Mushroom Classification Analysis

Table of contents

[Load and Inspect Dataset 1](#_Toc193812604)

[Overview 1](#_Toc193812605)

[Filter and Prepare Data 2](#_Toc193812606)

[Univariable Logistic Regression 2](#_Toc193812607)

[cap-diameter 3](#_Toc193812608)

[cap-shape 3](#_Toc193812609)

[cap-color 4](#_Toc193812610)

[does-bruise-or-bleed 4](#_Toc193812611)

[gill-attachment 5](#_Toc193812612)

[gill-color 5](#_Toc193812613)

[stem-height 6](#_Toc193812614)

[stem-width 6](#_Toc193812615)

[stem-color 7](#_Toc193812616)

[habitat 7](#_Toc193812617)

[season 8](#_Toc193812618)

[Multivariable Logistic Regression (Full Model) 8](#_Toc193812619)

[Exploration of Collinearity 8](#_Toc193812620)

[create forest plot 10](#_Toc193812621)

[Model Evaluation: ROC Curve and Classification Quality 13](#_Toc193812622)

## 

# Load and Inspect Dataset

## Overview

This project explores a dataset of mushrooms, classifying whether they are **edible or poisonous** based on their physical characteristics. We use **logistic regression** and **exploratory data analysis** to model the relationship between mushroom features and toxicity.

The dataset was originally sourced from [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/Mushroom) and contains **categorical variables** only.

Note: All features are categorical. The target variable is class — whether the mushroom is **edible (e)** or **poisonous (p)**.

# Filter and Prepare Data

Number of valid cases

| name | edible | poisonous |
| --- | --- | --- |
| gill-spacing | 16,238 | 19,768 |
| cap-surface | 20,474 | 26,475 |
| gill-attachment | 23,651 | 27,534 |
| ring-type | 25,769 | 32,829 |
| cap-diameter | 27,181 | 33,888 |
| cap-shape | 27,181 | 33,888 |
| cap-color | 27,181 | 33,888 |
| does-bruise-or-bleed | 27,181 | 33,888 |
| gill-color | 27,181 | 33,888 |
| stem-height | 27,181 | 33,888 |
| stem-width | 27,181 | 33,888 |
| stem-color | 27,181 | 33,888 |
| has-ring | 27,181 | 33,888 |
| habitat | 27,181 | 33,888 |
| season | 27,181 | 33,888 |

class cap-diameter cap-shape cap-color does-bruise-or-bleed gill-attachment gill-color stem-height stem-width stem-color habitat season

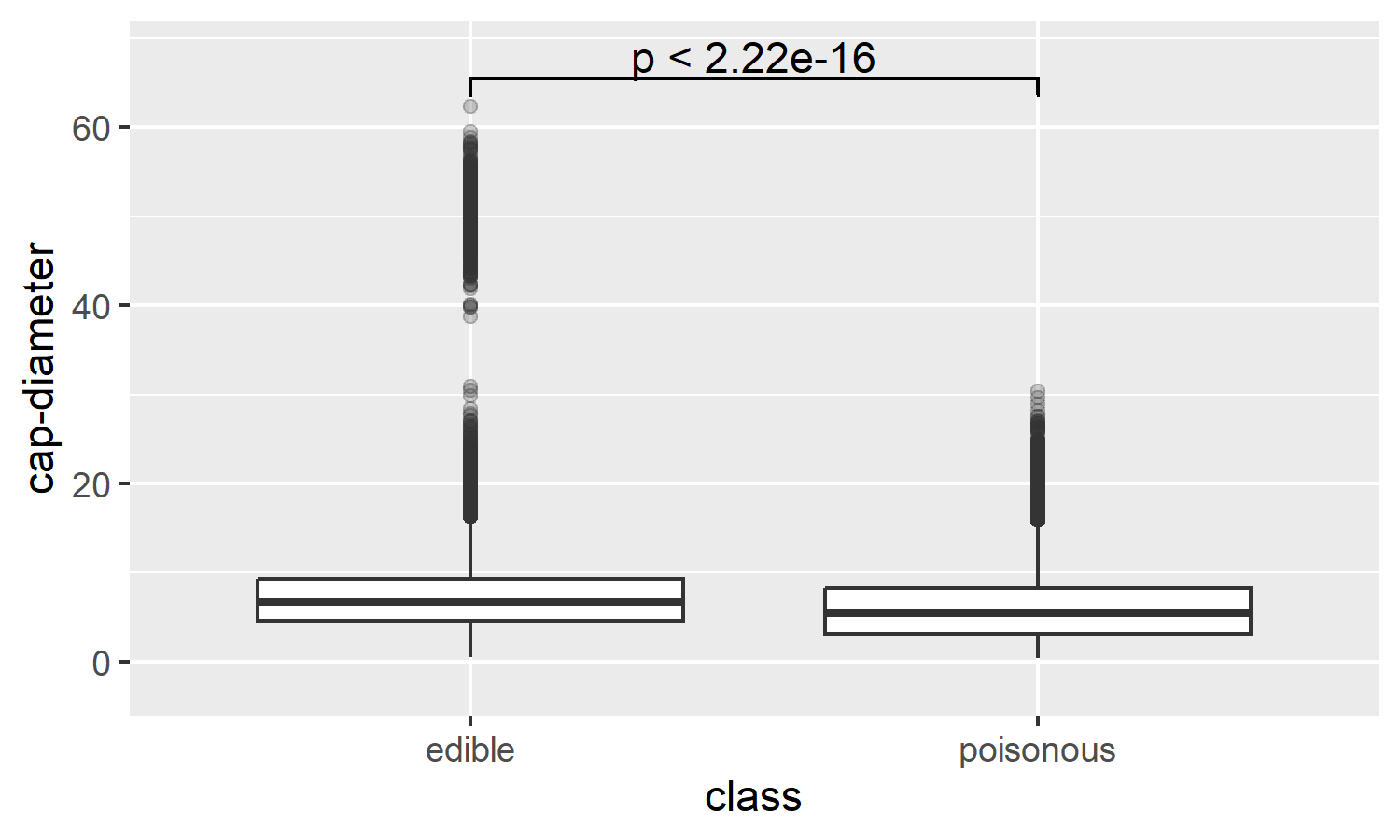
# Univariable Logistic Regression

Each predictor variable is individually tested to assess its relationship with the outcome variable, class (edible vs. poisonous), using logistic regression.

In this section, we:

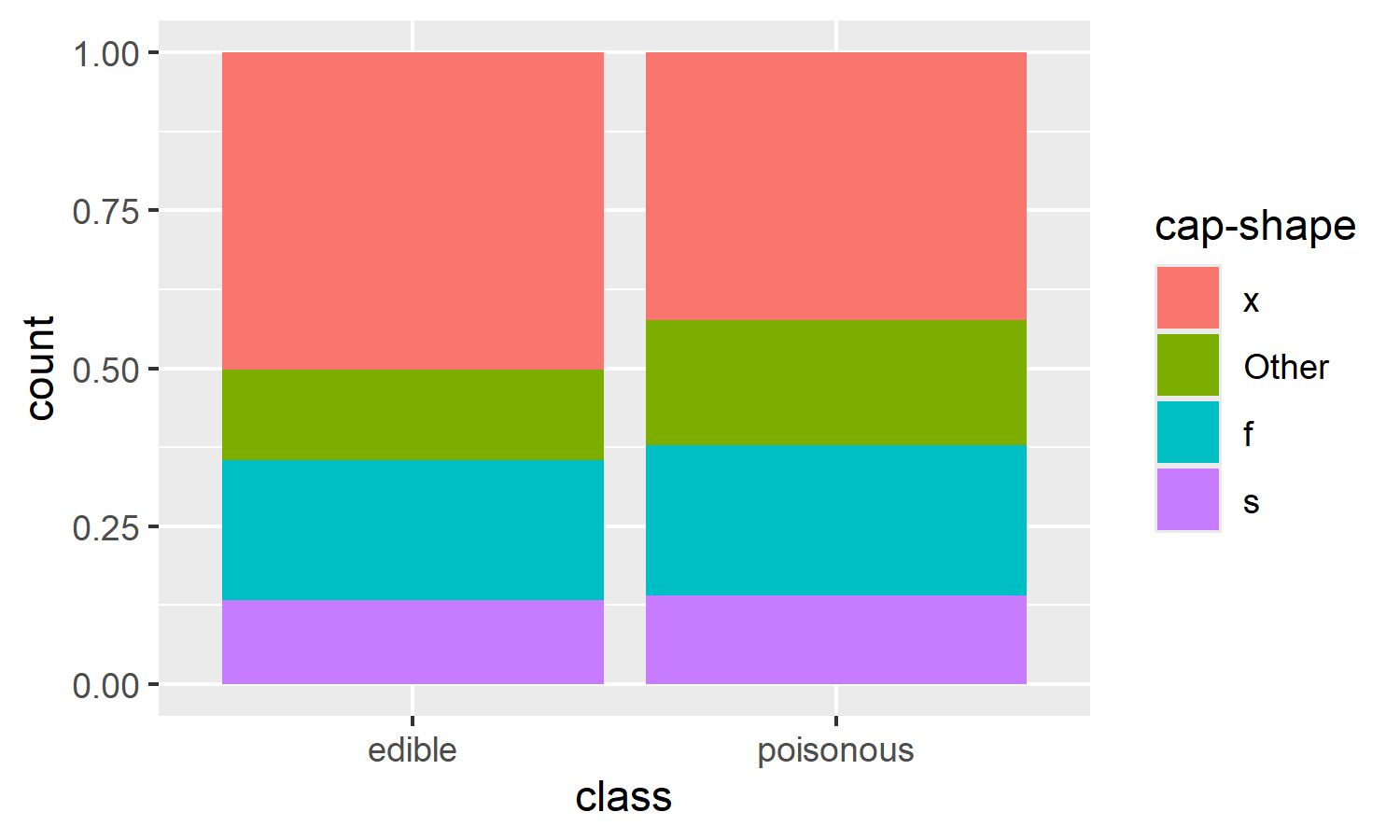
* Use **boxplots** for numeric variables and **stacked bar charts** for categorical variables to visually compare distributions across classes.
* Fit a **univariable logistic regression** model for each feature.
* Apply **Fisher’s exact test** to categorical predictors to evaluate their association with class.
* Perform **Type II ANOVA** on each logistic regression model to test for overall predictor significance.
* Format and display results using clean summary tables.

