```
[(base) [zg978@hlog-2 ~]$ PYTHONSTARTUP=descriptive_stats.py pyspark --deploy-mode clie]
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
  Using Python version 3.7.9 (default, Feb 5 2021 09:29:06)
SparkSession available as 'spark'.
root
 |-- id: integer (nullable = true)
 |-- host_neighbourhood: string (nullable = true)
 |-- neighbourhood_cleansed: string (nullable = true)
 |-- latitude_x: double (nullable = true)
 |-- longitude_x: double (nullable = true)
 |-- price: double (nullable = true)
 |-- number_of_reviews: integer (nullable = true)
 |-- review_scores_rating: double (nullable = true)
 |-- reviews_per_month: double (nullable = true)
 |-- crime_counts: integer (nullable = true)
 |-- PdId: long (nullable = true)
 |-- IncidntNum: integer (nullable = true)
 |-- Incident Code: integer (nullable = true)
 |-- Category: string (nullable = true)
 |-- Descript: string (nullable = true)
 |-- DayOfWeek: string (nullable = true)
 |-- Date: string (nullable = true)
 |-- PdDistrict: string (nullable = true)
 |-- Address: string (nullable = true)
 |-- location: string (nullable = true)
 |-- latitude_y: double (nullable = true)
 |-- longitude_y: double (nullable = true)
 |-- distance: double (nullable = true)
 |-- DBA Name: string (nullable = true)
 |-- Source Zipcode: string (nullable = true)
 |-- Neighborhoods - Analysis Boundaries: string (nullable = true)
 |-- Business Location: string (nullable = true)
 |-- Parking Tax: boolean (nullable = true)
 |-- Transient Occupancy Tax: boolean (nullable = true)
 |-- longitude_z29: double (nullable = true)
 |-- longitude_z30: double (nullable = true)
22/04/26 22:57:35 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.
            id host_neighbourhood ... longitude_z29 longitude_z30
                Duboce Triangle ... 37.773400
0
            958
                                                         -122.435530
                 Duboce Triangle ... 37.773400 -122.435530

Lower Haight ... 37.773400 -122.435530

Lower Haight ... 37.773400 -122.435530

Bernal Heights ... 37.742184 -122.423450
      30564588
35674827
1
2
3
      39969765
4
          5858
```

[(base) [zg978@hlog-2 ~]\$ module load python/gcc/3.7.9

```
Vinohrady
6409
      53519481
                                              37.781180
                                                            -122.493380
                                      . . .
6410
                   Upper Rockridge
      53586203
                                              37.782234
                                                            -122.412285
                                      . . .
6411
     53644011
                      Inner Sunset
                                              37.754158
                                                            -122.478960
6412
      53647393
                        Noe Valley
                                              37.783340
                                                            -122.417060
                                      . . .
[6413 rows x 31 columns]
neighbourhood_cleansed
                              object
latitude_x
                             float64
longitude_x
                             float64
price
                             float64
number_of_reviews
                               int32
review_scores_rating
                             float64
reviews_per_month
                             float64
crime_counts
                               int32
distance
                             float64
Parking Tax
                             float64
Transient Occupancy Tax
                             float64
dtype: object
latitude_x:
mean
        37.767068
         0.023146
std
min
        37.708370
25%
        37.751377
50%
        37.769440
75%
        37.786980
max
        37.807740
Name: latitude_x, dtype: float64
latitude_x mode: 37.76942
longitude_x:
       -122.431322
mean
std
          0.027356
       -122.509910
min
25%
       -122.444257
50%
       -122.424740
       -122.410957
75%
       -122.369030
max
Name: longitude_x, dtype: float64
longitude_x mode: -122.44992
price:
          210.687155
mean
          435.835111
std
min
            0.000000
25%
           95.000000
50%
          145.000000
75%
          227.000000
        25000.000000
max
Name: price, dtype: float64
price mode: 150.0
number_of_reviews:
mean
         56.151694
```

None

. . .

