Module 3

1. Cleaning and Wrangling Data Using a Spreadsheet
   1. 5 Traits of good data
      1. Accuracy
         1. Most significant
         2. Remove duplicates
         3. Remove nulls
      2. Completeness
         1. If the data set is complete or is filled with missing data.
      3. Timeliness
         1. Availability and accessibility of the available data.
      4. Reliability
      5. Relevance
         1. Consider if the data is relevant.
   2. Importing Data
      1. Excel by default works with xlmx, but it can work with csv and plain text.
      2. Using the text import widget
         1. Choose file type (Delimited, fixed width)
         2. Select if the data has headers.
         3. Select general for column data format as that’s most common.
2. The Fundamentals of Data Quality and Data Privacy
   1. Confidentiality
      1. Acknowledges that the customer personal information belongs to them.
         1. Sales forecast
         2. Employee information
         3. Patient Records
         4. Pi (Personal Information)
            1. Information that can be traced back to a specific individual.

emails, phone number etc.

* + - 1. PII (Personally identifiable information)
         1. Information that could be used to identify an individual

SS number and Driver’s License Number

* + - 1. SPI (Sensitive Personal Information)
         1. Information that necessarily identify an individual but contains private information that could be used to harm the individual.

Race, Sexual orientation, biometric or genetic information.

* 1. Collection and use
     1. The analyst must know the location of the company collecting the data and the location of the respondent.
        1. The general data protection regulation (GDPR)
           1. Specific to the EU
        2. Lei Geral de Protocao de Dados Pessoais (LGPD)
           1. Specific to Brazil.
        3. US doesn’t have a country wide principal law
           1. CCPA California created to protect citizens.
        4. HIPAA (Industry specific regulations)
           1. Govern the collection and disclosure of protected health information
           2. PCI standards govern credit cards data
           3. Failure to safeguard cardholder info can result in steep penalties.
  2. Compliance
     1. All companies within their respected regions must always abide by these regulations and must be readily available to its employees.
     2. Anonymous mean to exclude all data which ties it back to a particular individual.

In this lesson, you have learned the following information:

The Five Traits of Data Quality:

* Accuracy
* Completeness
* Reliability
* Relevance
* Timeliness

Importing Text:

* You can use the ‘Text Import Wizard’ to import data from other formats, such as plain text, or comma-separated value files.

The Three Fundamentals of Data Privacy:

* Confidentiality
* Collection and Use
* Compliance

1. Removing Duplicated and Inaccurate Data
   1. Types or inconsistency
      1. Spelling mistakes
      2. Extra white space
      3. Wrong case
   2. Data must be cleaned.
      1. To check for spelling in excel
         1. Select column, (Review), Spelling, Click change or select a different suggestion or ignore.
      2. Empty rows
         1. Select whole datasheet (CTRL+SHFT+END)
         2. DATA Tab
         3. Filter
         4. Select column
         5. Scroll down to (BLANKS)
         6. Highlighted in blue number on the left are the blank rows.\
      3. Removing Duplicate
         1. Do not select a column where you would expect duplicate data.
         2. Select appropriate column
         3. Method 1 (Recommended)
            1. Home tab
            2. Conditional Formatting

Highlight cell rules

Duplicate values.

Delete rows.

* + - 1. Method 2
         1. Data tab
         2. Remove Duplicates

Select rows you want to remove duplicates from.

* + 1. Changing the Case of text
       1. Some upper some lower or proper
          1. Use functions such as upper, lower and proper.’
       2. Insert a helper row.
          1. =PROPER(Cell)
       3. Insert a helper column
          1. =UPPER(Cell)
       4. Double clicking the fill handle will cause it to copy the formular all the way to the last row or column.
       5. With values from the helper column to replace the original values in the beginning. Use the paste values option.
    2. Changing data formats
       1. Open number formation.
       2. Select the format you want.
    3. Removing white spaces
       1. Select all rows.
       2. Find and replace function
          1. Find double space
          2. Replace with single space.
       3. Trim function
          1. Add a helper column

=TRIM(CELL)

Autofill

Paste Values

* + 1. Flash Fill
       1. Add helper column (For combining text)
          1. Just type the first name
          2. When typing the second name suggestion to fill will appear
          3. Press enter
    2. Text to Column
       1. Create new columns
          1. Select data
          2. DATA Tab

Text to column

Select delimited.

Select space

Select the new cell.

* + - 1. Extracting text same idea.
         1. Use

=LEFT (A2, SEARCH (“”, A2,1))

=RIGHT (A2, LEN(A2)-SEARCH (“”, A2,1))

Lab

Exercise 1: Removing Duplicated, Irrelevant or Inaccurate Data

In this exercise, you will learn how to deal with inaccurate data, how to remove empty rows, and how to remove duplicated data.

Task A: Check spelling

1. Download the file [**Customer\_demographics\_and\_sales\_Lab5.xlsx**](https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DA0130EN-SkillsNetwork/Hands-on%20Labs/Lab%205%20-%20Cleaning%20Data/Customer_demographics_and_sales_Lab5.xlsx). Upload and open it using Excel for the web.
2. Select column **L (CREDITCARD\_TYPE)**, then click **Review** tab, and select **Spelling**.
3. Click the correct suggestion to change the spelling.
   * **Note:** Don’t change ‘jcb’ spelling when doing the spell check. We will need 'jcb' for the Exercise 1 Task D.
4. Close the **Spelling** pane.

Graphical user interface, application

Description automatically generated

Task B: Remove empty rows

1. Press **CTRL+HOME**, then press **CTRL+SHIFT+END** to select the whole datasheet.
2. On the **Data** tab, click **Filter**.
3. Press **CTRL+HOME**, click the **filter arrow** in the **CUST\_NAME** column, and then click **Filter**.
4. Click the **Select All** checkbox to deselect all of them. Then select just **Blanks**, then **OK**.
5. Select **first row**, then press **CTRL+SHIFT+END** to select all rows.
6. Right-click the selected rows and then click **Delete Rows**.
7. Finally, on the **Data** tab, click **Clear**, then click **Filter**.

Graphical user interface, table

Description automatically generated

Task C: Remove duplicate rows

1. Select Column **T (ORDER\_ID)** since ORDER\_ID values are unique.
2. On the **Home** tab, click **Conditional Formatting> Highlight Cells Rules> Duplicate Values**, and then click **OK**.
3. Select the whole datasheet **(CTRL+SHIFT+END)**
4. On the **Data** tab, click **Remove Duplicates**.
5. In the Remove Duplicates dialog box, ensure that **Select all columns** is checked and that **My data has headers** is also checked, then click **OK**.
6. In the pop-up box informing you how many duplicate values were found and removed, click **OK**.

Graphical user interface, text, application

Description automatically generated

Task D: Use Find & Replace to correct misspelling

1. On the **Home** tab, click **Find & Select**.
2. Click **Find**. In Find what, type **jcb**, and click **Find All**.
3. Click **Replace**.
4. In Replace with, type **JCB**, click **Replace All**, and then click the **Close** icon.
5. On the **Home** tab, click **Conditional Formatting> Clear Rules> Clear Rules from Entire Sheet**.

Table

Description automatically generated

Exercise 2: Dealing with Inconsistencies in Data

In this exercise, you will learn how to change the case of text, how to change date formatting, and how to trim whitespace from data.

Task A: Use the PROPER function to change text from upper case to proper case

1. Select row **2**, then right-click it and choose **Insert Rows**.
2. In cell **A2**, type **=PROPER(A1)** and press **Enter**.
3. Hover over the bottom-right corner of cell **A2**, and drag the **Fill Handle** across to the last column.
   * If dragging across is too difficult with the mouse, then select the cells in the row 2 using **SHIFT+RIGHT ARROW**, then press **F2** to put the cursor focus back in cell **A2**, then hold **CTRL** while you press **Enter**.
4. Select row **2**, then press **CTRL+C**.
5. Select row **1**, Right-click and choose **Paste Options>Values**.
6. Select row **2**, right-click it and choose **Delete Rows**.

Task B: Use the UPPER function to change text from proper case to upper case

1. Select column **AG (Generation)**. Then right-click and choose **Insert Columns**. In cell **AG1**, type **Generation**.
2. In cell **AG2**, type **=UPPER(AH2)** and press **Enter**.
3. Hover over the bottom-right corner of cell **AG2** and double-click the **Fill Handle**.
4. Select column **AG**, then press **CTRL+C**.
5. Select column **AH**, right-click and choose **Paste Options>Values**.
6. Select column **AG**, right-click it and choose **Delete Columns**.

Task C: Use the LOWER function to change text from proper case to lower case

1. Select column **AC (T\_Type)**. Then right-click and choose **Insert Columns**. In cell **AC1**, type **T\_Type**.
2. In cell **AC2**, type **=LOWER(AD2)** and press **Enter**.
3. Hover over the bottom-right corner of cell **AC2** and double-click the **Fill Handle**.
4. Select column **AC**, then press **CTRL+C**.
5. Select column **AD**, right-click and choose **Paste Options>Values**.
6. Select column **AC**, right-click it and choose **Delete Columns**.

Task D: Change date formatting

1. Select column **Z (Order\_Ship\_Date)**.
2. On the **Home** tab, in the **Number** group click **Number Format> More Number Formats**.
3. In the Category list, select **Date**.
4. In the **Format Cells** box, under **Locale**, select **English (United States)**.
5. Under **Type**, select **Wednesday, March 14, 2012** and click **OK**.

Graphical user interface, text, application, email

Description automatically generated

Task E: Use Find & Replace to trim whitespace

1. Click **CTRL+HOME**.
2. Select all the data using **CTRL+SHIFT+END**.
3. On the **Home** tab, click **Find & Select**, then **Replace**.
4. In Find what, type **2 spaces**. In Replace with, type **1 space**.
5. Click **Find All**, then click **Replace All**.
6. Click the **Close** icon.

Exercise 3: More Excel Features for Cleaning Data

In this exercise, you will learn how to use the Flash Fill feature and functions in Excel to help clean data.

Task A: Use the Flash Fill feature to clean data:

1. Select column **A (Cust\_Name)**, right-click and choose **Insert Columns**.
2. In cell **A1** type **Customer\_Name** and press **Enter**.
3. In cell **A2**, type **Mr. Allen Perl** and press **Enter**.
4. Select column **A (Customer\_Name)**, on the **Data** tab, click **Flash Fill**.
5. Click **Undo** to undo this step.

**If you are using the desktop version of Excel, you could use the 'Text to Columns' feature to perform this next task (see the corresponding topic video for instructions).**

**If you are using ‘Excel for the web’ (the online version of Excel), the ‘Text to Columns’ feature is not available, but you can achieve the same results using functions, as shown in the steps below.**

Task B: Use LEFT, RIGHT, LEN, and SEARCH functions to clean data:

1. Select column **A (Cust\_Name)**, right-click and choose **Insert Columns**.
2. Select column **A** again, right-click and choose **Insert Columns**.
3. In cell **A1**, type **Customer\_Firstname** and in cell **B1**, type **Customer\_Lastname**.
4. Click **C1**, then on the **Home** tab, click **Format Painter**, then drag **across to A1 and B1**.
5. Double-click the **divider between columns A and B**.
6. In cell **A2** type **=LEFT(C2, SEARCH(" ",C2,1))** and press **Enter**.
7. In cell **B2** type **=RIGHT(C2,LEN(C2)-SEARCH(" ",C2,1))** and press **Enter**.
8. Double-click the **Fill Handle** on cell **A2**.
9. Double-click the **Fill Handle** on cell **B2**.

Congratulations! You have completed Lab 5, and you are ready for the next topic.

In this lesson, you have learned the following information:

* It’s important to remove any duplicated or inaccurate data, and it’s important to remove any empty rows in your dataset.
* There are several other types of data inconsistency that you may need to resolve, in order to properly clean your data:

1. Change the case of text
2. Fix date formatting errors
3. Trim whitepace from your data

* You can use the Flash Fill and Text to Columns features in Excel to manipulate and standardize your data, and functions can also be used to help manipulate and standardize your data.