

Food Nutrient Analysis

Exploratory Data Analysis on Food Nutrient Data By Prerna Banthiya

```
In [1]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
In [4]: data=pd.read_csv("/content/food data csv.csv")
data
```

Out[4]:

	ID	Description	Calories	Protein	TotalFat	Carbohydrate	Sodium	Sa
0	1001	BUTTER,WITH SALT	717.0	0.85	81.11		0.06	714.0
1	1002	BUTTER,WHIPPED,WITH SALT	717.0	0.85	81.11		0.06	827.0
2	1003	BUTTER OIL,ANHYDROUS	876.0	0.28	99.48		0.00	2.0
3	1004	CHEESE,BLUE	353.0	21.40	28.74		2.34	1395.0
4	1005	CHEESE,BRICK	371.0	23.24	29.68		2.79	560.0
...
7053	80200	FROG LEGS,RAW	73.0	16.40	0.30		0.00	58.0
7054	83110	MACKEREL,SALTED	305.0	18.50	25.10		0.00	4450.0
7055	90240	SCALLOP, (BAY&SEA),CKD,STMD	111.0	20.54	0.84		5.41	667.0
7056	90560	SNAIL,RAW	90.0	16.10	1.40		2.00	70.0
7057	93600	TURTLE,GREEN,RAW	89.0	19.80	0.50		0.00	68.0

7058 rows × 16 columns



EXPLORATORY DATA ANALYSIS

```
In [ ]: data.head(10)
```

	ID	Description	Calories	Protein	TotalFat	Carbohydrate	Sodium	Saturat
0	1001	BUTTER,WITH SALT	717.0	0.85	81.11	0.06	714.0	5
1	1002	BUTTER,WHIPPED,WITH SALT	717.0	0.85	81.11	0.06	827.0	5
2	1003	BUTTER OIL,ANHYDROUS	876.0	0.28	99.48	0.00	2.0	6
3	1004	CHEESE,BLUE	353.0	21.40	28.74	2.34	1395.0	7
4	1005	CHEESE,BRICK	371.0	23.24	29.68	2.79	560.0	7
5	1006	CHEESE,BRIE	334.0	20.75	27.68	0.45	629.0	7
6	1007	CHEESE,CAMEMBERT	300.0	19.80	24.26	0.46	842.0	7
7	1008	CHEESE,CARAWAY	376.0	25.18	29.20	3.06	690.0	7
8	1009	CHEESE,CHEDDAR	403.0	24.90	33.14	1.28	621.0	2
9	1010	CHEESE,CHESHIRE	387.0	23.37	30.60	4.78	700.0	7



In []: `data.tail(10)`

Out[]:

	ID	Description	Calories	Protein	TotalFat	Carbohydrate	Sodium	Se
7048	44203	COCKTAIL MIX,NON-ALCOHOLIC,CONCD,FRZ	287.0	0.08	0.01	71.60	0.0	
7049	44258	PUDDINGS,CHOC FLAVOR,LO CAL,REG,DRY MIX	365.0	10.08	3.00	74.42	3326.0	
7050	44259	PUDDINGS,ALL FLAVORS XCPT CHOC,LO CAL,REG,DRY MIX	351.0	1.60	0.10	86.04	1765.0	
7051	44260	PUDDINGS,ALL FLAVORS XCPT CHOC,LO CAL,INST,DRY...	350.0	0.81	0.90	84.66	4250.0	
7052	48052	VITAL WHEAT GLUTEN	370.0	75.16	1.85	13.79	29.0	
7053	80200	FROG LEGS,RAW	73.0	16.40	0.30	0.00	58.0	
7054	83110	MACKEREL,SALTED	305.0	18.50	25.10	0.00	4450.0	
7055	90240	SCALLOP, (BAY&SEA),CKD,STMD	111.0	20.54	0.84	5.41	667.0	
7056	90560	SNAIL,RAW	90.0	16.10	1.40	2.00	70.0	
7057	93600	TURTLE,GREEN,RAW	89.0	19.80	0.50	0.00	68.0	



In []:

data[40:50]

Out[]:

	ID	Description	Calories	Protein	TotalFat	Carbohydrate	Sodium
40	1041	CHEESE,TILSIT	340.0	24.41	25.98	1.88	753.0
41	1042	CHEESE,PAST PROCESS,AMERICAN,FORT W/ VITAMIN D	371.0	18.13	31.79	3.70	1671.0
42	1043	CHEESE,PAST PROCESS,PIMENTO	375.0	22.13	31.20	1.73	1428.0
43	1044	CHEESE,PAST PROCESS,SWISS	334.0	24.73	25.01	2.10	1370.0
44	1045	CHEESE FD,COLD PK,AMERICAN	331.0	19.66	24.46	8.32	966.0
45	1046	CHEESE FD,PAST PROCESS,AMERICAN,VITAMIN D FORT	330.0	16.86	25.63	8.56	1284.0
46	1047	CHEESE FD,PAST PROCESS,SWISS	323.0	21.92	24.14	4.50	1552.0
47	1048	CHEESE SPRD,PAST PROCESS,AMERICAN	290.0	16.41	21.23	8.73	1625.0
48	1049	CREAM,FLUID,HALF AND HALF	130.0	2.96	11.50	4.30	41.0
49	1050	CREAM,FLUID,LT (COFFEE CRM OR TABLE CRM)	195.0	2.70	19.31	3.66	40.0



In []:

data.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7058 entries, 0 to 7057
Data columns (total 16 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   ID          7058 non-null    int64  
 1   Description  7058 non-null    object  
 2   Calories     7057 non-null    float64 
 3   Protein      7057 non-null    float64 
 4   TotalFat     7057 non-null    float64 
 5   Carbohydrate 7057 non-null    float64 
 6   Sodium       6974 non-null    float64 
 7   SaturatedFat 6757 non-null    float64 
 8   Cholesterol  6770 non-null    float64 
 9   Sugar        5148 non-null    float64 
 10  Calcium      6922 non-null    float64 
 11  Iron         6935 non-null    float64 
 12  Potassium    6649 non-null    float64 
 13  VitaminC     6726 non-null    float64 
 14  VitaminE     4338 non-null    float64 
 15  VitaminD     4224 non-null    float64 
dtypes: float64(14), int64(1), object(1)
memory usage: 882.4+ KB
```

```
In [ ]: data.dtypes
```

```
Out[ ]: 0
        ID      int64
        Description   object
        Calories    float64
        Protein     float64
        TotalFat    float64
        Carbohydrate float64
        Sodium      float64
        SaturatedFat float64
        Cholesterol  float64
        Sugar       float64
        Calcium     float64
        Iron        float64
        Potassium   float64
        VitaminC    float64
        VitaminE    float64
        VitaminD    float64
```

dtype: object

```
In [ ]: data.columns
```

```
Out[ ]: Index(['ID', 'Description', 'Calories', 'Protein', 'TotalFat', 'Carbohydrate',
       'Sodium', 'SaturatedFat', 'Cholesterol', 'Sugar', 'Calcium', 'Iron',
       'Potassium', 'VitaminC', 'VitaminE', 'VitaminD'],
       dtype='object')
```

```
In [5]: data = data.drop('ID', axis=1)
```

```
In [ ]: data
```

Out[]:

	Description	Calories	Protein	TotalFat	Carbohydrate	Sodium	SaturatedF
0	BUTTER,WITH SALT	717.0	0.85	81.11	0.06	714.0	51.3
1	BUTTER,WHIPPED,WITH SALT	717.0	0.85	81.11	0.06	827.0	50.4
2	BUTTER OIL,ANHYDROUS	876.0	0.28	99.48	0.00	2.0	61.9
3	CHEESE,BLUE	353.0	21.40	28.74	2.34	1395.0	18.6
4	CHEESE,BRICK	371.0	23.24	29.68	2.79	560.0	18.7
...
7053	FROG LEGS,RAW	73.0	16.40	0.30	0.00	58.0	0.0
7054	MACKEREL,SALTED	305.0	18.50	25.10	0.00	4450.0	7.1
7055	SCALLOP, (BAY&SEA),CKD,STMD	111.0	20.54	0.84	5.41	667.0	0.2
7056	SNAIL,RAW	90.0	16.10	1.40	2.00	70.0	0.3
7057	TURTLE,GREEN,RAW	89.0	19.80	0.50	0.00	68.0	0.1

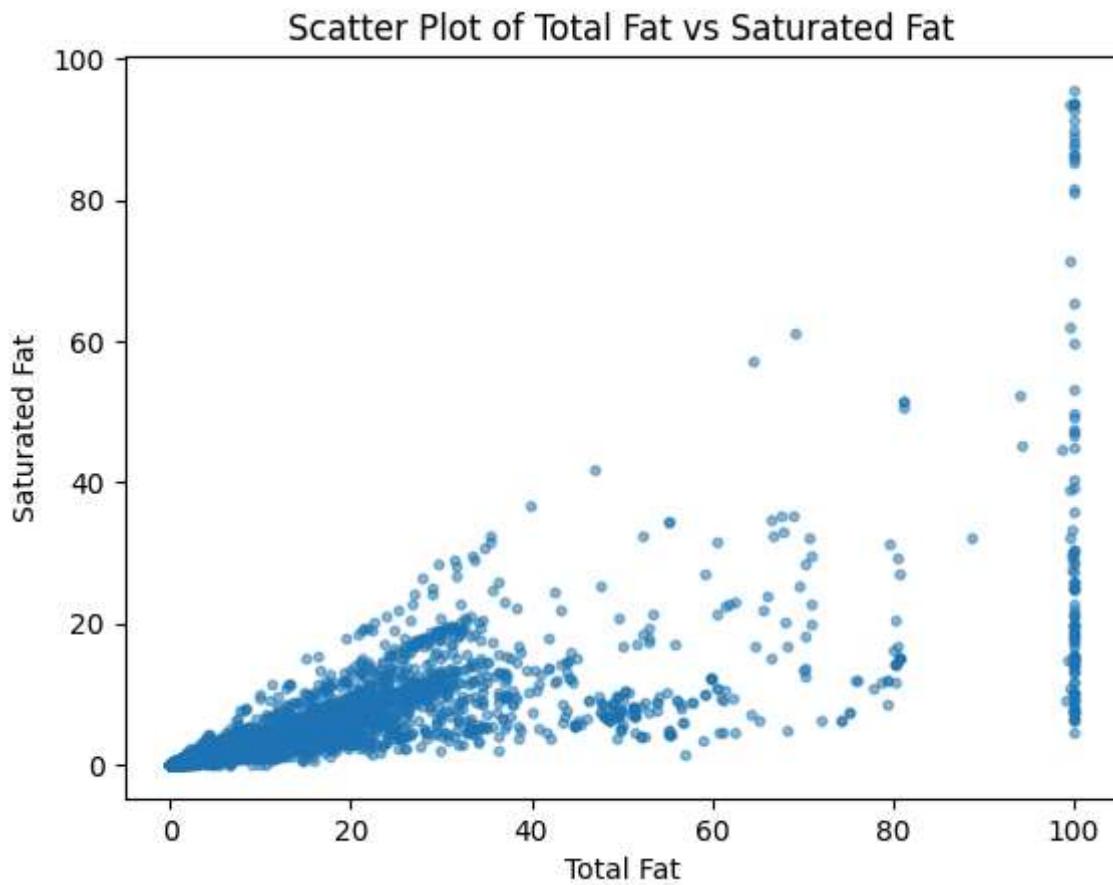
7058 rows × 15 columns



Data Visualization

In []: `plt.scatter(data['TotalFat'], data['SaturatedFat'], s=10, alpha=0.5)
plt.xlabel('Total Fat')
plt.ylabel('Saturated Fat')
plt.title('Scatter Plot of Total Fat vs Saturated Fat')`

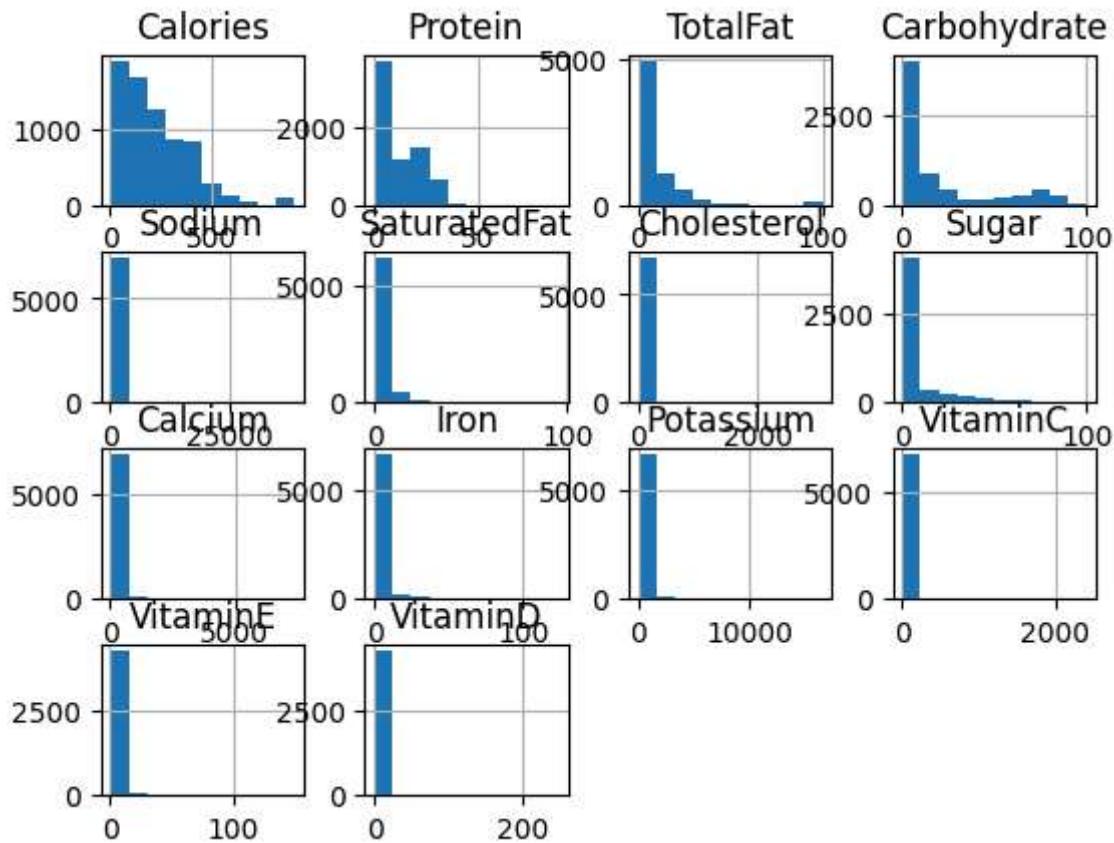
Out[]: Text(0.5, 1.0, 'Scatter Plot of Total Fat vs Saturated Fat')



It shows a positive correlation between Total Fat and Saturated Fat

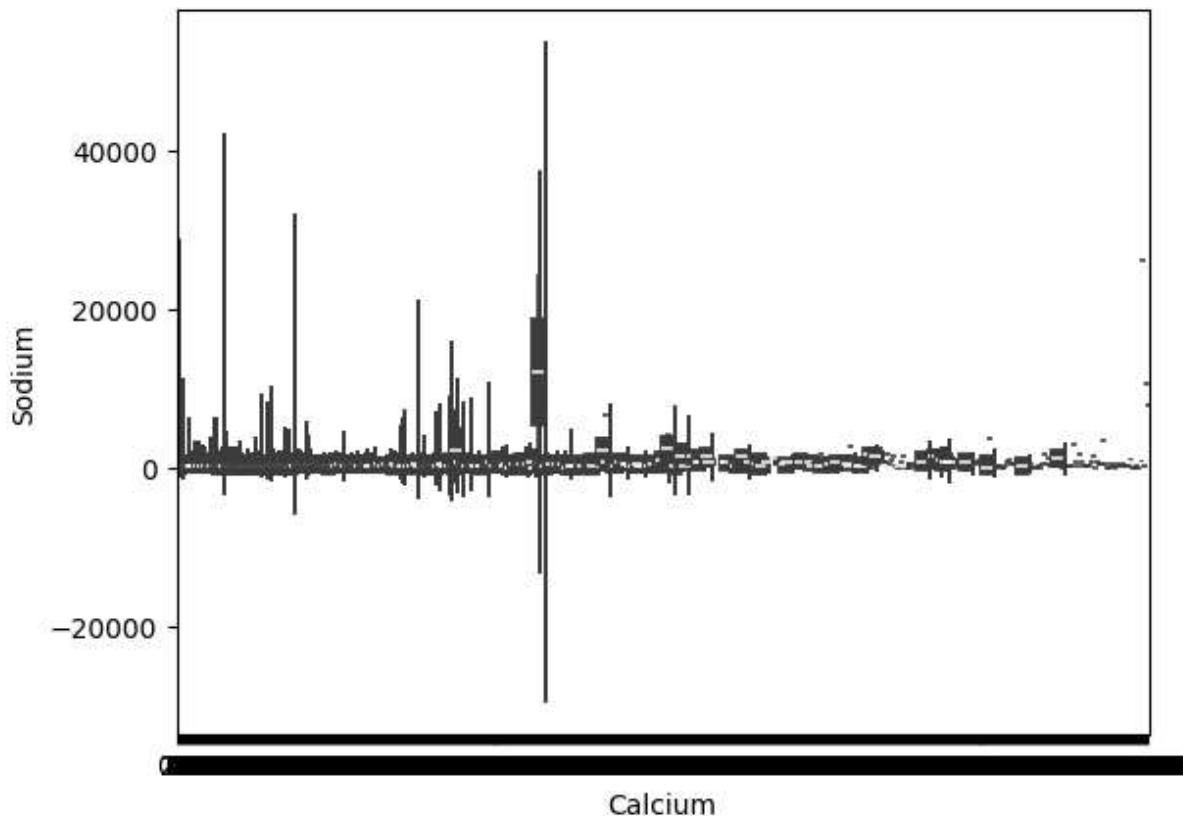
```
In [ ]: plt.figure(figsize=(10, 10))
data.hist()
```

```
Out[ ]: array([[<Axes: title={'center': 'Calories'}>,
   <Axes: title={'center': 'Protein'}>,
   <Axes: title={'center': 'TotalFat'}>,
   <Axes: title={'center': 'Carbohydrate'}>],
  [<Axes: title={'center': 'Sodium'}>,
   <Axes: title={'center': 'SaturatedFat'}>,
   <Axes: title={'center': 'Cholesterol'}>,
   <Axes: title={'center': 'Sugar'}>],
  [<Axes: title={'center': 'Calcium'}>,
   <Axes: title={'center': 'Iron'}>,
   <Axes: title={'center': 'Potassium'}>,
   <Axes: title={'center': 'VitaminC'}>],
  [<Axes: title={'center': 'VitaminE'}>,
   <Axes: title={'center': 'VitaminD'}>, <Axes: >, <Axes: >]],
 dtype=object)
<Figure size 1000x1000 with 0 Axes>
```



```
In [9]: sns.violinplot(x = 'Calcium', y = 'Sodium', data=data,split=True)
```

```
Out[9]: <Axes: xlabel='Calcium', ylabel='Sodium'>
```



```
In [12]: df1 = data.drop('Description', axis=1)
df1
```

Out[12]:

	Calories	Protein	TotalFat	Carbohydrate	Sodium	SaturatedFat	Cholesterol	Sugar
0	717.0	0.85	81.11	0.06	714.0	51.368	215.0	0.06
1	717.0	0.85	81.11	0.06	827.0	50.489	219.0	0.06
2	876.0	0.28	99.48	0.00	2.0	61.924	256.0	0.00
3	353.0	21.40	28.74	2.34	1395.0	18.669	75.0	0.50
4	371.0	23.24	29.68	2.79	560.0	18.764	94.0	0.51
...
7053	73.0	16.40	0.30	0.00	58.0	0.076	50.0	0.00
7054	305.0	18.50	25.10	0.00	4450.0	7.148	95.0	0.00
7055	111.0	20.54	0.84	5.41	667.0	0.218	41.0	0.00
7056	90.0	16.10	1.40	2.00	70.0	0.361	50.0	0.00
7057	89.0	19.80	0.50	0.00	68.0	0.127	50.0	0.00

7058 rows × 14 columns



```
In [14]: avg_values = df1.mean()
avg_values
```

Out[14]:

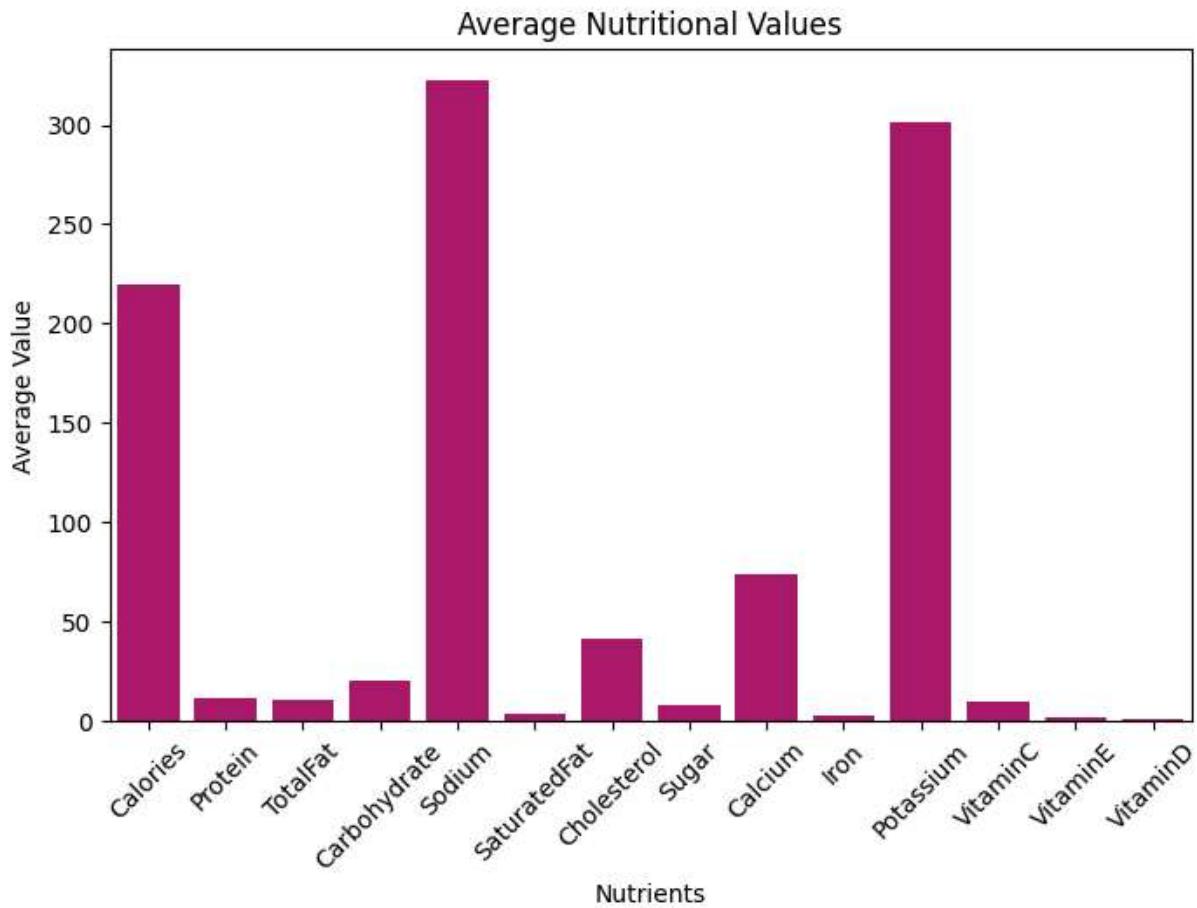
0

Calories	219.695338
Protein	11.710368
TotalFat	10.320614
Carbohydrate	20.697860
Sodium	322.059220
SaturatedFat	3.452267
Cholesterol	41.551994
Sugar	8.256540
Calcium	73.530627
Iron	2.828368
Potassium	301.357949
VitaminC	9.435980
VitaminE	1.487462
VitaminD	0.576918

dtype: float64

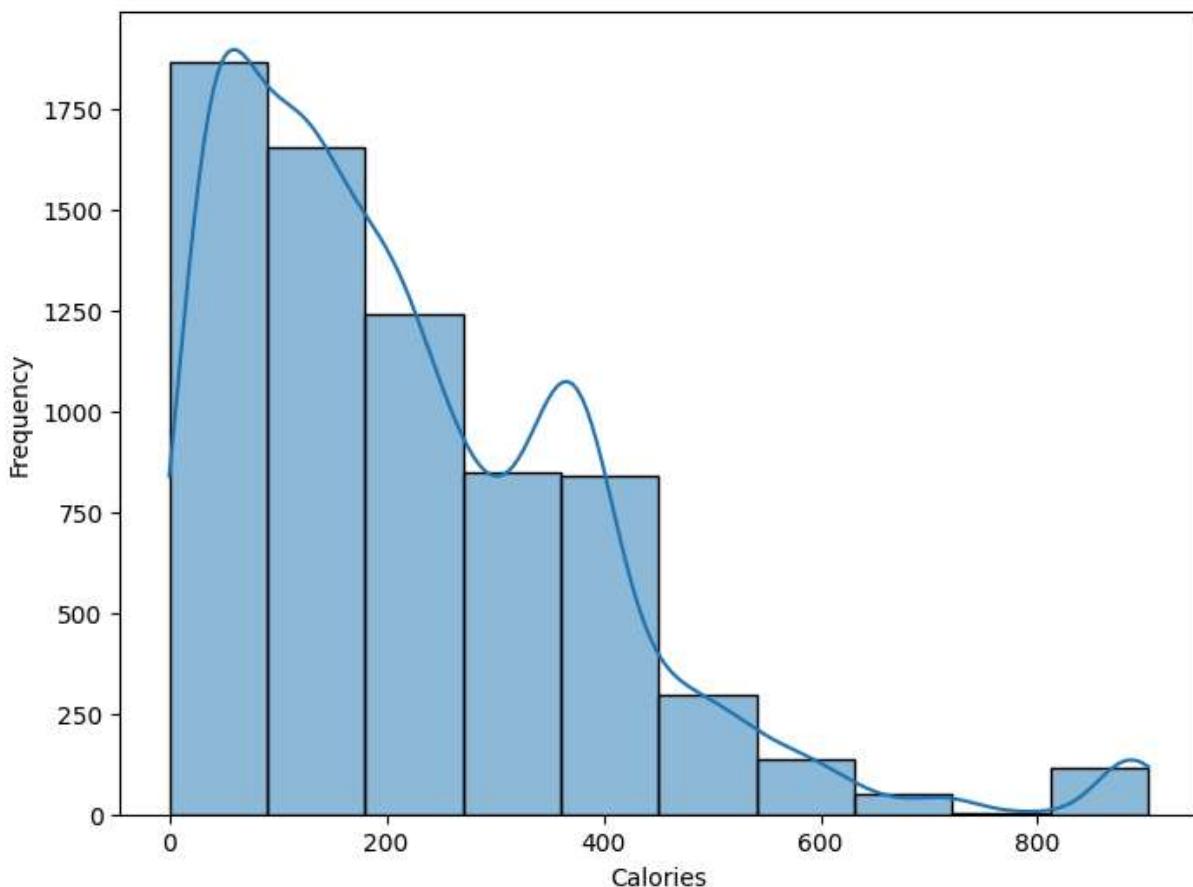
In [21]:

```
plt.figure(figsize=(8, 5))
sns.barplot(x=avg_values.index, y=avg_values.values,color="#c2026d")
plt.xticks(rotation=45)
plt.xlabel('Nutrients')
plt.ylabel('Average Value')
plt.title('Average Nutritional Values')
plt.show()
```



```
In [25]: plt.figure(figsize=(8, 6))
sns.histplot(data['Calories'], bins=10, kde=True)
plt.xlabel('Calories')
plt.ylabel('Frequency')
plt.title('Distribution of Calories')
plt.show()
```

Distribution of Calories



```
In [26]: df1.corr()
```

Out[26]:

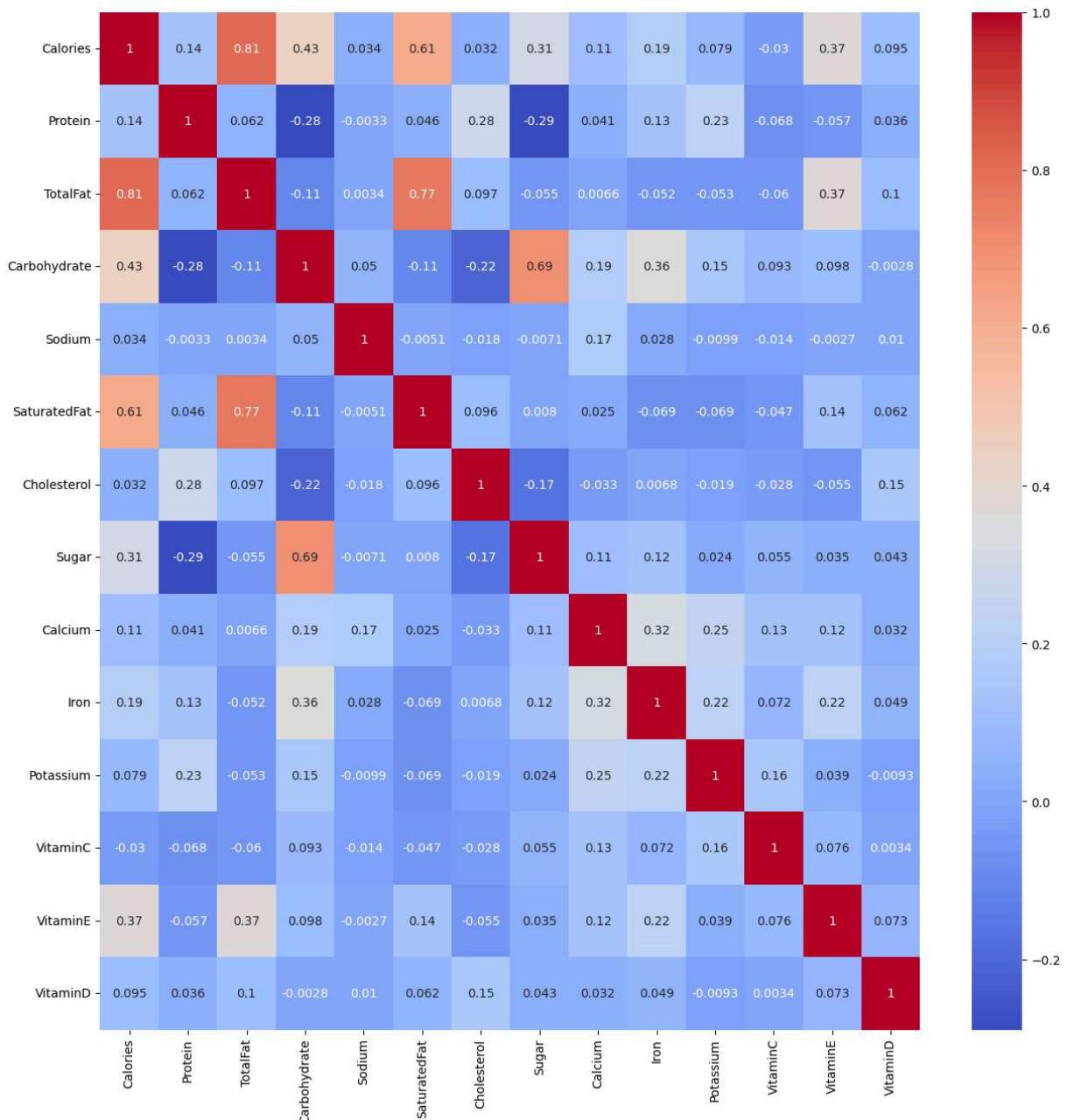
	Calories	Protein	TotalFat	Carbohydrate	Sodium	SaturatedFat	Chol
Calories	1.000000	0.135258	0.807770	0.434701	0.033703	0.611601	0.
Protein	0.135258	1.000000	0.061682	-0.284500	-0.003253	0.045784	0.
TotalFat	0.807770	0.061682	1.000000	-0.109399	0.003390	0.766142	0.
Carbohydrate	0.434701	-0.284500	-0.109399	1.000000	0.049544	-0.108676	-0.
Sodium	0.033703	-0.003253	0.003390	0.049544	1.000000	-0.005075	-0.
SaturatedFat	0.611601	0.045784	0.766142	-0.108676	-0.005075	1.000000	0.
Cholesterol	0.032433	0.280578	0.097111	-0.216070	-0.018348	0.096021	1.
Sugar	0.309989	-0.289221	-0.055459	0.688422	-0.007078	0.007981	-0.
Calcium	0.112560	0.041071	0.006585	0.187122	0.174784	0.024601	-0.
Iron	0.192506	0.133609	-0.051781	0.362023	0.027904	-0.068871	0.
Potassium	0.078807	0.225451	-0.052801	0.148615	-0.009881	-0.068748	-0.
VitaminC	-0.029628	-0.067523	-0.059612	0.093021	-0.013911	-0.046903	-0.
VitaminE	0.365777	-0.057482	0.370318	0.097550	-0.002742	0.138286	-0.
VitaminD	0.095231	0.035705	0.100754	-0.002758	0.010261	0.062197	0.



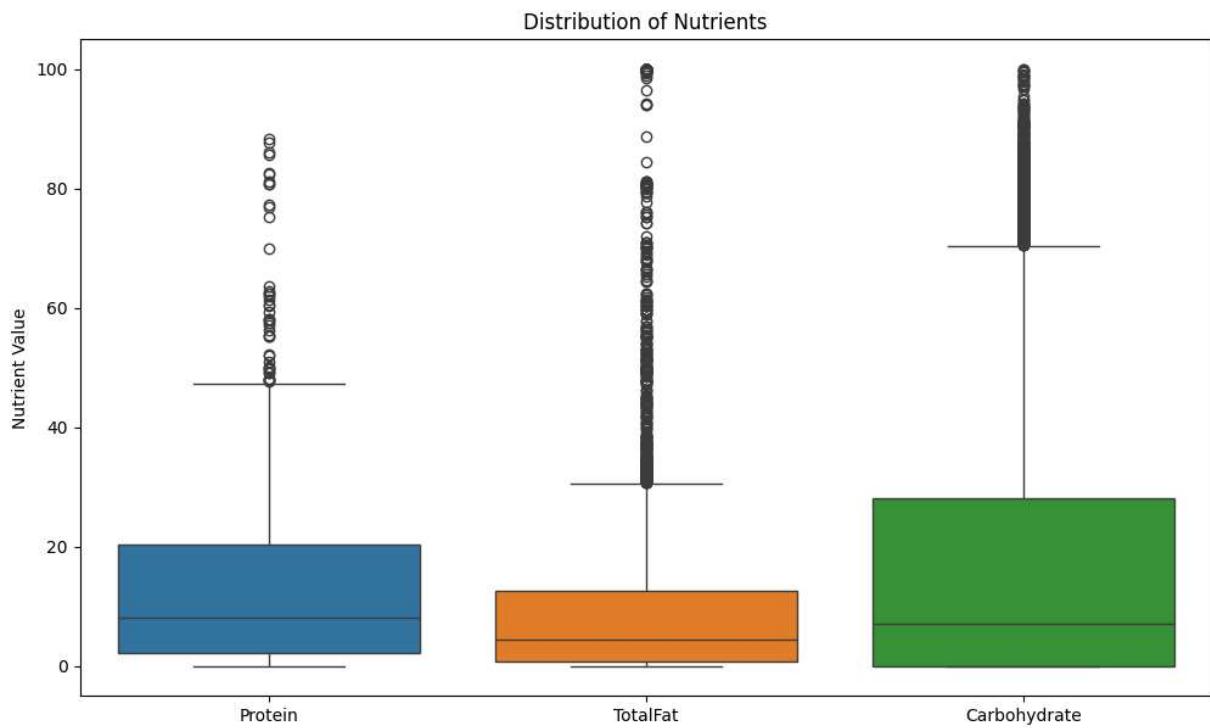
In [27]: `plt.figure(figsize=(15,15))
sns.heatmap(df1.corr(), cmap='coolwarm', annot=True)`

Out[27]: <Axes: >

food nutrient EDA



```
In [32]: plt.figure(figsize=(12, 7))
sns.boxplot(data=data[['Protein', 'TotalFat', 'Carbohydrate']])
plt.ylabel('Nutrient Value')
plt.title('Distribution of Nutrients')
plt.show()
```



missing values

In [34]: `data.info()`

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7058 entries, 0 to 7057
Data columns (total 15 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   Description  7058 non-null   object  
 1   Calories     7057 non-null   float64 
 2   Protein      7057 non-null   float64 
 3   TotalFat    7057 non-null   float64 
 4   Carbohydrate 7057 non-null   float64 
 5   Sodium       6974 non-null   float64 
 6   SaturatedFat 6757 non-null   float64 
 7   Cholesterol  6770 non-null   float64 
 8   Sugar        5148 non-null   float64 
 9   Calcium      6922 non-null   float64 
 10  Iron         6935 non-null   float64 
 11  Potassium    6649 non-null   float64 
 12  VitaminC    6726 non-null   float64 
 13  VitaminE    4338 non-null   float64 
 14  VitaminD    4224 non-null   float64 
dtypes: float64(14), object(1)
memory usage: 827.2+ KB
```

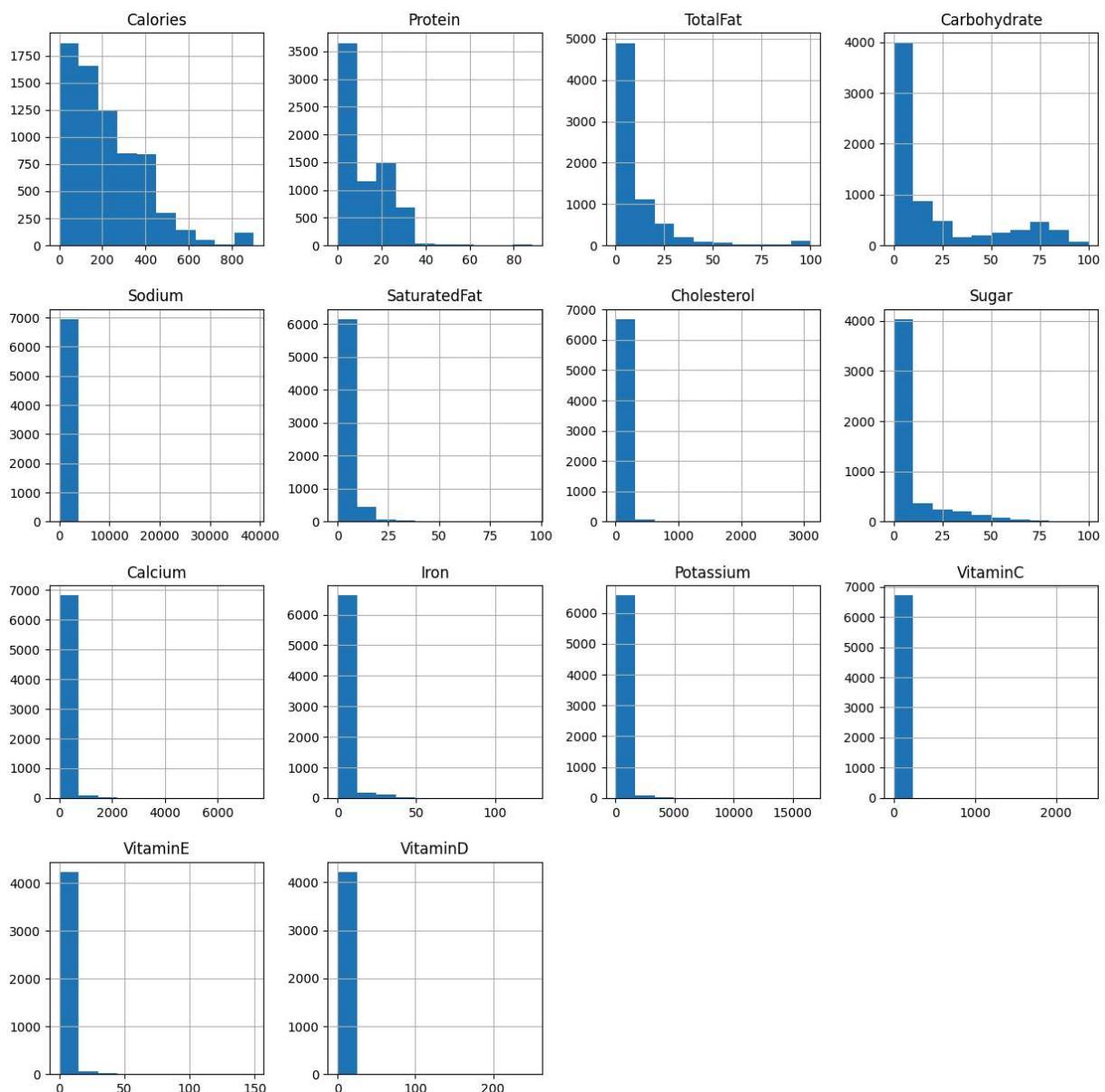
In [35]: `data.isna().sum()`

Out[35]:

	0
Description	0
Calories	1
Protein	1
TotalFat	1
Carbohydrate	1
Sodium	84
SaturatedFat	301
Cholesterol	288
Sugar	1910
Calcium	136
Iron	123
Potassium	409
VitaminC	332
VitaminE	2720
VitaminD	2834

dtype: int64In [37]: `data.hist(figsize=((15,15)))`

```
Out[37]: array([[<Axes: title={'center': 'Calories'}>,
   <Axes: title={'center': 'Protein'}>,
   <Axes: title={'center': 'TotalFat'}>,
   <Axes: title={'center': 'Carbohydrate'}>],
  [<Axes: title={'center': 'Sodium'}>,
   <Axes: title={'center': 'SaturatedFat'}>,
   <Axes: title={'center': 'Cholesterol'}>,
   <Axes: title={'center': 'Sugar'}>],
  [<Axes: title={'center': 'Calcium'}>,
   <Axes: title={'center': 'Iron'}>,
   <Axes: title={'center': 'Potassium'}>,
   <Axes: title={'center': 'VitaminC'}>],
  [<Axes: title={'center': 'VitaminE'}>,
   <Axes: title={'center': 'VitaminD'}>, <Axes: >, <Axes: >]],
 dtype=object)
```



```
In [38]: data.columns
```

```
Out[38]: Index(['Description', 'Calories', 'Protein', 'TotalFat', 'Carbohydrate',
       'Sodium', 'SaturatedFat', 'Cholesterol', 'Sugar', 'Calcium', 'Iron',
       'Potassium', 'VitaminC', 'VitaminE', 'VitaminD'],
      dtype='object')
```

```
In [39]: for i in ['Calories', 'Protein', 'TotalFat', 'Carbohydrate',
       'Sodium', 'SaturatedFat', 'Cholesterol', 'Sugar', 'Calcium', 'Iron',
       'Potassium', 'VitaminC', 'VitaminE', 'VitaminD']:
    data[i] = data[i].fillna(data[i].median())
```

```
In [41]: data.isna().sum()
```

Out[41]:

Description	0
Calories	0
Protein	0
TotalFat	0
Carbohydrate	0
Sodium	0
SaturatedFat	0
Cholesterol	0
Sugar	0
Calcium	0
Iron	0
Potassium	0
VitaminC	0
VitaminE	0
VitaminD	0

dtype: int64

In [42]:

```
data.head(10)
```

Out[42]:

	Description	Calories	Protein	TotalFat	Carbohydrate	Sodium	SaturatedFat
0	BUTTER,WITH SALT	717.0	0.85	81.11	0.06	714.0	51.368
1	BUTTER,WHIPPED,WITH SALT	717.0	0.85	81.11	0.06	827.0	50.489
2	BUTTER OIL,ANHYDROUS	876.0	0.28	99.48	0.00	2.0	61.924
3	CHEESE,BLUE	353.0	21.40	28.74	2.34	1395.0	18.669
4	CHEESE,BRICK	371.0	23.24	29.68	2.79	560.0	18.764
5	CHEESE,BRIE	334.0	20.75	27.68	0.45	629.0	17.410
6	CHEESE,CAMEMBERT	300.0	19.80	24.26	0.46	842.0	15.259
7	CHEESE,CARAWAY	376.0	25.18	29.20	3.06	690.0	18.584
8	CHEESE,CHEDDAR	403.0	24.90	33.14	1.28	621.0	21.092
9	CHEESE,CHESHIRE	387.0	23.37	30.60	4.78	700.0	19.475



label Encoding

In [43]:

`data.nunique()`

Out[43]:

	0
Description	7054
Calories	655
Protein	2415
TotalFat	2151
Carbohydrate	2758
Sodium	1196
SaturatedFat	3213
Cholesterol	287
Sugar	1566
Calcium	498
Iron	926
Potassium	885
VitaminC	529
VitaminE	485
VitaminD	113

dtype: int64

Scaling

```
In [77]: from sklearn.preprocessing import MinMaxScaler
minmax = MinMaxScaler()
```

```
In [78]: y = data['Description']
```

```
In [79]: x = data.drop('Description', axis=1)
```

```
In [80]: newx = minmax.fit_transform(x)
```

```
In [81]: type(newx)
```

Out[81]: numpy.ndarray

```
In [82]: df = pd.DataFrame(newx, columns=['Calories', 'Protein', 'TotalFat', 'Carbohydrate',
                                         'Sodium', 'SaturatedFat', 'Cholesterol', 'Sugar', 'Calcium', 'Iron',
                                         'Potassium', 'VitaminC', 'VitaminE', 'VitaminD'])
```

```
In [83]: df = pd.concat([y, df], axis=1)
df
```

Out[83]:

	Description	Calories	Protein	TotalFat	Carbohydrate	Sodium	Saturated
0	BUTTER,WITH SALT	0.794900	0.009624	0.8111	0.0006	0.018422	0.51
1	BUTTER,WHIPPED,WITH SALT	0.794900	0.009624	0.8111	0.0006	0.021338	0.52
2	BUTTER OIL,ANHYDROUS	0.971175	0.003170	0.9948	0.0000	0.000052	0.64
3	CHEESE,BLUE	0.391353	0.242301	0.2874	0.0234	0.035993	0.19
4	CHEESE,BRICK	0.411308	0.263134	0.2968	0.0279	0.014449	0.19
...
7053	FROG LEGS,RAW	0.080931	0.185688	0.0030	0.0000	0.001496	0.00
7054	MACKEREL,SALTED	0.338137	0.209466	0.2510	0.0000	0.114815	0.07
7055	SCALLOP,(BAY&SEA),CKD,STMD	0.123060	0.232563	0.0084	0.0541	0.017209	0.00
7056	SNAIL,RAW	0.099778	0.182292	0.0140	0.0200	0.001806	0.00
7057	TURTLE,GREEN,RAW	0.098670	0.224185	0.0050	0.0000	0.001754	0.00

7058 rows × 15 columns



Takeways- -Histogram plots represents the distributions of nutrient values, revealing spread and skewness.

-Boxplots indicate varying ranges and outliers for Protein, Total Fat, and Carbohydrate.

-Strong positive correlation between Calories and Total Fat (0.808), and Total Fat and Saturated Fat (0.766).

-Moderate positive correlation between Carbohydrate and Calories (0.435), and negative correlation with Protein (-0.285).

-Sodium has relatively low correlations with other nutrients.

Average Nutrient Values: -Calories, Total Fat, and Sodium have the highest average values.

-Vitamin E and Vitamin D have relatively low average values compared to other nutrients.