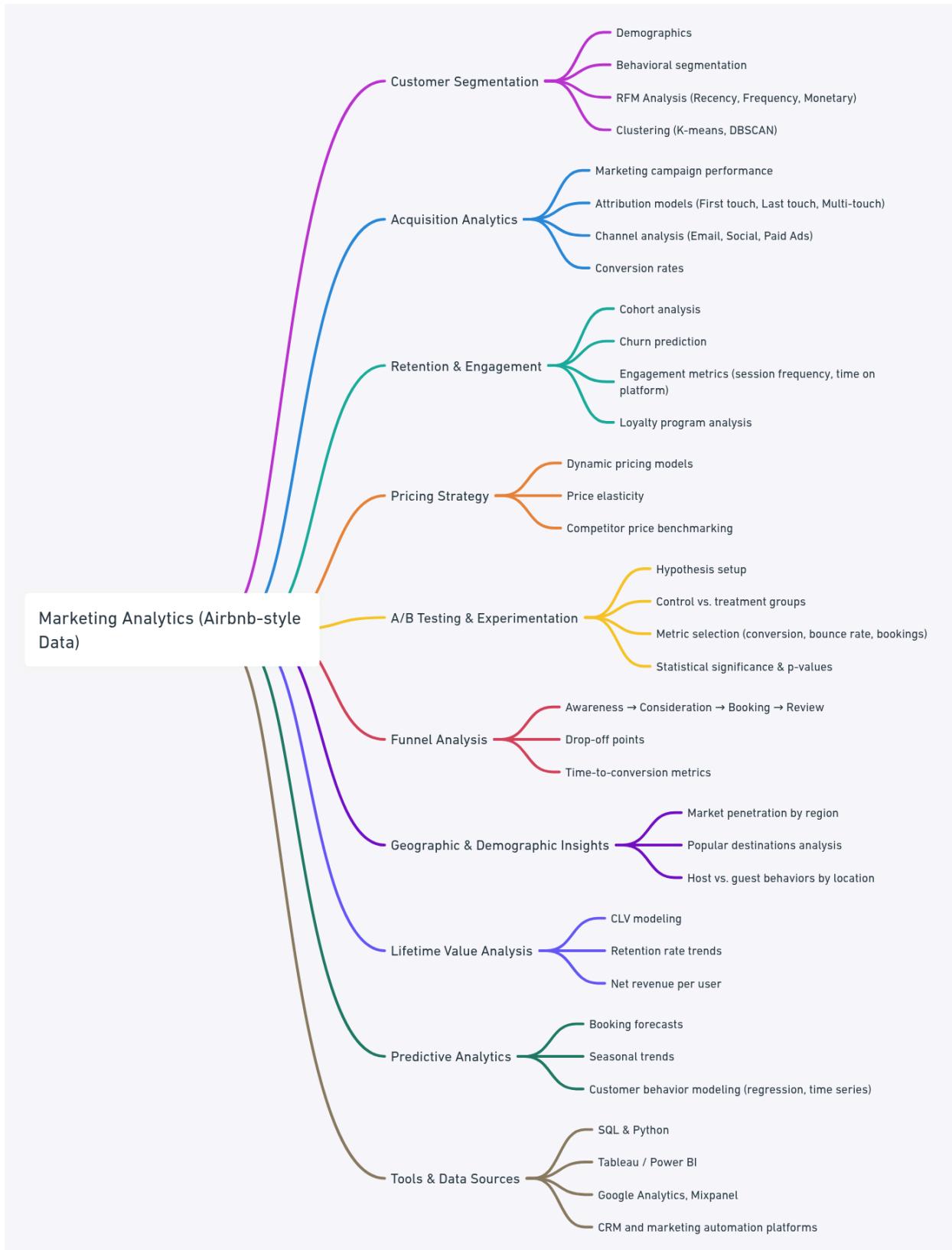


Analytics



Introduction to Data Analytics

Let's learn how to approach data analytics interviews and understand the different types of analytics interview questions that will show up for data analyst and data science roles.

0 of 3 Completed

Data Analytics Fundamentals: Causal Inference

In this course we'll go over the core concepts of causality, significance, and analyzing data. This is meant as a quick refresher and a high level overview of causal inference basics to eventually apply them in data analytics problems.

0 of 9 Completed

Diagnosing and Investigating Metrics

Investigating metrics is a type of product intuition problem that will come up frequently in interviews. Examples of this are typically phrased along the lines of - If X metric is up/down by Y percent, how would you investigate it?

0 of 12 Completed

Easy SQL Questions

Get started on tackling easy level SQL questions involving aggregations, joining multiple tables, and pulling data for beginning analytical reports.

0 of 12 Completed

Medium SQL Questions

Medium level SQL questions utilize more advanced concepts like sub-queries, window functions, and solving case study problems.

0 of 19 Completed

Measuring Success

Measuring the success of products is critical to data science and analytics interviews. Generally, this question is an encapsulation of every time a product manager or executive asks the question: "So, how is it doing?".

0 of 11 Completed

A/B Testing & Experiment Design

Let's start with a general framework for A/B testing. In practice, an A/B testing and experimentation all follow a step by step process of setting metrics and designing experiments.

1 of 10 Completed

Business Health Metrics

Business health metrics tackle case studies and questions focused on setting and defining core metrics for different business priorities.

Analytics Knowledge & Utilities (AKU)

ANALYTICS

Analytics, the systematic computational analysis of data or statistics, has become the bedrock of modern decision-making across all sectors.

It encompasses a wide range of job functions, responsibilities, and utilities, from interpreting historical data to predicting future outcomes.

Analytics Summary:

Core Knowledge & Applied Responsibilities

This section defines analytics and outlines its core purpose: to discover and communicate data patterns, transforming raw data into actionable insights that drive business decisions, optimization, and growth.

Data Patterns Discovery, Interpretation, & Communication

Analytics is the process of discovering, interpreting, and communicating significant patterns in data

Data Transformation, Insights, Decisions, Optimization, & Growth Goals

The primary goal is to transform raw data into actionable insights that can drive business decisions, optimize processes, and foster growth

Data Collection, Cleaning, Processing, & Modeling

Job functions within analytics are diverse, ranging from data collection and cleaning to the development of complex predictive models

Analytics Core Concepts:

Key Analytics Terminology & Definitions

This section breaks down the essential vocabulary of the analytics field, defining foundational terms such as the different types of analytics (descriptive, predictive, prescriptive), business intelligence, machine learning, and big data.

Descriptive Analytics

Summarizes historical data to understand what has happened.

Predictive Analytics

Uses statistical models and forecasting techniques to understand the future and answer "what could happen?".

Prescriptive Analytics

Suggests actions to take based on predictions, helping to answer "what should we do?".

Business Intelligence (BI)

Focuses on using data to understand business performance, often through dashboards and reports tracking Key Performance Indicators (KPIs).

Machine Learning (ML)

A subset of Artificial Intelligence (AI) that uses algorithms to learn from data and make predictions or decisions without being explicitly programmed.

Big Data

The large volume of data, both structured and unstructured, that inundates a business daily.

Data Governance

The overall management of the availability, usability, integrity, and security of data within an organization.

Frameworks & Methodologies: Structuring the Analytical Process

This section details the structured approaches and workflows used in analytics projects, covering foundational requirements, standard processes like CRISP-DM, and key statistical theories that guide effective data analysis.

Foundational Requirements

Core knowledge includes a strong understanding of statistical concepts, data handling techniques, and the ability to translate business problems into analytical questions.

Standard Analytics Workflow

A typical, iterative workflow includes Problem Definition, Data Collection, Data Preparation, Exploratory Data Analysis (EDA), Modeling, Evaluation, and Deployment.

Key Statistical Theories & Applications

Methods include descriptive statistics (summarizing data) and inferential statistics (drawing conclusions). Key applications include Hypothesis Testing, Regression Analysis, and A/B Testing.

Implementation Models & Frameworks

Frameworks like CRISP-DM (Cross-Industry Standard Process for Data Mining) provide a structured, six-phase approach to guide analytics projects from business understanding to deployment.

Skills, Tools & Technologies: The Analyst's Toolkit

This section provides an overview of the essential software and platforms in an analyst's toolkit, covering programming languages, data visualization tools, and the cloud and big data technologies used for data storage and processing.

Programming & Data Management

Languages: Proficiency in Python and R is essential for data analysis and ML, while SQL remains the standard for querying and managing relational databases.

Data Management & Engineering

Understanding of database architecture, ETL pipelines, and data warehousing concepts - including data storage, processing & pipelines; utilizing cloud platforms such as AWS, Google Cloud, and Azure offer scalable services for data storage, processing, and analysis.

Big Data & Transformation: Technologies like Apache Spark process massive datasets, while tools like dbt (data build tool) are used for transforming data directly within a data warehouse.

Statistics & Probability

Solid foundation in correlation, regression analysis, hypothesis testing, and root cause analysis.

Data Analysis & Science

Ability to perform segmentation, cohort analysis, and A/B testing, with advanced skills in predictive modeling and machine learning.

Visualization & Business Intelligence (BI)

Tools: Tableau, Power BI, and Looker are market-leading tools for creating interactive dashboards, charts, and reports that allow for visual data exploration.

Leadership & Strategic Business (Soft) Skills

Critical Thinking & Collaboration: The ability to critically evaluate problems, work effectively in teams, and translate analytical outcomes into business solutions.

Stakeholder Communication & Storytelling: Skill in communicating complex findings to non-technical audiences and using data storytelling techniques to persuade and drive action.

Responsibilities, Functions, & Application: Job Related Segments

This section illustrates the practical application of analytics across various business domains, showcasing real-world examples such as customer segmentation, operational optimization, fraud detection, and personalized recommendations.

Customer-Centric Analytics

Customer Segmentation: Grouping customers based on shared behaviors or characteristics to personalize marketing efforts.

Churn Prediction: Building models to identify customers at high risk of leaving, enabling proactive retention strategies.

Business Operations & Optimization

Operations Optimization: Using data to improve efficiency and reduce costs in areas like supply chain management and manufacturing.

Fraud Detection: Applying analytical models to identify and prevent fraudulent financial transactions in real-time.

Product & Marketing Analytics

Personalized Recommendations: Powering recommendation engines on platforms like Netflix and Amazon to enhance user experience.

Marketing Mix Modeling: Determining the effectiveness of various marketing channels to optimize advertising spend.

Community: Professional Networks & Resources

This section highlights the importance of professional networking and continuous learning by listing key industry associations, online forums, and other community resources that help analytics professionals stay connected and current.

Key Associations & Conferences

Organizations like INFORMS (Institute for Operations Research and the Management Sciences) and DAMA International offer professional development, standards, and networking opportunities.

Forums & Social Communities

Online communities are vital for problem-solving and knowledge sharing, including Stack Overflow for technical questions, Kaggle for competitions and datasets, and Reddit forums like r/datascience and r/analytics.

Content & Thought Leaders

Staying current requires following industry blogs, publications, and newsletters from prominent thought leaders and organizations.

Governance: Ethics, Regulations & Compliance

This section addresses the critical legal and ethical considerations in analytics, focusing on essential topics like data privacy regulations (GDPR, CCPA), algorithmic bias, and the need for transparent and accountable data governance.

Data Privacy & Compliance

Protecting personal and sensitive information is paramount. This involves strict adherence to data protection regulations like Europe's GDPR (General Data Protection Regulation) and the CCPA (California Consumer Privacy Act).

Algorithmic Bias & Fairness

A critical responsibility is to be aware of and mitigate biases inherent in data and algorithms to ensure fair and equitable outcomes.

Transparency & Accountability

Organizations must be transparent with stakeholders about how data is used and remain accountable for the decisions and impacts of their analytical models.

Outlook: Current Trends & Directions

This section explores the emerging technologies and methodologies shaping the analytics landscape, including the growing integration of AI, the demand for real-time processing, and the move toward data democratization and explainable AI.

AI Integration & Automation

Augmented Analytics: The use of AI/ML to augment human intelligence and automate data preparation, insight discovery, and analysis for all user levels.

Real-Time & Edge Processing

Real-Time Analytics: A growing demand for immediate insights that enables businesses to make faster, in-the-moment decisions.

Edge Analytics: The practice of processing data near its source (e.g., on IoT devices) to reduce latency and enable rapid response.

