Project 1 - Real Estate

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

Project Task: Week 1

Data Import and Preparation:

```
1. Import data.
2. Figure out the primary key and look for the requirement of indexing.
3. Gauge the fill rate of the variables and devise plans for missing value treatment. Please explicit
ly the reason for the treatment chosen for each variable.
```

Exploratory Data Analysis (EDA):

4.Perform debt analysis. You may take the following steps:

1. Import Data

In [2]:
```
df_train = pd.read_csv('p1train.csv')
```

In [3]:
```
df_test = pd.read_csv('p1test.csv')
```

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

Out[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')
```

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

Out[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')
```

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

Out[6]:
```
27321
```

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

Out[7]:
```
11709
```

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

Out[8]:

	UID	BLOCKID	SUMLEVEL	COUNTYID	STATEID	state	state_ab	city	place	type	...	female_age_mean	female_age_median
0	267822	NaN	140	53	36	New York	NY	Hamilton	Hamilton	City	...	44.48629	45.33333
1	246444	NaN	140	141	18	Indiana	IN	South Bend	Roseland	City	...	36.48391	37.58333
2	245683	NaN	140	63	18	Indiana	IN	Danville	Danville	City	...	42.15810	42.83333
3	279653	NaN	140	127	72	Puerto Rico	PR	San Juan	Guaynabo	Urban	...	47.77526	50.58333
4	247218	NaN	140	161	20	Kansas	KS	Manhattan	Manhattan City	City	...	24.17693	21.58333

5 rows × 80 columns

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

Out[9]:

	UID	BLOCKID	SUMLEVEL	COUNTYID	STATEID	state	state_ab	city	place	type	...	female_age_mean	female_age_r
0	255504	NaN	140	163	26	Michigan	MI	Detroit	Dearborn Heights City	CDP	...	34.78682	33
1	252676	NaN	140	1	23	Maine	ME	Auburn	Auburn City	City	...	44.23451	46
2	276314	NaN	140	15	42	Pennsylvania	PA	Pine City	Millerton	Borough	...	41.62426	44
3	248614	NaN	140	231	21	Kentucky	KY	Monticello	Monticello City	City	...	44.81200	48
4	286865	NaN	140	355	48	Texas	TX	Corpus Christi	Edroy	Town	...	40.66618	42

5 rows × 80 columns

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

Out[10]:

	UID	BLOCKID	SUMLEVEL	COUNTYID	STATEID	zip_code	area_code	lat	lng	ALand
count	27321.000000	0.0	27321.0	27321.000000	27321.000000	27321.000000	27321.000000	27321.000000	27321.000000	2.732100e+04
mean	257331.996303	NaN	140.0	85.646426	28.271806	50081.999524	596.507668	37.508813	-91.288394	1.295106e+08
std	21343.859725	NaN	0.0	98.333097	16.392846	29558.115660	232.497482	5.588268	16.343816	1.275531e+09
min	220342.000000	NaN	140.0	1.000000	1.000000	602.000000	201.000000	17.929085	-165.453872	4.113400e+04
25%	238816.000000	NaN	140.0	29.000000	13.000000	26554.000000	405.000000	33.899064	-97.816067	1.799408e+06
50%	257220.000000	NaN	140.0	63.000000	28.000000	47715.000000	614.000000	38.755183	-86.554374	4.866940e+06
75%	275818.000000	NaN	140.0	109.000000	42.000000	77093.000000	801.000000	41.380606	-79.782503	3.359820e+07
max	294334.000000	NaN	140.0	840.000000	72.000000	99925.000000	989.000000	67.074017	-65.379332	1.039510e+11

8 rows × 74 columns

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

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

```
 36  hi_median                    27053 non-null  float64
 37  hi_stdev                     27053 non-null  float64
 38  hi_sample_weight             27053 non-null  float64
 39  hi_samples                   27053 non-null  float64
 40  family_mean                  27023 non-null  float64
 41  family_median                27023 non-null  float64
 42  family_stdev                 27023 non-null  float64
 43  family_sample_weight         27023 non-null  float64
 44  family_samples               27023 non-null  float64
 45  hc_mortgage_mean             26748 non-null  float64
 46  hc_mortgage_median           26748 non-null  float64
 47  hc_mortgage_stdev            26748 non-null  float64
 48  hc_mortgage_sample_weight    26748 non-null  float64
 49  hc_mortgage_samples          26748 non-null  float64
 50  hc_mean                      26721 non-null  float64
 51  hc_median                    26721 non-null  float64
 52  hc_stdev                     26721 non-null  float64
 53  hc_samples                   26721 non-null  float64
 54  hc_sample_weight             26721 non-null  float64
 55  home_equity_second_mortgage  26864 non-null  float64
 56  second_mortgage              26864 non-null  float64
 57  home_equity                  26864 non-null  float64
 58  debt                         26864 non-null  float64
 59  second_mortgage_cdf          26864 non-null  float64
 60  home_equity_cdf              26864 non-null  float64
 61  debt_cdf                     26864 non-null  float64
 62  hs_degree                    27131 non-null  float64
 63  hs_degree_male               27121 non-null  float64
 64  hs_degree_female             27098 non-null  float64
 65  male_age_mean                27132 non-null  float64
 66  male_age_median              27132 non-null  float64
 67  male_age_stdev               27132 non-null  float64
 68  male_age_sample_weight       27132 non-null  float64
 69  male_age_samples             27132 non-null  float64
 70  female_age_mean              27115 non-null  float64
 71  female_age_median            27115 non-null  float64
 72  female_age_stdev             27115 non-null  float64
 73  female_age_sample_weight     27115 non-null  float64
 74  female_age_samples           27115 non-null  float64
 75  pct_own                      27053 non-null  float64
 76  married                      27130 non-null  float64
 77  married_snp                  27130 non-null  float64
```

```
 78  separated                    27130 non-null  float64
 79  divorced                     27130 non-null  float64
dtypes: float64(62), int64(12), object(6)
memory usage: 16.7+ MB
```

In [12]:
```
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
```

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

In [13]:
```
df_train.set_index(keys=['UID'],inplace=True)
df_test.set_index(keys=['UID'],inplace=True)
```

In [14]:
```
df_train.head()
```

Out[14]:

	BLOCKID	SUMLEVEL	COUNTYID	STATEID	state	state_ab	city	place	type	primary	...	female_age_mean	female_age_n
UID													
267822	NaN	140	53	36	New York	NY	Hamilton	Hamilton	City	tract	...	44.48629	45
246444	NaN	140	141	18	Indiana	IN	South Bend	Roseland	City	tract	...	36.48391	37
245683	NaN	140	63	18	Indiana	IN	Danville	Danville	City	tract	...	42.15810	42
279653	NaN	140	127	72	Puerto Rico	PR	San Juan	Guaynabo	Urban	tract	...	47.77526	50
247218	NaN	140	161	20	Kansas	KS	Manhattan	Manhattan City	City	tract	...	24.17693	21

5 rows × 79 columns

In [15]:
```
df_test.head()
```

Out[15]:

	BLOCKID	SUMLEVEL	COUNTYID	STATEID	state	state_ab	city	place	type	primary	...	female_age_mean	female_
UID													
255504	NaN	140	163	26	Michigan	MI	Detroit	Dearborn Heights City	CDP	tract	...	34.78682	
252676	NaN	140	1	23	Maine	ME	Auburn	Auburn City	City	tract	...	44.23451	
276314	NaN	140	15	42	Pennsylvania	PA	Pine City	Millerton	Borough	tract	...	41.62426	
248614	NaN	140	231	21	Kentucky	KY	Monticello	Monticello City	City	tract	...	44.81200	
286865	NaN	140	355	48	Texas	TX	Corpus Christi	Edroy	Town	tract	...	40.66618	

5 rows × 79 columns

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

In [16]:
```
missing_list_train = df_train.isnull().sum() * 100/len(df_train)
```

In [17]:
```
missing_values_df_train = pd.DataFrame(missing_list_train,columns=['Percentage of Missing Values'])
```

In [18]:
```
missing_values_df_train.sort_values(by='Percentage of Missing Values',inplace=True,ascending=False)
```

In [19]:
```
missing_values_df_train[missing_values_df_train['Percentage of Missing Values'] > 0 ][:10]
```

Out[19]:

	Percentage 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

In [20]:
```
missing_list_test = df_test.isnull().sum() *100/len(df_train)
```

In [21]:
```
missing_values_df_test = pd.DataFrame(missing_list_test,columns=['Percentage of Missing Values'])
```

In [22]:
```
missing_values_df_test.sort_values(by='Percentage of Missing Values',inplace=True,ascending=False)
```

In [23]:
```
missing_values_df_test[missing_values_df_test['Percentage of Missing Values']>0][:10]
```

Out[23]:

	Percentage 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

In [24]:
```
df_train.drop(columns=['BLOCKID','SUMLEVEL'],inplace=True)
# BLOCKID can be dropped, since it has 100% missing Values
# SUMLEVEL can be dropped, since it does not have any predictive power and no variance
```

In [25]:
```
df_test.drop(columns=['BLOCKID','SUMLEVEL'],inplace=True)
# BLOCKID can be dropped, since it has 43% missing Values
# SUMLEVEL can be dropped, since it does not have any predictive power and no variance
```

In [26]:
```
# Imputing the missing values in other columns 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_g
t_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_sampl
es', 'hc_mortgage_mean', 'hc_mortgage_median', 'hc_mortgage_stdev', 'hc_mortgage_sample_weight', 'hc_mortgage_sample
s', 'hc_mean', 'hc_median', 'hc_stdev', 'hc_samples', 'hc_sample_weight', 'home_equity_second_mortgage', 'second_mor
tgage', '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_sample
s', 'female_age_mean', 'female_age_median', 'female_age_stdev', 'female_age_sample_weight', 'female_age_samples', 'p
ct_own', 'married', 'married_snp', 'separated', 'divorced']
```

In [27]:
```
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_g
t_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_sampl
es', 'hc_mortgage_mean', 'hc_mortgage_median', 'hc_mortgage_stdev', 'hc_mortgage_sample_weight', 'hc_mortgage_sample
s', 'hc_mean', 'hc_median', 'hc_stdev', 'hc_samples', 'hc_sample_weight', 'home_equity_second_mortgage', 'second_mor
tgage', '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_sample
s', 'female_age_mean', 'female_age_median', 'female_age_stdev', 'female_age_sample_weight', 'female_age_samples', 'p
ct_own', 'married', 'married_snp', 'separated', 'divorced']
```

In [28]:
```
for col in df_train.columns:
    if col in (missing_train_cols):
        df_train[col].replace(np.nan, df_train[col].mean(),inplace=True)
```

In [29]:
```
for col in df_test.columns:
    if col in (missing_test_cols):
        df_test[col].replace(np.nan, df_test[col].mean(),inplace=True)
```

In [30]:
```
df_train.isnull().sum().sum()
```

Out[30]: 0

In [31]:
```
df_test.isnull().sum().sum()
```

Out[31]: 0

Exploratory Data Analysis (EDA):

4.Perform debt analysis. You may take the following steps:

```
a) Explore the top 2,500 locations where the percentage of households with a second mortgage is the highest a
nd percent ownership is above 10 percent. Visualize using geo-map. You may keep the upper limit for the perce
nt of households with a second mortgage to 50 percent

b) Use the following bad debt equation: Bad Debt = P (Second Mortgage ∩ Home Equity Loan) Bad Debt = second_m
ortgage + home_equity - home_equity_second_mortgage

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

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

e) Create a collated income distribution chart for family income, house hold income, and remaining income
```

a) Explore the top 2,500 locations where the percentage of households with a second mortgage is the highest and percent ownership is above 10 percent. Visualize using geo-map. You may keep the upper limit for the percent of households with a second mortgage to 50 percent

In [32]:
```
from pandasql import sqldf
```

In [33]:
```
q1 = "select place,pct_own,second_mortgage,lat,lng from df_train where pct_own > 0.10 and second_mortgage < 0.5 order
```

In [34]:
```
pysqldf = lambda q: sqldf(q, globals())
```

In [35]:
```
df_train_location_mort_pct = pysqldf(q1)
```

In [36]:
```
df_train_location_mort_pct.head()
```

Out[36]:

	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

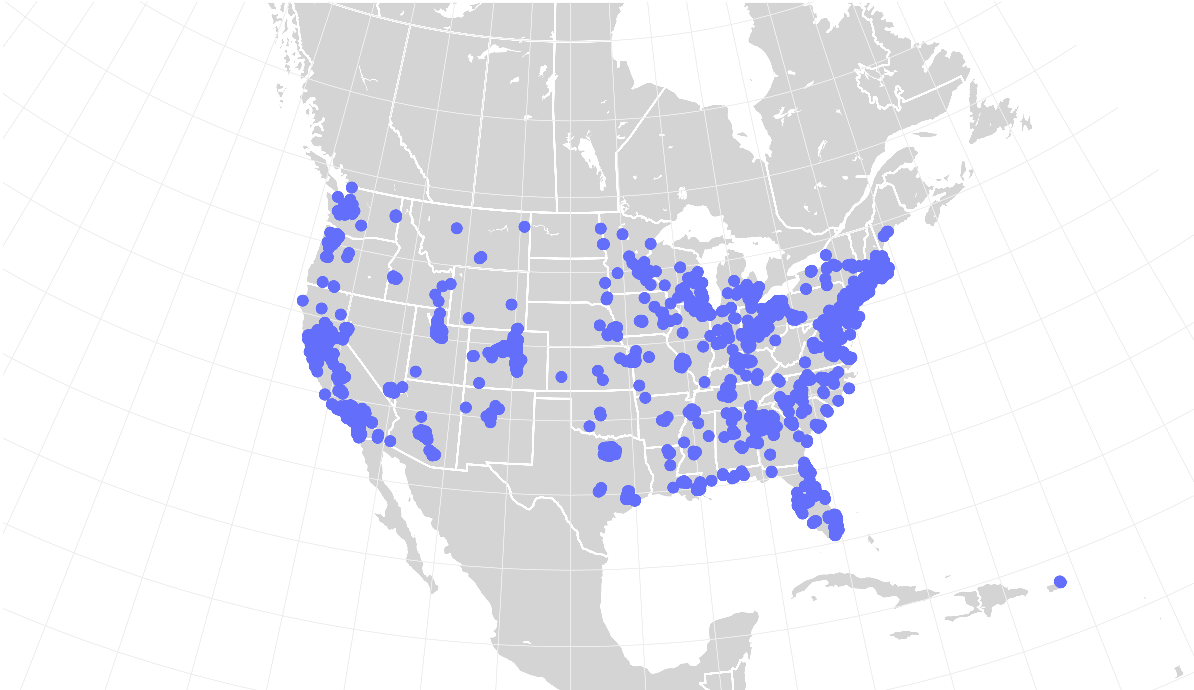
In [37]:
```
import plotly.express as px
```

In [38]:
```
import plotly.graph_objects as go
```

In [39]:
```
fig = go.Figure(data=go.Scattergeo(
    lat = df_train_location_mort_pct['lat'],
    lon = df_train_location_mort_pct['lng']),
    )
```

In [40]:
```
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 2500 locations with second mortgage is the highest and percent ownership is above 10 percent')
fig.show()
```

Top 2500 locations with second mortgage is the highest and percent ownership is above 10 percent



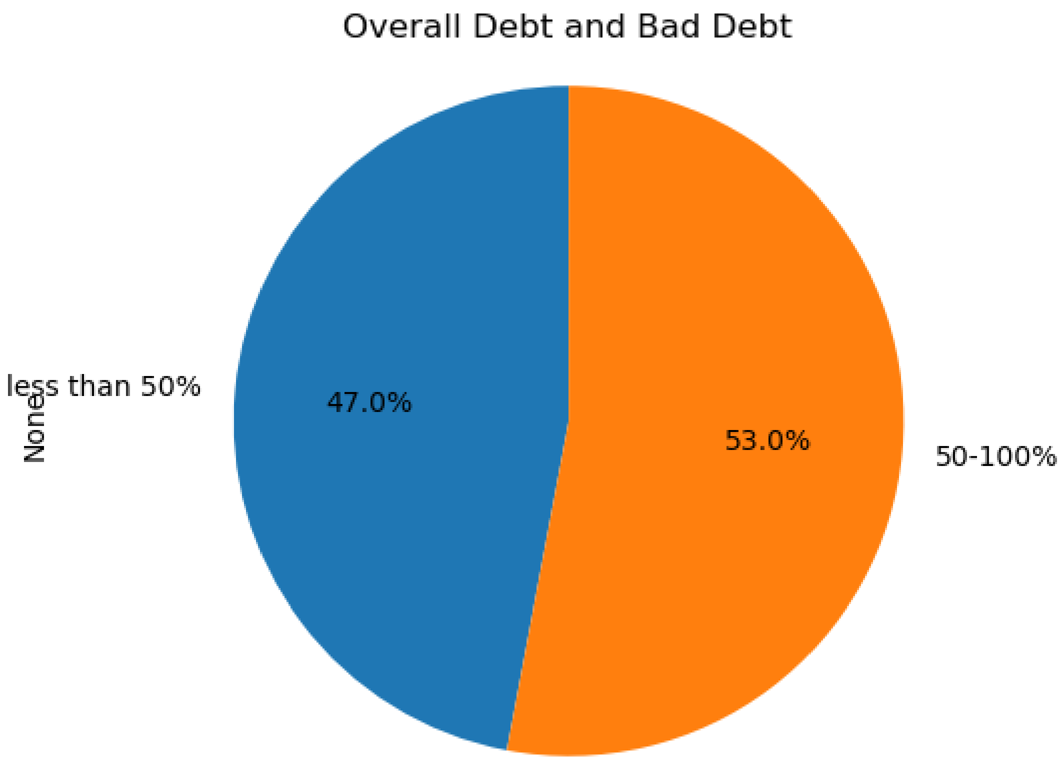
b) Use the following bad debt equation: Bad Debt = P (Second Mortgage ∩ Home Equity Loan) Bad Debt = second_mortgage + home_equity - home_equity_second_mortgage

In [41]:
```
df_train['bad_debt'] = df_train['second_mortgage']+df_train['home_equity']-df_train['home_equity_second_mortgage']
```

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

In [42]:
```
df_train['bins'] = pd.cut(df_train['bad_debt'],bins=[0,0.10,1],labels=['less than 50%','50-100%'])
```

In [43]:
```
df_train.groupby(['bins']).size().plot(kind='pie',subplots=True,startangle=90,autopct='%1.1f%%')
plt.axis('equal')
plt.title('Overall Debt and Bad Debt')
plt.show()
```



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

In [44]:
```
cols = []
df_train.columns
```

Out[44]:
```
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')
```

In [45]:
```
cols = ['second mortgage','home_equity','debt','bad_dept']
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])
```

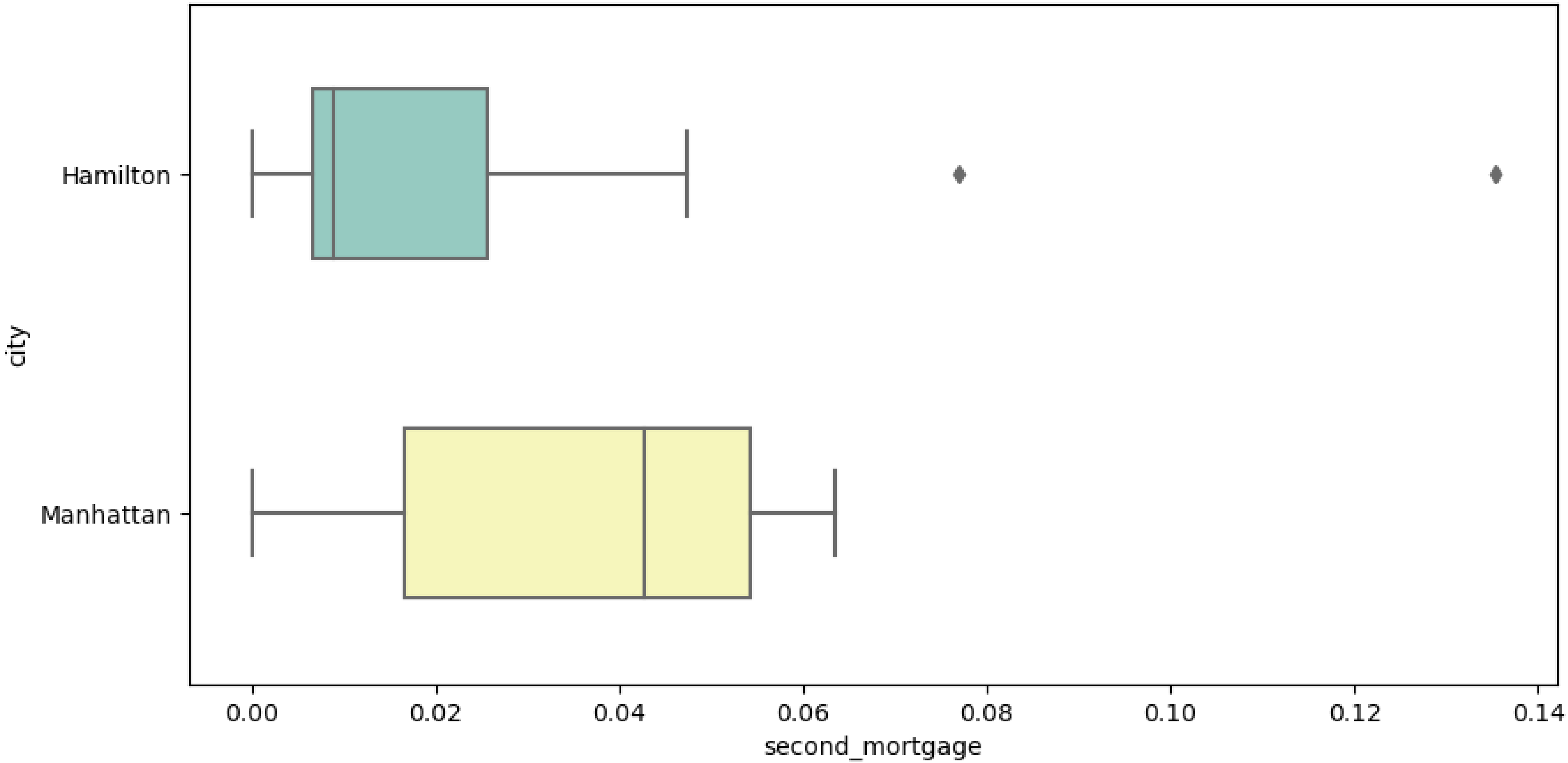
In [46]:
```
df_box_city.head()
```

Out[46]:

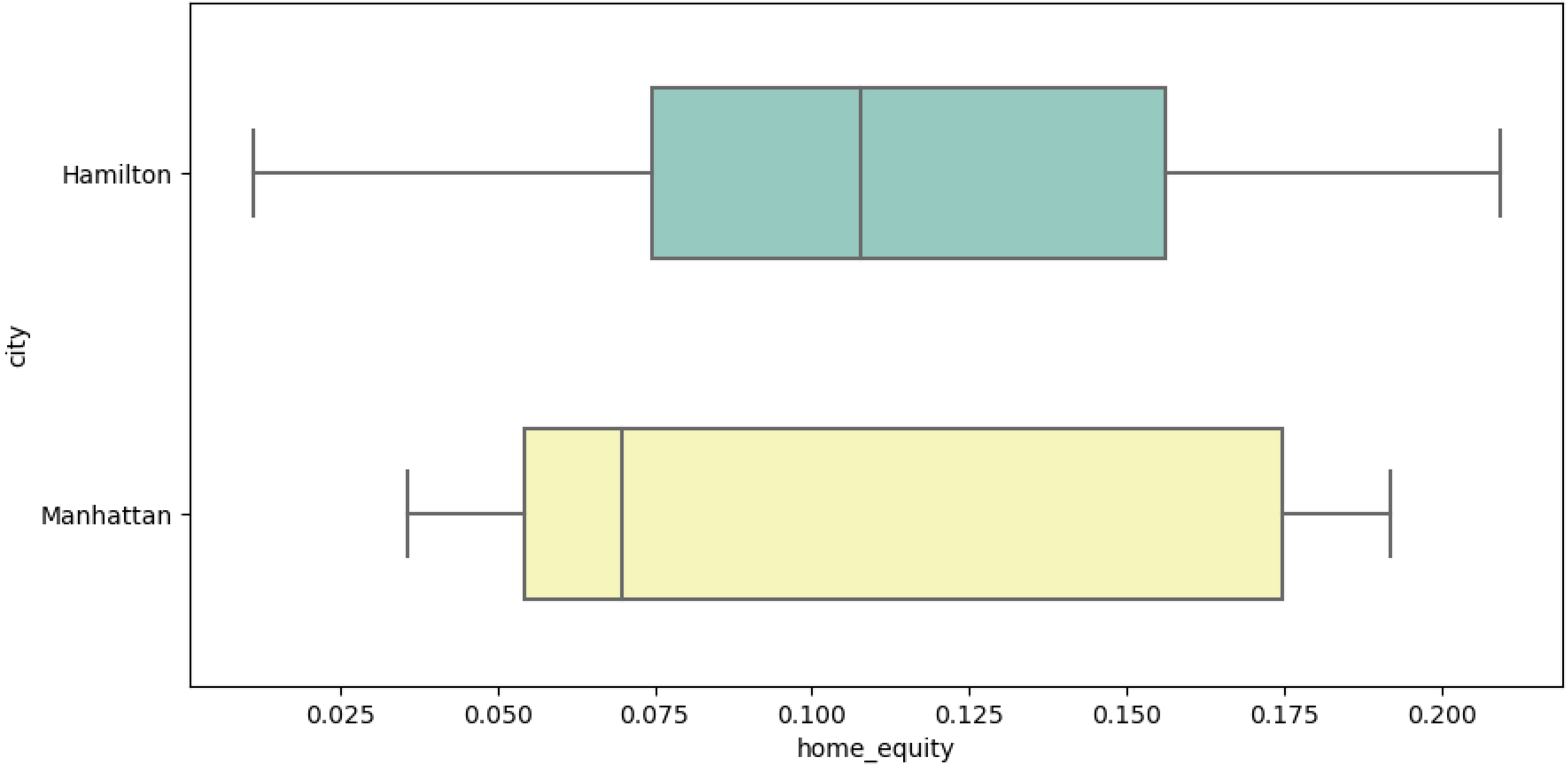
	COUNTYID	STATEID	state	state_ab	city	place	type	primary	zip_code	area_code	...	female_age_stdev	female_age_sa
UID													
267822	53	36	New York	NY	Hamilton	Hamilton	City	tract	13346	315	...	22.51276	
263797	21	34	New Jersey	NJ	Hamilton	Yardville	City	tract	8610	609	...	24.05831	
270979	17	39	Ohio	OH	Hamilton	Hamilton City	Village	tract	45015	513	...	22.66500	
259028	95	28	Mississippi	MS	Hamilton	Hamilton	CDP	tract	39746	662	...	22.79602	
270984	17	39	Ohio	OH	Hamilton	New Miami	Village	tract	45013	513	...	24.55724	

5 rows × 79 columns

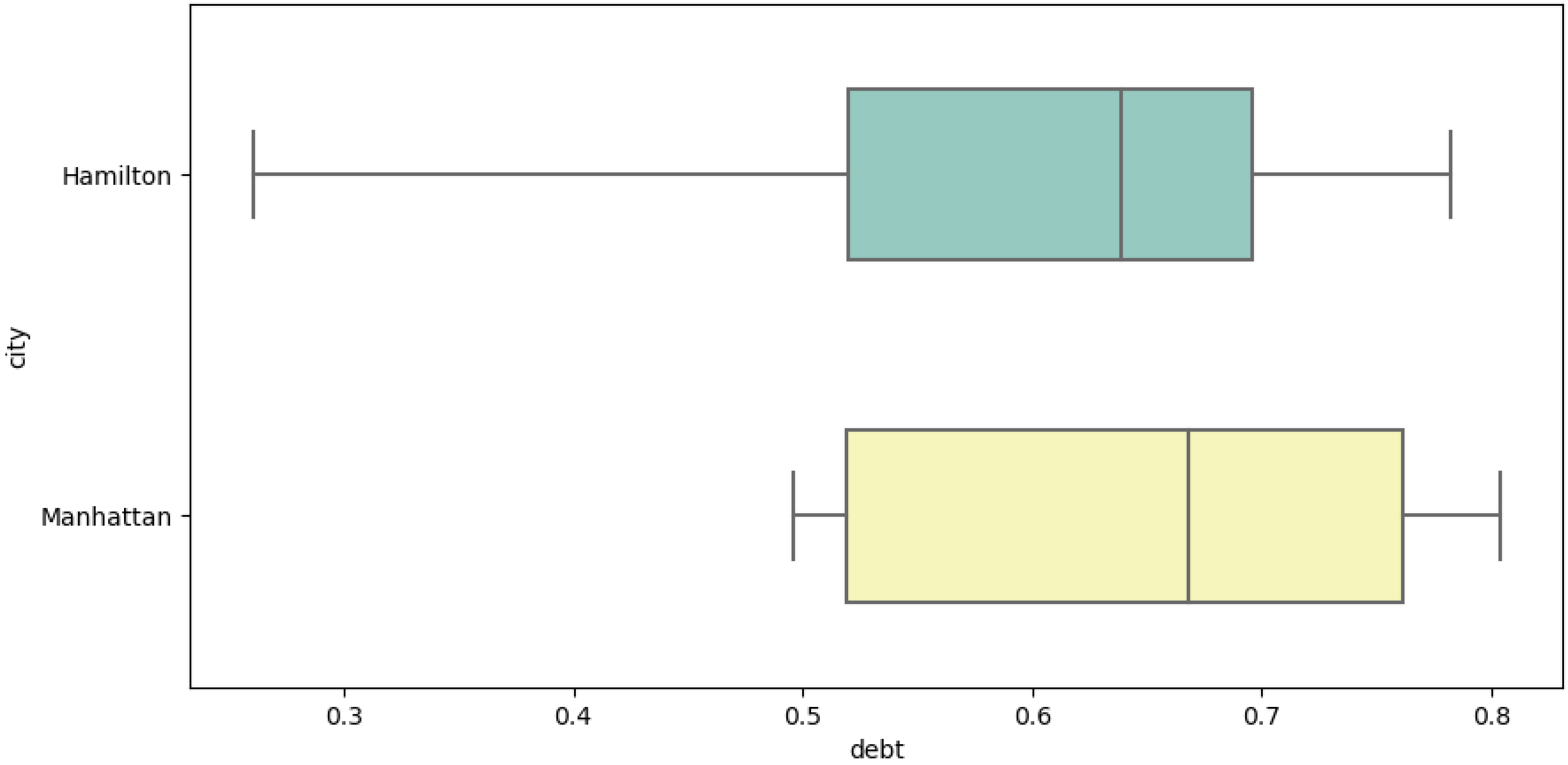
In [47]:
```
plt.figure(figsize=(10,5))
sns.boxplot(data=df_box_city,x='second_mortgage',y='city',width=0.5,palette='Set3')
plt.show()
```



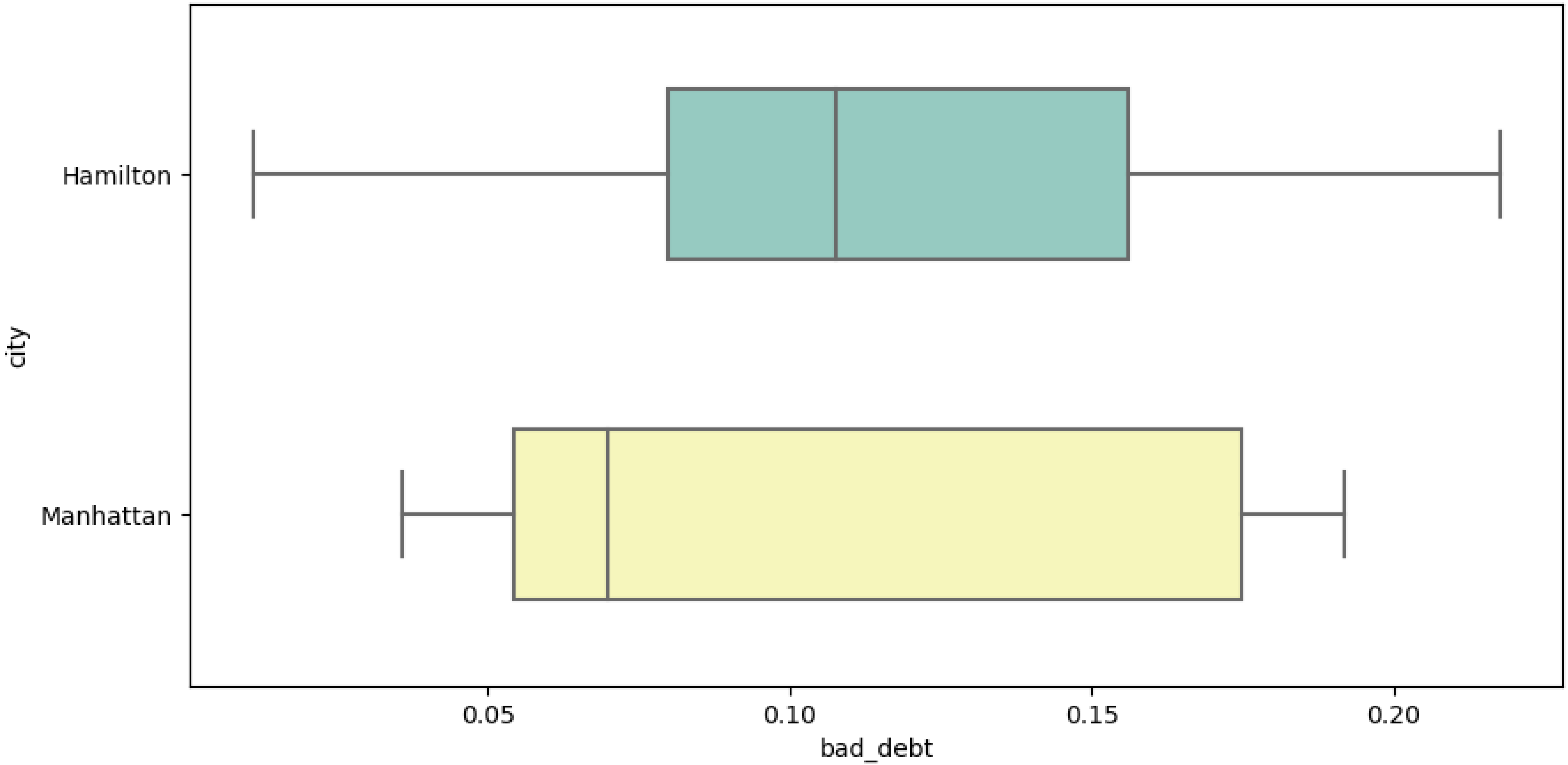
In [48]:
```
plt.figure(figsize=(10,5))
sns.boxplot(data=df_box_city,x='home_equity',y='city',width=0.5,palette='Set3')
plt.show()
```



In [49]:
```
plt.figure(figsize=(10,5))
sns.boxplot(data=df_box_city,x='debt',y='city',width=0.5,palette='Set3')
plt.show()
```



In [50]:
```
plt.figure(figsize=(10,5))
sns.boxplot(data=df_box_city,x='bad_debt',y='city',width=0.5,palette='Set3')
plt.show()
```

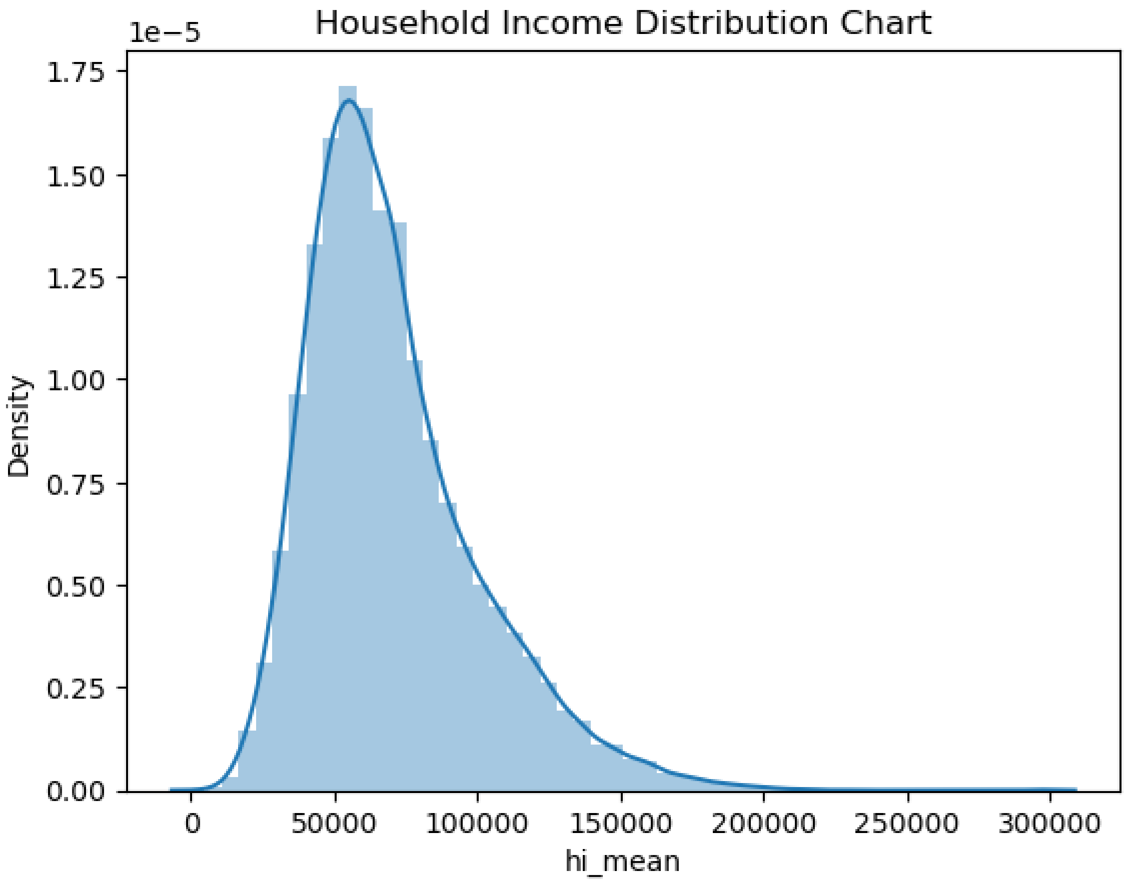


Manhattan has higher metrices compared to Hamilton

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

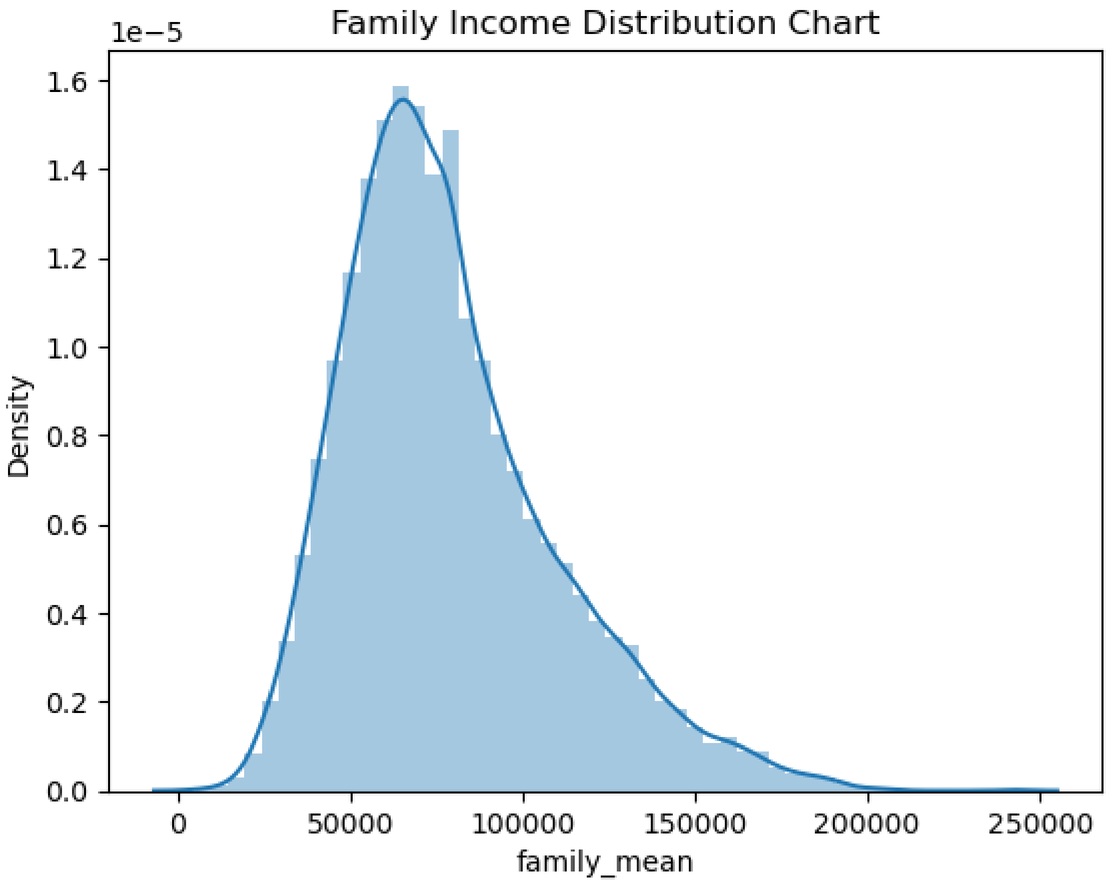
In [51]:
```
sns.distplot(df_train['hi_mean'])
plt.title('Household Income Distribution Chart')
plt.show()
```

```
C:\Users\Vinosh\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning:

`distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `d
isplot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
```



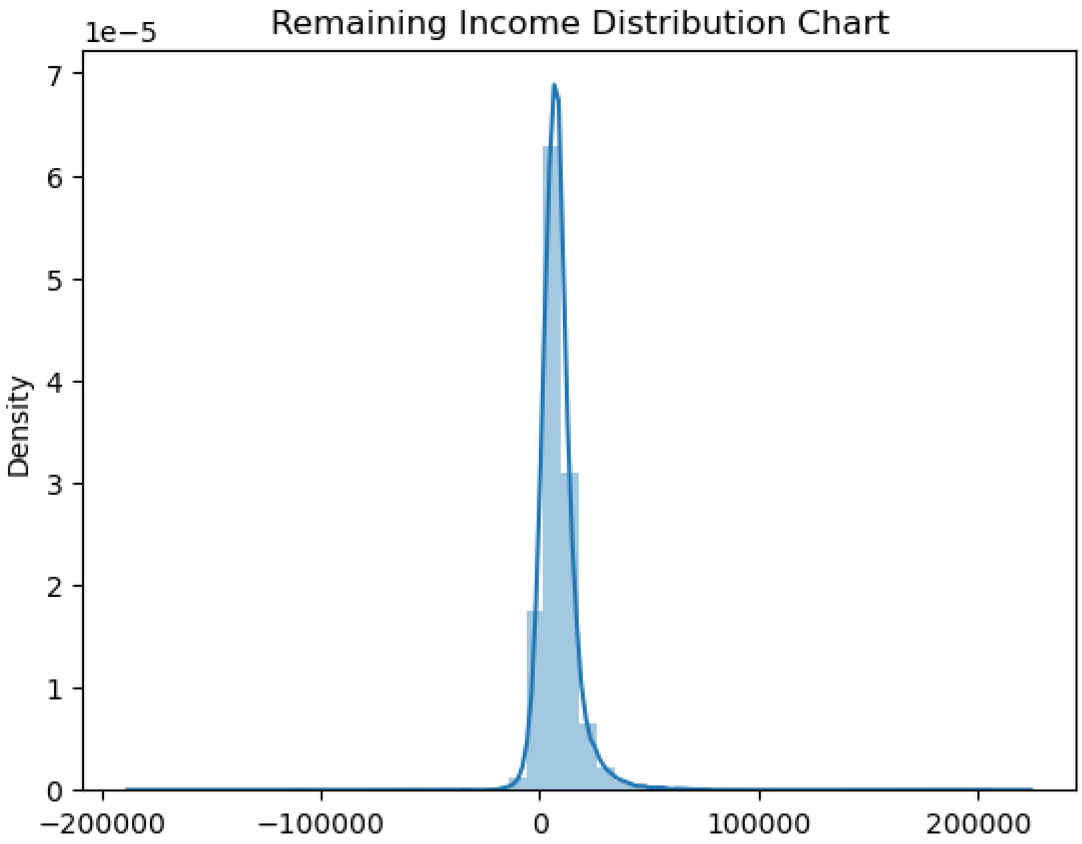
In [52]:
```
sns.distplot(df_train['family_mean'])
plt.title('Family Income Distribution Chart')
plt.show()
```

```
C:\Users\Vinosh\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning:

`distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `d
isplot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
```



In [53]:
```
sns.distplot(df_train['family_mean']-df_train['hi_mean'])
plt.title('Remaining Income Distribution Chart')
plt.show()
```

```
C:\Users\Vinosh\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning:

`distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `d
isplot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
```



Income Distribution has Normal Distribution

Project Task: Week 2

Exploratory Data Analysis (EDA):

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

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

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

   c) Visualize the findings using appropriate chart type
2. Create bins for population into a new variable by selecting appropriate class interval so that the number of categories don’t exceed 5 for the ease of analysis.

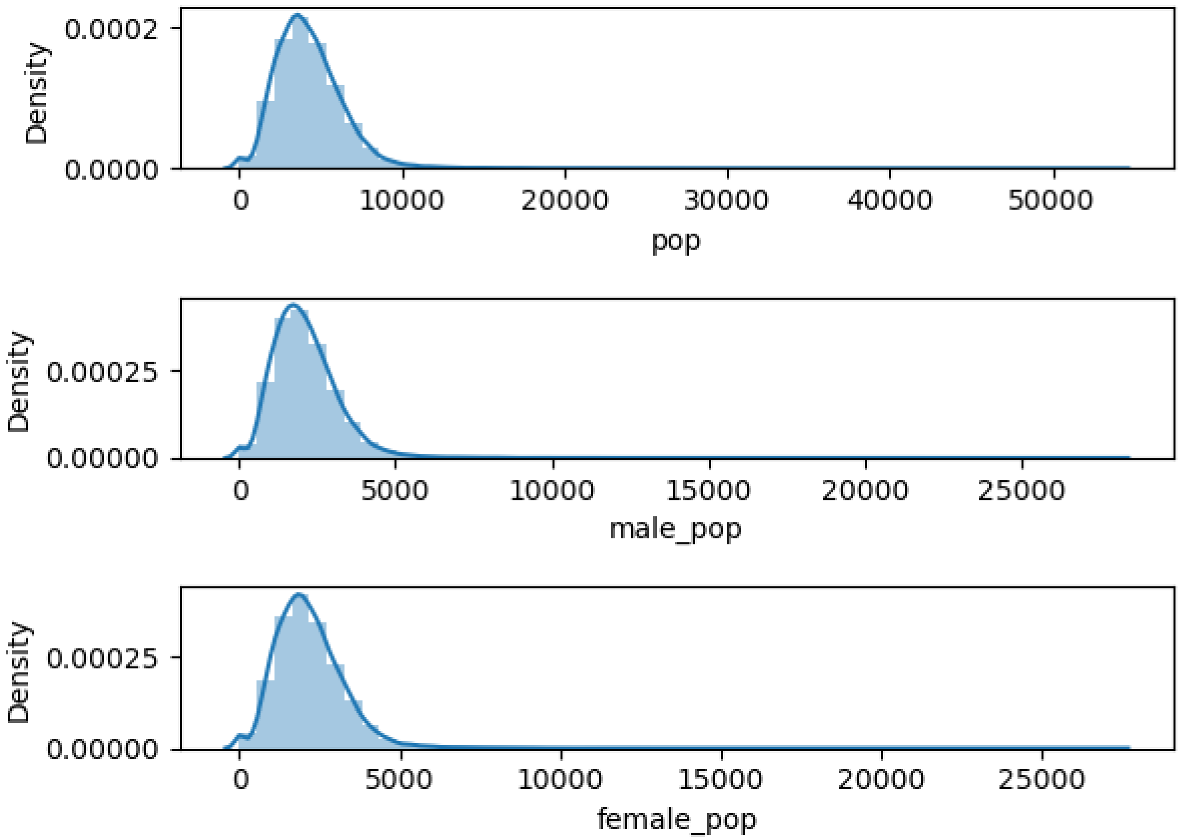
   a) Analyze the married, separated, and divorced population for these population brackets

   b) Visualize using appropriate chart type
3. Please detail your observations for rent as a percentage of income at an overall level, and for different states.
4. Perform correlation analysis for all the relevant variables by creating a heatmap. Describe your findings.

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

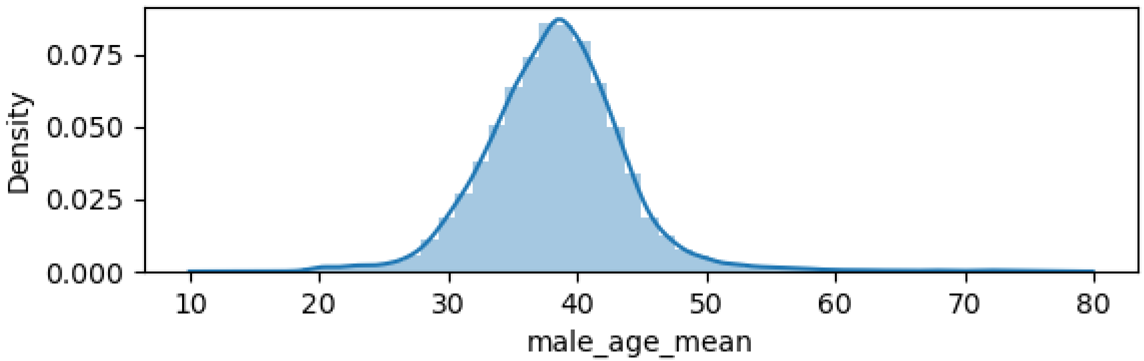
In [54]:
```
fig,(ax1,ax2,ax3)=plt.subplots(3,1)
sns.distplot(df_train['pop'],ax=ax1)
sns.distplot(df_train['male_pop'],ax=ax2)
sns.distplot(df_train['female_pop'],ax=ax3)
plt.subplots_adjust(wspace=0.8,hspace=0.8)
plt.show()
```

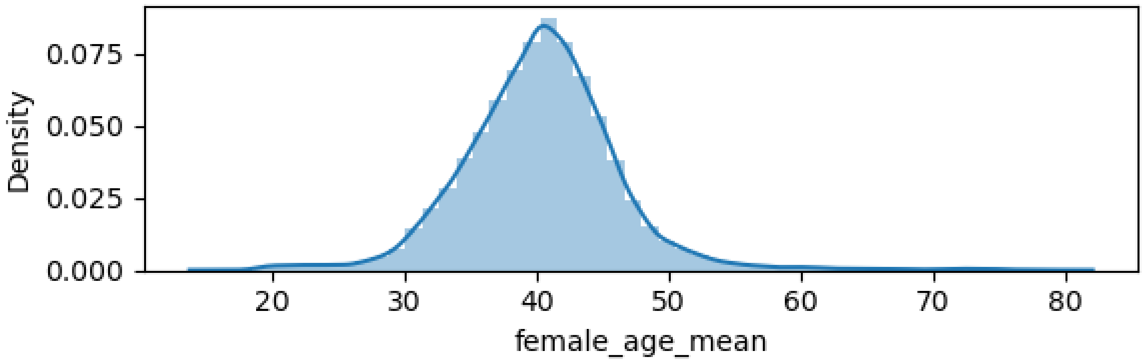
```
C:\Users\Vinosh\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning:

`distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `d
isplot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).

C:\Users\Vinosh\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning:

`distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `d
isplot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).

C:\Users\Vinosh\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning:

`distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `d
isplot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
```


Density
0.0002
0.0000
0
10000
20000
30000
40000
50000
pop
Density
0.00025
0.00000
0
5000
10000
15000
20000
25000
male_pop
Density
0.00025
0.00000
0
5000
10000
15000
20000
25000
female_pop

In [55]:
```
fig,(ax1,ax2)=plt.subplots(2,1)
sns.distplot(df_train['male_age_mean'],ax=ax1)
sns.distplot(df_train['female_age_mean'],ax=ax2)
plt.subplots_adjust(wspace=0.8,hspace=0.8)
plt.show()
```

```
C:\Users\Vinosh\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning:

`distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `d
isplot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).

C:\Users\Vinosh\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning:

`distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `d
isplot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
```





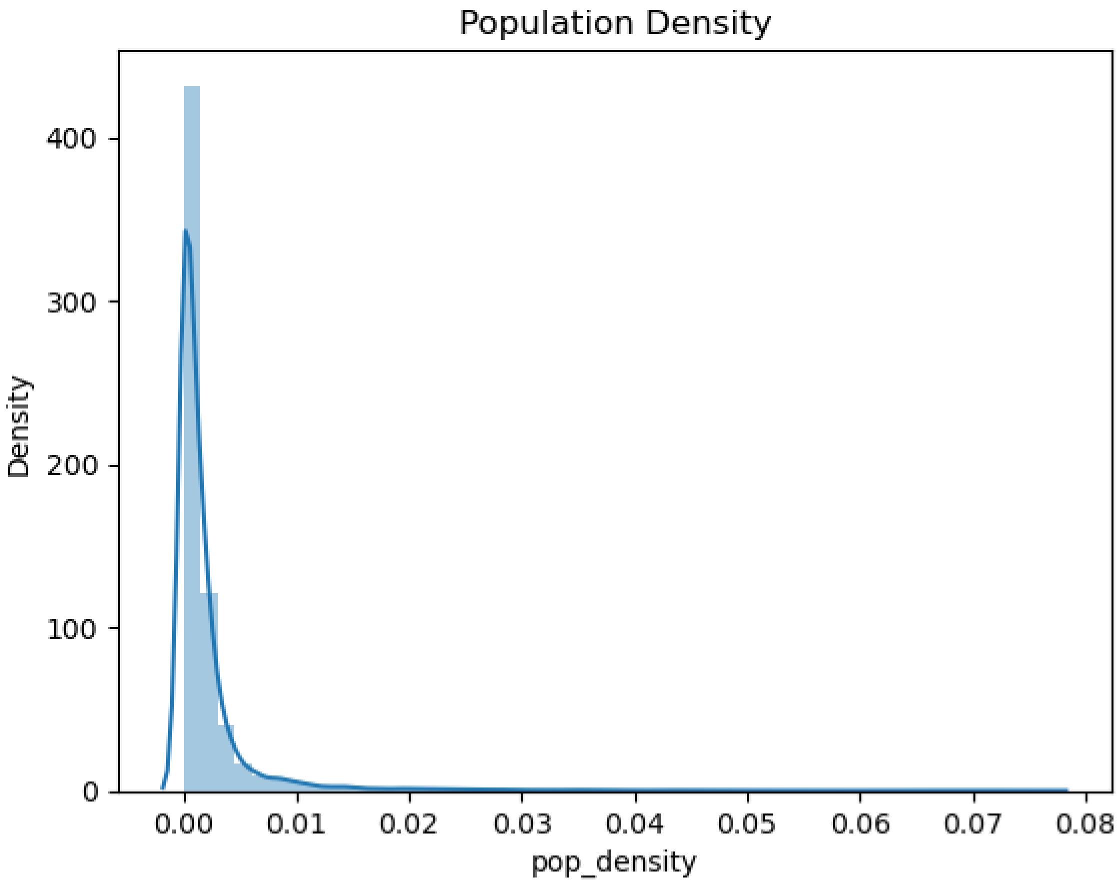
a) Use pop and ALand variables to create a new field called population density

In [56]:
```
df_train['pop_density'] = df_train['pop']/df_train['ALand']
```

In [57]:
```
df_test['pop_density'] = df_test['pop']/df_test['ALand']
```

In [58]:
```
sns.distplot(df_train['pop_density'])
plt.title('Population Density')
plt.show()
```

```
C:\Users\Vinosh\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning:

`distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `d
isplot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
```



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

In [59]:
```
df_train['age_median'] = (df_train['male_age_median']+df_train['female_age_median'])/2
```

In [60]:
```
df_test['age_median'] = (df_test['male_age_median']+df_test['female_age_median'])/2
```

In [61]:
```
df_train[['male_age_median','female_age_median','male_pop','female_pop','age_median']].head()
```

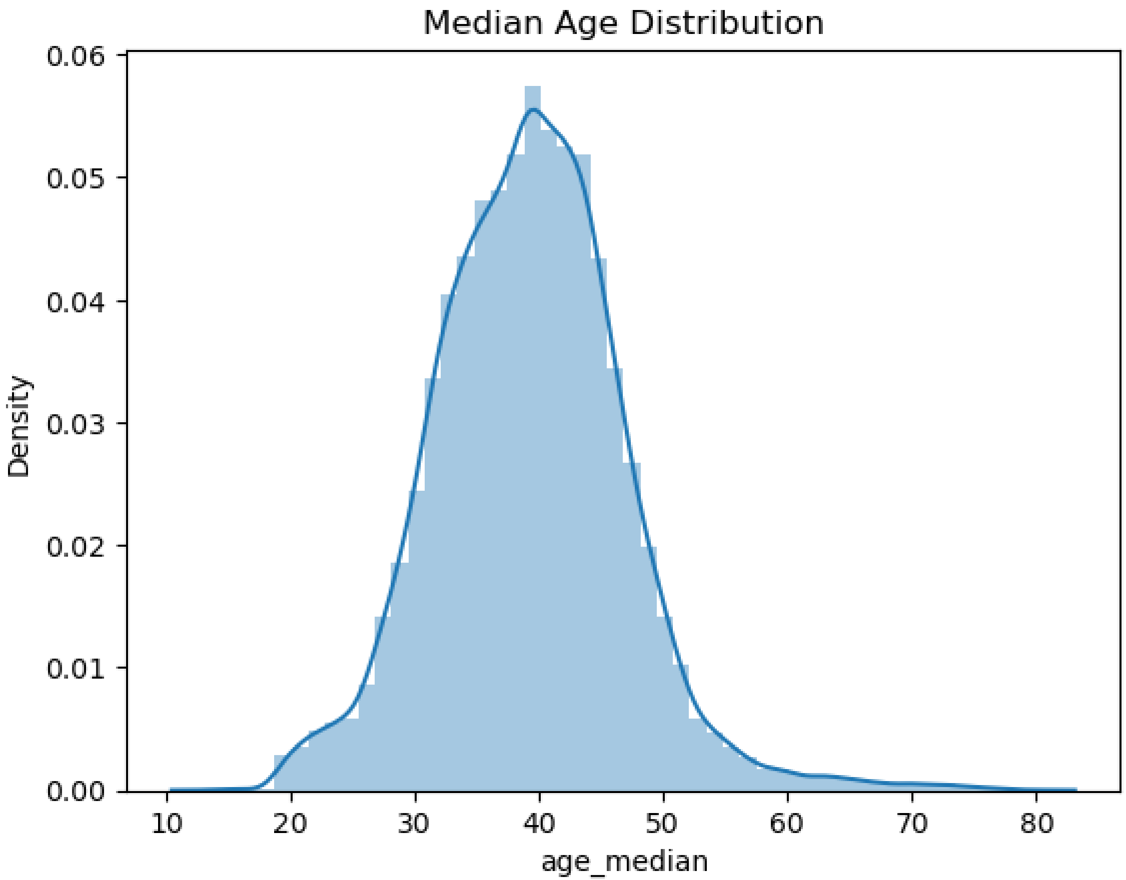
Out[61]:

	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

c) Visualize the findings using appropriate chart type

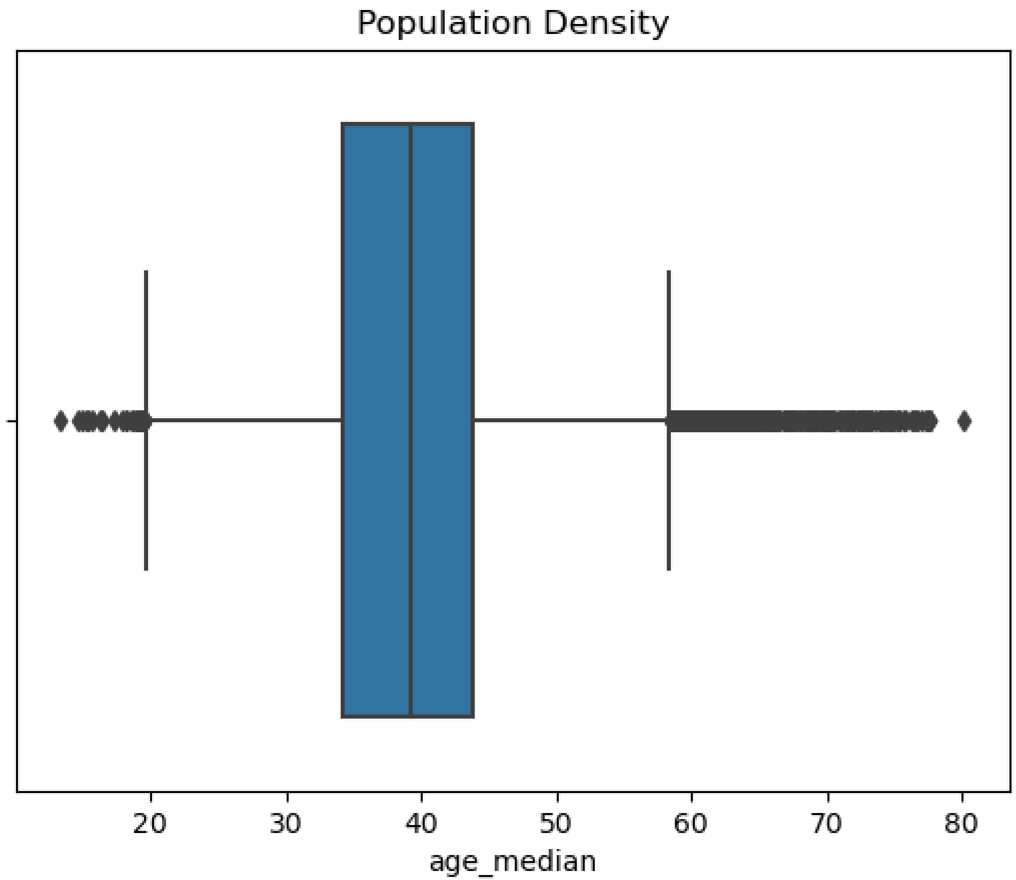
In [62]:
```
sns.distplot(df_train['age_median'])
plt.title('Median Age Distribution')
plt.show()
```

```
C:\Users\Vinosh\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning:

`distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `d
isplot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
```



In [63]:
```
sns.boxplot(df_train['age_median'])
plt.title('Population Density')
plt.show()
```

```
C:\Users\Vinosh\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning:

Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data
`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
```



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

In [64]:
```
df_train['pop'].describe()
```

Out[64]:
```
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
```

In [65]:
```
df_train['pop_bins'] = pd.cut(df_train['pop'],bins=5,labels=['very low','low','medium','high','very high'])
```

In [66]:
```
df_train[['pop','pop_bins']]
```

Out[66]:

	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 × 2 columns

In [67]:
```
df_train['pop_bins'].value_counts()
```

Out[67]:
```
very low     27058
low            246
medium           9
high             7
very high        1
Name: pop_bins, dtype: int64
```

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

In [68]:
```
df_train.groupby(by='pop_bins')[['married','separated','divorced']].count()
```

Out[68]:

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

In [69]:
```
df_train.groupby(by='pop_bins')[['married','separated','divorced']].agg(['mean','median'])
```

Out[69]:

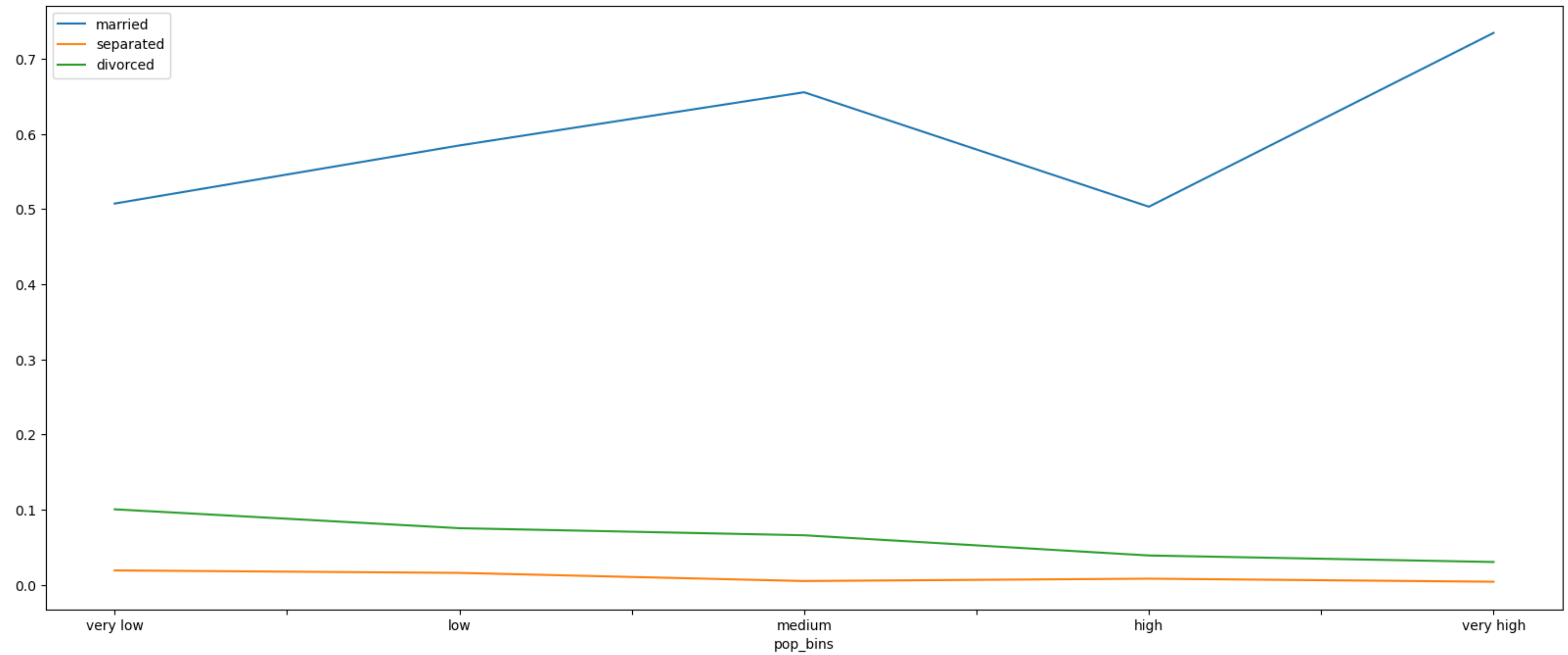
	married		separated		divorced	
	mean	median	mean	median	mean	median
pop_bins						
very low	0.507548	0.524680	0.019126	0.013650	0.100504	0.096020
low	0.584894	0.593135	0.015833	0.011195	0.075348	0.070045
medium	0.655737	0.618710	0.005003	0.004120	0.065927	0.064890
high	0.503359	0.335660	0.008141	0.002500	0.039030	0.010320
very high	0.734740	0.734740	0.004050	0.004050	0.030360	0.030360

1. Very high population group has more married people and less percantage of separated and divorced couples
2. In very low population groups, there are more divorced people

b) Visualize using appropriate chart type

In [70]:
```
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 1000x500 with 0 Axes>
```



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

In [71]:
```
rent_state_mean = df_train.groupby(by='state')['rent_mean'].agg(['mean'])
rent_state_mean.head()
```

Out[71]:

	mean
state	
Alabama	774.004927
Alaska	1185.763570
Arizona	1097.753511
Arkansas	720.918575
California	1471.133857

In [72]:
```
income_state_mean = df_train.groupby(by='state')['family_mean'].agg(['mean'])
income_state_mean.head()
```

Out[72]:

	mean
state	
Alabama	67030.064213
Alaska	92136.545109
Arizona	73328.238798
Arkansas	64765.377850
California	87655.470820

In [73]:
```
rent_perc_of_income = rent_state_mean['mean']/income_state_mean['mean']
rent_perc_of_income.head(10)
```

Out[73]:
```
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
```

In [74]:
```
sum(df_train['rent_mean'])/sum(df_train['family_mean'])
```

Out[74]:
```
0.013358170721473864
```

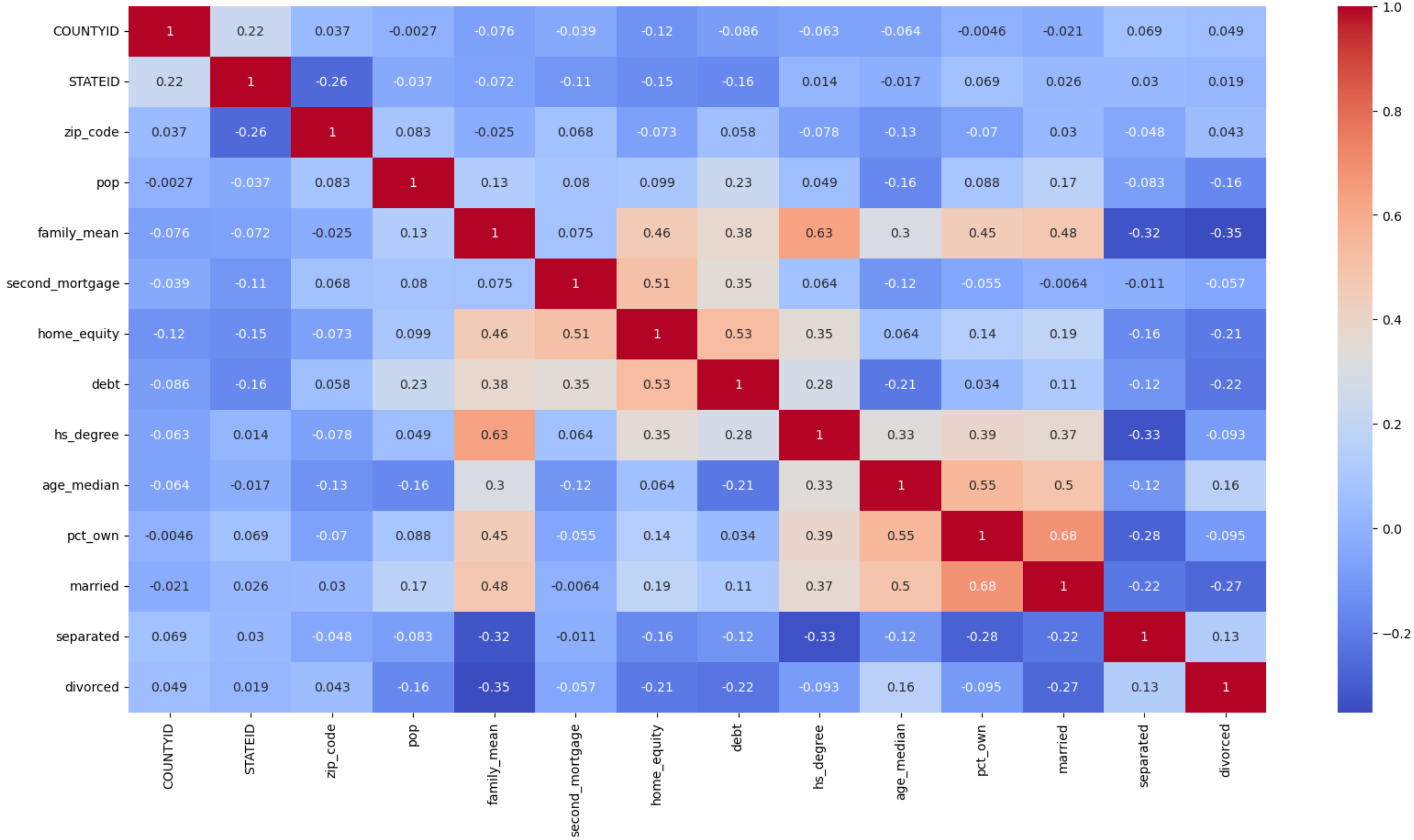
4. Perform correlation analysis for all the relevant variables by creating a heatmap. Describe your findings.

In [75]:
```
df_train.columns
```

Out[75]:
```
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')
```

In [76]:
```
cor = df_train[['COUNTYID','STATEID','zip_code','pop','family_mean','second_mortgage','home_equity','debt',
                'hs_degree','age_median','pct_own','married','separated','divorced']].corr()
```

In [77]:
```
plt.figure(figsize=(20,10))
sns.heatmap(cor,annot = True,cmap = 'coolwarm')
plt.show()
```



1. High positive correaltion is noticed between pop, male_pop and female_pop
2. High positive correaltion is noticed between rent_mean,hi_mean, family_mean,hc_mean

Project Task: Week 3

Data Pre-processing:

1. The economic multivariate data has a significant number of measured variables. The goal is to find where the measured variables depend on a number of smaller unobserved common factors or latent variables. 2. Each variable is assumed to be dependent upon a linear combination of the common factors, and the coefficients are known as loadings. Each measured variable also includes a component due to independent random variability, known as “specific variance” because it is specific to one variable. Obtain the common factors and then plot the loadings. Use factor analysis to find latent variables in our dataset and gain insight into the linear relationships in the data. Following are the list of latent variables:

• Highschool graduation rates

• Median population age

• Second mortgage statistics

• Percent own

• Bad debt expense

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

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

Out[79]:
```
array([[-1.12589165e-01,  1.95646462e-02, -2.39331065e-02,
        -6.27632576e-02,  4.23474724e-02],
       [-1.10186762e-01,  1.33506215e-02,  2.79651247e-02,
        -1.49825858e-01,  1.10838804e-01],
       [-8.28678641e-02,  5.16372369e-02, -1.36451867e-01,
        -4.98918626e-02, -1.04024839e-01],
       [ 1.80961146e-02,  1.92013750e-02,  5.81329804e-03,
         2.64842729e-02, -6.12442488e-03],
       [ 9.02324755e-02, -9.72544268e-02, -6.54601264e-02,
        -1.33145893e-01, -1.48594590e-01],
       [-1.07335681e-02, -4.12376813e-02,  1.45853484e-01,
         8.80433491e-03,  1.08227567e-01],
       [-4.28796974e-02, -2.09780212e-02,  3.66726851e-02,
        -9.45597345e-02,  5.91380498e-02],
       [-2.44243072e-03, -1.53245405e-02, -2.68300788e-03,
        -4.52473004e-02,  2.37240637e-02],
       [ 7.92164316e-02,  9.57453295e-01, -8.71151617e-02,
        -6.59924038e-03, -3.97273194e-02],
       [ 7.39808195e-02,  9.18750503e-01, -1.08834837e-01,
        -2.79371589e-02, -3.93153647e-02],
       [ 8.06598893e-02,  9.47839216e-01, -6.08006502e-02,
         1.53627080e-02, -3.86977273e-02],
       [ 7.70052110e-01,  9.84675423e-03, -3.71249731e-02,
         1.14949033e-01, -1.23784685e-01],
       [ 7.18615877e-01,  6.24980466e-03, -4.59787397e-02,
         1.09109686e-01, -1.35301910e-01],
       [ 7.07647243e-01,  2.46625402e-02, -1.00860864e-02,
         1.04472486e-01,  7.72381241e-02],
       [-1.34545492e-01,  3.36809296e-01, -4.87894961e-01,
        -4.15446193e-02,  3.17608528e-01],
       [ 2.31079707e-01,  4.37729793e-01, -6.40209212e-01,
        -2.52311017e-02,  3.47216227e-01],
       [-4.52068114e-02,  3.51263837e-02,  3.07537011e-02,
         4.44793494e-01, -1.63273406e-01],
       [-2.50717029e-02,  1.70166793e-02,  4.57227084e-02,
         6.76083841e-01, -1.55256749e-01],
       [-3.90694436e-02, -1.67460874e-02,  8.13962691e-02,
         8.36389104e-01, -9.18259801e-02],
       [-5.14161948e-02, -3.57207133e-02,  1.10795166e-01,
         9.25123723e-01, -4.44866476e-02],
       [-6.08589985e-02, -4.41860613e-02,  1.35794025e-01,
```

```
  9.53019910e-01, -2.21548664e-02],
 [-4.57771160e-02, -5.25526117e-02,  1.41019874e-01,
  9.32702625e-01, -5.86526030e-07],
 [-4.19486038e-02, -5.90387634e-02,  1.28851786e-01,
  8.87316685e-01,  1.05894282e-02],
 [-2.47894634e-02, -7.29670549e-02,  9.41510484e-02,
  7.79023672e-01,  2.95352803e-02],
 [ 2.12258447e-01,  4.65992339e-01, -6.14495943e-01,
 -2.47660016e-02,  3.66644520e-01],
 [ 2.33057238e-01,  4.47057843e-01, -6.28263418e-01,
 -2.71547710e-02,  3.43419607e-01],
 [ 7.85157086e-01,  4.91249255e-02,  1.44540484e-01,
 -2.05217626e-01, -1.54523360e-01],
 [ 7.10324884e-01,  4.99730438e-02,  1.32239991e-01,
 -2.19171864e-01, -2.10505572e-01],
 [ 8.61780938e-01,  4.35044832e-02,  1.65839096e-01,
 -1.19850813e-01,  3.16733557e-02],
 [-2.23443273e-01,  8.46259553e-01, -4.61177211e-02,
  6.88599244e-02,  2.27742316e-01],
 [ 1.43837555e-01,  9.53197411e-01,  2.27887433e-02,
 -4.57890463e-02,  1.00796445e-01],
 [ 8.30286472e-01,  3.42026001e-02,  1.61105999e-01,
 -2.04570321e-01, -7.48710515e-02],
 [ 7.94476569e-01,  2.83818596e-02,  1.51219547e-01,
 -2.07681490e-01, -9.12497107e-02],
 [ 8.11481647e-01,  4.32314892e-02,  1.43645559e-01,
 -1.07778260e-01,  5.79540142e-02],
 [-3.37741906e-01,  8.64927620e-01,  3.58933696e-02,
  9.07183945e-02,  4.46327252e-02],
 [ 5.03572666e-02,  9.35515342e-01,  1.51475401e-01,
 -2.51501271e-02, -9.34471599e-02],
 [ 9.78242236e-01, -3.31490281e-02, -1.05261173e-01,
  4.50364233e-02,  7.37362022e-02],
 [ 9.59137188e-01, -3.90023001e-02, -1.20630339e-01,
  4.52591414e-02,  6.64877229e-02],
 [ 8.14087168e-01,  2.23057306e-03,  7.66518502e-02,
  2.02747427e-02,  1.27634817e-01],
 [-4.15354001e-01,  7.18339595e-01,  3.40068079e-01,
 -7.18402753e-02, -2.77950528e-01],
 [ 7.64912732e-02,  7.24900618e-01,  2.74193199e-01,
 -4.83952665e-02, -3.52988266e-01],
 [ 9.10390865e-01, -5.36541210e-02, -4.68641894e-02,
```

```
         -7.64182289e-04,  1.63870465e-01],
        [ 8.73011863e-01, -5.30302292e-02, -5.89943137e-02,
         -1.58989763e-03,  1.52417542e-01],
        [ 7.55087660e-01, -3.56133725e-03,  5.39542547e-02,
          4.24181451e-03,  2.58043474e-01],
        [-1.23469884e-01,  6.07438127e-01,  6.33039219e-01,
         -2.14798817e-02,  2.47973916e-01],
        [-3.42866892e-01,  5.59526281e-01,  5.88213008e-01,
         -2.51533511e-02,  2.18419885e-01],
        [-1.60867218e-01, -1.53062636e-02, -1.57026580e-01,
          1.09243756e-01, -6.61660830e-01],
        [-1.37306763e-01, -2.17250667e-02, -1.58408927e-01,
          1.25156193e-01, -6.71630803e-01],
        [ 2.45096187e-01, -2.54584596e-02, -2.66691407e-02,
          9.53148453e-02, -6.42510825e-01],
        [ 2.03988663e-01,  7.85172830e-02, -3.01656217e-01,
          2.28379447e-02, -6.29223339e-01],
        [ 1.08926072e-01, -6.34332386e-02, -3.36565230e-02,
         -9.49480417e-02,  6.81473815e-01],
        [-2.63787620e-01, -6.43280757e-03, -3.58792210e-02,
         -9.37962495e-02,  6.47817018e-01],
        [-2.15717048e-01, -7.36588949e-02,  3.50113228e-01,
         -1.95201590e-02,  6.36783755e-01],
        [ 3.94306147e-01,  6.09565684e-02,  2.55337862e-01,
         -2.20362097e-01, -1.84248076e-01],
        [ 4.07877889e-01,  6.27256516e-02,  2.23926907e-01,
         -2.10028735e-01, -1.71989219e-01],
        [ 3.53156876e-01,  5.36715654e-02,  2.69603565e-01,
         -2.16933216e-01, -1.80072062e-01],
        [ 2.33537264e-01, -4.91732958e-02,  8.14450787e-01,
          9.36688926e-02,  3.27131934e-01],
        [ 2.40298202e-01, -3.38140117e-02,  8.31496951e-01,
          7.52417503e-02,  2.46323597e-01],
        [-6.71839476e-02,  6.58504524e-02,  5.86207673e-01,
          8.72955174e-02,  9.12541343e-02],
        [ 5.59835552e-02,  8.17918708e-01, -1.78458350e-01,
         -1.55949439e-02, -3.34299733e-02],
        [ 7.16426403e-02,  9.23428548e-01, -1.07142695e-01,
         -2.78635385e-02, -4.35991115e-02],
        [ 1.92496945e-01, -4.75870400e-02,  8.03173185e-01,
          1.43492708e-01,  3.33862150e-01],
        [ 1.87644433e-01, -3.29941014e-02,  8.58024490e-01,
```

```
         1.31329956e-01,  2.55679724e-01],
       [-1.02263656e-01,  6.03984253e-02,  4.72982250e-01,
         7.36848367e-02,  1.12273907e-01],
       [ 6.14776639e-02,  8.77962739e-01, -1.50410284e-01,
         2.20991026e-02, -4.17158180e-02],
       [ 7.83728211e-02,  9.54508776e-01, -5.91095904e-02,
         1.64800910e-02, -4.32590995e-02],
       [-3.24381953e-02,  1.11167165e-01,  7.84467411e-01,
        -4.37718523e-02, -2.80931237e-01],
       [ 1.76682388e-01,  1.90494238e-01,  5.61405488e-01,
        -1.20746165e-01, -1.32570786e-01],
       [-6.37386638e-02, -7.03047917e-02, -2.68934066e-01,
         1.28589791e-01,  1.88507855e-01],
       [-1.56051273e-01, -7.08033934e-02, -1.45964500e-01,
         1.24253731e-01,  1.46293109e-01],
       [-3.56716294e-01, -5.29910746e-02,  1.47771602e-01,
         2.87196184e-02,  1.13159575e-01],
       [ 2.42173831e-01, -2.86199123e-02, -3.25958340e-02,
         1.05027813e-01, -6.55406057e-01],
       [ 3.50196741e-01, -1.05016404e-02, -3.95274115e-01,
         5.92876782e-02,  2.91651780e-01],
       [ 2.25671545e-01, -3.42672754e-02,  8.92876622e-01,
         1.12426812e-01,  2.67065202e-01]])
```

Project Task: Week 4

Data Modeling :

1. Build a linear Regression model to predict the total monthly expenditure for home mortgages loan. Please refer ‘deplotment_RE.xlsx’. Column hc_mortgage_mean is predicted variable. This is the mean monthly mortgage and owner costs of specified geographical location. Note: Exclude loans from prediction model which have NaN (Not a Number) values for hc_mortgage_mean.

a) Run a model at a Nation level. If the accuracy levels and R square are not satisfactory proceed to below step.

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

c) Keep below considerations while building a linear regression model. Data Modeling :

• Variables should have significant impact on predicting Monthly mortgage and owner costs

- Utilize all predictor variable to start with initial hypothesis
- R square of 60 percent and above should be achieved
- Ensure Multi-collinearity does not exist in dependent variables
- Test if predicted variable is normally distributed

1. Build a linear Regression model to predict the total monthly expenditure for home mortgages loan. Please refer 'deplotment_RE.xlsx'. Column hc_mortgage_mean is predicted variable. This is the mean monthly mortgage and owner costs of specified geographical location. Note: Exclude loans from prediction model which have NaN (Not a Number) values for hc_mortgage_mean.

In [80]:
```
df_train.columns
```

Out[80]:
```
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')
```

In [81]:
```
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)
```

In [82]:
```
df_train['type'].unique()
```

Out[82]:
```
array([1, 2, 3, 4, 5, 6], dtype=int64)
```

In [83]:
```
df_test.replace(type_dict,inplace=True)
```

In [84]:
```
df_test['type'].unique()
```

Out[84]:
```
array([4, 1, 6, 3, 5, 2], dtype=int64)
```

In [85]:
```
feature_cols = ['COUNTYID','STATEID','zip_code','type','pop','family_mean','second_mortgage','home_equity','debt',
                'hs_degree','age_median','pct_own','married','separated','divorced']
```

In [86]:
```
x_train = df_train[feature_cols]
y_train = df_train['hc_mortgage_mean']
```

In [87]:
```
x_test = df_test[feature_cols]
y_test = df_test['hc_mortgage_mean']
```

In [88]:
```
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
```

In [89]:
```
x_train.head()
```

Out[89]:

UID	COUNTYID	STATEID	zip_code	type	pop	family_mean	second_mortgage	home_equity	debt	hs_degree	age_median	pct_own	ma
267822	53	36	13346	1	5230	67994.14790	0.02077	0.08919	0.52963	0.89288	44.666665	0.79046	0.5
246444	141	18	46616	1	2633	50670.10337	0.02222	0.04274	0.60855	0.90487	34.791665	0.52483	0.3
245683	63	18	46122	1	6881	95262.51431	0.00000	0.09512	0.73484	0.94288	41.833330	0.85331	0.6
279653	127	72	927	2	2700	56401.68133	0.01086	0.01086	0.52714	0.91500	49.750000	0.65037	0.4
247218	161	20	66502	1	5637	54053.42396	0.05426	0.05426	0.51938	1.00000	22.000000	0.13046	0.1

In [90]:
```
sc = StandardScaler()
x_train_scaled = sc.fit_transform(x_train)
x_test_scaled = sc.fit_transform(x_test)
```

a) Run a model at a Nation level. If the accuracy levels and R square are not satisfactory proceed to below step.

In [91]:
```
linreg = LinearRegression()
```

In [92]:
```
linreg.fit(x_train_scaled,y_train)
```

Out[92]:
```
LinearRegression()
```

In [93]:
```
y_pred = linreg.predict(x_test_scaled)
```

In [94]:
```
print('overall R2 score of linear regression model',r2_score(y_test,y_pred))
```

```
overall R2 score of linear regression model 0.7348210754610929
```

In [95]:
```
print('Overall RMSE of linear regression model',np.sqrt(mean_squared_error(y_test,y_pred)))
```

```
Overall RMSE of linear regression model 323.1018894984635
```

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

In [96]:
```
state = df_train['STATEID'].unique()
```

In [97]:
```
state[0:5]
```

Out[97]:
```
array([36, 18, 72, 20,  1], dtype=int64)
```

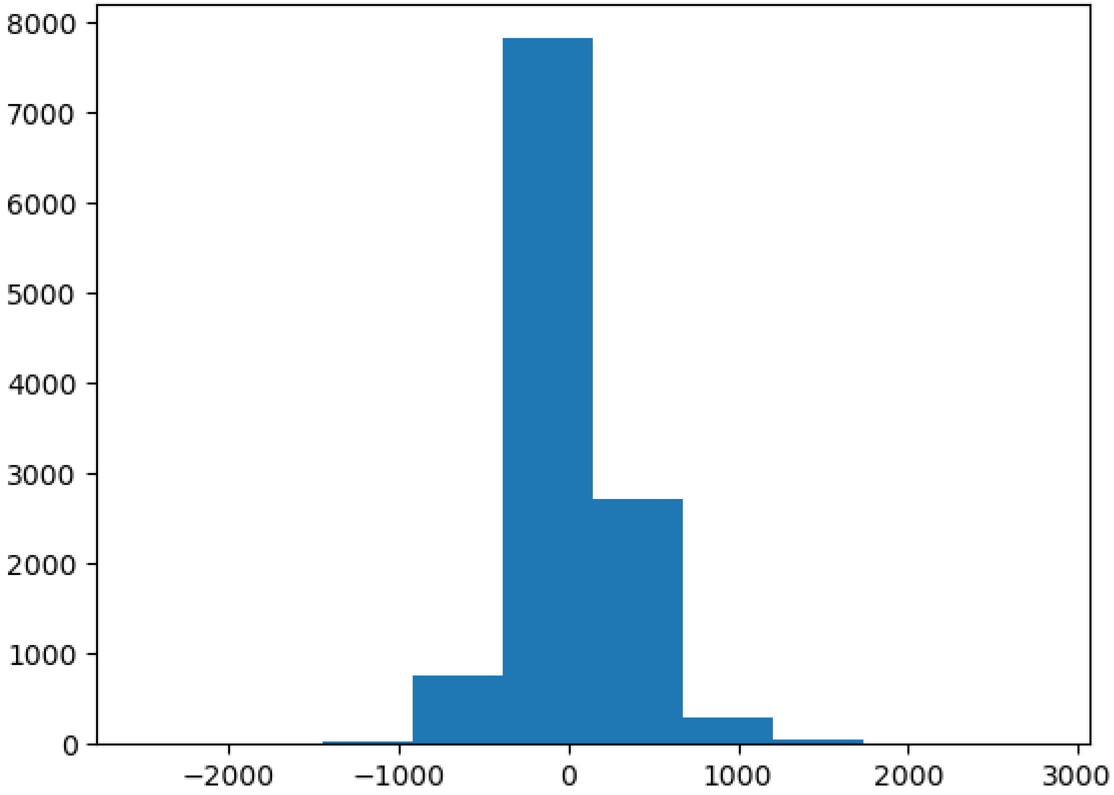
In [98]:
```
for i in [20,1,45]:
    print('State ID-',i)

    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)

    linreg.fit(x_train_scaled_nation,y_train_nation)
    y_pred_nation = linreg.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_
    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 :- 307.97188999314716


State ID- 1
Overall R2 score of Linear Regression model for state, 1 :- 0.8104382475484616
Overall RMSE of Linear Regression model for state, 1 :- 307.82758618484354


State ID- 45
Overall R2 score of Linear Regression model for state, 45 :- 0.7887446497855253
Overall RMSE of Linear Regression model for state, 45 :- 225.69615420724134
```

In [99]:
```
residuals = y_test-y_pred
residuals
```

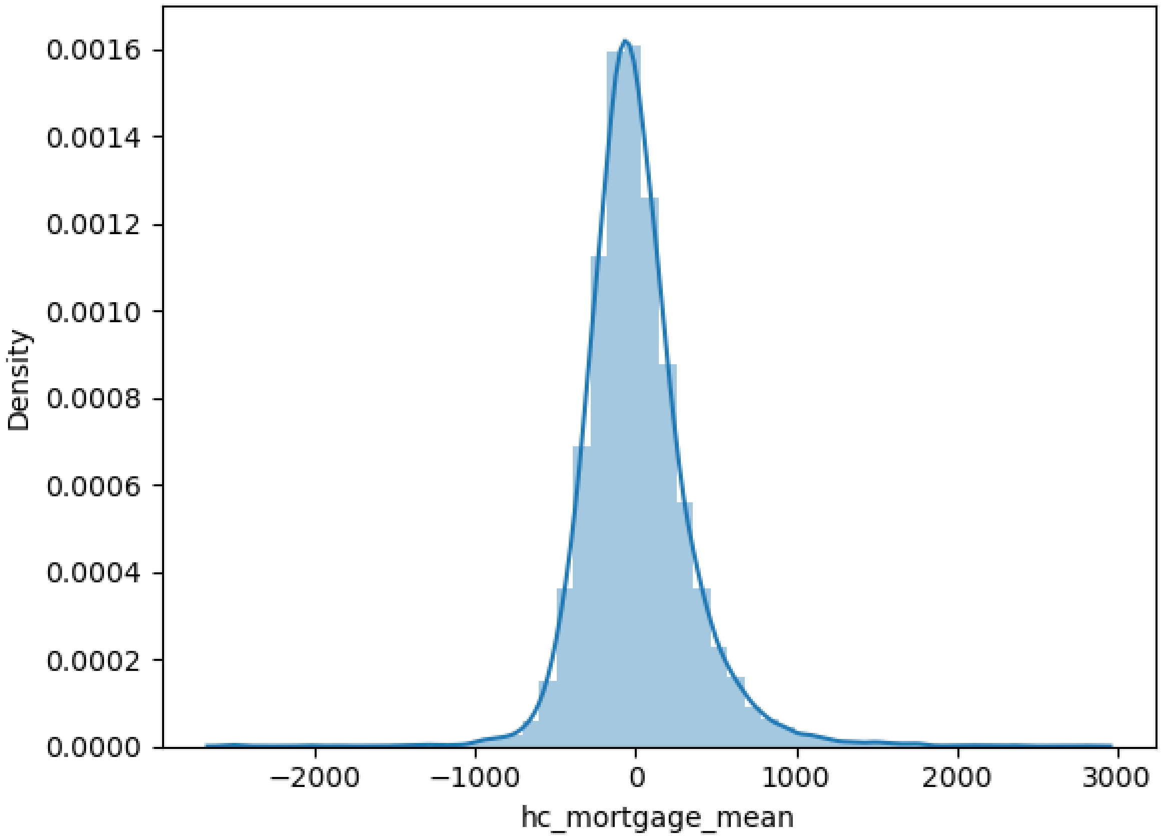
Out[99]:
```
UID
255504    281.969088
252676    -69.935775
276314    190.761969
248614   -157.290627
286865     -9.887017
             ...
238088    -67.541646
242811    -41.578757
250127   -127.427569
241096   -330.820475
287763    217.760642
Name: hc_mortgage_mean, Length: 11709, dtype: float64
```

In [100]:
```
plt.hist(residuals)
plt.show()
```



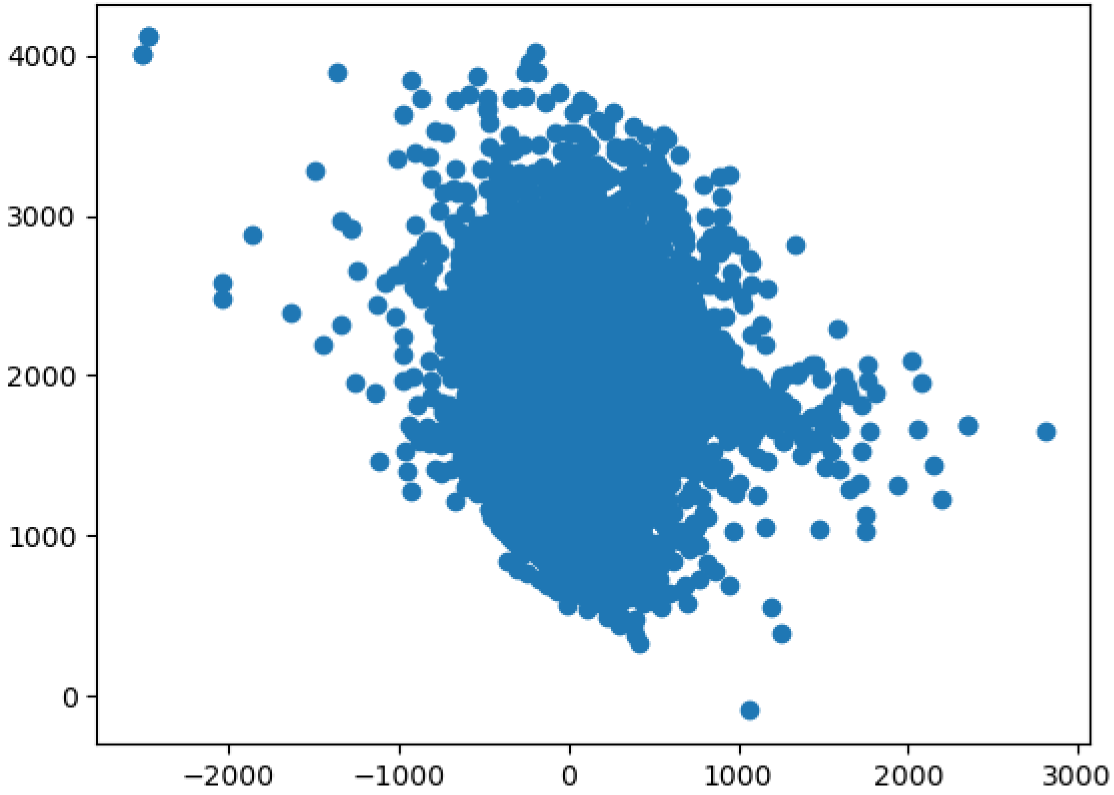
In [101]:
```
sns.distplot(residuals)
plt.show()
```

```
C:\Users\Vinosh\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning:

`distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `d
isplot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
```



The residuals is Normally Distributed

In [102]:
```
plt.scatter(residuals,y_pred)
plt.show()
```



In [ ]:
```
```