CA4022 - Assignment 1: Nutrition facts for Starbucks Menu dataset analysis using Apache Hive and Pig

Xi Zhang - 20100353

1. Introduction

Apache Pig is a high-level platform for creating programs that run on Apache Hadoop. It is designed to handle any kind of data—structured, semi-structured, or unstructured—using a simple scripting language known as Pig Latin, which abstracts the complexity of writing MapReduce programs. I utilised Apache Pig to perform preliminary data cleaning and transformation tasks. The data wrangling capabilities of Pig make it an ideal choice for preparing the datasets for in-depth analysis by filtering out inconsistencies, handling missing values, and reshaping the data into a more analysable form. Once cleaned and validated, the datasets will be transferred into Hive, a data warehouse infrastructure that facilitates reading, writing, and managing large datasets residing in distributed storage using SQL. Apache Hive provides a mechanism to project structure onto the data and query it using a SQL-like language called HiveQL. With Hive, we will execute aggregate functions, perform complex analyses, and extract meaningful patterns from the Starbucks datasets. The subsequent sections will delve into the processes undertaken within Apache Pig for data preparation, followed by an exploration of the analytical insights derived from PIg Latin and Hive queries.

2. Data Cleaning & Processing

Data cleaning and processing phase was critical in ensuring the reliability and accuracy of our analysis. The cleaning process for the Starbucks drinks dataset was conducted using Apache Pig's data manipulation capabilities. There was two major issues that I had to process in order to proceed with the next step for data analysis:

1) Converting from UTF-16 to UTF-8:

While loading in the Starbucks food dataset, I encountered a character encoding issue. The dataset was encoded in UTF-16, a format that is not natively compatible with Apache Pig and Hive. These tools are optimised for UTF-8 encoding. Therefore, I created a python code in order to change the encoding format.

2) Handling missing values:

- The raw data was initially loaded from a CSV file using Pig's LOAD function.
- The dataset contained placeholder characters ('-') representing missing values. To clean the data, a FOREACH operation was applied to iterate over each record. A conditional GENERATE statement was used to replace these placeholder characters with NULL, which is more suitable for representing missing data in subsequent analytical processes.
- The cleaned dataset was stored back into the filesystem in a designated cleaned data directory as a CSV file using Pig's STORE function. This ensures that the cleaned data is available for future analysis and can be easily accessed and queried.

With the cleaning and processing complete. We are ready to proceed to the next step of analysing the data with Apache Pig and Hive.

3. Data querying

I started with two straightforward queries in both Apache Pig and Hive, aiming to perform basic analysis and verify the accuracy of our data. By running these initial queries on both platforms, I was able to cross-check and confirm the consistency of the results. For the more advanced analysis, I turned to Hive, where I first loaded the cleaned datasets from Pig into new Hive tables. A particularly valuable command that I found was DESCRIBE, which displayed the structure of the Hive tables along with the column names. This was a simple yet useful command that helped me quickly grasp the structure of the tables.

I've chosen to focus on the theme of a healthy lifestyle, directing my queries towards extracting data that could inform and support nutritious dietary choices. These queries are designed to identify food and drink options that align with a balanced diet.

1) Simple queries:

I. Top 10 Beverages with the highest calories

This query selects the beverage name and calorie information from the drinks_cleaned table and displays the top ten entries with the highest calorie counts in descending order.
There are some slight differences in the outputs, this is due to the order that Pig and Hive have run and the output limit being set to 10.. Despite the variations in the last output, the calorie is still the same.

(a) Pig query result:

Outputs > Pig_outputs > Top10BevWithHighestCal_output > part-r-00000

```
1   Starbucks® Signature Hot Chocolate,430
2   White Chocolate Mocha,360
3   Cinnamon Dolce Frappuccino® Blended Coffee,350
4   Chocolate Smoothie,320
5   Hot Chocolate,320
6   Iced White Chocolate Mocha,300
7   Strawberry Smoothie,300
8   Caffè Mocha,290
9   Mocha Frappuccino® Blended Coffee,280
10  Cinnamon Dolce Latte,260
```

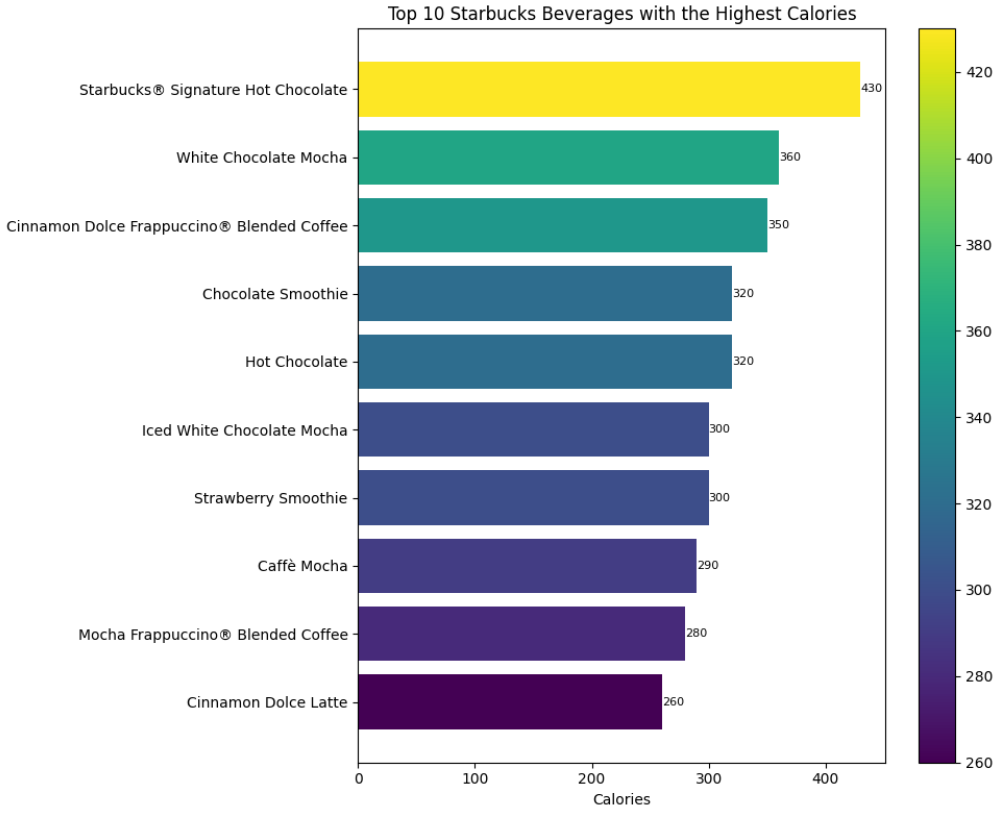
(b) Hive query result:

Outputs > Hive_outputs > Simple_Queries > Top10BevWithHighestCal

```
1   Starbucks® Signature Hot Chocolate,430
2   White Chocolate Mocha,360
3   Cinnamon Dolce Frappuccino® Blended Coffee,350
4   Hot Chocolate,320
5   Chocolate Smoothie,320
6   Strawberry Smoothie,300
7   Iced White Chocolate Mocha,300
8   Caffè Mocha,290
9   Mocha Frappuccino® Blended Coffee,280
10  Iced Coconutmilk Mocha Macchiato,260
```

This information is not only beneficial for personal health objectives but can also serve as a benchmark for Starbucks to consider when developing new, health-conscious alternatives for their menu.

- Visualisation graph:



From the output bar chart, we can see that the beverage with the highest calories is Hot chocolate(430 cal), followed by White chocolate mocha(360 cal). These two both contain chocolate which could be the reason they contain such a high calorie.

II. Average sugar for each beverage category

This query selects the beverage category and uses the AVG function to find the average sugar for each. The result is then sorted from highest to lowest. As we can see, the outputs are exactly the same which means that there are no errors encountered while carrying out the analysis.

(a) Pig query result

Outputs > Pig_outputs > AvgSugarForEachBevCat_output > part-m-00000 > part-r-00000

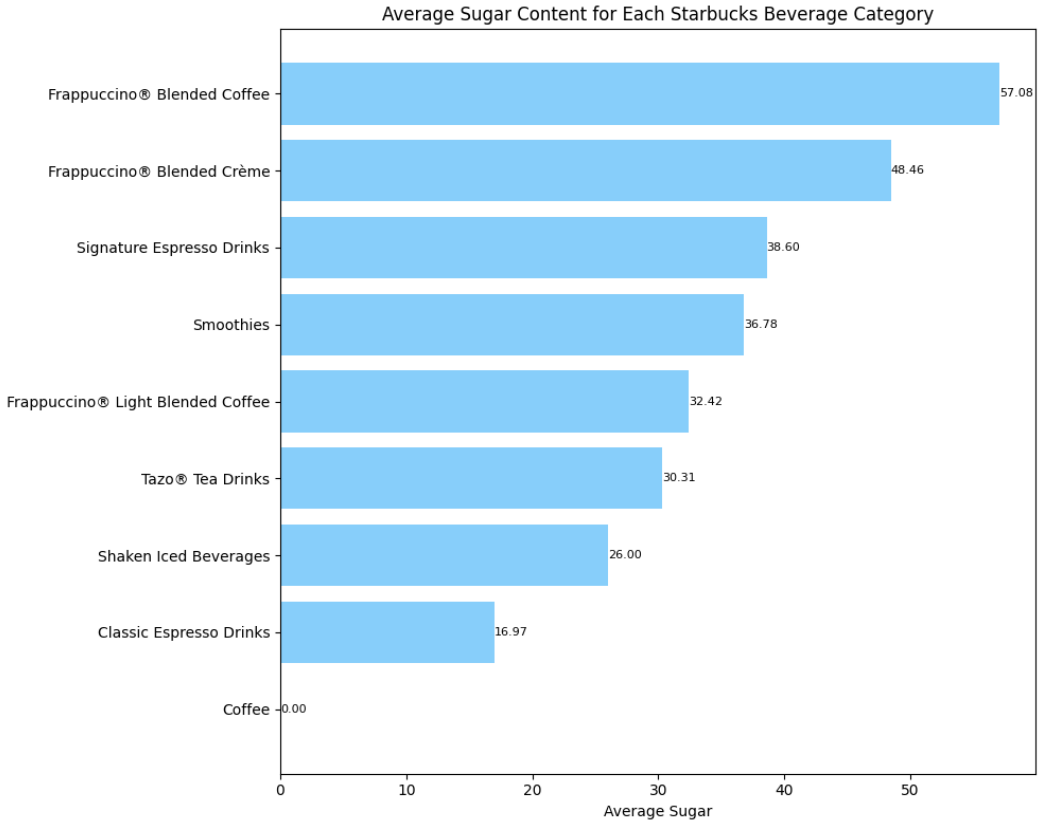
```
Frappuccino® Blended Coffee,57.083333333333336
Frappuccino® Blended Crème,48.46153846153846
Signature Espresso Drinks,38.6
Smoothies,36.77777777777778
Frappuccino® Light Blended Coffee,32.416666666666664
Tazo® Tea Drinks,30.307692307692307
Shaken Iced Beverages,26.0
Classic Espresso Drinks,16.96551724137931
Coffee,0.0
```

(b) Hive query result

```
Outputs > Hive_outputs > Simple_Queries > Avg_sugar_for_each_bev_cat
1   Frappuccino® Blended Coffee,57.083333333333336
2   Frappuccino® Blended Crème,48.46153846153846
3   Signature Espresso Drinks,38.6
4   Smoothies,36.77777777777778
5   Frappuccino® Light Blended Coffee,32.416666666666664
6   Tazo® Tea Drinks,30.307692307692307
7   Shaken Iced Beverages,26.0
8   Classic Espresso Drinks,16.96551724137931
9   Coffee,0.0
```

This information is especially valuable for those looking to enjoy Starbucks' offerings while maintaining a balanced diet or managing conditions such as diabetes.

- Visualisation graph:



From the output graph, we can see that the two beverage categories with the highest average sugar are Frappuccino Blended Coffee and Frappuccino Blended Creme. We also gather the fact that coffee is the most healthy as it contains 0 sugar.

2) Complex queries:

I. Beverages with Caffeine

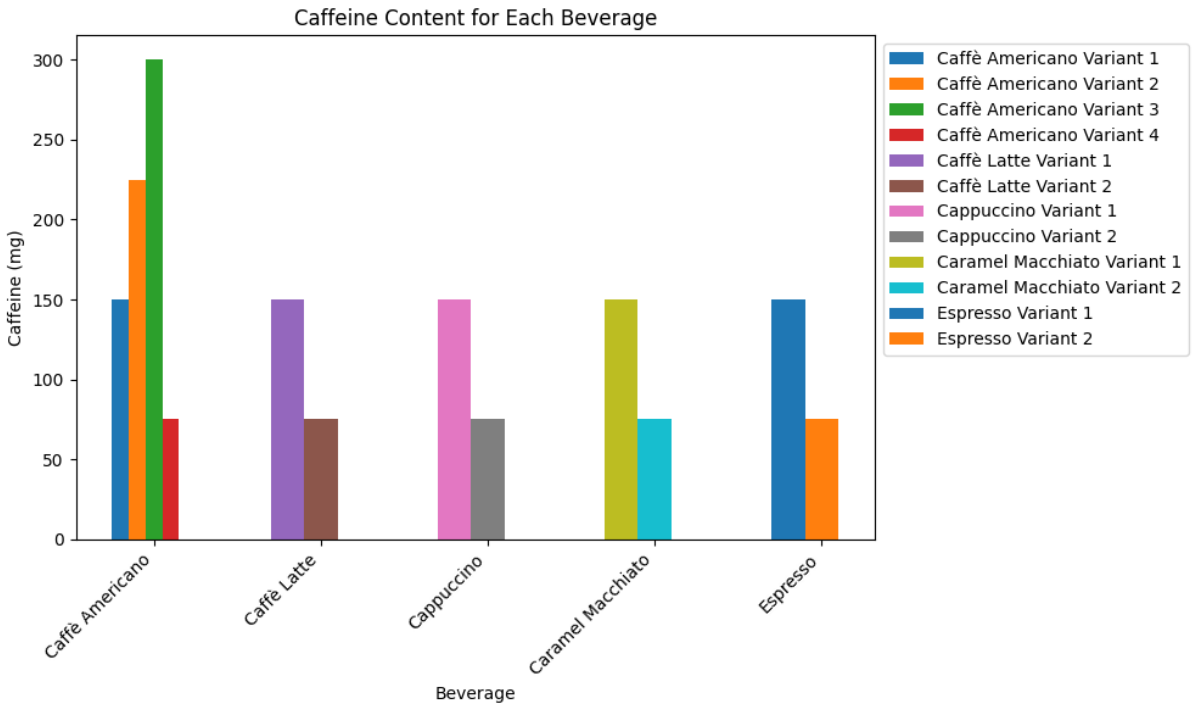
This query joins the expanded drinks dataset onto the drinks dataset. After joining it retrieves only the beverages with caffeine.

Outputs > Hive_outputs > Complex_Queries > Find_caffein

```
Beverage,Calories,Fat (g),Carb. (g),Fiber (g),Protein,Sodium,Caffeine (mg)
Caffè Americano,,,,,,,150
Caffè Americano,,,,,,,225
Caffè Americano,,,,,,,300
Caffè Americano,,,,,,,75
Caffè Latte,190,7,19,0,13,170,150
Caffè Latte,190,7,19,0,13,170,75
Cappuccino,120,4,12,0,8,100,150
Cappuccino,120,4,12,0,8,100,75
Caramel Macchiato,250,7,35,0,10,150,150
Caramel Macchiato,250,7,35,0,10,150,75
Espresso,,,,,,,150
Espresso,,,,,,,75
```

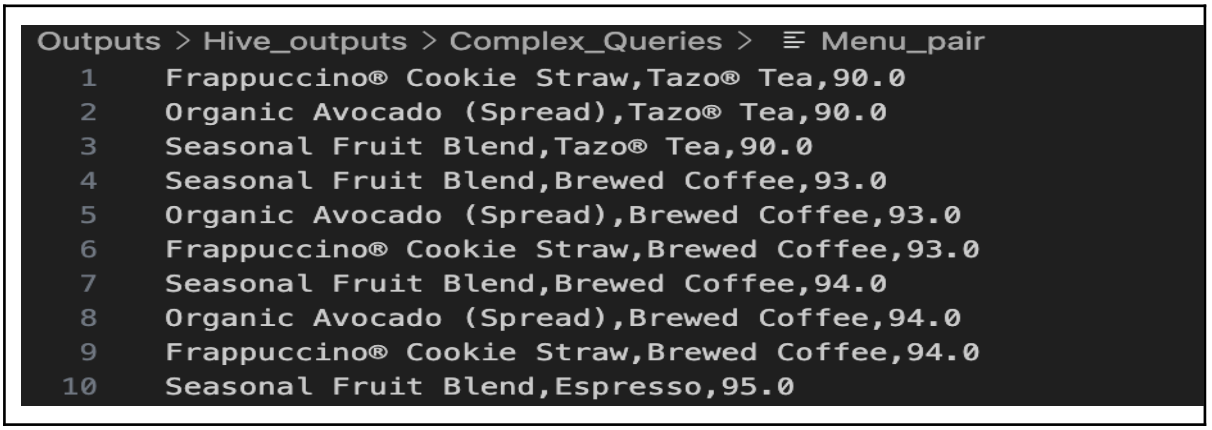
This information is beneficial to those who are sensitive to caffeine, looking to limit their intake for health reasons, or trying to avoid it altogether due to personal preferences or dietary restrictions will find it helpful to know which drinks contain caffeine and in what quantities.

- Visualisation graph:



II. Menu pair combinations

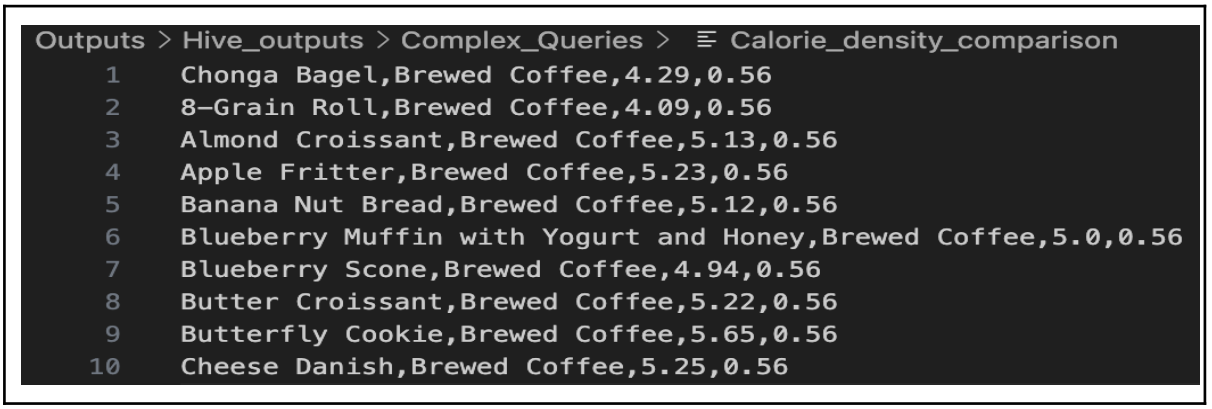
This query combines food items from the foods table and beverages from the expanded drinks table to calculate the total calorie count for each possible pair. It uses a CROSS JOIN to match every food item with every beverage item, creating a comprehensive list of combinations. The query then filters these combinations to include only those with a total calorie count of 500 or less and orders the results by total calories in ascending order.



To align with dietary considerations, specifically for those adhering to a calorie-restricted diet, the query applies a filter to retain only those pairings where the combined calorie count does not exceed 500. This threshold serves to highlight meal options suitable for consumers who are calorie-conscious.

III. Calorie Density comparison

This query performs a cross join between the foods table and expanded drinks table to calculate the caloric density for each food and drink item pairing. The resulting dataset includes rounded caloric densities for each possible combination of food and beverage items.



The output would be a valuable resource for consumers when they are looking to understand the energy content of their meal choices. This output can be implemented onto the Starbucks app to allow consumers to access easily.

IV. Sampling

This query chooses a random sample of 10% from each of the tables, with a condition that the calories is no more than 300.

Outputs > Hive_outputs > Complex_Queries > Sampling > drinks

```
Drink,Tazo® Bottled Tazoberry,150.0
Drink,Clover® Brewed Coffee,10.0
Drink,Iced Vanilla Latte,190.0
Drink,Mocha Light Frappuccino® Blended Coffee,140.0
```

Outputs > Hive_outputs > Complex_Queries > Sampling > drinks_expanded

```
Expanded Drink,Caffè Latte,70.0
Expanded Drink,Caffè Latte,110.0
Expanded Drink,Caffè Mocha (Without Whipped Cream),130.0
Expanded Drink,Caffè Mocha (Without Whipped Cream),170.0
Expanded Drink,Vanilla Latte (Or Other Flavoured Latte),250.0
Expanded Drink,Cappuccino,120.0
Expanded Drink,Espresso,10.0
Expanded Drink,Skinny Latte (Any Flavour),60.0
Expanded Drink,Hot Chocolate (Without Whipped Cream),130.0
Expanded Drink,Hot Chocolate (Without Whipped Cream),240.0
```

Outputs > Hive_outputs > Complex_Queries > Sampling > food

```
Food,Petite Vanilla Bean Scone,120.0
Food,The Classic - Bantam Bagels (2 Pack),200.0
Food,Justin's Classic Almond Butter,190.0
```

This sampling can be useful for when consumers are up for a surprise and can let the system randomly choose what they would be having.