Week 1 ; Data Import and Preparation:

1. Import data.

[75]:
```
import pandas as pd
#pd.set_option('display.max_rows',None)
pd.set_option('display.max_columns',None)
```

[77]:
```
df_train=pd.read_csv("/content/drive/MyDrive/Course 5 - Data Science Capstne␣
↪Project/Real Estate/Project 1/train.csv")
df_test=pd.read_csv("/content/drive/MyDrive/Course 5 - Data Science Capstne␣
↪Project/Real Estate/Project 1/test.csv")
```

[78]:
```
df_train.head()
```

[78]:
```
      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
2  245683      NaN       140        63       18      Indiana       IN
3  279653      NaN       140       127       72  Puerto Rico       PR
4  247218      NaN       140       161       20       Kansas       KS

         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
2    Danville         Danville   City   tract     46122        317  39.792202
3    San Juan         Guaynabo  Urban   tract       927        787  18.396103
4   Manhattan   Manhattan City   City   tract     66502        785  39.195573

         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
2 -86.515246   69561595.0   284193  6881      3643        3238  742.77365
3 -66.104169    1105793.0        0  2700      1141        1559  803.42018
4 -96.569366    2554403.0        0  5637      2586        3051  938.56493

   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
2        703.0   323.39011           291.85520         378.0     0.95238
3        782.0   297.39258           259.30316         368.0     0.94693
4        881.0   392.44096          1005.42886        1704.0     0.99286

   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
2     0.88624     0.79630     0.66667     0.39153     0.39153     0.28307
3     0.87151     0.69832     0.61732     0.51397     0.46927     0.35754
```

```
4     0.98247     0.91688     0.84740     0.78247     0.60974     0.55455

   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
2     0.15873               459           378   84942.68317    74964.0
3     0.32961               438           358   48733.67116    37845.0
4     0.44416              1725          1540   31834.15466    22497.0

      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
2  56811.62186        1155.20980      2488.0  95262.51431        85395.0
3  45100.54010         928.32193      1267.0  56401.68133        44399.0
4  34046.50907        1548.67477      1983.0  54053.42396        50272.0

   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
2   49292.67664             709.74925          1889.0        1506.06758
3   41082.90515             490.18479           729.0        1175.28642
4   39609.12605             244.08903           395.0        1192.58759

   hc_mortgage_median  hc_mortgage_stdev  hc_mortgage_sample_weight  \
0              1223.0          641.22898                  377.83135
1               784.0          482.27020                  316.88320
2              1361.0          731.89394                  699.41354
3              1101.0          428.98751                  261.28471
4              1125.0          327.49674                   76.61052

   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
2               1491.0  556.45986      532.0  184.42175       538.0
3                437.0  288.04047      247.0  185.55887       392.0
4                134.0  443.68855      444.0   76.12674       124.0

   hc_sample_weight  home_equity_second_mortgage  second_mortgage  \
0         499.29293                      0.01588          0.02077
1         189.60606                      0.02222          0.02222
2         323.35354                      0.00000          0.00000
3         314.90566                      0.01086          0.01086
4          79.55556                      0.05426          0.05426

   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
```

```
2    0.09512  0.73484          1.00000        0.46332  0.28704
3    0.01086  0.52714          0.53057        0.82530  0.73727
4    0.05426  0.51938          0.18332        0.65545  0.74967

   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
2    0.94288         0.94616           0.93952       39.38154
3    0.91500         0.90755           0.92043       48.64749
4    1.00000         1.00000           1.00000       26.07533

   male_age_median  male_age_stdev  male_age_sample_weight  male_age_samples  \
0         44.00000        22.97306               696.42136            2612.0
1         32.00000        20.37452               323.90204            1349.0
2         40.83333        22.89769               888.29730            3643.0
3         48.91667        23.05968               274.98956            1141.0
4         22.41667        11.84399              1296.89877            2586.0

  female_age_mean  female_age_median  female_age_stdev  \
0        44.48629           45.33333          22.51276
1        36.48391           37.58333          23.43353
2        42.15810           42.83333          23.94119
3        47.77526           50.58333          24.32015
4        24.17693           21.58333          11.10484

   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

   married_snp  separated  divorced
0      0.01882    0.01240   0.08770
1      0.01426    0.01426   0.09030
2      0.02830    0.01607   0.10657
3      0.02021    0.02021   0.10106
4      0.00000    0.00000   0.03109
```

[79]:

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call
drive.mount("/content/drive", force_remount=True).
```

[80]:

```
df_test.head()
```

```
[80]:       UID  BLOCKID  SUMLEVEL  COUNTYID  STATEID         state state_ab  \
      0  255504      NaN       140       163       26      Michigan       MI
      1  252676      NaN       140         1       23         Maine       ME
      2  276314      NaN       140        15       42  Pennsylvania       PA
      3  248614      NaN       140       231       21      Kentucky       KY
      4  286865      NaN       140       355       48         Texas       TX

                    city                    place      type primary  zip_code  \
      0          Detroit  Dearborn Heights City       CDP   tract     48239
      1           Auburn              Auburn City      City   tract      4210
      2        Pine City                Millerton   Borough   tract     14871
      3       Monticello          Monticello City      City   tract     42633
      4  Corpus Christi                     Edroy      Town   tract     78410

         area_code        lat         lng      ALand   AWater   pop  male_pop  \
      0        313  42.346422  -83.252823    2711280    39555  3417      1479
      1        207  44.100724  -70.257832   14778785  2705204  3796      1846
      2        607  41.948556  -76.783808  258903666   863840  3944      2065
      3        606  36.746009  -84.766870  501694825  2623067  2508      1427
      4        361  27.882462  -97.678586   13796057   497689  6230      3274

         female_pop   rent_mean  rent_median  rent_stdev  rent_sample_weight  \
      0        1938   858.57169        859.0   232.39082  276.07497
      1        1950   832.68625        750.0   267.22342           183.32299
      2        1879   816.00639        755.0   416.25699           141.39063
      3        1081   418.68937        385.0   156.92024            88.95960
      4        2956  1031.63763        997.0   326.76727           277.39844

         rent_samples  rent_gt_10  rent_gt_15  rent_gt_20  rent_gt_25  rent_gt_30  \
      0         424.0     1.00000     0.95696     0.85316     0.85316     0.85316
      1         245.0     1.00000     1.00000     0.86611     0.67364     0.30962
      2         217.0     0.97573     0.93204     0.78641     0.71845     0.63592
      3          93.0     1.00000     0.93548     0.93548     0.64516     0.55914
      4         624.0     0.72276     0.66506     0.53526     0.38301     0.18910

         rent_gt_35  rent_gt_40  rent_gt_50  universe_samples  used_samples  \
      0     0.85316     0.76962     0.63544               435           395
      1     0.30962     0.30962     0.27197               275           239
      2     0.47573     0.43689     0.32524               245           206
      3     0.46237     0.46237     0.36559               153            93
      4     0.16667     0.14263     0.11058               660           624

              hi_mean  hi_median     hi_stdev  hi_sample_weight  hi_samples  \
      0  48899.52121    38746.0  44392.20902         798.02401      1180.0
      1  72335.33234    61008.0  51895.81159         922.82969      1722.0
      2  58501.15901    51648.0  45245.27248         893.07759      1461.0
      3  38237.55059    31612.0  34527.61607         775.17947       957.0
```

```
4  114456.07790    94211.0  81950.95692          836.30759      2404.0

   family_mean  family_median  family_stdev  family_sample_weight  \
0   53802.87122        45167.0   43756.56479             464.30972
1   85642.22095        74759.0   49156.72870             482.99945
2   65694.06582        57186.0   44239.31893             619.73962
3   44156.38709        34687.0   34899.74300             535.21987
4  123527.02420       103898.0   72173.55823             507.42257

   family_samples  hc_mortgage_mean  hc_mortgage_median  hc_mortgage_stdev  \
0           769.0        1139.24548              1109.0          336.47710
1          1147.0        1533.25988              1438.0          536.61118
2          1084.0        1254.54462              1089.0          596.85204
3           689.0         862.65763               749.0          624.42157
4          1738.0        1996.41425              1907.0          740.21168

   hc_mortgage_sample_weight  hc_mortgage_samples    hc_mean  hc_median  \
0                  262.67011                474.0  488.51323      436.0
1                  373.96188                937.0  661.31296      668.0
2                  340.45884                552.0  397.44466      356.0
3                  299.56752                337.0  200.88113      180.0
4                  319.97570               1102.0  867.57713      804.0

    hc_stdev  hc_samples  hc_sample_weight  home_equity_second_mortgage  \
0  192.75147       271.0         189.18182                      0.06443
1  201.31365       510.0         279.69697                      0.01175
2  189.40372       664.0         534.16737                      0.01069
3   91.56490       467.0         454.85404                      0.00995
4  376.20236       642.0         333.91919                      0.00000

   second_mortgage  home_equity     debt  second_mortgage_cdf  \
0          0.06443      0.07651  0.63624              0.14111
1          0.01175      0.14375  0.64755              0.52310
2          0.01316      0.06497  0.45395              0.51066
3          0.00995      0.01741  0.41915              0.53770
4          0.00000      0.03440  0.63188              1.00000

   home_equity_cdf  debt_cdf  hs_degree  hs_degree_male  hs_degree_female  \
0          0.55087   0.51965    0.91047         0.92010           0.90391
1          0.26442   0.49359    0.94290         0.92832           0.95736
2          0.60484   0.83848    0.89238         0.86003           0.92463
3          0.80931   0.87403    0.60908         0.56584           0.65947
4          0.74519   0.52943    0.86297         0.87969           0.84466

   male_age_mean  male_age_median  male_age_stdev  male_age_sample_weight  \
0       33.37131         27.83333        22.36768               334.30978
1       43.88680         46.08333        22.90302               427.10824
```

```
2      39.81661     41.91667     24.29111          499.10080
3      41.81638     43.00000     24.65325          333.57733
4      42.13301     43.75000     22.69502          833.57435

   male_age_samples  female_age_mean  female_age_median  female_age_stdev\  0
             1479.0         34.78682           33.75000  21.58531
1            1846.0         44.23451           46.66667          22.37036
2            2065.0         41.62426           44.50000          22.86213
3            1427.0         44.81200           48.00000          21.03155
4            3274.0         40.66618           42.66667          21.30900

  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
2                453.11959              1879.0  0.81897  0.59961
3                263.94320              1081.0  0.84609  0.56953
4                709.90829              2956.0  0.79077  0.57620

   married_snp  separated  divorced
0      0.05910    0.03813   0.14299
1      0.02338    0.00000   0.13377
2      0.01746    0.01358   0.10026
3      0.05492    0.04694   0.12489
4      0.01726    0.00588   0.16379
```

```
[81]: df_train.shape
```

```
[81]: (27321, 80)
```

```
[82]: df_test.shape
```

```
[82]: (11709, 80)
```

2. Figure out the primary key and look for the requirement of indexing.

```
[83]: len(set(df_train['UID']).intersection(set(df_test['UID'])))
```

```
[83]: 123
```

So here 123 common UID in train and test data.

```
[85]: df_train.dtypes
```

```
[85] : UID              int64
       BLOCKID        float64
       SUMLEVEL         int64
       COUNTYID         int64
       STATEID          int64
```

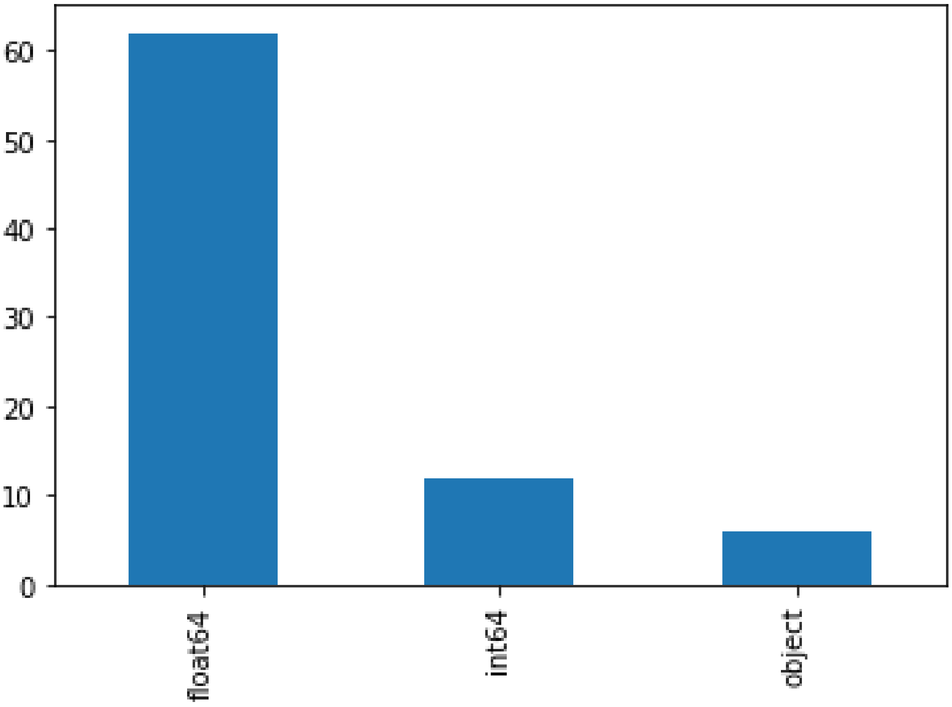
```
                   ...
pct_own           float64
married           float64
married_snp        float64
separated          float64
divorced          float64
Length: 80, dtype: object
```

[86]:

```
df_train.dtypes.value_counts().plot(kind="bar")
```

[86] :

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f069d8e39d0>
```



[87]:

```
df_train.describe(include="O")
```

[87]:

```
             state  state_ab     city          place   type  primary
count        27321     27321    27321          27321  27321    27321
unique          52        52     6916           9912      6        1
top     California        CA  Chicago  New York City   City    tract
freq          2926      2926      294            490  15237    27321
```

3. Gauge the fill rate of the variables and devise plans for missing value treatment. Please explain explicitly the reason for the treatment chosen for each variable.

