**Benefits and Limitations of Using Spreadsheets for Data Analysis**

**Benefits**:

1. **User-Friendly Interface**: Spreadsheets provide an intuitive way to organize data in a table format, making it easy to view and analyze the top results, such as the first 100 rows.
2. **Data Comparison**: Functions like VLOOKUP and XLOOKUP allow for quick comparisons between fields, facilitating simple data matching tasks.
3. **SQL Table Emulation**: Custom formulas can simulate SQL table functionality within Excel, enabling relational data management on a small scale.
4. **Versatile Tools**: Features like pivot tables, charts, and formulas support basic data summarization, visualization, and calculation tasks.
5. **Ideal for Small Datasets**: Spreadsheets handle datasets of up to 20,000 records effectively, making them a great choice for light, straightforward analyses.
6. **Accessibility**: Widely available and easy to learn, spreadsheets are accessible for beginners and professionals alike.

**Limitations**:

1. **Lack of Reproducibility**: Actions like filtering, sorting, or applying formulas lack a built-in log, making it difficult to retrace steps or reproduce analysis accurately.
2. **Performance Issues**: Spreadsheets struggle with large datasets; handling records exceeding 100,000 rows may result in crashes or slow performance.
3. **Limited Scalability**: They are not suitable for complex analyses, such as advanced statistical modeling or machine learning, due to limited functionality and computational power.
4. **Error-Prone**: Manual data entry and formula adjustments increase the risk of errors, especially for large or intricate datasets.
5. **Collaboration Challenges**: Multi-user editing in real-time can be cumbersome, leading to version control issues unless using advanced tools like Google Sheets.
6. **Automation Constraints**: Spreadsheets lack the automation and repeatability offered by programming languages like Python or R, limiting efficiency in iterative tasks.
7. **Data Integration**: Combining data from multiple sources is possible but can become tedious without tools like Power Query or dedicated software.

**2. When should we use the Waterfall and funnel chart?**

**Waterfall-**

- Shows how values increase/decrease over time

Used in: -

Budget Tracking

**Funnel: -**

Shows drop-off across process stages

Used in: -

Customer Journey

* 1. **When should we use donut chart and pie chart?**

**When to Use Pie Chart: -**

To show the relative proportion of different categories in a dataset.

When there are few categories (ideally less than 6).

When exact percentage comparisons are important.

When labels and values are easy to distinguish.

**Donut Chart**

**When to Use?**

When you want a modern, cleaner look with a center space.

To show proportions while allowing space in the center for extra labels or text.

When comparing multiple datasets (e.g., side-by-side donut charts).

For dashboard visualizations (less cluttered than pie charts).

**1. General Shortcuts**

| **Shortcut** | **Function** |
| --- | --- |
| Ctrl + N | Create a new workbook |
| Ctrl + O | Open an existing workbook |
| Ctrl + S | Save the workbook |
| Ctrl + P | Print the worksheet |
| Ctrl + Z | Undo last action |
| Ctrl + Y | Redo last undone action |
| Ctrl + C | Copy selected cells |
| Ctrl + X | Cut selected cells |
| Ctrl + V | Paste copied/cut cells |
| Ctrl + F | Find in the worksheet |
| Ctrl + H | Find and replace |

**2. Navigation Shortcuts**

| **Shortcut** | **Function** |
| --- | --- |
| Arrow Keys | Move between cells |
| Ctrl + Arrow Key | Jump to the edge of the data range |
| Ctrl + Home | Go to the **A1** cell |
| Ctrl + End | Go to the last used cell |
| Page Up / Page Down | Scroll up/down one screen |
| Alt + Tab | Switch between open workbooks |

**3. Cell Editing Shortcuts**

| **Shortcut** | **Function** |
| --- | --- |
| F2 | Edit selected cell |
| Alt + Enter | Insert a line break inside a cell |
| Ctrl + D | Copy value from the cell above |
| Ctrl + R | Copy value from the left cell |
| Ctrl + Shift + "+" | Insert a new row or column |
| Ctrl + "-" | Delete a row or column |

**4. Formatting Shortcuts**

| **Shortcut** | **Function** |
| --- | --- |
| Ctrl + B | Bold text |
| Ctrl + I | Italic text |
| Ctrl + U | Underline text |
| Ctrl + 1 | Open Format Cells dialog box |
| Ctrl + Shift + $ | Apply currency format |
| Ctrl + Shift + % | Apply percentage format |
| Ctrl + Shift + # | Apply date format |

**5. Selecting Data**

| **Shortcut** | **Function** |
| --- | --- |
| Ctrl + A | Select the entire worksheet |
| Ctrl + Shift + Arrow Key | Select data in a direction |
| Ctrl + Space | Select entire column |
| Shift + Space | Select entire row |

**6. Working with Formulas**

| **Shortcut** | **Function** |
| --- | --- |
| = (Equals sign) | Start a formula |
| Alt + = | AutoSum (Σ) selected cells |
| Ctrl + Shift + Enter | Enter an array formula |
| F4 | Repeat last action / Toggle absolute & relative references in formulas |

**7. Pivot Table & Charts**

| **Shortcut** | **Function** |
| --- | --- |
| Alt + N + V | Insert a Pivot Table |
| Alt + F1 | Create a chart from selected data |
| F11 | Create a chart in a new sheet |

**What is GitHub?**

GitHub is a **web-based platform** for **version control, collaboration, and code hosting**. It is built on **Git**, a distributed version control system that helps developers manage code efficiently.

**Why Should We Use GitHub?**

🔹 **1. Version Control**

* Tracks changes to your code over time.
* Allows you to revert to previous versions if needed.
* Helps in debugging by seeing historical changes.

**What is Power BI?**

Power BI is a **business intelligence (BI) tool** developed by Microsoft that helps users **analyze data, create reports, and generate interactive visualizations**. It enables users to connect to various data sources, transform raw data, and gain insights through dashboards and reports.

**Why Should We Use Power BI?**

**🔹 1. Data Connectivity**

* Connects to **Excel, SQL Server, Oracle, AWS, Azure, Google Analytics, APIs, and more**.
* Supports **real-time data** updates.

**🔹 2. Data Transformation & Modeling**

* **Power Query** helps in **data cleaning, merging, and transformation**.
* **DAX (Data Analysis Expressions)** allows advanced calculations and custom measures.

**🔹 3. Interactive Dashboards & Reports**

* Drag-and-drop interface for creating **dynamic charts, maps, and visuals**.
* Filters, slicers, and drill-through capabilities for deep insights.

**🔹 4. Automation & Scheduling**

* Automate report refreshes and scheduling.
* No manual data updates required.

**🔹 5. Cloud & On-Premise Access**

* Publish reports on **Power BI Cloud Service** for web access.
* Use **Power BI Desktop** for offline report building.
* Integrate with **Power BI Mobile App** for on-the-go analytics.

**🔹 6. AI-Powered Insights**

* Uses **AI-driven analytics** for trend detection and forecasting.
* Includes **natural language query (Q&A)** for quick insights.

**🔹 7. Collaboration & Sharing**

* Share reports securely across teams.
* Integrates with **Microsoft Teams, Excel, and SharePoint**.

**Who Uses Power BI?**

✅ **Data Analysts** – To analyze trends and patterns.  
✅ **Business Executives** – To track KPIs and company performance.  
✅ **Finance Teams** – To monitor budgets and financial reports.  
✅ **Sales & Marketing Teams** – To track customer insights and revenue trends

**What is Python?**

Python is a **high-level, interpreted programming language** known for its **simplicity, readability, and flexibility**. It is widely used in **web development, automation, AI, machine learning, and data analysis**.

**Why Use Python for Data Analysis?**

**🔹 1. Easy to Learn & Use**

* Simple **syntax**, making it beginner-friendly.
* Readable and clean code for data manipulation.

**🔹 2. Powerful Libraries for Data Analysis**

Python has **rich libraries** for handling, processing, and visualizing data:

| **Library** | **Purpose** |
| --- | --- |
| **pandas** | Data manipulation & analysis (like Excel for Python) |
| **NumPy** | Numerical computing (arrays, matrices, math functions) |
| **Matplotlib & Seaborn** | Data visualization (charts, graphs, heatmaps) |
| **Scikit-learn** | Machine learning & predictive analytics |
| **Statsmodels** | Statistical analysis & hypothesis testing |

**🔹 3. Handles Large Datasets Efficiently**

* Can process **millions of rows** (faster than Excel).
* Works well with **structured and unstructured data** (CSV, JSON, databases, etc.).

**🔹 4. Integration with Databases & Big Data**

* Connects with **SQL, MongoDB, Hadoop, Spark** for large-scale data processing.

**🔹 5. Automation & Reproducibility**

* Automate repetitive tasks like **data cleaning, transformation, and reporting**.
* Scripts can be reused and scheduled for **automated reporting**.

**🔹 6. Machine Learning & AI Capabilities**

* Python integrates with **TensorFlow, PyTorch, and Scikit-learn** for predictive modeling.
* Used in **data science, AI, and deep learning applications**.

**Who Uses Python for Data Analysis?**

✅ **Data Analysts & Scientists** – For exploring and visualizing data.  
✅ **Business Intelligence Teams** – To extract insights from large datasets.  
✅ **Financial Analysts** – To model financial forecasts.  
✅ **Engineers & Researchers** – To analyze experimental data.

**What is SQL?**

SQL (**Structured Query Language**) is a programming language used to **manage and query relational databases**. It allows users to **store, retrieve, manipulate, and analyze data** efficiently.

**Why Should We Use SQL?**

**🔹 1. Efficient Data Retrieval**

* Quickly fetch specific data using **SELECT queries**.
* Filter and sort data using **WHERE, ORDER BY, and GROUP BY**.

**🔹 2. Handles Large Datasets**

* Can manage **millions of records** efficiently (faster than Excel or CSV files).
* Optimized for structured data storage and retrieval.
* **Who Uses SQL?**
* ✅ **Data Analysts & Scientists** – To extract insights from databases.  
  ✅ **Software Engineers** – To store and manage application data.  
  ✅ **Business Intelligence Teams** – To generate reports for decision-making.  
  ✅ **Finance & Marketing Teams** – To track revenue, customer trends, and transactions.

**When to Use Python Over SQL?**

Python and SQL both play key roles in **data analysis and engineering**, but they serve different purposes. **SQL is best for querying and managing structured data**, while **Python is more powerful for complex processing, automation, and machine learning**.

**✅ Use Python When:**

**🔹 1. Advanced Data Manipulation & Cleaning**

* SQL is great for basic filtering, but Python (with **pandas**) is better for:  
  ✅ Handling missing values (**fillna, dropna**)  
  ✅ Complex transformations (e.g., regex-based string cleaning)  
  ✅ Handling **JSON, XML, or nested data** that SQL struggles with

**Example:**  
✅ Removing outliers from a dataset → Easier in Python than SQL

**🔹 2. Statistical Analysis & Machine Learning**

* SQL can **aggregate and summarize** data, but it can't do:  
  ✅ **Regression analysis** (e.g., linear regression)  
  ✅ **Predictive modeling** (e.g., machine learning)  
  ✅ **Time series forecasting**

**Example:**  
✅ Predicting sales trends based on historical data → Best in Python (with **scikit-learn**)

**🔹 3. Data Visualization**

* SQL **retrieves data**, but **it doesn’t visualize it**.
* Python (with **Matplotlib, Seaborn, Plotly**) helps create:  
  ✅ Interactive charts  
  ✅ Heatmaps & correlations  
  ✅ Geospatial maps

**Example:**  
✅ Creating a customer retention dashboard → Easier in Python than SQL

**🔹 4. Automation & Data Pipelines**

* SQL runs **queries manually** (or via stored procedures), but Python can:  
  ✅ Automate data extraction & transformations (**ETL pipelines**)  
  ✅ Schedule tasks (**with cron jobs, Apache Airflow**)  
  ✅ Handle API calls & real-time data ingestion

**Example:**  
✅ Automating daily stock market data updates → Best in Python

**🔹 5. Working with Multiple Data Sources**

* SQL works **only within relational databases**, but Python can:  
  ✅ Merge **CSV, Excel, JSON, APIs, and NoSQL (MongoDB, Firebase, etc.)**  
  ✅ Perform complex joins across **multiple data sources**

**Example:**  
✅ Merging customer data from SQL + product data from a NoSQL database → Best in Python

**✅ Use SQL When:**

🔹 Simple data retrieval (e.g., “Give me the top 10 customers”)  
🔹 Filtering, sorting, and grouping large datasets quickly  
🔹 Joining multiple tables efficiently  
🔹 Handling structured, relational data

**🚀 When to Use Both?**

1️⃣ **SQL** → Extract & filter data from the database  
2️⃣ **Python** → Perform advanced analysis & visualization

**Week 1**

**Basic Excel Questions**

**1. What are the most commonly used Excel functions for data analysis?**  
**Answer:**

* VLOOKUP / XLOOKUP – Lookup values from another table.
* INDEX-MATCH – More flexible than VLOOKUP for searching data.
* SUMIFS / COUNTIFS – Aggregate data based on multiple conditions.
* IF, IFS, IFERROR – Conditional logic functions.
* TEXT functions – LEFT, RIGHT, MID, LEN, TRIM, CONCATENATE (or TEXTJOIN).
* DATE functions – TODAY(), EOMONTH(), YEAR(), TEXT(date, "dd-mmm-yyyy").
* Pivot Tables – Summarize and analyze large datasets quickly.

**Intermediate Excel Questions**

**2. What is the difference between a Pivot Table and a normal table?**  
**Answer:**  
A **Pivot Table** dynamically summarizes, filters, and aggregates data without modifying the raw dataset. A **normal table** stores data in a structured way but lacks built-in analysis tools.

**3. How would you remove duplicate values from a dataset?**  
**Answer:**  
Use **"Remove Duplicates"** under the **Data** tab or use the UNIQUE() function (available in newer Excel versions).

**4. Explain how Power Query is useful for data analysis.**  
**Answer:**  
Power Query helps in:

* **Extracting** data from various sources (CSV, databases, APIs).
* **Transforming** data (cleaning, merging, pivoting/unpivoting).
* **Loading** data into Excel or Power BI.
* Automating **repeatable** data-cleaning tasks.

**Advanced Excel Questions**

**5. What is the difference between absolute, relative, and mixed cell references?**  
**Answer:**

* **Relative (A1)** – Changes when copied to another cell.
* **Absolute ($A$1)** – Stays fixed when copied.
* **Mixed ($A1 or A$1)** – Partially fixed (column or row stays constant).

**6. How would you handle large datasets efficiently in Excel?**  
**Answer:**

* Use **Tables** (Ctrl + T) for structured data.
* Use **Pivot Tables** instead of complex formulas.
* Minimize volatile functions like INDIRECT() and OFFSET().
* Avoid using full-column references (A:A), use specific ranges (A1:A1000).
* Turn off automatic calculation (File → Options → Formulas → Manual Calculation).

**7. How do you perform a dynamic lookup in Excel?**  
**Answer:**  
Use XLOOKUP (modern Excel) or INDEX-MATCH. Example:

excel

CopyEdit

=INDEX(B2:B10, MATCH(1001, A2:A10, 0))

(Looks up **1001** in Column A and returns corresponding value from Column B.)

**Scenario-Based Questions**

**8. Suppose you have sales data in Excel. How would you find total revenue by region?**  
**Answer:**

* Use **Pivot Table**: Drag "Region" to **Rows** and "Revenue" to **Values (Sum)**.
* Use SUMIFS():

excel

CopyEdit

=SUMIFS(Revenue\_Column, Region\_Column, "NSW")

(Gets revenue only for **NSW** region.)

**9. How would you visualize trends in sales over time?**  
**Answer:**

* **Line Chart** for time-series trends.
* **Bar/Column Chart** for comparisons.
* **Conditional Formatting** to highlight trends (e.g., green for increasing, red for decreasing).

**10. How do you handle missing data in an Excel dataset?**  
**Answer:**

* Use IFERROR() or IFNA() to replace errors.
* Use =IF(A2="", "Missing Data", A2).
* Use Power Query to filter or replace missing values.

# DAY 3

**Data Cleaning & Preparation Questions**

1. **What data quality issues do you see in this portfolio dataset? How would you clean them?**
   * Example answer: "I noticed that the Sale Price column has a property sold for **$1**, which is likely an error or missing data. I would check if it’s a data entry mistake or an outlier."
2. **How would you handle missing or inconsistent values in this portfolio dataset?**
   * Discuss techniques like **imputation, filtering, or flagging missing data**.
3. **How would you standardize suburb names in case they have inconsistent formats?**
   * Example: Convert "AUBURN NSW 2144" to just "AUBURN" using text functions or Power Query.
4. **How would you detect duplicate property records?**
   * Example approach: Identify duplicates based on Address, Sale Date, and Sale Price.

### **Descriptive Statistics**

1. **What are the main types of descriptive statistics?**
   * **Answer:** Measures of central tendency (**mean, median, mode**) and measures of dispersion (**range, variance, standard deviation**).
2. **What is the difference between mean, median, and mode? When would you use each?**
   * **Example answer:** The **mean** is the average, the **median** is the middle value, and the **mode** is the most frequent value.
   * **Use case:** Median is better when the data has outliers (e.g., property prices).
3. **What does standard deviation tell us about a dataset?**
   * **Answer:** It measures how much data varies from the mean. A high standard deviation means more variability.
4. **How would you identify outliers in a dataset?**
   * **Answer:**
     + **Using the IQR method**: An outlier is a value **outside Q1 - 1.5 × IQR or Q3 + 1.5 × IQR**.
     + **Using Z-score**: If Z-score > **3** or < **-3**, it’s an outlier.
5. **What is the difference between population and sample in statistics?**
   * **Answer:** A **population** includes all possible data points, while a **sample** is a subset of the population used for analysis.

**📌 Correlation & Regression**

1. **What is correlation, and how is it different from causation?**

* **Answer:** Correlation measures the strength and direction of a relationship between two variables, while causation means one variable directly affects another.

1. **What is the difference between Pearson and Spearman correlation?**

* **Pearson Correlation:** Measures **linear relationships** between variables.
* **Spearman Correlation:** Measures **monotonic relationships** (not necessarily linear).

1. **What is linear regression, and when would you use it?**

* **Answer:** Linear regression predicts a dependent variable using an independent variable (e.g., predicting house prices based on number of bedrooms).

1. **What is R-squared in regression analysis?**

* **Answer:** It represents the proportion of variance in the dependent variable explained by the independent variable(s). A higher **R²** means a better fit.

1. **What assumptions must hold for linear regression to work properly?**

* Linearity
* Independence of errors
* Normality of residuals
* Homoscedasticity (constant variance of errors)

**Week 4**

1. Difference between star and snowflake schema?

Refer <https://www.geeksforgeeks.org/difference-between-star-schema-and-snowflake-schema/> link

1. Practice from <https://www.examtopics.com/exams/microsoft/pl-300/view/> site for PL-300 certification.