STAT 456

Final report
(Starbucks Nutritional dataset)

Xiangyi Li

May 2, 2024

1 Summary

This report examines a Starbucks dataset containing the nutritional composition of 1147 Starbucks beverages with 15 variables. We are interested in the main ingredients of the beverages and some of the factors behind them, so we used PCA and FA methods to analyze the main ingredients and extract the hidden factors in the study. The study concluded that most of the Starbucks coffee beverages can be classified into four categories: carbohydrate, fat, caffeine, and dietary fiber. The shortcoming of this study is that we only used two methods of analysis, and there are many other ways to classify the nutritional composition of Starbucks beverages that could be more comprehensively interpreted. See below for a detailed report.

2 Dataset Introduce

2.1 Source

Official Starbucks Nutritional dataset from the pdf Starbucks Coffee Company Beverage Nutrition Information. Please view the data set information: https://github.com/rfordatascience/tidytuesday/blob/master/data/2021/2021-12-21/readme.md . Or, you can get the data from R code in appendix.

2.2 Variables

Discribe each variable in this dataset, including the type and the meaning and numbers of variables and observations I intend to use in the analysis.

This table shows the original variables in the dataset:

variable	**type of**	**meaning**
product_name	character	Product Name
size	character	short, tall, grande, venti
milk	double	Milk Type type of milk used
whip	double	Whip added or not (binary 0/1)
serv_size_m_L	double	Serving size in ml
calories	double	KCal
total_fat_g	double	Total fat grams
saturated_fat_g	double	Saturated fat grams
trans_fat_g	character	Trans fat grams
cholesterol_mg	double	Cholesterol mg
sodium_mg	double	Sodium milligrams
total_carbs_g	double	Total Carbs grams
fiber_g	character	Fiber grams
sugar_g	double	Sugar grams
caffeine_mg	double	Caffeine in milligrams

Table 1: Description of Original Variables

2.3 data exploration analysis

Start with outlier processing. removes outliers and missing values. The data dimension has changed from 1147×15 to 971×15, 176 outliers have been removed.

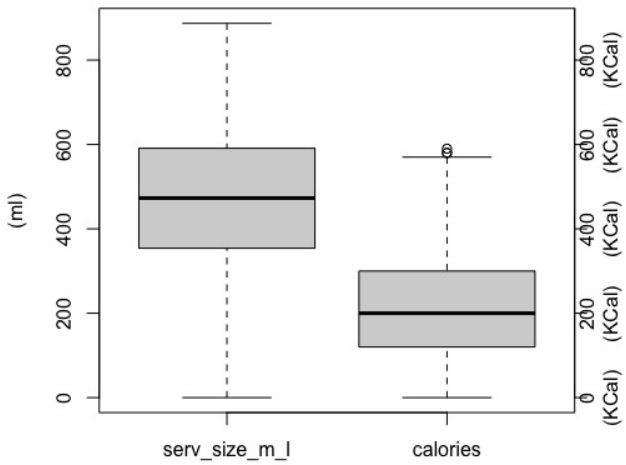
Summary Statistics

	serv_size_m_l	calories	total_fat_g	saturated_fat_g	trans_fat_g	cholesterol_mg
Min.	0	0.0	0.000	0.000	0.0000	0.00
1st Qu.	354	120.0	0.300	0.100	0.0000	0.00
Median	473	200.0	4.000	2.500	0.0000	5.00
Mean	449	214.3	5.531	3.482	0.1094	14.15
3rd Qu.	591	300.0	9.000	6.000	0.2000	25.00
Max.	887	590.0	23.000	16.000	0.5000	70.00

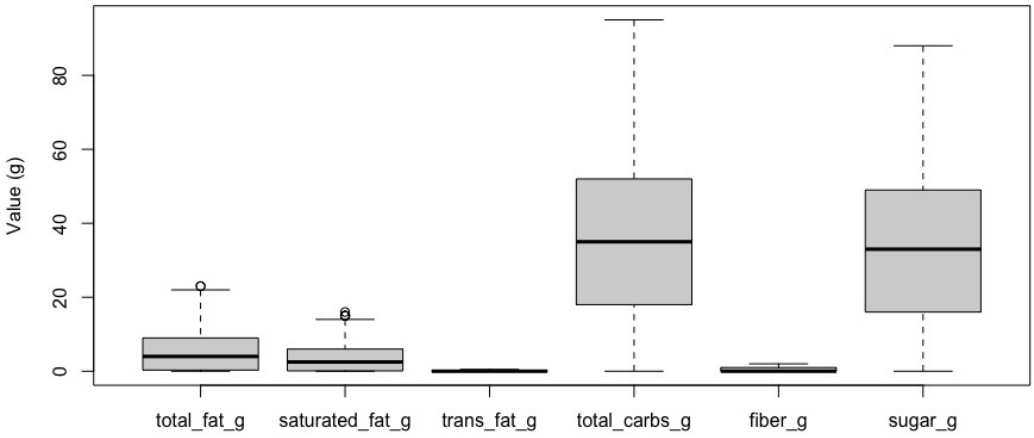
Table 2: Nutritional Information - Part 1

	sodium_mg	total_carbs_g	fiber_g	sugar_g	caffeine_mg
Min.	0.0	0.00	0.0000	0.00	0.00
1st Qu.	70.0	18.00	0.0000	16.00	40.00
Median	135.0	35.00	0.0000	33.00	75.00
Mean	137.2	36.29	0.3635	34.47	85.55
3rd Qu.	200.0	52.00	1.0000	49.00	140.00
Max.	370.0	95.00	2.0000	88.00	300.00

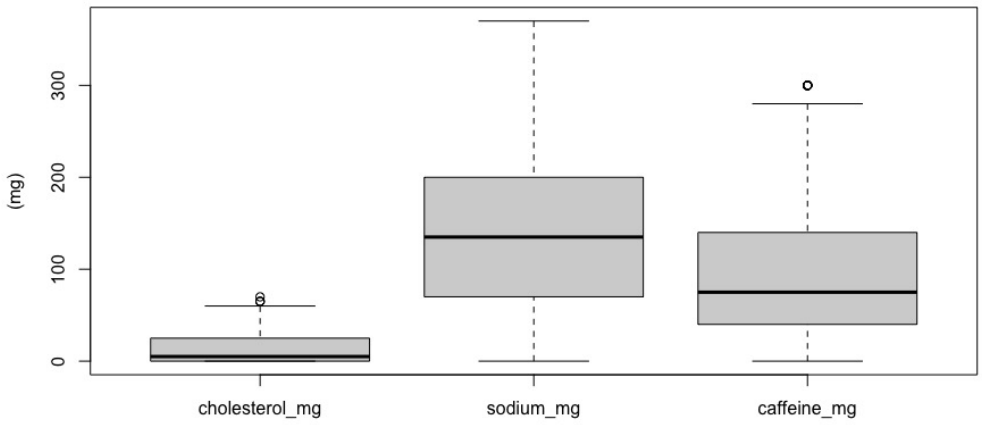
Table 3: Nutritional Information - Part 2



This graph tells us that the average amount of most drinks is around 450ml, the maximum is close to 890ml and the minimum is 0ml. This graph tells us that the average number of total calories for most drinks is around 210KCal, the maximum is close to 600KCal and the minimum is about 0KCal. This does fit with the fact that there are a lot of sweeter drinks, and there are also sugar-free drinks.



This graph shows the energy distribution of drinks for fat,sugar and fiber. You can see that sugar is the majority, and saturated fat is the main fat. Fiber and unsaturated fat are almost non-existent.



You can see in this graph the average amount of cholesterol and sodium and caffeine (mg). The cholesterol content is less, but there are still a few drinks with cholesterol content closer to 90mg. The sodium content is relatively high, up to 370mg. The average caffeine content is 85.5mg, but there are still a few drinks with a high caffeine content, close to 300mg.

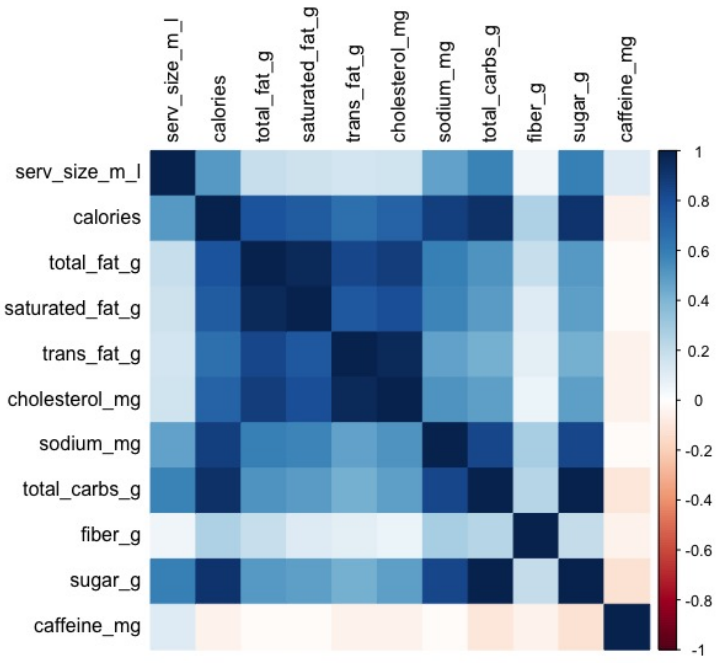


Figure 1: Correlation plot

From the figure, it can be seen that total_carb has a strong correlation with sugar and calories. Caffeine and fiber had little to no association with meditation. Surprisingly,serv_size was not associated with fat, and cholesterol was strongly associated with fat

Analysis Results

These plots depict the nutritional composition of Starbucks drinks, emphasizing sugar as the primary energy source. they show variations in fat, cholesterol, sodium, and caffeine content. Total carbohydrates correlate strongly with sugar and calories, while other components exhibit varying degrees of association. Notably, serving size does not affect fat content, but cholesterol strongly correlates with fat.

2.4 Questions of Interest

Why the methods I chose

1. Which variables are most important in describing differences in the nutritional composition of coffee drinks? (Use PCA ?)

2. Are there some common factors behind the main ingredients of the drinks?(such as high calorie ingredients and low calorie ingredients such as minerals?) (Use FA?)

3. Is there a difference between these two categorizations and how do I understand them? (Maybe we can do it together with the previous question?)

3 Multivariate Analysis Methods

3.1 PCA

I would like to know which variables are most important to account for differences in the nutrient content of coffee beverages.

First standardize the data, then calculate the principal components, check the proportion of variance explained by the principal components. This is the proportion of variance explained by the principal components of the first few PCA analyses:

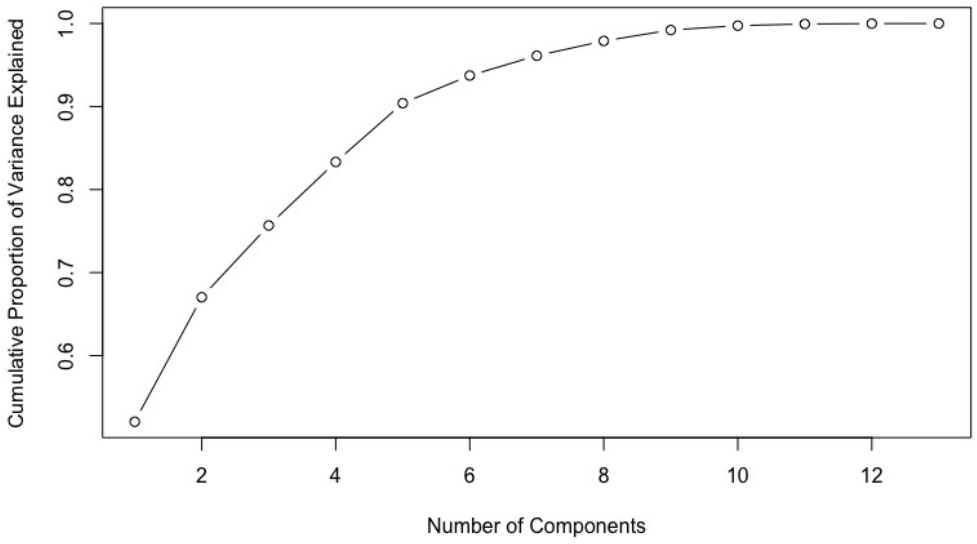


Figure 2: PCA components

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	2.6009	1.3959	1.0602	0.9970	0.9594	0.6577	0.5576
Proportion of Variance	0.5204	0.1499	0.0865	0.0765	0.0708	0.0333	0.0239
Cumulative Proportion	0.5204	0.6703	0.7567	0.8332	0.9040	0.9373	0.9612

Table 4: Principal Component Analysis Results

We can see that by the fourth principal component, the cumulative proportion has reached 83.32%. The retention explains 80% of the total variance, so we choose to retain the first four principal components and study them.

These are the first four principal component loadings:

	PC1	PC2	PC3	PC4
milk	0.1659	0.1411	0.0650	0.6051
whip	0.2748	0.3189	-0.0916	-0.2372
serv_size_m_l	0.1684	-0.4599	0.3135	-0.2042
calories	0.3696	-0.1564	-0.0070	0.0463
total_fat_g	0.3476	0.2403	0.0973	0.0638
saturated_fat_g	0.3343	0.2312	0.0854	0.0503
trans_fat_g	0.2966	0.3092	0.0425	-0.1590
cholesterol_mg	0.3267	0.2894	0.0259	-0.2055
sodium_mg	0.3082	-0.2788	-0.0902	0.0286
total_carbs_g	0.3174	-0.3655	-0.1131	-0.0016
fiber_g	0.1055	-0.0593	0.1915	0.6564
sugar_g	0.3112	-0.3699	-0.1521	-0.0800
caffeine_mg	-0.0320	-0.0135	0.8883	-0.1470

Table 5: Principal Components Analysis Results

Analysis Results

PC1 may represent the Calories or nutrient richness of the drinks.

The positive loadings of PC1 show the positive correlation between the variables calories, total_fat_g, saturated_fat_g, trans_fat_g, cholesterol_mg, sodium_mg, total_carbs_g, and sugar_g with PC1. This indicated that PC1 was mainly concerned with the nutrient density and calorie content of the drinks.

PC2 may represent the high fat content of drinks.

The positive loadings of PC2 show the positive correlation between the variables whip, total_fat_g, saturated_fat_g, trans_fat_g, cholesterol_mg. While the negative correlation between the variables serv_size_m_l, calories, fiber_g, total_carbs_g and PC2. This suggests that PC2 is more concerned with the fat content of the drinks, while showing a negative correlation with the portion size, calories and fiber content of the drinks.

PC3 is mainly concerned with the amount of caffeine in drinks.

The loading pattern of PC3 showed the strongest positive correlation between the variable caffeine_mg and PC3, with a loading value of 0.888. In addition, serv_size_m_l, fiber_g, and total_carbs_g also showed some degree of correlation with PC3, but with smaller negative loading values.

In addition, fiber_g and milk also showed some degree of correlation with PC4, but I think PC4 is mainly represent the protein in drinks.

The loading pattern of PC4 shows a positive correlation between the variables milk, total_fat_g, saturated_fat_g, trans_fat_g, cholesterol_mg, sodium_mg, total_carbs_g, and sugar_g, and PC4, with milk in particular having the largest value. This may mean that PC4 is mainly concerned with drinks. This may imply that PC4 is mainly concerned with the dairy and fat content of drinks, as well as other nutrients related to it.

3.2 FA

In order to discover the underlying structure and factors behind the data, to understand the correlations between the observed variables, and to find common underlying factors, we continued the comparative analysis using factor analysis (FA). We can compare and contrast the two methods Then.

The data are first analyzed and visualized in terms of correlations:

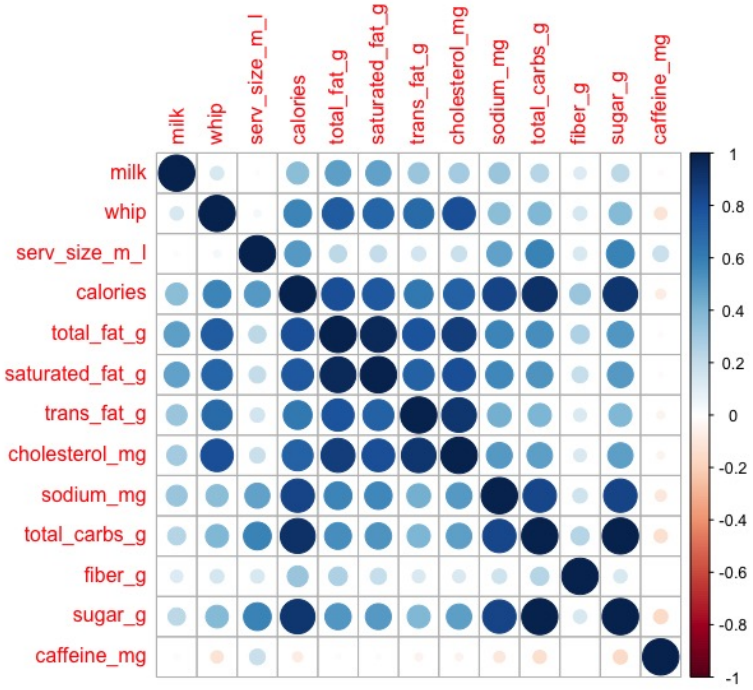


Figure 3: FA correlation

We can see that the overall correlation is quite high. Next, we selected 3 factors for factor analysis and obtained the following summary and factor loadings:

	ML1	ML3	ML2
milk	0.353	0.323	
whip	0.563	0.495	
serv_size_m_l	0.523	-0.258	-0.121
calories	0.989		
total_fat_g	0.768	0.623	0.133
saturated_fat_g	0.736	0.623	
trans_fat_g	0.583	0.538	
cholesterol_mg	0.674	0.568	
sodium_mg	0.839		-0.177
total_carbs_g	0.948	-0.260	-0.172
fiber_g	0.365	-0.224	0.901
sugar_g	0.925	-0.245	-0.285
caffeine_mg	-0.106		

Table 6: Loadings

Factor analysis with Call: `fa(r = df_cor, nfactors = 3, rotate = "none", scores = "Anderson", fm = "ml")`

Test of the hypothesis that 3 factors are sufficient. The degrees of freedom for the model is 42 and the objective function was 2.56.

The root mean square of the residuals (RMSA) is 0.06. The df corrected root mean square of the residuals is 0.08.

Analysis Results

Factor 1 may represent the high energy and carbohydrate content of the drinks.

Factor 1 shows a positive correlation between the variables milk, whip, serv_size_m_l, calories, total_fat_g, saturated_fat_g, trans_fat_g, cholesterol_mg, sodium_mg(0.839), total_carbs_g(0.948), fiber_g, sugar_g, and the Factor 1 This suggests that factor 1 is primarily a reflection of the fact that the variables are not in the same category. This indicated that Factor 1 mainly reflected the energy density and nutrient content of the drinks.

Factor 2 focuses on the portion size and fat content of the drinks, as well as other nutrients associated with it.

The loading pattern for Factor 2 shows a positive correlation between the variables milk(0.323), whip(0.495), total_fat_g(0.623), saturated_fat_g(0.623), trans_fat_g(0.538). This may mean that Factor 2 is primarily concerned with portion size and fat content.

Factor 3 is mainly concerned with fiber and other nutrients in drinks.

The loading pattern for Factor 3 shows a positive correlation between the fiber_g variable and Factor 3 (0.901). This suggests that Factor 3 is mainly concerned with fiber and other nutrients in drinks.

4 Conclusion with deeper interpretation

Using PCA, we categorized the principal components of the beverages into:

1 Calories/sugar(or how many calories the ingredients provide)

2 Fat

3 Caffeine

4 Protein (or milk)

This is a very reasonable principal component analysis result, but we can continue to see how the FA method analyzes the results differently.

The results of the FA were used to investigate the correlation between the ingredients of the drinks.

In ML1, we found that all calorie-providing ingredients were positive, which means that almost all of them were correlated with calories provided (except caffeine). This is consistent with our knowledge that sugar, fat and protein all provide calories to the body.

Saturated fats provide more calories, with sugar obviously contributing more, which can also be explained by the fact that the sugar content of the drink is higher than that of the other two, resulting in a higher and more relevant calorie contribution from sugar.

In ML2 we can see that both saturated and unsaturated fats are highly correlated with milk and whip, and in combination with PCA we can also see this component as highly energy dense

In ML3 we can see that the amount of fiber has a negative correlation with almost all of the other components, but since the correlations are very small, it can be seen that the fiber has little impression on the other nutrients, and in combination with the PCA analysis, it can be interpreted that the fiber component is only due to the additional additives in some beverages, and does not come from the fats, proteins, or caffeine.

4.1 Appendix

```r
# Get the Data
rm(list = ls()) # initialization
# Read in with tidytuesdayR package
# Install from CRAN via: install.packages("tidytuesdayR")
# This loads the readme and all the datasets for the week of interest
# Either ISO-8601 date or year/week works!
library(tidytuesdayR)
library(tidyverse)
library(ggplot2)
library(cowplot)
library(lubridate)
tuesdata <- tidytuesdayR::tt_load('2021-12-21')
tuesdata <- tidytuesdayR::tt_load(2021, week = 52)
starbucks <- tuesdata$starbucks;starbucks
table(starbucks $product_name)
starbucks_2 <- starbucks %>% group_by(product_name)
df <- starbucks
df$trans_fat_g <- as.numeric(df$trans_fat_g)
df$fiber_g <- as.numeric(df$fiber_g)
df$milk <- as.character(df$milk)
df$whip <- as.character(df$whip)
summary(df)
is_outlier <- function(x) {
  Q1 <- quantile(x, 0.25)
  Q3 <- quantile(x, 0.75)
  IQR <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR
  upper_bound <- Q3 + 1.5 * IQR
  x < lower_bound | x > upper_bound
}
clean_data <- function(df) {
  for (col in colnames(df)) {
    if (is.numeric(df[[col]])) {
      df <- df[!is.na(df[[col]]), ]
      df <- df[!is_outlier(df[[col]]), ]
    }
  }
  return(df)
}
df_clean <- clean_data(df);df_clean
## visualize
summary(df_clean[,c(5:15)])

boxplot(df_clean[,c(7,8,9,12,13,14)],ylab = "(g)")
boxplot(df_clean[,c(10,11,15)],ylab = "(mg)")
boxplot(df_clean[,c(5,6)],ylab = "(ml)")
par(new=TRUE)
```

```r
axis(side = 4, at = pretty(range(df_clean[,c(5,6)])), labels = sprintf(paste0("%s\n(KCal)"), pretty(range(df_clean[,c(5,6)]))))
mtext("Additional Value", side = 4, line = 3)

corrplot::corrplot(cor(df_clean_numeric), method = "color", col.lab = "black",tl.col = "black")
heatmap(correlation_matrix, col = colorRampPalette(c("blue", "white", "red"))(100))


## PCA

colnames(df)

df_pca <- df[,3:15]
df_pca <- na.omit(df_pca)
sapply(df_pca, class)
df_pca$milk <- as.numeric(as.character(df_pca$milk))
df_pca$whip <- as.numeric(as.character(df_pca$whip))

# scale
df_scaled <- scale(df_pca)
pca_result <- prcomp(df_scaled)
summary(pca_result)
pca_result$rotation[,1:4]
plot(cumsum(pca_result$sdev^2) / sum(pca_result$sdev^2), xlab = "Number of Components", ylab = "Cumulative Proportion of Variance Explained", type = "b")

cumulative_variance <- cumsum(pca_result$sdev^2) / sum(pca_result$sdev^2)
n_components <- which.max(cumulative_variance >= 0.8)


## MDS
df_scaled

## FA
library(psych)
library(corrplot)

df_cor <- cor(df_scaled)
corrplot(df_cor)
fa_result <- factanal(df_cor, factors = 4)
fa_result <- fa(df_cor, nfactors = 3, rotate = "none", fm = "ml",scores = "Anderson")

summary(fa_result)
print(fa_result$loadings)
fa.diagram(fa_result)
```


WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON