P-1

February 15, 2024

0.1 Real Estate Capstone Project

Project Task : Week1

Data Import and Preparation

1. Import Data

[1]:
```
# importing required libaries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

import warnings
warnings.filterwarnings('ignore')
```

[2]:
```
pd.set_option('display.max_columns',None)
```

[3]:
```
# import required dataset
df_train = pd.read_csv('train.csv')
df_test = pd.read_csv('test.csv')
```

[4]:
```
df_train.head(2)
```

[4]:
```
      UID  BLOCKID  SUMLEVEL  COUNTYID  STATEID     state state_ab  \
0  267822      NaN       140        53       36  New York       NY
1  246444      NaN       140       141       18   Indiana       IN

         city     place  type primary  zip_code  area_code        lat  \
0    Hamilton  Hamilton  City   tract     13346        315  42.840812
1  South Bend  Roseland  City   tract     46616        574  41.701441

         lng        ALand   AWater   pop  male_pop  female_pop  rent_mean  \
0 -75.501524  202183361.0  1699120  5230      2612        2618  769.38638
1 -86.266614    1560828.0   100363  2633      1349        1284  804.87924

   rent_median  rent_stdev  rent_sample_weight  rent_samples  rent_gt_10  \
0        784.0   232.63967           272.34441         362.0     0.86761
```

```
1       848.0   253.46747           312.58622         513.0     0.97410

   rent_gt_15  rent_gt_20  rent_gt_25  rent_gt_30  rent_gt_35  rent_gt_40  \
0     0.79155     0.59155     0.45634     0.42817     0.18592     0.15493
1     0.93227     0.69920     0.69920     0.55179     0.41235     0.39044

   rent_gt_50  universe_samples  used_samples      hi_mean  hi_median  \
0     0.12958               387           355  63125.28406    48120.0
1     0.27888               542           502  41931.92593    35186.0

      hi_stdev  hi_sample_weight  hi_samples  family_mean  family_median  \
0  49042.01206        1290.96240      2024.0  67994.14790        53245.0
1  31639.50203         838.74664      1127.0  50670.10337        43023.0

   family_stdev  family_sample_weight  family_samples  hc_mortgage_mean  \
0   47667.30119             884.33516          1491.0        1414.80295
1   34715.57548             375.28798           554.0         864.41390

   hc_mortgage_median  hc_mortgage_stdev  hc_mortgage_sample_weight  \
0              1223.0          641.22898                  377.83135
1               784.0          482.27020                  316.88320

   hc_mortgage_samples    hc_mean  hc_median   hc_stdev  hc_samples  \
0                867.0  570.01530      558.0  270.11299       770.0
1                356.0  351.98293      336.0  125.40457       229.0

   hc_sample_weight  home_equity_second_mortgage  second_mortgage  \
0         499.29293                      0.01588          0.02077
1         189.60606                      0.02222          0.02222

   home_equity     debt  second_mortgage_cdf  home_equity_cdf  debt_cdf  \
0      0.08919  0.52963              0.43658          0.49087   0.73341
1      0.04274  0.60855              0.42174          0.70823   0.58120

   hs_degree  hs_degree_male  hs_degree_female  male_age_mean  \
0    0.89288         0.85880           0.92434       42.48574
1    0.90487         0.86947           0.94187       34.84728

   male_age_median  male_age_stdev  male_age_sample_weight  male_age_samples  \
0             44.0        22.97306               696.42136            2612.0
1             32.0        20.37452               323.90204            1349.0

   female_age_mean  female_age_median  female_age_stdev  \
0         44.48629           45.33333          22.51276
1         36.48391           37.58333          23.43353

   female_age_sample_weight  female_age_samples  pct_own  married  \
```

```
0                685.33845              2618.0  0.79046  0.57851
1                267.23367              1284.0  0.52483  0.34886

   married_snp  separated  divorced
0      0.01882    0.01240    0.0877
1      0.01426    0.01426    0.0903
```

[5]:

```
df_test.head(2)
```

[5]:

```
      UID  BLOCKID  SUMLEVEL  COUNTYID  STATEID     state state_ab     city  \
0  255504      NaN       140       163       26  Michigan       MI  Detroit
1  252676      NaN       140         1       23     Maine       ME   Auburn

                  place  type primary  zip_code  area_code        lat  \
0  Dearborn Heights City   CDP   tract     48239        313  42.346422
1           Auburn City  City   tract      4210        207  44.100724

         lng     ALand   AWater   pop  male_pop  female_pop  rent_mean  \
0 -83.252823   2711280    39555  3417      1479        1938  858.57169
1 -70.257832  14778785  2705204  3796      1846        1950  832.68625

   rent_median  rent_stdev  rent_sample_weight  rent_samples  rent_gt_10  \
0        859.0   232.39082           276.07497         424.0         1.0
1        750.0   267.22342           183.32299         245.0         1.0

   rent_gt_15  rent_gt_20  rent_gt_25  rent_gt_30  rent_gt_35  rent_gt_40  \
0     0.95696     0.85316     0.85316     0.85316     0.85316     0.76962
1     1.00000     0.86611     0.67364     0.30962     0.30962     0.30962

   rent_gt_50  universe_samples  used_samples      hi_mean  hi_median  \
0     0.63544               435           395  48899.52121    38746.0
1     0.27197               275           239  72335.33234    61008.0

      hi_stdev  hi_sample_weight  hi_samples  family_mean  family_median  \
0  44392.20902         798.02401      1180.0  53802.87122        45167.0
1  51895.81159         922.82969      1722.0  85642.22095        74759.0

   family_stdev  family_sample_weight  family_samples  hc_mortgage_mean  \
0   43756.56479             464.30972           769.0        1139.24548
1   49156.72870             482.99945          1147.0        1533.25988

   hc_mortgage_median  hc_mortgage_stdev  hc_mortgage_sample_weight  \
0              1109.0          336.47710                  262.67011
1              1438.0          536.61118                  373.96188

   hc_mortgage_samples    hc_mean  hc_median    hc_stdev  hc_samples  \
0                474.0  488.51323      436.0  192.75147       271.0
```

```
1               937.0  661.31296      668.0  201.31365       510.0

   hc_sample_weight  home_equity_second_mortgage  second_mortgage  \
0         189.18182                      0.06443          0.06443
1         279.69697                      0.01175          0.01175

   home_equity     debt  second_mortgage_cdf  home_equity_cdf  debt_cdf  \
0      0.07651  0.63624              0.14111          0.55087   0.51965
1      0.14375  0.64755              0.52310          0.26442   0.49359

   hs_degree  hs_degree_male  hs_degree_female  male_age_mean  \
0    0.91047         0.92010           0.90391       33.37131
1    0.94290         0.92832           0.95736       43.88680

   male_age_median  male_age_stdev  male_age_sample_weight  male_age_samples  \
0         27.83333        22.36768               334.30978            1479.0
1         46.08333        22.90302               427.10824            1846.0

   female_age_mean  female_age_median  female_age_stdev  \
0         34.78682           33.75000          21.58531
1         44.23451           46.66667          22.37036

   female_age_sample_weight  female_age_samples  pct_own  married  \
0                 416.48097              1938.0  0.70252  0.28217
1                 532.03505              1950.0  0.85128  0.64221

   married_snp  separated  divorced
0      0.05910    0.03813   0.14299
1      0.02338    0.00000   0.13377
```

[6]:
```
df_train.shape
```

[6]:
```
(27321, 80)
```

[7]:
```
df_test.shape
```

[7]:
```
(11709, 80)
```

[8]:
```
df_train.describe()
```

[8]:
```
                 UID  BLOCKID  SUMLEVEL      COUNTYID       STATEID  \
count   27321.000000      0.0   27321.0  27321.000000  27321.000000
mean   257331.996303      NaN     140.0     85.646426     28.271806
std     21343.859725      NaN       0.0     98.333097     16.392846
min    220342.000000      NaN     140.0      1.000000      1.000000
25%    238816.000000      NaN     140.0     29.000000     13.000000
50%    257220.000000      NaN     140.0     63.000000     28.000000
```

```
75%    275818.000000      NaN     140.0    109.000000     42.000000
max    294334.000000      NaN     140.0    840.000000     72.000000

           zip_code     area_code           lat           lng          ALand  \
count  27321.000000  27321.000000  27321.000000  27321.000000  2.732100e+04
mean   50081.999524    596.507668     37.508813    -91.288394  1.295106e+08
std    29558.115660    232.497482      5.588268     16.343816  1.275531e+09
min      602.000000    201.000000     17.929085   -165.453872  4.113400e+04
25%    26554.000000    405.000000     33.899064    -97.816067  1.799408e+06
50%    47715.000000    614.000000     38.755183    -86.554374  4.866940e+06
75%    77093.000000    801.000000     41.380606    -79.782503  3.359820e+07
max    99925.000000    989.000000     67.074017    -65.379332  1.039510e+11

             AWater           pop      male_pop    female_pop     rent_mean  \
count  2.732100e+04  27321.000000  27321.000000  27321.000000  27007.000000
mean   6.521754e+06   4316.032685   2123.924820   2192.107866   1055.129032
std    2.186781e+08   2169.226173   1114.948893   1101.895160    437.430562
min    0.000000e+00      0.000000      0.000000      0.000000    117.150000
25%    0.000000e+00   2885.000000   1403.000000   1454.000000    743.153540
50%    2.756300e+04   4042.000000   1978.000000   2056.000000    953.193930
75%    5.239880e+05   5430.000000   2668.000000   2764.000000   1259.900165
max    2.453228e+10  53812.000000  27962.000000  27250.000000   3962.342290

        rent_median    rent_stdev  rent_sample_weight  rent_samples  \
count  27007.000000  27007.000000        27007.000000  27007.000000
mean    1007.672789    394.256202          295.979447    548.005702
std      443.797814    187.190303          272.203470    461.547524
min      104.000000     18.257420            0.343000      4.000000
25%      702.000000    263.662575          101.922785    221.000000
50%      897.000000    346.397060          219.210100    424.000000
75%     1198.000000    475.601650          408.709870    742.000000
max     3972.000000   1556.383030         3060.247900   6281.000000

         rent_gt_10    rent_gt_15    rent_gt_20    rent_gt_25    rent_gt_30  \
count  27007.000000  27007.000000  27007.000000  27007.000000  27007.000000
mean       0.957824      0.867134      0.739429      0.612959      0.499994
std        0.063186      0.109655      0.143799      0.160305      0.164006
min        0.000000      0.000000      0.000000      0.000000      0.000000
25%        0.940625      0.819330      0.662085      0.517115      0.396230
50%        0.977070      0.888160      0.758170      0.625000      0.503790
75%        1.000000      0.940680      0.837300      0.722290      0.608515
max        1.000000      1.000000      1.000000      1.000000      1.000000

         rent_gt_35    rent_gt_40    rent_gt_50  universe_samples  \
count  27007.000000  27007.000000  27007.000000      27321.000000
mean       0.411007      0.345424      0.254469        574.269390
std        0.160201      0.153217      0.137742        466.009996
```

min	0.000000	0.000000	0.000000	0.000000
25%	0.307095	0.243325	0.160775	250.000000
50%	0.408600	0.338620	0.242950	454.000000
75%	0.515145	0.440915	0.335690	771.000000
max	1.000000	1.000000	1.000000	6648.000000

	used_samples	hi_mean	hi_median	hi_stdev \
count	27321.000000	27053.000000	27053.000000	27053.000000
mean	528.533546	70441.191421	57580.508964	54429.005158
std	450.622720	30166.895308	29128.465950	17619.932892
min	0.000000	4999.846690	4790.000000	1825.741860
25%	209.000000	49149.660560	37424.000000	42093.741360
50%	408.000000	64020.023850	51278.000000	52213.886470
75%	718.000000	85812.383150	70734.000000	65329.560620
max	6094.000000	297142.857100	296897.000000	135902.619500

	hi_sample_weight	hi_samples	family_mean	family_median \
count	27053.000000	27053.000000	27023.000000	27023.000000
mean	923.580372	1607.974384	78987.539104	69279.801465
std	453.057675	751.096015	31386.178602	33472.030541
min	0.114260	3.000000	5374.842520	5278.000000
25%	600.290760	1096.000000	56859.372910	46166.000000
50%	863.714170	1519.000000	72876.445610	62416.000000
75%	1179.293470	2016.000000	96010.265100	84712.000000
max	10931.975610	20395.000000	242857.142900	242720.000000

	family_stdev	family_sample_weight	family_samples	hc_mortgage_mean \
count	27023.000000	27023.000000	27023.000000	26748.000000
mean	50728.337493	533.686966	1063.665988	1629.856392
std	14239.749880	290.603105	560.873112	623.206122
min	1825.741860	0.199960	3.000000	234.650000
25%	40887.774050	331.677595	687.000000	1158.312197
50%	49679.731230	490.868190	986.000000	1460.483290
75%	60415.096305	685.226575	1349.000000	1982.588285
max	111256.702500	6904.496890	14938.000000	4462.342290

	hc_mortgage_median	hc_mortgage_stdev	hc_mortgage_sample_weight \
count	26748.000000	26748.000000	26748.000000
mean	1551.455735	622.559191	287.552519
std	652.619435	238.068593	195.340264
min	237.000000	36.514840	0.198400
25%	1067.000000	440.432127	148.116155
50%	1371.000000	589.364540	253.549800
75%	1877.000000	788.063712	387.225985
max	4472.000000	1596.206270	4226.744200

	hc_mortgage_samples	hc_mean	hc_median	hc_stdev \

count	26748.000000	26721.000000	26721.000000	26721.000000
mean	669.827389	540.549473	513.383968	218.604647
std	464.411215	221.339933	231.392365	91.456509
min	1.000000	53.594610	53.000000	18.257420
25%	346.000000	389.284170	361.000000	154.444740
50%	590.000000	478.798920	449.000000	198.699610
75%	895.000000	631.398210	600.000000	266.510900
max	11670.000000	1700.179110	1702.000000	820.968550

	hc_samples	hc_sample_weight	home_equity_second_mortgage	\
count	26721.000000	26721.000000	26864.000000	
mean	370.284570	254.722233	0.025695	
std	250.727935	189.912748	0.031331	
min	2.000000	0.614040	0.000000	
25%	193.000000	120.818180	0.004990	
50%	327.000000	213.030300	0.018515	
75%	500.000000	342.572420	0.036943	
max	11330.000000	7107.064500	1.000000	

	second_mortgage	home_equity	debt	second_mortgage_cdf	\
count	26864.000000	26864.000000	26864.000000	26864.000000	
mean	0.029947	0.100847	0.629190	0.467957	
std	0.034134	0.069304	0.156267	0.294956	
min	0.000000	0.000000	0.000000	0.000000	
25%	0.007680	0.049247	0.538460	0.248910	
50%	0.022500	0.094400	0.648315	0.419310	
75%	0.042732	0.143492	0.737525	0.554115	
max	1.000000	1.000000	1.000000	1.000000	

	home_equity_cdf	debt_cdf	hs_degree	hs_degree_male	\
count	26864.000000	26864.000000	27131.000000	27121.000000	
mean	0.477485	0.499458	0.858459	0.852136	
std	0.256125	0.264138	0.112420	0.120746	
min	0.000000	0.000000	0.186520	0.000000	
25%	0.265270	0.281195	0.807890	0.795270	
50%	0.466705	0.491890	0.889040	0.883920	
75%	0.678620	0.718510	0.939580	0.941070	
max	1.000000	1.000000	1.000000	1.000000	

	hs_degree_female	male_age_mean	male_age_median	male_age_stdev	\
count	27098.000000	27132.000000	27132.000000	27132.000000	
mean	0.864931	38.339988	38.074193	21.500301	
std	0.112273	5.602570	7.874651	2.540576	
min	0.000000	12.145830	9.750000	0.962770	
25%	0.818025	35.020857	32.833330	20.581182	
50%	0.895935	38.336880	37.833330	21.906380	
75%	0.944650	41.402438	42.916670	22.954955	

```
max          1.000000     77.759920      80.166670      31.060950

       male_age_sample_weight  male_age_samples  female_age_mean  \
count            27132.000000      27132.000000     27115.000000
mean               535.457318       2138.719962        40.319803
std                312.922652       1104.593574         5.886317
min                  0.745760          3.000000        16.008330
25%                346.200508       1416.000000        36.892050
50%                490.967750       1986.000000        40.373320
75%                666.267472       2672.250000        43.567120
max              12017.070440      27962.000000        79.837390

       female_age_median  female_age_stdev  female_age_sample_weight  \
count       27115.000000      27115.000000              27115.000000
mean           40.355099         22.178745                544.238432
std             8.039585          2.540257                283.546896
min            13.250000          0.556780                  0.664700
25%            34.916670         21.312135                355.995825
50%            40.583330         22.514410                503.643890
75%            45.416670         23.575260                680.275055
max            82.250000         30.241270               6197.995200

       female_age_samples       pct_own       married   married_snp  \
count        27115.000000  27053.000000  27130.000000  27130.000000
mean          2208.761903      0.640434      0.508300      0.047537
std           1089.316999      0.226640      0.136860      0.037640
min              2.000000      0.000000      0.000000      0.000000
25%           1471.000000      0.502780      0.425102      0.020810
50%           2066.000000      0.690840      0.526665      0.038840
75%           2772.000000      0.817460      0.605760      0.065100
max          27250.000000      1.000000      1.000000      0.714290

          separated      divorced
count  27130.000000  27130.000000
mean       0.019089      0.100248
std        0.020796      0.049055
min        0.000000      0.000000
25%        0.004530      0.065800
50%        0.013460      0.095205
75%        0.027488      0.129000
max        0.714290      1.000000
```

