3

2

I am troubling with counting the number of counties using famous [cenus.csv](http://www.census.gov/popest/data/counties/totals/2015/CO-EST2015-alldata.html) data.

Task: Count number of counties in each state.

Facing comparing (I think) / Please read below?

I've tried this:

df = pd.read\_csv('census.csv')

dfd = df[:]['STNAME'].unique() //Gives out names of state

serr = pd.Series(dfd) // converting to series (from array)

After this, i've tried using two approaches:

1:

df[df['STNAME'] == serr] \*\*//ERROR: series length must match\*\*

2:

i = 0

for name in serr: //This generate error 'Alabama'

df['STNAME'] == name

for i in serr:

serr[i] == serr[name]

print(serr[name].count)

i+=1

Please guide me; it has been three days with this stuff.

[python](https://stackoverflow.com/questions/tagged/python) [pandas](https://stackoverflow.com/questions/tagged/pandas) [dataset](https://stackoverflow.com/questions/tagged/dataset) [data-science](https://stackoverflow.com/questions/tagged/data-science)

[share](https://stackoverflow.com/q/40523185)[improve this question](https://stackoverflow.com/posts/40523185/edit)

[edited Nov 10 '16 at 8:52](https://stackoverflow.com/posts/40523185/revisions)

[[https://i.stack.imgur.com/JTknr.png?s=32&g=1](https://stackoverflow.com/users/5014455/juanpa-arrivillaga)](https://stackoverflow.com/users/5014455/juanpa-arrivillaga)

[juanpa.arrivillaga](https://stackoverflow.com/users/5014455/juanpa-arrivillaga)

**39.2k**33977

asked Nov 10 '16 at 8:40

[[https://www.gravatar.com/avatar/618e4c0297ae7d1f8b2700dc57af2b6a?s=32&d=identicon&r=PG&f=1](https://stackoverflow.com/users/7140043/bakhtawar)](https://stackoverflow.com/users/7140043/bakhtawar)

[Bakhtawar](https://stackoverflow.com/users/7140043/bakhtawar)

**33**18

add a comment

3 Answers

[active](https://stackoverflow.com/questions/40523185/count-number-of-counties-per-state-using-python-census?answertab=active#tab-top)[oldest](https://stackoverflow.com/questions/40523185/count-number-of-counties-per-state-using-python-census?answertab=oldest#tab-top)[votes](https://stackoverflow.com/questions/40523185/count-number-of-counties-per-state-using-python-census?answertab=votes#tab-top)

6

Use groupby and aggregate COUNTY using nunique:

In [1]: import pandas as pd

In [2]: df = pd.read\_csv('census.csv')

In [3]: unique\_counties = df.groupby('STNAME')['COUNTY'].nunique()

Now the results

In [4]: unique\_counties

Out[4]:

STNAME

Alabama 68

Alaska 30

Arizona 16

Arkansas 76

California 59

Colorado 65

Connecticut 9

Delaware 4

District of Columbia 2

Florida 68

Georgia 160

Hawaii 6

Idaho 45

Illinois 103

Indiana 93

Iowa 100

Kansas 106

Kentucky 121

Louisiana 65

Maine 17

Maryland 25

Massachusetts 15

Michigan 84

Minnesota 88

Mississippi 83

Missouri 116

Montana 57

Nebraska 94

Nevada 18

New Hampshire 11

New Jersey 22

New Mexico 34

New York 63

North Carolina 101

North Dakota 54

Ohio 89

Oklahoma 78

Oregon 37

Pennsylvania 68

Rhode Island 6

South Carolina 47

South Dakota 67

Tennessee 96

Texas 255

Utah 30

Vermont 15

Virginia 134

Washington 40

West Virginia 56

Wisconsin 73

Wyoming 24

Name: COUNTY, dtype: int64

[share](https://stackoverflow.com/a/40523367)[improve this answer](https://stackoverflow.com/posts/40523367/edit)

[edited Nov 10 '16 at 8:59](https://stackoverflow.com/posts/40523367/revisions)

answered Nov 10 '16 at 8:50

[[https://i.stack.imgur.com/JTknr.png?s=32&g=1](https://stackoverflow.com/users/5014455/juanpa-arrivillaga)](https://stackoverflow.com/users/5014455/juanpa-arrivillaga)

[juanpa.arrivillaga](https://stackoverflow.com/users/5014455/juanpa-arrivillaga)

**39.2k**33977

add a comment

2

[juanpa.arrivillaga](https://stackoverflow.com/users/5014455) has a great solution. However, the code needs a minor modification.

The "counties" with 'SUMLEV' == 40 or 'COUNTY' == 0 should be filtered. Otherwise, all the number of counties are too big by one.

So, the correct answer should be:

unique\_counties = census\_df[census\_df['SUMLEV'] == 50].groupby('STNAME')['COUNTY'].nunique()

with the following result:

STNAME

Alabama 67

Alaska 29

Arizona 15

Arkansas 75

California 58

Colorado 64

Connecticut 8

Delaware 3

District of Columbia 1

Florida 67

Georgia 159

Hawaii 5

Idaho 44

Illinois 102

Indiana 92

Iowa 99

Kansas 105

Kentucky 120

Louisiana 64

Maine 16

Maryland 24

Massachusetts 14

Michigan 83

Minnesota 87

Mississippi 82

Missouri 115

Montana 56

Nebraska 93

Nevada 17

New Hampshire 10

New Jersey 21

New Mexico 33

New York 62

North Carolina 100

North Dakota 53

Ohio 88

Oklahoma 77

Oregon 36

Pennsylvania 67

Rhode Island 5

South Carolina 46

South Dakota 66

Tennessee 95

Texas 254

Utah 29

Vermont 14

Virginia 133

Washington 39

West Virginia 55

Wisconsin 72

Wyoming 23

Name: COUNTY, dtype: int64

[share](https://stackoverflow.com/a/47244785)[improve this answer](https://stackoverflow.com/posts/47244785/edit)

[edited Nov 12 '17 at 7:27](https://stackoverflow.com/posts/47244785/revisions)

[[https://i.stack.imgur.com/Bqpz6.png?s=32&g=1](https://stackoverflow.com/users/1961728/robincts)](https://stackoverflow.com/users/1961728/robincts)

[robinCTS](https://stackoverflow.com/users/1961728/robincts)

**4,816**82134

answered Nov 12 '17 at 2:11

[[https://www.gravatar.com/avatar/331a5e85405109a757eb1c080bc74279?s=32&d=identicon&r=PG&f=1](https://stackoverflow.com/users/8926389/dong-du)](https://stackoverflow.com/users/8926389/dong-du)

[Dong Du](https://stackoverflow.com/users/8926389/dong-du)

**21**2

add a comment

1

@Bakhtawar - This is a very simple way:

df.groupby(df['STNAME']).count().COUNTY

[how to sort pandas dataframe from one column](https://stackoverflow.com/questions/37787698/how-to-sort-pandas-dataframe-from-one-column)

[Ask Question](https://stackoverflow.com/questions/ask)



Come build the future with us

[View all 211 job openings!](https://stackoverflow.com/jobs/companies/amazon-us)

108

22

I have a data frame like this:

print(df)

0 1 2

0 354.7 April 4.0

1 55.4 August 8.0

2 176.5 December 12.0

3 95.5 February 2.0

4 85.6 January 1.0

5 152 July 7.0

6 238.7 June 6.0

7 104.8 March 3.0

8 283.5 May 5.0

9 278.8 November 11.0

10 249.6 October 10.0

11 212.7 September 9.0

As you can see, months are not in calendar order. So I created a second column to get the month number corresponding to each month (1-12). From there, how can I sort this data frame according to calendar months' order?

