**Alekhya Krishna Balivada Pyspark Coding Assessment (27-12-2023)**

**2Q. Explain ETL (Extract,Transform,Load) with pyspark**

**Ans:**

ETL (Extract,Transform,Load) is a process where we can extract the data from various sourses like databases, CSV files, APIs and so on then transform that data into the required format later on we can load that data into database.

ETL, which stands for extract, transform and load, is a data integration process that combines data from multiple data sources into a single, consistent data store that is loaded into a [data warehouse](https://www.ibm.com/topics/data-warehouse) or other target system.

Pyspark which is a python API for Apache Spark is widely used for performing ETL tasks.

**Extract:** During the extraction phase, we use various methods to connect to the sources and retrieving the data for further processes and transforming and loads the data. Each sourse might require specific connectors, APIs and databases.

**Transform:** In the transform stage, the raw data undergoes data processing. Here, the data is transformed and used for analytical use for data. This phase can involve the following tasks:

* Filtering, cleansing, eliminating redundant values, validating, and authenticating the data.
* Performing calculations based on the raw data. This can  include changing row and column headers for consistency and more.
* Formatting the data into tables or joined tables to match the schema of the target data warehouse.

**Load:** In this last step, the transformed data is moved from the staging area into a database. Typically, this involves an initial loading of all data, followed by further data changes, full refreshes to erase and replace data in the warehouse. Mostly ETL is automated and well-defined.

**3Q. Using Spark SQL – creating databases, tables**

**Ans:**

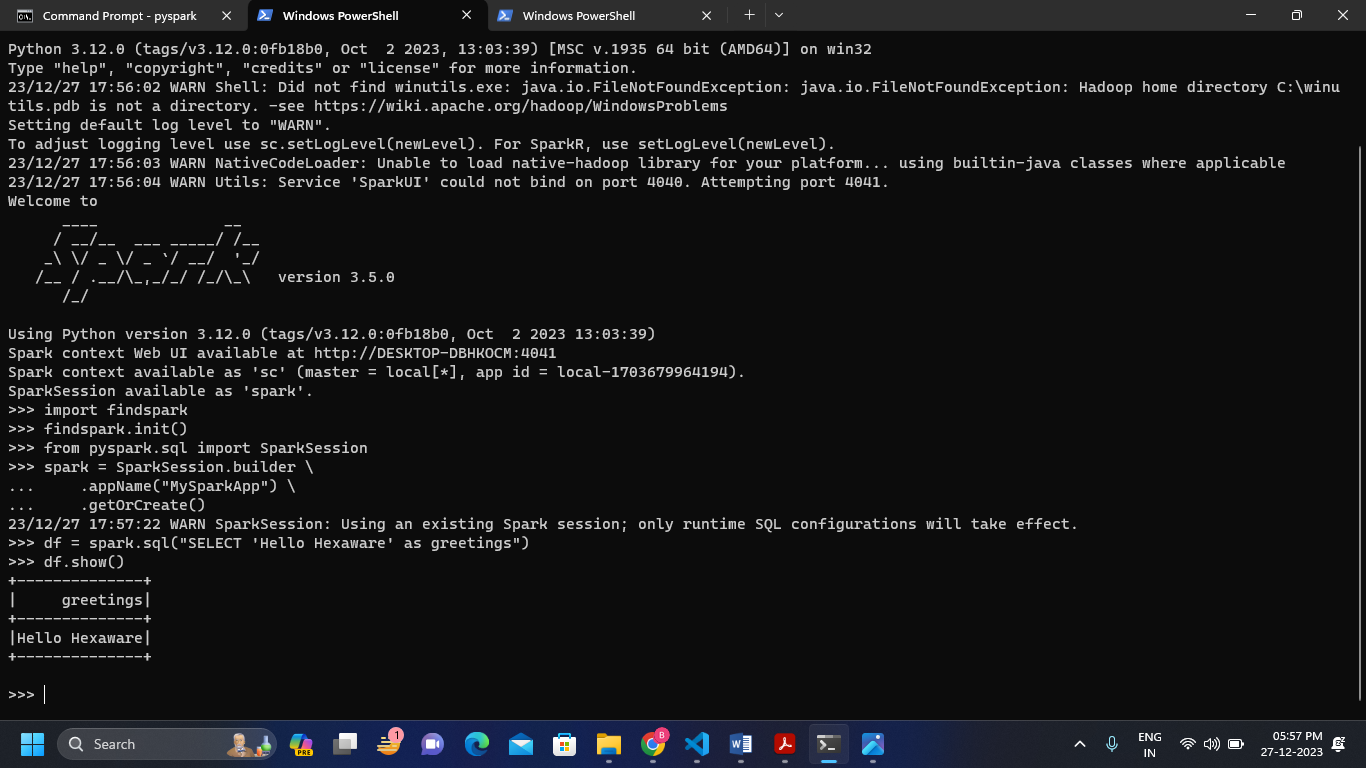
Step-1: Launch the Pyspark.

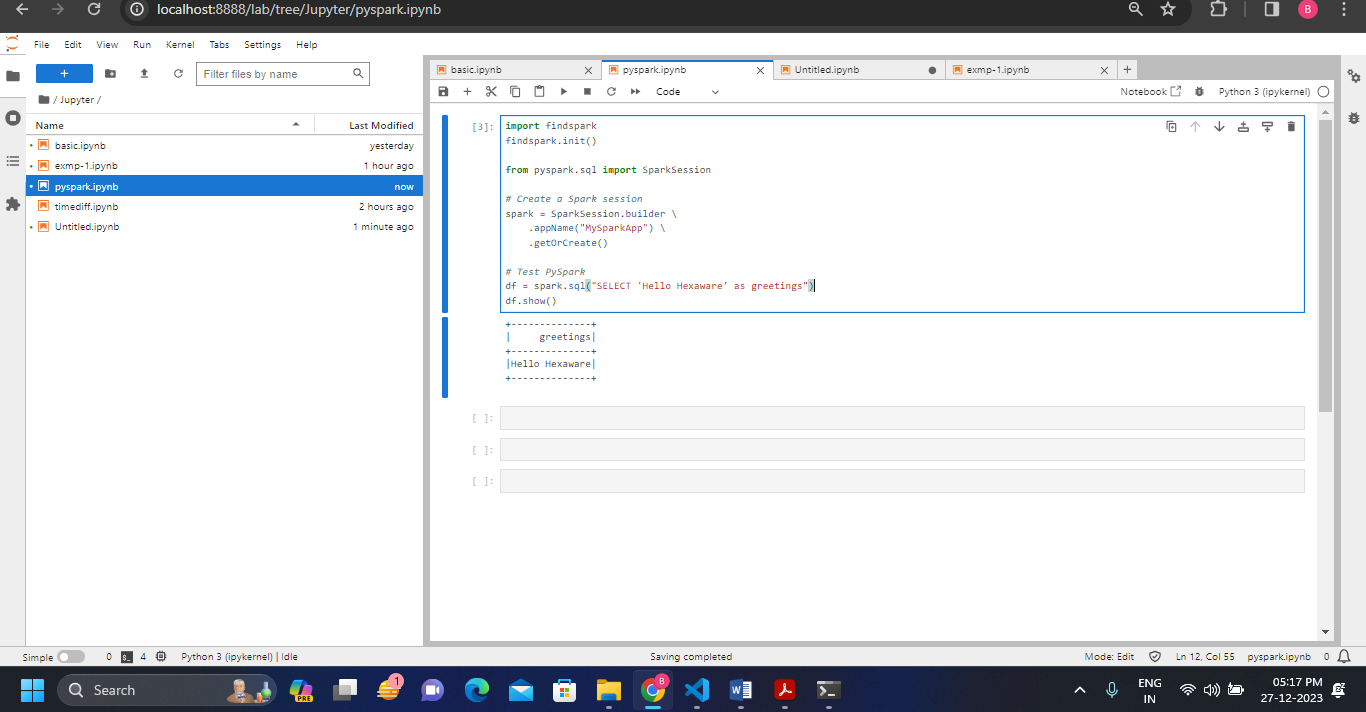
Step-2: Import the SparkSession from pyspark.sql

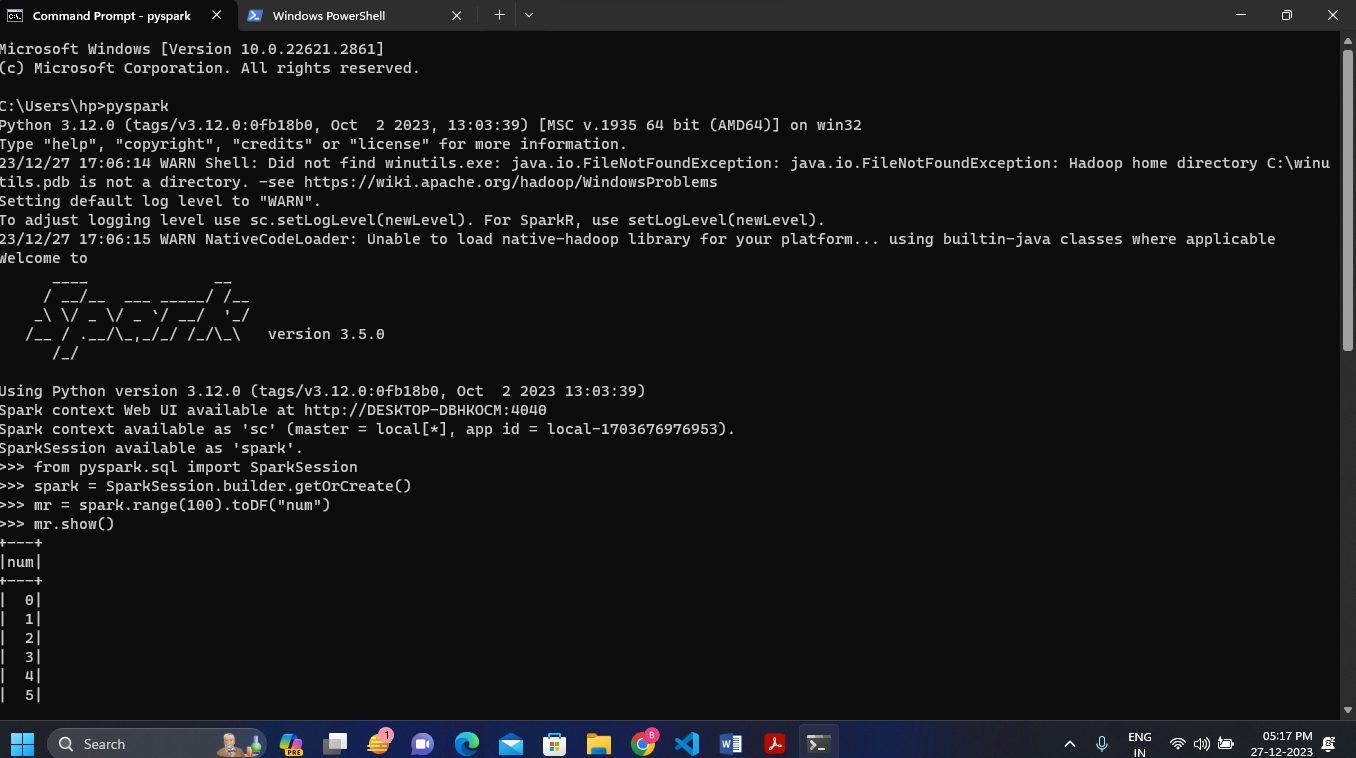
Step-3: Create a database by using getOrCreate command like “SparkSession.builder.getOrCreate ( )”.

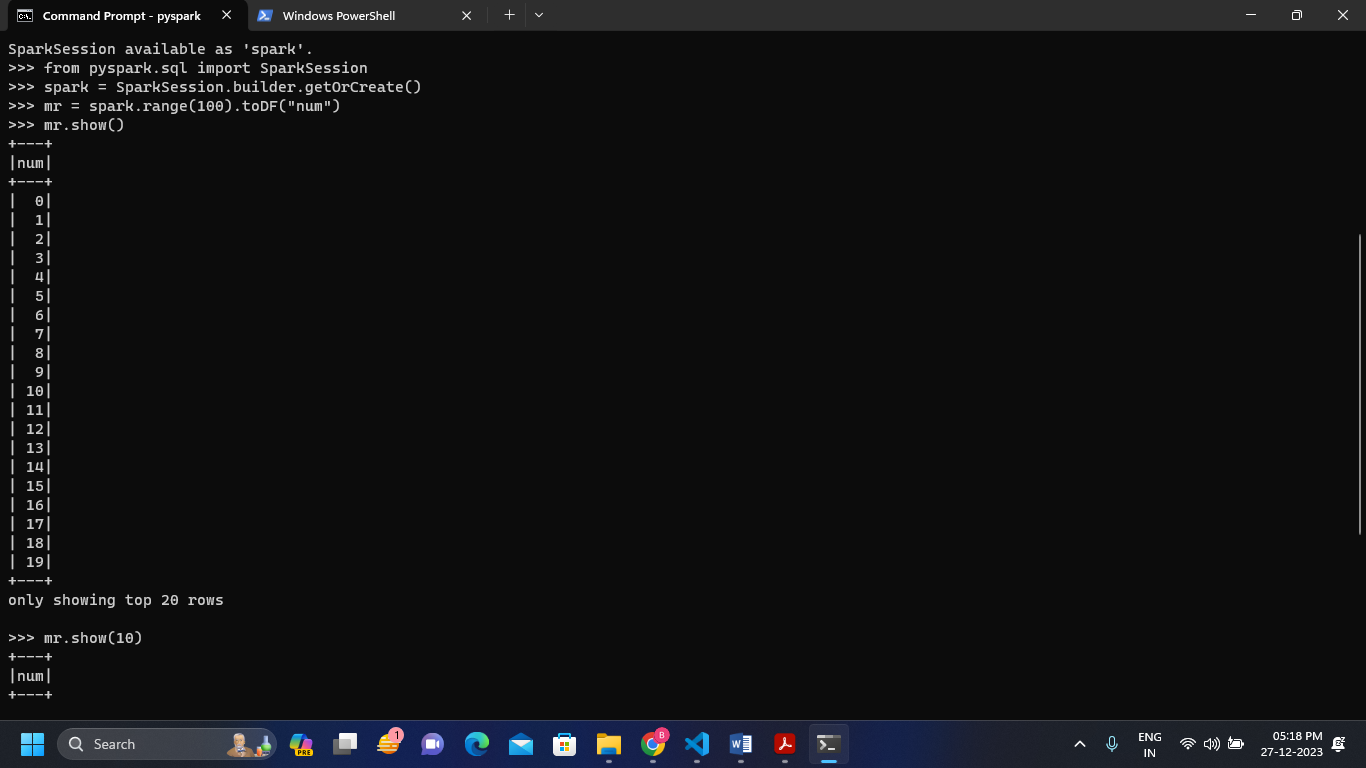
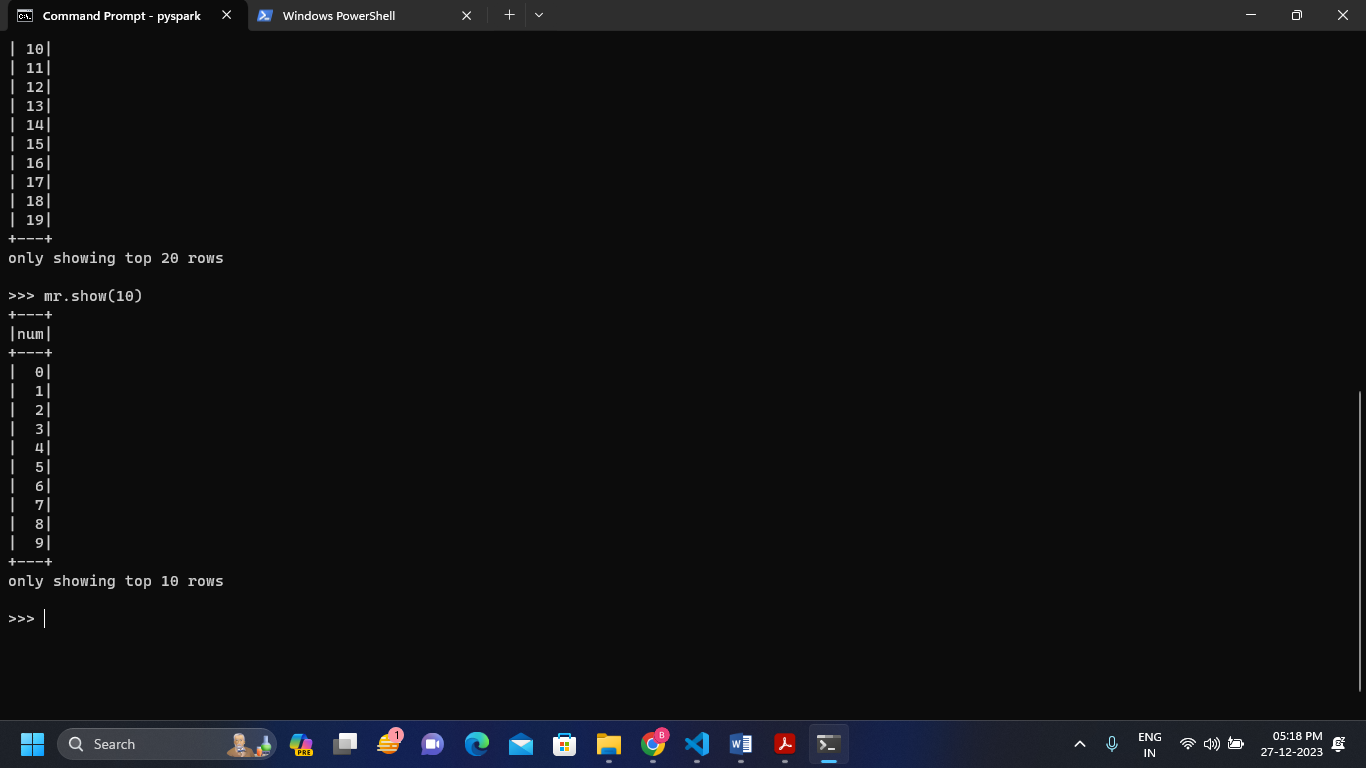
Step-4: Create the database.

In this example I created a dataframe and printed “hello hexaware” in it.

**In Command Prompt**

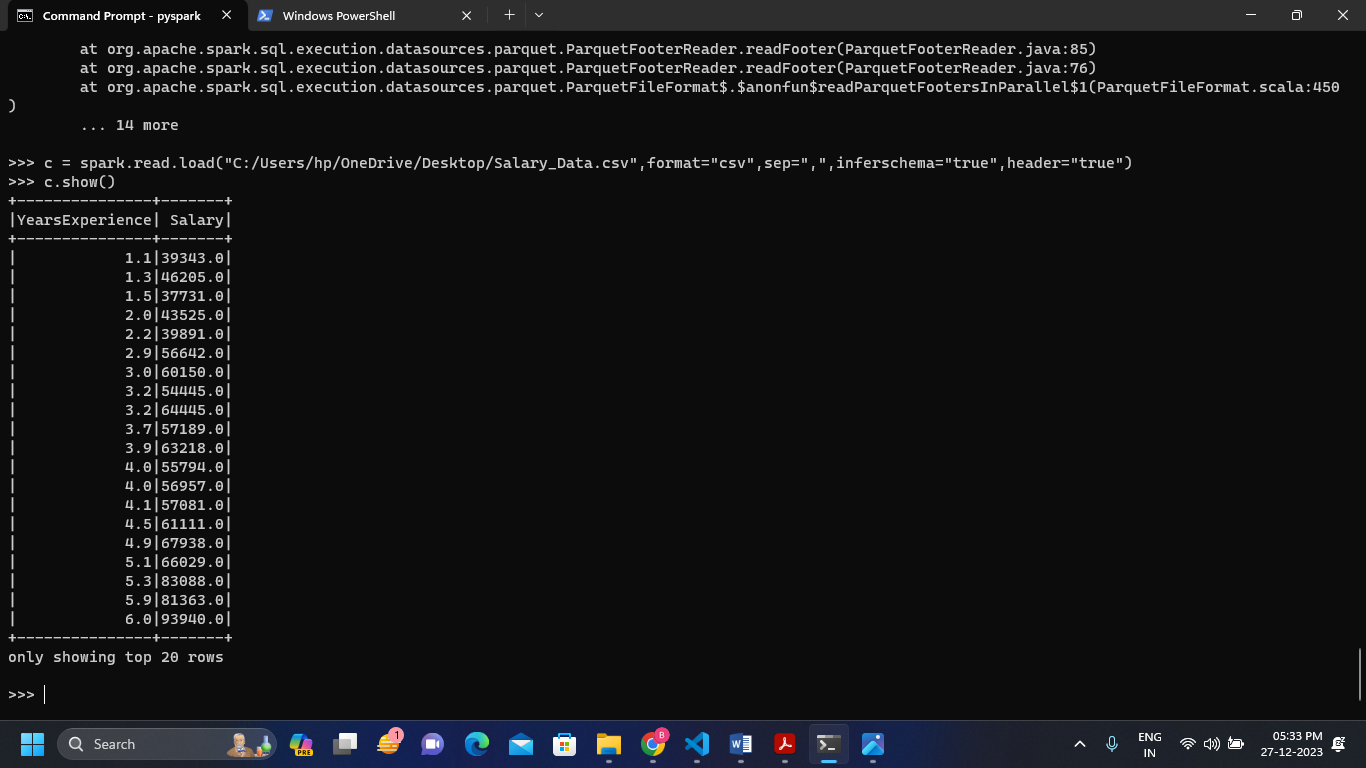
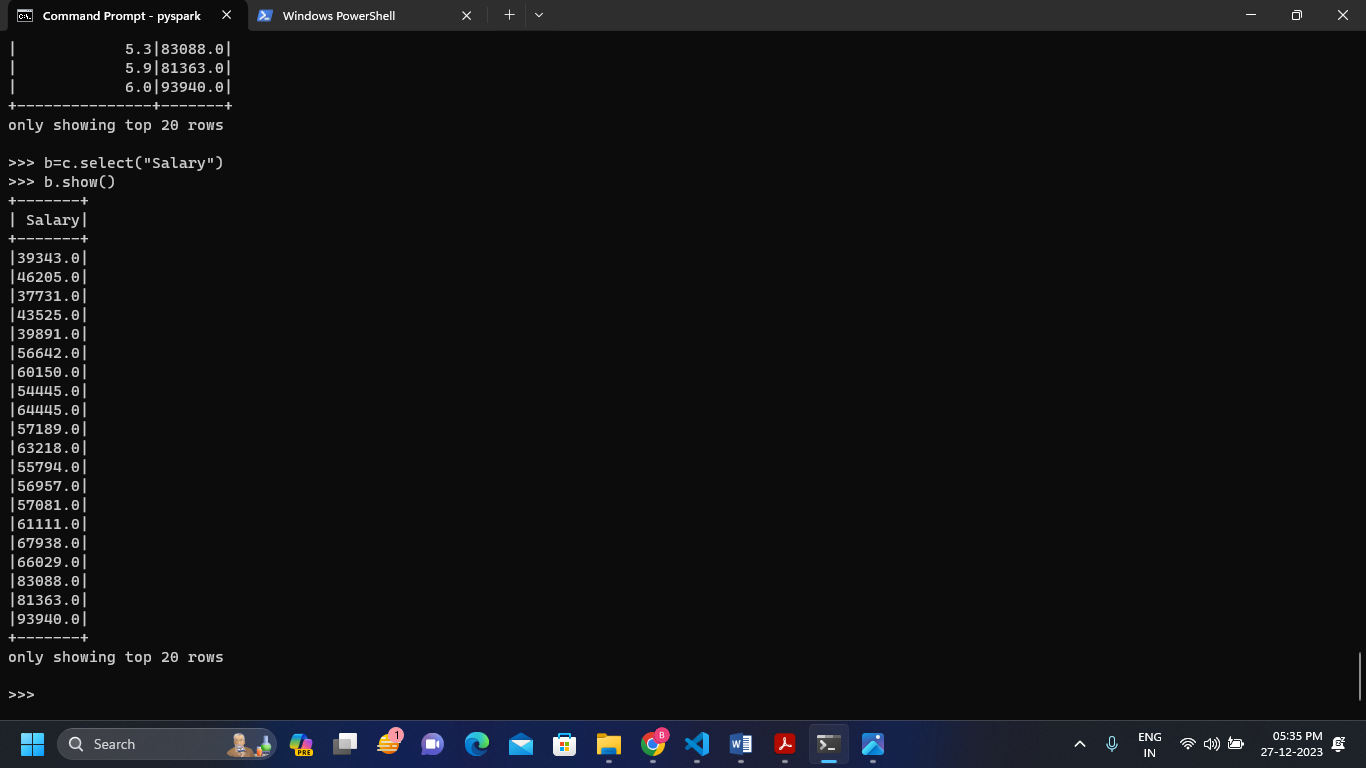
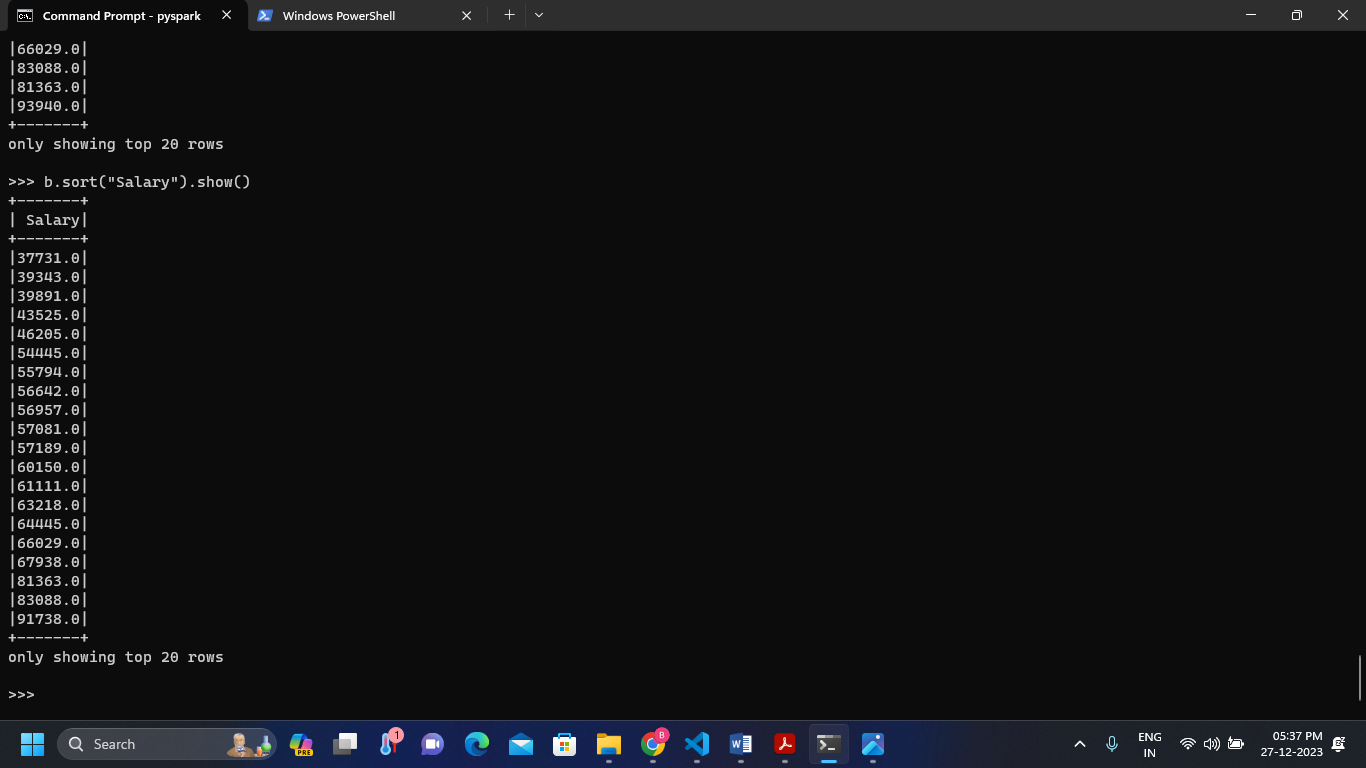
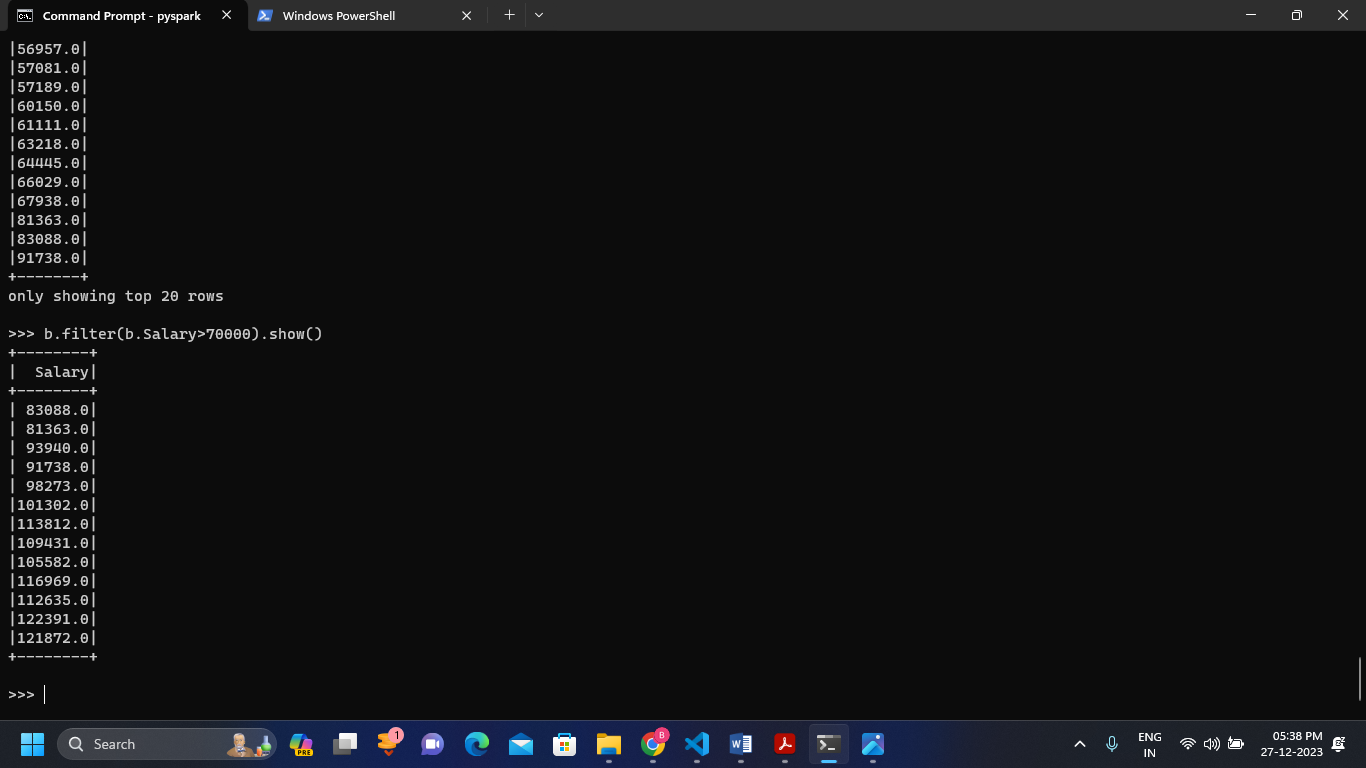
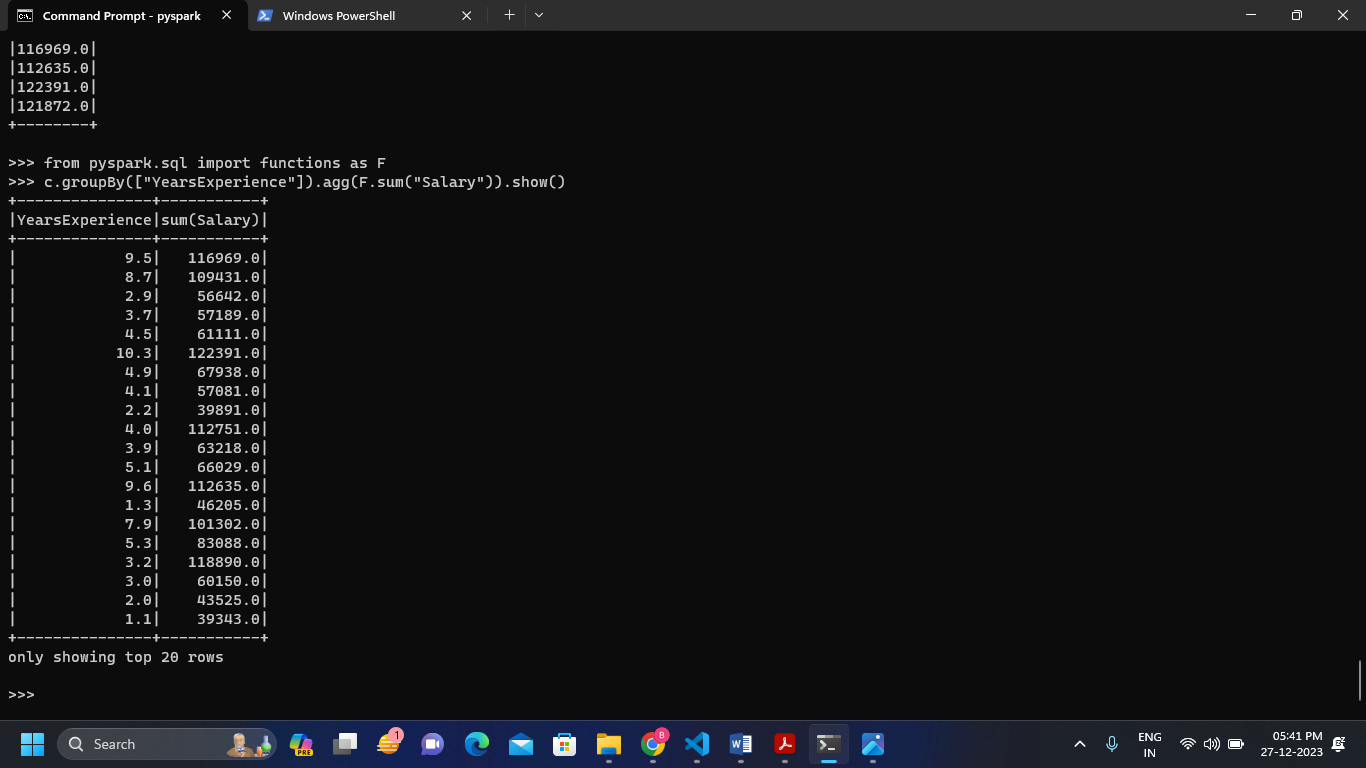
**In Jupyter-Lab**

In this example I created the dataframe as num and printed the range of 100 numbers which shows only the top 20 rows and 10 rows.



**4Q. Using Spark SQL – Transformations such as Filter, join, Simple aggregation and Group by**

**Ans:**

1. Displaying the CSV file and applying the transformations over it
2. Selecting the salary column and displaying it
3. Sorting the salary in ascending order.
4. Filtering the salary greater than 70000 and displaying it.
5. Using groupby and applying aggregate operations as per years of experience.

**Choice**

**1Q. Implements processing JSON and CSV data with Pyspark**

**Ans:**

We can read the CSV file by using the following command as

“spark.read.load

( “Path of the CSV file”,format=“CSV”,sep=”,”,inferschema=”true”,header=”true” )”

