

PS1 Part 1, Data Section

April 16, 2018

1 Laurence Warner

```
In [20]: import pandas as pd
import seaborn as sb
import matplotlib.pyplot as plt
%matplotlib inline
```

1.1 1

CPS data is stored and curated by NBER. But I couldn't find CSV data. Interested in how Prof. Soltoff got hold of CSV version.

1.2 2

Many key papers wanting to include variables on employment. Mellor and Milyo (2002) investigate effects on health. Also E.g. Burkhauser et al., 2011; Davis et al., 2015; Edwards & Lindstrom, 2017.

1.3 3

Monthly sample of 60,000 households.

1.4 4

```
In [4]: cps = pd.read_csv('jun_17_cps.csv')
```

```
/Users/laurencewarner/anaconda3/lib/python3.6/site-packages/IPython/core/interactiveshell.py:2
interactivity=interactivity, compiler=compiler, result=result)
```

```
In [5]: cps.head()
```

```
Out [5]:
```

	hrhhid	hrmonth	hryear4	hurespli	hufinal \
0	4797110019	6	2017	-1	UNOCCUPIED TENT OR TRAILER SITE
1	110177987986	6	2017	-1	UNOCCUPIED TENT OR TRAILER SITE
2	110327856469	6	2017	1	CAPI COMPLETE
3	110327856469	6	2017	1	CAPI COMPLETE
4	110327856469	6	2017	1	CAPI COMPLETE

	huspnish	hetenure		hehousut	hetelhhd	hetelavl	\
0	0	-1	UNOCCUPIED TENT SITE OR TRLR SITE		-1	-1	
1	0	-1	UNOCCUPIED TENT SITE OR TRLR SITE		-1	-1	
2	0	-1	HOUSE, APARTMENT, FLAT	YES		-1	
3	0	-1	HOUSE, APARTMENT, FLAT	YES		-1	
4	0	-1	HOUSE, APARTMENT, FLAT	YES		-1	

	...	pxdisrem	pxdisphy	pxdisdrs	pxdisout	hxfaminc	prdasian	pepdemp1	\
0	...	-1	-1	-1	-1	1	-1	-1	
1	...	-1	-1	-1	-1	1	-1	-1	
2	...	0	0	0	0	0	-1	-1	
3	...	0	0	0	0	0	-1	-1	
4	...	-1	-1	-1	-1	0	-1	-1	

	ptnmemp1	pepdemp2	ptnmemp2
0	-1	-1	-1
1	-1	-1	-1
2	-1	-1	-1
3	-1	-1	-1
4	-1	-1	-1

[5 rows x 387 columns]

In [6]: `cps.shape`

Out[6]: (148054, 387)

In [7]: `cps.describe()`

Out[7]:

	hrhhid	hrmonth	hryear4	hurespli	hetenure	\
count	1.480540e+05	148054.0	148054.0	148054.000000	148054.0	
mean	3.561019e+14	6.0	2017.0	1.044909	-1.0	
std	2.924787e+14	0.0	0.0	0.981399	0.0	
min	1.130714e+08	6.0	2017.0	-1.000000	-1.0	
25%	1.086131e+14	6.0	2017.0	1.000000	-1.0	
50%	2.788357e+14	6.0	2017.0	1.000000	-1.0	
75%	6.044004e+14	6.0	2017.0	1.000000	-1.0	
max	9.999322e+14	6.0	2017.0	13.000000	-1.0	

	hwhhwgt	hrnumhou	hrmis	huprscnt	\
count	126205.000000	148054.000000	148054.000000	148054.000000	
mean	2484.506678	2.858498	4.523343	0.633857	
std	1291.757648	1.873351	2.287926	0.946577	
min	182.902600	0.000000	1.000000	0.000000	
25%	1300.972800	2.000000	3.000000	0.000000	
50%	2829.153300	3.000000	5.000000	0.000000	
75%	3429.504700	4.000000	7.000000	1.000000	
max	9804.787300	15.000000	8.000000	9.000000	

	hrhhid2	...	pxdadtyp	pxmomtyp	\
count	148054.000000	...	148054.000000	148054.000000	
mean	5890.063159	...	1.033062	1.130817	
std	924.925554	...	4.969214	5.680917	
min	5011.000000	...	-1.000000	-1.000000	
25%	5011.000000	...	0.000000	0.000000	
50%	5112.000000	...	1.000000	1.000000	
75%	7011.000000	...	1.000000	1.000000	
max	7112.000000	...	53.000000	53.000000	

	pxcohab	pxdisear	pxdiseye	pxdisrem	\
count	148054.000000	148054.000000	148054.000000	148054.000000	
mean	1.636585	0.342794	0.354053	0.380780	
std	7.943847	4.778441	4.811441	4.898336	
min	-1.000000	-1.000000	-1.000000	-1.000000	
25%	-1.000000	-1.000000	-1.000000	-1.000000	
50%	1.000000	0.000000	0.000000	0.000000	
75%	1.000000	0.000000	0.000000	0.000000	
max	53.000000	43.000000	43.000000	43.000000	

	pxdisphy	pxdisdrs	pxdisout	hxfaminc
count	148054.000000	148054.000000	148054.000000	148054.000000
mean	0.377322	0.375444	0.380983	5.519013
std	4.878873	4.875035	4.892299	12.146658
min	-1.000000	-1.000000	-1.000000	0.000000
25%	-1.000000	-1.000000	-1.000000	0.000000
50%	0.000000	0.000000	0.000000	0.000000
75%	0.000000	0.000000	0.000000	1.000000
max	43.000000	43.000000	43.000000	43.000000

[8 rows x 158 columns]

In [8]: cols_num = ['hrmonth', 'hryear4', 'hrnumhou', 'huprscnt', 'peernh1o', 'prernwa', 'peern',

In [9]: cols_num

Out[9]: ['hrmonth',
'hryear4',
'hrnumhou',
'huprscnt',
'peernh1o',
'prernwa',
'peern',
'premp',
'puhrot2',
'pehractt']

In [10]: cps_num = cps[cols_num]

```
In [11]: cps_num.head()
```

```
Out[11]:
```

	hrmonth	hryear4	hrnumhou	huprsent	peernh1o	prernwa	peern	premp	\
0	6	2017	0	0	NaN	NaN	NaN	-1	
1	6	2017	0	0	NaN	NaN	NaN	-1	
2	6	2017	6	0	40.0	1600.0	NaN	1	
3	6	2017	6	0	12.0	360.0	NaN	1	
4	6	2017	6	0	NaN	NaN	NaN	-1	

	puhrot2	pehractt
0	-1	-1
1	-1	-1
2	-1	50
3	-1	30
4	-1	-1

```
In [13]: cps_num.isnull().sum()
```

```
Out[13]:
```

hrmonth	0
hryear4	0
hrnumhou	0
huprsent	0
peernh1o	142065
prernwa	134584
peern	147254
premp	0
puhrot2	0
pehractt	0
dtype:	int64

Table of descriptive statistics for these 8 numerical variables:

```
In [18]: desc = cps_num.describe()
desc
```

```
Out[18]:
```

	hrmonth	hryear4	hrnumhou	huprsent	peernh1o	\
count	148054.0	148054.0	148054.000000	148054.000000	5989.000000	
mean	6.0	2017.0	2.858498	0.633857	16.58289	
std	0.0	0.0	1.873351	0.946577	9.05162	
min	6.0	2017.0	0.000000	0.000000	1.12000	
25%	6.0	2017.0	2.000000	0.000000	10.50000	
50%	6.0	2017.0	3.000000	0.000000	14.00000	
75%	6.0	2017.0	4.000000	1.000000	20.00000	
max	6.0	2017.0	15.000000	9.000000	90.00000	

	prernwa	peern	premp	puhrot2	pehractt
count	13470.000000	800.000000	148054.000000	148054.000000	148054.000000
mean	935.669806	217.358450	-0.206546	-0.739278	14.451599
std	675.809701	260.890046	0.978440	2.082963	21.104730

min	0.000000	0.000000	-1.000000	-3.000000	-1.000000
25%	450.000000	50.000000	-1.000000	-1.000000	-1.000000
50%	750.000000	135.000000	-1.000000	-1.000000	-1.000000
75%	1230.760000	275.000000	1.000000	-1.000000	40.000000
max	2884.610000	1800.000000	1.000000	84.000000	150.000000

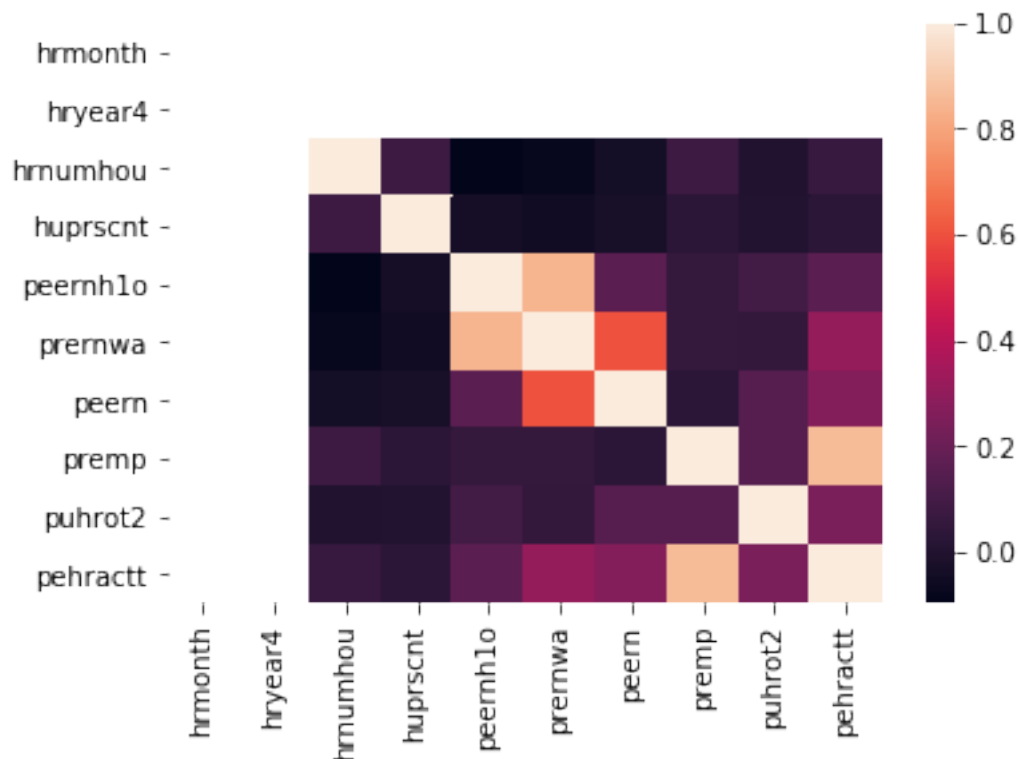
1.5 5

Visualizatio: heatmap of correlation between variables.

```
In [26]: corr = cps_num.corr()
```

```
In [27]: sb.heatmap(corr)
```

```
Out [27]: <matplotlib.axes._subplots.AxesSubplot at 0x11793beb8>
```



Almost all have positive correlation. Overtime amount (peerm) and weekly earnings (pre-rnwa) have a slightly less positive relationship. Maybe because people who work overtime are less wealthy.

1.6 6

Descriptive statistics conditional on sex.

```
In [28]: sex = cps['pesex']
```

```
In [30]: m = cps[sex == 'MALE']
        f = cps[sex == 'FEMALE']
```

```
In [33]: m.describe()
```

```
Out [33]:
```

	hrhhid	hrmonth	hryear4	hurespli	hetenure	hwhhwgt	\
count	6.125600e+04	61256.0	61256.0	61256.000000	61256.0	61256.000000	
mean	3.566636e+14	6.0	2017.0	1.323642	-1.0	2486.959261	
std	2.926331e+14	0.0	0.0	0.710668	0.0	1301.193180	
min	1.130714e+08	6.0	2017.0	-1.000000	-1.0	182.902600	
25%	1.090919e+14	6.0	2017.0	1.000000	-1.0	1292.329400	
50%	2.807591e+14	6.0	2017.0	1.000000	-1.0	2831.083700	
75%	6.045004e+14	6.0	2017.0	1.000000	-1.0	3438.242200	
max	9.998976e+14	6.0	2017.0	13.000000	-1.0	9804.787300	

	hrnumhou	hrmis	huprscnt	hrhhid2	...	\
count	61256.000000	61256.000000	61256.000000	61256.000000	...	
mean	3.291139	4.546640	0.656834	5882.711130	...	
std	1.677363	2.278692	0.914796	922.918262	...	
min	1.000000	1.000000	0.000000	5011.000000	...	
25%	2.000000	3.000000	0.000000	5011.000000	...	
50%	3.000000	5.000000	0.000000	5111.000000	...	
75%	4.000000	7.000000	1.000000	7011.000000	...	
max	15.000000	8.000000	9.000000	7112.000000	...	

	pxdadtyp	pxmomtyp	pxcohab	pxdisear	pxdiseye	\
count	61256.000000	61256.000000	61256.000000	61256.000000	61256.000000	
mean	1.350839	1.536078	2.230753	0.547881	0.559749	
std	5.210710	6.275466	8.981328	5.102527	5.136157	
min	0.000000	0.000000	-1.000000	-1.000000	-1.000000	
25%	1.000000	0.000000	0.000000	0.000000	0.000000	
50%	1.000000	1.000000	1.000000	0.000000	0.000000	
75%	1.000000	1.000000	1.000000	0.000000	0.000000	
max	53.000000	53.000000	53.000000	43.000000	43.000000	

	pxdisrem	pxdisphy	pxdisdrs	pxdisout	hxfaminc	\
count	61256.000000	61256.000000	61256.000000	61256.000000	61256.000000	
mean	0.589934	0.587110	0.581706	0.589591	6.251616	
std	5.229189	5.208846	5.195204	5.219521	12.953658	
min	-1.000000	-1.000000	-1.000000	-1.000000	0.000000	
25%	0.000000	0.000000	0.000000	0.000000	0.000000	
50%	0.000000	0.000000	0.000000	0.000000	0.000000	
75%	0.000000	0.000000	0.000000	0.000000	0.000000	
max	43.000000	43.000000	43.000000	43.000000	43.000000	

```
[8 rows x 158 columns]
```

```
In [34]: f.describe()
```

```

Out [34]:
      hrhhid  hrmonth  hryear4  hurespli  hetenure  hwhhwgt  \
count  6.494900e+04  64949.0  64949.0  64949.000000  64949.0  64949.000000
mean   3.567710e+14      6.0   2017.0      1.302514     -1.0   2482.193549
std    2.926646e+14      0.0      0.0      0.681292      0.0   1282.800751
min    1.130714e+08      6.0   2017.0     -1.000000     -1.0    221.103800
25%    1.087921e+14      6.0   2017.0      1.000000     -1.0   1310.883800
50%    2.805626e+14      6.0   2017.0      1.000000     -1.0   2827.184700
75%    6.052240e+14      6.0   2017.0      1.000000     -1.0   3420.896400
max    9.999322e+14      6.0   2017.0     13.000000     -1.0   9804.787300

      hrnumhou      hrmis      huprscnt      hrhhid2      ...      \
count  64949.000000  64949.000000  64949.000000  64949.000000  ...
mean    3.210442    4.542595    0.657531    5882.220866  ...
std     1.682402    2.284472    0.920314    924.232983  ...
min     1.000000    1.000000    0.000000    5011.000000  ...
25%     2.000000    3.000000    0.000000    5011.000000  ...
50%     3.000000    5.000000    0.000000    5111.000000  ...
75%     4.000000    7.000000    1.000000    7011.000000  ...
max     15.000000    8.000000    9.000000    7112.000000  ...

      pxdadtyp      pxmomtyp      pxcohab      pxdisear      pxdiseye      \
count  64949.000000  64949.000000  64949.000000  64949.000000  64949.000000
mean    1.417281    1.465411    1.963156    0.601087    0.615560
std     5.389659    5.884847    8.061966    5.175139    5.212158
min     0.000000    0.000000   -1.000000   -1.000000   -1.000000
25%     1.000000    0.000000    1.000000    0.000000    0.000000
50%     1.000000    1.000000    1.000000    0.000000    0.000000
75%     1.000000    1.000000    1.000000    0.000000    0.000000
max     53.000000   53.000000   53.000000   43.000000   43.000000

      pxdisrem      pxdisphy      pxdisdrrs      pxdisout      hxfaminc
count  64949.000000  64949.000000  64949.000000  64949.000000  64949.000000
mean    0.648016    0.642797    0.643613    0.648801    6.360036
std     5.305797    5.284114    5.288862    5.302050   13.028822
min    -1.000000   -1.000000   -1.000000   -1.000000    0.000000
25%     0.000000    0.000000    0.000000    0.000000    0.000000
50%     0.000000    0.000000    0.000000    0.000000    0.000000
75%     0.000000    0.000000    0.000000    0.000000    0.000000
max     43.000000   43.000000   43.000000   43.000000   43.000000

```

[8 rows x 158 columns]