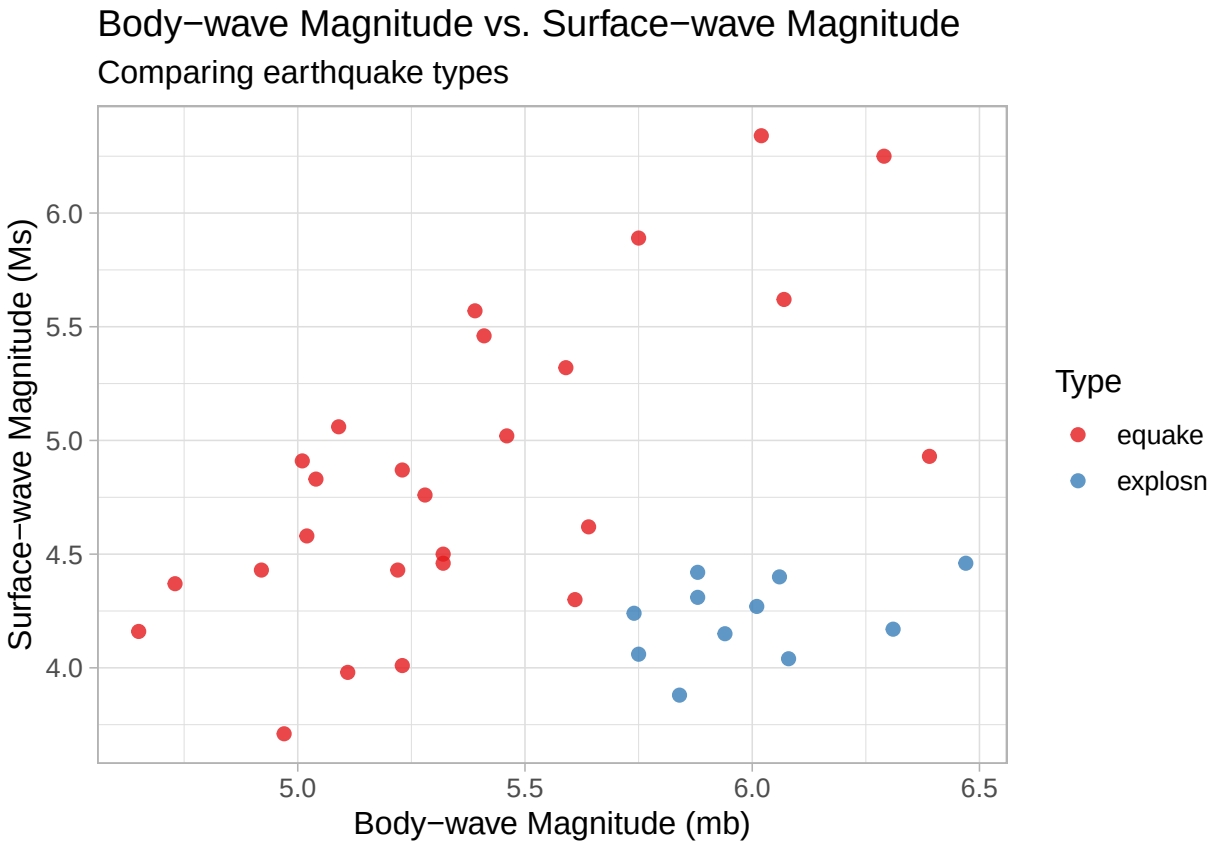
# Report_10883408

10883408

2024-04-08

## ML part A




**Graphical Analysis:**

The scatter plot created by this code would show each earthquake or explosion event as a point in the space defined by its body-wave and surface-wave magnitudes. The colors distinguish between different types of seismic events, which could be crucial for identifying patterns or clusters specific to earthquakes versus explosions.

**Clustering: If there are visible clusters or distinct areas predominantly occupied by one type of event, this could indicate that these two features (body-wave and surface-wave magnitudes) are effective for distinguishing between earthquakes and explosions.**

Overlaps: Significant overlapping of colors might suggest that the two features alone are not sufficient to distinguish between the event types without additional information or more complex modeling.

**Contextual Relevance:**

In the context of monitoring for unauthorized nuclear tests, this visualization helps in quickly assessing whether there are clear, distinguishable patterns in seismic readings that could indicate nuclear activities. Effective differentiation between natural seismic events (earthquakes) and man-made seismic events (nuclear explosions) is crucial for global security and monitoring compliance with international treaties such as the Comprehensive Nuclear-Test-Ban Treaty (CTBT).

**Numerical Summaries:**

While the provided code focuses on visual analysis, numerical summaries (like the mean, median, variance of mb and Ms for each type) would complement this by quantifying the central tendencies and dispersions. This could further aid in understanding how significantly the magnitudes for each type differ statistically.

