Project 5

Read in the dataset you will be working with:

```
coffee_ratings <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytues
day/master/data/2020/2020-07-07/coffee_ratings.csv')
```

Question: Is there a correlation between the processing method of coffee beans and their scoring by professionals?

Introduction: To answer this question, We will be using the **coffee_ratings** dataset. It is part of the *Coffee Quality Database* provided by Buzzfeed Data Scientist James LeDoux. The dataset includes over a thousand data points on various coffee bean varieties. The crucial columns we will need from the dataset are the coffee beans' professionally scored characteristics (aroma, flavor, aftertaste, acidity, body, balance), their overall score (total_cup_points), and what type of processing the beans underwent (processing_method).

Approach: We will use a PCA analysis to determine which characteristics are most indicative of processing method. After isolating the strongest principal components, we will use a k means clustering to find where the clusters for each processing method lie. The clustering will give us the final result on the strength and predictability of the processing method.

Analysis:

First we run a PCA to determine the best columns to run a kmeans on.

```
pca_fit <- coffee_ratings %>%
  na.omit() %>%
  select(where(is.numeric)) %>%
  scale() %>%
  prcomp()

summary(pca_fit)
```

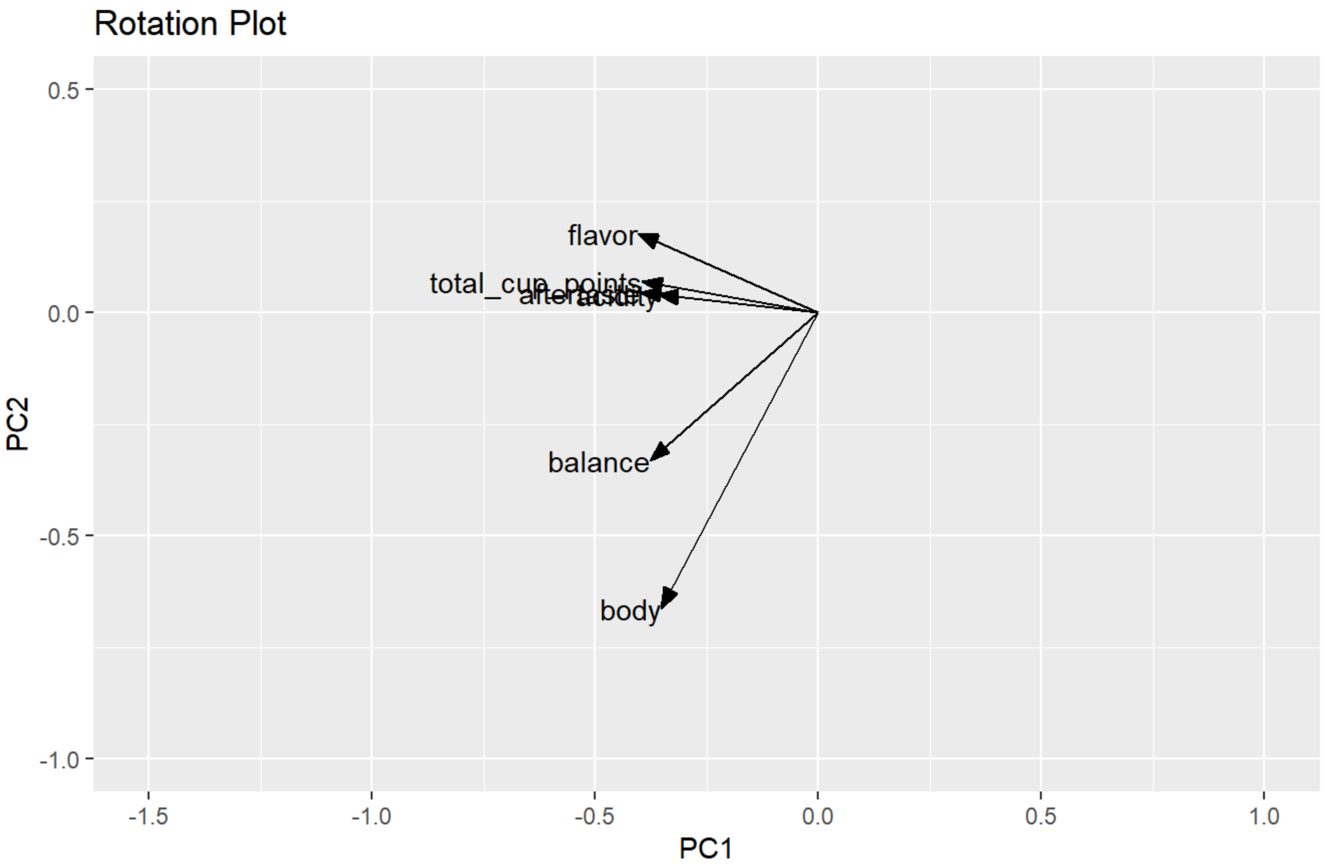
```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.5461 1.7507 1.4409 1.18830 1.04864 1.01270 0.92789
## Proportion of Variance 0.3412 0.1613 0.1093 0.07432 0.05788 0.05398 0.04532
## Cumulative Proportion  0.3412 0.5025 0.6118 0.68611 0.74398 0.79796 0.84328
##                            PC8     PC9    PC10    PC11    PC12    PC13   PC14
## Standard deviation     0.82072 0.77947 0.62754 0.61928 0.54460 0.47822 0.4159
## Proportion of Variance 0.03545 0.03198 0.02073 0.02018 0.01561 0.01204 0.0091
## Cumulative Proportion  0.87873 0.91070 0.93143 0.95162 0.96723 0.97926 0.9884
##                           PC15    PC16     PC17      PC18      PC19
## Standard deviation     0.35264 0.31096 0.003198 6.608e-17 2.541e-32
## Proportion of Variance 0.00654 0.00509 0.000000 0.000e+00 0.000e+00
## Cumulative Proportion  0.99491 1.00000 1.000000 1.000e+00 1.000e+00
```

Next, we filter down to the relevant columns for further processing. We plot a rotation plot of the new PCA to show strength of characteristics.

```
coffee_ratings <- coffee_ratings %>%
  select(
    total_cup_points, aroma, flavor, aftertaste, acidity, body, balance, processing_method
  ) %>%
  na.omit()

pca_fit <- coffee_ratings %>%
  select(where(is.numeric)) %>%
  scale() %>%
  prcomp()

arrow_style <- arrow(
  angle = 20, length = grid::unit(8, "pt"),
  ends = "first", type = "closed"
)

pca_fit %>%
  tidy(matrix = "rotation") %>%
  pivot_wider(
    names_from = "PC",
    values_from = "value",
    names_prefix = "PC"
  ) %>%
  ggplot(aes(PC1, PC2)) +
  geom_segment(
    xend = 0,
    yend = 0,
    arrow = arrow_style
  ) +
  geom_text(aes(label = column), hjust = 1) +
  xlim(-1.5, 1.0) +
  ylim(-1.0, 0.5) +
  coord_fixed() +
  labs(title = "Rotation Plot")
```

```
## Warning: Removed 1 rows containing missing values (geom_segment).
```

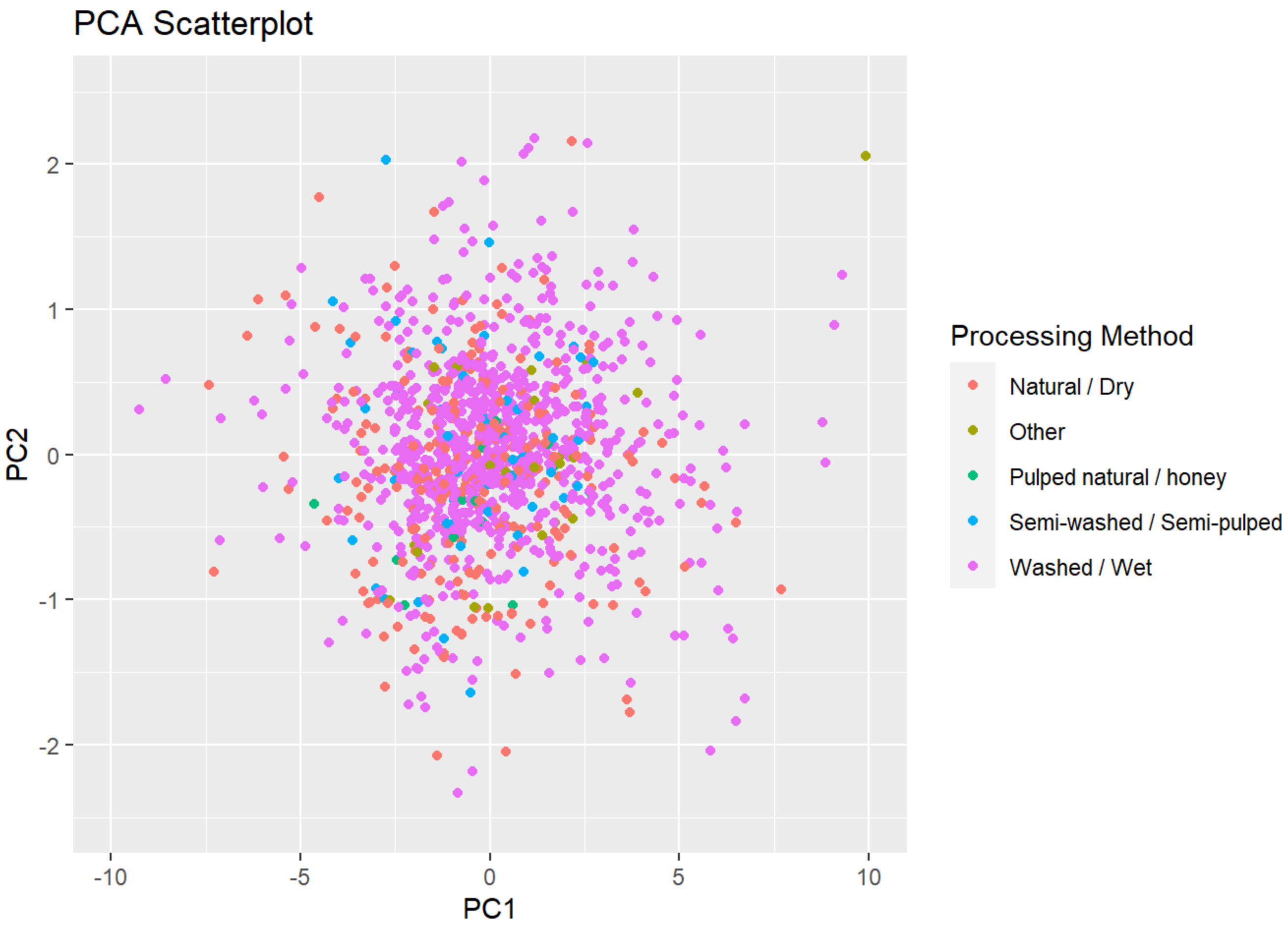
```
## Warning: Removed 1 rows containing missing values (geom_text).
```



We then plot a PCA scatter plot for comparison with the k means plot.

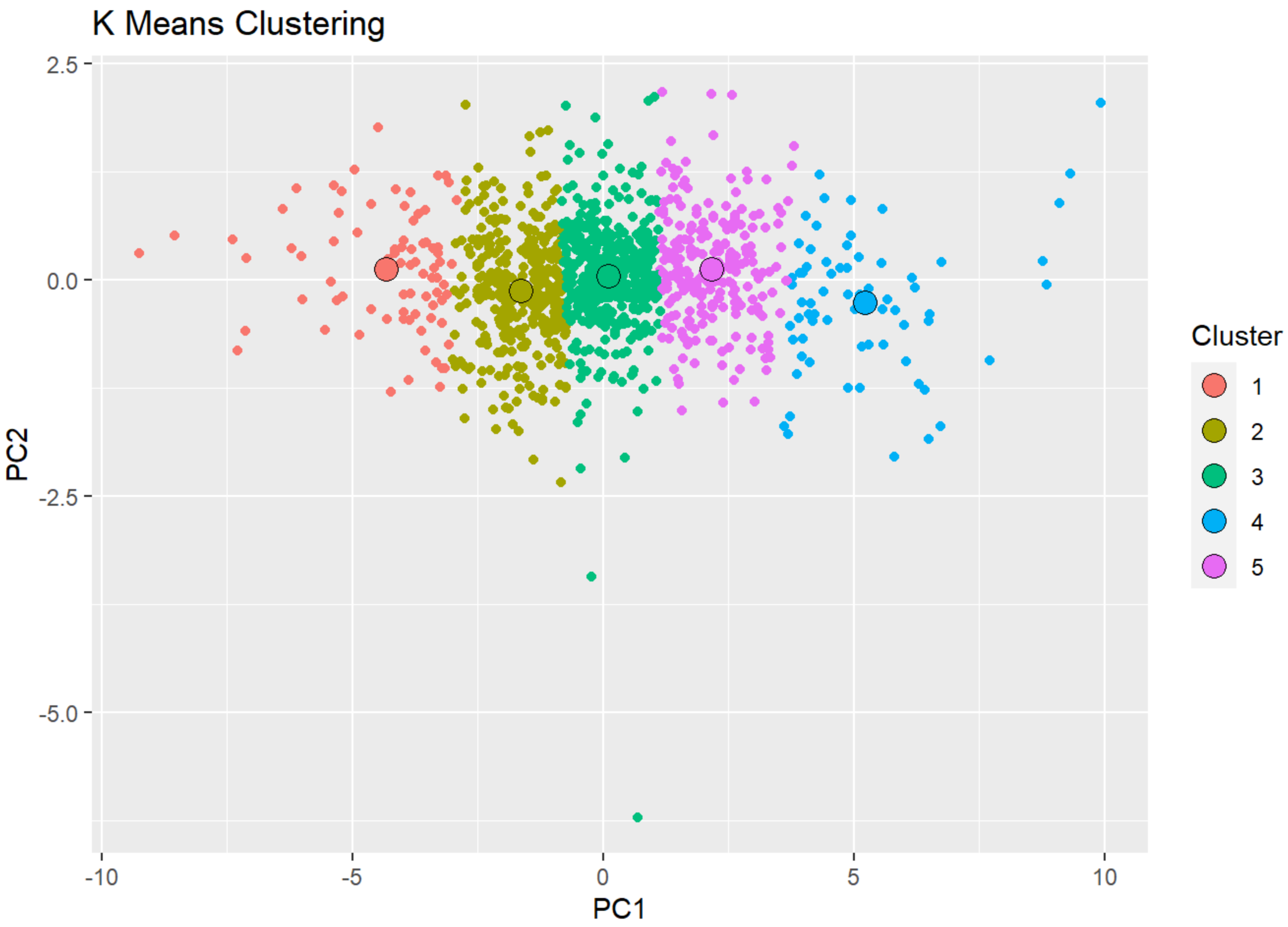
```
pca_fit %>%
  augment(coffee_ratings) %>%
  ggplot(aes(.fittedPC1, .fittedPC2)) +
  geom_point(aes(color = processing_method)) +
  xlim(-10, 10) +
  ylim(-2.5, 2.5) +
  labs(
    title = "PCA Scatterplot",
    x = "PC1",
    y = "PC2",
    color = "Processing Method"
  )
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



Finally, we run a k means with the correct number of clusters and plot. We can then compare this plot with the PCA scatterplot.

```
km_fit <- pca_fit$x[,1:2] %>%
  kmeans(centers = 5, nstart = 10)

km_fit %>%
  augment(pca_fit$x[,1:2]) %>%
  ggplot() +
  aes(PC1, PC2) +
  geom_point(
    aes(color = .cluster)
  ) +
  geom_point(
    data = tidy(km_fit),
    aes(fill = cluster),
    shape = 21, color = "black", size = 4
  ) +
  guides(color = "none") +
  labs(
    title = "K Means Clustering",
    fill = "Cluster"
  )
```



Discussion: Based on the results seen above, we can see that the PCA scatter plot does not show much differenciation among the processing methods. Still, we run a k means clustering to verify. Because the cluster does not follow any pattern seen in the PCA scatter plot, we can strongly conclude that a correlation does not exist between coffee ratings and the processing method of coffee bean. Critics do not have a preference for a particular processing method nor does a particular processing method particularly degrade or improve the result of a coffee bean.