atomcamp

Excel Masterclass 1: Manual

Data Cleaning & Introduction to Pivot Tables

This manual provides an overview and introduction of each topic that will be covered in the Excel masterclass on data cleaning & pivot tables. Please ensure you read through it to set a foundation for the class.

Introduction

- Overview of the Data Life Cycle
  - Data Collection: This initial stage involves gathering raw data from various sources. The quality and accuracy of data collected at this stage directly impact the effectiveness of subsequent stages.
  - Data Cleaning and Preparation: Critical for ensuring the reliability of the dataset, this stage involves removing errors, inconsistencies, and irrelevant data. Proper data cleaning enhances the accuracy of analysis and decision-making processes.
  - Data Analysis: At this stage, the cleaned data is examined to extract meaningful insights. This involves looking for patterns, correlations, and trends that can inform business decisions or scientific conclusions.
  - Data Visualization: Here, data is transformed into graphical representations like charts and graphs. Visualization makes it easier to understand complex data sets and communicate findings clearly.
  - Data Interpretation: This involves making sense of the data and its visualizations to draw actionable conclusions. It's about understanding the 'why' and 'how' behind the data.
  - Data Storage and Maintenance: Finally, data needs to be securely stored and maintained. This ensures its availability for future analysis and ensures that the data remains accurate and up-to-date.

Part 1: Data Cleaning in Excel
- Dataset Structuring
  - Proper dataset structuring involves organizing data in a format that is both logical and efficient for analysis. This includes defining columns clearly, ensuring each row represents a single record, and maintaining consistent formatting throughout the dataset.
- Data Shapes: Wide and Long Formats
  - Wide format datasets spread related data across multiple columns, with each column representing a different variable. This format is useful for data comparison purposes.
  - Long format datasets, on the other hand, stack data vertically, often consolidating multiple variables into a single column with corresponding values in another. This format is efficient for handling large datasets with repetitive measures.
- Unique/Primary Key Concepts
  - A unique or primary key is a specific piece of data in a column that uniquely identifies each row in a table. It's crucial for relational databases and for ensuring that each record in a dataset is distinct from all others, which is important for accurate data analysis and data integrity.
- Handling Duplicates
  - Managing duplicates involves identifying and removing or consolidating repeated entries in a dataset. This process is vital to prevent skewed analysis results and to ensure that each piece of data is represented accurately.
- Sort and Filter Functions
  - Sorting and filtering are fundamental functions in Excel used to organize data. Sorting rearranges the data based on specified criteria (like alphabetical order), while filtering allows for the display of only those rows that meet certain conditions, thereby facilitating focused analysis.

- Text Functions: LEFT, RIGHT, MID
  - The LEFT, RIGHT, and MID functions in Excel are used to extract specific segments of text from a cell. LEFT returns the first characters from the start of a string, RIGHT returns the last characters, and MID extracts a substring from the middle based on a specified position and length.
- The FIND Function
  - The FIND function in Excel is used to locate the position of a specified string within another string. This is particularly useful for parsing complex text data, allowing users to extract and analyze specific portions of data within a cell.
- The IF Function
  - The IF function allows for logical comparisons within Excel. It returns one value if a specified condition is true and another value if it's false. This function is integral for performing conditional analysis and decision-making within datasets.
- Text to Columns and CONCAT
  - The 'Text to Columns' feature in Excel is used to split text from a single cell into multiple columns based on a specified delimiter. The CONCAT function (or CONCATENATE in earlier versions) is used to combine text from different cells into one.
- Understanding Data Types
  - Excel supports various data types like numeric, text, date, and Boolean. Understanding and correctly using these data types is crucial for accurate data entry and analysis. It ensures that functions and formulas work correctly and that data is interpreted in the intended way.

Part 2: Introduction to Pivot Tables

- What are Pivot Tables?
  - Pivot Tables are one of Excel's most powerful features, used for summarizing, analyzing, exploring, and presenting data. They allow users to easily transform columns of data into a more readable and understandable format, often without using formulas.
- Creating Your First Pivot Table
  - Creating a pivot table involves selecting a range of data and choosing how to 'pivot' or rearrange this data. It typically involves specifying rows, columns, values, and filters to display the data in a summarized and organized manner.
- Basic Pivot Table Operations
  - Basic operations in pivot tables include sorting, filtering, and arranging fields. This enables users to explore different aspects of the data, highlight key information, and conduct a variety of summary calculations like sums and averages.
- Pivot Table Layout and Design
  - The layout and design of a pivot table can greatly affect its readability and impact. Excel offers various options to customize the appearance, such as adjusting field arrangements, applying styles, and formatting values for clearer presentation.