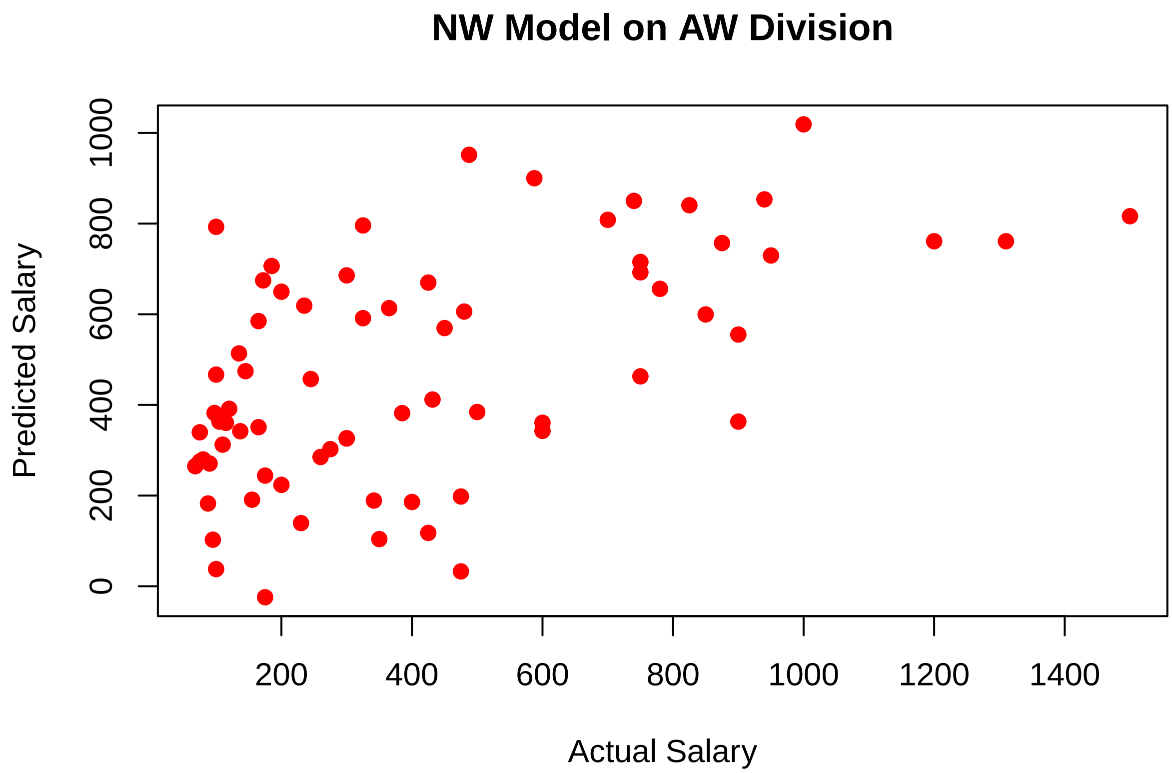
As expected the NE model works only when it comes to the NE division, but even in that case the scatterplot is too spread out, has clusters of points and does not give us a clear pattern. Naturally, it is just as non-functional if not worse when it comes to using for the other divisions.

```
# For NW Model:
predicted_salaries_NW_AE <- predict(larsNW, as.matrix(XAE), s = chosen_index_NW, type = "fit")$fit  # N
predicted_salaries_NW_AW <- predict(larsNW, as.matrix(XAW), s = chosen_index_NW, type = "fit")$fit  # N
predicted_salaries_NW_NE <- predict(larsNW, as.matrix(XNE), s = chosen_index_NW, type = "fit")$fit  # N
predicted_salaries_NW_NW <- predict(larsNW, as.matrix(XNW), s = chosen_index_NW, type = "fit")$fit  # N

# Scatterplot for NW model on AE division
plot(hittersAE$Salary, predicted_salaries_NW_AE,
     main = "NW Model on AE Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "blue", pch = 19)
```

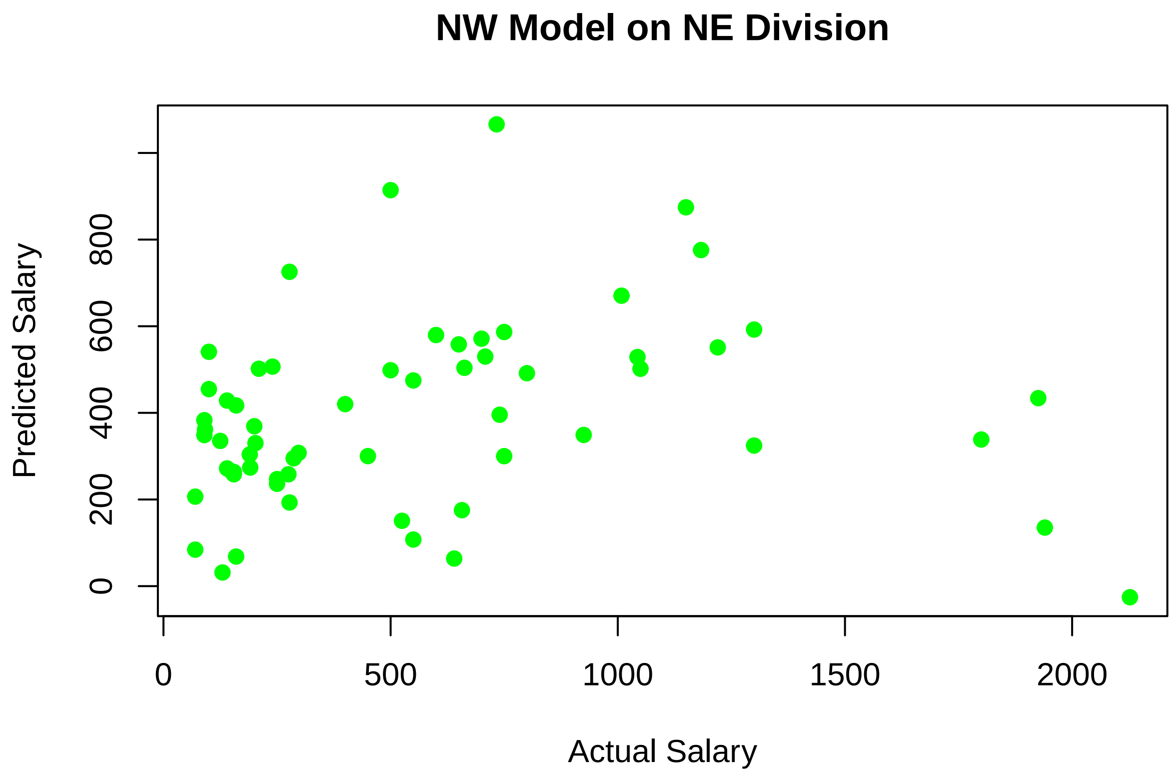


For NW Model:

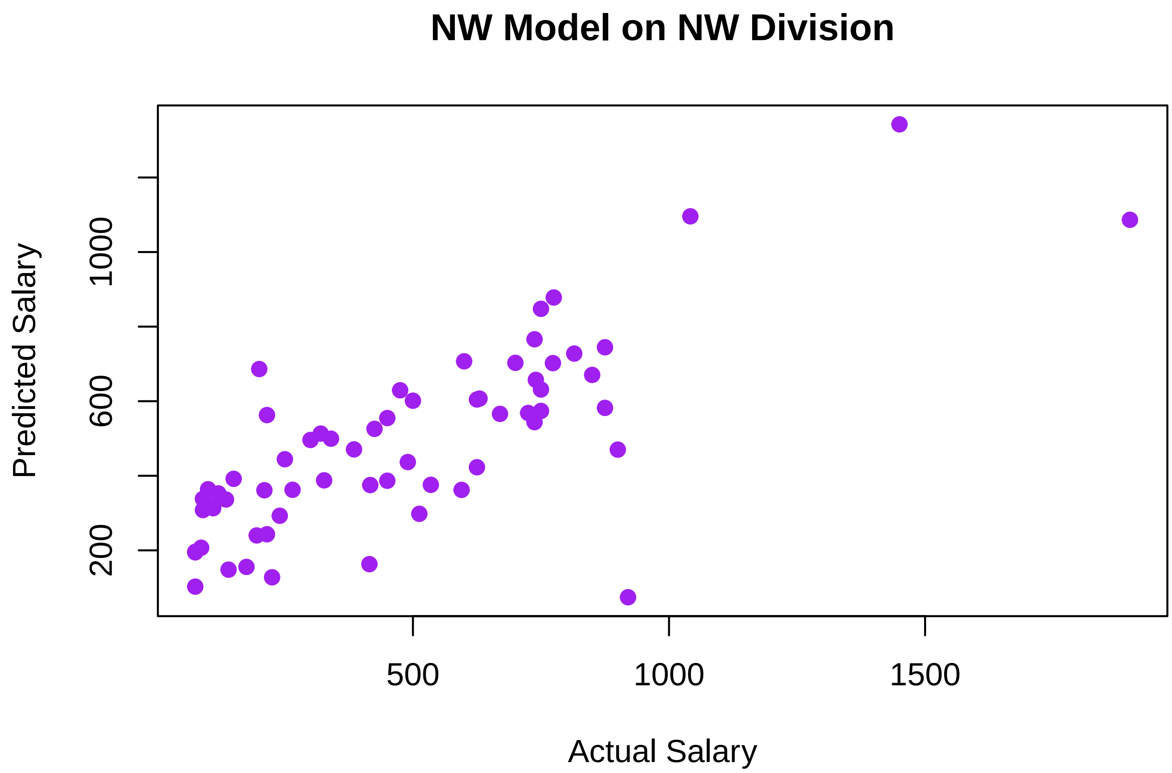
```
# Scatterplot for NW model on AW division
plot(hittersAW$Salary, predicted_salaries_NW_AW,
     main = "NW Model on AW Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "red", pch = 19)
```




```r
# Scatterplot for NW model on NE division
plot(hittersNE$Salary, predicted_salaries_NW_NE,
     main = "NW Model on NE Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "green", pch = 19)
```




```r
# Scatterplot for NW model on NW division
plot(hittersNW$Salary, predicted_salaries_NW_NW,
     main = "NW Model on NW Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "purple", pch = 19)
```

