Refining_Data

April 12, 2023

```
[1]: import numpy as np
     import pandas as pd
     import matplotlib.pyplot as plt
```

```
[ ]: datas = []
```

```
[8]: d0 = pd.read_csv("states0.csv")
     print(d0.shape[1])
     d0.head()
```

```
     11
```

```
[8]:    Unnamed: 0       State  TotalPop             Hispanic                White  \
     0           0     Alabama   4830620   3.7516156462584975%     61.878656462585%
     1           1      Alaska    733375    5.909580838323351%  60.910179640718574%
     2           2     Arizona   6641928   29.565921052631502%  57.120000000000026%
     3           3    Arkansas   2958208    6.215474452554738%    71.13781021897813%
     4           4  California  38421464   37.291874687968054%    40.21578881677474%

                    Black                Native                Asian  \
     0   31.25297619047618%   0.4532312925170065%   1.0502551020408146%
     1  2.8485029940119775%     16.39101796407186%    5.450299401197604%
     2  3.8509868421052658%      4.35506578947368%    2.876578947368419%
     3   18.968759124087573%   0.5229197080291965%   1.1423357664233578%
     4    5.677396405391911%  0.40529206190713685%   13.052234148776776%

                  Pacific              Income           GenderPop
     0  0.03435374149659865%  $43296.35860306644    2341093M_2489527F
     1   1.0586826347305378%  $70354.74390243902      384160M_349215F
     2  0.16763157894736833%  $54207.82095490716    3299088M_3342840F
     3  0.14686131386861315%  $41935.63396778917    1451913M_1506295F
     4  0.35141038442336353%  $67264.78230266465  19087135M_19334329F
```

```
[9]: d1 = pd.read_csv("states1.csv")
     print(d1.shape[1])
     d1.head()
```

```
     11
```

```
[9]:    Unnamed: 0                 State  TotalPop            Hispanic  \
     0           0              Colorado   5278906   20.78438003220608%
     1           1           Connecticut   3593222  15.604830917874388%
     2           2              Delaware    926454    8.82476635514019%
     3           3  District of Columbia    647484   9.165921787709499%
     4           4               Florida  19645772    21.3385426653884%

                      White                Black                Native  \
     0   69.89557165861504%   3.546376811594201%   0.5738325281803548%
     1    67.6770531400966%   10.34806763285027%  0.12620772946859898%
     2   64.63271028037383%  20.743925233644834%  0.25981308411214965%
     3  33.103910614525134%   51.77653631284915%  0.20055865921787713%
     4   59.08374880153398%  15.165675934803444%   0.2104506232023015%

                      Asian                Pacific              Income  \
     0   2.661996779388082%                    NaN  $64657.801787164906
     1   4.021980676328502%  0.018599033816425123%    $76146.5605875153
     2  3.2686915887850483%                    NaN   $61827.97663551402
     3  3.3832402234636865%  0.029608938547486034%   $75466.36363636363
     4  2.2831735378715257%   0.05151006711409391%  $50690.194986743794

                 GenderPop
     0   2648667M_2630239F
     1   1751607M_1841615F
     2     448413M_478041F
     3     306674M_340810F
     4  9600009M_10045763F
```

```
[10]: d2 = pd.read_csv("states2.csv")
      print(d2.shape[1])
      d2.head()
```

```
      11
```

```
[10]:    Unnamed: 0     State  TotalPop            Hispanic                White  \
     0           0   Georgia  10006693   8.418242207460397%   54.28630556974962%
     1           1    Hawaii   1406299   9.186708860759486%  25.032278481012657%
     2           2     Idaho   1616547  11.505369127516781%    83.1362416107383%
     3           3  Illinois  12873761  15.601733547351516%   60.85980738362764%
     4           4   Indiana   6568645   6.536744186046501%   78.43189368770771%

                      Black                Native                Asian  \
     0   32.08829841594277%  0.18758303525804798%   3.0976494634644895%
     1    2.052848101265823%   0.1449367088607596%    36.59208860759495%
     2  0.5667785234899323%    1.468120805369128%    1.135906040268457%
     3   17.108410914927717%  0.11842696629213499%     4.475377207062604%
     4    11.18697674418606%   0.1940863787375415%   1.5782724252491687%
```

```
                Pacific               Income          GenderPop
0  0.04660194174757282%  $50811.08205128205  4883331M_5123362F
1     8.75886075949367 2%  $73264.42628205128    709871M_696428F
2    0.1271812080536914%  $48017.31543624161    810464M_806083F
3   0.02003210272873195%  $59587.04887459807  6316899M_6556862F
4   0.03262458471760798%  $48616.22784810127  3235263M_3333382F
```

[11]:
```
d3 = pd.read_csv("states3.csv")
print(d3.shape[1])
d3.head()
```

```
11
```

[11]:
```
   Unnamed: 0      State  TotalPop             Hispanic                 White  \
0           0       Iowa   3093526    5.30364520048603%    87.71968408262464%
1           1     Kansas   2892987   11.644342105263148%    75.95828947368425%
2           2   Kentucky   4397353    3.222993688007212%     85.2307484220019%
3           3  Louisiana   4625253    4.866489361702128%   54.978546099290796%
4           4      Maine   1329100   1.4319088319088318%    93.70740740740736%

                Black                Native                Asian  \
0  3.2569866342648868%   0.2897934386391251%   1.699392466585662%
1     6.5678947368421%   0.7339473684210529%   2.331052631578946%
2   8.272317403065832%   0.1666366095581602%  1.1298467087466182%
3   36.32624113475175%  0.48430851063829816%   1.669060283687941%
4  1.1344729344729356%   0.7883190883190888%  0.9658119658119669%

                Pacific               Income          GenderPop
0  0.05516403402187123 5%   $53017.75304136253  1534595M_1558931F
1                   NaN  $53885.612648221344  1439862M_1453125F
2  0.046438232642019836%   $45285.80253623189  2164208M_2233145F
3  0.039184397163120555%   $44957.99376114082  2261156M_2364097F
4   0.01566951566951567%   $49181.97435897436    650081M_679019F
```

[12]:
```
d4 = pd.read_csv("states4.csv")
print(d4.shape[1])
d4.head()
```

```
11
```

[12]:
```
   Unnamed: 0          State  TotalPop             Hispanic  \
0           0       Maryland   5930538    8.47249820014399%
1           1  Massachusetts   6705586  11.461065573770476%
2           2       Michigan   9900571   4.634992732558134%
3           3      Minnesota   5419171   5.152923538230896%
4           4    Mississippi   2988081   2.842401215805473%
```

```
                White                Black                Native  \
0    52.679049676026%    30.6777537796976%  0.20309575233981278%
1  73.04105191256845%    6.83312841530056%  0.12827868852459007%
2  72.38172238372084%  17.633103197674423%  0.48441133720930313%
3  81.42706146926535%    5.65982008995502%    1.069040479760119%
4  53.28632218844981%  41.491945288753804%   0.3899696048632216%

                 Asian                Pacific              Income  \
0   5.325413966882652%   0.03628509719222463%  $78765.40072463769
1   5.835655737704914%    0.0198087431693989%  $72838.93672627235
2  2.4231104651162796%   0.01954941860465116%  $51201.83003663004
3   4.156071964017996%  0.032908545727136446%    $62820.833959429
4  0.8764437689969605%  0.015045592705167175%  $38909.91920731707

           GenderPop
0         2872643M_F
1  3249650M_3455936F
2  4861973M_5038598F
3  2692166M_2727005F
4  1451723M_1536358F
```

[13]:
```
d5 = pd.read_csv("states5.csv")
print(d5.shape[1])
d5.head()
```

```
11
```

[13]:
```
   Unnamed: 0          State  TotalPop             Hispanic  \
0           0       Missouri   6045448   4.037247838616718%
1           1        Montana   1014699  3.2688888888888896%
2           2       Nebraska   1869365   9.203759398496235%
3           3         Nevada   2798636  27.100883652430046%
4           4  New Hampshire   1324201  3.3219178082191796%

                White                Black                Native  \
0    77.508069164265%  14.122118155619594%  0.36332853025936646%
1  86.41555555555554%  0.4292592592592591%   7.0607407407407425%
2  81.13947368421056%   4.956203007518794%   0.8644736842105263%
3  53.23932253313698%   7.739617083946994%   1.0871870397643593%
4  91.31917808219184%  1.2277397260273974%  0.14280821917808229%

                 Asian               Pacific               Income  \
0  1.6244956772334296%   0.10165706051873193%   $49763.98772563177
1  0.5703703703703705%   0.07222222222222222%  $47645.682835820895
2  1.8590225563909788%   0.05714285714285715%  $55916.469696969696
3   7.095729013254786%    0.5745213549337267%  $55526.525073746314
```

```
4   2.191438356164382%  0.016095890410958904%    $68728.8595890411

           GenderPop
0  2964003M_3081445F
1          510163M_F
2    929606M_939759F
3  1407735M_1390901F
4    653484M_670717F
```

[14]:

```
d6 = pd.read_csv("states6.csv")
print(d6.shape[1])
d6.head()
```

```
11
```

[14]:

```
   Unnamed: 0           State  TotalPop             Hispanic  \
0           0      New Jersey   8904413   18.74950049950049%
1           1      New Mexico   2084117   45.28293172690762%
2           2        New York  19673174  17.241424747786684%
3           3  North Carolina   9845333   8.464762782128062%
4           4    North Dakota    721640   2.832682926829267%

                 White                Black                Native  \
0  56.488761238761285%  14.387862137862117%  0.11533466533466513%
1   40.69799196787147%  1.7550200803212852%    9.248594377510045%
2    56.4701050030883%  15.668046119003515%  0.32163887173152117%
3    64.5976508521419%    21.3951174573929%   1.0854905573468434%
4   87.44829268292683%  1.2843902439024397%    5.651219512195119%

                 Asian                Pacific               Income  \
0   8.159990009990018%  0.031318681318681325%   $76581.08341708542
1    1.23433734939759%   0.04277108433734938%   $47329.96787148595
2  7.8971587399629355%  0.023450689726168417%   $64290.74911292006
3   2.317457392906495%   0.05232611699677568%   $49937.46413697362
4  0.9619512195121945%                    NaN  $58188.112195121954

            GenderPop
0   4343027M_4561386F
1   1032414M_1051703F
2  9541801M_10131373F
3   4795408M_5049925F
4     367963M_353677F
```