**Justification:**

The choice of a scatter plot is justified as it allows stakeholders to visually parse the relationship between two continuous variables across categories. Given the high stakes involved in nuclear monitoring, quick visual assessments alongside rigorous statistical analysis are imperative. The plot facilitates this by providing a clear, immediate visual summary of the data as per the described features.


**Explanation of the Plot:**

The box plot with jitter provides a clear view of how the body-wave magnitudes are distributed within each type of seismic event. The box plot component shows the median (the middle line in the box), the 25th and 75th percentiles (the lower and upper hinges of the box), and potential outliers (points that fall outside 1.5 times the IQR from the hinges). The jittered points represent individual data points, offering a granular look at the data distribution and any potential anomalies or outliers.

**Justification of the Statements:**

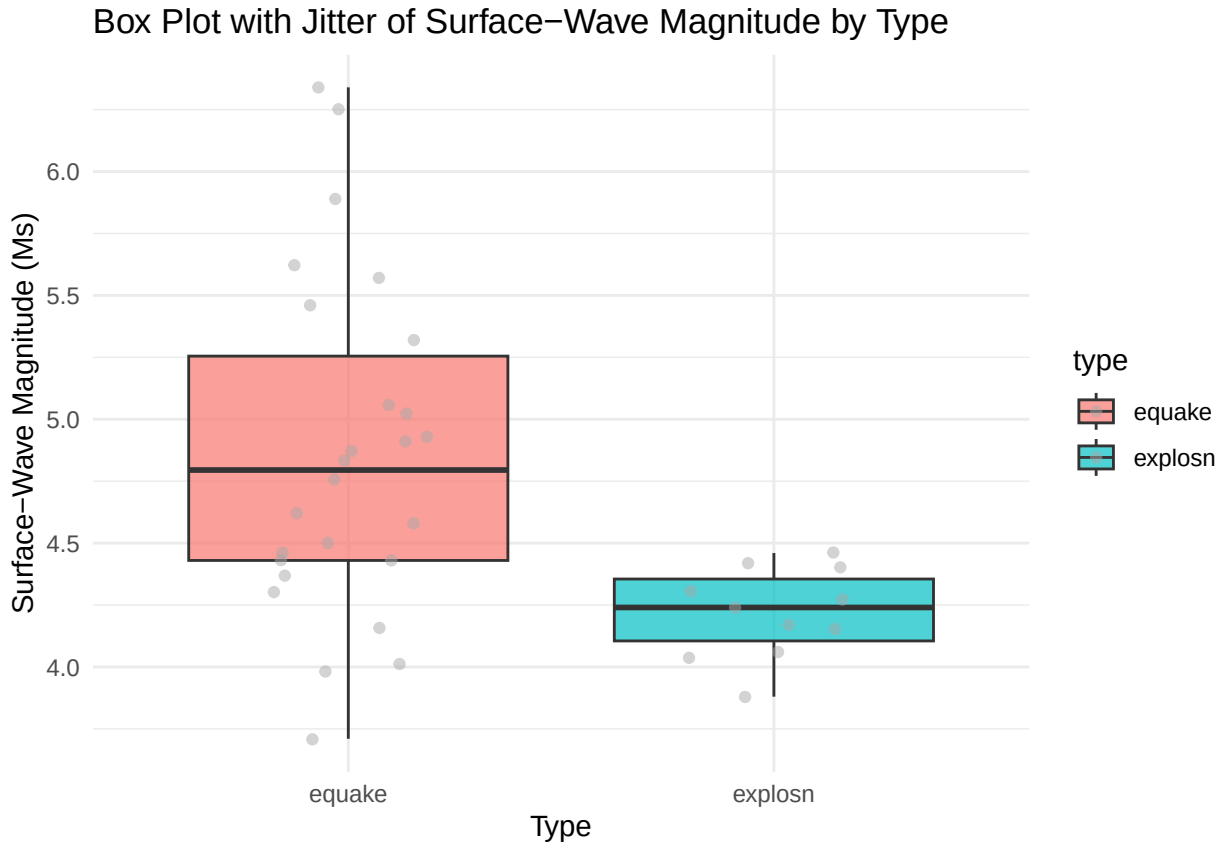
Distribution Insight: The plot allows stakeholders to quickly assess whether there are significant differences in the body-wave magnitudes between different types of seismic events. For example, if one type of event typically has higher magnitude readings, this could indicate a different energy release characteristic.

Outlier Detection: By visually representing both the summary statistics and the actual data points, the plot helps identify outliers or unusual observations that may warrant further investigation.

Decision Making: This type of visualization supports decision-making in seismology and geophysics by providing a straightforward way to compare seismic event types. This could be crucial for designing monitoring systems or for academic studies in seismology.

