sommelieR

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

4/10/2019

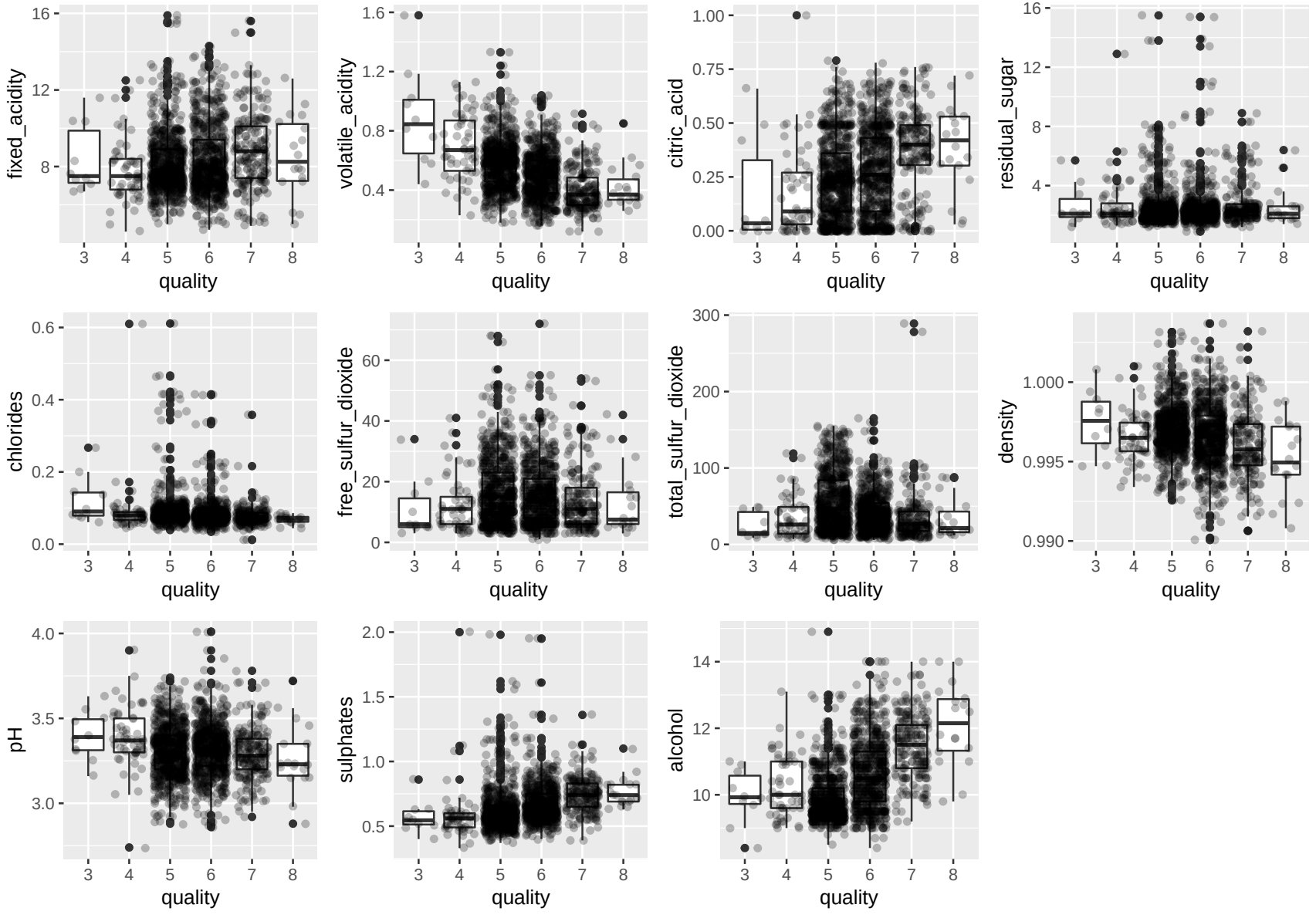
Introduction

- 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

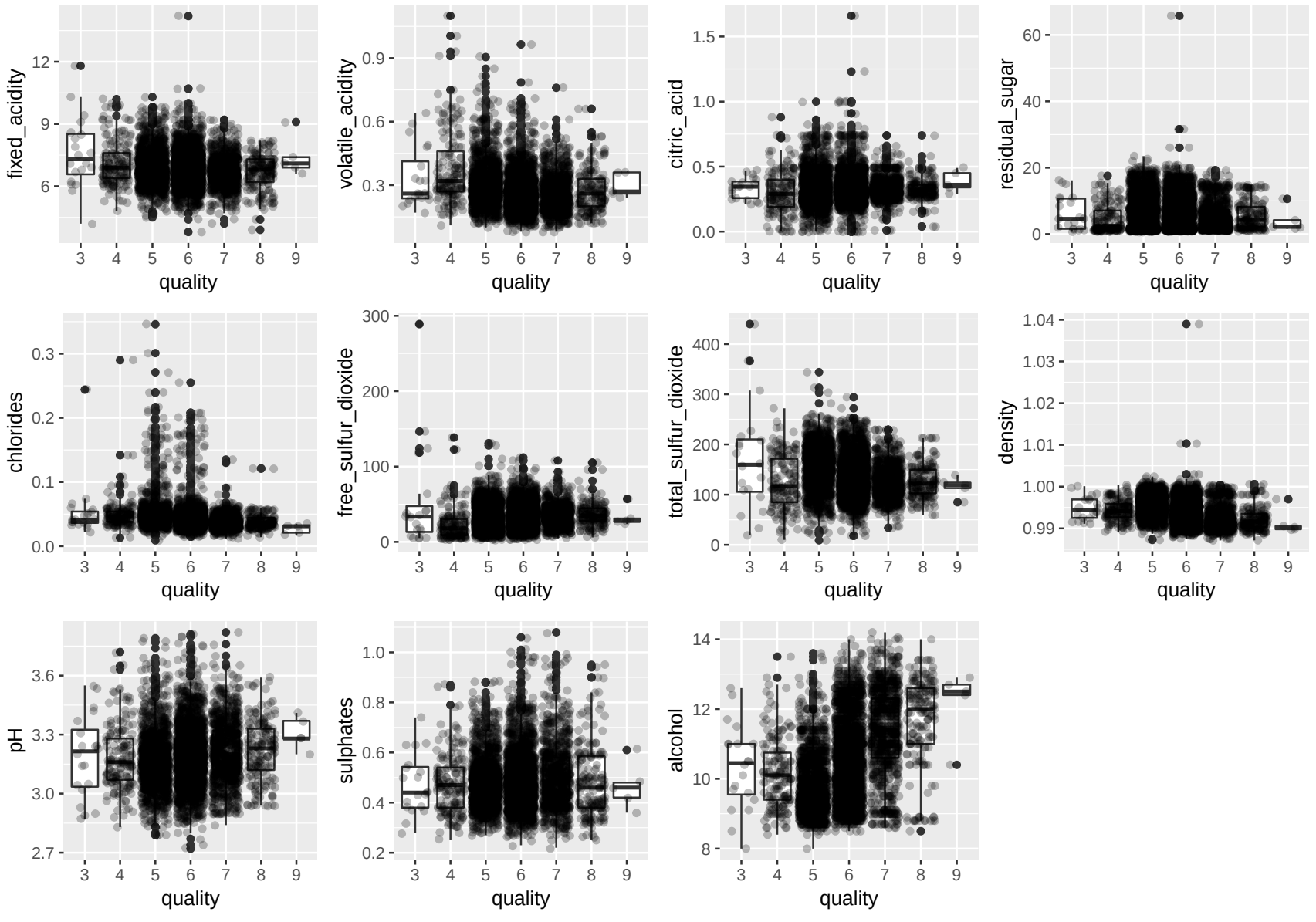
Project Aim

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

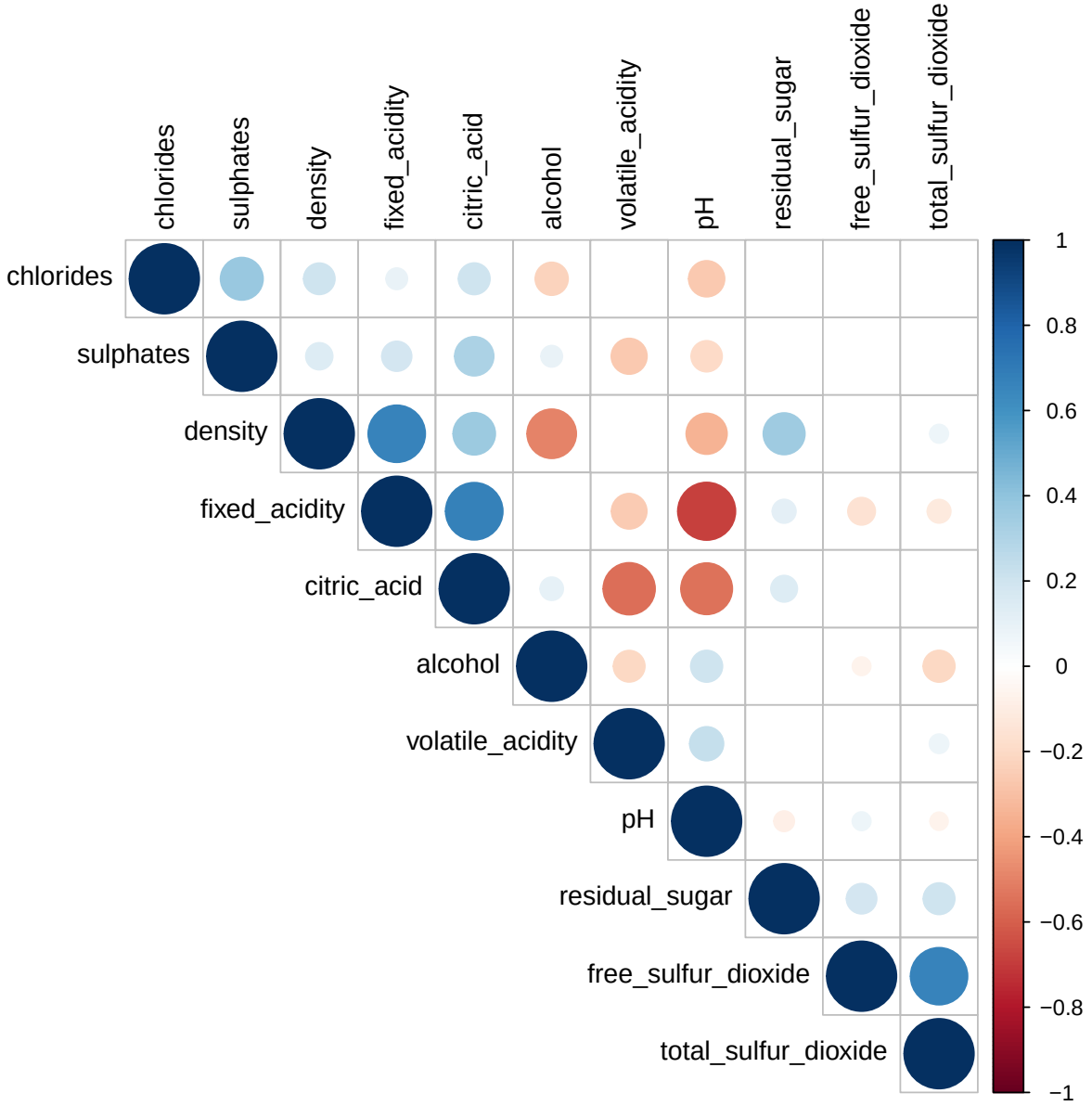
Looking At the Red Wine Data



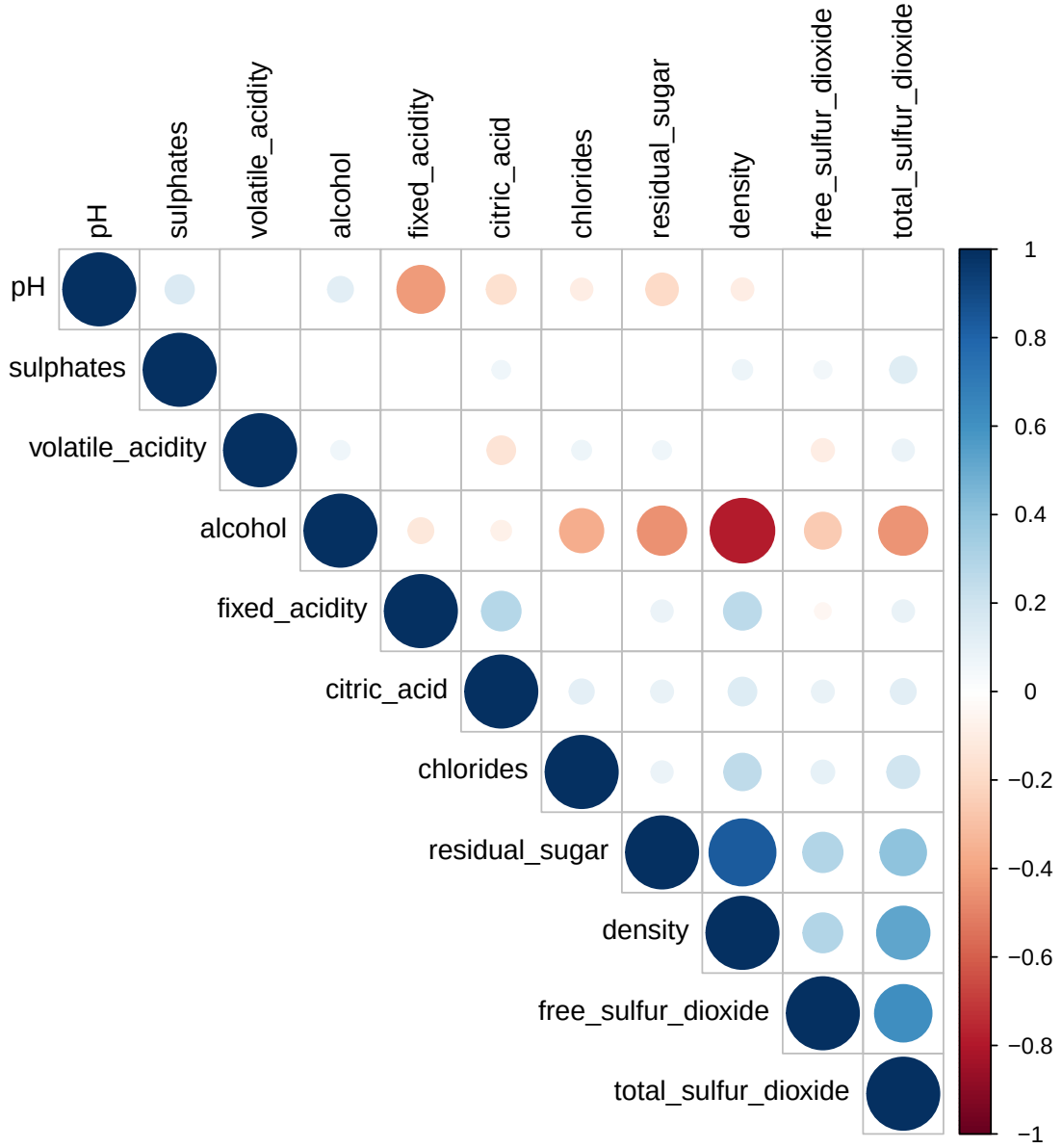
Looking At the White Wine Data



