```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
FOOD1 = pd.read_csv('/content/drive/My Drive/Colab Notebooks/1. Data/FOOD/FOOD-DATA-GROUP1.csv')
FOOD2 = pd.read_csv('/content/drive/My Drive/Colab Notebooks/1. Data/FOOD/FOOD-DATA-GROUP2.csv')
FOOD3 = pd.read_csv('/content/drive/My Drive/Colab Notebooks/1. Data/FOOD/FOOD-DATA-GROUP3.csv')
FOOD4 = pd.read_csv('/content/drive/My Drive/Colab Notebooks/1. Data/FOOD/FOOD-DATA-GROUP4.csv')
FOOD5 = pd.read_csv('/content/drive/My Drive/Colab Notebooks/1. Data/FOOD/FOOD-DATA-GROUP5.csv')
```

```
FOOD1.head()
```

	Unnamed: 0.1	Unnamed: 0	food	Caloric Value	Fat	Saturated Fats	Monounsaturated Fats	Polyunsaturated Fats	Carbohydrates	Sugars	...	Calcium	Co
0	0	0	cream cheese	51	5.0	2.9	1.3	0.200	0.8	0.500	...	0.008	14
1	1	1	neufchatel cheese	215	19.4	10.9	4.9	0.800	3.1	2.700	...	99.500	0
2	2	2	requeijao cremoso light catupiry	49	3.6	2.3	0.9	0.000	0.9	3.400	...	0.000	0
3	3	3	ricotta cheese	30	2.0	1.3	0.5	0.002	1.5	0.091	...	0.097	41
4	4	4	cream cheese low fat	30	2.3	1.4	0.6	0.042	1.2	0.900	...	22.200	0

5 rows × 37 columns

```
FOOD2.head()
```

	Unnamed: 0.1	Unnamed: 0	food	Caloric Value	Fat	Saturated Fats	Monounsaturated Fats	Polyunsaturated Fats	Carbohydrates	Sugars	...	Calcium	Co
0	0	0	eggnog	224	10.6	6.6	3.3	0.5	20.4	20.4	...	330.2	(
1	1	1	beer light	96	0.0	0.0	0.0	0.0	5.4	0.3	...	13.2	(
2	2	2	beer budweiser	12	0.0	0.0	0.0	0.0	0.9	0.0	...	1.2	(
3	3	3	weizenbier erdinger	220	18.0	13.0	1.0	0.0	0.0	0.0	...	0.0	(
4	4	4	beer light budweiser	9	0.0	0.0	0.0	0.0	0.4	0.0	...	0.9	(

5 rows × 37 columns

```
FOOD3.head()
```

	Unnamed: 0.1	Unnamed: 0	food	Caloric Value	Fat	Saturated Fats	Monounsaturated Fats	Polyunsaturated Fats	Carbohydrates	Sugars	...	Calcium	Co
0	0	0	nectarine	66	0.500	0.066	0.100	0.200	15.8	11.8	...	0.081	9
1	1	1	kiwifruit gold	51	0.200	0.008	0.099	0.051	12.8	10.0	...	13.800	(
2	2	2	prickly pear raw	8	0.072	0.000	0.000	0.000	1.9	0.2	...	34.200	(
3	3	3	pineapple	45	0.100	0.074	0.001	0.087	11.8	8.9	...	0.061	1
4	4	4	rowan	253	4.600	0.600	0.000	0.000	54.5	32.1	...	34.200	2

5 rows × 37 columns

```
FOOD4.head()
```

	Unnamed: 0.1	Unnamed: 0	food	Caloric Value	Fat	Saturated Fats	Monounsaturated Fats	Polyunsaturated Fats	Carbohydrates	Sugars	...	Calcium	Co
0	0	0	chocolate pudding fat free	105	0.3	0.0	0.00	0.000	23.6	17.8	...	44.100	0
1	1	1	tapioca pudding	143	4.3	1.1	2.80	0.088	23.9	16.4	...	78.100	0
2	2	2	tapioca pudding fat free	105	0.4	0.1	0.08	0.067	23.9	15.9	...	58.200	0
3	3	3	rice pudding	122	2.4	1.4	0.60	0.100	20.8	13.1	...	0.063	107
4	4	4	corn pudding	328	12.6	6.3	3.90	1.400	42.4	16.5	...	0.066	97

5 rows × 37 columns

```
FOOD5.head()
```

	Unnamed: 0.1	Unnamed: 0	food	Caloric Value	Fat	Saturated Fats	Monounsaturated Fats	Polyunsaturated Fats	Carbohydrates	Sugars	...	Calcium	C
0	0	0	margarine with yoghurt	88	9.8	1.9	5.6	2.0	0.073	0.0	...	2.8	
1	1	1	sunflower seed butter	99	8.8	0.7	6.2	1.6	3.700	1.7	...	10.2	
2	2	2	hazelnut oil	120	13.6	1.0	10.6	1.4	0.000	0.0	...	0.0	
3	3	3	menhaden fish oil	1966	218.0	66.3	58.2	74.5	0.000	0.0	...	0.0	
4	4	4	cod liver fish oil	123	13.6	3.1	6.4	3.1	0.000	0.0	...	0.0	

5 rows × 37 columns

DATA CLEANING AND MANIPULATION

```
# Check the dataset

FOOD1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 551 entries, 0 to 550
Data columns (total 37 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Unnamed: 0.1          551 non-null    int64
 1   Unnamed: 0            551 non-null    int64
 2   food                  551 non-null    object
 3   Caloric Value         551 non-null    int64
 4   Fat                   551 non-null    float64
 5   Saturated Fats        551 non-null    float64
 6   Monounsaturated Fats  551 non-null    float64
 7   Polyunsaturated Fats  551 non-null    float64
 8   Carbohydrates         551 non-null    float64
 9   Sugars                551 non-null    float64
 10  Protein               551 non-null    float64
 11  Dietary Fiber         551 non-null    float64
 12  Cholesterol           551 non-null    float64
 13  Sodium                551 non-null    float64
 14  Water                 551 non-null    float64
 15  Vitamin A             551 non-null    float64
 16  Vitamin B1            551 non-null    float64
 17  Vitamin B11           551 non-null    float64
 18  Vitamin B12           551 non-null    float64
 19  Vitamin B2            551 non-null    float64
 20  Vitamin B3            551 non-null    float64
 21  Vitamin B5            551 non-null    float64
 22  Vitamin B6            551 non-null    float64
 23  Vitamin C             551 non-null    float64
 24  Vitamin D             551 non-null    float64
```

```
 25  Vitamin E             551 non-null    float64
 26  Vitamin K             551 non-null    float64
 27  Calcium               551 non-null    float64
 28  Copper                551 non-null    float64
 29  Iron                  551 non-null    float64
 30  Magnesium             551 non-null    float64
 31  Manganese             551 non-null    float64
 32  Phosphorus            551 non-null    float64
 33  Potassium             551 non-null    float64
 34  Selenium              551 non-null    float64
 35  Zinc                  551 non-null    float64
 36  Nutrition Density     551 non-null    float64
dtypes: float64(33), int64(3), object(1)
memory usage: 159.4+ KB
```

```
FOOD2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 319 entries, 0 to 318
Data columns (total 37 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Unnamed: 0.1          319 non-null    int64
 1   Unnamed: 0            319 non-null    int64
 2   food                  319 non-null    object
 3   Caloric Value         319 non-null    int64
 4   Fat                   319 non-null    float64
 5   Saturated Fats        319 non-null    float64
 6   Monounsaturated Fats  319 non-null    float64
 7   Polyunsaturated Fats  319 non-null    float64
 8   Carbohydrates         319 non-null    float64
 9   Sugars                319 non-null    float64
 10  Protein               319 non-null    float64
 11  Dietary Fiber         319 non-null    float64
 12  Cholesterol           319 non-null    float64
 13  Sodium                319 non-null    float64
 14  Water                 319 non-null    float64
 15  Vitamin A             319 non-null    float64
 16  Vitamin B1            319 non-null    float64
 17  Vitamin B11           319 non-null    float64
 18  Vitamin B12           319 non-null    float64
 19  Vitamin B2            319 non-null    float64
 20  Vitamin B3            319 non-null    float64
 21  Vitamin B5            319 non-null    float64
 22  Vitamin B6            319 non-null    float64
 23  Vitamin C             319 non-null    float64
 24  Vitamin D             319 non-null    float64
 25  Vitamin E             319 non-null    float64
 26  Vitamin K             319 non-null    float64
 27  Calcium               319 non-null    float64
 28  Copper                319 non-null    float64
 29  Iron                  319 non-null    float64
 30  Magnesium             319 non-null    float64
 31  Manganese             319 non-null    float64
 32  Phosphorus            319 non-null    float64
 33  Potassium             319 non-null    float64
 34  Selenium              319 non-null    float64
 35  Zinc                  319 non-null    float64
 36  Nutrition Density     319 non-null    float64
dtypes: float64(33), int64(3), object(1)
memory usage: 92.3+ KB
```

```
FOOD3.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 571 entries, 0 to 570
Data columns (total 37 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Unnamed: 0.1          571 non-null    int64
 1   Unnamed: 0            571 non-null    int64
 2   food                  571 non-null    object
 3   Caloric Value         571 non-null    int64
 4   Fat                   571 non-null    float64
 5   Saturated Fats        571 non-null    float64
 6   Monounsaturated Fats  571 non-null    float64
 7   Polyunsaturated Fats  571 non-null    float64
 8   Carbohydrates         571 non-null    float64
 9   Sugars                571 non-null    float64
 10  Protein               571 non-null    float64
 11  Dietary Fiber         571 non-null    float64
 12  Cholesterol           571 non-null    float64
 13  Sodium                571 non-null    float64
 14  Water                 571 non-null    float64
 15  Vitamin A             571 non-null    float64
 16  Vitamin B1            571 non-null    float64
 17  Vitamin B11           571 non-null    float64
 18  Vitamin B12           571 non-null    float64
```

```
 19  Vitamin B2            571 non-null    float64
 20  Vitamin B3            571 non-null    float64
 21  Vitamin B5            571 non-null    float64
 22  Vitamin B6            571 non-null    float64
 23  Vitamin C             571 non-null    float64
 24  Vitamin D             571 non-null    float64
 25  Vitamin E             571 non-null    float64
 26  Vitamin K             571 non-null    float64
 27  Calcium               571 non-null    float64
 28  Copper                571 non-null    float64
 29  Iron                  571 non-null    float64
 30  Magnesium             571 non-null    float64
 31  Manganese             571 non-null    float64
 32  Phosphorus            571 non-null    float64
 33  Potassium             571 non-null    float64
 34  Selenium              571 non-null    float64
 35  Zinc                  571 non-null    float64
 36  Nutrition Density     571 non-null    float64
dtypes: float64(33), int64(3), object(1)
memory usage: 165.2+ KB
```

```
FOOD4.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 232 entries, 0 to 231
Data columns (total 37 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Unnamed: 0.1          232 non-null    int64
 1   Unnamed: 0            232 non-null    int64
 2   food                  232 non-null    object
 3   Caloric Value         232 non-null    int64
 4   Fat                   232 non-null    float64
 5   Saturated Fats        232 non-null    float64
 6   Monounsaturated Fats  232 non-null    float64
 7   Polyunsaturated Fats  232 non-null    float64
 8   Carbohydrates         232 non-null    float64
 9   Sugars                232 non-null    float64
 10  Protein               232 non-null    float64
 11  Dietary Fiber         232 non-null    float64
 12  Cholesterol           232 non-null    float64
 13  Sodium                232 non-null    float64
 14  Water                 232 non-null    float64
 15  Vitamin A             232 non-null    float64
 16  Vitamin B1            232 non-null    float64
 17  Vitamin B11           232 non-null    float64
 18  Vitamin B12           232 non-null    float64
 19  Vitamin B2            232 non-null    float64
 20  Vitamin B3            232 non-null    float64
 21  Vitamin B5            232 non-null    float64
 22  Vitamin B6            232 non-null    float64
 23  Vitamin C             232 non-null    float64
 24  Vitamin D             232 non-null    float64
 25  Vitamin E             232 non-null    float64
 26  Vitamin K             232 non-null    float64
 27  Calcium               232 non-null    float64
 28  Copper                232 non-null    float64
 29  Iron                  232 non-null    float64
 30  Magnesium             232 non-null    float64
 31  Manganese             232 non-null    float64
 32  Phosphorus            232 non-null    float64
 33  Potassium             232 non-null    float64
 34  Selenium              232 non-null    float64
 35  Zinc                  232 non-null    float64
 36  Nutrition Density     232 non-null    float64
dtypes: float64(33), int64(3), object(1)
memory usage: 67.2+ KB
```

```
FOOD5.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 722 entries, 0 to 721
Data columns (total 37 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Unnamed: 0.1          722 non-null    int64
 1   Unnamed: 0            722 non-null    int64
 2   food                  722 non-null    object
 3   Caloric Value         722 non-null    int64
 4   Fat                   722 non-null    float64
 5   Saturated Fats        722 non-null    float64
 6   Monounsaturated Fats  722 non-null    float64
 7   Polyunsaturated Fats  722 non-null    float64
 8   Carbohydrates         722 non-null    float64
 9   Sugars                722 non-null    float64
 10  Protein               722 non-null    float64
 11  Dietary Fiber         722 non-null    float64
 12  Cholesterol           722 non-null    float64
```

```
 13  Sodium                722 non-null    float64
 14  Water                 722 non-null    float64
 15  Vitamin A             722 non-null    float64
 16  Vitamin B1            722 non-null    float64
 17  Vitamin B11           722 non-null    float64
 18  Vitamin B12           722 non-null    float64
 19  Vitamin B2            722 non-null    float64
 20  Vitamin B3            722 non-null    float64
 21  Vitamin B5            722 non-null    float64
 22  Vitamin B6            722 non-null    float64
 23  Vitamin C             722 non-null    float64
 24  Vitamin D             722 non-null    float64
 25  Vitamin E             722 non-null    float64
 26  Vitamin K             722 non-null    float64
 27  Calcium               722 non-null    float64
 28  Copper                722 non-null    float64
 29  Iron                  722 non-null    float64
 30  Magnesium             722 non-null    float64
 31  Manganese             722 non-null    float64
 32  Phosphorus            722 non-null    float64
 33  Potassium             722 non-null    float64
 34  Selenium              722 non-null    float64
 35  Zinc                  722 non-null    float64
 36  Nutrition Density     722 non-null    float64
dtypes: float64(33), int64(3), object(1)
memory usage: 208.8+ KB
```

```
# Merge the tables

merged_data = pd.concat([FOOD1, FOOD2, FOOD3, FOOD4, FOOD5], ignore_index=True)
```

```
merged_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2395 entries, 0 to 2394
Data columns (total 37 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Unnamed: 0.1          2395 non-null   int64
 1   Unnamed: 0            2395 non-null   int64
 2   food                  2395 non-null   object
 3   Caloric Value         2395 non-null   int64
 4   Fat                   2395 non-null   float64
 5   Saturated Fats        2395 non-null   float64
 6   Monounsaturated Fats  2395 non-null   float64
 7   Polyunsaturated Fats  2395 non-null   float64
 8   Carbohydrates         2395 non-null   float64
 9   Sugars                2395 non-null   float64
 10  Protein               2395 non-null   float64
 11  Dietary Fiber         2395 non-null   float64
 12  Cholesterol           2395 non-null   float64
 13  Sodium                2395 non-null   float64
 14  Water                 2395 non-null   float64
 15  Vitamin A             2395 non-null   float64
 16  Vitamin B1            2395 non-null   float64
 17  Vitamin B11           2395 non-null   float64
 18  Vitamin B12           2395 non-null   float64
 19  Vitamin B2            2395 non-null   float64
 20  Vitamin B3            2395 non-null   float64
 21  Vitamin B5            2395 non-null   float64
 22  Vitamin B6            2395 non-null   float64
 23  Vitamin C             2395 non-null   float64
 24  Vitamin D             2395 non-null   float64
 25  Vitamin E             2395 non-null   float64
 26  Vitamin K             2395 non-null   float64
 27  Calcium               2395 non-null   float64
 28  Copper                2395 non-null   float64
 29  Iron                  2395 non-null   float64
 30  Magnesium             2395 non-null   float64
 31  Manganese             2395 non-null   float64
 32  Phosphorus            2395 non-null   float64
 33  Potassium             2395 non-null   float64
 34  Selenium              2395 non-null   float64
 35  Zinc                  2395 non-null   float64
 36  Nutrition Density     2395 non-null   float64
dtypes: float64(33), int64(3), object(1)
memory usage: 692.4+ KB
```

```
# Drop `Unnamed: 0.1`, `Unnamed: 0`

merged_data = merged_data.drop(['Unnamed: 0.1', 'Unnamed: 0'], axis=1)
```

```
# Count null values

merged_data.head()
```

	food	Caloric Value	Fat	Saturated Fats	Monounsaturated Fats	Polyunsaturated Fats	Carbohydrates	Sugars	Protein	Dietary Fiber	...	Calcium	Copp
0	cream cheese	51	5.0	2.9	1.3	0.200	0.8	0.500	0.9	0.0	...	0.008	14.1
1	neufchatel cheese	215	19.4	10.9	4.9	0.800	3.1	2.700	7.8	0.0	...	99.500	0.0
2	requeijao cremoso light catupiry	49	3.6	2.3	0.9	0.000	0.9	3.400	0.8	0.1	...	0.000	0.0
3	ricotta cheese	30	2.0	1.3	0.5	0.002	1.5	0.091	1.5	0.0	...	0.097	41.2
4	cream cheese low fat	30	2.3	1.4	0.6	0.042	1.2	0.900	1.2	0.0	...	22.200	0.0

5 rows × 35 columns

```
# Check if there any duplicated values

merged_data.duplicated().sum()
```

```
0
```

No duplicate data and missing values, GREAT! Now our data is ready for deeper analysis.

```
#Let's check the data once more

merged_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2395 entries, 0 to 2394
Data columns (total 35 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   food                  2395 non-null   object
 1   Caloric Value         2395 non-null   int64
 2   Fat                   2395 non-null   float64
 3   Saturated Fats        2395 non-null   float64
 4   Monounsaturated Fats  2395 non-null   float64
 5   Polyunsaturated Fats  2395 non-null   float64
 6   Carbohydrates         2395 non-null   float64
 7   Sugars                2395 non-null   float64
 8   Protein               2395 non-null   float64
 9   Dietary Fiber         2395 non-null   float64
 10  Cholesterol           2395 non-null   float64
 11  Sodium                2395 non-null   float64
 12  Water                 2395 non-null   float64
 13  Vitamin A             2395 non-null   float64
 14  Vitamin B1            2395 non-null   float64
 15  Vitamin B11           2395 non-null   float64
 16  Vitamin B12           2395 non-null   float64
 17  Vitamin B2            2395 non-null   float64
 18  Vitamin B3            2395 non-null   float64
 19  Vitamin B5            2395 non-null   float64
 20  Vitamin B6            2395 non-null   float64
 21  Vitamin C             2395 non-null   float64
 22  Vitamin D             2395 non-null   float64
 23  Vitamin E             2395 non-null   float64
 24  Vitamin K             2395 non-null   float64
 25  Calcium               2395 non-null   float64
 26  Copper                2395 non-null   float64
 27  Iron                  2395 non-null   float64
 28  Magnesium             2395 non-null   float64
 29  Manganese             2395 non-null   float64
 30  Phosphorus            2395 non-null   float64
 31  Potassium             2395 non-null   float64
 32  Selenium              2395 non-null   float64
 33  Zinc                  2395 non-null   float64
 34  Nutrition Density     2395 non-null   float64
dtypes: float64(33), int64(1), object(1)
memory usage: 655.0+ KB
```

CORRELATION AND DESCRIPTIVE ANALYSIS

```
merged_data.head()
```

	food	Caloric Value	Fat	Saturated Fats	Monounsaturated Fats	Polyunsaturated Fats	Carbohydrates	Sugars	Protein	Dietary Fiber	...	Calcium	Copp
0	cream cheese	51	5.0	2.9	1.3	0.200	0.8	0.500	0.9	0.0	...	0.008	14.1
1	neufchatel cheese	215	19.4	10.9	4.9	0.800	3.1	2.700	7.8	0.0	...	99.500	0.0
2	requeijao cremoso light catupiry	49	3.6	2.3	0.9	0.000	0.9	3.400	0.8	0.1	...	0.000	0.0
3	ricotta cheese	30	2.0	1.3	0.5	0.002	1.5	0.091	1.5	0.0	...	0.097	41.2
4	cream cheese low fat	30	2.3	1.4	0.6	0.042	1.2	0.900	1.2	0.0	...	22.200	0.0

5 rows × 35 columns

```
# Let's do Descriptive Statistics with the dataframe

merged_data.describe()
```

	Caloric Value	Fat	Saturated Fats	Monounsaturated Fats	Polyunsaturated Fats	Carbohydrates	Sugars	Protein	Dietary Fiber
count	2395.000000	2395.000000	2395.000000	2395.000000	2395.000000	2395.000000	2395.000000	2395.000000	2395.000000
mean	223.769520	10.176276	3.924917	4.133622	2.152844	18.589021	4.457459	13.400777	2.235790
std	384.728244	29.008915	19.502262	12.939587	7.145738	29.406134	13.339929	32.294246	5.404483
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	44.500000	0.300000	0.064000	0.058000	0.071000	0.500000	0.000000	0.800000	0.000000
50%	117.000000	2.100000	0.500000	0.500000	0.400000	6.800000	0.086000	3.500000	0.200000
75%	258.000000	9.400000	2.700000	3.400000	1.700000	25.050000	3.200000	13.300000	2.200000
max	6077.000000	550.700000	672.000000	291.100000	188.000000	390.200000	291.500000	560.300000	76.500000

8 rows × 34 columns

The high **standard deviation** values for many components (like calories, fat, and sugars) indicate that there is significant **variability in the nutritional content** of the food items in this dataset. This might be due to the **diverse range of foods included**. A significant gap between the 25th and 75th percentiles suggests a wide range of values.

```
# Filter only numerical columns

numerical_data = merged_data.select_dtypes(include=['float64', 'int64'])
```

```
# correlation heatmap

numerical_data.corr()
```

	Caloric Value	Fat	Saturated Fats	Monounsaturated Fats	Polyunsaturated Fats	Carbohydrates	Sugars	Protein	Dietary Fiber
Caloric Value	1.000000	0.901783	0.606614	0.845348	0.603871	0.297667	0.118609	0.748770	0.152123
Fat	0.901783	1.000000	0.551220	0.924344	0.626602	0.026412	0.019477	0.600596	-0.013629
Saturated Fats	0.606614	0.551220	1.000000	0.550169	0.328653	-0.027468	-0.009464	0.269882	-0.029532
Monounsaturated Fats	0.845348	0.924344	0.550169	1.000000	0.579191	-0.008268	0.022631	0.547464	-0.039129
Polyunsaturated Fats	0.603871	0.626602	0.328653	0.579191	1.000000	0.100891	0.066817	0.383249	0.034792
Carbohydrates	0.297667	0.026412	-0.027468	-0.008268	0.100891	1.000000	0.441794	-0.020191	0.522200
Sugars	0.118609	0.019477	-0.009464	0.022631	0.066817	0.441794	1.000000	-0.070012	0.082745
Protein	0.748770	0.600596	0.269882	0.547464	0.383249	-0.020191	-0.070012	1.000000	0.031205
Dietary Fiber	0.152123	-0.013629	-0.029532	-0.039129	0.034792	0.522200	0.082745	0.031205	1.000000
Cholesterol	0.269212	0.252769	0.112380	0.208068	0.144282	-0.068233	-0.035014	0.332728	-0.059909
Sodium	0.144128	0.127233	0.041113	0.132466	0.080922	0.095868	0.012686	0.090717	0.008106
Water	0.534724	0.422762	0.176465	0.376780	0.235754	0.026109	0.107630	0.684874	0.016601
Vitamin A	0.012179	-0.018181	-0.010742	-0.017392	-0.017198	0.001172	-0.016470	-0.015306	0.282110
Vitamin B1	0.391420	0.312557	0.133583	0.277243	0.223690	0.092383	-0.025965	0.436501	0.103537
Vitamin B11	0.008006	-0.000736	-0.002737	-0.002300	-0.000865	0.020390	0.010045	0.006482	0.021311
Vitamin B12	-0.002386	-0.000309	-0.002675	-0.003583	-0.007175	-0.019097	-0.009987	0.009545	-0.009463
Vitamin B2	0.305870	0.256995	0.116858	0.231726	0.157622	0.019674	0.003963	0.357212	-0.001940
Vitamin B3	0.693851	0.565265	0.246968	0.510710	0.371661	0.009399	-0.053692	0.876010	-0.000610
Vitamin B5	0.467535	0.386485	0.177847	0.361627	0.241877	0.016951	-0.023900	0.574255	0.037728
Vitamin B6	0.614840	0.490550	0.209555	0.441232	0.312299	0.015895	-0.037404	0.794186	0.061392
Vitamin C	-0.002313	-0.003270	-0.005958	-0.005336	-0.003573	0.010369	0.030435	-0.004345	0.026312
Vitamin D	-0.059747	-0.052703	-0.032253	-0.050024	-0.044386	0.003576	0.053124	-0.056172	0.052763
Vitamin E	0.270989	0.297108	0.156938	0.218090	0.255696	0.110054	0.143855	0.129316	0.085738
Vitamin K	-0.006541	-0.006595	-0.005200	0.016400	-0.003951	-0.002969	-0.001311	-0.005159	0.022527
Calcium	0.265974	0.198221	0.097436	0.148955	0.131101	0.188805	0.050227	0.240558	0.170630
Copper	0.025887	0.008352	0.001054	0.000493	0.001014	0.074046	0.010991	0.018976	0.200339
Iron	0.373881	0.274335	0.113436	0.249328	0.177821	0.166013	-0.009551	0.410093	0.190030
Magnesium	0.474511	0.310466	0.104273	0.259127	0.297089	0.361753	-0.004080	0.488868	0.469852
Manganese	0.057497	0.012966	-0.002582	0.007921	0.001212	0.101102	-0.001622	0.076323	0.227270
Phosphorus	0.735810	0.593405	0.256100	0.530300	0.406182	0.104912	-0.042079	0.873285	0.141830
Potassium	0.681601	0.503715	0.214213	0.447152	0.345841	0.245100	0.019411	0.780445	0.333958
Selenium	0.067144	0.030169	0.007729	0.027909	0.000388	0.048588	0.004823	0.109278	0.152436
Zinc	0.534415	0.452758	0.215322	0.413799	0.262224	0.015785	-0.050331	0.639852	0.062026
Nutrition Density	0.535323	0.422081	0.202325	0.358096	0.285149	0.323416	0.114739	0.455231	0.274237

34 rows × 34 columns

```
# Let's foccused on top 10 most correlated column

corr_matrix = numerical_data.corr().abs().unstack().sort_values(ascending=False).drop_duplicates()
corr_matrix.head(10)
```

		0
Caloric Value	**Caloric Value**	1.000000
Monounsaturated Fats	**Fat**	0.924344
Caloric Value	**Fat**	0.901783
Protein	**Vitamin B3**	0.876010
	Phosphorus	0.873285
Phosphorus	**Vitamin B3**	0.859884
Manganese	**Selenium**	0.848944
Caloric Value	**Monounsaturated Fats**	0.845348
Potassium	**Phosphorus**	0.826714
Vitamin B2	**Vitamin B12**	0.817085

dtype: float64

Due to the large number of columns, it's more efficient to focus on the most correlated ones. Both Monounsaturated Fats and Caloric Value show a strong correlation with Fat, while Protein and Manganese are highly correlated with Vitamin B3. By understanding these correlations, we can prioritize the columns or variables that are most relevant for analysis.

FEATURE ENGINEERING

```
merged_data.info()
```

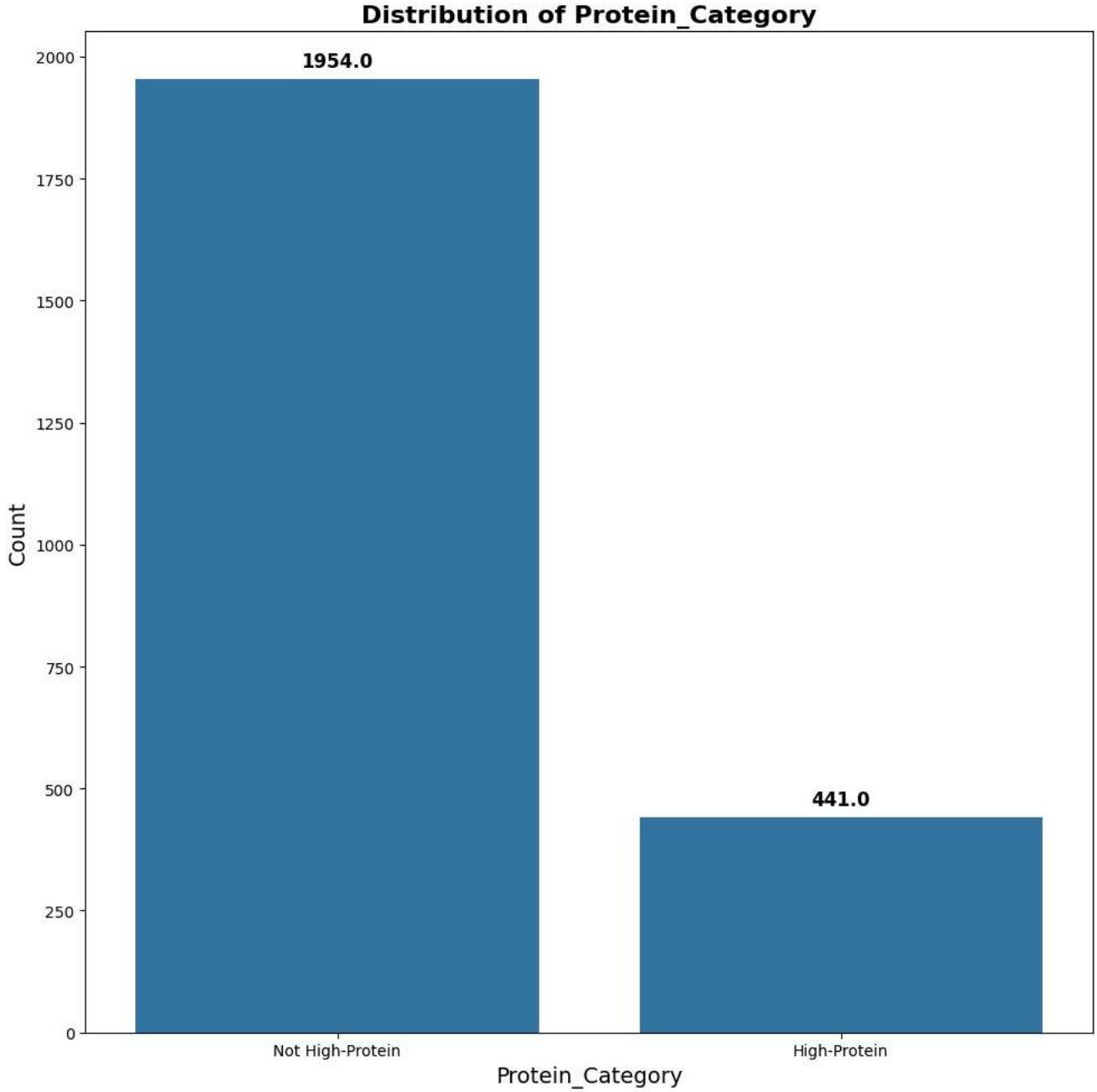
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2395 entries, 0 to 2394
Data columns (total 35 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   food                  2395 non-null   object
 1   Caloric Value         2395 non-null   int64
 2   Fat                   2395 non-null   float64
 3   Saturated Fats        2395 non-null   float64
 4   Monounsaturated Fats  2395 non-null   float64
 5   Polyunsaturated Fats  2395 non-null   float64
 6   Carbohydrates         2395 non-null   float64
 7   Sugars                2395 non-null   float64
 8   Protein               2395 non-null   float64
 9   Dietary Fiber         2395 non-null   float64
 10  Cholesterol           2395 non-null   float64
 11  Sodium                2395 non-null   float64
 12  Water                 2395 non-null   float64
 13  Vitamin A             2395 non-null   float64
 14  Vitamin B1            2395 non-null   float64
 15  Vitamin B11           2395 non-null   float64
 16  Vitamin B12           2395 non-null   float64
 17  Vitamin B2            2395 non-null   float64
 18  Vitamin B3            2395 non-null   float64
 19  Vitamin B5            2395 non-null   float64
 20  Vitamin B6            2395 non-null   float64
 21  Vitamin C             2395 non-null   float64
 22  Vitamin D             2395 non-null   float64
 23  Vitamin E             2395 non-null   float64
 24  Vitamin K             2395 non-null   float64
 25  Calcium               2395 non-null   float64
 26  Copper                2395 non-null   float64
 27  Iron                  2395 non-null   float64
 28  Magnesium             2395 non-null   float64
 29  Manganese             2395 non-null   float64
 30  Phosphorus            2395 non-null   float64
 31  Potassium             2395 non-null   float64
 32  Selenium              2395 non-null   float64
 33  Zinc                  2395 non-null   float64
 34  Nutrition Density     2395 non-null   float64
dtypes: float64(33), int64(1), object(1)
memory usage: 655.0+ KB
```

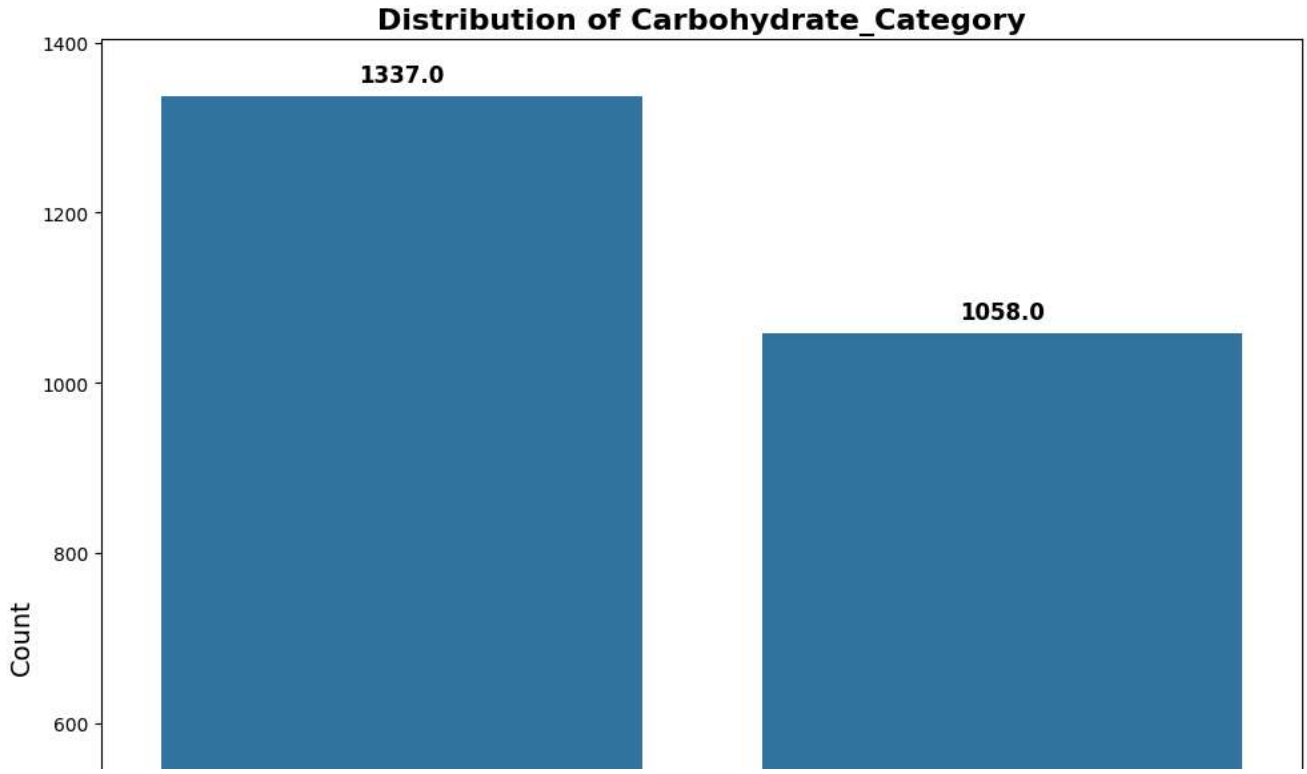
```
# Define functions to categorize food items
def categorize_protein(row):
    if row['Protein'] > 20:
        return 'High-Protein'
    else:
        return 'Not High-Protein'

def categorize_carb(row):
    if row['Carbohydrates'] < 10:
        return 'Low-Carb'
    else:
        return 'Not Low-Carb'

def categorize_fat(row):
    if row['Fat'] < 5:
        return 'Low-Fat'
    else:
        return 'Not Low-Fat'

def categorize_fiber(row):
    if row['Dietary Fiber'] > 5:
        return 'High-Fiber'
    else:
        return 'Not High-Fiber'

def categorize_sugar(row):
    if row['Sugars'] < 5:
        return 'Low-Sugar'
    else:
        return 'Not Low-Sugar'

def categorize_cal(row):
    if row['Caloric Value'] < 100:
        return 'Low-Cal'
    else:
        return 'Not Low-Cal'

def categorize_chol(row):
    if row['Cholesterol'] < 20:
        return 'Low-Chol'
    else:
        return 'Not Low-Chol'

def categorize_sodium(row):
    if row['Sodium'] < 10:
        return 'Low-Sodium'
    else:
        return 'Not Low-Sodium'

# Apply the functions to the DataFrame
merged_data['Protein_Category'] = merged_data.apply(categorize_protein, axis=1)
merged_data['Carbohydrate_Category'] = merged_data.apply(categorize_carb, axis=1)
merged_data['Fat_Category'] = merged_data.apply(categorize_fat, axis=1)
merged_data['Fiber_Category'] = merged_data.apply(categorize_fiber, axis=1)
merged_data['Sugar_Category'] = merged_data.apply(categorize_sugar, axis=1)
merged_data['Calorie_Category'] = merged_data.apply(categorize_cal, axis=1)
merged_data['Carbohydrate_Category'] = merged_data.apply(categorize_carb, axis=1)
merged_data['Cholesterol_Category'] = merged_data.apply(categorize_chol, axis=1)
merged_data['Sodium_Category'] = merged_data.apply(categorize_sodium, axis=1)

# Display the updated DataFrame
merged_data.head()
```

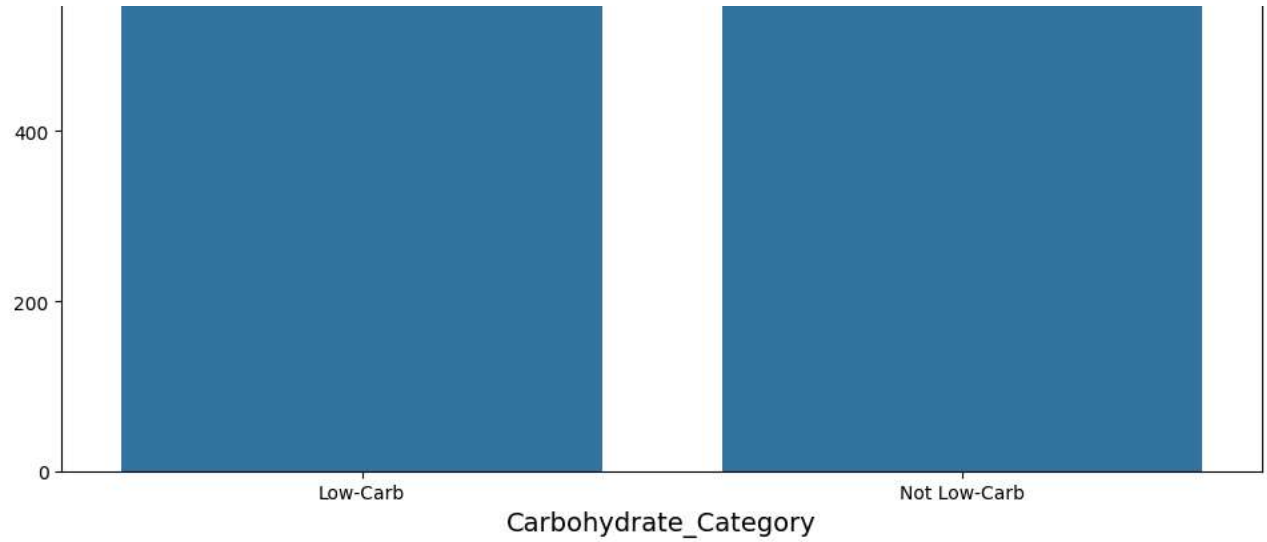
	food	Caloric Value	Fat	Saturated Fats	Monounsaturated Fats	Polyunsaturated Fats	Carbohydrates	Sugars	Protein	Dietary Fiber	...	Zinc	Nutriti Densi
0	cream cheese	51	5.0	2.9	1.3	0.200	0.8	0.500	0.9	0.0	...	0.039	7.0
1	neufchatel cheese	215	19.4	10.9	4.9	0.800	3.1	2.700	7.8	0.0	...	0.700	130.1
2	requeijao cremoso light catupiry	49	3.6	2.3	0.9	0.000	0.9	3.400	0.8	0.1	...	0.000	5.4
3	ricotta cheese	30	2.0	1.3	0.5	0.002	1.5	0.091	1.5	0.0	...	0.035	5.1
4	cream cheese low fat	30	2.3	1.4	0.6	0.042	1.2	0.900	1.2	0.0	...	0.053	27.0

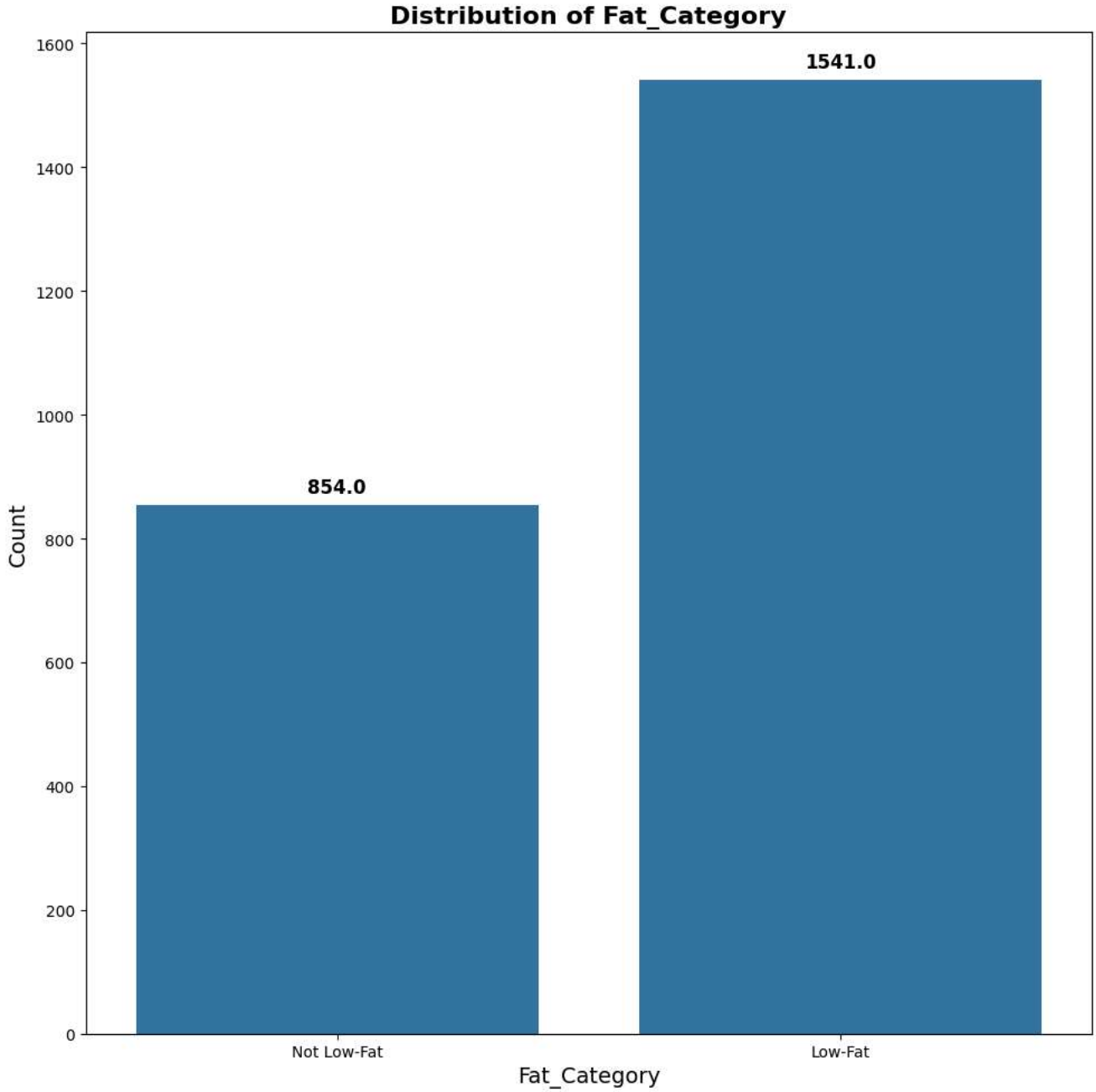
5 rows × 43 columns

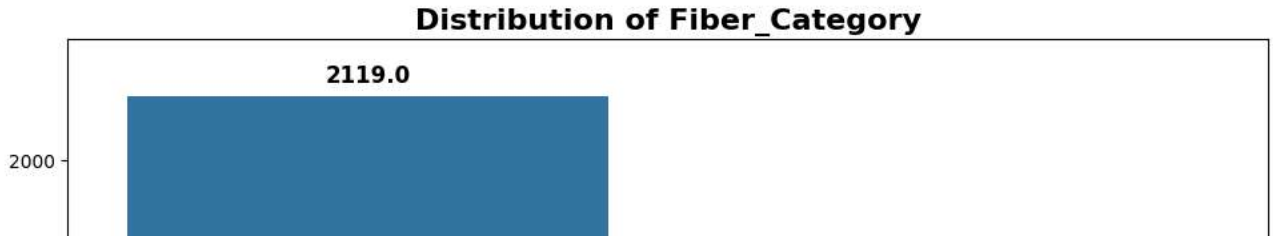
```
# Visualiza the categories

categories = [
    'Protein_Category', 'Carbohydrate_Category', 'Fat_Category',
    'Fiber_Category', 'Sugar_Category', 'Calorie_Category',
    'Cholesterol_Category', 'Sodium_Category'
]

# Create a plot for each category
for category in categories:
    plt.figure(figsize=(10, 10))

    # Create the count plot
    ax = sns.countplot(data=merged_data, x=category)

    # Enhance the plot aesthetics
    plt.title(f'Distribution of {category}', fontsize=16, weight='bold')
    plt.xlabel(category, fontsize=14)
    plt.ylabel('Count', fontsize=14)

    # Annotate bars with counts
    for p in ax.patches:
        height = p.get_height()
        ax.annotate(f'{height}',
                    xy=(p.get_x() + p.get_width() / 2, height),
                    xytext=(0, 5),  # 5 points vertical offset
                    textcoords='offset points',
                    ha='center', va='bottom',
                    fontsize=12, color='black', weight='bold')

