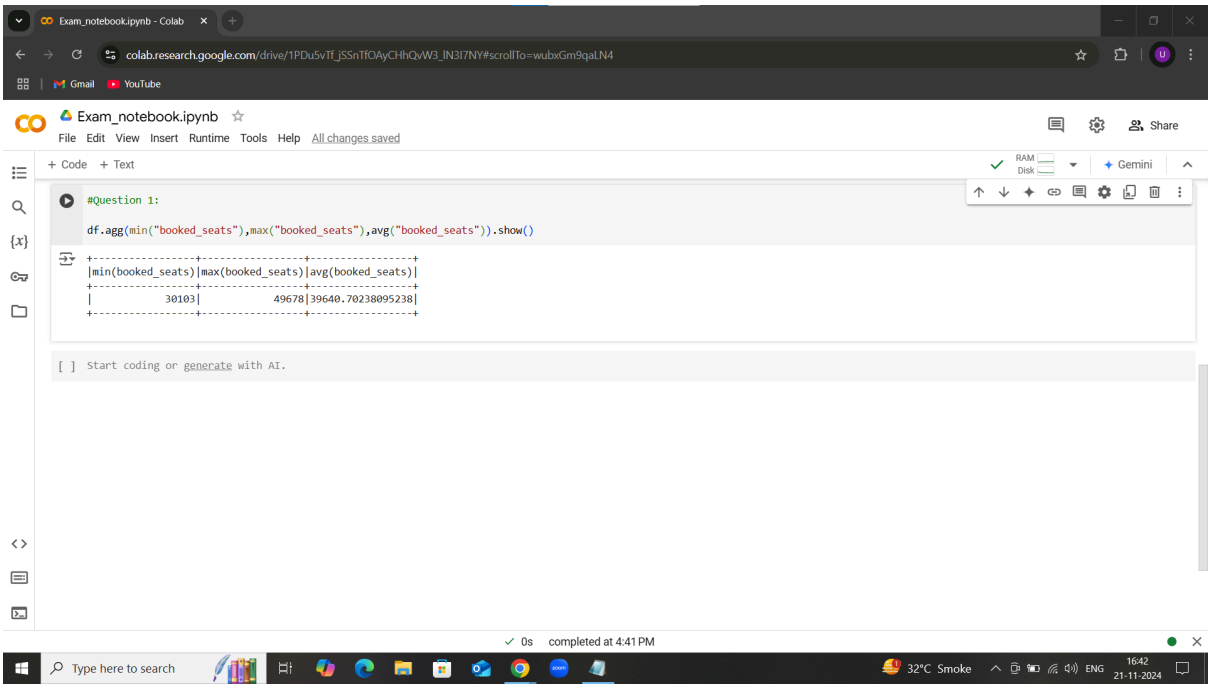
SPARK

USE RDD OR DATAFRAME:

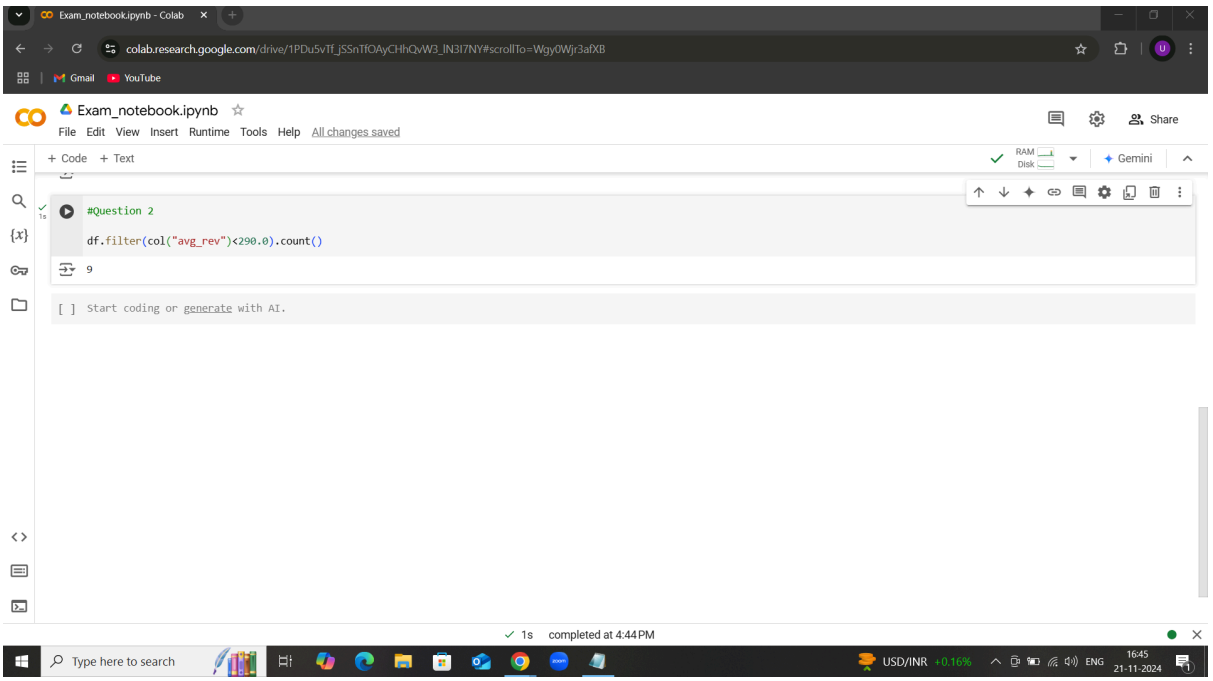
1:

```
df.agg(min("booked_seats"),max("booked_seats"),avg("booked_seats")).sho
w()
```



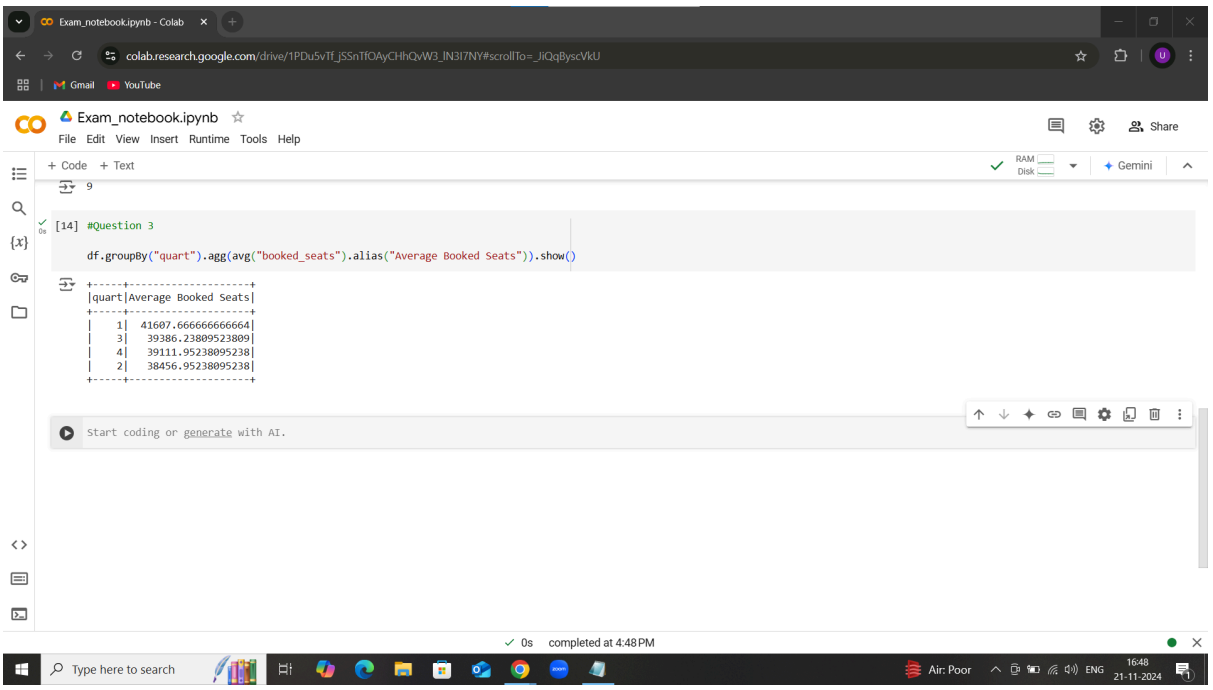
2:

```
df.filter(col("avg_rev")<290.0).count()
```



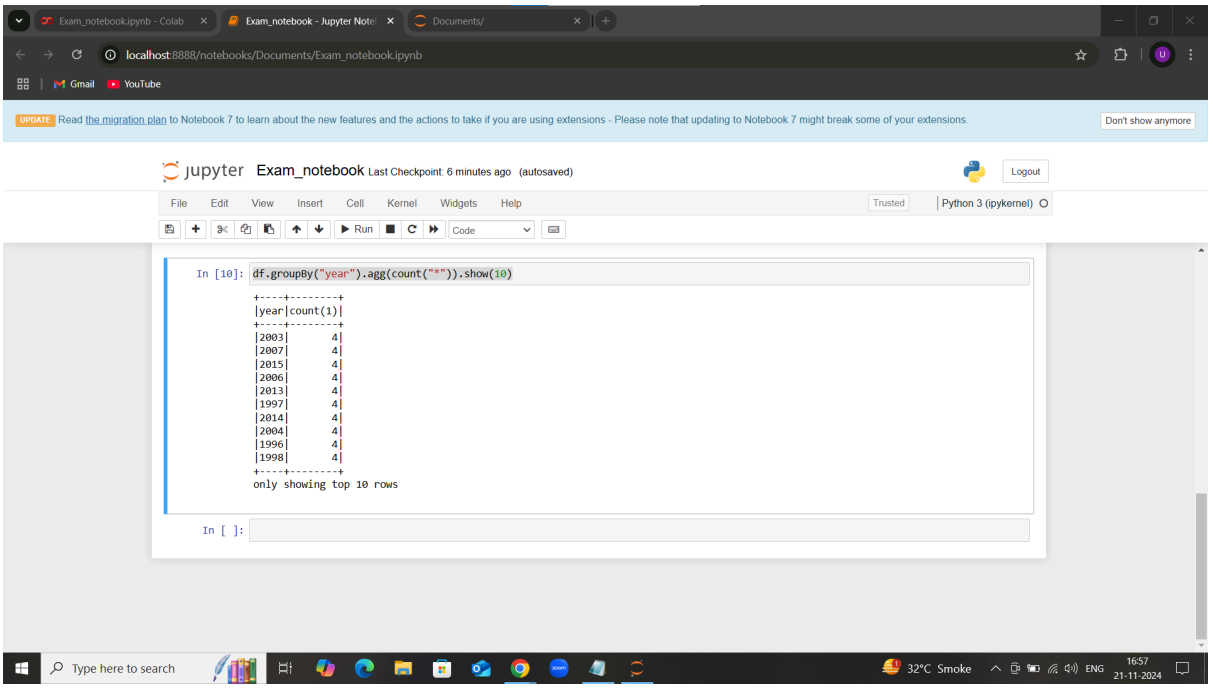
3:

```
df.groupBy("quart").agg(avg("booked_seats").alias("Average Booked Seats")).show()
```



