Mod 11 Live Session

This is the notebook - it starts with the Mod 10 notebook and then picks up with Mod 11 work

H. Diana McSpadden (hdm5s)

In [ ]:
```
import numpy as np
import pandas as pd
import prince
from scipy import stats
import os
#from pandas_profiling import ProfileReport
#os.chdir("/Users/jk8sd/Downloads")
```

In [ ]:
```
ahs = pd.read_csv('ahs_cleaned-1.csv', na_values=[-6, "'-9'"])
```

In [ ]:
```
ahs.columns
```

Out[ ]:
```
Index(['Unnamed: 0', 'DIVISION', 'TENURE', 'YRBUILT', 'UNITSIZE', 'HSHLDTYPE',
       'HHRACE', 'HHSEX', 'HINCP', 'TOTHCAMT', 'MARKETVAL', 'MAINTAMT',
       'FUSEBLOW', 'SEWBREAK', 'ROACH', 'RODENT', 'NOWIRE', 'PLUGS', 'COLD',
       'NOTOIL', 'NOWAT', 'FLOORHOLE', 'FNDCRUMB', 'PAINTPEEL', 'ROOFHOLE',
       'ROOFSAG', 'ROOFSHIN', 'WALLCRACK', 'WALLSIDE', 'WALLSLOPE', 'WINBOARD',
       'WINBROKE', 'LEAKI', 'MOLDBATH'],
      dtype='object')
```

In [ ]:
```
#profile = ProfileReport(ahs,
#                         title = 'American Housing Survey EDA',
#                         html = {'style': {'full_width': True}},
#                         minimal = False)
#profile.to_notebook_iframe()
```

In [ ]:
```
ahs['HINCP']
```

Out[ ]:
```
0        257000.0
1        201000.0
2             NaN
3         66900.0
4         35000.0
           ...
63180     74000.0
63181    207000.0
63182    158100.0
63183    130200.0
63184    120000.0
Name: HINCP, Length: 63185, dtype: float64
```

In [ ]:
```
ahs['RODENT'].value_counts()
```

Out[ ]:
```
No signs in the last 12 months            48821
Seen a few times in the last 12 months     4212
Seen monthly in the last 12 months          522
Seen daily in the last 12 months            474
Seen weekly in the last 12 months           426
Name: RODENT, dtype: int64
```

In [ ]:
```
ahs.groupby("RODENT").agg({'HINCP':'mean'})
```

Out[ ]:

	HINCP
RODENT	
No signs in the last 12 months	87738.246779
Seen a few times in the last 12 months	86156.387464
Seen daily in the last 12 months	51274.924051
Seen monthly in the last 12 months	82798.544061
Seen weekly in the last 12 months	64086.826291

In [ ]:
```
stats.f_oneway(ahs.query("RODENT=='No signs in the last 12 months'").HINCP.dropna(),
               ahs.query("RODENT=='Seen a few times in the last 12 months'").HINCP.drop
               ahs.query("RODENT=='Seen daily in the last 12 months'").HINCP.dropna(),
               ahs.query("RODENT=='Seen monthly in the last 12 months'").HINCP.dropna()
               ahs.query("RODENT=='Seen weekly in the last 12 months'").HINCP.dropna())
```

Out[ ]:
```
F_onewayResult(statistic=21.68467615110672, pvalue=6.703833330074091e-18)
```

In [ ]:
```
ahs['YRBUILT'].value_counts()
```

Out[ ]:
```
1970    9313
1980    9072
2000    8883
1990    7863
1960    6860
1950    6330
1919    3594
1940    3001
1920    2494
1930    1699
2010     846
2017     503
2016     489
2015     486
2014     444
2013     340
2012     320
2018     269
2011     251
2019     128
Name: YRBUILT, dtype: int64
```

In [ ]:
```
ahs['MARKETVAL'].describe()
```

Out[ ]:
```
count    3.839000e+04
mean     3.762769e+05
std      5.537866e+05
min      1.000000e+03
25%      1.404465e+05
50%      2.552730e+05
75%      4.359682e+05
max      9.999998e+06
Name: MARKETVAL, dtype: float64
```

In [ ]:
```
ahs[['MARKETVAL', 'YRBUILT']].corr()
```

Out[ ]:

	MARKETVAL	YRBUILT
MARKETVAL	1.00000	-0.00403
YRBUILT	-0.00403	1.00000

In [ ]:
```
ahs2 = ahs[['MARKETVAL', 'YRBUILT']].dropna()
stats.pearsonr(ahs2['MARKETVAL'], ahs2['YRBUILT'])
```

Out[ ]:
```
(-0.004029500232993765, 0.4298243664197942)
```

In [ ]:
```
ahs.columns
```

Out[ ]:
```
Index(['Unnamed: 0', 'DIVISION', 'TENURE', 'YRBUILT', 'UNITSIZE', 'HSHLDTYPE',
       'HHRACE', 'HHSEX', 'HINCP', 'TOTHCAMT', 'MARKETVAL', 'MAINTAMT',
       'FUSEBLOW', 'SEWBREAK', 'ROACH', 'RODENT', 'NOWIRE', 'PLUGS', 'COLD',
       'NOTOIL', 'NOWAT', 'FLOORHOLE', 'FNDCRUMB', 'PAINTPEEL', 'ROOFHOLE',
       'ROOFSAG', 'ROOFSHIN', 'WALLCRACK', 'WALLSIDE', 'WALLSLOPE', 'WINBOARD',
       'WINBROKE', 'LEAKI', 'MOLDBATH'],
      dtype='object')
```

In [ ]:
```
broken = ahs[['FUSEBLOW', 'SEWBREAK', 'ROACH', 'RODENT', 'NOWIRE', 'PLUGS', 'COLD',
       'NOTOIL', 'NOWAT', 'FLOORHOLE', 'FNDCRUMB', 'PAINTPEEL', 'ROOFHOLE',
       'ROOFSAG', 'ROOFSHIN', 'WALLCRACK', 'WALLSIDE', 'WALLSLOPE', 'WINBOARD',
       'WINBROKE', 'LEAKI', 'MOLDBATH']].dropna()
```

In [ ]:
```
MCA = prince.MCA(n_components=2)
MCA = MCA.fit(broken)
```

In [ ]:
```
pd.set_option('display.max_rows', 100)
MCA.column_coordinates(broken).sort_values(1)
```

Out[ ]:

	0	1
WALLSLOPE_Broken	4.859652	-2.910609
ROOFSAG_Broken	3.902624	-2.223199
ROOFHOLE_Broken	4.146945	-2.002111
ROOFSHIN_Broken	2.383514	-1.479640
WALLSIDE_Broken	3.081924	-1.271913
WINBOARD_Broken	3.453403	-1.062335
FNDCRUMB_Broken	1.821293	-0.457588
WINBROKE_Broken	2.177870	-0.445659
FLOORHOLE_Broken	4.284032	-0.297980
RODENT_Seen monthly in the last 12 months	1.356874	-0.139335
LEAKI_Not broken	-0.080362	-0.092084
FUSEBLOW_No fuses / breakers blown in the last 3 months	-0.071365	-0.070827
RODENT_No signs in the last 12 months	-0.115876	-0.058155
COLD_Not broken	-0.083473	-0.055314
NOTOIL_Not broken	-0.024595	-0.054145
NOWAT_Not broken	-0.022932	-0.050263
PAINTPEEL_Broken	3.404738	-0.049909
SEWBREAK_No breakdowns in the last 3 months	-0.024639	-0.047818
ROACH_No signs in the last 12 months	-0.080402	-0.037409
MOLDBATH_Not broken	-0.032766	-0.015206
WALLCRACK_Not broken	-0.109388	-0.010125
NOWIRE_Not broken	-0.017618	-0.007298
PLUGS_Not broken	-0.018019	-0.006925
PAINTPEEL_Not broken	-0.059399	0.000871
FLOORHOLE_Not broken	-0.042145	0.002931
WINBOARD_Not broken	-0.031694	0.009750
WINBROKE_Not broken	-0.076222	0.015597
FNDCRUMB_Not broken	-0.092293	0.023188
ROOFHOLE_Not broken	-0.054294	0.026213
WALLSLOPE_Not broken	-0.046273	0.027715
WALLSIDE_Not broken	-0.072108	0.029759
ROOFSAG_Not broken	-0.060365	0.034388
ROOFSHIN_Not broken	-0.076179	0.047291

	0	1
ROACH_Seen monthly in the last 12 months	0.917764	0.120397
ROACH_Seen daily in the last 12 months	2.536481	0.151762
WALLCRACK_Broken	2.236705	0.207032
NOWIRE_Broken	0.685449	0.283946
ROACH_Seen a few times in the last 12 months	0.343755	0.314518
RODENT_Seen a few times in the last 12 months	0.580372	0.367878
PLUGS_Broken	1.027688	0.394940
FUSEBLOW_1 fuse / breaker blown in the last 3 months	0.660564	0.637415
FUSEBLOW_2 fuses / breakers blown in the last 3 months	0.752397	0.720868
COLD_Broken	1.363430	0.903488
ROACH_Seen weekly in the last 12 months	1.212532	0.975465
FUSEBLOW_4 or more fuses / breakers blown in the last 3 months	1.584148	1.262412
RODENT_Seen daily in the last 12 months	3.184792	1.281264
LEAKI_Broken	1.118387	1.281520
RODENT_Seen weekly in the last 12 months	1.927030	1.498603
MOLDBATH_Broken	3.340377	1.550195
FUSEBLOW_3 fuses / breakers blown in the last 3 months	1.296649	1.881418
SEWBREAK_Two breakdowns in the last 3 months for 6 hours or more	3.190841	2.106429
NOWAT_Broken	1.089745	2.388503
SEWBREAK_Sewage system broke down in the last 3 months, but never for 6 hours or more	0.679774	2.653595
SEWBREAK_Four or more breakdowns in last three months for 6 hours or more	5.443641	3.215459
NOTOIL_Broken	2.186892	4.814384
SEWBREAK_One breakdown in the last 3 months for 6 hours or more	1.532865	5.588409
SEWBREAK_Three breakdowns in the last 3 months for 6 hours or more	4.994376	7.737863

In [ ]:
```
broken_house_index = MCA.row_coordinates(broken)
ahs_broken = ahs.join(broken_house_index) # add the feature to the original dataset
```

In [ ]:
```
ahs_broken.sort_values(0, ascending=False)
```

Out[ ]:

	Unnamed: 0	DIVISION	TENURE	YRBUILT	UNITSIZE	HSHLDTYPE	HHRACE	HHSEX	HINCP
5557	5557	West South Central	Owned or being bought by someone in your house...	1970	750 to 999 square feet	Married-couple family household	White only	Male	90000.0
60404	60404	South Atlantic	Occupied without payment of rent	1920	NaN	Nonfamily household	White only	Female	0.0
44482	44482	Pacific	Rented	1950	500 to 749 square feet	Other family household	Black only	Female	17400.0
34830	34830	West South Central	Owned or being bought by someone in your house...	1980	1,000 to 1,499 square feet	Nonfamily household	Black only	Female	32600.0
42001	42001	East South Central	Rented	1950	1,000 to 1,499 square feet	Other family household	Black only	Female	9600.0
...	...	...	...	...	...	...	...	...	...
63171	63171	South Atlantic	Rented	2015	1,000 to 1,499 square feet	Nonfamily household	White only	Male	144000.0
63175	63175	West South Central	Rented	2016	750 to 999 square feet	Other family household	Black only	Female	126000.0
63176	63176	Middle Atlantic	NaN	2000	750 to 999 square feet	NaN	NaN	NaN	NaN
63177	63177	South Atlantic	Rented	2017	1,500 to 1,999 square feet	Nonfamily household	White only	Male	125000.0
63179	63179	Middle Atlantic	NaN	1970	NaN	NaN	NaN	NaN	NaN

63185 rows × 36 columns

This captures us up to where we were last week

Now on to the new stuff

Let's plot the new broken-ness index ...

Start with loading libraries

In [ ]:
```
from matplotlib import pyplot as plt
import seaborn as sns
```

Goals:

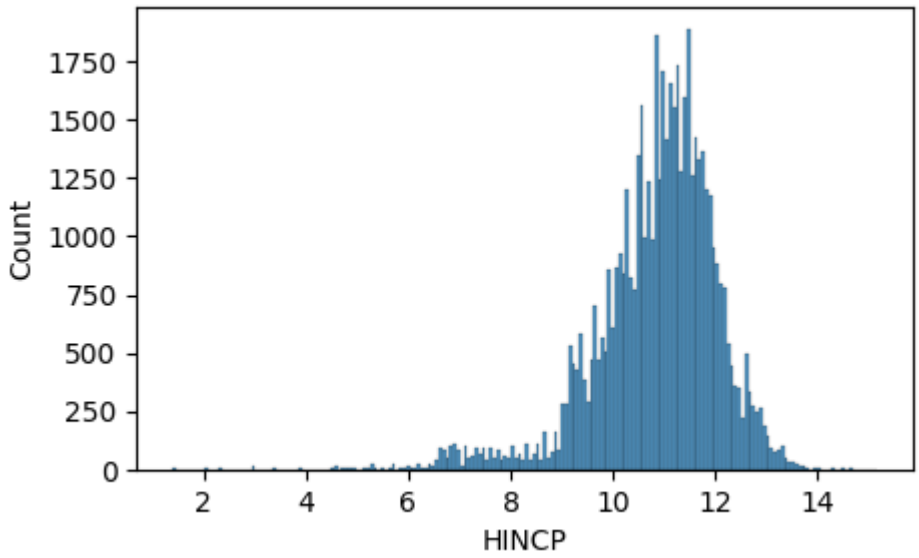
- The distributions of income, race, ownership, and housing expenses
- The relationship between these four features with each other, as well as the two indices of home disrepair that we built last week
- The time progression of average house price against year built
- Graph matrices in which each cell contains a graph that is specific to a census division or ownership status

First, distributions of income, race, ownership and housing expenses:

In [ ]:
```
# income is HINCP - continuous valued feature - use histogram
# set the plot size
plt.figure(figsize=(5, 3))
sns.histplot(np.log(ahs_broken['HINCP']), kde=False)
```

```
c:\Users\dianam\Anaconda3\envs\ds6001_mod10\lib\site-packages\pandas\core\arraylike.p
y:402: RuntimeWarning: divide by zero encountered in log
  result = getattr(ufunc, method)(*inputs, **kwargs)
c:\Users\dianam\Anaconda3\envs\ds6001_mod10\lib\site-packages\pandas\core\arraylike.p
y:402: RuntimeWarning: invalid value encountered in log
  result = getattr(ufunc, method)(*inputs, **kwargs)
```

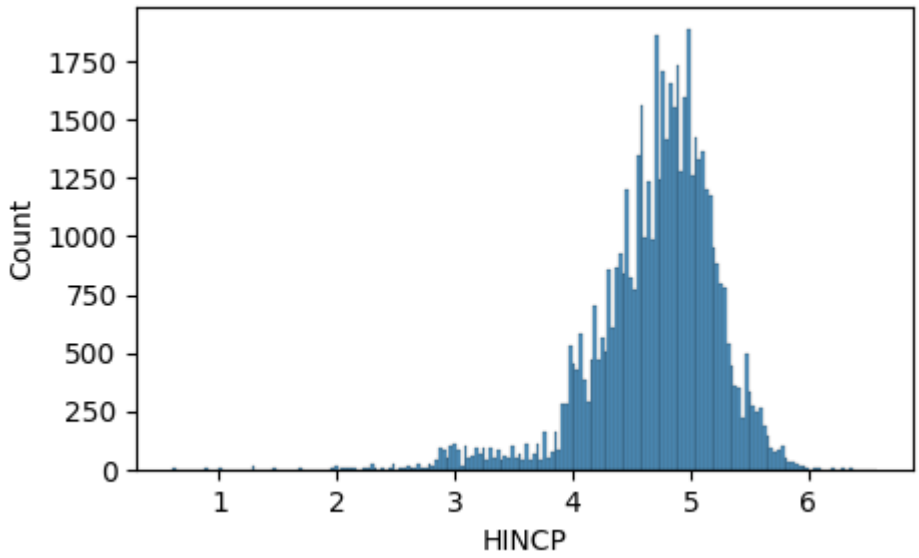
Out[ ]:
```
<AxesSubplot: xlabel='HINCP', ylabel='Count'>
```



In [ ]:
```
plt.figure(figsize=(5, 3))
sns.histplot(np.log(ahs_broken['HINCP'])/np.log(10), kde=False)
```

```
c:\Users\dianam\Anaconda3\envs\ds6001_mod10\lib\site-packages\pandas\core\arraylike.p
y:402: RuntimeWarning: divide by zero encountered in log
  result = getattr(ufunc, method)(*inputs, **kwargs)
c:\Users\dianam\Anaconda3\envs\ds6001_mod10\lib\site-packages\pandas\core\arraylike.p
y:402: RuntimeWarning: invalid value encountered in log
  result = getattr(ufunc, method)(*inputs, **kwargs)
```

Out[ ]:
```
<AxesSubplot: xlabel='HINCP', ylabel='Count'>
```

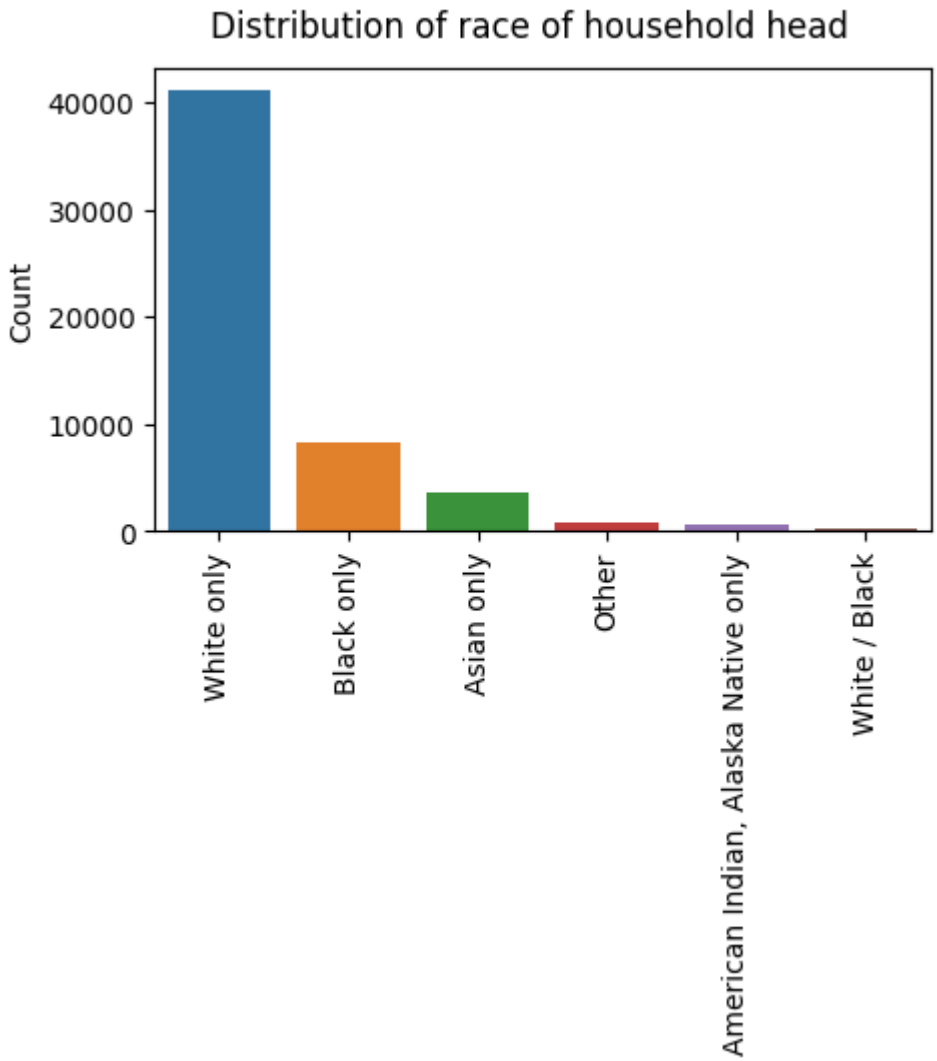


In [ ]:
```
# now race
ahs['HHRACE'].value_counts()
```

Out[ ]:
```
White only                              41116
Black only                               8215
Asian only                               3589
Other                                     751
American Indian, Alaska Native only       603
White / Black                             181
Name: HHRACE, dtype: int64
```

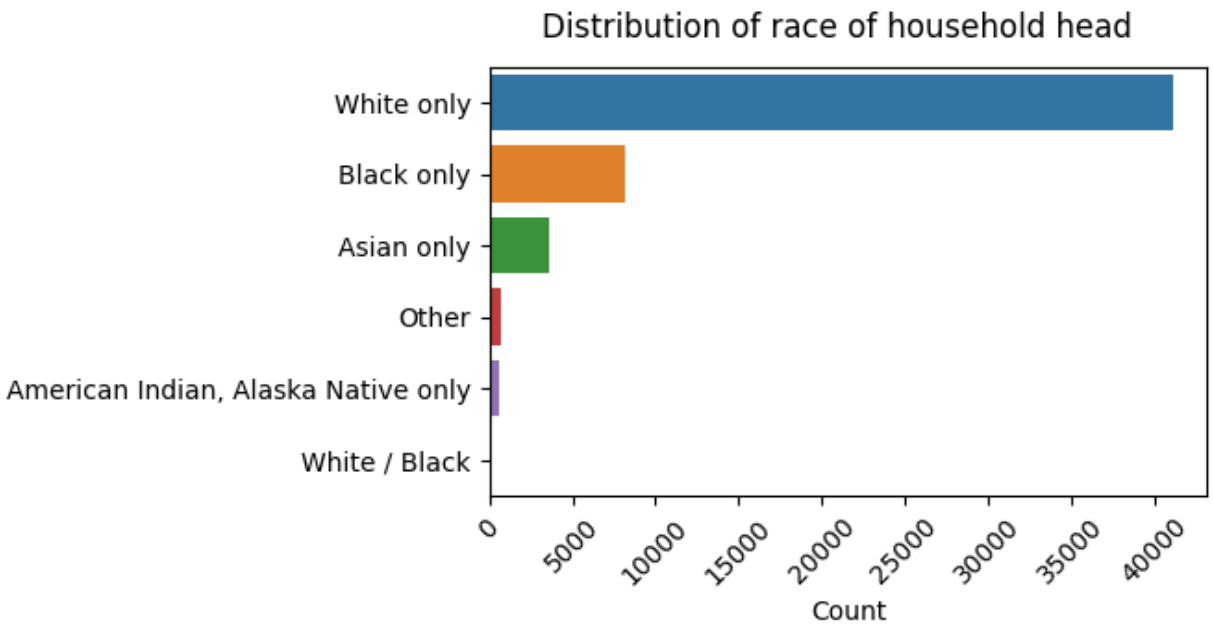
In [ ]:
```
# categorical feature - use bar chart
plt.figure(figsize=(5, 3))
sns.barplot(x=ahs['HHRACE'].value_counts().index, y=ahs['HHRACE'].value_counts())
# shift the x labels 90 degrees
plt.xticks(rotation=90)
# change the x axis label
plt.ylabel('Count')
plt.suptitle('Distribution of race of household head')
```

Out[ ]:
```
Text(0.5, 0.98, 'Distribution of race of household head')
```



In [ ]:
```
# make horizontal seaborn bar chart
plt.figure(figsize=(5, 3))
sns.barplot(y=ahs['HHRACE'].value_counts().index, x=ahs['HHRACE'].value_counts())
# shift the x labels 45 degrees
plt.xticks(rotation=45)
plt.xlabel('Count')
plt.suptitle('Distribution of race of household head')
```

Out[ ]:
```
Text(0.5, 0.98, 'Distribution of race of household head')
```



Housing expenses

In [ ]:
```
ahs.groupby('TENURE').agg({'TOTHCAMT':'mean'})
```

Out[ ]:

TENURE	TOTHCAMT
Occupied without payment of rent	199.145946
Owned or being bought by someone in your household	1666.237444
Rented	1328.444487

If you own you have higher expenses than if you are renting

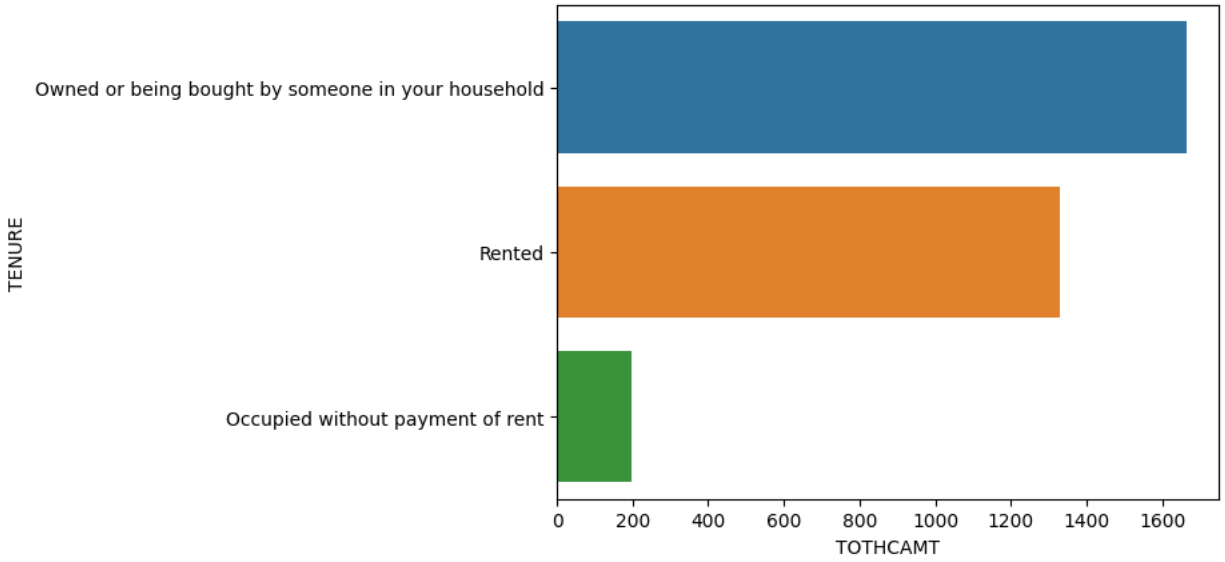
In [ ]:
```
ahs_means = ahs.groupby('TENURE').agg({'TOTHCAMT':'mean'}).reset_index().sort_values('
ahs_means
```

Out[ ]:

	TENURE	TOTHCAMT
1	Owned or being bought by someone in your house...	1666.237444
2	Rented	1328.444487
0	Occupied without payment of rent	199.145946

In [ ]:
```
sns.barplot(x="TOTHCAMT", y="TENURE", data=ahs_means)
```

Out[ ]:
```
<AxesSubplot: xlabel='TOTHCAMT', ylabel='TENURE'>
```



Indices of home disrepair

In [ ]:
```
ahs_broken.head(1).T
```

Out[ ]:

	0
Unnamed: 0	0
DIVISION	South Atlantic
TENURE	Owned or being bought by someone in your house...
YRBUILT	2000
UNITSIZE	2,000 to 2,499 square feet
HSHLDTYPE	Married-couple family household
HHRACE	White only
HHSEX	Male
HINCP	257000.0
TOTHCAMT	1642.0
MARKETVAL	280249.0
MAINTAMT	1022.0
FUSEBLOW	No fuses / breakers blown in the last 3 months
SEWBREAK	No breakdowns in the last 3 months
ROACH	No signs in the last 12 months
RODENT	No signs in the last 12 months
NOWIRE	Not broken
PLUGS	Not broken
COLD	Not broken
NOTOIL	Not broken
NOWAT	Not broken
FLOORHOLE	Not broken
FNDCRUMB	Not broken
PAINTPEEL	Not broken
ROOFHOLE	Not broken
ROOFSAG	Not broken
ROOFSHIN	Not broken
WALLCRACK	Not broken
WALLSIDE	Not broken
WALLSLOPE	Not broken
WINBOARD	Not broken
WINBROKE	Not broken
LEAKI	Not broken

	0
MOLDBATH	Not broken
0	-0.159316
1	-0.053928

In [ ]:
```
ahs_broken = ahs_broken.rename(columns={0:'disrepair',1:'structure_vs_utils'})
ahs_broken.head(1).T
```

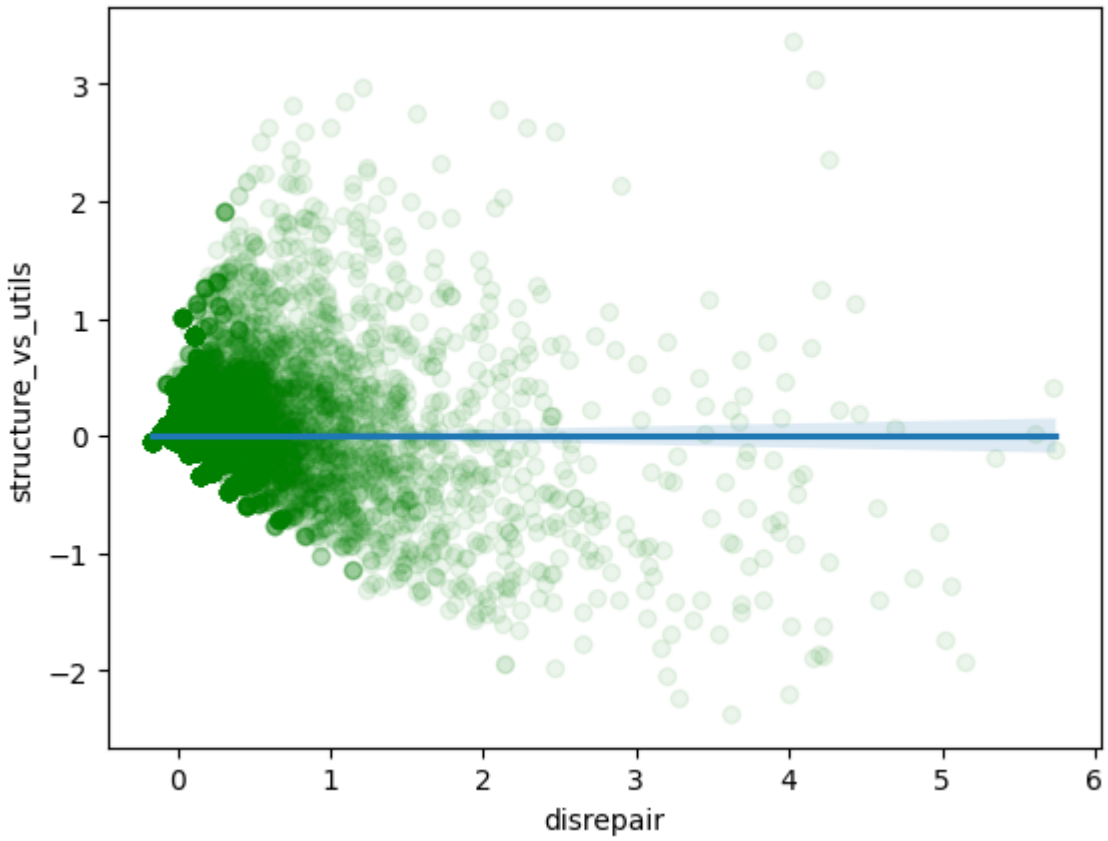
Out[ ]:

	0
Unnamed: 0	0
DIVISION	South Atlantic
TENURE	Owned or being bought by someone in your house...
YRBUILT	2000
UNITSIZE	2,000 to 2,499 square feet
HSHLDTYPE	Married-couple family household
HHRACE	White only
HHSEX	Male
HINCP	257000.0
TOTHCAMT	1642.0
MARKETVAL	280249.0
MAINTAMT	1022.0
FUSEBLOW	No fuses / breakers blown in the last 3 months
SEWBREAK	No breakdowns in the last 3 months
ROACH	No signs in the last 12 months
RODENT	No signs in the last 12 months
NOWIRE	Not broken
PLUGS	Not broken
COLD	Not broken
NOTOIL	Not broken
NOWAT	Not broken
FLOORHOLE	Not broken
FNDCRUMB	Not broken
PAINTPEEL	Not broken
ROOFHOLE	Not broken
ROOFSAG	Not broken
ROOFSHIN	Not broken
WALLCRACK	Not broken
WALLSIDE	Not broken
WALLSLOPE	Not broken
WINBOARD	Not broken
WINBROKE	Not broken
LEAKI	Not broken

	0
MOLDBATH	Not broken
disrepair	-0.159316
structure_vs_utils	-0.053928

In [ ]:
```
# do a scatter plot of the two MCA features (these will show two uncorrelated features
sns.regplot(x='disrepair', y='structure_vs_utils', data=ahs_broken, scatter_kws = {'co
# position the legend outside the plot
#plt.legend(bbox_to_anchor=(1.05, 1), loc=2, borderaxespad=0.)
```

Out[ ]:
```
<AxesSubplot: xlabel='disrepair', ylabel='structure_vs_utils'>
```



The time progression of average house price against year built

In [ ]:
```
ahs_line = ahs.groupby('YRBUILT').agg({'MARKETVAL':'mean'}).reset_index().sort_values(
ahs_line
```

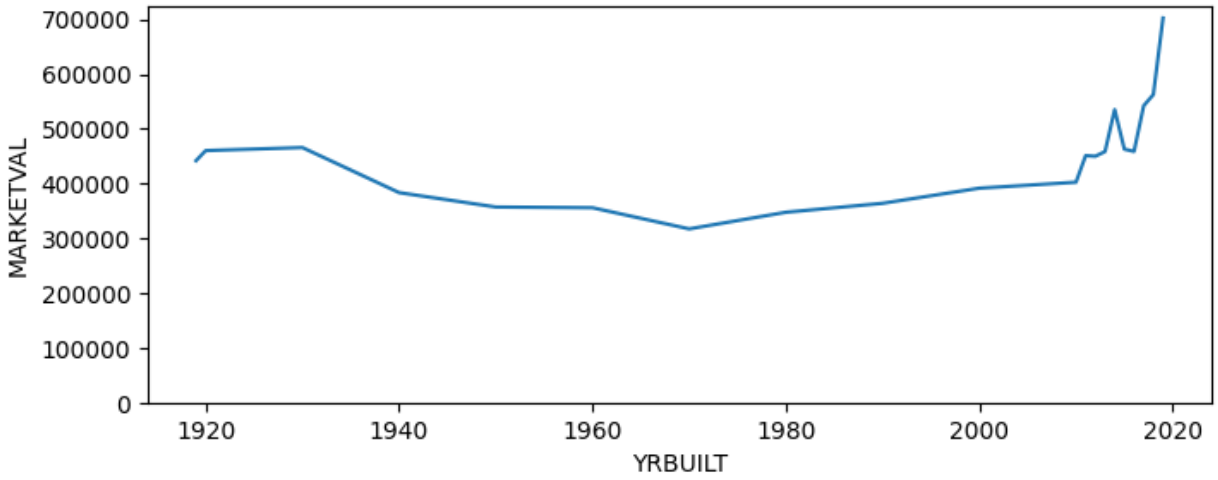
Out[ ]:

	YRBUILT	MARKETVAL
0	1919	441787.054923
1	1920	460242.827079
2	1930	465825.451481
3	1940	383461.994423
4	1950	357163.403623
5	1960	355834.658272
6	1970	317178.414418
7	1980	347564.607041
8	1990	363917.258788
9	2000	391545.044761
10	2010	402451.034810
11	2011	451071.786982
12	2012	449889.661836
13	2013	458420.516588
14	2014	535170.154122
15	2015	462851.737762
16	2016	458754.904605
17	2017	542170.035088
18	2018	562692.710526
19	2019	702470.655914

Before 2010 there is only information for the decade of the house

In [ ]:
```
# we want a line plot
plt.figure(figsize=(8, 3))
sns.lineplot(x='YRBUILT', y='MARKETVAL', data=ahs_line)
plt.ylim(0, 725000)
```

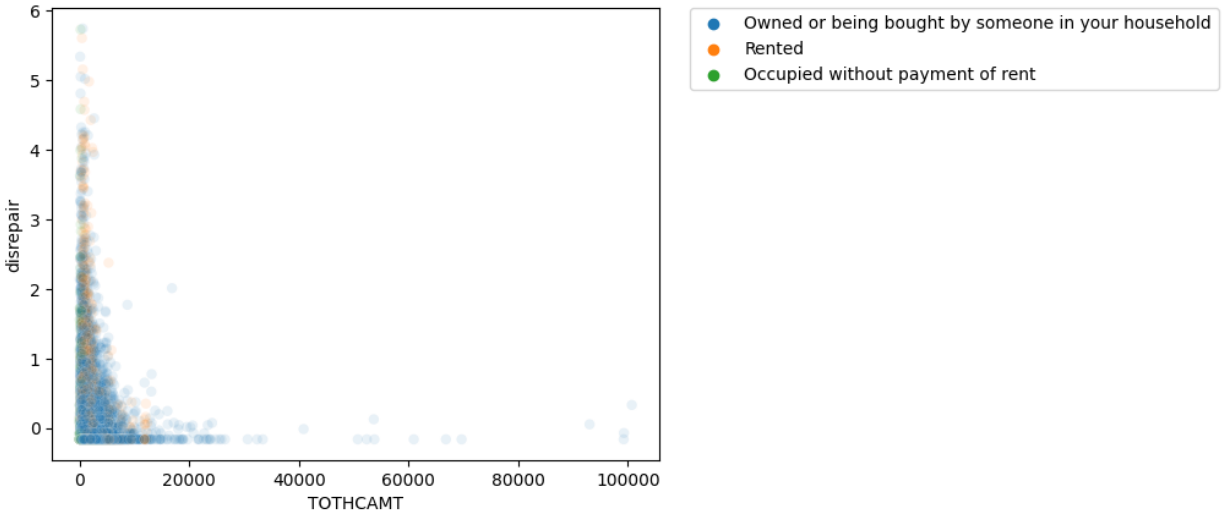
Out[ ]:
```
(0.0, 725000.0)
```



Graph matrices in which each cell contains a graph that is specific to a census division or ownership status

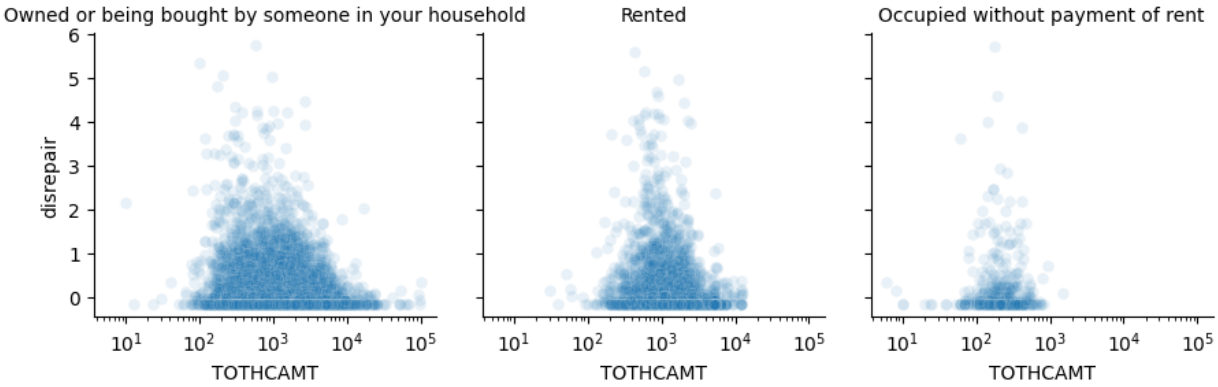
In [ ]:
```
# scatterplot for
sns.scatterplot(x='TOTHCAMT', y='disrepair', alpha=0.1, hue='TENURE', data=ahs_broken)
plt.legend(bbox_to_anchor=(1.05, 1), loc=2, borderaxespad=0.)
```

Out[ ]:
```
<matplotlib.legend.Legend at 0x17c759e41c0>
```



In [ ]:
```
# need a FacetGrid to plot multiple plots
g = sns.FacetGrid(ahs_broken, col="TENURE", height=3, aspect=1)
# plot the scatterplot on each facet
g.map(sns.scatterplot, "TOTHCAMT", "disrepair", alpha=0.1)
plt.xscale('log')
# tilt x labels 45 degrees
#g.set_xticklabels(rotation=45)
# set the titles
g.set_titles('{col_name}')
#plt.suptitle('Total housing costs vs. disrepair by tenure')
# put some padding between suptitle and subplots
#plt.subplots_adjust(top=0.1)
```

Out[ ]:
```
<seaborn.axisgrid.FacetGrid at 0x17c06c77c40>
```


Owned or being bought by someone in your household
Rented
Occupied without payment of rent
disrepair
TOTHCAMT
TOTHCAMT
TOTHCAMT

In [ ]: