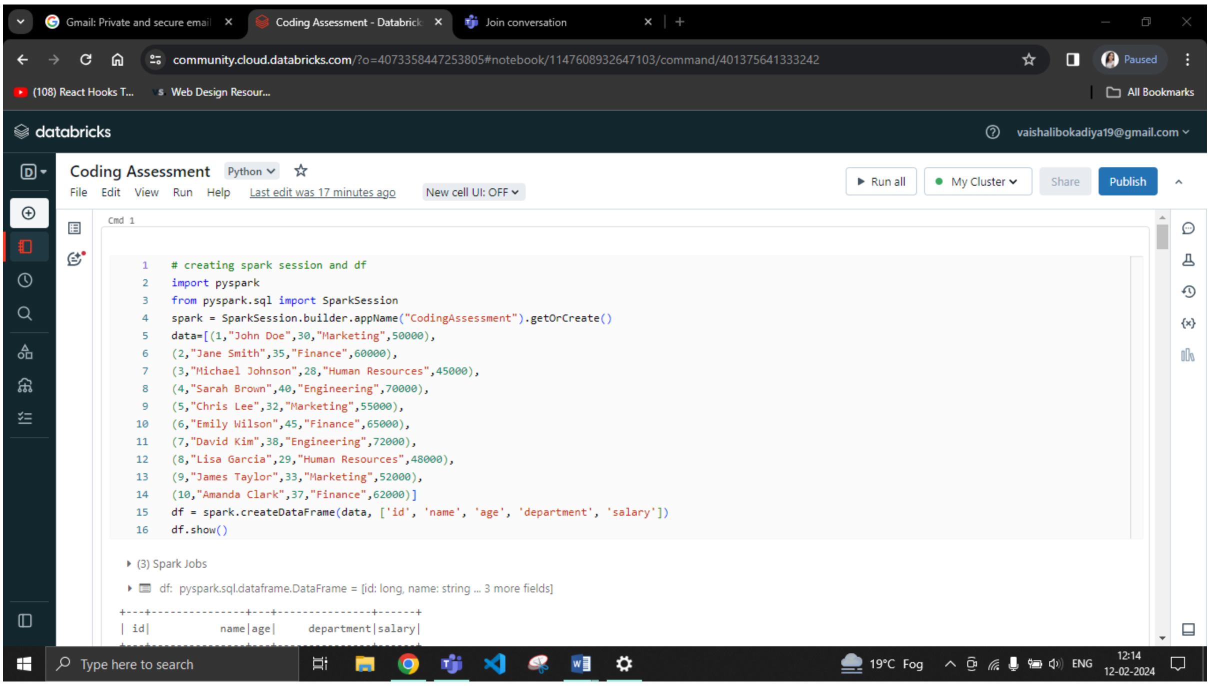
VAISHALI BOKADIYA

CODING ASSESSMENT

QUESTION 1

Execute Manipulating, Droping, Sorting, Aggregations, Joining, GroupBy DataFrames:

Creating session and df:



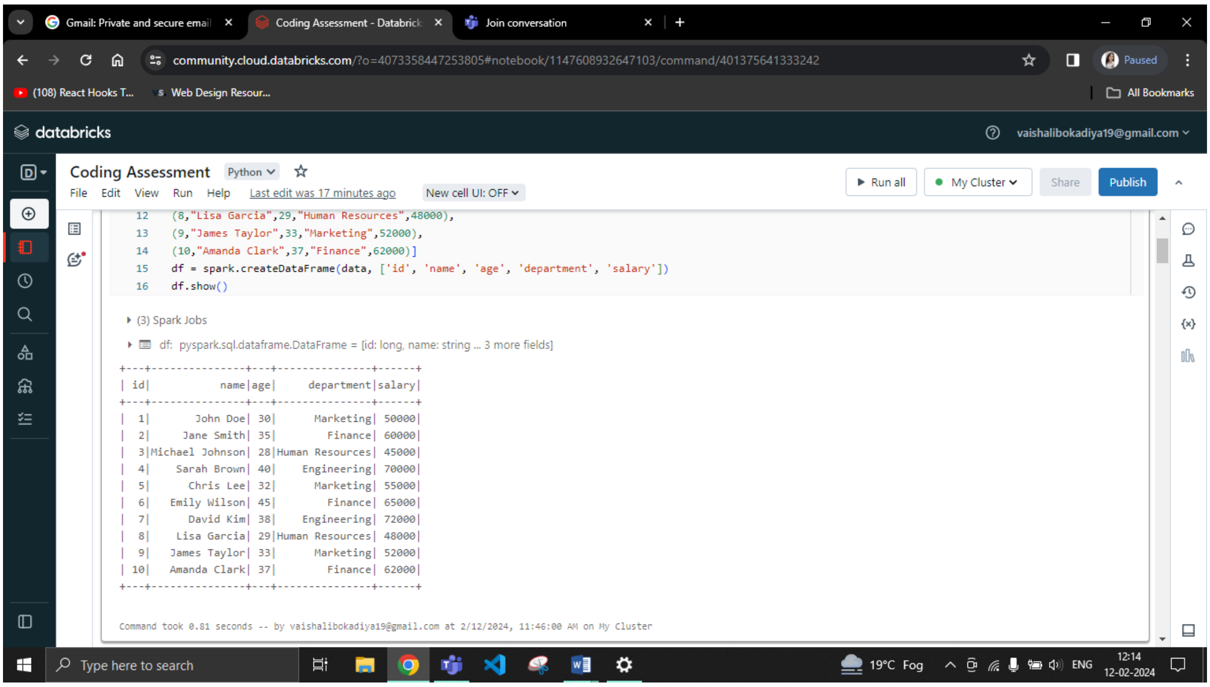
Cmd 1

```
# creating spark session and df
import pyspark
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("CodingAssessment").getOrCreate()
data=[(1,"John Doe",30,"Marketing",50000),
(2,"Jane Smith",35,"Finance",60000),
(3,"Michael Johnson",28,"Human Resources",45000),
(4,"Sarah Brown",40,"Engineering",70000),
(5,"Chris Lee",32,"Marketing",55000),
(6,"Emily Wilson",45,"Finance",65000),
(7,"David Kim",38,"Engineering",72000),
(8,"Lisa Garcia",29,"Human Resources",48000),
(9,"James Taylor",33,"Marketing",52000),
(10,"Amanda Clark",37,"Finance",62000)]
df = spark.createDataFrame(data, ['id', 'name', 'age', 'department', 'salary'])
df.show()
```

(3) Spark Jobs

df: pyspark.sql.dataframe.DataFrame = [id: long, name: string ... 3 more fields]

```
+---+---------------+---+---------------+------+
| id|           name|age|     department|salary|
```



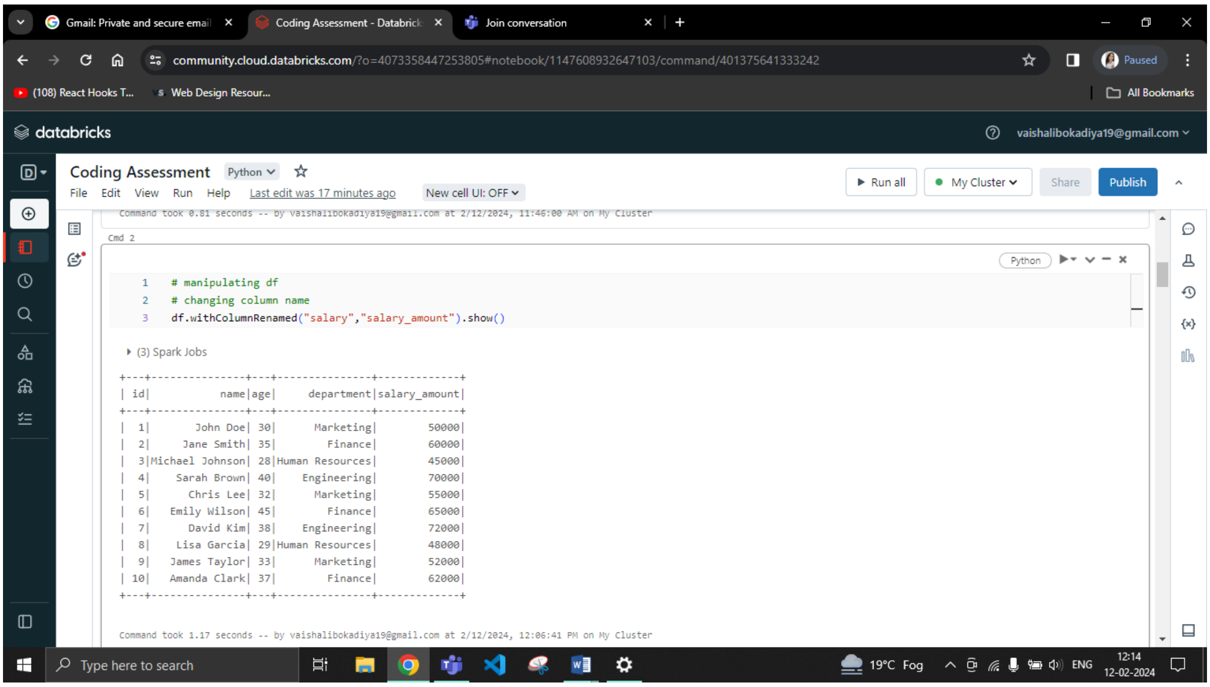
```
(8,"Lisa Garcia",29,"Human Resources",48000),
(9,"James Taylor",33,"Marketing",52000),
(10,"Amanda Clark",37,"Finance",62000)]
df = spark.createDataFrame(data, ['id', 'name', 'age', 'department', 'salary'])
df.show()
```

(3) Spark Jobs

df: pyspark.sql.dataframe.DataFrame = [id: long, name: string ... 3 more fields]

```
+---+---------------+---+---------------+------+
| id|           name|age|     department|salary|
+---+---------------+---+---------------+------+
|  1|       John Doe| 30|      Marketing| 50000|
|  2|     Jane Smith| 35|        Finance| 60000|
|  3|Michael Johnson| 28|Human Resources| 45000|
|  4|    Sarah Brown| 40|    Engineering| 70000|
|  5|      Chris Lee| 32|      Marketing| 55000|
|  6|   Emily Wilson| 45|        Finance| 65000|
|  7|      David Kim| 38|    Engineering| 72000|
|  8|    Lisa Garcia| 29|Human Resources| 48000|
|  9|   James Taylor| 33|      Marketing| 52000|
| 10|   Amanda Clark| 37|        Finance| 62000|
+---+---------------+---+---------------+------+

Command took 0.81 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 11:46:00 AM on My Cluster
```

Manipulating:

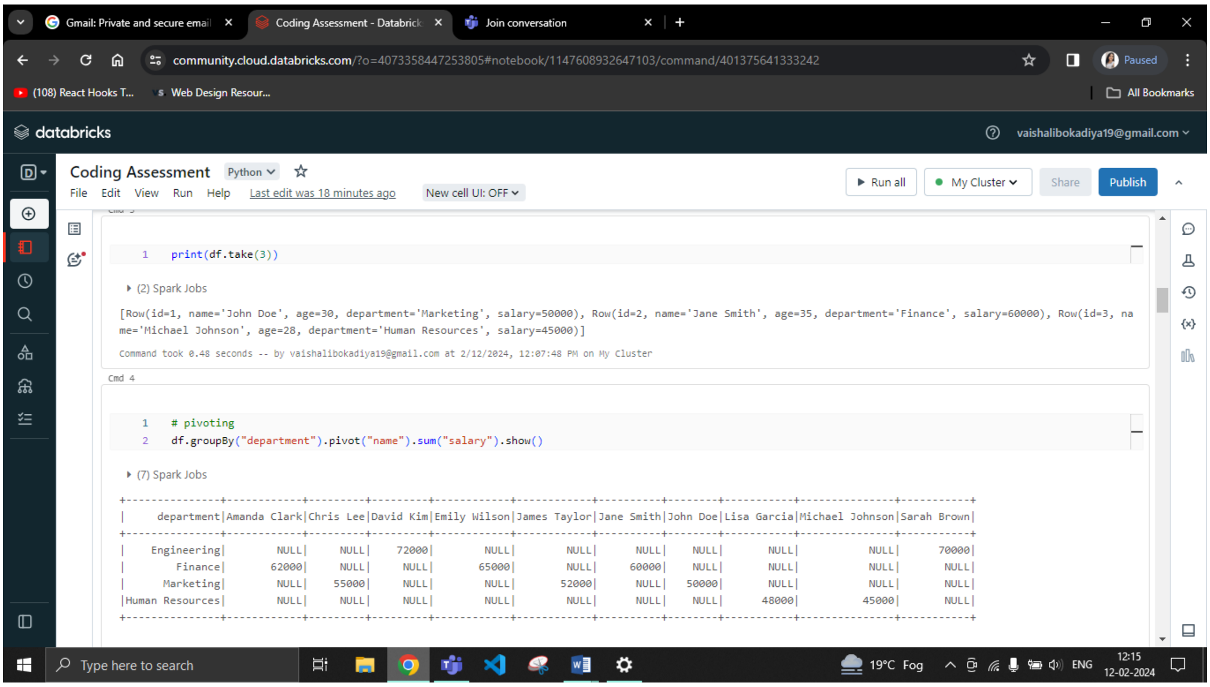


```
Command took 0.81 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 11:46:08 AM on My Cluster
```

```python
# manipulating df
# changing column name
df.withColumnRenamed("salary","salary_amount").show()
```

(3) Spark Jobs

```
+---+---------------+---+---------------+-------------+
| id|           name|age|     department|salary_amount|
+---+---------------+---+---------------+-------------+
|  1|       John Doe| 30|      Marketing|        50000|
|  2|     Jane Smith| 35|        Finance|        60000|
|  3|Michael Johnson| 28|Human Resources|        45000|
|  4|    Sarah Brown| 40|    Engineering|        70000|
|  5|      Chris Lee| 32|      Marketing|        55000|
|  6|   Emily Wilson| 45|        Finance|        65000|
|  7|      David Kim| 38|    Engineering|        72000|
|  8|    Lisa Garcia| 29|Human Resources|        48000|
|  9|   James Taylor| 33|      Marketing|        52000|
| 10|   Amanda Clark| 37|        Finance|        62000|
+---+---------------+---+---------------+-------------+

Command took 1.17 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:06:41 PM on My Cluster
```



```python
print(df.take(3))
```

(2) Spark Jobs

```
[Row(id=1, name='John Doe', age=30, department='Marketing', salary=50000), Row(id=2, name='Jane Smith', age=35, department='Finance', salary=60000), Row(id=3, na
me='Michael Johnson', age=28, department='Human Resources', salary=45000)]

Command took 0.48 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:07:48 PM on My Cluster
```

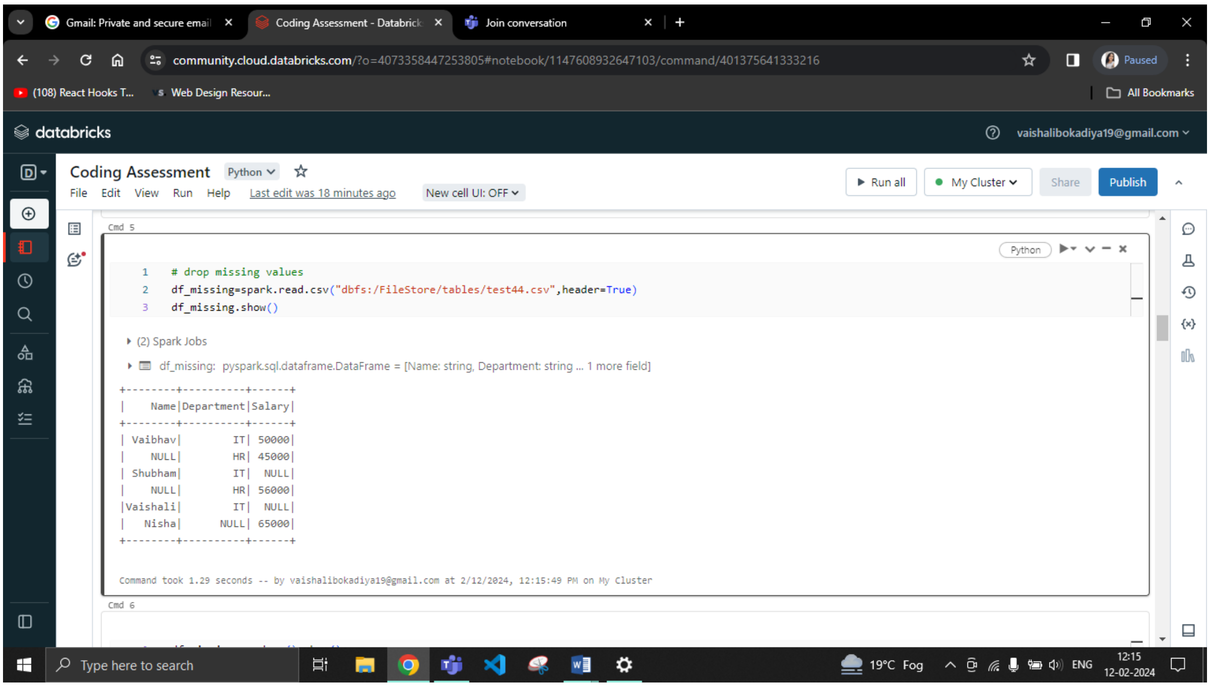
Cmd 4

```python
# pivoting
df.groupBy("department").pivot("name").sum("salary").show()
```

(7) Spark Jobs

```
+---------------+------------+---------+---------+------------+------------+----------+--------+-----------+---------------+-----------+
|     department|Amanda Clark|Chris Lee|David Kim|Emily Wilson|James Taylor|Jane Smith|John Doe|Lisa Garcia|Michael Johnson|Sarah Brown|
+---------------+------------+---------+---------+------------+------------+----------+--------+-----------+---------------+-----------+
|    Engineering|        NULL|     NULL|    72000|        NULL|        NULL|      NULL|    NULL|       NULL|           NULL|      70000|
|        Finance|       62000|     NULL|     NULL|       65000|        NULL|     60000|    NULL|       NULL|           NULL|       NULL|
|      Marketing|        NULL|    55000|     NULL|        NULL|       52000|      NULL|   50000|       NULL|           NULL|       NULL|
|Human Resources|        NULL|     NULL|     NULL|        NULL|        NULL|      NULL|    NULL|      48000|          45000|       NULL|
+---------------+------------+---------+---------+------------+------------+----------+--------+-----------+---------------+-----------+
```

Dropping missing values:

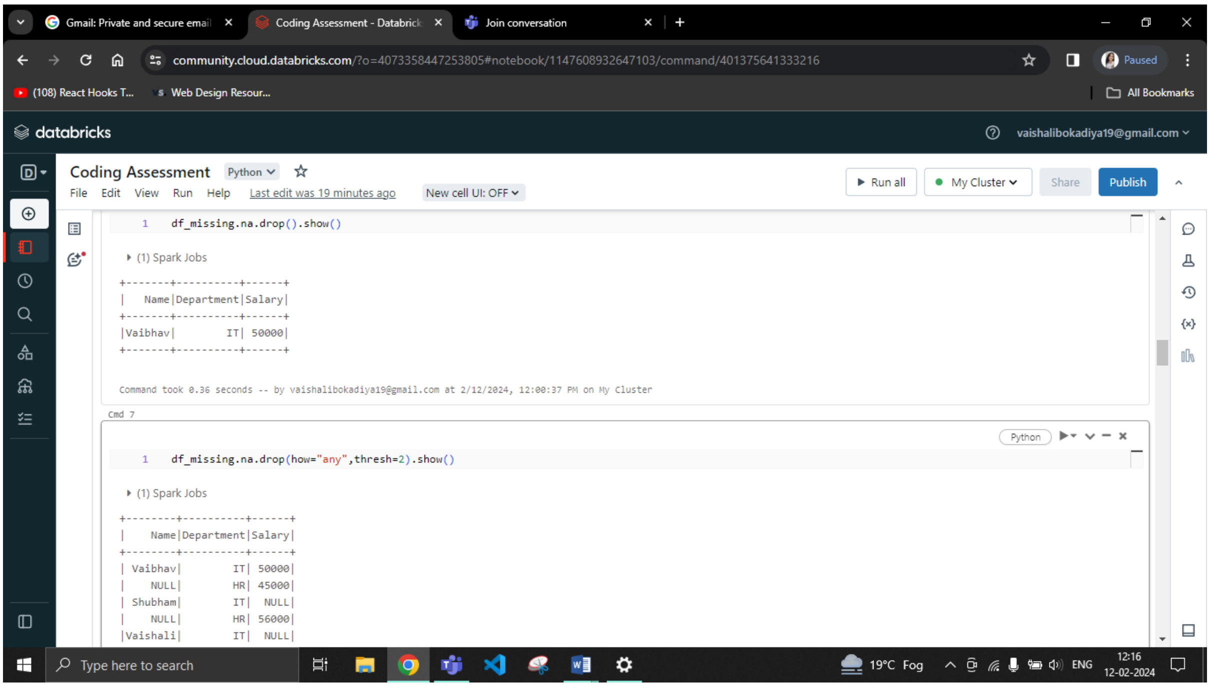


```
# drop missing values
df_missing=spark.read.csv("dbfs:/FileStore/tables/test44.csv",header=True)
df_missing.show()
```

(2) Spark Jobs

df_missing: 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 1.29 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:15:49 PM on My Cluster
```



```
df_missing.na.drop().show()
```

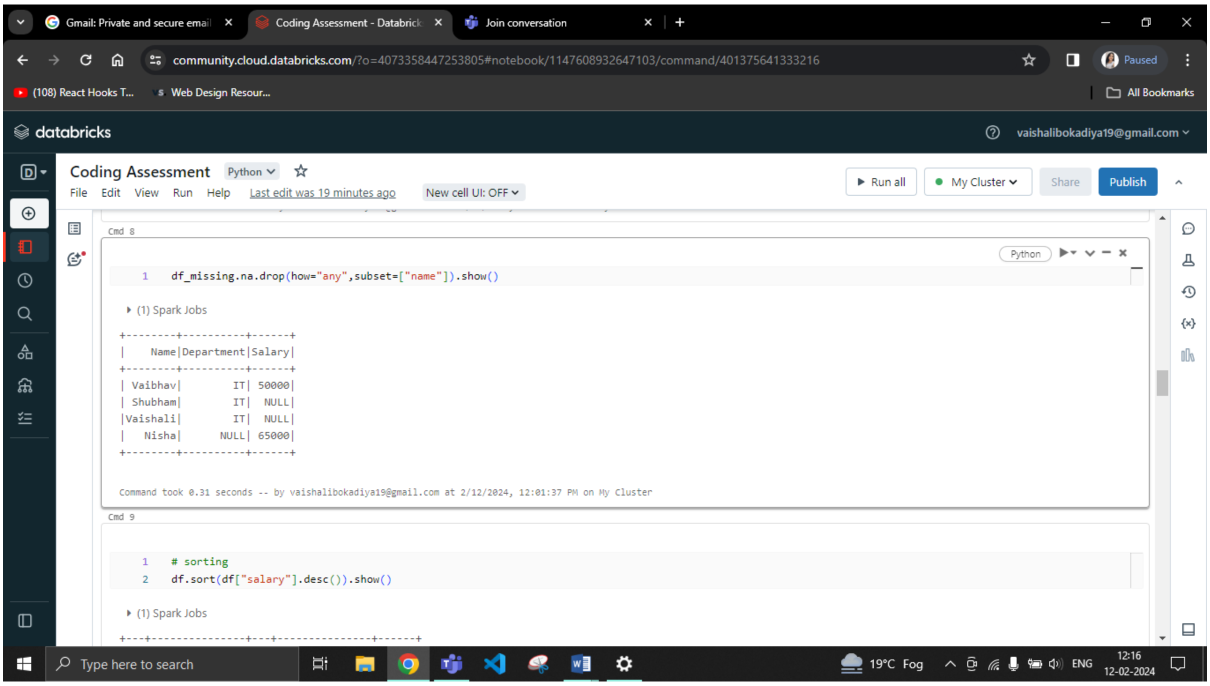
(1) Spark Jobs

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

Command took 0.36 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:00:37 PM on My Cluster
```

```
df_missing.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|
```



Cmd 8

```
df_missing.na.drop(how="any",subset=["name"]).show()
```

(1) Spark Jobs

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

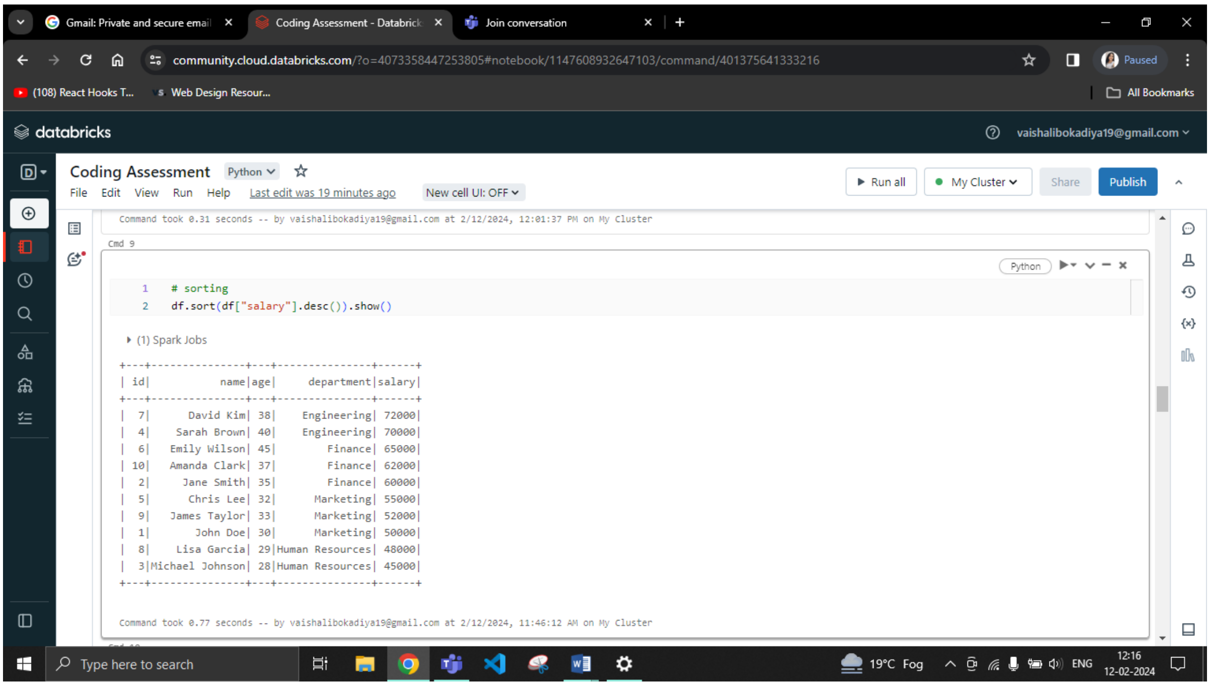
Command took 0.31 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:01:37 PM on My Cluster

Cmd 9

```
# sorting
df.sort(df["salary"].desc()).show()
```

(1) Spark Jobs

Sorting:



Command took 0.31 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:01:37 PM on My Cluster

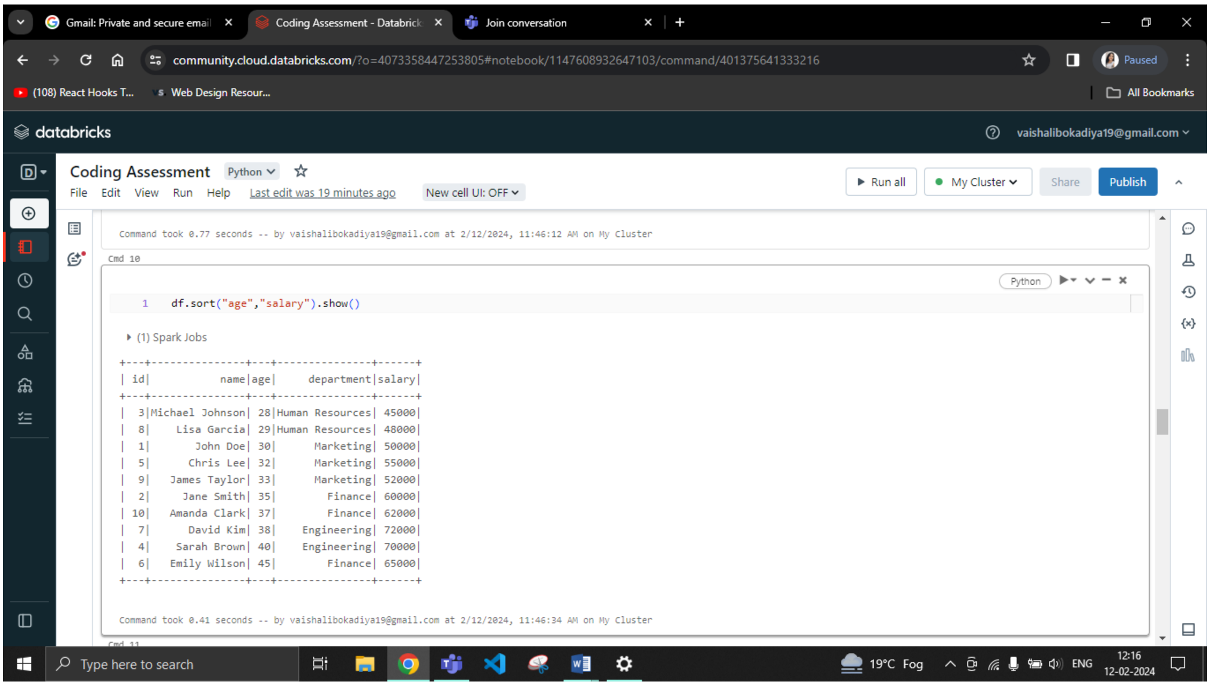
Cmd 9

```
# sorting
df.sort(df["salary"].desc()).show()
```

(1) Spark Jobs

```
+---+--------------+---+---------------+------+
| id|          name|age|     department|salary|
+---+--------------+---+---------------+------+
|  7|      David Kim| 38|    Engineering| 72000|
|  4|    Sarah Brown| 40|    Engineering| 70000|
|  6|   Emily Wilson| 45|        Finance| 65000|
| 10|   Amanda Clark| 37|        Finance| 62000|
|  2|     Jane Smith| 35|        Finance| 60000|
|  5|      Chris Lee| 32|      Marketing| 55000|
|  9|   James Taylor| 33|      Marketing| 52000|
|  1|       John Doe| 30|      Marketing| 50000|
|  8|    Lisa Garcia| 29|Human Resources| 48000|
|  3|Michael Johnson| 28|Human Resources| 45000|
+---+--------------+---+---------------+------+
```

Command took 0.77 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 11:46:12 AM on My Cluster



```
Command took 0.77 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 11:46:12 AM on My Cluster
```