Data Democratization & Architecture

Data Democratization: A movement toward empowering employees at all levels with access to data and user-friendly, self-service analytics tools.

Data Fabric: An emerging architectural approach that simplifies and integrates data management across complex cloud and on-premises environments.

Advanced Methodologies

Explainable AI (XAI): A rising field focused on developing models and techniques that can explain how complex AI systems arrive at their decisions, building trust and transparency.

Keyword & Phrase Index (KPI)

Most common keywords, terms, and phrases for analytics management roles.

Keyword Utilization Index (KUI)

Keyword

Count

data

240

analytics

144

marketing

134

experience

131

business

108

team

75

insights

67

product

63

Professional Development Roadmaps (PDR)

Skill Acquisition Sequences

Foundation to Intermediate (Years 1-3)

Statistical Fundamentals: Descriptive statistics, hypothesis testing, and basic regression

Programming Basics: SQL proficiency, Python/R fundamentals, and data manipulation

Visualization Skills: Chart selection, dashboard design, and storytelling techniques

Business Acumen: Industry knowledge, financial literacy, and stakeholder management

Intermediate to Advanced (Years 4-6)

Advanced Analytics: Machine learning, predictive modeling, and experimental design

Technical Leadership: Code review, methodology development, and junior mentoring

Strategic Thinking: Business strategy alignment, ROI measurement, and innovation management

Cross-functional Collaboration: Project management, stakeholder influence, and change management

Advanced to Expert (Years 7+)

Thought Leadership: Industry expertise, research publication, and conference speaking

Organizational Impact: Strategy development, budget management, and executive communication

Innovation Management: Emerging technology adoption, methodology development, and competitive advantage

Talent Development: Team building, succession planning, and knowledge transfer

Responsibilities and Functions

A Marketing Analytics Director is responsible for leading and managing a team of analysts to collect, analyze, and interpret marketing data, transforming it into actionable insights that drive marketing effectiveness and business outcomes. In the digital age, marketers have access to an abundance of data from various sources, and the Marketing Analytics Director plays a critical role in extracting meaningful patterns and trends from this data. Their key responsibilities and functions typically include:

Developing and Implementing Marketing Analytics Strategies: This involves defining key performance indicators (KPIs), establishing measurement frameworks, and outlining the analytical approaches to be used across various marketing channels.

Leading and Managing the Analytics Team: This includes setting team goals, providing guidance and mentorship, fostering a data-driven culture, and ensuring the team has the necessary resources and skills to succeed.

Analyzing Marketing Campaign Performance: This involves collecting and analyzing data from various sources, such as website traffic, social media engagement, email campaigns, and advertising platforms, to assess campaign effectiveness, identify trends, and uncover optimization opportunities.

Conducting Deep-Dive Data Analysis: This includes performing in-depth analysis to understand customer behavior, identify market trends, and uncover insights that can inform marketing strategies and product development.

Managing Marketing Performance Tracking Systems: This involves overseeing the implementation and maintenance of marketing analytics tools and technologies, ensuring data accuracy and consistent reporting.

Collaborating with Cross-Functional Teams: This includes working closely with marketing, sales, product, and IT teams to align on key business metrics, share insights, and ensure data-driven decision-making across the organization.

Communicating Insights and Recommendations: This involves presenting findings and recommendations to senior management and stakeholders in a clear and concise manner, translating complex data into actionable strategies.

Staying Up-to-Date with Industry Trends: This includes keeping abreast of the latest marketing analytics technologies, methodologies, and best practices to ensure the team is using the most effective tools and approaches.

Key Objectives and Performance Indicators

The key objectives of a Marketing Analytics Director are aligned with driving business growth and improving marketing effectiveness. Marketing analytics helps predict customer needs, optimize marketing spending, and ensure that every marketing dollar invested yields the best possible results. Some common objectives and associated performance indicators include:

Increase Marketing ROI: Measured by metrics such as return on ad spend (ROAS), customer acquisition cost (CAC), and customer lifetime value (CLTV).

Improve Customer Engagement: Measured by metrics such as website traffic, social media engagement, email open rates, and conversion rates.

Optimize Marketing Campaigns: Measured by metrics such as click-through rates (CTR), conversion rates, and cost per acquisition (CPA).

Enhance Customer Segmentation and Targeting: Measured by the effectiveness of targeted marketing campaigns and personalized messaging.

Drive Data-Driven Decision Making: Measured by the adoption of data-driven insights in marketing strategies and business decisions.

Analyst

Data Analysis & Reporting (DAR)

CORE TOPICS

Analytics Value Chain & Core Functions (AVC.CF)

BROAD TOPIC

SUB-TOPICS

DETAILED TOPICS

Data Acquisition

Event & API tracking, Fivetran/Airbyte pipelines

Raw landing tables, CDC streams

Data Engineering

dbt/SQL ELT, column-level lineage, tests

Curated marts, entity-resolution tables

Analytics & BI

Ad-hoc SQL/Python, descriptive stats, dashboards

Exploratory notebooks, Looker/Power BI reports

Modeling & ML

Feature engineering, MLflow experiments, MLOps

Forecasts, propensity scores, segmentations

Experimentation & Causal

A/B test design, CUPED, diff-in-diff

Lift reports, decision memos

Activation

Reverse-ETL, marketing pixels, product triggers

Audiences in CDP/CRM, automated actions

Governance & Quality

Data contracts, Monte Carlo/Soda monitors

SLA reports, incident retros

Storytelling & Strategy

Executive read-outs, OKR alignment

Roadmaps, North-star metrics, ROI models

Data Visualization

Visualization Tools

Tableau

- Connecting to data sources
- Calculated fields and parameters
- Dashboards and storyboards

Power BI

- Data modeling and DAX
- Report building
- Sharing and collaboration features

Visualization Principles

Design Best Practices

- Choosing chart types (bar, line, pie)
- Color theory and accessibility
- Avoiding misleading visuals

Storytelling with Data

- Structuring a narrative
- Highlighting key insights
- Tailoring visuals for stakeholders

3. Analytics & Analysis Domain

Core Skills and Competencies

Technical Skills

- Statistical Analysis [Skill] - Descriptive statistics, inferential statistics, hypothesis testing, regression analysis
- Data Modeling [Skill] - Conceptual modeling, predictive modeling, statistical modeling
- Programming Languages [Skill] - Python, R, SQL, SAS, MATLAB, Scala, Java

- Machine Learning [Skill] - Supervised learning, unsupervised learning, deep learning, ensemble methods
- Database Management [Skill] - SQL querying, database design, data warehousing, NoSQL databases

Business Skills

- Data Storytelling [Skill] - Narrative construction, insight communication, stakeholder engagement
- Business Intelligence [Skill] - KPI development, dashboard creation, performance measurement
- Analytics Strategy [Skill] - Analytics roadmap development, ROI measurement, data governance

Key Methodologies and Frameworks

- CRISP-DM [Framework] - Cross-Industry Standard Process for Data Mining (6 phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment)
- SEMMA [Framework] - Sample, Explore, Modify, Model, Assess (SAS-developed methodology)
- Descriptive Analytics [Methodology] - Historical data analysis, reporting, data aggregation
- Predictive Analytics [Methodology] - Forecasting, machine learning models, statistical modeling
- Prescriptive Analytics [Methodology] - Optimization, simulation, decision modeling

Essential Tools and Software Platforms

- Tableau [Tool] - Advanced data visualization, interactive dashboards, self-service analytics
- Python [Tool] - Versatile programming language with data science libraries (pandas, numpy, scikit-learn)

- R [Tool] - Open-source statistical computing with extensive package ecosystem
- SAS [Tool] - Enterprise statistical software with advanced analytics capabilities
- Google Cloud Platform [Tool] - BigQuery, Dataflow, Cloud ML Engine, Looker
- Snowflake [Tool] - Cloud data platform, data warehousing, analytics workloads

Sub-disciplines and Specializations

- Customer Analytics [Specialization] - Customer segmentation, lifetime value modeling, churn prediction
- Marketing Analytics [Specialization] - Campaign optimization, attribution modeling, A/B testing
- Financial Analytics [Specialization] - Risk modeling, fraud detection, credit scoring
- Data Science [Specialization] - Machine learning, artificial intelligence, experimental design
- Business Intelligence [Specialization] - Data warehousing, OLAP, reporting, dashboard development

Emerging Trends and Technologies

- Agentic AI [Trend] - Autonomous AI systems capable of independent decision-making
- Generative AI for Analytics [Trend] - AI-powered report generation, automated insight discovery
- Real-time Analytics [Trend] - Stream processing, edge computing, instantaneous decision-making
- MLOps [Trend] - Machine learning operations, model lifecycle management

1. ANALYSIS PLANNING & STRATEGY

Business Question Definition

Reverse-engineering from decisions needed

Stakeholder requirement gathering

Success criteria establishment

Scope and constraint documentation

Hypothesis Development

Formulating testable hypotheses

Identifying key variables and relationships

Defining success/failure metrics

Creating analysis roadmaps

Resource Planning

Data source identification

Tool and technology selection

Timeline and milestone setting

Team role assignment

2. DATA MANAGEMENT & QUALITY

Data Quality Assessment

Completeness checks (missing values, gaps)

Accuracy validation (range checks, business rules)

Consistency verification (cross-source validation)

Timeliness evaluation (data freshness)

Data quality scorecarding

Data Governance

Source documentation

Data lineage tracking

Access control and security

Version control implementation

Metadata management

Data Preparation Standards

ETL process documentation

Transformation logic standardization

Error handling procedures

Data validation checkpoints

3. ANALYTICAL FRAMEWORKS

Progressive Depth Analysis

Descriptive: What happened?

Summary statistics

Historical trending

Current state assessment

Diagnostic: Why did it happen?

Root cause analysis

Variance decomposition

Driver identification

Predictive: What will happen?

Forecasting models

Risk assessment

Scenario planning

Prescriptive: What should we do?

Optimization analysis

Recommendation development

Impact estimation

Segmentation & Cohort Analysis

Customer segmentation strategies

Behavioral cohort creation

Performance tier analysis

Geographic/demographic breakdowns

Time-based cohort tracking

Comparative Analysis

Period-over-period comparisons

Benchmark analysis (internal/external)

Competitive positioning

Best practice identification

Gap analysis

4. REPORTING STRUCTURES & STANDARDS

Report Architecture

Executive Dashboards

KPI summaries

Trend visualizations

Exception highlighting

Action triggers

Operational Reports

Detailed metrics

Process performance

Resource utilization

Bottleneck identification

Analytical Deep Dives

Investigation findings

Statistical analysis

Predictive insights

Recommendation details

Pyramid Principle Structure

Lead with conclusions/recommendations

Support with key evidence

Provide detailed analysis

Include technical appendices

Visual Design Standards

Chart type selection matrix

Color coding conventions

Layout templates

Annotation guidelines

Accessibility requirements

5. INSIGHT GENERATION METHODOLOGIES

Pattern Recognition Techniques

Trend identification

Seasonality detection

Anomaly flagging

Correlation analysis

Cluster identification

Variance Analysis Framework

Actual vs. Plan/Forecast

Volume/Price/Mix decomposition

Time series variance

Geographic variance

Product/Service variance

Statistical Rigor

Significance testing

Confidence intervals

Sample size validation

Bias identification

Uncertainty quantification

6. QUALITY ASSURANCE & VALIDATION

Analysis Validation

Peer review processes

Sensitivity testing
Scenario stress-testing
Back-testing against historical data
Cross-validation techniques
Documentation Standards
Methodology documentation
Assumption cataloging
Limitation disclosure
Source attribution
Calculation transparency
Reproducibility Framework
Code versioning
Environment documentation
Data snapshot preservation
Process step recording
Audit trail maintenance

7. STAKEHOLDER MANAGEMENT

Communication Frameworks
Audience Segmentation
Executive briefings
Manager updates
Technical deep-dives
Cross-functional sharing
Delivery Cadence
Real-time alerts
Daily operational updates
Weekly business reviews
Monthly strategic reports

Quarterly business reviews

Feedback Integration

Iterative review cycles

Requirement refinement

Finding validation sessions

Action plan development

Impact measurement

8. ACTION & IMPLEMENTATION

Recommendation Framework

Specific action identification

Impact quantification

Resource requirement estimation

Implementation timeline

Success metrics definition

Decision Support Tools

What-if scenario modeling

ROI calculators

Risk assessment matrices

Priority scoring models

Implementation roadmaps

Performance Tracking

KPI definition and tracking

Initiative impact measurement

Continuous improvement cycles

Learning documentation

Best practice capture

9. TECHNOLOGY & TOOLS UTILIZATION

Analytics Platform Management

Tool selection criteria
Platform integration strategies
Performance optimization
Automation implementation
Scalability planning
Self-Service Analytics
User training programs
Template development
Governance frameworks
Usage monitoring
Quality control measures

10. PROFESSIONAL DEVELOPMENT

Skill Building Areas
Statistical methods
Business domain expertise
Communication skills
Technology proficiency
Project management
Industry Best Practices
Professional standards adoption
Methodology updates
Tool evaluation
Network building
Continuous learning

4. Function-Specific Deep Dives

Marketing Analytics Specialization

Customer Analytics Excellence

Segmentation Mastery: Behavioral segmentation, psychographic profiling, and dynamic segmentation Analytics Vidhya365 Data Science

Lifecycle Analysis: Customer journey mapping, cohort analysis, and retention modeling 365 Data Science

Value Modeling: Customer Lifetime Value (CLV) calculation, predictive CLV modeling, and value optimization Analytics Vidhya365 Data Science

Attribution and Measurement

Multi-Touch Attribution: Data-driven attribution modeling, cross-device tracking, and incrementality testing Medium +3

Marketing Mix Modeling: Adstock and saturation curves, media effectiveness measurement, and budget optimization Adsmurai +3

Experimental Design: A/B testing for marketing campaigns, multivariate testing, and causal inference 365 Data Science

Advanced Applications

Personalization Engines: Real-time recommendation systems, dynamic content optimization, and behavioral targeting Ironhack

Campaign Optimization: Automated bidding algorithms, creative optimization, and budget allocation

Brand Analytics: Brand sentiment analysis, brand lift measurement, and competitive brand analysis Analytics Vidhya

Different Types of Data Analyst Interview Questions

Interview Query regularly analyzes the contents of data analyst interviews. By tagging common keywords and mapping them back to question topics for over 10K+ tech companies, we've found that SQL questions are asked most frequently.

In fact, in interviews for data analyst roles, SQL and data manipulation questions are asked 85% of the time.

Here are the types of technical interview questions data analysts get asked most frequently:

Behavioral interview questions

SQL and data processing

Data analytics case studies

Python, algorithms, and coding questions

Statistics and probability

A/B testing and experimentation

Product metrics

Additionally, for more traditional data analyst roles, expect interview questions around the following:

Excel

Data Visualization

Let's first dive into how to approach and answer behavioral interview questions.

Question	Topic	Difficulty	Ask Chance
----------	-------	------------	------------

Causal Inference Without A/B			
------------------------------	--	--	--

A/B Testing & Experimentation			
-------------------------------	--	--	--

Medium			
--------	--	--	--

Very High			
-----------	--	--	--

Department Expenses			
---------------------	--	--	--

SQL			
-----	--	--	--

Medium			
--------	--	--	--

Very High			
-----------	--	--	--

Algorithm Reliability			
-----------------------	--	--	--

ML Ops & Training Pipelines			
-----------------------------	--	--	--

Hard

Very High

This feature requires a user account

Sign up to access this feature.

feature

Access 1000+ data science interview questions

feature

30,000+ top company interview guides

feature

Unlimited code runs and submissions

Behavioral Interview Questions for Data Analysts

Behavioral questions in data analyst interviews ask about specific situations you've been in in which you had to apply specific skills or knowledge.

For many data analysts, behavioral questions can be fairly tough.

One tip: Always try to relate the question back to your experience and strengths.

1. Describe a time when you spotted an inconsistency. How did you respond?

Successful data analysts can help businesses identify anomalies and respond quickly.

For data sense questions, think about a time you spotted an inconsistency in data quality and how you eventually addressed it.

2. Talk about a time when you had to make a decision with a lot of uncertainty.

Interviewers want to see you demonstrate the following:

Decisiveness – Show the interviewer that you can make decisions and communicate your decision-making process.

Self-direction – Show that you are able to choose a path forward, deduce information, and create a plan of action.

Adaptability – Your response should show that you can adapt your decision-making in a challenging situation.

Here's an example answer: "In my previous job, I worked on a sales forecasting problem under a strict deadline. However, I was missing the most recent data due to a processing error and only had 3-year-old sales figures. My strategy was applying the growth factor to the data to establish correct correlation and variances. This strategy helped me deliver a close forecast and meet the deadline."

3. How would you convey insights and the methods you use to a non-technical audience?

Interviewers ask this question to see if you can make complex subjects accessible and that you have a knack for communicating insights in a way that persuades people. Here's a marketing analytics example response:

"I was working on a customer segmentation project. The marketing department wanted to better segment users. I worked on a presentation and knew the audience wouldn't understand some of the more complex segmenting strategies. I put together a presentation that talked about the benefits and potential trade-offs of segmenting options like K-means clustering. For each option, I created a slide to show how it worked, and after the presentation, we could have an informed discussion about which approach to use."

4. How do you set goals and achieve them? Give us an example.

Interviewers want to see that you can set manageable goals and understand your process for achieving them. Don't forget to mention the challenges you faced, which will make your response more dynamic and insightful. For example, you might say:

"Data visualization was something I struggled with in college. I didn't have a strong design eye, and my visualizations were hard to read. In my last job, I made it a goal to improve, and there were two strategies that were most helpful. I took an online data visualization course, and I built a clip file of my favorite visualizations. The course was great for building my domain knowledge. However, I felt I learned the most by building my clip file and breaking down what made a good visualization on my own."

5. Describe a time when you solved a conflict at work.

This question assesses your ability to remain objective at work, communicate effectively in challenging situations, and remain calm under fire. Here's an example response:

"In my previous job, I was the project manager on a dashboard project. One of the BI engineers wasn't meeting the deadlines I had laid out, and I brought that up with him. At first, he was defensive and angry with me. But I listened to his concerns about the deadlines and asked what I could do to help. I learned from our conversation that he had a full workload besides this project. I talked with the engineering manager, and we were able to reduce some of his workload. He caught up quickly, and we were able to finish the project on time."

6. Give an example of a situation when you have shown effectiveness, empathy, humbleness, and adaptability.

This is a leadership question in disguise. If you can relate a time you were an effective leader, chances are you will easily incorporate all of these traits. For example:

"I was the lead on a marketing analytics project. We had a critical deadline to meet, but we were in danger of missing the deadline due to a data processing error. The team morale was low, so I held a quick meeting to lay out a schedule, answer questions, and rally the team. That meeting gave the team the jolt it needed. We made the deadline, and I ensured leadership knew how hard each contributor had worked."

7. Give me an example of a time when you failed on a project.

This question tests your resilience, how you respond to adversity, and how you learn from your mistakes. You could say:

"I had to give a client a presentation about a data analytics project. I mistakenly assumed the audience had more technical knowledge than they did. The presentation was received with a

lot of blank stares. However, I knew the material about our findings was strong. I stopped for questions, and then jumped ahead to the visualizations and findings. This helped get the presentation on track, and by the end, the client was impressed. Now, whenever I have a presentation, I take time to understand the audience before I start working on it."

8. Talk about an occasion when you used logic to solve a problem.

A strong response to this question shows that you can solve problems creatively and that you don't just jump at the first or easiest solution. One tip: Illustrate your story with data to make it more credible.

Here's what you could say: "In my previous job, I was responsible for competitor research, and through my analysis, I noticed that our most significant competitors had increased sales by 5% during Q1. This deviated significantly from our sales forecasts for these accounts. I found that we needed to update our competitor sales models with more recent market research and historical data. I tested the model adjustments, and ultimately, I improved our forecasting accuracy by 15%."

9. What do you do if you disagree with your manager?

Interviewers ask this question to gauge your emotional maturity, see that you can remain objective, and gain insights into your communication skills. Avoid subjective examples, such as my boss being a micromanager. Instead, you could say:

"One time, I disagreed with my manager over the process for building a dashboard, as their approach was to jump straight into the execution. I knew that it would be better to perform some planning in advance rather than feeling our way through and reacting to roadblocks as they arose, so I documented a plan that could potentially save us time in development. That documentation and planning showed where pitfalls were likely to arise, and by solving for future issues, we could launch the new dashboard three weeks early."

10. How comfortable are you with presenting insights to stakeholders?

This question is asked to see how confident you are in your communication skills, and it provides insight into how you communicate complex technical ideas. With this question, talk about how you make data and analytics accessible. Try to answer these questions:

Do you create visualizations?

What do you do to prepare for a data analytics presentation?

What strategies do you use to make data more accessible?

What presentation tools do you use?

11. Talk about a time you were surprised by the results of an analytics project.

This question basically asks: Are you open to new ideas in your work? Analysts can get stuck trying to prove their hypothesis, even if the data says otherwise. A successful analyst is OK with being wrong and listens to the data. You could say:

“While working on a customer analytics project, I was surprised to find that a subsegment of our customer base wasn’t actually responding to our offers. We had lumped the subsegment into a larger customer bucket and had assumed that a broader segmentation wouldn’t make a difference. I relayed the insight to the marketing team, and we reduced churn among this subsegment.”

12. Why are you interested in working for this company?

This question is super common in analyst behavioral interviews. However, it still trips a lot of candidates up. Another variation of this question would be: why did you want to work in data analytics?

In your response, your goal should be to convey your passion for the work and discuss what excites you about the company/role. You might focus on the company’s culture, a mentor who inspired you, a recommendation you received, or someone in your network who’s connected with the company. A sample response:

“I’m excited by the possibility of using data to foster stronger social connections amongst friends and peers. I also like to ‘go fast’ and experiment, which fits into Meta’s innovative culture.”

13. Talk about a time when you had trouble communicating with stakeholders. How were you able to overcome it?

Interviewers ask questions like this to assess how you handle adversity and adapt. Don’t be afraid to share what went wrong. B. Describe what you learned and how you will apply it to future work.

Here's a sample answer for a data analyst role: "I presented a data analytics project to non-technical stakeholders, but my presentation was far too technical. I realized that the audience wasn't following the technical aspects, so I stopped and asked questions. I spent time clarifying the technical details until there were no questions left. I learned that it's important to tailor presentations to the audience, so before I start a presentation, I always consider the audience."

Question	Topic	Difficulty	Ask Chance
----------	-------	------------	------------

Causal Inference Without A/B			
------------------------------	--	--	--

A/B Testing & Experimentation			
-------------------------------	--	--	--

Medium			
--------	--	--	--

Very High			
-----------	--	--	--

Department Expenses			
---------------------	--	--	--

SQL			
-----	--	--	--

Medium			
--------	--	--	--

Very High			
-----------	--	--	--

Algorithm Reliability			
-----------------------	--	--	--

ML Ops & Training Pipelines			
-----------------------------	--	--	--

Hard			
------	--	--	--

Very High			
-----------	--	--	--

This feature requires a user account

Sign up to access this feature.

feature

Access 1000+ data science interview questions

feature

30,000+ top company interview guides

feature

Unlimited code runs and submissions

SQL Interview Questions for Data Analysts

SQL Interview Questions for Data Analysts

Question	Topic	Difficulty	Ask Chance
----------	-------	------------	------------

Causal Inference Without A/B

A/B Testing & Experimentation

Medium

Very High

Department Expenses

SQL

Medium

Very High

Algorithm Reliability

ML Ops & Training Pipelines

Hard

Very High

This feature requires a user account

Sign up to access this feature.

feature

Access 1000+ data science interview questions

feature

30,000+ top company interview guides

feature

Unlimited code runs and submissions

Data analysts use SQL to query data to solve complex business problems or find answers for other employees. In general, SQL data analyst questions focus on analytics and reporting:

Basic SQL Questions - These include the basics, e.g., definitions and beginner SQL queries.

Analytics Questions – For analytics-based questions, you might have to understand what kind of report or graph to build first and then write a query to generate that report. So, it's an extra step on top of a regular SQL question.

Reporting Questions – SQL reporting questions replicate the work many data or business analysts do daily, e.g., writing queries.

Reporting interview questions focus on writing a query to generate an already-known output, such as producing a report or a metric given some example table.

For analytics-based questions, you might have to understand what kind of report or graph to build first and then write a query to generate that report. So, it's an extra step on top of a regular SQL question.

Basic SQL Interview Questions

14. What are the different ways of handling NULL when querying a data set?

To handle such a situation, we can use three different operations:

IS NULL – This operator returns true if the column value is NULL.

IS NOT NULL – This operator returns true if the column value is not NULL.

<=> – This operator compares values, which (unlike the = operator) is true even for two NULL values.

15. What's the difference between UNION and UNION ALL? (Asked by Facebook)

UNION and UNION ALL are SQL operators used to concatenate 2 or more result sets. This allows us to write multiple SELECT statements, retrieve the desired results, and then combine them together into a final, unified set.

The main difference between UNION and UNION ALL is that:

UNION: only keeps unique records

UNION ALL: keeps all records, including duplicates

16. What is the difference between an SQL view and a table? (Asked by Kaiser Permanente)

A table is structured with columns and rows. A view is a virtual table extracted from a database by writing a query.

17. What's the difference between an INNER and OUTER JOIN?

The difference between an inner and outer join is that inner joins result in the intersection of two tables, whereas outer joins result in the union of two tables.

18. What is the difference between WHERE and HAVING?

The WHERE clause is used to filter rows before grouping, and HAVING is used to exclude records after grouping.

19. When do you use the CASE WHEN function?

CASE WHEN lets you write complex conditional statements on the SELECT clause and also allows you to pivot data from wide to long formats.

20. What is the difference between DELETE TABLE and TRUNCATE TABLE in SQL?

Although they're both used to delete data, a key difference is that DELETE is a Database Manipulation Language (DML) command, while TRUNCATE is a Data Definition Language (DDL) command.

Therefore, DELETE is used to remove specific data from a table, while TRUNCATE removes all the table rows without maintaining the table's structure.

Another difference is that DELETE can be used with the WHERE clause, but TRUNCATE cannot. In this case, DELETE TABLE would remove all the data from the table while maintaining the structure. TRUNCATE would delete the entire table.

21. How would you pull the date from a timestamp in SQL?

EXTRACT allows us to pull temporal data types like date, time, timestamp, and interval from date and time values.

22. Write an SQL query to select all employees' records with last names between "Bailey" and "Frederick."

For this question, assume the table is called "Employees" and the last name column is "LastName".

```
SELECT * FROM Employees WHERE LastName BETWEEN 'Bailey' AND 'Frederick'
```

23. What is the ISNULL function? When would you use it?

The ISNULL function returns an alternative value if an expression is NULL. Therefore, if you wanted to add a default value for NULL values, you would use ISNULL. For example, in the statement:

```
SELECT name, ISNULL(price, 50) FROM PRODUCTS
```

NULL price values would be replaced with 50.

Reporting SQL Questions

24. We have a table with an ID and name field. The table holds over 100 million rows, and we want to sample a random row in the table without throttling the database. Write a query to sample a row from this table randomly.

big_table

Column	Type
--------	------

id	INTEGER
----	---------

```
name VARCHAR
```

In most SQL databases, there exists a RAND() function, which normally we can call:

```
SELECT * FROM big_table  
ORDER BY RAND()
```

The function will randomly sort the rows in the table. This function works fine and is fast if you only have, let's say, around 1,000 rows. It might take a few seconds to run at 10K. And then at 100K, maybe you have to go to the bathroom or cook a meal before it finishes.

What happens at 100 million rows?

Someone in DevOps is probably screaming at you.

Random sampling is important in SQL with scale. We don't want to use the pre-built function because it wasn't meant for performance. But maybe we can re-purpose it for our own use case.

We know that the RAND() function actually returns a floating point between 0 and 1. So, if we were to instead call:

```
SELECT RAND()
```

We would get a random decimal point to some Nth degree of precision. RAND() essentially allows us to seed a random value. How can we use this to select a random row quickly?

Let's try to grab a random number using RAND() from our table that can be mapped to an ID. Given we have 100 million rows, we probably want a random number from 1 to 100 million. We can do this by multiplying our random seed from RAND() by the maximum number of rows in our table.

```
SELECT CEIL(RAND() * (  
    SELECT MAX(id) FROM big_table)
```

)

We use the CEIL function to round the random value to an integer. We must return to our existing table to get the value.

What happens if we have missing or skipped ID values, though? We can solve this by running the join on all the IDs that are greater or equal than our random value and selecting only the direct neighbor if a direct match is impossible.

Once one row is found, we stop (LIMIT 1). And we read the rows according to the index (ORDER BY id ASC). Now, our performance is optimal.

```
SELECT r1.id, r1.name
FROM big_table AS r1
INNER JOIN (
    SELECT CEIL(RAND() * (
        SELECT MAX(id)
        FROM big_table)
    ) AS id
) AS r2
ON r1.id >= r2.id
ORDER BY r1.id ASC
LIMIT 1
```

25. Given a table of job postings, write a query to break down the number of users that have posted their jobs once versus the number of users that have posted at least one job multiple times.

Hint: We want the value of two different metrics, the number of users that have posted their jobs once and the number of users that have posted at least one job multiple times. What does that mean exactly?

26. Write a query to get the current salary for each employee.

More context. Let's say we have a table representing a company payroll schema.

Due to an ETL error, the employees table, instead of updating the salaries every year when doing compensation adjustments, did an insert instead. The head of HR still needs the current salary of each employee.

27. Write a query to get the total amount spent on each item in the ‘purchases’ table by users that registered in 2023.

More context. Let’s say you work at Costco. Costco has a database with two tables. The first is users, composed of user information, including their registration date, and the second table is purchases, which has the entire item purchase history (if any) for those users.

Here’s a process you can use to solve this question:

You can use INNER JOIN or JOIN to connect tables users and purchases on the user_id column

You can filter the results by using the WHERE clause

Use GROUP BY to aggregate items and apply the SUM() function to calculate the amount spent

28. Write a query to get the cost of all transactions by user ordered by total cost descending.

Here’s a code solution:

```
SELECT
    u.name
    ,u.id AS user_id
    ,ROUND(SUM(p.price * t.quantity ) ,2) AS total_cost
FROM users u
INNER JOIN transactions t
    ON u.id = t.user_id
INNER JOIN products p
    ON p.id = t.product_id
GROUP BY u.name
```

ORDER BY total_cost DESC

29. Given a table of transactions and a table of users, write a query to determine if users tend to order more to their primary address versus other addresses.

Hint: This question has been asked in Amazon data analyst interviews, and the first step is getting data from the users table to the transactions table. This can be done using a JOIN based on a common column between the tables. How do we identify when the addresses match? We can use the CASE WHEN statement to produce a flag for further calculations. Finally, we need the percentage of all the transactions made to the primary address rounded to two decimals.

30. Write a query to get the top three users with the most upvotes on their comments.

You're provided with three tables representing a forum of users and their comments on posts and are asked to find the top three users with the most upvotes in the year 2020. Additionally, we're told that upvotes on deleted comments and upvotes that users make on their own comments don't matter.

Hint: The trickiest part about this question is performing your JOINs on the proper fields. If you join two of our tables on the wrong key, you could make things difficult, or even impossible, for yourself later on.

31. Write a query to identify customers who placed more than three transactions in 2019 and 2020.

In this question, you're given two transactions and users.

Hint: Start by joining the transactions and users tables. Use INNER JOIN or JOIN.

Analytics SQL Questions

32. Given a table of search results, write a query to compute a metric to measure the quality of the search results for each query.

search_results table

Column	Type
--------	------

query	VARCHAR
-------	---------

```
result_id      INTEGER  
position       INTEGER  
rating        INTEGER
```

You want to be able to compute a metric that measures the precision of the ranking system based on position. For example, if the results for dog and cat are....

query	result_id	position	rating	notes
dog	1000	1	2	picture of hotdog
dog	998	2	4	dog walking
dog	342	3	1	zebra
cat	123	1	4	picture of cat
cat	435	2	2	cat memes
cat	545	3	1	pizza shops

...we would rank 'cat' as having a better search result ranking precision than 'dog' based on the correct sorting by rating.

Write a query to create a metric to validate and rank the queries by their precision of search results, round the metric (avg_rating column) to 2 decimal places.

33. Given the two tables, write an SQL query that creates a cumulative distribution of a number of comments per user. Assume bin buckets class intervals of one.

Hint: What is a cumulative distribution exactly? What would the dataset look like if we were to imagine our output and figure out what we wanted to display on a cumulative distribution graph?

34. We are given a table of bank transactions with three columns: user_id, a deposit or withdrawal value (determined if the value is positive or negative), and created_at time for each transaction.

Write a query to get the total three-day rolling average for daily deposits.

Usually, if the problem states to solve for a moving/rolling average, we're given the dataset as a table with two columns: the date and the value.

This problem, however, is taken one step further with a table of just transactions with values conditioned to filtering for only deposits and removing records representing withdrawals denoted by a negative value (e.g., -10).

35. Given a table of user experiences representing each person's work experiences, write a query to determine if a data scientist gets promoted faster or if they switch jobs more frequently.

More context. Let's say we're interested in analyzing the career paths of data scientists. Job titles are bucketed into data scientist, senior data scientist, and data science manager. We're interested in determining if a data scientist who switches jobs more often gets promoted to a manager role faster than a data scientist who stays at one job longer.

This question has been asked in Google data analyst interviews, and it requires a bit of creative problem-solving to understand how we can prove or disprove the hypothesis. The hypothesis is that data scientists who end up switching jobs more often get promoted faster.

Therefore, in analyzing this dataset, we can prove this hypothesis by separating the data scientists into specific segments based on how often they jump into their careers. How would you do that?

36. Write a query to get the distribution of the number of conversations created by each user by day in the year 2020.

Our focus is getting our key metric of the number of new conversations created daily in a single query. To get this metric, we have to group by the date field, and then group by the distinct number of users messaged. Afterward, we can then group by the frequency value and get the total count of that as our distribution.

37. Write a query that could display the percentage of users on our forum that would be acting fraudulently in this manner.

More context. We're given three tables representing a forum of users and their comments on posts. We want to determine if users create multiple accounts to upvote their comments. What kind of metrics could we use to figure this out?

38. Uber users are complaining that the pick-up map is wrong. How would you verify how frequently this is actually happening?

Hint. What metric would help you investigate this problem?

39. What strategies could we try to implement to increase the outreach connection rate?

More context. Let's say that Facebook account managers cannot reach business owners after repeated calls to try to onboard them onto a new Facebook business product. Assume that we have training data on all of the account manager's outreach in terms of calls made, calls picked up, time of call, etc...

One option would be to investigate when calls are most likely to be connected. Could changing our approach here improve the connection rate?

40. You're analyzing churn on Facebook. How would you investigate if a disparity exists in retention on different Facebook platforms?

Follow-up question. How would you investigate the causes of such a disparity?

Data Analytics Case Study

Data analysis case study questions combine a rotating mix of product intuition, business estimation, and data analytics.

Case questions come up in interviews when the job responsibilities lean to more of a heavy analytics space with an emphasis on solving problems and producing insights for management.

Many times, data analysts will transition into a heavy analytics role when they're required to take on more scope around the product and provide insights that upper-level management can understand and interpret.

So, data analytics case study questions will focus on a particular problem, and you will be judged on how you break down the question, analyze the problem, and communicate your insights.

Here's an example data analytics case study question:

41. Given a table of Stack Overflow posts data, suggest three metrics to monitor the community's health.

Community members can create a post to ask a question, and other users can reply with answers or comments to that question. The community can express their support for the post by upvoting or downvoting.

post_analytics table:

Column	Type	Description
id	int	Primary key of posts table
user_idint		ID of the user who created the post
created_at	datetime	Timestamp of the post
title	string	Title of the post
body	string	Text content of the post
comment_count	int	Total number of the comments on a post
view_count	int	Total number of the views on a post
answer_count	int	Total number of answers on a post
upvotes	int	Total number of upvotes on the post

More context. You work at Stack Overflow on the community team that monitors the platform's health. Community members can create a post to ask a question, and other users can reply with answers or comments to that question. The community can express their support for the post by upvoting or downvoting.

42. Write the queries for these metrics in SQL.

This is a classic data analytics case study. A question like this is designed to assess your data intuition, product sense, and ability to isolate key metrics.

Remember: There isn't one correct answer, but usually, the conversation should head in a similar direction.

For example, this question asks about community health. Broadly, there are several metrics you'll want to consider: Growth rate, engagement, and user retention, which would provide insights into the community's health.

The challenge with this question is determining how to measure those metrics with the provided data.

43. Describe an analytics experiment that you designed. How were you able to measure success?

Case questions sometimes take the form of behavioral questions. Data analysts get tasked with experimenting with data to test new features or campaigns. Many behavioral questions will ask about experiments but also tap into how you approach measuring your results.

With questions like these, be sure to describe the objective of the experiment, even if it is a simple A/B test. Don't be afraid to get technical and explain the metrics and process you used to quantify the results.

44. An online marketplace introduces a new feature that lets buyers and sellers conduct audio chats. Write a query to indicate whether the feature is successful or not.

Bonus question. How would you measure the success of this new feature?

See a step-by-step solution to this data analytics case study problem.

45. Write a query to prove or disprove the hypothesis: CTR depends on the search result rating.

More context. You're given a table that represents search results from searches on Facebook. The query column is the search term, the position column represents each position the search result came in, and the rating column represents the human rating from 1 to 5, where 5 is high relevance, and 1 is low relevance.

Each row in the search_eventstable represents a single search, with the has_clicked column representing whether a user clicked on a result or not. We hypothesize that the CTR depends on the search result rating.

46. A revised new-user email journey boosts conversion rates from 40% to 43%. However, a few months prior, CVR was 45%. How would you investigate if the new email journey caused the increase in CVR?

See a step-by-step solution to this problem on YouTube.

Python Coding Questions for Data Analysts

Python coding questions for data analysts are usually pretty simple and not as difficult as the ones seen on Leetcode. Most interviewers want to test their basic knowledge of Python to the point that they can write scripts or some basic functions to move data between SQL and Excel or onto a dashboard. These can then be said to only be the basic Python interview questions.

Most data analysts never write production code, and their code is never under scrutiny because it's not holding a website up or performing some critical business function.

Therefore, most coding questions for data analyst interviews are generally easier and mostly test basic functions required for data manipulation. Pandas questions may also be asked in this round of the interview.

Here's an example Python coding question:

47. Write a function that can take a string and return a list of bigrams. (Asked by Indeed)

sentence = "Have free hours and love children?"

```
output = [  
    ('have', 'free'),  
    ('free', 'hours'),  
    ('hours', 'and'),
```

```
('and', 'love'),  
('love', 'children')  
]
```

Bigrams are two words that are placed next to each other. To parse them out of a string, we must first split the input string.

We would use the Python function `.split()` to create a list with each individual word as an input. Create another empty list that will eventually be filled with tuples.

Then, once we've identified each individual word, we need to loop through $k-1$ times (if k is the number of words in a sentence) and append the current word and subsequent word to make a tuple. This tuple gets added to a list that we eventually return. Remember to use the Python function `.lower()` to turn all the words into lowercase!

```
def find_bigrams(sentence):  
    input_list = sentence.split()  
    bigram_list = []  
  
    # Now we have to loop through each word  
    for i in range(len(input_list)-1):  
        #strip the whitespace and lower the word to ensure consistency  
        bigram_list.append((input_list[i].strip().lower(), input_list[i+1].strip().lower()))  
    return bigram_list
```

48. Explain negative indexing. What purpose does it serve?

Negative indexing is a function in Python that allows users to index arrays or lists from the last element. For example, the value `-1` returns the last element, while `-2` returns the second-to-last element. It is used to display data from the end of a list or to reverse a number or string.

Example of negative indexing:

```
a = "Python Data Analyst Questions"  
print (a[-1])  
>> s
```

49. What is a compound data type in Python?

Compound data structures are single variables that represent multiple values. Some of the most common in Python are:

Lists - A collection of values where the order is important.

Tuples - A sequence of values where the order is important.

Sets - A collection of values where membership in the set is important.

Dictionaries - A collection of key-value pairs where you can access values based on their keys.

50. What is the difference between Python lists, tuples, and sets? When should you use one over the other?

Lists, tuples, and sets are compound data types that serve a similar purpose: storing collections of items in Python. However, knowing the differences between each of them is crucial for computing and memory efficiency.

Lists are mutable collections that are ordered and allow duplicate elements. They are versatile and offer various operations, such as accessing, adding, and removing items. They are suitable when the order of items matters or when you need to change the collection over time.

Tuples are similar to lists in order collections. However, they are immutable, meaning you cannot change their content once defined. Tuples are faster than lists and can be used in situations where the content will remain constant.

Sets are unordered collections that do not allow duplicate elements. Because they are unordered, you cannot access elements by an index. Sets are faster than lists and tuples for membership testing, i.e., checking if an item is in the collection. They are also beneficial when removing duplicates from a collection or performing mathematical set operations such as union, intersection, and difference.

51. How would you find duplicate values in a dataset for a variable in Python?

You can check for duplicates using the Pandas duplicated() method. This will return a boolean series, which is TRUE only for unique elements.

```
DataFrame.duplicated(subset=None,keep='last')
```

52. What is list comprehension in Python? Provide an example.

List comprehension defines and creates a list based on an existing one. For example, if we wanted to separate all the letters in the word “retain” and make each letter a list item, we could use list comprehension:

```
r_letters = [ letter for letter in 'retain' ]  
print( r_letters)
```

We can also use list comprehension for filtering. For example, to get all the vowels in the word “retain,” we do the following:

```
vowels = [vowel for vowel in 'retain' if vowel in ('a', 'e', 'i', 'o', 'u')]  
print(vowels)
```

If you are concerned about duplicate values, you can opt for sets instead by replacing “[]” with “{}”.

```
unique_vowels = {vowel for vowel in 'retain' if vowel in ('a', 'e', 'i', 'o', 'u')}
```

```
print(unique_vowels)
```

53. What is sequence unpacking? Why is it important?

Sequence unpacking is a Python operation that allows you to de-structure the elements of a collection and assign them directly to variables without the need for iteration. It provides a terse method for mapping variables to the elements of a compound data structure. For example:

```
# instead of:
```

```
x = coordinates[0]  
y = coordinates[1]
```

```
# we can unpack a list:
```

```
x, y = coordinates
```

```
# we can also do the same for sets, tuples, and dictionaries.
```

We can even swap the elements of two variables without the use of a third variable:

```
a = 3
```

```
b = 2
```

```
a, b = b, a
```

```
assert a == 2
```

```
assert b == 3
```

```
# no assertion errors
```

If the size of a collection is unclear, you can use the * operator on a variable to assign all extra items to said variable:

```
food = ('apples', 'oranges', 'carrots', 'cabbages', 'lettuce')
```

```
apples, oranges, *vegetables = food
```

```
# apples = 'apples', oranges = 'oranges',
```

```
# vegetables = ('carrots', 'cabbages', 'lettuce')
```

54. Write a function that takes in a list of dictionaries with a key and a list of integers and returns a dictionary with the standard deviation of each list.

Hint: need to use the equation for standard deviation to answer this question. Using the equation allows us to take the sum of the square of the data value minus the mean over the total number of data points, all in a square root.

55. Given a list of timestamps in sequential order, return a list of lists grouped by week (7 days) using the first timestamp as the starting point.

This question sounds like it should be an SQL question, doesn't it? Weekly aggregation implies a form of GROUP BY in a regular SQL or pandas question. In either case, aggregation on a dataset of this form by week would be pretty trivial.

56. Given two strings A and B, return whether or not A can be shifted some number of times to get B.

Example:

```
A = 'abcde'
```

```
B = 'cdeab'
```

```
can_shift(A, B) == True
```

```
A = 'abc'
```

```
B = 'acb'
```

```
can_shift(A, B) == False
```

Hint: This problem is relatively simple if we work out the underlying algorithm that allows us to easily check for string shifts between the strings A and B.

57. Given two strings, string1 and string2, write a function is_subsequence to determine if string1 is a subsequence of string2.

Hint: Notice that in the subsequence problem set, one string in this problem will need to be traversed to check for the values of the other string. In this case, it is string2.

Statistics and Probability Interview Questions

Statistics and probability questions for data analysts will usually come up on an onsite round as a test of basic fundamentals.

Statistics questions are more likely than probability questions to show up, as statistics are the fundamental building blocks for many analyst formulas and calculations.

58. Given uniform distributions X and Y and the mean 0 and standard deviation 1 for both, what's the probability of $2X > Y$? (Asked by Snapchat)

Given that X and Y both have a mean of 0 and a standard deviation of 1, what does that indicate for the distributions of X and Y?

Let's look at this question a little closer.

We're given two normal distributions. The values can either be positive or negative, but each value is equally likely to occur. Since we know the mean is 0 and the standard deviation is 1, we understand that the distributions are also symmetrical across the Y-axis.

In this scenario, we are equally likely to randomly sample a value that is greater than 0 or less than 0 from the distribution.

Now, let's take examples of random values that we could get from each scenario. There are about six different scenarios here.

$X > Y$: Both positive

$X > Y$: Both negative

$X < Y$: Both positive

$X < Y$: Both negative

$X > Y$: X is positive, Y is negative

$X < Y$: X is negative, Y is positive

We can simulate a random sampling by equating that all six are equally likely to occur. If we play out each scenario and plug the variables into $2X > Y$, then we see about half of the time the statement is true, or 50%.

Why is this the case? Generally, let's return to the fact that both distributions are symmetrical across the Y-axis. We can intuitively understand that if both X and Y are random variables across the same distribution, we will see $2X$ as being, on average, double positive or double negative the value that Y is.

59. What is an unbiased estimator, and can you provide an example for a layman to understand?

To answer this question, consider how a biased estimator looks. Then, think about how an unbiased estimator differs. Ultimately, an estimator is unbiased if its expected value equals the true value of a parameter, meaning that the estimates are in line with the average.

60. Let's say we have a sample size of N. The margin of error for our sample size is 3. How many more samples would we need to decrease the margin of error to 0.3?

Hint: In order to decrease our margin of error, we'll probably have to increase our sample size. But by how much?

61. What's the Difference Between Correlation and Covariance?

Covariance measures the linear relationship of variables, while correlation measures the strength and direction of the relationship. Therefore, correlation is a function of covariance. For example, a correlation between two variables does not mean that the change in variable X caused the change in variable Y's value.

62. How would you describe probability distribution to a non-technical person?

Probability distributions represent random variables and associated probabilities of different outcomes. In essence, a distribution maps the probability of various outcomes.

For example, a distribution of test grades might look similar to a normal distribution, AKA bell curve, with the highest number of students receiving Cs and Bs and a smaller percentage of students failing or receiving a perfect score. In this way, the center of the distribution would be the highest, while outcomes at either end of the scale would fall lower and lower.

63. What is a non-normal distribution? Provide an example.

A probability distribution is abnormal if most observations do not cluster around the mean, forming the bell curve. An example of a non-normal probability distribution is a uniform distribution, in which all values are equally likely to occur within a given range. A random number generator set to produce only the numbers 1-5 would create such a non-normal distribution, as each value would be equally represented in your distribution after several hundred iterations.

64. What is the probability that it's raining in Seattle?

More context. You are about to get on a plane to Seattle. You call 3 random friends in Seattle and ask each other if it's raining. Each has a $\frac{2}{3}$ chance of telling you the truth and a $\frac{1}{3}$ chance of messing with you by lying. All 3 friends tell you that "yes" it is raining.

Hint: There are several ways to answer this question. Given that a frequentist approach operates on the set of known principles and variables given in the original problem, you can logically deduce that $P(\text{Raining}) = 1 - P(\text{Not Raining})$.

Since all three friends have given you the same answer as to whether or not it's raining, what can you determine about the relationship between $P(\text{Not Raining})$ and the probability that each of your friends is lying?

65. How would you design a function to detect anomalies if given a univariate dataset? What if the data is bivariate?

Before jumping into anomaly detection, discuss the meaning of a univariate dataset. Univariate means one variable. For example, travel time in hours from your city to 10 other cities is given in an example list below:

12, 27, 11, 41, 35, 22, 18, 43, 26, 10

This kind of single column data set is called a univariate dataset. Anomaly detection is a way to discover unexpected values in datasets. The anomaly means data exists that is different from the normal data. For example, you can see below the dataset where one data point is unexpectedly high intuitively:

12, 27, 11, 41, 35, 22, 76767676, 18, 43, 26, 10

66. You want to look at the mean and median for a dataset. When would you use one measure over the other? How do you calculate the confidence interval of each measure?

You should answer these questions in your response:

Which measure has the widest application?

What happens when the dataset has values that are way above or below most other values?

How would your choice of metric be influenced by the non-continuous data?

67. You have a biased and unbiased coin. You select a random coin and flip it two times. What is the probability that both flips result in the same side?

Hint: The first step in solving this problem is to separate it into two instances— one where you grab the fair coin and one where you grab the biased coin. Solve for the probabilities of flipping the same side separately for both.

68. What could be the cause of a capital approval rate decrease?

Capital approval rates have gone down compared to our overall approval rate. Let's say last week it was 85%, and the approval rate went down to 82% this week, which is a statistically significant reduction.

The first analysis shows that all approval rates stayed flat or increased over time when looking at the individual products.

Product 1: 84% to 85% week over week

Product 2: 77% to 77% week over week

Product 3: 81% to 82% week over week

Product 4: 88% to 88% week over week

Hint: This would be an example of Simpson's Paradox, a phenomenon in statistics and probability. Simpson's Paradox occurs when a trend shows in several groups but either disappears or is reversed when combining the data.

69. How would you explain confidence intervals?

In probability, confidence intervals refer to a range of values you expect your estimate to fall between if you rerun a test. Confidence intervals are a range that is equal to the mean of your estimate plus or minus the variation.

For example, if a presidential popularity poll had a confidence interval of 93%, encompassing a 50%-55% approval, it would be expected that, if you re-polled your sample 100 more times, 93 times the estimate would fall between the upper and lower values of your interval. Those other seven events would fall outside, which is to say either below 50% or above 55%. More polling would allow you to get closer to the true population average and narrow the interval.

70. You must draw two cards from a shuffled deck, one at a time. What's the probability that the second card is not an ace?

One question to add: does order matter here? Is drawing an ace on the second card the same as drawing an ace on the first and still drawing a second card? Let's see if we can solve this and prove it.

We can generalize to two scenarios when drawing two cards to get an ace:

Drawing an ace on the first card and an ace on the second card

Drawing not an ace on the first card and an ace on the second card

If we model the probability of the first scenario, we can multiply the two probabilities of each occurrence to get the actual probability.

Question	Topic	Difficulty	Ask Chance
Causal Inference Without A/B			
A/B Testing & Experimentation			
Medium			
Very High			
Department Expenses			
SQL			
Medium			
Very High			
Algorithm Reliability			
ML Ops & Training Pipelines			
Hard			
Very High			
This feature requires a user account			
Sign up to access this feature.			
feature			
Access 1000+ data science interview questions			

feature

Access 1000+ data science interview questions

feature

30,000+ top company interview guides

feature

Unlimited code runs and submissions

A/B Testing and Experimentation

A/B testing and experimentation questions for data analysts tend to explore the candidate's ability to conduct A/B tests properly. You should have a strong knowledge of p-values and confidence intervals and be able to assess the experiment's validity.

71. The PM checks the results of an A/B test (standard control and variant) and finds a .04 p-value. How would you assess the validity of the result? How would you assess the validity of the result?

In this particular question, you'll need to clarify the context of how the A/B test was set up and measured.

If we have an A/B test to analyze, there are two main ways in which we can look for invalidity. We could likely re-phrase the question: How do you correctly set up and measure an A/B test?

Let's start out by answering the first part of figuring out the validity of the setup of the A/B test:

1. How were the user groups separated?

Can we determine that the control and variant groups were sampled according to the test conditions?

If we're testing changes to a landing page to increase conversion, can we compare the two different users in the groups to see different metrics in which the distributions should look the same?

For example, if the groups were randomly bucketed, would the distribution of traffic from different attribution channels still look similar, or would the variant A traffic channel come primarily from Facebook ads and the variant B from email? If testing group B has more email traffic, that could be a biased test.

2. Were the variants equal in all other aspects?

The outside world often has a much larger effect on metrics than product changes do. Users can behave very differently depending on the day of the week, the time of year, the weather (especially for a travel company like Airbnb), or whether they learned about the website through an online ad or found it organically.

If variant A's landing page has a picture of the Eifel Tower and the submit button on the top of the page, and variant B's landing page has a large picture of an ugly man and the submit button on the bottom of the page, then we could get conflicting results based on the change to multiple features.

Measurement

Looking at the actual measurement of the p-value, we understand that the industry standard is .05, which means that 19 out of 20 times that we perform that test, we're going to be correct that there is a difference between the populations.

However, we have to note some things about the test in the measurement process.

What was the sample size of the test?

How long did it take before the product manager measured the p-value? Lastly, how did the product manager measure the p-value, and did they do so by continually monitoring the test?

If the product manager ran a T-test with a small sample size, they could very well easily get a p-value under 0.05. Often, the source of confusion in AB testing is how much time you need to make a conclusion about the results of an experiment.

The problem with using the p-value as a stopping criterion is that the statistical test that gives you a p-value assumes that you designed the experiment with a sample and effect size in mind. If we continuously monitor the development of a test and the resulting p-value, we are very likely to see an effect, even if there is none. The opposite error is common when you stop an experiment too early before an effect becomes visible.

The most important reason is that we perform a statistical test every time you compute a p-value, and the more you do it, the more likely you are to find an effect.

How long should we recommend running an experiment for then? To prevent a false negative (a Type II error), the best practice is to determine the minimum effect size that we care about and compute based on the sample size (the number of new samples that come every day) and the certainty you want, how long to run the experiment for, before starting the experiment.

72. How can you effectively design an A/B test? Are there times when A/B testing shouldn't be used?

Split testing fails when you have unclear goals. That's why it's imperative to start backward with that goal. Is it to increase conversions? Are you trying to increase engagement and time spent on the page? Once you have that goal, you can start experimenting with variables.

73. How much traffic would you need to drive to a page for the result of an A/B test to be statistically significant?

Statistical significance - having 95% confidence in the results - requires the right volume of data. That's why most A/B tests run for 2-8 weeks. Comparing metrics like conversions is fairly easy to calculate. In fact, most A/B tools have built-in calculators.

74. How would you conduct an experiment to test a new ETA estimate feature in Uber? How would you know if your results were significant?

Hint: This question asks you to think hypothetically about A/B testing. But the format is the same: Walk the interviewer through setting up the test and how you arrive at a statistically relevant result.

75. How would you explain P-value to someone who is non-technical?

The p-value is a fundamental concept in statistical testing. First, why does this kind of question matter? What an interviewer is looking for here is whether you can answer this question in a way that conveys your understanding of statistics and can also answer a question from a non-technical worker who doesn't understand why a p-value might matter.

For example, if you were a data scientist and explained to a PM that the ad campaign test has a .08 p-value, why should the PM care about this number?

76. Your company wants to test new marketing channels. How would you design an A/B test for the most efficient marketing spend?

The new channels include YouTube ads, Google search ads, Facebook ads, and direct mail campaigns.

First, you'd want to follow up with clarifying questions and make some assumptions. Let's assume, for example, that the most efficient means the lowest cost per conversion and that we've been asked to spend evenly across all platforms.

77. You want to run an experiment but find that the distribution of the dataset is not normal. What kind of analysis would you run, and how would you measure which variant won?

Understanding whether your data abides by or violates a normal distribution is an important first step in your subsequent data analysis.

This understanding will change which statistical tests you want to use if you need to look for statistical significance immediately. For example, you cannot run a t-test if your distribution is non-normal since this test uses mean/average as a way to find differences between groups.

78. You want to A/B test pricing levels for subscriptions. The PM asks you to design a two-week test. How do you approach this? How do you determine if the pricing increase is a good business decision?

Hint: Is A/B testing a price difference a good idea? Would it encourage users to opt out of your test if they saw different product prices?

Is there a better way to test pricing?

79. A survey shows that app users who use an optional location-sharing feature are “less happy” with the app as a whole. Is the feature actually causing users to be unhappy?

Causal relationships are hard to come by, and truly determining causality is tough as the world is full of confounding variables. Because of this, instead of causality, we can dissect the correlation between the location-sharing feature and the user unhappiness level.

At its core, this interview question tests how to dig into the science and statistics behind their assumption. The interviewer is essentially asking a small variation of a traditional experimental design with survey research and wants to know how you would either validate or disprove this claim.

Product Metrics Data Analyst Questions

Metrics is a common product analyst interview question subject, and you’ll also see this type of question in product-oriented data analyst roles. In general, these questions test your ability to choose metrics to investigate problems or measure success. These questions require a strong product sense to answer.

80. You’re given a list of marketing channels and their costs. What metrics would you use to determine the value of each marketing channel?

The first thing we’ll want to do when faced with an interview question like this one is to ask a few clarifying questions. Answer these questions first:

What is the company’s business model?

Is there one product or many?

Let’s say it’s a SaaS business that offers a free Studio model of their product but makes their money selling enterprise subscriptions. This gives us a better sense of how they’re approaching their customers. They’re saying: here’s a good free tool, but you can pay to make it even better.

How many marketing channels are there?

Imagine what your analysis would look like if the answer to this question was “a few.” Now imagine what your analysis would look like if the answer to this question was “hundreds.”

Are some marketing channels bigger than others? What’s the proportion?

Mode could be spending 90% of its marketing budget on Facebook Ads and 10% on affiliate marketing, or vice versa. We can’t know unless we ask.

What is meant by “the value of each marketing channel?”

Here’s where we start getting into the meat of the question.

81. A PM at Facebook comes to you and tells you that friend requests are down 10%. What do you do?

This question has been asked in Facebook data analyst interviews. See an example solution to this question on YouTube.

82. What are some reasons why the average number of comments per user would be decreasing, and what metrics would you look into?

More context. Let’s say you work for a social media company that has just done a launch in a new city. Looking at weekly metrics, you see a slight decrease in the average number of comments per user from January to March in this city. The company has been consistently growing new users in the city from January to March.

Let’s model an example scenario to help us see the data.

Jan: 10000 users, 30000 comments, 3 comments/user

Feb: 20000 users, 50000 comments, 2.5 comments/user

Mar: 30000 users, 60000 comments, 2 comments/user

We're given information that the total user count is increasing linearly, which means that the decreasing comments/user is not an effect of a declining user base creating a loss of network effects on the platform. What else can we hypothesize, then?

83. How would you measure the success of Facebook Groups?

Start here: What is the point of Facebook Groups? Primarily we could say Facebook Groups provides a way for Facebook users to connect with other users through a shared interest or real-life/offline relationship.

How could we use the goals of Facebook Groups to measure success?

84. What kind of analysis would you conduct to recommend UI changes?

More context. You have access to a set of tables summarizing user event data for a community forum app. You're asked to conduct a user journey analysis using this data with the eventual goal of improving the user interface.

85. How would you measure the success of Uber Eats?

See a step-by-step solution for this question on YouTube.

86. What success metrics would you be interested in for an advertising-driven consumer product?

With this question, you might define success in terms of advertising performance. A few metrics you might be interested in are:

CTR

CPC

Pageviews or daily actives (for apps)

Conversion rate

Number of purchases

Cost per conversion

87. How do success metrics change by product type?

Let's look at two examples: An eCommerce product like Groupon vs. a subscription product like Netflix.

E-commerce metrics tend to be related to conversions and sales. Therefore, you might be interested in the number of purchases, conversion rate, quarterly or monthly sales, and cost of goods sold.

Subscription products tend to focus more on subscriber costs and revenue, like churn rates, cost of customer acquisition, average revenue per user, lifetime value, and monthly recurring revenue.

88. Given a dataset of raw events, how would you come up with a measurement to define what a “session” is for the company?

More context. Let's say you're given event data from users on social networking sites like Facebook. A product manager is interested in understanding the average number of “sessions” that occur every day. However, the company has not technically defined what a “session” is yet.

The best the product manager can do is illustrate an example of a user browsing Facebook in the morning on their phone and then again during lunch as two distinct “sessions.” There must be a period of time when the user leaves Facebook to do another task before coming back again anew.

89. Some of the success metrics for the LinkedIn newsfeed algorithm are going up, while others are going down. What would you look at?

See a solution for this question on YouTube.

90. The number of products or subscriptions sold is declining. How would you investigate this problem?

This question provides you with a chance to show your expertise in analyzing sale metrics and KPIs. Some of the challenges you might bring up include competitor price analysis, examining core customer experiences, and investigating evolving customer desires. Your goal in your response should be to outline how you would perform root cause analysis.

Tip. Start with some clarifying questions like, What is the product? Who is the audience? How long has the decline in sales persisted?

91. You're asked to investigate how to improve search results. What metrics would you investigate? What would you look at to determine if current search results are effective?

More context. Specifically, we want to improve search results for people looking for things to do in San Francisco.

92. Let's say you work on the growth team at Facebook and are tasked with promoting Instagram from within the Facebook app. Where and how could you promote Instagram through Facebook?

This product question is more focused on growth and is very much used for Facebook's growth marketing analyst technical screen. Here are a couple of things that we have to remember.

Like usual product questions where we are analyzing a problem and coming up with a solution with data, we have to do the same with growth, except we have to come up with solutions in the form of growth ideas and provide data points for how they might support our hypothesis.

93. How would you measure success for Facebook Stories?

Measuring the success of Facebook Stories requires an integrated approach that examines how users interact with the feature and its impact on the platform. Key to this evaluation is understanding engagement levels, which are reflected through metrics such as the total number of story views and unique viewers, alongside interactions like replies and reactions. These figures are pivotal because they indicate not just how many people are watching, but how actively they are engaging with the content.

Excel Interview Questions

Excel is still a widely used tool by data analysts, and in interviews, Excel questions typically focus on advanced features. These questions might ask for definitions, or you may be required to perform some Excel tasks.

Data analysts should also have strong knowledge of data visualization. Data visualization interview questions typically focus on design and presenting data, and may be more behavioral in nature. Be prepared to talk about how you make data accessible on dashboards.

94. Explain the Excel VLOOKUP function. What are the limitations of VLOOKUP?

This function allows users to find data from one column, and return a corresponding value from another.

For example, if you were analyzing a spreadsheet of customer data, you might use VLOOKUP to find a customer name and the corresponding phone number.

One limitation of VLOOKUP is that it only looks to the right of the column you are analyzing. For example, you couldn't return a value from column A, if you used column B as the lookup column.

Another limitation is that VLOOKUP only returns the first value; if the spreadsheet contains duplicate records, you won't see any duplicates.

95. What is conditional formatting in Excel? When is a good time to use conditional formatting?

Conditional formatting allows users to change the appearance of a cell based on specified conditions.

Using conditional formatting, you can quickly highlight cells or ranges of cells, based on your conditions. Data analysts use conditional formatting to visualize data, identify patterns or trends, or detect potential issues.

96. What are your favorite data visualization tools?

Data analysts will be asked what tools they have experience with. Choose a few you're most comfortable with and explain the features you like.

97. What are some challenges you've experienced working with large volumes of data?

One tip: Think of questions like this in terms of Big Data's 5 Vs: volume, velocity, variety, veracity, and value.

98. Can you use multiple data formats in pivot tables?

Data can be imported from a variety of sources by selecting the Data tab and clicking Get External Data > From Other Sources. Excel worksheet data, data feeds, text files, and other such data formats can be imported, but you will need to create relationships between the imported tables and those in your worksheet before using them to create a pivot table.

99. When creating a visualization, you suspect data is missing. What do you do?

In your answer, provide an overview of your data validation process. For example, you might say, “The first step I would do would be to prepare a data validation report, which reveals why the data failed.” Then, you might talk through strategies for analyzing the dataset or techniques to process missing data, like deletion or mean/median/mode imputation.

Question	Topic	Difficulty	Ask Chance
Causal Inference Without A/B			
A/B Testing & Experimentation			
Medium			
Very High			
Department Expenses			
SQL			
Medium			
Very High			
Algorithm Reliability			
ML Ops & Training Pipelines			
Hard			
Very High			
This feature requires a user account			
Sign up to access this feature.			
feature			
Access 1000+ data science interview questions			

feature

30,000+ top company interview guides

feature

Unlimited code runs and submissions

Visualization Interview Questions

Data visualization involves presenting data in a graphical or pictorial format. This allows viewers to see data trends and patterns that may not be easy to understand in text-based data. Tableau, Power BI, and Python libraries such as Matplotlib and Seaborn are some of the most commonly used tools for data visualization.

100. Discuss your experience creating visualizations using Tableau, Power BI, or Python tools. What distinct features have you utilized in each?

This question requires you to detail your hands-on experience with the mentioned tools. It involves discussing specific features you have used in Tableau, Power BI, and Python, such as creating different types of charts, setting up dashboards, or using Python libraries like Matplotlib and Seaborn for custom visualizations.

101. What is DAX, and why is it important in Power BI?

DAX, or Data Analysis Expressions, is a library of functions and operators used to create formulas in Power BI, Analysis Services, and Power Pivot in Excel. These formulas, or expressions, are used to define custom calculations for tables and fields and to manipulate data within the model.

102. Imagine you're working on a sales report and have a table of daily sales data. You want to calculate the monthly sales total. How could you use DAX to do this?

This question tests your understanding of DAX time-intelligence functions. A suitable response could be:

"I would combine the SUM and CALCULATE functions and a Date table. First, I would create a measure using the SUM function to total the sales. Then, I would use the CALCULATE function and the DATESMTD (Dates Month to Date) function to calculate the monthly total. The DAX expression would look something like this:

Monthly Sales = CALCULATE(SUM(Sales[Daily Sales]), DATESMTD('Date'[Date]))“

103. Suppose a company has collected a large dataset on customer behavior, including demographics, transaction data, browsing history, and customer service interactions. You are tasked with presenting this data to the executive team, which does not comprise data professionals. How would you go about this?

This question assesses your ability to analyze complex datasets and create straightforward, impactful visualizations. Your response might include:

"Understanding the audience is key. For an executive summary, it's important to focus on high-level insights. I would start by performing exploratory data analysis to identify key trends and relationships within the data. From this, I could determine which aspects are most relevant to the executive team's interests and strategic goals.

For visualization, I would use a tool like Tableau or Power BI, which is known for its user-friendly, interactive dashboards. To make the data more digestible, I would utilize various chart types, such as bar graphs for categorical data comparison, line graphs for trend analysis, or pie charts for proportions.

To add an interactive element, I'd implement filters to allow executives to view data for different demographics, products, or time periods. Keeping the design clean and ensuring the visuals tell a clear story is crucial. For the presentation, I would walk them through the dashboard, explain key insights, and address any questions."

104. You are working for an e-commerce company that needs a real-time dashboard to monitor sales across various product categories. Would you use Tableau or Power BI for this task? How would you leverage the chosen tool's features to create the dashboard?

Your response should demonstrate your knowledge of Tableau and Power BI and ability to select the most appropriate tool for a specific task.

"For real-time sales monitoring, both Tableau and Power BI can be effective. However, if the company uses Microsoft's suite of products and requires extensive integration with these services, I would lean towards Power BI as it's part of the same ecosystem."

Power BI has robust real-time capabilities. I would leverage Power BI's DirectQuery feature to connect to the sales database, ensuring the data displayed on the dashboard is always up-to-date. The tool also allows for datasets that can be used to stream and update data continuously.

To visualize sales, I would design a dashboard that includes key metrics such as total sales, sales by product category, and changes in sales over time. I would also include slicers to allow users to filter data by region, time period, or other relevant dimensions.

Power BI also allows for creating alerts based on KPIs that could notify the team when a sales target is reached or when there are significant changes in sales trends."

Airbnb Business Analyst Interview Guide – Process, Questions & Tips

Next, you'll tackle a technical screen that combines a 30-minute HackerRank SQL assessment with a short case study or deck critique. You'll be challenged to write queries, interpret data, and solve real-world business problems—often referencing an Airbnb business analytics tool you've used or built. This stage tests your ability to extract insights from complex datasets, apply statistical reasoning, and communicate your findings clearly. Expect to demonstrate proficiency in SQL, Python, and data visualization, as well as your approach to A/B testing and business metric analysis. Your performance here is crucial, as it directly reflects your readiness to drive impact at scale.

SQL / Technical Questions

Expect to showcase your Airbnb advanced analytics chops with SQL questions that assess your ability to manipulate complex datasets, detect patterns, and deliver actionable insights for business decisions:

1. Find the total salary of slacking employees

To solve this, use an INNER JOIN to combine the employees and projects tables, filtering for employees who have at least one project assigned but no completed projects (End_dt IS NULL). Group by employee ID and use HAVING COUNT(p.End_dt) = 0 to identify slacking employees. Finally, sum their salaries using a subquery.

2. Write a query to get the average commute time for each commuter in New York

To solve this, use two subqueries: one to calculate the average commute time for each commuter in New York grouped by commuter_id, and another to calculate the overall average commute time across all commuters in New York. Use the TIMESTAMPDIFF function to calculate the duration of each ride in minutes, and then join the results to display both averages in the output.

3. Write a query to retrieve all user IDs whose transactions have exactly a 10-second gap

To solve this, use the LAG() and LEAD() window functions to calculate the time difference between consecutive transactions. Filter the results to include only those transactions with a 10-second gap and return the distinct user IDs in ascending order.

4. Find the average number of accepted friend requests for each age group

To solve this, use a RIGHT JOIN between the requests_accepted and age_groups tables to associate accepted friend requests with age groups. Calculate the average acceptance by dividing the count of accepted requests by the count of unique users in each age group, grouping by age_group, and ordering the results in descending order.

5. Cumulative Sales Since Last Restocking

To calculate cumulative sales since the last restocking, first identify the latest restocking date for each product using the MAX() function grouped by product_id. Then, use a window function SUM(...) OVER() to compute the running total of sales for each product after its last restocking date. Join the sales, products, and the derived table of last restocking dates, filtering sales that occurred after the last restocking date.

Case Study & Forecasting Questions

These questions test how you translate data into forward-looking strategy, often asking you to forecast metrics or build models that align with Airbnb, Inc. forecast and analysis efforts:

6. How would you build a dynamic pricing system for Airbnb based on demand and availability?

To build a dynamic pricing system, gather data on demand, availability, seasonality, and external factors like local events. Use machine learning models, such as regression or reinforcement learning, to predict optimal prices. Consider factors like user behavior, competitor pricing, and elasticity of demand while ensuring the system adapts to real-time changes.

7. How would you forecast revenue for the next year?

To forecast revenue for the next year, analyze historical revenue data for Facebook's various revenue streams, considering attributes like seasonality and trends. Depending on the behavior of each stream, use models such as classical time series forecasting, ARMA, or ARIMA to predict future revenue, and sum forecasts for all streams to estimate total revenue.

8. Given a task to predict hotel occupancy rates, how would you design the model?

To design the model, start by collecting relevant data such as historical occupancy rates, booking trends, seasonality, holidays, and external factors like local events or weather. Use this data to train a machine learning model, such as regression or time-series forecasting. Evaluate the model's performance using metrics like Mean Absolute Error (MAE) or Root Mean Square Error (RMSE) to ensure accuracy.

9. Call Center Resource Management

To address this problem, a predictive model such as a time-series forecasting model or a regression model can be used to predict call volumes and allocate agents accordingly. Metrics to evaluate the model include accuracy in predicting call volumes, agent utilization rates, and customer wait times. Over-allocation ensures better customer satisfaction but may increase

costs, while under-allocation risks longer wait times and lower customer satisfaction. Balancing these trade-offs is key to determining the optimal allocation strategy.

10. How would you build a function to return a list of daily forecasted revenue starting from Day 1 to the end of the quarter (Day N)?

To solve this, calculate the daily growth rate by dividing the difference between the total revenue target and Day 1 revenue by the number of days minus one. Then, iteratively compute the revenue for each day by adding the daily growth rate to the previous day's revenue, storing the results in a list.

Dashboard & Tooling Questions

You'll be asked to critique and design visualizations, evaluate metrics, and improve user understanding using tools like Minerva and Superset, often within the context of an Airbnb analytics dashboard or other Airbnb analysis software:

11. How would you visualize data with long tail text to effectively convey its characteristics and help extract actionable insights?

To visualize long-tail text data, start with frequency distribution plots like log-log scale histograms or Zipf plots to highlight keyword occurrences. Use semantic analysis techniques such as word clouds or clustering methods like t-SNE to uncover patterns. Integrate these insights with conversion metrics and temporal trends into dashboards for actionable business decisions.

12. Design a dashboard that provides personalized insights, sales forecasts, and inventory recommendations for shop owners

To design a merchant dashboard, prioritize metrics like sales trends (e.g., week-over-week revenue changes), inventory insights (e.g., days remaining for stock), and customer behavior (e.g., repeat purchase rates). Use adaptive visualizations tailored to merchant types, such as cohort charts for customer segmentation or smart banners highlighting actionable insights. Ensure scalability by leveraging metadata to personalize layouts and validate utility through merchant interaction tracking.

13. Interpreting Fraud Detection Trends

To interpret fraud detection graphs, focus on identifying anomalies, spikes, or patterns in fraudulent activities over time. Key insights include understanding the frequency, timing, and types of fraud, which can help refine detection algorithms and implement preventive measures. Use these insights to improve fraud detection processes by enhancing model accuracy, updating rules, and deploying real-time monitoring systems.

14. Critique a Minerva dashboard that shows booking trends by city. What would you improve?

Start by evaluating clarity and usefulness. Are city-level metrics aggregated at appropriate levels? Look for over-cluttered visualizations, inconsistent color schemes, or missing comparison baselines like year-over-year trends. Suggest adding filters for dates, guest demographics, and property types. Finally, recommend aligning KPI definitions across the Airbnb analytics dashboard so regional teams interpret the data consistently.

Behavioral & Values Questions

Here, interviewers are looking to understand how you collaborate, communicate, and live out Airbnb's values—especially in moments of ambiguity, cross-functional tension, or high-stakes decision-making:

15. How would you convey insights and the methods you use to a non-technical audience?

At Airbnb, business analysts frequently collaborate with design, operations, and regional teams that may not have technical expertise. You should describe how you structure your insights using storytelling, visualizations, and real-world analogies. For example, you might walk through how you used clustering to segment guests by travel behavior, then presented clear visuals and trade-offs to help a marketing team prioritize a campaign. Focus on simplifying without oversimplifying.

16. What do you tell an interviewer when they ask you what your strengths and weaknesses are?

Tailor your answer to qualities that align with Airbnb's values. For strengths, consider areas like strong SQL proficiency, experience with experimentation platforms, or a track record of turning ambiguous problems into structured analysis. For weaknesses, avoid clichés. Instead, be honest and share how you've addressed it. For example, you might say you previously over-indexed on perfection in dashboards but learned that speed-to-insight is more critical in fast-paced teams like Airbnb's regional operations.

17. Why Do You Want to Work With Us

Go beyond generic admiration. In 2025, Airbnb continues to lead in the intersection of travel, trust, and technology. Reference your alignment with Airbnb's mission and how you're excited to contribute to key initiatives, such as sustainable travel growth or optimizing the host onboarding experience. Mention interest in working with global teams, using real-time data to drive product or marketplace strategy, and leveraging Airbnb's strong analytics infrastructure to make meaningful impact.

18. Talk about a time when you had trouble communicating with stakeholders. How were you able to overcome it?

Airbnb analysts often sit between product, operations, and regional leadership. When there is disagreement or confusion, it is your role to bridge data and narrative. Provide a story where you had to identify why your insight wasn't resonating, then explain how you reframed the data or involved the stakeholders earlier in the analysis process. Focus on active listening, using visual tools, or scenario modeling to create alignment and improve decision-making.

How to Prepare for a Business Analyst Role at Airbnb

To excel in the Airbnb business analyst interview, you'll want to master both technical and strategic preparation. Start by honing your SQL and Tableau skills, as you'll be expected to analyze complex datasets and visualize insights using an airbnb business analytics tool. Practice writing advanced queries that extract actionable trends from booking, pricing, and review data—these are core to the technical screens and on-site exercises. Airbnb's interviewers value candidates who can not only manipulate data but also translate findings into business recommendations using clear, compelling dashboards.

Building your own forecasting projects is a powerful way to stand out. Leverage the open-source Airbnb review dataset, which contains millions of real guest reviews and listing

details from global cities. Use this dataset to practice exploratory data analysis, sentiment mining, and predictive modeling—skills that directly mirror the challenges you'll face at Airbnb. For example, you might forecast review scores or booking trends using Python and Tableau, then present your findings as you would to a cross-functional team. Practice with our AI Interviewer to gain more clarity on the approach to the ideal answers.

Equally important is your ability to tell stories with data and align with Airbnb's core values. Use the STAR method (Situation, Task, Action, Result) to structure your behavioral answers, and be ready to demonstrate how you "Champion the Mission" by connecting your work to Airbnb's vision of belonging and community impact.

Finally, simulate the interview environment with mock interviews and peer feedback. Benchmark your SQL speed and accuracy on data-driven platforms, and seek out realistic case studies to sharpen your business sense. This holistic, hands-on approach will help you approach the process with confidence and clarity, ready to make a measurable impact from day one.

SQL / Technical Questions

Expect to showcase your Airbnb advanced analytics chops with SQL questions that assess your ability to manipulate complex datasets, detect patterns, and deliver actionable insights for business decisions:

1. Find the total salary of slacking employees

To solve this, use an INNER JOIN to combine the employees and projects tables, filtering for employees who have at least one project assigned but no completed projects (End_dt IS NULL). Group by employee ID and use HAVING COUNT(p.End_dt) = 0 to identify slacking employees. Finally, sum their salaries using a subquery.

2. Write a query to get the average commute time for each commuter in New York

To solve this, use two subqueries: one to calculate the average commute time for each commuter in New York grouped by commuter_id, and another to calculate the overall average commute time across all commuters in New York. Use the TIMESTAMPDIFF function to

calculate the duration of each ride in minutes, and then join the results to display both averages in the output.

3. Write a query to retrieve all user IDs whose transactions have exactly a 10-second gap

To solve this, use the LAG() and LEAD() window functions to calculate the time difference between consecutive transactions. Filter the results to include only those transactions with a 10-second gap and return the distinct user IDs in ascending order.

4. Find the average number of accepted friend requests for each age group

To solve this, use a RIGHT JOIN between the requests_accepted and age_groups tables to associate accepted friend requests with age groups. Calculate the average acceptance by dividing the count of accepted requests by the count of unique users in each age group, grouping by age_group, and ordering the results in descending order.

5. Cumulative Sales Since Last Restocking

To calculate cumulative sales since the last restocking, first identify the latest restocking date for each product using the MAX() function grouped by product_id. Then, use a window function SUM(...) OVER() to compute the running total of sales for each product after its last restocking date. Join the sales, products, and the derived table of last restocking dates, filtering sales that occurred after the last restocking date.

Case Study & Forecasting Questions

These questions test how you translate data into forward-looking strategy, often asking you to forecast metrics or build models that align with Airbnb, Inc. forecast and analysis efforts:

6. How would you build a dynamic pricing system for Airbnb based on demand and availability?

To build a dynamic pricing system, gather data on demand, availability, seasonality, and external factors like local events. Use machine learning models, such as regression or

reinforcement learning, to predict optimal prices. Consider factors like user behavior, competitor pricing, and elasticity of demand while ensuring the system adapts to real-time changes.

7. How would you forecast revenue for the next year?

To forecast revenue for the next year, analyze historical revenue data for Facebook's various revenue streams, considering attributes like seasonality and trends. Depending on the behavior of each stream, use models such as classical time series forecasting, ARMA, or ARIMA to predict future revenue, and sum forecasts for all streams to estimate total revenue.

8. Given a task to predict hotel occupancy rates, how would you design the model?

To design the model, start by collecting relevant data such as historical occupancy rates, booking trends, seasonality, holidays, and external factors like local events or weather. Use this data to train a machine learning model, such as regression or time-series forecasting. Evaluate the model's performance using metrics like Mean Absolute Error (MAE) or Root Mean Square Error (RMSE) to ensure accuracy.

9. Call Center Resource Management

To address this problem, a predictive model such as a time-series forecasting model or a regression model can be used to predict call volumes and allocate agents accordingly. Metrics to evaluate the model include accuracy in predicting call volumes, agent utilization rates, and customer wait times. Over-allocation ensures better customer satisfaction but may increase costs, while under-allocation risks longer wait times and lower customer satisfaction. Balancing these trade-offs is key to determining the optimal allocation strategy.

10. How would you build a function to return a list of daily forecasted revenue starting from Day 1 to the end of the quarter (Day N)?

To solve this, calculate the daily growth rate by dividing the difference between the total revenue target and Day 1 revenue by the number of days minus one. Then, iteratively compute

the revenue for each day by adding the daily growth rate to the previous day's revenue, storing the results in a list.

Dashboard & Tooling Questions

You'll be asked to critique and design visualizations, evaluate metrics, and improve user understanding using tools like Minerva and Superset, often within the context of an Airbnb analytics dashboard or other Airbnb analysis software:

11. How would you visualize data with long tail text to effectively convey its characteristics and help extract actionable insights?

To visualize long-tail text data, start with frequency distribution plots like log-log scale histograms or Zipf plots to highlight keyword occurrences. Use semantic analysis techniques such as word clouds or clustering methods like t-SNE to uncover patterns. Integrate these insights with conversion metrics and temporal trends into dashboards for actionable business decisions.

12. Design a dashboard that provides personalized insights, sales forecasts, and inventory recommendations for shop owners

To design a merchant dashboard, prioritize metrics like sales trends (e.g., week-over-week revenue changes), inventory insights (e.g., days remaining for stock), and customer behavior (e.g., repeat purchase rates). Use adaptive visualizations tailored to merchant types, such as cohort charts for customer segmentation or smart banners highlighting actionable insights. Ensure scalability by leveraging metadata to personalize layouts and validate utility through merchant interaction tracking.

13. Interpreting Fraud Detection Trends

To interpret fraud detection graphs, focus on identifying anomalies, spikes, or patterns in fraudulent activities over time. Key insights include understanding the frequency, timing, and types of fraud, which can help refine detection algorithms and implement preventive measures. Use these insights to improve fraud detection processes by enhancing model accuracy, updating rules, and deploying real-time monitoring systems.

14. Critique a Minerva dashboard that shows booking trends by city. What would you improve?

Start by evaluating clarity and usefulness. Are city-level metrics aggregated at appropriate levels? Look for over-cluttered visualizations, inconsistent color schemes, or missing comparison baselines like year-over-year trends. Suggest adding filters for dates, guest demographics, and property types. Finally, recommend aligning KPI definitions across the Airbnb analytics dashboard so regional teams interpret the data consistently.

Behavioral & Values Questions

Here, interviewers are looking to understand how you collaborate, communicate, and live out Airbnb's values—especially in moments of ambiguity, cross-functional tension, or high-stakes decision-making:

15. How would you convey insights and the methods you use to a non-technical audience?

At Airbnb, business analysts frequently collaborate with design, operations, and regional teams that may not have technical expertise. You should describe how you structure your insights using storytelling, visualizations, and real-world analogies. For example, you might walk through how you used clustering to segment guests by travel behavior, then presented clear visuals and trade-offs to help a marketing team prioritize a campaign. Focus on simplifying without oversimplifying.

16. What do you tell an interviewer when they ask you what your strengths and weaknesses are?

Tailor your answer to qualities that align with Airbnb's values. For strengths, consider areas like strong SQL proficiency, experience with experimentation platforms, or a track record of turning ambiguous problems into structured analysis. For weaknesses, avoid clichés. Instead, be honest and share how you've addressed it. For example, you might say you previously over-indexed on perfection in dashboards but learned that speed-to-insight is more critical in fast-paced teams like Airbnb's regional operations.

17. Why Do You Want to Work With Us

Go beyond generic admiration. In 2025, Airbnb continues to lead in the intersection of travel, trust, and technology. Reference your alignment with Airbnb's mission and how you're excited to contribute to key initiatives, such as sustainable travel growth or optimizing the host onboarding experience. Mention interest in working with global teams, using real-time data to drive product or marketplace strategy, and leveraging Airbnb's strong analytics infrastructure to make meaningful impact.

18. Talk about a time when you had trouble communicating with stakeholders. How were you able to overcome it?

Airbnb analysts often sit between product, operations, and regional leadership. When there is disagreement or confusion, it is your role to bridge data and narrative. Provide a story where you had to identify why your insight wasn't resonating, then explain how you reframed the data or involved the stakeholders earlier in the analysis process. Focus on active listening, using visual tools, or scenario modeling to create alignment and improve decision-making.