Merging - due 10/19 *by Larissa & Charlotte*

Pre-Legalization Analysis

Exercise 1

We will begin by examining county-level data on arrests from California in 2009, which is derived from data provided by the Office of the California State Attorney General here. Download and import the file ca_arrests_2009.csv from the first link above (the second one is just to show you where it came from).

In [ ]:
```
import pandas as pd
cali = pd.read_csv("https://raw.githubusercontent.com/nickeubank/practicalc
```

Exercise 2

Use your data exploration skills to get a feel for this data. If you need to, you can find the original codebook here (This data is similar, but has been collapsed to one observation per county.)

In [ ]:
```
cali.sample(20)
cali.describe()
cali.isna().sum()
cali['COUNTY'].value_counts()
```

Out[ ]:
```
Alameda County            1
Santa Cruz County         1
Plumas County             1
Riverside County          1
Sacramento County         1
San Benito County         1
San Bernardino County     1
San Diego County          1
San Francisco County      1
San Joaquin County        1
San Luis Obispo County    1
San Mateo County          1
Santa Barbara County      1
Santa Clara County        1
Shasta County             1
Alpine County             1
Sierra County             1
```

```
Siskiyou County           1
Solano County             1
Sonoma County             1
Stanislaus County         1
Sutter County             1
Tehama County             1
Trinity County            1
Tulare County             1
Tuolumne County           1
Ventura County            1
Yolo County               1
Placer County             1
Orange County             1
Nevada County             1
Napa County               1
Amador County             1
Butte County              1
Calaveras County          1
Colusa County             1
Contra Costa County       1
Del Norte County          1
El Dorado County          1
Fresno County             1
Glenn County              1
Humboldt County           1
Imperial County           1
Inyo County               1
Kern County               1
Kings County              1
Lake County               1
Lassen County             1
Los Angeles County        1
Madera County             1
Marin County              1
Mariposa County           1
Mendocino County          1
Merced County             1
Modoc County              1
Mono County               1
Monterey County           1
Yuba County               1
Name: COUNTY, dtype: int64
```

Exercise 3

Figuring out what county has the most violent arrests isn't very meaningful if we don't normalize for size. A county with 10 people and 10 arrests for violent crimes is obviously worse than a county with 1,000,000 people an 11 arrests for violent crime.

To address this, also import nhgis_county_populations.csv from the directory we're working from.

In [ ]:
```
pop = pd.read_csv("https://raw.githubusercontent.com/nickeubank/practicalda
```

Exercise 4

Use your data exploration skills to get used to this data, and figure out how it relates to your 2009 arrest data.

In [ ]:
```
pop.sample(20)
pop.describe()
pop.isna().sum()
pop['COUNTY'].value_counts()
```

Out[ ]:
```
Washington County       60
Jefferson County        50
Franklin County         48
Jackson County          46
Lincoln County          46
                        ..
Bedford city             1
Kusilvak Census Area     1
Petersburg Borough       1
LaSalle Parish           1
Oglala Lakota County     1
Name: COUNTY, Length: 1959, dtype: int64
```

- They hold information on the population sizes of US counties in the time period 2005-2009. That relates to our first dataframe, as we have info on the county-level as well.

Exercise 5

Once you feel like you have a good sense of the relation between our arrest and population data, merge the two datasets.

In [ ]:
```
merged = pd.merge(cali, pop, on="COUNTY")
merged
```

Out[ ]:

	Unnamed: 0_x	COUNTY	VIOLENT	PROPERTY	F_DRUGOFF	F_SEXOFF	F_ALLOTHER	F_T
0	1682	Alameda County	4318	4640	5749	260	3502	1
1	1682	Alameda County	4318	4640	5749	260	3502	1
2	1683	Alpine County	8	4	2	1	1	
3	1683	Alpine County	8	4	2	1	1	
4	1684	Amador County	100	59	101	5	199	
...	...	...	...	...	...	...	...	
161	1737	Ventura County	2275	2425	2040	120	2083	
162	1738	Yolo County	585	634	614	39	662	
163	1738	Yolo County	585	634	614	39	662	
164	1739	Yuba County	354	368	211	39	257	
165	1739	Yuba County	354	368	211	39	257	

166 rows × 14 columns

Checking Your Merges

Exercise 6

When you merge data, you have to make some assumptions about the nature of the data you're working with. For example, you have to assume it's OK to connect variables that share the same value of your merging variables. Similarly, you have to make assumptions about whether your merge a 1-to-1 merge (meaning there is only one observation of the variable you're merging on in both datasets), a 1-to-many merge (meaning there is only one observation of the variable you're merging on in the first dataset, but multiple observations in the second). So before running a merge, answer the following questions:

What variable do you think will be consistent across these two datasets you can use for merging?

- The COUNTY variable.

Do you think there will be exactly 1 observation for each value in your arrest data?

- YES, we checked.

Do you think there will be exactly 1 observation for each value in your population data?

- No, apparently for some counties this is not the case.

Being correct in your assumptions about these things is very important. If you think there's only one observation per value of your merging variable in each dataset, but there are in fact 2, you'll end up with two observations for each value after the merge. Moreover, not only is the structure of your data now a mess, but the fact you were wrong means you didn't understand something about your data.

Because of the importance of this, pandas provides a utility for testing these assumptions when you do a merge: the validate keyword! Validate will accept "1:1", "1:m", "m:1", and "m:m". It will then check to make sure your merge matches the type of merge you think it is. I highly recommend always using this option (...and not just because I'm the one who added validate to pandas).

Repeat the merge you conducted above, but this time use the validate to make sure your assumptions about the data were correct.

In [ ]:
```
merged = pd.merge(cali, pop, on ="COUNTY", validate="one_to_one")
```

```
---------------------------------------------------------------------------
MergeError                                Traceback (most recent call last)
/var/folders/h9/g02cmrsn6y571zb1v96gs56c0000gn/T/ipykernel_4804/1740390590.
py in <module>
----> 1 merged = pd.merge(cali, pop, on ="COUNTY", validate="one_to_one")

~/miniforge3/lib/python3.9/site-packages/pandas/core/reshape/merge.py in me
rge(left, right, how, on, left_on, right_on, left_index, right_index, sort,
suffixes, copy, indicator, validate)
    105     validate: str | None = None,
    106 ) -> DataFrame:
--> 107     op = _MergeOperation(
    108         left,
    109         right,

~/miniforge3/lib/python3.9/site-packages/pandas/core/reshape/merge.py in __
init__(self, left, right, how, on, left_on, right_on, axis, left_index, rig
ht_index, sort, suffixes, copy, indicator, validate)
    708         # are in fact unique.
    709         if validate is not None:
--> 710             self._validate(validate)
    711
    712     def get_result(self) -> DataFrame:

~/miniforge3/lib/python3.9/site-packages/pandas/core/reshape/merge.py in _v
alidate(self, validate)
   1423                 )
   1424             elif not right_unique:
-> 1425                 raise MergeError(
   1426                     "Merge keys are not unique in right dataset; no
t a one-to-one merge"
   1427                 )

MergeError: Merge keys are not unique in right dataset; not a one-to-one me
rge
```

In [ ]:

```
print("Apparently this is not a one-to-one merge. As we figured before, the
```

```
Apparently this is not a one-to-one merge. As we figured before, the merge
key (COUNTY) in the right dataframe (population data) is not unique. Some c
ounties exist in several states. But we are anyways only interested in Cali
.
```

In [ ]:

```
print("Subset for the right time and the state of California and check agai
pop2009 = pop[pop['YEAR'] == '2005-2009'].copy()
ca_pop2009 = pop2009[pop2009.STATE == 'California'].copy()
merged_2009 = pd.merge(cali, ca_pop2009, on ='COUNTY', validate='1:1')
```

```
Subset for the right time and the state of California and check again if 1:
1 merging is now validated.
```

Now the 1:1 merging is validated!

Exercise 7 and 8

Now repeat your previous merge using both the validate keyword and the indicator keyword with how='outer'.

In [ ]:
```
merged_2009 = pd.merge(cali, ca_pop2009, on ="COUNTY", validate="one_to_mar
merged_2009
merged_2009._merge.value_counts()
```

Out[ ]:
```
both          56
left_only      2
right_only     2
Name: _merge, dtype: int64
```

- We find that we merged on 56 observations, while 2 rows from the right dataframe did not merge and 2 from the left dataframe did not merge.

Exercise 9

You should be able to get to the point that all counties in our arrest data merge with population data. Can you figure out why that did not happen? Can you fix the data so that they all merge to population data?

In [ ]:
```
merged_2009[merged_2009._merge == "left_only"]
```

Out[ ]:

	Unnamed: 0_x	COUNTY	VIOLENT	PROPERTY	F_DRUGOFF	F_SEXOFF	F_ALLOTHER	F_TO
7	1689.0	Del Norte County	144.0	104.0	79.0	13.0	97.0	4
13	1695.0	Inyo County	81.0	44.0	39.0	3.0	38.0	2(

In [ ]:
```
merged_2009[merged_2009._merge == "right_only"]
```

Out[ ]:

	Unnamed: 0_x	COUNTY	VIOLENT	PROPERTY	F_DRUGOFF	F_SEXOFF	F_ALLOTHER	F_TC
58	NaN	DelNorte County	NaN	NaN	NaN	NaN	NaN	
59	NaN	Injo County	NaN	NaN	NaN	NaN	NaN	

- We find that two rows from both dataframes have not merged. When investigating that further we find that both dataframes include two counties where they differ in the spelling. We need to correct for that so we can merge on them.

In [ ]:
```
ca_pop2009['COUNTY'] = ca_pop2009['COUNTY'].replace(['DelNorte County'],'De
ca_pop2009['COUNTY'] = ca_pop2009['COUNTY'].replace(['Injo County'],'Inyo C
```

In [ ]:
```
print("Let's try again")
merged_2009 = pd.merge(cali, ca_pop2009, on ="COUNTY", validate="one_to_one
merged_2009
merged_2009._merge.value_counts()
```

Out[ ]:
```
Let's try again
both          58
left_only      0
right_only     0
Name: _merge, dtype: int64
```

Comparing Arrest Rates

Exercise 10

Now that we have arrest counts and population data, we can calculate arrest rates. For each county, create a new variable called violent_arrest_rate_2009 that is the number of violent arrests for 2009 divided by the population of the county from 2005-2009, and an analogous variable for drug offenses (F_DRUGOFF).

In [ ]:
```
merged_2009["violent_arrest_rate_2009"] = merged_2009["VIOLENT"]/merged_200
merged_2009["violent_arrest_rate_2009"]
```

Out[ ]:
```
0     0.002963
1     0.006938
2     0.002629
3     0.002941
4     0.004533
5     0.002762
6     0.002930
7     0.005012
8     0.003240
9     0.004914
10    0.003729
11    0.003899
12    0.003743
13    0.004645
14    0.005493
15    0.002659
16    0.005389
```

```
17    0.002934
18    0.003609
19    0.003129
20    0.002027
21    0.003582
22    0.004824
23    0.004826
24    0.005130
25    0.004023
26    0.003420
27    0.002777
28    0.002174
29    0.002064
30    0.002144
31    0.004866
32    0.002861
33    0.003854
34    0.003817
35    0.004266
36    0.003284
37    0.004552
38    0.004849
39    0.002323
40    0.002060
41    0.003214
42    0.002492
43    0.002995
44    0.002587
45    0.004321
46    0.004256
47    0.003934
48    0.002928
49    0.004377
50    0.004695
51    0.003894
52    0.004669
53    0.005244
54    0.002869
55    0.002871
56    0.003031
57    0.004993
Name: violent_arrest_rate_2009, dtype: float64
```

In [ ]:
```
merged_2009["drug_offenses_2009"] = merged_2009["F_DRUGOFF"]/merged_2009["t
merged_2009["drug_offenses_2009"]
```

Out[ ]:
```
0     0.003946
1     0.001735
2     0.002655
3     0.002487
4     0.002642
5     0.001333
6     0.002851
7     0.002750
8     0.002291
9     0.003938
```

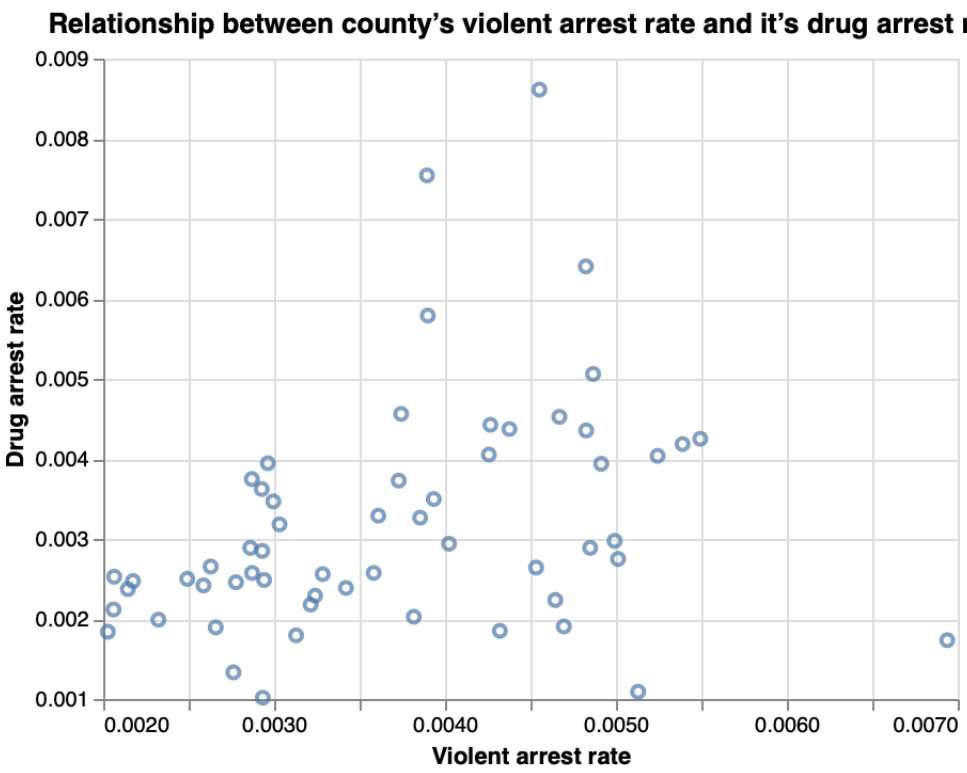
```
10    0.003729
11    0.005791
12    0.004562
13    0.002236
14    0.004251
15    0.001895
16    0.004185
17    0.001017
18    0.003290
19    0.001796
20    0.001840
21    0.002575
22    0.006405
23    0.004355
24    0.001091
25    0.002940
26    0.002388
27    0.002459
28    0.002473
29    0.002528
30    0.002373
31    0.005061
32    0.002888
33    0.003265
34    0.002027
35    0.004426
36    0.002560
37    0.008613
38    0.002889
39    0.001991
40    0.002119
41    0.002179
42    0.002501
43    0.003469
44    0.002419
45    0.001852
46    0.004054
47    0.003498
48    0.003623
49    0.004373
50    0.001907
51    0.007541
52    0.004525
53    0.004038
54    0.003748
55    0.002575
56    0.003182
57    0.002976
Name: drug_offenses_2009, dtype: float64
```

Exercise 11

Make a scatter plot that shows the relationship between each county's violent arrest rate and it's drug arrest rate.

In [ ]:
```
import altair as alt
alt.Chart(merged_2009, title="Relationship between county's violent arrest
    x=alt.X("violent_arrest_rate_2009", title="Violent arrest rate", scale=
    y=alt.Y("drug_offenses_2009", title="Drug arrest rate", scale=alt.Scale
)
```

Out[ ]:



Comparing with 2018 Arrests

Exercise 12

Just as we created violent arrest rates and drug arrest rates for 2009, now we want to do it for 2018. Using the data on 2018 arrests (also in the same repository we used before) and the same dataset of population data (you'll have to use population from 2013-2017, as 2018 population data has yet to be released), create a dataset of arrest rates.

