Working with Cohorts in Excel
Tutorial Script
2023-24 Data Collection

As an IPEDS keyholder, you will likely work with census files and other frozen data files to complete many of the IPEDS survey components. How you receive the data files depends on your specific institution. Some IPEDS keyholders have direct access to the data in the Student Information System and extract the frozen data files themselves, while others work with the central IT office or another department to obtain the data files.

The files used in this video came from the example institution's IT office. The IT office supplied the same data in two different formats; comma delimited, also called comma separated values or CSV files; and the tab delimited file, which is also called tab separated values or TSV files. The nice thing about CSV and TSV files is that a variety of software programs can open these file formats.

We will open both formats in Microsoft Excel to highlight differences in how the two formats open. Select File, Open and then Browse to the location of the files.

Select “All Files” to see the two file formats.

First, I will open the tab-delimited file by selecting the file and clicking OPEN.

This type of file uses tabs to mark, or delimit, the end of each data field, where comma delimited files uses commas as the marker. Since Excel recognized the file type, “delimited” is already selected, so click next.

Again, Excel recognizes the data as tab delimited so “tab” is already checked, indicating tab delimited data. Click Next.

Regardless of using tab delimited or comma delimited files, pay special attention in this step and double check whether your Key ID might have leading zeros. If there are leading zeros and you leave the data type as general, excel will automatically delete the leading zeros since it treats numerical information in the general format type fields as numbers. If your data has leading zeros you want to retain, change this from general to text to retain the leading zero. This file does not have a Key ID with leading zeros, so I’ll leave the general option selected and click finish. You can now see the data file open in excel.

The next file that I'll open is comma delimited but excel will automatically open the csv file without having to go through the steps from the tab delimited file.

This is the sample student data file for this exercise.

Understanding your data using the Data Dictionary. If you receive data from the IT office to use in your IPEDS submission you will likely also receive a data dictionary file. If not, you should work directly with the staff that supplied the data to create a data dictionary.

While I won’t cover data dictionary creation in this training, there are many resources available from AIR and other sources on this topic.

The data dictionary allows you or other staff to understand the raw data in the related data file. For example, some values in a data file may be codes, while others are descriptors.

Now that we have explored how to open comma delimited or tab delimited files using Excel, the actual data contained in the files can be explored. The data dictionary for this file will be helpful for identifying the data fields and the meaning of the values in each data field.

Open the file labeled Training Data Dictionary.

As you can see, the first field of data in our ample data file is IPEX ID. This field contains the unique ID for each student at many institutions this is also called Student ID.

Other fields include year, semester codes, registration status, level, student gender, and race, along with the degree intent codes and enrollment status, as well as Pell award, and federal loan figures. Scroll to right to see the final field, exclusion eligibility.

The data dictionary can be very helpful for understanding the data contained in the actual data files. If you have direct access to pull data from the Student Information System, you should create a data dictionary for the files you save.

To explore the data definitions, we will go to the data dictionary and lookup each of the elements by name.

As you can see the field values for Semester are 1, 2, 3, and 4. By using the data dictionary you can see that the value 4 listed in the data file is a coded value that represents the fall semester.

Now we will explore the values for “Registration status.”

There are codes such as HS for high school and CS for continuing student. As well as FH, which is the example institution’s code for first-time students that graduated high school in the 12 months prior to enrolling at the institution and FF, which is the code for first-time undergraduate students at this institution that did not enroll within 12 months after high school.

Basically, FH is a first-time student straight out of high school student or within the past year and FF is a first-time student who has been out of high school more than a year.

“Level” has descriptors like sophomores, juniors, seniors.

“Gender” is indicated as M for male and F for female. The “Race” field contains codes such as H for Hispanic, A for Asian, and B for Black.

One of the fields that is not always intuitive to those newly working with student data is degree intent. In essence, this is the intention of the student when they enrolled in this program of study. Some students may enroll with the intention of not graduating, as they are only interested in taking a few courses to assist in career advancement or for another reason.

Most institutions would have a code here for non-degree seeking students. But since this example institution only enrolls degree-seeking and certificate-seeking students, our data file here only contains degree-seeking and certificate-seeking students, and the resulting data dictionary only includes the codes for those students. In this case, the values of 0, 2, and 4.

The data dictionary listing for degree intent lists 0 for students intending to receive a certificate, 2 for those seeking an associate's degree, and 4 for four a bachelor 's degree.

Enrollment status is listed as “P” for part-time students and “F” for full-time students. Pell award and federal loan amounts are also listed. The “Exclusion Eligible” field lists whether a student originally

included in the IPEDS cohort can be excluded from certain calculations based on death, military, or church service, or serving in the Peace Corps.

Now I'll demonstrate how to do some quick data analysis using the pivot table function in Microsoft Excel 2016.

A common question asked of IPEDS Keyholders by others at an institution is "how many part-time versus full-time students are new to the institution this term?" An easy way to answer this question as well as others like race/ethnicity, gender, and/or registration status comparisons is to create a pivot table in Excel.

To start the creation of a pivot table, go to insert and click pivot tables, then OK. Drag ID to values and a count of Student IDs will be displayed, in this case 31,445. To look at full -time Cohorts versus part-time students bring the appropriate data field, enrollment status, to the row label.

Next, to view which students are full-time as well as first t-time in college, I'll bring registration status down to filter. Under "report filter" multiple items can be selected.

In this case, I select FF for full-time students that graduated high school more than 12 months ago and FH for full-time students that graduated high school within the last 12 months. Then select OK. As you can see, there are 3,523 first-time college students comprised of 2,665 full-time first-time students and 858 part-time first-time students.

To disaggregate this data by gender, male/female, we can go back and choose gender by bringing it to the column label. To make the data table a little easier to read, center the data. Now you can see the male versus female students in each of the full-time and part-time categories.

Another common question answered with this type of data is "how many students received Pell grant funding and the average amount of Pell funding students received.

So, using the current setup, we can adjust the pivot table by removing "Count of IPEXID" from the Sum Values section and replace it with "Pell Award". Then click on the "Pell Award" label to select the "Value Field Settings "option. Select "Average" in the new dialog window and click OK. As you can see, the data in the pivot table needs to be cleaned up slightly. Click the "Value Field Settings "on the "Pell Award" label and then click the "Number Format" button in the bottom left of the same dialog window that pops up and then currency. This will change the pivot table to display the average amount of Pell award for our full-time and part-time students disaggregated by gender.

A key point to be aware of when dealing with datasets that contain values you want to average, is that Excel will handle blank data fields differently than a field that contains a 0. In a nutshell, fields that contain zeros are included in the average, whereas fields that are blank are not treated as zero and not included in the numerator or the denominator for the calculation of the average. So, based on the data file only having blank fields for Pell Award, rather than 0, the resulting pivot table is a display of the average amount of Pell Award funding for students that received Pell Awards and not all students. If you want to display the average Pell Award funding for all students within the part-time/full-time categories and gender breakdown, the data file would need to be recoded to replace all blank fields in this column with a numeric value of zero.