

TRACK 11

Capstone, Portfolio & Interviews

End-to-end capstone, portfolio and GitHub, analyst and data scientist interview preparation with model answers, and a 90-day plan.

4 lessons

Version 2026.10

Updated October 2026

Created by **Varun Kumar**
www.skillnavora.in

Contents

Every entry is a clickable link. The PDF also has bookmarks for each lesson and section.

- [01 Portfolio, GitHub and Resume for Data Roles](#)
- [02 Data Analyst Interview Preparation: SQL, Excel, Power BI, Statistics and Case Questions](#)
- [03 Data Scientist, ML and GenAI Interview Preparation: Questions and Model Answers](#)
- [04 Your 90-Day Plan: Learning Roadmap, Job Search Strategy and Growing in Your First Role](#)

Portfolio, GitHub and Resume for Data Roles

Turn your learning into evidence employers trust. Choose three to five portfolio projects, structure them on GitHub with strong READMEs, publish dashboards and apps, write an ATS-friendly one-page resume with quantified bullets, and optimise LinkedIn and Naukri profiles for analyst, data scientist, data engineer and GenAI roles.

● Career track

✦ 2026 hiring practice in India

Beginner

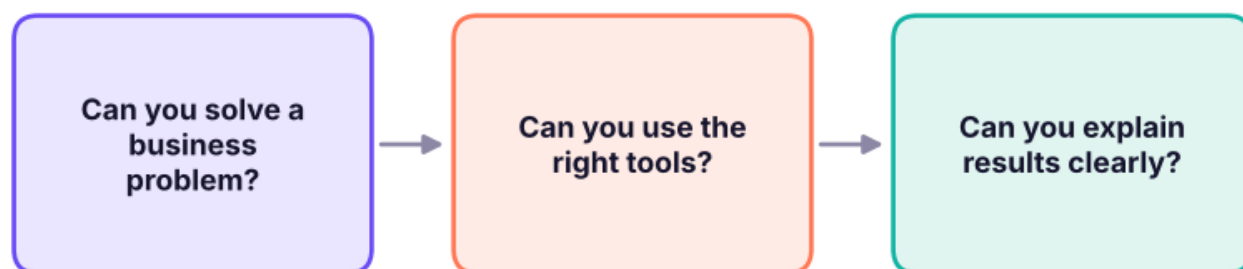
🕒 90 min

By the end of this lesson you can

- Select portfolio projects that match the target role
- Publish projects on GitHub with professional READMEs and live demos
- Write a one-page, ATS-friendly resume with quantified impact bullets
- Optimise LinkedIn and job-portal profiles and keywords

What hiring managers look for

What your portfolio must prove



A portfolio answers three questions for a recruiter in under two minutes.

Target role	Must-have evidence
Data / business analyst	SQL case study, Excel model, Power BI dashboard, EDA with recommendations
Data scientist	EDA, A/B test analysis, ML model with business threshold, explainability
Data engineer	Pipeline (ingest → transform → warehouse), dbt/SQL models, orchestration, tests
GenAI / AI engineer	RAG assistant with evaluation, tool-using agent, deployed demo, cost analysis

Pick 3–5 projects (from this handbook)

1. **SQL:** end-to-end sales case study (SQL lesson 15).
2. **Power BI:** Adventure Cycles dashboard with DAX and drill-through (Power BI capstone).
3. **Python EDA:** telecom churn EDA with recommendations (Python lesson 22) or retail analytics capstone (Python lesson 29).
4. **ML:** churn prediction with profit-based threshold and SHAP (ML lesson 28).
5. **GenAI:** NovaMart RAG support assistant with evaluation and guardrails (GenAI lessons 5, 9, 12).

Make at least one project your own: a dataset from data.gov.in, Kaggle or your workplace (anonymised) about a domain you care about.

GitHub structure and README

Text

```
churn-prediction/
├── README.md           ← the most important file
├── data/README.md      ← source, licence, how to download (do not commit large/private data)
├── notebooks/01_eda.ipynb
├── src/features.py, train.py, evaluate.py
├── reports/figures/    ← 2-4 key charts used in the README
├── requirements.txt
└── app/ (optional Streamlit demo)
```

README template: **Problem** (one paragraph, business framing) → **Data** → **Approach** (bullets) → **Results** (numbers + one chart) → **Recommendation** → **How to run** → **Next steps**. Put a live link (Streamlit Cloud, Hugging Face Spaces, Power BI publish-to-web with sample data, Tableau Public) at the top.

Resume: one page, quantified

Weak bullet	Strong bullet
"Worked on churn analysis using Python."	"Built a churn model (ROC-AUC 0.80) and targeting rule that reaches 2.3× more churners than random outreach, worth ≈₹60K/month in saved margin."
"Made Power BI dashboards."	"Designed a Power BI sales dashboard (star schema, 25 DAX measures) used weekly by 4 regional managers; cut reporting time from 6 hours to 30 minutes."
"Know SQL."	"Wrote 40+ SQL queries (CTEs, window functions) to analyse 6,300 orders; identified tables/bookcases as loss-making due to >20% discounts."

Formula: **action verb + what you did + tool + measurable result**. Sections: header (name, city, phone, email, LinkedIn, GitHub, portfolio), summary (2 lines, target role), skills (grouped), projects/experience, education, certifications. Use a simple single-column layout (ATS systems mis-read tables and graphics), PDF format, and mirror keywords from the job description honestly.

LinkedIn and job portals

- **Headline:** "Data Analyst | SQL · Power BI · Python | Turning sales data into decisions".
- **About:** 3 short paragraphs: who you are, what you build (link projects), what you are looking for.
- **Featured:** pin your best project and dashboard. Post a short write-up per project (problem, chart, insight).
- **Naukri/Instahyre:** complete every field, update weekly (recency boosts ranking), set salary and notice period honestly.

KEY TAKEAWAYS

- Choose 3–5 projects that match the target role and show business impact.
- Publish on GitHub with clear READMEs, charts and live demos.
- One-page ATS-friendly resume; every bullet = action + tool + quantified result.
- Keep LinkedIn and portals complete, keyword-aligned and active.

Quick quiz

4 questions · tap an answer to check it

1. What is the most important file in a portfolio repository?

A README.md

B requirements.txt

C .gitignore

D The largest notebook

Answer: A. Recruiters read the README first.

2. Which resume bullet is strongest?

A "Used Python for analysis."

B "Built a churn model (AUC 0.80) that reaches 2.3× more churners than random outreach."

C "Familiar with machine learning."

D "Responsible for dashboards."

Answer: B. Action, tool and quantified result.

3. Why avoid two-column graphic resumes for online applications?

A Applicant tracking systems may parse them incorrectly

B They look unprofessional

C They are too short

D Recruiters dislike colour

Answer: A. Keep layout simple.

4. How many portfolio projects are ideal?

A 3–5 strong, role-relevant projects

B 1

C 20 small ones

D None

Answer: A. Depth beats count.

EXERCISE Practice

(1) Choose your target role and list the 4 projects you will publish. (2) Rewrite three of your current resume bullets with the formula. (3) Write the README for one project using the template.

Show solution (worked answer)

A good result: a one-page resume, four public repositories with READMEs and charts, one live demo link and an updated LinkedIn headline.

Data Analyst Interview Preparation: SQL, Excel, Power BI, Statistics and Case Questions

Prepare for analyst interviews with the question types that appear most often, model answers and practice links: SQL coding rounds (joins, aggregation, window functions), Excel and Power BI questions, statistics and A/B testing, metric and product-sense cases, and behavioural questions with the STAR method.

● Career track

✦ 2026

Intermediate

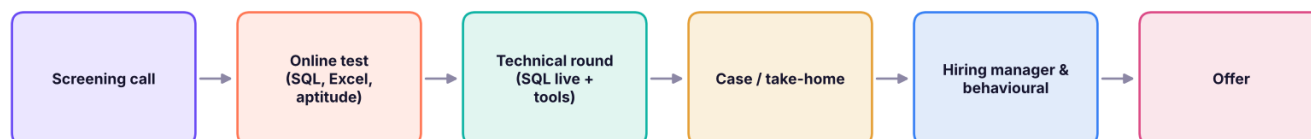
🕒 120 min

By the end of this lesson you can

- Solve common SQL interview patterns quickly and explain your approach
- Answer core Excel, Power BI and DAX questions
- Explain key statistics and A/B testing concepts clearly
- Structure metric, root-cause and product-sense cases
- Tell project stories with STAR

The interview loop

Analyst interview process



A typical analyst hiring process.

SQL patterns you must know

Python

```
import duckdb, pandas as pd
sales = pd.read_csv("superstore_sales.csv", parse_dates=["order_date"])
print(duckdb.sql("""
-- Top 2 products by sales in each region (window function)
WITH r AS (
    SELECT region, product_name, SUM(sales) AS sales,
           RANK() OVER (PARTITION BY region ORDER BY SUM(sales) DESC) AS rnk
    FROM sales GROUP BY region, product_name)
SELECT region, product_name, ROUND(sales) AS sales FROM r WHERE rnk ≤ 2 ORDER BY region,
rnk;
""").df())
```

OUTPUT

	region	product_name	sales
0	East	Apex Copier 347	4191180.0
1	East	Zenith Copier 719	3323372.0
2	North	Orbit Copier 746	3790940.0
3	North	Zenith Machine 401	3566453.0
4	South	Orbit Copier 746	5405517.0
5	South	Vertex Copier 262	4258600.0
6	West	Zenith Copier 719	4087068.0
7	West	Zenith Machine 401	4023468.0

Python

```
import duckdb, pandas as pd
sales = pd.read_csv("superstore_sales.csv", parse_dates=["order_date"])
print(duckdb.sql("""
-- Month-over-month growth with LAG
WITH m AS (SELECT DATE_TRUNC('month', order_date) AS month, SUM(sales) AS sales FROM sales
GROUP BY 1)
SELECT month, ROUND(sales) AS sales, ROUND(100.0 * (sales - LAG(sales) OVER (ORDER BY
month)) / LAG(sales) OVER (ORDER BY month), 1) AS mom_pct
FROM m ORDER BY month DESC LIMIT 4;
""").df())
```

OUTPUT

	month	sales	mom_pct
0	2025-12-01	7652267.0	-25.6
1	2025-11-01	10284516.0	67.5
2	2025-10-01	6140290.0	-27.5
3	2025-09-01	8468060.0	45.9

Pattern	Typical question
GROUP BY + HAVING	Customers with more than 5 orders
Joins (inner, left, anti)	Customers who never ordered
Window: ROW_NUMBER/RANK	Top N per group, de-duplicate latest record
Window: LAG/LEAD, running SUM	Growth, retention, cumulative totals
CASE WHEN	Buckets, conditional counts, pivots
Date functions	Monthly cohorts, days between events
Self join / recursive CTE	Manager hierarchy, consecutive days

Approach out loud: clarify the table grain → write the query step by step (CTEs) → check edge cases (NULLs, ties, duplicates) → validate with a small count.

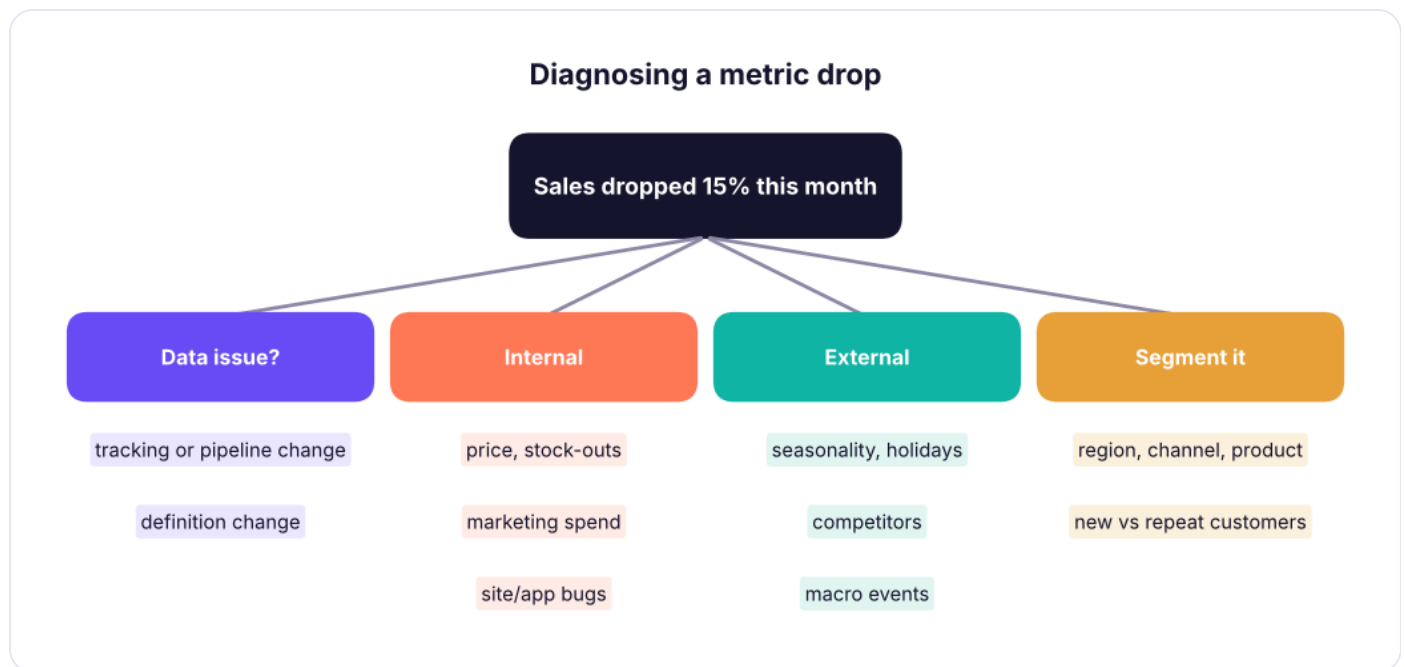
Excel and Power BI questions

Question	Short answer
VLOOKUP vs XLOOKUP vs INDEX-MATCH?	XLOOKUP looks left/right, has defaults and exact match by default; INDEX-MATCH works in older versions; VLOOKUP breaks when columns move
Pivot table vs Power Pivot?	Power Pivot uses a data model with relationships and DAX measures across tables
Calculated column vs measure (DAX)?	Columns are computed row by row and stored; measures are evaluated in filter context at query time
What does CALCULATE do?	Evaluates an expression in a modified filter context
Star schema?	Fact table of events joined to dimension tables; simpler, faster models
Import vs DirectQuery?	Import copies data (fast, scheduled refresh); DirectQuery queries the source live (fresh, slower)
Row-level security?	Roles with DAX filters so users see only their rows

Statistics and experiments

Be ready to explain: mean vs median, standard deviation, distributions, correlation vs causation, confidence intervals, p-values, Type I/II errors, sample size and A/B test pitfalls (peeking, SRM, novelty). See ML lessons 1–6 for model answers.

Case questions: a framework



Root-cause framework for "metric X dropped".

1. Clarify the metric and time frame.
2. Check data quality first.
3. Segment (where is the drop?).
4. Form hypotheses (internal vs external).
5. Validate with data.
6. Recommend action and how you would measure it.

Product-sense metrics: define a north-star metric, input metrics and guardrails (e.g. for a food-delivery app: weekly ordering customers; conversion, delivery time, order value; cancellations, complaints).

Behavioural: STAR

Situation (context) → **Task** (your goal) → **Action** (what *you* did, with tools) → **Result** (numbers, learning). Prepare 6 stories: a project you are proud of, a mistake you fixed, a conflict, a tight deadline, influencing without authority, learning a new tool fast.

Practice resources

W3Schools SQL quiz and exercises, LeetCode SQL 50, HackerRank SQL, StrataScratch, DataLemur, Kaggle Learn; mock interviews with peers; time yourself (15 minutes per medium SQL question).

KEY TAKEAWAYS

- Master SQL patterns: aggregation, joins, window functions, CTEs, dates; think aloud and check edge cases.
- Know Excel lookups/pivots and Power BI modelling, DAX, CALCULATE, star schema and RLS.
- Explain statistics and A/B testing simply.
- Use a structured framework for metric and product cases.
- Prepare STAR stories with quantified results.

Quick quiz

4 questions · tap an answer to check it

1. Which SQL feature finds the top 3 products in each category?

- A** A window function such as `RANK() OVER (PARTITION BY category ORDER BY sales DESC)`
- B GROUP BY only
- C DISTINCT
- D UNION

Answer: A. Then filter rank ≤ 3 .

2. Sales dropped 15%. What is the first check?

- A** Whether the data or tracking changed
- B Fire the marketing team
- C Build an ML model
- D Ignore it

Answer: A. Rule out data issues first.

3. What is a DAX measure?

- A** An expression evaluated at query time in the current filter context
- B A stored column
- C A table relationship
- D A Power Query step

Answer: A. Columns are stored row by row.

4. What does STAR stand for?

- A** Situation, Task, Action, Result
- B Skills, Tools, Analysis, Reporting
- C Start, Test, Assess, Repeat
- D None

Answer: A. Structure for behavioural answers.

EXERCISE Practice

(1) Write SQL for: customers who ordered in 2024 but not in 2025; the second-highest order per customer; 7-day moving average of sales. (2) Prepare two STAR stories. (3) Do a 30-minute mock case: "Delivery complaints doubled last week."

Show solution (worked answer)

Use NOT EXISTS / LEFT JOIN ... IS NULL for the anti-join, DENSE_RANK() = 2 for the second highest, and AVG(...) OVER (ORDER BY day ROWS BETWEEN 6 PRECEDING AND CURRENT ROW) for the moving average.

Data Scientist, ML and GenAI Interview Preparation: Questions and Model Answers

A compact question bank with model answers for data scientist, ML engineer and GenAI roles: Python and pandas, statistics, ML fundamentals, model evaluation, feature engineering, system and case design (churn, fraud, recommendations), MLOps, LLMs, RAG and agents, plus how to handle take-home assignments and live coding.

● Career track

✦ 2026

Advanced

🕒 120 min

By the end of this lesson you can

- Answer core Python, statistics and ML questions concisely
- Walk through an end-to-end ML system design
- Explain LLM, RAG and agent concepts and trade-offs
- Plan and present take-home assignments

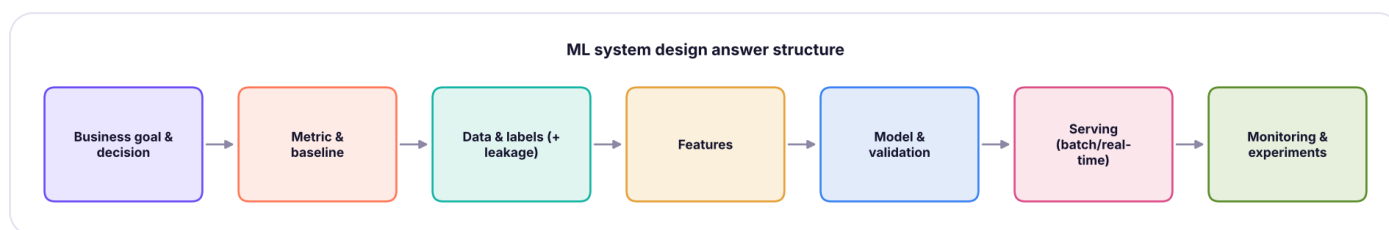
Python and pandas (quick answers)

Question	Answer
List vs tuple vs set vs dict?	Mutable sequence / immutable record / unique unordered / key→value; sets and dicts give O(1) lookups
<code>apply</code> vs vectorisation?	Vectorised column operations run in C and are much faster; avoid <code>apply(axis=1)</code> on large data
<code>merge</code> vs <code>join</code> vs <code>concat</code> ?	merge on columns, join on index, concat stacks
How to handle missing values?	Understand why, then drop, impute (median, group-wise, model) or flag; impute inside CV
Generator vs list?	Generators are lazy and memory-efficient, consumed once

Statistics and ML fundamentals

Question	Model answer
Bias - variance trade - off?	Simple models underfit (bias), complex models overfit (variance); choose complexity by validation error
L1 vs L2?	L1 zeroes coefficients (selection); L2 shrinks smoothly; Elastic Net mixes
Why cross - validation?	A single split is noisy; k - fold averages and supports tuning without touching the test set
Precision vs recall?	Precision: flagged that are right; recall: positives caught; choose by error costs
ROC - AUC vs PR - AUC?	PR - AUC is more informative for rare positives
Bagging vs boosting?	Parallel averaging (variance) vs sequential error correction (bias)
How does gradient descent work?	Iteratively step weights against the loss gradient with a learning rate
What is data leakage?	Using information unavailable at prediction time; prevent with pipelines, time - based splits and feature cut - offs
p - value?	Probability of data at least this extreme if H0 were true; not the probability H0 is true

ML system design: a template



Answer ML design questions in this order.

Example: **"Design a fraud detection system for UPI payments."** Real-time decision under 100 ms; label = confirmed fraud (delayed, imbalanced); features: amount vs user history, device/location change, merchant risk, velocity, graph links; model: gradient boosting + rules + anomaly scores; evaluate PR-AUC and recall at alert capacity; serve with a feature store and low-latency API; monitor drift and attack patterns; human review queue; A/B test thresholds.

LLM, RAG and agent questions

Question	Model answer
How do LLMs work?	Transformers trained to predict the next token, then instruction-tuned and aligned with feedback
RAG vs fine-tuning?	RAG for changing knowledge with citations; fine-tuning for behaviour/format
How to evaluate a RAG system?	Retrieval hit rate/MRR + answer correctness, faithfulness, abstention on a labelled set
How to reduce hallucinations?	Grounding, tools, "say I don't know", citations, validation, evaluation
What is an agent?	An LLM in a loop choosing tool actions toward a goal, with memory and stopping conditions
What is MCP?	Open protocol to connect assistants to tools/data via servers (tools, resources, prompts)
Prompt injection defence?	Separate data from instructions, least-privilege tools, approvals, output checks
Cost control?	Model routing, caching, batching, shorter context and outputs

Take-home assignments

1. Restate the problem and assumptions in the first lines of your report.
2. Do a focused EDA; do not chart everything.
3. Baseline first, then 1–2 better models; validate correctly.
4. Translate results into business impact and a recommendation.
5. Clean, runnable code (requirements, README); a 5-slide summary.
6. State limitations and next steps honestly.

Live coding tips

Clarify inputs/outputs and edge cases, start with a simple correct solution, test with a small example, then optimise and state Big-O (Python DSA lessons 26–28).

KEY TAKEAWAYS

- Answer concisely, then go deeper only if asked.
- Use the design template: goal → metric → data → features → model → serving → monitoring.
- Know LLM/RAG/agent trade-offs, evaluation and safety.
- Take-homes: baseline, correct validation, business recommendation, clean code, honest limitations.

Quick quiz

4 questions · tap an answer to check it

1. What should you state first in an ML system-design answer?

A The business goal and the decision the model supports

B The algorithm

C The GPU type

D The library version

Answer: A. Framing first.

2. Best metric for a fraud model with 0.5% fraud?

A PR-AUC and recall at the alert capacity

B Accuracy

C R^2

D MAE

Answer: A. Rare positives.

3. RAG or fine-tuning for policy documents that change monthly?

A RAG

B Fine-tuning

C Neither

D Training from scratch

Answer: A. Knowledge belongs in retrieval.

4. What must a take-home report include besides the model?

A Business impact, recommendation and limitations

B Only code

C Only accuracy

D Every chart possible

Answer: A. Decision-makers read conclusions.

EXERCISE Practice

(1) Answer "Design a product recommendation system for an e-commerce home page" using the template. (2) Explain RAG to a non-technical manager in 60 seconds. (3) Do one timed take-home on the telecom churn data (3 hours).

Show solution (worked answer)

Recommendations: goal = revenue per session; candidates via collaborative filtering/embeddings + popularity; ranking model with user, item and context features; offline metrics (recall@k, NDCG) + online A/B test on revenue; cold start with popular/contextual items; monitor diversity and drift.

Your 90-Day Plan: Learning Roadmap, Job Search Strategy and Growing in Your First Role

Turn this handbook into a job. A week-by-week 90-day study and project plan for analyst, data science and GenAI tracks, a weekly routine, a job-search funnel with referrals and outreach messages, salary research and negotiation basics, and habits that help you grow in the first 6 months on the job.

● Career track

✦ 2026

Beginner

⌚ 60 min

By the end of this lesson you can

- Follow a realistic 90-day roadmap with weekly goals
- Run a job-search funnel with referrals and targeted applications
- Prepare for salary discussions with market data
- Build habits for success in the first months of a data role

90-day roadmap

Weeks	Analyst track	Data science / ML track	GenAI track
1-2	Excel lessons 1-6	Python 1-10	Python 1-12
3-4	Excel 7-12 + SQL 1-5	Python 15-22 (NumPy, pandas, EDA)	Python 15-19, 23-24 (APIs)
5-6	SQL 6-15 + SQL case study	Stats ML 1-6 (statistics, A/B)	GenAI 1-4 (LLMs, prompts, APIs, embeddings)
7-8	Power BI 1-6 + dashboard project	ML 7-15 (workflow, regression, metrics, CV)	GenAI 5-8 (RAG, tools, agents, MCP)
9-10	Python 1-16, 20-22	ML 16-25 + one capstone	GenAI 9-12 + capstone app
11-12	AI analytics 1-4, portfolio polish	ML 26-28, portfolio polish	Portfolio, deployment, write-ups
13	Interview prep (Career 2) and applications	Interview prep (Career 3)	Interview prep (Career 3)

Weekly routine

Learn (8-10 h)

Lessons and quizzes; type the code yourself

Build (4-6 h)

Project work; one visible output per week

Share (1 h)

Post a LinkedIn update or README section

Practise (2 h)

SQL/DSA problems and mock interviews

Apply (2 h, from week 9)

Targeted applications and referral outreach

Review (30 min)

What worked, what to change next week

A sustainable weekly routine.



Job-search funnel

Stage	Target per week	Tips
Research companies and roles	15	Match your projects to their domain
Referral / recruiter outreach	10	Personalised messages; alumni and community contacts
Applications	10–15	Tailor resume keywords; apply within 3 days of posting
Screens and tests	2–4	Track results; review mistakes
Interviews	1–3	Prepare company - specific cases

Referral message template:

Text

```
Hi <Name>, I'm a data analyst (SQL, Power BI, Python) who built <project link> on <domain>.
I saw the <role> opening at <company> (job id). Could you refer me or share what the team
values most?
Happy to send my one-page resume. Thank you!
```

Track everything in a simple sheet: company, role, date, source, status, next step, notes.

Salary and offers

Research ranges on AmbitionBox, Glassdoor, LinkedIn Salary and peers for your city and experience. Ask for the full CTC breakdown (fixed, variable, joining bonus, benefits), negotiate on evidence (skills, competing offers, market data), and compare learning opportunities, team quality and growth, not only salary.

First 90 days in the job

1. Learn the business: products, customers, key metrics and their definitions.
2. Find the data: sources, owners, known quality issues; document as you go.
3. Deliver a small, useful win in the first month.
4. Communicate: weekly updates, clear charts with insight titles, assumptions stated.
5. Automate repetitive reports; share reusable SQL and notebooks.
6. Keep learning: one new skill per quarter (dbt, cloud, experimentation, GenAI).

KEY TAKEAWAYS

- Follow the 90-day roadmap with a weekly routine of learn, build, share, practise and apply.
- Run the job search as a funnel with referrals, tailored applications and tracking.
- Negotiate with market data and evaluate the whole offer.
- In the new role, learn the business, deliver early wins, communicate clearly and keep learning.

Quick quiz

3 questions · tap an answer to check it

1. What should each week of the plan produce?

A At least one visible output (project progress, write-up or dashboard)

B Only reading

C Only certificates

D Nothing until week 12

Answer: A. Evidence compounds.

2. Which outreach is most effective?

A Personalised referral requests linking a relevant project

B Generic mass messages

C Applying to every job without reading it

D Waiting for recruiters

Answer: A. Relevance wins.

3. What matters when comparing offers?

A Full CTC breakdown plus learning, team and growth

B Only the fixed salary

C Only the company logo

D Only the job title

Answer: A. Look at the whole package.

EXERCISE Practice

Write your personal 13-week plan with weekly goals, choose your capstone, create your job-tracker sheet, and send three referral messages this week.

Show solution (worked answer)

A good plan names the lessons, the project deliverable and the job-search target for every week, with time blocked in your calendar.