real estate

June 16, 2023

[1]:
```
# Importing Libraries
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
# Importing module
import warnings
# Warnings filter.
warnings.filterwarnings('ignore')
# Import the necessary libraries
import plotly.offline as pyo
import plotly.graph_objs as go
# Set notebook mode to work in offline
pyo.init_notebook_mode()
```

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

0.0.1 Descriptive Analysis

[3]:
```
train.head()
```

[3]:

	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]
```

[4]:
```
test.head()
```

[4]:
```
      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]
```

[5]:
```
train.describe()
```

[5]:

	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	67.074018	-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      0.027487      0.129000
max        1.000000      0.714290      0.714290      1.000000

[8 rows x 74 columns]
```

[6]:
```
test.describe()
```

[6]:
```
                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     38.618092    -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]
```

[7]:
```
train.columns
```

[7]:
```
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')
```

[8]:
```
test.columns
```

[8]:
```
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')
```

[9]:
```
# UID is unique userID value in the train and test dataset. So an index can be␣
↪created from the UID feature
train.set_index(keys=['UID'],inplace=True)#Set the DataFrame index using␣
↪existing columns.
test.set_index(keys=['UID'],inplace=True)
```

[10]:
```
# Handling Missing value
train.isnull().sum()/len(train)*100
```

[10]:
```
BLOCKID        100.000000
SUMLEVEL         0.000000
COUNTYID         0.000000
STATEID          0.000000
state            0.000000
                   …
pct_own          0.980930
married          0.699096
married_snp      0.699096
separated        0.699096
divorced         0.699096
Length: 79, dtype: float64
```

[11]:
```
train=train.drop(['BLOCKID','SUMLEVEL'],axis=1)
```

[12]:
```
test.isnull().sum()/len(test)*100
```

[12]:
```
BLOCKID        100.000000
SUMLEVEL         0.000000
COUNTYID         0.000000
STATEID          0.000000
state            0.000000
                   …
pct_own          1.041934
married          0.717397
```

```
married_snp      0.717397
separated        0.717397
divorced         0.717397
Length: 79, dtype: float64
```

[13]:
```
test=test.drop(['BLOCKID','SUMLEVEL'],axis=1)
```

[14]:
```
# Imputing  missing values with mean
missing_train_cols=[]
for col in train.columns:
    if 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']
```

[15]:
```
missing_test_cols=[]
for col in test.columns:
    if 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']
```

[16]:
```
# Missing cols are all numerical variables
for col in train.columns:
    if col in (missing_train_cols):
        train[col].replace(np.nan,train[col].mean(),inplace=True)
```

[17]:
```
for col in test.columns:
    if col in (missing_test_cols):
        test[col].replace(np.nan,test[col].mean(),inplace=True)
```

[18]:
```
train.isna().sum().sum()
```

[18]: 0

[19]:
```
test.isna().sum().sum()
```

[19]: 0

0.0.2 Week 1 Exploratory Data Analysis

[20]:
```
df = train[train['pct_own']>0.1]
df.shape
```

[20]: (26565, 77)

[21]:
```
df = df.sort_values(by='second_mortgage',ascending=False)
```

[22]:
```
pd.set_option('display.max_columns', None)
df.head()
```

[22]:

UID	COUNTYID	STATEID	state	state_ab	city	place	type	primary	zip_code	area_code	lat
289712	147	51	Virginia	VA	Farmville	Farmville	Town	tract	23901	434	37.297357
251185	27	25	Massachusetts	MA	Worcester	Worcester City	City	tract	1610	508	42.254262
269323	81	36	New York	NY	Corona	Harbor Hills	City	tract	11368	718	40.751809
251324	3	24	Maryland	MD	Glen Burnie	Glen Burnie	CDP	tract	21061	410	39.127273

```
235788  Egypt Lake-leto  City   tract     33614        813  28.029063

              lng      ALand  AWater   pop  male_pop  female_pop   rent_mean  \
UID
289712 -78.396452   413391.0       0  1733       609        1124   782.00000
251185 -71.800347   797165.0       0  2133      1139         994   942.32740
269323 -73.853582   169666.0       0  4181      2249        1932  1413.12357
251324 -76.635265  1110282.0       0  4866      1985        2881  1335.49818
235788 -82.495395  2050906.0  234794  5468      2784        2684   914.10322

        rent_median  rent_stdev  rent_sample_weight  rent_samples  rent_gt_10  \
UID
289712        781.0    22.95830            11.00000          11.0     1.00000
251185        953.0   304.34109           333.88019         645.0     0.98906
269323       1388.0   499.47343           205.65925         777.0     1.00000
251324       1335.0   336.92824           352.62444        1502.0     1.00000
235788        880.0   191.64962          1067.77502        1847.0     0.96619

        rent_gt_15  rent_gt_20  rent_gt_25  rent_gt_30  rent_gt_35  \
UID
289712     1.00000     1.00000     1.00000     1.00000     1.00000
251185     0.97813     0.86250     0.81563     0.68438     0.53281
269323     0.92664     0.80952     0.69241     0.58301     0.44659
251324     0.91545     0.77763     0.60186     0.49001     0.42011
235788     0.92794     0.80820     0.58925     0.44235     0.37140

        rent_gt_40  rent_gt_50  universe_samples  used_samples      hi_mean  \
UID
289712     1.00000     0.00000                11            11  33088.92156
251185     0.47500     0.39063               655           640  39036.18368
269323     0.37967     0.30245               821           777  56434.63436
251324     0.38016     0.23435              1502          1502  59466.62302
235788     0.22783     0.17350              1965          1804  48495.17313

        hi_median     hi_stdev  hi_sample_weight  hi_samples  family_mean  \
UID
289712    23236.0  19970.41249          16.33316        19.0  47067.92731
251185    29037.0  42317.65457         599.87224       768.0  50471.95789
269323    46106.0  47279.53535         674.74625       997.0  48558.91165
251324    50164.0  37351.26266        1293.31194      2068.0  64899.68626
235788    38340.0  41137.53473        1664.02791      2179.0  52332.06236

        family_median  family_stdev  family_sample_weight  family_samples  \
UID
289712        59954.0   24030.19608               5.33316             8.0
251185        40476.0   45794.28515             314.09134           432.0
269323        40462.0   35569.90113             630.41529           878.0
```

```
251324        50705.0   39727.56212             706.84520          1125.0
235788        39980.0   41386.75431             755.11681          1010.0

        hc_mortgage_mean  hc_mortgage_median  hc_mortgage_stdev  \
UID
289712        2249.50000              2249.0           182.57419
251185        1596.15811              1690.0           465.71234
269323        3037.81395              3320.0           888.70919
251324        1622.29005              1520.0           511.53797
235788        1641.00508              1462.0           774.11061

        hc_mortgage_sample_weight  hc_mortgage_samples    hc_mean  hc_median  \
UID
289712                    0.79359                  4.0  749.50000      749.0
251185                   30.05003                 96.0  589.73200      528.0
269323                   29.17150                138.0  751.81483      894.0
251324                  156.43774                496.0  452.77058      509.0
235788                   81.16409                169.0  446.96166      404.0

         hc_stdev  hc_samples  hc_sample_weight  home_equity_second_mortgage  \
UID
289712   36.51484         4.0           1.97980                      0.00000
251185  198.18324        17.0          10.43434                      0.43363
269323  269.48263        38.0          23.35354                      0.31818
251324  165.06276        70.0          49.29293                      0.27739
235788   86.60735        45.0          34.89899                      0.28972

        second_mortgage  home_equity     debt  second_mortgage_cdf  \
UID
289712          0.50000      0.00000  0.50000              0.00067
251185          0.43363      0.43363  0.84956              0.00100
269323          0.31818      0.40341  0.78409              0.00241
251324          0.30212      0.35689  0.87633              0.00289
235788          0.28972      0.38785  0.78972              0.00324

        home_equity_cdf  debt_cdf  hs_degree  hs_degree_male  \
UID
289712          1.00000   0.77776    1.00000         1.00000
251185          0.00468   0.08684    0.71803         0.68883
269323          0.00638   0.18540    0.58739         0.61499
251324          0.01131   0.05915    0.86185         0.85294
235788          0.00770   0.17505    0.92809         0.93188

        hs_degree_female  male_age_mean  male_age_median  male_age_stdev  \
UID
289712           1.00000       21.33803         19.25000         9.50021
251185           0.75828       30.99146         30.75000        18.15286
```

```
269323          0.55192       30.09851         29.58333        18.22005
251324          0.86732       29.07276         27.41667        19.97922
235788          0.92375       31.39914         29.08333        16.25854

        male_age_sample_weight  male_age_samples  female_age_mean  \
UID
289712               364.20985             609.0         19.58762
251185               255.90977            1139.0         30.60147
269323               483.12831            2249.0         29.80694
251324               475.95730            1985.0         32.53273
235788               613.84520            2784.0         34.53924

        female_age_median  female_age_stdev  female_age_sample_weight  \
UID
289712           19.16667           4.00258                 673.39577
251185           26.16667          19.21553                 262.09529
269323           27.66667          18.45616                 448.69061
251324           30.66667          19.61959                 694.10357
235788           28.58333          18.56943                 814.45000

        female_age_samples  pct_own  married  married_snp  separated  divorced
UID
289712              1124.0  0.62069  0.03612      0.01806    0.01806   0.00000
251185               994.0  0.20247  0.37844      0.11976    0.09341   0.10539
269323              1932.0  0.15618  0.44490      0.14555    0.02357   0.04066
251324              2881.0  0.22380  0.58250      0.08321    0.00000   0.01778
235788              2684.0  0.11618  0.36953      0.12876    0.09957   0.07339
```

[23]:
```
top_2500_second_mortgage_pctown_10 = df.head(2500)
top_2500_second_mortgage_pctown_10
```

[23]:
```
        COUNTYID  STATEID          state state_ab           city  \
UID
289712       147       51       Virginia       VA       Farmville
251185        27       25  Massachusetts       MA       Worcester
269323        81       36       New York       NY          Corona
251324         3       24       Maryland       MD     Glen Burnie
235788        57       12        Florida       FL           Tampa
…            …        …              …        …              …
229021        67        6     California       CA      Carmichael
261444       183       37  North Carolina       NC         Raleigh
225977        37        6     California       CA  Marina Del Rey
251433         5       24       Maryland       MD       Baltimore
230480        77        6     California       CA         Manteca

                place     type primary  zip_code  area_code        lat  \
UID
```

```
289712        Farmville     Town   tract     23901        434  37.297357
251185   Worcester City     City   tract      1610        508  42.254262
269323     Harbor Hills     City   tract     11368        718  40.751809
251324      Glen Burnie      CDP   tract     21061        410  39.127273
235788  Egypt Lake-leto     City   tract     33614        813  28.029063
…                  …      …      …         …          …      …
229021       Carmichael     City   tract     95608        916  38.617256
261444     Raleigh City  Village   tract     27606        919  35.757135
225977   Marina Del Rey     City   tract     90292        310  33.983203
251433         Lochearn      CDP   tract     21208        410  39.353095
230480     Manteca City     City   tract     95336        209  37.732143

               lng       ALand   AWater   pop  male_pop  female_pop  \
UID
289712  -78.396452    413391.0        0  1733       609        1124
251185  -71.800347    797165.0        0  2133      1139         994
269323  -73.853582    169666.0        0  4181      2249        1932
251324  -76.635265   1110282.0        0  4866      1985        2881
235788  -82.495395   2050906.0   234794  5468      2784        2684
…           …          …       …   …        …         …
229021 -121.337317   2453452.0        0  6388      3285        3103
261444  -78.704288   4014315.0   375580  6471      3506        2965
225977 -118.466139    902161.0   285884  3674      2084        1590
251433  -76.733315   1913598.0        0  2372      1049        1323
230480 -121.242902  99716769.0  1333851  6158      3063        3095

         rent_mean  rent_median  rent_stdev  rent_sample_weight  rent_samples  \
UID
289712   782.00000        781.0    22.95830            11.00000          11.0
251185   942.32740        953.0   304.34109           333.88019         645.0
269323  1413.12357       1388.0   499.47343           205.65925         777.0
251324  1335.49818       1335.0   336.92824           352.62444        1502.0
235788   914.10322        880.0   191.64962          1067.77502        1847.0
…            …          …          …                  …            …
229021   982.48589        874.0   345.27914           831.39667        1499.0
261444   987.07155        920.0   287.70421          1218.07615        2460.0
225977  2014.76772       2054.0   765.87674           211.00829        1308.0
251433  1902.92592       1864.0   371.03624            12.47686         112.0
230480  1301.87928       1260.0   607.57824           182.65634         412.0

        rent_gt_10  rent_gt_15  rent_gt_20  rent_gt_25  rent_gt_30  \
UID
289712     1.00000     1.00000     1.00000     1.00000     1.00000
251185     0.98906     0.97813     0.86250     0.81563     0.68438
269323     1.00000     0.92664     0.80952     0.69241     0.58301
251324     1.00000     0.91545     0.77763     0.60186     0.49001
235788     0.96619     0.92794     0.80820     0.58925     0.44235
```

```
…            …          …          …          …          …
229021     0.98722    0.88904    0.80699    0.73974    0.62340
261444     0.98175    0.91790    0.80886    0.60122    0.46177
225977     0.96177    0.84633    0.76376    0.57951    0.48471
251433     1.00000    1.00000    1.00000    0.90654    0.90654
230480     0.94802    0.85644    0.78218    0.58663    0.42574

        rent_gt_35  rent_gt_40  rent_gt_50  universe_samples  used_samples  \
UID
289712     1.00000     1.00000     0.00000                11            11
251185     0.53281     0.47500     0.39063               655           640
269323     0.44659     0.37967     0.30245               821           777
251324     0.42011     0.38016     0.23435              1502          1502
235788     0.37140     0.22783     0.17350              1965          1804
…            …          …          …                 …           …
229021     0.48218     0.44788     0.36113              1538          1487
261444     0.35317     0.31277     0.24631              2531          2302
225977     0.36162     0.31651     0.29281              1363          1308
251433     0.85047     0.80374     0.75701               148           107
230480     0.31931     0.24505     0.21040               497           404

            hi_mean  hi_median     hi_stdev  hi_sample_weight  hi_samples  \
UID
289712   33088.92156    23236.0  19970.41249          16.33316        19.0
251185   39036.18368    29037.0  42317.65457         599.87224       768.0
269323   56434.63436    46106.0  47279.53535         674.74625       997.0
251324   59466.62302    50164.0  37351.26266        1293.31194      2068.0
235788   48495.17313    38340.0  41137.53473        1664.02791      2179.0
…              …          …          …                 …           …
229021   60924.42356    45269.0  52642.08272        1578.38746      2535.0
261444   48917.03619    42560.0  42446.57656        1974.08399      2971.0
225977  122306.76210    87679.0  99344.45974         784.03896      2199.0
251433   81031.46025    70563.0  54301.37929         431.36409       867.0
230480   93433.24940    74648.0  69348.40202         769.38949      1715.0

         family_mean  family_median  family_stdev  family_sample_weight  \
UID
289712   47067.92731        59954.0   24030.19608               5.33316
251185   50471.95789        40476.0   45794.28515             314.09134
269323   48558.91165        40462.0   35569.90113             630.41529
251324   64899.68626        50705.0   39727.56212             706.84520
235788   52332.06236        39980.0   41386.75431             755.11681
…              …             …          …                  …
229021   70649.41206        59105.0   53672.88317             878.22961
261444   63058.72564        54776.0   39228.93828             488.87870
225977  153695.07560       168705.0   75332.51510             146.31854
251433   89756.95195        82142.0   51783.89571             249.78569
```

```
230480   93602.43101        76881.0   62539.58165             615.59620

        family_samples  hc_mortgage_mean  hc_mortgage_median  \
UID
289712             8.0        2249.50000              2249.0
251185           432.0        1596.15811              1690.0
269323           878.0        3037.81395              3320.0
251324          1125.0        1622.29005              1520.0
235788          1010.0        1641.00508              1462.0
…                  …                …                    …
229021          1592.0        1989.68075              1945.0
261444           817.0        1296.00864              1157.0
225977           637.0        3484.08913              4036.0
251433           580.0        1696.25710              1686.0
230480          1432.0        2503.93443              2342.0

        hc_mortgage_stdev  hc_mortgage_sample_weight  hc_mortgage_samples  \
UID
289712          182.57419                    0.79359                  4.0
251185          465.71234                   30.05003                 96.0
269323          888.70919                   29.17150                138.0
251324          511.53797                  156.43774                496.0
235788          774.11061                   81.16409                169.0
…                     …                          …                    …
229021          741.38518                  215.20179                740.0
261444          604.17462                  204.23737                397.0
225977         1222.36555                  100.85670                546.0
251433          647.48634                  183.61145                507.0
230480          880.19803                  198.44275                866.0

          hc_mean  hc_median   hc_stdev  hc_samples  hc_sample_weight  \
UID
289712  749.50000      749.0   36.51484         4.0           1.97980
251185  589.73200      528.0  198.18324        17.0          10.43434
269323  751.81483      894.0  269.48263        38.0          23.35354
251324  452.77058      509.0  165.06276        70.0          49.29293
235788  446.96166      404.0   86.60735        45.0          34.89899
…           …          …         …           …                …
229021  378.33570      357.0  149.76596       257.0         194.81444
261444  435.49901      397.0   82.96708        43.0          34.91919
225977  906.75439      782.0  486.87426       290.0         134.15205
251433  540.13764      554.0  232.21751       212.0         128.56768
230480  518.47460      466.0  282.26273       352.0         247.45455

        home_equity_second_mortgage  second_mortgage  home_equity     debt  \
UID
289712                      0.00000          0.50000      0.00000  0.50000
```

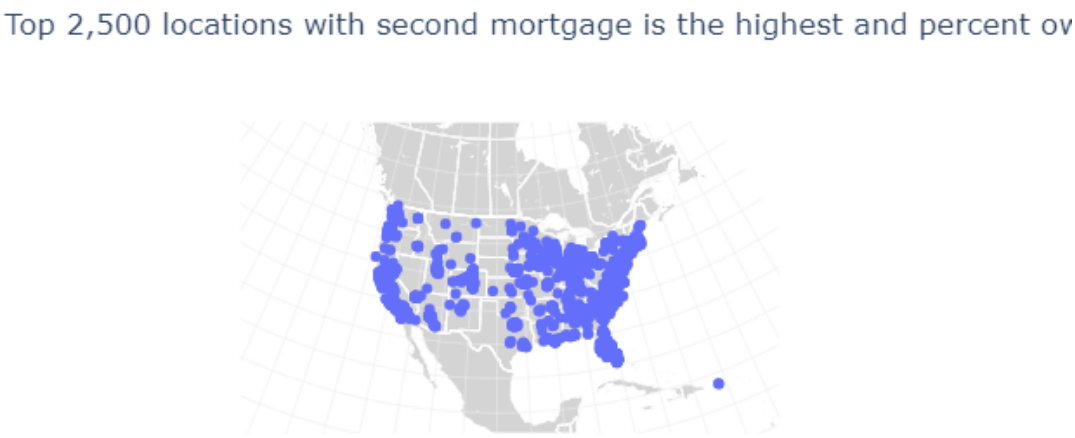
```
251185                    0.43363          0.43363      0.43363  0.84956
269323                    0.31818          0.31818      0.40341  0.78409
251324                    0.27739          0.30212      0.35689  0.87633
235788                    0.28972          0.28972      0.38785  0.78972
…                           …              …          …      …
229021                    0.06820          0.06820      0.27482  0.74223
261444                    0.06818          0.06818      0.09318  0.90227
225977                    0.05502          0.06818      0.18301  0.65311
251433                    0.06120          0.06815      0.13769  0.70515
230480                    0.06814          0.06814      0.16831  0.71100

        second_mortgage_cdf  home_equity_cdf  debt_cdf  hs_degree  \
UID
289712              0.00067          1.00000   0.77776    1.00000
251185              0.00100          0.00468   0.08684    0.71803
269323              0.00241          0.00638   0.18540    0.58739
251324              0.00289          0.01131   0.05915    0.86185
235788              0.00324          0.00770   0.17505    0.92809
…                     …              …       …        …
229021              0.12857          0.03592   0.27082    0.89663
261444              0.12858          0.47258   0.03936    0.96621
225977              0.12858          0.15169   0.48052    0.96709
251433              0.12884          0.28580   0.35585    0.90623
230480              0.12884          0.18870   0.34214    0.80677

        hs_degree_male  hs_degree_female  male_age_mean  male_age_median  \
UID
289712         1.00000           1.00000       21.33803         19.25000
251185         0.68883           0.75828       30.99146         30.75000
269323         0.61499           0.55192       30.09851         29.58333
251324         0.85294           0.86732       29.07276         27.41667
235788         0.93188           0.92375       31.39914         29.08333
…                …               …             …              …
229021         0.91061           0.88485       35.54012         32.16667
261444         0.97898           0.95033       28.13271         25.41667
225977         0.98074           0.95087       43.95378         41.91667
251433         0.91862           0.89746       45.28862         47.75000
230480         0.80107           0.81224       36.60759         37.75000

        male_age_stdev  male_age_sample_weight  male_age_samples  \
UID
289712         9.50021               364.20985             609.0
251185        18.15286               255.90977            1139.0
269323        18.22005               483.12831            2249.0
251324        19.97922               475.95730            1985.0
235788        16.25854               613.84520            2784.0
…                …                     …                …
```

```
229021        21.81258               835.48971            3285.0
261444        11.64197              1191.01348            3506.0
225977        19.00547               520.22767            2084.0
251433        24.81611               271.52252            1049.0
230480        22.99172               761.94470            3063.0

        female_age_mean  female_age_median  female_age_stdev  \
UID
289712         19.58762           19.16667           4.00258
251185         30.60147           26.16667          19.21553
269323         29.80694           27.66667          18.45616
251324         32.53273           30.66667          19.61959
235788         34.53924           28.58333          18.56943
…                 …                  …                 …
229021         42.55686           43.08333          21.33702
261444         29.19520           25.00000          13.44444
225977         44.81016           41.41667          18.58900
251433         50.74387           52.75000          24.90042
230480         37.42113           38.83333          22.82683

        female_age_sample_weight  female_age_samples  pct_own  married  \
UID
289712                 673.39577              1124.0  0.62069  0.03612
251185                 262.09529               994.0  0.20247  0.37844
269323                 448.69061              1932.0  0.15618  0.44490
251324                 694.10357              2881.0  0.22380  0.58250
235788                 814.45000              2684.0  0.11618  0.36953
…                          …                  …       …        …
229021                 805.77720              3103.0  0.41425  0.44711
261444                1044.70191              2965.0  0.12827  0.23974
225977                 343.62694              1590.0  0.44682  0.27404
251433                 301.08168              1323.0  0.84707  0.43002
230480                 744.85694              3095.0  0.67116  0.62787

        married_snp  separated  divorced
UID
289712      0.01806    0.01806   0.00000
251185      0.11976    0.09341   0.10539
269323      0.14555    0.02357   0.04066
251324      0.08321    0.00000   0.01778
235788      0.12876    0.09957   0.07339
…              …          …         …
229021      0.03273    0.00000   0.12335
261444      0.07685    0.00827   0.07165
225977      0.04473    0.02057   0.13162
251433      0.02822    0.00000   0.07223
230480      0.06491    0.01817   0.04890
```

```
[2500 rows x 77 columns]
```

