Real estate project

January 21, 2024

[1]:
```
import time
import random
from math import *
import operator
import pandas as pd
import numpy as np

# import plotting libraries
import matplotlib
import matplotlib.pyplot as plt
from pandas.plotting import scatter_matrix
%matplotlib inline

import seaborn as sns
sns.set(style="white", color_codes=True)
sns.set(font_scale=1.5)
```

[2]:
```
df_train=pd.read_csv("train.csv")
```

[3]:
```
df_test=pd.read_csv("test.csv")
```

[4]:
```
df_train.columns
```

[4]:
```
Index(['UID', 'BLOCKID', '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'],
dtype='object')
```

[5]:
```
df_test.columns
```

[5]:
```
Index(['UID', 'BLOCKID', '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'],
      dtype='object')
```

[6]:
```
len(df_train)
```

[6]: 27321

[7]:
```
len(df_test)
```

[7]: 11709

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

[8]:
```
      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  … female_age_mean  female_age_median  \
0   Hamilton        Hamilton   City  …        44.48629           45.33333
1 South Bend        Roseland   City  …        36.48391           37.58333
2   Danville        Danville   City  …        42.15810           42.83333
3   San Juan        Guaynabo  Urban  …        47.77526           50.58333
4  Manhattan  Manhattan City   City  …        24.17693           21.58333

   female_age_stdev  female_age_sample_weight  female_age_samples  pct_own  \
0          22.51276                 685.33845              2618.0  0.79046
1          23.43353                 267.23367              1284.0  0.52483
2          23.94119                 707.01963              3238.0  0.85331
3          24.32015                 362.20193              1559.0  0.65037
4          11.10484                1854.48652              3051.0  0.13046

   married  married_snp  separated  divorced
0  0.57851      0.01882    0.01240   0.08770
1  0.34886      0.01426    0.01426   0.09030
2  0.64745      0.02830    0.01607   0.10657
3  0.47257      0.02021    0.02021   0.10106
4  0.12356      0.00000    0.00000   0.03109

[5 rows x 80 columns]
```

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

[9]:
```
      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  … female_age_mean  \
0         Detroit  Dearborn Heights City      CDP  …        34.78682
1          Auburn            Auburn City     City  …        44.23451
2       Pine City              Millerton  Borough  …        41.62426
3      Monticello        Monticello City     City  …        44.81200
4  Corpus Christi                  Edroy     Town  …        40.66618

   female_age_median  female_age_stdev  female_age_sample_weight  \
0           33.75000          21.58531                 416.48097
1           46.66667          22.37036                 532.03505
2           44.50000          22.86213                 453.11959
3           48.00000          21.03155                 263.94320
4           42.66667          21.30900                 709.90829

   female_age_samples  pct_own  married  married_snp  separated  divorced
```

```
0              1938.0  0.70252  0.28217      0.05910     0.03813    0.14299
1              1950.0  0.85128  0.64221      0.02338     0.00000    0.13377
2              1879.0  0.81897  0.59961      0.01746     0.01358    0.10026
3              1081.0  0.84609  0.56953      0.05492     0.04694    0.12489
4              2956.0  0.79077  0.57620      0.01726     0.00588    0.16379

[5 rows x 80 columns]
```

[10]:
```
df_train.describe()
```

[10]:
```
                 UID  BLOCKID  SUMLEVEL      COUNTYID       STATEID  \
count   27321.000000      0.0   27321.0  27321.000000  27321.000000
mean   257331.996303      NaN     140.0     85.646426     28.271806
std     21343.859725      NaN       0.0     98.333097     16.392846
min    220342.000000      NaN     140.0      1.000000      1.000000
25%    238816.000000      NaN     140.0     29.000000     13.000000
50%    257220.000000      NaN     140.0     63.000000     28.000000
75%    275818.000000      NaN     140.0    109.000000     42.000000
max    294334.000000      NaN     140.0    840.000000     72.000000

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

       …  female_age_mean  female_age_median  female_age_stdev  \
count  …     27115.000000       27115.000000      27115.000000
mean   …        40.319803          40.355099         22.178745
std    …         5.886317           8.039585          2.540257
min    …        16.008330          13.250000          0.556780
25%    …        36.892050          34.916670         21.312135
50%    …        40.373320          40.583330         22.514410
75%    …        43.567120          45.416670         23.575260
max    …        79.837390          82.250000         30.241270

       female_age_sample_weight  female_age_samples       pct_own  \
count              27115.000000        27115.000000  27053.000000
mean                 544.238432         2208.761903      0.640434
std                  283.546896         1089.316999      0.226640
min                    0.664700            2.000000      0.000000
25%                  355.995825         1471.000000      0.502780
50%                  503.643890         2066.000000      0.690840
```

```
75%                680.275055        2772.000000      0.817460
max               6197.995200       27250.000000      1.000000

            married   married_snp     separated      divorced
count  27130.000000  27130.000000  27130.000000  27130.000000
mean       0.508300      0.047537      0.019089      0.100248
std        0.136860      0.037640      0.020796      0.049055
min        0.000000      0.000000      0.000000      0.000000
25%        0.425102      0.020810      0.004530      0.065800
50%        0.526665      0.038840      0.013460      0.095205
75%        0.605760      0.065100      0.027488      0.129000
max        1.000000      0.714290      0.714290      1.000000

[8 rows x 74 columns]
```

[11]:
```
df_test.describe()
```

[11]:
```
                 UID  BLOCKID  SUMLEVEL      COUNTYID       STATEID  \
count   11709.000000      0.0   11709.0  11709.000000  11709.000000
mean   257525.004783      NaN     140.0     85.710650     28.489196
std     21466.372658      NaN       0.0     99.304334     16.607262
min    220336.000000      NaN     140.0      1.000000      1.000000
25%    238819.000000      NaN     140.0     29.000000     13.000000
50%    257651.000000      NaN     140.0     61.000000     28.000000
75%    276300.000000      NaN     140.0    109.000000     42.000000
max    294333.000000      NaN     140.0    810.000000     72.000000

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

       …  female_age_mean  female_age_median  female_age_stdev  \
count  …     11613.000000       11613.000000      11613.000000
mean   …        40.111999          40.131864         22.148145
std    …         5.851192           7.972026          2.554907
min    …        15.360240          12.833330          0.737110
25%    …        36.729210          34.750000         21.270920
50%    …        40.196960          40.333330         22.472990
75%    …        43.496490          45.333330         23.549450
max    …        90.107940          90.166670         29.626680
```

```
       female_age_sample_weight  female_age_samples       pct_own  \
count                11613.000000        11613.000000  11587.000000
mean                   550.411243         2233.003186      0.634194
std                    280.992521         1072.017063      0.232232
min                      0.251910            3.000000      0.000000
25%                    363.225840         1499.000000      0.492500
50%                    509.103610         2099.000000      0.687640
75%                    685.883910         2800.000000      0.815235
max                   4145.557870        15466.000000      1.000000

            married   married_snp     separated      divorced
count  11625.000000  11625.000000  11625.000000  11625.000000
mean       0.505632      0.047960      0.019346      0.099191
std        0.139774      0.038693      0.021428      0.048525
min        0.000000      0.000000      0.000000      0.000000
25%        0.422020      0.020890      0.004500      0.064590
50%        0.525270      0.038680      0.013870      0.094350
75%        0.605660      0.065340      0.027910      0.128400
max        1.000000      0.714290      0.714290      0.362750

[8 rows x 74 columns]
```

[12]:
```
df_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27321 entries, 0 to 27320
Data columns (total 80 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   UID                          27321 non-null  int64
 1   BLOCKID                      0 non-null      float64
 2   SUMLEVEL                     27321 non-null  int64
 3   COUNTYID                     27321 non-null  int64
 4   STATEID                      27321 non-null  int64
 5   state                        27321 non-null  object
 6   state_ab                     27321 non-null  object
 7   city                         27321 non-null  object
 8   place                        27321 non-null  object
 9   type                         27321 non-null  object
 10  primary                      27321 non-null  object
 11  zip_code                     27321 non-null  int64
 12  area_code                    27321 non-null  int64
 13  lat                          27321 non-null  float64
 14  lng                          27321 non-null  float64
 15  ALand                        27321 non-null  float64
 16  AWater                       27321 non-null  int64
 17  pop                          27321 non-null  int64
```

```
18  male_pop                     27321 non-null  int64
19  female_pop                   27321 non-null  int64
20  rent_mean                    27007 non-null  float64
21  rent_median                  27007 non-null  float64
22  rent_stdev                   27007 non-null  float64
23  rent_sample_weight           27007 non-null  float64
24  rent_samples                 27007 non-null  float64
25  rent_gt_10                   27007 non-null  float64
26  rent_gt_15                   27007 non-null  float64
27  rent_gt_20                   27007 non-null  float64
28  rent_gt_25                   27007 non-null  float64
29  rent_gt_30                   27007 non-null  float64
30  rent_gt_35                   27007 non-null  float64
31  rent_gt_40                   27007 non-null  float64
32  rent_gt_50                   27007 non-null  float64
33  universe_samples             27321 non-null  int64
34  used_samples                 27321 non-null  int64
35  hi_mean                      27053 non-null  float64
36  hi_median                    27053 non-null  float64
37  hi_stdev                     27053 non-null  float64
38  hi_sample_weight             27053 non-null  float64
39  hi_samples                   27053 non-null  float64
40  family_mean                  27023 non-null  float64
41  family_median                27023 non-null  float64
42  family_stdev                 27023 non-null  float64
43  family_sample_weight         27023 non-null  float64
44  family_samples               27023 non-null  float64
45  hc_mortgage_mean             26748 non-null  float64
46  hc_mortgage_median           26748 non-null  float64
47  hc_mortgage_stdev            26748 non-null  float64
48  hc_mortgage_sample_weight    26748 non-null  float64
49  hc_mortgage_samples          26748 non-null  float64
50  hc_mean                      26721 non-null  float64
51  hc_median                    26721 non-null  float64
52  hc_stdev                     26721 non-null  float64
53  hc_samples                   26721 non-null  float64
54  hc_sample_weight             26721 non-null  float64
55  home_equity_second_mortgage  26864 non-null  float64
56  second_mortgage              26864 non-null  float64
57  home_equity                  26864 non-null  float64
58  debt                         26864 non-null  float64
59  second_mortgage_cdf          26864 non-null  float64
60  home_equity_cdf              26864 non-null  float64
61  debt_cdf                     26864 non-null  float64
62  hs_degree                    27131 non-null  float64
63  hs_degree_male               27121 non-null  float64
64  hs_degree_female             27098 non-null  float64
65  male_age_mean                27132 non-null  float64
```

```
 66  male_age_median             27132 non-null  float64
 67  male_age_stdev              27132 non-null  float64
 68  male_age_sample_weight      27132 non-null  float64
 69  male_age_samples            27132 non-null  float64
 70  female_age_mean             27115 non-null  float64
 71  female_age_median           27115 non-null  float64
 72  female_age_stdev            27115 non-null  float64
 73  female_age_sample_weight    27115 non-null  float64
 74  female_age_samples          27115 non-null  float64
 75  pct_own                     27053 non-null  float64
 76  married                     27130 non-null  float64
 77  married_snp                 27130 non-null  float64
 78  separated                   27130 non-null  float64
 79  divorced                    27130 non-null  float64
dtypes: float64(62), int64(12), object(6)
memory usage: 16.7+ MB
```

[13]:
```
df_test.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11709 entries, 0 to 11708
Data columns (total 80 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   UID                         11709 non-null  int64
 1   BLOCKID                     0 non-null      float64
 2   SUMLEVEL                    11709 non-null  int64
 3   COUNTYID                    11709 non-null  int64
 4   STATEID                     11709 non-null  int64
 5   state                       11709 non-null  object
 6   state_ab                    11709 non-null  object
 7   city                        11709 non-null  object
 8   place                       11709 non-null  object
 9   type                        11709 non-null  object
 10  primary                     11709 non-null  object
 11  zip_code                    11709 non-null  int64
 12  area_code                   11709 non-null  int64
 13  lat                         11709 non-null  float64
 14  lng                         11709 non-null  float64
 15  ALand                       11709 non-null  int64
 16  AWater                      11709 non-null  int64
 17  pop                         11709 non-null  int64
 18  male_pop                    11709 non-null  int64
 19  female_pop                  11709 non-null  int64
 20  rent_mean                   11561 non-null  float64
 21  rent_median                 11561 non-null  float64
 22  rent_stdev                  11561 non-null  float64
 23  rent_sample_weight          11561 non-null  float64
```

```
24  rent_samples                11561 non-null  float64
25  rent_gt_10                  11560 non-null  float64
26  rent_gt_15                  11560 non-null  float64
27  rent_gt_20                  11560 non-null  float64
28  rent_gt_25                  11560 non-null  float64
29  rent_gt_30                  11560 non-null  float64
30  rent_gt_35                  11560 non-null  float64
31  rent_gt_40                  11560 non-null  float64
32  rent_gt_50                  11560 non-null  float64
33  universe_samples            11709 non-null  int64
34  used_samples                11709 non-null  int64
35  hi_mean                     11587 non-null  float64
36  hi_median                   11587 non-null  float64
37  hi_stdev                    11587 non-null  float64
38  hi_sample_weight            11587 non-null  float64
39  hi_samples                  11587 non-null  float64
40  family_mean                 11573 non-null  float64
41  family_median               11573 non-null  float64
42  family_stdev                11573 non-null  float64
43  family_sample_weight        11573 non-null  float64
44  family_samples              11573 non-null  float64
45  hc_mortgage_mean            11441 non-null  float64
46  hc_mortgage_median          11441 non-null  float64
47  hc_mortgage_stdev           11441 non-null  float64
48  hc_mortgage_sample_weight   11441 non-null  float64
49  hc_mortgage_samples         11441 non-null  float64
50  hc_mean                     11419 non-null  float64
51  hc_median                   11419 non-null  float64
52  hc_stdev                    11419 non-null  float64
53  hc_samples                  11419 non-null  float64
54  hc_sample_weight            11419 non-null  float64
55  home_equity_second_mortgage 11489 non-null  float64
56  second_mortgage             11489 non-null  float64
57  home_equity                 11489 non-null  float64
58  debt                        11489 non-null  float64
59  second_mortgage_cdf         11489 non-null  float64
60  home_equity_cdf             11489 non-null  float64
61  debt_cdf                    11489 non-null  float64
62  hs_degree                   11624 non-null  float64
63  hs_degree_male              11620 non-null  float64
64  hs_degree_female            11604 non-null  float64
65  male_age_mean               11625 non-null  float64
66  male_age_median             11625 non-null  float64
67  male_age_stdev              11625 non-null  float64
68  male_age_sample_weight      11625 non-null  float64
69  male_age_samples            11625 non-null  float64
70  female_age_mean             11613 non-null  float64
71  female_age_median           11613 non-null  float64
```

```
 72  female_age_stdev             11613 non-null  float64
 73  female_age_sample_weight     11613 non-null  float64
 74  female_age_samples           11613 non-null  float64
 75  pct_own                      11587 non-null  float64
 76  married                      11625 non-null  float64
 77  married_snp                  11625 non-null  float64
 78  separated                    11625 non-null  float64
 79  divorced                     11625 non-null  float64
dtypes: float64(61), int64(13), object(6)
memory usage: 7.1+ MB
```

[14]:
```
#UID is unique userID value in the train and test dataset. So an index can be␣
 ↪created from the UID feature
df_train.set_index(keys=['UID'],inplace=True)#Set the DataFrame index using␣
 ↪existing columns.
df_test.set_index(keys=['UID'],inplace=True)
```

[15]:
```
df_train.head(2)
```

[15]:
```
        BLOCKID  SUMLEVEL  COUNTYID  STATEID     state state_ab        city  \
UID
267822      NaN       140        53       36  New York       NY    Hamilton
246444      NaN       140       141       18   Indiana       IN  South Bend

           place  type primary  …  female_age_mean  female_age_median  \
UID                             …
267822  Hamilton  City   tract  …         44.48629           45.33333
246444  Roseland  City   tract  …         36.48391           37.58333

        female_age_stdev  female_age_sample_weight  female_age_samples  \
UID
267822          22.51276                 685.33845              2618.0
246444          23.43353                 267.23367              1284.0

        pct_own  married  married_snp  separated  divorced
UID
267822  0.79046  0.57851      0.01882    0.01240    0.0877
246444  0.52483  0.34886      0.01426    0.01426    0.0903

[2 rows x 79 columns]
```

[16]:
```
df_test.head(2)
```

[16]:
```
        BLOCKID  SUMLEVEL  COUNTYID  STATEID     state state_ab     city  \
UID
255504      NaN       140       163       26  Michigan       MI  Detroit
252676      NaN       140         1       23     Maine       ME   Auburn
```

```
                       place  type primary  …  female_age_mean  \
UID                                          …
255504  Dearborn Heights City   CDP   tract  …         34.78682
252676            Auburn City  City   tract  …         44.23451

        female_age_median  female_age_stdev  female_age_sample_weight  \
UID
255504           33.75000          21.58531                 416.48097
252676           46.66667          22.37036                 532.03505

        female_age_samples  pct_own  married  married_snp  separated  divorced
UID
255504              1938.0  0.70252  0.28217      0.05910    0.03813   0.14299
252676              1950.0  0.85128  0.64221      0.02338    0.00000   0.13377

[2 rows x 79 columns]
```

[17]:
```python
#percantage of missing values in train set
missing_list_train=df_train.isnull().sum() *100/len(df_train)
missing_values_df_train=pd.DataFrame(missing_list_train,columns=['Percantage of missing values'])
missing_values_df_train.sort_values(by=['Percantage of missing values'],inplace=True,ascending=False)
missing_values_df_train[missing_values_df_train['Percantage of missing values'] >0][:10]
#BLOCKID can be dropped, since it is 100%missing values
```

[17]:
```
                          Percantage of missing values
BLOCKID                                     100.000000
hc_samples                                    2.196113
hc_mean                                       2.196113
hc_median                                     2.196113
hc_stdev                                      2.196113
hc_sample_weight                              2.196113
hc_mortgage_mean                              2.097288
hc_mortgage_stdev                             2.097288
hc_mortgage_sample_weight                     2.097288
hc_mortgage_samples                           2.097288
```

[18]:
```python
#percantage of missing values in test set
missing_list_test=df_test.isnull().sum() *100/len(df_train)
missing_values_df_test=pd.DataFrame(missing_list_test,columns=['Percantage of missing values'])
missing_values_df_test.sort_values(by=['Percantage of missing values'],inplace=True,ascending=False)
```

```
missing_values_df_test[missing_values_df_test['Percantage of missing values']␣
↪>0][:10]
#BLOCKID can be dropped
```

[18]:

```
                              Percantage of missing values
BLOCKID                                          42.857143
hc_samples                                        1.061455
hc_mean                                           1.061455
hc_median                                         1.061455
hc_stdev                                          1.061455
hc_sample_weight                                  1.061455
hc_mortgage_mean                                  0.980930
hc_mortgage_stdev                                 0.980930
hc_mortgage_sample_weight                         0.980930
hc_mortgage_samples                               0.980930
```

[19]:

```
df_train .drop(columns=['BLOCKID','SUMLEVEL'],inplace=True) #SUMLEVEL doest not␣
↪have any predictive power and no variance
```

[20]:

```
df_test .drop(columns=['BLOCKID','SUMLEVEL'],inplace=True) #SUMLEVEL doest not␣
↪have any predictive power
```

[21]:

```
# Imputing  missing values with mean
missing_train_cols=[]
for col in df_train.columns:
    if df_train[col].isna().sum() !=0:
         missing_train_cols.append(col)
print(missing_train_cols)
```

```
['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']
```

[22]:

```
# Imputing  missing values with mean
missing_test_cols=[]
for col in df_test.columns:
```

```
    if df_test[col].isna().sum() !=0:
         missing_test_cols.append(col)
print(missing_test_cols)
```

```
['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']
```

[23]:
```
# Missing cols are all numerical variables
for col in df_train.columns:
    if col in (missing_train_cols):
        df_train[col].replace(np.nan, df_train[col].mean(),inplace=True)
```

[24]:
```
# Missing cols are all numerical variables
for col in df_test.columns:
    if col in (missing_test_cols):
        df_test[col].replace(np.nan, df_test[col].mean(),inplace=True)
```

[25]:
```
df_train.isna().sum().sum()
```

[25]: 0

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

[26]: 0

[27]:
```
from pandasql import sqldf
q1 = "select place,pct_own,second_mortgage,lat,lng from df_train where pct_own␣
 ↪>0.10 and second_mortgage <0.5 order by second_mortgage DESC LIMIT 2500;"
pysqldf = lambda q: sqldf(q, globals())
df_train_location_mort_pct=pysqldf(q1)
```

[28]:
```
df_train_location_mort_pct.head()
```

[28]:

	place	pct_own	second_mortgage	lat	lng
0	Worcester City	0.20247	0.43363	42.254262	-71.800347

```
1     Harbor Hills  0.15618          0.31818  40.751809 -73.853582
2      Glen Burnie  0.22380          0.30212  39.127273 -76.635265
3   Egypt Lake-leto  0.11618          0.28972  28.029063 -82.495395
4      Lincolnwood  0.14228          0.28899  41.967289 -87.652434
```

