Unit I: Introduction to Data Analytics

1. Define Data Analytics. State its importance in business decision-making

Ans:- **Data Analytics** is the comprehensive process of inspecting, cleansing, transforming, and modeling raw data to discover useful information, identify patterns, draw logical conclusions, and support effective decision-making

Importance of Data Analytics in Business Decision-Making:-

1. Improved Decision-Making :- Data Analytics provides business with evidence-based insights to make better-informed decisions, reducing reliance on intuition or guesswork

2. Improved Efficiency and productivity:- By identifying inefficiencies and bottlenecks in process, data analytics helps streamline operations and optimize resource allocation.

3. Enhanced Customer understanding:- Analyzing customer data allows business to understand customer preferences, behaviors, and needs, leading to personalized experiences and increased customer satisfactions

4. Competitive Advantages:- Businesses that effectively leverage data analytics gain a competitive edge by understanding market trends, identifying opportunities, and adapting quickly to changing conditions

2. List and explain any phases in the data analytics life cycle.

=> The Data Analytics Life Cycle

The data analytics life cycle is a systematic, **cyclical** process for guiding an analytics project from start to finish. It provides a structured framework to ensure that the analysis is thorough and the results are aligned with business goals.

Here are the six key phases:

1. Phase 1: Problem Definition (Discovery)

This is the **most critical** phase. Before you write any code or look at any data, you must understand the business problem.

- **What it is:** Working with stakeholders (like managers) to understand their exact goals.
- **Key Question:** "What business question are we trying to answer?" or "What is the Key Performance Indicator (KPI) we need to improve?"
- **Example:** The goal is not just "look at sales data." A proper problem definition is: "Identify the top 3 reasons why customer churn (cancellation) has increased by 15% in the last quarter."

2. Phase 2: Data Collection (Acquisition)

Once you know the problem, you need the raw materials: data.

- **What it is:** This is the process of gathering all the necessary data from various sources.

- **Sources:** This data can come from internal company databases (SQL), customer relationship management (CRM) systems, website log files, or even external third-party sources (like social media data).
- **Example:** For our churn problem, you would collect customer account details, usage history, support-call logs, and customer feedback forms.

3. Phase 3: Data Preparation (Preprocessing)

This is the most **time-consuming** phase. Raw data is almost always "dirty" (messy, incomplete).

- **What it is:** Cleaning and transforming the raw data into a clean, usable format for analysis.
- **Key Activities:**
    - **Cleaning:** Handling missing values (nulls), removing duplicate entries.
    - **Transformation:** Standardizing formats (e.g., changing "Male" and "M" to just "Male").
    - **Feature Engineering:** Creating new variables from existing ones (e.g., calculating "customer age" from their "date of birth").

4. Phase 4: Data Modeling & Analysis

This is the "brain" of the project where you start finding answers.

- **What it is:** Applying statistical techniques, algorithms, and models to the clean data to discover patterns, correlations, and insights.
- **Example:** You might use a **classification algorithm** (like Logistic Regression or a Decision Tree) to build a model that can *predict* which customers are at high risk of churning. This is where you move from "what happened" to "why it happened" and "what will happen."

5. Phase 5: Data Visualization & Communication

Insights are useless if they are not understood.

- **What it is:** Translating your complex technical findings into a simple, clear, and easy-to-understand story for the business stakeholders.
- **Tools:** Using tools like Tableau, Power BI, or even simple charts (bar graphs, line charts) in Excel or Python.
- **Example:** Instead of showing the complex math, you present a **dashboard** with a clear bar chart: "Customers who made more than 3 support calls in a month have an 80% higher churn rate."

6. Phase 6: Deployment & Monitoring

The project is not finished when the report is sent. The solution must be put into action.

- **What it is:** Implementing your model or findings into the company's day-to-day operations and continuously monitoring its performance.
- **Example: Deployment:** The churn prediction model is integrated into the company's CRM. Now, the sales team gets an *automatic alert* when a customer is flagged as "high churn risk."

- **Monitoring:** You check the model's accuracy every month. Is it still correctly predicting churn? If not, it may need to be retrained (which starts the cycle over).

3. Explain types of data analytics.

1. Descriptive Analytics

This is the most common and fundamental type of analysis. It is the starting point.

- **Key Question: "What happened?"**
- **Explanation:** This type of analytics summarizes raw data from the past to describe what has occurred. It provides a clear snapshot of the past. It does **not** explain *why* it happened.
- **Example:** A college principal's dashboard shows that student attendance in the first-year computer department was **78%** last month. This is a fact; it just describes the past.
- **Business Example:** A sales report showing total sales revenue for the last quarter.

---

2. Diagnostic Analytics

Once you know *what* happened, the next logical question is to find out *why*.

- **Key Question: "Why did it happen?"**
- **Explanation:** This is the "drill-down" or "root cause analysis" phase. It takes the descriptive data and investigates to find the underlying causes of a particular outcome.
- **Example:** The principal, seeing the 78% attendance, digs deeper. The diagnostic analysis might find a **correlation**: the attendance for a specific subject (e.g., Applied Mathematics) was only 40% on Fridays, which brought the total average down.
- **Business Example:** Investigating *why* sales were low and finding that a new competitor launched a 50% off sale in the same period.

---

3. Predictive Analytics

This is where we start looking into the future. It uses the patterns found in past data to make forecasts.

- **Key Question: "What is likely to happen?"**
- **Explanation:** This type uses statistical models and machine learning techniques on historical data to forecast future trends. It is about probabilities, not certainties.
- **Example:** Using the attendance data from the last 3 years, the college builds a model. The model **predicts** that attendance will likely drop by 15% in the week after the annual college festival.
- **Business Example:** A bank using a customer's credit history to predict the likelihood of them defaulting on a loan in the future.

4. Prescriptive Analytics

This is the most advanced and most valuable stage. It doesn't just predict the future; it tells you what to do about it.

- **Key Question: "What should we do?"**
- **Explanation:** This type of analytics goes beyond prediction by recommending a specific course of action. It uses optimization and simulation algorithms to suggest the best possible outcome.
- **Example:** The model not only predicts the attendance drop but also **recommends** the best solution: "To maintain 80% attendance, it is best to schedule an 'Industry Expert Lecture' on the Friday after the festival."
- **Business Example:** A logistics app (like Google Maps) not only *predicts* traffic but *prescribes* the fastest route for you to take right now.

4. Explain measures of central tendency.

A **measure of central tendency** is a single, summary value that attempts to describe the center of a dataset. It represents a "typical" or "central" value for a distribution. They provide a quick and simple understanding of the data.

The three primary measures of central tendency are:

1. Arithmetic Mean
2. Median
3. Mode

1. Arithmetic Mean (Average)

This is the most common measure of central tendency.

- **Explanation:** The Mean is calculated by summing all the values in a dataset and then dividing by the total number of values.[3]
- **Formula:** For a dataset $x_1, x_2, ..., x_n with 'n' values:

$$\text{Mean}(\bar{x}) = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + ... + x_n}{n}$$

When to Use: It is excellent for data that is normally distributed (symmetrical) and has no extreme values

2. Median

This is the "middle" value of the dataset.

- **Explanation:** The Median is the value that separates the higher half of the data from the lower half. To find it, you **must** first sort the data in ascending or descending order.
- **Method:**
    1. Sort the data.
    2. **If 'n' (number of values) is odd:** The median is the middle value.[6]
    3. **If 'n' is even:** The median is the average of the two middle values.[7]
- **Formula (for position):**
    - Odd 'n': Position = $\frac{n+1}{2}$
    - Even 'n': Positions = $\frac{n}{2}$ and $\frac{n+1}{2}$
- **When to Use:** This is the best measure to use when your data is **skewed** (asymmetrical) or when it **contains outliers** (e.g., salary data), because the median is not affected by extreme values.[10] We call it a **robust** statistic

3. Mode

This is the "most frequent" value.

- **Explanation:** The Mode is the value that appears most often in a dataset. A dataset can have:
    - **No Mode:** All values are unique.
    - **One Mode:** Unimodal (e.g., {1, 2, **3**, **3**, 4})
    - **Two Modes:** Bimodal (e.g., {1, **2**, **2**, **3**, **3**, 4})
    - **Multiple Modes:** Multimodal

5. Explain data types used in data analytics.

1. Categorical Data (Qualitative)

As the name suggests, this type of data represents categories or labels. It describes qualities or characteristics. You cannot perform meaningful mathematical operations like 'addition' or 'average' on it.

It is divided into two sub-types:

A. Nominal Data

- **Explanation:** This is the simplest form. The data consists of "names" or "labels." There is **no natural order** or ranking between the categories.
- **Operations:** You can only count them (frequency) or find the **Mode**.
- **Examples:**
    - **Gender:** (Male, Female, Other)

- **Pincode:** (400001, 411038) — *Be careful here!* Even though a pincode uses numbers, you cannot "average" them. It's just a label for a region.
- **Color of a Car:** (Red, Blue, White)

B. Ordinal Data

- **Explanation:** This data also consists of labels, but unlike nominal data, there is a **clear order or rank** between them. However, the *exact difference* (or interval) between the categories is not defined or is unequal.
- **Operations:** You can count, find the **Mode**, and also find the **Median** (the middle value).
- **Examples:**
  - **Customer Rating:** (Poor, Average, Good, Excellent) — We know Excellent is better than Good, but we don't know *how much* better.
  - **Education Level:** (SSC, HSC, Undergraduate, Postgraduate)
  - **Star Rating:** (1 Star, 2 Stars, 3 Stars)

---

2. Numerical Data (Quantitative)

This data represents measurable quantities or numbers. You **can** perform meaningful mathematical operations (add, subtract, average) on this data.

It is also divided into two sub-types:

A. Discrete Data

- **Explanation:** This data can only take specific, fixed values. It is often data that you **count**. You cannot have a value in between two fixed points.
- **Operations:** You can use all statistical measures (Mean, Median, Mode).
- **Examples:**
  - **Number of students in a class:** (You can have 50 or 51 students, but **not** 50.5 students).
  - **Number of laptops sold by a store.**
  - **Number of heads in three coin tosses:** (0, 1, 2, or 3)

B. Continuous Data

- **Explanation:** This data can take **any value** within a given range. It is data that you **measure**. It is infinitely divisible.
- **Operations:** You can use all statistical measures.
- **Examples:**
  - **Temperature:** (It can be 32.5°C, 32.55°C, 32.551°C, etc.)
  - **Height or Weight:** (e.g., 170.2 cm)

- **Salary:** (e.g., ₹50,250.75)

6. Explain the central limit theorem with proof.

1. Statement of the Theorem (2 Marks)

The **Central Limit Theorem (CLT)** states that for *any* population, regardless of its original distribution (it can be normal, skewed, uniform, or anything else), the **sampling distribution of the sample mean** ($\bar{x}$) will be approximately a **normal distribution** (a bell curve) as the sample size ('n') becomes sufficiently large.

Two key parameters are defined:

1. Mean of the Sample Means ($\mu_{\bar{x}}$): The mean of the sampling distribution will be equal to the original population's mean (μ).

$$\mu_{\bar{x}} = \mu$$

2. Standard Deviation of the Sample Means (σ_x): The standard deviation of the sampling distribution will be the population's standard deviation (σ) divided by the square root of the sample size (n). This is also known as the Standard Error.

$$\sigma_x = \frac{\sigma}{\sqrt{n}}$$

---

2. Illustration (Visual 'Proof') (2 Marks)

This is the most important part to explain.

1. **Imagine a Population:** Start with a population that is **not** normally distributed. For example, a "right-skewed" distribution (like salaries in a company).
2. **Take Samples:**
   - If you take many small samples (e.g., n=2) and plot their means, the new distribution will still be skewed.
   - If you take many medium samples (e.g., n=10) and plot their means, the new distribution will start to look more symmetrical.
   - If you take many large samples (e.g., n=30 or more) and plot their means, the new distribution will look almost identical to a perfect **normal (bell-shaped) distribution**.

This diagram illustrates the concept perfectly:

---

3. Conditions and Importance (2 Marks)

You must state the "rules" for the theorem to work.

Conditions:

1. **Randomness:** The samples must be drawn randomly from the population.

2. **Independence:** The samples must be independent of each other (one sample doesn't affect the next).
3. **Sufficiently Large 'n':** This is the key. While the theorem just says "large 'n'," the accepted rule of thumb in statistics is:
    - **n >= 30**
    - (Note: If the original population is *already* normal, the CLT holds true for *any* sample size, even n=1.)

Importance (Why we care):

The CLT is vital because many populations in the real world are **not** normally distributed. The CLT allows us to "force" normality by using sample means. This lets us use all the tools of the normal distribution (like z-scores, p-values) to perform **hypothesis testing** and create **confidence intervals**, which are the core of making business decisions from data

Unit II: Statistical Analysis

7. Define the following term

a) Correlation:- Correlation is a statistical measure that describes the strength and direction of a linear relationship between two quantitative (numerical) variables

b) Regression :- **Regression** is a statistical method used to model the relationship between a **dependent variable** (the one you want to predict) and one or more **independent variables** (the factors used for prediction).

c) Skewness:- **Skewness** is a measure of the **asymmetry** (lack of symmetry) of a probability distribution or dataset. It indicates which way the "tail" of the distribution is pulled.

d) Kurtosis:- **Kurtosis** is a measure that describes the "tailedness" and "peakedness" of a distribution compared to a normal distribution. It tells you how much of the data is in the tails versus the center.

8. What is data cleaning? Why it is important in the Analytics. What are the different steps in Data cleaning Process?

a. Data Cleaning, also known as data cleansing or data wrangling, is the comprehensive process of **detecting, correcting, and removing** corrupt, inaccurate, incomplete, or irrelevant records from a dataset

b. The goal is to transform a "dirty," raw dataset into a "clean," reliable, and high-quality dataset that is fit for analysis and modeling.

c. The importance of data cleaning is summed up in one famous principle: **"Garbage In, Garbage Out" (GIGO):-**

1. **Improves Model Accuracy:** A clean dataset is the single biggest factor in building an accurate and reliable predictive model

2. **Prevents Incorrect Decisions:** Dirty data leads to misleading insights. A business might make a costly mistake (e.g., launching a product in the wrong city) because of analysis based on flawed data.

