Data Analytics (315326)

Model Question Paper — Set A · Model Answers

Q.1

a) Define Data Analytics; two applications. 2 M

Data Analytics is the process of collecting, cleaning, transforming and analysing raw data using statistical and computational methods to discover useful patterns and information, draw conclusions and support decision-making.

Applications (any two): Banking — fraud detection; Retail/E-commerce — product recommendation and sales forecasting; Healthcare — predicting disease risk; Transport — route optimisation; Sports — player performance analysis.

(Definition 1 M, two applications 1 M)

b) Skewness and its types. 2 M

Skewness means lack of symmetry in a frequency distribution. It shows whether the values are spread more towards the left or the right of the mean.

Types: i) Positive (right) skewness — long tail on the right, Mean > Median > Mode. ii) Negative (left) skewness — long tail on the left, Mean < Median < Mode. (Symmetric distribution: zero skewness.)

(Definition 1 M, types 1 M)

c) Central Limit Theorem. 2 M

Central Limit Theorem (CLT): If samples of size n are drawn at random from any population with mean μ and standard deviation σ, then as n becomes large, the sampling distribution of the sample mean $\bar{x}$ becomes approximately **normal** with mean μ and standard deviation (standard error) $\sigma/\sqrt{n}$ — whatever the shape of the population.

A sample size of **n ≥ 30** is generally considered large enough.

(Statement 1½ M, sample size ½ M)

d) Slicer in a PivotTable. 2 M

A **slicer** is a visual filter control in Excel, shown as a box of buttons (one button for each item of a field, e.g. East, West, North). Clicking a button filters the PivotTable / PivotChart to show only that item.

Use: Quick, interactive filtering of PivotTables, Pivot Charts and dashboards; it also shows clearly which filter is currently applied. One slicer can be connected to several PivotTables (Report Connections).

(Definition 1 M, use 1 M)

e) Embedded chart. 2 M

An **embedded chart** is a chart that is placed as an object **on the same worksheet** as its source data (it floats over the cells). It can be moved and resized with the mouse, and it is saved and printed with the worksheet.

The other option is a **chart sheet**, where the chart occupies a separate sheet of its own.

(Correct explanation 2 M)

f) plt.xlabel() and plt.ylabel(). 2 M

- **plt.xlabel('text')** — sets the label (title) of the x-axis, i.e. the horizontal axis, e.g. `plt.xlabel('Month')`.
- **plt.ylabel('text')** — sets the label of the y-axis, i.e. the vertical axis, e.g. `plt.ylabel('Sales (₹)')`.

Axis labels tell the reader what quantity and unit each axis shows. Font size etc. can be given: `plt.xlabel('Month', fontsize=12)`.

(Each function 1 M)

g) Formats for saving a plot; function. 2 M

Formats: PNG, PDF, SVG, JPG (also EPS).

Function: `plt.savefig('filename.ext')` — the format is taken from the extension, e.g. `plt.savefig('chart.png', dpi=300)`, `plt.savefig('chart.pdf')`.

(Two formats 1 M, function 1 M)

Q.2

a) Life cycle of Data Analytics. 4 M

The data analytics life cycle is a series of phases that a data project goes through. The work is **iterative** — a team may go back to an earlier phase when new information is found.

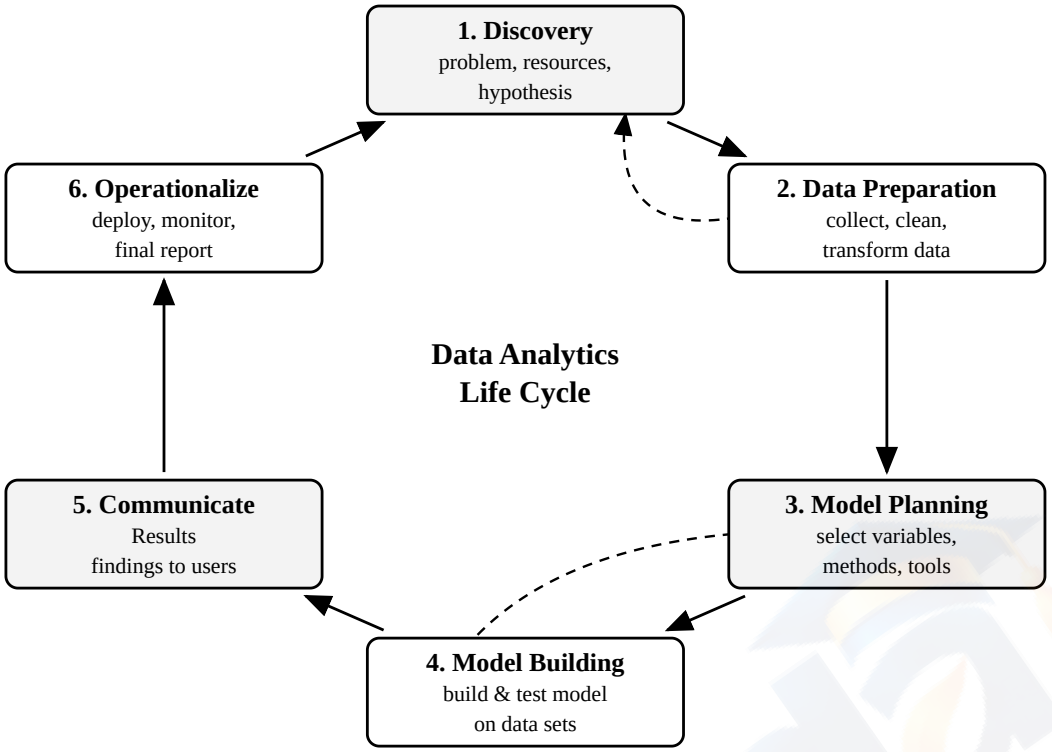


Fig.: Data Analytics life cycle (six phases; dashed arrows show going back to an earlier phase)

1. **Discovery:** Understand the business problem, frame it as an analytics problem, check available resources (people, data, tools, time) and form initial hypotheses.
2. **Data preparation:** Collect data from sources, set up an analytic sandbox, clean, transform and integrate data (ETL / ELT). Usually the most time-consuming phase.
3. **Model planning:** Explore data, select important variables and choose the techniques / models (e.g. regression, classification, clustering) and tools.
4. **Model building:** Divide data into training and testing sets, build the model, run it and check its accuracy.
5. **Communicate results:** Compare results with the goals set in discovery, and present key findings to stakeholders using charts and reports.
6. **Operationalize:** Deploy the model in a pilot / real environment, deliver final reports, code and documentation, and monitor performance.

(Diagram 2 M, explanation of phases 2 M)

b) Box plot and five-number summary. 4 M

Arranging the data in ascending order: 8, 11, 13, 14, 16, 17, 19, 20, 25 (n = 9)

Minimum = 8

Q1 = 1(n + 1)/4 th value = 2.50th value = 2th value + 0.50 × (3th − 2th) = 11 + 0.50 × (13 − 11) = **12**

Q2 (Median) = 2(n + 1)/4 th value = 2×10/4 = 5th value = **16**

Q3 = 3(n + 1)/4 th value = 7.50th value = 7th value + 0.50 × (8th − 7th) = 19 + 0.50 × (20 − 19) =

19.50

Maximum = 25

IQR = Q3 − Q1 = 19.50 − 12 = 7.50. Lower fence = Q1 − 1.5 × IQR = 0.75; upper fence = Q3 + 1.5 × IQR = 30.75. All values lie within the fences, so there is **no outlier**; whiskers go to the minimum and maximum.

Minimum	Q1	Median (Q2)	Q3	Maximum
8	12	16	19.50	25

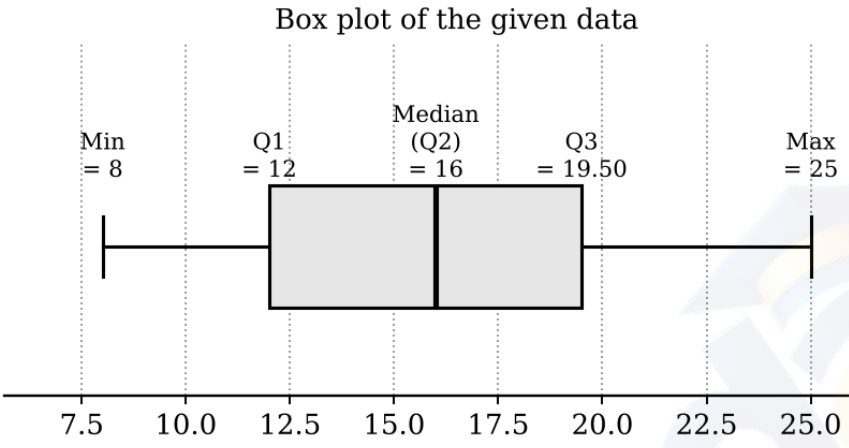


Fig.: Box plot (box from Q1 to Q3, line inside the box = median, whiskers to min and max)

(Quartiles 2 M, box plot 2 M. Quartiles are found by the (n + 1)/4 position method.)

c) Four types of data analytics. 4 M

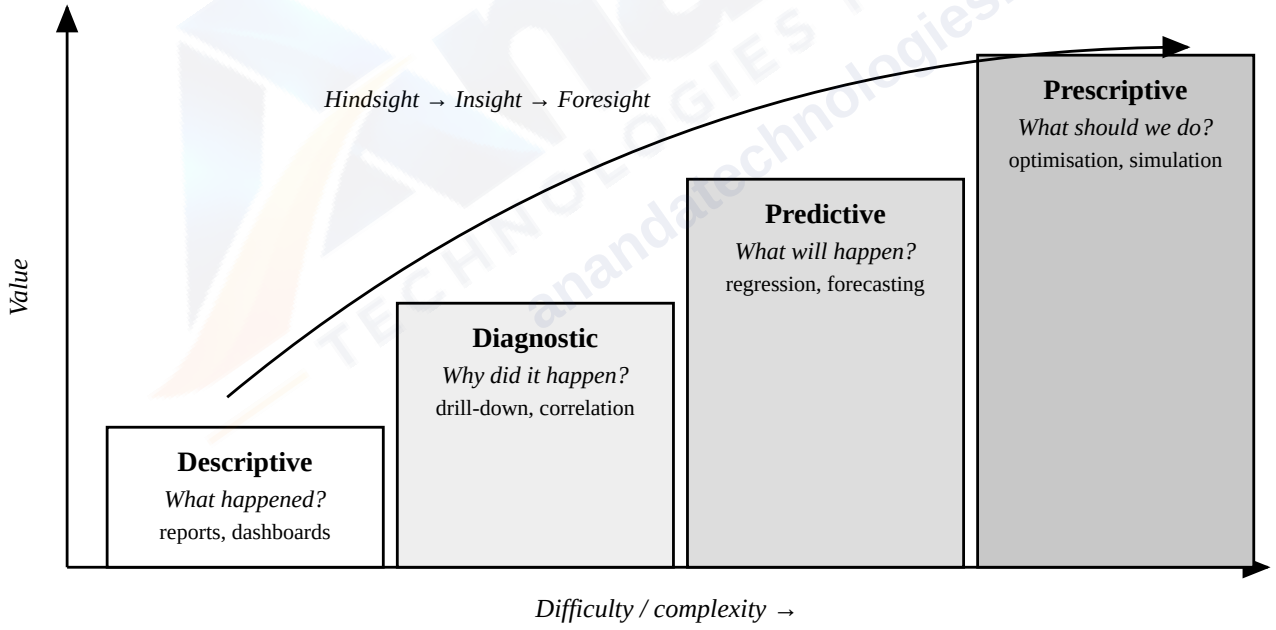


Fig.: Four types of data analytics — value and difficulty increase from descriptive to prescriptive

1. **Descriptive analytics — 'What happened?'** Summarises past data using totals, averages, percentages, charts and dashboards. Example: monthly sales report of a shop.
2. **Diagnostic analytics — 'Why did it happen?'** Finds causes of an event using drill-down, data mining and correlation. Example: finding why sales fell in March (stock shortage in one

region).

3. **Predictive analytics — 'What is likely to happen?'** Uses past data with statistical / machine-learning models such as regression to forecast. Example: forecasting next month's demand, predicting loan default.
4. **Prescriptive analytics — 'What should be done?'** Suggests the best action using optimisation and simulation. Example: deciding best price or delivery route; GPS suggesting the fastest route.

(Each type with example 1 M)

d) Exploding a pie chart slice. 4 M

Purpose: Exploding means pulling one slice out of the pie, away from the centre. It is used to **highlight / draw attention** to an important category (e.g. the largest share, or the item being discussed) without changing the data.

Steps in Excel:

1. Select the data (labels and values) and insert a pie chart: *Insert → Charts → Pie → 2-D Pie*.
2. Click once on the pie — all slices get selected. Click again on the required slice — only that slice is selected.
3. Drag the selected slice outward from the centre; **or** right-click it → *Format Data Point → Series Options* → set **Point Explosion** (e.g. 20%).
4. Release / close the pane. (Dragging after selecting the whole pie explodes all slices.) Add data labels if needed: *Chart Elements (+) → Data Labels*.

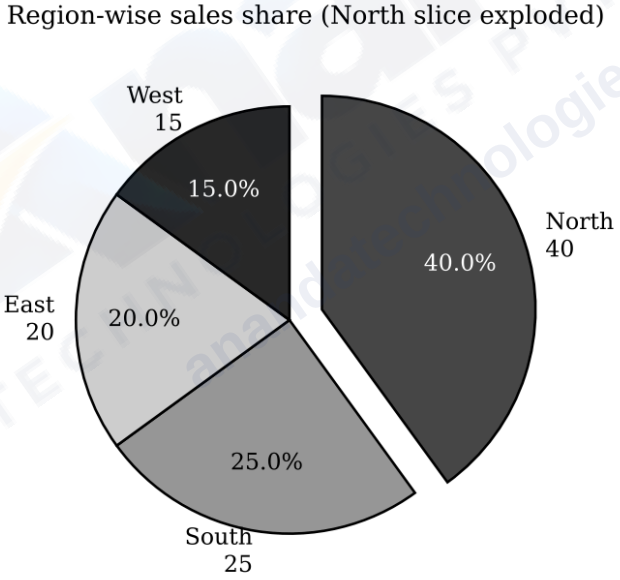


Fig.: Pie chart with one exploded slice

(Purpose 1 M, steps 3 M)

Q.3

a) Correlation; positive, negative, zero correlation. 4 M

Correlation is a statistical measure of the degree and direction of the linear relationship between two variables. It is measured by the correlation coefficient **r**, which lies between −1 and +1.

- **Positive correlation:** Both variables move in the same direction (one increases, other also increases), $0 < r \leq +1$. Example: study hours and marks.
- **Negative correlation:** Variables move in opposite directions, $-1 \leq r < 0$. Example: price of a product and its demand.
- **Zero (no) correlation:** No linear relation between the variables, $r \approx 0$. Example: shoe size and marks.

A **scatter diagram** plots each pair (x, y) as a point; the pattern of points shows the type of correlation.

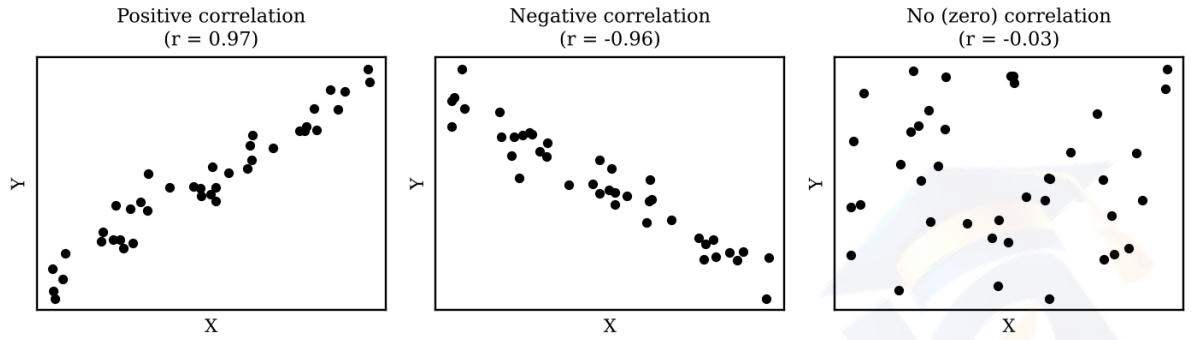


Fig.: Scatter diagrams for positive, negative and zero correlation

(Definition 1 M, three types with diagrams 3 M)

b) Steps to create a PivotTable (region-wise, product-wise sales). 4 M

Assume the data has columns Date, Region, Product and Amount with a header row (data of Q.6 c).

1. Click any cell inside the data range (or select the full range, e.g. A1:D9).
2. Go to **Insert → PivotTable**. In the *Create PivotTable* dialog check the Table/Range.
3. Choose where to place the report — **New Worksheet** (or Existing Worksheet with a cell address) and click **OK**. An empty PivotTable and the **PivotTable Fields** pane appear.
4. Drag **Region** to the **Rows** area and **Product** to the **Columns** area.
5. Drag **Amount** to the **Values** area. Excel shows *Sum of Amount* (use *Value Field Settings* to change Sum to Count/Average).
6. Format the numbers (Value Field Settings → Number Format) and give a suitable design from the *Design* tab.

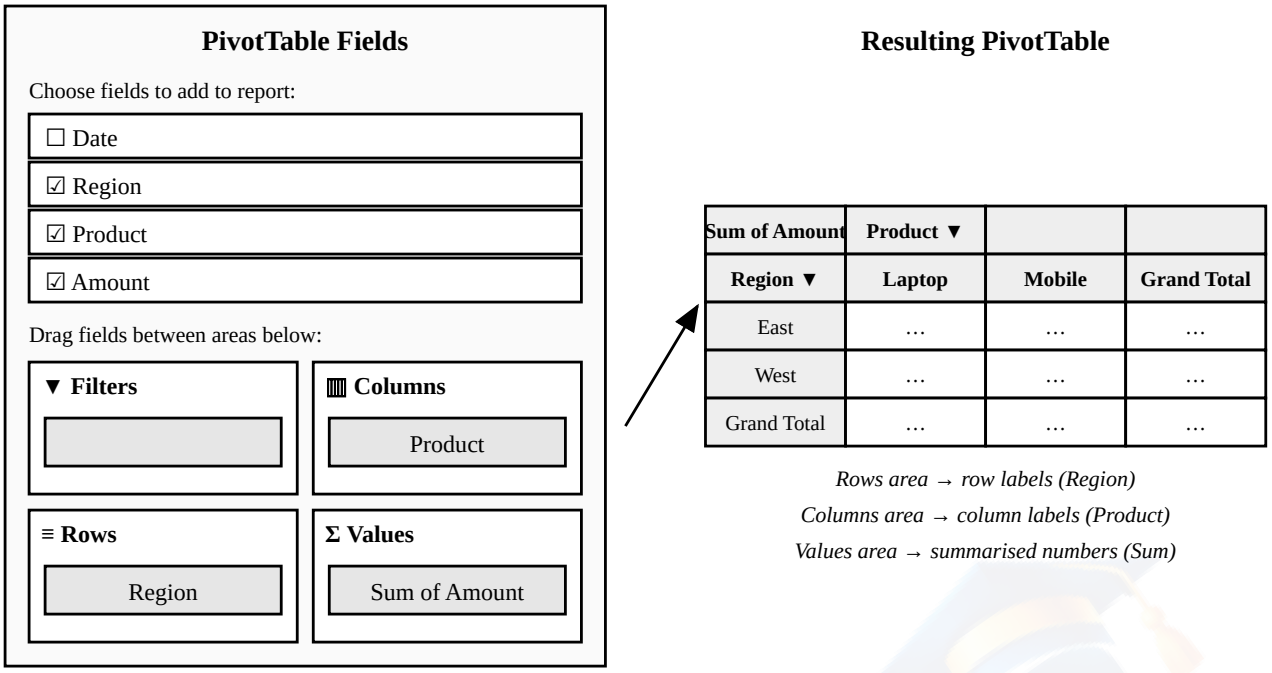


Fig.: PivotTable Fields pane — drag fields into the four areas

Resulting PivotTable (computed from the data):

Sum of Amount / Region	Laptop	Mobile	Grand Total
East	93,000	32,000	125,000
West	102,000	39,000	141,000
Grand Total	195,000	71,000	266,000

(Steps 3 M, layout / result 1 M)

c) Installing Matplotlib; four plots. 4 M

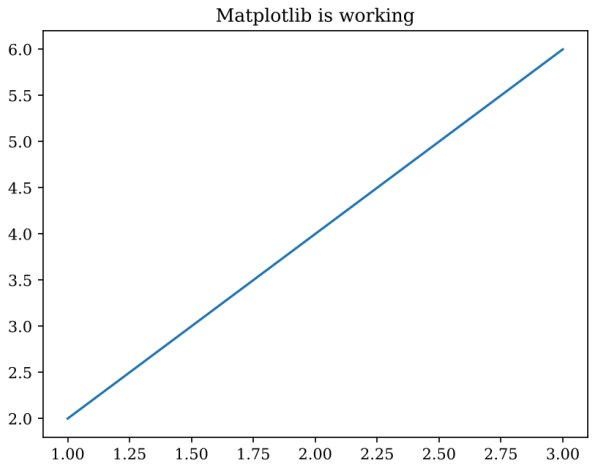
Steps:

1. Install Python (3.x) from python.org and tick *Add Python to PATH*. Check: `python --version`.
2. Make sure pip is available: `pip --version` (upgrade with `python -m pip install --upgrade pip`).
3. Install Matplotlib from the command prompt: `pip install matplotlib` (Anaconda users: `conda install matplotlib`; it is already included in Anaconda).
4. Check the installation by importing it and drawing a small plot:

```
import matplotlib
import matplotlib.pyplot as plt
print('Matplotlib version:', matplotlib.__version__)
plt.plot([1, 2, 3], [2, 4, 6])
plt.title('Matplotlib is working')
plt.show()
```

Output:

```
Matplotlib version: 3.11.2
```



Output window (plt.show())

Plots supported by Matplotlib (any four): line plot `plot()`, bar chart `bar()`/`barh()`, histogram `hist()`, scatter plot `scatter()`, pie chart `pie()`, box plot `boxplot()`, area / stack plot `stackplot()`.

(Installation steps 2 M, check 1 M, four plots 1 M)

d) Variance and standard deviation. 4 M

n = 7, Mean $\bar{x} = \Sigma x / n = 42 / 7 = 6$

x	$x - \bar{x}$	$(x - \bar{x})^2$
4	-2	4
8	2	4
6	0	0
5	-1	1
3	-3	9
7	1	1
9	3	9
$\Sigma x = 42$		$\Sigma(x - \bar{x})^2 = 28$

Variance $\sigma^2 = \Sigma(x - \bar{x})^2 / n = 28 / 7 = \mathbf{4}$

Standard deviation $\sigma = \sqrt{\sigma^2} = \sqrt{4} = \mathbf{2}$

Note: If the data is treated as a sample, $s^2 = \Sigma(x - \bar{x})^2/(n - 1) = 28/6 = 4.6667$ and $s = 2.1602$.

(Mean 1 M, table 1 M, variance 1 M, SD 1 M)

Q.4

a) Normal distribution and its properties. 4 M

A continuous random variable X follows a **normal distribution** with mean μ and standard deviation σ if its probability density function is

$$f(x) = (1 / (\sigma\sqrt{(2\pi)})) \cdot e^{-(x - \mu)^2 / (2\sigma^2)}, \quad -\infty < x < \infty$$ — written $X \sim N(\mu, \sigma^2)$

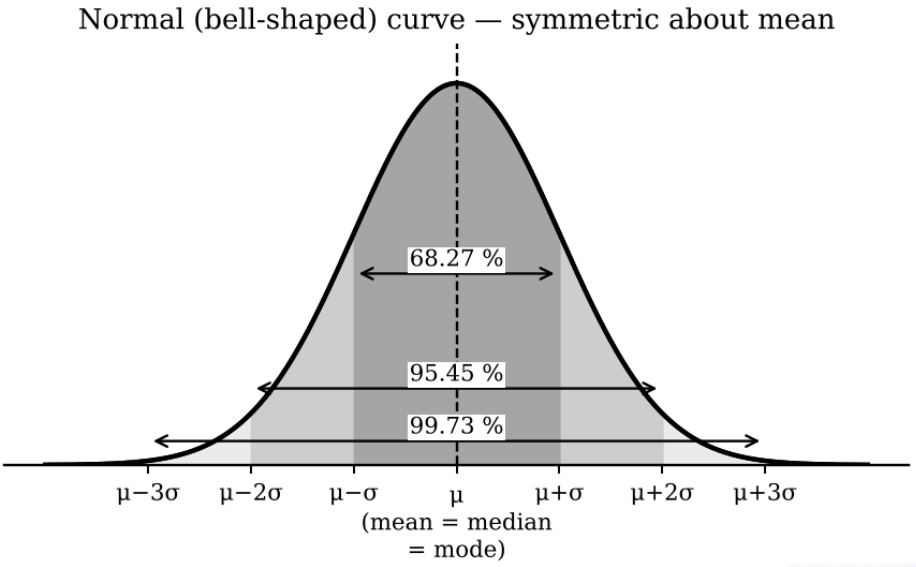


Fig.: Normal curve with areas within 1σ, 2σ and 3σ of the mean

Properties:

1. The curve is bell-shaped and symmetric about the mean μ.
2. Mean = Median = Mode (all at the centre).
3. Total area under the curve is 1; the curve never touches the x-axis (asymptotic).
4. Empirical rule: about 68.27% of values lie within μ ± σ, 95.45% within μ ± 2σ and 99.73% within μ ± 3σ.
5. Skewness = 0 and it is mesokurtic ($\beta_2 = 3$).

(Explanation 1 M, curve 1 M, properties 2 M)

b) Changing the chart type and moving the legend. 4 M

A) Changing the chart type (e.g. column chart → line chart):

1. Click the chart to select it; the *Chart Tools* (Chart Design and Format) tabs appear.
2. Click **Chart Design → Change Chart Type** (or right-click the chart → *Change Chart Type…*).
3. In the dialog choose the new type (Line, Bar, Pie …) and sub-type, see the preview and click **OK**. The data stays the same; only the way it is displayed changes.
4. To change the type of only one series (combo chart), right-click that series → *Change Series Chart Type*.

B) Moving the legend:

1. Select the chart and click the **Chart Elements (+)** button → point to **Legend** → choose Right, Top, Left or Bottom; **or** *Chart Design → Add Chart Element → Legend*.
2. For more control, right-click the legend → **Format Legend** → Legend Position. The legend can also be dragged with the mouse to any place inside the chart area.

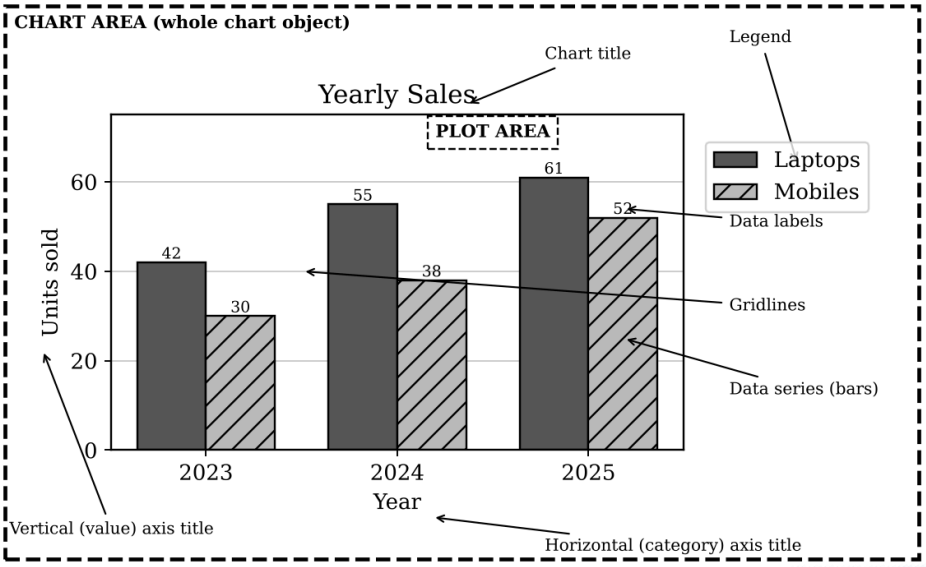


Fig.: Parts of an Excel chart (legend placed on the right)

(Changing type 2 M, moving legend 2 M)

c) Kurtosis and its types.

4 M

Kurtosis is a measure of the peakedness (and heaviness of tails) of a frequency curve compared with the normal curve. It is measured by $\beta_2 = \mu_4 / \mu_2^2$ (moment coefficient); excess kurtosis $\gamma_2 = \beta_2 - 3$.

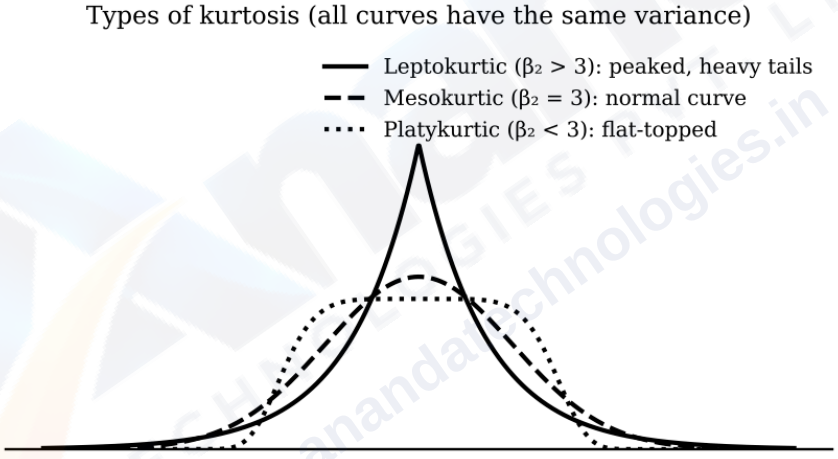


Fig.: Leptokurtic, mesokurtic and platykurtic curves

1. **Leptokurtic** ($\beta_2 > 3$): more peaked than normal, heavy tails; most values close to the mean.
2. **Mesokurtic** ($\beta_2 = 3$): normal (moderately peaked) curve.
3. **Platykurtic** ($\beta_2 < 3$): flatter than normal, thin tails; values widely spread.

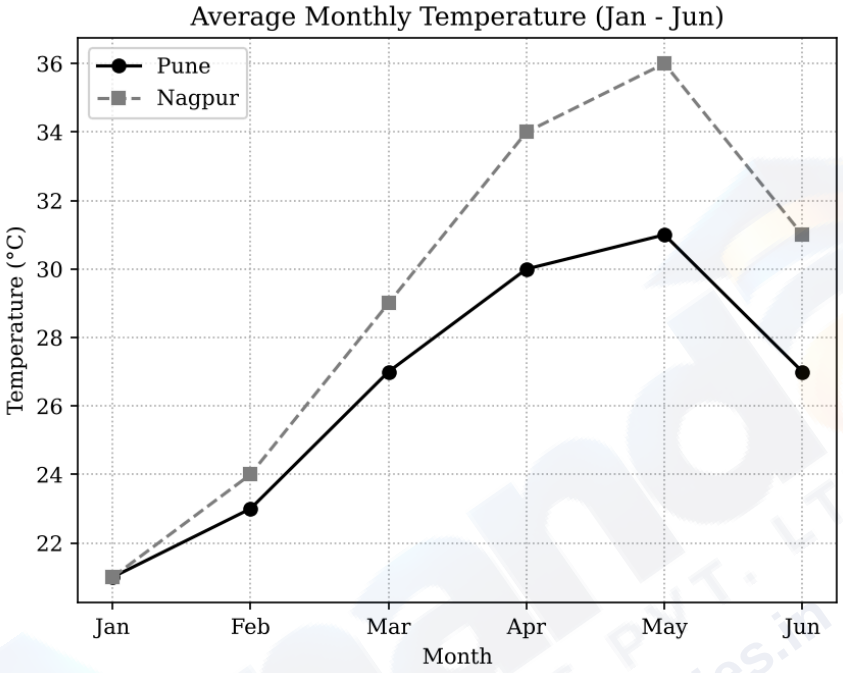
(Definition 1 M, diagram 1 M, three types 2 M)

d) Python program — line chart of temperatures.

4 M

```
# Line chart: monthly average temperature of two cities
import matplotlib.pyplot as plt

months = ['Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun']
pune   = [21, 23, 27, 30, 31, 27]
nagpur = [21, 24, 29, 34, 36, 31]
```

```
plt.plot(months, pune, marker='o', linestyle='-', color='black', label='Pune')
plt.plot(months, nagpur, marker='s', linestyle='--', color='gray',
label='Nagpur')

plt.title('Average Monthly Temperature (Jan - Jun)')
plt.xlabel('Month')
plt.ylabel('Temperature (°C)')
plt.legend()
plt.grid(True, linestyle=':')
plt.show()
```



Output window (plt.show())

(Correct program 3 M, title/labels/legend 1 M)

e) Sampling funnel. 4 M

A **sampling funnel** shows how we move step by step from the whole population to the small sample that is actually studied. At each level the number of units becomes smaller, like a funnel.

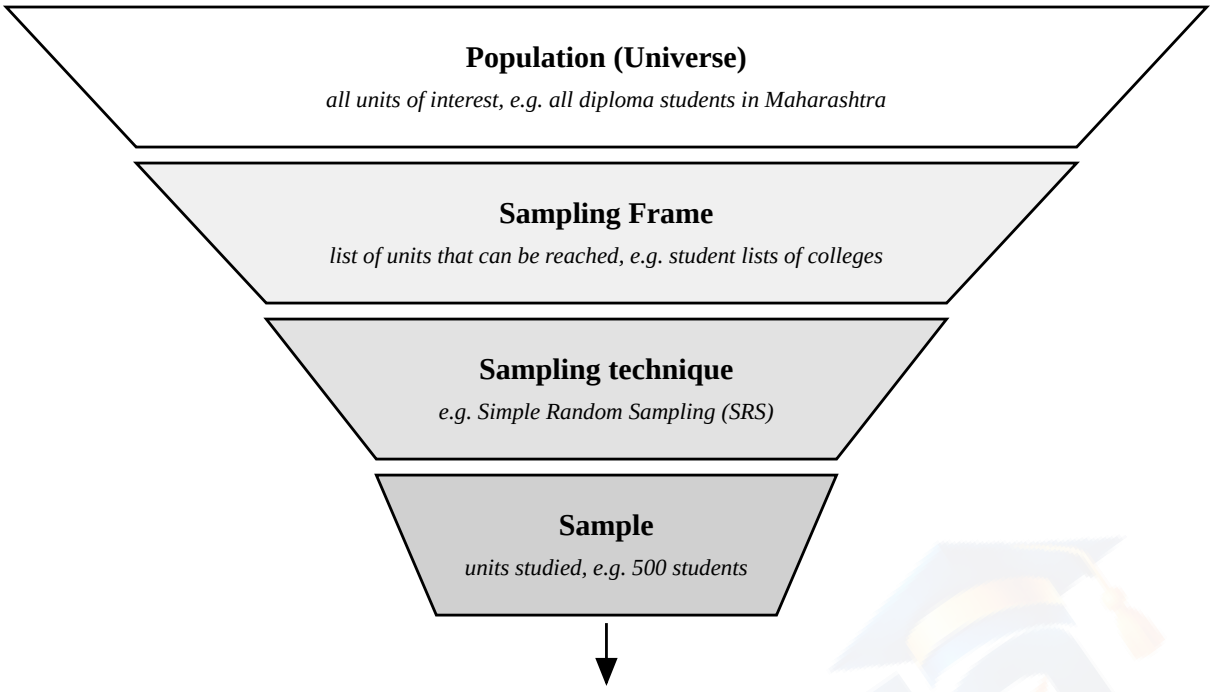


Fig.: Sampling funnel

1. **Population (universe):** the complete set of items / people about which we want information.
2. **Sampling frame:** the list or source from which units can actually be selected (e.g. college student lists). Units not in the frame cannot be selected.
3. **Sampling technique:** the method of selection — simple random, stratified, systematic, cluster sampling.
4. **Sample:** the selected units. Statistics such as $\bar{x}$ and s are calculated from it and used to estimate population parameters μ and σ.

(Diagram 2 M, explanation 2 M)

Q.5

a) Mean, median, mode and Karl Pearson's coefficient of skewness. 6 M

Mean: Mean $\bar{x} = \Sigma x / n = (12 + 15 + 18 + 15 + 20 + 22 + 15 + 25 + 18 + 30 + 16) / 11 = 206 / 11 =$ **18.73**

Median: Arranged in ascending order: 12, 15, 15, 15, 16, 18, 18, 20, 22, 25, 30
n = 11 (odd), median = ((n+1)/2)th = 6th value = **18**

Mode: 15 occurs 3 times (maximum), so **Mode = 15**

Standard deviation (needed for skewness):

x	$x - \bar{x}$	$(x - \bar{x})^2$
12	-6.73	45.2562
15	-3.73	13.8926
18	-0.73	0.5289

15	-3.73	13.8926
20	1.27	1.6198
22	3.27	10.7107
15	-3.73	13.8926
25	6.27	39.3471
18	-0.73	0.5289
30	11.27	127.0744
16	-2.73	7.4380
$\Sigma x = 206$		$\Sigma(x - \bar{x})^2 = 274.1818$

$\sigma = \sqrt{(\Sigma(x - \bar{x})^2 / n)} = \sqrt{(274.1818 / 11)} = \sqrt{24.9256} =$ **4.9926**

Karl Pearson's coefficient of skewness Sk = (Mean − Mode) / σ = (18.7273 − 15) / 4.9926 = **0.7466**

(Check using median: Sk = 3(Mean − Median)/σ = 3(18.7273 − 18) / 4.9926 = 0.4370, also positive.)

Comment: Sk is positive, so the distribution is **positively skewed** — Mean (18.73) > Median (18) > Mode (15); a few high marks (25, 30) pull the mean to the right.

(Mean, median, mode 3 M, SD 1 M, skewness 1 M, comment 1 M)

b) Chi-square test of association. 6 M

H_0 (null hypothesis): Area of residence and device used are independent (no association).
H_1: There is an association between area and device.

Expected frequency E = (Row total × Column total) / Grand total

Cell	O	E = (Row total × Column total)/N	O − E	$(O - E)^2$	$(O - E)^2/E$
Urban – Laptop	30	100×45/200 = 22.50	7.50	56.25	2.5000
Urban – Mobile	50	100×120/200 = 60	-10	100	1.6667
Urban – Desktop	20	100×35/200 = 17.50	2.50	6.25	0.3571
Rural – Laptop	15	100×45/200 = 22.50	-7.50	56.25	2.5000
Rural – Mobile	70	100×120/200 = 60	10	100	1.6667
Rural – Desktop	15	100×35/200 = 17.50	-2.50	6.25	0.3571
				$\chi^2 = \Sigma$	9.0476

χ^2 calculated = 9.048

Degrees of freedom = (r − 1)(c − 1) = (2 − 1)(3 − 1) = 2. Table value of χ^2 at 5% level for 2 d.f. = 5.991

Decision: χ^2 calculated (9.048) > χ^2 table (5.991), so H_0 is **rejected**. There is a **significant association** between area of residence and the device used for online study.

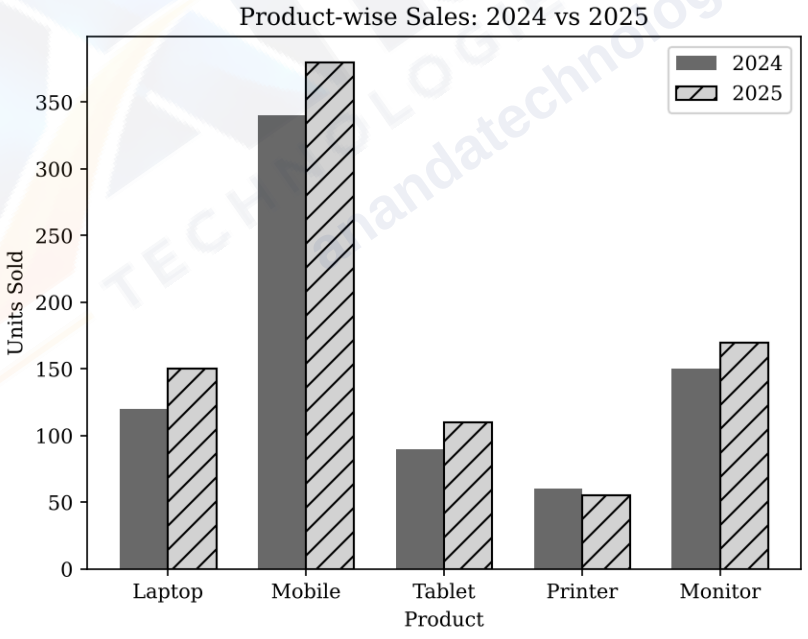
(Hypotheses 1 M, expected frequencies 2 M, χ^2 value 2 M, conclusion 1 M)

c) Python program — grouped bar chart, saved as PNG and PDF. 6 M

```
# Grouped bar chart: sales of five products in two years
import matplotlib.pyplot as plt
import numpy as np

products = ['Laptop', 'Mobile', 'Tablet', 'Printer', 'Monitor']
sales_2024 = [120, 340, 90, 60, 150]
sales_2025 = [150, 380, 110, 55, 170]

x = np.arange(len(products))        # 0, 1, 2, 3, 4
w = 0.35                             # width of each bar

plt.bar(x - w/2, sales_2024, width=w, color='dimgray', label='2024')
plt.bar(x + w/2, sales_2025, width=w, color='lightgray',
        edgecolor='black', hatch='//', label='2025')

plt.xticks(x, products)
plt.title('Product-wise Sales: 2024 vs 2025')
plt.xlabel('Product')
plt.ylabel('Units Sold')
plt.legend()

plt.savefig('sales_chart.png', dpi=150)   # PNG image
plt.savefig('sales_chart.pdf')            # PDF (vector) file
print('Chart saved as sales_chart.png and sales_chart.pdf')
plt.show()
```

Output:

```
Chart saved as sales_chart.png and sales_chart.pdf
```



Output: saved file sales_chart.png

Files created: sales_chart.pdf

Explanation: `np.arange()` gives the x positions; the two bar sets are shifted by ± w/2 so they stand side by side. `xticks()` puts the product names, `legend()` uses the *label* of each series, and `savefig()` writes the PNG and PDF files (format taken from the extension).

(Data and bars 3 M, title/labels/legend 2 M, saving 1 M)

Q.6

a) Steps to make an Excel dashboard; elements. 6 M

A **dashboard** is a single screen (sheet) that shows the key figures (KPIs) and charts of a business at a glance, with controls to filter them.

Steps:

1. **Plan:** decide the purpose, the users and the KPIs to show (e.g. total sales, orders, region-wise sales, monthly trend).
2. **Import and clean the data** into a sheet (Data tab → Get Data / copy-paste); remove duplicates, blanks and errors.
3. **Convert the data into an Excel Table** (Ctrl + T / Format as Table) so that formulas and charts grow with new data.
4. **Create the analysis** on a separate sheet — PivotTables, and formulas such as SUM, AVERAGE, SUMIFS, COUNTIFS for KPIs.
5. **Insert charts** — Pivot Charts / column, line, pie charts; use trendlines for trend analysis.
6. **Add interactive controls** — slicers and timelines (Insert → Slicer / Timeline), and connect them to all PivotTables (Report Connections).
7. **Arrange and format** everything on a 'Dashboard' sheet: title, KPI cards at the top, charts below; consistent colours; hide gridlines.
8. **Test and update:** check filters, refresh data (Data → Refresh All) and protect the sheet if needed.

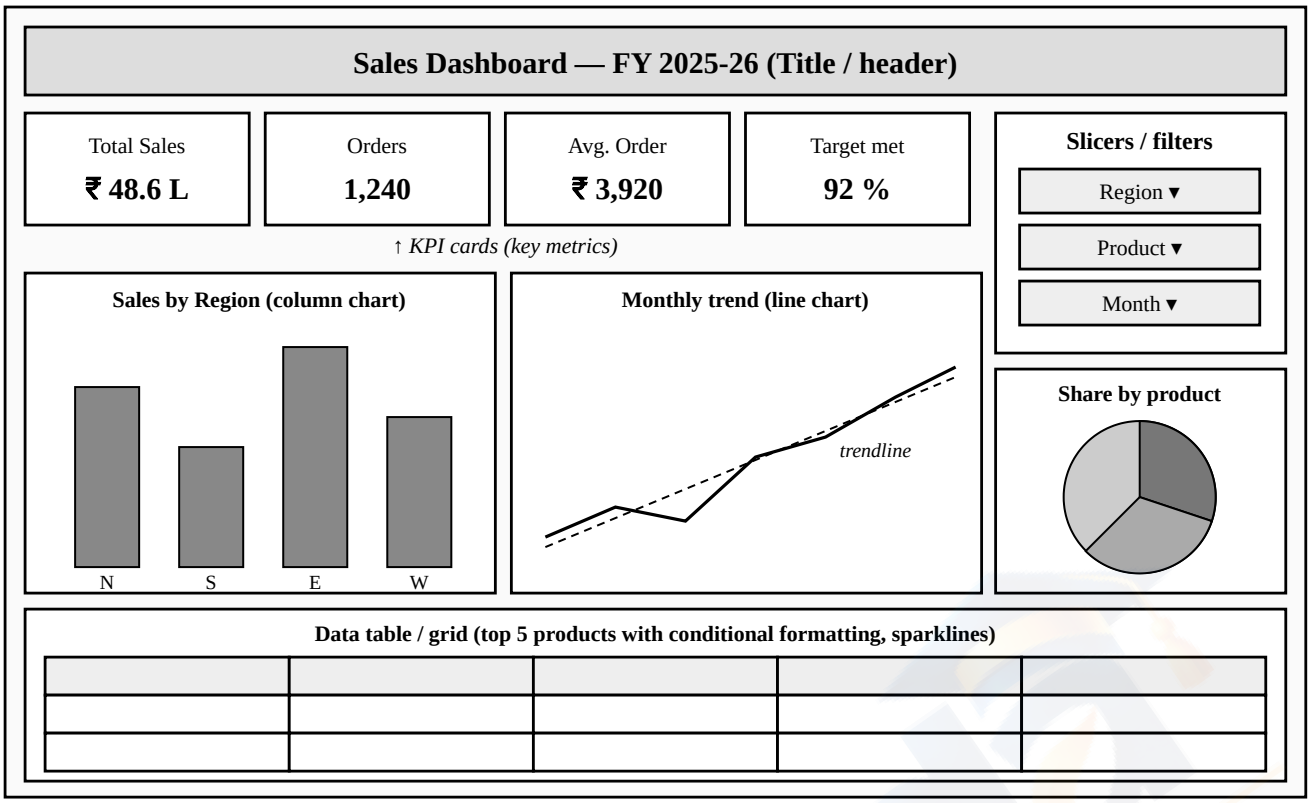


Fig.: Typical elements of an Excel dashboard (layout sketch)

Elements: title/header, KPI cards, slicers/filters, charts (column, line, pie), trendline, data table/grid with conditional formatting.

(Steps 4 M, sketch with elements 2 M)

b) Karl Pearson's correlation and regression line of y on x. 6 M

n = 5, $\bar{x}$ = 30/5 = 6, $\bar{y}$ = 160/5 = 32

x	y	dx = x − $\bar{x}$	dy = y − $\bar{y}$	dx^2	dy^2	dx·dy
2	15	-4	-17	16	289	68
4	25	-2	-7	4	49	14
6	30	0	-2	0	4	0
8	40	2	8	4	64	16
10	50	4	18	16	324	72
Σx = 30	Σy = 160			40	730	170

Correlation coefficient $r = \Sigma dx \cdot dy / \sqrt{(\Sigma dx^2 \cdot \Sigma dy^2)} = 170 / \sqrt{(40 \times 730)} = 170 / 170.8801 =$ **0.9948**

r is close to +1 → **very high positive correlation** between advertising and sales.

Regression coefficient $b_{yx} = \Sigma dx \cdot dy / \Sigma dx^2 = 170 / 40 =$ **4.2500**

Regression line of y on x: $y - \bar{y} = b_{yx}(x - \bar{x})$ → $y - 32 = 4.2500(x - 6)$ → **y = 6.5000 + 4.2500x**

Estimate for x = 12: y = 6.5000 + 4.2500 × 12 = **57.5000 ₹ lakh**

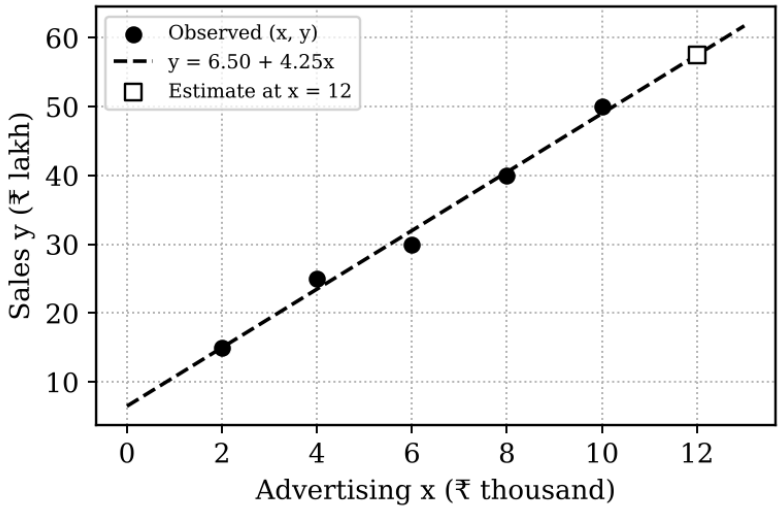


Fig.: Scatter diagram with fitted regression line

(r 3 M, regression line 2 M, estimate 1 M)

c) PivotTable with dates grouped by month; Pivot Chart. 6 M

Steps:

1. Click inside the data and choose **Insert → PivotTable → New Worksheet → OK**.
2. Drag **Date** to **Rows**, **Region** to **Columns** and **Amount** to **Values** (Sum of Amount).
3. **Group dates:** right-click any date in the row labels → **Group…** → in the Grouping dialog select **Months** (deselect Days; select Years also if data covers more than one year) → OK. (New Excel versions may group dates automatically.)
4. Click inside the PivotTable → **PivotTable Analyze → PivotChart** (or Insert → PivotChart) → choose **Clustered Column** → OK.
5. Add chart title and axis titles from *Chart Elements (+)*. Use the field buttons on the chart (Region ▾, Months ▾) to filter.
6. If the source data changes, click **Refresh** (PivotTable Analyze → Refresh) — both table and chart update.

Resulting PivotTable (Sum of Amount, ₹; computed from the given data):

Month / Region	East	West	Grand Total
Jan	60,000	18,000	78,000
Feb	48,000	73,000	121,000
Mar	17,000	50,000	67,000
Grand Total	125,000	141,000	266,000

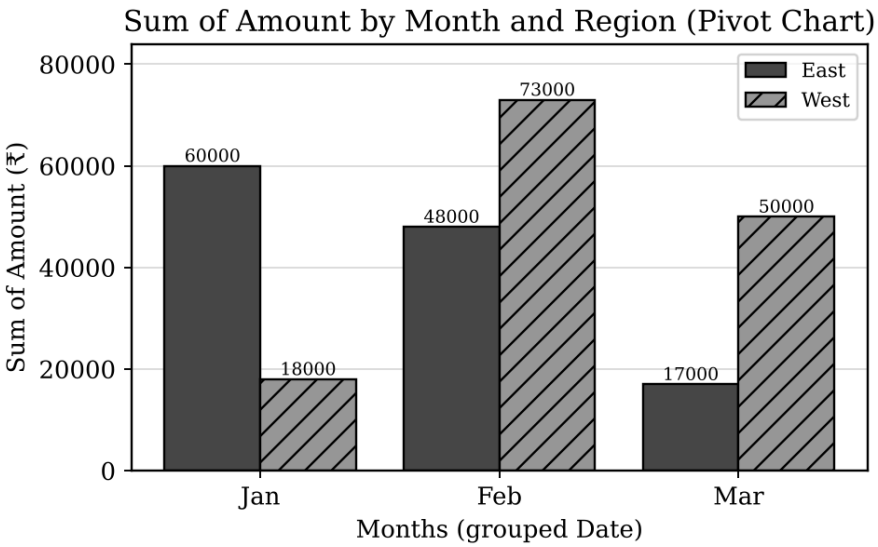


Fig.: Pivot Chart — clustered column chart of month-wise sales of each region

(Steps 3 M, PivotTable 1½ M, Pivot Chart 1½ M)

Note: This is a **model question paper prepared by Ananda Technologies** only for practice and exam preparation, based on the pattern of previous MSBTE question papers and the K-scheme syllabus. It is **not an official MSBTE paper**, and it does **not** mean that these questions will come in the exam. Answers follow standard textbook points; marks shown are indicative. Free for students — anandatechnologies.in