Final Project: Fast Food Nutritional Analysis

Sainath Sunkara

2025-06-30

Introduction

Fast food is one of the biggest thing people in the USA depend on for quick food on the go. This is because life has become more and more fast paced leaving people to attain food quickly at fast food restaurants. As good as this food may taste, it can heavily influence your health in the long term. The trend of fast food in this country can also relate to the growing obesity rates! I plan do use the 3 data sets I have collected to answer the research questions given below!

Research Questions and Background Information

1. **How do the nutritional contents vary among the top 6 fast-food chains?**
2. **Is there a relationship between the number of restaurant locations and the average calories served by these chains?**
3. **Is there a relationship between the sales volume of fast-food chains and their average calorie content?**

I plan on answering these three questions by loading in all the data, joining the data, and creating graphics and tables to further help. The first research question will show us how the different fast food chain differ in terms of average nutritional value. We will use a box plot and show the values for different nutritional values. The second question tries to find out the relationship between the number of locations a fast food join has and the average calories among their menu. We will try and use a point plot to visualize this data. The third research question tries to see if there is some sort of relationship between sales volume in millions and their calorie content. This will help us to see if there is some sort of correlation between the average calorie content and the amount of sales generated

Data Summary

For this project, I am using three main data sets-
The first data set is a "Fast Food Nutrition" data set which gives data on the particular fast food joint along with the nutritional values of items. We will work on this data set to clean and find the average calorie value across their menu. This data set was found on kaggle and i will primarily use the nutritional values given in this data set.

For the second data set, I scraped data from the Wikipedia page "https://en.wikipedia.org/wik/List_of_the_largest_fast_food_restaurant_chains". This data set gives information on the country, fast food joint, and number of locations. We will need to clean this table my normalizing the fast food join names to later join with other tables

The third data set I used is another data set from kaggle which gives us information on the top 50 fast food joint in USA. From this table, we will use the "Systemwide Sales (Millions - U.S Dollars)" to answer our third question

Hence to answer our questions we perform data cleaning, merging, and visualization.

- **Primary Dataset:** Fast Food Nutrition Menu (Kaggle)
  Attributes: Company, Calories, Fat, Sodium, Sugars, etc.
- **Dataset 1:** Wikipedia HTML Scrape
  Attributes: Fast Food Chain Name, Number of Locations
- **Dataset 2:** Top 50 Fast-Food Chains in the USA
  Attributes: Company, Systemwide Sales, Total Units, etc

We focus on calories, sodium, sugars, number of locations, and sales.

Exploratory Data Analysis

Table: Comparison of Fast Food Chains

Here we show a visual Figure 1 on the nutritional differences among different fast food joints. As it can be seen, some clearly dominate other, while some are pretty consistent. Burger king shows up as the highest value for all our unhealthy attributes showing us that the average unhealthy nutritional value is not that good. The heaviest differences can be seen in the trans fat value!we further see the transfat values in Table 1 to further inpec the data