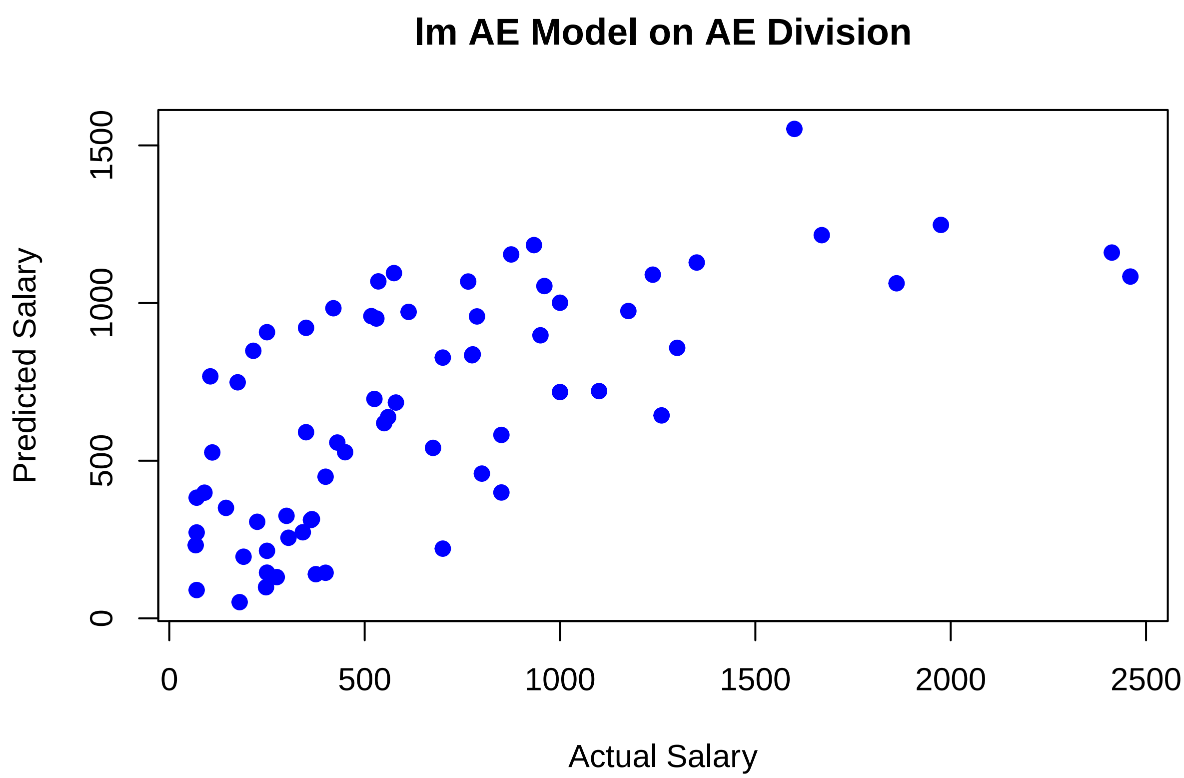


The NW model works somewhat only when predicting for the NW division. It seems to be worthless when applying it to any other division. This can be explained somewhat since significance and variety of the predictors of this model does not match at all with any other model.

Using lm Models

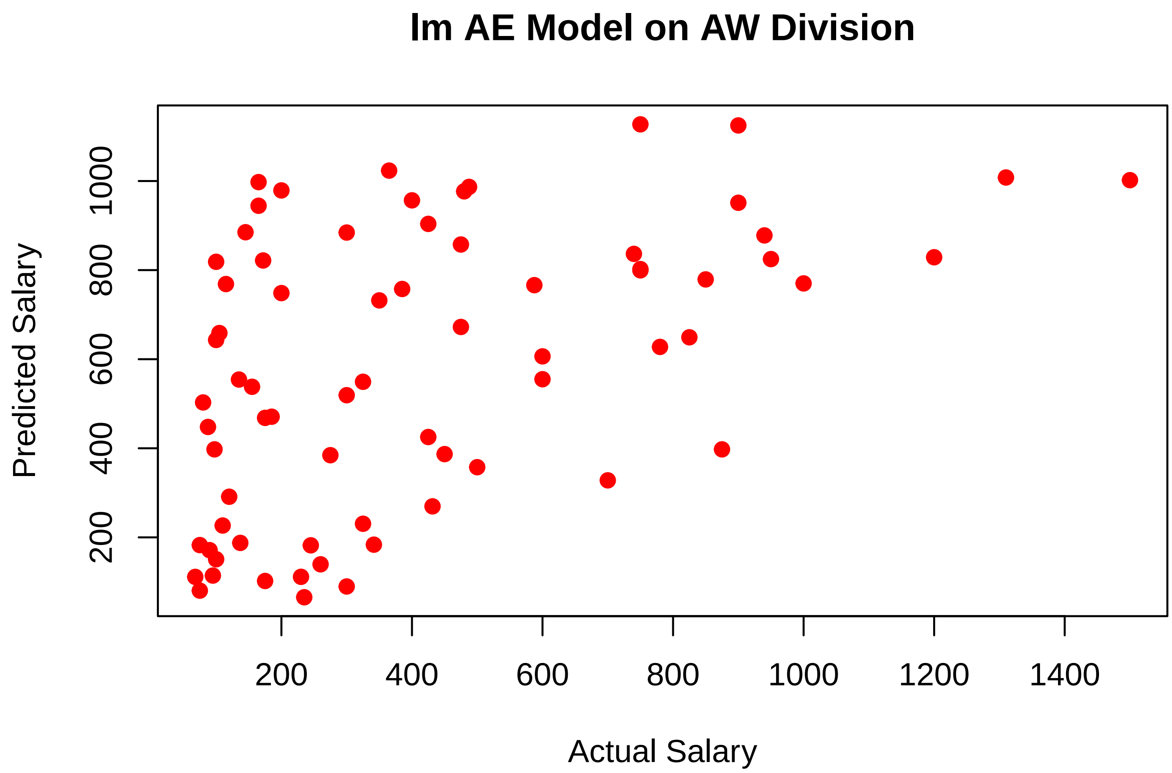
Applying the models of all the divisions onto all the divisions. Generating 16 scatterplots:

```
# AE lm Model
lm_AE <- lm(Salary ~ Hits + Walks, data = hittersAE)  # Add variables chosen by LARS
predicted_salaries_lm_AE_AE <- predict(lm_AE, hittersAE)  # lm AE model on AE division
predicted_salaries_lm_AE_AW <- predict(lm_AE, hittersAW)  # lm AE model on AW division
predicted_salaries_lm_AE_NE <- predict(lm_AE, hittersNE)  # lm AE model on AE division
predicted_salaries_lm_AE_NW <- predict(lm_AE, hittersNW)  # lm AE model on AW division

# Scatterplot for lm AE model on AE division
plot(hittersAE$Salary, predicted_salaries_lm_AE_AE,
     main = "lm AE Model on AE Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "blue", pch = 19)
```

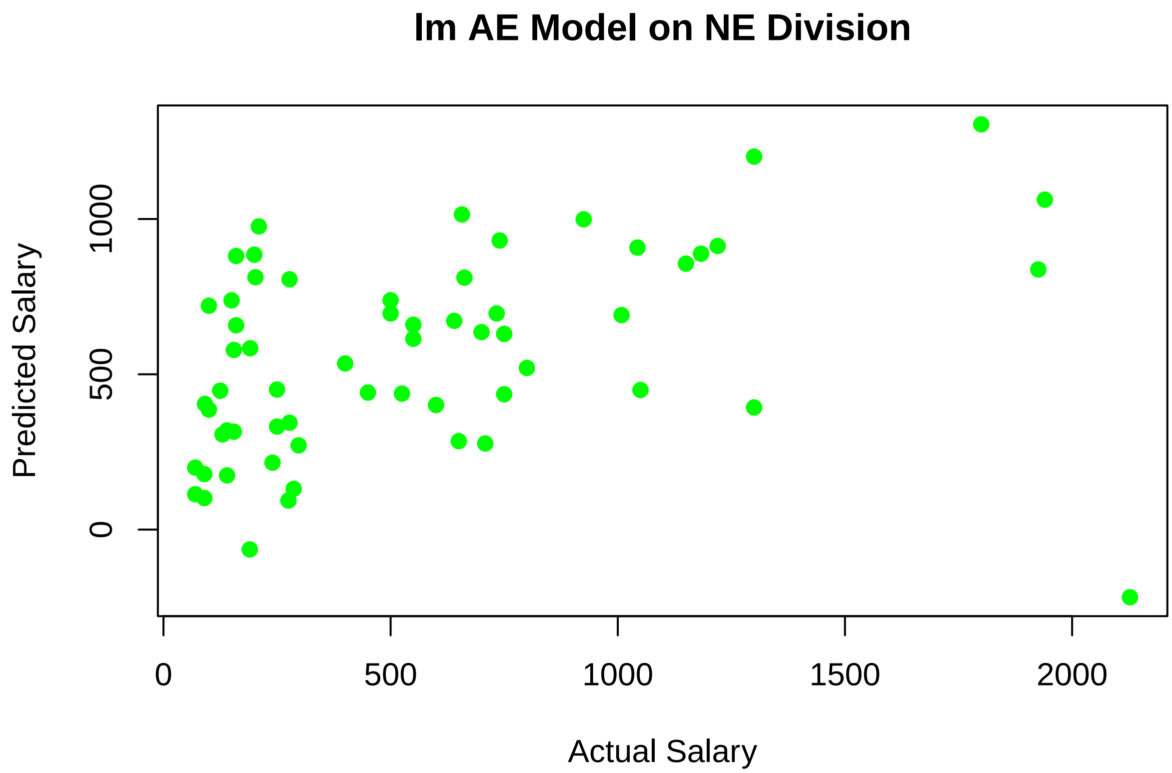


For AE Model:

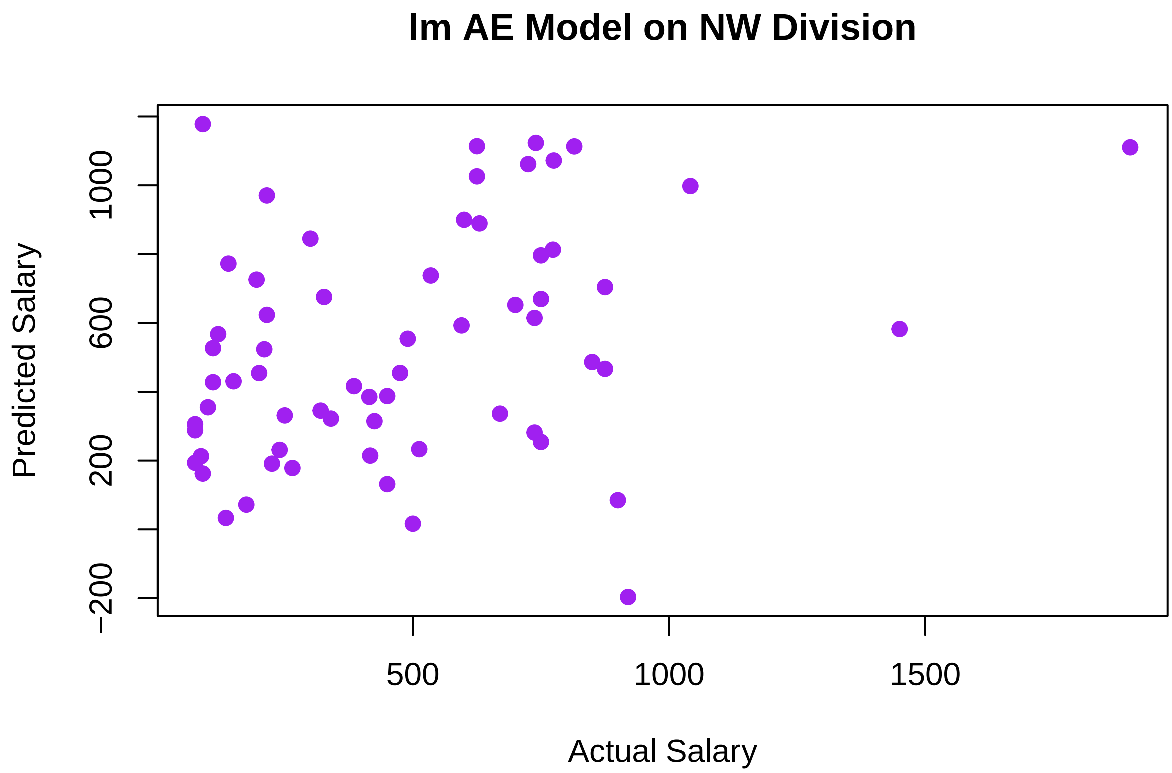
```
# Scatterplot for lm AE model on AW division
plot(hittersAW$Salary, predicted_salaries_lm_AE_AW,
     main = "lm AE Model on AW Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "red", pch = 19)
```




```r
# Scatterplot for lm AE model on NE division
plot(hittersNE$Salary, predicted_salaries_lm_AE_NE,
     main = "lm AE Model on NE Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "green", pch = 19)
```

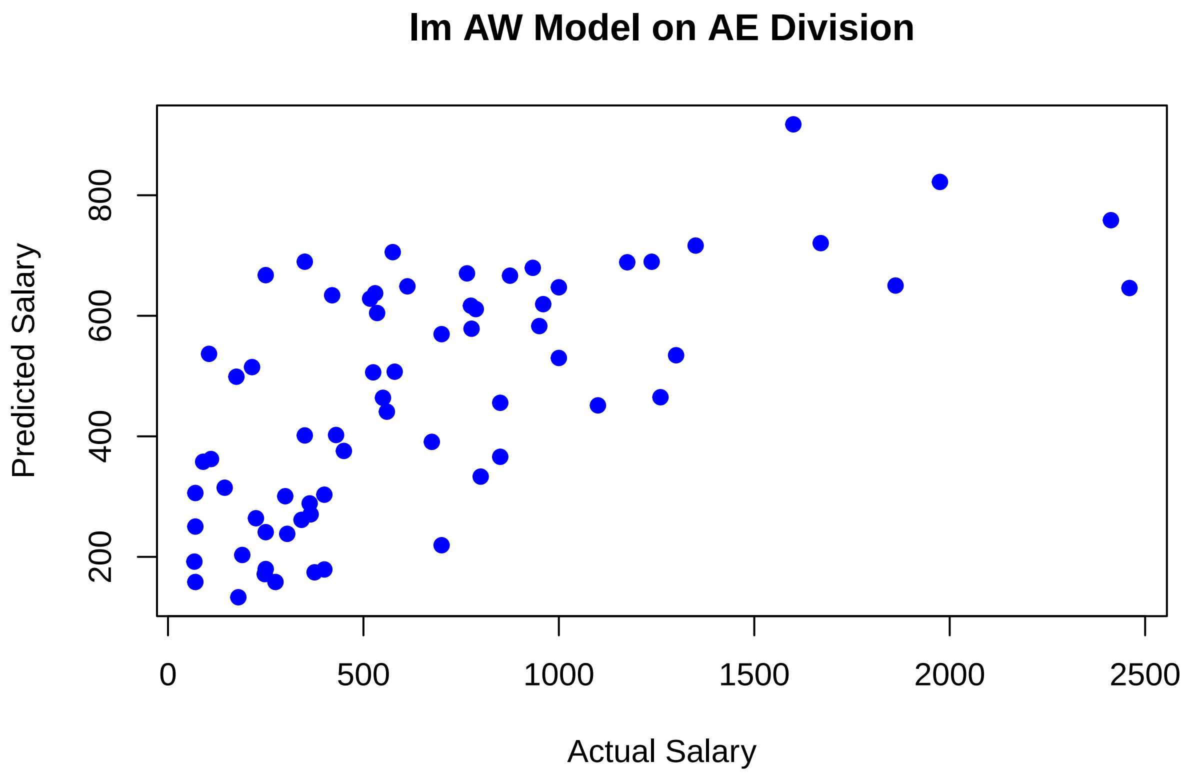



```r
# Scatterplot for lm AE model on NW division
plot(hittersNW$Salary, predicted_salaries_lm_AE_NW,
     main = "lm AE Model on NW Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "purple", pch = 19)
```



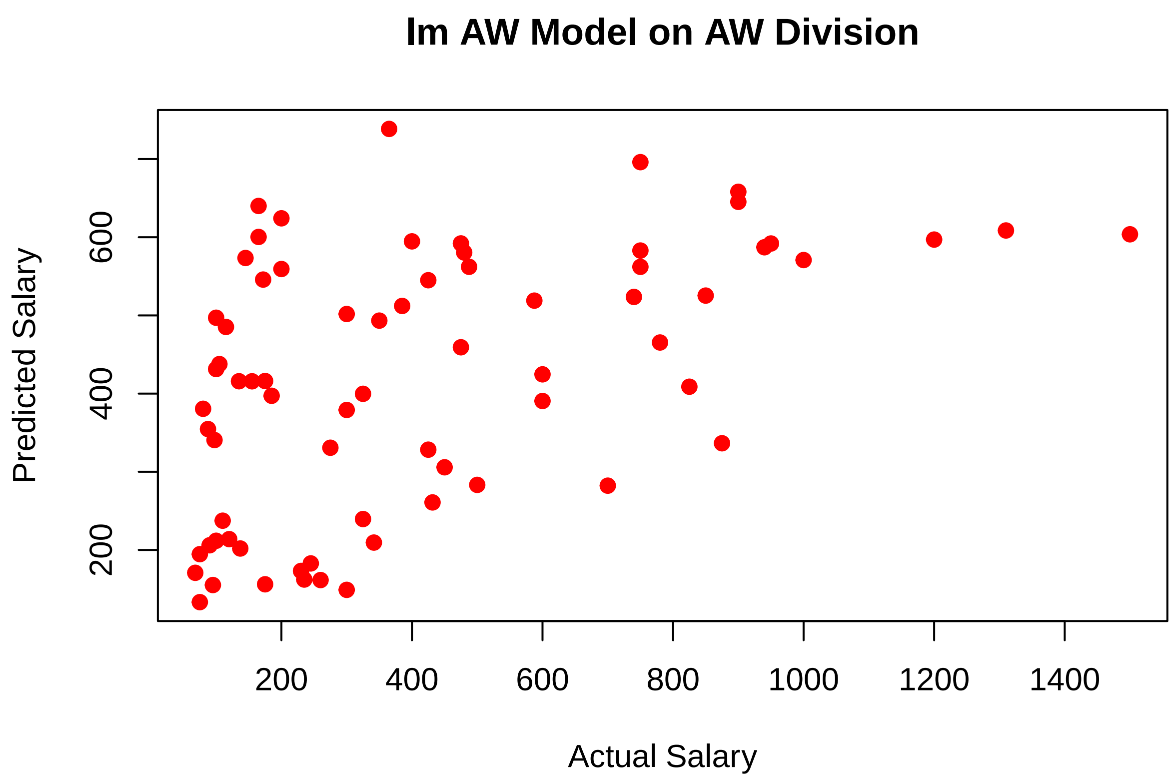
Similar to the lars model, the lm model for the AE divison also only seems to work well with predicting the AE division's salaries and falls apart entirely when trying to predict anything for the other divisions.

```r
# AW lm Model
lm_AW <- lm(Salary ~ Hits + Walks + PutOuts, data = hittersAW)  # Use variables chosen by LARS
predicted_salaries_lm_AW_AE <- predict(lm_AW, hittersAE)  # lm AW model on AE division
predicted_salaries_lm_AW_AW <- predict(lm_AW, hittersAW)  # lm AW model on AW division
predicted_salaries_lm_AW_NE <- predict(lm_AW, hittersNE)  # lm AW model on NE division
predicted_salaries_lm_AW_NW <- predict(lm_AW, hittersNW)  # lm AW model on NW division

# Scatterplot for lm AW model on AE division
plot(hittersAE$Salary, predicted_salaries_lm_AW_AE,
     main = "lm AW Model on AE Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "blue", pch = 19)
```

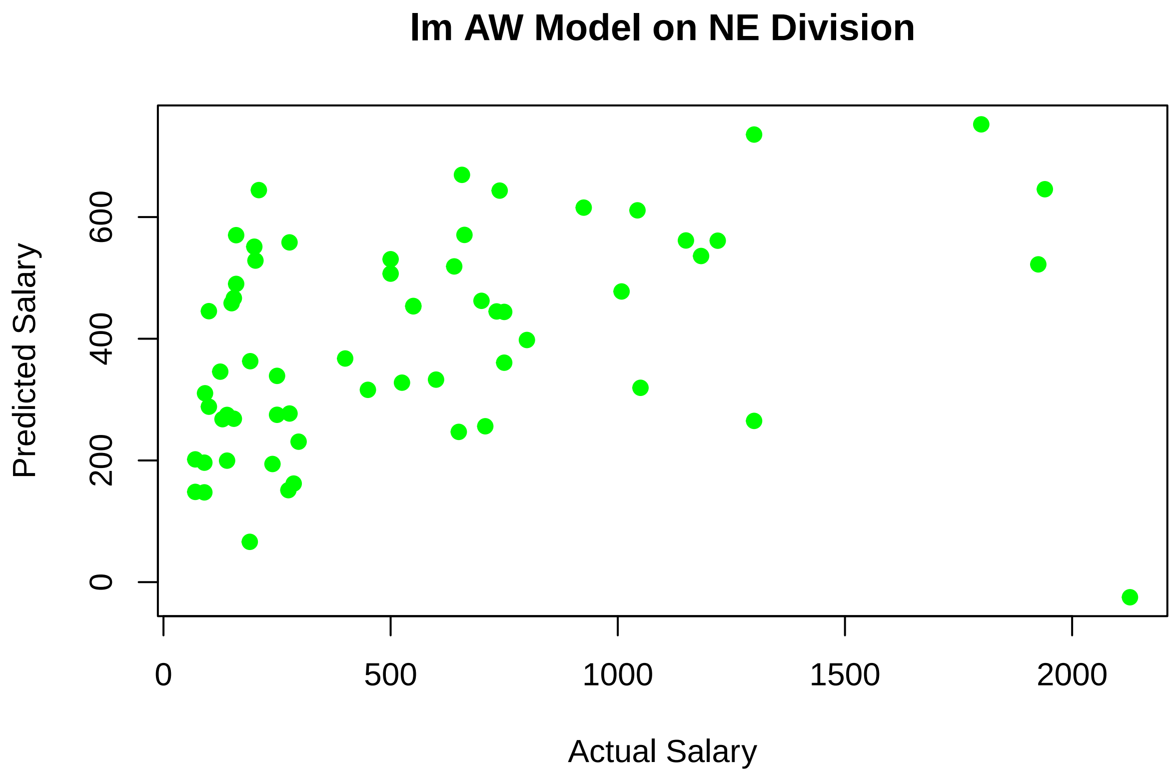


For AW Model:

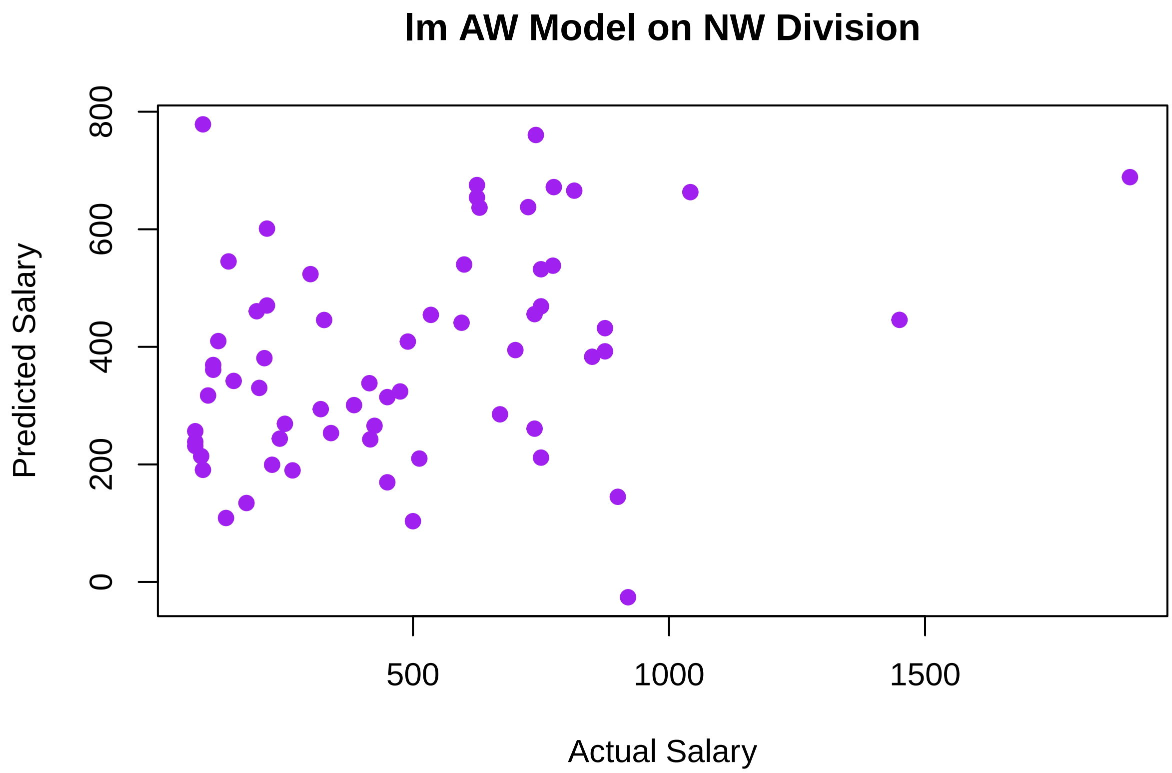
```
# Scatterplot for lm AW model on AW division
plot(hittersAW$Salary, predicted_salaries_lm_AW_AW,
     main = "lm AW Model on AW Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "red", pch = 19)
```




```r
# Scatterplot for lm AW model on NE division
plot(hittersNE$Salary, predicted_salaries_lm_AW_NE,
     main = "lm AW Model on NE Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "green", pch = 19)
```

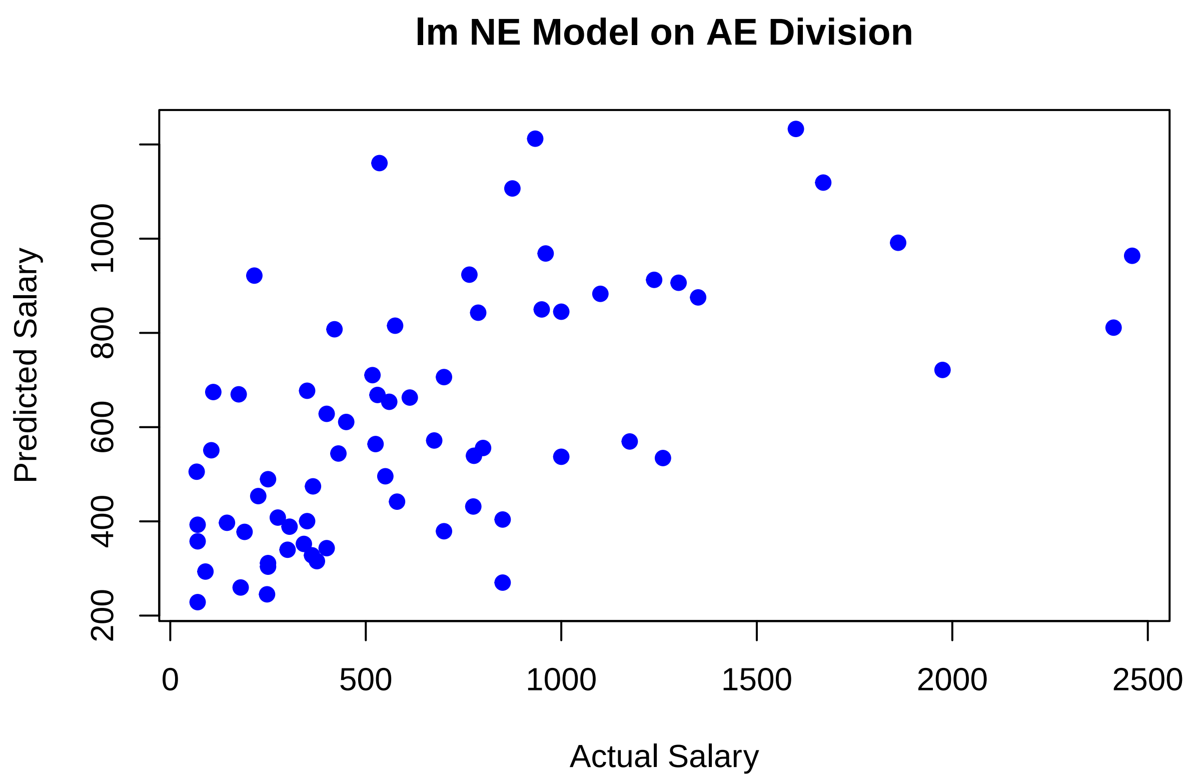



```r
# Scatterplot for lm AW model on NW division
plot(hittersNW$Salary, predicted_salaries_lm_AW_NW,
     main = "lm AW Model on NW Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "purple", pch = 19)
```



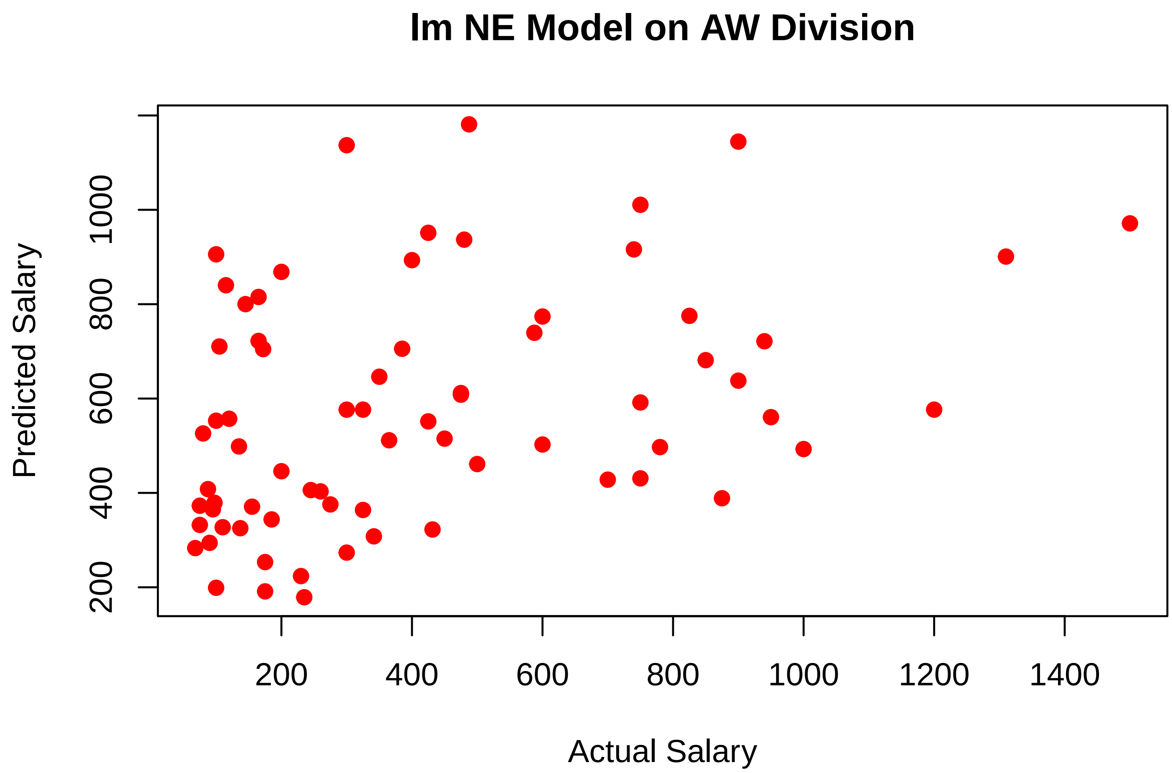
The AW lm model is best when applied to the AW division's salaries. However, the scatterplot is too spread out and has a lot of deviation from the 45 degree line to be a suitable model. Surprisingly, when applied to the AE division, the spread is less than it is for the AW division, but as expected the model still ends up working better for its own division rather than the other divisions. When it comes to the NE and NW divisions it does not work at all.

```
# NE lm Model
lm_NE <- lm(Salary ~ Errors + Walks, data = hittersNE)  # Use variables chosen by LARS
predicted_salaries_lm_NE_AE <- predict(lm_NE, hittersAE)  # lm NE model on AE division
predicted_salaries_lm_NE_AW <- predict(lm_NE, hittersAW)  # lm NE model on AW division
predicted_salaries_lm_NE_NE <- predict(lm_NE, hittersNE)  # lm NE model on NE division
predicted_salaries_lm_NE_NW <- predict(lm_NE, hittersNW)  # lm NE model on NW division

# Scatterplot for lm NE model on AE division
plot(hittersAE$Salary, predicted_salaries_lm_NE_AE,
     main = "lm NE Model on AE Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "blue", pch = 19)
```

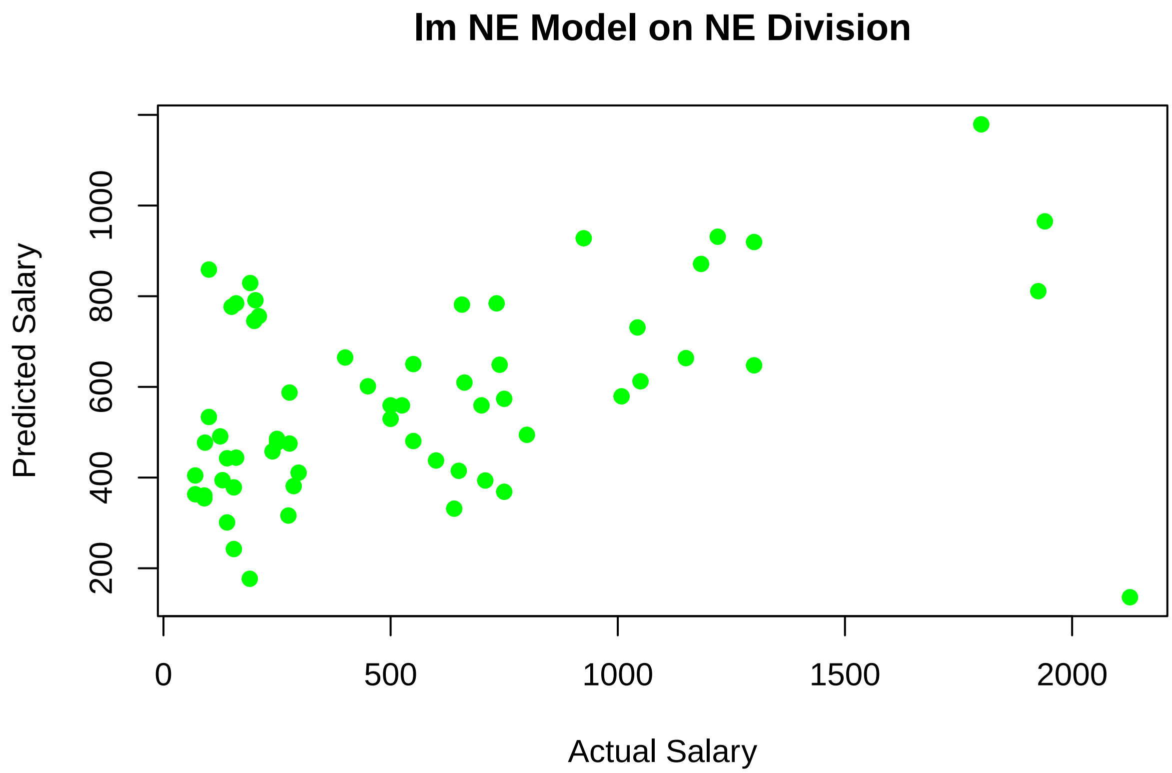


For NE Model:

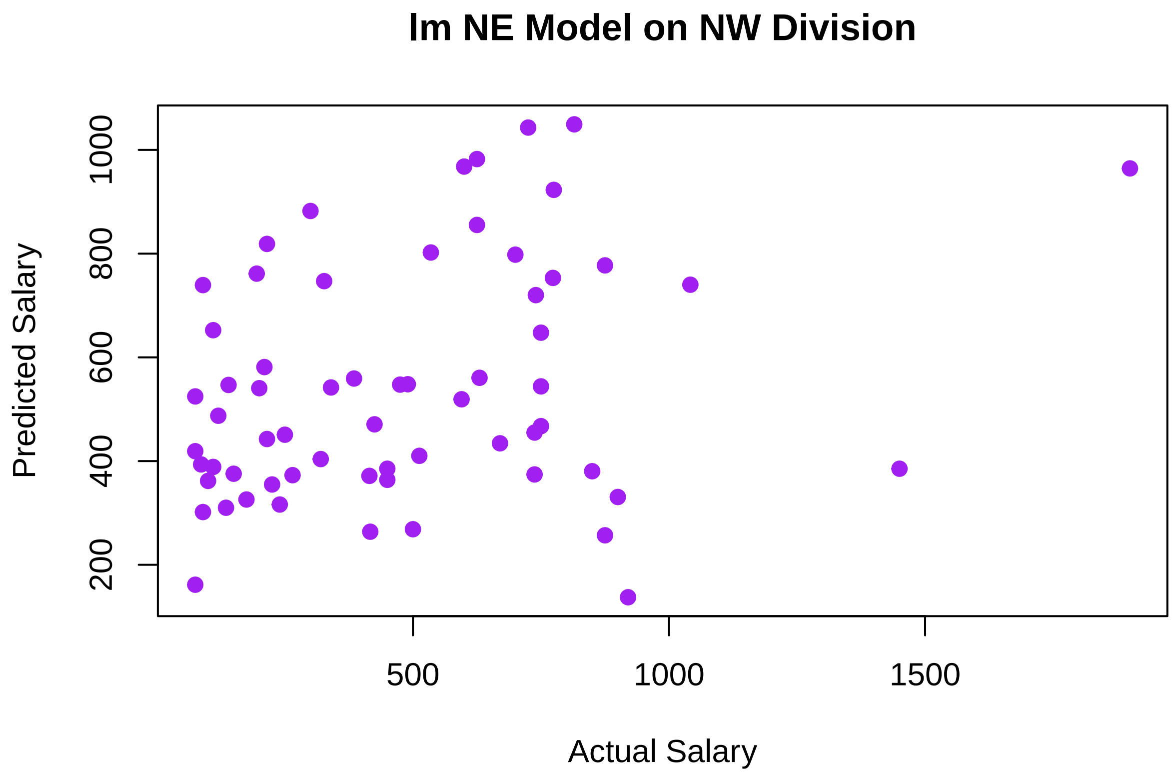
```
# Scatterplot for lm NE model on AW division
plot(hittersAW$Salary, predicted_salaries_lm_NE_AW,
     main = "lm NE Model on AW Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "red", pch = 19)
```




```r
# Scatterplot for lm NE model on NE division
plot(hittersNE$Salary, predicted_salaries_lm_NE_NE,
     main = "lm NE Model on NE Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "green", pch = 19)
```

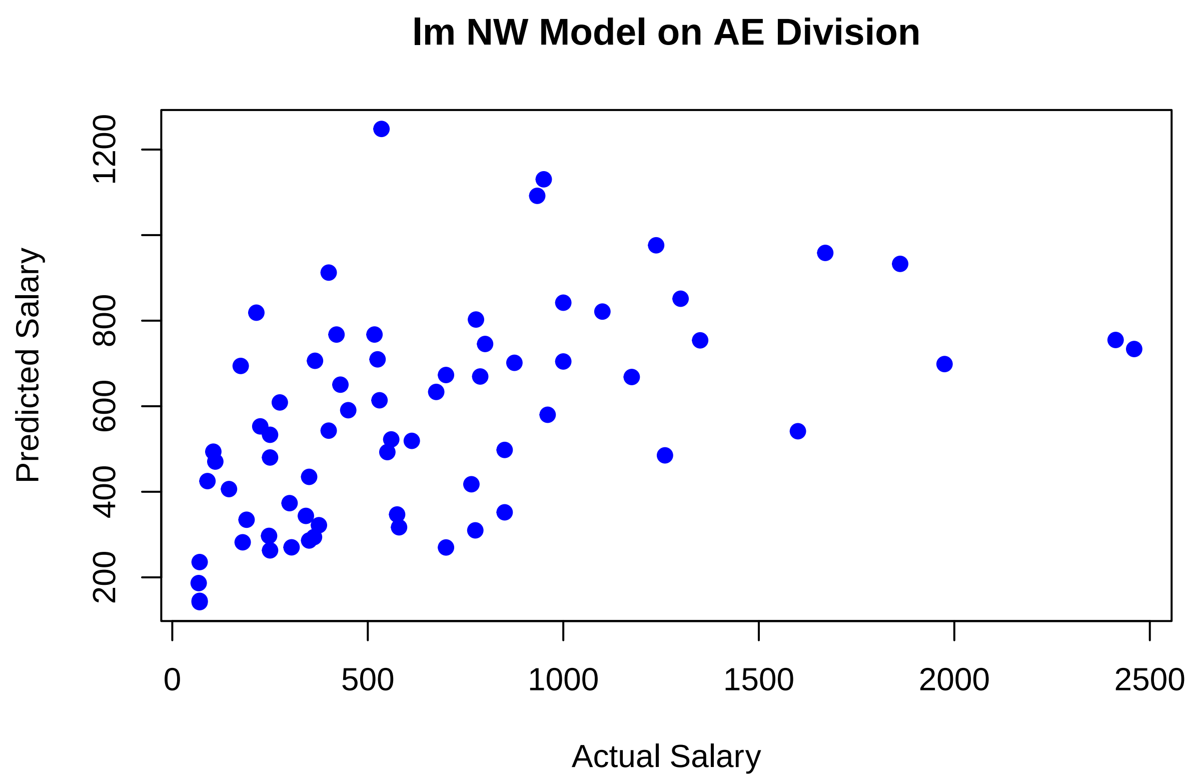



```r
# Scatterplot for lm NE model on NW division
plot(hittersNW$Salary, predicted_salaries_lm_NE_NW,
     main = "lm NE Model on NW Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "purple", pch = 19)
```



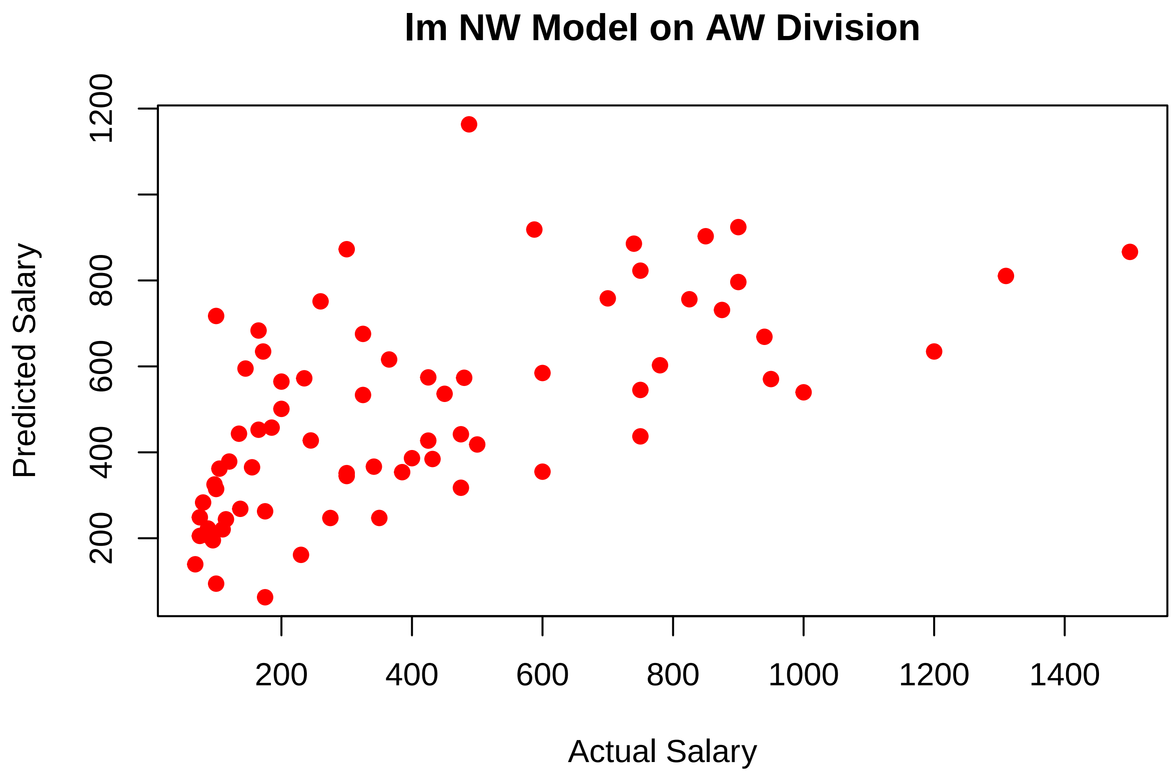
The lm NE model seems to work only when applied to the NE division, and similar to the lars model it is too spread out and in clusters to work as a good prediction model. It does not work at all when applied to other divisions.

```
# NW lm Model
lm_NW <- lm(Salary ~ HmRun + Walks + Years + RBI, data = hittersNW)  # Use variables chosen by LARS
predicted_salaries_lm_NW_AE <- predict(lm_NW, hittersAE)  # lm NW model on AE division
predicted_salaries_lm_NW_AW <- predict(lm_NW, hittersAW)  # lm NW model on AW division
predicted_salaries_lm_NW_NE <- predict(lm_NW, hittersNE)  # lm NW model on NE division
predicted_salaries_lm_NW_NW <- predict(lm_NW, hittersNW)  # lm NW model on NW division

# Scatterplot for lm NW model on AE division
plot(hittersAE$Salary, predicted_salaries_lm_NW_AE,
     main = "lm NW Model on AE Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "blue", pch = 19)
```

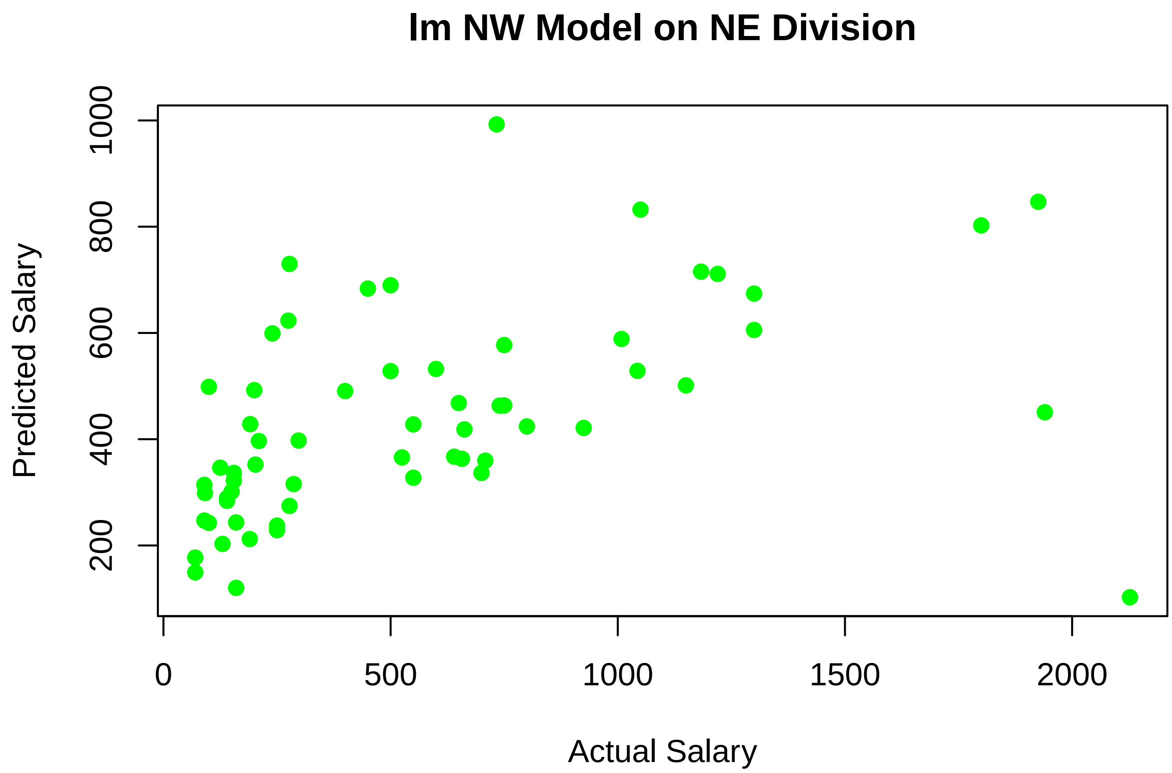


For NW Model:

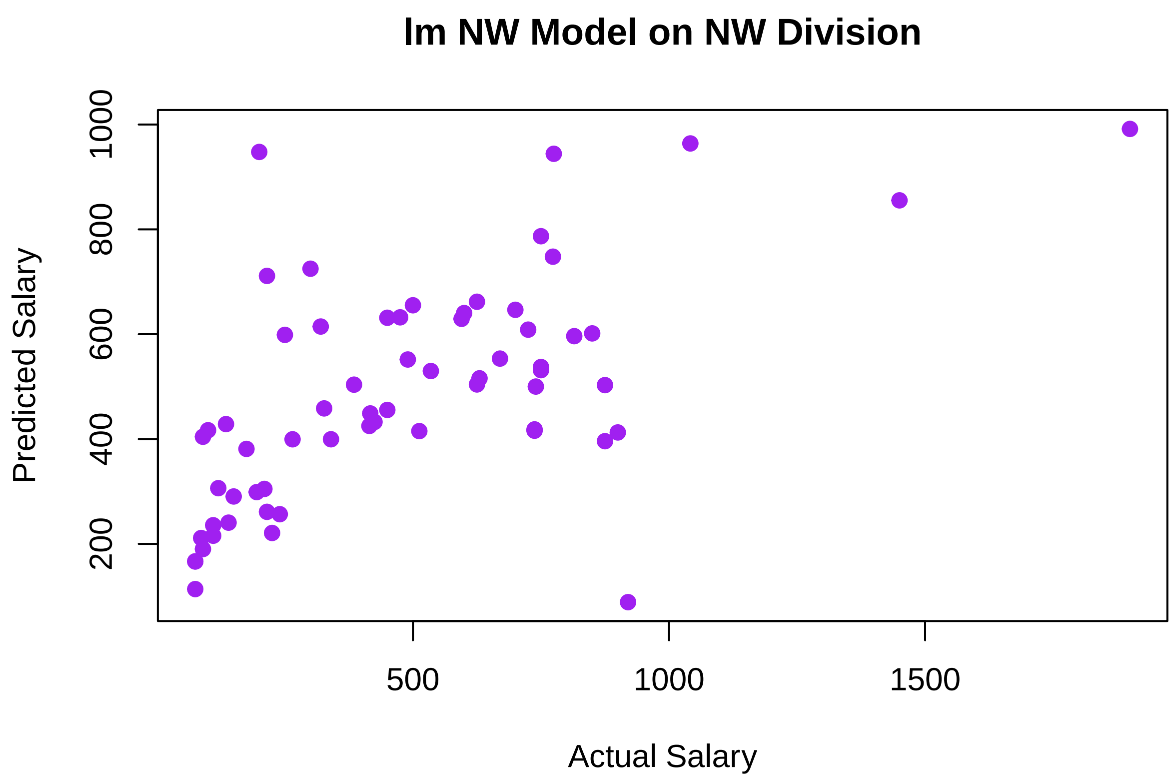
```
# Scatterplot for lm NW model on AW division
plot(hittersAW$Salary, predicted_salaries_lm_NW_AW,
     main = "lm NW Model on AW Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "red", pch = 19)
```




```r
# Scatterplot for lm NW model on NE division
plot(hittersNE$Salary, predicted_salaries_lm_NW_NE,
     main = "lm NW Model on NE Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "green", pch = 19)
```




```r
# Scatterplot for lm NW model on NW division
plot(hittersNW$Salary, predicted_salaries_lm_NW_NW,
     main = "lm NW Model on NW Division",
     xlab = "Actual Salary", ylab = "Predicted Salary", col = "purple", pch = 19)
```



The NW lm model does not function as well as the lars model did for the NW division, it shows only some correlation between the predicted salary and the actual salary. The model is not good enough to be very useful even for the NW division, but it does not work at all when it comes to applying it onto other divisions.

Comparing the Models

It is evident from all the scatterplots that in the case of both the lars models and the lm models, they only seem to work best when applied to their own respective divisions. Such as the AE models working best for the AE division data and so on. Hence, the best way to compare the lars and lm models would be to compare how the models fare for their respective division. Such as comparing the lars model for AE and the lm model for AE when applied to the AE division. The metrics we will be using to the comapare the models against each other will be MSE(Mean Squared Error) and R-squared.

```
# Function to calculate R-squared and MSE
calculate_metrics <- function(actual, predicted) {
  rss <- sum((actual - predicted)^2)  # Residual sum of squares
  tss <- sum((actual - mean(actual))^2)  # Total sum of squares
  r_squared <- 1 - rss/tss  # R-squared
  mse <- mean((actual - predicted)^2)  # Mean squared error
  return(list(r_squared = r_squared, mse = mse))
}

# Compare metrics for AE Division
```

```r
metrics_AE_lars <- calculate_metrics(hittersAE$Salary, predicted_salaries_AE_AE)
metrics_AE_lm <- calculate_metrics(hittersAE$Salary, predicted_salaries_lm_AE_AE)

# Compare metrics for AW Division
metrics_AW_lars <- calculate_metrics(hittersAW$Salary, predicted_salaries_AW_AW)
metrics_AW_lm <- calculate_metrics(hittersAW$Salary, predicted_salaries_lm_AW_AW)

# Compare metrics for NE Division
metrics_NE_lars <- calculate_metrics(hittersNE$Salary, predicted_salaries_NE_NE)
metrics_NE_lm <- calculate_metrics(hittersNE$Salary, predicted_salaries_lm_NE_NE)

# Compare metrics for NW Division
metrics_NW_lars <- calculate_metrics(hittersNW$Salary, predicted_salaries_NW_NW)
metrics_NW_lm <- calculate_metrics(hittersNW$Salary, predicted_salaries_lm_NW_NW)

# Print metrics
cat("AE Division: \n")
```

Creating a function to calculate the performance metrics and implementing it

```
## AE Division:
```

```r
cat("LARS R-squared:", metrics_AE_lars$r_squared, "MSE:", metrics_AE_lars$mse, "\n")
```

```
## LARS R-squared: 0.7894057 MSE: 61291.04
```

```r
cat("LM R-squared:", metrics_AE_lm$r_squared, "MSE:", metrics_AE_lm$mse, "\n\n")
```

```
## LM R-squared: 0.4539415 MSE: 158924
```

```r
cat("AW Division: \n")
```

```
## AW Division:
```

```r
cat("LARS R-squared:", metrics_AW_lars$r_squared, "MSE:", metrics_AW_lars$mse, "\n")
```

```
## LARS R-squared: 0.6141142 MSE: 42449.34
```

```r
cat("LM R-squared:", metrics_AW_lm$r_squared, "MSE:", metrics_AW_lm$mse, "\n\n")
```

```
## LM R-squared: 0.2491549 MSE: 82596.66
```

```r
cat("NE Division: \n")
```

```
## NE Division:
```

```r
cat("LARS R-squared:", metrics_NE_lars$r_squared, "MSE:", metrics_NE_lars$mse, "\n")
```

```
## LARS R-squared: 0.4653834 MSE: 136593.4
```

```r
cat("LM R-squared:", metrics_NE_lm$r_squared, "MSE:", metrics_NE_lm$mse, "\n\n")
```

```
## LM R-squared: 0.174233 MSE: 210981.8
```

```r
cat("NW Division: \n")
```

```
## NW Division:
```

```r
cat("LARS R-squared:", metrics_NW_lars$r_squared, "MSE:", metrics_NW_lars$mse, "\n")
```

```
## LARS R-squared: 0.5846886 MSE: 50676.58
```

```r
cat("LM R-squared:", metrics_NW_lm$r_squared, "MSE:", metrics_NW_lm$mse, "\n")
```

```
## LM R-squared: 0.3667765 MSE: 77266.36
```

Division	Lars R-squared	Lars MSE	LM R-squared	LM MSE
AE	0.7894057	61291.04	0.4539415	158924
AW	0.6141142	42449.34	0.2491549	82596.66
NE	0.4653834	136593.4	0.174233	210981.8
NW	0.5846886	50676.58	0.3667765	77266.36

This tells us that the lars models are consistently outperforming the lm models for all the divisions in both having a higher R-squared value and a lower MSE. Based on performance, lars is able to more meaningfully capture the relationships between the predictors and the salary across all the divisions.

Conclusions

Predicting between and within divisions: It is evident from all the data analysis that all the models, lars and lm work best when predicting within the divisions. None of the models of any divisions could work reliably when applied to any other division. Predicting between divisions does not yield useful results. As we saw from the bar plots earlier, the divisions all seem to have different significance and variety of predictors when it come to the salaries in their divisions.

Salary Distrubution: It makes sense that since the divisions have different predictors for the salary, their salary distributions would also differ. As we saw from the box plot earlier, the median salaries seem to be around the same level for all the divisions, with the it being slightly higher for the AE division with a larger spread. AW seems to a slightly lower median and less spread as well, but overall the median salaries seem to be around the same level.

How player performance influences salary differently across the divisions: Based on all of the models' performance metrics, it is evident the AE division seems to have the most correlation between predictors and actual salary. This means that player performance has the most impact on salary within the AE division as compared to the other divisions, this is evident since the lars model had an R-squared value of 0.79, which is much higher as compared to the other divisions. Conversely, the NE divisions has the lowest correlation between the predictors and the actual salary, which would explain why neither the lars model ot the lm model could work well enough. The AW and NW divisions also have a significantly lower R-squared value than the AE division, but it is still higher than the NE division. This effectively means that the AE division's salaries are correlated to the player performance most, followed by the AW division and the NW division, and the NE division has the least correlation between player performance and salary.