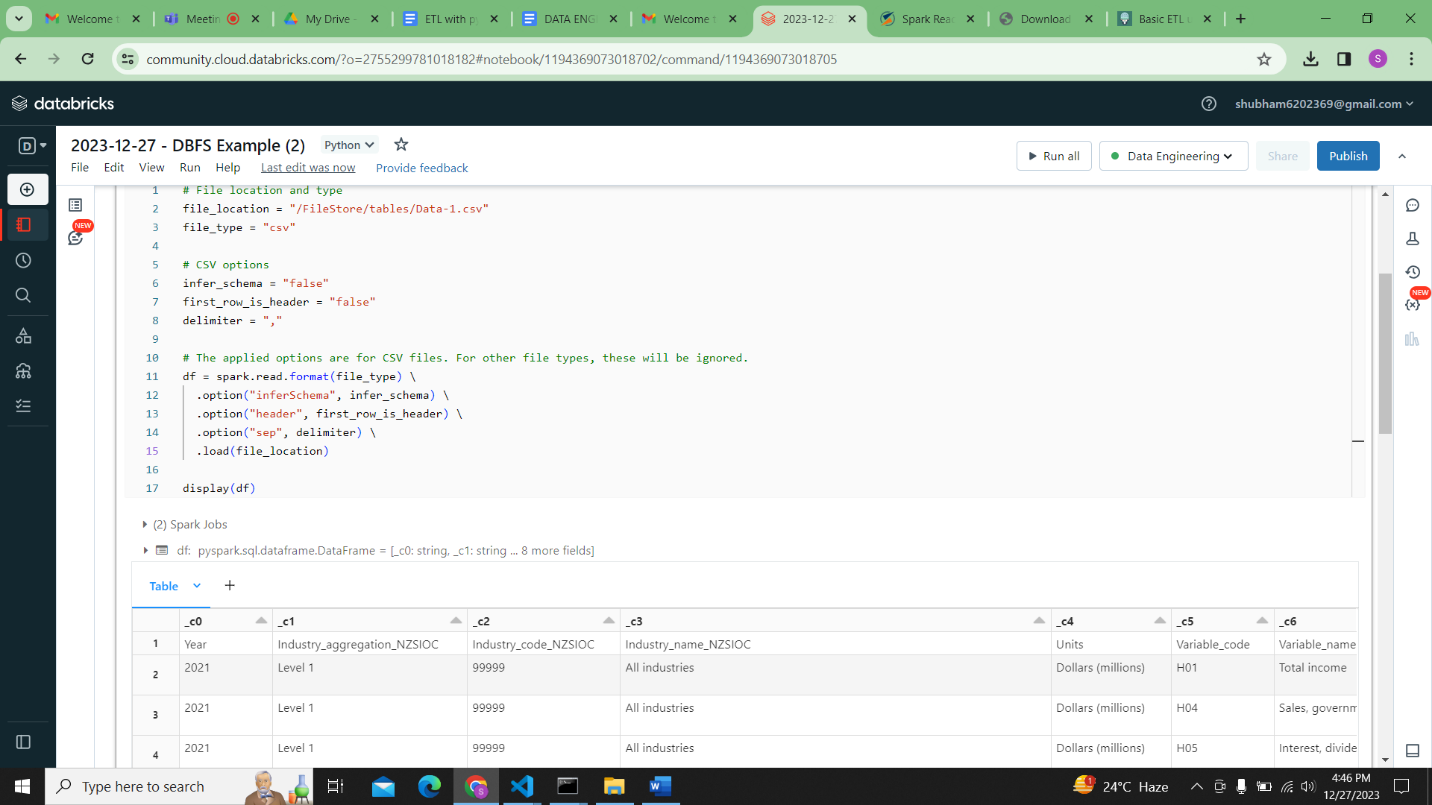
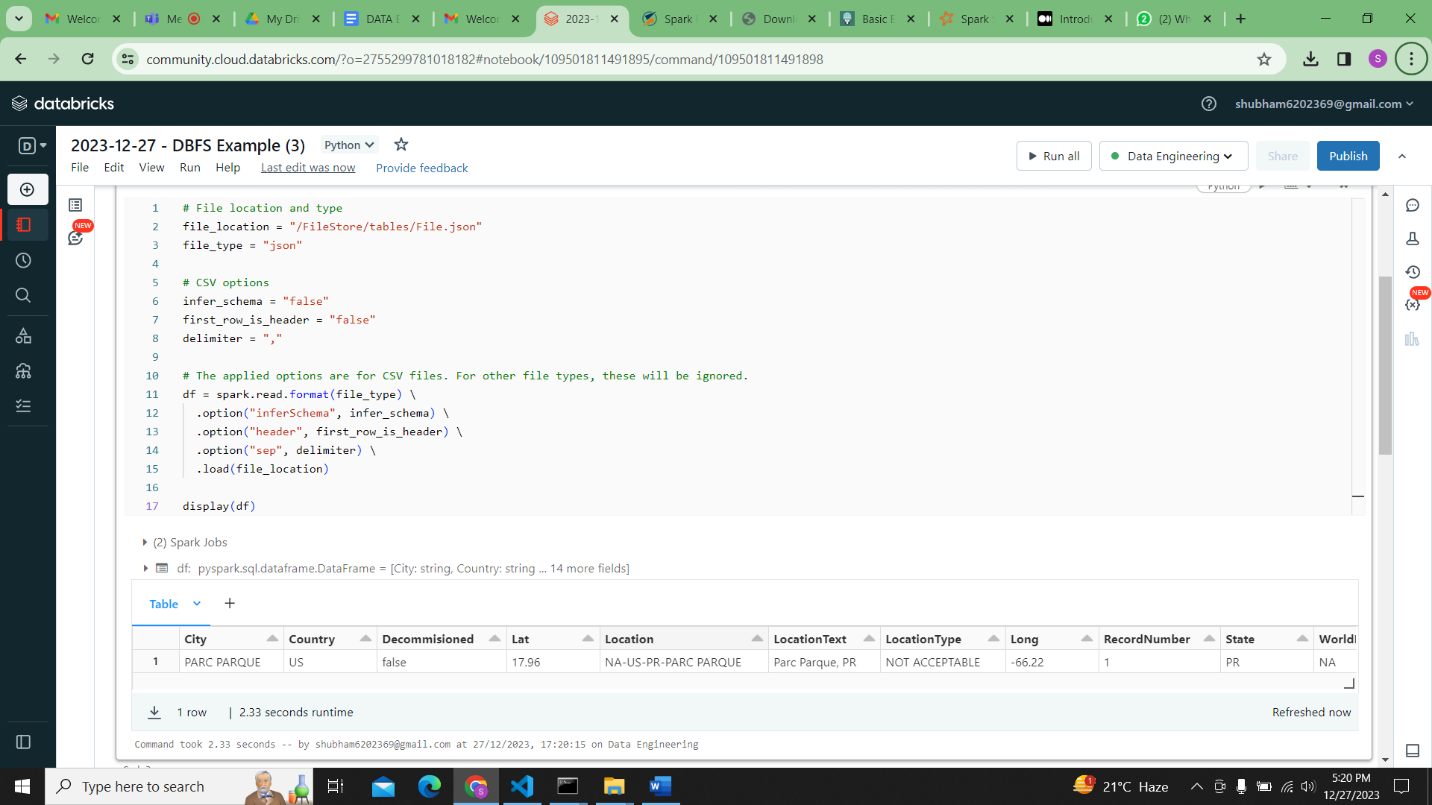
BIBA-PYSPARK CODING ASSESSMENT

Date: -27/12/2023

1. **Implement Processing JSON and CSV data with Pyspark.**

**Implementing csv file here**

Implementing JSON File

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

* ETL: 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 warehousing or other target system.
* Extract data from legacy systems
* Cleanse the data to improve data quality and establish consistency
* Load data into a target database
* How ETL works:

1. Extract: During data extraction, raw data is copied or exported from source locations to a staging area. Data management teams can extract data from a variety of data sources, which can be structured or unstructured. Those sources include but are not limited to:

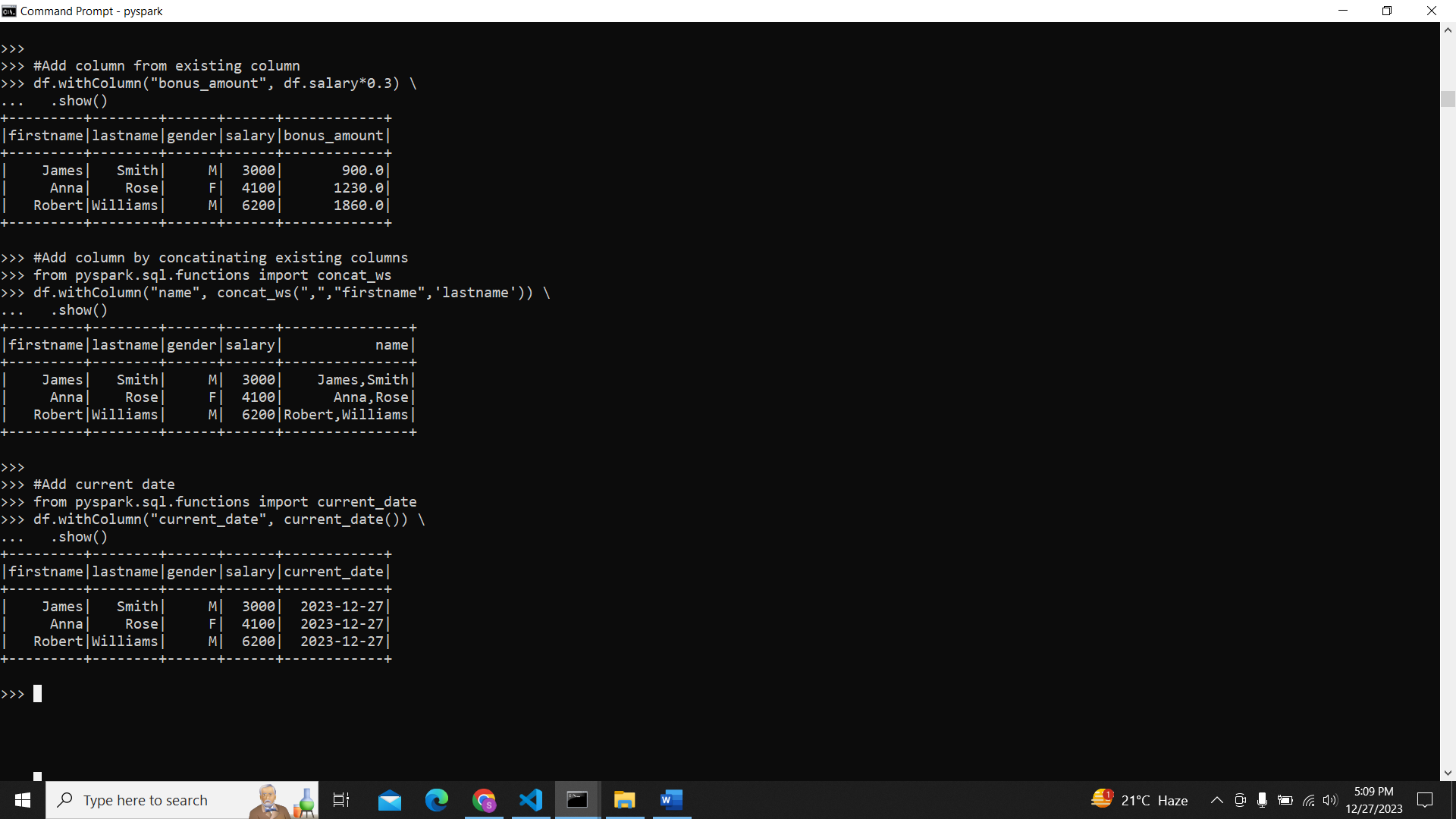
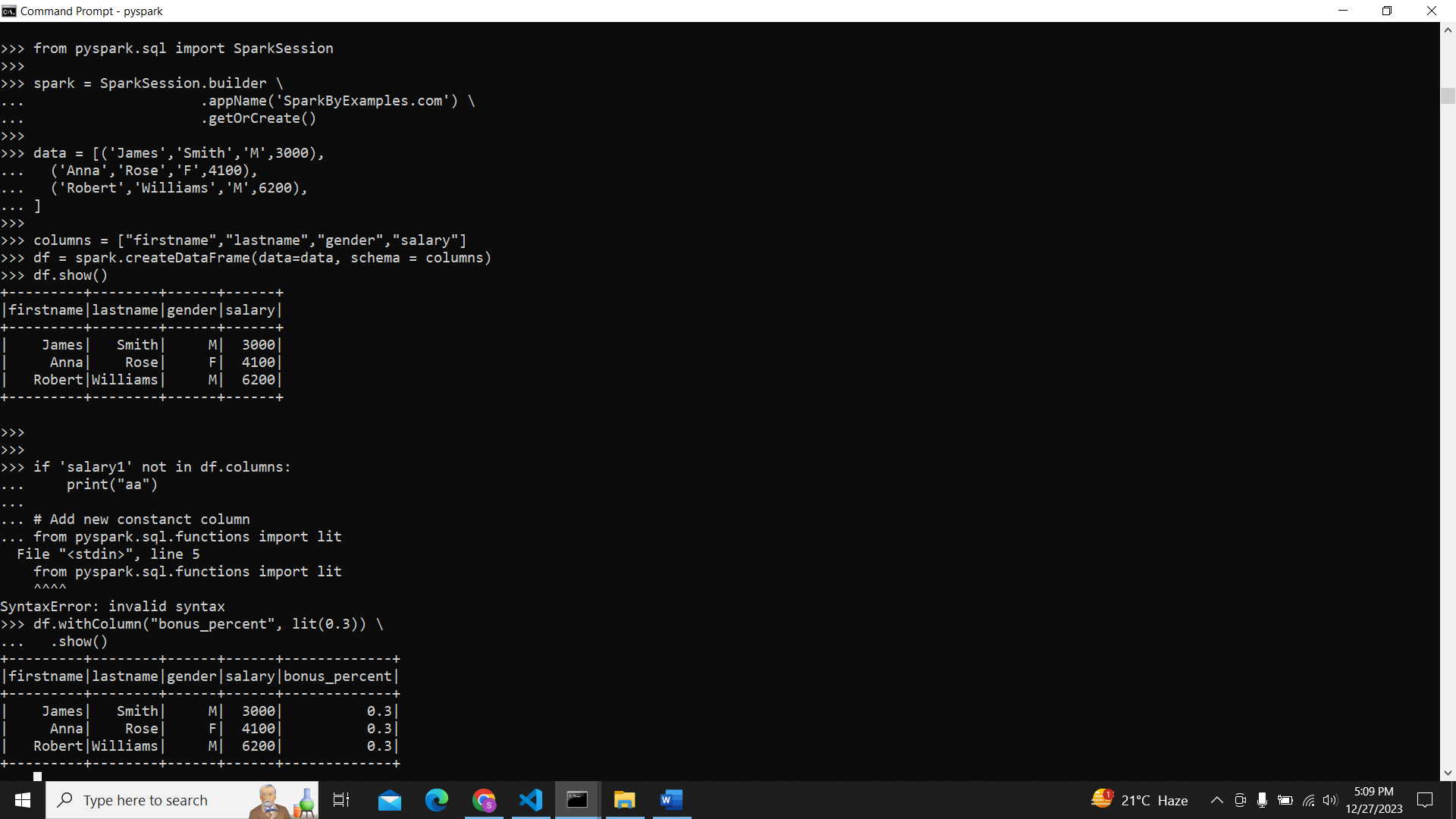
* SQL or [NoSQL](https://www.ibm.com/topics/nosql-databases) servers
* CRM and ERP systems
* Flat files
* Email
* Web pages

2.Transform: In the staging area, the raw data undergoes data processing. Here, the data is transformed and consolidated for its intended analytical use case. This phase can involve the following tasks:

* Filtering, cleansing, de-duplicating, validating, and authenticating the data.
* Performing calculations, translations, or summarizations based on the raw data. This can include changing row and column headers for consistency, converting currencies or other units of measurement, editing text strings, and more.

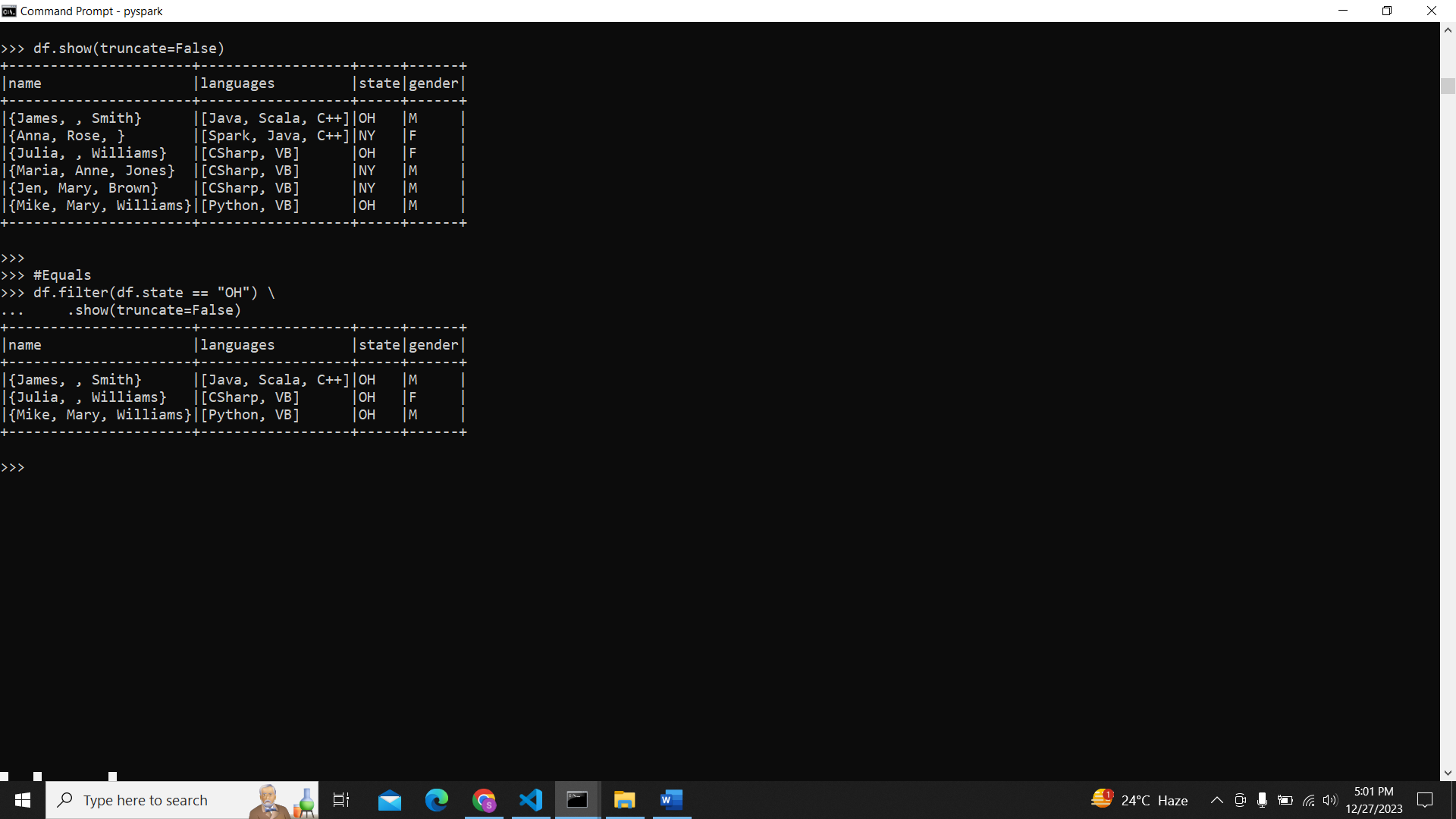
