Grade A Wine EDA/Pre-processing

Raam Pravin

2024-12-28

Importing necessary libraries, importing data, and readying data for analysis

```
setwd("~/Downloads")
redwine <- read.csv("winequality-red.csv")
whitewine <- read.csv("winequality-white.csv")

cat("Creating variable Names Red Wine")
```

```
## Creating variable Names Red Wine
```

```
redwine_seperated <- str_split_fixed(redwine$fixed.acidity.volatile.acidity.citric.acid.residual.sugar.

redwine_seperated <- data.frame(redwine_seperated)

cat("Creating variable Names White Wine")
```

```
## Creating variable Names White Wine
```

```
whitewine_seperated <- str_split_fixed(whitewine$fixed.acidity.volatile.acidity.citric.acid.residual.su

whitewine_seperated <- data.frame(whitewine_seperated)



redwine_seperated <- redwine_seperated %>%
  rename(fixed_acidity = 'X1')
redwine_seperated <- redwine_seperated %>%
  rename(volatile_acidity = 'X2')
redwine_seperated <- redwine_seperated %>%
  rename(citric_acid = 'X3')
redwine_seperated <- redwine_seperated %>%
  rename(residual_sugar = 'X4')
redwine_seperated <- redwine_seperated %>%
  rename(chlorides = 'X5')
redwine_seperated <- redwine_seperated %>%
  rename(free_sulfur_dioxide = 'X6')
redwine_seperated <- redwine_seperated %>%
  rename(total_sulfur_dioxide = 'X7')
redwine_seperated <- redwine_seperated %>%
  rename(density = 'X8')
redwine_seperated <- redwine_seperated %>%
  rename(pH = 'X9')
redwine_seperated <- redwine_seperated %>%
  rename(sulphates = 'X10')
```

```
redwine_seperated <- redwine_seperated %>%
  rename(alcohol = 'X11')
redwine_seperated <- redwine_seperated %>%
  rename(quality = 'X12')

whitewine_seperated <- str_split_fixed(whitewine$fixed.acidity.volatile.acidity.citric.acid.residual.su

whitewine_seperated <- data.frame(whitewine_seperated)

whitewine_seperated <- whitewine_seperated %>%
  rename(fixed_acidity = 'X1')
whitewine_seperated <- whitewine_seperated %>%
  rename(volatile_acidity = 'X2')
whitewine_seperated <- whitewine_seperated %>%
  rename(citric_acid = 'X3')
whitewine_seperated <- whitewine_seperated %>%
  rename(residual_sugar = 'X4')
whitewine_seperated <- whitewine_seperated %>%
  rename(chlorides = 'X5')
whitewine_seperated <- whitewine_seperated %>%
  rename(free_sulfur_dioxide = 'X6')
whitewine_seperated <- whitewine_seperated %>%
  rename(total_sulfur_dioxide = 'X7')
whitewine_seperated <- whitewine_seperated %>%
  rename(density = 'X8')
whitewine_seperated <- whitewine_seperated %>%
  rename(pH = 'X9')
whitewine_seperated <- whitewine_seperated %>%
  rename(sulphates = 'X10')
whitewine_seperated <- whitewine_seperated %>%
  rename(alcohol = 'X11')
whitewine_seperated <- whitewine_seperated %>%
  rename(quality = 'X12')


redwine_seperated <- apply(redwine_seperated,2,as.numeric)
whitewine_seperated <- apply(whitewine_seperated,2,as.numeric)

redwine_seperated <- data.frame(redwine_seperated)
whitewine_seperated <- data.frame(whitewine_seperated)



redwine_seperated$type <- 'red'
whitewine_seperated$type <- 'white'

redwine_seperated$type <- as.factor(redwine_seperated$type)
whitewine_seperated$type <- as.factor(whitewine_seperated$type)

wine <- full_join(redwine_seperated,whitewine_seperated)
```

Exploratory Data Analysis Getting Descriptive Statistics for Red and White Wine

```r
#Getting descriptive stats for Red Wine
skim(redwine_seperated)
```

Table 1: Data summary

Name	redwine_seperated
Number of rows	1599
Number of columns	13
Column type frequency:	
factor	1
numeric	12
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
type	0	1	FALSE	1	red: 1599

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
fixed_acidity	0	1	8.32	1.74	4.60	7.10	7.90	9.20	15.90	
volatile_acidity	0	1	0.53	0.18	0.12	0.39	0.52	0.64	1.58	
citric_acid	0	1	0.27	0.19	0.00	0.09	0.26	0.42	1.00	
residual_sugar	0	1	2.54	1.41	0.90	1.90	2.20	2.60	15.50	
chlorides	0	1	0.09	0.05	0.01	0.07	0.08	0.09	0.61	
free_sulfur_dioxide	0	1	15.87	10.46	1.00	7.00	14.00	21.00	72.00	
total_sulfur_dioxide	0	1	46.47	32.90	6.00	22.00	38.00	62.00	289.00	
density	0	1	1.00	0.00	0.99	1.00	1.00	1.00	1.00	
pH	0	1	3.31	0.15	2.74	3.21	3.31	3.40	4.01	
sulphates	0	1	0.66	0.17	0.33	0.55	0.62	0.73	2.00	
alcohol	0	1	10.42	1.07	8.40	9.50	10.20	11.10	14.90	
quality	0	1	5.64	0.81	3.00	5.00	6.00	6.00	8.00	

```r
#Basic Stats
summary(redwine_seperated)
```

```
##  fixed_acidity   volatile_acidity  citric_acid    residual_sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
##    chlorides       free_sulfur_dioxide total_sulfur_dioxide    density
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956
##  Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968
```

```
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037
##        pH          sulphates         alcohol         quality       type
##  Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000   red:1599
##  1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.310   Median :0.6200   Median :10.20   Median :6.000
##  Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636
##  3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

```
#Getting descriptive stats for White Wine
skim(whitewine_seperated)
```

Table 4: Data summary

Name	whitewine_seperated
Number of rows	4898
Number of columns	13
Column type frequency:	
factor	1
numeric	12
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
type	0	1	FALSE	1	whi: 4898

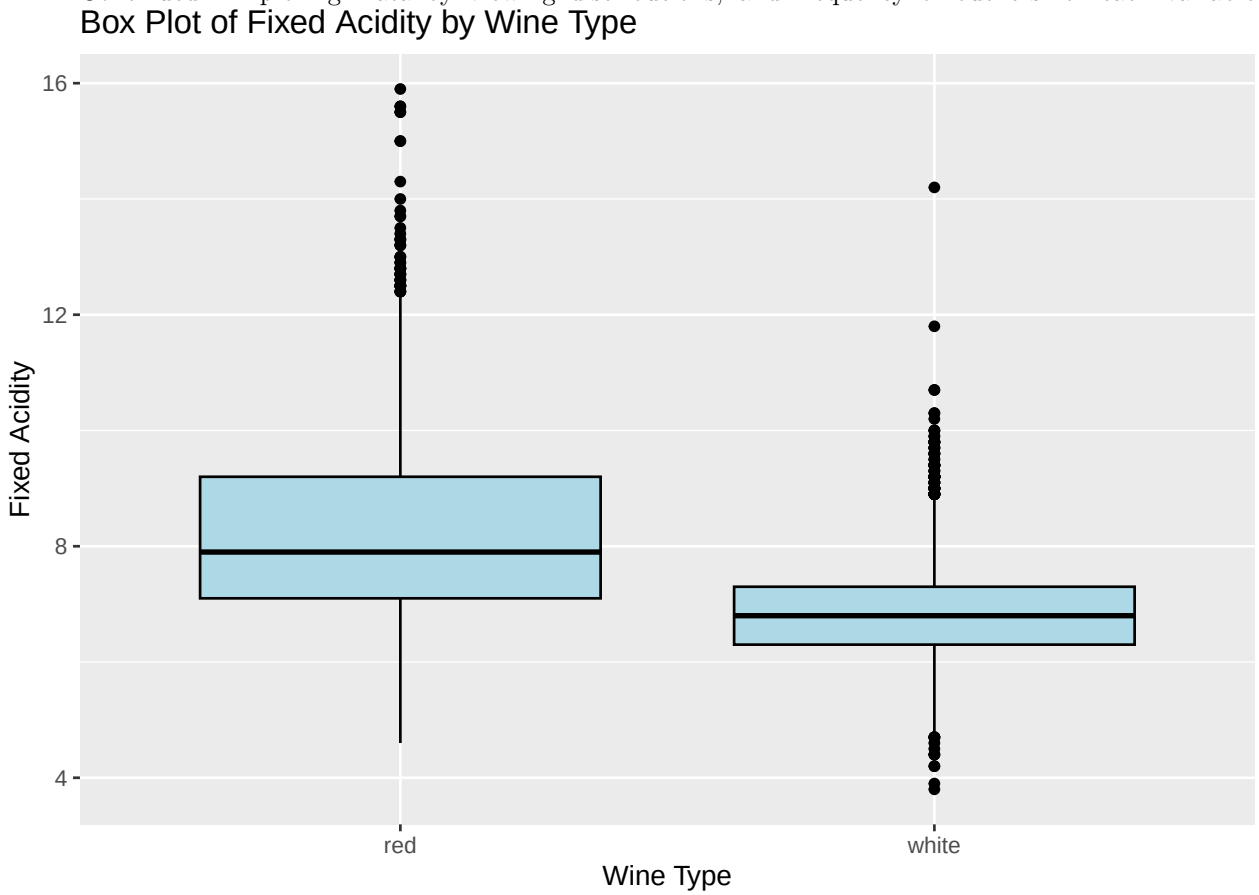
Variable type: numeric

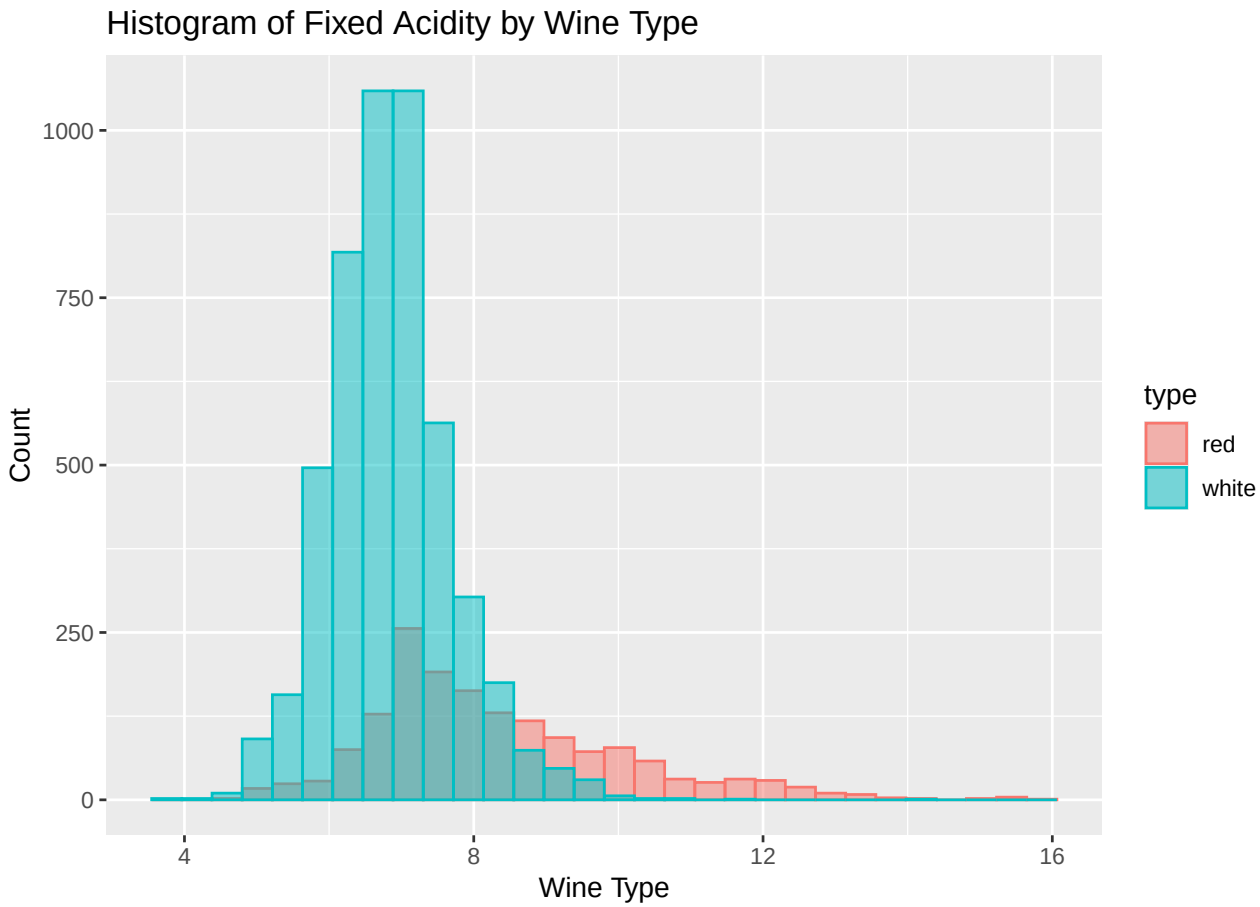
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
fixed_acidity	0	1	6.85	0.84	3.80	6.30	6.80	7.30	14.20	
volatile_acidity	0	1	0.28	0.10	0.08	0.21	0.26	0.32	1.10	
citric_acid	0	1	0.33	0.12	0.00	0.27	0.32	0.39	1.66	
residual_sugar	0	1	6.39	5.07	0.60	1.70	5.20	9.90	65.80	
chlorides	0	1	0.05	0.02	0.01	0.04	0.04	0.05	0.35	
free_sulfur_dioxide	0	1	35.31	17.01	2.00	23.00	34.00	46.00	289.00	
total_sulfur_dioxide	0	1	138.36	42.50	9.00	108.00	134.00	167.00	440.00	
density	0	1	0.99	0.00	0.99	0.99	0.99	1.00	1.04	
pH	0	1	3.19	0.15	2.72	3.09	3.18	3.28	3.82	
sulphates	0	1	0.49	0.11	0.22	0.41	0.47	0.55	1.08	
alcohol	0	1	10.51	1.23	8.00	9.50	10.40	11.40	14.20	
quality	0	1	5.88	0.89	3.00	5.00	6.00	6.00	9.00	

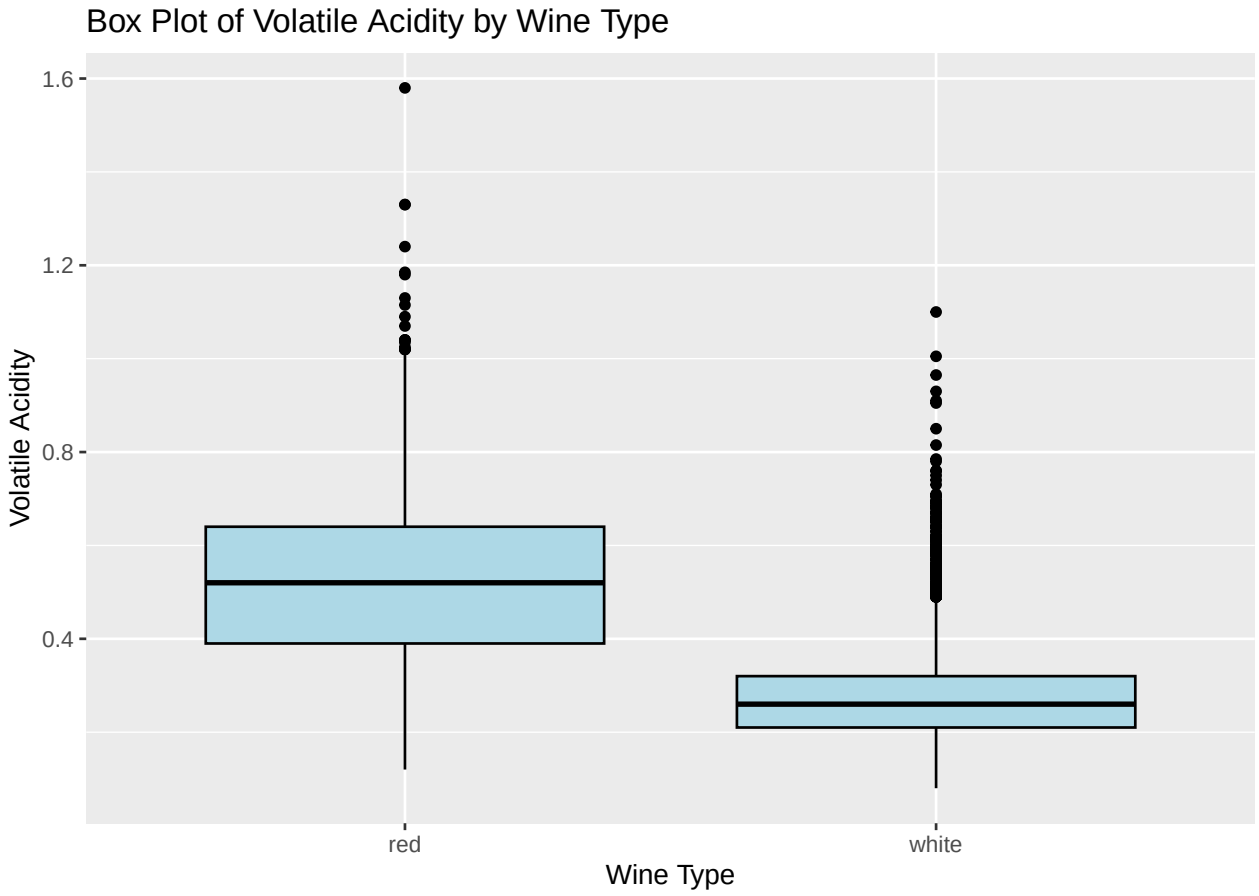
```
#Basic Stats
summary(whitewine_seperated)
```

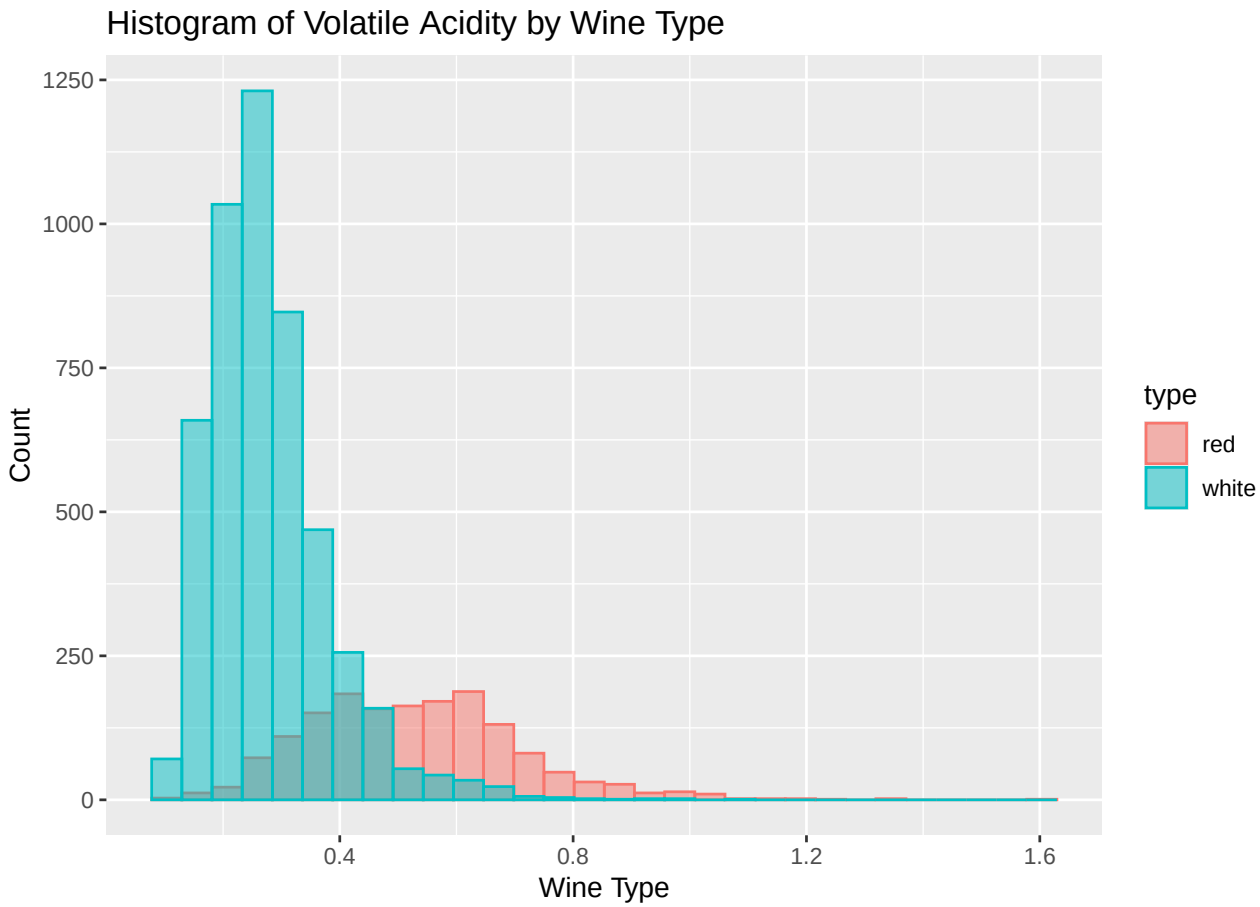
```
##  fixed_acidity    volatile_acidity  citric_acid     residual_sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
##  Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
##  Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391
##  3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900
##  Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800
##    chlorides       free_sulfur_dioxide total_sulfur_dioxide    density
##  Min.   :0.00900   Min.   :  2.00      Min.   :  9.0        Min.   :0.9871
##  1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0        1st Qu.:0.9917
##  Median :0.04300   Median : 34.00      Median :134.0        Median :0.9937
##  Mean   :0.04577   Mean   : 35.31      Mean   :138.4        Mean   :0.9940
##  3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0        3rd Qu.:0.9961
##  Max.   :0.34600   Max.   :289.00      Max.   :440.0        Max.   :1.0390
##        pH          sulphates         alcohol         quality         type
##  Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000   white:4898
##  1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.180   Median :0.4700   Median :10.40   Median :6.000
##  Mean   :3.188   Mean   :0.4898   Mean   :10.51   Mean   :5.878
##  3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40   3rd Qu.:6.000
##  Max.   :3.820   Max.   :1.0800   Max.   :14.20   Max.   :9.000
```

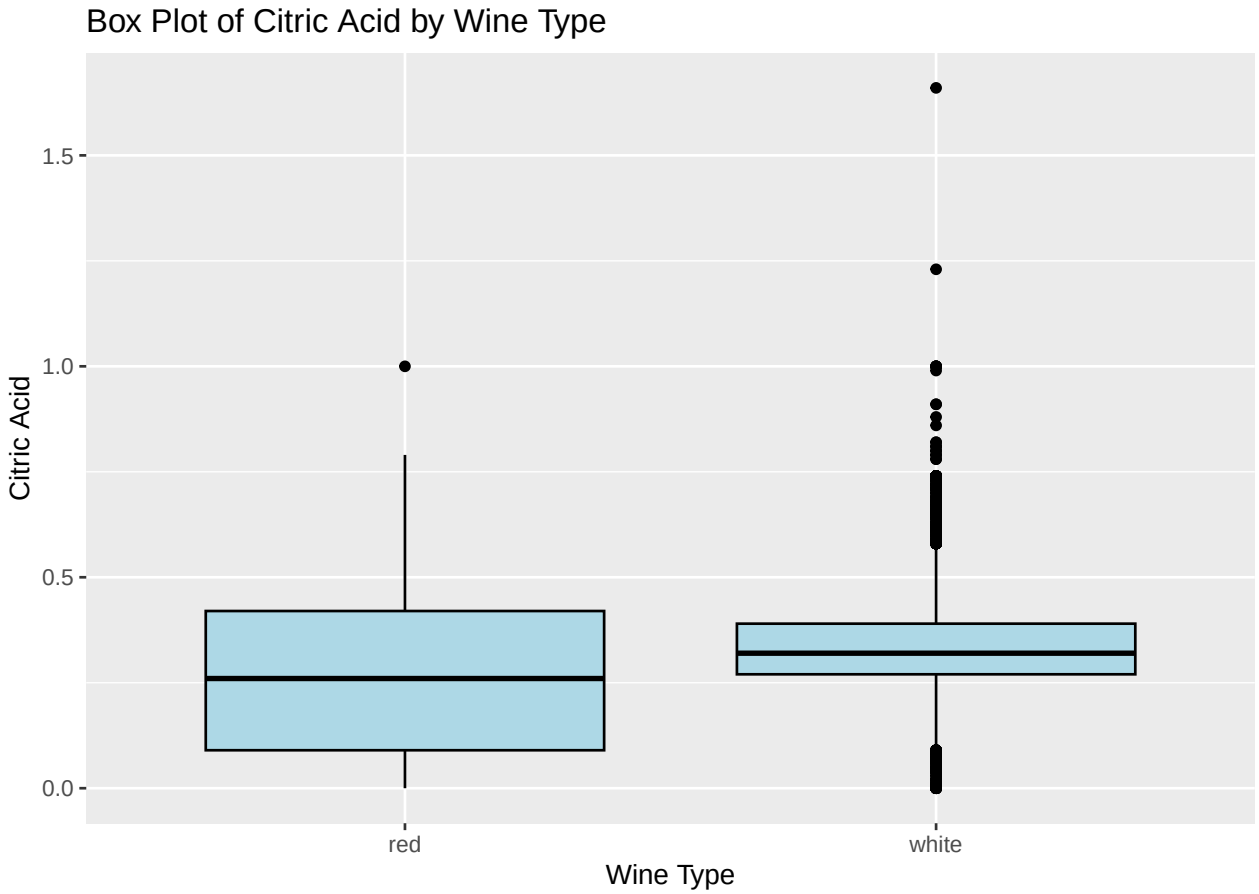
EDA Continued: Exploring Data by viewing distributions, and frequency of outliers for each variable

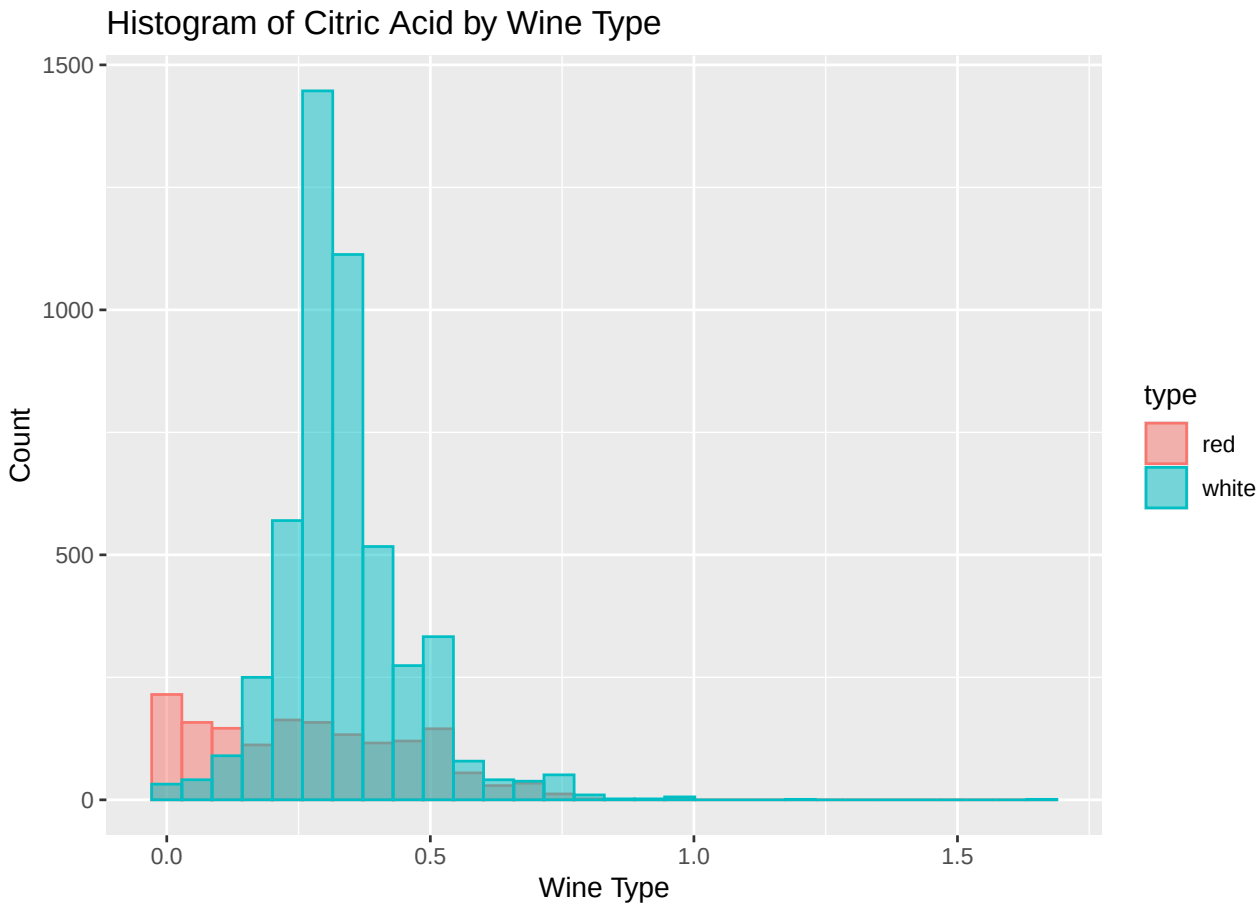


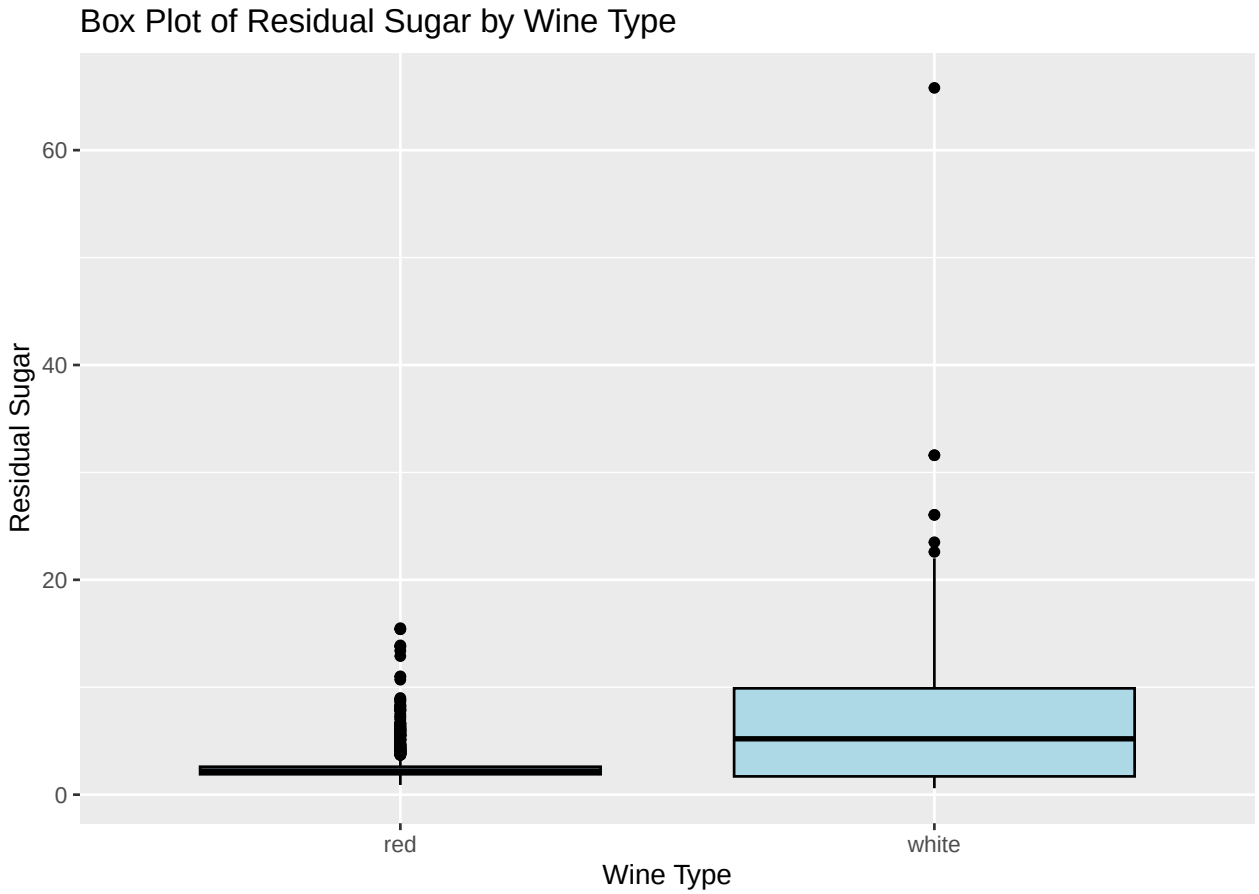

Histogram of Fixed Acidity by Wine Type
Count
1000
750
500
250
0
4
8
12
16
Wine Type
type
red
white

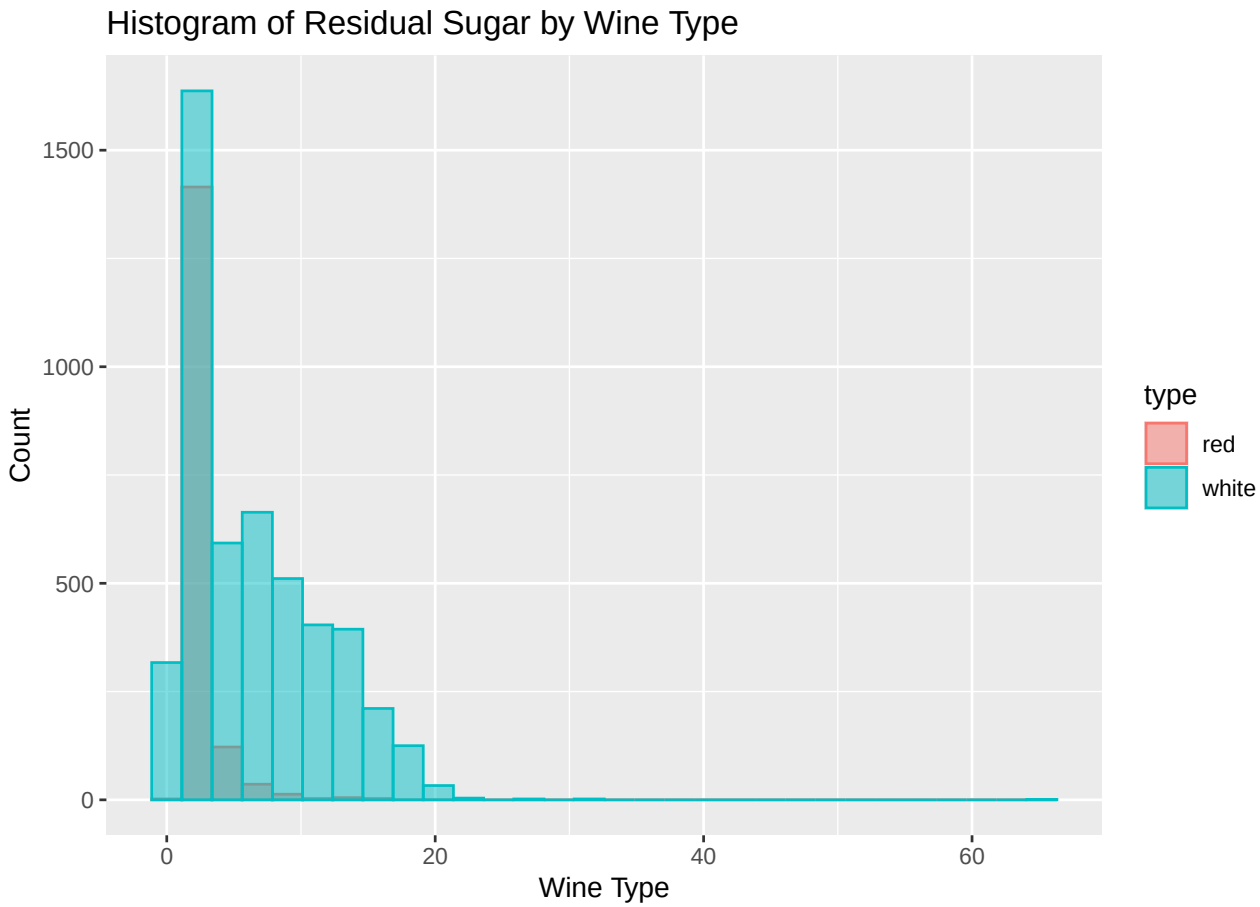

Box Plot of Volatile Acidity by Wine Type
Volatile Acidity
1.6
1.2
0.8
0.4
red
white
Wine Type

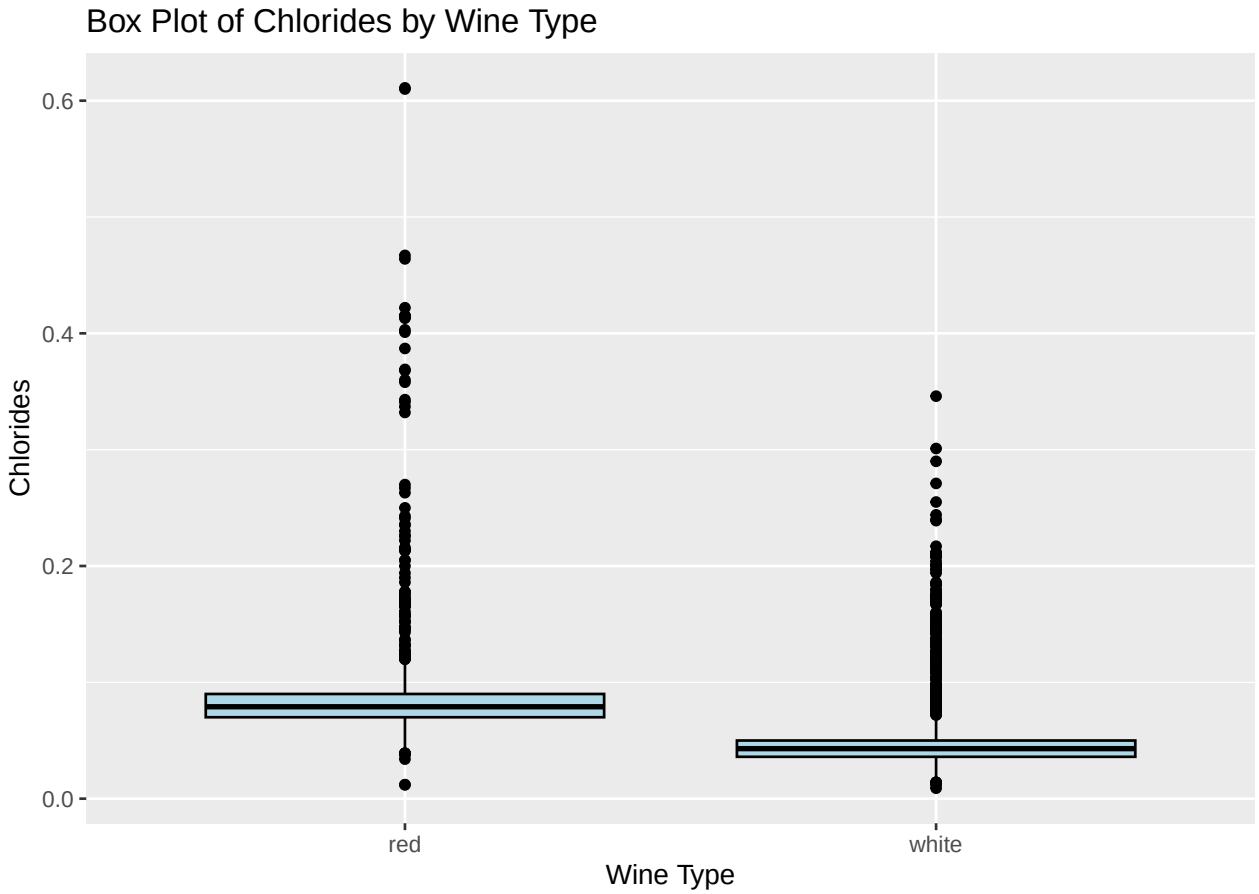

Histogram of Volatile Acidity by Wine Type
1250
1000
750
500
250
0
Count
0.4
0.8
1.2
1.6
Wine Type
type
red
white

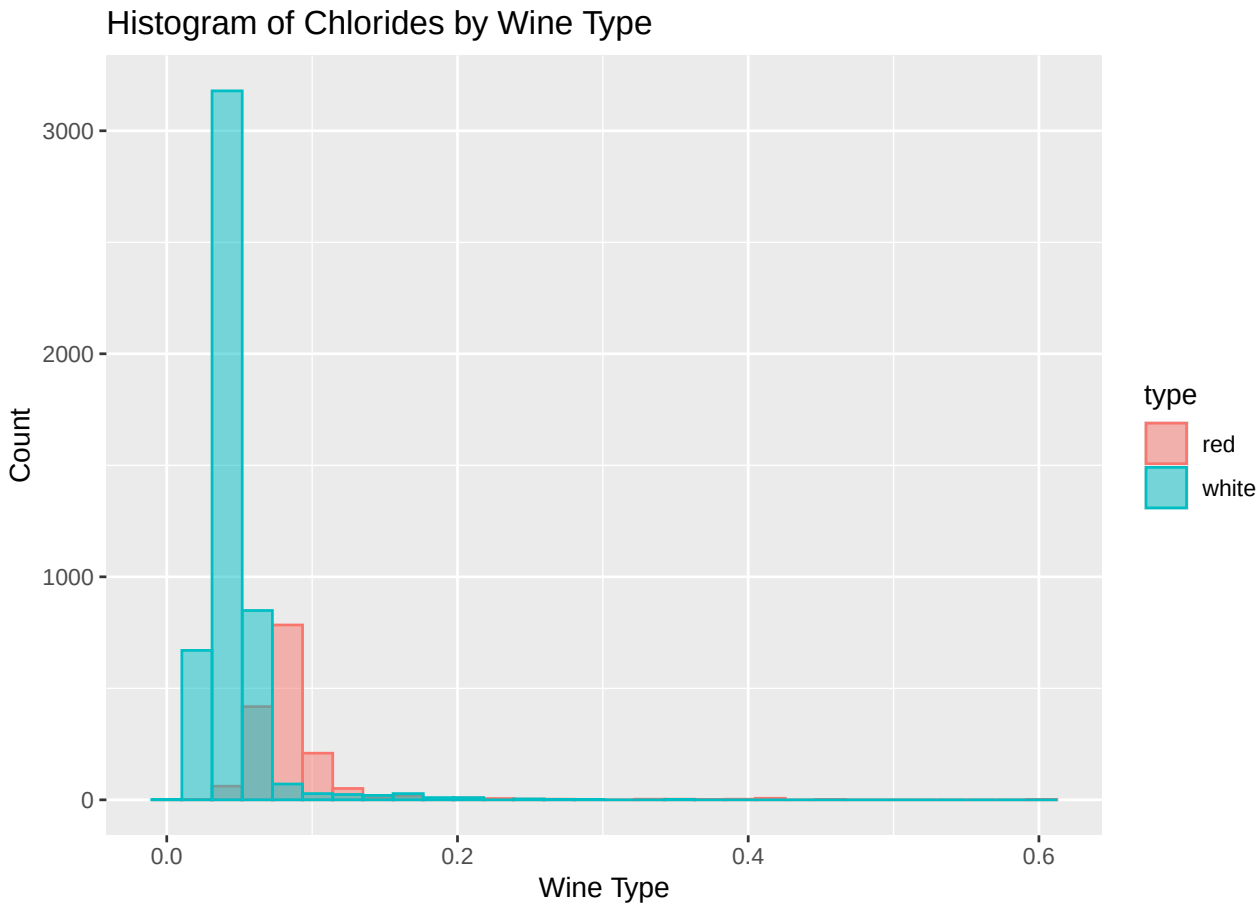

Box Plot of Citric Acid by Wine Type
Citric Acid
1.5
1.0
0.5
0.0
red
white
Wine Type

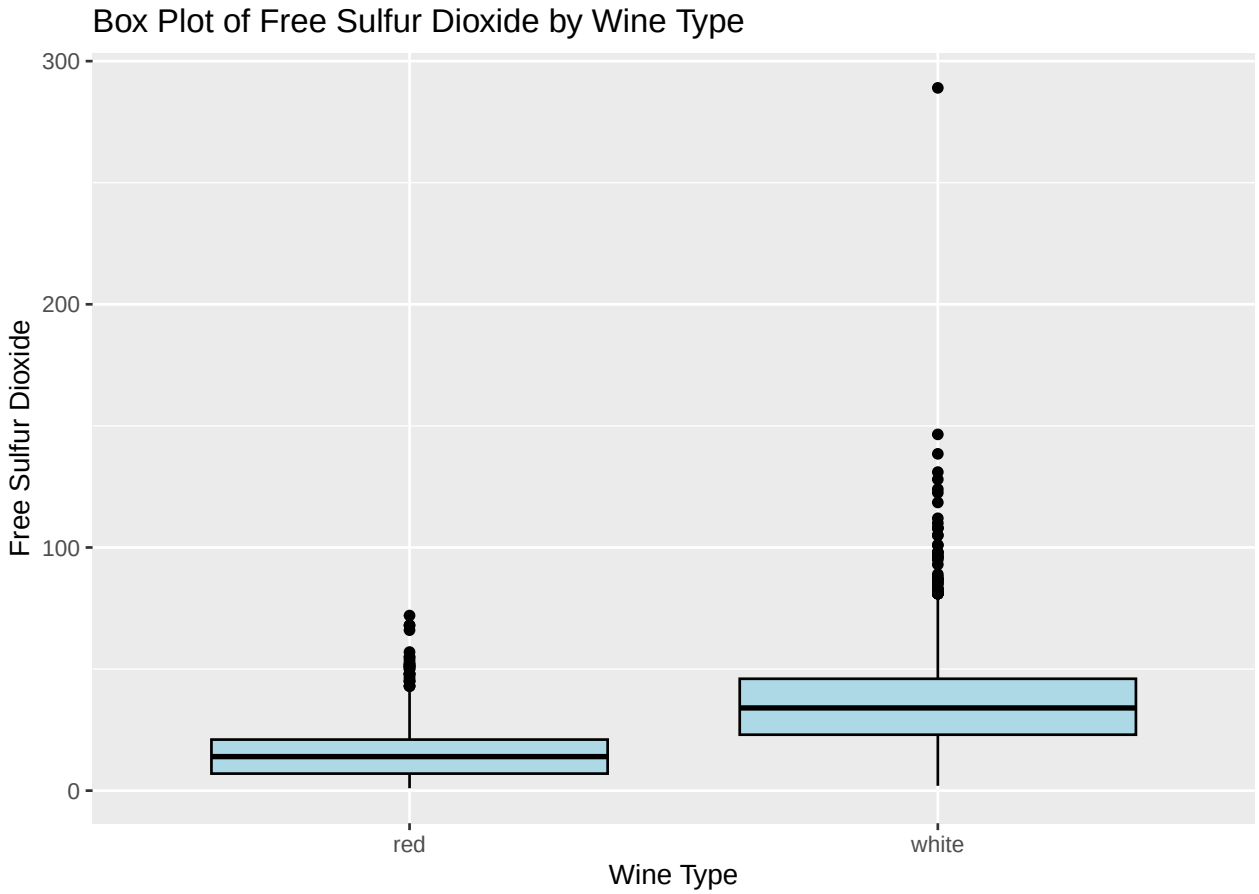

