CA4022 Assignment 1

Starbucks Nutrition Analysis Using Apache Pig and Hive

Seán O'Neill

20297411

Find Github repository here

1. Dataset

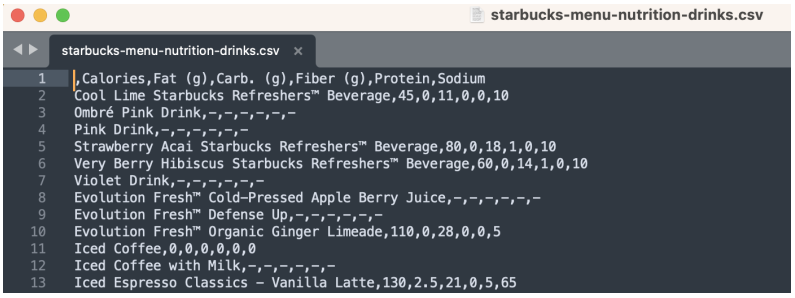
The dataset I used for this assignment is on Starbucks Nutrition data, which can be found on kaggle here. This comprises 3 datasets. These are:

- Drinks dataset. This is a table that contains all drinks available on the starbucks menu. It has a column for the name of the drink, the calorie count, the sodium count and a column for each of fat, carbohydrates, fibre and protein, all of which are measured in grams.
- Food dataset. This is the exact same as the food dataset, but does not contain any sodium data.
- Expanded drinks dataset. This is similar to the original drinks dataset, but contains less drinks and has more info available on each such as the category of each drink and the preparation of each drink, which is every variation of the drink available. It also contains more nutrition related data such as the vitamin A and iron content.

Apache Hive and Pig are very useful tools to process and analyse this type of data. I used Pig to clean and prepare the data for analysis, then used both Pig and Hive to execute simple queries related to this data and finally just Hive to complete more complex queries, which I will explain in this document.

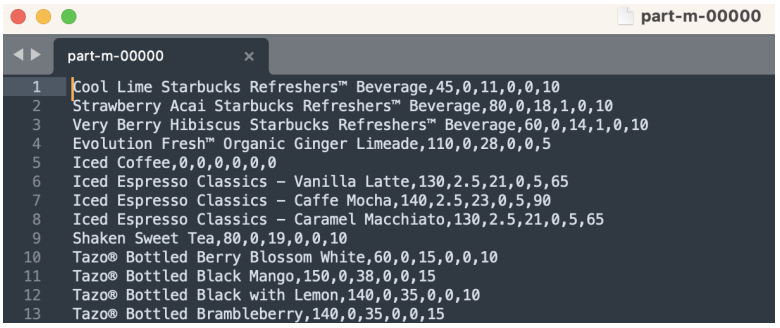
2. Cleaning the data

As stated above, I used Pig to clean the data. I stored the 3 csv files downloaded from kaggle in a folder called starbucks-raw-data. The three datasets did not require a major amount of cleaning, but each one needed its own unique technique to be cleaned. I began by cleaning the drinks dataset, which looked like this as a csv directly from Kaggle:



(a) Drinks dataset before cleaning

The first thing I noticed here that needed to be cleaned was the rows containing only '-'. This was the first thing I removed, as these would prove no use to my analysis and could cause problems when treating these columns as integers and floats. The next thing I did was remove the header rows as these also served no purpose to me. This is how the file looked as a csv after cleaning:



(b) Drinks dataset after cleaning

Once I had cleaned the drinks dataset, I cleaned the food and expanded drinks dataset, but the only necessary cleaning for the two was to remove the headers. Once that was done they were in a similar format to the file in (b).

3. Simple PIG and HIVE queries

3.1 Drinks with highest calories

My first query was a basic one. I wanted to find the drinks with the highest calories from the drinks table. I did this in both Pig and Hive and got the following results:

Pig:

```
(Starbucks® Signature Hot Chocolate,430)
(White Chocolate Mocha,360)
(Cinnamon Dolce Frappuccino® Blended Coffee,350)
(Chocolate Smoothie,320)
(Hot Chocolate,320)
```

Hive:

```
Starbucks® Signature Hot Chocolate      430
White Chocolate Mocha   360
Cinnamon Dolce Frappuccino® Blended Coffee      350
Chocolate Smoothie      320
Hot Chocolate   320
```

I got identical results from both Pig and Hive, which is a good sign, we can see here that the Signature Hot Chocolate has the highest calorie count, and by some distance.

3.2 Drinks with most variations

For this query, I used the extended drinks table to find what drink could be ordered in the most variations. These drinks can come in various sizes and with other alterations, such as the type of milk used. These are the results:

Pig:

```
(Caffè Latte,12)
(Caffè Mocha (Without Whipped Cream),12)
(Cappuccino,12)
(Caramel Macchiato,12)
(Coffee,12)
```

Hive:

```
Caffè Latte     12
Caffè Mocha (Without Whipped Cream)     12
Cappuccino      12
Caramel Macchiato       12
Coffee  12
```

Again we are returned with identical results in both Pig and Hive, with all of the top 5 drinks having 12 variations.

3.3 Drinks with fewest variations

Because all of the top five drinks came with the same variation count, I wanted to find the 5 with the least variations. I wanted to check that not every drink came with the exact same choice of alterations:

Pig:

```
(Espresso,2)
(Banana Chocolate Smoothie,3)
(Caramel,3)
(Iced Brewed Coffee (With Classic Syrup),3)
(Java Chip,3)
```

Hive:

```
Espresso        2
Banana Chocolate Smoothie       3
Caramel 3
Iced Brewed Coffee (With Classic Syrup) 3
Java Chip       3
```

Again, Identical results in Pig and Hive, also a sanity check that shows not every drink comes in 12 variations, with an Espresso only having 2.

4. Complex HIVE queries

4.1 Foods Low in Fat and High in Protein

With the number of people interested in keeping fit and scientific advancements, a lot of people know what it is they need to consume to keep in shape. A very important part of food for people looking to gain muscle is the amount of protein contained in the food. While consuming high protein meals, people may also want to keep the amount of fat they consume low. So for this query, I will use the food table to find the items with the lowest fat count that have above average protein in HIVE, finding these results:

```
Fresh Blueberries and Honey Greek Yogurt Parfait        2.5     14
Berry Trio Yogurt       2.5     14
Multigrain Bagel        4.0     17
Chonga Bagel    5.0     12
Sprouted Grain Vegan Bagel      6.0     12
```

(c) Food items with lowest fat content and above average protein

We can see here that the Blueberry and Honey Greek yoghurt was the best food item for high protein and low fat content, with just 2.5 grams of fat and 14 grams of protein.

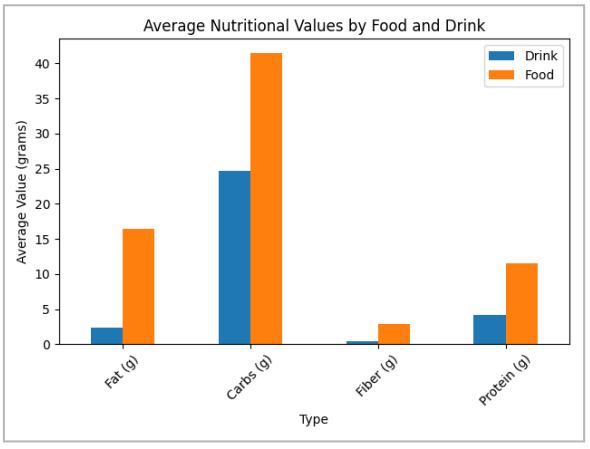
4.2 Food or Drink with Highest Combined Fat, Carbohydrates, Fibre and Protein

For this query, I am interested in finding the food or drink with the highest combined fat, carbs, fibre and protein content, as these are all measured in grams. The first thing to do here was to combine the food and drinks menu using UNION in Hive. When combining the two tables, there is no sodium column in the food table, so I added a column here and gave every food a sodium count of 0. In the combined table, I also added a 'type' column. This was simply a food or drink tag for each row. These are the top 5 results after running the query:

```
Lentils & Vegetable Protein Bowl with Brown Rice        food    153.0
Strawberries & Jam Sandwich     food    134.0
Za'atar Chicken & Lemon Tahini Salad    food    128.0
Green Goddess Avocado Salad     food    122.0
Roasted Turkey & Dill Havarti Sandwich  food    117.0
```

(d) Menu items with combined fat, carb, fibre and protein content

As we can see from the results, the *Lentils and Vegetable Protein Bowl with Brown Rice* food item has the highest combined gram count of my select parameters. This doesn't seem a surprise as it is marketed as a protein rich meal, with the addition of rice adding to the carbohydrate count. In fact, all of the top five in this query are food items, which would be expected as the drinks tend to be very low in fibre, fat and protein.



(d) Average Nutritional Value by Food and Drink

4.3 Drinks containing low fat milk from random sample of 40

For this query, I am looking at the extended drinks table. There are 242 entries in this table, and I want to take a random sample of 40 of them and find out which of these drinks contain nonfat milk. A lot of people like to order coffee with decreased fat milk for various personal reasons. I want to find out what type of choice these people are given, by finding out if there were suddenly only 40 random drinks left on the menu, how many items would be available to them. From these drinks from the random samples, I also want to print the total fat count to see if there truly is 0 grams of fat in these drinks.

I ran this query three times and got these three results:

A:

```
Caffè Latte     Tall Nonfat Milk        0.2
Tazo® Green Tea Latte   Short Nonfat Milk       0.2
```

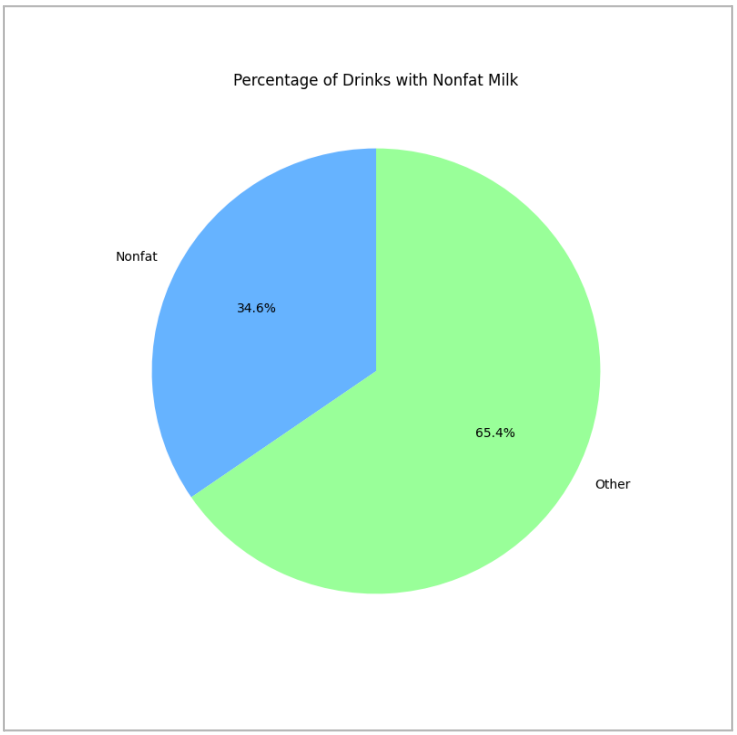
B:

```
Coffee  Tall Nonfat Milk        0.1
```

C:

```
Caramel Venti Nonfat Milk       0.1
```

So if just 40 random drinks were left on the Starbucks menu, it only seems like 1 - 2 of them will be a drink with Nonfat milk. We can also see that even without any fat added to the drink via the milk, it is very unlikely that the fat count of the drink is 0.



(e) Percentage of drinks with Nonfat milk