Import spark session:

A screenshot of a computer

Description automatically generated

Creating rdds:

A screenshot of a computer

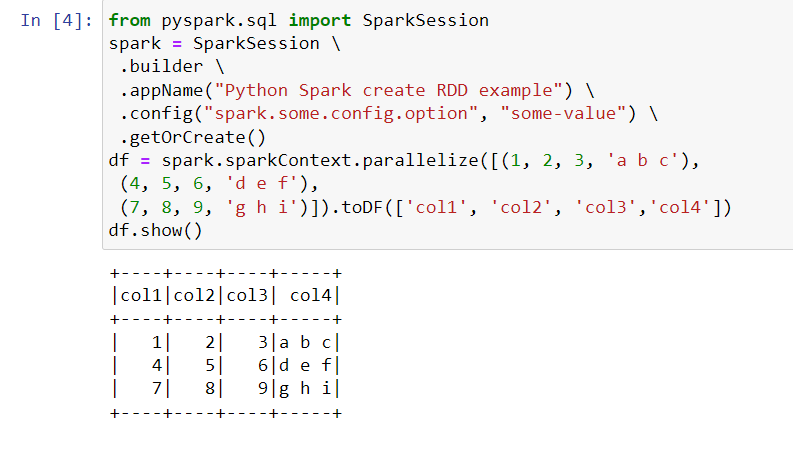
Description automatically generated

Using parallelize:

A screenshot of a computer

Description automatically generated

Create dataframes:

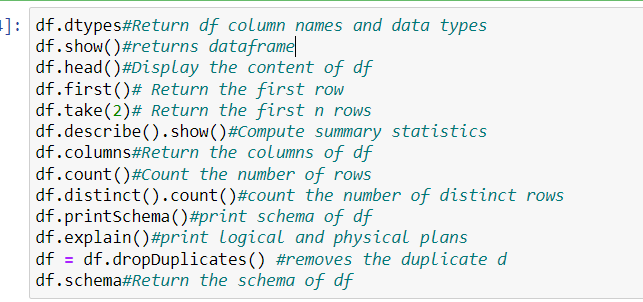


A screen shot of a computer

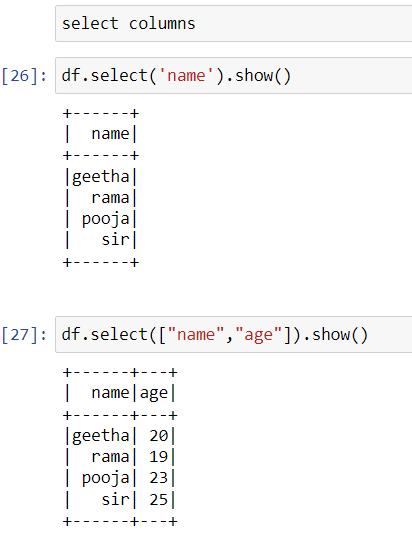
Description automatically generated

A screenshot of a computer

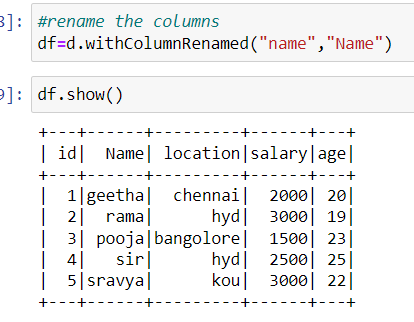
Description automatically generated



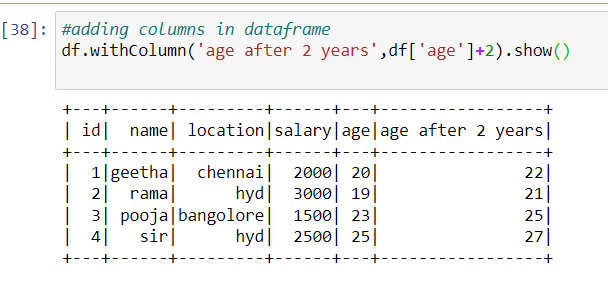
Selecting columns:



Renaming columns:



Adding columns:

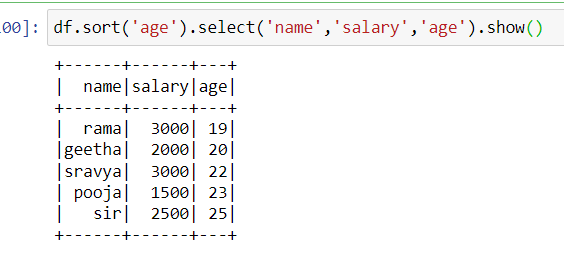


Drop columns:

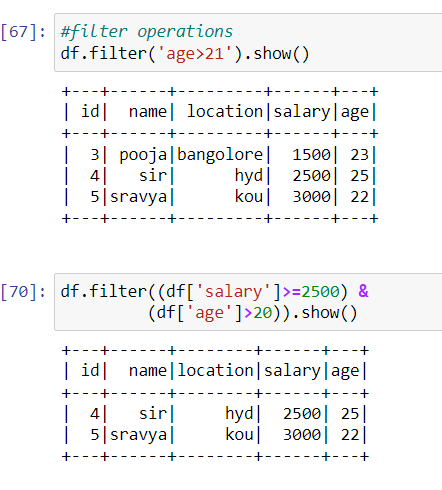
A screenshot of a computer code

Description automatically generated

Sort:



Filter:



Dropping null values (NA function):

It has 3 parameters:

how=any/all,thresh,subset



A screenshot of a computer

Description automatically generated

Filling missing values:

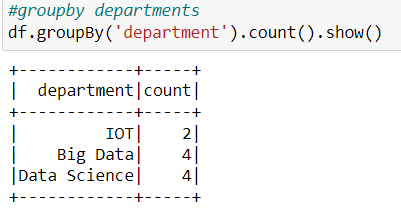
A screenshot of a computer code

Description automatically generated

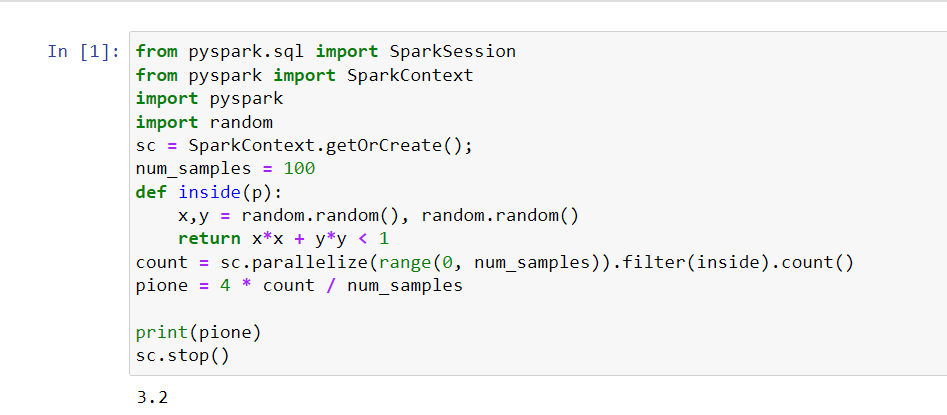
Groupby:

A screenshot of a computer

Description automatically generated



Program:



ETL :

ETL is procedure of migrating data from one system to another.

Extraction:is process of retrieving data out of homogenous or heterogenous sources.

Transformation:during data processing data is being cleaned and incorrect or inaccurate records are being modified or deleted.

Load:finally , the processed data is loaded into target systems.

Orders dataset:

A table of names on a white background

Description automatically generated

A screenshot of a computer

Description automatically generated

Final dataset:

A screenshot of a computer

Description automatically generated