

```
hadoop fs -mkdir -p /salmah/hive/airbnb
```

```
hadoop fs -chmod 765 /salmah/hive/airbnb
```

```
hadoop fs -put /home/salmah/Downloads/LA_Listings.csv /salmah/hive/airbnb/
```

```
hadoop fs -put /home/salmah/Downloads/NY_Listings.csv /salmah/hive/airbnb/
```

```
cd hive/bin
```

```
hive
```

```
create database airbnb;
```

```
show databases;
```

```

CREATE TABLE la_listings (
  listing_id STRING,
  name STRING,
  host_id STRING,
  host_name STRING,
  host_response_rate STRING,
  host_is_superhost STRING,
  host_total_listings_count INT,
  street STRING,
  city STRING,
  neighbourhood_cleansed STRING,
  state STRING,
  country STRING,
  latitude FLOAT,
  longitude FLOAT,
  property_type STRING,
  room_type STRING,
  accommodates INT,
  bathrooms FLOAT,
  bedrooms FLOAT,
  amenities STRING,
  price FLOAT,
  minimum_nights INT,
  maximum_nights INT,
  availability_365 INT,
  calendar_last_scraped STRING,
  number_of_reviews INT,
  last_review_date STRING,
  review_scores_rating FLOAT,
  review_scores_accuracy FLOAT,
  review_scores_cleanliness FLOAT,
  review_scores_checkin FLOAT,
  review_scores_communication FLOAT,
  review_scores_location FLOAT,
  review_scores_value FLOAT,
  reviews_per_month FLOAT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
TBLPROPERTIES ("skip.header.line.count"="1");

LOAD DATA INPATH '/salmah/hive/airbnb/LA_Listings.csv' INTO TABLE la_listings;
select * from la_listings limit 5;

```

```

CREATE TABLE ny_listings (
  listing_id STRING,
  name STRING,
  host_id STRING,
  host_name STRING,
  host_response_rate STRING,
  host_is_superhost STRING,
  host_total_listings_count INT,
  street STRING,
  city STRING,
  neighbourhood_cleansed STRING,
  state STRING,
  country STRING,
  latitude FLOAT,
  longitude FLOAT,
  property_type STRING,
  room_type STRING,
  accommodates INT,
  bathrooms FLOAT,
  bedrooms FLOAT,
  amenities STRING,
  price FLOAT,
  minimum_nights INT,
  maximum_nights INT,
  availability_365 INT,
  calendar_last_scraped STRING,
  number_of_reviews INT,
  last_review_date STRING,
  review_scores_rating FLOAT,
  review_scores_accuracy FLOAT,
  review_scores_cleanliness FLOAT,
  review_scores_checkin FLOAT,
  review_scores_communication FLOAT,
  review_scores_location FLOAT,
  review_scores_value FLOAT,
  reviews_per_month FLOAT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
TBLPROPERTIES ("skip.header.line.count"="1");

LOAD DATA INPATH '/salmah/hive/airbnb/NY_Listings.csv' INTO TABLE ny_listings;

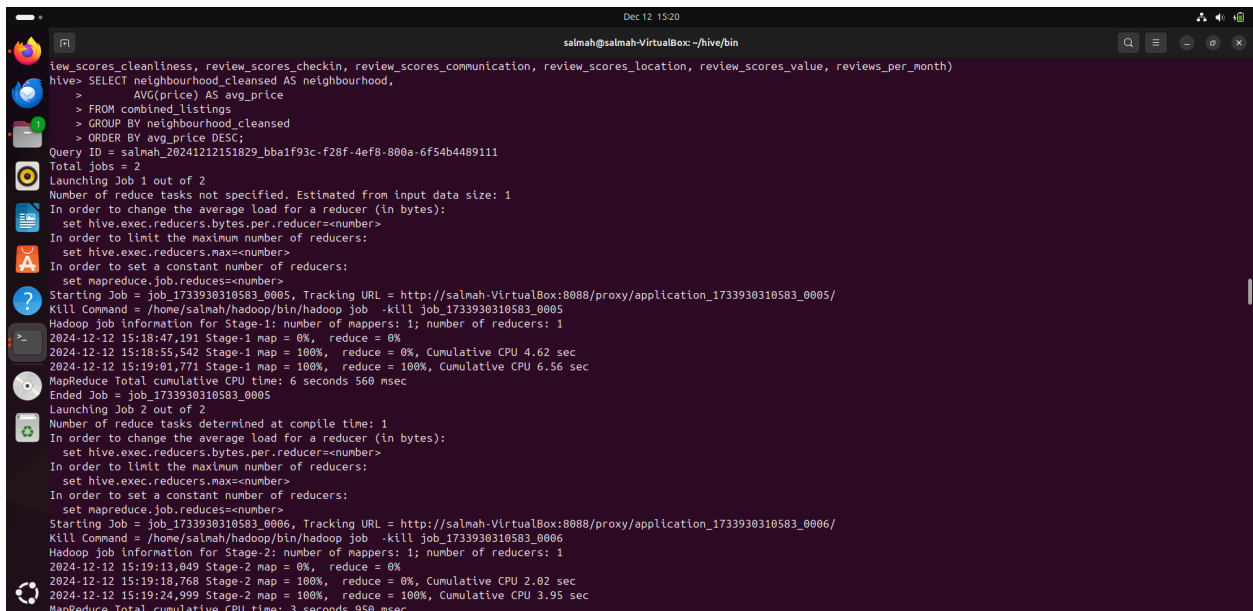
```

COMBINE LA AND NY

```
CREATE TABLE combined_listings AS
SELECT * FROM la_listings
UNION ALL
SELECT * FROM ny_listings;
```

AREA PRICING

```
SELECT neighbourhood_cleansed AS neighbourhood,
       AVG(price) AS avg_price
FROM combined_listings
GROUP BY neighbourhood_cleansed
ORDER BY avg_price DESC;
```



```
Dec 12 15:20
salmah@salmah-VirtualBox: ~/hive/bin

iew_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, review_scores_value, reviews_per_month)
hive> SELECT neighbourhood_cleansed AS neighbourhood,
>       AVG(price) AS avg_price
> FROM combined_listings
> GROUP BY neighbourhood_cleansed
> ORDER BY avg_price DESC;
Query ID = salmah_20241212151829_bba1f93c-f28f-4ef8-800a-6f54b4489111
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 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.reducers=<number>
Starting Job = job_1733930310583_0005, Tracking URL = http://salmah-VirtualBox:8088/proxy/application_1733930310583_0005/
Kill Command = /home/salmah/hadoop/bin/hadoop job -kill job_1733930310583_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-12-12 15:18:47,191 Stage-1 map = 0%, reduce = 0%
2024-12-12 15:18:55,542 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.62 sec
2024-12-12 15:19:01,771 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 6.56 sec
MapReduce Total cumulative CPU time: 6 seconds 560 msec
Ended Job = job_1733930310583_0005
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.reducers=<number>
Starting Job = job_1733930310583_0006, Tracking URL = http://salmah-VirtualBox:8088/proxy/application_1733930310583_0006/
Kill Command = /home/salmah/hadoop/bin/hadoop job -kill job_1733930310583_0006
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2024-12-12 15:19:13,049 Stage-2 map = 0%, reduce = 0%
2024-12-12 15:19:18,768 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 2.02 sec
2024-12-12 15:19:24,999 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 3.95 sec
MapReduce Total cumulative CPU time: 3 seconds 950 msec
```

```
Dec 12 15:20
salmah@salmah-VirtualBox: ~/hive/bin

MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 6.56 sec HDFS Read: 52495325 HDFS Write: 18807 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 3.95 sec HDFS Read: 23401 HDFS Write: 13908 SUCCESS
Total MapReduce CPU Time Spent: 10 seconds 510 msec

OK
Rolling Hills 2184.0
Bel-Air 1607.3193277310925
Country Club 1000.0
Beverly Crest 964.2893772893773
Walibu 964.245
Fort Wadsworth 800.0
Prince's Bay 667.5
Hollywood Hills West 660.0064239828694
Avalon 622.660465116279
Theater District 592.6285097192225
Downtown/Civic Center 591.0
Beverly Hills 572.078031212485
Florence-Firestone 542.6666666666666
Westlake Village 441.1666666666667
Southeast Antelope Valley 431.3333333333333
Unincorporated Santa Monica Mountains 426.09871244635195
Midtown 423.8252834943301
Century City 377.6666666666667
Encino 362.4313725490196
Brentwood 339.0110497237569
Palos Verdes Estates 335.63157894736844
Pacific Palisades 330.43870967741935
Woodrow 330.0
Tribeca 325.18729096989966
Manhattan Beach 318.84097859327215
Tottenville 305.8
Rancho Palos Verdes 300.48275862068965
La Canada Flintridge 298.85714285714283
Flatiron District 296.9855072463768
Chatsworth Reservoir 275.0
Hermosa Beach 269.7537878787879
Hollis Hills 267.5
Downtown 263.1811212266411
NoHo 252.06153846153848
```

```
Dec 12 15:21
salmah@salmah-VirtualBox: ~/hive/bin

Vermont Vista 64.9090909090909
Concord 64.8695652173913
Schuylerville 64.6666666666667
Tremont 64.47368421052632
Valinda 63.76923076923077
South San Gabriel 63.53333333333333
University Heights 63.529411764705884
Arden Heights 63.4
Bell 63.4
South Diamond Bar 63.0
Woodlawn 62.666666666666664
Morris Park 62.36
Corona 61.88235294117647
Mount Eden 61.76923076923077
Maywood 61.666666666666664
New Dorp Beach 61.4
Broadway-Manchester 61.13333333333333
Commerce 60.75
Grant City 59.25
Bronxdale 58.51724137931034
Northwest Palmdale 57.888888888888886
New Dorp 57.0
Hunts Point 56.407407407407405
Griffith Park 55.0
Ridge Route 55.0
Irwindale 52.5
Soundview 51.666666666666664
Lake View Terrace 51.5
East La Mirada 46.666666666666664
Charter Oak 44.76923076923077
Florence 43.42424242424242
Rancho Dominguez 42.5
Watts 39.833333333333336
Bull's Head 37.666666666666664
Harvard Park 29.76923076923077
Cudahy 20.0
Time taken: 57.415 seconds, Fetched: 491 row(s)
hive>
```

TOP CITY RENTAL

```
SELECT neighbourhood_cleansed AS neighbourhood,
       COUNT(*) AS total_listings
FROM combined_listings
GROUP BY neighbourhood_cleansed
ORDER BY total_listings DESC
LIMIT 10;
```

```
Dec 12 15:26
salmah@salmah-VirtualBox: ~/hive/bin

Total MapReduce CPU Time Spent: 0 msec
hive> SELECT neighbourhood_cleansed AS neighbourhood,
>       COUNT(*) AS total_listings
> FROM combined_listings
> GROUP BY neighbourhood_cleansed
> ORDER BY total_listings DESC
> LIMIT 10;
Query ID = salmah_20241212152501_a7114381-6874-4fe6-94bf-144f6436e0c5
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 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_1733930310583_0008, Tracking URL = http://salmah-VirtualBox:8088/proxy/application_1733930310583_0008/
Kill Command = /home/salmah/hadoop/bin/hadoop job -kill job_1733930310583_0008
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-12-12 15:25:07,687 Stage-1 map = 0%, reduce = 0%
2024-12-12 15:25:19,038 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.05 sec
2024-12-12 15:25:24,194 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 4.6 sec
MapReduce Total cumulative CPU time: 4 seconds 600 msec
Ended Job = job_1733930310583_0008
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_1733930310583_0009, Tracking URL = http://salmah-VirtualBox:8088/proxy/application_1733930310583_0009/
Kill Command = /home/salmah/hadoop/bin/hadoop job -kill job_1733930310583_0009
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2024-12-12 15:25:36,351 Stage-2 map = 0%, reduce = 0%
2024-12-12 15:25:40,455 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.43 sec
2024-12-12 15:25:45,614 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 3.33 sec
```

```
Dec 12 15:27
salmah@salmah-VirtualBox: ~/hive/bin

2024-12-12 15:25:07,687 Stage-1 map = 0%, reduce = 0%
2024-12-12 15:25:19,038 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.05 sec
2024-12-12 15:25:24,194 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 4.6 sec
MapReduce Total cumulative CPU time: 4 seconds 600 msec
Ended Job = job_1733930310583_0008
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_1733930310583_0009, Tracking URL = http://salmah-VirtualBox:8088/proxy/application_1733930310583_0009/
Kill Command = /home/salmah/hadoop/bin/hadoop job -kill job_1733930310583_0009
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2024-12-12 15:25:36,351 Stage-2 map = 0%, reduce = 0%
2024-12-12 15:25:40,455 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.43 sec
2024-12-12 15:25:45,614 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 3.33 sec
MapReduce Total cumulative CPU time: 3 seconds 330 msec
Ended Job = job_1733930310583_0009
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 4.6 sec HDFS Read: 52494942 HDFS Write: 15613 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 3.33 sec HDFS Read: 20300 HDFS Write: 175 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 930 msec
OK
Williamsburg 6017
Bedford-Stuyvesant 5529
Harlem 4189
Bushwick 3775
Hollywood 3507
Venice 3314
Hell's Kitchen 3313
Upper West Side 2990
East Village 2936
Upper East Side 2804
Time taken: 45.494 seconds, Fetched: 10 row(s)
hive>
```

ROOM PRICE BY TYPE

```
SELECT room_type, AVG(price) AS avg_price
FROM combined_listings
GROUP BY room_type
ORDER BY avg_price DESC;
```

```
Dec 12 15:31
salmah@salmah-VirtualBox: ~/hive/bin

Time taken: 45.494 seconds, Fetched: 10 row(s)
hive> SELECT room_type, AVG(price) AS avg_price
> FROM combined_listings
> GROUP BY room_type
> ORDER BY avg_price DESC;
Query ID = salmah_20241212152903_ce626fb7-4b0b-47f7-9f37-57cba0cb91c5
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 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_1733930310583_0010, Tracking URL = http://salmah-VirtualBox:8088/proxy/application_1733930310583_0010/
Kill Command = /home/salmah/hadoop/bin/hadoop job -kill job_1733930310583_0010
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-12-12 15:29:23,001 Stage-1 map = 0%, reduce = 0%
2024-12-12 15:30:12,653 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.65 sec
2024-12-12 15:30:24,786 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 10.33 sec
MapReduce Total cumulative CPU time: 10 seconds 330 msec
Ended Job = job_1733930310583_0010
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_1733930310583_0011, Tracking URL = http://salmah-VirtualBox:8088/proxy/application_1733930310583_0011/
Kill Command = /home/salmah/hadoop/bin/hadoop job -kill job_1733930310583_0011
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2024-12-12 15:30:48,914 Stage-2 map = 0%, reduce = 0%
2024-12-12 15:31:10,669 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 4.53 sec
2024-12-12 15:31:22,540 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 11.37 sec
MapReduce Total cumulative CPU time: 11 seconds 370 msec
Ended Job = job_1733930310583_0011
```

```
Dec 12 15:32
salmah@salmah-VirtualBox: ~/hive/bin

  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1733930310583_0010, Tracking URL = http://salmah-VirtualBox:8088/proxy/application_1733930310583_0010/
Kill Command = /home/salmah/hadoop/bin/hadoop job -kill job_1733930310583_0010
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-12-12 15:29:23,001 Stage-1 map = 0%, reduce = 0%
2024-12-12 15:30:12,653 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.65 sec
2024-12-12 15:30:24,786 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 10.33 sec
MapReduce Total cumulative CPU time: 10 seconds 330 msec
Ended Job = job_1733930310583_0010
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_1733930310583_0011, Tracking URL = http://salmah-VirtualBox:8088/proxy/application_1733930310583_0011/
Kill Command = /home/salmah/hadoop/bin/hadoop job -kill job_1733930310583_0011
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2024-12-12 15:30:48,914 Stage-2 map = 0%, reduce = 0%
2024-12-12 15:31:10,669 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 4.53 sec
2024-12-12 15:31:22,540 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 11.37 sec
MapReduce Total cumulative CPU time: 11 seconds 370 msec
Ended Job = job_1733930310583_0011
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 10.33 sec HDFS Read: 52495351 HDFS Write: 248 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 11.37 sec HDFS Read: 4808 HDFS Write: 126 SUCCESS
Total MapReduce CPU Time Spent: 21 seconds 700 msec
OK
Hotel room      359.0964912280702
Entire home/apt 227.12482515601462
Private room   102.11301265641502
Shared room    62.05795901469675
Time taken: 142.718 seconds, Fetched: 4 row(s)
hive>
```



```
SET hive.execution.engine=tez;  
SET tez.am.resource.memory.mb=2048;  
SET tez.am.dag.submit.time.timeout=60000;  
SET hive.tez.container.size=2048;  
SET tez.runtime.io.sort.mb=1024;
```