3. **Increases Efficiency:** When data is clean and standardized, it is much faster and easier for data scientists to work with, saving time and resources

d. Steps in the Data Cleaning Process

1. **Handling Missing Data (Null Values):** This is the most common problem. You have rows where some information is blank

2. **Handling Duplicate Data:-** You must identify and remove duplicate rows (e.g., the same customer or transaction entered twice). Duplicate data can skew your analysis, making certain categories seem larger than they are

3. **Handling Inaccurate & Inconsistent Data (Standardization):**

4. **Handling Outliers: Outliers** are extreme values that are very different from the rest of the data

9. Define ANOVA. State its Types

ANOVA stands for **Analysis of Variance**.

It is a statistical test used to determine if there are any statistically significant differences between the **means of three or more independent groups**

Core Principle: ANOVA works by comparing the variance (spread) **between** the groups to the variance **within** each group.

- If the variance *between* the groups is much larger than the variance *within* the groups, it implies that the groups are genuinely different

Types of ANOVA –

A. One-Way ANOVA

- **Explanation:** This is the simplest form. It is used when you have **one** categorical independent variable (also called a "factor") and **one** quantitative dependent variable.
- **Key Question:** "Do the means of my dependent variable differ across the 3+ levels (groups) of my single independent variable?"
- **Example:** You want to test if the **average exam score** (dependent variable) is different among students from **three different teaching divisions** (independent variable: 'Division' with levels A, B, C).

B. Two-Way ANOVA

- **Explanation:** This is used when you have **two** categorical independent variables (factors) and **one** quantitative dependent variable.
- **Key Question:** A Two-Way ANOVA tests three things:

1. The **Main Effect** of the first independent variable.
2. The **Main Effect** of the second independent variable.
3. The **Interaction Effect** of the two variables (i.e., does the effect of one variable *depend on* the level of the other variable?).

- **Example:** You want to test the **average crop yield** (dependent variable) based on:
  - Factor 1: **Type of Fertilizer** (Brand X, Brand Y)
  - Factor 2: **Type of Soil** (Sandy, Clay)
  - This test can tell you if fertilizer matters, if soil matters, and if (for example) Brand X fertilizer works *uniquely well* only in sandy soil (the interaction).

10. Name different type of Graphical Techniques.

1. Bar Chart

- **What it is:** A chart that uses rectangular bars (vertical or horizontal) where the length of the bar is proportional to the value it represents.
- **When to use:** It is used for **comparing categorical data**.
- **Example:** Comparing the sales figures of five different products.

---

2. Histogram

- **What it is:** This looks similar to a bar chart, but there are **no gaps** between the bars. It groups a single numerical variable into "bins" (intervals) and shows the frequency (count) of data points falling into each bin.
- **When to use:** To understand the **distribution (shape)** of numerical data.
- **Example:** Visualizing the distribution of exam marks (e.g., how many students scored between 80-90, 90-100, etc.).

---

3. Line Chart (or Line Graph)

- **What it is:** A chart that connects a series of data points (markers) with a continuous straight line.
- **When to use:** This is the best chart for visualizing **trends over a period of time** (Time-Series data).
- **Example:** Tracking the stock price of a company over the last 30 days.

---

4. Pie Chart

- **What it is:** A circular statistical graph divided into "slices" to illustrate numerical proportion. The angle of each slice is proportional to the quantity it represents.

- **When to use:** To show the **parts-of-a-whole** composition of a single categorical variable. It is best used for a small number of categories (e.g., 6 or fewer).
- **Example:** The percentage breakdown of a student's total monthly budget (e.g., 40% on rent, 20% on food, etc.).

---

5. Scatter Plot

- **What it is:** A graph that uses dots (or markers) to plot the values of **two different numerical variables**, with one variable on the X-axis and the other on the Y-axis.
- **When to use:** To observe and show the **relationship (correlation)** between two numerical variables.
- **Example:** Plotting a student's "hours spent studying" (X-axis) against their "final exam score" (Y-axis) to see if there is a positive correlation.

---

6. Box Plot (Box-and-Whisker Plot)

- **What it is:** A standardized way of displaying the distribution of data based on a five-number summary: minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum.
- **When to use:** Excellent for **identifying outliers** and understanding the spread (dispersion) and skewness of data. It's also very powerful for comparing the distributions of multiple groups.
- **Example:** Comparing the salary distributions across different departments in a company.

11. What is Degrees of Freedom and how it is calculated?

In simple terms, **Degrees of Freedom (df)** refers to the number of values or observations in a calculation that are **"free to vary"** or can be chosen independently, given a certain constraint.

How is it Calculated?

- For a Single Sample (e.g., one-sample t-test):

The formula is df = n - 1

Where 'n' is the sample size. The "1" represents the one parameter (the mean) that is estimated, which acts as a constraint.

- **Other Common Examples:**
  - **Two-Sample t-test (independent):** df = $n_1 + n_2 - 2$
  - **Chi-Square Goodness-of-Fit Test:** df = k - 1 (where 'k' is the number of categories)[4]
  - **Chi-Square Test for Independence:** df = (rows – 1) x (columns – 1)

12. State the procedure for Hypothesis testing.

Procedure for Hypothesis Testing

Here is the 6-step procedure you must follow:

1. State the Hypotheses (H_0 and H_a)

This is the first and most important step. You must define two competing statements about the population.

- **Null Hypothesis (H_0):** This is the "default" assumption, a statement of **no effect** or **no difference**. It always contains an equals sign (=, $\leq$, or $\geq$).
    - *Example: $H_0: \mu = 100$ (The average mark is 100)*
- **Alternative Hypothesis (H_a or H_1):** This is the claim you are trying to find evidence for. It is the opposite of the null and never contains an equals sign ($\neq$, <, or >).
    - *Example: $H_a: \mu \neq 100$ (The average mark is not 100)*

---

2. Set the Significance Level (α)

- The significance level, or **alpha (α)**, is the **threshold of 'reasonable doubt'**. It is the maximum probability of making a **Type I error** (rejecting a true null hypothesis) that you are willing to accept.
- **Common Values:** α is chosen *before* the test. The most common values are **0.05 (5%)** or 0.01 (1%).
- An α of 0.05 means you are willing to accept a 5% chance that you are wrong when you reject the null hypothesis.

---

3. Choose the Test and Calculate the Test Statistic

- Based on your data (sample size, data type, whether you know the population standard deviation), you choose the correct statistical test.
    - *Examples: **z-test, t-test, Chi-Square (χ^2) test, ANOVA**.*
- You then use a formula to summarize your sample data into a single number, called the **test statistic**. This statistic measures how far your sample result is from what you would expect if the null hypothesis (H_0) were true.

---

4. Determine the Decision Rule (Critical Value or P-value)

You must define your rule for rejecting H_0. There are two methods:

- **A) Critical Value Method:** You find a "critical value" from a statistical table (e.g., z-table, t-table) based on your α. This value creates a **"rejection region"**.

- **B) P-value Method:** This is the modern approach. The **p-value** is the probability of observing a test statistic as extreme as, or more extreme than, the one you calculated, *assuming the null hypothesis is true*.

---

5. Make the Statistical Decision

This is a binary decision based on the rule from Step 4.

- **Using Critical Value:** If your calculated test statistic (from Step 3) falls into the rejection region (beyond the critical value), you **Reject H_0**.
- **Using P-value:** This is the key rule:
    - If **p-value $\leq \alpha$**: You **Reject H_0**. (The result is statistically significant).
    - If **p-value > α**: You **Fail to Reject H_0**. (The result is not statistically significant).

---

6. State the Final Conclusion

This is the most important step. You must translate your statistical decision back into the context of the original problem.

- **If you Rejected H_0:** "There **is** sufficient statistical evidence to conclude that..." (and you state the alternative hypothesis in words).
- **If you Failed to Reject H_0:** "There **is not** sufficient statistical evidence to conclude that..." (and you state the alternative hypothesis in words).

Unit III: Data Analytics with Excel

13. Write down steps to create an excel dashboard.

Steps to Create an Excel Dashboard

1. Data Collection and Formatting as a Table

- **Step 1:** Consolidate your raw data into a single, clean sheet.
- **Step 2:** Ensure your data is clean (no blank rows/columns, correct formats).
- **Step 3:** This is the most important part. Select your entire data range and format it as an **Excel Table** (Shortcut: **Ctrl + T**).
- **Why?** An Excel Table is a "dynamic range." When you add new data (e.g., a new month's sales), the table expands, and your dashboard will refresh automatically.

2. Create a "Helper" Sheet with Pivot Tables

- **Step 1:** Create a new, separate sheet. Name it "Helper" or "Pivot".

- **Step 2:** Go back to your Data Table. On the "Table Design" tab, click "Summarize with PivotTable." Place this Pivot Table in your new "Helper" sheet.
- **Step 3:** Create a separate Pivot Table for **each key metric (KPI)** you want to visualize. For example:
    - One Pivot Table for "Total Sales by Region".
    - A second Pivot Table for "Sales by Product".
    - A third Pivot Table for "Sales Trend Over Time (by Month)".

3. Create Visualizations (PivotCharts)

- **Step 1:** Go to your "Helper" sheet.
- **Step 2:** Click on your first Pivot Table (e.g., "Sales by Region") and go to the "PivotTable Analyze" tab. Click on **"PivotChart"**.
- **Step 3:** Choose the correct chart type (e.g., a **Bar Chart** for "Sales by Region," a **Line Chart** for "Sales Trend Over Time").
- **Step 4:** Repeat this for every Pivot Table you created. You will now have 3-4 charts on your "Helper" sheet.

4. Design the Main Dashboard Sheet

- **Step 1:** Create another new, blank sheet. Name this one **"Dashboard"**.
- **Step 2:** For a professional look, remove the gridlines. Go to the **View** tab and uncheck the **Gridlines** box.
- **Step 3:** Add a clear title at the top of the sheet (e.g., "Annual Sales Dashboard").

5. Assemble and Link with Slicers (Interactivity)

This is the most important step that makes it a "dashboard."

- **Step 1 (Assemble):** Go to your "Helper" sheet. **Cut** (do not copy) each chart (Ctrl + X) and **Paste** (Ctrl + V) it onto your "Dashboard" sheet. Arrange them neatly.
- **Step 2 (Add Slicers):** Click on any one of your charts on the "Dashboard." Go to the "PivotTable Analyze" tab and click **"Insert Slicer"**.
- **Step 3 (Choose Filters):** Select the fields you want to filter by (e.g., "Region," "Year," "Product Category"). Click OK.
- **Step 4 (Link Slicers):** This is critical. By default, a slicer only controls *one* chart. You must link it to all of them.
    - Right-click the Slicer -> **"Report Connections..."**
    - In the pop-up box, check the boxes for **ALL** the Pivot Tables you created.
    - Repeat this for every slicer.

6. Final Formatting and Review

- **Step 1:** Arrange your Slicers neatly (e.g., in a horizontal menu at the top).
- **Step 2:** "Clean" your charts. Right-click on the "gray buttons" on your PivotCharts (field buttons) and select **"Hide all field buttons on chart"**.
- **Step 3:** Test your dashboard. Click on any button in a Slicer (e.g., "North Region"). All the charts on your dashboard should update at the same time.

14. Write down steps to create a pivot table

(Professor nods)

"A very good, practical question. The Pivot Table is perhaps the single most powerful tool in all of Excel for data analysis. It is the fastest way to summarize large amounts of data.

For a 6-mark answer, you must provide these clear, sequential steps.

Steps to Create a Pivot Table

Step 1: Prepare Your Data

- Before you begin, you must have "clean" source data.
- Your data must be in a table format, with **no blank rows or columns**.
- Each column must have a unique **header** (e.g., "Region", "Sales", "Date").

Step 2: Select Your Data

- Click any single cell **inside** your data table.
- (Optional but recommended): Format your data as an Excel Table (**Ctrl + T**). This makes your data "dynamic," so the Pivot Table can be refreshed easily when you add new rows.

Step 3: Insert the Pivot Table

- Go to the **Insert** tab on the Excel ribbon.
- On the far left, click the **PivotTable** button.
- A "Create PivotTable" dialog box will appear.

Step 4: Choose Your Data and Location

- In the dialog box, Excel will automatically confirm the data range (your table).
- It will ask where to place the Pivot Table. By default, it selects **"New Worksheet"**. This is the best practice.
- Click **OK**.

Step 5: Build the Report using the PivotTable Fields

- Excel will open a new sheet with a blank Pivot Table on the left and a **"PivotTable Fields"** pane on the right.

- This pane is the most important part. It has two sections:
    1. **Field List:** A list of all your column headers.
    2. **Areas (Layout):** Four empty boxes at the bottom.
- You must **drag and drop** your fields into these four areas:
    - **ROWS:** Fields you want to list down the side (e.g., drag "Product Category" here).
    - **COLUMNS:** Fields you want to list across the top (e.g., drag "Region" here).
    - **VALUES:** The numerical field you want to calculate (e.g., drag "Sales" here). It will default to Sum of Sales.
    - **FILTERS:** Fields you want to use as a master filter for the whole report (e.g., drag "Year" here).

Step 6: Summarize and Analyze

- As you drag fields into the areas, the Pivot Table on the left will instantly build and summarize your data.
- You can change the calculation in the VALUES area (e.g., from Sum to Average or Count) by clicking on it and selecting "Value Field Settings...".

15. Explain grouping of items in Pivot Charts with suitable examples.

1. Grouping Dates (Most Common Use)

This is the most common and useful grouping feature.

- **Scenario:** You have a data table with **daily sales** for an entire year (365 rows of data).
- **Problem:** If you create a Pivot Table and Chart of this, your line chart will have 365 individual points. This is too messy to read, and you cannot see any trend.
- **Solution (Grouping):**
    1. You build your Pivot Table (e.g., "Date" in Rows, "Sales" in Values).
    2. You **right-click** any of the date cells in the Pivot Table (e.g., "01-Jan-2024").
    3. Select **"Group..."** from the menu.
    4. A dialog box appears. You un-select "Days" and select **"Months"** and **"Years"**.
- **Result on the Pivot Chart:**
    - The Pivot Table instantly collapses the 365 daily rows into 12 monthly rows.
    - Your messy 365-point line chart **automatically updates** to become a clean, 12-point line chart. This now clearly shows you the sales trend from month to month, which is a far more useful insight.

---

2. Grouping Numbers (Creating a Frequency Distribution)

This is used to create "bins" or "brackets" from a continuous range of numbers.

- **Scenario:** You have a list of 1,000 students and their "Final Exam Marks" (ranging from 0 to 100).
- **Problem:** If you create a Pivot Chart (Bar Chart) of this, it will try to show a separate bar for every single mark (e.g., a bar for 71, 72, 73...). This is not a summary.
- **Solution (Grouping):**
    1. You build your Pivot Table (e.g., "Marks" in Rows, "Count of Students" in Values).
    2. You **right-click** any of the "Marks" cells in the Row area.
    3. Select **"Group..."** from the menu.
    4. The dialog box will ask for a "Starting at:", "Ending at:", and "By:" value.
    5. You set: **Starting at: 0**, **Ending at: 100**, **By: 10**.
- **Result on the Pivot Chart:**
    - The Pivot Table collapses the individual marks into groups: "0-9", "10-19", "20-29", and so on.
    - The Pivot Chart **automatically updates** to become a **Histogram**. It now shows you a bar chart with 10 bars, clearly displaying the *distribution* of marks (e.g., "200 students scored in the 80-89 bracket").

16. Describe the process of sorting and filtering in Pivot Tables.

1. Sorting in Pivot Tables (3 Marks)

This changes the order of your items, such as sorting from highest to lowest sales, or alphabetically.

There are two main ways to sort:

A) Sort by Value (e.g., Largest to Smallest): This is the most common. You want to see your top-performing products or regions.

- **Steps:**
    1. Go to your Pivot Table.
    2. Right-click on a number in your **Values** column (e.g., a cell under "Sum of Sales").
    3. Select **Sort** -> **Sort Largest to Smallest**. The entire list of items (e.g., products) will re-order based on their sales value.

B) Sort by Label (e.g., A to Z): This is used to sort the text-based row or column labels themselves, not their values.

- **Steps:**

1. Right-click on one of your **Row Labels** (e.g., a product name like "Laptop").
2. Select **Sort** -> **Sort A to Z**. This will sort your items alphabetically.

---

2. Filtering in Pivot Tables (3 Marks)

This lets you **hide** data you don't want to see, allowing you to focus on a specific subset (e.g., "only show data for the Pune region").

There are three main ways to filter:

A) The Label Filter (Basic):

- In your Pivot Table, click the drop-down arrow next to the 'Row Labels' or 'Column Labels' header.
- A menu will appear. You can uncheck the items (e.g., uncheck 'Mumbai' and 'Delhi') to hide them.

B) The Report Filter (Classic):

- In the **"PivotTable Fields"** pane (on the right), drag a field (e.g., "Year") into the **FILTERS** area.
- This adds a drop-down filter *above* your Pivot Table, allowing you to control the entire report.

C) Slicers (Modern & Recommended): This is the best method for creating interactive reports or dashboards.

- **Steps:**
    1. Click anywhere inside your Pivot Table.
    2. Go to the **PivotTable Analyze** tab on the ribbon.
    3. Click **Insert Slicer**.
    4. Select the field you want to filter by (e.g., "Region").
- This creates a set of user-friendly buttons that you can click to filter your data.

17. Illustrate how Excel can be used for trend analysis.

1. Method 1: Visual Analysis (Using Line Charts)

This is the simplest, most intuitive first step. A line chart is the best tool for visualizing data over time.

- **Step 1:** Organize your data in two columns: **Date** (e.g., Jan, Feb, Mar) and **Value** (e.g., Sales).
- **Step 2:** Select your data.
- **Step 3:** Go to the **Insert** tab and select a **Line Chart**.
- **Result:** You can immediately see the pattern:

- Is the line generally going up (**Uptrend**)?
- Is it going down (**Downtrend**)?
- Is it going up and down at regular intervals (**Seasonality**)?

---

2. Method 2: Statistical Analysis (Adding a Trendline)

This is the most important part. You add a statistical "line of best fit" over your chart to see the *average* trend, ignoring the small "noise."

- **Step 1:** Create the Line Chart as described above.
- **Step 2:** Right-click directly on the data line in your chart.
- **Step 3:** Select **Add Trendline...** from the menu.
- **Step 4:** The "Format Trendline" pane will open.
  - Select **Linear** for a simple, straight-line trend.
  - **Crucial Step:** Check the boxes for **Display Equation on chart** and **Display R-squared value on chart**.
- **Result:**
  - The **Trendline** shows the true, underlying direction.
  - The **Equation** (e.g., $y = 500x + 2000$) gives you a mathematical model to make predictions.
  - The **R-squared** value (e.g., 0.92) tells you how well the line fits your data (closer to 1 is better).

---

3. Method 3: Predictive Analysis (The "Forecast Sheet" Tool)

This is Excel's most powerful, one-click tool. It not only finds the trend but also projects it into the future with confidence.

- **Step 1:** Select your entire two-column data range (Date and Sales).
- **Step 2:** Go to the **Data** tab on the ribbon.
- **Step 3:** In the "Forecast" group, click the **Forecast Sheet** button.
- **Step 4:** A pop-up window will instantly show you a preview of the forecast. Click **Create**.
- **Result:** Excel creates a new worksheet with:
  1. Your original data.
  2. A **forecasted** set of data for future periods.

3. A chart that shows your past data (blue) and the future forecast (orange), along with **Upper and Lower Confidence Bounds**. This gives you a best-case, worst-case, and most-likely scenario.

These are the three levels of trend analysis you can perform in Excel.

18. Explain the role of Slicers in updating Pivot Table data.

A **Slicer** is a visual, interactive, user-friendly **filter** for a Pivot Table. Instead of using the traditional, clunky drop-down filter menus, Slicers provide a set of large, clickable buttons to filter your data.

---

2. The Role of Slicers (5 Marks)

The Slicer's primary role is to make filtering data **faster, easier, and more transparent**.

Here are its key functions:

- **1. Dynamic and Interactive Filtering:** This is its main job. When you click a button in a Slicer (e.g., you click "Pune" in a Region slicer), the Pivot Table instantly filters to show **only** the data for Pune. You can click multiple buttons (e.g., "Pune" and "Mumbai") to show data for both.
- **2. Visually Shows the Current Filter State:** This is a huge advantage. An old drop-down filter menu hides its selection (you just see a small filter icon). A Slicer, however, is always visible on the screen. The selected item (e.g., "Pune") stays highlighted, so anyone looking at the report can immediately see **what** is being filtered.
- **3. Central Control for Dashboards (Most Important Role):** This is the real power of Slicers. A single Slicer can be connected to **multiple Pivot Tables** (and their Pivot Charts) at the same time.
    - **Example:** Imagine you have a dashboard sheet with 4 different Pivot Charts (Sales by Region, Sales by Product, Sales by Month, etc.).
    - You can create one "Year" Slicer (with buttons "2023", "2024").
    - By using the **"Report Connections"** option, you link this one Slicer to all four Pivot Tables.
    - Now, when you click the "2024" button, **all four charts on your dashboard update at once**. This is impossible to do with a simple drop-down filter.

