Pandas Exercises II - due thu 09/16 | *by Raj and Charlotte*

Exercise 1

Data for these exercises can be found here. First, download US_ACS_2017_10pct_sample.dta.

Exercise 2

Now import US_ACS_2017_10pct_sample.dta into a pandas DataFrame. This can be done with the command pd.read_stata, which will read in files created in the program Stata (and which uses the file suffix .dta). This is a format commonly used by social scientists.

In [200…

```
import pandas as pd
acs = pd.read_stata('US_ACS_2017_10pct_sample.dta')
```

Exercise 3

We want to find out the number of rows.

In [201…

```
f'Our dataset has {len(acs)} rows'
```

Out[201…

```
'Our dataset has 319004 rows'
```

Exercise 4

We want to find out the number of columns.

In [202…

```
f'Our dataset has {len(acs.columns)} columns'
```

Out[202…

```
'Our dataset has 104 columns'
```

Exercise 5

Let's see what variables are in this dataset.

In [203…

```
acs.columns
```

Out[203…

```
Index(['year', 'datanum', 'serial', 'cbserial', 'numprec', 'subsamp', 'hhwt
',
       'hhtype', 'cluster', 'adjust',
       ...
       'migcounty1', 'migmet131', 'vetdisab', 'diffrem', 'diffphys', 'diffm
ob',
       'diffcare', 'diffsens', 'diffeye', 'diffhear'],
      dtype='object', length=104)
```

In [204…

```
print("The following variables are in our dataset:")
for c in acs.columns: print(c)
```

```
The following variables are in our dataset:
year
datanum
serial
cbserial
numprec
subsamp
hhwt
hhtype
cluster
adjust
cpi99
region
stateicp
statefip
countyicp
countyfip
metro
city
citypop
strata
gq
farm
ownershp
ownershpd
mortgage
mortgag2
mortamt1
mortamt2
respmode
pernum
cbpernum
perwt
slwt
```

```
famunit
sex
age
marst
birthyr
race
raced
hispan
hispand
bpl
bpld
citizen
yrnatur
yrimmig
language
languaged
speakeng
hcovany
hcovpriv
hinsemp
hinspur
hinstri
hcovpub
hinscaid
hinscare
hinsva
hinsihs
school
educ
educd
gradeatt
gradeattd
schltype
degfield
degfieldd
degfield2
degfield2d
empstat
empstatd
labforce
occ
ind
classwkr
classwkrd
looking
availble
inctot
ftotinc
incwage
incbus00
incss
incwelfr
incinvst
incretir
incsupp
incother
incearn
```

```
poverty
migrate1
migrate1d
migplac1
migcounty1
migmet131
vetdisab
diffrem
diffphys
diffmob
diffcare
diffsens
diffeye
diffhear
```

Exercise 4

That's a lot of variables, and definitely more than we need. In general, life is easier when working with these kinds of huge datasets if you can narrow down the number of variables a little. In this exercise, we will be looking at the relationship between education and wages, we need variables for:

- Age
- Income
- Education
- Employment status (is the person actually working)

These quantities of interest correspond to the following variables in our data: age, inctot, educ, and empstat.

Subset your data to just those variables.

In [205…

```
print("Subsetting dataset")
small_acs = acs.loc[:, (acs.columns == 'age') | (acs.columns == 'inctot') |
small_acs
```

```
Subsetting dataset
```

Out[205…

	age	educ	empstat	inctot
0	4	nursery school to grade 4	n/a	9999999
1	17	grade 11	employed	6000
2	63	4 years of college	employed	6150
3	66	grade 12	not in labor force	14000
4	1	n/a or no schooling	n/a	9999999
...	...	...	...	...
318999	33	4 years of college	employed	22130
319000	4	nursery school to grade 4	n/a	9999999
319001	20	grade 12	employed	5000
319002	47	5+ years of college	employed	240000
319003	33	5+ years of college	employed	48000

319004 rows × 4 columns

Exercise 5

Now that we have a more manageable number of variables, it’s often very useful to look at a handful of rows of your data. The easiest way to do this is probably the .head() method (which will show you the first five rows), or the tail() method, which will show you the last five rows.

But to get a good sense of your data, it’s often better to use the sample() command, which returns a random set of rows. As the first and last rows are sometimes not representative, a random set of rows can be very helpful. Try looking at a random sample of 20 rows (note: you don’t have to run .sample() ten times to get ten rows. Look at the .sample help file if you’re stuck.

In [206…

```
print("Here are 20 random rows sampled:")
small_acs.sample(20)
```

```
Here are 20 random rows sampled:
```

Out[206…

	age	educ	empstat	inctot
147530	51	grade 12	employed	27000
24995	47	1 year of college	employed	38000
55583	5	nursery school to grade 4	n/a	9999999
284935	69	2 years of college	not in labor force	29090
81740	2	n/a or no schooling	n/a	9999999
82959	37	5+ years of college	employed	19600
69238	14	grade 5, 6, 7, or 8	n/a	9999999
286707	50	5+ years of college	employed	119000
253232	19	1 year of college	not in labor force	750
93807	79	grade 5, 6, 7, or 8	not in labor force	25000
105609	30	grade 5, 6, 7, or 8	employed	125000
66158	71	grade 12	not in labor force	9300
99487	35	grade 12	employed	25000
68703	4	n/a or no schooling	n/a	9999999
301773	70	1 year of college	employed	24000
103486	30	grade 11	employed	30000
262674	45	4 years of college	employed	140000
42886	28	2 years of college	employed	12000
149289	67	grade 12	not in labor force	32400
208356	87	4 years of college	not in labor force	45400

Exercise 6

Do you see any immediate problems? Write them down with your partner.

- a lot of n/a's, i.e. not available data points (non-responsive or missings)
- a lot of extremely high values for the income variable (9999999) that are most likely to be missings that still need to be recoded because currently they are most likely to skew any analysis or descriptive statistics

---

Exercise 7

So let’s begin by dropping anyone who has inctot equal to 9999999.

In [207…

```
acs_new = small_acs.drop(small_acs[small_acs.inctot == 9999999].index)
acs_new.sample(20)
```

Out[207…

	age	educ	empstat	inctot
155499	16	grade 10	not in labor force	0
247780	51	1 year of college	not in labor force	15500
256723	42	2 years of college	not in labor force	0
293015	43	5+ years of college	employed	40000
203885	46	grade 12	employed	40000
8075	49	2 years of college	employed	35000
286732	17	grade 11	not in labor force	0
181116	16	grade 10	not in labor force	1200
121882	28	1 year of college	employed	44000
315793	36	grade 12	employed	52000
144741	22	grade 12	employed	12000
76095	18	grade 11	not in labor force	0
304845	69	4 years of college	not in labor force	0
298222	65	grade 12	not in labor force	1800
295141	26	grade 12	employed	37000
99312	36	4 years of college	employed	30000
60874	40	4 years of college	employed	60000
130945	62	1 year of college	employed	39000
33161	24	4 years of college	employed	26400
214201	57	grade 12	employed	36000

Exercise 8

OK, the other potential problem is that our data includes lots of people who are unemployed and people who are not in the labor force (this means they not only don't have a job, but also aren't looking for a job). For this analysis, we want to focus on the wages of people who are currently employed. So subset the dataset for the people for whom empstat is equal to "employed".

Note that our decision to only look at people who are employed impacts how we should interpret the relationship we estimate between education and income. Because we are only looking at employed people, we will be estimating the relationship between education and income for people who are employed. That means that if education affects the likelihood someone is employed, we won't capture that in this analysis. (Economists all this the "intensive margin", while looking at whether people get jobs in the first place is called the "extensive margin".)

In [208…

```
print("Subsetting the dataset for the people for whom empstat is equal to '
acs_emp = acs_new.loc[acs_new['empstat'] == "employed"]
acs_emp
```

```
Subsetting the dataset for the people for whom empstat is equal to “employe
d”, i.e. people that are employed.
```

Out[208…

	age	educ	empstat	inctot
1	17	grade 11	employed	6000
2	63	4 years of college	employed	6150
5	50	grade 12	employed	50000
9	17	grade 12	employed	2000
10	47	n/a or no schooling	employed	18000
...	...	...	...	...
318995	67	grade 12	employed	125000
318999	33	4 years of college	employed	22130
319001	20	grade 12	employed	5000
319002	47	5+ years of college	employed	240000
319003	33	5+ years of college	employed	48000

148758 rows × 4 columns

Exercise 9

Now let's turn to education. The educ variable seems to have a lot of discrete values. Let's see what values exist, and their distribution, using the value_counts() method. This is an extremely useful tool you'll use a lot! Try the following code (modified for the name of your dataset, of course)

In [209…

```
acs_emp['educ'].value_counts()
```

Out[209…

```
grade 12                     47815
4 years of college           33174
1 year of college            22899
5+ years of college          20995
2 years of college           14077
grade 11                      2747
grade 5, 6, 7, or 8           2092
grade 10                      1910
n/a or no schooling           1291
grade 9                       1290
nursery school to grade 4      468
Name: educ, dtype: int64
```

Exercise 10

There are a lot of values in here, so let's just check a couple. What is the average value of inctot for people whose highest grade level is "grade 12" (in the US, that is someone who has graduated high school)?

In [210…

```
acs_emp.groupby('educ', as_index=False)['inctot'].mean().round(2)
```

Out[210…

	educ	inctot
0	n/a or no schooling	32276.88
1	nursery school to grade 4	27592.65
2	grade 5, 6, 7, or 8	30684.20
3	grade 9	27171.91
4	grade 10	23018.80
5	grade 11	21541.69
6	grade 12	38957.76
7	1 year of college	43123.87
8	2 years of college	48679.31
9	4 years of college	75485.05
10	5+ years of college	110013.22

In [211…

```
print("Respondents whose highest grade is 12 on average earn $38,957.76 ann
```

```
Respondents whose highest grade is 12 on average earn $38,957.76 annually.
```

Exercise 11

What is the average income of someone who graduated college ("4 years of college")? What does that suggest is the value of getting a college degree after graduating high school?

In [212…

```
print("Respondents who graduated college on average earn $75,485.05 annuall
```

```
Respondents who graduated college on average earn $75,485.05 annually.
```

In [213…

```
x = 75485.05 - 38957.76
f'Finding both those averages seems to suggest that the value of getting a
```

Out[213…

```
'Finding both those averages seems to suggest that the value of getting a c
ollege degree after graduating high school is $36527.29, holding all else e
qual.'
```

Exercise 12

What is the average income for someone who has not finished high school? What does that suggest is the value of a high school diploma?

In [214…

```
acs_emp
```

Out[214…

	age	educ	empstat	inctot
1	17	grade 11	employed	6000
2	63	4 years of college	employed	6150
5	50	grade 12	employed	50000
9	17	grade 12	employed	2000
10	47	n/a or no schooling	employed	18000
...	...	...	...	...
318995	67	grade 12	employed	125000
318999	33	4 years of college	employed	22130
319001	20	grade 12	employed	5000
319002	47	5+ years of college	employed	240000
319003	33	5+ years of college	employed	48000

148758 rows × 4 columns

In [215…

```
p = (32276.88 + 27592.65 + 30684.20 + 27171.91 + 23018.80 + 21541.69)/6
p = round(p, 2)
p
f'Respondents who have not finished high school on average earn ${p} annual
```

Out[215…

```
'Respondents who have not finished high school on average earn $27047.69 an
nually.'
```

In [216…

```
d = 38957.76 - p
d = round(d, 2)
f'This seems to suggest that the value of getting a high school diploma is
```

Out[216…

```
'This seems to suggest that the value of getting a high school diploma is $
11910.07, holding all else equal.'
```

Exercise 13

Complete the following table:

- Average income for someone who has not finished high school: $27,047.69
- Average income for someone who only completed 9th grade: $27,171.91
- Average income for someone who only completed 10th grade: $23,018.80
- Average income for someone who only completed 11th grade: $21,541.69
- Average income for someone who finished high school (12th grade) but never started college: $38,957.76
- Average income for someone who completed 4 year of college (in the US, this means graduating college): $75,485.05

Exercise 14

Why do you think there is no benefit from moving from grade 9 to grade 10, or grade 10 to grade 11, but there is a huge benefit to moving from grade 11 to graduating high school (grade 12)?

- Moving from grade 9 to grade 10 or grade 11 does not yield a certificate or diploma (*potentially interesting side note: in Germany it does! When you graduate grade 10, you could potentially get a diploma, when you decide to leave school, it's called "Mittlere Reife"*). However, when you graduate high school, even though you might only know marginally more than when you were in grade 11, you get a diploma and that might distinguish you in the eyes of an employer. This reminded me of the "markets of lemons" in economics, when there is asymmetric information. In this way maybe a diploma is an indicator for an employer to see that yes, the employee is qualified to finish high school (i.e. disciplined, smart, driven, etc) and ergo qualified to work, whereas when you finish only with grade 11, while you might know only a little less, you lack that drive to get a diploma and might signal you are not a qualified employee. Overall, however, there might be other mechanisms at work too, like confounding factors. People who finish high school and people who only finish grade 11 might not be comparable and there might be other factors that differentiate them, apart from having a diploma. If you do not finish high school you might overall be less driven, less intelligent or just strive for a different career. Nevertheless, the effect of a diploma as a signal seems to be a good starting point to explain the question at hand.