sommelieR

Bryce Rowland, Seoyoon Cho, Taylor Lagler, Paloma Hauser,
Mike Nodzenski

4/10/2019

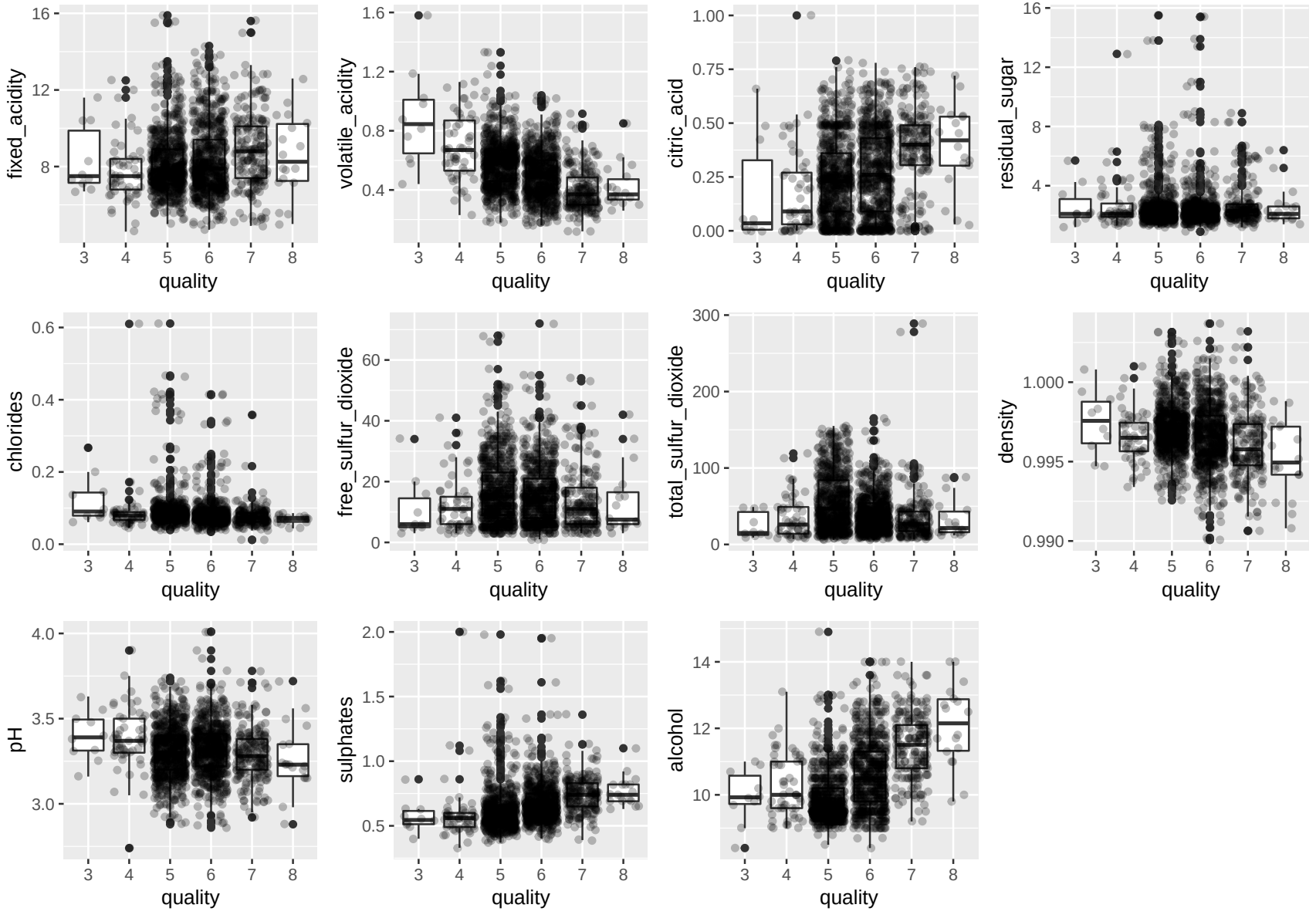
Introduction

- 2 datasets related to red and white Vinho Verde wines
- each contain 11 physiochemical variables
  - fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol
- 6,493 observations
- outcome variable: wine quality
  - ordinal variable ranging from 0-10
  - 0 is poor, 10 is excellent
  - classes are unbalanced

