

ANSWER KEY

Statistics, Probability & Power BI - Evaluation (Actual Paper, Set 2)

32 questions · 9 topics · correct answers with explanations

CodeWithMunnaX

TOPICS

1. Probability & Bayes' Theorem Q20, Q25, Q29
2. Descriptive Statistics & Z-scores Q11, Q15, Q26
3. Distributions & Sampling Q9, Q16, Q19, Q21, Q23, Q24
4. Hypothesis Testing Q8, Q10, Q12, Q27, Q28
5. Model Evaluation Metrics Q30
6. Power BI & BI Tools Q1, Q2, Q4, Q7, Q14, Q17, Q18
7. Data Visualisation & Storytelling Q3, Q22
8. Seaborn & Plotly Q5, Q6, Q13
9. Case Studies (Subjective) Q31, Q32

TOPIC 1 Probability & Bayes' Theorem

3 Qs

Q20

Single choice

A grocery app logs 400 baskets. Of these, 160 contain milk and 100 contain both milk and tea. For a shopper who already has milk in the basket, what is the probability that tea is also in the basket?

- ☐ A 0.25, which is 100 out of all 400 baskets
- ☒ B 0.625, which is 100 out of the 160 milk baskets ✓ Correct
- ☐ C 0.40, which is 160 out of all 400 baskets
- ☐ D 0.375, which is the 60 milk-only baskets out of 160

EXPLANATION

Conditional probability restricts attention to the baskets where the condition holds: $P(\text{tea} \mid \text{milk}) = P(\text{tea and milk}) / P(\text{milk}) = (100/400) / (160/400) = 100/160 = 0.625$. Dividing 100 by all 400 baskets gives the joint probability, 160/400 is just the probability of milk, and 60/160 = 0.375 is the share of milk baskets that have no tea.

Q25

Single choice

A bank's model flags 12% of all loan applications as risky. Among applicants who eventually default, who make up 4% of all applicants, the model flags 75%. For an applicant who has just been flagged, what is the probability of eventually defaulting?

- ☐ A 4%, the prior default rate before flagging
- ☐ B 75%, the model's flag rate for defaulters
- ☐ C 9%, from 0.75 multiplied by 0.12
- ☒ D 25%, from 0.75×0.04 divided by 0.12

✓ Correct

EXPLANATION

Bayes' theorem gives the posterior $P(\text{default} \mid \text{flagged}) = P(\text{flagged} \mid \text{default}) \times P(\text{default}) / P(\text{flagged}) = 0.75 \times 0.04 / 0.12 = 0.25$. The 75% figure is the likelihood, which answers the reverse question; 4% is the prior before any evidence is seen; and 9% results from multiplying the likelihood by the evidence term instead of dividing.

Q29

Single choice

A message filter uses the comparison form of Bayes' theorem to decide among three categories for an incoming message. For the observed words the prior and likelihood are: Promotional (prior 0.50, likelihood 0.02), Transactional (prior 0.30, likelihood 0.05) and Phishing (prior 0.20, likelihood 0.06). Which category has the highest posterior, and why can $P(\text{evidence})$ be skipped?

- ☐ A Promotional; it has the highest prior, and the evidence term only rescales the likelihoods
- ☐ B Phishing; it has the highest likelihood, and the evidence term would cancel out all the priors
- ☒ C Transactional; the evidence term is the same for every category, so prior $\times$ likelihood decides
- ☐ D Transactional; the evidence term is largest for it, which scales the other products down

✓ Correct

EXPLANATION

When hypotheses are only being compared, the posterior is proportional to likelihood $\times$ prior because the denominator $P(\text{evidence})$ is constant across the comparison. The products are $0.50 \times 0.02 = 0.010$ for Promotional, $0.30 \times 0.05 = 0.015$ for Transactional and $0.20 \times 0.06 = 0.012$ for Phishing, so Transactional wins. The highest likelihood (Phishing) or the highest prior (Promotional) alone does not decide the result, and the evidence term does not differ between categories.

TOPIC 2 Descriptive Statistics & Z-scores

3 Qs

Q11

Single choice

Five weekly sales figures have a sample mean of 40. Four of them are 38, 45, 41 and 36. What is the fifth figure, and what does this show about dividing by $n - 1$ when computing a sample variance?

A 40; the fixed mean pins it down, yet all five values vary freely so the divisor is 5

B 40; the fixed mean pins it down, so only four values vary and the divisor is 4

✓ Correct

C Not determinable; the fifth value is free to vary, so the divisor is 5

D Not determinable; the fifth value is free to vary, so the divisor is 4

EXPLANATION

Five values with a mean of 40 must total 200, and the known four total 160, so the fifth value is forced to be 40 (the deviations from the mean always sum to zero). Once the sample mean is computed from the data, only $n - 1 = 4$ observations can vary independently while the last is constrained, which is why the sample variance divides by the degrees of freedom $n - 1$ instead of n . Treating the fifth value as free contradicts the fixed mean, and keeping five free values would not justify an $n - 1$ divisor.

Q15

Single choice

A credit model uses two features: age (mean 35 years, standard deviation 5) and monthly income (mean INR 80,000, standard deviation INR 20,000). An applicant is 45 years old and earns INR 100,000 a month. After both features are standardised, which statement is correct?

A Income is further from its mean, because its raw gap of INR 20,000 exceeds the age gap

B Age is further from its mean ($z = 2$) than income is from its own ($z = 1$)

✓ Correct

C Both sit at $z = 1$, because standardising maps every feature to the same value

D The two cannot be compared, since the units still differ after standardising

EXPLANATION

The z-score $z = (x - \mu) / \sigma$ expresses each value in units of its own standard deviation: age gives $(45 - 35) / 5 = 2$ and income gives $(100,000 - 80,000) / 20,000 = 1$. So the applicant is more extreme on age even though the raw income gap looks far larger, which is exactly why standardising puts features with different scales on

a common scale. Standardising does not force every value to the same z , and the common scale removes the unit problem.

Q26

Single choice

A hiring portal reports that the mean salary offered to its freshers last quarter was INR 9.8 lakh, while the median salary was INR 6.4 lakh. What does this gap most plausibly indicate about the salary distribution?

- ☐ A Left-skewed, with a few very low offers pulling the mean downward
- ☒ B Right-skewed, with a few very high offers pulling the mean upward ✓ **Correct**
- ☐ C Symmetric, since the mean and the median both locate the centre
- ☐ D Categorical, so only the mode is a valid measure of the centre

EXPLANATION

When the mean is larger than the median, a handful of very large values are dragging the average to the right of the typical value, which is the pattern of right-skewed data with a long tail towards the right. A left-skewed set would show the mean below the median, and a symmetric set would show the two close together. Salary is numeric, so the mean and median are both valid here and the mode is not the only usable measure.

TOPIC 3 Distributions & Sampling

6 Qs

Q9

Multiple select

A forecasting team is mixing up discrete and continuous probability tools. Which of the following statements are correct?

Note: This is a Multiple Select Question (MSQ). The answer is marked incorrect if any correct option is missed or any incorrect option is selected.

- ☐ A A PDF value at a point equals the probability that the variable takes exactly that value
- ☒ B Summing a PMF over all outcomes gives 1, and integrating a PDF over its full range also gives 1 ✓ **Correct**
- ☒ C For a continuous variable the probability of any exact single value is zero, so probabilities belong to ranges ✓ **Correct**

- D** A normal variable has its highest single-value probability at the mean, because the curve peaks there

EXPLANATION

A continuous variable has infinitely many possible values, so the probability of one exact value is zero and probability is found as the area under the curve between two points. Total probability is 1 in both cases: a PMF sums to 1 over discrete outcomes and a PDF integrates to 1 over its range. A PDF value is a density, not a probability, and the peak of the normal curve at the mean is the highest density, not a non-zero single-value probability.

Q16

Single choice

The delivery times of a fleet are roughly bell-shaped with a mean of 30 minutes and a standard deviation of 4 minutes. Using the 68-95-99.7 rule, about what share of deliveries finish between 26 and 38 minutes?

- A** About 84%

- B** About 68%

- C** About 82%

✓ Correct

- D** About 95%

EXPLANATION

26 minutes is one standard deviation below the mean and 38 minutes is two standard deviations above it. Because the distribution is symmetric, half of the 68% band (34%) lies between 26 and 30 and half of the 95% band (47.5%) lies between 30 and 38, giving 81.5%, which is about 82%. 68% would cover only 26 to 34 minutes, 95% would cover 22 to 38 minutes, and 84% is the share of deliveries above 26 minutes without any upper cut-off at 38.

Q19

Single choice

A polling team's sample of 400 respondents gives a standard error of 2.0 points on average approval. The client wants the standard error cut to 1.0 point. Assuming the population standard deviation stays the same, what sample size is needed?

- A** 1,600 respondents; standard error falls with the square root of n

✓ Correct

- B** 800 respondents; the standard error falls in direct proportion to n

- C** 1,200 respondents; each extra 400 respondents removes 0.5 points

D 3,200 respondents; the standard error falls with the cube root of n

EXPLANATION

The standard error is $\sigma/\sqrt{n}$, so it shrinks with the square root of the sample size. Halving it from 2.0 to 1.0 requires $\sqrt{n}$ to double, which means n must grow by a factor of four, from 400 to 1,600. Doubling to 800 would only reduce the standard error by a factor of about 1.41, and the other two options rely on relationships that the formula does not contain.

Q21

Single choice

To judge nationwide interest in a new online course, a startup surveys only the students who happen to walk past its Bengaluru office on a Tuesday. Which description fits this sampling plan best?

A Simple random sampling, since each student has an equal chance

B Stratified sampling, since students naturally form groups by city

C Systematic sampling, since every k th passer-by would be selected

D Convenience sampling, which is highly biased for this claim

✓ Correct

EXPLANATION

Picking whoever is easiest to reach is convenience sampling, and it is highly biased when the goal is to say something about students across the whole country because one location and one day cannot represent them. Stratified sampling needs groups that already exist in the population and a proportional draw from each. Systematic sampling needs a fixed rule such as every k th person. Simple random sampling needs every student in the population to have an equal chance of selection, which students in other cities do not.

Q23

Single choice

A team draws 500 independent samples of 64 orders each from a population of order values that is strongly right-skewed with a standard deviation of INR 40. It then plots a histogram of the 500 sample means. Which description fits that histogram best?

A Right-skewed like the population itself, with spread near the population value of 40

B Roughly bell-shaped, centred near the population mean, with spread near $40/\sqrt{64} = 5$

✓ Correct

C Roughly bell-shaped, but with spread near the population standard deviation of 40

- D** Right-skewed like the population, but with spread shrunk to about $40/\sqrt{64} = 5$

EXPLANATION

By the central limit theorem the distribution of sample means approaches a normal distribution regardless of the shape of the original population, so the histogram of means is roughly bell-shaped and centred near the population mean. Its spread is the standard error $\sigma/\sqrt{n} = 40/\sqrt{64} = 5$, far narrower than the population's 40. The other options either keep the population's skew or keep its wider spread, and averaging changes both.

Q24

Single choice

A risk team needs a probability model for each of the situations below. In which situation can the binomial formula be applied directly?

- A** Number of defaulters among 15 loans, each independently defaulting with probability 0.04 ✓ **Correct**
- B** Number of calls reaching a help desk in an hour, when only the long-run hourly average is known
- C** Delivery time of one parcel, which can take any value within a continuous range
- D** Number of buyers among 10 visitors, where every purchase raises the next visitor's chance

EXPLANATION

The binomial distribution needs binary outcomes, a fixed number of trials, a constant success probability and independent trials; 15 loans with the same 4% default probability and independent behaviour meet all four. A count of calls known only through an average is the situation for the Poisson distribution, a delivery time is a continuous variable that needs a density over ranges, and visitors whose chances change after each purchase break the independence and constant-probability requirements.

TOPIC 4 Hypothesis Testing

5 Qs

Q8

Multiple select

A school tests the claim that average marks improved this year, with H_1 stating that marks increased. The analysis returns a p-value of 0.03. Which conclusions follow?

Note: This is a Multiple Select Question (MSQ). The answer is marked incorrect if any correct option is missed or any incorrect option is selected.

- A** H_1 has been proven true beyond any doubt

B H_0 is rejected at a significance level of 0.01

C H_0 is rejected at a significance level of 0.05

✓ Correct

D H_0 is also rejected at a significance level of 0.10

✓ Correct

EXPLANATION

The decision rule rejects H_0 when the p-value is smaller than the level of significance. Since 0.03 is below both 0.05 and 0.10, H_0 is rejected at those levels, but 0.03 is not below 0.01, so H_0 is not rejected at that stricter level. A low p-value is strong evidence against H_0 ; it does not prove that H_1 is true.

Q10

Multiple select

A data team must pick a test for each study below. Which studies call for a dependent (paired) t-test?

Note: This is a Multiple Select Question (MSQ). The answer is marked incorrect if any correct option is missed or any incorrect option is selected.

A Battery life of 16 phones of one model against 16 phones of a newer model

B Hours of sleep for the same 30 patients before and after a medication

✓ Correct

C Offline and online prices recorded for the very same 25 books

✓ Correct

D Average hourly output across three separate manufacturing plants

EXPLANATION

A paired test applies when the same group is measured twice or when matched items are compared, which covers the same patients before and after medication and the same books priced offline and online. Two different sets of phones are independent samples and need a two-sample independent t-test, and three plants involve more than two groups, which calls for ANOVA.

Q12

Single choice

A pharma lab tests whether a new tablet cures a viral infection, with H_0 stating that the tablet does not cure it. The sample leads the lab to reject H_0 , yet in reality the tablet has no curing effect. What has happened, and which quantity sets how often this mistake occurs?

A A Type II error, whose probability equals the significance level α

B A Type I error, whose probability equals β , the complement of power

C A Type I error, whose probability equals the significance level α

✓ Correct

D A Type II error, whose probability equals β , the complement of power

EXPLANATION

Rejecting a null hypothesis that is actually true is a Type I error, the false-positive case of declaring an innocent person guilty, and its probability is the level of significance α . A Type II error is the opposite mistake of failing to reject a false null hypothesis, and its probability is β , where power is $1 - \beta$.

Q27

Single choice

An analyst at a quick-commerce startup wants to know whether the share of returned orders depends on the city in which the order was placed. City and return status (returned or not returned) are both categorical. Which test fits this question?

A Paired t-test, since the same orders are measured twice

B One-way ANOVA, since there are more than two cities

C Chi-square test, since both variables are categorical

✓ Correct

D One-sample t-test, since the return rate is one number

EXPLANATION

A chi-square test checks for an association between two categorical variables, which is exactly the question of whether return status is independent of city. One-way ANOVA compares the means of a numeric variable across groups, a paired t-test needs the same subjects measured under two conditions, and a one-sample t-test compares one sample mean with a benchmark value; none of these tests whether two categorical variables are related.

Q28

Single choice

An analyst wants to know whether average delivery time differs across five cities and plans to run a separate two-sample t-test for every pair of cities. How many tests would that be, and what should be used instead?

A 10 tests; a chi-square test across all five cities instead

B 5 tests; one one-way ANOVA across all five cities instead

C 10 tests; one one-way ANOVA across all five cities instead

✓ **Correct**

D 20 tests; one one-way ANOVA across all five cities instead

EXPLANATION

Five cities form $5 \times 4 / 2 = 10$ distinct pairs, so pairwise testing would need ten t-tests, and running many tests inflates the chance of a false positive. A single one-way ANOVA compares the means of more than two groups at once and avoids this multiple comparison problem. Counting 5 or 20 tests miscounts the pairs, and chi-square is a test for two categorical variables, not for comparing mean delivery times.

TOPIC 5 Model Evaluation Metrics

1 Qs

Q30

Single choice

A fraud model screens 1,000 transactions, of which 20 are fraudulent. It flags 40 transactions, and 16 of those are genuinely fraudulent. A teammate says the accuracy of 97.2% proves the model is excellent. Which assessment is most defensible?

