Group Project Report

Fast Food Nutrition Info graphic

Subject: Introduction to Data Visusalization

Course Teacher: Dr. Fang XIE

Group 6

Group Members

Yanjing Zhan 2030026196

Wenhao Guo 2030026043

Ruoxin Wang 2030026170

Jiapeng Yu 2030026192

Contents

1. Introduction

Nowadays, "fast food" as a synonym for convenience, speed and simplicity fits perfectly into the fast-paced, minimalist trend of today's world. Fast food restaurants have become part of our modern lifestyle, offering us a wide variety of delicious options to cater to our more and more pressed for time. From iconic brands like McDonald's and Burger King to Taco Bell and Pizza Hut, these fast food restaurants have become an integral part of our food culture.

Although fast food has the advantage of satisfying the people's need for convenience and providing instant gratification, the impact of fast food on human health is complex. As consumers, we are often attracted to the enticing flavors and aromas of these fast foods and unknowingly consume excessive amounts of food due to the relatively self-contained packaging. Therefore, it is important for us to recognize the probable consequences of our dietary choices. The quick methods used in the preparation of fast food (frying, baking), the use of large quantities of additives, the lack of fresh vegetables and fruits, and the high levels of fat, sugar and salt have raised worries about the health effects.

The nutritional values provided by fast food restaurants are not limited to simple calorie counts, but provide important information about the micronutrient content of these foods. This information includes, but is not limited to, attributes such as calories, calories from fat, total fat, saturated fat, trans fat, cholesterol, sodium, carbohydrates, fiber, sugar, and protein of the food. The detailed descriptions of these nutrients allow us to have a more complete understanding of the nutritional value of foods allowing us to make more wise decisions about our eating habits. This comprehensive knowledge will be more helpful to our health, especially for those who are losing weight or maintaining a healthy diet.

The Fast Food Nutrition Dataset includes the nutritional information of six leading fast food companies - McDonald's, Burger King, Wendy's, KFC, Taco Bell and Pizza Hut. By analyzing the data, we are able to get a better understanding of the connotations of fast food consumption and its potential impact on public health. Our goal is to provide the general public with the necessary knowledge to help them make wiser choices for more sensible, nutritious, and healthier foods to maintain good health in the fast pace of life.

2. Data Analysis

Fast Food Nutrition Dataset is a live updated dataset, last updated on 10/17/2023, which provides nutritional information for 1147 kinds of food items, including burgers, sandwiches, chicken products, salads, and side dishes such as French fries. The provided nutritional information include 14 types:

(1) Company: The name of food company.
(2) Item: The name of the food item.
(3) Calories: The total number of calories in the food item.
(4) Calories from Fat: The number of calories that come from fat in the food item.
(5) Total Fat (g): The total amount of fat in grams present in the food item.
(6) Saturated Fat (g): The amount of saturated fat in grams present in the food item.
(7) Trans Fat (g): The amount of trans fat in grams present in the food item.
(8) Cholesterol (mg): The amount of cholesterol in milligrams present in the food item.
(9) Sodium (mg): The amount of sodium in milligrams present in the food item.
(10) Carbs (g): The total amount of carbohydrates in grams present in the food item.
(11) Fiber (g): The amount of dietary fiber in grams present in the food item.

(12) Sugars (g): The amount of sugar in grams present in the food item.

(13) Protein (g): The amount of protein in grams present in the food item.

(14) Weight Watchers Pnts: The Weight Watchers points value assigned to the food item.

3. Data Processing

After careful inspection of the dataset, a noticeable problem was identified - null values were present in different feature columns (Table 1). This phenomenon highlighted the imperative need for a remedial strategy to address these missing data issues. While transformation or estimation techniques are a viable solution, the nature of the missing values prompted a more careful analysis.

Table 1 Count for NULL Value

Column Name	Num of NULL
Company	0
Item	0
Calories	1
Calories from Fat	506
Total Fat (g)	57
Saturated Fat (g)	57
Trans Fat (g)	57
Cholesterol (mg)	1
Sodium (mg)	1
Carbs (g)	57
Fiber (g)	57
Sugars (g)	1
Protein (g)	57
Weight Watchers Pnts	261

Particularly striking are Weight Watchers Pnts and Calories from Fat, each of which have a large number of null entries, 261 and 506, respectively.

The large number of null values for Weight Watchers Pnts and Calories from Fat challenges the feasibility of traditional estimation techniques. Inserting median or mean values may lead to interpretation bias and affect the veracity of the data description. In fact, it could also raise concerns about replacing real data with falsified entries, thereby distorting the true picture reflected by the data set.a

In summation, removing Weight Watchers Pnts and Calories from Fat and is a reasonable solution. This approach guarantees the integrity of our analysis and ensures that the insights derived are still based on real data representations. By removing this column, we eliminate the potential bias that can be introduced by estimation and make the analysis more transparent and credible.

Table 2 Special Values Collection

Column Name	Num of Special Value	Value
Calories	1	‘ ‘
Total Fat (g)	2	‘ ‘ ‘’
Saturated Fat (g)	2	‘ ‘ ‘’
Trans Fat (g)	2	‘ ‘ ‘’
Cholesterol (mg)	2	‘ ‘ ‘<5’
Sodium (mg)	2	‘ ‘ ‘<1’
Carbs (g)	3	‘ ‘ ‘<1’ ‘’
Fiber (g)	3	‘ ‘ ‘<1’ ‘’
Sugars (g)	2	‘ ‘ ‘<1’
Protein (g)	2	‘ ‘ ‘’

When analyzing the Cholesterol (mg) feature column, we encountered problems with more than just NaN values. The dataset (Table 2) also had spaces (',' ') and "less than

5 (<5)". These two simple-looking symbols actually hide a significant problem - they increase the complexity of our analysis.

It is also worth pointing out that the space and less-than symbols are not ordinary characters, but camouflaged missing values. What they represent is an information gap that creates a greater challenge in dealing with incomplete data. This subtle issue adds another layer to our quest for accurate and reliable data.

In our exploration of the intricacies of data preprocessing, we encounter a wide variety of exceptions, each requiring a unique solution. Thus, we chose to replace NaN values, spaces, and values less than 5 by means, and values less than 1 by zeros. This treatment aims to maintain the integrity and reliability of the data while providing an accurate basis for analysis.

Table 3 Changed Column Names

Original Column Name	New Column Name
Company	Company
Item	Item
Calories	Calories
Total Fat (g)	TotalFat
Saturated Fat (g)	SaturatedFat
Trans Fat (g)	TramsFat
Cholesterol (mg)	Cholesterol
Sodium (mg)	Sodium
Carbs (g)	Carbs
Fiber (g)	Fiber
Sugars (g)	Sugars
Protein (g)	Protein

After completing the data cleaning, we faced another challenge. The column names in the original dataset contained spaces and parentheses, which did not conform to the

naming conventions of the R language. If reading directly from the file and using the default naming convention, although it was compliant, it would lead to the introduction of a large number of "_" in the column names, which would affect the readability of the dataset.

To address this issue, we have renamed all column names. The updated column names have been listed in Table 3. These new column names follow the R language naming convention and avoid the use of "_". In this way, we have ensured the accuracy and reliability of the data, while maintaining the readability and aesthetics of the dataset. Finally, the dataset was divided into six smaller datasets (Table 4) based on company names for ease of subsequent use.

Table 4 Divided dataset

Company Name	Num of Item
McDonald’s	328
KFC	218
Burger King	190
Pizza Hut	74
Taco Bell	183
Wendy’s	154

4. Tasks Abstraction

4.1. Nutrient Proportion

We drew a pie chart about the nutrient ratio, and you can intuitively see the six companies, the proportion of nutrients in each company, and you can intuitively feel the similarities and differences between them.