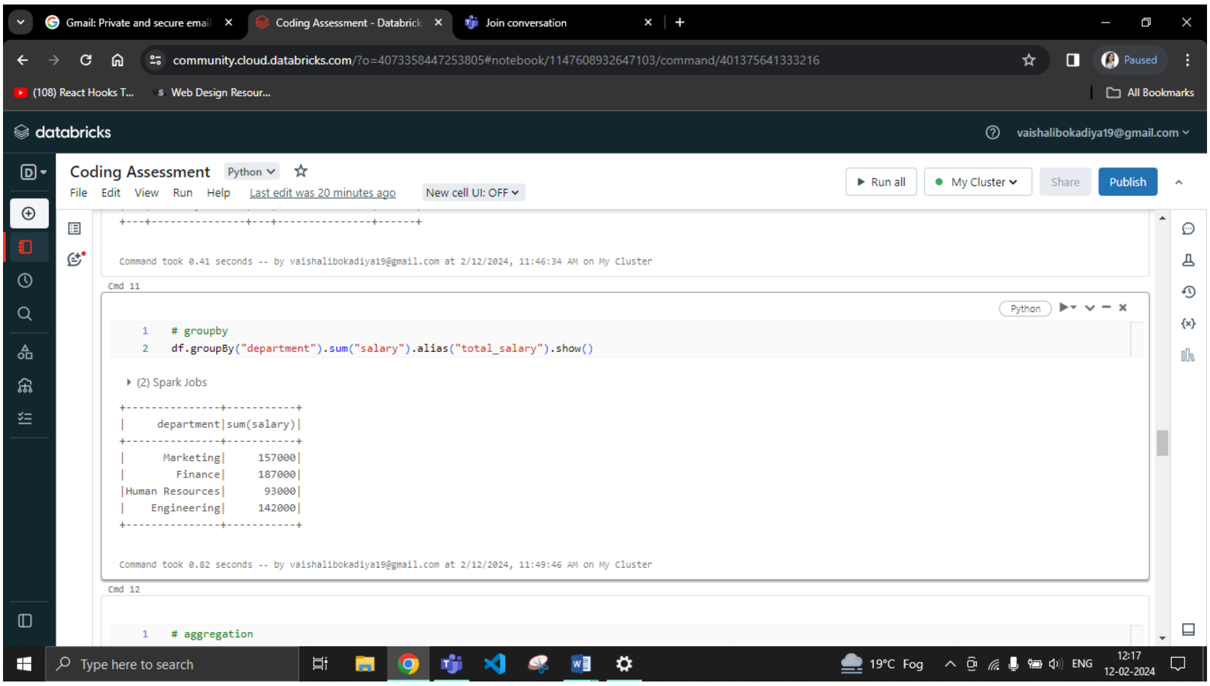
Cmd 10

```python
df.sort("age","salary").show()
```

(1) Spark Jobs

```
+---+---------------+---+---------------+------+
| id|           name|age|     department|salary|
+---+---------------+---+---------------+------+
|  3|Michael Johnson| 28|Human Resources| 45000|
|  8|    Lisa Garcia| 29|Human Resources| 48000|
|  1|       John Doe| 30|      Marketing| 50000|
|  5|      Chris Lee| 32|      Marketing| 55000|
|  9|   James Taylor| 33|      Marketing| 52000|
|  2|     Jane Smith| 35|        Finance| 60000|
| 10|   Amanda Clark| 37|        Finance| 62000|
|  7|      David Kim| 38|    Engineering| 72000|
|  4|    Sarah Brown| 40|    Engineering| 70000|
|  6|   Emily Wilson| 45|        Finance| 65000|
+---+---------------+---+---------------+------+

Command took 0.41 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 11:46:34 AM on My Cluster
```

GroupBy



```
Command took 0.41 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 11:46:34 AM on My Cluster
```

Cmd 11

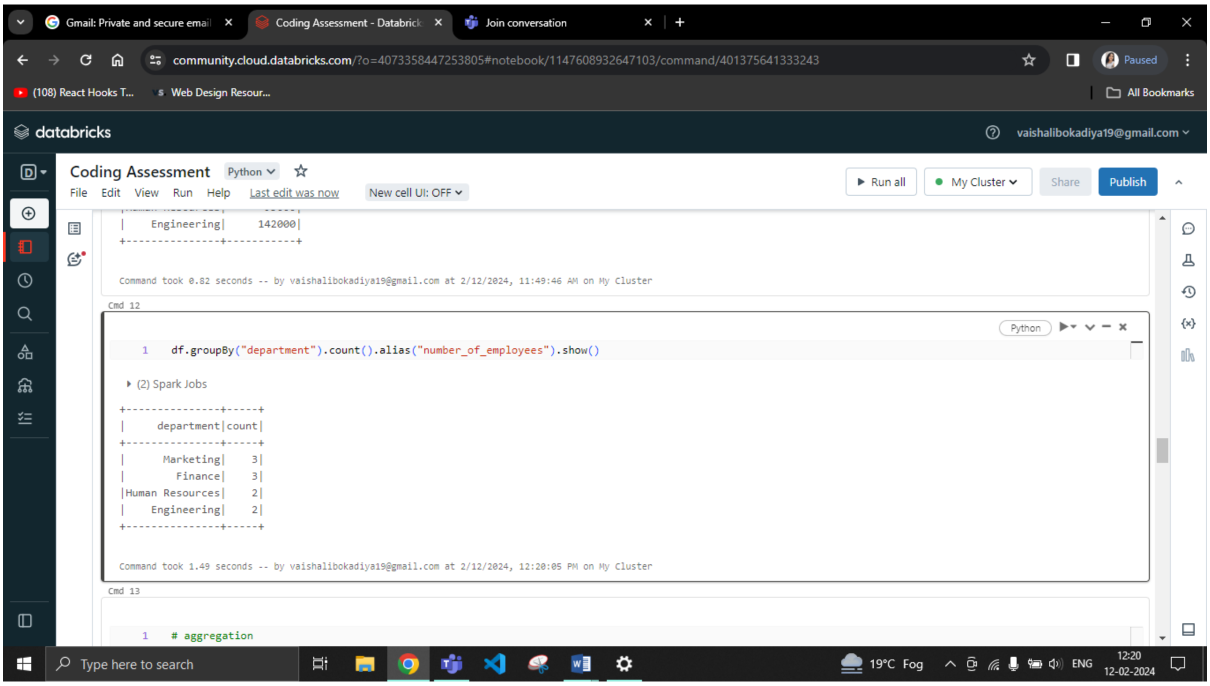
```python
# groupby
df.groupBy("department").sum("salary").alias("total_salary").show()
```

(2) Spark Jobs

```
+---------------+-----------+
|     department|sum(salary)|
+---------------+-----------+
|      Marketing|     157000|
|        Finance|     187000|
|Human Resources|      93000|
|    Engineering|     142000|
+---------------+-----------+

Command took 0.82 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 11:49:46 AM on My Cluster
```

Cmd 12

```python
# aggregation
```



```
|    Engineering|      142000|
+---------------+------------+

Command took 0.82 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 11:49:46 AM on My Cluster
```

Cmd 12

```
df.groupBy("department").count().alias("number_of_employees").show()
```