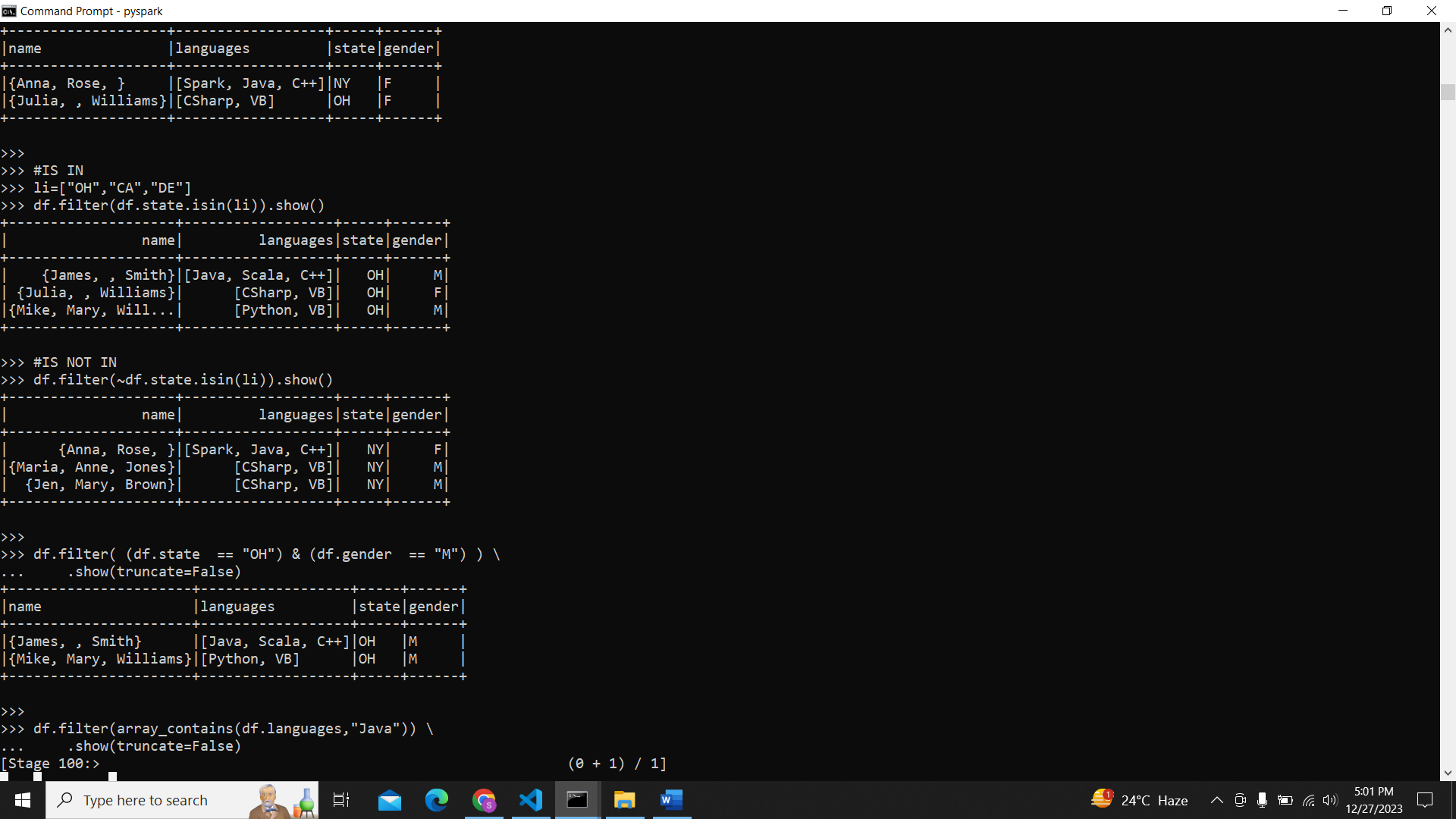
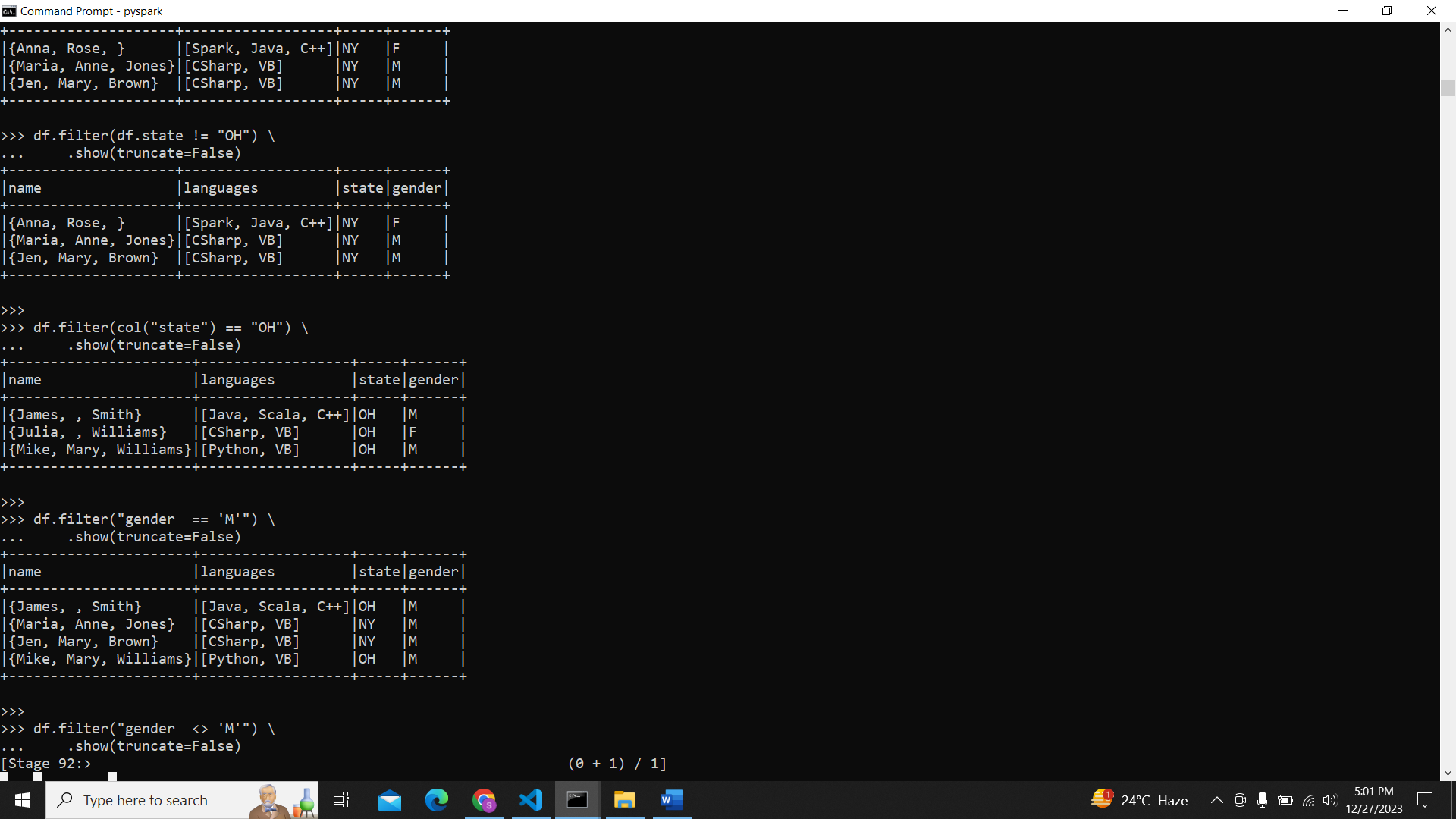
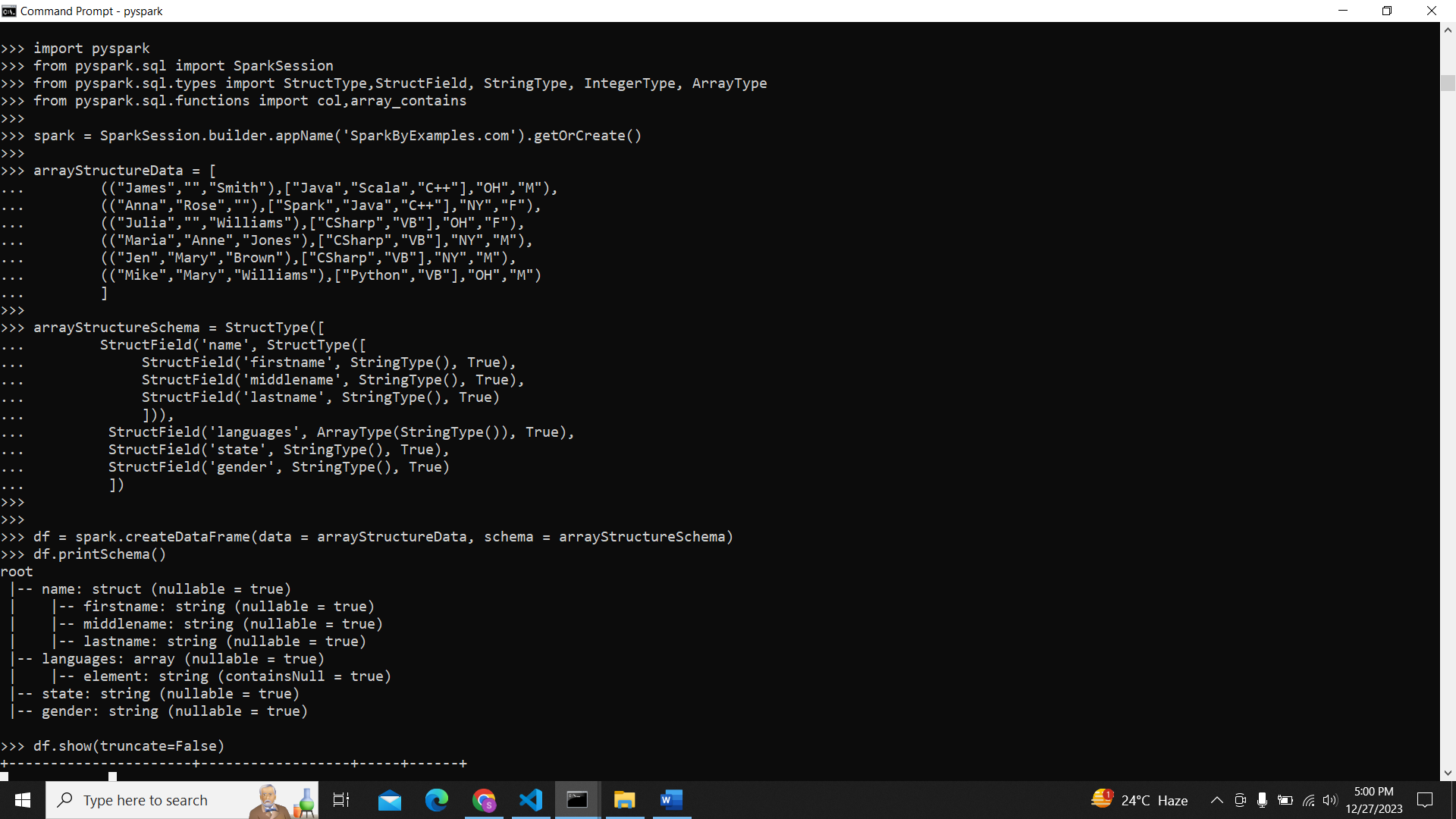
3.Load: In this last step, the transformed data is moved from the staging area into a target data warehouse. Typically, this involves an initial loading of all data, followed by periodic loading of incremental data changes and, less often, full refreshes to erase and replace data in the warehouse. For most organizations that use ETL, the process is automated, well-defined, continuous and batch-driven.

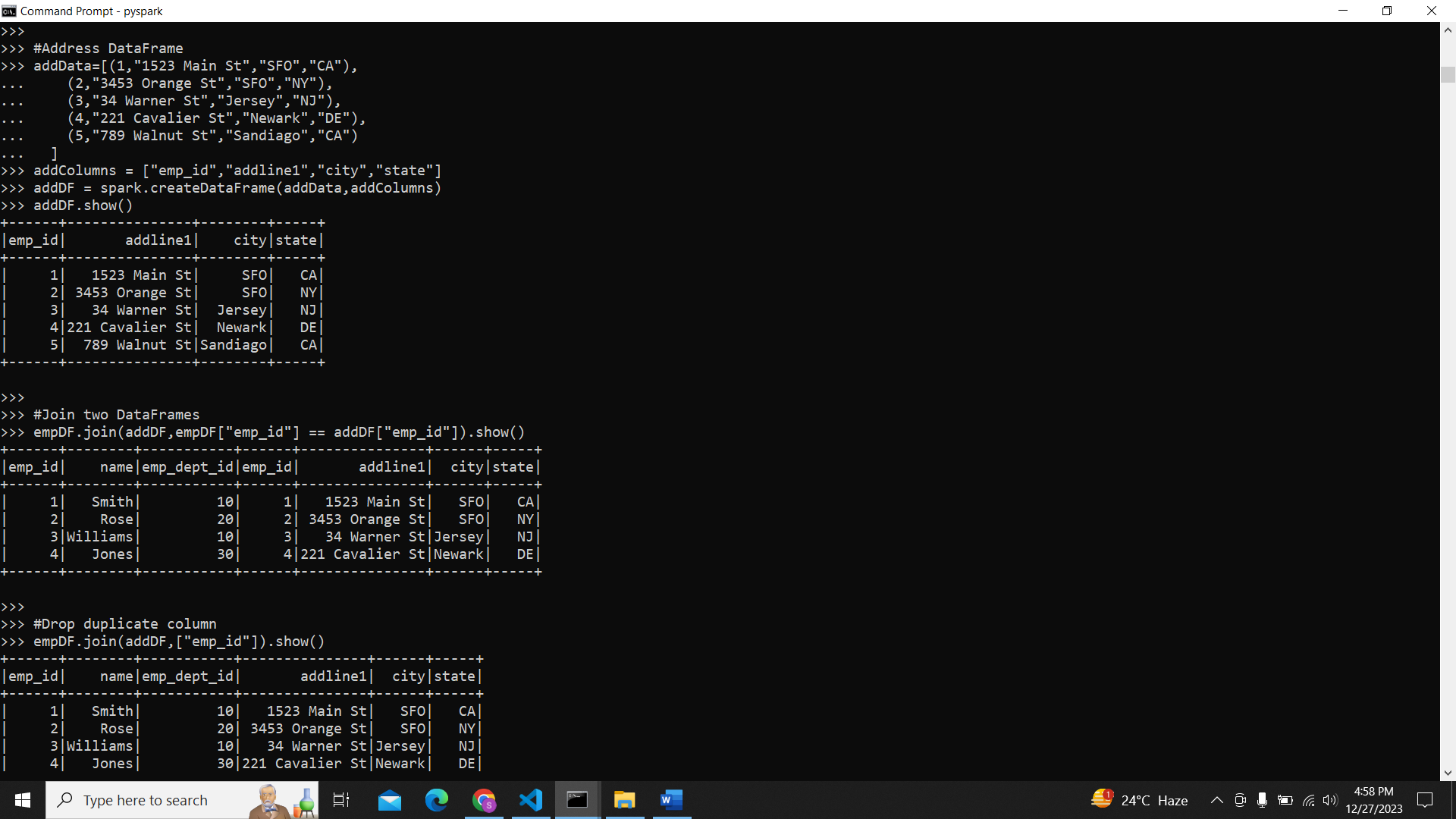
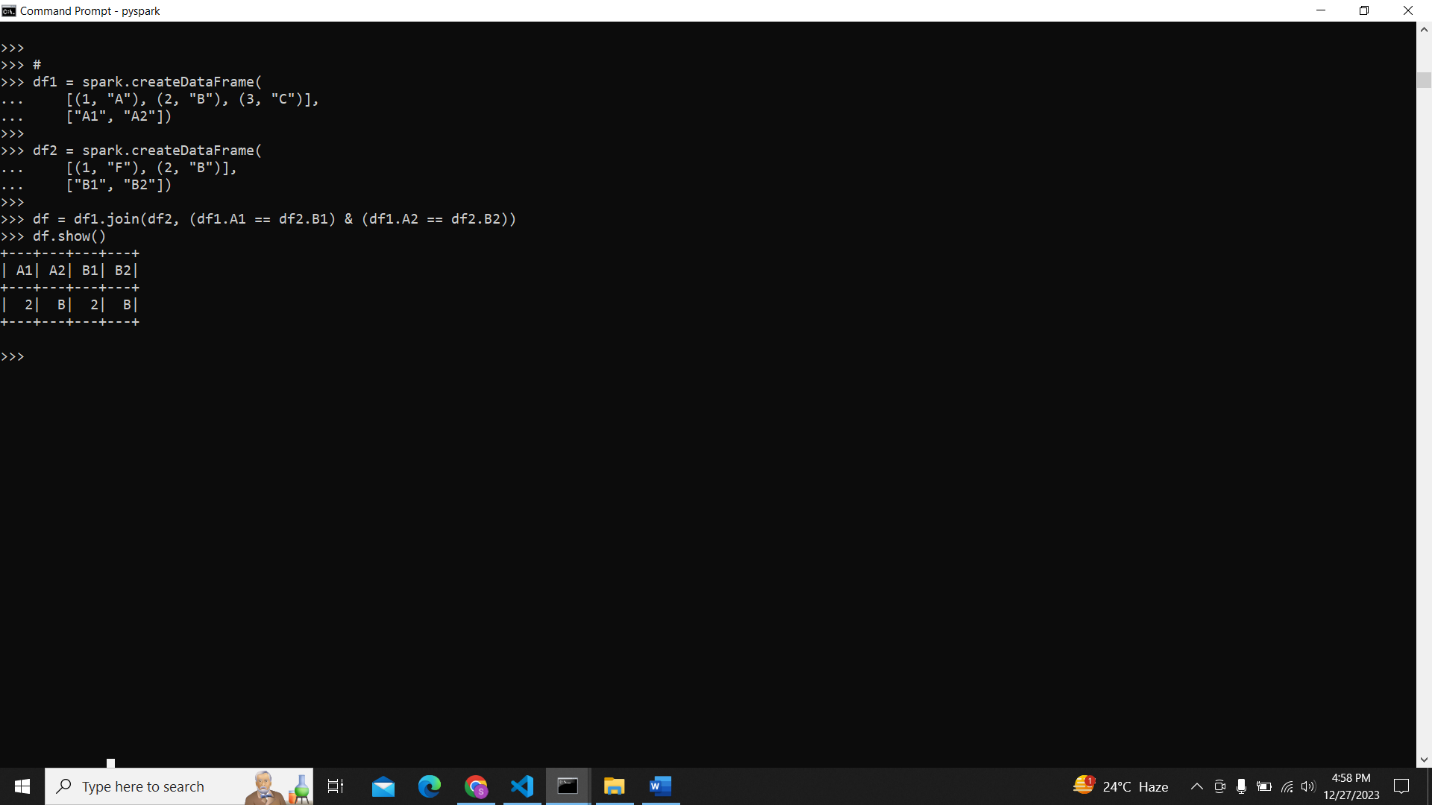
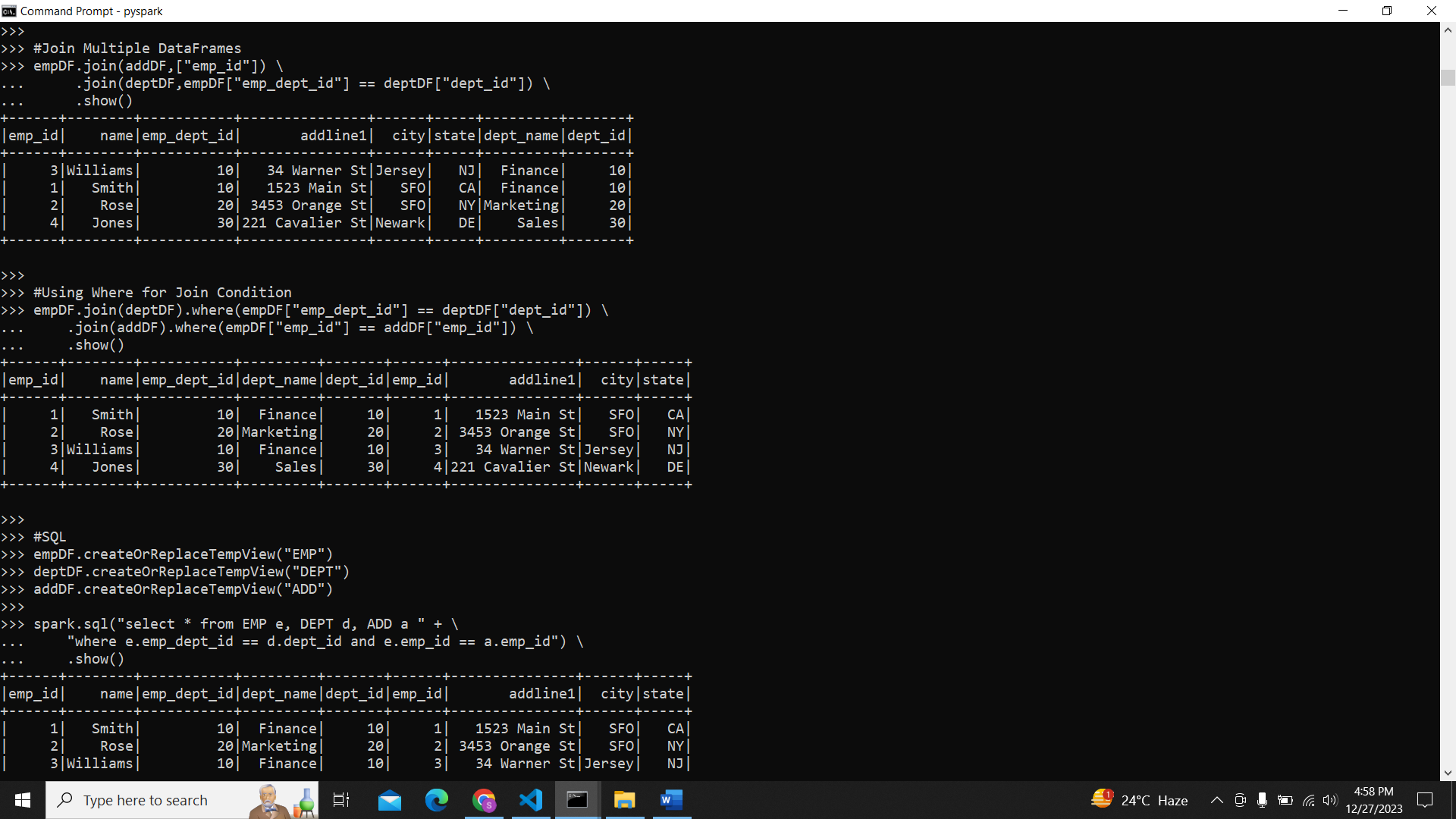
**ETL with Pyspark**

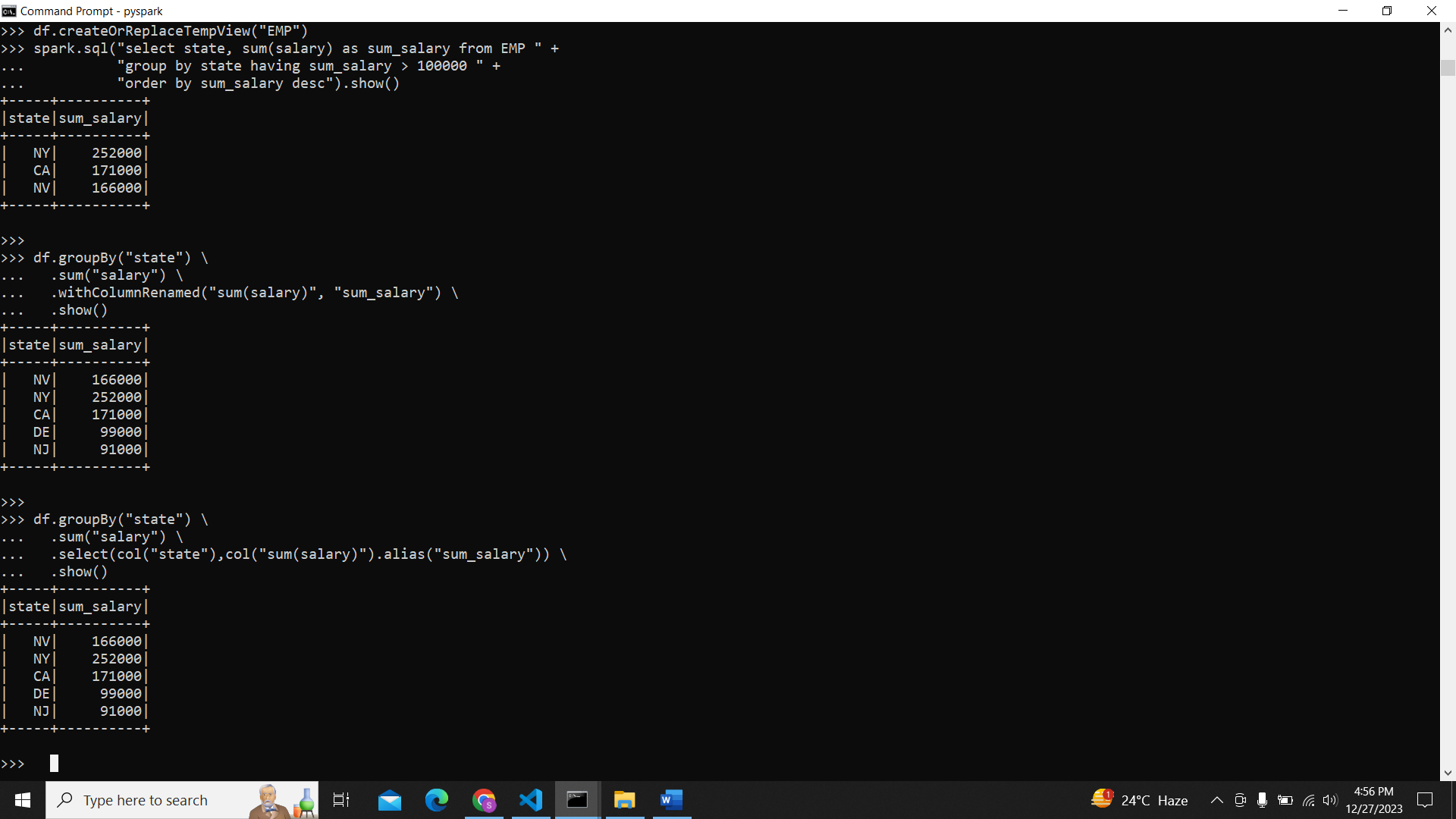
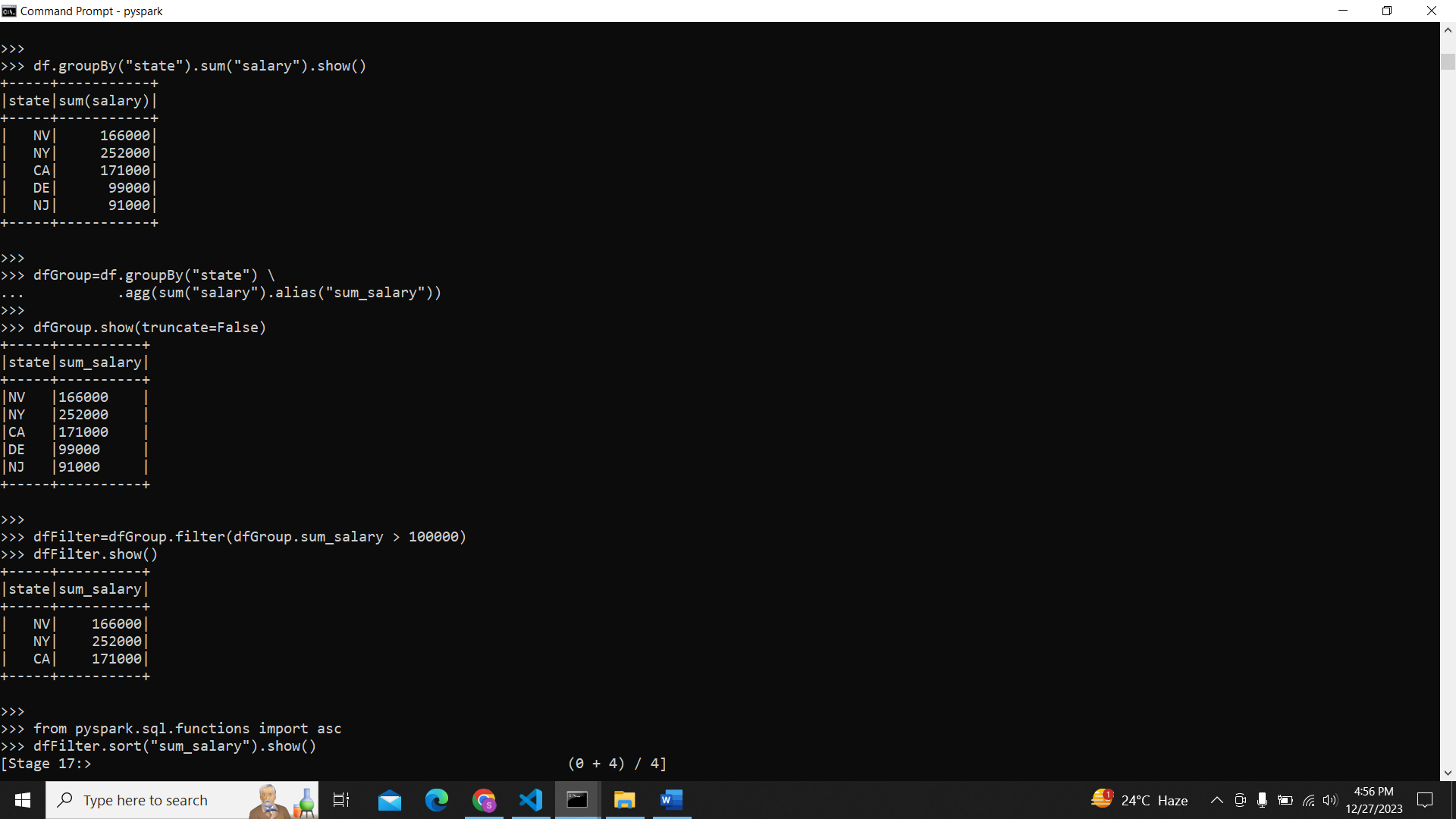
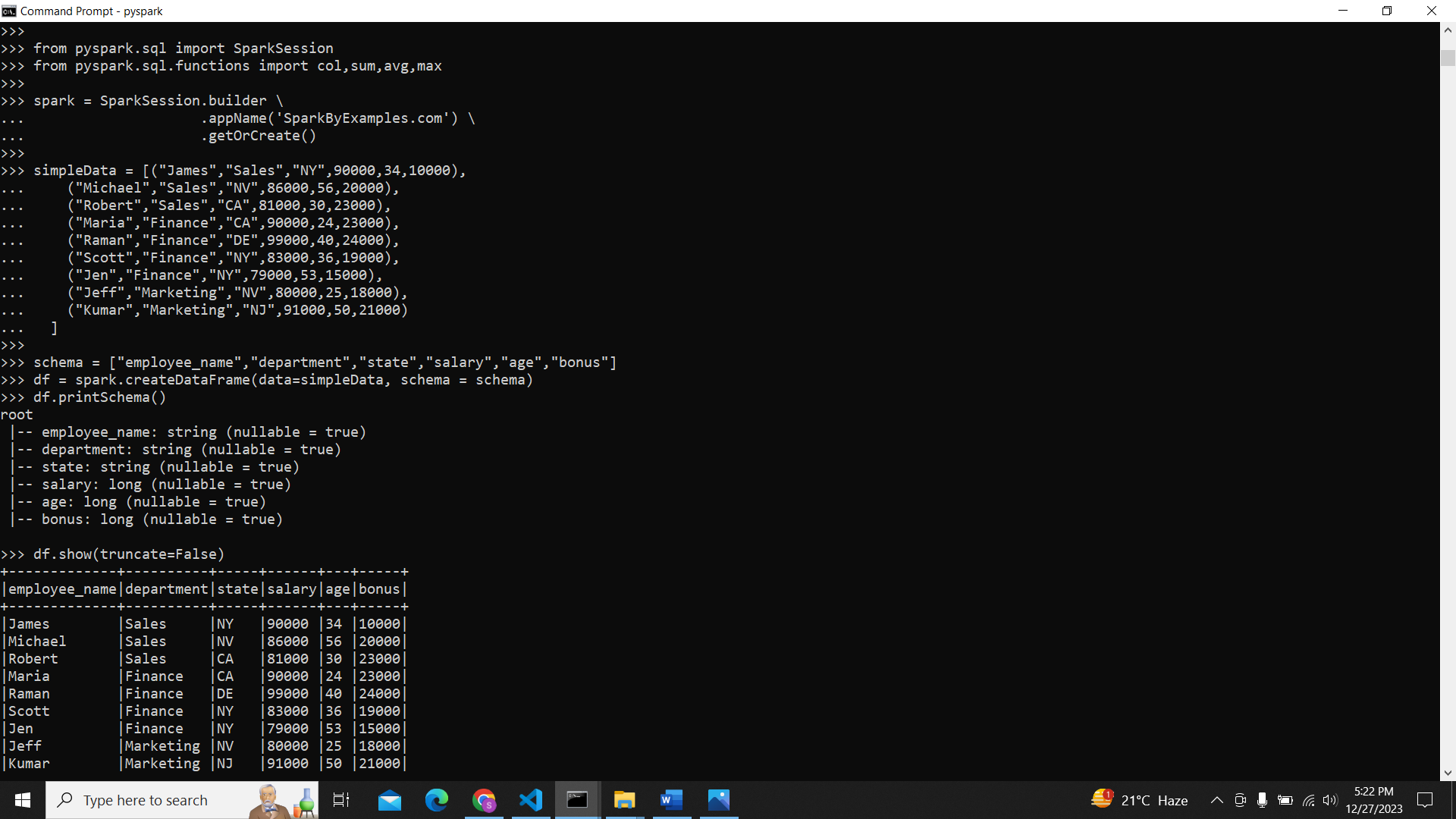
3. Using Spark SQL - Creating databases, tables

Using Spark SQL - Transformations such as Filter, Join, Simple Aggregations, GroupBy.

**Filter**

**Join**

 **GroupBy**

**Aggregations**