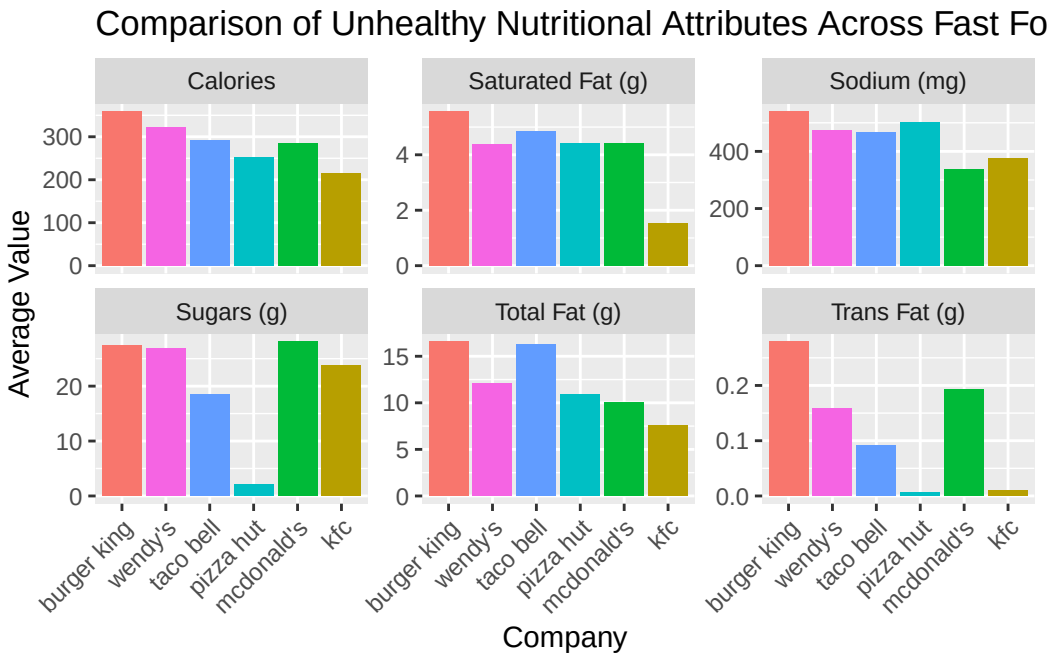
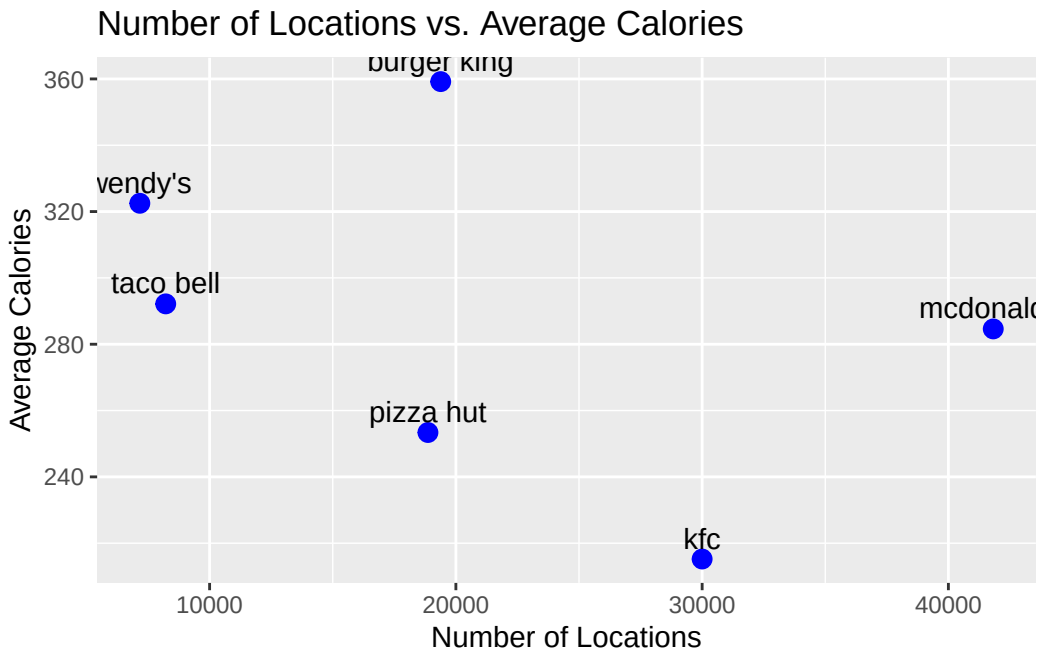


