Dani DiTomasso
Professor Eleish
Data Analytics
21 November 2024

Assignment 7 Written Response

Question 1

Upon loading the dataset into R Studio, the first step was to clean the database. Since no fields had any NA or missing values, it was not necessary to clear those rows from the overall table. However, any outliers found within the Absenteeism.time.in.hours variable were removed from the dataset before any statistical analysis was run. This variable was chosen in particular for outliers due to the fact that it is the main factor that was being studied in the dataset, which is Absenteeism at the workplace. Any other variables are recorded as potential factors that may influence why an individual may be absent from work, which was the main focus of this analysis.

For an exploratory analysis, 4 main histograms were created in order to gain an understanding of the dataset. These variables include “Absenteeism.time.in.hours”, “Age”, “Distance.from.Residence.to.Work”, and “Transportation.expense.” Each of these histograms, labeled as figures 1-4 below, provide a brief overview of the dataset and which values appear most frequently. These histograms revealed that the Age where people are most likely to be absent from work are 30 - 40 years old, and they are likely to be absent either 0-4 hours, or 8 hours from work. Additionally, The distance from residence to their place of work is likely either 25 miles or 50 miles, and the total transportation expense is most frequently around 175 or 225 to 300 USD.

Question 2

When conducting the statistical analysis, three different types of models were mainly utilized: these include Linear Models, SVM models with Linear and Polynomial kernels, and a KNN model. The first analysis, linear model, was conducted on the variable Absenteeism.time.in.hours, with the goal of finding the variables with the most statistical significance towards having some correlation to the amount of time an individual was absent from work. The first linear model included many variables, and as seen in figure 5, revealed that the variables of Social.smoker, Social.drinker, Disciplinary.failure, Age, Distance.from.Residence.to.Work, Transportation.expense, and Reason.for.absence all have a p-value less than 0.05, proving them statistically significant. However, in order to see if these variables had any direct effect or causation to the amount of time someone was absent from work, a correlation matrix was then conducted on the variables that were found to be statistically significant. As seen in figure 6, most variables had little to no effect on Absenteeism.time.in.hours, proving the critical statistical concept that correlation does not always mean causation to be true. Finally, a graph depicting absenteeism time vs transportation expense was created, including the distance from residence to work for each point (figure 7).

The next two models, SVM and KNN, were conducted based on the Seasons variable. Each of the seasons, represented by numbers 1-4, were run through training models and compared to other variables through the equation “Seasons ~ Absenteeism.time.in.hours + Disciplinary.failure + Age + Distance.from.Residence.to.Work + Transportation.expense + Reason.for.absence”. This equation was then put through both an SVM linear and polynomial kernel, and then passed through the “calc_metrics” function, created to calculate the Precision, F1, and Recall values of each of the matrices. After this, a KNN model was conducted using the same equation as above. The elbow method was used to determine the appropriate K value, which in this case, was 2 (figure 8). The KNN model was also then passed through the function of “calc_metrics” and then the values were then compared in a table-like structure that was printed within the console log. As seen in figure 9, this revealed that the Precision and Recall values of the KNN model were about 66%, whereas for both SVM models the values ranged from 13% to 26%. For this study, these values show that the KNN model is much more likely to correctly predict values compared to the SVM models, and therefore should be taken more heavily into consideration.

Question 3

As a conclusionary analysis, a visual representation of the KNN model (which was found to be the most likely to correctly predict values) was created. As seen in figure 10, the analysis shows a high frequency at the center of the diagram moving in a diagonal direction. Overall, a diagonal frequency result also generally refers to an accurate representation of values, in which the predicted and actual numbers are closely related to each other.