Effective Communication: The plot serves as an effective communication tool by presenting complex statistical data in a form that is easy to understand, even for those without deep statistical knowledge.

This combination of box plot and jitter plot is particularly effective in contexts where understanding the variability within and across categories is crucial. It is a powerful tool for exploratory data analysis, providing both an overview and a detailed look at the data distribution across different categories.




Justification of the Statements: Visualization of Variability: This plot is effective for visually assessing the variability and central tendencies of surface-wave magnitudes across different types of seismic events. The box plot provides a summary view, while the jittered points give detailed insight into the individual data entries.

Comparative Analysis: By displaying data for different types side by side, the plot facilitates direct comparisons between groups. For example, one might observe that nuclear explosions tend to have a tighter range of surface-wave magnitudes compared to earthquakes, which might exhibit a broader spread and potentially higher medians.

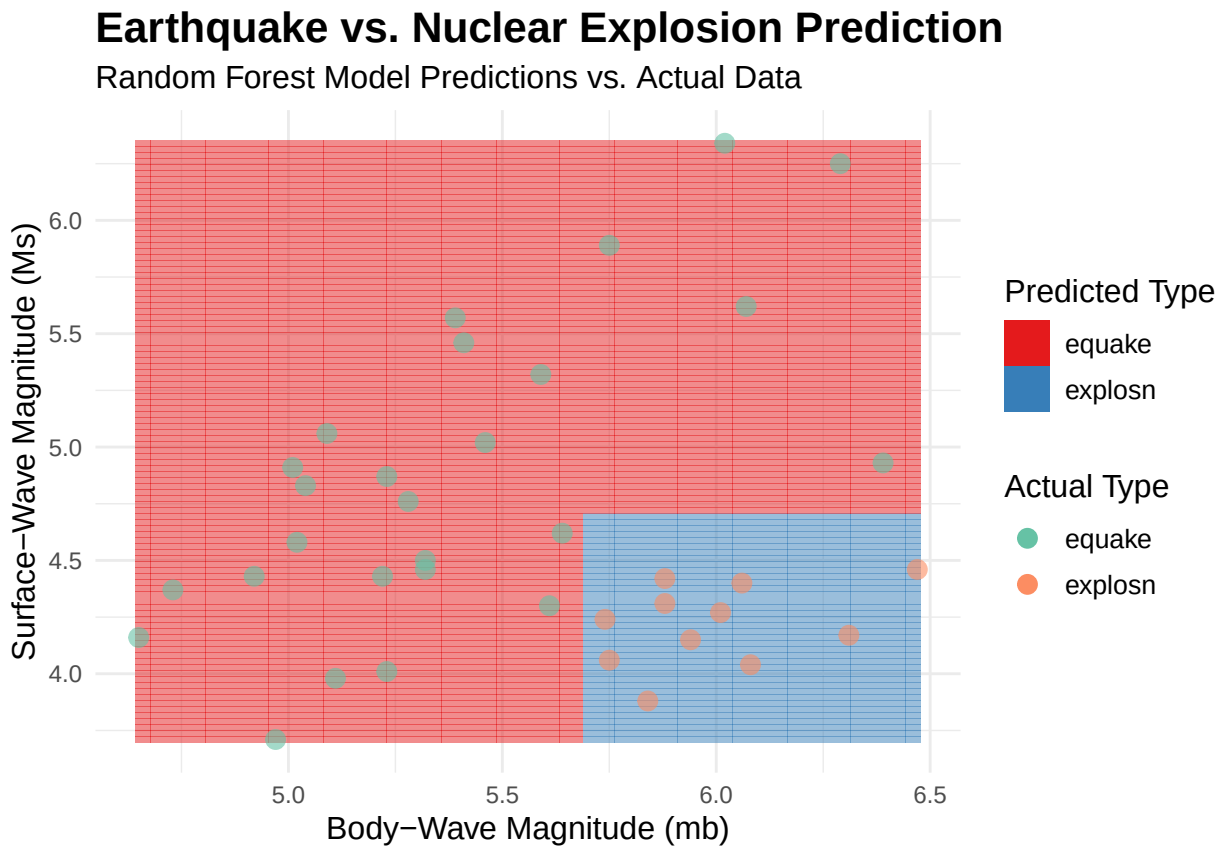
Outlier Detection: The visualization makes it easy to spot any anomalies or outliers in the data, which could indicate measurement errors or exceptional cases that may require further investigation.

Informative and Accessible: The addition of a clear title, axis labels, and a legend makes the plot accessible even to those who may not be familiar with the data or statistical plots, enabling stakeholders or a general audience to understand the findings at a glance.

Contextual Relevance: In the context of monitoring seismic activities, being able to distinguish between different types of events based on surface-wave magnitudes is crucial. Such plots not only aid in preliminary analyses but can also be instrumental in developing more sophisticated models or algorithms to automate the detection and classification of seismic events. This is particularly important in scenarios where quick decision-making is necessary, such as in early warning systems or nuclear treaty compliance monitoring.

In summary, the plot generated by this R code effectively uses graphical elements to present key statistical insights into seismic data, aiding in both detailed statistical analysis and high-level data overview. This approach is justified given its utility in exploratory data analysis, where understanding data distribution and anomalies is crucial.

## ML part B

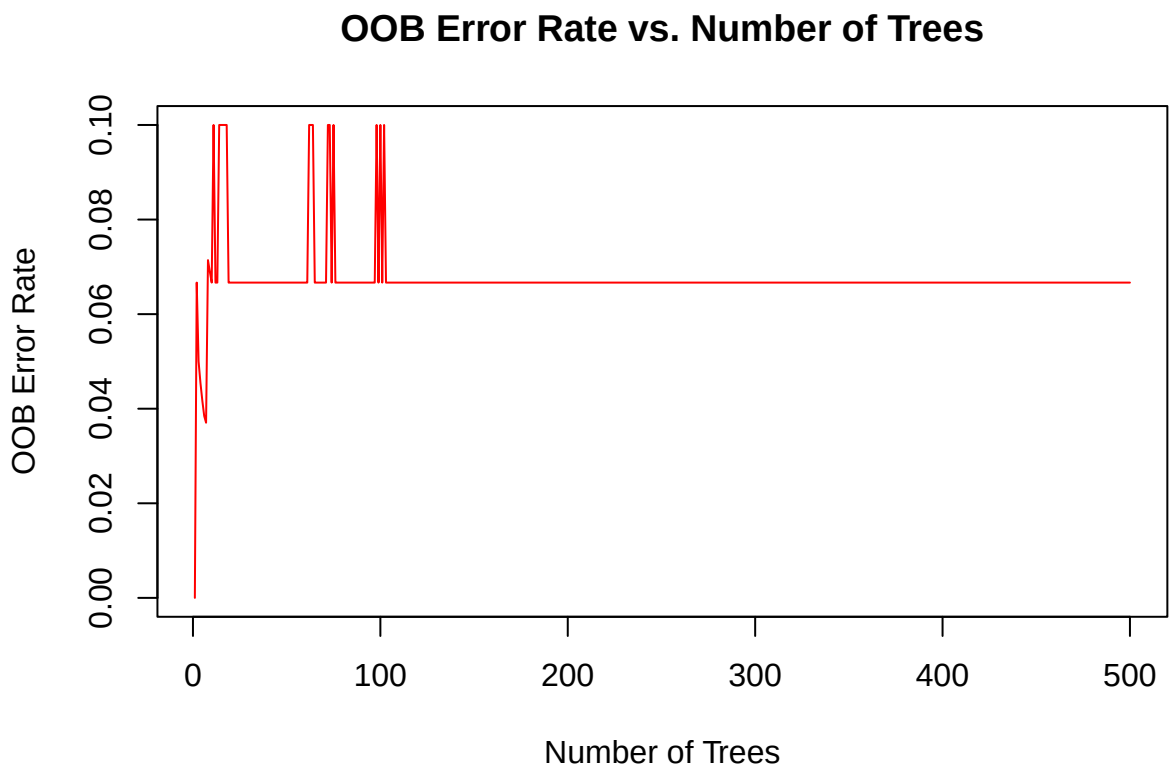



**error rate**

```
# Extract the Random Forest model from the caret object
rf <- rfModel$finalModel

# Error rate plot
plot(rf$err.rate[, "OOB"], type = "l", col = "red",
```

```
     xlab = "Number of Trees",
     ylab = "OOB Error Rate",
     main = "OOB Error Rate vs. Number of Trees")
```




Model Tuning ### Random Forest Training: Library: The randomForest package in R is used for constructing the Random Forest model. Data Partitioning: The dataset is divided into training (80%) and testing (20%) subsets using the createDataPartition function from the caret package, ensuring that the distribution of event types is balanced across both sets. Cross-validation: The model's hyperparameters are tuned using 5-fold cross-validation. This method splits the training set into five smaller sets; the model is trained on four of these while validating on the fifth, cycling through each subset. This process helps in optimizing model parameters while avoiding overfitting. Tuning the mtry Parameter: The mtry parameter, which determines the number of variables considered at each split in the tree, is tuned. Two values are tested: 1 and 2. This parameter significantly affects the model's performance and helps in finding a good balance between bias and variance.

