**PYSPARK**

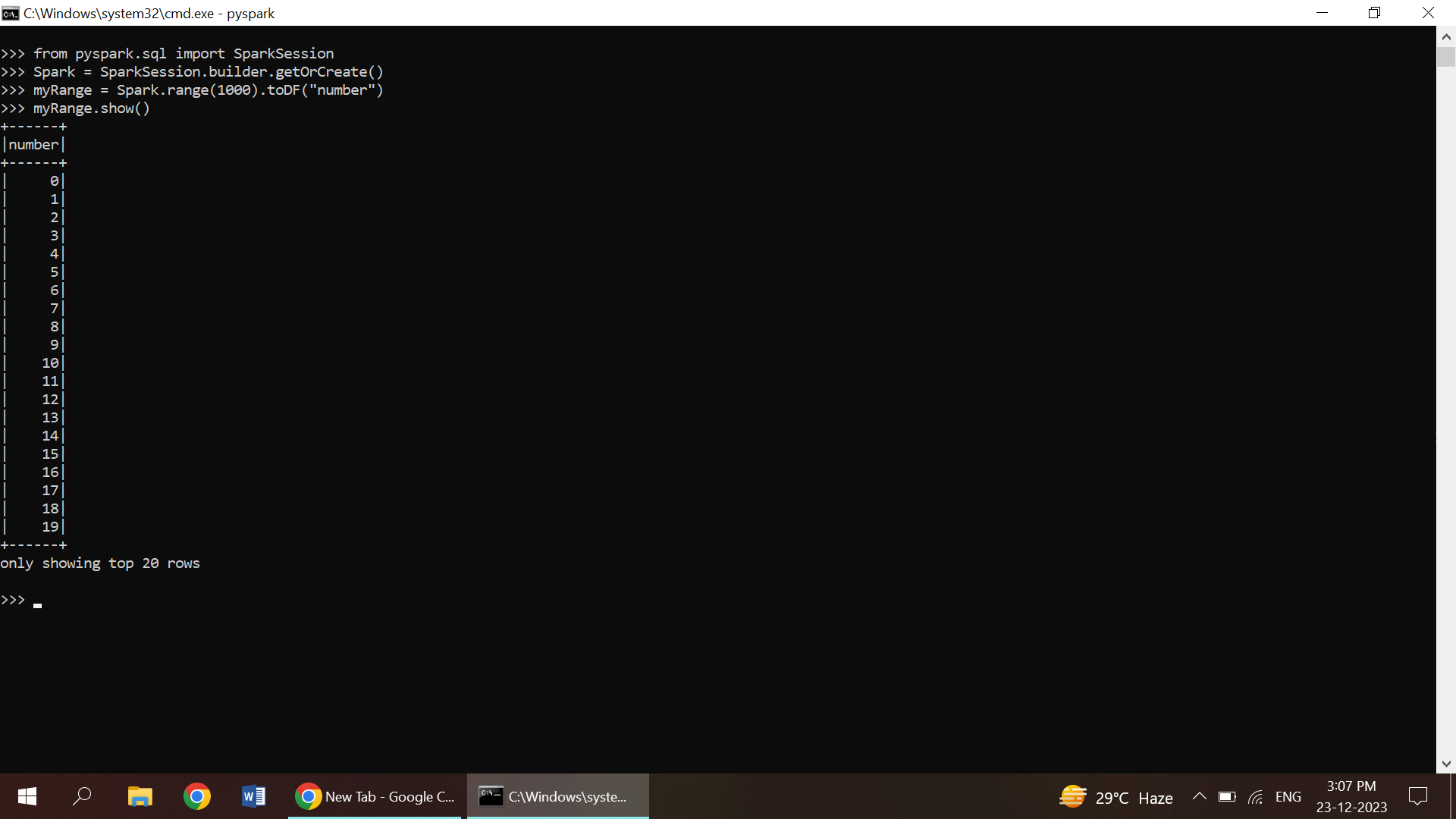
**COMMANDS**

S.R.TAANUSRI

22.12.2023

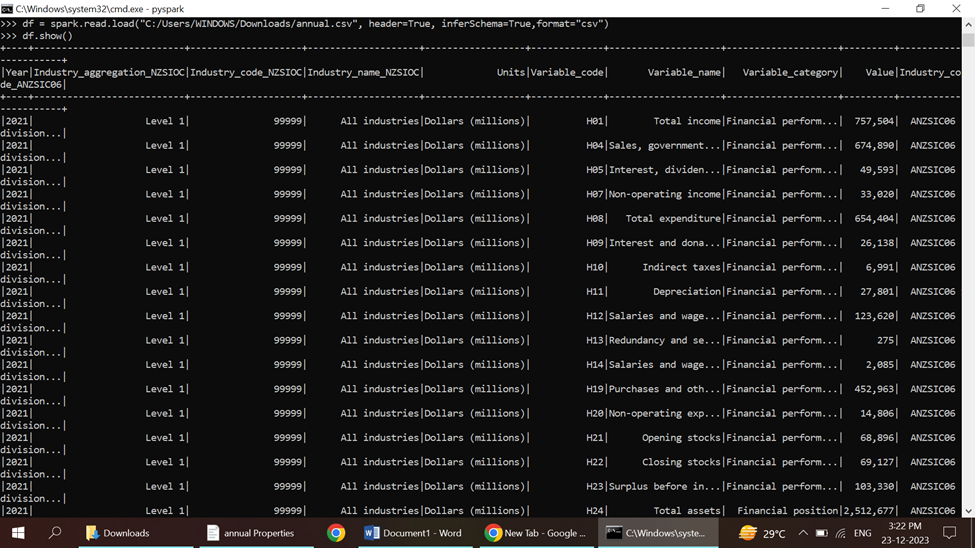
1.SHOW()

* This method is commonly used to display the contents of a Data Frame.
* This method prints the first few rows of the Data Frame to the console, providing a quick way to inspect the data.



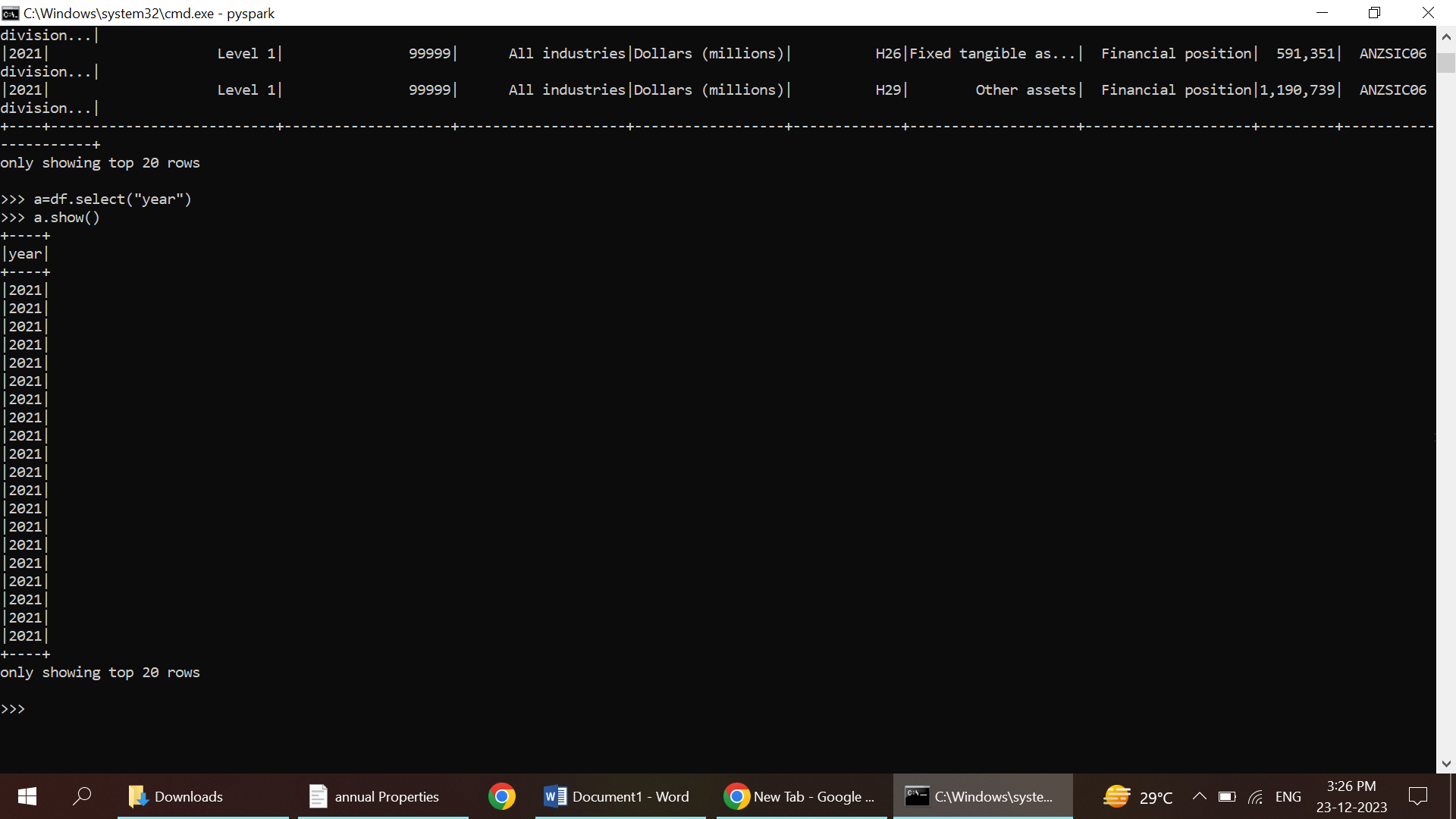
2.IMPORT DATA FROM CSV FILE

* The spark.read.csv method to read the CSV file into a Data Frame.
* Specify the path to the CSV file using the path parameter.
* Header: to indicate whether the first row contains column names
* InferSchema: to let Spark infer the data types of columns.



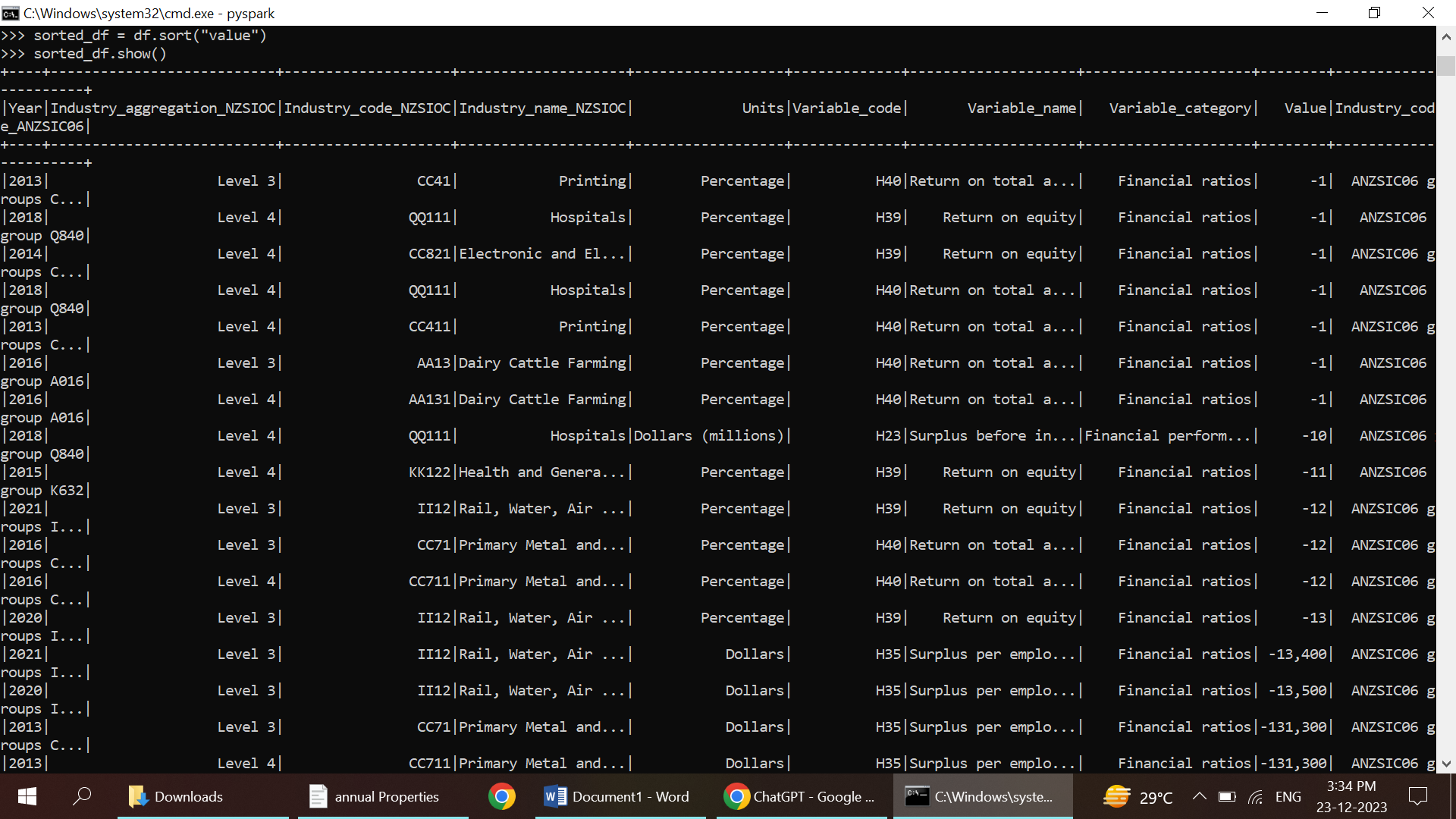
3.SELECT

* The select operation is used to choose specific columns from a Data Frame.
* It allows you to create a new Data Frame that includes only the columns you are interested in.



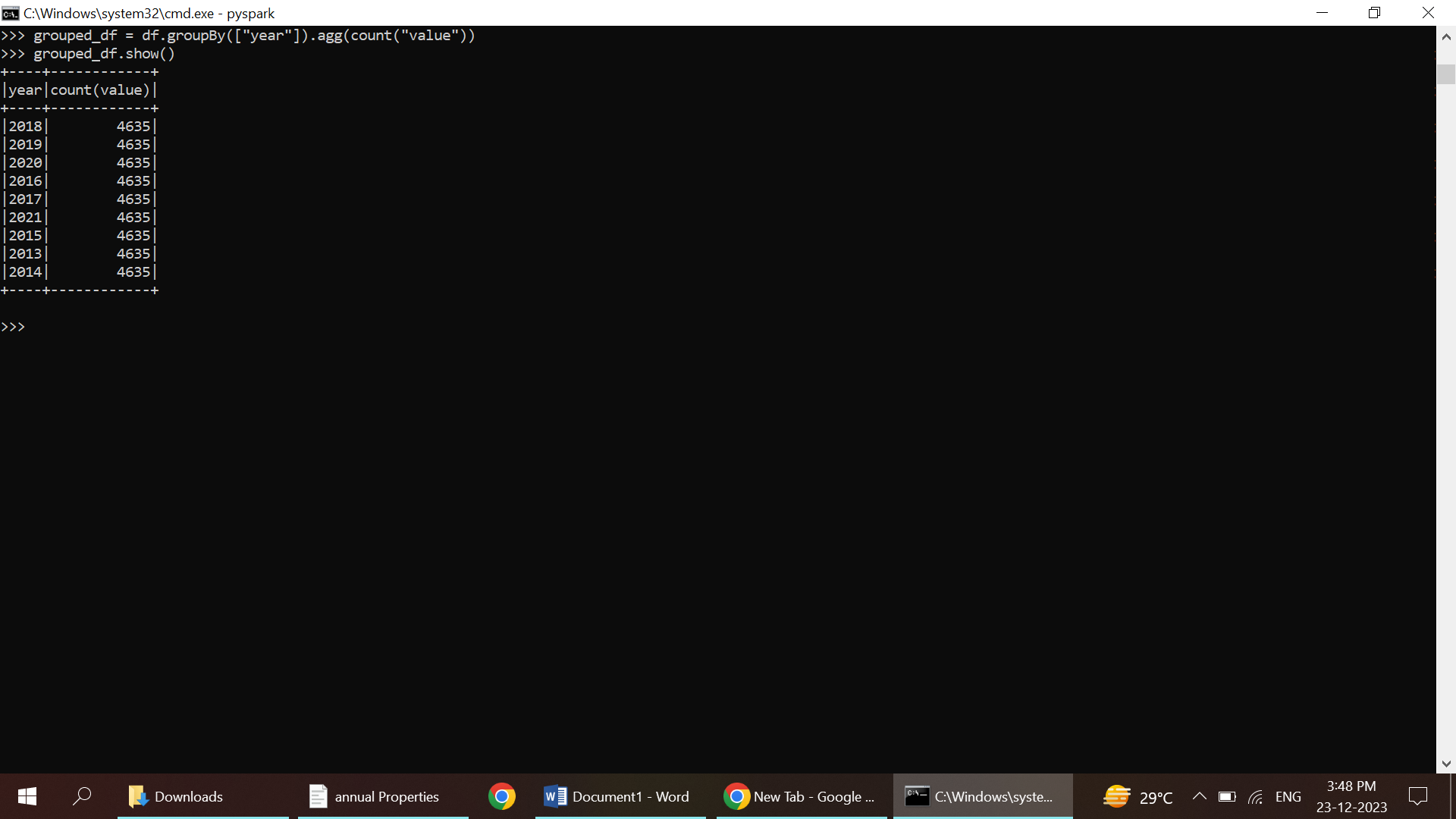
4.SORT

* The sort operation is used to sort the rows of a Data Frame based on one or more columns.
* It is equivalent to the SQL ORDER BY clause.
* The sort operation returns a new Data Frame with the rows sorted in ascending or descending order based on the specified columns.

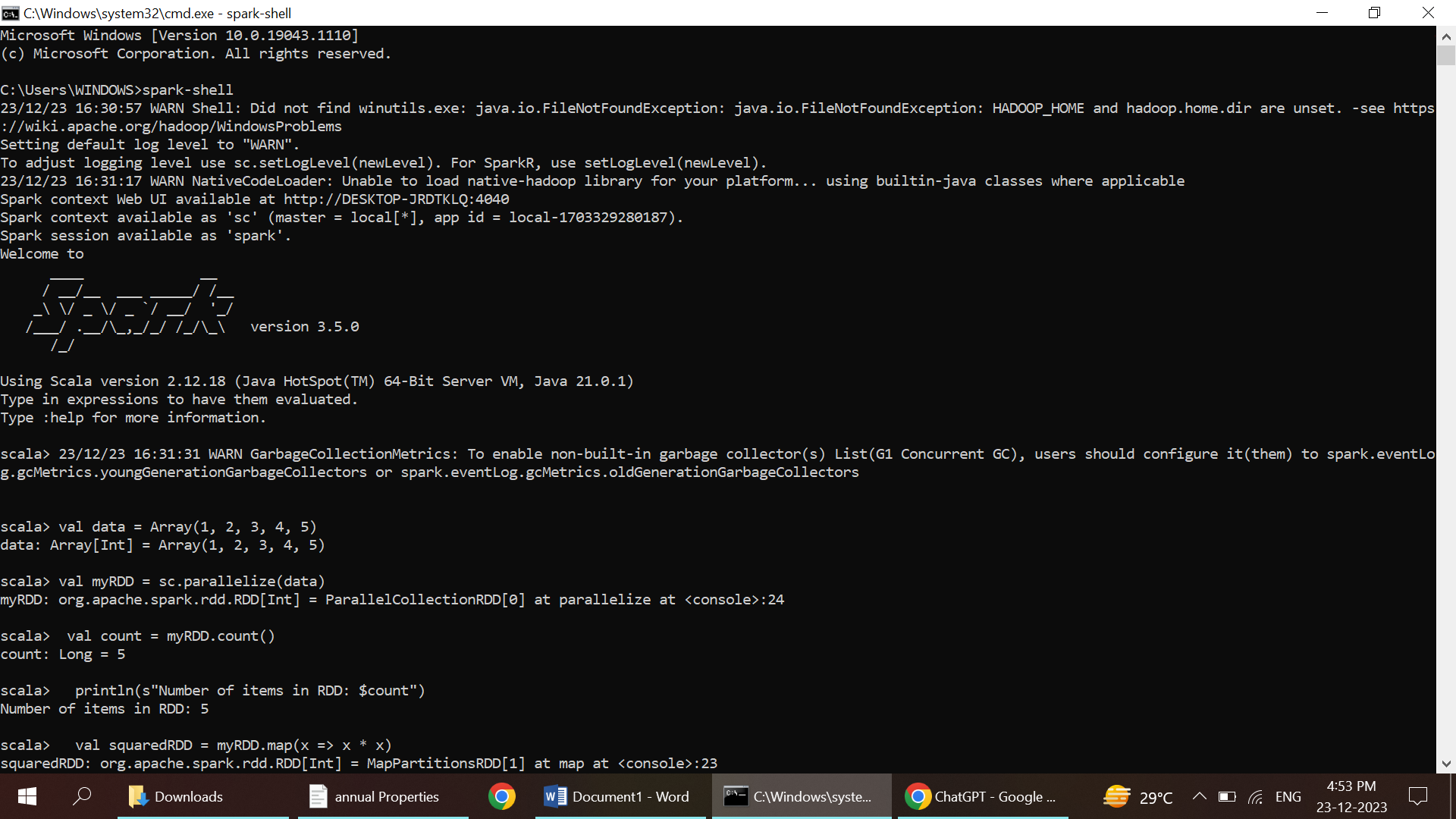


5.GROUP BY

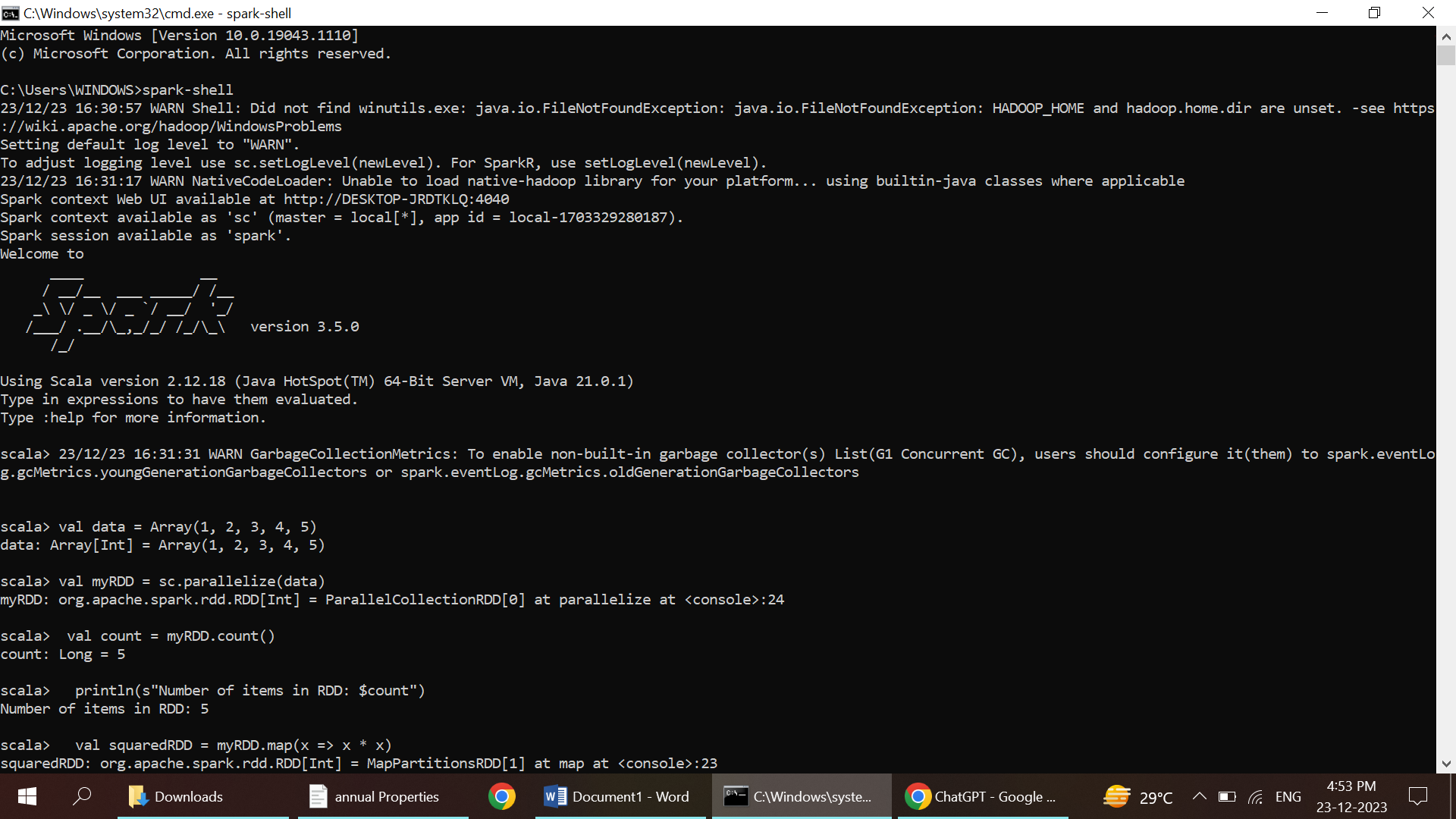
* The groupBy operation is used to group the rows of a Data Frame based on one or more columns.
* It is similar to the SQL GROUP BY clause.
* The groupBy operation returns a Grouped Data object, which you can then use to perform aggregate functions or apply other transformations on each group.



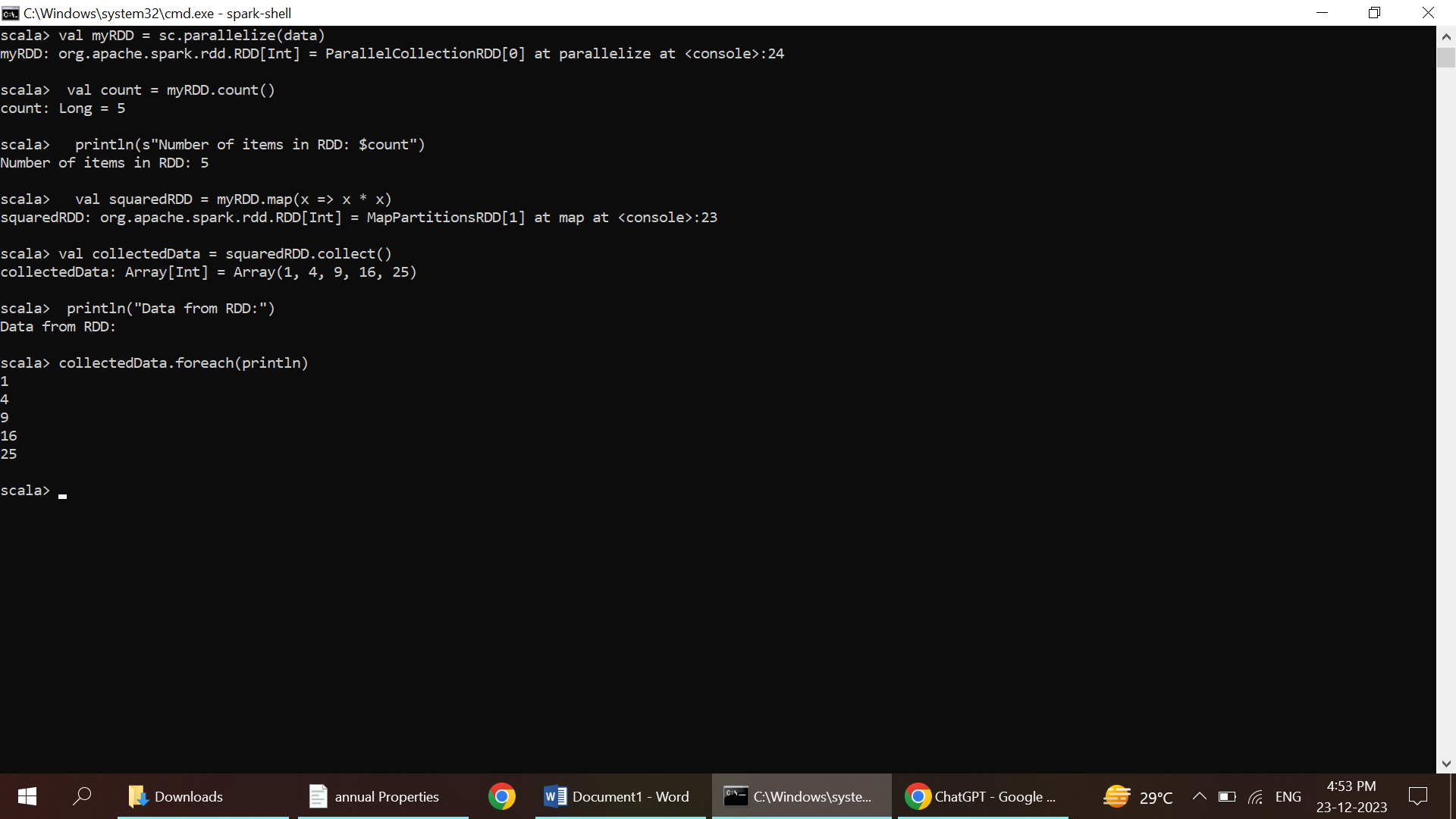
1.OPEN PYSPARK SHELL



2.CREATE RDD THROUGH PARELLELIZING



3.COUNT DATA



4.MAP DATA