[9]:
```
df_test.describe()
```

[9]:
```
                 UID  BLOCKID  SUMLEVEL      COUNTYID       STATEID  \
count   11709.000000      0.0   11709.0  11709.000000  11709.000000
mean   257525.004783      NaN     140.0     85.710650     28.489196
```

```
std      21466.372658       NaN      0.0     99.304334     16.607262
min     220336.000000       NaN    140.0      1.000000      1.000000
25%     238819.000000       NaN    140.0     29.000000     13.000000
50%     257651.000000       NaN    140.0     61.000000     28.000000
75%     276300.000000       NaN    140.0    109.000000     42.000000
max     294333.000000       NaN    140.0    810.000000     72.000000

           zip_code     area_code           lat           lng         ALand  \
count  11709.000000  11709.000000  11709.000000  11709.000000  1.170900e+04
mean   50123.418396    593.598514     37.405491    -91.340229  1.095500e+08
std    29775.134038    232.074263      5.625904     16.407818  7.624940e+08
min      601.000000    201.000000     17.965835   -166.770979  8.299000e+03
25%    25570.000000    404.000000     33.919813    -97.816561  1.718660e+06
50%    47362.000000    612.000000     38.618093    -86.643344  4.835000e+06
75%    77406.000000    787.000000     41.232973    -79.697311  3.204540e+07
max    99929.000000    989.000000     64.804269    -65.695344  5.520166e+10

             AWater           pop      male_pop    female_pop     rent_mean  \
count  1.170900e+04  11709.000000  11709.000000  11709.000000  11561.000000
mean   5.156069e+06   4367.205995   2152.510804   2214.695192   1054.143003
std    1.522649e+08   2121.779736   1086.382137   1086.438040    434.549555
min    0.000000e+00      0.000000      0.000000      0.000000    147.548100
25%    0.000000e+00   2937.000000   1433.000000   1484.000000    741.389730
50%    2.270900e+04   4119.000000   2010.000000   2090.000000    952.526270
75%    4.864500e+05   5474.000000   2690.000000   2792.000000   1259.756750
max    1.212570e+10  39454.000000  27962.000000  15466.000000   3962.342290

        rent_median    rent_stdev  rent_sample_weight  rent_samples  \
count  11561.000000  11561.000000         11561.00000  11561.000000
mean    1007.017646    394.613338           304.51603    563.476256
std      441.484366    189.193868           281.31471    474.563369
min      104.000000     18.257420             0.39279      3.000000
25%      704.000000    262.377940           103.86843    226.000000
50%      897.000000    349.497450           228.96877    441.000000
75%     1194.000000    475.718140           420.81563    763.000000
max     3972.000000   1720.718990          4112.12237   7634.000000

         rent_gt_10    rent_gt_15    rent_gt_20    rent_gt_25    rent_gt_30  \
count  11560.000000  11560.000000  11560.000000  11560.000000  11560.000000
mean       0.957482      0.867770      0.742615      0.614405      0.501188
std        0.063603      0.107789      0.142514      0.161556      0.165759
min        0.000000      0.000000      0.000000      0.000000      0.000000
25%        0.940410      0.820913      0.665775      0.517220      0.397740
50%        0.976970      0.889180      0.763485      0.628110      0.507090
75%        1.000000      0.939660      0.839375      0.726447      0.612313
max        1.000000      1.000000      1.000000      1.000000      1.000000
```

	rent_gt_35	rent_gt_40	rent_gt_50	universe_samples	\
count	11560.000000	11560.000000	11560.000000	11709.000000	
mean	0.412992	0.347003	0.255507	588.795969	
std	0.161312	0.153982	0.137658	477.469706	
min	0.000000	0.000000	0.000000	0.000000	
25%	0.307947	0.241998	0.160375	255.000000	
50%	0.412875	0.342330	0.243710	470.000000	
75%	0.517088	0.444723	0.340120	790.000000	
max	1.000000	1.000000	1.000000	7634.000000	

	used_samples	hi_mean	hi_median	hi_stdev	\
count	11709.000000	11587.000000	11587.000000	11587.000000	
mean	542.688189	70169.909595	57361.971779	54164.666604	
std	463.283992	30619.277296	29661.241996	17794.261539	
min	0.000000	4999.846690	4790.000000	1825.741860	
25%	216.000000	48814.166430	36953.500000	41662.440610	
50%	424.000000	63788.482430	51013.000000	51925.227180	
75%	741.000000	85416.924520	70484.500000	64897.947475	
max	7336.000000	221622.723500	242249.000000	124534.013900	

	hi_sample_weight	hi_samples	family_mean	family_median	\
count	11587.000000	11587.000000	11573.000000	11573.000000	
mean	935.084700	1624.344093	78684.992592	69049.818630	
std	457.759256	747.394839	31979.019465	34130.762923	
min	0.399920	3.000000	5374.842520	5278.000000	
25%	611.598530	1110.000000	56140.036620	45709.000000	
50%	877.368400	1530.000000	72809.895350	61971.000000	
75%	1194.786860	2031.000000	95623.665980	84319.000000	
max	8133.778720	12316.000000	242857.142900	242720.000000	

	family_stdev	family_sample_weight	family_samples	hc_mortgage_mean	\
count	11573.000000	11573.000000	11573.000000	11441.000000	
mean	50408.173385	540.262293	1073.081483	1636.445391	
std	14349.930513	289.029814	550.898356	634.770720	
min	1825.741860	0.266610	4.000000	349.500000	
25%	40413.475230	338.046690	694.000000	1152.337490	
50%	49401.698830	496.572350	996.000000	1463.893720	
75%	60297.436260	689.158350	1358.000000	1990.646240	
max	105579.486100	4888.944600	6658.000000	4462.342290	

	hc_mortgage_median	hc_mortgage_stdev	hc_mortgage_sample_weight	\
count	11441.000000	11441.000000	11441.000000	
mean	1559.639018	621.742098	289.285332	
std	664.567754	240.815700	197.175161	
min	349.000000	36.514840	0.595190	
25%	1068.000000	436.938690	147.242890	
50%	1374.000000	586.516070	255.414250	

```
75%          1885.000000        787.554270                 387.587270
max          4472.000000       1814.113980                1936.551660

       hc_mortgage_samples        hc_mean      hc_median       hc_stdev  \
count         11441.000000   11419.000000   11419.000000   11419.000000
mean            673.433004     538.906730     512.067869     217.949778
std             461.505232     226.307832     237.514474      93.108675
min               2.000000      53.594610      53.000000      18.257420
25%             343.000000     386.273775     357.000000     152.652175
50%             593.000000     474.995830     445.000000     198.361260
75%             908.000000     629.517360     598.000000     265.684575
max            5033.000000    1700.179110    1702.000000     782.862850

         hc_samples  hc_sample_weight  home_equity_second_mortgage  \
count  11419.000000      11419.000000                 11489.000000
mean     369.762326        255.189048                     0.025789
std      249.644673        190.267726                     0.030513
min        2.000000          0.491230                     0.000000
25%      189.000000        118.787880                     0.005060
50%      327.000000        212.090910                     0.018780
75%      501.000000        345.170125                     0.037270
max     3965.000000       2878.131310                     1.000000

       second_mortgage   home_equity           debt  second_mortgage_cdf  \
count     11489.000000  11489.000000   11489.000000         11489.000000
mean          0.030187      0.101570       0.631615             0.467226
std           0.033644      0.070412       0.157634             0.296905
min           0.000000      0.000000       0.000000             0.000000
25%           0.007790      0.049700       0.541060             0.246060
50%           0.022600      0.095440       0.650070             0.418330
75%           0.043150      0.143860       0.740560             0.553320
max           1.000000      1.000000       1.000000             1.000000

       home_equity_cdf      debt_cdf      hs_degree  hs_degree_male  \
count     11489.000000  11489.000000   11624.000000    11620.000000
mean          0.475517      0.494432       0.855912        0.849148
std           0.257148      0.264962       0.114424        0.122605
min           0.000000      0.000000       0.000000        0.000000
25%           0.263960      0.274550       0.802980        0.790218
50%           0.461850      0.487770       0.886430        0.881020
75%           0.676590      0.714090       0.940100        0.940182
max           1.000000      1.000000       1.000000        1.000000

       hs_degree_female  male_age_mean  male_age_median  male_age_stdev  \
count      11604.000000   11625.000000     11625.000000    11625.000000
mean           0.863003      38.149424        37.833111       21.431971
std            0.113205       5.579728         7.795907        2.582541
```

```
min          0.199710    17.009880     9.750000     0.737110
25%          0.813850    34.916000    32.666670    20.507130
50%          0.893695    38.200730    37.833330    21.884600
75%          0.944935    41.180250    42.583330    22.938350
max          1.000000    83.358330    83.333330    27.920410

       male_age_sample_weight  male_age_samples  female_age_mean  \
count            11625.000000      11625.000000     11613.000000
mean               542.945584       2168.064430        40.111999
std                296.016752       1074.723594         5.851192
min                  0.745760          4.000000        15.360240
25%                355.219790       1445.000000        36.729210
50%                499.653480       2020.000000        40.196960
75%                676.560290       2696.000000        43.496490
max              12017.070440      27962.000000        90.107940

       female_age_median  female_age_stdev  female_age_sample_weight  \
count       11613.000000      11613.000000              11613.000000
mean           40.131864         22.148145                550.411243
std             7.972026          2.554907                280.992521
min            12.833330          0.737110                  0.251910
25%            34.750000         21.270920                363.225840
50%            40.333330         22.472990                509.103610
75%            45.333330         23.549450                685.883910
max            90.166670         29.626680               4145.557870

       female_age_samples       pct_own       married   married_snp  \
count        11613.000000  11587.000000  11625.000000  11625.000000
mean          2233.003186      0.634194      0.505632      0.047960
std           1072.017063      0.232232      0.139774      0.038693
min              3.000000      0.000000      0.000000      0.000000
25%           1499.000000      0.492500      0.422020      0.020890
50%           2099.000000      0.687640      0.525270      0.038680
75%           2800.000000      0.815235      0.605660      0.065340
max          15466.000000      1.000000      1.000000      0.714290

          separated      divorced
count  11625.000000  11625.000000
mean       0.019346      0.099191
std        0.021428      0.048525
min        0.000000      0.000000
25%        0.004500      0.064590
50%        0.013870      0.094350
75%        0.027910      0.128400
max        0.714290      0.362750
```

[10]:

```
df_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27321 entries, 0 to 27320
Data columns (total 80 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   UID                         27321 non-null  int64
 1   BLOCKID                     0 non-null      float64
 2   SUMLEVEL                    27321 non-null  int64
 3   COUNTYID                    27321 non-null  int64
 4   STATEID                     27321 non-null  int64
 5   state                       27321 non-null  object
 6   state_ab                    27321 non-null  object
 7   city                        27321 non-null  object
 8   place                       27321 non-null  object
 9   type                        27321 non-null  object
 10  primary                     27321 non-null  object
 11  zip_code                    27321 non-null  int64
 12  area_code                   27321 non-null  int64
 13  lat                         27321 non-null  float64
 14  lng                         27321 non-null  float64
 15  ALand                       27321 non-null  float64
 16  AWater                      27321 non-null  int64
 17  pop                         27321 non-null  int64
 18  male_pop                    27321 non-null  int64
 19  female_pop                  27321 non-null  int64
 20  rent_mean                   27007 non-null  float64
 21  rent_median                 27007 non-null  float64
 22  rent_stdev                  27007 non-null  float64
 23  rent_sample_weight          27007 non-null  float64
 24  rent_samples                27007 non-null  float64
 25  rent_gt_10                  27007 non-null  float64
 26  rent_gt_15                  27007 non-null  float64
 27  rent_gt_20                  27007 non-null  float64
 28  rent_gt_25                  27007 non-null  float64
 29  rent_gt_30                  27007 non-null  float64
 30  rent_gt_35                  27007 non-null  float64
 31  rent_gt_40                  27007 non-null  float64
 32  rent_gt_50                  27007 non-null  float64
 33  universe_samples            27321 non-null  int64
 34  used_samples                27321 non-null  int64
 35  hi_mean                     27053 non-null  float64
 36  hi_median                   27053 non-null  float64
 37  hi_stdev                    27053 non-null  float64
 38  hi_sample_weight            27053 non-null  float64
 39  hi_samples                  27053 non-null  float64
 40  family_mean                 27023 non-null  float64
 41  family_median               27023 non-null  float64
 42  family_stdev                27023 non-null  float64
```

```
 43  family_sample_weight         27023 non-null  float64
 44  family_samples               27023 non-null  float64
 45  hc_mortgage_mean             26748 non-null  float64
 46  hc_mortgage_median           26748 non-null  float64
 47  hc_mortgage_stdev            26748 non-null  float64
 48  hc_mortgage_sample_weight    26748 non-null  float64
 49  hc_mortgage_samples          26748 non-null  float64
 50  hc_mean                      26721 non-null  float64
 51  hc_median                    26721 non-null  float64
 52  hc_stdev                     26721 non-null  float64
 53  hc_samples                   26721 non-null  float64
 54  hc_sample_weight             26721 non-null  float64
 55  home_equity_second_mortgage  26864 non-null  float64
 56  second_mortgage              26864 non-null  float64
 57  home_equity                  26864 non-null  float64
 58  debt                         26864 non-null  float64
 59  second_mortgage_cdf          26864 non-null  float64
 60  home_equity_cdf              26864 non-null  float64
 61  debt_cdf                     26864 non-null  float64
 62  hs_degree                    27131 non-null  float64
 63  hs_degree_male               27121 non-null  float64
 64  hs_degree_female             27098 non-null  float64
 65  male_age_mean                27132 non-null  float64
 66  male_age_median              27132 non-null  float64
 67  male_age_stdev               27132 non-null  float64
 68  male_age_sample_weight       27132 non-null  float64
 69  male_age_samples             27132 non-null  float64
 70  female_age_mean              27115 non-null  float64
 71  female_age_median            27115 non-null  float64
 72  female_age_stdev             27115 non-null  float64
 73  female_age_sample_weight     27115 non-null  float64
 74  female_age_samples           27115 non-null  float64
 75  pct_own                      27053 non-null  float64
 76  married                      27130 non-null  float64
 77  married_snp                  27130 non-null  float64
 78  separated                    27130 non-null  float64
 79  divorced                     27130 non-null  float64
dtypes: float64(62), int64(12), object(6)
memory usage: 16.7+ MB
```

[11]:
```
df_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27321 entries, 0 to 27320
Data columns (total 80 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   UID                          27321 non-null  int64
```

1	BLOCKID	0 non-null	float64
2	SUMLEVEL	27321 non-null	int64
3	COUNTYID	27321 non-null	int64
4	STATEID	27321 non-null	int64
5	state	27321 non-null	object
6	state_ab	27321 non-null	object
7	city	27321 non-null	object
8	place	27321 non-null	object
9	type	27321 non-null	object
10	primary	27321 non-null	object
11	zip_code	27321 non-null	int64
12	area_code	27321 non-null	int64
13	lat	27321 non-null	float64
14	lng	27321 non-null	float64
15	ALand	27321 non-null	float64
16	AWater	27321 non-null	int64
17	pop	27321 non-null	int64
18	male_pop	27321 non-null	int64
19	female_pop	27321 non-null	int64
20	rent_mean	27007 non-null	float64
21	rent_median	27007 non-null	float64
22	rent_stdev	27007 non-null	float64
23	rent_sample_weight	27007 non-null	float64
24	rent_samples	27007 non-null	float64
25	rent_gt_10	27007 non-null	float64
26	rent_gt_15	27007 non-null	float64
27	rent_gt_20	27007 non-null	float64
28	rent_gt_25	27007 non-null	float64
29	rent_gt_30	27007 non-null	float64
30	rent_gt_35	27007 non-null	float64
31	rent_gt_40	27007 non-null	float64
32	rent_gt_50	27007 non-null	float64
33	universe_samples	27321 non-null	int64
34	used_samples	27321 non-null	int64
35	hi_mean	27053 non-null	float64
36	hi_median	27053 non-null	float64
37	hi_stdev	27053 non-null	float64
38	hi_sample_weight	27053 non-null	float64
39	hi_samples	27053 non-null	float64
40	family_mean	27023 non-null	float64
41	family_median	27023 non-null	float64
42	family_stdev	27023 non-null	float64
43	family_sample_weight	27023 non-null	float64
44	family_samples	27023 non-null	float64
45	hc_mortgage_mean	26748 non-null	float64
46	hc_mortgage_median	26748 non-null	float64
47	hc_mortgage_stdev	26748 non-null	float64
48	hc_mortgage_sample_weight	26748 non-null	float64

```
 49  hc_mortgage_samples          26748 non-null  float64
 50  hc_mean                      26721 non-null  float64
 51  hc_median                    26721 non-null  float64
 52  hc_stdev                     26721 non-null  float64
 53  hc_samples                   26721 non-null  float64
 54  hc_sample_weight             26721 non-null  float64
 55  home_equity_second_mortgage  26864 non-null  float64
 56  second_mortgage              26864 non-null  float64
 57  home_equity                  26864 non-null  float64
 58  debt                         26864 non-null  float64
 59  second_mortgage_cdf          26864 non-null  float64
 60  home_equity_cdf              26864 non-null  float64
 61  debt_cdf                     26864 non-null  float64
 62  hs_degree                    27131 non-null  float64
 63  hs_degree_male               27121 non-null  float64
 64  hs_degree_female             27098 non-null  float64
 65  male_age_mean                27132 non-null  float64
 66  male_age_median              27132 non-null  float64
 67  male_age_stdev               27132 non-null  float64
 68  male_age_sample_weight       27132 non-null  float64
 69  male_age_samples             27132 non-null  float64
 70  female_age_mean              27115 non-null  float64
 71  female_age_median            27115 non-null  float64
 72  female_age_stdev             27115 non-null  float64
 73  female_age_sample_weight     27115 non-null  float64
 74  female_age_samples           27115 non-null  float64
 75  pct_own                      27053 non-null  float64
 76  married                      27130 non-null  float64
 77  married_snp                  27130 non-null  float64
 78  separated                    27130 non-null  float64
 79  divorced                     27130 non-null  float64
dtypes: float64(62), int64(12), object(6)
memory usage: 16.7+ MB
```

2. Figure out the primary key and look for the requirement of indexing. Unique and not null can only be used as Primary Key

[12]:
```
df_train.duplicated().value_counts()
```

[12]:
```
False    27161
True       160
dtype: int64
```

[13]:
```
df_test.duplicated().value_counts()
```

[13]:
```
False    11677
True        32
```

```
dtype: int64
```

Removing duplicates from Datasets

[14]:
```python
df_train.drop_duplicates(keep='first', inplace = True)
df_test.drop_duplicates(keep='first', inplace = True)
```

[15]:
```python
df_train.shape
```

[15]:
```
(27161, 80)
```

[16]:
```python
df_test.shape
```

[16]:
```
(11677, 80)
```

[17]:
```python
df_train.nunique()  == df_train.shape[0]
```

[17]:
```
UID             True
BLOCKID        False
SUMLEVEL       False
COUNTYID       False
STATEID        False
               …
pct_own        False
married        False
married_snp    False
separated      False
divorced       False
Length: 80, dtype: bool
```

[18]:
```python
df_test.nunique()  == df_test.shape[0]
```

[18]:
```
UID             True
BLOCKID        False
SUMLEVEL       False
COUNTYID       False
STATEID        False
               …
pct_own        False
married        False
married_snp    False
separated      False
divorced       False
Length: 80, dtype: bool
```

From above UID has Unique values hence UID can considered as Primary Key for dataset

[19]:
```python
#df_train = df_train.reset_index()
```

```
[20]: #df_test = df_test.reset_index()
```

```
[21]: #df_train
```

3. Missing value Treatment

```
[22]: #This flag will help us split the data back later
df_train['split']= 'Train'
df_test['split']= 'Test'
```

```
[23]: df_combined=df_train.append(df_test, ignore_index=True)
df_combined.head(2)
```

```
[23]:      UID  BLOCKID  SUMLEVEL  COUNTYID  STATEID     state state_ab  \
0  267822      NaN       140        53       36  New York       NY
1  246444      NaN       140       141       18   Indiana       IN

         city     place  type primary  zip_code  area_code        lat  \
0    Hamilton  Hamilton  City   tract     13346        315  42.840812
1  South Bend  Roseland  City   tract     46616        574  41.701441

         lng        ALand   AWater   pop  male_pop  female_pop  rent_mean  \
0 -75.501524  202183361.0  1699120  5230      2612        2618  769.38638
1 -86.266614    1560828.0   100363  2633      1349        1284  804.87924

   rent_median  rent_stdev  rent_sample_weight  rent_samples  rent_gt_10  \
0        784.0   232.63967           272.34441         362.0     0.86761
1        848.0   253.46747           312.58622         513.0     0.97410

   rent_gt_15  rent_gt_20  rent_gt_25  rent_gt_30  rent_gt_35  rent_gt_40  \
0     0.79155     0.59155     0.45634     0.42817     0.18592     0.15493
1     0.93227     0.69920     0.69920     0.55179     0.41235     0.39044

   rent_gt_50  universe_samples  used_samples      hi_mean  hi_median  \
0     0.12958               387           355  63125.28406    48120.0
1     0.27888               542           502  41931.92593    35186.0

      hi_stdev  hi_sample_weight  hi_samples  family_mean  family_median  \
0  49042.01206        1290.96240      2024.0  67994.14790        53245.0
1  31639.50203         838.74664      1127.0  50670.10337        43023.0

   family_stdev  family_sample_weight  family_samples  hc_mortgage_mean  \
0   47667.30119             884.33516          1491.0        1414.80295
1   34715.57548             375.28798           554.0         864.41390

   hc_mortgage_median  hc_mortgage_stdev  hc_mortgage_sample_weight  \
0              1223.0          641.22898                  377.83135
```

```
1              784.0          482.27020                 316.88320

   hc_mortgage_samples    hc_mean  hc_median   hc_stdev  hc_samples  \
0                867.0  570.01530      558.0  270.11299       770.0
1                356.0  351.98293      336.0  125.40457       229.0

   hc_sample_weight  home_equity_second_mortgage  second_mortgage  \
0         499.29293                      0.01588          0.02077
1         189.60606                      0.02222          0.02222

   home_equity     debt  second_mortgage_cdf  home_equity_cdf  debt_cdf  \
0      0.08919  0.52963              0.43658          0.49087   0.73341
1      0.04274  0.60855              0.42174          0.70823   0.58120

   hs_degree  hs_degree_male  hs_degree_female  male_age_mean  \
0    0.89288         0.85880           0.92434       42.48574
1    0.90487         0.86947           0.94187       34.84728

   male_age_median  male_age_stdev  male_age_sample_weight  male_age_samples  \
0             44.0        22.97306               696.42136            2612.0
1             32.0        20.37452               323.90204            1349.0

   female_age_mean  female_age_median  female_age_stdev  \
0         44.48629           45.33333          22.51276
1         36.48391           37.58333          23.43353

   female_age_sample_weight  female_age_samples  pct_own  married  \
0                 685.33845              2618.0  0.79046  0.57851
1                 267.23367              1284.0  0.52483  0.34886

   married_snp  separated  divorced  split
0      0.01882    0.01240    0.0877  Train
1      0.01426    0.01426    0.0903  Train
```

[24]:
```
df_combined.tail(2)
```

[24]:
```
          UID  BLOCKID  SUMLEVEL  COUNTYID  STATEID  state state_ab     city  \
38836  241096      NaN       140        27       19   Iowa       IA  Carroll
38837  287763      NaN       140       453       48  Texas       TX   Austin

                    place  type primary  zip_code  area_code        lat  \
38836        Carroll City  City   tract     51401        712  42.081366
38837  Sunset Valley City  Town   tract     78745        512  30.219013

             lng       ALand  AWater   pop  male_pop  female_pop  rent_mean  \
38836 -94.866175  11066759.0       0  5945      2732        3213  696.93368
38837 -97.774728   1990126.0       0  4117      2070        2047  950.09294
```

```
       rent_median  rent_stdev  rent_sample_weight  rent_samples  rent_gt_10  \
38836        576.0   595.16228           503.83775         590.0     0.96886
38837        864.0   333.82364           417.07457         675.0     1.00000

       rent_gt_15  rent_gt_20  rent_gt_25  rent_gt_30  rent_gt_35  rent_gt_40  \
38836     0.92042     0.83045     0.69723     0.62284     0.43772     0.33737
38837     0.97481     0.86074     0.73926     0.44593     0.38370     0.27852

       rent_gt_50  universe_samples  used_samples      hi_mean  hi_median  \
38836     0.33737               663           578  57877.26387    41838.0
38837     0.25778               682           675  58006.33817    44179.0

          hi_stdev  hi_sample_weight  hi_samples  family_mean  family_median  \
38836  49745.93715        1605.79897      2596.0  75066.29009        72135.0
38837  49189.98590         902.67611      1396.0  54913.24441        42469.0

       family_stdev  family_sample_weight  family_samples  hc_mortgage_mean  \
38836   47200.66016             782.93088          1568.0        1182.30365
38837   41016.08651             581.04758           877.0        1364.17379

       hc_mortgage_median  hc_mortgage_stdev  hc_mortgage_sample_weight  \
38836              1059.0          587.01032                  796.11244
38837              1318.0          463.57052                  217.49287

       hc_mortgage_samples    hc_mean  hc_median   hc_stdev  hc_samples  \
38836               1267.0  369.29903      334.0  133.20792       666.0
38837                456.0  550.78197      555.0  199.13527       258.0

       hc_sample_weight  home_equity_second_mortgage  second_mortgage  \
38836         556.40404                       0.0357           0.0357
38837         163.55556                       0.0000           0.0000

       home_equity     debt  second_mortgage_cdf  home_equity_cdf  debt_cdf  \
38836      0.07967  0.65546               0.3001          0.53579   0.47507
38837      0.05042  0.63866               1.0000          0.67315   0.51407

       hs_degree  hs_degree_male  hs_degree_female  male_age_mean  \
38836    0.91407         0.92428           0.90634       39.18219
38837    0.78685         0.80615           0.76820       35.56404

       male_age_median  male_age_stdev  male_age_sample_weight  \
38836            40.25        24.86317               636.20201
38837            35.00        21.67509               522.45931

       male_age_samples  female_age_mean  female_age_median  female_age_stdev  \
38836            2732.0         45.63179           48.16667          24.84209
```

```
38837           2070.0         35.99955          35.41667          20.68049

       female_age_sample_weight  female_age_samples  pct_own  married  \
38836                 693.82905              3213.0  0.83330  0.66699
38837                 559.30291              2047.0  0.52587  0.51922

       married_snp  separated  divorced split
38836      0.02738     0.0000   0.04694  Test
38837      0.08066     0.0252   0.10586  Test
```

[25]:
```
df_combined.shape
```

[25]:
```
(38838, 81)
```

[26]:
```
df_combined.isna().sum()
```

[26]:
```
UID                0
BLOCKID        38838
SUMLEVEL           0
COUNTYID           0
STATEID            0
               …
married          227
married_snp      227
separated        227
divorced         227
split              0
Length: 81, dtype: int64
```

[27]:
```
# Fill rate of the variables -> (1- missing %)
1-df_combined.isna().sum()/len(df_combined)
```

[27]:
```
UID            1.000000
BLOCKID        0.000000
SUMLEVEL       1.000000
COUNTYID       1.000000
STATEID        1.000000
                 …
married        0.994155
married_snp    0.994155
separated      0.994155
divorced       0.994155
split          1.000000
Length: 81, dtype: float64
```

[28]:
```
# BlOCKID is completly missing or Null in both train and test data. So we will
 ↪drop BLOCKID feature.
```

```
df_combined.drop(columns =['BLOCKID'], axis=1, inplace=True)
```

[29]:
```
df_combined.isna().sum()/len(df_combined)*100
```

[29]:
```
UID               0.000000
SUMLEVEL          0.000000
COUNTYID          0.000000
STATEID           0.000000
state             0.000000
                    …
married           0.584479
married_snp       0.584479
separated         0.584479
divorced          0.584479
split             0.000000
Length: 80, dtype: float64
```

[30]:
```
# Missing value greater than zero
col_check=df_combined.isna().sum().to_frame().reset_index()
null_col=col_check[col_check[0]>0]['index'].tolist()
null_col
```

[30]:
```
['rent_mean',
 'rent_median',
 'rent_stdev',
 'rent_sample_weight',
 'rent_samples',
 'rent_gt_10',
 'rent_gt_15',
 'rent_gt_20',
 'rent_gt_25',
 'rent_gt_30',
 'rent_gt_35',
 'rent_gt_40',
 'rent_gt_50',
 'hi_mean',
 'hi_median',
 'hi_stdev',
 'hi_sample_weight',
 'hi_samples',
 'family_mean',
 'family_median',
 'family_stdev',
 'family_sample_weight',
 'family_samples',
 'hc_mortgage_mean',
 'hc_mortgage_median',
```

```
 'hc_mortgage_stdev',
 'hc_mortgage_sample_weight',
 'hc_mortgage_samples',
 'hc_mean',
 'hc_median',
 'hc_stdev',
 'hc_samples',
 'hc_sample_weight',
 'home_equity_second_mortgage',
 'second_mortgage',
 'home_equity',
 'debt',
 'second_mortgage_cdf',
 'home_equity_cdf',
 'debt_cdf',
 'hs_degree',
 'hs_degree_male',
 'hs_degree_female',
 'male_age_mean',
 'male_age_median',
 'male_age_stdev',
 'male_age_sample_weight',
 'male_age_samples',
 'female_age_mean',
 'female_age_median',
 'female_age_stdev',
 'female_age_sample_weight',
 'female_age_samples',
 'pct_own',
 'married',
 'married_snp',
 'separated',
 'divorced']
```

[31]:
```
#If the feature have less than 8 unique value then I am consdering as␣
 ↪categorical else it will be continuous
for i in null_col:
    print(i)
    if df_combined[i].nunique()>8:      #Continuous data
        df_combined[i].fillna(df_combined[i].median(),inplace=True)     #Bcz␣
 ↪median is not impacted by outlier
    else:df_combined[i].fillna(df_combined[i].mode()[0],inplace=True) ␣
 ↪#Categorical data
```

```
rent_mean
rent_median
rent_stdev
```

```
rent_sample_weight
rent_samples
rent_gt_10
rent_gt_15
rent_gt_20
rent_gt_25
rent_gt_30
rent_gt_35
rent_gt_40
rent_gt_50
hi_mean
hi_median
hi_stdev
hi_sample_weight
hi_samples
family_mean
family_median
family_stdev
family_sample_weight
family_samples
hc_mortgage_mean
hc_mortgage_median
hc_mortgage_stdev
hc_mortgage_sample_weight
hc_mortgage_samples
hc_mean
hc_median
hc_stdev
hc_samples
hc_sample_weight
home_equity_second_mortgage
second_mortgage
home_equity
debt
second_mortgage_cdf
home_equity_cdf
debt_cdf
hs_degree
hs_degree_male
hs_degree_female
male_age_mean
male_age_median
male_age_stdev
male_age_sample_weight
male_age_samples
female_age_mean
female_age_median
female_age_stdev
```

```
female_age_sample_weight
female_age_samples
pct_own
married
married_snp
separated
divorced
```

[32]:
```
df_combined.isna().sum()/len(df_combined)*100
```

[32]:
```
UID            0.0
SUMLEVEL       0.0
COUNTYID       0.0
STATEID        0.0
state          0.0
              …
married        0.0
married_snp    0.0
separated      0.0
divorced       0.0
split          0.0
Length: 80, dtype: float64
```

[33]:
```
df_combined.shape
```

[33]:
```
(38838, 80)
```

[34]:
```
# As we have seen above we have 123 unique UID which are common in both train␣
↪and test data. so duplicate UID removing them.
df_combined.drop_duplicates(subset=['UID'],inplace=True)
df_combined.shape
```

[34]:
```
(38715, 80)
```

Exploratory Data Analysis (EDA): Perform debt analysis. You may take the following steps: a. Explore the top 2,500 locations where the percentage of households with a 'second mortgage' is the highest and percent ownership is above 10 percent. Visualize using geo-map. You may keep the upper limit for the percent of households with a second mortgage to 50 percent

[35]:
```
top_2500_loc=df_train[(df_train['second_mortgage']<0.50) &
                      (df_train['pct_own']>0.10) ].
↪sort_values(by='second_mortgage', ascending=False).head(2500)
```

[36]:
```
top_2500_loc=top_2500_loc[['state','city','state_ab','place','lat','lng']]
top_2500_loc.head()
```

[36]:
```
              state         city state_ab            place        lat  \
11980  Massachusetts    Worcester       MA   Worcester City  42.254262
26018       New York       Corona       NY     Harbor Hills  40.751809
7829        Maryland  Glen Burnie       MD      Glen Burnie  39.127273
2077         Florida        Tampa       FL  Egypt Lake-leto  28.029063
1701        Illinois      Chicago       IL       Lincolnwood  41.967289

             lng
11980 -71.800347
26018 -73.853582
7829  -76.635265
2077  -82.495395
1701  -87.652434
```

[37]:
```
!pip install geopandas
import warnings
warnings.filterwarnings('ignore')
```

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: geopandas in /usr/local/lib/python3.10/site-
packages (0.11.0)
Requirement already satisfied: pandas>=1.0.0 in /usr/local/lib/python3.10/site-
packages (from geopandas) (1.5.3)
Requirement already satisfied: shapely<2,>=1.7 in
/usr/local/lib/python3.10/site-packages (from geopandas) (1.8.2)
Requirement already satisfied: fiona>=1.8 in /usr/local/lib/python3.10/site-
packages (from geopandas) (1.8.21)
Requirement already satisfied: pyproj>=2.6.1.post1 in
/usr/local/lib/python3.10/site-packages (from geopandas) (3.3.1)
Requirement already satisfied: packaging in /usr/local/lib/python3.10/site-
packages (from geopandas) (22.0)
Requirement already satisfied: attrs>=17 in /usr/local/lib/python3.10/site-
packages (from fiona>=1.8->geopandas) (23.1.0)
Requirement already satisfied: certifi in /usr/local/lib/python3.10/site-
packages (from fiona>=1.8->geopandas) (2022.6.15)
Requirement already satisfied: click>=4.0 in /usr/local/lib/python3.10/site-
packages (from fiona>=1.8->geopandas) (8.1.3)
Requirement already satisfied: cligj>=0.5 in /usr/local/lib/python3.10/site-
packages (from fiona>=1.8->geopandas) (0.7.2)
Requirement already satisfied: click-plugins>=1.0 in
/usr/local/lib/python3.10/site-packages (from fiona>=1.8->geopandas) (1.1.1)
Requirement already satisfied: six>=1.7 in /usr/local/lib/python3.10/site-
packages (from fiona>=1.8->geopandas) (1.16.0)
Requirement already satisfied: munch in /usr/local/lib/python3.10/site-packages
(from fiona>=1.8->geopandas) (2.5.0)
Requirement already satisfied: setuptools in /usr/local/lib/python3.10/site-
packages (from fiona>=1.8->geopandas) (58.1.0)
```

```
Requirement already satisfied: python-dateutil>=2.8.1 in
/usr/local/lib/python3.10/site-packages (from pandas>=1.0.0->geopandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/site-
packages (from pandas>=1.0.0->geopandas) (2022.1)
Requirement already satisfied: numpy>=1.21.0 in /usr/local/lib/python3.10/site-
packages (from pandas>=1.0.0->geopandas) (1.23.5)

[notice] A new release of pip is
available: 23.3 -> 24.0
[notice] To update, run:
pip install --upgrade pip
```

[38]:

```
import geopandas as gpd
gdf = gpd.GeoDataFrame(top_2500_loc, geometry=gpd.points_from_xy(x=top_2500_loc.
 ↪lng, y=top_2500_loc.lat))
gdf
```

[38]:

```
                 state         city state_ab             place        lat  \
11980    Massachusetts    Worcester       MA    Worcester City  42.254262
26018         New York       Corona       NY     Harbor Hills  40.751809
7829          Maryland  Glen Burnie       MD       Glen Burnie  39.127273
2077           Florida        Tampa       FL   Egypt Lake-leto  28.029063
1701          Illinois      Chicago       IL        Lincolnwood  41.967289
…                  …            …      …                  …        …
17914  North Carolina      Raleigh       NC      Raleigh City  35.757135
25642         Maryland    Baltimore       MD          Lochearn  39.353095
24443       California      Manteca       CA      Manteca City  37.732143
26671     Pennsylvania Philadelphia       PA Philadelphia City  40.039070
8377           Florida   Cutler Bay       FL        Cutler Bay  25.550391

              lng                    geometry
11980  -71.800347   POINT (-71.80035 42.25426)
26018  -73.853582   POINT (-73.85358 40.75181)
7829   -76.635265   POINT (-76.63526 39.12727)
2077   -82.495395   POINT (-82.49540 28.02906)
1701   -87.652434   POINT (-87.65243 41.96729)
…           …                          …
17914  -78.704288   POINT (-78.70429 35.75713)
25642  -76.733315   POINT (-76.73331 39.35310)
24443 -121.242902  POINT (-121.24290 37.73214)
26671  -75.125135   POINT (-75.12514 40.03907)
8377   -80.347791   POINT (-80.34779 25.55039)