Figure 1

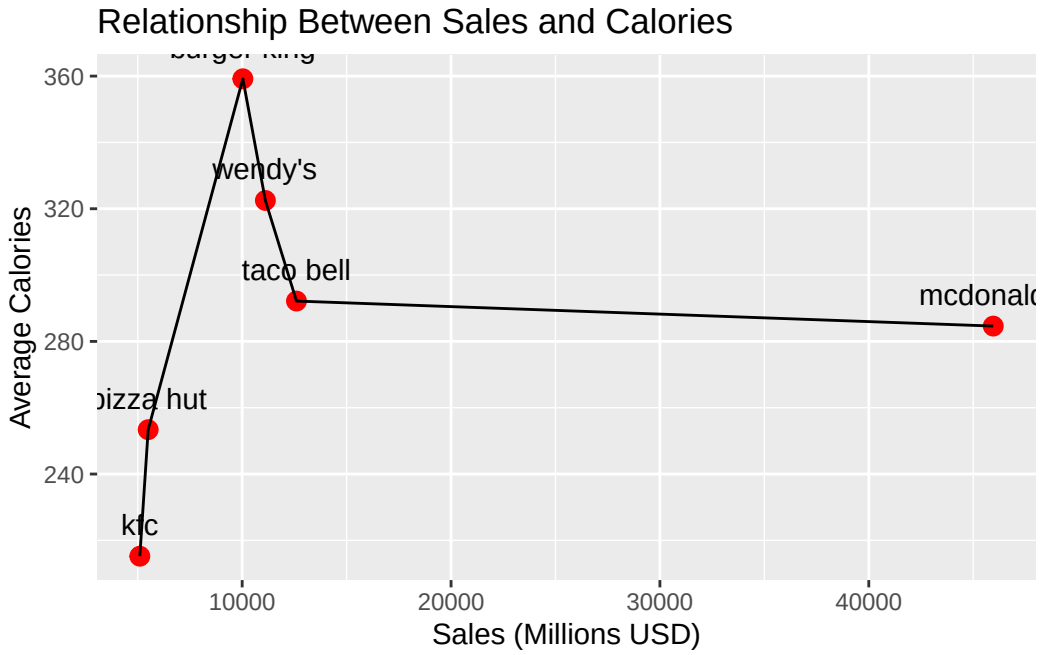
Table 1: Top 3 Fast Food Chains by Average Trans Fat Content

Company	Trans Fat (g)	Calories	Sales
burger king	0.2793296	359.1899	10033
wendy's	0.1590909	322.5000	11111
taco bell	0.0912698	292.1667	12600
mcdonald's	0.1935976	284.6189	45960
pizza hut	0.0067568	253.3784	5500

We use the visual below to see if there is any correlation on the number of locations and the average calories across the menu. The graph convinces us that the more the number of locations, the lesser the value of average calories across the menu. Mcdonalds and burger king are somewhat of outliers in this graphing.



To answer our third research question, we try to see if there is some sort of a relationship between the number of sales and the average calories across the fast food joints menu. Through the graphic visual, we understand that there isn't and particular correlation.



We use this table to better understand the information about different fast food joints nutritional value with their sales in millions.

Table 2: Comparison of Fast Food Chains: Nutrition, Locations, and Sales(mil)

Company	Calories	Sodium (mg)	Sugars (g)	Number.of.locations	Sales
burger king	359.1899	540.1117	27.324022	19384	10033
wendy's	322.5000	475.8117	26.941558	7166	11111
taco bell	292.1667	465.9611	18.522222	8218	12600
mcdonald's	284.6189	338.4604	28.103659	41822	45960
pizza hut	253.3784	501.7568	2.054054	18866	5500
kfc	215.2294	377.1101	23.756881	30000	5100

Conclusion

In conclusion, we see how nutritional facts differ among different fast food joints, the correlation between the average number of calories and location, and the relationship between sales and calories. From our analysis we found out that on the average, the more the number of locations, the lesser their menu average calories is. In the last research question we concluded that there is not specific correlation but we are able to visualize the data effectively.

```
#load the libraries first
library(dplyr)
library(tidyr)
library(stringr)
library(ggplot2)
library(tools)
library(readr)
library(rvest)
library(knitr)


# first we load all the required libraries
library(dplyr)
library(tidyr)
library(stringr)
library(ggplot2)
library(tools)
library(readr)
library(rvest)
library(knitr)
```

```r
# Import both the data sets we get from kaggle
df <- read_csv("~/Downloads/FastFoodNutritionMenuV2.csv")
sales_df <- read_csv("~/Downloads/Top 50 Fast-Food Chains in USA.csv")

#this is to scrape the table from wikipedia and import it
url <- "https://en.wikipedia.org/wiki/List_of_the_largest_fast_food_restaurant_chains"

page <- read_html(url)
tables <- page %>% html_table(fill = TRUE)
#Assign variabe to the table we need
fast_food_table <- tables[[4]]
#This is from a google codebase i found as a refernece, it helps in cleaning the column names
colnames(fast_food_table) <-
  make.names(colnames(fast_food_table), unique = TRUE)

# This is to clean the values of the name and make it consistent with the other data sets. Th
location_df_clean <- fast_food_table %>%
  mutate(Name = str_replace_all(Name, "’", "'"),
         Name = str_replace_all(Name, "‘", "'"),
         Name = str_squish(Name),
         Name = tolower(Name),
         Name = case_when(
           Name == "kfc" ~ "kfc",
           Name == "mcdonald's" ~ "mcdonald's",
           TRUE ~ Name
         )) %>%
  mutate(Number.of.locations = str_extract(Number.of.locations, "\\d{1,3}(,\\d{3})*|\\d+"),
         Number.of.locations = as.numeric(gsub(",", "", Number.of.locations)))

# This is done to clean and standardize the column names given in the first data set.

colnames(df) <- gsub("\n", " ", colnames(df))
colnames(df) <- gsub("\\s+", " ", colnames(df))
colnames(df) <- trimws(colnames(df))
# We work on the company name similar to how we did with the wikipedia data set to mutate the
df <- df %>%
  mutate(Company = str_replace_all(Company, "’", "'"),
         Company = str_replace_all(Company, "‘", "'"),
         Company = str_squish(Company),
         Company = tolower(Company))

# We gather all the unhealthy nutritional values.
```

```r
unhealthy_columns <- c('Calories', 'Total Fat (g)', 'Saturated Fat (g)', 'Trans Fat (g)', 'So

#This is a process obtained from google where it mutated all the values in the unhealthy colu
df <- df %>%
  mutate(across(all_of(unhealthy_columns), ~ parse_number(.)))

# this helps us calulate the averages or the mean across all the values
summary_df <- df %>%
  group_by(Company) %>%
  summarise(across(all_of(unhealthy_columns), ~ mean(., na.rm = TRUE))) %>%
  arrange(desc(Calories))

#we use the method left join to keep all the elements in the left data set and add the ones t
merged_df <- summary_df %>%
  left_join(location_df_clean, by = c("Company" = "Name")) %>%
  filter(!is.na(Number.of.locations))

# this is the second data frame we got from kaggle, similar to how we cleaned the name is the
sales_clean <- sales_df %>%
  mutate(Company = str_squish(`Fast-Food Chains`),
         Company = str_replace_all(Company, "’", "'"),
         Company = str_replace_all(Company, "‘", "'"),
         Company = tolower(Company),
         Company = case_when(
           Company == "kfc" ~ "kfc",
           Company == "mcdonald's" ~ "mcdonald's",
           TRUE ~ Company
         )) %>%
  filter(Company %in% merged_df$Company) %>%
  select(Company, Sales = `U.S. Systemwide Sales (Millions - U.S Dollars)`)


#Inititally used full join but due to a lot of NA values, used left join to keep all the elem
# Final merge
final_combined <- merged_df %>%
  left_join(sales_clean, by = "Company")

#this is to make the data for the initial resercrch questions.
summary_long <- summary_df %>%
  pivot_longer(cols = -Company, names_to = "Nutrient", values_to = "Value")
ggplot(summary_long, aes(x = reorder(Company, -Value), y = Value, fill = Company)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
```

```r
  facet_wrap(~ Nutrient, scales = "free_y") +
  labs(title = "Comparison of Unhealthy Nutritional Attributes Across Fast Food Chains",
       x = "Company", y = "Average Value") +
  # this is done as the text overlays if not angles, this is a method i got online from the c
  theme(axis.text.x = element_text(angle = 45, hjust = 1))


# we create a dislay of info for the 5 joints to show their exact trans fat values as we find
top_transfat <- final_combined %>%
  arrange(desc("Trans Fat (g)")) %>%
  select(Company, "Trans Fat (g)", Calories, Sales) %>%
  head(5)

# This neately displays the table
knitr::kable(top_transfat, caption = "Top 3 Fast Food Chains by Average Trans Fat Content")

ggplot(final_combined, aes(x = Number.of.locations, y = Calories, label = Company)) +
  geom_point(size = 3, color = "blue") +
  # this positions the text so the it isnt over the points and can be seen properly
  geom_text(vjust = -0.5) +
  labs(title = "Number of Locations vs. Average Calories",
       x = "Number of Locations", y = "Average Calories")



ggplot(final_combined, aes(x = Sales, y = Calories, label = Company)) +
  geom_point(size = 3, color = "red") +
  #again helps us position the text
  geom_text(vjust = -1) +
  geom_line() +
  labs(title = "Relationship Between Sales and Calories",
       x = "Sales (Millions USD)", y = "Average Calories")

# this table shows altered values and doesnt show raw data values from the data sets as all t
final_combined %>%
  select(Company, Calories, `Sodium (mg)`, `Sugars (g)`, Number.of.locations, Sales) %>%
  kable(caption = "Comparison of Fast Food Chains: Nutrition, Locations, and Sales(mil)")

# first we load all the required libraries
library(dplyr)
library(tidyr)
library(stringr)
```

```r
library(ggplot2)
library(tools)
library(readr)
library(rvest)
library(knitr)

# Import both the data sets we get from kaggle
df <- read_csv("~/Downloads/FastFoodNutritionMenuV2.csv")
sales_df <- read_csv("~/Downloads/Top 50 Fast-Food Chains in USA.csv")

#this is to scrape the table from wikipedia and import it
url <- "https://en.wikipedia.org/wiki/List_of_the_largest_fast_food_restaurant_chains"

page <- read_html(url)
tables <- page %>% html_table(fill = TRUE)
#Assign variabe to the table we need
fast_food_table <- tables[[4]]
#This is from a google codebase i found as a refernece, it helps in cleaning the column names
colnames(fast_food_table) <-
  make.names(colnames(fast_food_table), unique = TRUE)

# This is to clean the values of the name and make it consistent with the other data sets. Th
location_df_clean <- fast_food_table %>%
  mutate(Name = str_replace_all(Name, "’", "'"),
         Name = str_replace_all(Name, "‘", "'"),
         Name = str_squish(Name),
         Name = tolower(Name),
         Name = case_when(
           Name == "kfc" ~ "kfc",
           Name == "mcdonald's" ~ "mcdonald's",
           TRUE ~ Name
         )) %>%
  mutate(Number.of.locations = str_extract(Number.of.locations, "\\d{1,3}(,\\d{3})*|\\d+"),
         Number.of.locations = as.numeric(gsub(",", "", Number.of.locations)))

# This is done to clean and standardize the column names given in the first data set.

colnames(df) <- gsub("\n", " ", colnames(df))
colnames(df) <- gsub("\\s+", " ", colnames(df))
colnames(df) <- trimws(colnames(df))
# We work on the company name similar to how we did with the wikipedia data set to mutate the
df <- df %>%
```

```r
  mutate(Company = str_replace_all(Company, "’", "'"),
         Company = str_replace_all(Company, "‘", "'"),
         Company = str_squish(Company),
         Company = tolower(Company))

# We gather all the unhealthy nutritional values.
unhealthy_columns <- c('Calories', 'Total Fat (g)', 'Saturated Fat (g)', 'Trans Fat (g)', 'So

#This is a process obtained from google where it mutated all the values in the unhealthy colu
df <- df %>%
  mutate(across(all_of(unhealthy_columns), ~ parse_number(.)))

# this helps us calulate the averages or the mean across all the values
summary_df <- df %>%
  group_by(Company) %>%
  summarise(across(all_of(unhealthy_columns), ~ mean(., na.rm = TRUE))) %>%
  arrange(desc(Calories))

#we use the method left join to keep all the elements in the left data set and add the ones t
merged_df <- summary_df %>%
  left_join(location_df_clean, by = c("Company" = "Name")) %>%
  filter(!is.na(Number.of.locations))

# this is the second data frame we got from kaggle, similar to how we cleaned the name is the
sales_clean <- sales_df %>%
  mutate(Company = str_squish(`Fast-Food Chains`),
         Company = str_replace_all(Company, "’", "'"),
         Company = str_replace_all(Company, "‘", "'"),
         Company = tolower(Company),
         Company = case_when(
           Company == "kfc" ~ "kfc",
           Company == "mcdonald's" ~ "mcdonald's",
           TRUE ~ Company
         )) %>%
  filter(Company %in% merged_df$Company) %>%
  select(Company, Sales = `U.S. Systemwide Sales (Millions - U.S Dollars)`)


#Inititally used full join but due to a lot of NA values, used left join to keep all the elem
# Final merge
final_combined <- merged_df %>%
  left_join(sales_clean, by = "Company")
```

```r
#this is to make the data for the initial resercrch questions.
summary_long <- summary_df %>%
  pivot_longer(cols = -Company, names_to = "Nutrient", values_to = "Value")




ggplot(summary_long, aes(x = reorder(Company, -Value), y = Value, fill = Company)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  facet_wrap(~ Nutrient, scales = "free_y") +
  labs(title = "Comparison of Unhealthy Nutritional Attributes Across Fast Food Chains",
       x = "Company", y = "Average Value") +
  # this is done as the text overlays if not angles, this is a method i got online from the c
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# we create a dislay of info for the 5 joints to show their exact trans fat values as we find
top_transfat <- final_combined %>%
  arrange(desc("Trans Fat (g)")) %>%
  select(Company, "Trans Fat (g)", Calories, Sales) %>%
  head(5)

# This neately displays the table
knitr::kable(top_transfat, caption = "Top 3 Fast Food Chains by Average Trans Fat Content")

ggplot(final_combined, aes(x = Number.of.locations, y = Calories, label = Company)) +
  geom_point(size = 3, color = "blue") +
  # this positions the text so the it isnt over the points and can be seen properly
  geom_text(vjust = -0.5) +
  labs(title = "Number of Locations vs. Average Calories",
       x = "Number of Locations", y = "Average Calories")

ggplot(final_combined, aes(x = Sales, y = Calories, label = Company)) +
  geom_point(size = 3, color = "red") +
  #again helps us position the text
  geom_text(vjust = -1) +
  geom_line() +
  labs(title = "Relationship Between Sales and Calories",
       x = "Sales (Millions USD)", y = "Average Calories")
```

```r
final_combined %>%
  select(Company, Calories, `Sodium (mg)`, `Sugars (g)`, Number.of.locations, Sales) %>%
  kable(caption = "Comparison of Fast Food Chains: Nutrition, Locations, and Sales(mil)")
```

References

- Wikipedia. List of the largest fast food restaurant chains
- kaggle data set. nutrition menu data set
- Second kaggle data set Top 50 fast food chains
- Code documentation referred to link
- code documentation for ggplot link