ProjectEDA

March 31, 2024

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

[ ]:
```
data = pd.read_csv('C:/Users/ldmag/Documents/GitHub/Code-Assignments-Projects/
↪Assignments/STAT-501/data/USDataF23.txt', delimiter='\t')
data
```

[ ]:
```
          State Region  CountyIndex  UrbanIndicator  Population  LandArea  \
0     Illinois     NC            1               1       30987       7.0
1     Illinois     NC            2               1       11663       5.4
2     Illinois     NC            3               1       37093      12.4
3     Illinois     NC            4               1       36427       8.3
4     Illinois     NC            5               1       99581      34.7
..         …    …           …               …          …       …
217  Wisconsin     NC           29               0       10227       4.5
218  Wisconsin     NC           30               0       10993       6.0
219  Wisconsin     NC           31               0       15333       7.3
220  Wisconsin     NC           32               0       12388       6.5
221  Wisconsin     NC           33               1       18659       4.9

     PopulationDensity  PercentMaleDivorce  PercentFemaleDivorce  \
0          4426.714286                6.27                  9.16
1          2159.814815                5.53                  6.77
2          2991.370968                4.91                  7.87
3          4388.795181                5.62                  9.42
4          2869.769452                9.34                 12.13
..                 …                   …                     …
217        2272.666667                8.30                 10.05
218        1832.166667                7.56                  9.94
219        2100.410959                7.61                 11.62
220        1905.846154                7.96                  9.74
221        3807.959184                9.96                 11.29

     MedianIncome  … PercentCollegeGraduates  MedianHouseAge  \
0           48758  …                   19.88              42
```

```
1           53665  …                  32.89              37
2           59020  …                  46.21              54
3           56011  …                  47.44              38
4           35039  …                  18.61              57
..            …  …                    …                …
217         28892  …                  21.38              65
218         34061  …                  21.05              56
219         35301  …                  20.89              47
220         26873  …                  14.12              72
221         30031  …                   9.13              63

     RobberiesPerPopulation  AssaultsPerPopulation  BurglariesPerPopulation  \
0                     27.15                 117.64                   464.53
1                     15.21                  53.23                   349.81
2                     10.61                  53.04                   379.21
3                      4.98                  29.88                   174.32
4                    284.93                 467.69                  1373.09
..                      …                      …                        …
217                   17.71                  62.00                   557.97
218                   42.12                 126.36                   303.26
219                    0.00                  51.29                   393.21
220                   43.16                  38.36                   556.27
221                   65.18                 130.36                   570.31

     LarceniesPerPopulation  EducationSpending  EducationSpendingP2  \
0                   2310.57        12076.22650          12832.06795
1                   1528.52        11997.01572          12735.38179
2                   1943.78        12195.07184          12831.67355
3                   1324.80        12308.89898          12876.42623
4                   3685.31        12094.73855          12934.55714
..                      …                …                    …
217                 4268.89        11304.23505          11292.86114
218                 2299.72        11321.88194          10910.38999
219                 3766.81        11660.43733          11286.79777
220                 3155.42        11711.86064          11181.82714
221                 2579.98        11478.17277          11123.46465

       TestScore      RegionNew
0    1675.070862  North Central
1    1656.257029  North Central
2    1684.396377  North Central
3    1633.947073  North Central
4    1679.750452  North Central
..           …            …
217  1621.844589  North Central
218  1649.067925  North Central
219  1656.648874  North Central
```

```
220  1576.426195  North Central
221  1662.661341  North Central

[222 rows x 21 columns]
```

0.1 Summary statistics

[ ]:
```
data.describe()
```

[ ]:
```
       CountyIndex  UrbanIndicator    Population    LandArea  \
count   222.000000      222.000000  2.220000e+02  222.000000
mean     19.531532        0.716216  5.253160e+04   18.736486
std      13.497658        0.451852  2.039370e+05   24.849015
min       1.000000        0.000000  1.009200e+04    1.700000
25%       8.250000        0.000000  1.413600e+04    6.325000
50%      17.000000        1.000000  2.119750e+04   10.900000
75%      28.000000        1.000000  3.848875e+04   24.275000
max      52.000000        1.000000  2.783726e+06  233.000000

       PopulationDensity  PercentMaleDivorce  PercentFemaleDivorce  \
count         222.000000          222.000000            222.000000
mean         2602.892625            8.927072             11.896982
std          1699.106251            3.018184              3.418288
min            95.887354            2.870000              4.270000
25%          1500.149623            6.635000              9.387500
50%          2269.774193            8.870000             11.785000
75%          3226.623218           10.842500             14.307500
max         11947.321890           17.100000             21.910000

       MedianIncome  PercentCollegeGraduates  MedianHouseAge  \
count    222.000000               222.000000      222.000000
mean   33227.626126                21.429234       55.864865
std    11940.961621                13.664479       10.236556
min     8866.000000                 3.270000       32.000000
25%    23980.500000                11.772500       48.000000
50%    31260.000000                16.795000       55.500000
75%    40287.000000                28.460000       63.750000
max    84441.000000                71.230000       78.000000

       RobberiesPerPopulation  AssaultsPerPopulation  BurglariesPerPopulation  \
count              222.000000             222.000000               222.000000
mean               125.928423             316.205225               814.568649
std                215.840499             518.977192               614.266040
min                  0.000000               0.000000                31.010000
25%                 18.977500              65.247500               446.255000
50%                 45.595000             129.975000               627.005000
75%                121.532500             317.652500               999.970000
```

```
max                1362.780000            4932.500000              5000.380000

       LarceniesPerPopulation  EducationSpending  EducationSpendingP2  \
count              222.000000         222.000000           222.000000
mean              3314.132973       11199.911335         11122.205776
std               1502.916253         663.568663           916.388825
min                241.100000        9559.323299          8756.816518
25%               2270.690000       11034.959765         10919.704300
50%               3129.360000       11302.703355         11235.329440
75%               4277.342500       11526.462162         11478.408875
max               9888.570000       12768.877970         13608.895640

         TestScore
count   222.000000
mean   1620.471957
std      45.109578
min    1478.095408
25%    1592.523616
50%    1623.900065
75%    1651.747836
max    1754.276047
```

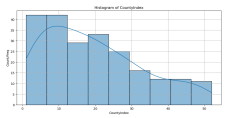
[ ]:
```
# no missing data
data.isnull().sum()
```

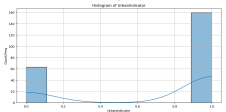
[ ]:
```
State                     0
Region                    0
CountyIndex               0
UrbanIndicator            0
Population                0
LandArea                  0
PopulationDensity         0
PercentMaleDivorce        0
PercentFemaleDivorce      0
MedianIncome              0
IncomeCategory            0
PercentCollegeGraduates   0
MedianHouseAge            0
RobberiesPerPopulation    0
AssaultsPerPopulation     0
BurglariesPerPopulation   0
LarceniesPerPopulation    0
EducationSpending         0
EducationSpendingP2       0
TestScore                 0
RegionNew                 0
dtype: int64
```

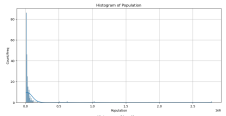
[ ]:
```
def plot_histogram(data):
    with warnings.catch_warnings():
        warnings.simplefilter('ignore', FutureWarning)
        cols = data.select_dtypes(include='number').columns
        colcount = len(cols)
        fig, axes = plt.subplots(colcount, 1, figsize=(10,5*colcount))

        for i, c in enumerate(cols):
            ax = axes[i]
            sns.histplot(data=data, x=c, kde=True, ax=ax)
            ax.set_title(f'Histogram of {c}')
            ax.set_xlabel(c)
            ax.set_ylabel('Count/Freq')
            ax.grid(True)

        plt.tight_layout()
        plt.show()

def plot_boxplots(data):
    with warnings.catch_warnings():
        warnings.simplefilter('ignore', FutureWarning)
        cols = data.select_dtypes(include='number').columns
        colcount = len(cols)
        fig, axes = plt.subplots(colcount, 1, figsize=(10,5*colcount))

        for i, c in enumerate(cols):
            ax = axes[i]
            sns.boxplot(data=data, y=c, ax=ax)
            ax.set_title(f'Boxplot of {c}')
            ax.set_ylabel(c)
            ax.grid(True)

        plt.tight_layout()
        plt.show()
```

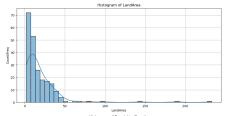
0.1.1 Histogram interpretation

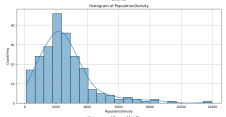
[ ]:
```
plot_histogram(data)
```

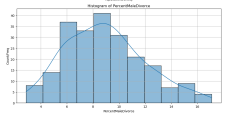

Histogram of CountyIndex
Count/Freq
CountyIndex

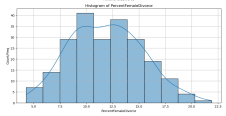

Histogram of UrbanIndicator
Count/Freq
UrbanIndicator

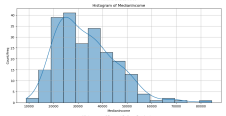

Histogram of Population
Count/Freq
Population

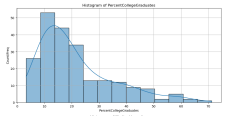

Histogram of LandArea
Count/Freq
LandArea

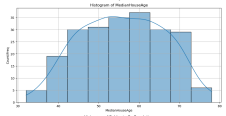

Histogram of PopulationDensity
Count/Freq
PopulationDensity

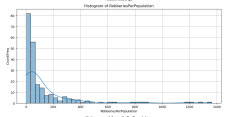

Histogram of PercentMaleDivorce
Count/Freq
PercentMaleDivorce

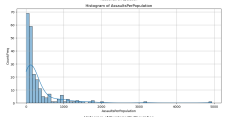

Histogram of PercentFemaleDivorce
Count/Freq
PercentFemaleDivorce

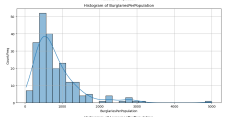

Histogram of MedianIncome
Count/Freq
MedianIncome


Histogram of PercentCollegeGraduates
Count/Freq
PercentCollegeGraduates

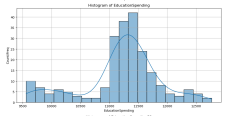

Histogram of MedianHouseAge
Count/Freq
MedianHouseAge

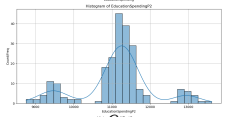

Histogram of NativeAmPerPopulation
Count/Freq
NativeAmPerPopulation

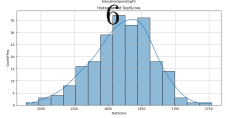

Histogram of AsianAmPerPopulation
Count/Freq
AsianAmPerPopulation


Histogram of HispanicsPerPopulation
Count/Freq
HispanicsPerPopulation


Histogram of LaCorneasPerPopulation
Count/Freq
LaCorneasPerPopulation


Histogram of EducationSpending
Count/Freq
EducationSpending


Histogram of EducationSpendingPC
Count/Freq
EducationSpendingPC


Histogram of TestScore
Count/Freq
TestScore

0.1.2 Histograms:

Ignore CountyIndex - these are basically continuous values that have no meaning statistically. Could use for interpretation. Ignore UrbanIndicator - these are categorical and only have a range of 0 - 1. Good for inference however.

Population: Heavy right skew, non-normal.

Land Area Heavy right skew, non-normal. Potential outliers present.

Population Density Heavy right skew, non-normal. Potential outliers present. Expect correlation with population and Land Area.

Male Divorce Non-normal, slight right skew. Bimodal.

Female Divorce Non-normal, slight right skew. Bimodal. Expect correlation between Male and Female divorce rates and with Income.

Median Income Non-normal, right skewed. Likely bimodal.

College Graduate Percentage Heavy right skew. Non-normal

Median House Age Bimodal. Slight left skew (?).

Robberies per 100k Heavy right skew. Expect outliers.

Assaults per 100k Heavy right skew. Expect outliers.

Burglaries per 100k Right skewed, non-normal. Expect outliers.

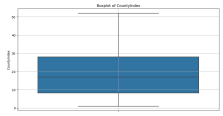
Larcenies per 100k Right skweed, asymmetric.

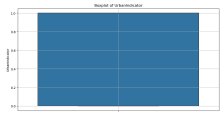
Education Spending Considered approximately normal, but appears asymmetric.

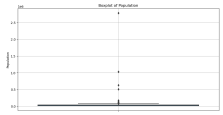
Education Spending 2 Considered approximately normal, but asymmetric - note the gaps in distribution, means there is an element of seasonality. Two sample test for this might work.

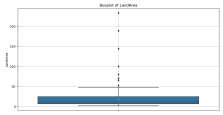
Test Scores Bimodal, close to normal, slight left skew.

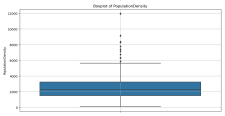
[ ]:
```
plot_boxplots(data)
```

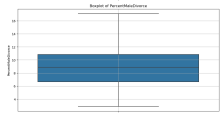

Boxplot of CrueltyArrides
CrueltyArrides

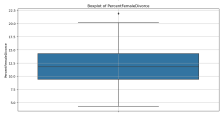

Boxplot of UrbanIndicator
UrbanIndicator

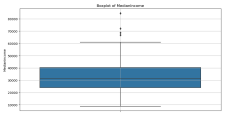

Boxplot of Population
Population

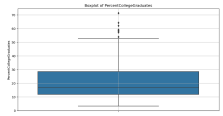

Boxplot of LandArea
LandArea


Boxplot of PopulationDensity
PopulationDensity

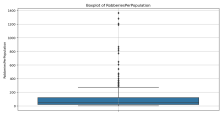

Boxplot of PercentMaleDivorce
PercentMaleDivorce

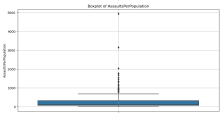

Boxplot of PercentFemaleDivorce
PercentFemaleDivorce

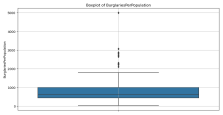

Boxplot of MedianIncome
MedianIncome

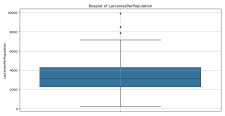

Boxplot of PercentCollegeGraduates
PercentCollegeGraduates

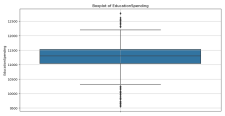

Boxplot of MedianHouseAge
MedianHouseAge

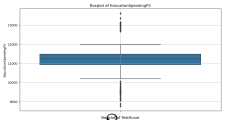

Boxplot of RobberiesPerPopulation
RobberiesPerPopulation

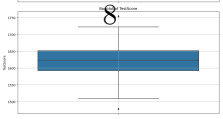

Boxplot of AssaultsPerPopulation
AssaultsPerPopulation


Boxplot of BurglariesPerPopulation
BurglariesPerPopulation


Boxplot of LarcenciesPerPopulation
LarcenciesPerPopulation


Boxplot of EducationSpending
EducationSpending


Boxplot of EducationSpendingSQ
EducationSpendingSQ


Boxplot of TestScore
TestScore

0.1.3 Boxplots (Outlier identification)

Ignore the first two.

Population 1 outlier present.

Land Area 3 outliers present.

Population Density 1 outlier present.

Male Divorce No outliers.

Female Divorce No outliers.

Median Income 1 outlier present.

College Graduates 1 outlier present.

Median House Age No outliers.

Robberies per 100k Approximately 10 outliers.

Assaults per 100k 2 outliers present.

Burglaries per 100k 1 outlier present.

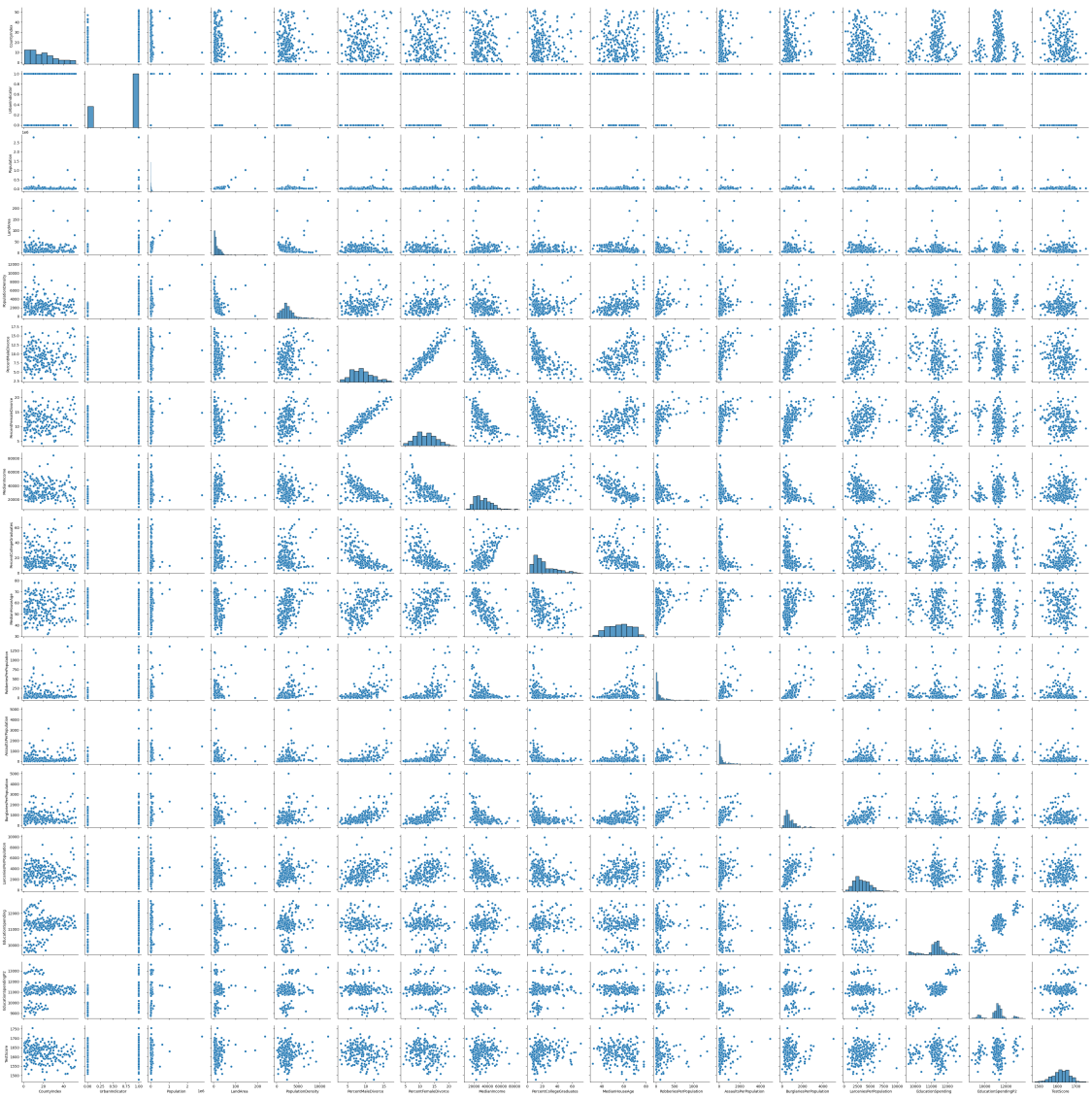
Larcenies per 100k 0-1 outlier present.

Education Spending Quite a few.

Education Spending P2 Quite a few, like the previous.

Test Score No outliers present.

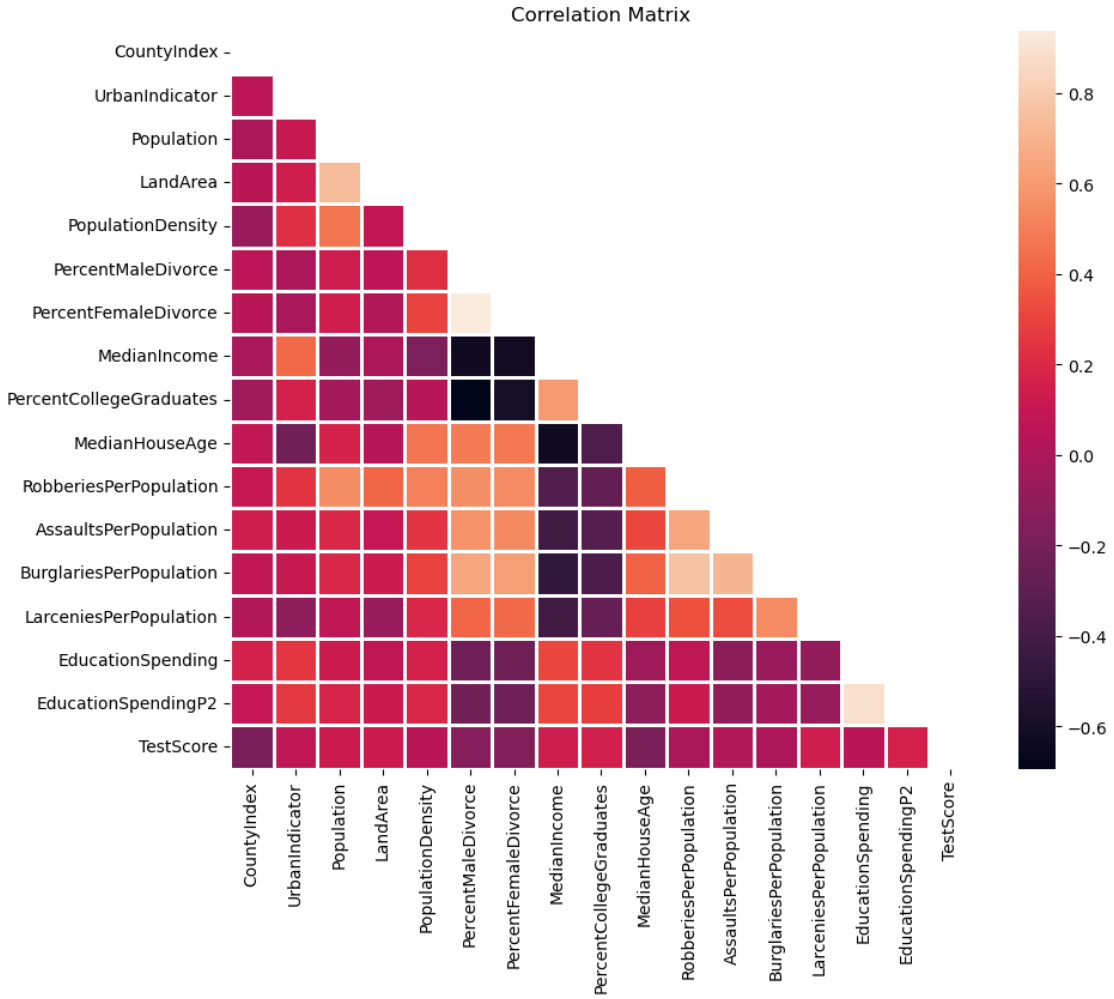
[ ]:
```
# pairplot
with warnings.catch_warnings():
    warnings.simplefilter('ignore', FutureWarning)
    sns.pairplot(data)
    plt.show()
```



Some variables appear to have a linear relationship with each other. Some linear, some non-linear.

0.2 Correlations

[ ]:
```
correlation = data.drop(['State', 'Region', 'RegionNew', 'IncomeCategory'],␣
  ↪axis=1).corr()
mask = np.triu(np.ones_like(correlation, dtype=bool))
plt.figure(figsize=(10,8))
labs = correlation.map(lambda v: v if v else '')
sns.heatmap(correlation, mask=mask, annot=labs, linewidths=1)
plt.title('Correlation Matrix')
plt.show()
```



Strong to Moderate Positive:

- LandArea : Population
- Robberies : Population
- Robberies : LandArea
- MedianHouseAge : Population
- Robberies : PopulationDensity
- FemaleDivorce : MaleDivorce
- MedianHouseAge : MaleDivorce
- Robberies : MaleDivorce
- Assaults : MaleDivorce
- Burglaries : MaleDivorce
- Larcenies : MaleDivorce
- MedianHouseAge : FemaleDivorce
- Robberies : FemaleDivorce
- Assaults : FemaleDivorce
- Burglaries : FemaleDivorce

- Larcenies : FemaleDivorce
- Graduates : MedianIncome
- Assaults : Robberies
- Burglaries : Robberies
- Burglaries : Assaults
- Larcenies : Burglaries
- EducationSpending : EducationSpendingP2

Strong to Moderate Negative:

- MedianIncome : MaleDivorce
- MedianIncome : FemaleDivorce
- Graduates : MaleDivorce
- Graduates : FemaleDivorce
- MedianHouseAge : MedianIncome
- Burglaries : MedianIncome
- Larcenies : MedianIncome

0.2.1 Recommendations

Could potentially implement hypothesis testing based on our own inferences on the sample provided to us, or conduct a one-way ANOVA to compare differences between means for a specific or multiple hypotheses. We could also attempt to predict values within range using Linear or Logistic regression - or attempt to cluster datapoints to make an inference on the population of this sample.