In [2]:
```
# imports
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
import seaborn as sns
```

Load Data

In [3]:
```
# Load Datasets
hca_users = pd.read_csv('datasets/Fake_datasets/hca_users.csv')
hca_food_items = pd.read_csv('datasets/Fake_datasets/hca_food_items.csv')
hca_recipes = pd.read_csv('datasets/Fake_datasets/hca_recipes.csv')
hca_food_logs = pd.read_csv('datasets/Fake_datasets/hca_food_logs.csv')
```

In [4]:
```
# print the shape
print("hca_users : "+str(hca_users.shape))
print("hca_food_items: "+str(hca_food_items.shape))
print("hca_recipes: "+str(hca_recipes.shape))
print("hca_food_logs: "+str(hca_food_logs.shape))
```

```
hca_users : (100, 29)
hca_food_items: (100, 16)
hca_recipes: (100, 18)
hca_food_logs: (100, 13)
```

In [5]:
```
hca_users_filtered = hca_users[["trainer_code","user_code","gender","cuisine_ids", "spent_on_meal", "medical_condition_ids", "aller
hca_food_items_filtered = hca_food_items[["fdc_id", "item_name", "food_category", "energy", "protein", "cholesterol", "carbohydrate
#hca_recipes_filtered = hca_recipes[["recipe_id","recipe_name", "prepare_time", "total_carbohydrate", "total_protein", "total_energ
hca_recipes_filtered = hca_recipes
hca_food_logs_filtered = hca_food_logs[["log_date","item_id","food_time","food_type","log_time","energy","protein","carbohydrate","
```

Preporocessing

Exploratory Data Analysis

In [6]:
```
# Checking for null values in hca_users
hca_users_filtered.isnull().sum()
```

Out[6]:
```
trainer_code             0
user_code                0
gender                   0
cuisine_ids              0
spent_on_meal            0
medical_condition_ids    0
allergic_condition_ids   0
dtype: int64
```

In [7]:
```
# Checking for null values in hca_food_items
hca_food_items_filtered.isnull().sum()
```

Out[7]:
```
fdc_id          0
item_name       0
food_category   0
energy          0
protein         0
cholesterol     0
carbohydrate    0
lipid_fat       0
dtype: int64
```

In [8]:
```
# Checking for null values in hca_recipes
hca_recipes_filtered.isnull().sum()
```

Out[8]:
```
Unnamed: 0           0
recipe_id            0
recipe_name          0
recipeSummary        0
num_of_serving       0
serving_size         0
item_unit            0
recipe_weight        0
recipe_type_id       0
recipe_notes         0
prepare_time         0
description          0
```

```
total_carbohydrate    0
total_protein         0
total_energy          0
total_lipid_fat       0
recipe_image          0
food_type             0
dtype: int64
```

In [10]:
```
# Checking for null values in hca_food_logs
hca_food_logs_filtered.isnull().sum()
```

Out[10]:
```
log_date         0
item_id          0
food_time        0
food_type        0
log_time         0
energy           0
protein          0
carbohydrate     0
lipid_fat        0
item_unit        0
item_quantity    0
created_by       0
dtype: int64
```

