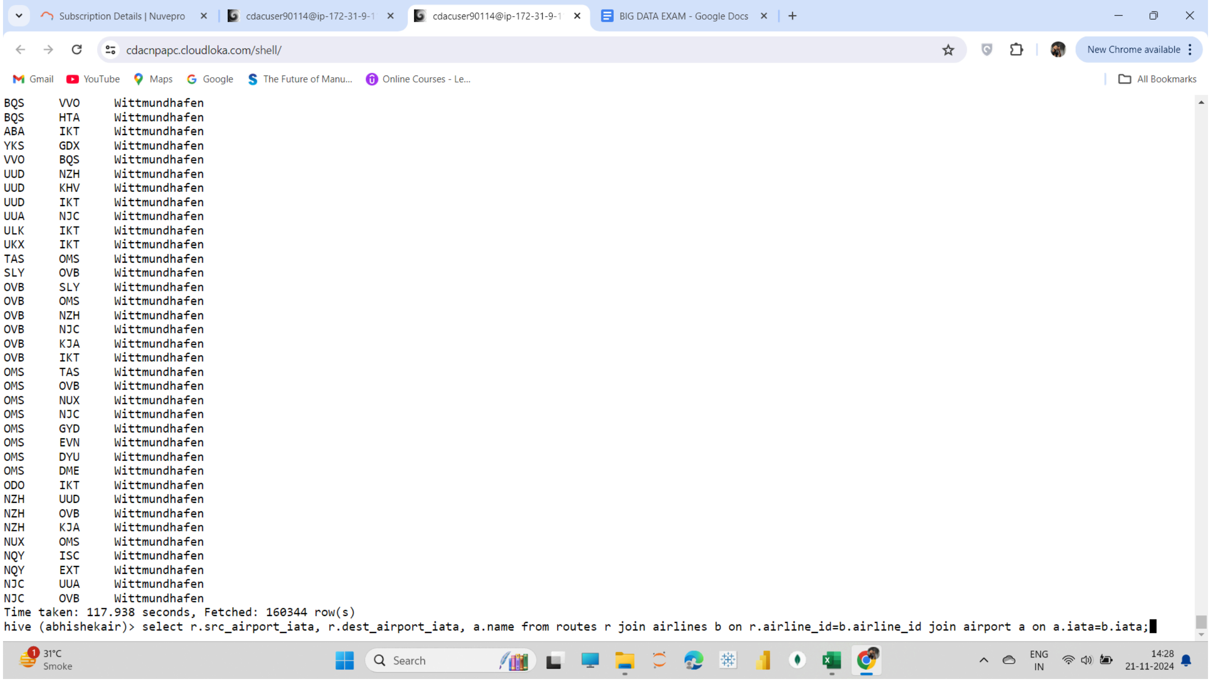
BIG DATA EXAM

Name : Abhishek Bhadarge
Roll No : 03

#HIVE
Question 1:

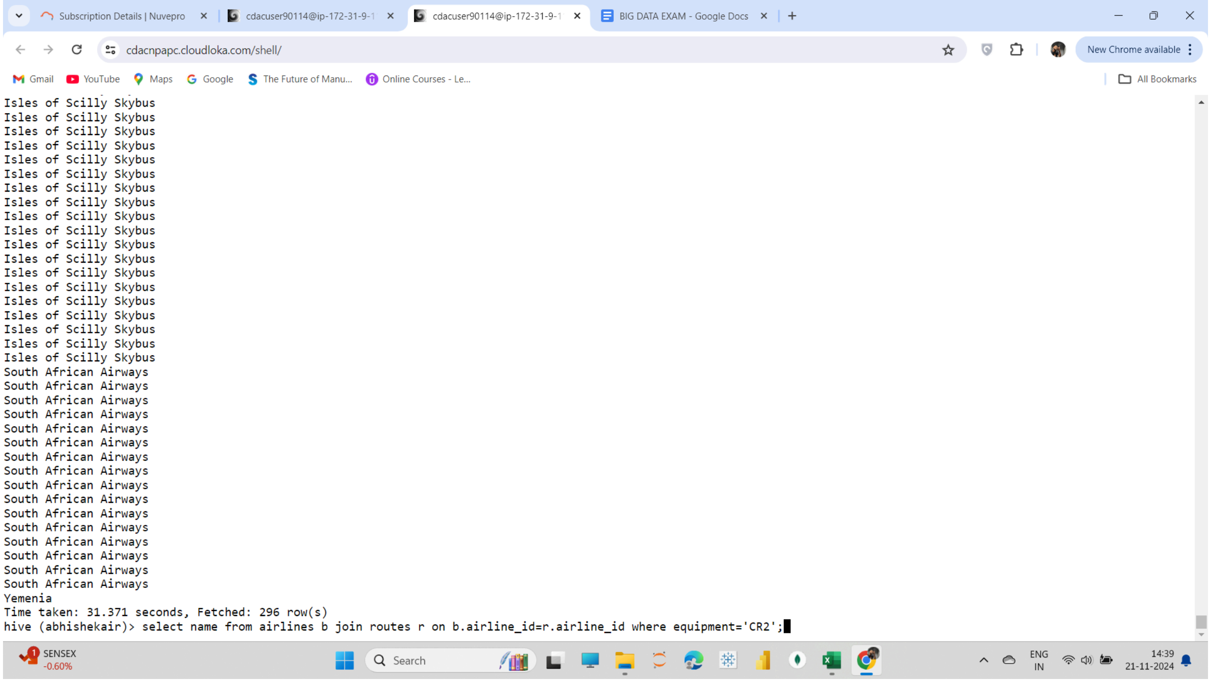
```
1.select r.src_airport_iata, r.dest_airport_iata, a.name from
routes r join airlines b on r.airline_id=b.airline_id join airport
a on a.iata=b.iata;
```



Subscription Details | Nuvepro
cdacuser90114@ip-172-31-9-1
cdacuser90114@ip-172-31-9-1
BIG DATA EXAM - Google Docs
cdacnpapc.cloudloka.com/shell/
New Chrome available
Gmail
YouTube
Maps
Google
The Future of Manu...
Online Courses - Le...
All Bookmarks
31°C
Smoke
Search
ENG
IN
14:28
21-11-2024

```
BQS     VVO     Wittmundhafen
BQS     HTA     Wittmundhafen
ABA     IKT     Wittmundhafen
YKS     GDX     Wittmundhafen
VVO     BQS     Wittmundhafen
UUD     NZH     Wittmundhafen
UUD     KHV     Wittmundhafen
UUD     IKT     Wittmundhafen
UUA     NJC     Wittmundhafen
ULK     IKT     Wittmundhafen
UKX     IKT     Wittmundhafen
TAS     OMS     Wittmundhafen
SLY     OVB     Wittmundhafen
OVB     SLY     Wittmundhafen
OVB     OMS     Wittmundhafen
OVB     NZH     Wittmundhafen
OVB     NJC     Wittmundhafen
OVB     KJA     Wittmundhafen
OVB     IKT     Wittmundhafen
OMS     TAS     Wittmundhafen
OMS     OVB     Wittmundhafen
OMS     NUX     Wittmundhafen
OMS     NJC     Wittmundhafen
OMS     GYD     Wittmundhafen
OMS     EVN     Wittmundhafen
OMS     DYU     Wittmundhafen
OMS     DME     Wittmundhafen
ODO     IKT     Wittmundhafen
NZH     UUD     Wittmundhafen
NZH     OVB     Wittmundhafen
NZH     KJA     Wittmundhafen
NUX     OMS     Wittmundhafen
NQY     ISC     Wittmundhafen
NQY     EXT     Wittmundhafen
NJC     UUA     Wittmundhafen
NJC     OVB     Wittmundhafen
Time taken: 117.938 seconds, Fetched: 160344 row(s)
hive (abhishekair)> select r.src_airport_iata, r.dest_airport_iata, a.name from routes r join airlines b on r.airline_id=b.airline_id join airport a on a.iata=b.iata;
```

```
2.select name from airlines b join routes r on
b.airline_id=r.airline_id where equipment='CR2';
```


Subscription Details | Nuvepro
cdacuser90114@ip-172-31-9-1
BIG DATA EXAM - Google Docs
cdacnpapc.cloudloka.com/shell/
New Chrome available
SENSEX
-0.60%
Search
14:39
21-11-2024

```
Isles of Scilly Skybus
Isles of Scilly Skybus
Isles of Scilly Skybus
Isles of Scilly Skybus
Isles of Scilly Skybus
Isles of Scilly Skybus
Isles of Scilly Skybus
Isles of Scilly Skybus
Isles of Scilly Skybus
Isles of Scilly Skybus
Isles of Scilly Skybus
Isles of Scilly Skybus
Isles of Scilly Skybus
Isles of Scilly Skybus
Isles of Scilly Skybus
Isles of Scilly Skybus
Isles of Scilly Skybus
Isles of Scilly Skybus
Isles of Scilly Skybus
South African Airways
South African Airways
South African Airways
South African Airways
South African Airways
South African Airways
South African Airways
South African Airways
South African Airways
South African Airways
South African Airways
South African Airways
South African Airways
South African Airways
South African Airways
South African Airways
Yemenia
Time taken: 31.371 seconds, Fetched: 296 row(s)
hive (abhishekair)> select name from airlines b join routes r on b.airline_id=r.airline_id where equipment='CR2';
```

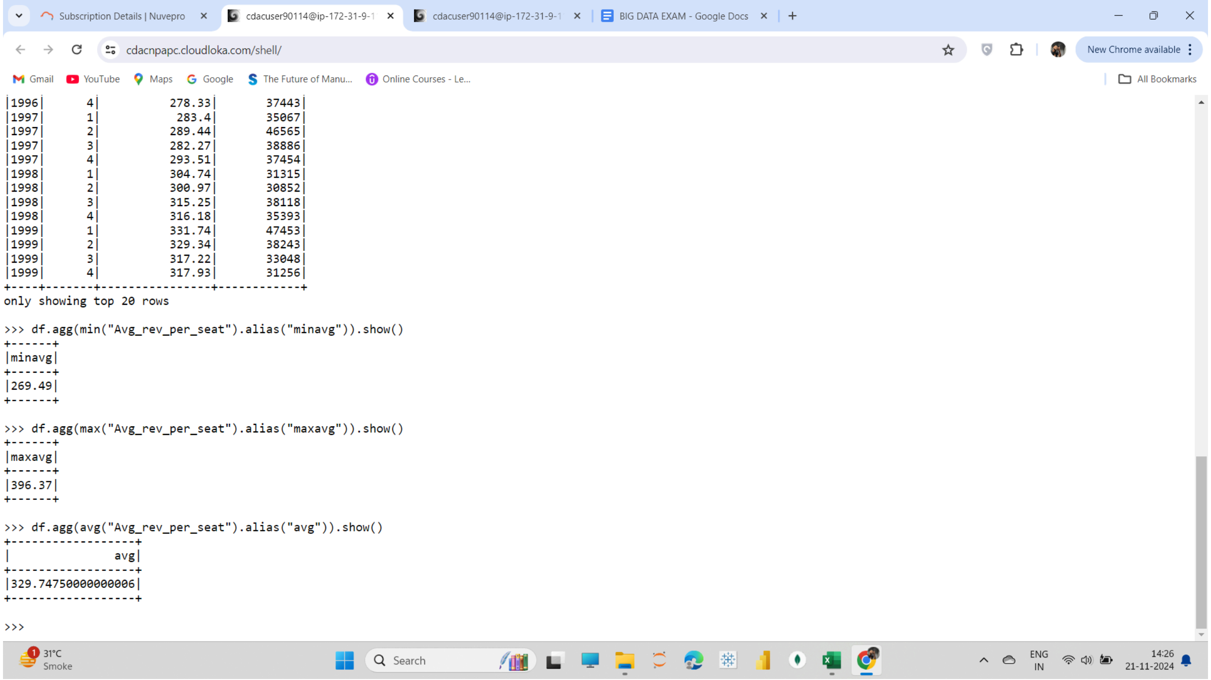
#SPARK

Question 2:

1.df.agg(min("Avg_rev_per_seat").alias("minavg")).show()

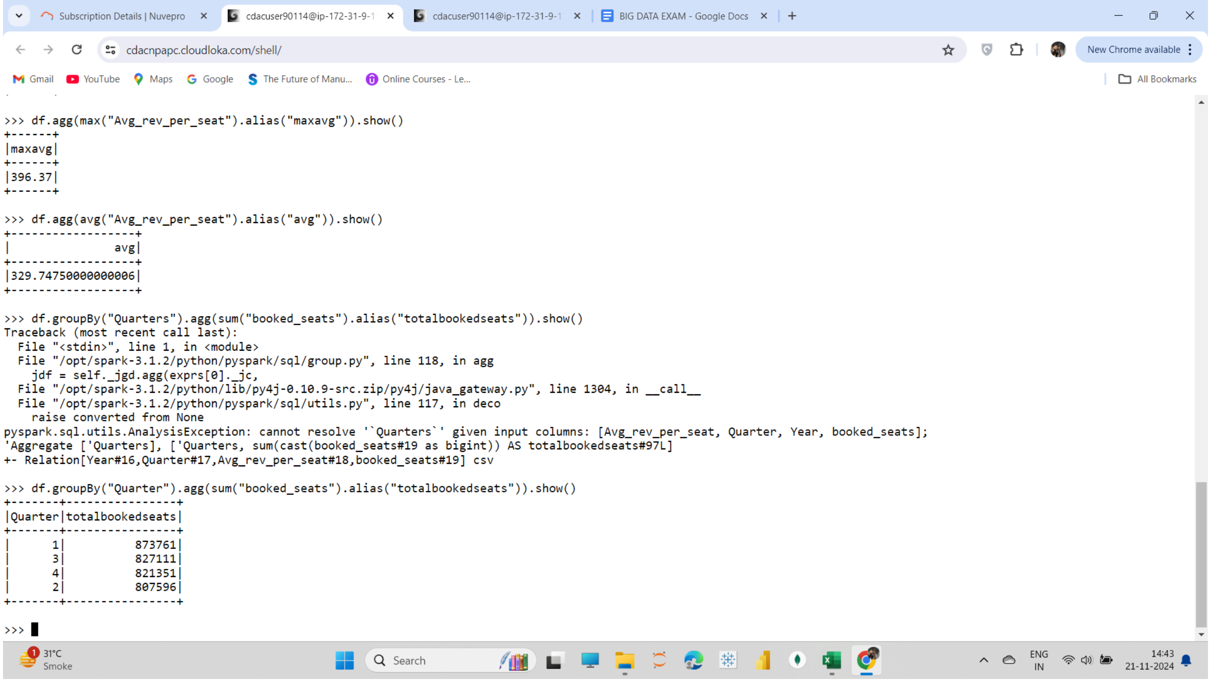
df.agg(max("Avg_rev_per_seat").alias("maxavg")).show()

df.agg(avg("Avg_rev_per_seat").alias("avg")).show()


Subscription Details | Nuvepro
cdacuser90114@ip-172-31-9-1
BIG DATA EXAM - Google Docs
cdacnpapc.cloudloka.com/shell/
New Chrome available
31°C
Smoke
Search
14:26
21-11-2024

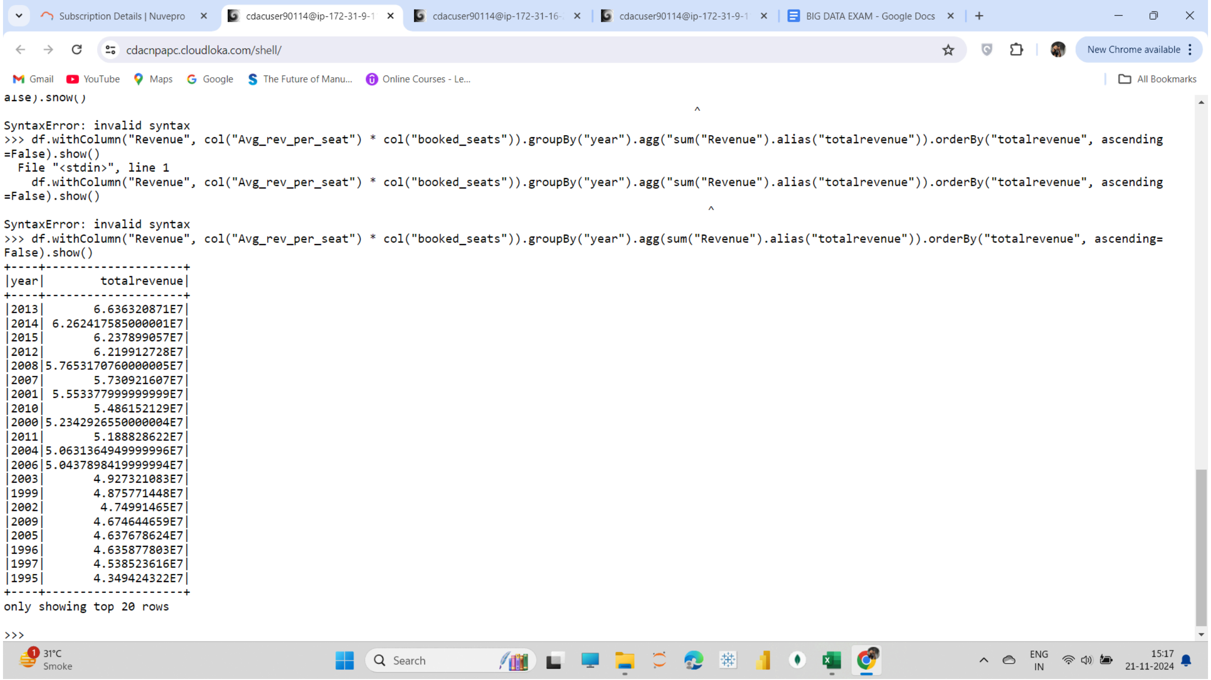
```
|1996|      4|          278.33|       37443|
|1997|      1|           283.4|       35067|
|1997|      2|          289.44|       46565|
|1997|      3|          282.27|       38886|
|1997|      4|          293.51|       37454|
|1998|      1|          304.74|       31315|
|1998|      2|          300.97|       30852|
|1998|      3|          315.25|       38118|
|1998|      4|          316.18|       35393|
|1999|      1|          331.74|       47453|
|1999|      2|          329.34|       38243|
|1999|      3|          317.22|       33048|
|1999|      4|          317.93|       31256|
+----+-------+----------------+------------+
only showing top 20 rows

>>> df.agg(min("Avg_rev_per_seat").alias("minavg")).show()
+------+
|minavg|
+------+
|269.49|
+------+

>>> df.agg(max("Avg_rev_per_seat").alias("maxavg")).show()
+------+
|maxavg|
+------+
|396.37|
+------+

>>> df.agg(avg("Avg_rev_per_seat").alias("avg")).show()
+------------------+
|               avg|
+------------------+
|329.74750000000006|
+------------------+

>>>
```

3.df.groupBy("Quarter").agg(sum("booked_seats").alias("totalbooked seats")).show()


Subscription Details | Nuvepro
cdacuser90114@ip-172-31-9-1
cdacuser90114@ip-172-31-9-1
BIG DATA EXAM - Google Docs
cdacnpapc.cloudloka.com/shell/
New Chrome available
Gmail
YouTube
Maps
Google
The Future of Manu...
Online Courses - Le...
All Bookmarks
31°C
Smoke
Search
ENG
IN
14:43
21-11-2024

```
>>> df.agg(max("Avg_rev_per_seat").alias("maxavg")).show()
+------+
|maxavg|
+------+
|396.37|
+------+

>>> df.agg(avg("Avg_rev_per_seat").alias("avg")).show()
+-----------------+
|              avg|
+-----------------+
|329.74750000000006|
+-----------------+

>>> df.groupBy("Quarters").agg(sum("booked_seats").alias("totalbookedseats")).show()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
  File "/opt/spark-3.1.2/python/pyspark/sql/group.py", line 118, in agg
    jdf = self._jgd.agg(exprs[0]._jc,
  File "/opt/spark-3.1.2/python/lib/py4j-0.10.9-src.zip/py4j/java_gateway.py", line 1304, in __call__
  File "/opt/spark-3.1.2/python/pyspark/sql/utils.py", line 117, in deco
    raise converted from None
pyspark.sql.utils.AnalysisException: cannot resolve '`Quarters`' given input columns: [Avg_rev_per_seat, Quarter, Year, booked_seats];
'Aggregate ['Quarters], ['Quarters, sum(cast(booked_seats#19 as bigint)) AS totalbookedseats#97L]
+- Relation[Year#16,Quarter#17,Avg_rev_per_seat#18,booked_seats#19] csv

>>> df.groupBy("Quarter").agg(sum("booked_seats").alias("totalbookedseats")).show()
+-------+----------------+
|Quarter|totalbookedseats|
+-------+----------------+
|      1|          873761|
|      3|          827111|
|      4|          821351|
|      2|          807596|
+-------+----------------+

>>> █
```

5.df.withColumn("Revenue", col("Avg_rev_per_seat") * col("booked_seats")).groupBy("year").agg(sum("Revenue").alias("totalrevenue")).orderBy("totalrevenue", ascending= False).show()


Subscription Details | Nuvepro
cdacuser90114@ip-172-31-9-1
cdacuser90114@ip-172-31-16-
cdacuser90114@ip-172-31-9-1
BIG DATA EXAM - Google Docs
cdacnpapc.cloudloka.com/shell/
New Chrome available
Gmail
YouTube
Maps
Google
The Future of Manu...
Online Courses - Le...
All Bookmarks
31°C
Smoke
Search
ENG
IN
15:17
21-11-2024

```
alse).show()
                                                                                                     ^
SyntaxError: invalid syntax
>>> df.withColumn("Revenue", col("Avg_rev_per_seat") * col("booked_seats")).groupBy("year").agg("sum("Revenue").alias("totalrevenue")).orderBy("totalrevenue", ascending
=False).show()
  File "<stdin>", line 1
    df.withColumn("Revenue", col("Avg_rev_per_seat") * col("booked_seats")).groupBy("year").agg("sum("Revenue").alias("totalrevenue")).orderBy("totalrevenue", ascending
=False).show()
                                                                                                       ^
SyntaxError: invalid syntax
>>> df.withColumn("Revenue", col("Avg_rev_per_seat") * col("booked_seats")).groupBy("year").agg(sum("Revenue").alias("totalrevenue")).orderBy("totalrevenue", ascending=
False).show()
+----+-------------------+
|year|       totalrevenue|
+----+-------------------+
|2013|       6.636320871E7|
|2014| 6.262417585000001E7|
|2015|       6.237899057E7|
|2012|       6.219912728E7|
|2008|5.7653170760000005E7|
|2007|       5.730921607E7|
|2001| 5.553377999999999E7|
|2010|       5.486152129E7|
|2000|5.2342926550000004E7|
|2011|       5.188828622E7|
|2004|5.0631364949999996E7|
|2006|5.0437898419999994E7|
|2003|       4.927321083E7|
|1999|       4.875771448E7|
|2002|        4.74991465E7|
|2009|       4.674644659E7|
|2005|       4.637678624E7|
|1996|       4.635877803E7|
|1997|       4.538523616E7|
|1995|       4.349424322E7|
+----+-------------------+
only showing top 20 rows

>>>
```

#Question 2(HIVE).

3. Select * from partitioned_routes where trim(upper(src_airport)=”LAX” limit 10);

2. Insert overwrite table partition (src_airport) select r.airline, r.src_airport, r.dest_airport where r.drc_airport=”JFK”;

1. Create table partition

#(SPARK)Question 2:

4.

```
rdd.map(lambda a:a[0].distinct)
```