EARTH QUAKE PREDICTION

BUSINESS ANALYTICS WITH R - GROUP 9

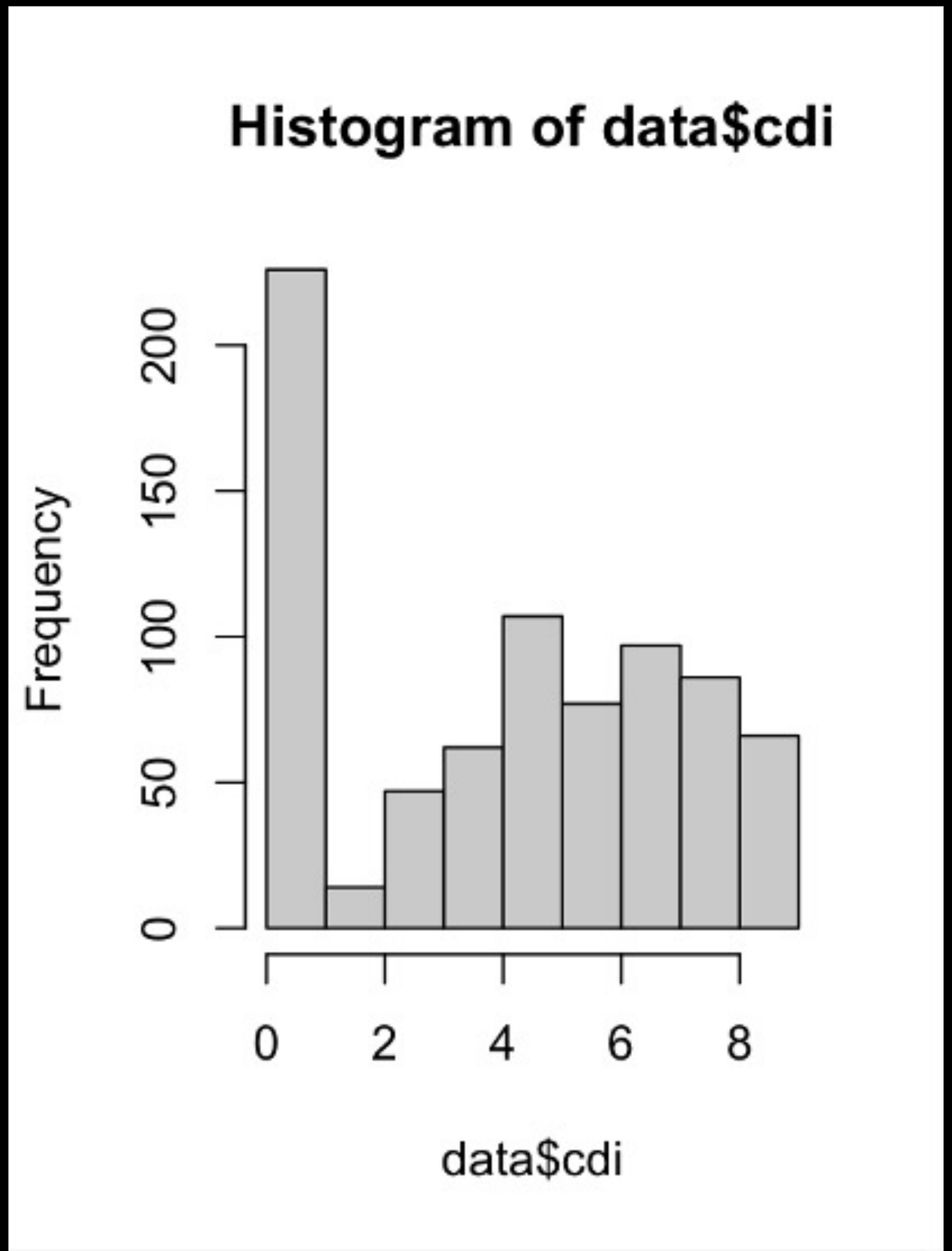
INTRODUCTION

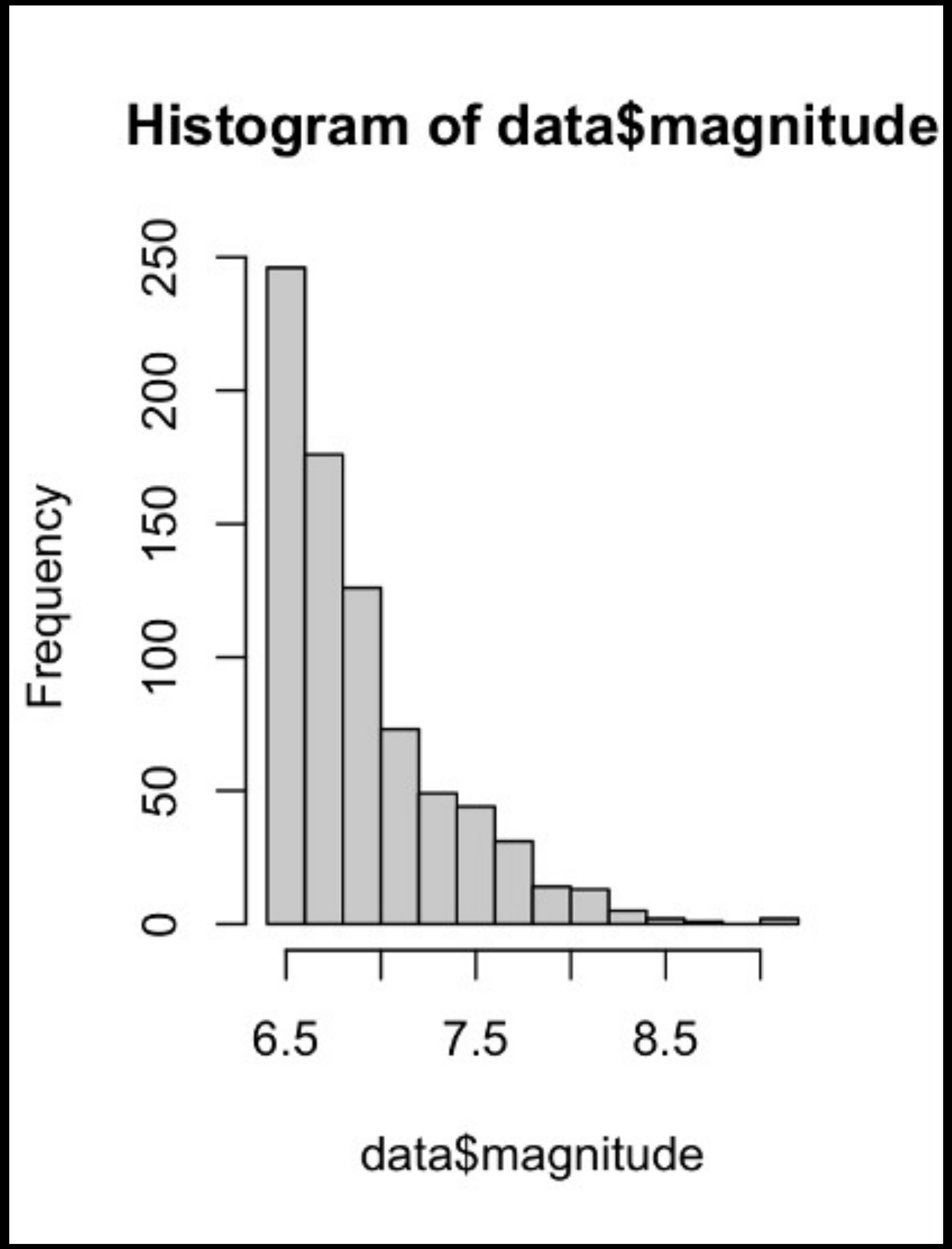
- The "Earthquake Dataset" is a collection of earthquake-related data from across the world.
- We have taken the data source from Kaggle. The dataset covers 782 earthquakes from 1/1/2001 to 1/1/2023.
- The goal of the project is to predict the areas at higher risk of earthquakes and estimate the likelihood of future seismic events.
- Ultimately, the aim is to develop a reliable and accurate predictive model that can help governments and emergency services prepare for and respond to earthquakes, potentially saving lives and reducing damage.

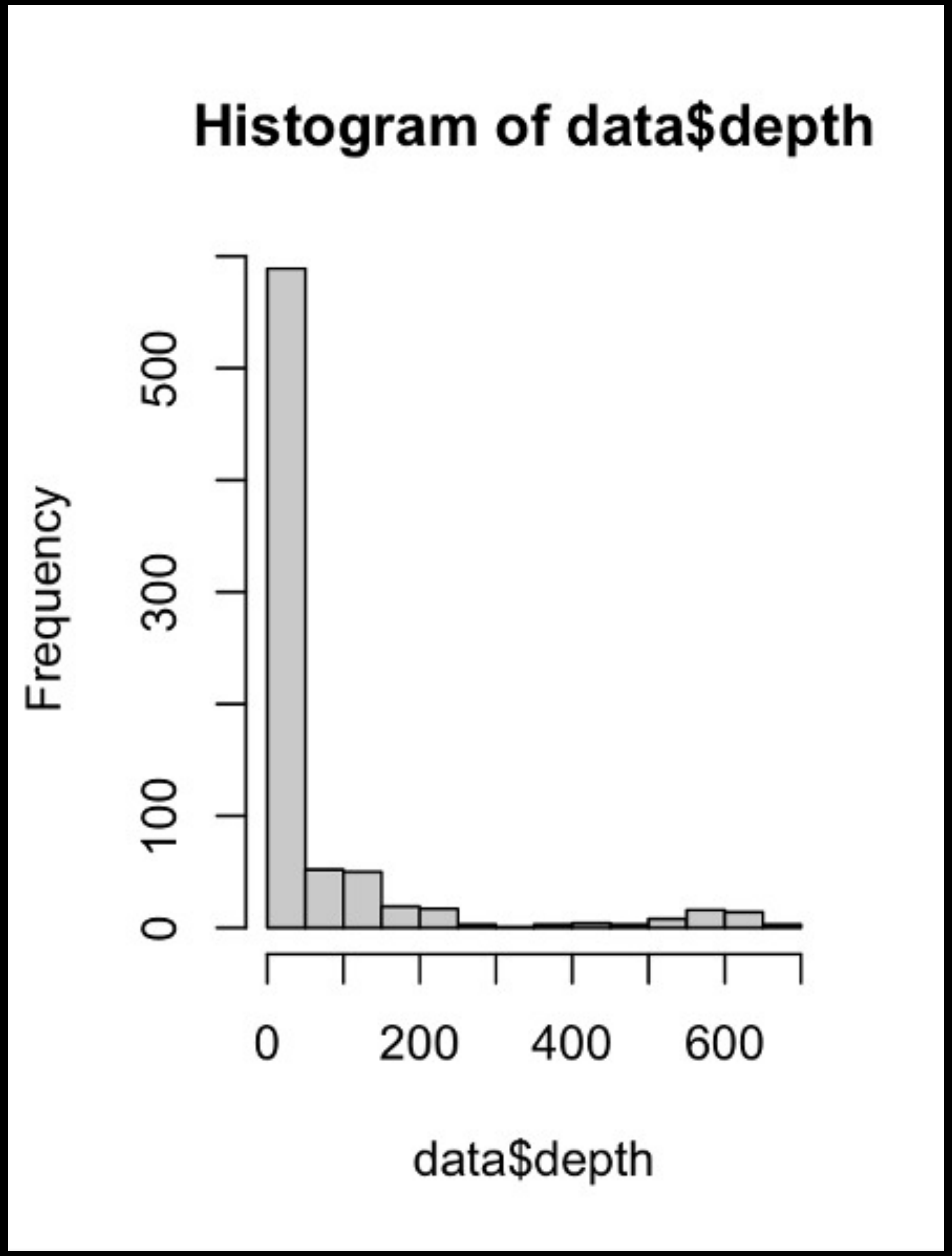
DATA EXPLORATION

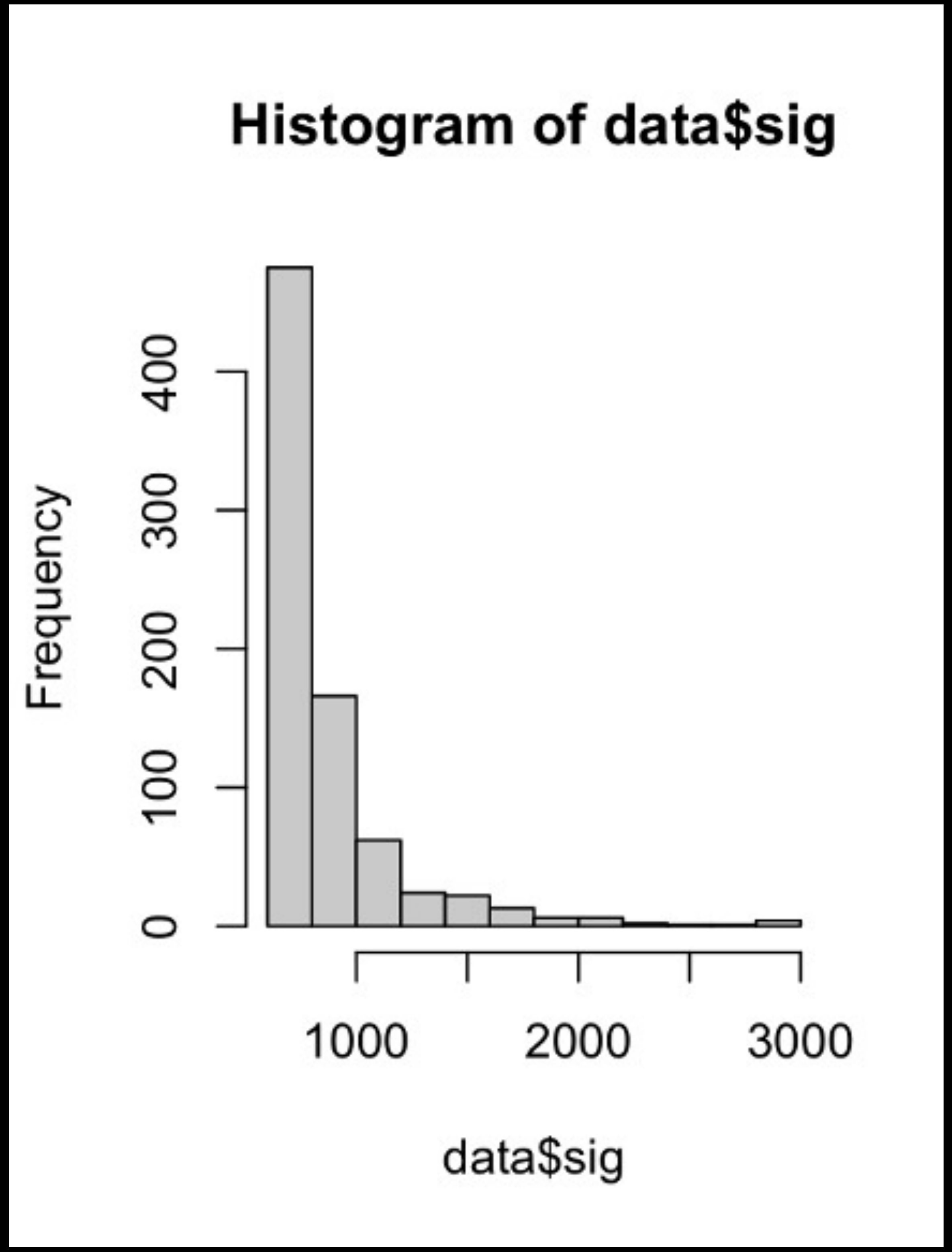
- **SOURCE**: Kaggle.com
- **DATA SET**: 782 earthquakes data of 23 years from 2001 to 2023
- **NO OF VARIABLES**: 19
- **FACTORS**: Magnitude, Significance, depth, locationResponse
- **VARIABLE**: cdi (intensity)
- **PREDICTOR VARIABLE**: Magnitude, Significance, depth, location

