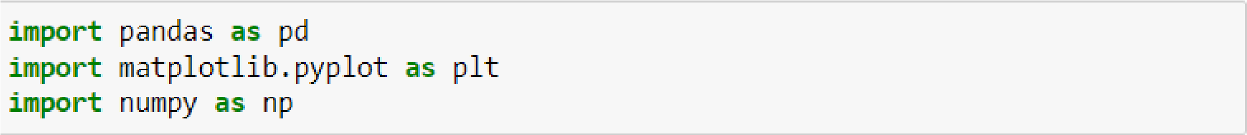
Import, Preprocess, and Visualize a Dataset Task

In this task, I imported a dataset, did some preparation processes, and visualized the results to answer some given questions.

1- Import the Dataset:

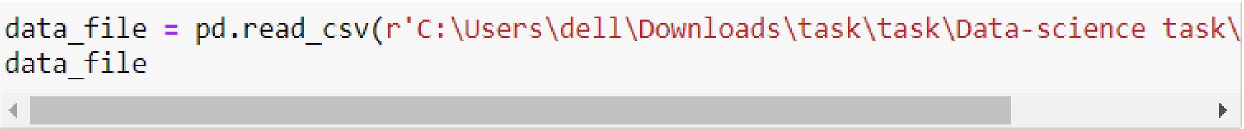
To import it, I preferred using the Pandas library and loading the dataset on a data frame to easily deal with it.

- **First, Import the Pandas library**



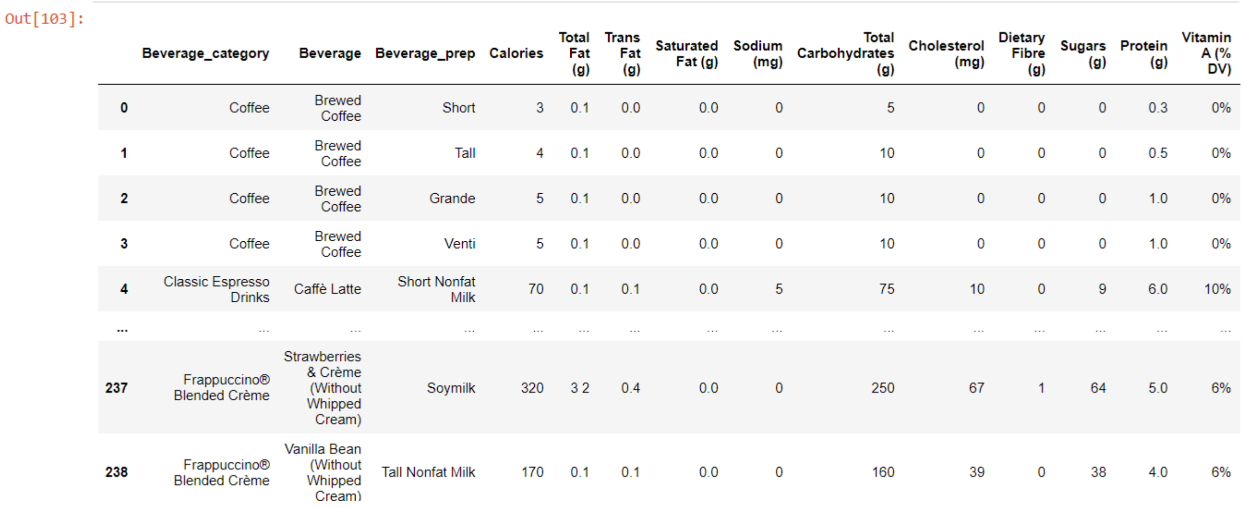
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
```

- **Then import the dataset using the read_csv() method**



```
data_file = pd.read_csv(r'C:\Users\dell\Downloads\task\task\Data-science task\
data_file
```

This function read a CSV file by giving it its path and loading it into a Pandas data frame.



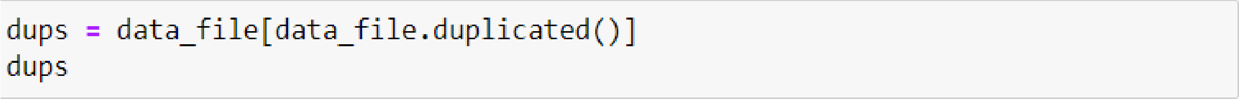
Out[103]:

	Beverage_category	Beverage	Beverage_prep	Calories	Total Fat (g)	Trans Fat (g)	Saturated Fat (g)	Sodium (mg)	Total Carbohydrates (g)	Cholesterol (mg)	Dietary Fibre (g)	Sugars (g)	Protein (g)	Vitamin A (% DV)
0	Coffee	Brewed Coffee	Short	3	0.1	0.0	0.0	0	5	0	0	0	0.3	0%
1	Coffee	Brewed Coffee	Tall	4	0.1	0.0	0.0	0	10	0	0	0	0.5	0%
2	Coffee	Brewed Coffee	Grande	5	0.1	0.0	0.0	0	10	0	0	0	1.0	0%
3	Coffee	Brewed Coffee	Venti	5	0.1	0.0	0.0	0	10	0	0	0	1.0	0%
4	Classic Espresso Drinks	Caffè Latte	Short Nonfat Milk	70	0.1	0.1	0.0	5	75	10	0	9	6.0	10%
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
237	Frappuccino® Blended Crème	Strawberries & Crème (Without Whipped Cream)	Soymilk	320	3 2	0.4	0.0	0	250	67	1	64	5.0	6%
238	Frappuccino® Blended Crème	Vanilla Bean (Without Whipped Cream)	Tall Nonfat Milk	170	0.1	0.1	0.0	0	160	39	0	38	4.0	6%

2- *Remove the duplicated rows from the Dataset:*

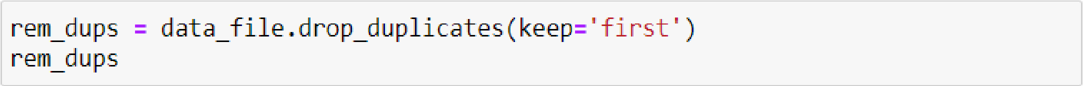
In this step, I checked if there were duplicated rows, and remove them.

- **First, Check the duplicated rows using the duplicated() method.**



```
dups = data_file[data_file.duplicated()]
dups
```

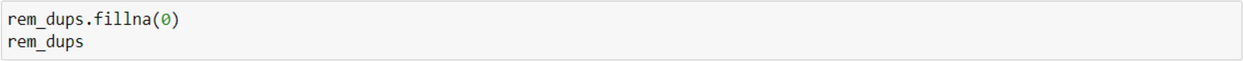
- **Then, drop these duplicated rows using the drop_duplicates() method and load the result into a new data frame.**
  Note that I used the (Keep) argument with a value ‘first’ to keep the first duplicated row and delete what is after.



```
rem_dups = data_file.drop_duplicates(keep='first')
rem_dups
```

3- *Fill the ‘Null’ values in the Dataset:*

In this step, I filled the null values with a “0” value using the fillna() method.

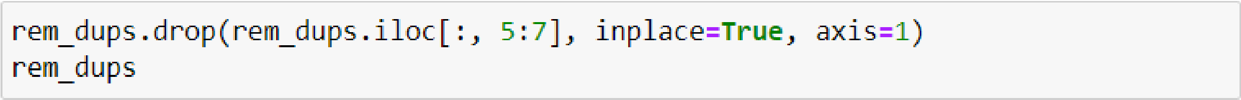


```
rem_dups.fillna(0)
rem_dups
```

4- *Drop the unnecessary columns from the Dataset:*

In this step, I dropped the “Trans Fat (g)” and “Saturated Fat (g)” columns because their values are summed in the “Total Fat (g)” columns, so I found them as duplicated data.

I dropped columns using the drop() method and give it the indexes of the columns that I wanted to drop.



```
rem_dups.drop(rem_dups.iloc[:, 5:7], inplace=True, axis=1)
rem_dups
```

5- *Visualize the results to answer some questions:*

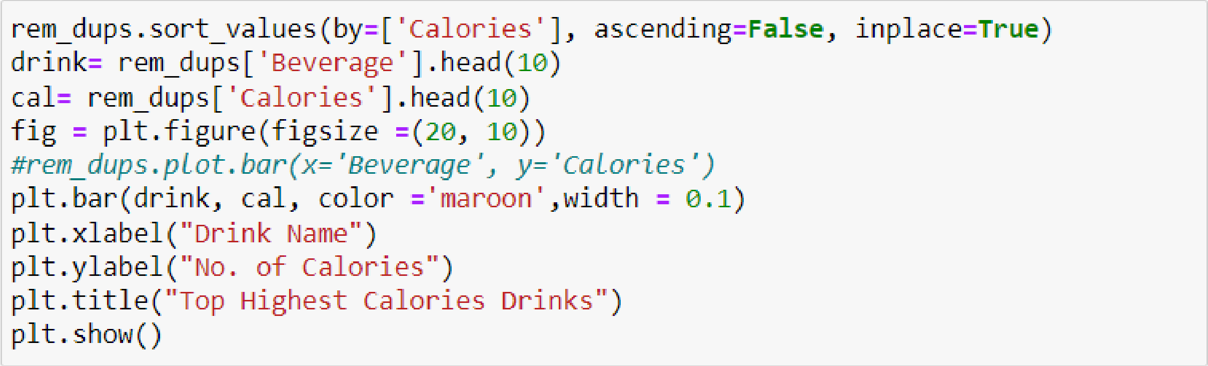
In this step, I wanted to answer Two Questions:

Q1. Which drink has the highest calories from the dataset?

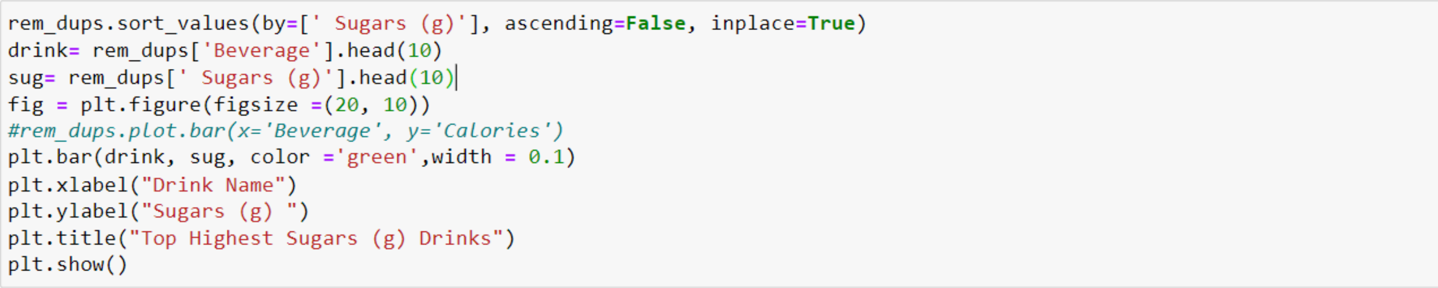
Q2. Highest Sugar Drink?

So, to solve them, I had to drown a Bar Char.

- **First, sort the data frame values using the sort_values() method to sort them based on a given column.**
- **To answer the first question, I had to sort the data by the "Calories" column. And by the "Sugars (g)" column to answer the second question.**
- **After that, define variables for the X and Y axis.**
- **Then, use the bar() method to drown the bar chart.**
- **Using xlabel(), and ylabel() methods, I gave these axes an understandable text.**
- **Using the title() method, I labeled the chart with meaningful text.**
- **Finally, I used the show() method to show the chart.**

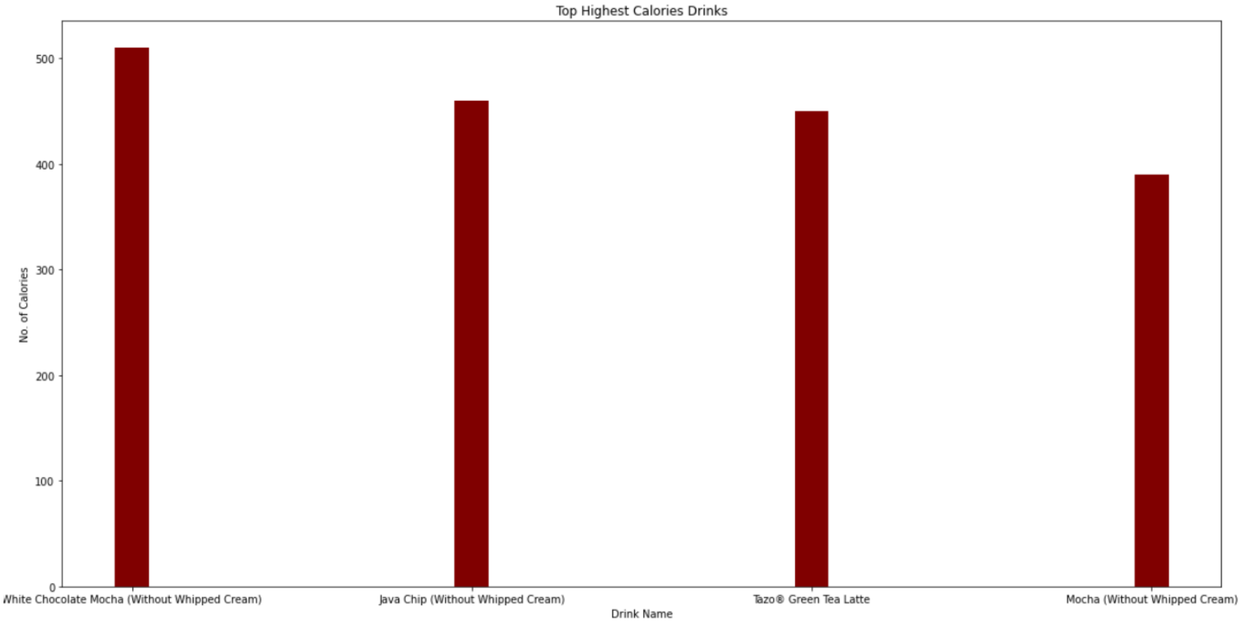


```
rem_dups.sort_values(by=['Calories'], ascending=False, inplace=True)
drink= rem_dups['Beverage'].head(10)
cal= rem_dups['Calories'].head(10)
fig = plt.figure(figsize =(20, 10))
#rem_dups.plot.bar(x='Beverage', y='Calories')
plt.bar(drink, cal, color ='maroon',width = 0.1)
plt.xlabel("Drink Name")
plt.ylabel("No. of Calories")
plt.title("Top Highest Calories Drinks")
plt.show()
```



```
rem_dups.sort_values(by=[' Sugars (g)'], ascending=False, inplace=True)
drink= rem_dups['Beverage'].head(10)
sug= rem_dups[' Sugars (g)'].head(10)
fig = plt.figure(figsize =(20, 10))
#rem_dups.plot.bar(x='Beverage', y='Calories')
plt.bar(drink, sug, color ='green',width = 0.1)
plt.xlabel("Drink Name")
plt.ylabel("Sugars (g) ")
plt.title("Top Highest Sugars (g) Drinks")
plt.show()
```

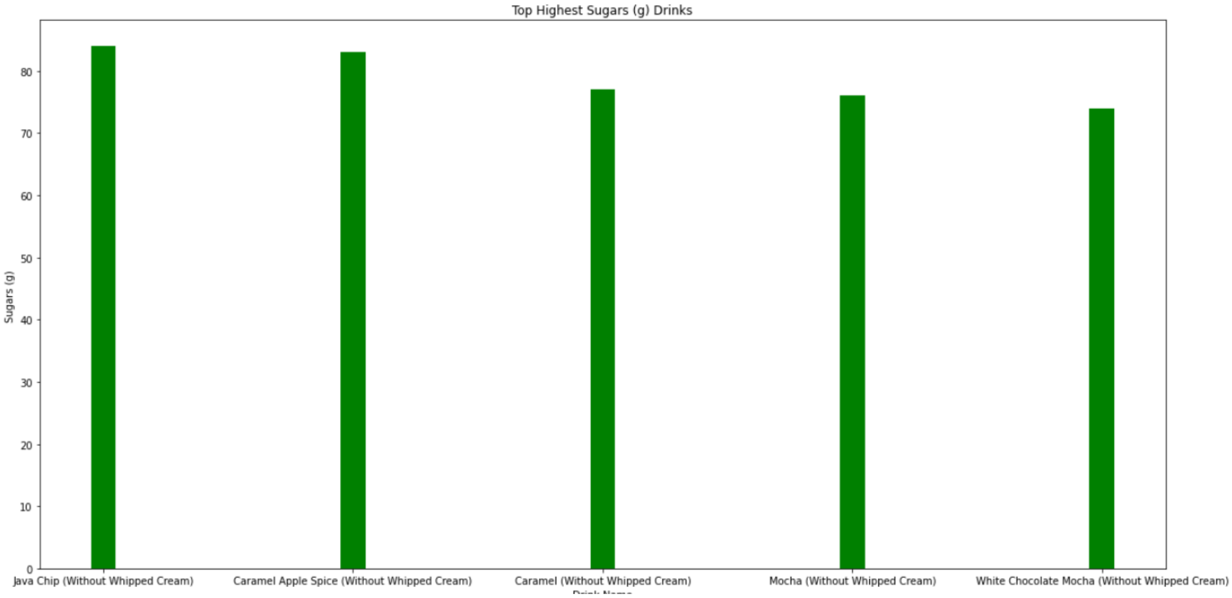
6- ***The Final Results:***



From this chart, I answered the first question:

Q1. Which drink has the highest calories from the dataset?

From the chart, we can find that “White Chocolate Mocha (Without Whipped Cream) is the top calories drink with more than 500 calory.



From this chart, I answered the second question:

Q2. Highest Sugar Drink?

From the chart, we can find that "Java Chip (Without Whipped Cream) is the top sugar drink with more than 80 gram.