Mahmoud Basha

mahmoudbuiltin@gmail.com

Data Science task

Contents

How to Run the code

There are two common methods that you can use to run the notebook. First method is to run on Cloud using google colab, second method is to run locally using Jupiter notebook.

To Run the notebook, I would suggest using google colab to avoid installing and managing python packages on local machine.

Google Colab

1 – upload the data set to google colab by dragging and dropping into the files tab on the left of the webpage.

2 – Install the required python packages on your virtual machine in google cloab by running the following commands

```
! pip install pandas

! pip install -U scikit-learn

! pip install numpy

! pip install seaborn

! pip install matplotlib
```

3- Change the path to the dataset in read_csv function in the second cell to ‘drinkMenu.csv’

4- now you can proceed to run the cells in the notebook sequentially

Jupyter Notebook

The upcoming steps are for windows machine, mac and linux will have very similar commands and steps.

1 – You need to make sure you have a working python installation on your machine

2 – you can use conda to install the python packages or you can use pip as in google colab section

3 – if you choose to use pip the commands you need to run in the command line prompt are as follows

```
 pip install pandas

pip install -U scikit-learn

pip install numpy

pip install seaborn

pip install matplotlib

pip install notebook
```

4 – to run jupyter notebook run the following command in the command line prompt

```
jupyter notebook
```

5 – your default browser will open up a new windows and you will be prompted to choose the notebook to open, navigate to the location where you downloaded my notebook and simply click on it to open.

6 – change the path of the dataset in the read_csv function as commented in the code to your local location where you saved the excel sheet.

7 – you can now proceed to run the code cells sequentially.

Documentation

Import dataset and basic data exploration

The dataset has 242 record and 18 feature/attribute.

3 of the 18 columns (Beverage_category, Beverage, Beverage_prep) are clear strings represented as pandas object data type. 9 of the 18 are numeric columns (Calories, Trans Fat (g), Saturated Fat (g), Sodium (mg), Total Carbohydrates (g), Cholesterol (mg), Dietary Fibre (g), Sugars (g), Protein (g)). The remaining 6 columns are of type object and need further inspection.

There exist only one null value in the 'Caffeine (mg)' in record 158.

'Caffeine (mg)' column has two obscure values -> 'varies' and 'Varies'. These two values are merged into a single value and then replaced by the mean of the 'Caffeine (mg)' column.

'Vitamin A (% DV) ', 'Vitamin C (% DV)', and ' Calcium (% DV) ' are stripped out of the percentage sign and converted to integers.

'Iron (% DV)' is stripped out of the percentage sign, converted to float then converted to integer; after checking the values of the column it was found that all numbers after decimal point were zeros thus it was safe to convert the column to integer.

' Total Fat (g)' column had an error in a value were it was written '3 2' were it should be '3.2' so it was fixed.

Data preparation

Fill null values

Checking the only record that has a null value (row 158) in the ‘Caffeine (mg)’ , it was found that the rest of the attributes values were normal and within expected range. Thus, I concluded that the missing value was MCAR (missing completely at Random) and proposed three methods to fix the missing value.

1 – delete the record with the null value

2- fill the missing value with the mean of the column, which cam out to be 89.5205

3- fill the missing value with machine learning (KNN imputer), which resulted a value of 86.90 to replace the null value.

Method 3 was used.

Remove duplicates

The data has no ID column so detecting duplicates imposed a challenge. Two different methods were proposed to remove the duplicates :

1 – use pandas function ’drop_duplicates()’ which marks two records as duplicates they are exactly identical in every field -> this method dropped zero records

2 – use the combination of the three columns ('Beverage_category','Beverage','Beverage_prep') to act as and identifier for the record, as it was assumed that any drink that has the same exact values for the three columns should also have the same nutritional value, and if these nutritional values are not the same it is only a result from an error from data entry or data collection. -> this method dropped 90 record.

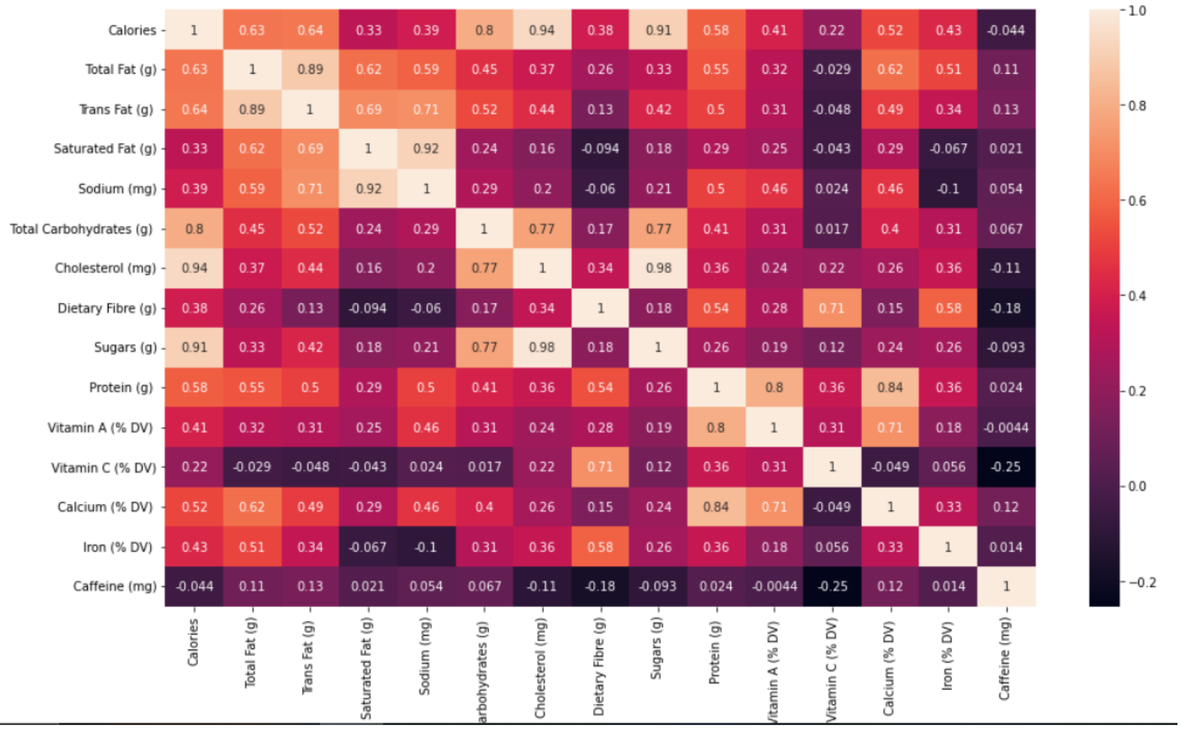
Method 2 was used

Drop unnecessary Columns

To measure column importance for dropping two methods were proposed:

1 - Removing features with low variance -> features with variances close to 1 were to be dropped ; the following features were observed to have variance close to 1 (' Calcium (% DV) ', ' Protein (g) ', ' Sodium (mg)', ' Sugars (g)', ' Total Fat (g)', 'Caffeine (mg)', 'Cholesterol (mg)', 'Iron (% DV) ', 'Saturated Fat (g)', 'Trans Fat (g) ', 'Vitamin A (% DV) ', 'Vitamin C (% DV)').

2 – Correlation heatmap -> features that are highly correlated can be assumed to be redundant and can therefore be dropped as they offer no new information. We can impose a certain threshold if the correlation value between two features exceeds it, one can be dropped.



None of the two methods were used as the dataset is not too big so we don’t have a curse of dimensionality and all columns can be quickly processed.

Data Visualizations: Using plots to answers this questions

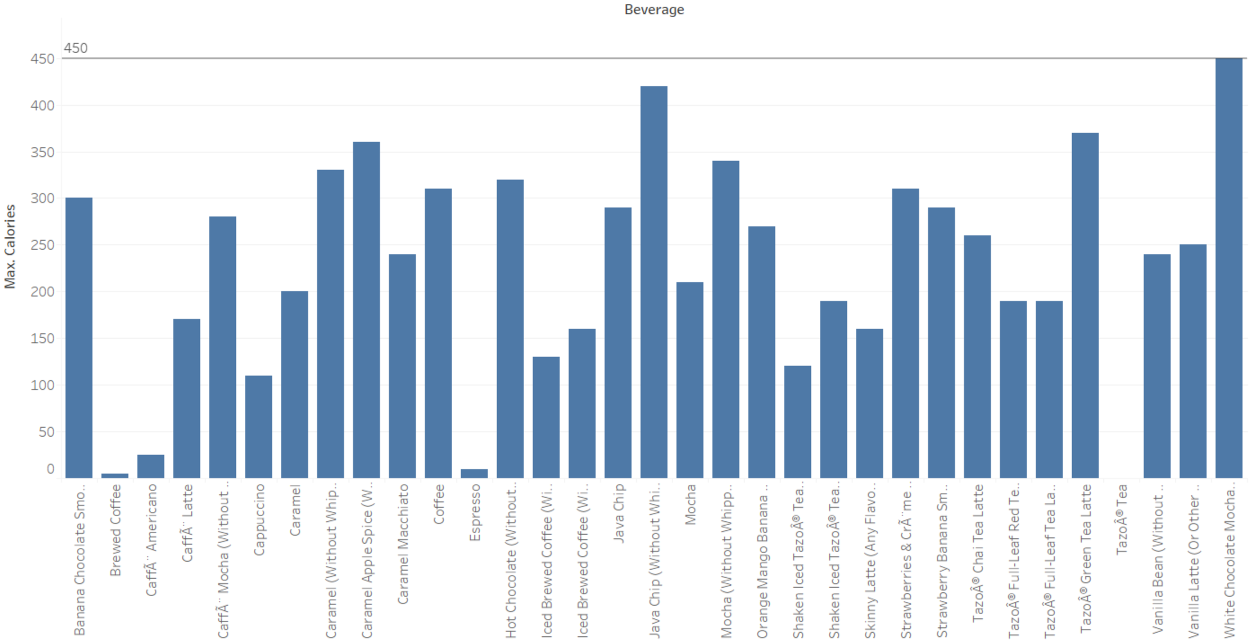
Two approaches were used in this part:

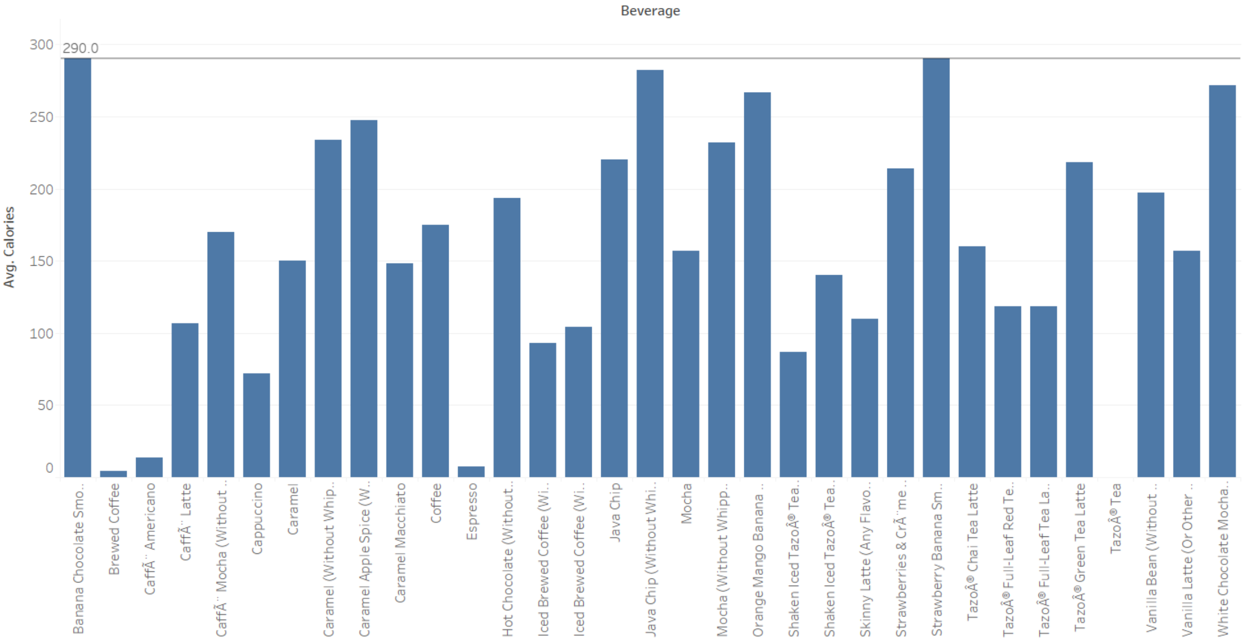
1 – Tableau -> was used as I find it very practical and straightforward when doing visualizations

2 – Python –> As the tableau usually requires a license to use commercially, python free packages can prove useful in a lot of situations.

Tableau

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

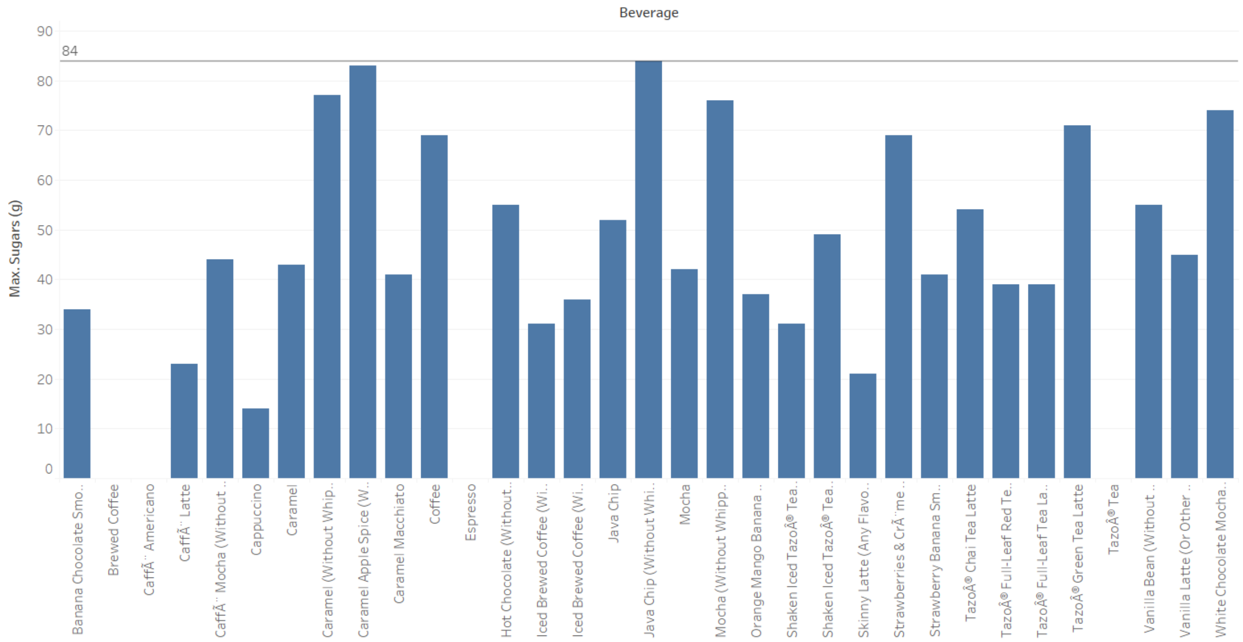


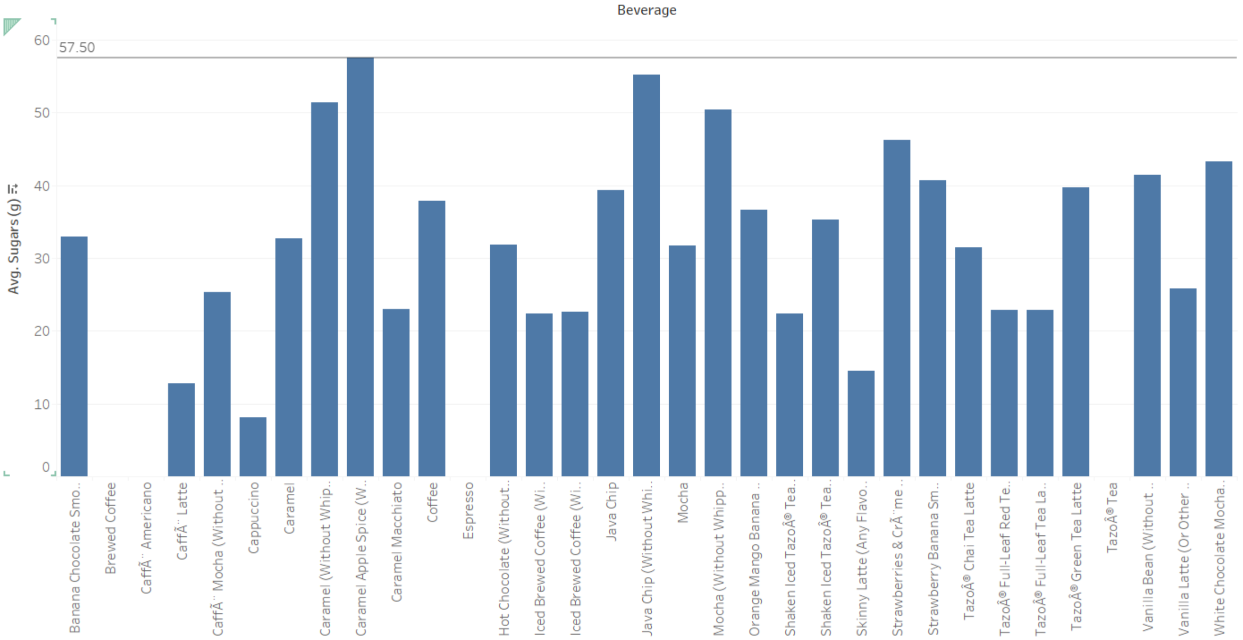


The drink that has the absolute highest calories is 'White Chocolate Mocha (Without Whipped Cream)'.

But on average 'Banana Chocolate Smoothie' and 'Strawberry Banana Smoothie' are the highest in calories with 'Java Chip' and 'White Chocolate Mocha (Without Whipped Cream)' being the runner ups.

Q2. Highest Sugar Drink ?



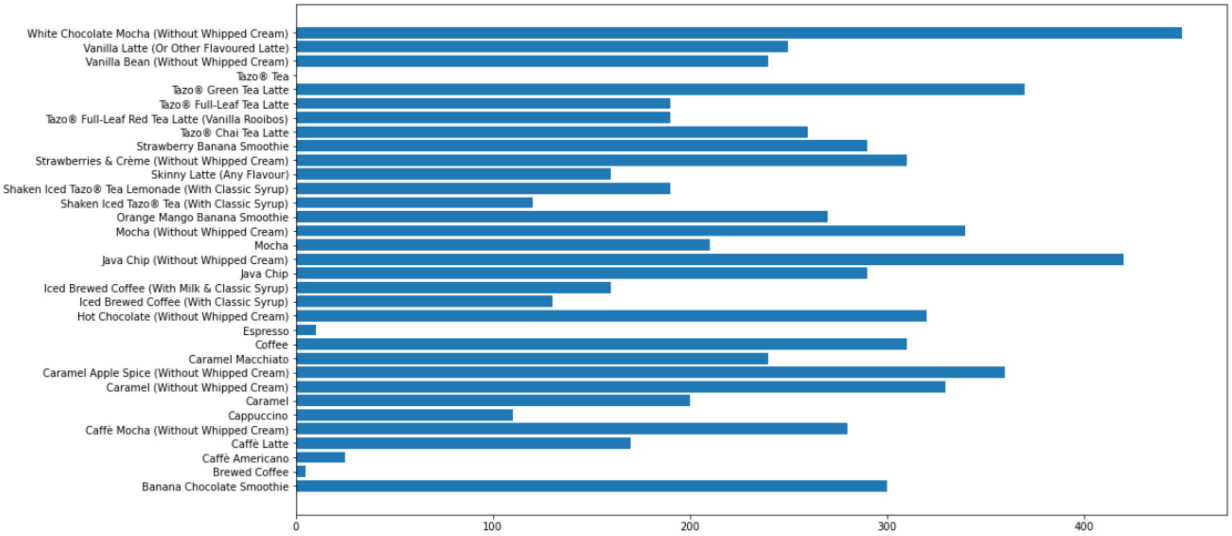


Drink that has absolute highest sugar is 'Java Chip (Without Whipped Cream)' with 'Caramel Apple Spice (Without Whipped Cream)' being very close.

On average 'Caramel Apple Spice (Without Whipped Cream)' is the highest in sugar

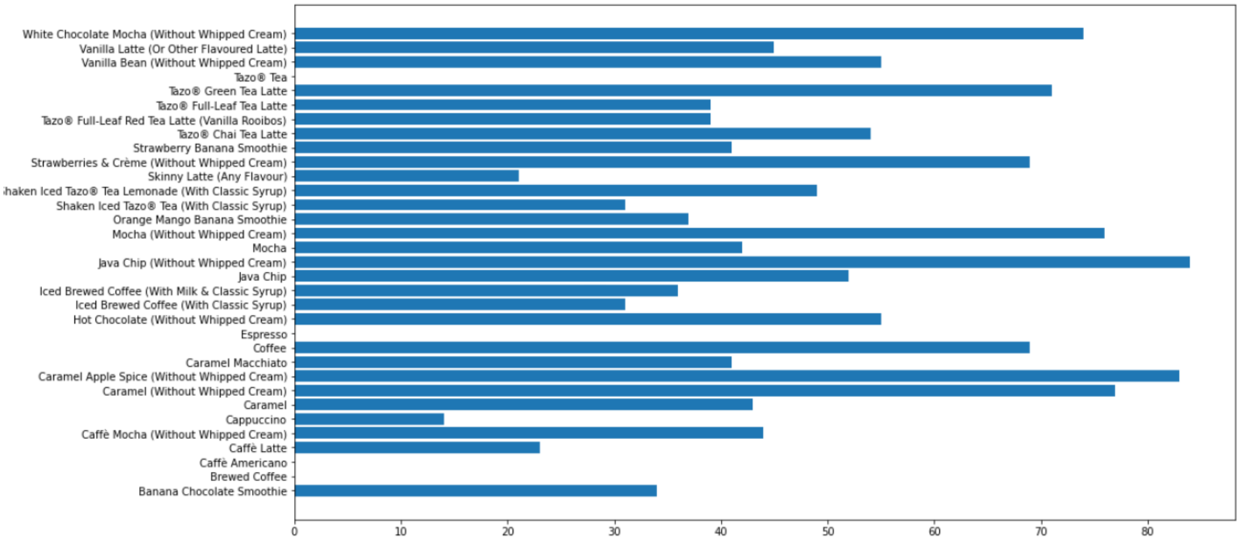
Matplotlib

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



We observe that the results is consistent with the tableau results, 'White Chocolate Mocha (Without Whipped Cream)' have the absolute highest value in Calories.

Q2. Highest Sugar Drink ?



We again observe that the drink with absolute max sugar is 'Java Chip (Without Whipped Cream)' which is consistent with tableau results.