DA 331 - Big Data Analytics : Tools & Techniques

Lab 2

Instructor:	Dr. Chiranjib Sur	**Time:**	Monday - 3:00-4:00 (5201) Tuesday - 2:00-3:00 (5201) Wednesday - 9:00-11:00 (MDSAI Lab)
Email:	chiranjib@iitg.ac.in	**Place:**	MDSAI Lab.

Problem 1:

Given the dataset, find the followings:

1. You want the biggest (average) trips, which location will you choose?
2. You want the biggest (average) trips, which location will you choose? (in pandas) Find the difference in running time.
3. Find the locations where the maximum number of passengers arrives.
4. Find the locations where the maximum number of passengers starts.
5. Find the locations where the maximum number of passengers starts in a day.

   You may need this to process the dates.

   ```
   from pyspark.sql.functions import year, month, day
   extracted_df = df.select(year(”date”).alias(”year”), /
   month(”date”).alias(”month”), day(”date”).alias(”day”))
   extracted_df.show()
   ```