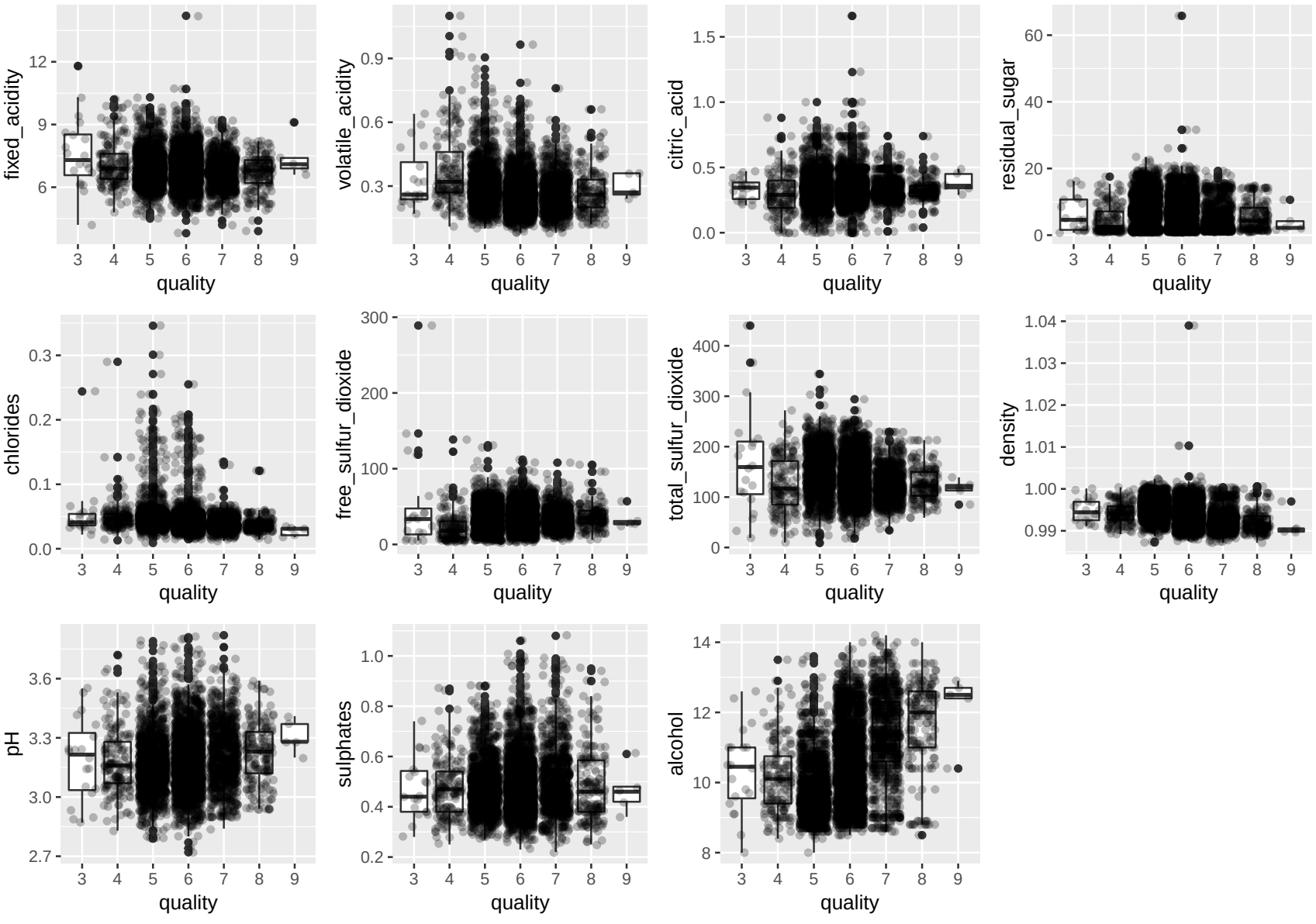
Project Aim

Is it possible to predict wine quality using some subset of the physiochemical variables?

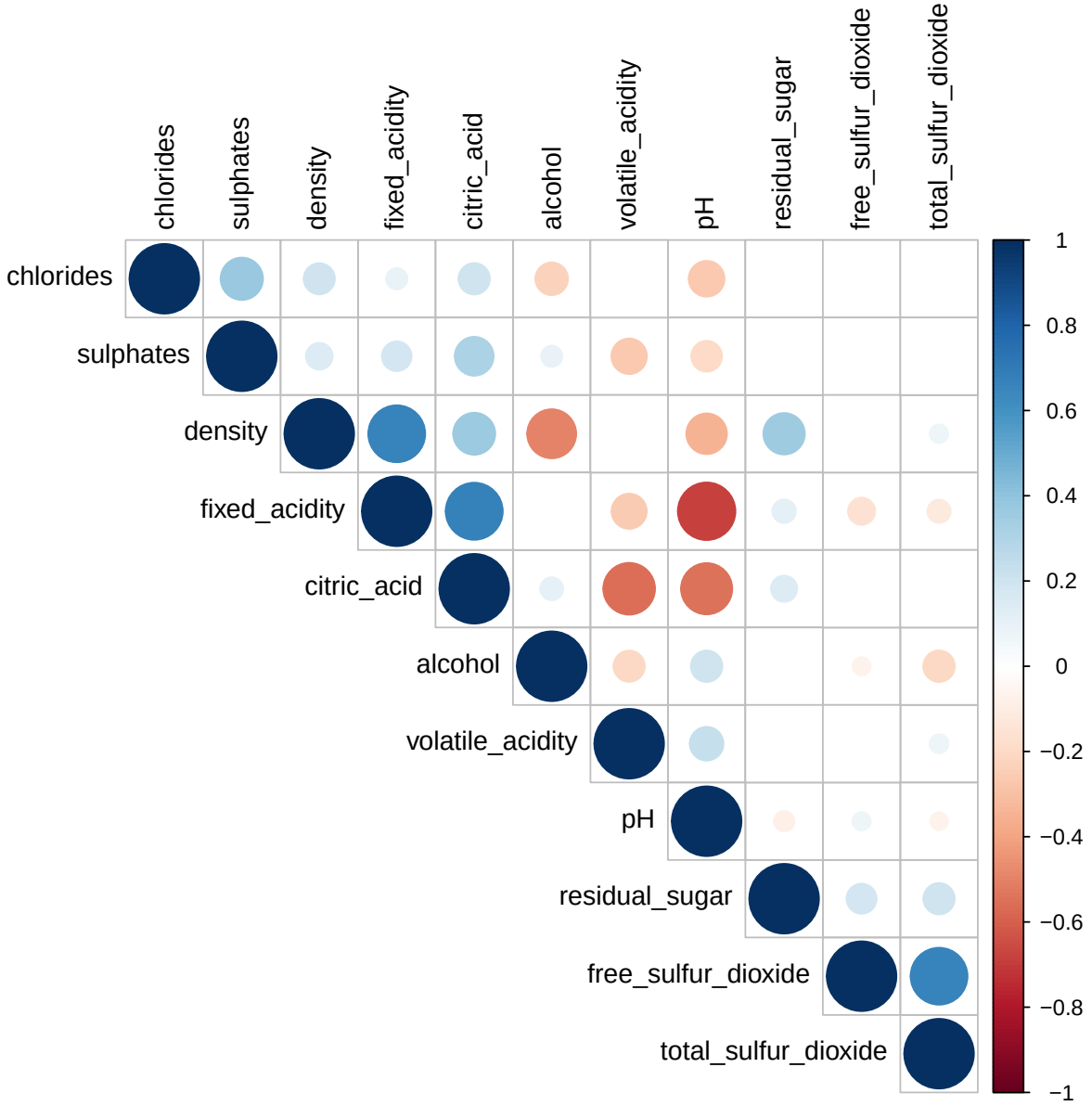
Looking At the Red Wine Data



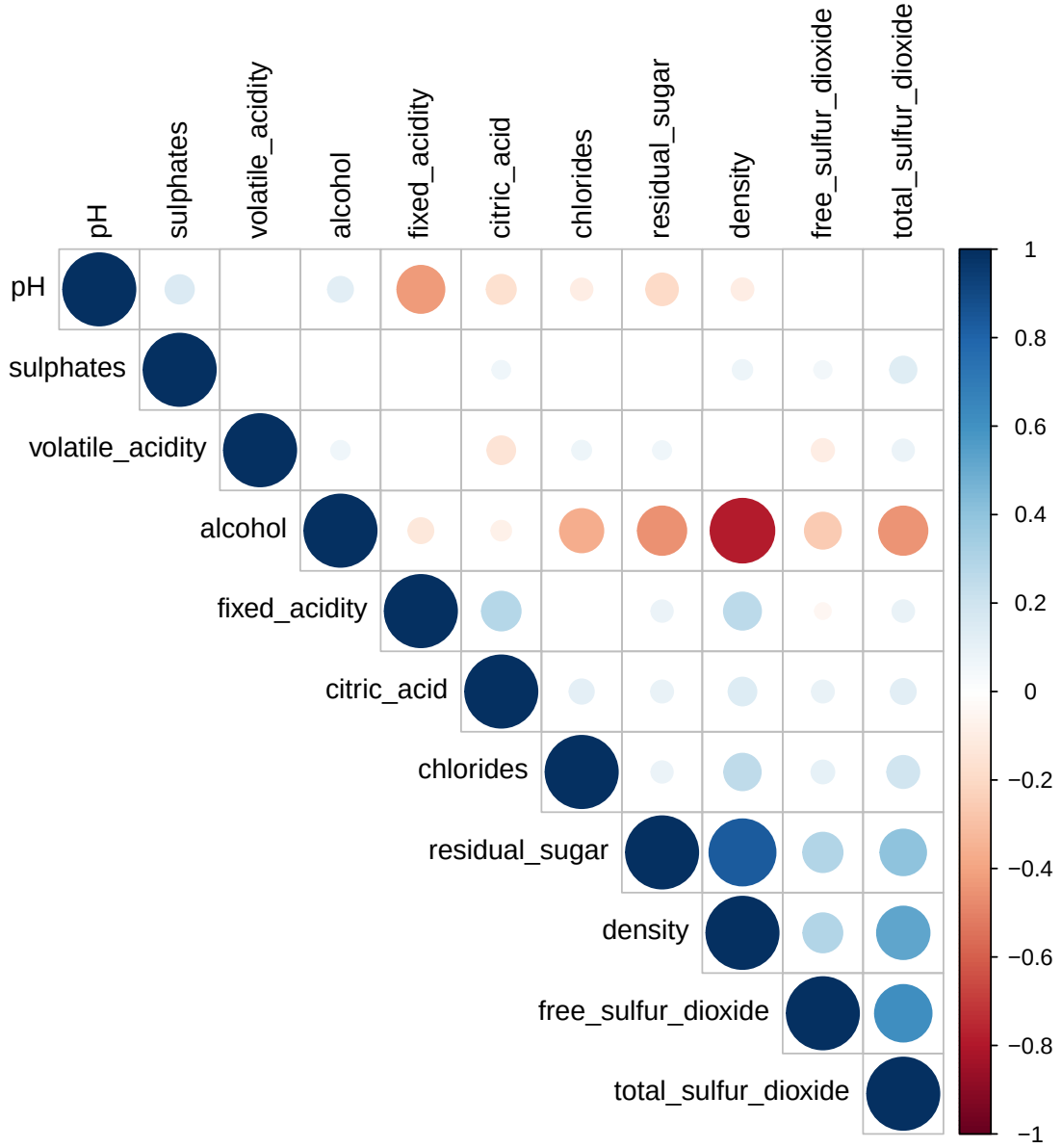
Looking At the White Wine Data



Correlations- Red Wine



Correlations- White Wine



Methods

- linear model
- partial proportional odds model
- multinomial model
- random forests

Variable Selection

For variable selection, we examined the correlations between predictors and considered best subsets found in R. For the likelihood based models, we included the following predictors:

- red wine: volatile acidity, total sulfur dioxide, pH, alcohol, sulfates
- white wine: pH, density, volatile_acidity, residual_sugar, alcohol

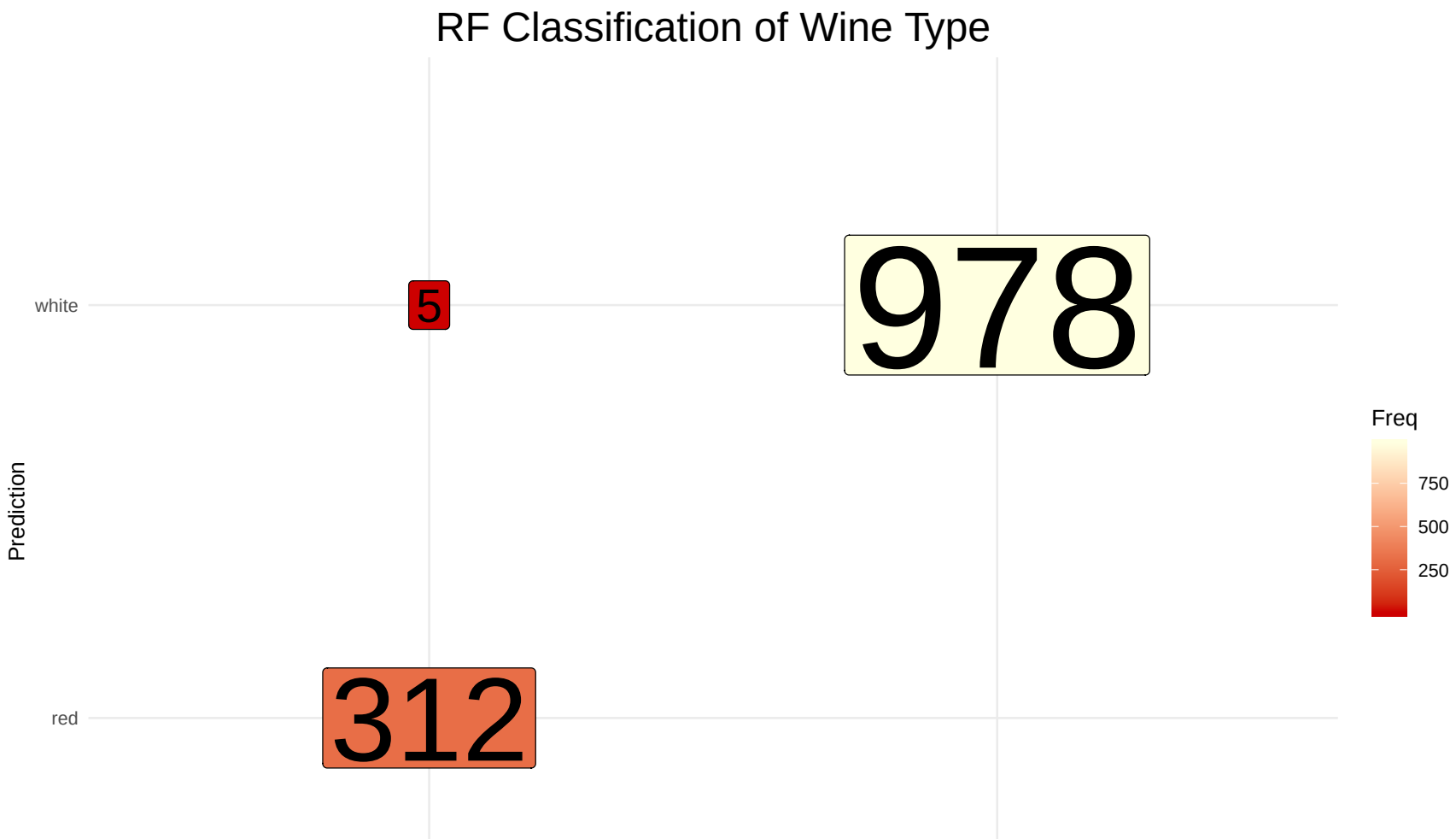
Results: Linear Model

Results: Multinomial Model

Results: Random Forests (wine type classification)

First, we merge the white and red wine datasets and use random forests to try to classify the wines into red or white

```
## Warning: Removed 1 rows containing missing values (geom_
```



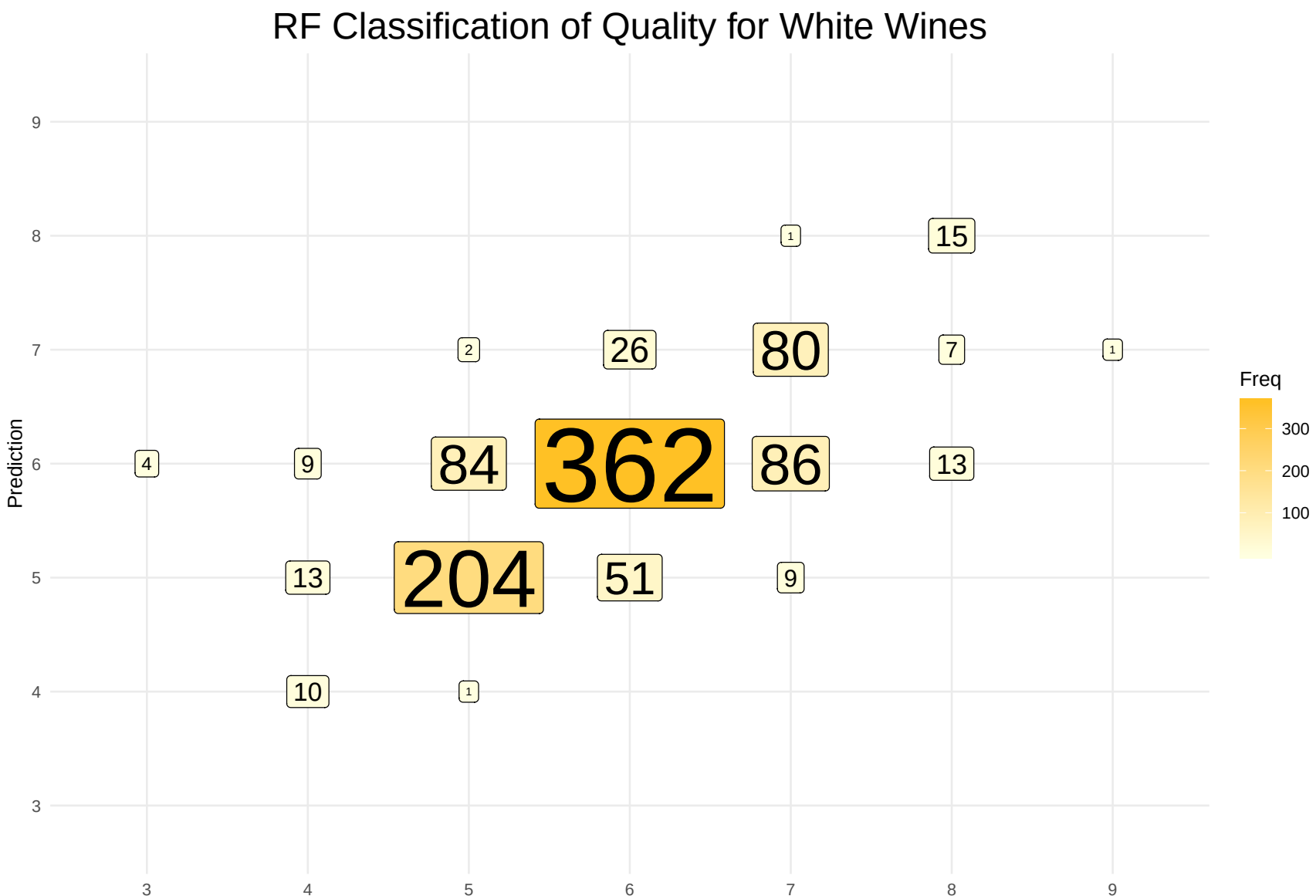
Results: Random Forests (red wine quality classification)

Next, we use random forests to classify each wine's wine quality score. We have 7 categories (scores 3-9) and we do this separately for the red and white wines.

```
## Warning: model fit failed for Resample02: mtry= 2 Error
##   Can't have empty classes in y.

## Warning: model fit failed for Resample02: mtry= 6 Error
##   Can't have empty classes in y.

## Warning: model fit failed for Resample02: mtry=11 Error
##   Can't have empty classes in y.

## Warning: model fit failed for Resample18: mtry= 2 Error
##   Can't have empty classes in y.

## Warning: model fit failed for Resample18: mtry= 6 Error
##   Can't have empty classes in y.
```

Results: Random Forests (white wine quality classification)

```
## Warning: Removed 30 rows containing missing values (geor
```



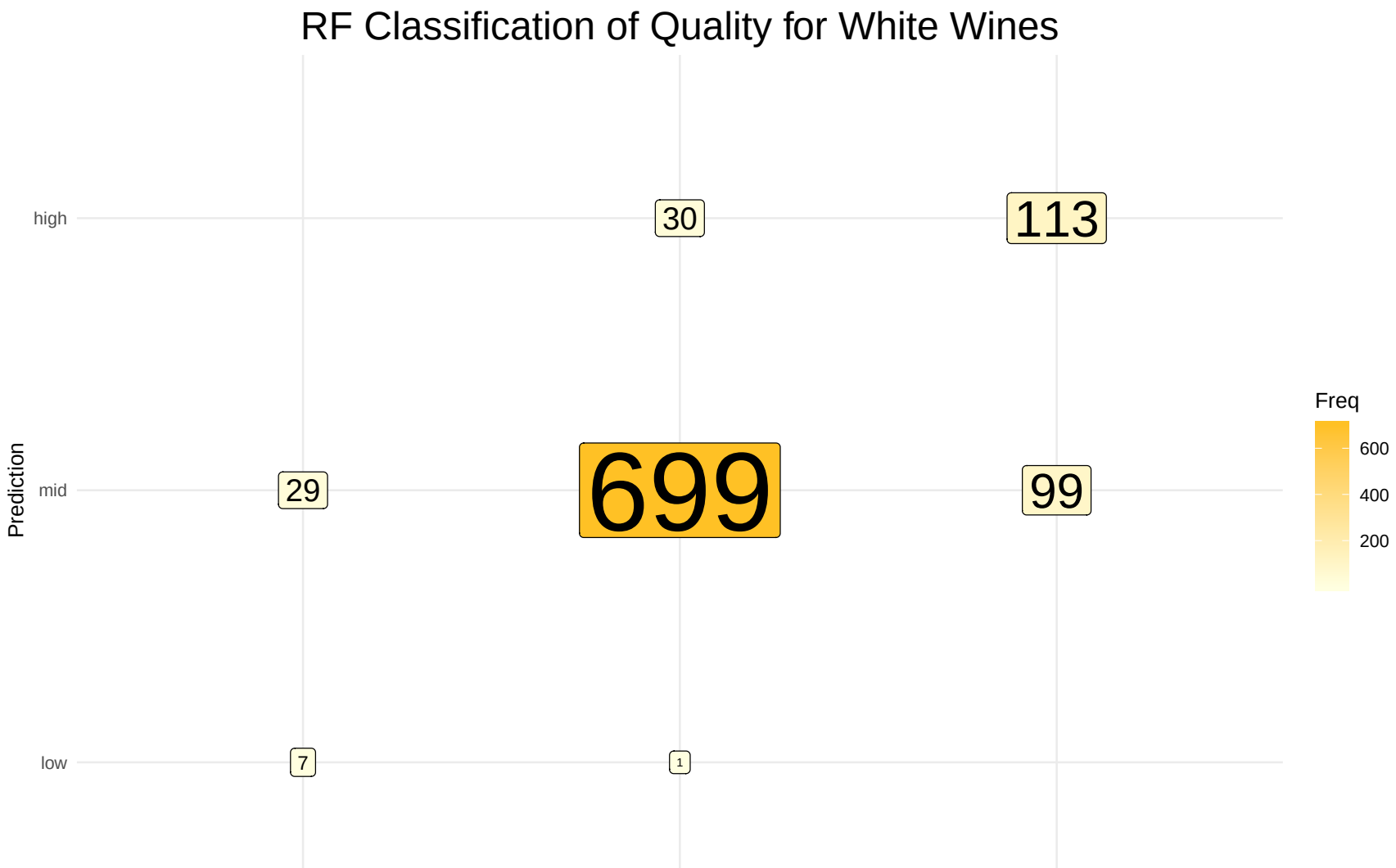
Results: Random Forests (grouped red wine quality classification)

We now group wine quality into three categories: low (scores 3-4), medium (5-6), high (7-9), and try to classify the wines into these three categories.

```
## Warning: Removed 4 rows containing missing values (geom_
```

Results: Random Forests (grouped white wine quality classification)

```
## Warning: Removed 2 rows containing missing values (geom_
```



Results: Accuracy

```
##
##   3   4   5   6   7   8
##   2  10 136 127  39   3

##
##   3   4   5   6   7   8   9
##   4  32 291 439 176  35   1

##       Accuracy AccuracyLower AccuracyUpper
##      0.7255521     0.6728923     0.7739405

##       Accuracy AccuracyLower AccuracyUpper
##      0.6860941     0.6559695     0.7150934

## y.test.red.grp
##  low  mid high
##   12  263   42
```

Discussion

(about interpretation or possible future directions)