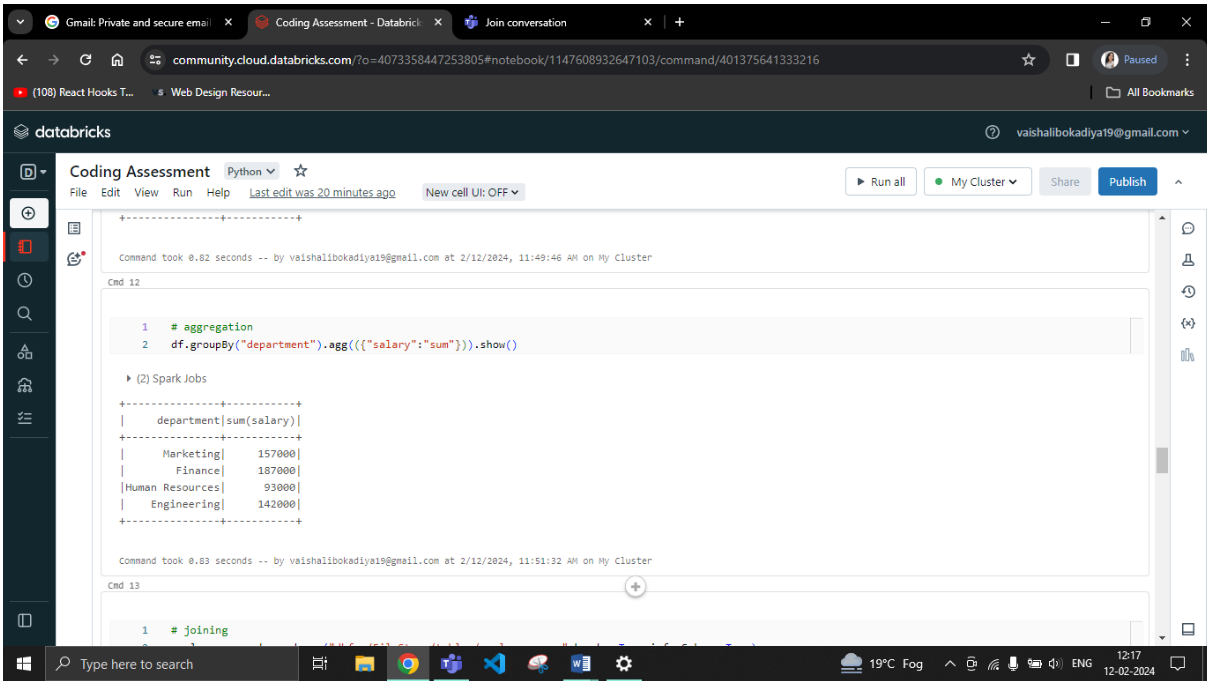
(2) Spark Jobs

```
+---------------+-----+
|     department|count|
+---------------+-----+
|      Marketing|    3|
|        Finance|    3|
|Human Resources|    2|
|    Engineering|    2|
+---------------+-----+

Command took 1.49 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:20:05 PM on My Cluster
```

Cmd 13

```
# aggregation
```

Aggregation:



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

Command took 0.82 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 11:49:46 AM on My Cluster
```

Cmd 12

```
# aggregation
df.groupBy("department").agg(({"salary":"sum"})).show()
```

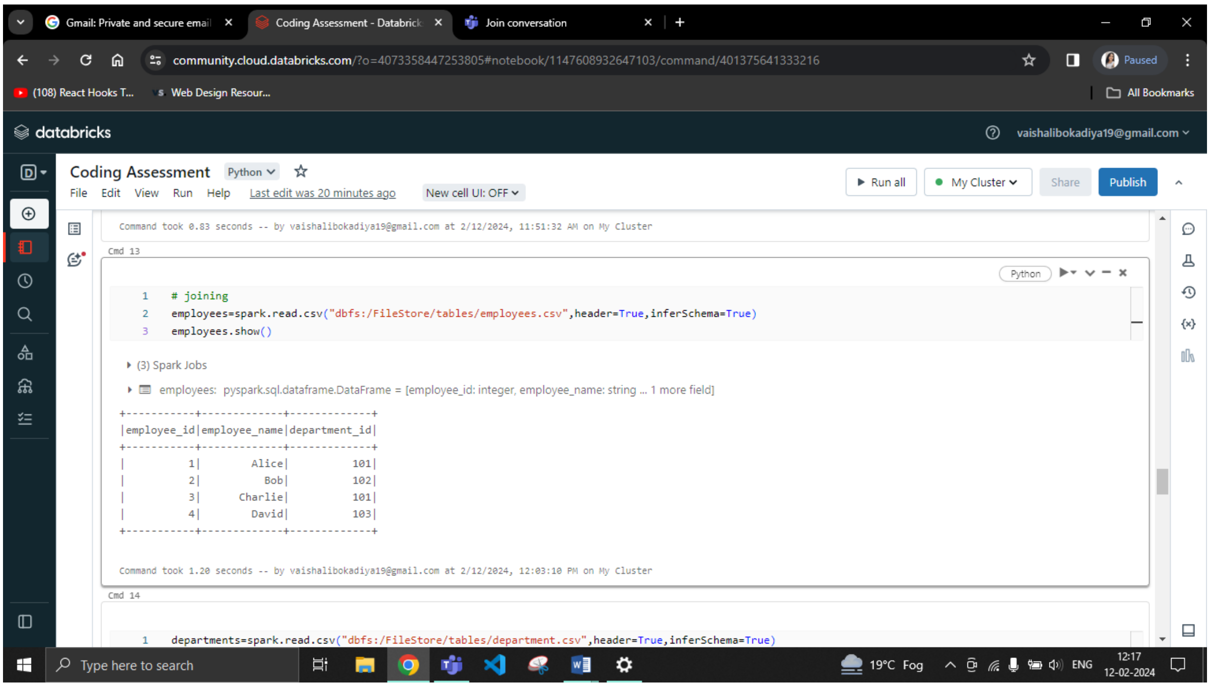
(2) Spark Jobs

```
+---------------+-----------+
|     department|sum(salary)|
+---------------+-----------+
|      Marketing|     157000|
|        Finance|     187000|
|Human Resources|      93000|
|    Engineering|     142000|
+---------------+-----------+

Command took 0.83 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 11:51:32 AM on My Cluster
```

Cmd 13

```
# joining
```

Joins



community.cloud.databricks.com/?o=4073358447253805#notebook/1147608932647103/command/401375641333216
databricks
vaishalibokadiya19@gmail.com
Coding Assessment
Python
File Edit View Run Help
Last edit was 20 minutes ago
New cell UI: OFF
Run all
My Cluster
Share
Publish

```
Command took 0.83 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 11:51:32 AM on My Cluster
```

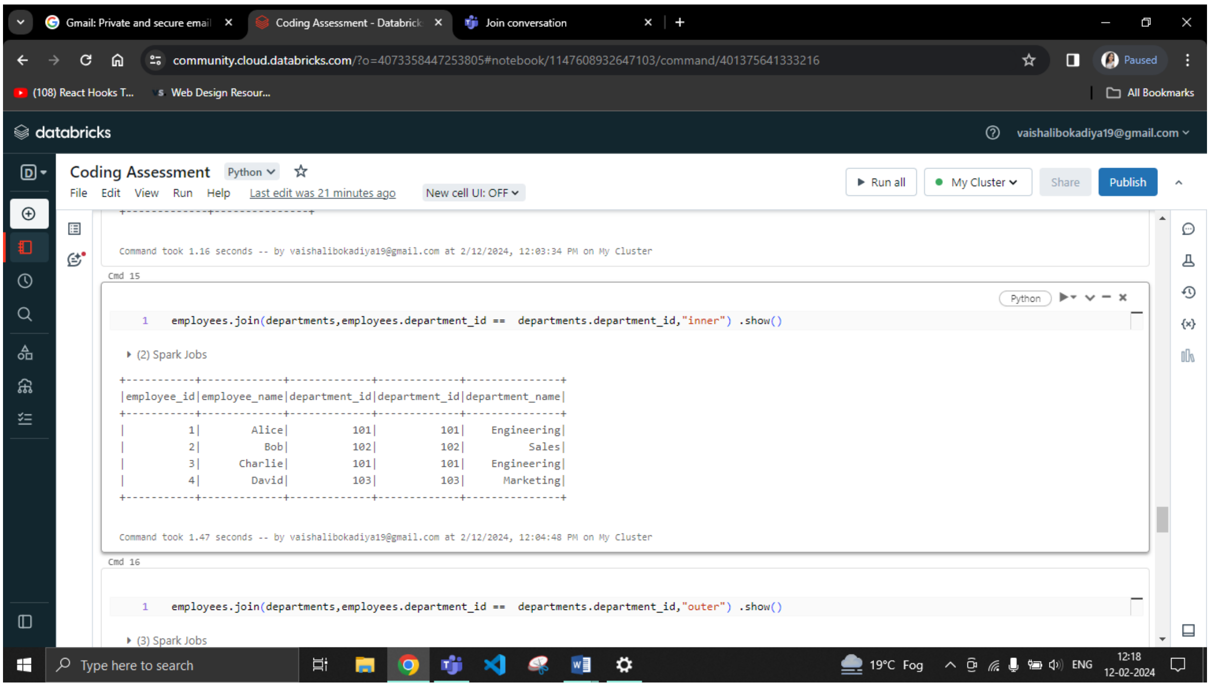
Cmd 13

