Lab 05	Programming with Spark Dataframe APIs and Spark-SQL
IT413 No SQL Databases, Winter'2021, DAIICT, Gandhinagar; pm_jat	

In this lab, we do some exercise based on Spark Dataframe APIs and Spark-SQL. You can work either in Java, Python, or Javascript.

For Java we continue using Eclipse and Maven. Following are dependencies that you would be requiring for Spark.

```
<dependencies>
    <dependency>
        <groupId>org.apache.spark</groupId>
        <artifactId>spark-core_2.12</artifactId>
        <version>2.4.7</version>
    </dependency>
    <dependency>
        <groupId>org.apache.spark</groupId>
        <artifactId>spark-sql_2.12</artifactId>
        <version>2.4.7</version>
    </dependency>
</dependencies>
```

Submission Required: Source Code of solutions of all exercises in a single compressed file.

Exercises:

General instruction: Have separate source file for each question. Given question number as source file name. Keep complete solution of one question in single source file.

Perform following computation using Spark dataframe API

You are given a house prices file houses.csv; the file contains 20 records and attributes: ID, area, beds, baths, zip, year, and price.

1. List houses that are having at least 3 beds and 2 bath rooms.
2. List house_id, year, zip that are with area > 2000 sq ft and price <= 400000.
3. Produce List of zip, price_per_sq_ft in descending order of price_per_sq_ft

You are given two data file as following

Order (OrderNo, CustomerID, OrderAmount)
Customer(CustomerID, Country, State)

4. Produce a Summary Report that list State, Total Revenue generated in that state. Perform following computation using Spark Dataframe API

Spark SQL

5. Solve questions 1 to 4 using Spark SQL. Place all questions solutions in a single source file.