```
In [10]: import findspark
            findspark.init()
            from pyspark.sql import SparkSession
           spark=SparkSession.builder.appName("example").getOrCreate()
In [11]: import datetime
           users = [
                     "id": 1,
                     "first_name": "Corrie",
"last name": "Van den Oord",
                     "email": "cvandenoord0@etsy.com",
"gender": "male",
"current_city": "Dallas",
"phone numbers": Row(mobile="+1 234 567 8901", home="+1 234 567 8911"),
                     "courses": [1, 2],
"is_customer": True,
                     "amount_paid": 1000.55,
"customer_from": datetime.date(2021, 1, 15),
"last_updated_ts": datetime.datetime(2021, 2, 10, 1, 15, 0)
                     "first_name": "Nikolaus",
"last_name": "Brewitt",
                     "email": "nbrewittl@dailymail.co.uk",
"gander": "male"
In [12]: from pyspark.sql import Row
In [13]: df=spark.createDataFrame(users)
In [15]: df.show()
           --+----+
            |amount paid|courses| current city|customer from|
                                                                                          email|first name|gender| id|is customer| last na
           me| last_updated_ts| _ phone_numbers| +------+
           | 1000.55| [1, 2]| Dallas| 2021-01-15|cvandenoord0@etsy...| Corrie| male| 1| rd|2021-02-10 01:15:00|{+1 234 567 8901,...| | 900.0| [3]| Houston| 2021-02-14|nbrewitt1@dailyma...| Nikolaus| male| 2| tt|2021-02-18 03:33:00|{+1 234 567 8923,...| | 850.55| [2, 4]| | 2021-01-21|openney2@vistapri...| Orelie|female| 3| ey|2021-03-15 15:16:55|{+1 714 512 9752,...| null| amaddocks3@home.nl| Ashbyl male| 4|
                                                                                                                                     true|Van den Oo
                                                                                                                                     true|
                                                                                                                                                   Brewi
                                                                                                                                     truel
                                                                                                                                                   Penn
           ey|261-03-13 13:10:35(+1 714 312 9732,...|

| null| []|San Fransisco| null| amaddocks3@home.pl|

| ks|2021-04-10 17:45:30| {null, null}|

| null| []| null| null|krome4@shutterfly...|

me|2021-04-02 00:55:18|{+1 817 934 7142,...|
                                                                  'null| amaddocks3@home.pl| Ashby| male| 4|
                                                                                                                                     false|
                                                                                                                                                 Maddoc
                                                                                                        Kurt|female| 5|
                                                                                                                                     false
                                                                                                                                                      Ro
            --+-----
In [16]: df\
           .select("id","current_city")\
            .where('current_city IS NOT NULL')\
            .show()
            | id| current_city|
                         Dallas|
               2
                         Houston
               31
               4|San Fransisco
In [18]: from pyspark.sql.types import *
           from pyspark.sql.functions import *
In [19]: df.select("id","customer from").where(col("customer from").isNull()).show()
            +---+
            | id|customer from|
            +---+
                     null|
             5 İ
```

```
In [20]: df.select("id","current_city","customer_from").orderBy("customer_from").show()
        | id| current_city|customer_from|
           4|San Fransiscol
           5
                                 null
                    null
                  Dallas
                            2021-01-15
           1
           3
                  Houston
                            2021-02-14
In [21]: df.select("id","current_city","customer_from").orderBy("current_city").show()
        +---+
        | id| current_city|customer_from|
                    nulli
           51
                                 nulli
                            2021-01-21
                  Dallas
                            2021-01-15
                 Houston
                            2021-02-14
           4|San Fransisco|
                                null
In [22]: df.select('id','current_city','customer_from').orderBy(df.customer_from.desc()).show()
        +---+-----
        | id| current_city|customer_from|
                Houston| 2021-02-14|
          3 |
                            2021-01-21
                  Dallas
                          2021-01-15
          11
                             null
           4|San Fransisco|
          5| null|
                                 nullj
In [25]: df.select("id","current_city","customer_from").orderBy(desc_nulls_last("customer_from")).show()
        | id| current_city|customer_from|
          2|
                 Houston|
                            2021-02-14|
           з|
                            2021-01-21
                  Dallasi
          1 İ
                            2021-01-15
           4|San Fransisco|
                                 null
        | 5| null|
                                 nullj
In [27]: df1 = spark.read.option("header",True).csv("zipcode.csv")
In [28]: df1.show()
        | id|zipcode| type|
                                         city|state|population|
                704|STANDARD|
                                         null|
                                                        30100|
                                                 PR I
           1|
                704
                     null|PASEO COSTA DEL SUR|
                                                         null
           3 İ
                709
                       nullį
                               BDA SAN LUISİ
                                                 PR
                                                         3700
                     UNIQUE | CINGULAR WIRELESS |
              76166
                                                        84000
                                                 TX
           5
              76177 STANDARD
                                         nullj
                                                 TX
                                                         null
           1 i
                704 I STANDARD I
                                         nulli
                                                 PR
                                                        30100
                704 STANDARD
                                         null
                                                 PR
                                                        30100
           1
In [30]: df1.drop_duplicates().show()
        +---+-----
                                       city|state|population|
        | id|zipcode| type|
                704|STANDARD|
                                               PRI
                                         null|
          11
                                                        301001
              76177 STANDARD
                                                 TXİ
                                         nulli
                                                         null
              76166 UNIOUE CINGULAR WIRELESS TX
                                                        84000 İ
```

```
In [33]: df1.drop_duplicates(["zipcode"]).show()
                               +---+
                              | id|zipcode| type|
                                                                                                                                            city|state|population|
                                                          704|STANDARD|
                                                                                                                                             nulli
                                                                                                                                                                                                    301001
                                       11
                                                           709| null| BDA SAN LULI
709| null| CINGULAR WIRELESS|
null|
                                                                                                                     BDA SAN LUIS
                                                                                                                                                                                                       3700
                                        4 |
                                                  76166
                                                                                                                                                                          ΤXİ
                                                                                                                                                                                                     84000
                                                                                                                                                                    TX
                                                 76177|STANDARD| null|
                                        5İ
                                                                                                                                                                                                     nulli
In [34]: df1.dropna().show()
                               | id|zipcode| type| city|state|population|
                               | 4| 76166|UNIQUE|CINGULAR WIRELESS| TX|
In [35]: df1.dropna("all").show()
                               l idlzincodel tynel
                                                                                                                                                citylstatelnonulationl
In [39]: df1.na.drop(subset="type").show()
                               | id|zipcode| type|
                                                                                                                                          city|state|population|
                                                           704 | STANDARD I
                                                                                                                                               null|
                                                                                                                                                                                                     30100 l
                                                                                                                                                                          PR I
                                                    7617 | STANDARD | TOTAL |
7617 | STANDARD | TOTAL |
764 | STANDARD | TOTAL |
764 | STANDARD | TOTAL |
764 | STANDARD | TOTAL |
764 | STANDARD | TOTAL |
764 | STANDARD | TOTAL |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
764 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
765 | STANDARD |
                                                                                                                                                                          TX
                                                                                                                                                                                                     84000
                                                                                                                                              ELESS| IA|
null| TX|
null| PR|
null| PR|
                                                                                                                                                                                                      null
                                      1 j
                                                        704 I STANDARD I
                                                                                                                                                                                                     30100
                                                          704 STANDARD
                                                                                                                                                                                                     30100
In [40]: dfl.na.drop(subset="population").show()
                               +---+-----
                               | id|zipcode| type| city|state|popu
                                                                                                                            city|state|population|
                                                          704|STANDARD| null|
709| null| BDA SAN LUIS|
                                        1|
                                                                                                                                                                          PR I
                                                                                                                                                                                                     30100|
                                        3 i
                                                           709| null| BDA SAN LUIS|
5166| UNIQUE|CINGULAR WIRELESS|
                                                                                                                                                                          PR İ
                                                                                                                                                                                                      3700
                                                 76166
                                                                                                                                                                          TX
                                                                                                                                                                                                     84000
                                        1
                                                           704 STANDARD
                                                                                                                                              null
                                                                                                                                                                                                     30100
                                       1İ
                                                           704 | STANDARD |
                                                                                                                                               nulli
                                                                                                                                                                          PR İ
                                                                                                                                                                                                     30100
 In [41]: df1.na.drop(thresh=5).show()
                               +---+
                                                                                                                                    city|state|population|
                                                  704|STANDARD| NULL | 1704|STANDARD| NULL | 1709| NULL | BDA SAN LUIS | PR | 176166 | UNIQUE | CINGULAR WIRELESS | TX | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | PR | 1704|STANDARD | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | 
                                                                                                                                                                                                     30100
                                      1 İ