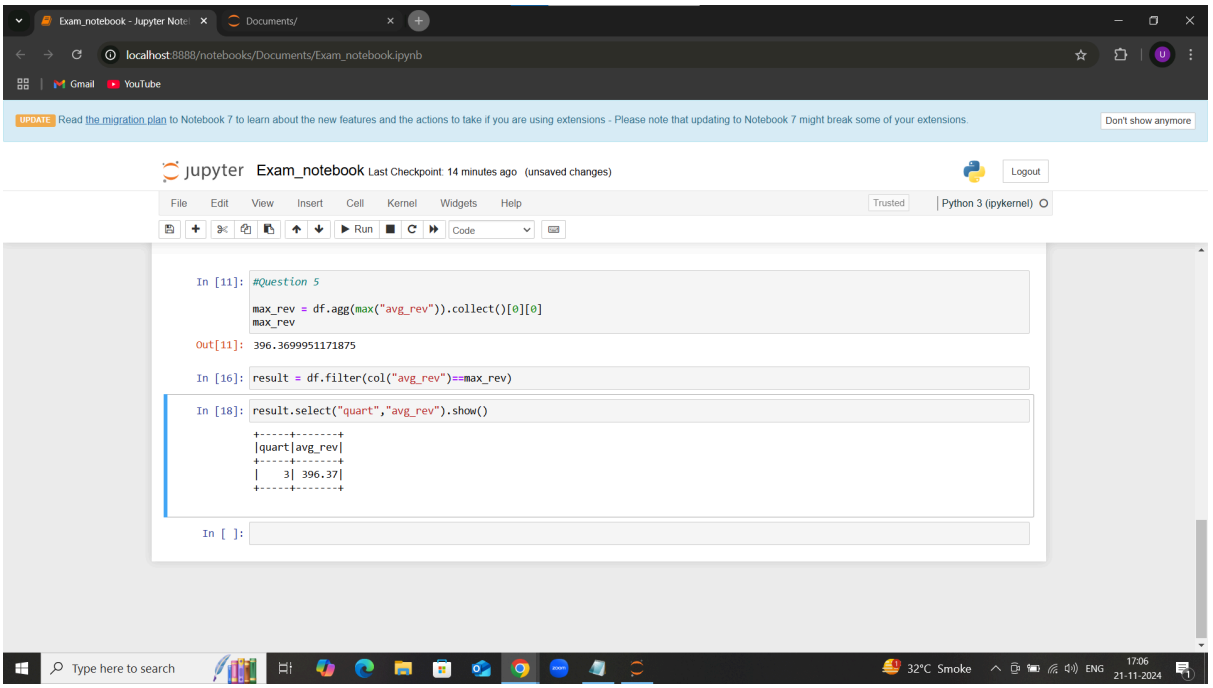
4:

df.groupBy("year").agg(count("*")).show(10)



5:

max_rev = df.agg(max("avg_rev")).collect()[0][0]
result = df.filter(col("avg_rev")==max_rev)
result.select("quart","avg_rev").show()



USE RDD ONLY

1:

Code:

```
#loading file
datardd = sc.textFile("user/cdacuser/training/airlines.csv")

#checking for header
datardd.take(5)

#eliminating the header
head = datardd.first()
eliminate = datardd.filter(lambda a: a!=head)
```

HIVE:

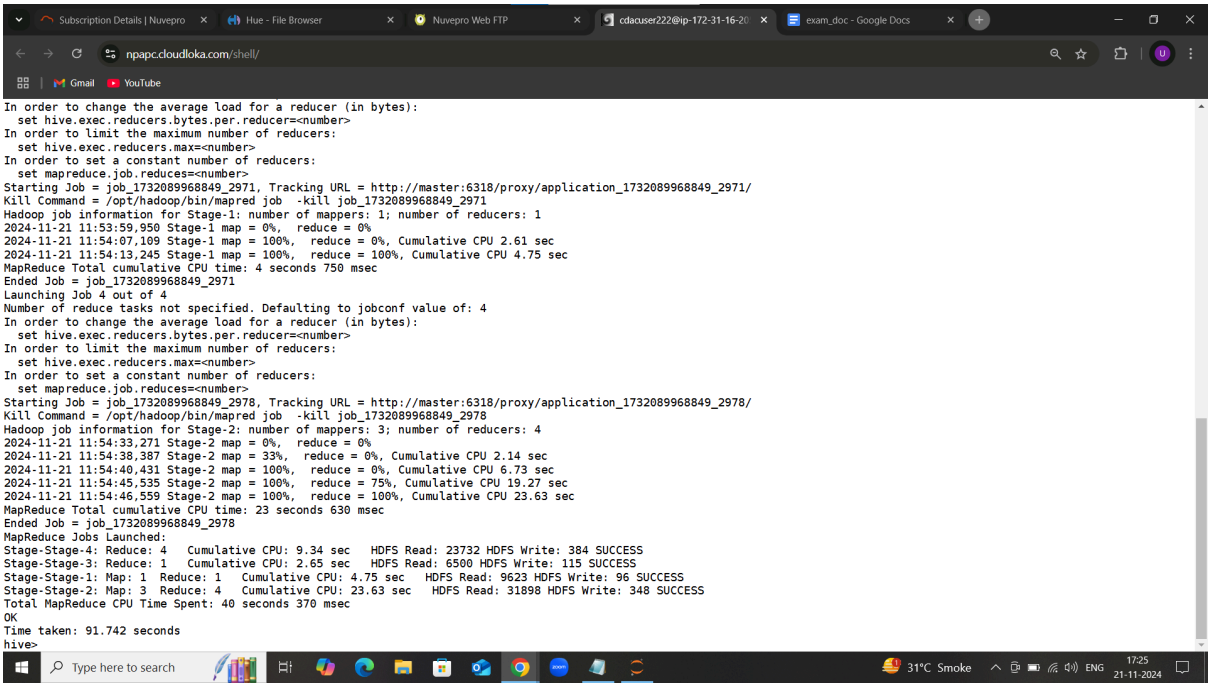
Question 1:

1:

Query:

```
select src_airport_id from routes where src_airport_id not in (select dest_airport_id
from routes) limit 5;
```

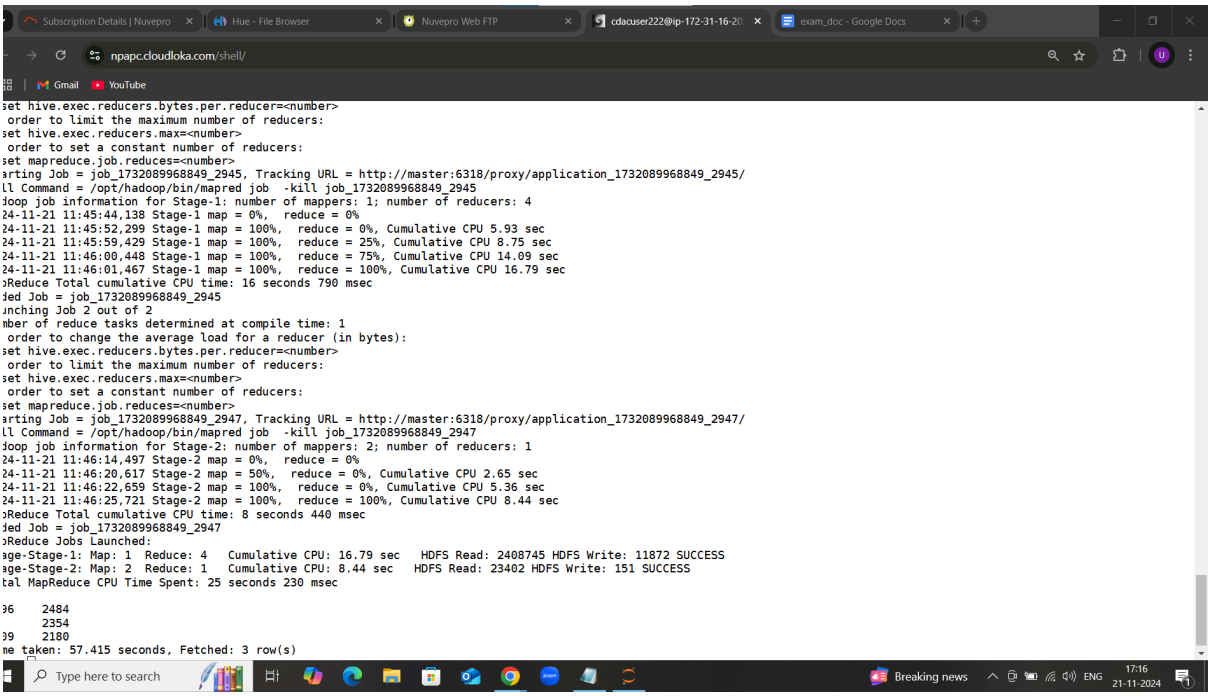
Output:



2:

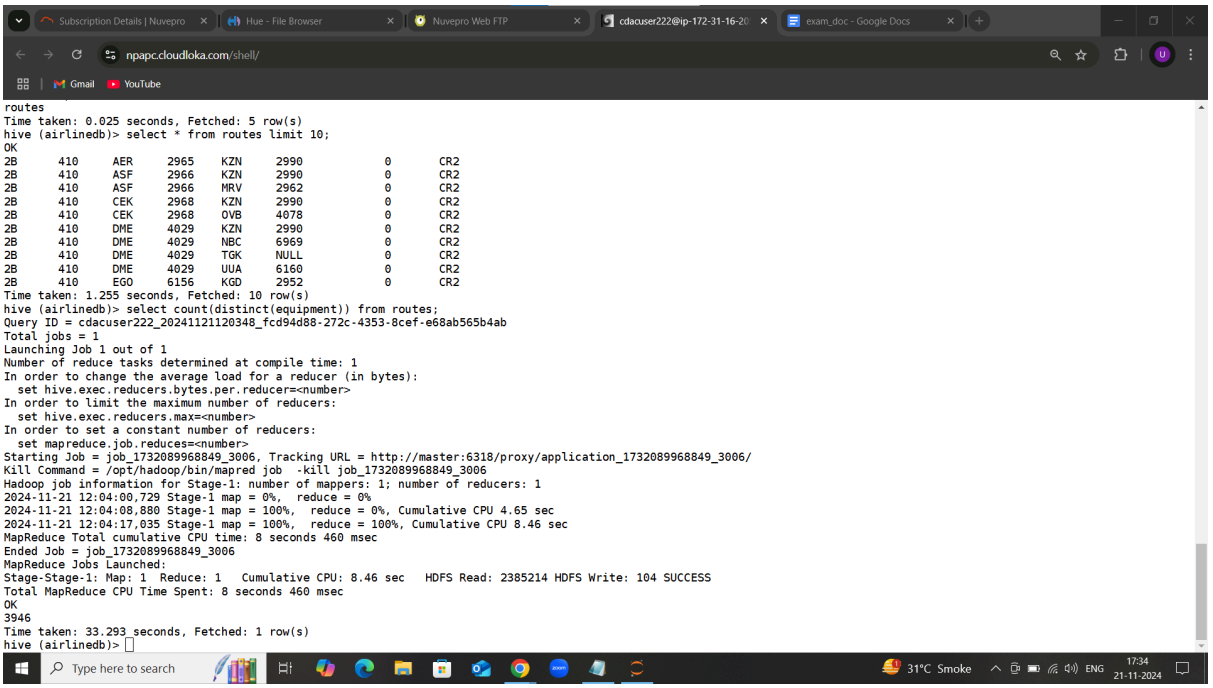
Query:
select airline_id, count(*) from routes group by airline_id order by count(*) desc limit 3;

Output:



3:
Query:
Select count(distinct(equipment)) from routes;

Output:



Question 2:

1.
Query:

create table routes_partitioned
(airline_iata string, airline_id int, src_airport_iata string,
src_airport_id int, dest_airport_iata string,
dest_airport_id int, codeshare string, stops int, equipment int)
partitioned by (dest_airport_iata) row format delimited
fields terminated by "," stored as textfile;