A Accuracy misleads: flagging nothing would score 98%, while precision is 40% and recall 80%

✓ **Correct**

B Accuracy is reliable here because the dataset is large, and a precision of 40% is acceptable

C The model is poor because its recall is only 40%, even though its precision is 80%

D Accuracy misleads because the classes are balanced, so precision and recall add no information

EXPLANATION

The confusion matrix is TP = 16, FP = 24, FN = 4 and TN = 956, so accuracy is $972/1000 = 97.2\%$, precision is $16/40 = 40\%$ and recall is $16/20 = 80\%$. With only 2% fraud, a model that never flags anything would already reach 98% accuracy, so accuracy is misleading on imbalanced data and precision and recall tell the real story: the model catches most fraud but most of its flags are false alarms. Reading recall as 40% and precision as 80% swaps the two metrics, trusting accuracy because the dataset is large ignores the imbalance, and calling the classes balanced is wrong since only 2% of transactions are fraudulent.

TOPIC 6 Power BI & BI Tools

7 Qs

Q1

Multiple select

A BI developer is building a sales dashboard on a shared data source. Which statements are correct?

Note: This is a Multiple Select Question (MSQ). The answer is marked incorrect if any correct option is missed or any incorrect option is selected.

- ☒ A Renaming Cust_ID to Customer ID once in the semantic model carries to every chart built on it ✓ **Correct**
- ☐ B A worksheet is a collection of dashboards combined into one interface
- ☐ C An extract connection sends a fresh query to the source database each time a dashboard is opened
- ☒ D A postal code looks numeric but is used to slice data, so it behaves as a dimension ✓ **Correct**

EXPLANATION

Fields that look numeric, such as postal codes and order IDs, are dimensions when they are used to slice or group the data, and the semantic model applies cosmetic transformations such as renaming fields once so that they flow to all downstream visualizations. A worksheet is a single chart and a dashboard combines several worksheets, so calling a worksheet a collection of dashboards reverses the relationship. A fresh query on every view describes a live connection, whereas an extract serves stored snapshots refreshed on a schedule.

Q2

Multiple select

Priya is preparing an order file in Power BI before building the dashboard. Which of the following tasks belong in Power Query, rather than in a DAX calculation after loading?

Note: This is a Multiple Select Question (MSQ). The answer is marked incorrect if any correct option is missed or any incorrect option is selected.

- ☒ A Replacing "US" with "United States" for consistent country names ✓ **Correct**
- ☐ B Creating a Total Units Sold measure by summing one numeric column
- ☒ C Promoting the first row so that it becomes the column headers ✓ **Correct**
- ☐ D Counting unique customers with a distinct-count measure over a column

EXPLANATION

Power Query cleans, transforms and structures data before the dashboard is created, and promoting the first row to headers and find-and-replace on values are both of that kind. Summing a column into a measure and counting distinct customers are calculations done with DAX after the data is loaded.

Q4

Multiple select

A company is shortlisting desktop-authored BI tools such as Tableau and Power BI. Which of the following are core components of the architecture of such tools?

Note: This is a Multiple Select Question (MSQ). The answer is marked incorrect if any correct option is missed or any incorrect option is selected.

- A** A public viewer link that stakeholders open to interact with dashboards ✓ **Correct**
- B** Server or cloud infrastructure that manages connections and refreshes ✓ **Correct**
- C** A desktop authoring app where analysts build charts and dashboards ✓ **Correct**
- D** A notebook kernel that re-runs every chart for each person who opens it

EXPLANATION

Desktop-authored BI tools such as Tableau and Power BI have a desktop authoring app for building charts and dashboards, server or cloud infrastructure that manages data connections and automatic updates in the background, and a public viewer through which stakeholders explore dashboards without technical knowledge. Re-running a notebook for every viewer is the manual Python approach that BI tools replace, not one of their components.

Q7

Multiple select

In a Power BI sales model, a calculated column $\text{Profit} = \text{Quantity} \times \text{Unit Price} - \text{Discount} - \text{Shipping Cost}$ is added to the sales table. Total Profit is a measure that sums the Profit column, Total Sales sums a Sales column, and Profit Margin Percentage is Total Profit divided by Total Sales. The table holds two orders.

Order	Quantity	Unit Price (INR)	Discount (INR)	Shipping Cost (INR)	Sales (INR)
O1	4	500	400	100	2,000
O2	2	300	150	60	600

Which statements are correct?

Note: This is a Multiple Select Question (MSQ). The answer is marked incorrect if any correct option is missed or any incorrect option is selected.

- A** Profit Margin Percentage over both orders shows 70%, the average of the two order margins
- B** Total Profit over both orders shows INR 1,890, the sum of the two Profit values ✓ **Correct**
- C** The Profit column evaluates row by row, giving INR 1,500 for O1 and INR 390 for O2 ✓ **Correct**
- D** Profit Margin Percentage over both orders shows about 72.7%, total profit over total sales ✓ **Correct**

EXPLANATION

A calculated column is computed per row: O1 gives $4 \times 500 - 400 - 100 = 1,500$ and O2 gives $2 \times 300 - 150 - 60 = 390$. Total Profit sums them to 1,890, and the margin measure divides total profit by total sales: $1,890 / 2,600$ is about 72.7%. Averaging the two order margins (75% and 65%) to get 70% is a different calculation, because the measure aggregates profit and sales first and divides afterwards.

Q14

Single choice

In a BI tool, an analyst drags fields onto a sales chart. Which of the following fields is a measure rather than a dimension?

- A** Profit, summed across the orders in a chart ✓ **Correct**
- B** Postal code, used to slice orders by area
- C** Region name, used as the chart's column header
- D** Order ID, used to label each individual row

EXPLANATION

Measures are the numeric quantities that get aggregated with operations such as sum or average and appear as the values on an axis, and profit is exactly that. Dimensions slice and organise the data into chart headers, axis labels and groups; region, postal code and order ID all play that role, even though postal code and order ID look numeric.

Q17

Single choice

A cookie shop's CookieTypes table has three rows: Choco (100 units sold, revenue INR 20 per cookie, cost INR 12 per cookie), Oatmeal (50 units, INR 15, INR 9) and Almond (40 units, INR 25, INR 10). The measure below is placed on a card visual.

```
Total Profit = SUMX(CookieTypes, [Unit_Sold] * ([Revenue_Per_Cookie] - [Cost_Per_Cookie]))
```

What value does the card show?

A INR 1,700

✓ Correct

B INR 1,837

C INR 2,050

D INR 3,750

EXPLANATION

SUMX works row by row and then aggregates: Choco gives $100 \times 8 = 800$, Oatmeal gives $50 \times 6 = 300$ and Almond gives $40 \times 15 = 600$, so the total is INR 1,700. INR 1,837 comes from multiplying total units by an averaged margin, INR 2,050 is only the total cost, and INR 3,750 is only the total revenue.

Q18

Single choice

A retail chain's CFO opens a revenue dashboard once each morning, and the data warehouse behind it holds billions of rows shared by many teams. Separately, a control-room wall must show payment-gateway failures as they happen. Which connection plan fits both needs?

A Live connection for the CFO view; extract with a scheduled refresh for the failure wall

B Live connection for both views, so that no viewer ever sees data that is out of date

C Extract for both views, refreshed once a day, to protect the shared warehouse

D Extract with scheduled refresh for the CFO view; live only for the failure wall

✓ Correct

EXPLANATION

Extract is the sensible default: a once-a-day revenue view tolerates a scheduled lag, renders fast from stored snapshots and spares the shared warehouse from a query on every view. A live connection is justified only where real-time data is mandatory, which is the failure wall. Making both live hammers the warehouse and raises cost, making both extracts leaves the wall stale, and reversing the roles puts the real-time need on the lagging extract.

TOPIC 7 Data Visualisation & Storytelling

2 Qs

Q3

Multiple select

A designer is simplifying a cluttered chart so that more of the ink carries actual data. Which of the following changes are consistent with removing clutter?

Note: This is a Multiple Select Question (MSQ). The answer is marked incorrect if any correct option is missed or any incorrect option is selected.

- ☒ A Keep any gridlines light gray and only where reading values needs them ✓ **Correct**
- ☐ B Add a 3D effect so that the bars stand out from the background
- ☒ C Label the lines directly on the chart instead of adding a legend ✓ **Correct**
- ☐ D Wrap the plot area in a heavy border to frame the message

EXPLANATION

Direct labelling removes the legend, which forces readers to juggle between legend and graph, and light gray gridlines used sparingly keep only the ink that helps read specific values. 3D effects are hard to interpret on a 2D surface and a heavy border is unnecessary framing, so both add decoration instead of data.

Q22

Single choice

A dashboard shows profit gains in green and losses in red, with no other cue. Roughly 8% of men cannot tell the two reliably apart. Which change keeps the meaning readable without relying on colour alone?

- ☐ A Move the legend to the top-left corner of the chart
- ☐ B Use a rainbow palette with a different hue per series
- ☐ C Make the greens and reds more strongly saturated
- ☒ D Add plus and minus signs next to the values ✓ **Correct**

EXPLANATION

Meaning should never be carried by colour alone, so adding symbols such as plus and minus signs (or patterns and different line styles) lets every viewer read the gain or loss. A rainbow palette lacks contrast and

hierarchy, a stronger red-green pair is still a red-green pair, and relocating the legend does nothing for the colour-only encoding.

TOPIC 8 Seaborn & Plotly

3 Qs

Q5

Multiple select

A dataset has five numerical columns and a categorical column, region, with four values. Which statements about choosing a seaborn visual for it are correct?

Note: This is a Multiple Select Question (MSQ). The answer is marked incorrect if any correct option is missed or any incorrect option is selected.

- ☐ A A facet grid suits discovering which numeric columns correlate, because it plots every possible pair on its own
- ☒ B A facet grid on region suits comparing one chosen relationship across the four regions, with the split variable picked manually ✓ Correct
- ☒ C A pair plot suits first-look exploration because it plots every numeric column against every other automatically ✓ Correct
- ☐ D A pair plot of the five numeric columns produces ten panels, because each unique pair appears once

EXPLANATION

A pair plot plots every numerical variable against every other one without the user choosing, which makes it the quick tool for exploratory analysis, whereas a facet grid needs the analyst to choose which variable to split on and so serves a specific, predetermined comparison such as one relationship across four regions. A pair plot of five numeric columns is a 5 x 5 grid of 25 panels (the diagonal shows distributions and the mirror panels repeat each other), not ten, and the automatic every-pair behaviour belongs to the pair plot, not the facet grid.

Q6

Multiple select

A data scientist writes the following call to share an interactive view with stakeholders.

```
fig = px.scatter(df, x="usage_a", y="usage_b", color="plant_type",
                 size="usage_c", hover_name="customer",
                 animation_frame="month", animation_group="customer")
```

Which statements about this call are correct?

Note: This is a Multiple Select Question (MSQ). The answer is marked incorrect if any correct option is missed or any incorrect option is selected.

- ☒ **A** The size argument scales each marker by the numerical column usage_c ✓ **Correct**
- ☐ **B** The hover_name argument colour-codes every point by the customer name
- ☒ **C** The points are colour-coded by the values in the plant_type column ✓ **Correct**
- ☒ **D** animation_group keeps each customer's point the same object across the frames ✓ **Correct**

EXPLANATION

In Plotly Express, color colour-codes points by a categorical column, size sets the marker size from a numerical variable (here usage_c), and animation_group keeps each object's identity across animation frames while animation_frame animates the figure over the values of month. hover_name only controls the name displayed when a user hovers over a point; it does not colour anything.

Q13

Single choice

A data scientist runs the call below on a DataFrame that has four numeric columns plus two categorical columns, region and plant_type.

```
sns.pairplot(df, hue="plant_type", corner=True)
```

How many subplot panels does the resulting figure contain?

- ☐ **A** 6 panels: only the unique off-diagonal scatters remain
- ☐ **B** 16 panels: corner=True only changes the colours of the grid
- ☐ **C** 36 panels: all six columns are plotted against each other
- ☒ **D** 10 panels: 4 diagonal plots plus 6 lower-triangle scatters ✓ **Correct**

EXPLANATION

A pair plot plots every numerical variable against every other numerical variable, so the four numeric columns give a 4 x 4 grid of 16 panels and the categorical columns only supply the colour grouping through hue. The diagonal shows each variable's distribution rather than a scatter plot, and corner=True removes the redundant mirror panels above it. That leaves the 4 diagonal panels plus 6 unique lower-triangle scatters, which is 10 panels.

TOPIC 9 Case Studies (Subjective)

2 Qs

Q31

Subjective

Problem Statement

Roadrunner Logistics runs two hubs, Pune and Chennai, and delivers with vans and bikes. The operations head wants to understand how distance, load and delivery time interact, whether the distance-to-time relationship differs by hub, and how to present the findings to the CEO and to three managers who will watch a daily dashboard.

Dataset

Twelve delivered routes:

hub	vehicle	distance_km	load_kg	delivery_min
Pune	Van	4.2	120	22
Pune	Van	8.5	180	31
Pune	Van	12.0	150	38
Pune	Bike	2.1	15	12
Pune	Bike	5.0	20	19
Pune	Bike	7.4	25	26
Chennai	Van	3.8	110	25
Chennai	Van	9.1	200	36
Chennai	Van	14.5	160	47
Chennai	Bike	1.9	10	11
Chennai	Bike	4.6	18	17
Chennai	Bike	6.8	22	23

Ready-to-run initializer (identical to the table above):

```
import pandas as pd
import seaborn as sns

df = pd.DataFrame({
    "hub": ["Pune"] * 6 + ["Chennai"] * 6,
    "vehicle": ["Van", "Van", "Van", "Bike", "Bike", "Bike"] * 2,
    "distance_km": [4.2, 8.5, 12.0, 2.1, 5.0, 7.4, 3.8, 9.1, 14.5, 1.9, 4.6, 6.8],
    "load_kg": [120, 180, 150, 15, 20, 25, 110, 200, 160, 10, 18, 22],
    "delivery_min": [22, 31, 38, 12, 19, 26, 25, 36, 47, 11, 17, 23],
})
```

Tasks

1. **Pair plot.** Write seaborn code that plots every numeric column of `df` against every other, colouring the points by `vehicle` and removing the redundant mirror panels. State how many panels the figure contains and what the diagonal panels show.

2. **Facet grid with styling.** Write seaborn code that draws `delivery_min` against `distance_km` as a scatter, with one column of panels per `hub` and points coloured by `vehicle`. Then apply a clean background style, a presentation-sized font context and removal of the plot borders.
3. **Choosing the visual.** The operations head asks whether the distance-to-time relationship differs between the two hubs. Explain why a facet grid suits this question better than a pair plot, when the pair plot is the better tool, and one trade-off of the facet grid.
4. **CEO view.** For a CEO-facing version of the Task 2 chart, write one quantifiable sentence the chart should deliver using numbers from the dataset, a headline-style title, what you would highlight in an accent colour and what you would make gray, two clutter elements you would remove, and one change that keeps the chart readable for colour-blind viewers.
5. **Dashboard design.** The three managers will watch the same view in a BI tool every morning. Choose between a live connection and an extract and justify the choice using refresh needs and database load. Classify `hub`, `vehicle` and `delivery_min` as dimensions or measures and name an aggregation for the measure. Say where renaming `delivery_min` to "Delivery Time (min)" should be done and why it needs to be done only once.

Submission Format

Paste your code for Tasks 1 and 2 and write Tasks 3 to 5 as structured text, all in the input box, clearly labelled Task 1 to Task 5.

Note: You may use any IDE (online: Google Colab, Jupyter Notebook; offline: VS Code, PyCharm, or similar) to write, run, and test your code. You must **disable all AI-based extensions** (GitHub Copilot, Tabnine, Codeium, etc.) before starting. AI tools of any kind are not permitted and will lead to disqualification. **Submit your complete code as a copy-pasted snippet in the input box — do not submit a URL, notebook link, GitHub repository, or Google Drive file.**

MODEL ANSWER

Question Explanation

This question tests whether you can use seaborn's pair plot and facet grid correctly, pick the right multi-dimension visual for a question, apply the design principles that make a chart land with a CEO, and make sound BI-tool decisions about connection type and the semantic model.

Short Explanation

Use `sns.pairplot(df, hue="vehicle", corner=True)` for 6 panels, `sns.relplot` with `col="hub"` and `hue="vehicle"` for the hub comparison, style it with `sns.set_style`, `sns.set_context` and `sns.despine`, pick the facet grid for a specific comparison, simplify the CEO chart around one quantified message, and use an extract with a semantic-model rename for the dashboard.

Long Explanation

Task 1: a pair plot plots every numerical variable against every other, so the three numeric columns give a 3 x 3 grid; hue colours by vehicle, and `corner=True` removes the mirror panels above the diagonal, leaving 3 diagonal distribution panels and 3 lower-triangle scatters, 6 panels in total. Task 2: `sns.relplot` handles scatter plots with a facet split: `col="hub"` creates one column of panels per hub and `hue="vehicle"` adds colour. `sns.set_style` sets the background (for example `whitegrid`), `sns.set_context` scales fonts for a presentation (talk) and `sns.despine` removes borders. Task 3: the question is about one chosen relationship across a category, which is what a facet grid is built for, since the analyst picks the split variable; a pair plot is automatic and best for exploration when no relationship has been chosen yet; the trade-off is that many categories would create too many panels. Task 4: a CEO needs the conclusion and the key number, so the chart should make exactly one point, for example that bikes finish every route of up to 7.4 km within 26 minutes while vans on routes of 12 km or more need 38 to 47 minutes. The title is a headline stating that

conclusion, the vans' long-distance points get an accent colour while everything else is gray, clutter such as borders, 3D effects and heavy gridlines is removed (and the legend replaced by direct labels), and plus/minus symbols, patterns or different marker shapes keep it readable for colour-blind viewers. Task 5: a morning view tolerates a scheduled lag, and an extract renders fast from stored snapshots while sparing the database from a live query on each view, so extract is the right default. hub and vehicle are dimensions that slice the data; delivery_min is a measure aggregated with an average. The rename belongs in the semantic model, which is built once and applies automatically to every downstream visualization.

Model Solution

```
import pandas as pd
import seaborn as sns

df = pd.DataFrame({
    "hub": ["Pune"] * 6 + ["Chennai"] * 6,
    "vehicle": ["Van", "Van", "Van", "Bike", "Bike", "Bike"] * 2,
    "distance_km": [4.2, 8.5, 12.0, 2.1, 5.0, 7.4, 3.8, 9.1, 14.5, 1.9, 4.6, 6.8],
    "load_kg": [120, 180, 150, 15, 20, 25, 110, 200, 160, 10, 18, 22],
    "delivery_min": [22, 31, 38, 12, 19, 26, 25, 36, 47, 11, 17, 23],
})

# Task 1: pair plot, coloured by vehicle, mirror panels removed
sns.pairplot(df, hue="vehicle", corner=True)

# Task 2: facet grid with styling
sns.set_style("whitegrid")
sns.set_context("talk")
sns.relplot(data=df, x="distance_km", y="delivery_min", col="hub", hue="vehicle")
sns.despine()
```

Task 1 answer. 6 panels: 3 diagonal panels showing the distribution of each numeric variable plus 3 lower-triangle scatters.

Task 3. A facet grid compares one chosen relationship (distance vs delivery time) across a category (hub), with the split variable chosen manually. A pair plot is automatic and best for exploration. Trade-off: too many panels with many categories.

Task 4. Message: bikes finish every route up to 7.4 km within 26 minutes, while vans on routes of 12 km or more need 38 to 47 minutes. Headline title stating that conclusion; van long-route points in an accent colour with everything else in gray; remove borders and 3D effects and label directly instead of using a legend; add marker shapes or symbols so colour is not the only cue.

Task 5. Extract (scheduled refresh is enough; lighter load on the database). hub and vehicle are dimensions; delivery_min is a measure, aggregated as an average. Rename in the semantic model, which is built once and applies to every downstream visualization.

Q32

Subjective

Scenario

SwiftBite, a quick-delivery startup, advertises "groceries in 15 minutes". Its analyst audits the claim with three pieces of evidence. Use a significance level of $\alpha = 0.05$ throughout.

Data

Evidence 1: delivery times. A random sample of 16 orders has a mean delivery time of $\bar{x} = 17.2$ minutes and a sample standard deviation of $s = 4.4$ minutes. The critical t-value at 15 degrees of freedom and $\alpha = 0.05$ is ± 2.131 for a two-tailed test and 1.753 for a one-tailed test.

Evidence 2: rider training. The same six riders were timed on a comparable route before and after a training programme.

Rider	Before (min)	After (min)
R1	18	12
R2	16	14
R3	20	17
R4	15	14
R5	19	15
R6	14	12

For a one-tailed test at $\alpha = 0.05$ with 5 degrees of freedom, the critical t-value is 2.015.

Evidence 3: late-delivery classifier. Of 200 orders, 20 were actually late. A model flagged 30 orders as late, and 15 of those flagged orders were actually late.

Tasks

- Two-tailed test.** The company claims the true mean delivery time is 15 minutes. State H_0 and H_1 , give the degrees of freedom, compute the standard error and the t-statistic, compare it with the critical value, and state your decision. Say what the decision does and does not tell the company.
- Directional re-test.** The operations head now says the real worry is deliveries taking longer than 15 minutes. Restate H_0 and H_1 for this directional claim, decide using the one-tailed critical value, and explain why the decision differs from Task 1 on identical data by describing where the rejection region sits in each design. State the probability of a Type I error under each design, and say at which step of the testing process the direction must be fixed.
- Rider training.** Name the appropriate test for Evidence 2 and justify why an independent two-sample test is unsuitable. State H_0 and H_1 for the claim that training reduces delivery time, compute the six differences (before minus after) with their mean and sample standard deviation, treat the differences as a single sample tested against 0 using the one-sample t formula, compute t, compare it with 2.015, and conclude in business terms.
- Classifier check.** Using Evidence 3, build the confusion matrix and compute accuracy, precision and recall. Show why accuracy alone is misleading here, and map false positives and false negatives to Type I and Type II errors, taking "the order is on time" as H_0 .

Submission Format

Write a structured response in the input box, clearly labelled Task 1 to Task 4, with all working shown.

Note: Write your answer in clear, concise English. Bullet points and structured sections are encouraged. Do not paste external links or screenshots — write the full response in the input box provided. AI tools of any kind are not permitted and will lead to disqualification.

MODEL ANSWER**Question Explanation**

This question tests whether you can run and interpret a one-sample t-test, see how the choice of one-tailed or two-tailed framing changes a borderline decision, recognise when a paired test is required, and read a confusion matrix on imbalanced data.

Short Explanation

Task 1 gives $t = 2.0$ against ± 2.131 so H_0 is not rejected; Task 2 gives 2.0 against 1.753 so H_0 is rejected; Task 3 is a paired test with t of about 4.11 against 2.015 so training works; Task 4 gives accuracy 90%, precision 50% and recall 75%, with accuracy no better than a model that flags nothing.

Long Explanation

Task 1: $H_0: \mu = 15$ and $H_1: \mu \neq 15$, with $df = 16 - 1 = 15$. The standard error is $4.4/\sqrt{16} = 1.1$ and $t = (17.2 - 15)/1.1 = 2.0$. Because $|2.0|$ is smaller than 2.131 , the statistic is inside the acceptance region and H_0 is not rejected; this means the sample does not give strong enough evidence against the 15-minute claim, not that the claim is proven. Task 2: with $H_1: \mu > 15$ the whole 5% rejection region sits in the upper tail, so the cut-off is 1.753 and $t = 2.0$ exceeds it, so H_0 is rejected. In the two-tailed design the 5% is split 2.5% per tail, which moves each cut-off out to 2.131 , and the same data now falls short. The Type I error probability is $\alpha = 5\%$ in both designs, and because hypotheses are framed first in the testing process, the direction has to be fixed before the test statistic is calculated. Task 3: the same riders are measured twice, so the samples are dependent and a paired t-test is used; an independent test would ignore that each pair belongs to one rider. With H_0 : no reduction and H_1 : training reduces time, the differences (before minus after) are 6, 2, 3, 1, 4 and 2, with mean 3.0. The squared deviations from 3.0 are 9, 1, 0, 4, 1 and 1, which sum to 16, so the sample variance is $16/5 = 3.2$ and s is about 1.79. Then $t = 3.0/(1.79/\sqrt{6})$ is about 4.11, which exceeds 2.015, so H_0 is rejected: the training is linked to significantly shorter delivery times. Task 4: $TP = 15$, $FP = 30 - 15 = 15$, $FN = 20 - 15 = 5$ and $TN = 200 - 15 - 15 - 5 = 165$. Accuracy = $(15 + 165)/200 = 90\%$, precision = $15/30 = 50\%$ and recall = $15/20 = 75\%$. A model flagging nothing would be right on all 180 on-time orders and also score 90%, so accuracy hides the weak precision. With H_0 as the order is on time, a false positive (flagging an on-time order) is a Type I error and a false negative (missing a late order) is a Type II error.

Model Solution

Task 1. $H_0: \mu = 15$ min; $H_1: \mu \neq 15$ min. $df = 15$. $SE = 4.4/\sqrt{16} = 1.1$. $t = (17.2 - 15)/1.1 = 2.0$. Since $|2.0| < 2.131$, fail to reject H_0 at 5%. The sample does not give enough evidence against the 15-minute promise, but this does not prove the true mean is exactly 15 minutes.

Task 2. $H_1: \mu > 15$ min ($H_0: \mu = 15$ min). The one-tailed critical value is 1.753 and $t = 2.0 > 1.753$, so reject H_0 . The two-tailed test splits the 5% into 2.5% in each tail, pushing the cut-off to 2.131 ; the one-tailed test puts all 5% in the upper tail, so a smaller cut-off applies. Type I error probability is 5% in both designs. The hypotheses, including the direction, are framed first, before the test statistic is computed.

Task 3. Dependent (paired) t-test, because the same riders are measured twice; an independent test ignores the pairing. H_0 : no reduction; H_1 : training reduces time. Differences (before - after): 6, 2, 3, 1, 4, 2; mean = 3.0; $s \approx 1.79$. $t = 3.0/(1.79/\sqrt{6}) \approx 4.11 > 2.015$, so reject H_0 : training is associated with significantly shorter delivery times.

Task 4. TP = 15, FP = 15, FN = 5, TN = 165. Accuracy = 90%, precision = 50%, recall = 75%. A model that flags nothing also scores 90% accuracy, so accuracy is misleading on this imbalanced data. False positive = Type I error; false negative = Type II error.

Statistics, Probability & Power BI - Evaluation (Actual Paper, Set 2) · Answer Key · CodeWithMunnaX