**Model Visualization**

Decision Surface Visualization: Grid Creation: A grid is created over the range of body and surface magnitudes found in the dataset. This grid allows for a comprehensive visualization of how the model classifies different combinations of these two features. Prediction on Grid: The Random Forest model predicts the type of event for each point on the grid, essentially plotting the decision boundaries between earthquake and explosion classifications. Plotting: Using ggplot2, the decision surface is visualized. The predictions for the grid are shown as a color-filled background (geom_tile), and actual data points are overlaid (geom_point) to show how well the predictions match reality. This visualization helps in understanding the model's classification rules visually. Model Evaluation ### Error Rate Computation: Out-of-Bag (OOB) Error: The OOB error rate is plotted against the number of trees in the forest. OOB error is a method of measuring prediction error of random forests, decision trees, and other models utilizing bootstrap aggregating to sub-

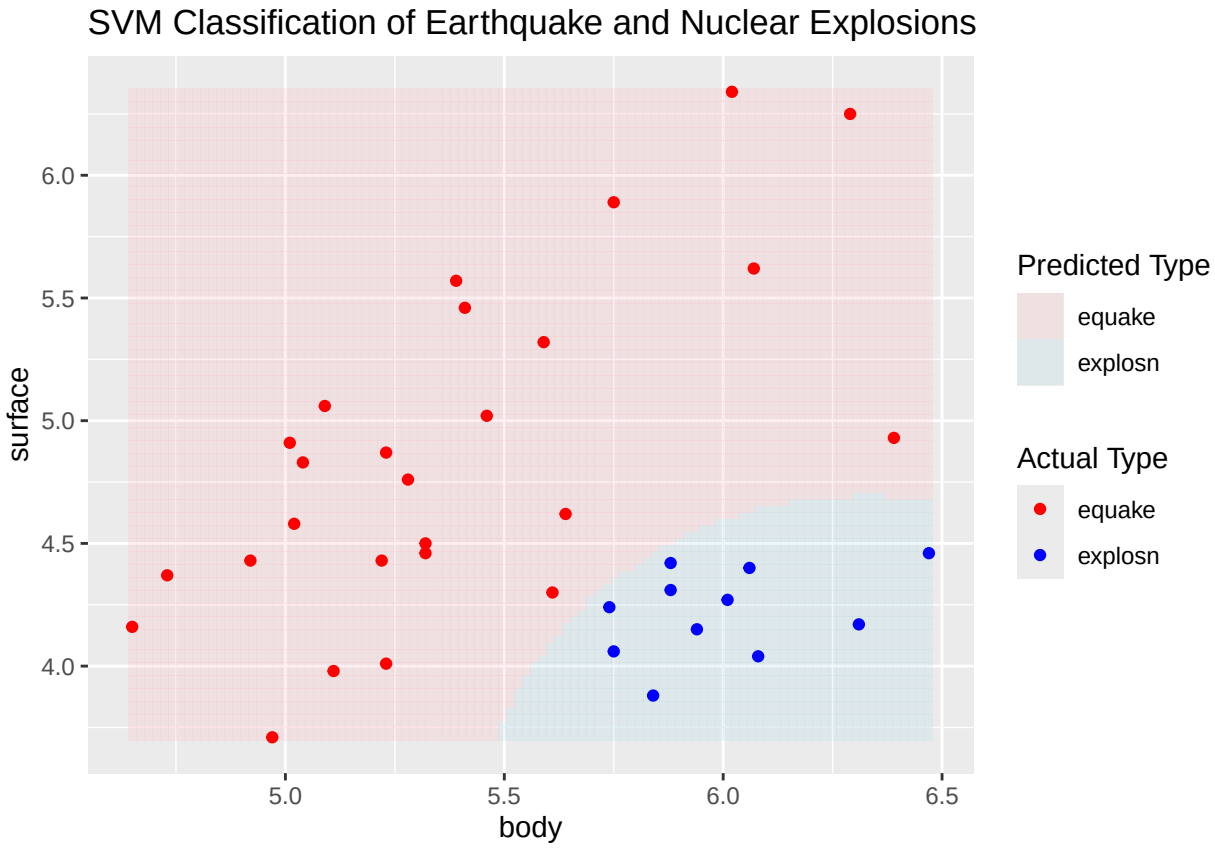
sample data sampled for the training process. OOB is the mean prediction error on each training sample x , using only the trees that did not have x in their bootstrap sample. Plot Details: The plot provides insight into how the model's accuracy evolves as more trees are added to the forest. Ideally, as the number of trees increases, the OOB error rate should decrease and then stabilize, indicating the model is neither underfitting nor overfitting.

### Additional Evaluation with Leave-One-Out Cross-Validation (LOOCV)

While not directly shown in the code snippets you provided, leave-one-out cross-validation (LOOCV) could be another method for evaluating the model. In LOOCV, the model is trained on all data points except one, which is used as the test set. This is repeated such that each data point serves as the test set exactly once. LOOCV provides a robust estimate of the model's performance but can be computationally intensive for larger datasets. It would be especially useful here to confirm the model's effectiveness due to the small size of the dataset.

