INTERVIEW PREP GUIDE

01 SQL

What Gets Tested

Topic	Interview Frequency
Window functions — RANK, ROW_NUMBER, DENSE_RANK, LAG, LEAD	Very High
JOINs — especially LEFT JOIN edge cases and row count validation	Very High
GROUP BY + HAVING	High
Subqueries & CTEs	High
Date/time functions — EXTRACT, DATE_TRUNC, DATEDIFF, DATE_FORMAT	High
Running totals & month-over-month growth	Medium
Data validation after JOINs — nulls, duplicates, shape checks	Medium

The Interview Approach

1. **Clarify:** Repeat the question. Confirm table names, key columns, and edge cases.
2. **Pseudocode:** Talk through your logic before writing a single line.
3. **Write:** Code cleanly. Alias every table. Format consistently.
4. **Test:** Walk through a mental example. Check NULLs, ties, and empty results.
5. **Explain:** Tell them what the query returns and why — not just how.

Practice Questions

Q1. Top 3 products by revenue in each category

Pattern: Window function + PARTITION BY

```
SELECT category, product_name, revenue
FROM (
  SELECT category, product_name, revenue,
         RANK() OVER (PARTITION BY category ORDER BY revenue DESC) AS rnk
  FROM sales
) t
WHERE rnk <= 3;

-- Use RANK (not ROW_NUMBER) to handle ties correctly.
-- Use DENSE_RANK if you never want gaps in rank numbers.
```

Q2. Month-over-month revenue growth

Pattern: LAG() for time-series comparison

```
SELECT
  month,
  revenue,
  LAG(revenue) OVER (ORDER BY month) AS prev_month_revenue,
  ROUND(
    (revenue - LAG(revenue) OVER (ORDER BY month))
    / LAG(revenue) OVER (ORDER BY month) * 100, 2
  ) AS mom_growth_pct
FROM monthly_revenue;
```

Q3. Users active in Jan but not in Feb

Pattern: NOT IN subquery — classic retention investigation

```
SELECT DISTINCT user_id
FROM orders
WHERE DATE_FORMAT(order_date, '%Y-%m') = '2024-01'
  AND user_id NOT IN (
    SELECT user_id FROM orders
    WHERE DATE_FORMAT(order_date, '%Y-%m') = '2024-02'
  );

-- Follow-up: What % of Jan users were lost?
-- Wrap in a CTE and divide lost users by total Jan users.
```

Q4. 7-day rolling average of daily signups

Pattern: Rolling window aggregate

```
SELECT
  signup_date,
  signups,
  AVG(signups) OVER (
    ORDER BY signup_date
    ROWS BETWEEN 6 PRECEDING AND CURRENT ROW
  ) AS rolling_7d_avg
FROM daily_signups;
```

Q5. Find and remove duplicate orders

Pattern: Data quality check — GROUP BY + HAVING, then deduplication

```
-- Find duplicates
SELECT order_id, COUNT(*) AS cnt
FROM orders
GROUP BY order_id
HAVING COUNT(*) > 1;

-- Keep only the most recent record per order
DELETE FROM orders
WHERE id NOT IN (
  SELECT MAX(id) FROM orders GROUP BY order_id
);
```

> **What interviewers actually care about in SQL rounds**
>
> Can you translate a business question into a query? Can you explain WHAT your query returns, not just HOW it works? Can you handle NULLs, ties, and edge cases without being prompted? Those three things separate candidates.

02 Python & EDA

Pandas — Most Tested Operations

Operation	Code Pattern
Filter rows	df[df['col'] > value]
GroupBy + aggregate	df.groupby('col').agg({'sales': 'sum'})
Merge two DataFrames	pd.merge(df1, df2, on='id', how='left')
Handle missing values	df['col'].fillna(df['col'].median(), inplace=True)
Apply custom function	df['col'].apply(lambda x: ...)
Detect duplicates	df.duplicated().sum()
Sort values	df.sort_values('revenue', ascending=False)
Select specific columns	df[['col1', 'col2']]
Rename columns	df.rename(columns={'old': 'new'}, inplace=True)
Reset index	df.reset_index(drop=True, inplace=True)

Common Interview Questions — Python

Q1. Clean a messy dataset

Walk through this order — interviewers want to see a systematic approach:

```
import pandas as pd

# Step 1: Understand the data
df.info()               # dtypes, null counts
df.describe()           # distributions
df.duplicated().sum()   # duplicate rows

# Step 2: Remove duplicates
df = df.drop_duplicates()

# Step 3: Handle nulls — think before you drop
df['revenue'].fillna(df['revenue'].median(), inplace=True)
df.dropna(subset=['user_id'], inplace=True)  # can't have null IDs

# Step 4: Fix data types
df['date'] = pd.to_datetime(df['date'])
df['price'] = df['price'].astype(float)
```

Q2. Walk me through how you'd do EDA on a new dataset

The framework — use this structure every time:

6. **Define the question:** What business problem am I trying to answer?
7. **Univariate analysis:** Distribution, central tendency, spread, and outliers for each key column.
8. **Bivariate analysis:** Correlations, scatter plots, cross-tabs. Is revenue correlated with discount?
9. **Multivariate analysis:** Segment breakdowns — revenue by region AND category together.
10. **Interpret:** What did I find? What new questions does this raise?

Q3. Find the top 10% of customers by spend

```
threshold = df['total_spend'].quantile(0.90)
top_customers = df[df['total_spend'] >= threshold]
print(f'{len(top_customers)} customers in top 10%')
```

Visualisation — Choosing the Right Chart

Chart Type	Best Used For
Bar chart	Comparing values across categories (revenue by region)
Line chart	Trends over time (DAU, MRR, weekly signups)
Histogram	Distribution of a single numeric variable
Boxplot	Outlier detection and spread comparison across groups
Heatmap	Correlation matrix or null value map across columns
Scatter plot	Relationship between two numeric variables

> **Interview tip — Python**
> Don't memorise syntax. Know the logic. If you forget .fillna(), say 'I'd impute with the median using Pandas fillna.' Interviewers care that you know WHY, not just what to type.

03 Excel & Power BI

Excel — What Gets Tested

Function / Feature	What Interviewers Ask
INDEX + MATCH	Why is this better than VLOOKUP? When do you use it?
VLOOKUP	Debug: why is this returning #N/A?
SUMIFS / COUNTIFS	Total revenue for Region=North AND Q3 only
Pivot Tables	Build a revenue-by-city-by-category summary from raw data
Nested IFs / IFS function	Bucket customers into High / Medium / Low value tiers
Conditional Formatting	Highlight top 10% of values or flag cells below threshold
Data Validation	Prevent incorrect entries in a shared reporting template
EOMONTH / DATEDIF	Calculate tenure, subscription age, days to renewal

The VLOOKUP vs INDEX-MATCH Answer

You will get asked this. Have a clear answer ready:

> **Sample Answer**
> VLOOKUP locks you into looking left-to-right from the first column — if the key column isn't leftmost, it breaks. INDEX-MATCH works in any direction, handles column insertions without breaking, and is faster on large datasets. I default to INDEX-MATCH for anything beyond a quick lookup.

VLOOKUP #N/A Debugging Checklist

- **Trailing spaces:** Use TRIM() on the key column in both sheets.
- **Data type mismatch:** One column is text, the other is a number. Force consistent types.
- **Wrong column index:** Count from the lookup column, not column A.
- **Approximate match on:** The last argument should be 0 (exact match) for most use cases.
- **Lookup range not fixed:** Use $ to lock the range when dragging the formula down.

Power BI — What Gets Tested

Topic	What to Be Able to Do
Data modeling	Set up star schema — fact and dimension tables, define relationships and cardinality
DAX — CALCULATE	Override filter context (e.g. revenue for all regions despite a slicer)
DAX — Time intelligence	YTD, MTD, DATEADD, SAMEPERIODLASTYEAR measures
Power Query	Clean and transform data before loading into the model
Slicers + drill-through	Build interactive reports with cross-filtering and drill-through pages
Measures vs calculated columns	Know when to use each and why measures are preferred

Common DAX Question: YoY Growth Measure

```
Revenue YoY % =
DIVIDE(
  [Total Revenue] - CALCULATE([Total Revenue],
SAMEPERIODLASTYEAR('Date'[Date])),
  CALCULATE([Total Revenue], SAMEPERIODLASTYEAR('Date'[Date]))
)
```

> **Power BI interview tip**
> Most interviewers won't test obscure DAX. They want to know: Can you model data correctly? Do you understand the difference between a measure and a calculated column? Can you build a dashboard that actually answers a business question? Focus on those three.

04 Business Cases & RCA

The Framework — Use This Every Time

Step	What You Do	Common Mistake to Avoid
Clarify	Restate the problem. Ask: which metric, what time period, what's the baseline?	Jumping into analysis without confirming the question
Structure (MECE)	Break the metric into exhaustive, non-overlapping sub-components	Listing random hypotheses with no structure
Hypothesise	State 2–3 likely causes before touching any data	Going straight to data with no prior thinking
Analyse	Name the queries or charts you'd run and what you expect to find	Describing output without explaining what you'd look for
Recommend	Clear action + expected impact + how you'd measure success	Vague 'we should investigate further'

Case Type 1: Metric Drop

> **Classic Question**
> DAU dropped 18% last Tuesday. Walk me through how you'd investigate.

11. **Is the data real?** Check the pipeline first — did logging break? (Always ask this.)
12. **Segment:** Which platform (iOS/Android/Web)? Which region? New vs returning users?
13. **External factors:** App update, competitor campaign, holiday, payment downtime?
14. **Funnel:** Did open rate drop, or did users open but not engage? Where exactly did they fall off?
15. **Product changes:** Was any feature released or config changed around that date?
16. **State your hypothesis:** Lead with your most likely root cause and reasoning before the interviewer asks.

Case Type 2: Retention & Cohort Analysis

> **Classic Question**
> D30 retention dropped from 45% to 38%. Walk me through your approach.

- **Cohort definition:** Group users by signup month, not calendar month. These are different things.
- **Segment the drop:** Is it all cohorts or a specific one? Which acquisition channel?
- **Activation check:** Did churned users complete the 'aha moment'? Lower activation = structural onboarding problem.
- **Feature usage:** Which features correlate with retained users vs churned? Build this in SQL.
- **Recommend:** Match the intervention to the root cause — onboarding, notifications, pricing, or product gap.

Case Type 3: A/B Test Decision

> **Classic Question**
> Email B: 18% conversion vs 15% control (p=0.04, statistically significant). But average order value dropped ₹200. Do you ship?

17. **Calculate total revenue impact:** Higher conversion × lower AOV — which wins at scale? Do the math.
18. **Check guardrail metrics:** Is engagement (opens, clicks) lower? Why would conversion rise but AOV drop?
19. **Segment:** Does Email B work better for new users? Worse for high-value repeat customers?
20. **Business context:** Are we optimising for volume or margin right now? That changes the answer.
21. **Recommendation:** Don't ship broadly. Run a follow-up test on the segment where Email B wins without AOV drop.

RCA Frameworks

Framework	Best For	How to Use in Interviews
5 Whys	Operational problems with a clear chain of causation	Keep asking 'why' until you hit a root cause, not a symptom
Fishbone (Ishikawa)	Multi-factor problems across people, process, data, product, external	Sketch on paper; fill in branches for each cause category
MECE Tree	Any metric drop or business strategy question	Break the metric into exhaustive, non-overlapping components
Hypothesis → Test	Experimentation and ambiguous data problems	State what you believe, then what data would confirm or reject it

A/B Testing — Key Concepts to Know

Concept	What to Be Able to Explain
Hypothesis	Control vs treatment; what you expect to change and why
Primary metric	The single metric the test is designed to move
Guardrail metrics	Metrics that must not degrade — e.g. return rate, complaints
Statistical significance	$p < 0.05$ means a 5% chance the result is due to random variation
Statistical vs practical significance	A result can be significant but too small to be worth shipping
Sample size & power	Bigger MDE = smaller sample needed. Don't start without calculating this.
Peeking problem	Stopping a test early when you see significance inflates false positive rate
Sample ratio mismatch	Traffic split not matching intended ratio — test is compromised

05 Resume & Portfolio

Resume — The Non-Negotiables

Do This	Avoid This
Lead every bullet with an action verb (Analysed, Built, Reduced, Identified)	'Responsible for creating reports'
Quantify impact — '15% reduction in churn', '2x faster reporting'	Vague bullets: 'Worked on dashboards'
1 page if under 3 years of experience	2-page resume with no industry experience
Mention tools with context — SQL (MySQL/BigQuery), Python (Pandas), Tableau	Listing tools without explaining how you used them
Include a Projects section if you have no work experience	Leaving the resume empty of any data work
Tailor keywords from the JD for each application	Sending the exact same resume to every company

Resume Bullet Formula

The Formula

Action Verb + What You Did + Tool Used + Impact / Outcome Example: Analysed 10K customer transactions using SQL window functions and Python to identify a 23% churn concentration in the 30–60 day user cohort, informing a targeted retention campaign.

Before vs After — Fixing Weak Bullets

Before (weak):

Worked on Python project for customer data.

After:

Built end-to-end EDA pipeline in Python (Pandas, Seaborn) on 15K customer records, identifying 3 high-value segments and recommending a personalised discount strategy estimated to improve repeat purchase rate by 12%.

Before (weak):

Used Power BI to make dashboards.

After:

Designed 3 interactive Power BI dashboards with DAX time-intelligence measures (YTD, MoM), enabling the marketing team to cut weekly reporting time from 4 hours to 30 minutes.

Portfolio Projects — How to Frame Them

Project Type	What to Emphasise in Your Write-up
Excel Sales Dashboard	RCA approach, data cleaning judgment, decisions enabled by the dashboard
Python EDA	End-to-end process: question → clean → analyse → insight → recommendation
SQL Case Study	Business problem translated to queries; complexity of logic used
Power BI Dashboard	Data modeling, DAX measures, interactivity, and the business question answered
AI-Assisted Analysis	How you used GenAI + the moments you overrode it with your own judgment

GitHub README Structure

Every project repo needs a README that answers these in order:

22. **Business Problem:** What question were you answering?
23. **Dataset:** Size, source, key columns.
24. **Approach:** Steps taken — EDA, cleaning, analysis, modelling.
25. **Tools Used:** Link to the actual files.
26. **Key Insights:** 3–5 bullet findings with numbers.
27. **Business Recommendation:** What should the company do, and what's the expected impact?

06 Behavioral & HR

STAR Method

Letter	What to Cover	Length
S — Situation	Context: project, team, the problem you faced	2–3 sentences
T — Task	Your specific responsibility in that situation	1 sentence
A — Action	What YOU did — say 'I', not 'we'. Most detail goes here.	3–4 sentences
R — Result	Outcome, quantified if possible	1–2 sentences

Top Questions + Sample Answers

1. Tell me about yourself.

Formula: Current status + relevant skills/experience + why this role.

> **Sample Answer**
> I'm a data analyst with hands-on experience in SQL, Python, Excel, and Power BI. I've worked on end-to-end analytics projects — from cleaning messy datasets and building dashboards to running cohort analyses and designing A/B test frameworks. I'm excited about this role because [specific reason tied to the company or team].

2. Describe a time you worked with messy or incomplete data.

> **Structure your answer around: what the issue was → how you diagnosed it → what you decided to do → what the result was**
> Situation: I was analysing a dataset with 50K transactions where 8% had data quality issues — duplicates, nulls, and inconsistent formats. Action: I categorised the issues first using df.info() and profiling, then decided treatment per issue type — for duplicates I used timestamps to distinguish data errors from legitimate repeat transactions; for nulls I imputed with medians by category rather than dropping rows. Result: Retained 97% of records. Validated the cleaned dataset by cross-checking aggregated totals against the source.

3. Tell me about a project you're proud of.

Use this structure — it works for any project:

- The specific business problem you solved — not just the tools you used
- Your process: how you approached it, what you found, what surprised you
- The recommendation or decision you enabled
- What you'd do differently — this shows self-awareness

4. Tell me about a time you disagreed with someone.

> **Show you can push back with data, not ego**
>
> Frame it as: I had a data-backed concern → I raised it constructively → the team considered it → outcome improved. Even if you didn't 'win', the maturity to engage constructively is what they're evaluating.

5. Tell me about a time you failed.

Be genuine. Pick something real but not catastrophic. The structure:

- What happened and what went wrong
- What you should have done differently
- What you learned and how you applied it after

6. Where do you see yourself in 3 years?

Show ambition but stay grounded. A strong answer: 'I want to go deep on product analytics — owning end-to-end analysis for a business unit, from dashboards to experiment design to influencing strategy. I'd like to be the analyst who connects data to real decisions, not just reporting.'

Questions to Ask the Interviewer

Ask at least 2. These make you stand out:

- What does a typical analysis request look like for this team? How is work handed to analysts?
- What data stack are you working with, and are there plans to expand tooling?
- What's a recent decision this team made that was directly driven by data?
- What separates analysts who grow quickly here from those who plateau?
- What are the biggest data challenges the team is working through right now?

Salary Negotiation — India Benchmarks

Role	Freshers (0–1 yr)	1–2 Years Experience
Data Analyst	₹3.5 – 5.5 LPA	₹5.5 – 8 LPA
Business Analyst	₹4 – 6.5 LPA	₹6 – 10 LPA
Product Analyst	₹5 – 8 LPA	₹8 – 13 LPA
SQL / BI Analyst	₹3 – 5 LPA	₹5 – 7.5 LPA

The Negotiation Script

"I've researched market benchmarks for this role in [city], and based on my skills and the work I've done, I was hoping we could discuss a figure closer to [X]. Is there flexibility there?" Always anchor above your target. Never accept the first offer without asking at least once.