Project #1: US Census Data Cleaning

Importing libraries

In [ ]:
```
import glob
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

Listing csv files

In [ ]:
```
csv_files = glob.glob('*.{}'.format('csv'))
csv_files
```

Out[ ]:
```
['states0.csv',
 'states1.csv',
 'states2.csv',
 'states3.csv',
 'states4.csv',
 'states5.csv',
 'states6.csv',
 'states7.csv',
 'states8.csv',
 'states9.csv']
```

Reading and combining csv files

In [ ]:
```
combinedData=pd.DataFrame()

for file in csv_files:
    data=pd.read_csv(file)
    combinedData=pd.concat([combinedData, data], ignore_index=True)

combinedData.to_csv("us_census.csv", index=False)
```

Dropping 'Unnamed: 0' column

In [ ]:
```
#drop unnamed: 0 column
combinedData.drop(combinedData.columns[0], axis=1, inplace=True)
combinedData.to_csv("us_census.csv", index=False)
# DO NOT RUN THIS CODE
```

Reading us_census.csv

In [ ]:
```
combinedData=pd.read_csv("us_census.csv")
```

Checking dtypes of columns

In [ ]:
```
#dtypes of the columns
combinedData.dtypes
```

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

Displaying first 5 rows

In [ ]:
```
#head of the data
combinedData.head()
```

Out[ ]:

	State	TotalPop	Hispanic	White	Black	
0	Alabama	4830620	3.7516156462584975%	61.878656462585%	31.25297619047618%	0.45323
1	Alaska	733375	5.909580838323351%	60.910179640718574%	2.8485029940119775%	16.391
2	Arizona	6641928	29.565921052631502%	57.120000000000026%	3.8509868421052658%	4.355
3	Arkansas	2958208	6.215474452554738%	71.13781021897813%	18.968759124087573%	0.52291
4	California	38421464	37.291874687968054%	40.21578881677474%	5.677396405391911%	0.405292

Converting dtype of objects into appropriate dtypes for data manipulation

In [ ]:
```
combinedData['Hispanic']=combinedData['Hispanic'].str.replace('%','')
combinedData['Hispanic']=combinedData['Hispanic'].astype('float64')

combinedData['White']=combinedData['White'].str.replace('%','')
combinedData['White'] = combinedData['White'].astype('float64')

combinedData['Black']=combinedData['Black'].str.replace('%','')
combinedData['Black'] = combinedData['Black'].astype('float64')

combinedData['Native']=combinedData['Native'].str.replace('%','')
combinedData['Native'] = combinedData['Native'].astype('float64')

combinedData['Asian']=combinedData['Asian'].str.replace('%','')
combinedData['Asian'] = combinedData['Asian'].astype('float64')
```

```
combinedData['Pacific']=combinedData['Pacific'].str.replace('%','')
combinedData['Pacific'] = combinedData['Pacific'].astype('float64')

combinedData['Income']=combinedData['Income'].str.replace('$','')
combinedData['Income'] = combinedData['Income'].astype('float64')
```

Splitting male and female population values into two columns

In [ ]:
```
combinedData[['MalePop', 'FemalePop']]=combinedData['GenderPop'].str.split('_', exp

combinedData['MalePop']=combinedData['MalePop'].str.replace('M','')
combinedData['MalePop'] = combinedData['MalePop'].astype('int64')

combinedData['FemalePop']=combinedData['FemalePop'].str.replace('F','')
combinedData['FemalePop']=combinedData['FemalePop'].replace('', np.nan).astype(floa
combinedData['FemalePop'] = combinedData['FemalePop'].astype('int64')

combinedData.drop(['GenderPop'], axis=1, inplace=True)
```

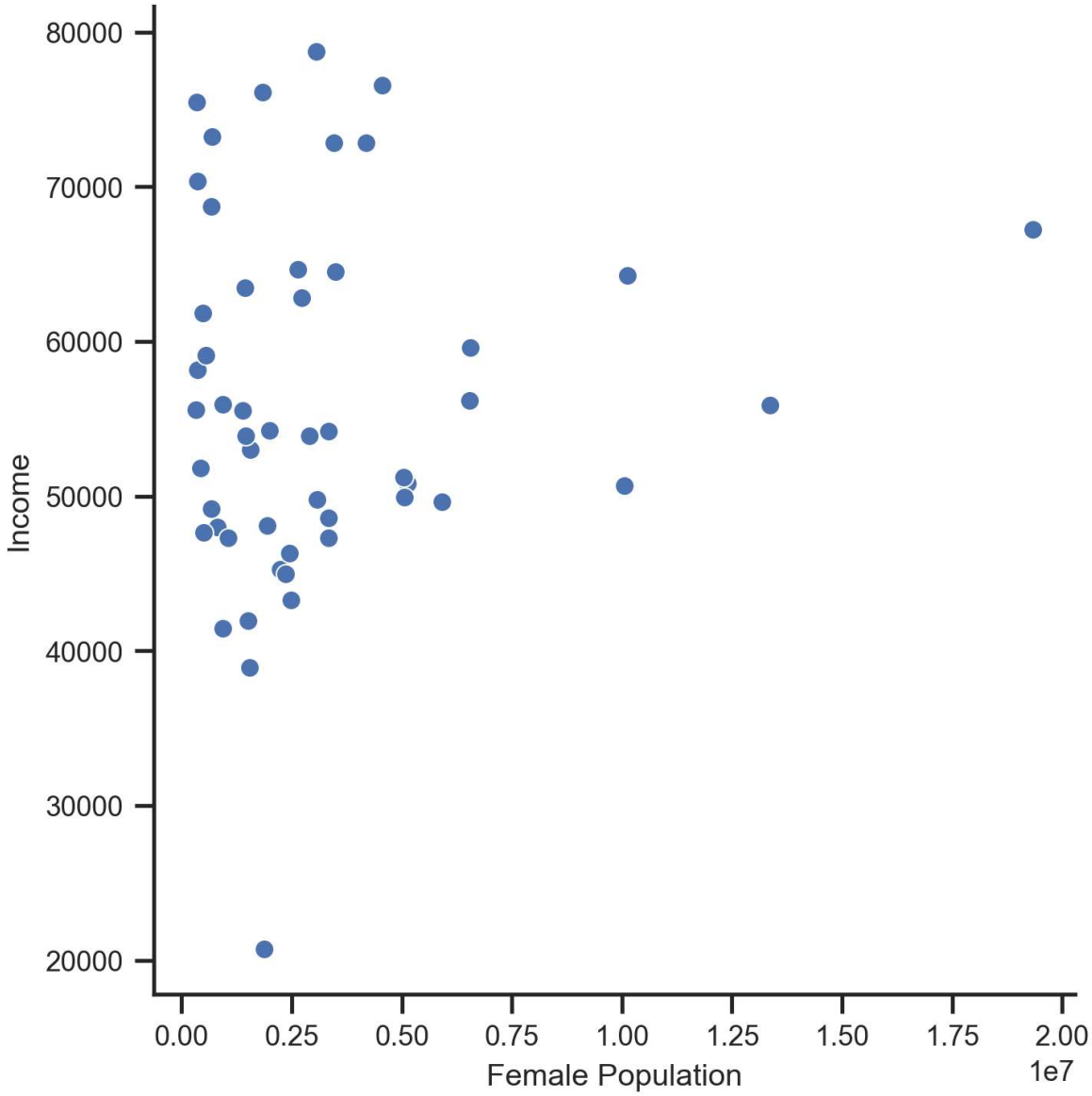
Dropping duplicate values

In [ ]:
```
combinedData.drop_duplicates(inplace=True)
```

Scatter plot of Income against Female Population

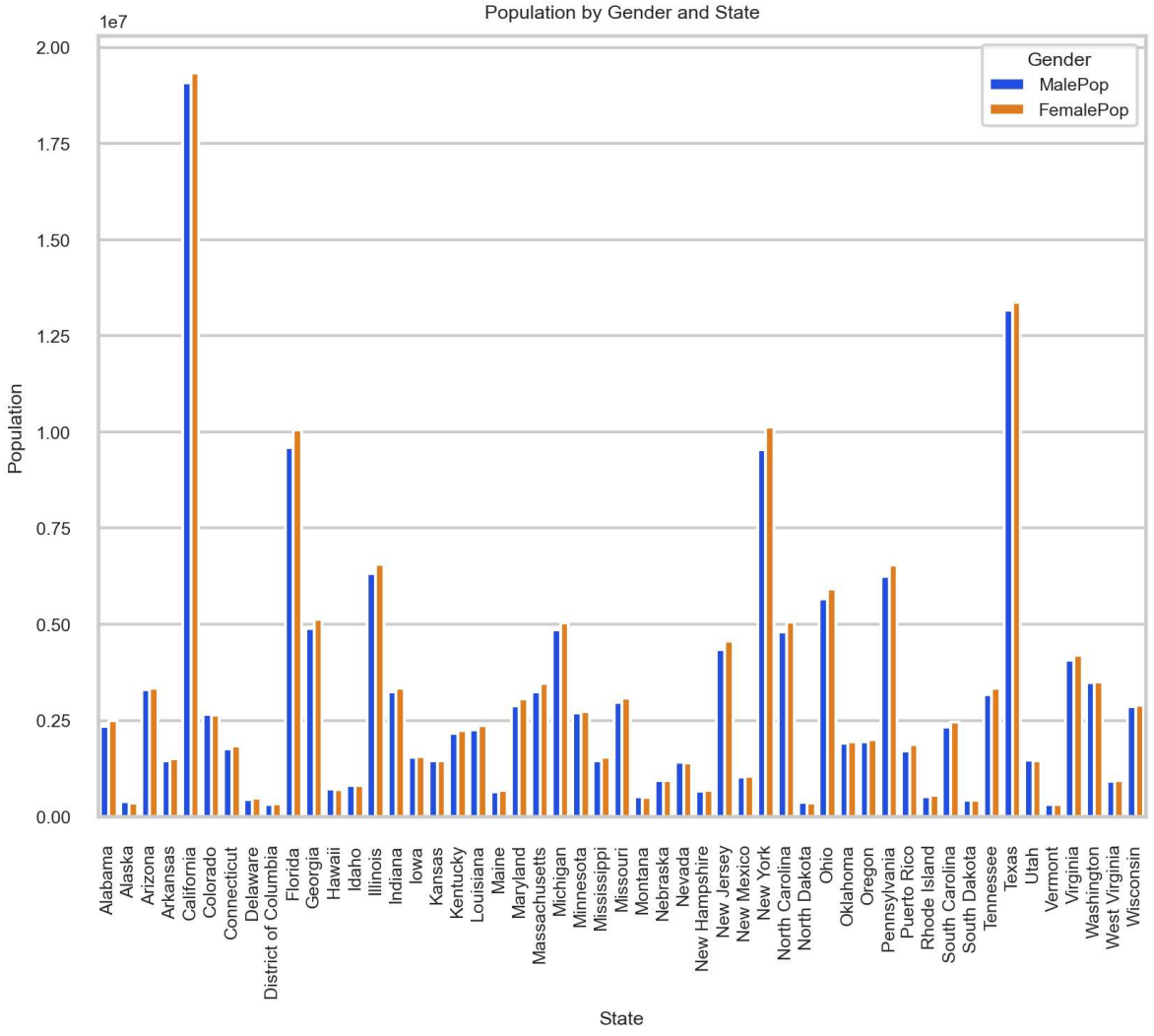
In [ ]:
```
sns.set(rc={'figure.dpi':300, 'savefig.dpi':300})
sns.set_theme()
sns.set_context('notebook', font_scale=0.8)
sns.set_style('ticks')
plot=sns.relplot(x='FemalePop', y='Income', data=combinedData)
#labeling the axis
plot.set_axis_labels("Female Population", "Income")
```

Out[ ]:
```
<seaborn.axisgrid.FacetGrid at 0x1d501682fd0>
```



Bar plot of Population by Gender and State

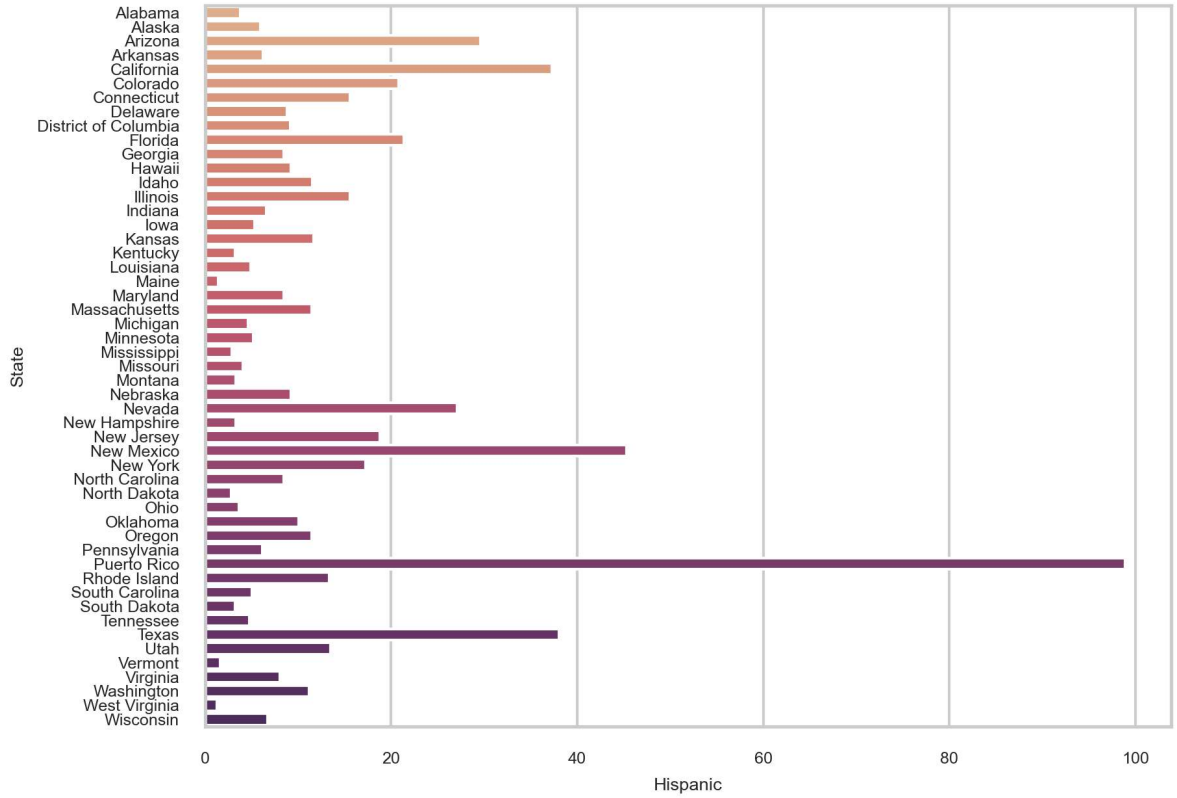
In [ ]:
```
data = sns.load_dataset('tips')

state_gender_data = combinedData.groupby(
    ['State'])[['MalePop', 'FemalePop']].sum().reset_index()

state_gender_data_melted = pd.melt(state_gender_data, id_vars=['State'], value_vars
                                   'MalePop', 'FemalePop'], var_name='Gender', valu

sns.set(style="whitegrid")
sns.set_context('notebook', font_scale=0.55)
ax = sns.barplot(x='State', y='Population', hue='Gender',
                 data=state_gender_data_melted, palette='bright')

ax.set(xlabel='State', ylabel='Population',
       title='Population by Gender and State')
plt.xticks(rotation=90)

plt.show()
```



Bar plot of Hispanics per State

In [ ]:
```
sns.set_context('notebook', font_scale=0.55)
sns.barplot(x='Hispanic', y='State', data=combinedData, palette='flare')
```

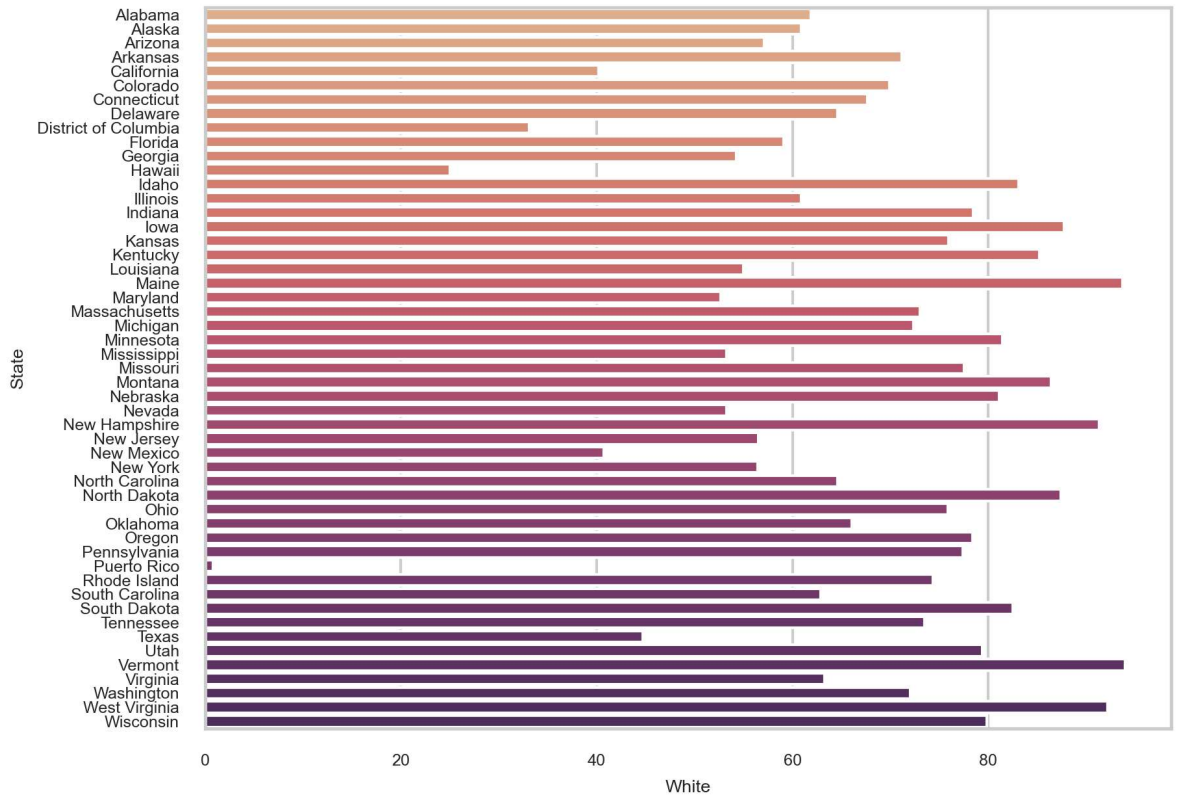
Out[ ]:
```
<Axes: xlabel='Hispanic', ylabel='State'>
```



Bar plot of Whites per State

In [ ]:
```
sns.barplot(x='White', y='State', data=combinedData, palette='flare')
```

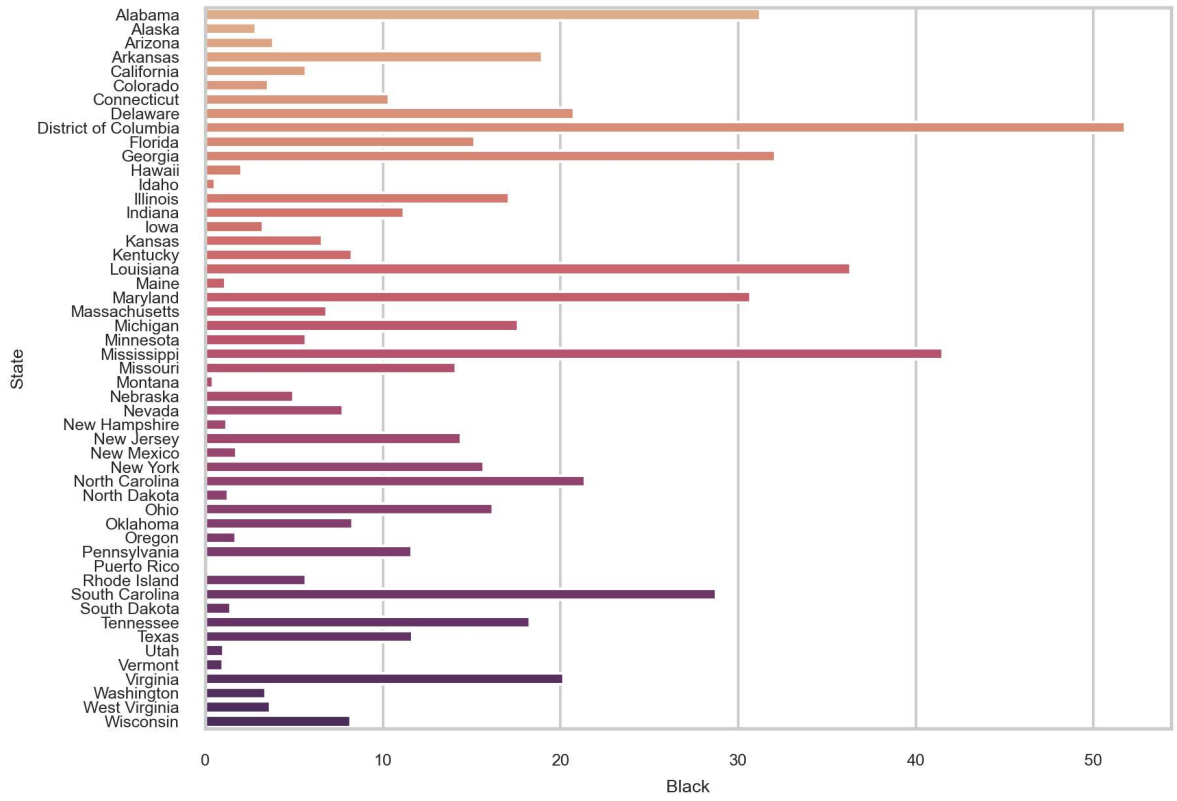
Out[ ]:
```
<Axes: xlabel='White', ylabel='State'>
```



Bar plot of Blacks per State

In [ ]:
```
sns.barplot(x='Black', y='State', data=combinedData, palette='flare')
```

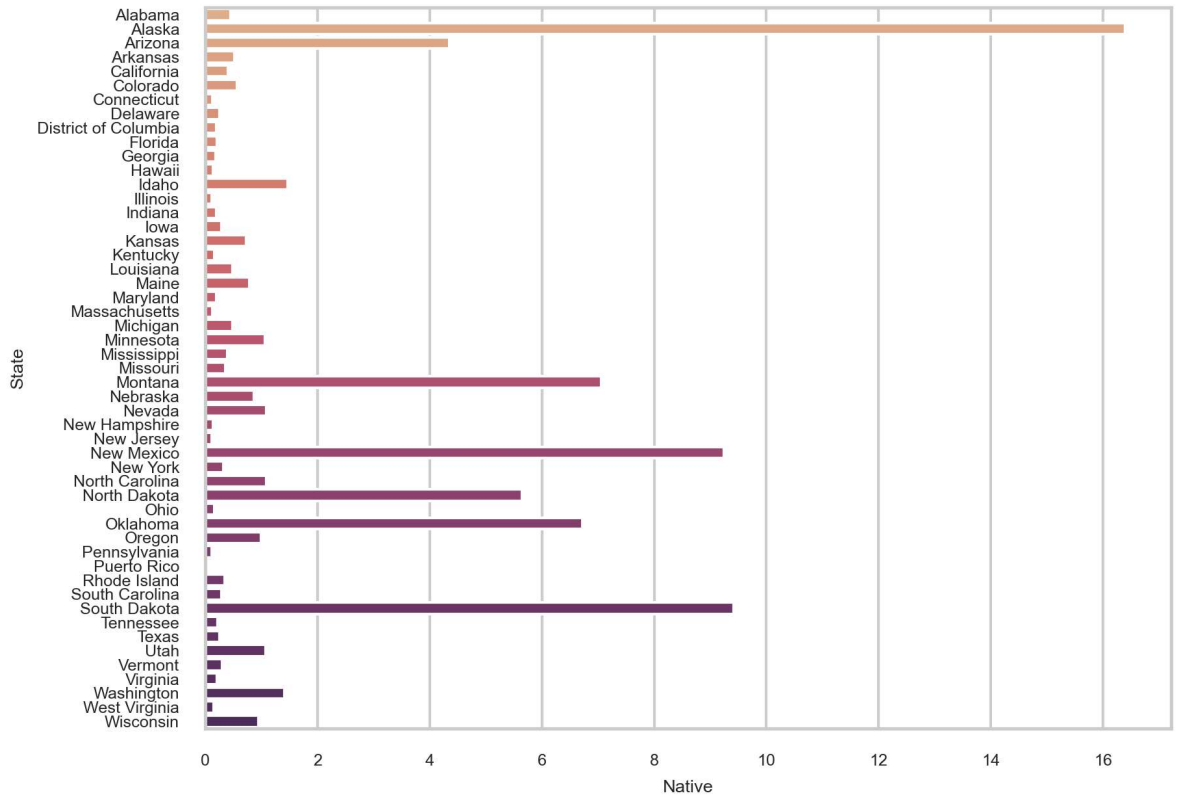
Out[ ]:
```
<Axes: xlabel='Black', ylabel='State'>
```



Bar plot of Natives per State

In [ ]:
```
sns.barplot(x='Native', y='State', data=combinedData, palette='flare')
```

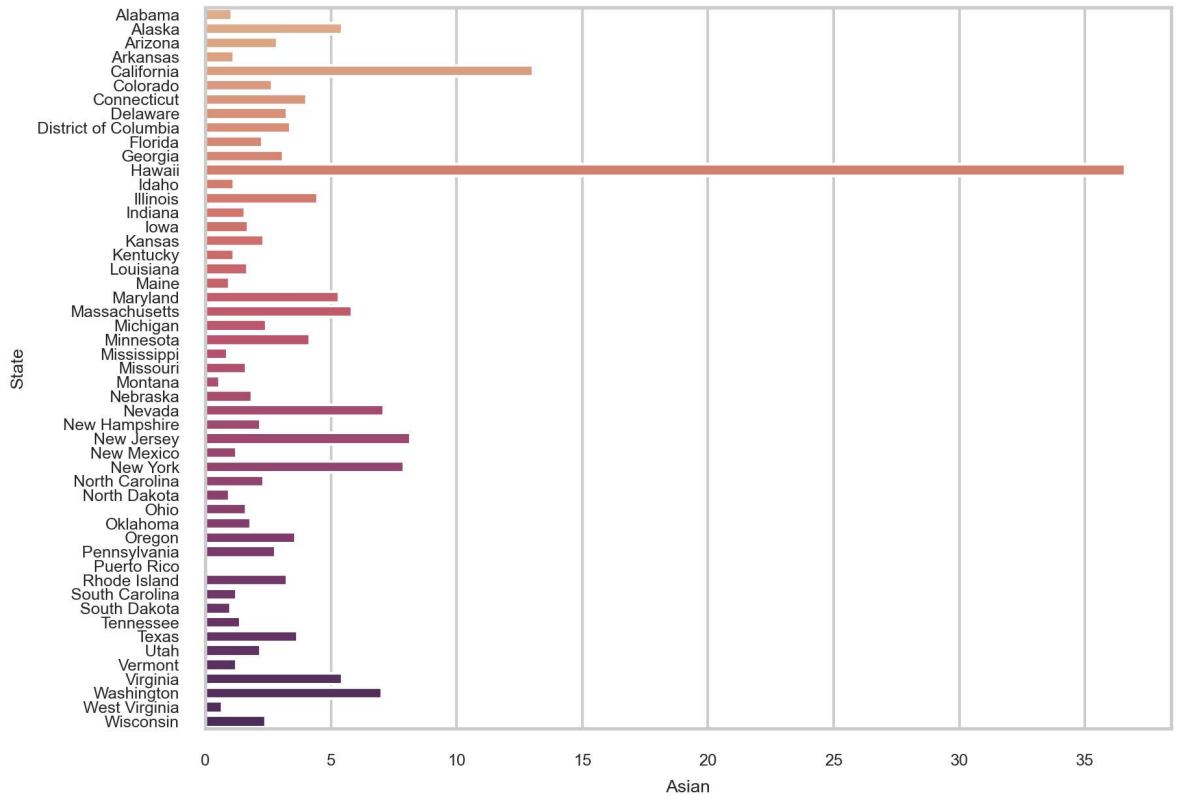
Out[ ]:
```
<Axes: xlabel='Native', ylabel='State'>
```



Bar plot of Asians per State

In [ ]:
```
sns.barplot(x='Asian', y='State', data=combinedData, palette='flare')
```

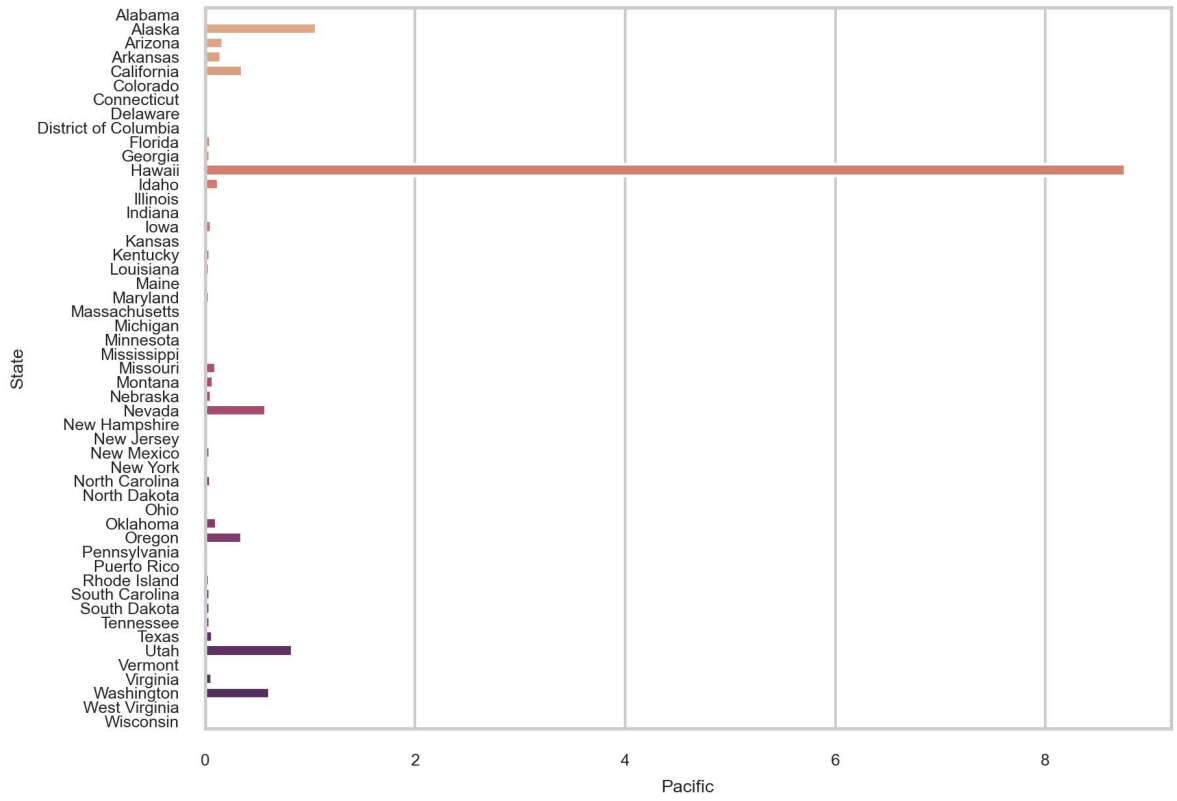
Out[ ]:
```
<Axes: xlabel='Asian', ylabel='State'>
```



Bar plot of Pacifics per State

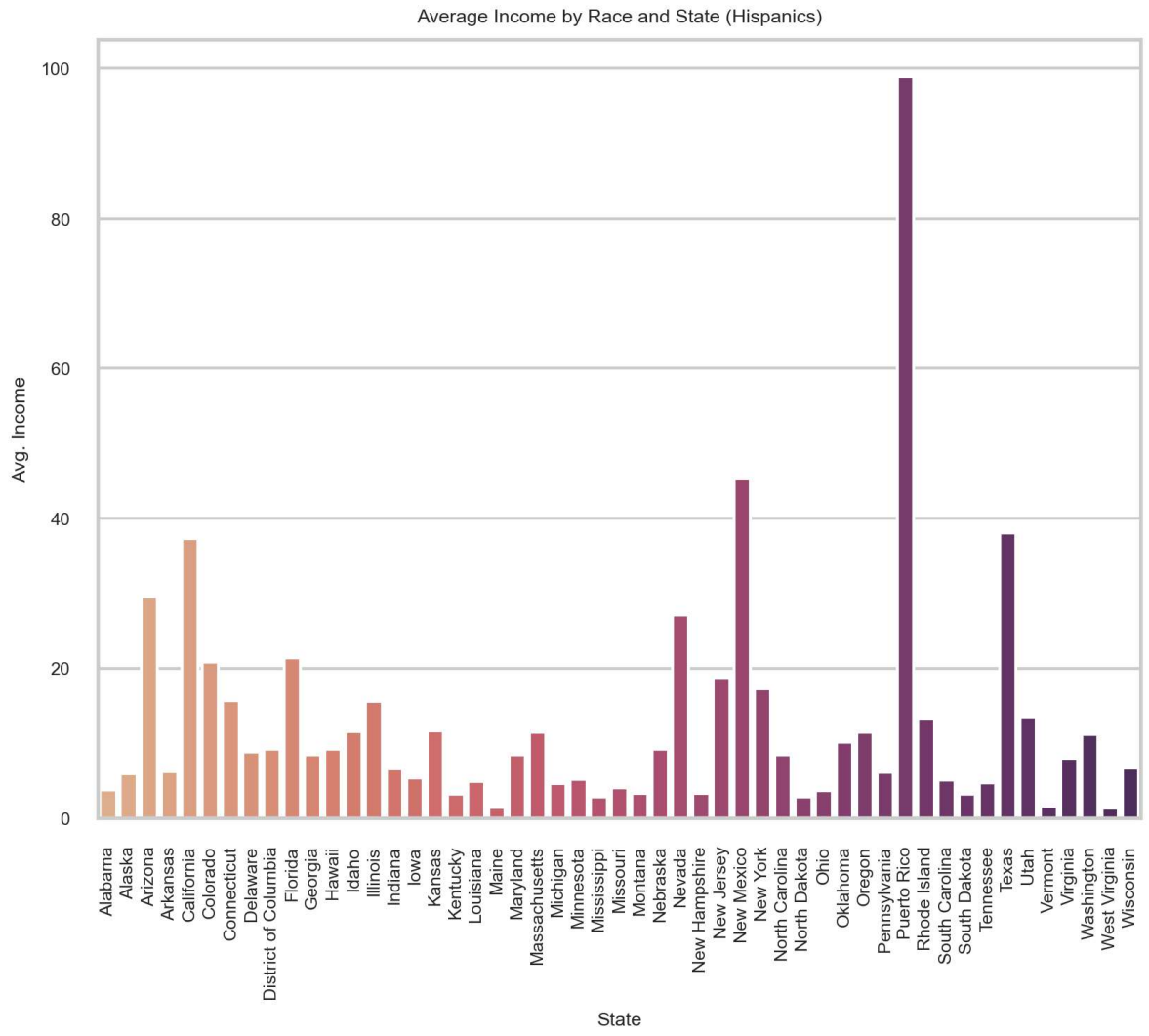
In [ ]:
```
sns.barplot(x='Pacific', y='State', data=combinedData, palette='flare')
```

Out[ ]:
```
<Axes: xlabel='Pacific', ylabel='State'>
```



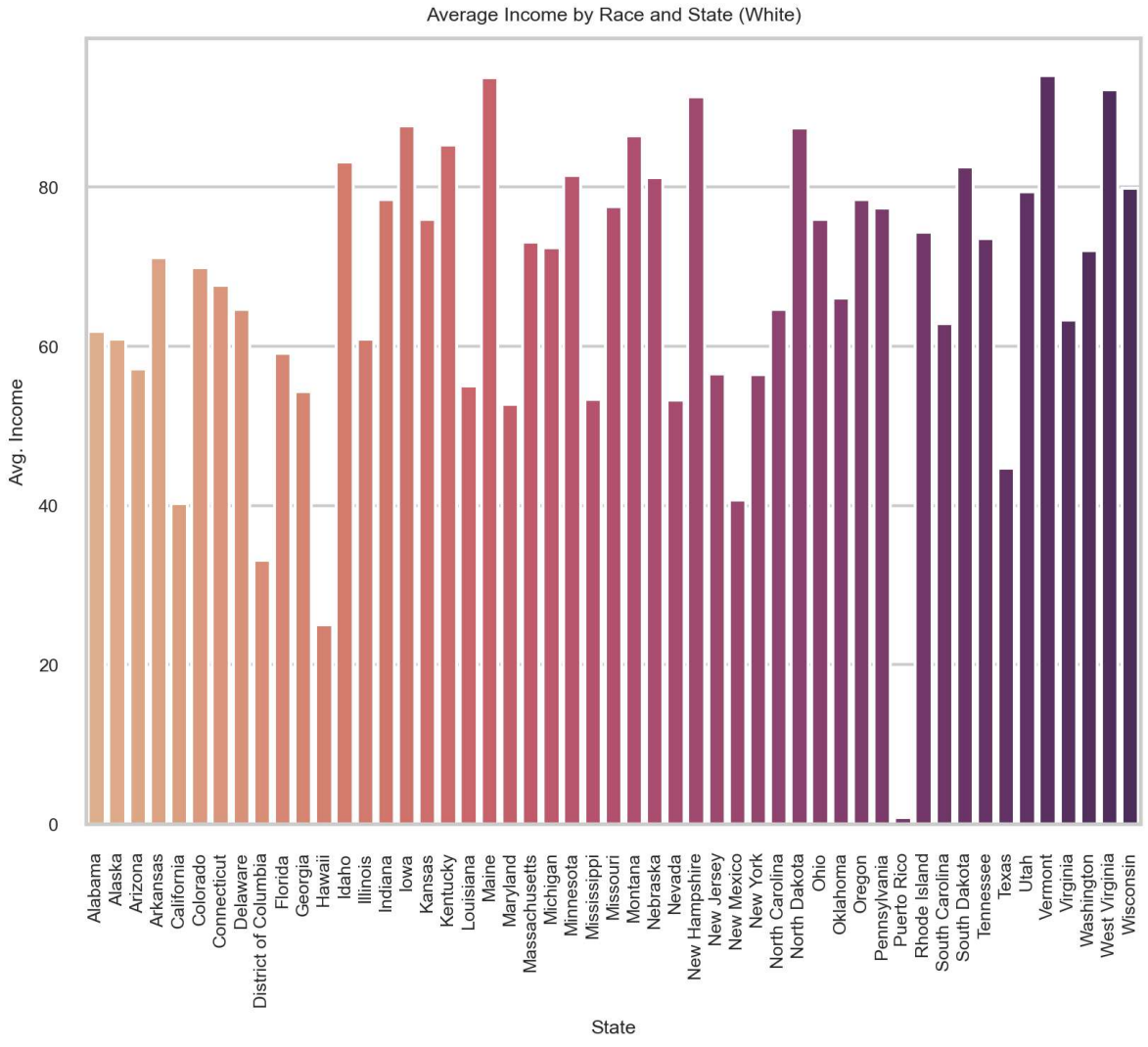
Bar plot of Average Income of Hispanics by State

In [ ]:
```
grouped_data = combinedData.groupby('State')[
    ['Hispanic', 'Income']].mean().reset_index()

