GR 5072: Week 4 Activity with pandas

We will be replicating the findings in the "Estimating the Labor Market Returns of the General Educational Development (GED)" paper, which requires us to clean and prepare the data for analysis before we finally fit the relevant regression models. We will go through this step to clean the data:

Data analysis steps:

1. Check for missing data, applying listwise deletion if required
2. Prepare variables for analysis using a multiple regression methodology
    - For categorical variables, make dummy variables as required
    - For numeric variables, fix or transform as necessary
    - **Don't forget to always double check your work!**
3. Explore data / check descriptive statistics
4. Fit models and interpret results

Step 1: We will apply listwise deletion, so drop the following from the sample:

- 2011 earnings <= 0 (F3ERN2011 <= 0)
- 2011 earnings >= 200000 (F3ERN2011 >= 200000)
- Hours worked in 2011 < 0 (F3C02 < 0)
- Drop those who were enrolled in school in 2011 since their estimate for earnings will be confounded (F3JUNEDSTAT < 3]
- Drop those with missing gender (BYS14 < 0)
- Drop those missing ethnicity (BYRACE < 0)
- Drop those missing the Cognitive ability test score (BYTXCSTD < 0)
- Drop those missing region (F3REGION < 0)
- Drop rows who are missing the BYP61 variable (BYP61 < -.25)
- Drop those missing maternal education (BYMOTHED < 0)
- Drop those missing on the job training (F3B35 < 0)

Step 2: Create dummies for the following variables:

- BYSEX is 2=female, 1=male (create a new dummy "female" which is 1 = female, 0=male)
- BYRACE: We need the following dummies: white, black, Asian, Hispanic, and other (see the pdf for how these are coded)
- BYP61, 1=absence of biological parent from home
- BYMOTHED, we just need 8 for the 8 different categories
- F3REGION
- F3EVERDO: this variable is 1 if someone dropped out. Lets make a dummy called "high school graduation" that is 1 if they didn't drop out / graduate
- F3EVRGED, make a dummy "ged" that is 1 if this variable is 1 (I know its redundant, but this variable name is shorter/easier to use)

Step 2: Create the following numeric column:

- Years of post-secondary education, which depends on F3ATTAINMENT as follows:
    - 1 if F3ATTAINMENT = 4
    - 2 if F3ATTAINMENT = 5
    - 4 if F3ATTAINMENT = 6
    - 5 if F3ATTAINMENT = 7
    - 6 if F3ATTAINMENT = 8
    - 8 if F3ATTAINMENT = 10
    - Otherwise this is 0

Step 3: Explore data, look at descriptive statistics

Step 4: Fit the two regression models. One has a DV of income in 2011, the other has hours worked in 2011