```
[(base) [zg978@hlog-2 ~]$ PYTHONSTARTUP=clean_airbnb.py pyspark --deploy-mode client
Python 3.7.9 (default, Feb 5 2021, 09:29:06)
[GCC 4.8.5 20150623 (Red Hat 4.8.5-39)] on linux
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newL
evel).
Welcome to
   /__/___/___//__
_\ \/ _ \/ _ \/ _ \/ _/ \/ _/
/__ / .__/\_, _/_/ /_\\ version 2.4.0-cdh6.3.4
Using Python version 3.7.9 (default, Feb 5 2021 09:29:06)
SparkSession available as 'spark'.
root
  |-- id: integer (nullable = true)
  -- listing_url: string (nullable = true)
  |-- scrape_id: long (nullable = true)
  |-- last_scraped: timestamp (nullable = true)
  -- name: string (nullable = true)
  |-- description: string (nullable = true)
  -- neighborhood_overview: string (nullable = true)
  |-- picture_url: string (nullable = true)
  |-- host_id: integer (nullable = true)
  |-- host_url: string (nullable = true)
  |-- host_name: string (nullable = true)
  |-- host_since: timestamp (nullable = true)
  |-- host_location: string (nullable = true)
  |-- host_about: string (nullable = true)
  -- host_response_time: string (nullable = true)
  |-- host_response_rate: string (nullable = true)
  -- host_acceptance_rate: string (nullable = true)
  |-- host_is_superhost: string (nullable = true)
  -- host_thumbnail_url: string (nullable = true)
  |-- host_picture_url: string (nullable = true)
  |-- host_neighbourhood: string (nullable = true)
  -- host_listings_count: integer (nullable = true)
  |-- host_total_listings_count: integer (nullable = true)
  -- host_verifications: string (nullable = true)
  -- host_has_profile_pic: string (nullable = true)
  |-- host_identity_verified: string (nullable = true)
  -- neighbourhood: string (nullable = true)
  |-- neighbourhood_cleansed: string (nullable = true)
  -- neighbourhood_group_cleansed: string (nullable = true)
  -- latitude: double (nullable = true)
  |-- longitude: double (nullable = true)
  -- property_type: string (nullable = true)
  |-- room_type: string (nullable = true)
  -- accommodates: integer (nullable = true)
  |-- bathrooms: string (nullable = true)
  |-- bathrooms_text: string (nullable = true)
  |-- bedrooms: integer (nullable = true)
  |-- beds: integer (nullable = true)
  -- amenities: string (nullable = true)
  |-- price: string (nullable = true)
  |-- minimum_nights: integer (nullable = true)
  |-- maximum_nights: integer (nullable = true)
  |-- minimum_minimum_nights: integer (nullable = true)
```

|-- maximum_minimum_nights: integer (nullable = true)

```
|-- maximum_nights_avg_ntm: double (nullable = true)
 |-- calendar_updated: string (nullable = true)
 -- has_availability: string (nullable = true)
 |-- availability_30: integer (nullable = true)
 |-- availability_60: integer (nullable = true)
 |-- availability_90: integer (nullable = true)
 |-- availability_365: integer (nullable = true)
 |-- calendar last scraped: timestamp (nullable = true)
 |-- number_of_reviews: integer (nullable = true)
 -- number_of_reviews_ltm: integer (nullable = true)
 |-- number_of_reviews_130d: integer (nullable = true)
 |-- first review: timestamp (nullable = true)
 |-- last_review: timestamp (nullable = true)
 |-- review_scores_rating: double (nullable = true)
 |-- review scores accuracy: double (nullable = true)
 |-- review_scores_cleanliness: double (nullable = true)
 |-- review_scores_checkin: double (nullable = true)
 |-- review_scores_communication: double (nullable = true)
 |-- review_scores_location: double (nullable = true)
 |-- review_scores_value: double (nullable = true)
 |-- license: string (nullable = true)
 |-- instant_bookable: string (nullable = true)
 |-- calculated_host_listings_count: integer (nullable = true)
 |-- calculated_host_listings_count_entire_homes: integer (nullable = true)
 |-- calculated host listings count private rooms: integer (nullable = true)
 |-- calculated_host_listings_count_shared_rooms: integer (nullable = true)
 |-- reviews_per_month: double (nullable = true)
22/04/26 22:49:26 WARN util.Utils: Truncated the string representation of a plan since
it was too large. This behavior can be adjusted by setting 'spark.debug.maxToStringFi
elds' in SparkEnv.conf.
6413
4768
Nan Percentage: 0.25651021362856696
/opt/cloudera/parcels/CDH-6.3.4-1.cdh6.3.4.p0.6626826/lib/spark/python/pyspark/shell.p
v:19: FutureWarning: The default value of regex will change from True to False in a fu
ture version. In addition, single character regular expressions will*not* be treated a
s literal strings when regex=True.
  An interactive shell.
id
                            int32
host_neighbourhood
                           object
neighbourhood_cleansed
                           object
latitude
                          float64
```

22/04/26 22:49:32 WARN scheduler.TaskSetManager: Stage 3 contains a task of very large size (491 KB). The maximum recommended task size is 100 KB.

float64

float64

float64

float64

int32

longitude

number of reviews

reviews_per_month

dtype: object

review_scores_rating

price