Live Exercise

H. Diana McSpadden (hdm5s)

In [ ]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import prince
from scipy import stats
import os
from ydata_profiling import ProfileReport
```

In [ ]:

```
# UVa\ds6001\mod10\live\ahs_cleaned-1.csv
ahs = pd.read_csv('ahs_cleaned-1.csv', na_values=[-6, "'-9'"])
```

In [ ]:

```
ahs.head(2).T
```

Out[ ]:

	0	1
Unnamed: 0	0	1
DIVISION	South Atlantic	New England
TENURE	Owned or being bought by someone in your house...	Owned or being bought by someone in your house...
YRBUILT	2000	1970
UNITSIZE	2,000 to 2,499 square feet	3,000 to 3,999 square feet
HSHLDTYPE	Married-couple family household	Nonfamily household
HHRACE	White only	White only
HHSEX	Male	Female
HINCP	257000.0	201000.0
TOTHCAMT	1642.0	1049.0
MARKETVAL	280249.0	1000270.0
MAINTAMT	1022.0	295.0
FUSEBLOW	No fuses / breakers blown in the last 3 months	No fuses / breakers blown in the last 3 months
SEWBREAK	No breakdowns in the last 3 months	No breakdowns in the last 3 months
ROACH	No signs in the last 12 months	No signs in the last 12 months
RODENT	No signs in the last 12 months	No signs in the last 12 months
NOWIRE	Not broken	Not broken
PLUGS	Not broken	Not broken
COLD	Not broken	Not broken
NOTOIL	Not broken	Not broken
NOWAT	Not broken	Not broken
FLOORHOLE	Not broken	Not broken
FNDCRUMB	Not broken	Not broken
PAINTPEEL	Not broken	Not broken
ROOFHOLE	Not broken	Not broken
ROOFSAG	Not broken	Not broken
ROOFSHIN	Not broken	Not broken
WALLCRACK	Not broken	Not broken
WALLSIDE	Not broken	Not broken
WALLSLOPE	Not broken	Not broken
WINBOARD	Not broken	Not broken
WINBROKE	Not broken	Not broken

	0	1
LEAKI	Not broken	Not broken
MOLDBATH	Not broken	Not broken

In [ ]:

```
ahs
```

Out[ ]:

	Unnamed: 0	DIVISION	TENURE	YRBUILT	UNITSIZE	HSHLDTYPE	HHRACE	HHSEX	HINCP
0	0	South Atlantic	Owned or being bought by someone in your house...	2000	2,000 to 2,499 square feet	Married-couple family household	White only	Male	257000.0
1	1	New England	Owned or being bought by someone in your house...	1970	3,000 to 3,999 square feet	Nonfamily household	White only	Female	201000.0
2	2	West South Central	NaN	1970	750 to 999 square feet	NaN	NaN	NaN	NaN
3	3	West South Central	Owned or being bought by someone in your house...	1970	2,000 to 2,499 square feet	Married-couple family household	White only	Male	66900.0
4	4	West North Central	Rented	1970	750 to 999 square feet	Nonfamily household	Black only	Female	35000.0
...	...	...	...	...	...	...	...	...	...
63180	63180	East North Central	Owned or being bought by someone in your house...	2016	4,000 square feet or more	Nonfamily household	White only	Male	74000.0
63181	63181	South Atlantic	Owned or being bought by someone in your house...	2018	1,500 to 1,999 square feet	Married-couple family household	White only	Male	207000.0
63182	63182	South Atlantic	Owned or being bought by someone in your house...	2018	2,000 to 2,499 square feet	Married-couple family household	White only	Female	158100.0

	Unnamed: 0	DIVISION	TENURE	YRBUILT	UNITSIZE	HSHLDTYPE	HHRACE	HHSEX	HINCP
63183	63183	South Atlantic	Owned or being bought by someone in your house...	2018	2,500 to 2,999 square feet	Married-couple family household	White only	Male	130200.0
63184	63184	South Atlantic	Owned or being bought by someone in your house...	2016	3,000 to 3,999 square feet	Married-couple family household	White only	Female	120000.0

In [ ]:

```
ahs = ahs.iloc[:,1:]
ahs.index.name = 'row_id'
```

In [ ]:

```
ahs
```

Out[ ]:

	DIVISION	TENURE	YRBUILT	UNITSIZE	HSHLDTYPE	HHRACE	HHSEX	HINCP	TOTHCAM
row_id									
0	South Atlantic	Owned or being bought by someone in your house...	2000	2,000 to 2,499 square feet	Married-couple family household	White only	Male	257000.0	1642.
1	New England	Owned or being bought by someone in your house...	1970	3,000 to 3,999 square feet	Nonfamily household	White only	Female	201000.0	1049.
2	West South Central	NaN	1970	750 to 999 square feet	NaN	NaN	NaN	NaN	Na
3	West South Central	Owned or being bought by someone in your house...	1970	2,000 to 2,499 square feet	Married-couple family household	White only	Male	66900.0	671.
4	West North Central	Rented	1970	750 to 999 square feet	Nonfamily household	Black only	Female	35000.0	680.
...	...	...	...	...	...	...	...	...	.
63180	East North Central	Owned or being bought by someone in your house...	2016	4,000 square feet or more	Nonfamily household	White only	Male	74000.0	6171.
63181	South Atlantic	Owned or being bought by someone in your house...	2018	1,500 to 1,999 square feet	Married-couple family household	White only	Male	207000.0	2520.
63182	South Atlantic	Owned or being bought by someone	2018	2,000 to 2,499 square feet	Married-couple family household	White only	Female	158100.0	1896.

	DIVISION	TENURE	YRBUILT	UNITSIZE	HSHLDTYPE	HHRACE	HHSEX	HINCP	TOTHCAM
row_id									
		in your house...							
63183	South Atlantic	Owned or being bought by someone in your house...	2018	2,500 to 2,999 square feet	Married-couple family household	White only	Male	130200.0	2008.
63184	South Atlantic	Owned or being bought by someone in your house...	2016	3,000 to 3,999 square feet	Married-couple family household	White only	Female	120000.0	2122.
63185									

In [ ]:

```
ahs.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 63185 entries, 0 to 63184
Data columns (total 33 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   DIVISION   63185 non-null  object 
 1   TENURE     54455 non-null  object 
 2   YRBUILT    63185 non-null  int64  
 3   UNITSIZE   57629 non-null  object 
 4   HSHLDTYPE  54455 non-null  object 
 5   HHRACE     54455 non-null  object 
 6   HHSEX      54455 non-null  object 
 7   HINCP      54455 non-null  float64
 8   TOTHCAMT   54455 non-null  float64
 9   MARKETVAL  38390 non-null  float64
 10  MAINTAMT   32972 non-null  float64
 11  FUSEBLOW   54435 non-null  object 
 12  SEWBREAK   54355 non-null  object 
 13  ROACH      54455 non-null  object 
 14  RODENT     54455 non-null  object 
 15  NOWIRE     63035 non-null  object 
 16  PLUGS      63035 non-null  object 
 17  COLD       50479 non-null  object 
 18  NOTOIL     54417 non-null  object 
 19  NOWAT      53731 non-null  object 
 20  FLOORHOLE  63185 non-null  object 
 21  FNDCRUMB   41861 non-null  object 
 22  PAINTPEEL  63185 non-null  object 
 23  ROOFHOLE   41939 non-null  object 
 24  ROOFSAG    42094 non-null  object 
 25  ROOFSHIN   41956 non-null  object 
 26  WALLCRACK  63185 non-null  object 
 27  WALLSIDE   42168 non-null  object 
 28  WALLSLOPE  42202 non-null  object 
 29  WINBOARD   42373 non-null  object 
 30  WINBROKE   42339 non-null  object 
 31  LEAKI      54455 non-null  object 
 32  MOLDBATH   53820 non-null  object 
dtypes: float64(4), int64(1), object(28)
memory usage: 15.9+ MB
```

In [ ]:

```
ahs.describe().T
```

Out[ ]:

	count	mean	std	min	25%	50%	75%	max
YRBUILT	63185.0	1970.509646	26.429845	1919.0	1950.0	1970.0	1990.00	2019.0
HINCP	54455.0	87066.124176	100064.851607	-5000.0	27500.0	60000.0	111000.00	3876000.0
TOTHCAMT	54455.0	1517.628739	1783.335753	0.0	670.0	1164.0	1892.50	100700.0
MARKETVAL	38390.0	376276.939750	553786.639374	1000.0	140446.5	255273.0	435968.25	9999998.0
MAINTAMT	32972.0	874.907710	1357.366635	-9.0	2.0	460.5	1016.00	9998.0

In [ ]:

```
profile = ProfileReport(ahs, title='AHS Profiling Report', html={'style':{'full_width'
```

In [ ]:

```
profile.to_notebook_iframe()
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]
Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]
Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]
```

Overview

Dataset statistics

Number of variables	33
Number of observations	63185
Missing cells	365855
Missing cells (%)	17.5%
Duplicate rows	478
Duplicate rows (%)	0.8%
Total size in memory	15.9 MiB
Average record size in memory	264.0 B

Variable types

Categorical	28
Numeric	5

Alerts

Dataset has 478 (0.8%) duplicate rows	**Duplicates**
`HINCP` is highly overall correlated with `TOTHCAMT`	**High correlation**
`TOTHCAMT` is highly overall correlated with `HINCP` and 1 other fields (HINCP, MARKETVAL)	**High correlation**

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 [ ]:

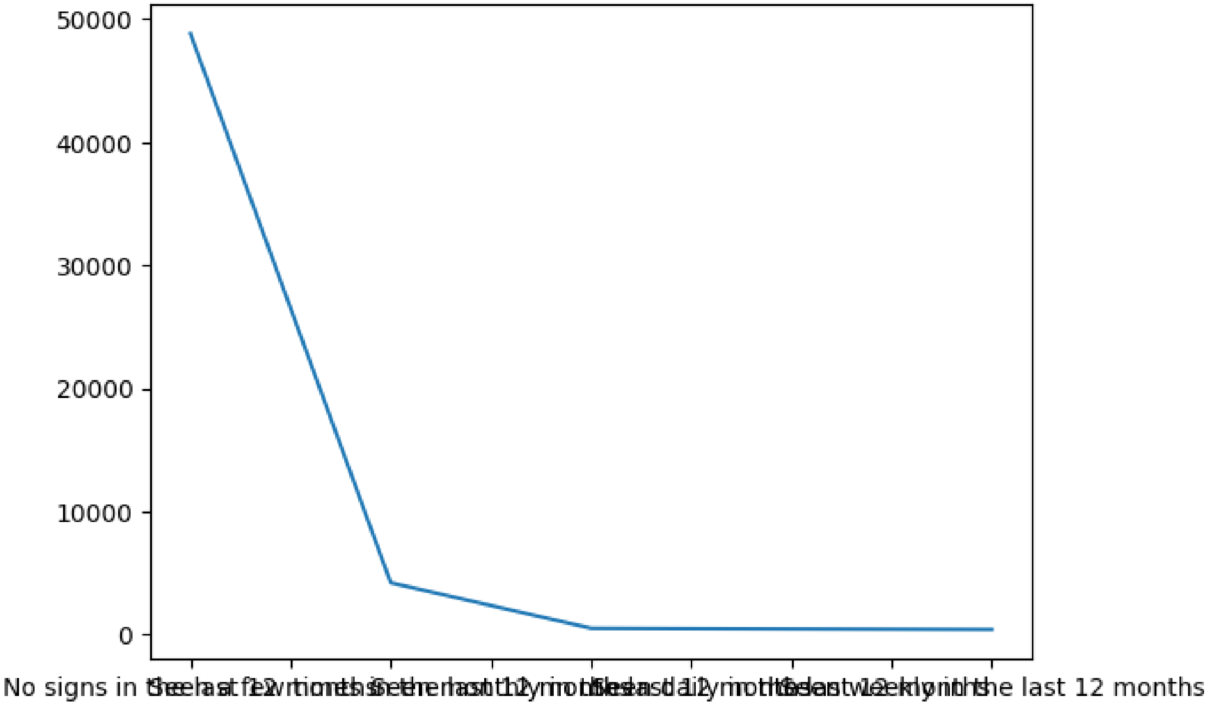
```
%matplotlib inline
```

In [ ]:

```
ahs['RODENT'].value_counts().plot()
```

Out[ ]:

```
<AxesSubplot: >
```



Now some ANOVA TEST

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.dropna(),
    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.groupby("RODENT").HINCP.mean().sort_values(ascending=False).to_frame('mean_inc').
```

Out[ ]:

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

In [ ]:

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

Out[ ]:

	MARKETVAL	YRBUILT
MARKETVAL	1.00000	-0.00403
YRBUILT	-0.00403	1.00000

In [ ]:

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

Out[ ]:

	MARKETVAL	YRBUILT
MARKETVAL	1.000000	0.083796
YRBUILT	0.083796	1.000000

In [ ]:

```

```

In [ ]:

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

Out[ ]:

```
(-0.004029500232993765, 0.4298243664197942)
```

MCA

In [ ]:

```
broken = ahs[['FUSEBLOW','ROACH','RODENT','NOWIRE','PLUGS']].dropna()
broken.shape
```

Out[ ]:

```
(54435, 5)
```

In [ ]:

```
# sample 5000 rows from broken
broken = broken.sample(4000, random_state=42)
```

In [ ]:

```
prince_mca = prince.MCA(n_components=2)
prince_mca = prince_mca.fit(broken)
#gss_mca = prince_mca.transform(gss_cat)
prince_mca.row_coordinates(broken)
```

Out[ ]:

row_id	0	1
37340	0.342858	0.115637
26264	1.115773	-1.465172
43884	-0.175239	-0.011974
35036	-0.175239	-0.011974
9002	0.118970	0.677848
...	...	...
59549	-0.175239	-0.011974
23758	-0.111287	0.235141
6106	-0.175239	-0.011974
36339	-0.111287	0.235141
52994	-0.175239	-0.011974

4000 rows × 2 columns

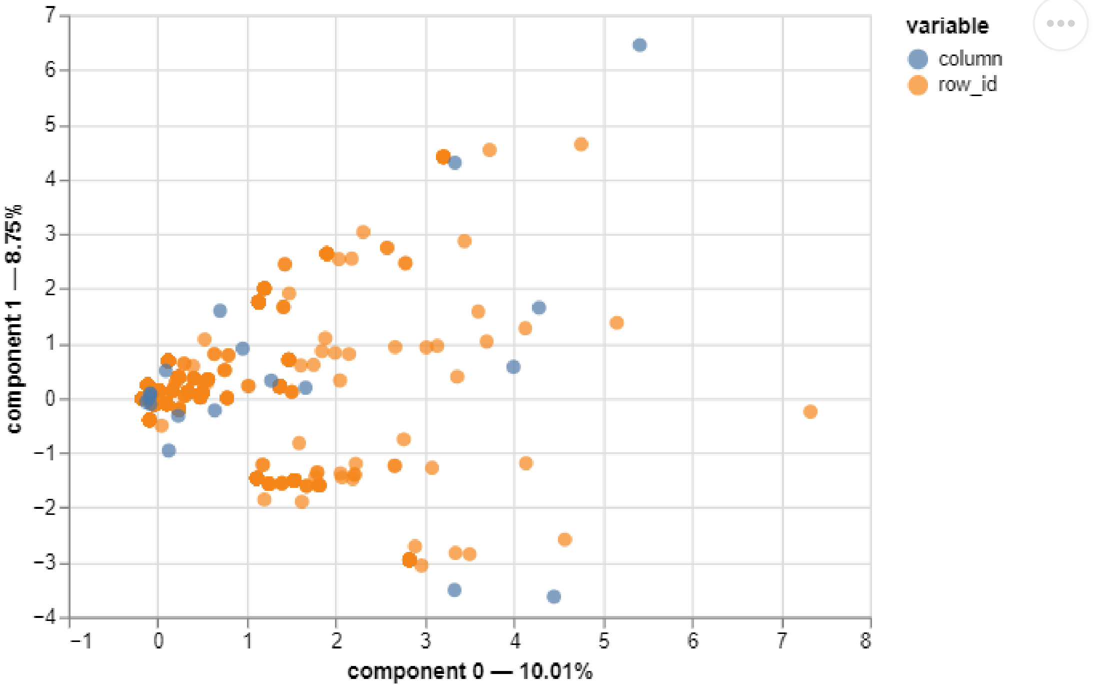
In [ ]:

```
plt = prince_mca.plot(
    broken,
    x_component=0,
    y_component=1
)
```

In [ ]:

```
plt
```

Out[ ]:



In [ ]: