Data Cleaning and Manipulation with Excel

## Installation

**Getting Started**

Here are the instructions on how to get Excel;

1. We get a free license through SMU: <https://www.smu.edu/oit/services/microsoft365>. You can use Excel online or download the application; I recommend downloading the app, because there is some additional functionality.
2. Download the excel workbook from the GitHub repository under Week1: <https://github.com/Statistics-and-Data-Science-SMU/Intro-Data-Project/tree/main/Week1>
3. That’s it! You’re all set up.

**Workbook vs CSV**

* What is a CSV? CSV stands for Comma Separated Values; files of this type have the ".csv" extension. It's a common data type, where data is stored as text separated by commas. You can open and edit these files in excel, but if you save a file as a CSV it'll change the formatting back to the default comma-separated format. If you want to save your tables/graphs, make sure to save your files as an Excel workbook.
* What is an Excel workbook? Files with the ".xls" or "xlsx" (they mean the same thing) are Excel workbooks. Data in this format is saved in a tabular form with rows/columns.

**Data Cleaning**

Remove Duplicates:

* Using "Remove Duplicates" to clean the Orders and Returns tables.
* Identifying and handling missing data (e.g., blank cells).
* Splitting Columns (Text to Columns function)
* Formatting Cells

**Data Analysis with Formulas**

* Skipping these basic ones, assuming they know these
  + Basic formulas (sum, average, count, min, max, etc.)
  + Logical functions (if, and, or)
* Lookup functions:
  + V Lookup – You have an Orders table, and you want to find the customer name associated with a specific Customer ID.
    - =VLOOKUP(*customer\_id*, Orders!A:Y, 7, FALSE)
      * *Customer\_id* is the id you want to look up
      * Orders!A:Y is the range of your table, with the first column containing ID
      * 7 specifies that we want to return the 7th column (Customer Name)
      * False indicates an exact match
  + H Lookup if our table is set up horizontally; same logic
  + X Lookup – ex. Find the shipping cost for a specific Order ID
    - =XLOOKUP(*order\_id*, Orders!A:A, Orders!F:F, "Not Found")
      * *Order\_id* is the order to look up
      * Orders!A:A is the column to search for the Order ID
      * Orders!F:F is the column to return the shipping cost
      * “Not Found” is the value returned if it’s not found

**Task:**

**I want to find the total cost of all returned items, using Lookup functions.**

Steps:

1. Go to the **Returns** window
2. Convert the values there into a table
   1. Go to **insert** tab, press **table**, select the range =$A$1:$B$1635
3. Create new column for cost
   1. In Cell **C1**, type “**Cost**”
4. Find the Cost of each return
   1. =XLOOKUP([@[Order ID]], Orders!Y1:Y1956, Orders!E1:E1956, "Not Found")

**Pivot Tables**

Pivot tables are one of Excel’s most powerful tools for data analysis. They allow you to summarize and reorganize complex data in a way that’s easy to interpret.

**Creating a Pivot Table**

1. **Select Your Data**
   * First, select the data you want to summarize. Ensure your table has proper headers.
   * Go to the **Insert** tab and click on **PivotTable**.
   * In the pop-up window, Excel will automatically select the data range. Choose whether you want the pivot table on a new worksheet or in the existing one, then press **OK**.
2. **Setting Up the Pivot Table**
   * You’ll see the **PivotTable Field List** on the right side, where you can drag and drop fields into different areas (Rows, Columns, Values, Filters).
   * **Rows**: This is where you group your data. For example, dragging **Product Category** into Rows will group your data by product category.
   * **Values**: This is where you put the numerical data you want to summarize, like Sales, Quantity, or Profit. You can change the aggregation method (Sum, Average, Count, etc.) by clicking on the drop-down arrow next to the field name.
   * **Columns**: Use this area to break your data down into more detailed subcategories, such as by Region or Ship Mode.
   * **Filters**: This area lets you filter your pivot table based on specific criteria, like a particular Product Category or Date range.

**Example: Analyzing Total Sales by Region and Product Category**

1. **Drag** Region to the **Rows** area.
2. **Drag** Product Category to the **Columns** area.
3. **Drag** Sales to the **Values** area.
4. Excel will automatically summarize the total sales for each region and product category, making it easy to identify trends or areas of high performance.

**Example: Counting the Number of Orders by Customer Segment**

1. **Drag** Customer Segment to the **Rows** area.
2. **Drag** Order ID to the **Values** area.
3. Change the aggregation from **Sum** to **Count** by clicking the drop-down arrow next to Order ID and selecting **Value Field Settings > Count**.
4. This will give you the total number of orders placed by each customer segment.

**Sorting and Filtering a Pivot Table**

* You can easily sort the data by selecting a column and choosing **Sort A to Z** or **Sort Z to A**.
* You can filter data within the pivot table using the dropdown arrows that appear next to the Row or Column labels.

**Refreshing a Pivot Table**

* If you update the original data source, your pivot table won’t automatically update. You’ll need to refresh it by clicking on the **PivotTable Analyze** tab and selecting **Refresh**.

**Charts**

Charts are essential for visualizing data and making insights easier to understand. Excel offers a variety of chart types to help you represent data visually.

**Creating a Basic Chart**

1. **Select Your Data**
   * Highlight the data you want to visualize, including headers (for labels).
   * Go to the **Insert** tab and choose the chart type you want from the **Charts** group.
2. **Choosing the Right Chart Type**
   * **Column/Bar Chart**: Ideal for comparing categories (e.g., Total Sales by Product Category).
   * **Line Chart**: Best for showing trends over time (e.g., Profit over several months).
   * **Pie Chart**: Useful for showing proportions of a whole (e.g., Market share of different regions).
   * **Scatter Plot**: Good for displaying correlations between two variables (e.g., Sales vs. Profit).
   * **Combo Chart**: Combine two chart types, such as a line chart and a column chart, to show different aspects of your data together.

**Example: Creating a Column Chart to Show Sales by Region**

1. **Select** the data for Region and Sales from your dataset.
2. Go to **Insert** and click on **Column Chart** under the **Charts** group.
3. Excel will generate a column chart showing sales for each region.
4. Customize the chart using the **Chart Tools** (Design and Format tabs), where you can adjust colors, labels, and titles.

**Customizing Your Chart**

* **Chart Title**: Add or modify the title of your chart by clicking on it and typing a new name.
* **Axis Titles**: Add axis titles by selecting the chart and going to the **Chart Elements** (the plus sign next to the chart), then check the **Axis Titles** box.
* **Legend**: You can turn the legend on or off, or move it by selecting **Legend** under **Chart Elements**.
* **Data Labels**: If you want to display the values directly on the chart, check the **Data Labels** box.

**Example: Creating a Line Chart to Show Monthly Sales Trends**

1. Select the Order Date and Sales columns.
2. Go to **Insert** and choose a **Line Chart**.
3. Excel will create a line chart that shows the trend of sales over time.
4. You can format the chart by changing the line colors, adding markers, or customizing the time intervals on the x-axis using the **Chart Tools** tab.

**Combo Charts**

* **Combo charts** allow you to plot different types of data together (e.g., Sales and Profit).

1. Select your data (e.g., Sales and Profit).
2. Go to the **Insert** tab, click **Combo Chart**, and choose the desired combo (e.g., Column for Sales, Line for Profit).
3. This chart helps in comparing two related metrics in a single visual.

**Formatting Tips**

* **Change Chart Colors**: In the **Design** tab, you can choose from Excel’s predefined color themes or customize colors.
* **Change Chart Type**: If you feel the chart type isn’t right for your data, you can switch it by clicking **Change Chart Type** in the **Design** tab.
* **Chart Style**: Apply preset styles using the **Chart Styles** gallery under the **Design** tab to make your chart more visually appealing.

**Example: Creating a Pie Chart for Market Share by Region**

1. Select the Region and Sales data.
2. Go to **Insert** and choose **Pie Chart**.
3. Excel will generate a pie chart showing how much each region contributes to total sales.
4. Customize it by adding **Data Labels** to show percentages.

**Dynamic Charts with Pivot Tables**

* You can create **Pivot Charts** by adding a chart to a pivot table. These charts automatically update when you filter or modify the pivot table.

1. Create a **Pivot Table**.
2. Go to the **PivotTable Analyze** tab and click **PivotChart**.
3. Choose a chart type, and it will dynamically update as you change the pivot table.