As before, be careful with your merges!!!

In [ ]:
```
print("Load and get a feel for the new data.")
cali18 = pd.read_csv("https://raw.githubusercontent.com/nickeubank/practica
cali18.sample(20)
cali18.describe()
cali18.isna().sum()
cali18['COUNTY'].value_counts()
```

```
Load and get a feel for the new data.
Alameda County            1
```

Out[ ]:

```
Out[ ]:
```

Santa Cruz County	1
Plumas County	1
Riverside County	1
Sacramento County	1
San Benito County	1
San Bernardino County	1
San Diego County	1
San Francisco County	1
San Joaquin County	1
San Luis Obispo County	1
San Mateo County	1
Santa Barbara County	1
Santa Clara County	1
Shasta County	1
Alpine County	1
Sierra County	1
Siskiyou County	1
Solano County	1
Sonoma County	1
Stanislaus County	1
Sutter County	1
Tehama County	1
Trinity County	1
Tulare County	1
Tuolumne County	1
Ventura County	1
Yolo County	1
Placer County	1
Orange County	1
Nevada County	1
Napa County	1
Amador County	1
Butte County	1
Calaveras County	1
Colusa County	1
Contra Costa County	1
Del Norte County	1
El Dorado County	1
Fresno County	1
Glenn County	1
Humboldt County	1
Imperial County	1
Inyo County	1
Kern County	1
Kings County	1
Lake County	1
Lassen County	1
Los Angeles County	1
Madera County	1
Marin County	1
Mariposa County	1
Mendocino County	1
Merced County	1
Modoc County	1
Mono County	1
Monterey County	1
Yuba County	1

```
Name: COUNTY, dtype: int64
```

In [ ]:
```
pop2018 = pop[pop['YEAR'] == '2013-2017'].copy()
ca_pop2018 = pop2018[pop2018.STATE == 'California'].copy()
```

```
Adjust population dataframe for the time frame and the state of California.
```

In [ ]:
```
ca_pop2018['COUNTY'] = ca_pop2018['COUNTY'].replace(['DelNorte County'],'De
ca_pop2018['COUNTY'] = ca_pop2018['COUNTY'].replace(['Injo County'],'Inyo C
```

```
Adjust the names of the counties to prevent the merging failure from the fi
rst instance to repeat here.
```

In [ ]:
```
merged_2018 = pd.merge(cali, ca_pop2018, on ="COUNTY", validate="one_to_one
merged_2018
merged_2018._merge.value_counts()
```

```
Merge and run diagnostics and validation checks.
```

Out[ ]:
```
both          58
left_only      0
right_only     0
Name: _merge, dtype: int64
```

All merging went fine!

Exercise 13

Now merge the two county-level datasets so you have one row for every county, and variables for violent arrest rates in 2018, violent arrest rates in 2009, felony drug arrest rates in 2018, and felony drug arrest rates in 2009.

In [ ]:
```
merged_2018["violent_arrest_rate_2018"] = merged_2018["VIOLENT"]/merged_201
merged_2018["violent_arrest_rate_2018"]
```

```
Create violent arrest rate variable, standardized over population for the y
ear 2018
```

Out[ ]:
```
0     0.002650
1     0.006650
2     0.002681
3     0.002846
4     0.004683
5     0.002700
6     0.002648
7     0.005247
8     0.003081
9     0.004505
10    0.003723
11    0.003712
12    0.003329
13    0.004452
14    0.004882
```

```
15    0.002597
16    0.005445
17    0.003209
18    0.003495
19    0.002933
20    0.001917
21    0.003624
22    0.004743
23    0.004372
24    0.005212
25    0.003699
26    0.003197
27    0.002603
28    0.002135
29    0.001947
30    0.001899
31    0.005341
32    0.002473
33    0.003546
34    0.003562
35    0.003995
36    0.002988
37    0.004199
38    0.004451
39    0.002174
40    0.001894
41    0.002917
42    0.002255
43    0.002756
44    0.002593
45    0.004853
46    0.004342
47    0.003676
48    0.002713
49    0.004127
50    0.004457
51    0.003731
52    0.004986
53    0.004758
54    0.002969
55    0.002683
56    0.002752
57    0.004743
Name: violent_arrest_rate_2018, dtype: float64
```

In [ ]:

```
merged_2018["drug_offenses_2018"] = merged_2018["F_DRUGOFF"]/merged_2018["t
merged_2018["drug_offenses_2018"]
```

Create felony drug arrest rate variable, standardized over population for t
he year 2018

Out[ ]:

```
0     0.003528
1     0.001663
2     0.002707
3     0.002407
4     0.002730
5     0.001304
```

```
6     0.002576
7     0.002879
8     0.002178
9     0.003610
10    0.003723
11    0.005513
12    0.004057
13    0.002143
14    0.003778
15    0.001851
16    0.004228
17    0.001112
18    0.003186
19    0.001684
20    0.001741
21    0.002605
22    0.006297
23    0.003946
24    0.001109
25    0.002703
26    0.002232
27    0.002305
28    0.002428
29    0.002384
30    0.002101
31    0.005554
32    0.002497
33    0.003004
34    0.001892
35    0.004145
36    0.002329
37    0.007945
38    0.002651
39    0.001863
40    0.001948
41    0.001977
42    0.002263
43    0.003191
44    0.002426
45    0.002080
46    0.004135
47    0.003269
48    0.003358
49    0.004124
50    0.001810
51    0.007226
52    0.004832
53    0.003664
54    0.003878
55    0.002406
56    0.002888
57    0.002827
Name: drug_offenses_2018, dtype: float64
```

In [ ]:

```
print("Subset the merged_2009 dataframe for the rows we want:")
merged_2009small = merged_2009.loc[:, (merged_2009.columns == 'COUNTY') | (
merged_2009small
```

```
Subset the merged_2009 dataframe for the rows we want:
```

Out[ ]:

	COUNTY	VIOLENT	F_DRUGOFF	violent_arrest_rate_2009	drug_offenses_2009
0	Alameda County	4318	5749	0.002963	0.003946
1	Alpine County	8	2	0.006938	0.001735
2	Amador County	100	101	0.002629	0.002655
3	Butte County	641	542	0.002941	0.002487
4	Calaveras County	211	123	0.004533	0.002642
5	Colusa County	58	28	0.002762	0.001333
6	Contra Costa County	2976	2895	0.002930	0.002851
7	Del Norte County	144	79	0.005012	0.002750
8	El Dorado County	570	403	0.003240	0.002291
9	Fresno County	4377	3508	0.004914	0.003938
10	Glenn County	104	104	0.003729	0.003729
11	Humboldt County	503	747	0.003899	0.005791
12	Imperial County	599	730	0.003743	0.004562
13	Inyo County	81	39	0.004645	0.002236
14	Kern County	4290	3320	0.005493	0.004251
15	Kings County	390	278	0.002659	0.001895
16	Lake County	349	271	0.005389	0.004185
17	Lassen County	101	35	0.002934	0.001017
18	Los Angeles County	35319	32193	0.003609	0.003290

19	Madera County	453	260	0.003129	0.001796
20	Marin County	500	454	0.002027	0.001840
21	Mariposa County	64	46	0.003582	0.002575
22	Mendocino County	415	551	0.004824	0.006405
23	Merced County	1169	1055	0.004826	0.004355
24	Modoc County	47	10	0.005130	0.001091
25	Mono County	52	38	0.004023	0.002940
26	Monterey County	1385	967	0.003420	0.002388
27	Napa County	367	325	0.002777	0.002459
28	Nevada County	211	240	0.002174	0.002473
29	Orange County	6145	7524	0.002064	0.002528
30	Placer County	712	788	0.002144	0.002373
31	Plumas County	100	104	0.004866	0.005061
32	Riverside County	5825	5881	0.002861	0.002888
33	Sacramento County	5302	4492	0.003854	0.003265
34	San Benito County	209	111	0.003817	0.002027
35	San Bernardino County	8474	8793	0.004266	0.004426
36	San Diego County	9812	7648	0.003284	0.002560
37	San Francisco County	3629	6867	0.004552	0.008613
38	San Joaquin County	3223	1920	0.004849	0.002889
	San Luis				

39	Obispo County	609	522	0.002323	0.001991
40	San Mateo County	1446	1487	0.002060	0.002119
41	Santa Barbara County	1292	876	0.003214	0.002179
42	Santa Clara County	4309	4325	0.002492	0.002501
43	Santa Cruz County	753	872	0.002995	0.003469
44	Shasta County	464	434	0.002587	0.002419
45	Sierra County	14	6	0.004321	0.001852
46	Siskiyou County	189	180	0.004256	0.004054
47	Solano County	1599	1422	0.003934	0.003498
48	Sonoma County	1359	1682	0.002928	0.003623
49	Stanislaus County	2211	2209	0.004377	0.004373
50	Sutter County	426	173	0.004695	0.001907
51	Tehama County	236	457	0.003894	0.007541
52	Trinity County	65	63	0.004669	0.004525
53	Tulare County	2183	1681	0.005244	0.004038
54	Tuolumne County	160	209	0.002869	0.003748
55	Ventura County	2275	2040	0.002871	0.002575
56	Yolo County	585	614	0.003031	0.003182
57	Yuba County	354	211	0.004993	0.002976

In [ ]:

```
print("Subset the merged_2018 dataframe for the rows we want:")
merged_2018small = merged_2018.loc[:, (merged_2018.columns == 'COUNTY') | (
merged_2018small
```

```
Subset the merged_2018 dataframe for the rows we want:
```

Out[ ]:

Out[ ]:

	COUNTY	VIOLENT	F_DRUGOFF	violent_arrest_rate_2018	drug_offenses_2018
0	Alameda County	4318	5749	0.002650	0.003528
1	Alpine County	8	2	0.006650	0.001663
2	Amador County	100	101	0.002681	0.002707
3	Butte County	641	542	0.002846	0.002407
4	Calaveras County	211	123	0.004683	0.002730
5	Colusa County	58	28	0.002700	0.001304
6	Contra Costa County	2976	2895	0.002648	0.002576
7	Del Norte County	144	79	0.005247	0.002879
8	El Dorado County	570	403	0.003081	0.002178
9	Fresno County	4377	3508	0.004505	0.003610
10	Glenn County	104	104	0.003723	0.003723
11	Humboldt County	503	747	0.003712	0.005513
12	Imperial County	599	730	0.003329	0.004057
13	Inyo County	81	39	0.004452	0.002143
14	Kern County	4290	3320	0.004882	0.003778
15	Kings County	390	278	0.002597	0.001851
16	Lake County	349	271	0.005445	0.004228
17	Lassen County	101	35	0.003209	0.001112
18	Los Angeles County	35319	32193	0.003495	0.003186
19	Madera County	453	260	0.002933	0.001684
20	Marin County	500	454	0.001917	0.001741
21	Mariposa County	64	46	0.003624	0.002605
22	Mendocino County	415	551	0.004743	0.006297
23	Merced	1169	1055	0.004372	0.003946

	County				
24	Modoc County	47	10	0.005212	0.001109
25	Mono County	52	38	0.003699	0.002703
26	Monterey County	1385	967	0.003197	0.002232
27	Napa County	367	325	0.002603	0.002305
28	Nevada County	211	240	0.002135	0.002428
29	Orange County	6145	7524	0.001947	0.002384
30	Placer County	712	788	0.001899	0.002101
31	Plumas County	100	104	0.005341	0.005554
32	Riverside County	5825	5881	0.002473	0.002497
33	Sacramento County	5302	4492	0.003546	0.003004
34	San Benito County	209	111	0.003562	0.001892
35	San Bernardino County	8474	8793	0.003995	0.004145
36	San Diego County	9812	7648	0.002988	0.002329
37	San Francisco County	3629	6867	0.004199	0.007945
38	San Joaquin County	3223	1920	0.004451	0.002651
39	San Luis Obispo County	609	522	0.002174	0.001863
40	San Mateo County	1446	1487	0.001894	0.001948
41	Santa Barbara County	1292	876	0.002917	0.001977
42	Santa Clara County	4309	4325	0.002255	0.002263
43	Santa Cruz County	753	872	0.002756	0.003191
	Shasta				

44	County	464	434	0.002593	0.002426
45	Sierra County	14	6	0.004853	0.002080
46	Siskiyou County	189	180	0.004342	0.004135
47	Solano County	1599	1422	0.003676	0.003269
48	Sonoma County	1359	1682	0.002713	0.003358
49	Stanislaus County	2211	2209	0.004127	0.004124
50	Sutter County	426	173	0.004457	0.001810
51	Tehama County	236	457	0.003731	0.007226
52	Trinity County	65	63	0.004986	0.004832
53	Tulare County	2183	1681	0.004758	0.003664
54	Tuolumne County	160	209	0.002969	0.003878
55	Ventura County	2275	2040	0.002683	0.002406
56	Yolo County	585	614	0.002752	0.002888
57	Yuba County	354	211	0.004743	0.002827

In [ ]:
```
merged_total = pd.merge(merged_2009small, merged_2018small, on ="COUNTY", v
merged_total
```

Out[ ]:

	COUNTY	VIOLENT_x	F_DRUGOFF_x	violent_arrest_rate_2009	drug_offenses_2009
0	Alameda County	4318	5749	0.002963	0.003946
1	Alpine County	8	2	0.006938	0.001735
2	Amador County	100	101	0.002629	0.002655
3	Butte County	641	542	0.002941	0.002487
4	Calaveras County	211	123	0.004533	0.002642
5	Colusa County	58	28	0.002762	0.001333

6	Contra Costa County	2976	2895	0.002930	0.002851
7	Del Norte County	144	79	0.005012	0.002750
8	El Dorado County	570	403	0.003240	0.002291
9	Fresno County	4377	3508	0.004914	0.003938
10	Glenn County	104	104	0.003729	0.003729
11	Humboldt County	503	747	0.003899	0.005791
12	Imperial County	599	730	0.003743	0.004562
13	Inyo County	81	39	0.004645	0.002236
14	Kern County	4290	3320	0.005493	0.004251
15	Kings County	390	278	0.002659	0.001895
16	Lake County	349	271	0.005389	0.004185
17	Lassen County	101	35	0.002934	0.001017
18	Los Angeles County	35319	32193	0.003609	0.003290
19	Madera County	453	260	0.003129	0.001796
20	Marin County	500	454	0.002027	0.001840
21	Mariposa County	64	46	0.003582	0.002575
22	Mendocino County	415	551	0.004824	0.006405
23	Merced County	1169	1055	0.004826	0.004355
24	Modoc County	47	10	0.005130	0.001091
25	Mono County	52	38	0.004023	0.002940
26	Monterey County	1385	967	0.003420	0.002388

27	Napa County	367	325	0.002777	0.002459
28	Nevada County	211	240	0.002174	0.002473
29	Orange County	6145	7524	0.002064	0.002528
30	Placer County	712	788	0.002144	0.002373
31	Plumas County	100	104	0.004866	0.005061
32	Riverside County	5825	5881	0.002861	0.002888
33	Sacramento County	5302	4492	0.003854	0.003265
34	San Benito County	209	111	0.003817	0.002027
35	San Bernardino County	8474	8793	0.004266	0.004426
36	San Diego County	9812	7648	0.003284	0.002560
37	San Francisco County	3629	6867	0.004552	0.008613
38	San Joaquin County	3223	1920	0.004849	0.002889
39	San Luis Obispo County	609	522	0.002323	0.001991
40	San Mateo County	1446	1487	0.002060	0.002119
41	Santa Barbara County	1292	876	0.003214	0.002179
42	Santa Clara County	4309	4325	0.002492	0.002501
43	Santa Cruz County	753	872	0.002995	0.003469
44	Shasta County	464	434	0.002587	0.002419
45	Sierra County	14	6	0.004321	0.001852
46	Siskiyou County	189	180	0.004256	0.004054

47	Solano County	1599	1422	0.003934	0.003498
48	Sonoma County	1359	1682	0.002928	0.003623
49	Stanislaus County	2211	2209	0.004377	0.004373
50	Sutter County	426	173	0.004695	0.001907
51	Tehama County	236	457	0.003894	0.007541
52	Trinity County	65	63	0.004669	0.004525
53	Tulare County	2183	1681	0.005244	0.004038
54	Tuolumne County	160	209	0.002869	0.003748
55	Ventura County	2275	2040	0.002871	0.002575
56	Yolo County	585	614	0.003031	0.003182
57	Yuba County	354	211	0.004993	0.002976

Exercise 14

Did drug arrests go down from 2009 to 2018? (they sure better! This is what's called a "sanity check" of your data and analysis. If you find drug arrests went up, you know something went wrong with your code or your understanding of the situations.

In [ ]:
```
print("Check if drug arrests went down from 2009 to 2018:")
diff = merged_total["drug_offenses_2018"] - merged_total["drug_offenses_20(
avg_diff = diff.mean()
print(f"The average change from 2009 to 2018 is {avg_diff}, standardized ov
```

```
Check if drug arrests went down from 2009 to 2018:
The average change from 2009 to 2018 is -0.0001416050877654095 2, standardiz
ed over the total population.
```

On average when substracting the drug_offenses_2009 from the drug_offenses_2018 we find that the difference is negative, i.e. on average the rates were higher in 2009. The difference is fairly marginal, but we have to be aware of the standardization we conducted over the total population. Let's see how many more arrests we can expect for a population of 100,000:

In [ ]:
```
scale = avg_diff*100000
print(f"For a population of 100,000 on average {scale.round(2)*(-1)} more c
```

```
For a population of 100,000 on average 14.16 more drug arrests were recorde
d in 2009 vs. 2018.
```

Exercise 15

Now we want to look at whether violent crime decreased following drug legalization. Did the average violent arrest rate decrease? By how much? (Note: We're assuming that arrest rates are proportionate to crime rates. If policing increased so that there were more arrests per crime committed, that would impact our interpretation of these results. But this is just an exercise, so...)

In [ ]:
```
print("Check if violent arrests went down from 2009 to 2018:")
diff_violent = merged_total["violent_arrest_rate_2018"] - merged_total["vio
avg_diff_violent = diff_violent.mean()
print(f"The average change from 2009 to 2018 is {avg_diff_violent}, standar
scale2 = avg_diff_violent*100000
print(f"For a population of 100,000 on average {scale2.round(2)*(-1)} more
```

```
Check if violent arrests went down from 2009 to 2018:
The average change from 2009 to 2018 is -0.00014546435087355145, standardiz
ed over the total population.
For a population of 100,000 on average 14.55 more violent arrests were reco
rded in 2009 vs. 2018.
```

FINDING

We find that both rates go down from 2009 to 2018.

Exercise 16

So let's split our sample into two groups: high drug arrests in 2009, and low drug arrests in 2009 (cut the sample at the average drug arrest rate in 2009).

In [ ]:
```
mean_drug_2009 = merged_total["drug_offenses_2009"].mean()
mean_drug_2009
```

Out[ ]:
```
0.003191448015158338
```

In [ ]:
```
print("Splitting the sample. First into those with low drug arrests in 2009
low = merged_total.loc[merged_total["drug_offenses_2009"] < mean_drug_2009]
low
```

```
Splitting the sample. First into those with low drug arrests in 2009:
```

Out[ ]: COUNTY VIOLENT x F_DRUGOFF x violent_arrest_rate_2009 drug_offenses_2009 \

Out[ ]:

	COUNTY	VIOLENT_x	P_DRUGOFF_x	violent_arrest_rate_2009	drug_offenses_2009	V
1	Alpine County	8	2	0.006938	0.001735	
2	Amador County	100	101	0.002629	0.002655	
3	Butte County	641	542	0.002941	0.002487	
4	Calaveras County	211	123	0.004533	0.002642	
5	Colusa County	58	28	0.002762	0.001333	
6	Contra Costa County	2976	2895	0.002930	0.002851	
7	Del Norte County	144	79	0.005012	0.002750	
8	El Dorado County	570	403	0.003240	0.002291	
13	Inyo County	81	39	0.004645	0.002236	
15	Kings County	390	278	0.002659	0.001895	
17	Lassen County	101	35	0.002934	0.001017	
19	Madera County	453	260	0.003129	0.001796	
20	Marin County	500	454	0.002027	0.001840	
21	Mariposa County	64	46	0.003582	0.002575	
24	Modoc County	47	10	0.005130	0.001091	
25	Mono County	52	38	0.004023	0.002940	
26	Monterey County	1385	967	0.003420	0.002388	
27	Napa County	367	325	0.002777	0.002459	
28	Nevada County	211	240	0.002174	0.002473	
29	Orange County	6145	7524	0.002064	0.002528	
30	Placer County	712	788	0.002144	0.002373	

32	Riverside County	5825	5881	0.002861	0.002888
34	San Benito County	209	111	0.003817	0.002027
36	San Diego County	9812	7648	0.003284	0.002560
38	San Joaquin County	3223	1920	0.004849	0.002889
39	San Luis Obispo County	609	522	0.002323	0.001991
40	San Mateo County	1446	1487	0.002060	0.002119
41	Santa Barbara County	1292	876	0.003214	0.002179
42	Santa Clara County	4309	4325	0.002492	0.002501
44	Shasta County	464	434	0.002587	0.002419
45	Sierra County	14	6	0.004321	0.001852
50	Sutter County	426	173	0.004695	0.001907
55	Ventura County	2275	2040	0.002871	0.002575
56	Yolo County	585	614	0.003031	0.003182
57	Yuba County	354	211	0.004993	0.002976

In [ ]:

```
print("Splitting the sample. Now into those with high drug arrests in 2009:
hi = merged_total.loc[merged_total["drug_offenses_2009"] > mean_drug_2009]
hi
```

```
Splitting the sample. Now into those with high drug arrests in 2009:
```

Out[ ]:

	COUNTY	VIOLENT_x	F_DRUGOFF_x	violent_arrest_rate_2009	drug_offenses_2009
0	Alameda County	4318	5749	0.002963	0.003946
	Fresno				

9	County	4377	3508	0.004914	0.003938
10	Glenn County	104	104	0.003729	0.003729
11	Humboldt County	503	747	0.003899	0.005791
12	Imperial County	599	730	0.003743	0.004562
14	Kern County	4290	3320	0.005493	0.004251
16	Lake County	349	271	0.005389	0.004185
18	Los Angeles County	35319	32193	0.003609	0.003290
22	Mendocino County	415	551	0.004824	0.006405
23	Merced County	1169	1055	0.004826	0.004355
31	Plumas County	100	104	0.004866	0.005061
33	Sacramento County	5302	4492	0.003854	0.003265
35	San Bernardino County	8474	8793	0.004266	0.004426
37	San Francisco County	3629	6867	0.004552	0.008613
43	Santa Cruz County	753	872	0.002995	0.003469
46	Siskiyou County	189	180	0.004256	0.004054
47	Solano County	1599	1422	0.003934	0.003498
48	Sonoma County	1359	1682	0.002928	0.003623
49	Stanislaus County	2211	2209	0.004377	0.004373
51	Tehama County	236	457	0.003894	0.007541
52	Trinity County	65	63	0.004669	0.004525
53	Tulare County	2183	1681	0.005244	0.004038

54	Tuolumne County	160	209	0.002869	0.003748

Exercise 16 continue

Now we can ask: did violent crime fall more from 2009 to 2018 in the counties that had lots of drug arrests in 2009 (where legalization likely had more of an effect) than in counties with fewer drug arrests in 2009 (where legalization likely mattered less)? Calculate this difference-in-differences:

(the change in violent crime rate for counties with lots of drug arrests in 2009) - (the change in violent crime rate for counties with few drug arrests in 2009)

In [ ]:
```
diff_violent_low = low["violent_arrest_rate_2009"] - low["violent_arrest_ra
diff_violent_low_mean = diff_violent_low.mean()
print(f"From 2009 to 2018 the average violent crime rate for counties with
```

```
From 2009 to 2018 the average violent crime rate for counties with few drug
arrests decreased by 0.000123878476924112, standardized over the total popu
lation.
```

In [ ]:
```
diff_violent_hi =  hi["violent_arrest_rate_2009"] - hi["violent_arrest_rate
diff_violent_hi_mean = diff_violent_hi.mean()
print(f"From 2009 to 2018 the average violent crime rate for counties with
```

```
From 2009 to 2018 the average violent crime rate for counties with many dru
g arrests decreased by 0.00017831241992704626, standardized over the total
population.
```

In [ ]:
```
diff_in_diff = diff_violent_hi_mean - diff_violent_low_mean
diff_in_diff
print(f"From 2009 to 2018 the average violent arrest rate declined by {diff
```

```
From 2009 to 2018 the average violent arrest rate declined by 5.44339430029
34256e-05 more in absolute terms in counties with many drug arrests compare
d to those with few arrests. I.e. the counties with more drug arrests in 20
09 were more impacted.
```

To make this number more accessible and understandable we will scale it by 100,000 to see the impact per 100,000 inhabitants:

In [ ]:
```
prop = diff_in_diff*100000
print(f"From 2009 to 2018 in absolute terms the violent arrest rate decline
```

```
From 2009 to 2018 in absolute terms the violent arrest rate declined by 5 c
ases more in counties with many drug arrests compared to those with few arr
ests. I.e. the counties with more drug arrests in 2009 were more impacted.
```

Exercise 17

Hmmm... we showed that there was a greater absolute decline in violent arrest rates in counties more impacted by drug legalization. But was there also a greater proportionate decline?

Calculate:

(the percentage change in violent crime rate for counties with lots of drug arrests in 2009) - (the percentage change in violent crime rate for counties with few drug arrests in 2009)

In [ ]:
```
low_perc = (diff_violent_low_mean/low["violent_arrest_rate_2009"].mean()) *
low_perc
print(f"On average violent arrests decreased by {low_perc.round(2)}% in cou
```

```
On average violent arrests decreased by 3.64% in counties with few drug arr
ests.
```

In [ ]:
```
hi_perc = (diff_violent_hi_mean/low["violent_arrest_rate_2009"].mean()) * 1
hi_perc
print(f"On average violent arrests decreased by {hi_perc.round(2)}% in cour
```

```
On average violent arrests decreased by 5.24% in counties with many drug ar
rests.
```

In [ ]:
```
prop = hi_perc/low_perc
print(f"Proportionally, from 2009 to 2018 the violent arrest rates declined
```

```
Proportionally, from 2009 to 2018 the violent arrest rates declined on aver
age 1.44 times more in counties with many drug arrests compared to those wi
th few arrests. I.e. the counties with more drug arrests in 2009 were more
impacted.
```