    # Adjust layout to prevent clipping
    plt.tight_layout()

    # Show the plot
    plt.show()
```

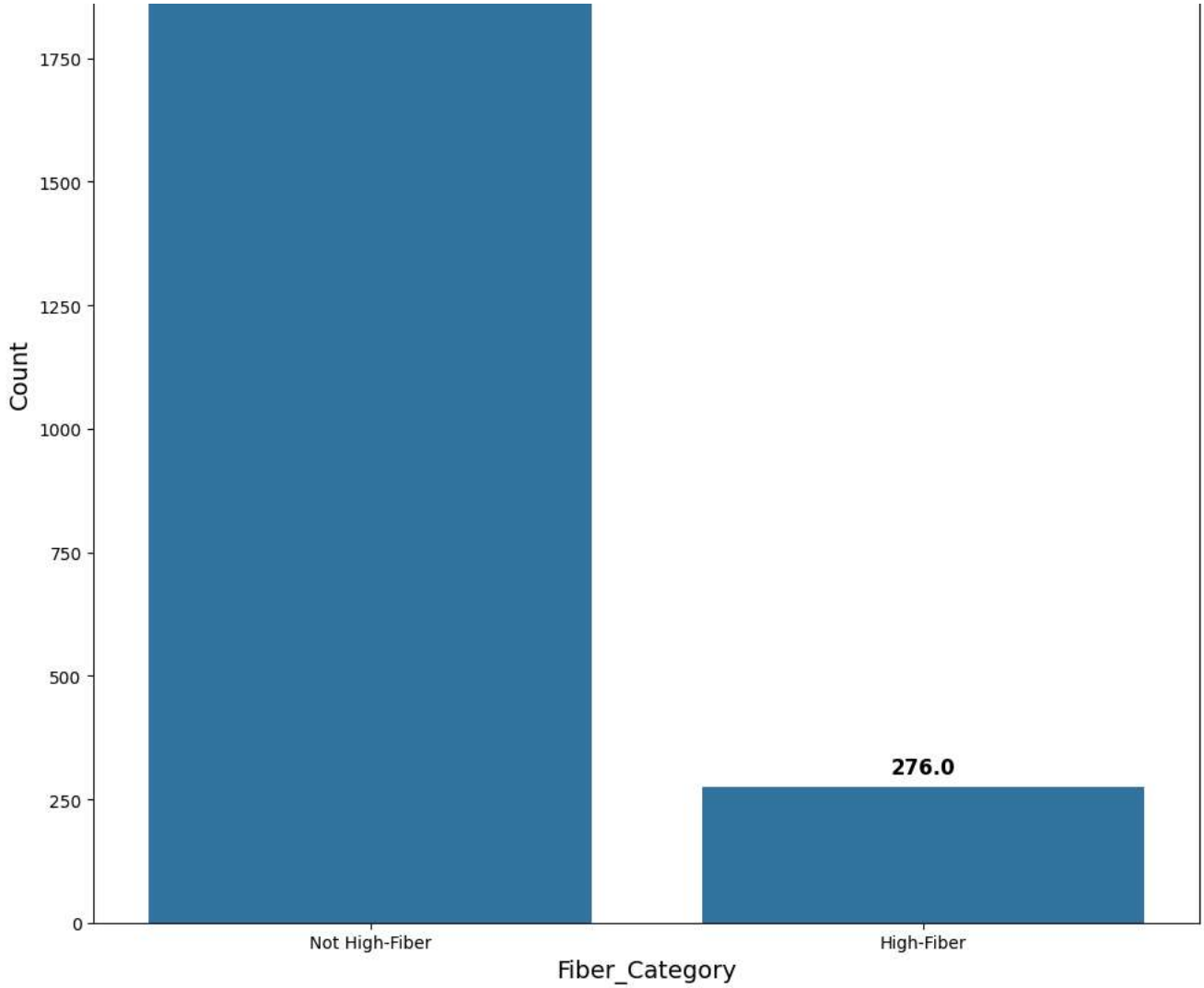

Distribution of Protein_Category
1954.0
441.0
Count
Protein_Category
Not High-Protein
High-Protein
0
250
500
750
1000
1250
1500
1750
2000

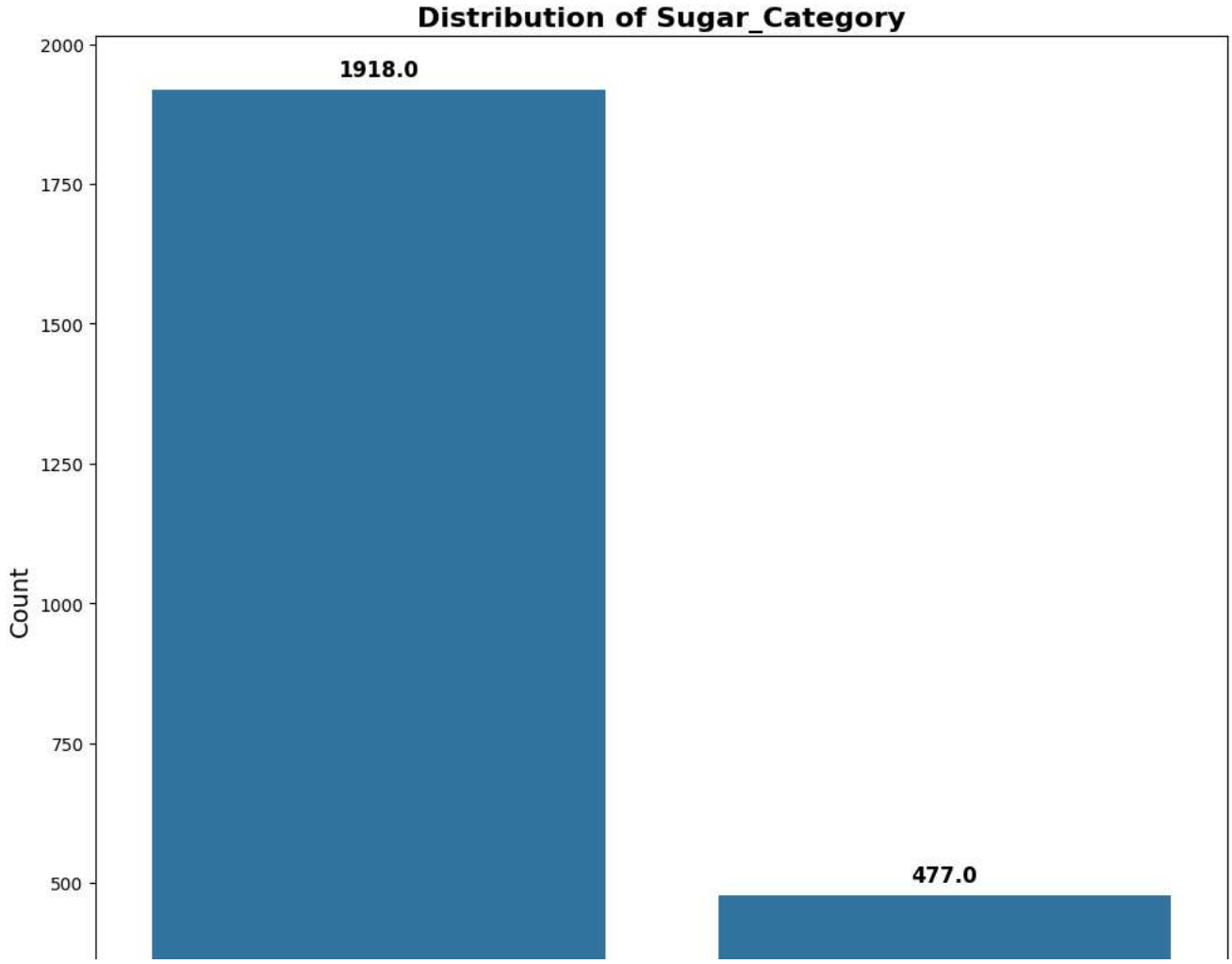

Distribution of Carbohydrate_Category
1337.0
1058.0
Count
600
800
1000
1200
1400


400
200
0
Low-Carb
Not Low-Carb
Carbohydrate_Category


Distribution of Fat_Category
1600
1400
1200
1000
800
600
400
200
0
Count
854.0
1541.0
Not Low-Fat
Low-Fat
Fat_Category


Distribution of Fiber_Category
2119.0
2000


1750
1500
1250
1000
750
500
250
0
Count
276.0
Not High-Fiber
High-Fiber
Fiber_Category


Distribution of Sugar_Category
2000
1750
1500
1250
1000
750
500
Count
1918.0
477.0