[88]:

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

[89]:

```
df_combined=df_train.append(df_test, ignore_index=True)
df_combined.head()
```

[89] :

```
      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
2  245683      NaN       140        63       18      Indiana       IN
3  279653      NaN       140       127       72  Puerto Rico       PR
4  247218      NaN       140       161       20       Kansas       KS

         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
2    Danville         Danville   City   tract     46122        317  39.792202
3    San Juan         Guaynabo  Urban   tract       927        787  18.396103
4   Manhattan   Manhattan City   City   tract     66502        785  39.195573

         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
2 -86.515246   69561595.0   284193  6881      3643        3238  742.77365
3 -66.104169    1105793.0        0  2700      1141        1559  803.42018
4 -96.569366    2554403.0        0  5637      2586        3051  938.56493

   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
2        703.0   323.39011           291.85520         378.0     0.95238
3        782.0   297.39258           259.30316         368.0     0.94693
4        881.0   392.44096          1005.42886        1704.0     0.99286

   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
2     0.88624     0.79630     0.66667     0.39153     0.39153     0.28307
3     0.87151     0.69832     0.61732     0.51397     0.46927     0.35754
4     0.98247     0.91688     0.84740     0.78247     0.60974     0.55455

   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
2     0.15873               459           378  84942.68317    74964.0
3     0.32961               438           358  48733.67116    37845.0
```

```
4    0.44416              1725          1540  31834.15466    22497.0

      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
2  56811.62186        1155.20980      2488.0  95262.51431        85395.0
3  45100.54010         928.32193      1267.0  56401.68133        44399.0
4  34046.50907        1548.67477      1983.0  54053.42396        50272.0

   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
2   49292.67664             709.74925          1889.0        1506.06758
3   41082.90515             490.18479           729.0        1175.28642
4   39609.12605             244.08903           395.0        1192.58759

   hc_mortgage_median  hc_mortgage_stdev  hc_mortgage_sample_weight  \
0              1223.0          641.22898                  377.83135
1               784.0          482.27020                  316.88320
2              1361.0          731.89394                  699.41354
3              1101.0          428.98751                  261.28471
4              1125.0          327.49674                   76.61052

   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
2               1491.0  556.45986      532.0  184.42175       538.0
3                437.0  288.04047      247.0  185.55887       392.0
4                134.0  443.68855      444.0   76.12674       124.0

   hc_sample_weight  home_equity_second_mortgage  second_mortgage  \
0         499.29293                      0.01588          0.02077
1         189.60606                      0.02222          0.02222
2         323.35354                      0.00000          0.00000
3         314.90566                      0.01086          0.01086
4          79.55556                      0.05426          0.05426

   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
2      0.09512  0.73484              1.00000          0.46332   0.28704
3      0.01086  0.52714              0.53057          0.82530   0.73727
4      0.05426  0.51938              0.18332          0.65545   0.74967

   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
```

```
2    0.94288        0.94616        0.93952      39.38154
3    0.91500        0.90755        0.92043      48.64749
4    1.00000        1.00000        1.00000      26.07533

   male_age_median  male_age_stdev  male_age_sample_weight  male_age_samples  \
0         44.00000        22.97306               696.42136            2612.0
1         32.00000        20.37452               323.90204            1349.0
2         40.83333        22.89769               888.29730            3643.0
3         48.91667        23.05968               274.98956            1141.0
4         22.41667        11.84399              1296.89877            2586.0

   female_age_mean  female_age_median  female_age_stdev  \
0         44.48629           45.33333          22.51276
1         36.48391           37.58333          23.43353
2         42.15810           42.83333          23.94119
3         47.77526           50.58333          24.32015
4         24.17693           21.58333          11.10484

   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

   married_snp  separated  divorced  split
0      0.01882    0.01240   0.08770  Train
1      0.01426    0.01426   0.09030  Train
2      0.02830    0.01607   0.10657  Train
3      0.02021    0.02021   0.10106  Train
4      0.00000    0.00000   0.03109  Train
```

```
[90]: df_combined.tail()
```

```
[90]:
          UID  BLOCKID  SUMLEVEL  COUNTYID  STATEID          state  state_ab  \
39025  238088      NaN       140       105       12        Florida        FL
39026  242811      NaN       140        31       17       Illinois        IL
39027  250127      NaN       140         9       25  Massachusetts        MA
39028  241096      NaN       140        27       19           Iowa        IA
39029  287763      NaN       140       453       48          Texas        TX

           city                place     type primary  zip_code  area_code  \
39025  Lakeland      Crystal Springs     City   tract     33810        863
39026   Chicago         Chicago City  Village   tract     60609        773
39027  Lawrence    Methuen Town City     City   tract      1841        978
39028   Carroll         Carroll City     City   tract     51401        712
39029    Austin  Sunset Valley City     Town   tract     78745        512
```

```
             lat        lng       ALand   AWater   pop  male_pop  female_pop  \
39025  28.226068 -82.068886  92582775.0  1166617  5611      2697        2914
39026  41.804936 -87.667304    327029.0        0  2695      1504        1191
39027  42.737778 -71.131761   5225804.0   393810  7392      3669        3723
39028  42.081366 -94.866175  11066759.0        0  5945      2732        3213
39029  30.219013 -97.774728   1990126.0        0  4117      2070        2047

        rent_mean  rent_median  rent_stdev  rent_sample_weight  rent_samples  \
39025  1458.82449       1603.0   566.90682            29.43733          99.0
39026   700.53513        661.0   254.66700           480.86455         592.0
39027  1069.70567       1138.0   488.13975           207.29615         506.0
39028   696.93368        576.0   595.16228           503.83775         590.0
39029   950.09294        864.0   333.82364           417.07457         675.0

       rent_gt_10  rent_gt_15  rent_gt_20  rent_gt_25  rent_gt_30  rent_gt_35  \
39025     1.00000     1.00000     1.00000     0.62626     0.62626     0.35354
39026     1.00000     0.90034     0.85911     0.63058     0.53952     0.41237
39027     0.85375     0.83004     0.77273     0.56324     0.47431     0.33399
39028     0.96886     0.92042     0.83045     0.69723     0.62284     0.43772
39029     1.00000     0.97481     0.86074     0.73926     0.44593     0.38370

       rent_gt_40  rent_gt_50  universe_samples  used_samples      hi_mean  \
39025     0.18182     0.09091               147            99  57723.48180
39026     0.35223     0.19931               618           582  35249.76522
39027     0.30237     0.02569               539           506  89549.15374
39028     0.33737     0.33737               663           578  57877.26387
39029     0.27852     0.25778               682           675  58006.33817

       hi_median     hi_stdev  hi_sample_weight  hi_samples  family_mean  \
39025    48192.0  41301.62188        1636.68434      2496.0  70786.81912
39026    27396.0  28889.72217         683.94534       838.0  38912.54156
39027    75357.0  66560.76837        1339.55365      2739.0  99484.96572
39028    41838.0  49745.93715        1605.79897      2596.0  75066.29009
39029    44179.0  49189.98590         902.67611      1396.0  54913.24441

       family_median  family_stdev  family_sample_weight  family_samples  \
39025        59194.0   40582.36046             945.85894          1685.0
39026        32554.0   29796.19973             415.51917           555.0
39027        89050.0   62721.62266             853.61856          1986.0
39028        72135.0   47200.66016             782.93088          1568.0
39029        42469.0   41016.08651             581.04758           877.0

       hc_mortgage_mean  hc_mortgage_median  hc_mortgage_stdev  \
39025        1269.83033              1119.0          689.35735
39026        1406.83478              1224.0          621.89533
39027        1791.63902              1794.0          656.68467
```

```
39028         1182.30365                 1059.0            587.01032
39029         1364.17379                 1318.0            463.57052

       hc_mortgage_sample_weight  hc_mortgage_samples    hc_mean  hc_median  \
39025                  608.62709               1024.0  536.66053      500.0
39026                   62.54709                139.0  487.66419      465.0
39027                  548.16568               1634.0  654.78088      612.0
39028                  796.11244               1267.0  369.29903      334.0
39029                  217.49287                456.0  550.78197      555.0

        hc_stdev  hc_samples  hc_sample_weight  home_equity_second_mortgage  \
39025  267.25752      1325.0         914.89899                      0.02043
39026  220.16444        81.0          47.09727                      0.05909
39027  256.84182       566.0         299.83838                      0.02727
39028  133.20792       666.0         556.40404                      0.03570
39029  199.13527       258.0         163.55556                      0.00000

       second_mortgage  home_equity     debt  second_mortgage_cdf  \
39025          0.03619      0.04044  0.43593              0.29592
39026          0.05909      0.08182  0.63182              0.16199
39027          0.02727      0.13545  0.74273              0.37297
39028          0.03570      0.07967  0.65546              0.30010
39029          0.00000      0.05042  0.63866              1.00000

       home_equity_cdf  debt_cdf  hs_degree  hs_degree_male  hs_degree_female  \
39025          0.71860   0.85762    0.92097         0.95007           0.89480
39026          0.52552   0.52957    0.54890         0.49817           0.60965
39027          0.29411   0.26972    0.94057         0.94000           0.94105
39028          0.53579   0.47507    0.91407         0.92428           0.90634
39029          0.67315   0.51407    0.78685         0.80615           0.76820

       male_age_mean  male_age_median  male_age_stdev  male_age_sample_weight  \
39025       51.03535         55.50000        22.41099               704.65208
39026       32.94145         29.83333        20.52061               408.44261
39027       35.85743         34.91667        22.49430               880.48254
39028       39.18219         40.25000        24.86317               636.20201
39029       35.56404         35.00000        21.67509               522.45931

       male_age_samples  female_age_mean  female_age_median  female_age_stdev  \
39025            2697.0         53.51255           59.58333          23.23426
39026            1504.0         33.14169           32.83333          20.24698
39027            3669.0         43.53905           43.66667          23.17995
39028            2732.0         45.63179           48.16667          24.84209
39029            2070.0         35.99955           35.41667          20.68049

       female_age_sample_weight  female_age_samples  pct_own  married  \
39025                 699.33353              2914.0  0.93121  0.65969
```

```
39026              306.63915               1191.0  0.33122  0.42882
39027              900.13903               3723.0  0.84372  0.50269
39028              693.82905               3213.0  0.83330  0.66699
39029              559.30291               2047.0  0.52587  0.51922

       married_snp  separated  divorced split
39025      0.02135    0.02135   0.08780  Test
39026      0.07781    0.02829   0.05305  Test
39027      0.00108    0.00108   0.07294  Test
39028      0.02738    0.00000   0.04694  Test
39029      0.08066    0.02520   0.10586  Test
```

[91]:
```
df_combined.shape
```

[91]:
```
(39030, 81)
```

[92]:
```
df_combined.isna().sum()
```

[92]:
```
UID                0
BLOCKID        39030
SUMLEVEL           0
COUNTYID           0
STATEID            0
               ...
married          275
married_snp      275
separated        275
divorced         275
split              0
Length: 81, dtype: int64
```

[93]:
```
# Fill rate of the variables -> (1- missing %)
1-df_combined.isna().sum()/len(df_combined)
```

[93]:
```
UID            1.000000
BLOCKID        0.000000
SUMLEVEL       1.000000
COUNTYID       1.000000
STATEID        1.000000
                 ...
married        0.992954
married_snp    0.992954
separated      0.992954
divorced       0.992954
split          1.000000
Length: 81, dtype: float64
```

[94]:
```
# 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)
```

[95]:
```
df_combined.isna().sum()/len(df_combined)*100
```

[95]:
```
UID            0.000000
SUMLEVEL       0.000000
COUNTYID       0.000000
STATEID        0.000000
state          0.000000
                 ...
married        0.704586
married_snp    0.704586
separated      0.704586
divorced       0.704586
split          0.000000
Length: 80, dtype: float64
```

[96]:
```
# 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
```

[96]:
```
['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']
```

[97]:
```
#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
```

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

```
[98]: 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
```

```
[99]: df_combined.shape
```

```
[99]: (39030, 80)
```

```
[100]: # Drop duplicate observations
       df_combined.drop_duplicates(inplace=True)
       df_combined.shape
```

```
[100]: (38838, 80)
```

```
[101]: # 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
```

```
[101]: (38715, 80)
```

Exploratory Data Analysis (EDA):

4. 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

[102]:

```
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)
```

[103]:

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

[103]:

```
               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
```

[104]:

```
!pip install geopandas
import warnings
warnings.filterwarnings('ignore')
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Requirement already satisfied: geopandas in /usr/local/lib/python3.8/dist-
packages (0.12.2)
Requirement already satisfied: pandas>=1.0.0 in /usr/local/lib/python3.8/dist-
packages (from geopandas) (1.3.5)
Requirement already satisfied: pyproj>=2.6.1.post1 in
/usr/local/lib/python3.8/dist-packages (from geopandas) (3.4.1)
Requirement already satisfied: shapely>=1.7 in /usr/local/lib/python3.8/dist-
packages (from geopandas) (2.0.1)
Requirement already satisfied: packaging in /usr/local/lib/python3.8/dist-
packages (from geopandas) (23.0)
Requirement already satisfied: fiona>=1.8 in /usr/local/lib/python3.8/dist-
packages (from geopandas) (1.9.0)
Requirement already satisfied: certifi in /usr/local/lib/python3.8/dist-packages
(from fiona>=1.8->geopandas) (2022.12.7)
Requirement already satisfied: attrs>=19.2.0 in /usr/local/lib/python3.8/dist-
packages (from fiona>=1.8->geopandas) (22.2.0)
Requirement already satisfied: click~=8.0 in /usr/local/lib/python3.8/dist-
packages (from fiona>=1.8->geopandas) (8.1.3)
Requirement already satisfied: munch>=2.3.2 in /usr/local/lib/python3.8/dist-
```

```
packages (from fiona>=1.8->geopandas) (2.5.0)
Requirement already satisfied: click-plugins>=1.0 in
/usr/local/lib/python3.8/dist-packages (from fiona>=1.8->geopandas) (1.1.1)
Requirement already satisfied: cligj>=0.5 in /usr/local/lib/python3.8/dist-
packages (from fiona>=1.8->geopandas) (0.7.2)
Requirement already satisfied: numpy>=1.17.3 in /usr/local/lib/python3.8/dist-
packages (from pandas>=1.0.0->geopandas) (1.21.6)
Requirement already satisfied: python-dateutil>=2.7.3 in
/usr/local/lib/python3.8/dist-packages (from pandas>=1.0.0->geopandas) (2.8.2)
Requirement already satisfied: pytz>=2017.3 in /usr/local/lib/python3.8/dist-
packages (from pandas>=1.0.0->geopandas) (2022.7.1)
Requirement already satisfied: six in /usr/local/lib/python3.8/dist-packages
(from munch>=2.3.2->fiona>=1.8->geopandas) (1.15.0)
```

