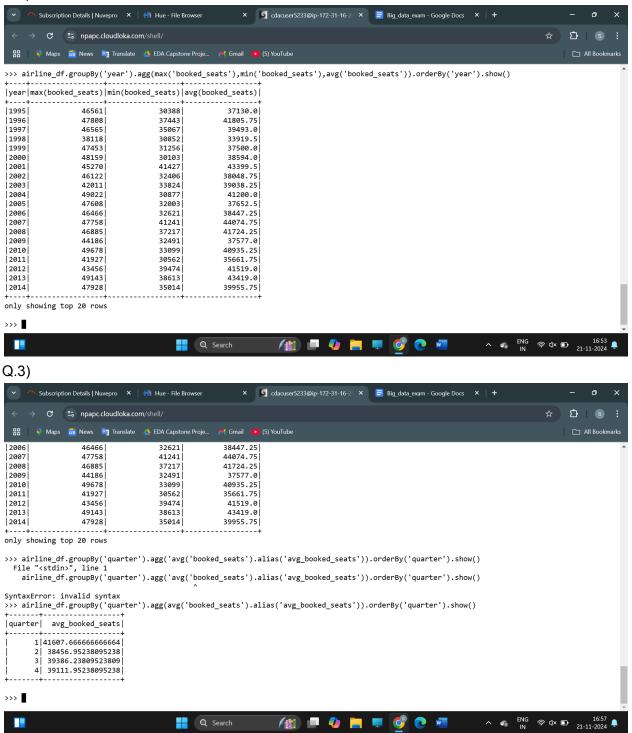


SPARK QUESTION 2, USing DataFrame.

Q.1)

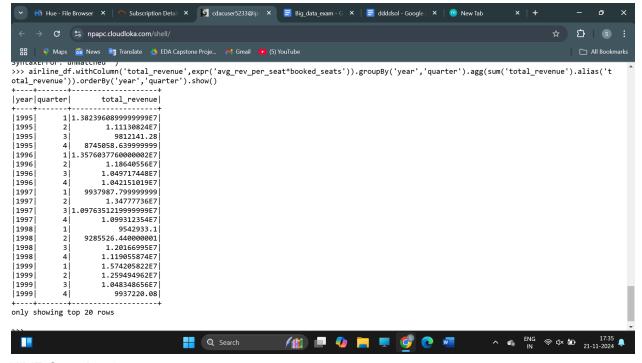


```
× G cdacuser5233@ip-172-31-16-2 × Big_data_exam - Google Docs × +
  \leftarrow \rightarrow \mathbf{C} \overset{\mathbf{c}_{5}}{} npapc.cloudloka.com/shell/
  🔡 🕴 Maps 🧰 News 🔯 Translate 🍐 EDA Capstone Proje... 🎽 Gmail 🔼 (5) YouTube
        4 39111.95238095238
>>> airline_df.withColumn('avg_revenue',expr('avg_rev_per_seat<290')).count('avg_revenue').show()
  File "<stdin>", line 1
   airline_df..withColumn('avg_revenue',expr('avg_rev_per_seat<290')).count('avg_revenue').show()</pre>
SyntaxError: invalid syntax
>>> airline_df.withColumn('avg_revenue',expr('avg_rev_per_seat<290')).count('avg_revenue').show()
Traceback (most recent call last):
File "<stdin>", line 1, in <module>
TypeError: count() takes 1 positional argument but 2 were given
>>> airline_df.withColumn('avg_revenue',expr('avg_rev_per_seat<290')).count().show()
Traceback (most recent call last):
File "<stdin>", line 1, in <module>
AttributeError: 'int' object has no attribute 'show'
>>> airline_df.withColumn('avg_revenue',expr('avg_rev_per_seat<290')).count('avg_revenue')
Traceback (most recent call last):
   File "<stdin>", line 1, in <module>
TypeError: count() takes 1 positional argument but 2 were given
>>> airline_df.withColumn('avg_revenue',expr('avg_rev_per_seat<290')).agg(count('avg_revenue')).show()
|count(avg_revenue)|
>>>
   Q Search
                                                                                                                                         🦈 🗘 🗖
4)
                                                             Subscription Details | Nuvepro × | 🔒 Hue - File Browser
      → C % npapc.cloudloka.com/shell/
                                                                                                                                                 立 ∣ s

☐ All Bookmarks

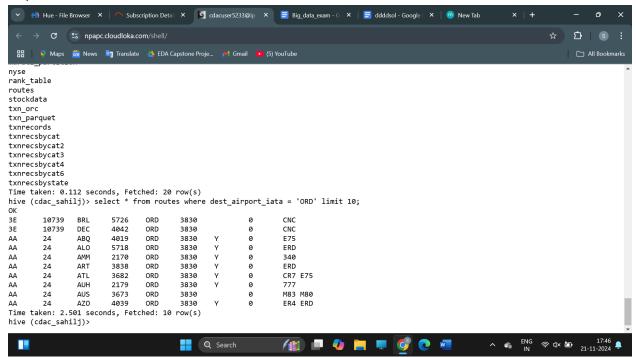
>>> airline_df.groupBy('year').agg(count('year')).orderBy('year').show()
|year|count(year)|
11995
 1996
 1997
                  4
 1998
                  4|
4|
4|
 1999
 2000
 2001
 2002
                  4
                  4
 2003
 2004
 2005
                  4 |
4 |
 2006
                  4
 2007
                  4
 2008
 2009
 2010
 2011
                  4
                  4
 2012
 2013
2014
                 4
only showing top 20 rows
>>>
  Q Search
```

5)



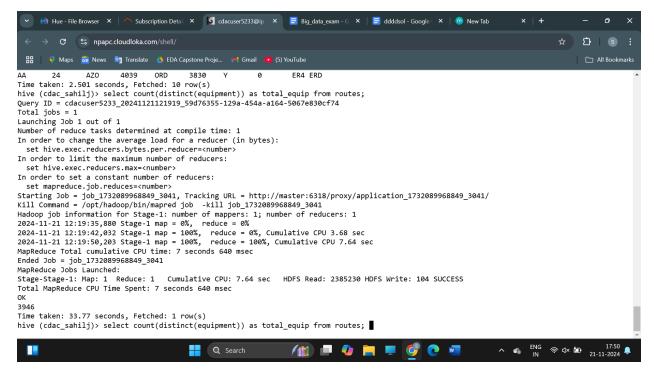
HIVE Question 2

Q3)



Question 1) in

3)



Question 1)

2)