Unit IV: Data Visualization

19. Describe the steps to move and resize an embedded chart. (4M)

1. Steps to Move an Embedded Chart (2 Marks)

1. **Select the Chart:** Click anywhere on the chart's border or in a blank white space inside the chart area. (Be careful not to click on a chart element, like a bar or the title, as this will select that element instead).

2. **Get the Move Cursor:** Hover your mouse pointer over the chart's border. The cursor will change from a normal pointer to a **four-headed arrow** ✚.
3. **Click and Drag:** Once the four-headed arrow appears, click and hold the left mouse button.
4. **Move and Release:** Drag the chart across the worksheet to its new location. Release the mouse button to drop it in place.

---

2. Steps to Resize an Embedded Chart (2 Marks)

1. **Select the Chart:** Click on the chart to make it active. When selected, you will see a frame around it with small circles at the corners and on the sides. These are called **resize handles**.
2. **Get the Resize Cursor:** Move your mouse pointer directly over one of these resize handles. The cursor will change into a **two-headed arrow** (e.g., ↔ or diagonally).
3. **Click and Drag:**
    - **To resize proportionally (Recommended):** Click and drag a **corner handle** diagonally. This keeps the chart's height and width ratio the same, so it doesn't look stretched.
    - **To resize in one direction:** Click and drag a **side handle** (top, bottom, left, or right). This will stretch or squash the chart in that direction.
4. **Release:** When the chart is the size you want, release the mouse button.

20. State the steps to chart non-adjacent cells with example. (2M)

Steps:

1. **Select the first data range** (e.g., A1:A10) by clicking and dragging your mouse.
2. **Press and hold the Ctrl key** on your keyboard.
3. While holding Ctrl, **select the second data range** (e.g., C1:C10) with your mouse.
4. **Release the Ctrl key**. Now, both non-adjacent ranges are selected.
5. Go to the **Insert** tab and choose your desired chart (e.g., Bar Chart).

21. What is meant by Exploding a Slice in a Pie Chart? (2M)

Exploding a slice in a pie chart is a visual technique where one or more slices are moved slightly away from the center of the pie

The single, most important purpose of this is to **add emphasis** to that specific slice. It makes that piece of data (e.g., the 'Top Performer' or a 'Problem Area') stand out from the rest for the person reading the chart."

22. What is the use of a Legend and marker in a chart? (2M)

1. Legend

A **Legend** is a key or a guide that **identifies the data series** in a chart. It is essential when you are plotting more than one set of data on the same chart.

- **Use:** It links the colors, patterns, or symbols in the chart to the data's name. For example, in a bar chart comparing sales for 2023 and 2024, the legend would show:
    - **Blue Bar = 2023 Sales**
    - **Orange Bar = 2024 Sales**

2. Marker

A **Marker** is a symbol (like a circle, square, or diamond) that is used to represent an **individual data point** on a chart.

- **Use:** Markers are most commonly seen on **Line Charts** and **Scatter Plots**. They pinpoint the exact location of a specific value (e.g., the sales for the month of "March") on the chart, and the line simply connects these markers.

23. Explain the steps with example i)format chart title and legend in Excel, ii) Formatting and aligning numbers and text in a chart. (4M)

i) Format Chart Title and Legend (2M)

To Format the Chart Title:

1. **Select:** Click once on the "Chart Title" text to select its box.
2. **Edit Text:** Click again to place your cursor inside and type your new title (e.g., "Monthly Sales Report").
3. **Format Look: Double-click** the title's box. The **"Format Chart Title" pane** opens.
    - Use **"Fill & Line"** (paint bucket icon) to change the background color or border.
    - Use **"Text Options"** to change the font color, add a text outline, or apply effects like shadows.

To Format the Legend:

1. **Select: Double-click** the Legend box to select it and open the **"Format Legend" pane**.
2. **Format Position:** Click the **"Legend Options"** icon (three bars). This is the most important part. You can change the **Position** of the legend (e.g., move it to the Top, Bottom, or Right of the chart).
3. **Format Look:** You can also use the "Fill & Line" options to put a border or background color on the legend box.

---

ii) Formatting and Aligning Numbers and Text (2M)

This refers to the labels on your axes or the data itself.

To Format Numbers (e.g., on the Y-Axis):

1. **Select: Double-click** the axis with the numbers you want to format (e.g., the Y-axis). The **"Format Axis" pane** will open.
2. **Format Number:** Click the **"Axis Options"** icon (three bars). Scroll down to the **"Number"** section.
3. **Example:** Here, you can change the **Category** from General to Currency to add a "₹" symbol (e.g., ₹1,00,000) or set the number of **Decimal Places** to 0.

To Align Text (e.g., on the X-Axis):

1. **Select: Double-click** the text-based axis (e.g., the X-axis with category names). The **"Format Axis" pane** opens.
2. **Align Text:** Click the **"Text Options"** tab. Select the **"Textbox"** icon (looks like a page).
3. **Example:** Here you can change the **"Text direction"** or use **"Custom Angle"** to rotate your labels (e.g., to 45 degrees) so they don't overlap.

24. Write steps to create pie chart with example. (4M)

Steps to Create a Pie Chart

1. **Select Your Data:**
   - Click and drag your mouse to select the **entire data range** you want to chart.
   - In our example, you would select cells A1 to B5 (including the headers).
2. **Go to the Insert Tab:**
   - On the Excel ribbon at the top, click the **Insert** tab.
3. **Find the Pie Chart Icon:**
   - In the "Charts" section, look for the Pie Chart icon (a circle divided into slices).
   - Click this icon.
4. **Choose Your Chart Type:**
   - A drop-down menu will appear showing different pie chart styles (e.g., 2D Pie, 3D Pie, Doughnut).
   - Click on the **2D Pie** to insert the basic chart onto your sheet.
5. **Add Data Labels (Recommended):**
   - The chart will appear, but it's not very useful without numbers.
   - Click the **+** icon on the top-right of the chart.
   - Check the box for **Data Labels**.

- To show percentages, click the small arrow next to "Data Labels" -> **More Options...**, and check the box for **Percentage**.

Unit V: Data Visualization using Python

25. Explain the procedure to install and set up Matplotlib in Python.

1. Prerequisite: Check Python and pip

Before you install anything, you must make sure Python and its package manager, **pip**, are installed.

- Open your **Command Prompt** (on Windows) or **Terminal** (on Mac/Linux).
- Type python --version and press Enter. You should see a version number.

2. Installation
In your Command Prompt or Terminal, run the following command

pip install matplotlib

3. "Setup" and Verification

Step 1: Open a Python Script or Interpreter

- You can open any Python IDE (like VS Code, PyCharm) or just type python in your terminal to open the simple interpreter.

Step 2: Import the Library

- To verify the installation, type:

```
import matplotlib

print(matplotlib.__version__)
```

26. Describe the steps to add titles, labels, and legends to a plot using matplotlib with example

Key Matplotlib Functions

1. **Title:** You use the plt.title() function to add a main title to the top of the chart.
2. **Labels:** You use plt.xlabel() and plt.ylabel() to add descriptive labels to the X-axis and Y-axis, respectively.
3. **Legend:** This is a two-step process:
    - **Step 1:** You must add a label parameter inside your plotting command (e.g., plt.plot(..., label='Data 1')) for each data series you plot.
    - **Step 2:** You must call the plt.legend() function to tell Matplotlib to display the legend on the chart.

Python

```
# Step 1: Import the library
import matplotlib.pyplot as plt
import numpy as np

# Step 2: Create some sample data
x = np.array([1, 2, 3, 4, 5])
y1 = x * 2
y2 = x ** 2

# Step 3: Create the plot.
# Notice the 'label' parameter here. This is for the legend.
plt.plot(x, y1, label='Linear (x*2)', marker='o')
plt.plot(x, y2, label='Quadratic (x^2)', marker='x')

# --- This is the answer to your question ---

# Step 4: Add the Title and Labels
plt.title("Growth Comparison: Linear vs. Quadratic")  # Adds the main title
plt.xlabel("Input Value (X)")                  # Adds the X-axis label
plt.ylabel("Calculated Value (Y)")             # Adds the Y-axis label

# Step 5: Add the Legend
# This command finds all the 'label' parameters from Step 3
# and displays them in a box.
plt.legend()

# --- End of answer ---

# Step 6: Show the final plot
plt.grid(True) # Adds a grid for easier reading
plt.show()
```

27. Describe the process of changing figure size and aspect ratio in Matplotlib with example.

The figsize Parameter

- The **figsize** parameter is a **tuple** (a pair of numbers) that defines the (width, height) of your chart.
- The units are in **inches**.
- This single parameter controls both the **size** and the **aspect ratio**.

Aspect Ratio is not a separate setting. It is simply the **ratio** of the width to the height you provide in figsize.

- figsize=(8, 4) creates an 8-inch wide, 4-inch tall plot. The aspect ratio is 2:1 (a "wide" plot).
- figsize=(6, 6) creates a 6-inch wide, 6-inch tall plot. The aspect ratio is 1:1 (a "square" plot).

Python Code:-

```
# Step 1: Import the library
import matplotlib.pyplot as plt
import numpy as np

# Step 2: Create sample data
x = np.linspace(0, 10, 100) # 100 points from 0 to 10
y = np.sin(x)

# --- This is the answer to your question ---

# Step 3: Create the figure AND set its size.
# We are creating a plot that is 10 inches wide and 4 inches tall.
# This results in a "wide" aspect ratio, perfect for a time-series.
fig, ax = plt.subplots(figsize=(10, 4))

# --- End of answer ---

# Step 4: Plot your data as usual
ax.plot(x, y)
ax.set_title("My Wide Sine Wave Plot (10x4 inches)")
ax.set_xlabel("Time")
ax.set_ylabel("Amplitude")

# Step 5: Show the plot
plt.show()
```

28. Explain how to customize axis limits, ticks, and labels. Show with example

Key Functions for Customization

1. **Axis Limits (xlim, ylim):**
    - This controls the **min and max range** of the axis.
    - **Functions:** plt.xlim(min, max) and plt.ylim(min, max).
    - **Use:** To "zoom in" on a specific area or to "zoom out" to give your data some padding.
2. **Axis Ticks (xticks, yticks):**
    - This controls **where the tick marks are placed** on the axis.
    - **Function:** plt.xticks(list_of_tick_positions)
    - **Use:** To specify the exact points you want to mark (e.g., [0, 10, 20, 30]).
3. **Axis Tick Labels (xticks, yticks):**
    - This controls the **text that appears at the tick marks**.
    - **Function:** plt.xticks(ticks=positions_list, labels=labels_list)
    - **Use:** This is how you change numerical labels to text (e.g., changing 1, 2, 3 to "Jan", "Feb", "Mar").

Python Code:-

```
import matplotlib.pyplot as plt


# Data for 4 quarters
x = [1, 2, 3, 4]
y = [150, 200, 175, 225]


# Create the plot
plt.plot(x, y, 'o-', color='red', linewidth=2)
plt.title("Quarterly Sales Report (Customized)")
plt.xlabel("Quarter")
plt.ylabel("Sales (in Crores)")


# --- This is the answer to your question ---
```

```
# 1. Customize Axis Limits
# Give the plot some padding.
# X-axis will go from 0 to 5 (instead of 1 to 4)
# Y-axis will go from 100 to 250 (instead of 150 to 225)
plt.xlim(0, 5)
plt.ylim(100, 250)

# 2. Customize Ticks and Labels (X-Axis)
# We want ticks at positions 1, 2, 3, 4
# But we want to label them with text.
tick_positions_x = [1, 2, 3, 4]
tick_labels_x = ["Q1-2024", "Q2-2024", "Q3-2024", "Q4-2024"]
plt.xticks(ticks=tick_positions_x, labels=tick_labels_x)

# 3. Customize Ticks (Y-Axis)
# We only want ticks at 100, 150, 200, 250.
tick_positions_y = [100, 150, 200, 250]
plt.yticks(ticks=tick_positions_y)

# --- End of answer ---

plt.grid(True, linestyle='--') # Add a light grid
plt.show()
```

29. Discuss how to save a plot in PNG format with a higher resolution.

The savefig() Function and dpi Parameter

1. **Function:** plt.savefig("filename.png")
    - This is the function you call to save the plot.
    - You must call this **before** you call plt.show(). If you call it after plt.show(), Matplotlib will save a blank image.
2. **Parameter: dpi (Dots Per Inch)**
    - This is the parameter that controls the resolution.
    - A higher dpi value creates a larger, higher-quality PNG file.
    - **Default dpi:** Often 100 or 72. This is low resolution, "screen quality."
    - **High dpi:** For a report or presentation, you should use **dpi=300** or dpi=600.
3. **Parameter: bbox_inches='tight' (Recommended)**
    - This is an optional but highly useful parameter.
    - It tells Matplotlib to "crop" the saved image to remove any excess white space around your chart.

```
import matplotlib.pyplot as plt

import numpy as np


# 1. Create your plot as usual

x = np.linspace(0, 10, 100)

y = np.sin(x)


plt.plot(x, y)

plt.title("My High-Resolution Sine Wave")

plt.xlabel("X-Axis")

plt.ylabel("Y-Axis")


# --- This is the answer ---


# 2. Save the figure BEFORE showing it

# We are saving it as a PNG with 300 DPI and a tight layout.

plt.savefig(
```

```
    "my_high_res_plot.png",  # The filename and extension
    dpi=300,                 # Set the resolution
    bbox_inches='tight'      # Crop extra white space
)

# --- End of answer ---

# 3. Now, you can show the plot
plt.show()

print("Plot saved successfully to my_high_res_plot.png")
```

30. Explain the concept of interactive visualizations using Matplotlib widgets.

1. The Concept of Interactive Visualizations (2 Marks)

An **interactive visualization** is a plot that allows a user to **change the parameters** of the data being displayed in real-time, without having to re-run the code.

The goal is to move from a static image to a dynamic tool for exploration. Instead of just *looking* at a chart, you *use* the chart to ask "what if?" questions.

In Matplotlib, this is achieved using **Widgets**. These are graphical control elements (like sliders, buttons, or radio buttons) that you add to your plot window.

The process is:

1. A user manipulates a **Widget** (e.g., drags a slider).
2. The widget fires an **Event**.
3. This event calls a special **update function** that you write.
4. The update function re-calculates the data and redraws the plot.

Python Code:-

```
import matplotlib.pyplot as plt
from matplotlib.widgets import Slider
import numpy as np

# Create the figure and the main plot
```

```
fig, ax = plt.subplots()
plt.subplots_adjust(bottom=0.25) # Leave space at the bottom

# Plot the initial, default line
t = np.arange(0.0, 1.0, 0.001)
initial_freq = 3
line, = ax.plot(t, np.sin(2*np.pi*initial_freq*t))

# Create the axes *for the slider*
# [left, bottom, width, height]
slider_ax = plt.axes([0.25, 0.1, 0.65, 0.03])
freq_slider = Slider(
    ax=slider_ax,
    label='Frequency (Hz)',
    valmin=0.1,
    valmax=30.0,
    valinit=initial_freq
)
def update(val):
    new_freq = freq_slider.val
    new_y_data = np.sin(2*np.pi*new_freq*t)
    line.set_ydata(new_y_data) # Update the line's data
    fig.canvas.draw_idle()    # Redraw the plot
freq_slider.on_changed(update)
plt.show()
```

Measures of Central Tendency (Python Example)

```
import numpy as np
from statistics import mean, median, mode
# Create sample dataset
data = [10, 12, 15, 12, 18, 20, 12, 25]
```

```
# Calculate measures
mean_value = mean(data)
median_value = median(data)
mode_value = mode(data)

# Display results
print("Data:", data)
print("Mean =", mean_value)
print("Median =", median_value)
print("Mode =", mode_value)
```

Data Types in Data Analytics (Python Demo)

```
import pandas as pd

# Create a small dataset
data = {
   "Gender": ["Male", "Female", "Male", "Female"],
   "Age": [22, 25, 24, 23],
   "Height_cm": [175.5, 160.2, 169.8, 158.4]
}

df = pd.DataFrame(data)

# Display data types
print(df)
print("\nData Types:\n", df.dtypes)
```

Central Limit Theorem (Python Illustration)

```
import numpy as np
import matplotlib.pyplot as plt

# Create a non-normal population (right-skewed)
population = np.random.exponential(scale=2, size=10000)

# Take many random samples and store their means
sample_means = []
for i in range(1000):
   sample = np.random.choice(population, size=30, replace=True)
   sample_means.append(np.mean(sample))

# Plot both population and sampling distribution
plt.figure(figsize=(10,5))
```

```
plt.subplot(1,2,1)
plt.hist(population, bins=30, color='lightblue', edgecolor='black')
plt.title("Original Population (Skewed)")

plt.subplot(1,2,2)
plt.hist(sample_means, bins=30, color='lightgreen', edgecolor='black')
plt.title("Sampling Distribution of Means (CLT → Normal)")

plt.tight_layout()
plt.show()
```

Correlation & Regression (Python Example)

```
import numpy as np
from scipy import stats

# Example data (study hours vs marks)
hours = np.array([2, 4, 6, 8, 10])
marks = np.array([40, 50, 60, 70, 80])

# Calculate correlation coefficient
corr = np.corrcoef(hours, marks)[0, 1]
print("Correlation Coefficient (r):", corr)

# Perform simple linear regression
slope, intercept, r_value, p_value, std_err = stats.linregress(hours, marks)
print(f"Regression Line: Marks = {intercept:.2f} + {slope:.2f} * Hours")
```

Data Cleaning Steps (Python Example)

```
import pandas as pd
import numpy as np

# Create raw data with issues
data = {
   "Name": ["A", "B", "B", "C", None],
   "Age": [20, 21, 21, np.nan, 22],
   "Marks": [85, 90, 90, 88, 92]
}

df = pd.DataFrame(data)
print("Original Data:\n", df)

# Remove duplicates
```

```
df = df.drop_duplicates()

# Handle missing values
df["Name"].fillna("Unknown", inplace=True)
df["Age"].fillna(df["Age"].mean(), inplace=True)

# Display cleaned data
print("\nCleaned Data:\n", df)
```

ANOVA (Python Example)

```
import scipy.stats as stats

# Create sample data for 3 groups
group_A = [82, 85, 88, 90, 87]
group_B = [78, 80, 75, 79, 77]
group_C = [92, 94, 89, 96, 91]

# Perform one-way ANOVA
f_statistic, p_value = stats.f_oneway(group_A, group_B, group_C)

# Display results
print("F-statistic:", f_statistic)
print("p-value:", p_value)

if p_value < 0.05:
    print("Reject H0 → Significant difference between group means")
else:
    print("Fail to reject H0 → No significant difference")
```

Graphical Techniques (Python Examples)

```
import matplotlib.pyplot as plt
import numpy as np

# Bar chart
products = ["A", "B", "C", "D"]
sales = [100, 120, 80, 150]
plt.bar(products, sales)
plt.title("Sales by Product")
plt.xlabel("Product")
plt.ylabel("Sales")
plt.show()
```

```
# Histogram
marks = np.random.randint(40, 100, 50)
plt.hist(marks, bins=5, color='orange', edgecolor='black')
plt.title("Distribution of Marks")
plt.xlabel("Marks Range")
plt.ylabel("Frequency")
plt.show()
```

Hypothesis Testing (Python Example)

```
from scipy import stats
import numpy as np

# Sample data (average marks of 10 students)
sample = [52, 48, 50, 47, 53, 49, 46, 51, 50, 48]
mu_0 = 50

# Perform one-sample t-test
t_stat, p_value = stats.ttest_1samp(sample, mu_0)
print("t-statistic:", t_stat)
print("p-value:", p_value)

if p_value < 0.05:
    print("Reject H0 → Sample mean differs significantly from population mean")
else:
    print("Fail to reject H0 → No significant difference")
```

Data Visualization (Combined Concepts)

```
import matplotlib.pyplot as plt
import numpy as np

months = ["Jan", "Feb", "Mar", "Apr", "May"]
sales_2024 = [100, 120, 140, 130, 150]
sales_2025 = [110, 125, 135, 145, 160]

plt.plot(months, sales_2024, marker='o', label='2024 Sales')
plt.plot(months, sales_2025, marker='s', label='2025 Sales')
plt.title("Monthly Sales Comparison (2024 vs 2025)")
plt.xlabel("Month")
plt.ylabel("Sales (in ₹000)")
plt.legend()
plt.grid(True, linestyle='--')
plt.show()
```