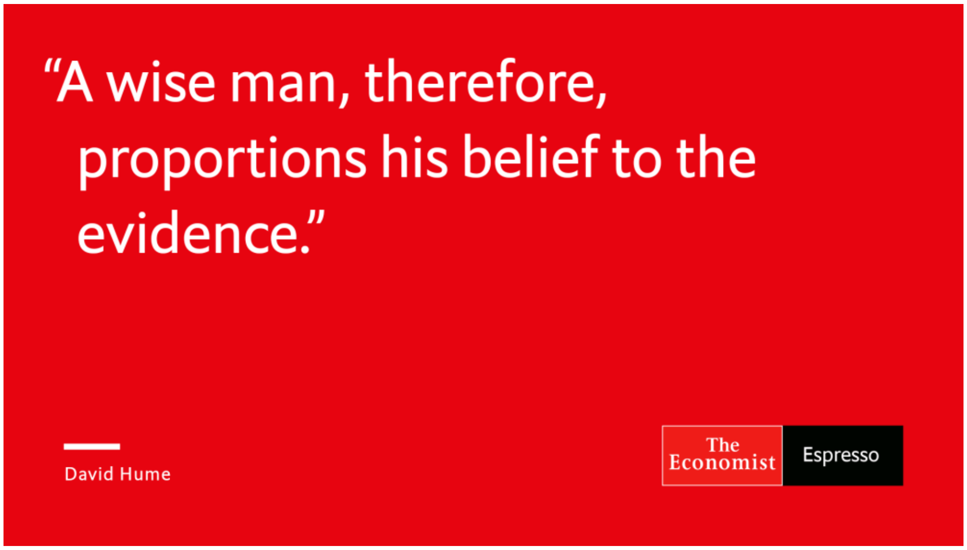
CHAPTER 2 Integrated Case: Industry Analysis



"A wise man, therefore, proportions his belief to the evidence."

David Hume

The Economist
Espresso

2.1 Learning Objectives

A common problem when dealing with financial accounting data is outliers. Outliers are abnormal observations compared to other observations in a dataset. This often means observations that are either extremely large or extremely small. It is up to the analyst to decide what is considered abnormal. Knowing how to deal with outliers is critical for the correct execution of data analytics and is something developed with experience and by leveraging accounting knowledge. In this chapter, we will approach financial accounting data from the standpoint of a financial analyst.

By the end of this chapter, students should be able to apply all stages of CRISP-DM in the context of an integrated case. Students will use the financial data of firms competing in the same industry to answer a specific business question. The outline of the chapter maps to the following learning objectives:

1. Define the business problem and convert it into a data analytics problem.
2. Understand data - load, review, and clean data while leveraging accounting knowledge.
3. Prepare data - create new variables.
4. Model the data analytics problem and evaluate results using accounting knowledge.
5. Communicate results.
6. Improvise - create variations of an existing problem.

2.2 Students: Advance Preparation

If you have not already done so, review [Appendix C: Alteryx Checklist]to familiarize yourself with the resources you can refer to while reading the textbook and working on the problems.

Use the following hyperlink to access Appendix D: Student Preparation Worksheet and focus on the topics Formatting Data, Filtering Data, and Summarizing Data.

2.3 Firm Size and Profitability

2.3.1 Business Understanding

There are two ways to perform analytics: deductive (start with the problem) and inductive (start with the data). Rather than exploring the data, this chapter will focus on starting with the business question. Because this is the first attempt at applying CRISP-DM, we will start with the following example.

You work for a senior financial analyst, and your first assignment is to leverage the financial data that you have for all firms in the same industry as Apple to answer the following question:

- *Are large technology firms more profitable than small firms?*

The problem is relatively simple, but in order to answer this question using data analytics, there are some important questions that you will need to clarify.

1. What are technology firms?
   - For this assignment, technology firms are defined as all firms competing in the same industry as Apple.
2. Will we examine the question during one fiscal year or over several years?
   - For the context of this initial assignment, we are going to …
3. What measure are we supposed to use to capture firm size? Should we use sales, assets, employees, market share, market valuation, or something else?
   - To classify firms based on size, we will use …
4. How are we going to split companies in terms of size? For example, if we choose to use sales for classifying firms in terms of size, what is the cut-off point we should use to make the classification? Should we use a specific figure (e.g., 100 million) or the industry average or median or third quartile? Should we split the industry into just two groups or multiple groups?
   - We will classify firms in … group(s) using the …
5. What measure of profitability are we supposed to use (gross profit margin, net profit margin, return on assets, return on equity), and how should we specify it? (See Appendix A for a review of some of the most commonly used financial accounting ratios.)
   - To measure profitability, we will use the … and define this as follows …

Who should be responsible for answering these questions, the financial analyst or the data scientist? Does it matter? Think about your answers to these questions before you move on to the next section!

2.3.1.1 Convert into a Data Analytics problem

It is likely that a data scientist working on a financial accounting problem is not familiar with the different measures of firm size and profitability. Therefore, we need to leverage our domain specific knowledge (i.e., financial accounting) to convert this financial accounting problem into a data analytics problem.

Leveraging the answers to the questions above, we can convert this financial accounting problem into a data analytics problem as follows:

1. We will treat all firms competing in the same NAICS industry classification as Apple as technology firms (i.e., NAICS = 334220). For more information on NAICS codes, see https://www.census.gov/naics/.
2. For the context of this initial assignment, we will focus on just one year, 2016.
3. Use sales to define company size.
4. Use median sales to classify companies as large or small. Companies with sales equal to or greater than the industry median will be classified as large, and those with less than the industry median will be classified as small.
5. Use gross margin, $(Sales-COGS)/Sales$, to measure profitability.
6. Now, we can answer the question, “Are large technology firms more profitable than small firms?”

2.3.2 Data Understanding

At least initially, it is a good idea to use a checklist like Appendix C: Alteryx Checklist. Below is a summary of the key tasks related to starting a new project in Alteryx:

I. Create a folder for the project.

II. Set this folder as the working directory. If your working directory contains many files, it may make sense to create subfolders. For example, you may have a subfolder containing your data and have named this folder ‘data.’. To reference this folder, you can use relative paths. For example, ../data/ means a folder called data in the working directory. Review the documentation in Appendix C: Alteryx Checklist for working with relative pathways in Alteryx.

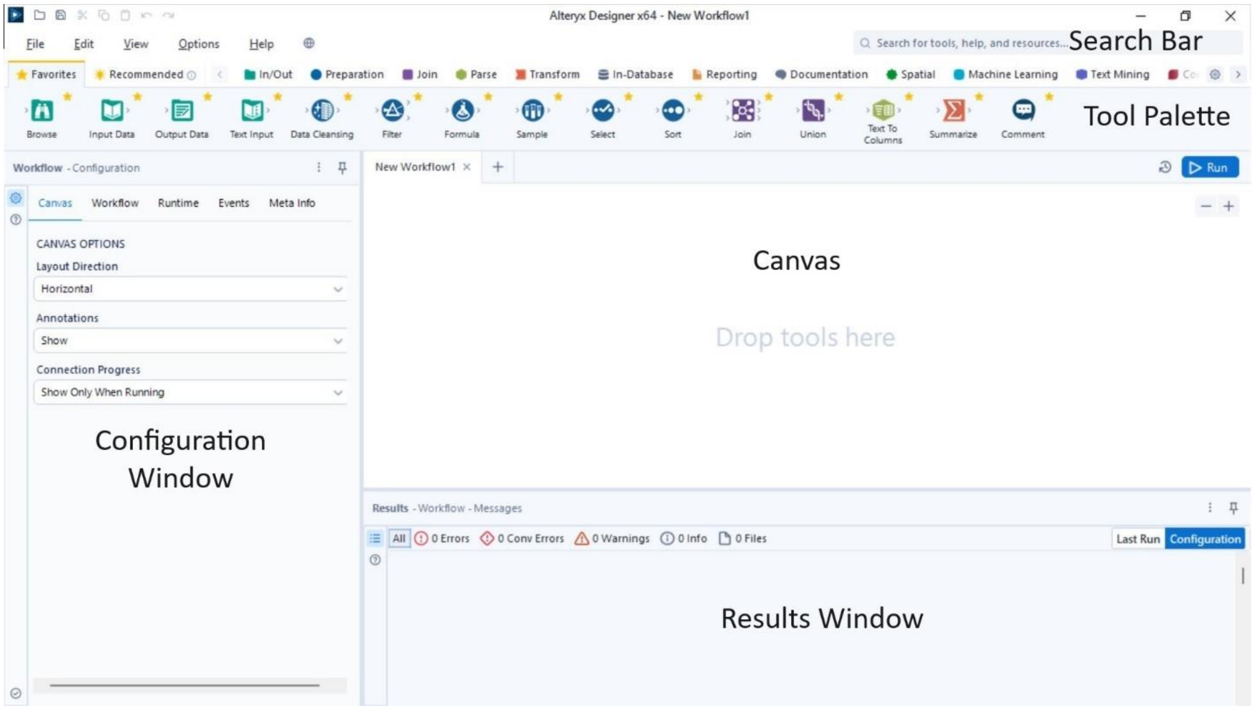
III. Download and save the project data inside the working directory. The dataset for this exercise (industryAnalysis_2000_2023.csv) is financial data for publicly traded firms from 2000 to 2023. The data has been extracted from Standard and Poor’s Compustat, a financial and market information database of active and inactive publicly traded companies. Investors, universities, and financial analysts use Compustat. The database contains historical data for firms since the 1950s.

As shown in Figure 2.1, Alteryx Designer is made up of four primary sections:

1. Tool Palette – Operations in Alteryx are performed by tools. The tools are represented

by icons with a descriptive symbol on them and color coded into categories. Each tab in Alteryx contains the most common tools in that category, and additional tools can be found using the Search Bar.

2. Canvas – The Canvas is the area used to build a workflow. A workflow is the steps the data goes through as it gets input, manipulated, and analyzed. Drag a tool from the Tool Palette or the Search Bar results and drop it onto the Canvas to add it to the workflow. Connect the output anchor of a tool to one or more input anchors of the next tool(s) in the workflow.
3. Configuration Window – In order for tools to work correctly they need to be configured once they are placed on the Canvas. Once a tool is on the Canvas, click on it and configure it in the Configuration Window.
4. Results Window – To execute the steps represented by the tools on the Canvas, click on the Run button on the top right of the Canvas. The results of the execution, including any error messages, will appear in the Results Window. You can see the dataset as it passes from tool to tool by clicking on the output anchor of a tool. When you do so, the current state of the dataset will display in the Results Window.



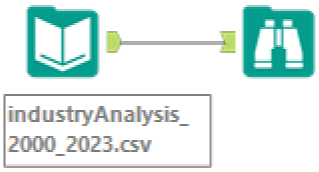
CANVAS OPTIONS

Layout Direction: Horizontal

Annotations: Show

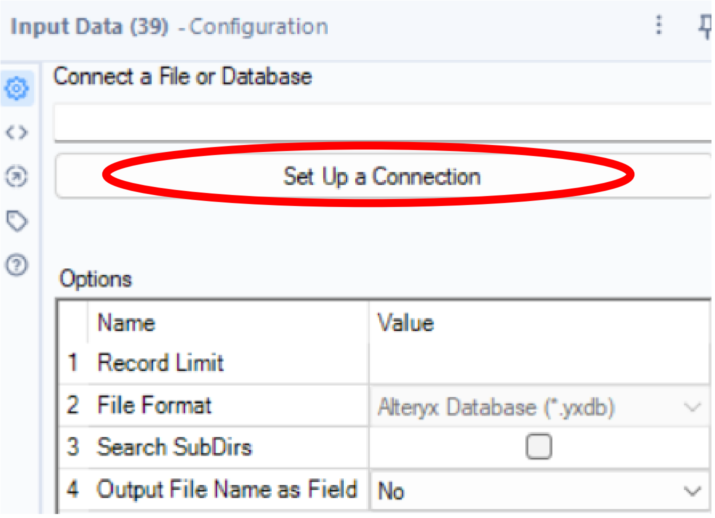
Connection Progress: Show Only When Running

Figure 2.1: Alteryx Designer



Build the workflow shown on the left by following the steps below. This workflow will be used to review the industry analysis dataset.

Input the data using an Input Data Tool. Drag an Input Data Tool onto the Canvas from the In/Out category on the Tool Palette. Configure the connection following the steps below to read the dataset industryAnalysis_2000_2023.csv.

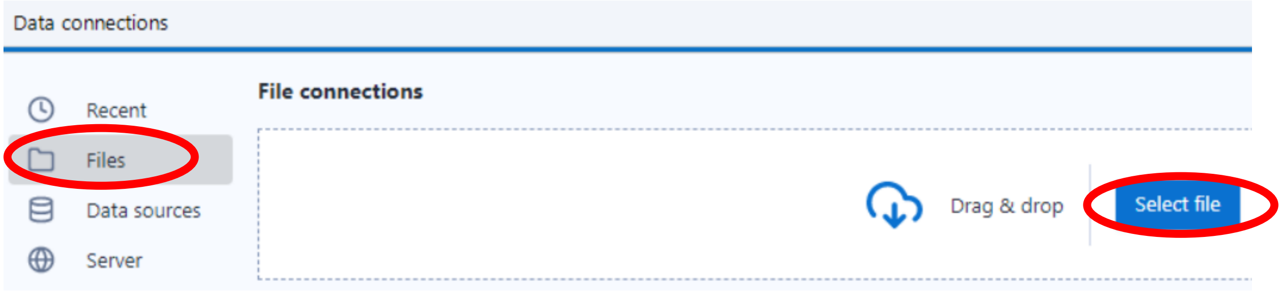


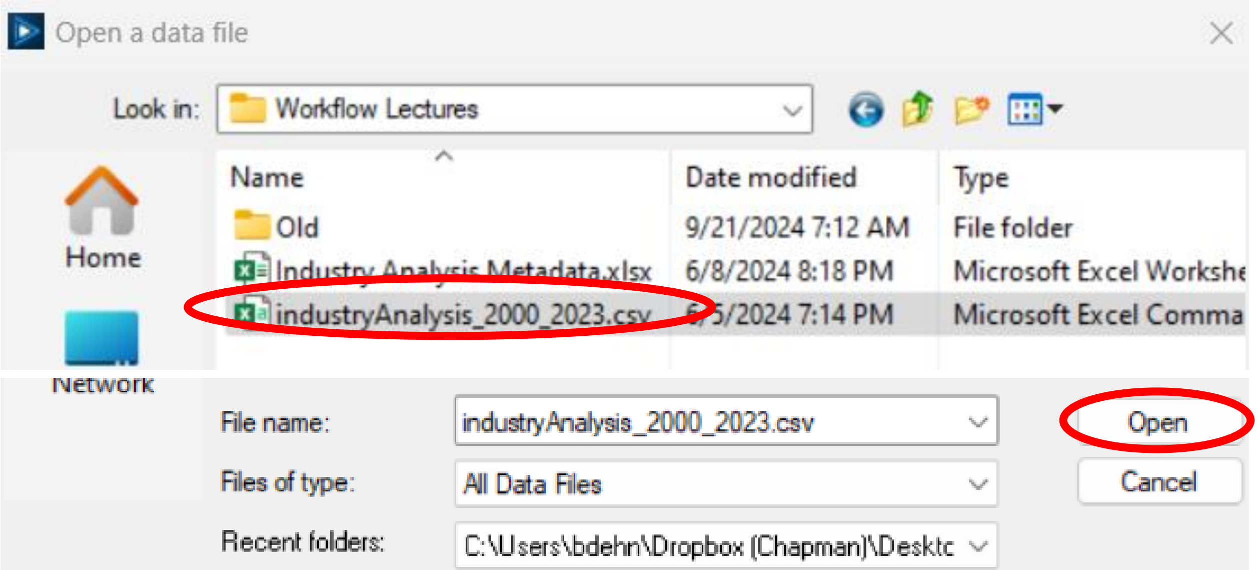
Input Data (39) - Configuration

Connect a File or Database

Options

	Name	Value
1	Record Limit	
2	File Format	Alteryx Database (*.yxdb)
3	Search SubDirs	□
4	Output File Name as Field	No





Open a data file

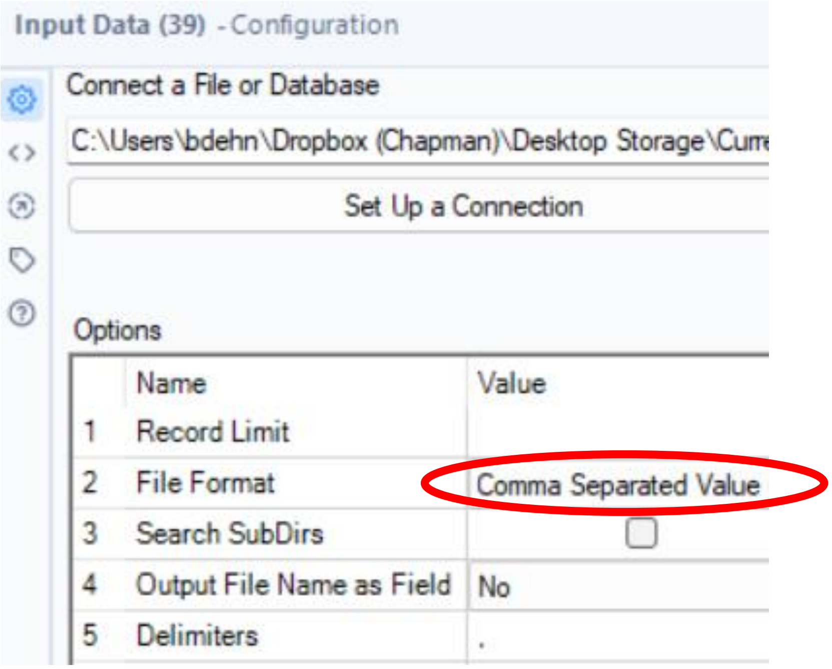
Look in: Workflow Lectures

Name	Date modified	Type
Old	9/21/2024 7:12 AM	File folder
Industry Analysis Metadata.xlsx	6/8/2024 8:18 PM	Microsoft Excel Workshe
industryAnalysis_2000_2023.csv	6/5/2024 7:14 PM	Microsoft Excel Comma

File name: industryAnalysis_2000_2023.csv

Files of type: All Data Files

Recent folders: C:\Users\bdehn\Dropbox (Chapman)\Desktc



Input Data (39) - Configuration

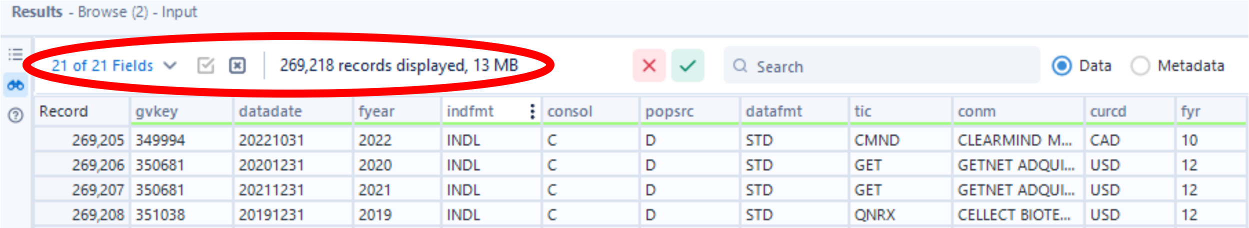
Connect a File or Database

C:\Users\bdehn\Dropbox (Chapman)\Desktop Storage\Curre

Options

	Name	Value
1	Record Limit	
2	File Format	Comma Separated Value
3	Search SubDirs	□
4	Output File Name as Field	No
5	Delimiters	,

Connect a Browse Tool to the output anchor of the Input Data Tool and run the workflow. Click on the Browse Tool to examine the data and a sample of observations in the Results Window. There are 269,218 records and 21 fields in the dataset.