```
# joining
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 1.20 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:03:10 PM on My Cluster
```

Cmd 14

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



community.cloud.databricks.com/?o=4073358447253805#notebook/1147608932647103/command/401375641333216
databricks
vaishalibokadiya19@gmail.com
Coding Assessment
Python
File Edit View Run Help
Last edit was 21 minutes ago
New cell UI: OFF
Run all
My Cluster
Share
Publish

```
Command took 1.16 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:03:34 PM on My Cluster
```

Cmd 15

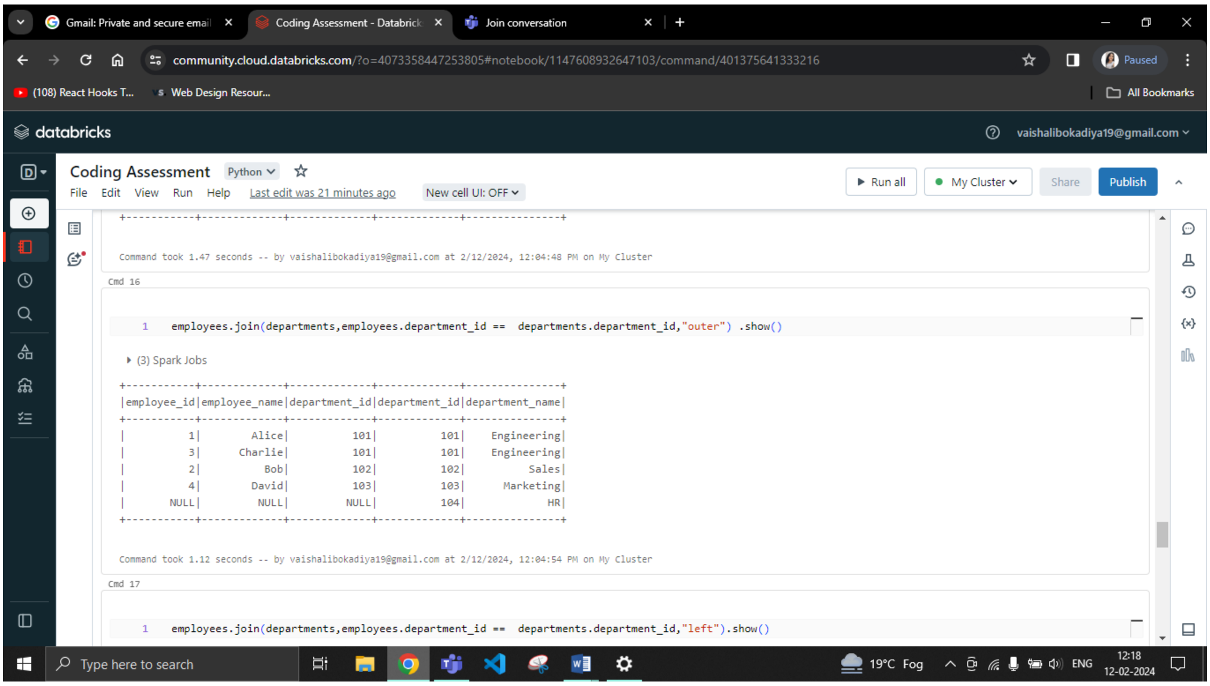
```
employees.join(departments,employees.department_id ==  departments.department_id,"inner") .show()
```

```
(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 1.47 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:04:48 PM on My Cluster
```

Cmd 16

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

```
(3) Spark Jobs
```



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

```
Command took 1.47 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:04:48 PM on My Cluster
```

Cmd 16

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

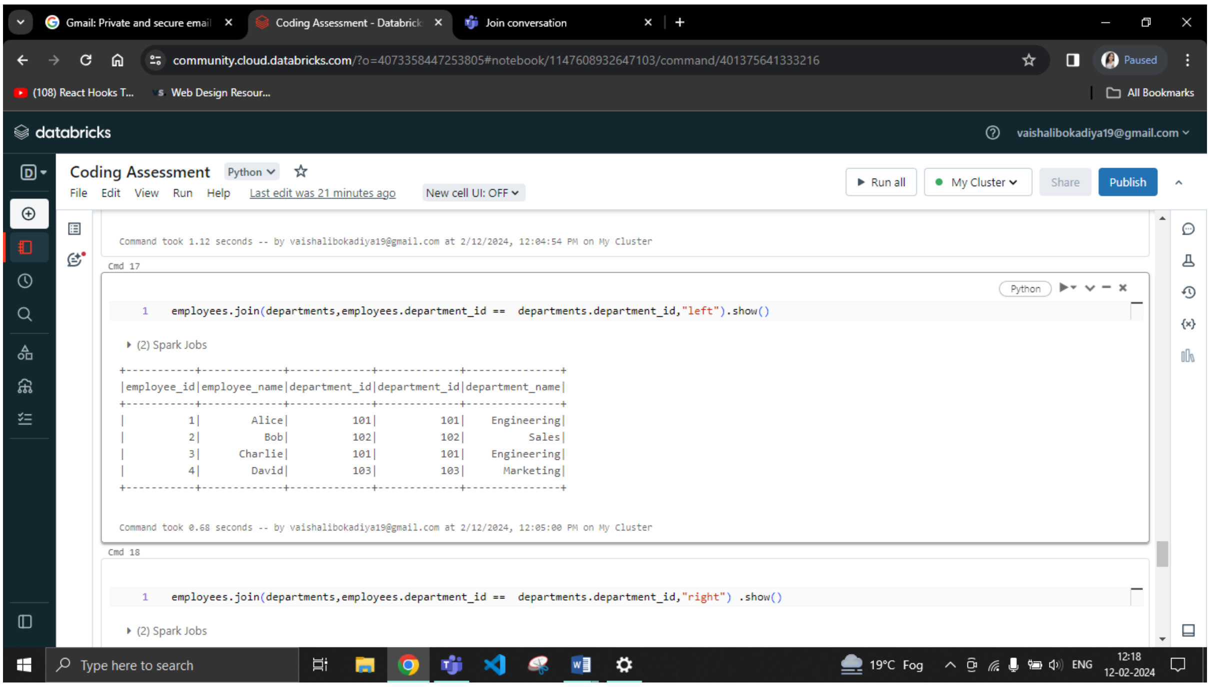
▸ (3) Spark Jobs

```
+-----------+-------------+-------------+-------------+---------------+
|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 1.12 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:04:54 PM on My Cluster
```

Cmd 17

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



```
Command took 1.12 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:04:54 PM on My Cluster
```

Cmd 17

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

▸ (2) Spark Jobs

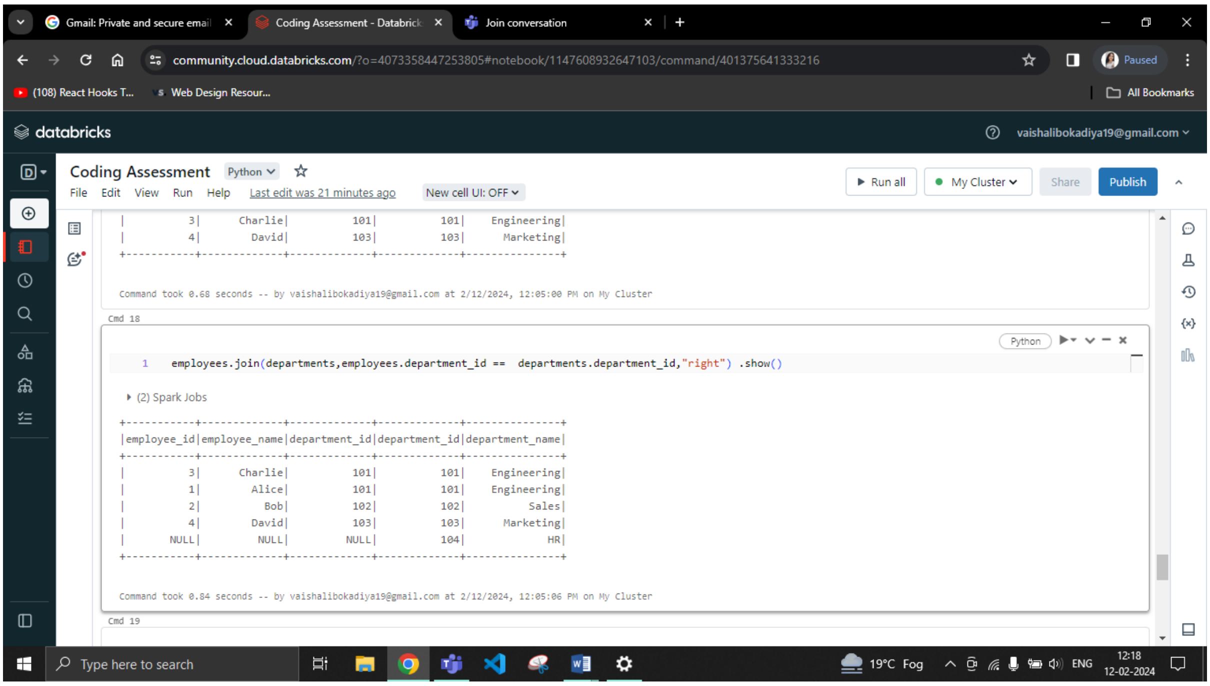
```
+-----------+-------------+-------------+-------------+---------------+
|employee_id|employee_name|department_id|department_id|department_name|
+-----------+-------------+-------------+-------------+---------------+
|          1|        Alice|          101|          101|    Engineering|
|       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

```
Command took 0.68 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:05:00 PM on My Cluster
```

Cmd 18

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

▸ (2) Spark Jobs



```
|          3|     Charlie|          101|          101|   Engineering|
|          4|       David|          103|          103|     Marketing|
+-----------+------------+-------------+-------------+--------------+

Command took 0.68 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:05:00 PM on My Cluster
```

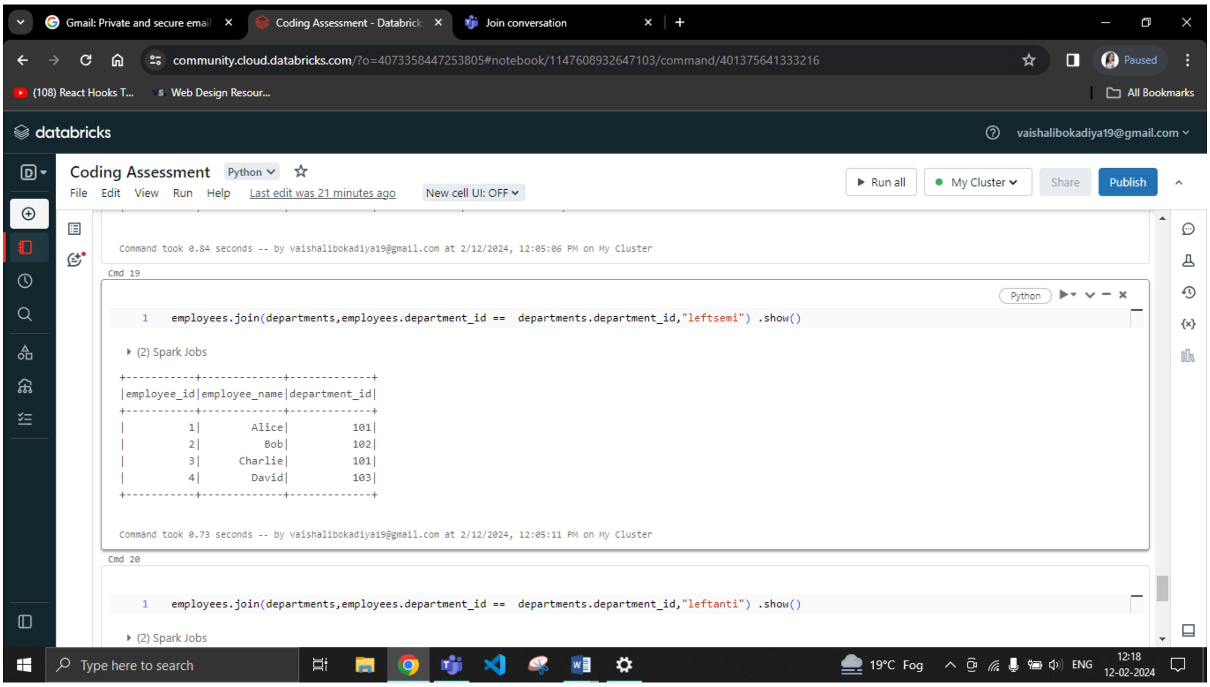
Cmd 18

```
employees.join(departments,employees.department_id ==  departments.department_id,"right") .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 0.84 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:05:06 PM on My Cluster
```

Cmd 19



```
Command took 0.84 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:05:06 PM on My Cluster
```

Cmd 19

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

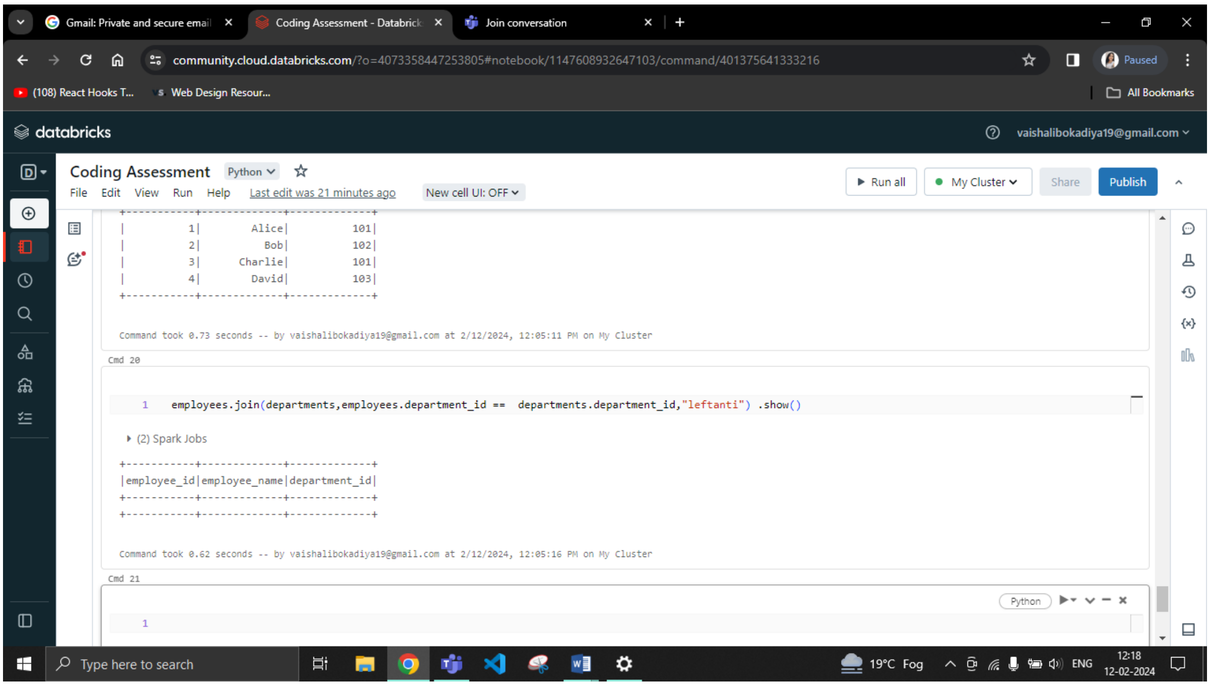
(2) Spark Jobs

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

Command took 0.73 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:05:11 PM on My Cluster
```

Cmd 20

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

(2) Spark Jobs



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

Command took 0.73 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:05:11 PM on My Cluster

Cmd 20

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

(2) Spark Jobs

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

Command took 0.62 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:05:16 PM on My Cluster

Cmd 21