DATA PREPARATION: HOW TO WRANGLE, ENRICH, AND PROFILE DATA
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRODUCTION</td>
<td>2</td>
</tr>
<tr>
<td>GETTING STARTED</td>
<td>3</td>
</tr>
<tr>
<td>EXERCISE 1: DATA IMPORT AND WRANGLING</td>
<td>6</td>
</tr>
<tr>
<td>EXERCISE 2: STRUCTURE YOUR STORY</td>
<td>14</td>
</tr>
<tr>
<td>Detailed Instructions</td>
<td>15</td>
</tr>
<tr>
<td>EXERCISE 4: IMPORT A JSON DATASET</td>
<td>22</td>
</tr>
<tr>
<td>EXERCISE 5: IMPORT MULTIPLE TABLES</td>
<td>31</td>
</tr>
<tr>
<td>CONCLUSION</td>
<td>38</td>
</tr>
</tbody>
</table>
INTRODUCTION

Distance learning is a major new trend in education that gives people the flexibility to learn and pick up new skills at their own speed. In this workshop, you will examine data on the value of online education in the United States. For the sake of this scenario, let’s assume that you made a significant investment in a distance learning program last year. Unfortunately, you did not have enough data to accurately determine the value of this new type of education. To remedy this, you decided to investigate the value of pursuing online education.

GETTING STARTED

Start by launching MicroStrategy Desktop. If you have not downloaded it yet, you can find the software at www.microstrategy.com

1. Double click the MicroStrategy Desktop icon.

2. Click on New Dossier on the Welcome screen.

You will now enter the brand-new dossier authoring interface. Before we get started building a dossier, let’s briefly familiarize ourselves with the interface.
1. **Toolbar:** Provides controls to redo or undo an action, refresh the dossier, add dossier elements (including datasets, pages, chapters, visualizations, on-canvas filters, text, images, and links), change the format, and share your dossier. This is also where you can take advantage of brand-new responsive design functionality and preview your dossier in mobile and presentation mode.

2. **Panel Control:** The icons on the far left-hand side of the interface are used to show or hide different panels.

3. **Contents Panel:** With dossiers, you can organize your information into a chapter and page structure, allowing you to tell a story about your data. The Contents Panel is where you add chapters and pages to your dossier. Click on the editor icon in the Panel Control section to show or hide this panel.

4. **Dataset Panel:** The Dataset Panel shows the names and the elements of the datasets you have imported for your dossier. You can add additional datasets, delete datasets, and link two different datasets here. You can also add new metrics, attributes, links, etc. by right-clicking on the data elements in the panel. To show or hide this panel, click on the dataset icon in the Panel Control section.

5. **Editor, Filter, and Formatting Panels**
   a. **Editor Panel:** This is where you can drag and drop attributes and metrics to add data to your dossiers. Click on the editor icon in the Panel Control section to show or hide this panel.

   b. **Filter Panel:** This is where you can add filters to your dossier at the chapter level, and filter data based on the values of any attribute or metric to change your view to better understand your data. Simple checkboxes, sliders,
calendars, and radio buttons make filters easy to use, and the impact of your selections is instantly visible. MicroStrategy also supports adding visualizations as a filter. Click on the filter icon in the Panel Control section to show or hide this panel.

c. **Format Panel:** With dossiers, you can apply formatting options to enhance a user’s ability to view and read the information in your visualizations. Formatting options vary depending on the type of visualization you are formatting. Click on the format icon in the Panel Control section to show or hide this panel.

6. **Visualization Gallery:** Use the Visualization Gallery to quickly select the visualization you want to use to display your data, or import a custom visualization with just a few clicks. Click on the visualization icon in the Panel Control section to show or hide this panel.
EXERCISE 1: DATA IMPORT AND WRANGLING

Overview
In this exercise, you will create a dossier and use data wrangling to prepare your data. To get started, you will import the dataset Web Tables—Enrollment in Distance Education Courses, by State_ Fall 2012.xlsx into MicroStrategy Desktop.