DATA VISUALIZATION

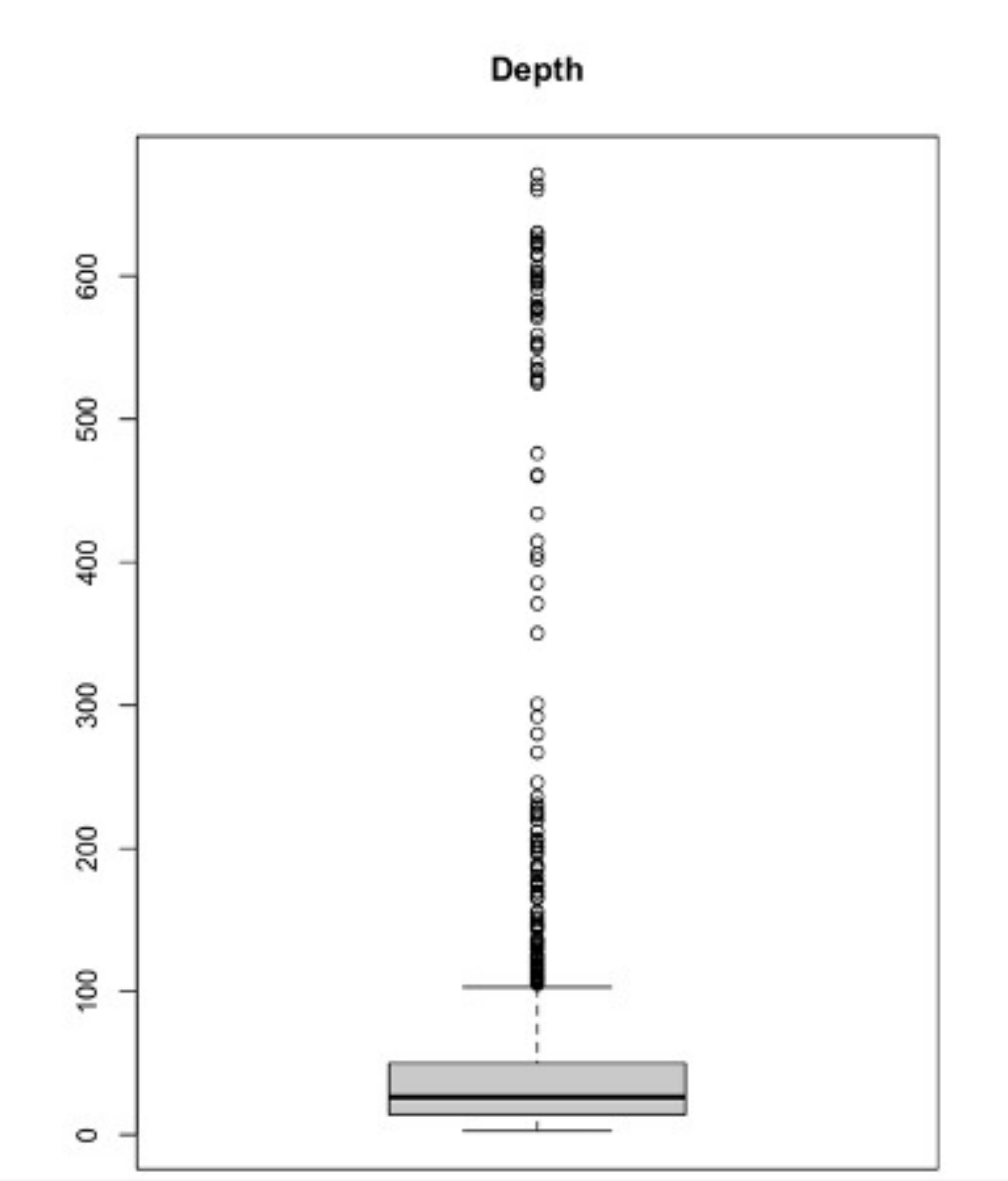


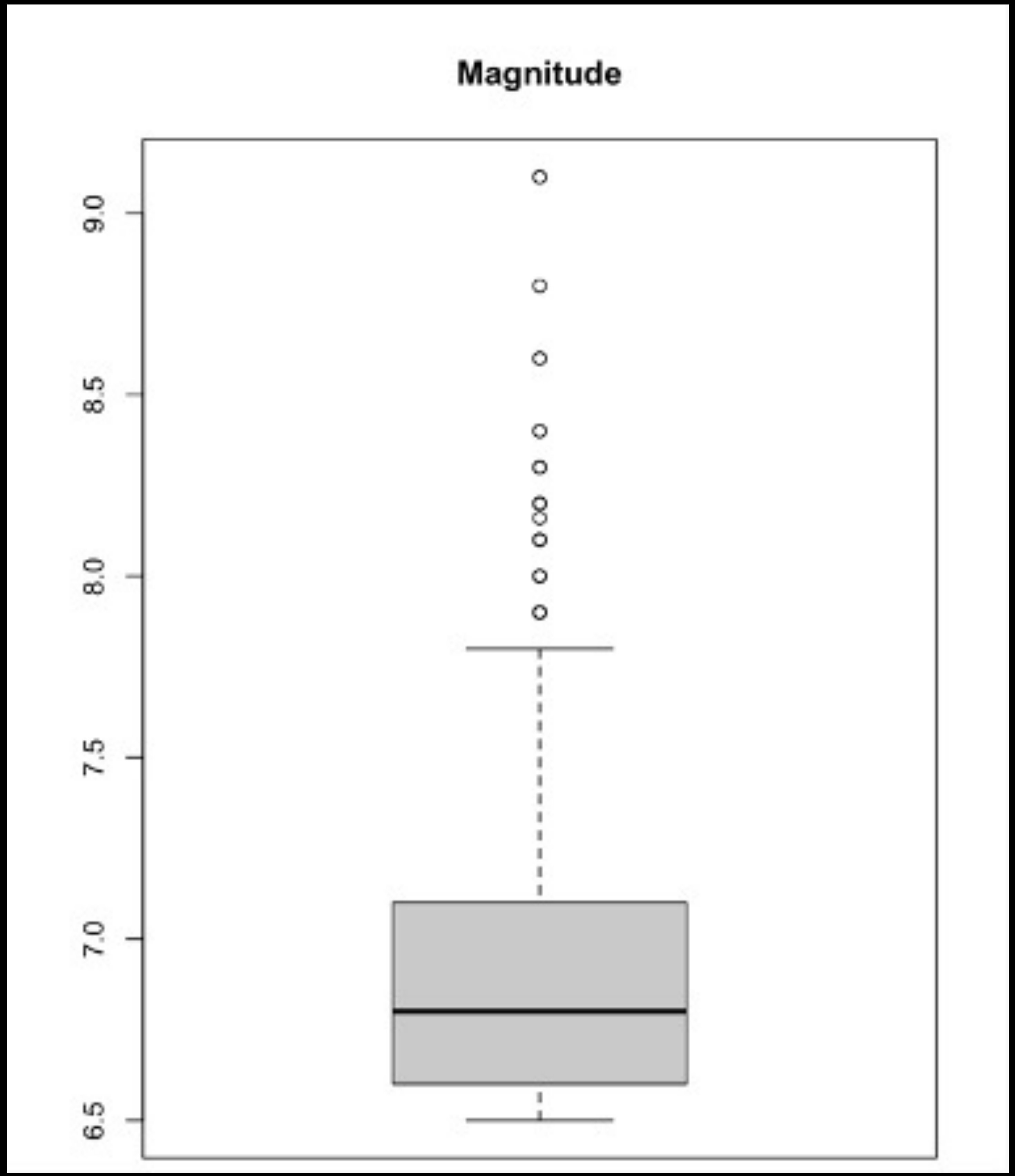


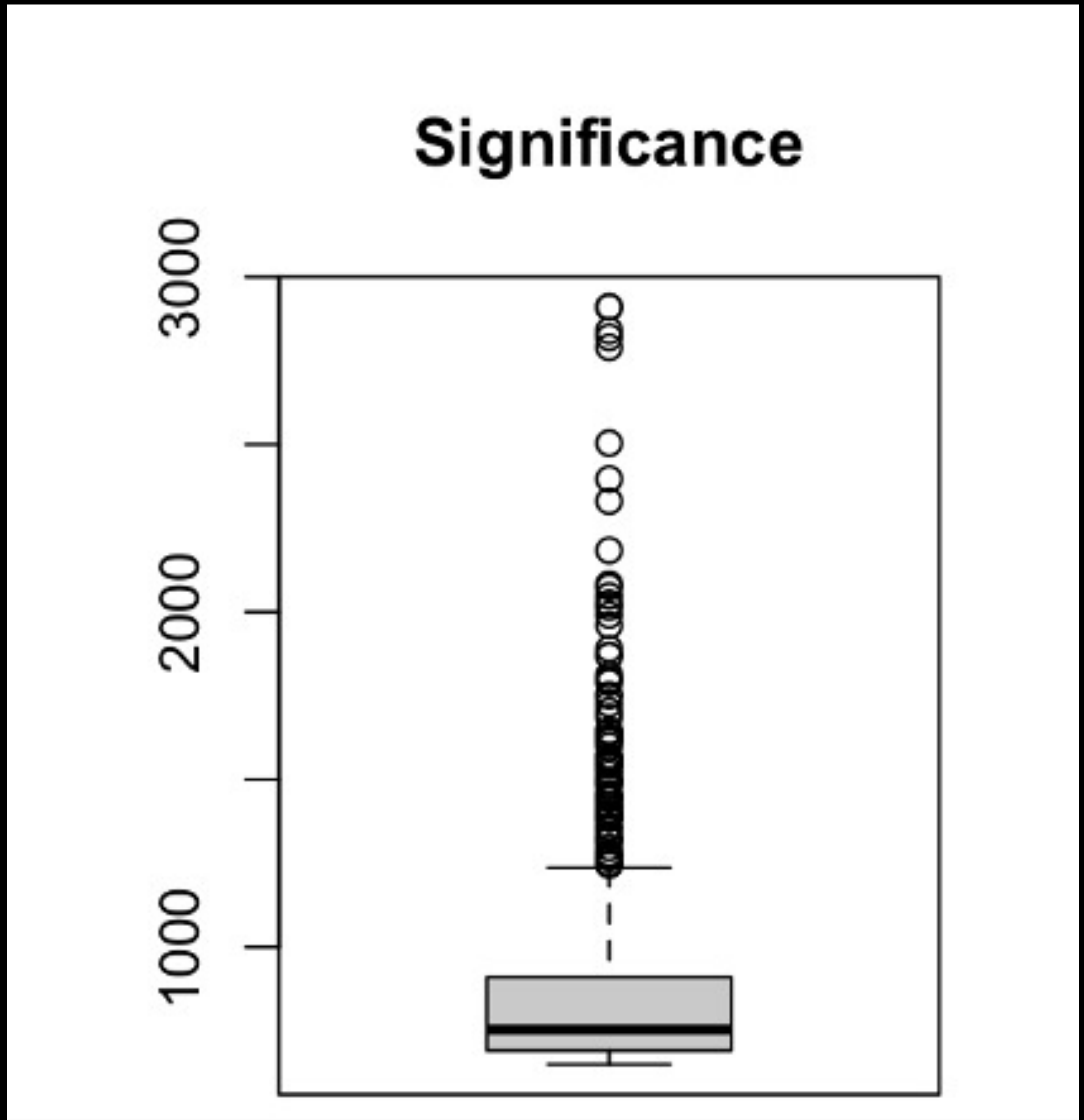




CHECKING FOR OUTLIERS







METHODOLOGY

- Splitting the data into training and testing with 123 seed for reproducib results

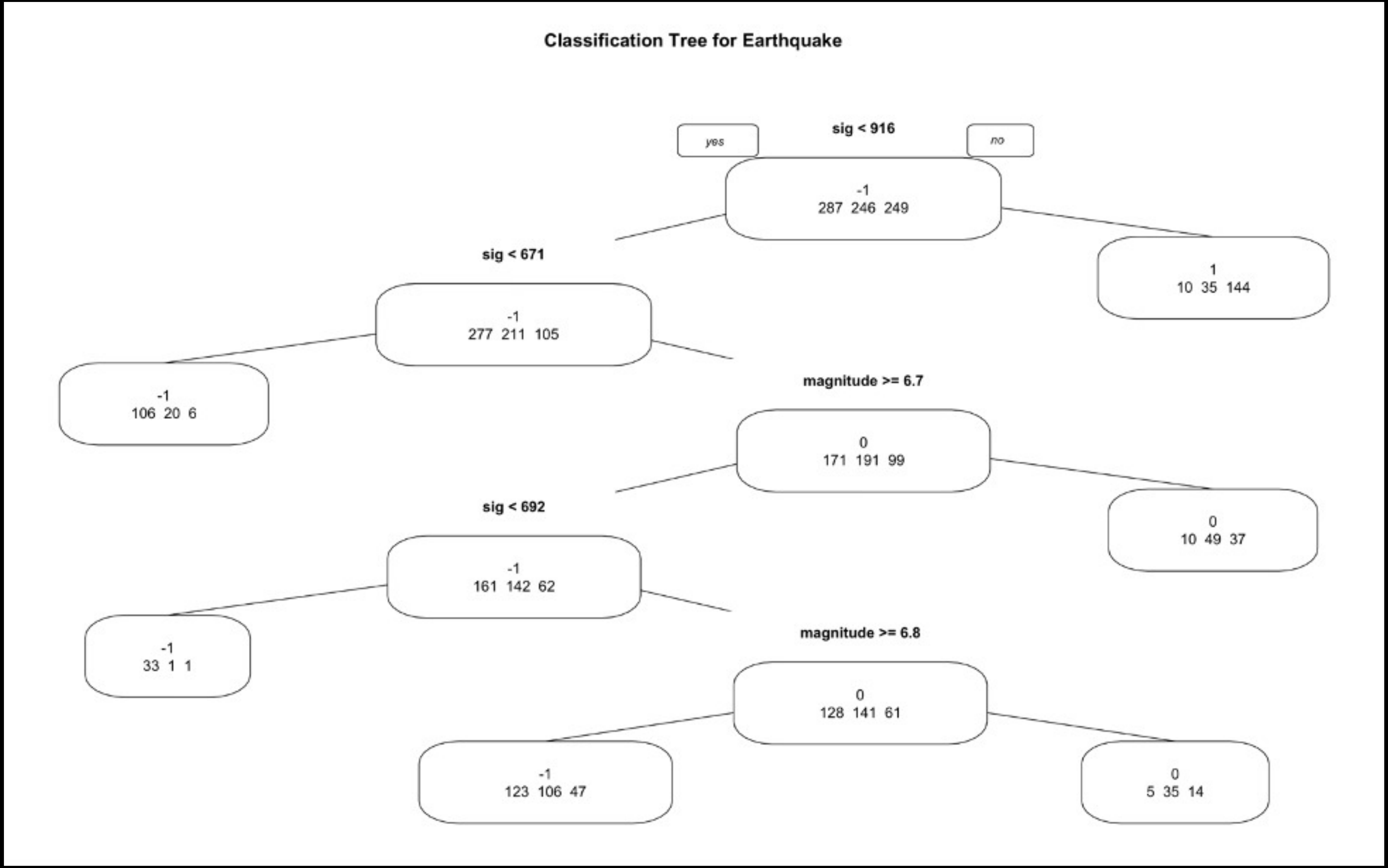
  Train data: 521 rows

  Test data: 261 rows

  METHODS USED:

- Decision Tree Model
- Logistic Regression
- K - Nearest Neighbors
- Naive Bayes Classifier
- Support Vector Machines

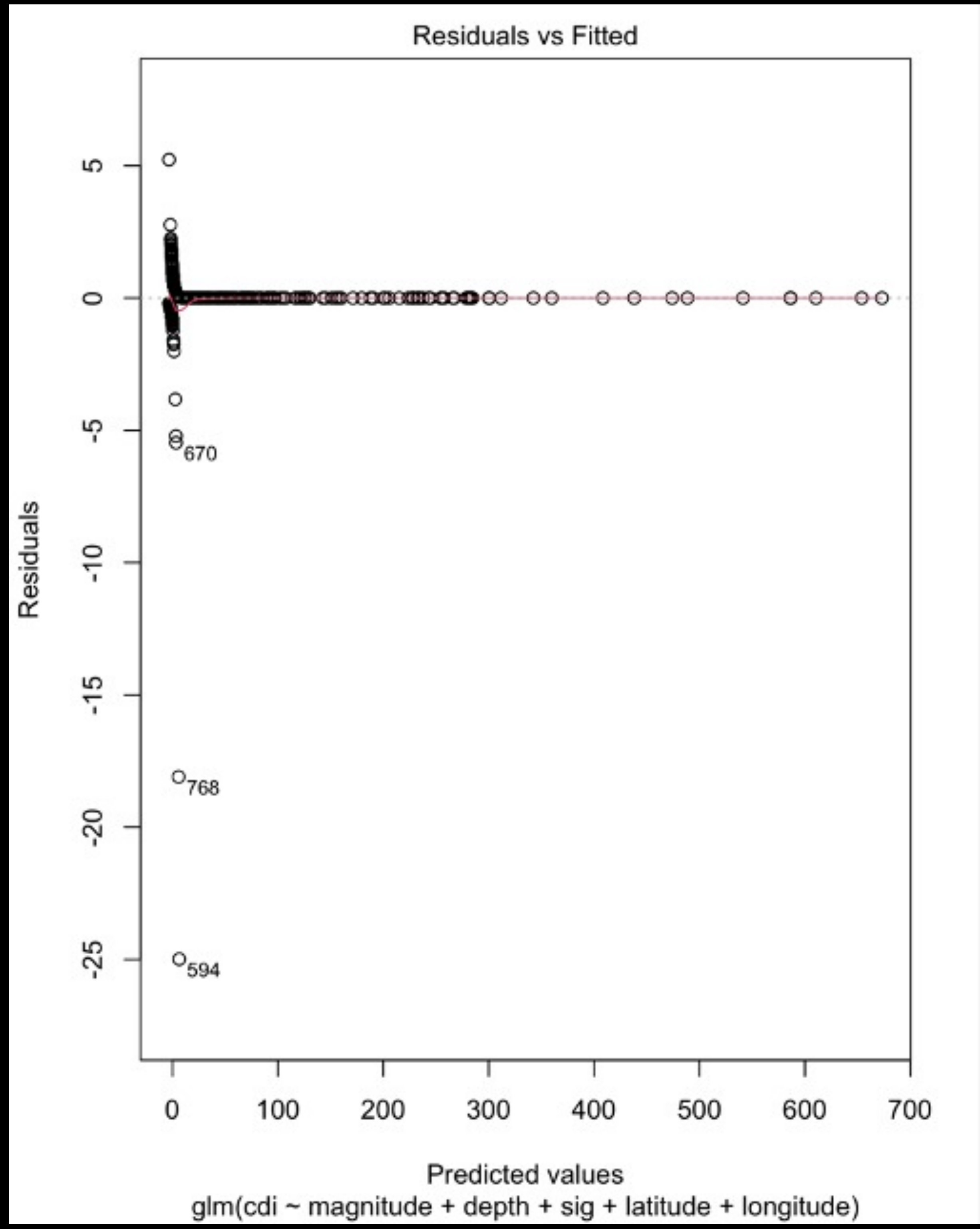
MODEL EVALUATION - DECISION TREE

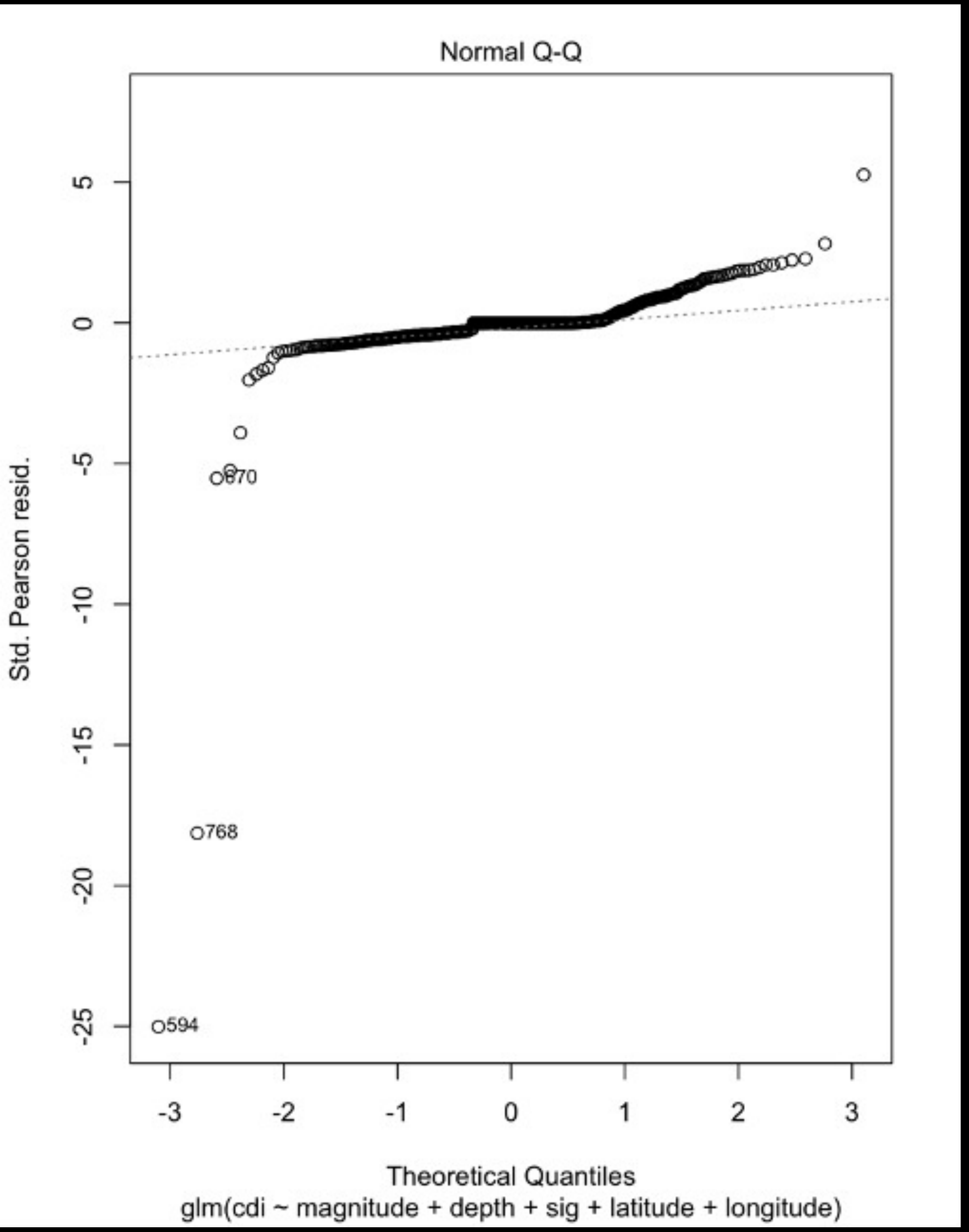


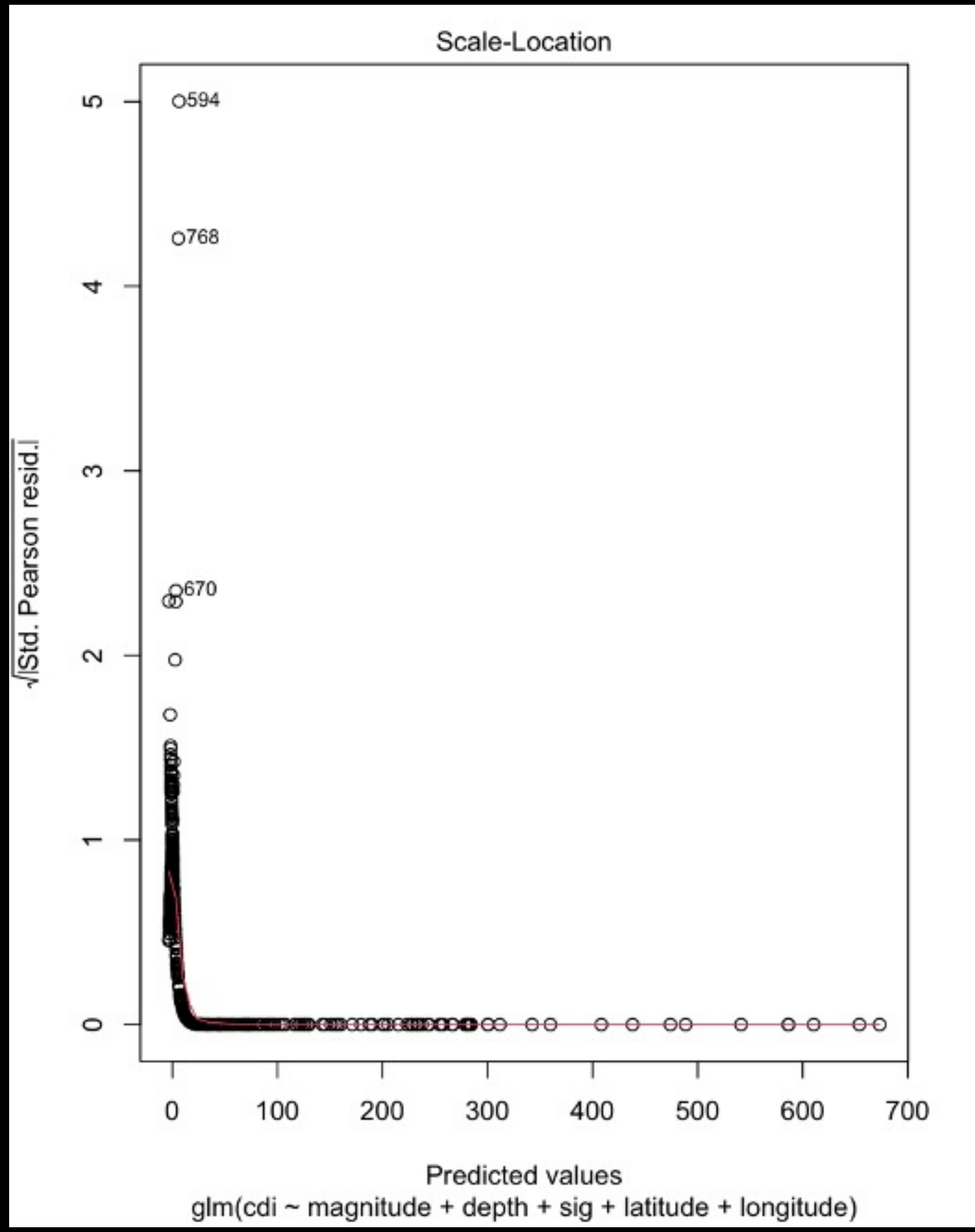
CONFUSION MATRIX FOR DECISION TREE

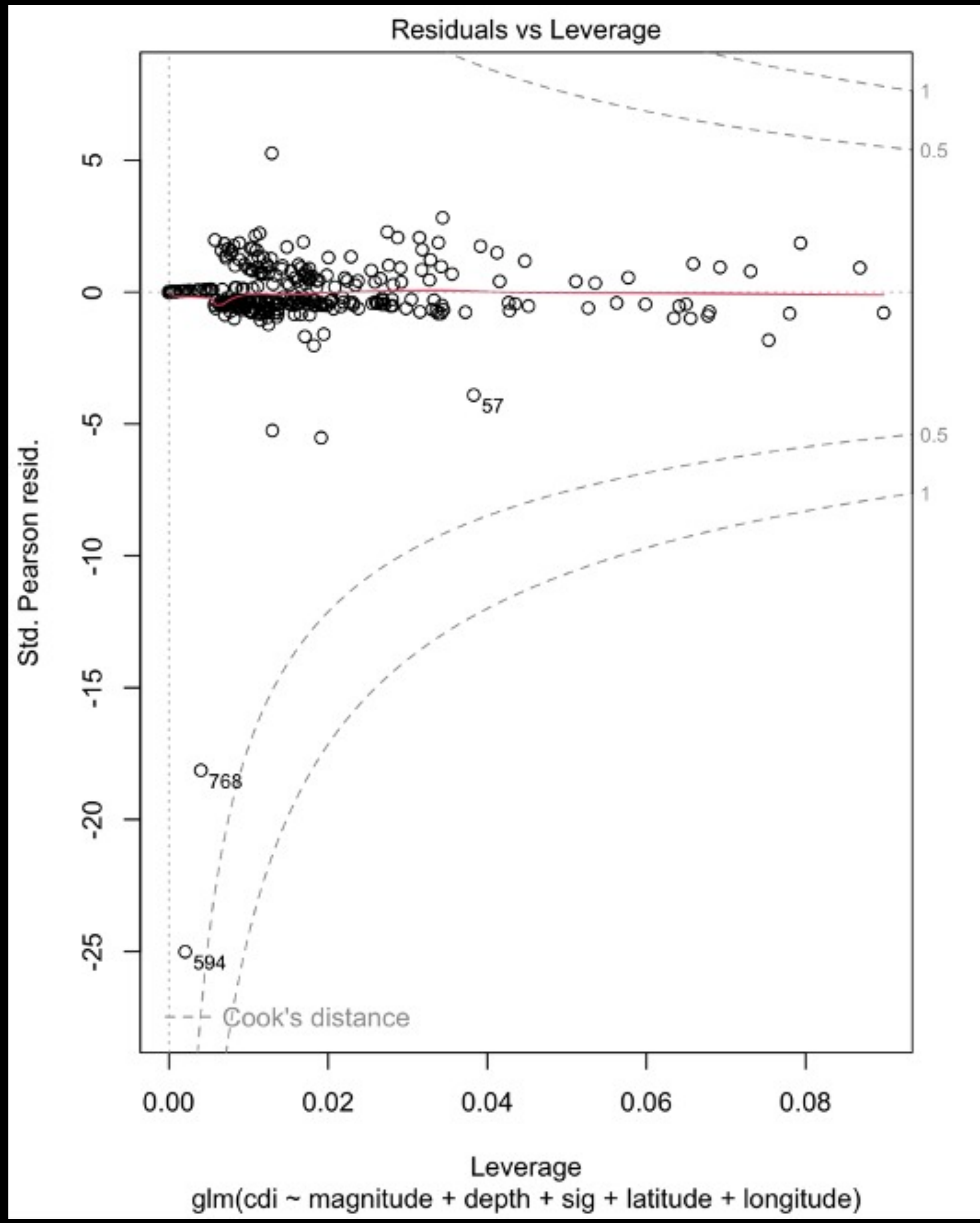
```
> # data in data.test
> data.pred <- predict(fit, data.test, type = "class")
> data.actual <- data.test$cdi
> cm1 <- confusionMatrix(data.pred, data.actual, positive = "1")
> cm1
Confusion Matrix and Statistics

          Reference
Prediction -1  0  1
        -1 87 46 15
        0   3 30 20
        1   5 14 41

Overall Statistics

               Accuracy : 0.6054
                 95% CI : (0.5432, 0.6651)
    No Information Rate : 0.364
    P-Value [Acc > NIR] : 2.204e-15

                  Kappa : 0.399

 Mcnemar's Test P-Value : 1.670e-09

Statistics by Class:

                     Class: -1 Class: 0 Class: 1
Sensitivity             0.9158   0.3333   0.5395
Specificity             0.6325   0.8655   0.8973
Pos Pred Value          0.5878   0.5660   0.6833
Neg Pred Value          0.9292   0.7115   0.8259
Prevalence              0.3640   0.3448   0.2912
Detection Rate          0.3333   0.1149   0.1571
Detection Prevalence    0.5670   0.2031   0.2299
Balanced Accuracy       0.7742   0.5994   0.7184
```

MODEL EVALUATION - LOGISTIC REGRESSION:





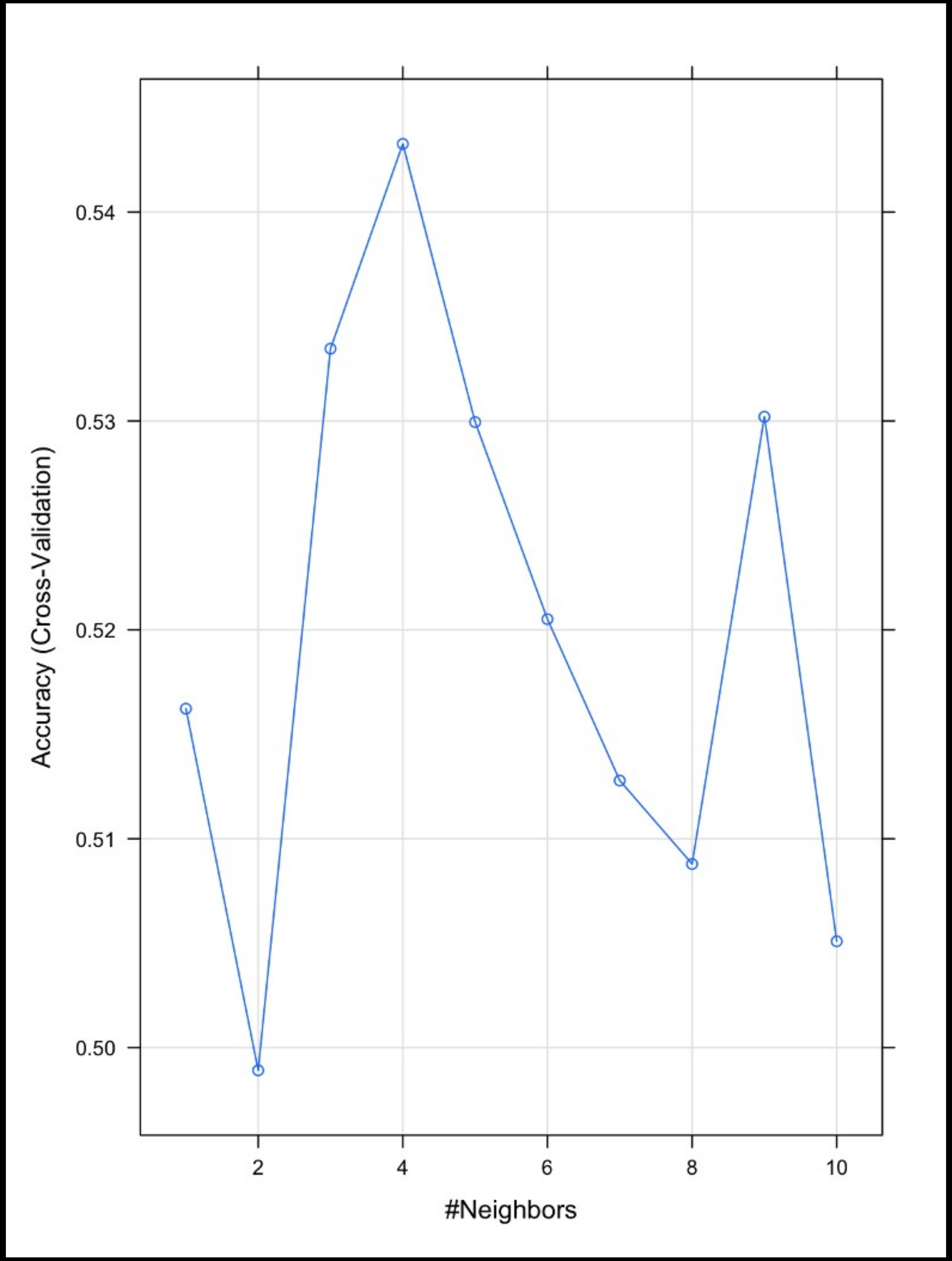




```
> # compute predicted probabilities for data.test
> logitPredict <- predict(logit.reg, data.test, type = "response")
> logitPredictClass <- cut(logitPredict, breaks = c(0.03703, 0.21527, 0.69781 ,1), labels = c("-1", "0", "1"))
>
> # evaluate classifier on data.test
> actual <- data.test$cdi
> predict <- logitPredictClass
> cm2 <- confusionMatrix(predict, actual, positive = "1")
> cm2
Confusion Matrix and Statistics

          Reference
Prediction -1  0  1
        -1 57  6  2
        0  33 23  9
        1   4 61 65

Overall Statistics

               Accuracy : 0.5577
                 95% CI : (0.495, 0.619)
    No Information Rate : 0.3615
    P-Value [Acc > NIR] : 9.698e-11

                  Kappa : 0.3466

 Mcnemar's Test P-Value : 1.582e-12

Statistics by Class:

                     Class: -1 Class: 0 Class: 1
Sensitivity             0.6064  0.25556   0.8553
Specificity             0.9518  0.75294   0.6467
Pos Pred Value          0.8769  0.35385   0.5000
Neg Pred Value          0.8103  0.65641   0.9154
Prevalence              0.3615  0.34615   0.2923
Detection Rate          0.2192  0.08846   0.2500
Detection Prevalence    0.2500  0.25000   0.5000
Balanced Accuracy       0.7791  0.50425   0.7510
```

CONFUSION MATRIX FOR LOGISTIC REGRESSION MODEL

MODEL EVALUATION - K - NEAREST NEIGHBOURS:



CONFUSION MATRIX FOR KNN MODEL

```
> # Evaluate classifier performance on testing data
> actual <- data.test$cdi
> knnPredict <- predict(knnFit, data.test)
> cm3 <- confusionMatrix(knnPredict, actual, positive = "1")
> cm3
Confusion Matrix and Statistics

          Reference
Prediction -1  0  1
        -1 65 27 21
        0  22 50 18
        1   8 13 37

Overall Statistics

               Accuracy : 0.5824
                 95% CI : (0.52, 0.6429)
    No Information Rate : 0.364
    P-Value [Acc > NIR] : 6.451e-13

                  Kappa : 0.3661

 Mcnemar's Test P-Value : 0.06744

Statistics by Class:

                     Class: -1 Class: 0 Class: 1
Sensitivity             0.6842   0.5556   0.4868
Specificity             0.7108   0.7661   0.8865
Pos Pred Value          0.5752   0.5556   0.6379
Neg Pred Value          0.7973   0.7661   0.8079
Prevalence              0.3640   0.3448   0.2912
Detection Rate          0.2490   0.1916   0.1418
Detection Prevalence    0.4330   0.3448   0.2222
Balanced Accuracy       0.6975   0.6608   0.6867
```

MODEL EVALUATION - NAIVE BAYES CLASSIFICATION

```
> # Evaluate Performance using Confusion Matrix
> actual <- data.test$cdi
> # predict class probability
> nbPredict <- predict(fit.nb, data.test, type = "raw")
> # predict class membership
> nbPredictClass <- predict(fit.nb, data.test, type = "class")
> cm4 <- confusionMatrix(nbPredictClass, actual, positive = "1")
> cm4
Confusion Matrix and Statistics

          Reference
Prediction -1  0  1
        -1 76 66 25
        0   9 13 13
        1  10 11 38

Overall Statistics

               Accuracy : 0.4866
                 95% CI : (0.4245, 0.549)
    No Information Rate : 0.364
    P-Value [Acc > NIR] : 3.370e-05

                  Kappa : 0.2162

 Mcnemar's Test P-Value : 8.328e-11

Statistics by Class:

                     Class: -1 Class: 0 Class: 1
Sensitivity             0.8000  0.14444   0.5000
Specificity             0.4518  0.87135   0.8865
Pos Pred Value          0.4551  0.37143   0.6441
Neg Pred Value          0.7979  0.65929   0.8119
Prevalence              0.3640  0.34483   0.2912
Detection Rate          0.2912  0.04981   0.1456
Detection Prevalence    0.6398  0.13410   0.2261
Balanced Accuracy       0.6259  0.50789   0.6932
```

MODEL EVALUATION - SUPPORT VECTOR MACHINE

```
> # Make predictions on the test set
> pred <- predict(model, data.test)
> actual <- data.test$cdi
> cm5 <- confusionMatrix(pred, actual, positive = "1")
> cm5
Confusion Matrix and Statistics

          Reference
Prediction -1  0  1
        -1 94 58 31
        0   0 21  9
        1   1 11 36

Overall Statistics

               Accuracy : 0.5785
                 95% CI : (0.5161, 0.6392)
    No Information Rate : 0.364
    P-Value [Acc > NIR] : 1.574e-12

                  Kappa : 0.3532

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: -1 Class: 0 Class: 1
Sensitivity             0.9895  0.23333   0.4737
Specificity             0.4639  0.94737   0.9351
Pos Pred Value          0.5137  0.70000   0.7500
Neg Pred Value          0.9872  0.70130   0.8122
Prevalence              0.3640  0.34483   0.2912
Detection Rate          0.3602  0.08046   0.1379
Detection Prevalence    0.7011  0.11494   0.1839
Balanced Accuracy       0.7267  0.59035   0.7044
```

RESULTS: COMPARISON ACROSS DIFFERENT METHODS CONSIDERING -1 AS THE "NEGATIVE" CLASS

```
> ##### compare across different methods ####
> # compare across different methods (considering Class: -1 as the "negative" class)
> result1 <- rbind(cm1$byClass[1, c("Sensitivity", "Specificity", "Balanced Accuracy")],
+                  cm2$byClass[1, c("Sensitivity", "Specificity", "Balanced Accuracy")],
+                  cm3$byClass[1, c("Sensitivity", "Specificity", "Balanced Accuracy")],
+                  cm4$byClass[1, c("Sensitivity", "Specificity", "Balanced Accuracy")],
+                  cm5$byClass[1, c("Sensitivity", "Specificity", "Balanced Accuracy")])
> row.names(result1) <- c("Decision Tree", "Logistic Reg", "KNN", "Naive Bayes", "SVM")
> result1
              Sensitivity Specificity Balanced Accuracy
Decision Tree   0.9157895   0.6325301         0.7741598
Logistic Reg    0.6063830   0.9518072         0.7790951
KNN             0.6526316   0.7168675         0.6847495
Naive Bayes     0.8000000   0.4518072         0.6259036
SVM             0.9894737   0.4638554         0.7266646
```

CONSIDERING 0 AS THE “NETURAL” CLASS

```
> # compare across different methods (considering Class: 0 as the "neutral" class)
> result2 <- rbind(cm1$byClass[2, c("Sensitivity", "Specificity", "Balanced Accuracy")],
+                  cm2$byClass[2, c("Sensitivity", "Specificity", "Balanced Accuracy")],
+                  cm3$byClass[2, c("Sensitivity", "Specificity", "Balanced Accuracy")],
+                  cm4$byClass[2, c("Sensitivity", "Specificity", "Balanced Accuracy")],
+                  cm5$byClass[2, c("Sensitivity", "Specificity", "Balanced Accuracy")])
> row.names(result2) <- c("Decision Tree", "Logistic Reg", "KNN", "Naive Bayes", "SVM")
> result2
              Sensitivity Specificity Balanced Accuracy
Decision Tree   0.3333333   0.8654971         0.5994152
Logistic Reg    0.2555556   0.7529412         0.5042484
KNN             0.5000000   0.7543860         0.6271930
Naive Bayes     0.1444444   0.8713450         0.5078947
SVM             0.2333333   0.9473684         0.5903509
```

CONSIDERING 1 AS THE "POSITIVE" CLASS:

```
> # compare across different methods (considering Class: 1 as the "positive" class)
> result3 <- rbind(cm1$byClass[3, c("Sensitivity", "Specificity", "Balanced Accuracy")],
+                  cm2$byClass[3, c("Sensitivity", "Specificity", "Balanced Accuracy")],
+                  cm3$byClass[3, c("Sensitivity", "Specificity", "Balanced Accuracy")],
+                  cm4$byClass[3, c("Sensitivity", "Specificity", "Balanced Accuracy")],
+                  cm5$byClass[3, c("Sensitivity", "Specificity", "Balanced Accuracy")])
> row.names(result3) <- c("Decision Tree", "Logistic Reg", "KNN", "Naive Bayes", "SVM")
> result3
              Sensitivity Specificity Balanced Accuracy
Decision Tree   0.5394737   0.8972973         0.7183855
Logistic Reg    0.8552632   0.6467391         0.7510011
KNN             0.5131579   0.8594595         0.6863087
Naive Bayes     0.5000000   0.8864865         0.6932432
SVM             0.4736842   0.9351351         0.7044097
```

ADVANTAGES

- Improved understanding of earthquake risks - better prioritize resource allocation
- More accurate and reliable earthquake prediction - save lives and reduce damage
- Importance of multiple metrics in performance evaluation - comprehensive understanding

IMPLICATIONS

- More targeted and effective strategies for disaster preparedness and response
- More accurate and reliable earthquake prediction models - global scale.
- Improve building codes and safety regulations - reducing the impact of earthquakes more effective prevention and mitigation strategies.
- Facilitate international collaboration on disaster response.

FUTURE SCOPE

- **Early Warning System**
- **Evacuation Planning**
- **Resource Allocation**
- **Infrastructure Development**
- **Insurance Planning**
- **Education and Awareness**


THANK YOU