```
In [224]: import pandas as pd
           import numpy as np
           import matplotlib.pyplot as plt
           import seaborn as sns
           from matplotlib import style
In [225]: df0 = pd.read csv('states0.csv')
           df1 = pd.read_csv('states1.csv')
           df2 = pd.read csv('states2.csv')
           df3 = pd.read_csv('states3.csv')
           df4 = pd.read_csv('states4.csv')
           df5 = pd.read csv('states5.csv')
           df6 = pd.read csv('states6.csv')
           df7 = pd.read_csv('states7.csv')
           df8 = pd.read csv('states8.csv')
           df9 = pd.read csv('states9.csv')
In [226]: df0
Out[226]:
               Unnamed:
                            State
                                  TotalPop
                                                      Hispanic
                                                                             White
            0
                         Alabama
                                   4830620
                                           3.7516156462584975%
                                                                  61.878656462585%
                                                                                    31.25297619047
            1
                      1
                                            5.909580838323351% 60.910179640718574%
                           Alaska
                                   733375
                                                                                   2.8485029940119
                                                               57.120000000000026%
                          Arizona
                                   6641928
                                           29.565921052631502%
                                                                                   3.8509868421052
                         Arkansas
                                   2958208
                                            6.215474452554738%
                                                                71.13781021897813%
                                                                                   18.968759124087
                         California
                                  38421464
                                           37.291874687968054%
                                                                40.21578881677474%
                                                                                    5.677396405391
                         Colorado
                                   5278906
                                            20.78438003220608%
                                                                69.89557165861504%
                                                                                    3.546376811594
```

```
In [227]: combined_df = pd.concat([df1,df2,df3,df4,df5,df6,df7,df8,df9] , ignore_index=T
```

```
In [228]: combined_df.to_csv('combined_df' , index = False)
```

In [229]: df = pd.read_csv('combined_df')
df

Out[229]:

	Unnamed: 0	State	TotalPop	Hispanic	White	
0	0	Colorado	5278906	20.78438003220608%	69.89557165861504%	3.546
1	1	Connecticut	3593222	15.604830917874388%	67.6770531400966%	10.34{
2	2	Delaware	926454	8.82476635514019%	64.63271028037383%	20.7439
3	3	District of Columbia	647484	9.165921787709499%	33.103910614525134%	51.776
4	4	Florida	19645772	21.3385426653884%	59.08374880153398%	15.1656
5	5	Georgia	10006693	8.418242207460397%	54.28630556974962%	32.088
6	0	Georgia	10006693	8.418242207460397%	54.28630556974962%	32.088
7	1	Hawaii	1406299	9.186708860759486%	25.032278481012657%	2.0528
8	2	Idaho	1616547	11.505369127516781%	83.1362416107383%	0.56677

In [230]: df = df.iloc[:,1:]
df

)0	1.4319088319088318%	93.70740740740736%	1.1344729344729356%	0.7883190883190888%	0.9658119658
38	8.47249820014399%	52.679049676026%	30.6777537796976%	0.20309575233981278%	5.3254139668
38	8.47249820014399%	52.679049676026%	30.6777537796976%	0.20309575233981278%	5.3254139668
36	11.461065573770476%	73.04105191256845%	6.83312841530056%	0.12827868852459007%	5.8356557377
′1	4.634992732558134%	72.38172238372084%	17.633103197674423%	0.48441133720930313%	2.42311046511
′1	5.152923538230896%	81.42706146926535%	5.65982008995502%	1.069040479760119%	4.1560719640
31	2.842401215805473%	53.28632218844981%	41.491945288753804%	0.3899696048632216%	0.87644376899
18	4.037247838616718%	77.508069164265%	14.122118155619594%	0.36332853025936646%	1.62449567723
18	4.037247838616718%	77.508069164265%	14.122118155619594%	0.36332853025936646%	1.62449567723
)9	3.2688888888888896%	86.4155555555554%	0.4292592592592591%	7.0607407407407425%	0.57037037037
35	9.203759398496235%	81.13947368421056%	4.956203007518794%	0.8644736842105263%	1.85902255639
36	27.100883652430046%	53.23932253313698%	7.739617083946994%	1.0871870397643593%	7.0957290132 🔻
4					

```
In [231]: df.info()
            <class 'pandas.core.frame.DataFrame'>
            RangeIndex: 54 entries, 0 to 53
            Data columns (total 10 columns):
             #
                  Column
                               Non-Null Count
                                                   Dtype
                                                   _ _ _ _ _
             0
                               54 non-null
                                                   object
                  State
                  TotalPop
                               54 non-null
                                                   int64
             1
             2
                  Hispanic
                               54 non-null
                                                   object
             3
                               54 non-null
                                                   object
                  White
             4
                               54 non-null
                                                   object
                  Black
             5
                               54 non-null
                                                   object
                  Native
             6
                  Asian
                               54 non-null
                                                   object
             7
                  Pacific
                               50 non-null
                                                   object
             8
                  Income
                               54 non-null
                                                   object
             9
                  GenderPop 54 non-null
                                                   object
            dtypes: int64(1), object(9)
            memory usage: 4.3+ KB
In [232]: df.isnull()
Out[232]:
                       TotalPop Hispanic White
                                                 Black Native Asian Pacific Income GenderPop
                 State
              0 False
                           False
                                     False
                                           False
                                                  False
                                                          False
                                                                 False
                                                                          True
                                                                                 False
                                                                                             False
              1 False
                           False
                                     False
                                           False
                                                  False
                                                          False
                                                                 False
                                                                         False
                                                                                 False
                                                                                              False
                           False
                                                          False
              2 False
                                     False
                                           False
                                                  False
                                                                 False
                                                                          True
                                                                                 False
                                                                                              False
                False
                                           False
              3
                           False
                                     False
                                                  False
                                                          False
                                                                 False
                                                                         False
                                                                                 False
                                                                                              False
                 False
                           False
                                     False
                                           False
                                                  False
                                                          False
                                                                 False
                                                                         False
                                                                                 False
                                                                                              False
                                           False
                 False
                           False
                                     False
                                                  False
                                                          False
                                                                 False
                                                                         False
                                                                                 False
                                                                                              False
                 False
                           False
                                     False
                                           False
                                                  False
                                                          False
                                                                 False
                                                                         False
                                                                                 False
                                                                                              False
              7 False
                                           False
                                                          False
                                                                         False
                                                                                 False
                                                                                             False
                           False
                                     False
                                                  False
                                                                 False
                 False
                                           False
                                                                         False
                                                                                 False
                                                                                              False
                           False
                                     False
                                                  False
                                                          False
                                                                 False
                 False
                                           False
                                                                                 False
                                                                                              False
                           False
                                     False
                                                  False
                                                          False
                                                                 False
                                                                         False
                False
                           False
                                     False
                                           False
                                                                                              False
                                                  False
                                                          False
                                                                 False
                                                                         False
                                                                                 False
In [233]: df = df.dropna()
In [234]: | df = df.drop_duplicates()
```

```
In [235]: df.info()
          <class 'pandas.core.frame.DataFrame'>
          Int64Index: 42 entries, 1 to 53
          Data columns (total 10 columns):
               Column
           #
                           Non-Null Count
                                           Dtype
                           _____
                                           ----
                           42 non-null
                                           object
           0
               State
                          42 non-null
                                           int64
           1
               TotalPop
           2
               Hispanic
                          42 non-null
                                           object
           3
                           42 non-null
               White
                                           object
           4
                          42 non-null
               Black
                                           object
           5
                          42 non-null
               Native
                                           object
           6
                          42 non-null
                                           object
               Asian
           7
               Pacific
                          42 non-null
                                           object
           8
               Income
                          42 non-null
                                           object
           9
               GenderPop 42 non-null
                                           object
          dtypes: int64(1), object(9)
          memory usage: 3.6+ KB
In [236]: | df = df.reset_index(drop =True)
In [237]: | df[['Male', 'Female']] = df['GenderPop'].str.split('_',expand = True)
In [238]: df['Male'] = df['Male'].str.replace('M','')
In [239]: |df['Female'] = df['Female'].str.replace('F','')
In [240]: df['Male'] = df['Male'].astype(int)
In [241]: df["Female"][12] = 2900655
          C:\Users\mosai\AppData\Local\Temp\ipykernel_7092\2094601370.py:1: SettingWith
          CopyWarning:
          A value is trying to be set on a copy of a slice from a DataFrame
          See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/s
          table/user guide/indexing.html#returning-a-view-versus-a-copy (https://panda
          s.pydata.org/pandas-docs/stable/user guide/indexing.html#returning-a-view-ver
          sus-a-copy)
            df["Female"][12] = 2900655
```

```
miniproject - Jupyter Notebook
In [242]: |df["Female"][14] = 2900655
          C:\Users\mosai\AppData\Local\Temp\ipykernel 7092\3515725951.py:1: SettingWith
          CopyWarning:
          A value is trying to be set on a copy of a slice from a DataFrame
          See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/s
          table/user guide/indexing.html#returning-a-view-versus-a-copy (https://panda
          s.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-ver
          sus-a-copy)
             df["Female"][14] = 2900655
In [243]: df["Female"][18] = 2900655
          C:\Users\mosai\AppData\Local\Temp\ipykernel 7092\3063036996.py:1: SettingWith
          CopyWarning:
          A value is trying to be set on a copy of a slice from a DataFrame
          See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/s
          table/user_guide/indexing.html#returning-a-view-versus-a-copy (https://panda
          s.pydata.org/pandas-docs/stable/user guide/indexing.html#returning-a-view-ver
          sus-a-copy)
             df["Female"][18] = 2900655
In [244]: | df['Female']
Out[244]: 0
                  1841615
                   340810
           1
          2
                 10045763
           3
                  5123362
          4
                   696428
          5
                   806083
          6
                  6556862
          7
                  3333382
          8
                  1558931
          9
                  2233145
          10
                  2364097
          11
                   679019
          12
                  2900655
          13
                  3455936
          14
                  2900655
          15
                  2727005
          16
                  1536358
          17
                  3081445
          18
                  2900655
In [245]: |df['Female'] = df['Female'].astype(int)
In [246]: | df['Female'].mean()
```

