EMMA ding

emmading.com
info@emmading.com

Data-Driven Interview Preparation for Data Scientists

Challenges for Data Scientists

Preparing for data science interviews is hard. There are three specific challenges/questions that many data scientists face:

- **There are too many things to prepare.**
  - How do you know where to start?
- **Each one takes a lot of time to become proficient in.**
  - How in-depth should your studying be?
- **I've studied, but I still don't feel confident enough.**
  - How do I know when I'm ready?

If we could answer these three questions, preparing for interviews would be a lot easier. So, how do we do that?

Why the Data-Driven Approach?

Here are the three main benefits of using this approach:

▼ **You spend time on what really matters.**

- Most people interview with more than 3 companies to land an offer. As you do more interviews, you'll realize that fundamental concepts appear frequently in interviews, and many companies ask similar interview questions.
- Instead of starting from scratch with every interview, the data-driven approach has you studying those **frequently asked concepts.** You are spending time on things that will appear in multiple interviews.

▼ **You don't have to know everything.**

- Candidates who get offers may not have gotten all the questions correct. What people who get offers do have is **a strong mastery of the fundamentals.**
- The data-driven approach assures you that you don't have to know every little thing. By studying the most frequently asked questions, you learn what you need to know for interviews and don't waste time on trying to learn absolutely everything.

▼ **You have flexibility.**

- You don't have to have a certain amount of time. The data-driven approach works with any time limit.
- If you have limited time, you can focus on just the most frequently asked questions to ensure you have the broadest coverage of interview topics in the least amount of time.
- If you have more time, you can study and prepare for less frequently asked questions to increase your chances of success.

Goals

There are two things that would make interview prep a lot easier:

1. Know **exactly what you need to prepare** and **the level of depth required.**
2. **Be confident** that you will ace the interview. Showing up with confidence increases your chance of success.

⚖️ So now the question is, is there an interview preparation method that can help us achieve these things?

The Data-Driven Approach

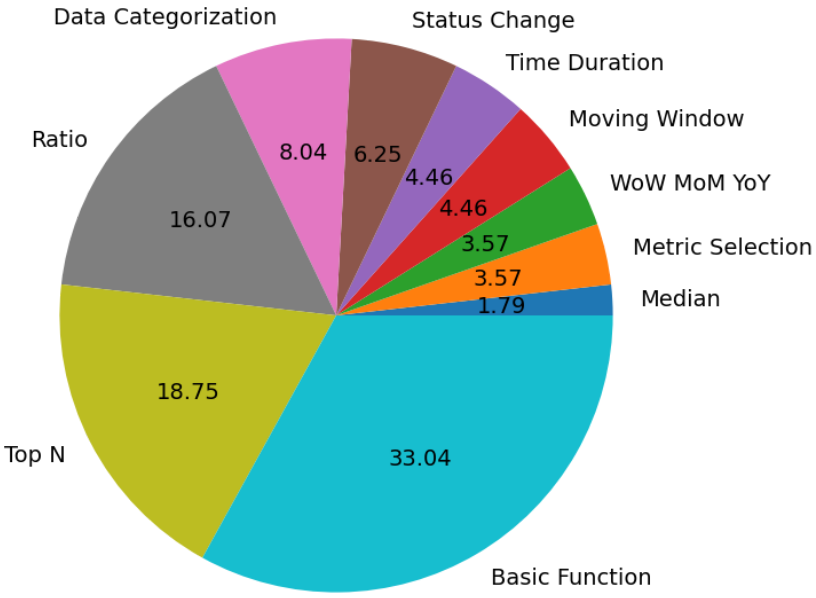
The data-driven approach is a system that helps you study systematically and cover the necessary material at the appropriate depth while using the **least amount of time** possible.

▼ **How Does It Work?**

The data-driven approach works by using data to structure your preparation. The key is to focus on **actual interview questions**. By analyzing real interview questions and categorizing the most frequently asked topics, you can focus your time and energy on the most important aspects.

▼ **Example**

Let’s look at prepping for SQL interviews as an example. After analyzing 150+ real SQL interview questions from 30+ companies, we have the following breakdown of questions by category.



The data-driven method works by **focusing on the low-hanging fruit first.**

That means for SQL we would study categories in the following order:

1. Basic Function
2. Top N
3. Ratio
4. Data Categorization
5. Status Change
6. Time Duration
7. Moving Window
8. WoW, MoM, YoY
9. Metric Selection
10. Median

As you can tell from the pie chart, after studying just the first three categories (Basic Functions, Top N, and Ratio) you would have achieved **over 65% coverage of SQL questions!**

Using this method means that you can achieve the most coverage in the least amount of time. It ensures that you are practicing what is most likely to be on the interview and not just studying randomly.

Distribution of Interview Questions

Of course, collecting and analyzing interview questions is time-consuming, but don't worry!

Below you will find my breakdown of real interview questions for SQL, statistics, machine learning, and behavioral questions!

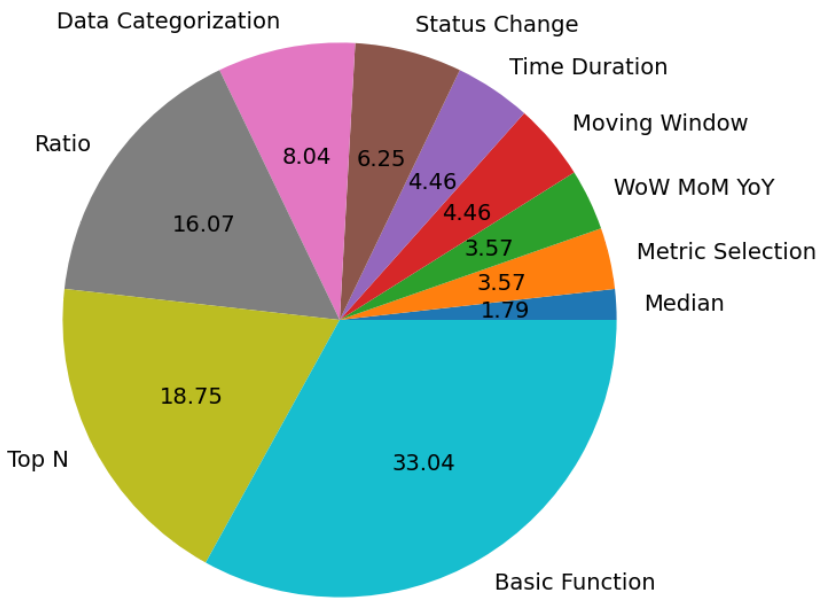
> 👇 These charts can get you well on your way to effective and efficient interview preparation.

SQL

Distribution of SQL interview questions by analyzing 150+ real SQL interview questions from 30+ companies.

Top 3 categories:

- **Basic Function**: These problems focus on fundamental SQL functions like aggregations, filtering, and sorting to manipulate and retrieve data from a database.
- **Top N**: Top N SQL problems involve identifying and extracting a specific number of records with the highest or lowest values based on a particular column or set of columns.
- **Ratio Problems**: Ratio SQL problems require calculating and analyzing proportions or ratios between different values in a database, often involving relationships between multiple columns or tables.

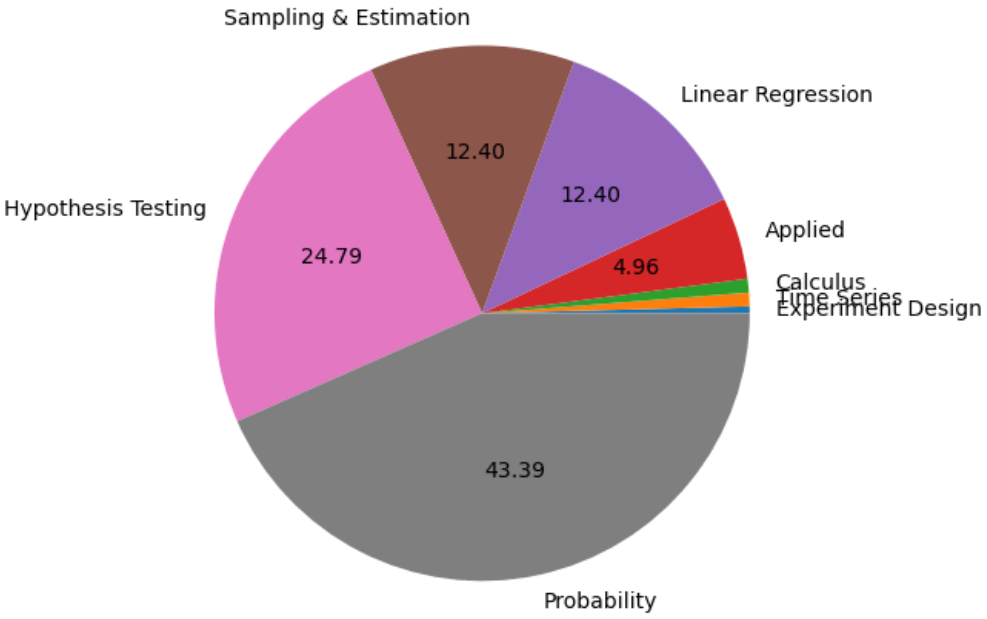


Statistics

Distribution of statistics questions from 90+ real statistics interview questions from 30+ companies.

Top 3 categories:

- **Probability**: Probability statistics problems involve analyzing the likelihood of events occurring and calculating probabilities based on fundamental principles of probability theory.
- **Hypothesis Testing**: Hypothesis testing statistics problems require making decisions about population parameters based on sample data and determining whether the observed results are statistically significant.
- **Sampling & Estimation**: Sampling and estimation statistics problems focus on obtaining information about a population by taking representative samples and making inferences about population parameters using statistical techniques.



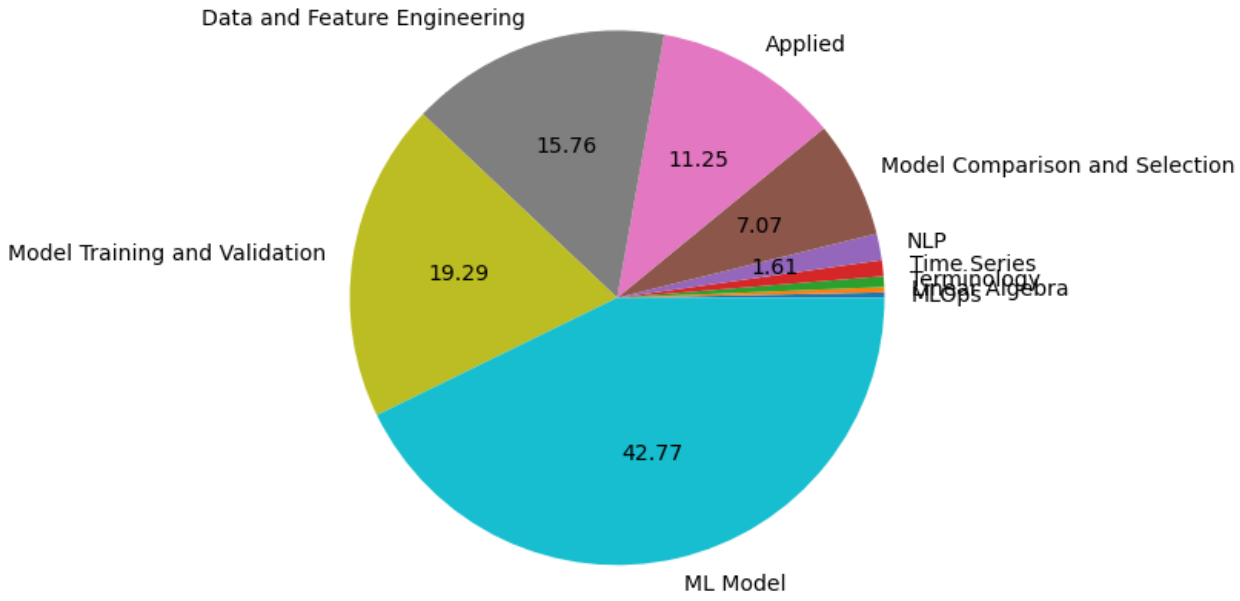
Machine Learning

Analysis based on 190+ ML interview questions from 50+ companies.

Top 3 categories:

- **ML Models**: ML models refer to questions related to details of machine learning models, such as decision trees, neural networks, or support vector machines.
- **Model Training and Validation**: Model training and validation involve the process of optimizing and assessing the performance of ML models by iteratively fitting the model to training data and evaluating its performance on validation or test data to ensure generalization and avoid overfitting.

- **Data and Feature Engineering**: Data and feature engineering involve the process of selecting, transforming, and creating meaningful features from raw data to improve the performance and interpretability of ML models, often including tasks like data cleaning, normalization, dimensionality reduction, and creating new features based on domain knowledge.

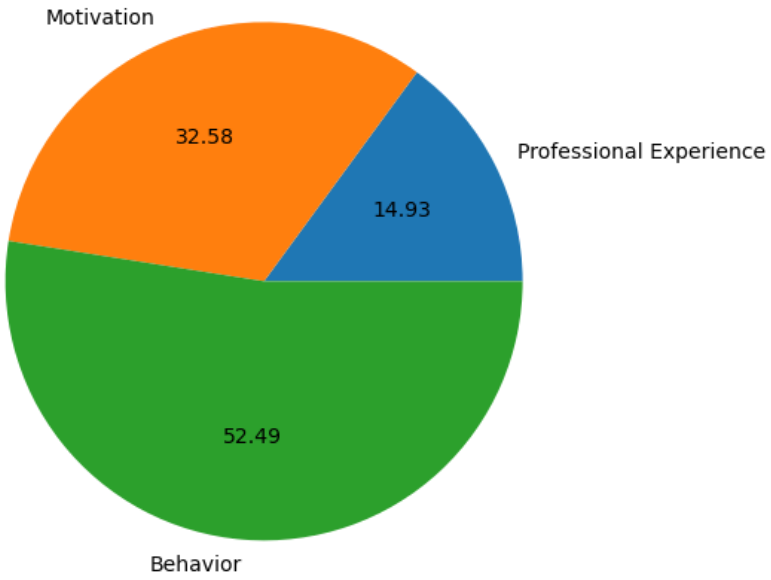


Behavioral Questions

Analysis based on 120+ behavioral interview questions from 30+ companies.

Top 3 categories:

- **Behavior**: Behavioral interview problems focus on assessing your past behaviors and actions in specific situations to evaluate your skills, abilities, and how you handle various challenges.
- **Motivation**: Motivation-based interview problems aim to understand your drive, aspirations, and the factors that inspire you professionally, enabling the interviewer to gauge your alignment with the role and organization.
- **Professional Experience**: Professional experience-based interview problems delve into your previous work accomplishments, projects, and responsibilities to gain insights into your skills, expertise, and how well they align with the requirements of the position.


Motivation
32.58
Professional Experience
14.93
52.49
Behavior