Auditmetrics AI v6.3

Getting Started and Minicourse

A step by step learning exercise

Companies need to learn how to develop actionable strategies by data mining the troves of data they are collecting. AI guided statistical audits can benefit small business in many ways including determining customer actions, simplifying their processes, and decreasing levels of risk.

There is available for distribution several test data files for further practice that can be obtained by contacting us at info@auditmetrics.com

Changes in Version 6 Software

Version 5 data access architecture was based on a direct read of MS Excel spreadsheets and MS Access relational databases. Though this design is user friendly, over time it proved to be problematic. As we expanded our reach to larger data sets, version 5 proved to be untenable.

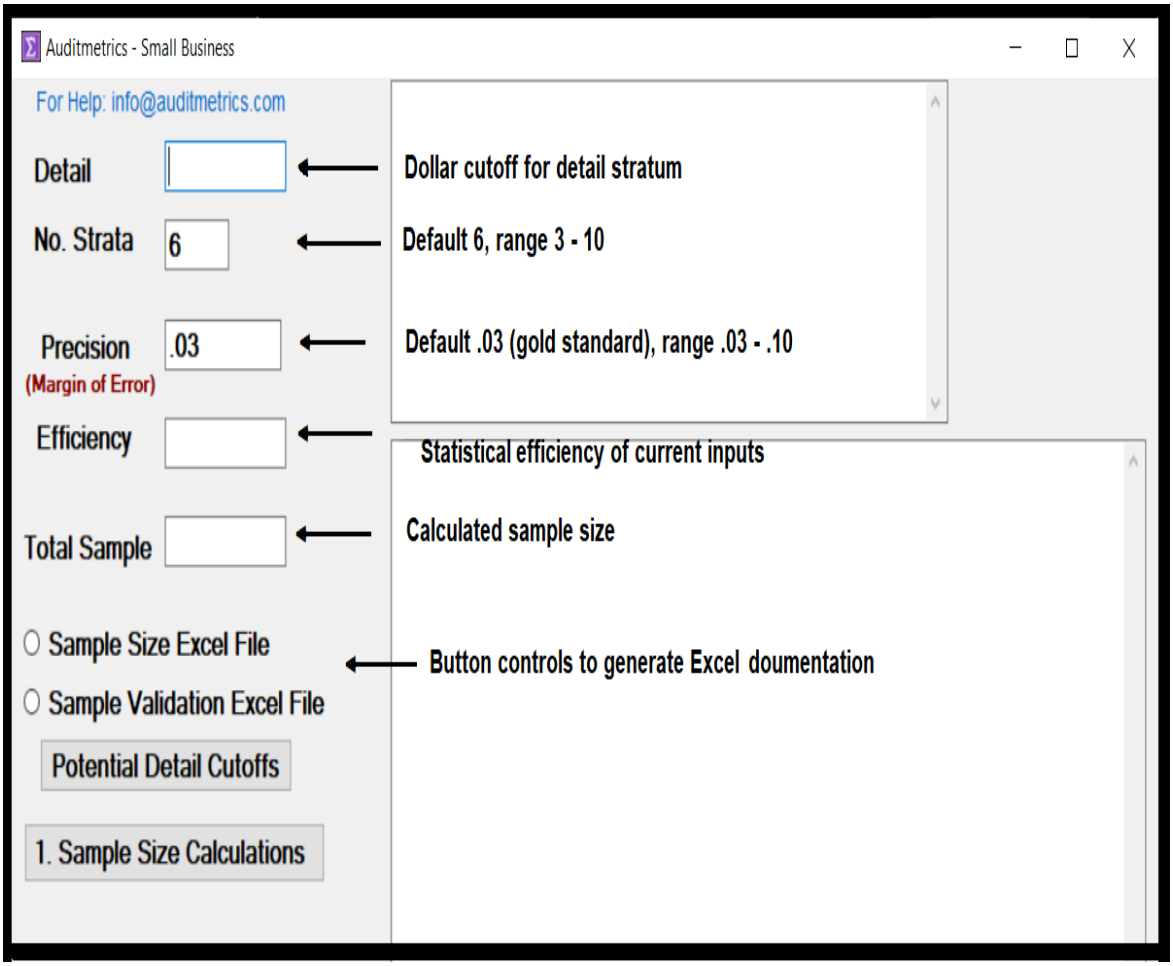
With large data sets the V5 original architecture turned out to have technical and practical limitations. Excel spreadsheet size limitation is a million records. So as we needed to access accounts with several million records, Excel as a direct data read became untenable. Of course an Access relational database can fill the bill but it has some performance limitations. The user friendly nature of setting up tables and data relationships does come with considerable operational overhead. For example, sampling an account of 10 million records considerably slows processing performance.

Large corporations are not subject to such limitations. They can scale up their servers with more sophisticated and efficient server side data handling software such as Oracle and MS SQL Server. But the primary goal of Statistical Audit–AI is to be a resource for small businesses. That more than likely means operations on a desktop or laptop computer.

Another problem is upgrades of MS Office and Windows do come with issues. Going from an older version to an upgrade means at best minor changes in coding but at times whole segments of code eventually becoming outdated. We therefore decided to get back to basics and use a standardized international data structure. All major database software and statistical software can generate that same common universal data structure, text files. Text files are the most efficient data source for input-output (IO) functions.

With that conversion, auditmetrics V6 can process 5 million records in a matter of seconds on a laptop greatly outperforming V5. This allows greater flexibility, quicker action, better privacy and less cost for small businesses to monitor customers and markets.

Text files are sometimes called ASCII files, American Standard Code for Information Interchange. ASCII is the traditional name for the encoding system; the Internet Assigned Numbers Authority (IANA) prefers the updated name **US-ASCII**.



In this document is listed a step by step process of installing Auditmetrics and using test data to actually implement a statistical audit. Resources needed are:

MS Office Excel and Access
Windows NotePad text editor
or one of your choice

Be sure to go through ALL of the steps. Failure to do so may leave out critical information.

Those statements that precede by a "!" are critical points to remember

1. Installing Auditmetrics

To install on a desktop download from Microsoft Store OR go to the install flash drive and click on the windows installer either SetupSB.msi or SetupPro.msi. Once installed, for SB V6.4 an icon will be placed on the desktop. If you are on a company network it is best to contact IT.

2. Starting Auditmetrics

Click on the Auditmetrics icon and examine the main screen. You will see three data entry boxes: Detail, No. Strata and Precision. These are the basic inputs in designing a random sample for the statistical audit. Detail is the stratum of the largest dollar volume that is examined at 100% of the transactions. No. Strata is the total number of strata segments. Precision is equivalent to what the pollsters call margin of error. The default is 3% which is the gold standard for precision. In technical terms you are indicating that you want to be 95% confident that an estimate from a sample would be within 3% of the true book value. 95% confidence means that out of 100 random samples, 95 would be within 3% of the true book value. It is your mix of these three inputs that determine the required random sample size.

3. Test Data – Excel and Text File

There is included a test file, "Test_Population.xlsx". It is the book of transactions we will use to obtain a random sample. First, open the Excel file. The first row of the spreadsheet contains variable names. The column named amount is the transaction in which we are interested. Auditmetrics requires only four variables. Any other variables are relevant for a particular audit but not necessarily required by Auditmetrics. There is also a test Access file that is discussed in Appendix I in the text book..

Auditmetrics requires four variables, if not present, an error message will be displayed. They are indicated in red on the spreadsheet.

1. **Amount** – The transaction of interest in the analysis.

2. **Absamt** – Absolute value of each transaction. This variable **must** be sorted in ascending order. This is to handle credits.

3. **Transaction_ID** – an identifier for each transaction of the account, in this case it is a record count.

4. **DataSet** – A name to identify this dataset, valuable for internal controls e. g. date and other account info .i.e AcmeAcct020519

5. **Primary Key- Optional** – If a dataset is from a relational database with a primary key that links the various data tables, it is prudent to include this variable in the audit population to be sampled.

Variables 1 to 4 must be in the data file and spelled exactly as above. Any other variables are those relevant for the specific audit to be conducted.

With the required variables above, letter case and order do not matter and absamt **must** be sorted in ascending order. If a dataset is from a relational database with a primary key that links various data tables it is prudent to include it in the total book to be sampled. In terms of statistics nomenclature, total book is the "audit population" from which a sample is to be derived.

Transaction_ID has value when the dataset is the merging of several data sources. For example, if you have two data sets of 1000 each and use MS Access to merge them into one file then Transaction_ID 1 to 1000 is from the first dataset and 1001 to 2000 from the second dataset. Transaction_ID also has value when filing with the IRS. For more details review *Appendix IV – Random Sampling and IRS Directives* in the book.

4. Data Input – The Tab Delimited Text File

A spreadsheet is not useful for data processing unless you have specialized software for that purpose. Such software tends to be inefficient with several limitations that is explained in the book. However it is a simple matter to save the Excel file as a tab delimited text file which is universally accepted by a multitude of products including all major database and statistical software. For Excel It is just a simple matter of using "save as" and select "tab delimited text" for file output.

To look at the text file you should use a text editor. You can use your favorite text editor or Notepad that is available in all versions of Windows. If you are using the professional version of Auditmetrics and dealing with millions of records, it may be wise to invest in a more robust text editor such as IDM's Ultraedit®. It has a nice feature of making tabs in the tab delimited file visible.

> **! Transfer Excel test data to a Tab Delimited text file using "save as". Auditmetrics can handle all Excel currency formats including parenthesis for negative numbers. The only exception is negative numbers in red type only, which will not transfer as negative.**
>
> **!When sharing data between several software products it is best to use Excel's General format for amount and absamt.**

Auditmetrics has a powerful parser to handle input data. A problem may arise when sharing data with other software. Each may have its own quirks. The best policy, while in Excel, would be to always format amount and absamt as *General* rather than currency or accounting format. Auditmetrics can handle these formats but it is best to keep things simple.

5. Let's Get Started

Create on your computer a folder called Auditmetrics_Test and place the test data file. You can park your files of interest on any folder you wish to create. Click on the Auditmetrics icon to get started. Before calculating sample size you need three input values that are displayed on the screen:

Detail Cutoff

The first step is to determine high end statistical outliers, or in the terminology of stratified audit sampling, the "detail stratum." This is the stratum in which one does not rely on a sample but reviews 100% of all transactions. Eliminating the largest transactions from sampling results in a reduction of the variability (standard error) of the remaining transactions from which a sample will be drawn. This enhances statistical efficiency in that the detail stratum allows a direct review of all transaction with the greatest economic impact.

To start, click on the tab **"Potential Detail Cutoff".** AI will provide a value for detail cutoff. A rule of thumb is that the detail stratum should represent approximately 1/3 of total dollar volume but Auditmetrics uses a more statistical analytical approach and for this data it is $7,500.

Also displayed are the upper percentile rankings of dollar volume. The auditor should initially spend time getting a sense of the distribution characteristics of the account to be audited. The auditor can vary detail cutoffs to determine if there is a more efficient input combination. The percentile rankings on the screen help in getting a sense of where the initial cutoff is located. As we cover the other inputs in determining proper sample size, you will find that you may get better results by tweaking all inputs.

> **!** ***Professional Version, Auditmetrics-AI, has a button in the upper right hand corner that allows the "Potential Detail Cutoff" tab to also display Benford Law's first digit and second digit assessment. It is a useful forensic accounting tool to detect possible fraud.***

You will also notice that what is also displayed is the interquartile range. This is the range of dollar amount that contains the middle 50% of the total audit population. Of the 25,152 transactions one half are between $14 and $205, 25% less than $14 and 25% above $205. This will give the auditor an additional detailed look of sample segments for possible errors or other potential problems. Each Segment may represent different products or services that require different benchmarks, staffing, marketing needs and monitoring methods in guiding performance.

When a file to be sampled is primarily dollar transactions it is a survey of the flow of dollars. What about customers or clients? When this basic unit is aggregated then the analysis is an economic impact assessment. The auditor can look at different market segments based on the level of economic activity. If the file to be sampled contains other data such as account id or a primary key from a relational database then the analysis can have access to other important data such as zip code and customer demographics. The auditor can then segment the audit population into different market segments. Those segments can be geographic, demographic or level of economic activity.

Of course the statistical audit depends on a validly drawn sample. Critical is how representative is that sample? Auditmetrics–AI statistical analytics guides you through the complex mathematics to assure such proper sampling. The goal is for you to have a sample that meets established statistical standards of the IRS, AICPA and the Multistate Tax Commission. Yes, anyone can draw a random sample but will it hold up under scrutiny? Will it provide tightly focused statistical estimates or

estimates that act erratically? Valid Statistical analytics can be assured by an efficient sample by means of the Auditmetrics AI assistance operating in the background.

The Number of Strata

Stratification is the process of dividing the population of transactions into segments (strata) based on a certain characteristic. In sampling based on dollars, one would stratify the population based on the dollar amount of the transaction. A stratified random sample will yield more precise results than an unrestricted random sample of the same size. Six strata is a default**.**

Precision

The default precision or margin of error on the screen is 3%. If the audit is for a formal submission to the IRS or state revenue agency, precision gold standard would be 3%. If you are conducting an internal audit and just want to get a preliminary look at the data then choose less precise values such as 5% or 7%. This will result in a smaller sample size.

Do a run with detail 7500 with precision and number of strata defaults. Then
Explore the interaction of detail, precision and number of strata impact on sample size.

Efficiency Factor

You should notice on the screen the measure "Efficiency Factor". The statistical issues surrounding this measure are better left to the book, but the higher the efficiency the better.

!Once the three inputs are decided then select "1. Sample Size Calculations" tab.

6. Generating the Sample

If the sample specifications displayed on the screen are acceptable, the next step is to generate the sample. Select tabs **"2. Select Random Sample"** and **"3. Sample Validation" .** That is all that is required and two random sample files will be generated: "SampleData.txt" and "SampleData.csv". The .csv file is a comma separated variable file which is a text file that can be directly read into Excel and saved as an Excel workbook. The .txt file uses a tab as a variable separator. Both files will be saved in the same folder that had originally been set up for the audit population data.

7. Documenting and Recording Results

Once you are satisfied with the sample as displayed on the screen, then spreadsheet templates should be generated that both can be used to document and record audit results. As part of deciding on a final design acceptance, does the actual precision of the randomly selected sample actually match the original input precision used to determine sample size? Remember random means it is possible to have a sample with values that are outside of the original precision.

Statistical audits have an advantage over other types of sampling environments. The auditor selects samples from computerized accounting systems. Such systems can automatically summarize descriptions of the total book such as account totals and other measures. Therefore key audit population parameters are known. Suppose the precision in designing an audit sample is **set at 4%** of total dollar volume. A validity check would be to determine if a total dollar estimate derived from that sample does indeed fall within the 4% precision of the actual book total. Auditmetrics does this analysis and displays on the screen:

Observed sample precision under 0.04 no need to resample

OR

Observed sample precision over 0.04 need to resample

If the precision test fails then re-run "**1. Sample Size Calculation"** to start the process again. Redo steps 2 and 3 until the observed precision of the sample matches or is better than the precision used during sample design.

! This is an overall sample validation based on precision, the next validation test is a strata by strata statistical test.

The second validation test a pass/fail 95% confidence interval for each stratum. Its results will on the screen. If one or more strata fail then start again to calculate another sample and point and click until both validation test #1 and test #2 are passed.

! Only Proceed when as below all strata are designated "OK" and precision test is passed.

n	Mean	SD.	Total $		
69	22.74	22.78	1569	ok	ok
84	141.33	69.6	11871	ok	ok
89	418.15	174.03	37216	ok	ok
129	1001.66	404.51	129215	ok	ok
172	2108.62	955.52	362682	ok	ok
125	4412.92	2562.87	551614	ok	ok
451	30657.11	68197.67	13826355		

Validation #1- Observed precision under 0.03 no need to resample

Validation #2- Strata specific test passed.

8. Generating Excel Templates

After a specific sample design is decided upon, the next step is to document the sample with an Excel spreadsheet. You will now do one final run but this time it should be done with one of the following radio buttons selected.

The steps for documenting the sample:

- Sample Size Excel File
- o Sample Validation Excel File

Select button **"1. Sample Size Calculation"** , **"2. Select Random Sample",** and **"3. Sample Validation".** The sample calculation step will run again but this time generating an Excel file that will document sample specifics that should be shared with all interested parties.

The steps for sample validation and recording audit results:

- o Sample Size Excel File
- Sample Validation Excel File

Then again select button **"1. Sample Size Calculation" "2. Select Random Sample"** and **"3. Sample Validation"**

! The spreadsheets are protected so that you cannot inadvertently write over a critical formula. If you need to make alterations then go to review at the top menu and select "unprotect sheet"

The spreadsheet exhibited below is a segment from the ***"sample" section*** from the validation spreadsheet. Each strata validity test has been passed. The column **"Amount Error"** is filled in by the auditor after totaling those transactions from the sample that are in error. The spreadsheet contains all of the Excel formulas to calculate the 95% confidence interval around the sample estimate of $33,500 of taxes owed. It should be noted that the 95% confidence interval displayed is a one sided confidence interval which is consistent with IRS directives.

Sample		Audit Results	Validity Check
Sample Total Value	Sample Size	Amount Error	
$1,011	41	$28	pass
$7,636	89	$225	pass
$14,486	94	$436	pass
$31,217	127	$916	pass
$58,512	138	$152	pass
$632,464	284	$6,325	
$112,861	**489**		
$745,325	773		
		Sample Error Rate =	0.025
		Overall Rate=	0.020
Detail Error	**$6,325**		
Sample Estimate:			
Lower Bound	$33,110		
Mid-Point	**$33,500**		
Upper Bound	$33,891		

9. Data Mining as a Sequel

Auditmetrics gets you started in sampling an account based on dollars. In business, dollars are the life blood of survival. With its templates, Auditmetrics can determine which dollars do not meet sufficient performance, in both absolute and percentage terms. You may have noticed that when a political poll is discussed in addition to the quantitative percentage result, most will discuss "what are the internals" of the poll. That means is how do the overall results relate to important breakdown factors such as gender, race, economic status and age. In statistics this is called crosstabs which exposes the underlying dynamics that help to plan future action. The original dollars is a quantitative variable while the factor breakdowns are called attributes.

What does this mean for small business? We are entering in a high tech commercial environment where huge commercial entities can marshal vast sophisticated programming to correct past deficiencies and search for potential opportunities. Auditmetrics helps the smaller enterprise to use accepted analytical tools that can fill in the void to survive in the modern economy. It is basically an analytical level playing field.

Variable Sampling -Variable sampling involves quantitative measurable amounts and the result is rated on a continuous scale that measures the degree of conformity. Variable sampling is about checking "how much". For Auditmetrics we start with dollars as our variable.

Attribute Sampling- In attribute sampling the data result either conforms or does not conform. It is a method of measuring quality that consists of noting the presence or absence of some characteristic in each of the units under consideration. Attribute sampling checks "how many Conform". An example would be "how many transactions are from urban vs suburban areas"?

Excel functions ="frequency" and pivot tables are valuable tools in market analysis. Pivot table is a spreadsheet functionality that allows you to arrange and categorize attributes. It can be used to breakdown revenue by geographic categories, age breakdowns etc.

A histogram is a graphical representation (chart) of distribution data. A frequency distribution displays the number of data points that fall within specified ranges in a sample, for example dollar ranges. Dollar ranges are valuable for day-to-day marketing, histograms are commonly used in finance. Since finance affects every single business, understanding how to read, create, and manipulate data in the form of a histogram and pivot table is critical for business owners and marketers.

Below is excerpted from a random sample that an auditor used to monitor the supply chain for a small manufacturing company:

Transactio	amount	absamt	PERIOD	Vendor	Strata	Error
22	421.43	421.43	3/1/2017	Jones & Sons	3	yes
24351	550	550	2/1/2016	ACE Dist.	3	no
8585	51	51	4/1/2015	ABC Co.	1	no
12345	7000	7000	3/1/2015	Smith Bros.	5	no
24376	10500	10500	6/1/2017	Acme Inc.	6	yes
14666	6034.48	6034.48	1/1/2015	Acme Inc.	5	no
80	9000	9000	2/1/2017	Smith Bros.	6	no
	.	.	.	.	.	.
	.	.	.	.	.	.

In conducting the audit a record was kept by the auditor to monitor which transactions contained errors. It was determined if a delivery was incomplete, faulty, delayed, the wrong supply or price etc. All costly delays in doing business. The Auditmetrics sample and templates documents the scope of the problem. But more is needed to guide the auditor towards a solution.

Using an Excel pivot table the auditor can now target corrective measures. For example Jones & Sons has the highest error rate. A true attribute percentage would be how many transactions are in error. The percentage displayed below is based on how many dollars are in error providing a more detailed economic impact statement. Therefore it is a dollar variable calculation broken down by the attribute error. This is the start of the detective work.

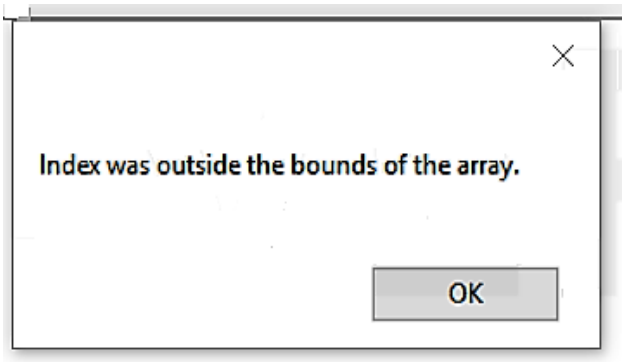
	Error By Vendor			
Sum of amount	**Error**			**Percent Error**
Company	**no**	**yes**	**Grand Total**	**By Vendor**
ABC Co.	$3,880,847	$166,861	$4,047,708	**4%**
ACE Dist.	$1,548,288	$159,263	$1,707,552	**9%**
Acme Inc.	$3,391,568	$398,518	$3,790,085	**11%**
Flower Inc.	$1,701,054	$205,547	$1,906,601	**11%**
Jones & Sons	$1,696,507	$405,821	$2,102,328	**19%**
Smith Bros.	$1,656,511	$112,436	$1,768,947	**6%**
Grand Total	**$13,874,774**	**$1,448,446**	**$15,323,221**	

10. **Data Sources and Trouble Shooting**

In the text book, we discuss obtaining tab delimited data from Excel, MS Access, explained in Appendix II, and QuickBooks. For most businesses, these are the most common data transfer vehicles other than specialized tailored accounting systems. For large business the other potential data sources would probably require the input of an IT administrator or database manager. Despite the multitude of data inputs the output sample files are two, a tab delimited file and "comma separated variable" (CSV) text file. The CSV file has the advantage of being directly read in by Excel and can be immediately saved as an Excel workbook.

The sample data in this exercise was medical claims data contained in a single Excel spreadsheet. When dealing with large datasets or the need to merge several datasets, MS Access provides an easy way to do more complex data manipulations. If you are not familiar with Access then this is a good time to review Appendix II especially how to import/export tab delimited text files.

The dataset for analysis requires a rectangular matrix with each variable a column and each individual transaction a row. There are times when you think your data is all set for analysis and then you get the message:

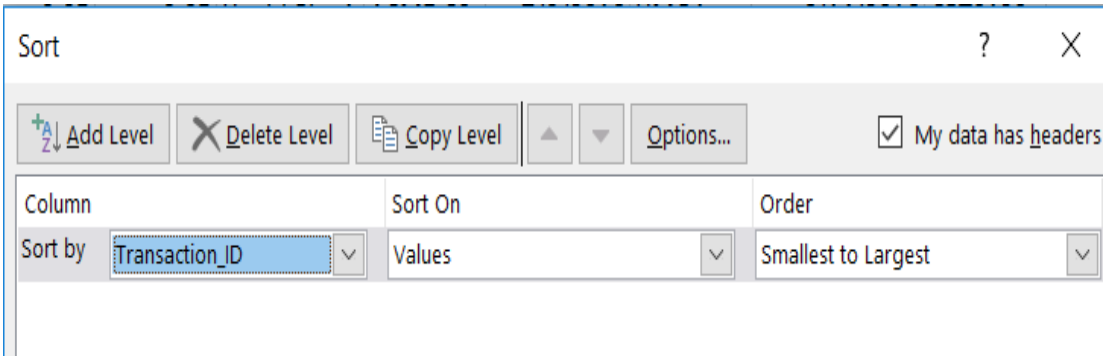


This is a message thrown off by the operating system and the usual problem is there is something wrong with the data. Processing cannot continue. What generated this error message is that the original Excel file contained blank cells (row 15414) with no data or only partial data (row 15430):

Excel is a file manager not a database manager so when it is solely used as a data source, it may not be in the form of a proper rectangular data matrix with the columns as variables and rows containing individual data points. In this case the blank and partial rows are throwing off the Auditmetrics data parser. Sorting the file will help in determine if each row is complete.

	A	B	C	D	F
	DataSet	**Transaction**	**Account**	**Name**	**Amount**
15413	ERG	15410	107691	Acme Co.	15
15414					
15415	ERG	15412	310789	XYZ Co.	35
15428	ERG	15425	107691	Acme Co.	45
15429	ERG	15426	310789	XYZ Co.	12
15430	Workshop				
15431	ERG	15427	107691	Acme Co.	25
15432	ERG	15428	310789	XYZ Co.	32
15433	ERG	15429	310789	XYZ Co.	15

Solution Use Excel Data Sort:



Make sure that "my data" has headers" is checked. Once you sort the spreadsheet you will notice the partial and bad lines either move to the bottom or top depending on the data and filter mechanism.

	A	B	C	D	F
	DataSet	**Transaction**	**Account**	**Name**	**Amount**
15413	ERG	15410	107691	Acme	15
15414	ERG	15412	310789	XYZ Co.	35
15415	ERG	15425	107691	Acme	45
15428	ERG	15426	310789	XYZ Co.	12
15429	ERG	15427	107691	Acme	25
15430	ERG	15428	310789	XYZ Co.	32
15431	ERG	15429	310789	XYZ Co.	15
15432					
15433	Workshop				

You can now eliminate the bad data and are left with a rectangular data matrix to generate a tab text file. This methodology is like the method used to derive data from QuickBooks standard reports described in the text.