Simplyx

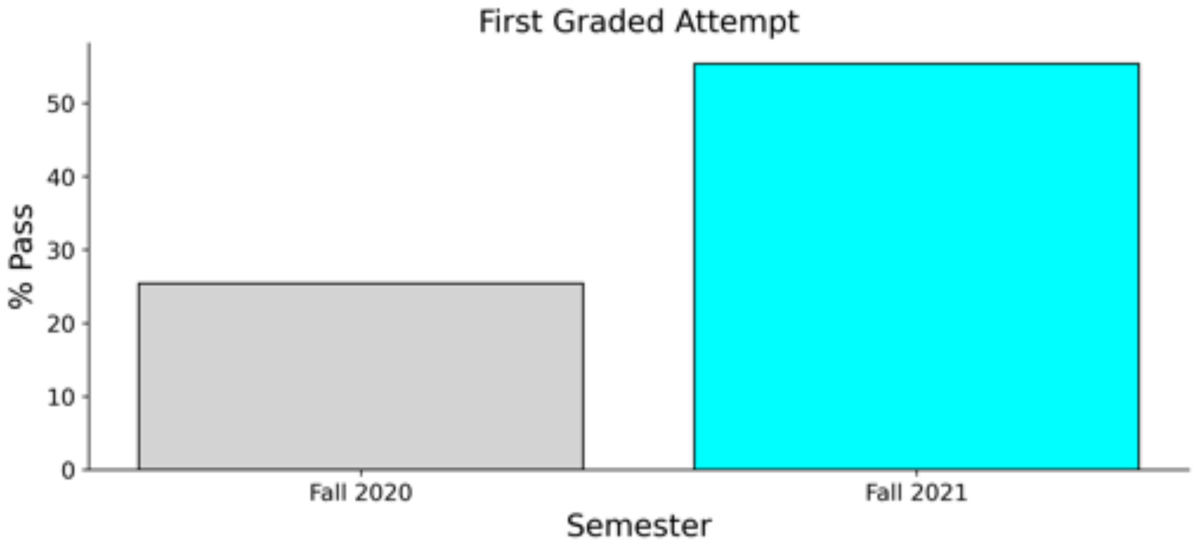
Data Analysis with Alteryx

Peter Kinder
Kai Larsen

Preface

Whether using a cell phone, charging a credit card, driving a car or adjusting a thermostat, the average person is progressing through daily life while leaving a trail of data via an ever-growing variety of activities. According to Statista Research Department, the aggregate of these data trails is projected to reach 181 zettabytes by 2025 [1]. Put into context, the total data storage of a typical computer is around single terabyte, or one billionth of a single zettabyte. With 7.8 Billion people on planet earth, that is about 23 terrabytes per person. In response to this vast trove of available data, the business community has recognized the potential value and is adjusting accordingly. As of 2022, the U.S. Bureau of Labor Statistics projects that the number of jobs related to data analysis will grow by 25% over the decade, which is considered to be "much faster than average" [2]. To help fuel this growth, a number of organizations have developed software to assist data analysts. This booklet will focus on one of those software packages, Alteryx, and seeks to prepare you for the Alteryx Designer Core Certification. In addition to this booklet, Alteryx also provides a prep guide, practice test, weekly challenges and documentation that will be referenced throughout the booklet. It is highly recommended that these resources are utilized in tandem with this booklet.

At this point, it is appropriate to discuss why this booklet and its associated courseware were created given that Alteryx provides preparatory resources. The main motivation was in giving soon-to-be data analysts a greater hands-on approach. The authors of this booklet have taught and assisted in teaching an undergraduate business analytics course for decades and felt that there was room for improvement. Historically, only about 25% of our business students were able to obtain the certification. In response, this standalone courseware named **Simplyx** was developed and delivered promising results; 55.4% of students obtained the certification on their first graded attempt. Compared to the 25.4% of students who obtained the certification on their first graded attempt during the past fall semester of 2020, this was a marked improvement.

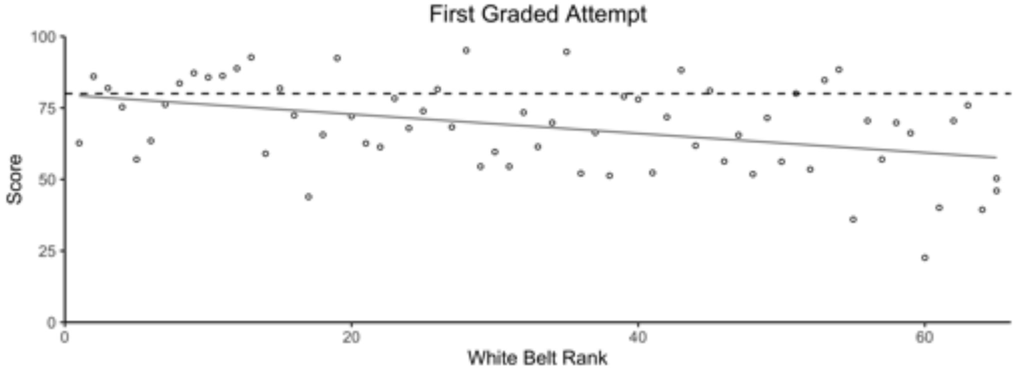


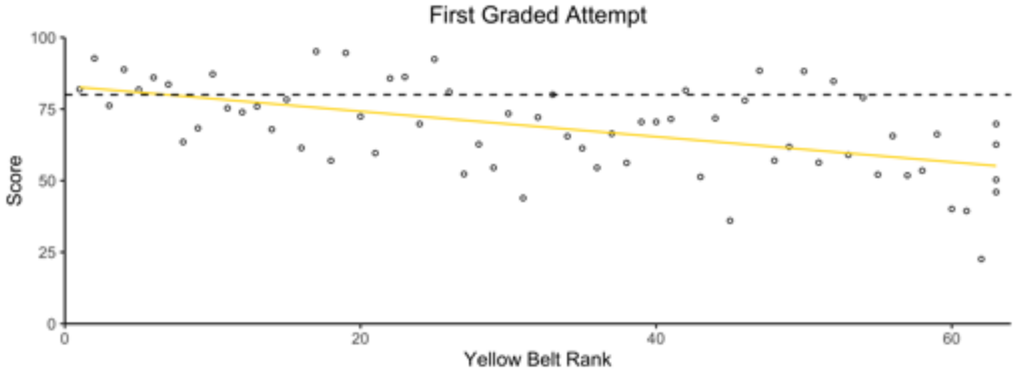
In the following semester, due to the impressive increase in the number of students who obtained the certification, Simplyx was refined to collect usage data with the goal of improving Simplyx. However, due to a different instructor teaching the course, and only a limited portion of Simplyx completed (about 75%) before the first graded attempt, only 26.8% of students obtained the certification. Yet, there were some valuable takeaways from the usage data that will hopefully motivate you:

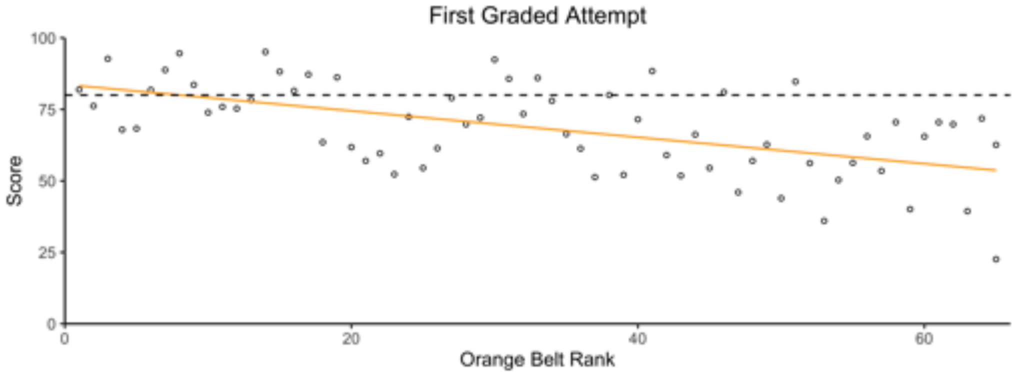
1. A single student completed all of the content prior to their first graded attempt, and that student obtained the certification with a score of 92.7%; third best out of a class of 67 students with the highest score being 95.1%.

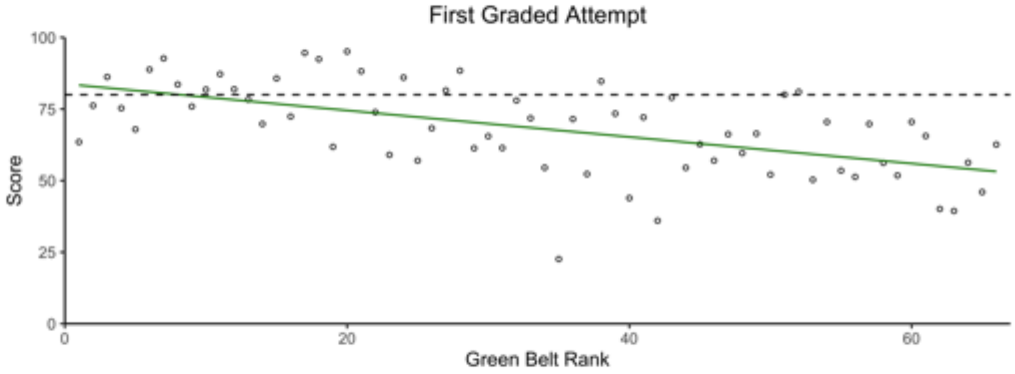
2. The content is organized into "belts" with the purple belt being the second to last belt. By the time of the first graded attempt, only 27 out of 67 students had attempted at least one exercise from the purple belt. From the group of students who attempted at least one exercise from the purple belt, 51.9% of students obtained the certification. From the group of students who had not attempted at least one exercise from the purple belt, only 10% of students obtained the certification.

3. Ranking each student for each belt, first by completeness and second by earliest time of first attempt, it appears that the ranking is moderately negatively correlated (approximately -0.50) with a certification score. That is, the less complete and the later the start of each belt, the lower the certification score. The takeaway is that you should start early and aim to complete every exercise. For a visual representation of this relationship, the following graphs with a dashed line at the passing certification score of 80 were generated:









4. Each belt contains exercises that are straightforward and focused on specific tools as well as exercises that combine tools in a more complex data analysis exercise. Upon investigating the exercises that the fewest students completed, all of them were the more complex data analysis exercises. When comparing the group of students who completed these exercises to the group of students that did not for each belt, the percentage of students who passed the certification varied significantly.

White Belt

	# Students	% Pass
Not Completed	13	7.7
Completed	54	31.5

Yellow Belt

	# Students	% Pass
Not Completed	38	13.2
Completed	29	44.8

Orange Belt

	# Students	% Pass
Not Completed	33	12.1
Completed	34	41.2

Green Belt

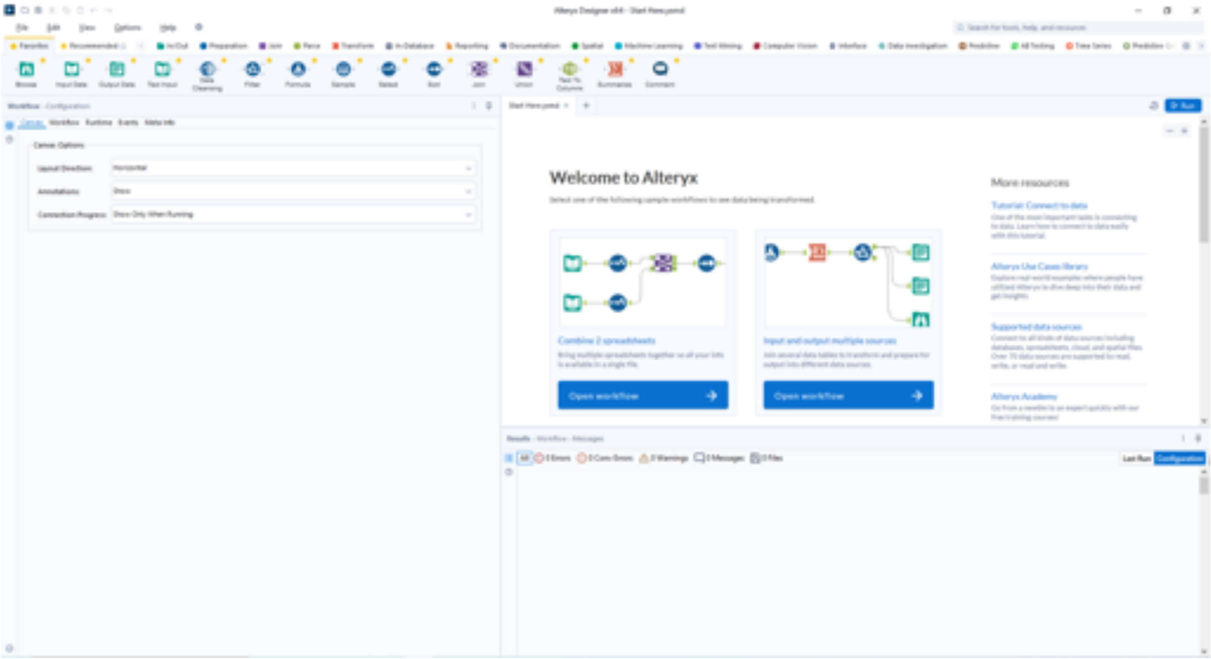
	# Students	% Pass
Not Completed	41	9.8
Completed	26	53.8

Overall, it appears that the more hands-on resource of the courseware improved the likelihood of students obtaining the certification. Furthermore, it looks like students that started early and finished all of the exercises for each belt performed better than students that did not. Hopefully, this booklet will improve students outcomes' even more.

Table of Contents

1 Introduction

In the highly digitized world we live in today, many of us take the utility of computers for granted. Looking back a century from now, the advent of an early IBM mechanical punched card tabulator during the 1920s helped establish statistics and data analysis as a discipline in the United States [3]. More recently, in 1993, a new programming language called R was developed that would go on to become a cornerstone of data analysis that is still widely used to this day. While R is an effective way to analyze data, it requires a familiarity with the programming language that can be a barrier to users. This barrier was effectively lowered by Alteryx, which presents a graphical user interface (GUI) largely built on R allowing users to analyze large amounts of data on low code platform. A user is presented with the following screen (click here for interactive walk through):

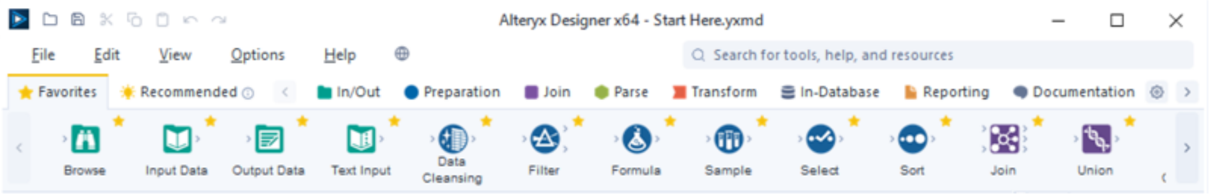


What the reading is currently looking at is referred to as a **Workflow**. For a first time user, this can be a bit overwhelming, so let's break it down into pieces. Starting at the top of the GUI, a user will find a typical Windows menu bar with file, edit, view, options, help, and language tabs. The entirety of the available choices is beyond the scope of this booklet, but it is worthwhile to point out a couple highlights:

- Under the file tab, a user can save their workflow. It is recommended that a user save frequently, as *Alteryx does not automatically save a user's workflow.* Additionally, a user can quickly save their workflow by pressing **Ctrl-S**. *While reading along, take note of the mentioned keyboard shortcuts in preparation for a future certification attempt.*

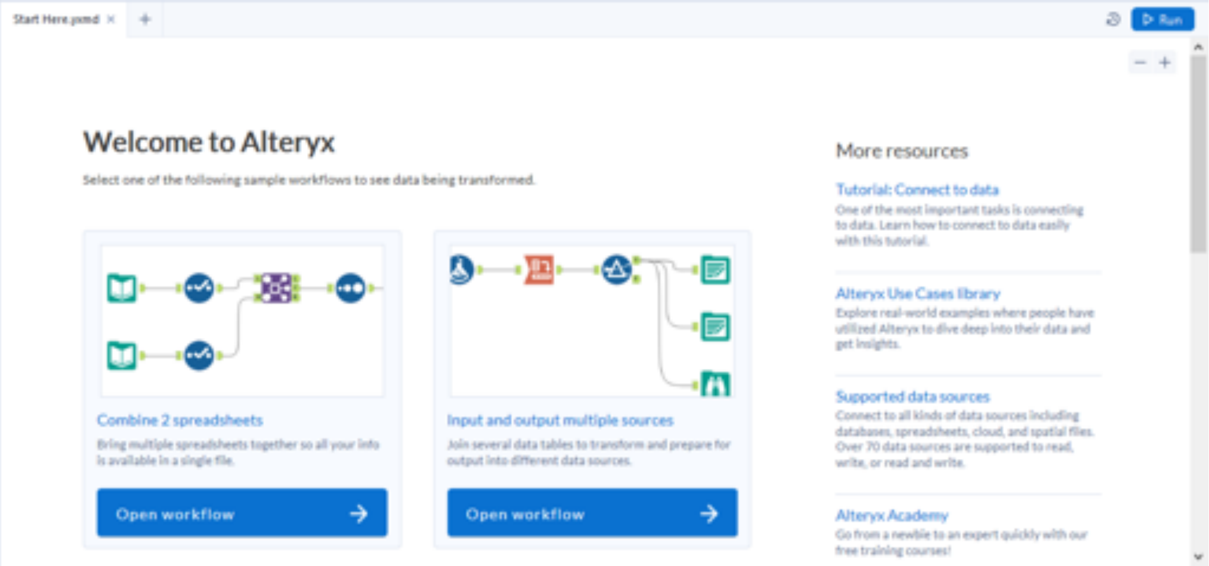
- Under the options tab, a user can export their workflow. Essentially, this is the same as saving a workflow, except that the data that is being used in the workflow can be packaged along with the workflow itself. This can be extremely useful when sharing a workflow or if a user has multiple devices.

In addition to the Windows menu, a user will find what is known as a **Tool Pallete** containing the various tools that Alteryx offers. Below is a closer look at the Tool Pallete:



As can be seen, the individual tools are represented by unique icons in the area with a light blue background. Just above that area is a menu of tool categories that allow the user to navigate to the different tools quickly. Alteryx contains over 270 tools, so this arrangement can be quite handy. But, the tools covered in this booklet and on the certification all located in the visible categories in the above image: In/Out, Preparation, Join, Parse, Transform and Documentation. Lastly, *right-clicking on most tools opens a submenu for help with that particular tool.* Clicking Help in that submenu opens an internet browser window to the documentation of that specific tool. Clicking Open Example opens a new workflow tab with some examples of that tool's usage.

The next area of interest is referred to as the **Canvas**. The Canvas is the area where a user places tools to perform the data analysis. Below is a closer look at the Canvas:



Welcome to Alteryx

Select one of the following sample workflows to see data being transformed.

Combine 2 spreadsheets

Bring multiple spreadsheets together so all your info is available in a single file.

Open workflow →

Input and output multiple sources

Join several data tables to transform and prepare for output into different data sources.

Open workflow →

More resources

Tutorial: Connect to data

One of the most important tasks is connecting to data. Learn how to connect to data easily with this tutorial.

Alteryx Use Cases library

Explore real-world examples where people have utilized Alteryx to dive deep into their data and get insights.

Supported data sources

Connect to all kinds of data sources including databases, spreadsheets, cloud, and spatial files. Over 70 data sources are supported to read, write, or read and write.

Alteryx Academy

Go from a newbie to an expert quickly with our free training courses!

In the graphic above, the Alteryx default "Start Here.yxmd" workflow is loaded, but the Canvas is mainly a blank white space where tools can be added. There are primarily four distinct ways that tools can be added to the Canvas:

1. Drag and drop from the Tool Pallete.
2. Right-click and Insert from dropdown.
3. Right-click and Paste (or **Ctrl-V**) a copied tool.
4. Right-click on an existing tool and Insert After.

There are a couple other unique ways to add specific tools to the Canvas, but those will be addressed when covering those tools. Another key aspect of the Canvas is the blue Run button in the upper right-hand side of the Canvas. This executes the workflow, and can also be invoked by pressing **Ctrl-R**. There is also a vertical scroll bar and a zoom in/out button on the Canvas. A user can build as large a workflow as necessary, and these features allow a user to navigate large workflows. If a workflow is large enough, a horizontal scroll bar will also appear at the bottom of the Canvas. The last element present on every Canvas is the + tab button near the top left-hand side. This allow a user to open up a new blank Canvas, which constitutes an empty workflow. A user can have multiple Workflows open at the same time, and it should be noted that *when a user performs an action such as run, save or export, that action is performed on the visible Workflow.*

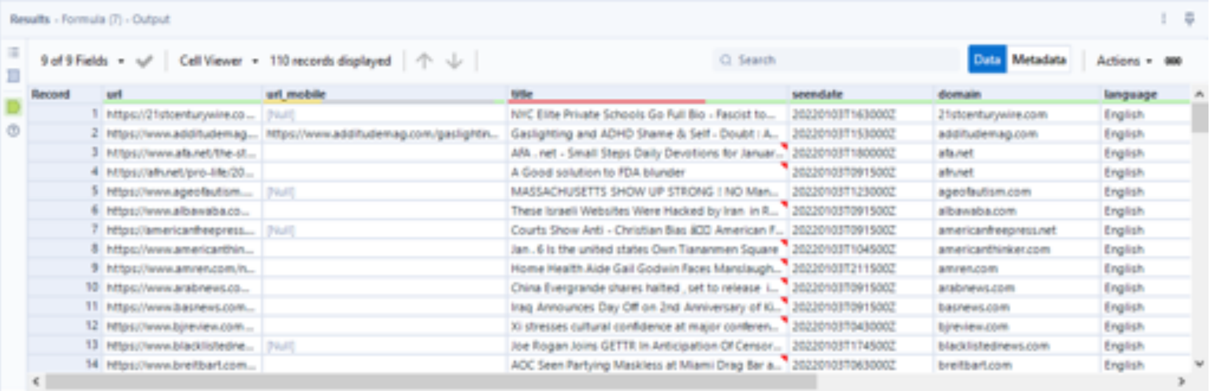
While the Canvas is the blank white space, the default Workflow loaded in the most recent graphic allows for some further detail. Seen in the workflow is a collection of tools connected by **Connectors**; the lines connecting each individual tool. When a user runs the Workflow, the actions that each tool performs on the data is executed in order from the leftmost side of a continuous connection of tools. Think of the data as a stream flowing along a path, being modified at each tool. As a user is building a workflow, the connections between tools can be created two ways:

1. Dragging and dropping a tool in close proximity to an existing tool.

2. Clicking on an output anchor and dragging a connector to an input anchor.

The second method of connecting tools brings up the question of an anchor, which is a small green node seen on the side of tools. Each tool has either an **Input Anchor** (on the left side of a tool) or an **Output Anchor** (on the right side of a tool) or both. Some tools have multiple input anchors and output anchors. Additionally, some anchors are unlabeled and some are labeled; what each of these labels means will be addressed when covering those tools. Lastly, by clicking on the anchors, a user can view the data at that particular point in the Workflow process. For example, clicking on an input anchor will allow a user to view the data before the tool has modified the data, and clicking on output anchor will reveal the data after the modification.

Directly below the Canvas is the Results Window, where a user can inspect the actual data. Below is a closer look with some sample data highlighting important concepts:



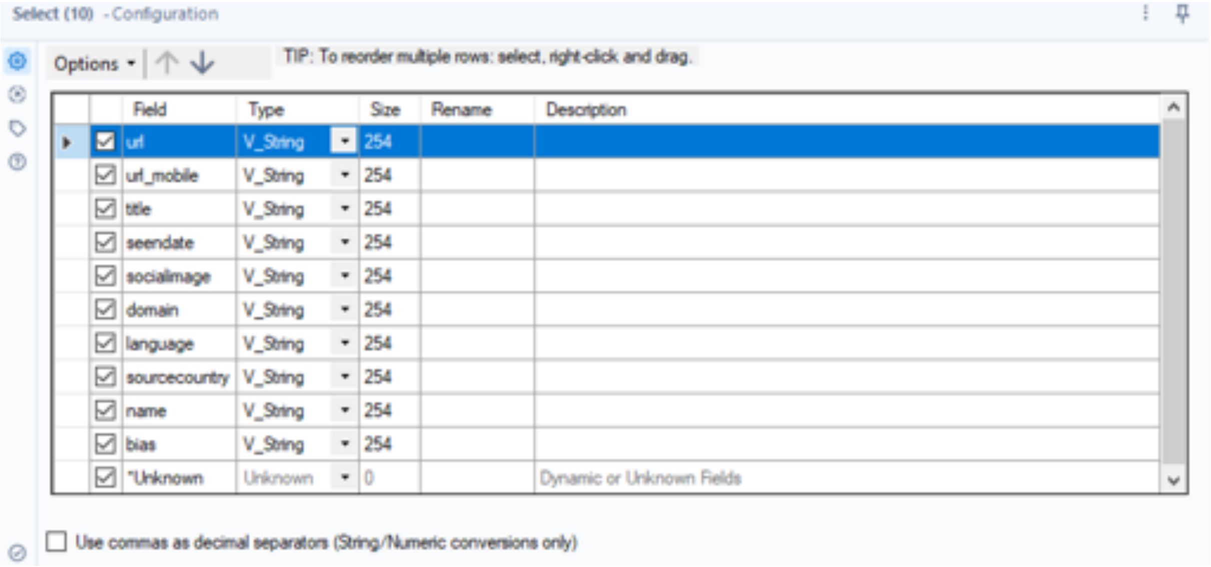
Record	url	url_mobile	title	seendate	domain	language
1	https://21stcenturywire.co...	[Null]	NYC Elite Private Schools Go Full Bio - Fascist to...	20220103T163000Z	21stcenturywire.com	English
2	https://www.additudemag...	https://www.additudemag.com/gaslightin...	Gaslighting and ADHD Shame & Self - Doubt : A...	20220103T153000Z	additudemag.com	English
3	https://www.afa.net/the-st...		AFA . net - Small Steps Daily Devotions for Januar...	20220103T180000Z	afa.net	English
4	https://afn.net/pro-life/20...		A Good solution to FDA blunder	20220103T091500Z	afn.net	English
5	https://www.ageofautism...	[Null]	MASSACHUSETTS SHOW UP STRONG ! NO Man...	20220103T123000Z	ageofautism.com	English
6	https://www.albawaba.co...		These Israeli Websites Were Hacked by Iran in R...	20220103T091500Z	albawaba.com	English
7	https://americanfreepress...	[Null]	Courts Show Anti - Christian Bias [illegible] American F...	20220103T091500Z	americanfreepress.net	English
8	https://www.americanthin...		Jan . 6 Is the united states Own Tiananmen Square	20220103T104500Z	americanthinker.com	English
9	https://www.amren.com/n...		Home Health Aide Gail Godwin Faces Manslaugh...	20220103T211500Z	amren.com	English
10	https://www.arabnews.co...		China Evergrande shares halted , set to release i...	20220103T091500Z	arabnews.com	English
11	https://www.basnews.com...		Iraq Announces Day Off on 2nd Anniversary of Ki...	20220103T091500Z	basnews.com	English
12	https://www.bjreview.com...		Xi stresses cultural confidence at major conferen...	20220103T043000Z	bjreview.com	English
13	https://www.blacklistedne...	[Null]	Joe Rogan Joins GETTR In Anticipation Of Censor...	20220103T174500Z	blacklistednews.com	English
14	https://www.breitbart.com...		AOC Seen Partying Maskless at Miami Drag Bar a...	20220103T063000Z	breitbart.com	English

The most obvious aspect of the Results Window is the actual data. The data is organized by **Record** (a row of data) and **Field** (a column of data). Generally speaking, a record is either a unit of analysis or a unit of observation and the field is a characteristic of that unit. Going further in-depth is beyond the scope of this booklet, but in the graphic above, it is reasonable to infer that a record represents a news article. It is also apparent that there is a color coding scheme related to the data. According to the official Alteryx documentation, the each color represents the following:

- Green (Ok): Indicates no identified quality issues.
- Yellow (Null): Indicates values that are null, excluding empty values.
- Red (Not Ok): Indicates leading/trailing white spaces, or embedded new lines.
- Gray (Empty): Indicates values that contain at least one string with no values.

This color scheme is important to understand in order to quickly assess the quality of the data and to perform potential alterations. Greater detail related to the percentage of the field that has these characteristics can be viewed by hovering over the field title. While hovering over a field title, a user will notice a vertical ellipse appear on the right side of the title. Clicking on this ellipse allows the user to perform actions such a Data Cleanse, Filter, and Sort. This can be helpful when inspecting the data, but *performing these actions does not carry through for the rest of the workflow.* A user can also right-click on the data or click the Actions dropdown in the upper right-hand side to copy the data. That data can then be pasted elsewhere or, by right-clicking the canvas, pasted as a Text Input tool into the actual workflow. Lastly, a user can click the Metadata button to see the data types of each field. Data types will be covering an upcoming section, but it is worth noting that *one of the main reasons why a tool does not perform as a user intends relates to an incorrect data type.*

The final portion of the Alteryx GUI that warrants discussion is the **Configuration Window**. The Configuration Window is specific to the tool that the user has selected, and controls how the tool will modify the data. Below is one example:


Select (10) - Configuration

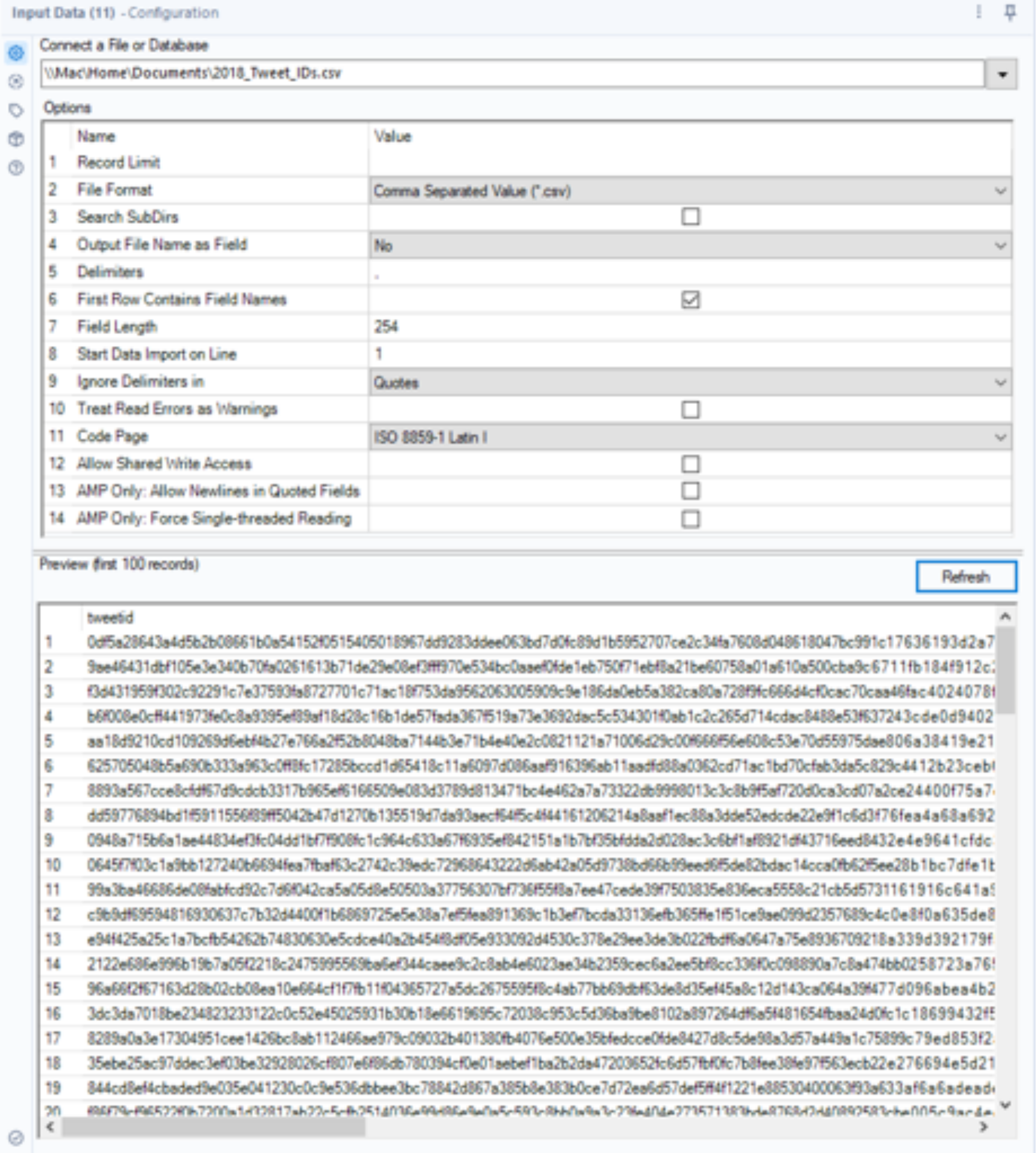
Options

TIP: To reorder multiple rows: select, right-click and drag.

	Field	Type	Size	Rename	Description
☑	url	V_String	254		
☑	url_mobile	V_String	254		
☑	title	V_String	254		
☑	seendate	V_String	254		
☑	socialimage	V_String	254		
☑	domain	V_String	254		
☑	language	V_String	254		
☑	sourcecountry	V_String	254		
☑	name	V_String	254		
☑	bias	V_String	254		
☑	*Unknown	Unknown	0		Dynamic or Unknown Fields

☐ Use commas as decimal separators (String/Numeric conversions only)

For comparison's sake, below is another example:



\\Mac\Home\Documents\2018_Tweet_IDs.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	,
6	First Row Contains Field Names	☑
7	Field Length	254
8	Start Data Import on Line	1
9	Ignore Delimiters in	Quotes
10	Treat Read Errors as Warnings	□
11	Code Page	ISO 8859-1 Latin I
12	Allow Shared Write Access	□
13	AMP Only: Allow Newlines in Quoted Fields	□
14	AMP Only: Force Single-threaded Reading	□

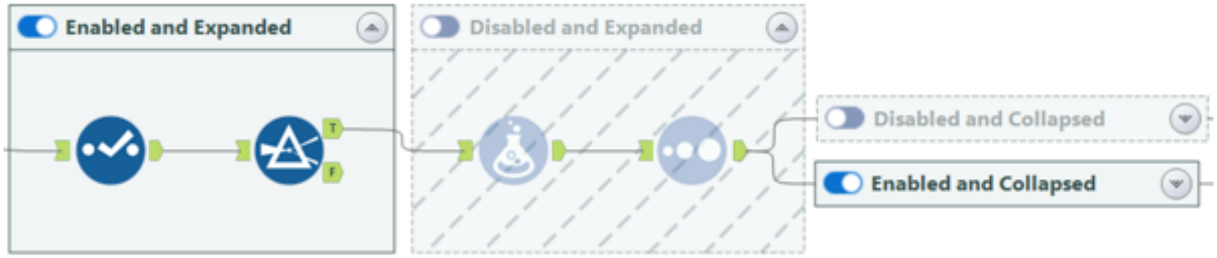
	tweetid
1	0df5a28643a4d5b2b08661b0a54152f0515405018967dd9283ddee063bd7d0fc89d1b5952707ce2c34fa7608d048618047bc991c17636193d2a7
2	9ae46431dbf105e3e340b70fa0261613b71de29e08ef3ff970e534bc0aaef0fde1eb750f71ebf8a21be60758a01a610a500cba9c6711fb184f912c
3	f3d431959f302c92291c7e37593fa8727701c71ac18f753da9562063005909c9e186da0eb5a382ca80a728f9fc666d4cf0cac70caa46fac4024078
4	b6f008e0cff441973fe0c8a9395ef89af18d28c16b1de57fada367f519a73e3692dac5c534301f0ab1c2c265d714cdac8488e53f637243cde0d9402
5	aa18d9210cd109269d6ebf4b27e766a2f52b8048ba7144b3e71b4e40e2c0821121a71006d29c00f666f56e608c53e70d55975dae806a38419e21
6	625705048b5a690b333a963c0ff8fc17285bccd1d65418c11a6097d086aaf916396ab11aadfd88a0362cd71ac1bd70cfab3da5c829c4412b23ceb
7	8893a567cce8cfdf67d9cdcb3317b965ef6166509e083d3789d813471bc4e462a7a73322db9998013c3c8b9f5af720d0ca3cd07a2ce24400f75a7
8	dd59776894bd1f5911556f89ff5042b47d1270b135519d7da93aecf64f5c4f44161206214a8aaf1ec88a3dde52edcde22e9f1c6d3f76fea4a68a692
9	0948a715b6a1ae44834ef3fc04dd1bf7f908fc1c964c633a67f6935ef842151a1b7bf35bfdda2d028ac3c6bf1af8921df43716eed8432e4e9641cfdc
10	0645f7f03c1a9bb127240b6694fea7fbaf63c2742c39edc72968643222d6ab42a05d9738bd66b99eed6f5de82bdac14cca0fb62f5ee28b1bc7dfe1b
11	99a3ba46686de08fabfcd92c7d6f042ca5a05d8e50503a37756307bf736f55f8a7ee47cede39f7503835e836eca5558c21cb5d5731161916c641a
12	c9b9df69594816930637c7b32d4400f1b6869725e5e38a7ef5fea891369c1b3ef7bcda33136efb365ffe1f51ce9ae099d2357689c4c0e8f0a635de8
13	e94f425a25c1a7bcfb54262b74830630e5cdce40a2b454f8df05e933092d4530c378e29ee3de3b022fbdf6a0647a75e8936709218a339d392179f
14	2122e686e996b19b7a05f2218c2475995569ba6ef344caee9c2c8ab4e6023ae34b2359cec6a2ee5bf8cc336f0c098890a7c8a474bb0258723a76
15	96a66f2f67163d28b02cb08ea10e664cf1f7fb11f04365727a5dc2675595f8c4ab77bb69dbf63de8d35ef45a8c12d143ca064a39f477d096abea4b2
16	3dc3da7018be234823233122c0c52e45025931b30b18e6619695c72038c953c5d36ba9be8102a897264df6a5f481654fbaa24d0fc1c18699432f5
17	8289a0a3e17304951cee1426bc8ab112466ae979c09032b401380fb4076e500e35bfedcce0fde8427d8c5de98a3d57a449a1c75899c79ed853f2
18	35ebe25ac97ddec3ef03be32928026cf807e6f86db780394cf0e01aebef1ba2b2da47203652fc6d57fbf0fc7b8fee38fe97f563ecb22e276694e5d21
19	844cd8ef4cbaded9e035e041230c0c9e536dbbee3bc78842d867a385b8e383b0ce7d72ea6d57def5ff4f1221e88530400063f93a633af6a6adead
20	f86f79cf96522f0b7200a1d32817ab22c5cfb2514036e99d86e9e0a5c593c8bb0a9a3c23fed04e273571383bde8768d2d40892583cbe005c9ac4e

Since each Configuration Window is specific to the selected tool, the Configuration Window will be addressed when covering those tools. However, in the first graphic in this Introduction section, a view of what the Configuration Window looks like when no tools are selected is visible. The options available in that view are beyond the scope of this booklet, expect that it is recommended that new users leave the Canvas Options as is.

With a general overview of the Alteryx GUI complete, you is ready to start using Simplyx. Visit the Simplyx Content page to download the Introduction Workflow and complete the associated exercises while finishing the Introduction section.

1.1 Tool Container

One of many advantages of using Alteryx for data analysis is the ability to view a complex sequence of data modifications in an organized visual manner. The Tool Container is a tool that is used in combination with other tools to enhance this benefit of Alteryx, and also provides additional utility to a user. Below are examples of typical Tool Containers in their different possible states:



Starting on the left, the Tool Container named Enabled and Expanded is just that. When a Tool Container enabled and expanded, the primary purpose is organizing the Workflow. When a user runs the Workflow, the tools located in the Tool Container perform their actions on the data, and the data continues to run through the workflow. This is the default state of a Tool Container when it is added to a Workflow.

Moving to the right, the Tool Container named Disabled and Expanded is an example of a Tool container that has been disabled but still displays the contents of the Tool Container. When a tool container is disabled, the tools located in the Tool Container *do not* perform their actions on the data, and the data *does not* continue to run through the workflow. This can be beneficial to a user when working with a large dataset and constructing a complex workflow because it can reduce the computations required and reduce runtime.

On the topic of runtime, and separate from the Tool Container, a user can also right-click on most tools and select Cache and Run Workflow. This runs the workflow, and caches the data at the tool that was right-clicked. The user can then run the workflow again without having to preform the computations leading up to the cache, which can save an immense amount of time when working with large datasets.

Back on the topic of the Tool Container, the two Tool Containers on the far right are examples of a disabled and collapsed Tool Container as well as an enabled and collapsed Tool Container. To collapse or expand the Tool Container, a user can click the icon in the upper right-hand side and can be helpful when organizing complex Workflows. *It is important to understand that whether a Tool Container is collapsed or expanded has no bearing on whether the Tool Container is enabled or disabled.* Failing to realize this can result in a significant loss of time for the user.

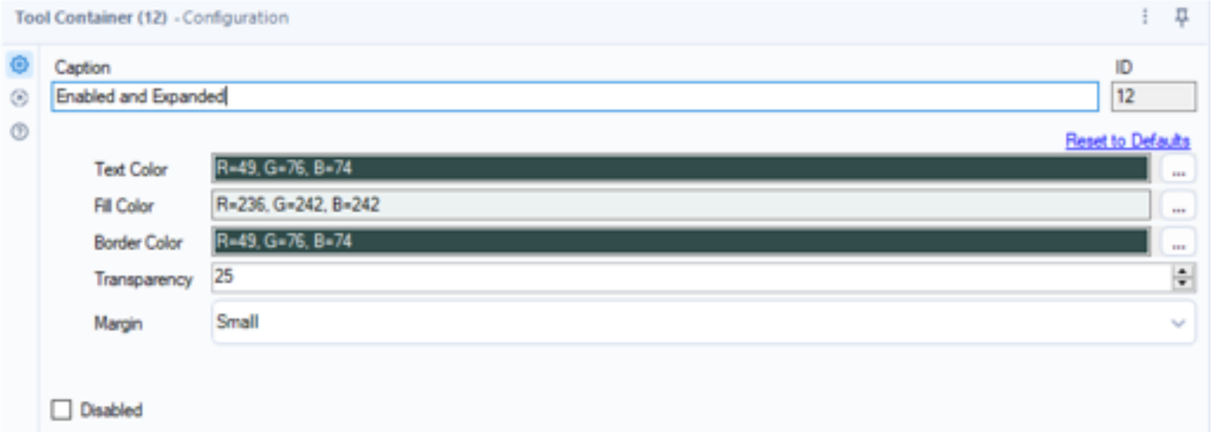
To add tools to a Tool Container, there are two possible actions a user can take:

1. Select existing tools on the Canvas, right-click and select Add to New Container.

2. Add a Tool Container to the Canvas, and drag and drop tools into it.

When deleting a Tool Container, a user can right-click on the Tool Container. At this point, a user can delete the Tool Container *and* the tools within it by selecting Delete. However, if a user only wants to remove the Tool Container, **Delete Container Only** should be selected. If a user accidentally performs an unwanted action, the user can press **Ctrl-Z** to quickly undo that action.

Lastly, and probably most importantly, a user can enable or disable a Tool Container by clicking the slider in the upper left-hand side. When the slider is to the **left**, the Tool Container is **disabled**. When the slider is to the **right**, the Tool Container is **enabled**. The Tool Container can also be disabled or enabled by selecting or deselecting a checkbox located in the Configuration Window, and doing so will be reflected in the Tool Container's slider. For reference, below is the Configuration Window:



Tool Container (12) - Configuration

Caption: Enabled and Expanded

ID: 12

Reset to Defaults

Text Color: R=49, G=76, B=74

Fill Color: R=236, G=242, B=242

Border Color: R=49, G=76, B=74

Transparency: 25

Margin: Small

☐ Disabled

1.2 Data Types

As mentioned earlier, *one of the main reasons why a tool does not perform as a user intends relates to an incorrect data type.* Alteryx supports a variety of data types including text, numeric, time, blobs and spatial objects. For the purposes of this book, you does not need to familiarize themself with blobs or spatial objects, but it is critical to understand the other data type categories. Alteryx provides a comprehensive overview of Data Types that you **must understand in detail**. When reviewing each data type category, carefully consider what makes each sub data type different from the others.

However, for the purposes of this booklet and while working with Alteryx for the first time, there are a couple simple rules of thumb the user can follow:

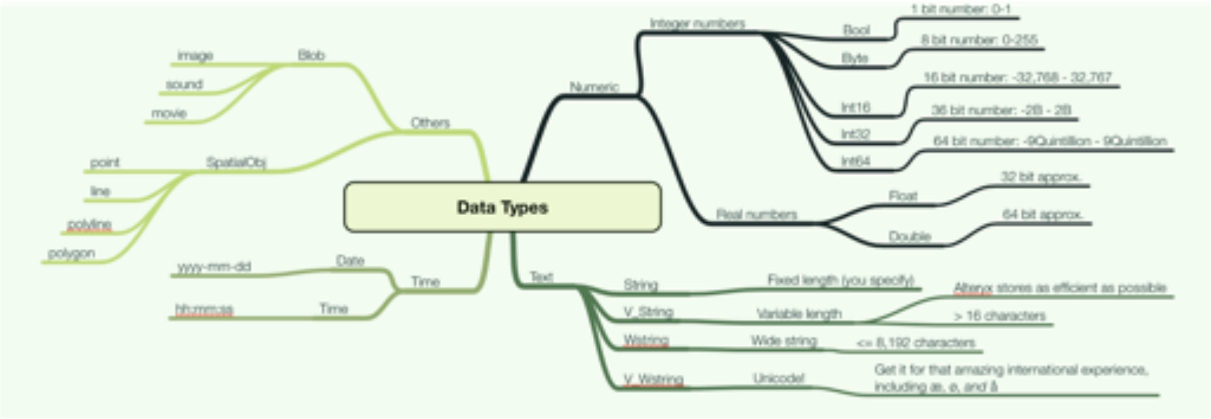
1. When working with a field of text data, use the **V_WString** data type. This data type is the most flexible type of string.

2. When working with a field of whole numbers, use the **Int64** data type. This data type if the most flexible type of integer.

3. When working with a field of numbers with decimals, use the **Double** data type. This data type if the most flexible numeric data type.

4. When unsure whether to use a Int64 or Double data type, use the Double data type.

For quick reference, the visualization below can be very helpful:



1.3 File Types

When performing data analysis, the most common file type that will be encountered is a **CSV**. A CSV stands for comma separated values and uses a "," as a delimiter of a single record. One very important consideration concerning Alteryx and CSV is that *all of the data will be read in as text*, regardless the data type is appears to be. However, there is a great deal of other file types that Alteryx supports too, including its own file type, the **yxdb**. The yxdb file type is the most efficient file type to use when working with Alteryx. Another great benefit of the yxdb file type is that when it is read into Alteryx, the *data types it was output with are preserved*. All in all, for the purposes of this booklet, you should be familiar with the following file types and their extensions:

File Type	Extension
Alteryx Analytic App	.yxwz
Alteryx Database	.yxdb
Alteryx Field Types	.yxft
Alteryx Packaged Workflow	.yxzp
Alteryx Workflow	.yxmd
Comma Separated Values	.csv
Microsoft Excel 1997-2003	.xls
Microsoft Excel	.xlsx
Microsoft Excel Macro-Enabled	.xlsm
Text	.txt
Zip	.zip

At this point, you have hopefully already download the Introduction Workflow from Simplyx and encountered a Zip file. If not, download the Introduction Workflow now. To unzip the Zip file, right-click on the file and select Extract All if using a Windows computer. If viewing the Zip from a Mac, double-click the file to unzip it. Within the unzipped folder, there will be a yxmd file and other folders. Double click on the yxmd file to open it in Alteryx. Alternatively, open Alteryx and drag and drop the yxmd onto the Canvas, or select File in the Windows Menu and then Open Workflow.

2 White Belt

Like a white belt in karate, fundamentals are the focus of the Simplyx white belt. Before being able to analyze data, it first must be read in. After reading in the data, it must be converted to the correct data types and cleaned. Then, a quick inspection of some of the descriptive statistics can save a user a considerable amount of time. Lastly, the often-overlooked process of writing the data to a new file bears great importance, and is critical to understand when working with other data analysts.

At this point, it is recommended that you visit the Simplyx Content page to download the White Belt Workflow and complete the exercises while reading along.

2.1 Input

Garbage in, garbage out is a common refrain when analyzing data. That is, if the input is flawed, the output will be flawed. When working with Alteryx, this is important to remember because a small error at the beginning of the data analysis process can snowball into a large problem. Thus, understanding how to use the **Text Input** and **Input Data** tools are important considerations when analyzing data with Alteryx.

The Text Input tool allows a user to quickly add a limited amount of data to a workflow. This functionality is best suited for rapidly testing how tools will modify the data in addition to some lookup tables associated with tools discussed later in this booklet. It is also beneficial when sharing Workflows due to that fact that the data in a Text Input tool saves to the actual workflow, as opposed to data in an Input Data tool.

Interactive Lesson

Text Input Tool Interactive Lesson

Now that you have completed the Text Input Tool Interactive Lesson, you should be ready to start experimenting with the tool. Some finer points related to the Text Input tool are are as follows:

- The Text Input tool is limited to **10,000 total cells on import**, but a user can manually add more cells if needed.
- A user can right-click the Text Input tool to **Convert to Macro Input** tool.

Lastly, while discussed in the Text Input Interactive Lesson, the Text Input tool will automatically configure the data to the **smallest data type available**. As mentioned before, one of the main reasons why a tool does not perform as a user intends relates to an incorrect data type.

Exercise Checkpoint

Complete Exercise 1 & 2

You should now be accustomed to the Text Input tool. Alteryx also provides Text Input Tool Mastery and Text Input Tool Documentation for further detail. *To be ready for the Alteryx certification, please read both and make sure it all makes sense to you.* Take notes on any questions you may have.

While Text Input tools have a large variety of uses that a user will come to know when using Alteryx, the Input Data tool will probably be the primary way a user imports data to a workflow when analyzing data. When a user wants to include the data in a Data Input tool, a user can navigate to the Windows Menu at the top of Alteryx, select Options and choose Export Workflow.

Interactive Lesson

Input Data Tool Interactive Lesson 1

As was the case in Input Data Tool Interactive Lesson 1, the three main file types that will be covered in this booklet are CSV, Microsoft Excel, and Alteryx Database file types. It is worth reiterating the concept of truncated data, especially regarding the CSV file type. A user should always review the Messages in the Results Window to ensure that none of the data is being truncated. Seeing this message when analyzing data can have extremely negative effects because the truncated data being read in is garbage, resulting in a garbage output. Additionally, a user needs to recognize that the Preview functionality only displays the first **100 records**.

Another important concept not reviewed in detail in Input Data Interactive Lesson 1 is importing multiple files with a single Input Data tool. Additionally, a user can import multiple files with a single Input Data tool. For example, say there were several of yxdb files in the Document directory (folder) and the user wanted read them all in. The user could use a *** wildcard** in the file path, which would end up looking as such:

\\Mac\Home\Documents\*.yxdb

Or, say that a user wanted all of the CSV files in the Document directory:

\\Mac\Home\Documents\*.csv

In these cases, all of the yxdb or csv files in the Documents directory would be read in. But, say that there were files that we named like SampleFileXX, where X were some number or character. Then, a user could use the following to read in only those yxdb files:

\\Mac\Home\Documents\SampleFile*.yxdb

The * wildcard can be very helpful when working with multiple datasets, but can be too flexible in some cases. Say that a user only wanted the files that were named like SampleFileX but not SampleFileXX. In this case, a user would want to use a ? wildcard and the following path:

\\Mac\Home\Documents\SampleFile?.yxdb

Interactive Lesson

Input Data Tool Interactive Lesson 2

Exercise Checkpoint

Complete Exercise 3 - 7

You should now be accustomed to the Input Data tool. Alteryx also provides the Input Data Tool Mastery and Input Data Tool Documentation for further detail. *To be ready for the Alteryx certification, please read both and make sure it all makes sense to you.* Take notes on any questions you may have.

2.2 Output

Working with a number of students in the past, mastery of the **Output Data** tool has often been overlooked resulting in a loss of what should be relatively easy points on the certification. While this booklet will focus on the Data Output tool with respect to CSV, Microsoft Excel, and Alteryx Database file types, you should recognize that the tool can write data to a file **or database**.

Interactive Lesson

Output Data Tool Interactive Lesson

When writing a file to a CSV, there is no particular limit as to how many records can be written, but when working with a Microsoft Excel file, only **1,048,576 records can be written** [4]. This limit also applies to using Excel, which is one of the reasons why Alteryx is superior to Excel when working in the age of big data. This limit also applies to using Excel, which is one of the reasons why Alteryx is superior to Excel when working in the age of big data.

Regarding the Code Page option, this option is essentially how the data is encoded, and can result in a loss of data. ISO 8859-1 Latin I is generally acceptable when working with data in English, but Unicode (UTF-8 or UTF-16) is broader code page and can be selected if there are any doubts.

Lastly, when building a Workflow, a user will often write data at certain points along the way. A user can then branch off the Workflow in an efficient manner by converting the Output Data tool to an Input Data tool. To do so, a user can:

1. Right-click the Output Data tool in the Workflow.
2. Select **Convert To Input Data**.
3. Configure the resulting Input Data tool.

While the Simplyx exercises do not include use of the Output Data tool due to the technical inability to assess the use of the tool, you should make a point to understand the Output Data tool. In addition to the documentation linked to in the earlier part of this section, a couple other very helpful resources include the Output Data Tool Mastery and the Output Data Tool Documentation. *To be ready for the Alteryx certification, please read both and make sure it all makes sense to you.* Take notes on any questions you may have.

2.3 Browse

Two significant advantages of Alteryx are efficiency and the ability to work with big data. In select cases, the **Browse** tool can further enhance these two advantages. However, as with many tools, having incorrect data types is very often associated with a tool not performing as a user intends.

Without a Browse tool, the Results Window will only display up to **1 MB of data**, which can give users the false impression that they are able to inspect all the data.

Interactive Lesson

Browse Tool Interactive Lesson

As noted in the Browse Tool Interactive Lesson, since the Browse tool loads all the data in the Results Window, the runtime of the Workflow can be negatively impacted resulting in an unnecessary loss of time. So, a user should use a Browse tool when appropriate, but avoid overloading the Workflow with Browse tools. Additionally, it is worth highlighting that a user can right-click on the Browse tool and:

- **Convert to Macro Output** tool.
- **Convert to Output Data** tool.

Exercise Checkpoint

Complete Exercise 8 - 12

Revisit **Exercise 12** in Simplyx and inspect the Browse tool without changing the data type to an appropriate numerical type. Notice how at first glance it looks like the Browse tool is providing relevant descriptive statistics. Upon a more detailed inspection, a user will notice that the Browse tool is misleading a user when an incorrect data type is used. Alteryx also provides Browse Tool Mastery and Browse Tool Documentation for further detail. *To be ready for the Alteryx certification, please read both and make sure it all makes sense to you.* Take notes on any questions you may have.

2.4 Select

By now you should know that having incorrect data types is very often associated with a tool not performing as a user intends. Thankfully, the **Select** tool allows a user to modify a field's data type, among its other functionalities. More generally speaking, the Select tool allows a user to modify the metadata of the data, including whether a field is or is not included and the name, data type and size of the data.

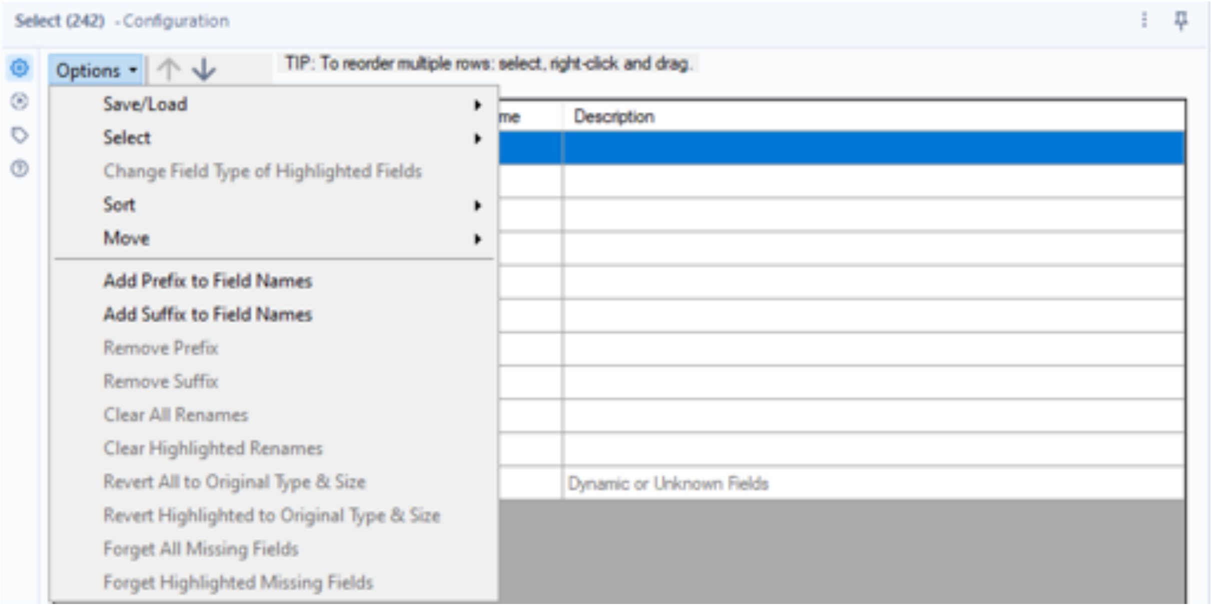
Interactive Lesson

Data Type Interactive Lesson

Interactive Lesson

Select Tool Interactive Lesson

Additionally, a user can click the Options dropdown in the upper left-hand side of the Select tool Configuration Window:



Select (242) - Configuration

TIP: To reorder multiple rows: select, right-click and drag.

Options
Save/Load
Select
Change Field Type of Highlighted Fields
Sort
Move
Add Prefix to Field Names
Add Suffix to Field Names
Remove Prefix
Remove Suffix
Clear All Renames
Clear Highlighted Renames
Revert All to Original Type & Size
Revert Highlighted to Original Type & Size
Forget All Missing Fields
Forget Highlighted Missing Fields
me
Description
Dynamic or Unknown Fields

While some of the options are straightforward, a couple options warrant some discussion. The Select option will allow a user to check or uncheck all the Field checkboxes, which can save a user a great deal of time when the data has many Fields. Down at the bottom of the options list, a user can **Forget All Missing Fields**. A **Missing Field** is a field that the Select tool is configured for but is not flowing through the Workflow into the tool. This will usually be colored yellow and tends to occur when using a previously made Workflow on another dataset.

Exercise Checkpoint

Complete Exercise 13 - 17

You should now be accustomed to the Select tool. Alteryx also provides Select Tool Mastery and Select Tool Documentation for further detail. *To be ready for the Alteryx certification, please read both and make sure it all makes sense to you.* Take notes on any questions you may have.

2.5 DateTime

For an individual who has not worked with a DateTime data type before, the **DateTime** tool can have a bit of a steep learning curve.

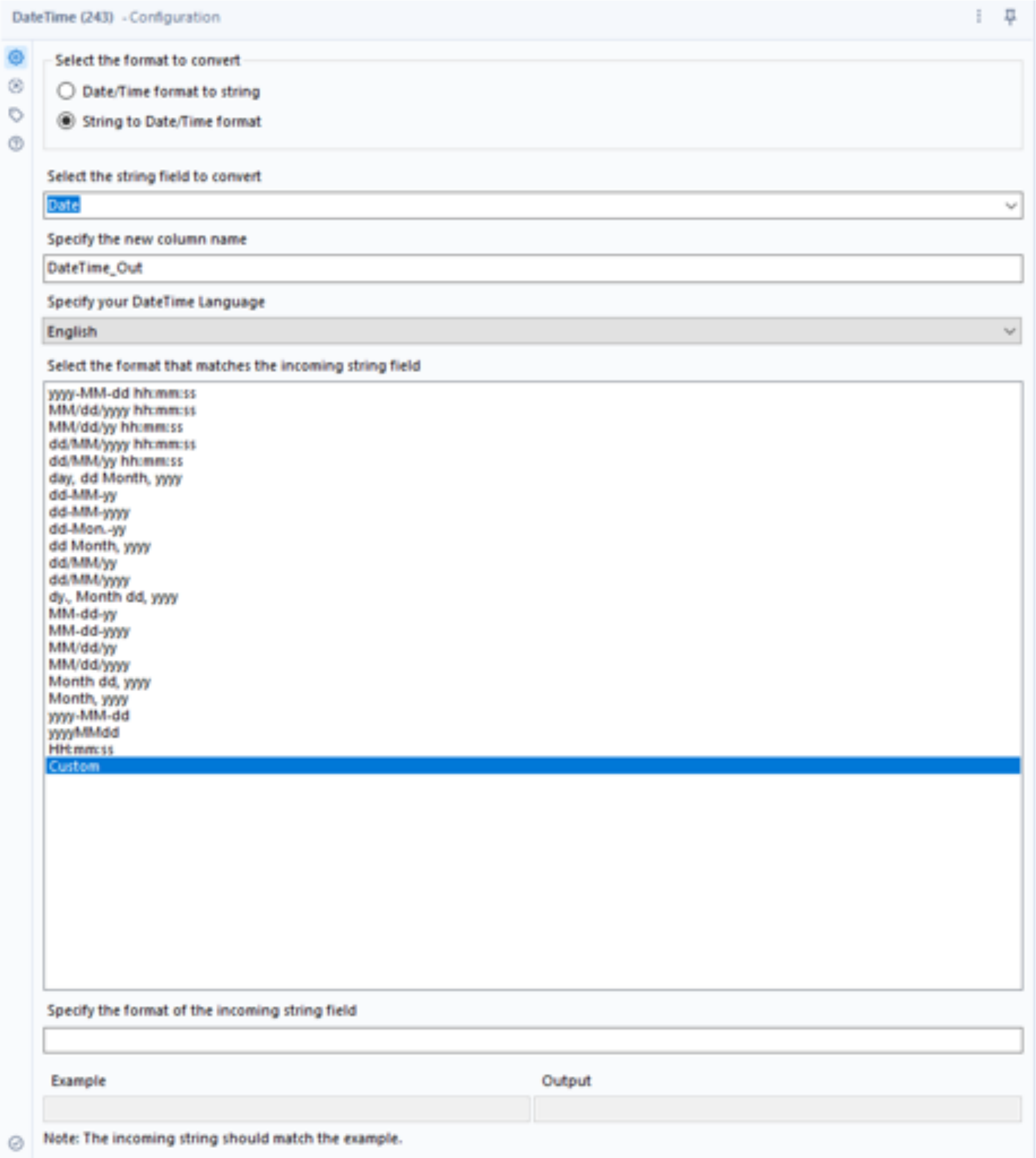
Video Lesson

DateTime Tool Interactive Lesson

A user reading in a csv file may be dismayed to find that there are many ways to format a date and time, and that Alteryx will treat their date, time, or datetime as a string. It may be tempting to use the Select tool to convert a text data type to a Date, Time, or DateTime data type. But, unless the data is in the ISO format of **yyyy-mm-dd** for a Date, **HH:MM:SS** for a Time, or **yyyy-mm-dd HH:MM:SS** for a DateTime, a user will encounter an error and end up with null values.

When the data is not in ISO format, a user can set the DateTime tool to convert the String to Date/Time format. Why would a user convert to DateTime data type? Having a field in a DateTime format allows a user to do math with it, such as add or subtract time or calculate the difference between two dates. Furthermore, a user can make the conversion the other way from Date/Time format to string, allowing a user to format the data into more readable/desirable formats.

Consider **Exercise 20**. The exercise instructs a user to convert a field named Date to the generic format of “Sunday, November 12, 2017.” To do so, a user must convert the field to a DateTime data type and then back into a text data type. So, a user must add a DateTime tool to the Canvas and connect it to the output anchor of the Text Input tool. Then, a typical DateTime tool Configuration Window should appear:



DateTime (243) - Configuration

Select the format to convert

○ Date/Time format to string

● String to Date/Time format

Select the string field to convert

Date

Specify the new column name

DateTime_Out

Specify your DateTime Language

English

Select the format that matches the incoming string field

```
yyyy-MM-dd hh:mm:ss
MM/dd/yyyy hh:mm:ss
MM/dd/yy hh:mm:ss
dd/MM/yyyy hh:mm:ss
dd/MM/yy hh:mm:ss
day, dd Month, yyyy
dd-MM-yy
dd-MM-yyyy
dd-Mon.-yy
dd Month, yyyy
dd/MM/yy
dd/MM/yyyy
dy., Month dd, yyyy
MM-dd-yy
MM-dd-yyyy
MM/dd/yy
MM/dd/yyyy
Month dd, yyyy
Month, yyyy
yyyy-MM-dd
yyyyMMdd
HH:mm:ss
Custom
```

Specify the format of the incoming string field

Example

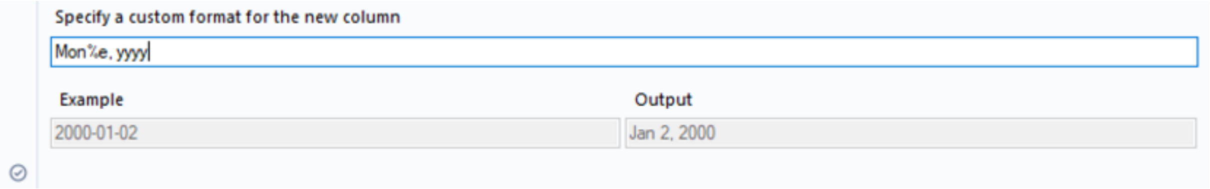
Output

Note: The incoming string should match the example.

The first consideration for a user is to ensure that the **String to Date/Time format** radio button option is selected. A user also must make sure that the correct field in the **Select the string to convert** dropdown is selected. Now, the incoming format of the text data type version of the data must be matched in the **Select the format that matches the incoming string field** list. At more comprehensive list of this notation can be found in the documentation linked to at the end of this section, but for now a user can just select MM/dd/yyyy.

Now, a user should run the Workflow and observe in the Results Window that a new field has been created with a Date data type. A user can then chain (add another)

DateTime tool to convert the Date data type back into the desired format with a text data type. For this, a user will have to select **Custom** in the DateTime tool. It is left to you to consult the documentation regarding which notation should be used, but a general example of doing so can be seen below:



Specify a custom format for the new column: Mon%e, yyyy

Example: 2000-01-02

Output: Jan 2, 2000

Exercise Checkpoint

Complete Exercise 18 - 20

You should now be accustomed to the Select tool. Alteryx also provides DateTime Tool Mastery, DateTime Tool Documentation, and DateTime Function Documentation for further detail. *To be ready for the Alteryx certification, please read both and make sure it all makes sense to you.* Take notes on any questions you may have.

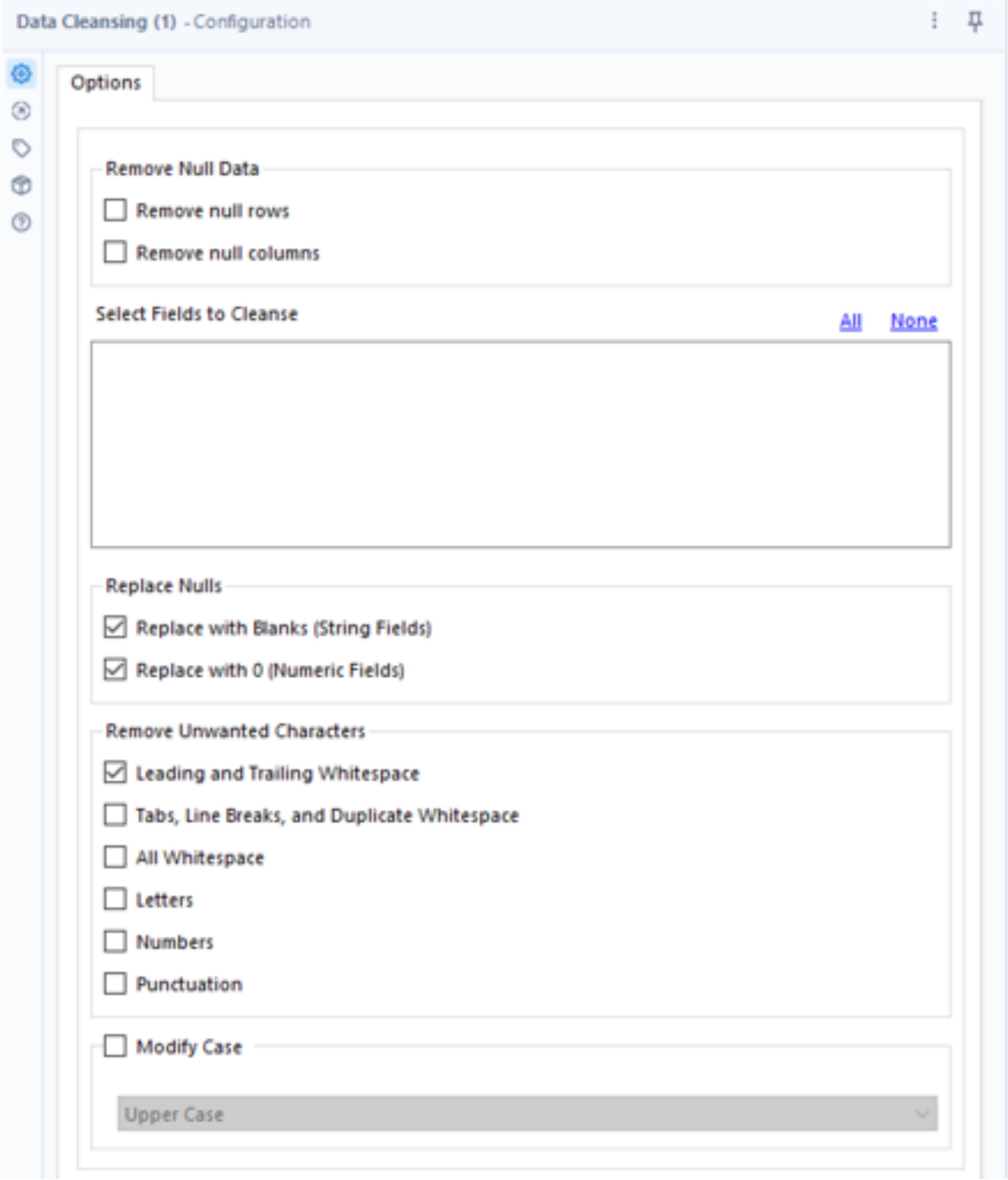
2.6 Data Cleansing

Youmay look a one Word and another word and conclude that it is the same WORD. A computer on the other hand will consider word, Word and WORD to be three distinct words. This small example highlights one of the important reasons why a user must clean their data, and a quick way to do so is with the **Data Cleanse** tool. A couple other distinctions that you should recognize are the following:

1. Tabs, Line Breaks, and Duplicate Whitespace can affect parsing.
2. Null and 0 values are treated differently in numerical calculations.
3. Nulls and Blanks (Empty Strings) are different.
    - An Empty String is a string with length 0.
    - A Null is a missing value.

Luckily, the Data Cleanse tool is one of the most intuitive tools Alteryx has to offer. However, its drawback is that it does slow down the runtime of a workflow, so like the Browse tool, it is not advised that a user have too many Data Cleansing tools. It should also be noted that the Data Cleansing tool applies all the selected actions on all of the selected fields. If a user wants to perform different actions on different field, multiple Data Cleaning tools are required.

After adding a Data Cleaning tool, a user will encounter the following Configuration Window:



Data Cleansing (1) - Configuration

Options

Remove Null Data

☐ Remove null rows

☐ Remove null columns

Select Fields to Cleanse

All None

Replace Nulls

☑ Replace with Blanks (String Fields)

☑ Replace with 0 (Numeric Fields)

Remove Unwanted Characters

☑ Leading and Trailing Whitespace

☐ Tabs, Line Breaks, and Duplicate Whitespace

☐ All Whitespace

☐ Letters

☐ Numbers

☐ Punctuation

☐ Modify Case

Upper Case

When connected to an Output Anchor of another tool, the **Select Field to Cleanse** area of the Configuration Window will be populated with fields that a user can select. Then, a user can choose to **Replace Nulls** and/or **Remove Unwanted Characters**. The options of those actions are relatively straightforward, but the user is encouraged to experiment in Alteryx.

If a user wants to Remove Null Rows or Remove Null Columns, it must be realized that the Data Cleaning tool will **only** apply this action to rows/columns that only contains null values. A single non-null value will result in the row/column remaining in the data.

Lastly, a user can **Modify Case**, and has the following three options:

1. **UPPER CASE**
2. **lower case**
3. **Title Case**

Exercise Checkpoint

Complete Exercise 20 - 25

You should now be accustomed to the Select tool. Alteryx also provides Data Cleansing Tool Mastery and Data Cleansing Tool Documentation for further detail. *To be ready for the Alteryx certification, please read both and make sure it all makes sense to you.* Take notes on any questions you may have.

3 Yellow Belt

Congratulations! You have earned your Simplyx white belt and developed the fundamentals for working with Alteryx! Now it is time to step it up a level and earn your yellow belt. By earning your yellow belt, you will learn how to perform a number of actions on that data that will be used in nearly every data analysis and will prepare the data for more complex analysis and manipulation. With this newfound knowledge, you will be able to flip the data as you please with the Sort tool, and chop it into slices with the the Filter, Select Records, Sample and Unique tools.

At this point, it is recommended that you visit the Simplyx Content page to download the Yellow Belt Workflow and complete the exercises while reading along.

3.1 Sort

Consider the values 9, 100, and 20. Sort them from largest to smallest. You hopefully ordered them 100, 20, 9. But what if we had told you that these values did not have a numeric data type and instead had a text data type? Using the **Sort** tool to sort 9, 100, and 20 from largest to smallest with a text data type would produce 9, 20, 100. Why is this? It is being applied based on the ASCII value of the first character of the text. In short, sorting Abe, Abi, and Ben from largest to smallest will produce Ben, Abi, Abe because the first character is evaluated first, followed by the second, etc. For Abi and Abe, only on the third character does the sort order become apparent. When sorting a text data type column containing 9, 100, and 20, the whole sort order is determined by the first character (9, 1, and 2).

Interactive Lesson

Sort Tool Interactive Lesson

Thankfully, sorting mistakes became less common after Alteryx recently updated the Sort tool to select **Use Dictionary Order** by default. But a user should always carefully consider the data type and how the sort will be applied to avoid mistakes on what normally would be easy problems. A couple tips when it comes to sorting include:

1. Use a Browse tool to inspect **all** the data after a sort.
2. Convert all text data type data to **upper/lower case**.
3. Beware of **punctuation** and remove it if it's not relevant.

Lastly, it can be a bit confusing at first what ascending and descending mean for different data types, so review the following table:

Data Type	Ascending	Descending
Text	A -> Z	Z -> A
Numeric	Small -> Large	Large -> Small
DateTime	Most Recent -> Least Recent	Least Recent -> Most Recent

Exercise Checkpoint

Complete Exercise 1 - 7

You should now be accustomed to the Select tool. Alteryx also provides the Sort Tool Mastery and Sort Tool Documentation for further detail. *To be ready for the Alteryx certification, please read both and make sure it all makes sense to you.* Take notes on any questions you may have.

3.2 Filter

Consider the words different, Different, and DIFFERENT. If asked to apply a filter for the words that equal DIFFERENT, how many words you think would satisfy the filter condition? While a computer will consider the three words to be different, the Alteryx Filter tool will consider them to be the same word. This is an important consideration when using that filter tool that does not generally apply when working with other Alteryx tools and data in general.

Interactive Lesson

Filter Tool Interactive Lesson

In the Filter Tool Interactive Lesson, the concept of Boolean logic with the AND/OR operators was briefly mentioned. While beyond the scope of this booklet, it is essential to take the Filter Tool Interactive lesson a step further. Consider the following sentence:

"Alteryx is the best data analytics software"

Now, consider the two following conditions:

(contains "Alteryx" **AND** contains "tech") **OR** (contains "best" **AND** contains "data")

contains "Alteryx" **AND** contains "tech" **OR** contains "best" **AND** contains "data"

Are these two conditions equivalent? The only difference between the two statements is the presence of paratheses. However, parentheses are significant to evaluating a Boolean expression (statement). In the first condition, Alteryx first evaluates the expressions within the parentheses, resulting in:

(false) **OR** (true) = false **OR** true = true

In the second condition, the expression reduces to:

false **AND** true **OR** true **AND** true = false

The expression evaluates to false in the second condition due to the false AND true portion of the Boolean expression. Boolean logic (algebra) can be a bit confusing to a first-time user, but to oversimplify it, evaluate a Boolean expression until all the subexpressions in the parentheses have been evaluated. If the resulting expression with no parentheses includes a false as well as the AND operator, then the entire expression is false. Otherwise, it is true. We recommend actively using parentheses even if you are great at Boolean logic because the person you share your Workflow with may not be.

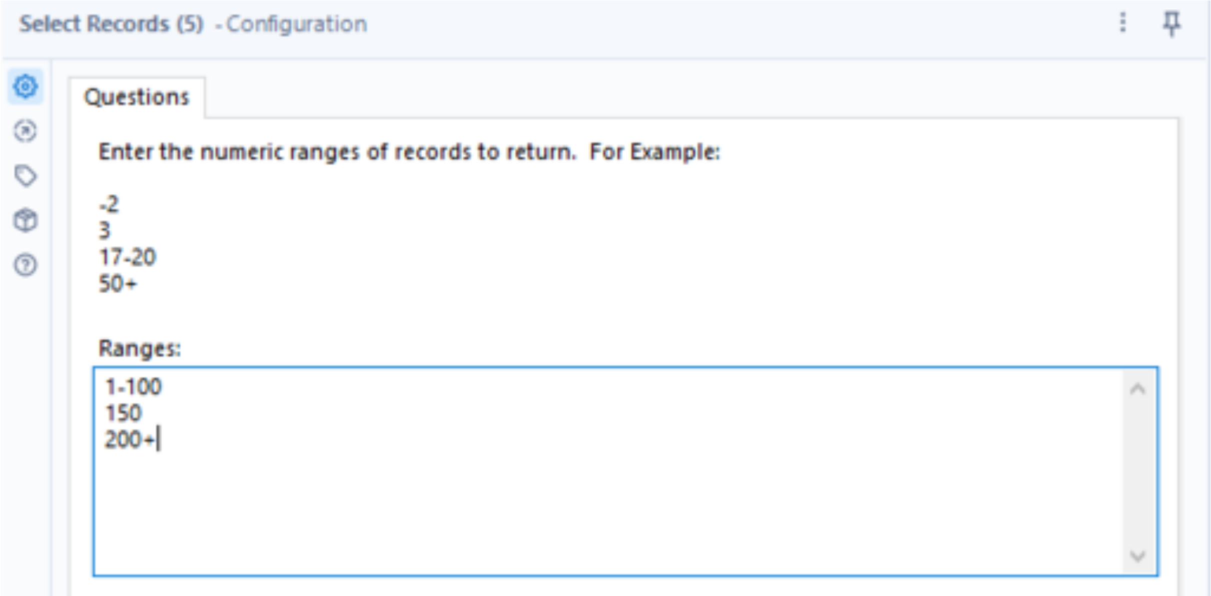
Exercise Checkpoint

Complete Exercise 8 - 12

You should now be accustomed to the Filter tool. Alteryx also provides the Filter Tool Mastery and Filter Tool Documentation for further detail. *To be ready for the Alteryx certification, please read both and make sure it all makes sense to you.* Take notes on any questions you may have.

3.3 Select Records

The **Select Records** tool is a tool that allows a user to isolate a subset of the data based on the record index. For clarity, the record index is the value that can be observed on the left-hand side of a Results Window when inspecting data. The Select Records tool is one of the most straightforward tools included in Alteryx, and upon adding the tool to the Canvas a user will see a Configuration Window similar to the following:



Select Records (5) - Configuration

Questions

Enter the numeric ranges of records to return. For Example:

```
-2
3
17-20
50+
```

Ranges:

```
1-100
150
200+
```

A user can enter any combination of ranges in the **Ranges** area of the Configuration Window. In the graphic above, a user would have the records with an index between 1-100 and 150, and any index greater than 200 continue to flow through the Workflow.

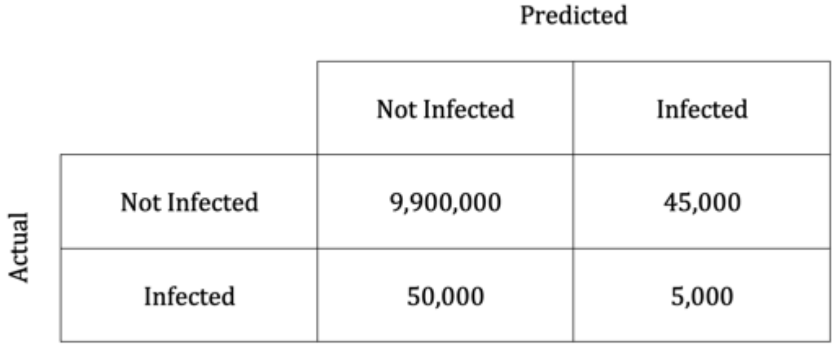
Exercise Checkpoint

Complete Exercise 13 - 18

You should now be accustomed to the Select Records tool. Alteryx also provides the Select Records Tool Documentation for further detail. *To be ready for the Alteryx certification, please read the documentation and make sure it all makes sense to you.* Take notes on any questions you may have.

3.4 Sample

As the COVID-19 pandemic has shown, knowing whether one is infected with a virus is an important ability for a society. But imagine that a model had been created to predict whether an individual was infected, and that model's performance could be measured by the following confusion matrix:



		Predicted	
		Not Infected	Infected
Actual	Not Infected	9,900,000	45,000
	Infected	50,000	5,000

The model would be 99.05% accurate, which might seem like a pretty good model. However, the model only predicted 5,000 people had COVID-19 out of 55,000 people who actually had COVID-19, resulting in a 9.10% sensitivity. Furthermore, out of the 50,000 people who were predicted to have COVID-19, only 5,000 of them actually had COVID-19, resulting in a 10% precision. Knowing this, do you still think the model is a good model?

This conflicting information is what is known as the accuracy paradox and is generally attributable to an imbalance between the positive class (infected) and negative class (not infected). Thankfully, the Sample tool can help a user address this accuracy paradox and other issues during their data analysis.

Interactive Lesson

Sample Tool Interactive Lesson

One very important consideration mentioned in the Sample Tool Interactive Lesson was that the **1 in N chance to include each row** does not guarantee a specific number of rows will be included. Additionally, the **Group by column** option for the Sample tool is **case sensitive**, so if grouping by the words different, Different, and DIFFERENT, there would be 3 distinct groups.

Exercise Checkpoint

Complete Exercise 19 - 23

You should now be accustomed to the Sample tool. Alteryx also provides the Sample Tool Mastery and Sample Tool Documentation for further detail. *To be ready for the Alteryx certification, please read both and make sure it all makes sense to you.* Take notes on any questions you may have.

3.5 Unique

Consider again the COVID-19 pandemic and image that early on in the pandemic you were in charge of allocating resources to the states hit hardest by COVID-19. You are given monthly reports from each state that catalog the number of positive tests for COVID-19 and upon your analysis have following summary of COVID-19 infections per state:

State	# of Infections
Wyoming	215,760
New York	91,734
California	70,924
Florida	60,211
Texas	55,984
⋮	⋮
Vermont	10,115
North Dakota	6,672
South Dakota	5,987

According to the summary, Wyoming is by far the hardest hit state and you decide to divert millions of dollars in government assistance to combat the virus. However, the public is skeptical of your decision, and an independent investigation determines that Wyoming actually has one of the lowest number of COVID-19 infections nationwide. It turns out that a glitch in Wyoming's health records system resulted in a large number of duplicate records of COVID-19 infections. Since these duplicates were not excluded from the summations, the number of COVID-19 infections in Wyoming was highly inflated in your analysis leading to a gross misallocation of resources.

Duplicate data can lead to the garbage in, garbage out conundrum, but the **Unique** tool can help a user address this issue and derive quick insight into the data.

Interactive Lesson

Unique Tool Interactive Lesson

Again, consider the words different, Different, and DIFFERENT. Are these the same words? When considered by the Unique tool, these words are in fact unique. Or consider the text values of 2022-08-26 and 08/26/2022? These two values both represent the same date, right? Well, the Unique tool will interpret these as two unique dates unless both are converted to DateTime format first.

Exercise Checkpoint

Complete Exercise 24 - 28

You should now be accustomed to the Unique tool. Alteryx also provides the Unique Tool Mastery and Unique Tool Documentation for further detail. *To be ready for the Alteryx certification, please read both and make sure it all makes sense to you.* Take notes on any questions you may have.

4 Orange Belt

Up to this point, you have learned how to read, inspect, clean, manipulate and write data using Alteryx. However, have you noticed that all the actions that you have learned so far have affected either an entire record or field? And more granularly, those effects were performed on the whole cell of each row or record? But say for instance you wanted to sample your data by state and zip code, but the data only had a field with the full address? How would you sample only part of a record? And what if some of the addresses had state abbreviations, while other addresses have the full state name?

While earning your orange belt, you will develop skills that will allow you to overcome these challenges. At this point, it is recommended that you visit the Simplyx Content page to download the Orange Belt Workflow and complete the exercises while reading along.

4.1 Find Replace

When performing a data analysis, you will frequently use data from a multiple sources and the data may lack consistency. Consider the following two addresses:

1234 Ralphie Road, Boulder, CO 80301

1234 Ralphie Road, Boulder, Colorado 80301

If you want to modify your data so that there is only one record per address, how would you go about that? Presumably you would want to use the Unique or Sample tool, but will that work? One of the many uses of the **Find Replace** tool is the ability to modify specific portions of cell, which would allow you to find CO and replace it with Colorado, or vice versa.

Interactive Lesson

Find Replace Tool Interactive Lesson

Whenever using the Find Replace tool always remember that the tool will **only work with text data types!** Trying to use the Find Replace tool with numeric or datetime data types can leave a user banging their head against the desk. Additionally, it is very important that a user realizes that the data that is connected to the **F (find anchor)** is the data that will be replaced, while the data that is connected to the **R (replace anchor)** is the data that will be replacing the found data. Lasty, another relatively common mistake is having duplicate values in **Find Values** field of the data that will be replacing the found data. For example, consider the following lookup table:

State Abbv.	State Name
HI	Hawaii
HI	Hawai'i

If you were using the above lookup table to replace the state abbreviation HI with the state name in another data set, what would be the resulting value for the state name? Would it be Hawaii or Hawai'i? Well, it would end up being Hawai'i because Alteryx will move from the top row of the lookup table to the bottom and perform the find replace on the data. Since Hawai'i comes after Hawaii, Hawai'i will be the end result. But the main takeaway is to **avoid duplicate values** in the Find values field of the data that will be replacing the match.

Exercise Checkpoint

Complete Exercise 1 - 4

You should now be accustomed to the Unique tool. Alteryx also provides the Find Replace Tool Mastery and Find Replace Tool Documentation for further detail. *To be ready for the Alteryx certification, please read both and make sure it all makes sense to you.* Take notes on any questions you may have.

4.2 Text-to-Columns

While earning your yellow belt, you learned how to slice the data into subsets, but you were only able to slice it by record. While it is convenient when the data you are working with has been organized in a manner where this is sufficient, much of the data that you encounter in the real word will not be so neat. Consider again two addresses:

1234 Ralphie Road, Boulder, CO 80301

5678 Leeds Lane, Boulder, CO 80302

Imagine that these addresses are tied to home prices, and you want to sample the 100 most expensive homes from each zip code for a home renovation promotion. Since the zip code of the address is not in its own row, you can't group by zip code and use a Sample tool. Thankfully, the **Text-to-Columns** tool allows you to separate individual pieces of data into multiple columns, allowing you to perform your analysis.

Interactive Lesson

Text-to-Columns Tool Interactive Lesson

While mentioned in the Text-to-Columns Tool Interactive Lesson and perhaps not quite understood at this point, it is recommended that you **almost always** use the Record ID tool in conjunction with the Text-to-Columns tool. However, the Record ID tool is covered in the next section, so this is not necessary for the Simplyx exercises. Additionally, in the interest of reducing your Workflow's runtime and saving you time, remember that **multiple delimiters** can be used in a single Text-to-Columns tool. Lastly, there are a couple common delimiters that occur frequently in data, but are not obvious to represent:

1. A Tab is represented by **\t**
2. A Newline is represented by **\n**
3. A Space is represented by **\s**

Exercise Checkpoint

Complete Exercise 5 - 15

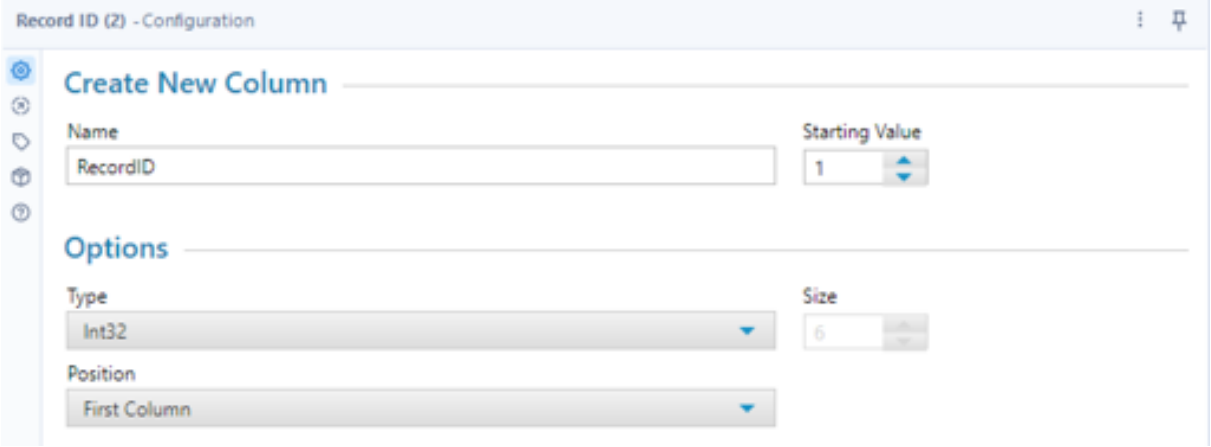
You should now be accustomed to the Unique tool. Alteryx also provides the Text-to-Columns Tool Mastery and Text-to-Columns Tool Documentation for further detail. *To be ready for the Alteryx certification, please read both and make sure it all makes sense to you.* Take notes on any questions you may have.

5 Green Belt

Now that you have earned your orange belt and arrived at the green belt, you are halfway there in terms of number of tools that you have mastered. Well done! But the level of difficulty is about to go up a bit, so prepare yourself. While earning the green belt, you will learn how to group your data in more advanced manners, generate new data from old, and apply formulas that will allow you to derive greater insight from your data. This can be a challenge to many but developing an understanding of these tools will go a long way in your journey to obtain your final certification and become a data analyst. At this point, it is recommended that you visit the Simplyx Content page to download the Green Belt Workflow and complete the exercises while reading along.

5.1 Record ID

As mentioned in the previous section, you almost always want to use a **Record ID** tool in conjunction with the Text-to-Columns tool when splitting to rows. The action of adding a Record ID can seem a bit trivial at this point, but it can be of great use when transforming your data. When adding a Record ID tool to your Workflow, you will be presented with the following Configuration Window:



Record ID (2) - Configuration

Create New Column

Name: RecordID

Starting Value: 1

Options

Type: Int32

Size: 6

Position: First Column

The configuration of the Record ID tool is rather intuitive. A new column (either the first or last column, determined by **Position**) will be created that indexes the number of records in your dataset. The first number of the index is the **Starting Value**. For example, if you wanted the first row to be given the value 101, you would set the starting value to this number. The second row would then be given the value 102, etc.

However, there is one aspect of the Record ID that isn't entirely obvious. When changing the Type of the Record ID tool in the Configuration Window to String, the Record ID tool will add leading 0s so that each record is the selected Size. For instance, if the Size was 6 and the Staring Value was 1, the first value in the newly created column would be: 000001

Exercise Checkpoint

Complete Exercise 1 - 4

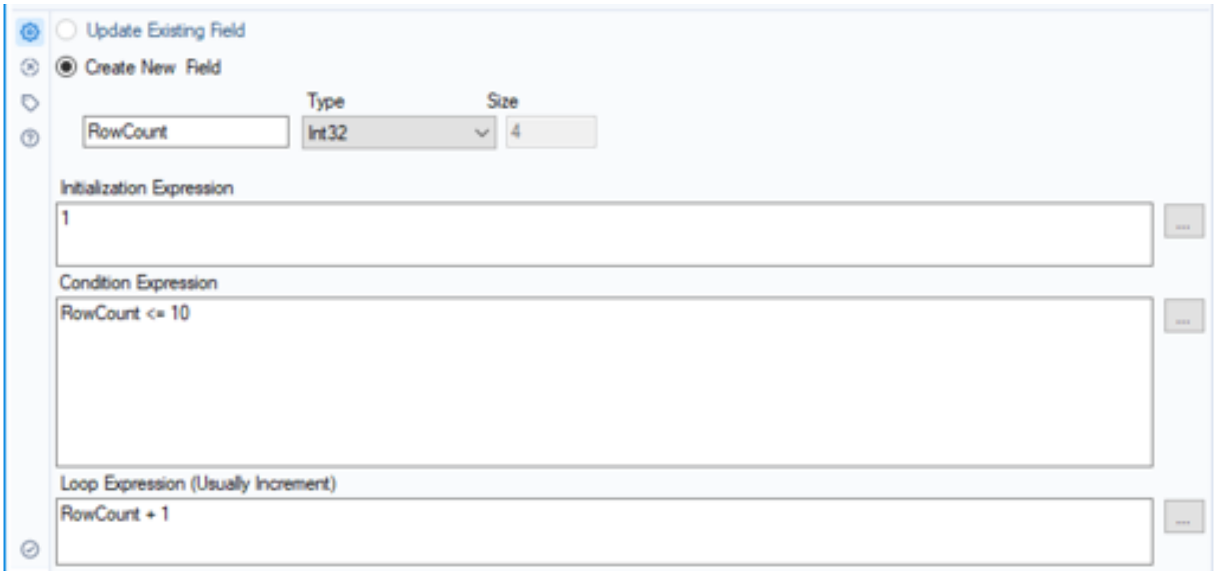
You should now be accustomed to the Unique tool. Alteryx also provides the Record ID Tool Mastery and Record ID Tool Documentation for further detail. *To be ready for the Alteryx certification, please read both and make sure it all makes sense to you.* Take notes on any questions you may have.

5.2 Generate Rows

Occasionally you will need to generate additional records to perform your data analysis. To do this, the Generate Rows tool uses a while loop to construct new rows. Consider the following while loop pseudocode (code that is not in any particular programming language):

```
RowCount = 1
while RowCount <= 10:
    create new record with value RowCount
    RowCount = RowCount + 1
```

Let's step through this and compare it to the Generate Row's Configuration Window:



○ Update Existing Field

● Create New Field

	Type	Size
RowCount	Int32	4

Initialization Expression

1

Condition Expression

RowCount <= 10

Loop Expression (Usually Increment)

RowCount + 1

1. On line 1 RowCount is set to 1
   - In Generate Rows, this is the **Initialize Expression**.
2. On line 2, RowCount is checked to make sure it is less than or equal to 10
   - In Generate Rows, this is the **Condition Expression**.
3. On line 3, a new record is generated with the value RowCount
4. On line 4, RowCount is set to itself plus 1
   - In Generate Rows, this is the **Loop Expression**.
5. The loop moves back to line 2, and steps 2 – 4 repeat until RowCount is greater than 10

Our favorite use-case for the Generate Rows tool is generate a set of successive dates.

Video Lesson

Generate Rows Tool Interactive Lesson

One important consideration when using Generate Rows is making sure that your Conditional Expression will eventually be satisfied. Failing to do so can lead to an infinite loop, and cause Alteryx to crash. So, **ensure that you frequently save your Workflow to avoid losing any prior progress**.

Exercise Checkpoint

Complete Exercise 5 - 8

You should now be accustomed to the Unique tool. Alteryx also provides the Generate Rows Tool Mastery and Generate Rows Tool Documentation for further detail. *To be ready for the Alteryx certification, please read both and make sure it all makes sense to you.* Take notes on any questions you may have.

5.3 Tile

Quite often you will be given data that requires you to group the data in a variety of ways. Grouping, often called clustering, is a key data analyst function. We group student course scores into letter grades, where both a student with a 95 and a student with a 96 would end up in the same group, termed recipients of an A.

Whether that grouping is by number of records, a sum of record values, the standard deviation of record values or their uniqueness, the Tile tool provides a quick and efficient way to do so.

Video Lesson

Tile Tool Interactive Lesson

There are a few important considerations when using the Tile tool. The first consideration has to do **Equal Sum** option. When using this option, consider whether your data includes null values, as the tile tool will essentially consider a null value to be 0. This can lead a user to believe that a grouping is larger than it actually should be.

The second consideration is when using the Unique option. The built in Unique option is essentially using the Unique tool, so all the previously discussed considerations apply, such as white space and cases.

Exercise Checkpoint

Complete Exercise 9 - 11

You should now be accustomed to the Unique tool. Alteryx also provides the Tile Tool Mastery and Tile Tool Documentation for further detail. *To be ready for the Alteryx certification, please read both and make sure it all makes sense to you.* Take notes on any questions you may have.

5.4 Formula

The **Formula** tool is one of the most dynamic tools that Alteryx provides for a user to modify their data. At the heart of this tool are **expressions** which consist of functions, variables, operators, and syntax. For example, consider the following expression:

abs([loan_amount]) + 100

The example expression can be broken down into the following components:

1. abs() is the **function** often used to turn negative numbers positive
2. [loan_amount] is the **variable** (column) to be turned positive, if needed
3. + is the **operator**
4. 1000 is the **syntax**

Interactive Lesson

Expressions Interactive Lesson

One important consideration when using the Formula tool is when a user has multiple expressions in the same tool. By clicking the + icon, a user can add a new expression within the Formula tool. The expression then **run in order from top to bottom**. So, when running the workflow, the expression at the top runs first and modifies the data. Then, the expression immediately below to the top expression runs second and modifies the data. This order continues until all the expressions in the Formula tool have executed.

Another important consideration is the data types that a user is working with when using the Formula tool. The wrong data type can cause unintended consequences or prevent the Formula tool from working. When using the Formula tool, pay attention to the **Data Preview** to ensure the formula is going to modify the data as you intend.

Lastly, when working with data, it is important that a user realize that there is a difference between null values and empty strings. Both of these are used as a placeholders for nothing to the computer. You may have experienced this difference earlier on when using the Filter tool but understanding the difference can be critical when using the Formula tool.

Interactive Lesson

Null & Empty Interactive Lesson

In the Null & Empty Interactive lesson, you learned how to use the IsNull() and IsEmpty() functions, but occasionally you will want to check if the data is not null or not empty. To do this, you can use the **!** operator to negate the function. That is, if you wanted to check if something is not null, you would use the following:

!IsNull([column_name])

Or say for instance that you wanted to check that the value of a cell doesn't equal something. You could use the sub-expression [column_name] != "Something". This ability to work with text is another valuable attribute of the Formula tool. You may have come across earlier problems where you only wanted to remove a certain character like the $ symbol, but the Data Cleanse tool made this difficult to accomplish. When confronted with challenges such as this (and others related to text), the Formula tool can be of great use.

Interactive Lesson

String Functions Interactive Lesson

Exercise Checkpoint

Complete Exercise 12 - 13

Now, consider that you had a bunch of data on sales and the amounts were formatted as $10,000 but you wanted to subtract 1000 from this? First you would need to remove the punctuation (dollar sign and comma), which you could do with the Data Cleanse tool, or as you just learned, with the Formula tool. Then, it is critical that you change the data type to a numeric data type before you try to add to it. Once the data type has been adjusted, you can perform numeric functions on the data.

Interactive Lesson

Numeric Functions Interactive Lesson

Exercise Checkpoint

Complete Exercise 14 - 18

Up to this point, you have learned how to apply expressions to the entire column of data. But what if you wanted to change certain data values if they meet one condition and leave them the same if they didn't. This ability can save a user an immense amount of time, and in many cases is the only way a user can accomplish what they desire to do.

Interactive Lesson

Conditional Statements Interactive Lesson

Exercise Checkpoint

Complete Exercise 19 - 28

You should now be accustomed to the Formula tool. Alteryx also provides the Formula Tool Mastery and Formula Tool Documentation for further detail. *To be ready for the Alteryx certification, please read both and make sure it all makes sense to you.* Take notes on any questions you may have.

5.5 Multi-field Formula

While the Formula tool is helpful when wanting to modify a single field, often a user will want to apply the same formula to many different fields. A user can include many formulas in a Formula tool or use many Formula tools. However, this can become very tedious and cumbersome quite fast. Thankfully though, Alteryx provides the **Multi-field Formula** tool that allows a user to apply a single formula to many fields at the same time.

Interactive Lesson

Multi-field Formula Tool Interactive Lesson

One of the most common issues when it comes to the Multi-field Formula too is the **Copy Out Fields and Add** option. When this is selected, new fields with either a suffix or prefix are created after applying the specific formula to the data. Often, a user expects to use the tool and inspects their data afterwards in the original fields only to see no change. To apply the formula to the original fields, this option must be deselected.

Exercise Checkpoint

Complete Exercise 29 - 32

You should now be accustomed to the Multi-field Formula tool. Alteryx also provides the Multi-field Formula Tool Mastery and Multi-field Formula ToolDocumentation for further detail. *To be ready for the Alteryx certification, please read both and make sure it all makes sense to you.* Take notes on any questions you may have.

5.6 Multi-row Formula

Another common intent for a user is to combine the values of two or more rows into a single expression. This can be very useful when working with data that includes snapshot from dates. For example, calculating running totals or moving averages are very common tasks when working with data. To allow a user to perform these calculations, the **Multi-row Formula** tool can often be the only solution within Alteryx.

Interactive Lesson

Multi-row Formula Tool Interactive Lesson

Probably the most important consideration when using the Multi-row Formula tool is how a user has sorted their rows prior to using the Multi-row Formula tool. The Multi-row Formula tool work from the top down, so not having the rows correctly sorted can lead to incorrect answers that are hard to detect. So, ensure that you have sorted your data correctly before using the Multi-row Formula tool.

Additionally, in the Multi-row Formula Interactive Lesson, you learned about the **Value for Rows that don't Exist** option. The correct configuration for this option can be very important in certain scenarios, so choosing the correct option is important. However, there is a bit a nuance when using the **Group By** option.

First, a user should **first sort the data by the group identifier**, and then sort each of the groups however they believe they should be for the Multi-row Formula tool to perform its intended action. This breaks the data into easier to inspect blocks that allow a user to more easily determine if the Multi-row Formula tool modified the data as intended. Otherwise, if the groups are scattered, assessing how one row changed based on others can be much more difficult.

Second, after sorting the data by groups, for the first group there is quite obviously a row that does not exist if prior rows are being used in the expression. For example, if Row-1 is being used to calculate a value in the 1st record, obviously no record exists before the 1st. But what happens when the Multi-row Formula tool gets to the second group? Say the first group is 20 rows, and the Multi-row Formula tool is calculating the value for the 21st row. Does the Row-1 from the perspective of the 21st row exist? A user can see that there is indeed a 20th row. But in fact, since the Group By option has been selected and the 20th row is in a different group than the 21st row, the Row-1 record **does not exist** from the perspective of the 21st row.

Exercise Checkpoint

Complete Exercise 33 - 35

You should now be accustomed to the Multi-row Formula tool. Alteryx also provides the Multi-row Formula Tool Mastery and Multi-row Formula Tool Documentation for further detail. *To be ready for the Alteryx certification, please read both and make sure it all makes sense to you.* Take notes on any questions you may have.

6 Purple Belt

While earning your orange belt, you learned how to find values in one dataset and replace or append values from another dataset with the Find Replace tool. The skill of being able to combine multiple datasets is critical when analyzing data, as it is often the case that multiple datasets contain relevant data. This ability to join, union or append datasets is the focus of the purple belt. Understanding the fundamentals of these actions is beneficial even beyond Alteryx. For example, working with data while using Python or SQL incorporates joins, unions and appends. So, take care to understand what is happening when using these tools, as doing so will serve you well when working with data. At this point, it is recommended that you visit the Simplyx Content page to download the Purple Belt Workflow and complete the exercises while reading along.

6.1 Dynamic Rename

At this point in your studies, you have encountered dataset with many fields. Often, the names for those fields can be difficult to work with or require modification for numerous reasons. To perform these modifications effectively without manually making the adjustments using the Select tool, the **Dynamic Rename** tool provides a user with ability to do so with ease.

Video Lesson

Dynamic Rename Tool Video Lesson

Exercise Checkpoint

Complete Exercise 1 - 5

You should now be accustomed to the Dynamic Rename tool. Alteryx also provides the Dynamic Rename Tool Mastery and Dynamic Rename Tool Documentation for further detail. *To be ready for the Alteryx certification, please read both and make sure it all makes sense to you.* Take notes on any questions you may have.

6.2 Append

A very common application of data analysis is to score individual data items and then match those data items to others with a similar score. As a result, a recommendation can be served to a user. For example, program recommendations on streaming services try to score their subscribers and then match you to programs from their vast libraries. Inherent in this process is appending data to create combinations that can then be evaluated. For users of Alteryx, the **Append** tool allows users to create these combinations in preparation for future analysis.

Video Lesson

Append Tool Video Lesson

When appending one data set to another, what you are doing is performing a **Cartesian Join** (also known as a Cross Join). Essentially, for every row in one data set, you are joining it to the entire other data set. As a result, **the number of rows in the joined data set is the number of rows in the first data set times the number of rows in the second data set**. So, if the first data set had 10 rows, and the second had 20 rows, the joined data set would have 200 rows. This value is known as the **Cartesian Product**.

It is also possible to append a dataset to itself. In this case, there will be a number of **self matches** and duplicates. For example, consider the following example:

ID	**Name**	**Age**
1	John	20
2	Betty	21

If you used the Append tool to perform a Cartesian Join, you would have the following table:

ID	**Name**	**Age**	**Source ID**	**Source Name**	**Source Age**
1	John	20	1	John	20
1	John	20	2	Betty	21
2	Betty	21	1	John	20
2	Betty	21	2	Betty	21

As can be seen, John joined to John and Betty joined to Betty. Furthermore, John joined to Betty and Betty joined to John. There is little if any value to be had by comparing someone to themselves, and performing the same comparison twice is a waste. While this may seem insignificant with such a small example, keeping self matches and duplicates can be a huge burden with large datasets. For example, consider appending a dataset with 1,000,000 records to itself, which would result in one trillion records (1,000,000,000,000). By eliminating self matches and duplicates, this large number can be roughly cut in half.

To remove self matches and duplicates, a user can add a record id to each record (as was done above, named ID). Then, after appending the dataset to itself, a user can filter the data to remove records where the record id from the original row is less than the record id of all the appended rows (a Filter tool with the condition ID ¡ Source ID). To check that you have done this correctly, if you let n be the number of rows in the dataset, the resulting number of records in the joined dataset should be reduced to:

$$\mathbf{\frac{n \cdot (n - 1)}{2}}$$

Exercise Checkpoint

Complete Exercise 6 - 10

You should now be accustomed to the Append tool. Alteryx also provides the Append Tool Mastery and Append Tool Documentation for further detail. *To be ready for the Alteryx certification, please read both and make sure it all makes sense to you.* Take notes on any questions you may have.

6.3 Join

The Cartesian Join via the Append tool covered in the previous section can be too broad of a join, and instead joining datasets based on commonly held records can often be a more apt solution. To more precisely joining data, the **Join** tool is used.

Interactive Lesson

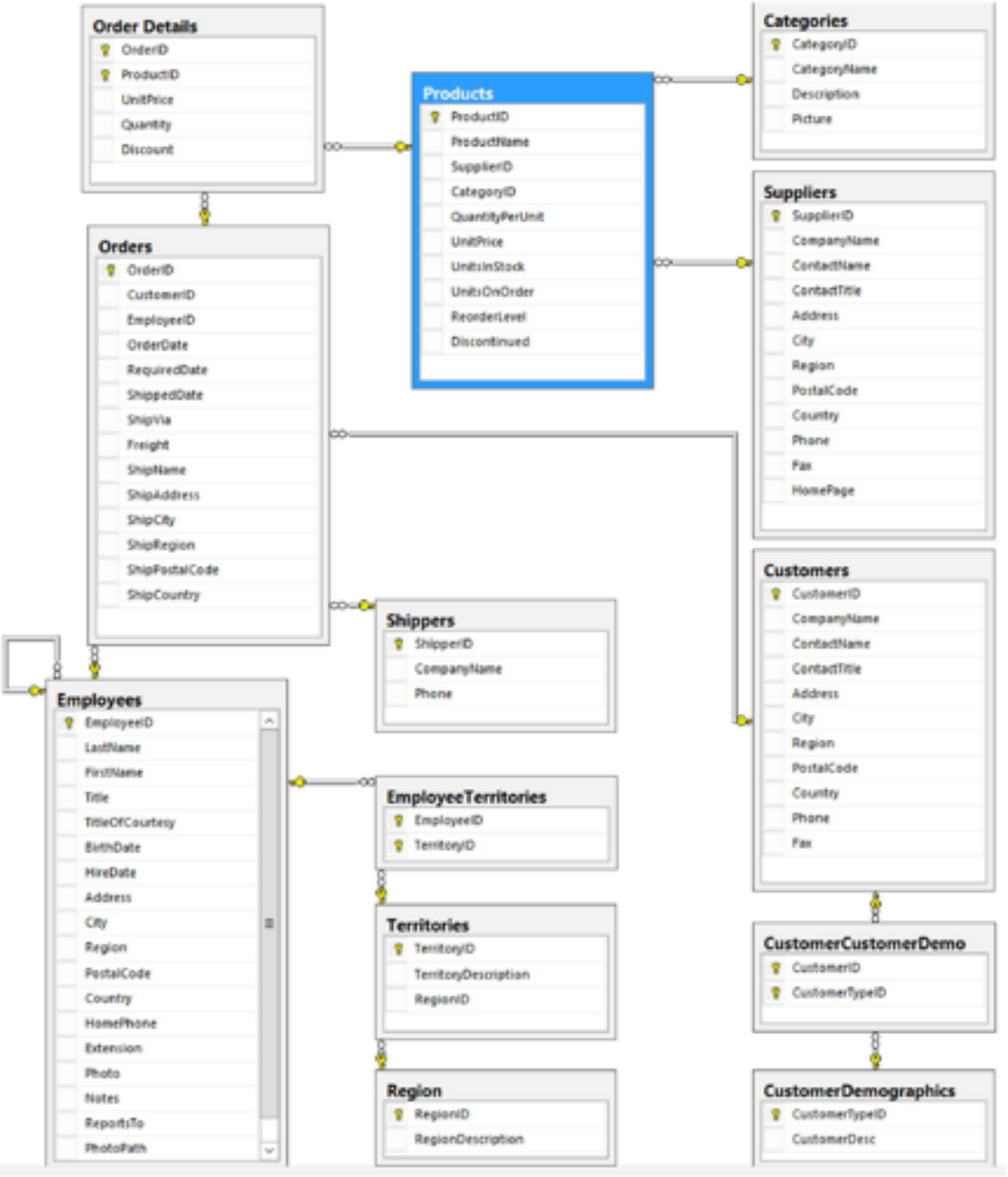
Join Tool Interactive Lesson

It is very important to remember that specific fields used to join datasets **must be the same data type**. Additionally, there are some finer details to consider when performing a join. It is possible for a user to join datasets by certain fields, but to be performing the join incorrectly. To avoid this mistake, consider the following:

A **Primary Key**, which is a **unique identifier** for a dataset (the primary key for you in various databases will tend to be your Student Identifier or Social Security Number). That is, the primary key can only appear once in the data set. Typically, a field that includes the term id will be the primary key, but this is not always the case. Furthermore, a primary key isn't necessarily a single row and instead can be a combination of rows. For example, if a dataset contained a record for each day, and there was a day, month and year column, the combinations of these columns could theoretically serve as the primary key. Lastly, there can only be **one primary key** in a data set.

A **Foreign Key** is a field or combination of fields in another dataset that matches the primary key in other dataset. However, unlike a primary key, a foreign key can occur multiple times in a dataset. Another difference is that there can be **multiple foreign keys** in a dataset. After identifying a primary key in one dataset and the corresponding foreign key in another, a user can correctly perform a join. We will use these terms loosely in this booklet, but when working with databases, such keys are carefully defined and specified to allow the organization and integration of data.

So, how can you identify a primary and foreign key in datasets. Well, as stated earlier, a good strategy is to look for a field with the term id in it. But often a user will be provided with a data dictionary or an **Entity-Relationship Diagram (ERD)**. A sample ERD diagram can be seen below:



In the above example, each table represent an individual dataset. Each label in the tables represents a field found in that particular dataset. For example, the EmployeeTerritories dataset has two fields: EmployeeID and TerritoryID. In each table, there are little key icons, which represent the primary keys for a dataset. On a standalone basis, the keys represent the primary key for each table. So, for the Shippers dataset, the primary key is ShipperID. For the EmployeeTerritories dataset, the combination of EmployeeID and TerritoryID represent its primary key.

When two datasets are considered together, it is possible to determine what is the primary key and what is the foreign key. The first element to notice are the edges connecting the datasets with a key and an infinity sign. This is signifying a one-to-many relationship, from which the existence of a foreign key and primary key can be inferred. To understand this, consider joining the Territories and EmployeeTerritories dataset. As can be observed on the edge, the Territories data set will have the primary key and the EmployeeTerritories will have the foreign key (a copy of Territories' primary key). Way to distinguish whether the same field (TerritoryeId) is a primary or foreign key in a date set/table is sometimes to look at the name of that table. The table with the name mirroring the name of the table is often the primary key (TerritoryId is the primary key for the Territory table). Also, if TerritoryId has repeating values in one of the tables, that field is the foreign key.

Now consider joining the Region and Territories datasets. Here the Region dataset will have the primary key and the Territories dataset will have the foreign key. The primary key and foreign key will be the RegionID. For this example, it is important to note that a field **does not need to have the key icon to be a foreign key**.

Now that you have learned how to correctly join two datasets, what types of joins you can perform needs to be discussed. While the important data can be the data that joins on the primary and foreign key, there is often data that will not be joined. However, just because the data did not join to the other dataset doesn't mean that it is not important to your analysis. Please inspect the following types of joins found in the documentation.

Exercise Checkpoint

Complete Exercise 11 - 17

You should now have learned that the Join tool outputs to either the J(oin) anchor, where you will find the records that the Join tool was able to combine (they had a value in common). The L(eft Unjoin) anchor contains the rows in the Left input table that did not have a match in the right table, and the R(ight Unjoin) anchor contains the rows in the Right input table that did not have a match in the left table.

You should now be accustomed to the Join tool. Alteryx also provides the Join Tool Mastery and Join Tool Documentation for further detail. *To be ready for the Alteryx certification, please read both and make sure it all makes sense to you.* Take notes on any questions you may have.

6.4 Union

In the previous section, you may have been confused how you could perform a left, right or full outer join. If you thought that this could not be done only using the Join tool, then you were correct. To account for this, as well as other application, the **Union** tool allows you to join two datasets in a different manner. The Join tool allows you to join two datasets **horizontally**, but the Union tool allows you to join two datasets **vertically**. That is, you can join the datasets by stacking them on each other according to field names, positions, or manually.

Interactive Lesson

Union Tool Interactive Lesson

To perform a **left outer join** using the Join tool and Union tool, first join the data using the Join tool. Then connect the left (L) output anchor and the join (J) output anchor of the Join tool to the input anchor of the Union tool. To perform a **right outer join**, connect the right (R) output anchor and the join (J) output anchor of the Join tool to the input anchor of the Union tool. To perform a **full outer join**, connect the right (R), left (L), and join (J) output anchors to the input anchor of the Union tool.

Exercise Checkpoint

Complete Exercise 18 - 25

You should now be accustomed to the Union tool. Alteryx also provides the Union Tool Mastery and Union Tool Documentation for further detail. *To be ready for the Alteryx certification, please read both and make sure it all makes sense to you.* Take notes on any questions you may have.

7 Blue Belt

In the previous belt you learned how to manipulate multiple datasets into a single dataset. Sometimes, this dataset – or even datasets that have not been joined or unioned together – are not in the right format for future evaluation and will need to be transformed. Even more often, the data you wish to extract insight from won't be structured in a manner that allows you to perform your desired actions. Overcoming these obstacles and summarizing your findings are the two core aspects that you will master while earning your blue belt. At this point, it is recommended that you visit the Simplyx Content page to download the Blue Belt Workflow and complete the exercises while reading along.

7.1 Summarize

Imagine that you were given a dataset and asked to report to your boss some valuable metrics that your company could act upon. What types of metrics would you report? If the dataset contains numeric data, perhaps some totals, means, medians and modes. If the dataset contains textual data, perhaps the frequency of key terms or non-answers. Metrics like these and others are quickly determined using the **Summarize** tool.

Interactive Lesson

Summarize Tool Interactive Lesson

Exercise Checkpoint

Complete Exercise 1 - 10

You should now be accustomed to the Summarize tool. Alteryx also provides the Summarize Tool Mastery and Summarize Tool Documentation for further detail. *To be ready for the Alteryx certification, please read both and make sure it all makes sense to you.* Take notes on any questions you may have.

7.2 Cross Tab

Often, a single row of data will represent a single event, such as a sale. For example, a sale could include price, date, and salesperson. Now imagine that you have a dataset with many rows of this nature, and your boss requests a report listing the total amount of sales for each person for each day? How could you present this in a readable fashion? Perhaps each column would be a day and each row would represent the total sales for a particular salesperson. This task, among many others, is one possible application of moving **vertical data fields onto a horizontal axis** that is accomplished using the **Cross Tab** tool.

Interactive Lesson

Cross Tab Tool Interactive Lesson

An important consideration to remember when using the Cross Tab tool is that if no grouping is used, the Cross Tab tool will only generate **two** rows of data (one being the field names). If grouping is used, the number of rows (including the field names) will equal the number of groups.

Exercise Checkpoint

Complete Exercise 11 - 15

You should now be accustomed to the Cross Tab tool. Alteryx also provides the Cross Tab Tool Mastery and Cross Tab Tool Documentation for further detail. *To be ready for the Alteryx certification, please read both and make sure it all makes sense to you.* Take notes on any questions you may have.

7.3 Transpose

Another common case is the opposite of the case suggested for the Cross Tab tool. Imagine that your boss has been receiving weekly reports of daily sales for each employee with dates as the column and employees as the row. The end of the year rolls around and your boss requests that you calculate annual sales for each employee. You join each weekly report on the employee and have a dataset with 365 daily sales columns. You could use a formula tool and manually enter in 365 columns, or you move **horizontal data fields onto a vertical axis** with the **Transpose** tool. Then, a simple use of the Summarize tool would calculate annual sales per employee.

Interactive Lesson

Transpose Tool Interactive Lesson

The Transpose tool is used very frequently in tandem with the Summarize tool. A rule of thumb when deciding which **Key Column** (if any) you want to select is to consider how you would want to group the data with the Summarize tool. So, with the example above, the key column you would want to select would be the employee (i.e., the name or employee id).

Exercise Checkpoint

Complete Exercise 11 - 15

You should now be accustomed to the Transpose tool. Alteryx also provides the Transpose Tool Mastery and Transpose Tool Documentation for further detail. *To be ready for the Alteryx certification, please read both and make sure it all makes sense to you.* Take notes on any questions you may have.

References

[1] https://www.statista.com/statistics/871513/worldwide-data-created

[2] https://www.bls.gov/ooh/math/operations-research-analysts.htm

[3] https://ww2.amstat.org/asa175/statcomputing.cfm

[4] https://support.microsoft.com/en-us/office/excel-specifications-and-limits-1672b34d-7043-467e-8e27-269d656771c3