### cap-diameter



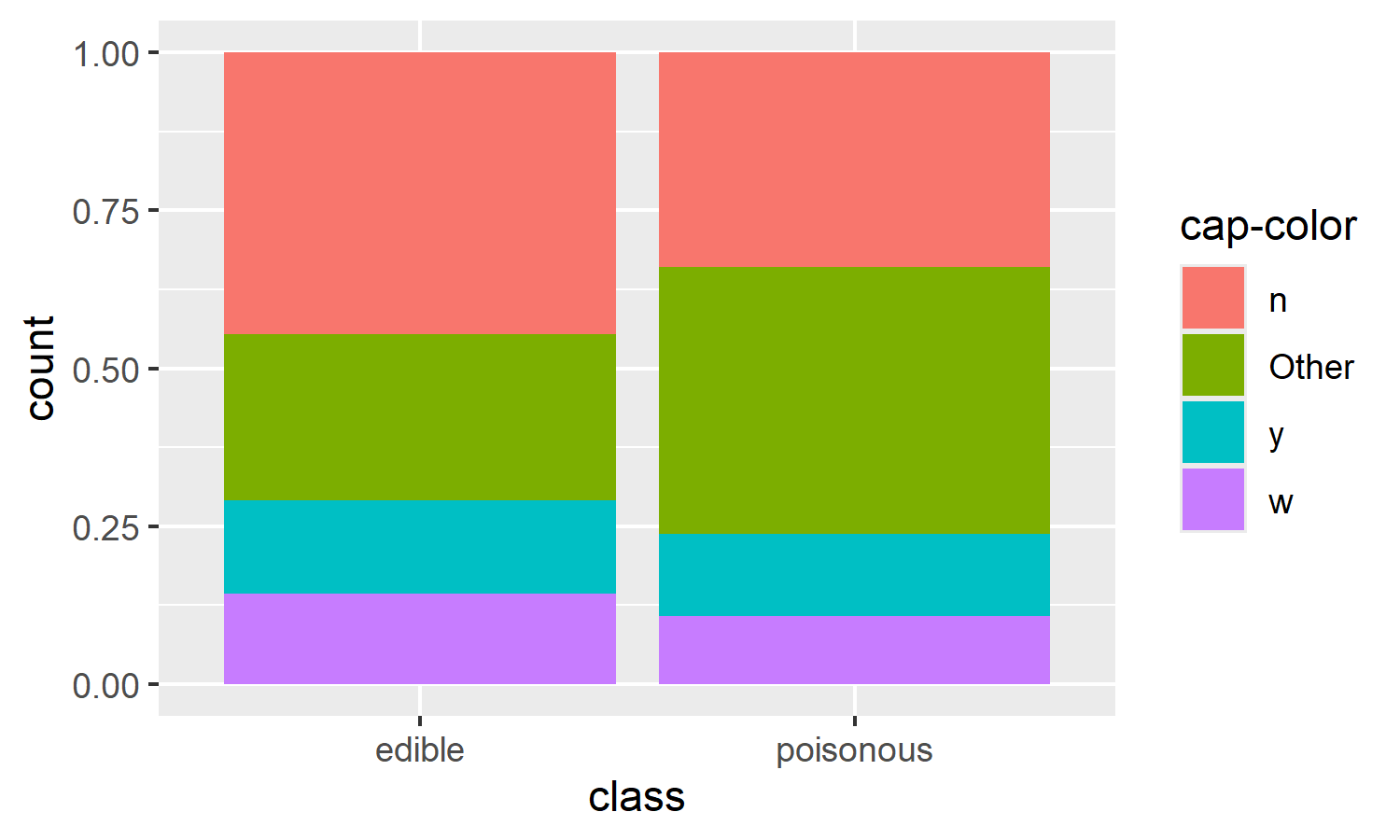
| term | statistic | df | p.value |
| --- | --- | --- | --- |
| `cap-diameter` | 1,346.734 | 1 | 0.001 |

### cap-shape



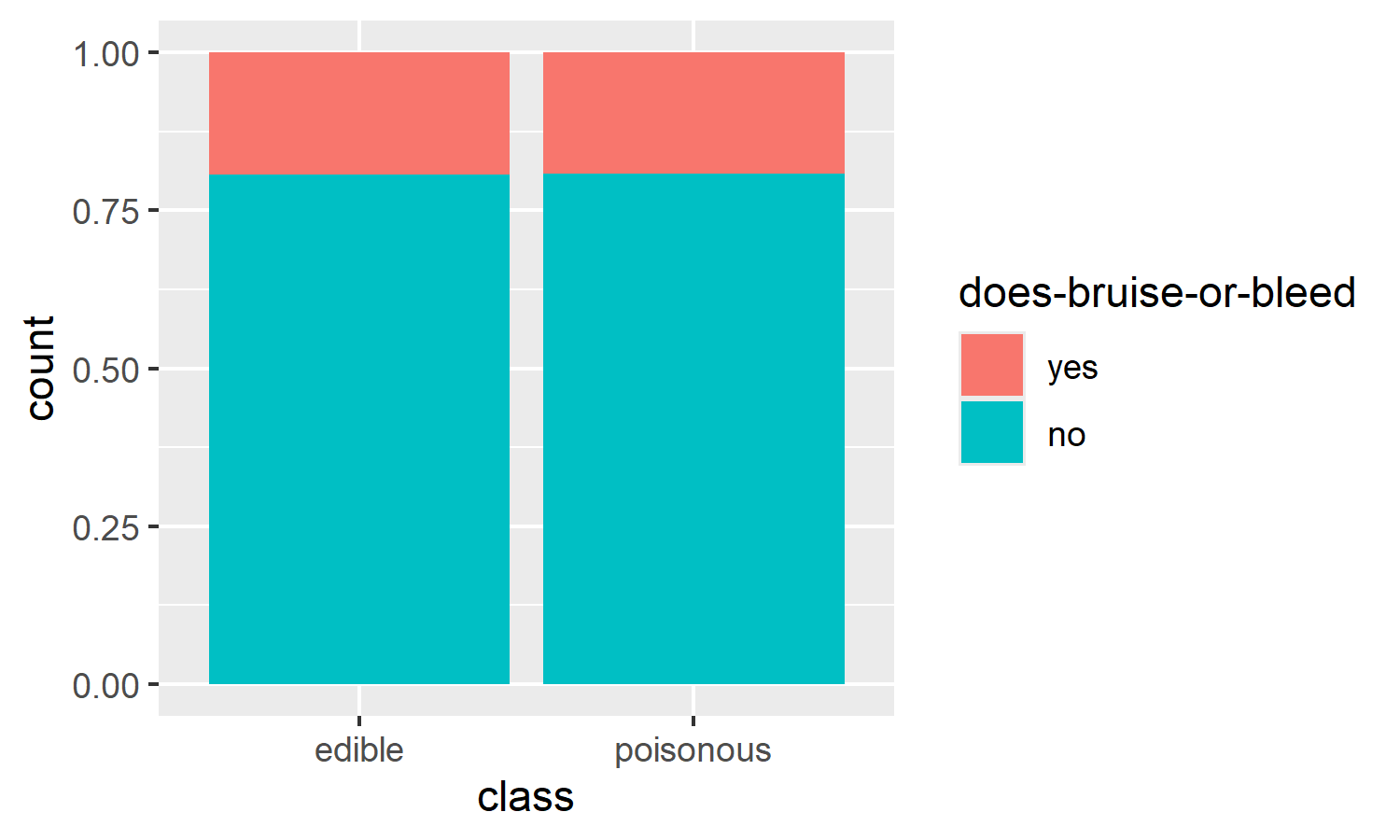
| term | statistic | df | p.value |
| --- | --- | --- | --- |
| `cap-shape` | 427.6002 | 3 | 0.001 |

### cap-color



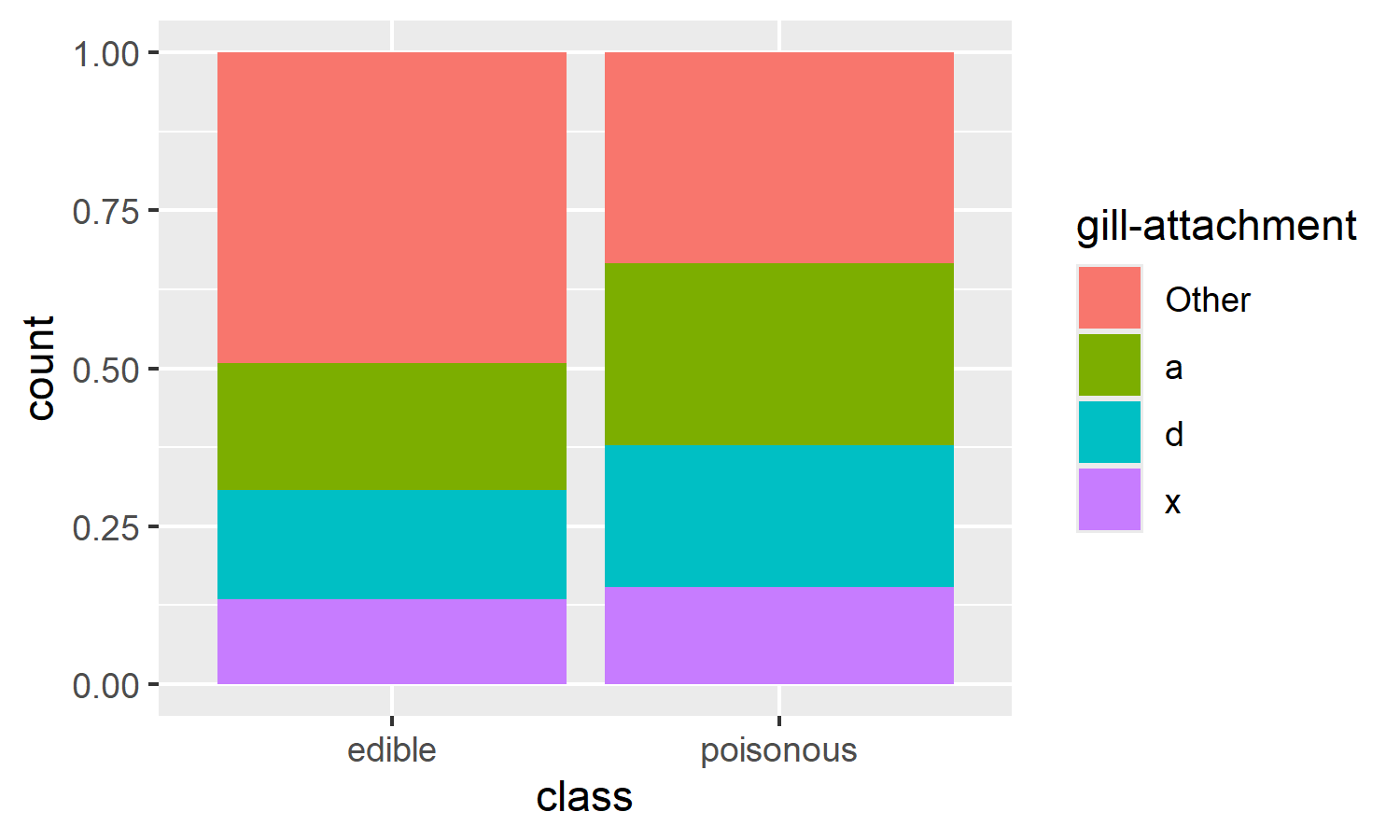
| term | statistic | df | p.value |
| --- | --- | --- | --- |
| `cap-color` | 1,465.465 | 3 | 0.001 |

### does-bruise-or-bleed



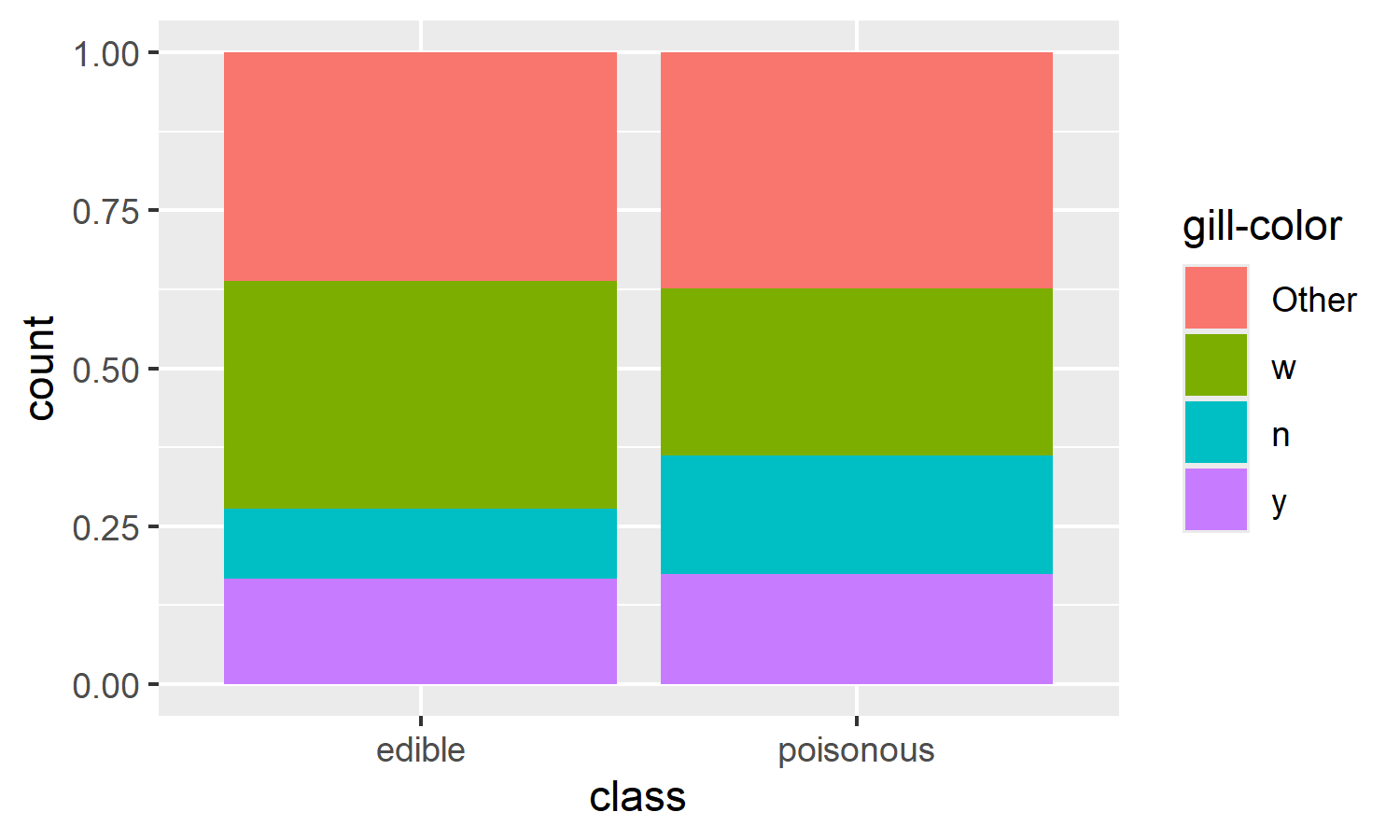
| term | statistic | df | p.value |
| --- | --- | --- | --- |
| `does-bruise-or-bleed` | 0.2421258 | 1 | 0.623 |

### gill-attachment



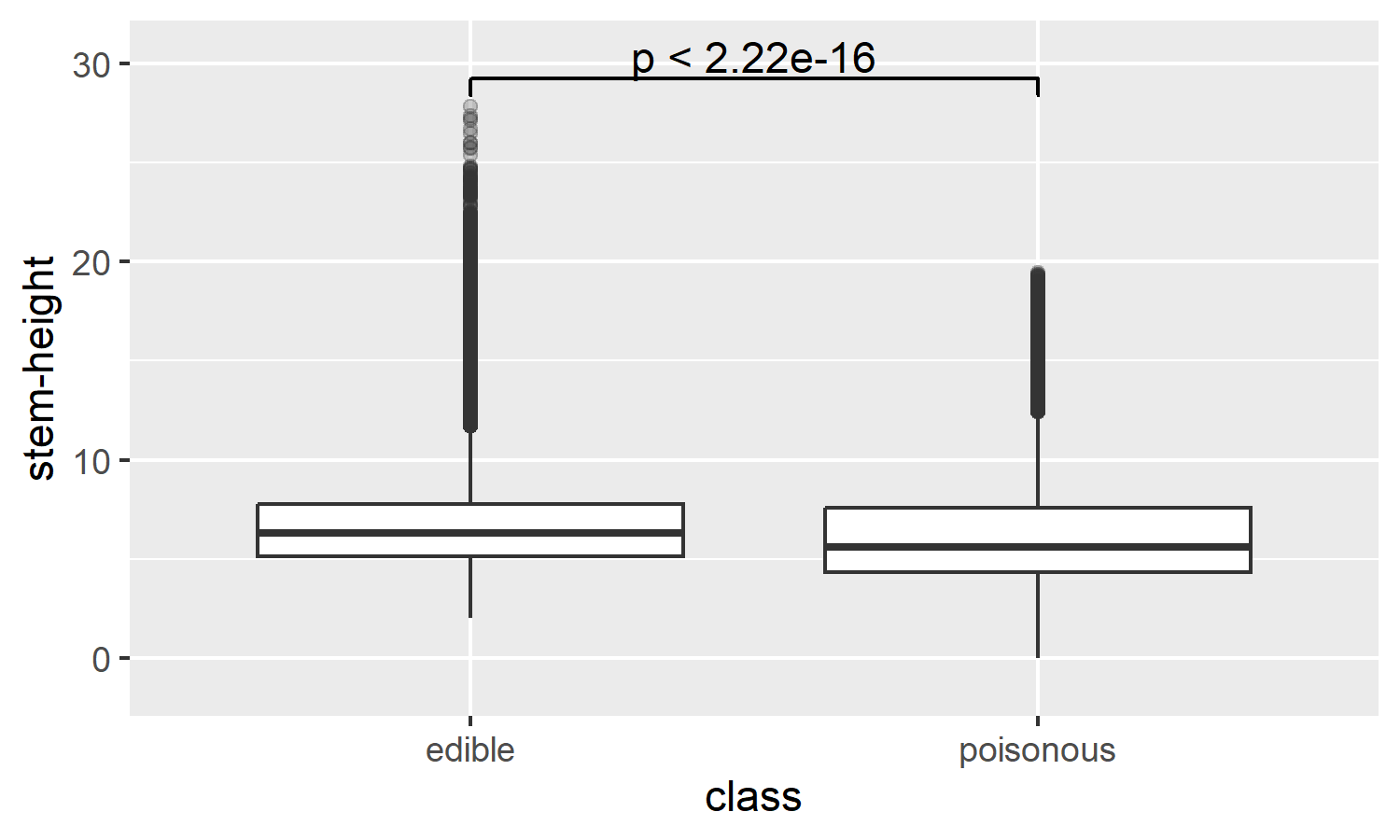
| term | statistic | df | p.value |
| --- | --- | --- | --- |
| `gill-attachment` | 1,398.169 | 3 | 0.001 |

### gill-color



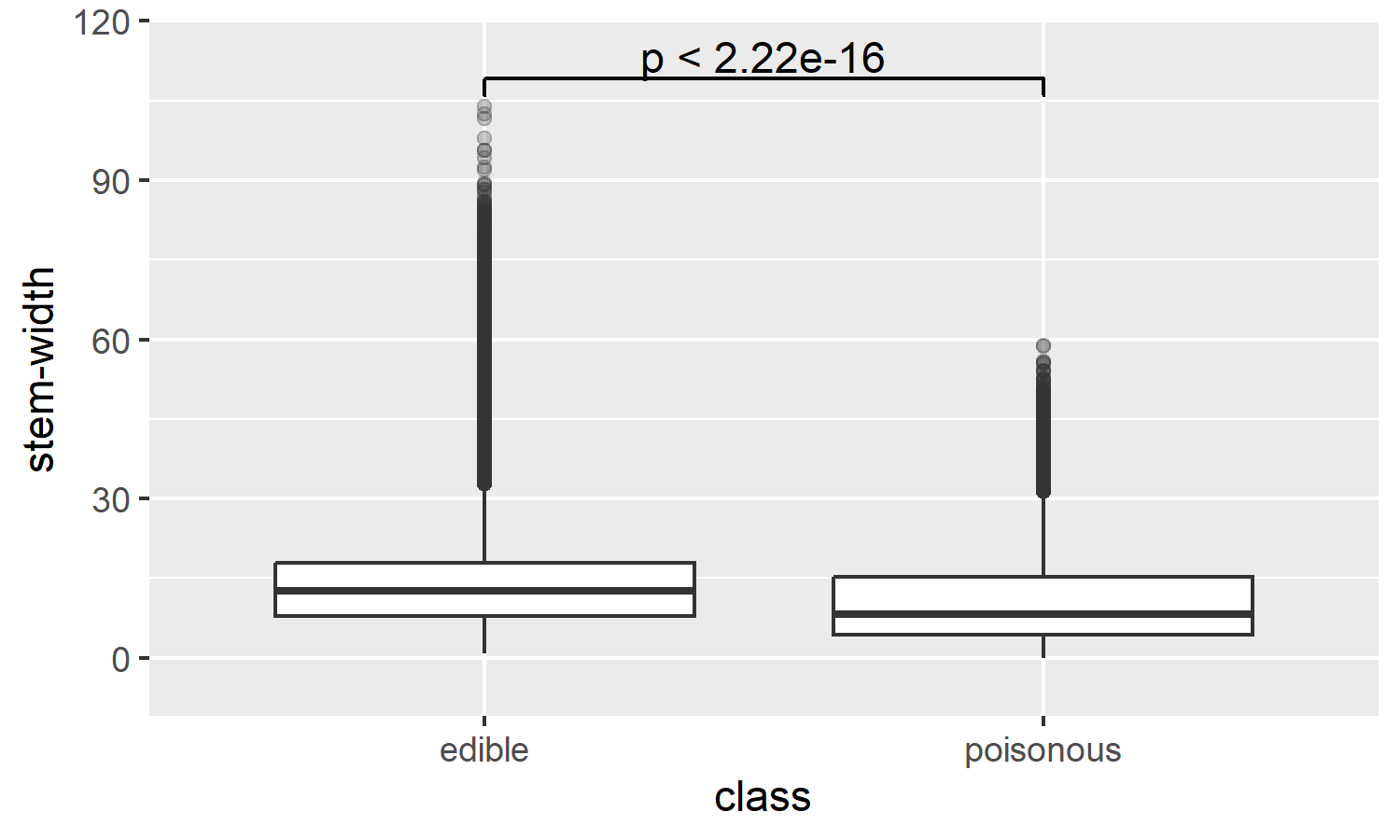
| term | statistic | df | p.value |
| --- | --- | --- | --- |
| `gill-color` | 896.6245 | 3 | 0.001 |

