Assignment

1. Write PySpark code to create a new dataframe with the data given below having 2 columns (‘season’) and (‘windspeed’).

[Datatypes of the column names can be inferred]

```
# Data
 [("Spring", 12.3),
  ("Summer", 10.5),
  ("Autumn", 8.2),
  ("Winter", 15.1)]
```

2. Consider the library management dataset located at the following path **(/public/trendytech/datasets/library_data.json)**. Using PySpark, load the data into a Dataframe and enforce schema using StructType.

3. Given the dataset **(/public/trendytech/datasets/train.csv)**, create a Dataframe using PySpark and perform the following operations

a) Drop the columns passenger_name and age from the dataset.

b) Count the number of rows after removing duplicates of columns train_number and ticket_number.

c) Count the number of unique train names.

4. You are working as a Data Engineer in a large retail company. The company has a dataset named "sales_data.json" that contains sales records from various stores. The dataset is stored in JSON format and may have some corrupt or malformed records due to occasional data quality issues.

Your task is to read the "sales_data.json" dataset **(/public/trendytech/datasets/sales_data.json)** using PySpark, utilizing different read modes to handle corrupt records. You need to create a Dataframe using pyspark and perform the following operations:

1. Read the dataset using the "permissive" mode and count the number of records read.

2. Read the dataset using the "dropmalformed" mode and display the number of malformed records.
3. Read the dataset using the "failfast" mode.

5. You have a hospital dataset with the following fields:

- patient_id (integer): Unique identifier for each patient.
- admission_date (date): The date the patient was admitted to the hospital. (MM-dd-yyyy)
- discharge_date (date): The date the patient was discharged from the hospital. (yyyy-MM-dd)
- diagnosis (string): The diagnosed medical condition of the patient.
- doctor_id (integer): The identifier of the doctor responsible for the patient's care.
- total_cost (float): The total cost of the hospital stay for the patient.

Using PySpark, load the data into a Dataframe and perform the following operations on the hospital dataset **(/public/trendytech/datasets/hospital.csv)**:

1. Drop the "doctor_id" column from the dataset.
2. Rename the "total_cost" column to "hospital_bill".
3. Add a new column called "duration_of_stay" that represents the number of days a patient stayed in the hospital. (hint: The duration should be calculated as the difference between the "discharge_date" and "admission_date" columns.)
4. Create a new column called "adjusted_total_cost" that calculates the adjusted total cost based on the diagnosis as follows:
   If the diagnosis is "Heart Attack", multiply the hospital_bill by 1.5.
   If the diagnosis is "Appendicitis", multiply the hospital_bill by 1.2.
   For any other diagnosis, keep the hospital_bill as it is.
5. Select the "patient_id", "diagnosis", "hospital_bill", and "adjusted_total_cost" columns.