[15]:

```
d7 = pd.read_csv("states7.csv")
print(d7.shape[1])
d7.head()
```

```
11
```

```
[15]:    Unnamed: 0         State  TotalPop             Hispanic  \
      0           0          Ohio  11575977  3.6720843250595037%
      1           1      Oklahoma   3849733    10.0799043062201%
      2           2        Oregon   3939233  11.441212121212132%
      3           3  Pennsylvania  12779559   6.128013741411624%
      4           4   Puerto Rico   3583073   98.89357384441935%

                        White                Black                  Native  \
      0    75.90306018361096%  16.207276436586163%    0.16888813328799712%
      1    66.05942583732046%   8.314736842105255%      6.716842105263157%
      2    78.39551515151517%   1.730787878787877%     1.0002424242424257%
      3    77.38385384134914%  11.633947532791995%    0.11926920674578385%
      4   0.7736189402480265%  0.0925591882750846%  0.0028184892897406984%

                         Asian                 Pacific              Income  \
      0    1.6210812648758952%   0.022645358721523304%   $49655.24846625767
      1    1.8011483253588516%     0.10622009569377985%   $48100.85426653883
      2     3.594909090909088%      0.3453333333333332%   $54271.90181818182
      3    2.7977514053716495%   0.019394128669581522%   $56170.46451005025
      4   0.07519729425028186%  0.0012401352874859078%  $20720.538285714287

                 GenderPop
      0  5662893M_5913084F
      1  1906944M_1942789F
      2  1948453M_1990780F
      3  6245344M_6534215F
      4  1713860M_1869213F
```

```
[16]: d8 = pd.read_csv("states8.csv")
      print(d8.shape[1])
      d8.head()
```

```
      11
```

```
[16]:    Unnamed: 0           State  TotalPop             Hispanic  \
      0           0    Rhode Island   1053661  13.356666666666678%
      1           1  South Carolina   4777576   5.056684981684991%
      2           2    South Dakota    843190  3.2396396396396376%
      3           3       Tennessee   6499615   4.720026972353339%
      4           4           Texas  26538614   38.04673809068304%

                        White                Black                Native  \
      0    74.32541666666665%    5.68291666666667%   0.3462500000000001%
      1   62.888736263736185%   28.75091575091577%   0.2923992673992673%
      2    82.50090090090092%  1.4238738738738752%    9.417567567567566%
      3    73.49008766014822%  18.283816587997297%  0.22663519892110592%
      4   44.687908934379145%   11.65004782858236%  0.26114405969007126%
```

```
                 Asian                Pacific               Income  \
0  3.2474999999999983%  0.035833333333333335%  $59125.270833333336
1   1.249175824175822%  0.046978021978021964%  $46296.807763401106
2  1.0193693693693688%   0.04189189189189189%   $51805.40540540541
3  1.4072825354012126%   0.04315576534052599%  $47328.083616587355
4  3.6696958102161825%   0.06881576430074614%  $55874.522600500095

             GenderPop
0      510388M_543273F
1    2322409M_2455167F
2      423477M_419713F
3    3167756M_3331859F
4  13171316M_13367298F
```

[17]:

```python
d9 = pd.read_csv("states9.csv")
print(d9.shape[1])
d9.head()
```

```
11
```

[17]:

```
   Unnamed: 0          State  TotalPop             Hispanic  \
0           0           Utah   2903379  13.468376068376063%
1           1        Vermont    626604  1.6092896174863391%
2           2       Virginia   8256630     8.0110164981373%
3           3     Washington   6985464   11.140968858131506%
4           4  West Virginia   1851420   1.290909090909089%

                 White                Black                Native  \
0   79.40683760683764%  1.0179487179487194%   1.0813675213675222%
1   93.98306010928961%  0.9808743169398909%    0.301639344262295%
2  63.271048430015945%   20.17599787120807%  0.21245343267695582%
3   72.03840830449816%   3.384429065743947%   1.4107266435986163%
4   92.17623966942146%  3.6628099173553723%  0.15268595041322316%

                 Asian               Pacific              Income  \
0   2.196068376068376%   0.8259829059829059%  $63488.91780821918
1  1.2387978142076501%  0.03060109289617486%  $55602.96721311475
2   5.455242150079845%  0.06471527408195847%  $72866.01341201717
3   7.022006920415224%    0.609896193771627%  $64493.76768377254
4  0.6824380165289253%  0.02644628099173554%  $41437.11157024794

           GenderPop
0  1459229M_1444150F
1    308573M_318031F
2  4060948M_4195682F
3  3487725M_3497739F
```

```
4    913631M_937789F
```

```
[22]: set(d0.columns) == set(d1.columns)
```

```
[22]: True
```

```
[23]: set(d1.columns) == set(d2.columns)
```

```
[23]: True
```

```
[24]: set(d2.columns) == set(d3.columns)
```

```
[24]: True
```

```
[25]: set(d4.columns) == set(d5.columns)
```

```
[25]: True
```

```
[26]: set(d5.columns) == set(d6.columns)
```

```
[26]: True
```

```
[27]: set(d6.columns) == set(d7.columns)
```

```
[27]: True
```

```
[28]: set(d8.columns) == set(d9.columns)
```

```
[28]: True
```

```
[29]: set(d7.columns) == set(d8.columns)
```

```
[29]: True
```

So They have same columns just not in order

```
[33]: df = pd.concat([d0, d1], axis=0)
```

```
[34]: df
```

```
[34]:    Unnamed: 0                State  TotalPop             Hispanic  \
      0           0              Alabama   4830620   3.7516156462584975%
      1           1               Alaska    733375    5.909580838323351%
      2           2              Arizona   6641928   29.565921052631502%
      3           3             Arkansas   2958208    6.215474452554738%
      4           4           California  38421464   37.291874687968054%
      5           5             Colorado   5278906    20.78438003220608%
      0           0             Colorado   5278906    20.78438003220608%
```

```
1           1           Connecticut   3593222  15.604830917874388%
2           2              Delaware    926454    8.82476635514019%
3           3  District of Columbia    647484   9.165921787709499%
4           4               Florida  19645772    21.3385426653884%
5           5               Georgia  10006693   8.418242207460397%

                 White                Black                Native  \
0     61.878656462585%   31.25297619047618%   0.4532312925170065%
1  60.910179640718574%  2.8485029940119775%    16.39101796407186%
2  57.120000000000026%  3.8509868421052658%     4.35506578947368%
3   71.13781021897813%  18.968759124087573%   0.5229197080291965%
4   40.21578881677474%   5.677396405391911%  0.40529206190713685%
5   69.89557165861504%   3.546376811594201%   0.5738325281803548%
0   69.89557165861504%   3.546376811594201%   0.5738325281803548%
1    67.6770531400966%   10.34806763285027%  0.12620772946859898%
2   64.63271028037383%  20.743925233644834%  0.25981308411214965%
3  33.103910614525134%   51.77653631284915%  0.20055865921787713%
4   59.08374880153398%  15.165675934803444%   0.2104506232023015%
5   54.28630556974962%   32.08829841594277%  0.18758303525804798%

                 Asian                Pacific               Income  \
0  1.0502551020408146%   0.03435374149659865%   $43296.35860306644
1   5.450299401197604%    1.0586826347305378%   $70354.74390243902
2   2.876578947368419%   0.16763157894736833%   $54207.82095490716
3  1.1423357664233578%   0.14686131386861315%   $41935.63396778917
4  13.052234148776776%   0.35141038442336353%   $67264.78230266465
5   2.661996779388082%                    NaN  $64657.801787164906
0   2.661996779388082%                    NaN  $64657.801787164906
1   4.021980676328502%  0.018599033816425123%    $76146.5605875153
2  3.2686915887850483%                    NaN   $61827.97663551402
3  3.3832402234636865%  0.029608938547486034%   $75466.36363636363
4  2.2831735378715257%   0.05151006711409391%  $50690.194986743794
5  3.0976494634644895%  0.046601941747572824%   $50811.08205128205

             GenderPop
0    2341093M_2489527F
1      384160M_349215F
2    3299088M_3342840F
3    1451913M_1506295F
4  19087135M_19334329F
5    2648667M_2630239F
0    2648667M_2630239F
1    1751607M_1841615F
2      448413M_478041F
3      306674M_340810F
4   9600009M_10045763F
5    4883331M_5123362F
```

[35]:
```
df = pd.concat([df, d1], axis=0)
```

[36]:
```
df = pd.concat([df, d2], axis=0)
```

[37]:
```
df = pd.concat([df, d3], axis=0)
```

[38]:
```
df = pd.concat([df, d4], axis=0)
```

[39]:
```
df = pd.concat([df, d5], axis=0)
```

[40]:
```
df = pd.concat([df, d6], axis=0)
```

[41]:
```
df = pd.concat([df, d7], axis=0)
```

[42]:
```
df = pd.concat([df, d8], axis=0)
```

[43]:
```
df = pd.concat([df, d9], axis=0)
```

[44]:
```
df
```

[44]:
```
    Unnamed: 0          State  TotalPop             Hispanic  \
0            0        Alabama   4830620  3.7516156462584975%
1            1         Alaska    733375   5.909580838323351%
2            2        Arizona   6641928  29.565921052631502%
3            3       Arkansas   2958208   6.215474452554738%
4            4     California  38421464  37.291874687968054%
..         …           …       …                  …
1            1        Vermont    626604  1.6092896174863391%
2            2       Virginia   8256630     8.0110164981373%
3            3     Washington   6985464  11.140968858131506%
4            4  West Virginia   1851420   1.290909090909089%
5            5      Wisconsin   5742117   6.683333333333334%

                  White                Black                Native  \
0      61.878656462585%    31.25297619047618%   0.4532312925170065%
1   60.910179640718574%  2.8485029940119775%    16.39101796407186%
2   57.120000000000026%  3.8509868421052658%     4.35506578947368%
3    71.13781021897813%  18.968759124087573%   0.5229197080291965%
4    40.21578881677474%   5.677396405391911%  0.40529206190713685%
..                  …                  …                   …
1    93.98306010928961%  0.9808743169398909%    0.301639344262295%
2   63.271048430015945%   20.17599787120807%  0.21245343267695582%
3    72.03840830449816%   3.384429065743947%   1.4107266435986163%
4    92.17623966942146%  3.6628099173553723%  0.15268595041322316%
5    79.86400862068966%   8.195186781609202%   0.9536637931034483%

                  Asian                Pacific                Income  \
0   1.0502551020408146%   0.03435374149659865%   $43296.35860306644
1    5.450299401197604%    1.0586826347305378%   $70354.74390243902
2    2.876578947368419%   0.16763157894736833%   $54207.82095490716
3   1.1423357664233578%   0.14686131386861315%   $41935.63396778917
4   13.052234148776776%   0.35141038442336353%   $67264.78230266465
..                   …                      …                    …
1   1.2387978142076501%   0.03060109289617486%   $55602.96721311475
2    5.455242150079845%   0.06471527408195847%   $72866.01341201717
3    7.022006920415224%     0.60989619377162 7%   $64493.76768377254
4   0.6824380165289253%   0.02644628099173554%   $41437.11157024794
5    2.404238505747124%  0.020833333333333332%  $53898.889208633096

              GenderPop
0     2341093M_2489527F
1       384160M_349215F
2     3299088M_3342840F
3     1451913M_1506295F
4   19087135M_19334329F
..                    …
1       308573M_318031F
2     4060948M_4195682F
3     3487725M_3497739F
4       913631M_937789F
5     2851385M_2890732F

[66 rows x 11 columns]
```

Every Dataset have 6 rows, and there were 10 files so it is right

[53]:
```
df = df.drop_duplicates(subset='State')
len(df)
```

[53]:
```
51
```

[54]:
```
df = df.drop(df.columns[0], axis=1)
df.head()
```

[54]:
```
        State  TotalPop             Hispanic                White  \
0     Alabama   4830620   3.7516156462584975%     61.878656462585%
1      Alaska    733375    5.909580838323351%   60.91017964071857 4%
2     Arizona   6641928   29.565921052631502%  57.120000000000026%
3    Arkansas   2958208    6.215474452554738%    71.13781021897813%
4  California  38421464   37.291874687968054%    40.21578881677474%

                Black               Native                Asian  \
0   31.25297619047618%   0.4532312925170065%  1.0502551020408146%
1  2.8485029940119775%    16.39101796407186%   5.450299401197604%
```

```
2  3.8509868421052658%     4.35506578947368%   2.876578947368419%
3  18.968759124087573%   0.5229197080291965%  1.1423357664233578%
4   5.677396405391911%  0.40529206190713685%  13.052234148776776%

                Pacific              Income            GenderPop
0  0.03435374149659865%  $43296.35860306644    2341093M_2489527F
1   1.0586826347305378%  $70354.74390243902      384160M_349215F
2  0.16763157894736833%  $54207.82095490716    3299088M_3342840F
3  0.14686131386861315%  $41935.63396778917    1451913M_1506295F
4  0.35141038442336353%  $67264.78230266465  19087135M_19334329F
```

```
[55]: df.dtypes
```

```
[55]: State        object
      TotalPop      int64
      Hispanic     object
      White        object
      Black        object
      Native       object
      Asian        object
      Pacific      object
      Income       object
      GenderPop    object
      dtype: object
```

```
[57]: df.isnull().sum()
```

```
[57]: State        0
      TotalPop     0
      Hispanic     0
      White        0
      Black        0
      Native       0
      Asian        0
      Pacific      4
      Income       0
      GenderPop    0
      dtype: int64
```

```
[58]: type(df['State'])
```

```
[58]: pandas.core.series.Series
```

```
[62]: df['Hispanic'] = df['Hispanic'].str.rstrip('%').astype(float)
```

```
[66]: df.head()
```

```
[66]:         State  TotalPop   Hispanic                White                Black  \
      0     Alabama   4830620   3.751616    61.878656462585%    31.25297619047618%
      1      Alaska    733375   5.909581  60.910179640718574%   2.8485029940119775%
      2     Arizona   6641928  29.565921  57.120000000000026%   3.8509868421052658%
      3    Arkansas   2958208   6.215474   71.13781021897813%   18.968759124087573%
      4  California  38421464  37.291875   40.21578881677474%    5.677396405391911%

                    Native                 Asian              Pacific  \
      0   0.4532312925170065%   1.0502551020408146%  0.03435374149659865%
      1     16.39101796407186%    5.450299401197604%   1.0586826347305378%
      2      4.35506578947368%    2.876578947368419%  0.16763157894736833%
      3    0.5229197080291965%   1.1423357664233578%  0.14686131386861315%
      4   0.40529206190713685%  13.052234148776776%   0.35141038442336353%

                   Income            GenderPop
      0  $43296.35860306644     2341093M_2489527F
      1  $70354.74390243902       384160M_349215F
      2  $54207.82095490716     3299088M_3342840F
      3  $41935.63396778917     1451913M_1506295F
      4  $67264.78230266465   19087135M_19334329F
```

```
[67]: df.dtypes
```

```
[67]: State          object
      TotalPop        int64
      Hispanic      float64
      White          object
      Black          object
      Native         object
      Asian          object
      Pacific        object
      Income         object
      GenderPop      object
      dtype: object
```

```
[71]: df['White'] = df['White'].str.rstrip('%').astype(float)
      df['Black'] = df['Black'].str.rstrip('%').astype(float)
      df['Asian'] = df['Asian'].str.rstrip('%').astype(float)
      df['Pacific'] = df['Pacific'].str.rstrip('%').astype(float)
      df['Native'] = df['Native'].str.rstrip('%').astype(float)
      df.head()
```

```
[71]:         State  TotalPop   Hispanic      White      Black     Native  \
      0     Alabama   4830620   3.751616  61.878656  31.252976   0.453231
      1      Alaska    733375   5.909581  60.910180   2.848503  16.391018
      2     Arizona   6641928  29.565921  57.120000   3.850987   4.355066
      3    Arkansas   2958208   6.215474  71.137810  18.968759   0.522920
```

```
4  California  38421464  37.291875  40.215789   5.677396   0.405292

       Asian   Pacific               Income            GenderPop
0   1.050255  0.034354  $43296.35860306644    2341093M_2489527F
1   5.450299  1.058683  $70354.74390243902      384160M_349215F
2   2.876579  0.167632  $54207.82095490716    3299088M_3342840F
3   1.142336  0.146861  $41935.63396778917    1451913M_1506295F
4  13.052234  0.351410  $67264.78230266465  19087135M_19334329F
```

[72]:
```
df.dtypes
```

[72]:
```
State         object
TotalPop       int64
Hispanic     float64
White        float64
Black        float64
Native       float64
Asian        float64
Pacific      float64
Income        object
GenderPop     object
dtype: object
```

[74]:
```
df[['males', 'females']] = df['GenderPop'].str.split('_').apply(pd.Series)
df['males'] = df['males'].str[:-1].astype(int)
df['females'] = df['females'].str[:-1].astype(int)
df.head()
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
/tmp/ipykernel_19814/3498029282.py in <module>
      1 df[['males', 'females']] = df['GenderPop'].str.split('_').apply(pd.
 ↪Series)
      2 df['males'] = df['males'].str[:-1].astype(int)
----> 3 df['females'] = df['females'].str[:-1].astype(int)
      4 df.head()

~/anaconda3/lib/python3.9/site-packages/pandas/core/generic.py in astype(self,␣
 ↪dtype, copy, errors)
   5910         else:
   5911             # else, only a single dtype is given
-> 5912             new_data = self._mgr.astype(dtype=dtype, copy=copy,␣
 ↪errors=errors)
   5913             return self._constructor(new_data).__finalize__(self,␣
 ↪method="astype")
   5914
```

```
~/anaconda3/lib/python3.9/site-packages/pandas/core/internals/managers.py in astype(self, dtype, copy, errors)
    417 
    418     def astype(self: T, dtype, copy: bool = False, errors: str = "raise") -> T:
--> 419         return self.apply("astype", dtype=dtype, copy=copy, errors=errors)
    420 
    421     def convert(

~/anaconda3/lib/python3.9/site-packages/pandas/core/internals/managers.py in apply(self, f, align_keys, ignore_failures, **kwargs)
    302                     applied = b.apply(f, **kwargs)
    303                 else:
--> 304                     applied = getattr(b, f)(**kwargs)
    305             except (TypeError, NotImplementedError):
    306                 if not ignore_failures:

~/anaconda3/lib/python3.9/site-packages/pandas/core/internals/blocks.py in astype(self, dtype, copy, errors)
    578         values = self.values
    579 
--> 580         new_values = astype_array_safe(values, dtype, copy=copy, errors=errors)
    581 
    582         new_values = maybe_coerce_values(new_values)

~/anaconda3/lib/python3.9/site-packages/pandas/core/dtypes/cast.py in astype_array_safe(values, dtype, copy, errors)
   1290 
   1291     try:
-> 1292         new_values = astype_array(values, dtype, copy=copy)
   1293     except (ValueError, TypeError):
   1294         # e.g. astype_nansafe can fail on object-dtype of strings

~/anaconda3/lib/python3.9/site-packages/pandas/core/dtypes/cast.py in astype_array(values, dtype, copy)
   1235 
   1236     else:
-> 1237         values = astype_nansafe(values, dtype, copy=copy)
   1238 
   1239     # in pandas we don't store numpy str dtypes, so convert to object

~/anaconda3/lib/python3.9/site-packages/pandas/core/dtypes/cast.py in astype_nansafe(arr, dtype, copy, skipna)
   1152         # work around NumPy brokenness, #1987
   1153         if np.issubdtype(dtype.type, np.integer):
-> 1154             return lib.astype_intsafe(arr, dtype)
```

```
   1155
   1156         # if we have a datetime/timedelta array of objects

~/anaconda3/lib/python3.9/site-packages/pandas/_libs/lib.pyx in pandas._libs.li.
 ↪astype_intsafe()

ValueError: invalid literal for int() with base 10: ''
```

[75]:
```
df.head()
```

[75]:
```
        State  TotalPop   Hispanic      White      Black     Native  \
0     Alabama   4830620   3.751616  61.878656  31.252976   0.453231
1      Alaska    733375   5.909581  60.910180   2.848503  16.391018
2     Arizona   6641928  29.565921  57.120000   3.850987   4.355066
3    Arkansas   2958208   6.215474  71.137810  18.968759   0.522920
4  California  38421464  37.291875  40.215789   5.677396   0.405292

       Asian   Pacific              Income            GenderPop     males  \
0   1.050255  0.034354  $43296.35860306644    2341093M_2489527F   2341093
1   5.450299  1.058683  $70354.74390243902      384160M_349215F    384160
2   2.876579  0.167632  $54207.82095490716    3299088M_3342840F   3299088
3   1.142336  0.146861  $41935.63396778917    1451913M_1506295F   1451913
4  13.052234  0.351410  $67264.78230266465  19087135M_19334329F  19087135

     females
0   2489527F
1    349215F
2   3342840F
3   1506295F
4  19334329F
```

[76]:
```
df.dtypes
```

[76]:
```
State         object
TotalPop       int64
Hispanic     float64
White        float64
Black        float64
Native       float64
Asian        float64
Pacific      float64
Income        object
GenderPop     object
males          int64
females       object
dtype: object
```

[82]:
```
df['females'] = df['females'].str.extract(r'(\d+)F', expand=False).astype(float)
```

[80]:
```
df
```

[80]:
```
                  State  TotalPop   Hispanic      White      Black     Native  \
0               Alabama   4830620   3.751616  61.878656  31.252976   0.453231
1                Alaska    733375   5.909581  60.910180   2.848503  16.391018
2               Arizona   6641928  29.565921  57.120000   3.850987   4.355066
3              Arkansas   2958208   6.215474  71.137810  18.968759   0.522920
4            California  38421464  37.291875  40.215789   5.677396   0.405292
5              Colorado   5278906  20.784380  69.895572   3.546377   0.573833
1           Connecticut   3593222  15.604831  67.677053  10.348068   0.126208
2              Delaware    926454   8.824766  64.632710  20.743925   0.259813
3  District of Columbia    647484   9.165922  33.103911  51.776536   0.200559
4               Florida  19645772  21.338543  59.083749  15.165676   0.210451
5               Georgia  10006693   8.418242  54.286306  32.088298   0.187583
1                Hawaii   1406299   9.186709  25.032278   2.052848   0.144937
2                 Idaho   1616547  11.505369  83.136242   0.566779   1.468121
3              Illinois  12873761  15.601734  60.859807  17.108411   0.118427
4               Indiana   6568645   6.536744  78.431894  11.186977   0.194086
5                  Iowa   3093526   5.303645  87.719684   3.256987   0.289793
1                Kansas   2892987  11.644342  75.958289   6.567895   0.733947
2              Kentucky   4397353   3.222994  85.230748   8.272317   0.166637
3             Louisiana   4625253   4.866489  54.978546  36.326241   0.484309
4                 Maine   1329100   1.431909  93.707407   1.134473   0.788319
5              Maryland   5930538   8.472498  52.679050  30.677754   0.203096
1         Massachusetts   6705586  11.461066  73.041052   6.833128   0.128279
2              Michigan   9900571   4.634993  72.381722  17.633103   0.484411
3             Minnesota   5419171   5.152924  81.427061   5.659820   1.069040
4           Mississippi   2988081   2.842401  53.286322  41.491945   0.389970
5              Missouri   6045448   4.037248  77.508069  14.122118   0.363329
1               Montana   1014699   3.268889  86.415556   0.429259   7.060741
2              Nebraska   1869365   9.203759  81.139474   4.956203   0.864474
3                Nevada   2798636  27.100884  53.239323   7.739617   1.087187
4         New Hampshire   1324201   3.321918  91.319178   1.227740   0.142808
5            New Jersey   8904413  18.749500  56.488761  14.387862   0.115335
1            New Mexico   2084117  45.282932  40.697992   1.755020   9.248594
2              New York  19673174  17.241425  56.470105  15.668046   0.321639
3        North Carolina   9845333   8.464763  64.597651  21.395117   1.085491
4          North Dakota    721640   2.832683  87.448293   1.284390   5.651220
5                  Ohio  11575977   3.672084  75.903060  16.207276   0.168888
1              Oklahoma   3849733  10.079904  66.059426   8.314737   6.716842
2                Oregon   3939233  11.441212  78.395515   1.730788   1.000242
3          Pennsylvania  12779559   6.128014  77.383854  11.633948   0.119269
4           Puerto Rico   3583073  98.893574   0.773619   0.092559   0.002818
5          Rhode Island   1053661  13.356667  74.325417   5.682917   0.346250
1        South Carolina   4777576   5.056685  62.888736  28.750916   0.292399
```

```
2         South Dakota    843190   3.239640  82.500901   1.423874   9.417568
3            Tennessee   6499615   4.720027  73.490088  18.283817   0.226635
4                Texas  26538614  38.046738  44.687909  11.650048   0.261144
5                 Utah   2903379  13.468376  79.406838   1.017949   1.081368
1              Vermont    626604   1.609290  93.983060   0.980874   0.301639
2             Virginia   8256630   8.011016  63.271048  20.175998   0.212453
3           Washington   6985464  11.140969  72.038408   3.384429   1.410727
4        West Virginia   1851420   1.290909  92.176240   3.662810   0.152686
5            Wisconsin   5742117   6.683333  79.864009   8.195187   0.953664

       Asian   Pacific               Income            GenderPop     males  \
0   1.050255  0.034354   $43296.35860306644    2341093M_2489527F   2341093
1   5.450299  1.058683   $70354.74390243902      384160M_349215F    384160
2   2.876579  0.167632   $54207.82095490716    3299088M_3342840F   3299088
3   1.142336  0.146861   $41935.63396778917    1451913M_1506295F   1451913
4  13.052234  0.351410   $67264.78230266465  19087135M_19334329F  19087135
5   2.661997       NaN  $64657.801787164906    2648667M_2630239F   2648667
1   4.021981  0.018599    $76146.5605875153    1751607M_1841615F   1751607
2   3.268692       NaN   $61827.97663551402      448413M_478041F    448413
3   3.383240  0.029609   $75466.36363636363      306674M_340810F    306674
4   2.283174  0.051510  $50690.194986743794   9600009M_10045763F   9600009
5   3.097649  0.046602   $50811.08205128205    4883331M_5123362F   4883331
1  36.592089  8.758861   $73264.42628205128      709871M_696428F    709871
2   1.135906  0.127181   $48017.31543624161      810464M_806083F    810464
3   4.475377  0.020032   $59587.04887459807    6316899M_6556862F   6316899
4   1.578272  0.032625   $48616.22784810127    3235263M_3333382F   3235263
5   1.699392  0.055164   $53017.75304136253    1534595M_1558931F   1534595
1   2.331053       NaN  $53885.612648221344    1439862M_1453125F   1439862
2   1.129847  0.046438   $45285.80253623189    2164208M_2233145F   2164208
3   1.669060  0.039184   $44957.99376114082    2261156M_2364097F   2261156
4   0.965812  0.015670   $49181.97435897436      650081M_679019F    650081
5   5.325414  0.036285   $78765.40072463769           2872643M_F   2872643
1   5.835656  0.019809   $72838.93672627235    3249650M_3455936F   3249650
2   2.423110  0.019549   $51201.83003663004    4861973M_5038598F   4861973
3   4.156072  0.032909     $62820.833959429    2692166M_2727005F   2692166
4   0.876444  0.015046   $38909.91920731707    1451723M_1536358F   1451723
5   1.624496  0.101657   $49763.98772563177    2964003M_3081445F   2964003
1   0.570370  0.072222  $47645.682835820895            510163M_F    510163
2   1.859023  0.057143  $55916.469696969696      929606M_939759F    929606
3   7.095729  0.574521  $55526.525073746314    1407735M_1390901F   1407735
4   2.191438  0.016096    $68728.8595890411      653484M_670717F    653484
5   8.159990  0.031319   $76581.08341708542    4343027M_4561386F   4343027
1   1.234337  0.042771   $47329.96787148595    1032414M_1051703F   1032414
2   7.897159  0.023451   $64290.74911292006   9541801M_10131373F   9541801
3   2.317457  0.052326   $49937.46413697362    4795408M_5049925F   4795408
4   0.961951       NaN  $58188.112195121954      367963M_353677F    367963
5   1.621081  0.022645   $49655.24846625767    5662893M_5913084F   5662893
```

```
1   1.801148  0.106220   $48100.85426653883    1906944M_1942789F   1906944
2   3.594909  0.345333   $54271.90181818182    1948453M_1990780F   1948453
3   2.797751  0.019394   $56170.46451005025    6245344M_6534215F   6245344
4   0.075197  0.001240  $20720.538285714287    1713860M_1869213F   1713860
5   3.247500  0.035833  $59125.270833333336     510388M_543273F     510388
1   1.249176  0.046978  $46296.807763401106    2322409M_2455167F   2322409
2   1.019369  0.041892   $51805.40540540541     423477M_419713F     423477
3   1.407283  0.043156  $47328.083616587355    3167756M_3331859F   3167756
4   3.669696  0.068816  $55874.522600500095  13171316M_13367298F  13171316
5   2.196068  0.825983   $63488.91780821918    1459229M_1444150F   1459229
1   1.238798  0.030601   $55602.96721311475     308573M_318031F     308573
2   5.455242  0.064715   $72866.01341201717    4060948M_4195682F   4060948
3   7.022007  0.609896   $64493.76768377254    3487725M_3497739F   3487725
4   0.682438  0.026446   $41437.11157024794     913631M_937789F     913631
5   2.404239  0.020833  $53898.889208633096    2851385M_2890732F   2851385
```

```
      females
0   2489527.0
1    349215.0
2   3342840.0
3   1506295.0
4  19334329.0
5   2630239.0
1   1841615.0
2    478041.0
3    340810.0
4  10045763.0
5   5123362.0
1    696428.0
2    806083.0
3   6556862.0
4   3333382.0
5   1558931.0
1   1453125.0
2   2233145.0
3   2364097.0
4    679019.0
5         NaN
1   3455936.0
2   5038598.0
3   2727005.0
4   1536358.0
5   3081445.0
1         NaN
2    939759.0
3   1390901.0
4    670717.0
```

```
5   4561386.0
1   1051703.0
2  10131373.0
3   5049925.0
4    353677.0
5   5913084.0
1   1942789.0
2   1990780.0
3   6534215.0
4   1869213.0
5    543273.0
1   2455167.0
2    419713.0
3   3331859.0
4  13367298.0
5   1444150.0
1    318031.0
2   4195682.0
3   3497739.0
4    937789.0
5   2890732.0
```

[84]:
```
df['females'] = df['females'].fillna(0).astype(int)
```

[85]:
```
df.head()
```

[85]:
```
        State  TotalPop   Hispanic      White      Black     Native  \
0     Alabama   4830620   3.751616  61.878656  31.252976   0.453231
1      Alaska    733375   5.909581  60.910180   2.848503  16.391018
2     Arizona   6641928  29.565921  57.120000   3.850987   4.355066
3    Arkansas   2958208   6.215474  71.137810  18.968759   0.522920
4  California  38421464  37.291875  40.215789   5.677396   0.405292

       Asian   Pacific              Income            GenderPop     males  \
0   1.050255  0.034354  $43296.35860306644    2341093M_2489527F   2341093
1   5.450299  1.058683  $70354.74390243902      384160M_349215F    384160
2   2.876579  0.167632  $54207.82095490716    3299088M_3342840F   3299088
3   1.142336  0.146861  $41935.63396778917    1451913M_1506295F   1451913
4  13.052234  0.351410  $67264.78230266465  19087135M_19334329F  19087135

    females
0   2489527
1    349215
2   3342840
3   1506295
4  19334329
```

```
[86]: df = df.drop('GenderPop', axis=1)
```

```
[87]: df.head()
```

```
[87]:        State  TotalPop   Hispanic      White      Black     Native  \
      0    Alabama   4830620   3.751616  61.878656  31.252976   0.453231
      1     Alaska    733375   5.909581  60.910180   2.848503  16.391018
      2    Arizona   6641928  29.565921  57.120000   3.850987   4.355066
      3   Arkansas   2958208   6.215474  71.137810  18.968759   0.522920
      4 California  38421464  37.291875  40.215789   5.677396   0.405292

            Asian   Pacific               Income     males   females
      0  1.050255  0.034354  $43296.35860306644   2341093   2489527
      1  5.450299  1.058683  $70354.74390243902    384160    349215
      2  2.876579  0.167632  $54207.82095490716   3299088   3342840
      3  1.142336  0.146861  $41935.63396778917   1451913   1506295
      4 13.052234  0.351410  $67264.78230266465  19087135  19334329
```

```
[90]: df.dtypes
```

```
[90]: State         object
      TotalPop       int64
      Hispanic     float64
      White        float64
      Black        float64
      Native       float64
      Asian        float64
      Pacific      float64
      Income        object
      males          int64
      females        int64
      dtype: object
```

```
[91]: df['Income'] = df['Income'].str.replace("$",'').astype(float)
      df.head()
```

```
/tmp/ipykernel_19814/392498448.py:1: FutureWarning: The default value of regex
will change from True to False in a future version. In addition, single
character regular expressions will *not* be treated as literal strings when
regex=True.
  df['Income'] = df['Income'].str.replace("$",'').astype(float)
```

```
[91]:        State  TotalPop   Hispanic      White      Black     Native  \
      0    Alabama   4830620   3.751616  61.878656  31.252976   0.453231
      1     Alaska    733375   5.909581  60.910180   2.848503  16.391018
      2    Arizona   6641928  29.565921  57.120000   3.850987   4.355066
      3   Arkansas   2958208   6.215474  71.137810  18.968759   0.522920
```

```
4  California  38421464  37.291875  40.215789   5.677396   0.405292

       Asian   Pacific        Income     males   females
0   1.050255  0.034354  43296.358603   2341093   2489527
1   5.450299  1.058683  70354.743902    384160    349215
2   2.876579  0.167632  54207.820955   3299088   3342840
3   1.142336  0.146861  41935.633968   1451913   1506295
4  13.052234  0.351410  67264.782303  19087135  19334329
```

[92]:

```
df.dtypes
```

[92]:

```
State        object
TotalPop      int64
Hispanic    float64
White       float64
Black       float64
Native      float64
Asian       float64
Pacific     float64
Income      float64
males         int64
females       int64
dtype: object
```

[93]:

```
df['State'] = df['State'].astype(str)
df.dtypes
```

[93]:

```
State        object
TotalPop      int64
Hispanic    float64
White       float64
Black       float64
Native      float64
Asian       float64
Pacific     float64
Income      float64
males         int64
females       int64
dtype: object
```

[94]:

```
df.head()
```

[94]:

```
        State  TotalPop   Hispanic      White      Black     Native  \
0     Alabama   4830620   3.751616  61.878656  31.252976   0.453231
1      Alaska    733375   5.909581  60.910180   2.848503  16.391018
2     Arizona   6641928  29.565921  57.120000   3.850987   4.355066
3    Arkansas   2958208   6.215474  71.137810  18.968759   0.522920
```

```
4  California  38421464  37.291875  40.215789   5.677396   0.405292

       Asian   Pacific        Income     males   females
0   1.050255  0.034354  43296.358603   2341093   2489527
1   5.450299  1.058683  70354.743902    384160    349215
2   2.876579  0.167632  54207.820955   3299088   3342840
3   1.142336  0.146861  41935.633968   1451913   1506295
4  13.052234  0.351410  67264.782303  19087135  19334329
```

[ ]: