数据准备：

```
# 第一题
spark = SparkSession.builder\
    .appName('four')\
    .master('local[*]')\
    .getOrCreate()

sc = spark.sparkContext
data_rdd = sc.parallelize([('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)])

# peer_id, id_1, id_2, year.
schema = StructType()\
    .add('peer_id', StringType(), False)\
    .add('id_1', StringType(), False)\
    .add('id_2', StringType(), False)\
    .add('year', IntegerType(), False)

data_df = data_rdd.toDF(schema)
```

第一问及结果：

```
# 1.For each peer_id, get the year when peer_id contains id_2, for example for 'ABC17969(AB)' year is 2022.
result_one = data_df.where(col('peer_id').contains(col('id_2'))).select('year')
result_one.show()
```

five

```
+----+
|year|
+----+
|2022|
|2023|
+----+
```

第二问及结果：

```
# 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:
data_df.where((col('year') <= 2022) & (col('peer_id') == 'ABC17969(AB)')).groupby('peer_id', 'year').agg(count('peer_id').alias('num')).select('peer_id','year', 'num').show()
data_df.where((col('year') <= 2023) & (col('peer_id') == 'AE686(AE)')).groupby('peer_id', 'year').agg(count('peer_id').alias('num')).select('peer_id','year', 'num').show()
```

five ×

```
24/03/19 16:26:45 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
+------------+----+---+
|     peer_id|year|num|
+------------+----+---+
|ABC17969(AB)|2022|  4|
|ABC17969(AB)|2021|  1|
+------------+----+---+

+---------+----+---+
|  peer_id|year|num|
+---------+----+---+
|AE686(AE)|2023|  1|
|AE686(AE)|2021|  3|
|AE686(AE)|2022|  2|
+---------+----+---+
```

第三问及结果：

```
# 比如给的是三
window = Window.orderBy(desc('year')).rowsBetween(Window.unboundedPreceding, Window.currentRow)
# long 为 ABC17969(AB)
long_res_df = long_res_df.select('peer_id','year','num',sum('num').over(window).alias('s'),row_number().over(window).alias('rk'))

# #情况一：当数字大于3， 直接返回年限
long_res_df.where(col('num') >= 3).select('peer_id','year').show()
# #情况二：当部分和大于三， 返回多个年限

short_res_df = short_res_df.select('peer_id','year','num',sum('num').over(window).alias('s'),row_number().over(window).alias('rk'))
res = short_res_df.where((col('num') < 3) & (col('s') >= 3)).select('rk')
short_res_df.where(col('rk') <= res['rk']).select('peer_id','year').show()
```

Run five ×

```
24/03/19 17:33:34 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applica
+------------+----+
|     peer_id|year|
+------------+----+
|ABC17969(AB)|2022|
+------------+----+

+---------+----+
|  peer_id|year|
+---------+----+
|AE686(AE)|2023|
|AE686(AE)|2022|
|AE686(AE)|2021|
+---------+----+
```