Overall Data Science
Review Topic List

This list is a high-level overview of all things related to data science that are important to have an understanding for in order to (1) feel confident in data science skills, (2) be successful in data science job hunting & (3) be successful in a data science job. Not every job will make use of every one of these concepts, but all have potential to be so, or be tested during an interview process.

Legend

Abbreviation	Description	Notes
FK (for topic X, Y, Z)	Foundational Knowledge - topic/skill that provides foundations for understanding more complex topics or carrying out other responsibilities as a data scientist	Parentheticals describe which more complex topics the FK is required for
DSIK	Data Science Interview (Knowledge) - commonly tested topic during conversational portions of data science interviews	Topics can be addressed in several formats: - Q & A (conversational "quizzing") - Whiteboarding (pseudoalgorithm)
DSICC	Data Science Interview (Code Challenge) - commonly tested topic addressed in code challenges	Code challenges can come in the forms of: - Take home - Live coding (on laptop, whiteboard, etc.) - live coding usually involves less complex challenges (basic SQL, python coding, etc.)
NTH	Nice To Have - Not absolutely required, but will be helpful to know, more likely to get a job	Can help data scientists stand out during interview process, and will help facilitate more in-depth learning & greater complexity of development
OPN	Optional - Completely optional	Can help data scientists stand out during interview process

Topic	Subtopics	Overall Priority	Relevance (legend)	Notes
Coding	Python - Basics - Data structures - OOP - Libraries	Very High	- FK (for most of DS implementation) - DSICC - DSIK (sometimes)	- Sometimes topics can be tested like OOP, BigO, multi threading, etc.
	SQL - Order of operations - Joins - MySQL - CloudSQL	Very High	- FK (for data wrangling & management) - DSICC - DSIK	- Sometimes topics can be tested verbally like SQL order of operations, join vs union, etc.
	R	Low	- OPN (assuming python)	- Becoming optional in DS
	Bash / command line	Medium	- NTH	- 'Required' for git use for the most part - Especially useful for interacting with cloud technology
	Git - Push - Pull - Rebase - Forking	High	- FK (for collaborative work streams) - NTH	- Not often tested during interviews, but may be required to download data for code challenge - Will likely be used in day-to-da work in DS job

Topic	Subtopics	Overall Priority	Relevance (legend)	Notes
Statistics	Basic estimates - Mean - Median - Mode - Percentiles - Boxplots - Variability vs Standard deviation	High	- FK (for cohort analytics) - DSIK - DSICC	Can be tested both on a knowledge basis, or expected during code challenge
	Probability - Calculating probability with different scenarios - Probability & cumulative density functions, kernels	High	- FK (for segmentation, clustering, hyperparameter optimization, a/b testing) - DSIKCC	
	Hypothesis testing - H0, Ha - Meaning & interpretation of p-value - Significant testing (t-test, z-test, chi-square) - Confidence intervals - Bootstrapping / resampling - Sample sizes / power analysis - Effect size - A/B testing - Multi arm a bandit	Very High	- FK (for A/B testing, academic research with health data, marketing strategy decision making, for hyperparameter optimization) - DSIK - DSICC	
	Concepts / rules - Types of bias - Central limit theorem - Sample vs Population estimates - Bootstrapping - Permutation test	Very High	- FK (for building models) - DSIK	

Topic	Subtopics	Overall Priority	Relevance (legend)	Notes
	Distributions - Normal / Gaussian - Binomial - Poison - Exponential - Plotting - Using for: A/B testing, assumption-testing	Very High	- FK (for data preparation & determining validity of prediction results) - DSIK	Need to know assumptions / description of each, what they're used for, etc.
Data Processing	- Importing CSV, JSON, txt, etc - Categorical & dummy variables - Time stamps & date	Very High	- FK (for most of data science implementations) - DSICC	- Many different libraries & standards for time stamps and they interact differently with various plotting libraries
Data Exploration	Topics - Examining data (nulls, completeness, etc.) - Imputing data - Data point density	Very High	- FK (for examining feature relationships, discovering bias, time series data that is incomplete) - DSICC	- Imputing data mostly relates to time series data
	Python coding - Matplotlib - Seaborn - Plotly	Very High	- FK - DSICC	
	Approaches - Feature comparison by plotting - Regression - Scatterplots - Histograms - Sub-segmentation	Very High	- FK - DSICC	

Topic	Subtopics	Overall Priority	Relevance (legend)	Notes
Modeling	Concepts - Simplicity vs need for complexity - Bias vs variance - Supervised vs unsupervised vs reinforcement vs semi supervised - How to select model type	Very High	- FK (for determining MVP = minimum viable product, resource requirements with DevOps) - DSIK	- How bias vs variance trade off affects model parameter tuning & structure selection - Over concepts will be HEAVILY tested in knowledge-based interview stages
(Supervised)	Regression - Assumptions - Least squares linear regression - Understanding of multiple linear regression - Dummy variables - Testing / scoring - Regularization (L1 & L2) - Polynomial / Spline, etc.	Very High	- FK (feature engineering & dimensionality reduction) - DSICC - DSIK - Some NTH	- WIll be tested heavily in interviews both conceptually (bias/variance, scoring methods, assumptions) & coding - Advanced approaches (polynomial / splines are NTH)
(Supervised - Regression)	Evaluation & scoring - Residuals (error): absolute error, squared error, MSE, RMSE, A1C, B1C - Parameter selection (Regularization)	Very High	- FK (for model selection & hyperparameter optimization) - DSIK	- WIll be tested heavily in interviews both conceptually
(Supervised)	Classification - Class balancing - Naive Bayes (using probability) - Logistic regression - Gradient Boosted Classifier - Odds ratio	Very High	- FK (feature engineering, clustering) - DSIK - DSICC	- Foundational knowledge that is very important, and also commonly required for implementation in take home assignments or code challenges

Topic	Subtopics	Overall Priority	Relevance (legend)	Notes
(Supervised - classification)	Classification evaluation - Confusion matrix - Precision - Recall (sensitivity) - Specificity - AUC - Lift	Very High	- FK - DSIK - DSICC	- Extremely high priority for DS interview - knowledge testing!! Especially Precision, recall, specificity, AUC
(Supervised)	Decision Trees - Single vs Ensemble - Purity (Gini, entropy, etc.) - Information Gain (split decision) - Parameters (leaf depth, randomization, etc.) - Bagging - Random Forest - Boosting	Very High	- FK (feature engineer, clustering) - DSIK - DSICC	- Foundational knowledge that is very important, and also commonly required for implementation in take home assignments or code challenges
(Unsupervised)	Clustering - Standardization / PCA beforehand - Distance metrics - N-cluster selection - KNN - Kmeans - SVD - Use cases, advantages, disadvantages - Hierarchical	Very High	- FK (for recommendation engines, feature engineer) - DSIK	- Foundational knowledge that is very important, and also commonly required for implementation in take home assignments or code challenges

Topic	Subtopics	Overall Priority	Relevance (legend)	Notes
(Supervised/ Semisupervised)	Neural Networks - Forward propagation - Backward propagation - Hidden layer - Activation functions (tanh, sigmoid, ReLu) - Types (CNN, RNN, LSTM, Autoencoder, transformers, GAN, +++) - Tensorfow / Keras	High - especially for MLE	- FK - DSIK	- Usually tested conceptually. For MLE interviews, you will definitely need to be intimately familiar with deep learning at large, including these topics, know how to code, deep dive into parameter discussion, etc.
Model Selection	Model Comparison - Simplicity vs complexity - Bias vs Variance - Data completeness & availability - Goal - Reproducibility - Personalization - Model Versioning	High	- FK (for selecting MVP, for reproducibility improvement, for versioning) - DSIK	- Concepts tested to help interviewers know what your approach to problem solving with DS is
	Model outcome interpretation - Over/under-fitting - Train vs validation vs Test scores - Feature importance - Hyperparameter optimization	Very High	- FK - DSIK	- WIll be tested heavily (concepts) during conversational parts of DS interviews
Text data	Natural Language Processing (NLP) - tf-idf - Stop words - Tokens	High/medium	- FK - DSIK - NTH	- May be asked about familiarity, especially if it’s relevant to the job

Topic	Subtopics	Overall Priority	Relevance (legend)	Notes
Data Tooling	Big Data Tools - Cloud platforms (AWS, GCP, Azure) - Spark - HDFS (old, but still asked about) - Dask - Multi-threading (python)	Medium	- FK - NTH	- May be asked about familiarity, especially if it's relevant to the job
	Spark - Resilient Distributed Dataset - Lazy evaluation - MapReduce - Spark SQL - PySpark - Spark Mlib	Medium	- FK - NTH	- Differences between Spark & HDFS/MapReduce are common test questions
	Databases - Centralized vs distributed - On-prem vs Cloud - Relational vs Non-relational (NoSQL, document-based, etc.)	High	- NTH - DSIK	- May be asked about familiarity, especially if it's relevant to the job
	Cloud Tools - Databases - Virtual machines - Service Accounts - Ingestion pipelines (e.g. AWS Kinesis/firehose, GCP Dataflow)	Medium/High	- NTH	- May be asked about familiarity, especially if it's relevant to the job
	Other Tools - API / django/ flask - Pub/sub (or similar) - Tensorflow / Keras	Low	- NTH - OPN	- May be asked about familiarity, especially if it's relevant to the job