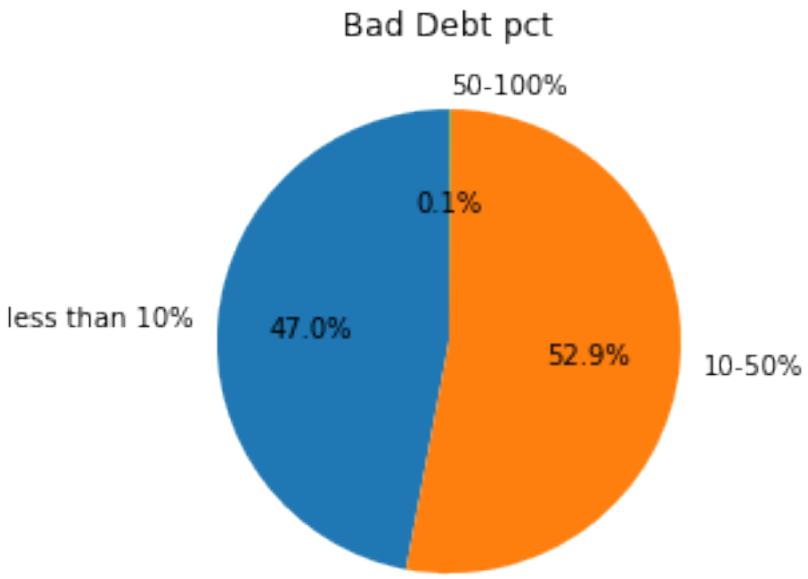
[24]:
```
import plotly.express as px
import plotly.graph_objects as go
```

[25]:
```
# Visualization 1 (Geo-Map):
fig = go.Figure(data=go.Scattergeo(
    lat = top_2500_second_mortgage_pctown_10['lat'],
    lon = top_2500_second_mortgage_pctown_10['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()
```

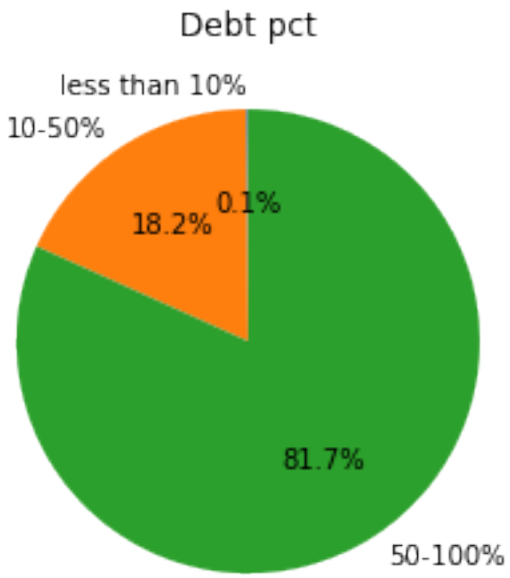


[26]:
```
train['bad_debt']=train['second_mortgage']+train['home_equity']-train['home_equity_second_mort
```

[27]:
```
# Visualization 2:
train['bins_bad_debt'] = pd.cut(train['bad_debt'],bins=[0,0.1,.5,1],␣
 ↪labels=["less than 10%","10-50%","50-100%"])
train.groupby(['bins_bad_debt']).size().
 ↪plot(kind='pie',subplots=True,startangle=90, autopct='%1.1f%%')
plt.title('Bad Debt pct')
plt.ylabel("")

plt.show()
```



[28]:
```
# Visualization 3:
train['bins_debt'] = pd.cut(train['debt'],bins=[0,0.1,.5,1], labels=["less than␣
↪10%","10-50%","50-100%"])
train.groupby(['bins_debt']).size().
↪plot(kind='pie',subplots=True,startangle=90, autopct='%1.1f%%')
plt.title('Debt pct')
plt.ylabel("")

plt.show()
```



[29]:
```
cols=['second_mortgage','home_equity','debt','bad_debt']
df_box_hamilton=train.loc[train['city'] == 'Hamilton']
df_box_manhattan=train.loc[train['city'] == 'Manhattan']
df_box_city=pd.concat([df_box_hamilton,df_box_manhattan])
df_box_city.head(4)
```

[29]:

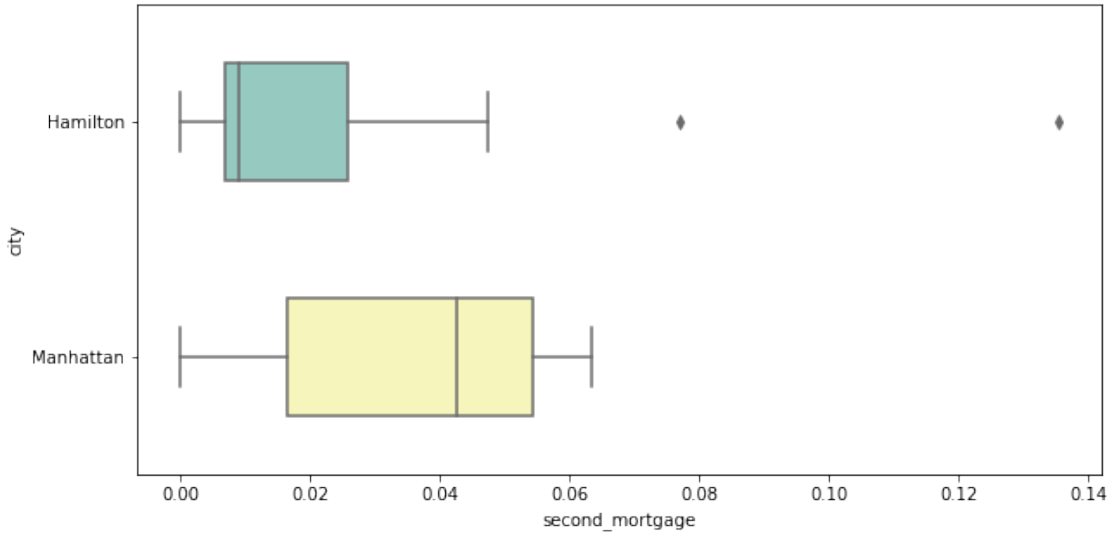
UID	COUNTYID	STATEID	state	state_ab	city	place	type	primary	zip_code	area_code	lat	lng	ALand	AWater	pop	male_pop	female_pop	rent_mean
267822	53	36	New York	NY	Hamilton	Hamilton	City	tract	13346	315	42.840812	-75.501524	202183361.0	1699120	5230	2612	2618	769.38638
263797	21	34	New Jersey	NJ	Hamilton	Yardville	City	tract	8610	609	40.206266	-74.675274	4623635.0	75545	5050	1926	3124	1299.55492
270979	17	39	Ohio	OH	Hamilton	Hamilton City	Village	tract	45015	513	39.364028	-84.570717	3598447.0	112290	4615	2087	2528	687.22347
259028	95	28	Mississippi	MS	Hamilton	Hamilton	CDP	tract	39746	662	33.759514	-88.377770	235934245.0	710507	3783	1829	1954	659.65320

```
        rent_median  rent_stdev  rent_sample_weight  rent_samples  rent_gt_10  \
UID
267822        784.0   232.63967           272.34441         362.0     0.86761
263797       1106.0   476.90596           273.50240         982.0     1.00000
270979        719.0   277.66094           534.12791         687.0     0.98399
259028        755.0   161.98765           137.40404         147.0     0.92174

        rent_gt_15  rent_gt_20  rent_gt_25  rent_gt_30  rent_gt_35  \
UID
267822     0.79155     0.59155     0.45634     0.42817     0.18592
263797     0.84746     0.68114     0.60169     0.36335     0.25530
270979     0.88501     0.77001     0.62591     0.51092     0.42504
259028     0.84348     0.69565     0.20870     0.20870     0.20870

        rent_gt_40  rent_gt_50  universe_samples  used_samples      hi_mean  \
UID
267822     0.15493     0.12958               387           355  63125.28406
263797     0.23835     0.17055              1000           944  80521.77955
270979     0.33770     0.22853               693           687  53074.46754
259028     0.20870     0.13913               221           115  56218.87091

        hi_median     hi_stdev  hi_sample_weight  hi_samples   family_mean  \
UID
267822    48120.0  49042.01206        1290.96240      2024.0   67994.14790
263797    61619.0  64319.32971        1476.17237      2721.0  117179.87740
270979    39319.0  48747.89548        1202.21734      1726.0   59345.58535
259028    47965.0  51357.62464         810.61414      1301.0   64115.06976

        family_median  family_stdev  family_sample_weight  family_samples  \
UID
267822        53245.0   47667.30119             884.33516          1491.0
263797       105448.0   62810.85492             353.10227          1144.0
270979        43927.0   48015.86057             744.58085          1088.0
259028        55171.0   48135.02541             585.00426          1019.0

        hc_mortgage_mean  hc_mortgage_median  hc_mortgage_stdev  \
UID
267822        1414.80295              1223.0          641.22898
263797        1865.82107              1672.0          874.41806
270979        1011.30380               995.0          246.23596
259028        1151.71231              1011.0          469.03313

        hc_mortgage_sample_weight  hc_mortgage_samples    hc_mean  hc_median  \
UID
267822                  377.83135                867.0  570.01530      558.0
263797                  431.82729               1283.0  774.11639      780.0
```

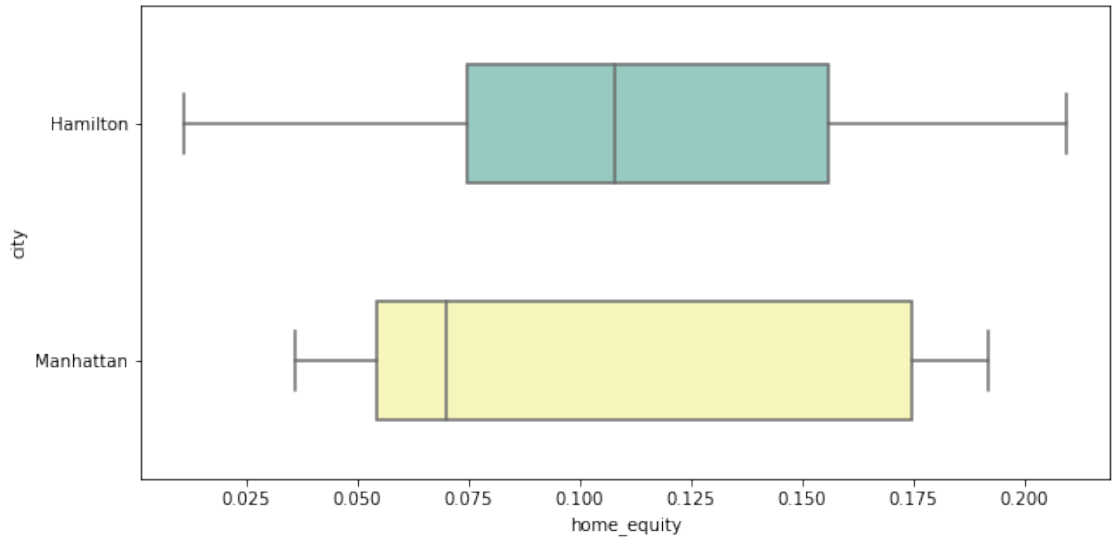
```
270979                514.75102               729.0  320.70619      318.0
259028                280.60828               409.0  386.92921      329.0

         hc_stdev  hc_samples  hc_sample_weight  home_equity_second_mortgage  \
UID
267822  270.11299       770.0         499.29293                      0.01588
263797  183.95710       438.0         224.36364                      0.01859
270979  101.22659       304.0         270.16162                      0.04743
259028  186.69331       671.0         536.61202                      0.00000

        second_mortgage  home_equity     debt  second_mortgage_cdf  \
UID
267822          0.02077      0.08919  0.52963              0.43658
263797          0.03021      0.16909  0.74550              0.34594
270979          0.04743      0.15005  0.70571              0.21952
259028          0.00000      0.02130  0.37870              1.00000

        home_equity_cdf  debt_cdf  hs_degree  hs_degree_male  \
UID
267822          0.49087   0.73341    0.89288         0.85880
263797          0.18660   0.26368    0.94187         0.96176
270979          0.24346   0.35449    0.83930         0.85614
259028          0.79686   0.90663    0.82507         0.79527

        hs_degree_female  male_age_mean  male_age_median  male_age_stdev  \
UID
267822           0.92434       42.48574         44.00000        22.97306
263797           0.93029       44.17886         43.75000        21.65283
270979           0.82438       35.35287         35.83333        19.97726
259028           0.85023       35.35016         33.25000        22.58558

        male_age_sample_weight  male_age_samples  female_age_mean  \
UID
267822               696.42136            2612.0         44.48629
263797               446.96441            1926.0         52.81825
270979               502.14915            2087.0         35.13247
259028               444.45947            1829.0         37.53793

        female_age_median  female_age_stdev  female_age_sample_weight  \
UID
267822           45.33333          22.51276                 685.33845
263797           55.00000          24.05831                 732.58443
270979           31.66667          22.66500                 565.32725
259028           35.91667          22.79602                 483.01311

        female_age_samples  pct_own  married  married_snp  separated  \
UID
```

```
267822              2618.0  0.79046  0.57851       0.01882     0.01240
263797              3124.0  0.64400  0.56377       0.01980     0.00990
270979              2528.0  0.61278  0.47397       0.04419     0.02663
259028              1954.0  0.83241  0.58678       0.01052     0.00000

        divorced  bad_debt  bins_bad_debt bins_debt
UID
267822   0.08770   0.09408  less than 10%   50-100%
263797   0.04892   0.18071         10-50%   50-100%
270979   0.13741   0.15005         10-50%   50-100%
259028   0.11721   0.02130  less than 10%    10-50%
```

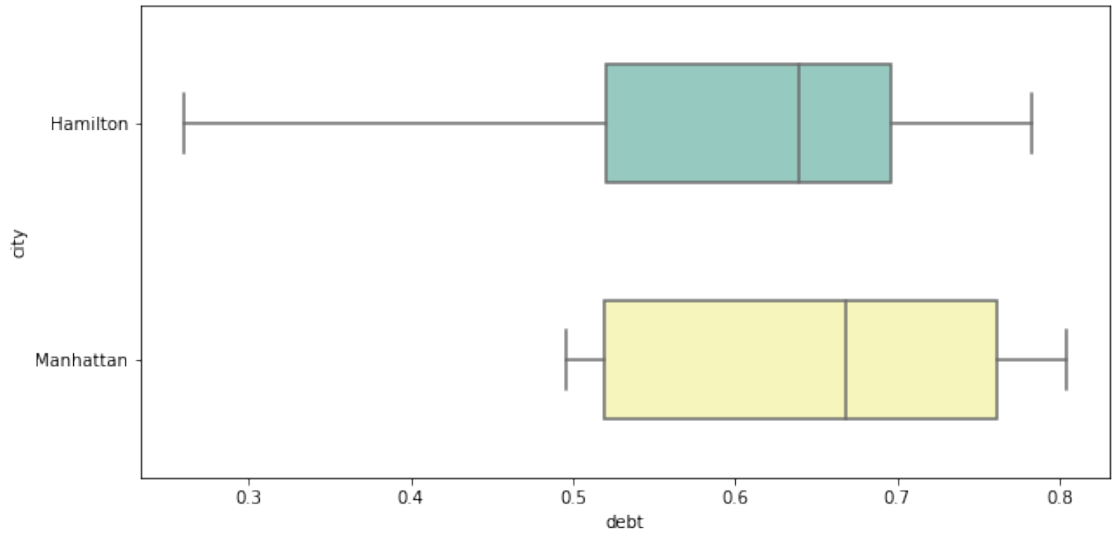
[30]:
```
# Visualization 4:
plt.figure(figsize=(10,5))
sns.boxplot(data=df_box_city,x='second_mortgage', y='city',width=0.
↪5,palette="Set3")
plt.show()
```



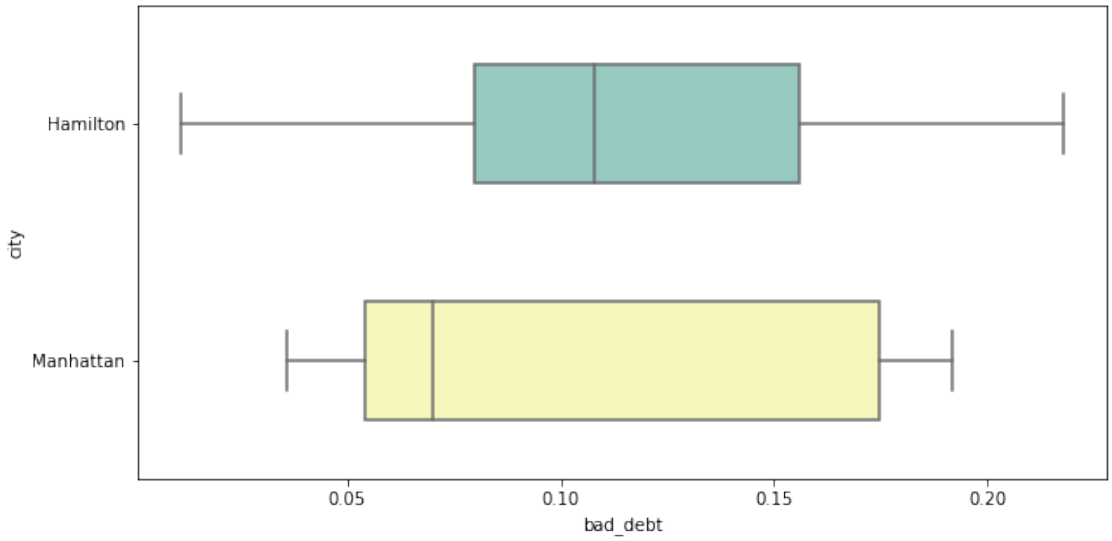
[31]:
```
# Visualization 5:
plt.figure(figsize=(10,5))
sns.boxplot(data=df_box_city,x='home_equity', y='city',width=0.5,palette="Set3")
plt.show()
```



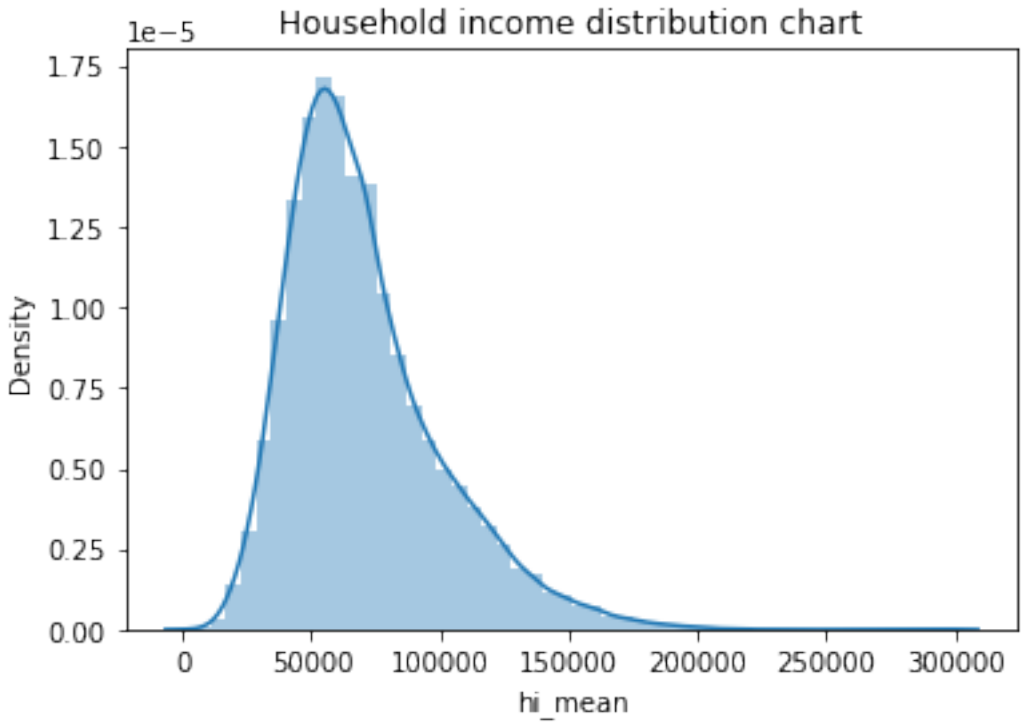
[32]:
```
# Visualization 6:
plt.figure(figsize=(10,5))
sns.boxplot(data=df_box_city,x='debt', y='city',width=0.5,palette="Set3")
plt.show()
```



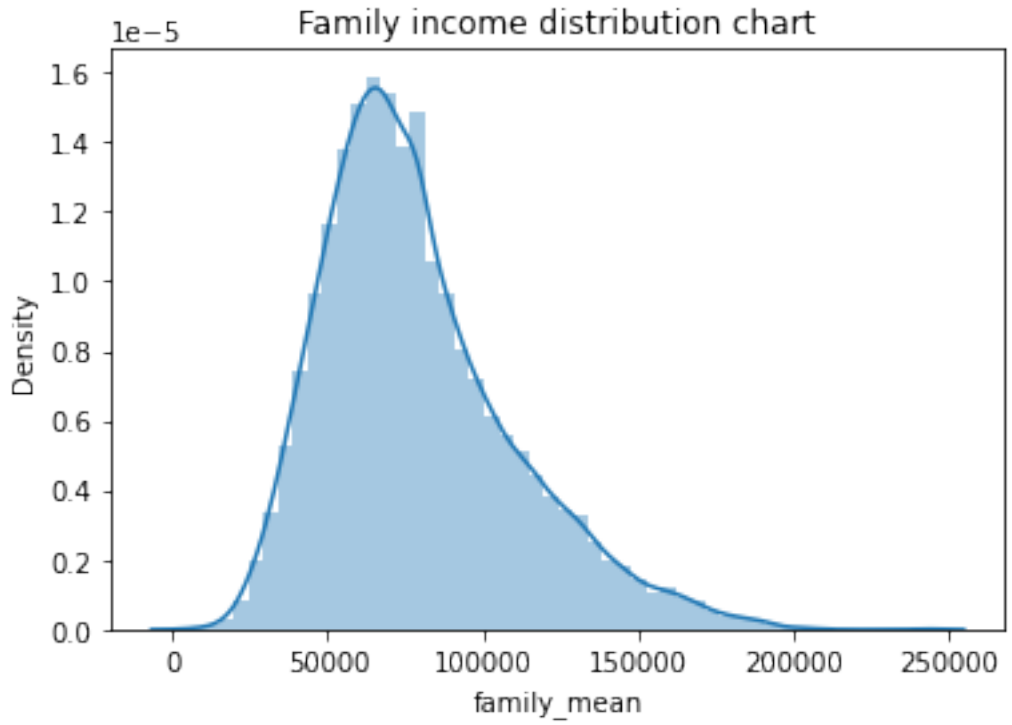
[33]:
```
# Visualization 7:
plt.figure(figsize=(10,5))
sns.boxplot(data=df_box_city,x='bad_debt', y='city',width=0.5,palette="Set3")
plt.show()
```



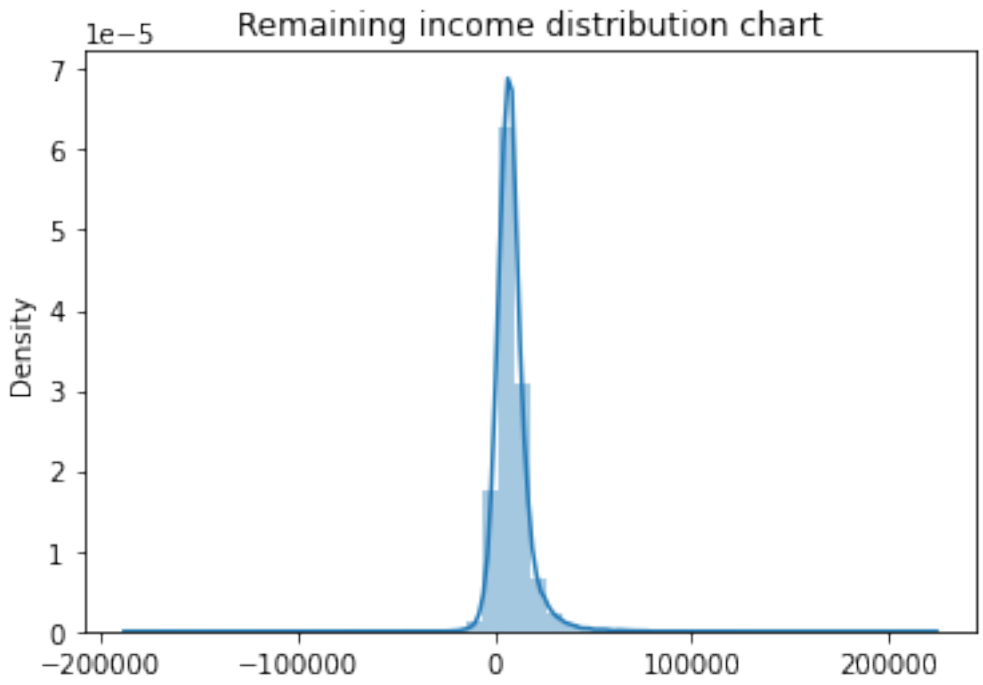
[34]:
```
# Visualization 8:
sns.distplot(train['hi_mean'])
plt.title('Household income distribution chart')
plt.show()
```



[35]:
```
# Visualization 9:
sns.distplot(train['family_mean'])
plt.title('Family income distribution chart')
plt.show()
```

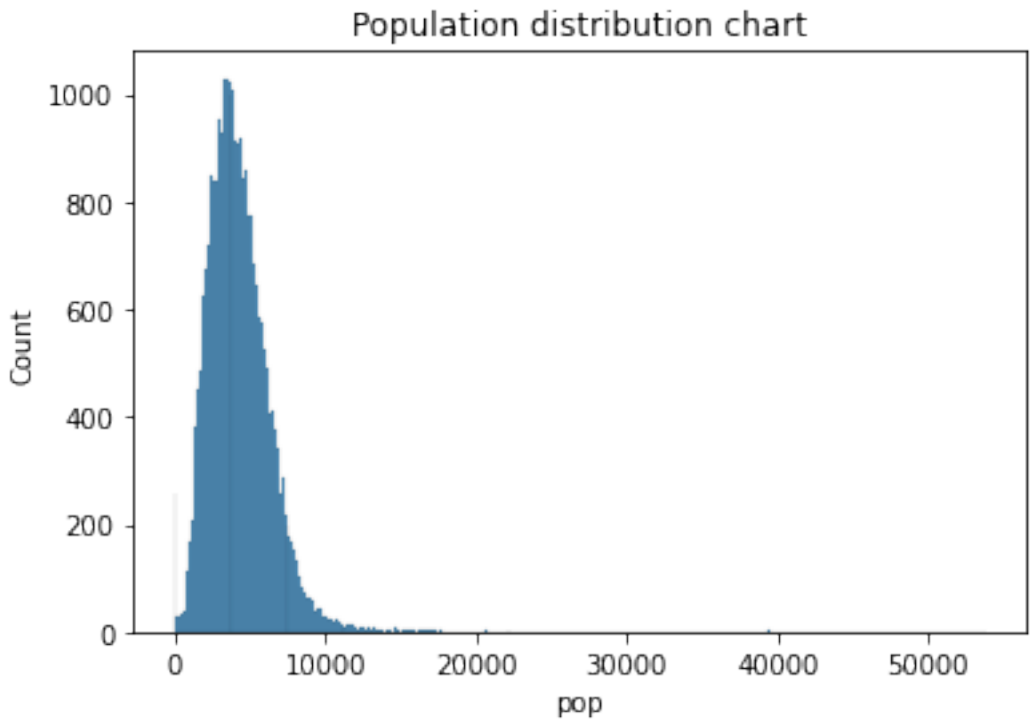


[36]:
```
# Visualization 10:
sns.distplot(train['family_mean']-train['hi_mean'])
plt.title('Remaining income distribution chart')
plt.show()
```

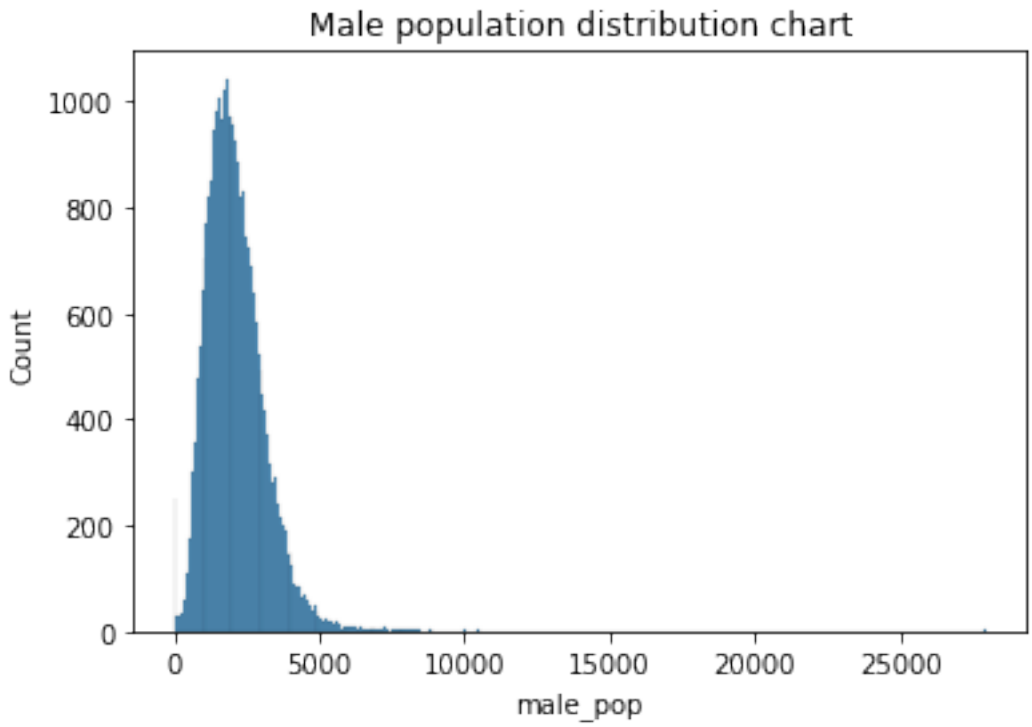


0.0.3 Week 2 Exploratory Data Analysis:

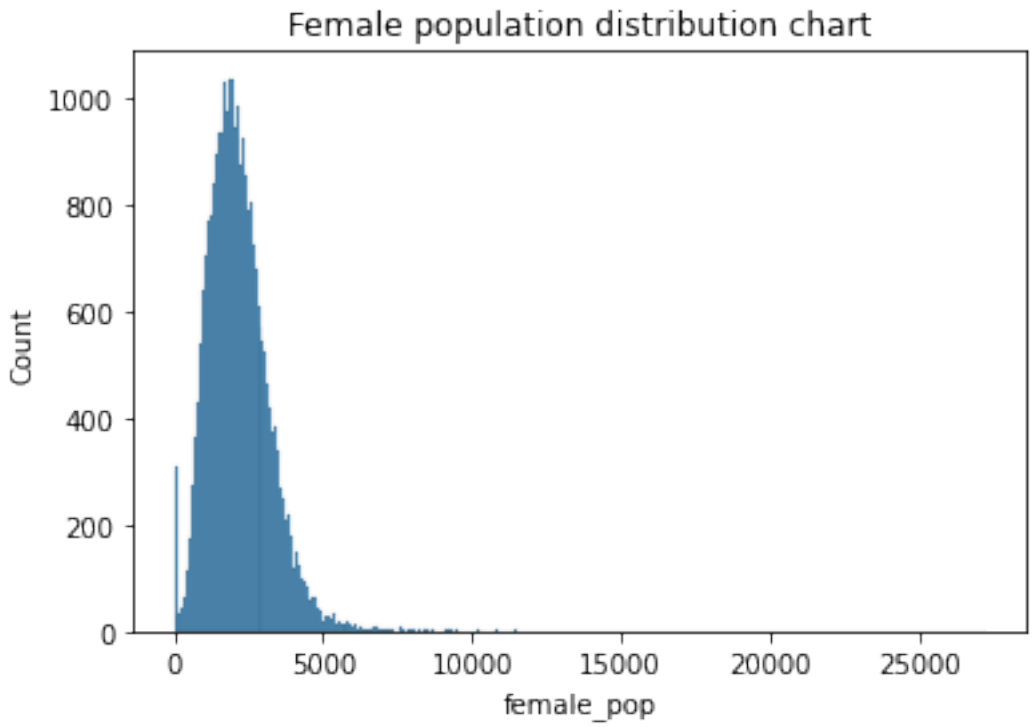
[37]:
```
# Visualization 11:
sns.histplot(train['pop'])
plt.title('Population distribution chart')
plt.show()
```



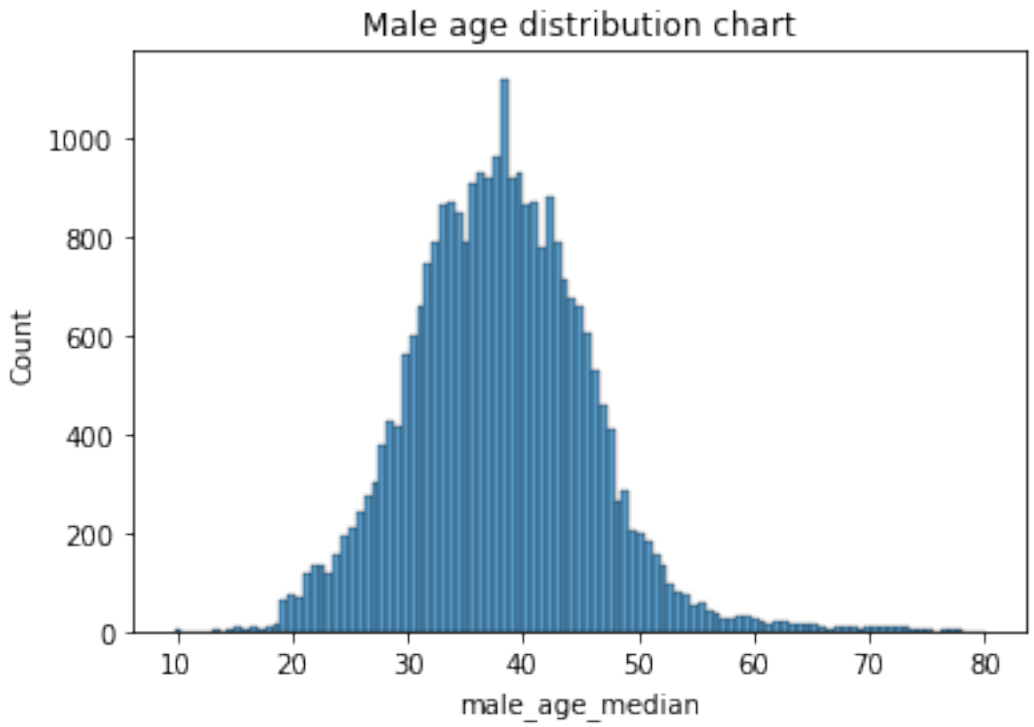
[38]:
```
# Visualization 12:
sns.histplot(train['male_pop'])
plt.title('Male population distribution chart')
plt.show()
```



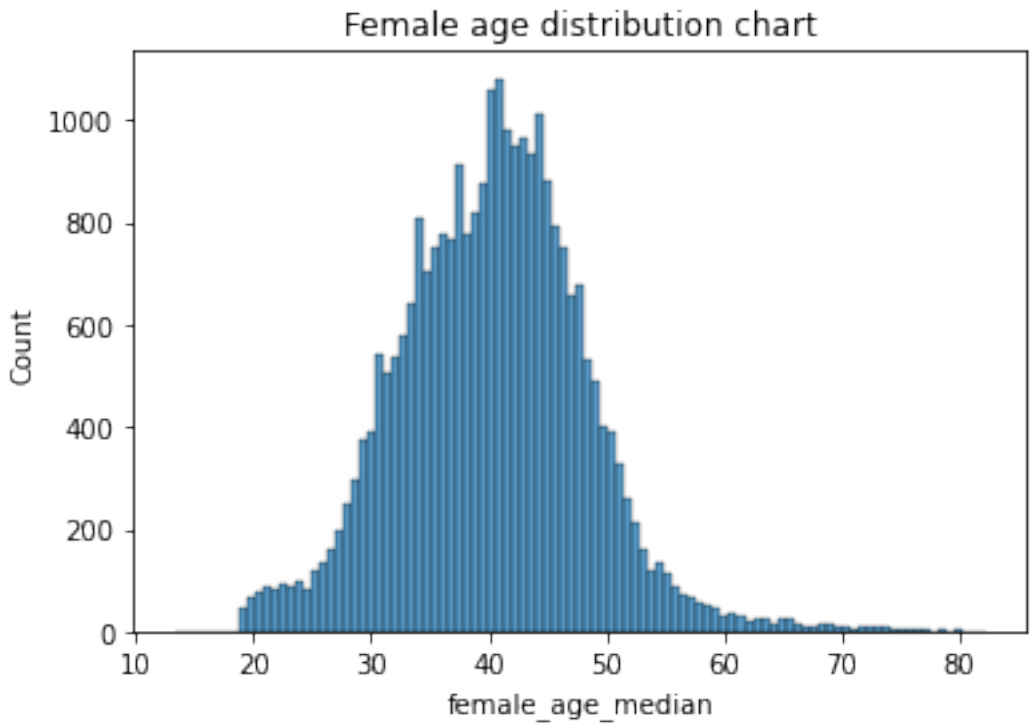
[39]:
```
# Visualization 13:
sns.histplot(train['female_pop'])
plt.title('Female population distribution chart')
plt.show()
```



[40]:
```
# Visualization 14:
sns.histplot(train['male_age_median'])
plt.title('Male age distribution chart')
plt.show()
```



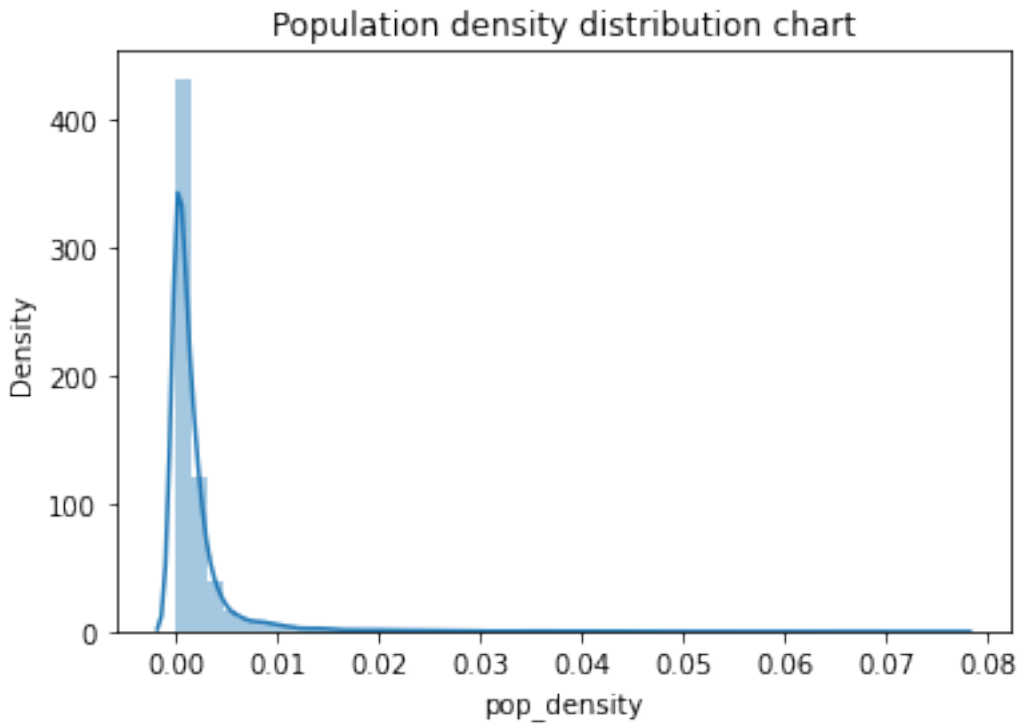
[41]:
```
# Visualization 15:
sns.histplot(train['female_age_median'])
plt.title('Female age distribution chart')
plt.show()
```



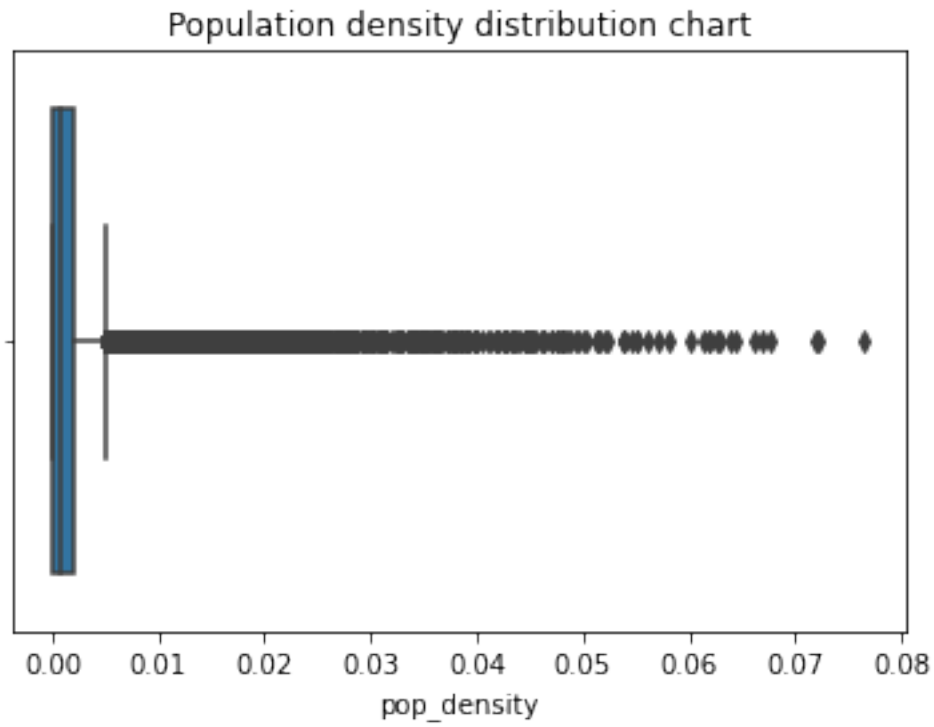
[42]:
```
train["pop_density"]=train["pop"]/train["ALand"]
```

[43]:
```
test["pop_density"]=test["pop"]/test["ALand"]
```

[44]:
```
# Visualization 16:
sns.distplot(train['pop_density'])
plt.title('Population density distribution chart')
plt.show()
```



[45]:
```
# Visualization 17:
sns.boxplot(train['pop_density'])
plt.title('Population density distribution chart')
plt.show()
```



[46]:
```
train["median_age"]=(train["male_age_median"]+train["female_age_median"])/2
```

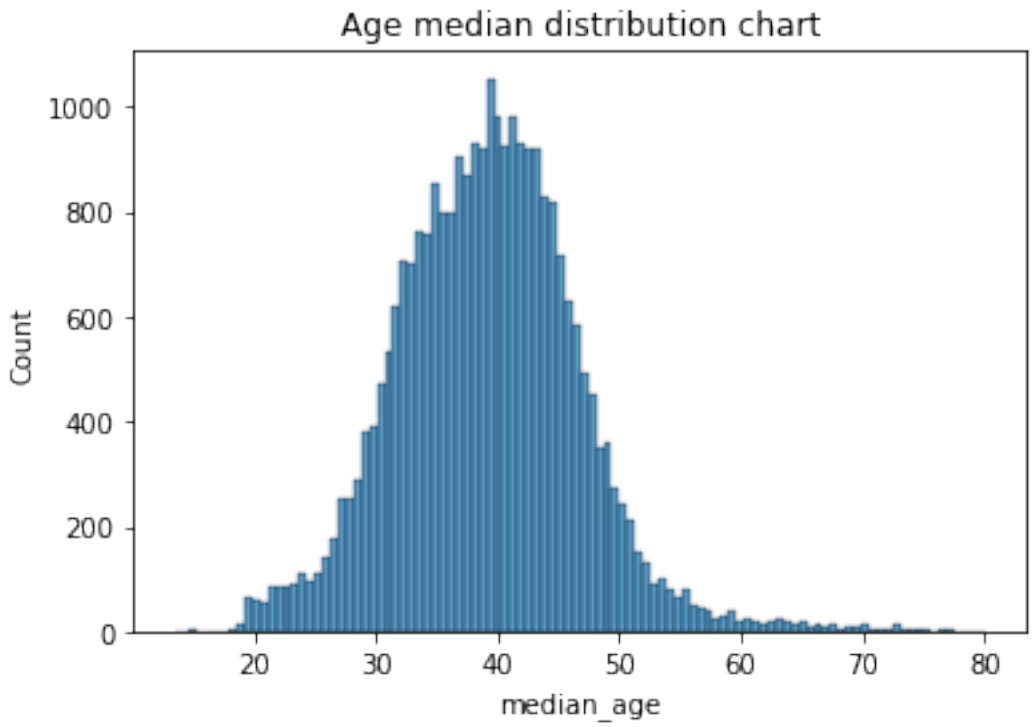
[47]:
```
test["median_age"]=(test["male_age_median"]+test["female_age_median"])/2
```

[48]:
```
train[['male_age_median','female_age_median','male_pop','female_pop','median_age']].
↪head()
```

[48]:

UID	male_age_median	female_age_median	male_pop	female_pop	median_age
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

[49]:
```
# Visualization 18:
sns.histplot(train['median_age'])
plt.title('Age median distribution chart')
plt.show()
```



[50]:
```
train["pop"].describe()
```

[50]:
```
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
```

[51]:
```
train['pop_bins']=pd.cut(train['pop'],bins=5,labels=['very␣
↪low','low','medium','high','very high'])
```

[52]:
```
train[['pop','pop_bins']]
```

[52]:

UID	pop	pop_bins
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]
```

[53]:
```
train['pop_bins'].value_counts()
```

[53]:
```
very low     27058
low            246
medium           9
high             7
very high        1
Name: pop_bins, dtype: int64
```

[54]:
```
train.groupby(by='pop_bins')[['married','separated','divorced']].count()
```

[54]:

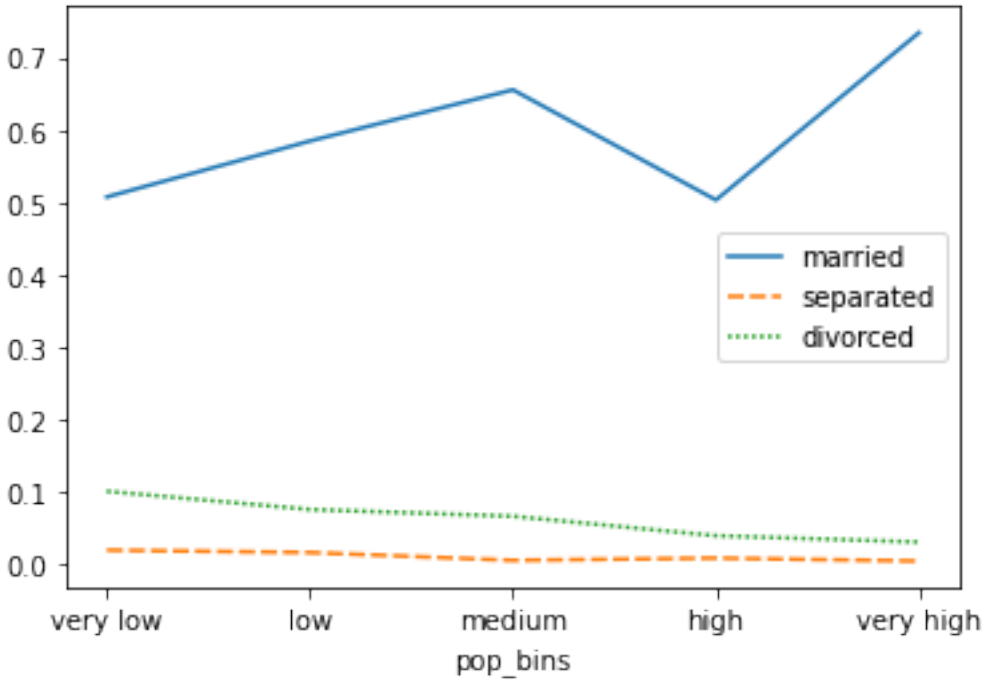
pop_bins	married	separated	divorced
very low	27058	27058	27058
low	246	246	246
medium	9	9	9
high	7	7	7
very high	1	1	1

[55]:
```
train.groupby(by='pop_bins')[['married','separated','divorced']].agg(["mean",␣
↪"median"])
```

[55]:

	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

[56]:
```
# Visualization 19:
pop_bin_married=train.
↪groupby(by='pop_bins')[['married','separated','divorced']].agg(["mean"])
sns.lineplot(data=pop_bin_married)
plt.show()
```



[57]:
```
rent_state_mean=train.groupby(by='state')['rent_mean'].agg(["mean"])
rent_state_mean.head()
```

[57]:
```
                  mean
state
Alabama       774.004927
Alaska       1185.763570
Arizona      1097.753511
Arkansas      720.918575
California   1471.133857
```

[58]:
```
income_state_mean=train.groupby(by='state')['family_mean'].agg(["mean"])
income_state_mean.head()
```

[58]:
```
                   mean
state
Alabama      67030.064213
Alaska       92136.545109
Arizona      73328.238798
Arkansas     64765.377850
California   87655.470820
```

[59]:
```
rent_perc_of_income=rent_state_mean['mean']/income_state_mean['mean']
rent_perc_of_income.head(10)
```

[59]:
```
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
```

[60]:
```
#overall level rent as a percentage of income
sum(train['rent_mean'])/sum(train['family_mean'])
```

[60]:
```
0.013358170721473864
```

[61]:
```
#Correlation analysis and heatmap
train[["COUNTYID","STATEID","zip_code",␣
 ↪"type","pop","family_mean",'second_mortgage', 'home_equity',␣
 ↪'debt','hs_degree','median_age','pct_own', 'married','separated',␣
 ↪'divorced']].corr()
```

[61]:

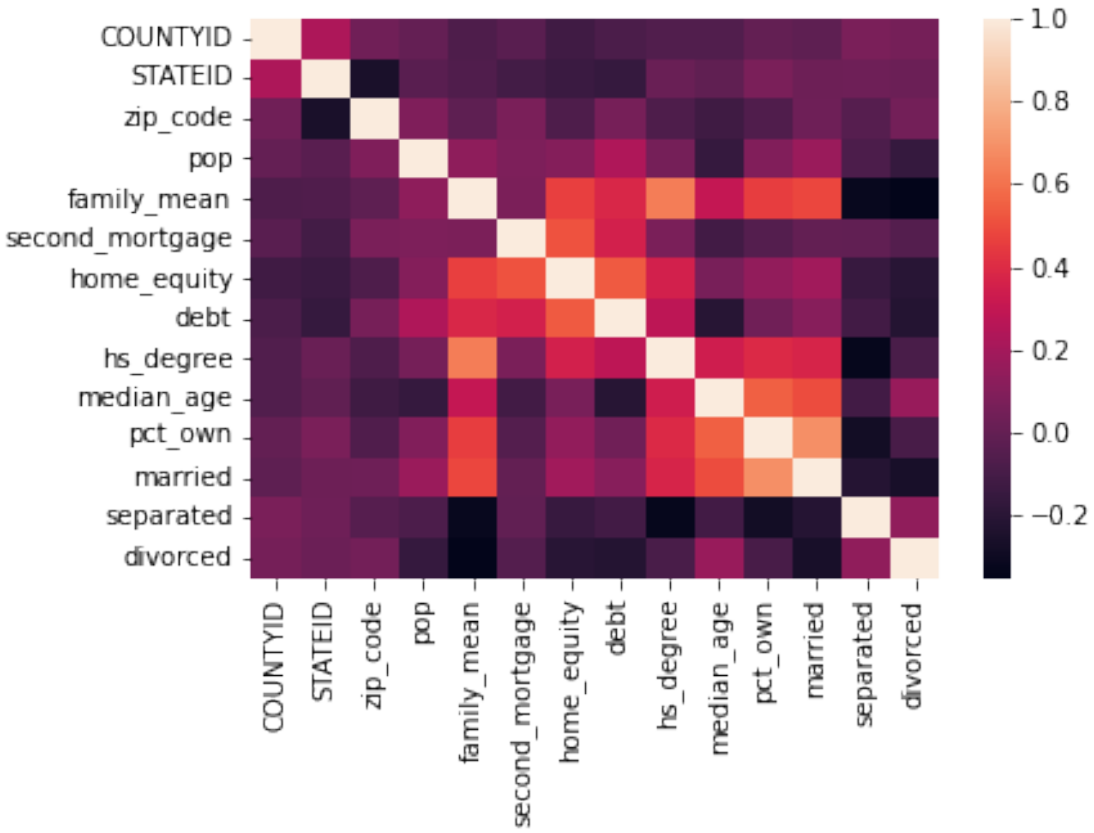
	COUNTYID	STATEID	zip_code	pop	family_mean	\
COUNTYID	1.000000	0.224549	0.036527	-0.002662	-0.075688	
STATEID	0.224549	1.000000	-0.261465	-0.036599	-0.071612	
zip_code	0.036527	-0.261465	1.000000	0.083058	-0.024658	
pop	-0.002662	-0.036599	0.083058	1.000000	0.128173	
family_mean	-0.075688	-0.071612	-0.024658	0.128173	1.000000	
second_mortgage	-0.039283	-0.112512	0.067693	0.079675	0.074703	
home_equity	-0.123939	-0.145301	-0.073191	0.099352	0.458973	
debt	-0.086231	-0.160532	0.057775	0.231013	0.378871	
hs_degree	-0.062703	0.014132	-0.077672	0.049238	0.634493	
median_age	-0.063521	-0.017172	-0.126150	-0.162499	0.300215	
pct_own	-0.004632	0.069314	-0.069965	0.088457	0.450961	
married	-0.021428	0.025763	0.030217	0.167656	0.480095	
separated	0.069059	0.030409	-0.048023	-0.083182	-0.323433	
divorced	0.048850	0.018748	0.043310	-0.160931	-0.353274	

	second_mortgage	home_equity	debt	hs_degree	\
COUNTYID	-0.039283	-0.123939	-0.086231	-0.062703	
STATEID	-0.112512	-0.145301	-0.160532	0.014132	
zip_code	0.067693	-0.073191	0.057775	-0.077672	
pop	0.079675	0.099352	0.231013	0.049238	
family_mean	0.074703	0.458973	0.378871	0.634493	
second_mortgage	1.000000	0.510460	0.351298	0.064412	

```
home_equity              0.510460     1.000000  0.532062   0.354566
debt                     0.351298     0.532062  1.000000   0.279957
hs_degree                0.064412     0.354566  0.279957   1.000000
median_age              -0.116616     0.063776 -0.213281   0.334228
pct_own                 -0.054530     0.140941  0.034207   0.390815
married                 -0.006438     0.189763  0.108496   0.370706
separated               -0.010731    -0.155198 -0.119073  -0.333321
divorced                -0.056991    -0.207202 -0.222350  -0.092984

                 median_age   pct_own   married  separated  divorced
COUNTYID          -0.063521 -0.004632 -0.021428   0.069059  0.048850
STATEID           -0.017172  0.069314  0.025763   0.030409  0.018748
zip_code          -0.126150 -0.069965  0.030217  -0.048023  0.043310
pop               -0.162499  0.088457  0.167656  -0.083182 -0.160931
family_mean        0.300215  0.450961  0.480095  -0.323433 -0.353274
second_mortgage   -0.116616 -0.054530 -0.006438  -0.010731 -0.056991
home_equity        0.063776  0.140941  0.189763  -0.155198 -0.207202
debt              -0.213281  0.034207  0.108496  -0.119073 -0.222350
hs_degree          0.334228  0.390815  0.370706  -0.333321 -0.092984
median_age         1.000000  0.546692  0.495153  -0.116763  0.164205
pct_own            0.546692  1.000000  0.683960  -0.284877 -0.095413
married            0.495153  0.683960  1.000000  -0.219686 -0.267833
separated         -0.116763 -0.284877 -0.219686   1.000000  0.133244
divorced           0.164205 -0.095413 -0.267833   0.133244  1.000000
```

[62]:
```
# Visualization 20:
sns.heatmap(train[["COUNTYID","STATEID","zip_code",␣
↪"type","pop","family_mean",'second_mortgage', 'home_equity',␣
↪'debt','hs_degree','median_age','pct_own', 'married','separated',␣
↪'divorced']].corr())
```

[62]: <AxesSubplot:>



0.0.4 Data Pre-processing:

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

[63]:
```
from sklearn.decomposition import FactorAnalysis
```

[64]:
```
fa = FactorAnalysis(n_components=5,random_state=11)
```

[65]:
```
train_transformed = fa.fit_transform(train.
↪select_dtypes(exclude=('object','category')))
```

[66]:
```
train_transformed.shape
```

[66]:
```
(27321, 5)
```

[67]:
```
train_transformed
```

[67]:
```
array([[ 0.05640687, -0.05073008,  1.25002287, -0.3262312 ,  0.18142577],
       [-0.10015645,  0.01442735,  0.11011385, -0.95809506,  0.58805728],
       [-0.04710979, -0.0094559 ,  0.13106345,  0.45168299,  0.90054992],
       …,
       [ 0.93167634, -0.37995383, -0.96907522,  0.41947942,  0.30372135],
       [-0.08682288,  0.00848632, -0.88563901,  3.03163033,  1.15593997],
       [-0.09529886,  0.01164864, -1.3315217 , -0.69048312, -0.11200755]])
```

[68]:
```
x_train = pd.read_csv('train.csv')
x_test = pd.read_csv('test.csv')
```

[69]:
```
x_train.drop(['BLOCKID','SUMLEVEL'],axis=1,inplace=True)
```

[70]:
```
x_train.dropna(axis=0,inplace=True)
x_train.head()
```

[70]:

	UID	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
0	267822	53	36	New York	NY	Hamilton	Hamilton	City	tract	13346	315	42.840812	-75.501524	202183361.0	1699120	5230	2612	2618	769.38638	784.0
1	246444	141	18	Indiana	IN	South Bend	Roseland	City	tract	46616	574	41.701441	-86.266614	1560828.0	100363	2633	1349	1284	804.87924	848.0
2	245683	63	18	Indiana	IN	Danville	Danville	City	tract	46122	317	39.792202	-86.515246	69561595.0	284193	6881	3643	3238	742.77365	703.0
3	279653	127	72	Puerto Rico	PR	San Juan	Guaynabo	Urban	tract	927	787	18.396103	-66.104169	1105793.0	0	2700	1141	1559	803.42018	782.0
4	247218	161	20	Kansas	KS	Manhattan	Manhattan City	City	tract	66502	785	39.195573	-96.569366	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
```

[71]:
```
x_train.drop_duplicates(inplace=True)
```

[72]:
```
x_train.shape
```

[72]:
```
(26585, 78)
```

[73]:
```
x_test.head()
```

[73]:

	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	27.882461	-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
```

[74]:
```
x_test.shape
```

[74]:
```
(11709, 80)
```

[75]:
```
x_test.drop(['BLOCKID','SUMLEVEL'],axis=1,inplace=True)
```

[76]:
```
x_test.isna().sum()
```

[76]:
```
UID              0
COUNTYID         0
STATEID          0
state            0
state_ab         0
                ..
pct_own        122
married         84
```

```
married_snp     84
separated       84
divorced        84
Length: 78, dtype: int64
```

[77]:
```
x_test.dropna(axis=0,inplace=True)
```

[78]:
```
x_test.drop_duplicates(inplace=True)
```

[79]:
```
x_test.shape
```

[79]:
```
(11355, 78)
```

[80]:
```
imp_feature = x_train.select_dtypes(exclude=('object','category'))
```

[81]:
```
imp_feature.head()
```

[81]:

	UID	COUNTYID	STATEID	zip_code	area_code	lat	lng
0	267822	53	36	13346	315	42.840812	-75.501524
1	246444	141	18	46616	574	41.701441	-86.266614
2	245683	63	18	46122	317	39.792202	-86.515246
3	279653	127	72	927	787	18.396103	-66.104169
4	247218	161	20	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
```

[82]:
```
imp_feature.shape
```

[82]:
```
(26585, 72)
```

[83]:
```
to_drop = ['UID','COUNTYID', 'STATEID', 'zip_code', 'area_code', 'lat', 'lng']
```

[84]:
```
for col in imp_feature.columns:
    if col in to_drop:
        imp_feature.drop(col,axis=1,inplace=True)
```

[85]:
```
imp_feature.head()
```

[85]:
```
         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
```

[86]:
```
x_train_features =␣
↪imp_feature[['pop','rent_median','hi_median','family_median','hc_mean','second_mortgage','h
```

[87]:
```
x_train_features.head()
```

[87]:

	pop	rent_median	hi_median	family_median	hc_mean	second_mortgage	home_equity	debt	hs_degree	pct_own	married	separated	divorced
0	5230	784.0	48120.0	53245.0	570.01530	0.02077	0.08919	0.52963	0.89288	0.79046	0.57851	0.01240	0.08770
1	2633	848.0	35186.0	43023.0	351.98293	0.02222	0.04274	0.60855	0.90487	0.52483	0.34886	0.01426	0.09030
2	6881	703.0	74964.0	85395.0	556.45986	0.00000	0.09512	0.73484	0.94288	0.85331	0.64745	0.01607	0.10657
3	2700	782.0	37845.0	44399.0	288.04047	0.01086	0.01086	0.52714	0.91500	0.65037	0.47257	0.02021	0.10106
4	5637	881.0	22497.0	50272.0	443.68855	0.05426	0.05426	0.51938	1.00000	0.13046	0.12356	0.00000	0.03109

[88]:
```
x_train_features.shape
```

[88]: (26585, 13)

[89]:
```
y_train = imp_feature['hc_mortgage_mean']
```

[90]:
```
x_test_feature =␣
↪x_test[['pop','rent_median','hi_median','family_median','hc_mean','second_mortgage','home_e
```

[91]:
```
from sklearn.linear_model import LinearRegression
le = LinearRegression()
```

[92]:
```
le.fit(x_train_features,y_train)
```

[92]: LinearRegression()

[93]:
```
y_pred = le.predict(x_test_feature)
```

[94]:
```
y_test = x_test['hc_mortgage_mean']
```

[95]:
```
from sklearn.metrics import r2_score,mean_squared_error
```

[96]:
```
r2_score(y_test,y_pred)
```

[96]: 0.8073813546881963

[97]:
```
np.sqrt(mean_squared_error(y_test,y_pred))
```

[97]: 277.0451838858074

[ ]:
```
# Visualization 21:
sns.distplot(y_pred)
plt.show()
```