localhost:8888/notebooks/Week-4/Mini project/US Census Data Cleaning/miniproject.ipynb

Out[246]: 3107844.8571428573

In [247]: df.head()

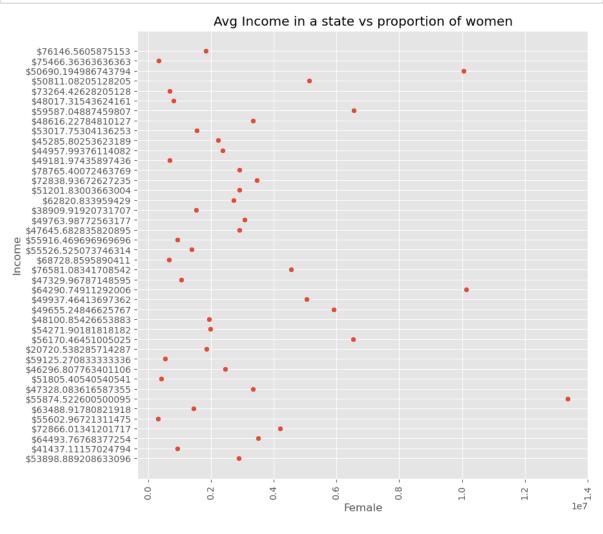
Out[247]:									
		State	TotalPop	Hispanic	White	Black			
	0	Connecticut	3593222	15.604830917874388%	67.6770531400966%	10.34806763285027%	0.1		
	1	District of Columbia	647484	9.165921787709499%	33.103910614525134%	51.77653631284915%	0.2		
	2	Florida	19645772	21.3385426653884%	59.08374880153398%	15.165675934803444%	0.		
	3	Georgia	10006693	8.418242207460397%	54.28630556974962%	32.08829841594277%	0.1		
	4	Hawaii	1406299	9.186708860759486%	25.032278481012657%	2.052848101265823%	0.		
	4						•		
<pre>In [248]: df['Hispanic'] = df['Hispanic'].str.replace("%","").astype(float)</pre>									
In [249]:	<pre>df['White'] = df['White'].str.replace("%","").astype(float)</pre>								
In [250]:	<pre>df['Black'] = df['Black'].str.replace("%","").astype(float)</pre>								
In [251]:	<pre>[251]: df['Native'] = df['Native'].str.replace("%","").astype(float)</pre>								
In [252]:	df	['Asian'] :	= df['Asi	an'].str.replace("	%","").astype(floa	t)			
In [253]:	df	['Pacific'] = df['F	Cacific'].str.repla	ce(<mark>"%",""</mark>).astype(float)			

```
In [254]: df.info()
          <class 'pandas.core.frame.DataFrame'>
          RangeIndex: 42 entries, 0 to 41
          Data columns (total 12 columns):
           #
               Column
                           Non-Null Count
                                           Dtype
                           -----
                                           ----
           0
               State
                           42 non-null
                                           object
               TotalPop
                          42 non-null
                                           int64
           1
           2
               Hispanic
                          42 non-null
                                           float64
           3
               White
                          42 non-null
                                           float64
           4
                          42 non-null
                                           float64
               Black
           5
                          42 non-null
                                           float64
               Native
           6
               Asian
                          42 non-null
                                           float64
           7
               Pacific
                          42 non-null
                                           float64
           8
                          42 non-null
                                           object
               Income
           9
               GenderPop 42 non-null
                                           object
           10
               Male
                           42 non-null
                                           int32
               Female
                           42 non-null
                                           int32
           11
          dtypes: float64(6), int32(2), int64(1), object(3)
          memory usage: 3.7+ KB
 In [ ]:
```

Graph-1

The first visualization your boss wants you to make is a scatterplot that shows average income in a state vs proportion of women in that state.

```
In [255]: style.use('ggplot')
    plt.figure(figsize = (9,9))
    sns.scatterplot(x=df['Female'],y=df['Income'])
    plt.xticks(rotation = 90)
    plt.title("Avg Income in a state vs proportion of women")
    plt.show()
```



In [256]: df.head()

Out[256]:

	State	TotalPop	Hispanic	White	Black	Native	Asian	Pacific	
0	Connecticut	3593222	15.604831	67.677053	10.348068	0.126208	4.021981	0.018599	\$76
1	District of Columbia	647484	9.165922	33.103911	51.776536	0.200559	3.383240	0.029609	\$754
2	Florida	19645772	21.338543	59.083749	15.165676	0.210451	2.283174	0.051510	\$5069
3	Georgia	10006693	8.418242	54.286306	32.088298	0.187583	3.097649	0.046602	\$508
4	Hawaii	1406299	9.186709	25.032278	2.052848	0.144937	36.592089	8.758861	\$732
4									•

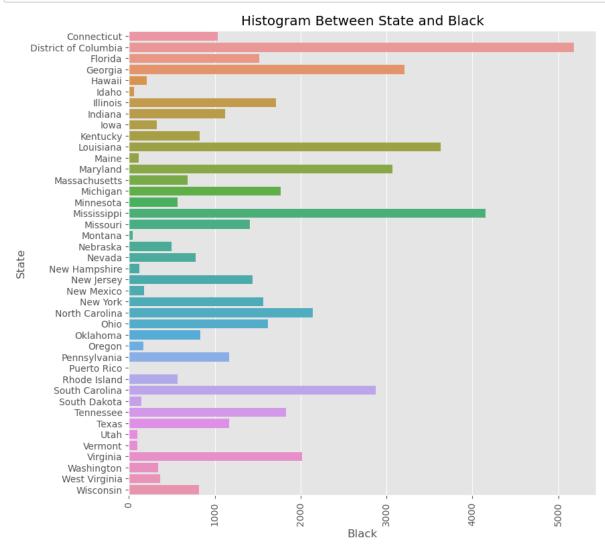
```
In [266]: # Converting From % to normal numbers
            df["Hispanic"] = df["Hispanic"] * 100
In [267]: style.use('ggplot')
            plt.figure(figsize = (9,9))
             sns.barplot(x = df['Hispanic'] , y= df['State'])
             plt.xticks(rotation = 90)
             plt.title("Histogram Between State and Hispanic")
            plt.show()
                                               Histogram Between State and Hispanic
                      Connecticut -
                District of Columbia
                         Florida
                         Georgia
                         Hawaii
                          Idaho
                          Illinois
                         Indiana
                          lowa
                        Kentucky
                       Louisiana
                          Maine
                       Maryland
                    Massachúsetts
                        Michigan
                       Minnesota
                       Mississippi
                        Missouri
                        Montana
                       Nebraska
                        Nevada
                   New Hampshire
                      New Jersey
                     New Mexico
                       New York
                    North Carolina
                           Ohio
                       Oklahoma
                         Oregon
```

```
In [268]: # Converting From % to normal numbers
df["White"] = df["White"] * 100
```

```
In [269]: style.use('ggplot')
              plt.figure(figsize = (9,9))
              sns.barplot(x = df['White'] , y= df['State'])
              plt.xticks(rotation = 90)
              plt.title("Histogram Between State and White")
              plt.show()
                         MILLINESULA
                         Mississippi
                           Missouri
                          Montana
                          Nebraska
               State
                           Nevada
                     New Hampshire
                        New Jersey
New Mexico
                          New York
                      North Carolina
                              Ohio
                         Oklahoma
                           Oregon
                       Pennsylvania
                        Puerto Rico
                       Rhode Island
                     South Carolina
                      South Dakota
                         Tennessee
                             Texas
                             Utah
                          Vermont
                           Virginia
                        Washington
                       West Virginia
                         Wisconsin
                                                                  4000
                                                                                  0009
                                                                                                  8000
                                                                       White
```

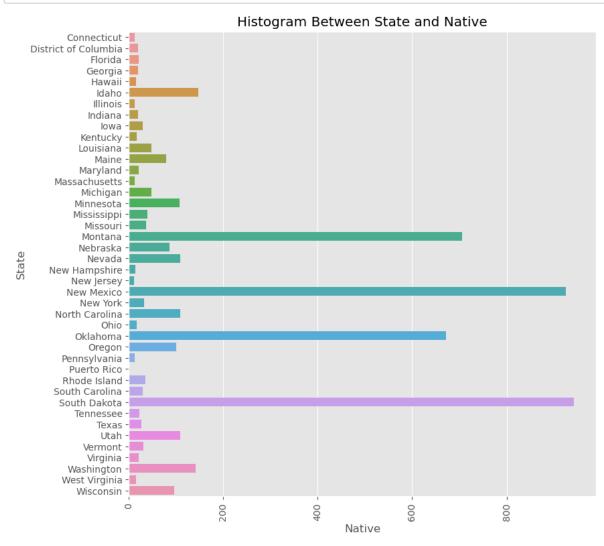
```
In [270]: # Converting From % to normal numbers
df["Black"] = df["Black"] * 100
```

```
In [271]:
    style.use('ggplot')
    plt.figure(figsize = (9,9))
    sns.barplot(x = df['Black'] , y= df['State'])
    plt.xticks(rotation = 90)
    plt.title("Histogram Between State and Black")
    plt.show()
```



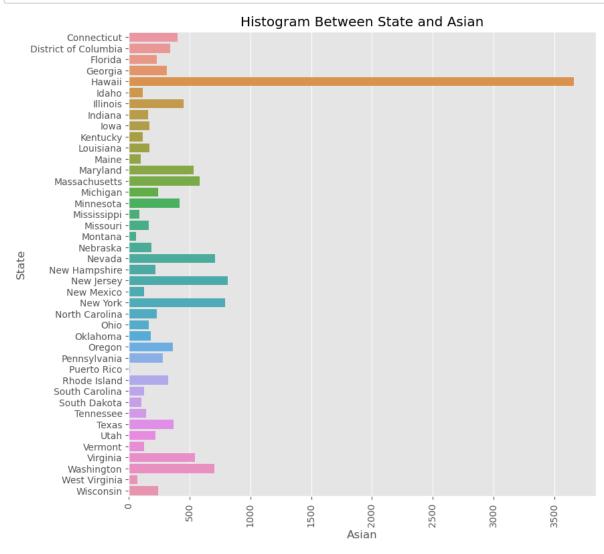
```
In [272]: # Converting From % to normal numbers
df["Native"] = df["Native"] * 100
```

```
In [273]: style.use('ggplot')
   plt.figure(figsize = (9,9))
   sns.barplot(x = df['Native'] , y= df['State'])
   plt.xticks(rotation = 90)
   plt.title("Histogram Between State and Native")
   plt.show()
```



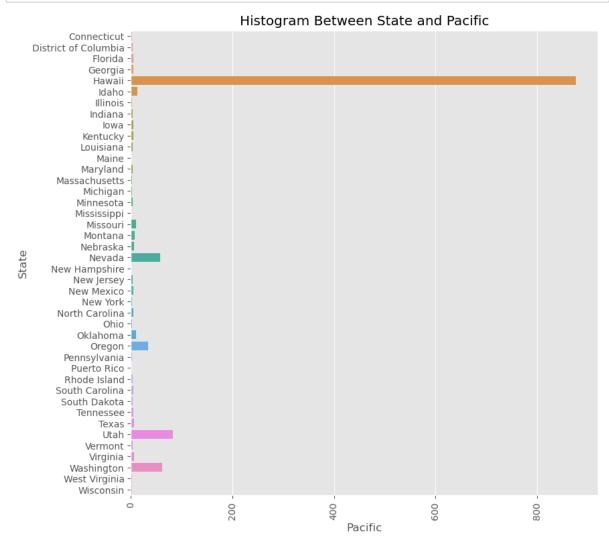
```
In [274]: # Converting From % to normal numbers
df["Asian"] = df["Asian"] * 100
```

```
In [275]: style.use('ggplot')
   plt.figure(figsize = (9,9))
   sns.barplot(x = df['Asian'] , y= df['State'])
   plt.xticks(rotation = 90)
   plt.title("Histogram Between State and Asian")
   plt.show()
```



```
In [276]: # Converting From % to normal numbers
df["Pacific"] = df["Pacific"] * 100
```

```
In [277]:
    style.use('ggplot')
    plt.figure(figsize = (9,9))
    sns.barplot(x = df['Pacific'] , y= df['State'])
    plt.xticks(rotation = 90)
    plt.title("Histogram Between State and Pacific")
    plt.show()
```



```
In [ ]:
```