You will then use MicroStrategy’s data wrangling functionality to prepare the data in the Web Tables—Enrollment in Distance Education Courses, by State_ Fall 2012.xlsx. There are some discrepancies in the way the data in each column is presented. In order to do this you will need to parse and edit the columns.

3. Click on the Save icon MicroStrategy Desktop toolbar and save your dossier as Online Education Data Wrangle.

4. From the menu bar, click the Add Data icon in the tool bar, and select New Data to open the Connect to Your Data interface.

5. For this example, you will import data from an Excel file. Select the File from Disk option.
**TIP**: With MicroStrategy you can instantly connect to nearly any data source, from traditional Excel spreadsheets, to cloud-based applications, big data sources, and even social media sources like Facebook and Twitter. MicroStrategy offers over 80+ data import options for even easier analysis.

6. Click **Choose Files** to browse your local machine and import a spreadsheet.

7. Browse to the **Web Tables—Enrollment in Distance Education Courses, by State_ Fall 2012** file and then click **Open**.

This file should be in the supporting files folder provided to you by the instructor before this session.
TIP: MicroStrategy allows users to import multiple sheets at the same time. This allows for easy uploading of entire Excel worksheets, without requiring you to delete sheets that aren’t used for your visualization. Each sheet will automatically be treated as a different table within the dataset, allowing you to create joins between tables.

8. Click Prepare Data.
9. Ensure that you select only the Table 2 workbook in Web Tables—Enrollment in Distance Education Courses, by State_Fall 2012.xlsx.
10. Click Select.
After the data loads, you will be taken to the Preview window, which contains information about the attributes and metrics included in the different datasets.

11. In order to view the actual columns in the dataset, click on the Web Tables—Enrollment in Distance Education Courses, by State_ Fall 2012.xlsx table in the Preview window.

12. Select the carrot on the “Web Tables—Enrollment in Distance Education Courses, by State_ Fall 2012.xlsx” table, and select Parse.

13. In the box next to Skip, type 4, then click Apply in the upper-left corner of the Parse Your Data interface.
Notice that all the rows are automatically moved up.

14. Next, click the blue **Apply** button located in the bottom-right corner of the interface to apply the changes and return to the **Preview** window.

**Wrangle Your Data**

15. Click the **Wrangle** icon located in the upper-left corner of the **Preview** window.

---

**TIP**: MicroStrategy automatically maps the imported rows and columns as attributes and metrics. Attributes provide a business model with context for reporting and analysis. Metrics represent business measures and key performance indicators (KPIs). If changes to any content or data types of the attributes or metrics are needed, they can be performed manually in the **Preview** window. Note that each attribute and metric is clearly marked. Attributes are represented by a **blue icon**, while metrics are represented by an **orange icon**.

---

Often, imported datasets contain errors, missing records, or incorrect formatting that can make analysis difficult. Rather than searching through each record in Excel and
making individual modifications, we can make changes directly within MicroStrategy by using the Data Wrangling tool. The Preview window shows data for the first 50 rows in the selected dataset. It is designed to give you a general idea of the type of data contained within a specific dataset.

16. Click on the first column, Regions.

17. In the Select Function dropdown menu, select Delete Row, then in Where Cell Contains text box, type “United States”.

18. Click the Apply button in the upper side of the window - notice that this function removes the first row.

19. Select the Regions column again. From the Select Function dropdown menu, click Text Selector.
20. You will notice the Text Selector window appears below the data.

![Text Selector Window](image)

21. Click **District of** in the text selector window. A window will appear above the data.

![Text Selector Window with District of Highlighted](image)

22. Click **Delete Row Included in Selector** on the top-right section.
23. After you delete the “**District of**” row, in the Region window, find the attribute element “**Columbia**”, and hover on it.
24. Click on **Edit**, and replace “**Columbia**” with “**District of Columbia**”. Click **Apply to All**.

![Edit Region](image)

As you saw, the row for “District of” did not contain any data, making it unnecessary for our analysis. On the other hand, the row “Columbia” contained all the data we needed, which is why we renamed it “District of Columbia”.

25. To exit the Data Wrangling tool, click on the **blue Apply** button at the bottom of the window to return to the preview window.
26. Click **Finish** in the **preview** window.

**TIP:** Because the data wrangling tool stores every action in a script, you can revert any of your changes. To do this, look at the History Script window in the top right of the Data Wrangling interface. You can select individual changes to revert using the undo/redo buttons, revert the entire dataset, or simply change the last action you performed.
EXERCISE 2: STRUCTURE YOUR STORY

Create a Table of Content

Now that you have imported your data, you are ready to put together a compelling story for your meeting. Dossiers allow you to build a modern and interactive analytical book of your business. Each dossier consists of individual pages arranged in easy-to-navigate chapters. On each page, you can tell compelling stories about your data with powerful visualizations such as graphs and maps. To provide a guided experience for end users, dossiers include a Table of Contents that help you logically organize content into chapters and pages.

Create a **Table of Contents** with the following structure:

- The first chapter should be called **State Education**. It should have a single page labelled **Online Education per State**.
- The second chapter should be called **Institution**. It should have one page called **Enrollment per Institution**.

When finished the Table of Contents, it will look like the following:

![Table of Contents Example]

Complete Exercise 2 on your own, or use the detailed steps that follow. If completing on your own, continue to Exercise 3 after saving your dossier.
Detailed Instructions

27. Right-click on Chapter 1. In the Table of Contents, and rename it State Education.

28. Similarly, right-click on Page 1, and rename it Online Education per State – You want the first page of your dossier to show key performance indicators.

29. Click the Insert chapter icon on the toolbar to insert a new chapter into your dossier – notice that adding a new chapter also adds a page in that chapter.

30. Name the new Chapter Institution, and the new page to Enrollment per Institution.

31. Save your dossier.

32. In the Visualization Gallery on the right-hand side of the Desktop interface, click the Map icon.
33. In the Web Tables—Enrollment in Distance Education Courses, by State_Fall 2012.xlsx dataset, right-click on the Regions attribute, navigate to Define Geography, and select State from the dropdown menu.

34. Click Ok.

35. Drag the Regions attribute from the dataset to the Geo Attribute drop zone located in the Editor Panel.

36. Open the Format Panel – by clicking the gear icon next to the Editor Panel icon.
37. In the **Graphic type** menu, select **Area**.

38. In the **Map style** dropdown menu, select **Terrain**.

39. Navigate back to the **Editor Panel**.

40. Drag the **Undergrad Students FT Enrollment Percent** metric from the dataset to the **Color By** drop zone.
41. Right-click the metric in the drop zone, and select **Calculation**.
42. Divide the selected metric by 100, and click **OK**.

Since the data type that is coming from the data source does not present the numbers as percent, we will need to reformat these values.

43. To add a threshold, right-click the **Undergrad Students FT Enrollment Percent** metric, and select **Threshold**.

44. In the Threshold window, select **Red-Orange-Green** from the Color dropdown. Click **OK**.

45. Add a new visualization by clicking the **Insert Visualization** icon located in the top of the window.

Next, we’ll move the new visualization below the existing map visualization.
Click on the title bar of the newly created visualization and drag it underneath the map visualization – when dragging the visualization, you will notice a blue horizontal line, this line indicates the location of the visualization.

46. From the Visualization Gallery, select the **Bar Chart** icon.

47. Drag and drop **Undergrad Enrolled Full-Time** and **Graduate Enrolled Full-Time** metrics from the dataset into the **Vertical** drop zone.

48. Similarly, drag the **Regions** attribute to the **Horizontal** drop zone.

49. Move the **Metric Names** (located in the vertical drop zone) label to the **Break By** drop zone.
50. Move the **Metric Names** label from the **Break By** drop zone to the **Color By**.

51. Right-click in the bar chart horizontal gridlines, and select **Hide** from the available options.
52. Double-click on the title of Visualization 1, type **Percent of Undergraduate Students per State**, and hit enter on your keyboard.

53. Repeat the process for Visualization 2, renaming it **Undergraduate vs. Graduate Students per State**.

54. To view the dossier, click on the **Presentation Mode** icon located in the upper-right corner of the interface.

55. To return to edit mode, click the blue pencil icon.
EXERCISE 4: IMPORT A JSON DATASET

Now that we have seen the breakdown of distance education in the United States by State. Let’s evaluate the type of institution and location that have the largest enrollment of online education in the US. In this exercise, you will learn how to bring JSON files into MicroStrategy Desktop.

Import Institutions.json Dataset

56. From the menu bar, click on the Add Data icon in the tool bar, and select New Data to open the Data Source interface.

57. For this exercise, you will import data from a JSON file. Select the File from Disk option.

58. Click Choose Files to browse your local machine and select the file you want to import.

59. Browse to Institutions.json, and then click Open.

This file should be in the supporting files folder provided to you before this session.

60. Click Prepare Data.

In the preview window, you notice this JSON file holds attributes and metrics, similar to the previous Excel file. When you right-click on Percent DL Enrollment and select Change Data Type, you will notice this metric is a Double data type, while Total Enrollment and Distance Students are Integers.
61. Select the carrot in the Institutions.json table. Select **Define Relationships**.

62. Define the relationship as below. In each **State**, there are many **Institutions** (one-to-many), and there is one **Type-of-Institution** per each **Institution** (one-to-one).
63. Click **Save** in the **Define Relationships** window.

64. Click **Finish** in the **Preview** window.

You will now add new visualizations in the **Enrollment per Institution** Page.

65. Select the **Bubble Chart** from the **Visualization Gallery**.

66. From the Institutions.json dataset, drag the **Distance Students** metric, and drop it in the **Vertical** drop zone.

67. Similarly, drag the **Total Enrollment** metric to the **Horizontal** drop zone.

68. Drag and drop the **Type-of-Institution** attribute to the **Color By** drop zone.

69. Next, drop the **Year** and **State** attributes in the **Break By** drop zone (the **Type-of-Institution** attribute should already be there).
70. Lastly, right-click in the Percent DL Enrollment metric in the Datasets panel (not the Visualization panel), and point to Aggregate by, and then select Average.

71. Right-click on the newly created metric, select Number Format, and change from Fixed to Percentage.

72. Add Avg (Percent DL Enrollment) metric to the Size By drop zone.
73. Insert a new visualization and drag it beneath the existing bubble chart.

74. Change the newly created visualization into a **Heat Map** by clicking on the **Heat Map** icon in the **Visualization Gallery**.
75. Place the **Year** and **Institution** attributes inside the **Grouping** drop zone.

76. Drag the **Distance Students** metric in the **Size By** drop zone.

77. Place the **Percent DL Enrollment** metric in the **Color By** drop zone.

78. Right-click on the **Percent DL Enrollment** metric inside the **Color By** box and click on **Thresholds**.

79. Select **Green** from the color dropdown.

80. In the **Based on** dropdown, select **Lowest**.
81. Change the bands to 25, 50, 75, and 90 units. Click OK.

82. Right-click **Percent DL Enrollment** again in the Color By drop zone, and change the **Number Format** to **Percentage**.

Your visualization should now look like the image below:
83. Expand your heat map by clicking the expand icon that appears when you hover over the titlebar.

The expanded heat map should look like the image below:

84. Restore the Heat Map visualization back to the previous size.

85. Rename Visualization 1 **Enrollment Distribution per Type of Institution**.

86. Rename Visualization 2 **Distance Education Enrollment 2012 to 2015**.

87. To view your dossier in **Presentation Mode**, click on the icon located in the upper-right hand corner of the interface.
88. Click on the **Edit** icon at the top-right to go back to the authoring interface.
EXERCISE 5: IMPORT MULTIPLE TABLES

Next you want to evaluate the type of individuals who opt for online education. In this section, you will determine the income level and work schedule of Americans studying online by creating two new charts. You will also explore how to handle multiple tables as one single dataset.

89. From the menu bar, click the Add Data icon in the toolbar and select New Data

90. Select the File from Disk option.

91. Click Choose Files to browse your local machine and import a spreadsheet.

92. Browse to the DL_data.xlsx, and then click Open.

This file should be in the supporting files folder provided to you before this session.

93. Next, click on the Prepare Data button.

94. Unclick the All option, and then select the rest of the workbooks.

95. Click Select.
After the data loads, you will be taken to the Preview window, which contains information about the attributes and metrics included in the different datasets.

In the preview window, you see six tables. Each table correspond to the worksheet you selected in the previous window. The third table (Age), contains a meaningless attribute called Column 7.

96. Right-click on that column, and select **Do Not Import**.

You may notice there is a **Dependency status** attribute in the Dependency_stat table and a **Dep stat** attribute in the Age table. You will now map these 2 columns together since we know that they are the same attribute with names.

97. Right-click on the **Dep stat** attribute in the Age table, rename it **Dependency status**, and click enter on your keyboard.

98. A notification window will appear asking you whether you want to map this attribute to Dependency status, click **Map?** to map the attributes.
A small link icon ⬅️ appears next to **Dependency status**. It indicates that these two attributes are now linked.

The first table, **Type of Institution**, contains an erroneous attribute called **Percent whose entire program was distance education**. This column needs to be converted to a metric.

99. Right-click on **Percent whose entire program was distance education** attribute column, and select **Convert to Metric**.

As soon as you convert this column to metric, it will be renamed **Percent whose entire program was distance education(2)**.
100. Right-click on this metric, and select Change Data Type to Double.

101. Right-click once again, and select Rename.
102. Change the name to Percent whose entire program was distance education.
103. Click Finish to return to the dossier authoring interface.
104. Create a new chapter called Demography with a page named Income.

105. Select Combo Chart visualization.
106. From the DL_data dataset, place **Dependency Income level** attribute in the vertical drop zone.

107. Drag and drop the metrics **Percent who took any distance education course in 2011-12** and **Percent whose entire program was distance education** in the horizontal drop zone.

108. Move the **Metric Names** label to the **Break By** and **Color By** drop zones.

109. Right-click on the **Percent who took any distance education course in 2011-12** metric in the horizontal drop zone, and select **Change Shape**, then choose **Tick**.
110. Right-click on the visualization’s vertical grid lines, and select Hide.

111. Right click in the **Percent whose entire program was distance education** metric, and select Change Shape, then **Square**.

112. Rename the visualization **Income Level of Distance Education Students**.
113. Add a new visualization to the Income page.
114. Insert a Sequence Sunburst Chart Visualization from the Custom session of the Visualization Gallery.

115. From the DL_data dataset, place the Work and Type of institution Attributes in the Attribute drop zone.
116. Drag and drop the Percent whose entire program was distance education metric in the Metric drop zone.

117. Rename the visualization Working Professionals Pursuing Online Education Entirely.

118. Move the Sequence Sunburst visualization under the first visualization
CONCLUSION

Congratulations! You have completed the Data Preparation for Dossier Desktop workshop. Visit the MicroStrategy community for D3 and advanced visualizations in Dossier. The community can be found at https://www.community.microstrategy.com

For further MicroStrategy Dossier information, including resources and certifications in the MicroStrategy Community, please visit:
https://www.microstrategy.com/us/services/education/