**​Preparing Data**

**Objectives**

Data needs to be cleaned before it can be analyzed. Cleaning the data makes it easier to read and interpret, ensures consistency and accurate results, and enables a better decision-making process.

**Background / Scenario**

The rapidly increasing volume and complexity of data are due to growing mobile data traffic, cloud computing traffic, and adoption of technologies including the Internet of Things and Artificial Intelligence. Data in its raw form is useless unless it can be cleaned and analyzed to provide information that leads to knowledge and wisdom for actionable intelligence. Businesses are quickly discovering that they do not actually want data itself—they want the information and knowledge that comes from data, which they can use to make better decisions.

**Required Resources**

* Mobile device or PC/laptop with an internet connection and Microsoft Excel or similar spreadsheet program

**Instructions**

**Part 1: Explore a Sample Set of Data**

**Step 1: Open the Bike Sales\_Prepare\_Lab 3.4.7.xlsx**

1. Download the sample Excel file entitled **Bike Sales\_Prepare\_Lab 3.4.7.xlsx** and explore it.

**Step 2: Expand Datasheet Columns as Necessary**

You will notice that some of the data is truncated in the columns. Expand the columns so you can review the data.

**Step 3: Review the Data.**

1. Review the raw data to locate any data that could skew the data analysis.

List any issues in the data that can affect your analysis.

**Data Cleaning**

**Step 1: Finding Duplicates**

The data set provided may contain duplicate entries. One of the processes of cleaning data is to find and remove these. In the Bike Sales dataset, the only column that cannot have the same value more than once is column A, Sales\_Order #.

1. Select column A to check for duplicate data here.
2. With column A selected, click the **Conditional Formatting** button in the **Home** toolbar and select **Highlight Cell Rules** > **Duplicate Values**, then click **Done**.  
   Conditional formatting should find two pairs of duplicates. Cells A2 and A3 both have the same sales order number of 261695. Also, cells A8 and A9 have the same sales order number of 261701. These types of duplicate entries can easily occur during manual data entry or when copying and pasting data into a worksheet.

Review the duplicated entries.

In the case of cells **A2** and **A3**, it appears that the Sales Order# of 261695 was incorrectly entered into cell **A3**. As a data analyst, I would need to go to the source of the data and verify the Sales Order number.

For cells **A8** and **A9**, a close review shows that both rows are exactly the same. Most likely in this case a sales entry was entered twice.

**Step 2: Fixing and Removing Duplicates**

When duplicate data entries are identified, they need to be reviewed carefully before they are removed, so that relevant data is not accidentally deleted.

1. To correct the duplicate entry in cell **A3**, change the Sales Order# to **261696**. The conditional formatting should be automatically removed by Excel.
2. To correct duplicate rows 8 and 9, one of the rows needs to be removed.  
   1) One way to remove a duplicate entry is to select a row and delete.  
   2) If there are many duplicates that need to be removed in a large dataset, the Remove Duplicates tool can be used.
3. Click the **Remove Duplicates** tool in the **Data** tool bar.
4. In the **Remove Duplicates** dialog box, select the **Sales\_Order**# column and ensure that the checkbox **My data has headers** is selected. Click **OK** to continue.

Excel will remove the second instance of each set of duplicate rows.

**Step 3: Finding Empty Cells**

There are many reasons why a cell might be blank. It could be human error from manual data entry or it could be a result of copying data from other sources. Context is key when determining what to do with empty cells. Sometimes, as a data analyst, I will need to fill every blank cell in the data with the same constant value. Other times, there may be clues as to what should be in an empty cell from the surrounding data. Analysts may also have to go back to the source of the data to discover what the missing values should be.

To find empty cells, the **Conditional Formatting** tool can be used.

1. Select the entire sheet by clicking the arrow in the top left corner of the worksheet to the left of column A.
2. Click on the **Conditional Formatting** tool in the **Styles** tool bar and select **Highlight Cell Rules** > **Text That Contains**.
3. In the **Conditional Formatting** window, under **Rule Type**, select **Blanks**. Change **Format** to **Red fill with dark Red text** and click **Done**.  
   Any blank cells in the worksheet should now be filled in green. There should be four highlighted cells: C12, G17, M23, and N24.  
   A data analyst would likely review the blank cells to see if the missing data can be obtained. If it cannot, the only option may be to delete the rows with the missing data.

**Step 4: Data Parsing from Text to Column**

I noticed some of the cells have multiple data elements separated by a data delimiter like a comma. For example, consider the column **Product\_Description**. You can parse the data in this column so that each part of the product description is displayed in its own column. You will use the **Text to Columns** function to achieve this.

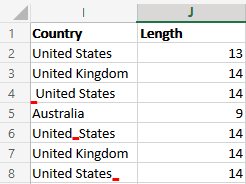
You will parse the data in the column **Product\_Description** to move the bike size and color into separate columns.

1. Start by adding a new blank column to the right of column M, **Product\_Description**. This new column becomes column N.
2. Highlight column M, **Product\_Description**.
3. In the **Data** toolbar, click the **Text to Columns** button in **Data Tools** ribbon.
4. In the **Text to Columns** window, select **Comma** as the only delimiter and click **Apply**.
5. All of the bike sizes should now be moved into column N.
6. Name column N “**Size**”.  
   Next, repeat this process to remove the color of the bike from the **Production\_Description** column.
7. Add a new column again to the right of column M.
8. Use the **Text to Column** tool again, but this time set the delimiter to **Space**.  
   This moves the color of the bikes to the new column so that all that is left in column M now are the bike models.
9. Name the new column “**Color**”.
10. Change the column name of column M from Product\_Description to **Model**.
11. You should now have three columns that each display one part of the bike’s description.

**Step 5: Removing Extra Spaces**

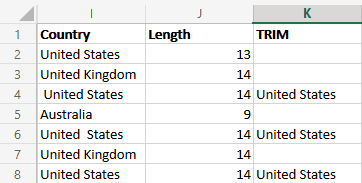
When data is pasted from external sources into an Excel worksheet, there is a good chance that cells may contain extra spaces that will need to be removed so that searches and queries will generate accurate results. The TRIM function is used to eliminate excess spaces and tab spaces in Excel worksheet cells. In this step, you will use the TRIM functions to clean up the data.

1. Insert two new blank columns to the right of column I, **Country**. These will become columns J and K.
2. Name the two new columns Length and TRIM, respectively.
3. In the Length column, in cell **J2**, enter the function **=LEN(I2)** to see how many characters are in cell **I2**.  
   The result should be 13, which is the number of characters in “United States” if counting the space in the middle.
4. Now copy this **LEN** function from cell **J2** down through cell **J8**.  
   Notice that the other cells with United States are showing a length of 14 characters and not 13. This is because each of these cells contains an extra space.  
   In cell **I4** the extra space is in front. In cell **I6** the extra space is between the words United and States. In cell **I8** the extra space is not readily noticeable, but it is at the end.

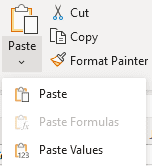


To remove the extra spaces in these cells, use the TRIM function in column K.

1. In cell **K4**, enter **=TRIM(I4)**. The function removes the leading space from cell **I4**.
2. Copy the TRIM function to cells **K6** and **K8**. The function removes the spaces from these cells.



1. To ensure that any extra spaces in the Country column are removed, copy and paste the **TRIM** function to all the cells in column K.  
   All extra spaces in column M are now removed. The values in the TRIM column now need to be pasted to the Country column.
2. Select cells **K2** through **K89** and copy them.
3. Select cells **I2** through **I89** and click the down arrow under the **Paste** tool and then select **Paste Values**.

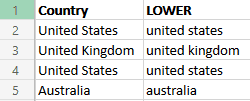


This pastes only the values (country names) into column I, and not the formulas that were in Column K.

1. Delete the Length and TRIM columns, as they are no longer needed.

**Step 6: Changing Case**

1. Add a new column to the right of column I, Country. This new column will become column J.
2. Label the new column LOWER.
3. In cell **J2**, enter the function**=LOWER(I2)**  
   The result should be “united states,” all in lower case.
4. Copy and paste the formula to cells **I3** through **I89** to make all the country names lower case.



1. Copy all of the country names in the LOWER column, column J, and use **Paste** > **Paste Values** to paste them to the Country column, column I.
2. Once finished, delete the LOWER column as it is no longer needed.

**Step 7: Highlight Possible Errors**

In this step, you will highlight all the Unit Costs and Unit Prices that are zero. This type of data is false and will skew the dataset, so it is important to find these errors and fix them.

1. Select the Unit\_Cost column and the Unit\_Price column.
2. Click the **Conditional Formatting** tool > **Highlight Cell Rules** > **Equal to**.
3. In the Conditional Formatting window, enter **0** under the Equal to box and click **Done** to highlight all the bikes with a unit cost or unit prices of zero.  
   Cell **O9** under Unit Cost and Cell **P9** under Unit\_Price should be highlighted in red.  
   A data analyst would need to determine what values to input into these cells or to delete these two rows of data. In this example, we know the values that should be in these cells, because they are in other rows of the sheet.
4. Correct the issue by entering $1252 in cell **O9** and $769 in Cell **P9**.

**Step 8: Find and Replace**

we want to replace F and M in the Customer\_Gender column with the full words Female and Male for readability.

1. Select the **Customer\_Gender** column.
2. Click the **Find & Select** button and select **Replace**.
3. In the Find and Replace dialog box enter the following information.  
   1) **Find what**: F  
   2) **Replace with**: Female  
   3) Expand **Search Options** and under **Within**: select **Selection**
4. Click **Find all**.  
   All cells with a value of F are highlighted.
5. Click **Replace all**.  
   50 matches should have been replaced.
6. To find and replace “M” with “Male,” use the same process, except under **Search Options** make sure **Within**: is set to **Selection** and that **Match case** is checked. This will prevent Excel from changing the “m” in Female to Male, How many replacements were made from **M** to **Male**?

**Step 9: Spell Check**

The feature of checking spelling is available in the **Review** tab. You can check the spelling by cell, column, row, or sheet. Spell check will ensure that spelling errors don’t cause results of searches or queries to be incorrect.

1. Select all columns with text values.
2. In the **Review** toolbar, click the **Spelling tool**.  
   Words not found in the dictionary are displayed in the **Spelling** dialog box that appears on the right side of the worksheet.
3. Click **Ignore** for any of the column names that are identified as not in the dictionary.
4. Select **Change All** for any words found that are misspelled. (**Change All** is found under the drop arrow at the right end of the **Suggestions box**.)

**Step 10: Remove Formatting.**

Sometimes the data set may have formatting embedded within the spreadsheet. The formatting can be as simple as color or text alignment. There may also be a logical condition applied to your cells using Excel's conditional formatting option that gives many cells different formatting than the rest of the sheet.

In this example, the alignment formatting of the **Age\_Group** column will be removed.

1. Select the **Age\_Group** column.
2. With the column selected click the **Clear** button in the **Home** toolbar and select **Clear Formats**.  
   This will remove the center alignment formation of the **Age\_Group** column.

**Save My Work and Close Excel**

1. Save the spreadsheet as Bike Sales\_Prepare\_Lab 3.4.7\_fixed.
2. Close Excel.

© 2017 - 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public