[105]:
```
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
```

[105]:
```
                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
5478       California Marina Del Rey      CA    Marina Del Rey  33.983204
25642        Maryland     Baltimore       MD          Lochearn  39.353095
26671    Pennsylvania  Philadelphia       PA Philadelphia City  40.039070
24443      California       Manteca       CA      Manteca City  37.732143

              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)
5478  -118.466139  POINT (-118.46614 33.98320)
25642  -76.733315   POINT (-76.73331 39.35310)
26671  -75.125135   POINT (-75.12514 40.03907)
24443 -121.242902  POINT (-121.24290 37.73214)

[2500 rows x 7 columns]
```

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

[106]:

```
#Bad Debt = second_mortgage + home_equity - home_equity_second_mortgage
df_combined['bad_debt'] = df_combined['second_mortgage'] +␣
 ↪df_combined['home_equity'] - df_combined['home_equity_second_mortgage']
df_combined.head()
```

[106]:

```
      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
2  245683       140        63       18      Indiana       IN    Danville
3  279653       140       127       72  Puerto Rico       PR    San Juan
4  247218       140       161       20       Kansas       KS   Manhattan

            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
2        Danville   City   tract     46122        317  39.792202 -86.515246
3        Guaynabo  Urban   tract       927        787  18.396103 -66.104169
4  Manhattan City   City   tract     66502        785  39.195573 -96.569366

         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
2   69561595.0   284193  6881      3643        3238  742.77365        703.0
3    1105793.0        0  2700      1141        1559  803.42018        782.0
4    2554403.0        0  5637      2586        3051  938.56493        881.0

   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
2   323.39011           291.85520         378.0     0.95238     0.88624
3   297.39258           259.30316         368.0     0.94693     0.87151
4   392.44096          1005.42886        1704.0     0.99286     0.98247

   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
2     0.79630     0.66667     0.39153     0.39153     0.28307     0.15873
3     0.69832     0.61732     0.51397     0.46927     0.35754     0.32961
4     0.91688     0.84740     0.78247     0.60974     0.55455     0.44416

   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
2               459           378  84942.68317    74964.0  56811.62186
3               438           358  48733.67116    37845.0  45100.54010
4              1725          1540  31834.15466    22497.0  34046.50907
```

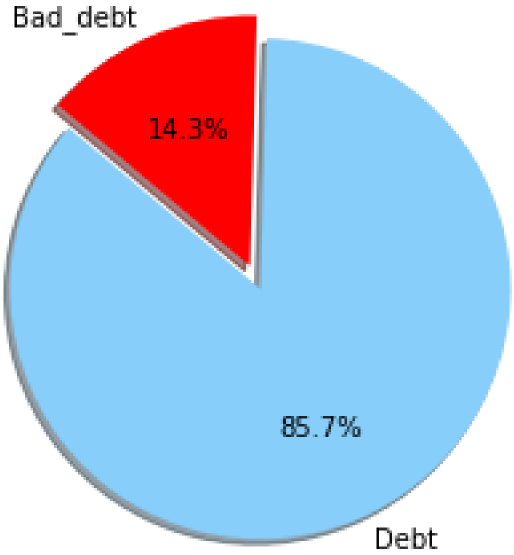
```
   hi_sample_weight  hi_samples  family_mean  family_median  family_stdev   \
0        1290.96240      2024.0  67994.14790        53245.0   47667.30119
1         838.74664      1127.0  50670.10337        43023.0   34715.57548
2        1155.20980      2488.0  95262.51431        85395.0   49292.67664
3         928.32193      1267.0  56401.68133        44399.0   41082.90515
4        1548.67477      1983.0  54053.42396        50272.0   39609.12605

   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
2             709.74925          1889.0        1506.06758              1361.0
3             490.18479           729.0        1175.28642              1101.0
4             244.08903           395.0        1192.58759              1125.0

   hc_mortgage_stdev  hc_mortgage_sample_weight  hc_mortgage_samples   \
0          641.22898                  377.83135                867.0
1          482.27020                  316.88320                356.0
2          731.89394                  699.41354               1491.0
3          428.98751                  261.28471                437.0
4          327.49674                   76.61052                134.0

     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
2  556.45986      532.0  184.42175       538.0         323.35354
3  288.04047      247.0  185.55887       392.0         314.90566
4  443.68855      444.0   76.12674       124.0          79.55556

   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
2                      0.00000          0.00000      0.09512  0.73484
3                      0.01086          0.01086      0.01086  0.52714
4                      0.05426          0.05426      0.05426  0.51938

   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
2              1.00000          0.46332   0.28704    0.94288          0.94616
3              0.53057          0.82530   0.73727    0.91500          0.90755
4              0.18332          0.65545   0.74967    1.00000          1.00000

   hs_degree_female  male_age_mean  male_age_median  male_age_stdev  \
0           0.92434       42.48574         44.00000        22.97306
1           0.94187       34.84728         32.00000        20.37452
2           0.93952       39.38154         40.83333        22.89769
3           0.92043       48.64749         48.91667        23.05968
```

```
4             1.00000       26.07533          22.41667         11.84399

   male_age_sample_weight  male_age_samples  female_age_mean  \
0               696.42136            2612.0         44.48629
1               323.90204            1349.0         36.48391
2               888.29730            3643.0         42.15810
3               274.98956            1141.0         47.77526
4              1296.89877            2586.0         24.17693

   female_age_median  female_age_stdev  female_age_sample_weight  \
0           45.33333          22.51276                 685.33845
1           37.58333          23.43353                 267.23367
2           42.83333          23.94119                 707.01963
3           50.58333          24.32015                 362.20193
4           21.58333          11.10484                1854.48652

   female_age_samples  pct_own  married  married_snp  separated  divorced  \
0              2618.0  0.79046  0.57851      0.01882    0.01240   0.08770
1              1284.0  0.52483  0.34886      0.01426    0.01426   0.09030
2              3238.0  0.85331  0.64745      0.02830    0.01607   0.10657
3              1559.0  0.65037  0.47257      0.02021    0.02021   0.10106
4              3051.0  0.13046  0.12356      0.00000    0.00000   0.03109

   split  bad_debt
0  Train   0.09408
1  Train   0.04274
2  Train   0.09512
3  Train   0.01086
4  Train   0.05426
```

c. Create pie charts to show overall debt and bad debt

[107]:

```
import matplotlib.pyplot as plt
labels = 'Debt', 'Bad_debt'
sizes = [df_combined['debt'].mean()*100, df_combined['bad_debt'].mean()*100]
colors = [ 'lightskyblue','red']
explode = (0.1, 0)  # explode 1st slice

#Plot
plt.pie(sizes,explode=explode,labels=labels, colors=colors,
autopct='%1.1f%%', shadow=True, startangle=140)

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



d. Create Box and whisker plot and analyze the distribution for 2nd mortgage, home equity, goo

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

```
[108] :       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
        2  245683       140        63       18      Indiana        IN    Danville
        3  279653       140       127       72  Puerto Rico        PR    San Juan
        4  247218       140       161       20       Kansas        KS   Manhattan

                    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
        2        Danville   City   tract     46122        317  39.792202 -86.515246
        3        Guaynabo  Urban   tract       927        787  18.396103 -66.104169
        4  Manhattan City   City   tract     66502        785  39.195573 -96.569366

                 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
        2   69561595.0   284193  6881      3643        3238  742.77365        703.0
        3    1105793.0        0  2700      1141        1559  803.42018        782.0
        4    2554403.0        0  5637      2586        3051  938.56493        881.0

           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
2   323.39011            291.85520         378.0       0.95238    0.88624
3   297.39258            259.30316         368.0       0.94693    0.87151
4   392.44096           1005.42886        1704.0       0.99286    0.98247

   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
2     0.79630     0.66667     0.39153     0.39153     0.28307     0.15873
3     0.69832     0.61732     0.51397     0.46927     0.35754     0.32961
4     0.91688     0.84740     0.78247     0.60974     0.55455     0.44416

   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
2               459           378   84942.68317    74964.0  56811.62186
3               438           358   48733.67116    37845.0  45100.54010
4              1725          1540   31834.15466    22497.0  34046.50907

   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
2        1155.20980      2488.0  95262.51431        85395.0   49292.67664
3         928.32193      1267.0  56401.68133        44399.0   41082.90515
4        1548.67477      1983.0  54053.42396        50272.0   39609.12605

   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
2             709.74925          1889.0        1506.06758              1361.0
3             490.18479           729.0        1175.28642              1101.0
4             244.08903           395.0        1192.58759              1125.0

   hc_mortgage_stdev  hc_mortgage_sample_weight   hc_mortgage_samples  \
0          641.22898                  377.83135                 867.0
1          482.27020                  316.88320                 356.0
2          731.89394                  699.41354                1491.0
3          428.98751                  261.28471                 437.0
4          327.49674                   76.61052                 134.0

     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
2  556.45986      532.0  184.42175       538.0          323.35354
3  288.04047      247.0  185.55887       392.0          314.90566
4  443.68855      444.0   76.12674       124.0           79.55556
```

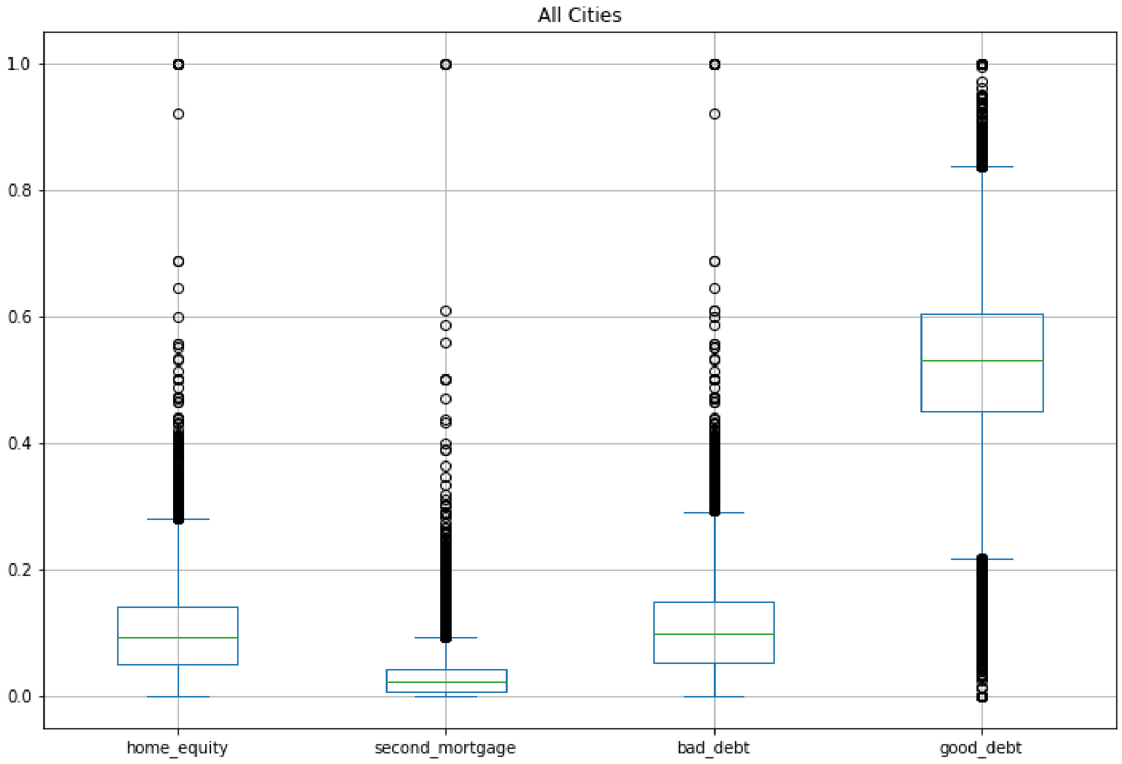
```
   home_equity_second_mortgage  second_mortgage  home_equity     debt  \
0                      0.01588          0.02077      0.08919  0.52963
1                      0.02222          0.02222      0.04274  0.60855
2                      0.00000          0.00000      0.09512  0.73484
3                      0.01086          0.01086      0.01086  0.52714
4                      0.05426          0.05426      0.05426  0.51938

   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
2              1.00000          0.46332   0.28704    0.94288         0.94616
3              0.53057          0.82530   0.73727    0.91500         0.90755
4              0.18332          0.65545   0.74967    1.00000         1.00000

   hs_degree_female  male_age_mean  male_age_median  male_age_stdev  \
0           0.92434       42.48574         44.00000        22.97306
1           0.94187       34.84728         32.00000        20.37452
2           0.93952       39.38154         40.83333        22.89769
3           0.92043       48.64749         48.91667        23.05968
4           1.00000       26.07533         22.41667        11.84399

   male_age_sample_weight  male_age_samples  female_age_mean  \
0               696.42136            2612.0         44.48629
1               323.90204            1349.0         36.48391
2               888.29730            3643.0         42.15810
3               274.98956            1141.0         47.77526
4              1296.89877            2586.0         24.17693

   female_age_median  female_age_stdev  female_age_sample_weight  \
0           45.33333          22.51276                 685.33845
1           37.58333          23.43353                 267.23367
2           42.83333          23.94119                 707.01963
3           50.58333          24.32015                 362.20193
4           21.58333          11.10484                1854.48652

   female_age_samples  pct_own  married  married_snp  separated  divorced  \
0              2618.0  0.79046  0.57851      0.01882    0.01240   0.08770
1              1284.0  0.52483  0.34886      0.01426    0.01426   0.09030
2              3238.0  0.85331  0.64745      0.02830    0.01607   0.10657
3              1559.0  0.65037  0.47257      0.02021    0.02021   0.10106
4              3051.0  0.13046  0.12356      0.00000    0.00000   0.03109

   split  bad_debt  good_debt
0  Train   0.09408    0.43555
1  Train   0.04274    0.56581
2  Train   0.09512    0.63972
```

```
3  Train    0.01086    0.51628
4  Train    0.05426    0.46512
```

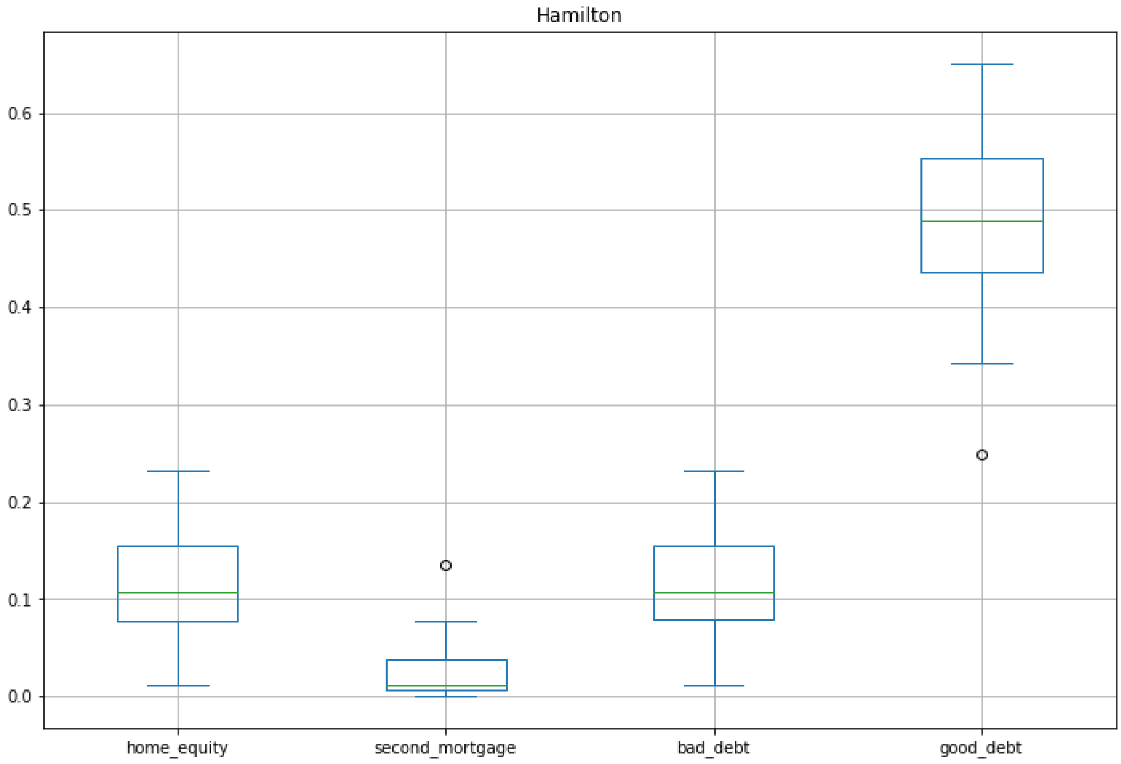
[109]:
```
df_combined.columns
```

[109] :
```
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')
```

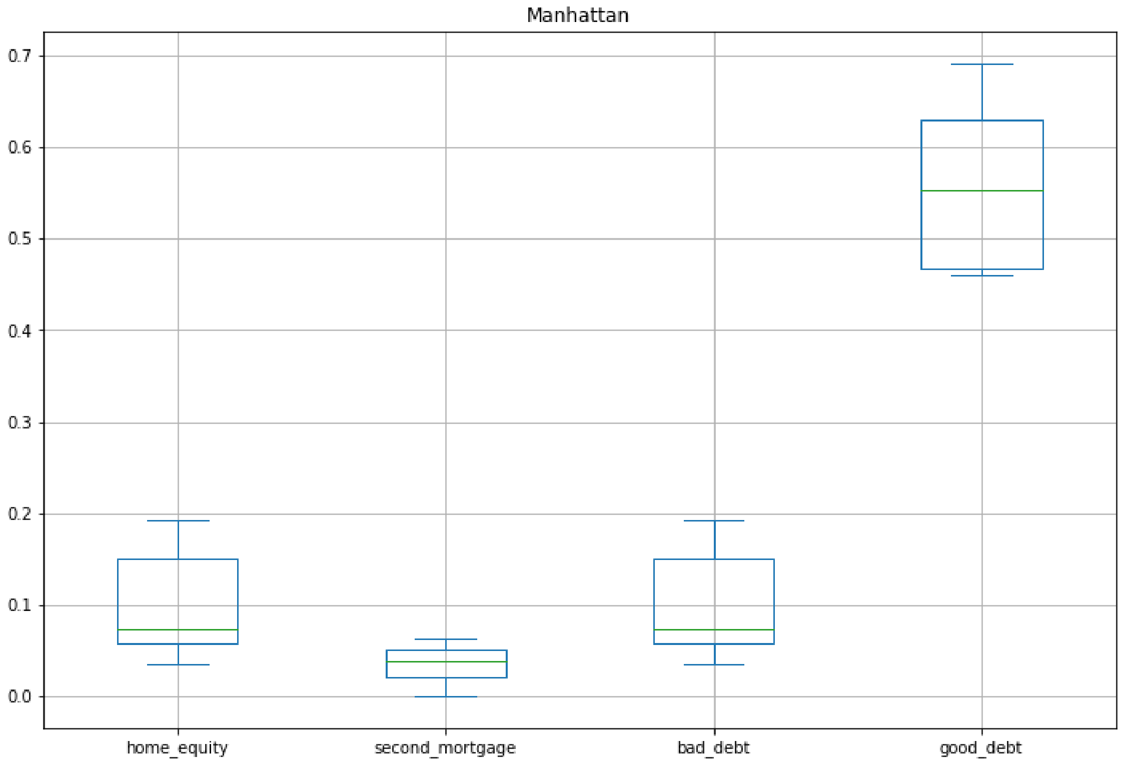
[110]:
```
all_cities = df_combined[["home_equity","second_mortgage","bad_debt","good_debt"]]
all_cities.plot.box(figsize=(12,8),grid=True)
plt.title('All Cities')
plt.show()
```



[111]:
```
hamilton = df_combined[df_combined['city']=='Hamilton']
hamilton = hamilton[['home_equity','second_mortgage','bad_debt', 'good_debt']]
hamilton.plot.box(figsize=(12,8),grid=True)
plt.title('Hamilton')
plt.show()
```

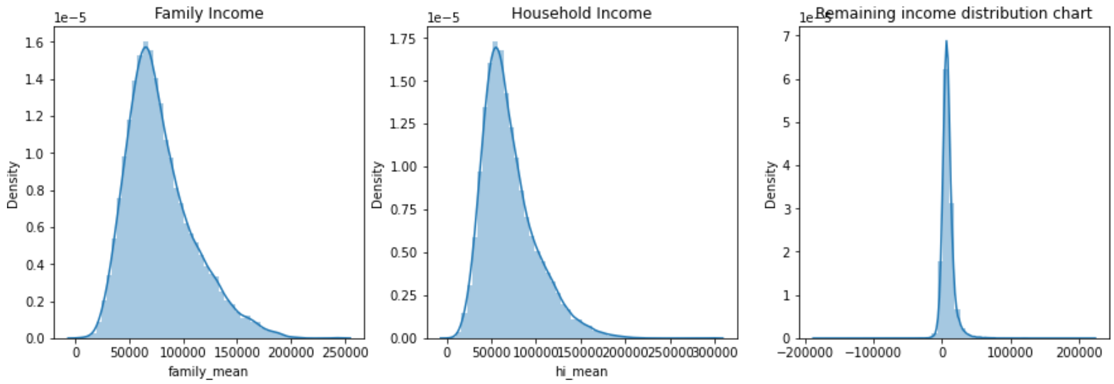


[112]:
```
Manhattan = df_combined[df_combined['city']=='Manhattan']
Manhattan = Manhattan[['home_equity','second_mortgage','bad_debt', 'good_debt']]
Manhattan.plot.box(figsize=(12,8),grid=True)
plt.title('Manhattan')
plt.show()
```



e. Create a collated income distribution chart for family income, house hold income, and remai

[51]:
```
import seaborn as sns
plt.figure(figsize=(15,10))

plt.subplot(2,3,1)
sns.distplot(df_train['family_mean'])
plt.title('Family Income')
plt.subplot(2,3,2)
sns.distplot(df_train['hi_mean'])
plt.title('Household Income')
plt.subplot(2,3,3)
sns.distplot(df_train['family_mean']-df_train['hi_mean'])
plt.title('Remaining income distribution chart')
plt.show()
```



Project Task: Week 2 Exploratory Data Analysis (EDA):

1. 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):
   a. Use pop and ALand variables to create a new field called population density

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

```
[113]: df_combined.head()
```

```
[113]:      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
2  245683       140        63       18        Indiana       IN    Danville
3  279653       140       127       72    Puerto Rico       PR    San Juan
4  247218       140       161       20         Kansas       KS   Manhattan

            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
2        Danville   City   tract     46122        317  39.792202  -86.515246
3        Guaynabo  Urban   tract       927        787  18.396103  -66.104169
4  Manhattan City   City   tract     66502        785  39.195573  -96.569366

         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
2   69561595.0   284193  6881      3643        3238  742.77365        703.0
3    1105793.0        0  2700      1141        1559  803.42018        782.0
4    2554403.0        0  5637      2586        3051  938.56493        881.0

   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
```

```
2   323.39011          291.85520        378.0      0.95238   0.88624
3   297.39258          259.30316        368.0      0.94693   0.87151
4   392.44096         1005.42886       1704.0      0.99286   0.98247

   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
2     0.79630     0.66667     0.39153     0.39153     0.28307     0.15873
3     0.69832     0.61732     0.51397     0.46927     0.35754     0.32961
4     0.91688     0.84740     0.78247     0.60974     0.55455     0.44416

   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
2               459           378  84942.68317    74964.0  56811.62186
3               438           358  48733.67116    37845.0  45100.54010
4              1725          1540  31834.15466    22497.0  34046.50907

   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
2        1155.20980      2488.0  95262.51431        85395.0   49292.67664
3         928.32193      1267.0  56401.68133        44399.0   41082.90515
4        1548.67477      1983.0  54053.42396        50272.0   39609.12605

   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
2             709.74925          1889.0        1506.06758              1361.0
3             490.18479           729.0        1175.28642              1101.0
4             244.08903           395.0        1192.58759              1125.0

   hc_mortgage_stdev  hc_mortgage_sample_weight  hc_mortgage_samples  \
0          641.22898                  377.83135                867.0
1          482.27020                  316.88320                356.0
2          731.89394                  699.41354               1491.0
3          428.98751                  261.28471                437.0
4          327.49674                   76.61052                134.0

     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
2  556.45986      532.0  184.42175       538.0         323.35354
3  288.04047      247.0  185.55887       392.0         314.90566
4  443.68855      444.0   76.12674       124.0          79.55556

   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
2                         0.00000         0.00000      0.09512  0.73484
3                         0.01086         0.01086      0.01086  0.52714
4                         0.05426         0.05426      0.05426  0.51938

   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
2              1.00000          0.46332   0.28704    0.94288         0.94616
3              0.53057          0.82530   0.73727    0.91500         0.90755
4              0.18332          0.65545   0.74967    1.00000         1.00000

   hs_degree_female  male_age_mean  male_age_median  male_age_stdev  \
0           0.92434       42.48574         44.00000        22.97306
1           0.94187       34.84728         32.00000        20.37452
2           0.93952       39.38154         40.83333        22.89769
3           0.92043       48.64749         48.91667        23.05968
4           1.00000       26.07533         22.41667        11.84399

   male_age_sample_weight  male_age_samples  female_age_mean  \
0               696.42136            2612.0         44.48629
1               323.90204            1349.0         36.48391
2               888.29730            3643.0         42.15810
3               274.98956            1141.0         47.77526
4              1296.89877            2586.0         24.17693

   female_age_median  female_age_stdev  female_age_sample_weight  \
0           45.33333          22.51276                 685.33845
1           37.58333          23.43353                 267.23367
2           42.83333          23.94119                 707.01963
3           50.58333          24.32015                 362.20193
4           21.58333          11.10484                1854.48652

   female_age_samples  pct_own  married  married_snp  separated  divorced  \
0              2618.0  0.79046  0.57851      0.01882    0.01240   0.08770
1              1284.0  0.52483  0.34886      0.01426    0.01426   0.09030
2              3238.0  0.85331  0.64745      0.02830    0.01607   0.10657
3              1559.0  0.65037  0.47257      0.02021    0.02021   0.10106
4              3051.0  0.13046  0.12356      0.00000    0.00000   0.03109

   split  bad_debt  good_debt
0  Train   0.09408    0.43555
1  Train   0.04274    0.56581
2  Train   0.09512    0.63972
3  Train   0.01086    0.51628
4  Train   0.05426    0.46512
```

b. Use male_age_median, female_age_median, male_pop, and female_pop to create a new field call

[114]:
```
# Weighted average
#  median_age=((male_age_median  *  male_pop)+(female_age_median*female_pop))/
 ↪(male_pop+female_pop)
#              =((40*10)+(50*30))/40
#              =(400+1500)/40
#              =190/4
#              =47.5
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'])
```

[115]:
```
df_combined.head()
```

[115]:
```
      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
2  245683       140        63       18      Indiana       IN    Danville
3  279653       140       127       72  Puerto Rico       PR    San Juan
4  247218       140       161       20       Kansas       KS   Manhattan

            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
2        Danville   City   tract     46122        317  39.792202 -86.515246
3        Guaynabo  Urban   tract       927        787  18.396103 -66.104169
4  Manhattan City   City   tract     66502        785  39.195573 -96.569366

         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
2   69561595.0   284193  6881      3643        3238  742.77365        703.0
3    1105793.0        0  2700      1141        1559  803.42018        782.0
4    2554403.0        0  5637      2586        3051  938.56493        881.0

   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
2   323.39011           291.85520         378.0     0.95238     0.88624
3   297.39258           259.30316         368.0     0.94693     0.87151
4   392.44096          1005.42886        1704.0     0.99286     0.98247

   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
2     0.79630     0.66667     0.39153     0.39153     0.28307     0.15873
```

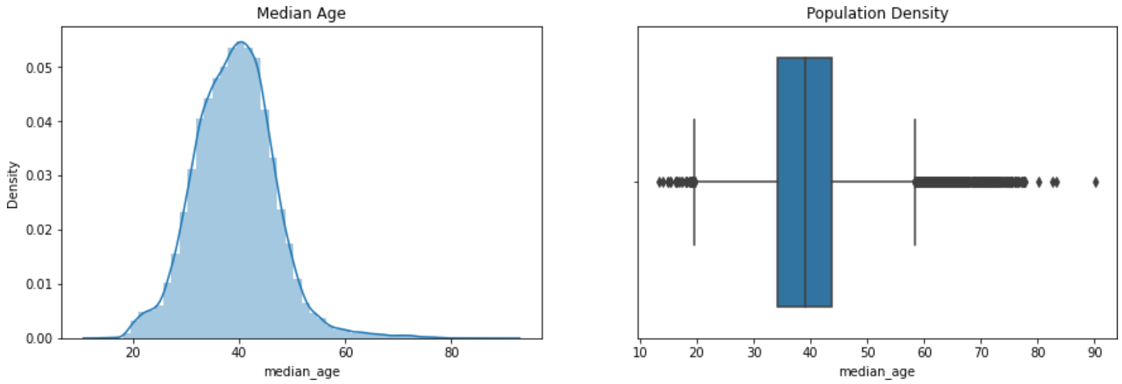
```
3     0.69832     0.61732     0.51397     0.46927     0.35754     0.32961
4     0.91688     0.84740     0.78247     0.60974     0.55455     0.44416

   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
2               459           378   84942.68317    74964.0  56811.62186
3               438           358   48733.67116    37845.0  45100.54010
4              1725          1540   31834.15466    22497.0  34046.50907

   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
2        1155.20980      2488.0  95262.51431        85395.0   49292.67664
3         928.32193      1267.0  56401.68133        44399.0   41082.90515
4        1548.67477      1983.0  54053.42396        50272.0   39609.12605

   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
2             709.74925          1889.0        1506.06758              1361.0
3             490.18479           729.0        1175.28642              1101.0
4             244.08903           395.0        1192.58759              1125.0

   hc_mortgage_stdev  hc_mortgage_sample_weight  hc_mortgage_samples   \
0          641.22898                  377.83135                867.0
1          482.27020                  316.88320                356.0
2          731.89394                  699.41354               1491.0
3          428.98751                  261.28471                437.0
4          327.49674                   76.61052                134.0

     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
2  556.45986      532.0  184.42175       538.0         323.35354
3  288.04047      247.0  185.55887       392.0         314.90566
4  443.68855      444.0   76.12674       124.0          79.55556

   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
2                      0.00000            0.00000      0.09512  0.73484
3                      0.01086            0.01086      0.01086  0.52714
4                      0.05426            0.05426      0.05426  0.51938

   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
2             1.00000       0.46332   0.28704   0.94288        0.94616
3             0.53057       0.82530   0.73727   0.91500        0.90755
4             0.18332       0.65545   0.74967   1.00000        1.00000

   hs_degree_female  male_age_mean  male_age_median  male_age_stdev  \
0           0.92434       42.48574         44.00000        22.97306
1           0.94187       34.84728         32.00000        20.37452
2           0.93952       39.38154         40.83333        22.89769
3           0.92043       48.64749         48.91667        23.05968
4           1.00000       26.07533         22.41667        11.84399

   male_age_sample_weight  male_age_samples  female_age_mean  \
0               696.42136            2612.0         44.48629
1               323.90204            1349.0         36.48391
2               888.29730            3643.0         42.15810
3               274.98956            1141.0         47.77526
4              1296.89877            2586.0         24.17693

   female_age_median  female_age_stdev  female_age_sample_weight  \
0           45.33333          22.51276                 685.33845
1           37.58333          23.43353                 267.23367
2           42.83333          23.94119                 707.01963
3           50.58333          24.32015                 362.20193
4           21.58333          11.10484                1854.48652

   female_age_samples  pct_own  married  married_snp  separated  divorced  \
0              2618.0  0.79046  0.57851      0.01882    0.01240   0.08770
1              1284.0  0.52483  0.34886      0.01426    0.01426   0.09030
2              3238.0  0.85331  0.64745      0.02830    0.01607   0.10657
3              1559.0  0.65037  0.47257      0.02021    0.02021   0.10106
4              3051.0  0.13046  0.12356      0.00000    0.00000   0.03109

   split  bad_debt  good_debt  median_age
0  Train   0.09408    0.43555   44.667430
1  Train   0.04274    0.56581   34.722748
2  Train   0.09512    0.63972   41.774472
3  Train   0.01086    0.51628   49.879012
4  Train   0.05426    0.46512   21.965629
```

c. Visualize the findings using appropriate chart type

[120]:
```
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()
```



2. 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.

[121]:
```
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()
```

[121]:
```
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

[122]:
```
df_combined.groupby(by='pop_bins')[['married','separated','divorced']].count()
```

[122]:
```
           married  separated  divorced
pop_bins
very low     38350      38350     38350
low            348        348       348
medium          12         12        12
high             4          4         4
very high        1          1         1
```

[123]:
```
df_combined.groupby(by='pop_bins')[['married','separated','divorced']].
↪agg(["mean", "median"])
```

[123]:

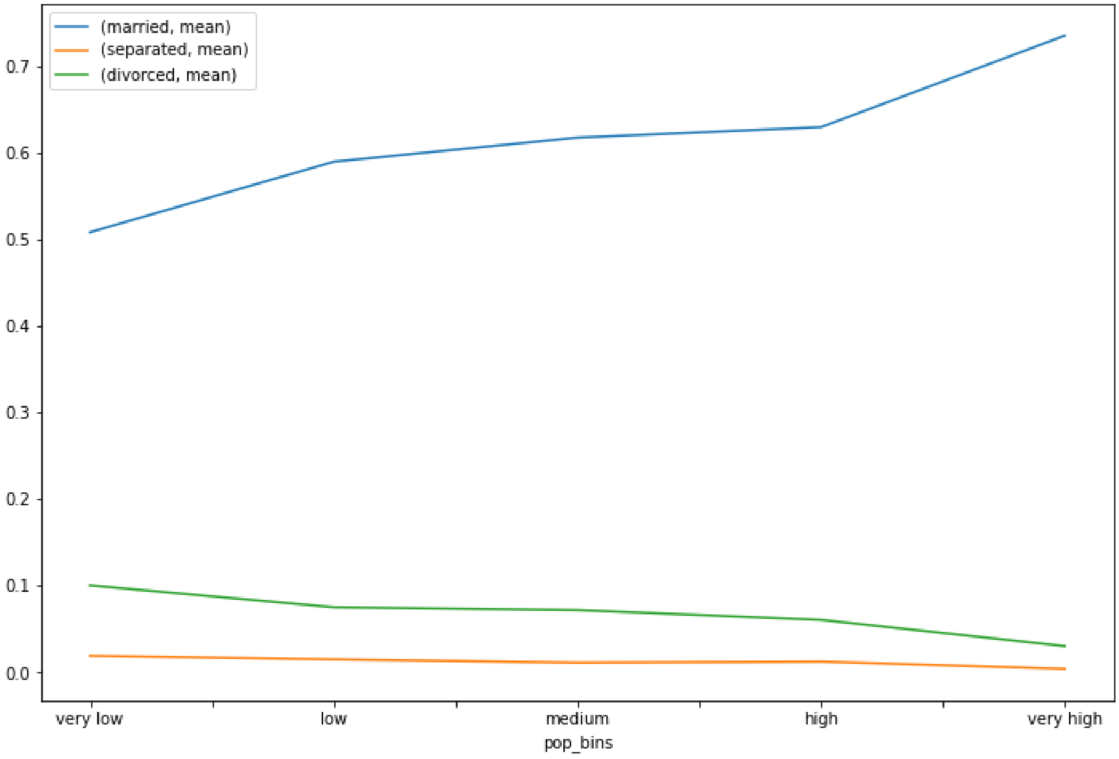
```
           married             separated           divorced
              mean    median      mean    median      mean   median
pop_bins
very low  0.508000 0.526210  0.019127 0.013580  0.100325  0.09510
low       0.589247 0.601815  0.014929 0.010255  0.075192  0.06934
medium    0.617047 0.605765  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
```

b. Visualize using appropriate chart type

[124]:

```
plt.figure(figsize=(12,8))
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()
```

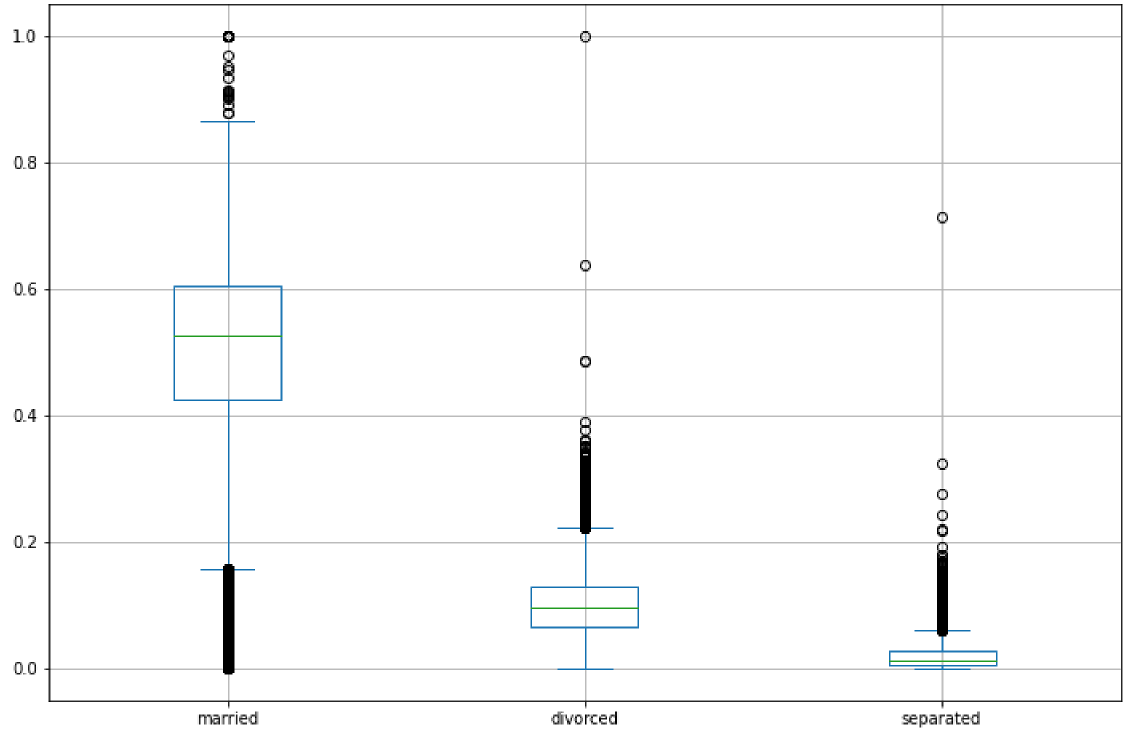
```
<Figure size 864x576 with 0 Axes>
```



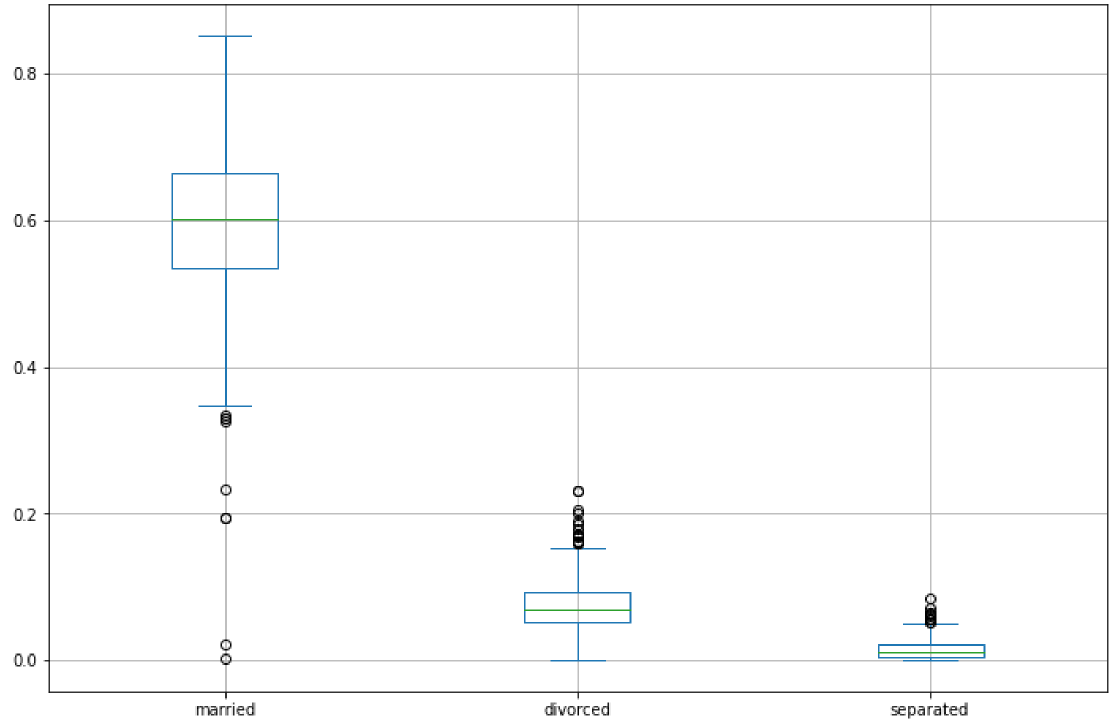
[126]:

```
df_combined.groupby(by="pop_bins")[["married","divorced", "separated"]].plot.
 ↪box(figsize=(12,8),grid="True")
```

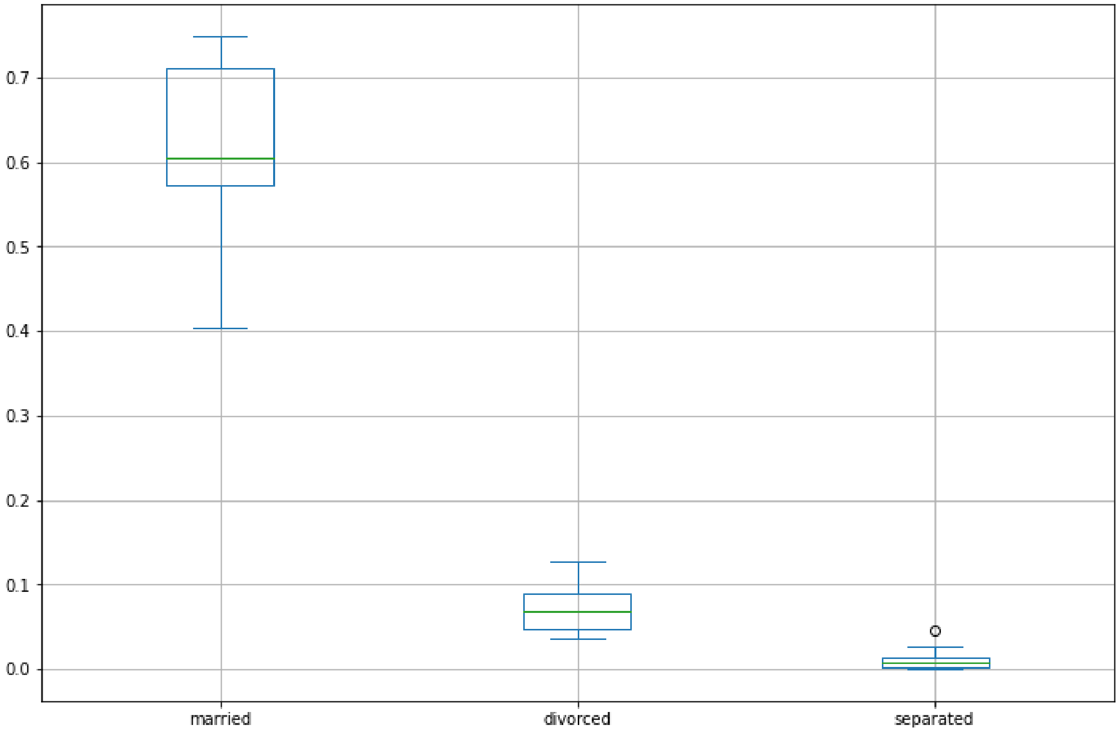
```
plt.show()
```

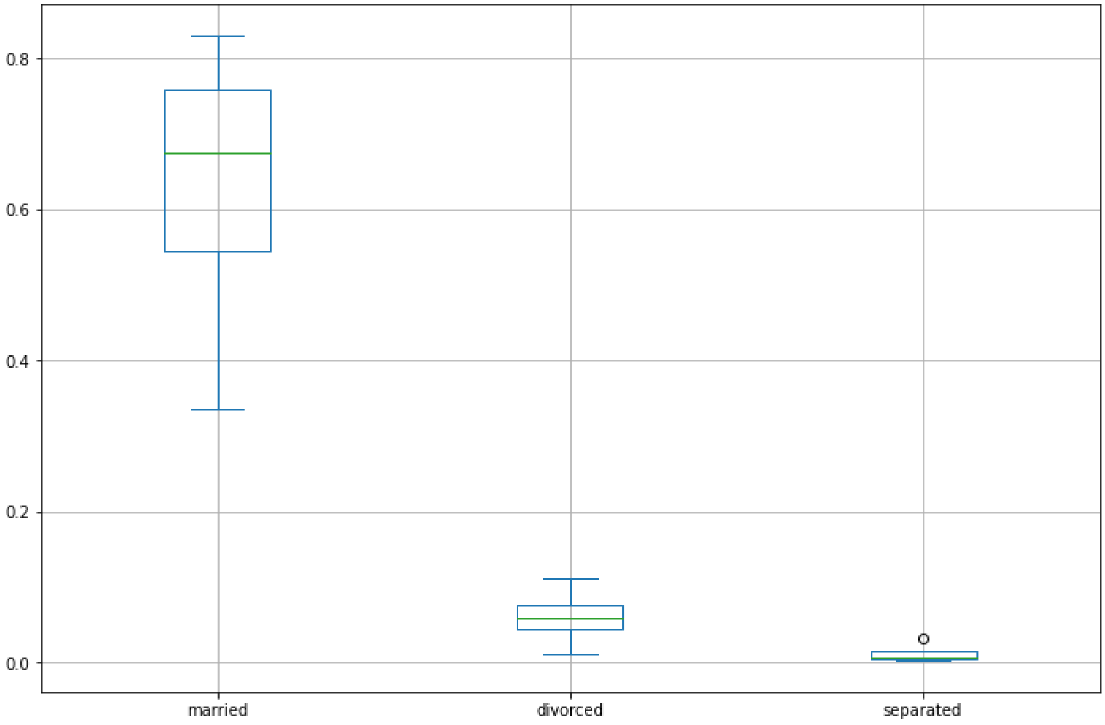


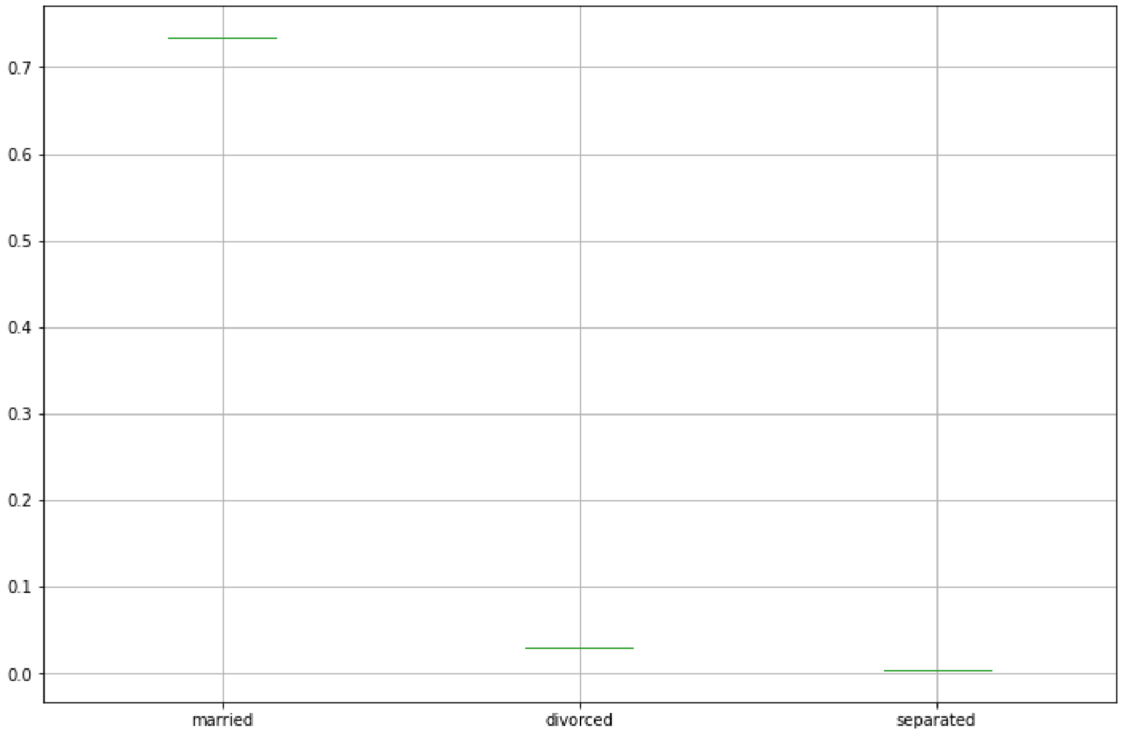
1.0
0.8
0.6
0.4
0.2
0.0
married
divorced
separated



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



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

[127]:
```
rent_state_mean = df_combined.groupby(by="state")["rent_mean"].agg(["mean"])
rent_state_mean.head()
```

[127]:
```
                  mean
state
Alabama     765.872557
Alaska     1190.093590
Arizona    1084.510940
Arkansas    716.544987
California 1466.020465
```

[128]:
```
income_state_mean=df_combined.groupby(by="state")["family_mean"].agg(["mean"])
income_state_mean.head()
```

[128]:
```
                  mean
state
Alabama   65311.510962
Alaska    91911.137520
Arizona   73014.068487
Arkansas  64234.705963
```

```
California     87711.550734
```

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

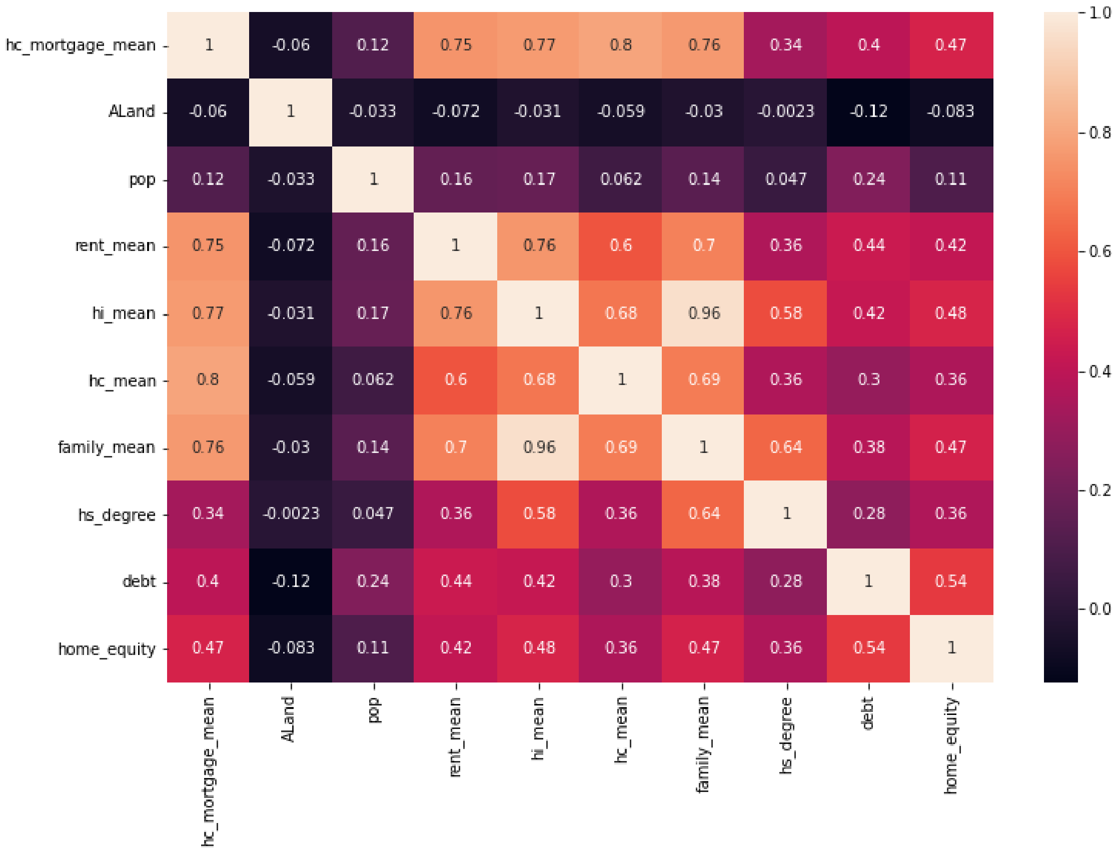
```
[129]: state
       Alabama                 1.172646
       Alaska                  1.294831
       Arizona                 1.485345
       Arkansas                1.115511
       California              1.671411
       Colorado                1.359697
       Connecticut             1.272141
       Delaware                1.311538
       District of Columbia    1.357450
       Florida                 1.576101
       Name: mean, dtype: float64
```

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

```
[130]: 0.013351543786573208
```

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

```
[131]: 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()
```

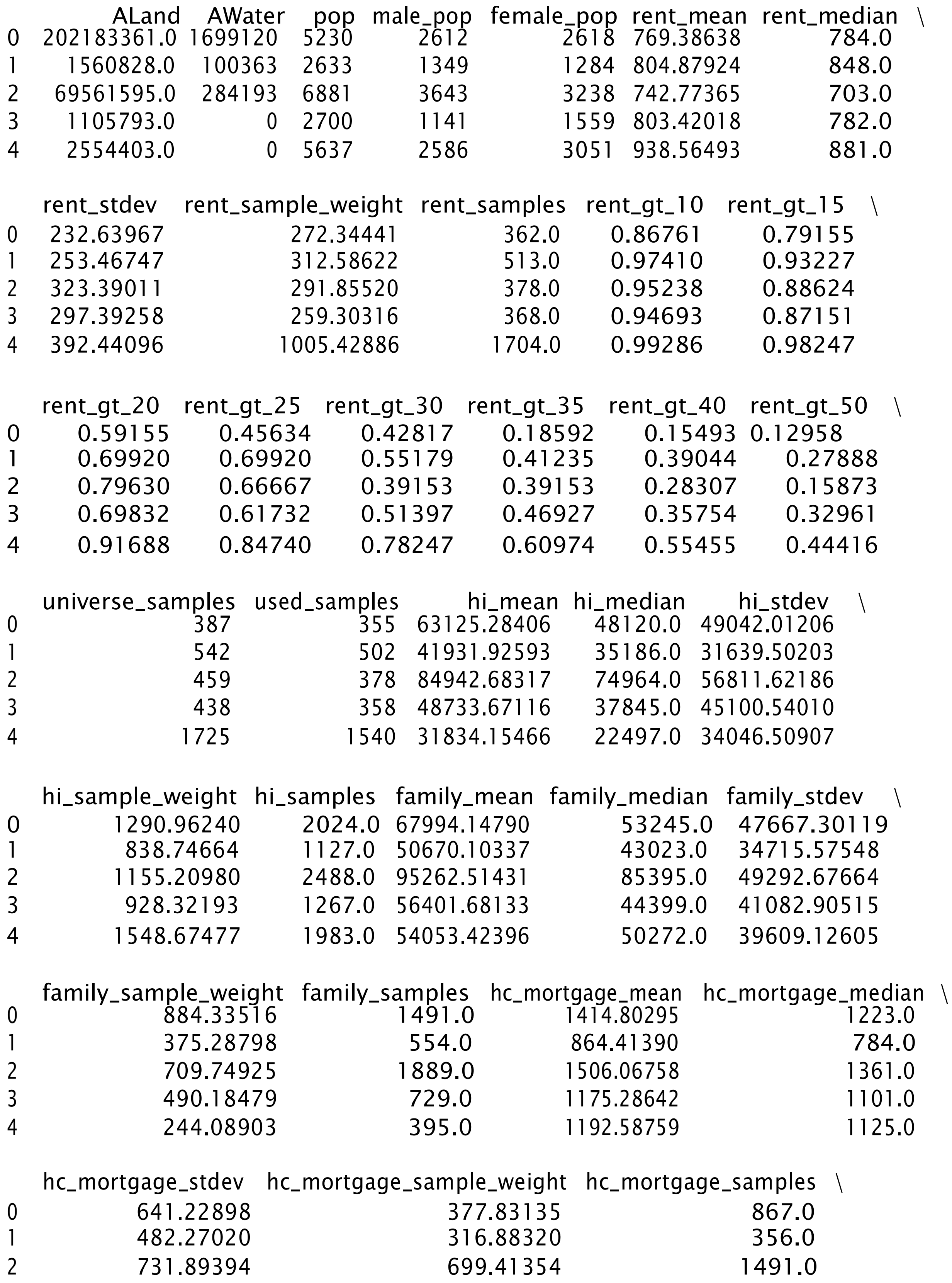


rent_mean, hi_mean, hc_mean, family_mean has a good correlation with the target i.e- hc_mortagage_mean

[132]:
```
train = df_combined[df_combined['split'] == 'Train']
test = df_combined[df_combined['split'] == 'Test']
```

[133]:
```
train.head()
```

[133]:
```
      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
2  245683       140        63       18       Indiana       IN    Danville
3  279653       140       127       72   Puerto Rico       PR    San Juan
4  247218       140       161       20        Kansas       KS   Manhattan

             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
2         Danville   City   tract     46122        317  39.792202 -86.515246
3         Guaynabo  Urban   tract       927        787  18.396103 -66.104169
4   Manhattan City   City   tract     66502        785  39.195573 -96.569366
```



```
          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
2   69561595.0   284193  6881      3643        3238  742.77365        703.0
3    1105793.0        0  2700      1141        1559  803.42018        782.0
4    2554403.0        0  5637      2586        3051  938.56493        881.0

   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
2   323.39011           291.85520         378.0     0.95238     0.88624
3   297.39258           259.30316         368.0     0.94693     0.87151
4   392.44096          1005.42886        1704.0     0.99286     0.98247

   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
2     0.79630     0.66667     0.39153     0.39153     0.28307     0.15873
3     0.69832     0.61732     0.51397     0.46927     0.35754     0.32961
4     0.91688     0.84740     0.78247     0.60974     0.55455     0.44416

   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
2               459           378  84942.68317    74964.0  56811.62186
3               438           358  48733.67116    37845.0  45100.54010
4              1725          1540  31834.15466    22497.0  34046.50907

   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
2        1155.20980      2488.0  95262.51431        85395.0   49292.67664
3         928.32193      1267.0  56401.68133        44399.0   41082.90515
4        1548.67477      1983.0  54053.42396        50272.0   39609.12605

   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
2             709.74925          1889.0        1506.06758              1361.0
3             490.18479           729.0        1175.28642              1101.0
4             244.08903           395.0        1192.58759              1125.0

   hc_mortgage_stdev  hc_mortgage_sample_weight  hc_mortgage_samples  \
0          641.22898                  377.83135                867.0
1          482.27020                  316.88320                356.0
2          731.89394                  699.41354               1491.0
```

```
3        428.98751                  261.28471                  437.0
4        327.49674                   76.61052                  134.0

    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
2  556.45986      532.0  184.42175       538.0          323.35354
3  288.04047      247.0  185.55887       392.0          314.90566
4  443.68855      444.0   76.12674       124.0           79.55556

   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
2                      0.00000           0.00000      0.09512  0.73484
3                      0.01086           0.01086      0.01086  0.52714
4                      0.05426           0.05426      0.05426  0.51938

   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
2               1.00000          0.46332   0.28704    0.94288          0.94616
3               0.53057          0.82530   0.73727    0.91500          0.90755
4               0.18332          0.65545   0.74967    1.00000          1.00000

   hs_degree_female  male_age_mean   male_age_median  male_age_stdev   \
0           0.92434        42.48574          44.00000         22.97306
1           0.94187        34.84728          32.00000         20.37452
2           0.93952        39.38154          40.83333         22.89769
3           0.92043        48.64749          48.91667         23.05968
4           1.00000        26.07533          22.41667         11.84399

   male_age_sample_weight   male_age_samples   female_age_mean   \
0               696.42136             2612.0          44.48629
1               323.90204             1349.0          36.48391
2               888.29730             3643.0          42.15810
3               274.98956             1141.0          47.77526
4              1296.89877             2586.0          24.17693

  female_age_median    female_age_stdev  female_age_sample_weight    \
0         45.33333            22.51276                  685.33845
1         37.58333            23.43353                  267.23367
2         42.83333            23.94119                  707.01963
3         50.58333            24.32015                  362.20193
4         21.58333            11.10484                 1854.48652

   female_age_samples    pct_own  married    married_snp  separated   divorced    \
0              2618.0   0.79046  0.57851        0.01882     0.01240    0.08770
```

```
1            1284.0  0.52483  0.34886      0.01426     0.01426   0.09030
2            3238.0  0.85331  0.64745      0.02830     0.01607   0.10657
3            1559.0  0.65037  0.47257      0.02021     0.02021   0.10106
4            3051.0  0.13046  0.12356      0.00000     0.00000   0.03109

   split  bad_debt  good_debt  median_age  pop_bins
0  Train   0.09408    0.43555   44.667430  very low
1  Train   0.04274    0.56581   34.722748  very low
2  Train   0.09512    0.63972   41.774472  very low
3  Train   0.01086    0.51628   49.879012  very low
4  Train   0.05426    0.46512   21.965629  very low
```

[134]:
```
test.head()
```

[134]:
```
          UID  SUMLEVEL  COUNTYID  STATEID         state  state_ab  \
27321  255504       140       163       26      Michigan        MI
27322  252676       140         1       23         Maine        ME
27323  276314       140        15       42  Pennsylvania        PA
27324  248614       140       231       21      Kentucky        KY
27325  286865       140       355       48         Texas        TX

                 city                  place     type primary  zip_code  \
27321         Detroit  Dearborn Heights City      CDP   tract     48239
27322          Auburn            Auburn City     City   tract      4210
27323       Pine City              Millerton  Borough   tract     14871
27324      Monticello        Monticello City     City   tract     42633
27325  Corpus Christi                  Edroy     Town   tract     78410

       area_code        lat        lng        ALand   AWater   pop  male_pop  \
27321        313  42.346422 -83.252823    2711280.0    39555  3417      1479
27322        207  44.100724 -70.257832   14778785.0  2705204  3796      1846
27323        607  41.948556 -76.783808  258903666.0   863840  3944      2065
27324        606  36.746009 -84.766870  501694825.0  2623067  2508      1427
27325        361  27.882462 -97.678586   13796057.0   497689  6230      3274

       female_pop   rent_mean  rent_median  rent_stdev  rent_sample_weight  \
27321        1938   858.57169        859.0   232.39082           276.07497
27322        1950   832.68625        750.0   267.22342           183.32299
27323        1879   816.00639        755.0   416.25699           141.39063
27324        1081   418.68937        385.0   156.92024            88.95960
27325        2956  1031.63763        997.0   326.76727           277.39844

       rent_samples  rent_gt_10  rent_gt_15  rent_gt_20  rent_gt_25  \
27321         424.0     1.00000     0.95696     0.85316     0.85316
27322         245.0     1.00000     1.00000     0.86611     0.67364
27323         217.0     0.97573     0.93204     0.78641     0.71845
27324          93.0     1.00000     0.93548     0.93548     0.64516
```

```
27325         624.0   0.72276   0.66506   0.53526   0.38301

       rent_gt_30  rent_gt_35  rent_gt_40  rent_gt_50  universe_samples  \
27321     0.85316     0.85316     0.76962     0.63544               435
27322     0.30962     0.30962     0.30962     0.27197               275
27323     0.63592     0.47573     0.43689     0.32524               245
27324     0.55914     0.46237     0.46237     0.36559               153
27325     0.18910     0.16667     0.14263     0.11058               660

       used_samples       hi_mean  hi_median     hi_stdev  hi_sample_weight  \
27321           395   48899.52121    38746.0  44392.20902         798.02401
27322           239   72335.33234    61008.0  51895.81159         922.82969
27323           206   58501.15901    51648.0  45245.27248         893.07759
27324            93   38237.55059    31612.0  34527.61607         775.17947
27325           624  114456.07790    94211.0  81950.95692         836.30759

       hi_samples   family_mean  family_median  family_stdev  \
27321      1180.0   53802.87122        45167.0   43756.56479
27322      1722.0   85642.22095        74759.0   49156.72870
27323      1461.0   65694.06582        57186.0   44239.31893
27324       957.0   44156.38709        34687.0   34899.74300
27325      2404.0  123527.02420       103898.0   72173.55823

       family_sample_weight  family_samples  hc_mortgage_mean  \
27321             464.30972           769.0        1139.24548
27322             482.99945          1147.0        1533.25988
27323             619.73962          1084.0        1254.54462
27324             535.21987           689.0         862.65763
27325             507.42257          1738.0        1996.41425

       hc_mortgage_median  hc_mortgage_stdev  hc_mortgage_sample_weight  \
27321              1109.0          336.47710                  262.67011
27322              1438.0          536.61118                  373.96188
27323              1089.0          596.85204                  340.45884
27324               749.0          624.42157                  299.56752
27325              1907.0          740.21168                  319.97570

       hc_mortgage_samples    hc_mean  hc_median   hc_stdev  hc_samples  \
27321                474.0  488.51323      436.0  192.75147       271.0
27322                937.0  661.31296      668.0  201.31365       510.0
27323                552.0  397.44466      356.0  189.40372       664.0
27324                337.0  200.88113      180.0   91.56490       467.0
27325               1102.0  867.57713      804.0  376.20236       642.0

       hc_sample_weight  home_equity_second_mortgage  second_mortgage  \
27321         189.18182                      0.06443          0.06443
27322         279.69697                      0.01175          0.01175
```

```
27323      534.16737                        0.01069          0.01316
27324      454.85404                        0.00995          0.00995
27325      333.91919                        0.00000          0.00000

       home_equity     debt  second_mortgage_cdf  home_equity_cdf  debt_cdf  \
27321      0.07651  0.63624              0.14111          0.55087   0.51965
27322      0.14375  0.64755              0.52310          0.26442   0.49359
27323      0.06497  0.45395              0.51066          0.60484   0.83848
27324      0.01741  0.41915              0.53770          0.80931   0.87403
27325      0.03440  0.63188              1.00000          0.74519   0.52943

       hs_degree  hs_degree_male  hs_degree_female  male_age_mean  \
27321    0.91047         0.92010           0.90391       33.37131
27322    0.94290         0.92832           0.95736       43.88680
27323    0.89238         0.86003           0.92463       39.81661
27324    0.60908         0.56584           0.65947       41.81638
27325    0.86297         0.87969           0.84466       42.13301

       male_age_median  male_age_stdev  male_age_sample_weight  \
27321         27.83333        22.36768               334.30978
27322         46.08333        22.90302               427.10824
27323         41.91667        24.29111               499.10080
27324         43.00000        24.65325               333.57733
27325         43.75000        22.69502               833.57435

       male_age_samples  female_age_mean  female_age_median  female_age_stdev  \
27321            1479.0         34.78682           33.75000          21.58531
27322            1846.0         44.23451           46.66667          22.37036
27323            2065.0         41.62426           44.50000          22.86213
27324            1427.0         44.81200           48.00000          21.03155
27325            3274.0         40.66618           42.66667          21.30900

       female_age_sample_weight  female_age_samples  pct_own  married  \
27321                 416.48097              1938.0  0.70252  0.28217
27322                 532.03505              1950.0  0.85128  0.64221
27323                 453.11959              1879.0  0.81897  0.59961
27324                 263.94320              1081.0  0.84609  0.56953
27325                 709.90829              2956.0  0.79077  0.57620

       married_snp  separated  divorced split  bad_debt  good_debt  \
27321      0.05910    0.03813   0.14299  Test   0.07651    0.55973
27322      0.02338    0.00000   0.13377  Test   0.14375    0.50380
27323      0.01746    0.01358   0.10026  Test   0.06744    0.38651
27324      0.05492    0.04694   0.12489  Test   0.01741    0.40174
27325      0.01726    0.00588   0.16379  Test   0.03440    0.59748

       median_age  pop_bins
```

```
27321   31.189053  very low
27322   46.382991  very low
27323   43.147420  very low
27324   45.155104  very low
27325   43.235983  very low
```

Project Task: Week 3 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

[135]:

```
!pip install factor_analyzer
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Requirement already satisfied: factor_analyzer in /usr/local/lib/python3.8/dist-
packages (0.4.1)
Requirement already satisfied: scipy in /usr/local/lib/python3.8/dist-packages
(from factor_analyzer) (1.7.3)
Requirement already satisfied: numpy in /usr/local/lib/python3.8/dist-packages
(from factor_analyzer) (1.21.6)
Requirement already satisfied: pandas in /usr/local/lib/python3.8/dist-packages
(from factor_analyzer) (1.3.5)
Requirement already satisfied: pre-commit in /usr/local/lib/python3.8/dist-
packages (from factor_analyzer) (3.0.3)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.8/dist-
packages (from factor_analyzer) (1.0.2)
Requirement already satisfied: pytz>=2017.3 in /usr/local/lib/python3.8/dist-
packages (from pandas->factor_analyzer) (2022.7.1)
Requirement already satisfied: python-dateutil>=2.7.3 in
/usr/local/lib/python3.8/dist-packages (from pandas->factor_analyzer) (2.8.2)
Requirement already satisfied: cfgv>=2.0.0 in /usr/local/lib/python3.8/dist-
packages (from pre-commit->factor_analyzer) (3.3.1)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.8/dist-
packages (from pre-commit->factor_analyzer) (6.0)
Requirement already satisfied: nodeenv>=0.11.1 in /usr/local/lib/python3.8/dist-
```

```
packages (from pre-commit->factor_analyzer) (1.7.0)
Requirement already satisfied: identify>=1.0.0 in /usr/local/lib/python3.8/dist-
packages (from pre-commit->factor_analyzer) (2.5.17)
Requirement already satisfied: virtualenv>=20.10.0 in
/usr/local/lib/python3.8/dist-packages (from pre-commit->factor_analyzer)
(20.17.1)
Requirement already satisfied: threadpoolctl>=2.0.0 in
/usr/local/lib/python3.8/dist-packages (from scikit-learn->factor_analyzer)
(3.1.0)
Requirement already satisfied: joblib>=0.11 in /usr/local/lib/python3.8/dist-
packages (from scikit-learn->factor_analyzer) (1.2.0)
Requirement already satisfied: setuptools in /usr/local/lib/python3.8/dist-
packages (from nodeenv>=0.11.1->pre-commit->factor_analyzer) (57.4.0)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.8/dist-
packages (from python-dateutil>=2.7.3->pandas->factor_analyzer) (1.15.0)
Requirement already satisfied: distlib<1,>=0.3.6 in
/usr/local/lib/python3.8/dist-packages (from virtualenv>=20.10.0->pre-
commit->factor_analyzer) (0.3.6)
Requirement already satisfied: platformdirs<3,>=2.4 in
/usr/local/lib/python3.8/dist-packages (from virtualenv>=20.10.0->pre-
commit->factor_analyzer) (2.6.2)
Requirement already satisfied: filelock<4,>=3.4.1 in
/usr/local/lib/python3.8/dist-packages (from virtualenv>=20.10.0->pre-
commit->factor_analyzer) (3.9.0)
```

[136]:

```
import numpy as np
from sklearn.decomposition import FactorAnalysis
from factor_analyzer import FactorAnalyzer
```

[137]:

```
df_train.describe().T
```

[137]:

```
                count           mean           std       min            25%  \
UID           27321.0  257331.996303  21343.859725  220342.0  238816.000000
BLOCKID           0.0            NaN           NaN       NaN            NaN
SUMLEVEL      27321.0     140.000000      0.000000     140.0     140.000000
COUNTYID      27321.0      85.646426     98.333097       1.0      29.000000
STATEID       27321.0      28.271806     16.392846       1.0      13.000000
…                 …              …             …         …              …
pct_own       27053.0       0.640434      0.226640       0.0       0.502780
married       27130.0       0.508300      0.136860       0.0       0.425102
married_snp   27130.0       0.047537      0.037640       0.0       0.020810
separated     27130.0       0.019089      0.020796       0.0       0.004530
divorced      27130.0       0.100248      0.049055       0.0       0.065800

                        50%            75%           max
UID           257220.000000  275818.000000  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           0.690840       0.817460       1.00000
married           0.526665       0.605760       1.00000
married_snp       0.038840       0.065100       0.71429
separated         0.013460       0.027488       0.71429
divorced          0.095205       0.129000       1.00000

[74 rows x 8 columns]
```

Project Task: Week 4 Data Modeling :

1. 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. Data Modeling :

- 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

[140]:

```
train.columns
```

[140] :

```
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', 'median_age', 'pop_bins'],
       dtype='object')
```

[141]:

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

[141]:

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

[142]:

```

```

[143]:

```
test.replace(type_dict,inplace=True)
```

[144]:

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

[144]:

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

[145]:

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

[145]:

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

[146]:

```
feature_cols=['COUNTYID','STATEID','zip_code','type','pop',␣
 ↪'family_mean','second_mortgage', 'home_equity', 'debt','hs_degree',
               'pct_own', 'married','separated', 'divorced']
```

[147]:

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

[148]:

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

[149]:

```
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
```

[150]:

```
X_train.head()
```

[150]:

```
   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
2        63       18     46122     1  6881  95262.51431          0.00000
3       127       72       927     2  2700  56401.68133          0.01086
4       161       20     66502     1  5637  54053.42396          0.05426
```

```
   home_equity     debt  hs_degree  pct_own  married  separated  divorced
0      0.08919  0.52963    0.89288  0.79046  0.57851    0.01240   0.08770
1      0.04274  0.60855    0.90487  0.52483  0.34886    0.01426   0.09030
2      0.09512  0.73484    0.94288  0.85331  0.64745    0.01607   0.10657
3      0.01086  0.52714    0.91500  0.65037  0.47257    0.02021   0.10106
4      0.05426  0.51938    1.00000  0.13046  0.12356    0.00000   0.03109
```

[151]:
```
X_test.head()
```

[151] :
```
       COUNTYID  STATEID  zip_code  type   pop   family_mean  second_mortgage  \
27321       163       26     48239     4  3417   53802.87122          0.06443
27322         1       23      4210     1  3796   85642.22095          0.01175
27323        15       42     14871     6  3944   65694.06582          0.01316
27324       231       21     42633     1  2508   44156.38709          0.00995
27325       355       48     78410     3  6230  123527.02420          0.00000

       home_equity     debt  hs_degree  pct_own  married  separated  divorced
27321      0.07651  0.63624    0.91047  0.70252  0.28217    0.03813   0.14299
27322      0.14375  0.64755    0.94290  0.85128  0.64221    0.00000   0.13377
27323      0.06497  0.45395    0.89238  0.81897  0.59961    0.01358   0.10026
27324      0.01741  0.41915    0.60908  0.84609  0.56953    0.04694   0.12489
27325      0.03440  0.63188    0.86297  0.79077  0.57620    0.00588   0.16379
```

[152]:
```
sc = StandardScaler()
X_train_scaled = sc.fit_transform(X_train)
X_test_scaled = sc.fit_transform(X_test)
```

a. Run a model at a Nation level. If the accuracy levels and R square are not satisfactory pro

[153]:
```
lr = LinearRegression()
lr.fit(X_train_scaled, y_train)
```

[153]:
```
LinearRegression()
```

[154]:
```
y_pred= lr.predict(X_test_scaled)
```

R square of 60 percent and above should be achieved

[155]:
```
r2_score(y_test,y_pred)
```

[155]:
```
0.7381882934134452
```

[156]:
```
mean_absolute_error(y_test, y_pred)
```

[156]:
```
233.8696569414009
```

[157]:
```
mean_squared_error(y_test, y_pred)
```

[157]:
```
103818.40486733473
```

[158]:
```
np.sqrt(mean_squared_error(y_test,y_pred))
```

[158]:
```
322.20863561880947
```

[159]:
```
r2_score(y_train,  lr.predict(X_train_scaled))
```

[159]:
```
0.734344756627955
```

[160]:
```
lr.coef_
```

[160]:
```
array([ -28.50842455,  -21.7100607 ,  -22.98370175,  -57.43101333,
         -4.78426374,  558.7402445 ,   -0.55955638,   70.89657588,
         12.81271881, -113.18431746, -176.51983734,    8.10645154,
          5.24214879,  -55.79637445])
```

[161]:
```
X_train.columns
```

[161]:
```
Index(['COUNTYID', 'STATEID', 'zip_code', 'type', 'pop', 'family_mean',
       'second_mortgage', 'home_equity', 'debt', 'hs_degree', 'pct_own',
       'married', 'separated', 'divorced'],
      dtype='object')
```

b. Run another model at State level. There are 52 states in USA.

[162]:
```
state = train['STATEID'].unique()
state
```

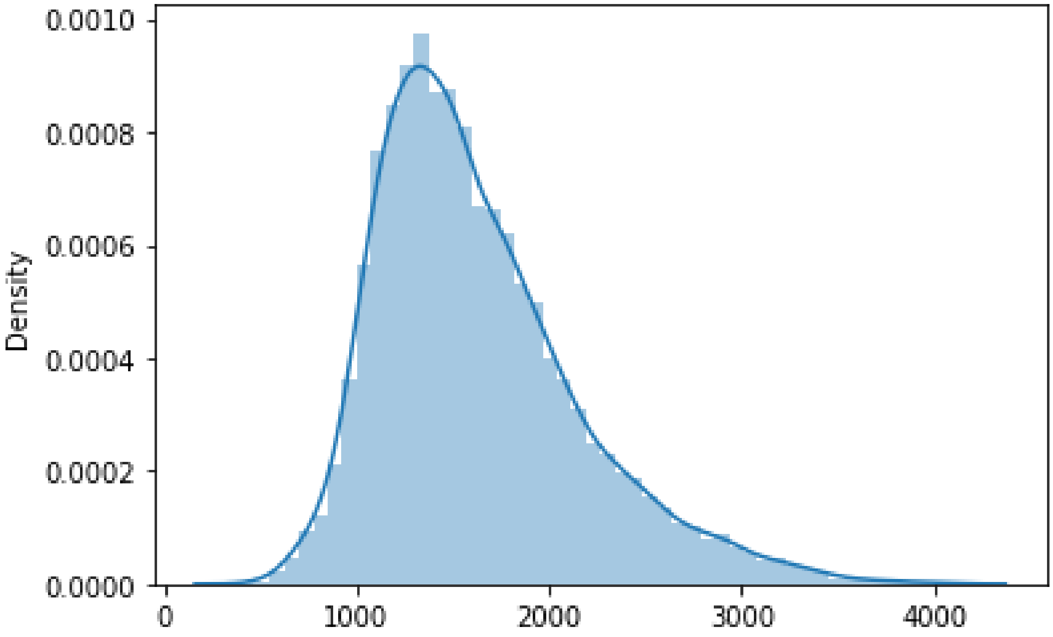
[162]:
```
array([36, 18, 72, 20,  1, 48, 45,  6,  5, 24, 17, 19, 47, 32, 22,  8, 44,
       28, 34, 41,  4, 12, 55, 42, 37, 51, 26, 39, 40, 13, 16, 46, 27, 29,
       53, 56,  9, 54, 21, 25, 11, 15, 30,  2, 33, 49, 50, 31, 38, 35, 23,
       10])
```

[163]:
```
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 :- 0.7458953509562303
Overall RMSE of linear regression model for state, 11 :- 238.52276788095125


State ID- 1
Overall R2 score of linear regression model for state, 1 :- 0.8086161640279984
Overall RMSE of linear regression model for state, 1 :- 311.532907203562


State ID- 29
Overall R2 score of linear regression model for state, 29 :- 0.7090032526359473
Overall RMSE of linear regression model for state, 29 :- 270.06841264277546
```

Test if predicted variable is normally distributed

[164]:
```
sns.distplot(y_pred)
plt.show()
```



Data Reporting:

2. Create a dashboard in tableau by choosing appropriate chart types and metrics useful for the business. The dashboard must entail the following:
   a. Box plot of distribution of average rent by type of place (village, urban, town, etc.).
   b. Pie charts to show overall debt and bad debt.
   c. 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.
   d. Heat map for correlation matrix.
   e. Pie chart to show the population distribution across different types of places (village, urban, town etc.)

0.0.1 PLEASE REFER TABLEAU FILE FOR DASHBOARD AND VISUALIZATION CREATED FOR DATA REPORTING.

0.0.2 Link : https://public.tableau.com/app/profile/santhosh.tn/viz/RealEstateSimplilearn_1678

[ ]: