Live Coding - Module 9

Rachel Holman

In [1]:
```
import numpy as np
import pandas as pd
import os
os.chdir("/Users/rachelholman/Desktop/MSDS/DS6001 - Application of DS/Module 8:
```

In [2]:
```
cases = pd.read_csv('data100k.csv')
cases.head(3).T
```

Out[2]:

	0	1	2
person_id	102090000000110	343221000000125	343221000000125
HearingDate	2019-02-28	2009-12-07	2011-01-20
CodeSection	A.46.2-862	B.46.2-301	A.46.2-707
codesection	covered elsewhere	covered elsewhere	covered elsewhere
ChargeType	Misdemeanor	Misdemeanor	Misdemeanor
chargetype	Misdemeanor	Misdemeanor	Misdemeanor
Class	1	1	3
DispositionCode	Guilty	Guilty	Guilty
disposition	Conviction	Conviction	Conviction
Plea	NaN	NaN	NaN
Race	Black(Non-Hispanic)	Black(Non-Hispanic)	Black(Non-Hispanic)
Sex	Male	Female	Female
fips	25	540	540
convictions	True	True	True
arrests	False	False	True
felony10	False	False	False
sevenyear	False	True	True
tenyear	False	True	True
within7	True	False	False
within10	True	False	True
class1_2	False	False	False
class3_4	False	False	False
expungable	Automatic (pending)	Not eligible	Not eligible
old_expungable	False	False	False
expungable_no_lifetimelimit	Automatic (pending)	Not eligible	Not eligible
reason	Conviction of misdemeanor charges listed in 19...	Conviction of misdemeanor charges that are not...	Conviction of misdemeanor charges that are not...
sameday	False	False	False
lifetime	False	False	False

In [3]:
```
cases['Race'].value_counts()
```

Out[3]:
```
White Caucasian(Non-Hispanic)                    114421
Black(Non-Hispanic)                               80173
White Caucasian (Non-Hispanic)                    41679
Black (Non-Hispanic)                              33254
Hispanic                                           9319
White                                              3527
Other(Includes Not Applicable.. Unknown)           3452
Asian Or Pacific Islander                          2787
Black                                              2200
MISSING                                            1022
Unknown (Includes Not Applicable.. Unknown)         785
Other (Includes Not Applicable.. Unknown)           615
American Indian                                     302
Unknown                                              54
Asian or Pacific Islander                             7
American Indian Or Alaskan Native                     1
Name: Race, dtype: int64
```

In [4]:
```
cases['Race'].unique()
```

Out[4]:
```
array(['Black(Non-Hispanic)', 'Hispanic', 'White Caucasian(Non-Hispanic)',
       'MISSING', 'Asian Or Pacific Islander', 'Black (Non-Hispanic)',
       'White Caucasian (Non-Hispanic)',
       'Other(Includes Not Applicable.. Unknown)',
       'Other (Includes Not Applicable.. Unknown)', 'Black', 'White',
       'Unknown (Includes Not Applicable.. Unknown)', 'American Indian',
       'Unknown', 'Asian or Pacific Islander',
       'American Indian Or Alaskan Native'], dtype=object)
```

In [5]:
```
replace_map = {'Black(Non-Hispanic)':'Black (Non-Hispanic)',
               'Hispanic':'Hispanic',
               'White Caucasian(Non-Hispanic)':'White (Non-Hispanic)',
               'MISSING':'Missing/Other/Unknown',
               'Asian Or Pacific Islander':'Asian or Pacific Islander',
               'Black (Non-Hispanic)':'Black (Non-Hispanic)',
               'White Caucasian (Non-Hispanic)':'White (Non-Hispanic)',
               'Other(Includes Not Applicable.. Unknown)':'Missing/Other/Unknowr
               'Other (Includes Not Applicable.. Unknown)':'Missing/Other/Unknow
               'Black':'Black (Non-Hispanic)',
               'White':'White (Non-Hispanic)',
               'Unknown (Includes Not Applicable.. Unknown)':'Missing/Other/Unkr
               'American Indian':'American Indian or Alaskan Native',
               'Unknown':'Missing/Other/Unknown',
               'Asian or Pacific Islander':'Asian or Pacific Islander',
               'American Indian Or Alaskan Native':'American Indian or Alaskan N
cases['Race'] = cases['Race'].replace(replace_map)
cases['Race'].value_counts()
```

Out[5]:
```
White (Non-Hispanic)                159627
Black (Non-Hispanic)                115627
Hispanic                              9319
Missing/Other/Unknown                 5928
Asian or Pacific Islander             2794
American Indian or Alaskan Native      303
Name: Race, dtype: int64
```

In [6]:
```
cases['disposition'].value_counts()
```

Out[6]:
```
Conviction            188521
Dismissed             103867
Deferral Dismissal      1210
Name: disposition, dtype: int64
```

In [7]:
```
cases_convict = cases.query("disposition == 'Conviction'")
cases_convict.shape
```

Out[7]:
```
(188521, 28)
```

In [8]:
```
cases_convict.head(3).T
```

Out[8]:

	0	1	2
person_id	102090000000110	343221000000125	343221000000125
HearingDate	2019-02-28	2009-12-07	2011-01-20
CodeSection	A.46.2-862	B.46.2-301	A.46.2-707
codesection	covered elsewhere	covered elsewhere	covered elsewhere
ChargeType	Misdemeanor	Misdemeanor	Misdemeanor
chargetype	Misdemeanor	Misdemeanor	Misdemeanor
Class	1	1	3
DispositionCode	Guilty	Guilty	Guilty
disposition	Conviction	Conviction	Conviction
Plea	NaN	NaN	NaN
Race	Black (Non-Hispanic)	Black (Non-Hispanic)	Black (Non-Hispanic)
Sex	Male	Female	Female
fips	25	540	540
convictions	True	True	True
arrests	False	False	True
felony10	False	False	False
sevenyear	False	True	True
tenyear	False	True	True
within7	True	False	False
within10	True	False	True
class1_2	False	False	False
class3_4	False	False	False
expungable	Automatic (pending)	Not eligible	Not eligible
old_expungable	False	False	False
expungable_no_lifetimelimit	Automatic (pending)	Not eligible	Not eligible
reason	Conviction of misdemeanor charges listed in 19...	Conviction of misdemeanor charges that are not...	Conviction of misdemeanor charges that are not...
sameday	False	False	False
lifetime	False	False	False

In [9]:
```
cases_convict_race = cases_convict.groupby(['CodeSection','Race', 'fips']).size
cases_convict_race = cases_convict_race.rename({0:'count'},axis=1)
cases_convict_race
```

Out[9]:

	CodeSection	Race	fips	count
0	01-2007	White (Non-Hispanic)	51	1
1	1	Black (Non-Hispanic)	550	3
2	1	White (Non-Hispanic)	550	1
3	1-12	Black (Non-Hispanic)	650	27
4	1-12	White (Non-Hispanic)	650	6
...	...	...	...	...
27500	Z.18.2-91	White (Non-Hispanic)	840	2
27501	Z.18.2-91; 26	Black (Non-Hispanic)	700	1
27502	Z.18.2-95	Black (Non-Hispanic)	67	1
27503	Z.18.2-95	Black (Non-Hispanic)	83	1
27504	Z18.2-47	Black (Non-Hispanic)	730	1

27505 rows × 4 columns

In [10]:
```
cases_convict_race = pd.pivot_table(cases_convict_race,
                                    index=['CodeSection', 'fips'],
                                    columns=['Race'],
                                    values=['count'],
                                    fill_value=0).reset_index()
cases_convict_race
```

Out[10]:

	CodeSection	fips						
Race			American Indian or Alaskan Native	Asian or Pacific Islander	Black (Non-Hispanic)	Hispanic	Missing/Other/Unknown	His
0	01-2007	51	0	0	0	0	0	
1	1	550	0	0	3	0	0	
2	1-12	650	0	0	27	0	0	
3	1-200	29	0	0	1	0	0	
4	1-200	105	0	0	0	0	0	
...	...	...	...	...	...	...	...	
18700	Z.18.2-91	840	0	0	0	0	0	
18701	Z.18.2-91; 26	700	0	0	1	0	0	
18702	Z.18.2-95	67	0	0	1	0	0	
18703	Z.18.2-95	83	0	0	1	0	0	
18704	Z18.2-47	730	0	0	1	0	0	

18705 rows × 8 columns

In [11]:
```
cases_convict_race.columns = ['CodeSection','fips','aian','api','black','hisp',
cases_convict_race.columns
```

Out[11]:
```
Index(['CodeSection', 'fips', 'aian', 'api', 'black', 'hisp', 'unknown',
       'white'],
      dtype='object')
```

In [12]:
```
cases_convict_race['totalconvictions']= cases_convict_race['aian']+\
    cases_convict_race['api']+cases_convict_race['black']+cases_convict_race['h
    cases_convict_race['unknown']+cases_convict_race['white']
```

In [47]:
```
cases_convict_race = cases_convict_race.query("totalconvictions > 10")
cases_convict_race['aian_pct']=cases_convict_race['aian']/cases_convict_race['t
cases_convict_race['api_pct']=cases_convict_race['api']/cases_convict_race['tot
cases_convict_race['black_pct']=cases_convict_race['black']/cases_convict_race[
cases_convict_race['hisp_pct']=cases_convict_race['hisp']/cases_convict_race['t
cases_convict_race['unknown_pct']=cases_convict_race['unknown']/cases_convict_r
cases_convict_race['white_pct']=cases_convict_race['white']/cases_convict_race[
cases_convict_race
```

Out[47]:

	CodeSection	fips	aian	api	black	hisp	unknown	white	totalconvictions	aian_pct	
2	1-12	650	0	0	27	0	0	6	33	0.0	(
22	10-172	760	0	0	9	0	0	2	11	0.0	(
29	10-2	730	0	0	16	0	0	3	19	0.0	(
36	10-256	800	0	0	10	0	0	4	14	0.0	(
55	10-42	550	0	0	8	0	0	17	25	0.0	(
...	...	...	...	...	...	...	...	...	...	...	
18572	NO DMV	810	0	1	19	1	1	21	43	0.0	(
18582	NODMV	810	0	0	5	0	0	6	11	0.0	(
18692	Z.18.2-91	710	0	0	11	0	0	3	14	0.0	(
18696	Z.18.2-91	760	0	0	8	0	0	3	11	0.0	(
18698	Z.18.2-91	810	0	1	13	0	0	23	37	0.0	

2596 rows × 16 columns

In [48]:
```
cases_convict_race.sort_values(['black_pct'], ascending=False)
```

Out[48]:

	CodeSection	fips	aian	api	black	hisp	unknown	white	totalconvictions	aian_pct	a
5599	18.2-374.1:1	117	0	0	15	0	0	0	15	0.0	
4754	18.2-308.2	550	0	0	22	0	0	0	22	0.0	
11454	4.1-322	510	0	0	13	0	0	0	13	0.0	
8709	18.2-94	710	0	0	18	0	0	0	18	0.0	
4755	18.2-308.2	590	0	0	15	0	0	0	15	0.0	
...	...	...	...	...	...	...	...	...	...	...	
1938	18.2-172	77	0	0	0	0	0	18	18	0.0	
15368	A.18.2-266	27	0	0	0	0	0	17	17	0.0	
17255	B.46.2-301	167	0	0	0	0	0	64	64	0.0	
15380	A.18.2-266	51	0	0	0	0	0	14	14	0.0	
8932	18.2-96	173	0	0	0	0	0	12	12	0.0	

2596 rows × 16 columns

Census demographic population data

In [49]:
```
url_race = "https://demographics.coopercenter.org/sites/demographics/files/medi
url_hisp = "https://demographics.coopercenter.org/sites/demographics/files/medi
```

In [50]:
```
race_pop = pd.read_excel(url_race, skiprows=[0,1,2,3,5,6,7])
race_pop = race_pop[['FIPS', 'Jurisdiction', 'Total Population',
                     'White Alone', 'African American Alone', 'Asian Alone',
                     'Other Races Alone', 'Two or more races']]
race_pop = race_pop.rename({'FIPS':'fips',
                            'Jurisdiction':'locality',
                            'Total Population':'total_pop',
                            'White Alone':'white_pop',
                            'African American Alone':'black_pop',
                            'Asian Alone':'asian_pop',
                            'Other Races Alone':'other_pop',
                            'Two or more races':'two_pop'}, axis=1)
race_pop.head(10)
```

Out[50]:

	fips	locality	total_pop	white_pop	black_pop	asian_pop	other_pop	two_pop
0	1	Accomack County	32316	21899	9304	257	293	563
1	3	Albemarle County	109330	89388	10600	6051	483	2808
2	5	Alleghany County	14860	13783	698	46	56	277
3	7	Amelia County	13145	10050	2688	80	85	242
4	9	Amherst County	31605	24299	6041	180	305	780
5	11	Appomattox County	15911	12521	2951	61	56	322
6	13	Arlington County	236842	177639	22922	25945	1840	8496
7	15	Augusta County	75558	70102	3518	490	223	1225
8	17	Bath County	4147	3853	201	20	9	64
9	19	Bedford County	78997	70750	5697	975	299	1276

In [51]:
```
hisp_pop = pd.read_excel(url_hisp, skiprows=[0,1,2,3,5,6,7,8,9])
hisp_pop = hisp_pop[['FIPS', 'Unnamed: 6']]
hisp_pop = hisp_pop.rename({'FIPS':'fips',
                            'Unnamed: 6':'hisp_pop'}, axis=1)
hisp_pop
```

Out[51]:

	fips	hisp_pop
0	1	2955
1	3	6313
2	5	238
3	7	418
4	9	767
...	...	...
128	800	4300
129	810	38235
130	820	1966
131	830	1069
132	840	5127

133 rows × 2 columns

Merging in pandas

Step 1: a merge to test for problems

In [52]:
```
race_hisp_pop = pd.merge(race_pop, hisp_pop,
                         on= 'fips',
                         how= 'outer',
                         validate= 'one_to_one',
                         indicator= 'matched')
race_hisp_pop['matched'].value_counts()
```

Out[52]:
```
both          133
left_only       0
right_only      0
Name: matched, dtype: int64
```

Step 2: Once we know we are OK, do an inner merge

In [53]:
```
race_hisp_pop = pd.merge(race_pop, hisp_pop,
                         on= 'fips',
                         how= 'inner')
race_hisp_pop
```

Out[53]:

	fips	locality	total_pop	white_pop	black_pop	asian_pop	other_pop	two_pop	hisp
0	1	Accomack County	32316	21899	9304	257	293	563	
1	3	Albemarle County	109330	89388	10600	6051	483	2808	
2	5	Alleghany County	14860	13783	698	46	56	277	
3	7	Amelia County	13145	10050	2688	80	85	242	
4	9	Amherst County	31605	24299	6041	180	305	780	
...	...	...	...	...	...	...	...	...	
128	800	Suffolk city	92108	47993	39246	1754	475	2640	
129	810	Virginia Beach city	449974	303182	91161	33241	2907	19483	3
130	820	Waynesboro city	22630	18265	3040	382	149	794	
131	830	Williamsburg city	14954	11013	2391	904	104	542	
132	840	Winchester city	28078	22862	3174	764	279	999	

133 rows × 9 columns

In [54]:
```
cases_pop = pd.merge(cases_convict_race, race_hisp_pop,
                     on='fips',
                     how='outer',
                     validate= 'many_to_one',
                     indicator= 'matched')
cases_pop['matched'].value_counts()
```

Out[54]:
```
both          2596
right_only      13
left_only        0
Name: matched, dtype: int64
```

In [55]:
```
cases_pop.query("matched == 'left_only'")['fips'].unique()
```

Out[55]:
```
array([], dtype=int64)
```

In [56]:
```
replace_map={761:760,
             762:760,
             763:760,
             764:760,
             711:710,
             712:710,
             701:700,
             702:700,
             122:121}
cases_convict_race['fips']= cases_convict_race['fips'].replace(replace_map)
```

In [57]:
```
cases_pop.query("matched=='right_only'")['fips'].unique()
```

Out[57]:
```
array([ 17,  45,  91,  95, 115, 181, 580, 660, 678, 683, 685, 720, 735])
```

In [58]:
```
cases_pop = pd.merge(cases_convict_race, race_hisp_pop,
                     on='fips',
                     how='inner')
cases_pop.T
```

Out[58]:

	0	1	2	3	4	5	6	7
CodeSection	1-12	18.2-103	18.2-108	18.2-111	18.2-172	18.2-178	18.2-248	18.2 248.
fips	650	650	650	650	650	650	650	650
aian	0	0	0	0	0	0	0	0
api	0	1	0	0	0	2	0	0
black	27	91	9	10	10	17	32	15
hisp	0	0	0	0	0	0	0	0
unknown	0	0	1	0	0	0	0	0
white	6	36	2	10	4	9	2	0
totalconvictions	33	128	12	20	14	28	34	2
aian_pct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
api_pct	0.0	0.007812	0.0	0.0	0.0	0.071429	0.0	0.0
black_pct	0.818182	0.710938	0.75	0.5	0.714286	0.607143	0.941176	0.71428
hist_pct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
unknown_pct	0.0	0.0	0.083333	0.0	0.0	0.0	0.0	0.0
white_pct	0.181818	0.28125	0.166667	0.5	0.285714	0.321429	0.058824	0.28571
hisp_pct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
locality	Hampton city	Hampton city	Hampton city	Hampton city	Hampton city	Hampton city	Hampton city	Hampto cit
total_pop	134510	134510	134510	134510	134510	134510	134510	13451
white_pop	55478	55478	55478	55478	55478	55478	55478	5547
black_pop	68976	68976	68976	68976	68976	68976	68976	6897
asian_pop	3282	3282	3282	3282	3282	3282	3282	328
other_pop	1010	1010	1010	1010	1010	1010	1010	101
two_pop	5764	5764	5764	5764	5764	5764	5764	576
hisp_pop	8368	8368	8368	8368	8368	8368	8368	836

24 rows × 2596 columns

In [61]:
```
cases_pop['disparity_white'] = cases_pop['white_pct']/(cases_pop['white_pop']/c
cases_pop['disparity_black'] = cases_pop['black_pct']/(cases_pop['black_pop']/c
cases_pop['disparity_asian'] = cases_pop['api_pct']/(cases_pop['asian_pop']/cas
cases_pop['disparity_hisp'] = cases_pop['hisp_pct']/(cases_pop['hisp_pop']/case
```

In [65]:
```
results = cases_pop.sort_values('disparity_black', ascending=False).reset_index
results.loc[0,]
```

Out[65]:

index	2157
CodeSection	18.2-57
fips	35
aian	0
api	0
black	4
hisp	0
unknown	0
white	10
totalconvictions	14
aian_pct	0.0
api_pct	0.0
black_pct	0.285714
hist_pct	0.0
unknown_pct	0.0
white_pct	0.714286
hisp_pct	0.0
locality	Carroll County
total_pop	29791
white_pop	28964
black_pop	263
asian_pop	68
other_pop	113
two_pop	383
hisp_pop	1147
disparity_white	0.73468
disparity_black	32.363933
disparity_asian	0.0
disparity_hisp	0.0

Name: 0, dtype: object

In [69]:
```
cases_pop_nodrive = cases_pop[['46.2' not in str(x) for x in cases_pop['CodeSec
cases_pop_nodrive.sort_values('disparity_hisp', ascending=False).reset_index().
```

Out[69]:

index	2197
CodeSection	18.2-248
fips	5
aian	0
api	0
black	1
hisp	4
unknown	0
white	11
totalconvictions	16
aian_pct	0.0
api_pct	0.0
black_pct	0.0625
hist_pct	0.25
unknown_pct	0.0
white_pct	0.6875
hisp_pct	0.25
locality	Alleghany County
total_pop	14860
white_pop	13783
black_pop	698
asian_pop	46
other_pop	56
two_pop	277
hisp_pop	238
disparity_white	0.741221
disparity_black	1.330587
disparity_asian	0.0
disparity_hisp	15.609244

Name: 0, dtype: object

In [ ]: