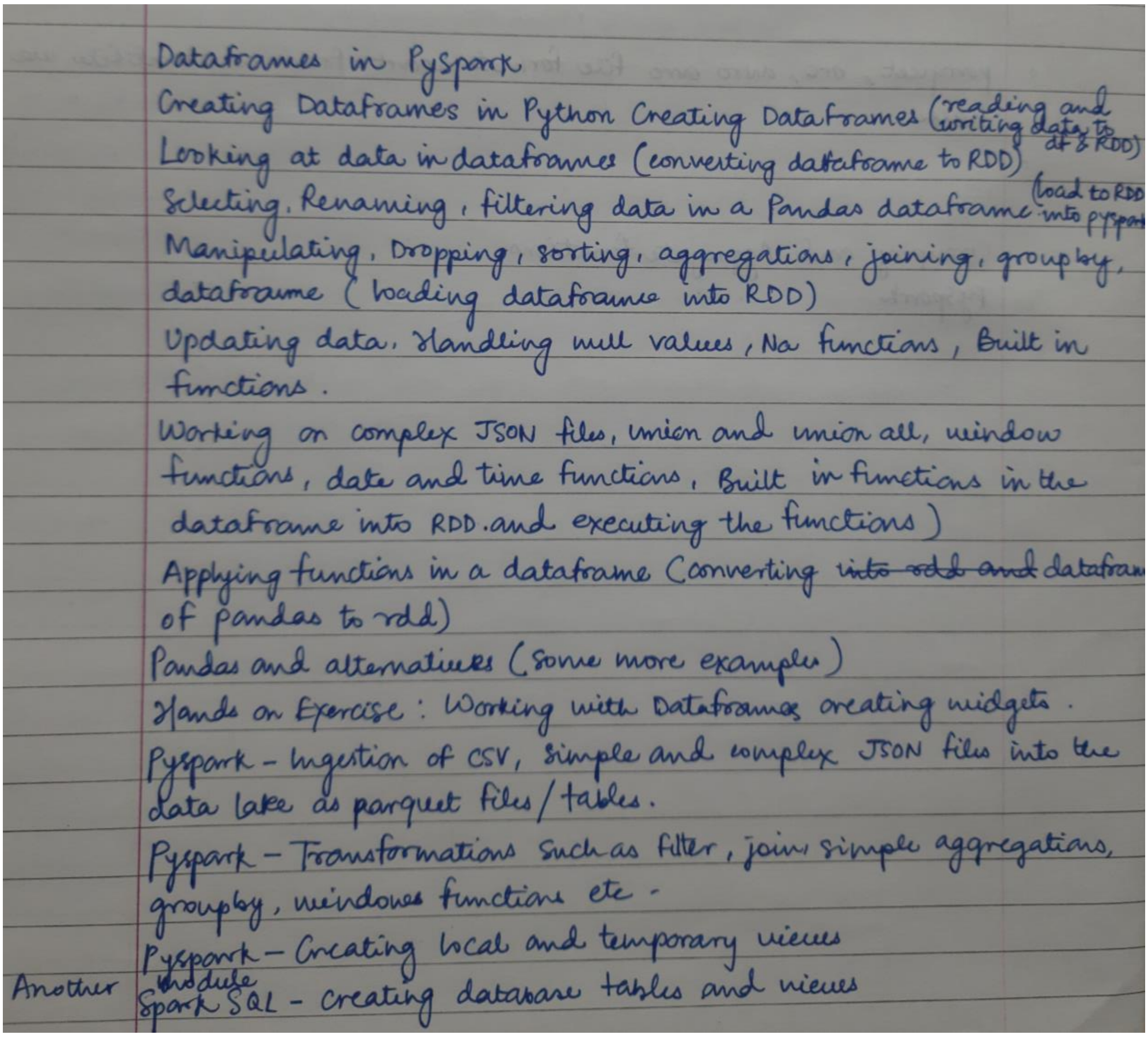
VAISHALI BOKADIYA

PYSPARK DAY 4 ASSESSMENT

Notes:



Dataframes in Pyspark

Creating Dataframes in Python Creating Dataframes (reading and writing data to df & RDD)

Looking at data in dataframes (converting dataframe to RDD)

Selecting, Renaming, filtering data in a Pandas dataframe (load to RDD into pyspark

Manipulating, Dropping, sorting, aggregations, joining, groupby, dataframe (loading dataframe into RDD)

Updating data, Handling null values, Na functions, Built in functions.

Working on complex JSON files, union and union all, window functions, date and time functions, Built in functions in the dataframe into RDD.and executing the functions)

Applying functions in a dataframe (Converting ~~into rdd and~~ datafram of pandas to rdd)

Pandas and alternatives (some more examples)

Hands on Exercise: Working with Dataframes creating widgets.

Pyspark – Ingestion of CSV, simple and complex JSON files into the data lake as parquet files/tables.

Pyspark – Transformations such as filter, join, simple aggregations, groupby, windows functions etc -

Pyspark – Creating local and temporary views

Another module
Spark SQL - creating database tables and views

Parquet, orc, avro are file formats apart from csv, textfile.

Read Text file into Pyspark DataFrame.

⇒ spark.read.format("text").load("output.txt") ~~or~~
OR spark.read.csv or spark.read.csv or spark.read.text.

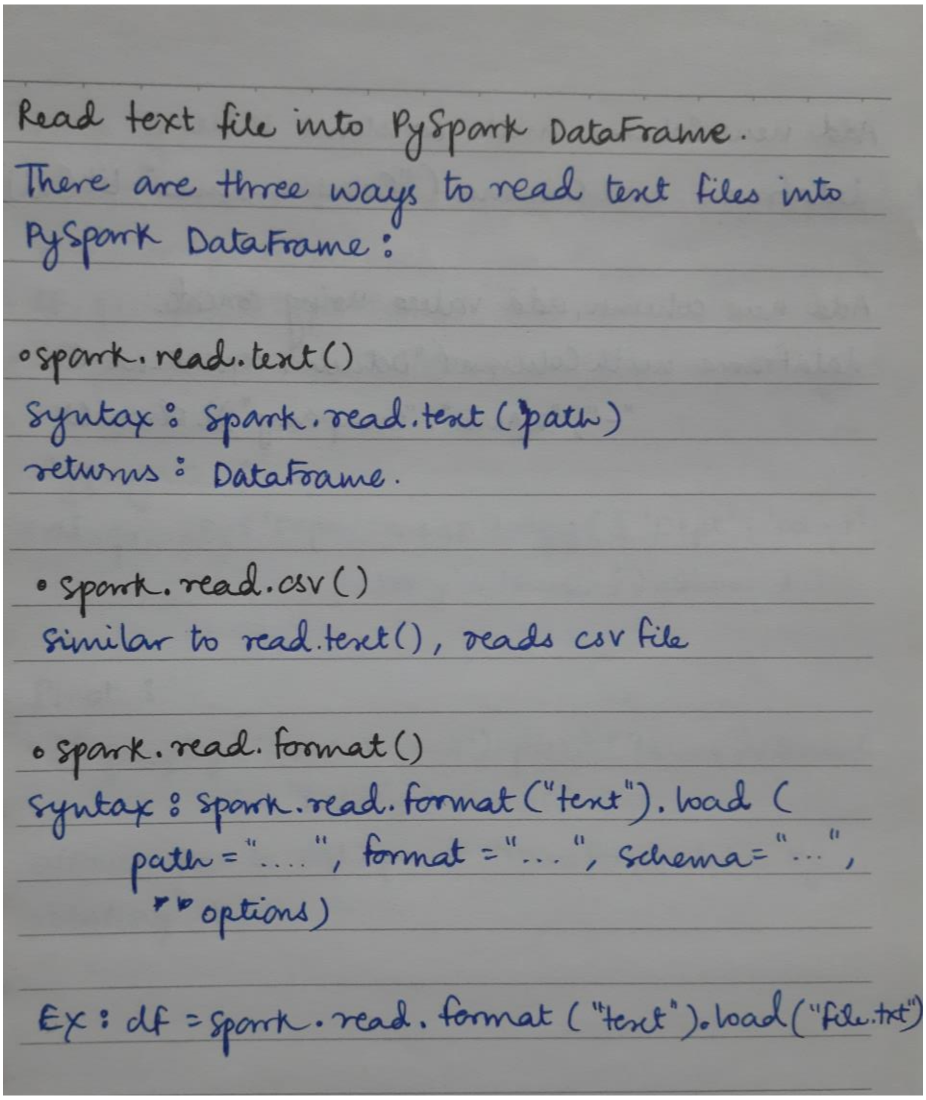
Process of execution :
It converts text file into dataframe,
when it runs on the pyspark notebook, it stores as RDD.

ex: program
→ from pyspark.sql import sparkSession
(imports spark session of library from pyspark.sql module)

→ spark = SparkSession.builder.getOrCreate()
(inorder to create a session of spark to run the program, initialization starts from "SparkSession.builder.getOrCreate()" command)

Here program will start using specific method
→ df = spark.read.format("text").load("output.txt)
(using selectExpr, run expression and save as column name and splits method do split the columns and show is giving member of rows)

→ df.selectExpr("split(value," ") as name-of-column").show(4,False)



Read text file into PySpark DataFrame.

There are three ways to read text files into PySpark DataFrame:

- spark.read.text()

syntax: spark.read.text (path)
returns: DataFrame.

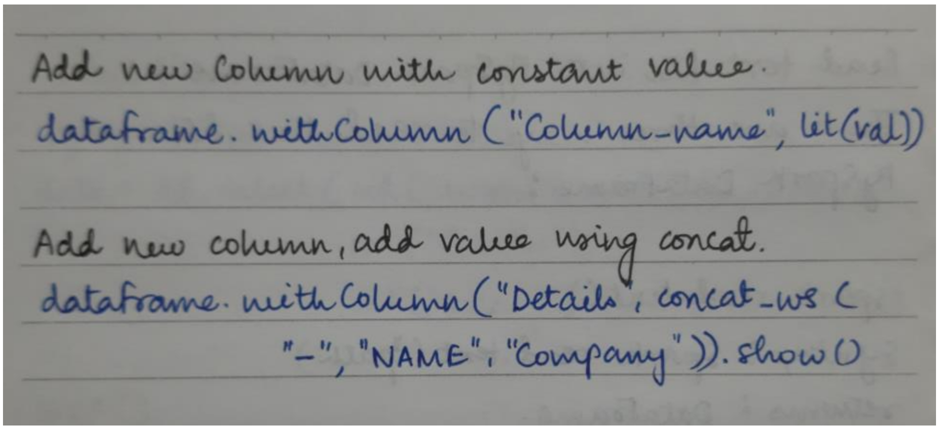
- spark.read.csv()

Similar to read.text(), reads csv file

- spark.read.format()

```
syntax: spark.read.format("text").load (
        path="....", format="...", schema="...",
         **options)
```

```
Ex: df = spark.read.format ("text").load("file.txt")
```

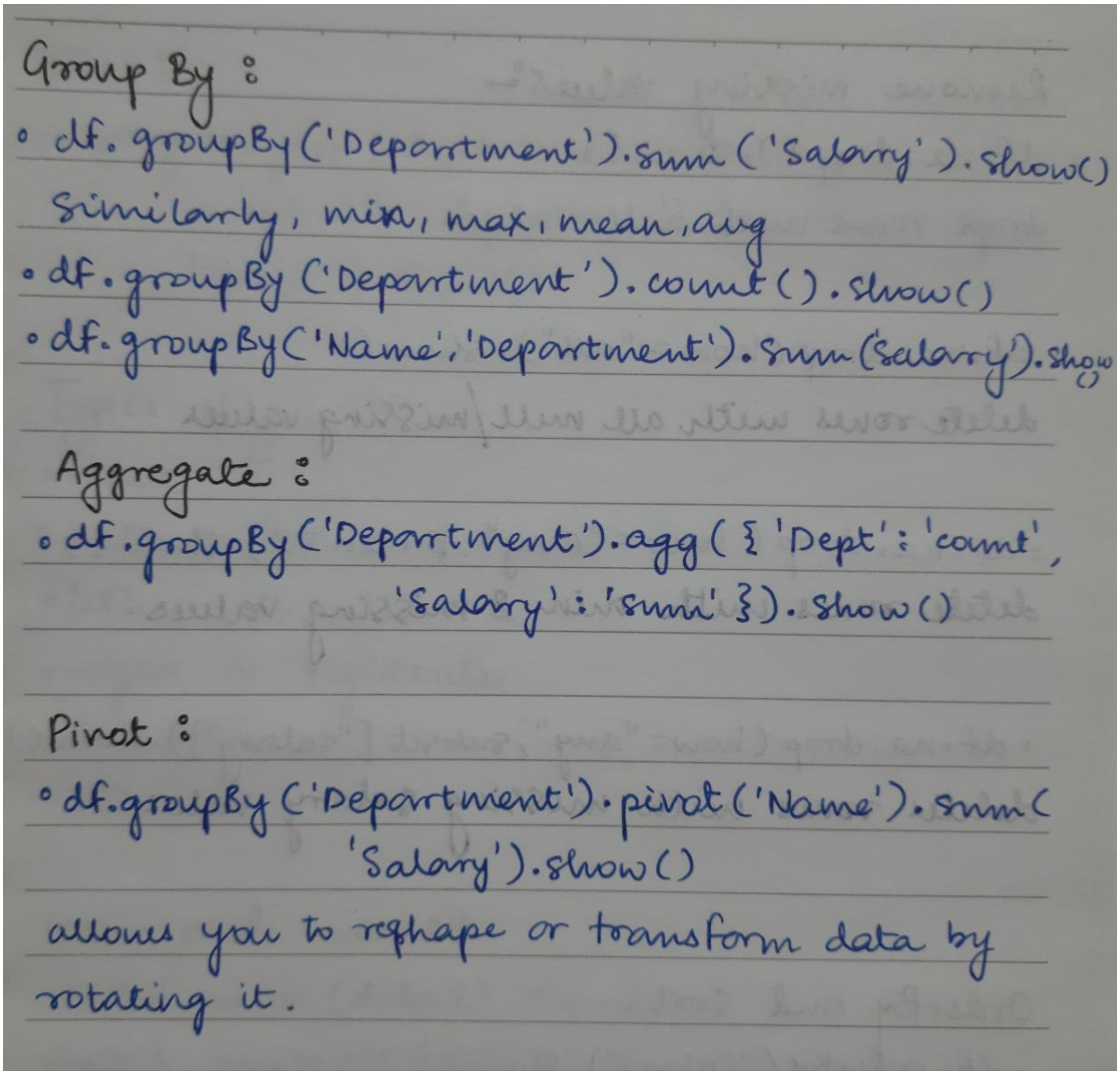


Add new Column with constant value.

```
dataframe.withColumn ("Column_name", lit(val))
```

Add new column, add values using concat.

```
dataframe.withColumn("Details", concat_ws (
                "-", "NAME", "Company")).show()
```



Group By :

- df.groupBy('Department').sum('Salary').show()

  similarly, min, max, mean, avg
- df.groupBy('Department').count().show()
- df.groupBy('Name','Department').sum('Salary').show()

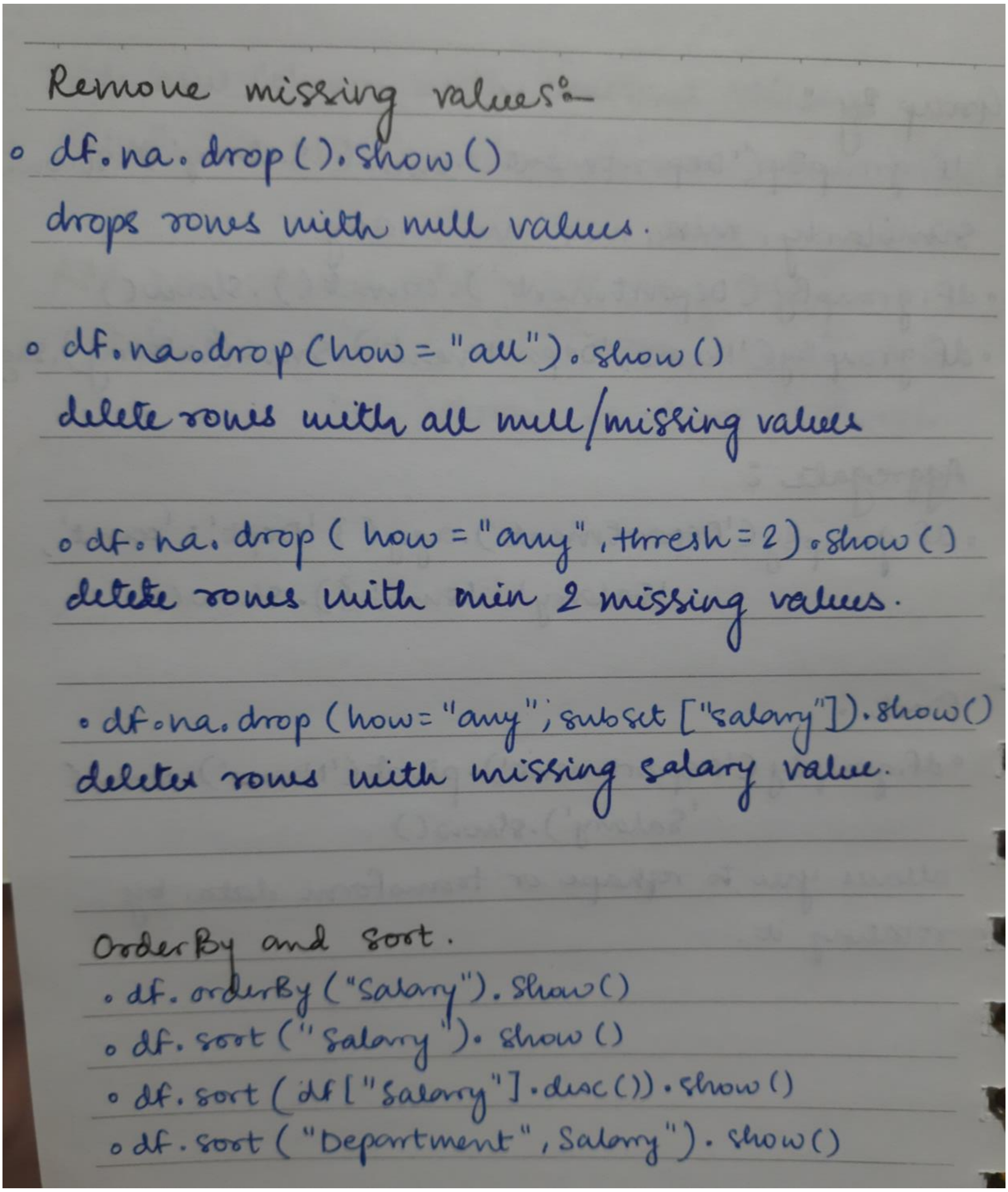
Aggregate :

- df.groupBy('Department').agg({'Dept':'count',
  'Salary':'sum'}).show()

Pivot :

- df.groupBy('Department').pivot('Name').sum(
  'Salary').show()

allows you to reshape or transform data by rotating it.



Remove missing values:-

- df.na.drop().show()

  drops rows with null values.

- df.na.drop(how = "all").show()

  delete rows with all null/missing values

- df.na.drop(how = "any", thresh = 2).show()

  delete rows with min 2 missing values.

- df.na.drop (how="any"; subset ["salary"]).show()

  deletes rows with missing salary value.

OrderBy and sort.

- df.orderBy("Salary").Show()
- df.sort("Salary").show()
- df.sort(df["Salary"].disc()).show()
- df.sort("Department", Salary").show()

Remove missing values:-

- df.na.drop().show()

drops rows with null values.

- df.na.drop(how="all").show()

delete rows with all null/missing values

- df.na.drop(how="any", thresh=2).show()

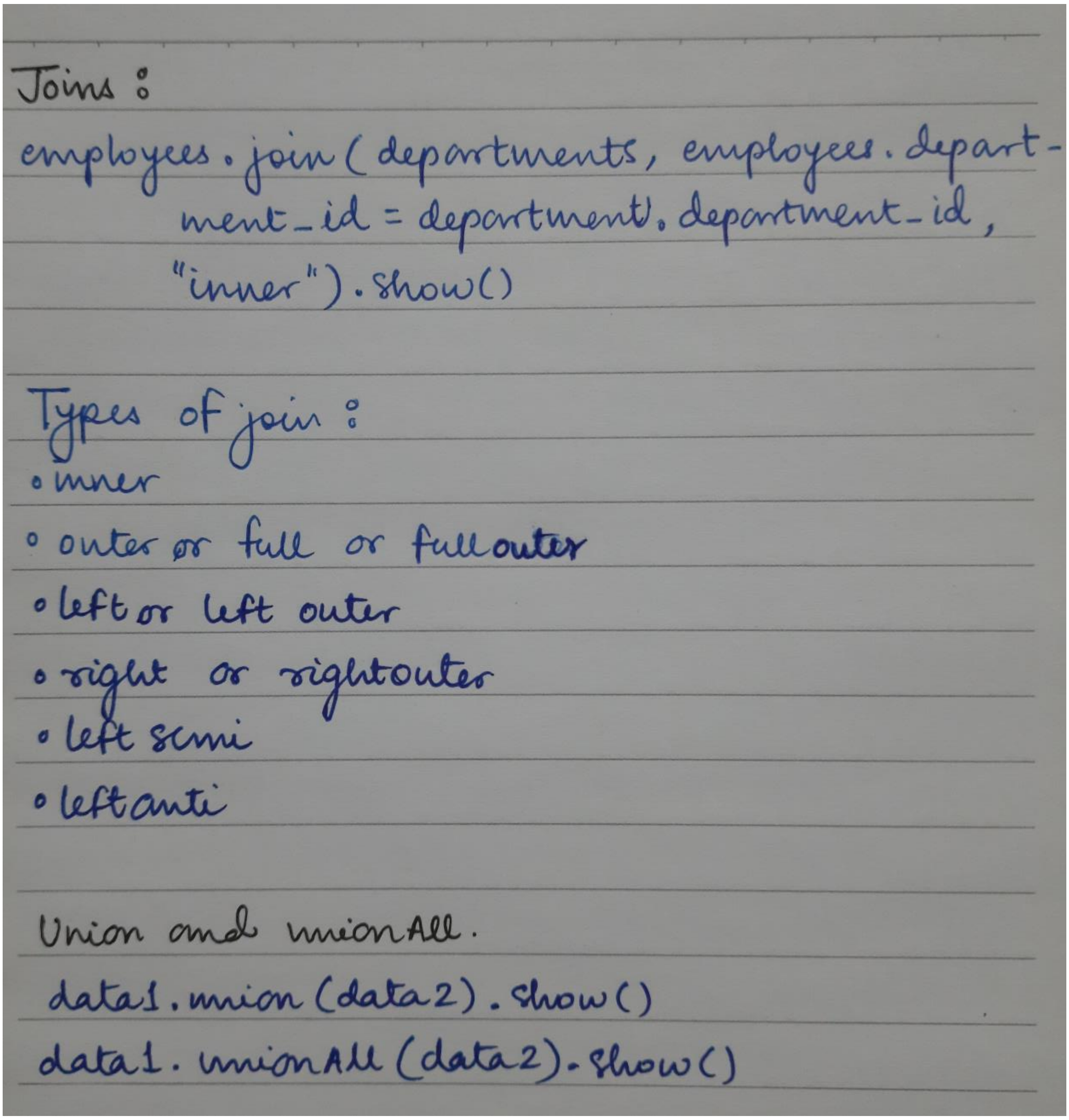
delete rows with min 2 missing values.

- df.na.drop(how="any"; subset ["salary"]).show()

deletes rows with missing salary value.

OrderBy and sort.

- df.orderBy("Salary").Show()
- df.sort("Salary").show()
- df.sort(df["Salary"].disc()).show()
- df.sort("Department", Salary").show()



Joins :

```
employees.join(departments, employees.department_id = departments.department_id,
        "inner").show()
```

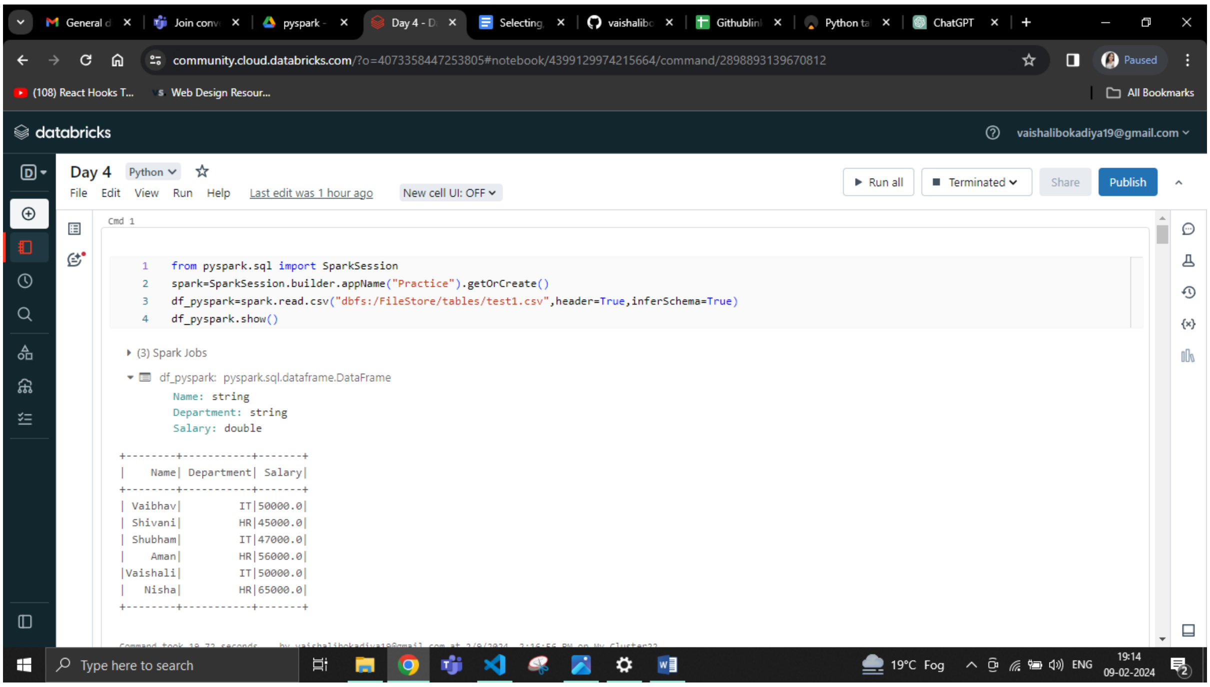
Types of join :

- inner
- outer or full or fullouter
- left or left outer
- right or rightouter
- left semi
- leftanti

Union and unionAll.

```
data1.union(data2).show()
data1.unionAll(data2).show()
```

Practice:

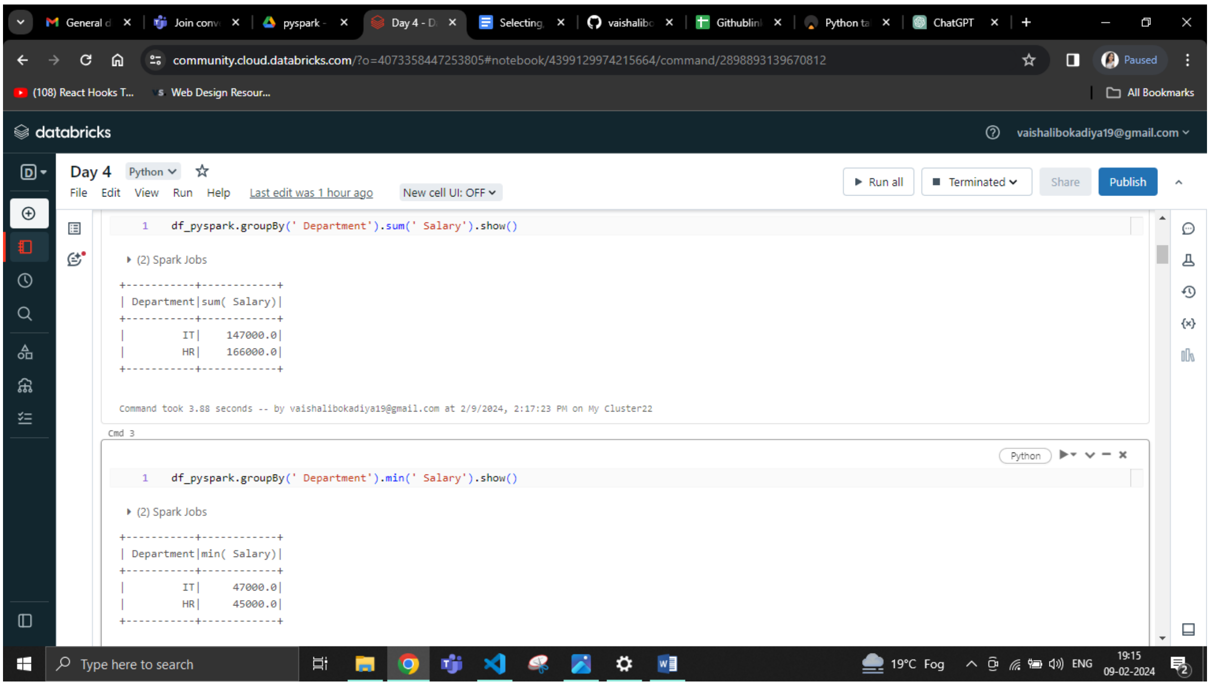


Cmd 1

```python
from pyspark.sql import SparkSession
spark=SparkSession.builder.appName("Practice").getOrCreate()
df_pyspark=spark.read.csv("dbfs:/FileStore/tables/test1.csv",header=True,inferSchema=True)
df_pyspark.show()
```

(3) Spark Jobs

```
df_pyspark:  pyspark.sql.dataframe.DataFrame
      Name:  string
      Department:  string
      Salary:  double

+--------+----------+-------+
|    Name| Department| Salary|
+--------+----------+-------+
| Vaibhav|         IT|50000.0|
| Shivani|         HR|45000.0|
| Shubham|         IT|47000.0|
|    Aman|         HR|56000.0|
|Vaishali|         IT|50000.0|
|   Nisha|         HR|65000.0|
+--------+----------+-------+

Command took 10.72 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 2:16:56 PM on My Cluster22
```



```python
df_pyspark.groupBy(' Department').sum(' Salary').show()
```

(2) Spark Jobs

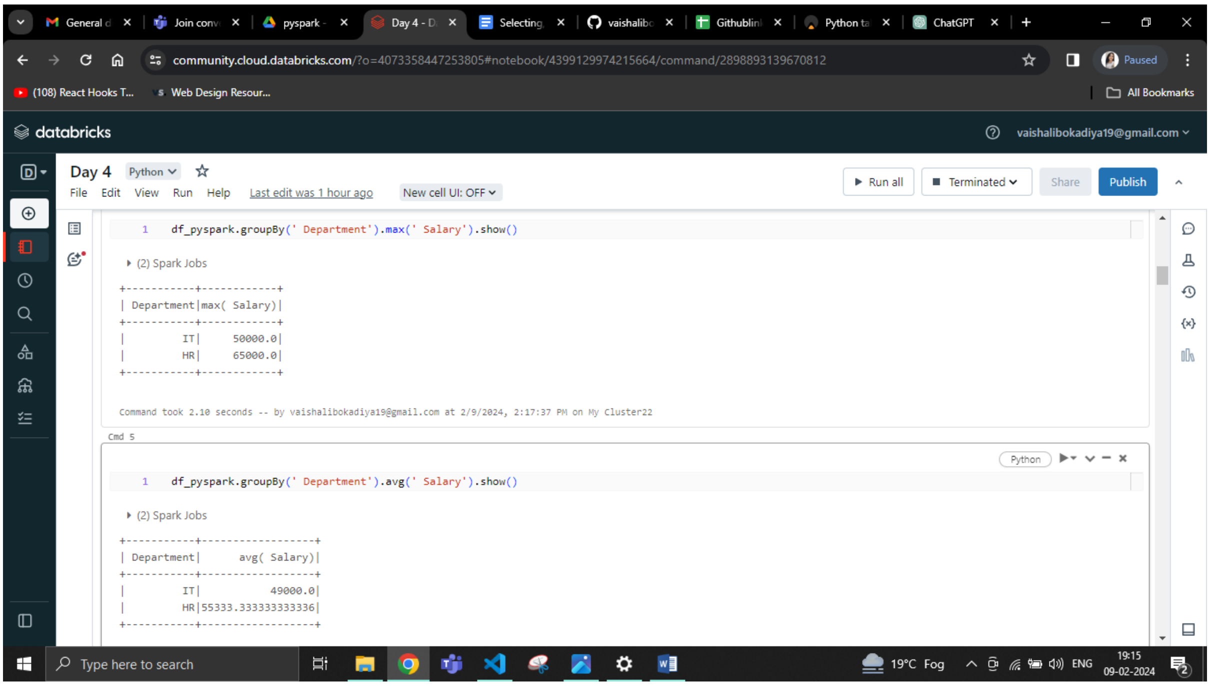
```
+-----------+-----------+
| Department|sum( Salary)|
+-----------+-----------+
|         IT|   147000.0|
|         HR|   166000.0|
+-----------+-----------+

Command took 3.88 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 2:17:23 PM on My Cluster22
```

Cmd 3

```python
df_pyspark.groupBy(' Department').min(' Salary').show()
```

(2) Spark Jobs

```
+-----------+-----------+
| Department|min( Salary)|
+-----------+-----------+
|         IT|    47000.0|
|         HR|    45000.0|
+-----------+-----------+
```



```
df_pyspark.groupBy(' Department').max(' Salary').show()
```

(2) Spark Jobs

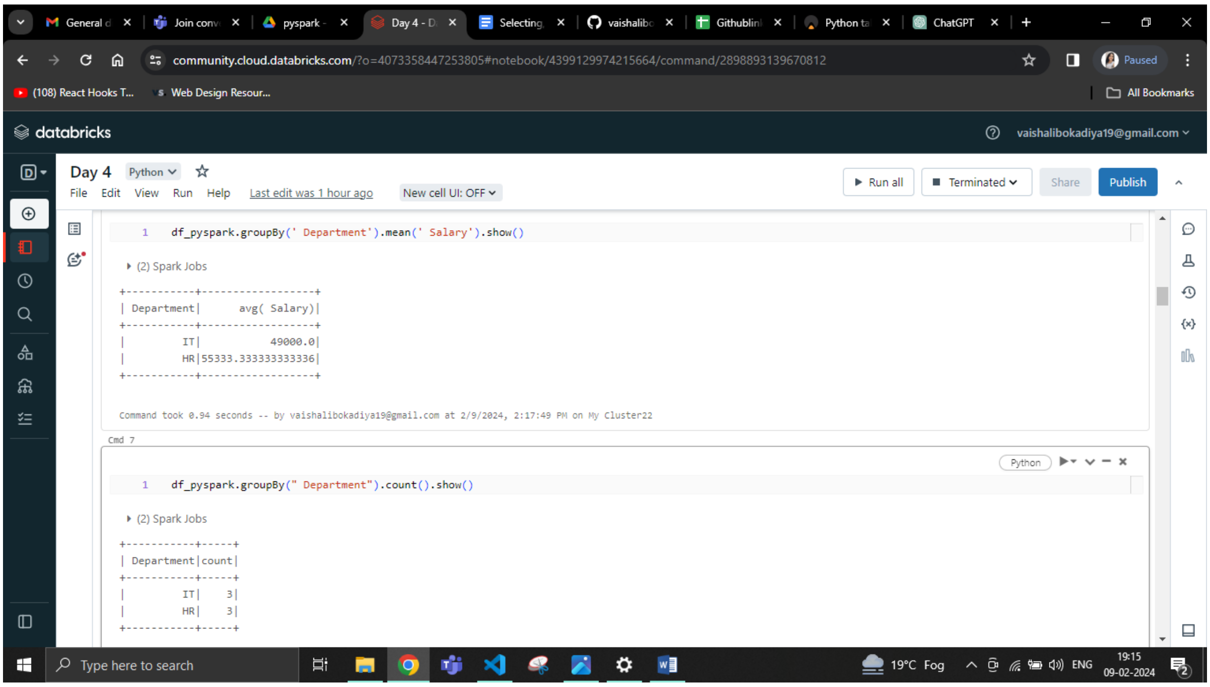
```
+----------+-----------+
| Department|max( Salary)|
+----------+-----------+
|         IT|     50000.0|
|         HR|     65000.0|
+----------+-----------+

Command took 2.10 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 2:17:37 PM on My Cluster22
```

Cmd 5

```
df_pyspark.groupBy(' Department').avg(' Salary').show()
```

(2) Spark Jobs

```
+----------+-----------------+
| Department|      avg( Salary)|
+----------+-----------------+
|         IT|          49000.0|
|         HR|55333.333333333336|
+----------+-----------------+
```



```
df_pyspark.groupBy(' Department').mean(' Salary').show()
```

(2) Spark Jobs

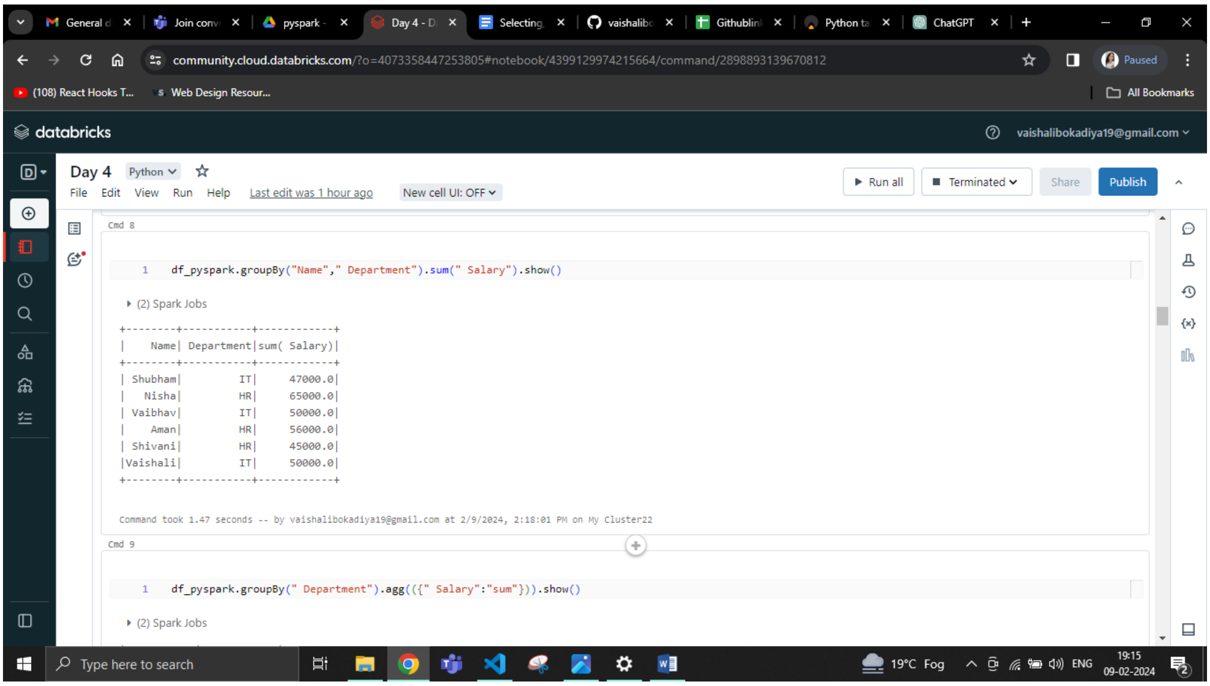
```
+----------+-----------------+
| Department|      avg( Salary)|
+----------+-----------------+
|         IT|          49000.0|
|         HR|55333.333333333336|
+----------+-----------------+

Command took 0.94 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 2:17:49 PM on My Cluster22
```

Cmd 7

```
df_pyspark.groupBy(" Department").count().show()
```

(2) Spark Jobs

```
+----------+-----+
| Department|count|
+----------+-----+
|         IT|    3|
|         HR|    3|
+----------+-----+
```



community.cloud.databricks.com/?o=4073358447253805#notebook/4399129974215664/command/2898893139670812
databricks
vaishalibokadiya19@gmail.com
Day 4
Python
File
Edit
View
Run
Help
Last edit was 1 hour ago
New cell UI: OFF
Run all
Terminated
Share
Publish
Cmd 8

```
df_pyspark.groupBy("Name"," Department").sum(" Salary").show()
```

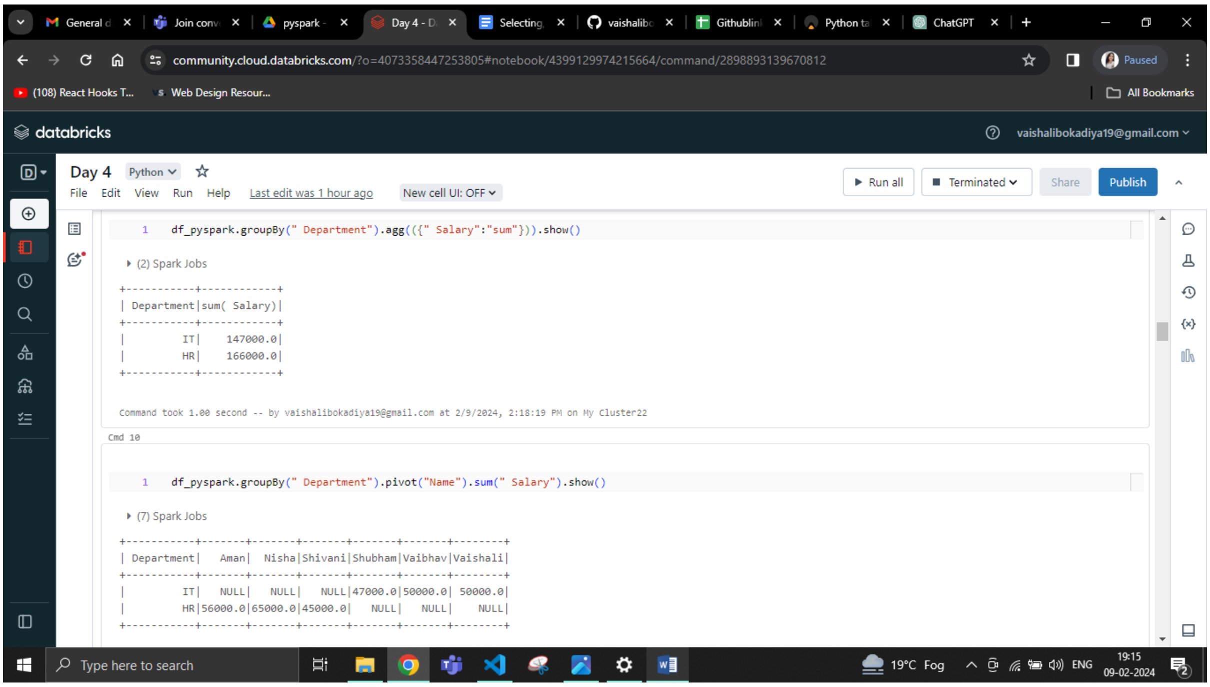
(2) Spark Jobs

```
+--------+----------+-----------+
|    Name| Department|sum( Salary)|
+--------+----------+-----------+
| Shubham|         IT|     47000.0|
|   Nisha|         HR|     65000.0|
| Vaibhav|         IT|     50000.0|
|    Aman|         HR|     56000.0|
| Shivani|         HR|     45000.0|
|Vaishali|         IT|     50000.0|
+--------+----------+-----------+
```

```
Command took 1.47 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 2:18:01 PM on My Cluster22
```

```
df_pyspark.groupBy(" Department").agg(({" Salary":"sum"})).show()
```

(2) Spark Jobs



community.cloud.databricks.com/?o=4073358447253805#notebook/4399129974215664/command/2898893139670812
databricks
vaishalibokadiya19@gmail.com
Day 4
Python
File
Edit
View
Run
Help
Last edit was 1 hour ago
New cell UI: OFF
Run all
Terminated
Share
Publish

```
df_pyspark.groupBy(" Department").agg(({" Salary":"sum"})).show()
```

(2) Spark Jobs

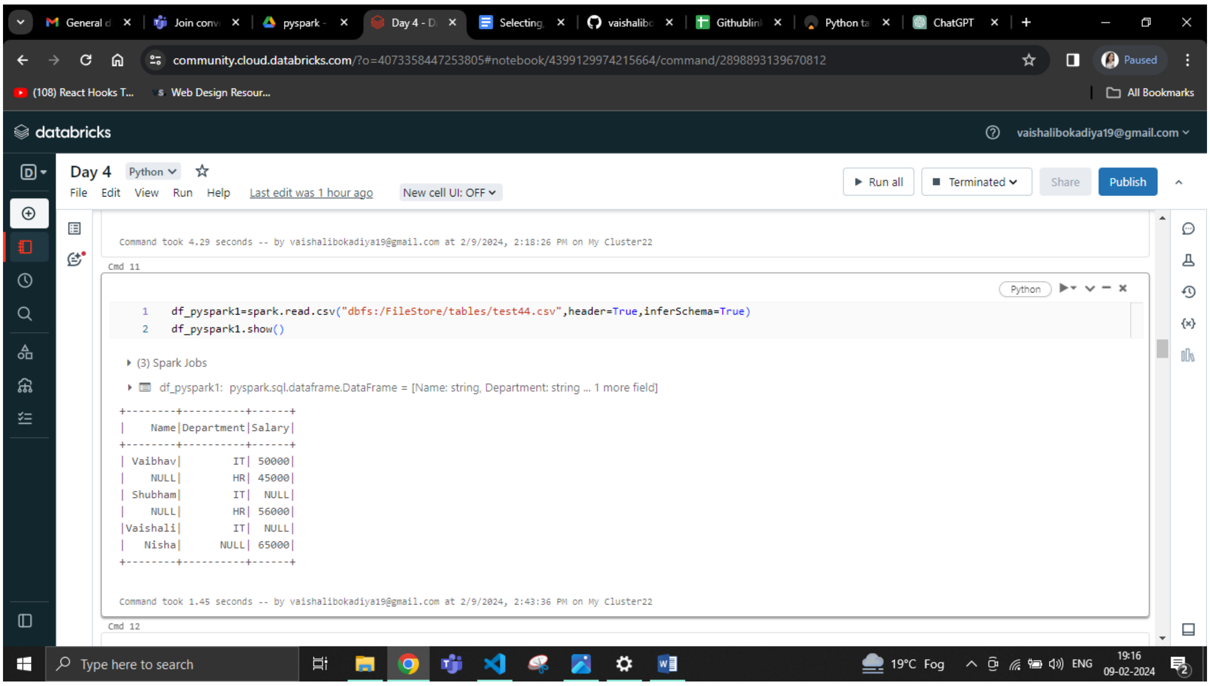
```
+-----------+-----------+
| Department|sum( Salary)|
+-----------+-----------+
|         IT|    147000.0|
|         HR|    166000.0|
+-----------+-----------+
```

```
Command took 1.00 second -- by vaishalibokadiya19@gmail.com at 2/9/2024, 2:18:19 PM on My Cluster22
```

```
df_pyspark.groupBy(" Department").pivot("Name").sum(" Salary").show()
```

(7) Spark Jobs

```
+-----------+-------+-------+-------+-------+-------+--------+
| Department|   Aman|  Nisha|Shivani|Shubham|Vaibhav|Vaishali|
+-----------+-------+-------+-------+-------+-------+--------+
|         IT|   NULL|   NULL|   NULL|47000.0|50000.0| 50000.0|
|         HR|56000.0|65000.0|45000.0|   NULL|   NULL|    NULL|
+-----------+-------+-------+-------+-------+-------+--------+
```



Command took 4.29 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 2:18:26 PM on My Cluster22

Cmd 11

```
df_pyspark1=spark.read.csv("dbfs:/FileStore/tables/test44.csv",header=True,inferSchema=True)
df_pyspark1.show()
```

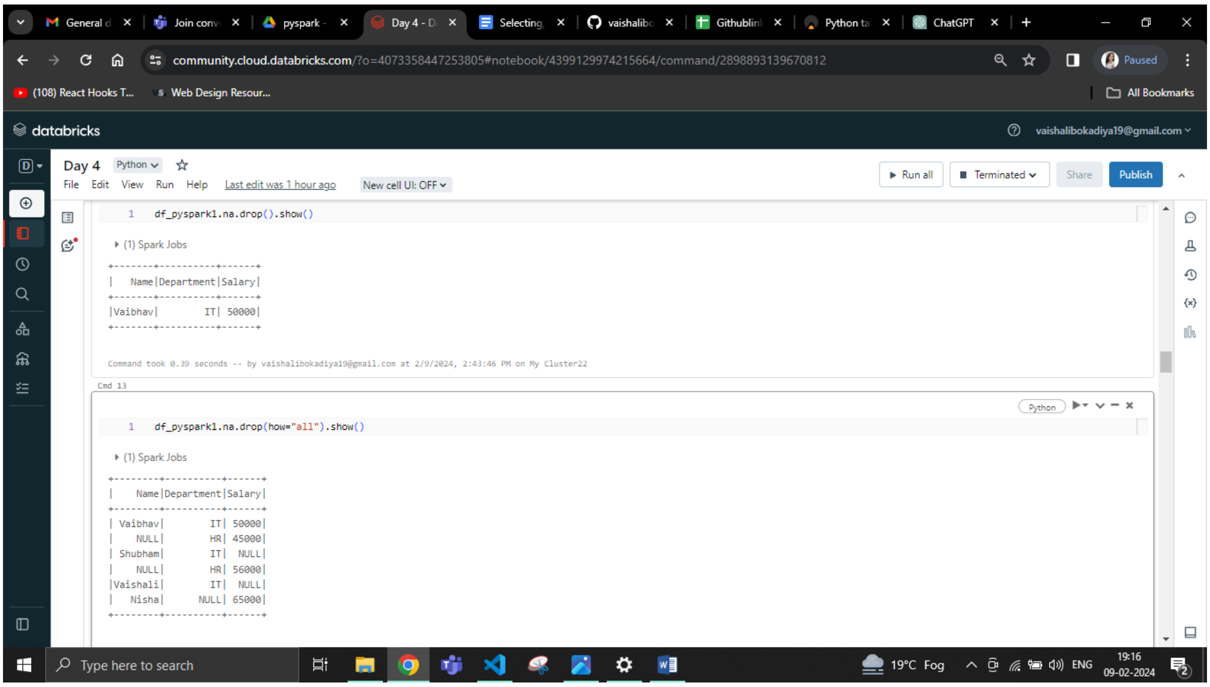
(3) Spark Jobs

df_pyspark1: pyspark.sql.dataframe.DataFrame = [Name: string, Department: string ... 1 more field]

```
+--------+----------+------+
|    Name|Department|Salary|
+--------+----------+------+
| Vaibhav|        IT| 50000|
|    NULL|        HR| 45000|
| Shubham|        IT|  NULL|
|    NULL|        HR| 56000|
|Vaishali|        IT|  NULL|
|   Nisha|      NULL| 65000|
+--------+----------+------+
```

Command took 1.45 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 2:43:36 PM on My Cluster22

Cmd 12



```
df_pyspark1.na.drop().show()
```

(1) Spark Jobs

```
+-------+----------+------+
|   Name|Department|Salary|
+-------+----------+------+
|Vaibhav|        IT| 50000|
+-------+----------+------+
```

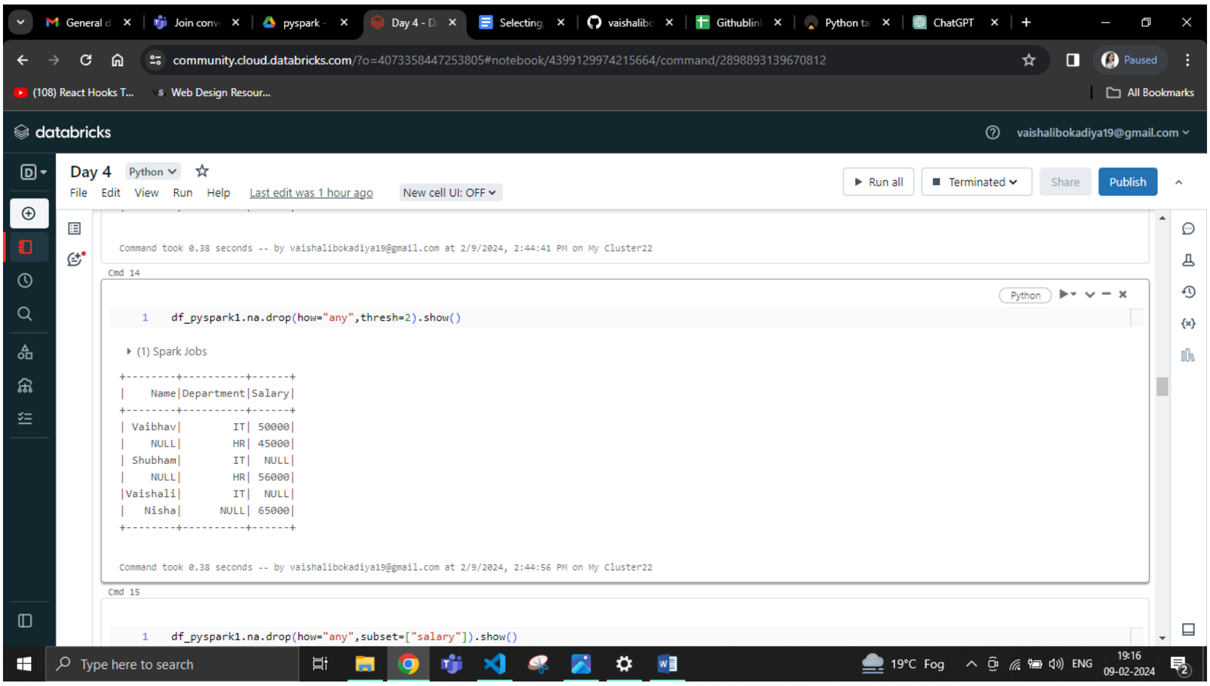
Command took 0.39 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 2:43:46 PM on My Cluster22

Cmd 13

```
df_pyspark1.na.drop(how="all").show()
```

(1) Spark Jobs

```
+--------+----------+------+
|    Name|Department|Salary|
+--------+----------+------+
| Vaibhav|        IT| 50000|
|    NULL|        HR| 45000|
| Shubham|        IT|  NULL|
|    NULL|        HR| 56000|
|Vaishali|        IT|  NULL|
|   Nisha|      NULL| 65000|
+--------+----------+------+
```



```
Command took 0.38 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 2:44:41 PM on My Cluster22
```

Cmd 14

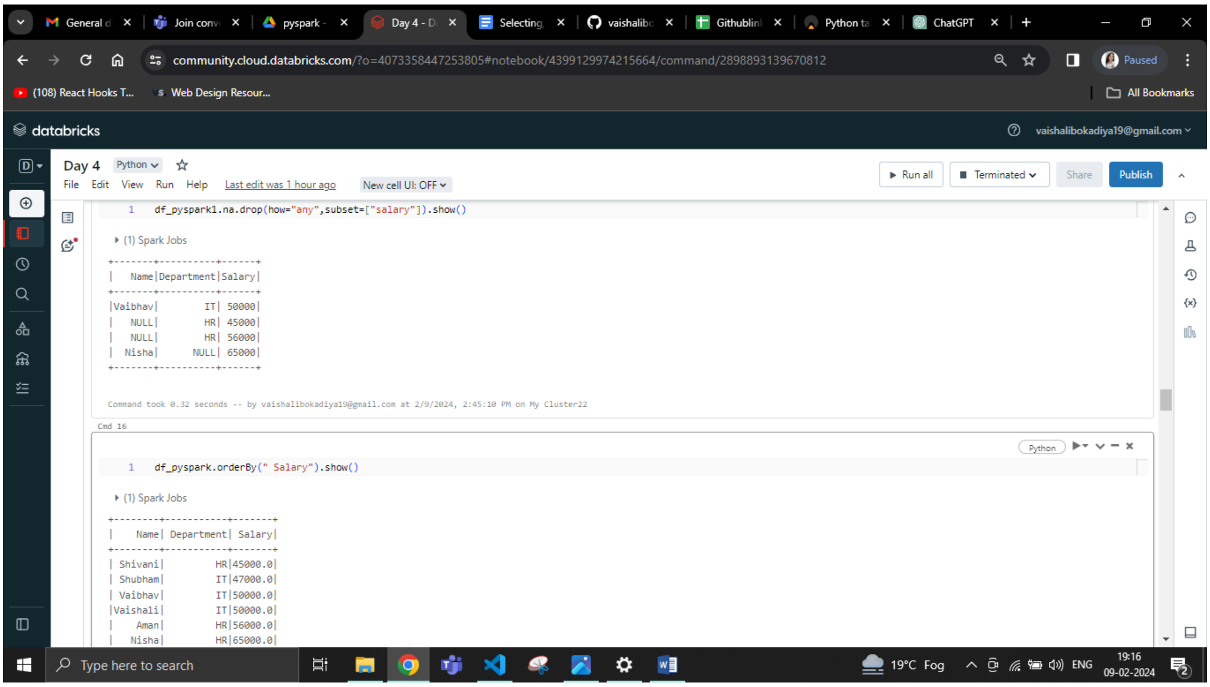
```python
df_pyspark1.na.drop(how="any",thresh=2).show()
```

(1) Spark Jobs

```
+--------+----------+------+
|    Name|Department|Salary|
+--------+----------+------+
| Vaibhav|        IT| 50000|
|    NULL|        HR| 45000|
| Shubham|        IT|  NULL|
|    NULL|        HR| 56000|
|Vaishali|        IT|  NULL|
|   Nisha|      NULL| 65000|
+--------+----------+------+

Command took 0.38 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 2:44:56 PM on My Cluster22
```

Cmd 15

```python
df_pyspark1.na.drop(how="any",subset=["salary"]).show()
```



```python
df_pyspark1.na.drop(how="any",subset=["salary"]).show()
```

(1) Spark Jobs

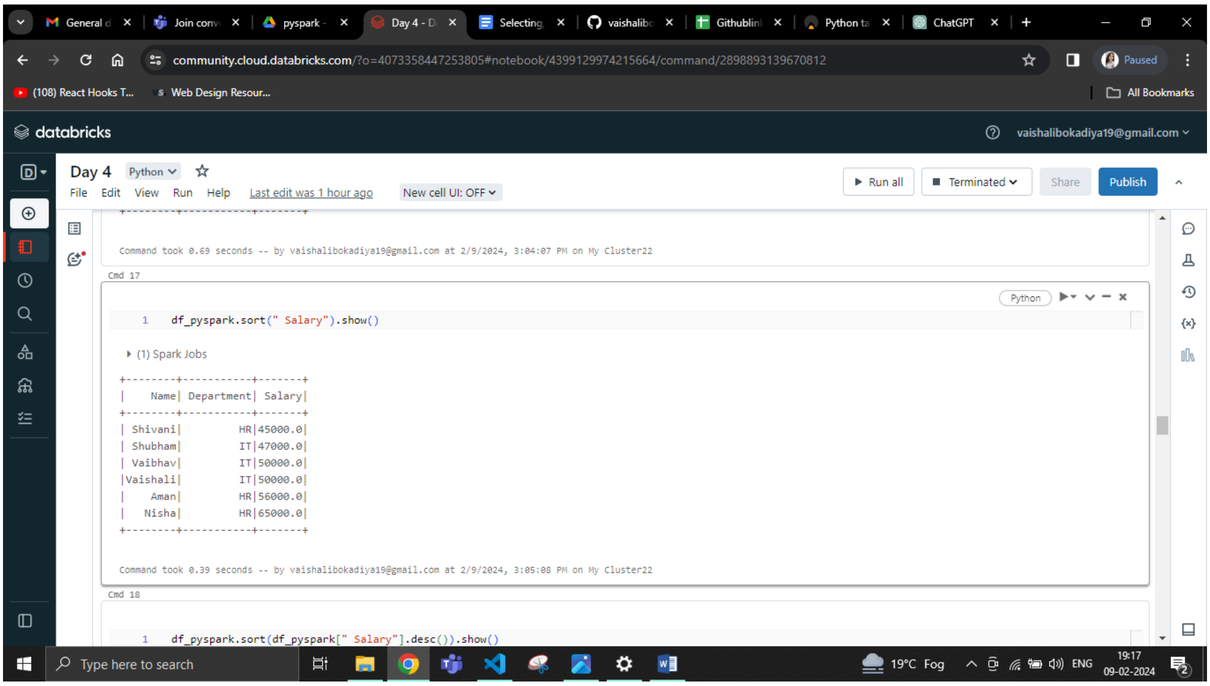
```
+-------+----------+------+
|   Name|Department|Salary|
+-------+----------+------+
|Vaibhav|        IT| 50000|
|   NULL|        HR| 45000|
|   NULL|        HR| 56000|
|  Nisha|      NULL| 65000|
+-------+----------+------+

Command took 0.32 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 2:45:10 PM on My Cluster22
```

Cmd 16

```python
df_pyspark.orderBy(" Salary").show()
```

(1) Spark Jobs

```
+--------+----------+-------+
|    Name| Department| Salary|
+--------+----------+-------+
| Shivani|        HR|45000.0|
| Shubham|        IT|47000.0|
| Vaibhav|        IT|50000.0|
|Vaishali|        IT|50000.0|
|    Aman|        HR|56000.0|
|   Nisha|        HR|65000.0|
```



```
Command took 0.69 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 3:04:07 PM on My Cluster22
```

Cmd 17

```
df_pyspark.sort(" Salary").show()
```

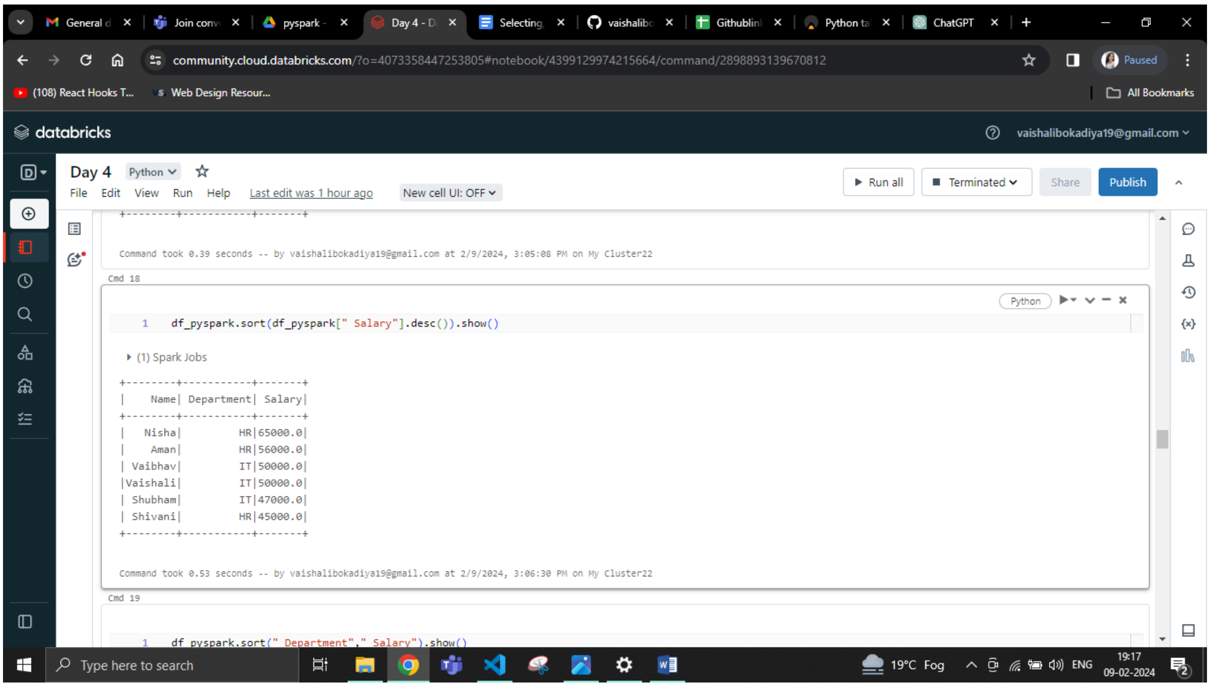
(1) Spark Jobs

```
+--------+-----------+-------+
|    Name| Department| Salary|
+--------+-----------+-------+
| Shivani|         HR|45000.0|
| Shubham|         IT|47000.0|
| Vaibhav|         IT|50000.0|
|Vaishali|         IT|50000.0|
|    Aman|         HR|56000.0|
|   Nisha|         HR|65000.0|
+--------+-----------+-------+
```

```
Command took 0.39 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 3:05:08 PM on My Cluster22
```

Cmd 18

```
df_pyspark.sort(df_pyspark[" Salary"].desc()).show()
```



```
Command took 0.39 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 3:05:08 PM on My Cluster22
```

Cmd 18

```
df_pyspark.sort(df_pyspark[" Salary"].desc()).show()
```

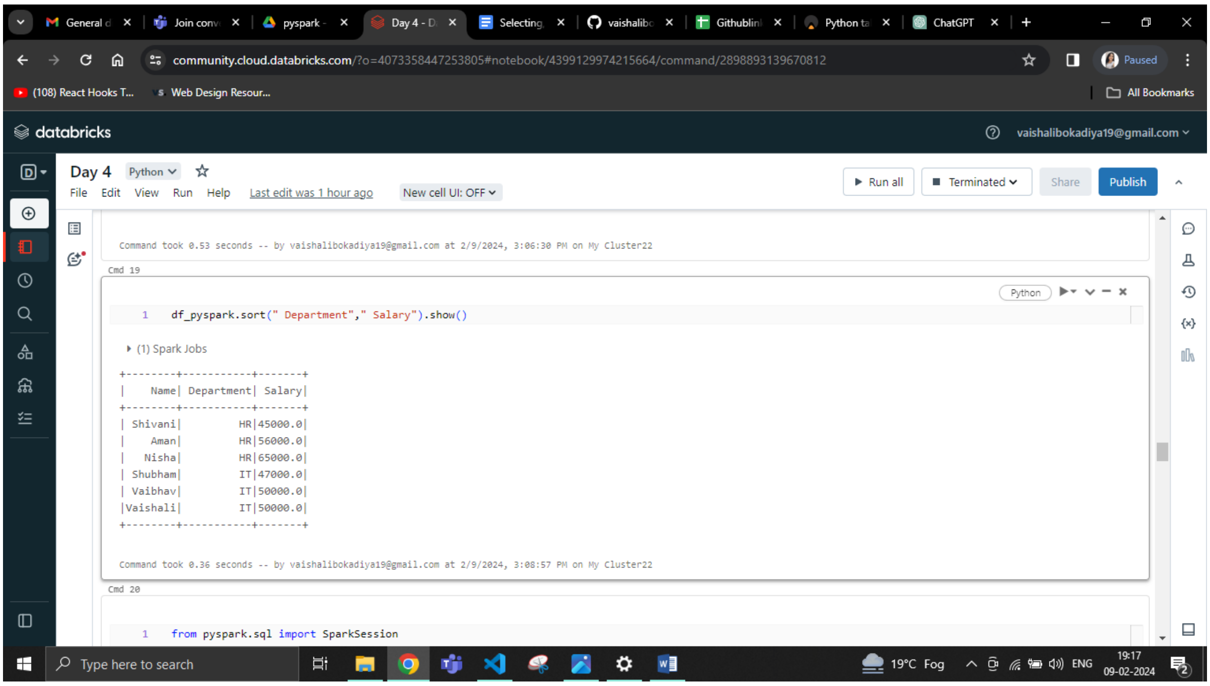
(1) Spark Jobs

```
+--------+-----------+-------+
|    Name| Department| Salary|
+--------+-----------+-------+
|   Nisha|         HR|65000.0|
|    Aman|         HR|56000.0|
| Vaibhav|         IT|50000.0|
|Vaishali|         IT|50000.0|
| Shubham|         IT|47000.0|
| Shivani|         HR|45000.0|
+--------+-----------+-------+
```

```
Command took 0.53 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 3:06:30 PM on My Cluster22
```

Cmd 19

```
df_pyspark.sort(" Department"," Salary").show()
```



```
Command took 0.53 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 3:06:30 PM on My Cluster22
```

Cmd 19

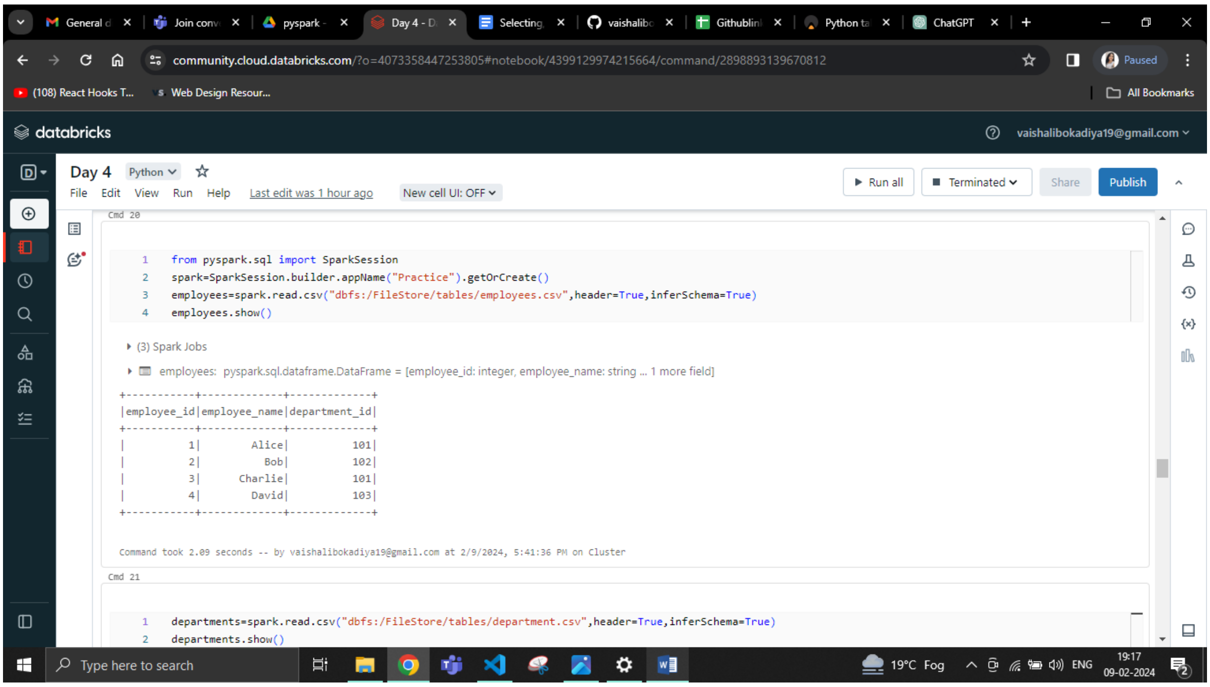
```python
df_pyspark.sort(" Department"," Salary").show()
```

(1) Spark Jobs

```
+--------+----------+-------+
|    Name| Department| Salary|
+--------+----------+-------+
| Shivani|         HR|45000.0|
|    Aman|         HR|56000.0|
|   Nisha|         HR|65000.0|
| Shubham|         IT|47000.0|
| Vaibhav|         IT|50000.0|
|Vaishali|         IT|50000.0|
+--------+----------+-------+

Command took 0.36 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 3:08:57 PM on My Cluster22
```

Cmd 20

```python
from pyspark.sql import SparkSession
```



Cmd 20

```python
from pyspark.sql import SparkSession
spark=SparkSession.builder.appName("Practice").getOrCreate()
employees=spark.read.csv("dbfs:/FileStore/tables/employees.csv",header=True,inferSchema=True)
employees.show()
```

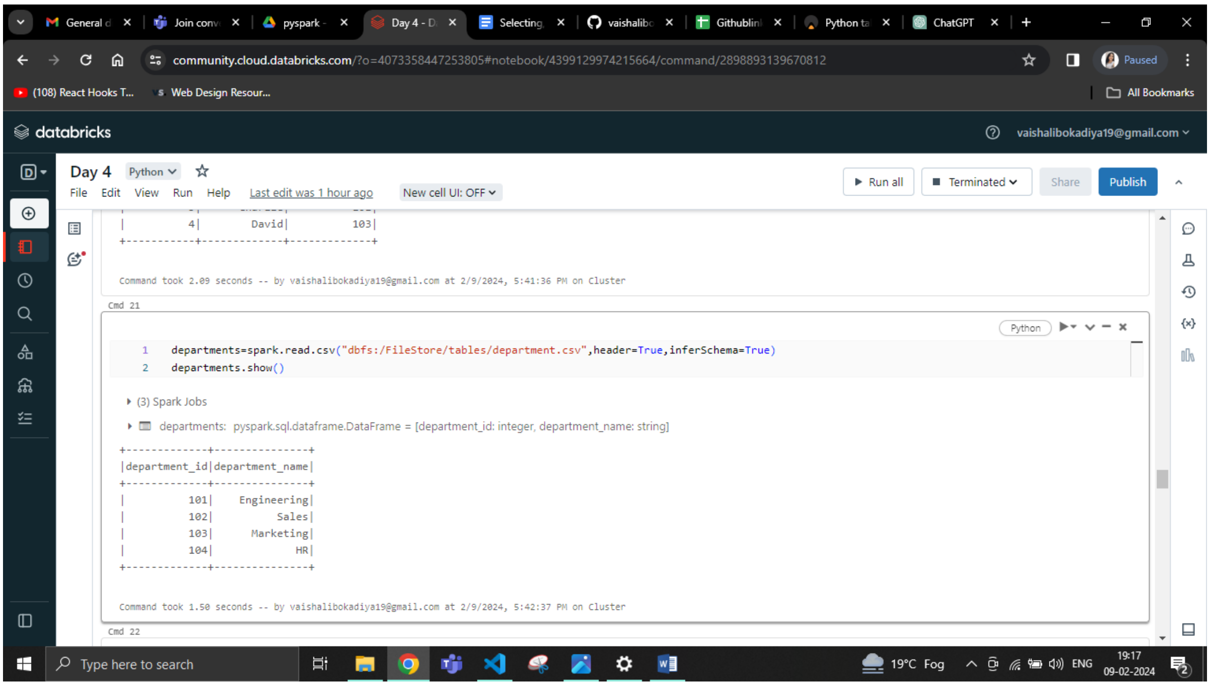
(3) Spark Jobs

employees: pyspark.sql.dataframe.DataFrame = [employee_id: integer, employee_name: string ... 1 more field]

```
+-----------+-------------+-------------+
|employee_id|employee_name|department_id|
+-----------+-------------+-------------+
|          1|        Alice|          101|
|          2|          Bob|          102|
|          3|      Charlie|          101|
|          4|        David|          103|
+-----------+-------------+-------------+

Command took 2.09 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 5:41:36 PM on Cluster
```

Cmd 21

```python
departments=spark.read.csv("dbfs:/FileStore/tables/department.csv",header=True,inferSchema=True)
departments.show()
```



community.cloud.databricks.com/?o=4073358447253805#notebook/4399129974215664/command/2898893139670812
databricks
vaishalibokadiya19@gmail.com
Day 4
Python
File Edit View Run Help
Last edit was 1 hour ago
New cell UI: OFF
Run all
Terminated
Share
Publish

```
|          4|       David|          103|
+-----------+------------+------------+

Command took 2.09 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 5:41:36 PM on Cluster
```

Cmd 21

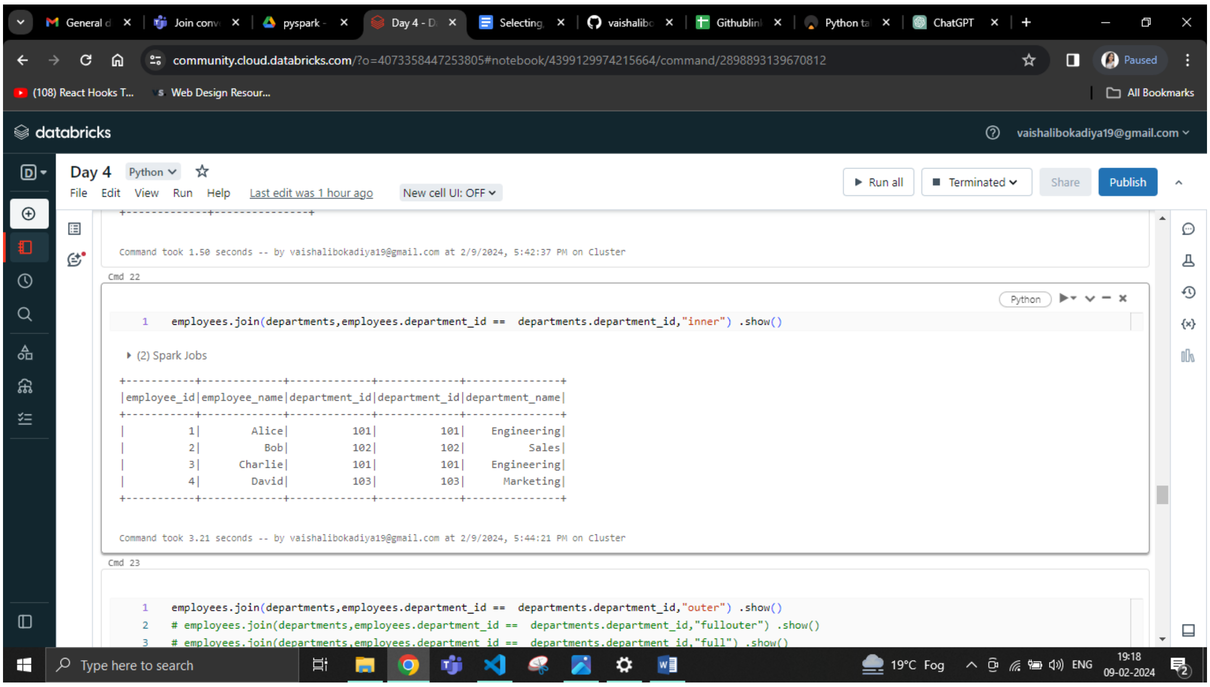
```
departments=spark.read.csv("dbfs:/FileStore/tables/department.csv",header=True,inferSchema=True)
departments.show()
```

(3) Spark Jobs

departments: pyspark.sql.dataframe.DataFrame = [department_id: integer, department_name: string]

```
+-------------+---------------+
|department_id|department_name|
+-------------+---------------+
|          101|    Engineering|
|          102|          Sales|
|          103|      Marketing|
|          104|             HR|
+-------------+---------------+

Command took 1.50 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 5:42:37 PM on Cluster
```

Cmd 22



community.cloud.databricks.com/?o=4073358447253805#notebook/4399129974215664/command/2898893139670812
databricks
vaishalibokadiya19@gmail.com
Day 4
Python
File Edit View Run Help
Last edit was 1 hour ago
New cell UI: OFF
Run all
Terminated
Share
Publish

```
Command took 1.50 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 5:42:37 PM on Cluster
```

Cmd 22

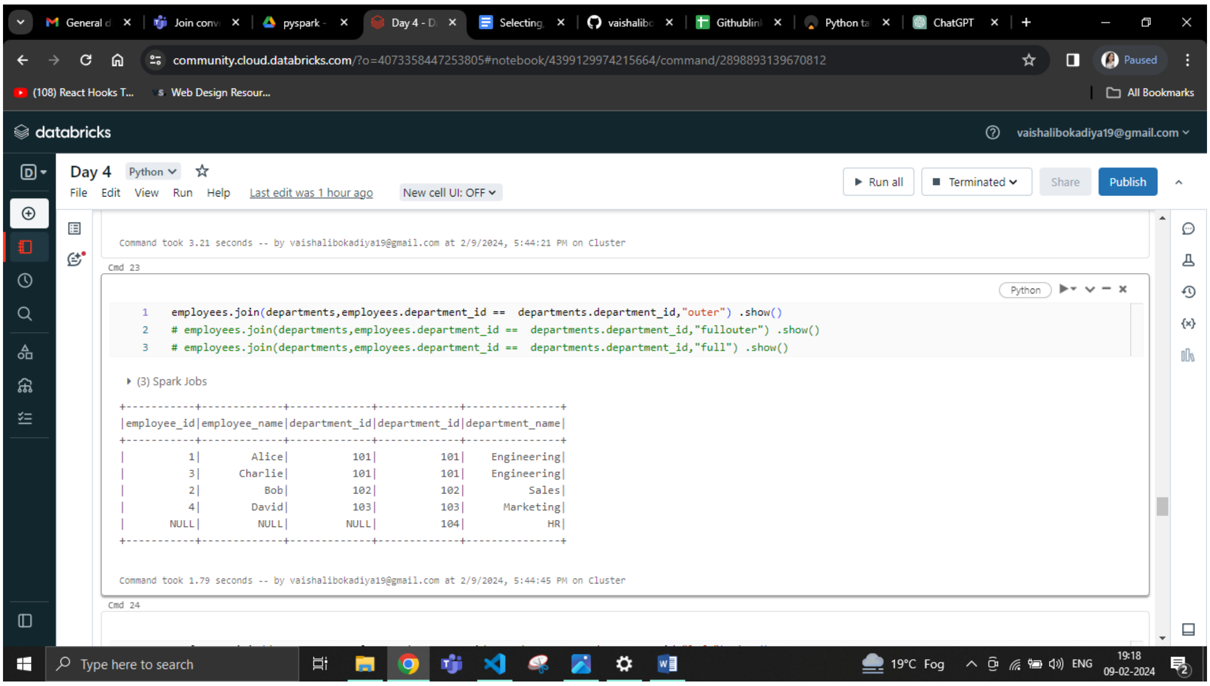
```
employees.join(departments,employees.department_id ==  departments.department_id,"inner") .show()
```

(2) Spark Jobs

```
+-----------+-------------+-------------+-------------+---------------+
|employee_id|employee_name|department_id|department_id|department_name|
+-----------+-------------+-------------+-------------+---------------+
|          1|        Alice|          101|          101|    Engineering|
|          2|          Bob|          102|          102|          Sales|
|          3|      Charlie|          101|          101|    Engineering|
|          4|        David|          103|          103|      Marketing|
+-----------+-------------+-------------+-------------+---------------+

Command took 3.21 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 5:44:21 PM on Cluster
```

Cmd 23

```
employees.join(departments,employees.department_id ==  departments.department_id,"outer") .show()
# employees.join(departments,employees.department_id ==  departments.department_id,"fullouter") .show()
# employees.join(departments,employees.department_id ==  departments.department_id,"full") .show()
```



community.cloud.databricks.com/?o=4073358447253805#notebook/4399129974215664/command/2898893139670812

```
Command took 3.21 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 5:44:21 PM on Cluster
```

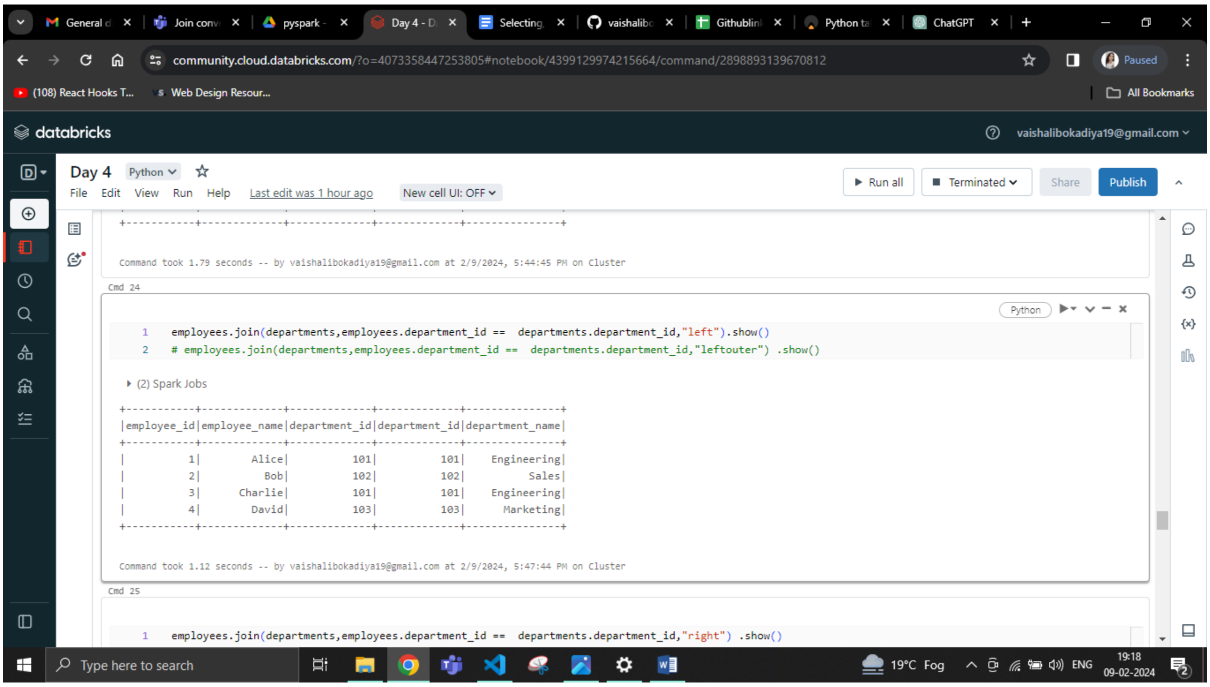
Cmd 23

```python
employees.join(departments,employees.department_id ==  departments.department_id,"outer") .show()
# employees.join(departments,employees.department_id ==  departments.department_id,"fullouter") .show()
# employees.join(departments,employees.department_id ==  departments.department_id,"full") .show()
```

(3) Spark Jobs

```
+-----------+-------------+-------------+-------------+---------------+
|employee_id|employee_name|department_id|department_id|department_name|
+-----------+-------------+-------------+-------------+---------------+
|          1|        Alice|          101|          101|    Engineering|
|          3|      Charlie|          101|          101|    Engineering|
|          2|          Bob|          102|          102|          Sales|
|          4|        David|          103|          103|      Marketing|
|       NULL|         NULL|         NULL|          104|             HR|
+-----------+-------------+-------------+-------------+---------------+


Command took 1.79 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 5:44:45 PM on Cluster
```

Cmd 24



community.cloud.databricks.com/?o=4073358447253805#notebook/4399129974215664/command/2898893139670812

```
Command took 1.79 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 5:44:45 PM on Cluster
```

Cmd 24

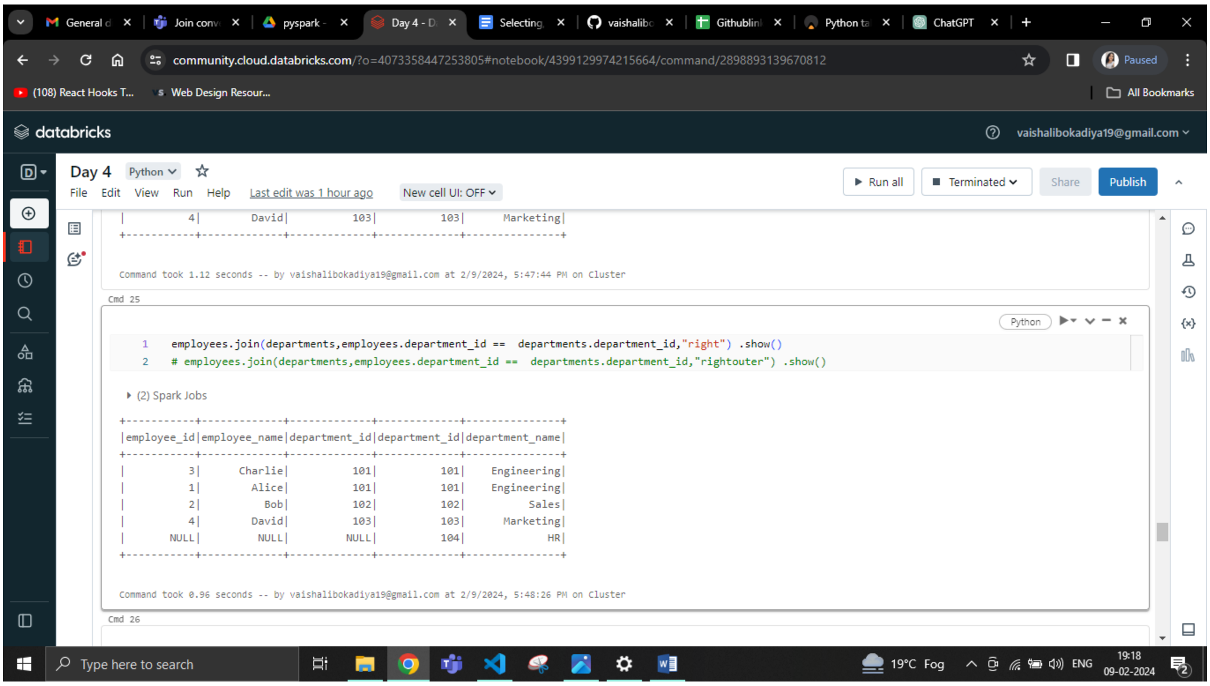
```python
employees.join(departments,employees.department_id ==  departments.department_id,"left").show()
# employees.join(departments,employees.department_id ==  departments.department_id,"leftouter") .show()
```

(2) Spark Jobs

```
+-----------+-------------+-------------+-------------+---------------+
|employee_id|employee_name|department_id|department_id|department_name|
+-----------+-------------+-------------+-------------+---------------+
|          1|        Alice|          101|          101|    Engineering|
|          2|          Bob|          102|          102|          Sales|
|          3|      Charlie|          101|          101|    Engineering|
|          4|        David|          103|          103|      Marketing|
+-----------+-------------+-------------+-------------+---------------+


Command took 1.12 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 5:47:44 PM on Cluster
```

Cmd 25

```python
employees.join(departments,employees.department_id ==  departments.department_id,"right") .show()
```



```
|          4|       David|          103|          103|     Marketing|
+-----------+------------+-------------+-------------+--------------+
```

Command took 1.12 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 5:47:44 PM on Cluster

Cmd 25

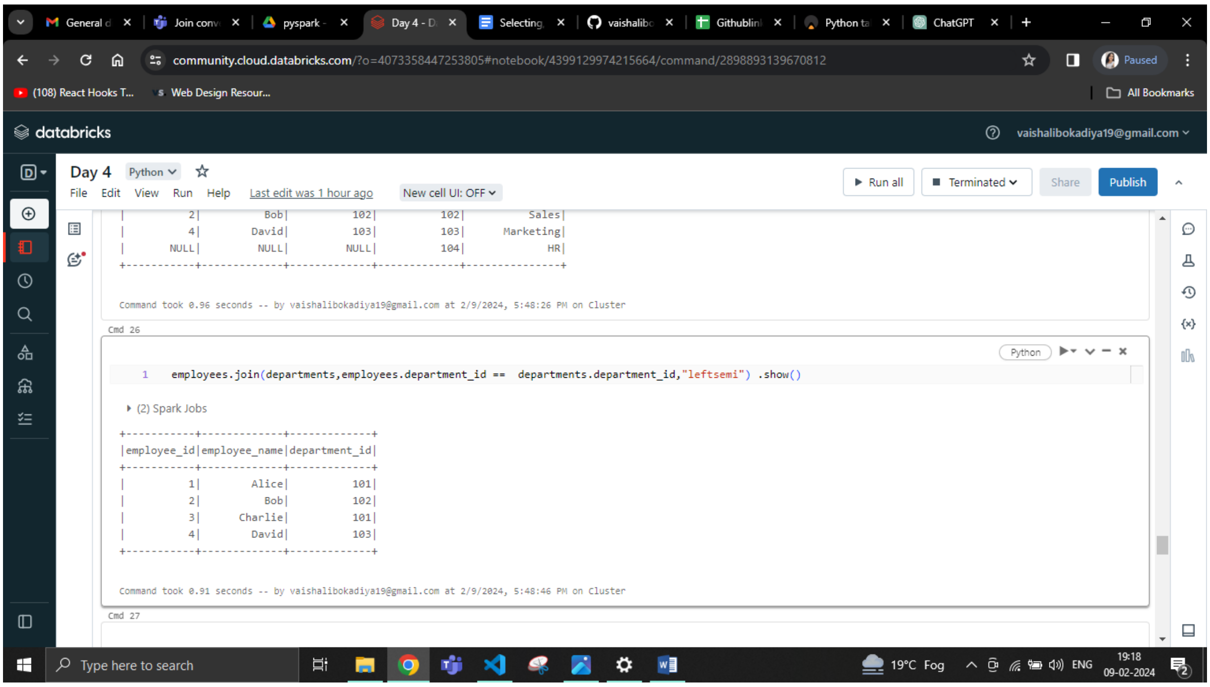
```python
employees.join(departments,employees.department_id ==  departments.department_id,"right") .show()
# employees.join(departments,employees.department_id ==  departments.department_id,"rightouter") .show()
```

(2) Spark Jobs

```
+-----------+-------------+-------------+-------------+---------------+
|employee_id|employee_name|department_id|department_id|department_name|
+-----------+-------------+-------------+-------------+---------------+
|          3|      Charlie|          101|          101|    Engineering|
|          1|        Alice|          101|          101|    Engineering|
|          2|          Bob|          102|          102|          Sales|
|          4|        David|          103|          103|      Marketing|
|       NULL|         NULL|         NULL|          104|             HR|
+-----------+-------------+-------------+-------------+---------------+
```

Command took 0.96 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 5:48:26 PM on Cluster

Cmd 26



```
|          2|          Bob|          102|          102|          Sales|
|          4|        David|          103|          103|      Marketing|
|       NULL|         NULL|         NULL|          104|             HR|
+-----------+-------------+-------------+-------------+---------------+
```

Command took 0.96 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 5:48:26 PM on Cluster

Cmd 26

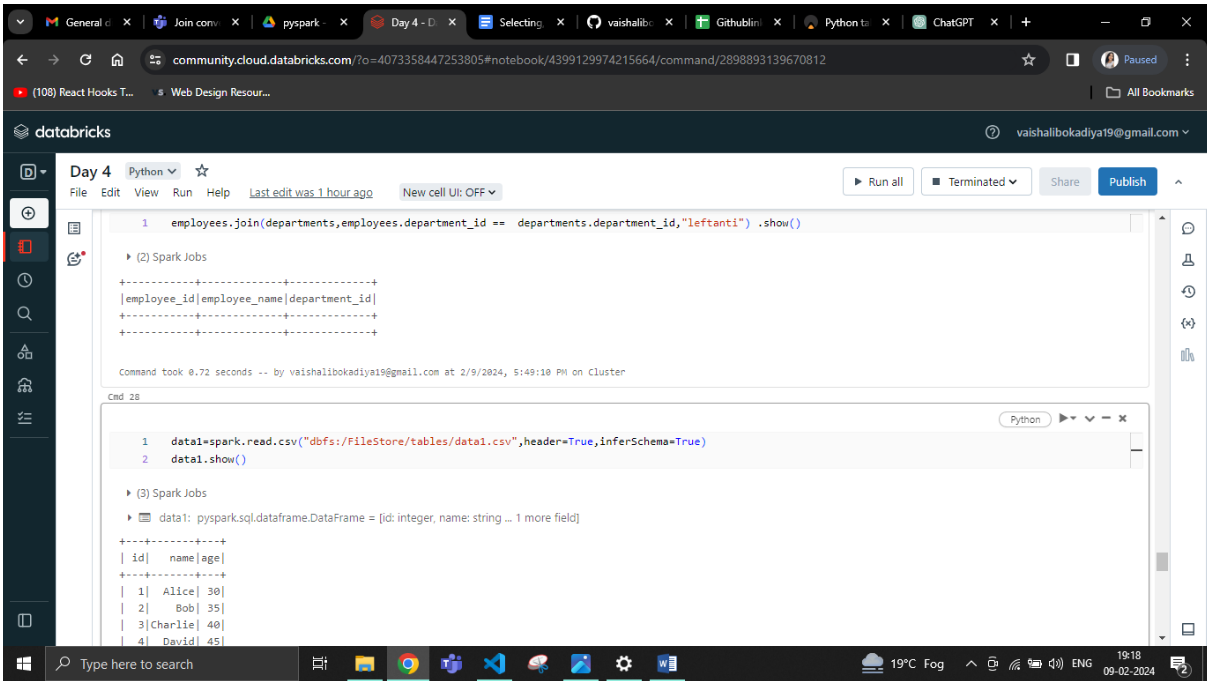
```python
employees.join(departments,employees.department_id ==  departments.department_id,"leftsemi") .show()
```

(2) Spark Jobs

```
+-----------+-------------+-------------+
|employee_id|employee_name|department_id|
+-----------+-------------+-------------+
|          1|        Alice|          101|
|          2|          Bob|          102|
|          3|      Charlie|          101|
|          4|        David|          103|
+-----------+-------------+-------------+
```

Command took 0.91 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 5:48:46 PM on Cluster

Cmd 27



```
employees.join(departments,employees.department_id ==  departments.department_id,"leftanti") .show()
```

(2) Spark Jobs

```
+-----------+-------------+-------------+
|employee_id|employee_name|department_id|
+-----------+-------------+-------------+
+-----------+-------------+-------------+

Command took 0.72 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 5:49:10 PM on Cluster
```

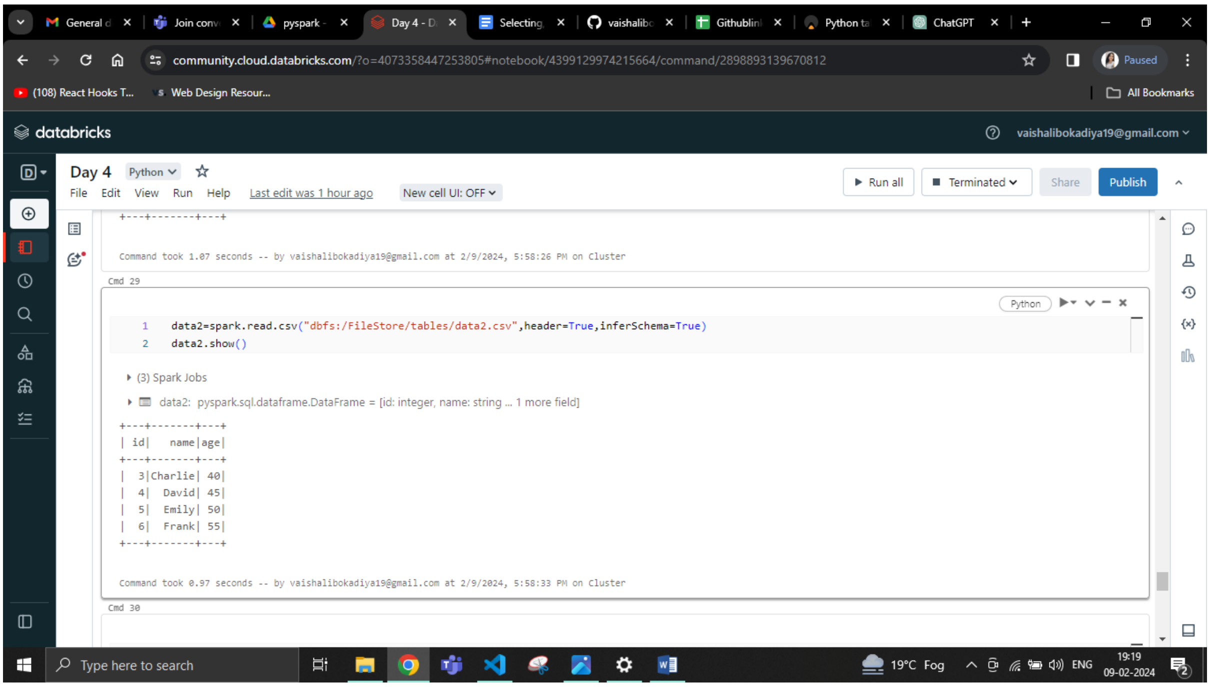
Cmd 28

```
data1=spark.read.csv("dbfs:/FileStore/tables/data1.csv",header=True,inferSchema=True)
data1.show()
```

(3) Spark Jobs

data1: pyspark.sql.dataframe.DataFrame = [id: integer, name: string ... 1 more field]

```
+---+-------+---+
| id|   name|age|
+---+-------+---+
|  1|  Alice| 30|
|  2|    Bob| 35|
|  3|Charlie| 40|
|  4|  David| 45|
```



```
+---+-------+---+

Command took 1.07 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 5:58:26 PM on Cluster
```

Cmd 29

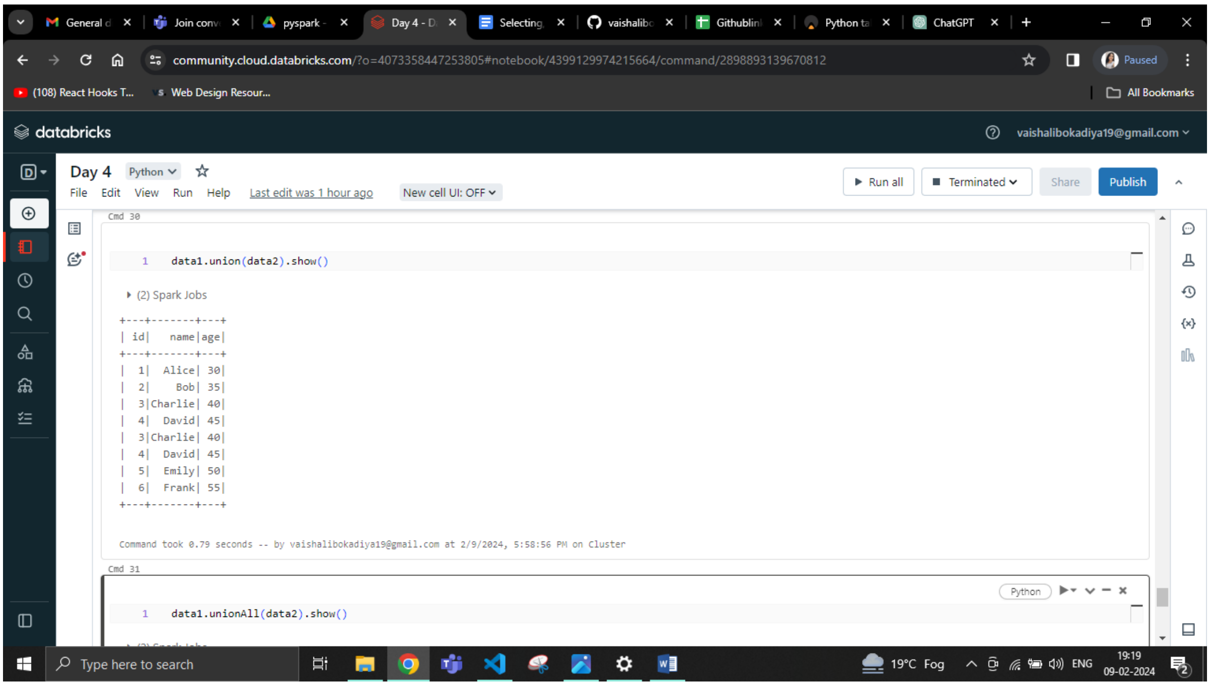
```
data2=spark.read.csv("dbfs:/FileStore/tables/data2.csv",header=True,inferSchema=True)
data2.show()
```

(3) Spark Jobs

data2: pyspark.sql.dataframe.DataFrame = [id: integer, name: string ... 1 more field]

```
+---+-------+---+
| id|   name|age|
+---+-------+---+
|  3|Charlie| 40|
|  4|  David| 45|
|  5|  Emily| 50|
|  6|  Frank| 55|
+---+-------+---+

Command took 0.97 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 5:58:33 PM on Cluster
```

Cmd 30



Cmd 30

```
data1.union(data2).show()
```

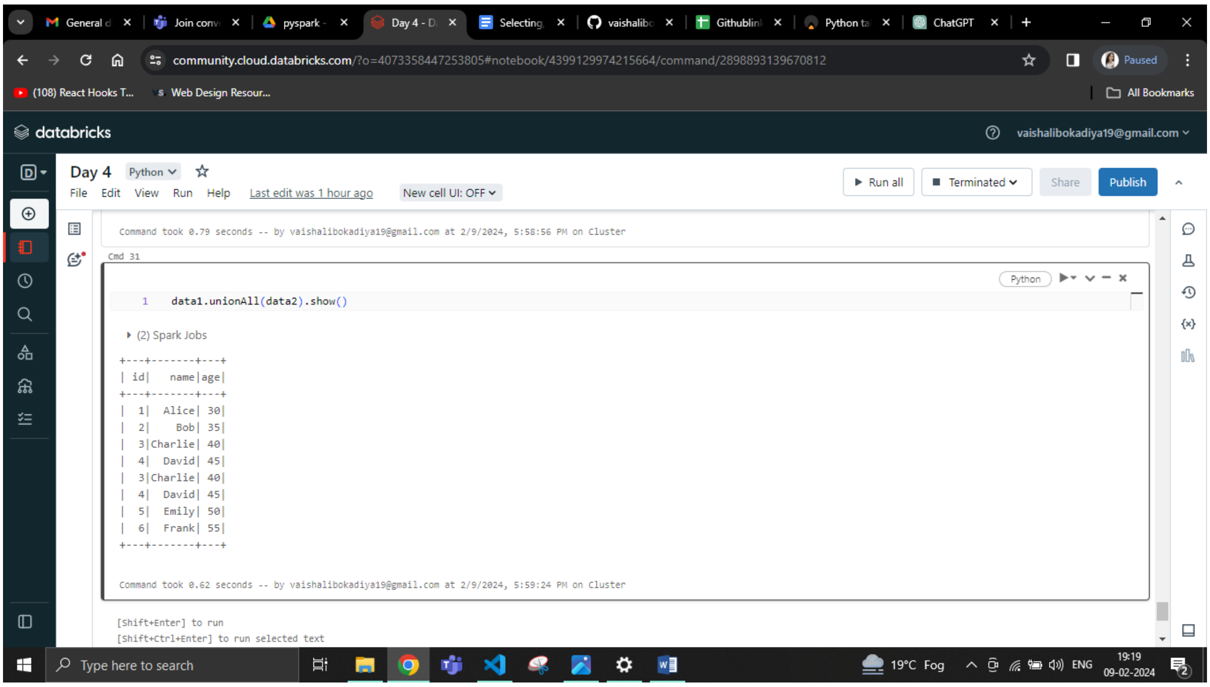
(2) Spark Jobs

```
+---+-------+---+
| id|   name|age|
+---+-------+---+
|  1|  Alice| 30|
|  2|    Bob| 35|
|  3|Charlie| 40|
|  4|  David| 45|
|  3|Charlie| 40|
|  4|  David| 45|
|  5|  Emily| 50|
|  6|  Frank| 55|
+---+-------+---+
```

Command took 0.79 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 5:58:56 PM on Cluster

Cmd 31

```
data1.unionAll(data2).show()
```



Command took 0.79 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 5:58:56 PM on Cluster

Cmd 31

```
data1.unionAll(data2).show()
```

(2) Spark Jobs

```
+---+-------+---+
| id|   name|age|
+---+-------+---+
|  1|  Alice| 30|
|  2|    Bob| 35|
|  3|Charlie| 40|
|  4|  David| 45|
|  3|Charlie| 40|
|  4|  David| 45|
|  5|  Emily| 50|
|  6|  Frank| 55|
+---+-------+---+
```

Command took 0.62 seconds -- by vaishalibokadiya19@gmail.com at 2/9/2024, 5:59:24 PM on Cluster

[Shift+Enter] to run
[Shift+Ctrl+Enter] to run selected text