### stem-height



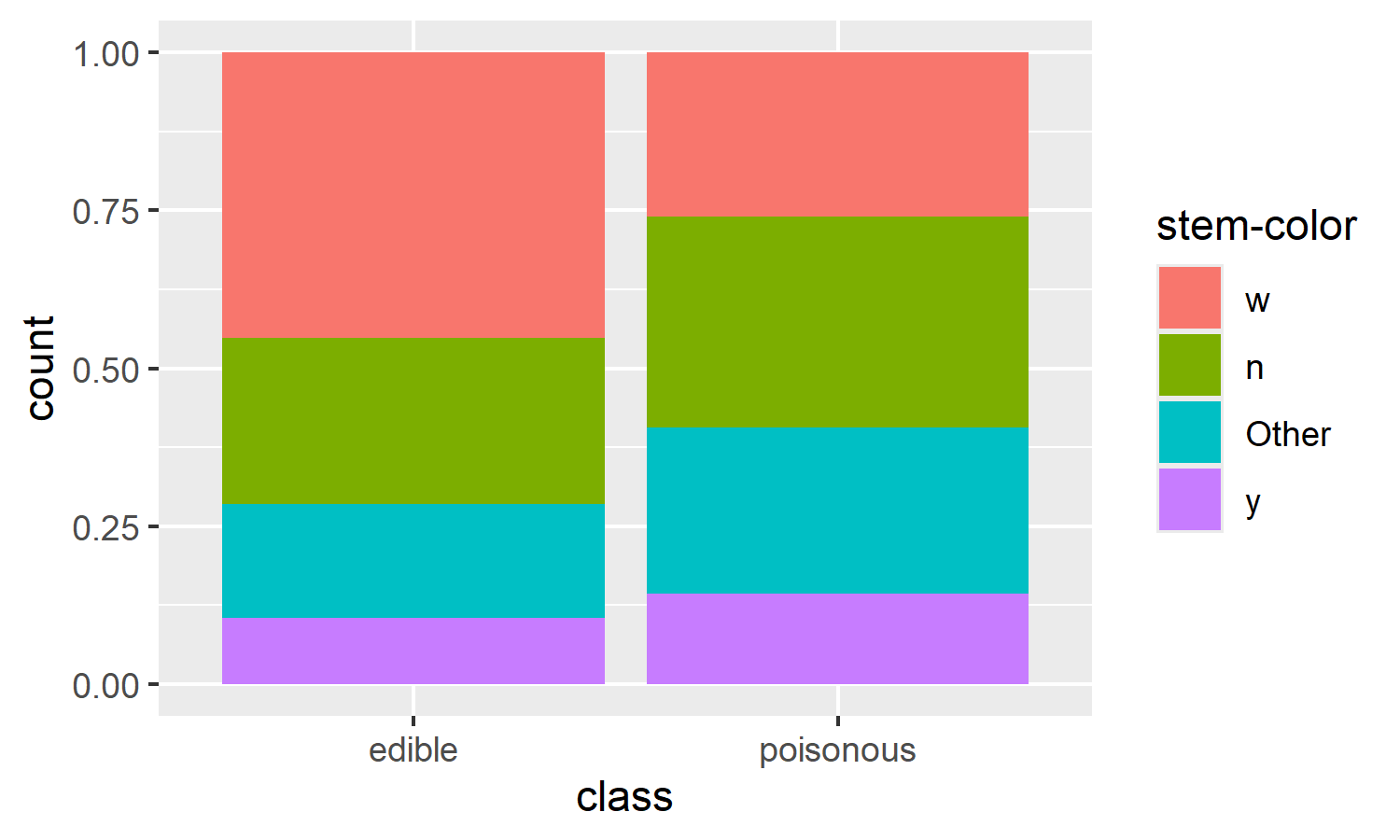
| term | statistic | df | p.value |
| --- | --- | --- | --- |
| `stem-height` | 746.53 | 1 | 0.001 |

### stem-width



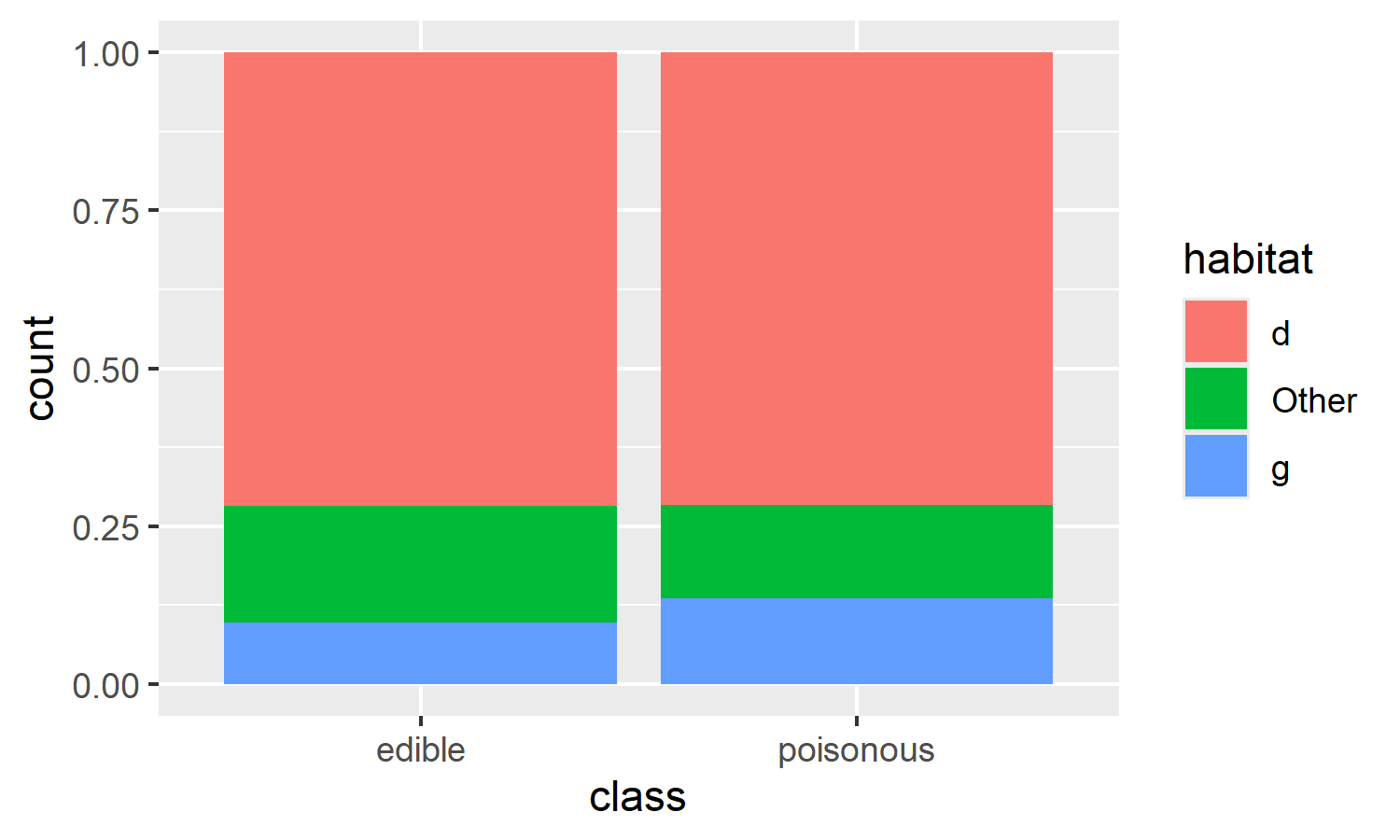
| term | statistic | df | p.value |
| --- | --- | --- | --- |
| `stem-width` | 1,574.721 | 1 | 0.001 |

### stem-color



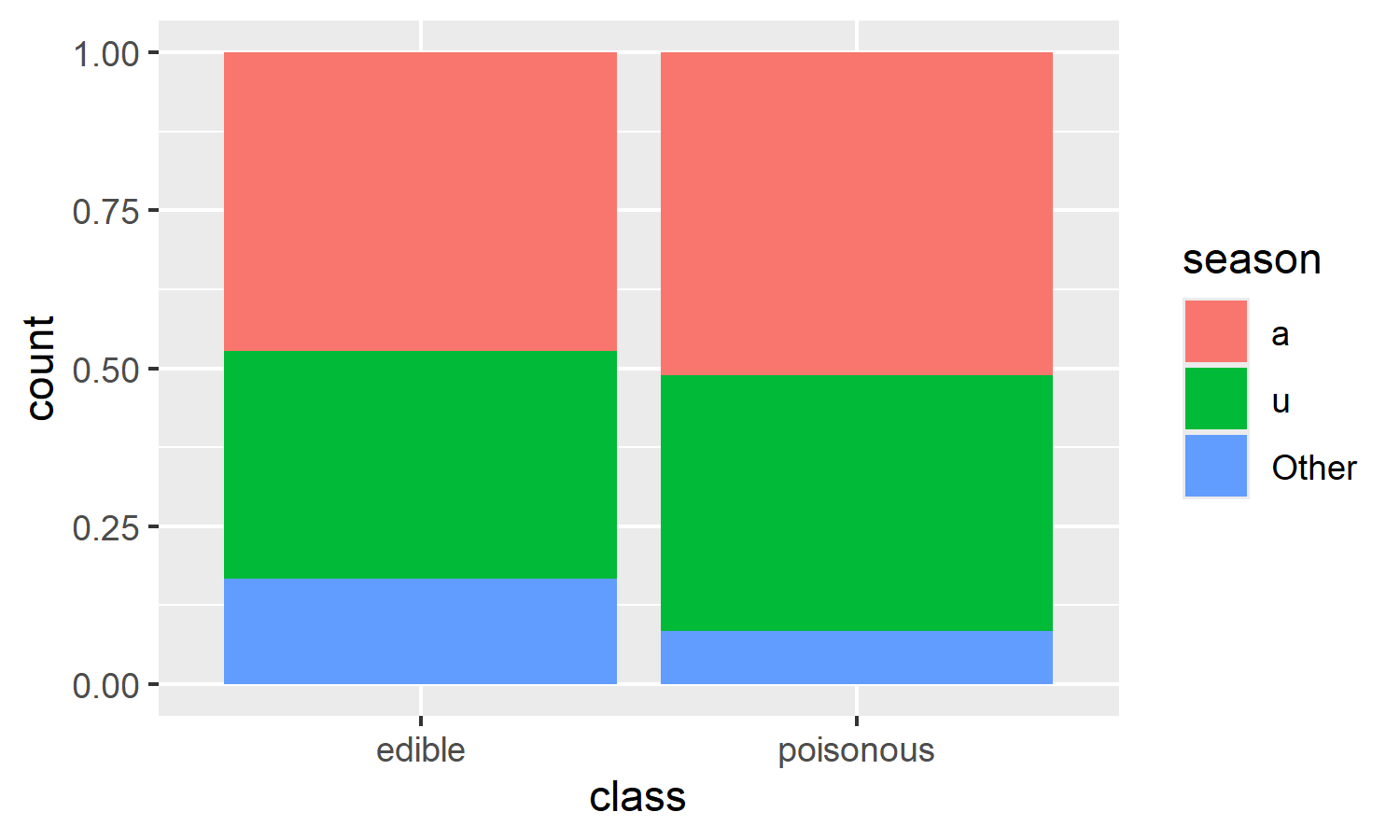
| term | statistic | df | p.value |
| --- | --- | --- | --- |
| `stem-color` | 2,093.702 | 3 | 0.001 |

### habitat



| term | statistic | df | p.value |
| --- | --- | --- | --- |
| habitat | 248.3861 | 2 | 0.001 |

### season



| term | statistic | df | p.value |
| --- | --- | --- | --- |
| season | 837.5709 | 2 | 0.001 |

We assessed each variable’s association with the class outcome using univariable logistic regression.

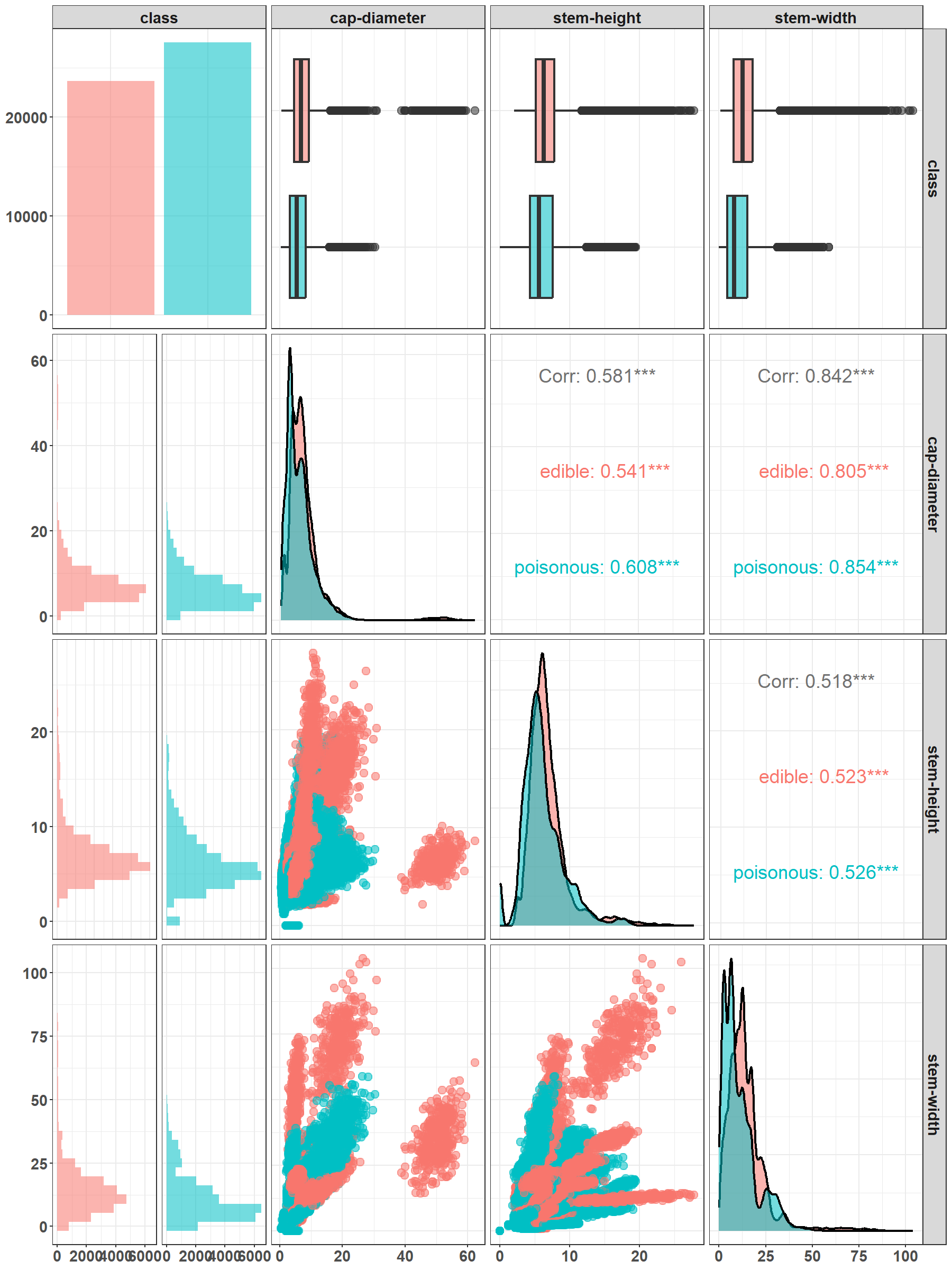
* **Cap diameter, cap shape, cap color, gill attachment, stem height, and stem width** were found to be **highly significant** predictors of edibility.
* **does-bruise-or-bleed** was **not statistically significant** (*p* = 0.623), suggesting it may not be a strong predictor when considered on its own.

These findings guided the selection of variables for the multivariable logistic regression model.

# Multivariable Logistic Regression (Full Model)

## Exploration of Collinearity

Before fitting the full multivariable model, we explored potential **collinearity** between numeric predictors using pairwise plots.



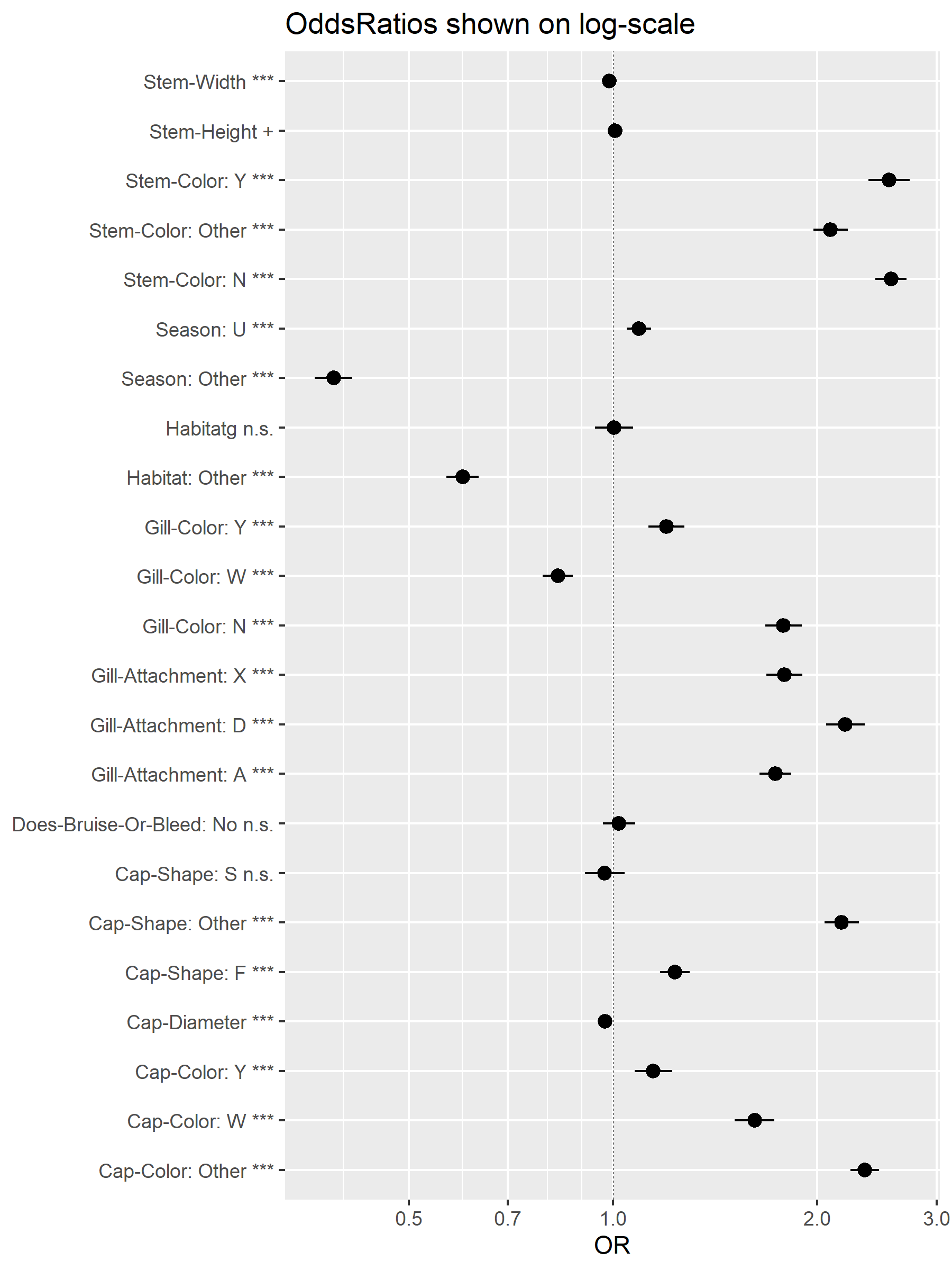
**Stem width** and **cap diameter** show strong positive correlation (*r* ≈ 0.84), especially for edible mushrooms. **Cap diameter** and **stem height** are also moderately correlated (*r* ≈ 0.58). Correlations are slightly stronger within each class, especially among edible mushrooms.

These relationships suggest some level of multicollinearity, which we account for during model fitting.

| Label | OR | CI\_low | CI\_high |
| --- | --- | --- | --- |
| Cap-Diameter \*\*\* | 0.9731974 | 0.9680589 | 0.9782986 |
| Cap-Shape: Other \*\*\* | 2.1731971 | 2.0509701 | 2.3030531 |
| Cap-Shape: F \*\*\* | 1.2346543 | 1.1739952 | 1.2984910 |
| Cap-Shape: S n.s. | 0.9727583 | 0.9096685 | 1.0401991 |
| Cap-Color: Other \*\*\* | 2.3496504 | 2.2388300 | 2.4661903 |
| Cap-Color: Y \*\*\* | 1.1474222 | 1.0762332 | 1.2233512 |
| Cap-Color: W \*\*\* | 1.6169140 | 1.5109076 | 1.7304853 |
| Does-Bruise-Or-Bleed: No n.s. | 1.0205228 | 0.9659432 | 1.0781916 |
| Gill-Attachment: A \*\*\* | 1.7357327 | 1.6452891 | 1.8312390 |
| Gill-Attachment: D \*\*\* | 2.2015763 | 2.0637329 | 2.3489338 |
| Gill-Attachment: X \*\*\* | 1.7896909 | 1.6843394 | 1.9018433 |
| Gill-Color: W \*\*\* | 0.8293225 | 0.7878306 | 0.8730025 |
| Gill-Color: N \*\*\* | 1.7842123 | 1.6764962 | 1.8991423 |
| Gill-Color: Y \*\*\* | 1.1985933 | 1.1279107 | 1.2737973 |
| Stem-Height + | 1.0071092 | 0.9992731 | 1.0149882 |
| Stem-Width \*\*\* | 0.9883341 | 0.9855995 | 0.9910652 |
| Stem-Color: N \*\*\* | 2.5711114 | 2.4387763 | 2.7109810 |
| Stem-Color: Other \*\*\* | 2.0937253 | 1.9766785 | 2.2178974 |
| Stem-Color: Y \*\*\* | 2.5542895 | 2.3811202 | 2.7406026 |
| Habitat: Other \*\*\* | 0.6004431 | 0.5684661 | 0.6341787 |
| Habitatg n.s. | 1.0042526 | 0.9415911 | 1.0712042 |
| Season: U \*\*\* | 1.0921927 | 1.0485387 | 1.1376850 |
| Season: Other \*\*\* | 0.3874101 | 0.3634492 | 0.4128643 |

# create forest plot

To visualize the final logistic regression results, we created a **forest plot** showing the **odds ratios (ORs)** and **95% confidence intervals** for each predictor. Each line represents a predictor (or level) from the model. The **dot** shows the estimated OR, while the **line** represents the confidence interval. A **dashed line at OR = 1** represents the null effect — values crossing this line indicate non-significant predictors. The Y-axis is shown on a **log scale**, which is standard for odds ratios. This plot makes it easy to visually identify which variables significantly **increase** or **decrease** the odds of a mushroom being poisonous.



**Cap diameter, stem width, and stem height** are highly significant predictors (**p < 0.001**) with ORs far from 1, indicating strong associations with mushroom edibility. Several levels of **stem color, cap color, gill color, and gill attachment** also show strong effects. For example:

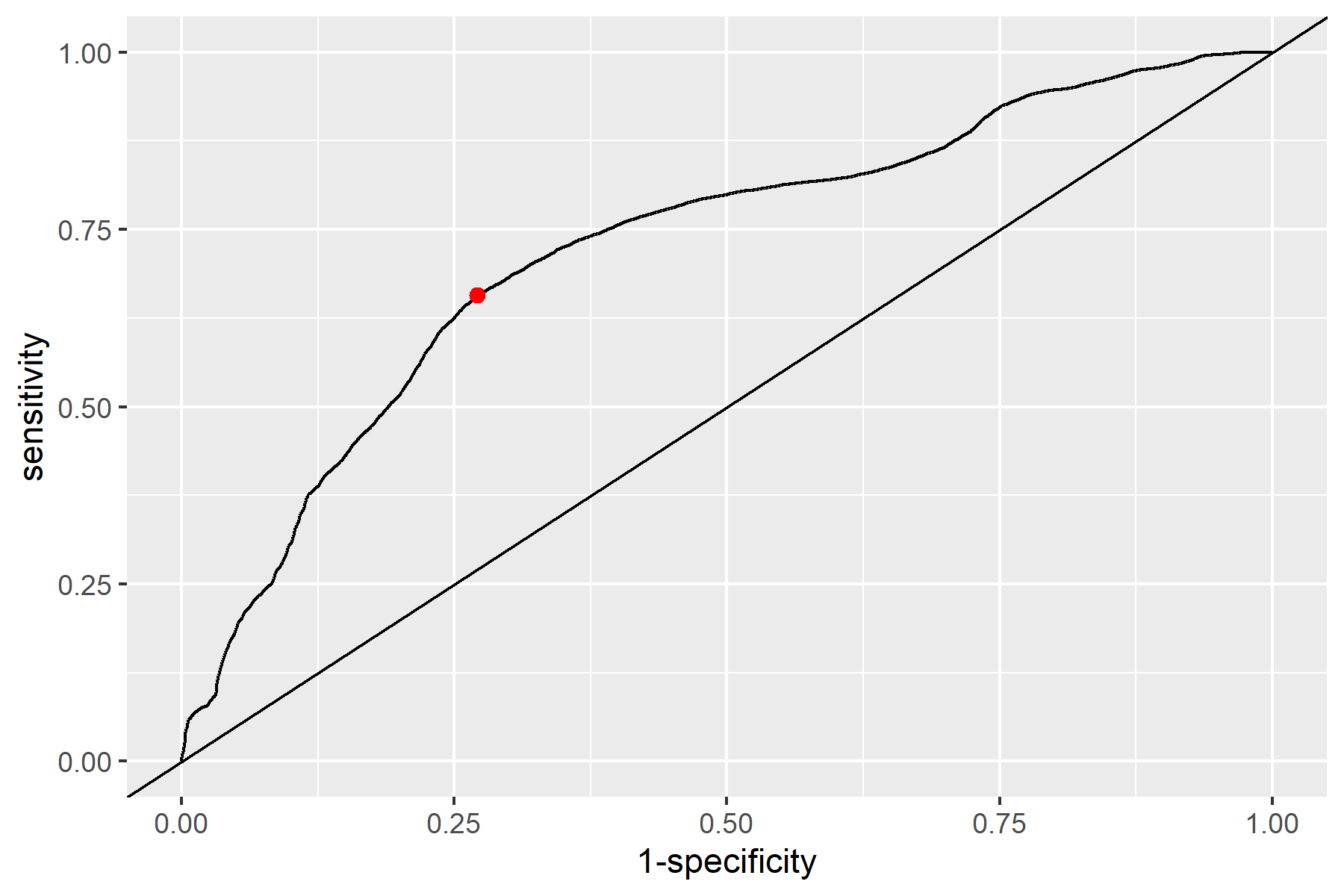
**Stem-Color: Y**, **Cap-Color: W**, and **Gill-Color: W** are strongly associated with higher or lower odds of being poisonous. **Does-Bruise-Or-Bleed: No** and **Cap-Shape: S** are not statistically significant (indicated by “n.s.”), as their confidence intervals include 1. A **logarithmic scale** is used on the X-axis to represent the multiplicative nature of odds ratios. Values **to the right of 1** suggest increased odds of being poisonous, while **values below 1** suggest decreased odds. Confidence intervals that **do not cross 1** represent statistically significant predictors.

This visualization highlights which mushroom characteristics most influence the probability of being poisonous, after adjusting for all other variables in the model. These results offer valuable insight for feature selection, risk interpretation, and biological relevance.

