

Average:

```
airlineDF.groupBy("Year","Quarter").agg(avg("booked_seats").alias("Average_booked_seats"))
).show()
```

→ **Minimum:**

```
>>> airlineDF.groupBy("Year","Quarter").agg(min("booked_seats").alias("Minimum_booked_seats")).limit(4).show()
+-----+-----+-----+
|Year|Quarter|Minimum_booked_seats|
+-----+-----+-----+
|1998|2|30852|
+-----+-----+-----+

>>> airlineDF.groupBy("Year","Quarter").agg(max("booked_seats").alias("Maximum_booked_seats")).limit(1).show()
+-----+-----+-----+
|Year|Quarter|Maximum_booked_seats|
+-----+-----+-----+
|1998|2|30852|
+-----+-----+-----+

>>> airlineDF.groupBy("Year","Quarter").agg(max("booked_seats",ascending=False).alias("Maximum_booked_seats")).limit(1).show()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: max() got an unexpected keyword argument 'ascending'
>>> airlineDF.groupBy("Year","Quarter").agg(max("booked_seats").alias("Maximum_booked_seats")).orderBy("Maximum_booked_seats",ascending=False).limit(1).show()
+-----+-----+-----+
|Year|Quarter|Maximum_booked_seats|
+-----+-----+-----+
|2010|1|45678|
+-----+-----+-----+

>>> airlineDF.groupBy("Year","Quarter").agg(avg("booked_seats").alias("Average_booked_seats")).show()
+-----+-----+-----+
|Year|Quarter|Average_booked_seats|
+-----+-----+-----+
|1998|2|30852.0|
|2015|2|44871.0|
|2001|1|43853.0|
|1998|1|31315.0|
|2002|3|46122.0|
|2014|4|47928.0|
|2000|4|30103.0|
|2003|2|33824.0|
|2013|2|39315.0|
|2012|4|42987.0|
|2007|1|44307.0|
+-----+-----+-----+
```

Maximum:

```
>>> airlineDF.groupBy("Year","Quarter").agg(min("booked_seats").alias("Minimum_booked_seats")).limit(1).show()
+-----+-----+-----+
|Year|Quarter|Minimum_booked_seats|
+-----+-----+-----+
|1998|2|30852|
+-----+-----+-----+

>>> airlineDF.groupBy("Year","Quarter").agg(max("booked_seats").alias("Maximum_booked_seats")).limit(1).show()
+-----+-----+-----+
|Year|Quarter|Maximum_booked_seats|
+-----+-----+-----+
|1998|2|30852|
+-----+-----+-----+

>>> airlineDF.groupBy("Year","Quarter").agg(max("booked_seats",ascending=False).alias("Maximum_booked_seats")).limit(1).show()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: max() got an unexpected keyword argument 'ascending'
>>> airlineDF.groupBy("Year","Quarter").agg(max("booked_seats").alias("Maximum_booked_seats")).orderBy("Maximum_booked_seats",ascending=False).limit(1).show()
+-----+-----+-----+
|Year|Quarter|Maximum_booked_seats|
+-----+-----+-----+
|2010|1|45678|
+-----+-----+-----+

>>> airlineDF.groupBy("Year","Quarter").agg(avg("booked_seats").alias("Average_booked_seats")).show()
+-----+-----+-----+
|Year|Quarter|Average_booked_seats|
+-----+-----+-----+
|1998|2|30852.0|
|2015|2|44871.0|
|2001|1|43853.0|
|1998|1|31315.0|
|2002|3|46122.0|
|2014|4|47928.0|
|2000|4|30103.0|
|2003|2|33824.0|
|2013|2|39315.0|
|2012|4|42987.0|
|2007|1|44307.0|
+-----+-----+-----+
```

Average:

```
>>> airlineDF.groupBy("Year","Quarter").agg(max("booked_seats",ascending=False).alias("Maximum_booked_seats")).limit(1).show()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: max() got an unexpected keyword argument 'ascending'
>>> airlineDF.groupBy("Year","Quarter").agg(max("booked_seats").alias("Maximum_booked_seats")).orderBy("Maximum_booked_seats",ascending=False).limit(1).show()
+-----+-----+-----+
|Year|Quarter|Maximum_booked_seats|
+-----+-----+-----+
|2010|1|45678|
+-----+-----+-----+

>>> airlineDF.groupBy("Year","Quarter").agg(avg("booked_seats").alias("Average_booked_seats")).show()
+-----+-----+-----+
|Year|Quarter|Average_booked_seats|
+-----+-----+-----+
|1998|2|30852.0|
|2015|2|44871.0|
|2001|1|43853.0|
|1998|1|31315.0|
|2002|3|46122.0|
|2014|4|47928.0|
|2000|4|30103.0|
|2003|2|33824.0|
|2013|2|39315.0|
|2012|4|42987.0|
|2007|1|44307.0|
|1999|2|36243.0|
|2003|3|40426.0|
|1997|3|38886.0|
|1999|4|31256.0|
|2000|1|37783.0|
|2009|3|37001.0|
|1996|2|43026.0|
|2008|1|46885.0|
|2009|1|44186.0|
+-----+-----+-----+
only showing top 20 rows

>>>
```

2.



```
airlineDF.filter(col("Avg_rev_per_seat")<290).count()
```

```
Subscription Details | Nuvepro x cdcuser124@ip-172-31-16-20 x +
npapc.cloudloka.com/shell/

+-----+
|Year|Quarter|Maximum_booked_seats|
+-----+
|2010|1|49678|
+-----+

>>> airlineDF.groupBy("Year","Quarter").agg(avg("booked_seats").alias("Average_booked_seats")).show()

+-----+
|Year|Quarter|Average_booked_seats|
+-----+
|1998|2|30852.0|
|2015|2|44871.0|
|2001|1|43853.0|
|1998|1|31315.0|
|2002|3|46122.0|
|2014|4|47928.0|
|2000|4|30103.0|
|2003|2|33824.0|
|2013|2|39315.0|
|2012|4|42987.0|
|2007|1|44307.0|
|1999|2|38243.0|
|2003|3|40420.0|
|1997|3|38886.0|
|1999|4|31256.0|
|2000|3|37785.0|
|2009|3|37001.0|
|1996|2|43020.0|
|2008|1|46885.0|
|2009|1|44186.0|
+-----+

only showing top 20 rows

>>> df.filter(col("Avg_rev_per_seat")<290).count()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'df' is not defined
>>> airlineDF.filter(col("Avg_rev_per_seat")<290).count()
9
>>>
```

3.



```
airlineDF.groupBy("Quarter").avg("booked_seats").count()
```

```
airlineDF.groupBy("Quarter").avg("booked_seats").show()
```

```
Subscription Details | Nuvepro x cdcuser124@ip-172-31-16-20 x +
npapc.cloudloka.com/shell/

+-----+
|2013|2|39315.0|
|2012|4|42987.0|
|2007|1|44307.0|
|1999|2|38243.0|
|2003|3|40420.0|
|1997|3|38886.0|
|1999|4|31256.0|
|2000|3|37785.0|
|2009|3|37001.0|
|1996|2|43020.0|
|2008|1|46885.0|
|2009|1|44186.0|
+-----+

only showing top 20 rows

>>> df.filter(col("Avg_rev_per_seat")<290).count()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'df' is not defined
>>> airlineDF.filter(col("Avg_rev_per_seat")<290).count()
9
>>> airlineDF.groupBy("Quarter").agg("booked_seats").count()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
    File "/opt/spark-3.1.2/python/pyspark/sql/group.py", line 117, in agg
      assert all(isinstance(c, Column) for c in exprs), "all exprs should be Column"
AssertionError: all exprs should be Column
>>> airlineDF.groupBy("Quarter").avg("booked_seats").count()
4
>>> airlineDF.groupBy("Quarter").avg("booked_seats").show()

+-----+
|Quarter| avg(booked_seats)|
+-----+
|1|41607.666666666664|
|3| 39386.23809523809|
|4| 39111.95238095238|
|2| 38456.95238095238|
+-----+

>>>
```

4.



```
airlineDF.select("Year").distinct().count()
```

```

1999| 4| 31256.0|
2000| 3| 37785.0|
2009| 3| 37001.0|
1996| 2| 43020.0|
2008| 1| 46885.0|
2009| 1| 44186.0|
+-----+
only showing top 20 rows

>>> df.filter(col("Avg_rev_per_seat")<290).count()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'df' is not defined
>>> airlineDF.filter(col("Avg_rev_per_seat")<290).count()
9
>>> airlineDF.groupBy("Quarter").agg("booked_seats").count()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
  File "/opt/spark-3.1.2/python/pyspark/sql/group.py", line 117, in agg
    assert all(isinstance(c, Column) for c in exprs), "all exprs should be Column"
AssertionError: all exprs should be Column
>>> airlineDF.groupBy("Quarter").avg("booked_seats").count()
4
>>> airlineDF.groupBy("Quarter").avg("booked_seats").show()
+-----+
|Quarter| avg(booked_seats)|
+-----+
|1|41607.666666666664|
|3| 39386.23809523809|
|4| 39111.95238095238|
|2| 38456.95238095238|
+-----+

>>> df.select("Year").distinct().count()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'df' is not defined
>>> airlineDF.select("Year").distinct().count()
21
>>>

```

5.



```
airlineDF.groupBy("Year","Quarter").agg(sum("Avg_rev_per_seat")).orderBy("Year","Quarter",ascending=False).limit(1).show()
```

```

2008| 1| 333.29|
2009| 1| 313.82|
+-----+
only showing top 20 rows

>>> airlineDF.groupBy("Year","Quarter").agg(sum("Avg_rev_per_seat")).orderBy("Year","Quarter").show()
+-----+
|Year|Quarter|sum(Avg_rev_per_seat)|
+-----+
|1995| 1| 296.9|
|1995| 2| 296.8|
|1995| 3| 287.51|
|1995| 4| 287.78|
|1996| 1| 283.97|
|1996| 2| 275.78|
|1996| 3| 269.49|
|1996| 4| 278.33|
|1997| 1| 283.4|
|1997| 2| 289.44|
|1997| 3| 282.27|
|1997| 4| 293.51|
|1998| 1| 304.74|
|1998| 2| 300.97|
|1998| 3| 315.25|
|1998| 4| 316.18|
|1999| 1| 331.74|
|1999| 2| 329.34|
|1999| 3| 317.22|
|1999| 4| 317.93|
+-----+
only showing top 20 rows

>>> airlineDF.groupBy("Year","Quarter").agg(sum("Avg_rev_per_seat")).orderBy("Year","Quarter",ascending=False).limit(1).show()
+-----+
|Year|Quarter|sum(Avg_rev_per_seat)|
+-----+
|2015| 4| 362.56|
+-----+

>>>

```

Hive

Q1.

1.



hive

set hive.cli.print.current.db=true;

use cdac_vinay;

show tables;

```
cdacuser124@ip-172-31-16-20:~$ hive
SLF4J: class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/hive-3.1.1/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Hive Session ID = cbb15cbe-5826-4f2e-94fe-12a5c2deb875

Logging initialized using configuration in jar:file:/opt/hive-3.1.1/lib/hive-common-3.1.1.jar!/Hive-log4j2.properties Async: true
Hive Session ID = c1541c22-ccea-4142-96f5-265110f61256
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive> set hive.cli.print.current.db=true;
hive (default)> use cdac_vinay;
OK
Time taken: 0.342 seconds
hive (cdac_vinay)> show tables;
OK
airlines
airport
codeshare_src
customer
indianairports
nysse
parking
routes
txn_src
txn_parquet
txnrecords
txnnrcsbycat
txnnrcsbycat3
txnnrcsbycat4
txnnrcsbycat5
zerostops_src
Time taken: 0.099 seconds, Fetched: 16 row(s)
hive (cdac_vinay)>
```

select ap.name from airport ap join routes r on ap.airport_id = r.src_airport_id where r.dest_airport_id is null limit 10;

```
hive (default)> use cdac_vinay;
OK
Time taken: 0.365 seconds
hive (cdac_vinay)> select ap.name from airport ap join routes r on ap.airport_id = r.src_airport_id where r.dest_airport_id is null limit 10;
Query ID = cdacuser124_28241121115347_dceec9f9-7fbd-41b7-8d58-1b572c8be22a
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Defaulting to jobconf value of: 4
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reducers=number
Starting Job = job_1732089968849_2972, Tracking URL = http://master:6318/proxy/application_1732089968849_2972/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2972
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 4
2024-11-21 11:54:01,744 Stage-1 map = 0%, reduce = 0%, Cumulative CPU 6.56 sec
2024-11-21 11:54:09,983 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 12.71 sec
2024-11-21 11:54:15,099 Stage-1 map = 100%, reduce = 25%, Cumulative CPU 16.04 sec
2024-11-21 11:54:17,140 Stage-1 map = 100%, reduce = 75%, Cumulative CPU 21.46 sec
2024-11-21 11:54:22,259 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 26.74 sec
MapReduce Total cumulative CPU time: 26 seconds 740 msec
Ended Job = job_1732089968849_2972
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 4 Cumulative CPU: 26.74 sec HDFS Read: 3154551 HDFS Write: 1411 SUCCESS
Total MapReduce CPU Time Spent: 26 seconds 740 msec
OK
Thule Air Base
Cape Town Intl
Mozila
Gran Canaria
Nouakchott
Makale
Provence
Esenboga
Licenciado Benito Juarez Intl
Lynden Pindling Intl
Time taken: 36.892 seconds, Fetched: 10 row(s)
hive (cdac_vinay)>
```

2.



```
select r.src   airport iata, r.dest   airport iata, a.name from routes r join airlines a on r.airline_id
= a.airline order by r.airline_id desc limit 3;
```

```
In order to set a constant number of reducers:
set mapreduce.job.reduces=number>
Starting Job = job_1732089968849_3002, Tracking URL = http://master:6318/proxy/application_1732089968849_3002/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_3002
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 4
2024-11-21 12:03:12.104 Stage-1 map = 0%, reduce = 0%
2024-11-21 12:03:19.303 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 6.81 sec
2024-11-21 12:03:21.346 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 13.27 sec
2024-11-21 12:03:26.453 Stage-1 map = 100%, reduce = 25%, Cumulative CPU 17.67 sec
2024-11-21 12:03:27.470 Stage-1 map = 100%, reduce = 50%, Cumulative CPU 21.76 sec
2024-11-21 12:03:29.513 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 30.57 sec
MapReduce Total cumulative CPU time: 30 seconds 570 msec
Ended Job = job_1732089968849_3002
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
set hive.exec.reducers.bytes.per.reducer=number>
In order to limit the maximum number of reducers:
set hive.exec.reducers.max=number>
In order to set a constant number of reducers:
set mapreduce.job.reduces=number>
Starting Job = job_1732089968849_3005, Tracking URL = http://master:6318/proxy/application_1732089968849_3005/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_3005
Hadoop job information for Stage-2: number of mappers: 2; number of reducers: 1
2024-11-21 12:03:42.470 Stage-2 map = 0%, reduce = 0%
2024-11-21 12:03:49.628 Stage-2 map = 50%, reduce = 0%
2024-11-21 12:03:50.653 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 9.21 sec
2024-11-21 12:03:57.790 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 12.52 sec
MapReduce Total cumulative CPU time: 12 seconds 520 msec
Ended Job = job_1732089968849_3005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 4 Cumulative CPU: 30.57 sec HDFS Read: 2726804 HDFS Write: 2891787 SUCCESS
Stage-Stage-2: Map: 2 Reduce: 1 Cumulative CPU: 12.52 sec HDFS Read: 2904183 HDFS Write: 213 SUCCESS
Total MapReduce CPU Time Spent: 43 seconds 90 msec
OK
DAG
CCU Rainbow Air Polynesia
DAG JSR Rainbow Air Polynesia
DAG CXB Rainbow Air Polynesia
Time taken: 59.844 seconds, Fetched: 3 row(s)
master6318/proxy/application_1732089968849_3005/
```

3.



```
select count(distinct(equipment)) from routes;
```

```
Time taken: 31.017 seconds, Fetched: 341 row(s)
hive (cdac_vinay)> select count(*) from routes group by airline order by airline_id;
FAILED: SemanticException [Error 10004]: Line 1:50 Invalid table alias or column reference 'airline_id': (possible column names are: _c0)
hive (cdac_vinay)> desc routes;
OK
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         string
Time taken: 0.038 seconds, Fetched: 9 row(s)
hive (cdac_vinay)> select count(distinct(equipment)) from routes;
Query ID = cdacuser124_20241121121005_61ba91cc-d70f-4b86-8b39-214654ba8f1e
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
set hive.exec.reducers.bytes.per.reducer=number>
In order to limit the maximum number of reducers:
set hive.exec.reducers.max=number>
In order to set a constant number of reducers:
set mapreduce.job.reduces=number>
Starting Job = job_1732089968849_3024, Tracking URL = http://master:6318/proxy/application_1732089968849_3024/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_3024
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-11-21 12:10:15.768 Stage-1 map = 0%, reduce = 0%
2024-11-21 12:10:23.922 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.98 sec
2024-11-21 12:10:34.109 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 8.15 sec
MapReduce Total cumulative CPU time: 8 seconds 150 msec
Ended Job = job_1732089968849_3024
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 8.15 sec HDFS Read: 2385299 HDFS Write: 184 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 150 msec
OK
3946
Time taken: 39.363 seconds, Fetched: 1 row(s)
hive (cdac_vinay)>
```

```
select * from routes where dest airport iata = 'ORD' limit 10;
```



```
Subscription Details | NuoPro x cdauser124@ip-172-31-16-20 x +
ngapc.cloudiloka.com/shell/
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-11-21 12:10:15,768 Stage-1 map = 0%, reduce = 0%
2024-11-21 12:10:23,522 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.98 sec
2024-11-21 12:10:34,109 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 8.15 sec
MapReduce Total cumulative CPU time: 8 seconds 150 msec
Ended Job = job_1732889968849_3024
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 8.15 sec HDFS Read: 2385299 HDFS Write: 104 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 150 msec
OK
$946
Time taken: 33.363 seconds, Fetched: 1 row(s)
hive (cdac_vinay)> create external table routes_partitioned (airline_iata string, airline_id int, src_airport_iata string, src_airport_id int, dest_airport_iata string,
codeshare string, stops int, equipment string)
> partition by (dest_airport_id int)
> row format delimited
> fields terminated by ','
> store as textfile;
FAILED: ParseException line 2:0 missing EOF at 'partition' near ')'
hive (cdac_vinay)> create external table routes_partitioned (airline_iata string, airline_id int, src_airport_iata string, src_airport_id int, dest_airport_iata string,
codeshare string, stops int, equipment string)
> partition(dest_airport_id int)
> row format delimited
> fields terminated by ','
> store as textfile;
FAILED: ParseException line 2:0 missing EOF at 'partition' near ')'
hive (cdac_vinay)> select * from routes where dest_airport_iata = 'ORD' limit 10;
OK
3E 10739 BRL 5726 ORD 3830 0 CNC
3E 10739 DEC 4042 ORD 3830 0 CNC
AA 24 ABQ 4019 ORD 3830 Y 0 E75
AA 24 ALD 5718 ORD 3830 Y 0 ERD
AA 24 AMM 2170 ORD 3830 Y 0 340
AA 24 ART 3838 ORD 3830 Y 0 ERD
AA 24 ATL 3682 ORD 3830 Y 0 CR7 E75
AA 24 AUH 2170 ORD 3830 Y 0 777
AA 24 AUS 3673 ORD 3830 Y 0 MS3 M80
AA 24 AZO 4039 ORD 3830 Y 0 ER4 ERD
Time taken: 1.371 seconds, Fetched: 10 row(s)
hive (cdac_vinay)>
```

4.

