PYSPARK PRACTICE

🡪Creating a Pyspark session.

🡪Giving the required credentials to connect to the AWS.

🡪Importing data from a s3 bucket to spark environment.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

🡪Displaying a particular column of the dataset

A screen shot of a computer screen

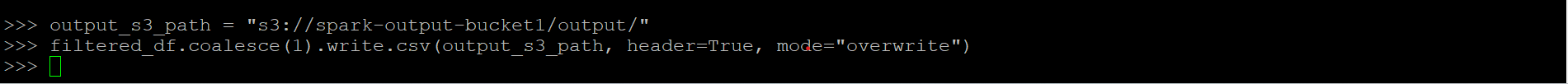
Description automatically generated

A screen shot of a computer

Description automatically generated

🡪Define the output S3 path where you want to store the results.

🡪Writing the filtered Data Frame to a new S3 file in CSV format.



🡪Below we can see that the dataframe we filtered stored in the specified s3 location.

A screenshot of a computer

Description automatically generated

🡪Below is the output file of Avg ratings<3

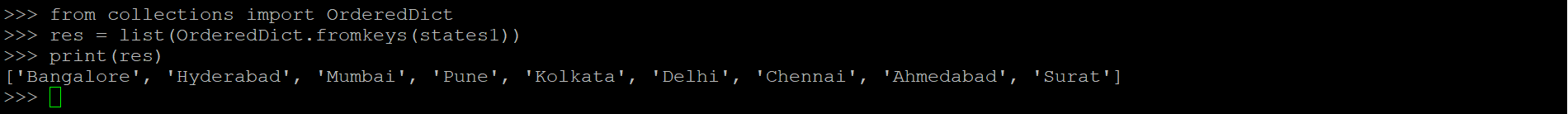
A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

🡪below we perform the distinct values in a column.

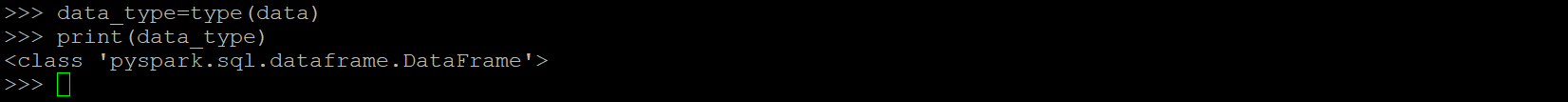


🡪Date incrementation

A black background with many small lights

Description automatically generated

🡪To check the type of the data

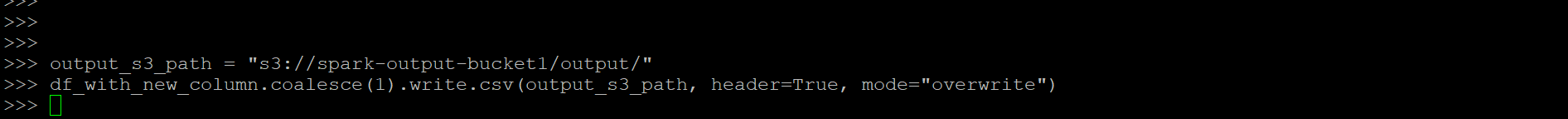


🡪To add a new column date with constant value

A screen shot of a computer

Description automatically generated

🡪storing the output in the s3



🡪New column with date

A screenshot of a computer

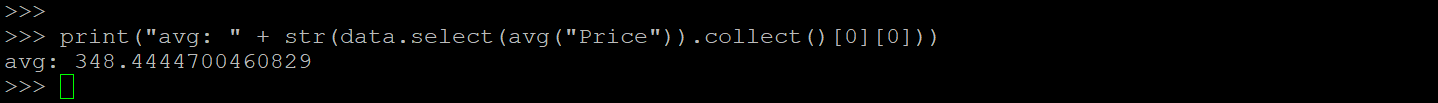
Description automatically generated

🡪**approx\_count\_distinct**: Estimates the approximate number of distinct values in a column.

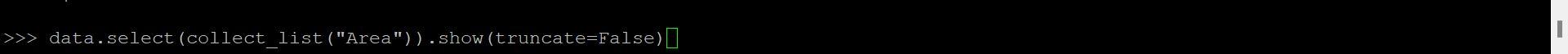
A black screen with white text

Description automatically generated

🡪**avg**: Calculates the average of a numerical column.



🡪**collect\_list**: Collects the elements of a column into a list.



A screen shot of a computer screen

Description automatically generated

🡪**collect\_set**: Collects the unique elements of a column into a set.

A computer screen with a keyboard

Description automatically generated

🡪**countDistinct**: Counts the distinct values of selected columns.

A black background with a black square

Description automatically generated

* **first**: Returns the first element of a column.
* **last**: Returns the last element of a column.
* **kurtosis**: Computes the kurtosis of a numerical column.
* **max**: Finds the maximum value in a numerical column.
* **min**: Finds the minimum value in a numerical column.
* **mean**: Computes the mean of a numerical column.
* **skewness**: Computes the skewness of a numerical column.
* **stddev**: Calculates the standard deviation of a numerical column.
* **stddev\_samp**: Calculates the sample standard deviation of a numerical column.
* **stddev\_pop**: Calculates the population standard deviation of a numerical column.
* **sum**: Calculates the sum of a numerical column.
* **sumDistinct**: Calculates the sum of distinct values in a numerical column.
* **variance**: Computes the variance of a numerical column.
* **var\_samp**: Computes the sample variance of a numerical column.
* **var\_pop**: Computes the population variance of a numerical column.

A screenshot of a computer

Description automatically generated

A computer screen with text on it

Description automatically generated