[python](https://stackoverflow.com/questions/tagged/python) [pandas](https://stackoverflow.com/questions/tagged/pandas) [sorting](https://stackoverflow.com/questions/tagged/sorting)

[share](https://stackoverflow.com/q/37787698)[improve this question](https://stackoverflow.com/posts/37787698/edit)

[edited Feb 1 at 22:39](https://stackoverflow.com/posts/37787698/revisions)

[[https://www.gravatar.com/avatar/8adc0e0b5afd1ff077dd16a48492c1e0?s=32&d=identicon&r=PG&f=1](https://stackoverflow.com/users/8468264/d-kennetz)](https://stackoverflow.com/users/8468264/d-kennetz)

[d\_kennetz](https://stackoverflow.com/users/8468264/d-kennetz)

**2,298**3927

asked Jun 13 '16 at 10:44

[[https://i.stack.imgur.com/eoaOn.jpg?s=32&g=1](https://stackoverflow.com/users/6428638/sachila-ranawaka)](https://stackoverflow.com/users/6428638/sachila-ranawaka)

[Sachila Ranawaka](https://stackoverflow.com/users/6428638/sachila-ranawaka)

**20.1k**42651

add a comment

3 Answers

[active](https://stackoverflow.com/questions/37787698/how-to-sort-pandas-dataframe-from-one-column?answertab=active#tab-top)[oldest](https://stackoverflow.com/questions/37787698/how-to-sort-pandas-dataframe-from-one-column?answertab=oldest#tab-top)[votes](https://stackoverflow.com/questions/37787698/how-to-sort-pandas-dataframe-from-one-column?answertab=votes#tab-top)

150

Use [sort\_values](http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.sort_values.html) to sort the df by a specific column's values:

In [18]:

df.sort\_values('2')

Out[18]:

0 1 2

4 85.6 January 1.0

3 95.5 February 2.0

7 104.8 March 3.0

0 354.7 April 4.0

8 283.5 May 5.0

6 238.7 June 6.0

5 152.0 July 7.0

1 55.4 August 8.0

11 212.7 September 9.0

10 249.6 October 10.0

9 278.8 November 11.0

2 176.5 December 12.0

If you want to sort by two columns, pass a list of column labels to sort\_values with the column labels ordered according to sort priority. If you use df.sort\_values(['2', '0']), the result would be sorted by column 2 then column 0. Granted, this does not really make sense for this example because each value in df['2'] is unique.

[share](https://stackoverflow.com/a/37787724)[improve this answer](https://stackoverflow.com/posts/37787724/edit)

[edited Aug 1 '18 at 15:08](https://stackoverflow.com/posts/37787724/revisions)

[[https://i.stack.imgur.com/kHN76.jpg?s=32&g=1](https://stackoverflow.com/users/3585557/steven-c-howell)](https://stackoverflow.com/users/3585557/steven-c-howell)

[Steven C. Howell](https://stackoverflow.com/users/3585557/steven-c-howell)

**4,371**43563

answered Jun 13 '16 at 10:45

[[https://www.gravatar.com/avatar/3f9be2c2958e208c8d9b629ac43c9c42?s=32&d=identicon&r=PG](https://stackoverflow.com/users/704848/edchum)](https://stackoverflow.com/users/704848/edchum)

[EdChum](https://stackoverflow.com/users/704848/edchum)

**183k**34398334

* thanks @EdChum. actually the column 2 dtype is float64. so when i try this it give's me an integer is required. how can i convert this column from float to int – [Sachila Ranawaka](https://stackoverflow.com/users/6428638/sachila-ranawaka" \o "20,105 reputation) [Jun 13 '16 at 10:55](https://stackoverflow.com/questions/37787698/how-to-sort-pandas-dataframe-from-one-column#comment63041839_37787724)
* I don't understand your problem, what is the error here? is it the column name? the dtype doesn't matter here try df.sort\_values(df.columns[2]) – [EdChum](https://stackoverflow.com/users/704848/edchum" \o "183,385 reputation) [Jun 13 '16 at 10:56](https://stackoverflow.com/questions/37787698/how-to-sort-pandas-dataframe-from-one-column#comment63041898_37787724)
* edit your question with the output from df.columns.tolist() this will tell you what the actual column names are – [EdChum](https://stackoverflow.com/users/704848/edchum" \o "183,385 reputation) [Jun 13 '16 at 11:02](https://stackoverflow.com/questions/37787698/how-to-sort-pandas-dataframe-from-one-column#comment63042100_37787724)
* 11

This will sort the dataframe in ascending order. If you want to do it in descending order use: df.sort\_values('2', ascending=False ) – [Anwarvic](https://stackoverflow.com/users/5612363/anwarvic" \o "1,419 reputation) [Jun 3 '18 at 12:43](https://stackoverflow.com/questions/37787698/how-to-sort-pandas-dataframe-from-one-column#comment88342995_37787724)

* 1

@JKC you need to pass a list: df.sort\_values(['col\_1','col\_2']) – [EdChum](https://stackoverflow.com/users/704848/edchum" \o "183,385 reputation) [Jul 30 '18 at 9:32](https://stackoverflow.com/questions/37787698/how-to-sort-pandas-dataframe-from-one-column#comment90147015_37787724)

[show **2** more comments](https://stackoverflow.com/questions/37787698/how-to-sort-pandas-dataframe-from-one-column)

**Question 7**

Which county has had the largest absolute change in population within the period 2010-2015? (Hint: population values are stored in columns POPESTIMATE2010 through POPESTIMATE2015, you need to consider all six columns.)

e.g. If County Population in the 5 year period is 100, 120, 80, 105, 100, 130, then its largest change in the period would be |130-80| = 50.

*This function should return a single string value.*

In [12]:



**def** answer\_seven():

**return** "YOUR ANSWER HERE"

**Question 8**

In this datafile, the United States is broken up into four regions using the "REGION" column.

Create a query that finds the counties that belong to regions 1 or 2, whose name starts with 'Washington', and whose POPESTIMATE2015 was greater than their POPESTIMATE 2014.

*This function should return a 5x2 DataFrame with the columns = ['STNAME', 'CTYNAME'] and the same index ID as the census\_df (sorted ascending by index).*

In [\*]:



**def** answer\_eight():

census\_df['SUMLEV'].unique

df = census\_df[census\_df['SUMLEV'] == 50]

df= df.set\_index(['STNAME', 'CTYNAME'])

df['Pop\_years'] = df['POPESTIMATE2015']**-** df['POPESTIMATE2014']

y = df['REGION'] **<** 3

z = df['Pop\_years'] **>** 0

df = y **&** z

df.loc('Washington County')

**return** df

​

answer\_eight()

5

4

Let us say we have the following table

and I want to find max and min value for every row for a set of specific columns (let's say CENSUS2010POP, ESTIMATESBASE1010, POPESTIMATE2010). How to do it with Pandas?

[python](https://stackoverflow.com/questions/tagged/python) [pandas](https://stackoverflow.com/questions/tagged/pandas)

[share](https://stackoverflow.com/q/41439419)[improve this question](https://stackoverflow.com/posts/41439419/edit)

asked Jan 3 '17 at 8:37

[YohanRoth](https://stackoverflow.com/users/3849781/yohanroth)

**980**1919

* 1

for finding max value from column.. refer [stackoverflow.com/questions/15741759/…](http://stackoverflow.com/questions/15741759/find-maximum-value-of-a-column-and-return-the-corresponding-row-values-using-pan) – [Harsha Biyani](https://stackoverflow.com/users/3457761/harsha-biyani) [Jan 3 '17 at 8:40](https://stackoverflow.com/questions/41439419/how-to-select-max-and-min-value-in-a-row-for-selected-columns#comment70084881_41439419)

* @HarshaBiyani I know hot to find max value in column... but I need it for a row and consider only few columns in a row – [YohanRoth](https://stackoverflow.com/users/3849781/yohanroth" \o "980 reputation) [Jan 3 '17 at 8:40](https://stackoverflow.com/questions/41439419/how-to-select-max-and-min-value-in-a-row-for-selected-columns#comment70084902_41439419)
* 1

Possible duplicate of [Python Pandas max value of selected columns](http://stackoverflow.com/questions/20033111/python-pandas-max-value-of-selected-columns) – [RomanPerekhrest](https://stackoverflow.com/users/3185459/romanperekhrest" \o "56,180 reputation) [Jan 3 '17 at 8:42](https://stackoverflow.com/questions/41439419/how-to-select-max-and-min-value-in-a-row-for-selected-columns#comment70084953_41439419)

add a comment

1 Answer

[active](https://stackoverflow.com/questions/41439419/how-to-select-max-and-min-value-in-a-row-for-selected-columns?answertab=active#tab-top)[oldest](https://stackoverflow.com/questions/41439419/how-to-select-max-and-min-value-in-a-row-for-selected-columns?answertab=oldest#tab-top)[votes](https://stackoverflow.com/questions/41439419/how-to-select-max-and-min-value-in-a-row-for-selected-columns?answertab=votes#tab-top)

5

I think you need [min](http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.min.html) and [max](http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.max.html):

df\_subset=df.set\_index('CTYNAME')[['CENSUS2010POP', 'ESTIMATESBASE1010', 'POPESTIMATE2010']]

df1 = df\_subset.min(axis=1)

print (df1)

df2= df\_subset.max(axis=1)

print (df2)

EDIT:

df = pd.DataFrame({'CTYNAME':['Alabama','Autauga County','Baldwin County','Barbour County'],

'CENSUS2010POP':[4,5,6,2],

'ESTIMATESBASE1010':[7,8,9,3],

'POPESTIMATE2010':[1,3,5,5]})

print (df)

CENSUS2010POP CTYNAME ESTIMATESBASE1010 POPESTIMATE2010

0 4 Alabama 7 1

1 5 Autauga County 8 3

2 6 Baldwin County 9 5

3 2 Barbour County

df\_subset=df.set\_index('CTYNAME')[['CENSUS2010POP', 'ESTIMATESBASE1010', 'POPESTIMATE2010']]

df1 = df\_subset.max(axis=1) - df\_subset.min(axis=1)

print (df1)

CTYNAME

Alabama 6

Autauga County 5

Baldwin County 4

Barbour County 3

dtype: int64

print (df1.nlargest(1).reset\_index(name='top1'))

CTYNAME top1

0 Alabama 6

[share](https://stackoverflow.com/a/41439470)[improve this answer](https://stackoverflow.com/posts/41439470/edit)

[edited Jan 3 '17 at 9:11](https://stackoverflow.com/posts/41439470/revisions)

answered Jan 3 '17 at 8:41

[jezrael](https://stackoverflow.com/users/2901002/jezrael)

**356k**26321397

* how to do it to preserve entire CTYNAME? I want to find the difference between max and min and then pick CTYNAME with largest difference... – [YohanRoth](https://stackoverflow.com/users/3849781/yohanroth" \o "980 reputation) [Jan 3 '17 at 8:55](https://stackoverflow.com/questions/41439419/how-to-select-max-and-min-value-in-a-row-for-selected-columns#comment70085340_41439470)
* Also how to do it without creating duplicated subset – [YohanRoth](https://stackoverflow.com/users/3849781/yohanroth" \o "980 reputation) [Jan 3 '17 at 8:59](https://stackoverflow.com/questions/41439419/how-to-select-max-and-min-value-in-a-row-for-selected-columns#comment70085437_41439470)
* Do you think max difference in one column of 'CENSUS2010POP', 'ESTIMATESBASE1010', 'POPESTIMATE2010' ? – [jezrael](https://stackoverflow.com/users/2901002/jezrael" \o "356,368 reputation) [Jan 3 '17 at 9:00](https://stackoverflow.com/questions/41439419/how-to-select-max-and-min-value-in-a-row-for-selected-columns#comment70085469_41439470)
* Please check edit, is it correct? – [jezrael](https://stackoverflow.com/users/2901002/jezrael" \o "356,368 reputation) [Jan 3 '17 at 9:12](https://stackoverflow.com/questions/41439419/how-to-select-max-and-min-value-in-a-row-for-selected-columns#comment70085849_41439470)

Pandas Idioms

print(df.head())

df.where(df['Quantity']==0)

.rename(columns = {'Weight': 'Weight(oz.}))

print(df.drop(df[df['Quantity'] == 0].index).rename(columns={'Weight': 'Weight (oz.)'}))

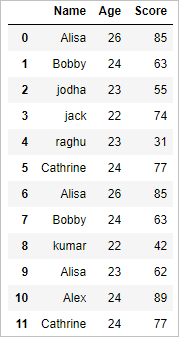
**Drop or delete column in python pandas**

In this tutorial we will learn how to  drop or delete column in python pandas by index,  drop column in pandas by name and drop column in python pandas by position. Lets see example of each.

**Create Dataframe**

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13 | import pandas as pd  import numpy as np    #Create a DataFrame  d = {      'Name':['Alisa','Bobby','jodha','jack','raghu','Cathrine',              'Alisa','Bobby','kumar','Alisa','Alex','Cathrine'],      'Age':[26,24,23,22,23,24,26,24,22,23,24,24],           'Score':[85,63,55,74,31,77,85,63,42,62,89,77]}    df = pd.DataFrame(d,columns=['Name','Age','Score'])  df |

The resultant dataframe will be

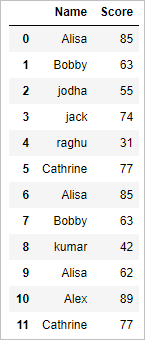


**Drop a column by name:**

Lets see an example of how to drop a column by name in python pandas

|  |  |
| --- | --- |
| 1  2  3 | # drop a column based on name    df.drop('Age',axis=1) |

The above code drops the column named ‘Age’, the argument axis=1 denotes column, so the resultant dataframe will be

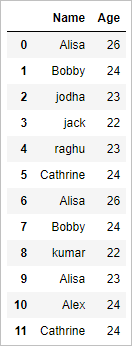


**Drop a column based on column index:**

Let’s see an example on dropping the column by its index in python pandas

|  |  |
| --- | --- |
| 1  2  3 | # drop a column based on column index    df.drop(df.columns[2],axis=1) |

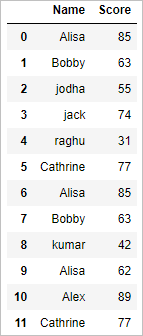
In the above example column with index 2 is dropped(3rd column). So the resultant dataframe will be



**Delete a column based on column name:**

|  |  |
| --- | --- |
| 1  2  3  4 | # delete a column    del df['Age']  df |

In the above example column with the name ‘Age’ is deleted. So the resultant dataframe will be



In an exercise, I was asked to merge 3 DataFrames with inner join (df1+df2+df3 = mergedDf), then in another question I was asked to tell how many entries I've lost when performing this 3-way merging.

#DataFrame1

df1 = pd.DataFrame(columns=["Goals","Medals"],data=[[5,2],[1,0],[3,1]])

df1.index = ['Argentina','Angola','Bolivia']

print(df1)

Goals Medals

Argentina 5 2

Angola 1 0

Bolivia 3 1

#DataFrame2

df2 = pd.DataFrame(columns=["Dates","Medals"],data=[[1,0],[2,1],[2,2])

df2.index = ['Venezuela','Africa']

print(df2)

Dates Medals

Venezuela 1 0

Africa 2 1

Argentina 2 2

#DataFrame3

df3 = pd.DataFrame(columns=["Players","Goals"],data=[[11,5],[11,1],[10,0]])

df3.index = ['Argentina','Australia','Belgica']

print(df3)

Players Goals

Argentina 11 5

Australia 11 1

Spain 10 0

#mergedDf

mergedDf = pd.merge(df1,df2,how='inner',left\_index=True, right\_index=True)

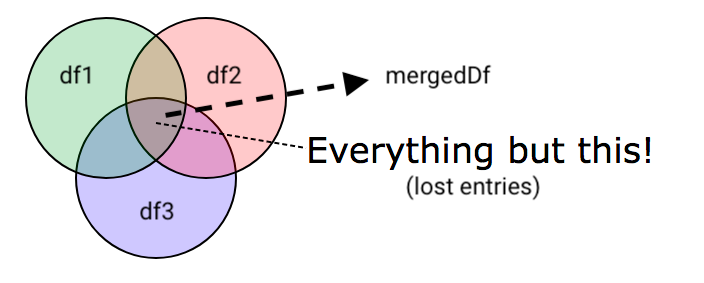
mergedDf = pd.merge(mergedDf,df3,how='inner',left\_index=True, right\_index=True)

print(mergedDF)

Goals\_X Medals\_X Dates Medals\_Y Players Goals\_Y

Argentina 5 2 2 2 11 2

#Calculate number of lost entries by code

I tried to merge everything with outer join and then subtracting the mergedDf, but I don't know how to do this, can anyone help me? [](https://i.stack.imgur.com/vQDFD.png)

[python](https://stackoverflow.com/questions/tagged/python) [pandas](https://stackoverflow.com/questions/tagged/pandas) [dataframe](https://stackoverflow.com/questions/tagged/dataframe)

[share](https://stackoverflow.com/q/53227491)|[improve this question](https://stackoverflow.com/posts/53227491/edit)

[edited Nov 9 '18 at 15:17](https://stackoverflow.com/posts/53227491/revisions)

Agustín Clemente

asked Nov 9 '18 at 14:21

[[https://lh3.googleusercontent.com/-o_bFqIGDRdg/AAAAAAAAAAI/AAAAAAAAABQ/V2L2aRdYG1w/photo.jpg?sz=32](https://stackoverflow.com/users/6161201/agust%c3%adn-clemente)](https://stackoverflow.com/users/6161201/agust%c3%adn-clemente)

[Agustín Clemente](https://stackoverflow.com/users/6161201/agust%c3%adn-clemente)Agustín Clemente

**370**1514

* Please post a [**Minimal**, Complete, and Verifiable example](https://stackoverflow.com/help/mcve). – [Alex](https://stackoverflow.com/users/1953800/alex) [Nov 9 '18 at 14:23](https://stackoverflow.com/questions/53227491/number-of-missing-entries-when-merging-dataframes#comment93341715_53227491)

add a comment |

3 Answers 3

[active](https://stackoverflow.com/questions/53227491/number-of-missing-entries-when-merging-dataframes?answertab=active#tab-top) [oldest](https://stackoverflow.com/questions/53227491/number-of-missing-entries-when-merging-dataframes?answertab=oldest#tab-top) [votes](https://stackoverflow.com/questions/53227491/number-of-missing-entries-when-merging-dataframes?answertab=votes#tab-top)

2

Solution with outer join and parameter indicator, last count rows with no both in both indicator columns a and b by sum of True values (processes like 1s):

mergedDf = pd.merge(df1,df2,how='outer',left\_index=True, right\_index=True, indicator='a')

mergedDf = pd.merge(mergedDf,df3,how='outer',left\_index=True, right\_index=True, indicator='b')

print(mergedDf)

Goals\_x Medals\_x Dates Medals\_y a Players Goals\_y \

Africa NaN NaN 2.0 1.0 right\_only NaN NaN

Angola 1.0 0.0 NaN NaN left\_only NaN NaN

Argentina 5.0 2.0 2.0 2.0 both 11.0 5.0

Australia NaN NaN NaN NaN NaN 11.0 1.0

Belgica NaN NaN NaN NaN NaN 10.0 0.0

Bolivia 3.0 1.0 NaN NaN left\_only NaN NaN

Venezuela NaN NaN 1.0 0.0 right\_only NaN NaN

b

Africa left\_only

Angola left\_only

Argentina both

Australia right\_only

Belgica right\_only

Bolivia left\_only

Venezuela left\_only

missing = ((mergedDf['a'] != 'both') & (mergedDf['b'] != 'both')).sum()

print (missing)

6

Another solution is use inner join and sum filtered values of each index which not matched mergedDf.index:

mergedDf = pd.merge(df1,df2,how='inner',left\_index=True, right\_index=True)

mergedDf = pd.merge(mergedDf,df3,how='inner',left\_index=True, right\_index=True)

vals = mergedDf.index

print (vals)

Index(['Argentina'], dtype='object')

dfs = [df1, df2, df3]

missing = sum((~x.index.isin(vals)).sum() for x in dfs)

print (missing)

6

Anoter solution if unique values in each index:

dfs = [df1, df2, df3]

L = [set(x.index) for x in dfs]

#https://stackoverflow.com/a/25324329/2901002

missing = len(set.union(\*L) - set.intersection(\*L))

print (missing)

6

[share](https://stackoverflow.com/a/53228308)|[improve this answer](https://stackoverflow.com/posts/53228308/edit)

[edited Nov 10 '18 at 7:59](https://stackoverflow.com/posts/53228308/revisions)

answered Nov 9 '18 at 15:08

[[https://i.stack.imgur.com/hMDvl.jpg?s=32&g=1](https://stackoverflow.com/users/2901002/jezrael)](https://stackoverflow.com/users/2901002/jezrael)

[jezrael](https://stackoverflow.com/users/2901002/jezrael)jezrael

**360k**26327407

add a comment |

1

You can passing True to the indicator in merge

df1=pd.DataFrame({'A':[1,2,3],'B':[1,1,1]})

df2=pd.DataFrame({'A':[2,3],'B':[1,1]})

df1.merge(df2,on='A',how='inner')

Out[257]:

A B\_x B\_y

0 2 1 1

1 3 1 1

df1.merge(df2,on='A',how='outer',indicator =True)

Out[258]:

A B\_x B\_y \_merge

0 1 1 NaN left\_only

1 2 1 1.0 both

2 3 1 1.0 both

mergedf=df1.merge(df2,on='A',how='outer',indicator =True)

Then with value\_counts you know how many you lost when do inner , since only the both will keep when how='inner'

mergedf['\_merge'].value\_counts()

Out[260]:

both 2

left\_only 1

right\_only 0

Name: \_merge, dtype: int64

For 3 df and filter with both merge columns words is both

df1.merge(df2, on='A',how='outer',indicator =True).rename(columns={'\_merge':'merge'}).merge(df3, on='A',how='outer',indicator =True)

[share](https://stackoverflow.com/a/53227679)|[improve this answer](https://stackoverflow.com/posts/53227679/edit)

[edited Nov 9 '18 at 15:28](https://stackoverflow.com/posts/53227679/revisions)

answered Nov 9 '18 at 14:31

[[https://i.stack.imgur.com/MfHh3.jpg?s=32&g=1](https://stackoverflow.com/users/7964527/wen-ben)](https://stackoverflow.com/users/7964527/wen-ben)

[Wen-Ben](https://stackoverflow.com/users/7964527/wen-ben)Wen-Ben

**126k**83872

* OP need I was asked to merge 3 DataFrames with inner join (df1+df2+df3 = mergedDf) – [jezrael](https://stackoverflow.com/users/2901002/jezrael" \o "360,183 reputation) [Nov 9 '18 at 15:08](https://stackoverflow.com/questions/53227491/number-of-missing-entries-when-merging-dataframes#comment93343320_53227679)
* @jezrael update – [Wen-Ben](https://stackoverflow.com/users/7964527/wen-ben) [Nov 9 '18 at 15:28](https://stackoverflow.com/questions/53227491/number-of-missing-entries-when-merging-dataframes#comment93343969_53227679)

add a comment |

1

I've found a simple but effective solution:

Merging the 3 DataFrames, inner and outer:

df1 = Df1()

df2 = Df2()

df3 = Df3()

inner = pd.merge(pd.merge(df1,df2,on='<Common column>',how='inner'),df3,on='<Common column>',how='inner')

outer = pd.merge(pd.merge(df1,df2,on='<Common column>',how='outer'),df3,on='<Common column>',how='outer')

Now, the number of missed entries (rows) is:

return (len(outer)-len(inner))

**How It Works**

For example, per capita GDP is a country’s [gross domestic product (GDP)](http://www.investinganswers.com/node/1223) per person. The formula for per capita [GDP](http://www.investinganswers.com/node/1223) is:

Per Capita GDP = Gross Domestic Product/Population

So, let's assume that Country XYZ has $100 trillion in gross domestic product and 250 million people. According to the formula, Country XYZ's net debt per capita is:

Per Capita GDP = $100,000,000,000,000/250,000,000 = $400,000

**Why It Matters**

Per capita measures help [analysts](http://www.investinganswers.com/node/5331) and investors get a better feel for whether a company, country, or other entity is productive, efficient, or profitable. For instance, the per capita measure of GDP indicates whether the country’s workforce is generally becoming more or less productive – that is, whether the country’s workforce is efficiently producing goods and services that consumers want.