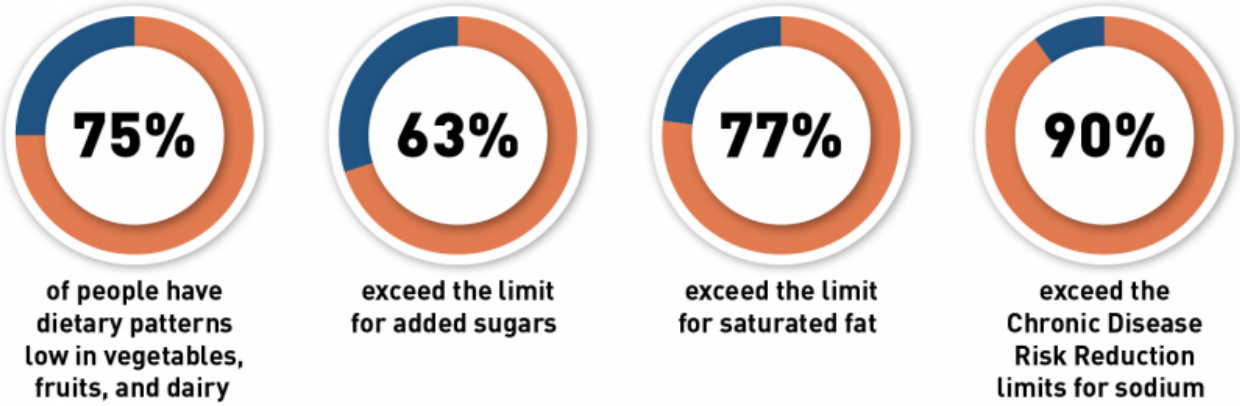
UNDERSTANDING THE RELATIONSHIP BETWEEN NUTRITIONAL CONTENT AND CEREAL RATINGS: A DATA-DRIVEN APPROACH

OVERVIEW

According to the US Food Drug Administration (FDA)



In an era where quick meals are paramount, breakfast cereals stand out for their convenience. However, with this convenience comes a question of nutritional value, which varies greatly among different cereals. The US Food and Drug Administration (FDA) is amplifying efforts to educate consumers about nutritionally deficient options. The label "healthy" is not just a marketing term; it's a regulated claim that can only be used if the product meets specific nutritional criteria. Against the backdrop of heightened health consciousness and the crucial role of breakfast cereals in dietary patterns—particularly in underprivileged homes where food diversity is scant—our project seeks to navigate the cereal aisle effectively, ensuring food security and combating malnutrition (https://www.fda.gov/food/food-labeling-nutrition/use-term-healthy-food-labeling).

Problem Statement

Many households rely on breakfast cereals, but knowing which options are truly healthy is challenging, especially in households with limited food choices.

Aim

To simplify cereal selection by analyzing nutritional content, categorize cereals as ‘Healthy’ or ‘Unhealthy’ and predict consumer ratings.

Objectives:

- Analysis of nutritional content provided.
- Predict ratings of cereals based on their nutritional makeup, providing a reliable guide for consumers.
- Systematically categorize cereals into 'Healthy' and 'Unhealthy' groups based on established nutritional criteria.
- Share insights from the analysis to enhance consumer understanding of cereal nutrition.

Data Description

80 Cereal data set contains the nutrition information from 77 unique cereals from 7 different manufacturers. The data also provides information about the serving sizes, that is, how many cups make a serving per cereal and the measurements of several nutritional items present in it.

The data set was taken from the Kaggle competition page - https://www.kaggle.com/datasets/crawford/80-cereals/data

The dataset contains a list of 77 different cereals, their manufacturer, the measurement of food nutrients present, the display shelf, the weight of each serving, the number of cups per serving, and the ratings of each cereal. The description of the fields is as follows:

- name: Name of cereal
- mfr: Manufacturer of cereal
    - A = American Home Food Products
    - G = General Mills
    - K = Kelloggs
    - N = Nabisco
    - P = Post
    - Q = Quaker oats
    - R = Ralston Purina

- type:
    - C = Cold
    - H = Hot
- calories: calories per serving
- protein: grams of protein
- fat: grams of fat
- sodium: milligrams of sodium
- fiber: grams of dietary fiber
- carbo: grams of complex carbohydrates
- sugars: grams of sugars
- potass: milligrams of potassium
- vitamins: vitamins and minerals - 0, 25, 100, indicating the typical percentage of FDA recommended
- shelf: display shelf (1, 2, or 3, counting from the floor)
- weight: weight in ounces of one serving
- cups: number of cups in one serving
- rating: a rating of the cereals (Possibly from Consumer Reports)

Below is a preview of the dataset for the study.

	name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
0	100% Bran	N	C	70	4	1	130	10.0	5.0	6	280	25	3	1.0	0.33	68.402973
1	100% Natural Bran	Q	C	120	3	5	15	2.0	8.0	8	135	0	3	1.0	1.00	33.983679
2	All-Bran	K	C	70	4	1	260	9.0	7.0	5	320	25	3	1.0	0.33	59.425505
3	All-Bran with Extra Fiber	K	C	50	4	0	140	14.0	8.0	0	330	25	3	1.0	0.50	93.704912
4	Almond Delight	R	C	110	2	2	200	1.0	14.0	8	-1	25	3	1.0	0.75	34.384843

PROCESS OUTLINE

- Data Cleaning
- Exploratory Data Analysis
- Data Preprocessing
- Model Selection & Training
- Z-Score Normalization
- Evaluation
- Deployment

DATA CLEANING

Two important steps taken include:

- **Renaming manufacturer abbreviations for clarity**

```
# Replace the manufacturers initials with the Manufacturers full name in the mfr column
cereal["mfr"] = cereal["mfr"].replace(['N', 'Q', 'K', 'R', 'G', 'P', 'A'],\
        ['Nabisco', 'Quaker Oats', 'Kelloggs', 'Ralston Purina',\
          'General Mills', 'Post', 'American Home Food Products'])
```

- **Rectifying negative nutrient values by replacing them with the mean.**

```
print(cereal[cereal.select_dtypes(exclude='object')<0].count())
```

```
name        0
mfr         0
type        0
calories    0
protein     0
fat         0
sodium      0
fiber       0
carbo       1
sugars      1
potass      2
vitamins    0
shelf       0
weight      0
cups        0
rating      0
dtype: int64
```

```
# Replace negative nutrient values with the mean
for col in cereal.select_dtypes(exclude='object').columns:
    cereal.replace({col:-1}, np.mean(cereal[col]), inplace=True)
print(cereal[cereal.select_dtypes(exclude='object')<0].count())
```

```
name        0
mfr         0
type        0
calories    0
protein     0
fat         0
sodium      0
fiber       0
carbo       0
sugars      0
potass      0
vitamins    0
shelf       0
weight      0
cups        0
rating      0
dtype: int64
```

EXPLORATORY DATA ANALYSIS (EDA)

In the Exploratory Data Analysis phase, we took a comprehensive dive into the dataset. Here, the aim was to get a better understanding, uncover the underlying and irregular patterns and gain robust insights from the data.

Univariate Analysis

We examined each nutritional feature individually to establish a baseline understanding. This provided a foundation for understanding each feature on its own before considering interactions and relationships with other features.

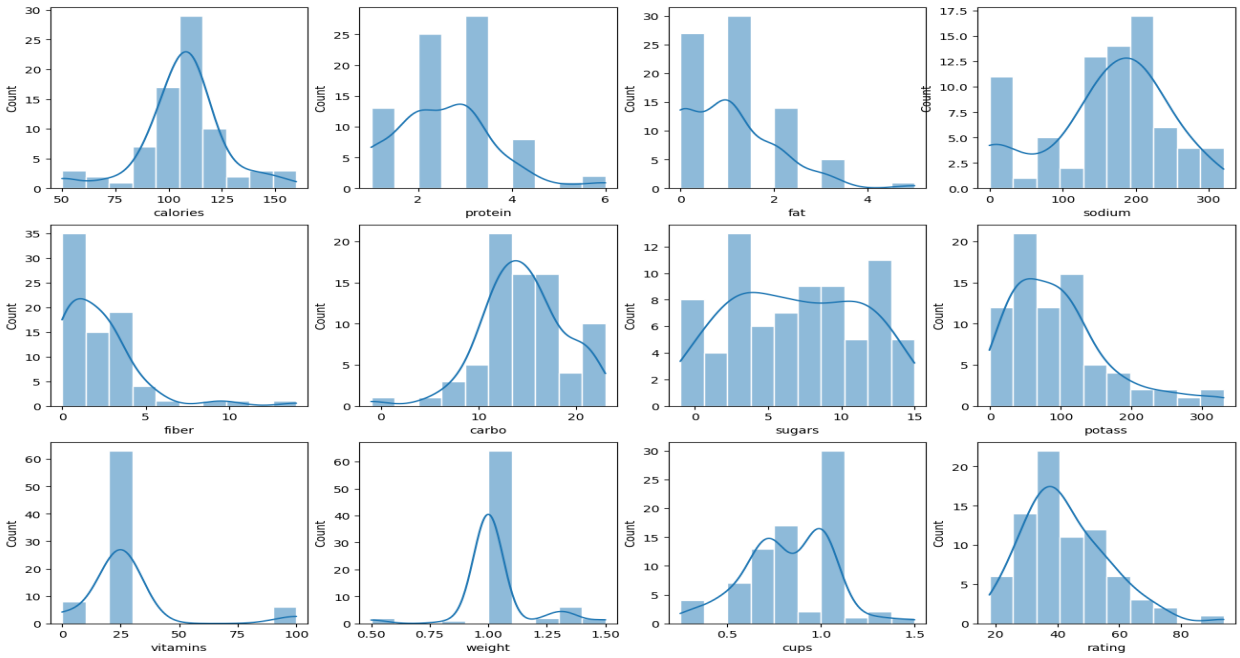


Fig. 1

From this, the following was observed:

- More than 30 of the products are ranging from 100-120 calories
- Over 50 products have a protein content ranging from 2-3g
- Over 55 products have a fat content ranging from 0-2g
- More than 44 products have a sodium content ranging from 150-200mg
- About 35 products have a fibre content of 0-2g
- More than 50 products contain carbohydrate of 12-18g
- More than 15 products have a sugar content of 2-3g, and about 12 products have a sugar content of 12-13g

- About 60 products have a vitamin content of 25%

Heat Map

The heat map illustrates feature correlations, highlighting the relationships between various nutrients and rating.

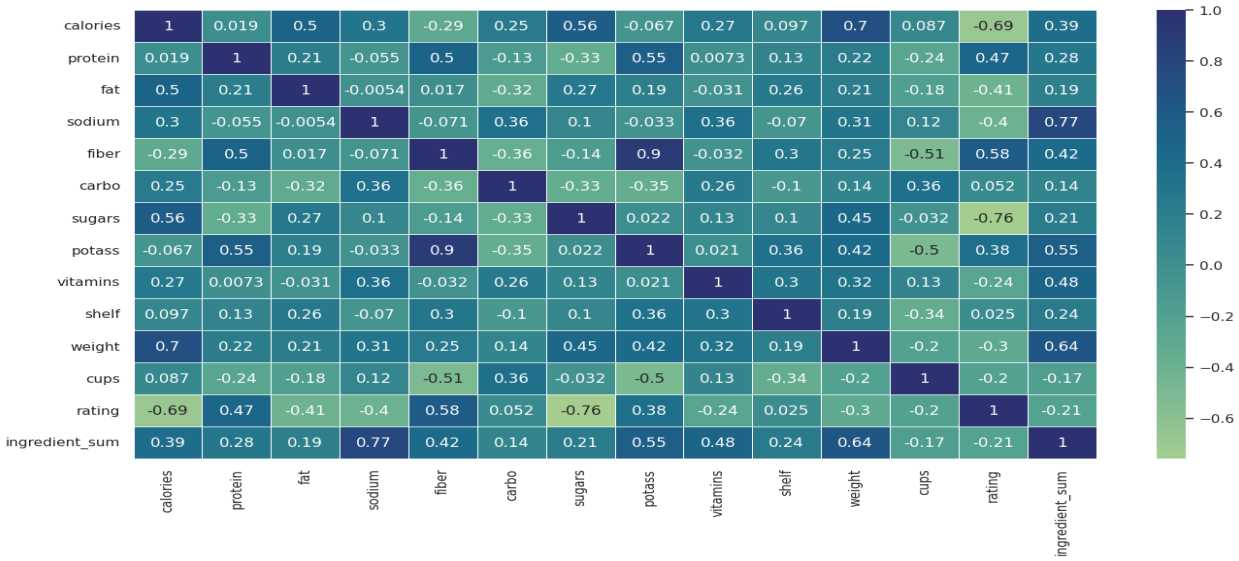


Fig. 2

This shows that the highest correlation seen is 0.9 which is between fiber and Potass.
Other correlations ranging from 0.5-0.7 are seen in calories and weight, rating and fiber, calories and fat, calories and sugar, protein and fiber, and protein and potass.

Fig. 3

This shows that rating has a positive moderate relationship with fiber, protein and potassium and a high negative relationship with sugars and calories. Carbohydrates, shelf, cups and vitamins have a very weak relationship with the customer ratings.

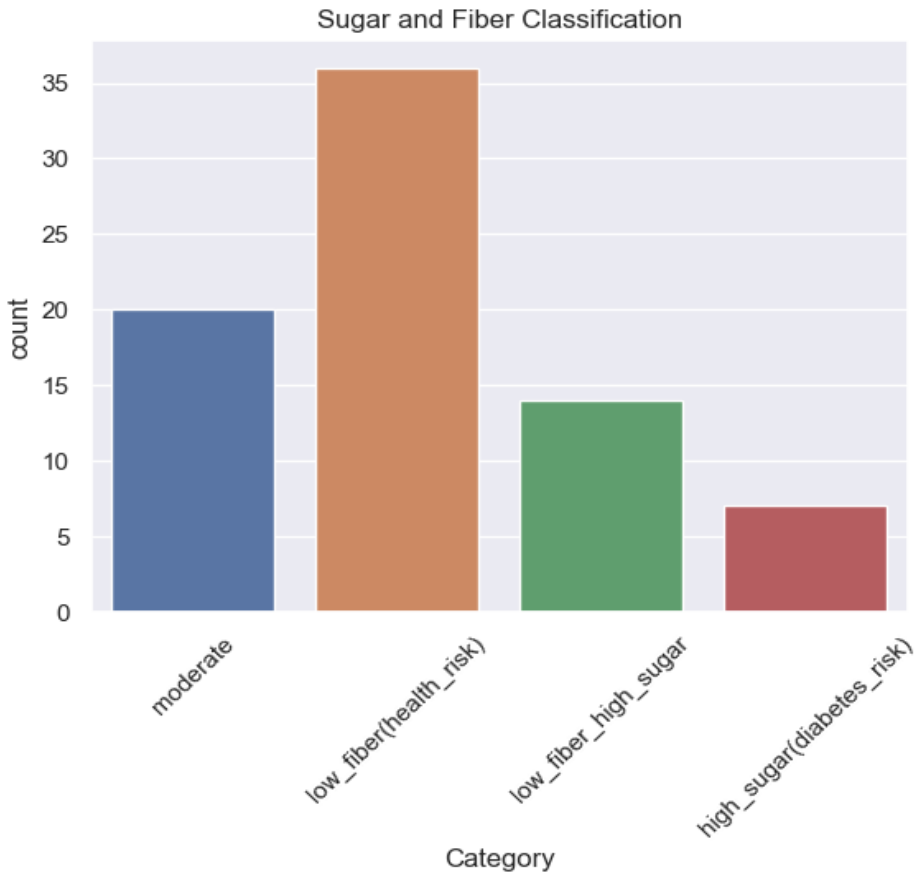


Fig. 4

From our nutrient analysis, we inferred that 46.8% of the cereals in this dataset have low Fiber(Health Risk) while 26% of the cereals in the dataset are both low in fiber and high in sugar which shows half of the cereals have low fiber content. 9.1% of the cereals in the dataset have high sugar(Diabetes Risk) which is the lowest relatively.

An article stated that less than 3g of fiber is not healthy enough while another article stated that a serving size of healthy cereal should not contain more than 10g of sugar.

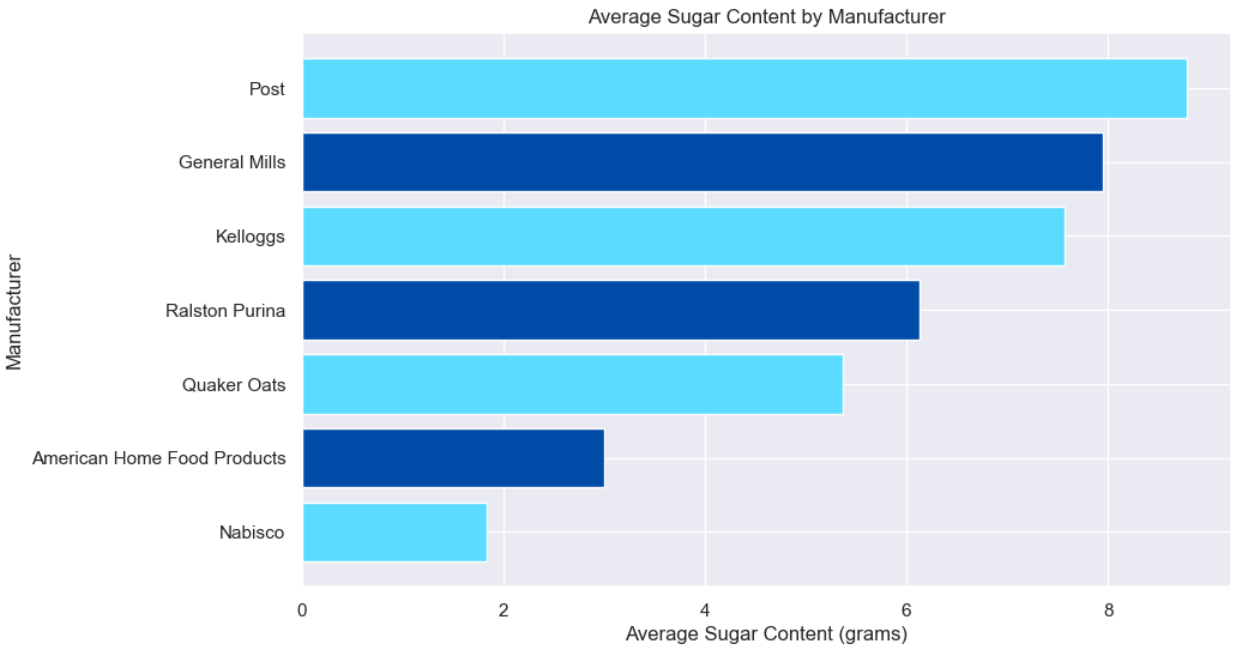


Fig. 5

From the above visualization, it can be observed that Post cereal manufacturers have the highest sugar content while Nabisco has the lowest sugar content when compared to other products.

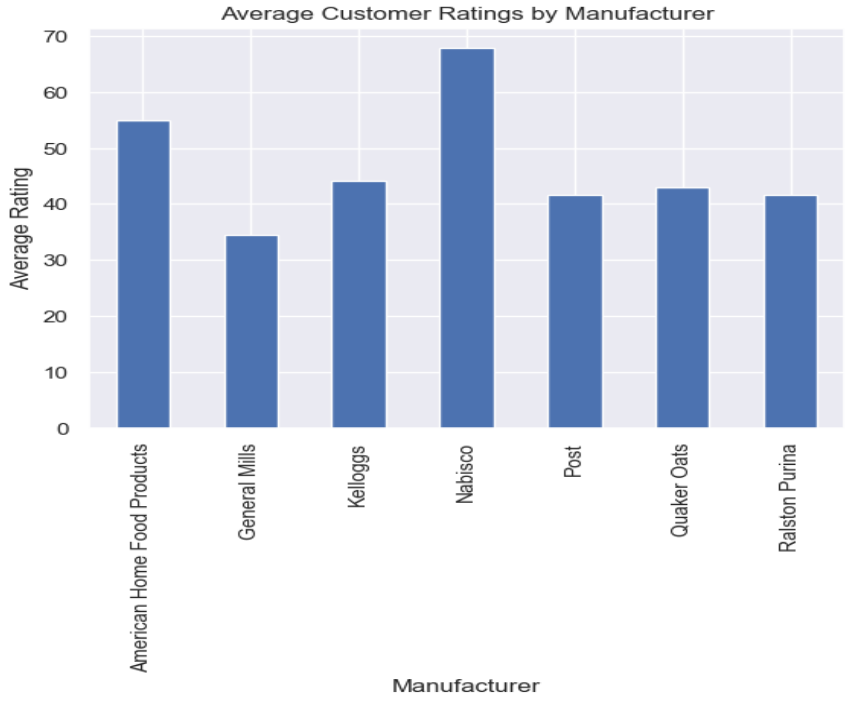


Fig. 6

Concurrently, it can be observed from the visualisation in Fig.6 that Nabisco has the highest customer rating while General Mills has the lowest customer rating.

```
#categorizing cerals based on their ratings
bins = [0, 20, 40, 60, 80, 100]
labels =  ['Very Poor', 'Poor', 'Average', 'Good', 'Excellent']
cereal['rating_category'] = pd.cut(cereal['rating'], labels=labels, bins=bins)
```

A rating category was created by mapping the ratings. Pandas cut() function, which takes in the 'rating' column as an array to be binned, was used.

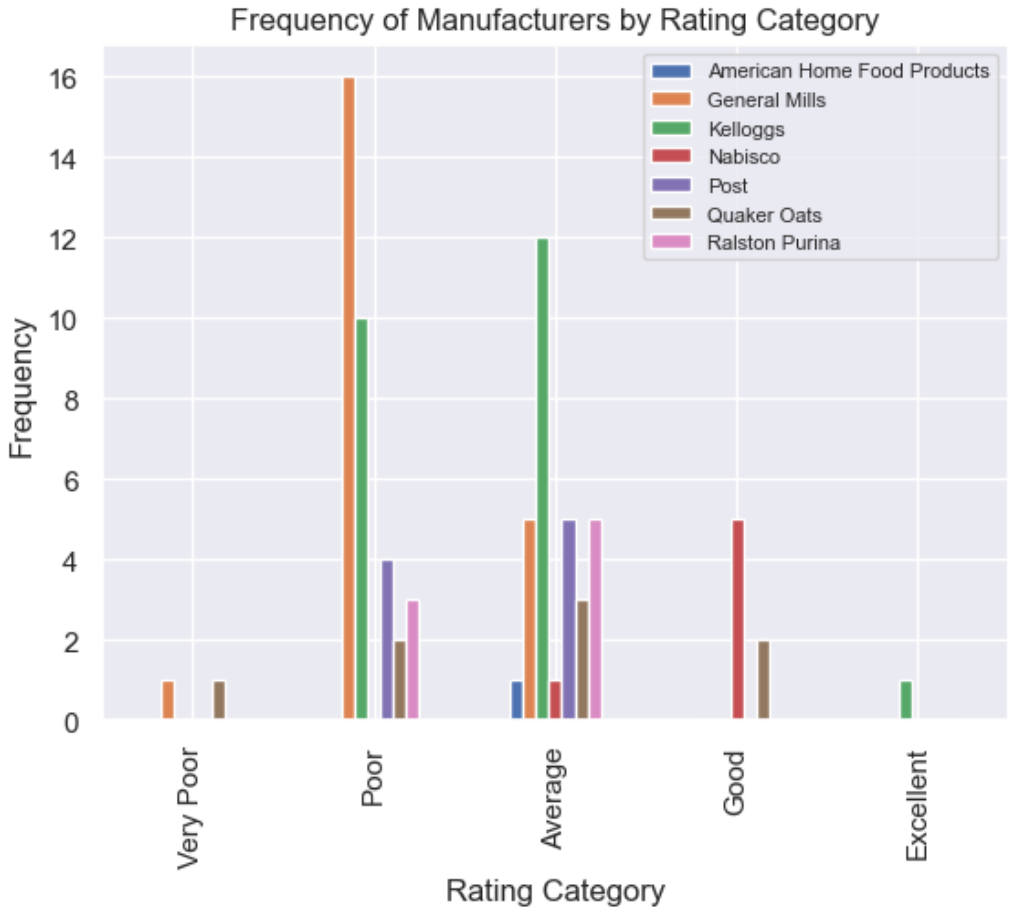


Fig.7

With further analysis, we can see that only Kelloggs cereal was considered worthy of the title, 'Excellent', though it was only one Cereal. However, Nabisco and Quaker Oat were not far behind, with Nabisco having five(5) out of six(6) of their cereals as considered good. Over half of Kelloggs cereals are considered average while the rest except one(1), were rated poor. The majority of General Mills cereals are in the Poor category. The rest of the manufacturer's cereals;

Ralston Purina's, Quaker Oats', Post's, and American Home Food Products' were mostly found to be poor or average or both in the rating category.

The analysis was further narrowed down to the name of the cereals rated “Excellent” and “Very Poor” along with their food nutrients.

```
excellent_cereal = cereal.nlargest(1, columns='rating')
print(excellent_cereal)
```

```
                        name       mfr type  calories  protein  fat  sodium  \
3  All-Bran with Extra Fiber  Kelloggs    C        50        4    0     140

   fiber  carbo  sugars  potass  vitamins  shelf  weight  cups     rating  \
3   14.0    8.0     0.0   330.0        25      3     1.0   0.5  93.704912

  category sugar_fiber rating_category
3     Baby    moderate       Excellent
```

```
very_poor_cereals = cereal.nsmallest(2, columns='rating')
print(very_poor_cereals)
```

```
                    name            mfr type  calories  protein  fat  sodium  \
10          Cap'n'Crunch    Quaker Oats    C       120        1    2     220
12  Cinnamon Toast Crunch  General Mills    C       120        1    3     210

    fiber  carbo  sugars  potass  vitamins  shelf  weight  cups     rating  \
10    0.0   12.0    12.0    35.0        25      2     1.0  0.75  18.042851
12    0.0   13.0     9.0    45.0        25      2     1.0  0.75  19.823573

   category             sugar_fiber rating_category
10    Adult    low_fiber_high_sugar       Very Poor
12    Adult  low_fiber(health_risk)       Very Poor
```

The cereal rated as Excellent is All-Bran with Extra Fiber manufactured by Kelloggs.
The cereals rated as the poorest are Cap'n'Crunch manufactured by Quaker Oats and Cinnamon Toast Crunch manufactured by General Mills.

DATA PREPROCESSING

After the EDA, the next phase is model building and implementation. First, the data was separated into features and target, the categorical variable was encoded using OneHotEncoder, multicollinearity was removed, and the data was split into train and test.

```
#  Define your features (X) and target (y)

X = cereal_df.drop(['name', 'mfr', 'potass','rating', 'shelf'], axis = 1)
# Potassium was dropped due to its high correlation with fiber (multicollinearity)

# Transform 'type' to categorical with OneHotEncoder

encoder = OneHotEncoder(sparse=False)

# Fit and transform the 'type' column
encoded = encoder.fit_transform(X[['type']])

# Create a DataFrame from the encoded values with column names
encoded_df = pd.DataFrame(encoded, columns=encoder.get_feature_names_out(['type']))

# Concatenate the new DataFrame with the original DataFrame
X = pd.concat([X, encoded_df], axis=1)

# Drop the original 'type' column if needed
X = X.drop(['type'], axis=1)
y = cereal_df['rating']

X.head()
```

MODEL BUILDING AND PERFORMANCE

Different machine learning models were deployed. The models include linear regression, ridge regression and lasso regression. These three models performed well, but ridge regression had the best outcome with R2score of 0.995, Mean ABsolute error of 0.75 and Root Mean Square Error of 1.022.

Z-Score Normalization

Attempting to enhance the model performance, the Z-score normalization was applied but observed a negligible impact on the outcomes.

The metric scores of the models are shown below.

	Model	MAE	MAE_norm	RMSE	RMSE_norm	R2_Score	R2_Score_norm
0	Linear Regression	0.719044	0.719044	1.042886	1.042886	0.995049	0.995049
1	Ridge Regression	0.748836	0.759653	1.022282	1.001196	0.995242	0.995437
2	Lasso Regression	0.718284	0.718829	1.041871	1.041894	0.995058	0.995058

MODEL DEPLOYMENT

The best performing model was deployed on Streamlit, due to its ease of use in deploying machine learning models. This allowed us to interactively present our results through a user-friendly interface.

Here is the link to the deployed model: https://cerealratings.streamlit.app/?utm_medium=social

CONCLUSION

This project evaluated the nutritional value of popular breakfast cereals by conducting an extensive Exploratory Data Analysis to scrutinize nutrient content.
We subsequently developed a machine learning model that forecasts consumer ratings.

CHALLENGES FACED

- Getting standard cereal requirement for all nutrients was difficult as this information is not readily available on the FDA website. Resorted to using articles based on other research.
- With a relatively small dataset, the risk of overfitting increases. We may require more data to build a more robust model.
- The information about how the cereal ratings are determined is not readily available. Hence, the variability in cereal ratings may not be consistent across different levels of predictor variables, leading to heteroscedasticity.
- Some predictor variables in the dataset display multicollinearity and can confound the individual impact they have on the dependent variable, in this case, cereal ratings. For example, potassium and fiber have a strong correlation coefficient of 0.9. Due to this high correlation and the common nutritional emphasis on fiber, potassium was excluded from the model to help isolate the unique contribution of each variable to the cereal ratings.

RECOMMENDATION

However, to increase the predictive power and generalizability of the model, we recommend an expanded data collection encompassing a wider variety of cereal brands. Also, as new cereal products are introduced, and nutritional contents are regulated, the model must be regularly updated with new data and adjusted to maintain its relevance and accuracy.