In [1]:

```
import pandas as pd
import numpy as np
import sys
import math
```

In [2]:

```
#Statcrunch Cereal Dataset
statcrunch = pd.read_csv('Statcrunch Cereal Dataset.csv')
#statcrunch.fillna(-1, inplace=True)
statcrunch.head()
```

Out[2]:

	Name	Manufacturer	Target	Type	Shelf	Calories	Cups	Weight	Protein	Fat	Sodium
0	100% Bran	Nabisco	adult	cold	top	70	0.33	1.0	4	1	130
1	100% Natural Bran	Quaker Oats	adult	cold	top	120	1.00	1.0	3	5	15
2	All-Bran	Kelloggs	adult	cold	top	70	0.33	1.0	4	1	260
3	All-Bran Extra Fiber	Kelloggs	adult	cold	top	50	0.50	1.0	4	0	140
4	Almond Delight	Ralston Purina	adult	cold	top	110	0.75	1.0	2	2	200

In [3]:

```
# Normalizing each column so that all the serving sizes are 1.5 cup, average
for index, row in statcrunch.iterrows():
    rate = 1.5 / row['Cups']
    statcrunch.loc[index, 'Calories'] = rate*row['Calories']
    statcrunch.loc[index, 'Cups'] = 1.5
    statcrunch.loc[index, 'Weight'] = rate*row['Weight']
    statcrunch.loc[index, 'Protein'] = rate*row['Protein']
    statcrunch.loc[index, 'Fat'] = rate*row['Fat']
    statcrunch.loc[index, 'Sodium'] = rate*row['Sodium']
    statcrunch.loc[index, 'Fiber'] = rate*row['Fiber']
    statcrunch.loc[index, 'Carbs'] = rate*row['Carbs']
    statcrunch.loc[index, 'Sugars'] = rate*row['Sugars']
    if (not math.isnan(row['Potassium'])):
        statcrunch.loc[index, 'Potassium'] = rate*row['Potassium']

    statcrunch.loc[index, 'Vitamins'] = rate*row['Vitamins']

    # Other database has it as "Kellogg", so change all "Kelloggs" to "Kellog
    if(row['Manufacturer'] == "Kelloggs"):
        statcrunch.loc[index, 'Manufacturer'] = "Kellogg"
statcrunch = statcrunch.round(2)
statcrunch
```

Out[3]:

	Name	Manufacturer	Target	Type	Shelf	Calories	Cups	Weight	Protein	Fa
0	100% Bran	Nabisco	adult	cold	top	318.18	1.5	4.55	18.18	4.5
1	100% Natural Bran	Quaker Oats	adult	cold	top	180.00	1.5	1.50	4.50	7.5
2	All-Bran	Kellogg	adult	cold	top	318.18	1.5	4.55	18.18	4.5
3	All-Bran Extra Fiber	Kellogg	adult	cold	top	150.00	1.5	3.00	12.00	0.0
4	Almond Delight	Ralston Purina	adult	cold	top	220.00	1.5	2.00	4.00	4.0
5	Apple Cinnamon Cheerios	General Mills	child	cold	bottom	220.00	1.5	2.00	4.00	4.0
6	Apple Jacks	Kellogg	child	cold	middle	165.00	1.5	1.50	3.00	0.0
7	Basic 4	General Mills	adult	cold	top	260.00	1.5	2.66	6.00	4.0

In [4]:

```
# Get all manufacturers in statcrunch dataset
cereal_man = list(statcrunch["Manufacturer"])
for i in range(len(cereal_man)):
    cereal_man[i] = cereal_man[i].lower()
cereal_man = list(set(cereal_man))
cereal_man
```

Out[4]:

```
['ralston purina',
 'post',
 'nabisco',
 'general mills',
 'kellogg',
 'quaker oats']
```

In [5]:

```
# USDA Products dataset, contains ingredients for many products
usda_products = pd.read_csv('./USDA-products.csv')
usda_products.fillna(-1, inplace=True)
usda_products.head()

usda_cereal_comp = pd.DataFrame(columns= usda_products.columns.values)

# Check if manufacturer in set of cereal manufacturers, if so append to usda_
for index, row in usda_products.iterrows():
    man = str(row['manufacturer']).lower()
    for cman in cereal_man:
        if cman in man:
            usda_cereal_comp = usda_cereal_comp.append(row, ignore_index = Tr
usda_cereal_comp
```

					LLC	13
1	45001990	EREWHON, STRAWBERRY CRISP	LI	41653012118	Post Foods, LLC	2017 22
2	45001993	ATTUNEFOODS, EREWHON, HONEY CRISPY BROWN RICE ...	LI	41653012101	Post Foods, LLC	2017 17
3	45002074	EREWHON, ORGANIC CINNAMON GRAHAMS, HONEY	LI	75940390009	Post Foods, LLC	2018 04
4	45004756	POST, HONEY BUNCHES OF OATS, FRUIT BLENDS CERE...	LI	884912002181	Post Consumer Brands, LLC	2018 02

In [6]:

```
# clean usda_cereal_names to have only cereal products

cereal_names = list(statcrunch["Name"])
reg_cereal_names = list(set(cereal_names))
for i in range(len(cereal_names)):
    cereal_names[i] = cereal_names[i].lower()
cereal_names = list(set(cereal_names))


usda_cereal_names = pd.DataFrame(columns= usda_products.columns.values)

for index, row in usda_cereal_comp.iterrows():
    cereal = row['long_name'].lower()
    for c in cereal_names:
        if (c in cereal):
            usda_cereal_names = usda_cereal_names.append(row, ignore_index =

usda_cereal_names
```

Out[6]:

	NDB_Number	long_name	data_source	gtin_upc	manufacturer	date_r
0	45001989	CORN FLAKES	LI	41653012293	Post Foods, LLC	20
1	45004756	POST, HONEY BUNCHES OF OATS, FRUIT BLENDS CERE...	LI	884912002181	Post Consumer Brands, LLC	20 (
2	45083031	SWEETENED PUFFED WHEAT CEREAL	LI	884912117625	Post Foods, LLC	20
3	45083061	SHREDDED WHEAT	LI	884912181701	Post Consumer Brands, LLC	20 2
		SHREDDED WHEAT SPOON			Post Foods	20

