Coffee Ratings

What we know

1. The two most economically important varieties of coffee plant are the Arabica and the Robusta; ~60% of the coffee produced worldwide is Arabica and ~40% is Robusta. Arabica beans consist of 0.8–1.4% caffeine and Robusta beans consist of 1.7–4% caffeine.

2. Coffee cupping (or tasting) procedure involves deeply sniffing the coffee, then loudly slurping the coffee so it spreads to the back of the tongue. The coffee taster attempts to measure aspects of the coffee's taste, specifically the body (the texture or mouthfeel, such as oiliness), sweetness, acidity (a sharp and tangy feeling, like when biting into an orange), flavour (the characters in the cup), and aftertaste.

Imports

```
library(tidyverse)
library(tidytuesdayR)
library(scales)
library(dbplyr)
library(ggplot2)
theme_set(theme_light())
```

Fetching and processing the data

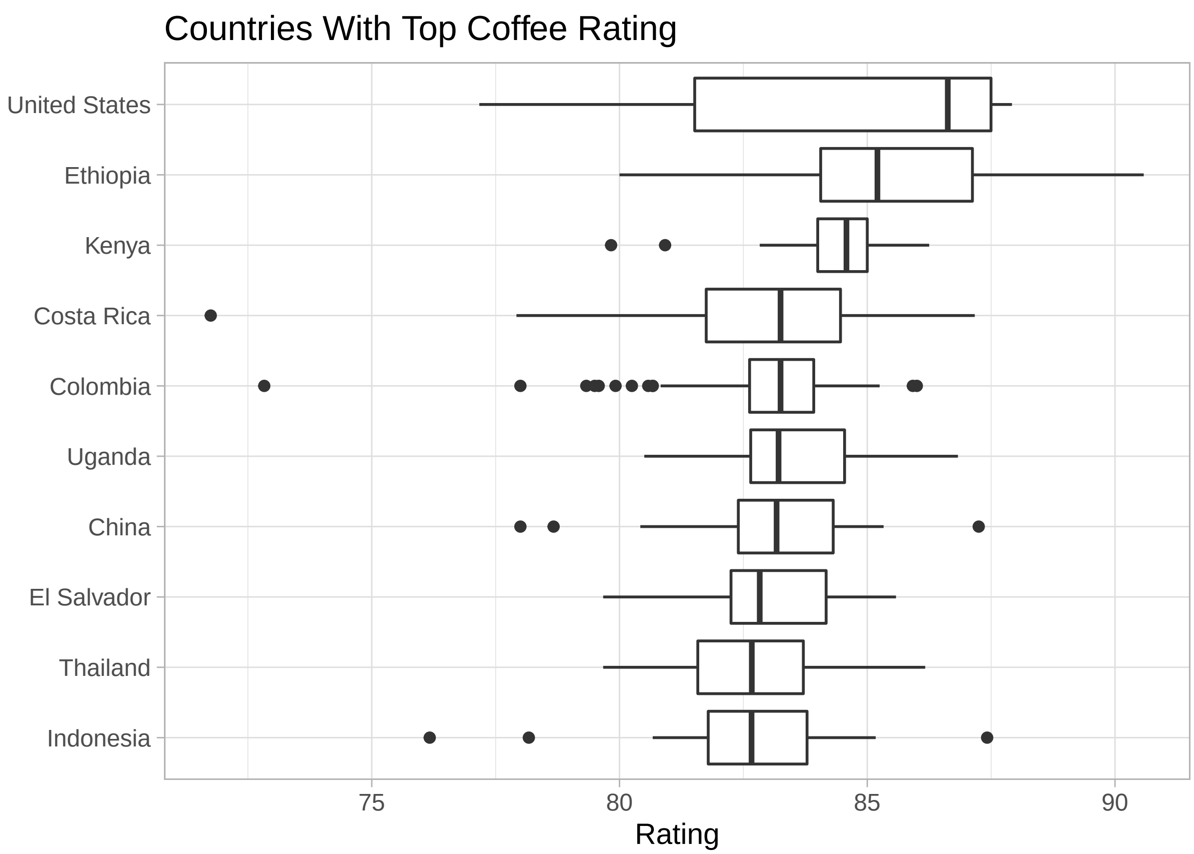
```
coffee_ratings <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/


## Add id
coffee_ratings <-
coffee_ratings %>%
mutate(coffee_id = row_number())

## Check for null percentages in each column
coffee_ratings %>%
  summarise(across(
    everything(),
    ~ mean(!is.na(.))
  )) %>%
  gather() %>%
  arrange(desc(value))
```

```
## # A tibble: 44 x 2
##    key                 value
##    <chr>               <dbl>
```

```
##   1 total_cup_points       1
##   2 species                1
##   3 number_of_bags         1
##   4 bag_weight             1
##   5 in_country_partner     1
##   6 grading_date           1
##   7 aroma                  1
##   8 flavor                 1
##   9 aftertaste             1
##  10 acidity                1
## # ... with 34 more rows
```

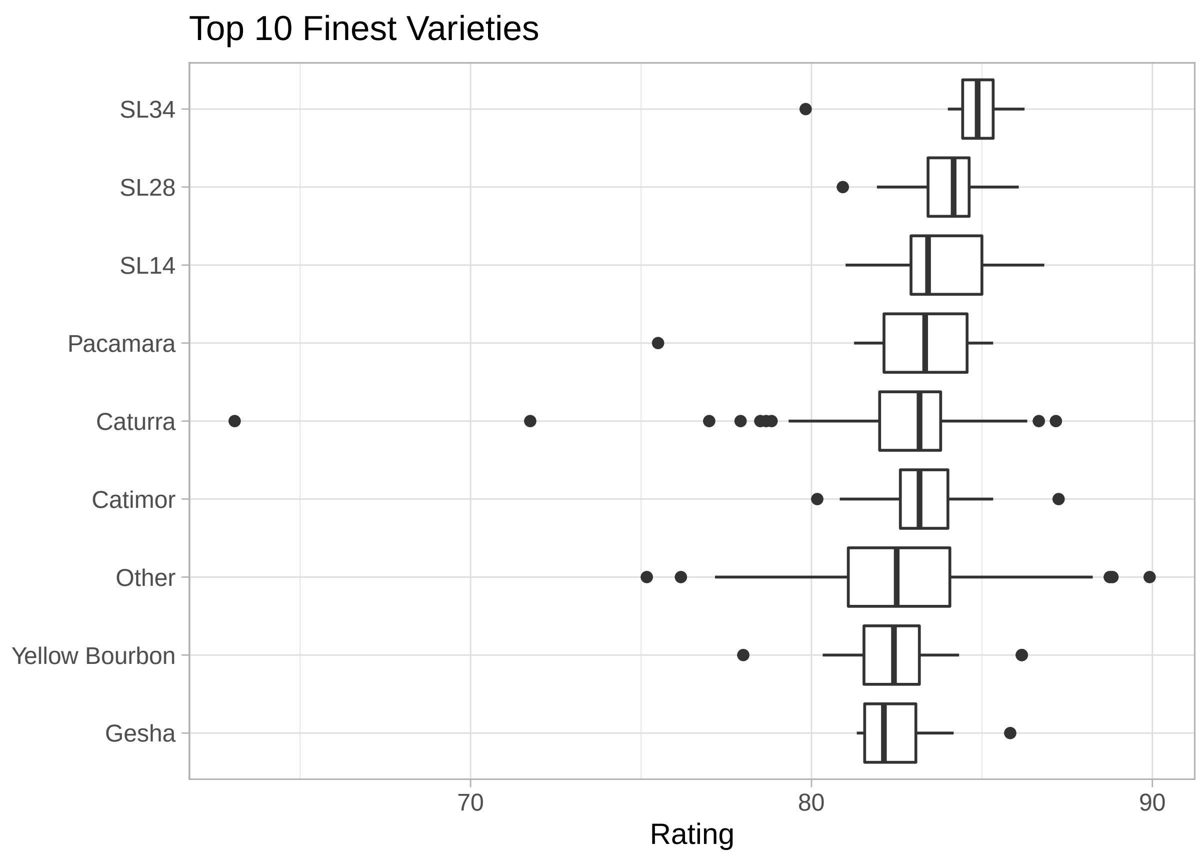
Which countries produce finest coffee

```
### Top rate countries
top_10_rated_countries <- coffee_ratings %>%
  group_by(country_of_origin)  %>%
  summarise(
    rating = weighted.mean(total_cup_points),
    number_off_products = n()
  ) %>%
  filter(number_off_products > 5) %>%
  arrange(desc(rating))  %>%
  head(10)

coffee_ratings %>%
  filter(!is.na(country_of_origin),
         country_of_origin %in% list(top_10_rated_countries$country_of_origin)[[1]]) %>%
  mutate(country_of_origin = fct_reorder(country_of_origin, total_cup_points)) %>%
  ggplot(aes(total_cup_points, country_of_origin)) +
  geom_boxplot()  +
  labs(title=str_to_title("Countries with top coffee rating"),
       x = "Rating",
       y = "")
```



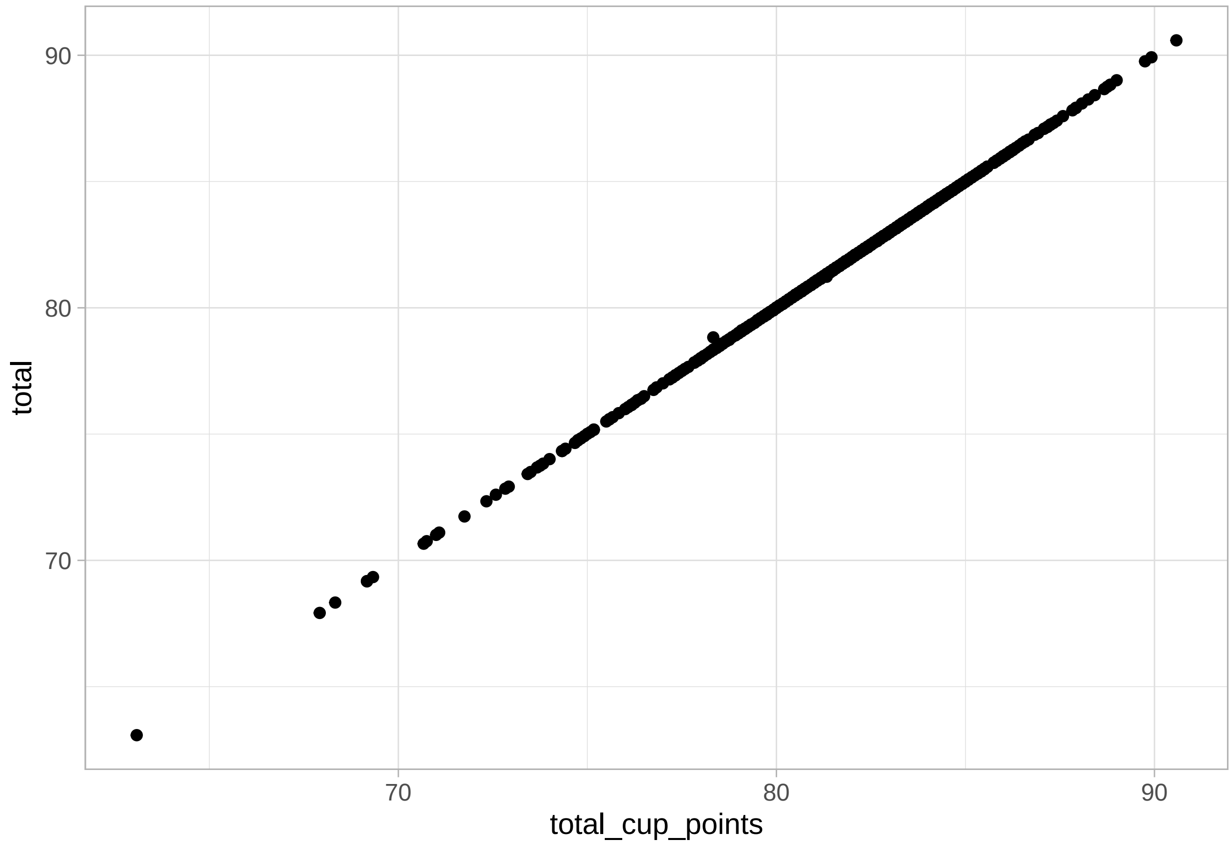
Which varieties of coffee are the finest

```
### Top rate varieties
top_10_rated_varieties <- coffee_ratings %>%
  group_by(variety)  %>%
  summarise(
    rating = weighted.mean(total_cup_points),
    number_off_products = n()
  ) %>%
  filter(number_off_products > 5) %>%
  arrange(desc(rating))  %>%
  head(10)

coffee_ratings %>%
  filter(!is.na(variety),
         variety %in% list(top_10_rated_varieties$variety)[[1]],
         total_cup_points > 10 ) %>%
  mutate(variety = fct_reorder(variety, total_cup_points)) %>%
  ggplot(aes(total_cup_points, variety)) +
  geom_boxplot()  +
  labs(title=str_to_title("Top 10 Finest varieties"),
       x = "Rating",
       y = "")
```



Check what contributes to total cup points calculation

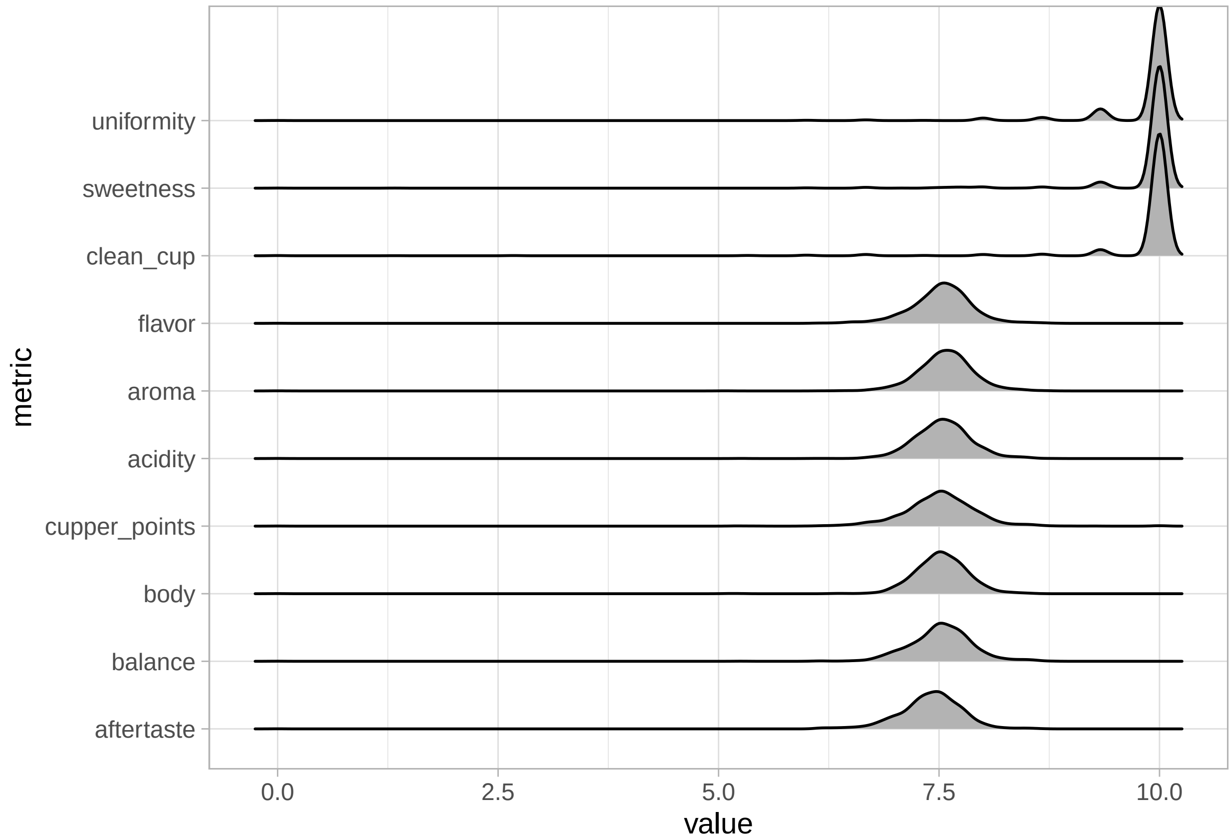
```
coffee_metrics <-
  coffee_ratings %>%
  pivot_longer(aroma:cupper_points, names_to = "metric", values_to = "value")


coffee_metrics  %>%
  group_by(coffee_id, total_cup_points)  %>%
  summarize(total = sum(value)) %>%
  filter(total >= 60) %>%
  ggplot(aes(total_cup_points, total)) +
  geom_point()
```



So that concludes that columns aroma:cupper_points are what contributes to total cup points

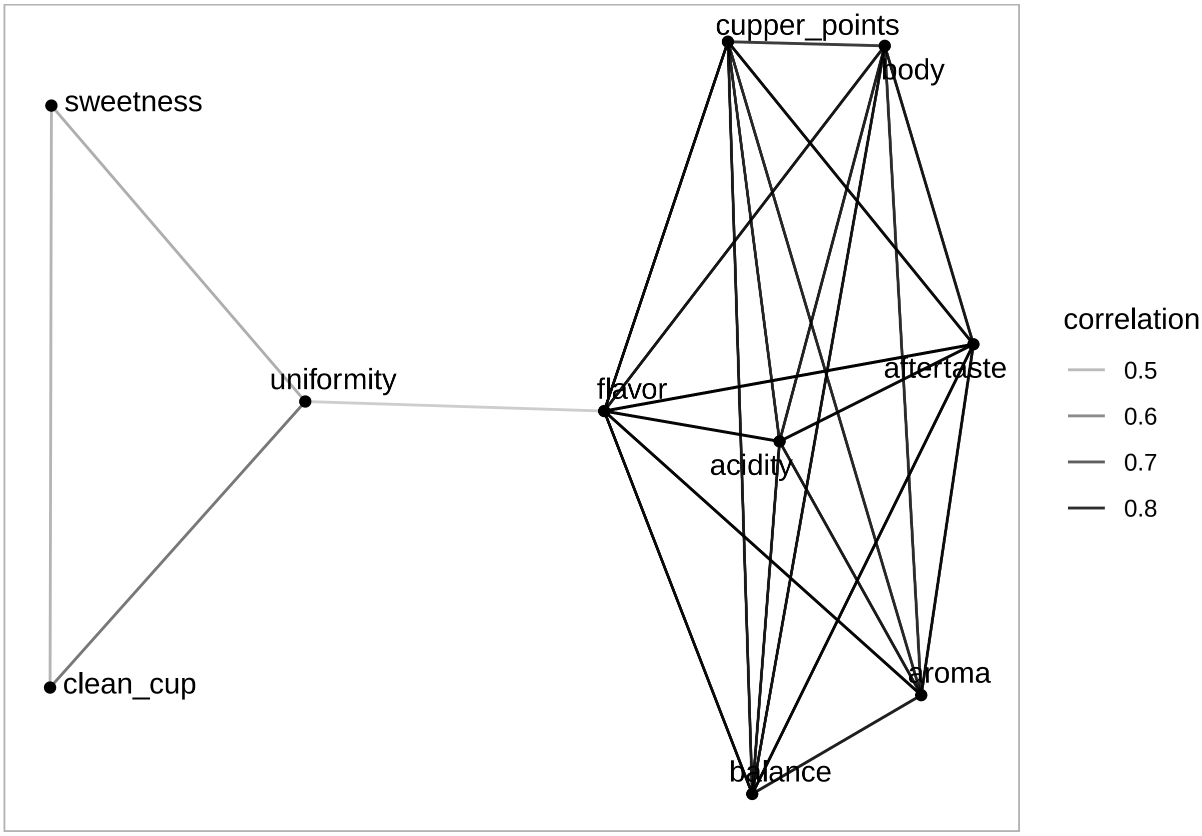
How do these features correlate

```
library(ggridges)

coffee_metrics %>%
  mutate(metric = fct_reorder(metric, value)) %>%
  ggplot(aes(x=value, y=metric)) +
  geom_density_ridges()
```



Most of metrics occur as displayed, we expect a strong corralation of flavor: aftertastean, also strong correlation uniformity:clean_cup and some relation between the two groups

```
library(widyr)
library(ggraph)
library(igraph)

coffee_metrics %>%
  pairwise_cor(metric, coffee_id, value, sort = TRUE) %>%
  head(50) %>%
  graph_from_data_frame() %>%
  ggraph() +
  geom_edge_link(aes(edge_alpha = correlation)) +
  geom_node_point() +
  geom_node_text(aes(label = name), repel = TRUE)
```



Both of the graphs above would suggest for example coffee with good flavor would have good aroma, aftertaste and balance.. etc

However, a good flavored coffee might or not be uniform and you can never predict wether it is clean or sweet