                                                                                                                                                                                                     30100
 In [42]: df1.na.fill(30000).show()
                                    id|zipcode| type|
                                                                                                                                               city|state|population|
                                                            704|STANDARD|
                                                                                                                                                      null|
                                                                                                                                                                                                           30100|
                                                           704| null|PASEO COSTA DEL SUR|
709| null| BDA SAN LUIS|
                                                                                                                                                                                                              null
3700
                                        2 |
                                                                                                                                                                                 PR
                                                                                                                                                                                 PR
                                        3 İ
                                                     76166
                                                                            UNIQUE | CINGULAR WIRELESS
                                                                                                                                                                                 TX
                                                                                                                                                                                                           84000
                                        5 j
                                                    76177 | STANDARD |
704 | STANDARD |
                                                                                                                                                      nullj
                                                                                                                                                                                тхі
                                                                                                                                                                                                           null
30100
                                                                                                                                                                                 PR
                                                                                                                                                      nulli
                                                            704 STANDARD
                                                                                                                                                      null
                                                                                                                                                                                PR
                                                                                                                                                                                                           30100
```

```
In [46]: df1.fillna("Mumbai", subset="city").show()
                                              city|state|population|
         | id|zipcode| type|
            1|
                  704 | STANDARD |
                                             Mumbai|
                                                       PR I
                                                                30100|
                       null|PASEO COSTA DEL SUR|
null| BDA SAN LUTSI
            2 j
3 j
                                                       PR İ
                  704
                                                                null
                  709
                                                       PR
                                                                3700
            4
                76166
                       UNIQUE
                                 CINGULAR WIRELESS
                                                        тхі
                                                                84000
                76177 | STANDARD |
            5 |
1 |
                                             Mumbail
                                                       TX
                                                                null
                  704 STANDARD
                                             Mumbai
                                                        PR
                                                                30100
            1
                  704 STANDARD
                                             Mumbai
                                                       PR
                                                                30100
In [50]: df1.fillna({"type":"VIP","city":"Mumbai","population":30000}).show()
         +---+
         | id|zipcode| type|
                                             city|state|population|
                  704|STANDARD|
                                            Mumbail
                                                       PR I
                                                                30100|
            11
                           VIP PASEO COSTA DEL SUR
            з ј
                  709 j
                           VIP
                                      BDA SAN LUIS
                                                      PR 
                                                                3700
In [51]: df1.show()
                                            -----+
          | id|zipcode| type|
                                              city|state|population|
                        STANDARD| null|
null|PASEO COSTA DEL SUR|
null| BDA SAN LUIS|
             1|
                   704|STANDARD|
                                                        PR I
                                                                30100|
             2
                   704 İ
                                                        PR i
                                                                 null
                   709
                                                        PR
                                                                 3700
                                 CINGULAR WIRELESS
                 76166 UNIQUE
76177 STANDARD
             4
                                                        ТХ
                                                                84000
             5 |
1 |
                                                nulli
                                                        TX
                                                                 null
                   704 STANDARD
                                                nullj
                                                                30100
             1
                   704 STANDARD
                                                nullj
                                                        PR
                                                                30100
In [53]: df1=df1.drop_duplicates()
In [55]: df.select("id","current city","customer from").replace("","LA").show()
          | id| current_city|customer_from|
          | 1|
                      Dallas| 2021-01-15|
 In [9]: df.groupBy("Year").count().sort(col("Year").desc()).show()
          |Year|count|
          |2014| 8362|
          2013 | 6158
          2012 6164
          |2011| 6216|
|2010| 6192|
          2009 6312
          2008 6481
          2007 6367
In [11]: df.write\
          .mode("overwrite")\
          .martitionBy("Year")\
.option("path","/home/labuser/1PySpark/Day 3/babynames")\
          .saveAsTable("babynames_year")
```

