The statistical learning course has provided us with new, powerful methods of classification that have expanded our knowledge of the concept. As we discovered these methods in R, the team decided to explore them further by choosing the classification assignment for our project. When surveying the data available for classification, we reviewed each dataset provided by performing a short EDA and researching the data's purpose when collected. This process proved the Absenteeism at Work (Absenteeism) data set to have the most model potential and relevancy to a real world problem. Therefore, our project aims to use algorithms we learned during this program to properly classify the Absenteeism dataset classes with the lowest possible error metric.

Our exploratory data analysis (EDA) initially inspected the data in four parts: checking for missing data, exploring each variable for distribution and sparsity, ensuring there were no zero-variance or correlated variables, and reviewing our predictor variable. To check data quality, we tested for missing data; this resulted in no missing values and thus no transformations or data cleaning was needed at this stage. Then we reviewed each variable in two steps, both with the objective to view the distribution and sparsity of the variable. First we built a table to view the number of responses for each unique observation in the variable, then we created a histogram to view the distribution of those observations in a graph. Though some of these variables had left skews, others right, others normal, the goal was to discover irregular data or sparsity. The two variables that stood out for sparsity were the Disciplinary.failure and Social.smoker variables (Appendix 1); since these variables were binary with a majority of zero responses, we chose to remove these two from the dataset for modeling moving forward.

The next step of the EDA was to review the variables for zero variance and correlation. There were no variables with near zero variance that would interfere with our models. Then we used a corrplot to test for correlation between variables that might cause model issues due to multicollinearity. This plot (Appendix 2) demonstrated a high correlation between age and service time. To refrain from potential discrimination issues, we excluded age from the analysis and kept service time. We also saw a correlation between BMI, weight, and height. This makes intuitive sense, as BMI is a calculation based on weight and height. Since BMI effectively represents the impact of both weight and height, both weight and height will be excluded from the analysis and BMI retained.

We also took a further look into some relationships between a few of the Absenteeism variables. First we created a scatterplot of Absenteeism.time.in.hours and the factorized version of Reason.for.absence (Appendix 3); no discernible patterns appeared to explain why someone was absence in relationship to when they were absent. For absences less than a week in duration, we created a bar chart see if there was a relationship between the number of hours absent and the distance between home and work (Appendix 4). Again, there were no discernible patterns between how far an employee lived from work and how long they were absent. We used a box plot to test for relationships between worker productivity and number of hours absent (Appendix 5) and there were very few outliers, however they should not affect the classification modeling as they were not significant outliers. A scatterplot comparing time off and amount of Service.time did not provide any data patterns (Appendix 6), indicating no relationship between how long an employee has worked with the company and their absent time. In the bar chart comparing the Month.of.absence to Absenteeism.time.in hours and colored by the factorized Day.of.the.week

(Appendix 7), we found there were more absences in March and July as well as more time taken off on Mondays and Tuesdays than another days of the week. Though these patterns were interesting, we concluded that they would not affect our model.

Before moving on to preprocessing, we took a further look into our predictor variable, absenteeism.time.in.hours. We found in the tables and histograms part of the EDA earlier that there were a significant amount of zero values in the variable (Appendix 8), which seemed odd to us. We figured that if the time someone was absent was zero, they didn't take time off, but wanted to dig into this trend a bit deeper before leaving the data in the model. Therefore, we tested to see if there was a relationship between the reason for absence being zero and the time absent being zero and found that all the zero value reasons for absence correlated with the time taken equal to zero. We decided that since it was a valid subset of the data, the employees who did not take time off will be useful in the classification modeling, we left those zero values in the data set.

Based on the EDA, there was minimal pre-processing needed to prepare the dataset for modeling. Per the correlation and sparsity analysis, we removed the Age, Weight, Height, Disciplinary.failure, and Social.smoker variables from the data set to form the absentee.sparse dataset in the interim. Then we looked back at the Absenteeism.time.in.hours variable and realized it needed to be a categorical variable for our classification models. From our EDA, we found that the variable ranged from 0 to 120 hours, with the majority of the data in the lower number of hours, about 8-10 hours or less. Since we wanted the number of days an employee was absent in our variable, we made two levels: any values of Absenteeism.time.in.hours that were less than 16 (the total absentee time equal to less than two working days) one factor, and

any values equal to or greater than 16. The less than 16 hours factor is the majority class (as "D" in the models), and the minority class is the greater than or equal to 16 hours (as "W" in the models). At this point, the data is ready to use initial classification models using the processed absentee.work dataset.

Before running our initial models, we chose three metrics to determine the success of each initial, unoptimized classification model. Accuracy determines the proportion of all predictions the model makes correctly, or the proportion of true positives, in the dataset overall. Sensitivity is the proportion of true positives that are correctly identified by the model. Specificity is the proportion of true negatives correctly identified by the model. Since we are trying to identify the number of employees who will be absent for more than 16 hours, we decided to put a premium on specificity performance.

Our goal was classification, and thus we chose the following five models that each aim to classify data: classification and regression trees (decision trees) that use partitions to separate the classes in the data; random forests that use repetitive trees that are able to avoid variable dominance that can appear in decision trees; linear discriminant analysis (LDA) that classifies data with a linear line into distributions within the data; quadratic discriminant analysis (QDA) which takes into account non-equal variances that LDA does not; and finally support vector machines (SVM) that create linear planes to separate data into classes (note that we used both C and Nu parameters for SVM thus making six total models). In each of these models, we made a 60% training / 40% test split in the data to train and test the models, respectively. Each model was created in a loop that ran for 50 replications. The random forests were run with 100 trees, and for SVM-C we used a c-value parameter of 0.25, and SVM-Nu the nu-value parameter was

0.15. For each repetition we created a (yhat) prediction for each model, then calculated three matrices (for each metric) each with a column for each model and a row for each repetition. These matrices were relabelled and transformed into a box plot for ease of analysis.

Observing the accuracy metric box plots, the SVM-C model has the greatest overall accuracy at around 0.913. The random forests and LDA had similar mean accuracies but had larger quantiles and an outlier each, while the SVM-Nu had an average performance. The sensitivity box plot proved that SVM-C is again the superior model, classifying all of the true positives for a 100% sensitivity accuracy. Again, the random forests and LDA came close, but still had larger quantiles, while the SVM-Nu had an average performance . The final metric specificity, which shows the proportion of true negatives correctly identified in the model found SVM-Nu to be the winner (note QDA slightly won out against SVM-Nu but it did very poorly in accuracy and sensitivity and thus was not named winner here). SVM-C was unable to correctly classify any member of the minor class. Let's review what's going on here.

LDA finds the optimal position to place a line between the distributions by modeling both 'sides' of the data with the same covariances. This line will be placed to minimize the error. SVM places a line between the distributions as well, but then it creates additional boundaries around that line. The boundaries, or margins, are mirrored on each side of the line. The margins are maximized to show the largest separation between the distributions while still allowing for misclassified points in the margins. By creating these additional margins anchored by support vector points on the line, the points that may have been misclassified with a singular linear line in LDA are contained within the margins. SVM accounts for any misclassifications in that margin, so the distributions on each side of the margins have a very low error rate.

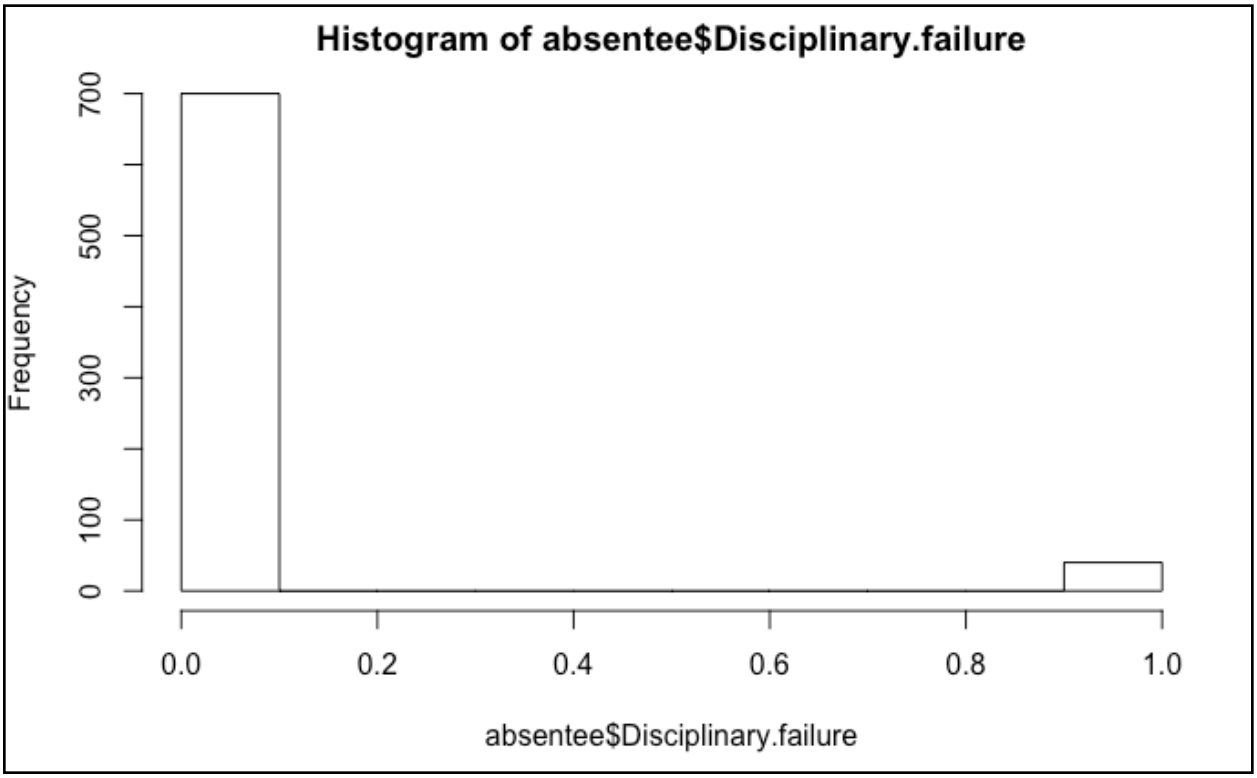
The difference between SVM-C and -nu comes down to how the C and nu values create the margins. Both values act as constraints on those margins and make a determination of which points will be the support vectors, but the SVM-nu has the advantage of using a parameter nu between 0 and 1 to control the number of support vectors. The SVM-C model does not control the number of support vectors and can range between 0 and infinity. Therefore, nu provides upper and lower bounds on the number of support vectors that lie on the wrong side of the hyperplane line in SVM implementation that SVM-C does not contain. Though SVM-C does well predicting overall accuracy in our initial model, it does not do as well predicting the minor class; essentially its accuracy is so good because it is predicting all observations to be in the major class. Based on our classification goals and comparatively great specificity performance, we chose the SVM-nu model to further optimize.

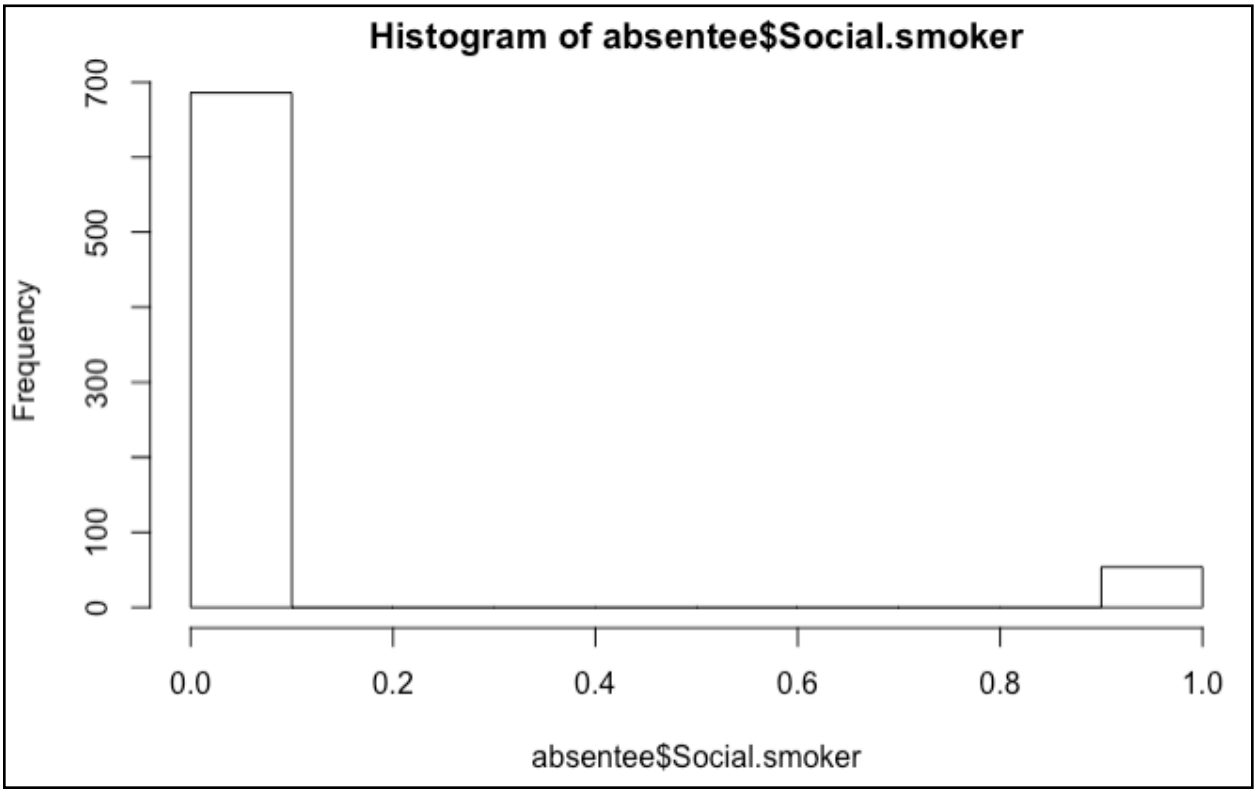
The Nu optimization parameter is trying to give a sense of how many "danger zone" observations exist on the margin of the dividing boundary. By optimizing that parameter within 50 different replication training/test data models, we gave ourselves the best chance of testing the model's predictive ability. Our optimal Nu value varied between 0.01 and 0.1. We utilized 30 different Nu's in every replication within that range. We could not use a range above 0.1 without experiencing model errors ("feasible" Nu values). We used a 60/40 training/test set split like in the initial model loop and a stratified group sampling method. To optimize the Nu value specifically, we also used a stratified training/test data split of 60/40 within the larger Training data set and ran SVM to find that optimal value. We used 10-fold Cross Validation to further optimize our models. We did also run SVM with the C regularization parameter and optimized it to see whether we could improve its ability to predict the minor class. In our runs, utilizing the

same general optimization techniques, we did not see a difference from base to optimized models for the C parameter SVM models (Appendices 12-14).

For our optimized SVM-Nu model, our overall accuracy for our absenteeism prediction decreased slightly from the base model to the optimized model (Appendix 15). We dropped about 2% on average and have an 'optimized' accuracy between 85% and 90% at the worst/best. On average, we saw a decrease of 7% in sensitivity for our optimized model (Appendix 16); we were a little bit worse in predicting employees would be absent for less than 16 hours. However, this is not our major focus. The specificity got better with optimization by roughly 5% (Appendix 17). This indicates the "true negatives" (employees who were absent for 16 hours or more) were predicted much more accurately. Optimizing the Nu value to draw the nonlinear hyperplane boundary and separate our observations helps classify absenteeism time more appropriately. In general, we also know that SVM is robust to large outliers, so we will have a better ability to predict employees who are absent for a higher amount of hours. The Confusion Matrices from base to optimized models show a lower accuracy rate overall (Appendix 18). However, we are more accurate where it counts: predicting the employees who are absent for more than 16 hours.

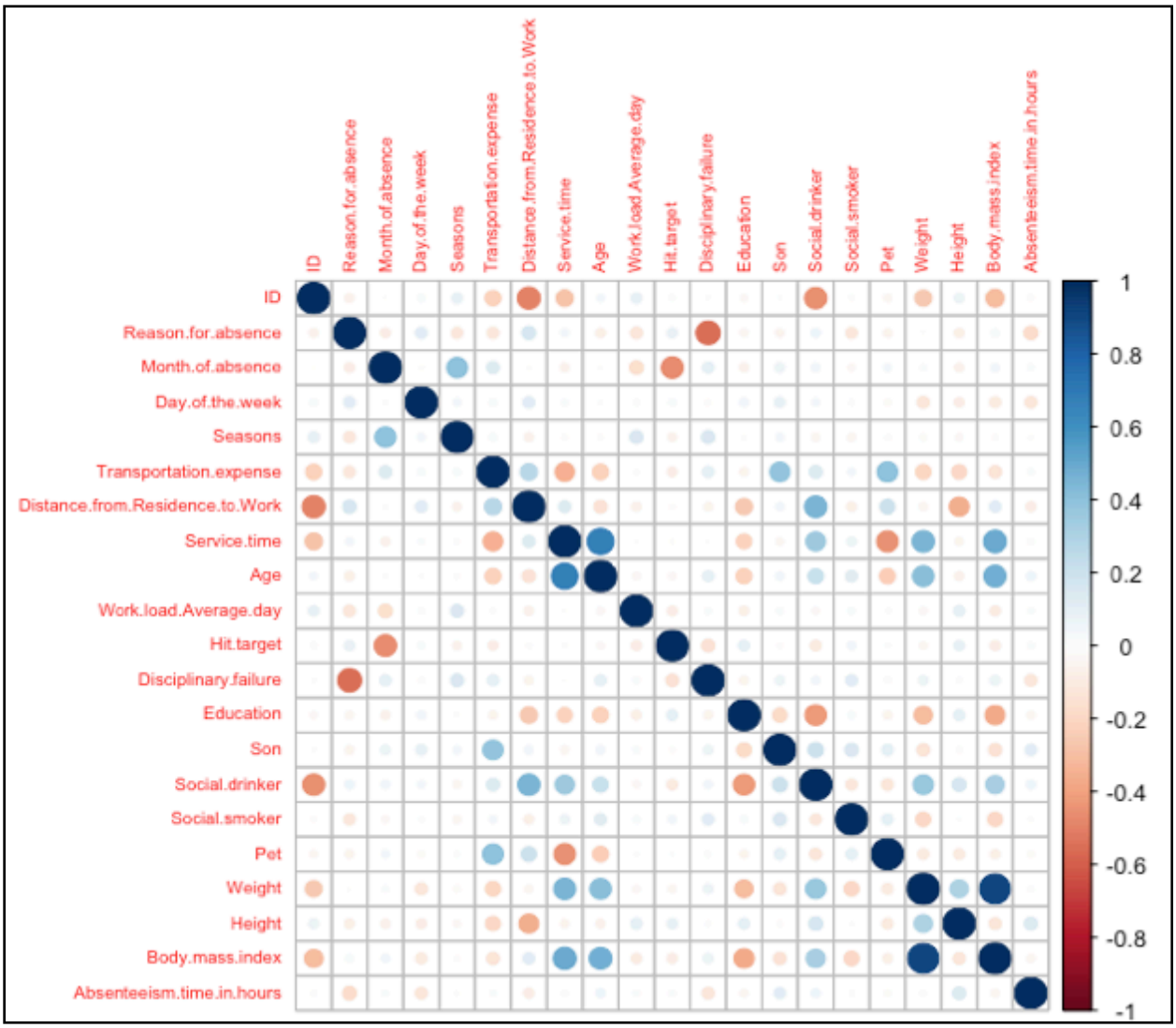
APPENDIX 1: Sparsity of the Disciplinary.failure & Social.smoker Variables





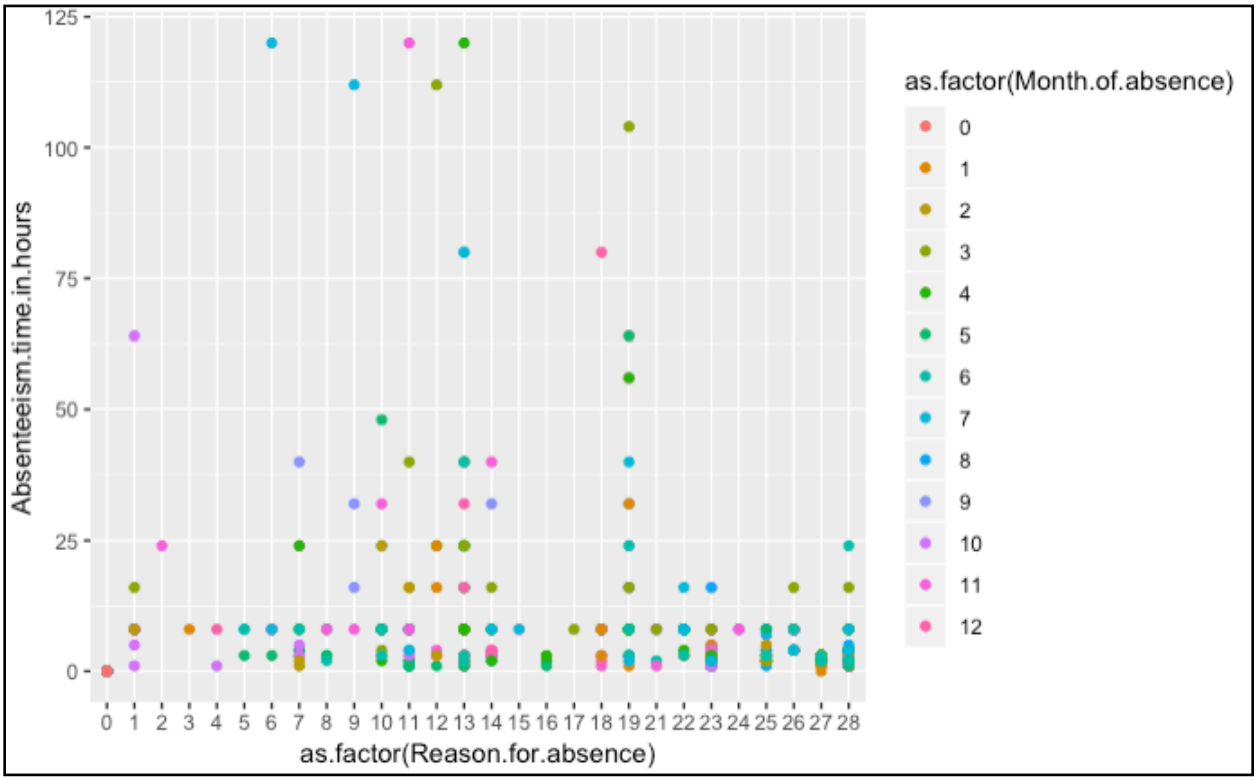
APPENDIX 2: Corrplot of Absenteeism Dataset Showing Variable Correlations

Note: The darker the circle, the more correlation; each variable is highly correlated to itself.

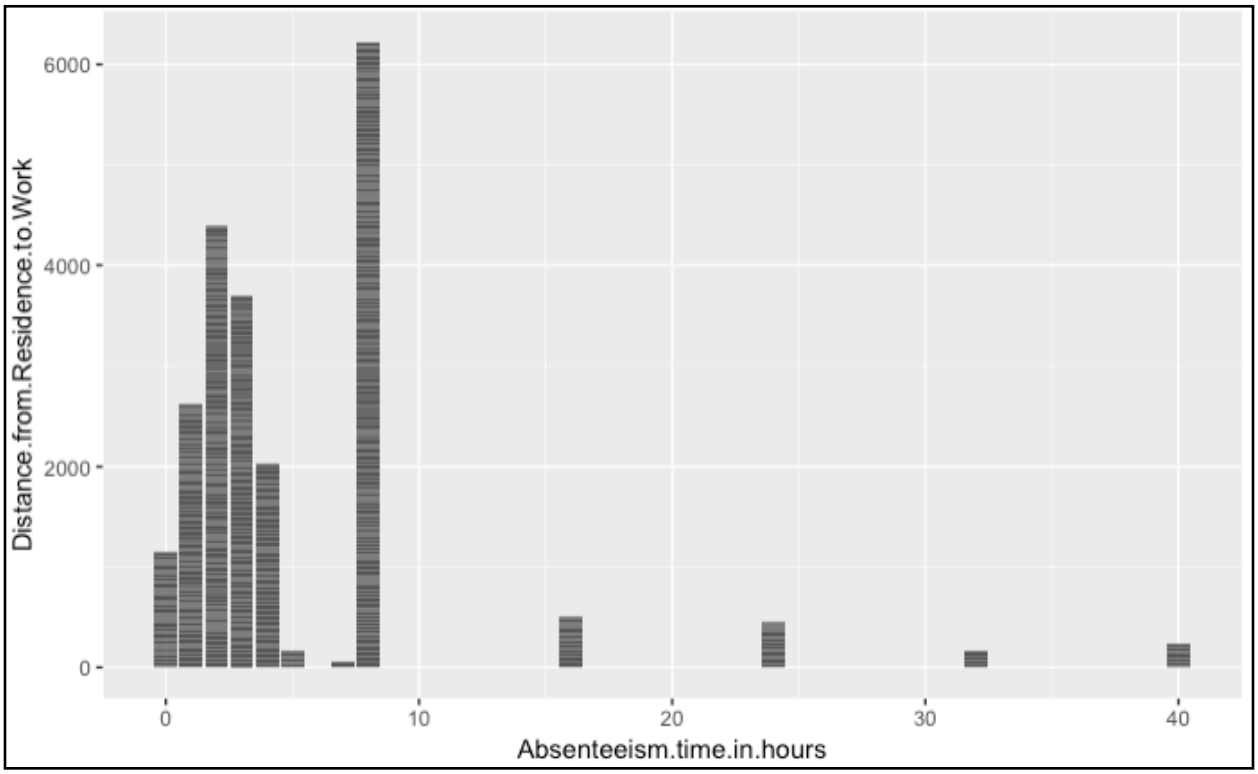


APPENDIX 3: Scatterplot of Absenteeism.time.in.hours & Reason.for.absence

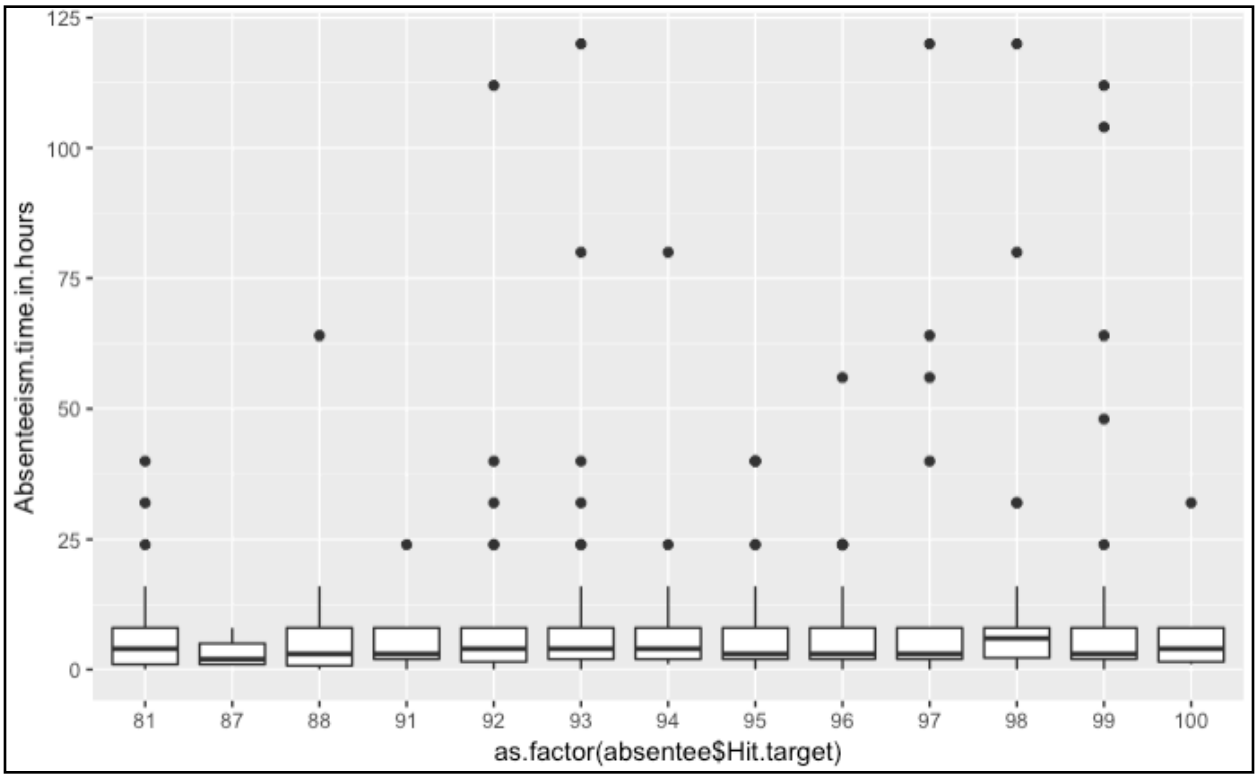
(Colored by Month.of.Absence)



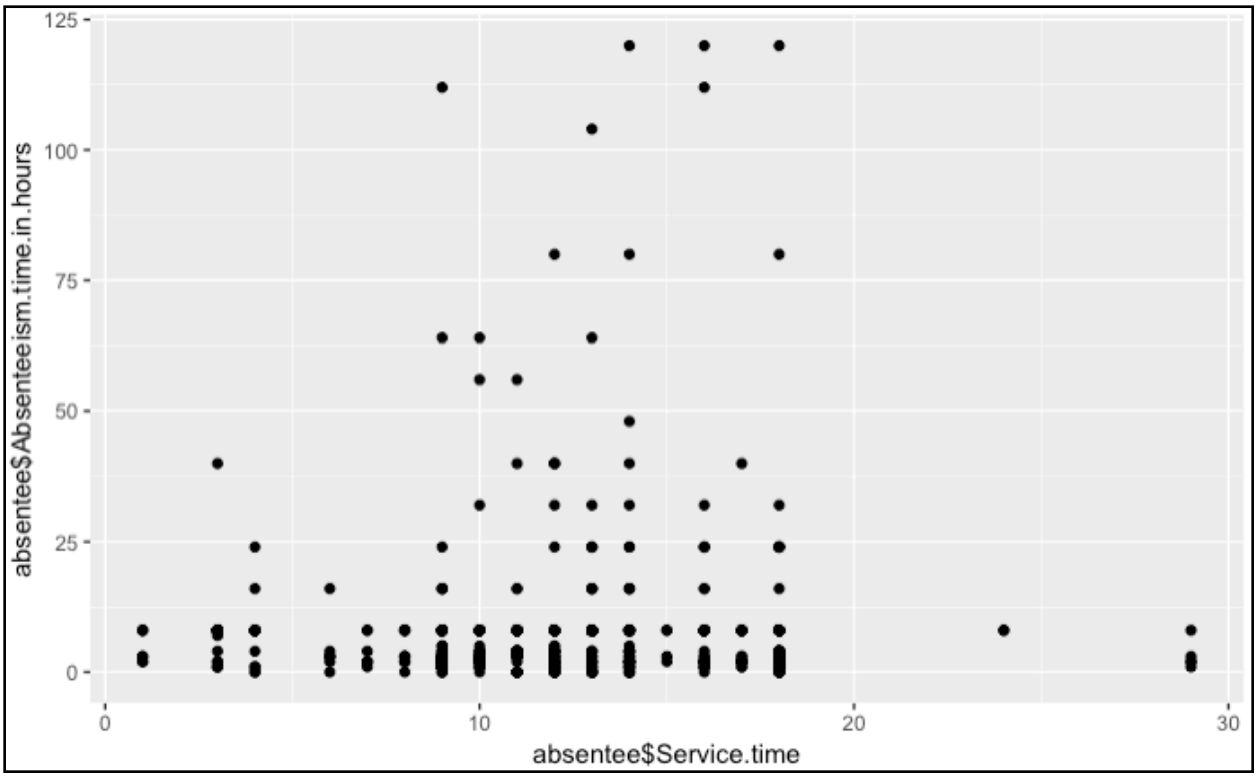
APPENDIX 4: Histogram of Absenteeism.time.in.hours & Distance.from.residence.to.work



APPENDIX 5: Box Plot of Absenteeism.time.in.hours & Hit.target

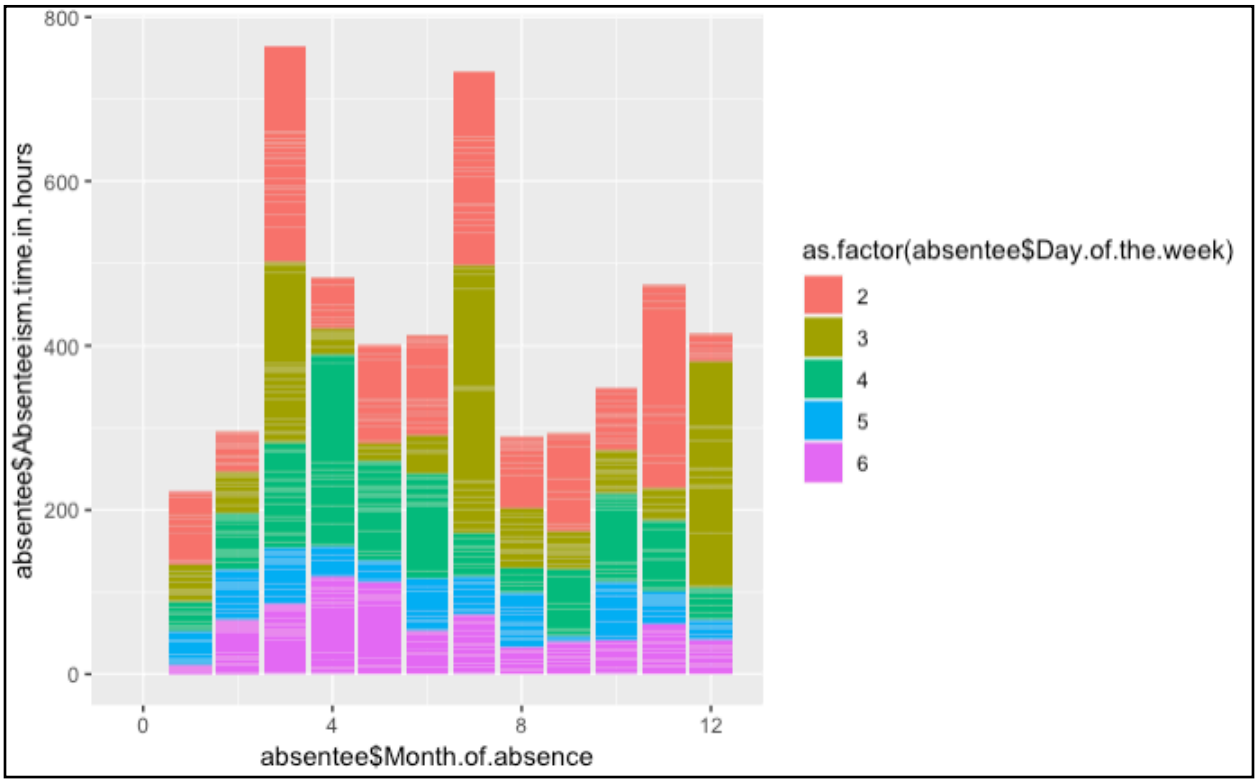


APPENDIX 6: Scatterplot of Absenteeism.time.in.hours & Service.time

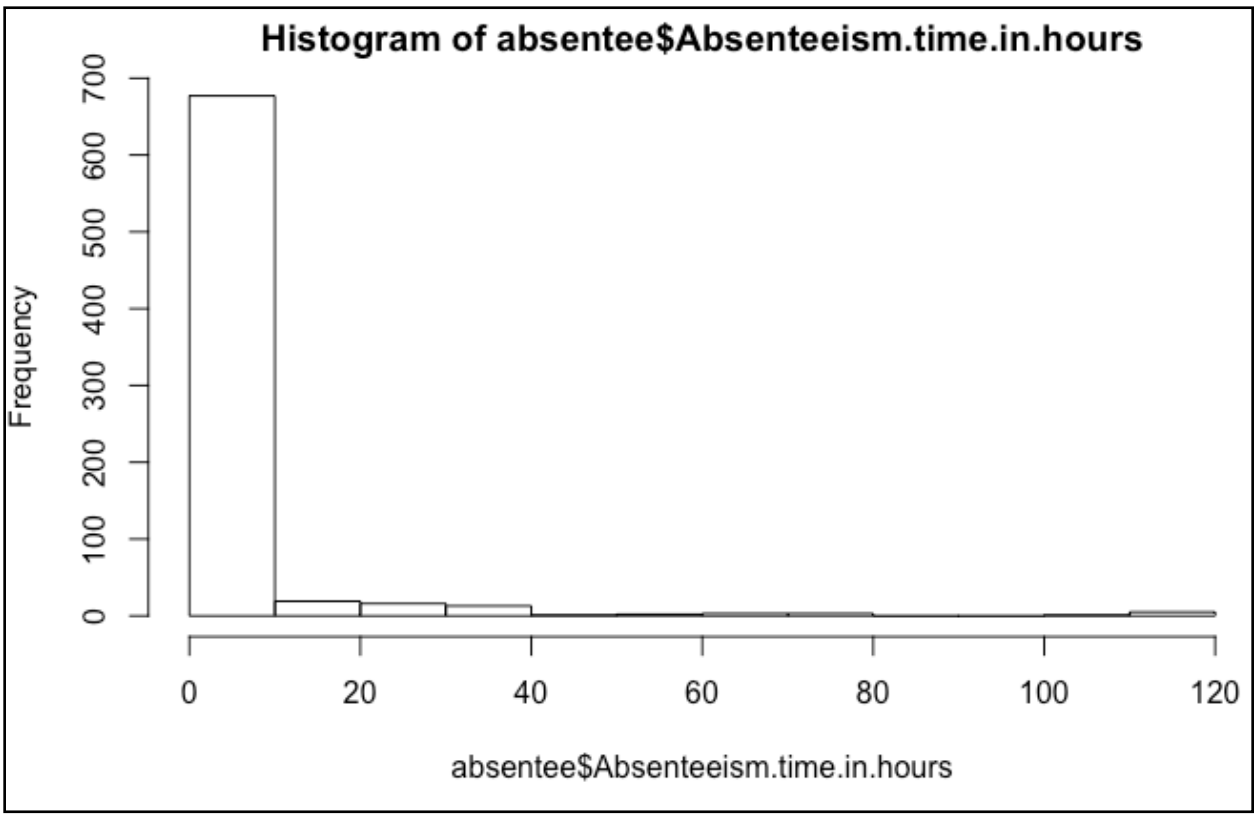


APPENDIX 7: Absenteeism.time.in.hours & Month.of.absence

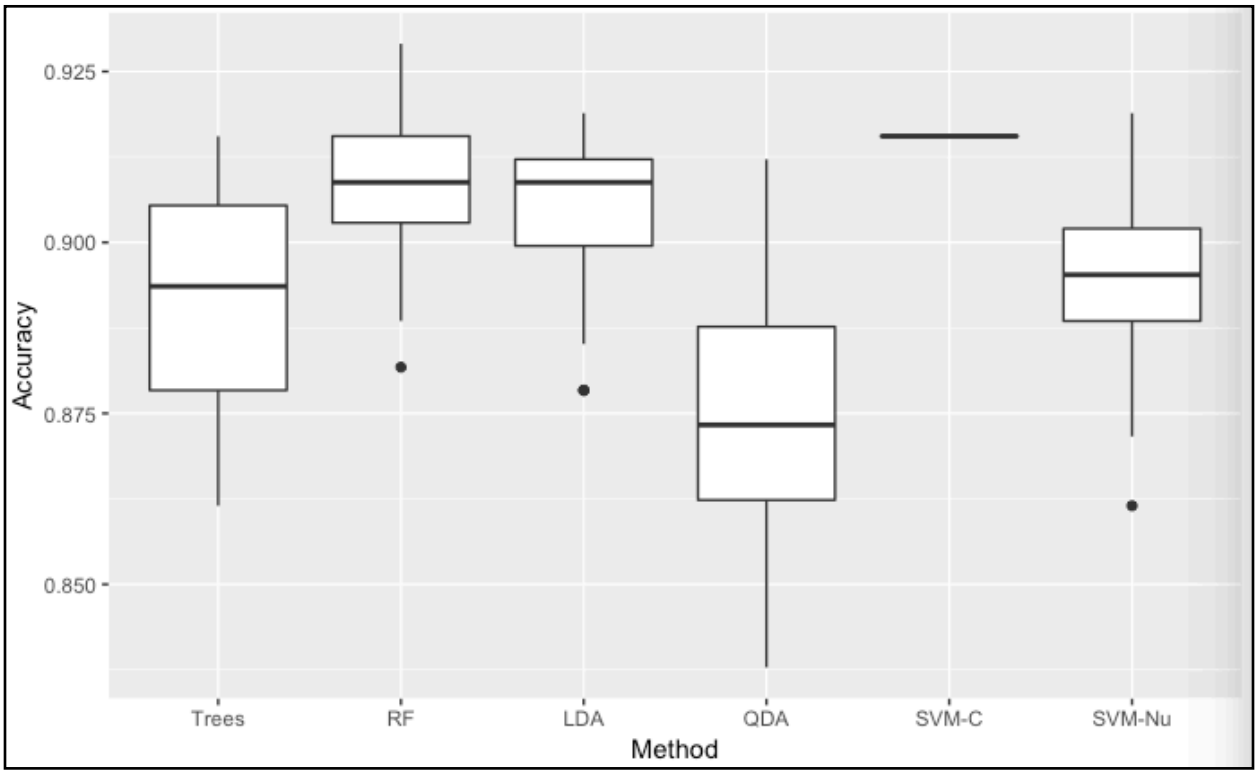
(Colored by Day.of.the.Week)



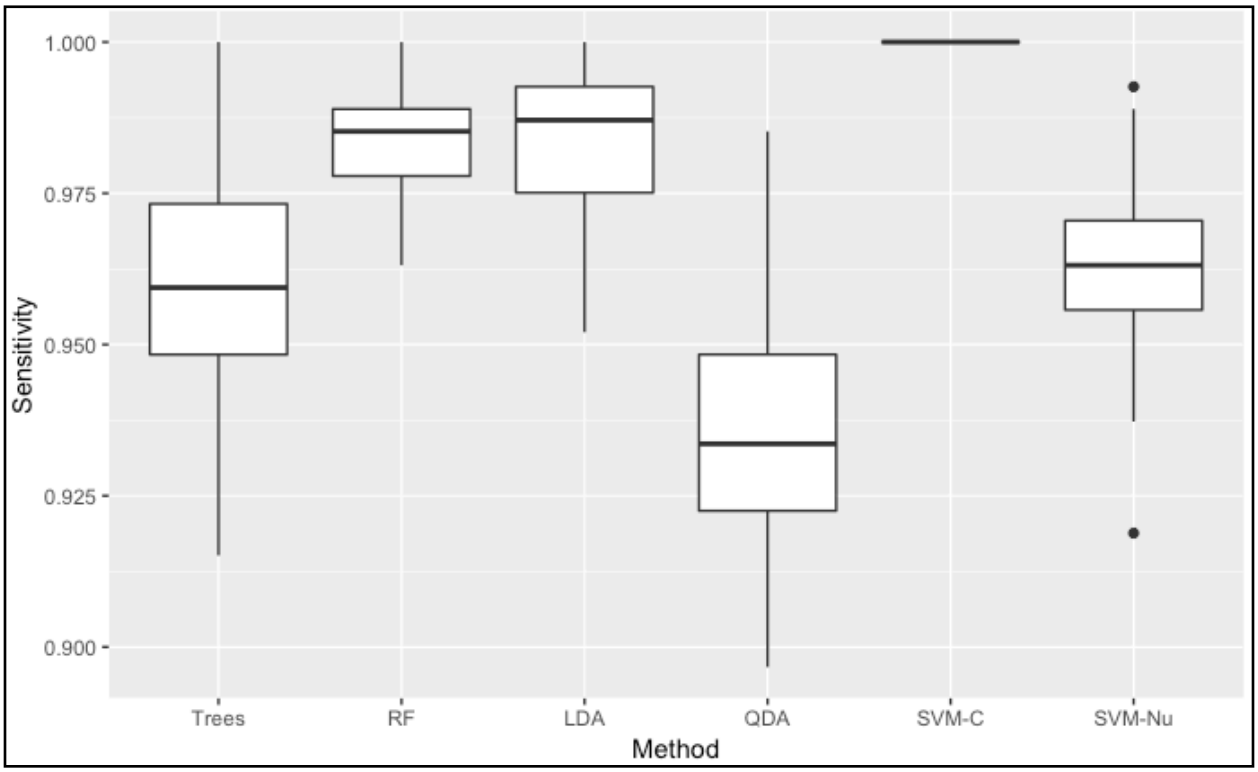
APPENDIX 8: Histogram of the Absenteeism.time.in.hours Variable



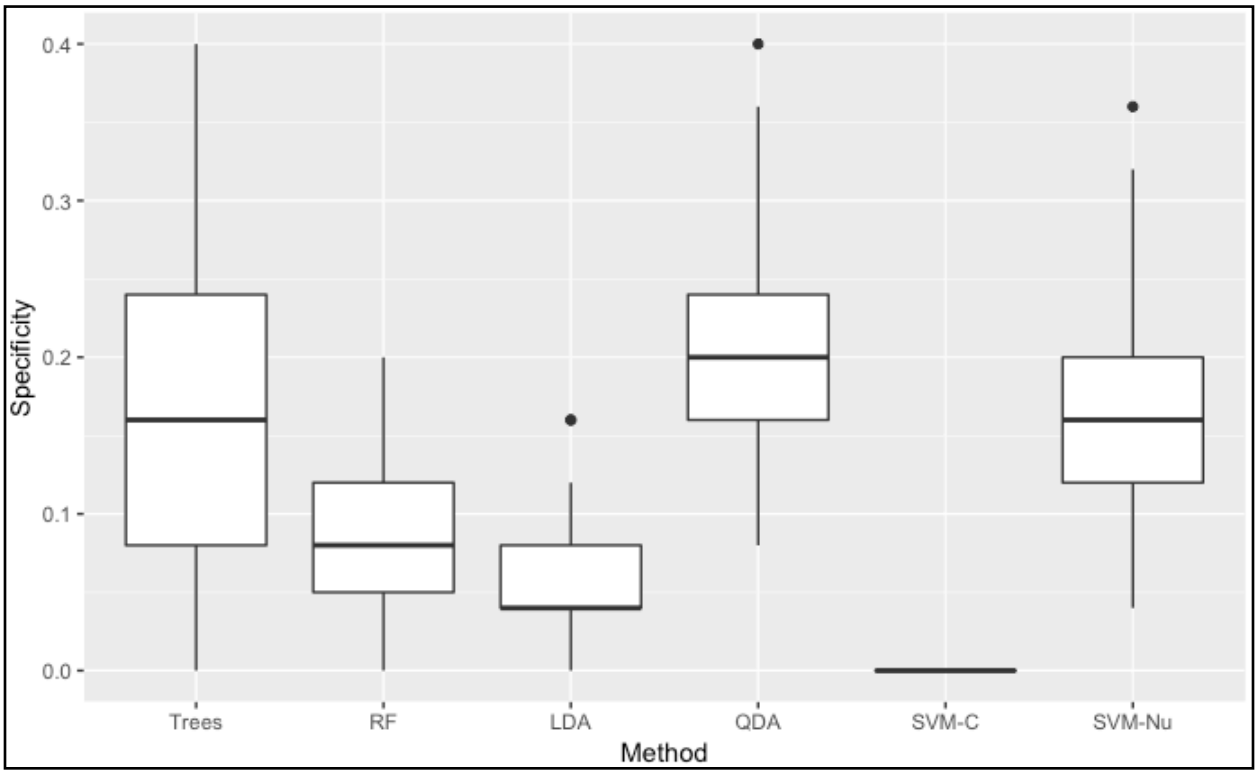
APPENDIX 9: Accuracy of Initial Models



APPENDIX 10: Sensitivity of Initial Models

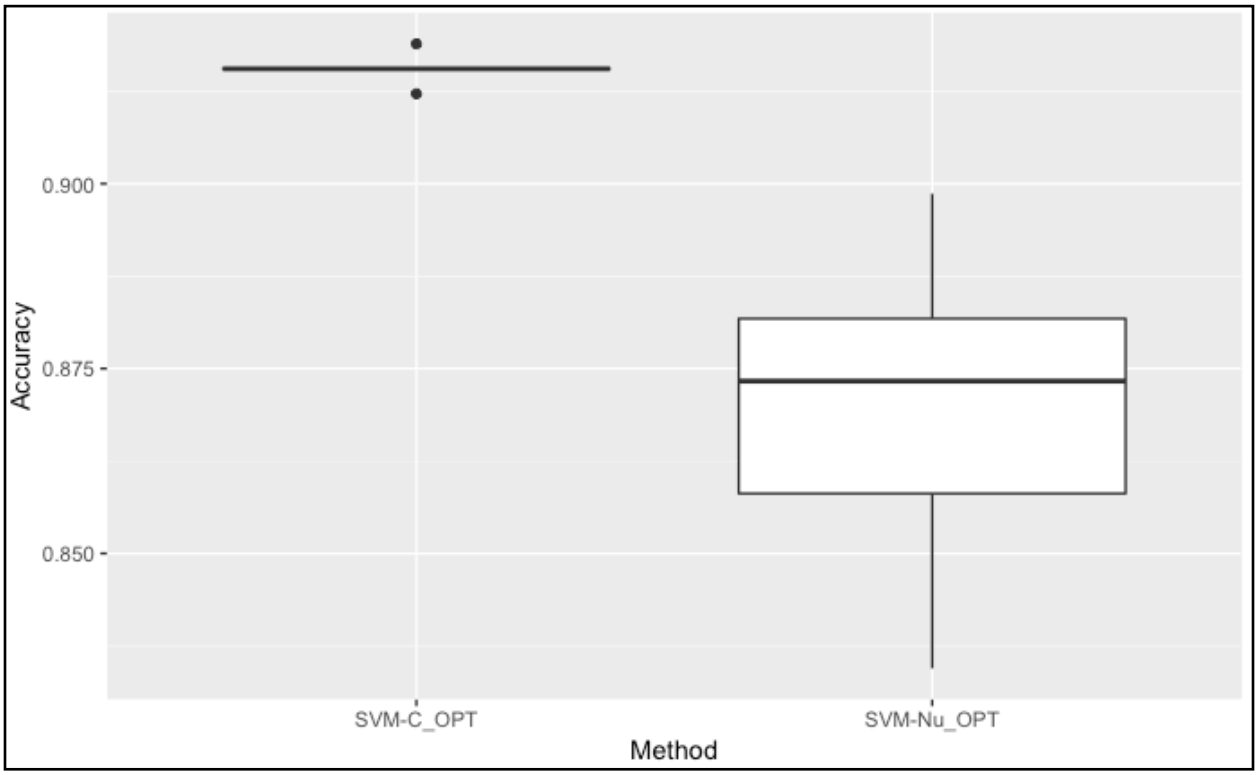


APPENDIX 11: Specificity of Initial Models



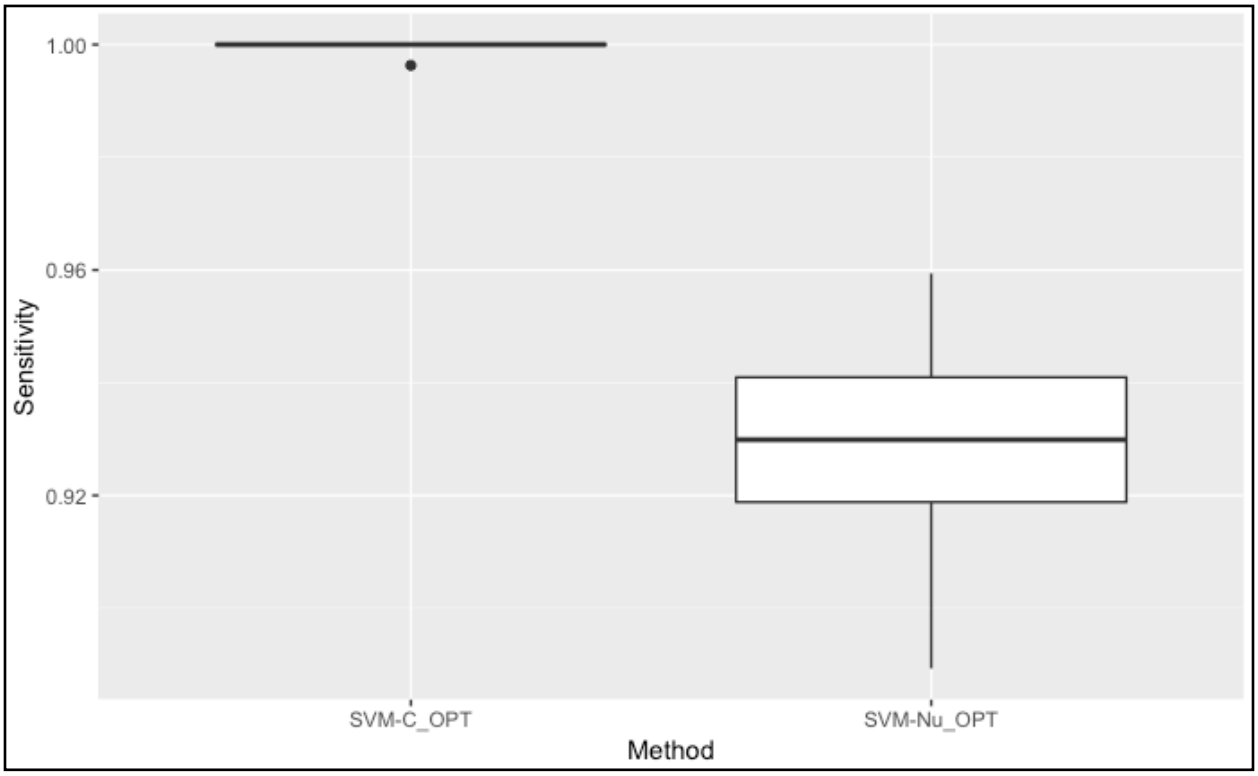
APPENDIX 12: Accuracy for SVM-C and -Nu Optimized Models

Note SVM-C shows the same results for the optimized model that it did for the initial.



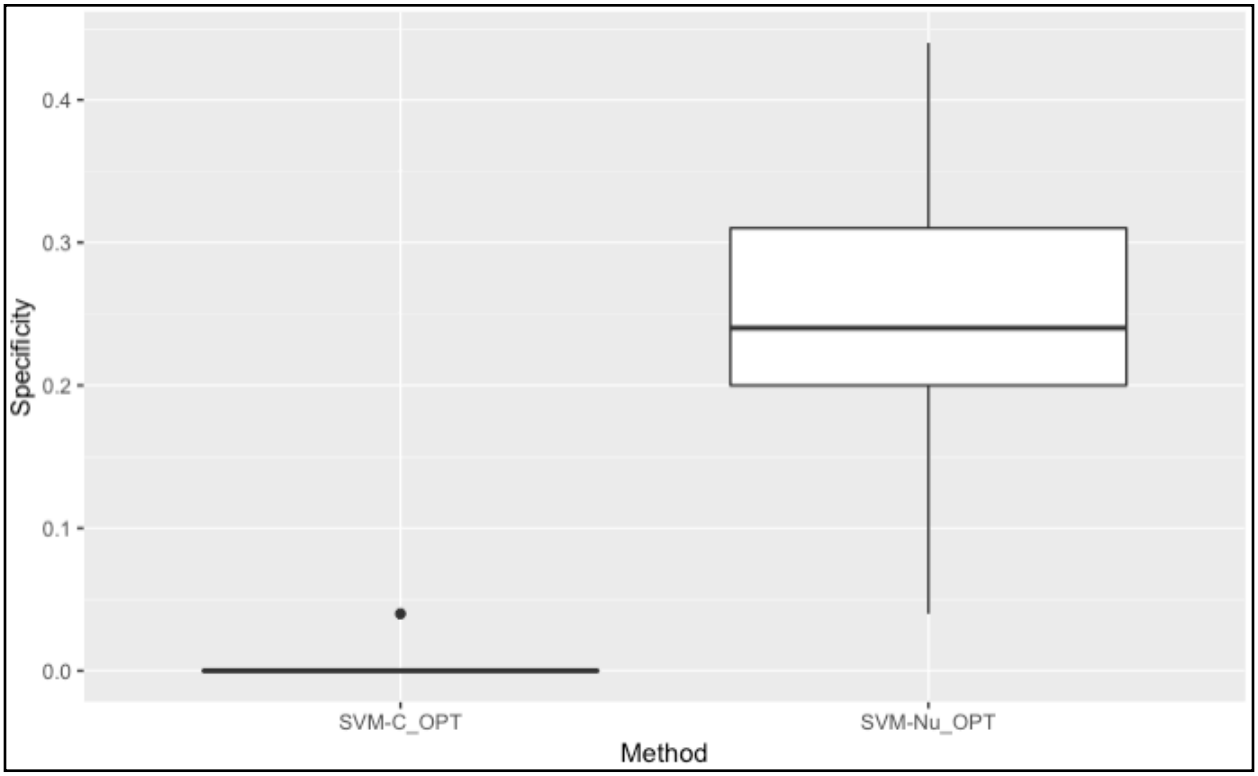
APPENDIX 13: Sensitivity for SVM-C and -Nu Optimized Models

Note SVM-C shows the same results for the optimized model that it did for the initial.

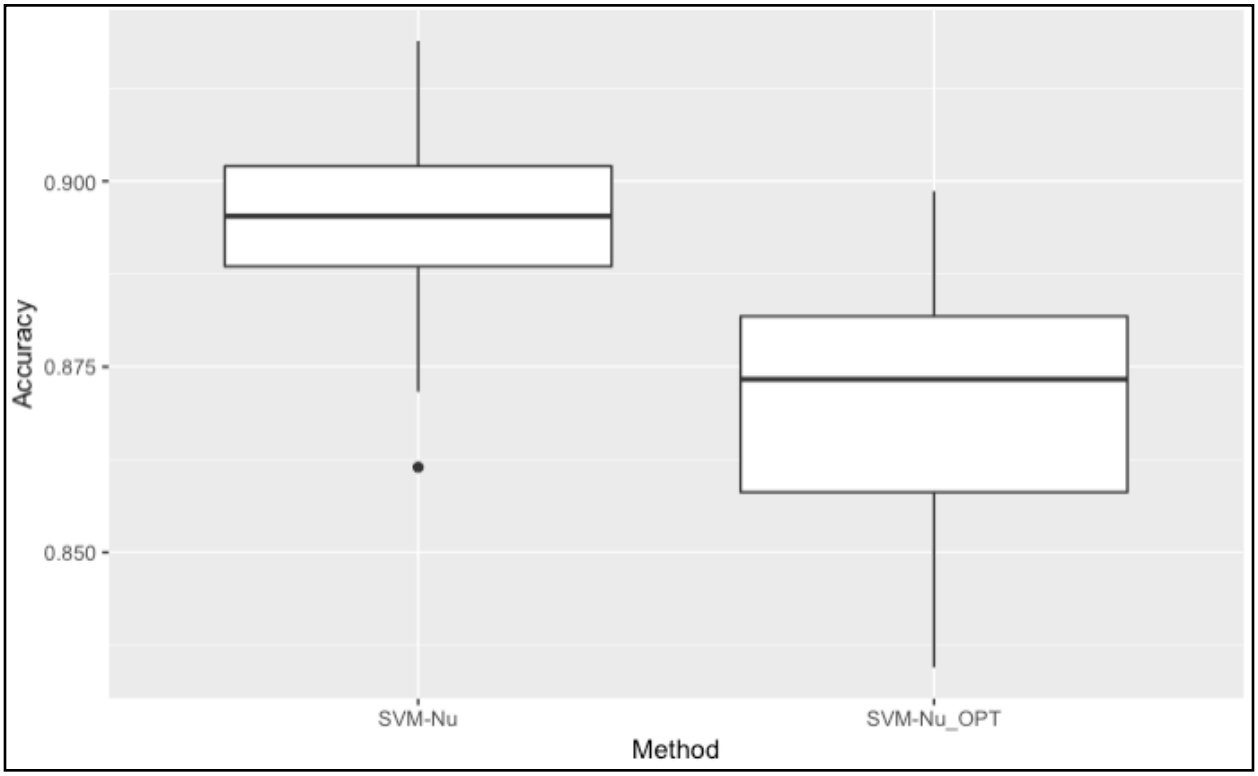


APPENDIX 14: Specificity for SVM-C and -Nu Optimized Models

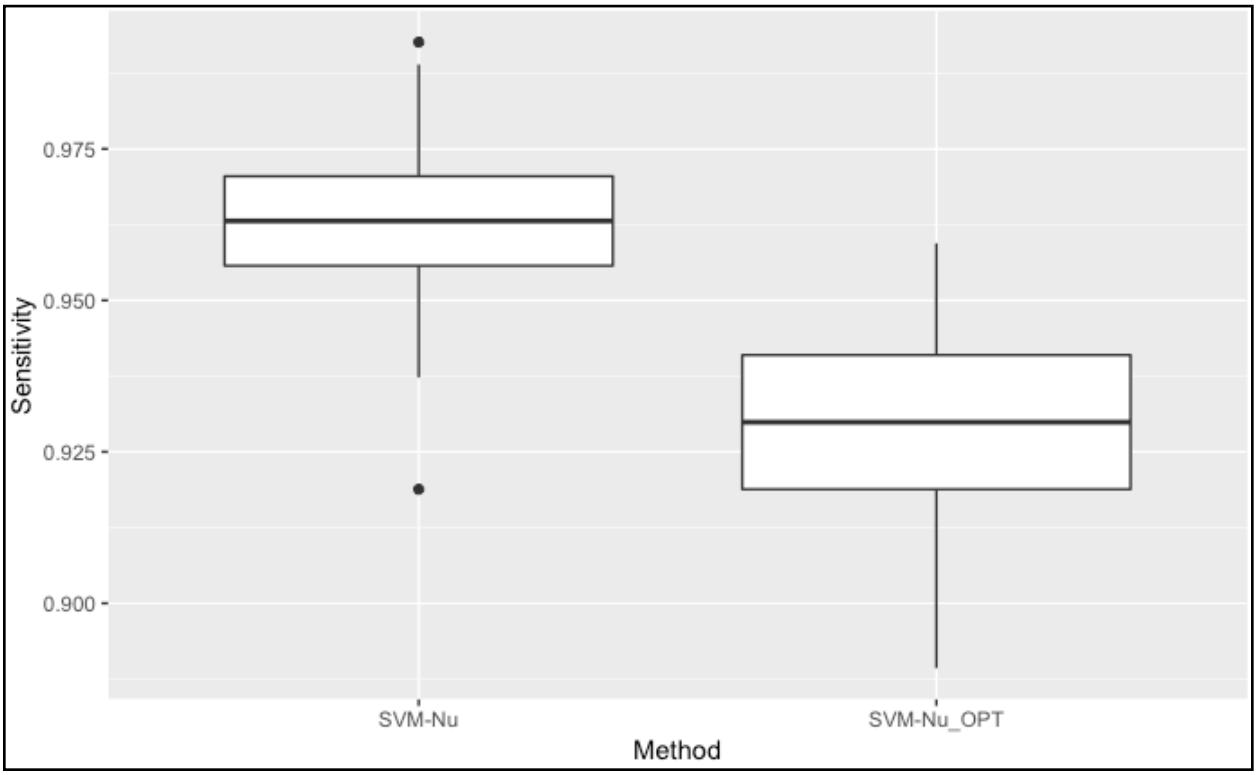
Note SVM-C shows the same results for the optimized model that it did for the initial.



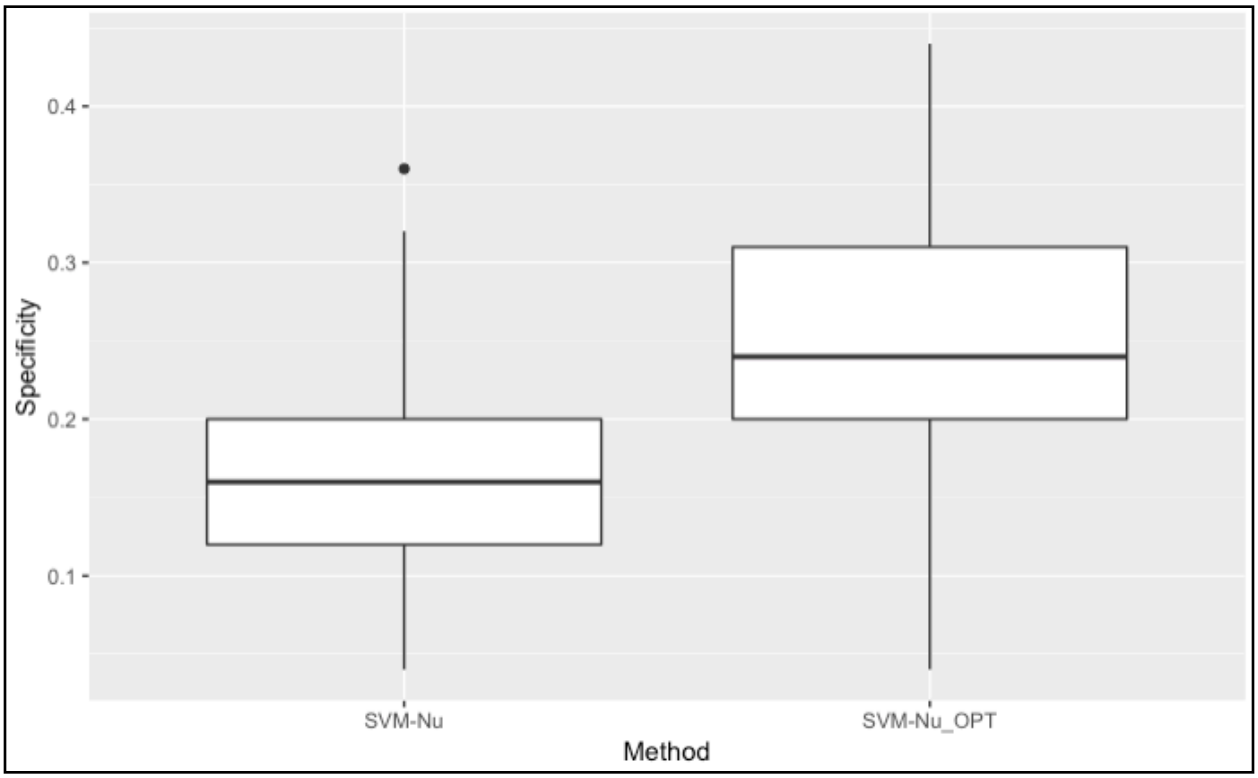
APPENDIX 15: Accuracy of Optimized Model



APPENDIX 16: Sensitivity of Optimized Model



APPENDIX 17: Specificity of Optimized Model



APPENDIX 18: Confusion Matrices of Initial & Optimized Models (Respectively)

```
Confusion Matrix and Statistics

          Reference
Prediction   D   W
         D 252  18
         W  19   7
```

```
Confusion Matrix and Statistics

          Reference
Prediction   D   W
         D 258  20
         W  13   5
```