DA 331 - Big Data Analytics : Tools & Techniques

Lab 2

Instructor: Dr. Chiranjib Sur

Problem 1:

Given the dataset, find the followings:

1. Find the average number of household for different geolocation in Spark?
2. Compare the time for execution for the above in Spark and in Pandas? Use the ”import time; # record start time ; start = time.time()” code to record the time.
3. Which geolocation has the highest median income in Spark?
4. Compare the time for execution for the above in Spark and in Pandas? Use the ”import time; # record start time ; start = time.time()” code to record the time.
5. Find the average median_house_value for different housing_median_age groups in Spark.
6. Which location (latitude and longitude) has the highest population?
7. Calculate the ratio of total_bedrooms to total_rooms for each record and find the average ratio.
8. Find the correlation between median_income and median_house_value in Spark.