Excel

Created	@November 14, 2024 4:47 PM
Tags	

Chapter 1: Data Cleaning in Excel

Remove Duplicates and Blank Rows

- Removing Duplicates:
    1. Go to Data > Remove Duplicates.
    2. Select the columns where you want to remove duplicates, then click OK.
- Removing Blank Rows:
    1. Go to Home > Find & Select > Go to Special > Blanks.
    2. Select the blank rows, right-click, and choose Delete > Delete Entire Row.

Handling Null Values

- Find Null Values:
    1. Go to Home > Find & Select > Find.
    2. Type 0 or any placeholder representing null values, then press Find All.
    3. Select the rows and right-click to delete.
- Replace Null Values:
    1. Go to Home > Find & Select > Find.
    2. Replace 0 with an empty space (just press Replace All).

Highlighting Duplicates and Empty Cells

- Highlight Duplicates:
    1. Go to Home > Conditional Formatting > Highlight Cell Rules > Duplicate Values.
- Highlight Empty Cells:

1. Go to Home > Conditional Formatting > Highlight Cell Rules > Text That Contains.
2. Select Blanks and choose a format (e.g., Green fill with dark green text).

Chapter 2: Excel Functions for Data Manipulation

Data Parsing from Text to Columns

- Text to Columns:
  1. Add a blank column to the right of the column you want to parse.
  2. Highlight the column, go to Data > Text to Columns.
  3. Select Delimited and choose Comma.
  4. Click Finish.

Removing Extra Spaces

- Using TRIM Function:
  1. Insert new columns and use =LEN(I2) to calculate the length of text in column.
  2. Use =TRIM(I4) to remove extra spaces and copy down the column.

Changing Text Cases

- Upper Case:Use =UPPER(cell address).
- Lower Case:Use =LOWER(cell address).
- Sentence Case:Use =PROPER(cell address).

Spell Check

- Spell Check:Select all text-based columns, go to Review > Spelling.

Chapter 3: Data Management and Formatting

Removing Format

- Clear Formats:
  1. Select the column, go to Home > Clear > Clear Formats.

Sorting Data

Sorting data in Excel is a crucial process that helps you arrange your data in a specific order, making it easier to analyze and visualize.

- Sort Data:

  Steps to Sort Data:

  1. Select All Data:
     - Select the entire data range (all columns and rows you wish to sort).
     - You can do this quickly by pressing Ctrl + A.
  2. Go to the Sort Function:
     - Navigate to the Home tab in the ribbon.
     - Click on the Sort & Filter button in the toolbar.
  3. Select Custom Sort:
     - Choose Custom Sort from the drop-down menu.
     - In the Sort dialog box that appears, check the box that says "My data has headers," ensuring the first row is treated as a header.
  4. Choose Column for Sorting:
     - In the Sort by drop-down list, select the column by which you want to sort.
     - Choose whether to sort Ascending (smallest to largest) or Descending (largest to smallest).
  5. Add Additional Sorting Levels:
     - If you want to sort by multiple columns, click on the Add Level button.
     - Select additional columns (such as sub-categories) in the next level and specify whether to sort them in Ascending or Descending order.
  6. Apply the Sort:
     - Click OK to apply the sort.
  7. Clear Sorting:

- To remove the sort, select the column you're sorting by and click on Sort & Filter again. Then choose Clearto revert to the original order.

Filter Data

- Apply Filters:
    1. Use Ctrl+A to select all data.
    2. Go to Home > Format as Table, then select My table has headers.
    3. You can filter by clicking the drop-down arrows at the top of each column.

Freeze Top Row

- Freeze Top Row:
    1. Go to View > Freeze Panes > Freeze Top Row.

Hide/Unhide Columns

- Hide Columns:
    1. Select the columns to hide, right-click, and choose Hide.
    2. To unhide, select adjacent columns, right-click, and choose Unhide.

Chapter 4: Conditional Formatting in Excel

Conditional formatting in Excel allows you to automatically format cells based on their values, helping you to visualise trends and patterns in your data more easily.

How to Apply Conditional Formatting:

1. Select the Data Range:
    - Highlight the columns or rows where you want to apply conditional formatting.
2. Access Conditional Formatting:
    - Go to the Home tab.
    - In the Styles group, click on Conditional Formatting.
3. Choose Formatting Style:
    - There are several types of conditional formatting you can apply:

    - Highlight Cell Rules: This can be used to highlight cells that meet specific conditions, such as values that are greater than a certain number or duplicate values.
    - Top/Bottom Rules: Highlights the top or bottom values based on criteria (e.g., top 10% or bottom 10).
    - Data Bars: Adds horizontal bars to cells, similar to bar graphs, to show data magnitude visually.
    - Color Scales: Applies color gradients to cells, where higher values might be green and lower values red, for instance.
    - Icon Sets: Displays icons (like arrows or circles) based on cell values, making it easy to see trends.
4. Apply and Customize Formatting:
    - After selecting your desired rule, you can customize the formatting options by choosing specific thresholds or color scales.
5. Clear Conditional Formatting:
    - If you want to remove conditional formatting, go back to the Conditional Formatting dropdown, and select Clear Rules. You can remove the rules from the entire sheet or specific columns.

Chapter 5: **VLOOKUP and Combining Datasets**

VLOOKUP is a powerful function used to combine two datasets by finding matching values in one column and returning related data from another column.

How to Use VLOOKUP:

1. **Select the Data Range**:
    - Ensure the datasets you want to combine have one common column (for example, ID numbers, product names, etc.).
2. **Write the VLOOKUP Formula**:
    - The syntax for VLOOKUP is:

```
=VLOOKUP(lookup_value, table_array, col_index_num, [range_lookup])
```

- **lookup_value**: The value you are searching for.
- **table_array**: The range of cells containing the data (usually from another sheet).
- **col_index_num**: The column number from which to retrieve the data.
- **[range_lookup]**: FALSE for exact match, TRUE for approximate match.

3. **Example**:
   - Let's say you have two datasets: one with Product IDs and Product Names, and another with Product IDs and Prices. To add the price information to the first dataset, you would use the following formula:

```
=VLOOKUP(A2, Product_Price_Table, 2, FALSE)
```

   - This looks for the value in cell A2 (Product ID) in the second dataset and returns the corresponding price from the second column.
4. **Fixing References with Dollar Signs**:
   - Use the `$` symbol to make cell references absolute (e.g., `$A$1`), so that when you copy the formula down, the references do not change.

Chapter 6: Formulas and Functions

1. Basic Functions

- **Concatenate**: `=CONCAT(A2, ": ", N2, " ", L2, " - ", TEXT(B2, "mm/dd/yyyy"))`

Combining data from one or more columns into a single column. The result of above example will be, “000261695: 4 Mountain Bikes – 12/01/2021”

- **COUNT, COUNTBLANK, COUNTA**:

- `COUNT`: Counts the number of cells in the selected range.
- `COUNTBLANK`: Counts number of empty cells within a specified range.
- `COUNTA`: Counts the number of non-empty cells within a range or list of arguments. This includes cells with text, numbers, or any other type of data.

2. Conditional Functions

- **COUNTIF**: `=COUNTIF(D1:D10, ">5")`

Counts the number of cells within a range that meet a specific condition (criteria).

- **SUMIF**: `=SUMIF(A1:A10, ">5", B1:B10)`

Creates a sum based on criteria.

3. Text and Date Functions

- **LEFT, RIGHT, MID**:Extract specific characters from a string. Example: `=LEFT(A2, 3)` extracts the first three characters from cell A2.
- **MINIF, MAXIF, AVERAGEIF**: These functions return minimum, maximum, and average values based on criteria.

Note: All the functions below are useful for data cleaning.

#NAME	the cell range name in the formula is not defined
#N/A	the cell contains no data
#NULL	the cell range cannot be understood
#REF	the cell or the range of cells referred to by the formula does not exist
#NUM	an argument used in the formula is invalid
#DIV/0!	an attempt was made to divide the value of the cell by zero
#VALUE	the formula has an incorrect/invalid function

Chapter 7: Pivot Tables and Charts

1. Pivot Table

- **Create Pivot Table**:
  1. Go to **Insert** > **PivotTable**.

2. Select your data range and click **OK**.
3. Drag fields to the Rows, Columns, and Values areas to summarize data.

2. Pivot Chart

- **Create Pivot Chart**:
  1. After creating the pivot table, go to **Insert** > **PivotChart** to add a visual representation of your pivot table data.

Chapter 8: **Descriptive Statistics and Data Analysis**

Descriptive Statistics in Excel help you summarize and describe the important features of a dataset. It involves basic measures such as the mean, median, mode, standard deviation, and range.

How to Perform Descriptive Statistics in Excel:

1. **Use the Data Analysis Toolpak**:
   - First, ensure the **Data Analysis Toolpak** is enabled.
   - Go to **File > Options > Add-ins**.
   - Select **Excel Add-ins** and click **Go**. Then, check the box next to **Analysis ToolPak** and click **OK**.
2. **Calculate Descriptive Statistics**:
   - Go to the **Data** tab.
   - In the **Analysis** group, click **Data Analysis**.
   - Select **Descriptive Statistics** and click **OK**.
3. **Input Data**:
   - Select the range of data you want to analyze.
   - Check the box for **Summary Statistics**.
4. **Interpret Results**:
   - Excel will output the following descriptive statistics:
     - **Mean**: The average of the data.

- **Median**: The middle value of the dataset.
- **Mode**: The most frequent value.
- **Standard Deviation**: Measures the spread of data.
- **Range**: The difference between the largest and smallest values.
- **Kurtosis**: The "tailedness" of the data distribution.
- **Skewness**: The asymmetry of the data distribution.

Correlation Analysis

- **Pearson Correlation**: Measures the linear relationship between two variables.
  - **Formula**: `=CORREL(range1, range2)`
  - **Interpretation**: Values range from -1 (perfect negative correlation) to +1 (perfect positive correlation). A value close to 0 indicates no correlation.

Chapter 9: Regression Analysis

Regression Analysis:

- For more advanced analysis, use the **Regression** tool found under **Data Analysis**.
- Select **Regression** and input your dependent and independent variables.
- The output will include coefficients, p-values, R-squared values, and more, which help you understand relationships between variables.

1. **Coefficients**

- **What they are**: Coefficients represent the estimated change in the dependent variable for each unit change in an independent variable, assuming all other variables remain constant. These are the values that define the equation of the regression line (or hyperplane, in multiple regression).

2. **P-Values**

- **What they are**: The p-value helps to assess the statistical significance of each coefficient. It tests the null hypothesis that a particular coefficient is equal to zero (no effect).

- **Interpretation**:
  - **Small p-value (typically < 0.05)**: Indicates strong evidence against the null hypothesis, suggesting that the independent variable is likely to have a significant impact on the dependent variable.
  - **Large p-value (typically > 0.05)**: Suggests weak evidence against the null hypothesis, implying that the independent variable may not significantly affect the dependent variable.
  - **Example**: If the p-value for "GPA" is 0.02, this suggests that GPA has a statistically significant relationship with the dependent variable at a 5% significance level.

3. **R-Squared (R^2)**

- **What it is**: R-squared is a statistical measure that represents the proportion of the variance in the dependent variable that is explained by the independent variables in the model.
- **Interpretation**:
  - **Value range**: R-squared ranges from 0 to 1.
  - **High R-squared value (close to 1)**: Means that the model explains a large proportion of the variance in the dependent variable. This indicates a good fit of the data to the regression model.
  - **Low R-squared value (close to 0)**: Means that the model does not explain much of the variability in the dependent variable. This indicates a poor fit.
  - **Example**: An R-squared value of 0.85 means that 85% of the variance in the dependent variable can be explained by the model, while the remaining 15% is unexplained (due to other factors or random variation).

4. **Adjusted R-Squared**

- **What it is**: Adjusted R-squared adjusts R-squared for the number of predictors in the model. Unlike R-squared, which always increases with more variables (even if they don't improve the model), adjusted R-squared penalizes the inclusion of irrelevant variables.

- **Interpretation**:
  - **Higher adjusted R-squared** indicates a better model fit when accounting for the number of predictors.
  - **Example**: If adjusted R-squared is 0.75, it means that 75% of the variance in the dependent variable is explained, accounting for the number of predictors used in the model.

5. **Standard Error of the Estimate (SE)**

- **What it is**: This measures the accuracy of the predictions made by the regression model. It represents the standard deviation of the residuals (the differences between the observed values and the predicted values).
- **Interpretation**:
  - A lower standard error means that the model's predictions are closer to the actual values.
  - **Example**: If the standard error is 2, it means the predicted values of the dependent variable will typically be within 2 units of the actual observed values.

6. **Confidence Intervals for the Coefficients**

- **What they are**: Confidence intervals provide a range of values within which the true population coefficient is likely to fall, with a specified level of confidence (usually 95%).
- **Interpretation**:
  - A 95% confidence interval means you can be 95% confident that the true coefficient lies within this range.
  - **Example**: If the confidence interval for a coefficient is (0.2, 0.8), it suggests that the true effect of the independent variable on the dependent variable is between 0.2 and 0.8, with 95% certainty.

7. **F-Statistic and F-Test**

- **What it is**: The F-statistic tests the overall significance of the regression model. It compares the fit of your model with a model that has no predictors (just the mean of the dependent variable).

- **Interpretation**:
  - A **large F-statistic** and a **small p-value** suggest that the model provides a better fit than a model with no predictors (i.e., your independent variables are collectively significant).
  - **Example**: If the F-statistic is high and the p-value is low (e.g., p-value < 0.05), it suggests that at least one of the predictors is statistically significant in explaining the dependent variable.

8. **T-Statistic and T-Test for Each Coefficient**

- **What it is**: The t-statistic tests whether a specific coefficient is significantly different from zero. It is calculated as the ratio of the estimated coefficient to its standard error.
- **Interpretation**:
  - A **larger absolute t-statistic** indicates stronger evidence that the coefficient is significantly different from zero.
  - **Example**: If the t-statistic for a coefficient is 3.5 and the p-value is 0.01, you can conclude that the coefficient is statistically significant at the 1% level.