# Model Evaluation: ROC Curve and Classification Quality

After building the multivariable logistic regression model, we assess its **predictive performance** using the **ROC curve** and classification plots based on the predicted probabilities (p\_poison).

| Variable | value |
| --- | --- |
| threshold | 0.5571802 |
| specificity | 0.7287218 |
| sensitivity | 0.6575143 |



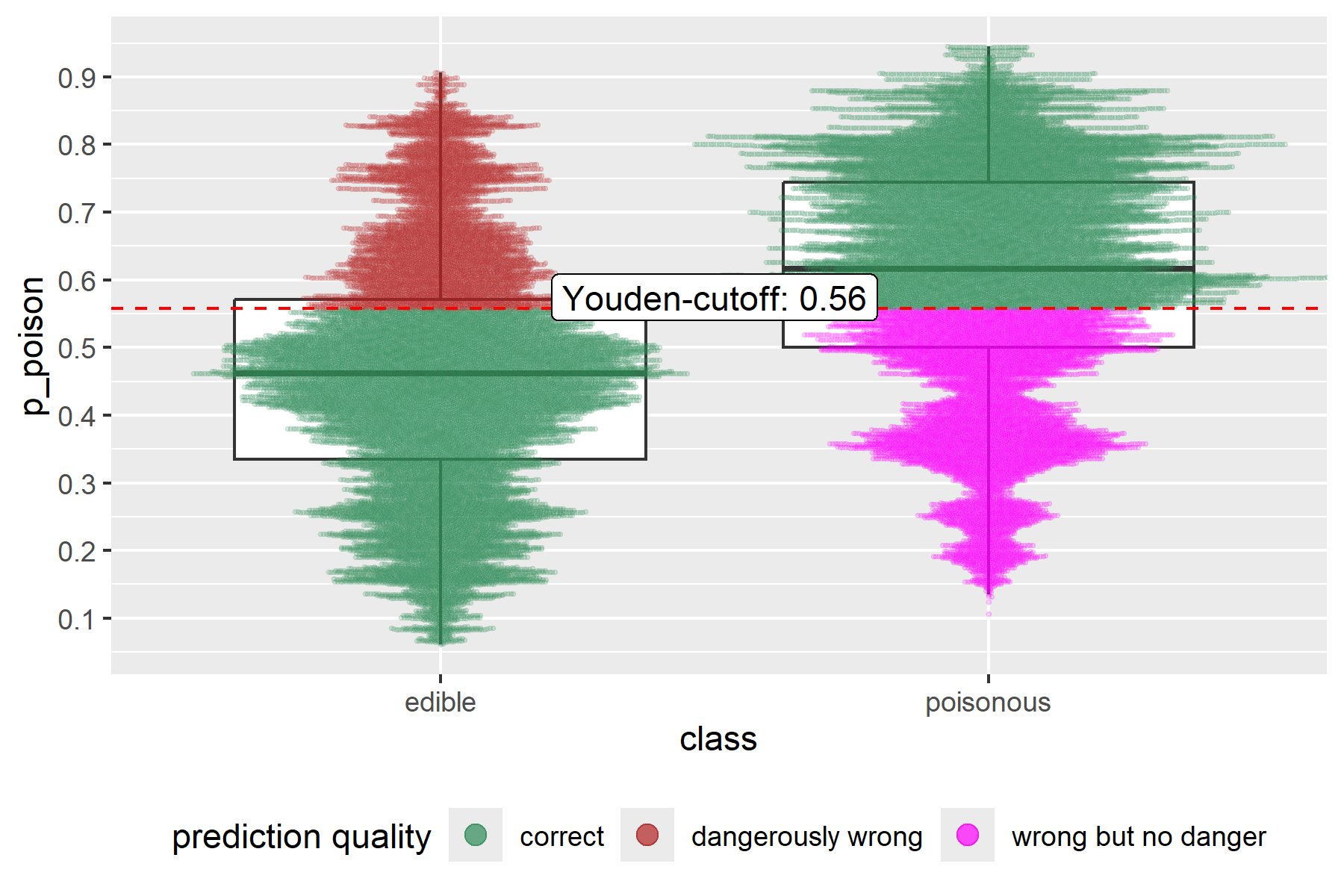
The **ROC curve** (Receiver Operating Characteristic) shows how well the model distinguishes between edible and poisonous mushrooms.

The **red dot** marks the **Youden’s optimal threshold (≈ 0.56)**, which balances **sensitivity** (true positive rate) and **specificity** (true negative rate). This threshold will be used to classify a mushroom as poisonous or edible based on its predicted probability.

**Youden’s optimal threshold = 0.56**

**Sensitivity = 0.66** → ~66% of poisonous mushrooms were correctly identified.

**Specificity = 0.73** → ~73% of edible mushrooms were correctly identified.



The violin/boxplot shows the distribution of predicted probabilities for each class:

Most **poisonous** mushrooms have high predicted probabilities (green = correct).

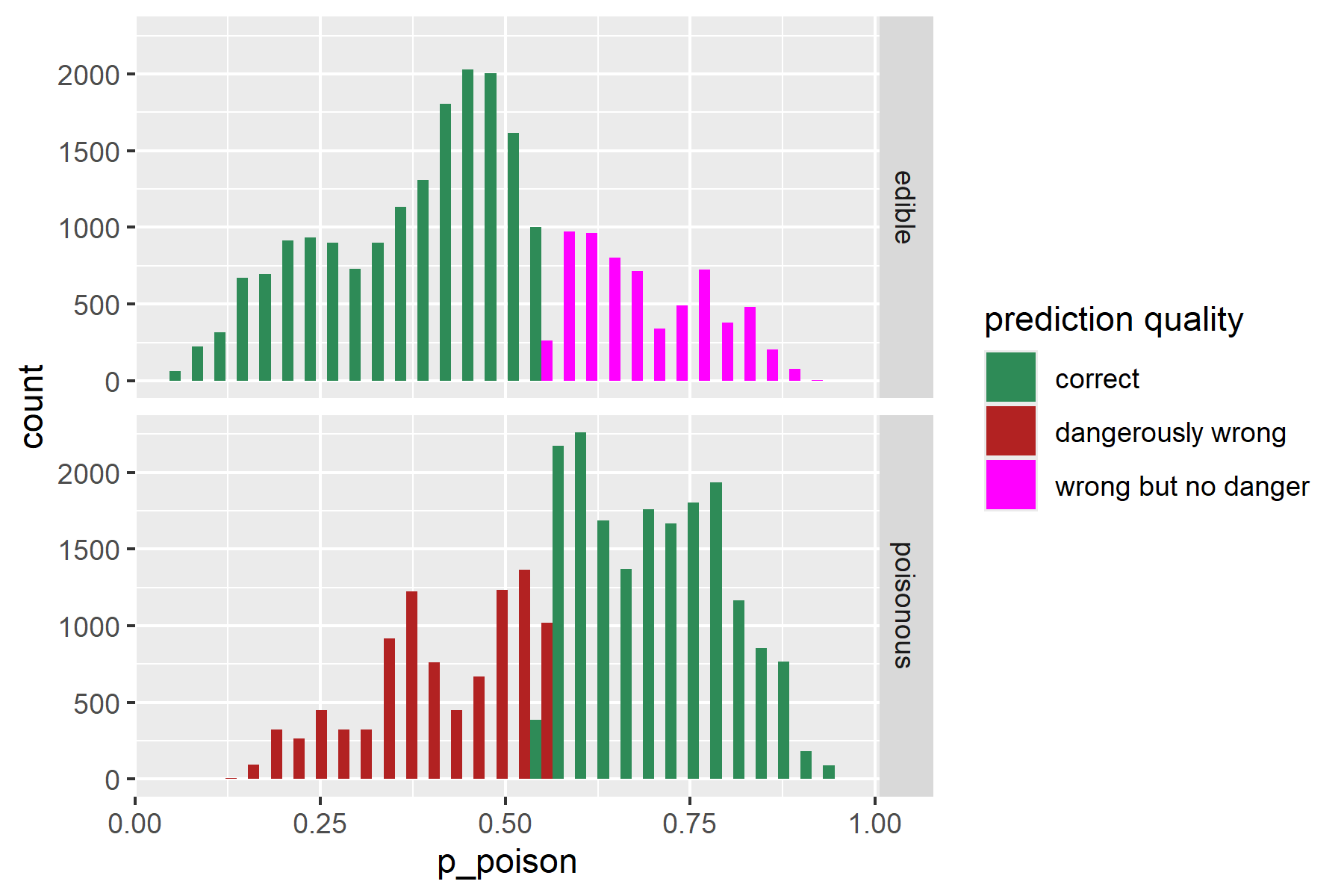
Most **edible** mushrooms have lower predicted probabilities (also green = correct).

**“Dangerously wrong”** predictions (edible predicted as poisonous or vice versa) are color-coded:

**Red** = edible predicted as poisonous (false positive → not dangerous).

**Magenta** = poisonous predicted as edible (false negative → dangerous).

This helps visualize the **model’s misclassification risk**.



The histogram breaks down how predictions are distributed across the probability range.

Green bars show correct predictions, while red and magenta indicate different types of errors.

it can be quickly identifyed **problematic prediction regions** near the cutoff (e.g. 0.5–0.6).