In [1]:
```
from pyspark.sql import SparkSession
import pyspark.sql.functions as F
import pyspark.sql.types as T
import numpy as np
import pandas as pd
```

In [2]:
```
spark = SparkSession.builder.getOrCreate()
```

In [3]:
```
spark
```

Out[3]:

SparkSession - in-memory

SparkContext

Spark UI

Version	v3.1.2
Master	local[*]
AppName	pyspark-shell

Manipulating data

In [4]:
```
startup_data = spark.read.csv(
"Desktop/7082 CEM - Big Data/startup data.csv",
inferSchema=True,
header=True,)
```

In [5]:
```
startup_data.columns
```

Out[5]:
```
['Unnamed: 0',
 'state_code',
 'latitude',
 'longitude',
 'zip_code',
 'id',
 'city',
 'Unnamed: 6',
 'name',
 'labels',
 'founded_at',
 'closed_at',
 'first_funding_at',
 'last_funding_at',
 'age_first_funding_year',
 'age_last_funding_year',
 'age_first_milestone_year',
 'age_last_milestone_year',
 'relationships',
 'funding_rounds',
 'funding_total_usd',
 'milestones',
 'state_code.1',
 'is_CA',
 'is_NY',
 'is_MA',
 'is_TX',
 'is_otherstate',
 'category_code',
 'is_software',
 'is_web',
 'is_mobile',
```

```
 'is_enterprise',
 'is_advertising',
 'is_gamesvideo',
 'is_ecommerce',
 'is_biotech',
 'is_consulting',
 'is_othercategory',
 'object_id',
 'has_VC',
 'has_angel',
 'has_roundA',
 'has_roundB',
 'has_roundC',
 'has_roundD',
 'avg_participants',
 'is_top500',
 'status']
```

In [6]:
```
startup_data1 = startup_data.dropDuplicates()
```

In [7]:
```
startup_data1.count()
```

Out[7]:
```
923
```

In [8]:
```
startup_data2 = startup_data.drop('Unnamed: 0','id','Unnamed: 6','labels','st
                                  'is_MA','is_TX','is_otherstate','is_softwar
                                  'is_advertising', 'is_gamesvideo','is_ecomm
                                  'is_othercategory', 'object_id')
```

In [9]:
```
startup_data2
```

Out[9]:
```
DataFrame[state_code: string, latitude: double, longitude: double, zip_code: s
tring, city: string, name: string, founded_at: string, closed_at: string, firs
t_funding_at: string, last_funding_at: string, age_first_funding_year: double,
age_last_funding_year: double, age_first_milestone_year: double, age_last_mile
stone_year: double, relationships: int, funding_rounds: int, funding_total_us
d: bigint, milestones: int, category_code: string, has_VC: int, has_angel: in
t, has_roundA: int, has_roundB: int, has_roundC: int, has_roundD: int, avg_par
ticipants: double, is_top500: int, status: string]
```

Handling missing values

In [10]:
```
startup_data2_sel = startup_data1.select('state_code','latitude','longitude',
 'closed_at','first_funding_at','last_funding_at','age_first_funding_year','a
 'age_first_milestone_year','age_last_milestone_year','relationships','fundin
 'milestones','category_code','has_VC','has_angel','has_roundA','has_roundB',
 'avg_participants','is_top500','status')
```

Findind missing values in the data

In [11]:
```
from pyspark.sql.functions import isnan, when, count, col
startup_data2_sel.select([count(when(isnan(c) | col(c).isNull(), c)).alias(c)
```

```
+----------+--------+---------+----+----+----------+---------+---------------
+---------------+----------------------+--------------------+---------------
--------+-----------------------+-------------+--------------+---------------
-+----------+-------------+------+---------+----------+----------+----------+-
---------+---------------+---------+------+
|state_code|latitude|longitude|city|name|founded_at|closed_at|first_funding_at
|last_funding_at|age_first_funding_year|age_last_funding_year|age_first_milest
one_year|age_last_milestone_year|relationships|funding_rounds|funding_total_us
d|milestones|category_code|has_VC|has_angel|has_roundA|has_roundB|has_roundC|h
as_roundD|avg_participants|is_top500|status|
+----------+--------+---------+----+----+----------+---------+---------------
```

```
+---------------+----------------------+--------------------+----------------
--------+----------------------+-------------+-------------+----------------
-+----------+-------------+------+---------+----------+----------+----------+-
---------+----------------+---------+------+
|          0|       0|         0|   0|   0|          0|       588|                 0
|              0|                      0|                    0|
152|                    152|            0|             0|                  0|
0|            0|     0|         0|          0|          0|          0|         0|
0|         0|     0|
+----------+--------+---------+----+----+----------+---------+----------------
+---------------+----------------------+--------------------+----------------
--------+----------------------+-------------+-------------+----------------
-+----------+-------------+------+---------+----------+----------+----------+-
---------+----------------+---------+------+
```

Three features has have nulls

In [12]:
```
startup_data2_sel.describe("age_first_milestone_year","age_last_milestone_yea
```

```
+-------+-----------------------+-----------------------+---------+
|summary|age_first_milestone_year|age_last_milestone_year|closed_at|
+-------+-----------------------+-----------------------+---------+
|  count|                    771|                    771|      335|
|   mean|      3.0553530479896254|     4.7544225680933865|     null|
| stddev|      2.9770571428977237|     3.2121071562092283|     null|
|    min|                -14.1699|                -7.0055| 1/1/2001|
|    max|                 24.6849|                24.6849| 9/8/2013|
+-------+-----------------------+-----------------------+---------+
```

In [13]:
```
# replacing empty data in 'close_at' - string data type feature with n/a
df_Startup = startup_data2_sel.na.fill('n/a','closed_at')
```

In [14]:
```
df_Startup.show()
```

```
+----------+------------------+-------------------+-------------+-------------
-------+----------+---------+----------------+---------------+----------------
------+----------------------+-----------------------+----------------------+
-------------+--------------+-----------------+----------+----------------+---
---+---------+----------+----------+----------+----------+----------------+---
------+--------+
|state_code|          latitude|          longitude|         city|
name|founded_at|closed_at|first_funding_at|last_funding_at|age_first_funding_y
ear|age_last_funding_year|age_first_milestone_year|age_last_milestone_year|rel
ationships|funding_rounds|funding_total_usd|milestones|   category_code|has_VC
|has_angel|has_roundA|has_roundB|has_roundC|has_roundD|avg_participants|is_top
500|  status|
+----------+------------------+-------------------+-------------+-------------
-------+----------+---------+----------------+---------------+----------------
------+----------------------+-----------------------+----------------------+
-------------+--------------+-----------------+----------+----------------+---
---+---------+----------+----------+----------+----------+----------------+---
------+--------+
|        CA|         33.708708|        -117.852069|    Santa Ana|
Mophie| 5/24/2005|      n/a|        8/1/2006|       8/1/2006|
1.189|                 1.189|                  2.189|                 7.7753|
5|             1|          1000000|         3|        hardware|     1|
0|         0|         0|         0|         0|             2.0|        1|acqui
red|
|        CA|          37.54827|       -121.988572|      Fremont|  Mendocino S
oftware|  1/1/2003| 2/7/2008|       9/12/2005|       1/1/2007|
2.6986|                 4.0027|                   1.0|                   1.0|
2|             2|         19700000|         1|        software|     1|
0|         0|         1|         0|         0|             4.0|        1|  clo
sed|
|        GA|         33.956215|        -83.987962|Lawrenceville|  Nalace Corp
oration|  1/1/2012| 5/1/2013|       4/19/2012|      4/19/2012|
0.2986|                 0.2986|                  0.2986|                 0.2986|
```

```
1|            1|           19000|         1|       ecommerce|     0|
1|         0|         0|         0|         0|             1.0|        0|  clo
sed|
|        CA|          37.36883|        -122.03635|    Sunnyvale|           Fl
ashSoft|11/25/2009|      n/a|       6/28/2011|      6/28/2011|
1.589|               1.589|                1.1863|                2.5452|
6|            1|         3000000|         2|        software|     0|
0|         1|         0|         0|         0|             2.0|        1|acqui
red|
|        CA|        59.3352318|        18.0571206|  Santa Clara|           Xe
lerated|  1/1/2000|      n/a|       4/20/2005|      8/14/2008|
5.3041|               8.6247|                null|                null|
4|            3|        53000000|         0|   semiconductor|     0|
0|         0|         0|         1|         1|          4.6667|        1|acqui
red|
|        NY|         40.750519|        -73.993494|     New York|           P
anvidea|  1/1/2007|      n/a|       3/19/2010|      3/19/2010|
3.2137|               3.2137|                3.0027|                4.6219|
5|            1|         2700000|         2|     games_video|     0|
0|         1|         0|         0|         0|             2.0|        1|acqui
red|
|        WA|        30.6324797|        -86.9843446|     Kirkland|           Cl
earwire| 10/1/2003|      n/a|        5/6/2008|      2/27/2013|
4.6|              9.4164|                5.2575|                9.7753|
19|             4|       5700000000|         2|          mobile|     0|
0|         0|         0|         0|         0|            4.25|        1|acqui
red|
|        MA|          37.09024|        -95.712891|    Cambridge|        N2N C
ommerce|  1/1/2006| 1/1/2008|        1/3/2008|       1/3/2008|
2.0055|               2.0055|                 0.0|                 0.0|
1|            1|         30000000|         1|       ecommerce|     1|
0|         0|         0|         0|         0|             2.0|        1|acqui
red|
|        MA|         42.504817|        -71.195611|   Burlington|
Certeon|  1/1/2003| 2/1/2013|       7/16/2007|       5/1/2012|
4.5397|                9.337|                2.0027|                2.0027|
2|            2|         19000000|         1|        software|     1|
0|         0|         1|         0|         0|             3.5|        1|  clo
sed|
|        CA|         37.406914|        -122.09037|Mountain View|
Ardian|  1/1/2003|      n/a|       1/1/2005|      3/23/2009|
2.0027|               6.2274|                4.0027|               10.7616|
8|            3|         64080000|         2|         biotech|     0|
0|         1|         1|         1|         0|             4.0|        1|acqui
red|
|        WA|        47.6399006|       -122.1914274|      Seattle|           Li
vemocha|  9/1/2007|      n/a|        1/1/2008|      2/25/2013|
0.3342|               5.4904|                2.7507|                 5.589|
14|             5|         19389998|         2|       education|     1|
0|         1|         1|         1|         0|             1.5|        1|acqui
red|
|        CA|32.988246999999994|        -117.080769|    San Diego|Sequoia Commu
nica...|  1/1/2000| 1/1/2011|       11/9/2004|       6/2/2009|
4.8603|               9.4247|                null|                null|
3|            7|         74000000|         0|   semiconductor|     1|
0|         0|         0|         1|         1|             2.0|        1|  clo
sed|
|        CA|          37.552262|        -122.292146|    San Mateo|      Servo S
oftware|  1/1/2008|      n/a|      12/15/2009|     12/15/2009|
1.9562|               1.9562|                1.9562|                2.0027|
11|             1|          3011408|         3|        software|     0|
0|         0|         1|         0|         0|             1.0|        1|acqui
red|
|        CA|        37.7611016|       -122.4160008|San Francisco|           W
eAre.Us| 12/1/2007|4/23/2012|        8/1/2008|       8/1/2008|
0.6685|               0.6685|                -0.9151|                 0.863|
7|            1|           50000|         2|     advertising|     0|
1|         0|         0|         0|         0|             3.0|        0|  clo
sed|
```

```
|        CA|         37.780883|        -122.395257|San Francisco|
Reddit|  1/1/2005|      n/a|        6/1/2005|       6/1/2005|
0.4137|               0.4137|                6.8329|                   8.874|
9|             1|           100000|         3|             web|     0|
1|         0|         0|         0|         0|             1.0|        1|acqui
red|
|        TX|         32.997114|         -96.676137|   Richardson|          App
Trigger|  1/1/2002|      n/a|      12/13/2007|     12/13/2007|
5.9507|               5.9507|                  null|                    null|
1|             1|         21500000|         0|public_relations|     0|
0|         0|         0|         1|         0|             4.0|        1|acqui
red|
|        CA|         37.417002|-122.07871000000002|Mountain View|       Kace n
etworks|  1/1/2003|      n/a|       7/25/2005|      5/22/2006|
2.5644|                3.389|                8.1452|                   8.1452|
7|             2|         11000000|         1|        software|     1|
0|         0|         1|         0|         0|             1.5|        1|acqui
red|
|        TX|         32.960431|          -96.83026|      Addison|          Chr
onicity|  9/1/2006| 4/1/2012|      10/10/2006|      5/26/2011|
0.1068|               4.7342|                  null|                    null|
0|             4|         19550000|         0|      consulting|     1|
0|         1|         0|         0|         0|             1.0|        1|  clo
sed|
|        CA|         37.445586|        -122.161929|    Palo Alto|       Partic
le Code|  5/1/2010|      n/a|       6/30/2010|      6/30/2010|
0.1644|               0.1644|                1.4822|                   1.4822|
3|             1|          3000000|         1|        software|     1|
0|         0|         0|         0|         0|             1.0|        0|acqui
red|
|        MA|         42.375518| -71.27229200000002|      Waltham|      Navic N
etworks|  1/3/2000|      n/a|       2/14/2000|      2/26/2001|
0.1151|               1.1507|                  null|                    null|
7|             3|         42000000|         0|     advertising|     0|
0|         1|         1|         1|         0|          2.6667|        1|acqui
red|
+----------+------------------+-------------------+-------------+-------------
-------+----------+---------+----------------+---------------+----------------
------+---------------------+-------------------------+-----------------------+
-------------+--------------+------------------+----------+----------------+---
---+---------+----------+----------+----------+----------+----------------+---
------+--------+
only showing top 20 rows
```

Impuing 2 other features (double data type)

In [15]:
```
from pyspark.ml.feature import Imputer

Imputer = Imputer(
    inputCols=["age_first_milestone_year","age_last_milestone_year"],
    outputCols=["{}_imputed".format(c) for c in ["age_first_milestone_year","
    ).setStrategy("mean")
```

In [16]:
```
Ndf_startup = Imputer.fit(df_Startup).transform(df_Startup)
```

In [17]:
```
Ndf_startup.show()
```

```
+----------+------------------+-------------------+-------------+-------------
-------+----------+---------+----------------+---------------+----------------
------+---------------------+-------------------------+-----------------------+
-------------+--------------+------------------+----------+----------------+---
---+---------+----------+----------+----------+----------+----------------+---
------+--------+-------------------------------+---------------------------
--+
|state_code|          latitude|          longitude|         city|
name|founded_at|closed_at|first_funding_at|last_funding_at|age_first_funding_y
ear|age_last_funding_year|age_first_milestone_year|age_last_milestone_year|rel
```

```
ationships|funding_rounds|funding_total_usd|milestones|   category_code|has_VC
|has_angel|has_roundA|has_roundB|has_roundC|has_roundD|avg_participants|is_top
500|  status|age_first_milestone_year_imputed|age_last_milestone_year_imputed|
+----------+-----------------+------------------+------------+------------
-------+---------+---------+---------------+--------------+---------------
------+---------------------+-------------------------+-----------------------+
-------------+--------------+-----------------+----------+----------------+---
---+---------+----------+----------+----------+----------+----------------+---
------+--------+---------------------------------+-----------------------------
--+
|        CA|         33.708708|        -117.852069|    Santa Ana|
Mophie| 5/24/2005|      n/a|        8/1/2006|       8/1/2006|
1.189|                1.189|                   2.189|                  7.7753|
5|             1|          1000000|         3|        hardware|     1|
0|         0|         0|         0|         0|              2.0|        1|acqui
red|                           2.189|                          7.7753|
|        CA|          37.54827|        -121.988572|      Fremont|  Mendocino S
oftware|  1/1/2003| 2/7/2008|       9/12/2005|       1/1/2007|
2.6986|                4.0027|                      1.0|                     1.0|
2|             2|         19700000|         1|        software|     1|
0|         0|         1|         0|         0|              4.0|        1|  clo
sed|                           1.0|                         1.0|
|        GA|         33.956215|         -83.987962|Lawrenceville|  Nalace Corp
oration|  1/1/2012| 5/1/2013|       4/19/2012|      4/19/2012|
0.2986|                0.2986|                   0.2986|                  0.2986|
1|             1|            19000|         1|       ecommerce|     0|
1|         0|         0|         0|         0|              1.0|        0|  clo
sed|                           0.2986|                         0.2986|
|        CA|          37.36883|         -122.03635|    Sunnyvale|           Fl
ashSoft|11/25/2009|      n/a|       6/28/2011|      6/28/2011|
1.589|                1.589|                   1.1863|                  2.5452|
6|             1|          3000000|         2|        software|     0|
0|         1|         0|         0|         0|              2.0|        1|acqui
red|                           1.1863|                          2.5452|
|        CA|        59.3352318|         18.0571206|  Santa Clara|           Xe
lerated|  1/1/2000|      n/a|       4/20/2005|      8/14/2008|
5.3041|                8.6247|                      null|                    null|
4|             3|         53000000|         0|   semiconductor|     0|
0|         0|         0|         1|         1|           4.6667|        1|acqui
red|              3.0553530479896254|             4.7544225680933865|
|        NY|         40.750519|         -73.993494|     New York|            P
anvidea|  1/1/2007|      n/a|       3/19/2010|      3/19/2010|
3.2137|                3.2137|                   3.0027|                  4.6219|
5|             1|          2700000|         2|     games_video|     0|
0|         1|         0|         0|         0|              2.0|        1|acqui
red|                           3.0027|                          4.6219|
|        WA|        30.6324797|        -86.9843446|     Kirkland|           Cl
earwire| 10/1/2003|      n/a|        5/6/2008|      2/27/2013|
4.6|               9.4164|                     5.2575|                  9.7753|
19|             4|       5700000000|         2|          mobile|     0|
0|         0|         0|         0|         0|             4.25|        1|acqui
red|                           5.2575|                          9.7753|
|        MA|          37.09024|         -95.712891|    Cambridge|        N2N C
ommerce|  1/1/2006| 1/1/2008|        1/3/2008|       1/3/2008|
2.0055|                2.0055|                      0.0|                     0.0|
1|             1|         30000000|         1|       ecommerce|     1|
0|         0|         0|         0|         0|              2.0|        1|acqui
red|                           0.0|                         0.0|
|        MA|         42.504817|         -71.195611|   Burlington|
Certeon|  1/1/2003| 2/1/2013|       7/16/2007|       5/1/2012|
4.5397|                 9.337|                   2.0027|                  2.0027|
2|             2|         19000000|         1|        software|     1|
0|         0|         1|         0|         0|              3.5|        1|  clo
sed|                           2.0027|                          2.0027|
|        CA|         37.406914|         -122.09037|Mountain View|
Ardian|  1/1/2003|      n/a|        1/1/2005|      3/23/2009|
2.0027|                6.2274|                   4.0027|                 10.7616|
8|             3|         64080000|         2|         biotech|     0|
0|         1|         1|         1|         0|              4.0|        1|acqui
```

```
red|                          4.0027|                         10.7616|
|        WA|        47.6399006|       -122.1914274|      Seattle|           Li
vemocha|  9/1/2007|      n/a|        1/1/2008|      2/25/2013|
0.3342|               5.4904|                2.7507|                   5.589|
14|             5|         19389998|         2|       education|     1|
0|         1|         1|         1|         0|             1.5|        1|acqui
red|                          2.7507|                           5.589|
|        CA|32.988246999999994|        -117.080769|    San Diego|Sequoia Commu
nica...|  1/1/2000| 1/1/2011|       11/9/2004|       6/2/2009|
4.8603|               9.4247|                  null|                   null|
3|             7|         74000000|         0|   semiconductor|     1|
0|         0|         0|         1|         1|             2.0|        1|  clo
sed|             3.0553530479896254|             4.7544225680933865|
|        CA|         37.552262|        -122.292146|    San Mateo|      Servo S
oftware|  1/1/2008|      n/a|      12/15/2009|     12/15/2009|
1.9562|               1.9562|                1.9562|                   2.0027|
11|             1|          3011408|         3|        software|     0|
0|         0|         1|         0|         0|             1.0|        1|acqui
red|                          1.9562|                           2.0027|
|        CA|        37.7611016|       -122.4160008|San Francisco|            W
eAre.Us| 12/1/2007|4/23/2012|        8/1/2008|       8/1/2008|
0.6685|               0.6685|                -0.9151|                    0.863|
7|             1|            50000|         2|     advertising|     0|
1|         0|         0|         0|         0|             3.0|        0|  clo
sed|                         -0.9151|                            0.863|
|        CA|         37.780883|        -122.395257|San Francisco|
Reddit|  1/1/2005|      n/a|       6/1/2005|       6/1/2005|
0.4137|               0.4137|                6.8329|                   8.874|
9|             1|           100000|         3|             web|     0|
1|         0|         0|         0|         0|             1.0|        1|acqui
red|                          6.8329|                           8.874|
|        TX|         32.997114|         -96.676137|   Richardson|          App
Trigger|  1/1/2002|      n/a|      12/13/2007|     12/13/2007|
5.9507|               5.9507|                  null|                   null|
1|             1|         21500000|         0|public_relations|     0|
0|         0|         0|         1|         0|             4.0|        1|acqui
red|             3.0553530479896254|             4.7544225680933865|
|        CA|         37.417002|-122.07871000000002|Mountain View|       Kace n
etworks|  1/1/2003|      n/a|       7/25/2005|      5/22/2006|
2.5644|                3.389|                8.1452|                   8.1452|
7|             2|         11000000|         1|        software|     1|
0|         0|         1|         0|         0|             1.5|        1|acqui
red|                          8.1452|                           8.1452|
|        TX|         32.960431|          -96.83026|      Addison|          Chr
onicity|  9/1/2006| 4/1/2012|      10/10/2006|      5/26/2011|
0.1068|               4.7342|                  null|                   null|
0|             4|         19550000|         0|      consulting|     1|
0|         1|         0|         0|         0|             1.0|        1|  clo
sed|             3.0553530479896254|             4.7544225680933865|
|        CA|         37.445586|        -122.161929|    Palo Alto|       Partic
le Code|  5/1/2010|      n/a|       6/30/2010|      6/30/2010|
0.1644|               0.1644|                1.4822|                   1.4822|
3|             1|          3000000|         1|        software|     1|
0|         0|         0|         0|         0|             1.0|        0|acqui
red|                          1.4822|                           1.4822|
|        MA|         42.375518| -71.27229200000002|      Waltham|      Navic N
etworks|  1/3/2000|      n/a|       2/14/2000|      2/26/2001|
0.1151|               1.1507|                  null|                   null|
7|             3|         42000000|         0|     advertising|     0|
0|         1|         1|         1|         0|          2.6667|        1|acqui
red|             3.0553530479896254|             4.7544225680933865|
+----------+------------------+--------------------+-------------+-------------
-------+----------+---------+---------------+---------------+---------------
------+---------------------+-----------------------+----------------------+
-------------+--------------+-----------------+----------+----------------+---
---+---------+----------+----------+----------+----------+----------------+---
------+--------+---------------------------------+-----------------------------
--+
```

only showing top 20 rows

In [18]:
```
Ndf_startup.columns
```

Out[18]:
```
['state_code',
 'latitude',
 'longitude',
 'city',
 'name',
 'founded_at',
 'closed_at',
 'first_funding_at',
 'last_funding_at',
 'age_first_funding_year',
 'age_last_funding_year',
 'age_first_milestone_year',
 'age_last_milestone_year',
 'relationships',
 'funding_rounds',
 'funding_total_usd',
 'milestones',
 'category_code',
 'has_VC',
 'has_angel',
 'has_roundA',
 'has_roundB',
 'has_roundC',
 'has_roundD',
 'avg_participants',
 'is_top500',
 'status',
 'age_first_milestone_year_imputed',
 'age_last_milestone_year_imputed']
```

In [19]:
```
Ndf_startup_sel = Ndf_startup.select('age_first_milestone_year','age_last_mil
 'age_first_milestone_year_imputed','age_last_milestone_year_imputed')
```

In [20]:
```
Ndf_startup_sel.describe('age_first_milestone_year','age_first_milestone_year
```

```
+-------+------------------------+--------------------------------+
|summary|age_first_milestone_year|age_first_milestone_year_imputed|
+-------+------------------------+--------------------------------+
|  count|                     771|                             923|
|   mean|      3.0553530479896254|              3.055353047989626|
| stddev|      2.9770571428977237|             2.7206149036792846|
|    min|                -14.1699|                        -14.1699|
|    max|                 24.6849|                         24.6849|
+-------+------------------------+--------------------------------+
```

In [21]:
```
Ndf_startup_sel.describe('age_last_milestone_year','age_first_milestone_year_
```

```
+-------+-----------------------+--------------------------------+
|summary|age_last_milestone_year|age_first_milestone_year_imputed|
+-------+-----------------------+--------------------------------+
|  count|                    771|                             923|
|   mean|     4.7544225680933865|              3.055353047989626|
| stddev|     3.2121071562092283|             2.7206149036792846|
|    min|                -7.0055|                        -14.1699|
|    max|                24.6849|                         24.6849|
+-------+-----------------------+--------------------------------+
```

In [22]:
```
Cstartup_data = Ndf_startup.drop('age_first_milestone_year','age_last_milesto
```

Ensuring there are no nulls in the data left

In [23]:
```
Cstartup_data.columns
```

Out[23]:
```
['state_code',
 'latitude',
 'longitude',
 'city',
 'name',
 'founded_at',
 'closed_at',
 'first_funding_at',
 'last_funding_at',
 'age_first_funding_year',
 'age_last_funding_year',
 'relationships',
 'funding_rounds',
 'funding_total_usd',
 'milestones',
 'category_code',
 'has_VC',
 'has_angel',
 'has_roundA',
 'has_roundB',
 'has_roundC',
 'has_roundD',
 'avg_participants',
 'is_top500',
 'status',
 'age_first_milestone_year_imputed',
 'age_last_milestone_year_imputed']
```

In [24]:
```
Cstartup_data_sel = Cstartup_data.select('state_code','latitude','longitude',
'first_funding_at','last_funding_at','age_first_funding_year','age_last_fundi
'funding_total_usd','milestones','category_code','has_VC','has_angel','has_ro
'avg_participants','is_top500','status','age_first_milestone_year_imputed','a
```

In [25]:
```
Cstartup_data_sel.select([count(when(isnan(c) | col(c).isNull(), c)).alias(c)
```

```
+----------+--------+---------+----+----+----------+---------+----------------
+---------------+----------------------+---------------------+-------------+--
------------+-----------------+----------+-------------+------+---------+-----
-----+----------+----------+----------+----------------+---------+------+-----
---------------------------+--------------------------------+
|state_code|latitude|longitude|city|name|founded_at|closed_at|first_funding_at
|last_funding_at|age_first_funding_year|age_last_funding_year|relationships|fu
nding_rounds|funding_total_usd|milestones|category_code|has_VC|has_angel|has_r
oundA|has_roundB|has_roundC|has_roundD|avg_participants|is_top500|status|age_f
irst_milestone_year_imputed|age_last_milestone_year_imputed|
+----------+--------+---------+----+----+----------+---------+----------------
+---------------+----------------------+---------------------+-------------+--
------------+-----------------+----------+-------------+------+---------+-----
-----+----------+----------+----------+----------------+---------+------+-----
---------------------------+--------------------------------+
|         0|       0|        0|   0|   0|         0|        0|               0
|              0|                      0|                     0|            0|
0|                 0|         0|            0|     0|        0|          0|
0|          0|         0|                0|        0|     0|
0|                              0|
+----------+--------+---------+----+----+----------+---------+----------------
+---------------+----------------------+---------------------+-------------+--
------------+-----------------+----------+-------------+------+---------+-----
-----+----------+----------+----------+----------------+---------+------+-----
---------------------------+--------------------------------+
```

In [26]:
```
Cstartup_data_sel.printSchema()
```

```
root
 |-- state_code: string (nullable = true)
 |-- latitude: double (nullable = true)
```

```
 |-- longitude: double (nullable = true)
 |-- city: string (nullable = true)
 |-- name: string (nullable = true)
 |-- founded_at: string (nullable = true)
 |-- closed_at: string (nullable = false)
 |-- first_funding_at: string (nullable = true)
 |-- last_funding_at: string (nullable = true)
 |-- age_first_funding_year: double (nullable = true)
 |-- age_last_funding_year: double (nullable = true)
 |-- relationships: integer (nullable = true)
 |-- funding_rounds: integer (nullable = true)
 |-- funding_total_usd: long (nullable = true)
 |-- milestones: integer (nullable = true)
 |-- category_code: string (nullable = true)
 |-- has_VC: integer (nullable = true)
 |-- has_angel: integer (nullable = true)
 |-- has_roundA: integer (nullable = true)
 |-- has_roundB: integer (nullable = true)
 |-- has_roundC: integer (nullable = true)
 |-- has_roundD: integer (nullable = true)
 |-- avg_participants: double (nullable = true)
 |-- is_top500: integer (nullable = true)
 |-- status: string (nullable = true)
 |-- age_first_milestone_year_imputed: double (nullable = true)
 |-- age_last_milestone_year_imputed: double (nullable = true)
```

Explorative Data Analysis

In [27]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [28]:
```
Cstartup_data.describe().show()
```

```
+-------+----------+-----------------+-------------------+-------+--------+---
-------+---------+----------------+---------------+----------------------+----
-----------------+-----------------+------------------+--------------------+--
----------------+-------------+------------------+--------------------+--------
----------+--------------------+--------------------+--------------------+-------
-----------+------------------+--------+--------------------------------+-----
--------------------------+
|summary|state_code|         latitude|          longitude|   city|    name|fou
nded_at|closed_at|first_funding_at|last_funding_at|age_first_funding_year|age_
last_funding_year|    relationships|    funding_rounds|   funding_total_usd|
milestones|category_code|            has_VC|          has_angel|        has_ro
undA|         has_roundB|         has_roundC|         has_roundD|  avg_partici
pants|         is_top500|  status|age_first_milestone_year_imputed|age_last_mi
lestone_year_imputed|
+-------+----------+-----------------+-------------------+-------+--------+---
-------+---------+----------------+---------------+----------------------+----
-----------------+-----------------+------------------+--------------------+--
----------------+-------------+------------------+--------------------+--------
----------+--------------------+--------------------+--------------------+-------
-----------+------------------+--------+--------------------------------+-----
--------------------------+
|  count|       923|              923|                923|    923|     923|
923|      923|            923|            923|                   923|
923|              923|               923|                923|                 9
23|          923|               923|                923|                923|
923|                923|                923|                923|
923|     923|                                923|                            923|
|   mean|      null|38.51744208819068|-103.53921224312029|   null|    null|
null|     null|            null|           null|     2.235630010834236|   3.93
14557963163588|7.710725893824486|2.3109425785482123| 2.541974909209101E7|1.841
8201516793067|         null|0.3261105092091008|0.25460455037919827|0.508125677
```

```
1397616|0.39219934994582883|0.23293607800650054|0.09967497291440953|2.83858613
21776837|  0.80931744312026|    null|               3.055353047989626|
4.754422568093385|
| stddev|      null|3.741497102584239| 22.394166739935237|   null|    null|
null|     null|            null|           null|      2.51044853951302|   2.96
79098466072684|7.265776000021265|1.3909217125220041|1.8963436448879448E8|1.322
6320231659667|         null|0.4690424277763106|0.43587487838886535|0.500205009
4043888| 0.4885054405250472| 0.4229310277822581| 0.2997285512560348|1.87460094
94380236|0.3930522835477896|    null|              2.7206149036792846|
2.935417824359152|
|    min|        AR|        25.752358|       -122.7569561|  Acton|#waywire|
1/1/1984| 1/1/2001|        1/1/2000|       1/1/2001|               -9.0466|
-9.0466|                  0|                  1|               11000|
0|  advertising|                  0|                  0|                  0|
0|                  0|                  0|                1.0|
0|acquired|                -14.1699|                        -7.0055|
|    max|        WV|        59.3352318|         18.0571206|Zeeland|   xkoto|
9/9/2010|      n/a|        9/9/2009|       9/9/2011|               21.8959|
21.8959|                63|                10|          5700000000|
8|          web|                  1|                  1|                  1|
1|                  1|                  1|              16.0|
1|  closed|                24.6849|                        24.6849|
+-------+----------+-----------------+-------------------+-------+--------+---
-------+---------+----------------+---------------+----------------------+----
-----------------+-----------------+------------------+--------------------+--
----------------+-------------+------------------+------------------+--------
----------+--------------------+--------------------+------------------+-------
----------+--------------------+--------+------------------------------+-----
-------------------------+
```

In [29]:
```
num_cols = ['age_first_funding_year','age_last_funding_year', 'relationships'
Cstartup_data.select(num_cols).describe().show()
```

```
+-------+----------------------+---------------------+-----------------+------
------------+
|summary|age_first_funding_year|age_last_funding_year|    relationships|    fu
nding_rounds|
+-------+----------------------+---------------------+-----------------+------
------------+
|  count|                   923|                  923|              923|
923|
|   mean|     2.235630010834236|   3.9314557963163588|7.710725893824486|2.3109
425785482123|
| stddev|      2.51044853951302|   2.9679098466072684|7.265776000021265|1.3909
217125220041|
|    min|               -9.0466|              -9.0466|                0|
1|
|    max|               21.8959|              21.8959|               63|
10|
+-------+----------------------+---------------------+-----------------+------
------------+
```

In [30]:
```
Cstartup_data_sel = Cstartup_data.select('age_first_funding_year', 'age_last_
                'age_first_milestone_year_imputed', 'age_last_milestone_year_
                'milestones','is_top500','avg_participants','funding_rounds',
                'category_code','has_VC', 'state_code', 'latitude','longitude
```

In [31]:
```
Cstartup_data_sel.describe('age_first_funding_year', 'age_last_funding_year',
                'age_first_milestone_year_imputed', 'age_last_milestone_year_
                'milestones','avg_participants','funding_rounds','has_angel',
                'status','category_code','has_VC').show()
```

```
+-------+----------------------+---------------------+-----------------------
--------+-------------------------------+--------------------+---------------
-+------------------+------------------+------------------+------------------
+--------+-------------+------------------+
|summary|age_first_funding_year|age_last_funding_year|age_first_milestone_year
```

```
_imputed|age_last_milestone_year_imputed|   funding_total_usd|    relationship
s|        milestones|  avg_participants|    funding_rounds|          has_angel
|  status|category_code|            has_VC|
+-------+---------------------+--------------------+----------------------
-------+-------------------------------+-------------------+---------------
-+-----------------+-----------------+-----------------+------------------
+--------+-------------+------------------+
|  count|                  923|                 923|
923|                           923|                923|                923|
923|               923|               923|               923|     923|
923|               923|
|   mean|     2.235630010834236|   3.9314557963163588|               3.0553530
47989626|              4.754422568093385| 2.541974909209101E7|7.71072589382448
6|1.8418201516793067|2.8385861321776837|2.3109425785482123|0.25460455037919827
|    null|         null|0.3261105092091008|
| stddev|      2.51044853951302|   2.9679098466072684|              2.72061490
36792846|              2.935417824359152|1.8963436448879448E8|7.26577600002126
5|1.3226320231659667|1.8746009494380236|1.3909217125220041|0.43587487838886535
|    null|         null|0.4690424277763106|
|    min|               -9.0466|             -9.0466|
-14.1699|                        -7.0055|               11000|
0|                  0|                1.0|                1|                  0
|acquired|  advertising|                  0|
|    max|               21.8959|             21.8959|
24.6849|                        24.6849|          5700000000|               63
|                 8|                16.0|               10|                  1|
closed|          web|                  1|
+-------+---------------------+--------------------+----------------------
-------+-------------------------------+-------------------+---------------
-+-----------------+-----------------+-----------------+------------------
+--------+-------------+------------------+
```

In [32]:
```
Cstartup_data_sel.stat.corr('age_first_funding_year','age_first_milestone_yea
```

Out[32]:
```
0.49620450175394337
```

In [33]:
```
Cstartup_data_sel.stat.corr('age_last_funding_year','age_last_milestone_year_
```

Out[33]:
```
0.560272743212842
```

In [34]:
```
Cstartup_data_sel.stat.crosstab('status','funding_rounds').show()
```

```
+---------------------+---+---+---+---+---+---+---+---+---+
|status_funding_rounds|  1| 10|  2|  3|  4|  5|  6|  7|  8|
+---------------------+---+---+---+---+---+---+---+---+---+
|             acquired|158|  1|179|135| 73| 32| 11|  7|  1|
|               closed|159|  0|101| 32| 17|  8|  2|  6|  1|
+---------------------+---+---+---+---+---+---+---+---+---+
```

In [35]:
```
Cstartup_data_sel.stat.crosstab('status', 'has_VC').show()
```

```
+-------------+---+---+
|status_has_VC|  0|  1|
+-------------+---+---+
|     acquired|414|183|
|       closed|208|118|
+-------------+---+---+
```

In [36]:
```
Cstartup_data_sel.stat.crosstab('status', 'category_code').show()
```

```
+--------------------+----------+---------+----------+-------+---------+-----
-----+---------+---------+----------+-------+-------+----------+--------+----
--+-----------+-------------+-------+---------+------+-----+---------------+--
--+-----+-----------+----------------+-----------+------+--------+-----------
-+------+--------+------+-------------+------+---+
```

```
|status_category_code|advertising|analytics|automotive|biotech|cleantech|consu
lting|ecommerce|education|enterprise|fashion|finance|games_video|hardware|heal
th|hospitality|manufacturing|medical|messaging|mobile|music|network_hosting|ne
ws|other|photo_video|public_relations|real_estate|search|security|semiconducto
r|social|software|sports|transportation|travel|web|
+--------------------+-----------+---------+----------+-------+---------+-----
-----+---------+---------+----------+-------+-------+-----------+--------+----
--+-----------+-------------+-------+---------+------+-----+---------------+--
--+-----+-----------+----------------+-----------+------+--------+------------
-+------+--------+------+--------------+------+---+
|            acquired|         45|       16|         1|     22|       10|
2|       11|        3|        56|      5|      4|         31|      11|     3|
1|            0|      4|        7|    52|    6|             24|   7|    2|
5|              10|          1|     7|      15|           24|     8|     101|
1|             2|     7| 93|
|              closed|         17|        3|         1|     12|       13|
1|       14|        1|        17|      3|      2|         21|      16|     0|
0|            2|      3|        4|    27|    0|             10|   1|    9|
2|              15|          2|     5|       4|           11|     6|      52|
0|             0|     1| 51|
+--------------------+-----------+---------+----------+-------+---------+-----
-----+---------+---------+----------+-------+-------+-----------+--------+----
--+-----------+-------------+-------+---------+------+-----+---------------+--
--+-----+-----------+----------------+-----------+------+--------+------------
-+------+--------+------+--------------+------+---+
```

In [37]:
```
Cstartup_data_sel.stat.crosstab('status','state_code').show()
```

```
+-----------------+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+-
--+
|status_state_code| AR| AZ| CA| CO| CT| DC| FL| GA| ID| IL| IN| KY| MA| MD| ME
| MI| MN| MO| NC| NH| NJ| NM| NV| NY| OH| OR| PA| RI| TN| TX| UT| VA| WA| WI|
WV|
+-----------------+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+-
--+
|         acquired|  0|  1|332| 14|  0|  2|  2|  6|  0|  9|  1|  1| 64|  5|  1
|  0|  3|  1|  2|  1|  3|  0|  1| 77|  0|  6|  6|  2|  2| 23|  1|  7| 24|  0|
0|
|           closed|  1|  1|156|  5|  4|  2|  4|  5|  1|  9|  1|  1| 19|  2|  1
|  3|  2|  1|  5|  1|  4|  1|  1| 29|  6|  1| 11|  1|  1| 19|  2|  6| 18|  1|
1|
+-----------------+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+-
--+
```

In [38]:
```
Cstartup_data_sel.groupBy("state_code").count().show()
```

```
+----------+-----+
|state_code|count|
+----------+-----+
|        AZ|    2|
|        MN|    5|
|        NJ|    7|
|        DC|    4|
|        OR|    7|
|        VA|   13|
|        RI|    3|
|        KY|    2|
|        NH|    2|
|        MI|    3|
|        NV|    2|
|        WI|    1|
|        ID|    1|
|        CA|  488|
|        CT|    4|
|        NC|    7|
```

```
|          MD|    7|
|          MO|    2|
|          IL|   18|
|          ME|    2|
+-----------+-----+
only showing top 20 rows
```

In [39]:
```
import pandas as pd
import pyspark.sql as sparksql
```

In [40]:
```
import six
for i in Cstartup_data_sel.columns:
    if not( isinstance(Cstartup_data_sel.select(i).take(1)[0][0], six.string_
        print("age_first_milestone_year_imputed", i, Cstartup_data_sel.stat.c
```

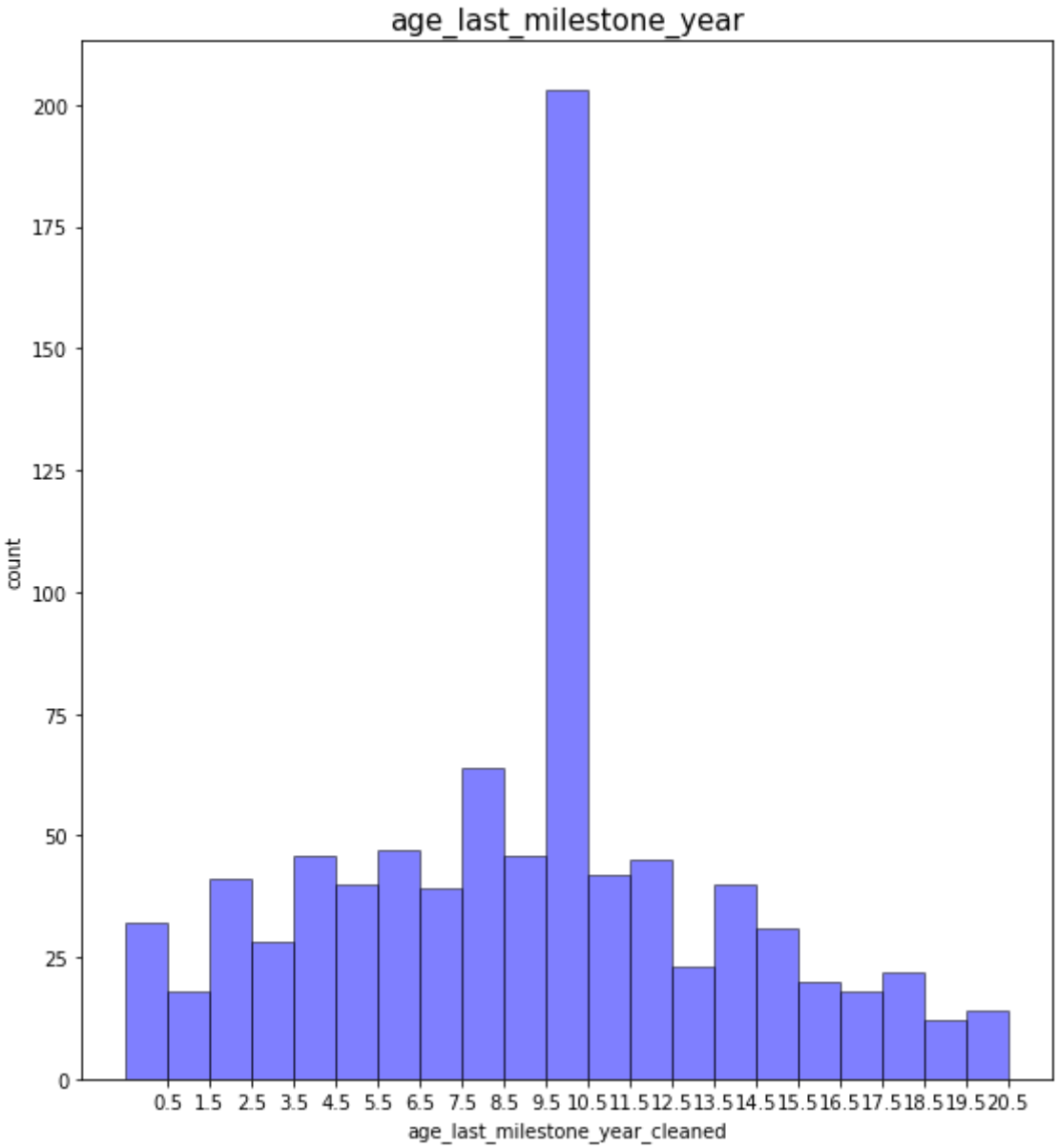
```
age_first_milestone_year_imputed age_first_funding_year 0.4962045017539433
age_first_milestone_year_imputed age_last_funding_year 0.6093921304405192
age_first_milestone_year_imputed age_first_milestone_year_imputed 1.0
age_first_milestone_year_imputed age_last_milestone_year_imputed 0.77748417002
22529
age_first_milestone_year_imputed funding_total_usd 0.06377772243076384
age_first_milestone_year_imputed relationships 0.22837566992186503
age_first_milestone_year_imputed milestones -0.04280525952031101
age_first_milestone_year_imputed is_top500 0.1361279313014484
age_first_milestone_year_imputed avg_participants 0.05147086263331164
age_first_milestone_year_imputed funding_rounds 0.17749013505354233
age_first_milestone_year_imputed has_angel -0.2616945975630935
age_first_milestone_year_imputed has_VC 0.09924098477619024
age_first_milestone_year_imputed latitude -0.06317072864624129
age_first_milestone_year_imputed longitude -0.04708812975749799
```

In [41]:
```
Cstartup_data_sel.groupBy("age_first_funding_year").sum('age_first_funding_ye
```

```
+---------------------+--------------------------+
|age_first_funding_year|sum(age_first_funding_year)|
+---------------------+--------------------------+
|                6.0055|                    6.0055|
|                0.4658|                    0.4658|
|                1.4082|                    2.8164|
|                9.7315|                    9.7315|
|               -0.4959|                   -0.9918|
|                 1.063|                     2.126|
|                1.7644|                    1.7644|
|                2.1233|                    2.1233|
|                5.0082|                    5.0082|
|                3.7534|                    3.7534|
|                3.3315|                    3.3315|
|                5.2027|                   10.4054|
|                6.6027|                    6.6027|
|                2.9507|                    2.9507|
|                1.6055|                    1.6055|
|               -0.3342|                   -0.3342|
|                   0.0|                       0.0|
|                7.5315|                    7.5315|
|                2.0658|                    2.0658|
|                5.1863|                    5.1863|
+---------------------+--------------------------+
only showing top 20 rows
```

In [42]:
```
var = 'age_last_milestone_year_imputed'
plot_data = Cstartup_data.select(var).toPandas()
x= plot_data[var]
bins = np.arange(-0.30, 10.50, 0.5)

hist, bin_edges = np.histogram(x,bins) # make the histogram
```

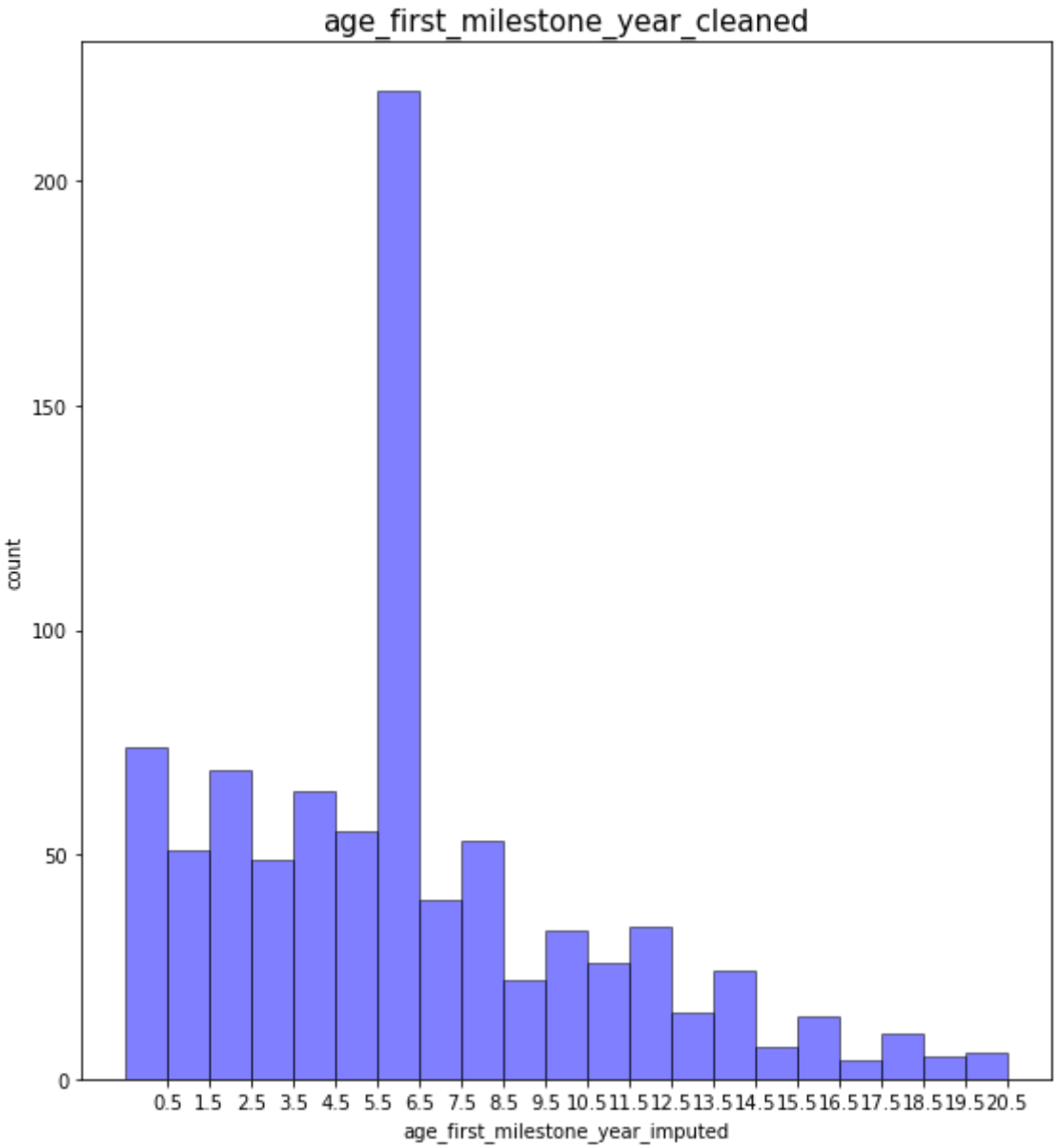
```
fig = plt.figure(figsize=(20, 10))
ax = fig.add_subplot(1, 2, 2)
# Plot the histogram heights against integers on the x axis
ax.bar(range(len(hist)),hist,width=1,alpha=0.5,ec ='black', color='blue')
# # Set the ticks to the middle of the bars
ax.set_xticks([0.5+i for i,j in enumerate(hist)])


plt.xlabel('age_last_milestone_year_cleaned')
plt.ylabel('count')
plt.title('age_last_milestone_year'.format(var), size=15)
plt.show()
```



In [43]:

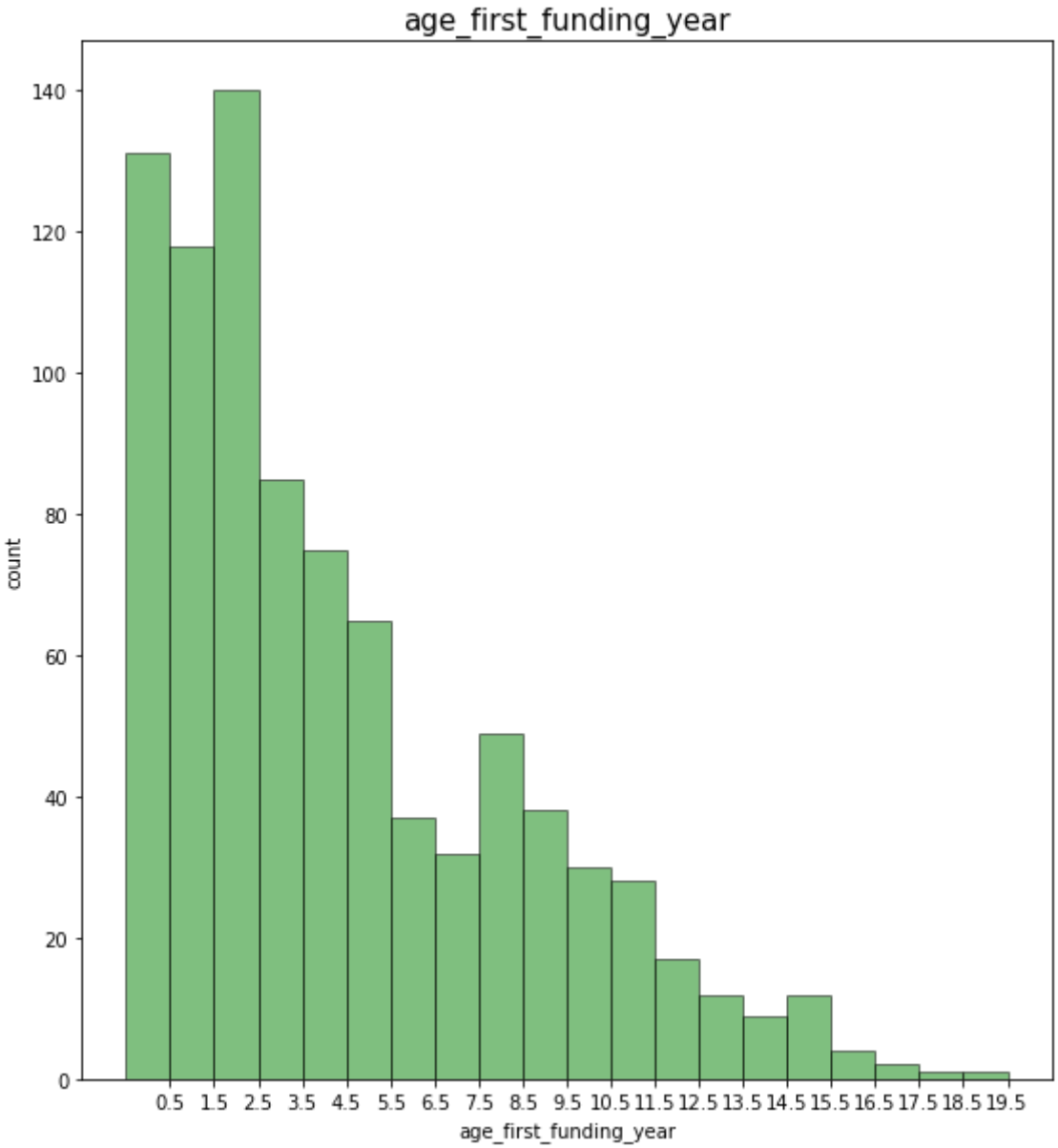
```
var = 'age_first_milestone_year_imputed'
plot_data = Cstartup_data.select(var).toPandas()
x= plot_data[var]
bins = np.arange(-0.30, 10.50, 0.5)

hist, bin_edges = np.histogram(x,bins) # make the histogram

fig = plt.figure(figsize=(20, 10))
ax = fig.add_subplot(1, 2, 2)
# Plot the histogram heights against integers on the x axis
ax.bar(range(len(hist)),hist,width=1,alpha=0.5,ec ='black', color='blue')
# # Set the ticks to the middle of the bars
ax.set_xticks([0.5+i for i,j in enumerate(hist)])
```

```
plt.xlabel('age_first_milestone_year_imputed')
plt.ylabel('count')
plt.title('age_first_milestone_year_cleaned'.format(var), size=15)
plt.show()
```



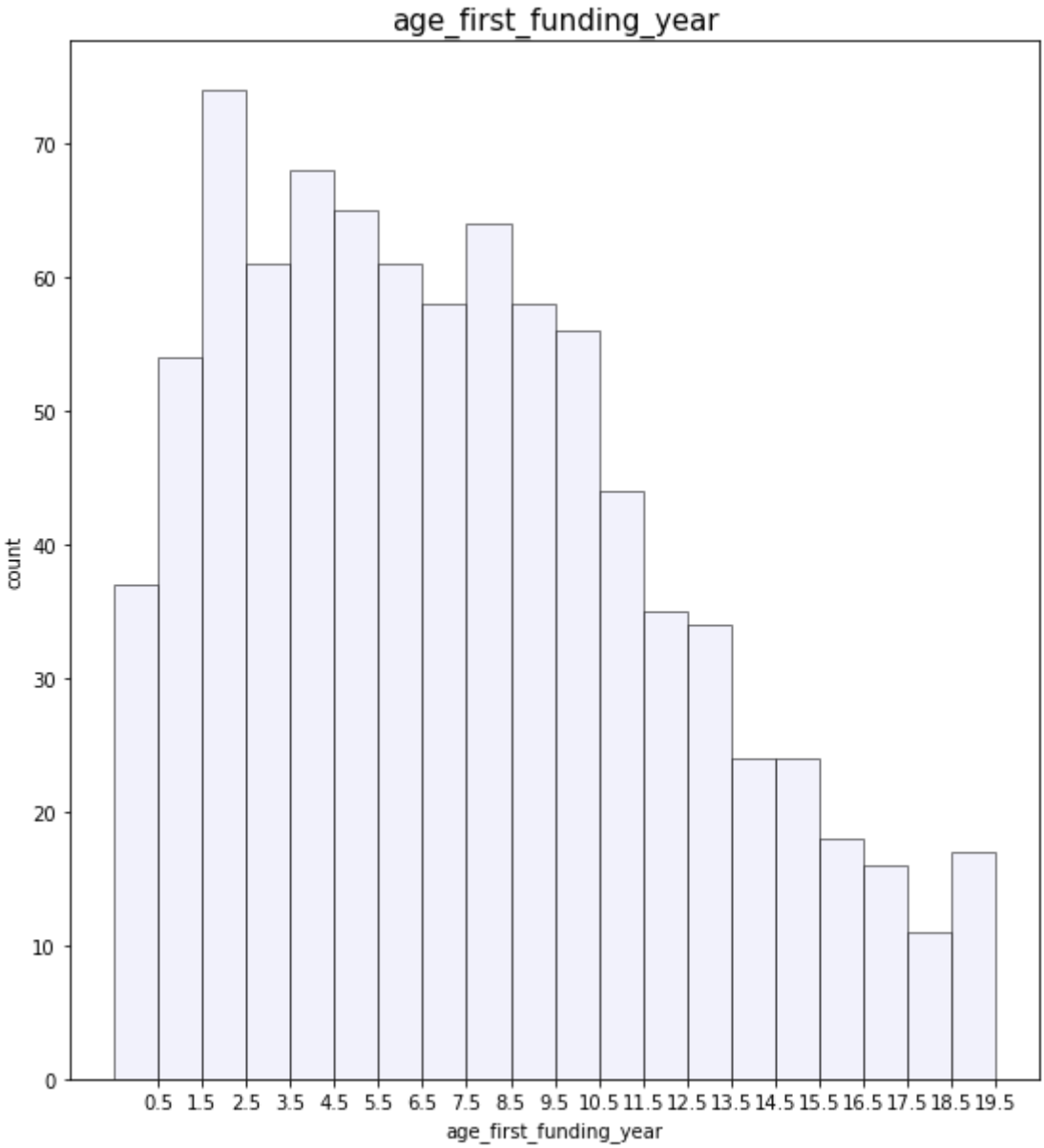
In [44]:

```
var = 'age_first_funding_year'
plot_data = Cstartup_data.select(var).toPandas()
x= plot_data[var]
bins = np.arange(-0.30, 10, 0.5)

hist, bin_edges = np.histogram(x,bins) # make the histogram

fig = plt.figure(figsize=(20, 10))
ax = fig.add_subplot(1, 2, 2)
# Plot the histogram heights against integers on the x axis
ax.bar(range(len(hist)),hist,width=1,alpha=0.5,ec ='black', color='green')
# # Set the ticks to the middle of the bars
ax.set_xticks([0.5+i for i,j in enumerate(hist)])


plt.xlabel('age_first_funding_year')
plt.ylabel('count')
plt.title('age_first_funding_year'.format(var), size=15)
plt.show()
```



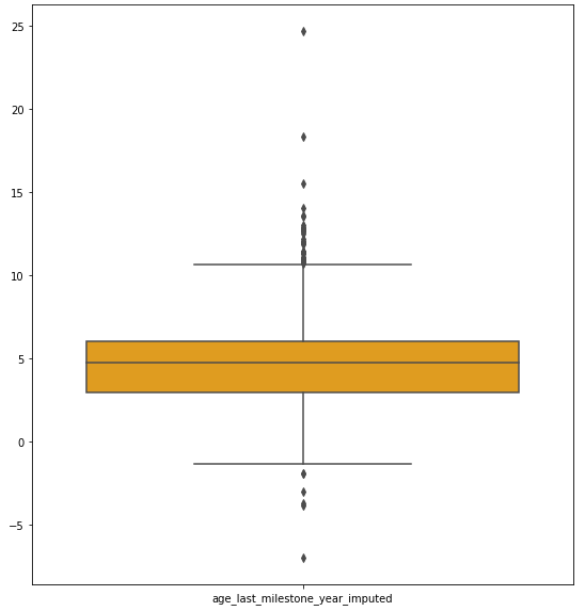
In [45]:

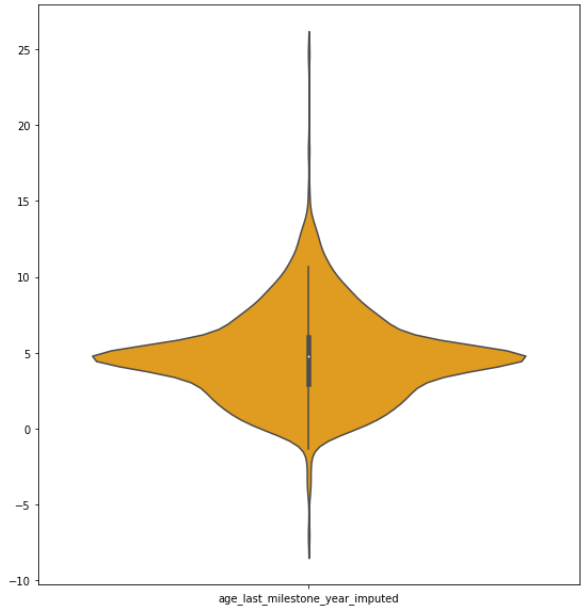
```
var = 'age_last_funding_year'
plot_data = Cstartup_data.select(var).toPandas()
x= plot_data[var]
bins = np.arange(-0.30, 10, 0.5)

hist, bin_edges = np.histogram(x,bins) # make the histogram

fig = plt.figure(figsize=(20, 10))
ax = fig.add_subplot(1, 2, 2)
# Plot the histogram heights against integers on the x axis
ax.bar(range(len(hist)),hist,width=1,alpha=0.5,ec ='black', color='lavender')
# # Set the ticks to the middle of the bars
ax.set_xticks([0.5+i for i,j in enumerate(hist)])


plt.xlabel('age_first_funding_year')
plt.ylabel('count')
plt.title('age_first_funding_year'.format(var), size=15)
plt.show()
```



In [46]:

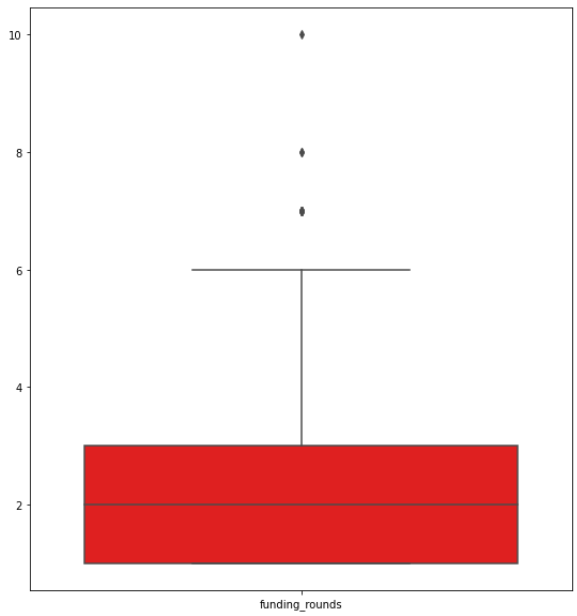
```
import seaborn as sns

var = 'age_last_milestone_year_imputed'
x = Cstartup_data.select(var).toPandas()

fig = plt.figure(figsize=(20, 10))
ax = fig.add_subplot(1, 2, 1)
ax = sns.boxplot(data=x, color='orange')

ax = fig.add_subplot(1, 2, 2)
ax = sns.violinplot(data=x, color='orange')
```

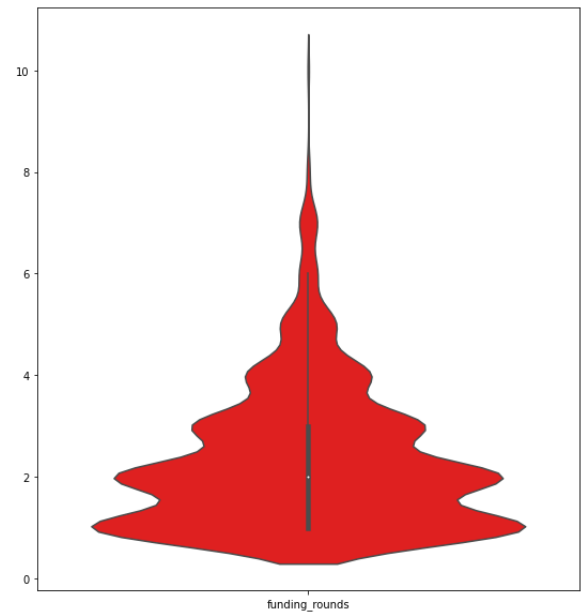




In [47]:

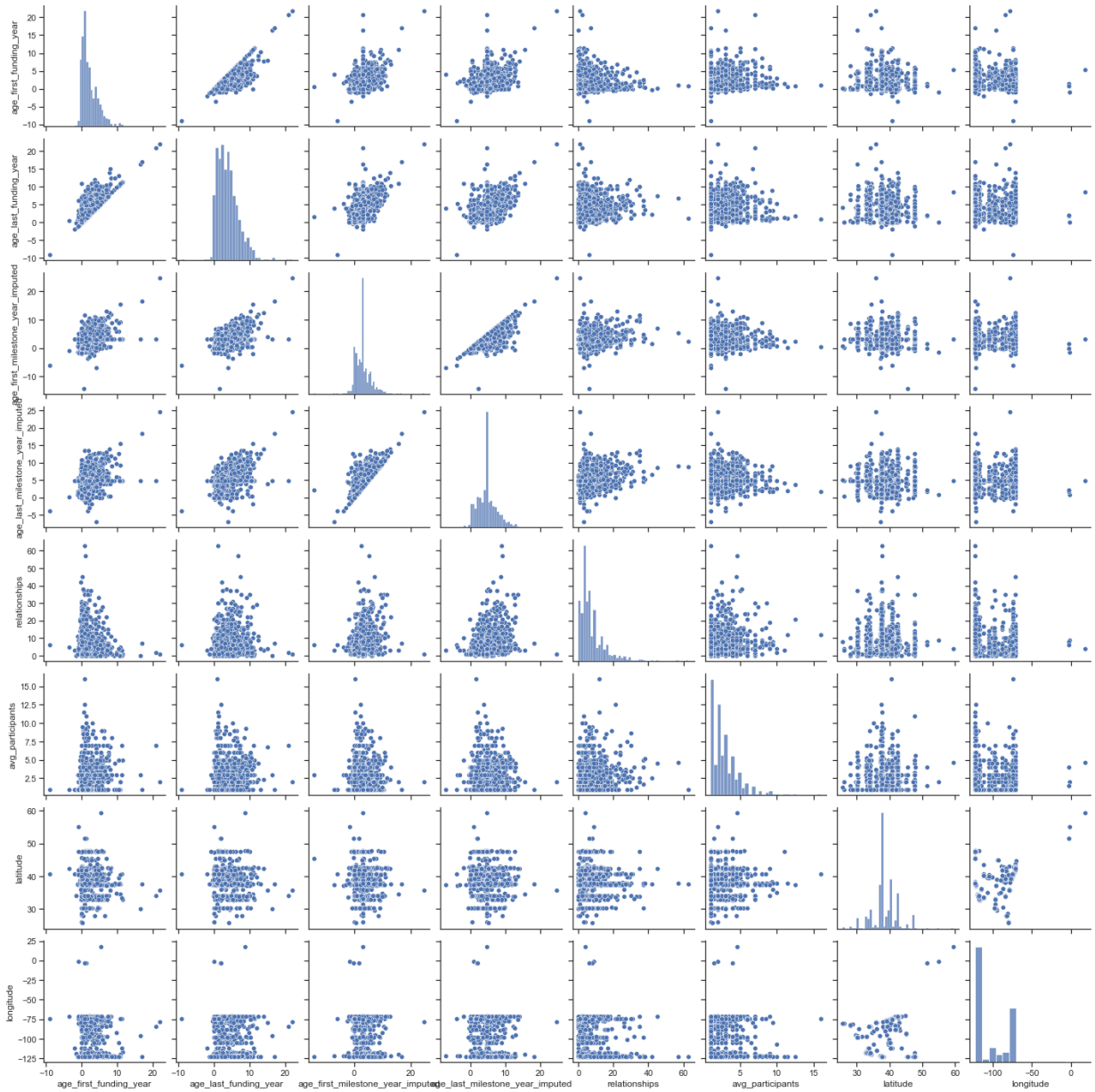
```
import seaborn as sns

var = 'funding_rounds'
x = Cstartup_data.select(var).toPandas()

fig = plt.figure(figsize=(20, 10))
ax = fig.add_subplot(1, 2, 1)
ax = sns.boxplot(data=x, color= 'red')

ax = fig.add_subplot(1, 2, 2)
ax = sns.violinplot(data=x, color='red')
```





In [48]:

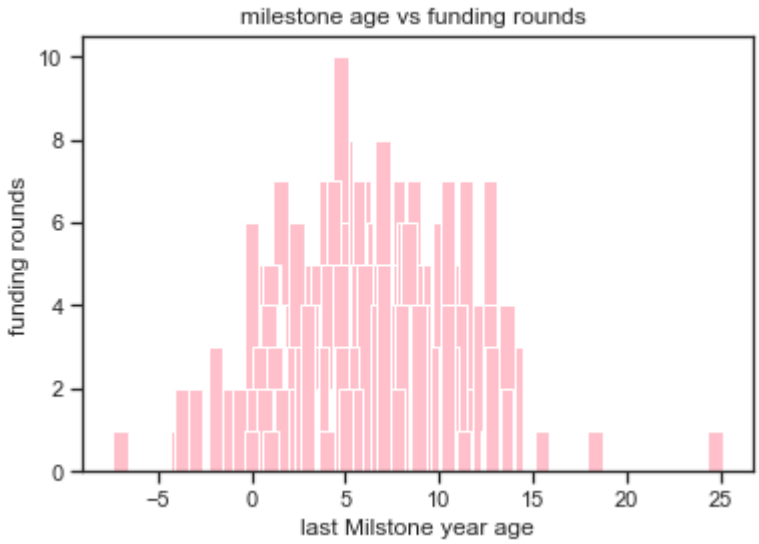
```
Cstartup_data_scatter_plot = Cstartup_data.select('age_first_funding_year', '
                'age_first_milestone_year_imputed', 'age_last_milestone_year_
                'avg_participants','category_code', 'state_code', 'latitude',
sns.set(style="ticks")
sns.pairplot(Cstartup_data_scatter_plot.toPandas())
plt.show()
```



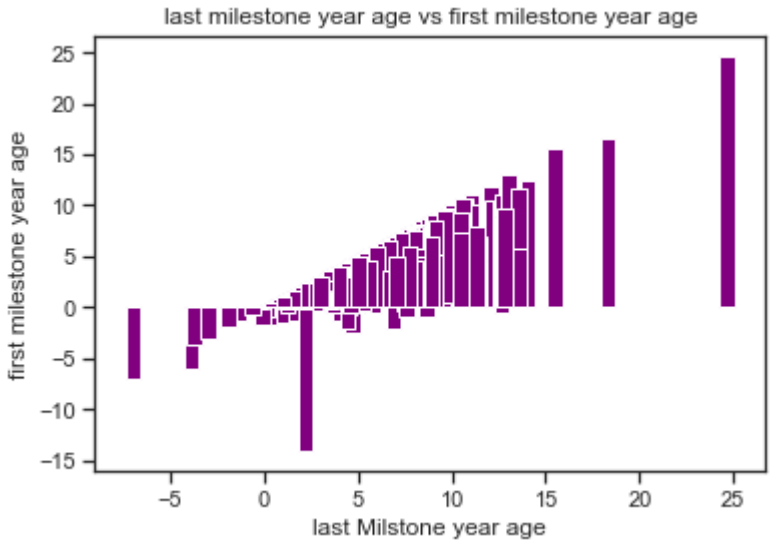
In [49]:
```
Cstartup_data_sel.stat.corr('age_first_milestone_year_imputed','age_last_mile
```

Out[49]:
```
0.7774841700222529
```

In [50]:
```
x = Cstartup_data_sel.toPandas()["age_last_milestone_year_imputed"]
y = Cstartup_data_sel.toPandas()["funding_rounds"]
plt.bar(x,y, color='pink')
plt.xlabel("last Milstone year age")
plt.ylabel("funding rounds")
plt.title("milestone age vs funding rounds ")
plt.show()
```



In [51]:
```
x = Cstartup_data_sel.toPandas()["age_last_milestone_year_imputed"]
y = Cstartup_data_sel.toPandas()["age_first_milestone_year_imputed"]
plt.bar(x,y, color='purple')
plt.xlabel("last Milstone year age")
plt.ylabel("first milestone year age")
plt.title("last milestone year age vs first milestone year age")
plt.show()
```



Linear regression

In [52]:
```
Cstartup_data_sel = Cstartup_data.select('age_first_funding_year', 'age_last_
                'age_first_milestone_year_imputed', 'age_last_milestone_year_
                'milestones','is_top500','avg_participants','funding_rounds',
                'category_code','has_VC', 'state_code', 'latitude','longitude
```

In [53]:
```
from pyspark.ml.feature import VectorAssembler
```

In [54]:
```
import six
for i in Cstartup_data_sel.columns:
    if not( isinstance(Cstartup_data_sel.select(i).take(1)[0][0], six.string_
        print("age_first_milestone_year_imputed", i, Cstartup_data_sel.stat.c
```

```
age_first_milestone_year_imputed age_first_funding_year 0.4962045017539432
age_first_milestone_year_imputed age_last_funding_year 0.6093921304405191
age_first_milestone_year_imputed age_first_milestone_year_imputed 1.0
age_first_milestone_year_imputed age_last_milestone_year_imputed 0.77748417002
```

```
22527
age_first_milestone_year_imputed funding_total_usd 0.06377772243076382
age_first_milestone_year_imputed relationships 0.22837566992186503
age_first_milestone_year_imputed milestones -0.04280525952031101
age_first_milestone_year_imputed is_top500 0.1361279313014484
age_first_milestone_year_imputed avg_participants 0.051470862633311625
age_first_milestone_year_imputed funding_rounds 0.1774901350535423
age_first_milestone_year_imputed has_angel -0.2616945975630934
age_first_milestone_year_imputed has_VC 0.09924098477619021
age_first_milestone_year_imputed latitude -0.06317072864624133
age_first_milestone_year_imputed longitude -0.04708812975749797
```

In [55]:
```
vectorAssembler = VectorAssembler(inputCols = ['age_first_funding_year', 'age
                'age_first_milestone_year_imputed', 'funding_total_usd','rela
                'milestones','is_top500','avg_participants','funding_rounds',
                'latitude','longitude'], outputCol = 'fs')
Cstartup_data_sel= vectorAssembler.transform(Cstartup_data_sel)
Cstartup_data_sel = Cstartup_data_sel.select(['fs','age_last_milestone_year_i
Cstartup_data_sel.show(3)
```

```
+--------------------+-------------------------------+
|                  fs|age_last_milestone_year_imputed|
+--------------------+-------------------------------+
|[1.189,1.189,2.18...|                         7.7753|
|[2.6986,4.0027,1....|                            1.0|
|[0.2986,0.2986,0....|                         0.2986|
+--------------------+-------------------------------+
only showing top 3 rows
```

In [56]:
```
splits = Cstartup_data_sel.randomSplit([0.7, 0.3])
train_df = splits[0]
test_df = splits[1]
```

In [57]:
```
from pyspark.ml.regression import LinearRegression
```

In [58]:
```
lr = LinearRegression(featuresCol = 'fs', labelCol='age_last_milestone_year_i
lr_model = lr.fit(train_df)
print("Coefficients: " + str(lr_model.coefficients))
print("Intercept: " + str(lr_model.intercept))
```

```
Coefficients: [0.0,0.14889497807648583,0.637903931758232,0.0,0.0,0.49368171016
44687,0.0,0.0,0.0,0.0,0.0,0.0,0.0]
Intercept: 1.2989255515482627
```

In [59]:
```
trainingSummary = lr_model.summary
print("RMSE: %f" % trainingSummary.rootMeanSquaredError)
print("r2: %f" % trainingSummary.r2)
```

```
RMSE: 1.620708
r2: 0.681388
```