melted_data = pd.melt(grouped_data, id_vars=['State'], value_vars=[
                      'Hispanic'], var_name='Race', value_name='AvgIncome')

sns.set_style("whitegrid")
ax=sns.barplot(x="State", y="AvgIncome", data=melted_data, palette='flare')
ax.set(xlabel='State', ylabel='Avg. Income',
       title='Average Income by Race and State (Hispanics)')
plt.xticks(rotation=90)

plt.show()
```



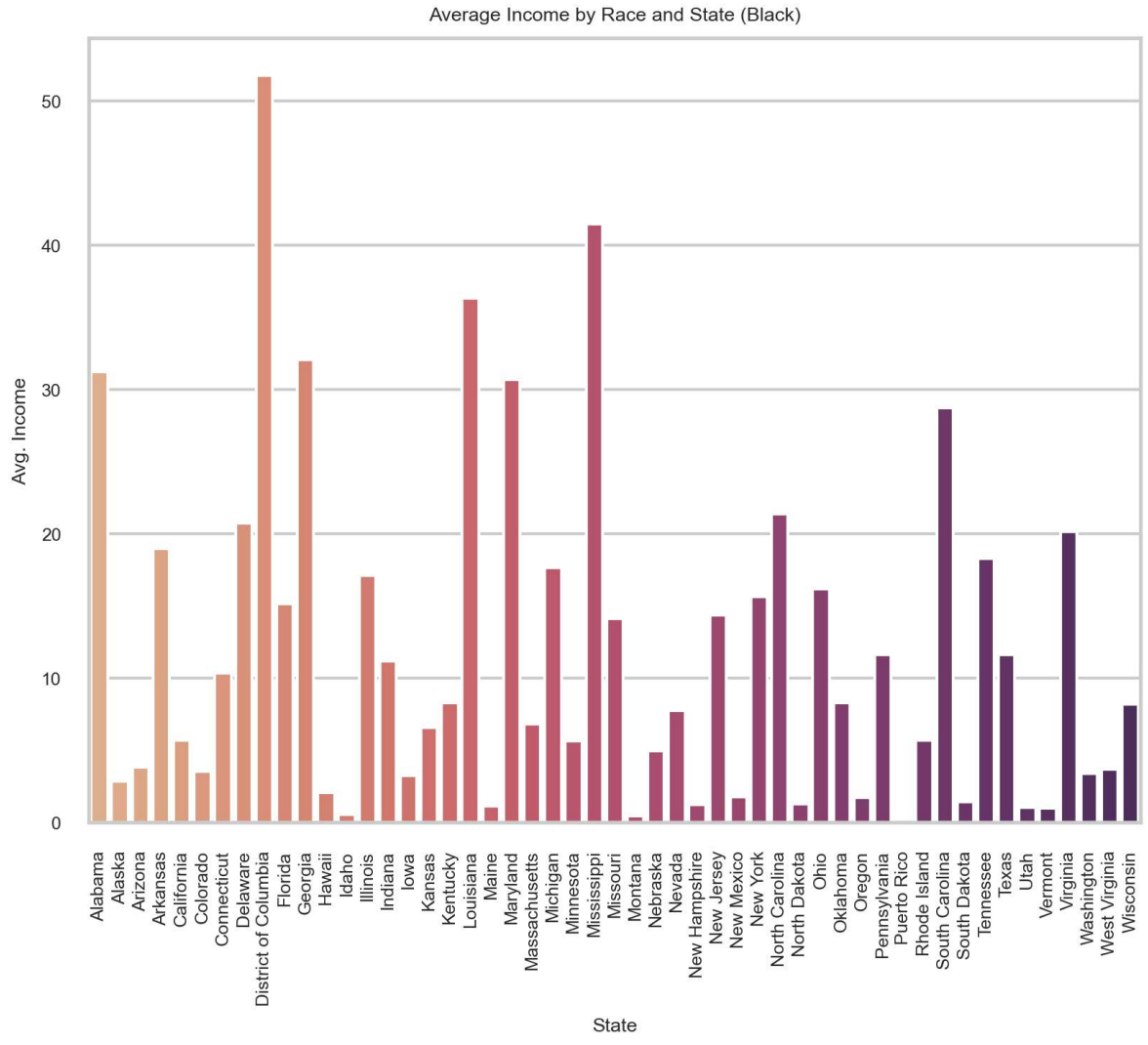
Bar plot of Average Income of Whites by State

In [ ]:
```
grouped_data = combinedData.groupby('State')[
    ['White', 'Income']].mean().reset_index()

melted_data = pd.melt(grouped_data, id_vars=['State'], value_vars=[
                      'White'], var_name='Race', value_name='AvgIncome')

sns.set_style("whitegrid")
ax=sns.barplot(x="State", y="AvgIncome", data=melted_data, palette='flare')
ax.set(xlabel='State', ylabel='Avg. Income',
       title='Average Income by Race and State (White)')
plt.xticks(rotation=90)

plt.show()
```



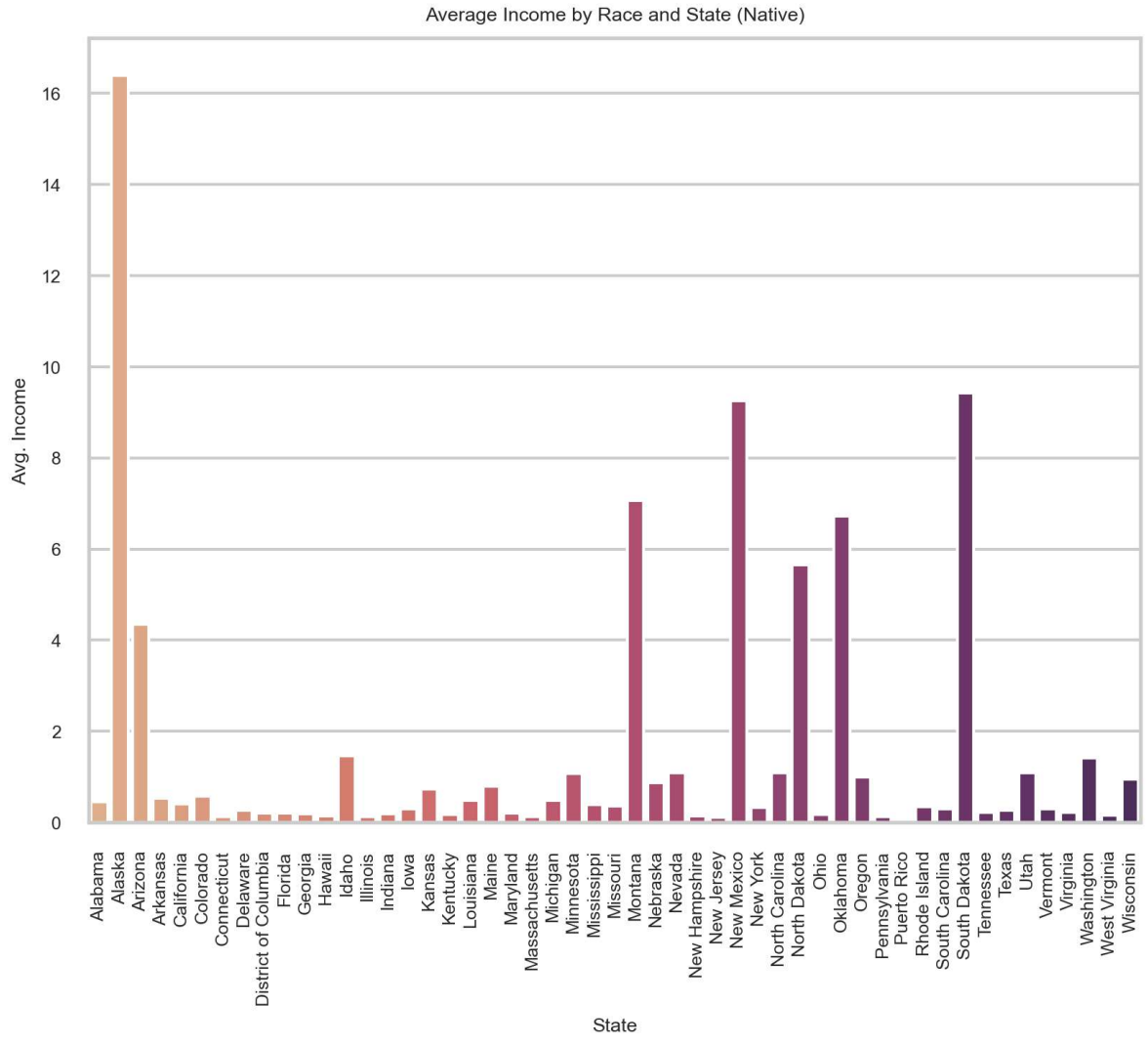
Bar plot of Average Income of Blacks by State

In [ ]:
```
grouped_data = combinedData.groupby('State')[
    ['Black', 'Income']].mean().reset_index()

melted_data = pd.melt(grouped_data, id_vars=['State'], value_vars=[
                      'Black'], var_name='Race', value_name='AvgIncome')

sns.set_style("whitegrid")
ax=sns.barplot(x="State", y="AvgIncome", data=melted_data, palette='flare')
ax.set(xlabel='State', ylabel='Avg. Income',
       title='Average Income by Race and State (Black)')
plt.xticks(rotation=90)

plt.show()
```



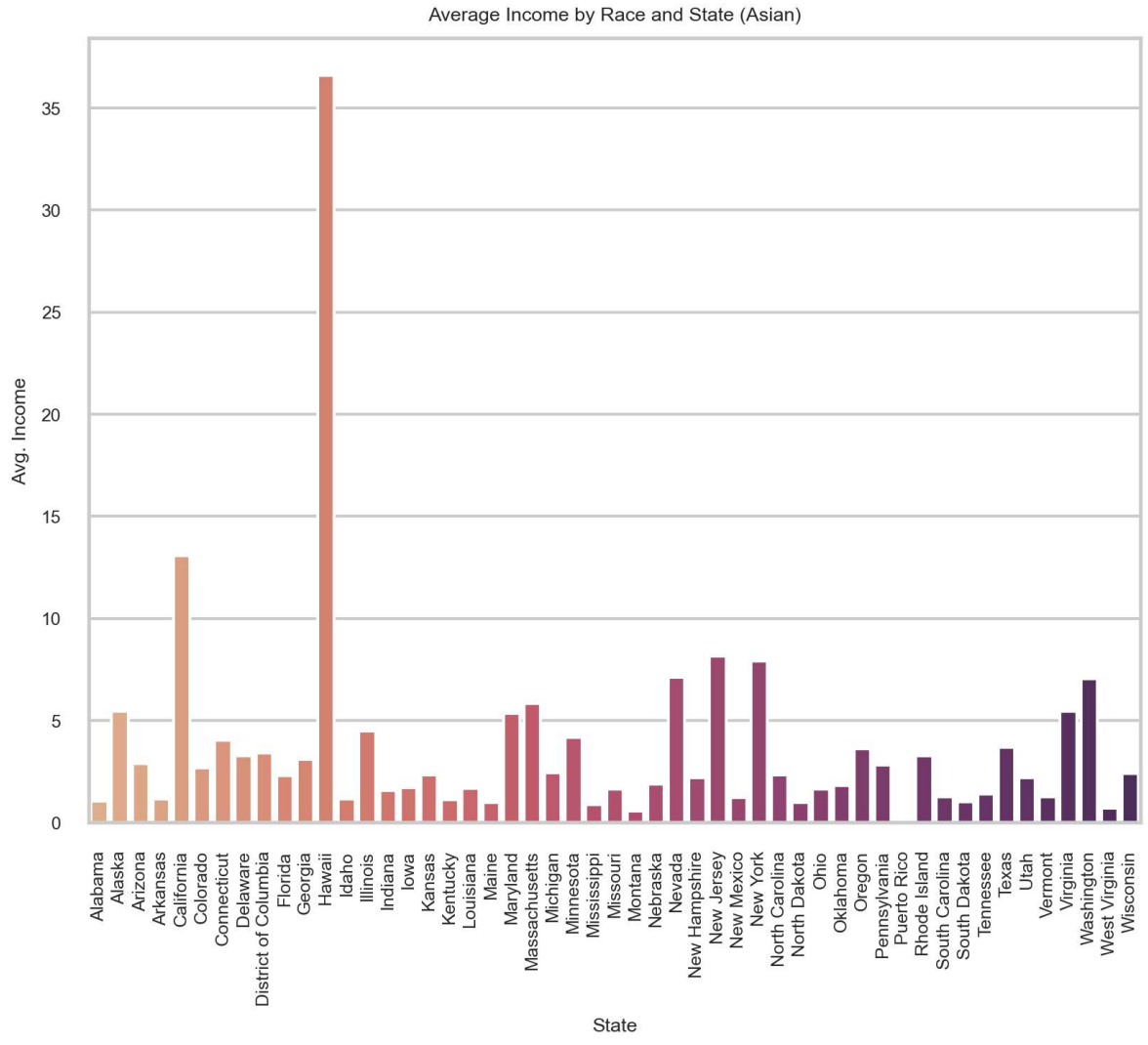
Bar plot of Average Income of Natives by State

In [ ]:
```
grouped_data = combinedData.groupby('State')[
    ['Native', 'Income']].mean().reset_index()

melted_data = pd.melt(grouped_data, id_vars=['State'], value_vars=[
                      'Native'], var_name='Race', value_name='AvgIncome')

sns.set_style("whitegrid")
ax=sns.barplot(x="State", y="AvgIncome", data=melted_data, palette='flare')
ax.set(xlabel='State', ylabel='Avg. Income',
       title='Average Income by Race and State (Native)')
plt.xticks(rotation=90)

plt.show()
```



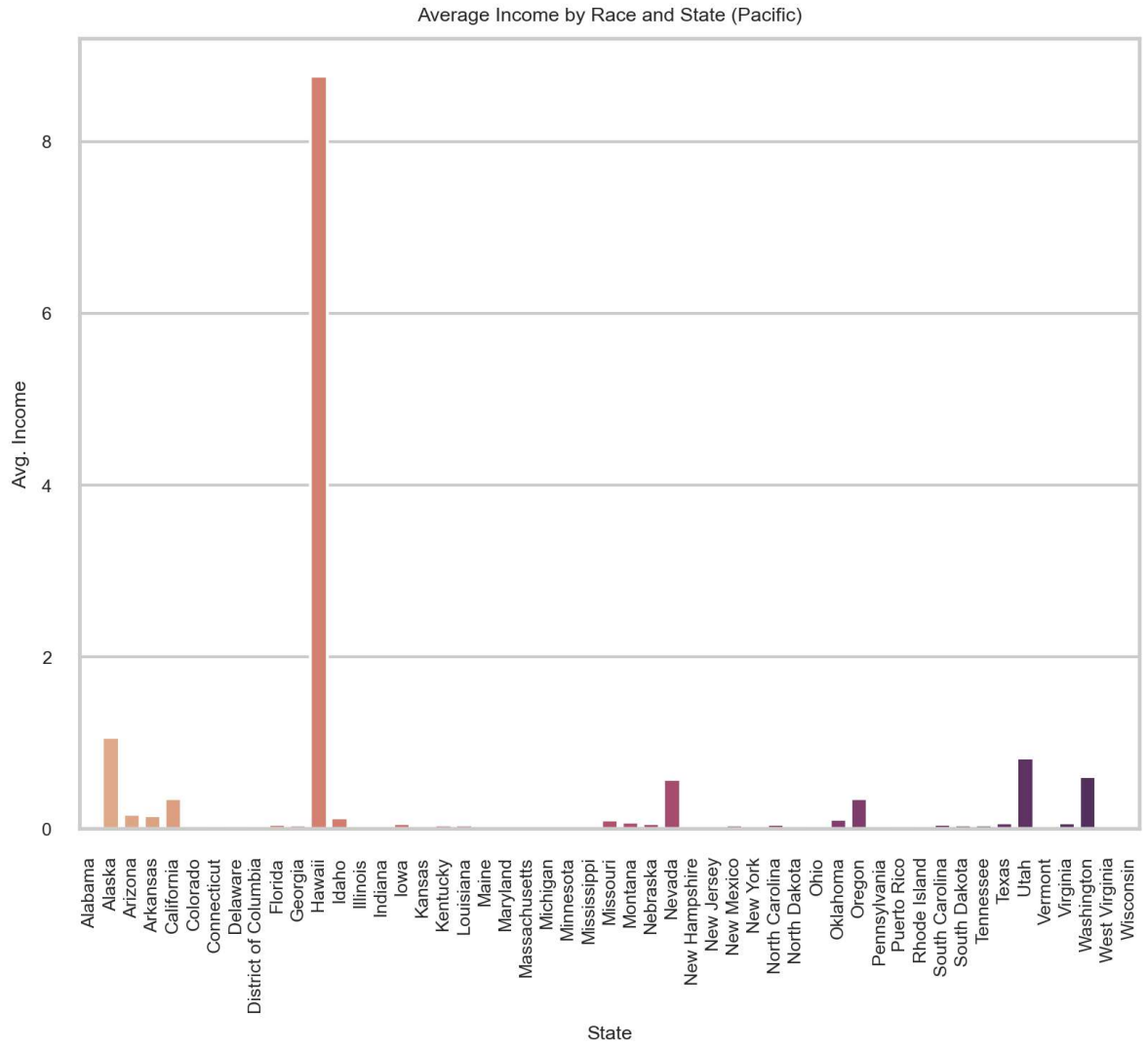
Bar plot of Average Income of Asians by State

In [ ]:
```
grouped_data = combinedData.groupby('State')[
    ['Asian', 'Income']].mean().reset_index()

melted_data = pd.melt(grouped_data, id_vars=['State'], value_vars=[
                      'Asian'], var_name='Race', value_name='AvgIncome')

sns.set_style("whitegrid")
ax=sns.barplot(x="State", y="AvgIncome", data=melted_data, palette='flare')
ax.set(xlabel='State', ylabel='Avg. Income',
       title='Average Income by Race and State (Asian)')
plt.xticks(rotation=90)

plt.show()
```



Bar plot of Average Income of Pacifics by State

In [ ]:
```
grouped_data = combinedData.groupby('State')[
    ['Pacific', 'Income']].mean().reset_index()

melted_data = pd.melt(grouped_data, id_vars=['State'], value_vars=[
                      'Pacific'], var_name='Race', value_name='AvgIncome')

sns.set_style("whitegrid")
ax=sns.barplot(x="State", y="AvgIncome", data=melted_data, palette='flare')
ax.set(xlabel='State', ylabel='Avg. Income',
       title='Average Income by Race and State (Pacific)')
plt.xticks(rotation=90)

plt.show()
```



Histogram of Frequency Distribution of Race per State

In [ ]:
```
race_data = combinedData[['Hispanic', 'White',
                          'Black', 'Native', 'Asian', 'Pacific']]

sns.histplot(data=race_data, x='Hispanic', kde=True)
sns.histplot(data=race_data, x='White', kde=True)
sns.histplot(data=race_data, x='Black', kde=True)
sns.histplot(data=race_data, x='Native', kde=True)
sns.histplot(data=race_data, x='Asian', kde=True)
sns.histplot(data=race_data, x='Pacific', kde=True)

plt.title('Frequency Distribution of Race')
plt.xlabel('Percentage')
plt.ylabel('Frequency')
plt.legend(labels=['Hispanic', 'White', 'Black', 'Native', 'Asian', 'Pacific'])
```

Out[ ]:
```
<matplotlib.legend.Legend at 0x1d508d7f550>
```