PySpark Module 4 Assignment

Problem Statement:

Consider yourself as a Data Scientist at a prestigious firm, the clients of which are leading MNCs. Presently, you are given with the data from one of the clients. You have to process the data and also find insights using Apache Spark in Python.

New York City Airbnb Open Data: This dataset is publicly available. This data file includes all needed information to find out more about hosts, geographical availability, and necessary metrics to make predictions and draw conclusions.

Lab Environment: Jupyter notebook

Tasks to be Done

Using this dataset, we will perform the following tasks:

1. Load the dataset
2. Print the first 10 rows
3. Find the total number of private rooms in the 'room_type' column
4. Find the max, min, and average of the price column
5. Find the number of rooms available for booking for less than 200 days a year (use the 'availability_365' column)
6. Find 10 host places that have the most number of reviews

Dataset Description:

- id: The listing ID
- name: The name of the listing
- host_id: The host ID
- host_name: The name of the host
- neighbourhood_group: Location
- neighbourhood: Area
- latitude: The latitude coordinates
- longitude: The longitude coordinates
- room_type: The listing space type

- price: Price in dollars
- minimum_nights: The amount of nights minimum
- number_of_reviews: The number of reviews
- last_review: The latest review
- reviews_per_month: The number of reviews per month
- calculated_host_listings_count: The amount of listing per host
- availability_365: The number of days when listing is available for booking