Figures

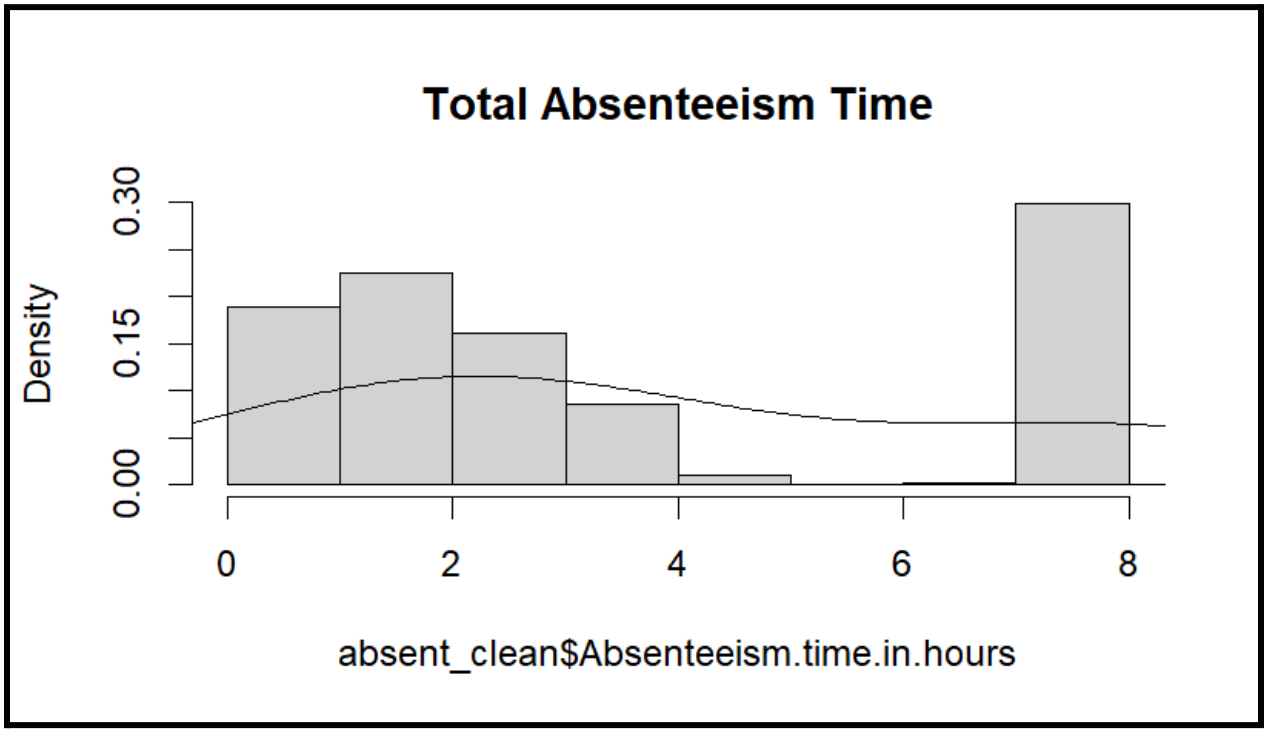

Figure 1: Histogram of total Absenteeism Time in Hours, based on the density of the total time spent absent from work. (1)

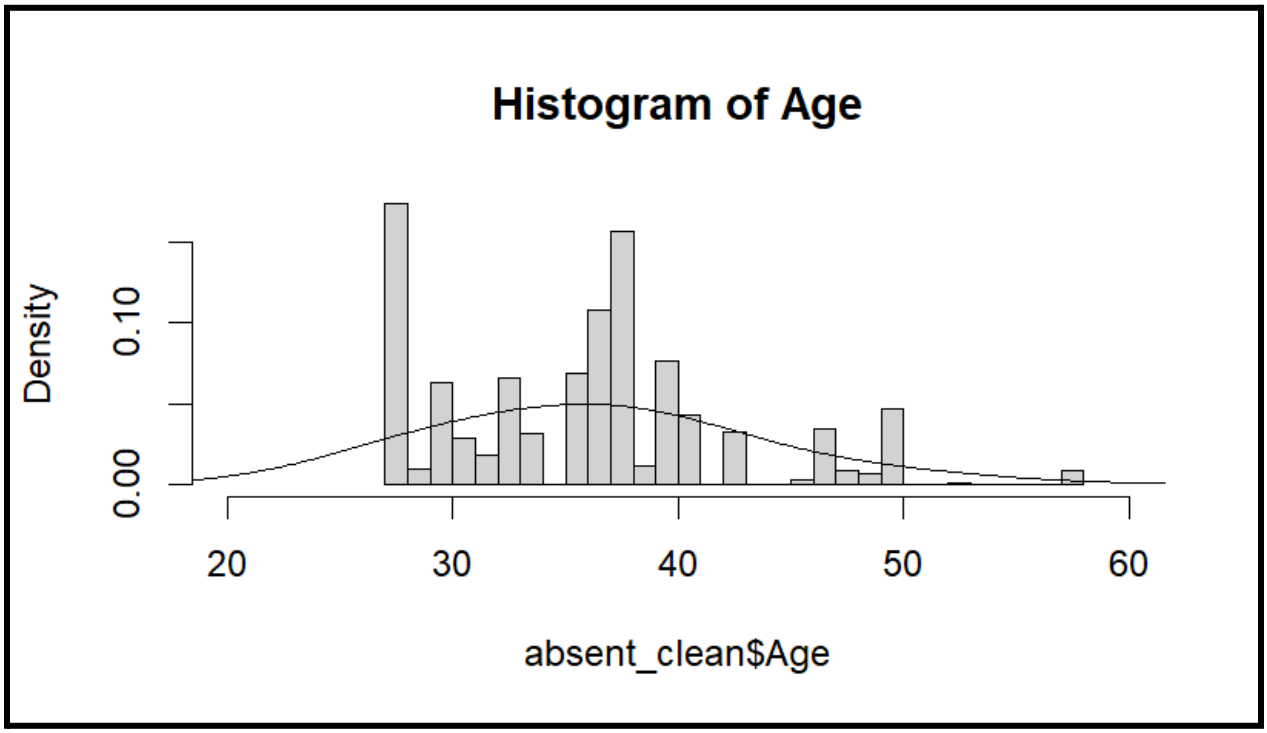

Figure 2: Histogram of Age for those who are absent from work. (1)

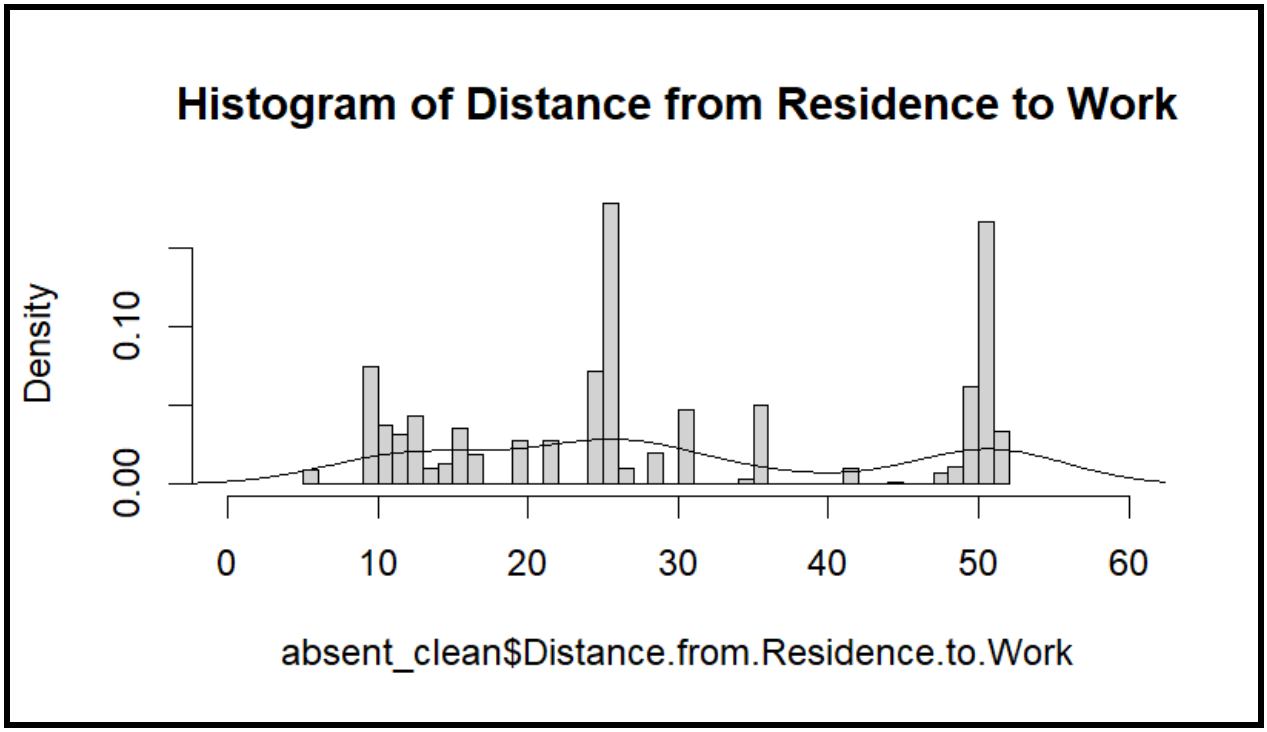

Figure 3: Histogram of the Distance from the individual's residence to their place of work. (1)

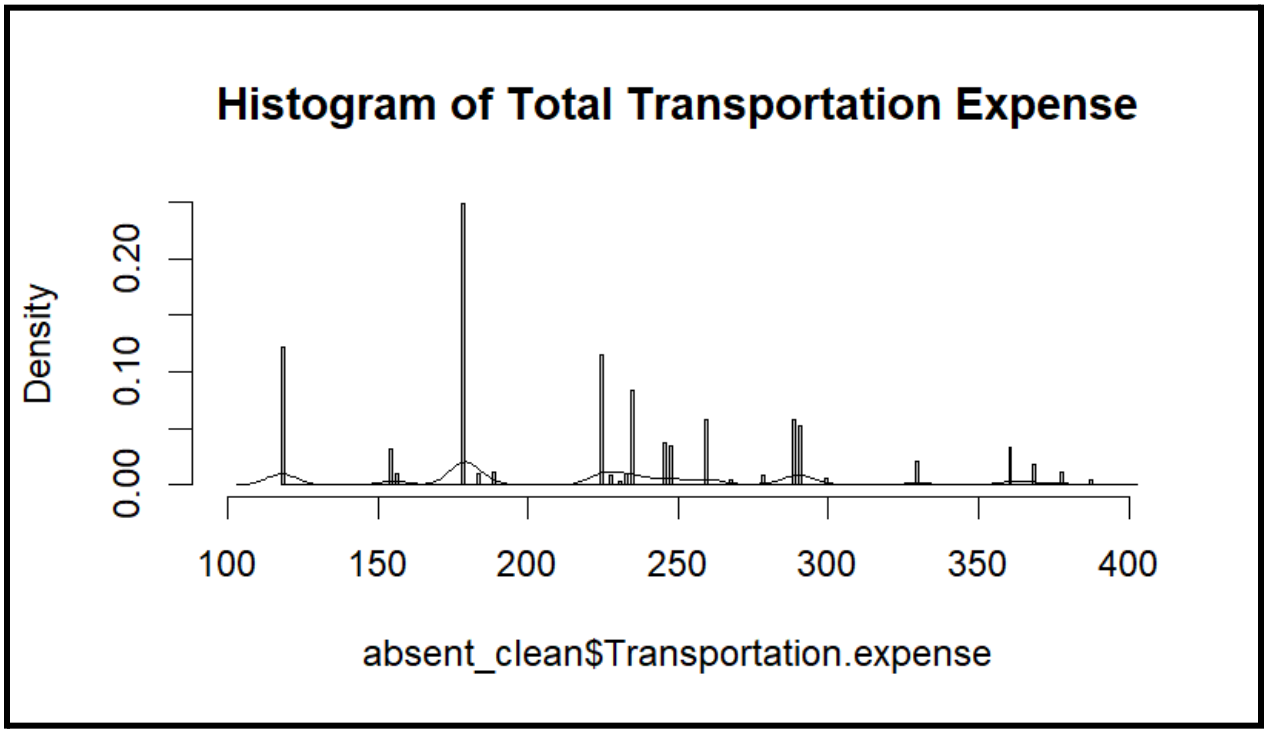

Figure 4: Histogram of the Total Transportation Expense for individuals going to and from their place of work in USD. (1)

```
Call:
lm(formula = Absenteeism.time.in.hours ~ Social.smoker + Social.drinker +
    Body.mass.index + Pet + Education + Disciplinary.failure +
    Work.load.Average.day + Age + Distance.from.Residence.to.Work +
    Transportation.expense + Day.of.the.week + Seasons + Reason.for.absence,
    data = absent_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6357 -1.8535 -0.3923  1.5969 12.8888

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                    5.898210   1.603536   3.678 0.000253 ***
Social.smoker                  1.116973   0.456567   2.446 0.014678 *
Social.drinker                 1.110637   0.293356   3.786 0.000167 ***
Body.mass.index                0.057118   0.033450   1.708 0.088170 .
Pet                           -0.169198   0.096243  -1.758 0.079191 .
Education                      0.134463   0.190584   0.706 0.480720
Disciplinary.failure          -8.634420   0.596384 -14.478  < 2e-16 ***
Work.load.Average.day          0.002639   0.002990   0.883 0.377691
Age                           -0.059089   0.022255  -2.655 0.008113 **
Distance.from.Residence.to.Work -0.020237   0.009347  -2.165 0.030730 *
Transportation.expense         0.011917   0.001942   6.135 1.44e-09 ***
Day.of.the.week               -0.105950   0.080299  -1.319 0.187463
Seasons                       -0.100169   0.102440  -0.978 0.328504
Reason.for.absence            -0.168649   0.016560 -10.184  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.918 on 682 degrees of freedom
Multiple R-squared:  0.2967,    Adjusted R-squared:  0.2832
F-statistic: 22.13 on 13 and 682 DF,  p-value: < 2.2e-16
```

Figure 5: Image of the console log that determines the significant variables that have a correlation to the Absenteeism.time.in.hours variable. (2)

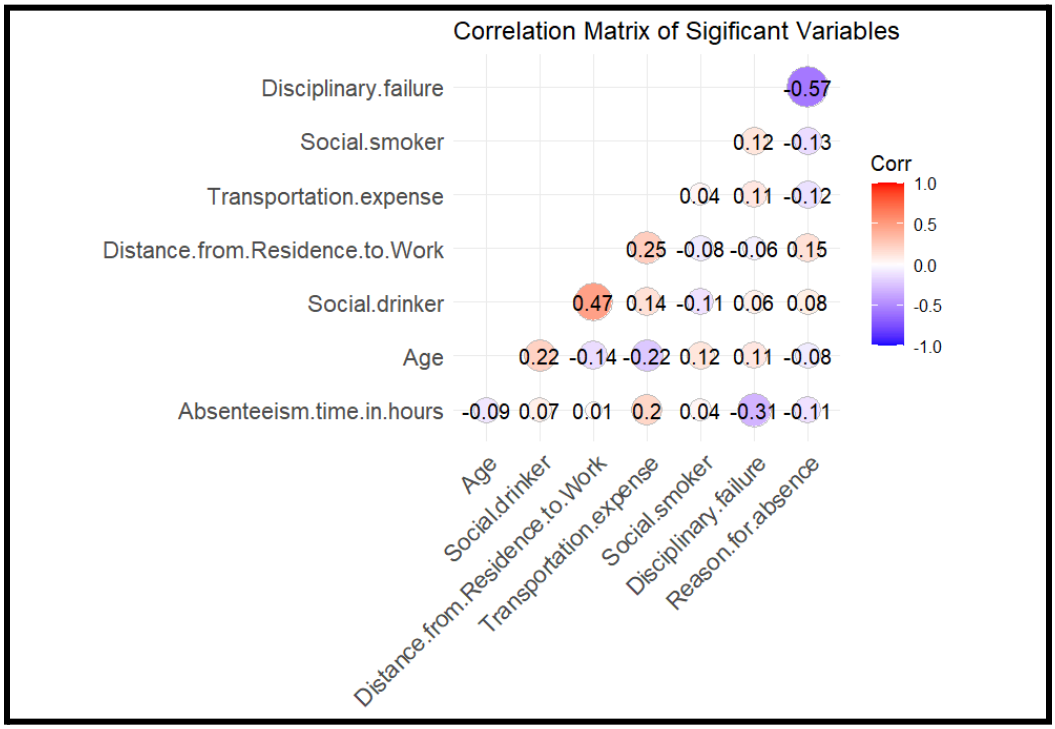

Figure 6: Correlation matrix run on significant variables determined for Absenteeism time in hours in the linear model. (2)

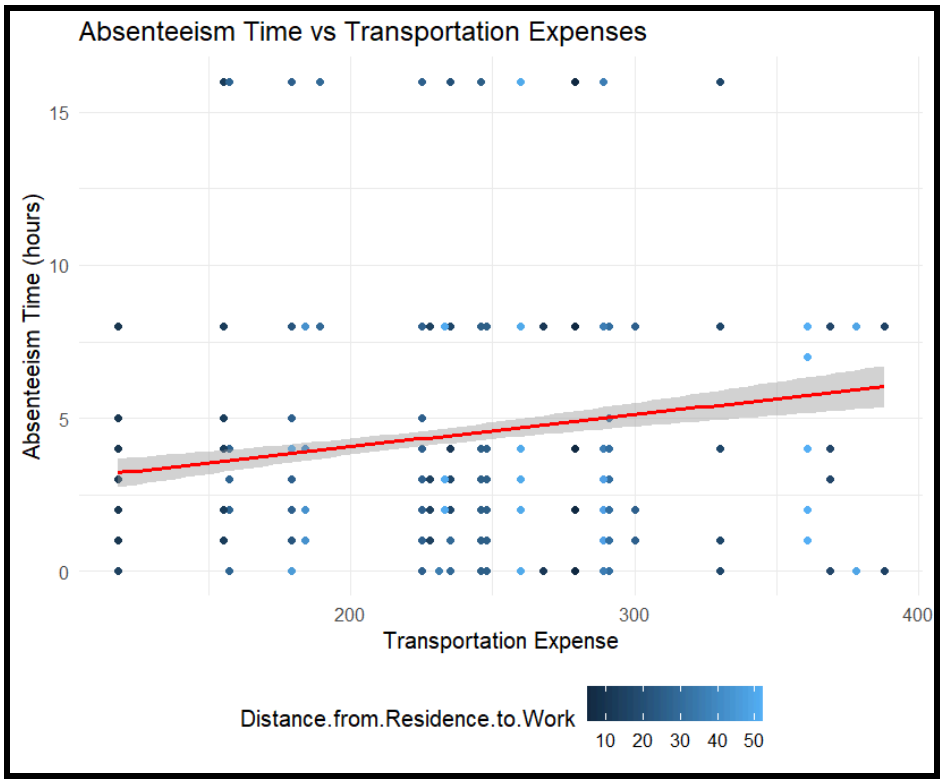

Figure 7: Scatterplot depicting Absenteeism Time based on Transportation Expenses, with each value depicting the distance each individual lived from their place of work. (2)

