**Name:** V Venkata Sri Prasad

**Batch:** Data Engineering

**Spark Coding Assignment**

**Question:**

Execute Manipulating, Droping, Sorting, Aggregations, Joining, GroupeBy DataFrames

**Procedure:**

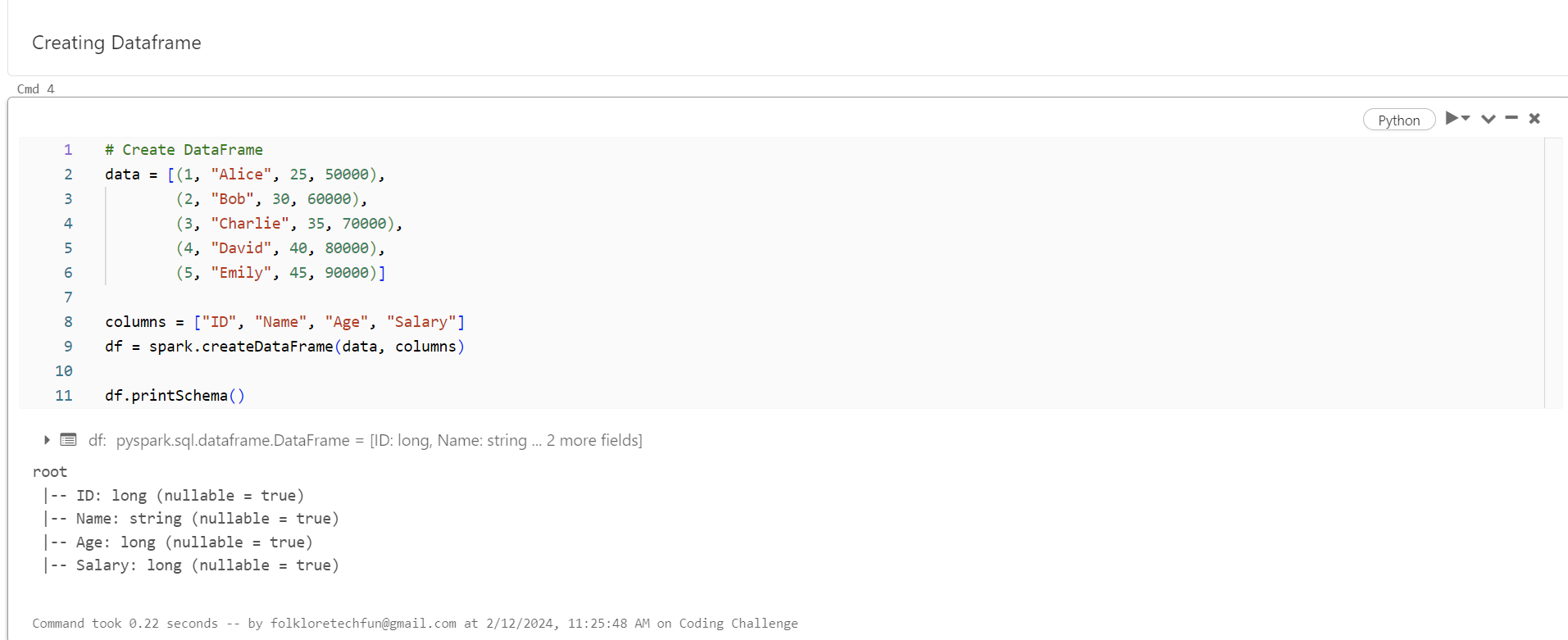
1. **Creating a Spark Session:**

A SparkSession is the entry point to Spark that allows interaction with Spark functionality.



1. **Creating a Dataframe:**

A DataFrame in Spark is a distributed collection of data organized into named columns. It provides a higher-level API for working with structured and semi-structured data, enabling easy manipulation, querying, and analysis.



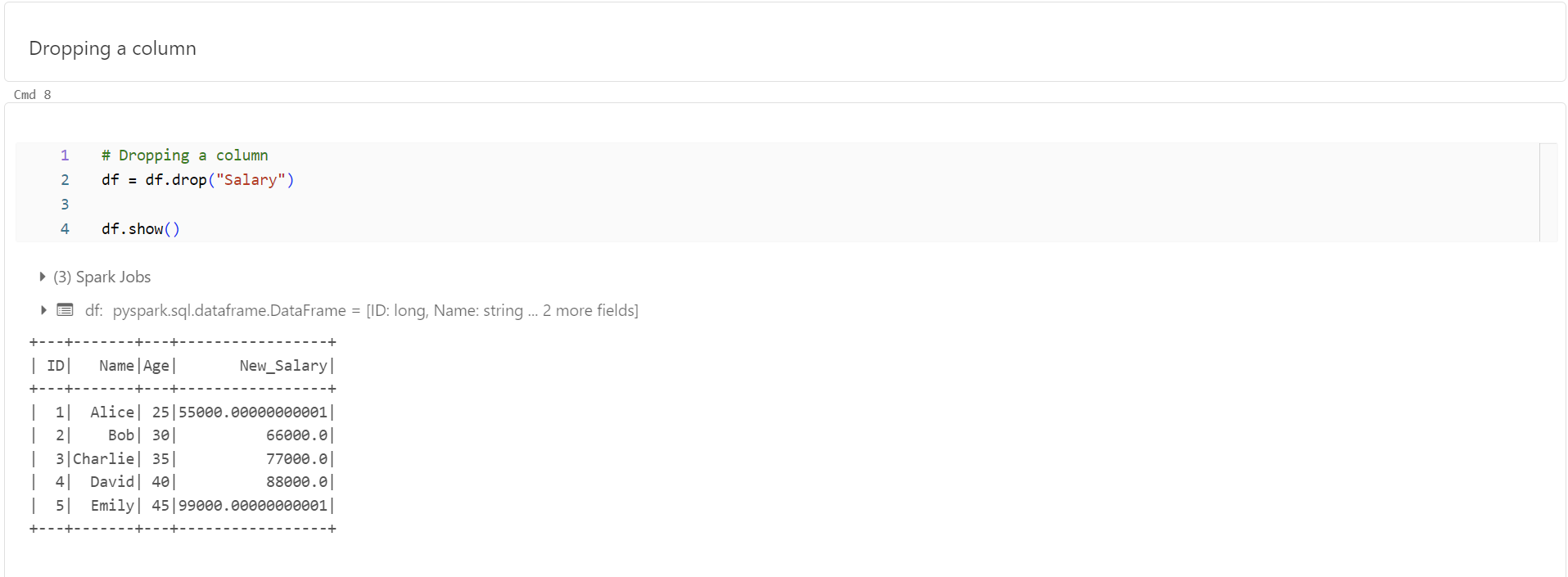
1. **Manipulating data:**

Adds a new column to the DataFrame, which can involve any transformation on existing columns.



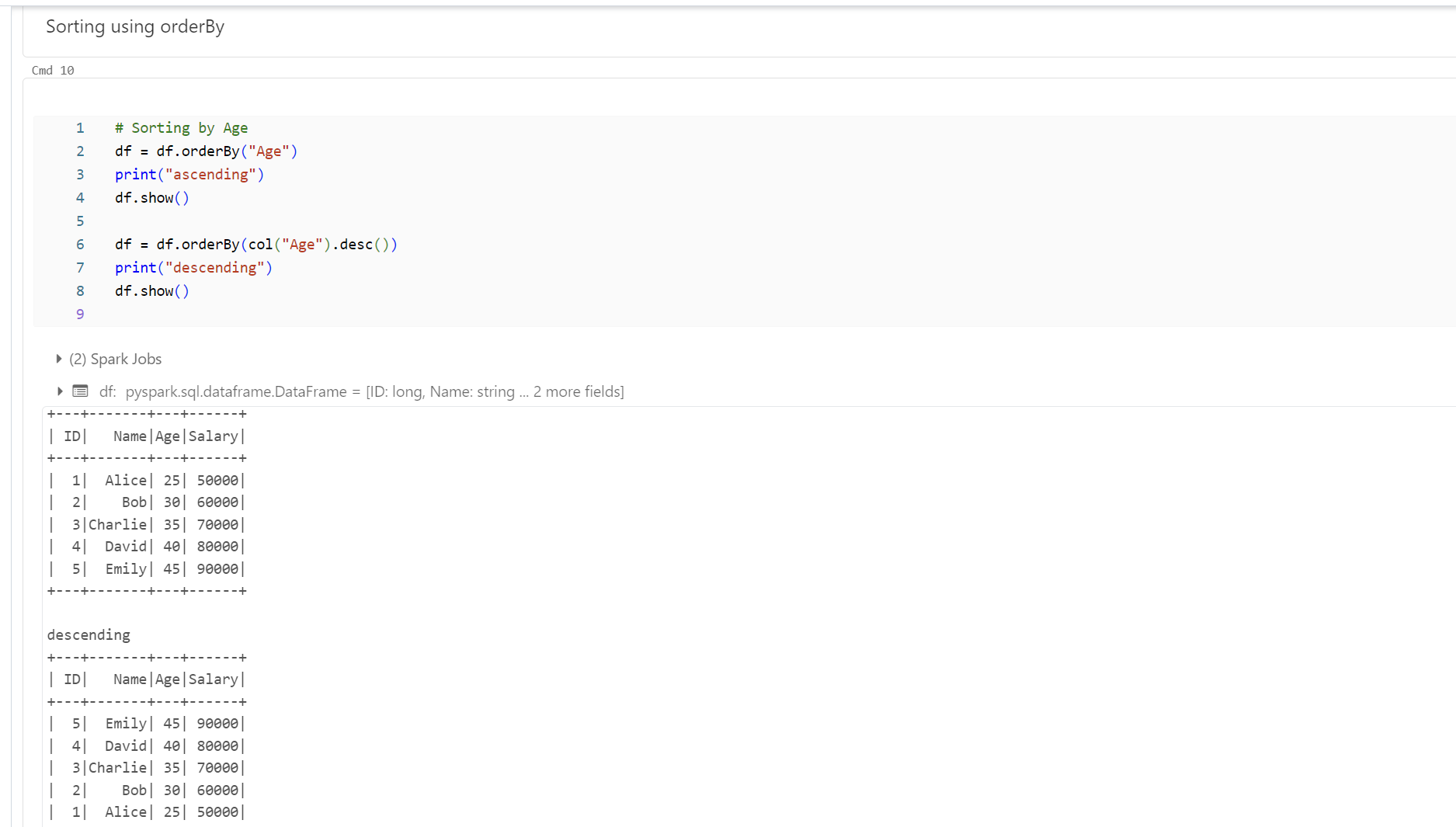
1. **Dropping a column:**

Removes a column from the DataFrame, typically if it's not required for further analysis.



1. **Sorting:**

Arranges the DataFrame's rows based on the values of one or more columns, facilitating better data examination.



1. **Aggregations:**

Computes summary statistics (e.g., average, sum, min, max) across the entire DataFrame or within groups.



1. **GroupBy:**

Splits the DataFrame into groups based on specified criteria (e.g., age) and applies aggregate functions to each group.

