Mcdonald's Case_Study

In this file lot of attributes are given and basis that we will do neccassry EDA.

Load the necessary libraries. Import and load the dataset with a Mcd

.

[1]:
```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
sns.set(color_codes=True)
```

1.Let us see the first 10 records of the dataset to know what are the variables.

[2]:
```python
Mcd=pd.read_csv('Mcdonald .csv')
Mcd.head()
```

Out[2]:

	Category	Item	Serving Size	Calories	Calories from Fat	Total Fat	Total Fat (% Daily Value)	Saturated Fat	Saturated Fat (% Daily Value)	Trans Fat	...
0	Breakfast	Egg McMuffin	4.8 oz (136 g)	300	120	13.0	20	5.0	25	0.0	...
1	Breakfast	Egg White Delight	4.8 oz (135 g)	250	70	8.0	12	3.0	15	0.0	...
2	Breakfast	Sausage McMuffin	3.9 oz (111 g)	370	200	23.0	35	8.0	42	0.0	...
3	Breakfast	Sausage McMuffin with Egg	5.7 oz (161 g)	450	250	28.0	43	10.0	52	0.0	...
4	Breakfast	Sausage McMuffin with Egg Whites	5.7 oz (161 g)	400	210	23.0	35	8.0	42	0.0	...

5 rows × 24 columns

2. The summary of the dataset

> Count ,Mean, Standard deviation, percentile and min_max value

[3]:
```python
Mcd.describe()
```

Out[3]:

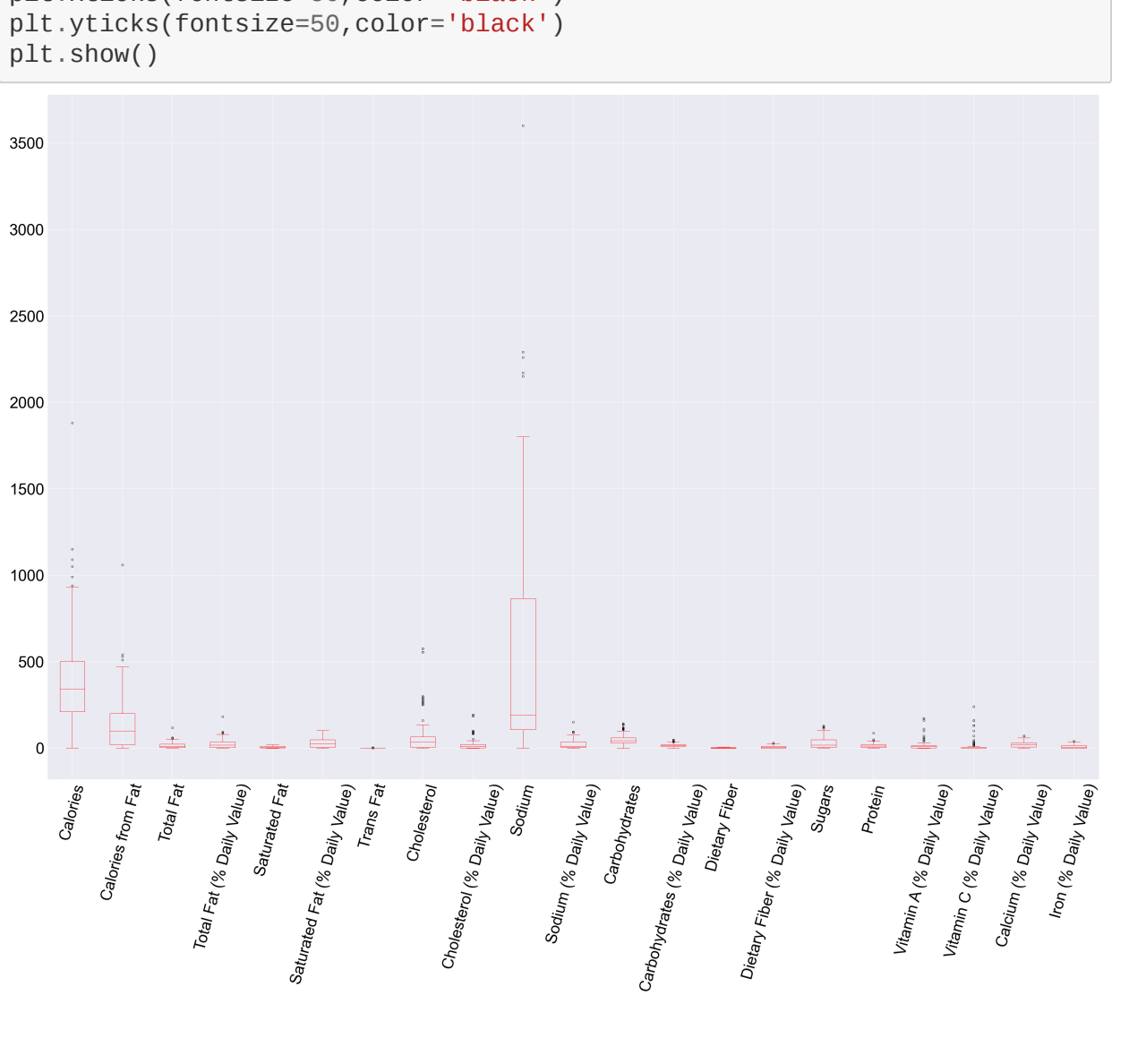
	Calories	Calories from Fat	Total Fat	Total Fat (% Daily Value)	Saturated Fat	Saturated Fat (% Daily Value)	Trans Fat	Cho
count	260.000000	260.000000	260.000000	260.000000	260.000000	260.000000	260.000000	260
mean	368.269231	127.096154	14.165385	21.815385	6.007692	29.965385	0.203846	54
std	240.269886	127.875914	14.205998	21.885199	5.321873	26.639209	0.429133	87
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	(
25%	210.000000	20.000000	2.375000	3.750000	1.000000	4.750000	0.000000	5
50%	340.000000	100.000000	11.000000	17.000000	5.000000	24.000000	0.000000	35
75%	500.000000	200.000000	22.250000	35.000000	10.000000	48.000000	0.000000	65
max	1880.000000	1060.000000	118.000000	182.000000	20.000000	102.000000	2.500000	575

8 rows × 21 columns

3. Finding out the outliers for each variable

> Our aim to finding outliers is to check the variation in the data.
>
> We will plot variables in x-axis and mean,25%,75%,max,min on y axis and anything apart from these are ouliers

[4]:
```python
plt.figure(figsize=(60,40))
Mcd.boxplot(rot=75,color='red')
plt.xticks(fontsize=50,color='black')
plt.yticks(fontsize=50,color='black')
plt.show()
```



By seeing the above graph we observed that all the variables are having outliers

> It means there are lot inconsistency present in the data

4.Now let us look at the and information of data given

> whi.ch would help to know the row and columns size and data type

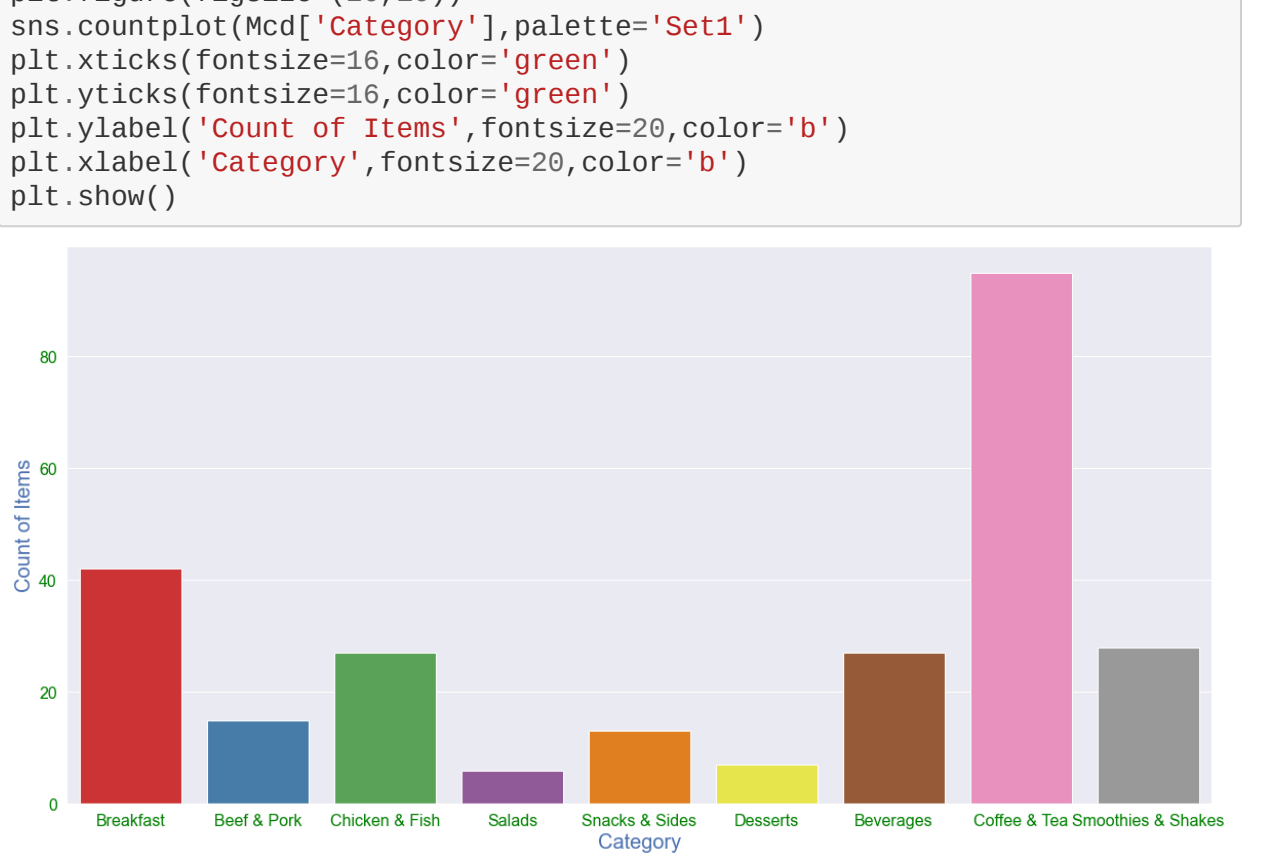
[5]:
```python
Mcd.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 260 entries, 0 to 259
Data columns (total 24 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   Category                       260 non-null    object
 1   Item                           260 non-null    object
 2   Serving Size                   260 non-null    object
 3   Calories                       260 non-null    int64
 4   Calories from Fat              260 non-null    int64
 5   Total Fat                      260 non-null    float64
 6   Total Fat (% Daily Value)      260 non-null    int64
 7   Saturated Fat                  260 non-null    float64
 8   Saturated Fat (% Daily Value)  260 non-null    int64
 9   Trans Fat                      260 non-null    float64
 10  Cholesterol                    260 non-null    int64
 11  Cholesterol (% Daily Value)    260 non-null    int64
 12  Sodium                         260 non-null    int64
 13  Sodium (% Daily Value)         260 non-null    int64
 14  Carbohydrates                  260 non-null    int64
 15  Carbohydrates (% Daily Value)  260 non-null    int64
 16  Dietary Fiber                  260 non-null    int64
 17  Dietary Fiber (% Daily Value)  260 non-null    int64
 18  Sugars                         260 non-null    int64
 19  Protein                        260 non-null    int64
 20  Vitamin A (% Daily Value)      260 non-null    int64
 21  Vitamin C (% Daily Value)      260 non-null    int64
 22  Calcium (% Daily Value)        260 non-null    int64
 23  Iron (% Daily Value)           260 non-null    int64
dtypes: float64(3), int64(18), object(3)
memory usage: 48.9+ KB
```

5.Food categories having the highest and lowest varieties.

> We will check the no. of Items available under each category
>
> Since we have the categorical value so we would use countplot

[6]:
```python
plt.figure(figsize=(20,10))
sns.countplot(Mcd['Category'],palette='Set1')
plt.xticks(fontsize=16,color='green')
plt.yticks(fontsize=16,color='green')
plt.ylabel('Count of Items',fontsize=20,color='b')
plt.xlabel('Category',fontsize=20,color='b')
plt.show()
```



We can see that Coffee & Tea and Breakest are on the higher side where as Salads and Desserts are on the lower side

6. Finding the correlation among varaibles

> To check is there any realtion between two variables
>
> We would come to a conclusion that how the two variables are related (+vely or -vely)
>
> Let us create a dataFrame as corr.

[7]:
```python
corr=Mcd.corr()
corr.head(5)
```

Out[7]:

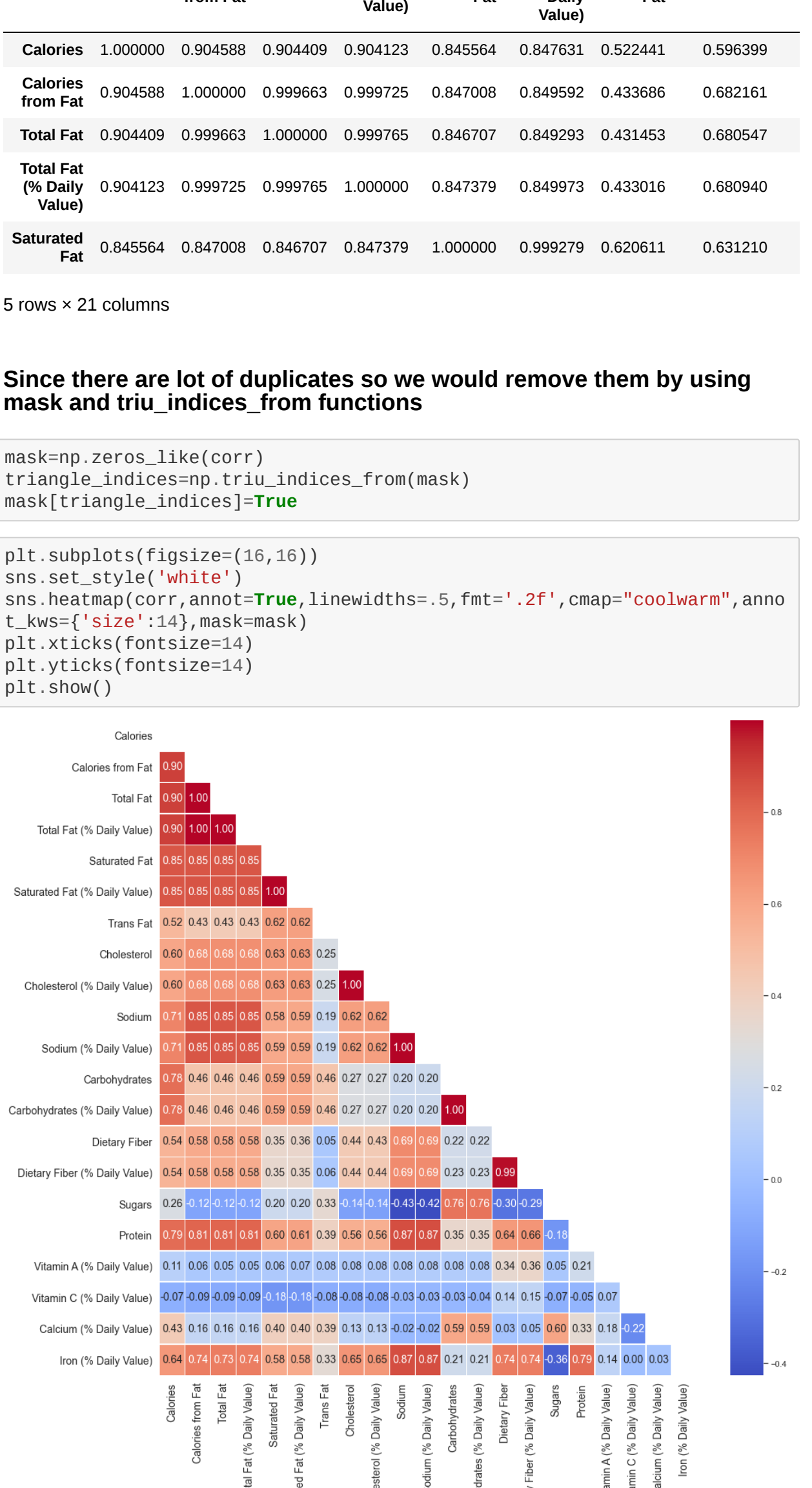
	Calories	Calories from Fat	Total Fat	Total Fat (% Daily Value)	Saturated Fat	Saturated Fat (% Daily Value)	Trans Fat	Cholesterol	Ch
Calories	1.000000	0.904588	0.904409	0.904123	0.845564	0.847631	0.522441	0.596399	
Calories from Fat	0.904588	1.000000	0.999663	0.999725	0.847008	0.849592	0.433686	0.682161	
Total Fat	0.904409	0.999663	1.000000	0.999765	0.846707	0.849293	0.431453	0.680547	
Total Fat (% Daily Value)	0.904123	0.999725	0.999765	1.000000	0.847379	0.849973	0.433016	0.680940	
Saturated Fat	0.845564	0.847008	0.846707	0.847379	1.000000	0.999279	0.620611	0.631210	

5 rows × 21 columns

Since there are lot of duplicates so we would remove them by using mask and triu_indices_from functions

[8]:
```python
mask=np.zeros_like(corr)
triangle_indices=np.triu_indices_from(mask)
mask[triangle_indices]=True
```

[14]:
```python
plt.subplots(figsize=(16,16))
sns.set_style('white')
sns.heatmap(corr,annot=True,linewidths=.5,fmt='.2f',cmap="coolwarm",annot_kws={'size':14},mask=mask)
plt.xticks(fontsize=14)
plt.yticks(fontsize=14)
plt.show()
```



The heatmap shows that variables having values more than +ve 0.8 are strongly related and values with -ve pointers indicate that variables are negatively related , let us point out the major relations

> Total Fat and Calroies are +vely related
>
> Saturated Fat and Calories are +vely related
>
> Sodium and Fat are +vely related
>
> Carbohydrate and Calories are +vely related (Not strongly value =0.78)
>
> Protein is +vely related to Calories, Fat and Sodium
>
> Iron(% Daily Value) is +vely related to Sodium and Protein

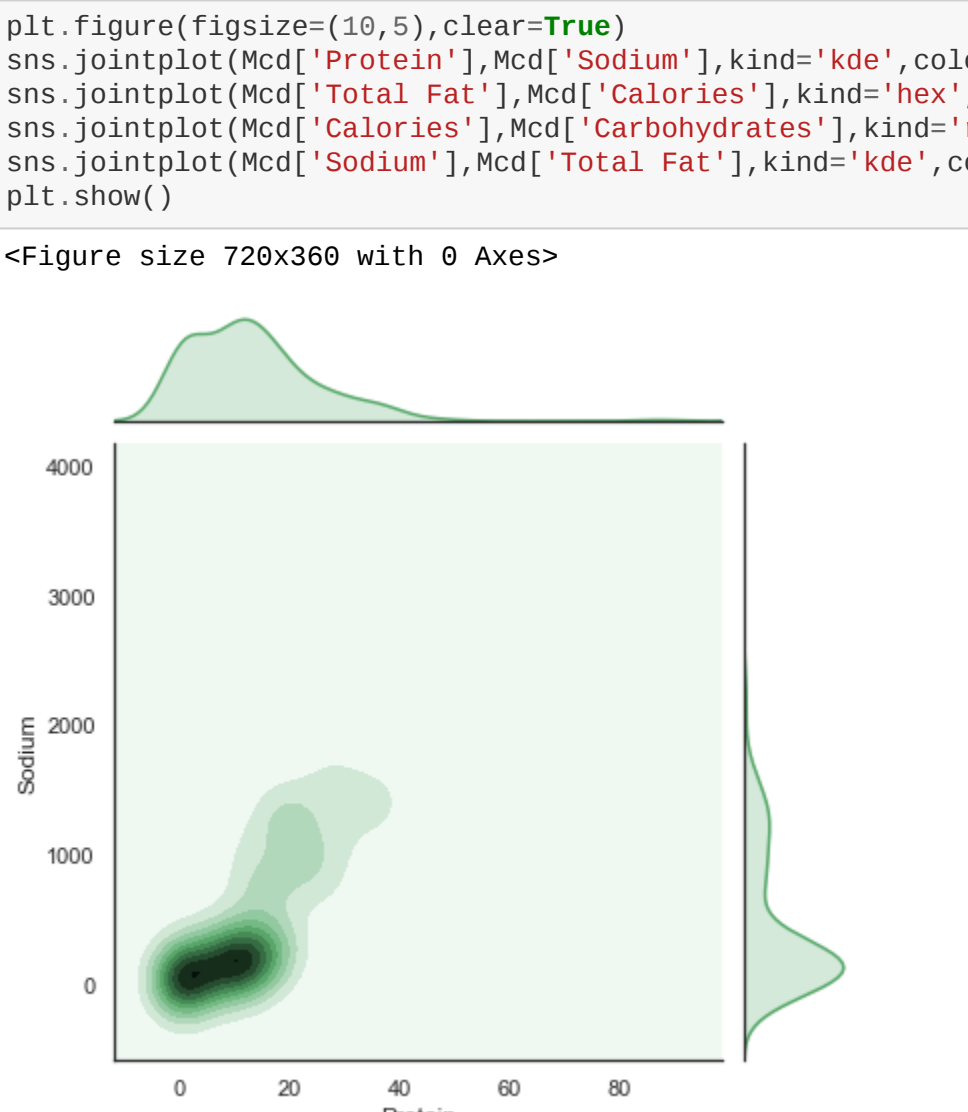
Note

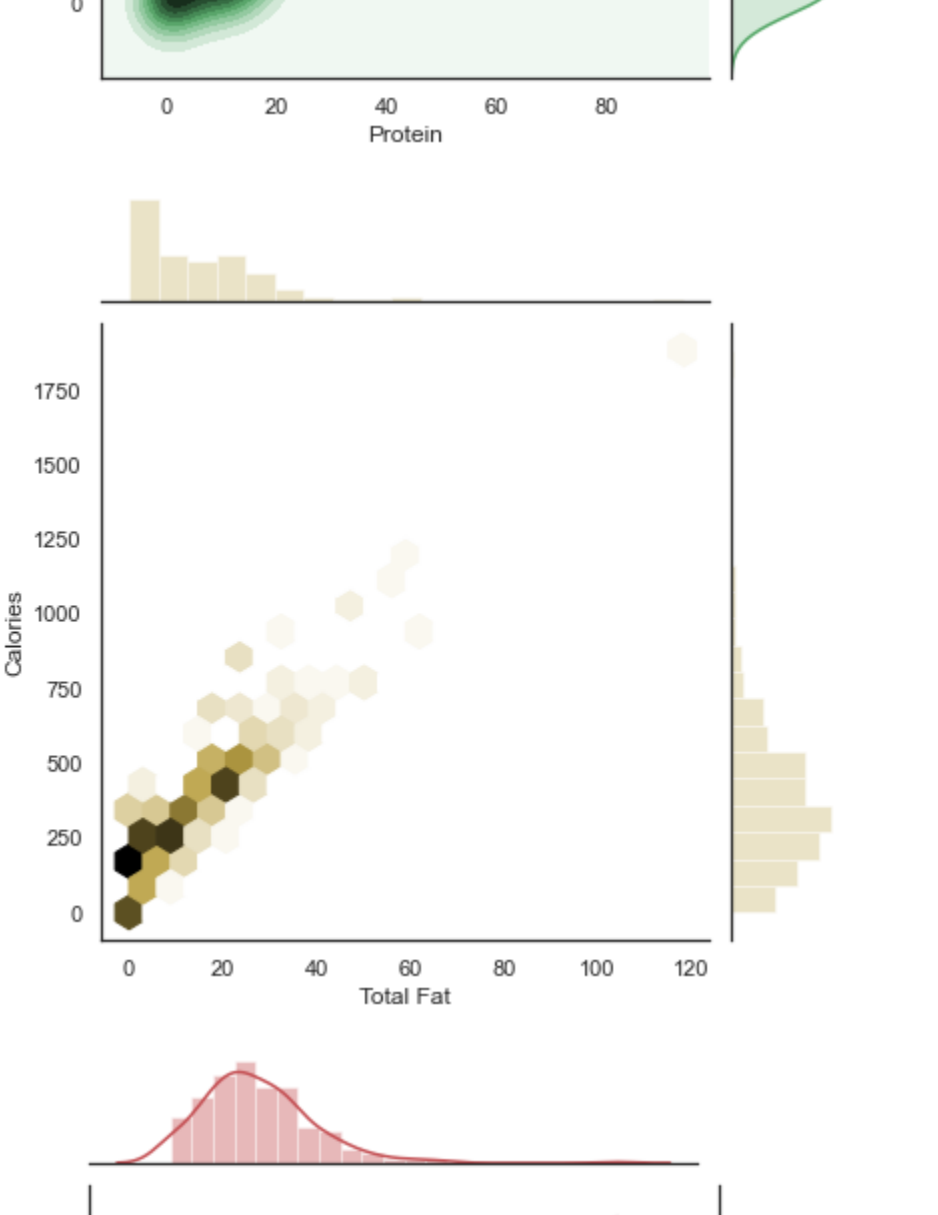
>Vitamin A,B and C dont have strong relation with any other variable

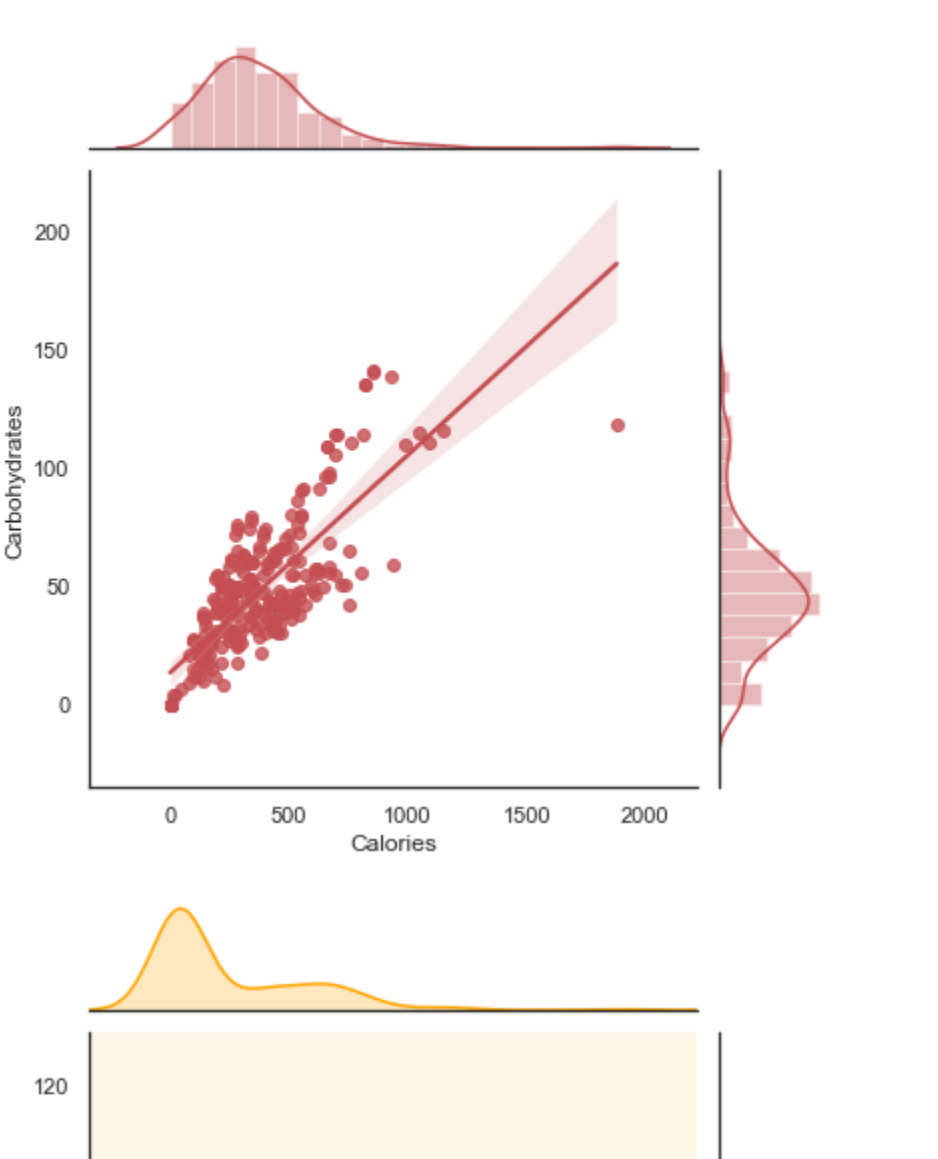
Now let us plot the graph for these correlated variables

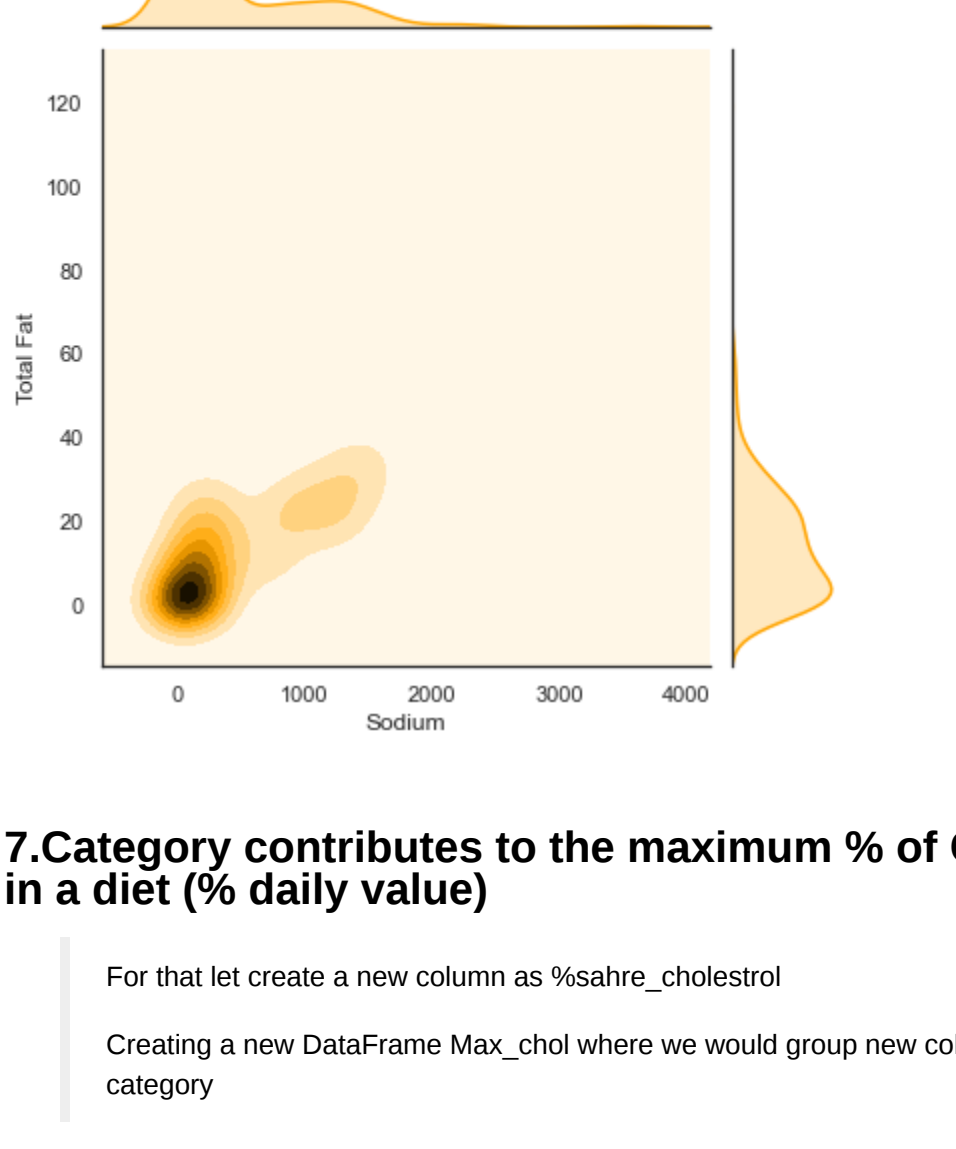
[133]:
```python
plt.figure(figsize=(10,5),clear=True)
sns.jointplot(Mcd['Protein'],Mcd['Sodium'],kind='kde',color='g')
sns.jointplot(Mcd['Total Fat'],Mcd['Calories'],kind='hex',color='y')
sns.jointplot(Mcd['Calories'],Mcd['Carbohydrates'],kind='reg',color='r')
sns.jointplot(Mcd['Sodium'],Mcd['Total Fat'],kind='kde',color='orange')
plt.show()
```

```
<Figure size 720x360 with 0 Axes>
```









7.Category contributes to the maximum % of Cholesterol in a diet (% daily value)

> For that let create a new column as %sahre_cholestrol
>
> Creating a new DataFrame Max_chol where we would group new coloumn and category

[11]:
```python
Mcd['%share_cholestrol']=Mcd['Cholesterol (% Daily Value)']/Mcd['Cholesterol (% Daily Value)'].sum()*100
Max_chol=Mcd[['%share_cholestrol','Category']].groupby('Category').sum()
Max_chol
```

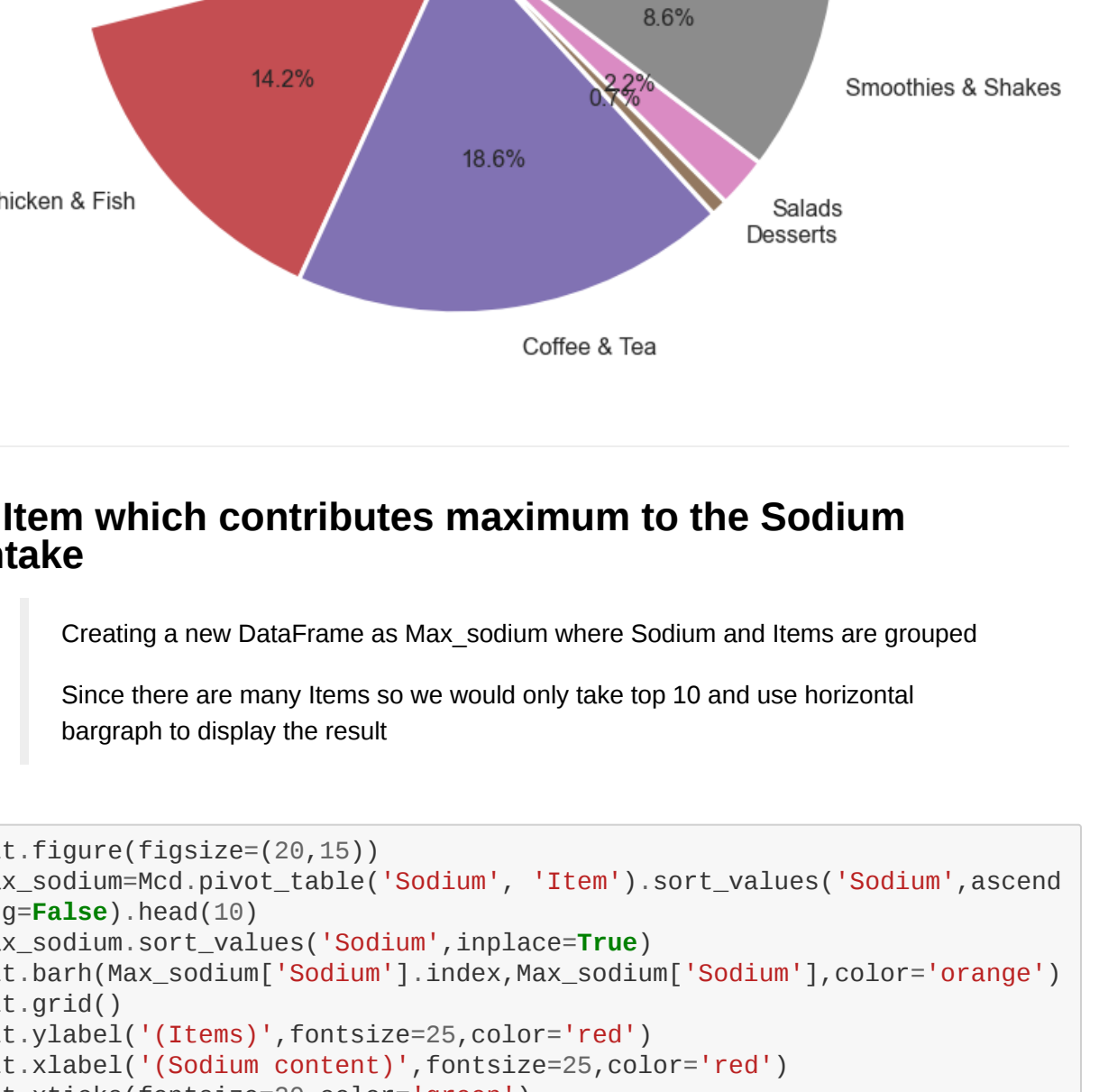
Out[11]:

Category	%share_cholestrol
Beef & Pork	9.075701
Beverages	0.104559
Breakfast	44.751150
Chicken & Fish	14.240903
Coffee & Tea	18.632371
Desserts	0.711000
Salads	2.174822
Smoothies & Shakes	8.615642
Snacks & Sides	1.693852

Plotting pie chart to show the % share of cholestrol(% daily value) category wise

> Max_chol dataFrame shows that Breakfast carries the max % i.e 44.8 Let use explode attribute to give a better view o the graph, [Breakfast is at the 3rd position]

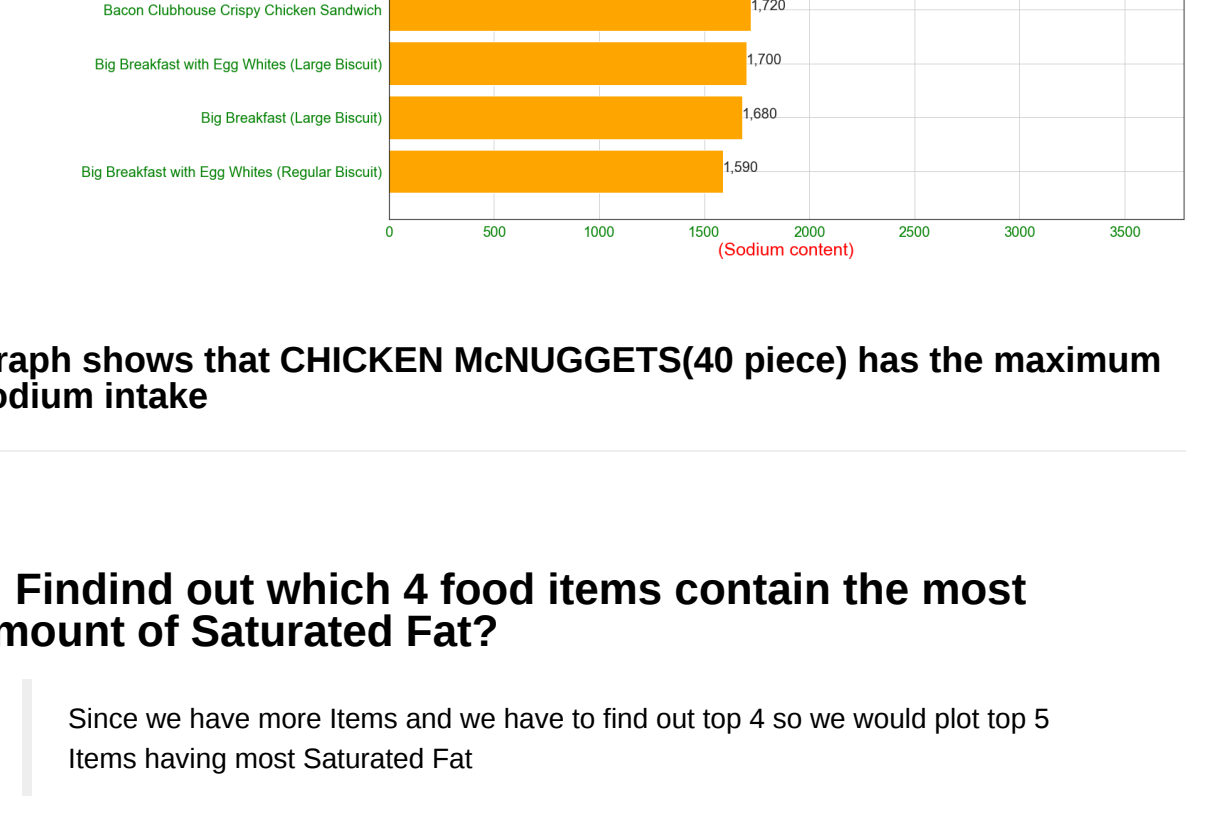
[12]:
```python
explode=[0,0,0.3,0,0,0,0,0,0]
plt.pie(Max_chol['%share_cholestrol'],labels=Max_chol['%share_cholestrol'].index,radius=3,
autopct='%0.1f%%',explode=explode,wedgeprops = {'linewidth': 3},textprops={'fontsize': 16})
plt.pie([1],colors='w',radius=0.2)
plt.show()
```



8.Item which contributes maximum to the Sodium intake

> Creating a new DataFrame as Max_sodium where Sodium and Items are grouped
>
> Since there are many Items so we would only take top 10 and use horizontal bargraph to display the result

[93]:
```python
plt.figure(figsize=(20,15))
Max_sodium=Mcd.pivot_table('Sodium', 'Item').sort_values('Sodium',ascending=False).head(10)
Max_sodium.sort_values('Sodium',inplace=True)
plt.barh(Max_sodium['Sodium'].index,Max_sodium['Sodium'],color='orange')
plt.grid()
plt.ylabel('(Items)',fontsize=25,color='red')
plt.xlabel('(Sodium content)',fontsize=25,color='red')
plt.xticks(fontsize=20,color='green')
plt.yticks(fontsize=20,color='green')
for pY,pX in enumerate(Max_sodium['Sodium']):
    plt.annotate("{:,}".format(pX),xy=(pX,pY),fontsize=20)
plt.show()
```



Graph shows that CHICKEN McNUGGETS(40 piece) has the maximum sodium intake

9. Findind out which 4 food items contain the most amount of Saturated Fat?

> Since we have more Items and we have to find out top 4 so we would plot top 5 Items having most Saturated Fat

[82]:
```python
plt.figure(figsize=(40,10))
Item_Fat=Mcd[['Item','Saturated Fat']].groupby('Item').sum().sort_values('Saturated Fat',ascending=False).head(5)
sns.barplot(Item_Fat['Saturated Fat'].index,Item_Fat['Saturated Fat'])
plt.grid()
plt.ylabel('(Saturated Fat)',fontsize=30,color='red')
plt.xlabel('(Items)',fontsize=30,color='red')
plt.xticks(fontsize=23,color='green')
plt.yticks(fontsize=23,color='green')
plt.show()
```

From above graph we conclude that Chicken McNuggets (40 piece),Frappe chocolate chip(Large),Big Breakfast with Hotcakes (Large Biscuit) and McFlurry with M&M's Candies(Medium) have the most saturated Fat i.e, 20