Short Report 2 Code

Anna Ursin and Kunwu Lyu

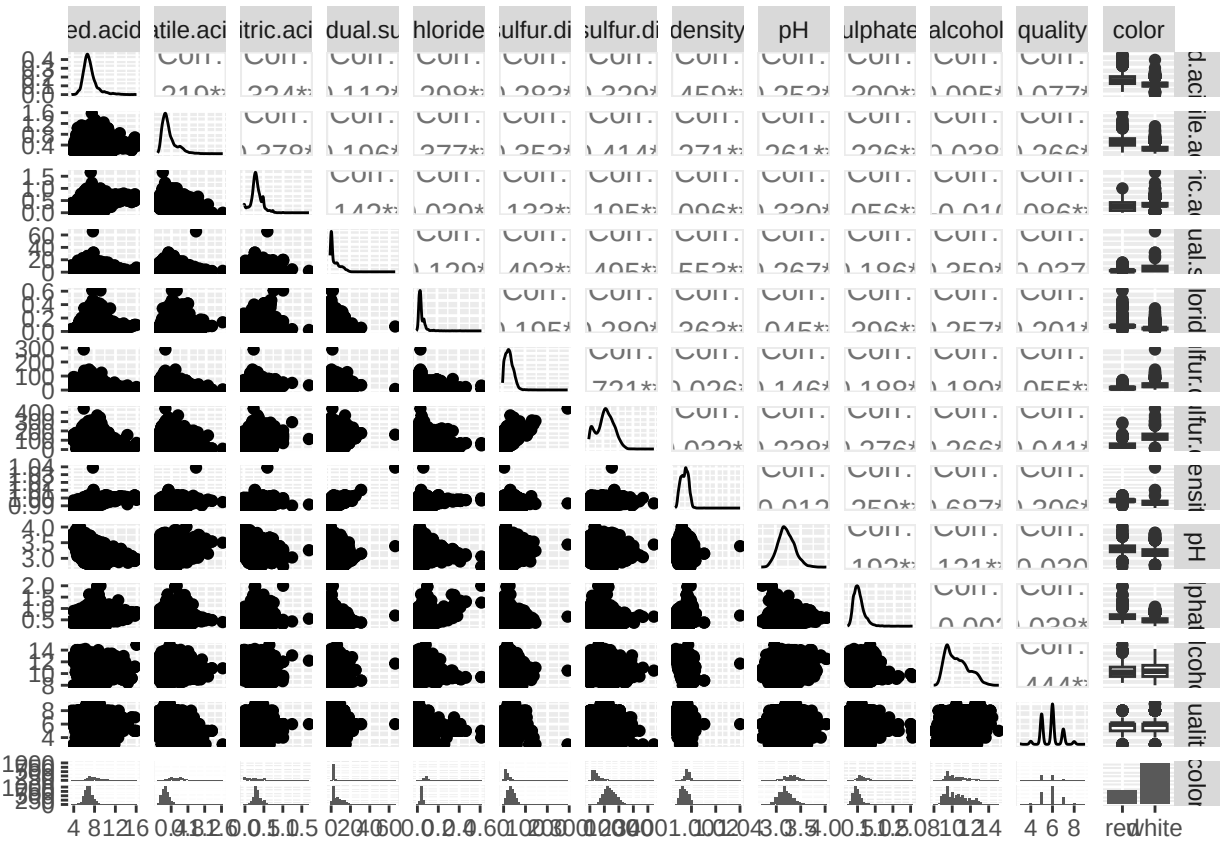
2024-11-07

Data Wrangling

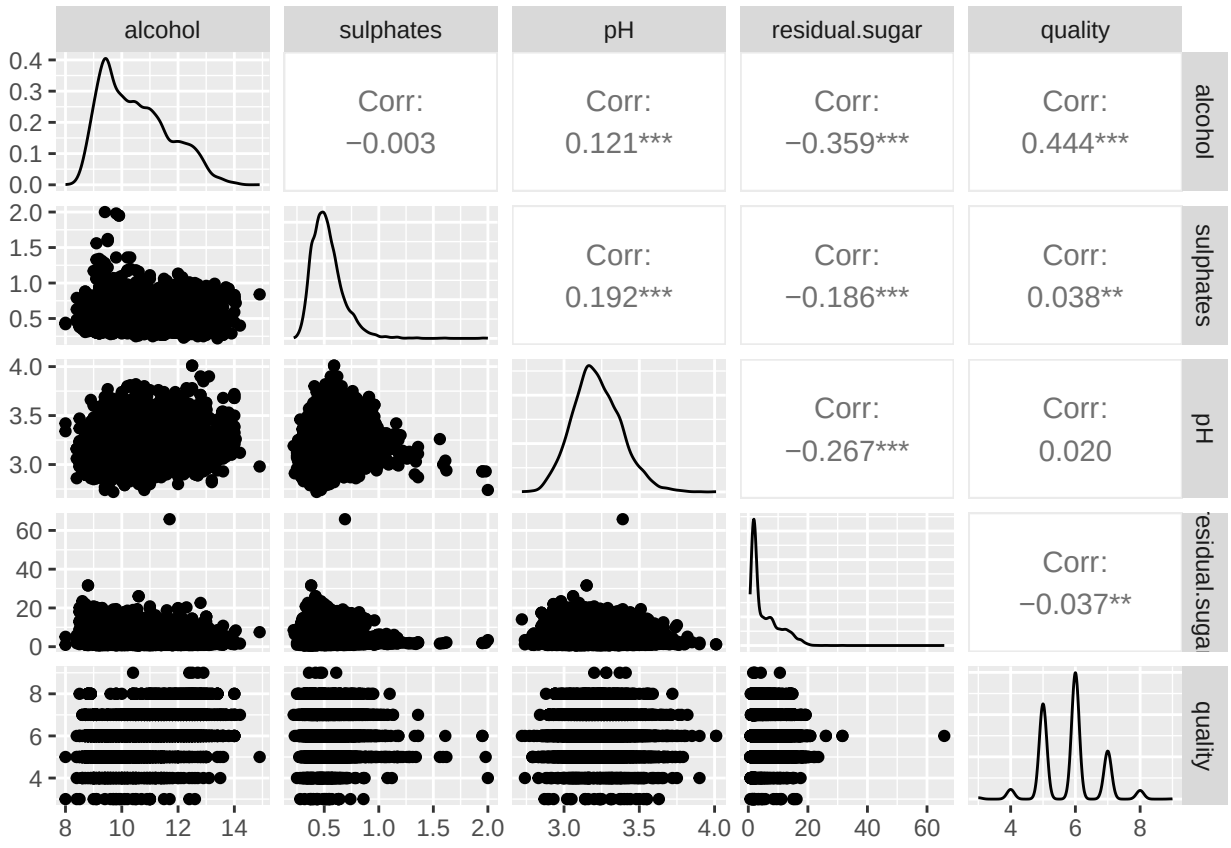
```
## Data Wrangling & Scatterplot Matrix
wine_dat <- read.csv("https://www.math.carleton.edu/ckelling/data/wine_project.csv")
glimpse(wine_dat)
```

```
## Rows: 6,497
## Columns: 13
## $ fixed.acidity        <dbl> 7.4, 7.8, 7.8, 11.2, 7.4, 7.4, 7.9, 7.3, 7.8, 7.5~
## $ volatile.acidity     <dbl> 0.700, 0.880, 0.760, 0.280, 0.700, 0.660, 0.600, ~
## $ citric.acid          <dbl> 0.00, 0.00, 0.04, 0.56, 0.00, 0.00, 0.06, 0.00, 0~
## $ residual.sugar       <dbl> 1.9, 2.6, 2.3, 1.9, 1.9, 1.8, 1.6, 1.2, 2.0, 6.1,~
## $ chlorides            <dbl> 0.076, 0.098, 0.092, 0.075, 0.076, 0.075, 0.069, ~
## $ free.sulfur.dioxide  <dbl> 11, 25, 15, 17, 11, 13, 15, 15, 9, 17, 15, 17, 16~
## $ total.sulfur.dioxide <dbl> 34, 67, 54, 60, 34, 40, 59, 21, 18, 102, 65, 102,~
## $ density              <dbl> 0.9978, 0.9968, 0.9970, 0.9980, 0.9978, 0.9978, 0~
## $ pH                   <dbl> 3.51, 3.20, 3.26, 3.16, 3.51, 3.51, 3.30, 3.39, 3~
## $ sulphates            <dbl> 0.56, 0.68, 0.65, 0.58, 0.56, 0.56, 0.46, 0.47, 0~
## $ alcohol              <dbl> 9.4, 9.8, 9.8, 9.8, 9.4, 9.4, 9.4, 10.0, 9.5, 10.~
## $ quality              <int> 5, 5, 5, 6, 5, 5, 5, 7, 7, 5, 5, 5, 5, 5, 5, 5, 7~
## $ color                <chr> "red", "red", "red", "red", "red", "red", "red", ~
```

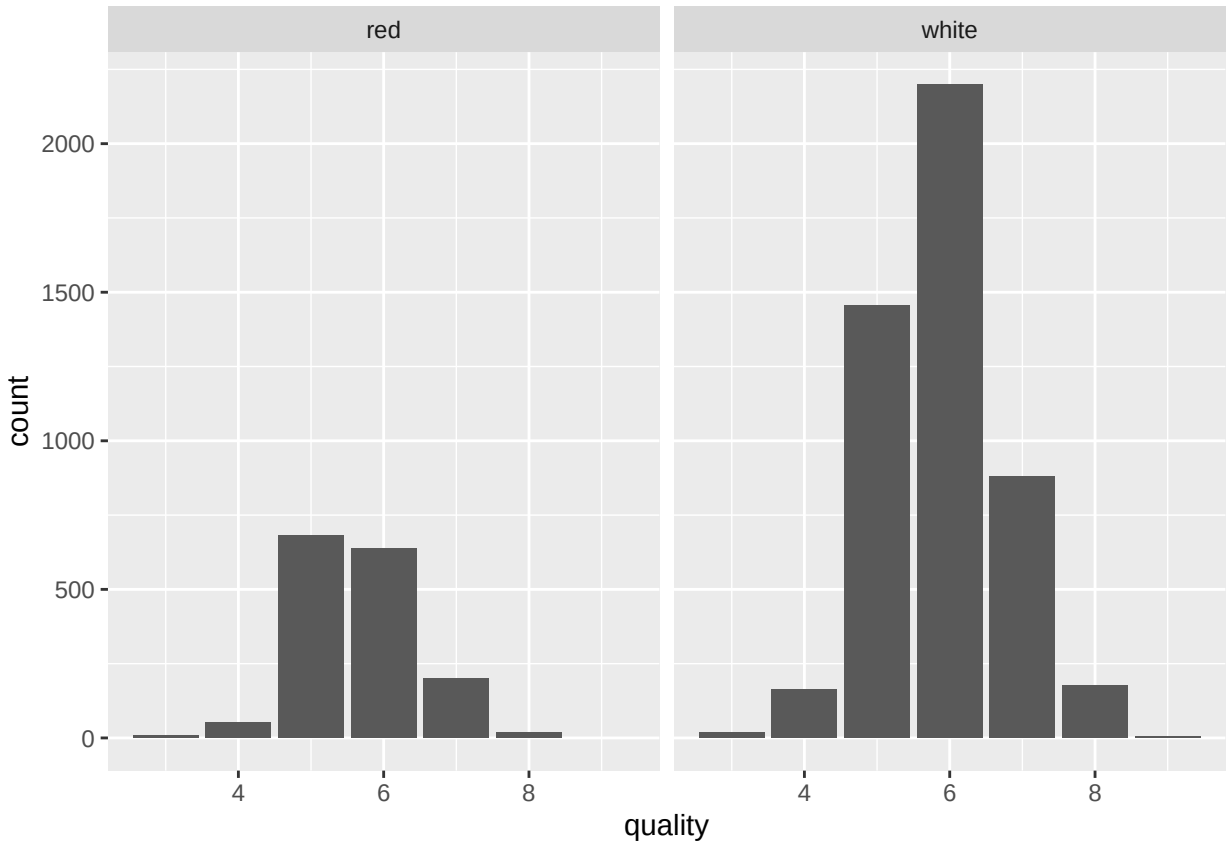
```
ggpairs(wine_dat) # for all
```




```
ggpairs(data = wine_dat, columns =
          c("alcohol", "sulphates", "pH",
            "residual.sugar", "quality")) # for interested vars
```




```
## dist of quality scores
ggplot(wine_dat, aes(x = quality)) +
  geom_bar() +
  facet_grid(cols = vars(color))
```

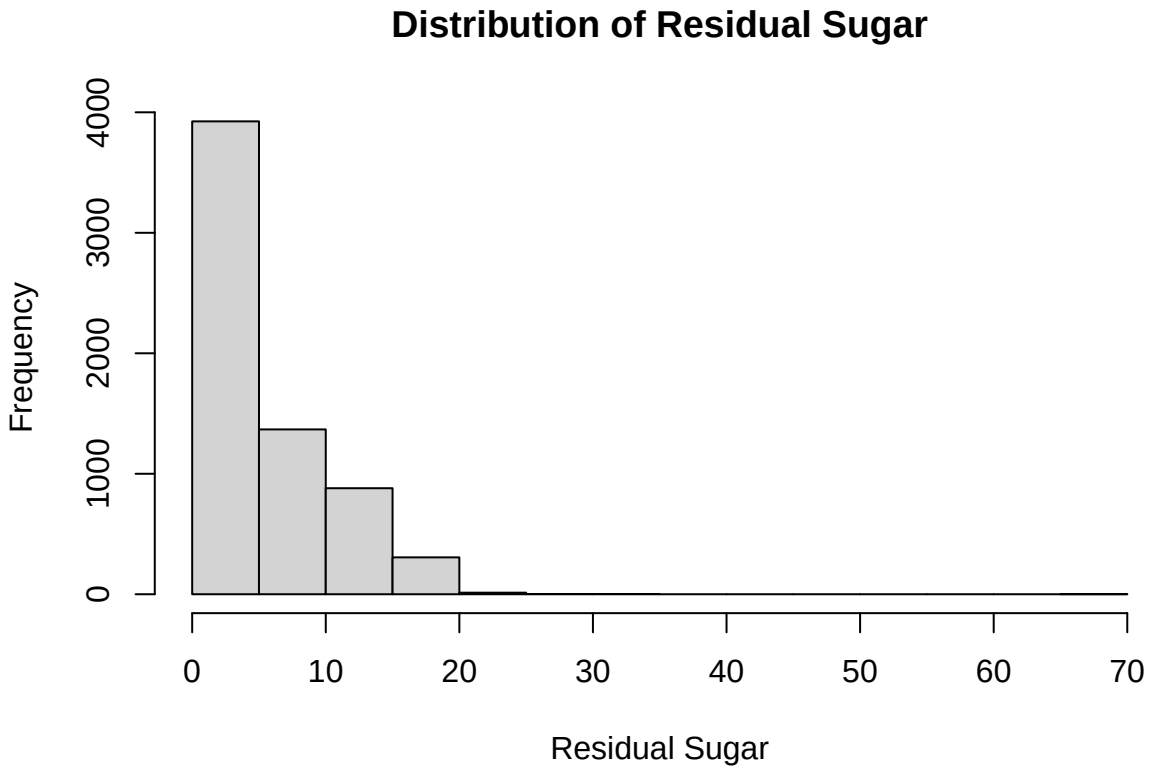



```
  labs(title = "Distribution of Wine Quality Scores")
```

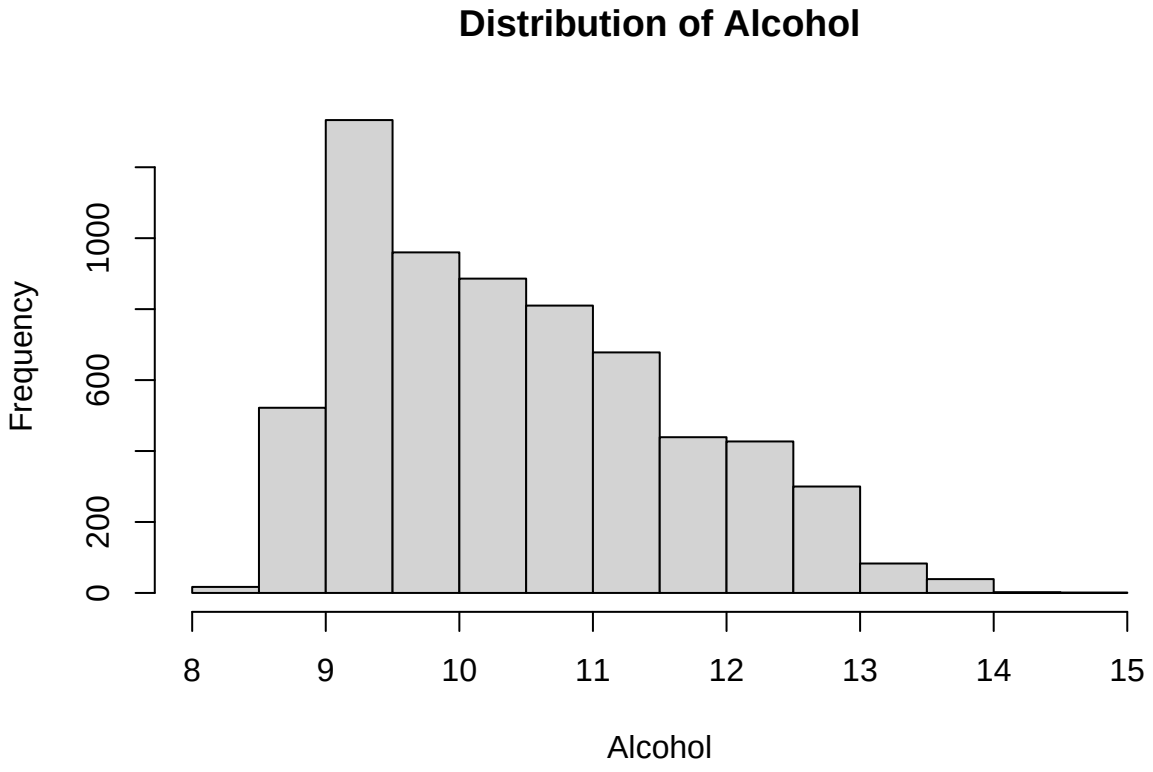
```
## $title
## [1] "Distribution of Wine Quality Scores"
##
## attr(,"class")
## [1] "labels"
```

EDA Plot

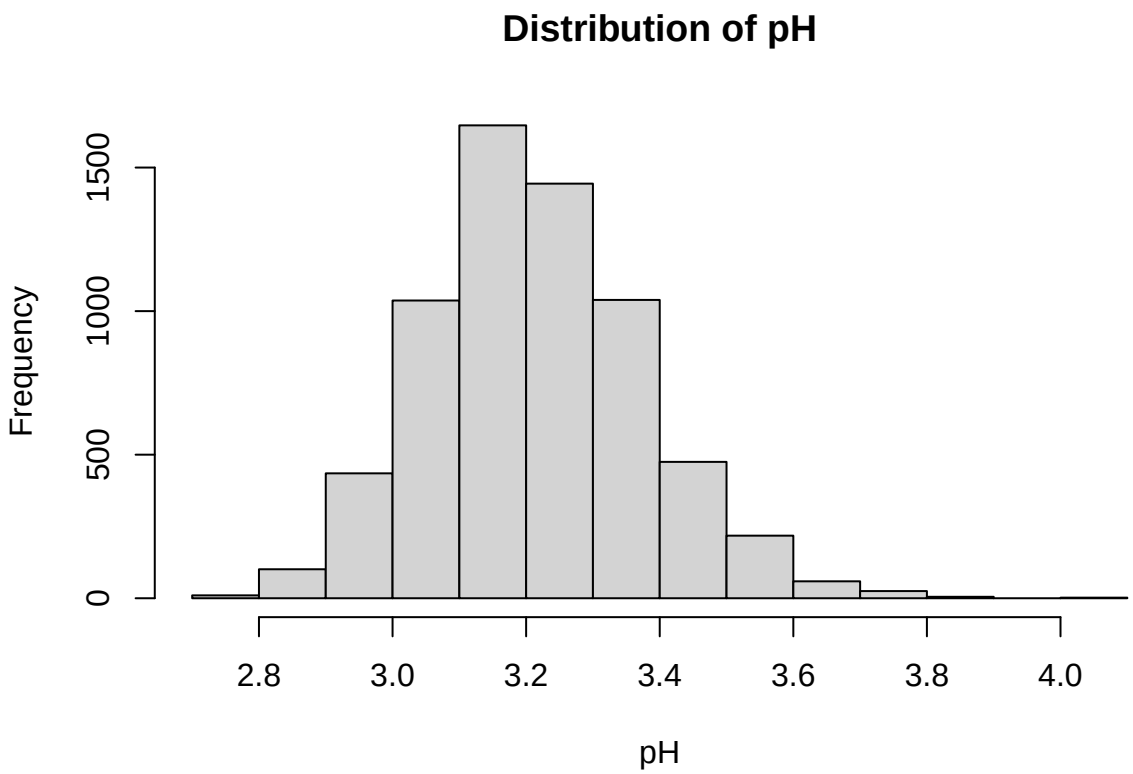
```
#Variable Distributions through Histograms
hist(wine_dat$residual.sugar,
     main = "Distribution of Residual Sugar", xlab = "Residual Sugar")
```



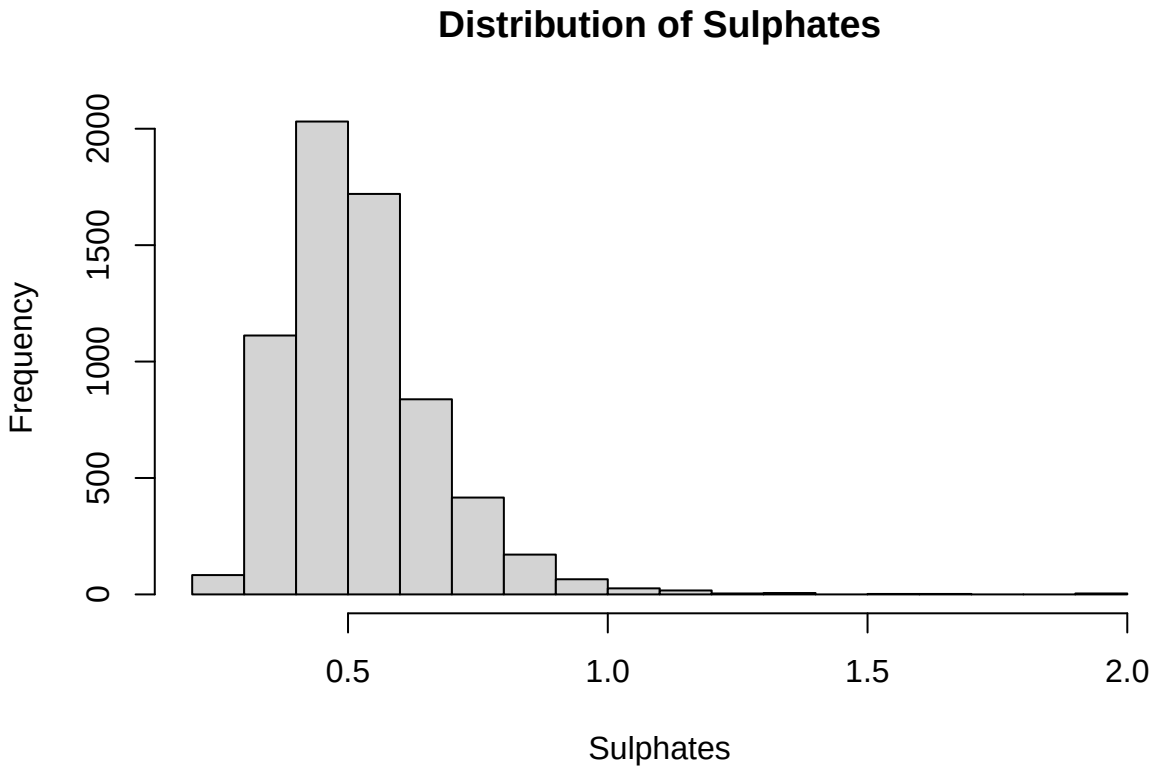

```r
hist(wine_dat$alcohol, main = "Distribution of Alcohol", xlab = "Alcohol")
```



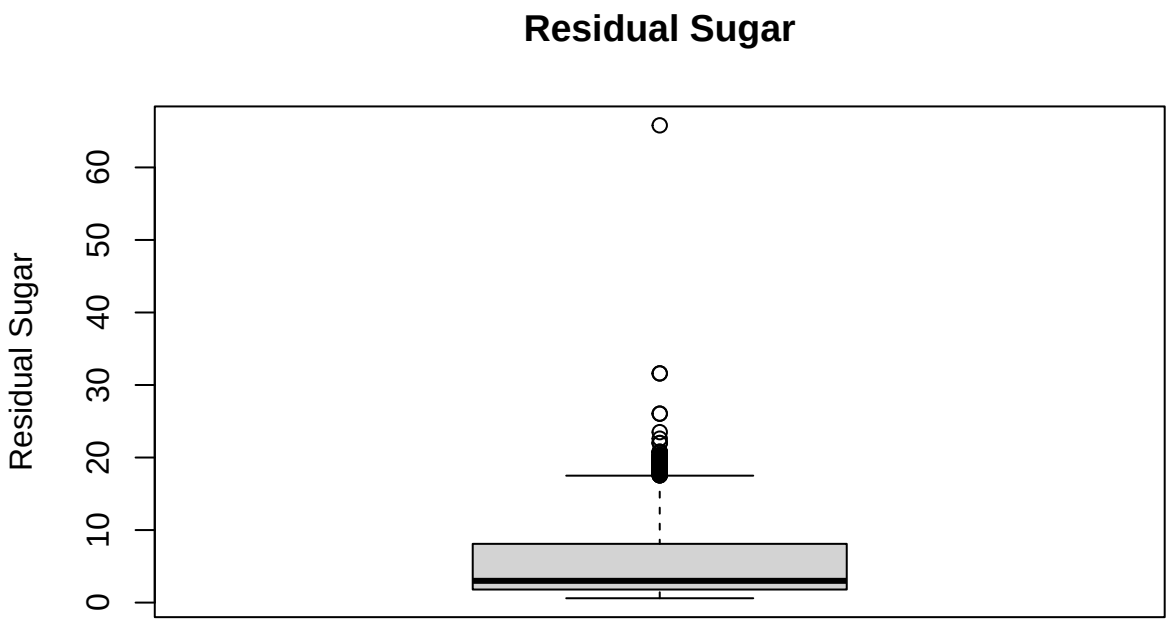

```r
hist(wine_dat$pH, main = "Distribution of pH", xlab = "pH")
```



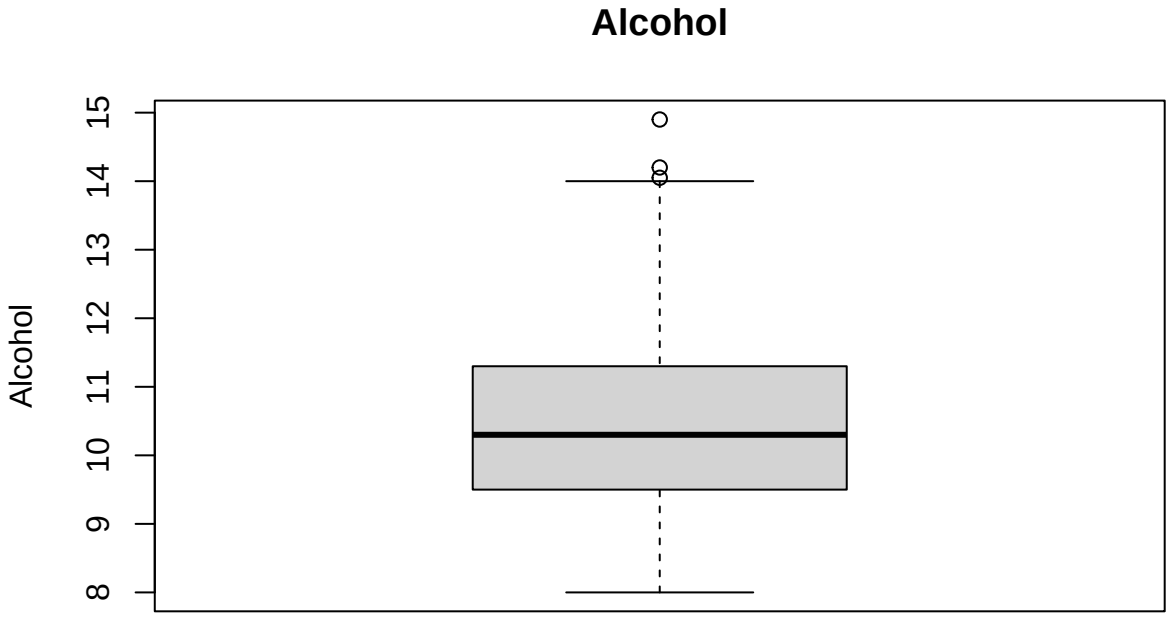

```r
hist(wine_dat$sulphates, main = "Distribution of Sulphates", xlab = "Sulphates")
```



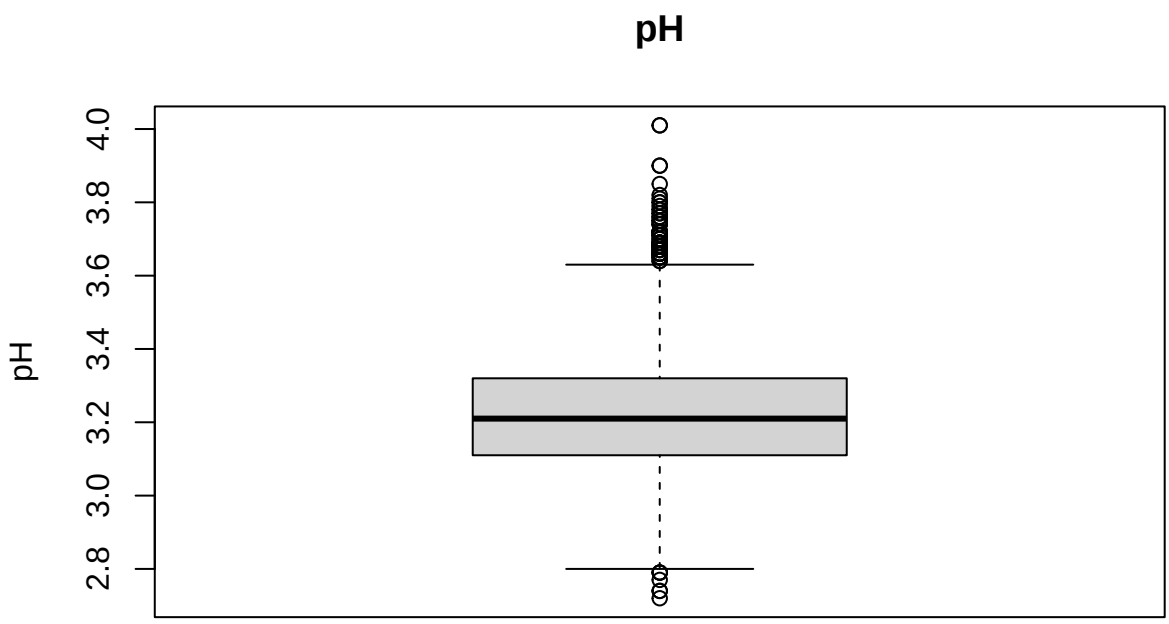

```r
#Variable Distributions through Boxplots
boxplot(wine_dat$residual.sugar, main = "Residual Sugar", ylab = "Residual Sugar")
```



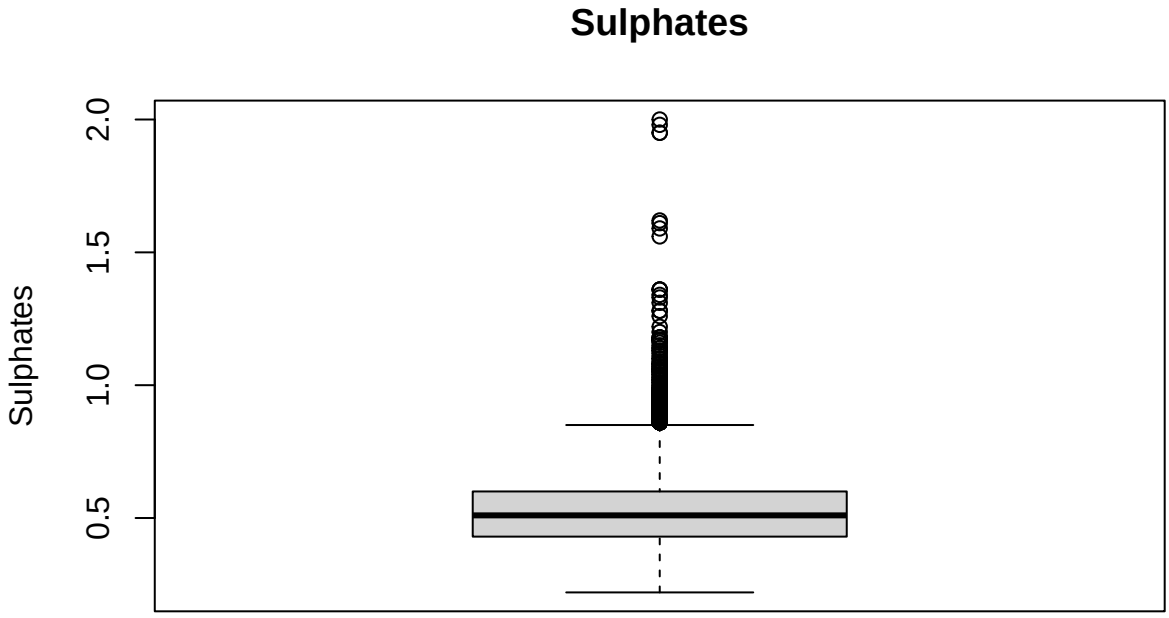

```
boxplot(wine_dat$alcohol, main = "Alcohol", ylab = "Alcohol")
```




```
boxplot(wine_dat$pH, main = "pH", ylab = "pH")
```

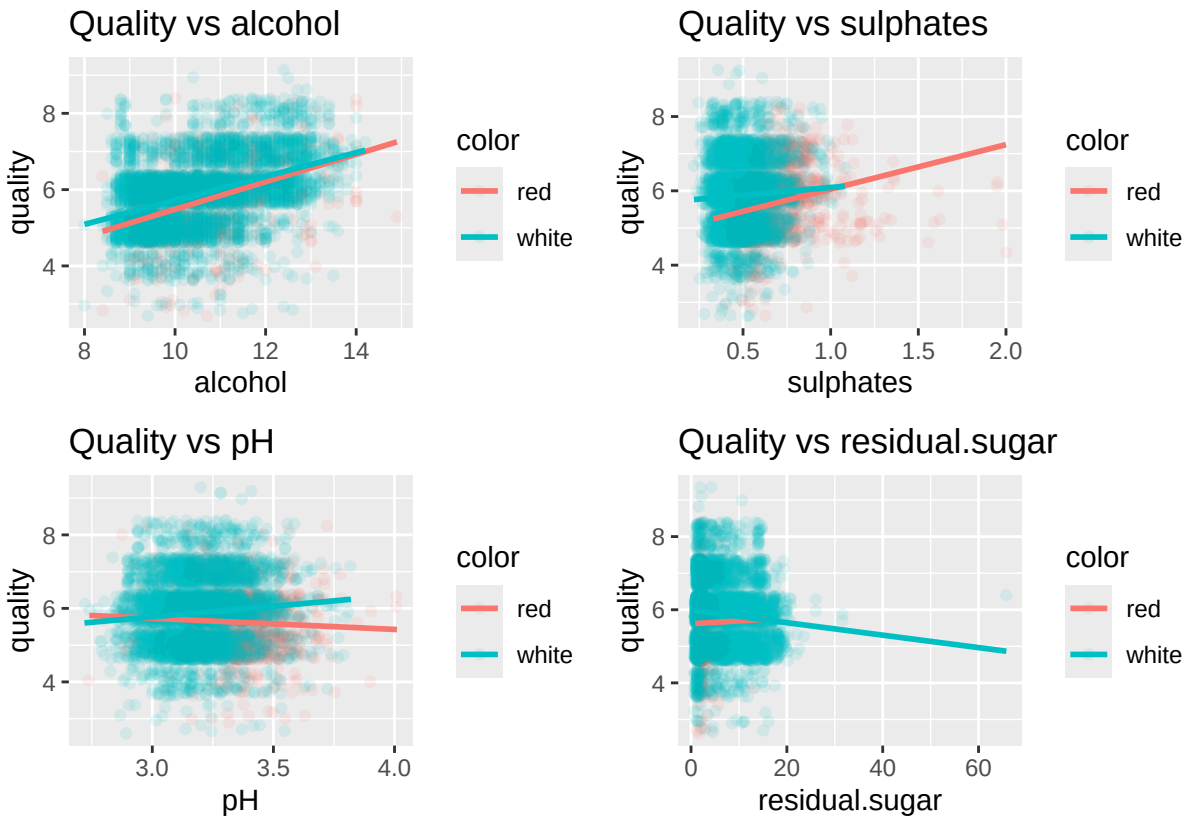



```
boxplot(wine_dat$sulphates, main = "Sulphates", ylab = "Sulphates")
```

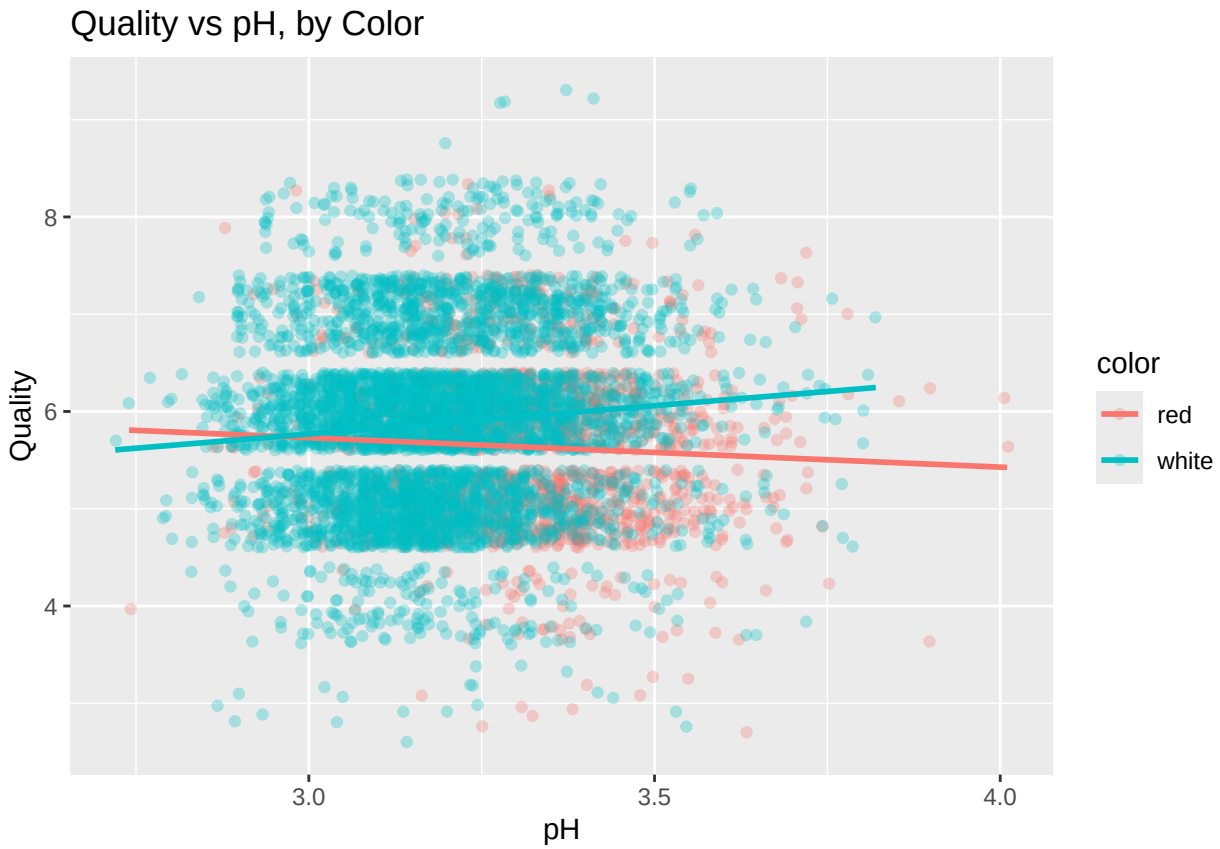



```
## scatterplots for predictors vs wine quality
create_plot <- function(var_name) {
  ggplot(data = wine_dat,
         aes_string(x = var_name, y = "quality", color = "color")) +
    geom_point(alpha = 0.1, position = position_jitter()) +
    geom_smooth(method = "lm", se = FALSE) +
    labs(title = paste("Quality vs", var_name))
}

variable_names <- c("alcohol", "sulphates", "pH", "residual.sugar")
plot_list <- map(variable_names, create_plot)

plot_grid <- wrap_plots(plot_list, ncol = 2)
```

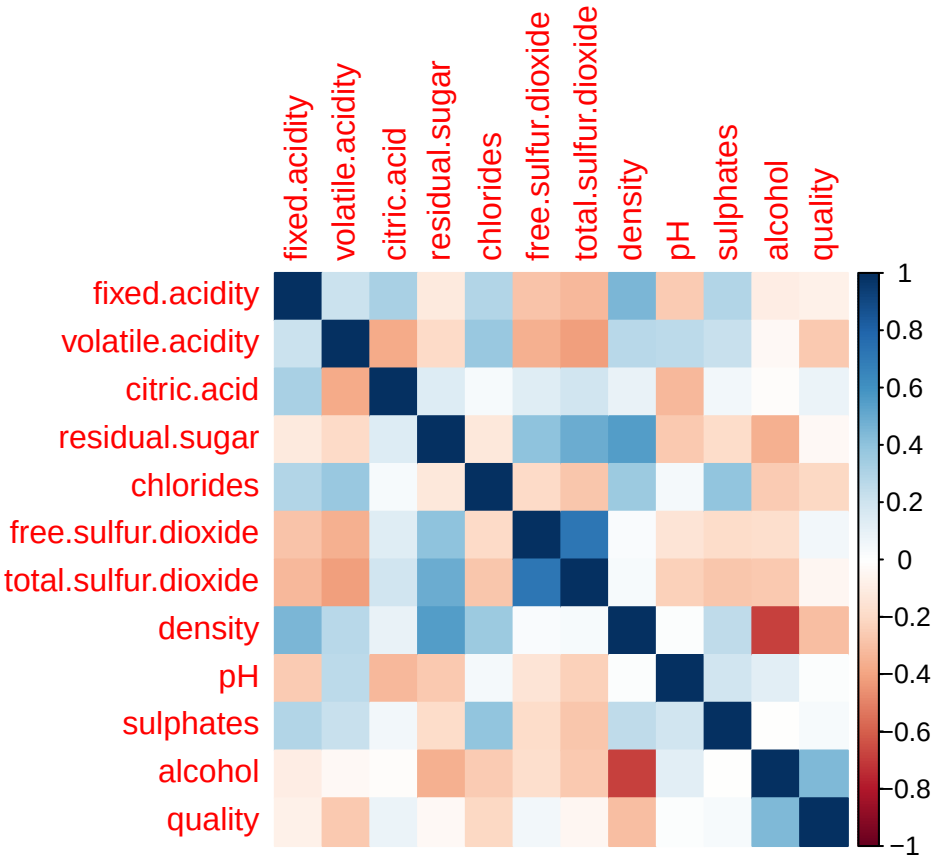
```
plot_grid
```




```
## scat plot for report
ggplot(data = wine_dat, aes(x = pH, y = quality, color = color)) +
  geom_point(alpha = 0.3, position = position_jitter()) +
  geom_smooth(method = "lm", se = F) +
  labs(title = "Quality vs pH, by Color", x = "pH", y = "Quality")
```




```
## correlation matrix
wine_numeric <- wine_dat %>% dplyr::select(-color)

corr_matrix <- cor(wine_numeric, use = "complete.obs")
corrplot(corr_matrix, method = "color")
```



Regression Model

```
## Main Effects Model
wine_lm_base <- lm(quality ~ alcohol + sulphates + residual.sugar + pH, data = wine_dat)
summary(wine_lm_base)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + sulphates + residual.sugar +
##     pH, data = wine_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4939 -0.5010 -0.0346  0.4919  3.0992
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.842152   0.221545   8.315  < 2e-16 ***
## alcohol         0.366868   0.008635  42.486  < 2e-16 ***
## sulphates       0.417133   0.066463   6.276 3.69e-10 ***
## residual.sugar  0.028000   0.002252  12.434  < 2e-16 ***
## pH             -0.076713   0.062563  -1.226     0.22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7717 on 6492 degrees of freedom
```

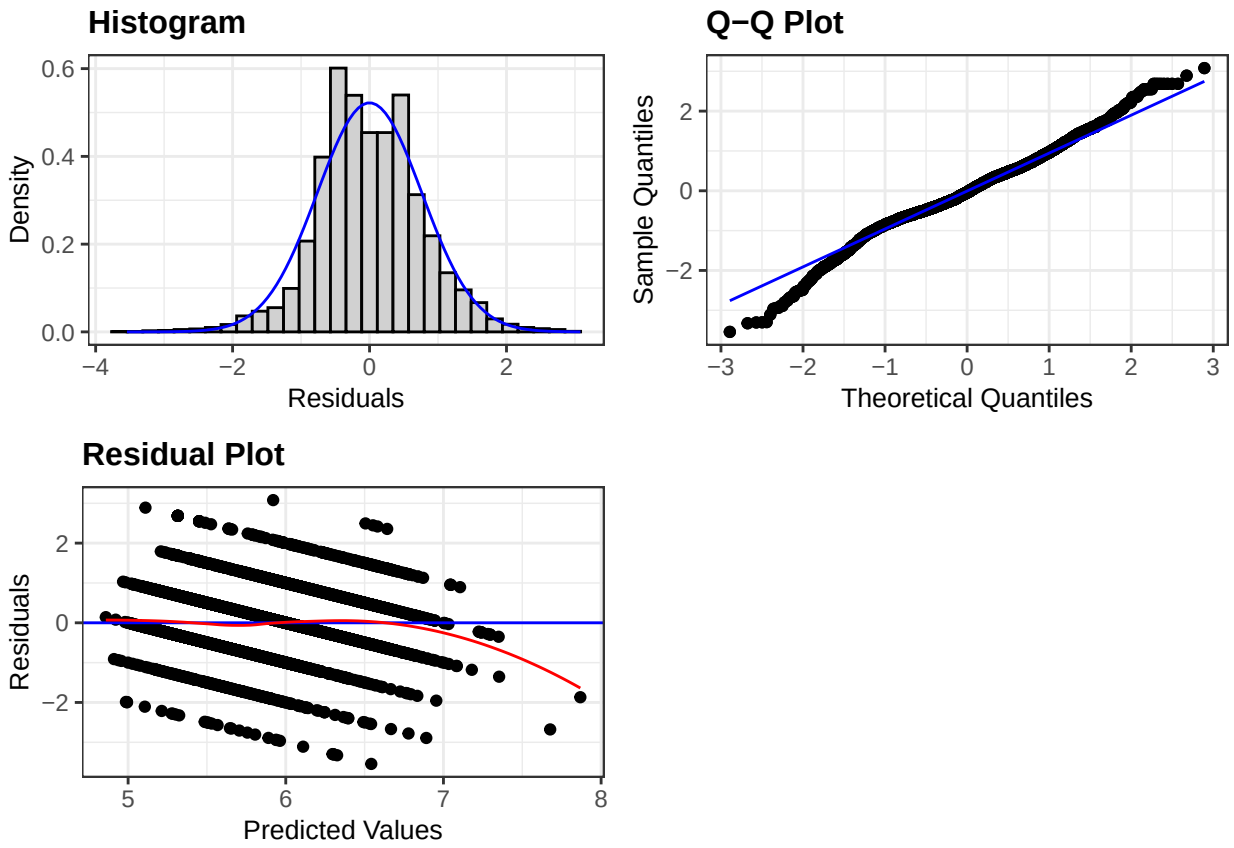
```
## Multiple R-squared:  0.2195, Adjusted R-squared:  0.219
## F-statistic: 456.3 on 4 and 6492 DF,  p-value: < 2.2e-16
```

```
## Interaction Model
wine_lm_interaction <- lm(quality ~ alcohol + sulphates + residual.sugar + pH*color,
                          data = wine_dat)
summary(wine_lm_interaction)
```

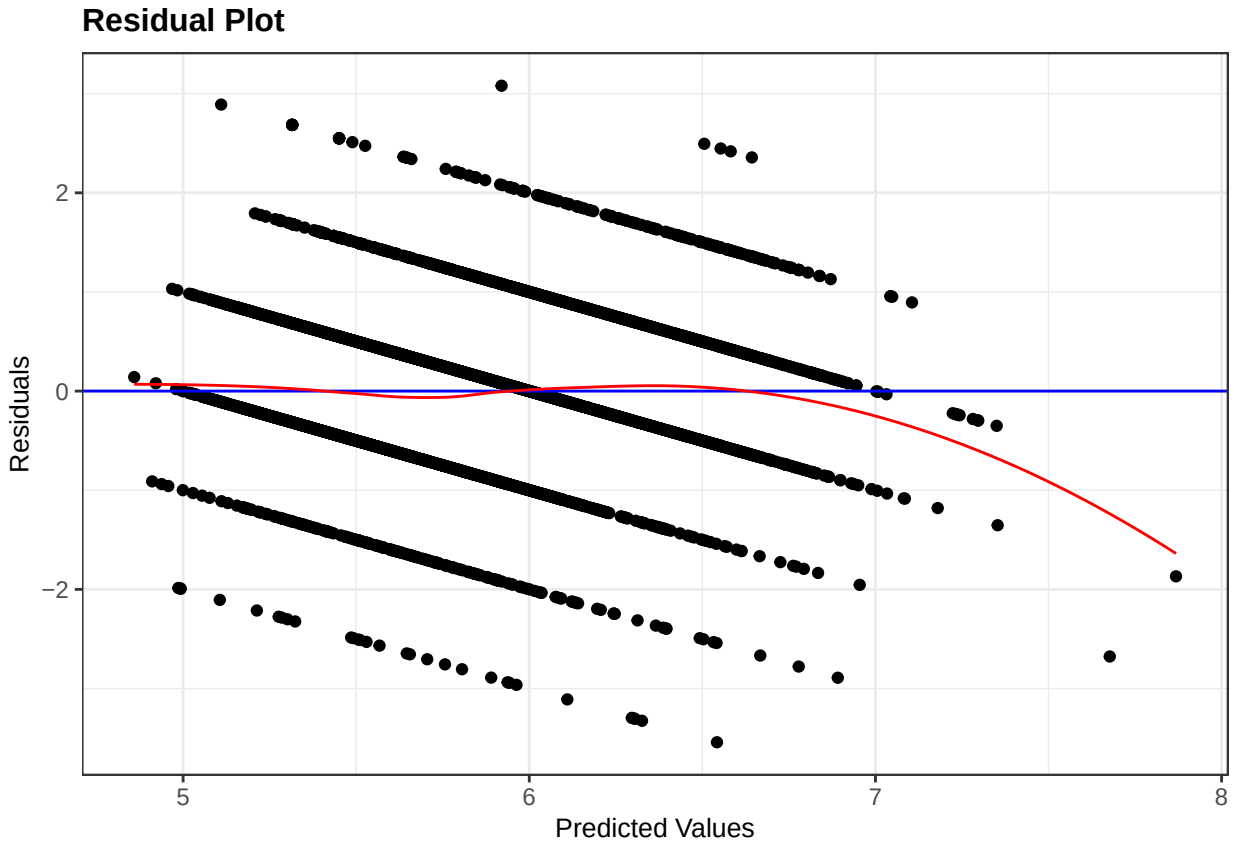
```
##
## Call:
## lm(formula = quality ~ alcohol + sulphates + residual.sugar +
##     pH * color, data = wine_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5422 -0.4971 -0.0404  0.4843  3.0800
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.620360   0.426412   8.490  < 2e-16 ***
## alcohol         0.356333   0.008710  40.913  < 2e-16 ***
## sulphates       0.624165   0.074126   8.420  < 2e-16 ***
## residual.sugar  0.023240   0.002348   9.896  < 2e-16 ***
## pH             -0.654821   0.125466  -5.219 1.85e-07 ***
## colorwhite     -2.926864   0.483888  -6.049 1.54e-09 ***
## pH:colorwhite   0.963311   0.146385   6.581 5.05e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7644 on 6490 degrees of freedom
## Multiple R-squared:  0.2345, Adjusted R-squared:  0.2338
## F-statistic: 331.3 on 6 and 6490 DF,  p-value: < 2.2e-16
```

Residual Plots

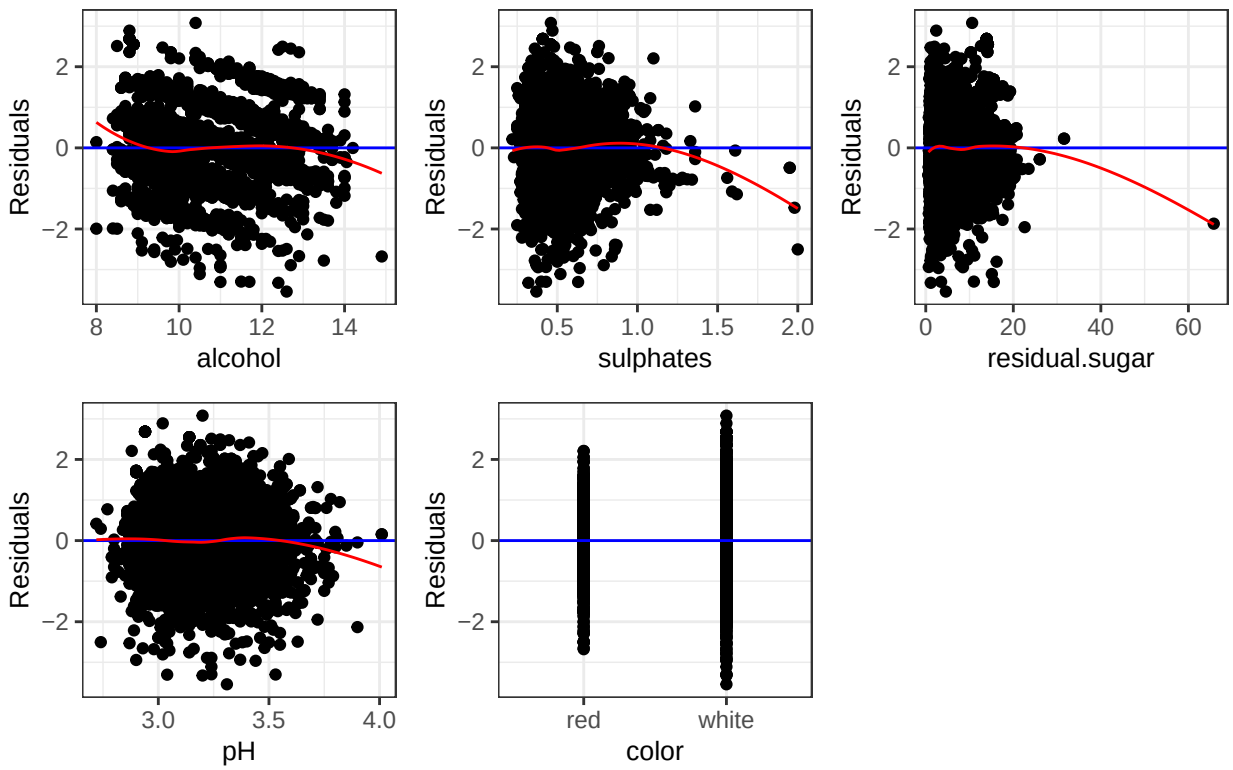
```
## Residual Plots with ggResidPanel
resid_panel(wine_lm_interaction, plots = c("hist", "qq", "resid"), smoother = T)
```




```
resid_panel(wine_lm_interaction, plots = "resid", smoother = TRUE)
```

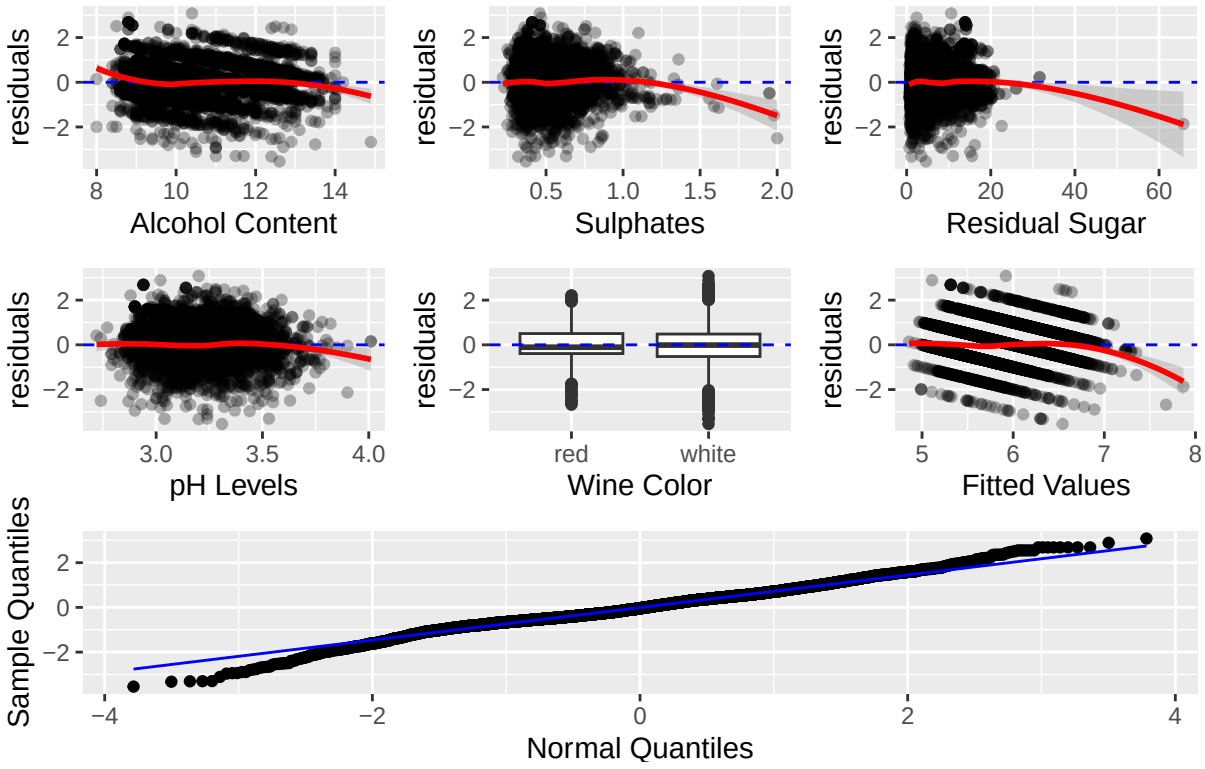



```
resid_xpanel(wine_lm_interaction, smoother = TRUE)
```



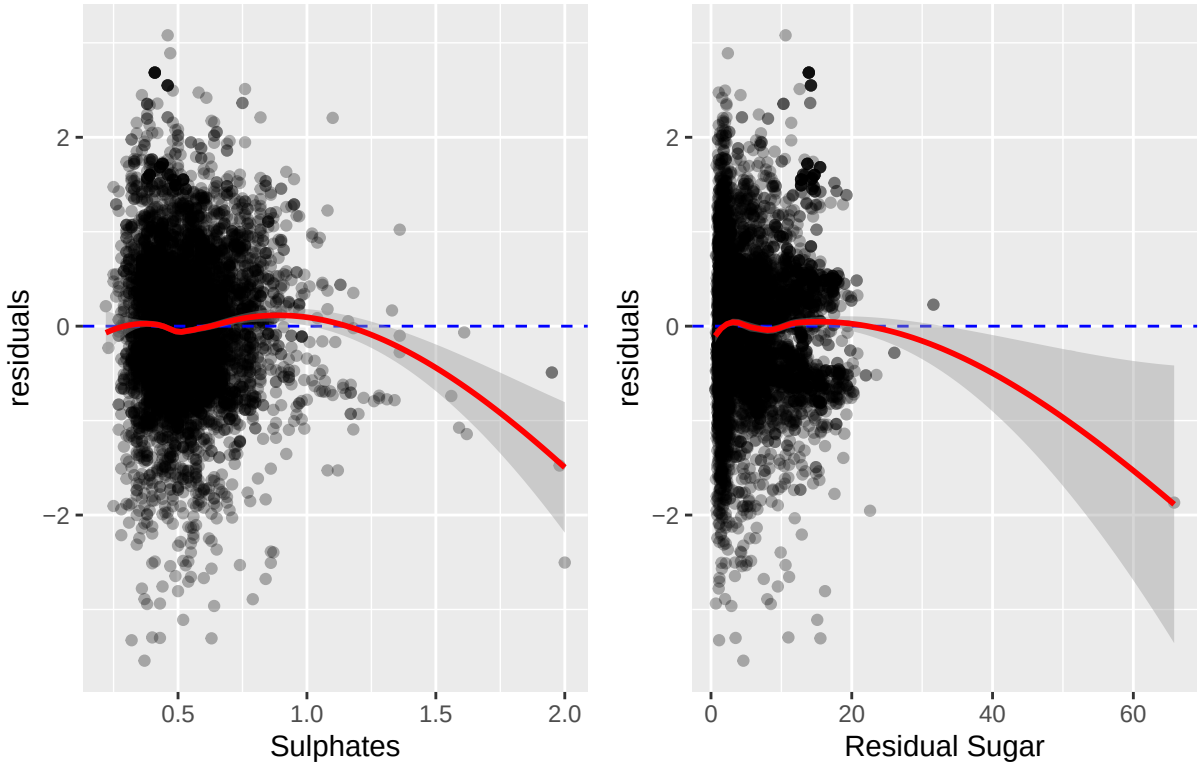

```
## Function to create residual plots
create_res_plot <- function(data) {
  # augment data for plotting
  wine_lm_aug <- augment(data, newdata = wine_dat)

  #determine if it's transfomred model
  if (identical(data, wine_lm_interaction)) {
    wine_lm_aug$var_sul <- wine_lm_aug$sulphates
    wine_lm_aug$var_sug <- wine_lm_aug$residual.sugar
    sul_name <- "Sulphates"
    sug_name <- "Residual Sugar"
  } else {
    wine_lm_aug$var_sul <- log(wine_lm_aug$sulphates)
    wine_lm_aug$var_sug <- log(wine_lm_aug$residual.sugar)
    sul_name <- "Log(Sulphates)"
    sug_name <- "Log(Residual Sugar)"
  }
  # residual plot
  wine_lm_res1 <- ggplot(wine_lm_aug, aes(x = alcohol, y = .resid)) +
    geom_point(alpha = 0.3) +
    geom_hline(yintercept = 0, linetype = "dashed", color = "blue") +
    geom_smooth(method = "loess", color = "red") +
    labs(x = "Alcohol Content", y = "residuals")

  wine_lm_res2 <- ggplot(wine_lm_aug, aes(x = var_sul, y = .resid)) +
    geom_point(alpha = 0.3) +
    geom_hline(yintercept = 0, linetype = "dashed", color = "blue") +
```

```
    geom_smooth(method = "loess", color = "red") +
    labs(x = sul_name, y = "residuals")

  wine_lm_res3 <- ggplot(wine_lm_aug, aes(x = var_sug, y = .resid)) +
    geom_point(alpha = 0.3) +
    geom_hline(yintercept = 0, linetype = "dashed", color = "blue") +
    geom_smooth(method = "loess", color = "red") +
    labs(x = sug_name, y = "residuals")

  wine_lm_res4 <- ggplot(wine_lm_aug, aes(x = pH, y = .resid)) +
    geom_point(alpha = 0.3) +
    geom_hline(yintercept = 0, linetype = "dashed", color = "blue") +
    geom_smooth(method = "loess", color = "red") +
    labs(x = "pH Levels", y = "residuals")

  wine_lm_res5 <- ggplot(wine_lm_aug, aes(x = color, y = .resid)) +
    geom_boxplot() +
    geom_hline(yintercept = 0, linetype = "dashed", color = "blue") +
    labs(x = "Wine Color", y = "residuals")

  wine_lm_res6 <- ggplot(wine_lm_aug, aes(x = .fitted, y = .resid)) +
    geom_point(alpha = 0.3) +
    geom_hline(yintercept = 0, linetype = "dashed", color = "blue") +
    geom_smooth(method = "loess", color = "red") +
    labs(x = "Fitted Values", y = "residuals")

  # normal qq plot
  wine_lm_qq <- ggplot(wine_lm_aug, aes(sample = .resid))+
    geom_qq() +
    geom_qq_line(color = "blue") +
    labs(y = "Sample Quantiles", x = "Normal Quantiles")

  combined_plot <- (wine_lm_res1 | wine_lm_res2 | wine_lm_res3) /
    (wine_lm_res4 | wine_lm_res5 | wine_lm_res6) / wine_lm_qq
  combined_plot +
    plot_layout(guides = 'collect') +
    plot_annotation(title = "Residual Plot and Normal Q-Q Plot of Wine Quality MLR")
}

# Residual Plot for non-transfomred model
create_res_plot(wine_lm_interaction)
```




```
# Selected Residual Plots for report
wine_lm_aug <- augment(wine_lm_interaction, newdata = wine_dat)
wine_lm_res2 <- ggplot(wine_lm_aug, aes(x = sulphates, y = .resid)) +
  geom_point(alpha = 0.3) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "blue") +
  geom_smooth(method = "loess", color = "red") +
  labs(x = "Sulphates", y = "residuals")
wine_lm_res3 <- ggplot(wine_lm_aug, aes(x = residual.sugar, y = .resid)) +
  geom_point(alpha = 0.3) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "blue") +
  geom_smooth(method = "loess", color = "red") +
  labs(x = "Residual Sugar", y = "residuals")
combined_plot <- (wine_lm_res2 | wine_lm_res3)
combined_plot +
  plot_layout(guides = 'collect') +
  plot_annotation(title = "Selected Residual Plot of Wine Quality MLR")
```

Selected Residual Plot of Wine Quality MLR

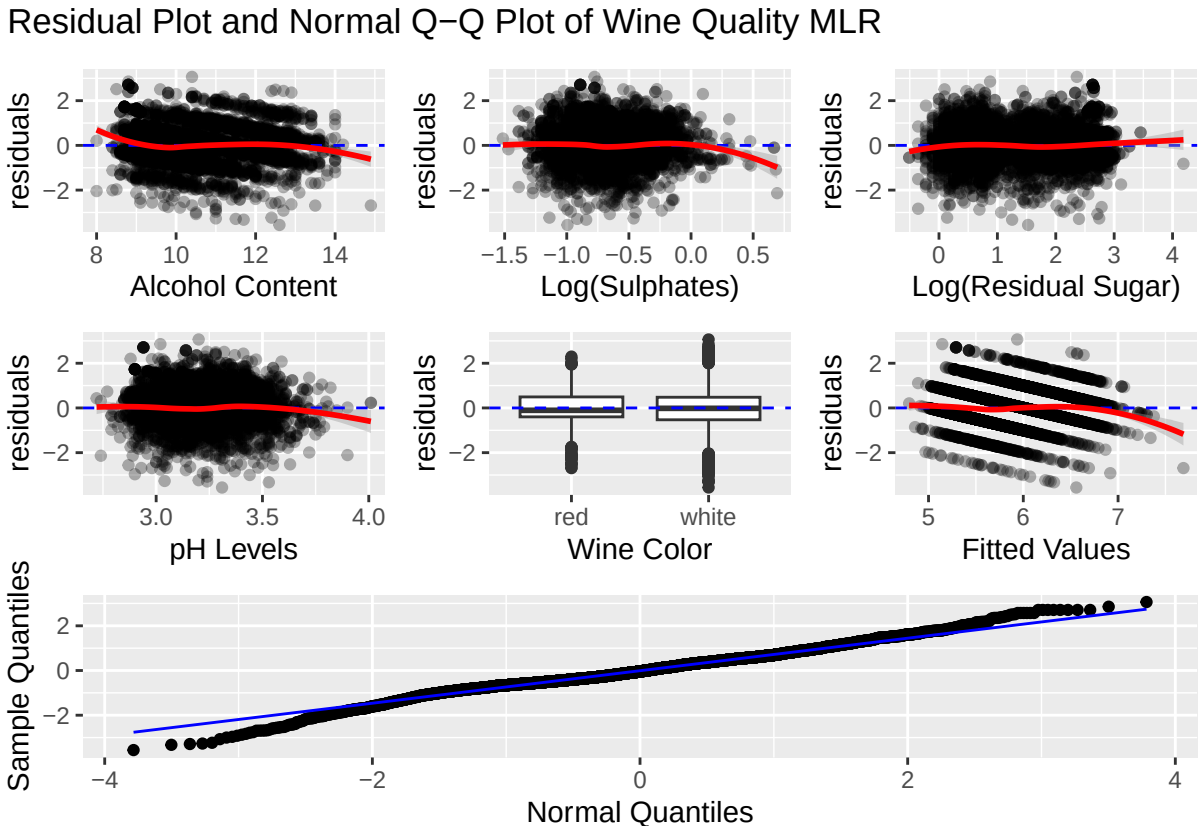



```
## transformed model
wine_lm_interaction_tran <- lm(quality ~ alcohol +
                                 log(sulphates) + log(residual.sugar) +
                                 pH*color, data = wine_dat)
summary(wine_lm_interaction_tran)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + log(sulphates) + log(residual.sugar) +
##     pH * color, data = wine_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5592 -0.4992 -0.0395  0.4820  3.0638
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.315962   0.414703  10.407  < 2e-16 ***
## alcohol               0.352826   0.008495  41.533  < 2e-16 ***
## log(sulphates)        0.369156   0.042944   8.596  < 2e-16 ***
## log(residual.sugar)   0.124058   0.012416   9.992  < 2e-16 ***
## pH                   -0.694190   0.124708  -5.567 2.70e-08 ***
## colorwhite           -2.947484   0.479598  -6.146 8.43e-10 ***
## pH:colorwhite         0.972984   0.145300   6.696 2.32e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7642 on 6490 degrees of freedom
```

```
## Multiple R-squared:  0.2349, Adjusted R-squared:  0.2342
## F-statistic: 332.1 on 6 and 6490 DF,  p-value: < 2.2e-16
```

```
## residual plots
create_res_plot(wine_lm_interaction_tran)
```




```
## confidence interval for coefficients
confint(wine_lm_interaction_tran)
```

```
##                          2.5 %     97.5 %
## (Intercept)          3.5030078  5.1289159
## alcohol              0.3361723  0.3694787
## log(sulphates)       0.2849707  0.4533413
## log(residual.sugar)  0.0997192  0.1483972
## pH                  -0.9386585 -0.4497216
## colorwhite          -3.8876554 -2.0073132
## pH:colorwhite        0.6881475  1.2578207
```

```
confint(wine_lm_interaction_tran)*log(2) # un-transforming logged
```

```
##                           2.5 %     97.5 %
## (Intercept)          2.42809997  3.5550936
## alcohol              0.23301691  0.2561032
## log(sulphates)       0.19752662  0.3142323
## log(residual.sugar)  0.06912008  0.1028611
## pH                  -0.65062853 -0.3117233
## colorwhite          -2.69471735 -1.3913635
## pH:colorwhite        0.47698750  0.8718549
```

```
                                        # predictors to the origianl scale
```

Model Comparison

```
# nested transformation model
wine_lm_base_tran <- lm(quality ~ alcohol + log(sulphates) +
                          log(residual.sugar) + pH, data = wine_dat)

summary(wine_lm_base_tran)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + log(sulphates) + log(residual.sugar) +
##     pH, data = wine_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5236 -0.5056 -0.0318  0.4949  3.0854
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.339614   0.226036  10.351  < 2e-16 ***
## alcohol              0.361335   0.008491  42.555  < 2e-16 ***
## log(sulphates)       0.232490   0.038734   6.002 2.05e-09 ***
## log(residual.sugar)  0.147694   0.012078  12.229  < 2e-16 ***
## pH                  -0.109696   0.062844  -1.746   0.0809 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.772 on 6492 degrees of freedom
## Multiple R-squared:  0.2189, Adjusted R-squared:  0.2184
## F-statistic: 454.8 on 4 and 6492 DF,  p-value: < 2.2e-16
```

```
#model comparison with transformed model
anova(wine_lm_base_tran, wine_lm_interaction_tran)
```

```
## Analysis of Variance Table
##
## Model 1: quality ~ alcohol + log(sulphates) + log(residual.sugar) + pH
## Model 2: quality ~ alcohol + log(sulphates) + log(residual.sugar) + pH *
##     color
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   6492 3869.5
## 2   6490 3790.0  2     79.44 68.016 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## AIC's
AIC(wine_lm_interaction)
```

```
## [1] 14955.76
```

```
AIC(wine_lm_interaction_tran)
```

```
## [1] 14952.05
```

```r
AIC(wine_lm_base_tran)
```

```
## [1] 15082.82
```

Confidence Intervals

```r
# linear comb for white wine ph
wine_lm_interaction_tran$coefficients[5] -> redPH
wine_lm_interaction_tran$coefficients[7] -> whtPH_int

whtPH <- (redPH + whtPH_int)[[1]]

vcov(wine_lm_interaction_tran)[c(5, 7), c(5, 7)] -> ph_cov

# point estimate
ph_se <- sqrt(ph_cov[[1, 1]] + ph_cov[[2, 2]] + 2 * ph_cov[[1, 2]])

test_stat <- whtPH/ph_se # t-stat
2*pt(test_stat, df = 6497 - 7, lower.tail = F) # p-val
```

```
## [1] 0.0001759616
```

```r
# confidence interval
wht_ci <- whtPH + c(-1, 1)*qt(0.975, df = 6497 - 7)*ph_se

# untransformed predictor
wine_lm_interaction_tran$coefficients[3] * log(2)
```

```
## log(sulphates)
##      0.2558794
```

```r
wine_lm_interaction_tran$coefficients[4] * log(2)
```

```
## log(residual.sugar)
##          0.08599059
```

Multicolinearity

```r
#checking for VIF - must use base model, not interaction model
vif(wine_lm_base)
```

```
##        alcohol      sulphates residual.sugar             pH
##       1.156899       1.066847       1.252074       1.103660
```

```r
vif(lm(quality ~ alcohol + sulphates + residual.sugar + pH + color,
       data = wine_dat)) # Might be this one; the base model doesn't have a color term
```

```
##        alcohol      sulphates residual.sugar             pH          color
##       1.194803       1.313478       1.373106       1.164569       1.563968
```

```r
vif(lm(quality ~ alcohol + log(sulphates) + log(residual.sugar) + pH + color,
       data = wine_dat)) # or the transformed one
```

```
##             alcohol      log(sulphates) log(residual.sugar)                  pH
##            1.139456            1.330182            1.270040            1.168728
##               color
##            1.517815
```

Influential Statistics

```
#figuring out outliers based on scatterplot matrix
mean_values <- wine_dat %>%
  summarize(across(c(alcohol, sulphates, pH, residual.sugar, citric.acid, fixed.acidity),
                   mean, na.rm = TRUE))
mean_values
```

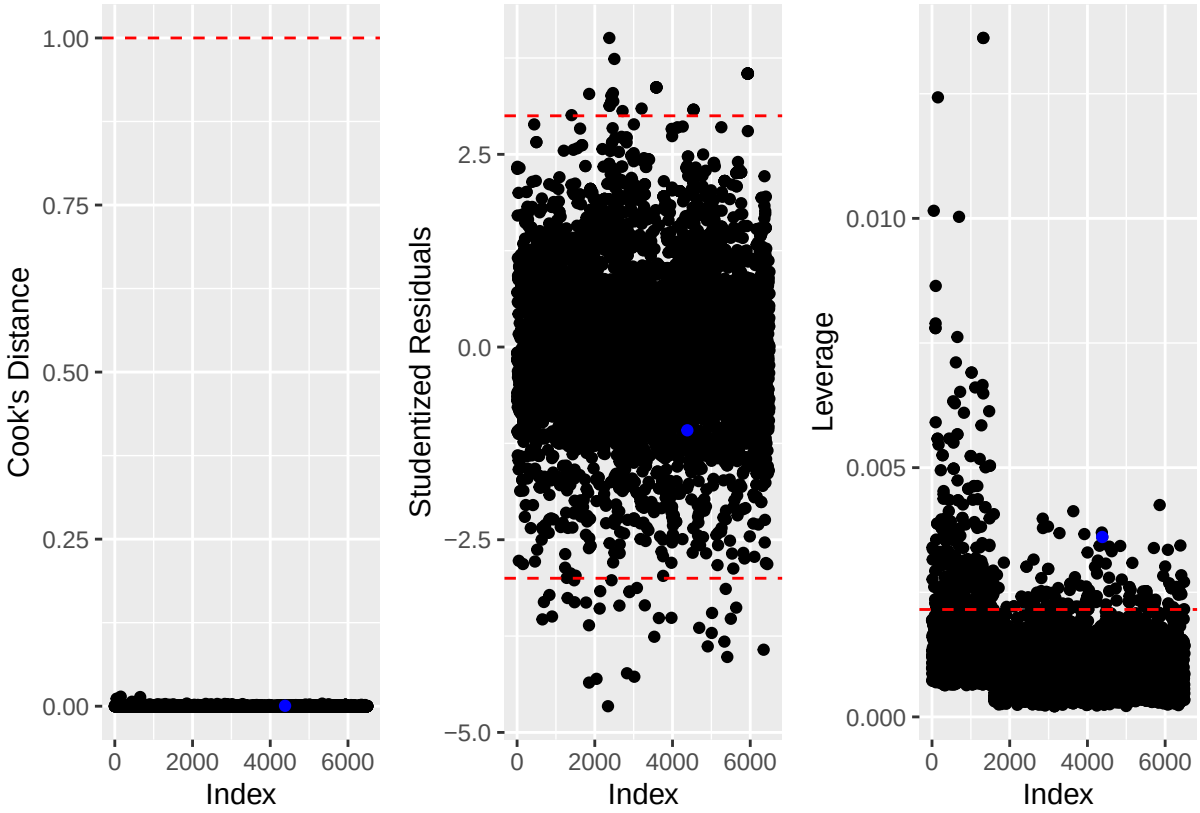
```
##   alcohol sulphates       pH residual.sugar citric.acid fixed.acidity
## 1 10.4918 0.5312683 3.218501       5.443235   0.3186332      7.215307
```

```
row_num <- 4381
case_values <- wine_dat[row_num,
                        c("alcohol", "sulphates", "pH", "residual.sugar",
                          "citric.acid", "fixed.acidity")]
case_values
```

```
##      alcohol sulphates   pH residual.sugar citric.acid fixed.acidity
## 4381    11.7      0.69 3.39           65.8         0.6           7.8
```

```
#This case has a very large residual sugar
```

```
# checking for influential observations
wine_dat_row <- wine_dat %>% mutate(case = row_number())
AUGwine_lm_interaction_tran <- augment(wine_lm_interaction_tran,
                                       data = wine_dat_row)

# Cook's Distance
ggplot(data = AUGwine_lm_interaction_tran, aes(x = case, y = .cooksd)) +
  geom_point() +
  geom_point(data = AUGwine_lm_interaction_tran %>%
               filter(case == 4381), color = "blue") +
  geom_hline(yintercept=1, linetype="dashed", color = "red")  +
  labs(y = "Cook's Distance", x = "Index") -> cookDistancePlot

# Studentized Residuals
ggplot(AUGwine_lm_interaction_tran, aes(x = case, y = .std.resid)) + geom_point() +
  geom_point(data = AUGwine_lm_interaction_tran %>%
               filter(case == 4381), color = "blue") +
  geom_hline(yintercept=3, linetype="dashed", color = "red")  +
  geom_hline(yintercept=-3, linetype="dashed", color = "red")  +
  labs(y = "Studentized Residuals", x = "Index") -> studResiPlot

# Leverages
ggplot(AUGwine_lm_interaction_tran, aes(x = case, y = .hat)) +
  geom_point() +
  geom_point(data = AUGwine_lm_interaction_tran %>%
               filter(case == 4381), color = "blue") +
  geom_hline(yintercept=2*7/6497, linetype="dashed", color = "red")  +
  labs(y = "Leverage", x = "Index") -> leveragePlot

cookDistancePlot|studResiPlot|leveragePlot # for report
```




```
# Selection of influential stats
high_L <- AUGwine_lm_interaction_tran %>% dplyr::select(case, .hat) %>%
  arrange(desc(.hat)) %>%
  filter(.hat > 2*7/6497) # greater than twice the average
high_SR <- AUGwine_lm_interaction_tran %>% dplyr::select(case, .std.resid) %>%
  arrange(desc(.std.resid)) %>%
  filter(.std.resid > 3) # greater than 3 se's
high_CD <- AUGwine_lm_interaction_tran %>% dplyr::select(case, .cooksd) %>%
  arrange(desc(.cooksd)) %>%
  filter(.cooksd > 1) # greater than 1

# List of influential stats
high_L
```

```
## # A tibble: 380 x 2
##     case    .hat
##    <int>   <dbl>
##  1  1317 0.0136
##  2  1322 0.0136
##  3   152 0.0124
##  4    46 0.0102
##  5   696 0.0100
##  6    96 0.00865
##  7    93 0.00789
##  8    87 0.00780
##  9    92 0.00780
## 10   653 0.00762
## # i 370 more rows
```

```
high_SR
```

```
## # A tibble: 29 x 2
##     case .std.resid
##    <int>      <dbl>
##  1  2374       4.01
##  2  2504       3.74
##  3  5932       3.55
##  4  5933       3.55
##  5  5934       3.55
##  6  5935       3.55
##  7  5936       3.55
##  8  5937       3.55
##  9  5938       3.55
## 10  5940       3.55
## # i 19 more rows
```

```
high_CD
```

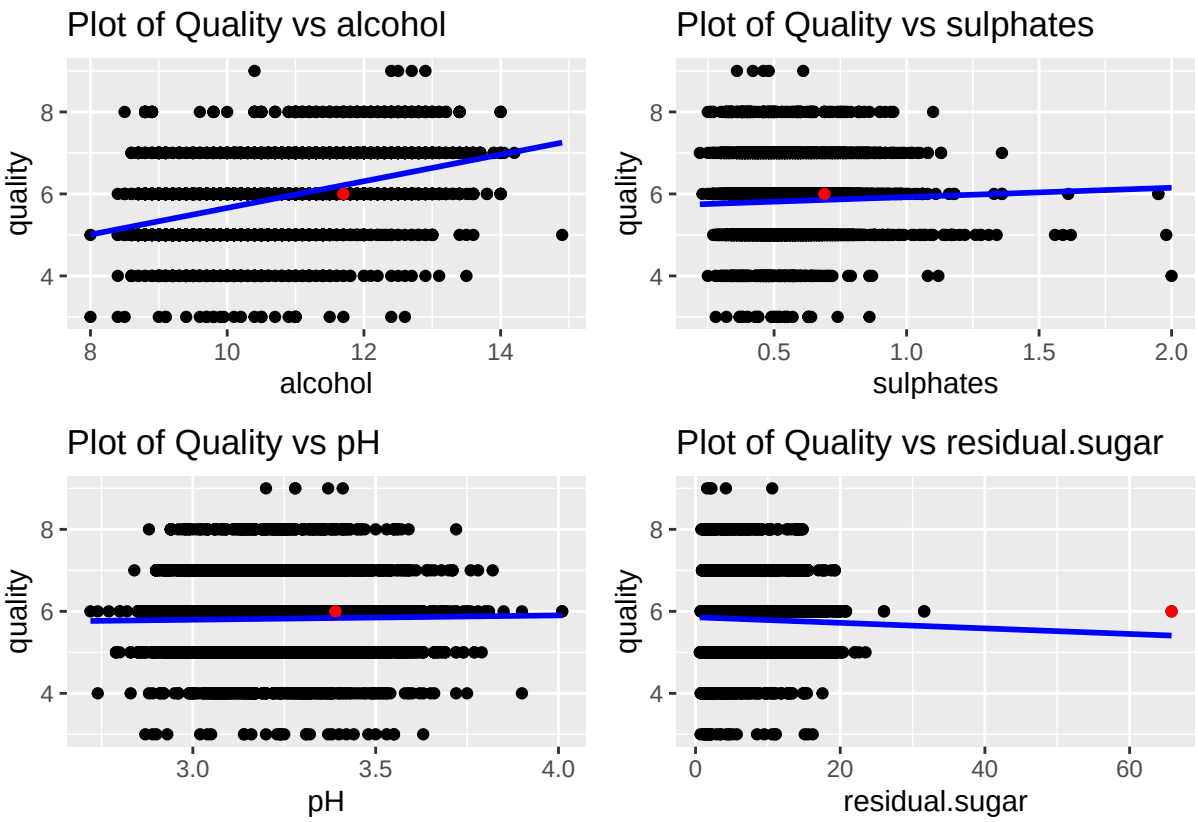
```
## # A tibble: 0 x 2
## # i 2 variables: case <int>, .cooksd <dbl>
```

```
## Reg on no outlier
wine_dat_no_outliers <- wine_dat %>%
  filter(residual.sugar != 65.8)
winenooutlierslm <- lm(quality ~ alcohol + sulphates + residual.sugar + pH*color,
                       data = wine_dat_no_outliers)
summary(winenooutlierslm) #doesn't change conclusion
```

```
##
## Call:
## lm(formula = quality ~ alcohol + sulphates + residual.sugar +
##     pH * color, data = wine_dat_no_outliers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5451 -0.4979 -0.0397  0.4817  3.0750
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.600042   0.426319   8.444  < 2e-16 ***
## alcohol         0.358163   0.008737  40.994  < 2e-16 ***
## sulphates       0.627085   0.074105   8.462  < 2e-16 ***
## residual.sugar  0.024378   0.002391  10.194  < 2e-16 ***
## pH             -0.655899   0.125416  -5.230 1.75e-07 ***
## colorwhite     -2.962047   0.483900  -6.121 9.83e-10 ***
## pH:colorwhite   0.973157   0.146380   6.648 3.21e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7641 on 6489 degrees of freedom
## Multiple R-squared:  0.2352, Adjusted R-squared:  0.2345
## F-statistic: 332.6 on 6 and 6489 DF,  p-value: < 2.2e-16
```

```
# visuals of (not) removing the outlier
create_inf_plot <- function(var_name) {
```

```
  ggplot(data = wine_dat,
         aes_string(x = var_name, y = "quality")) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE, color = "blue") +
    geom_point(data = wine_dat %>% filter(residual.sugar == 65.8),
               color = "red") +
    geom_smooth(data = wine_dat %>% filter(residual.sugar == 65.8),
                method = "lm", se = FALSE, color = "red") +
    labs(title = paste("Plot of Quality vs", var_name))
}

plot_list2 <- map(variable_names, create_inf_plot)
plot_grid2 <- wrap_plots(plot_list2, ncol = 2)
plot_grid2
```



Sequential Selection

```
set.seed(67393937)

upper_model <- lm(quality ~ (pH + alcohol + log(residual.sugar) + color + log(sulphates) +
                               fixed.acidity + volatile.acidity + citric.acid + chlorides +
                               free.sulfur.dioxide + total.sulfur.dioxide + density)^2,
                  data = wine_dat)
lower_model <- lm(quality ~ 1, data = wine_dat)

#backward selection
backwardSelectModel <- stepAIC(upper_model, scope = list(lower = lower_model,
```

```
                                                    upper = upper_model),
                              direction = "backward")
summary(backwardSelectModel)

#forward selection
forwardSelectModel <- stepAIC(lower_model, scope = list(lower = lower_model,
                                                    upper = upper_model),
                              direction = "forward")
summary(forwardSelectModel)

anova(wine_lm_interaction_tran, backwardSelectModel)
anova(wine_lm_interaction_tran, forwardSelectModel)
```