Programming Assignment 4

Name: Jimmy Nguyen Date: December 5th, 2022 Class/Professor: INST447/Sigalo

---

Research Questions:

1. Is there an association between the amount of protein and energy in a food?
2. Which food is the most nutritious?
3. Which food is the least nutritious?

In [4]:
```
# import packages
import pandas as pd
import numpy as np
import requests
from pandas.io.json import json_normalize #special package in pandas
import json
import matplotlib.pyplot as plt
%matplotlib inline
```

Section A: FoodDataCentral API

In [5]:
```
apiKey = "YBx4jCFqAh5pYnDh41zIMeD9NKQKzt92MOzn6yyY"
params = {'api_key': apiKey}
```

- **Use your Food Data Central API Key to send an API request to search the database for the following list of Thanksgiving dinner foods: turkey, macaroni and cheese, mashed potatoes, bread stuffing, ham, sweet potato souffle, cranberry sauce, mixed vegetables, apple pie, pecan pie.**
- **Create a dataframe with the average nutrient values for energy, carbs, protein, fiber, and fat for the foods in the list above.**

In [6]:
```
counter = 0
food_list = ["turkey", "macaroni and cheese", "mashed potatoes", "bread stuffing", "ha

for food in food_list:

    counter += 1
    response = requests.post(
        r'https://api.nal.usda.gov/fdc/v1/search',
        params = params,
        json = {'generalSearchInput': food}
        )

    item = response.json()

    l = [i for i in range(len(item['foods']))]
```

```
    for i in l:
        if i == 0:
            rdf = pd.json_normalize(item['foods'][i]['foodNutrients'])
        else:
            df = pd.json_normalize(item['foods'][i]['foodNutrients'])
            rdf = pd.concat([rdf,df])

    rdf['nutrientName'].replace('Energy','Energy',inplace=True)
    rdf['nutrientName'].replace('Carbohydrate, by difference','Carbs',inplace=True)
    rdf['nutrientName'].replace('Protein','Protein',inplace=True)
    rdf['nutrientName'].replace('Fiber, total dietary','Fiber',inplace=True)
    rdf['nutrientName'].replace('Total lipid (fat)','Fat',inplace=True)
    rdf['nutrientName'].replace('Sodium, Na','Sodium',inplace=True)
    rdf['nutrientName'].replace('Fatty acids, total saturated','Sat_Fatty_Acids',inpla
    rdf['nutrientName'].replace('Calcium, Ca','Calcium',inplace=True)
    rdf['nutrientName'].replace('Iron, Fe','Iron',inplace=True)
    rdf['nutrientName'].replace('Sugars, total including NLEA','Sugar',inplace=True)
    rdf['nutrientName'].replace('Cholesterol','Chol',inplace=True)
    rdf['nutrientName'].replace('Fatty acids, total trans','Trans_Fatty_Acids',inplace
    rdf['nutrientName'].replace('Vitamin C, total ascorbic acid','VitaminC',inplace=Tr
    rdf['nutrientName'].replace('Vitamin A, IU','VitaminA',inplace=True)
    rdf['nutrientName'].replace('Potassium, K','Potassium',inplace=True)
    rdf['nutrientName'].replace('Fatty acids, total polyunsaturated','Unsat_Fatty_Acid

    agg_food = rdf[['nutrientName','value']].groupby(['nutrientName']).agg('mean')
    agg_food = agg_food.reset_index()
    agg_food.columns = ["nutrientName","Value"]
    agg_food = agg_food[agg_food["nutrientName"].isin(['Energy','Carbs','Protein','Fib
    agg_food["Food"] = food

    new=agg_food.pivot_table(index=["Food"], columns=['nutrientName'], values='Value')
    new.reset_index(inplace=True)

    if counter ==1:
        all_foods = new
    else:
        all_foods = pd.concat([all_foods,new])

all_foods
```

Out[6]:

	nutrientName	Food	Carbs	Energy	Fat	Fiber	Protein
0	0	turkey	7.726800	393.138462	11.490800	0.457447	16.309400
1	0	macaroni and cheese	21.293000	337.333333	7.169800	1.040000	6.768000
2	0	mashed potatoes	39.653542	227.346154	4.151250	3.089583	4.389375
3	0	bread stuffing	27.434200	361.586207	9.924400	2.394000	8.527600
4	0	ham	8.709600	286.032787	8.660400	0.602381	13.732400
5	0	sweet potato souffle	24.002653	226.904762	6.355714	2.628571	3.046531
6	0	cranberry sauce	26.694375	219.115942	0.948571	1.880851	1.860200
7	0	mixed vegetables	9.677551	49.040816	0.046250	2.185714	1.696327
8	0	apple pie	37.568085	340.307692	13.170213	2.295652	2.942766
9	0	pecan pie	42.379600	500.481481	24.097600	2.757143	4.743000

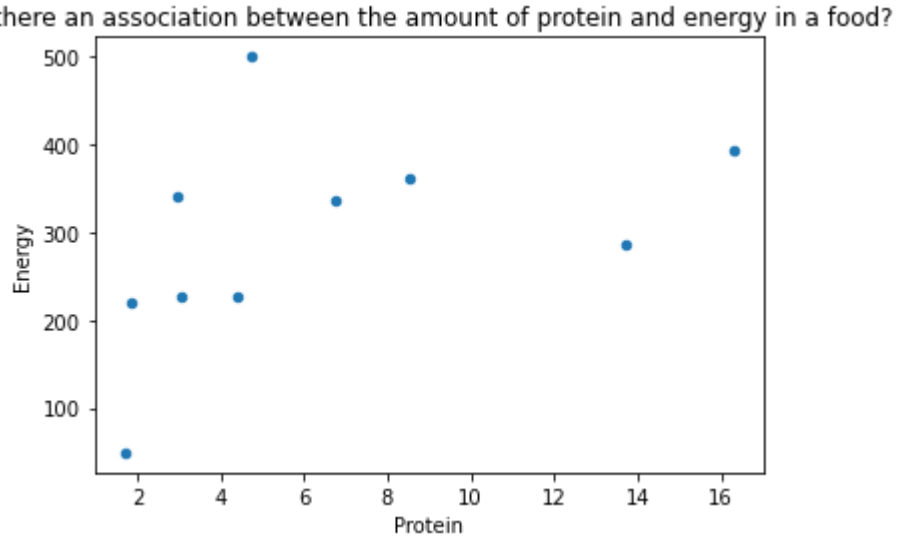
Consider the following research question: *Is there an association between the amount of protein and energy in a food?* Create a plot to show this relationship (your plot must include a title & proper axis labels). Interpret the results. Minimum 2 sentences.

In [7]:
```
# RQ1: Is there an association between the amount of protein and energy in a food?
# Create a scatterplot
all_foods.plot.scatter(x='Protein',y='Energy')

# Add axis labels
plt.xlabel("Protein")
plt.ylabel("Energy")

# add title
plt.title("Is there an association between the amount of protein and energy in a food?
```

Out[7]:
```
Text(0.5, 1.0, 'Is there an association between the amount of protein and energy in a
food?')
```



Upon further observation of the colloration between the amount of protein and energy in a food, there seems to be no exact corellation between the two variables. To explain, the increase in

Protein (x-axis) does not linearly (or any other types of relations) affect the amount of energy (y-axis) that a food has.

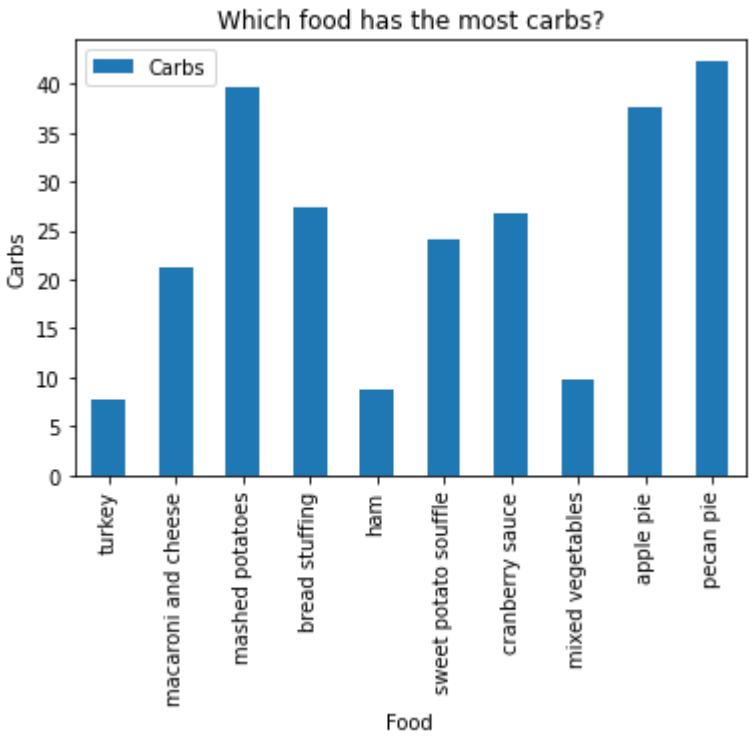
Create bar graphs to show the nutrient concentration for each food item. Each bar graph should show the concentrations of a single nutrient across all food items. Your plot must include a title & proper axis labels. Interpret the results. (You should have 5 graphs and 5 interpretations. Minimum 3 sentences for each interpretation.).

In [8]:
```
# Create a scatterplot
all_foods.plot.bar(x='Food',y='Carbs')

# Add axis labels
plt.xlabel("Food")
plt.ylabel("Carbs")

# add title
plt.title("Which food has the most carbs?")
```

Out[8]:
```
Text(0.5, 1.0, 'Which food has the most carbs?')
```



As the graph states, pecan pie ranks 1st (when we are deciding which food has the most carbs) following up with apple pie in 2nd and mashed poratoes 3rd. These observations are very reasonable as potatoes in general are a great source of carbs. For pies, the ingredients needed to build the outside crust (some ingredients inside as well) are made up of carbs so it also makes sense. One surprising take that I had while observing was that ham, turkey, and mixed vegetables also had some sources of carbs as well.

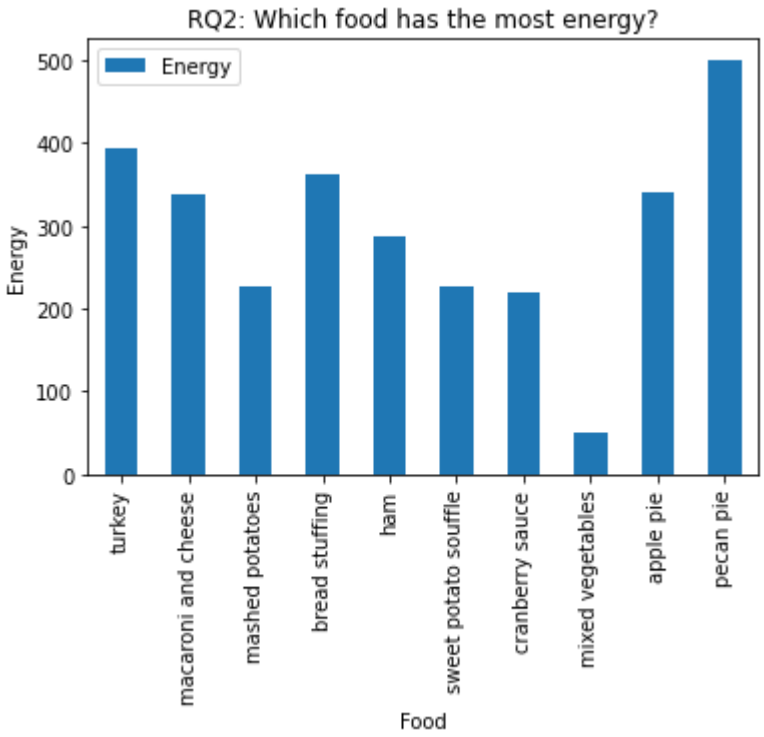
In [9]:
```
# Create a scatterplot
all_foods.plot.bar(x='Food',y='Energy')

# Add axis labels
```

```
plt.xlabel("Food")
plt.ylabel("Energy")

# add title
plt.title("RQ2: Which food has the most energy?")
```

Out[9]:
```
Text(0.5, 1.0, 'RQ2: Which food has the most energy?')
```



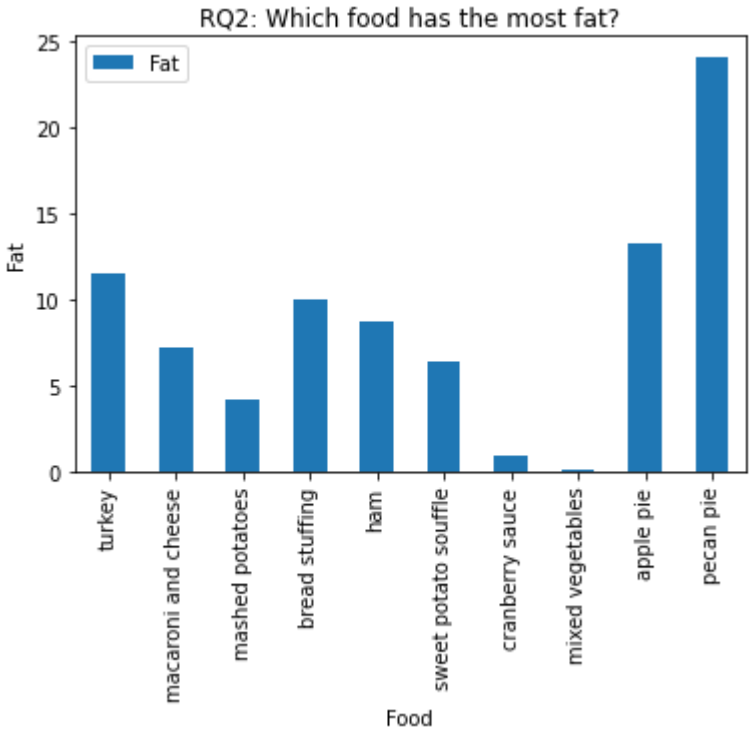
Next up, pecan pie ranks 1st following up with turkey in 2nd and bread stuffings 3rd. These observations are interesting as they vary the same throughout most of the foods in the list. For pecan pie, I'm assuming that the pie uses a variety of ingredients which could give it its energy-densed advantage. It was surprising to find out that turkey had ranked 2nd in top sources of energy. My assumption is that the amount of fat in turkey could contribute it towards being higher in the energy department.

In [10]:
```
# Create a scatterplot
all_foods.plot.bar(x='Food',y='Fat')

# Add axis labels
plt.xlabel("Food")
plt.ylabel("Fat")

# add title
plt.title("RQ2: Which food has the most fat?")
```

Out[10]:
```
Text(0.5, 1.0, 'RQ2: Which food has the most fat?')
```



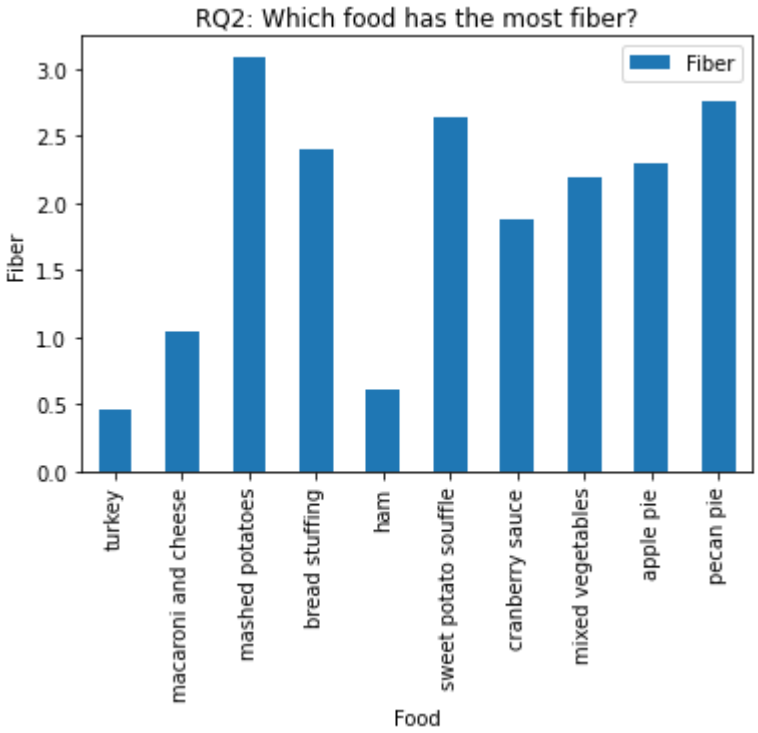
Thirdly, pecan pie ranks 1st once again, following up with apple pie in 2nd and turkey 3rd. These observations are very reasonable as potatoes in general are a great source of carbs. For pies, the ingredients needed to build the outside crust (some ingredients inside as well) are made up of carbs so it also makes sense. One surprising take that I had while observing was that ham, turkey, and mixed vegetables also had some sources of carbs as well.

In [11]:
```
# Create a scatterplot
all_foods.plot.bar(x='Food',y='Fiber')

# Add axis labels
plt.xlabel("Food")
plt.ylabel("Fiber")

# add title
plt.title("RQ2: Which food has the most fiber?")
```

Out[11]:
```
Text(0.5, 1.0, 'RQ2: Which food has the most fiber?')
```



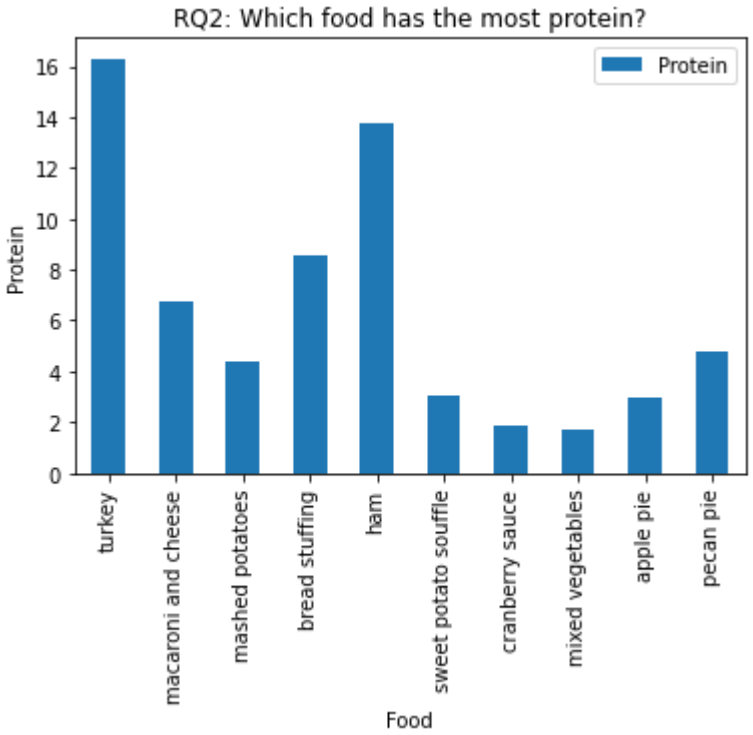
For our fourth graph, mashed potatoes ranks 1st, following up with pecan pie in 2nd and sweet potato souffle 3rd. These observations are very interesting as I had no clue that meals with potatoes (or carbs) had so much fiber in them. Upon further research, I found out that potatoes are a great source of carbs and fiber together if we exclude the butter and all other components that makes potatoes taste great. It was interesting to see that cranberry sauce had high amount of fiber in them as well. Additionally, ham and turkey was surprisingly low in the fiber department.

In [12]:
```
# Create a scatterplot
all_foods.plot.bar(x='Food',y='Protein')

# Add axis labels
plt.xlabel("Food")
plt.ylabel("Protein")

# add title
plt.title("RQ2: Which food has the most protein?")
```

Out[12]:
```
Text(0.5, 1.0, 'RQ2: Which food has the most protein?')
```



Lastly, turkey ranks 1st following up with ham in 2nd and bread stuffing 3rd. These observations are very reasonable as turkey and ham in general are a great source of proteins. Other additional sources of protein would be brocolli, chicken, beef, and pork. One surprising take that I had while observing was that bread stuffing also had a decent amount of protein as well. Upon further research, I've come to the conclusion that depending on the ingredients used to make the bread stuffing, the amount of proteins could increase/decrease based on those ingredients.

Answer the following questions:

- **Based on the graphs above, which of the Thanksgiving foods would you say is the healthiest/most nutritious, and why? There is no wrong answer here. Minimum 5 sentences.**
- **Based on the graphs above, which of the Thanksgiving foods would you say is the least healthiest/least nutritious, and why? There is no wrong answer here. Minimum 5 sentences.**

- *RQ2: The most healthiest Thanksgiving food would have to be the pecan pie (assuming that this is an average person with no health problems and a normal diet). The reason why the pecan pie would be the best is because pecan pie is high in carbs, energy, fat, and fiber; All of which are craved by the human body in order to store energy. Carbs are great source of energy which your body could tap into immediately; This is the reason why marathon runners stack up on carbs before a race. Additionally energy, fat, and fiber are all great as well because of the way your body needs these categories to process a person's digestive system. Of course in the long term, this would not be the healthiest meal as it would spike your blood levels.*

- *RQ3: The least healthiest Thanksgiving food would have to be the mixed vegetables (and personally my least favorite). This might come to as a surprise but mixed vegetable are needed in the sense that they reduce your hunger and handle other nutritional factors that aren't in these five categories. The main reason why its the least healthiest is because of its inability to deliver in the carbs, energy, fat, and protein department. Mixed vegetables are however great in order ways as it reduces the amount of food inside your body with the amount of fiber that it has. But since this is Thanksgiving, I would personally rather be thankful for the food rather than wasting it down the drain. :)*

Additional take: *Based on the graphs above, I would say that it depends on your normal activity routine (in addition with your diet). The reason why these two attributes are important to carefully assess first is because depending on a specific person's routine and diet, there are some foods that would make to be the most healthiest for the person and some that are the least healthiest. Take two examples: Person A who is an athlete, and Person B who works a normal 9-5 job. Person A would greatly benefit from having a pecan pie since it is so energy densed (as well as in other categories) while Person B would suffer and possibly gain weight from consuming too much pecan pie.*

Section B: Thanksgiving Food-Related Tweets

In [13]:
```
food_tweets = pd.read_csv('food_tweets_mentions.csv')
food_tweets
```

Out[13]:

	tweet_id	food_mention	num_retweets	num_favorites
0	1	turkey	955	621
1	2	turkey	318	124
2	3	turkey	357	787
3	4	turkey	615	328
4	5	turkey	764	672
...	...	...	...	...
477	478	pumpkin pie	837	303
478	479	pumpkin pie	130	514
479	480	pumpkin pie	450	847
480	481	pumpkin pie	260	292
481	482	pumpkin pie	750	453

482 rows × 4 columns

Calculate the following descriptive statistics. (These do NOT need to be saved to a new dataframe. Simply use functions to print out these statistics)

- **Calculate the average number of retweets per tweet**

- **Calculate the average number of favorites per tweet**
- **Calculate the maximum number of retweets for a single tweet**
- **Calculate the maximum number of favorites for a single tweet**
- **In a narrative (no bullet points)**
  - **Report the descriptive statistics above in a well-written paragraph.**

In [14]:
```
food_tweets[['food_mention', 'num_retweets']].groupby(['food_mention']).agg('mean').re
```

Out[14]:

	food_mention	num_retweets
0	apple pie	405.818182
1	bean casserole	480.789474
2	bread stuffing	508.904762
3	cranberry sauce	384.285714
4	dinner rolls	569.500000
5	gravy	386.500000
6	ham	455.966667
7	macaroni and cheese	518.291667
8	mashed potatoes	471.750000
9	mixed vegetables	502.823129
10	pecan pie	503.545455
11	pumpkin pie	472.433333
12	salad	475.076923
13	sweet potato souffle	625.333333
14	turkey	539.678571

In [15]:
```
food_tweets[['food_mention', 'num_favorites']].groupby(['food_mention']).agg('mean').r
```

Out[15]:

	food_mention	num_favorites
0	apple pie	331.272727
1	bean casserole	465.421053
2	bread stuffing	511.190476
3	cranberry sauce	388.571429
4	dinner rolls	688.000000
5	gravy	382.166667
6	ham	536.200000
7	macaroni and cheese	519.614583
8	mashed potatoes	482.250000
9	mixed vegetables	503.544218
10	pecan pie	627.000000
11	pumpkin pie	497.900000
12	salad	422.000000
13	sweet potato souffle	408.333333
14	turkey	478.035714

In [16]:
```
food_tweets[['tweet_id', 'num_retweets']].groupby(['tweet_id']).max().reset_index()
```

Out[16]:

	tweet_id	num_retweets
0	1	955
1	2	318
2	3	357
3	4	615
4	5	764
...	...	...
477	478	837
478	479	130
479	480	450
480	481	260
481	482	750

482 rows × 2 columns

In [17]:
```
food_tweets[['tweet_id', 'num_favorites']].groupby(['tweet_id']).max().reset_index()
```

Out[17]:

	tweet_id	num_favorites
0	1	621
1	2	124
2	3	787
3	4	328
4	5	672
...	...	...
477	478	303
478	479	514
479	480	847
480	481	292
481	482	453

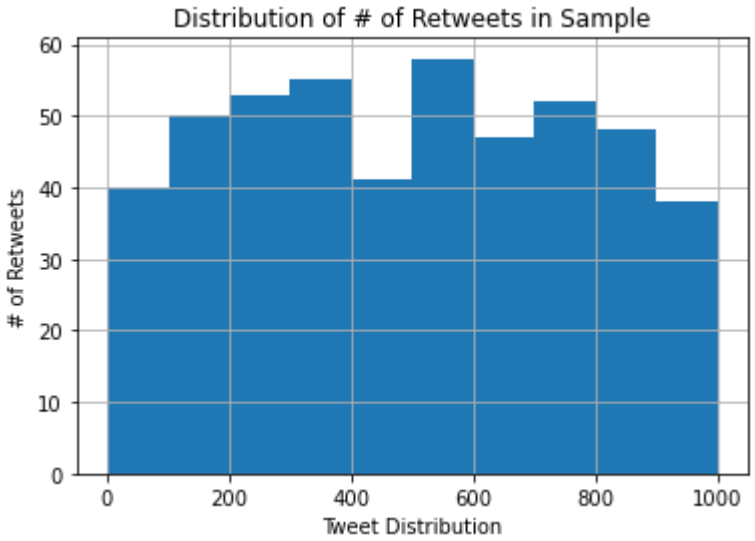
482 rows × 2 columns

For the average number of retweets per tweet, most tweet topic range from between 350 and 650. The lowest variable would have to be 'cranberry sauce' which came in to be around 384 retweets while the highest variable, sweet potato souffle, came in to have about 625 # of retweets. For the average number of favorites per tweet, most range from around 330 to 650. The lowest variable would have to be 'apple' which came in to be around 331 favorites while the highest variable, dinner rolls, came in to have about 688 favorites. For the following tweets, the max value of number of retweets as well as number of favorites fell in favor for the same tweet_id which is 482. This count came out to be 1,000 retweets and 999 favorites.

Create a histogram that shows the distribution of the number of retweets in the tweets sample. Your plot must include a title & proper axis labels. Interpret the results. Minimum 1 sentence.

In [18]:
```
h1 = food_tweets["num_retweets"].hist()
h1.set_xlabel("Tweet Distribution")
h1.set_ylabel("# of Retweets")
h1.set_title("Distribution of # of Retweets in Sample")
```

Out[18]:
```
Text(0.5, 1.0, 'Distribution of # of Retweets in Sample')
```

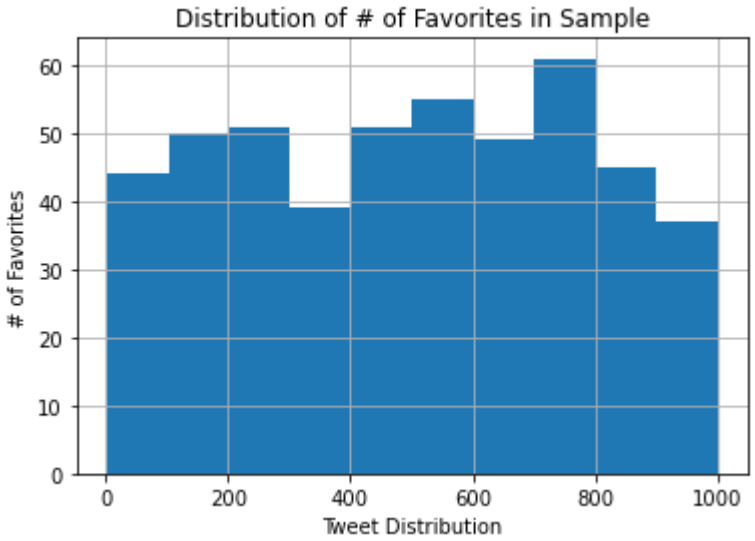


The distribution of retweets in the sample seems to follow an ordinary histogram curve with a few minor exclusions.

Create a histogram that shows the distribution of the number of favorites in the tweets sample. Your plot must include a title & proper axis labels. Interpret the results. Minimum 1 sentence.

In [19]:
```
h2 = food_tweets["num_favorites"].hist()
h2.set_xlabel("Tweet Distribution")
h2.set_ylabel("# of Favorites")
h2.set_title("Distribution of # of Favorites in Sample")
```

Out[19]:
```
Text(0.5, 1.0, 'Distribution of # of Favorites in Sample')
```



The distribution of favorites in the sample does not seem to follow an ordinary histogram curve at all if any.

Subset the tweets dataset to those tweets where the number of retweets is greater than 200 AND the number of favorites is greater than 50.

- **Name this dataframe food_tweets_subset**

In [20]:
```
food_tweets_subset = food_tweets.loc[food_tweets['num_retweets'] > 200]
food_tweets_subset = food_tweets.loc[food_tweets['num_favorites'] > 50]
food_tweets_subset
```

Out[20]:

	tweet_id	food_mention	num_retweets	num_favorites
0	1	turkey	955	621
1	2	turkey	318	124
2	3	turkey	357	787
3	4	turkey	615	328
4	5	turkey	764	672
...	...	...	...	...
477	478	pumpkin pie	837	303
478	479	pumpkin pie	130	514
479	480	pumpkin pie	450	847
480	481	pumpkin pie	260	292
481	482	pumpkin pie	750	453

457 rows × 4 columns

Using the subsetted dataset (food_tweets_subset), create an aggregated dataframe that includes the following:

- **Number of tweets for each food item mentioned. Name this column "num_tweets".**
- **Total number of retweets for each food item mentioned. Name this column "total_num_retweets".**
- **Average number of favorites for each food item mentioned. Name this column "avg_num_favorites".**
- **Name the final dataset food_tweets_subset_agg**
  - **Hint: To get the final dataset, you can do this in multiple steps, or aggregate multiple columns in a single step. See https://www.shanelynn.ie/summarising-aggregation-and-grouping-data-in-python-pandas/ Links to an external site.for more info.**

In [54]:
```
food_tweets_subset_agg = food_tweets_subset.groupby('food_mention').agg(
    num_tweets = ('tweet_id', 'count'),
    total_num_retweets = ('num_retweets', sum),
    avg_num_favorites = ('num_favorites', 'mean')
)

food_tweets_subset_agg = food_tweets_subset_agg.reset_index()
food_tweets_subset_agg
```

Out[54]:

	food_mention	num_tweets	total_num_retweets	avg_num_favorites
0	apple pie	10	3647	362.500000
1	bean casserole	17	7940	516.176471
2	bread stuffing	20	9816	535.000000
3	cranberry sauce	6	2139	449.333333
4	dinner rolls	8	4556	688.000000
5	gravy	11	4087	414.727273
6	ham	58	26419	554.275862
7	macaroni and cheese	89	45689	558.876404
8	mashed potatoes	3	1509	626.666667
9	mixed vegetables	142	71922	520.450704
10	pecan pie	11	5539	627.000000
11	pumpkin pie	28	13244	530.857143
12	salad	12	5359	455.583333
13	sweet potato souffle	15	9380	408.333333
14	turkey	27	14463	495.481481

Section C: Combine FoodDataCenter API & Thanksgiving Food-Related Tweets Data

Merge the Food Data Central API data (final dataset from Section A) and Thanksgiving Twitter API data (final dataset from Section B) based on food items that appear in BOTH datasets.

In [80]:
```
# Renaming all_foods 'Food' columns to 'food_mention'
all_foods.rename(columns = {"Food":"food_mention"}, inplace = True)

# Start the merging based on 'food_mention', excluding non-inclusive variables using
final_df = pd.merge(all_foods, food_tweets_subset_agg, how = 'inner', on = ['food_ment

# Print final df
final_df
```

Out[80]:

	food_mention	Carbs	Energy	Fat	Fiber	Protein	num_tweets	total_num_retwe
0	turkey	7.726800	393.138462	11.490800	0.457447	16.309400	27	14
1	macaroni and cheese	21.293000	337.333333	7.169800	1.040000	6.768000	89	45
2	mashed potatoes	39.653542	227.346154	4.151250	3.089583	4.389375	3	1
3	bread stuffing	27.434200	361.586207	9.924400	2.394000	8.527600	20	9
4	ham	8.709600	286.032787	8.660400	0.602381	13.732400	58	26
5	sweet potato souffle	24.002653	226.904762	6.355714	2.628571	3.046531	15	9
6	cranberry sauce	26.694375	219.115942	0.948571	1.880851	1.860200	6	2
7	mixed vegetables	9.677551	49.040816	0.046250	2.185714	1.696327	142	71
8	apple pie	37.568085	340.307692	13.170213	2.295652	2.942766	10	3
9	pecan pie	42.379600	500.481481	24.097600	2.757143	4.743000	11	5

Create a visualization that answers the following question:

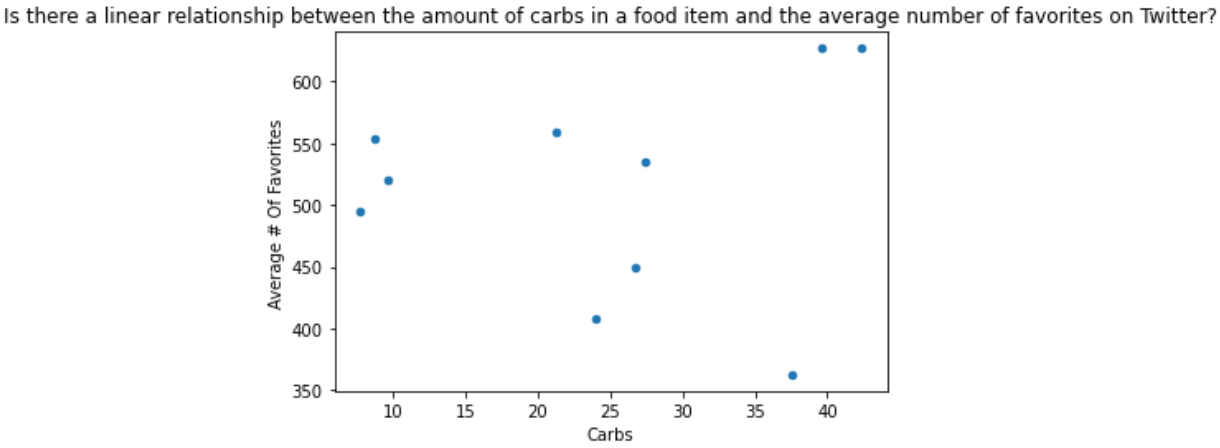
- **Is there a linear relationship between the amount of carbs in a food item (according to the FoodData Central API) and the average number of favorites on Twitter?**
  - **In a narrative (no bullet points): Describe the relationship (or lack thereof) shown in the visualization. Minimum 1 sentence.**
  - **Your plot must include a title & proper axis labels.**

In [84]:
```
# Create a scatterplot
final_df.plot.scatter(x = 'Carbs',y = 'avg_num_favorites')

# Add axis labels
plt.xlabel("Carbs")
plt.ylabel("Average # Of Favorites")

# add title
plt.title("Is there a linear relationship between the amount of carbs in a food item a
```

Out[84]:
```
Text(0.5, 1.0, 'Is there a linear relationship between the amount of carbs in a food
item and the average number of favorites on Twitter?')
```



Unfortunately, there is no linear relationship with the amount of carbs in a food when compared with the average number of favorited tweets mentioning the food. As the graph above depicts, the points between the average number of carbs for each food item and the average number of favorites for each food item are completely scattered.

In [ ]: