EDUSHARK TRAINING

Companion to the 12-Week Workbook

Exercises & Solutions

80 practice problems with worked solutions, one set per week.

How to use this book.

Read the corresponding chapter in the main textbook first, then attempt each exercise on paper or in your editor before peeking at the solution. Solutions are immediately after each problem – resist the temptation. The sets get harder as the week goes on.

Contents

1 Data Foundations & Excel 1

2 SQL Fundamentals 3

3 Advanced SQL 5

4 Statistics Foundations 7

5 Advanced Statistics & EDA 9

6 R Programming 11

7 Advanced R & Visualisation 13

8 R Shiny 15

9 Power BI 17

10 Tableau 19

11 Capstone 21

12 Career Prep 23

Chapter 1

Data Foundations & Excel

Exercise 1.1 Classify the data

For each value below, name its measurement type (numerical-discrete, numerical-continuous, categorical-nominal, categorical-ordinal, date/time): (a) ZIP code, (b) blood pressure, (c) movie rating 1–5, (d) customer email address, (e) signup timestamp, (f) number of children in a household.

Solution

(a) categorical-nominal – although it looks numeric, you don't average ZIP codes. (b) numerical-continuous. (c) categorical-ordinal. (d) categorical-nominal (identifier). (e) date/time. (f) numerical-discrete.

Exercise 1.2 SMART up a vague brief

Rewrite this stakeholder request as a SMART analytical question: "Can you check if the new pricing is working?"

Solution

"Compared to the 30-day period before launch, did the May 2026 price revision change weekly gross margin for SKUs in the Apparel category by more than 2 percentage points, holding promotional activity constant, and is the difference statistically significant?"

Exercise 1.3 Excel formula

Write an Excel formula that returns "high" if the cell `B2` contains a value at least 1.5 standard deviations above the mean of the range `B2:B500`, otherwise "normal".

Solution

```
=IF(B2>=AVERAGE($B$2:$B$500)+1.5*STDEV.S($B$2:$B$500), "high","normal")
```

Exercise 1.4 KPI design

A coffee chain has 50 stores. Pick three leading and three lagging KPIs you would put on the regional manager's daily dashboard.

Solution

Leading: customers per hour, average ticket time, NPS for stores surveyed this week. *Lagging:* same-store revenue YoY, gross margin, total transactions. Pair each lagging with its driver.

Exercise 1.5 Pivot Table workflow

Describe in five steps how to convert a flat sales table (Date, Region, Category, Amount) into a pivot table showing total amount by Region (rows) and Category (columns).

Solution

(1) Select the table; `Ctrl+T` to convert to a Table. (2) `Insert > PivotTable` on a new sheet. (3) Drag Region into *Rows*. (4) Drag Category into *Columns*. (5) Drag Amount into *Values*; ensure *Sum of Amount*.

Chapter 2

SQL Fundamentals

Exercise 2.1 First SELECT

Write a query returning the email and signup date of customers in India who signed up in 2026, ordered by signup date.

Solution

```sql
SELECT email, signup_date
FROM   customers
WHERE  country = 'India'
  AND  signup_date BETWEEN '2026-01-01' AND '2026-12-31'
ORDER BY signup_date;
```

Exercise 2.2 Group and filter

Return regions where total completed-order revenue exceeded £100,000, sorted by revenue descending.

Solution

```sql
SELECT   region, SUM(amount) AS revenue
FROM     orders
WHERE    status = 'completed'
GROUP BY region
HAVING   SUM(amount) > 100000
ORDER BY revenue DESC;
```

Exercise 2.3 LEFT JOIN

Return every customer with the count of their orders, including customers with zero orders.

Solution

```sql
SELECT  c.customer_id, c.first_name, c.last_name,
        COUNT(o.order_id) AS n_orders
FROM    customers c
LEFT JOIN orders o ON o.customer_id = c.customer_id
GROUP BY c.customer_id, c.first_name, c.last_name;
```

Exercise 2.4 CASE WHEN

Classify products with price < 500 as “budget”, 500–1999 as “standard”, 2000+ as “premium”.

Solution

```sql
SELECT product_id, name, price,
       CASE
         WHEN price < 500  THEN 'budget'
         WHEN price < 2000 THEN 'standard'
         ELSE 'premium'
       END AS tier
FROM   products;
```

Exercise 2.5 NULL gotcha

Why does `WHERE phone = NULL` return no rows even when nulls exist? What is the correct test?

Solution

SQL treats `NULL` as “unknown”; comparing anything to it – even `NULL = NULL` – yields `UNKNOWN`, which the `WHERE` filter discards. Use `phone IS NULL` (or `IS NOT NULL`).

Chapter 3

Advanced SQL

Exercise 3.1 Above-average price

Return products whose price is above the overall average price.

Solution

```sql
SELECT name, price
FROM   products
WHERE  price > (SELECT AVG(price) FROM products);
```

Exercise 3.2 Top-N per group

Return the top-3 products by revenue *per category*.

Solution

```sql
WITH ranked AS (
  SELECT category_id, product_id, revenue,
         DENSE_RANK() OVER (PARTITION BY category_id
                            ORDER BY revenue DESC) AS rnk
  FROM   product_revenue
)
SELECT * FROM ranked WHERE rnk <= 3;
```

Exercise 3.3 Running total

Add a running total and a 3-month rolling average to a monthly revenue table.

Solution

```sql
SELECT month_start, revenue,
       SUM(revenue) OVER (ORDER BY month_start) AS running_total,
       AVG(revenue) OVER (ORDER BY month_start
            ROWS BETWEEN 2 PRECEDING AND CURRENT ROW) AS rolling_3m
FROM   monthly_revenue;
```

Exercise 3.4 Recursive CTE

Return every employee and their level in the org hierarchy.

Solution See chapter 3 for the recursive CTE pattern.

Exercise 3.5 Anti-pattern hunt

List three SQL anti-patterns that hurt query performance.

Solution

`SELECT *` when you need a few columns; functions on indexed columns (`WHERE YEAR(d)=2026`); correlated subqueries on large tables; OR across different indexed columns; joining on string columns when integer keys exist.

CHAPTER 4

Statistics Foundations

Exercise 4.1 Mean or median?

Salaries at a small startup are £30k, 32k, 35k, 38k, 42k, 250k (the founder). Which measure of central tendency would you report and why?

Solution

Median (£36.5k). The founder is a clear outlier; the mean (£71.2k) misrepresents typical compensation.

Exercise 4.2 Empirical rule

A factory's bottle-fill weight is normally distributed with $\mu = 500$g and $\sigma = 4$g. What percentage of bottles weigh between 492 and 508g?

Solution

492 and 508 are $\mu \pm 2\sigma$. By the empirical rule, approximately 95%.

Exercise 4.3 Z-score interpretation

A customer's monthly spend has $z = 2.3$ relative to your customer base. What does that imply, and what action would you take?

Solution

She is roughly 2.3 SDs above the average spender (top 1%). Flag for the VIP / retention programme; consider an account manager assignment.

Exercise 4.4 Bayes drill

A test is 98% accurate. The disease prevalence is 0.5%. Compute the probability that a positive test indicates disease.

Solution

Let D = disease, T = positive test. $P(D|T) = \frac{P(T|D)P(D)}{P(T)} = \frac{0.98 \cdot 0.005}{0.98 \cdot 0.005 + 0.02 \cdot 0.995} \approx 0.20$. Roughly 20%.

Exercise 4.5 Sampling

You must survey a 12-million-customer base. Pick a sampling strategy and justify it.

Solution

Stratified random sampling, stratified by region (and possibly customer tier), with sample sizes proportional to stratum size. This preserves representativeness in subgroups while keeping the overall sample small (∼1,000–2,000 for a 3% margin of error at 95% confidence).

Chapter 5

Advanced Statistics & EDA

Exercise 5.1 Frame a hypothesis

A B2B sales team claims a new email cadence boosts response rate. Write H_0 and H_1.

Solution

H_0: response rate under the new cadence equals the old. H_1: response rate under the new cadence is higher (one-tailed) – choose two-tailed if you genuinely don't know the direction.

Exercise 5.2 Pick the test

You have two independent customer groups (n=200 each) and want to know if their mean order values differ. Which test, and why?

Solution

Welch's two-sample t-test. Two independent groups, numeric outcome, unknown population SDs, no assumption of equal variance.

Exercise 5.3 Correlation interpretation

$r = 0.84$ between ice-cream sales and drownings. Causation?

Solution

No. The confounder is summer / temperature, which drives both.

Exercise 5.4 EDA routine

Outline the first seven steps you would run on a new dataset.

Solution Shape, missingness, univariate plots, bivariate plots, correlations, outliers, draft three hypotheses. See chapter 5.

Exercise 5.5 A/B test mistake

After day 3, version B's conversion rate jumps to 1.4× A's. You declare victory and ship. What did you do wrong?

Solution

Peeked early. With a pre-committed sample size and significance level, stopping when p dips below α inflates Type I error. Run the test to its planned end (or use a proper sequential-testing method).

Chapter 6

R Programming

Exercise 6.1 Vector basics

Create a numeric vector of the first 10 odd numbers and compute its mean and standard deviation in one line each.

Solution

```r
v <- seq(1, 19, by = 2)
mean(v)   # 10
sd(v)     # 6.0553...
```

Exercise 6.2 dplyr pipeline

From a `orders` data frame, return the top 5 customers by total revenue in 2026.

Solution

```r
orders %>%
  filter(year(order_date) == 2026) %>%
  group_by(customer_id) %>%
  summarise(revenue = sum(amount)) %>%
  arrange(desc(revenue)) %>%
  slice_head(n = 5)
```

Exercise 6.3 Reshape

You have monthly columns `jan, feb, ..., dec`. Pivot to a long format with columns `month, revenue`.

Solution

```r
sales_long <- sales_wide %>%
  pivot_longer(cols = jan:dec, names_to = "month", values_to = "revenue")
```

Exercise 6.4 Custom function

Write `growth_rate(x, y)` that returns the percent change from `x` to `y` and handles `x == 0`.

Solution

```r
growth_rate ← function(x, y) {
  if (is.na(x) || x == 0) return(NA_real_)
  (y - x) / x * 100
}
```

Exercise 6.5 Joins

Write code that returns customers who placed at least one order, with their order count and lifetime revenue.

Solution

```r
customers %>%
  inner_join(orders, by = "customer_id") %>%
  group_by(customer_id, first_name, last_name) %>%
  summarise(n_orders = n(),
            lifetime = sum(amount), .groups = "drop")
```

Chapter 7

Advanced R & Visualisation

Exercise 7.1 Basic ggplot

Plot mpg vs weight in `mtcars`, coloured by number of cylinders.

Solution

```
ggplot(mtcars, aes(wt, mpg, colour = factor(cyl))) +
  geom_point(size = 3) + theme_minimal()
```

Exercise 7.2 Facet

Faceted line chart: revenue per month, one panel per region.

Solution

```
ggplot(sales, aes(month, revenue)) +
  geom_line() +
  facet_wrap(~ region, scales = "free_y") +
  theme_minimal()
```

Exercise 7.3 Regex

Extract the gmail addresses from a vector of mixed addresses.

Solution

```
str_subset(emails, "@gmail\\.com$")
```

Exercise 7.4 Dates

Given a date column, add columns for year-month, weekday name, and a `is_weekend` flag.

Solution

```
df %>% mutate(
  ym         = floor_date(date, "month"),
  weekday    = wday(date, label = TRUE),
  is_weekend = wday(date) %in% c(1, 7)
)
```

Exercise 7.5 across()

Compute the mean of every numeric column in `iris`, by Species.

Solution

```
iris %>% group_by(Species) %>%
  summarise(across(where(is.numeric), mean))
```

Chapter 8

R Shiny

Exercise 8.1 Minimal app

Write the smallest Shiny app that displays a slider 1–100 and a text output “You picked: X”.

Solution

```
ui ← fluidPage(
  sliderInput("n", "Pick", 1, 100, 50),
  textOutput("out")
)
server ← function(input, output, session) {
  output$out ← renderText(paste("You picked:", input$n))
}
shinyApp(ui, server)
```

Exercise 8.2 Reactive vs observe

When would you use `reactive()` vs `observe()`?

Solution

`reactive()` when you need a value that other code consumes; `observe()` for side effects (writing files, updating widgets) where no return value is needed.

Exercise 8.3 Filtered table

Show a DT table that filters by region. Cache the filter in a reactive.

Solution See chapter 8 Day 3 example.

Exercise 8.4 shinydashboard skeleton

Sketch a 3-tab dashboard layout in shinydashboard.

Solution See chapter 8 Day 5 listing.

Exercise 8.5 Deployment

What three things should you do before publishing to shinyapps.io?

Solution

Pre-compute data into .rds; pin packages with renv; wrap risky code in `tryCatch` with friendly notifications.

CHAPTER 9

Power BI

Exercise 9.1 Import vs DirectQuery

Name one situation where each is the right choice.

Solution

Import for a 2 GB historical dataset that updates nightly – fastest queries, daily refresh schedule. *DirectQuery* for a 20-billion-row SQL warehouse you can't extract – live query, slower visuals but always current.

Exercise 9.2 Power Query step list

Outline the steps to clean a CSV with messy headers, mixed-case region names, and dates as strings.

Solution

Promote first row to headers → change column types → trim/clean text → format `region` to proper case → change date string to Date type → remove empty rows → rename columns to snake_case.

Exercise 9.3 Star schema

Why is one wide flat table worse than a star schema for Power BI?

Solution

VertiPaq's columnar storage is built for the star pattern. Wide flat tables explode model size, kill cross-filtering performance, and force duplicate dimension data.

Exercise 9.4 Time intelligence

Write a YTD measure for `Total Revenue`.

Solution

```
YTD Revenue := TOTALYTD([Total Revenue], DateTable[Date])
```

Exercise 9.5 CALCULATE

Write a DAX measure "Revenue from Top 5 Customers".

Solution See chapter 9 listing.

CHAPTER 10

Tableau

Exercise 10.1 Dimensions vs measures

Classify these fields: order date, region, units sold, profit margin, customer email.

Solution

Order date – dimension (date). Region – dimension. Units sold – measure. Profit margin – measure. Customer email – dimension.

Exercise 10.2 Calculated field

Write a Tableau calculation for “

Solution `SUM([Sales]) / TOTAL(SUM([Sales]))`

Exercise 10.3 Filter action

Why use a filter action instead of a regular filter on a dashboard?

Solution

A filter action lets a click on one chart filter the others on the same dashboard – an interactive drill-down. A regular filter applies to a fixed worksheet only.

Exercise 10.4 Show Me

Two measures, no dimensions. What does Show Me recommend?

Solution Scatter plot.

Exercise 10.5 Tool choice

Your stakeholder wants to embed an interactive what-if model with custom maths. Tableau, Power BI or R Shiny?

Solution

R Shiny. Custom logic and reactive what-if simulations are exactly where Tableau and Power BI hit their limits.

CHAPTER 11

Capstone

Exercise 11.1 Scope a project

Write a one-paragraph proposal for a retail analytics capstone.

Solution

"Build an interactive R Shiny dashboard for a multi-store retail business to monitor weekly KPIs (revenue, units, AOV, gross margin) by region and category. Data source: 12 months of synthetic point-of-sale extracts. Dashboard supports filters for date range, region and category, and contains a forecasting tab with a simple 4-week revenue projection per region. Audience: regional managers."

Exercise 11.2 Repository structure

Sketch a directory tree for a serious analytics project.

Solution See chapter 11 Day 5 example.

Exercise 11.3 Find the surprise

Why is finding one counter-intuitive insight critical for the capstone?

Solution

It is the headline of your story. "Revenue went up" is forgettable; "Tuesdays out-sell Fridays in tier-2 cities" is what an interviewer repeats to a colleague.

Exercise 11.4 Performance

List three quick wins to speed up a slow Shiny app.

Solution

Pre-aggregate data into `.rds` files; memoise heavy reactives with `bindCache()`; remove unused packages; lazy-load tabs.

Exercise 11.5 Storytelling

Outline the 7-slide structure for your final capstone presentation.

Solution See chapter 12 Day 6.

Chapter 12

Career Prep

Exercise 12.1 ATS-friendly resume

Why are tables in a Word resume often a bad idea?

Solution

Many ATS parsers don't read tables correctly; columns can get scrambled or content dropped, killing your match score before a human sees the resume.

Exercise 12.2 STAR

Write a STAR answer (4–6 sentences) for "Tell me about a time you used data to influence a decision."

Solution

S: During my capstone for EDUSHARK, I analysed a retailer's 12-month POS data. *T*: My mentor asked which underperforming category to revive. *A*: I ran a margin-adjusted ranking and noticed Beverages had high units but the lowest margin contribution, while a small "Specialty Tea" sub-category had 3× the margin per unit and was sold out 22% of days. I built a Shiny app that surfaced this inventory pattern. *R*: My mentor recommended re-stocking Specialty Tea and reducing SKUs in Beverages – a simulated 9% gross-margin lift.

Exercise 12.3 Mock SQL question

"Return the second-highest salary per department." Write the query.

Solution

```sql
WITH ranked AS (
  SELECT department_id, salary,
         DENSE_RANK() OVER (PARTITION BY department_id
                            ORDER BY salary DESC) AS rnk
  FROM   employees
)
SELECT department_id, salary
FROM   ranked
WHERE  rnk = 2;
```

Exercise 12.4 Hard interview question

"Explain p-value to a non-technical executive in 30 seconds."

Solution

"Assume the change we made does nothing. The p-value tells me how often I'd see data this extreme by random chance alone. A p-value below 5% is the analyst's threshold for saying the change probably did do something."

Exercise 12.5 Three questions to ask interviewers

Write three thoughtful questions that work in any analyst interview.

Solution

(1) What does a great first 90 days look like for this role? (2) Walk me through a recent analysis the team owned end-to-end – what went well, what would you do differently? (3) Where do you see the team's analytics maturity in 18 months, and what needs to change to get there?