[29]:
```
import plotly.express as px
import plotly.graph_objects as go
```

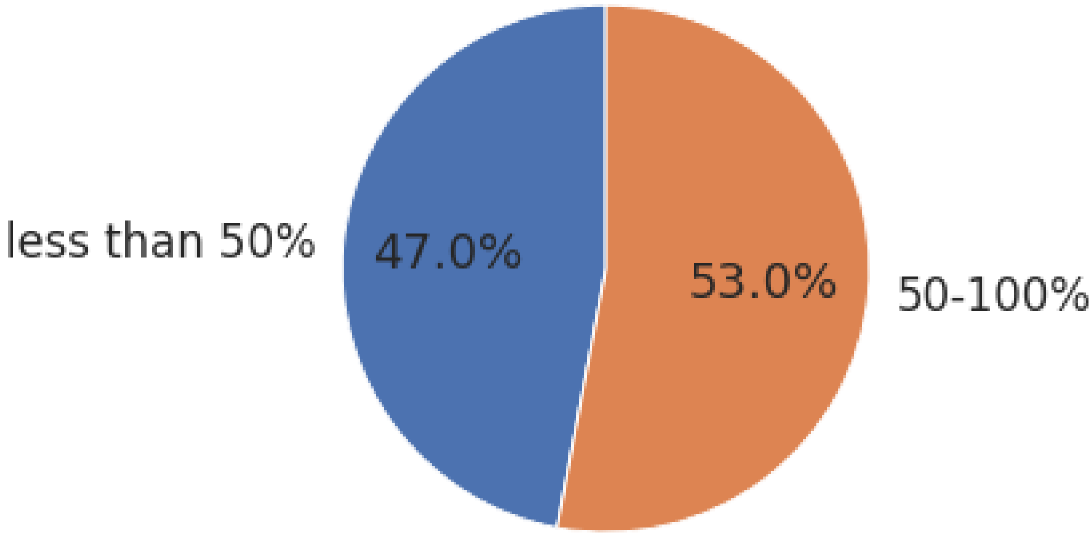
[30]:
```
fig = go.Figure(data=go.Scattergeo(
    lat = df_train_location_mort_pct['lat'],
    lon = df_train_location_mort_pct['lng']),
    )
fig.update_layout(
    geo=dict(
        scope = 'north america',
        showland = True,
        landcolor = "rgb(212, 212, 212)",
        subunitcolor = "rgb(255, 255, 255)",
        countrycolor = "rgb(255, 255, 255)",
        showlakes = True,
        lakecolor = "rgb(255, 255, 255)",
        showsubunits = True,
        showcountries = True,
        resolution = 50,
        projection = dict(
            type = 'conic conformal',
            rotation_lon = -100
        ),
        lonaxis = dict(
            showgrid = True,
            gridwidth = 0.5,
            range= [ -140.0, -55.0 ],
            dtick = 5
        ),
        lataxis = dict (
            showgrid = True,
            gridwidth = 0.5,
            range= [ 20.0, 60.0 ],
            dtick = 5
        )
    ),
    title='Top 2,500 locations with second mortgage is the highest and percent␣
 ↪ownership is above 10 percent')
fig.show()
```

Top 2,500 locations with second mortgage is the highest and percent ownership is above 10 perce

[31]:
```
###Use the following bad debt equation: Bad Debt = P (Second Mortgage   Home␣
↪Equity Loan) Bad Debt = second_mortgage + home_equity -␣
↪home_equity_second_mortgage c) Create pie charts to show overall debt and␣
↪bad debt
```

[32]:
```
df_train['bad_debt']=df_train['second_mortgage']+df_train['home_equity']-df_train['home_equity
```

[33]:
```
df_train['bins'] = pd.cut(df_train['bad_debt'],bins=[0,0.10,1], labels=["less␣
↪than 50%","50-100%"])
df_train.groupby(['bins']).size().plot(kind='pie',subplots=True,startangle=90,␣
↪autopct='%1.1f%%')
plt.axis('equal')

plt.show()
#df.plot.pie(subplots=True,figsize=(8, 3))
```



[34]:
```
####Create Box and whisker plot and analyze the distribution for 2nd mortgage,␣
↪home equity, good debt, and bad debt for different cities
```

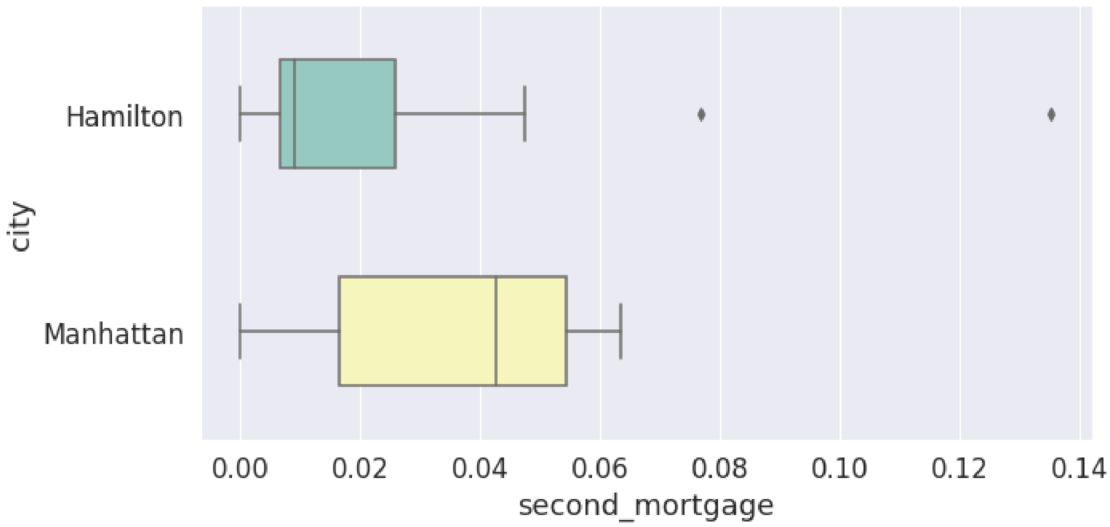
[35]:
```
cols=[]
df_train.columns
```

[35]:
```
Index(['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',
       'bad_debt', 'bins'],
      dtype='object')
```

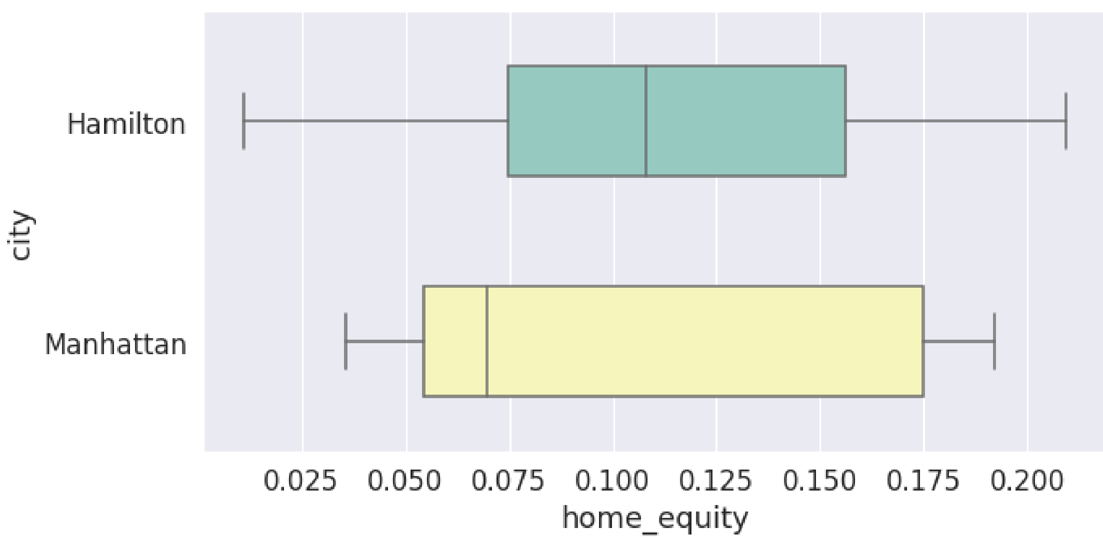
[36]:
```
#Taking Hamilton and Manhattan cities data
cols=['second_mortgage','home_equity','debt','bad_debt']
df_box_hamilton=df_train.loc[df_train['city'] == 'Hamilton']
df_box_manhattan=df_train.loc[df_train['city'] == 'Manhattan']
df_box_city=pd.concat([df_box_hamilton,df_box_manhattan])
df_box_city.head(4)
```

[36]:
```
        COUNTYID  STATEID        state state_ab      city          place  \
UID
267822        53       36     New York       NY  Hamilton       Hamilton
263797        21       34   New Jersey       NJ  Hamilton      Yardville
270979        17       39         Ohio       OH  Hamilton  Hamilton City
259028        95       28  Mississippi       MS  Hamilton       Hamilton

           type primary  zip_code  area_code  …  female_age_stdev  \
UID                                           …
267822     City   tract     13346        315  …          22.51276
263797     City   tract      8610        609  …          24.05831
270979  Village   tract     45015        513  …          22.66500
259028      CDP   tract     39746        662  …          22.79602

        female_age_sample_weight  female_age_samples  pct_own  married  \
UID
267822                 685.33845              2618.0  0.79046  0.57851
263797                 732.58443              3124.0  0.64400  0.56377
270979                 565.32725              2528.0  0.61278  0.47397
259028                 483.01311              1954.0  0.83241  0.58678

        married_snp  separated  divorced  bad_debt            bins
UID
267822      0.01882    0.01240   0.08770   0.09408  less than 50%
263797      0.01980    0.00990   0.04892   0.18071        50-100%
270979      0.04419    0.02663   0.13741   0.15005        50-100%
259028      0.01052    0.00000   0.11721   0.02130  less than 50%

[4 rows x 79 columns]
```

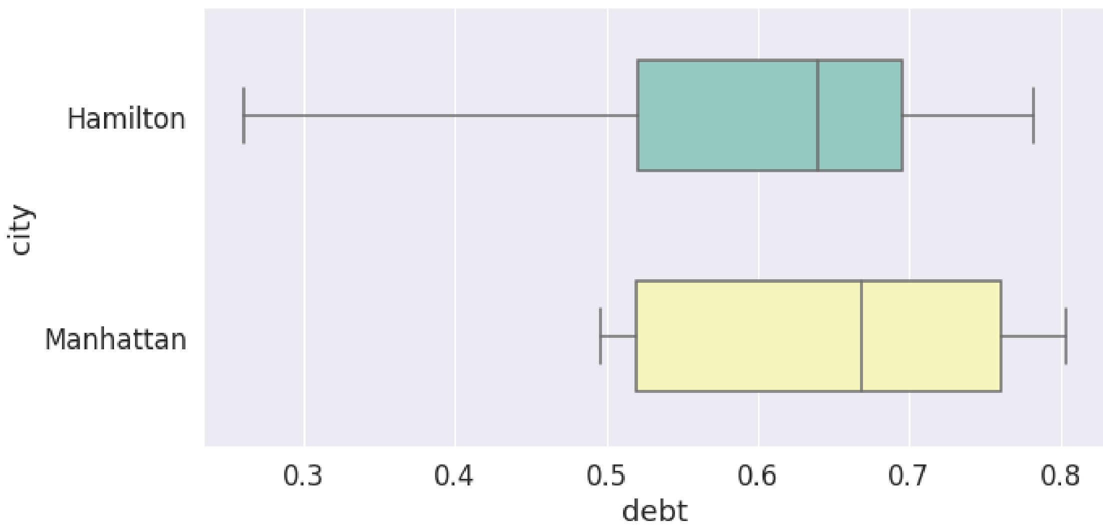
[37]:
```
plt.figure(figsize=(10,5))
sns.boxplot(data=df_box_city,x='second_mortgage', y='city',width=0.
 ↪5,palette="Set3")
plt.show()
```



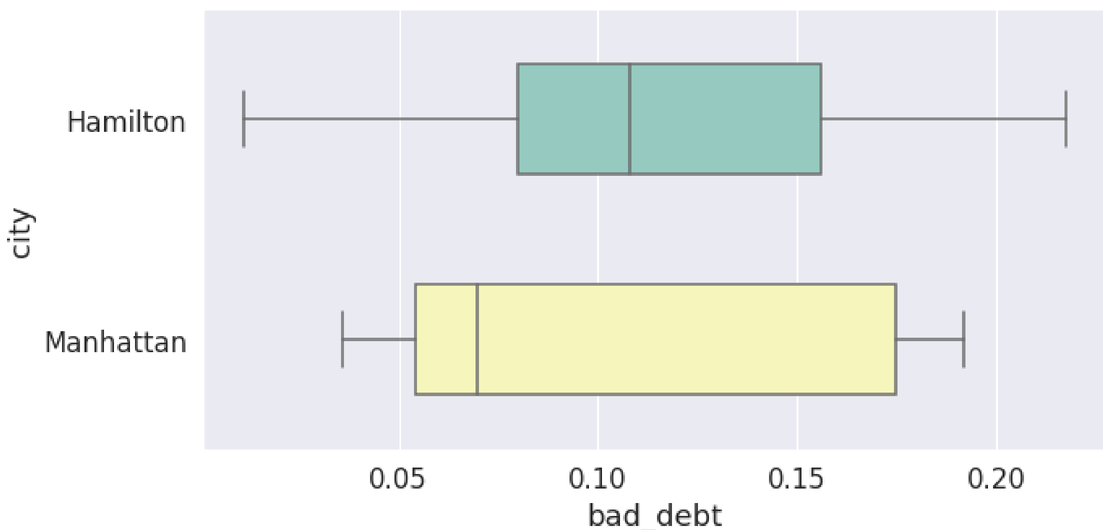
[38]:
```
plt.figure(figsize=(10,5))
sns.boxplot(data=df_box_city,x='home_equity', y='city',width=0.5,palette="Set3")
plt.show()
```



[39]:
```
plt.figure(figsize=(10,5))
sns.boxplot(data=df_box_city,x='debt', y='city',width=0.5,palette="Set3")
plt.show()
```



[40]:
```
plt.figure(figsize=(10,5))
sns.boxplot(data=df_box_city,x='bad_debt', y='city',width=0.5,palette="Set3")
plt.show()
```

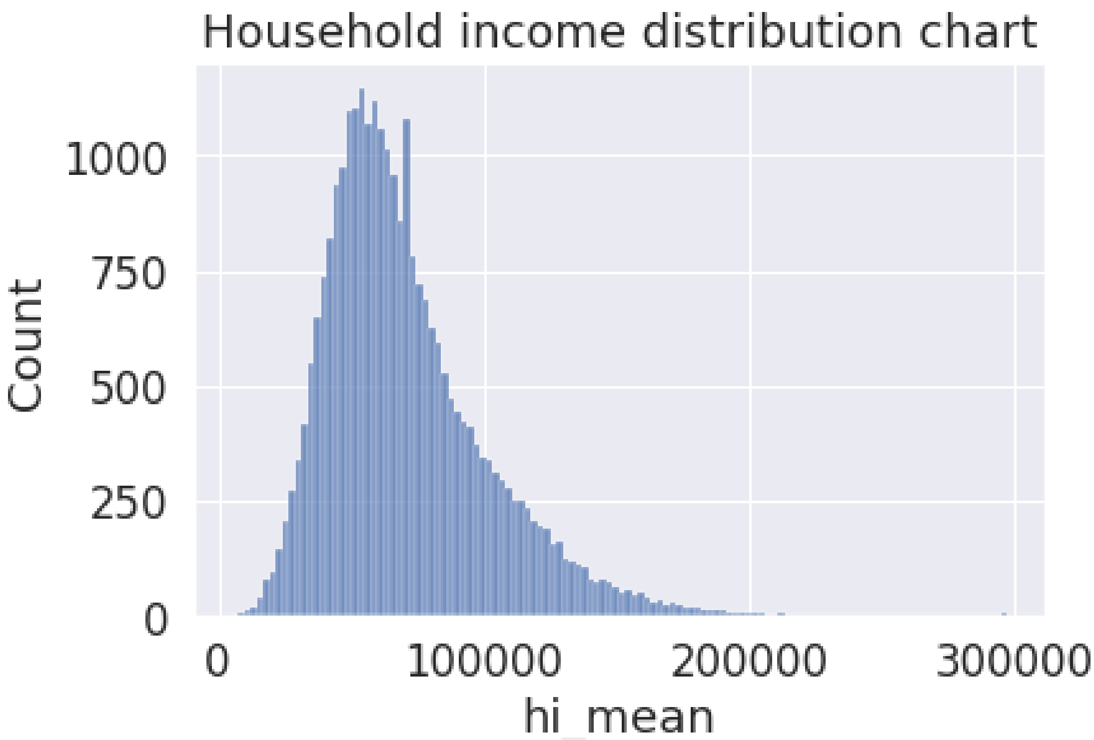


[41]:
```
#Manhattan has higher metrics compared to Hamilton
```

[42]:
```
#Now create a collated income distribution chart for family income, house hold␣
↪income, and remaining income
```

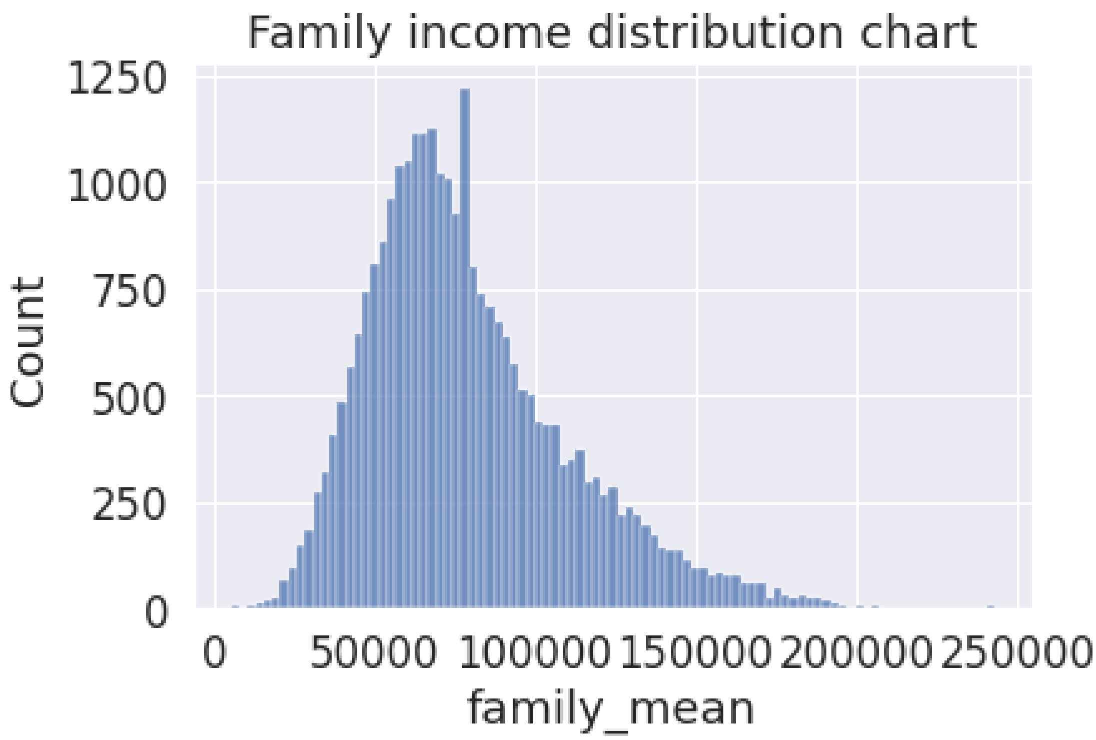
[43]:

```python
sns.histplot(df_train['hi_mean'])
plt.title('Household income distribution chart')
plt.show()
```

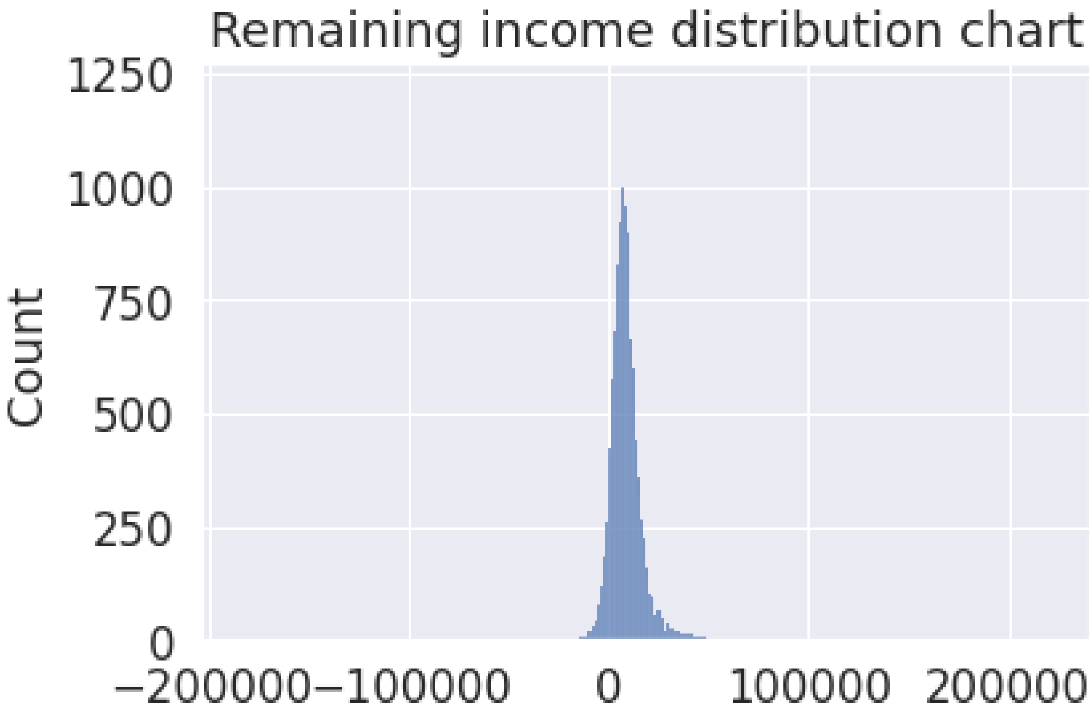


[44]:

```python
sns.histplot(df_train['family_mean'])
plt.title('Family income distribution chart')
plt.show()
```

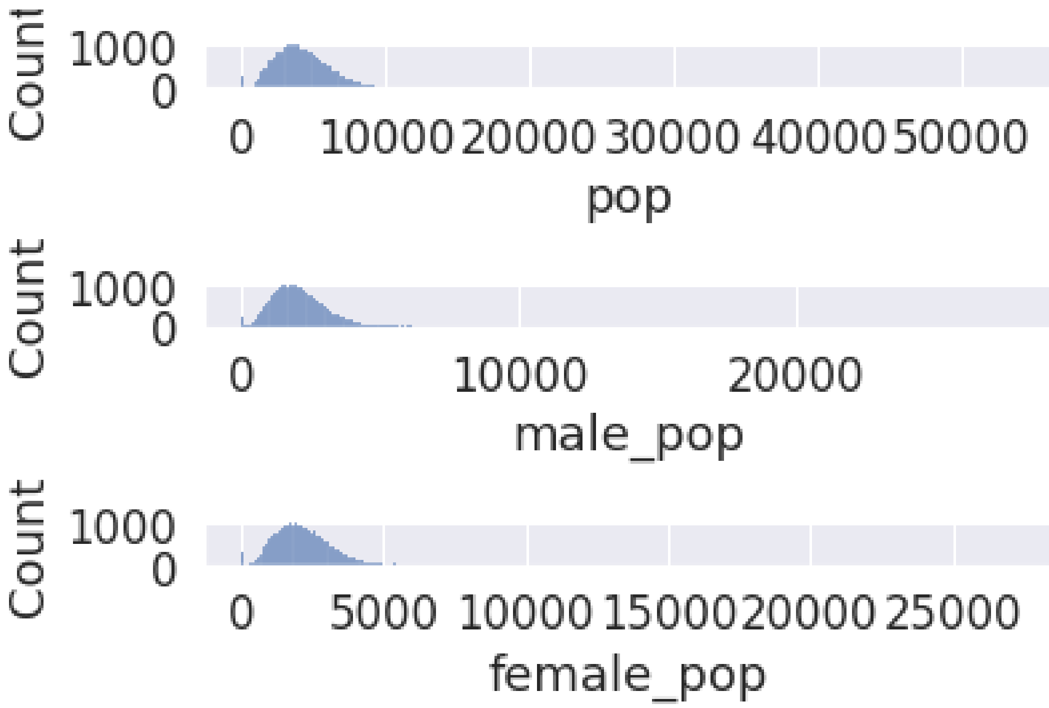


[45]:
```
sns.histplot(df_train['family_mean']-df_train['hi_mean'])
plt.title('Remaining income distribution chart')
plt.show()
```