-122.471430

37.746952

6408

std

min

25%

50%

92.574375

1.000000

4.000000

16.000000

53503353

```
25%
          4.000000
50%
         16.000000
75%
         63.000000
        877.000000
max
Name: number_of_reviews, dtype: float64
number_of_reviews mode: 1
review_scores_rating:
      4.734937
mean
std
        0.546967
min
        0.000000
25%
        4.710000
50%
        4.890000
75%
        5.000000
        5.000000
max
Name: review_scores_rating, dtype: float64
review_scores_rating mode: 5.0
reviews_per_month:
        2.055786
mean
std
         5.745753
min
         0.010000
25%
          0.210000
50%
          0.670000
75%
          2.130000
        162.000000
max
Name: reviews_per_month, dtype: float64
reviews_per_month mode: 0.04
crime_counts:
mean
      3811.146178
std
       1438.106772
min
        468.000000
25%
        3143.500000
50%
        4472.000000
75%
        4765.000000
        5325.000000
max
Name: crime_counts, dtype: float64
crime_counts mode: 4450
distance:
mean 0.000793
std
       0.000596
min
       0.000008
25%
        0.000367
50%
        0.000629
75%
        0.001062
        0.006581
max
Name: distance, dtype: float64
distance mode: 0.0001303868588
neighbourhood_cleansed:
neighbourhood_cleansed
Downtown/Civic Center
                         489
Mission
                         450
Western Addition
                         352
```

| South of Mark | et 294 | 4 |
|---------------|--------------|---|
| Bernal Height | s 265 | 5 |
| Castro/Upper | | |
| Outer Sunset | 239 | |
| Haight Ashbur | | |
| Nob Hill | 229 | |
| Noe Valley | 201 | |
| | | |
| Marina | 150 | |
| North Beach | 138 | |
| Inner Richmon | | |
| Outer Richmon | d 133 | 3 |
| Bayview | 127 | 7 |
| Potrero Hill | 120 | 6 |
| Russian Hill | 123 | 3 |
| Pacific Heigh | | |
| Inner Sunset | 112 | |
| Financial Dis | | |
| Excelsior | 103 | |
| Outer Mission | | |
| | 95 | |
| Parkside | | |
| Chinatown | 9: | |
| Ocean View | 78 | |
| West of Twin | | |
| Glen Park | 49 | 9 |
| Visitacion Va | lley 42 | 2 |
| Twin Peaks | 42 | 2 |
| Lakeshore | 37 | 7 |
| Crocker Amazo | n 29 | 9 |
| Presidio Heig | | |
| Seacliff | 1: | |
| Diamond Heigh | | |
| Presidio | | _ |
| | 5 mls | |
| Golden Gate P | dIK ' | 4 |
| dtype: int64 | | |
| | | |
| latitude_x: | | |
| latitude_x | | |
| | 17 | |
| | 17 | |
| 37.789310 | 14 | |
| 37.799881 | 11 | |
| 37.790560 | 10 | |
| | •• | |
| 37.756440 | 1 | |
| 37.756460 | 1 | |
| 37.756510 | 1 | |
| | 1 | |
| 37.730320 | 1 | |
| | | |
| Length: 3651, | dtype: int64 | |
| | | |
| longitude_x: | | |
| longitude_x | | |
| -122.44991 | 17 | |
| -122.44992 | 17 | |
| -122.40588 | 13 | |
| -122.40710 | | |
| -122.41952 | 11 | |
| | | |
| -122.44102 | 1 | |
| -122.44102 | 1 | |
| | | |
| -122.44098 | 1 | |
| | | |

```
longitude_x:
longitude_x
-122.44991
             17
-122.44992
             17
-122.40588
             13
-122.40710 13
-122.41952
             11
-122.44102
              1
              1
-122.44100
              1
-122.44098
-122.44096
-122.36903
              1
Length: 3664, dtype: int64
Parking Tax:
Parking Tax
0.0
      5046
        30
1.0
dtype: int64
Transient Occupancy Tax:
Transient Occupancy Tax
0.0
      4881
1.0
   195
dtype: int64
>>>
```