Histogram of Citric Acid by Wine Type
1500
1000
500
0
Count
0.0
0.5
1.0
1.5
Wine Type
type
red
white

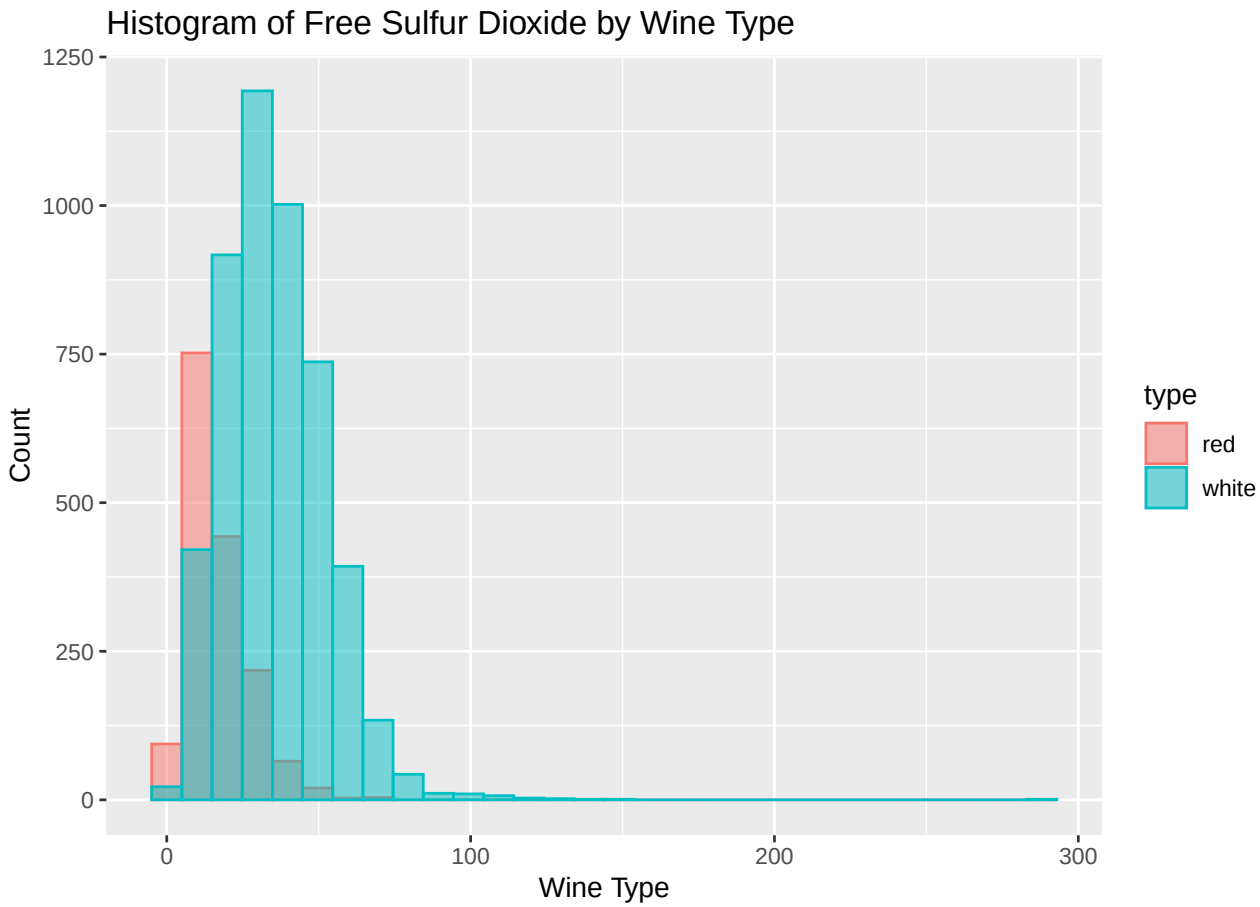

Box Plot of Residual Sugar by Wine Type
Residual Sugar
60
40
20
0
red
white
Wine Type

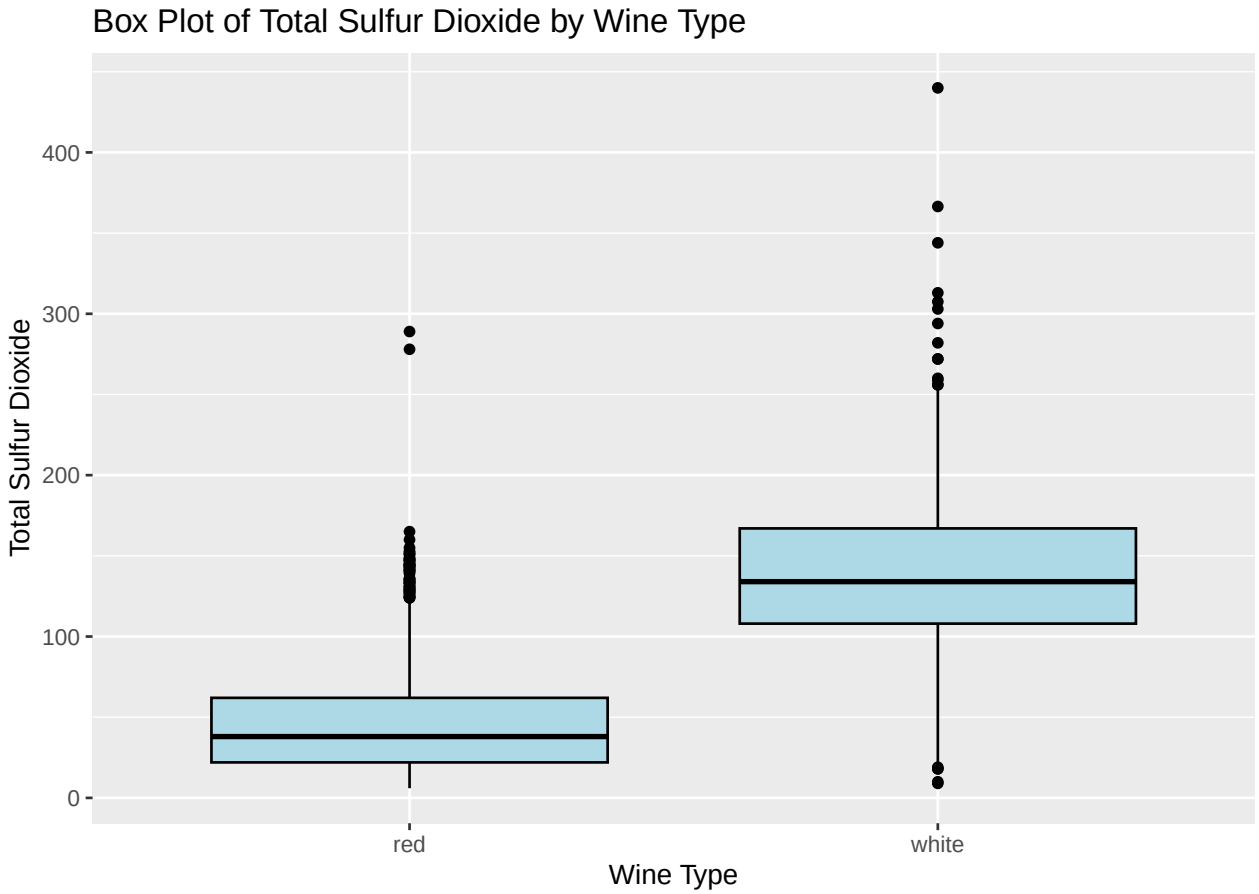

Histogram of Residual Sugar by Wine Type
Count
1500
1000
500
0
0
20
40
60
Wine Type
type
red
white

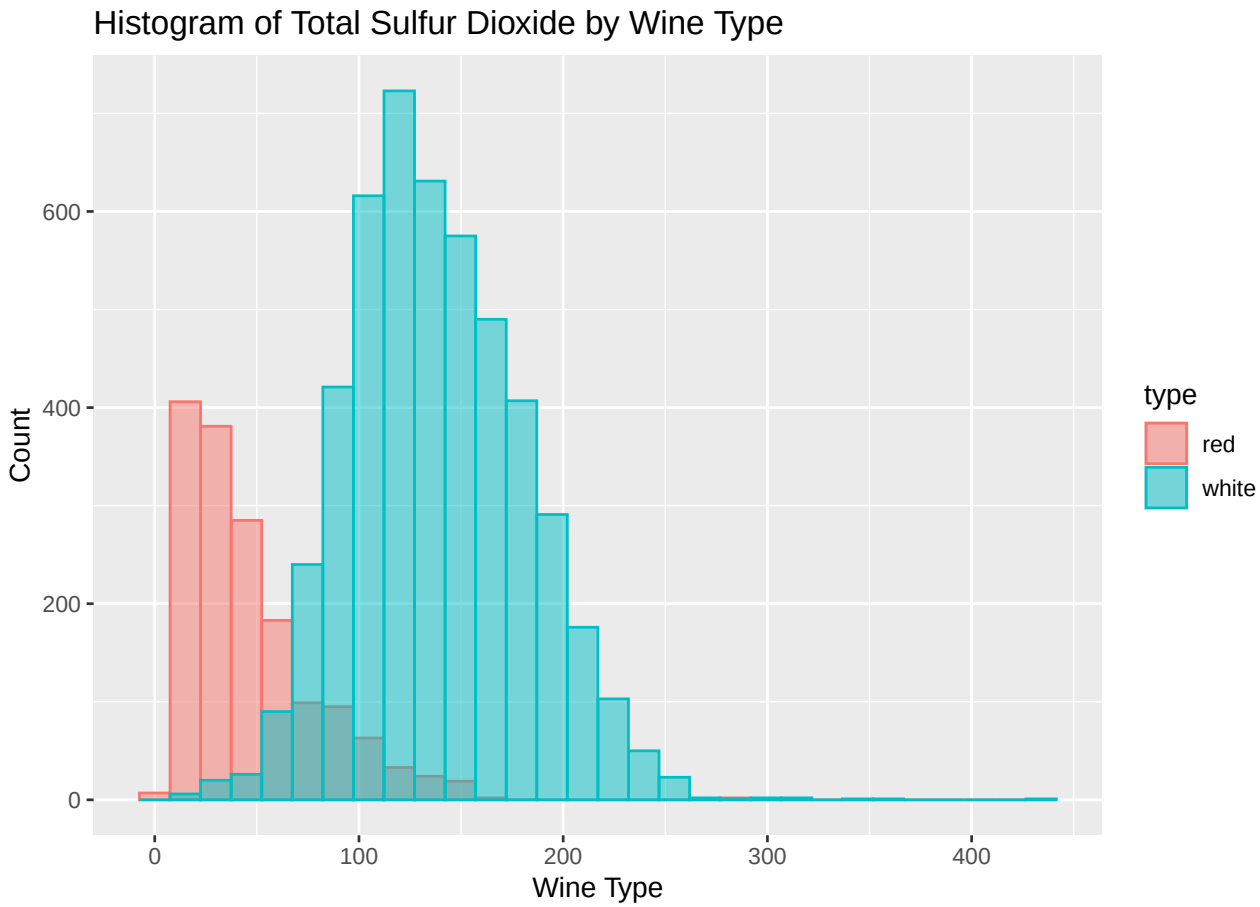

Box Plot of Chlorides by Wine Type
Chlorides
0.6
0.4
0.2
0.0
red
white
Wine Type

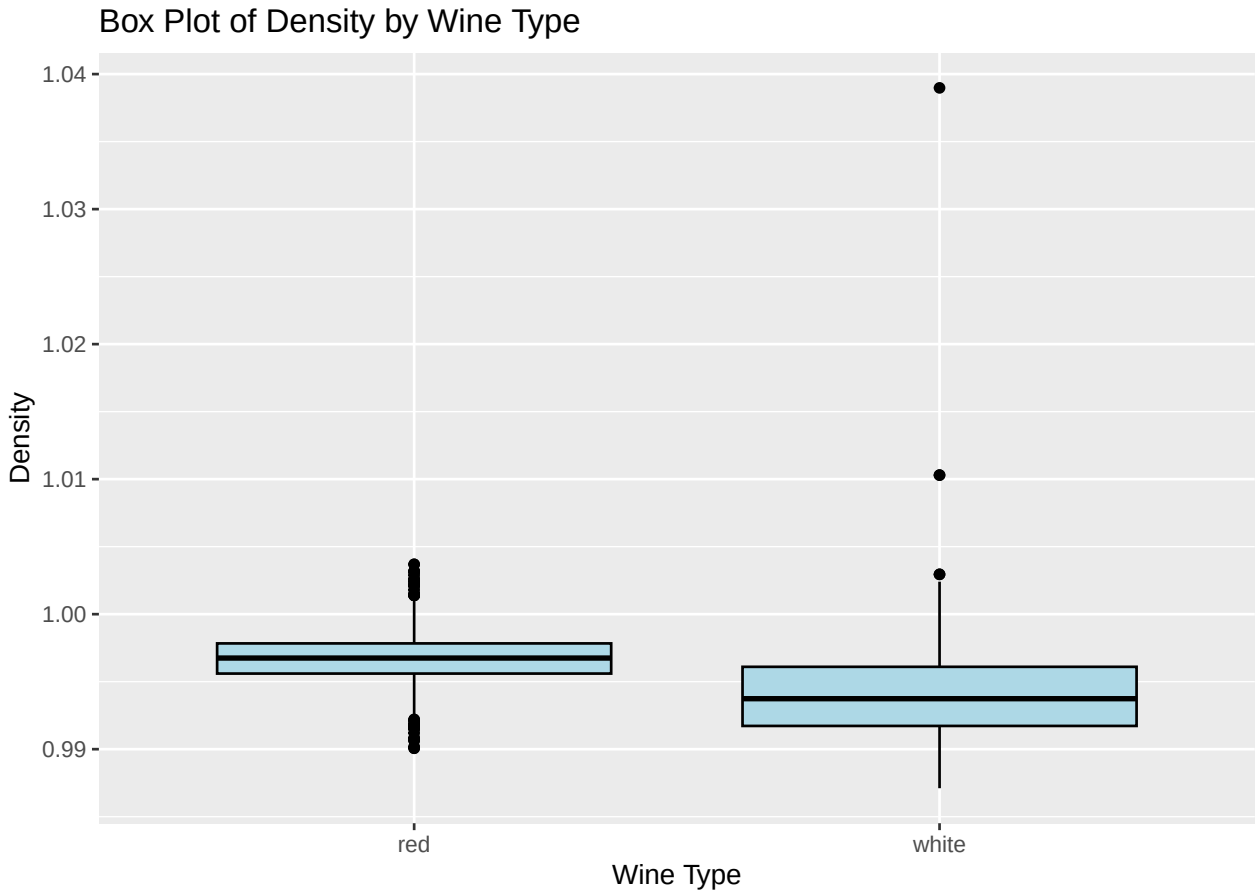

Histogram of Chlorides by Wine Type
Count
3000
2000
1000
0
0.0
0.2
0.4
0.6
Wine Type
type
red
white

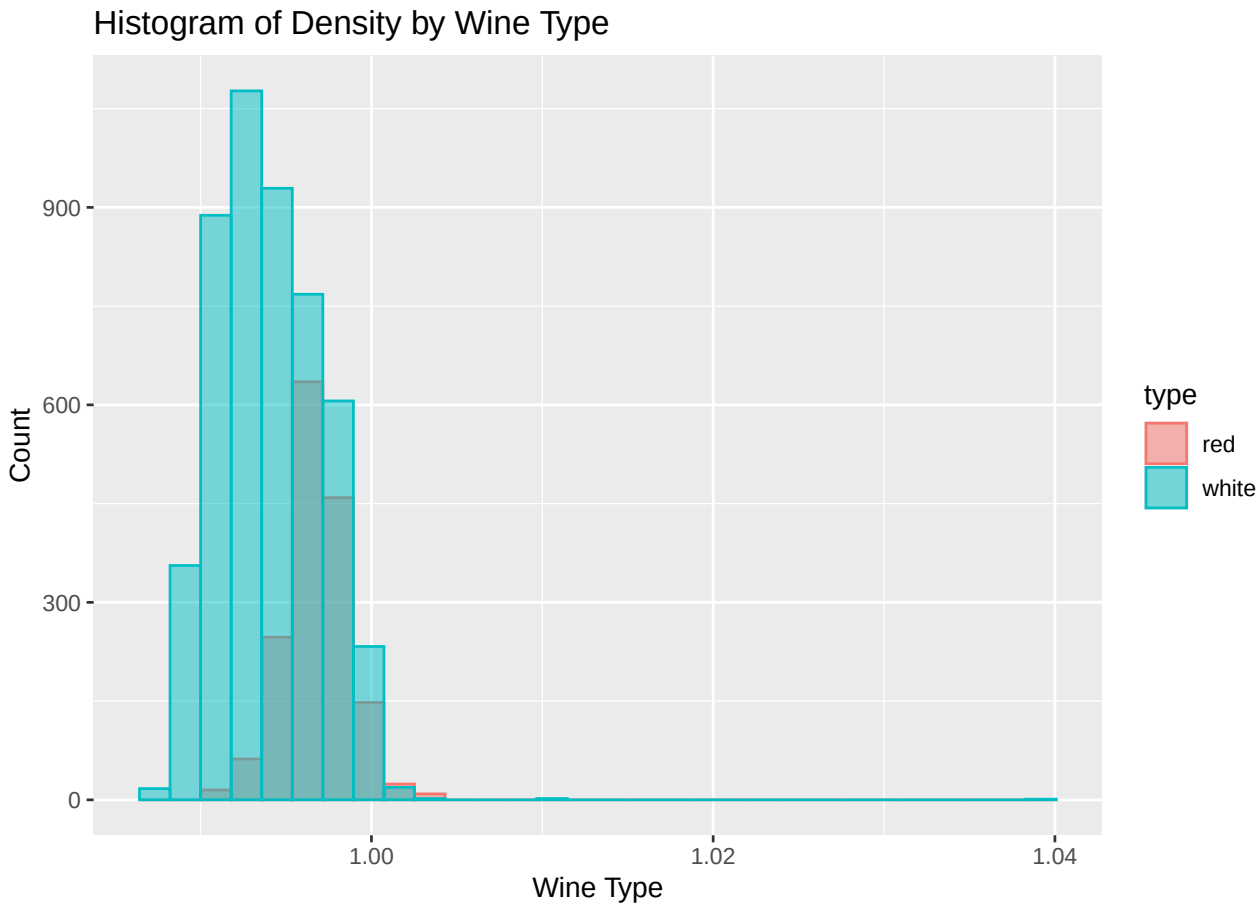

Box Plot of Free Sulfur Dioxide by Wine Type
Free Sulfur Dioxide
300
200
100
0
red
white
Wine Type

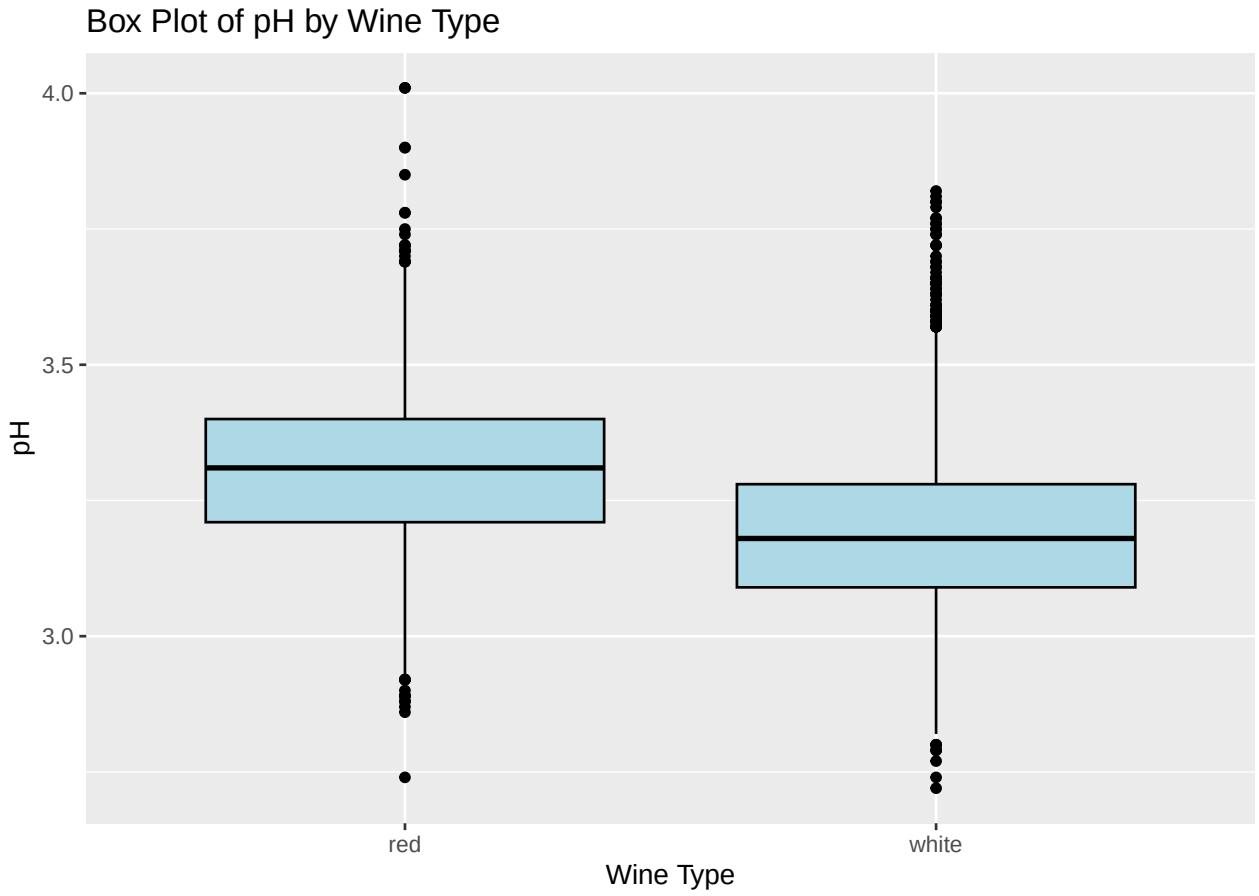

Histogram of Free Sulfur Dioxide by Wine Type
1250
1000
750
500
250
0
Count
0
100
200
300
Wine Type
type
red
white

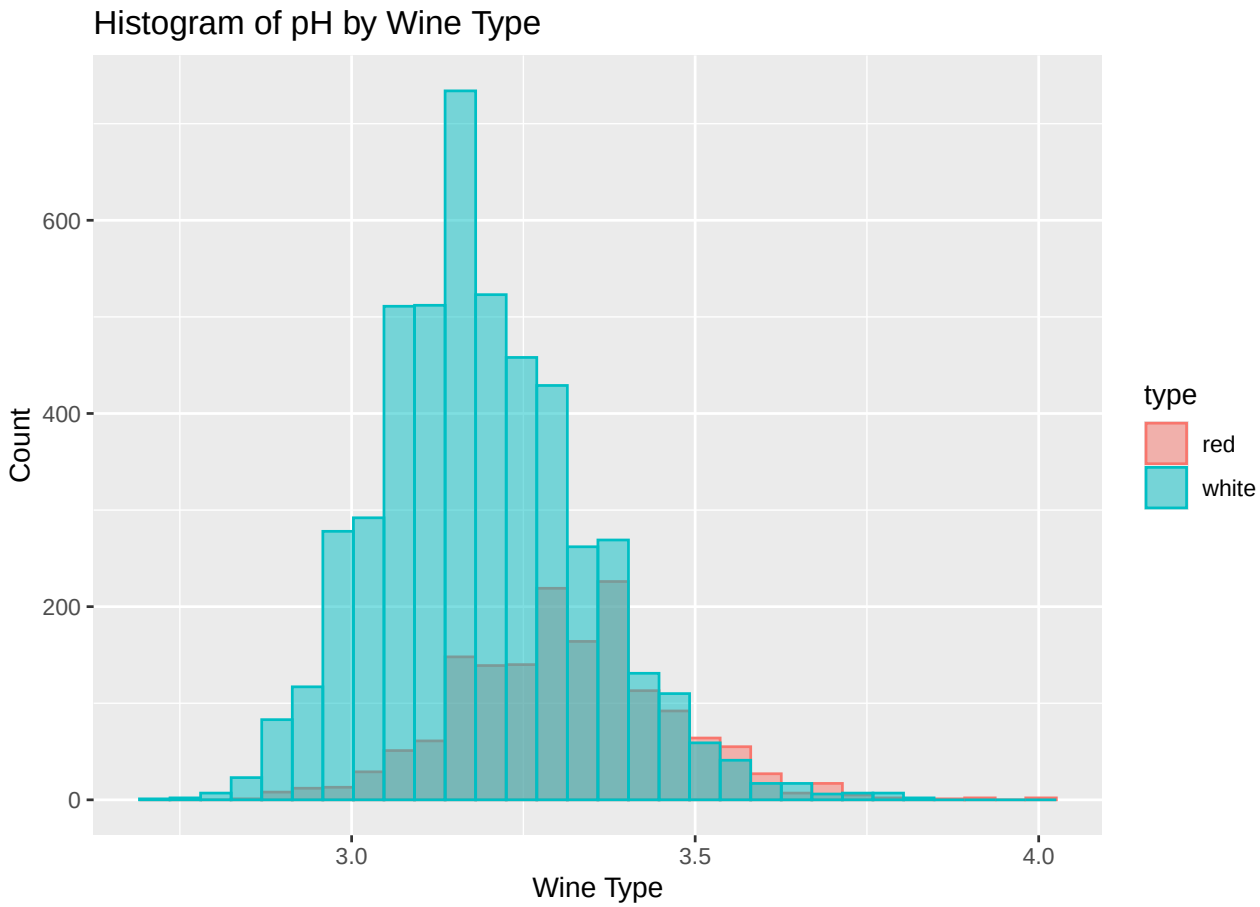

Box Plot of Total Sulfur Dioxide by Wine Type
Total Sulfur Dioxide
400
300
200
100
0
red
white
Wine Type

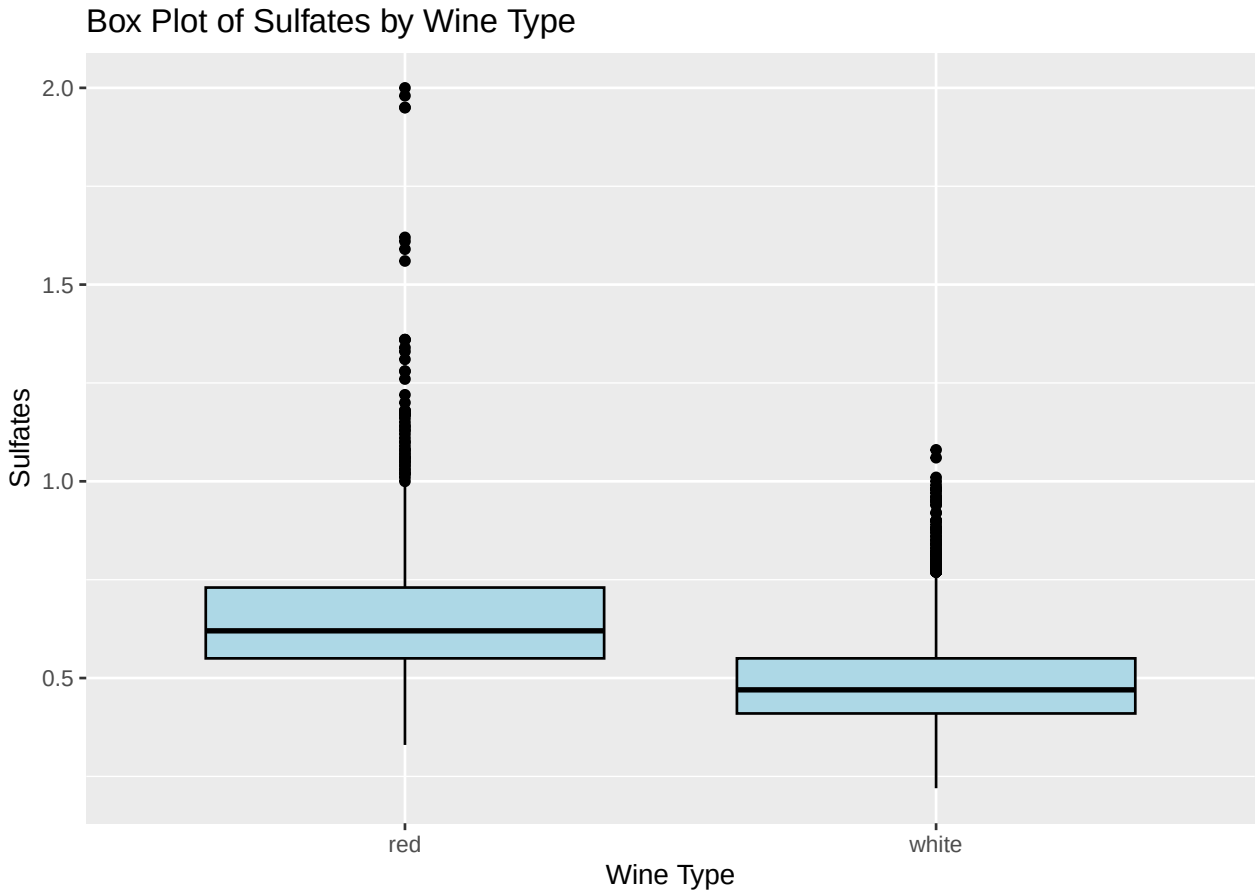

Histogram of Total Sulfur Dioxide by Wine Type
Count
600
400
200
0
0
100
200
300
400
Wine Type
type
red
white

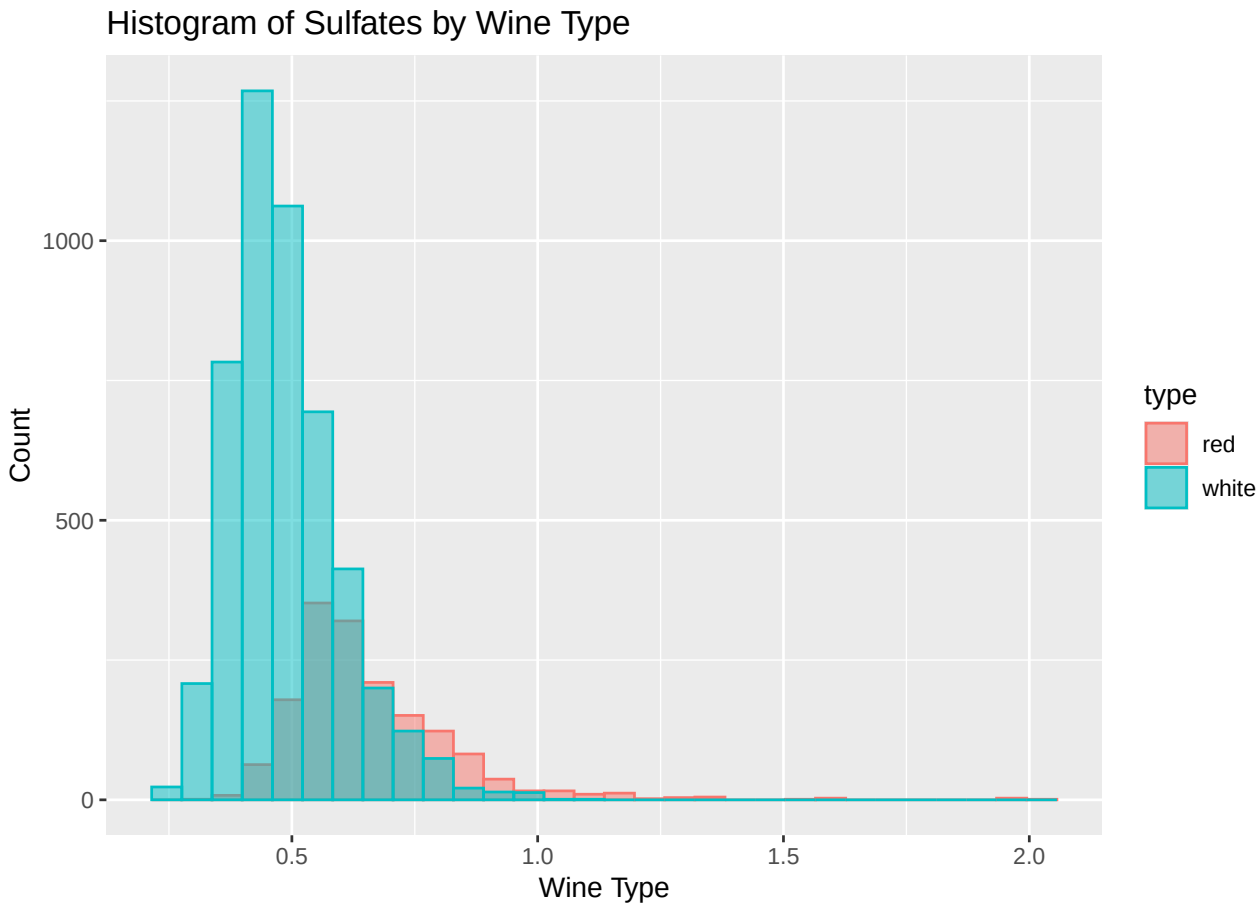

Box Plot of Density by Wine Type
1.04
1.03
1.02
1.01
1.00
0.99
Density
red
white
Wine Type

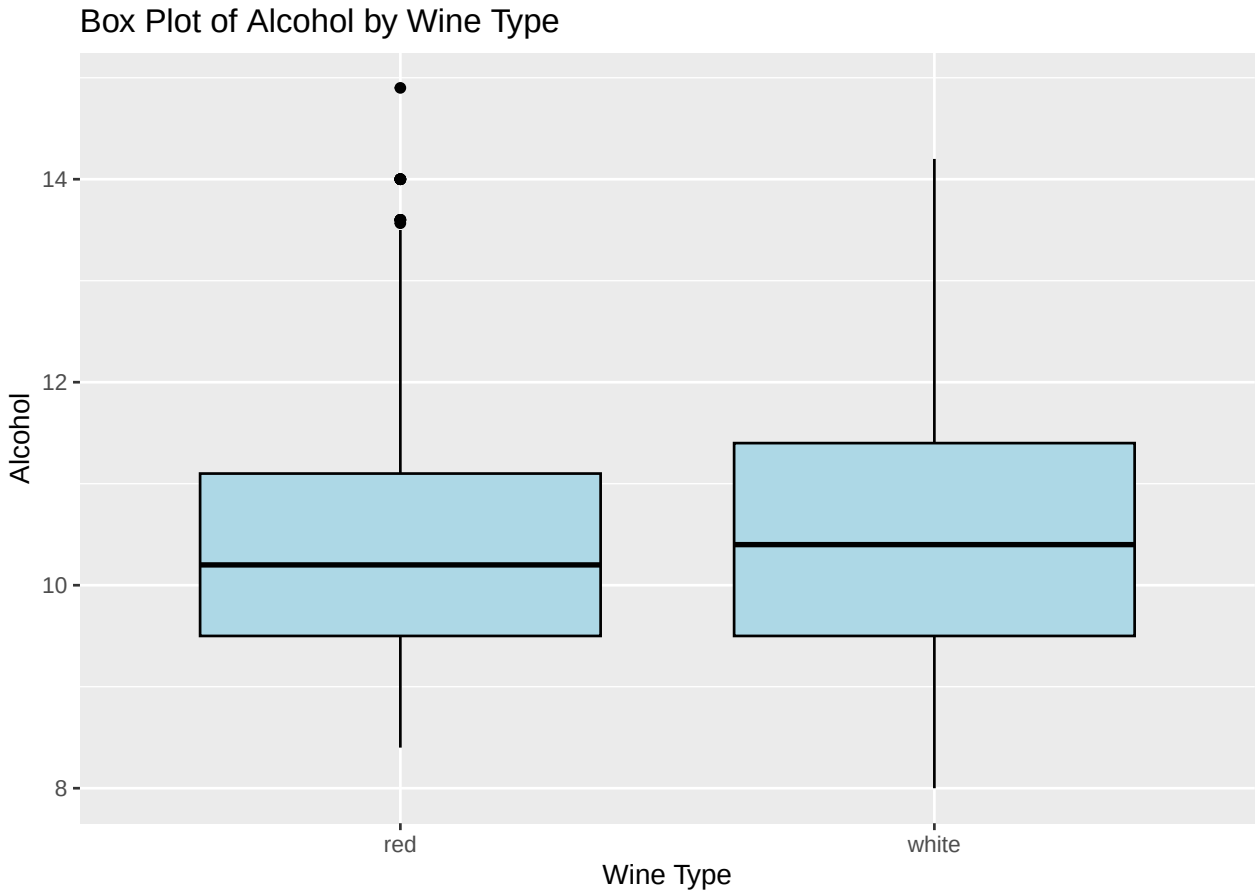

Histogram of Density by Wine Type
Count
900
600
300
0
1.00
1.02
1.04
Wine Type
type
red
white

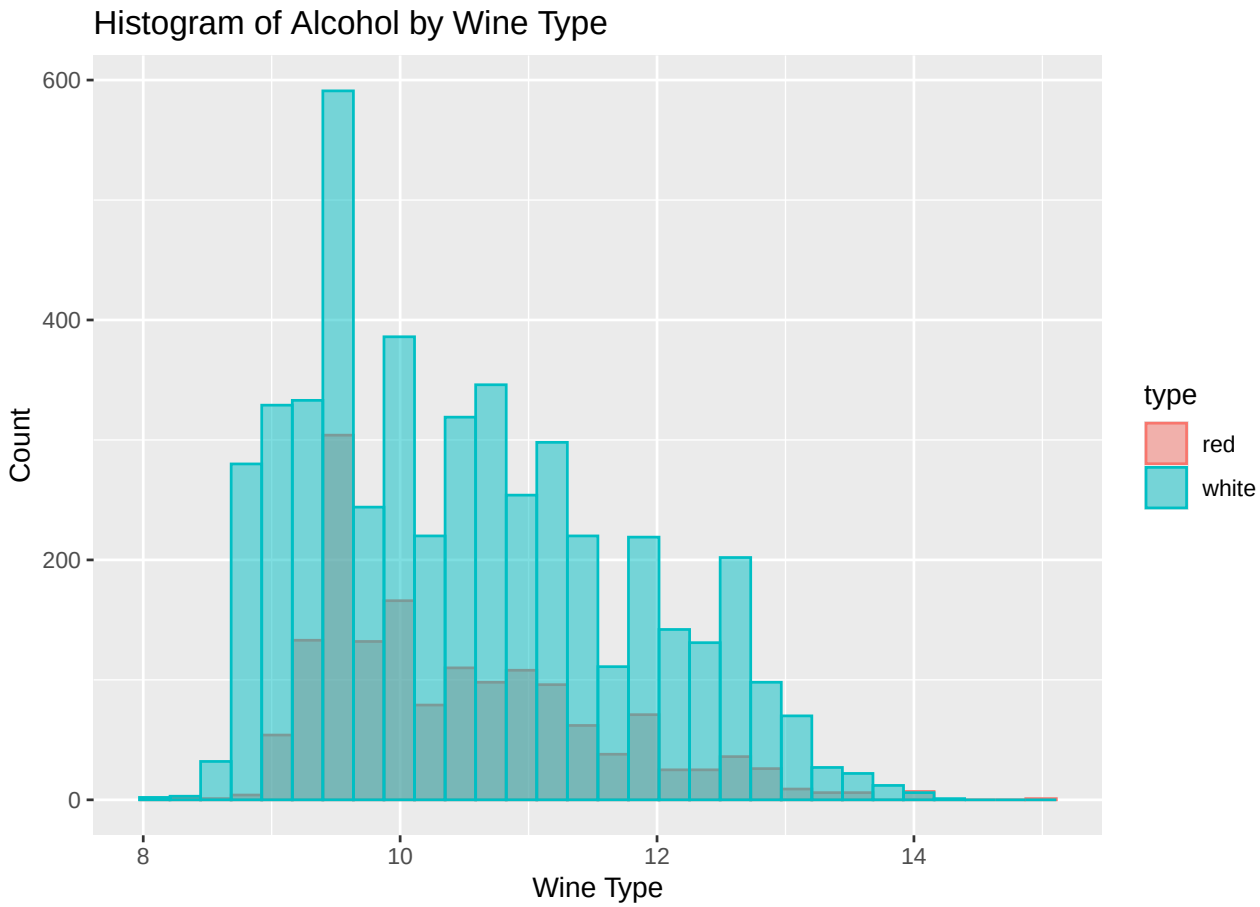

Box Plot of pH by Wine Type
4.0
3.5
3.0
pH
red
white
Wine Type

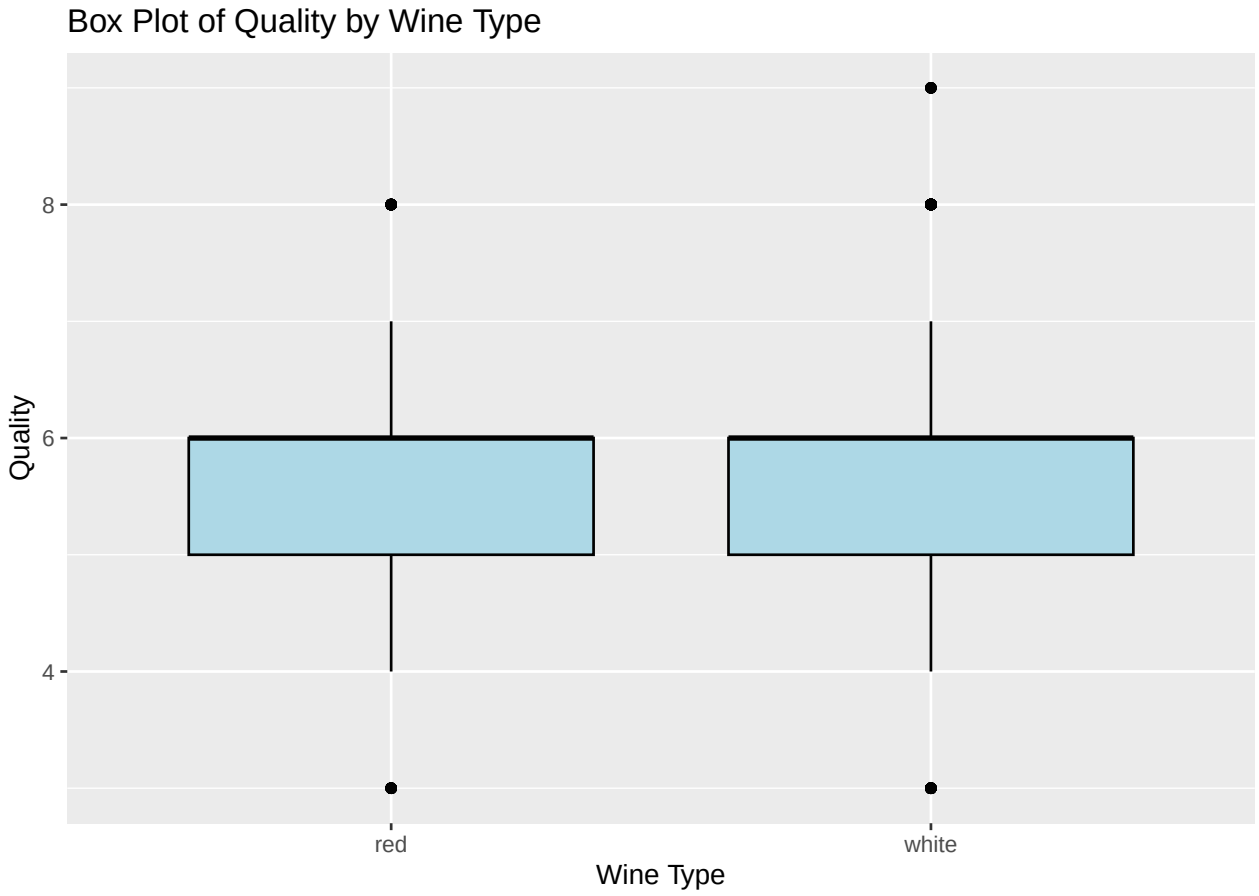

Histogram of pH by Wine Type
Count
600
400
200
0
3.0
3.5
4.0
Wine Type
type
red
white

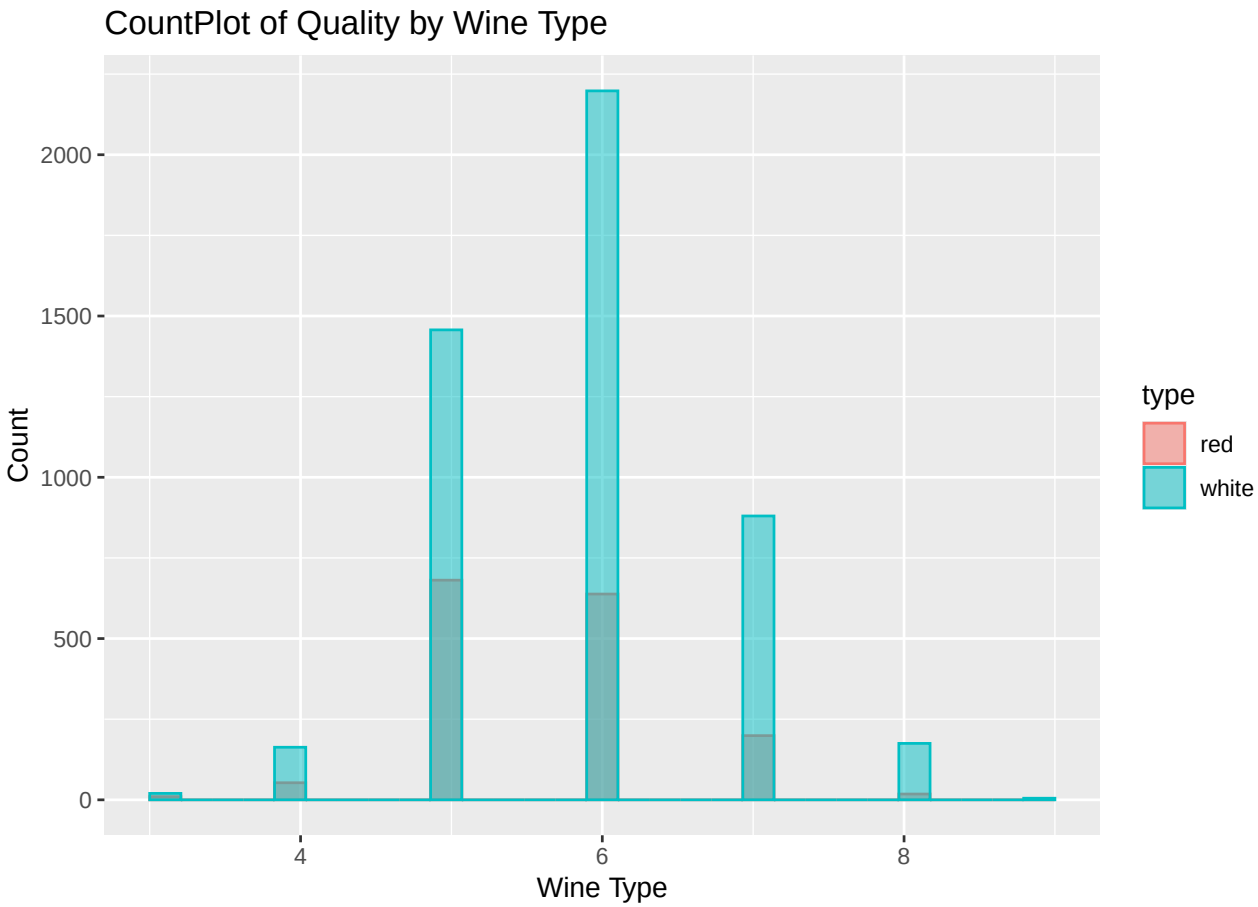

Box Plot of Sulfates by Wine Type
Sulfates
2.0
1.5
1.0
0.5
red
white
Wine Type


Histogram of Sulfates by Wine Type
Count
1000
500
0
0.5
1.0
1.5
2.0
Wine Type
type
red
white

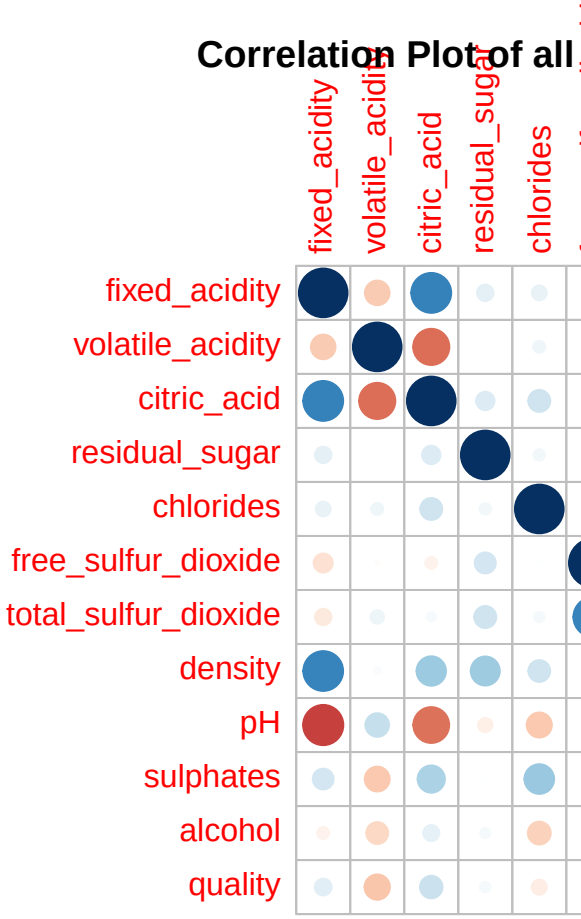

Box Plot of Alcohol by Wine Type
Alcohol
14
12
10
8
red
white
Wine Type


Histogram of Alcohol by Wine Type
600
400
200
0
Count
8
10
12
14
Wine Type
type
red
white

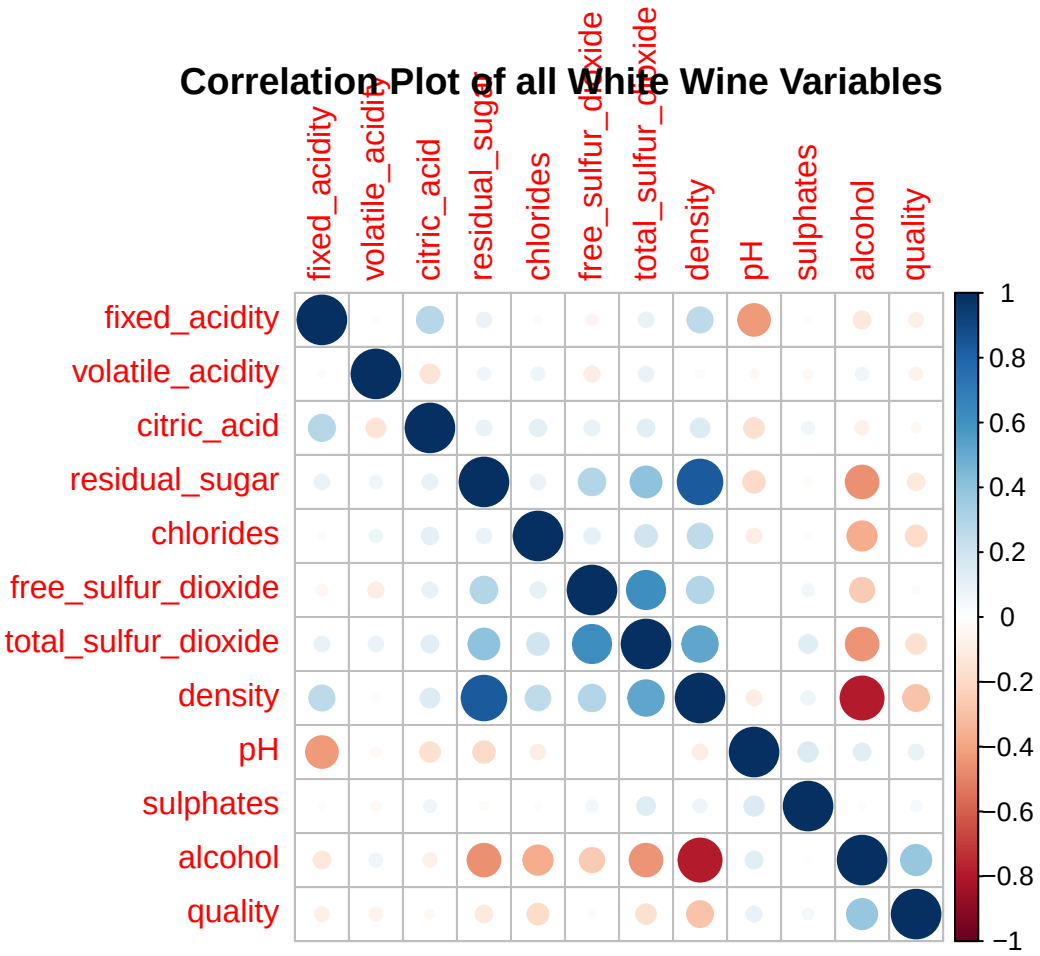

Box Plot of Quality by Wine Type
Quality
8
6
4
red
white
Wine Type




