Exploratory Data Analysis of McDonald's Nutrition Facts

By Tsion Woldeselassie

Abstract

This project analyzes the nutritional content of McDonald's menu items using the publicly available McDonald's Nutrition Facts dataset from Kaggle. The dataset contains 260 menu items and 24 features including calories, fat, sodium, protein, and vitamins. The goal was to explore how different nutrients relate to calorie content, visualize these relationships, and clean the data by handling outliers and missing values. I found that Total Fat and Saturated Fat are the strongest predictors of calories, while Vitamin C is the only nutrient with a negative correlation to calories. After applying the 1.5×IQR rule to remove outliers and replacing them with column means, the standard deviations dropped significantly across most features while the medians stayed mostly the same. This project demonstrates the importance of data exploration and preprocessing before any modeling or deeper analysis is done.

Introduction

Fast food nutrition has become an important public health topic. Many people eat at McDonald's regularly without knowing exactly how many calories or how much fat they are consuming. Understanding the nutritional makeup of menu items can help people make better choices and also gives us a good dataset to practice data analysis techniques on.

The goal of this project is to explore the McDonald's nutrition dataset, find which nutrients are most related to calorie content, visualize those relationships, and clean the data properly. This is relevant because in real-world data science, raw data always needs to be explored and cleaned before it can be used for predictions or decisions.

Literature Review

Nutritional data analysis has been used in many public health studies to understand eating habits and their effects on health. Researchers have shown that calorie intake is strongly linked to fat consumption, especially saturated fat, which aligns with basic nutritional science. Studies on fast food nutrition specifically have found that menu items tend to be calorie-dense and high in sodium and fat compared to home-cooked meals.

In data science, exploratory data analysis (EDA) is a well-established first step before any modeling. Techniques like correlation heatmaps, scatter plots, and box plots help analysts understand the structure of the data. Handling outliers using the IQR method is a standard preprocessing technique that reduces the effect of extreme values on statistical summaries and future models.

Methodology

Data Source: The dataset was downloaded from Kaggle (McDonald's Nutrition Facts) and contains 260 rows and 24 columns covering various menu categories like Breakfast, Beef & Pork, Beverages, and Salads.

Tools Used: All analysis was done in R using the following packages: ggplot2 for plotting, corrplot for the heatmap, and base R functions for statistical calculations.

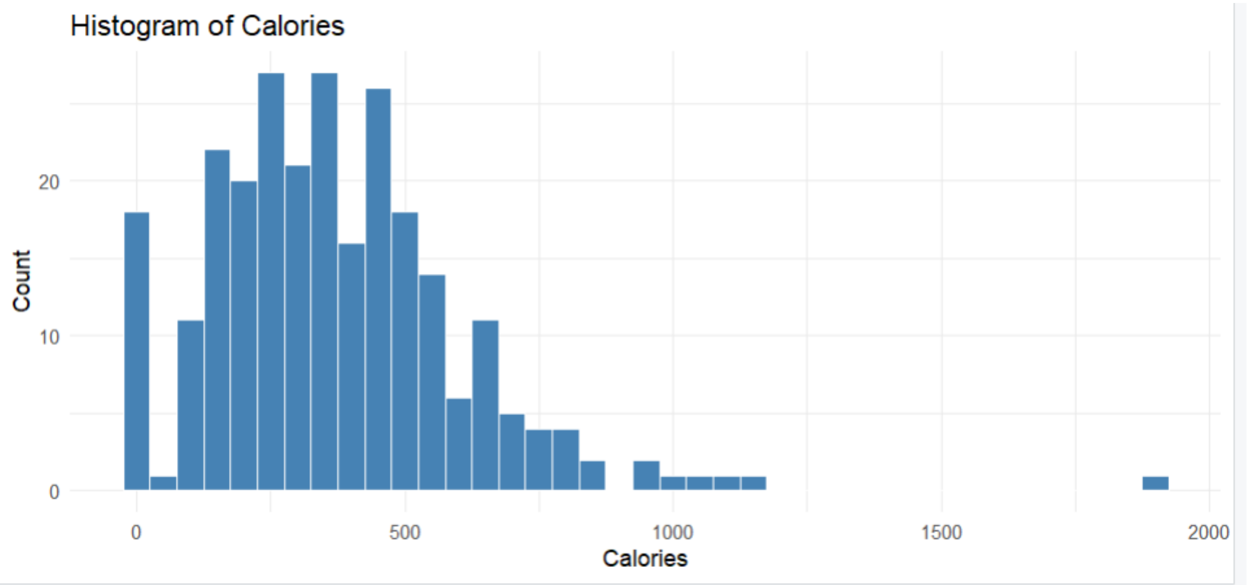
Preprocessing Steps: The dataset had 3 non-numeric columns (Category, Item, Serving Size) which were removed before numerical analysis, leaving 21 numeric features. Outliers were detected using the 1.5×IQR rule and replaced with NaN, then NaN values were filled in with the column mean.

Visualizations Used:

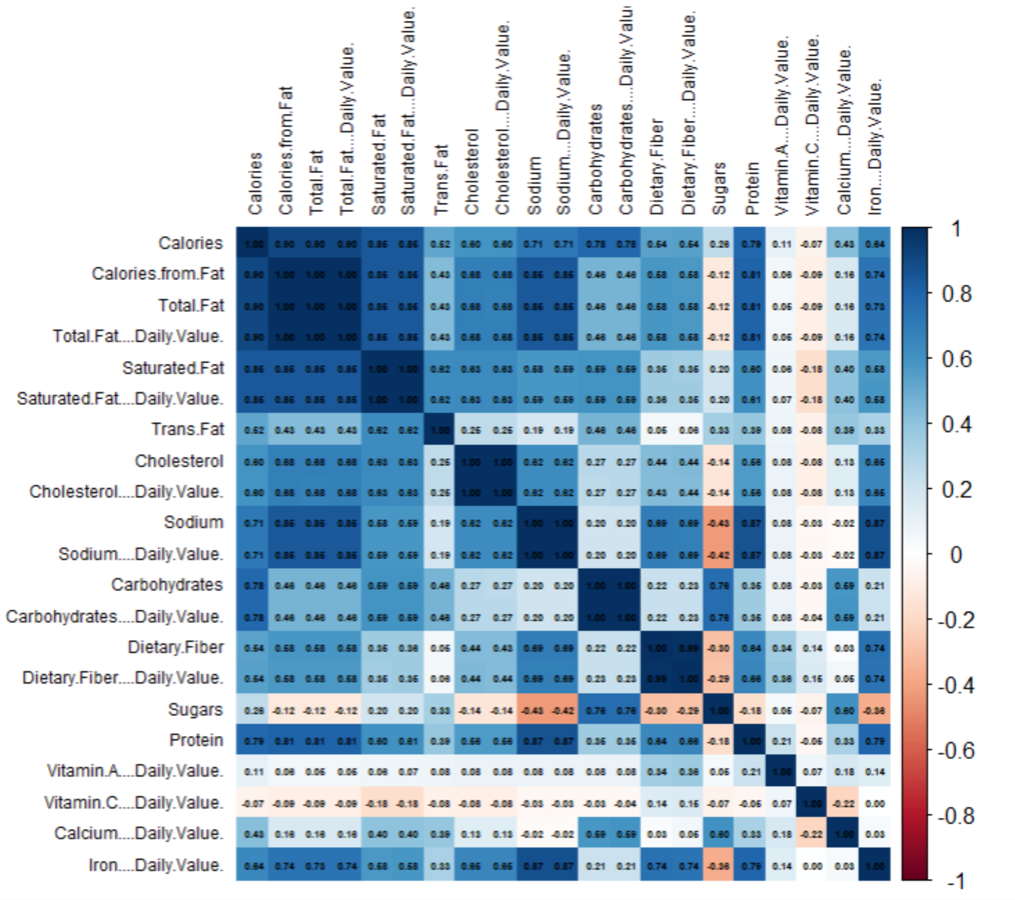
- Histogram to see the distribution of Calories
- Correlation heatmap to find relationships between all numeric features
- Scatter plots to visualize Calories vs selected features
- Box plots to show spread and outliers in key features

Results

Part 1 – Data Exploration: The histogram of Calories showed a right-skewed distribution, with most items between 0 and 500 calories and a peak around 300–400. A few items approached 2000 calories, pulling the tail to the right.

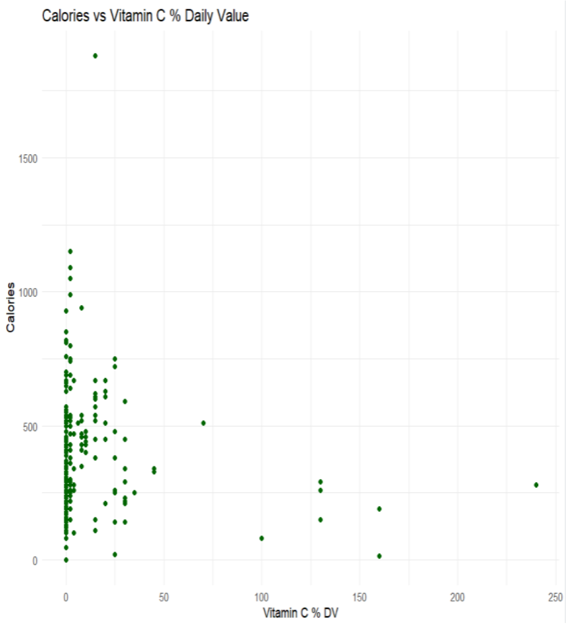


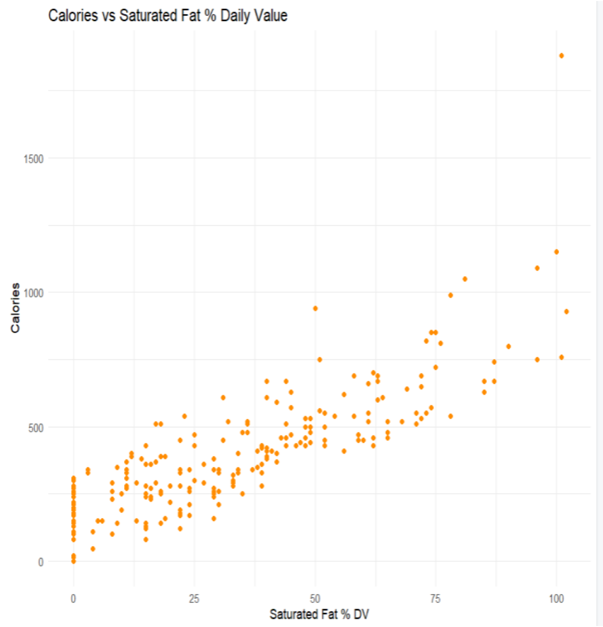
The 21×21 correlation heatmap showed that most fat-related features have strong positive correlations with Calories. The diagonal is always 1.0 because any feature compared to itself is a perfect match.

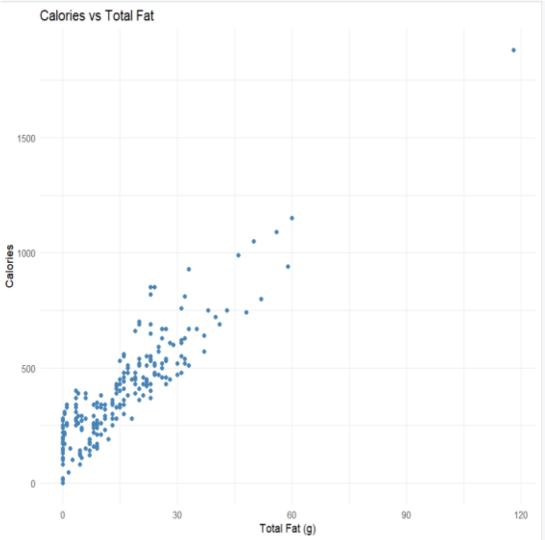


The features with the 2nd and 3rd highest positive correlations with Calories were Total Fat (0.904) and Saturated Fat % Daily Value (0.848). The only feature with a negative correlation was Vitamin C % Daily Value (−0.069).

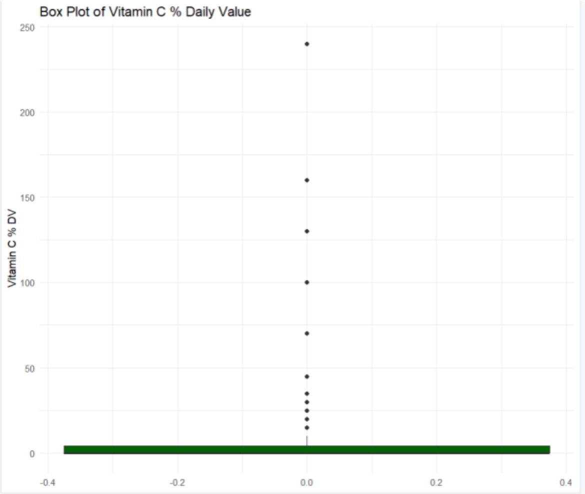
Part 2 – Plots: The scatter plot of Calories vs Total Fat showed a strong upward linear trend. Calories vs Saturated Fat % DV also showed a clear positive trend but with slightly more spread. Calories vs Vitamin C showed no real trend, points were scattered with no pattern.

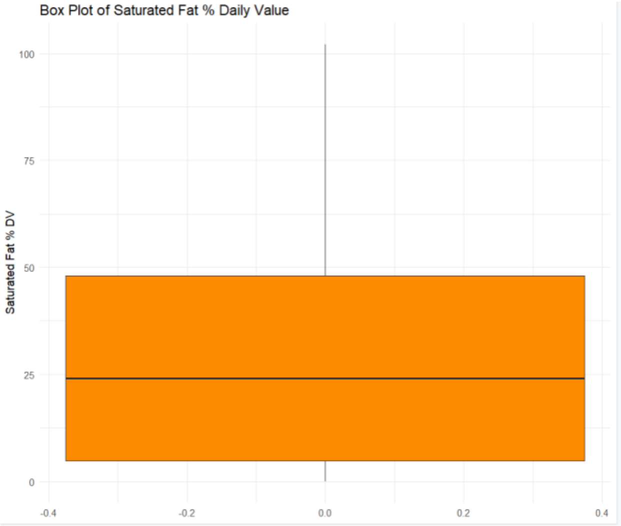

Calories vs Vitamin C % Daily Value

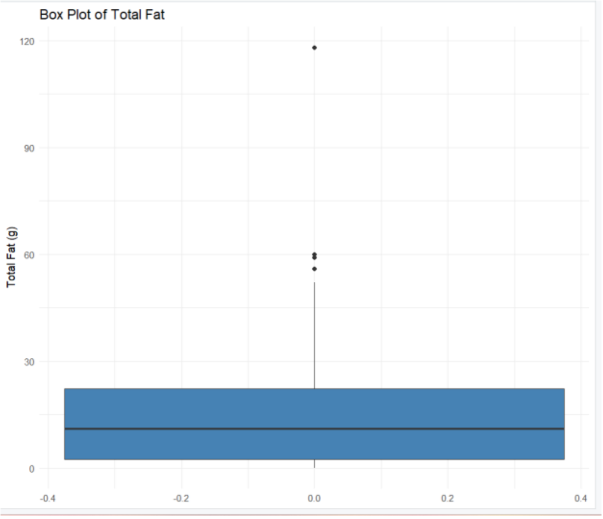

Calories vs Saturated Fat % Daily Value


Calories vs Total Fat

The box plots showed that Total Fat had 3 high outliers (up to 120g), Vitamin C was heavily concentrated at 0 with many outlier dots above, and Saturated Fat % DV had a fairly wide spread with no outliers.







Part 3 – Preprocessing: Before cleaning, Calories had a median of 340 and SD of 240.27. After replacing outliers with the column mean, the median stayed at 340 but the SD dropped to 199.05. Trans Fat went from SD of 0.43 down to 0.00 after its 56 outliers were replaced. Cholesterol's SD dropped from 87.27 to 29.85, and Vitamin C's SD dropped dramatically from 26.35 to 2.08.

Feature	**Median Before**	**Median After**	**SD Before**	**SD After**
Calories	**340**	**340**	**240.27**	**199.05**
Trans.Fat	**0**	**0**	**0.43**	**0.00**
Cholesterol	**35**	**34.67**	**87.27**	**29.85**

Feature	Median Before	Median After	SD Before	SD After
Vitamin.C	0	0	26.35	2.08
Sodium	190	190	577.03	498.40

Discussion/Analysis

The results make a lot of sense from a nutritional perspective. Fat contains 9 calories per gram, which is more than carbohydrates or protein (both 4 calories per gram), so it's expected that Total Fat and Saturated Fat would have the strongest relationship with Calories. Vitamin C being the only negatively correlated feature also makes sense because Vitamin C is found mostly in fruits and vegetables, which tend to be low in calories, while the high-calorie McDonald's items like burgers and milkshakes have almost no Vitamin C.

One limitation is that the negative correlation for Vitamin C is very weak (−0.069), meaning it's barely a real relationship. It's close enough to zero that we can't say Vitamin C truly predicts lower calories.

Another thing worth noting is that the IQR outlier method flagged Trans Fat (56 outliers) and Vitamin C (46 outliers) the most, not because those values are errors, but because most items have zero for those features, making any non-zero value look extreme statistically. This is a limitation of the IQR method when data is heavily zero-inflated. Replacing those outliers with the mean pulled the SD down significantly, which makes the data look more uniform than it really is.

Conclusion

This project explored McDonald's nutritional data using R. I found that Total Fat and Saturated Fat are the strongest predictors of calorie content, while Vitamin C is the only feature with a negative relationship to calories. After cleaning the data using the IQR rule and mean imputation, the medians stayed stable but standard deviations dropped across most features, showing that outlier removal reduces variability in the data. Overall, this project showed how important it is to explore and clean data before drawing conclusions, and how simple visualizations can reveal meaningful patterns in real-world datasets.

References

1. Kaggle. (2017). *McDonald's Nutrition Facts Dataset*. Retrieved from https://www.kaggle.com/mcdonalds/nutrition-facts

2. Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

3. Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer.

4. R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.