Quiz 2

1. Given three DataFrame df1,df2 and df, please choose one operation below which can generate df by using df1 and df2.

```
>>> df1
                    >>> df2
     lkey value
                        rkey value
                         foo
0
     foo
                 1
                    0
                                    5
                                    6
                    1
1
                 2
                        bar
     bar
                                    7
                    2
                 3
                        baz
2
     baz
                                    8
                    3
                         foo
3
     foo
         value left rkey
  lkey
                              value right
0
   foo
                    1
                        foo
1
                                          8
   foo
                    1
                        foo
2
   foo
                    5
                        foo
                                          5
3
   foo
                    5
                        foo
                                          8
                    2
4
                                          6
   bar
                        bar
5
                    3
                                          7
   baz
                        baz
```

df1.merge(df2, left_on='lkey', right_on='rkey')

df2.merge(df1, left_on='lkey', right_on='rkey')

df1.merge(df2, left_on='lkey', right_on='rkey', suffixes=('_left', '_right'))

df2.merge(df1, left_on='lkey', right_on='rkey', suffixes=('_left', '_right'))

df1.merge(df2, left_on='lkey', right_on='rkey', suffixes=(False, False))

df2.merge(df1, left_on='lkey', right_on='rkey', suffixes=(True, False))

2.

Given a DataFrame, apply the z-score normalization

In [3]: df
Out[3]:

a b c d e

a b c d e
 1 4 7 8 7
 2 5 8 8 3
 3 6 9 8 2

.

•		а	b	С	d	е
	0	-1.224745	-1.224745	-1.224745	NaN	1.38873
	1	0.000000	0.000000	0.000000	NaN	-0.46291
	2	1.224745	1.224745	1.224745	NaN	-0.92582
0		а	b		c d	е
	0	-1.224745	-1.224745	-1.22474	5 8	1.38873
	1	0.000000	0.000000	0.00000	0 8	-0.46291
	2	1.224745	1.224745	1.22474	5 8	-0.92582
		а	b		c d	е
	0	-1.224745	-1.224745			e 1.38873
	0			-1.22474	5 8	
		-1.224745 2.000000	-1.224745 5.000000	-1.22474 8.00000	5 8 0 8	1.38873
	1	-1.224745 2.000000	-1.224745 5.000000	-1.22474 8.00000	5 8 0 8	1.38873 -0.46291
	1	-1.224745 2.000000	-1.224745 5.000000	-1.22474 8.00000	5 8 0 8	1.38873 -0.46291 -0.92582
	1	-1.224745 2.000000 1.224745	-1.224745 5.000000 1.224745 b	-1.22474 8.00000 1.22474	5 8 0 8 5 8	1.38873 -0.46291 -0.92582 e
	1 2	-1.224745 2.000000 1.224745 a	-1.224745 5.000000 1.224745 b	-1.22474 8.00000 1.22474 c	5 8 0 8 5 8 d	1.38873 -0.46291 -0.92582 e 1.38873
	1 2	-1.224745 2.000000 1.224745 a -1.224745	-1.224745 5.000000 1.224745 b	-1.224745 8.000000 1.224745 c	5 8 0 8 5 8 d NaN NaN	1.38873 -0.46291 -0.92582 e 1.38873 -0.46291

3.

For text content standardization, the correct order should be:

Tokenization, Stemming, Lemmatization

0	Tokenization, Lemmatization, Stemming
0	Here is bug from WebCMS, don't choose me.
0	None of Above
•	Both a and b

4. Which kind of data can be considered as dirty data?

	Incomplete data
	Duplicate data
	Inconsistent data
•	All of above