Removing Duplicate Values in Excel

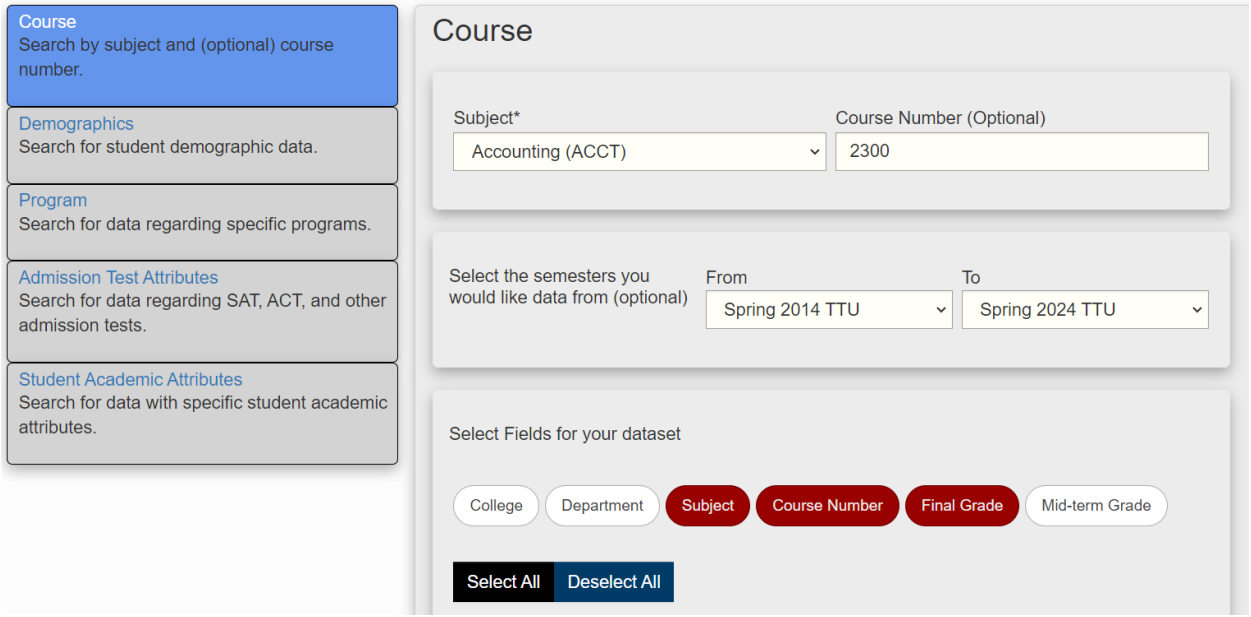
Problem Statement:

Student data downloaded from Red Raider Registry can contain duplicate values. These duplicate Unique Identifiers can result from searching courses over time or result from multiple course attempts by a student. This tutorial will provide steps on how to identify if the data set contains duplicate values and how to remove those duplicate values from the data set.

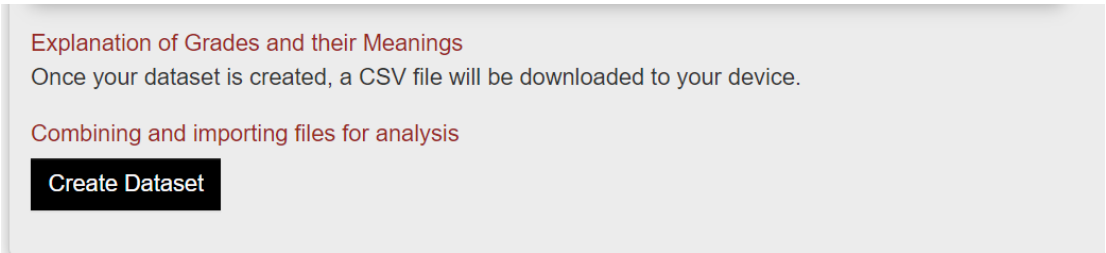
This example shows how to review and identify duplicate unique identifiers (students) in downloaded course data over the last ten (10) years.

Step 1- Download Dataset from Red Raider Registry

A. Select the fields required for the dataset

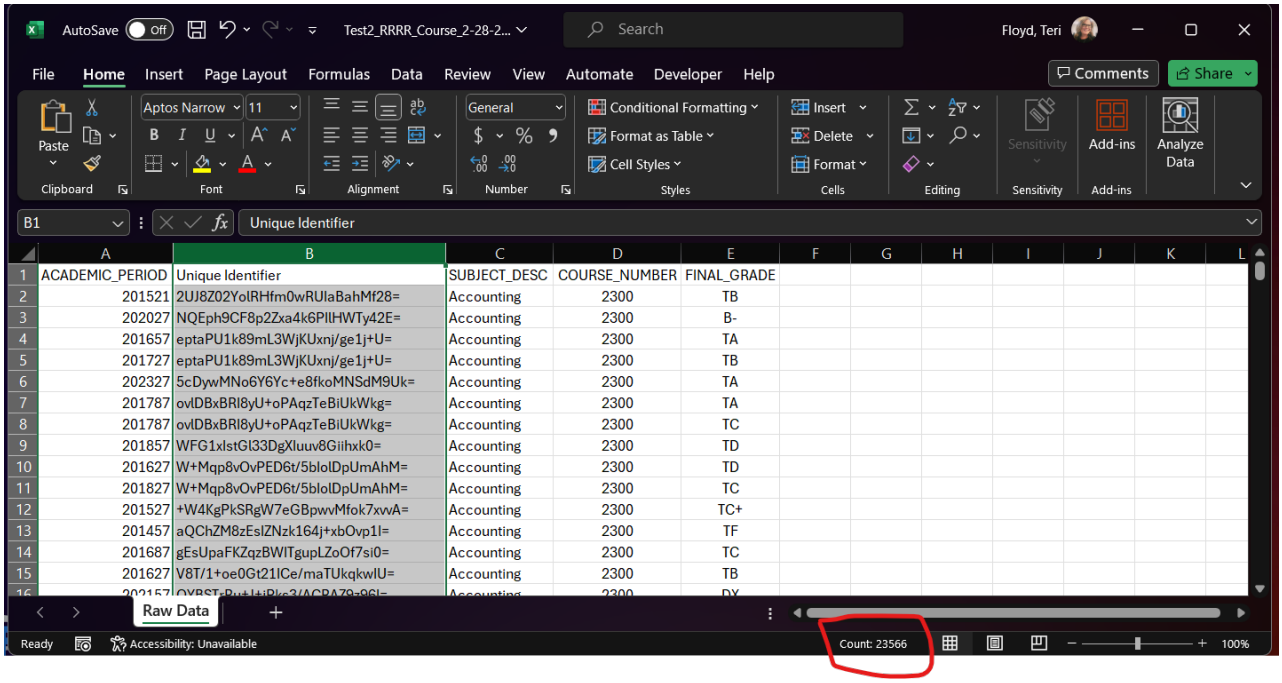


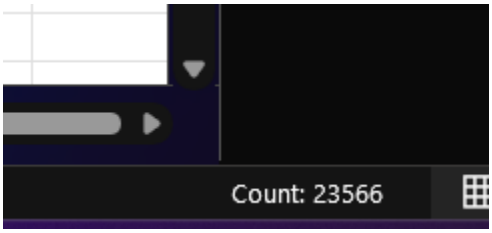
B. Click on Create Dataset



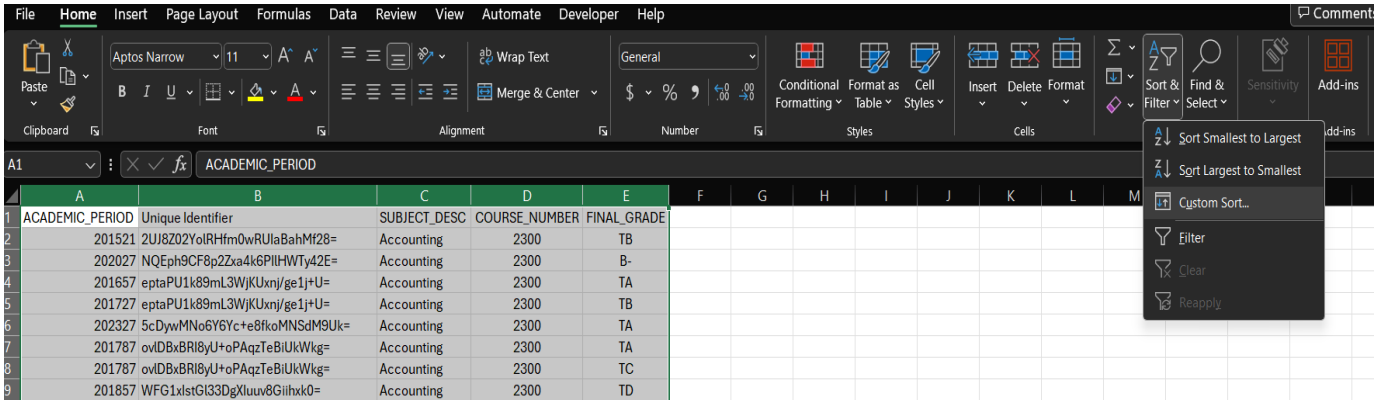
Step 2- Review Dataset and Sort

A. Review the row count. Click on column B to see how many rows were counted. In this example, there is a count of 23,566 rows. Note that this includes the column names in row one, so there are 23,565 rows of student data.

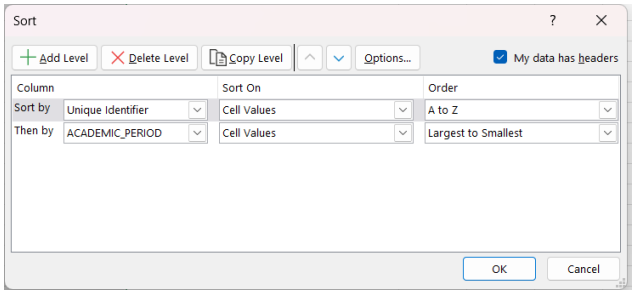




B. To identify students that have retaken a course it is important to sort the data. Select all columns in the worksheet and then select Home > Editing > Sort & Filter > Custom Sort.



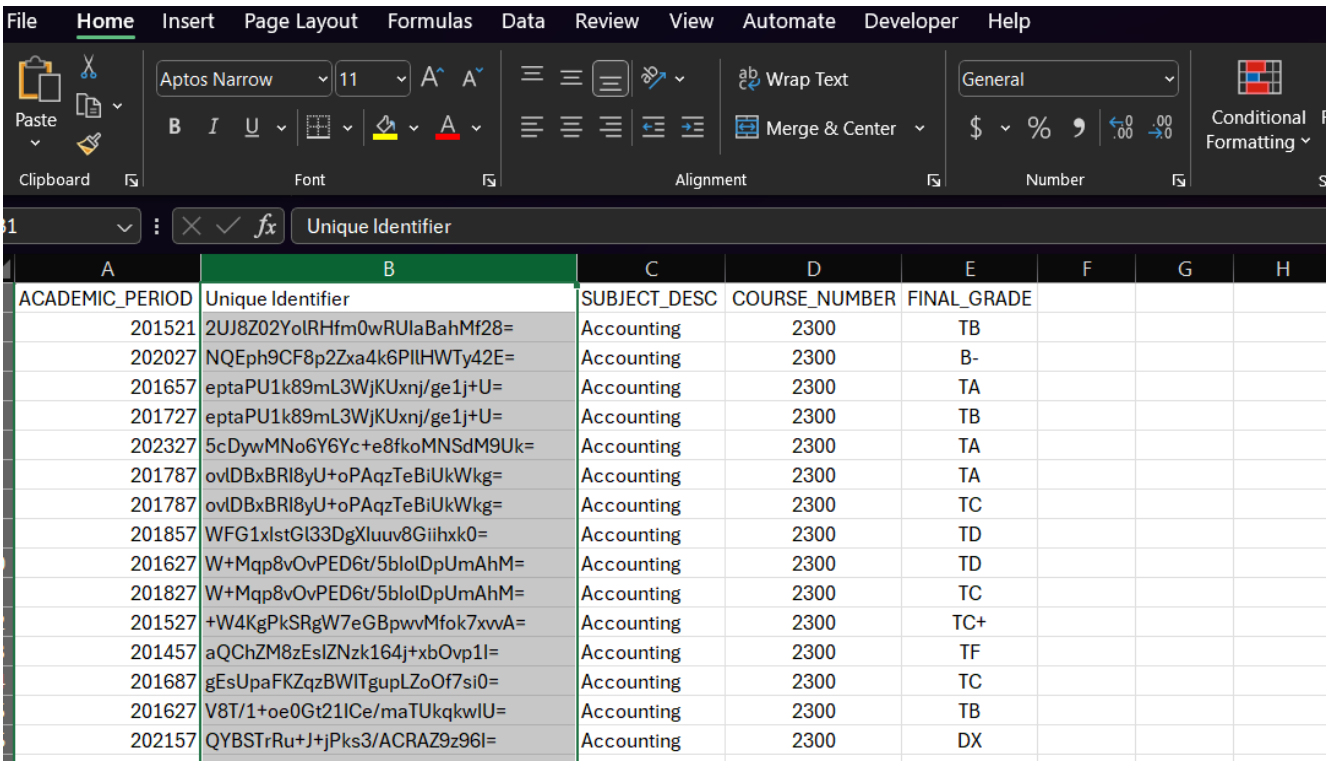
C. Sort by unique identifier (A to Z) and academic period in descending order (largest to smallest). The reason for sorting first by alphabetical order then by academic period in descending order is to identify the last TTU term that a student attempted the class. That way once duplicate values are removed, the data shows the most recent student information.



D. The data is now displayed in the proper order to identify duplicate values (repeat students) and when they took the course (most recent shows first).

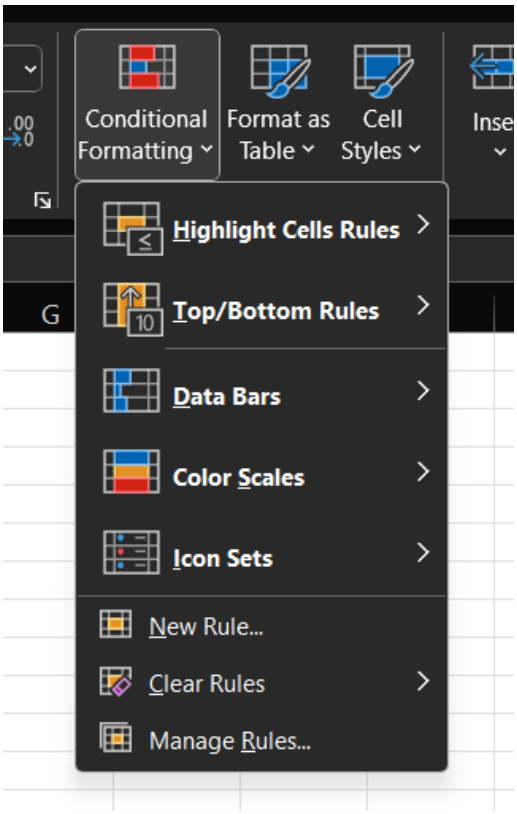
Step 3- Identify Duplicates in the Data Set

A. Select the cells you want to check for duplicates. In this example, we are selecting the Unique Identification column to check for duplicate values.

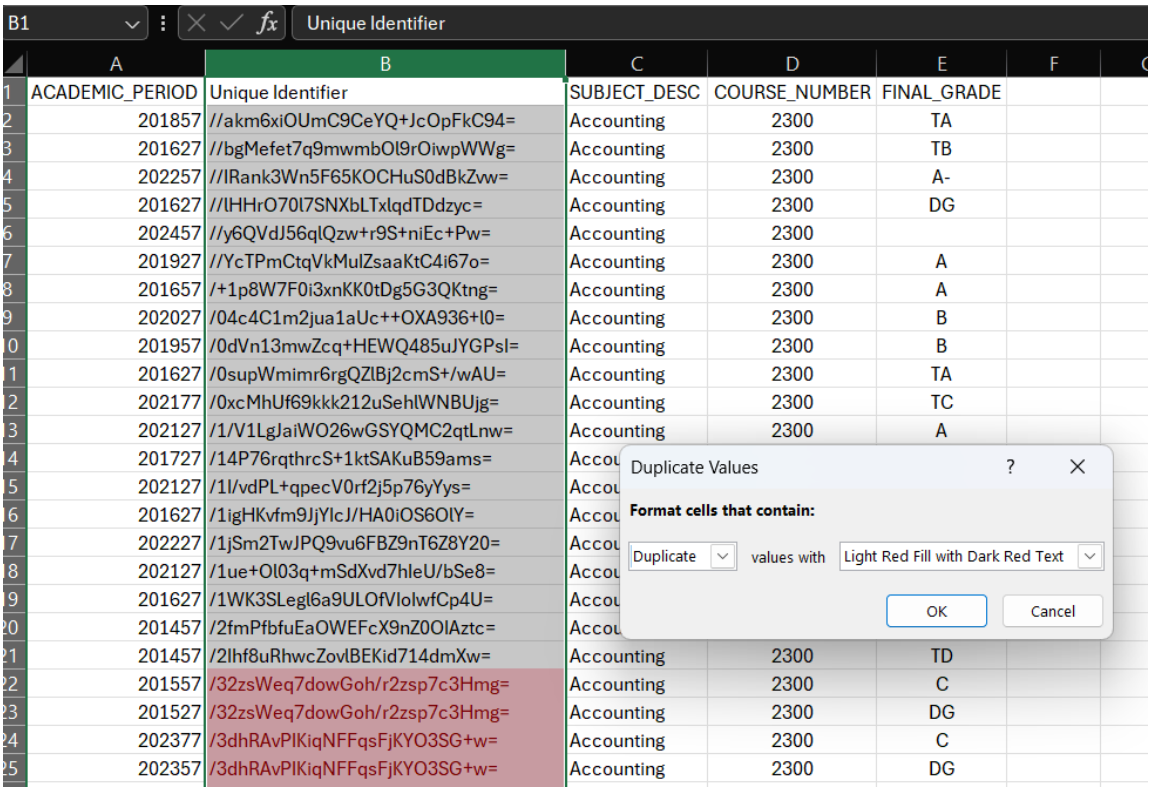


	A	B	C	D	E	F	G	H
	ACADEMIC_PERIOD	Unique Identifier	SUBJECT_DESC	COURSE_NUMBER	FINAL_GRADE			
	201521	2UJ8Z02YolRHfm0wRUlaBahMf28=	Accounting	2300	TB			
	202027	NQEph9CF8p2Zxa4k6PIlHWTy42E=	Accounting	2300	B-			
	201657	eptaPU1k89mL3WjKUxnj/ge1j+U=	Accounting	2300	TA			
	201727	eptaPU1k89mL3WjKUxnj/ge1j+U=	Accounting	2300	TB			
	202327	5cDywMNo6Y6Yc+e8fkoMNSdM9Uk=	Accounting	2300	TA			
	201787	ovlDBxBRl8yU+oPAqzTeBiUkWkg=	Accounting	2300	TA			
	201787	ovlDBxBRl8yU+oPAqzTeBiUkWkg=	Accounting	2300	TC			
	201857	WFG1xlstGl33DgXluuv8Giihxk0=	Accounting	2300	TD			
	201627	W+Mqp8vOvPED6t/5blolDpUmAhM=	Accounting	2300	TD			
	201827	W+Mqp8vOvPED6t/5blolDpUmAhM=	Accounting	2300	TC			
	201527	+W4KgPkSRgW7eGBpwvMfok7xvvA=	Accounting	2300	TC+			
	201457	aQChZM8zEsIZNzk164j+xbOvp1l=	Accounting	2300	TF			
	201687	gEsUpaFKZqzBWITgupLZoOf7si0=	Accounting	2300	TC			
	201627	V8T/1+oe0Gt21ICe/maTUkqkwlU=	Accounting	2300	TB			
	202157	QYBSTrRu+J+jPks3/ACRAZ9z96l=	Accounting	2300	DX			

B. Select Home > Conditional Formatting > Highlight Cells Rules > Duplicate Values



C. In the box next to values with, pick the formatting you want to apply to the duplicate values, and then select OK.

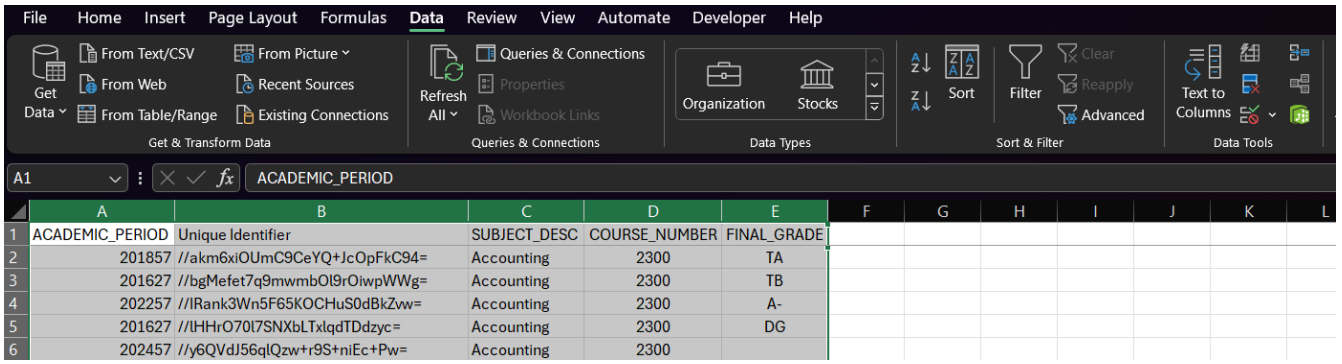


	A	B	C	D	E	F
1	ACADEMIC_PERIOD	Unique Identifier	SUBJECT_DESC	COURSE_NUMBER	FINAL_GRADE	
2	201857	//akm6xiOUmC9CeYQ+JcOpFkC94=	Accounting	2300	TA	
3	201627	//bgMefet7q9mwmbOl9rOiwpWWg=	Accounting	2300	TB	
4	202257	//IRank3Wn5F65KOCHuS0dBkZvw=	Accounting	2300	A-	
5	201627	//lHHrO70l7SNXbLTxlqdTDdzyc=	Accounting	2300	DG	
6	202457	//y6QVdJ56qlQzw+r9S+niEc+Pw=	Accounting	2300		
7	201927	//YcTPmCtqVkMulZsaaKtC4i67o=	Accounting	2300	A	
8	201657	/+1p8W7F0i3xnKK0tDg5G3QKtng=	Accounting	2300	A	
9	202027	/04c4C1m2jua1aUc++OXA936+l0=	Accounting	2300	B	
10	201957	/0dVn13mwZcq+HEWQ485uJYGPsl=	Accounting	2300	B	
11	201627	/0supWmimr6rgQZlBj2cmS+/wAU=	Accounting	2300	TA	
12	202177	/0xcMhUf69kkk212uSehlWNBUjg=	Accounting	2300	TC	
13	202127	/1/V1LgJaiWO26wGSYQMC2qtLnw=	Accounting	2300	A	
14	201727	/14P76rqthrcS+1ktSAKuB59ams=	Accou			
15	202127	/1l/vdPL+qpecV0rf2j5p76yYys=	Accou			
16	201627	/1igHKvfm9JjYlcJ/HA0iOS6OlY=	Accou			
17	202227	/1jSm2TwJPQ9vu6FBZ9nT6Z8Y20=	Accou			
18	202127	/1ue+Ol03q+mSdXvd7hleU/bSe8=	Accou			
19	201627	/1WK3SLegl6a9ULOfVlolwfCp4U=	Accou			
20	201457	/2fmPfbfuEaOWEFcX9nZ0OlAztc=	Accou			
21	201457	/2lhf8uRhwcZovlBEKid714dmXw=	Accounting	2300	TD	
22	201557	/32zsWeq7dowGoh/r2zsp7c3Hmg=	Accounting	2300	C	
23	201527	/32zsWeq7dowGoh/r2zsp7c3Hmg=	Accounting	2300	DG	
24	202377	/3dhRAvPIKiqNFFqsFjKYO3SG+w=	Accounting	2300	C	
25	202357	/3dhRAvPIKiqNFFqsFjKYO3SG+w=	Accounting	2300	DG	

Step 4 – Remove the Duplicate Values

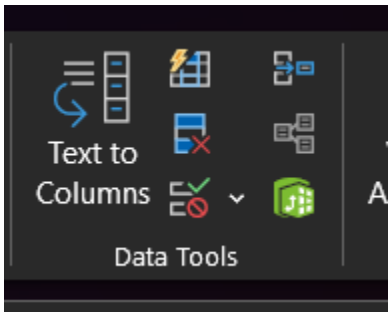
Please note that when you use Remove Duplicates, the duplicate data is permanently deleted. It is good practice to copy the original data to another worksheet before deleting duplicates, so you don't accidentally lose information.

A. Select the range of cells that has duplicate values you want to remove.

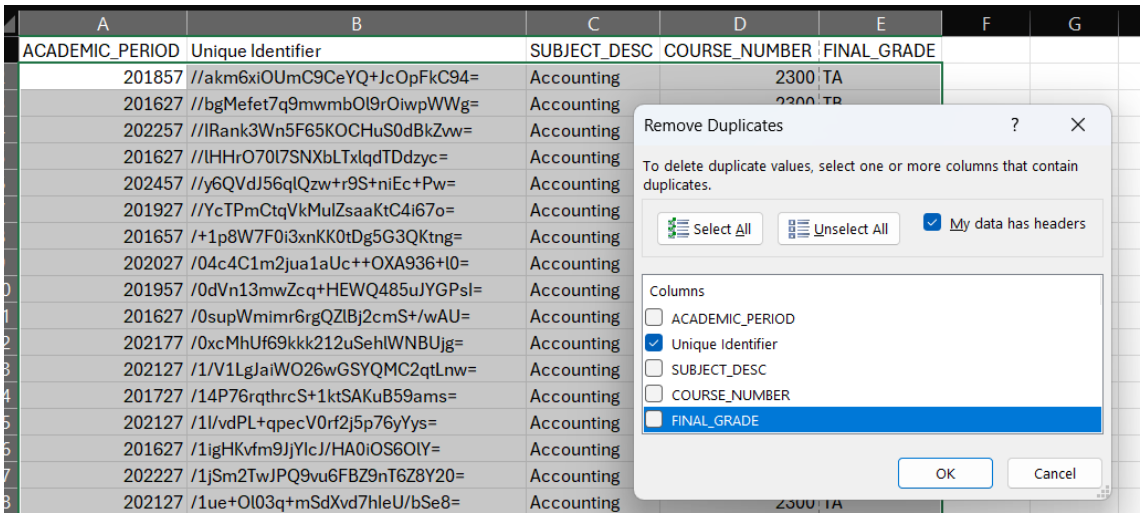


	A	B	C	D	E
1	ACADEMIC_PERIOD	Unique Identifier	SUBJECT_DESC	COURSE_NUMBER	FINAL_GRADE
2	201857	//akm6xiOUmC9CeYQ+JcOpFkC94=	Accounting	2300	TA
3	201627	//bgMefet7q9mwmbOl9rOiwpWWg=	Accounting	2300	TB
4	202257	//IRank3Wn5F65KOCHuS0dBkZvw=	Accounting	2300	A-
5	201627	//lHHrO70l7SNXbLTxlqdTDdzyc=	Accounting	2300	DG
6	202457	//y6QVdJ56qlQzw+r9S+niEc+Pw=	Accounting	2300	

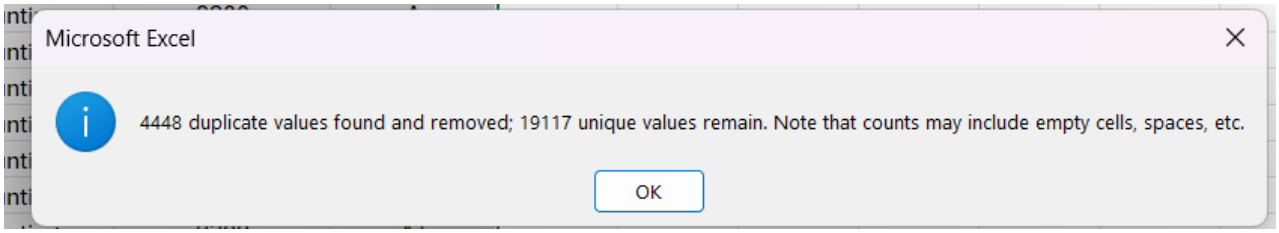
B. Select Data > Data Tools > Remove Duplicates



C. Click on the Remove Duplicates icon and select the columns that contain duplicates. In this test, we wanted to remove duplicate Unique Identifiers so that was the only column checked. Click OK.



D. Excel will remove all duplicates that occur after the first occurrence and a notification will pop up indicating how many duplicates were found / removed and how many unique values remain. Click OK.



E. Now the data has been deduplicated, it is another good practice to double check row counts to ensure that there was no additional data lost.

23,565 rows of raw data (contains duplicates)
- 4,448 duplicate values

19,117 unique values