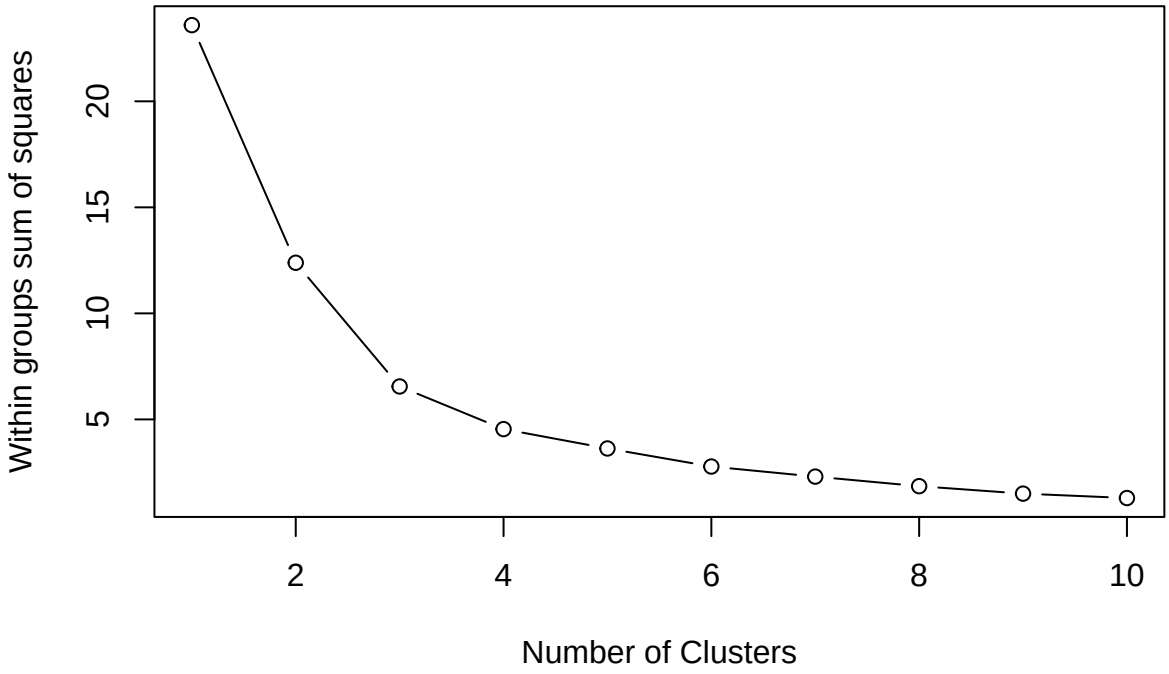
### Summary

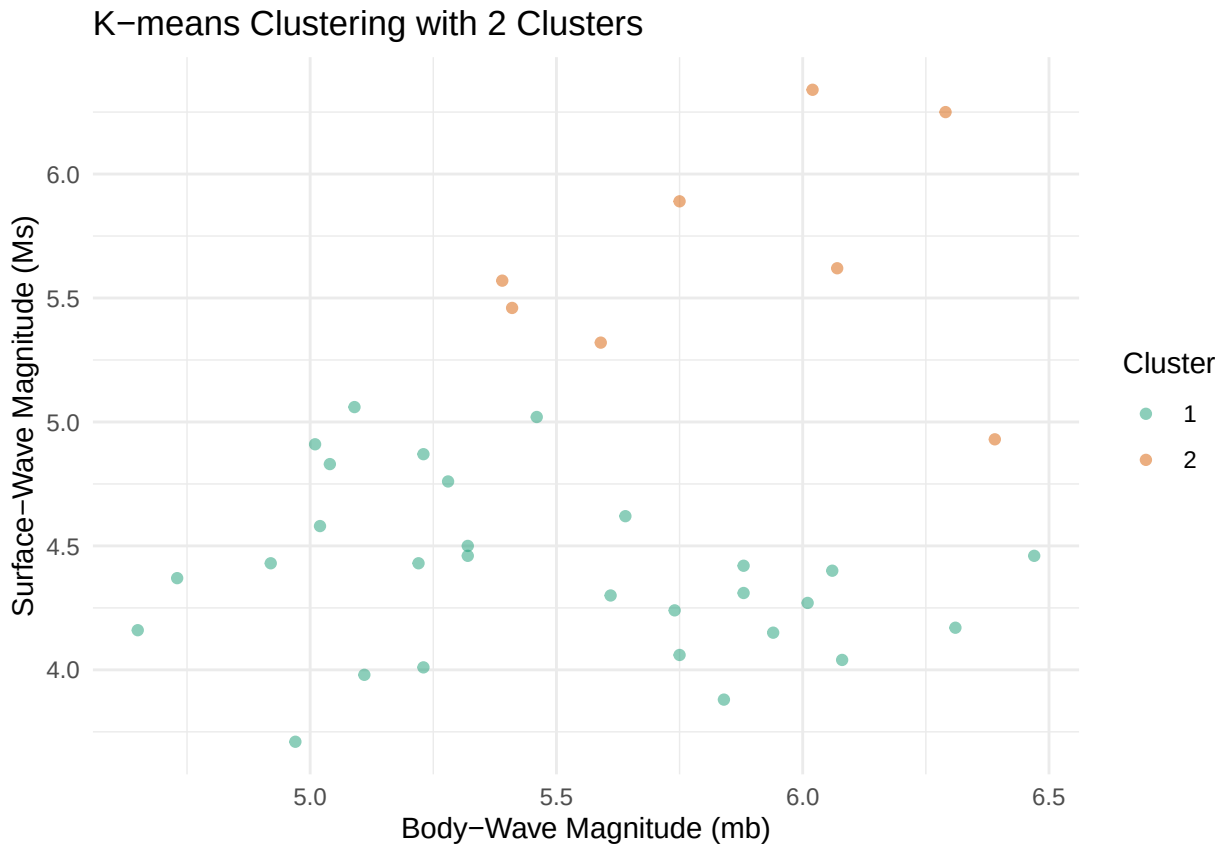
The combination of model tuning, visualization, and evaluation techniques provides a comprehensive approach to understanding and validating the Random Forest model. This methodical approach ensures that the model is both accurate and generalizable, capable of distinguishing between different types of seismic events effectively. The visualization of the decision boundary offers an intuitive way to interpret the model's performance, while the error rate plot and potential LOOCV provide quantitative measures of model accuracy.

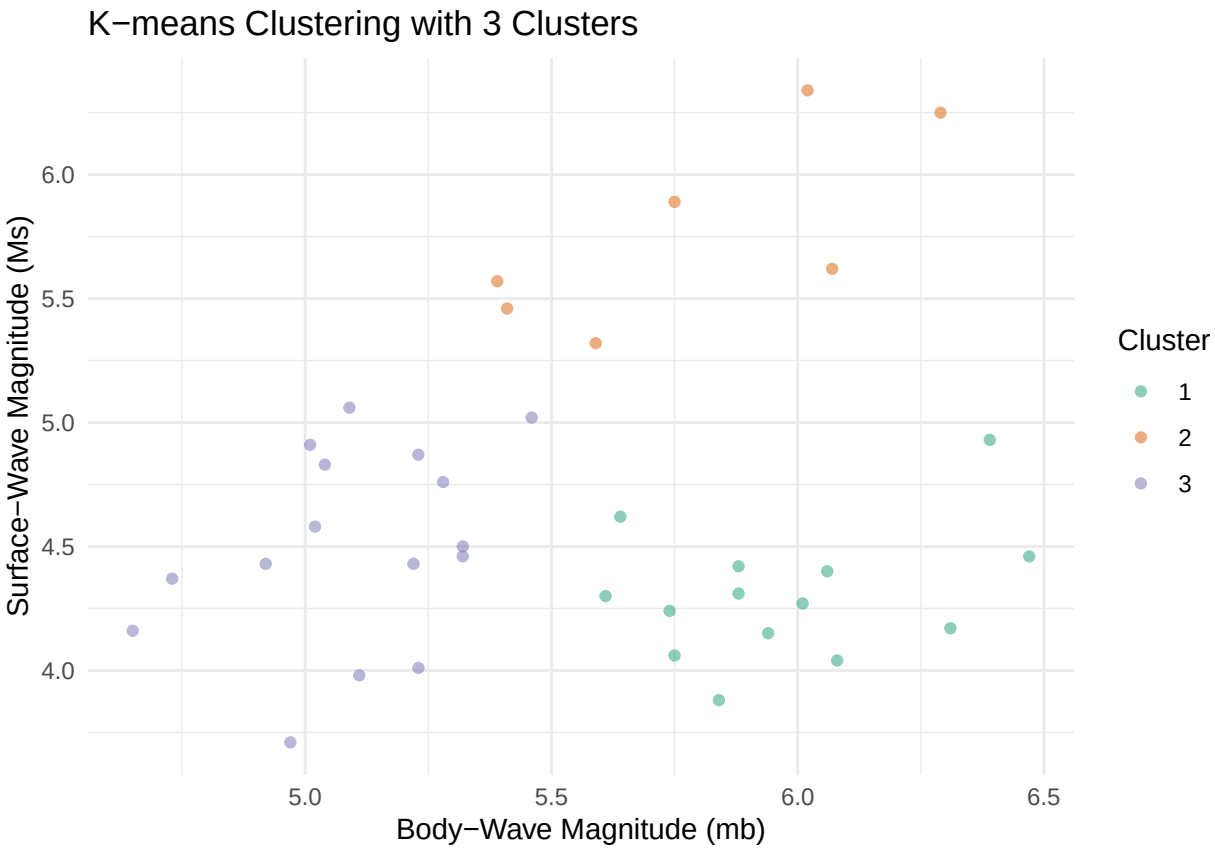


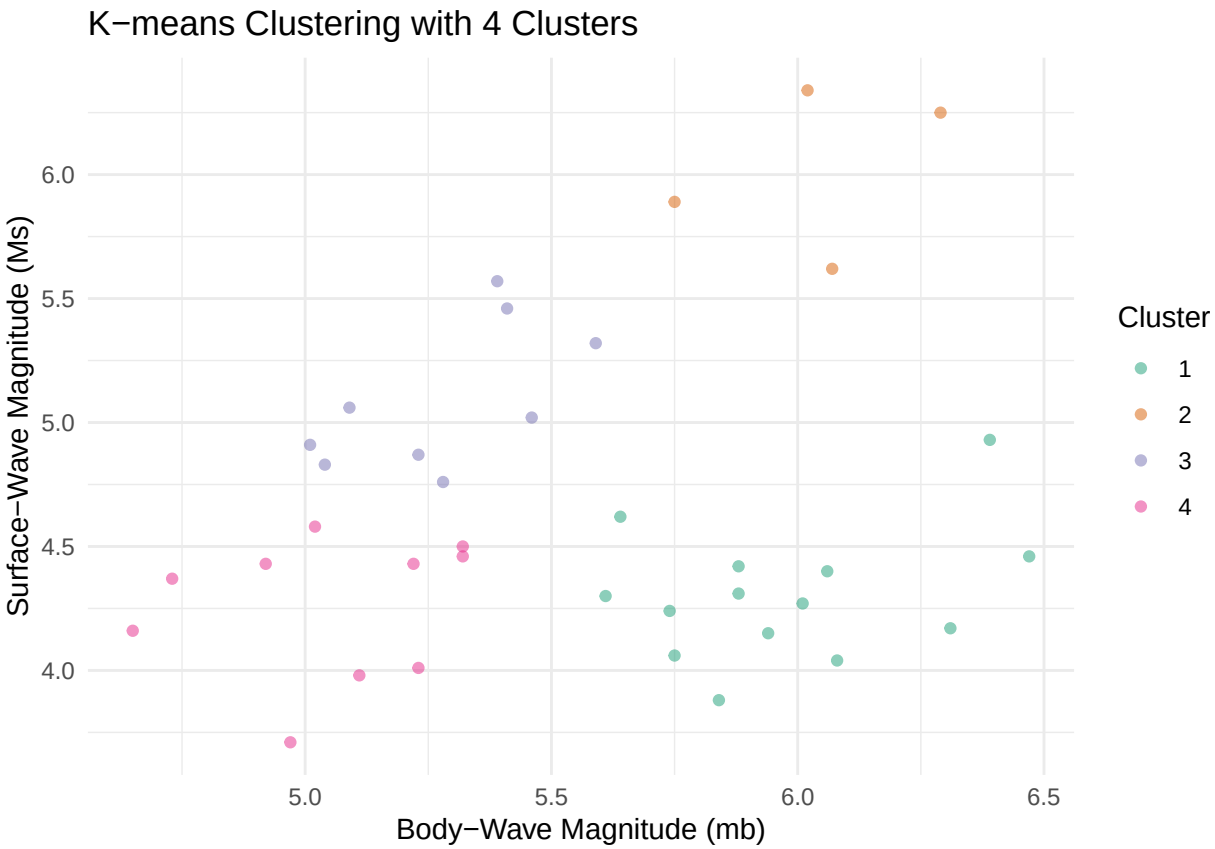

```
##   sigma   C
## 8   0.1 100
```

## Part D


Within groups sum of squares
5
10
15
20
2
4
6
8
10
Number of Clusters


K−means Clustering with 2 Clusters
Surface−Wave Magnitude (Ms)
6.0
5.5
5.0
4.5
4.0
5.0
5.5
6.0
6.5
Body−Wave Magnitude (mb)
Cluster
1
2


K−means Clustering with 3 Clusters
Surface−Wave Magnitude (Ms)
6.0
5.5
5.0
4.5
4.0
5.0
5.5
6.0
6.5
Body−Wave Magnitude (mb)
Cluster
1
2
3


K−means Clustering with 4 Clusters
Surface−Wave Magnitude (Ms)
6.0
5.5
5.0
4.5
4.0
5.0
5.5
6.0
6.5
Body−Wave Magnitude (mb)
Cluster
1
2
3
4