[2500 rows x 7 columns]
```

- Use the following bad debt equation: Bad Debt = P (Second Mortgage ∩ Home Equity Loan)

Bad Debt = second_mortgage + home_equity - home_equity_second_mortgage

[39]:
```
df_combined['bad_debt'] = df_combined['second_mortgage'] +␣
↪df_combined['home_equity'] - df_combined['home_equity_second_mortgage']
df_combined.head(10)
```

[39]:
```
      UID  SUMLEVEL  COUNTYID  STATEID           state state_ab  \
0  267822       140        53       36        New York       NY
1  246444       140       141       18         Indiana       IN
2  245683       140        63       18         Indiana       IN
3  279653       140       127       72     Puerto Rico       PR
4  247218       140       161       20          Kansas       KS
5  221087       140        79        1         Alabama       AL
6  286689       140       337       48           Texas       TX
7  280558       140        45       45  South Carolina       SC
8  269138       140        81       36        New York       NY
9  227164       140        37        6      California       CA

                  city                       place   type primary  zip_code  \
0             Hamilton                    Hamilton   City   tract     13346
1           South Bend                    Roseland   City   tract     46616
2             Danville                    Danville   City   tract     46122
3             San Juan                    Guaynabo  Urban   tract       927
4            Manhattan              Manhattan City   City   tract     66502
5              Trinity                     Trinity   Town   tract     35673
6               Nocona                 Nocona City   Town   tract     76255
7              Taylors                  Tigerville   City   tract     29687
8  South Richmond Hill               New York City   City   tract     11419
9  Rancho Palos Verdes  Palos Verdes Estates City   City   tract     90275

   area_code        lat         lng        ALand   AWater   pop  male_pop  \
0        315  42.840812  -75.501524  202183361.0  1699120  5230      2612
1        574  41.701441  -86.266614    1560828.0   100363  2633      1349
2        317  39.792202  -86.515246   69561595.0   284193  6881      3643
3        787  18.396103  -66.104169    1105793.0        0  2700      1141
4        785  39.195573  -96.569366    2554403.0        0  5637      2586
5        256  34.519582  -87.151801   78402217.0   487343  5475      2564
6        940  33.842814  -97.784340  663218412.0  3122513  1947       994
7        864  35.136763  -82.294817  160338537.0  1912842  3476      1658
8        718  40.688610  -73.830597     157581.0        0  3530      1778
9        310  33.755867 -118.407590    3565039.0  1123792  4139      2086

   female_pop  rent_mean  rent_median  rent_stdev  rent_sample_weight  \
0        2618  769.38638        784.0   232.63967           272.34441
1        1284  804.87924        848.0   253.46747           312.58622
2        3238  742.77365        703.0   323.39011           291.85520
3        1559  803.42018        782.0   297.39258           259.30316
4        3051  938.56493        881.0   392.44096          1005.42886
5        2911  605.10246        684.0   230.15912           272.10405
```

6	953	661.76963	674.0	230.48928	125.45345
7	1818	784.36272	729.0	401.67621	94.04990
8	1752	1438.85143	1501.0	444.91460	76.80713
9	2053	2104.29576	1856.0	838.73396	48.12378

	rent_samples	rent_gt_10	rent_gt_15	rent_gt_20	rent_gt_25	rent_gt_30 \
0	362.0	0.86761	0.79155	0.59155	0.45634	0.42817
1	513.0	0.97410	0.93227	0.69920	0.69920	0.55179
2	378.0	0.95238	0.88624	0.79630	0.66667	0.39153
3	368.0	0.94693	0.87151	0.69832	0.61732	0.51397
4	1704.0	0.99286	0.98247	0.91688	0.84740	0.78247
5	287.0	0.80139	0.74564	0.74564	0.58188	0.23345
6	153.0	0.78431	0.71242	0.69935	0.66013	0.64052
7	124.0	1.00000	1.00000	1.00000	0.83871	0.83871
8	332.0	1.00000	0.93578	0.93578	0.82875	0.80428
9	391.0	0.96675	0.96675	0.91304	0.83632	0.64450

	rent_gt_35	rent_gt_40	rent_gt_50	universe_samples	used_samples \
0	0.18592	0.15493	0.12958	387	355
1	0.41235	0.39044	0.27888	542	502
2	0.39153	0.28307	0.15873	459	378
3	0.46927	0.35754	0.32961	438	358
4	0.60974	0.55455	0.44416	1725	1540
5	0.23345	0.23345	0.08014	359	287
6	0.64052	0.63399	0.63399	182	153
7	0.57258	0.52419	0.52419	146	124
8	0.71254	0.63609	0.43425	332	327
9	0.61637	0.58824	0.46036	418	391

	hi_mean	hi_median	hi_stdev	hi_sample_weight	hi_samples \
0	63125.28406	48120.0	49042.01206	1290.96240	2024.0
1	41931.92593	35186.0	31639.50203	838.74664	1127.0
2	84942.68317	74964.0	56811.62186	1155.20980	2488.0
3	48733.67116	37845.0	45100.54010	928.32193	1267.0
4	31834.15466	22497.0	34046.50907	1548.67477	1983.0
5	56912.14107	44873.0	40121.43988	1391.84595	2095.0
6	57872.25064	43761.0	52036.76167	523.50554	793.0
7	74276.59665	59504.0	68335.13833	741.68039	1398.0
8	69482.99919	44906.0	62747.61391	510.47908	804.0
9	119148.78380	98399.0	91993.70081	595.05678	1557.0

	family_mean	family_median	family_stdev	family_sample_weight \
0	67994.14790	53245.0	47667.30119	884.33516
1	50670.10337	43023.0	34715.57548	375.28798
2	95262.51431	85395.0	49292.67664	709.74925
3	56401.68133	44399.0	41082.90515	490.18479
4	54053.42396	50272.0	39609.12605	244.08903

```
5   60875.74450       48032.0   39750.92905          1064.00539
6   68632.82777       56405.0   48917.69947           332.78813
7   84050.66542       69529.0   60389.84940           492.90740
8   69349.72400       51123.0   56330.89786           469.48412
9  135702.84030      124446.0   76150.66062           321.70488

   family_samples  hc_mortgage_mean  hc_mortgage_median  hc_mortgage_stdev  \
0          1491.0        1414.80295              1223.0          641.22898
1           554.0         864.41390               784.0          482.27020
2          1889.0        1506.06758              1361.0          731.89394
3           729.0        1175.28642              1101.0          428.98751
4           395.0        1192.58759              1125.0          327.49674
5          1641.0        1137.05215              1141.0          377.26160
6           564.0        1339.98441              1016.0          734.84378
7          1027.0        1891.72540              1767.0         1109.67216
8           753.0        2941.26980              2792.0          892.72056
9          1155.0        3306.26240              3302.0         1137.02429

   hc_mortgage_sample_weight  hc_mortgage_samples    hc_mean  hc_median  \
0                  377.83135                867.0  570.01530      558.0
1                  316.88320                356.0  351.98293      336.0
2                  699.41354               1491.0  556.45986      532.0
3                  261.28471                437.0  288.04047      247.0
4                   76.61052                134.0  443.68855      444.0
5                  482.59538                759.0  338.91273      326.0
6                  132.40505                210.0  484.73723      435.0
7                  272.42931                622.0  391.71253      308.0
8                   59.90830                324.0  966.47211      954.0
9                  110.26388                702.0  971.13374      820.0

    hc_stdev  hc_samples  hc_sample_weight  home_equity_second_mortgage  \
0  270.11299       770.0         499.29293                      0.01588
1  125.40457       229.0         189.60606                      0.02222
2  184.42175       538.0         323.35354                      0.00000
3  185.55887       392.0         314.90566                      0.01086
4   76.12674       124.0          79.55556                      0.05426
5  157.69587       977.0         823.46465                      0.00000
6  291.44606       401.0         274.48824                      0.00000
7  291.09124       630.0         503.74471                      0.03355
8  224.02324       148.0          71.39181                      0.02331
9  491.04684       437.0         190.98724                      0.01229

   second_mortgage  home_equity     debt  second_mortgage_cdf  \
0          0.02077      0.08919  0.52963              0.43658
1          0.02222      0.04274  0.60855              0.42174
2          0.00000      0.09512  0.73484              1.00000
3          0.01086      0.01086  0.52714              0.53057
```

4	0.05426	0.05426	0.51938	0.18332
5	0.00000	0.05991	0.43721	1.00000
6	0.00000	0.00000	0.34370	1.00000
7	0.03355	0.09665	0.49681	0.31734
8	0.02331	0.11441	0.68644	0.41132
9	0.02809	0.21247	0.61633	0.36543

	home_equity_cdf	debt_cdf	hs_degree	hs_degree_male	hs_degree_female	\
0	0.49087	0.73341	0.89288	0.85880	0.92434	
1	0.70823	0.58120	0.90487	0.86947	0.94187	
2	0.46332	0.28704	0.94288	0.94616	0.93952	
3	0.82530	0.73727	0.91500	0.90755	0.92043	
4	0.65545	0.74967	1.00000	1.00000	1.00000	
5	0.62900	0.85639	0.80537	0.84111	0.77123	
6	1.00000	0.92825	0.84475	0.84056	0.84880	
7	0.45654	0.78390	0.86265	0.82111	0.90000	
8	0.37760	0.40090	0.76310	0.79669	0.73226	
9	0.09640	0.56452	0.98606	0.98635	0.98578	

	male_age_mean	male_age_median	male_age_stdev	male_age_sample_weight	\
0	42.48574	44.00000	22.97306	696.42136	
1	34.84728	32.00000	20.37452	323.90204	
2	39.38154	40.83333	22.89769	888.29730	
3	48.64749	48.91667	23.05968	274.98956	
4	26.07533	22.41667	11.84399	1296.89877	
5	38.81194	41.41667	21.52576	565.96518	
6	39.36384	40.00000	23.08255	245.14423	
7	46.63912	53.08333	22.60861	411.56696	
8	34.08697	30.66667	19.57786	460.16923	
9	45.09668	47.33333	24.60028	524.26788	

	male_age_samples	female_age_mean	female_age_median	female_age_stdev	\
0	2612.0	44.48629	45.33333	22.51276	
1	1349.0	36.48391	37.58333	23.43353	
2	3643.0	42.15810	42.83333	23.94119	
3	1141.0	47.77526	50.58333	24.32015	
4	2586.0	24.17693	21.58333	11.10484	
5	2564.0	37.06814	36.41667	22.88689	
6	994.0	42.18601	42.75000	23.40326	
7	1658.0	46.22879	49.75000	21.76534	
8	1778.0	37.27535	37.33333	20.27963	
9	2086.0	46.41178	50.50000	24.77630	

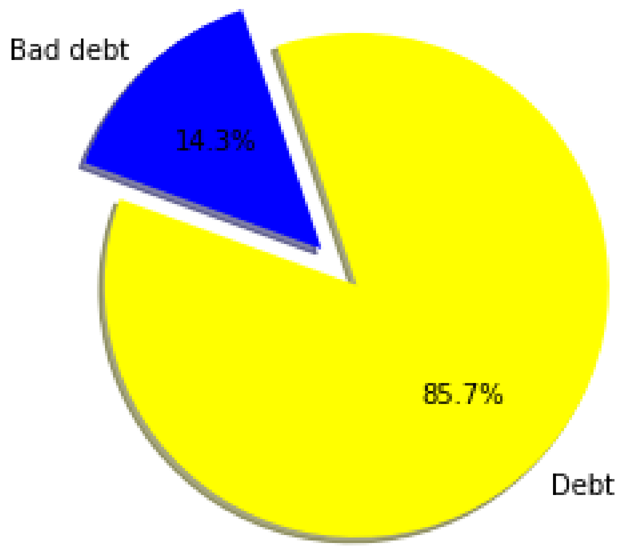
	female_age_sample_weight	female_age_samples	pct_own	married	\
0	685.33845	2618.0	0.79046	0.57851	
1	267.23367	1284.0	0.52483	0.34886	
2	707.01963	3238.0	0.85331	0.64745	

3	362.20193	1559.0	0.65037	0.47257
4	1854.48652	3051.0	0.13046	0.12356
5	708.76625	2911.0	0.83215	0.58503
6	240.99337	953.0	0.77658	0.63974
7	461.22601	1818.0	0.89931	0.73197
8	413.66078	1752.0	0.59602	0.52974
9	439.44640	2053.0	0.73651	0.65905

	married_snp	separated	divorced	split	bad_debt
0	0.01882	0.01240	0.08770	Train	0.09408
1	0.01426	0.01426	0.09030	Train	0.04274
2	0.02830	0.01607	0.10657	Train	0.09512
3	0.02021	0.02021	0.10106	Train	0.01086
4	0.00000	0.00000	0.03109	Train	0.05426
5	0.00680	0.00000	0.16910	Train	0.05991
6	0.01410	0.01410	0.09744	Train	0.00000
7	0.07850	0.05587	0.05587	Train	0.09665
8	0.13016	0.02309	0.05318	Train	0.11441
9	0.03370	0.00514	0.04911	Train	0.22827

Create pie charts to show overall debt and bad debt

[40]:
```
labels = 'Debt', 'Bad debt'
sizes = [df_combined['debt'].mean()*100, df_combined['bad_debt'].mean()*100]
colors = ['yellow', 'blue']
explode = (0.2, 0)  # explode 1st slice

#Plot
plt.pie(sizes,explode=explode,labels=labels, colors=colors,
autopct='%1.1f%%', shadow=True, startangle=160)

plt.axis('equal')
plt.show()
```



- Create Box and whisker plot and analyze the distribution for 2nd mortgage, home equity, good debt, and bad debt for different cities

[41]:
```
df_combined['good_debt']=df_combined['debt']-df_combined['bad_debt']
df_combined.head(2)
```

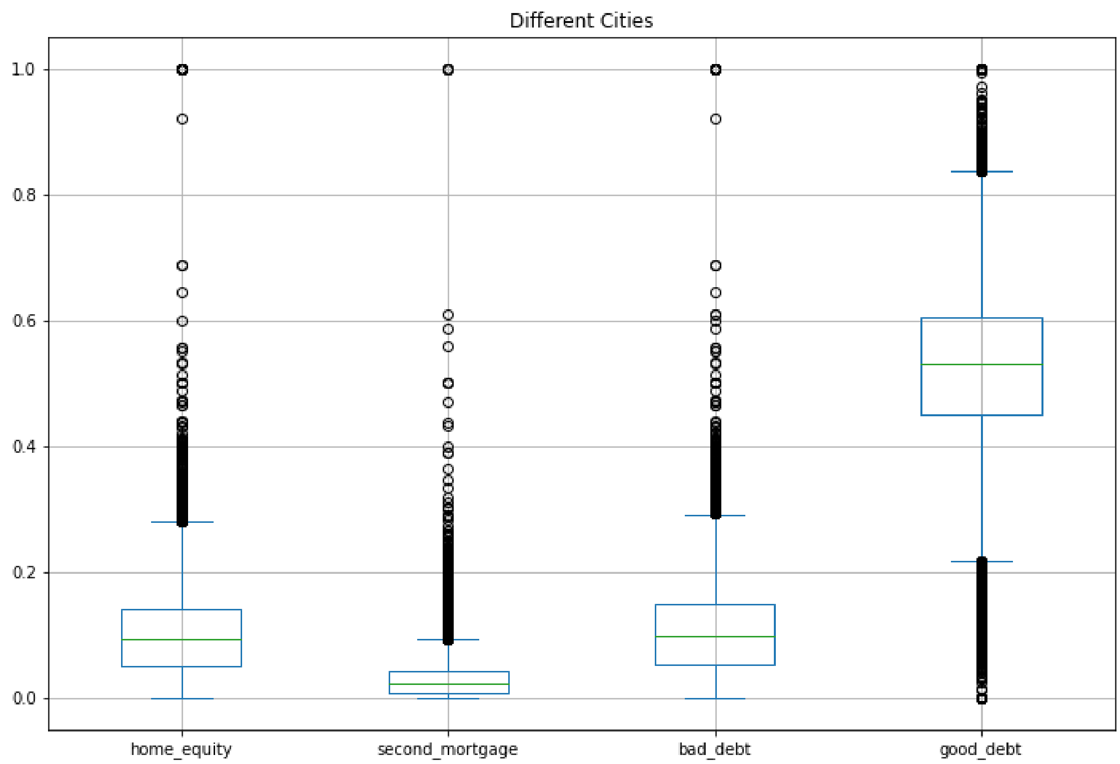
[41]:
```
      UID  SUMLEVEL  COUNTYID  STATEID      state state_ab        city  \
0  267822       140        53       36   New York       NY    Hamilton
1  246444       140       141       18    Indiana       IN  South Bend

      place  type primary  zip_code  area_code        lat        lng  \
0  Hamilton  City   tract     13346        315  42.840812 -75.501524
1  Roseland  City   tract     46616        574  41.701441 -86.266614

         ALand   AWater   pop  male_pop  female_pop  rent_mean  rent_median  \
0  202183361.0  1699120  5230      2612        2618  769.38638        784.0
1    1560828.0   100363  2633      1349        1284  804.87924        848.0

   rent_stdev  rent_sample_weight  rent_samples  rent_gt_10  rent_gt_15  \
0   232.63967           272.34441         362.0     0.86761     0.79155
1   253.46747           312.58622         513.0     0.97410     0.93227

   rent_gt_20  rent_gt_25  rent_gt_30  rent_gt_35  rent_gt_40  rent_gt_50  \
0     0.59155     0.45634     0.42817     0.18592     0.15493     0.12958
1     0.69920     0.69920     0.55179     0.41235     0.39044     0.27888

   universe_samples  used_samples      hi_mean  hi_median     hi_stdev  \
```

```
0             387           355  63125.28406    48120.0  49042.01206
1             542           502  41931.92593    35186.0  31639.50203

   hi_sample_weight  hi_samples  family_mean  family_median  family_stdev  \
0        1290.96240      2024.0  67994.14790        53245.0   47667.30119
1         838.74664      1127.0  50670.10337        43023.0   34715.57548

   family_sample_weight  family_samples  hc_mortgage_mean  hc_mortgage_median  \
0             884.33516          1491.0        1414.80295              1223.0
1             375.28798           554.0         864.41390               784.0

   hc_mortgage_stdev  hc_mortgage_sample_weight  hc_mortgage_samples  \
0          641.22898                  377.83135                867.0
1          482.27020                  316.88320                356.0

     hc_mean  hc_median   hc_stdev  hc_samples  hc_sample_weight  \
0  570.01530      558.0  270.11299       770.0         499.29293
1  351.98293      336.0  125.40457       229.0         189.60606

   home_equity_second_mortgage  second_mortgage  home_equity     debt  \
0                      0.01588          0.02077      0.08919  0.52963
1                      0.02222          0.02222      0.04274  0.60855

   second_mortgage_cdf  home_equity_cdf  debt_cdf  hs_degree  hs_degree_male  \
0              0.43658          0.49087   0.73341    0.89288         0.85880
1              0.42174          0.70823   0.58120    0.90487         0.86947

   hs_degree_female  male_age_mean  male_age_median  male_age_stdev  \
0           0.92434       42.48574             44.0        22.97306
1           0.94187       34.84728             32.0        20.37452

   male_age_sample_weight  male_age_samples  female_age_mean  \
0               696.42136            2612.0         44.48629
1               323.90204            1349.0         36.48391

   female_age_median  female_age_stdev  female_age_sample_weight  \
0           45.33333          22.51276                 685.33845
1           37.58333          23.43353                 267.23367

   female_age_samples  pct_own  married  married_snp  separated  divorced  \
0              2618.0  0.79046  0.57851      0.01882    0.01240    0.0877
1              1284.0  0.52483  0.34886      0.01426    0.01426    0.0903

   split  bad_debt  good_debt
0  Train   0.09408    0.43555
1  Train   0.04274    0.56581
```

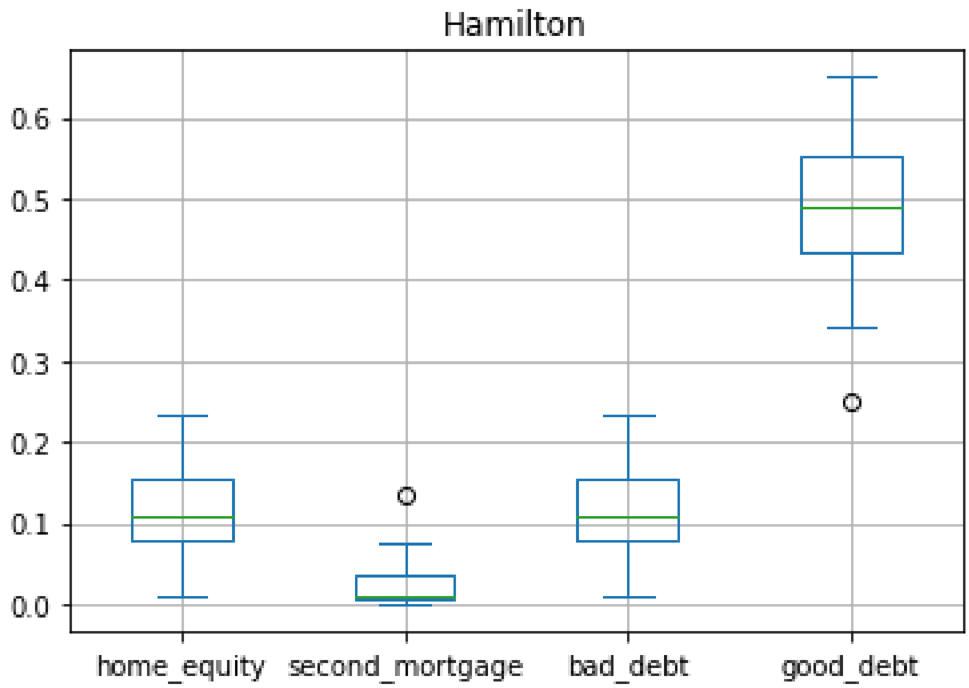
[42]:
```
df_combined.columns
```

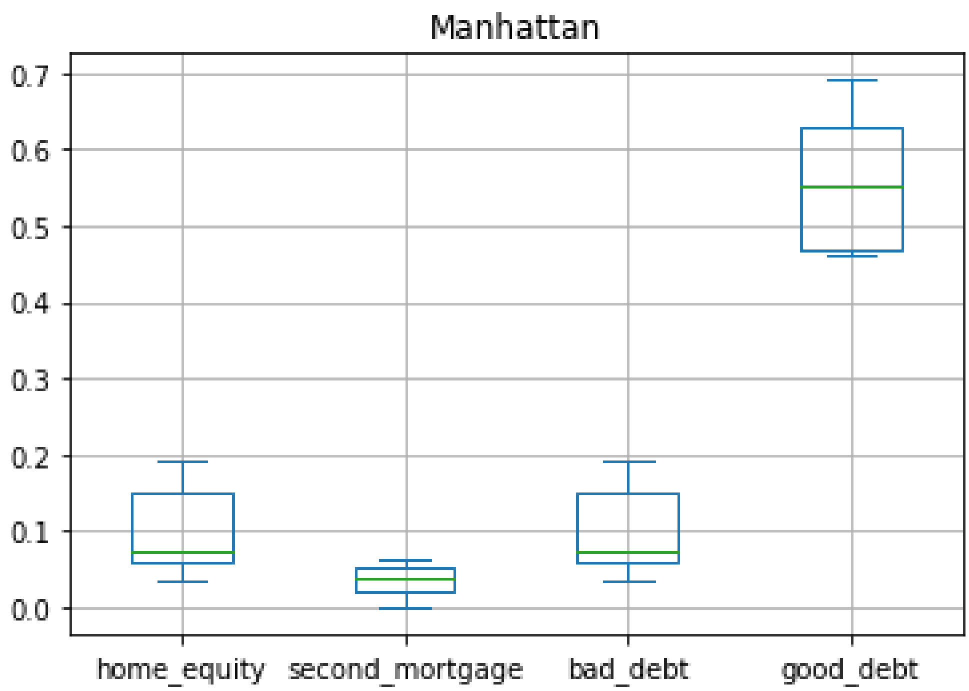
[42]:
```
Index(['UID', 'SUMLEVEL', 'COUNTYID', 'STATEID', 'state', 'state_ab', 'city',
       'place', 'type', 'primary', 'zip_code', 'area_code', 'lat', 'lng',
       'ALand', 'AWater', 'pop', 'male_pop', 'female_pop', 'rent_mean',
       'rent_median', 'rent_stdev', 'rent_sample_weight', 'rent_samples',
       'rent_gt_10', 'rent_gt_15', 'rent_gt_20', 'rent_gt_25', 'rent_gt_30',
       'rent_gt_35', 'rent_gt_40', 'rent_gt_50', 'universe_samples',
       'used_samples', 'hi_mean', 'hi_median', 'hi_stdev', 'hi_sample_weight',
       'hi_samples', 'family_mean', 'family_median', 'family_stdev',
       'family_sample_weight', 'family_samples', 'hc_mortgage_mean',
       'hc_mortgage_median', 'hc_mortgage_stdev', 'hc_mortgage_sample_weight',
       'hc_mortgage_samples', 'hc_mean', 'hc_median', 'hc_stdev', 'hc_samples',
       'hc_sample_weight', 'home_equity_second_mortgage', 'second_mortgage',
       'home_equity', 'debt', 'second_mortgage_cdf', 'home_equity_cdf',
       'debt_cdf', 'hs_degree', 'hs_degree_male', 'hs_degree_female',
       'male_age_mean', 'male_age_median', 'male_age_stdev',
       'male_age_sample_weight', 'male_age_samples', 'female_age_mean',
       'female_age_median', 'female_age_stdev', 'female_age_sample_weight',
       'female_age_samples', 'pct_own', 'married', 'married_snp', 'separated',
       'divorced', 'split', 'bad_debt', 'good_debt'],
      dtype='object')
```

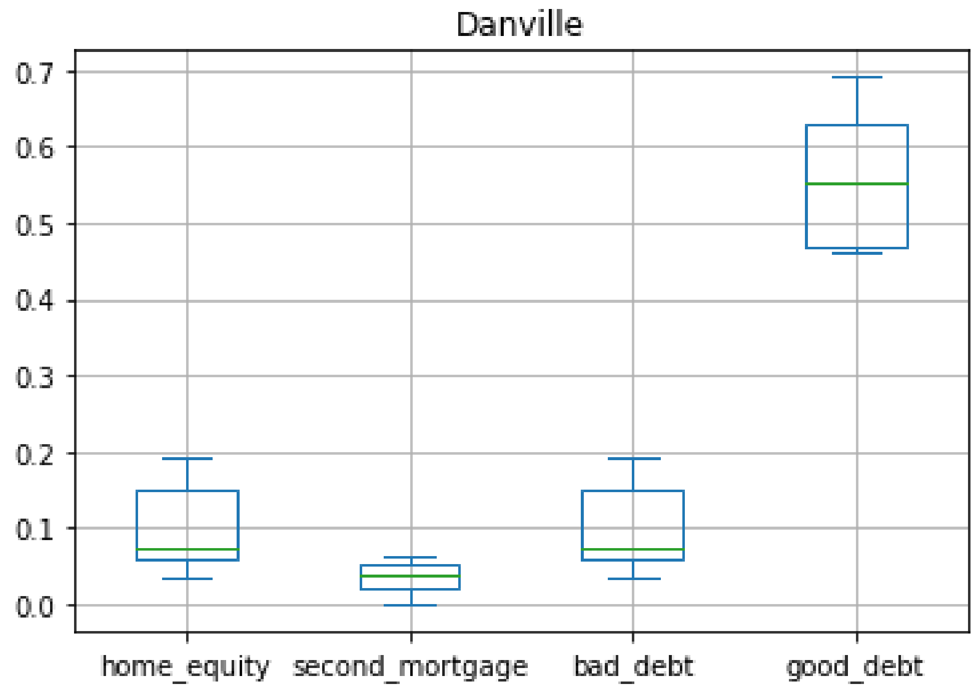
[43]:
```
diff_cities = df_combined[['home_equity','second_mortgage','bad_debt',␣
↪'good_debt']]
diff_cities.plot.box(figsize=(12,8),grid=True)
plt.title('Different Cities')
plt.show()
```



[44]:
```
hamilton = df_combined[df_combined['city']=='Hamilton']
hamilton = hamilton[['home_equity','second_mortgage','bad_debt', 'good_debt']]
hamilton.plot.box(grid=True)
plt.title('Hamilton')
plt.show()

Manhattan = df_combined[df_combined['city']=='Manhattan']
Manhattan = Manhattan[['home_equity','second_mortgage','bad_debt', 'good_debt']]
Manhattan.plot.box(grid=True)
plt.title('Manhattan')
plt.show()

Danville = df_combined[df_combined['city']=='Danville']
Danville = Danville[['home_equity','second_mortgage','bad_debt', 'good_debt']]
Manhattan.plot.box(grid=True)
plt.title('Danville')
plt.show()
```

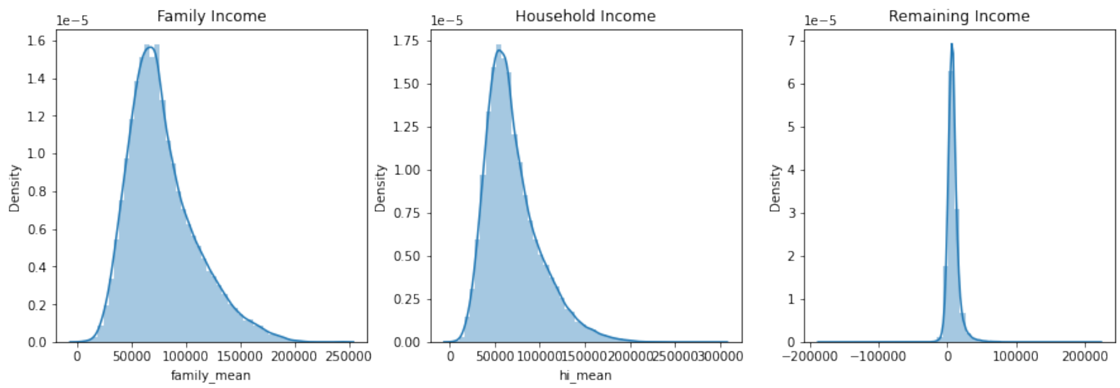

Hamilton
0.6
0.5
0.4
0.3
0.2
0.1
0.0
home_equity
second_mortgage
bad_debt
good_debt


Manhattan
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
home_equity
second_mortgage
bad_debt
good_debt



- Create a collated income distribution chart for family income, house hold income, and remaining income

[45]:
```
plt.figure(figsize=(15,10))

plt.subplot(2,3,1)
sns.distplot(df_combined['family_mean'])
plt.title('Family Income')

plt.subplot(2,3,2)
sns.distplot(df_combined['hi_mean'])
plt.title('Household Income')

plt.subplot(2,3,3)
sns.distplot(df_combined['family_mean']-df_combined['hi_mean'])
plt.title('Remaining Income')

plt.show()
```



5. Perform EDA and come out with insights into population density and age. You may have to derive new fields (make sure to weight averages for accurate measurements): • Use pop and ALand variables to create a new field called population density

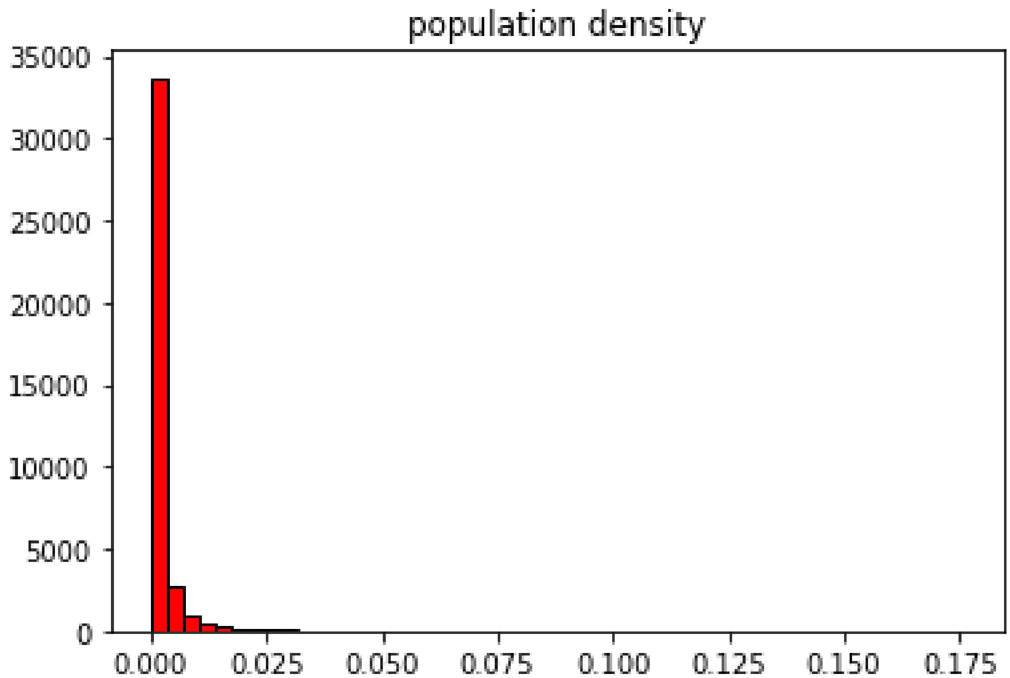
[46]:
```
df_combined['population_density'] = df_combined['pop']/df_combined['ALand']
```

[47]:
```
df_combined.head(2)
```

[47]:
```
      UID  SUMLEVEL  COUNTYID  STATEID     state state_ab        city  \
0  267822       140        53       36  New York       NY    Hamilton
1  246444       140       141       18   Indiana       IN  South Bend

      place  type primary  zip_code  area_code        lat        lng  \
0  Hamilton  City   tract     13346        315  42.840812 -75.501524
1  Roseland  City   tract     46616        574  41.701441 -86.266614

         ALand   AWater   pop  male_pop  female_pop  rent_mean  rent_median  \
0  202183361.0  1699120  5230      2612        2618  769.38638        784.0
1    1560828.0   100363  2633      1349        1284  804.87924        848.0

   rent_stdev  rent_sample_weight  rent_samples  rent_gt_10  rent_gt_15  \
0   232.63967           272.34441         362.0     0.86761     0.79155
1   253.46747           312.58622         513.0     0.97410     0.93227

   rent_gt_20  rent_gt_25  rent_gt_30  rent_gt_35  rent_gt_40  rent_gt_50  \
0     0.59155     0.45634     0.42817     0.18592     0.15493     0.12958
1     0.69920     0.69920     0.55179     0.41235     0.39044     0.27888

   universe_samples  used_samples      hi_mean  hi_median     hi_stdev  \
0               387           355  63125.28406    48120.0  49042.01206
1               542           502  41931.92593    35186.0  31639.50203

   hi_sample_weight  hi_samples  family_mean  family_median  family_stdev  \
```

```
0      1290.96240     2024.0  67994.14790       53245.0   47667.30119
1       838.74664     1127.0  50670.10337       43023.0   34715.57548

   family_sample_weight  family_samples  hc_mortgage_mean  hc_mortgage_median  \
0             884.33516          1491.0        1414.80295              1223.0
1             375.28798           554.0         864.41390               784.0

   hc_mortgage_stdev  hc_mortgage_sample_weight  hc_mortgage_samples  \
0          641.22898                  377.83135                867.0
1          482.27020                  316.88320                356.0

     hc_mean  hc_median   hc_stdev  hc_samples  hc_sample_weight  \
0  570.01530      558.0  270.11299       770.0         499.29293
1  351.98293      336.0  125.40457       229.0         189.60606

   home_equity_second_mortgage  second_mortgage  home_equity     debt  \
0                      0.01588          0.02077      0.08919  0.52963
1                      0.02222          0.02222      0.04274  0.60855

   second_mortgage_cdf  home_equity_cdf  debt_cdf  hs_degree  hs_degree_male  \
0              0.43658          0.49087   0.73341    0.89288         0.85880
1              0.42174          0.70823   0.58120    0.90487         0.86947

   hs_degree_female  male_age_mean  male_age_median  male_age_stdev  \
0           0.92434       42.48574             44.0        22.97306
1           0.94187       34.84728             32.0        20.37452

   male_age_sample_weight  male_age_samples  female_age_mean  \
0               696.42136            2612.0         44.48629
1               323.90204            1349.0         36.48391

   female_age_median  female_age_stdev  female_age_sample_weight  \
0           45.33333          22.51276                 685.33845
1           37.58333          23.43353                 267.23367

   female_age_samples  pct_own  married  married_snp  separated  divorced  \
0              2618.0  0.79046  0.57851      0.01882    0.01240    0.0877
1              1284.0  0.52483  0.34886      0.01426    0.01426    0.0903

   split  bad_debt  good_debt  population_density
0  Train   0.09408    0.43555            0.000026
1  Train   0.04274    0.56581            0.001687
```

[48]:
```python
plt.hist(df_combined['population_density'], bins=50, color='red',␣
 ↪edgecolor='black')
plt.title('population density')
plt.show()
```



- Use male_age_median, female_age_median, male_pop, and female_pop to create a new field called median age

[49]:
```
df_combined['median_age']=((df_combined['male_age_median'] *␣
↪df_combined['male_pop'])
+(df_combined['female_age_median']*df_combined['female_pop']))/
↪(df_combined['male_pop']+df_combined['female_pop'])
```

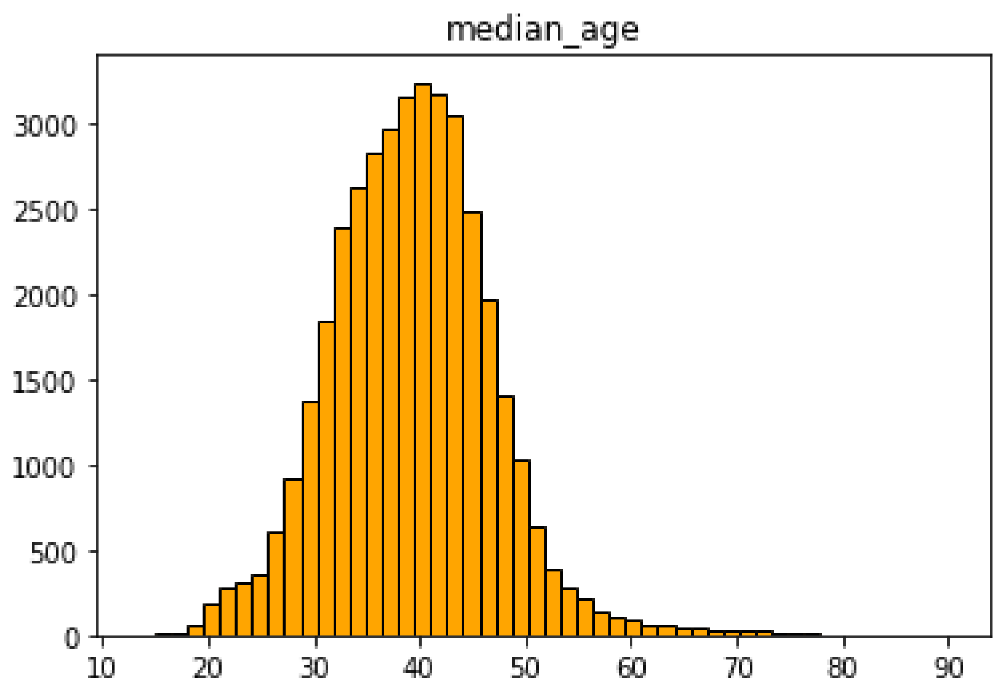
[50]:
```
df_combined.head(2)
```

[50]:
```
      UID  SUMLEVEL  COUNTYID  STATEID     state state_ab        city  \
0  267822       140        53       36  New York       NY    Hamilton
1  246444       140       141       18   Indiana       IN  South Bend

      place  type primary  zip_code  area_code        lat        lng  \
0  Hamilton  City   tract     13346        315  42.840812 -75.501524
1  Roseland  City   tract     46616        574  41.701441 -86.266614

         ALand   AWater   pop  male_pop  female_pop  rent_mean  rent_median  \
0  202183361.0  1699120  5230      2612        2618  769.38638        784.0
1    1560828.0   100363  2633      1349        1284  804.87924        848.0

   rent_stdev  rent_sample_weight  rent_samples  rent_gt_10  rent_gt_15  \
0   232.63967           272.34441         362.0     0.86761     0.79155
1   253.46747           312.58622         513.0     0.97410     0.93227
```

```
   rent_gt_20  rent_gt_25  rent_gt_30  rent_gt_35  rent_gt_40  rent_gt_50  \
0     0.59155     0.45634     0.42817     0.18592     0.15493     0.12958
1     0.69920     0.69920     0.55179     0.41235     0.39044     0.27888

   universe_samples  used_samples      hi_mean  hi_median     hi_stdev  \
0               387           355  63125.28406    48120.0  49042.01206
1               542           502  41931.92593    35186.0  31639.50203

   hi_sample_weight  hi_samples  family_mean  family_median  family_stdev  \
0        1290.96240      2024.0  67994.14790        53245.0   47667.30119
1         838.74664      1127.0  50670.10337        43023.0   34715.57548

   family_sample_weight  family_samples  hc_mortgage_mean  hc_mortgage_median  \
0             884.33516          1491.0        1414.80295              1223.0
1             375.28798           554.0         864.41390               784.0

   hc_mortgage_stdev  hc_mortgage_sample_weight  hc_mortgage_samples  \
0          641.22898                  377.83135                867.0
1          482.27020                  316.88320                356.0

     hc_mean  hc_median   hc_stdev  hc_samples  hc_sample_weight  \
0  570.01530      558.0  270.11299       770.0         499.29293
1  351.98293      336.0  125.40457       229.0         189.60606

   home_equity_second_mortgage  second_mortgage  home_equity     debt  \
0                      0.01588          0.02077      0.08919  0.52963
1                      0.02222          0.02222      0.04274  0.60855

   second_mortgage_cdf  home_equity_cdf  debt_cdf  hs_degree  hs_degree_male  \
0              0.43658          0.49087   0.73341    0.89288         0.85880
1              0.42174          0.70823   0.58120    0.90487         0.86947

   hs_degree_female  male_age_mean  male_age_median  male_age_stdev  \
0           0.92434       42.48574             44.0        22.97306
1           0.94187       34.84728             32.0        20.37452

   male_age_sample_weight  male_age_samples  female_age_mean  \
0               696.42136            2612.0         44.48629
1               323.90204            1349.0         36.48391

   female_age_median  female_age_stdev  female_age_sample_weight  \
0           45.33333          22.51276                 685.33845
1           37.58333          23.43353                 267.23367

   female_age_samples  pct_own  married  married_snp  separated  divorced  \
0              2618.0  0.79046  0.57851      0.01882    0.01240    0.0877
```

```
1              1284.0   0.52483   0.34886      0.01426    0.01426    0.0903

   split  bad_debt  good_debt  population_density  median_age
0  Train   0.09408    0.43555            0.000026   44.667430
1  Train   0.04274    0.56581            0.001687   34.722748
```

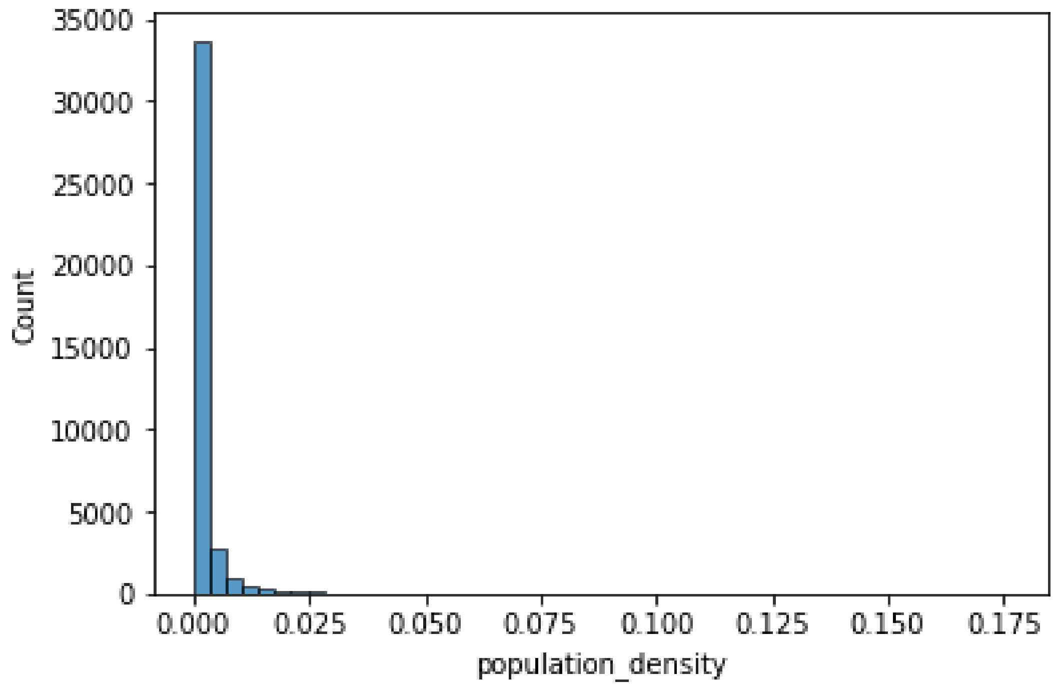
[51]:
```python
plt.hist(df_combined['median_age'], bins=50, color='orange', edgecolor='black')
plt.title('median_age')
plt.show()
```



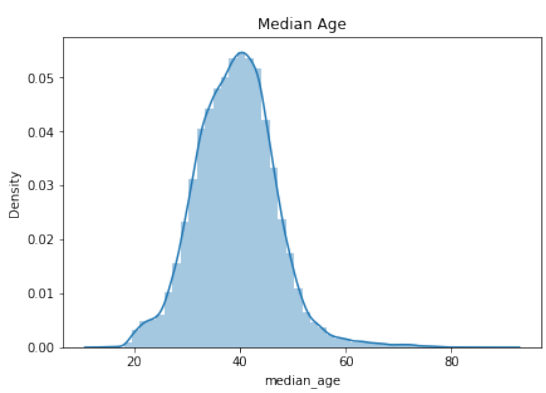
- Visualize the findings using appropriate chart type

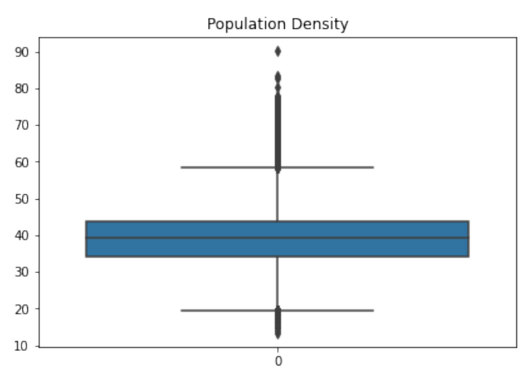
[52]:
```python
sns.histplot(df_combined['population_density'], bins=50)
```

[52]: <AxesSubplot: xlabel='population_density', ylabel='Count'>



[53]:
```
plt.figure(figsize=(15,10))
plt.subplot(2,2,1)
sns.distplot(df_combined['median_age'])
plt.title('Median Age')
plt.subplot(2,2,2)
sns.boxplot(df_combined['median_age'])
plt.title('Population Density')
plt.show()
```





6. Create bins for population into a new variable by selecting appropriate class interval so that the number of categories don't exceed 5 for the ease of analysis.

[54]:
```
df_combined['pop_bins']=pd.cut(df_combined['pop'],bins=5,labels=['very␣
↪low','low','medium','high','very high'])
df_combined['pop_bins'].value_counts()
```

[54]:
```
very low     38350
low            348
medium          12
high             4
very high        1
Name: pop_bins, dtype: int64
```

a. Analyze the married, separated, and divorced population for these population brackets

[55]:
```
df_combined.groupby(by='pop_bins')[['married','separated','divorced']].count()
```

[55]:

pop_bins	married	separated	divorced
very low	38350	38350	38350
low	348	348	348
medium	12	12	12
high	4	4	4
very high	1	1	1

[56]:
```
df_combined.groupby(by='pop_bins')[['married','separated','divorced']].
↪agg(["mean", "median"])
```

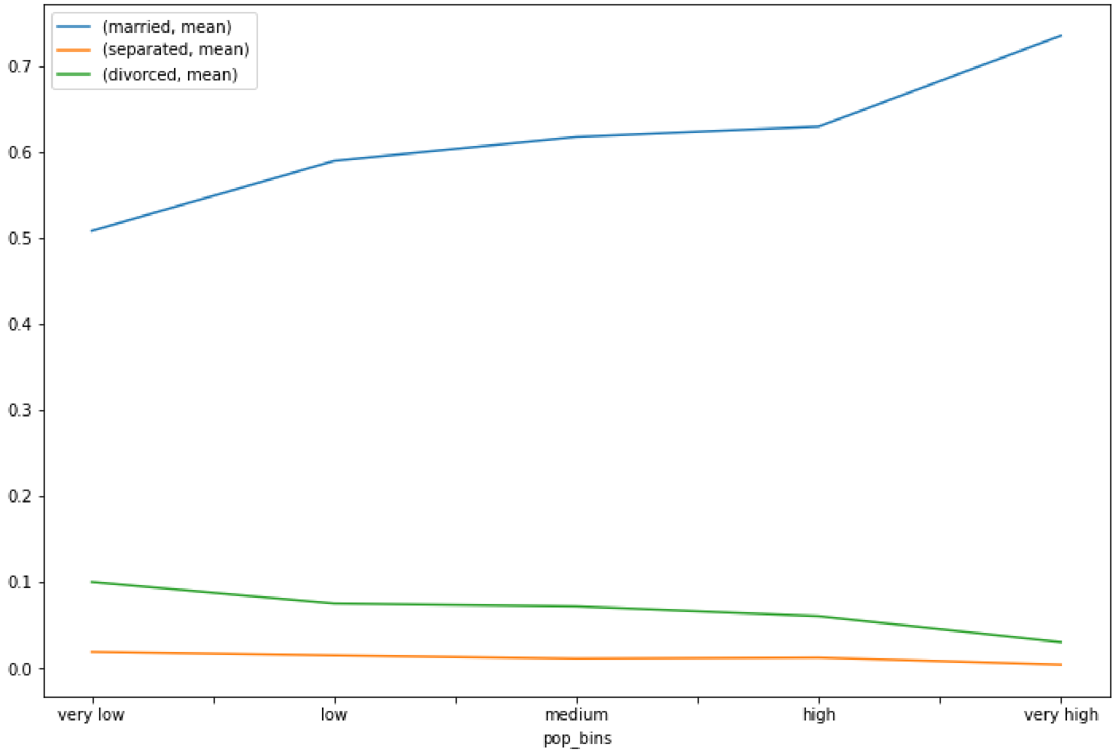
[56]:

	married		separated		divorced	
	mean	median	mean	median	mean	median
pop_bins						
very low	0.508002	0.526710	0.019127	0.013580	0.100325	0.09510
low	0.589247	0.601815	0.014929	0.010255	0.075192	0.06934
medium	0.617047	0.605765	0.011203	0.007745	0.071870	0.06909
high	0.629132	0.675095	0.012372	0.007340	0.060562	0.05987
very high	0.734740	0.734740	0.004050	0.004050	0.030360	0.03036

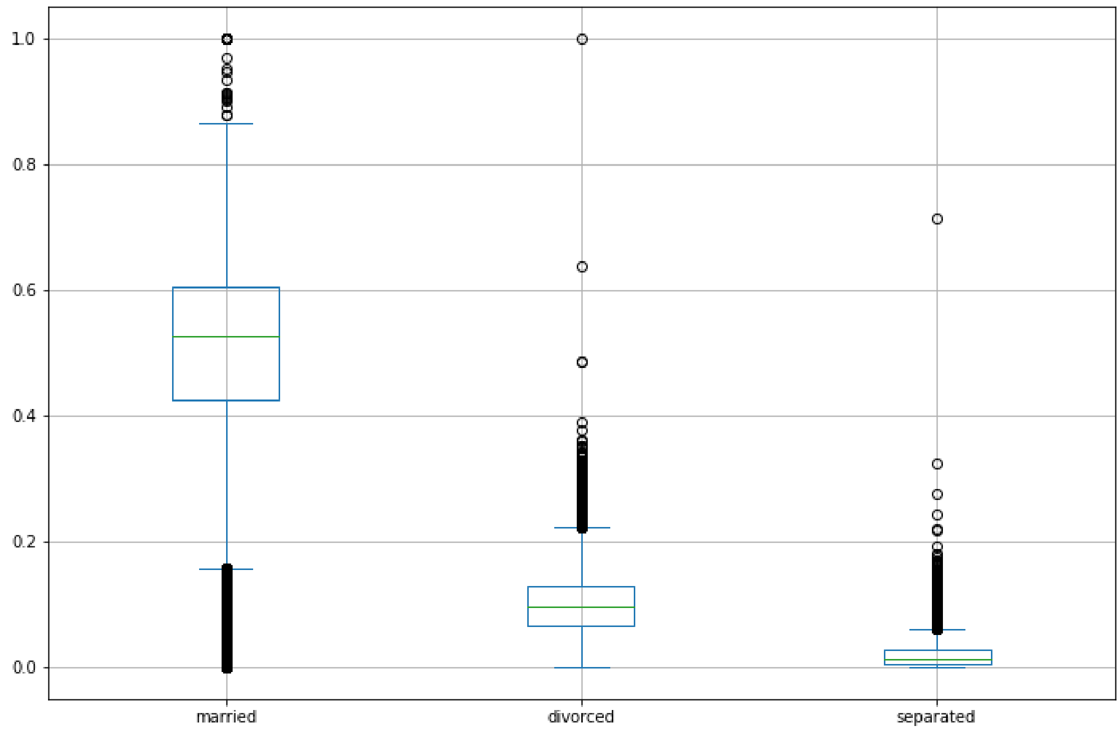
- Visualize using appropriate chart type

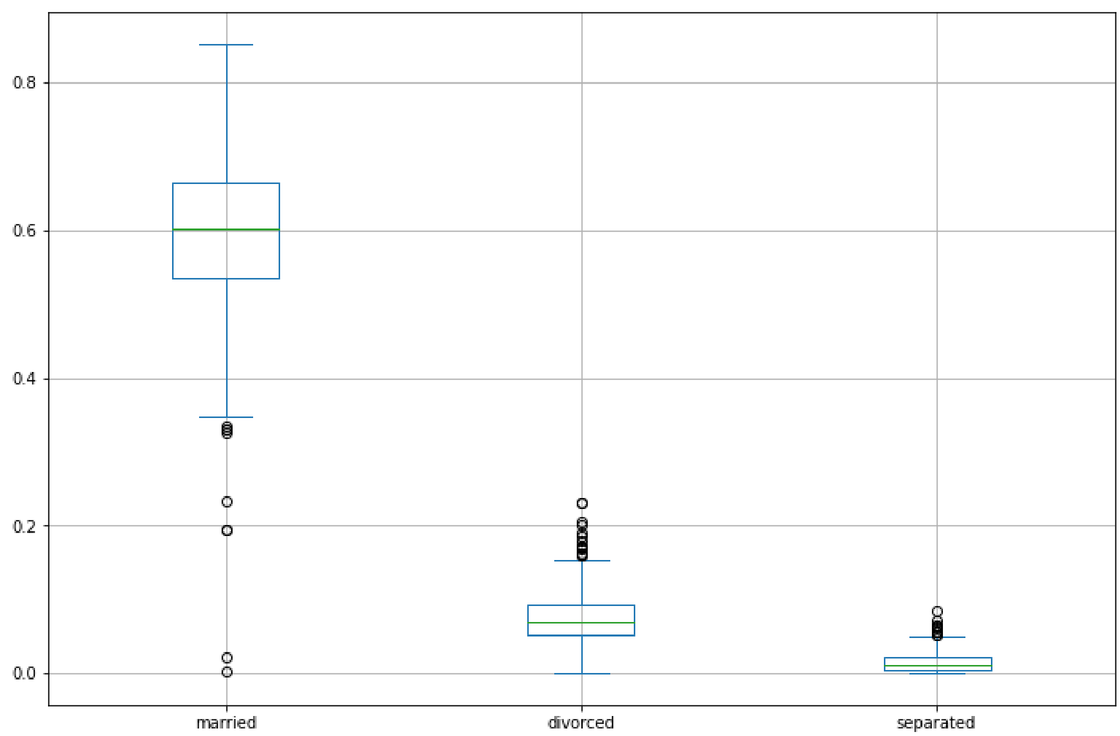
[57]:
```
plt.figure(figsize=(10,5))
pop_bin_married=df_combined.
↪groupby(by='pop_bins')[['married','separated','divorced']].agg(["mean"])
pop_bin_married.plot(figsize=(12,8))
plt.legend(loc='best')
plt.show()
```

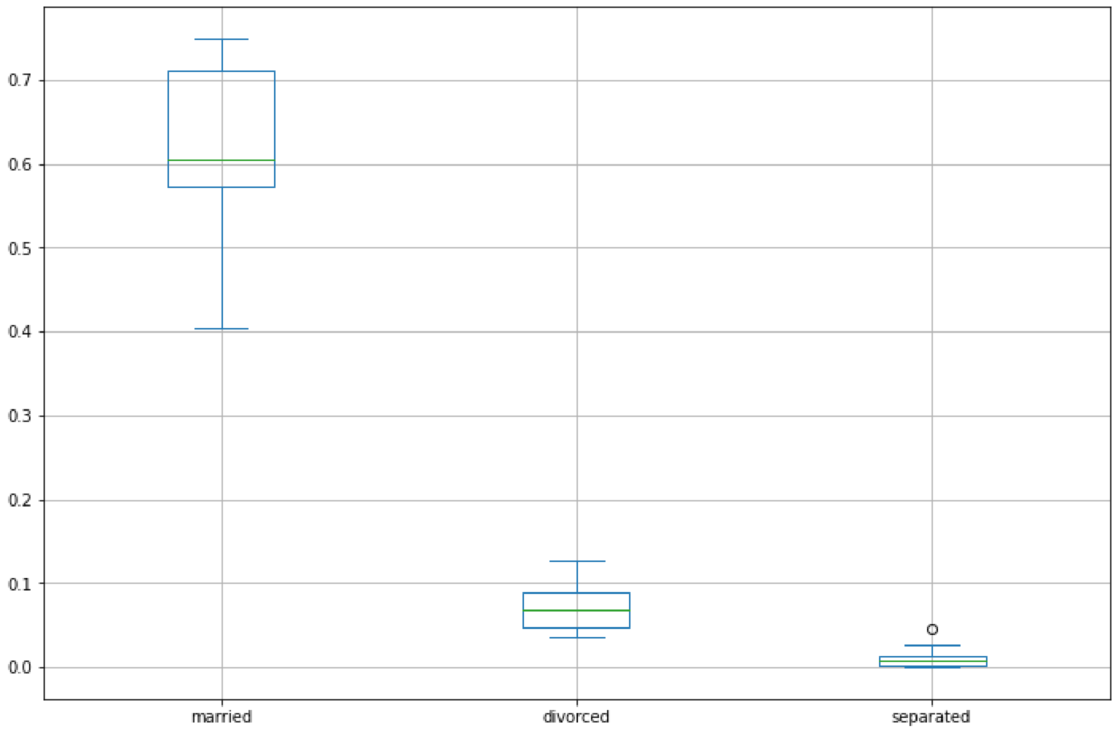
<Figure size 720x360 with 0 Axes>

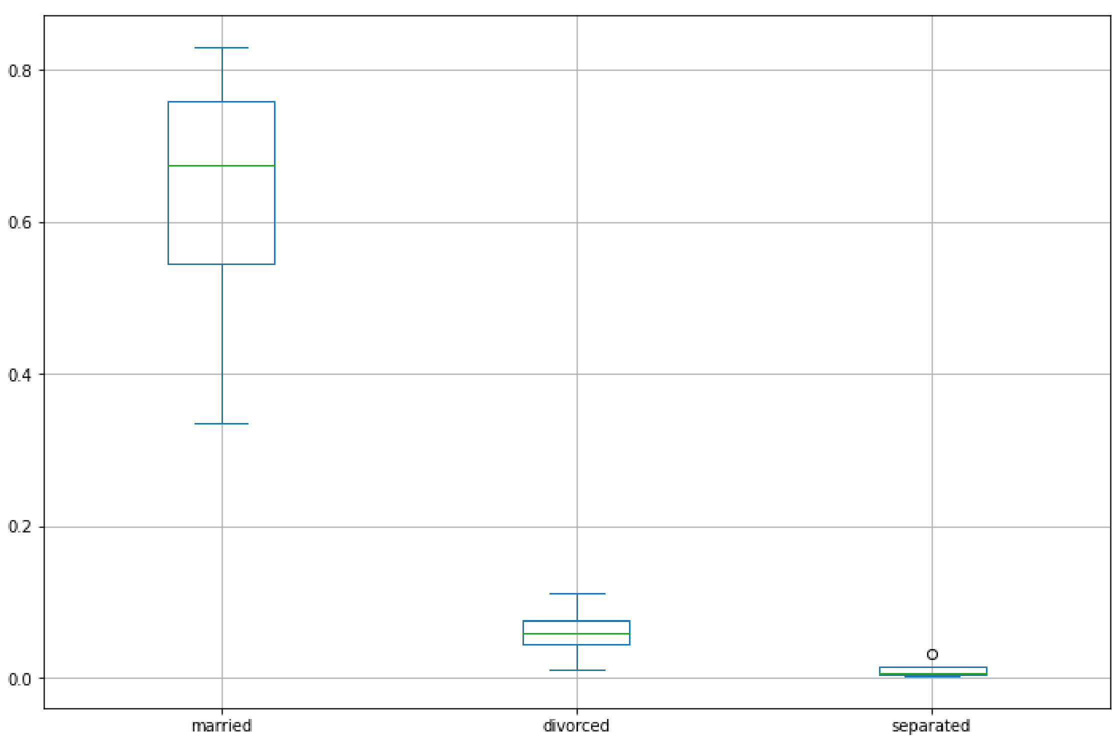


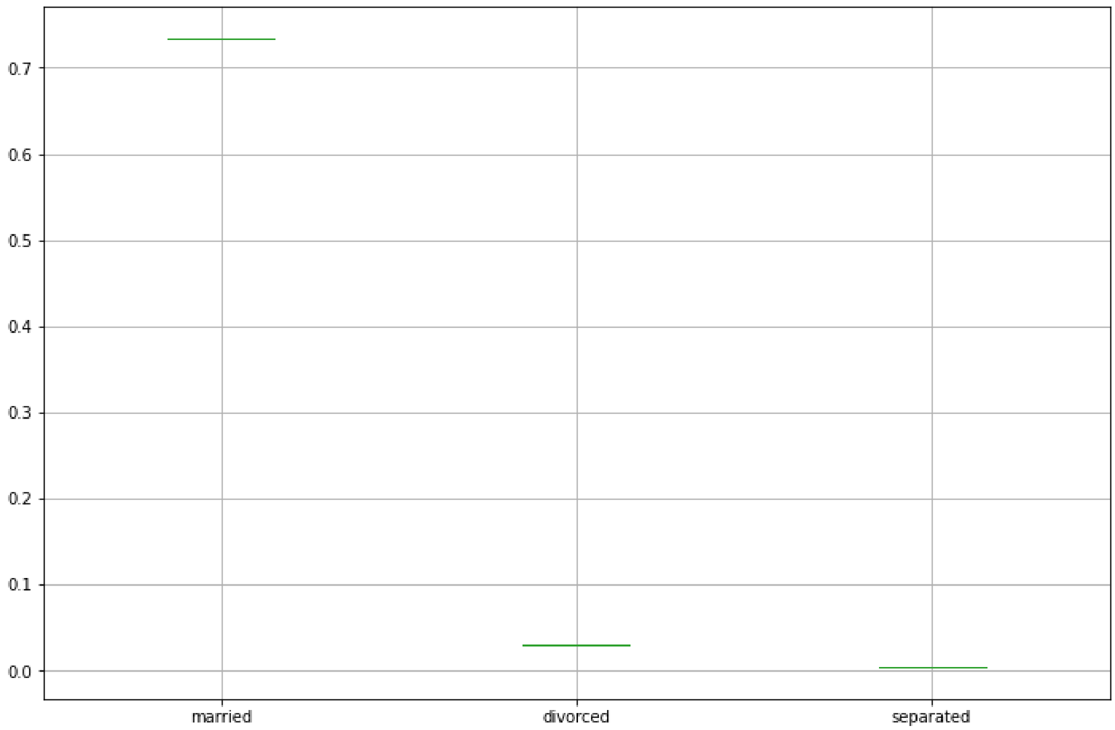
[58]:
```
df_combined.groupby(by='pop_bins')[['married','divorced', 'separated']].plot.
    ↪box(figsize=(12,8),grid='True')
plt.show()
```


1.0
0.8
0.6
0.4
0.2
0.0
married
divorced
separated


0.8
0.6
0.4
0.2
0.0
married
divorced
separated


0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
married
divorced
separated


0.8
0.6
0.4
0.2
0.0
married
divorced
separated



7. Please detail your observations for rent as a percentage of income at an overall level, and for different states.

[59]:
```
rent_state_mean = df_combined.groupby(by='state')['rent_mean'].agg(["mean"])
rent_state_mean.head(10)
```

[59]:

state	mean
Alabama	765.872568
Alaska	1190.093590
Arizona	1084.510968
Arkansas	716.544999
California	1466.020481
Colorado	1192.839715
Connecticut	1313.616792
Delaware	1102.107261
District of Columbia	1454.149546
Florida	1142.518799

[60]:
```
income_state_mean=df_combined.groupby(by='state')['family_mean'].agg(["mean"])
income_state_mean.head(10)
```

[60]:

```
                                mean
state
```

```
Alabama                  65311.673394
Alaska                   91911.137520
Arizona                  73014.362099
Arkansas                 64234.797753
California               87711.782288
Colorado                 87728.719535
Connecticut             103260.529612
Delaware                 84031.947372
District of Columbia    107123.968906
Florida                  72490.529377
```

[61]:
```
rent_perc_of_income=rent_state_mean['mean']/income_state_mean['mean']*100
rent_perc_of_income.head(10)
```

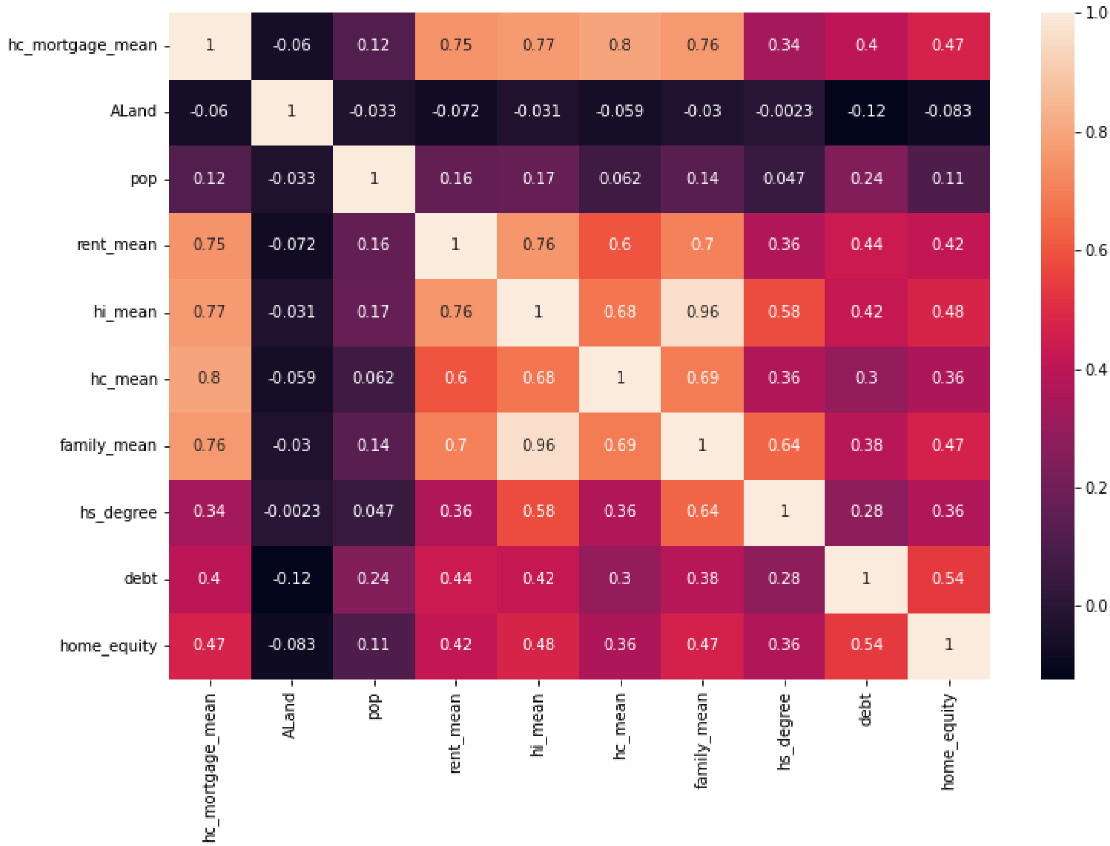
[61]:
```
state
Alabama                 1.172643
Alaska                  1.294831
Arizona                 1.485339
Arkansas                1.115509
California              1.671407
Colorado                1.359691
Connecticut             1.272138
Delaware                1.311534
District of Columbia    1.357446
Florida                 1.576094
Name: mean, dtype: float64
```

[62]:
```
sum(df_combined['rent_mean'])/sum(df_combined['family_mean'])
```

[62]:
```
0.013351500156256637
```

8. Perform correlation analysis for all the relevant variables by creating a heatmap. Describe your findings.

[63]:
```
plt.figure(figsize=(12,8))
sns.
 ↪heatmap(data=df_combined[['hc_mortgage_mean','ALand','pop','rent_mean','hi_mean','hc_mean',
                             'hs_degree','debt','home_equity']].corr(),annot=True)
plt.show()
```



[64]:
```
df_combined.to_csv('P-1.csv')
```

rent_mean, hi_mean, hc_mean, family_mean has a good correlation with the target i.e- hc_mortagage_mean

[65]:
```
train = df_combined[df_combined['split'] == 'Train']
test = df_combined[df_combined['split'] == 'Test']
```

[66]:
```
train.head(2)
```

[66]:
```
      UID  SUMLEVEL  COUNTYID  STATEID     state state_ab        city  \
0  267822       140        53       36  New York       NY    Hamilton
1  246444       140       141       18   Indiana       IN  South Bend

      place  type primary  zip_code  area_code        lat        lng  \
0  Hamilton  City   tract     13346        315  42.840812 -75.501524
1  Roseland  City   tract     46616        574  41.701441 -86.266614

         ALand   AWater   pop  male_pop  female_pop  rent_mean  rent_median  \
0  202183361.0  1699120  5230      2612        2618  769.38638        784.0
```

```
1   1560828.0   100363  2633      1349        1284  804.87924        848.0

   rent_stdev  rent_sample_weight  rent_samples  rent_gt_10  rent_gt_15  \
0   232.63967           272.34441         362.0     0.86761     0.79155
1   253.46747           312.58622         513.0     0.97410     0.93227

   rent_gt_20  rent_gt_25  rent_gt_30  rent_gt_35  rent_gt_40  rent_gt_50  \
0     0.59155     0.45634     0.42817     0.18592     0.15493     0.12958
1     0.69920     0.69920     0.55179     0.41235     0.39044     0.27888

   universe_samples  used_samples      hi_mean  hi_median     hi_stdev  \
0               387           355  63125.28406    48120.0  49042.01206
1               542           502  41931.92593    35186.0  31639.50203

   hi_sample_weight  hi_samples  family_mean  family_median  family_stdev  \
0        1290.96240      2024.0  67994.14790        53245.0   47667.30119
1         838.74664      1127.0  50670.10337        43023.0   34715.57548

   family_sample_weight  family_samples  hc_mortgage_mean  hc_mortgage_median  \
0             884.33516          1491.0        1414.80295              1223.0
1             375.28798           554.0         864.41390               784.0

   hc_mortgage_stdev  hc_mortgage_sample_weight  hc_mortgage_samples  \
0          641.22898                  377.83135                867.0
1          482.27020                  316.88320                356.0

     hc_mean  hc_median   hc_stdev  hc_samples  hc_sample_weight  \
0  570.01530      558.0  270.11299       770.0         499.29293
1  351.98293      336.0  125.40457       229.0         189.60606

   home_equity_second_mortgage  second_mortgage  home_equity     debt  \
0                      0.01588          0.02077      0.08919  0.52963
1                      0.02222          0.02222      0.04274  0.60855

   second_mortgage_cdf  home_equity_cdf  debt_cdf  hs_degree  hs_degree_male  \
0              0.43658          0.49087   0.73341    0.89288         0.85880
1              0.42174          0.70823   0.58120    0.90487         0.86947

   hs_degree_female  male_age_mean  male_age_median  male_age_stdev  \
0           0.92434       42.48574             44.0        22.97306
1           0.94187       34.84728             32.0        20.37452

   male_age_sample_weight  male_age_samples  female_age_mean  \
0              696.42136            2612.0         44.48629
1              323.90204            1349.0         36.48391

   female_age_median  female_age_stdev  female_age_sample_weight  \
```

```
0           45.33333         22.51276                   685.33845
1           37.58333         23.43353                   267.23367

   female_age_samples  pct_own  married  married_snp  separated  divorced  \
0              2618.0  0.79046  0.57851      0.01882    0.01240    0.0877
1              1284.0  0.52483  0.34886      0.01426    0.01426    0.0903

   split  bad_debt  good_debt  population_density  median_age  pop_bins
0  Train   0.09408    0.43555            0.000026   44.667430  very low
1  Train   0.04274    0.56581            0.001687   34.722748  very low
```

[67]:
```
test.head(2)
```

[67]:
```
          UID  SUMLEVEL  COUNTYID  STATEID     state state_ab     city  \
27161  255504       140       163       26  Michigan       MI  Detroit
27162  252676       140         1       23     Maine       ME   Auburn

                      place  type primary  zip_code  area_code        lat  \
27161  Dearborn Heights City   CDP   tract     48239        313  42.346422
27162             Auburn City  City   tract      4210        207  44.100724

             lng       ALand   AWater   pop  male_pop  female_pop  rent_mean  \
27161 -83.252823   2711280.0    39555  3417      1479        1938  858.57169
27162 -70.257832  14778785.0  2705204  3796      1846        1950  832.68625

       rent_median  rent_stdev  rent_sample_weight  rent_samples  rent_gt_10  \
27161        859.0   232.39082           276.07497         424.0         1.0
27162        750.0   267.22342           183.32299         245.0         1.0

       rent_gt_15  rent_gt_20  rent_gt_25  rent_gt_30  rent_gt_35  rent_gt_40  \
27161     0.95696     0.85316     0.85316     0.85316     0.85316     0.76962
27162     1.00000     0.86611     0.67364     0.30962     0.30962     0.30962

       rent_gt_50  universe_samples  used_samples      hi_mean  hi_median  \
27161     0.63544               435           395  48899.52121    38746.0
27162     0.27197               275           239  72335.33234    61008.0

          hi_stdev  hi_sample_weight  hi_samples  family_mean  family_median  \
27161  44392.20902         798.02401      1180.0  53802.87122        45167.0
27162  51895.81159         922.82969      1722.0  85642.22095        74759.0

       family_stdev  family_sample_weight  family_samples  hc_mortgage_mean  \
27161   43756.56479             464.30972           769.0        1139.24548
27162   49156.72870             482.99945          1147.0        1533.25988

       hc_mortgage_median  hc_mortgage_stdev  hc_mortgage_sample_weight  \
27161              1109.0          336.47710                  262.67011
```

```
27162             1438.0         536.61118                  373.96188

       hc_mortgage_samples    hc_mean  hc_median   hc_stdev  hc_samples  \
27161                474.0  488.51323      436.0  192.75147       271.0
27162                937.0  661.31296      668.0  201.31365       510.0

       hc_sample_weight  home_equity_second_mortgage  second_mortgage  \
27161         189.18182                      0.06443          0.06443
27162         279.69697                      0.01175          0.01175

       home_equity     debt  second_mortgage_cdf  home_equity_cdf  debt_cdf  \
27161      0.07651  0.63624              0.14111          0.55087   0.51965
27162      0.14375  0.64755              0.52310          0.26442   0.49359

       hs_degree  hs_degree_male  hs_degree_female  male_age_mean  \
27161    0.91047         0.92010           0.90391       33.37131
27162    0.94290         0.92832           0.95736       43.88680

       male_age_median  male_age_stdev  male_age_sample_weight  \
27161         27.83333        22.36768               334.30978
27162         46.08333        22.90302               427.10824

       male_age_samples  female_age_mean  female_age_median  female_age_stdev  \
27161            1479.0         34.78682           33.75000          21.58531
27162            1846.0         44.23451           46.66667          22.37036

       female_age_sample_weight  female_age_samples  pct_own  married  \
27161                 416.48097              1938.0  0.70252  0.28217
27162                 532.03505              1950.0  0.85128  0.64221

       married_snp  separated  divorced split  bad_debt  good_debt  \
27161      0.05910    0.03813   0.14299  Test   0.07651    0.55973
27162      0.02338    0.00000   0.13377  Test   0.14375    0.50380

       population_density  median_age  pop_bins
27161            0.001260   31.189053  very low
27162            0.000257   46.382991  very low
```

0.1.1 Project Task: Week 2

0.1.2 Data Pre-processing:

1. The economic multivariate data has a significant number of measured variables. The goal is to find where the measured variables depend on a number of smaller unobserved common factors or latent variables.

2. Each variable is assumed to be dependent upon a linear combination of the common factors, and the coefficients are known as loadings. Each measured variable also includes a component

due to independent random variability, known as "specific variance" because it is specific to one variable. Obtain the common factors and then plot the loadings. Use factor analysis to find latent variables in our dataset and gain insight into the linear relationships in the data.

Following are the list of latent variables:

- Highschool graduation rates
- Median population age
- Second mortgage statistics
- Percent own
- Bad debt expense

[68]:
```
!pip install factor_analyzer
```

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: factor_analyzer in ./.local/lib/python3.10/site-
packages (0.5.1)
Requirement already satisfied: pandas in /usr/local/lib/python3.10/site-packages
(from factor_analyzer) (1.5.3)
Requirement already satisfied: scipy in /usr/local/lib/python3.10/site-packages
(from factor_analyzer) (1.9.3)
Requirement already satisfied: numpy in /usr/local/lib/python3.10/site-packages
(from factor_analyzer) (1.23.5)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.10/site-
packages (from factor_analyzer) (1.3.1)
Requirement already satisfied: python-dateutil>=2.8.1 in
/usr/local/lib/python3.10/site-packages (from pandas->factor_analyzer) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/site-
packages (from pandas->factor_analyzer) (2022.1)
Requirement already satisfied: joblib>=1.1.1 in /usr/local/lib/python3.10/site-
packages (from scikit-learn->factor_analyzer) (1.2.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in
/usr/local/lib/python3.10/site-packages (from scikit-learn->factor_analyzer)
(3.1.0)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/site-
packages (from python-dateutil>=2.8.1->pandas->factor_analyzer) (1.16.0)

[notice] A new release of pip is
available: 23.3 -> 24.0
[notice] To update, run:
pip install --upgrade pip
```

[69]:
```
import numpy as np
from sklearn.decomposition import FactorAnalysis
from factor_analyzer import FactorAnalyzer
```

[70]:
```
df_train.describe().T
```

[70]:
```
                count           mean           std       min            25%  \
UID           27161.0  257328.592209  21342.667653  220342.0  238826.000000
BLOCKID           0.0            NaN           NaN       NaN            NaN
SUMLEVEL      27161.0     140.000000      0.000000     140.0     140.000000
COUNTYID      27161.0      85.660322     98.373195       1.0      29.000000
STATEID       27161.0      28.267185     16.385918       1.0      13.000000
…                 …              …             …         …            …
pct_own       26954.0       0.642269      0.224184       0.0       0.505040
married       27011.0       0.509312      0.135701       0.0       0.426550
married_snp   27011.0       0.047344      0.037156       0.0       0.020825
separated     27011.0       0.019073      0.020744       0.0       0.004555
divorced      27011.0       0.100385      0.048808       0.0       0.066015

                         50%            75%           max
UID            257212.000000  275810.000000  294334.00000
BLOCKID                  NaN            NaN           NaN
SUMLEVEL          140.000000     140.000000     140.00000
COUNTYID           63.000000     109.000000     840.00000
STATEID            28.000000      42.000000      72.00000
…                        …              …            …
pct_own             0.691585       0.817673       1.00000
married             0.527230       0.606055       1.00000
married_snp         0.038770       0.064895       0.71429
separated           0.013460       0.027460       0.71429
divorced            0.095330       0.129030       1.00000

[74 rows x 8 columns]
```

[71]:
```
#fa = FactorAnalyzer(n_factors=5)
#fa.fit_transform(df_train.select_dtypes(exclude= ('object','category')))
#fa.loadings_
```

0.2 Data Modeling :

3. Build a linear Regression model to predict the total monthly expenditure for home mortgages loan. Please refer deplotment_RE.xlsx. Column hc_mortgage_mean is predicted variable. This is the mean monthly mortgage and owner costs of specified geographical location. Note: Exclude loans from prediction model which have NaN (Not a Number) values for hc_mortgage_mean.
   a) Run a model at a Nation level. If the accuracy levels and R square are not satisfactory proceed to below step.
   b) Run another model at State level. There are 52 states in USA.
   c) Keep below considerations while building a linear regression model:

- Variables should have significant impact on predicting Monthly mortgage and owner costs
- Utilize all predictor variable to start with initial hypothesis
- R square of 60 percent and above should be achieved

- Ensure Multi-collinearity does not exist in dependent variables
- Test if predicted variable is normally distributed

[72]:
```
train.columns
```

[72]:
```
Index(['UID', 'SUMLEVEL', 'COUNTYID', 'STATEID', 'state', 'state_ab', 'city',
       'place', 'type', 'primary', 'zip_code', 'area_code', 'lat', 'lng',
       'ALand', 'AWater', 'pop', 'male_pop', 'female_pop', 'rent_mean',
       'rent_median', 'rent_stdev', 'rent_sample_weight', 'rent_samples',
       'rent_gt_10', 'rent_gt_15', 'rent_gt_20', 'rent_gt_25', 'rent_gt_30',
       'rent_gt_35', 'rent_gt_40', 'rent_gt_50', 'universe_samples',
       'used_samples', 'hi_mean', 'hi_median', 'hi_stdev', 'hi_sample_weight',
       'hi_samples', 'family_mean', 'family_median', 'family_stdev',
       'family_sample_weight', 'family_samples', 'hc_mortgage_mean',
       'hc_mortgage_median', 'hc_mortgage_stdev', 'hc_mortgage_sample_weight',
       'hc_mortgage_samples', 'hc_mean', 'hc_median', 'hc_stdev', 'hc_samples',
       'hc_sample_weight', 'home_equity_second_mortgage', 'second_mortgage',
       'home_equity', 'debt', 'second_mortgage_cdf', 'home_equity_cdf',
       'debt_cdf', 'hs_degree', 'hs_degree_male', 'hs_degree_female',
       'male_age_mean', 'male_age_median', 'male_age_stdev',
       'male_age_sample_weight', 'male_age_samples', 'female_age_mean',
       'female_age_median', 'female_age_stdev', 'female_age_sample_weight',
       'female_age_samples', 'pct_own', 'married', 'married_snp', 'separated',
       'divorced', 'split', 'bad_debt', 'good_debt', 'population_density',
       'median_age', 'pop_bins'],
      dtype='object')
```

[73]:
```
train['type'].unique()
```

[73]:
```
array(['City', 'Urban', 'Town', 'CDP', 'Village', 'Borough'], dtype=object)
```

[74]:
```
type_dict={'type':{'City':1, 'Urban':2, 'Town':3, 'CDP':4, 'Village':5,␣
 ↪'Borough':6}}
train.replace(type_dict,inplace=True)
```

[75]:
```
test.replace(type_dict,inplace=True)
```

[76]:
```
train['type'].unique()
```

[76]:
```
array([1, 2, 3, 4, 5, 6])
```

[77]:
```
test['type'].unique()
```

[77]:
```
array([4, 1, 6, 3, 5, 2])
```

[78]:
```
feature_cols=['COUNTYID','STATEID','zip_code','type','pop',␣
 ↪'family_mean','second_mortgage', 'home_equity', 'debt','hs_degree',
```

```
                'pct_own', 'married','separated', 'divorced']
```

[79]:
```
X_train = train[feature_cols]
y_train = train['hc_mortgage_mean']
```

[80]:
```
X_test = test[feature_cols]
y_test = test['hc_mortgage_mean']
```

[81]:
```
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score,␣
 ↪mean_absolute_error,mean_squared_error,accuracy_score
```

[82]:
```
X_train.head(2)
```

[82]:
```
   COUNTYID  STATEID  zip_code  type   pop  family_mean  second_mortgage  \
0        53       36     13346     1  5230  67994.14790          0.02077
1       141       18     46616     1  2633  50670.10337          0.02222

   home_equity     debt  hs_degree  pct_own  married  separated  divorced
0      0.08919  0.52963    0.89288  0.79046  0.57851    0.01240    0.0877
1      0.04274  0.60855    0.90487  0.52483  0.34886    0.01426    0.0903
```

[83]:
```
X_test.head(2)
```

[83]:
```
       COUNTYID  STATEID  zip_code  type   pop  family_mean  second_mortgage  \
27161       163       26     48239     4  3417  53802.87122          0.06443
27162         1       23      4210     1  3796  85642.22095          0.01175

       home_equity     debt  hs_degree  pct_own  married  separated  divorced
27161      0.07651  0.63624    0.91047  0.70252  0.28217    0.03813   0.14299
27162      0.14375  0.64755    0.94290  0.85128  0.64221    0.00000   0.13377
```

[84]:
```
sc = StandardScaler()
X_train_scaled = sc.fit_transform(X_train)
X_test_scaled = sc.fit_transform(X_test)
```

[85]:
```
lr = LinearRegression()
lr.fit(X_train_scaled, y_train)
```

[85]:
```
LinearRegression()
```

[86]:
```
y_pred= lr.predict(X_test_scaled)
```

[87]:
```
r2_score(y_test,y_pred)
```

[87]:
```
0.7381843831191806
```

R Square of above 60 % is achieved.

[88]:
```
mean_absolute_error(y_test, y_pred)
```

[88]: 233.87107809549642

[89]:
```
mean_squared_error(y_test, y_pred)
```

[89]: 103820.22842724771

[90]:
```
np.sqrt(mean_squared_error(y_test,y_pred))
```

[90]: 322.21146538763594

[91]:
```
r2_score(y_train, lr.predict(X_train_scaled))
```

[91]: 0.7343400491358771

[92]:
```
lr.coef_
```

[92]:
```
array([ -28.50905152,  -21.7110459 ,  -22.98421445,  -57.43072313,
         -4.78167778,  558.73814723,   -0.56122567,   70.89003828,
         12.81881543, -113.18538434, -176.51471006,    8.1107273 ,
          5.24319521,  -55.79370511])
```

[93]:
```
X_train.columns
```

[93]:
```
Index(['COUNTYID', 'STATEID', 'zip_code', 'type', 'pop', 'family_mean',
       'second_mortgage', 'home_equity', 'debt', 'hs_degree', 'pct_own',
       'married', 'separated', 'divorced'],
      dtype='object')
```

[94]:
```
state = train['STATEID'].unique()
state
```

[94]:
```
array([36, 18, 72, 20,  1, 48, 45,  6,  5, 24, 17, 19, 47, 32, 22,  8, 44,
       28, 34, 41,  4, 12, 55, 42, 37, 51, 26, 39, 40, 13, 16, 46, 27, 29,
       53, 56,  9, 54, 21, 25, 11, 15, 30,  2, 33, 49, 50, 31, 38, 35, 23,
       10])
```

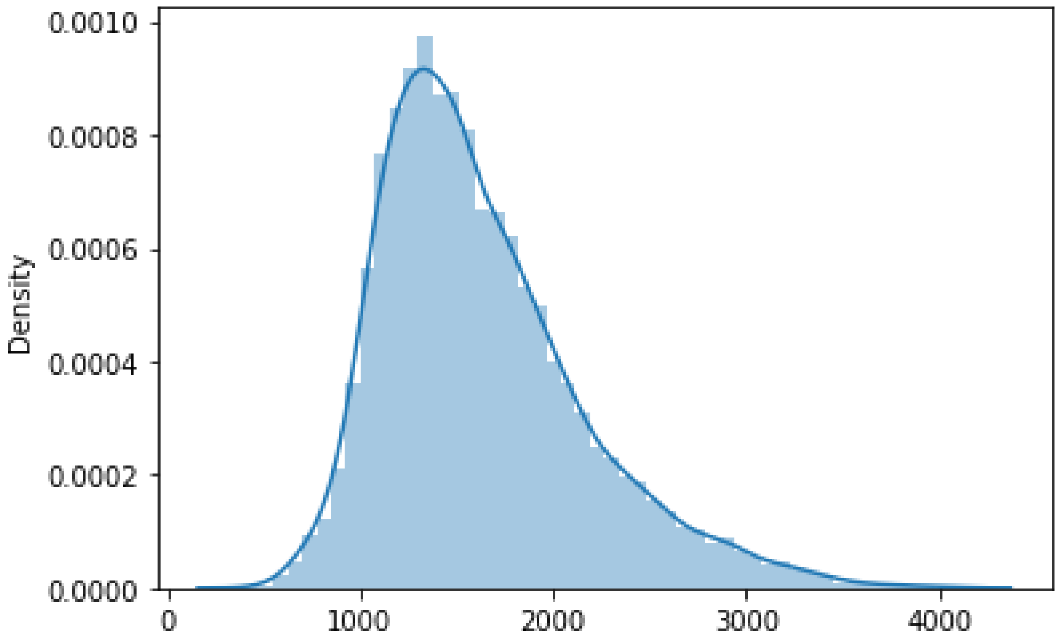
[95]:
```
for i in [11,1,29]:
    print("State ID-",i)

    X_train_nation = train[train['COUNTYID'] == i][feature_cols]
    y_train_nation = train[train['COUNTYID'] == i]['hc_mortgage_mean']

    X_test_nation = test[test['COUNTYID'] == i][feature_cols]
    y_test_nation = test[test['COUNTYID'] == i]['hc_mortgage_mean']
```

```
    X_train_scaled_nation = sc.fit_transform(X_train_nation)
    X_test_scaled_nation = sc.fit_transform(X_test_nation)

    lr.fit(X_train_scaled_nation,y_train_nation)
    y_pred_nation = lr.predict(X_test_scaled_nation)

    print("Overall R2 score of linear regression model for state,",i,":-"␣
↪,r2_score(y_test_nation,y_pred_nation))
    print("Overall RMSE of linear regression model for state,",i,":-" ,np.
↪sqrt(mean_squared_error(y_test_nation,y_pred_nation)))
    print("\n")
```

```
State ID- 11
Overall R2 score of linear regression model for state, 11 :- 0.7459039215483687
Overall RMSE of linear regression model for state, 11 :- 238.51906236063815


State ID- 1
Overall R2 score of linear regression model for state, 1 :- 0.80861461310093
Overall RMSE of linear regression model for state, 1 :- 311.5346317169071


State ID- 29
Overall R2 score of linear regression model for state, 29 :- 0.7089947086337807
Overall RMSE of linear regression model for state, 29 :- 270.07228257987407
```

[96]:

```
sns.distplot(y_pred)
plt.show()
```



0.3 Data Reporting:

4. Create a dashboard in tableau by choosing appropriate chart types and metrics useful for the business. The dashboard must entail the following:

- Box plot of distribution of average rent by type of place (village, urban, town, etc.).
- Pie charts to show overall debt and bad debt.
- Explore the top 2,500 locations where the percentage of households with a second mortgage is the highest and percent ownership is above 10 percent. Visualize using geo-map.
- Heat map for correlation matrix.
- Pie chart to show the population distribution across different types of places (village, urban, town etc.).

realestatetab.png