**Snack Analysis Project:**

**Challenges Faced:**

**1.While attempting to convert a Pandas DataFrame (starbucks3) into a PySpark DataFrame using the following command:**

starbucks3\_sdf=spark.createDataFrame(starbucks3)

an error occurred:

**PySparkTypeError: [CANNOT\_MERGE\_TYPE] Can not merge type DoubleType and LongType.**

The Excel file **starbucks\_drinkMenu\_expanded.xlsx** contains columns with mixed numeric data types:

For example, the same column might have:

one row with an integer (120) → inferred as LongType  
 another with a float (180.0) → inferred as DoubleType

**2.Problem: Schema Mismatches Across Multiple Datasets**

While integrating three different Starbucks datasets, inconsistencies in **column names and data types** caused merging issues.

In each file column names varies like ,

Fat and Fat(g), Total Fat,

Total Carbohydrates (g), carbs (g)

Protein, Protein (g)

Even if some columns are with same names but has different datatypes in each file of starbucks

Calories→ double in one file, but long in others  
 Fat/Fat (g) → string in one file, double in others  
 Protein → string in one file, long in second file, double in third file

These inconsistencies made it impossible to union the dataframes directly, as Spark requires aligned schemas.