Downloaded the dataset from <https://www.kaggle.com/datasets/kimjihoo/coronavirusdataset> and created producer to connect to confluent cloud kafka and stream data to mongoDB

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

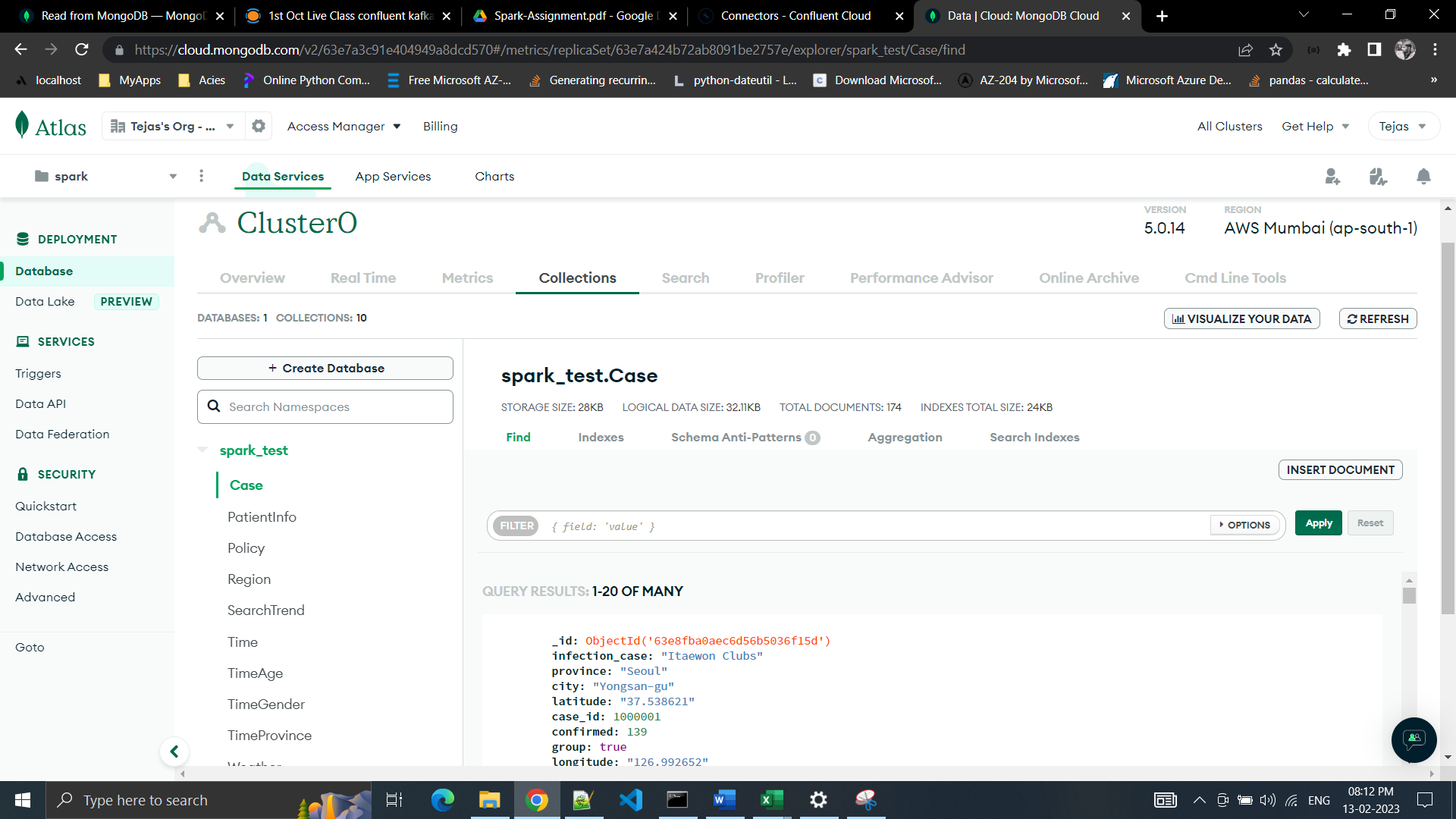
Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a computer

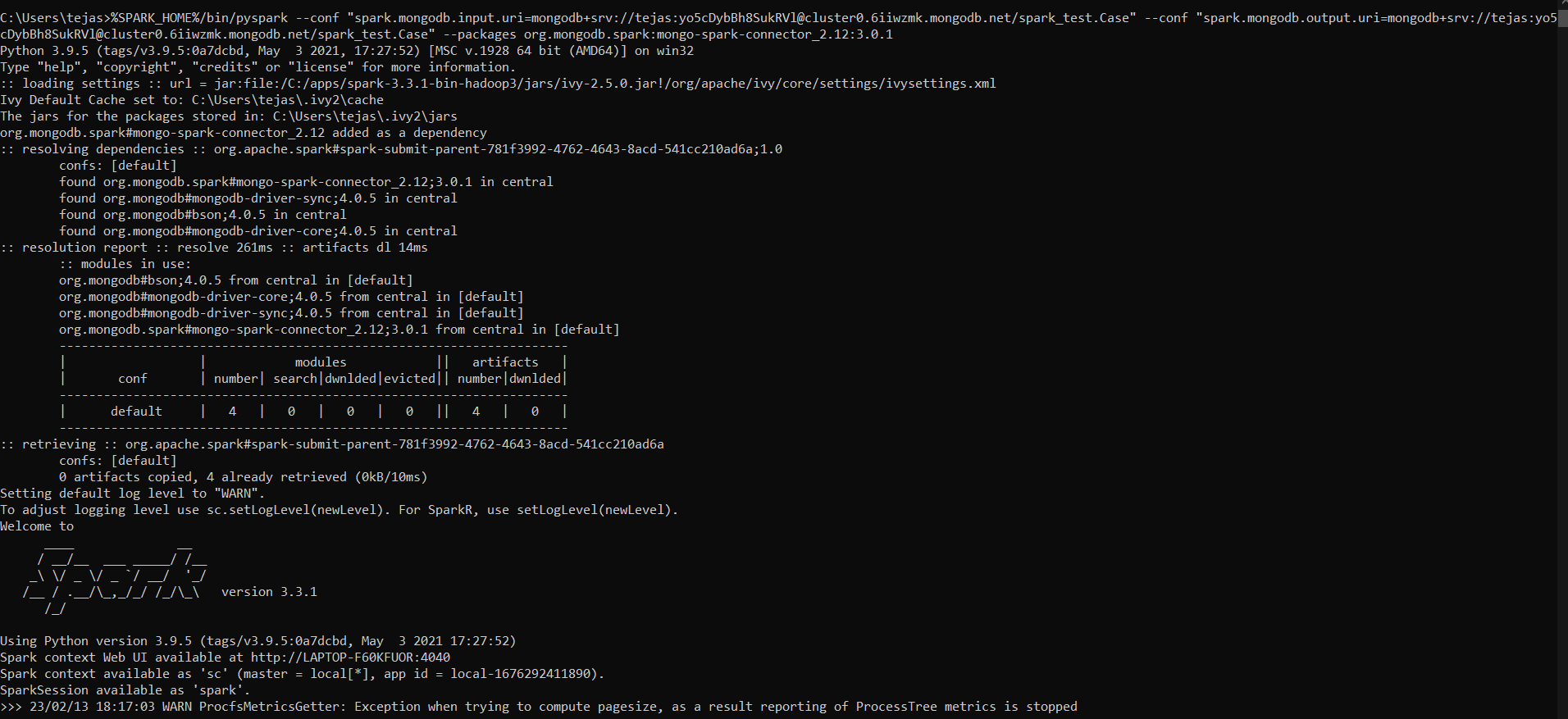
Description automatically generatedA screenshot of a computer

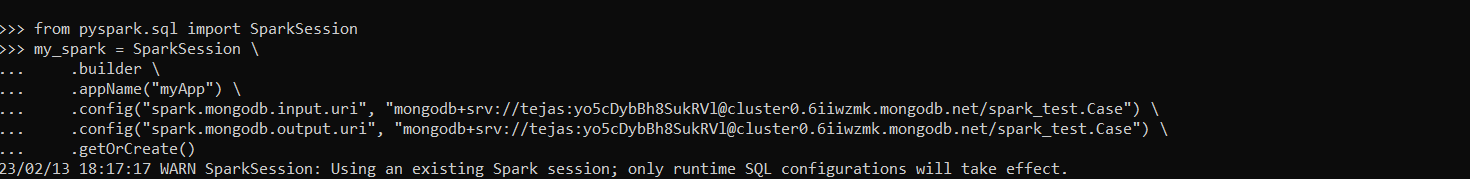
Description automatically generated



***Collect your data as a pyspark dataframe and perform different operations***.

%SPARK\_HOME%/bin/pyspark --conf "spark.mongodb.input.uri=mongodb+srv://tejas:yo5cDybBh8SukRVl@cluster0.6iiwzmk.mongodb.net/spark\_test.Case" --conf "spark.mongodb.output.uri=mongodb+srv://tejas:yo5cDybBh8SukRVl@cluster0.6iiwzmk.mongodb.net/spark\_test.Case" --packages org.mongodb.spark:mongo-spark-connector\_2.12:3.0.1

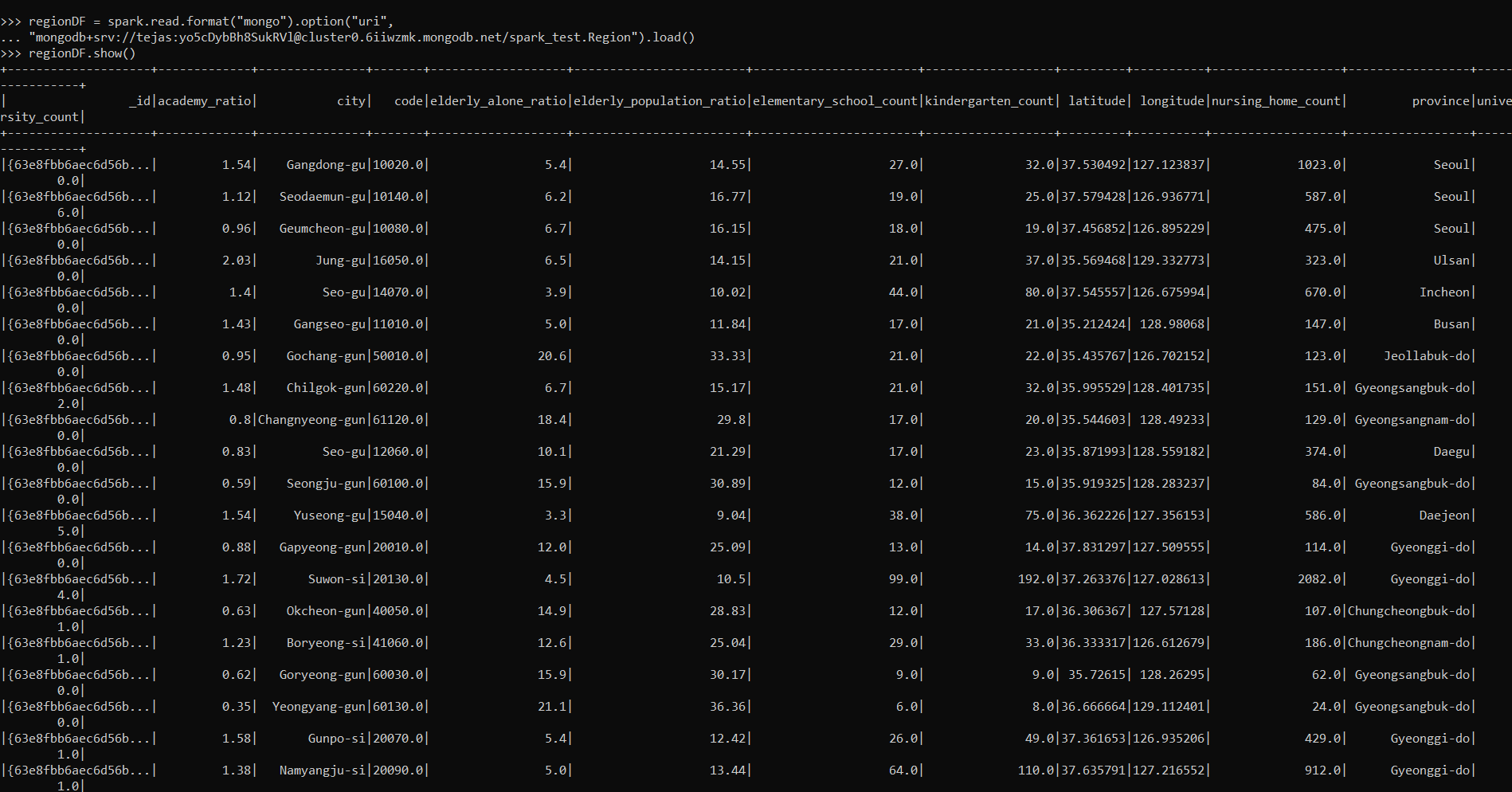




Read the data, show it and Count the number of records.

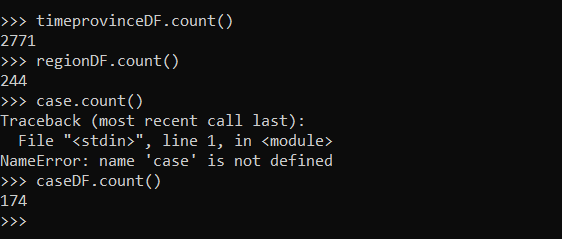
Graphical user interface

Description automatically generated with low confidence



Text

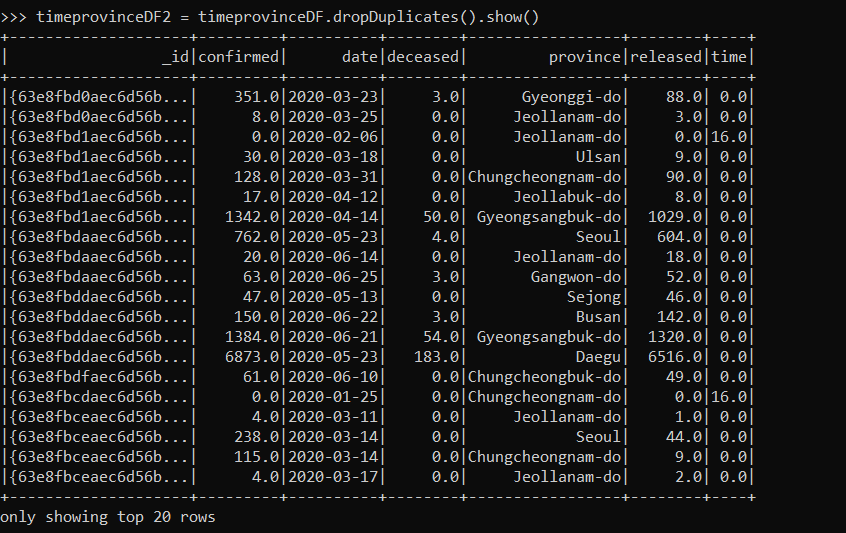
Description automatically generated with medium confidence



Describe the data with a describe function. If there is any duplicate value drop it. Use limit function for showcasing a limited number of records.

A picture containing graphical user interface

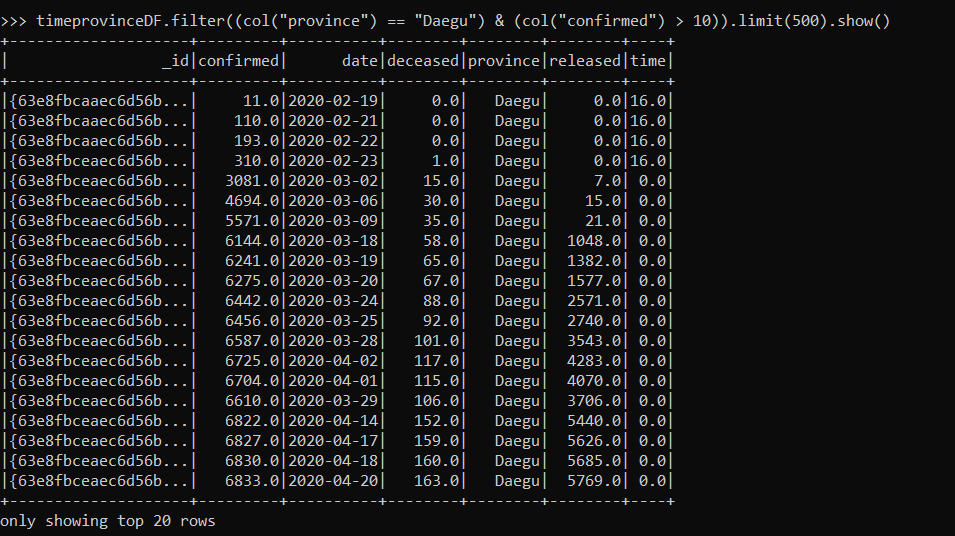
Description automatically generated



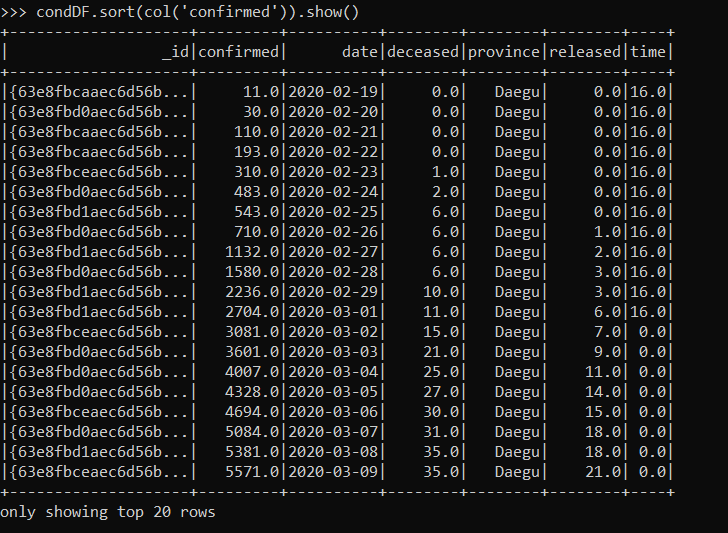
Text

Description automatically generated

Filter the data based on different columns or variables and do the best analysis.



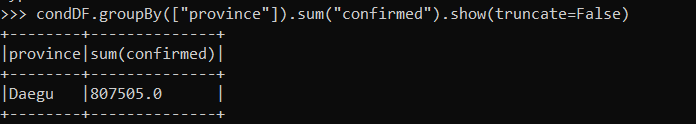
Sort the number of confirmed cases. Confirmed column is there in the dataset. Check with descending sort also.



Table

Description automatically generated with medium confidence

Use group by on top of province and city column and agg it with sum of confirmed cases.

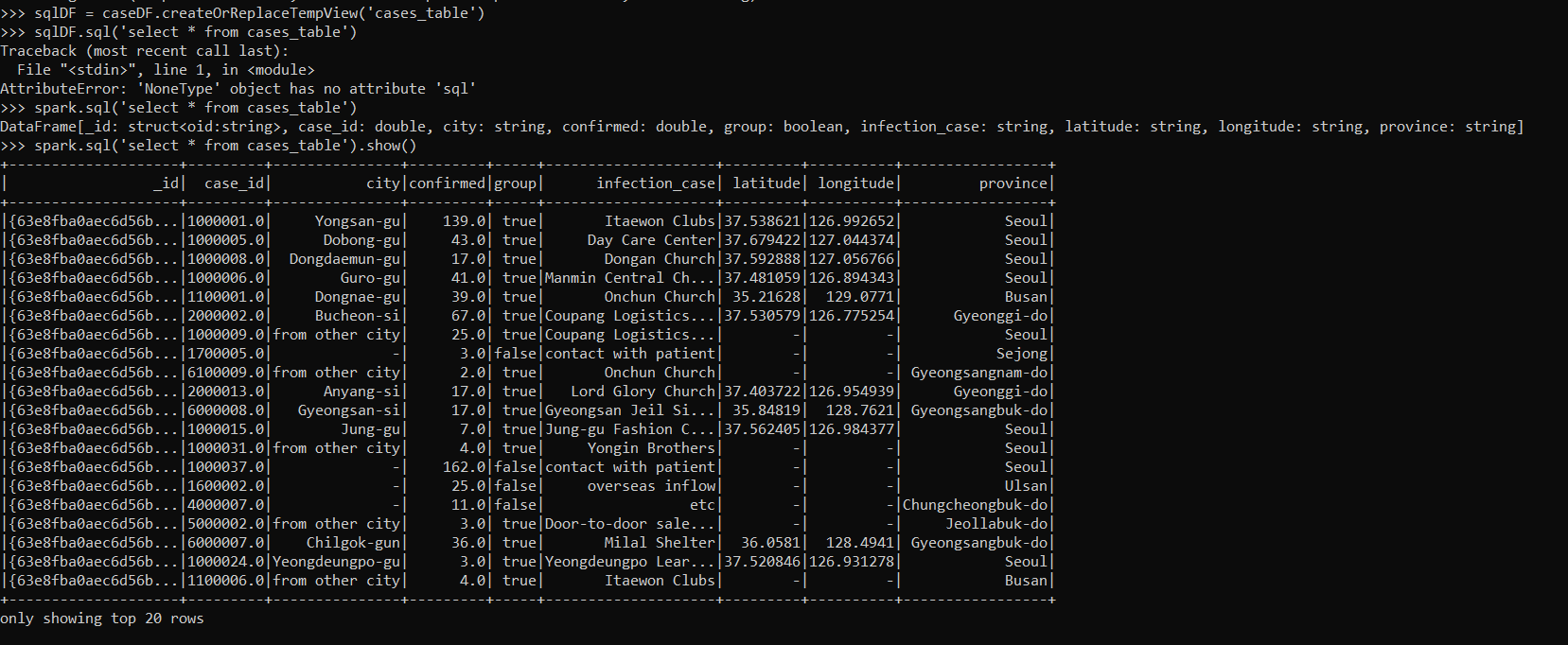


Using Joins

Graphical user interface, text

Description automatically generated with medium confidence

USING SQL to query Data



Creating UDF

A picture containing text

Description automatically generated