```
## Examining the CountPlot of Wine Type, the purpose of this investigation
##     since less than 25% of wines in both red and white wines have high wine
##     ratings greater than 7, Grade A wine is defined as those with an rating of 7
##     or greater. The purpose of this investigation is to predict Grade A wine
##     and be able to define characteristics of Grade A wine
```



EDA Continued: Checking Correlation of all Variables with correlation plot



Since summary statistics showed that variables were on different scales and there was a lot of difference observed between variables must apply scaling method to normalize data.

```
#Applying Min-max scaling
min_max_scaling_white <- preProcess(whitewine_seperated[1:11], method = "range")
white_wine_scaled <- predict(min_max_scaling_white,whitewine_seperated)
white_wine_scaled$quality <- as.factor(white_wine_scaled$quality)

min_max_scaling_red <- preProcess(redwine_seperated[1:11], method = "range")
red_wine_scaled <- predict(min_max_scaling_red,redwine_seperated)
red_wine_scaled$quality <- as.factor(red_wine_scaled$quality)


redwine_randomforest_columns12 <- c("fixed_acidity", "volatile_acidity",
                                    "citric_acid","residual_sugar","chlorides",
                                    "free_sulfur_dioxide","total_sulfur_dioxide",
                                "density","pH","sulphates", "alcohol", "quality")
whitewine_randomforest_columns12 <- c("fixed_acidity", "volatile_acidity",
                                      "citric_acid","residual_sugar","chlorides",
                                      "free_sulfur_dioxide","total_sulfur_dioxide",
                                "density","pH","sulphates", "alcohol", "quality")

red_wine_scaled <- red_wine_scaled[, redwine_randomforest_columns12,
                                   drop = FALSE ]
```

```
white_wine_scaled <- white_wine_scaled[, whitewine_randomforest_columns12,
                                       drop = FALSE ]
```

Since summary statistics showed that variables were on different scales and there was a lot of difference observed between variables must apply scaling method to normalize data.

Splitting Data into 31 random train/test to effectively evaluate model performance

```
# Randomly shuffling the data and dividing into train/test
white_wine_indexes <- sample(2, nrow(white_wine_scaled),
                             replace = TRUE, prob = c(0.8,0.2))
white_wine_train <- white_wine_scaled[white_wine_indexes==1,]
white_wine_test <- white_wine_scaled[white_wine_indexes==2,]


red_wine_indexes <- sample(2, nrow(red_wine_scaled),
                           replace = TRUE, prob = c(0.8,0.2))
red_wine_train <- red_wine_scaled[red_wine_indexes==1,]
red_wine_test <- red_wine_scaled[red_wine_indexes==2,]



# Set up 30 random train/test splits for white and red wine data
set.seed(123) # for reproducibility

# Generate indexes for 30 iterations
white_wine_indexes_list <- replicate(31, sample(2,
                                                nrow(white_wine_scaled),
                                                replace = TRUE,
                                                prob = c(0.8, 0.2)),
                                     simplify = FALSE)

red_wine_indexes_list <- replicate(31, sample(2, nrow(red_wine_scaled),
                                              replace = TRUE,
                                              prob = c(0.8, 0.2)),
                                   simplify = FALSE)

# Vectorized approach with lapply
white_wine_train_list <- lapply(white_wine_indexes_list, function(index) white_wine_scaled[index == 1, ]
white_wine_test_list <- lapply(white_wine_indexes_list, function(index) white_wine_scaled[index == 2, ]]

red_wine_train_list <- lapply(red_wine_indexes_list, function(index) red_wine_scaled[index == 1, ])
red_wine_test_list <- lapply(red_wine_indexes_list, function(index) red_wine_scaled[index == 2, ])
```

Since data is very unbalanced with Grade A Wine representing less than 25% of respective wine types randomly sampling with replacement fom original data to synthetically replicate minority class of Grade A Wine in both red wine and white wine data so that model can pick up complex relationships

```
# Defining oversampling functions
oversample_data_red <- function(my_data) {
  data <- my_data
  return(ovun.sample(quality ~ ., data = data, method = "over", N = 2150)$data)
}

oversample_data_white <- function(my_data) {
```

```r
  data <- my_data
  return(ovun.sample(quality ~ ., data = data, method = "over", N = 6150)$data)
}


# Applying oversampling to all training sets
oversampled_red_wine_train_list <- lapply(red_wine_train_list,
                                          oversample_data_red)

oversampled_white_wine_train_list <- lapply(white_wine_train_list,
                                            oversample_data_white)
```