Unemployment

Exploration of Unemployment in the US in 2015

Hypothesis: Gender, Race, and Occupation Type have an impact on predicting unemployment in a specific US County

Null Hypothesis: Gender, Race, and Occupation does NOT have an impact on predicting unemployment in a specific US County

Preparation

Import libraries

In [153]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
import statsmodels.api as sm
import seaborn as sns
```

Import the .csv data set

In [17]:
```
df = pd.read_csv('acs2015_county_data.csv')
df
```

Out[17]:

	CensusId	State	County	TotalPop	Men	Women	Hispanic	White	Black	Native	...
0	1001	Alabama	Autauga	55221	26745	28476	2.6	75.8	18.5	0.4	...
1	1003	Alabama	Baldwin	195121	95314	99807	4.5	83.1	9.5	0.6	...
2	1005	Alabama	Barbour	26932	14497	12435	4.6	46.2	46.7	0.2	...
3	1007	Alabama	Bibb	22604	12073	10531	2.2	74.5	21.4	0.4	...
4	1009	Alabama	Blount	57710	28512	29198	8.6	87.9	1.5	0.3	...
...	...	...	...	...	...	...	...	...	...	...	...
3215	72145	Puerto Rico	Vega Baja	56858	27379	29479	96.4	3.4	0.1	0.0	...
3216	72147	Puerto Rico	Vieques	9130	4585	4545	96.7	2.9	0.0	0.0	...
3217	72149	Puerto Rico	Villalba	24685	12086	12599	99.7	0.0	0.0	0.0	...
3218	72151	Puerto Rico	Yabucoa	36279	17648	18631	99.8	0.2	0.0	0.0	...
3219	72153	Puerto Rico	Yauco	39474	19047	20427	99.5	0.5	0.0	0.0	...

3220 rows × 37 columns

Columns

It looks like the data set is structured in the following

1. County ID; State; Country
2. Total Pop
3. Sex
4. Race
5. Income, Income Per Capita, & respective errors
6. Poverty & Child Poverty %
7. Occupation Type
8. Commute Type
9. Type of Employment

In [69]:
```
df.columns
```

Out[69]:
```
Index(['CensusId', 'State', 'County', 'TotalPop', 'Men', 'Women', 'Hisp
anic',
       'White', 'Black', 'Native', 'Asian', 'Pacific', 'Citizen', 'Inco
me',
       'IncomeErr', 'IncomePerCap', 'IncomePerCapErr', 'Poverty',
       'ChildPoverty', 'Professional', 'Service', 'Office', 'Constructi
on',
       'Production', 'Drive', 'Carpool', 'Transit', 'Walk', 'OtherTrans
p',
       'WorkAtHome', 'MeanCommute', 'Employed', 'PrivateWork', 'PublicW
ork',
       'SelfEmployed', 'FamilyWork', 'Unemployment'],
      dtype='object')
```

Shape

In this data set, there are 37 total attributes & 3220 observations

In [19]:
```
df.shape
```

Out[19]:
```
(3220, 37)
```

Describe

Summarize all attributes

In [20]:
```
df.describe()
```

Out[20]:

	CensusId	TotalPop	Men	Women	Hispanic	White	
count	3220.000000	3.220000e+03	3.220000e+03	3.220000e+03	3220.000000	3220.000000	3220.(
mean	31393.605280	9.940935e+04	4.889694e+04	5.051241e+04	11.011522	75.428789	8.(
std	16292.078954	3.193055e+05	1.566813e+05	1.626620e+05	19.241380	22.932890	14.:
min	1001.000000	8.500000e+01	4.200000e+01	4.300000e+01	0.000000	0.000000	0.(
25%	19032.500000	1.121800e+04	5.637250e+03	5.572000e+03	1.900000	64.100000	0.!
50%	30024.000000	2.603500e+04	1.293200e+04	1.305700e+04	3.900000	84.100000	1.!
75%	46105.500000	6.643050e+04	3.299275e+04	3.348750e+04	9.825000	93.200000	9.(
max	72153.000000	1.003839e+07	4.945351e+06	5.093037e+06	99.900000	99.800000	85.!

8 rows × 35 columns

Dataframe Types

Two attritributes are objects; the rest are either in64 or float64

In [21]:
```
df.dtypes
```

Out[21]:
```
CensusId              int64
State                object
County               object
TotalPop              int64
Men                   int64
Women                 int64
Hispanic            float64
White               float64
Black               float64
Native              float64
Asian               float64
Pacific             float64
Citizen               int64
Income              float64
IncomeErr           float64
IncomePerCap          int64
IncomePerCapErr       int64
Poverty             float64
ChildPoverty        float64
Professional        float64
Service             float64
Office              float64
Construction        float64
Production          float64
Drive               float64
Carpool             float64
Transit             float64
Walk                float64
OtherTransp         float64
WorkAtHome          float64
MeanCommute         float64
Employed              int64
PrivateWork         float64
PublicWork          float64
SelfEmployed        float64
FamilyWork          float64
Unemployment        float64
dtype: object
```

In [63]:
```
df_clean = df[['TotalPop', 'Men', 'Women', 'Hispanic',
        'White', 'Black', 'Native', 'Asian', 'Pacific', 'Citizen', 'Incom
e',
         'IncomePerCap', 'Poverty', 'ChildPoverty',
         'Professional', 'Service', 'Office', 'Construction',
        'Production',
        'Employed', 'Unemployment']]
df_clean
```

Out[63]:

	TotalPop	Men	Women	Hispanic	White	Black	Native	Asian	Pacific	Citizen	...	Inco
0	55221	26745	28476	2.6	75.8	18.5	0.4	1.0	0.0	40725	...	
1	195121	95314	99807	4.5	83.1	9.5	0.6	0.7	0.0	147695	...	
2	26932	14497	12435	4.6	46.2	46.7	0.2	0.4	0.0	20714	...	
3	22604	12073	10531	2.2	74.5	21.4	0.4	0.1	0.0	17495	...	
4	57710	28512	29198	8.6	87.9	1.5	0.3	0.1	0.0	42345	...	
...	...	...	...	...	...	...	...	...	...	...	...	
3215	56858	27379	29479	96.4	3.4	0.1	0.0	0.0	0.0	43656	...	
3216	9130	4585	4545	96.7	2.9	0.0	0.0	0.0	0.0	7085	...	
3217	24685	12086	12599	99.7	0.0	0.0	0.0	0.0	0.0	18458	...	
3218	36279	17648	18631	99.8	0.2	0.0	0.0	0.1	0.0	27924	...	
3219	39474	19047	20427	99.5	0.5	0.0	0.0	0.0	0.0	30661	...	

3220 rows × 21 columns

In [70]:
```
df_clean.columns
```

Out[70]:
```
Index(['TotalPop', 'Men', 'Women', 'Hispanic', 'White', 'Black', 'Nativ
e',
       'Asian', 'Pacific', 'Citizen', 'Income', 'IncomePerCap', 'Povert
y',
       'ChildPoverty', 'Professional', 'Service', 'Office', 'Constructi
on',
       'Production', 'Employed', 'Unemployment'],
      dtype='object')
```

Top 5 Highest Unemployment Rate by County

In [164]:
```
top_unemployed_counties = df[['State', 'County', 'TotalPop','Unemploymen
t']].sort_values(by=['Unemployment'], ascending=False)
top_unemployed_counties.head(10)
```

Out[164]:

	State	County	TotalPop	Unemployment
3142	Puerto Rico	Adjuntas	18962	36.5
3183	Puerto Rico	Lares	28727	35.2
3179	Puerto Rico	Jayuya	15890	31.7
3196	Puerto Rico	Orocovis	22595	31.2
3158	Puerto Rico	Cataño	26680	30.8
3208	Puerto Rico	San Sebastián	40471	29.4
2376	South Dakota	Corson	4149	29.4
3213	Puerto Rico	Utuado	31474	28.8
2412	South Dakota	Oglala Lakota	14153	28.7
81	Alaska	Kusilvak Census Area	7914	28.6

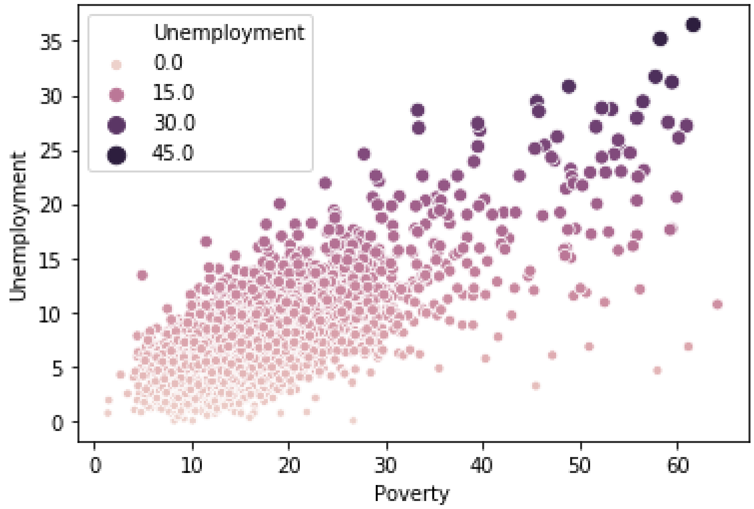
Takeaway: As you can see, 7 of the most unemployed counties are located within Puerto Rico and 2 of the most unemployed counties are in South Dakota.

Data Visualizations

Scatter Line - Poverty % vs Unemployment

In [165]:
```
sns.scatterplot(data=df_clean, x="Poverty", y="Unemployment", size = 'Unemployment', hue = 'Unemployment')
```

Out[165]: <matplotlib.axes._subplots.AxesSubplot at 0x7fa837e537d0>

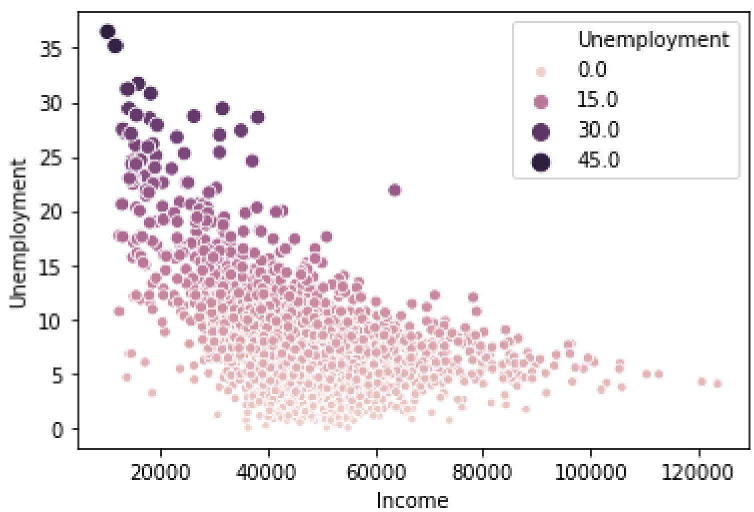


Counties w/ higher Poverty % tend to have a higher Unemployment rate

Scatterplot - Income vs Unemployment

In [135]:
```
sns.scatterplot(data=df_clean, x="Income", y="Unemployment", size = 'Unemployment', hue = 'Unemployment')
```

Out[135]: <matplotlib.axes._subplots.AxesSubplot at 0x7fa844940390>

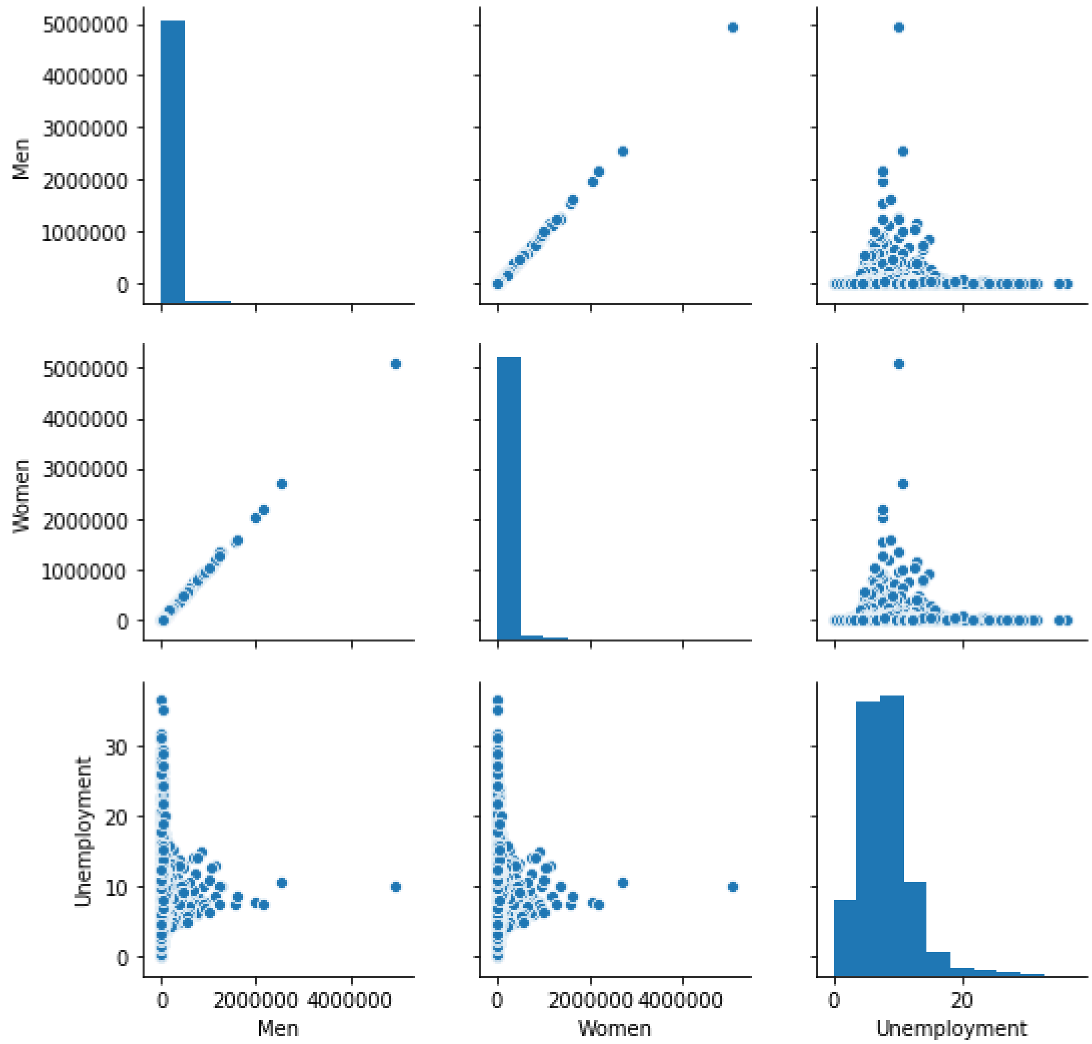


Counties w/ a lower Income tend to have a higher unemployment rate

Pairplot on Sex

In [148]:
```
sns.pairplot(df_clean[['Men', 'Women', 'Unemployment']])
```

Out[148]: <seaborn.axisgrid.PairGrid at 0x7fa86003f110>

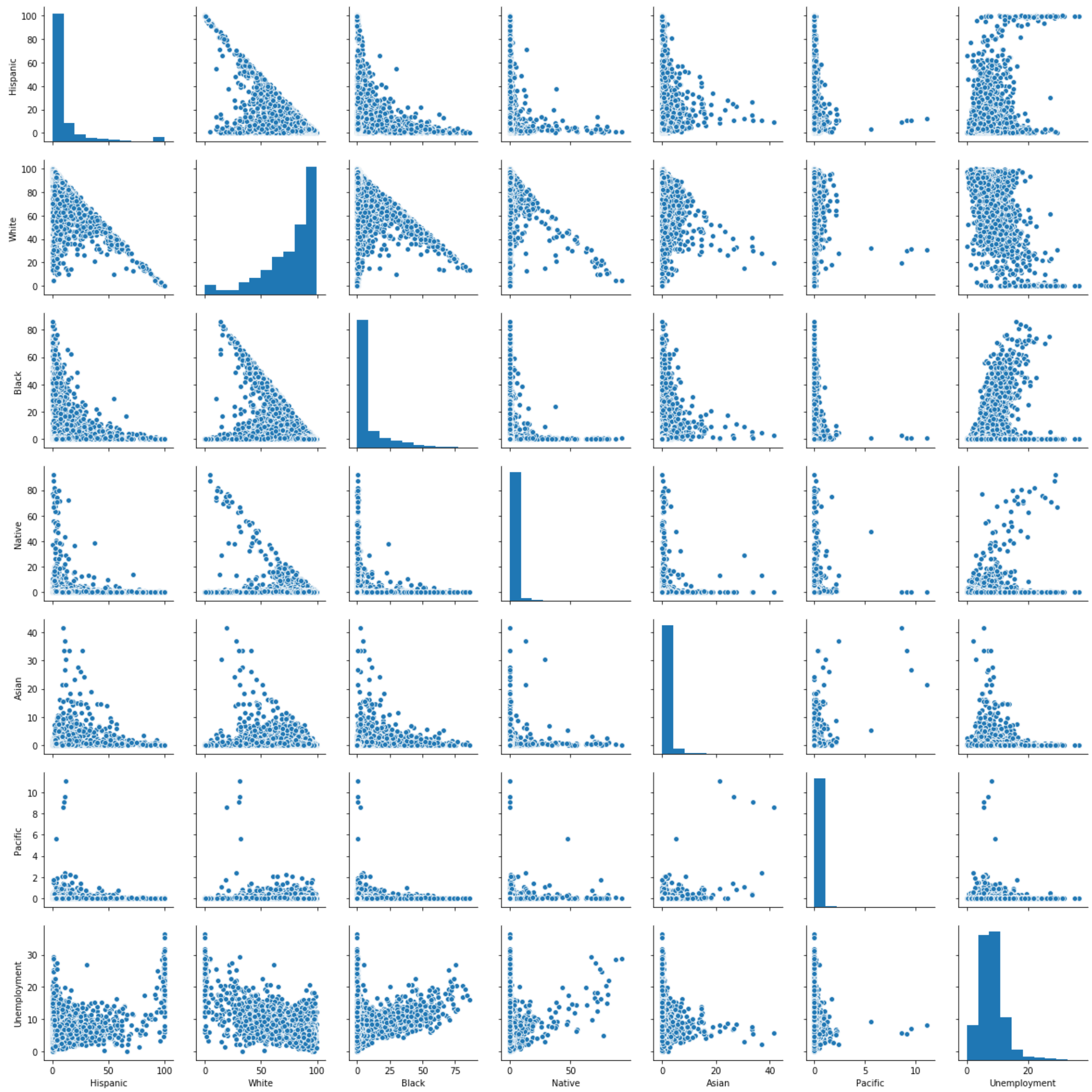


Takeaway: It seems Gender does not have an impact on unemployment

Pairplot on Ethnicity

In [152]:
```
sns.pairplot(df_clean[['Hispanic',
       'White', 'Black', 'Native', 'Asian', 'Pacific', 'Unemployment']])
```

Out[152]: <seaborn.axisgrid.PairGrid at 0x7fa861d6aa10>

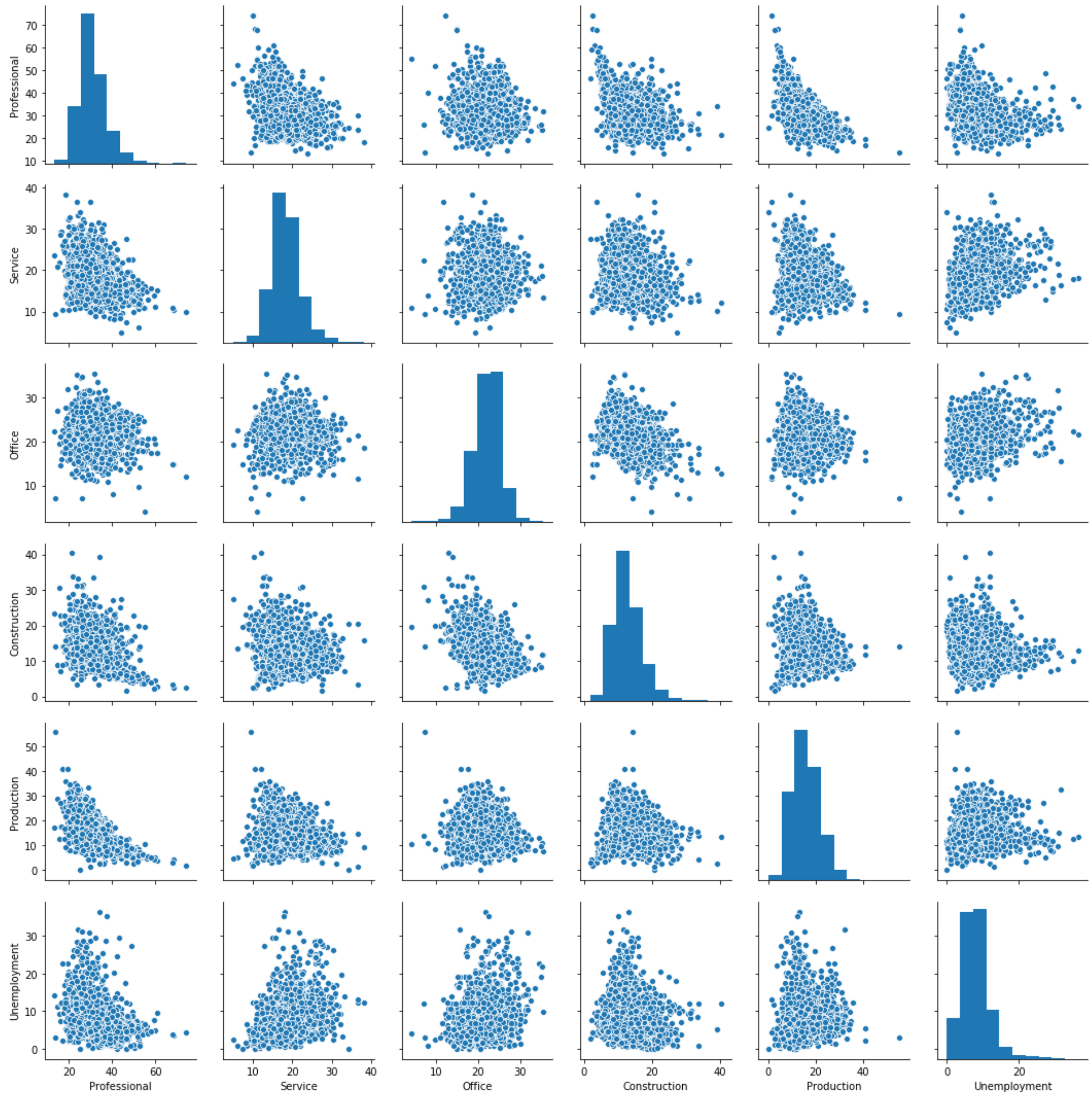


Takeaway: Communities with less diversity (i.e. lacking Asian and Pacific communities) are more likely to have a higher unemployment rate

Pairplot on Occupation

In [150]:
```
sns.pairplot(df_clean[['Professional', 'Service', 'Office', 'Constructio
n',
        'Production', 'Unemployment']])
```

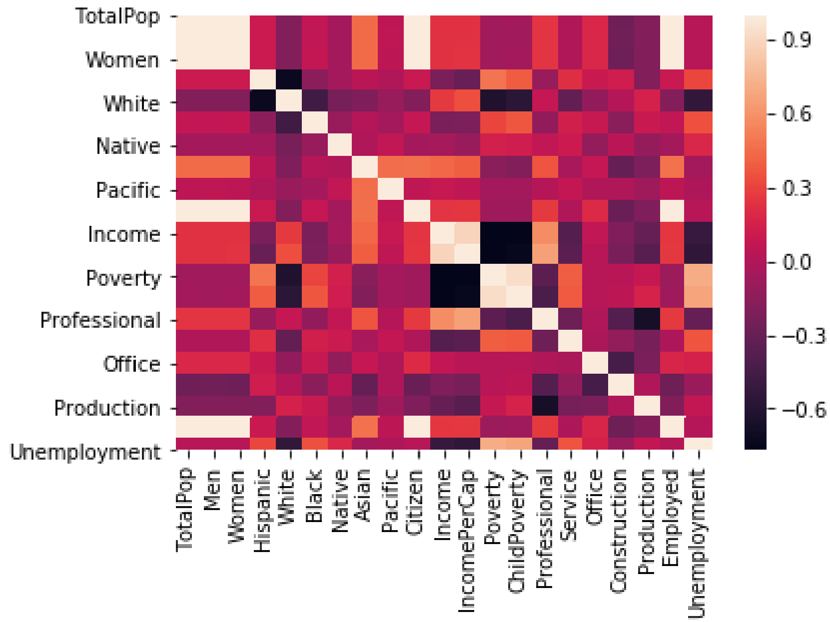
Out[150]: <seaborn.axisgrid.PairGrid at 0x7fa86056e110>



Takeaway: There is a negative correlation between the unemployment rate and the percentage of construction jobs. Also from 2011 onwards there has been a significant increase in the spending on the construction sector in the US budget.

In [138]:
```
sns.heatmap(df_clean.corr())
```

Out[138]: `<matplotlib.axes._subplots.AxesSubplot at 0x7fa85c3158d0>`



Based on the correlation heat map above, it seems like income, poverty, and occupation type are highly correlated

Pivot Table on Sex vs County

In [55]:
```
pivot = pd.pivot_table(data=df_clean,
                       columns=['County'],
                       values=['Men', 'Women', 'TotalPop'],
                       aggfunc = 'sum'
                       )
pivot.reindex(['Men', 'Women', 'TotalPop'])
```

Out[55]:

County	Abbeville	Acadia	Accomack	Ada	Adair	Adams	Addison	Adjuntas	Aguada	Agu
Men	12308	30023	16117	208879	36220	409098	18355	9266	19912	
Women	12689	32140	16998	208622	37854	406893	18588	9696	20691	
TotalPop	24997	62163	33115	417501	74074	815991	36943	18962	40603	

3 rows × 1928 columns

Prepartion for Linear Regression Model

In [87]:
```
x = df_clean[['TotalPop', 'Men', 'Women', 'Hispanic', 'White', 'Black', 'Native',
        'Asian', 'Pacific', 'Citizen', 'Income', 'IncomePerCap', 'Poverty',
        'ChildPoverty', 'Professional', 'Service', 'Office', 'Construction',
        'Production', 'Employed']]
y = df_clean[['Unemployment']]
```

Ran into NaN Error

Need to drop NaN values

In [81]:
```
df_clean.isna().sum()
```

Out[81]:
```
TotalPop        0
Men             0
Women           0
Hispanic        0
White           0
Black           0
Native          0
Asian           0
Pacific         0
Citizen         0
Income          1
IncomePerCap    0
Poverty         0
ChildPoverty    1
Professional    0
Service         0
Office          0
Construction    0
Production      0
Employed        0
Unemployment    0
dtype: int64
```

In [84]:
```
df_clean = df_clean.dropna()
```

In [85]:
```
df_clean.isna().sum()
```

Out[85]:
```
TotalPop         0
Men              0
Women            0
Hispanic         0
White            0
Black            0
Native           0
Asian            0
Pacific          0
Citizen          0
Income           0
IncomePerCap     0
Poverty          0
ChildPoverty     0
Professional     0
Service          0
Office           0
Construction     0
Production       0
Employed         0
Unemployment     0
dtype: int64
```

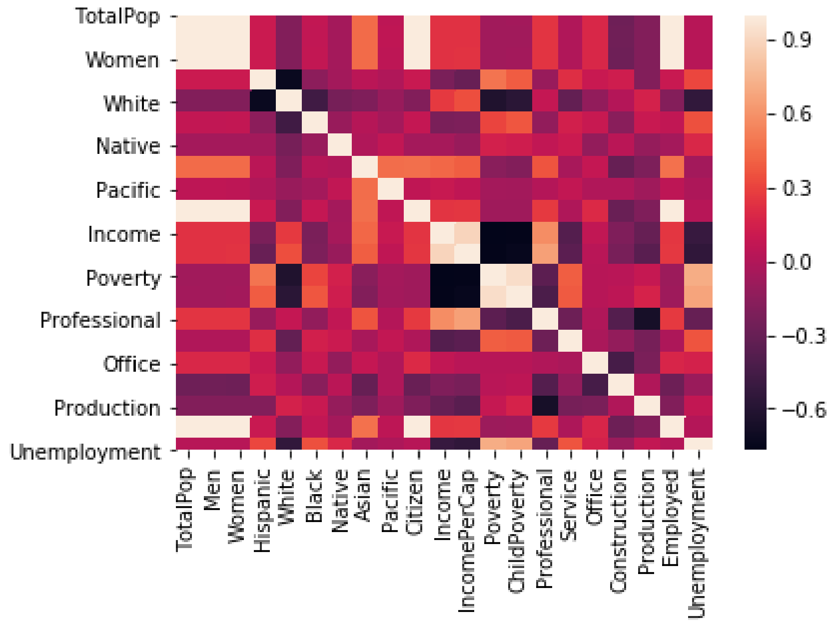
Rechecking the Correlation

In [93]:
```
df_clean.corr()['Unemployment'].sort_values(ascending=False)
```

Out[93]:
```
Unemployment     1.000000
Poverty          0.712419
ChildPoverty     0.678441
Service          0.365371
Black            0.352943
Hispanic         0.321536
Native           0.187386
Office           0.161331
Production       0.079907
Citizen          0.031346
Women            0.031068
TotalPop         0.030313
Men              0.029522
Employed         0.014002
Pacific         -0.015888
Asian           -0.055315
Construction    -0.091779
Professional    -0.300318
Income          -0.509054
White           -0.540146
IncomePerCap    -0.547239
Name: Unemployment, dtype: float64
```

In [166]:
```
sns.heatmap(df_clean.corr())
```

Out[166]: <matplotlib.axes._subplots.AxesSubplot at 0x7fa8407e46d0>



Fitting the Linear Regression Model

In [90]:
```
est = sm.OLS(y,x)
est2 = est.fit()
print(est2.summary())
```

```
                                 OLS Regression Results
=======================================================================================
================
Dep. Variable:           Unemployment   R-squared (uncentered):
0.916
Model:                            OLS   Adj. R-squared (uncentered):
0.915
Method:                 Least Squares   F-statistic:
1827.
Date:                Sun, 03 Apr 2022   Prob (F-statistic):
0.00
Time:                        09:08:45   Log-Likelihood:
-7683.9
No. Observations:                3218   AIC:
1.541e+04
Df Residuals:                    3199   BIC:
1.552e+04
Df Model:                          19
Covariance Type:            nonrobust
=======================================================================================
=========
                   coef    std err          t      P>|t|      [0.025
0.975]
---------------------------------------------------------------------------------------
---------
TotalPop     -5.749e-07   1.92e-06     -0.300      0.764   -4.33e-06
3.18e-06
Men           1.568e-05   1.04e-05      1.501      0.134    -4.8e-06
3.62e-05
Women        -1.625e-05   1.06e-05     -1.527      0.127   -3.71e-05
4.62e-06
Hispanic         0.0076      0.032      0.233      0.815      -0.056
0.071
White            0.0018      0.033      0.054      0.957      -0.062
0.066
Black            0.0431      0.033      1.320      0.187      -0.021
0.107
Native           0.0734      0.035      2.074      0.038       0.004
0.143
Asian            0.0296      0.046      0.649      0.517      -0.060
0.119
Pacific         -0.1833      0.160     -1.147      0.251      -0.496
0.130
Citizen       1.853e-05   3.11e-06      5.965      0.000    1.24e-05
2.46e-05
Income        5.082e-05   9.71e-06      5.233      0.000    3.18e-05
6.99e-05
IncomePerCap -6.065e-05   2.04e-05     -2.969      0.003      -0.000
-2.06e-05
Poverty          0.3239      0.020     16.123      0.000       0.284
0.363
ChildPoverty    -0.0149      0.012     -1.211      0.226      -0.039
0.009
Professional    -0.0761      0.035     -2.200      0.028      -0.144
-0.008
Service          0.0732      0.034      2.141      0.032       0.006
0.140
```

```
Office             0.1101      0.035      3.162      0.002       0.042
0.178
Construction      -0.0531      0.035     -1.537      0.124      -0.121
0.015
Production         0.0136      0.034      0.394      0.693      -0.054
0.081
Employed      -2.177e-05   5.14e-06     -4.233      0.000   -3.19e-05
-1.17e-05
=========================================================================
=======
Omnibus:                      298.199   Durbin-Watson:
1.684
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              1
542.600
Skew:                           0.283   Prob(JB):
0.00
Kurtosis:                       6.344   Cond. No.
1.14e+16
=========================================================================
=======

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[2] The smallest eigenvalue is 5.9e-18. This might indicate that there
are
strong multicollinearity problems or that the design matrix is singula
r.
```

We need to clean up the data some more. Some of these columns are not needed.

We will remove TotalPop & Employed

In [104]:
```
#Rerun with TotalPop & Employed removed

x1 = df_clean[['Men', 'Women', 'Hispanic', 'White', 'Black', 'Native',
       'Asian', 'Pacific', 'Citizen', 'Income', 'IncomePerCap', 'Povert
y',
       'ChildPoverty', 'Professional', 'Service', 'Office', 'Constructio
n',
       'Production']]
y = df_clean[['Unemployment']]

est3 = sm.OLS(y,x1)
est4 = est3.fit()
print(est4.summary())
```

```
                              OLS Regression Results
==============================================================================
================
Dep. Variable:           Unemployment   R-squared (uncentered):
0.915
Model:                            OLS   Adj. R-squared (uncentered):
0.915
Method:                 Least Squares   F-statistic:
1917.
Date:                Sun, 03 Apr 2022   Prob (F-statistic):
0.00
Time:                        10:18:32   Log-Likelihood:
-7692.9
No. Observations:                3218   AIC:
1.542e+04
Df Residuals:                    3200   BIC:
1.553e+04
Df Model:                          18
Covariance Type:            nonrobust
==============================================================================
=========
                   coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
---------
Men           4.095e-06   1.03e-05      0.398      0.691   -1.61e-05
2.43e-05
Women        -2.401e-05   1.11e-05     -2.169      0.030   -4.57e-05
-2.31e-06
Hispanic         0.0039      0.032      0.121      0.904      -0.060
0.068
White           -0.0028      0.033     -0.085      0.932      -0.067
0.062
Black            0.0387      0.033      1.182      0.237      -0.026
0.103
Native           0.0687      0.035      1.939      0.053      -0.001
0.138
Asian            0.0065      0.045      0.144      0.886      -0.082
0.095
Pacific         -0.1304      0.160     -0.817      0.414      -0.444
0.183
Citizen       1.693e-05   3.09e-06      5.477      0.000    1.09e-05
2.3e-05
Income        5.097e-05   9.74e-06      5.235      0.000    3.19e-05
7.01e-05
IncomePerCap -7.387e-05   2.02e-05     -3.650      0.000      -0.000
-3.42e-05
Poverty          0.3172      0.020     15.799      0.000       0.278
0.357
ChildPoverty    -0.0127      0.012     -1.032      0.302      -0.037
0.011
Professional    -0.0709      0.035     -2.047      0.041      -0.139
-0.003
Service          0.0849      0.034      2.486      0.013       0.018
0.152
Office           0.1236      0.035      3.554      0.000       0.055
0.192
```

```
Construction     -0.0461      0.035     -1.335      0.182     -0.114
0.022
Production        0.0190      0.034      0.552      0.581     -0.049
0.087
========================================================================
=======
Omnibus:                      289.920   Durbin-Watson:
1.671
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              1
491.440
Skew:                           0.267   Prob(JB):
0.00
Kurtosis:                       6.292   Cond. No.
1.16e+06
========================================================================
=======

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[2] The condition number is large, 1.16e+06. This might indicate that t
here are
strong multicollinearity or other numerical problems.
```

New Train Data Columns

In [102]:
```
x1.columns
```

Out[102]:
```
Index(['Men', 'Women', 'Hispanic', 'White', 'Black', 'Native', 'Asian',
       'Pacific', 'Citizen', 'Income', 'IncomePerCap', 'Poverty',
       'ChildPoverty', 'Professional', 'Service', 'Office', 'Constructi
on',
       'Production'],
      dtype='object')
```

In [160]:
```
x2 = df_clean[['TotalPop','Men', 'Women', 'Hispanic', 'White', 'Black',
'Native',
       'Asian', 'Pacific', 'Citizen', 'Income', 'IncomePerCap', 'Povert
y',
       'ChildPoverty', 'Professional', 'Service', 'Office', 'Constructio
n',
       'Production', 'Unemployment']]

# Changing integers to percent
x2['Men %'] = x2['Men']/x2['TotalPop']
x2['Women %'] = x2['Women']/x2['TotalPop']
x2['Citizen %'] = x2['Citizen']/x2['TotalPop']

# Race - floats to percent
x2['Hispanic'] = x2['Hispanic']/100
x2['White'] = x2['White']/100
x2['Black'] = x2['Black']/100
x2['Native'] = x2['Native']/100
x2['Pacific'] = x2['Pacific']/100

# Occupation Type - floats to percent
x2['Poverty'] = x2['Poverty']/100
x2['ChildPoverty'] = x2['ChildPoverty']/100
x2['Professional'] = x2['Professional']/100
x2['Service'] = x2['Service']/100
x2['Office'] = x2['Office']/100
x2['Construction'] = x2['Construction']/100
x2['Production'] = x2['Production']/100

# Unemployment - float to percent
x2['Unemployment'] = x2['Unemployment']/100

y2 = x2[['Unemployment']]
x2 = x2[['Men', 'Women', 'Hispanic', 'White', 'Black', 'Native',
       'Asian', 'Pacific', 'Citizen', 'Income', 'IncomePerCap', 'Povert
y',
       'ChildPoverty', 'Professional', 'Service', 'Office', 'Constructio
n',
       'Production']]
```

In [161]:
```
est5 = sm.OLS(y,x2)
est6 = est5.fit()
print(est6.summary())
```

```
                                 OLS Regression Results                                
=======================================================================================
================
Dep. Variable:           Unemployment   R-squared (uncentered):                   
0.915
Model:                            OLS   Adj. R-squared (uncentered):              
0.915
Method:                 Least Squares   F-statistic:                              
1917.
Date:                Sun, 03 Apr 2022   Prob (F-statistic):                       
0.00
Time:                        11:34:16   Log-Likelihood:                          
-7692.9
No. Observations:                3218   AIC:                                  
1.542e+04
Df Residuals:                    3200   BIC:                                  
1.553e+04
Df Model:                          18                                                  
Covariance Type:            nonrobust                                                  
=======================================================================================
=========
                   coef    std err          t      P>|t|      [0.025      
0.975]
---------------------------------------------------------------------------------------
---------
Men           4.095e-06   1.03e-05      0.398      0.691   -1.61e-05    
2.43e-05
Women        -2.401e-05   1.11e-05     -2.169      0.030   -4.57e-05   
-2.31e-06
Hispanic         0.3915      3.246      0.121      0.904      -5.972       
6.755
White           -0.2785      3.283     -0.085      0.932      -6.715       
6.158
Black            3.8713      3.275      1.182      0.237      -2.551      
10.293
Native           6.8750      3.545      1.939      0.053      -0.076      
13.826
Asian            0.0065      0.045      0.144      0.886      -0.082       
0.095
Pacific        -13.0448     15.967     -0.817      0.414     -44.351      
18.261
Citizen       1.693e-05   3.09e-06      5.477      0.000    1.09e-05     
2.3e-05
Income        5.097e-05   9.74e-06      5.235      0.000    3.19e-05    
7.01e-05
IncomePerCap -7.387e-05   2.02e-05     -3.650      0.000      -0.000   
-3.42e-05
Poverty         31.7227      2.008     15.799      0.000      27.786      
35.660
ChildPoverty    -1.2715      1.232     -1.032      0.302      -3.686       
1.143
Professional    -7.0920      3.465     -2.047      0.041     -13.886      
-0.298
Service          8.4915      3.416      2.486      0.013       1.793      
15.190
Office          12.3559      3.477      3.554      0.000       5.539      
19.172
```

```
Construction    -4.6148      3.457     -1.335      0.182     -11.394
2.164
Production       1.9036      3.448      0.552      0.581      -4.858
8.665
========================================================================
=======
Omnibus:                      289.920   Durbin-Watson:
1.671
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              1
491.440
Skew:                           0.267   Prob(JB):
0.00
Kurtosis:                       6.292   Cond. No.
1.16e+08
========================================================================
=======

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[2] The condition number is large, 1.16e+08. This might indicate that t
here are
strong multicollinearity or other numerical problems.
```

Takeaway from LR Model

We were hoping to use this model to run a test model on the US Census 2017 County data, however as shown above, the p-values indicate that some features are not as significant as we initially hypothesized.