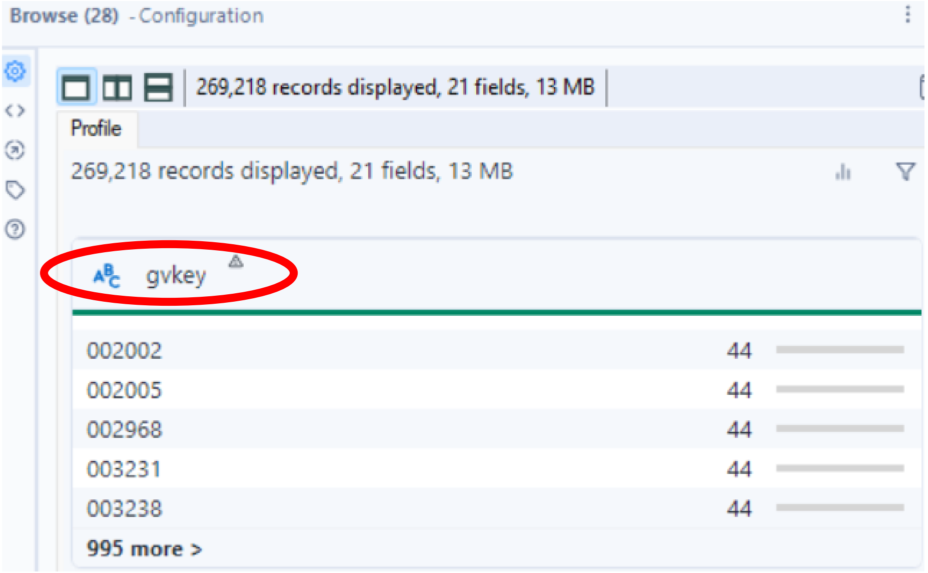


Results - Browse (2) - Input

21 of 21 Fields   269,218 records displayed, 13 MB

Record	gvkey	datadate	fyear	indfmt	consol	popsrc	datafmt	tic	conm	curcd	fyr
269,205	349994	20221031	2022	INDL	C	D	STD	CMND	CLEARMIND M...	CAD	10
269,206	350681	20201231	2020	INDL	C	D	STD	GET	GETNET ADQUI...	USD	12
269,207	350681	20211231	2021	INDL	C	D	STD	GET	GETNET ADQUI...	USD	12
269,208	351038	20191231	2019	INDL	C	D	STD	QNRX	CELLECT BIOTE...	USD	12

The Configuration Window displays the data type and the most common values for each variable. It is essential to understand what each variable stands for. If you do not know what the names of the variables mean, ask before you do any analysis!

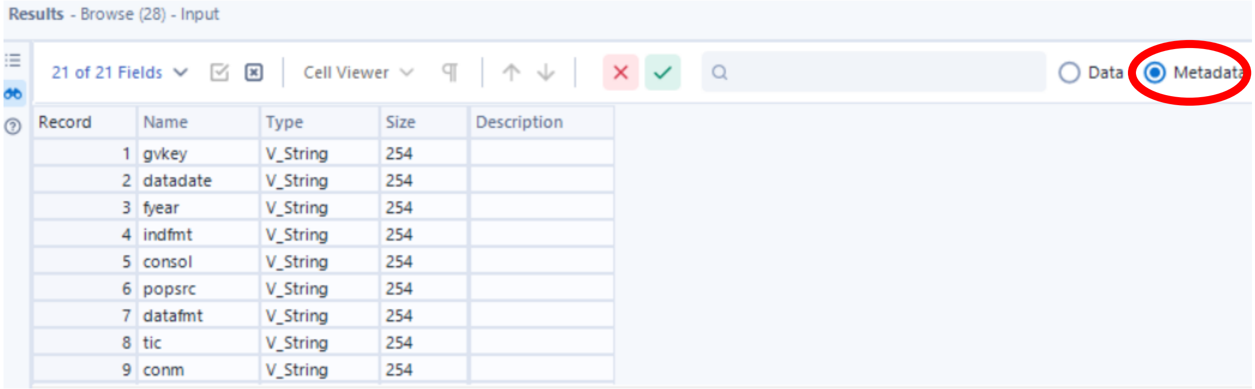


269,218 records displayed, 21 fields, 13 MB

269,218 records displayed, 21 fields, 13 MB

gvkey	
002002	44
002005	44
002968	44
003231	44
003238	44
995 more >	

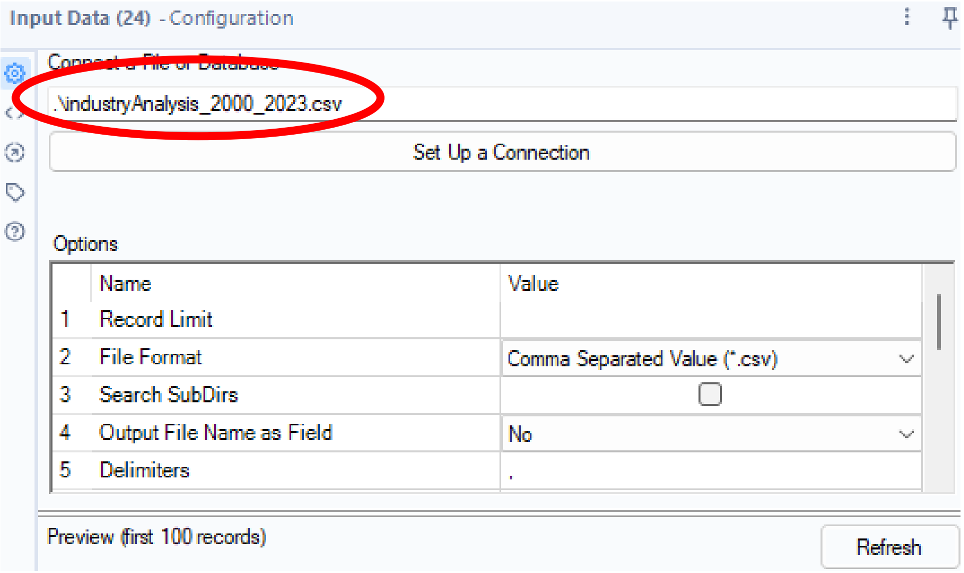
In the Results Window, when you click on the Metadata radio button, a simple data dictionary is displayed with information on field names, data types, data source, and field descriptions. Metadata can be thought of as a description of the dataset that serves as a reference regarding the fields and attributes described by the data.



Record	Name	Type	Size	Description
1	gvkey	V_String	254	
2	datadate	V_String	254	
3	fyear	V_String	254	
4	indfmt	V_String	254	
5	consol	V_String	254	
6	popsrc	V_String	254	
7	datafmt	V_String	254	
8	tic	V_String	254	
9	conm	V_String	254	

Although many items in a data dictionary can be classified as metadata, data dictionaries and metadata are not identical. Data dictionaries generally contain only some of the metadata necessary for understanding and navigating data elements and databases and, thus, contain only a subset of the metadata found in a robust metadata system.

After running a workflow and verifying that that the data has been imported correctly, it is a good idea to set relative pathways. When relative pathways are set, the Input Data Tool will find the dataset on any computer in any location as long as the dataset and workflow are in the same folder. For information on how to set relative pathways, see Appendix C: Alteryx Checklist section **C.3**. After setting relative pathways, the Input Data Tool configuration window will display the filename preceded by ".\" instead of the full path, thus indicating a relative pathway.



Input Data (24) - Configuration

Connect a File or Database

.\industryAnalysis_2000_2023.csv

Options

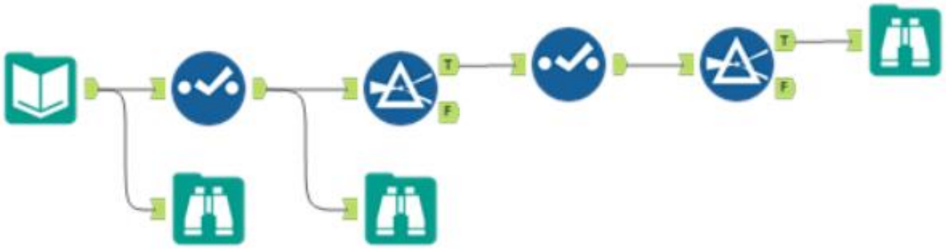
	Name	Value
1	Record Limit	
2	File Format	Comma Separated Value (*.csv)
3	Search SubDirs	☐
4	Output File Name as Field	No
5	Delimiters	,

Preview (first 100 records)

When we import a .csv file, the data types of all fields are V_String. As we proceed, we must change the data types as per the data dictionary below:

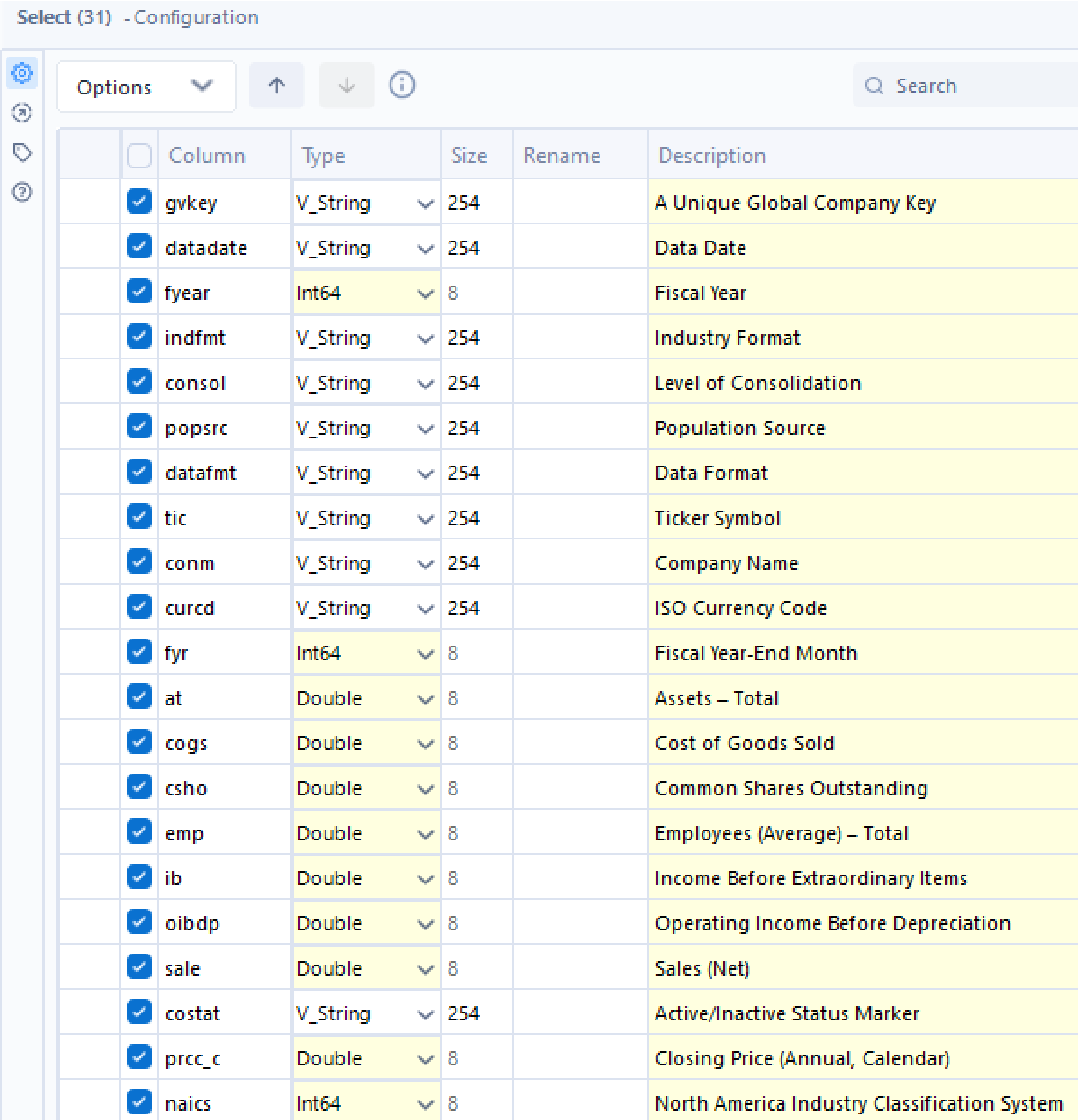
Name	Type	Description
gvkey	String	A Unique Global Company Key
datadate	Date	Data Date
fyear	Integer	Fiscal Year
indfmt	String	Industry Format
consol	String	Level of Consolidation
popsrc	String	Population Source
datafmt	String	Data Format
tic	String	Ticker Symbol
conm	String	Company Name
curcd	String	ISO Currency Code
fyr	Integer	Fiscal Year-End Month
at	Double	Assets – Total
cogs	Double	Cost of Goods Sold
csho	Double	Common Shares Outstanding
emp	Double	Employees (Average) – Total
ib	Double	Income Before Extraordinary Items
oibdp	Double	Operating Income Before Depreciation
sale	Double	Sales (Net)
costat	String	Active/Inactive Status Marker
prcc_c	Double	Closing Price (Annual, Calendar)
naics	Integer	North America Industry Classification System

We will continue building the workflow shown below by adding Select and Filter Tools as described on the following pages.



First, drag a Select Tool from the Preparation Tools category onto the Canvas and connect it to the Input Data Tool. Click on the Select Tool and in the Configuration Window, configure the tool as shown below. The description is optional but including it will add this metadata information to the dataset in the workflow if desired.

Note that we cannot change the datadate field to the **date** data type at this time, because the variable's format (yyyyMMdd) does not follow the Alteryx date format (yyyy-MM-dd). We will learn how to handle this using the DateTime Tool in a later chapter.

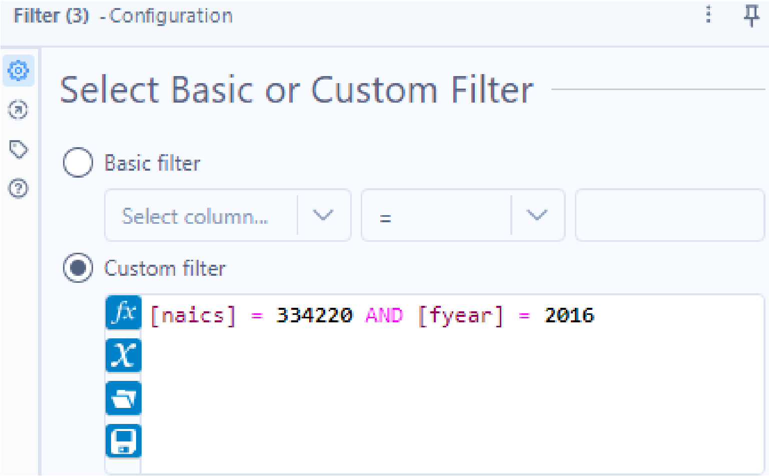


Select (31) - Configuration

□	Column	Type	Size	Rename	Description
☑	gvkey	V_String	254		A Unique Global Company Key
☑	datadate	V_String	254		Data Date
☑	fyear	Int64	8		Fiscal Year
☑	indfmt	V_String	254		Industry Format
☑	consol	V_String	254		Level of Consolidation
☑	popsrc	V_String	254		Population Source
☑	datafmt	V_String	254		Data Format
☑	tic	V_String	254		Ticker Symbol
☑	conm	V_String	254		Company Name
☑	curcd	V_String	254		ISO Currency Code
☑	fyr	Int64	8		Fiscal Year-End Month
☑	at	Double	8		Assets – Total
☑	cogs	Double	8		Cost of Goods Sold
☑	csho	Double	8		Common Shares Outstanding
☑	emp	Double	8		Employees (Average) – Total
☑	ib	Double	8		Income Before Extraordinary Items
☑	oibdp	Double	8		Operating Income Before Depreciation
☑	sale	Double	8		Sales (Net)
☑	costat	V_String	254		Active/Inactive Status Marker
☑	prcc_c	Double	8		Closing Price (Annual, Calendar)
☑	naics	Int64	8		North America Industry Classification System

The original dataset has financial data for firms from many industries and years. Given our decision to focus on technology companies competing in the same industry as Apple (NAICS = 334220), fiscal year 2016, using *sales* to control for size, and working with *gross margin*, it makes sense to focus on a subset of relevant data. We will use a Filter Tool and a Select Tool to reduce the dataset to only technology companies and a subset of the fields as shown below.

Drag a Filter Tool from the Preparation Tools category onto the Canvas and connect it to the Select Tool. Click on the Filter Tool and in the Configuration Window configure the tool as shown below.


Filter (3) - Configuration

Select Basic or Custom Filter

○ Basic filter

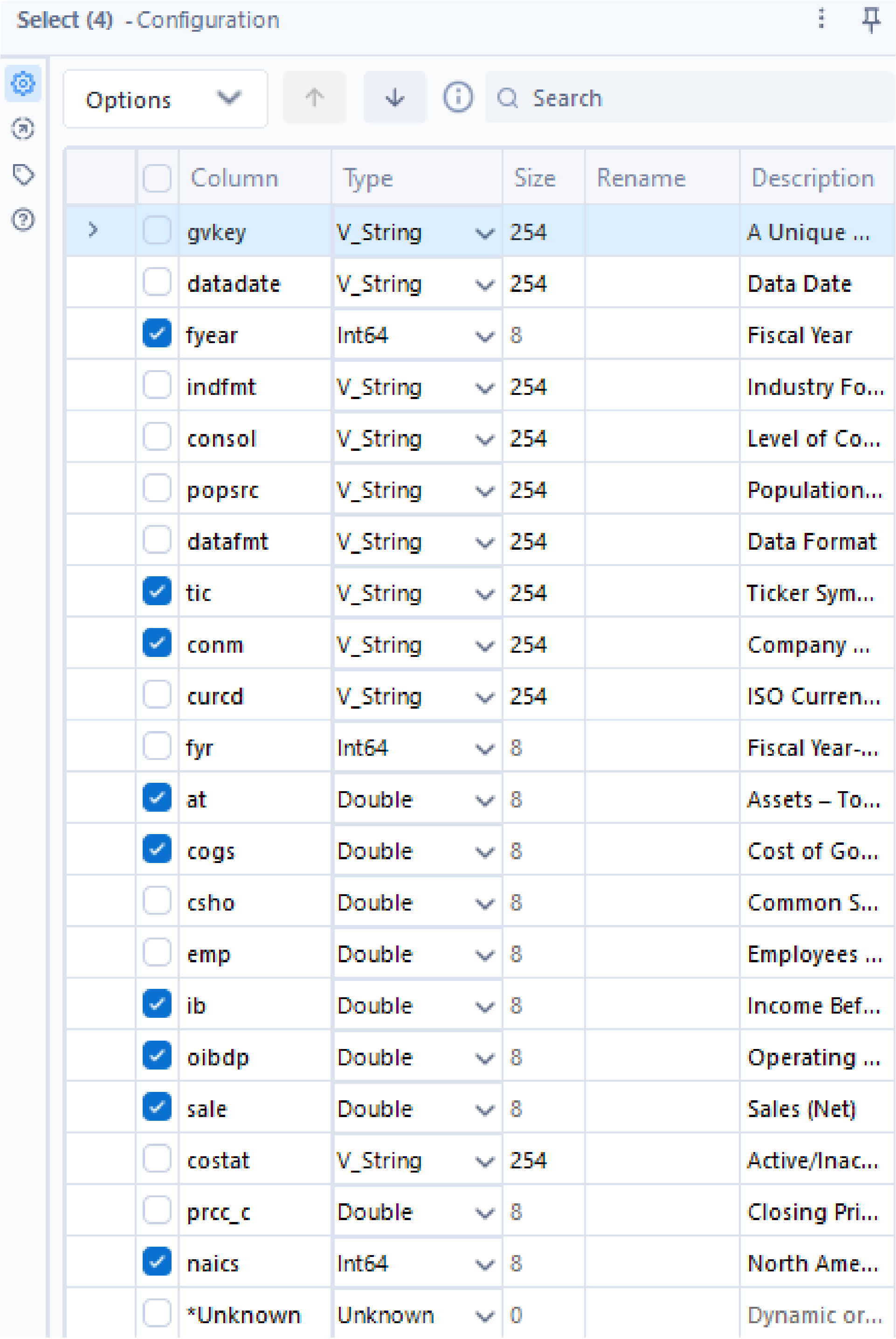
Select column... =

● Custom filter

```
[naics] = 334220 AND [fyear] = 2016
```

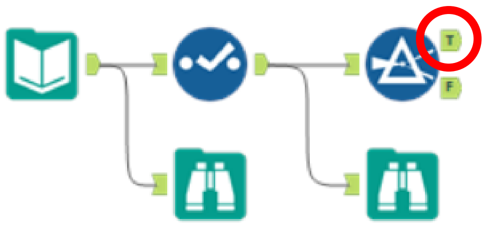
We will work with only a subset of the fields in the additional analysis, and we can use a Select Tool to choose which variables to keep.

Drag a Select Tool from the Preparation Tools category onto the Canvas and connect it to the Filter Tool. Click on the Select Tool and in the Configuration Window configure the tool as shown below then run the workflow.

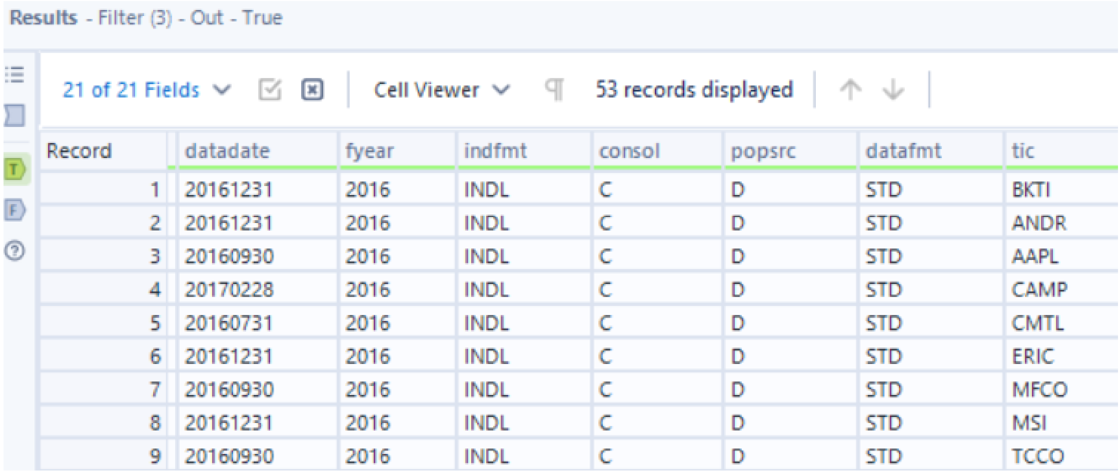


	□	Column	Type	Size	Rename	Description
>	□	gvkey	V_String	254		A Unique ...
	□	datadate	V_String	254		Data Date
	☑	fyear	Int64	8		Fiscal Year
	□	indfmt	V_String	254		Industry Fo...
	□	consol	V_String	254		Level of Co...
	□	popsrc	V_String	254		Population...
	□	datafmt	V_String	254		Data Format
	☑	tic	V_String	254		Ticker Sym...
	☑	conm	V_String	254		Company ...
	□	curcd	V_String	254		ISO Curren...
	□	fyr	Int64	8		Fiscal Year-...
	☑	at	Double	8		Assets – To...
	☑	cogs	Double	8		Cost of Go...
	□	csho	Double	8		Common S...
	□	emp	Double	8		Employees ...
	☑	ib	Double	8		Income Bef...
	☑	oibdp	Double	8		Operating ...
	☑	sale	Double	8		Sales (Net)
	□	costat	V_String	254		Active/Inac...
	□	prcc_c	Double	8		Closing Pri...
	☑	naics	Int64	8		North Ame...
	□	*Unknown	Unknown	0		Dynamic or...

Run the workflow. There are now 53 records and 9 fields in the dataset. The Browse Tool is the best way to get full information on the dataset as it passes from tool to tool. However, sometimes we just need a quick peak at the data. Clicking on an output anchor of a tool in the workflow, such as the "T" (true) output anchor of the Filter Tool displays a portion of the dataset.



A portion of the results window is shown below. Looking through the fields and the data preview, we can see that the dataset is much smaller now, with just 53 records, all with NAICS = 334220 and fiscal year 2016.


Results - Filter (3) - Out - True

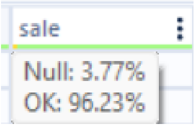
21 of 21 Fields | Cell Viewer | 53 records displayed

Record	datadate	fyear	indfmt	consol	popsrc	datafmt	tic
1	20161231	2016	INDL	C	D	STD	BKTI
2	20161231	2016	INDL	C	D	STD	ANDR
3	20160930	2016	INDL	C	D	STD	AAPL
4	20170228	2016	INDL	C	D	STD	CAMP
5	20160731	2016	INDL	C	D	STD	CMTL
6	20161231	2016	INDL	C	D	STD	ERIC
7	20160930	2016	INDL	C	D	STD	MFCO
8	20161231	2016	INDL	C	D	STD	MSI
9	20160930	2016	INDL	C	D	STD	TCCO

Below each field name and before the first record is a "data quality bar." The colors indicate the status of the data in that field.

Green (OK): The column contains values without leading or trailing white spaces.
Red (Not OK): The column contains values with leading or trailing white space.
Yellow (Null): The column contains records with no values.
Gray (Empty): The column contains strings with no values but includes something else such as white space or embedded new lines.
Diagonal lines (Partial results): The column contains more than 1 MB of data. The quality bar only reports the quality of the data currently displayed in the Results window.

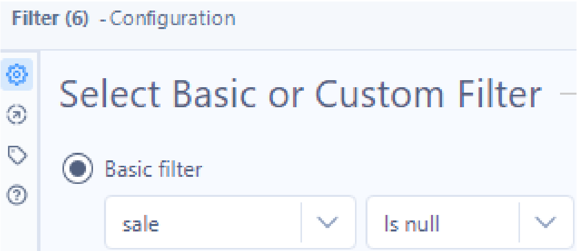
When you hover your cursor over the data quality bar it gives you a breakdown of the data issues. Placing your cursor over the data quality bar for the sales field reveals that 3.77% of the values are null.


sale

Null: 3.77%
OK: 96.23%

We can use another Filter Tool to examine the companies with null values in the sale field.

Drag a Filter Tool from the Preparation Tools category onto the Canvas and connect it to the Select Tool. Click on the Filter Tool and in the Configuration Window configure the tool as shown below.



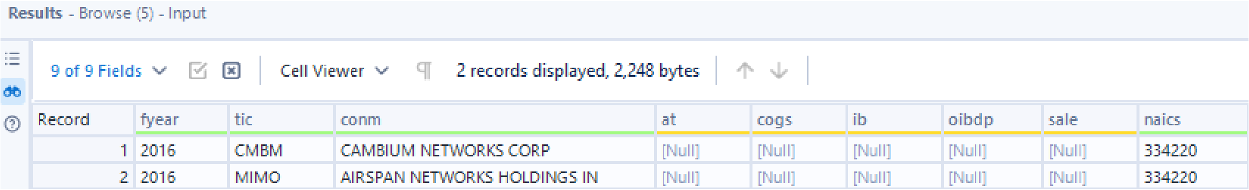
Filter (6) - Configuration

Select Basic or Custom Filter

● Basic filter

sale | Is null

Run the workflow. The records with missing values are:



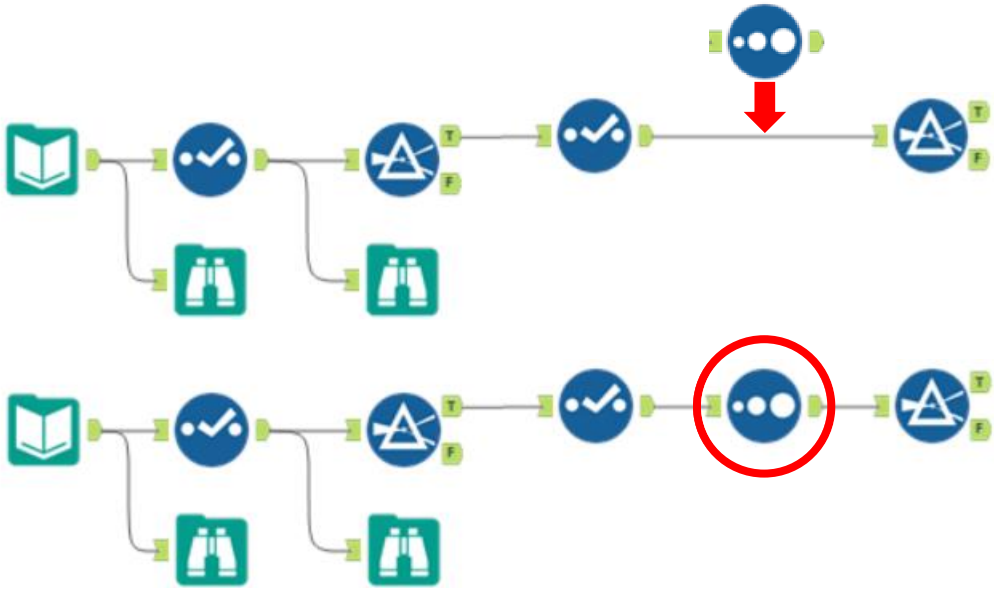
Results - Browse (5) - Input

9 of 9 Fields | Cell Viewer | 2 records displayed, 2,248 bytes

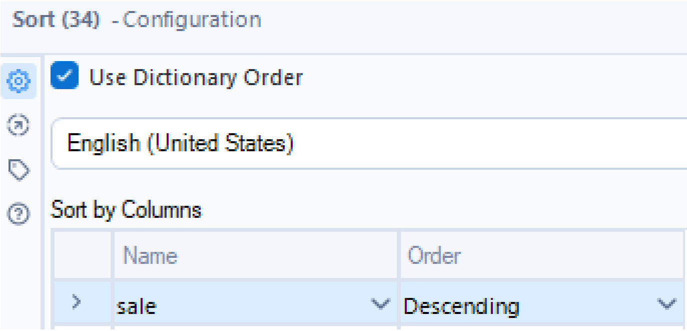
Record	fyear	tic	conm	at	cogs	ib	oibdp	sale	naics
1	2016	CMBM	CAMBIUM NETWORKS CORP	[Null]	[Null]	[Null]	[Null]	[Null]	334220
2	2016	MIMO	AIRSPAN NETWORKS HOLDINGS IN	[Null]	[Null]	[Null]	[Null]	[Null]	334220

It is often useful to examine the top and bottom values in a dataset. A Sample Tool can be used to select only a subset of records depending on how the dataset is sorted and how the tool is configured. We will add a Sort Tool and a Sample Tool to the workflow to examine the top three and bottom three records in the dataset.

Drag a Sort Tool from the Preparation Tools category onto the Canvas and place it on the connection between the second Select Tool and Filter Tool. The Sort Tool should automatically be inserted into the workflow, connected to the surrounding tools.



Click on the Sort Tool and in the Configuration Window configure the tool as shown below.



Sort (34) - Configuration

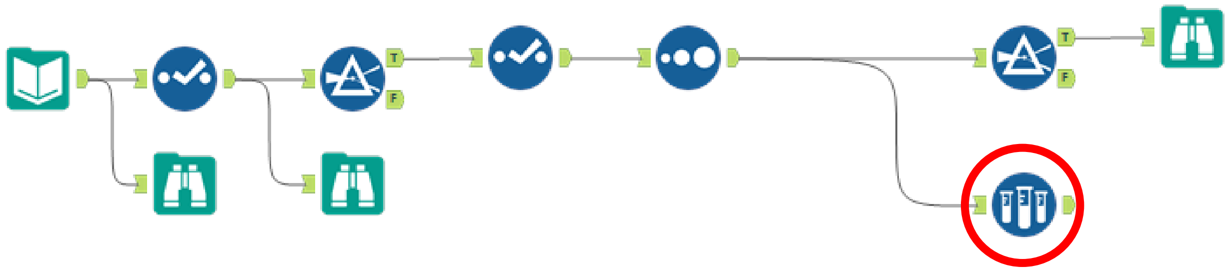
☑ Use Dictionary Order

English (United States)

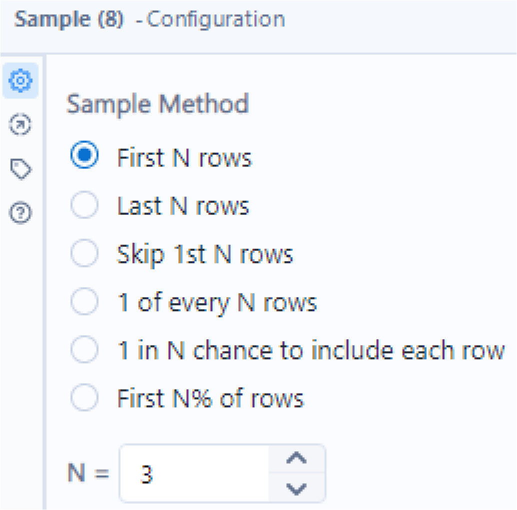
Sort by Columns

	Name	Order
>	sale	Descending

This sorts the records from largest to smallest, as measured by Net Sales. To choose only a subset of these records we can add a Sample Tool to the workflow after the Sort Tool.



Drag a Sample Tool from the Preparation Tools category onto the Canvas and connect it to the Sort Tool. Click on the Sample Tool and in the Configuration Window and configure the tool as shown below.



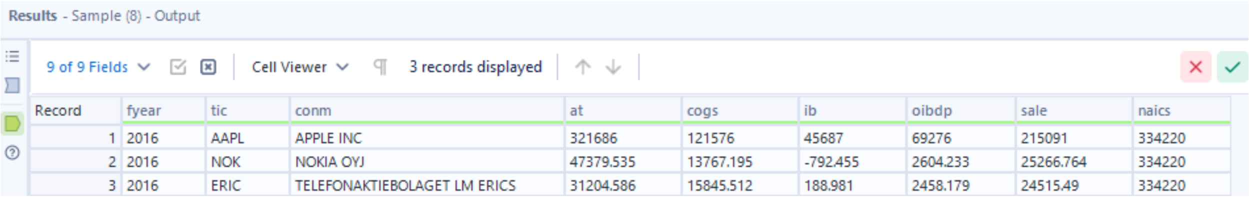
Sample (8) - Configuration

Sample Method

● First N rows
○ Last N rows
○ Skip 1st N rows
○ 1 of every N rows
○ 1 in N chance to include each row
○ First N% of rows

N = 3

Run the workflow. The resulting dataset contains only the three largest companies.



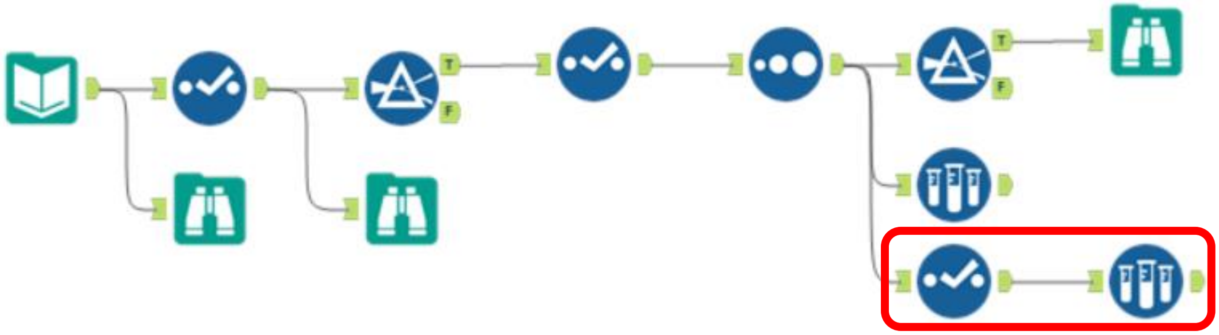
Results - Sample (8) - Output

9 of 9 Fields | Cell Viewer | 3 records displayed

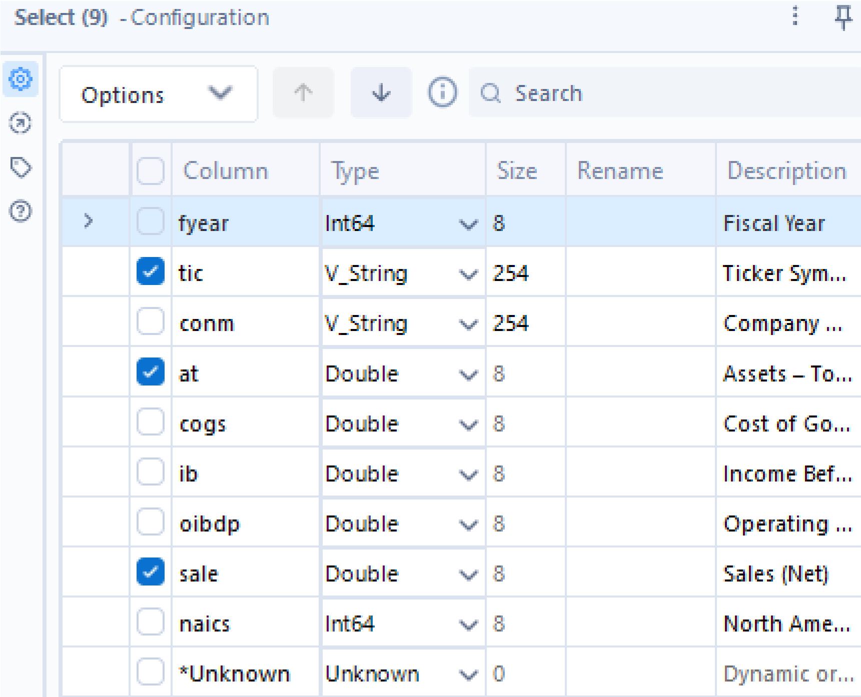
Record	fyear	tic	conm	at	cogs	ib	oibdp	sale	naics
1	2016	AAPL	APPLE INC	321686	121576	45687	69276	215091	334220
2	2016	NOK	NOKIA OYJ	47379.535	13767.195	-792.455	2604.233	25266.764	334220
3	2016	ERIC	TELEFONAKTIEBOLAGET LM ERICS	31204.586	15845.512	188.981	2458.179	24515.49	334220

Changing the sample method to the last N rows in the Configuration Window would result in a dataset with the three companies with the lowest sales. It is useful to click on the input anchor of the Sample Tool to see precisely which dataset is being sampled. In a large and complex workflow this is sometimes easier than finding the output anchor of the tool that the Sample Tool is connected to.

Any time a user is interested in fewer fields in the results a Select Tool can be used to choose fewer fields than the total number of available fields. Perhaps we only need the variables ticker symbol (tic), net sales (sale), and total assets (at). We can add another Select Tool to the workflow to choose only these three fields, and a Sample Tool to examine the top three observations in this reduced dataset.



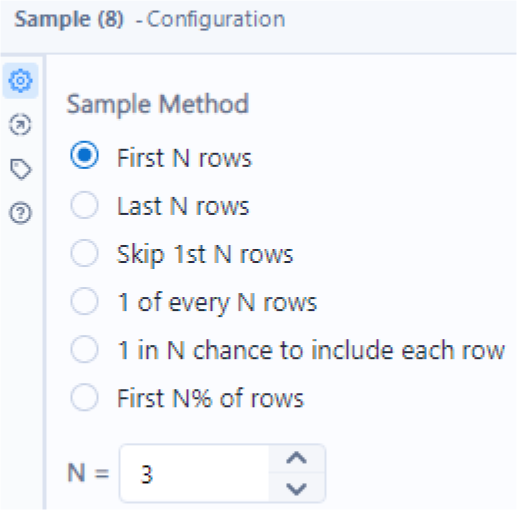
Drag a Select Tool from the Preparation Tools category onto the Canvas and connect it to the Sort Tool. Click on the Select Tool and in the Configuration Window configure the tool as shown below.



Select (9) - Configuration

	☐	Column	Type	Size	Rename	Description
>	☐	fyear	Int64	8		Fiscal Year
	☑	tic	V_String	254		Ticker Sym...
	☐	conm	V_String	254		Company ...
	☑	at	Double	8		Assets – To...
	☐	cogs	Double	8		Cost of Go...
	☐	ib	Double	8		Income Bef...
	☐	oibdp	Double	8		Operating ...
	☑	sale	Double	8		Sales (Net)
	☐	naics	Int64	8		North Ame...
	☐	*Unknown	Unknown	0		Dynamic or...

Drag another Sample Tool from the Preparation Tools category onto the Canvas and connect it to the Select Tool. Click on the Filter Tool and in the Configuration Window configure the tool as shown below and then run the workflow.



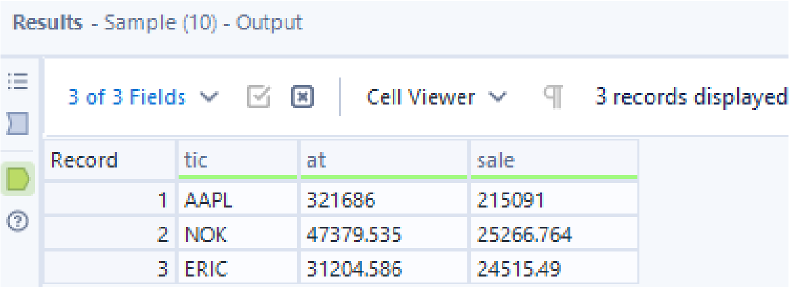
Sample (8) - Configuration

Sample Method

● First N rows
○ Last N rows
○ Skip 1st N rows
○ 1 of every N rows
○ 1 in N chance to include each row
○ First N% of rows

N = 3

As before, the resulting dataset contains the three largest companies, but now it only has the three fields chosen by the Select Tool.

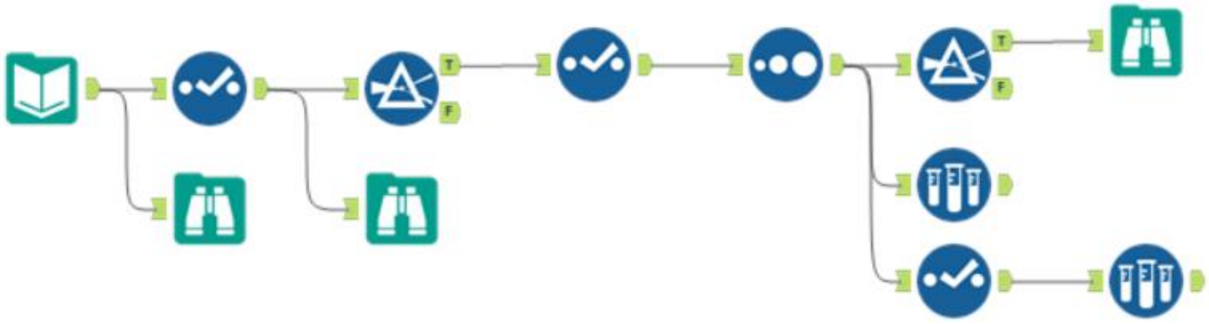


Results - Sample (10) - Output

3 of 3 Fields | Cell Viewer | 3 records displayed

Record	tic	at	sale
1	AAPL	321686	215091
2	NOK	47379.535	25266.764
3	ERIC	31204.586	24515.49

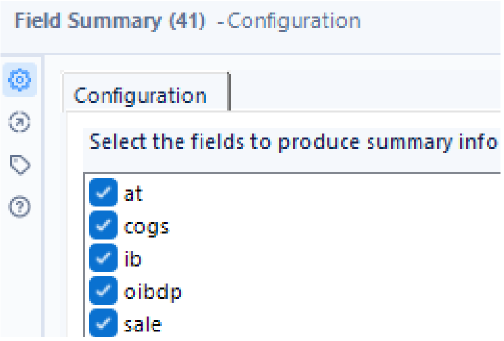
Your workflow should look like the one shown below. Save your workflow file because you will use it again to complete In-Chapter Practice Problem 2.3.2 and later in the chapter.



In-Chapter Practice Problem 2.3.2 (ICPP 2.3.2)

Download and unpack the Alteryx Workflow Package *2.3.2 ICPP.yxzp*. Modify the workflow *2.3.2 ICPP.yxmd* as follows.

1. Using Sort and Sample Tools, show the top six observations in the dataset, based on size (sales).
2. Using Sort and Sample Tools, show the bottom five observations in the dataset, based on size (sales).
3. Using Select, Sort, and Sample Tools, show the top 3 observations in the dataset, based on income before extraordinary items (*ib*). Display only the fields ticker symbol, sales, and income before extraordinary items, in that order.
4. Using Select, Sort, and Sample Tools, show the bottom 5 observations in the dataset, based on size (sales). Display only the fields company name, fiscal year, sales, and cost of goods sold, in that order.
5. Use a Field Summary Tool to generate summary statistics for the financial variables following the configuration shown below. Note that no other, non-financial variables, are selected.



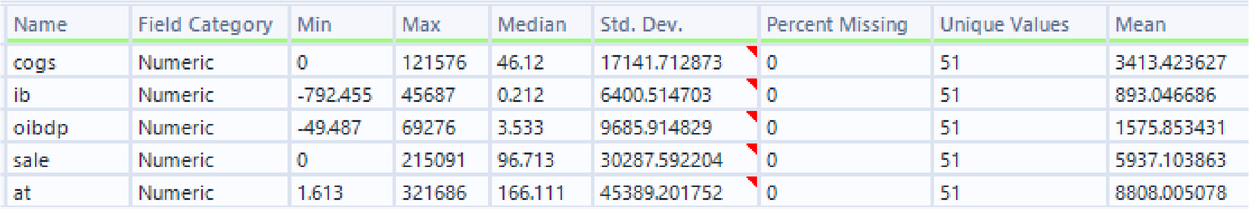
Field Summary (41) - Configuration

Configuration

Select the fields to produce summary info

☑ at
☑ cogs
☑ ib
☑ oibdp
☑ sale

The output of the first 9 columns of the "O" output anchor should look like this:



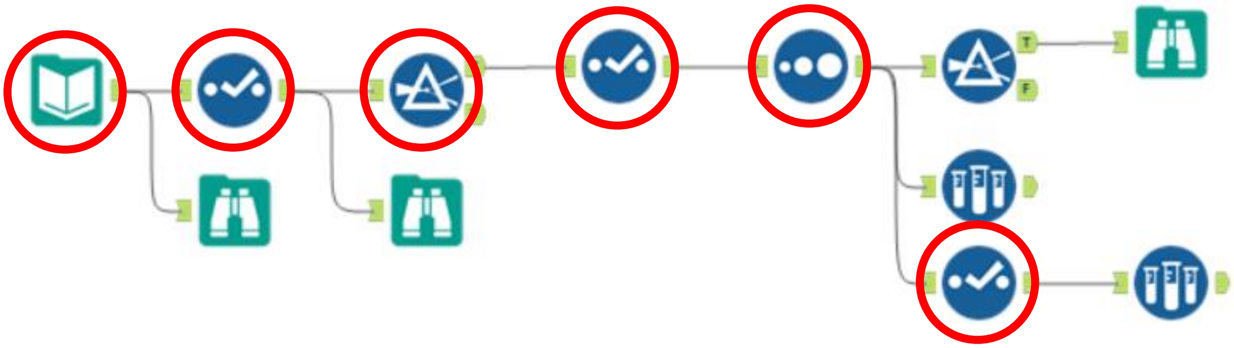
Name	Field Category	Min	Max	Median	Std. Dev.	Percent Missing	Unique Values	Mean
cogs	Numeric	0	121576	46.12	17141.712873	0	51	3413.423627
ib	Numeric	-792.455	45687	0.212	6400.514703	0	51	893.046686
oibdp	Numeric	-49.487	69276	3.533	9685.914829	0	51	1575.853431
sale	Numeric	0	215091	96.713	30287.592204	0	51	5937.103863
at	Numeric	1.613	321686	166.111	45389.201752	0	51	8808.005078

2.3.3 Critical Thinking

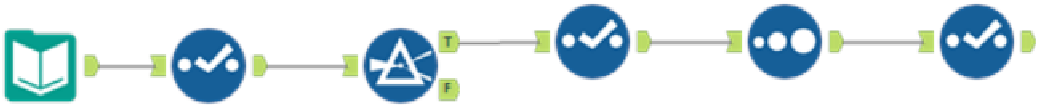
Summary statistics show at least one firm with zero *sales* and *cogs*. That does not make sense from an accounting perspective. One option would be removing the firms with zero *sales* or *cogs*. However, when we create profitability ratios, we divide by sales. If sales are small, the ratio will be very large. Therefore, it makes sense to remove firms with sales of less than $1 million. Because the data is expressed in millions, this means removing all companies with sales less than 1.

Start this section with the workflow from section 2.3.2. Before you make any changes to your workflow, save the workflow with a new name using "Save As" from the file menu.

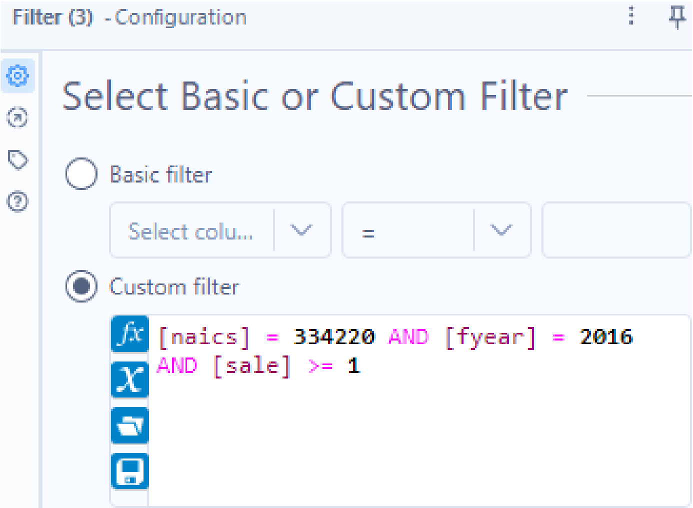
Delete the extra tools used to examine the data or subsets of the data. Keep only the six tools circled below.



Move the Select Tool in line with the Sort Tool and your workflow should now look like this.



We can use the first Filter Tool to remove the observations with sales of less than 1 million. Change the configuration of the Filter Tool in the workflow as shown below. The Filter Tool allows us to keep the observations that are in Apple's industry, for fiscal year 2016, and sales >= 1. This will remove three observations from the dataset.


Filter (3) - Configuration

Select Basic or Custom Filter

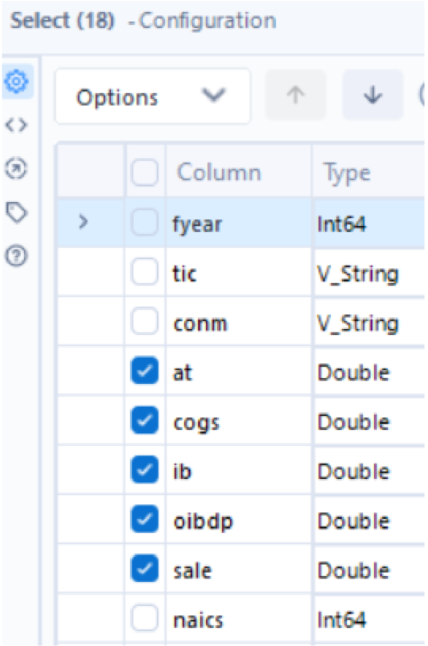
○ Basic filter

Select colu... =

● Custom filter

```
[naics] = 334220 AND [fyear] = 2016
AND [sale] >= 1
```

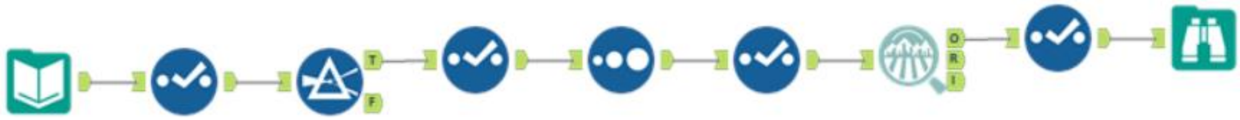
The Select Tool at the end of the workflow needs to be reconfigured as shown below to include the fields we will use for further analysis.

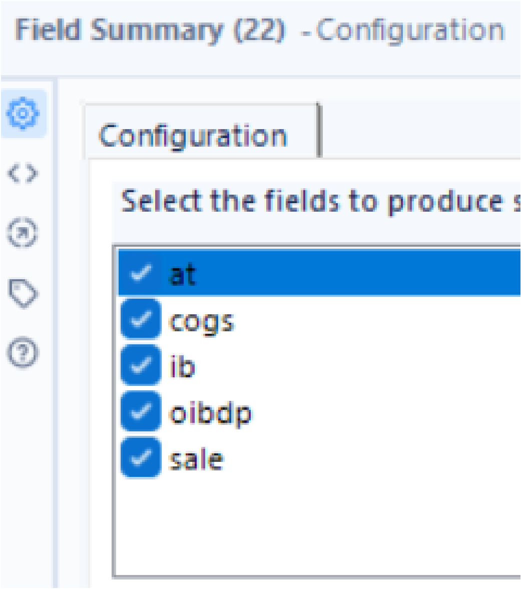

Select (18) - Configuration

Options

□	Column	Type
□	fyear	Int64
□	tic	V_String
□	conm	V_String
☑	at	Double
☑	cogs	Double
☑	ib	Double
☑	oibdp	Double
☑	sale	Double
□	naics	Int64

Add a Field Summary Tool to generate summary statistics for the financial variables and a Select Tool to make the output easier to read. Follow the configurations shown below and run the workflow.



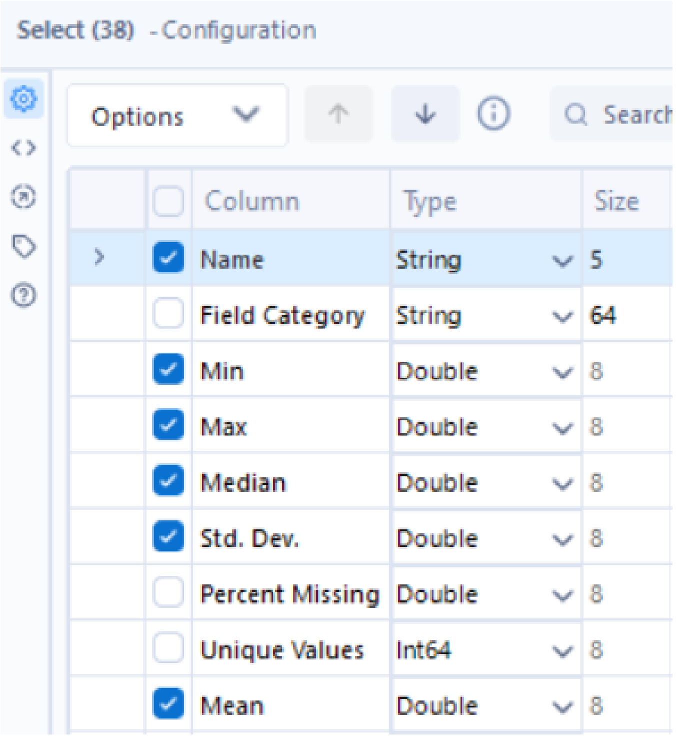


Field Summary (22) - Configuration

Configuration

Select the fields to produce s

- ☑ at
- ☑ cogs
- ☑ ib
- ☑ oibdp
- ☑ sale

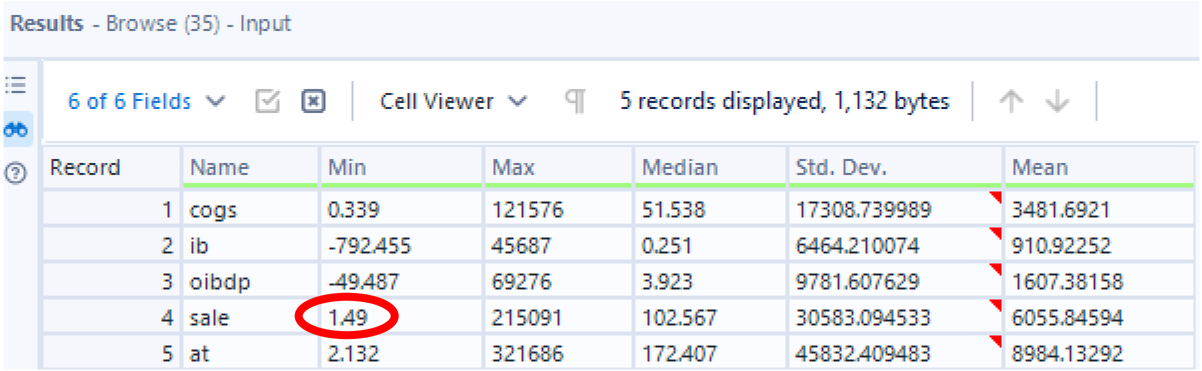


Select (38) - Configuration

	☐ Column	Type	Size
›	☑ Name	String	5
	☐ Field Category	String	64
	☑ Min	Double	8
	☑ Max	Double	8
	☑ Median	Double	8
	☑ Std. Dev.	Double	8
	☐ Percent Missing	Double	8
	☐ Unique Values	Int64	8
	☑ Mean	Double	8

Note: No other variables are selected.

After running the workflow, the summary statistics show that the smallest sales observation is over one million (1.49 = $1.49 million).



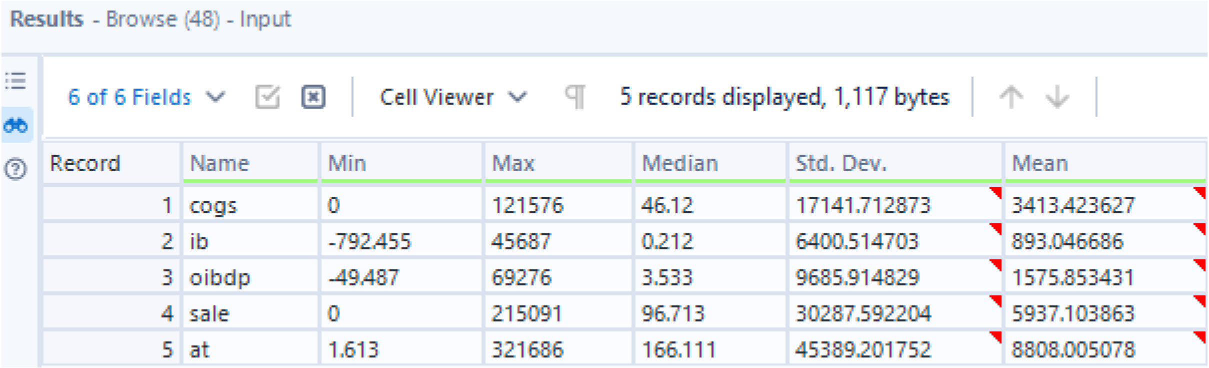
Results - Browse (35) - Input

5 records displayed, 1,132 bytes

Record	Name	Min	Max	Median	Std. Dev.	Mean
1	cogs	0.339	121576	51.538	17308.739989	3481.6921
2	ib	-792.455	45687	0.251	6464.210074	910.92252
3	oibdp	-49.487	69276	3.923	9781.607629	1607.38158
4	sale	1.49	215091	102.567	30583.094533	6055.84594
5	at	2.132	321686	172.407	45832.409483	8984.13292

In-Chapter Practice Problem 2.3.3

Repeat the above analysis by imposing the constraint that sales should simply be larger than zero. Compare your results with those above.



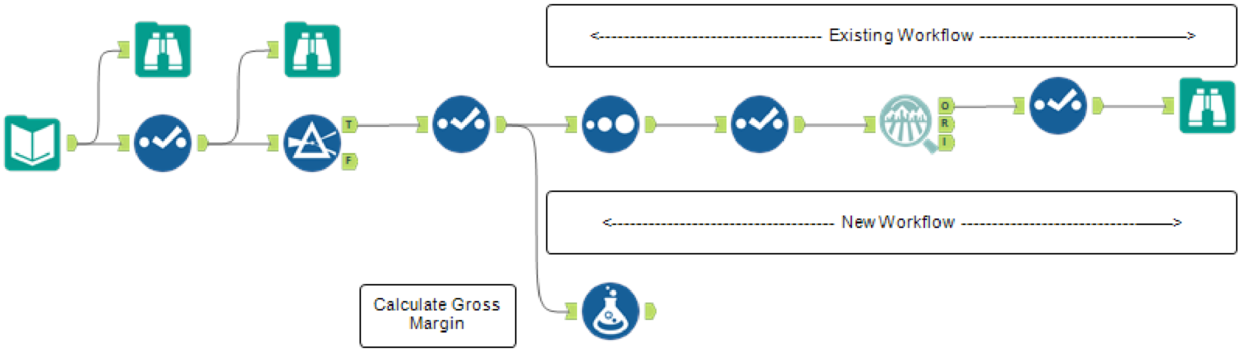
Results - Browse (48) - Input

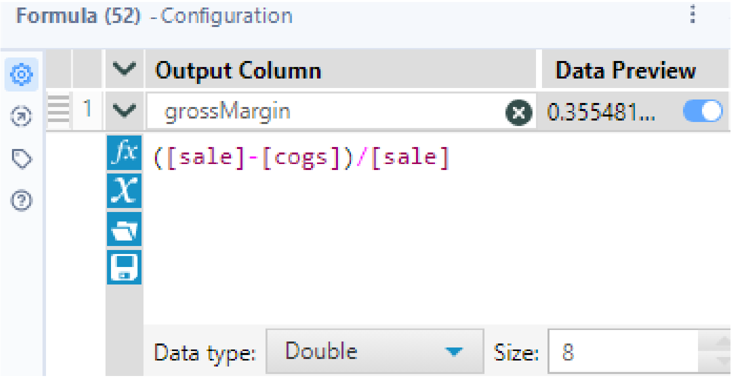
6 of 6 Fields | Cell Viewer | 5 records displayed, 1,117 bytes

Record	Name	Min	Max	Median	Std. Dev.	Mean
1	cogs	0	121576	46.12	17141.712873	3413.423627
2	ib	-792.455	45687	0.212	6400.514703	893.046686
3	oibdp	-49.487	69276	3.533	9685.914829	1575.853431
4	sale	0	215091	96.713	30287.592204	5937.103863
5	at	1.613	321686	166.111	45389.201752	8808.005078

2.3.4 Data Preparation

Add a Formula Tool to the workflow after the second Select Tool to calculate Gross Margin. Click on the Formula Tool and configure it in the Configuration Window as shown below.





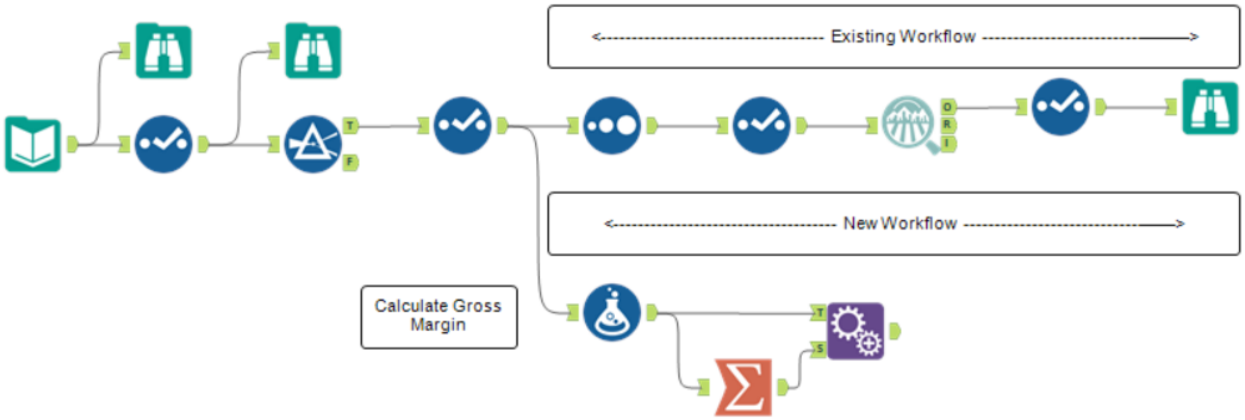
Formula (52) - Configuration

	Output Column	Data Preview
1	grossMargin	0.355481...

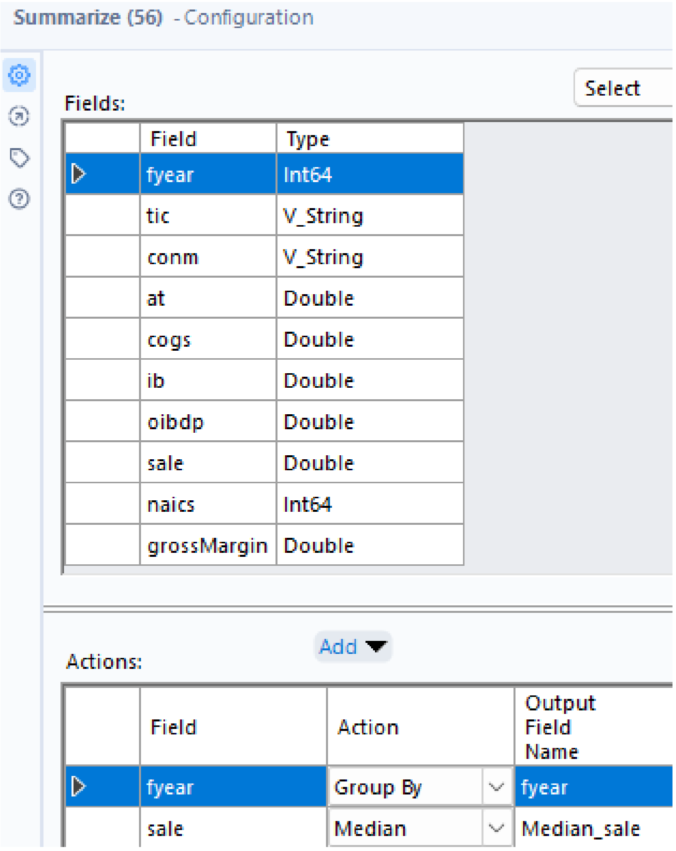
```
([sale]-[cogs])/[sale]
```

Data type: Double Size: 8

To distinguish between large and small firms, we will create a size variable. The new variable will take the value *large* if the firm's sales are equal to or above the industry median and the value *small* if the sales are below the median. But first, we need to add a new variable to the dataset equal to the industry median. Use a Summarize Tool to calculate median sales.



Add a Summarize Tool to the workflow after the Formula Tool and configure it as shown below. Even though our dataset only contains data for the fiscal year 2016, we will group by the fiscal year variable. This makes the workflow more powerful, because it can be reused with other datasets that contain more than one year.



Summarize (56) - Configuration

Select

Fields:

	Field	Type
▷	fyear	Int64
	tic	V_String
	conm	V_String
	at	Double
	cogs	Double
	ib	Double
	oibdp	Double
	sale	Double
	naics	Int64
	grossMargin	Double

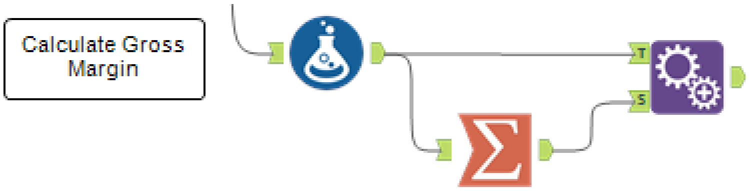
Actions:

Add ▼

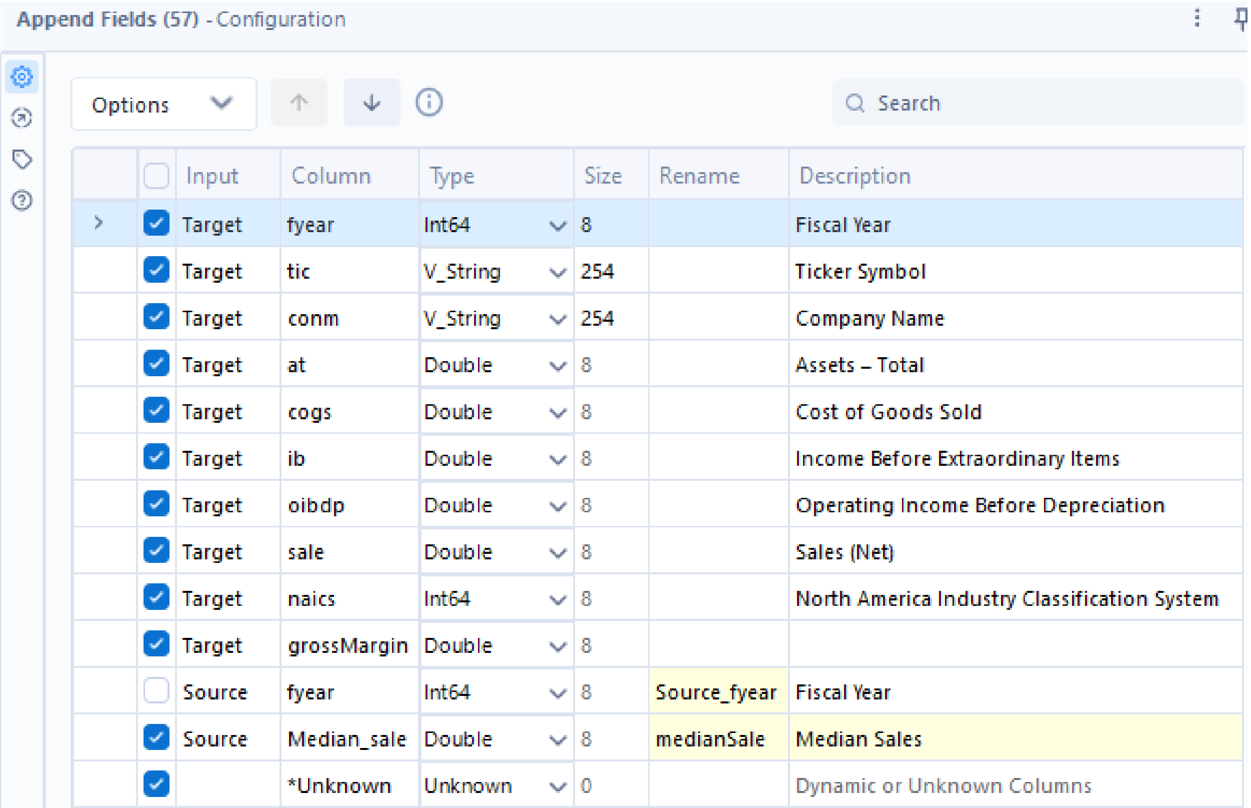
	Field	Action	Output Field Name
▷	fyear	Group By	fyear
	sale	Median	Median_sale

Firms with sales over around $103 million are considered large, and we can use this to create the large firm classification.

Add an Append Fields Tool to the workflow to add the median sales value to the dataset. Connect the Formula Tool to the “T” input anchor and the Summarize Tool to the “S” input anchor.

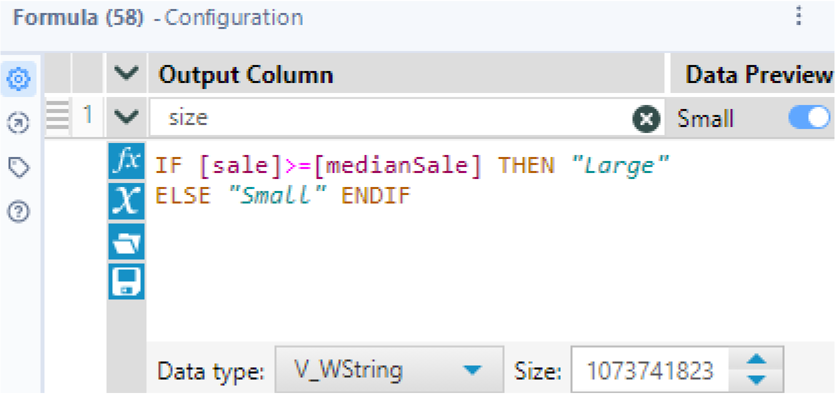


Configure the Append Fields Tool as shown below.


Append Fields (57) - Configuration

	☐	Input	Column	Type	Size	Rename	Description
>	☑	Target	fyear	Int64	8		Fiscal Year
	☑	Target	tic	V_String	254		Ticker Symbol
	☑	Target	conm	V_String	254		Company Name
	☑	Target	at	Double	8		Assets – Total
	☑	Target	cogs	Double	8		Cost of Goods Sold
	☑	Target	ib	Double	8		Income Before Extraordinary Items
	☑	Target	oibdp	Double	8		Operating Income Before Depreciation
	☑	Target	sale	Double	8		Sales (Net)
	☑	Target	naics	Int64	8		North America Industry Classification System
	☑	Target	grossMargin	Double	8		
	☐	Source	fyear	Int64	8	Source_fyear	Fiscal Year
	☑	Source	Median_sale	Double	8	medianSale	Median Sales
	☑		*Unknown	Unknown	0		Dynamic or Unknown Columns

Add another Formula Tool to the workflow after the Append Fields Tool. Configure the Formula Tool as shown below to categorize the firms as large or small.



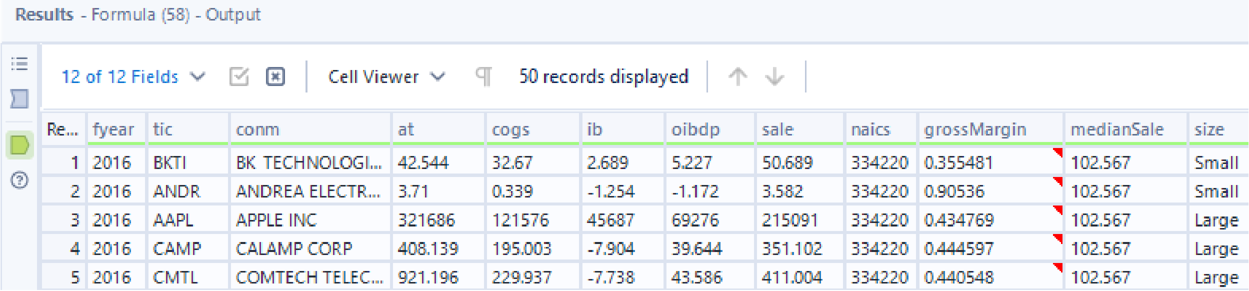
Formula (58) - Configuration

	Output Column	Data Preview
1	size	Small

```
IF [sale]>=[medianSale] THEN "Large"
ELSE "Small" ENDIF
```

Data type: V_WString Size: 1073741823

Before we start modeling our solution, we will take a look at the first few observations, as well as summary statistics. From the output anchor of the second Formula Tool:



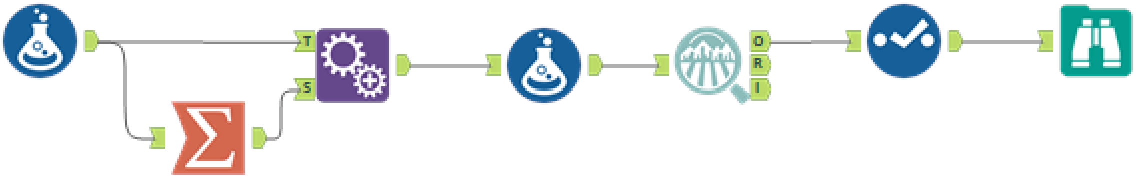
Results - Formula (58) - Output

12 of 12 Fields Cell Viewer 50 records displayed

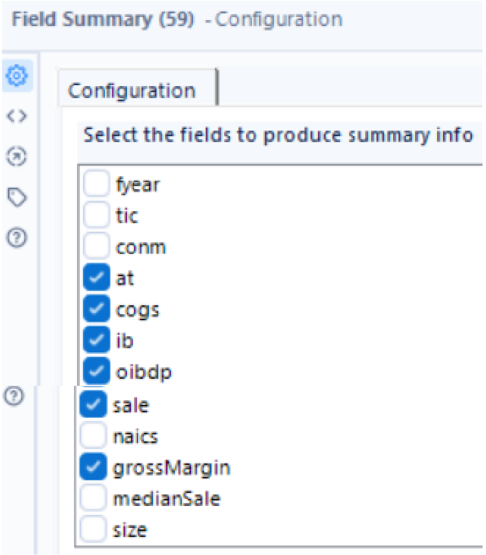
Re...	fyear	tic	conm	at	cogs	ib	oibdp	sale	naics	grossMargin	medianSale	size
1	2016	BKTI	BK TECHNOLOGI...	42.544	32.67	2.689	5.227	50.689	334220	0.355481	102.567	Small
2	2016	ANDR	ANDREA ELECTR...	3.71	0.339	-1.254	-1.172	3.582	334220	0.90536	102.567	Small
3	2016	AAPL	APPLE INC	321686	121576	45687	69276	215091	334220	0.434769	102.567	Large
4	2016	CAMP	CALAMP CORP	408.139	195.003	-7.904	39.644	351.102	334220	0.444597	102.567	Large
5	2016	CMTL	COMTECH TELEC...	921.196	229.937	-7.738	43.586	411.004	334220	0.440548	102.567	Large

We can see that Apple, with sales greater than 103 million, has been classified as *large*, and firms with sales below 103 million, as *small*.

The Field Summary Tool is an easy way to calculate summary statistics for a dataset. Add a Field Summary Tool, a Select Tool, and a Browse Tool to the workflow as shown and configure the tools in their Configuration Windows as shown below.



In the **Field Summary Tool** Configuration Window unselect fyear, tic, conm, naics, medianSale, and size.



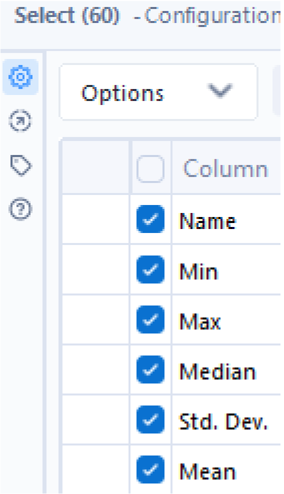
Field Summary (59) - Configuration

Configuration

Select the fields to produce summary info

- ☐ fyear
- ☐ tic
- ☐ conm
- ☑ at
- ☑ cogs
- ☑ ib
- ☑ oibdp
- ☑ sale
- ☐ naics
- ☑ grossMargin
- ☐ medianSale
- ☐ size

In the **Select Tool** Configuration Window select only the fields shown.

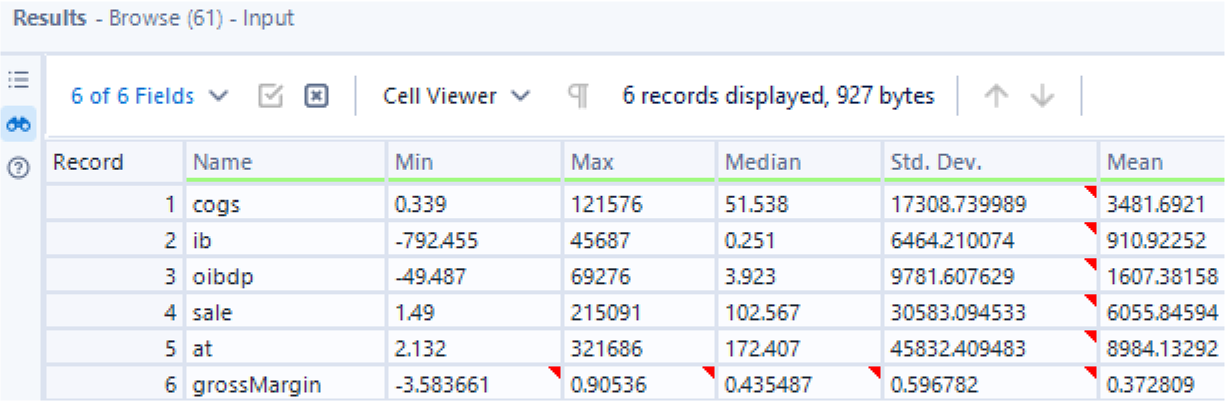


Select (60) - Configuration

Options

☐ Column
☑ Name
☑ Min
☑ Max
☑ Median
☑ Std. Dev.
☑ Mean

Run the Workflow. The Results Window of the Browse Tool will display basic summary statistics for the selected fields in the dataset.



Results - Browse (61) - Input

6 of 6 Fields | Cell Viewer | 6 records displayed, 927 bytes

Record	Name	Min	Max	Median	Std. Dev.	Mean
1	cogs	0.339	121576	51.538	17308.739989	3481.6921
2	ib	-792.455	45687	0.251	6464.210074	910.92252
3	oibdp	-49.487	69276	3.923	9781.607629	1607.38158
4	sale	1.49	215091	102.567	30583.094533	6055.84594
5	at	2.132	321686	172.407	45832.409483	8984.13292
6	grossMargin	-3.583661	0.90536	0.435487	0.596782	0.372809

Therefore, the average (mean) gross margin for firms in the same industry as Apple is around 37.3% and the median is around 43.5%.

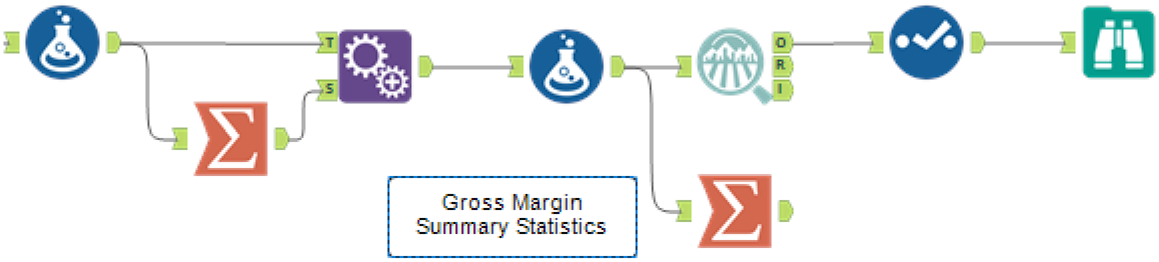
In-Chapter Practice Problem 2.3.4

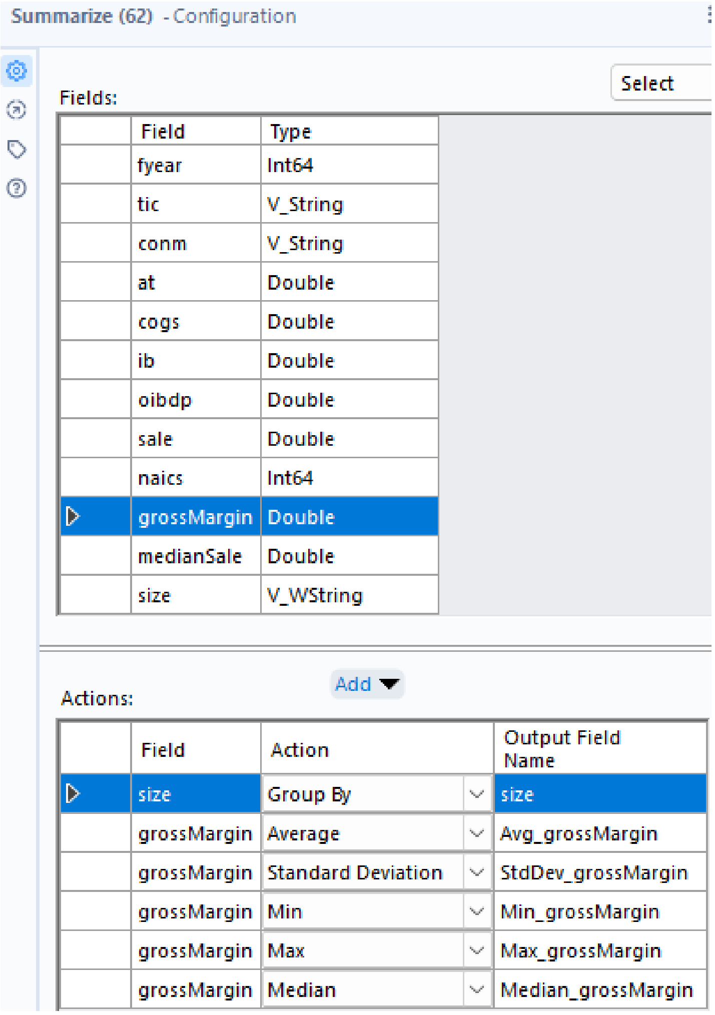
What does it indicate when the mean of a variable like *gross margin* is smaller than the median?

2.3.5 Model

To answer whether large firms are more profitable than small firms, we will generate summary statistics of gross margin for each group. First, we must group by firm size. This means splitting the dataset into two subsets. One for small firms and another for large firms. Second, generate aggregate statistical measures for each group (e.g., mean, standard deviation, median).

Use a Summarize Tool to generate summary statistics for *gross margin* across large and small firms as shown below.




Summarize (62) - Configuration

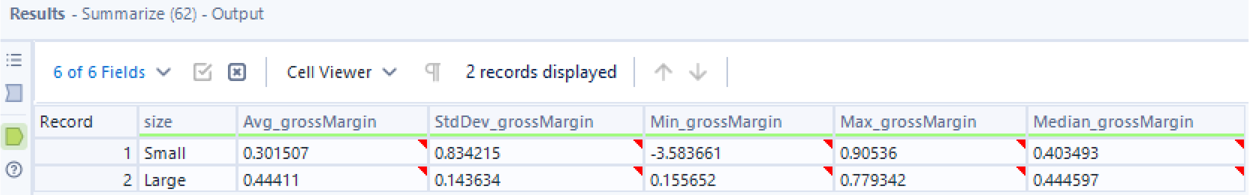
Fields:

	Field	Type
	fyear	Int64
	tic	V_String
	conm	V_String
	at	Double
	cogs	Double
	ib	Double
	oibdp	Double
	sale	Double
	naics	Int64
▷	grossMargin	Double
	medianSale	Double
	size	V_WString

Actions:

	Field	Action	Output Field Name
▷	size	Group By	size
	grossMargin	Average	Avg_grossMargin
	grossMargin	Standard Deviation	StdDev_grossMargin
	grossMargin	Min	Min_grossMargin
	grossMargin	Max	Max_grossMargin
	grossMargin	Median	Median_grossMargin

The output shows:



Record	size	Avg_grossMargin	StdDev_grossMargin	Min_grossMargin	Max_grossMargin	Median_grossMargin
1	Small	0.301507	0.834215	-3.583661	0.90536	0.403493
2	Large	0.44411	0.143634	0.155652	0.779342	0.444597

The large companies appear to have a higher average gross margin, 44.4% compared to 30.2% for small firms.

In-Chapter Practice Problem 2.3.5

Interpret the minimum value of gross margin for small and large firms.

2.3.6 Evaluation - Critical Thinking

At first glance, the results seem to indicate that,

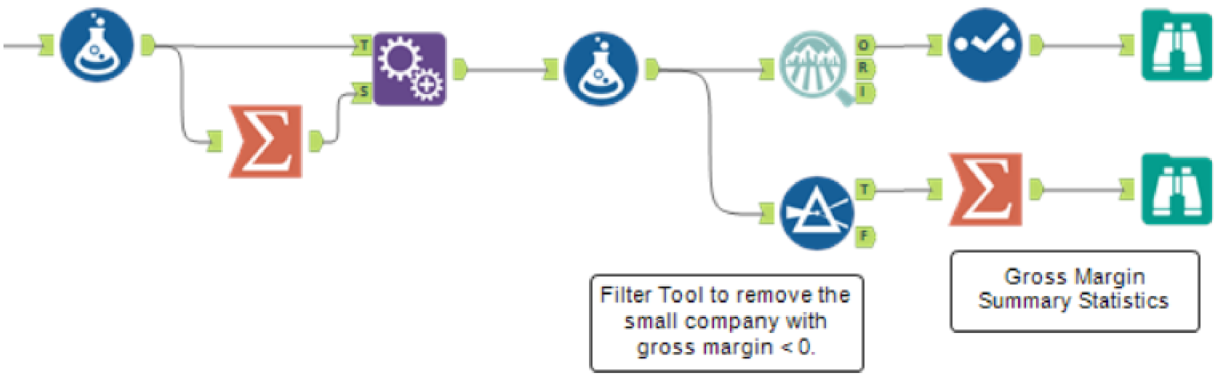
1. the average gross margin of small firms is smaller than the average gross margin of large firms, and
2. the median gross margin of small firms is smaller than the median gross margin of large firms.

Before we communicate our results to stakeholders, reviewing the results by leveraging accounting knowledge is a good idea. In particular, we want to pay attention to the gross margin of the small firms.

The average gross margin of the small firms could be due to an unusually low gross margin for one of the small companies. The minimum gross margin for the small companies is -3.58, or -358%. This means that one company is selling their inventory at around 22% of its cost. It is not reasonable to include this abnormal situation with the other companies when calculating the average.

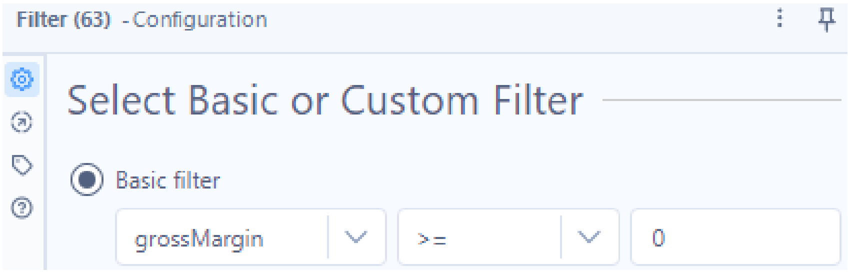
In statistics we learned that this type of observation is an *outlier*. At the very least we want to exclude companies with a negative gross margin from our analysis. Similarly to how we excluded very small companies previously, we will do this with a Filter Tool.

Add a Filter Tool to the workflow before the Summarize Tool to remove the observations with gross margin less than zero. Configure the Filter Tool in the workflow as shown below and run the workflow.



Filter Tool to remove the small company with gross margin < 0.

Gross Margin Summary Statistics



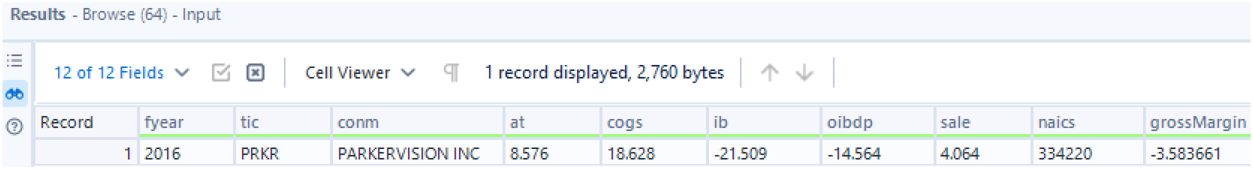
Filter (63) - Configuration

Select Basic or Custom Filter

● Basic filter

grossMargin | >= | 0

The Filter Tool's "T" output anchor (true) contains the dataset with the 49 companies that do not have a negative gross margin. The Filter Tool's "F" output anchor (false) contains the dataset with the 1 company that has a negative gross margin:



Results - Browse (64) - Input

12 of 12 Fields | Cell Viewer | 1 record displayed, 2,760 bytes

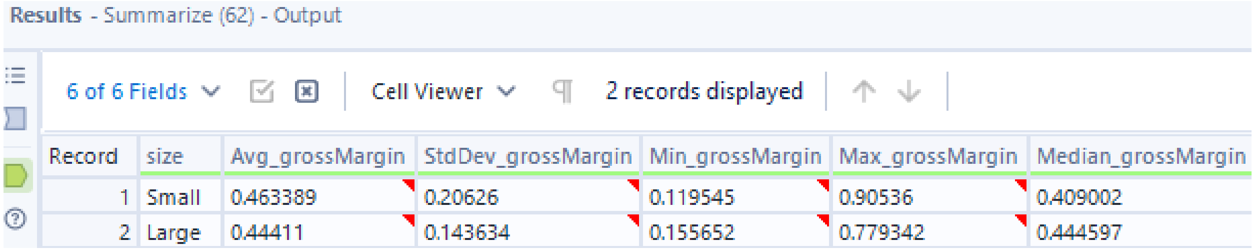
Record	fyear	tic	conm	at	cogs	ib	oibdp	sale	naics	grossMargin
1	2016	PRKR	PARKERVISION INC	8.576	18.628	-21.509	-14.564	4.064	334220	-3.583661

In-Chapter Practice Problem 2.3.6

1. What is the accounting interpretation and implication of the gross margin for this firm? Think in terms of its sales versus the cost of goods sold.
2. From an economic standpoint, does this value constitute an outlier? Remember that outliers are abnormal when compared to other observations in a dataset.

2.3.7 Re-Evaluate the Model

After removing the company with a negative gross margin, the summary statistics on the output anchor of the Summarize Tool show a different picture of the profitability of the large and small companies.



Results - Summarize (62) - Output

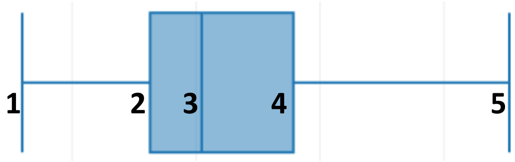
Record	size	Avg_grossMargin	StdDev_grossMargin	Min_grossMargin	Max_grossMargin	Median_grossMargin
1	Small	0.463389	0.20626	0.119545	0.90536	0.409002
2	Large	0.44411	0.143634	0.155652	0.779342	0.444597

The large companies now have a lower mean gross margin, 44.4% compared to 46.3% for small firms. However, the median gross margin for the large companies is still higher than for the small companies, 44.4% compared to 40.9%.

2.4 Visualization

Given that some people may not feel comfortable with the statistical analysis, it is a good idea - whenever possible - to generate appropriate visualizations. In the following sections, we will introduce the creation of box and whisker plots using the Interactive Chart Tool in Alteryx.

Box and whisker plots, commonly known as boxplots, are useful because they tell users a lot of information about the underlying distribution of the displayed variable. There are two common types of boxplots and the first is displayed below.



Each of the five vertical lines in the boxplot corresponds to a characteristic of the distribution, as follows:

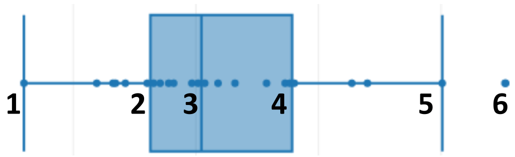
1. Minimum Value
2. 25th Percentile Value
3. Median Value
4. 75th Percentile Value
5. Maximum Value

The second type of boxplot is similar to the first, but it contains information about the individual datapoints in the distribution, which are plotted as points on the boxplot. It is useful because it indicates outliers, using a common rule for outliers based on the interquartile range (IQR), where IQR is the distance between the 25th and 75th Percentiles, or first and third quartiles.

To identify outliers, we can use the first quartile (Q1), third quartile (Q3), and the inter-quartile range (IQR = Q3 - Q1) as follows:

- Any value below the lower bound (*LB*) is an outlier.
- Where (LB = Q1 - 1.5 × IQR)
- Similarly, any value above the upper bound (*UB*) is an outlier.
- Where (UB = Q3 + 1.5 × IQR)

Any outliers will be shown in the boxplot as below the lower whisker or above the upper whisker.

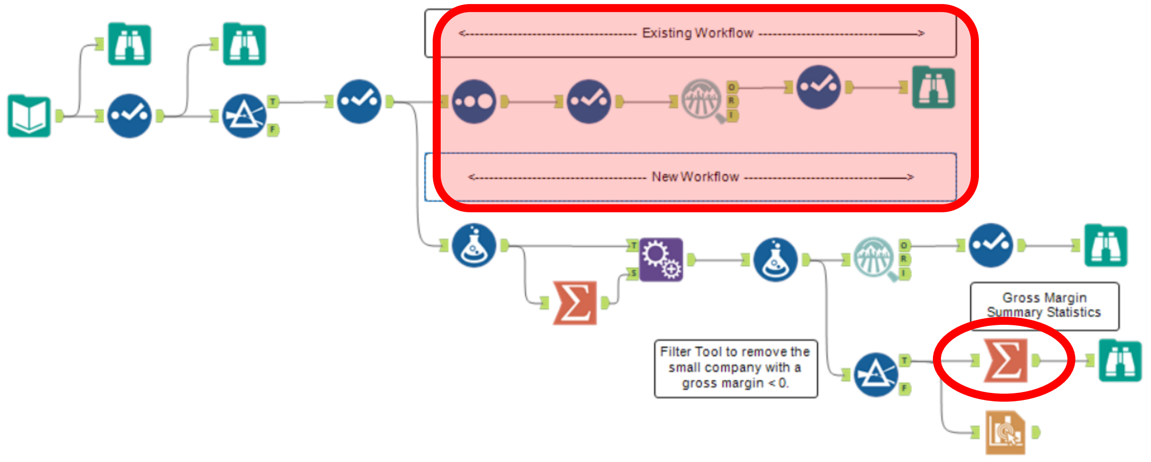


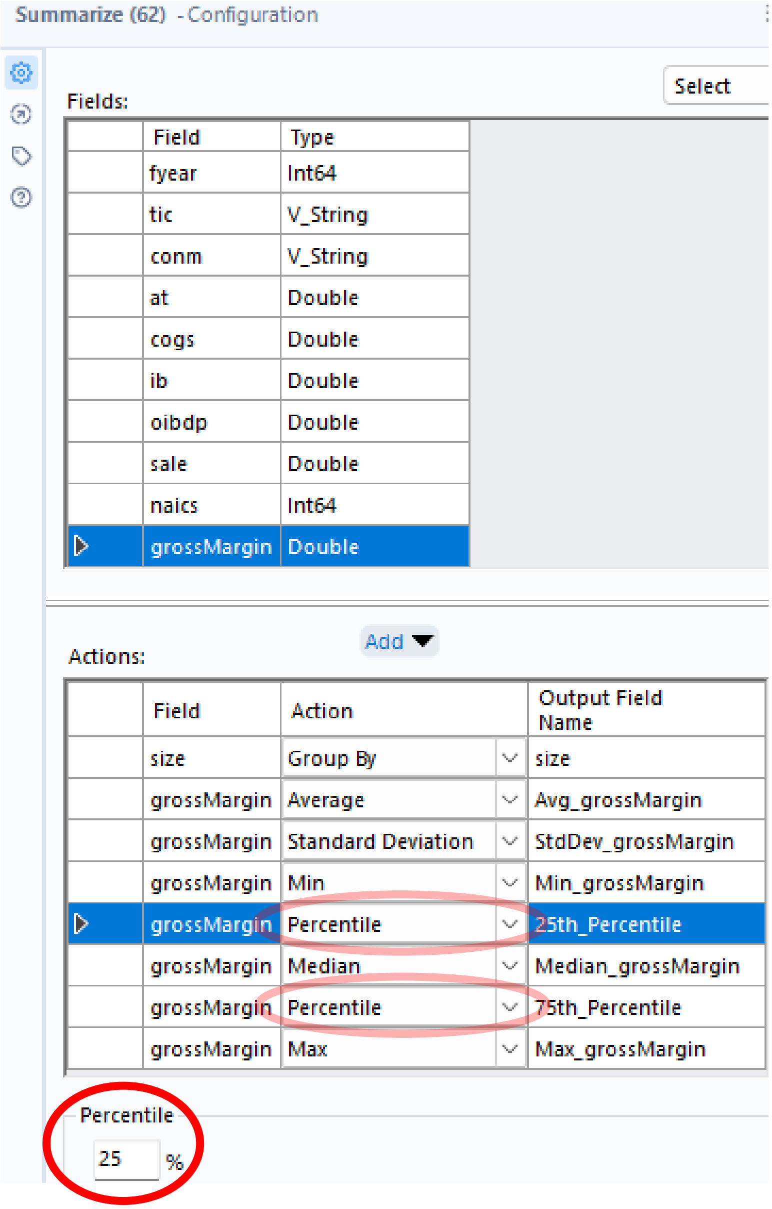
Each of the six points in the boxplot corresponds to a characteristic of the distribution, as follows:

1. Smallest value within the lower bound; therefore, not an outlier. Any observations less than the lower bound are outliers and would show up as a single point, such as #6.
2. 25th Percentile Value
3. Median Value
4. 75th Percentile Value
5. Largest value within the upper bound; therefore, not an outlier. Any observations greater than the upper bound are outliers and show up as a single point, such as #6.
6. An example of an outlier, in this case an observation above the upper bound.

2.4.1 Boxplot in Alteryx

Use Save As from the File menu to save the workflow with a new name. Delete the workflow tools from section 2.3.3 as indicated below and reconfigure the second Summarize Tool to calculate the 25th and 75th percentiles for comparison to the boxplot. To set the percentile, there is a box on the bottom of the Configuration Window. Set it to 25 and 75 for the 25th and 75th percentiles, respectively.


Existing Workflow
New Workflow
Gross Margin Summary Statistics
Filter Tool to remove the small company with a gross margin < 0.



Fields:

	Field	Type
	fyear	Int64
	tic	V_String
	conm	V_String
	at	Double
	cogs	Double
	ib	Double
	oibdp	Double
	sale	Double
	naics	Int64
▷	grossMargin	Double

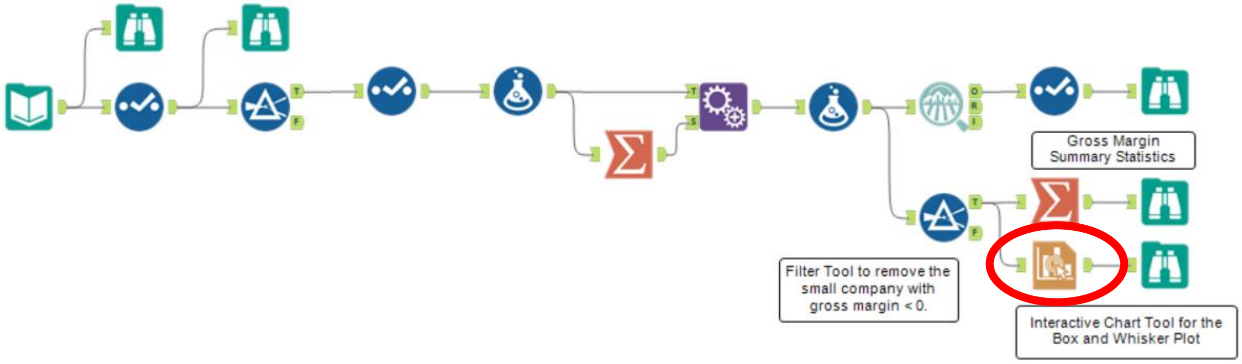
Actions:

	Field	Action	Output Field Name
	size	Group By	size
	grossMargin	Average	Avg_grossMargin
	grossMargin	Standard Deviation	StdDev_grossMargin
	grossMargin	Min	Min_grossMargin
▷	grossMargin	Percentile	25th_Percentile
	grossMargin	Median	Median_grossMargin
	grossMargin	Percentile	75th_Percentile
	grossMargin	Max	Max_grossMargin

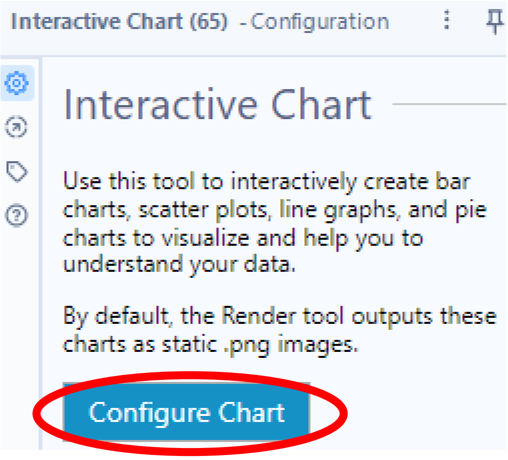
Percentile

25 %

Add an Interactive Chart Tool to the workflow on the “T” output anchor of the second Filter Tool as shown below.


Gross Margin Summary Statistics
Filter Tool to remove the small company with gross margin < 0.
Interactive Chart Tool for the Box and Whisker Plot

To launch the Interactive Chart Window, click on the Interactive Chart Tool and then click on the Configure Chart button in the Configuration Window.



Interactive Chart (65) - Configuration

Interactive Chart

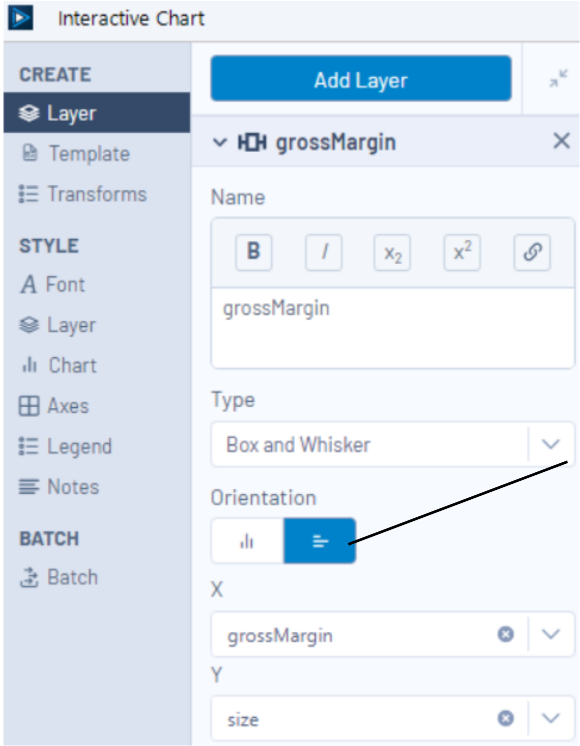
Use this tool to interactively create bar charts, scatter plots, line graphs, and pie charts to visualize and help you to understand your data.

By default, the Render tool outputs these charts as static .png images.

Configure the Interactive Chart Tool as shown below. In the Configuration Window, click on the “Configure Chart” button.

To create a Box and Whisker plot in the Interactive Chart Window:

1. Under CREATE, click on Layer and then the Add Layer button.
2. Name the layer, replacing the “layer 0” text in the Name Box.
3. Select “Box and Whisker” from the Type drop down list.
4. Select grossMargin for the X-axis and Size for the Y-axis as shown below.
5. Under CREATE, click on Layer and use the configuration shown below.



Name: grossMargin

Type: Box and Whisker

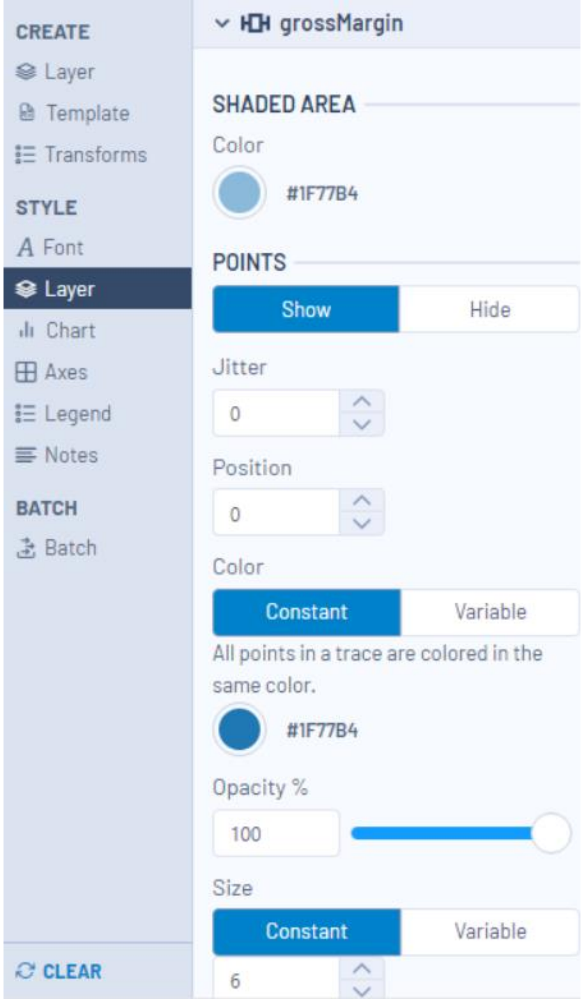
Orientation

X: grossMargin

Y: size

Sometimes the orientation option does not appear immediately. It will eventually appear as you continue configuring the box and whisker plot.

Layer Configuration



grossMargin

SHADED AREA

Color

#1F77B4

POINTS

[✓ Show] Hide

Jitter

0

Position

0

Color

[✓ Constant] Variable

All points in a trace are colored in the same color.

#1F77B4

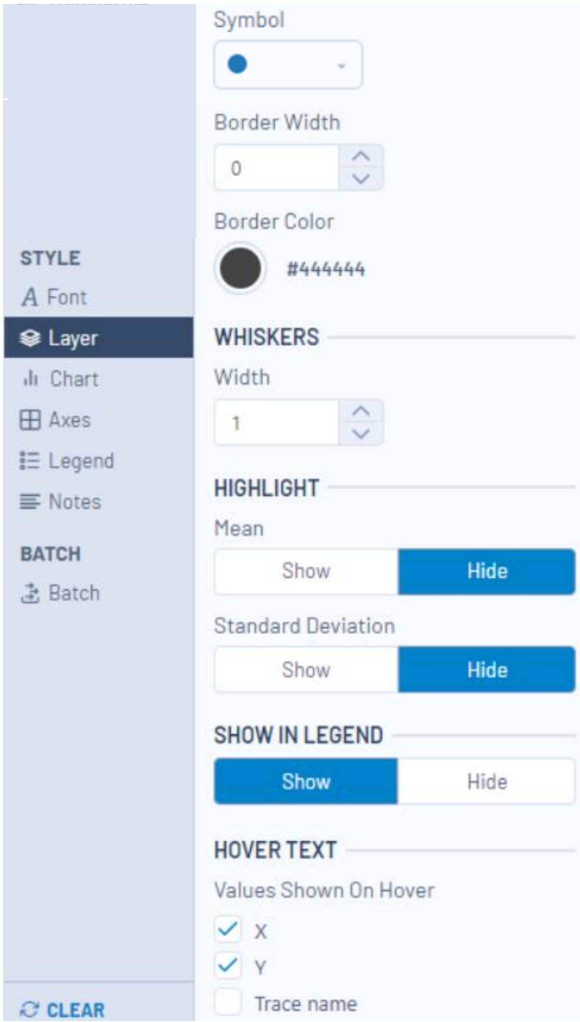
Opacity %

100

Size

[✓ Constant] Variable

6



Symbol

●

Border Width

0

Border Color

#444444

WHISKERS

Width

1

HIGHLIGHT

Mean

Show [✓ Hide]

Standard Deviation

Show [✓ Hide]

SHOW IN LEGEND

[✓ Show] Hide

HOVER TEXT

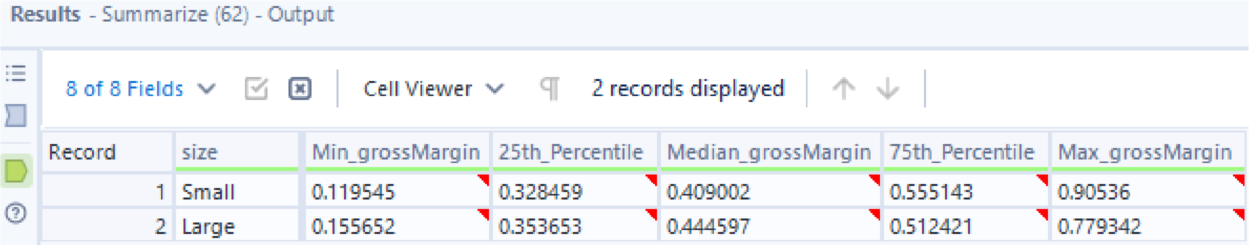
Values Shown On Hover

☑ X

☑ Y

□ Trace name

The resulting boxplots match the summary statistics from the output anchor of the Summarize Tool.



Results - Summarize (62) - Output

Record	size	Min_grossMargin	25th_Percentile	Median_grossMargin	75th_Percentile	Max_grossMargin
1	Small	0.119545	0.328459	0.409002	0.555143	0.90536
2	Large	0.155652	0.353653	0.444597	0.512421	0.779342

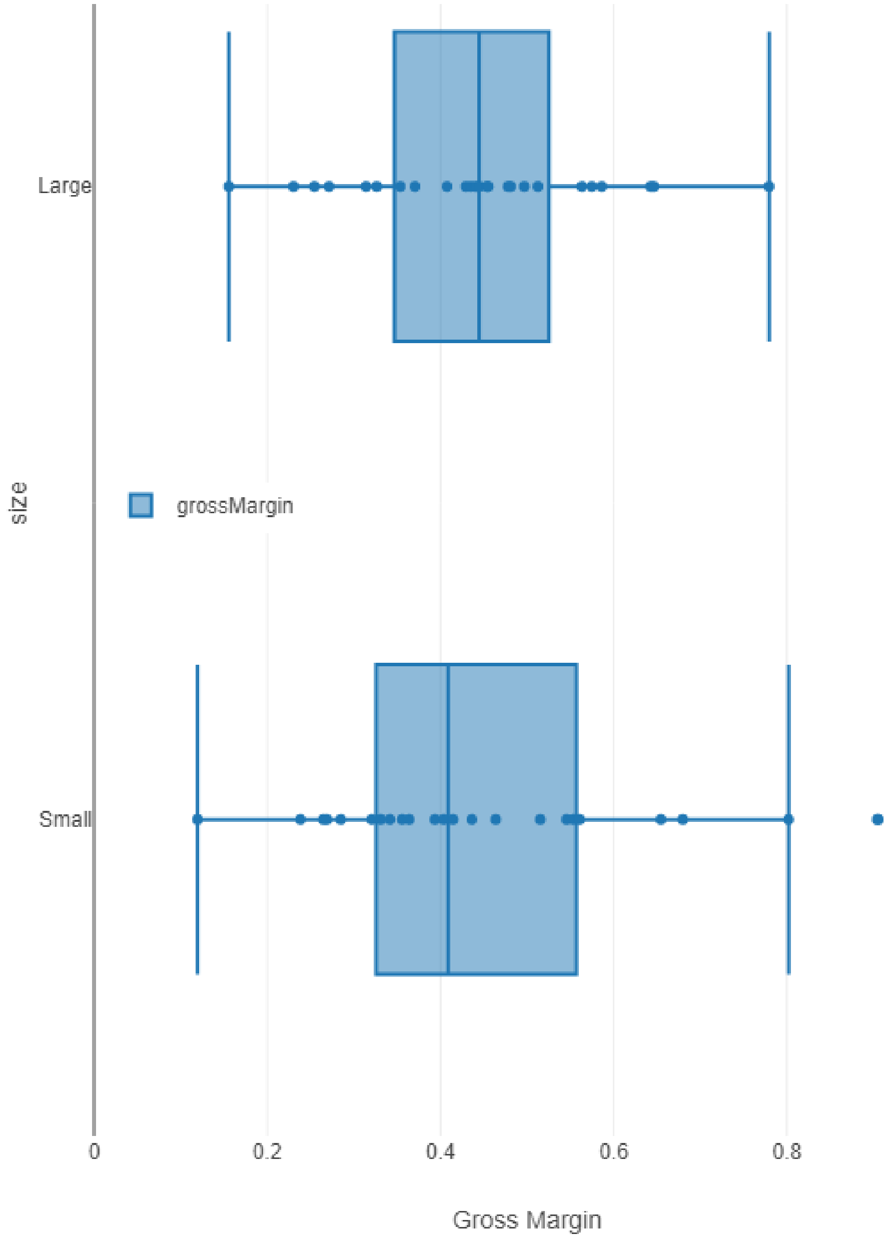


Figure 2.2: Graphical Comparison of Gross Margin

📝 In-Chapter Practice Problem 2.4.1

Communicating results, insights, and ideas is fundamental in data analysis. What is the main message that you want to communicate with this graph?

Hint: Compare the two boxed areas. They capture the 50% of observations around the median. Look at the width of these boxes. The dispersion (variance) is much wider among small firms. The standard deviation and inter-quartile range are two measures of dispersion.

The visualizations complement the summary statistics by showing that the centrality measures mean and median are close to one another. However,.

In-Chapter Practice Problem 2.4.2

Download and unpack the Alteryx Workflow Package *2.4.2 ICPP.yxzp*. Run the workflow 2.4.2 ICPP. Use the above rule that identifies outliers based on *ub*=Q3+1.5*IQR and *lb*=Q1-1.5*IQR to answer the following questions:

1. In the original dataset with 50 companies, does the company with grossMargin of -3.58 constitute an outlier?
2. Are there other outliers in the dataset?
3. Where do the outliers appear on the boxplot in Figure 2.3?

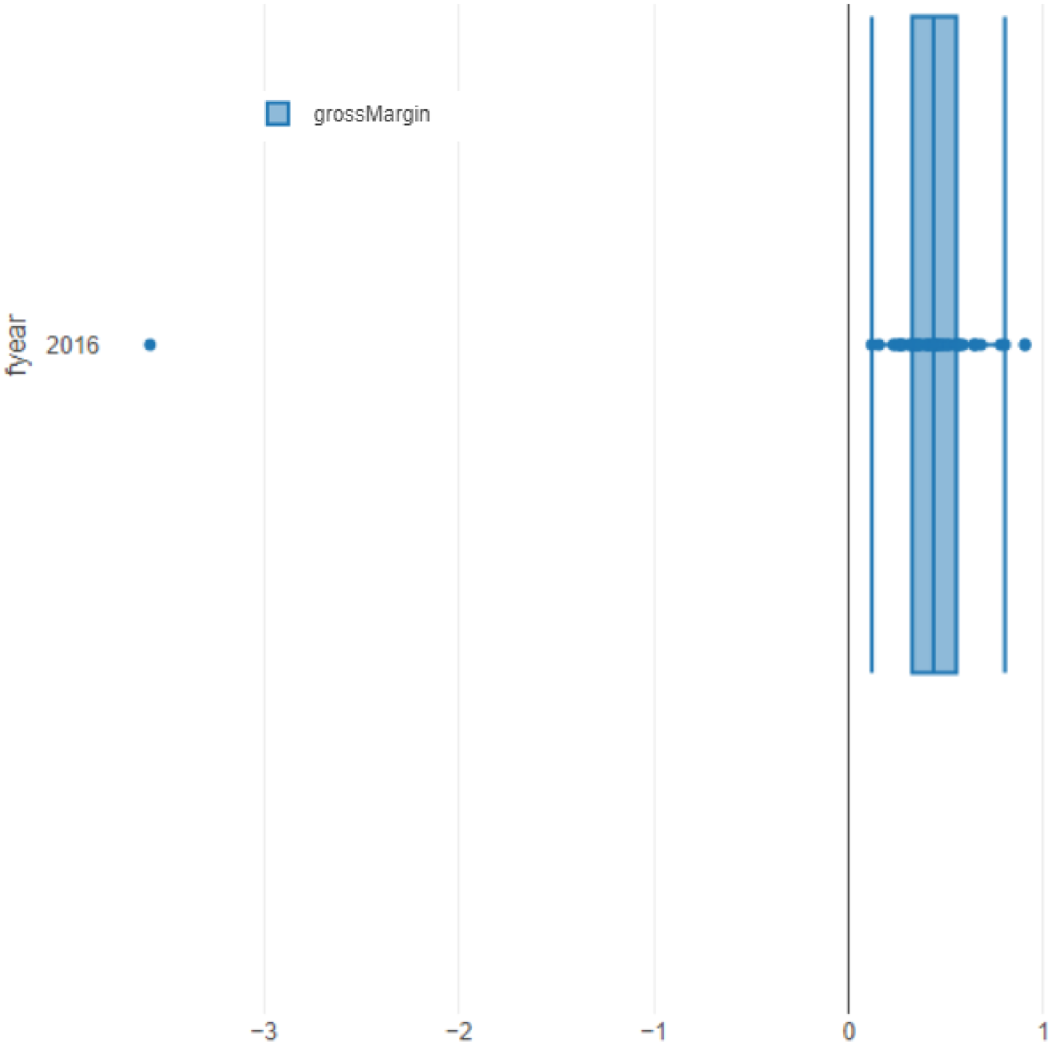


Figure 2.3: Graphical Comparison of Gross Margin

2.5 Communicate Key Findings (BLUF)

Most decision-makers have a good grasp of their field and need analysts to provide information and recommendations without unnecessary detail. Most of the problems in this textbook that require a memo summarizing findings limit that memo to a single page. These memos should be written with the conclusion first, or **bottom line up front** (BLUF). BLUF is a communication strategy originating from the military that prioritizes presenting crucial information first. This approach respects the reader's time and aids in quick decision-making, making it valuable in various professional settings beyond the military. By delivering key messages promptly, BLUF helps cut through information overload and convey essential points effectively.

Implementing BLUF requires careful organization of thoughts to ensure clear communication throughout the entire memo. This method reduces the recipient's effort in understanding and acting on the information, making it particularly useful in fast-paced environments. The BLUF technique is also valuable in content marketing, helping marketers capture audience attention quickly in a crowded information landscape. Despite its military origins, BLUF can enhance various types of writing by improving the overall effectiveness and efficiency of communication.

We would like to communicate two key points to the financial analyst in response to the question: *Are large technology firms more profitable than small firms?*

1. The overall level of gross margin relative to sales is quite similar between large and small firms.
2. However, large firms show less dispersion in their relative gross margin than small firms. Generally, we can interpret this to mean that the large firms' gross margins are close to one another, while the gross margin values for the small firms are spread across a broader range of values.

2.6 Take Away Points

Successful financial professionals use their knowledge to probe interesting audit, tax, managerial, and financial accounting questions, leveraging data analytics. They need to work with data scientists (computer scientists and statisticians) to address some of their more sophisticated questions.

Converting an accounting problem into a data analytics problem that a data scientist understands is critical for the correct execution of the analysis. In general, the trademarks of successful tech-savvy accounting professionals are their ability to translate business problems into data analytics problems and their ability to translate data analysis findings into accounting insights and implications.

2.7 Preview for Next Chapter

In the next chapter we will employ our new data analytics skills on a consulting project with the Accounting & Finance Analytics Consulting Group for the Chapman Investment Fund to delve into the impact of cost leadership and product differentiation strategies on investment decisions within the technology sector. The Chapman Investment Fund seeks to leverage the Accounting & Finance Analytics Consulting Group advanced analytics to classify potential investment opportunities based on their core business strategies—whether they focus on minimizing costs or differentiating their products.

The primary learning objective for this module remains consistent: to underscore the critical role of financial accounting data in analyzing industry trends and evaluating company performance. You will develop skills to identify significant outliers and understand their impact on your analyses. This chapter is designed to broaden your understanding and equip you with the knowledge to make well-informed decisions, supported by comprehensive financial data. Prepare for a journey into critical thinking, data-driven insights, and the strategic pursuit of investment acumen.

End-of-Chapter Practice Problems

The End-of-Chapter Practice Problems aim to apply some concepts and tools introduced in the chapter. The datasets for the End-of-Chapter Practice Problems are available on the textbook website

End-of-Chapter Practice Problem 2.1

1. Replicate the analysis in section 2.3 (the workflow build ending in 2.3.8) for 2020. Make a copy of your workflow with the 2016 analysis and modify it for 2020.
2. Communicate the key findings of your analysis and compare them to the findings from 2016.

End-of-Chapter Practice Problem 2.2

1. Replicate the analysis in section 2.3 for another industry. For example, perform the analysis for all firms in the same industry as Tesla for 2020. Make a copy of your workflow with the 2020 Apple analysis and modify it for Tesla. Run a web search for the NAICS of Tesla. You will find several websites that will tell you that the NAICS of Tesla is 336111.
2. Communicate the key findings of your analysis.