Tatiana Leonovich

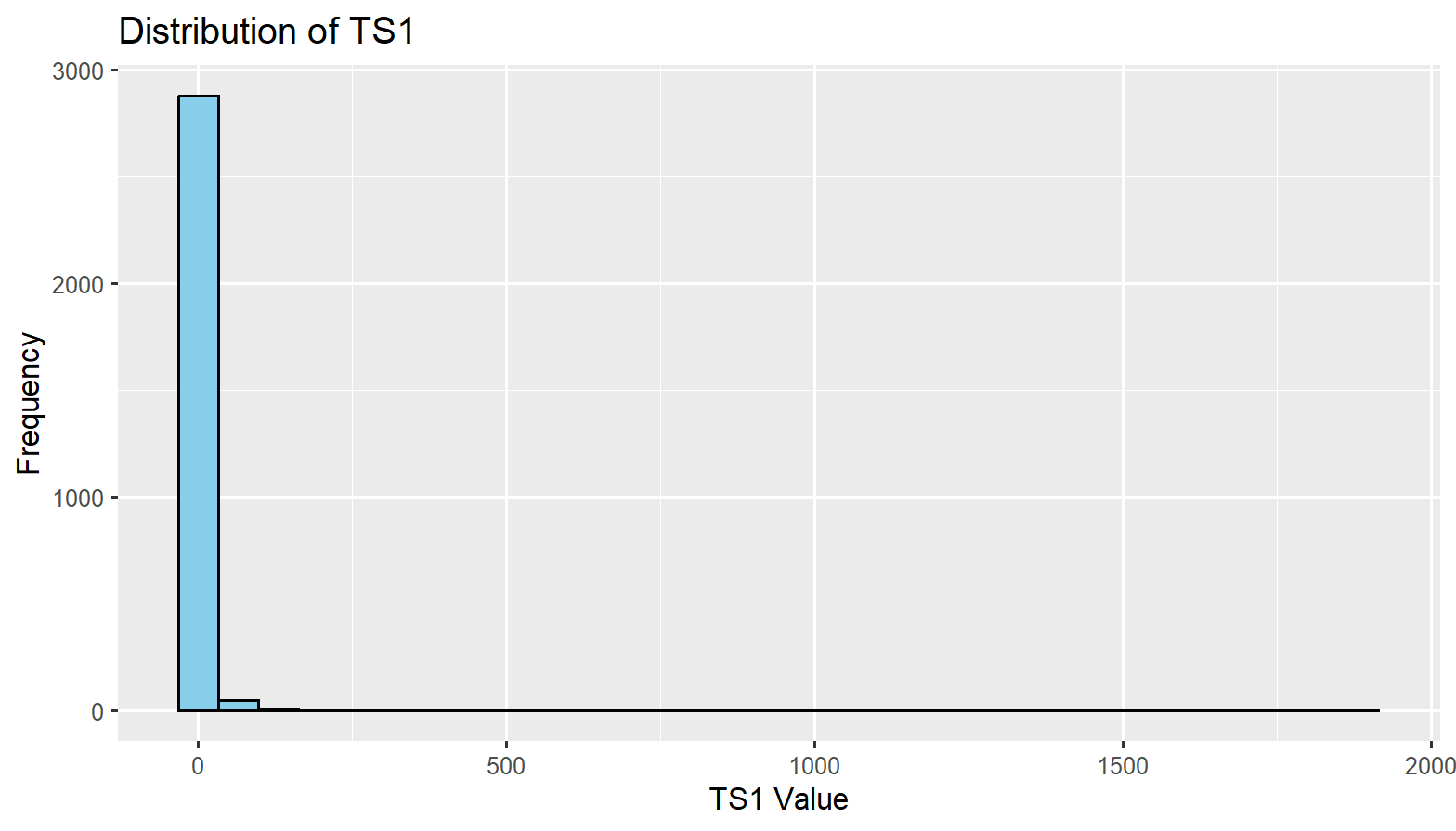
Dr. Ahmed Eleish

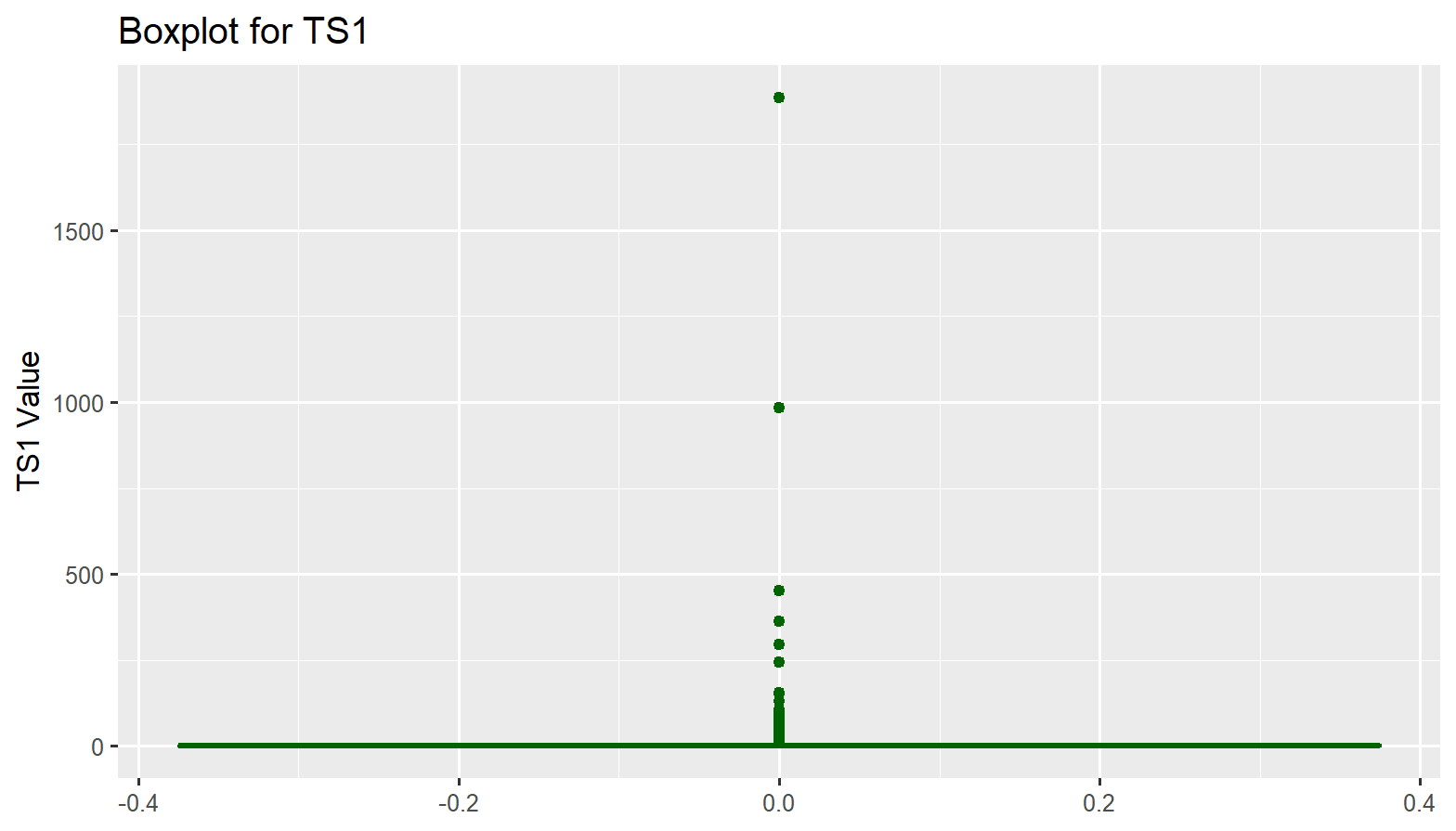
Data Analytics 4000

Assignment 7

For my assignment I used the facebook economy data set.

1. For the IDLink variable, the minimum value is 477, the first quartile is 21,317, the median is 41,333, and the maximum is 57,960. The mean value is 35,268, indicating the central tendency of these IDs. For the TS1 to TS77 variables, I have observed various trends. For most of these variables, the minimum value is 0, indicating that there are instances where these time-series variables do not have recorded values. The first quartile shows a relatively low frequency of recorded events, typically at 0 or 1. The median value usually hovers around 2, suggesting that in the central half of the data, the values are fairly concentrated near 1 or 2. The mean values range from around 4.9 for TS1 to 28.64 for TS77, which indicates that the distribution has a slight skew towards higher values for some variables, especially towards the later variables like TS66 to TS77.For the higher percentiles (third quartile), the values also tend to be higher, with most time-series variables peaking at values between 8 and 16, but some even reaching as high as 1886, which suggests possible outliers or rare, extreme events. I calculated the IQR and removed the outliers from the data.





1b. For the Model Development, Validation, and Optimization project, I chose two models that encompass different types of machine learning tasks, linear regression and random forest classification. These models are suitable for various real-world datasets and allow for comprehensive testing and evaluation. These models were selected to examine both a traditional statistical approach (Linear Regression) and a more complex, non-parametric approach (Random Forest) to predict the outcome variable, TS1, based on the predictors TS2, TS3, and TS4. While linear regression is effective when the relationships between variables are linear, it might not perform well if the relationships are non-linear or if there are complex interactions between the predictors. Random Forest is an ensemble learning method that can handle both linear and non-linear relationships, as well as capture interactions between variables automatically. By using Random Forest, I aimed to evaluate a more robust, non-parametric model that can work well with high-dimensional data and reveal non-linear trends that linear regression may miss.

First, the linear regression.The numerical variables (TS1, TS2, TS3, TS4) were scaled using standardization to normalize the data, making sure that no single variable dominates the model due to differences in scale. Any categorical variables were converted into factors to ensure proper handling by the models. To begin, I built a linear regression model using TS1 as the dependent variable and TS2, TS3, and TS4 as independent variables. I then evaluated the model's performance on the test data using Root Mean Squared Error (RMSE) as the metric to assess model fit.

1c. The intercept is 0.23798, which is statistically significant with a very small p-value (less than 2e-16), suggesting that the baseline value of TS1 is significant. The coefficient for TS2 is 0.97691, with a very high t-value (45.241) and a p-value less than 2e-16. This means that for every unit increase in TS2, TS1 increases by approximately 0.977, and the effect is highly significant. The coefficient for TS3 is -0.25497, with a negative sign and a significant t-value (-11.570), indicating that an increase in TS3 is associated with a decrease in TS1, and this effect is statistically significant. The coefficient for TS4 is -0.05039, with a t-value of -4.781 and a very small p-value (1.86e-06), showing a statistically significant negative relationship between TS4 and TS1.

The RSE is 0.8889, indicating that the model's predictions deviate from the actual values by approximately 0.89 units on average. The R-squared value is 0.7366, meaning the model explains 73.66% of the variance in TS1. This is a fairly strong fit, suggesting that the linear relationship between the predictors and the target is reasonably good. The adjusted R-squared value is 0.7362, which accounts for the number of predictors in the model and suggests that the model remains strong even after penalizing for additional variables. The F-statistic of 1963 with a p-value less than 2.2e-16 further confirms the overall significance of the model.

For the random forest, the confusion matrix displays the predicted vs. actual values of category\_var for each class (0 through 7). Each cell on the diagonal represents the number of correct predictions for that class. For example, the model correctly predicted all instances for class 0, as indicated by the value 283 in the (0, 0) position. The model achieved an accuracy of 1, meaning it classified all the test instances correctly. This indicates that the Random Forest model is performing perfectly on the test data. The confidence interval for accuracy is between 0.9929 and 1, further indicating a high level of certainty in the model's performance. The Kappa statistic is 1, which suggests perfect agreement between the predicted and actual categories, further confirming the model's exceptional performance.

The Random Forest model has demonstrated outstanding classification performance on the test set, achieving perfect classification accuracy across all classes. This could be due to the model's ability to handle complex relationships and interactions within the data. The 100% sensitivity, specificity, and predictive values for all classes show that the model is reliable and highly effective for this classification task.