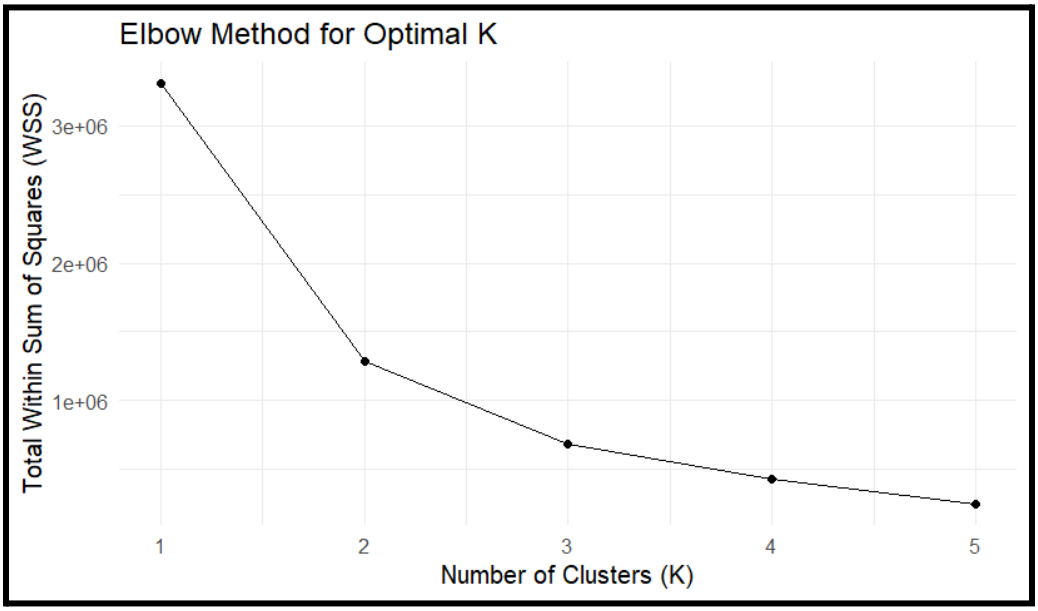

Figure 8: A plot to determine the optimal K value for a KNN training model - K=2. (2)

```
> print(all_metrics)
              Model Precision    Recall        F1
1     SVM (Linear) 0.1356707 0.2446283       NaN
2 SVM (Polynomial) 0.1623132 0.2606949       NaN
3         KNN (k=2) 0.6640954 0.6673283 0.6637036
> |
```

Figure 9: Image of console log that returns the Precision, Recall, and F1 for the SVM models and KNN models. (2)

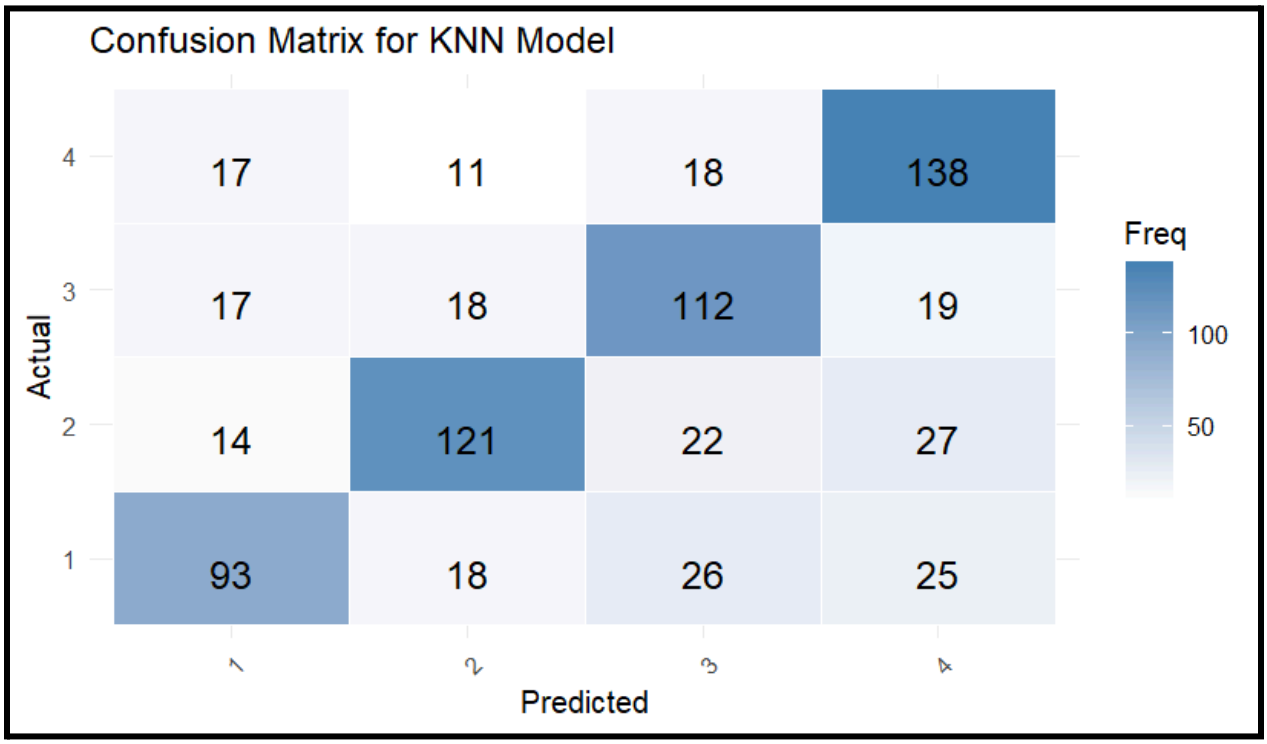

Figure 10: Plot of the confusion matrix for the KNN model. (3)

References

"Ggcorrplot Package - RDocumentation." *Rdocumentation.org*, 2023, www.rdocumentation.org/packages/ggcorrplot/versions/0.1.4.1. Accessed 22 Nov. 2024.