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

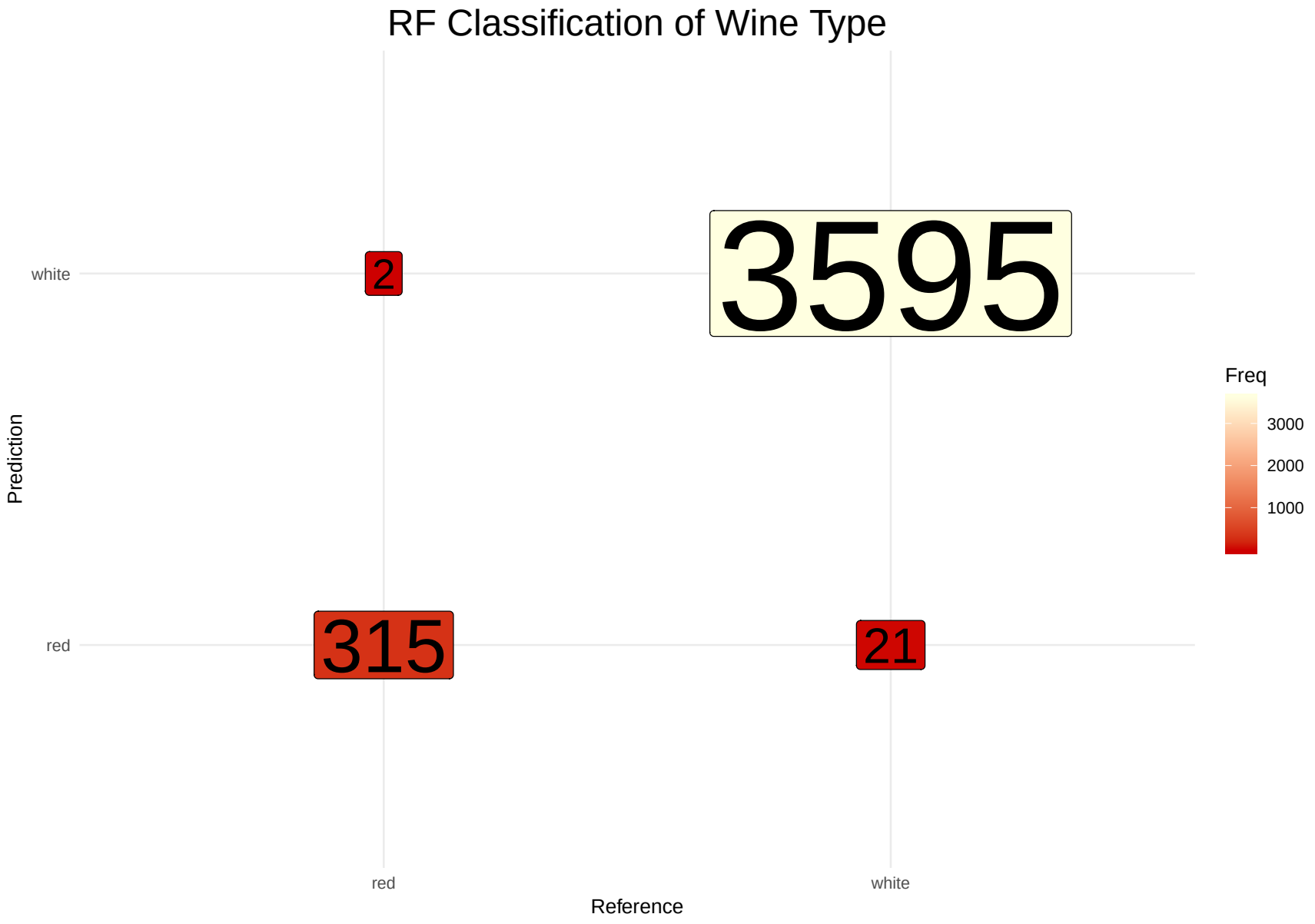
Results: Random Forests (wine type classification)

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

Table 1: Classification Results for Wine Type

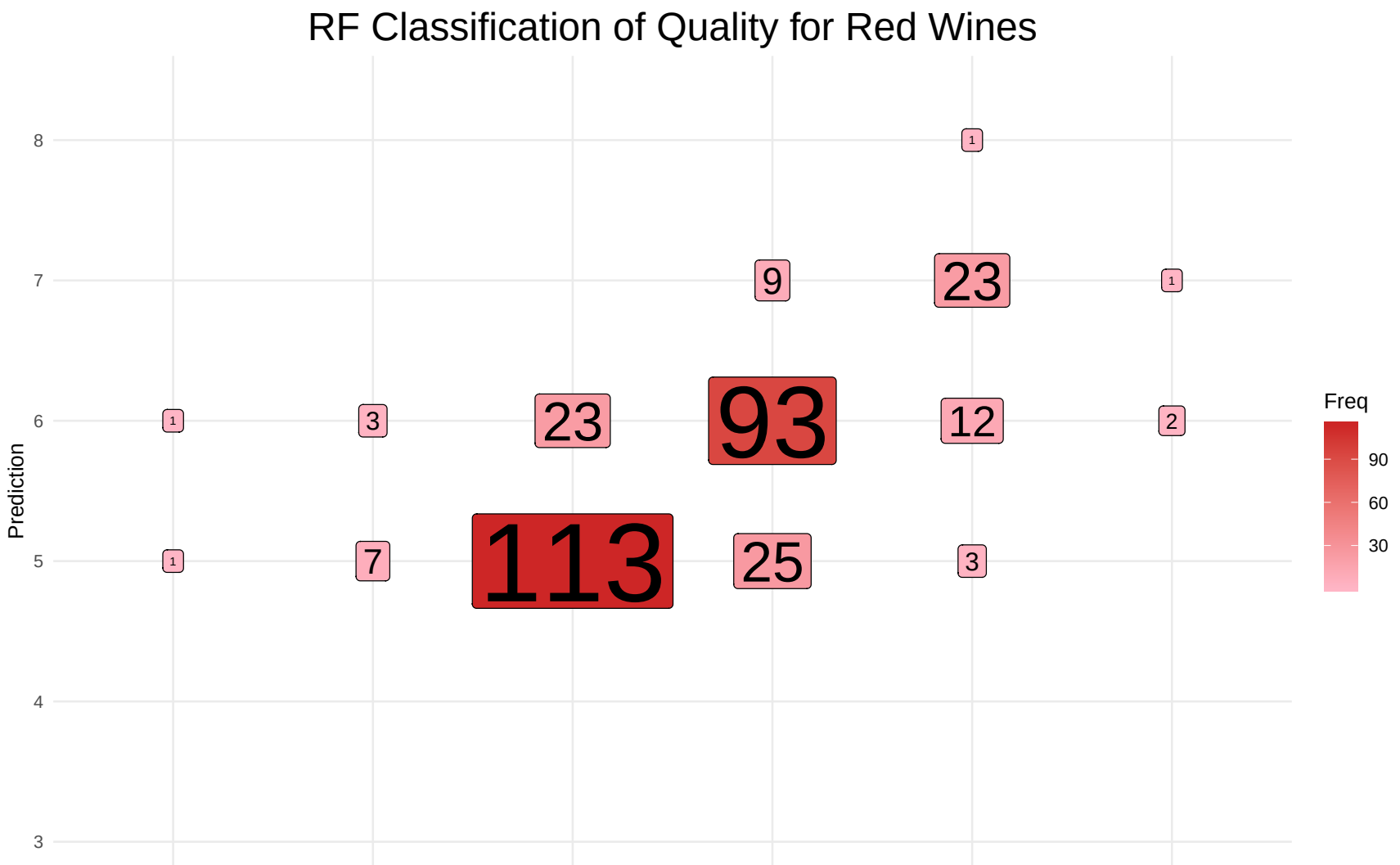
	Overall Results		Percent Correct by Category	
	Prediction Accuracy	Kappa	Red	White
Random Forest	0.9942	0.9616	99.37	99.42

Results: Random Forests (wine type classification)

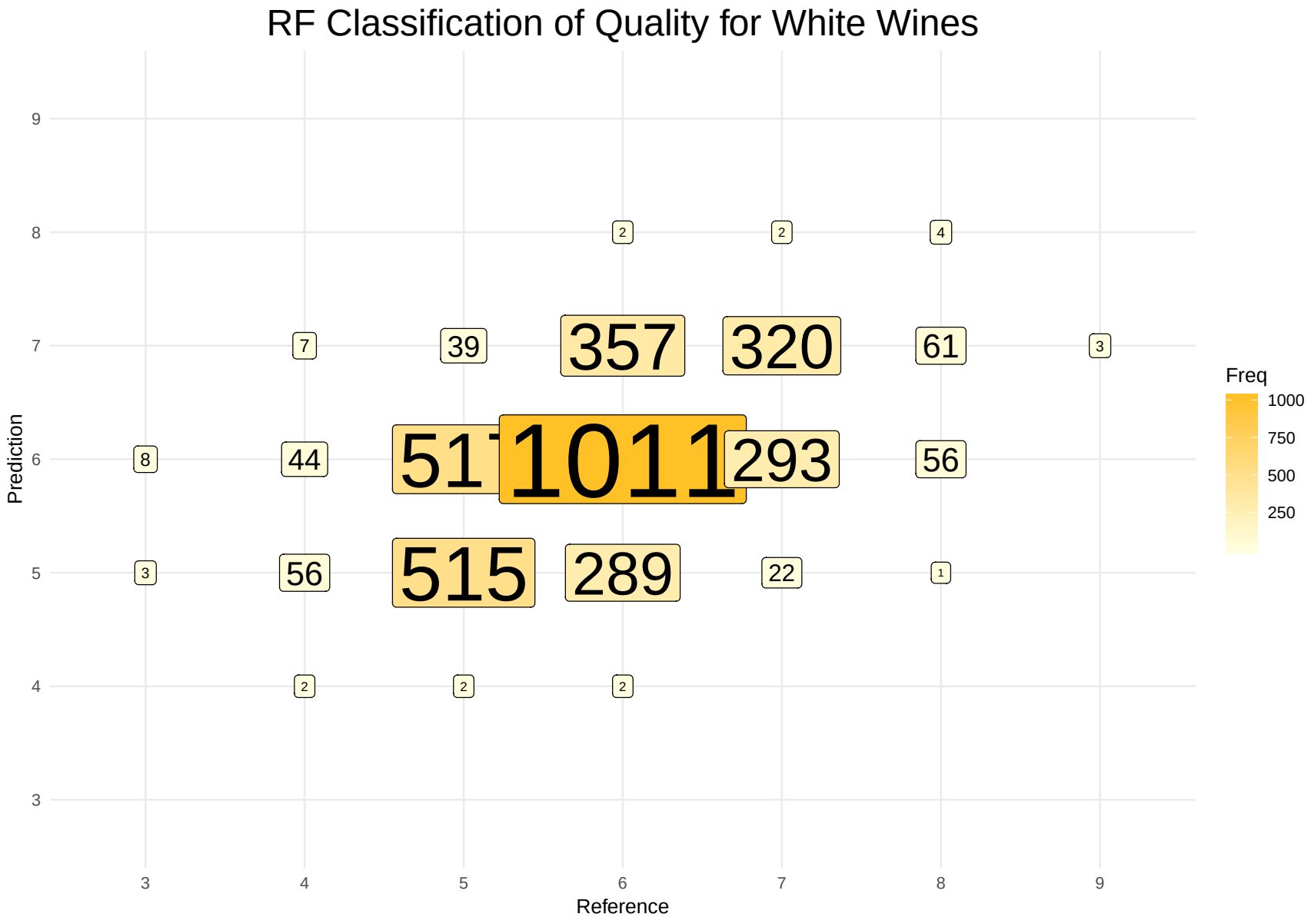


Results: Random Forests (red wine quality classification)

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.

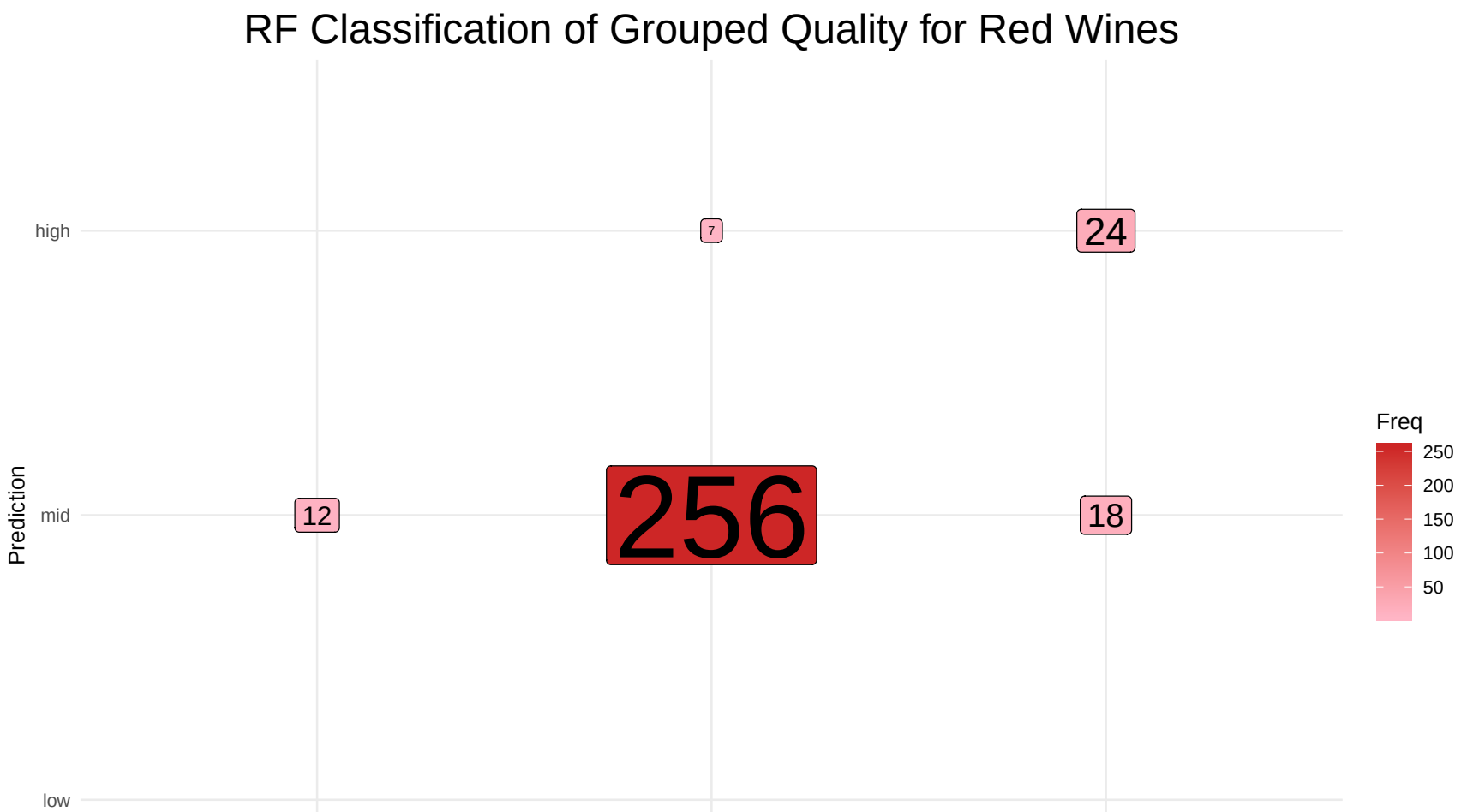


Results: Random Forests (white wine quality classification)

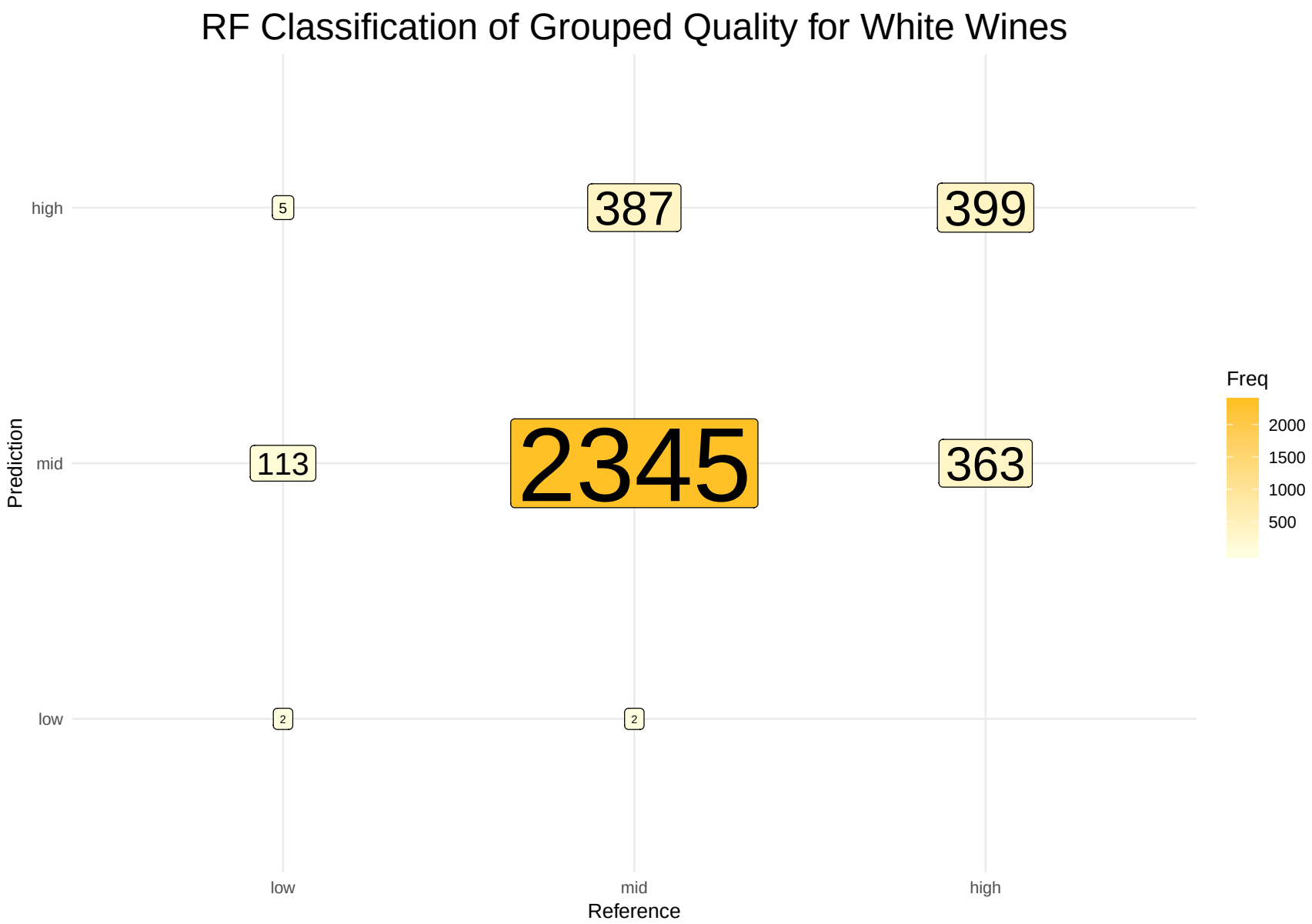


Results: Random Forests (grouped red wine quality classification)

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.

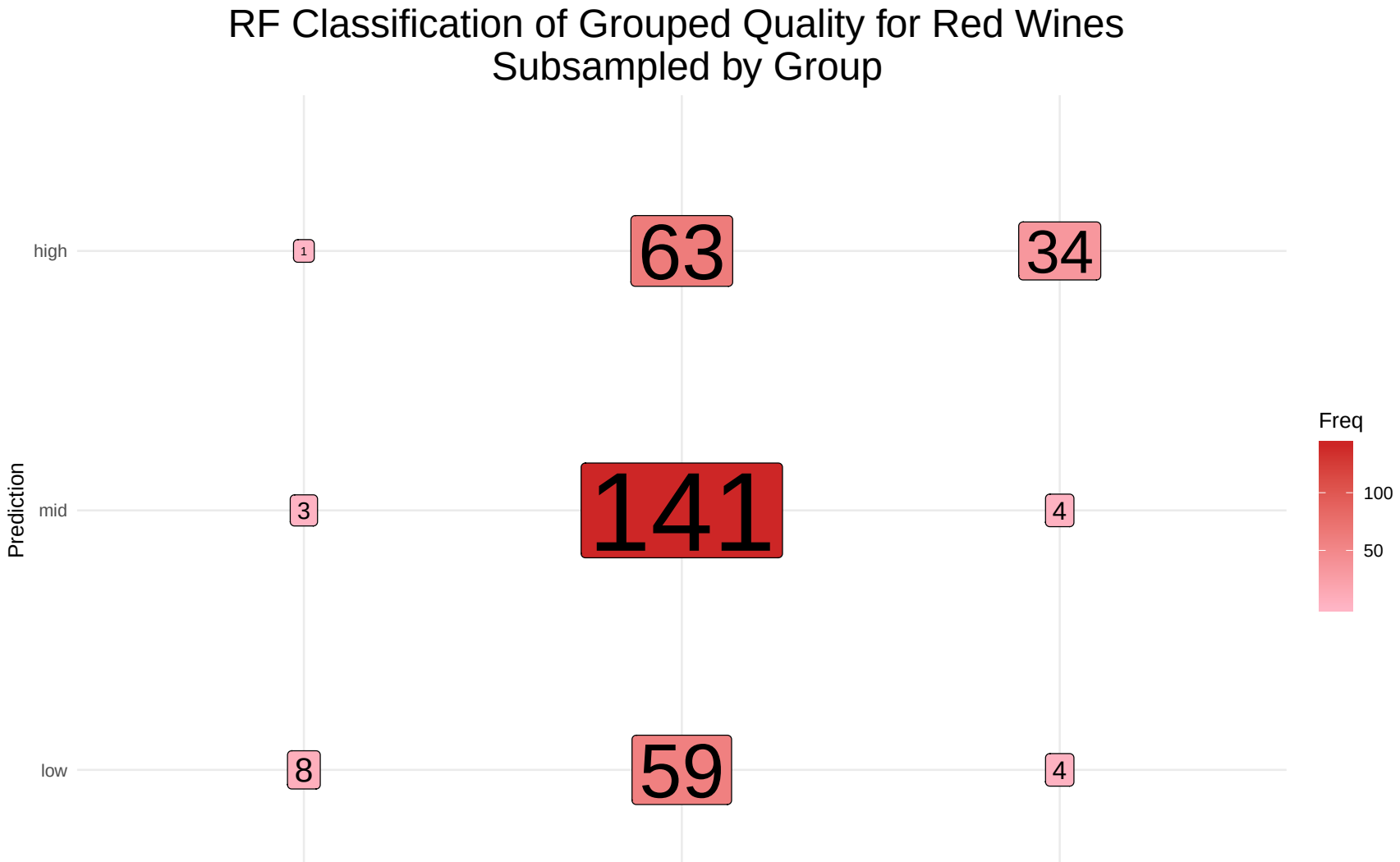


Results: Random Forests (grouped white wine quality classification)

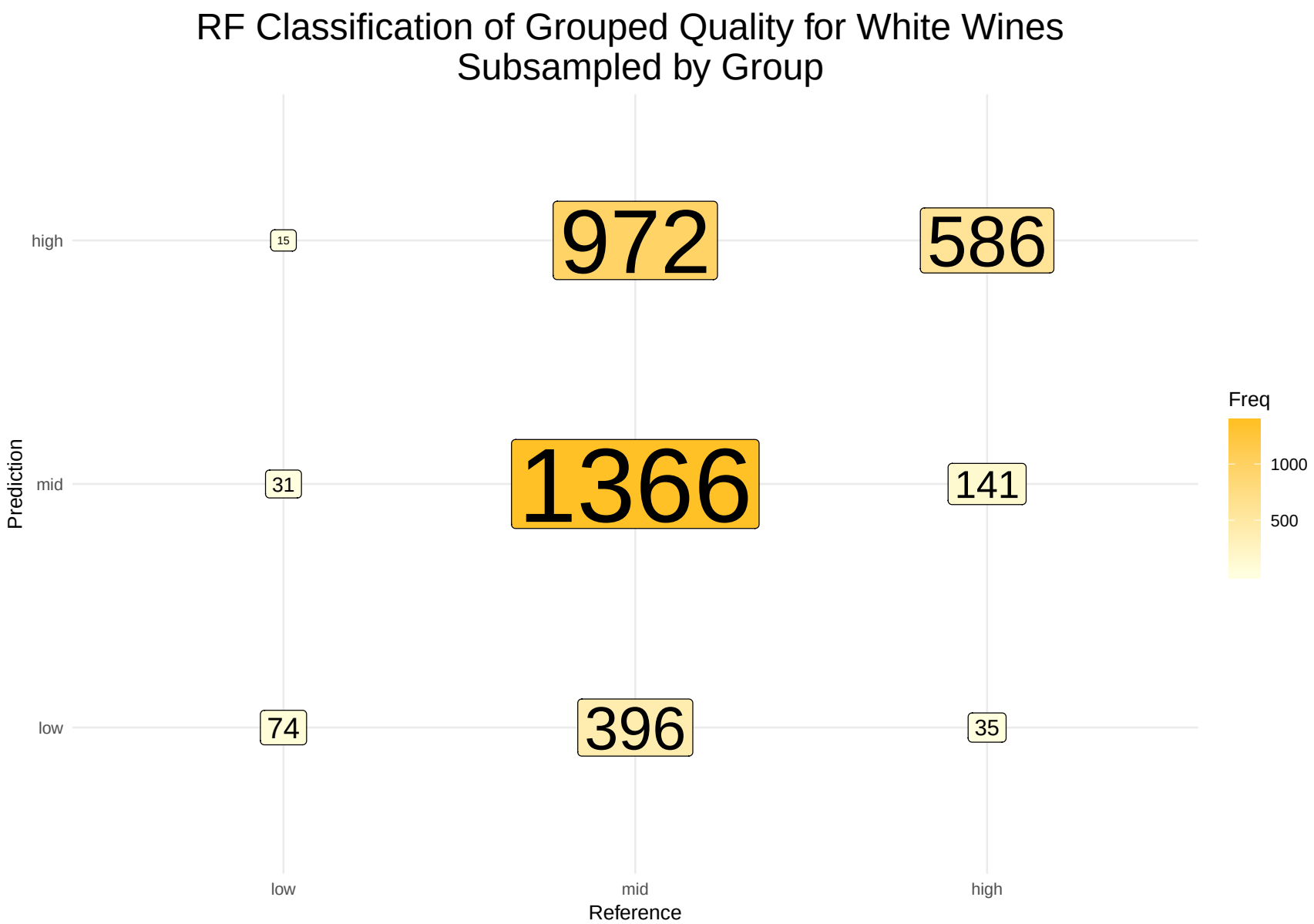


Results: Random Forests with Subsampling (grouped red wine quality classification)

We now perform the same grouped classification, but with subsampling to balance the groups.



Results: Random Forests with Subsampling (grouped red wine quality classification)



Comparison of Results: Red Wine, Ungrouped

Table 2: Comparison of Results for Red Wine

	Overall Results		Percent Correct by Category					
	Prediction Accuracy	Kappa	3	4	5	6	7	8
Random Forest	0.7224	0.5495	0	0	83.09	73.23	58.97	0
Method 2	NA	NA	NA	NA	NA	NA	NA	NA
Method 3	NA	NA	NA	NA	NA	NA	NA	NA
Method 4	NA	NA	NA	NA	NA	NA	NA	NA

Comparison of Results: White Wine, Ungrouped

Table 3: Comparison of Results for White Wine

	Overall Results		Percent Correct by Category						
	Prediction Accuracy	Kappa	3	4	5	6	7	8	9
Random Forest	0.5122	0.2422	0	1.83	48	60.87	50.24	3.28	0
Method 2	NA	NA	NA	NA	NA	NA	NA	NA	NA
Method 3	NA	NA	NA	NA	NA	NA	NA	NA	NA
Method 4	NA	NA	NA	NA	NA	NA	NA	NA	NA

Comparison of Results: Red Wine, Grouped

Table 4: Comparison of Results for Red Wine with Grouped Quality

	Overall Results		Percent Correct by Category		
	Prediction Accuracy	Kappa	Low	Mid	High
Random Forest	0.8833	0.5107	0.00	97.34	57.14
RF Subsampled	0.5773	0.2495	66.67	53.61	80.95
Method 3	NA	NA	NA	NA	NA
Method 4	NA	NA	NA	NA	NA
Method 5	NA	NA	NA	NA	NA

Comparison of Results: White Wine, Grouped

Table 5: Comparison of Results for White Wine with Grouped Quality

	Overall Results		Percent Correct by Category		
	Prediction Accuracy	Kappa	Low	Mid	High
Random Forest	0.7594	0.3390	1.67	85.77	52.36
RF Subsampled	0.5603	0.2446	61.67	49.96	76.90
Method 3	NA	NA	NA	NA	NA
Method 4	NA	NA	NA	NA	NA
Method 5	NA	NA	NA	NA	NA

Discussion

(about interpretation or possible future directions)