Use Scala (Spark) API to implement the following task.

Given the spark Dataframe as:

|  |
| --- |
| [  ('ABC17969(AB)', '1', 'ABC17969', 2022),  ('ABC17969(AB)', '2', 'CDC52533', 2022),  ('ABC17969(AB)', '3', 'DEC59161', 2023),  ('ABC17969(AB)', '4', 'F43874', 2022),  ('ABC17969(AB)', '5', 'MY06154', 2021),  ('ABC17969(AB)', '6', 'MY4387', 2022),  ('AE686(AE)', '7', 'AE686', 2023),  ('AE686(AE)', '8', 'BH2740', 2021),  ('AE686(AE)', '9', 'EG999', 2021),  ('AE686(AE)', '10', 'AE0908', 2021),  ('AE686(AE)', '11', 'QA402', 2022),  ('AE686(AE)', '12', 'OM691', 2022)  ] |

Schema: peer\_id, id\_1, id\_2, year.

Process:

1. For each peer\_id, get the year when peer\_id contains id\_2, for example for ‘ABC17969(AB)’ year is 2022.
2. Given a size number, for example 3. For each peer\_id count the number of each year (which is smaller or equal than the year in step1).

For example, for ‘ABC17969(AB)’, the count should be:

|  |
| --- |
| 2022, 4  2021, 1 |

(2023 is bigger than 2022, hence do not include)

1. Order the value in step 2 by year and check if the count number of the first year is bigger or equal than the given size number. If yes, just return the year.

If not, plus the count number from the biggest year to next year until the count number is bigger or equal than the given number. For example, for ‘AE686(AE)’, the year is 2023, and count are:

|  |
| --- |
| 2023, 1  2022, 2,  2021, 3 |

As 1(*2023 count*) + 2(*2022 count*) >= 3 (*given size number*), the output would be 2023, 2022.

The final expected output for the given example would be:

|  |
| --- |
| [  ('ABC17969(AB)', 2022),  (' AE686(AE)', 2022),  (' AE686(AE)', 2023),  ] |

Requirement:

1. Share your implement code;
2. Write unit test;
3. Add screenshot for the run result.

Tips:

If the given number is 5, dataframe is:

|  |
| --- |
| [  ('AE686(AE)', '7', 'AE686', 2022),  ('AE686(AE)', '8', 'BH2740', 2021),  ('AE686(AE)', '9', 'EG999', 2021),  ('AE686(AE)', '10', 'AE0908', 2023),  ('AE686(AE)', '11', 'QA402', 2022),  ('AE686(AE)', '12', 'OA691', 2022),  ('AE686(AE)', '12', 'OB691', 2022),  ('AE686(AE)', '12', 'OC691', 2019),  ('AE686(AE)', '12', 'OD691', 2017)  ] |

Then output would be:

|  |
| --- |
| [  ('AE686(AE)',2022),  ('AE686(AE)',2021)  ] |

If give number is 7, the output would be:

|  |
| --- |
| [  ('AE686(AE)',2022),  ('AE686(AE)',2021),  ('AE686(AE)',2019)  ] |