Recipies

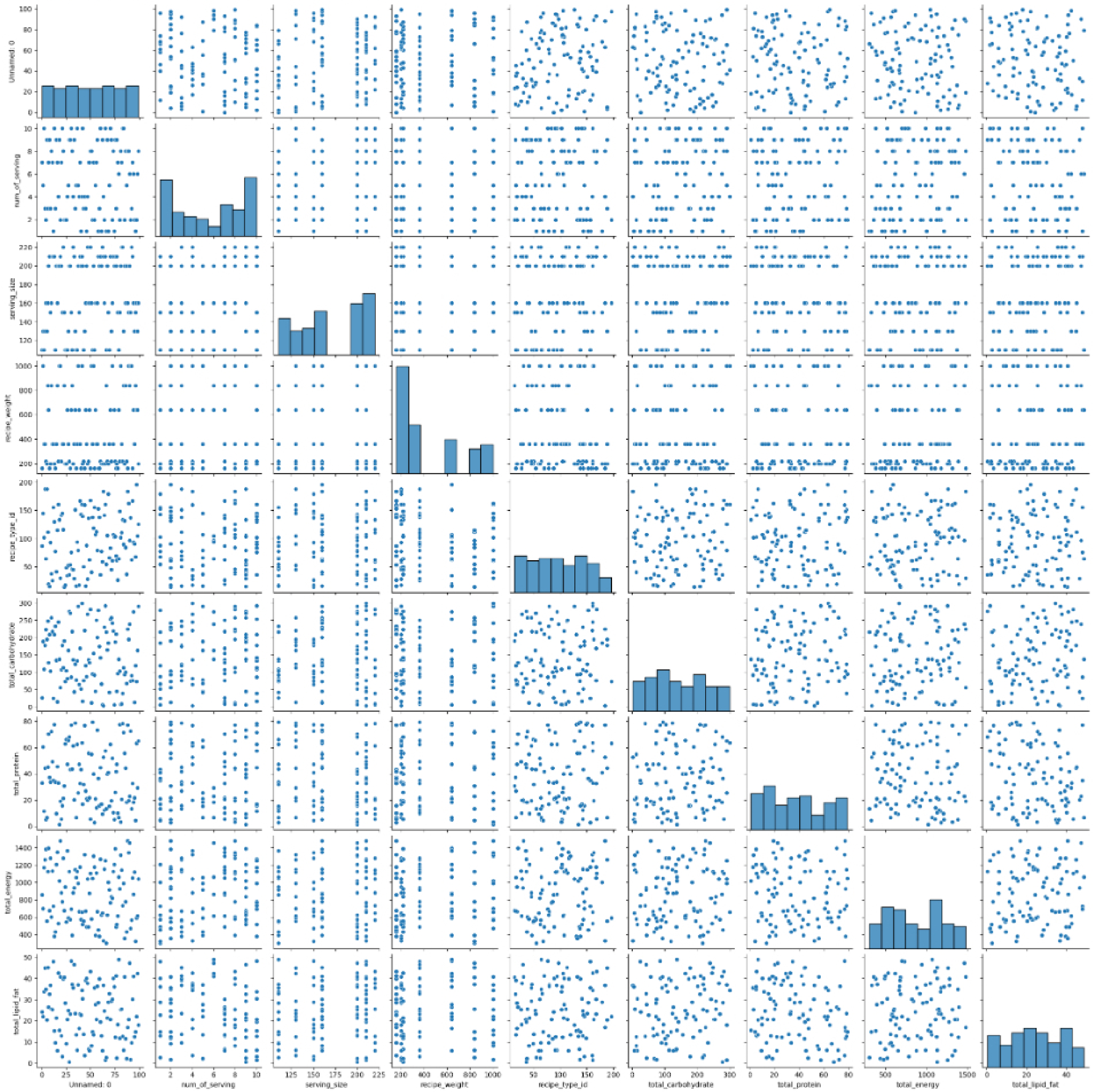
In [11]:
```
# Investigate all the elements within each Feature (Finding Categorical Variables) for hca_recipes_filtered

for column in hca_recipes_filtered:
    unique_values = np.unique(hca_recipes_filtered[column])
    nr_values = len(unique_values)
    if nr_values <= 5:
        print("The number of values for feature {} is: {} -- {}".format(column, nr_values, unique_values))
    else:
        print("The number of values for feature {} is: {}".format(column, nr_values))
```

```
The number of values for feature Unnamed: 0 is: 100
The number of values for feature recipe_id is: 100
The number of values for feature recipe_name is: 10
The number of values for feature recipeSummary is: 10
The number of values for feature num_of_serving is: 10
The number of values for feature serving_size is: 7
The number of values for feature item_unit is: 2 -- ['gm' 'ml']
The number of values for feature recipe_weight is: 7
The number of values for feature recipe_type_id is: 79
The number of values for feature recipe_notes is: 10
The number of values for feature prepare_time is: 5 -- ['10 min' '15 min' '25 min' '30 min' '5 min']
The number of values for feature description is: 10
The number of values for feature total_carbohydrate is: 100
The number of values for feature total_protein is: 100
The number of values for feature total_energy is: 96
The number of values for feature total_lipid_fat is: 99
The number of values for feature recipe_image is: 100
The number of values for feature food_type is: 1 -- ['recipe']
```

In [12]:
```
# Visualize the data using seaborn Pairplots
g = sns.pairplot(hca_recipes_filtered)
```

```
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
```

```
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

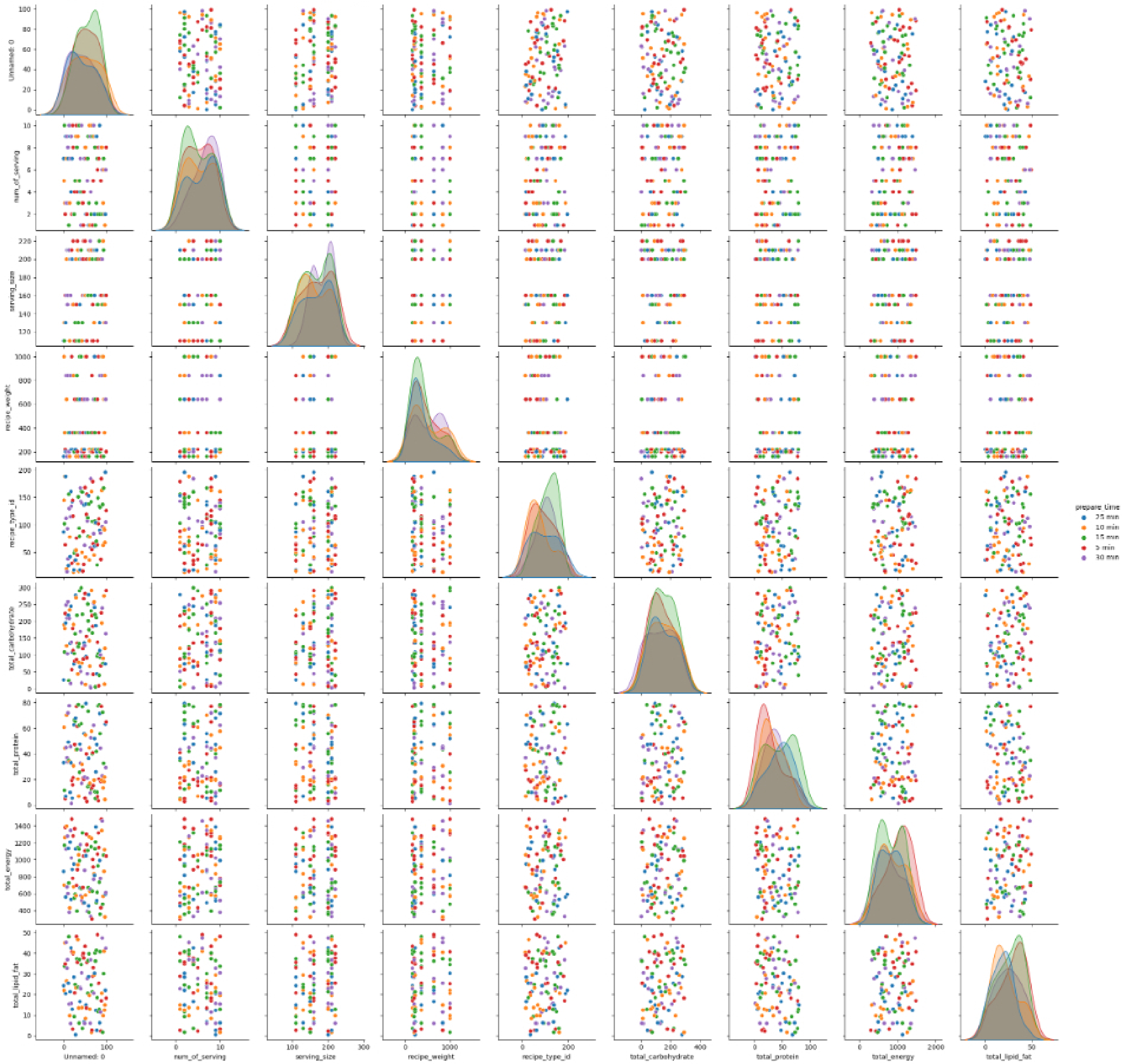


In [13]:

```python
# Visualize the data using seaborn Pairplots
g = sns.pairplot(hca_recipes_filtered, hue = 'prepare_time')
```

```
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

```
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



Food Items

In [14]:

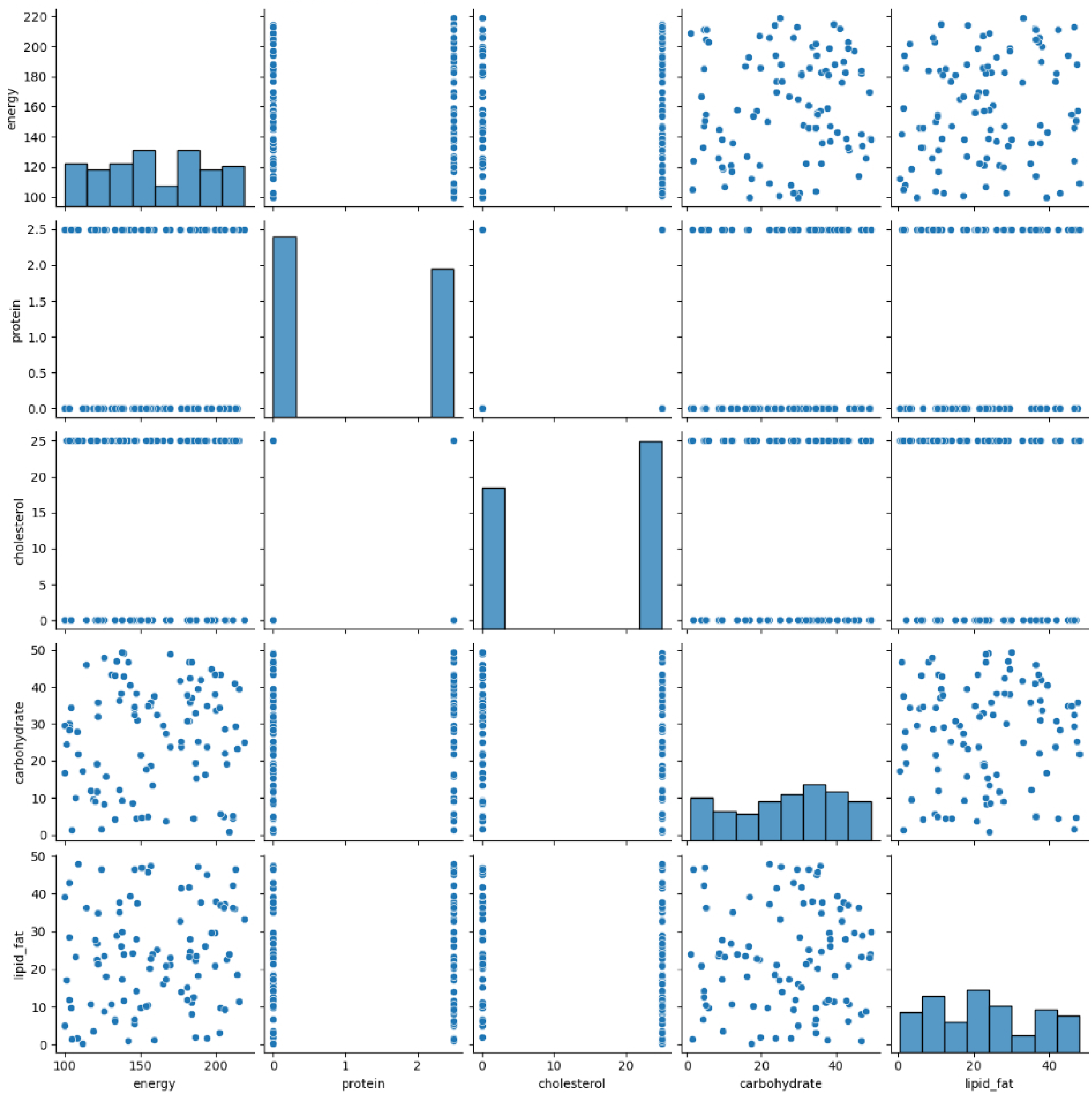
```python
# Investigate all the elements within each Feature (Finding Categorical Variables) for hca_food_items_filtered
for column in hca_food_items_filtered:
    unique_values = np.unique(hca_food_items_filtered[column])
    nr_values = len(unique_values)
    if nr_values <= 10:
        print("The number of values for feature {} is: {} -- {}".format(column, nr_values, unique_values))
    else:
        print("The number of values for feature {} is: {}".format(column, nr_values))
```

```
The number of values for feature fdc_id is: 100
The number of values for feature item_name is: 19
The number of values for feature food_category is: 2 -- ['BBQ & Cheese Sauce' 'Ice Cream & Frozen Yogurt']
The number of values for feature energy is: 71
The number of values for feature protein is: 2 -- [0.  2.5]
The number of values for feature cholesterol is: 2 -- [ 0 25]
The number of values for feature carbohydrate is: 98
```

```
The number of values for feature carbohydrate is: 98
The number of values for feature lipid_fat is: 99
```

In [15]:
```python
# Visualize the data using seaborn Pairplots

g = sns.pairplot(hca_food_items_filtered)
```

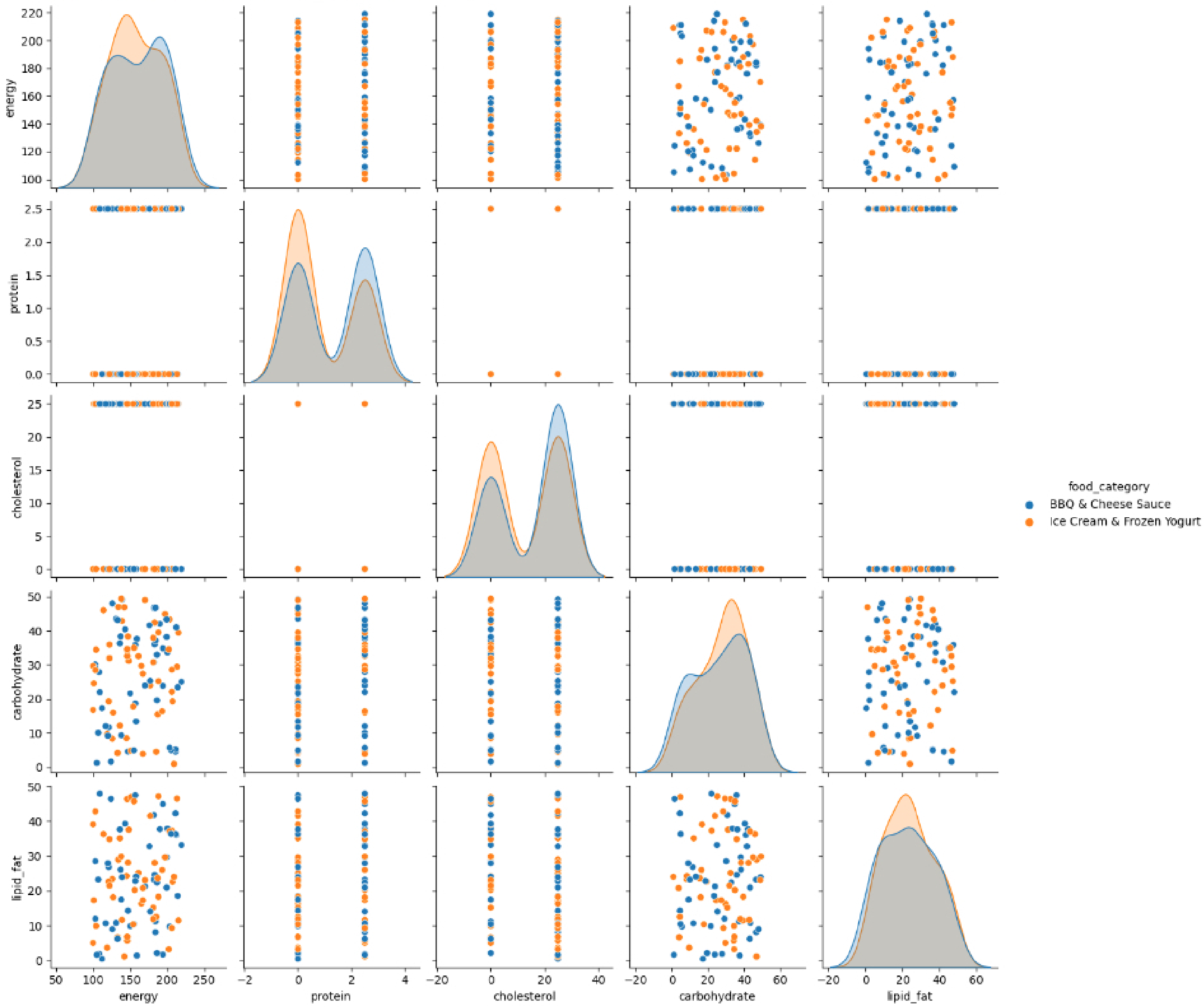
```
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



In [16]:
```python
# Visualize the data using seaborn Pairplots
g = sns.pairplot(hca_food_items_filtered, hue = 'food_category')
```

```
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
```

```
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



Food Log

In [17]:

```python
# Investigate all the elements within each Feature (Finding Categorical Variables) for hca_recipes_filtered

for column in hca_food_logs_filtered:
    unique_values = np.unique(hca_food_logs_filtered[column])
    nr_values = len(unique_values)
    if nr_values <= 10:
        print("The number of values for feature {} is: {} -- {}".format(column, nr_values, unique_values))
    else:
        print("The number of values for feature {} is: {}".format(column, nr_values))
```

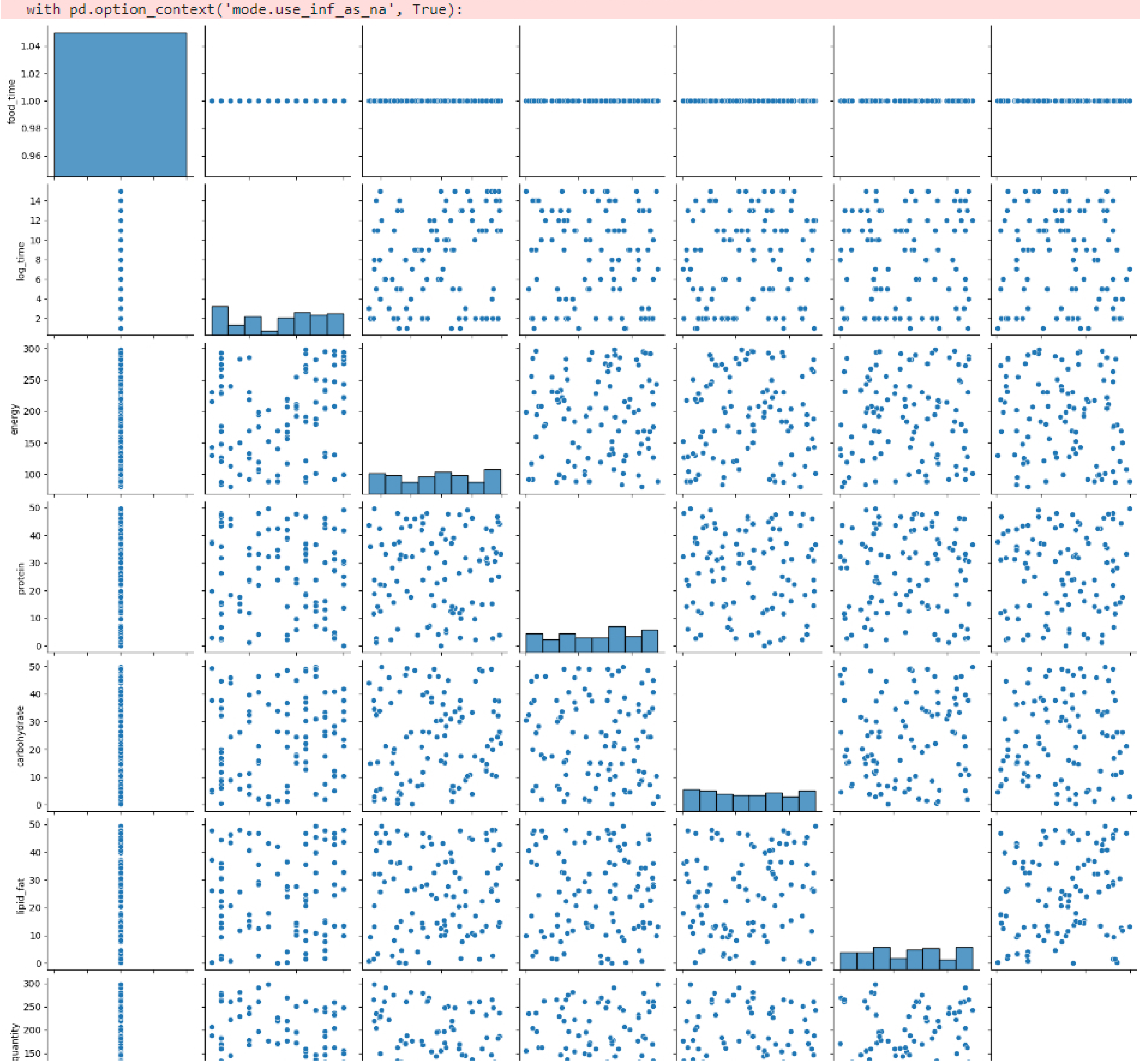
```
The number of values for feature log_date is: 20
The number of values for feature item_id is: 76
The number of values for feature food_time is: 1 -- [1]
The number of values for feature food_type is: 2 -- ['food_item' 'recipe']
The number of values for feature log_time is: 15
The number of values for feature energy is: 86
The number of values for feature protein is: 99
```

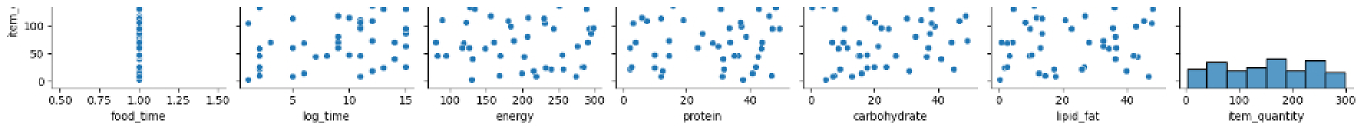
```
The number of values for feature carbohydrate is: 100
The number of values for feature lipid_fat is: 98
The number of values for feature item_unit is: 4 -- ['1 medium' '1 slice' 'gm' 'ml']
The number of values for feature item_quantity is: 84
The number of values for feature created_by is: 61
```

In [18]:

```python
# Visualize the data using seaborn Pairplots

g = sns.pairplot(hca_food_logs_filtered)
```

```
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```





In [19]:

```python
# Visualize the data using seaborn Pairplots
g = sns.pairplot(hca_food_logs_filtered, hue = 'food_type')
```

```
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Dinesh\Projects\HCA\env\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and w
ill be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

Saving Filtered Datasets

In [20]:
```
hca_users.to_csv('datasets/Fake_datasets/filtered/hca_users.csv')
hca_food_items.to_csv('datasets/Fake_datasets/filtered/hca_food_items.csv')
hca_recipes.to_csv('datasets/Fake_datasets/filtered/hca_recipes.csv')
hca_food_logs.to_csv('datasets/Fake_datasets/filtered/hca_food_logs.csv')
```

In [ ]:
```

```