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
0    foo      1   0   foo      5
1    bar      2   1   bar      6
2    baz      3   2   baz      7
3    foo      5   3   foo      8
   lkey  value_left rkey  value_right
0  foo            1  foo            5
1  foo            1  foo            8
2  foo            5  foo            5
3  foo            5  foo            8
4  bar            2  bar            6
5  baz            3  baz            7
```

○   `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
0	1	4	7	8	7
1	2	5	8	8	3
2	3	6	9	8	2

.

○ (selected)

	a	b	c	d	e
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

○

	a	b	c	d	e
0	-1.224745	-1.224745	-1.224745	8	1.38873
1	0.000000	0.000000	0.000000	8	-0.46291
2	1.224745	1.224745	1.224745	8	-0.92582

○

	a	b	c	d	e
0	-1.224745	-1.224745	-1.224745	8	1.38873
1	2.000000	5.000000	8.000000	8	-0.46291
2	1.224745	1.224745	1.224745	8	-0.92582

○

	a	b	c	d	e
0	-1.224745	-1.224745	-1.224745	NaN	1.38873
1	2.000000	5.000000	8.000000	NaN	-0.46291
2	1.224745	1.224745	1.224745	NaN	-0.92582

3.

For text content standardization, the correct order should be:

○   Tokenization, Stemming, Lemmatization

○	Tokenization, Lemmatization, Stemming
○	Here is bug from WebCMS, don't choose me.
○	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