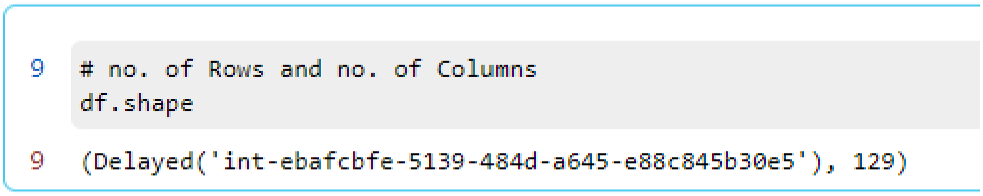
Name: Giorgos Moysiadis
Internship Batch Code: LISUM02
Submission Date: 2/8/2021
Submitted to: Data Glacier

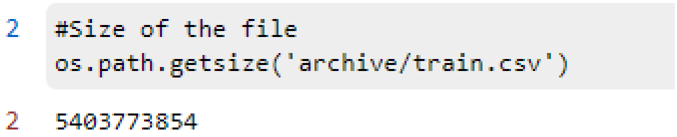
I used Kaggle's "The Shifts Weather Prediction Dataset" that consisted of around 3 million data points (jupyter notebook couldn't process the number) and 129 columns.



```
9   # no. of Rows and no. of Columns
    df.shape

9   (Delayed('int-ebafcbfe-5139-484d-a645-e88c845b30e5'), 129)
```

The file size was around 5 GB



```
2   #Size of the file
    os.path.getsize('archive/train.csv')

2   5403773854
```

After applying the yaml schema the final train.csv.gz is around divided into 162 parts and a total size of 1.6 GB