Math 656 HW1

Jeff Gould

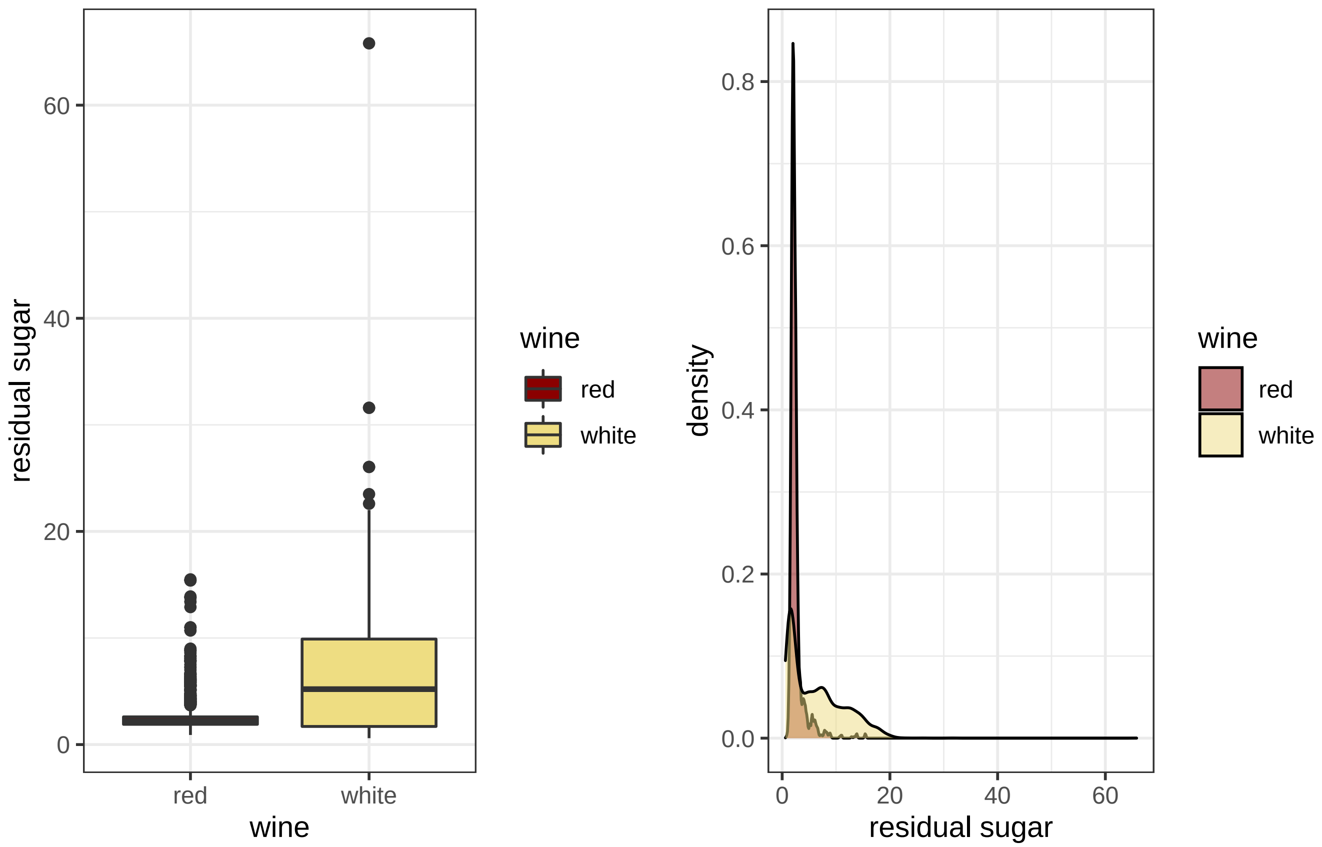
8/29/2020

First, load the two wine data sets and merge:

```
red_wine <- foreign::read.arff("winequality-red.arff") %>%
  mutate(wine = "red")

white_wine <- foreign::read.arff("winequality-white.arff") %>%
  mutate(wine = "white")

wine_all <- bind_rows(red_wine, white_wine)
```

First I look at the distribution of residual sugar for red and white wine. White wine tends to be sweeter than red, so I expected to see a higher amount of sugar. Sure enough, this was the case. I plot the box plot and the density for each:

```
box <- ggplot(data = wine_all) +
  geom_boxplot(aes(x = wine, y = `residual sugar`, fill = wine)) +
  theme_bw() +
  scale_fill_manual(values = c("red4", "lightgoldenrod"))

density_plot <- ggplot(data = wine_all) +
  geom_density(aes(x = `residual sugar`, fill = wine), alpha = 0.5) +
  theme_bw() +
  scale_fill_manual(values = c("red4", "lightgoldenrod"))
```

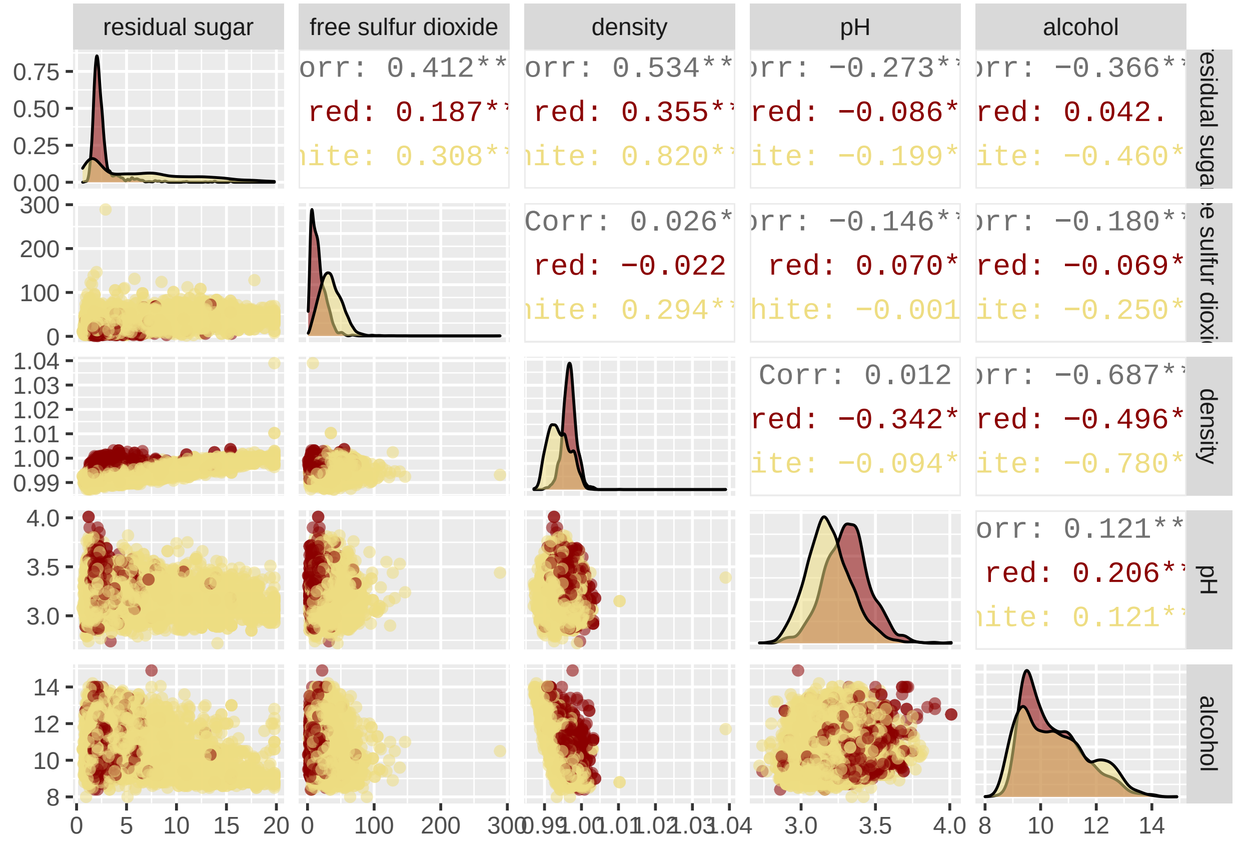


We see that this is the case. Red wines have a much narrower distribution of residual sugar compared to their white counterparts, and see that white wines usually have much more sugar. There are also a few outlier values for the white wine, which I found made some subsequent plots tough to read. So I winsorized the residual sugar tail to account for this

```
white_wine$`residual sugar` = DescTools::Winsorize(white_wine$`residual sugar`, probs = c(0,.995))
wine_all <- bind_rows(red_wine, white_wine)
```

Here we have matrix plots of the relationships between residual sugar, free sulfur dioxide, density, pH, and alcohol

```
GGally::ggpairs(wine_all[sample(c(1:nrow(wine_all)), size = nrow(wine_all)),],
                columns = c(4,6, 8,9,11), aes(alpha = 0.05, color = wine)) +
  scale_fill_manual(values = c("red4", "lightgoldenrod")) +
  scale_color_manual(values = c("red4", "lightgoldenrod"))
```



We again see the outliers at white wine in regards to density and free sulfur dioxide, and looking at the graphs we see that most of those are the same wines the we winsorized residual sugar for. However I don't want to overclean the data for this elementary analysis.

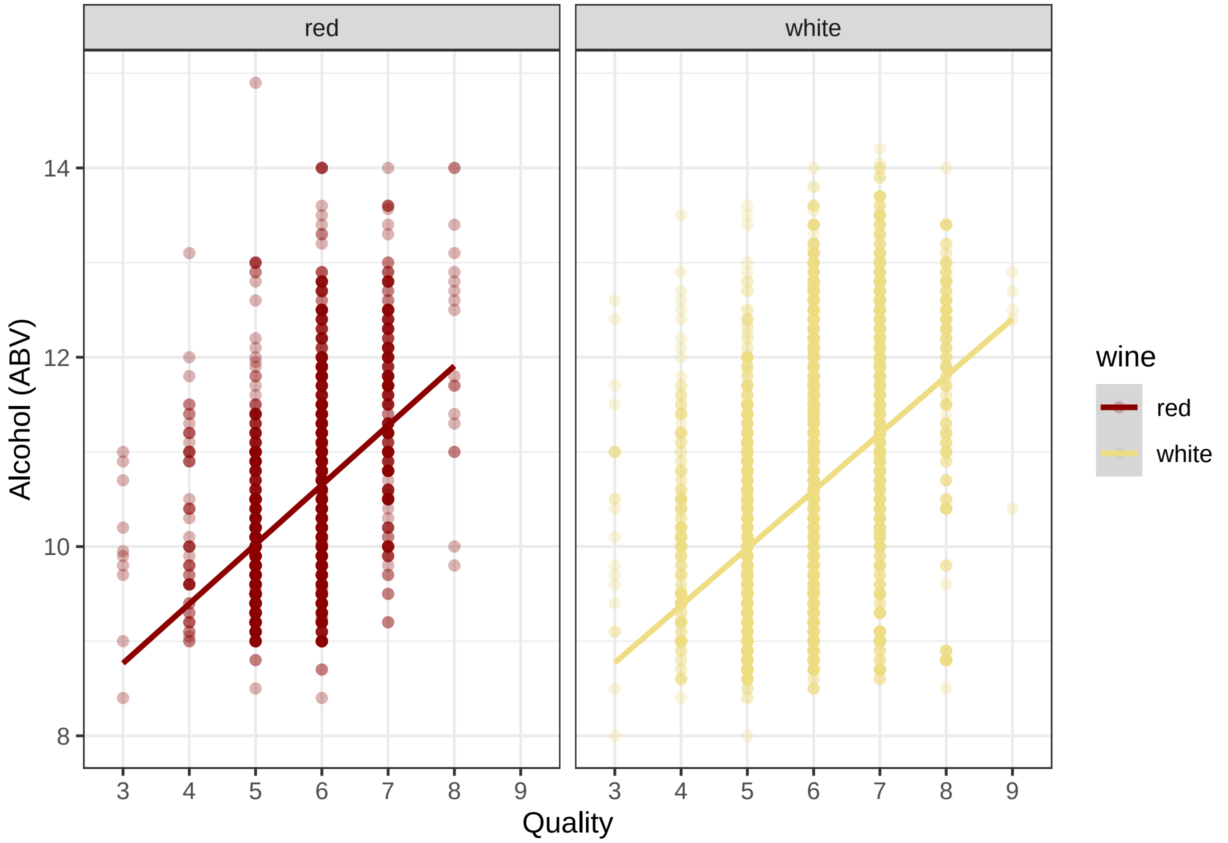
We see that strongest overall correlation is betweenresidual and density, with a correlation of -0.687. Chemically this makes intuitive sense, as the density of ethanol is lower than that of water. There is also a storng correlation between density and residual sugar, and while is is partially caused by the outliers in the white wine data set, the correlation of 0.355 for red wine is still the strongest for red wine outside of the relationship between alcohol and density. This again makes intuitive sense if one is familiar with the fermentation process. A higher residual sugar content usually means that the wine was not fermented as long as it could have been, leading to a lower alcohol content. Additionally, more sugar dissolved in a fluid increases its density.

Some other note of interest

- White wine tends to have a higher sulfur content. This could be important for people who might have reactions to sulfur
- Red wine tends to have a higher density than white wine. This aligns with common perception of red wine being "heavier" than white, or that it has a higher tannin content

Lastly, lets look at the relationships between wine quality and alcohol content:

```r
ggplot(wine_all, aes(x = quality, y = alcohol, color = wine, group = wine)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = lm, formula = y ~ x, level = 0) +
  facet_wrap(facets = "wine") +
  scale_color_manual(values = c("red4", "lightgoldenrod")) +
  theme_bw() +
  labs(x = "Quality", y = "Alcohol (ABV)")
```



```
require(magrittr)
wine_all %$%
  cor(as.numeric(quality), alcohol)
```

```
## [1] 0.4443185
```

```
wine_all %>%
  filter(wine == "red") %$%
  cor(as.numeric(quality), alcohol)
```

```
## [1] 0.4761663
```

```
wine_all %>%
  filter(wine == "white") %$%
  cor(as.numeric(quality), alcohol)
```

```
## [1] 0.4355747
```

We see that there is a moderate linear relationship between the quality of wine and it's alcohol content, and the total correlation between the two is 0.444, or $R^2 = 0.1974189$.