100 Key Concepts to Know for Data Science Interviews 🚀

Last Updated: April 04, 2022
Author: Dan Lee (Dan@DataInterview.com) | LinkedIn

Message to Candidates:

Hello Candidate 👋 ,

Dan here 👨‍🏫! I am a former **PayPal** and **Google** data scientist, who runs an interview prep platform called datainterview.com for candidates seeking BIE, DA, DS, DE and MLE roles.

When preparing for data science interviews, it's often confusing what topics to study given the breadth of topics one must master to become a skilled data scientist.

But, regardless of whether role, level, domain, or company you are preparing interviews for, there is still a basic set of concepts you must master.

Based on my personal database of 1,000+ interview questions pooled across small to large tech companies, domains from general consulting to product, and entry to senior levels, I've compiled this list of top **100 fundamental topics** you need to study to prepare for **ANY data science and MLE interviews** 📚

In the near future, I will add more topics in this document.

Happy interviewing! Feel free to email me at dan@datainterview.com if you have any questions!

Share the ❤️ by sharing this doc :)

– Dan

100 Key Concepts to Know for Data Science Interviews 🚀

Statistics	Machine Learning	Coding (R, Python)	SQL
Basic Statistical Concepts 1. Univariate statistics - mean, median, mode 2. Standard deviation and variance 3. Covariance and correlation 4. Population and sample 5. Nominal, ordinal and continuous, discrete data types 6. Outlines 7. The Simpson's Paradox 8. Selection Bias **Hypothesis Testing** 9. Hypothesis Statements 10. Z-Test 11. T-Test 12. T-Test for sample means 13. T-Test for proportions 14. Paired and unpaired T-Tests 15. Variance test 16. ANOVA 17. Chi-Squared test	**ML Foundations** 46. Variance & Bias trade-off 47. Predictability vs interpretability trade-off 48. Feature selection 49. Feature engineering 50. Model validation 51. Curse of Dimensionality 52. Data Leakage 53. Classification problem 54. Regression problem **ML Algos** 55. Regularized Regression 56. Decision Tree 57. Random Forest 58. XGBoost 59. Bagging vs Boosting 60. Variable importance from tree models 61. Principal Component Analysis 62. Hyperparameter Tuning 63. K-Means Clustering 64. Hierarchical Clustering	**Algorithms & Data Structures** 74. Array problems 75. Math problems 76. String problems 77. Matrix problems 78. Palindrome 79. Rotate a matrix by 180 degrees **Data Manipulation** 80. Calculate using statistical functions 81. Aggregation 82. Lags 83. Group By 84. Filtering 85. JOINs 86. Sorting	**Clauses** 87. SELECT 88. DISTINCT 89. ORDER BY 90. ALIAS 91. WHERE 92. NULL 93. Aggregations (SUM, MIN, MAX, COUNT, AVG) 94. HAVING 95. JOINs 96. SETs 97. Subqueries 98. PARTITION BY 99. RANK 100. CASE

18. Goodness of Fit test for categorical data
19. Nominal, ordinal and continuous, discrete data types
20. Pairwise tests
21. T-Test assumptions
22. Non-parametric tests
23. Type 1 & 2 Errors

Probability & Distributions

24. The Bayes Theorem
25. Conditional probability
26. Normal distribution
27. Uniform distribution
28. Bernoulli distribution
29. Binomial distribution
30. Geometric distribution
31. Poisson distribution
32. Exponential distribution
33. Deriving the mean and variance of distributions
34. Central Limit Theorem
35. The Birthday problem
36. Card probability problems
37. Die roll problems

Regression Modeling

38. OLS regression

65. The Elbow Technique
66. Neural Networks
67. Cross-Validation
68. AUC vs Accuracy
69. Imbalanced class problem

Productionization

70. Model checks
71. Model productionization steps
72. Online model evaluation
73. Designing scalable model systems

39. Confidence vs prediction intervals 40. Logistic regression 41. Regression model assumptions 42. Model diagnostic checks 43. R-Square vs R-Square Adjusted 44. AIC, BIC, CP Statistics 45. Model Interpretation			