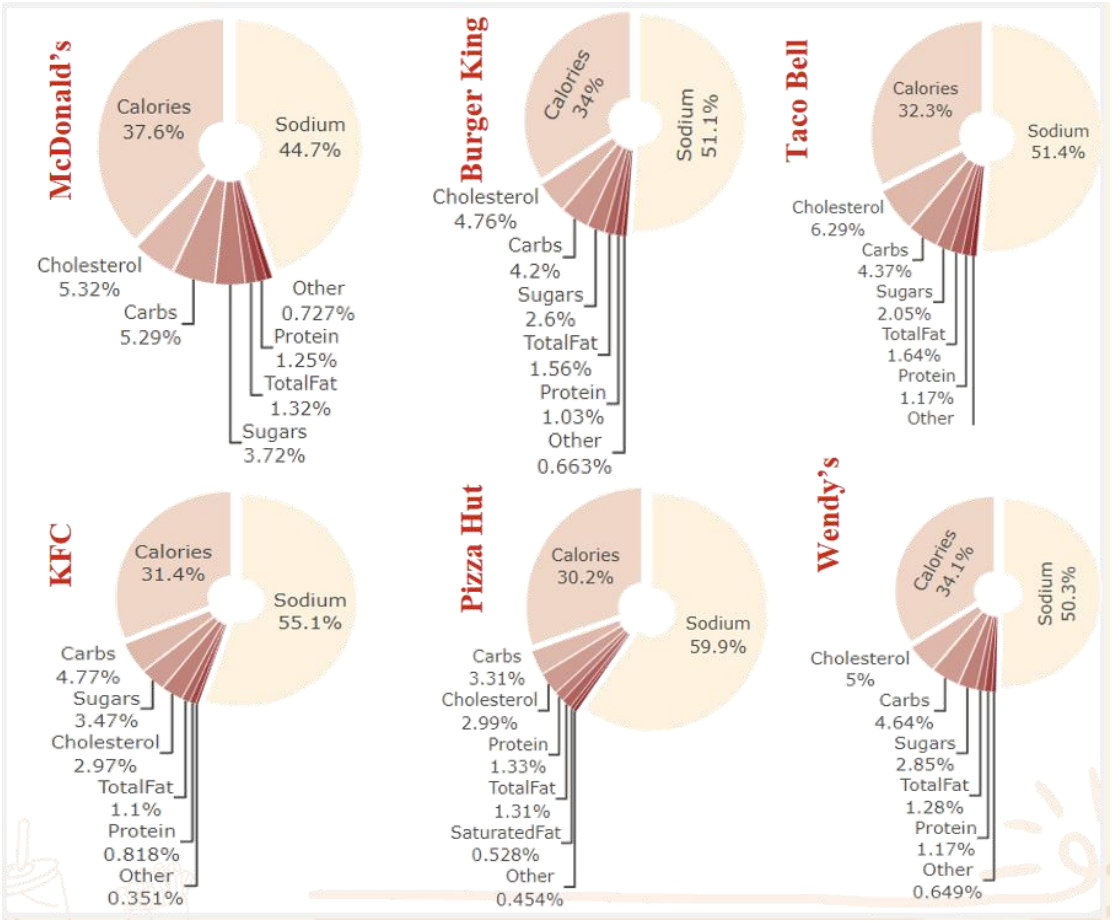
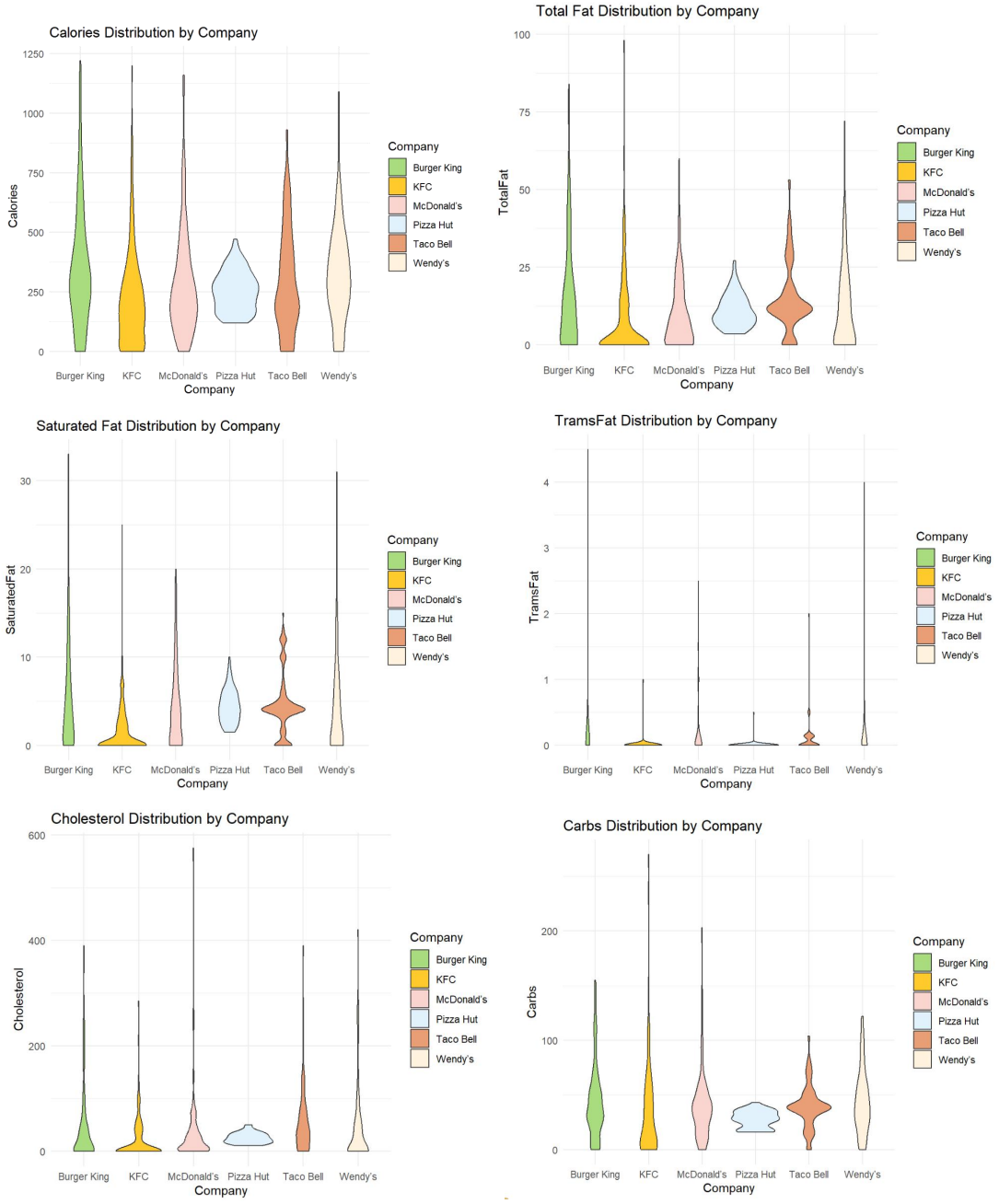


Figure. 1 Nutrition Proportion -- Pie Chart

As shown in Figure. 1, the nutritional profile of these fast food meals is characterize by a high ratio of sodium to calories, with a large portion of the calories coming from fat. Pizza Hut had the highest sodium content at 52.5 per cent. If these meals are consumed on a regular basis, the sodium content can affect health. These meals are low in protein, no more than 1.17 per cent, and may not be ideal for a balanced diet. Consumption of these meals should be moderated, especially for people with dietary

restrictions or health problems (e.g. high blood pressure) related to high sodium intake. Balancing these meals with other nutrient-dense, low-sodium, high-protein foods can be beneficial in maintaining a healthy, balanced diet.

4.2. Ingredient Distribution of Each Company



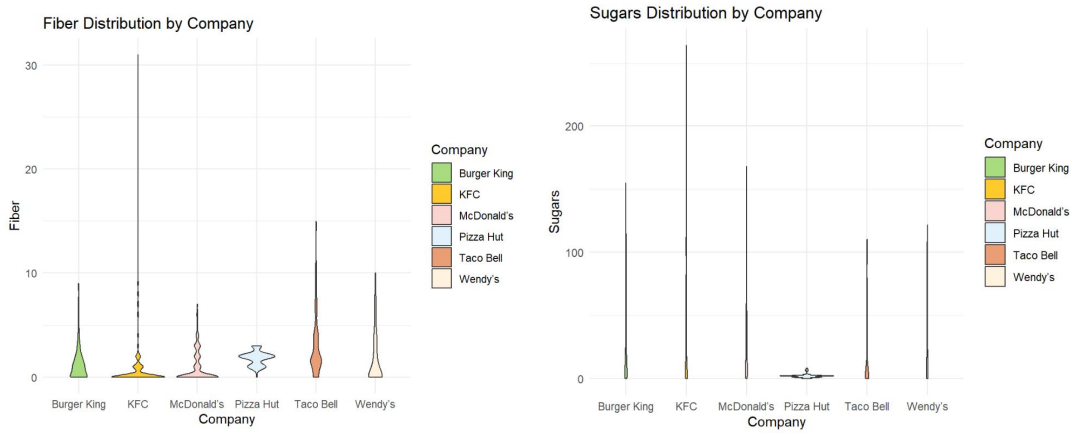


Figure. 2 Ingredient Distribution of Each Company -- Violin Plot

This set of violin charts presents the distribution of nutrients for fast food companies such as Burger King, KFC, McDonald's, Pizza Hut, Taco Bell and Wendy's. Using these charts, it seems that we are able to offer some insights into the distribution of nutrients in these chains. For calories, there are significant differences in content from company to company, with some companies having a wider distribution, showing a wide range of calorie content on their menus. For fat calories, some chains showed a more centralised distribution, suggesting that they are consistent in terms of fat calories, while others showed a wider distribution, suggesting a greater variety of fat content on their menus. With regard to total fat, the chart mainly shows the distribution of total fat content across companies, some of which have a narrower distribution, suggesting that their products are relatively consistent in terms of fat content. With regard to saturated fat, a specific type of fat associated with cardiovascular health risks, there are also differences in the distribution of saturated fat across companies, some of which have a wider range of saturated fat content.

4.3. Correlation among Ingredients

-- Heatmap and Density Plot

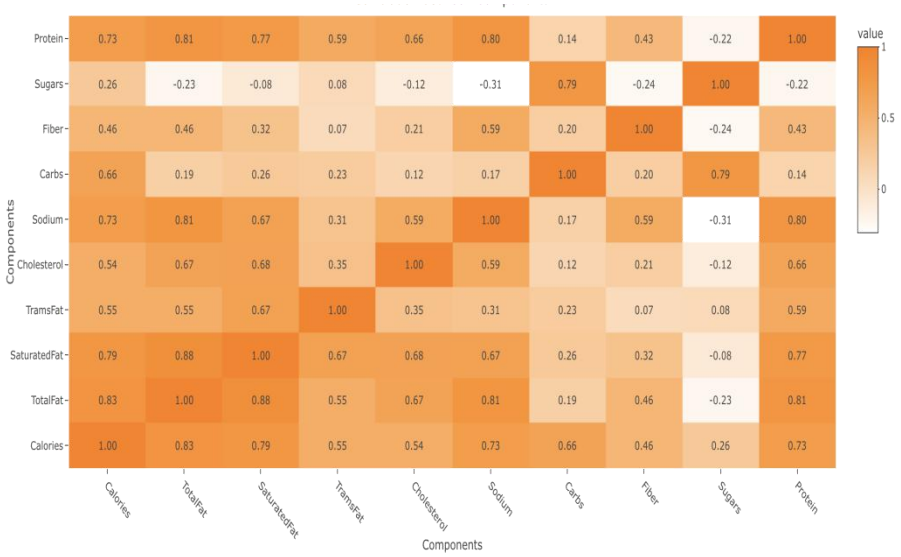


Figure. 3 Heat Map

The heatmap, which is a graphical representation of data where individual values contained in a matrix are represented as colors. This particular heatmap displays the correlation between different nutritional components in food items.

From the heatmap, we can make several observations:

- There are positive correlations among different types of fats (Total Fat, Saturated Fat, and Trans Fat) and Calories from Fat, which is expected since they all contribute to the total caloric content from fats.
- Cholesterol might show a positive correlation with types of fats, which makes sense as many high-fat foods are also high in cholesterol.
- Sodium does not seem to have a strong correlation with fats or cholesterol, suggesting that foods high in sodium are not necessarily high in fats or cholesterol.
- Carbohydrates, Sugars, and Fiber may have varying degrees of correlation with each other, which would be expected as they are all types of

carbohydrates.

- Protein seems to be less correlated with other components, which may indicate that protein content is independent of the fat, carb, and sugar content of the food items analyzed.

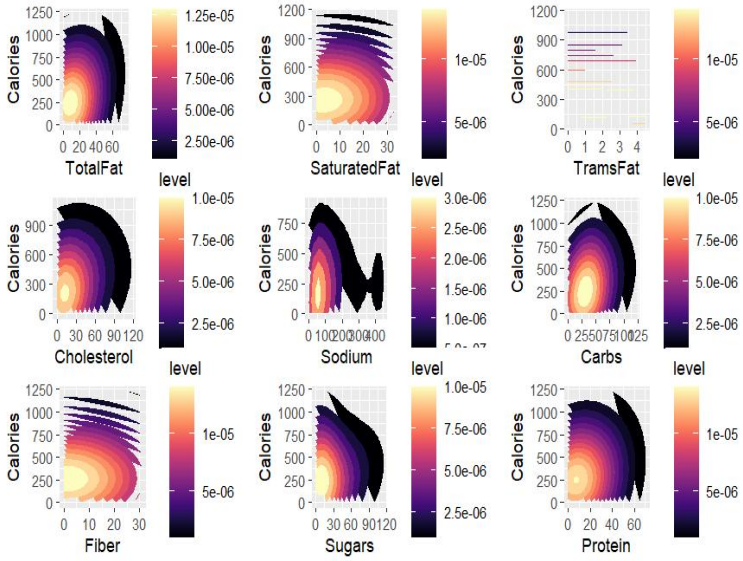


Figure. 4 Density Plot

These particular two-dimensional Kernel Density Estimation (KDE) plots are utilized in data visualization to generate a smooth curve based on a set of data. In this context, Calories are represented on the vertical axis, while various nutritional components are depicted on the horizontal axis. The color bars on the right side of each plot indicate the density of data points at different values, with darker colors signifying a higher concentration of data.

These plots offer insights into the distribution of different nutritional components in relation to calorie content. The Total Fat vs. Calories plot, for example, exhibits a dense region at lower calorie levels, suggesting that many food items in this range have lower total fat content. The plots for Sodium, Carbs, Sugars, and Protein illustrate how these components are dispersed across the calorie spectrum. For instance, a high density in the Protein plot at higher calorie levels suggests that food items with more calories tend to have higher protein content.

4.4. Distribution of Fast Food Restaurants in the US

As shown in Figure.5 and Figure.6, the results indicate a significant disparity in the distribution of fast-food chains across provinces and states, with some chains exhibiting a stronger presence in particular regions. For instance, the density of stores in populous states like Texas and California is notably higher for all chains, reflecting the correlation between population density and store counts.

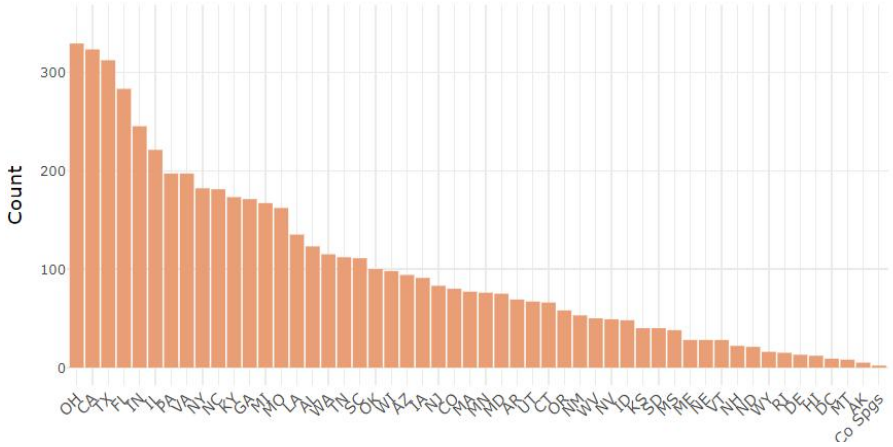


Figure. 5 Histogram of Stores Amount by States

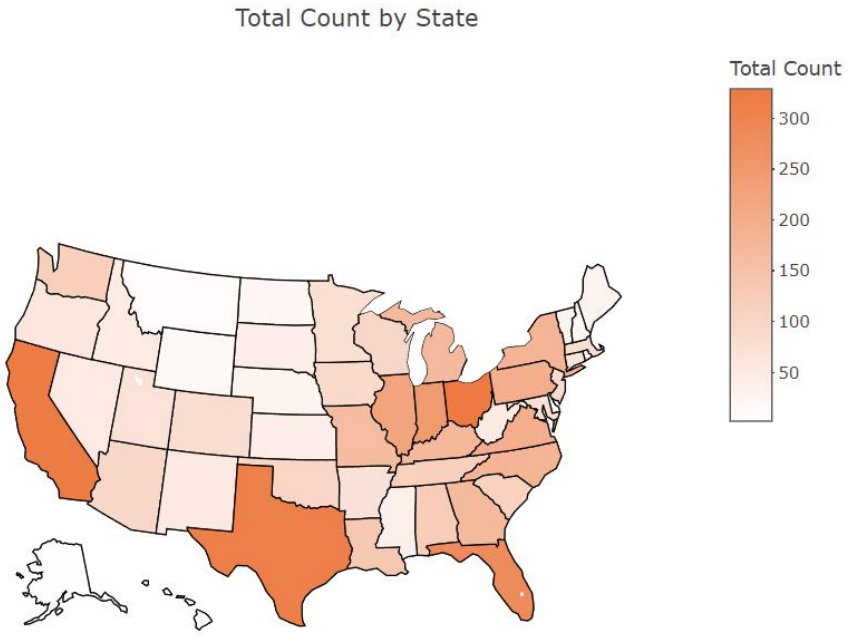


Figure. 6 Distribution Map

4.5. The Unhealth Food -- 3D Scatter Plot

These graphs visualize the relationship between the calorie, carbohydrate and sugar content of each fast food restaurant's menu. There is a wide distribution of points on all axes, suggesting a variety of calorie, carbohydrate and sugar content on the menu. From Figure 7 we can see that Chocolate Triple Thick is the one food item on the McDonald's menu that contains the most calories, sugar, and carbohydrates.

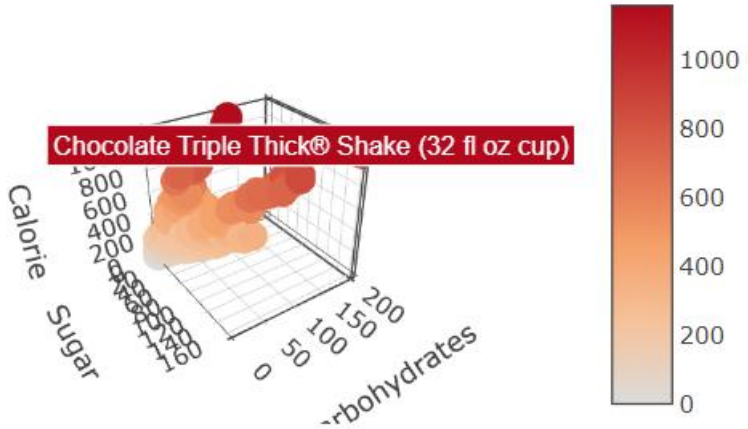


Figure. 7 Calories, sugar, carbohydrates in McDonalds's items

4.6. Total Fat vs Calories

-- Histogram & Scatter Plot

The visualisation shows a detailed scatterplot with histograms along the axes depicting the relationship between the calorie content of the food and the total fat. The scatterplot shows a direct positive correlation, i.e. the higher the fat content, the higher the calorie content. The histogram reflects the frequency distribution of the entire data set, showing a higher concentration of foods with lower calorie and fat content and fewer foods with both higher calorie and fat content. This nuanced representation of the data suggests that a significant number of foods manage to keep their calorie and fat content relatively low, which may be vital information for those

seeking a balanced diet. From Figure 8 we can see that KFC's Potato Salad contains the most fat and calories.

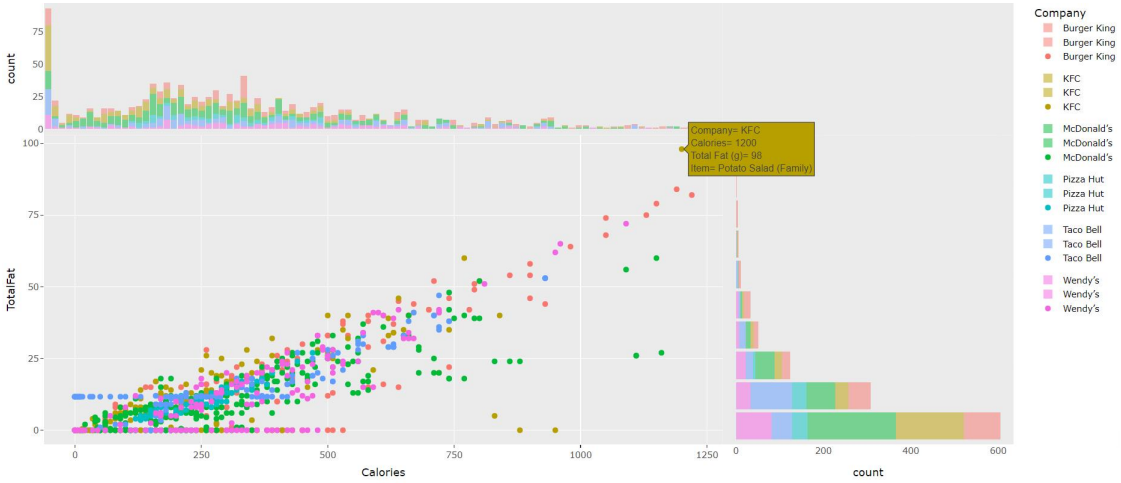


Figure. 8 Histogram & scatter Plot

5. Visualization design and the Corresponding Theories /Principles

1. Pie charts showing the proportion of 9 nutrients for each company:

(1) Visual Channels and Color Mapping: The pie charts use color and angle to represent nutrient proportions. However, the limited visual channels make it difficult to compare nutrient proportions across multiple companies. Therefore, we applied color mapping, representing the nine nutrients from least to most using varying color gradients.

(2) Interactivity and Expressiveness: Pie charts often lack interactivity, limiting detailed information or interactive exploration. Thus, we introduced hover functionality to display specific values, enhancing interactivity and effectiveness. To improve expressiveness, we employed different color

gradients to distinguish nutrient proportions, enhancing the charts' capability to display multiple nutrient distributions.

2. Frequency distribution plots (Violin plots) for each company's nutrients:

(1) Visual Channels and Color Mapping: Violin plots integrate box plots and kernel density plots, demonstrating nutrient distributions and data concentration.

(2) Interactivity and Expressiveness: Violin plots can offer interactivity by hovering over specific values or filtering data, facilitating in-depth data exploration. They also display central tendencies and dispersion, providing comprehensive insights.

(3) Effectiveness: When comparing nutrient distributions among multiple companies, violin plots effectively showcase these differences and similarities. By paralleling each nutrient's distribution for every company, it allows clear comparisons of nutrient usage preferences and proportions among companies.

3. Density plots for comparing the correlation between different components:

(1) Visual Channels and Color Mapping: Density plots use color gradients to represent correlations, providing clarity for readers to understand and analyze.

(2) Interactivity and Expressiveness: Density plots offer some interactivity for individual or limited component correlation analysis. To enhance interactivity, hover functionalities were implemented to display specific values.

(3) Effectiveness: Density plots effectively illustrate correlations in cases with a limited number of components. By separately displaying correlations between each pair of components, they are highly effective and clear.

4. 3D scatter plots for carbohydrates, calories, and sugar usage by each company:

(1) Visual Channels and Color Mapping: 3D scatter plots utilize three axes to represent the relationship between the three components. Each point's position on these axes depicts the corresponding component values, displaying

relationships through positional changes.

(2) Interactivity and Expressiveness: 3D scatter plots can possess interactivity, such as hovering over specific values or selecting companies for examination. They showcase relationships among three components for individual data points and visually enable accurate data point analysis.

(3) Effectiveness: 3D scatter plots effectively demonstrate relationships among three components and, based on point distribution, display correlations among these components.

5. Bar chart showing the number of fast-food companies in each U.S. state + Choropleth map:

(1) Visual Channels and Color Mapping: The bar chart illustrates the number of fast-food companies in each state, while the map provides geographic information, using color gradients to represent company quantities across states.

(2) Interactivity and Expressiveness: Both the map and bar chart offer interactivity, allowing users to explore data by hovering over values or selecting specific companies for inspection.

(3) Effectiveness: The bar chart vividly displays the quantity of outlets in each state, avoiding y-axis truncation to ensure data integrity. The choropleth map effectively represents outlet quantities using color depth, facilitating quick comprehension.

6. Conclusion

Our comprehensive analysis of fast-food nutritional content, particularly focusing on calories and fat, highlighted crucial insights. We found a consistent pattern across various fast-food chains where items with higher fat content also tended to have significantly higher caloric values. This relationship is critical for consumer health, as it underscores the impact of fat content on overall calorie consumption. Our conclusion stresses the need for consumers to be aware of these nutritional aspects to make healthier dietary choices. By opting for items with lower fat and calorie content, consumers can better manage their calorie intake, contributing to a healthier lifestyle. This understanding is vital in today's fast-paced world where fast-food consumption is prevalent, emphasizing the need for nutritional literacy and conscious eating habits.