In [7]:

```
# remove duplicates w exact name matching
usda_cereal_names.drop_duplicates(['long_name'], inplace= True, keep= 'first'
print(usda_cereal_names.shape)
usda_cereal_names
```

```
(383, 8)
```

In [8]:

```
# add ingredients column to statscrunch
statcrunch["Ingredients"] = " "

for index, row in statcrunch.iterrows():
    name = row["Name"].lower()
    #print("Name:", name)
    for indx, ing_row in usda_cereal_names.iterrows():
        if(name in ing_row['long_name'].lower()):
            #print(ing_row['long_name'].lower())
            statcrunch['Ingredients'][index] = ing_row['ingredients_english']
statcrunch

#rename to clean_data
clean_data = statcrunch
clean_data
```

Out[8]:

Name	Manufacturer	Target	Type	Shelf	Calories	Cups	Weight	Protein	Fa

In [9]:

```
'''
For each “category” (Sugar, Fats etc.), we first normalize all the values to
For the undesirable categories like Sugar, Fats, Calories and Sodium,
where lower numbers are preferred, we reverse the values in that category
by getting the maximum in that category then doing (maximum - value) for all
From there, for each cereal we loop through the normalized values in each cat
and multiply each of them by the given gain (given from the input) and add al

'''


cal_data = clean_data.copy()

# reverse category
def rev(cata):
    #print('before', cata)
    result = []
    max_val = max(cata)
    #print(max_val)
    for x in range(0,len(cata)):
        result.append(max_val - cata[x])
    #print('after', cata)
    return result

# normalize category to 0-100
def norm(cata):
    max_val = max(cata)
    min_val = min(cata)
    nom_noms = max_val - min_val
    for x in range(0,len(cata)):
        cata[x] = 100*((cata[x] - min_val)/nom_noms)
    return cata


def ordered_cereal(ordered_elements):
    for col in ordered_elements:
        # if undesirable category, reverse values
        if col == 'Sugar' or col == 'Calories' or col == 'Fat' or col == 'Soc
            #call rev
            cal_data[col] = rev(clean_data[col])
        # call normalize for all categories
        cal_data[col] = norm(cal_data[col])

    #hard coded gains-testing
    gains = [.2,.2,.14,.17,.13,.1,.06]
    gains_dict = {}
    for i in range(0,len(ordered_elements)):
        gains_dict[ordered_elements[i]] = gains[i]

    result = [];
    for index, cereal in cal_data.iterrows():
        results = 0
        for cat in ordered_elements:
            #print(cal_data)
            if(math.isnan(cal_data[cat][index])):
                results += 0
            else:
```

```
        results += (cal_data[cat][index] * gains_dict[cat])

    result.append(results)

  # append all normalized categories
  df3 = pd.DataFrame(result)
  clean_data['calories_norm'] = cal_data['Calories']
  clean_data['sugars_norm'] = cal_data['Sugars']
  clean_data['fat_norm'] = cal_data['Fat']
  clean_data['protein_norm'] = cal_data['Protein']
  clean_data['fiber_norm'] = cal_data['Fiber']
  clean_data['vitamins_norm'] = cal_data['Vitamins']
  clean_data['sodium_norm'] = cal_data['Sodium']

  cal_data['health_score'] = result
  clean_data['Health Score'] = 0
  #cal_data.append(df3)

# sample order
cereal_order = ['Calories', 'Sugars', 'Fat', 'Protein', 'Fiber', "Vitamins",
#order_healthy_cereals(cereal_order)

ordered_cereal(cereal_order)

cal_data.sort_values(by =['health_score'], inplace= True, ascending = False)
#clean_data
cal_data
clean_data
```

```
C:\ProgramData\Anaconda3\lib\site-packages\ipykernel_launcher.py:30: Settin
gWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/
stable/indexing.html#indexing-view-versus-copy (http://pandas.pydata.org/pa
ndas-docs/stable/indexing.html#indexing-view-versus-copy)
```

Out[9]:

	Name	Manufacturer	Target	Type	Shelf	Calories	Cups	Weight	Protein	Fa
0	100% Bran	Nabisco	adult	cold	top	318.18	1.5	4.55	18.18	4.5
1	100% Natural Bran	Quaker Oats	adult	cold	top	180.00	1.5	1.50	4.50	7.5
2	All-Bran	Kellogg	adult	cold	top	318.18	1.5	4.55	18.18	4.5
3	All-Bran Extra Fiber	Kellogg	adult	cold	top	150.00	1.5	3.00	12.00	0.0
4	Almond Delight	Ralston Purina	adult	cold	top	220.00	1.5	2.00	4.00	4.0
5	Apple Cinnamon Cheerios	General Mills	child	cold	bottom	220.00	1.5	2.00	4.00	4.0

	Name	Manufacturer	Target	Type	Shelf	Calories	Cups	Weight	Protein	Fa
6	Apple Jacks	Kellogg	child	cold	middle	165.00	1.5	1.50	3.00	0.0
7	Basic 4	General Mills	adult	cold	top	260.00	1.5	2.66	6.00	4.0
8	Bran Chex	Ralston Purina	adult	cold	bottom	201.49	1.5	2.24	4.48	2.2
9	Bran Flakes	Post	adult	cold	top	201.49	1.5	2.24	6.72	0.0
10	Cap'n'Crunch	Quaker Oats	child	cold	middle	240.00	1.5	2.00	2.00	4.0
11	Cheerios	General Mills	child	cold	bottom	132.00	1.5	1.20	7.20	2.4
12	Cinnamon Toast Crunch	General Mills	child	cold	middle	240.00	1.5	2.00	2.00	6.0
13	Clusters	General Mills	adult	cold	top	330.00	1.5	3.00	9.00	6.0
14	Cocoa Puffs	General Mills	child	cold	middle	165.00	1.5	1.50	1.50	1.5
15	Corn Chex	Ralston Purina	adult	cold	bottom	165.00	1.5	1.50	3.00	0.0
16	Corn Flakes	Kellogg	adult	cold	bottom	150.00	1.5	1.50	3.00	0.0
17	Corn Pops	Kellogg	child	cold	middle	165.00	1.5	1.50	1.50	0.0
18	Count Chocula	General Mills	child	cold	middle	165.00	1.5	1.50	1.50	1.5
19	Cracklin' Oat Bran	Kellogg	adult	cold	top	330.00	1.5	3.00	9.00	9.0
20	Cream of Wheat (Quick)	Nabisco	adult	hot	middle	150.00	1.5	1.50	4.50	0.0

	Name	Manufacturer	Target	Type	Shelf	Calories	Cups	Weight	Protein	Fa
21	Crispix	Kellogg	adult	cold	top	165.00	1.5	1.50	3.00	0.0
22	Crispy Wheat & Raisins	General Mills	adult	cold	top	200.00	1.5	2.00	4.00	2.0
23	Double Chex	Ralston Purina	adult	cold	top	200.00	1.5	2.00	4.00	0.0
24	Froot Loops	Kellogg	child	cold	middle	165.00	1.5	1.50	3.00	1.5
25	Frosted Flakes	Kellogg	child	cold	bottom	220.00	1.5	2.00	2.00	0.0
26	Frosted Mini-Wheats	Kellogg	adult	cold	middle	187.50	1.5	1.88	5.62	0.0
27	Fruit & Fibre	Post	adult	cold	top	268.66	1.5	2.80	6.72	4.4
28	Fruitful Bran	Kellogg	adult	cold	top	268.66	1.5	2.98	6.72	0.0
29	Fruity Pebbles	Post	child	cold	middle	220.00	1.5	2.00	2.00	2.0
...	...	...	...	...	...	...	...	...	...	..
46	Multi-Grain Cheerios	General Mills	adult	cold	bottom	150.00	1.5	1.50	3.00	1.5
47	Nut&Honey Crunch	Kellogg	adult	cold	middle	268.66	1.5	2.24	4.48	2.2
48	Nutri-Grain Almond-Raisin	Kellogg	adult	cold	top	313.43	1.5	2.98	6.72	4.4
49	Nutri-grain Wheat	Kellogg	adult	cold	top	135.00	1.5	1.50	4.50	0.0
50	Oatmeal Raisin Crisp	General Mills	adult	cold	top	390.00	1.5	3.75	9.00	6.0
51	Post Nat. Raisin Bran	Post	adult	cold	top	268.66	1.5	2.98	6.72	2.2
52	Product 19	Kellogg	adult	cold	top	150.00	1.5	1.50	4.50	0.0
53	Puffed Rice	Quaker Oats	adult	cold	top	75.00	1.5	0.75	1.50	0.0

	Name	Manufacturer	Target	Type	Shelf	Calories	Cups	Weight	Protein	Fa
54	Puffed Wheat	Quaker Oats	adult	cold	top	75.00	1.5	0.75	3.00	0.0
55	Quaker Oat Squares	Quaker Oats	adult	cold	top	300.00	1.5	3.00	12.00	3.0
56	Quaker Oatmeal	Quaker Oats	adult	hot	bottom	223.88	1.5	2.24	11.19	4.4
57	Raisin Bran	Kellogg	adult	cold	middle	240.00	1.5	2.66	6.00	2.0
58	Raisin Nut Bran	General Mills	adult	cold	top	300.00	1.5	3.00	9.00	6.0
59	Raisin Squares	Kellogg	adult	cold	top	270.00	1.5	3.00	6.00	0.0
60	Rice Chex	Ralston Purina	adult	cold	bottom	146.02	1.5	1.33	1.33	0.0
61	Rice Krispies	Kellogg	child	cold	bottom	165.00	1.5	1.50	3.00	0.0
62	Shredded Wheat	Nabisco	adult	cold	bottom	120.00	1.5	1.24	3.00	0.0
63	Shredded Wheat 'n'Bran	Nabisco	adult	cold	bottom	201.49	1.5	2.24	6.72	0.0
64	Shredded Wheat spoon size	Nabisco	adult	cold	bottom	201.49	1.5	2.24	6.72	0.0
65	Smacks	Kellogg	child	cold	middle	220.00	1.5	2.00	4.00	2.0
66	Special K	Kellogg	adult	cold	bottom	165.00	1.5	1.50	9.00	0.0
67	Strawberry Fruit Wheats	Nabisco	child	cold	middle	135.00	1.5	1.50	3.00	0.0
68	Total Corn Flakes	General Mills	adult	cold	top	165.00	1.5	1.50	3.00	1.5

	Name	Manufacturer	Target	Type	Shelf	Calories	Cups	Weight	Protein	Fa
69	Total Raisin Bran	General Mills	adult	cold	top	210.00	1.5	2.25	4.50	1.5
70	Total Whole Grain	General Mills	adult	cold	top	150.00	1.5	1.50	4.50	1.5
71	Triples	General Mills	adult	cold	top	220.00	1.5	2.00	4.00	2.0
72	Trix	General Mills	child	cold	middle	165.00	1.5	1.50	1.50	1.5
73	Wheat Chex	Ralston Purina	adult	cold	bottom	223.88	1.5	2.24	6.72	2.2
74	Wheaties	General Mills	child	cold	bottom	150.00	1.5	1.50	4.50	1.5
75	Wheaties Honey Gold	General Mills	child	cold	bottom	220.00	1.5	2.00	4.00	2.0

76 rows × 26 columns

In [10]:

```
#Given Code
#import requests
import pickle
import pandas as pd
from time import sleep, time
from random import randint
from bs4 import BeautifulSoup
from IPython.core.display import clear_output
```

In [11]:

```
# scrapping method, provide the raw path
def get_products(raw_path):
    url = open(raw_path,encoding="utf8")
    #print(url)

    page_html = BeautifulSoup(url.read())

    # get all products
    containers = page_html.find_all( class_ = 'search-result-gridview-items f
    # list items
    bk_containers = containers[0].find_all('li', class_ = 'Grid-col')

    titles=[]
    ratings=[]
    counts = []
    A_prices = []

    #print(len(bk_containers))
    for i in bk_containers:
        title = i.find(attrs = {'data-type':'itemTitles'}).get_text()
        #print(title)
        titles.append(title)

        rating = i.find('span', class_ = 'seo-avg-rating').get_text()
        #print(rating)
        ratings.append(rating)

        count = i.find('span', class_ = 'seo-review-count').get_text()
        #print(count)
        counts.append(count)

        price = i.find('span', class_ = 'price-main-block')
        #print(price)

        A_price = price.find('span', class_ = 'visuallyhidden').get_text()
        #print(A_price)
        A_prices.append(A_price)

    return {0 : titles, 1: ratings, 2: counts, 3: A_prices}
```

In [12]:

```
# SCRAPE ORGANIC FROM WALMART
titles_organic = []
ratings_organic = []
counts_organic = []
A_prices_organic = []

lst_org = get_products(r'HTMLpages/oraganic_0.html')
titles_organic += lst_org[0]
ratings_organic += lst_org[1]
counts_organic += lst_org[2]
A_prices_organic += lst_org[3]


lst_org_two = get_products(r'HTMLpages/oraganic_1.html')
titles_organic += lst_org_two[0]
ratings_organic += lst_org_two[1]
counts_organic += lst_org_two[2]
A_prices_organic += lst_org_two[3]

lst_org_three = get_products(r'HTMLpages/oraganic_2.html')
titles_organic += lst_org_three[0]
ratings_organic += lst_org_three[1]
counts_organic += lst_org_three[2]
A_prices_organic += lst_org_three[3]

# make dataframe
organic = pd.DataFrame({'Title': titles_organic,
'Rating': ratings_organic,
'Rating Count': counts_organic,
'Price': A_prices_organic,
})
organic
```

Out[12]:

	Title	Rating	Rating Count	Price
0	Annie's Certified Organic Cocoa Bunnies Cereal...	4.9	19	$3.28
1	Kashi Dark Cocoa Karma Breakfast Cereal 16.1 oz	4.9	13	$2.98
2	Nature's Path Organic Granola Pumpkin Seed & F...	4.9	169	$2.98
3	Kashi by Kids Honey Cinnamon Super Food Combos...	4	3	$3.68
4	Cascadian Farm Organic Granola Oats & Honey Ce...	4.7	149	$2.78
5	Love Crunch Organic Granola Dark Chocolate & R...	4.8	186	$3.28
6	Cascadian Farm Organic Cereal, Fruitful O's, 1...	4.5	12	$3.28
7	Kashi Heart to Heart Breakfast Oat Cereal Warm...	3.8	6	$2.99
8	Nature's Path Organic Heritage Flakes Cereal, ...	4.5	189	$7.12
9	Kashi by Kids Super Food Combos Organic Cocoa ...	4.7	3	$3.68
10	Cascadian Farm Organic Berry Vanilla Cereal, 1...	4.5	11	$2.99
11	Love Crunch Organic Granola Dark Chocolate & P...	4.9	77	$3.87
12	Cascadian Farm Lemon Blueberry Granola, 11.5 o...	5	8	$3.98
13	Cascadian Farm Organic Strawberry Granola, 10....	5	2	$3.98

	Title	Rating	Rating Count	Price
14	Nature's Path Organic Gluten-Free Cereal Mesa ...	4.8	94	$7.12
15	Natures Path Organic Gluten-Free Breakfast Cer...	5	2	$6.27
16	Kashi Kids Bites Berry Organic Snack Bites 5.6 oz	4.2	5	$3.48
17	Great Value Organic Breakfast Cereal, Toasted ...	5	8	$2.98
18	Bear Naked Organic White Chocolate Macademia G...	4.2	6	$5.98
19	Bear Naked Organic Chocolate Hazelnut Granola ...	5	2	$15.99
20	Cascadian Farm Organic Raisin Bran Cereal, 12 oz.	4.5	23	$3.58
21	Great Value Organic Honey Crunch & Oats Cereal...	0	0	$2.98
22	(2 Pack) Kashi Heart to Heart Organic Oat Cere...	4.5	57	$5.42
23	Cascadian Farm Organic Morning Fiber Cereal, 1...	4	10	$2.99
24	Cascadian Farm Organic Multi Grain Cereal, 12....	4	6	$3.81
25	Cascadian Farm Organic Cereal, Honey Nut O's, ...	3.4	5	$2.99
26	Kashi Sprouted Grain Breakfast Cereal 9.5 oz box	4.8	11	$3.64
27	Cascadian Farm Organic Graham Crunch Cereal, 9...	4.4	7	$2.99
28	Food To Live Certified Organic Buckwheat Groa...	5	12	$16.99
29	Nature's Path Whole Os Organic Cereal Gluten F...	4.4	60	$7.12
...	...	...	...	...
90	Nature's Path Organic EnviroKidz Koala Crisp C...	4.7	18	$33.26
91	Nature's Path Qia Super Flakes Cereal, Cocoa C...	4.6	28	$7.10
92	Envirokidz Organic Cereal - Koala Crisp - Pack...	0	0	$85.16
93	Made Good Granola Minis - Chocolate Chip - pac...	0	0	$52.67
94	Made Good Granola Minis - Apple Cinnamon - pac...	0	0	$90.65
95	Golden Temple Granola Organic Granola - Fruit ...	0	0	$98.78
96	Arrowhead Mills Organic Spelt Flakes - Pack of...	0	0	$95.10
97	Kashi Breakfast Cereal, Autumn Wheat, 16.3 Oz	4.1	11	$12.99
98	Maker Overnight Oats - Banana and Coffee - Ca...	0	0	$105.31
99	Arrowhead Mills Organic Gluten Free Cereal - S...	0	0	$71.49
100	Arrowhead Mills Cereal - Rice And Shine - Glut...	0	0	$63.95
101	Love Crunch Organic Granola Apple Crumble 11.5 oz	4.9	20	$5.10
102	Weetabix Organic Cereal - Case of 12 - 14 oz.	0	0	$113.55
103	Arrowhead Mills Organic Gluten Free Cereal, Sp...	0	0	$9.80
104	6 Pack : One Degree Organic Foods Sprout...	0	0	$44.30
105	Cascadian Farm Granola, French Vanilla Almond,...	0	0	$17.32
106	Nature's Path Natures Path Organic Gluten Free...	4.7	60	$4.81
107	New England Naturals Organic High Protein Gran...	5	2	$7.52
108	Nature's Path Organic Flax Plus Red Berry Crun...	0	0	$68.62

	Title	Rating	Rating Count	Price
109	Love Crunch Apple Crumble Premium Organic Gran...	0	0	$33.42
110	EnviroKidz Choco Chimps Organic Cereal Chocola...	0	0	$70.98
111	Nature's Path Organic Flax Plus Raisin Bran Ce...	5	1	$57.39
112	Bob's Red Mill, Organic Whole Grain Oat Groats...	0	0	$29.22
113	en Free Selections Sunrise Crunchy Honey Cerea...	4.3	15	$34.97
114	Jovial Organic Einkorn Wheat Berries, 16.0 Ounce	0	0	$8.10
115	Bytewise Organic Puffed Rice Cereal / Murmure,...	0	0	$8.00
116	Evoke Non-GMO Muesli, Antioxidant, Goji Berrie...	0	0	$5.99
117	Nature's Path Organic Qia Original, 7.9 OZ	4.8	30	$6.98
118	One Degree Organic Foods Granola, Sprouted Org...	0	0	$44.61
119	Grandy Oats Coconut Granola Super Hemp Blend, ...	0	0	$17.18

120 rows × 4 columns

In [13]:

```python
# SCRAPE VEGAN FROM WALMART
titles_vegan = []
ratings_vegan = []
counts_vegan = []
A_prices_vegan = []

lst_veg = get_products(r'HTMLpages/vegan_0.html')
titles_vegan += lst_veg[0]
ratings_vegan += lst_veg[1]
counts_vegan += lst_veg[2]
A_prices_vegan += lst_veg[3]


lst_veg_two = get_products(r'HTMLpages/vegan_1.html')
titles_vegan += lst_veg_two[0]
ratings_vegan += lst_veg_two[1]
counts_vegan += lst_veg_two[2]
A_prices_vegan += lst_veg_two[3]

lst_veg_three = get_products(r'HTMLpages/oraganic_2.html')
titles_vegan += lst_veg_three[0]
ratings_vegan += lst_veg_three[1]
counts_vegan += lst_veg_three[2]
A_prices_vegan += lst_veg_three[3]

#make dataframe
vegan = pd.DataFrame({'Title': titles_vegan,
'Rating': ratings_vegan,
'Rating Count': counts_vegan,
'Price': A_prices_vegan,
})
vegan
```

Out[13]:

	Title	Rating	Rating Count	Price
0	Kashi GOLEAN Breakfast Cereal Chocolate Crunch...	4.4	15	$2.92
1	Kashi by Kids Super Food Combos Organic Cocoa ...	4.7	3	$3.68
2	Kashi GOLEAN Toasted Berry Crisp Breakfast Cer...	4.6	37	$2.77
3	Kashi Berry Fruitful Breakfast Cereal 15.6 oz box	4.6	38	$2.98
4	Kashi Breakfast Cereal Cinnamon French Toast 1...	0	0	$3.28
5	Kashi Golean Crunch Peanut Butter Breakfast Ce...	4.9	7	$3.28
6	(2 Pack) Kashi 7 Whole Grain Non-GMO Breakfast...	0	0	$5.78
7	(2 Pack) Kashi Organic Biscuits Breakfast Cere...	4.8	49	$5.78
8	Kashi Sprouted Grain Breakfast Cereal 9.5 oz box	4.8	11	$3.64
9	(2 Pack) Kashi Organic Breakfast Cereal, Straw...	5	1	$7.07
10	Food To Live Certified Organic Buckwheat Groa...	5	12	$16.99
11	Nature's Path Organic Chia Plus Coconut Chia G...	4.9	86	$28.88
12	Nature's Path Organic Flax Plus Multibran Flak...	4.4	72	$3.64
13	Arrowhead Mills Puffed Rice Cereal, 6 oz, (Pac...	4.5	13	$33.48

	Title	Rating	Rating Count	Price
14	Arrowhead Mills Puffed Rice Breakfast Cereal, ...	4.5	36	$32.74
15	Purely Elizabeth Original Ancient Grain Granol...	0	0	$37.37
16	Made Good Granola Minis, Strawberry, 3.4 Oz, 6 Ct	3	1	$25.38
17	Barbara's Bakery Shredded Wheat Cereal, 13 oz,...	4.5	2	$29.95
18	Love Grown Chocolate Power O's, 10 Oz, Box, 6-...	4	2	$4.21
19	Love Grown Sea Stars Cereal, 7 Oz, Box, 6-Pack...	4.7	3	$4.48
20	Nature's Path Organic Optimum Power Blueberry ...	4.3	74	$31.66
21	UNCLE SAM ORIGINAL CEREAL (UNIT)	5	2	$14.79
22	Purely Elizabeth Original Grain-Free Granola, ...	5	1	$49.72
23	Nature's Path Organic Fruit Juice Sweetened Co...	4.6	32	$46.32
24	Arrowhead Mills Organic Spelt Flakes, 12 oz (P...	0	0	$55.36
25	Arrowhead Mills Puffed Corn Cereal, 6 oz, (Pac...	4.5	6	$23.88
26	, Granola minis, Og2, Strawberry, Pack of 6, S...	0	0	$37.89
27	Uncle Sam Cereal Cereal - Original - Family Si...	0	0	$96.36
28	Love Grown Power O's Chocolate Cereal, 10 oz, ...	0	0	$29.94
29	Made Good Granola Minis, Apple Cinnamon, 3.4 Oz	0	0	$25.38
...	...	...	...	...
90	Nature's Path Organic EnviroKidz Koala Crisp C...	4.7	18	$33.26
91	Nature's Path Qia Super Flakes Cereal, Cocoa C...	4.6	28	$7.10
92	Envirokidz Organic Cereal - Koala Crisp - Pack...	0	0	$85.16
93	Made Good Granola Minis - Chocolate Chip - pac...	0	0	$52.67
94	Made Good Granola Minis - Apple Cinnamon - pac...	0	0	$90.65
95	Golden Temple Granola Organic Granola - Fruit ...	0	0	$98.78
96	Arrowhead Mills Organic Spelt Flakes - Pack of...	0	0	$95.10
97	Kashi Breakfast Cereal, Autumn Wheat, 16.3 Oz	4.1	11	$12.99
98	Maker Overnight Oats - Banana and Coffee - Ca...	0	0	$105.31
99	Arrowhead Mills Organic Gluten Free Cereal - S...	0	0	$71.49
100	Arrowhead Mills Cereal - Rice And Shine - Glut...	0	0	$63.95
101	Love Crunch Organic Granola Apple Crumble 11.5 oz	4.9	20	$5.10
102	Weetabix Organic Cereal - Case of 12 - 14 oz.	0	0	$113.55
103	Arrowhead Mills Organic Gluten Free Cereal, Sp...	0	0	$9.80
104	6 Pack : One Degree Organic Foods Sprout...	0	0	$44.30
105	Cascadian Farm Granola, French Vanilla Almond,...	0	0	$17.32
106	Nature's Path Natures Path Organic Gluten Free...	4.7	60	$4.81
107	New England Naturals Organic High Protein Gran...	5	2	$7.52
108	Nature's Path Organic Flax Plus Red Berry Crun...	0	0	$68.62

	Title	Rating	Rating Count	Price
109	Love Crunch Apple Crumble Premium Organic Gran...	0	0	$33.42
110	EnviroKidz Choco Chimps Organic Cereal Chocola...	0	0	$70.98
111	Nature's Path Organic Flax Plus Raisin Bran Ce...	5	1	$57.39
112	Bob's Red Mill, Organic Whole Grain Oat Groats...	0	0	$29.22
113	en Free Selections Sunrise Crunchy Honey Cerea...	4.3	15	$34.97
114	Jovial Organic Einkorn Wheat Berries, 16.0 Ounce	0	0	$8.10
115	Bytewise Organic Puffed Rice Cereal / Murmure,...	0	0	$8.00
116	Evoke Non-GMO Muesli, Antioxidant, Goji Berrie...	0	0	$5.99
117	Nature's Path Organic Qia Original, 7.9 OZ	4.8	30	$6.98
118	One Degree Organic Foods Granola, Sprouted Org...	0	0	$44.61
119	Grandy Oats Coconut Granola Super Hemp Blend, ...	0	0	$17.18

120 rows × 4 columns

In [14]:

```python
# SCRAPE GLUTEN FROM WALMART
titles_gluten = []
ratings_gluten = []
counts_gluten = []
A_prices_gluten = []

lst_glu = get_products(r'HTMLpages/gluten_0.html')
titles_gluten += lst_glu[0]
ratings_gluten += lst_glu[1]
counts_gluten += lst_glu[2]
A_prices_gluten += lst_glu[3]


lst_glu_two = get_products(r'HTMLpages/gluten_1.html')
titles_gluten += lst_glu_two[0]
ratings_gluten += lst_glu_two[1]
counts_gluten += lst_glu_two[2]
A_prices_gluten += lst_glu_two[3]

lst_glu_three = get_products(r'HTMLpages/gluten_2.html')
titles_gluten += lst_glu_three[0]
ratings_gluten += lst_glu_three[1]
counts_gluten += lst_glu_three[2]
A_prices_gluten += lst_glu_three[3]

# make dataframe
gluten = pd.DataFrame({'Title': titles_gluten,
'Rating': ratings_gluten,
'Rating Count': counts_gluten,
'Price': A_prices_gluten,
})
gluten
```

Out[14]:

	Title	Rating	Rating Count	Price
0	Cheerios, Gluten Free, Breakfast Cereal, 18 oz...	4.8	406	$3.64
1	Post Fruity Pebbles Gluten Free Breakfast Cere...	4.7	62	$5.98
2	Apple Cinnamon Cheerios, Gluten Free Cereal, 2...	4.7	37	$3.64
3	Cinnamon Chex Cereal, Gluten Free, 19.6 oz	4.5	8	$3.00
4	Post Cocoa Pebbles Gluten Free Breakfast Cerea...	5	5	$5.98
5	Rice Chex Cereal, Gluten Free, 18 oz	4.8	60	$3.00
6	Very Berry Cheerios Cereal, Gluten Free, 19.5 oz	4.5	12	$3.64
7	Corn Chex Cereal, Gluten Free, 18 oz	4.7	31	$3.00
8	Chocolate Chex Cereal, Gluten Free, 21.1 oz	4.9	10	$3.00
9	Lucky Charms Gluten Free Breakfast Cereal, 32 ...	5	3	$5.98
10	Honey Nut Chex Cereal, Gluten Free, 20.3 oz	4.4	7	$3.00
11	Frosted Cheerios Cereal, Gluten Free, 19.5 oz	4.7	11	$3.53
12	Chocolate Peanut Butter Cheerios, Cereal, 20.3 oz	4.8	39	$3.64
13	Maple Cheerios Cereal, Gluten Free, 19.8 oz	4.9	38	$3.64

	Title	Rating	Rating Count	Price
14	Malt-O-Meal Breakfast Cereal, Crispy Rice, 36 ...	4.3	34	$4.98
15	Fruity Cheerios, Cereal with Oats, Gluten Free...	4.4	52	$3.52
16	Post Peanut Butter & Cocoa Pebbles Breakfast C...	3.4	15	$3.98
17	Malt-O-Meal Gluten Free Cereal, Fruity Dyno Bi...	4.5	18	$8.12
18	Multi Grain Cheerios Gluten Free Cereal, 9 oz Box	4.8	24	$2.98
19	Multi Grain Cheerios Gluten Free Multigrain Ce...	4.8	64	$3.64
20	KIND Gluten Free Breakfast Granola, Oats, Hone...	4.7	18	$3.98
21	Lucky Charms Gluten Free Breakfast Cereal, 10....	4.6	5	$2.98
22	Honey Nut Cheerios Gluten Free Cereal, 15.4 oz...	5	1	$3.49
23	Cheerios Cups, Gluten Free Cereal, Whole Grain...	5	1	$1.48
24	Fruity Cheerios, Cereal with Oats, Gluten Free...	4.4	52	$3.29
25	Cascadian Farm Organic Berry Vanilla Cereal, 1...	4.5	11	$2.99
26	Malt-O-Meal Gluten Free Breakfast Cereal, Coco...	4.8	121	$8.12
27	Lucky Charms, Gluten Free, Cereal, Family Size...	4.8	16	$7.00
28	Udis Au Naturel Gluten Free Granola Wildflower...	4.3	15	$4.48
29	Kind Cinnamon Oat Clusters Granola, 11 Oz, Pac...	5	11	$29.15
...	...	...	...	...
90	Bakery on Main Gourmet Naturals Gluten Free Ap...	0	0	$40.10
91	WholeMe Cinnamon Banana Chip Clusters, 8 oz	0	0	$45.09
92	6 PACKS : Udis Gluten Free Granola, Cranberry,...	0	0	$52.25
93	Envirokidz Organic Amazon Frosted Flakes Cerea...	0	0	$53.86
94	Flax4Life Gluten Free Flax Cranberry Orange Sn...	0	0	$36.30
95	Back To Nature Almond Chia Clusters Granola, 1...	0	0	$36.78
96	Rhinestone Bow Shirts White M (12)	0	0	$13.58
97	Purely Elizabeth Probiotic Granola Gluten Free...	0	0	$14.13
98	Barbara's Honest O's Cereal. Original, 8 Oz	0	0	$6.68
99	Love Grown Foods Strawberry Raspberry Hot Oats...	0	0	$12.18
100	Freedom Foods Pro Teen Crunch 10.6 Ounce	0	0	$31.72
101	Flax4Life Gluten Free Flax Banana Coconut Snac...	5	1	$44.73
102	Purely Elizabeth Original Ancient Grain Granol...	0	0	$43.16
103	General Mills Chocolate Cheerious Gluten Free ...	0	0	$24.99
104	Love Grown Cocoa Goodness Oat Clusters, 12 oz....	0	0	$42.62
105	Bakery on Main Triple Berry Fiber Power Granol...	0	0	$36.12
106	Modern Oats Mango Blackberry Oatmeal, 2.6 Oz, ...	0	0	$28.16
107	Bakery on Main Maple Multigrain Muffin Instant...	0	0	$36.04
108	Love Grown Strawberry Raspberry Hot Oats, 2.22...	0	0	$22.56

	Title	Rating	Rating Count	Price
109	Love Grown Raisin Almond Crunch Oat Clusters, ...	5	1	$7.97
110	Honey Nut Cheerios, Gluten Free	0	0	$19.06
111	Bakery On Main Variety Pack Instant Oatmeal, 1...	0	0	$10.62
112	Purely Elizabeth Probiotic Granola Gluten Free...	0	0	$26.62
113	Nature`S Path Flax Plus With Cinnamon 32 Oz	4.9	21	$61.02
114	General Mills Apple Cinnamon Cheerios Gluten F...	0	0	$24.99
115	Gluten Free Cereal Mix (4 oz, ZIN: 524846)	0	0	$3.90
116	Gluten Free Cereal Mix (8 oz, ZIN: 524847)	0	0	$5.90
117	Gluten Free Cereal Mix (16 oz, ZIN: 524848)	0	0	$6.39
118	Gluten Free Cereal Mix (4 oz, ZIN: 524846) - 2...	0	0	$7.49
119	Kay's Naturals Protein Cereal French Vanilla 1...	0	0	$9.99

120 rows × 4 columns

In [15]:

```
# for all the categories, add as a row in clean data
clean_data["Organic"] = False
for index,row in clean_data.iterrows():
    clean_data_name = row['Name'].lower()
    for idx,rw in organic.iterrows():
        organic_name = rw['Title'].lower()
        if(clean_data_name in organic_name or organic_name in clean_data_name
            clean_data['Organic'][index] = True

clean_data["Vegan"] = False
for index,row in clean_data.iterrows():
    clean_data_name = row['Name'].lower()
    for idx,rw in vegan.iterrows():
        vegan_name = rw['Title'].lower()
        if(clean_data_name in vegan_name or vegan_name in clean_data_name):
            clean_data['Vegan'][index] = True

clean_data["Gluten Free"] = False
for index,row in clean_data.iterrows():
    clean_data_name = row['Name'].lower()
    for idx,rw in gluten.iterrows():
        gluten_name = rw['Title'].lower()
        if(clean_data_name in gluten_name or gluten_name in clean_data_name):
            clean_data['Gluten Free'][index] = True

clean_data.sort_values(by =['Health Score'], inplace= True, ascending = False
clean_data
```

```
C:\ProgramData\Anaconda3\lib\site-packages\ipykernel_launcher.py:8: Setti
ngWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-doc
s/stable/indexing.html#indexing-view-versus-copy (http://pandas.pydata.or
g/pandas-docs/stable/indexing.html#indexing-view-versus-copy)

C:\ProgramData\Anaconda3\lib\site-packages\ipykernel_launcher.py:16: Sett
ingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-doc
s/stable/indexing.html#indexing-view-versus-copy (http://pandas.pydata.or
g/pandas-docs/stable/indexing.html#indexing-view-versus-copy)
  app.launch_new_instance()
C:\ProgramData\Anaconda3\lib\site-packages\ipykernel_launcher.py:24: Sett
ingWithCopyWarning:
```

In [16]:

```
# INGREDIENT ANALYSIS
# attempt in analyzing ingredients

# in top 50% healthiest cereals, common ingredients of the top 5 ingredients
ingredients_ordered = clean_data["Ingredients"]
ing_ord_len = int(len(ingredients_ordered)/2)
ingredients_top_50 = ingredients_ordered[0:ing_ord_len]
ingredients_bottom_50 = ingredients_ordered[ing_ord_len::]
cleaned_ing_top = []
cleaned_ing_bottom = []
for ing in ingredients_top_50:
    if(ing != " "):
        lst = ing.split(",")
        for i in range(0,len(lst)):
            lst[i] = lst[i].strip()
        cleaned_ing_top.append(lst)

for ing in ingredients_bottom_50:
    if(ing != " "):
        lst2 = ing.split(",")
        for i in range(0,len(lst2)):
            lst2[i] = lst2[i].strip()
        cleaned_ing_bottom.append(lst2)

top_50_dict = {}
for ing in cleaned_ing_top:
    # get top 5 ingredients for each product
    for idv in ing[0:5]:
        if(top_50_dict.get(idv) == None):
            top_50_dict[idv] = 1
        else:
            top_50_dict[idv] += 1

bottom_50_dict = {}
for ing in cleaned_ing_bottom:
    for idv in ing[0:5]:
        if(bottom_50_dict.get(idv) == None):
            bottom_50_dict[idv] = 1
        else:
            bottom_50_dict[idv] += 1

print(top_50_dict)
print(bottom_50_dict)
```

```
{'sugar': 13, 'wheat': 2, 'dextrose': 2, 'honey': 2, 'contains 2% or less
of vegetable oil (hydrogenated or partially hydrogenated soybean)': 2, 'p
uffed rice': 1, 'ferrous sulfate (a source of iron)': 1, 'niacinamide*':
1, 'citric acid': 1, 'thiamin mononitrate*': 1, 'milled corn': 1, 'whole
grain oat flour': 2, 'wheat flour': 1, 'rice': 4, 'corn flour': 1, 'whole
wheat flour': 1, 'rice flour': 1, 'whole grain corn': 1, 'corn meal': 1,
'salt': 4, 'brown sugar syrup': 1, 'whole grain wheat': 4, 'raisins': 2,
'wheat bran': 1, 'corn syrup. vitamin e (mixed tocopherols) added to pres
erve freshness.vitamins and minerals: calcium carbonate': 1, 'zinc and ir
on (mineral nutrients)': 1, 'rice chex : whole grain rice': 1, 'molasses.
vitamin e (mixed tocopherols) added to preserve freshness.vitamins and mi
nerals: calcium carbonate': 2, 'enriched flour bleached (wheat flour': 1,
'malted barley flour': 1, 'niacin': 1, 'ferrous sulfate': 1, 'thiamin mon
```

```
onitrate': 1, 'corn bran': 1, 'corn syrup': 2, 'soy protein isolate': 1,
'soluble corn fiber': 1, 'peanuts': 1, 'fructose': 2, 'contains 2% or les
s of: natural and artificial flavor': 1, 'gelatin': 1, 'red 40': 1, 'whol
e grain brown rice': 1, 'vegetable oil (soybean and palm oil with tbhq fo
r freshness)': 1, 'whole grain rice': 1}
{'cereal (whole grain corn': 1, 'sugar': 19, 'corn meal': 4, 'corn syru
p': 5, 'cocoa processed with alkali': 2, 'milled corn': 4, 'malt flavor':
1, '2% or less of salt. bht for freshness. vitamins and minerals: iron':
1, 'vitamin c (sodium ascorbate': 1, 'whole grain corn': 1, 'whole grain
oats': 3, 'whole grain wheat': 7, 'yellow corn grits': 1, 'rice flour':
3, 'canola oil': 2, 'fructose': 1, 'corn starch': 1, 'brown sugar syrup':
4, 'salt': 4, 'raisins': 2, 'wheat bran': 2, 'corn flour blend (whole gra
in yellow corn flour': 1, 'degerminated yellow corn flour)': 1, 'wheat fl
our': 2, 'whole grain oat flour': 2, 'contains 2% or less of salt': 2, 'c
ontains 2% or less of molasses': 1, 'cluster (whole grain oats': 1, 'ric
e': 2, 'hydrogenated vegetable oil (coconut and palm kernel oils)': 1, 'c
ontains less than 0.5% of natural and artificial flavor': 1, 'contains 2%
or less of milled corn': 1, 'brown rice syrup': 1, 'contains 2% or less o
f malt flavor': 1, 'bht for freshness. vitamins and minerals: iron': 1,
'whole grain yellow corn flour': 1, 'oat fiber': 1, 'molasses': 1}
```

In [17]:

```
clean = clean_data.to_json(orient='records')
clean
```

Out[17]:

```
'[{"Name":"100% Bran","Manufacturer":"Nabisco","Target":"adult","Type":"c
old","Shelf":"top","Calories":318.18,"Cups":1.5,"Weight":4.55,"Protein":1
8.18,"Fat":4.55,"Sodium":590.91,"Fiber":45.45,"Carbs":22.73,"Sugars":27.2
7,"Potassium":1272.73,"Vitamins":113.64,"Rating":68,"Ingredients":" ","ca
lories_norm":58.4307692308,"sugars_norm":87.013401404,"fat_norm":66.64222
8739,"protein_norm":100.0,"fiber_norm":100.0,"vitamins_norm":56.82,"sodiu
m_norm":50.0,"Health Score":0,"Organic":false,"Vegan":false,"Gluten Fre
e":false},{"Name":"Nutri-Grain Almond-Raisin","Manufacturer":"Kellogg","T
arget":"adult","Type":"cold","Shelf":"top","Calories":313.43,"Cups":1.
5,"Weight":2.98,"Protein":6.72,"Fat":4.48,"Sodium":492.54,"Fiber":6.72,"C
arbs":47.01,"Sugars":15.67,"Potassium":291.04,"Vitamins":55.97,"Rating":4
1,"Ingredients":" ","calories_norm":59.2427350427,"sugars_norm":50.0,"fat
_norm":67.1554252199,"protein_norm":32.7859237537,"fiber_norm":14.7854785
479,"vitamins_norm":27.985,"sodium_norm":58.3236025791,"Health Score":
0,"Organic":false,"Vegan":false,"Gluten Free":false},{"Name":"Quaker Oat
Squares","Manufacturer":"Quaker Oats","Target":"adult","Type":"cold","She
lf":"top","Calories":300.0,"Cups":1.5,"Weight":3.0,"Protein":12.0,"Fat":
3.0,"Sodium":405.0,"Fiber":6.0,"Carbs":42.0,"Sugars":18.0,"Potassium":33
0.0,"Vitamins":75.0,"Rating":50,"Ingredients":" ","calories_norm":61.5384
615385,"sugars_norm":57.4345883854,"fat_norm":78.0058651026,"protein_nor
```

In [ ]: