## Exercises[¶](https://www.practicaldatascience.org/html/exercises/Exercise_missing.html#Exercises)

### Exercise 1[¶](https://www.practicaldatascience.org/html/exercises/Exercise_missing.html#Exercise-1)

Today, we will be using the ACS data we used during out first pandas exercise to examine the US income distribution, and how it varies by race. Note that because the US income distribution has a very small number of people with *extremely* high incomes, and the ACS is just a sample of Americans, the far right tail of the distribution will not be very well estimated. However, this data should suffice for helping to understand wealth inequality in the United States.

To begin, load the ACS Data we used in our first pandas exercise. That [data can be found here](https://github.com/nickeubank/MIDS_Data/tree/master/US_AmericanCommunitySurvey). We’ll be working with US\_ACS\_2017\_10pct\_sample.dta.

### Exercise 2[¶](https://www.practicaldatascience.org/html/exercises/Exercise_missing.html#Exercise-2)

Let’s begin by calculating the mean US incomes from this data (recall that income is stored in the inctot variable). Store the result "EX2\_AVG\_INCOME".

### Exercise 3[¶](https://www.practicaldatascience.org/html/exercises/Exercise_missing.html#Exercise-3)

Hmmm… That doesn’t look right. The average American is definitely not earning 1.7 million dollars a year. Let’s look at the values of inctot using value\_counts(). Do you see a problem?

Now use value\_counts() with the argument normalize=True to see proportions of the sample that report each value instead of the count of people in each category. What percentage of our sample has an income of 9,999,999? Store that proportion (between 0 and 1) as "EX3\_SHARE\_MAKING\_9999999". What percentage has an income of 0? Store that proportion as "EX3\_SHARE\_MAKING\_ZERO".

### Exercise 4[¶](https://www.practicaldatascience.org/html/exercises/Exercise_missing.html#Exercise-4)

As we discussed before, the ACS uses a value of 9999999 to denote that income information is not available for someone. The problem with using this kind of “sentinel value” is that pandas doesn’t understand that this is supposed to denote missing data, and so when it averages the variable, it doesn’t know to ignore 9999999.

To help out pandas, use the replace command to replace all values of 9999999 with np.nan.

### Exercise 5[¶](https://www.practicaldatascience.org/html/exercises/Exercise_missing.html#Exercise-5)

Now that we’ve properly labeled our missing data as np.nan, let’s calculate the average US income once more. Save the result as EX5\_AVG\_INCOME.

### Exercise 6[¶](https://www.practicaldatascience.org/html/exercises/Exercise_missing.html#Exercise-6)

OK, now we’ve been able to get a reasonable average income number. As we can see, a major advantage of using np.nan is that pandas knows that np.nan observations should just be ignored when we are calculating means.

But it’s not enough to just get rid of the people who had inctot values of 9999999. We also need to know why those values were missing. Suppose, for example, that the value of 9999999 was used for anyone who made more than 100,000 dollars: if we just dropped those people, then our estimate of average income wouldn’t mean much, would it?

So let’s make sure we understand *why* data is missing for some people. If you recall from our last exercise, it seemed to be the case that most of the people who had incomes of 9999999 were children. Let’s make sure that’s true by looking at the distribution of the variable age for people for whom inctot is missing (i.e. subset the data to people with inctot missing, then look at the values of age with value\_counts()).

Then do the opposite: look at the distribution of the age variable for people who whom inctot is *not* missing.

Can you determine when 9999999 was being used? Is it ok we’re excluding those people from our analysis?

Note: In this data, Python doesn’t understand age is a number; it thinks it is a string because the original data has categories like “90 (90+ in 1980 and 1990)” and “less than 1 year old”. So you can’t just use min() or max(). We’ll discuss converting string variables into numbers in a future class.

### Exercise 7[¶](https://www.practicaldatascience.org/html/exercises/Exercise_missing.html#Exercise-7)

Great, so now we know why those people had missing data, and we’re ok with excluding them.

But as we previously noted, there are also a lot of observations of zero income in our data, and it’s not clear that we want everyone with a zero-income *should* be included in this average, since those may be people who are retired, or in school.

Let’s limit our attention to people who are currently working by subsetting to only employed respondents. We can do this using empstat. Remember you can use value\_counts() to see what values of empstat are in the data!

### Exercise 8[¶](https://www.practicaldatascience.org/html/exercises/Exercise_missing.html#Exercise-8)

Now let’s estimate the racial income gap in the United States. What is the average salary for employed Black Americans, and what is the average salary for employed White Americans? Store the results as "EX8\_AVG\_INCOME\_WHITE" and "EX8\_AVG\_INCOME\_BLACK".

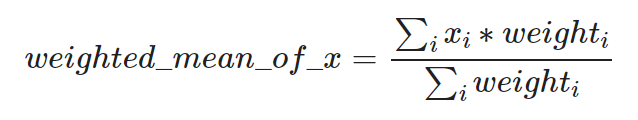
In percentage terms (between 0 and 100), how much more does the average White American make than the average Black American? Store the result as "EX8\_RACIAL\_DIFFERENCE". (Be careful to use the right reference group! Percentage differences aren’t symmetric.)

**Note:** these values are not quite accurate estimates. As we’ll discuss in later lessons, to get completely accurate estimates from the ACS we have to take into account how people were selected to be interviewed. But you get pretty good estimates in most cases even without weights – your estimate of the racial wage gap without weights is within 5% of the corrected value.

**Note:** This is actually an underestimate of the wage gap. The US Census treats Hispanic respondents as a sub-category of “white”, so in pooling what most Americans think of as “White” respondents (but which Census thinks of as “White, Non-Hispanic”) with Hispanic respondents (who tend to earn less), we get an underestimate of the average white salary in the US. But let’s ignore that nuance for the moment and just compare people who are coded as White and Black in the race variable).

### Exercise 9[¶](https://www.practicaldatascience.org/html/exercises/Exercise_missing.html#Exercise-9)

As noted above, these estimates are not actually *quite* correct because we aren’t using survey weights. To calculate a weighted average that takes into account survey weights, you need to use the following formula:



(As you can see, when  is constant for all observations, this just simplifies to our normal formula for mean values. It is only when weights vary across individuals that weights must be explicitly addressed).

In this data, weights are stored in the variable perwt, which is the number of people for which each observation is a stand-in (the inverse of that observations sampling probability).

Using the formula, re-calculate the *weighted* average income for both populations and store them as EX9\_AVG\_INCOME\_WHITE and EX9\_AVG\_INCOME\_BLACK

### Exercise 10[¶](https://www.practicaldatascience.org/html/exercises/Exercise_missing.html#Exercise-10)

While all ethnic distinctions are socially constructed, and so on some level these distinctions are all deeply problematic, as previously discussed the way White is coded in the race variable is inconsistent with what most Americans think of when they hear the term “White”. For most people, “White” is though of as a category that is distinct from being Hispanic or Latino (categories which are also usually conflated in American popular discussion). With that in mind, most researchers working with US Census data split “White” into “White, Hispanic” and “White, Non-Hispanic” using race *and* hispan.

So now calculate the weighted average income gap between *non-Hispanic* White Americans and Black Americans. What percentage (between 0 and 100) more do employed White non-Hispanic Americans earn than employed Black Americans? Store as "EX10\_WAGE\_GAP".

### Exercise 11[¶](https://www.practicaldatascience.org/html/exercises/Exercise_missing.html#Exercise-11)

### Is that greater or less than the difference you found in Exercise 8? Why do you think that’s the case?