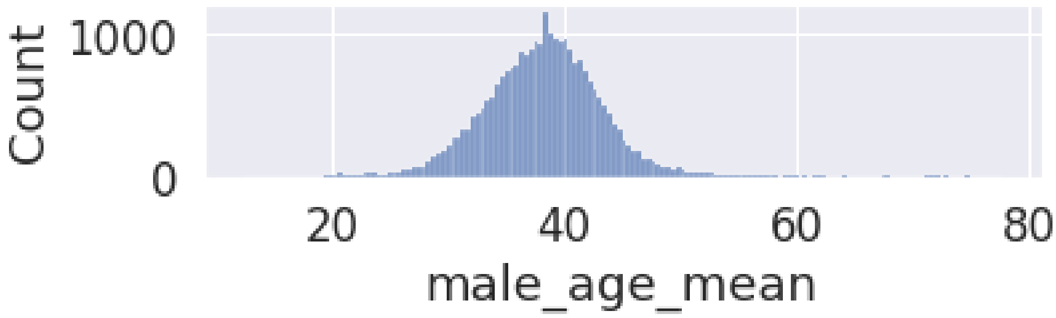


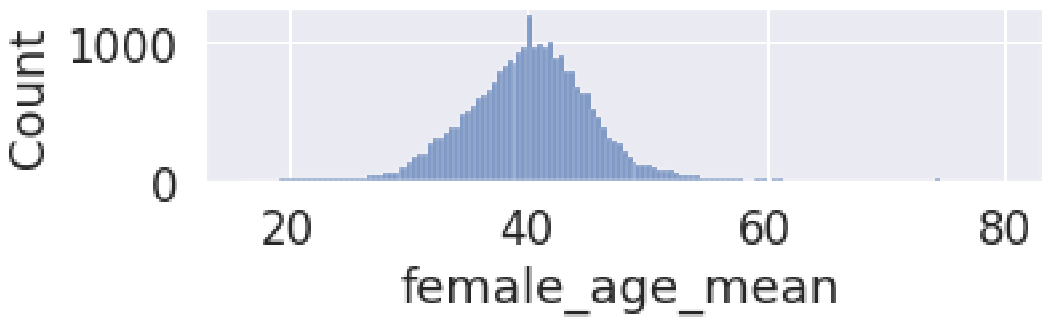
[46]:
```
##Perform EDA and come out with insights into population density and age. You
↪may have to derive new fields (make sure to weight averages for accurate
↪measurements):
```

[47]:
```
#plt.figure(figsize=(25,10))
fig,(ax1,ax2,ax3)=plt.subplots(3,1)
sns.histplot(df_train['pop'],ax=ax1)
sns.histplot(df_train['male_pop'],ax=ax2)
sns.histplot(df_train['female_pop'],ax=ax3)
plt.subplots_adjust(wspace=0.8,hspace=0.8)
plt.tight_layout()
plt.show()
```



[48]:
```
#plt.figure(figsize=(25,10))
fig,(ax1,ax2)=plt.subplots(2,1)
sns.histplot(df_train['male_age_mean'],ax=ax1)
sns.histplot(df_train['female_age_mean'],ax=ax2)
plt.subplots_adjust(wspace=0.8,hspace=0.8)
plt.tight_layout()
plt.show()
```



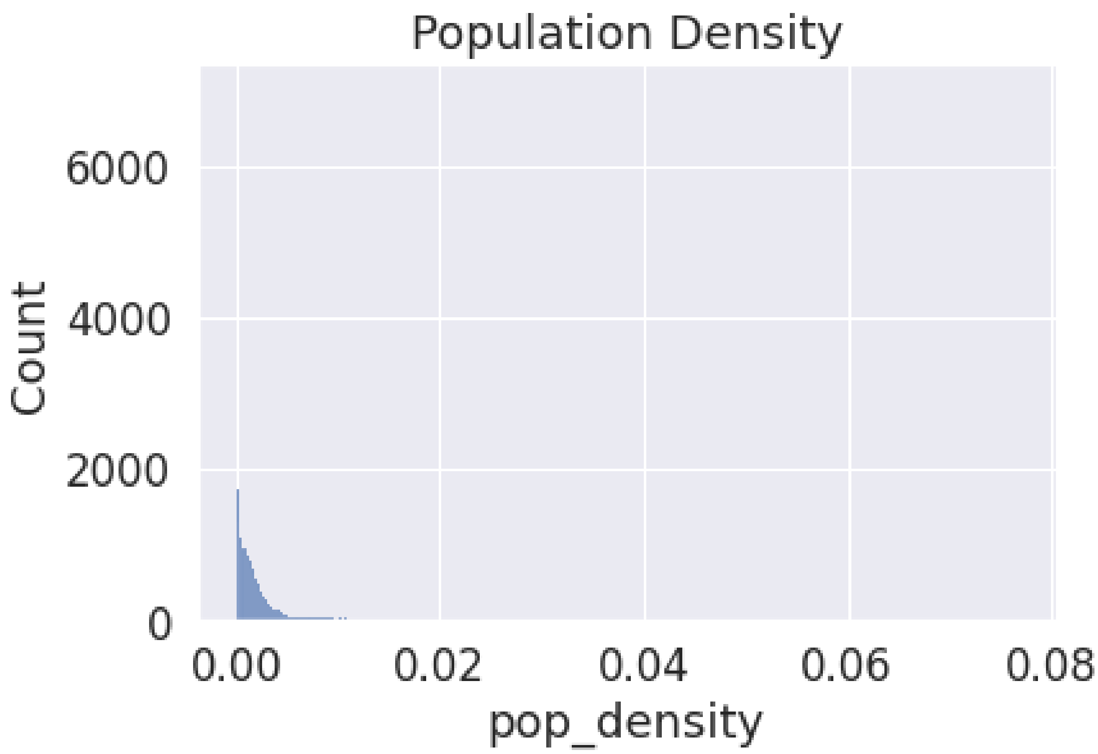


[49]:
```
##Population density:
```

[50]:
```
df_train['pop_density']=df_train['pop']/df_train['ALand']
```

[51]:
```
df_test['pop_density']=df_test['pop']/df_test['ALand']
```

[52]:
```
sns.histplot(df_train['pop_density'])
plt.title('Population Density')
plt.show() # Very less density is noticed
```



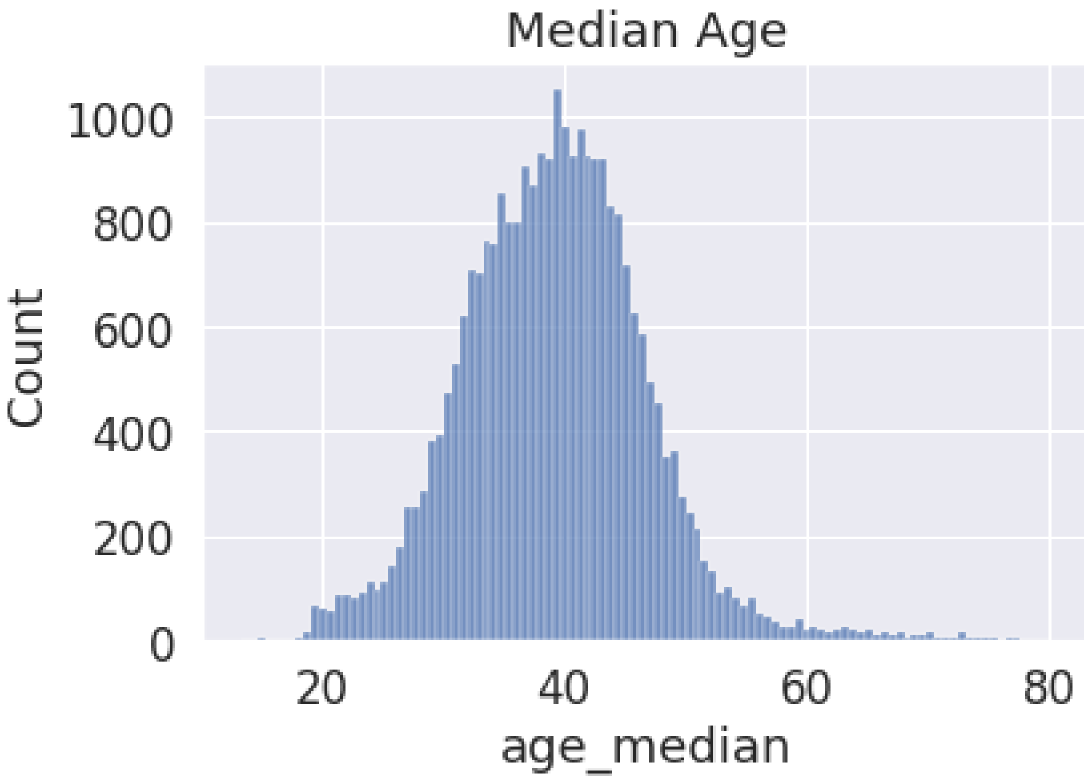
[53]:
```
df_train['age_median']=(df_train['male_age_median']+df_train['female_age_median'])/
  ↪2
df_test['age_median']=(df_test['male_age_median']+df_test['female_age_median'])/
  ↪2
```

[54]:
```
df_train[['male_age_median','female_age_median','male_pop','female_pop','age_median']].
  ↪head()
```

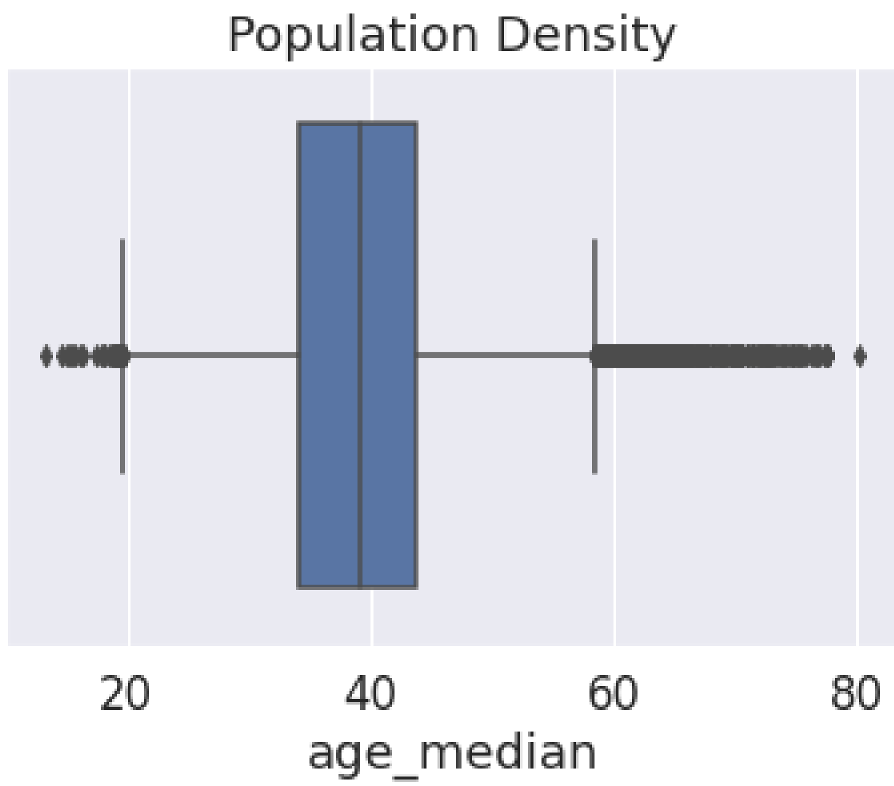
[54]:

	male_age_median	female_age_median	male_pop	female_pop	age_median
UID					
267822	44.00000	45.33333	2612	2618	44.666665
246444	32.00000	37.58333	1349	1284	34.791665
245683	40.83333	42.83333	3643	3238	41.833330
279653	48.91667	50.58333	1141	1559	49.750000
247218	22.41667	21.58333	2586	3051	22.000000

[55]:
```
sns.histplot(df_train['age_median'])
plt.title('Median Age')
plt.show()
# Age of population is mostly between 20 and 60
# Majority are of age around 40
# Median age distribution has a gaussian distribution
# Some right skewness is noticed
```



[56]:
```
sns.boxplot(x=df_train["age_median"])
plt.title('Population Density')
plt.show()
```



[57]:
```
#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.
```

[58]:
```
df_train['pop'].describe()
```

[58]:
```
count    27321.000000
mean      4316.032685
std       2169.226173
min          0.000000
25%       2885.000000
50%       4042.000000
75%       5430.000000
max      53812.000000
Name: pop, dtype: float64
```

[59]:
```
df_train['pop_bins']=pd.cut(df_train['pop'],bins=5,labels=['very␣
↪low','low','medium','high','very high'])
```

[60]:
```
df_train[['pop','pop_bins']]
```

[60]:

```
           pop  pop_bins
UID
267822    5230  very low
246444    2633  very low
245683    6881  very low
279653    2700  very low
247218    5637  very low
…          …        …
279212    1847  very low
277856    4155  very low
233000    2829  very low
287425   11542       low
265371    3726  very low

[27321 rows x 2 columns]
```

[61]:

```python
df_train['pop_bins'].value_counts()
```

[61]:

```
very low     27058
low            246
medium           9
high             7
very high        1
Name: pop_bins, dtype: int64
```

[62]:

```python
#Analyze the married, separated, and divorced population for these population␣
↪brackets
```

[63]:

```python
df_train.groupby(by='pop_bins')[['married','separated','divorced']].count()
```

[63]:

pop_bins	married	separated	divorced
very low	27058	27058	27058
low	246	246	246
medium	9	9	9
high	7	7	7
very high	1	1	1

[64]:

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

[64]:

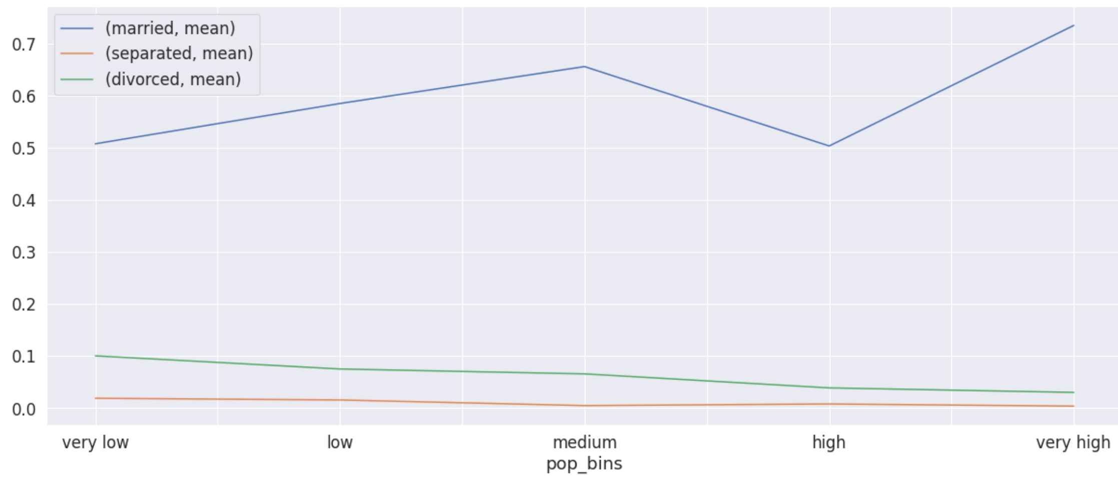
	married		separated		divorced	
	mean	median	mean	median	mean	median
pop_bins						
very low	0.507548	0.524680	0.019126	0.013650	0.100504	0.096020
low	0.584894	0.593135	0.015833	0.011195	0.075348	0.070045
medium	0.655737	0.618710	0.005003	0.004120	0.065927	0.064890

```
high       0.503359  0.335660  0.008141  0.002500  0.039030  0.010320
very high  0.734740  0.734740  0.004050  0.004050  0.030360  0.030360
```

[65]:
```
#Very high population group has more married people and less percantage of␣
↪separated and divorced couples
#In very low population groups, there are more divorced people
```

[66]:
```
plt.figure(figsize=(10,5))
pop_bin_married=df_train.
↪groupby(by='pop_bins')[['married','separated','divorced']].agg(["mean"])
pop_bin_married.plot(figsize=(20,8))
plt.legend(loc='best')
plt.show()
```

```
<Figure size 720x360 with 0 Axes>
```



[67]:
```
rent_state_mean=df_train.groupby(by='state')['rent_mean'].agg(["mean"])
rent_state_mean.head()
```

[67]:
```
                   mean
state
Alabama      774.004927
Alaska      1185.763570
Arizona     1097.753511
Arkansas     720.918575
California  1471.133857
```

[68]:
```
income_state_mean=df_train.groupby(by='state')['family_mean'].agg(["mean"])
income_state_mean.head()
```

[68]:
```
                   mean
state
Alabama     67030.064213
Alaska      92136.545109
Arizona     73328.238798
Arkansas    64765.377850
California  87655.470820
```

[69]:
```
rent_perc_of_income=rent_state_mean['mean']/income_state_mean['mean']
rent_perc_of_income.head(10)
```

[69]:
```
state
Alabama                 0.011547
Alaska                  0.012870
Arizona                 0.014970
Arkansas                0.011131
California              0.016783
Colorado                0.013529
Connecticut             0.012637
Delaware                0.012929
District of Columbia    0.013198
Florida                 0.015772
Name: mean, dtype: float64
```

[70]:
```
#overall level rent as a percentage of income
sum(df_train['rent_mean'])/sum(df_train['family_mean'])
```

[70]:
```
0.013358170721473864
```

[71]:
```
#Creating a heatmap - First perform a correlation analysis for all the relevant␣
↪variables by creating a heatmap. Describe your findings.
```

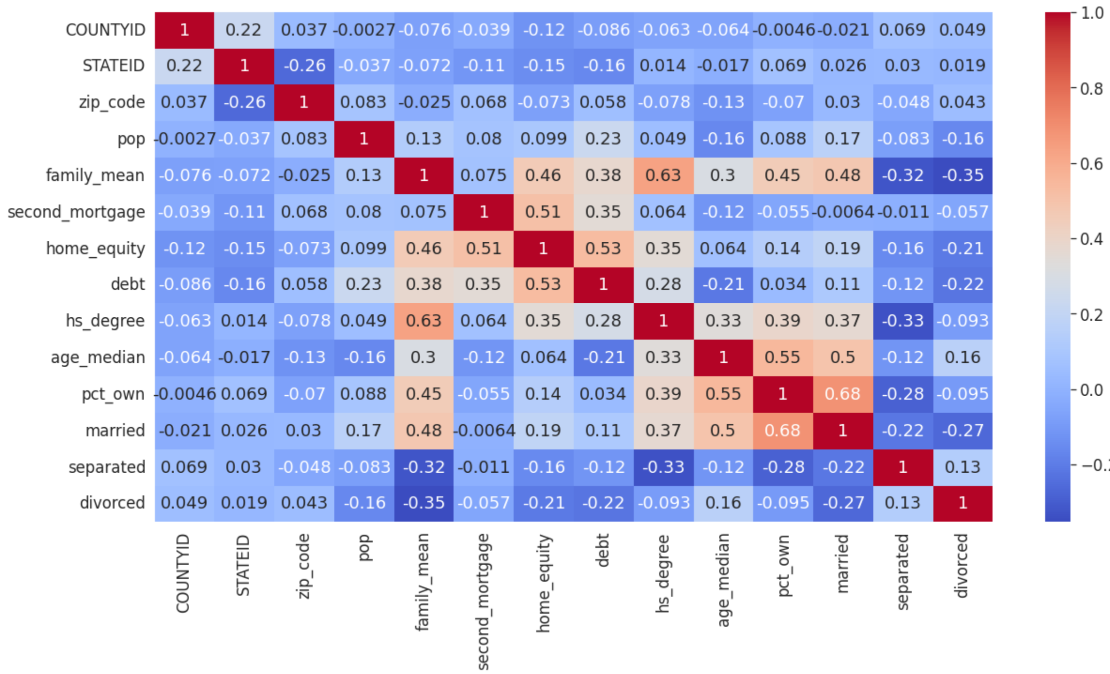
[72]:
```
df_train.columns
```

[72]:
```
Index(['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',
       'bad_debt', 'bins', 'pop_density', 'age_median', 'pop_bins'],
      dtype='object')
```

[73]:
```
cor=df_train[['COUNTYID','STATEID','zip_code','type','pop', 'family_mean',
         'second_mortgage', 'home_equity', 'debt','hs_degree',
           'age_median','pct_own', 'married','separated', 'divorced']].
 ↪corr(numeric_only = True)
```

[74]:
```
plt.figure(figsize=(20,10))
sns.heatmap(cor,annot=True,cmap='coolwarm')
plt.show()
```



[77]:
```
#pip install factor_analyzer
```

[78]:
```
from sklearn.decomposition import FactorAnalysis
from factor_analyzer import FactorAnalyzer
```

[79]:
```
fa=FactorAnalyzer(n_factors=5)
fa.fit_transform(df_train.select_dtypes(exclude= ('object','category')))
fa.loadings_
```

```
[79]: array([[-1.12589166e-01,  1.95646469e-02, -2.39331075e-02,
        -6.27632607e-02,  4.23474744e-02],
       [-1.10186763e-01,  1.33506224e-02,  2.79651241e-02,
        -1.49825866e-01,  1.10838810e-01],
       [-8.28678662e-02,  5.16372375e-02, -1.36451870e-01,
        -4.98918639e-02, -1.04024844e-01],
       [ 1.80961151e-02,  1.92013754e-02,  5.81329842e-03,
         2.64842737e-02, -6.12442420e-03],
       [ 9.02324716e-02, -9.72544295e-02, -6.54601286e-02,
        -1.33145898e-01, -1.48594600e-01],
       [-1.07335682e-02, -4.12376815e-02,  1.45853486e-01,
         8.80433435e-03,  1.08227569e-01],
       [-4.28796970e-02, -2.09780212e-02,  3.66726851e-02,
        -9.45597353e-02,  5.91380508e-02],
       [-2.44243026e-03, -1.53245407e-02, -2.68300860e-03,
        -4.52473026e-02,  2.37240649e-02],
       [ 7.92164350e-02,  9.57453322e-01, -8.71151645e-02,
        -6.59923899e-03, -3.97273181e-02],
       [ 7.39808218e-02,  9.18750507e-01, -1.08834839e-01,
        -2.79371584e-02, -3.93153641e-02],
       [ 8.06598905e-02,  9.47839212e-01, -6.08006518e-02,
         1.53627083e-02, -3.86977274e-02],
       [ 7.70052117e-01,  9.84675441e-03, -3.71249741e-02,
         1.14949036e-01, -1.23784685e-01],
       [ 7.18615871e-01,  6.24980548e-03, -4.59787395e-02,
         1.09109683e-01, -1.35301910e-01],
       [ 7.07647240e-01,  2.46625401e-02, -1.00860874e-02,
         1.04472485e-01,  7.72381211e-02],
       [-1.34545491e-01,  3.36809298e-01, -4.87894967e-01,
        -4.15446210e-02,  3.17608526e-01],
       [ 2.31079698e-01,  4.37729788e-01, -6.40209206e-01,
        -2.52310980e-02,  3.47216209e-01],
       [-4.52068092e-02,  3.51263837e-02,  3.07536934e-02,
         4.44793474e-01, -1.63273397e-01],
       [-2.50717060e-02,  1.70166790e-02,  4.57227195e-02,
         6.76083876e-01, -1.55256759e-01],
       [-3.90694449e-02, -1.67460878e-02,  8.13962758e-02,
         8.36389127e-01, -9.18259830e-02],
       [-5.14161951e-02, -3.57207138e-02,  1.10795173e-01,
         9.25123747e-01, -4.44866482e-02],
       [-6.08590000e-02, -4.41860607e-02,  1.35794005e-01,
         9.53019868e-01, -2.21548605e-02],
       [-4.57771151e-02, -5.25526122e-02,  1.41019878e-01,
         9.32702643e-01, -5.84027206e-07],
       [-4.19486043e-02, -5.90387634e-02,  1.28851779e-01,
         8.87316675e-01,  1.05894318e-02],
       [-2.47894628e-02, -7.29670553e-02,  9.41510479e-02,
```

```
  7.79023677e-01,  2.95352832e-02],
[ 2.12258452e-01,  4.65992342e-01, -6.14495952e-01,
 -2.47660037e-02,  3.66644520e-01],
[ 2.33057236e-01,  4.47057842e-01, -6.28263420e-01,
 -2.71547703e-02,  3.43419598e-01],
[ 7.85157094e-01,  4.91249263e-02,  1.44540486e-01,
 -2.05217628e-01, -1.54523361e-01],
[ 7.10324877e-01,  4.99730445e-02,  1.32239993e-01,
 -2.19171859e-01, -2.10505570e-01],
[ 8.61780947e-01,  4.35044840e-02,  1.65839097e-01,
 -1.19850814e-01,  3.16733594e-02],
[-2.23443274e-01,  8.46259538e-01, -4.61177255e-02,
  6.88599225e-02,  2.27742304e-01],
[ 1.43837554e-01,  9.53197410e-01,  2.27887411e-02,
 -4.57890460e-02,  1.00796441e-01],
[ 8.30286484e-01,  3.42026007e-02,  1.61106001e-01,
 -2.04570324e-01, -7.48710500e-02],
[ 7.94476569e-01,  2.83818600e-02,  1.51219548e-01,
 -2.07681490e-01, -9.12497124e-02],
[ 8.11481664e-01,  4.32314902e-02,  1.43645560e-01,
 -1.07778262e-01,  5.79540220e-02],
[-3.37741907e-01,  8.64927628e-01,  3.58933697e-02,
  9.07183967e-02,  4.46327274e-02],
[ 5.03572688e-02,  9.35515338e-01,  1.51475398e-01,
 -2.51501269e-02, -9.34471576e-02],
[ 9.78242255e-01, -3.31490286e-02, -1.05261178e-01,
  4.50364252e-02,  7.37362101e-02],
[ 9.59137190e-01, -3.90023001e-02, -1.20630340e-01,
  4.52591414e-02,  6.64877214e-02],
[ 8.14087175e-01,  2.23057317e-03,  7.66518494e-02,
  2.02747429e-02,  1.27634820e-01],
[-4.15353990e-01,  7.18339589e-01,  3.40068072e-01,
 -7.18402754e-02, -2.77950515e-01],
[ 7.64912621e-02,  7.24900636e-01,  2.74193212e-01,
 -4.83952617e-02, -3.52988287e-01],
[ 9.10390859e-01, -5.36541212e-02, -4.68641894e-02,
 -7.64182699e-04,  1.63870457e-01],
[ 8.73011839e-01, -5.30302287e-02, -5.89943105e-02,
 -1.58989853e-03,  1.52417522e-01],
[ 7.55087654e-01, -3.56133769e-03,  5.39542538e-02,
  4.24181353e-03,  2.58043466e-01],
[-1.23469881e-01,  6.07438115e-01,  6.33039198e-01,
 -2.14798871e-02,  2.47973912e-01],
[-3.42866884e-01,  5.59526272e-01,  5.88212992e-01,
 -2.51533551e-02,  2.18419886e-01],
[-1.60867217e-01, -1.53062630e-02, -1.57026579e-01,
  1.09243755e-01, -6.61660827e-01],
```

```
[-1.37306750e-01, -2.17250636e-02, -1.58408924e-01,
  1.25156188e-01, -6.71630775e-01],
[ 2.45096181e-01, -2.54584613e-02, -2.66691395e-02,
  9.53148494e-02, -6.42510844e-01],
[ 2.03988660e-01,  7.85172835e-02, -3.01656215e-01,
  2.28379467e-02, -6.29223346e-01],
[ 1.08926095e-01, -6.34332373e-02, -3.36565286e-02,
 -9.49480507e-02,  6.81473859e-01],
[-2.63787620e-01, -6.43280865e-03, -3.58792218e-02,
 -9.37962473e-02,  6.47817012e-01],
[-2.15717040e-01, -7.36588939e-02,  3.50113235e-01,
 -1.95201620e-02,  6.36783786e-01],
[ 3.94306143e-01,  6.09565680e-02,  2.55337864e-01,
 -2.20362097e-01, -1.84248082e-01],
[ 4.07877888e-01,  6.27256518e-02,  2.23926912e-01,
 -2.10028739e-01, -1.71989229e-01],
[ 3.53156875e-01,  5.36715658e-02,  2.69603573e-01,
 -2.16933223e-01, -1.80072075e-01],
[ 2.33537258e-01, -4.91732971e-02,  8.14450775e-01,
  9.36688860e-02,  3.27131928e-01],
[ 2.40298209e-01, -3.38140101e-02,  8.31496980e-01,
  7.52417601e-02,  2.46323614e-01],
[-6.71839517e-02,  6.58504562e-02,  5.86207698e-01,
  8.72955278e-02,  9.12541403e-02],
[ 5.59835567e-02,  8.17918701e-01, -1.78458350e-01,
 -1.55949432e-02, -3.34299735e-02],
[ 7.16426423e-02,  9.23428548e-01, -1.07142697e-01,
 -2.78635377e-02, -4.35991111e-02],
[ 1.92496941e-01, -4.75870408e-02,  8.03173181e-01,
  1.43492705e-01,  3.33862149e-01],
[ 1.87644432e-01, -3.29941013e-02,  8.58024497e-01,
  1.31329957e-01,  2.55679730e-01],
[-1.02263657e-01,  6.03984262e-02,  4.72982255e-01,
  7.36848388e-02,  1.12273910e-01],
[ 6.14776651e-02,  8.77962734e-01, -1.50410285e-01,
  2.20991027e-02, -4.17158184e-02],
[ 7.83728232e-02,  9.54508789e-01, -5.91095917e-02,
  1.64800929e-02, -4.32590992e-02],
[-3.24381953e-02,  1.11167166e-01,  7.84467415e-01,
 -4.37718523e-02, -2.80931231e-01],
[ 1.76682387e-01,  1.90494241e-01,  5.61405500e-01,
 -1.20746164e-01, -1.32570786e-01],
[-6.37386572e-02, -7.03047927e-02, -2.68934074e-01,
  1.28589793e-01,  1.88507866e-01],
[-1.56051271e-01, -7.08033939e-02, -1.45964503e-01,
  1.24253733e-01,  1.46293114e-01],
[-3.56716296e-01, -5.29910752e-02,  1.47771603e-01,
```

```
          2.87196182e-02,  1.13159574e-01],
        [ 2.42173829e-01, -2.86199123e-02, -3.25958329e-02,
          1.05027813e-01, -6.55406061e-01],
        [ 3.50196764e-01, -1.05016404e-02, -3.95274133e-01,
          5.92876760e-02,  2.91651803e-01],
        [ 2.25671536e-01, -3.42672776e-02,  8.92876593e-01,
          1.12426798e-01,  2.67065190e-01]])
```

[80]:
```
##Data Modeling : Linear Regression
#Build a linear Regression model to predict the total monthly expenditure for␣
↪home mortgages loan.
```

[81]:
```
df_train.columns
```

[81]:
```
Index(['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',
       'bad_debt', 'bins', 'pop_density', 'age_median', 'pop_bins'],
      dtype='object')
```

[82]:
```
df_train['type'].unique()
type_dict={'type':{'City':1,
                   'Urban':2,
                   'Town':3,
                   'CDP':4,
                   'Village':5,
                   'Borough':6}
          }
df_train.replace(type_dict,inplace=True)
```

[83]:
```
df_train['type'].unique()
```

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

[84]:
```
df_test.replace(type_dict,inplace=True)
```

[85]:
```
df_test['type'].unique()
```

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

[86]:
```
feature_cols=['COUNTYID','STATEID','zip_code','type','pop', 'family_mean',
         'second_mortgage', 'home_equity', 'debt','hs_degree',
           'age_median','pct_own', 'married','separated', 'divorced']
```

[87]:
```
x_train=df_train[feature_cols]
y_train=df_train['hc_mortgage_mean']
```

[88]:
```
x_test=df_test[feature_cols]
y_test=df_test['hc_mortgage_mean']
```

[89]:
```
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
```

[90]:
```
x_train.head()
```

[90]:
```
        COUNTYID  STATEID  zip_code  type   pop  family_mean  second_mortgage  \
UID
267822        53       36     13346     1  5230  67994.14790          0.02077
246444       141       18     46616     1  2633  50670.10337          0.02222
245683        63       18     46122     1  6881  95262.51431          0.00000
279653       127       72       927     2  2700  56401.68133          0.01086
247218       161       20     66502     1  5637  54053.42396          0.05426

        home_equity     debt  hs_degree  age_median  pct_own  married  \
UID
267822      0.08919  0.52963    0.89288   44.666665  0.79046  0.57851
246444      0.04274  0.60855    0.90487   34.791665  0.52483  0.34886
245683      0.09512  0.73484    0.94288   41.833330  0.85331  0.64745
279653      0.01086  0.52714    0.91500   49.750000  0.65037  0.47257
247218      0.05426  0.51938    1.00000   22.000000  0.13046  0.12356

        separated  divorced
UID
267822    0.01240   0.08770
246444    0.01426   0.09030
245683    0.01607   0.10657
279653    0.02021   0.10106
```

```
247218    0.00000    0.03109
```

[91]:
```
sc=StandardScaler()
x_train_scaled=sc.fit_transform(x_train)
x_test_scaled=sc.fit_transform(x_test)
```

[92]:
```
#Run a model at a Nation level. If the accuracy levels and R square are not␣
↪satisfactory proceed to below step.
```

[94]:
```
linereg=LinearRegression()
linereg.fit(x_train_scaled,y_train)
```

[94]:
```
LinearRegression()
```

[95]:
```
y_pred=linereg.predict(x_test_scaled)
```

[96]:
```
print("Overall R2 score of linear regression model", r2_score(y_test,y_pred))
print("Overall RMSE of linear regression model", np.
↪sqrt(mean_squared_error(y_test,y_pred)))
```

```
Overall R2 score of linear regression model 0.7348210754610929
Overall RMSE of linear regression model 323.1018894984635
```

[97]:
```
##The Accuracy and R2 score are good, but still will investigate the model␣
↪performance at state level
```

[98]:
```
state=df_train['STATEID'].unique()
state[0:5]
#Picking a few iDs 20,1,45,6
```

[98]:
```
array([36, 18, 72, 20,  1])
```

[99]:
```
for i in [20,1,45]:
    print("State ID-",i)

    x_train_nation=df_train[df_train['COUNTYID']==i][feature_cols]
    y_train_nation=df_train[df_train['COUNTYID']==i]['hc_mortgage_mean']

    x_test_nation=df_test[df_test['COUNTYID']==i][feature_cols]
    y_test_nation=df_test[df_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)

    linereg.fit(x_train_scaled_nation,y_train_nation)
    y_pred_nation=linereg.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- 20
Overall R2 score of linear regression model for state, 20 :- 0.6046603766461809
Overall RMSE of linear regression model for state, 20 :- 307.9718899931476


State ID- 1
Overall R2 score of linear regression model for state, 1 :- 0.8104382475484615
Overall RMSE of linear regression model for state, 1 :- 307.8275861848436


State ID- 45
Overall R2 score of linear regression model for state, 45 :- 0.7887446497855252
Overall RMSE of linear regression model for state, 45 :- 225.69615420724134
```

[100]:
```
# To check the residuals
```

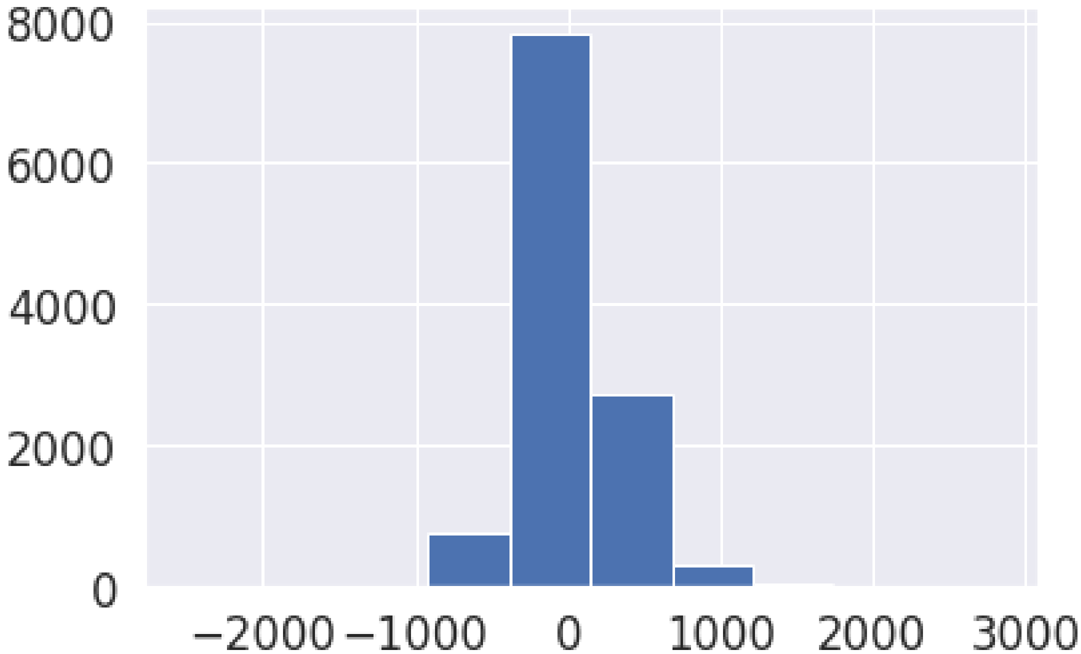
[101]:
```
residuals=y_test-y_pred
residuals
```

[101]:
```
UID
255504    281.969088
252676    -69.935775
276314    190.761969
248614   -157.290627
286865     -9.887017
             …
238088    -67.541646
242811    -41.578757
250127   -127.427569
241096   -330.820475
287763    217.760642
Name: hc_mortgage_mean, Length: 11709, dtype: float64
```

[102]:
```
plt.hist(residuals) # Normal distribution of residuals
```

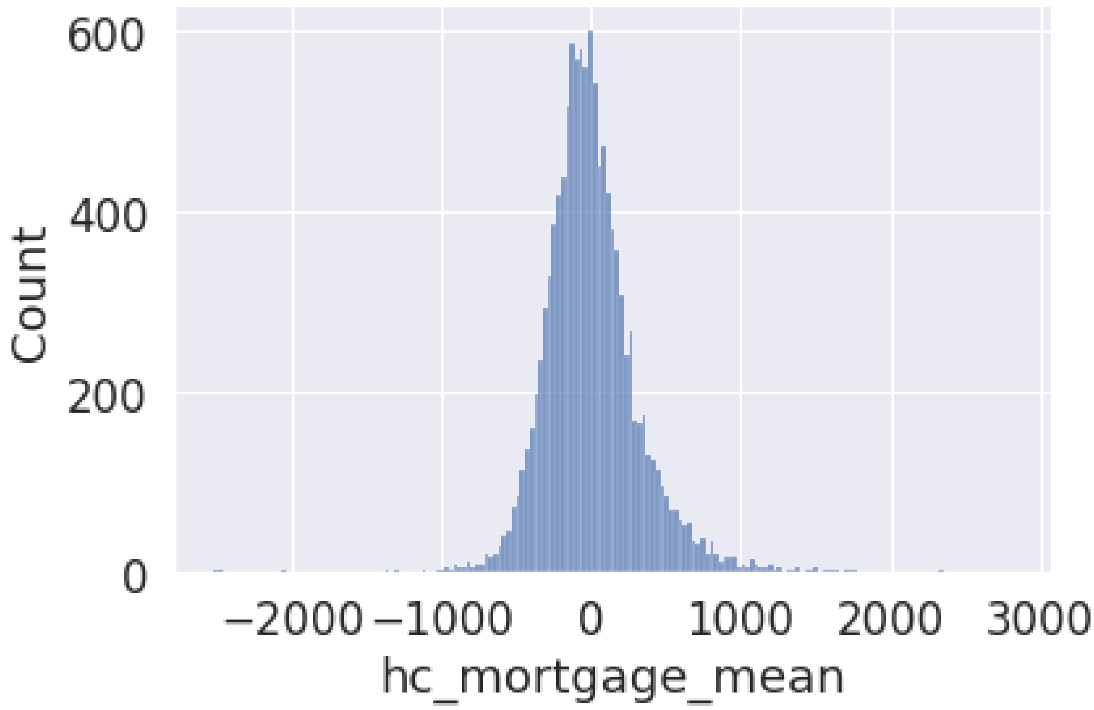
[102]:
```
(array([6.000e+00, 3.000e+00, 2.900e+01, 7.670e+02, 7.823e+03, 2.716e+03,
        3.010e+02, 4.900e+01, 1.200e+01, 3.000e+00]),
 array([-2515.04284233, -1982.92661329, -1450.81038425,  -918.69415521,
         -386.57792617,   145.53830287,   677.65453191,  1209.77076095,
```

```
        1741.88698999,  2274.00321903,  2806.11944807]),
 <BarContainer object of 10 artists>)
```



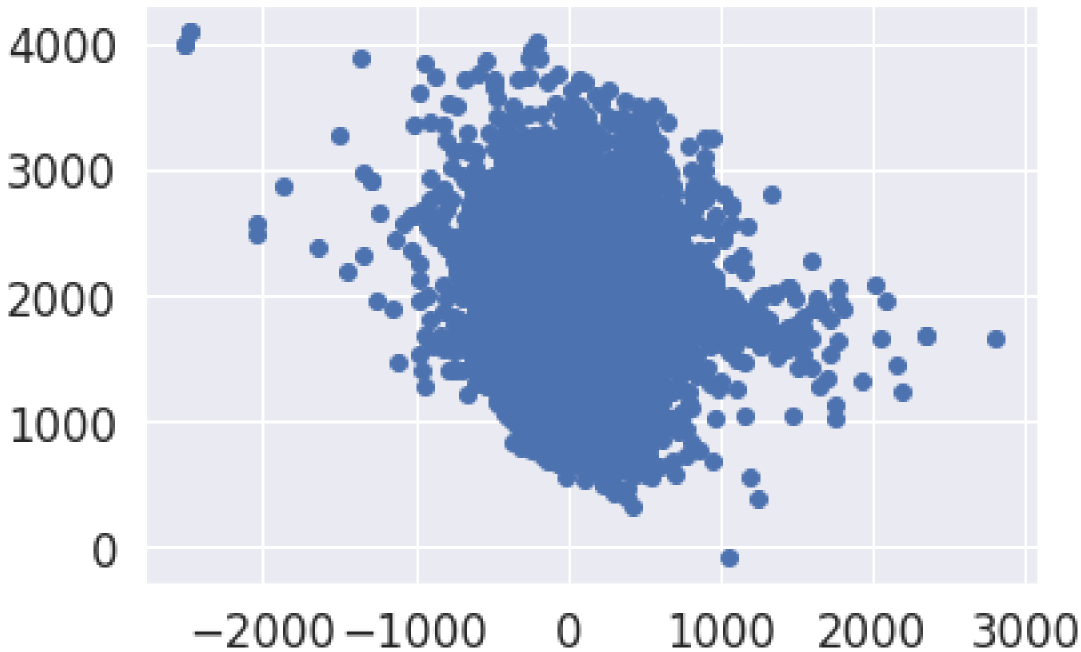
[103]:
```
sns.histplot(residuals)
```

[103]:
```
<AxesSubplot: xlabel='hc_mortgage_mean', ylabel='Count'>
```



[104]:
```
plt.scatter(residuals,y_pred)
# Same variance and residuals does not have correlation with predictor
# Independance of residuals
```

[104]: <matplotlib.collections.PathCollection at 0x7f3d5f9eb1c0>


4000
3000
2000
1000
0
−2000
−1000
0
1000
2000
3000

```
[ ]:
```