EDA

In [1]:
```
from IPython.display import display
from IPython.display import Image
import matplotlib
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import seaborn as sns
import datetime
import pickle
%matplotlib inline
sns.set()
sns.set_context('poster')
sns.set_style("darkgrid")
```

In [2]:
```
start = datetime.datetime.time(datetime.datetime.now())
```

Preliminary EDA on 2010 Data

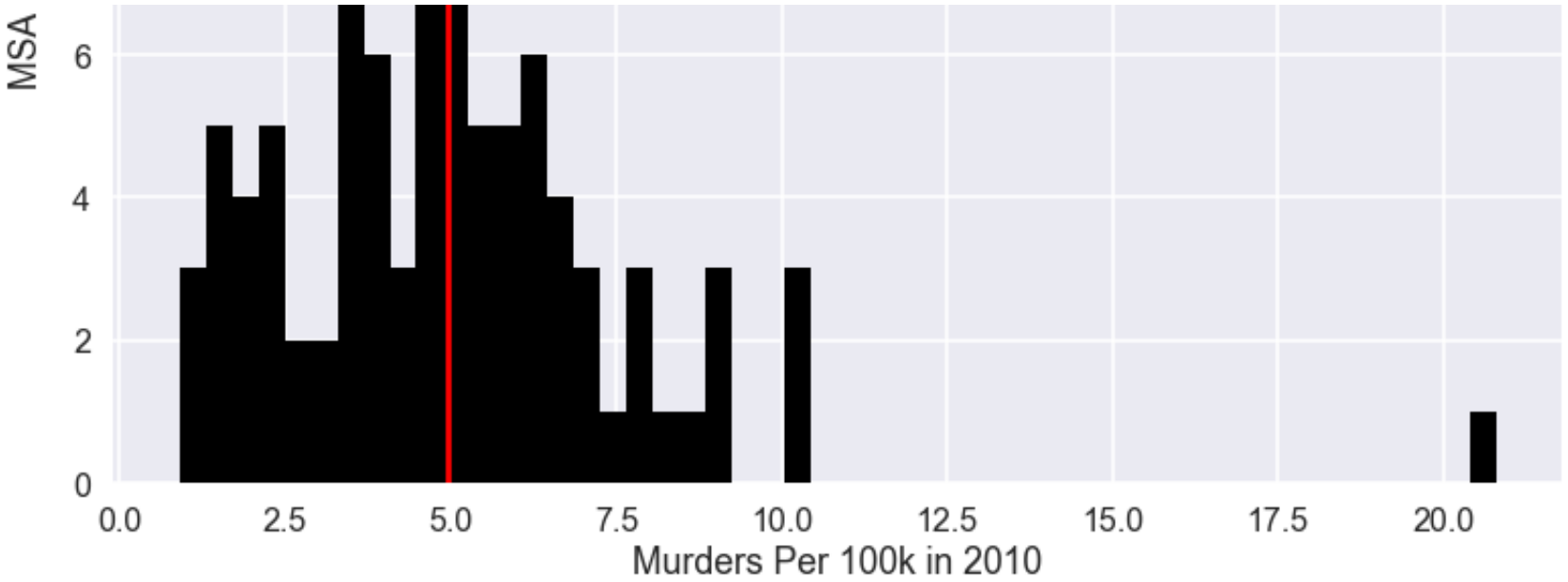
Data Import

In [3]:
```
df_2010 = pd.read_csv('../data/merged/eda_2010.csv')
df_2010.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 94 entries, 0 to 93
Columns: 191 entries, year to murder_per_100_k
dtypes: float64(173), int64(17), object(1)
memory usage: 140.3+ KB
```

Histogram

In [4]:
```
with sns.axes_style("darkgrid"):
    plt.hist(df_2010.murder_per_100_k.values, bins=50, facecolor='black', label='2010 data')
    plt.axvline(df_2010.murder_per_100_k.mean(), 0, 1, color='r', label='Mean')
    plt.xlabel("Murders Per 100k in 2010")
    plt.ylabel("MSA Count")
    plt.title("Murders Per 100k Histogram - 2010 data")
    plt.legend()
```



In [5]:
```
# drop column 145 because it contains blanks
df_2010 = df_2010.drop(
    ['poverty_married-couple_family_with_related_children_under_5_years_only'], axis=1)
```

In [6]:
```
relevant_cols = ['family_households_married-couple_family',
                 'family_household_married_couple_family_with_own_children_under_18_years',
                 'family_households_female_householder_no_husband_present',
                 'family_households_female_householder_no_husband_present_with_own_children_under_18
                 'now_married_except_separated',
                 'less_than_high_school_diploma',
                 'high_school_graduate_or_higher',
                 'unmarried_portion_of_women_15_to_50_years_who_had_a_birth_in_past_12_months',
                 'civilian_noninst_population_18_to_64_years_with_a_disability',
                 'civilian_noninst_population_65_years_and_older_with_a_disability',
                 'industry_transportation_and_warehousing_and_utilities',
                 'median_household_income_(dollars)',
                 'households_with_supplemental_security_income',
                 'households_with_food_stamp_snap_benefits',
                 'median_family_income_(dollars)',
                 'percentage_married-couple_family',
                 'percentage_female_householder_no_husband_present_family',
                 'poverty_all_families',
                 'poverty_all_families_with_related_children_under_18_years',
                 'poverty_all_families_with_related_children_under_18_years_with_related_children_ur
                 'poverty_all_people',
                 'poverty_65_years_and_over',
                 'no_telephone_service_available',
                 'house_median_value_(dollars)',
                 'murder_per_100_k']
```

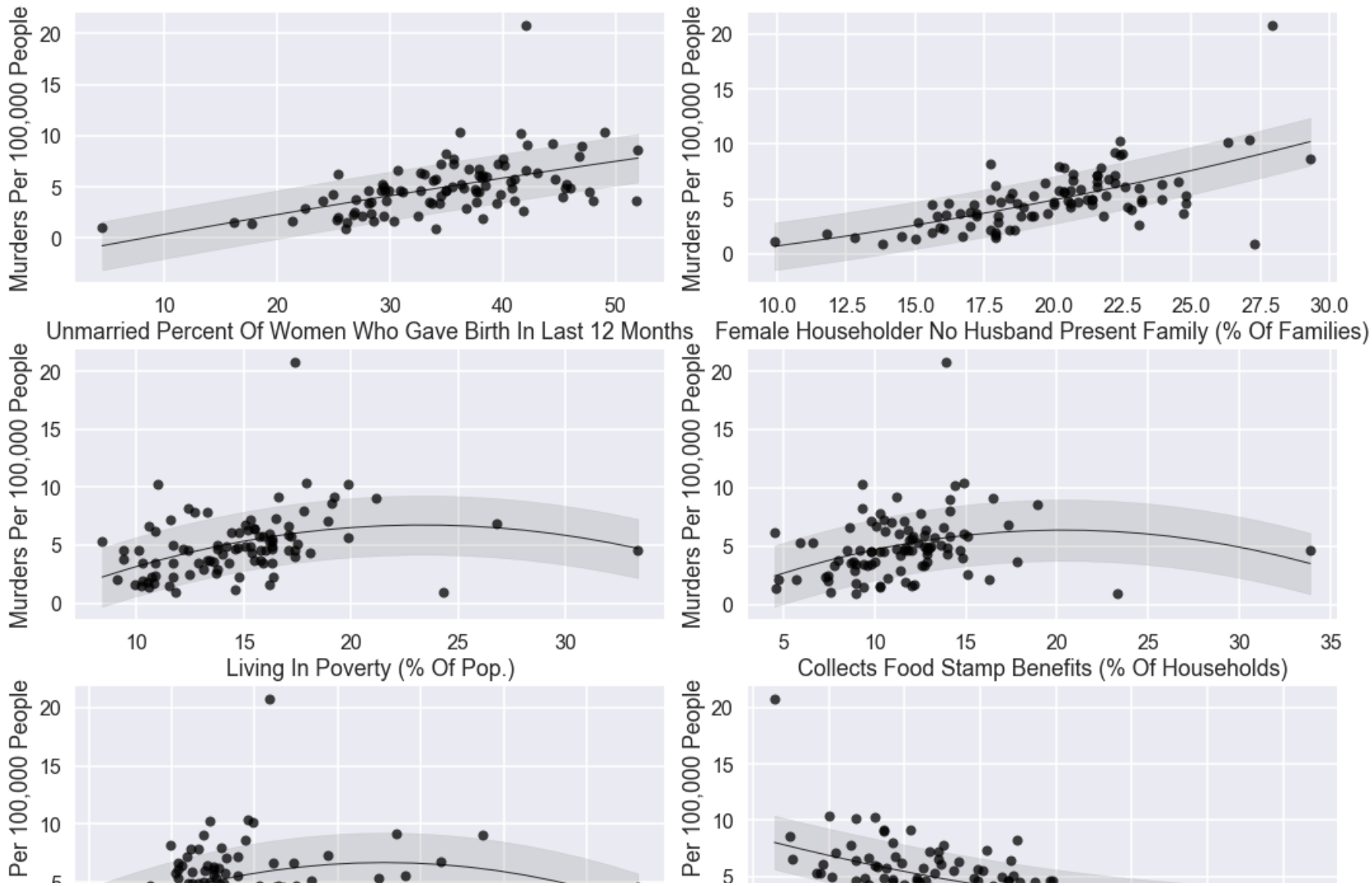
In [7]:
```
selected_cols = ['unmarried_portion_of_women_15_to_50_years_who_had_a_birth_in_past_12_months',
                 'percentage_female_householder_no_husband_present_family',
                 'poverty_all_people',
                 'households_with_food_stamp_snap_benefits',
                 'less_than_high_school_diploma',
                 'now_married_except_separated',
                 'percentage_married-couple_family',
                 'house_median_value_(dollars)',
                 'murder_per_100_k']
```

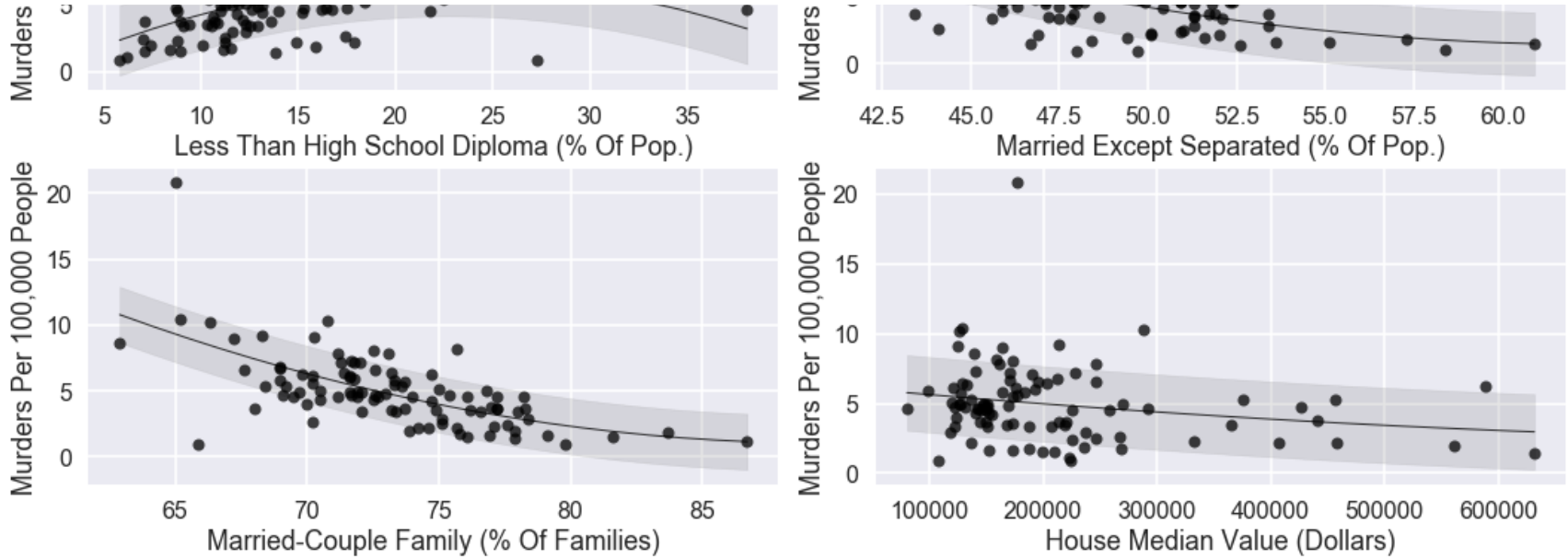
In [8]:
```
selected_col_x_vals = ['Unmarried Percent Of Women Who Gave Birth In Last 12 Months',
                       'Female Householder No Husband Present Family (% Of Families)',
                       'Living In Poverty (% Of Pop.)',
                       'Collects Food Stamp Benefits (% Of Households)',
                       'Less Than High School Diploma (% Of Pop.)',
                       'Married Except Separated (% Of Pop.)',
                       'Married-Couple Family (% Of Families)',
                       'House Median Value (Dollars)',
                       'Murders Per 100,000 People']
```

Scatter Plots

In [9]:
```
fig, ax = plt.subplots(4, 2, figsize=(15, 15))
plt.tight_layout()
x_vals = np.linspace(0, 1, 100)
x_vals = x_vals.reshape(len(x_vals),1)
ax = ax.ravel()
y = df_2010.murder_per_100_k
for i in range(0, len(selected_cols)-1):
    x = df_2010[selected_cols[i]]
    params = np.polyfit(x, y, 2)
    xp = np.linspace(x.min(), x.max(), 20)
    yp = np.polyval(params, xp)
    ax[i].plot(xp, yp, 'k', alpha=0.8, linewidth=1)
```

```
    ax[i].plot(x, y, 'o', markersize=8, alpha=0.75, color='black')
    sig = np.std(y - np.polyval(params, x))
    ax[i].fill_between(xp, yp - sig, yp + sig, color='gray', alpha=0.2)
    ax[i].set_xlabel(selected_col_x_vals[i])
    ax[i].set_ylabel(selected_col_x_vals[-1])
```





Check Very Strong Correlations

In [10]:
```
corrs = df_2010.corr().abs()
```

In [11]:
```
cols_using = ['family_households_married-couple_family',
              'family_household_married_couple_family_with_own_children_under_18_years',
              'family_households_female_householder_no_husband_present',
              'family_households_female_householder_no_husband_present_with_own_children_under_18',
              'now_married_except_separated',
              'less_than_high_school_diploma',
              'high_school_graduate_or_higher',
              'unmarried_portion_of_women_15_to_50_years_who_had_a_birth_in_past_12_months',
              'civilian_noninst_population_18_to_64_years_with_a_disability',
              'civilian_noninst_population_65_years_and_older_with_a_disability',
              'industry_transportation_and_warehousing_and_utilities',
              'median_household_income_(dollars)',
              'households_with_supplemental_security_income',
              'households_with_food_stamp_snap_benefits',
              'median_family_income_(dollars)',
              'percentage_married-couple_family',
              'percentage_female_householder_no_husband_present_family',
              'poverty_all_families',
              'poverty_all_families_with_related_children_under_18_years',
              'poverty_all_families_with_related_children_under_18_years_with_related_children_under
              'poverty_all_people',
              'poverty_65_years_and_over',
              'no_telephone_service_available',
              'house_median_value_(dollars)',
              'murder_per_100_k'
             ]
```

In [12]:
```
corrs_used = corrs[cols_using]
corrs_used = corrs_used.rename(columns={'unmarried_portion_of_women_15_to_50_years_who_had_a_birth_i
                                        'unmarried_women_who_had_a_birth_in_past_12_months'})
```

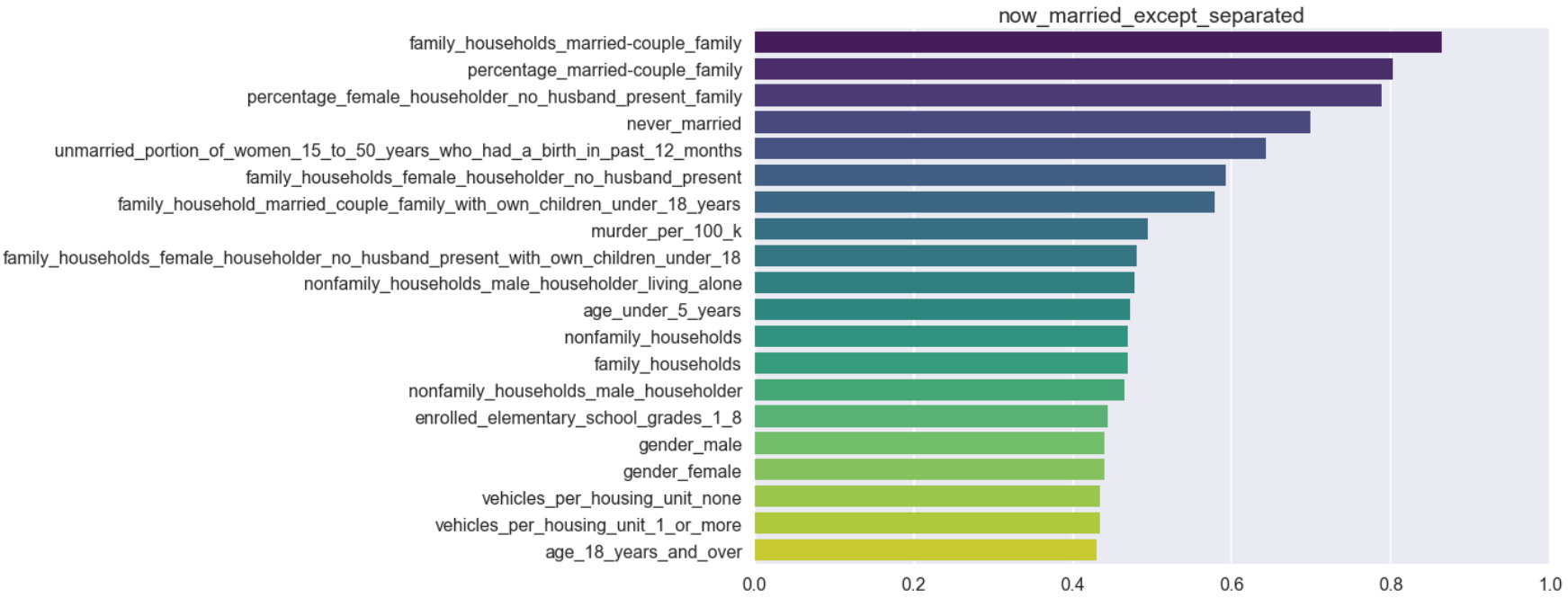
In [13]:
```
cols_important = ['now_married_except_separated',
                  'less_than_high_school_diploma',
                  'unmarried_women_who_had_a_birth_in_past_12_months',
                  'households_with_food_stamp_snap_benefits',
                  'percentage_married-couple_family',
                  'percentage_female_householder_no_husband_present_family',
                  'poverty_all_people',
                  'house_median_value_(dollars)',
                  'murder_per_100_k'
]
```

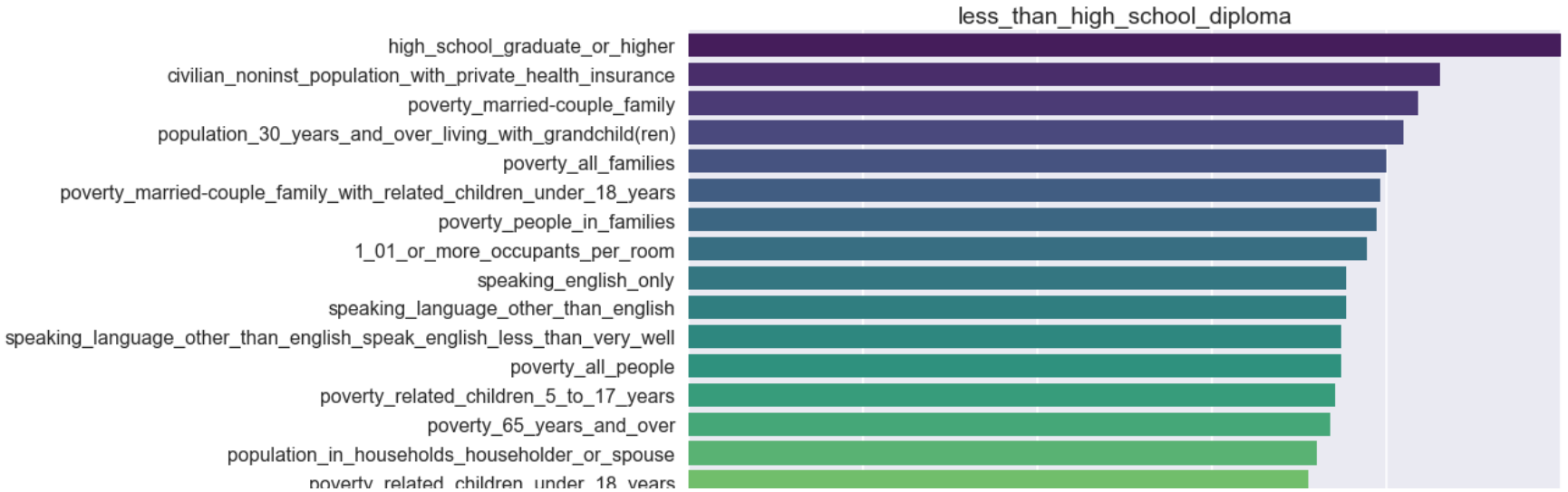
In [14]:
```
feat = dict()

top_n = 21

for n,i in corrs_used.iteritems():
    i_sorted = i.sort_values(ascending=False)
    topkeys = i_sorted.keys()[:top_n]
    top_cors = [(k,i[k]) for k in topkeys if n != k]
    feat[n] = top_cors
```

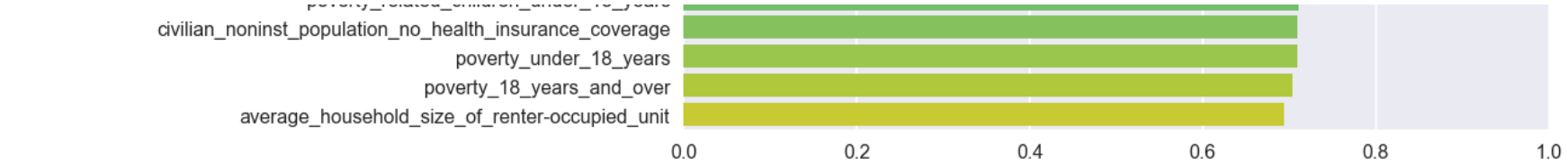
In [15]:
```
n_plots = len(cols_important)

for i,f in enumerate(cols_important):
    items = feat[f]
    labels = list(zip(*items))[0]
    scores = list(zip(*items))[1]

    if True:
        plt.figure(i)

    sns.barplot(scores, labels, palette='viridis')
    plt.xlim(0.,1.0)
    plt.title(f)
```

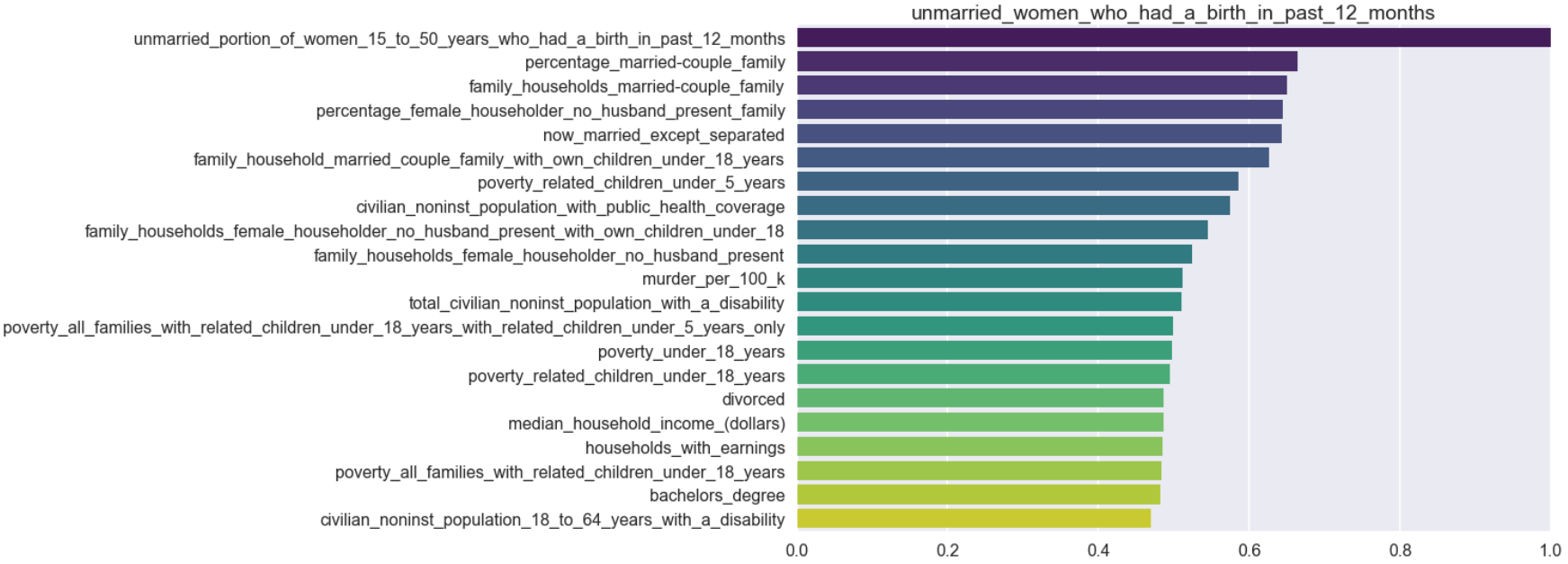
```
/Users/ilanaiden/anaconda/lib/python3.6/site-packages/seaborn/categorical.py:1428: FutureWarning:
```

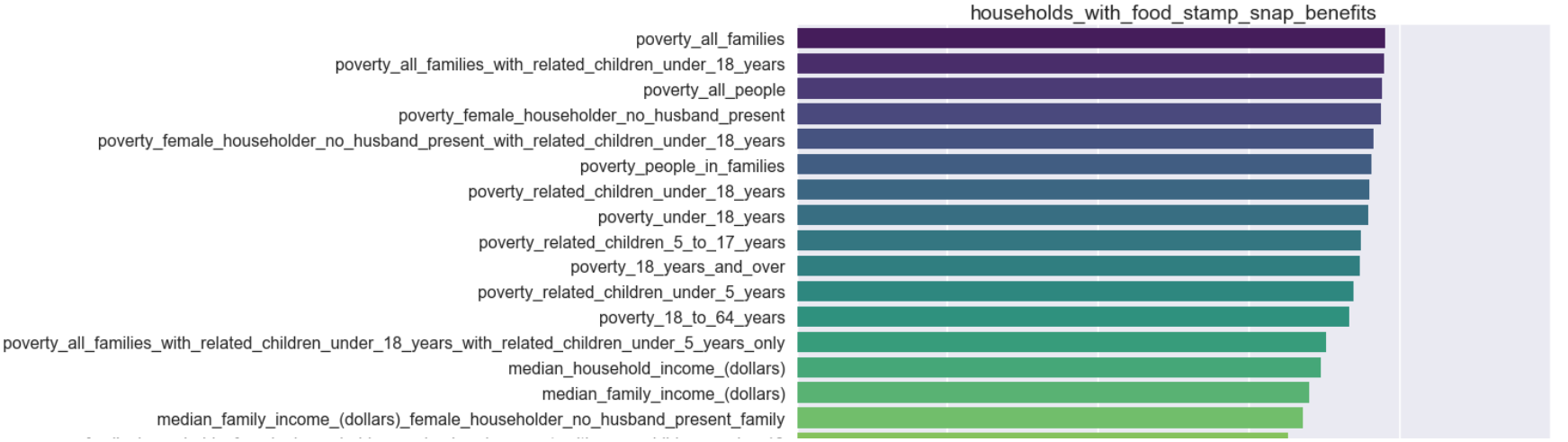
```
/Users/tianjdor/anaconda/lib/python3.6/site-packages/seaborn/categorical.py:1428: FutureWarning:
remove_na is deprecated and is a private function. Do not use.
  stat_data = remove_na(group_data)
```

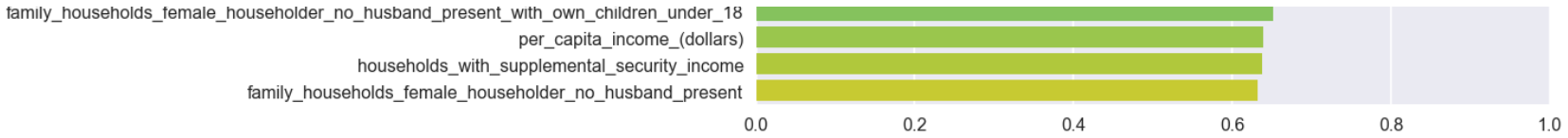


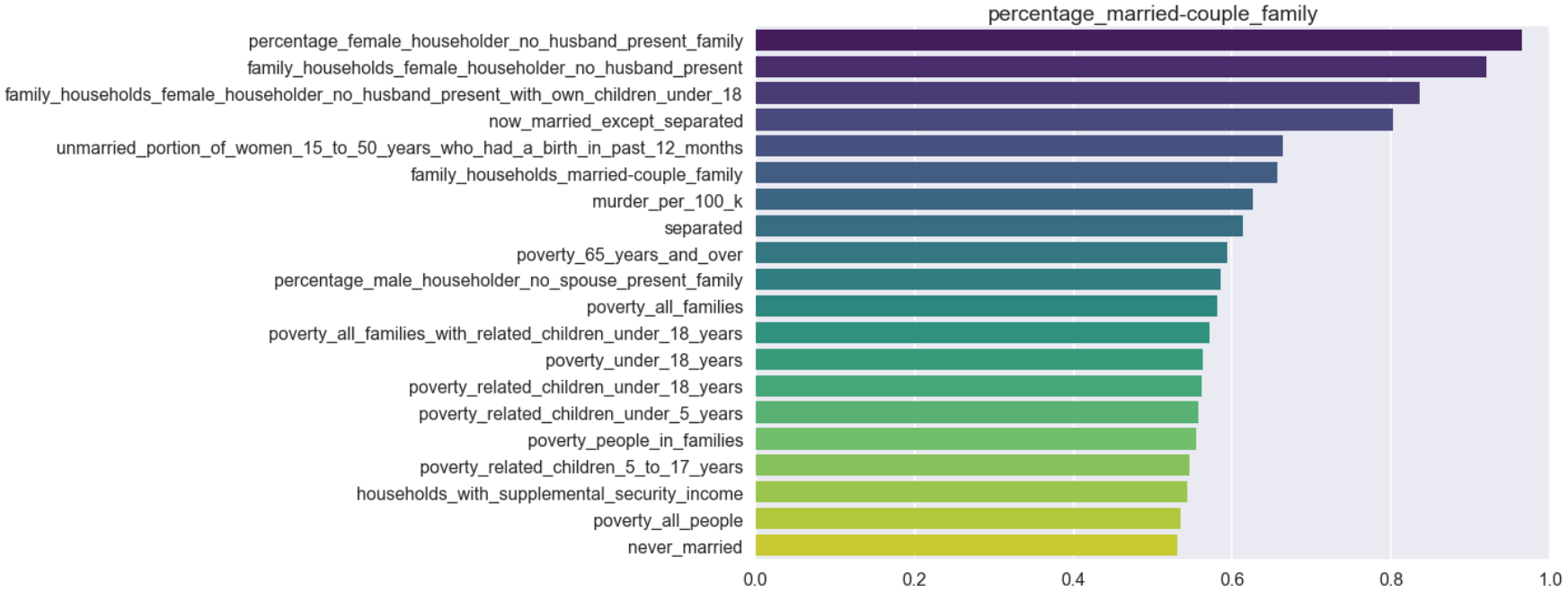


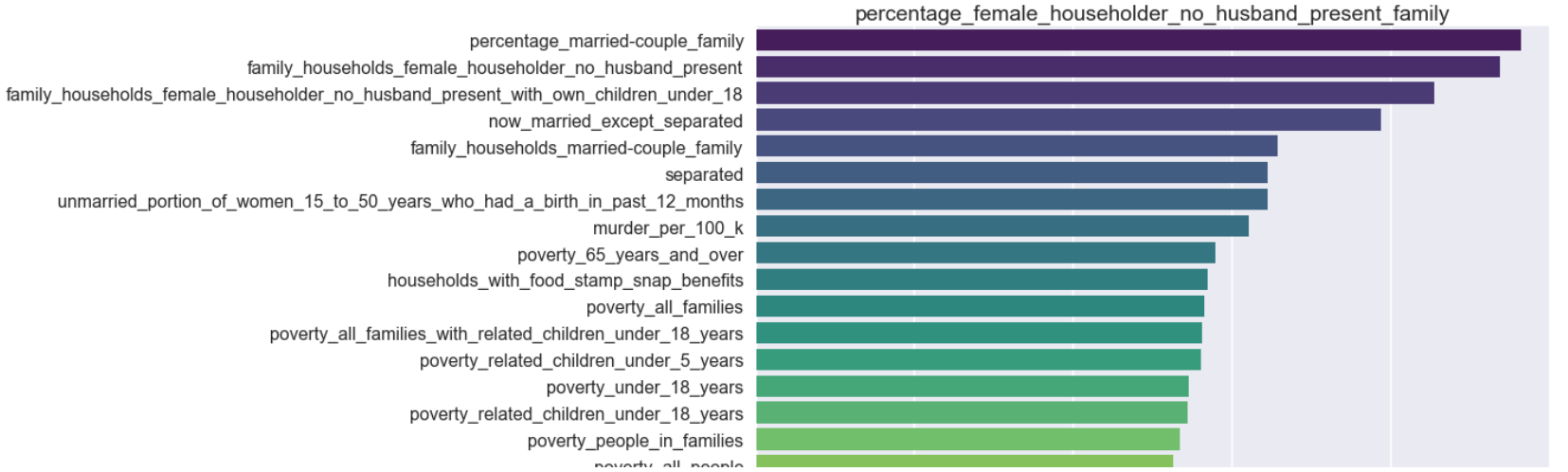

civilian_noninst_population_no_health_insurance_coverage
poverty_under_18_years
poverty_18_years_and_over
average_household_size_of_renter-occupied_unit
0.0
0.2
0.4
0.6
0.8
1.0

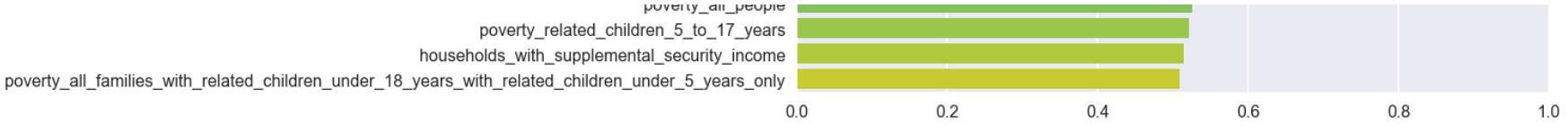

unmarried_women_who_had_a_birth_in_past_12_months
unmarried_portion_of_women_15_to_50_years_who_had_a_birth_in_past_12_months
percentage_married-couple_family
family_households_married-couple_family
percentage_female_householder_no_husband_present_family
now_married_except_separated
family_household_married_couple_family_with_own_children_under_18_years
poverty_related_children_under_5_years
civilian_noninst_population_with_public_health_coverage
family_households_female_householder_no_husband_present_with_own_children_under_18
family_households_female_householder_no_husband_present
murder_per_100_k
total_civilian_noninst_population_with_a_disability
poverty_all_families_with_related_children_under_18_years_with_related_children_under_5_years_only
poverty_under_18_years
poverty_related_children_under_18_years
divorced
median_household_income_(dollars)
households_with_earnings
poverty_all_families_with_related_children_under_18_years
bachelors_degree
civilian_noninst_population_18_to_64_years_with_a_disability
0.0
0.2
0.4
0.6
0.8
1.0

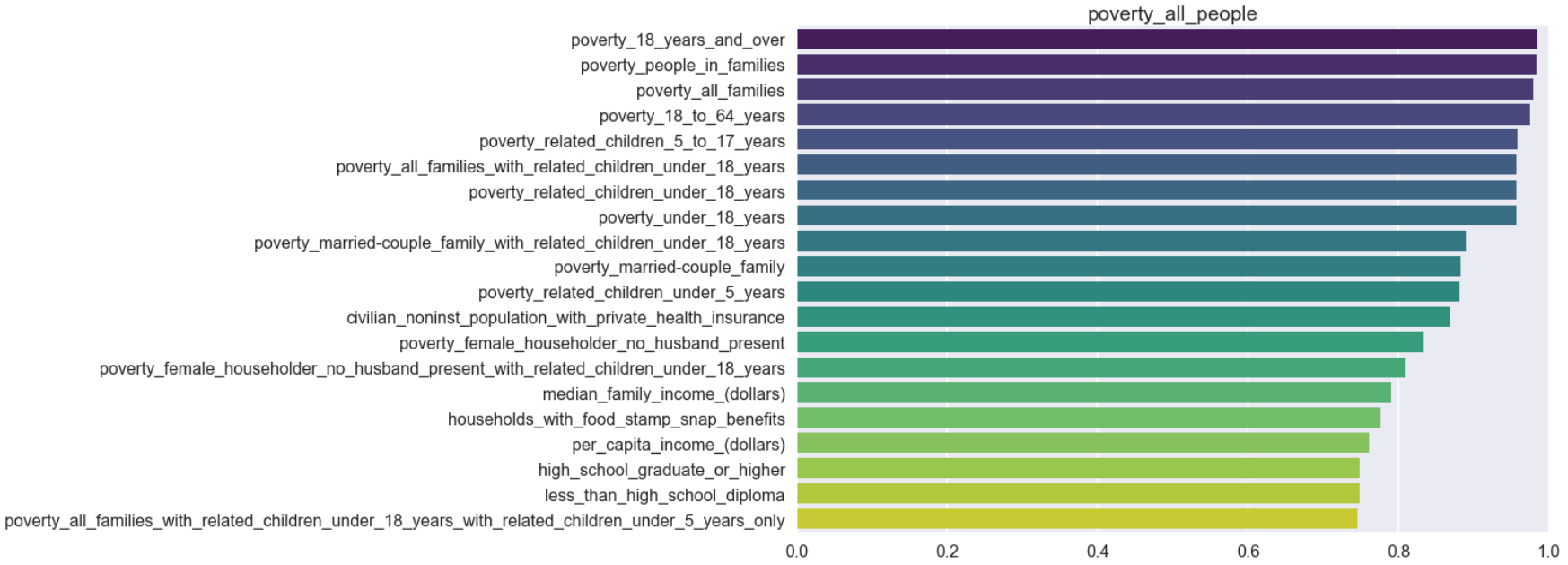

households_with_food_stamp_snap_benefits
poverty_all_families
poverty_all_families_with_related_children_under_18_years
poverty_all_people
poverty_female_householder_no_husband_present
poverty_female_householder_no_husband_present_with_related_children_under_18_years
poverty_people_in_families
poverty_related_children_under_18_years
poverty_under_18_years
poverty_related_children_5_to_17_years
poverty_18_years_and_over
poverty_related_children_under_5_years
poverty_18_to_64_years
poverty_all_families_with_related_children_under_18_years_with_related_children_under_5_years_only
median_household_income_(dollars)
median_family_income_(dollars)
median_family_income_(dollars)_female_householder_no_husband_present_family

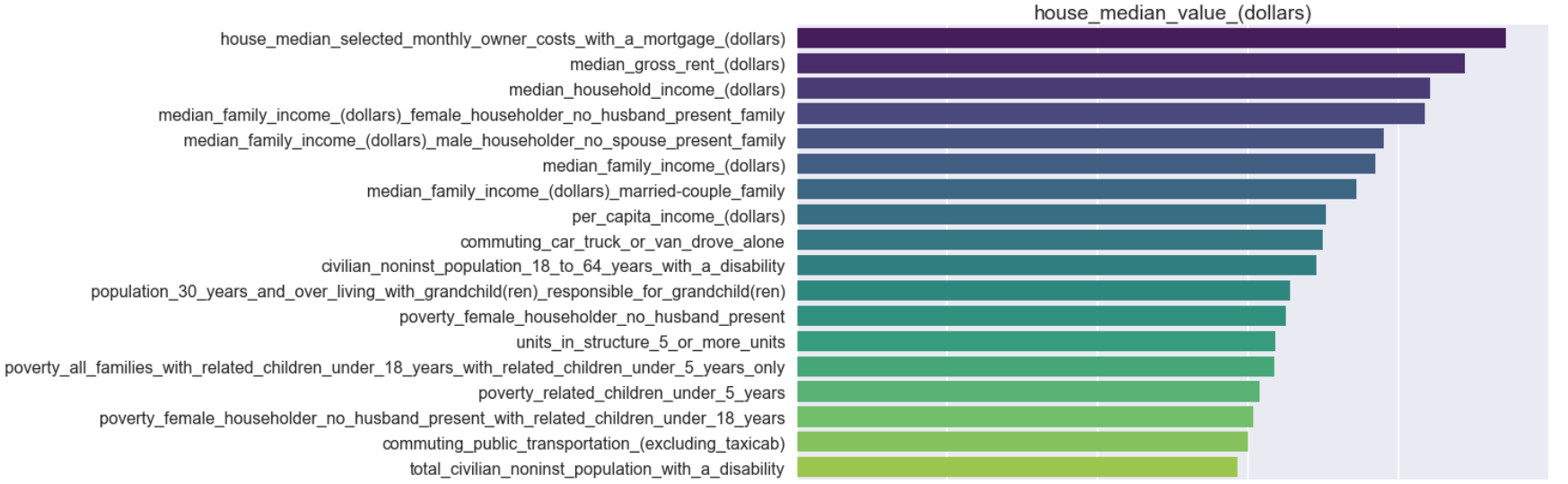

family_households_female_householder_no_husband_present_with_own_children_under_18
per_capita_income_(dollars)
households_with_supplemental_security_income
family_households_female_householder_no_husband_present
0.0
0.2
0.4
0.6
0.8
1.0

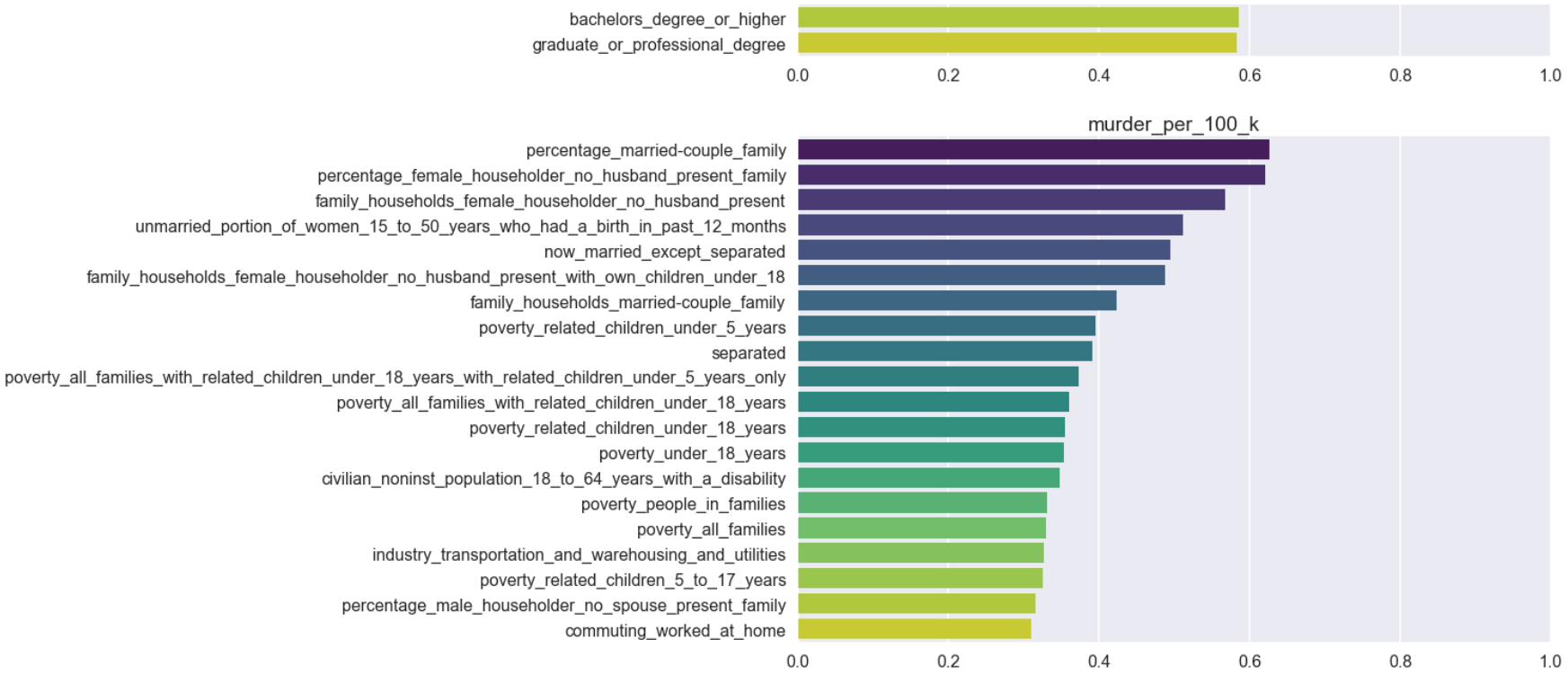

percentage_married-couple_family
percentage_female_householder_no_husband_present_family
family_households_female_householder_no_husband_present
family_households_female_householder_no_husband_present_with_own_children_under_18
now_married_except_separated
unmarried_portion_of_women_15_to_50_years_who_had_a_birth_in_past_12_months
family_households_married-couple_family
murder_per_100_k
separated
poverty_65_years_and_over
percentage_male_householder_no_spouse_present_family
poverty_all_families
poverty_all_families_with_related_children_under_18_years
poverty_under_18_years
poverty_related_children_under_18_years
poverty_related_children_under_5_years
poverty_people_in_families
poverty_related_children_5_to_17_years
households_with_supplemental_security_income
poverty_all_people
never_married
0.0
0.2
0.4
0.6
0.8
1.0


percentage_female_householder_no_husband_present_family
percentage_married-couple_family
family_households_female_householder_no_husband_present
family_households_female_householder_no_husband_present_with_own_children_under_18
now_married_except_separated
family_households_married-couple_family
separated
unmarried_portion_of_women_15_to_50_years_who_had_a_birth_in_past_12_months
murder_per_100_k
poverty_65_years_and_over
households_with_food_stamp_snap_benefits
poverty_all_families
poverty_all_families_with_related_children_under_18_years
poverty_related_children_under_5_years
poverty_under_18_years
poverty_related_children_under_18_years
poverty_people_in_families


poverty_related_children_5_to_17_years
households_with_supplemental_security_income
poverty_all_families_with_related_children_under_18_years_with_related_children_under_5_years_only
0.0
0.2
0.4
0.6
0.8
1.0

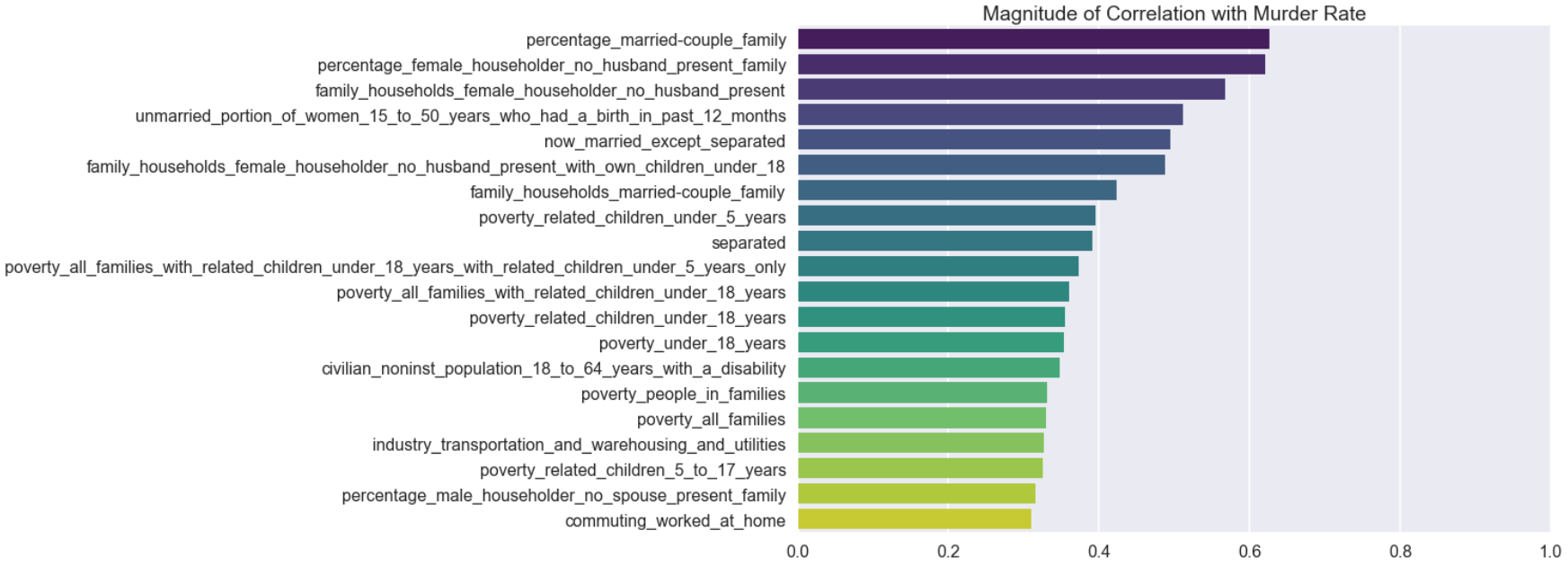

poverty_all_people
poverty_18_years_and_over
poverty_people_in_families
poverty_all_families
poverty_18_to_64_years
poverty_related_children_5_to_17_years
poverty_all_families_with_related_children_under_18_years
poverty_related_children_under_18_years
poverty_under_18_years
poverty_married-couple_family_with_related_children_under_18_years
poverty_married-couple_family
poverty_related_children_under_5_years
civilian_noninst_population_with_private_health_insurance
poverty_female_householder_no_husband_present
poverty_female_householder_no_husband_present_with_related_children_under_18_years
median_family_income_(dollars)
households_with_food_stamp_snap_benefits
per_capita_income_(dollars)
high_school_graduate_or_higher
less_than_high_school_diploma
poverty_all_families_with_related_children_under_18_years_with_related_children_under_5_years_only
0.0
0.2
0.4
0.6
0.8
1.0


house_median_value_(dollars)
house_median_selected_monthly_owner_costs_with_a_mortgage_(dollars)
median_gross_rent_(dollars)
median_household_income_(dollars)
median_family_income_(dollars)_female_householder_no_husband_present_family
median_family_income_(dollars)_male_householder_no_spouse_present_family
median_family_income_(dollars)
median_family_income_(dollars)_married-couple_family
per_capita_income_(dollars)
commuting_car_truck_or_van_drove_alone
civilian_noninst_population_18_to_64_years_with_a_disability
population_30_years_and_over_living_with_grandchild(ren)_responsible_for_grandchild(ren)
poverty_female_householder_no_husband_present
units_in_structure_5_or_more_units
poverty_all_families_with_related_children_under_18_years_with_related_children_under_5_years_only
poverty_related_children_under_5_years
poverty_female_householder_no_husband_present_with_related_children_under_18_years
commuting_public_transportation_(excluding_taxicab)
total_civilian_noninst_population_with_a_disability


bachelors_degree_or_higher
graduate_or_professional_degree
0.0
0.2
0.4
0.6
0.8
1.0
murder_per_100_k
percentage_married-couple_family
percentage_female_householder_no_husband_present_family
family_households_female_householder_no_husband_present
unmarried_portion_of_women_15_to_50_years_who_had_a_birth_in_past_12_months
now_married_except_separated
family_households_female_householder_no_husband_present_with_own_children_under_18
family_households_married-couple_family
poverty_related_children_under_5_years
separated
poverty_all_families_with_related_children_under_18_years_with_related_children_under_5_years_only
poverty_all_families_with_related_children_under_18_years
poverty_related_children_under_18_years
poverty_under_18_years
civilian_noninst_population_18_to_64_years_with_a_disability
poverty_people_in_families
poverty_all_families
industry_transportation_and_warehousing_and_utilities
poverty_related_children_5_to_17_years
percentage_male_householder_no_spouse_present_family
commuting_worked_at_home
0.0
0.2
0.4
0.6
0.8
1.0

In [16]:
```
items = feat['murder_per_100_k']
labels = list(zip(*items))[0]
scores = list(zip(*items))[1]

sns.barplot(scores, labels, palette='viridis')
plt.xlim(0.,1.0)
plt.title('Magnitude of Correlation with Murder Rate');
```

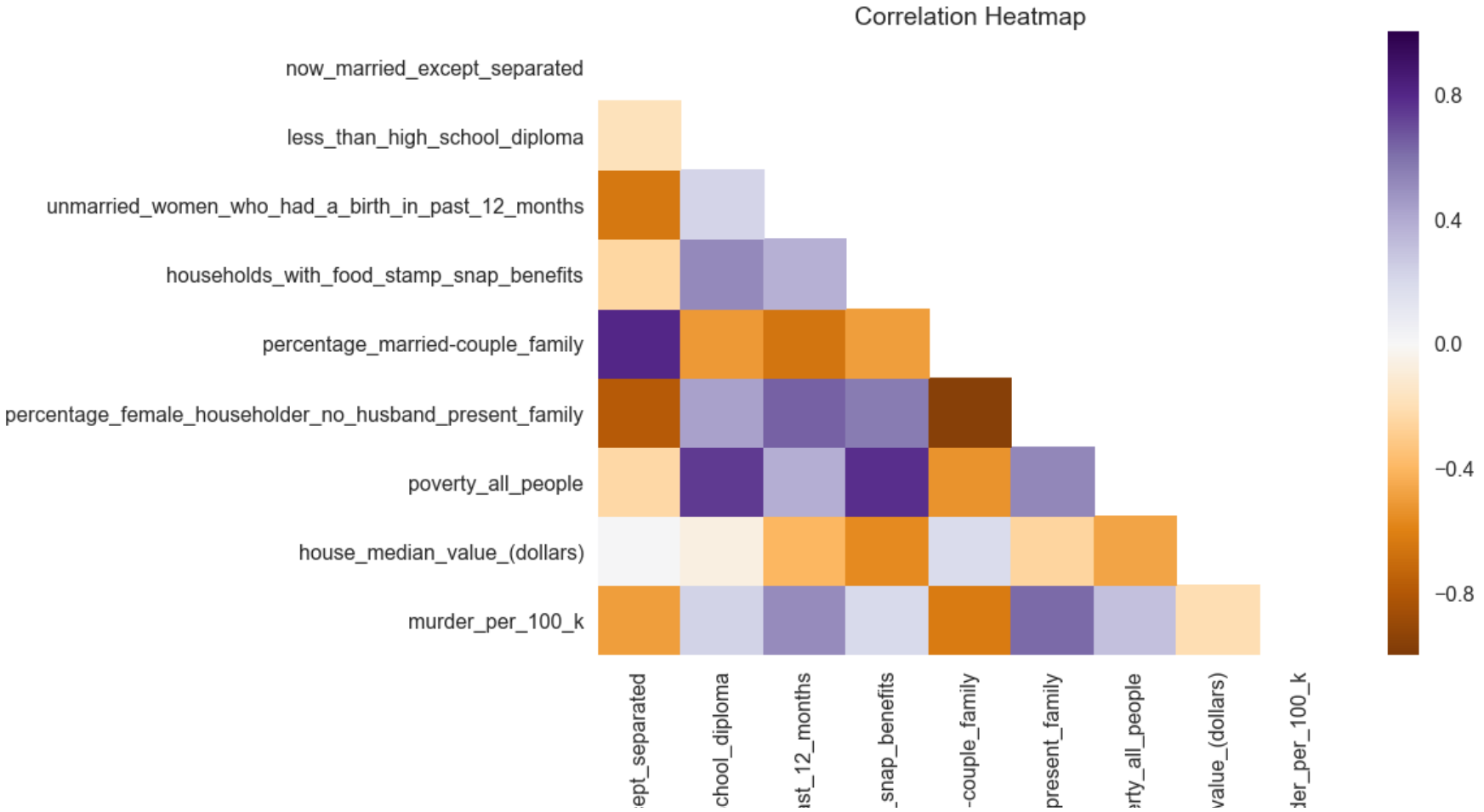
```
/Users/ilanjdor/anaconda/lib/python3.6/site-packages/seaborn/categorical.py:1428: FutureWarning:
remove_na is deprecated and is a private function. Do not use.
  stat_data = remove_na(group_data)
```

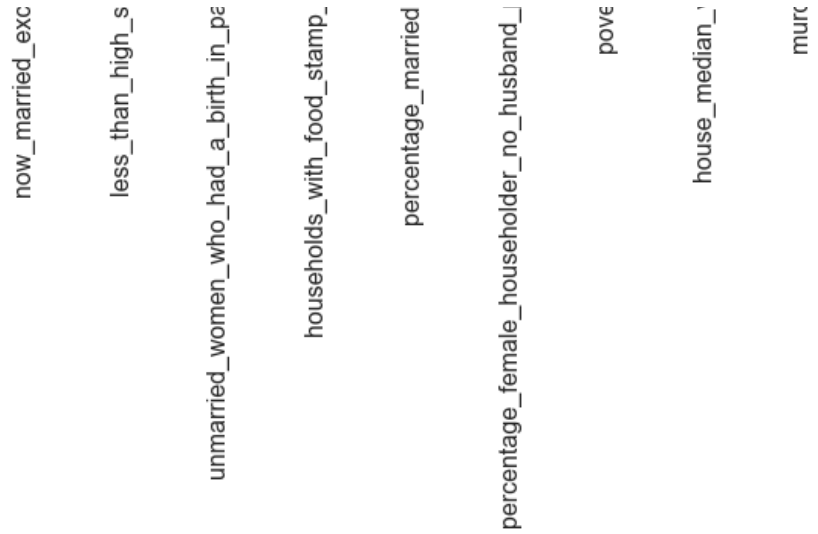


Correlation Heatmap

In [17]:
```
df_used = df_2010.rename(columns={'unmarried_portion_of_women_15_to_50_years_who_had_a_birth_in_past
                                  'unmarried_women_who_had_a_birth_in_past_12_months'})[cols_important]
```

In [18]:
```
mask = np.zeros_like(df_used.corr())
mask[np.triu_indices_from(mask)] = True
with sns.axes_style("white"):
    sns.heatmap(df_used.corr(), cmap='PuOr', mask=mask)
    plt.title('Correlation Heatmap');
```





EDA on All Data

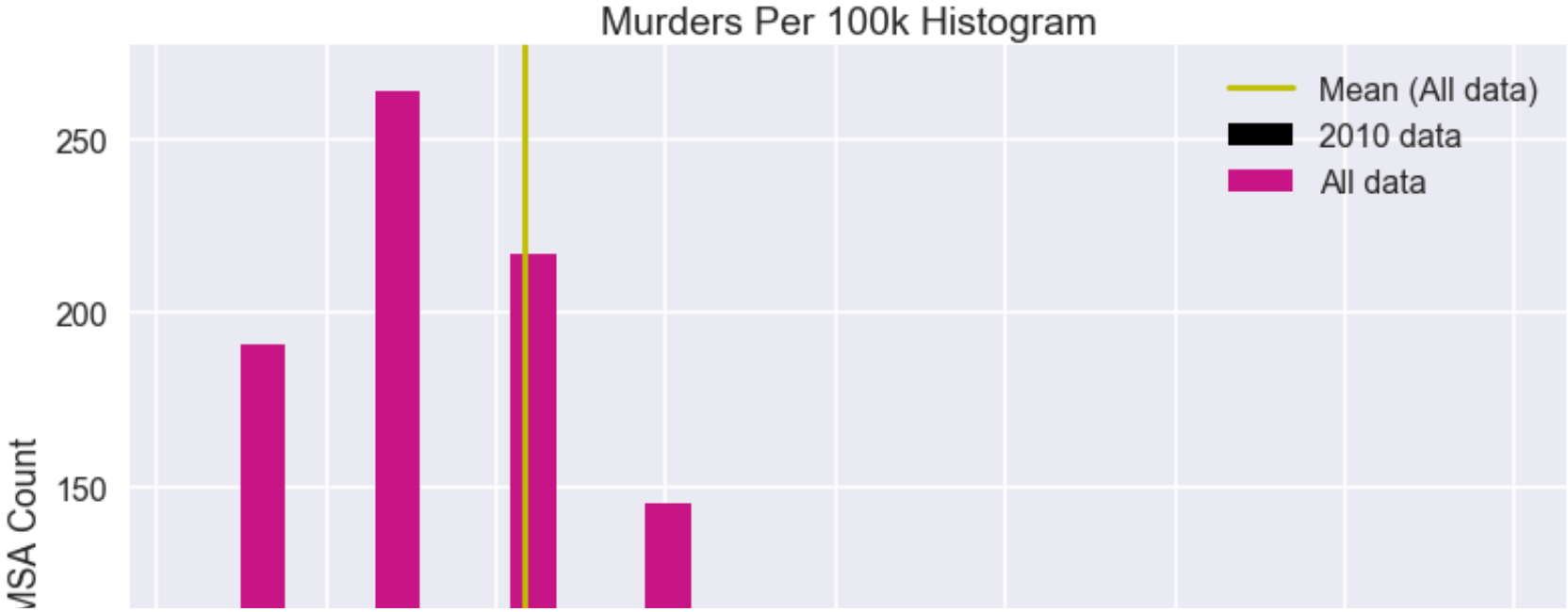
Data Import

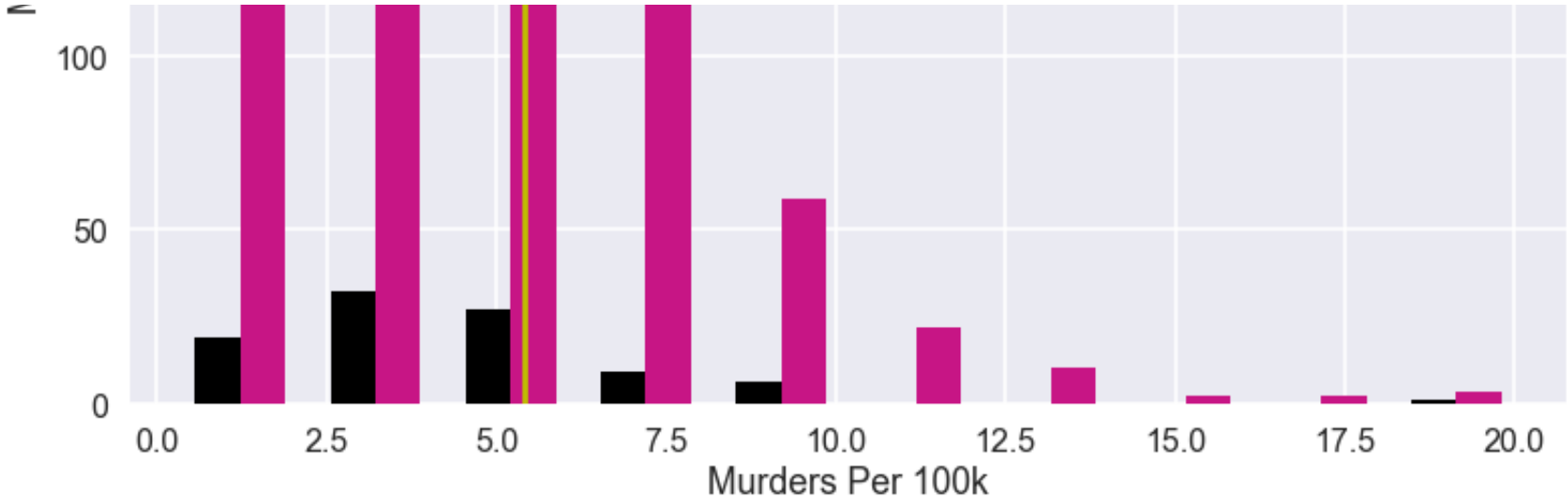
In [19]:
```
df_all = pd.read_csv('../data/merged/all_data_2006_to_2016.csv')
df_all.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 928 entries, 0 to 927
Data columns (total 13 columns):
MSA_orig                                                                928 non-null objec
t
MSA_corr                                                                928 non-null objec
t
MSA_abbr                                                                928 non-null objec
t
year                                                                    928 non-null int64
now_married_except_separated                                            928 non-null float
64
less_than_high_school_diploma                                           928 non-null float
64
unmarried_portion_of_women_15_to_50_years_who_had_a_birth_in_past_12_months    928 non-null float
64
households_with_food_stamp_snap_benefits                                928 non-null float
64
percentage_married-couple_family                                        928 non-null float
64
percentage_female_householder_no_husband_present_family                 928 non-null float
64
poverty_all_people                                                      928 non-null float
64
house_median_value_(dollars)                                            928 non-null int64
murder_per_100_k                                                        928 non-null float
64
dtypes: float64(8), int64(2), object(3)
memory usage: 94.3+ KB
```

Histogram

In [20]:
```
with sns.axes_style("darkgrid"):
    np.random.seed(0)
    fig, ax = plt.subplots()

    a_heights, a_bins = np.histogram(df_2010['murder_per_100_k'])
    b_heights, b_bins = np.histogram(df_all['murder_per_100_k'], bins=a_bins)

    width = (a_bins[1] - a_bins[0])/3

    ax.bar(a_bins[:-1], a_heights, width=width, facecolor='black',
           label='2010 data')
    ax.bar(b_bins[:-1]+width, b_heights, width=width, facecolor='mediumvioletred',
           label='All data')
    plt.axvline(df_all.murder_per_100_k.mean(), 0, 1, color='y', label='Mean (All data)')
    plt.xlabel("Murders Per 100k")
    plt.ylabel("MSA Count")
    plt.title("Murders Per 100k Histogram")
    plt.legend()
```

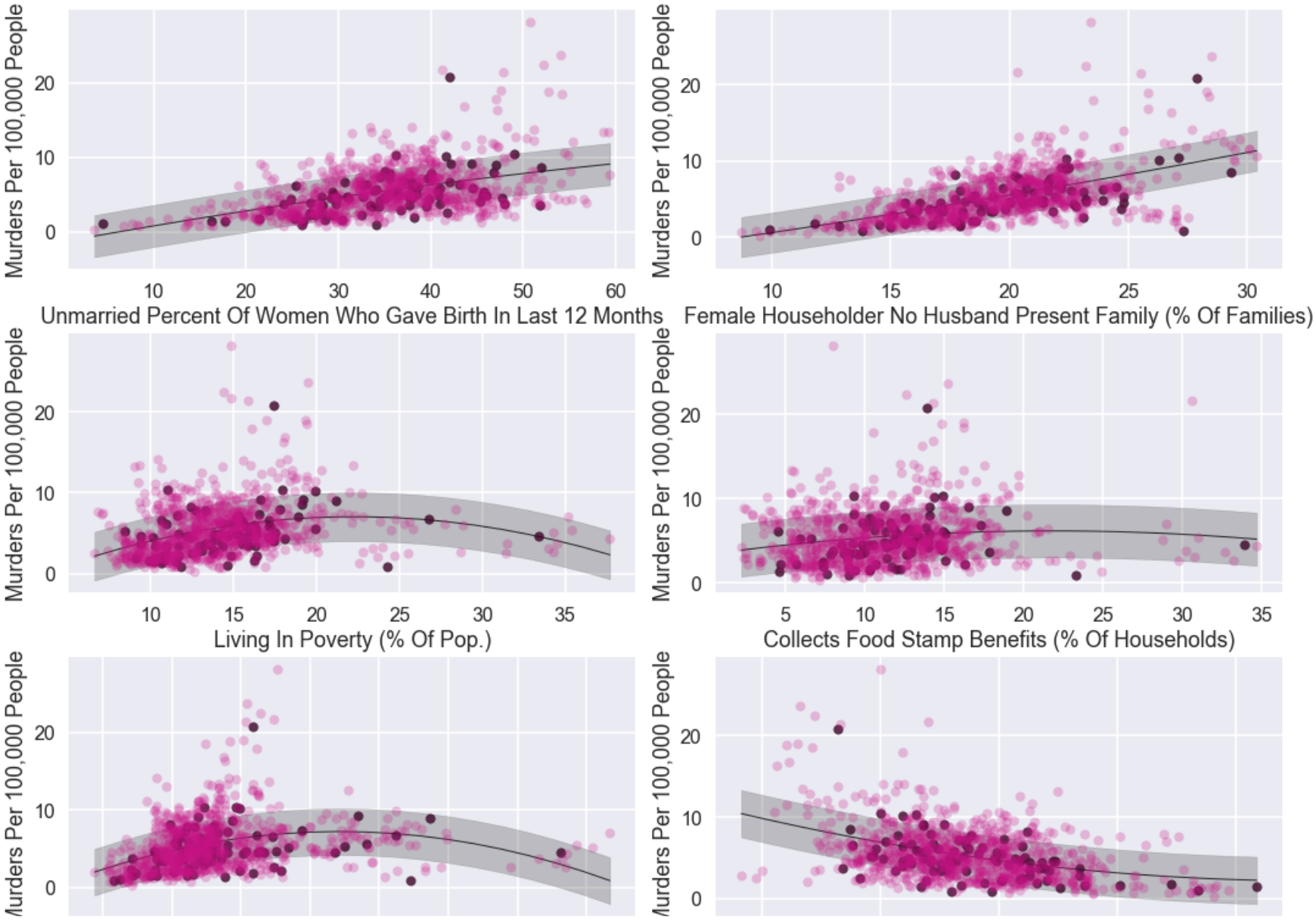


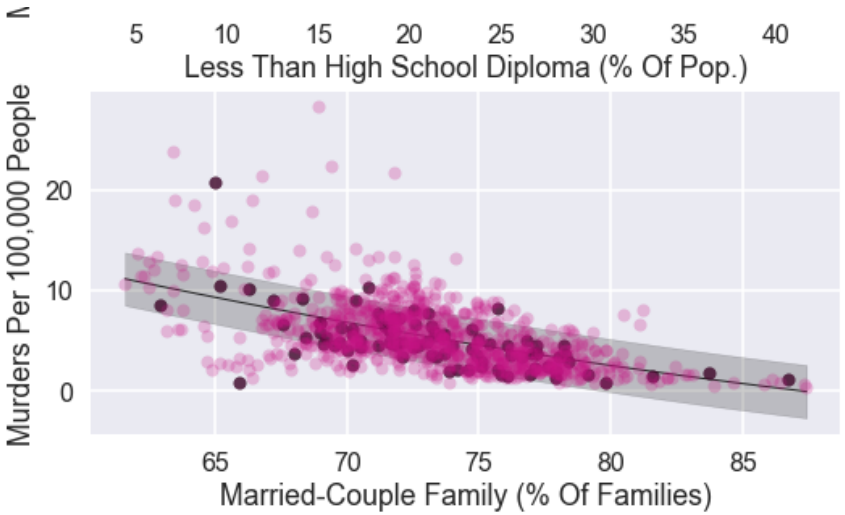


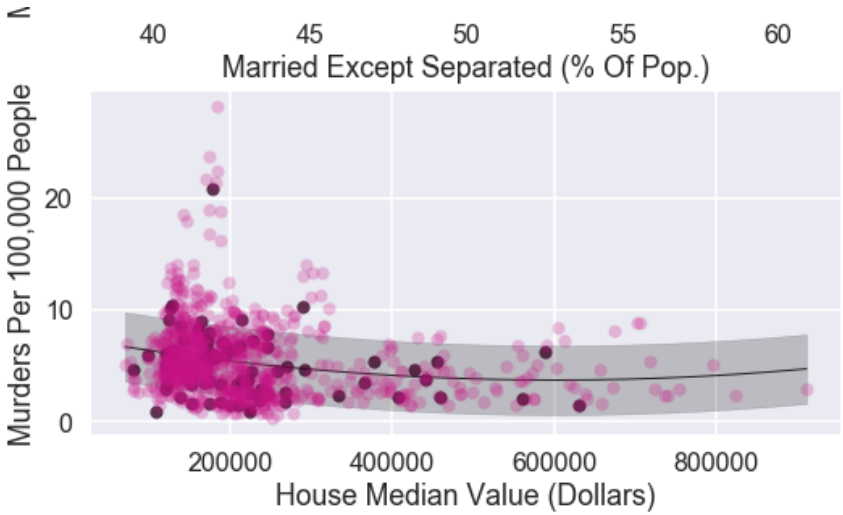
Scatter Plots

In [21]:
```
fig, ax = plt.subplots(4, 2, figsize=(15, 15))
plt.tight_layout()
x_vals = np.linspace(0, 1, 100)
x_vals = x_vals.reshape(len(x_vals),1)
ax = ax.ravel()
y_2010 = df_2010.murder_per_100_k
y_all = df_all.murder_per_100_k
for i in range(0, len(selected_cols)-1):
    x_2010 = df_2010[selected_cols[i]]
    x_all = df_all[selected_cols[i]]
    params = np.polyfit(x_all, y_all, 2)
    xp_all = np.linspace(x_all.min(), x_all.max(), 20)
    yp_all = np.polyval(params, xp_all)
    ax[i].plot(x_2010, y_2010, 'o', markersize=8, alpha=0.75, color='black')
    ax[i].plot(xp_all, yp_all, 'k', alpha=0.8, linewidth=1)
    ax[i].plot(x_all, y_all, 'o', markersize=8, alpha=0.25, color='mediumvioletred')
    sig = np.std(y_all - np.polyval(params, x_all))
```

```
    ax[i].fill_between(xp_all, yp_all - sig, yp_all + sig, color='k', alpha=0.2)
    ax[i].set_xlabel(selected_col_x_vals[i])
    ax[i].set_ylabel(selected_col_x_vals[-1])
```




Less Than High School Diploma (% Of Pop.)
Murders Per 100,000 People
Married-Couple Family (% Of Families)


Married Except Separated (% Of Pop.)
Murders Per 100,000 People
House Median Value (Dollars)

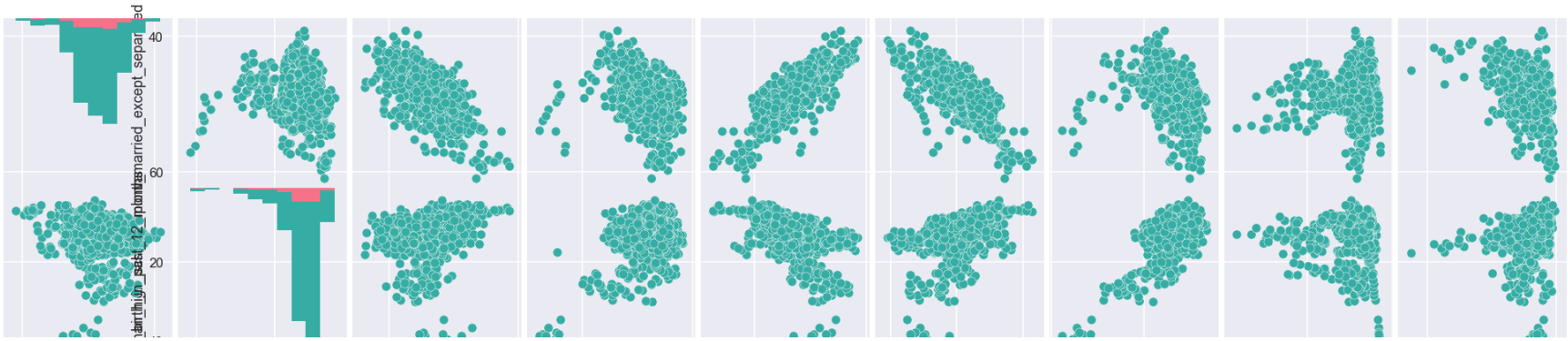
Pair plots

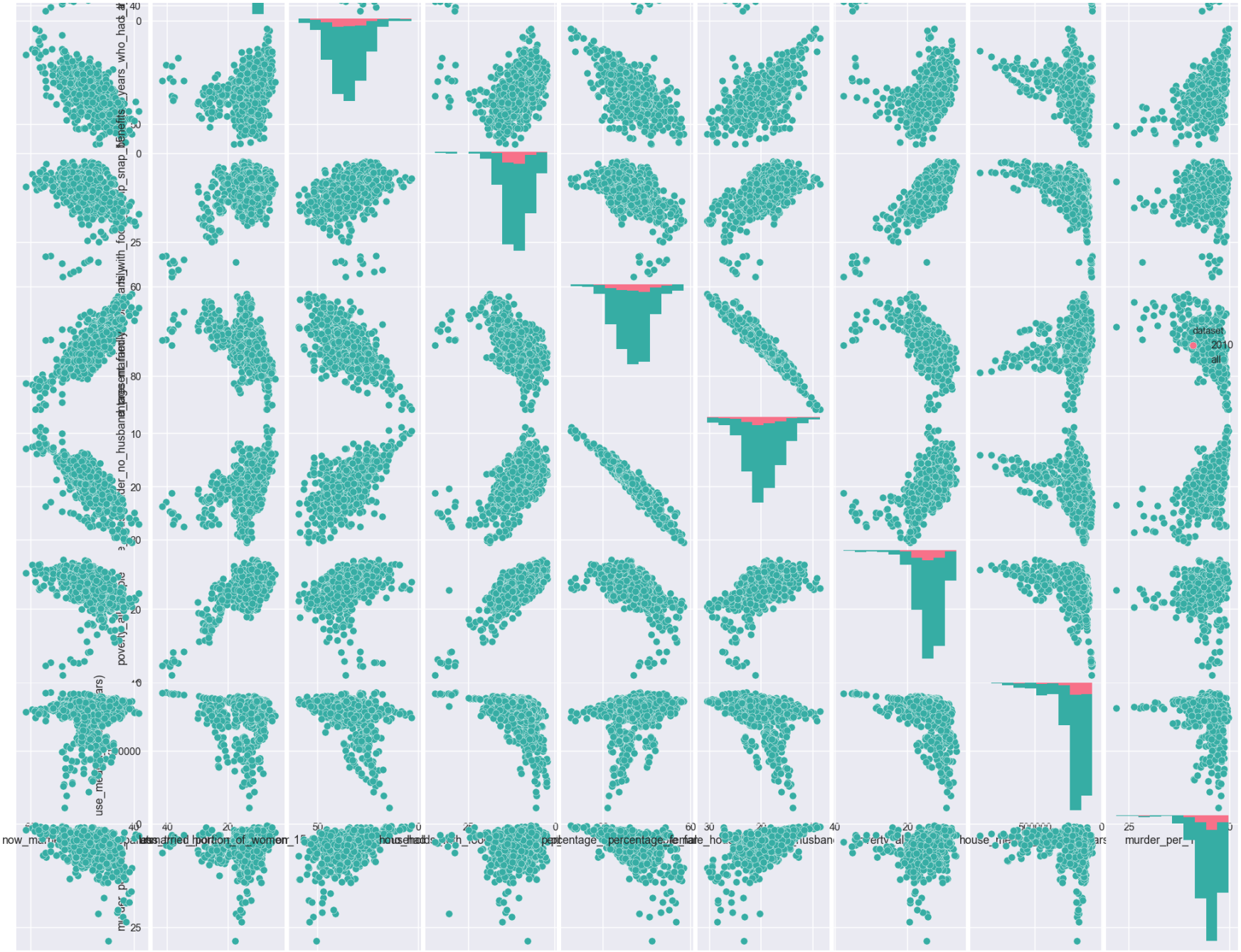
In [22]:
```
df_all_2 = df_all.drop(['MSA_orig', 'MSA_corr', 'MSA_abbr', 'year'], axis=1).copy()
df_2010_2 = df_2010[list(df_all_2)].copy()
df_all_2['dataset'] = 'all'
df_2010_2['dataset'] = '2010'
df_both = pd.DataFrame(np.vstack((df_2010_2, df_all_2)))
df_both.columns = list(df_all_2)
cols = [df_both.columns[-1]] + [col for col in df_both if col != df_both.columns[-1]]
df_both = df_both[cols]
df_both.head()
```

Out[22]:

	dataset	now_married_except_separated	less_than_high_school_diploma	unmarried_portion_of_women_15_to_50_years_who_had_a_birt
0	2010	47.9	10.7	
1	2010	46.7	8.9	
2	2010	46.9	13.2	
3	2010	51.3	12.3	
4	2010	48.7	12.5	

In [23]:
```
# we chose not to include on the website due to the illegible labels
sns.pairplot(df_both, hue="dataset", palette="husl");
```




dataset
2010
all

In [24]:
```
def print_runtime():
    hours = int(str(end)[0:2])-int(str(start)[0:2])
    minutes = int(str(end)[3:5])-int(str(start)[3:5])
    seconds = int(str(end)[6:8])-int(str(start)[6:8])
    if hours < 0:
        hours = hours + 24
    if minutes < 0:
        minutes = minutes + 60
        hours = hours - 1
    if seconds < 0:
        seconds = seconds + 60
        minutes = minutes - 1
    print(hours, "hrs", minutes, "mins", seconds, "secs")
```

In [25]:
```
end = datetime.datetime.time(datetime.datetime.now())
```

In [26]:
```
print_runtime()
```

```
0 hrs 0 mins 31 secs
```