ToolTechSavvy

Data Science Mastery Roadmap

Complete Learning Checklist

Phase 1: Mathematical Foundations

Statistics & Probability

■ Descriptive statistics (mean, median, mode, variance, standard deviation)
■ Probability theory fundamentals
■ Probability distributions (Normal, Binomial, Poisson, Exponential)
■ Conditional probability and Bayes' theorem
■ Random variables and expected values
■ Central Limit Theorem
■ Law of Large Numbers
■ Sampling methods and techniques
■ Population vs sample statistics
■ Confidence intervals
■ Hypothesis testing (t-tests, z-tests, chi-square)
■ P-values and statistical significance
■ Type I and Type II errors
■ A/B testing fundamentals
■ Statistical power analysis
■ ANOVA (Analysis of Variance)
■ Correlation vs causation
■ Covariance and correlation coefficients

Linear Algebra

■ Vectors and vector operations
■ Matrices and matrix operations
■ Matrix multiplication and transpose

- Identity and inverse matrices
- Determinants
- Eigenvalues and eigenvectors
- Singular Value Decomposition (SVD)
- Matrix factorization
- Vector spaces and linear independence
- Dot products and cross products

Calculus

- Derivatives and differentiation rules
- Partial derivatives
- Chain rule
- Gradient and gradient descent
- Integration basics
- Multivariable calculus
- Optimization techniques
- Local and global minima/maxima

Phase 2: Programming Fundamentals

Python Programming

- Python syntax and basic data types
- Variables and operators
- Control flow (if, elif, else)
- Loops (for, while)
- Functions and lambda expressions
- List comprehensions
- Dictionaries and sets
- Tuples and sequences
- String manipulation
- File I/O operations
- Exception handling (try, except, finally)
- Object-oriented programming (classes, objects)
- Inheritance and polymorphism
- Modules and packages
- Virtual environments
- PIP package management
- Working with JSON and CSV files
- Regular expressions (regex)

Data Structures & Algorithms

- Lists and arrays
- Linked lists
- Stacks and queues
- Hash tables and dictionaries
- Trees and binary trees
- Graphs and graph traversal
- Sorting algorithms (bubble, merge, quick)
- Search algorithms (binary search, linear search)
- Time complexity (Big O notation)
- Space complexity analysis

- Recursion
- Dynamic programming basics

SQL & Databases

- Database fundamentals and RDBMS concepts
- SQL syntax and basic queries
- SELECT statements and filtering (WHERE)
- Sorting and ordering (ORDER BY)
- Aggregate functions (COUNT, SUM, AVG, MIN, MAX)
- GROUP BY and HAVING clauses
- JOIN operations (INNER, LEFT, RIGHT, FULL)
- Subqueries and nested queries
- UNION and set operations
- Creating and modifying tables (DDL)
- Inserting, updating, and deleting data (DML)
- Indexes and query optimization
- Views and stored procedures
- Window functions
- Common Table Expressions (CTEs)
- Database normalization
- Working with MySQL, PostgreSQL, or SQLite
- NoSQL basics (MongoDB, Cassandra)

Phase 3: Data Manipulation & Analysis

NumPy

- NumPy arrays and array creation
- Array indexing and slicing
- Array reshaping and transposing
- Broadcasting rules
- Mathematical operations on arrays
- Statistical functions (mean, std, var)
- Linear algebra operations
- Random number generation
- Saving and loading arrays
- Array manipulation (concatenate, split, stack)

Pandas

- Series and DataFrames
- Reading data (CSV, Excel, JSON, SQL)
- Writing data to various formats
- Data inspection (head, tail, info, describe)
- Selecting and filtering data
- Boolean indexing
- loc and iloc indexing
- Handling missing data (dropna, fillna)
- Data type conversion (astype)
- Renaming columns and indexes
- Sorting data
- GroupBy operations and aggregation
- Pivot tables and cross-tabulation
- Merging and joining DataFrames
- Concatenating DataFrames
- Reshaping data (melt, stack, unstack)
- Applying functions (apply, map, applymap)
- String methods for text data

- DateTime operations and time series
- Handling duplicates
- Data sampling and resampling

Phase 4: Data Visualization

Matplotlib

- Basic plotting concepts
- Line plots
- Scatter plots
- Bar charts and histograms
- Pie charts
- Box plots and violin plots
- Subplots and figure management
- Customizing plots (colors, styles, markers)
- Adding labels, titles, and legends
- Axis formatting and scaling
- Saving figures
- Advanced plot customization

Seaborn

- Seaborn themes and color palettes
- Distribution plots (distplot, histplot, kdeplot)
- Categorical plots (barplot, countplot, boxplot)
- Relationship plots (scatterplot, lineplot)
- Regression plots (regplot, lmplot)
- Matrix plots (heatmap, clustermap)
- Multi-plot grids (FacetGrid, PairGrid)
- Pair plots for multivariate analysis
- Violin plots and swarm plots
- Customizing Seaborn visualizations

Advanced Visualization

- Plotly for interactive visualizations
- Plotly Express for quick plots
- Creating interactive dashboards
- Bokeh for web-based visualizations

- Geographical data visualization
- Network graphs and relationships
- 3D plotting
- Animated visualizations
- Chart selection for different data types
- Dashboard creation with Streamlit
- Dashboard creation with Dash
- Tableau fundamentals
- Power BI basics

Phase 5: Exploratory Data Analysis (EDA)

Data Profiling

- Understanding data structure and schema
- Identifying data types
- Checking for missing values
- Detecting duplicates
- Summary statistics generation
- Distribution analysis
- Identifying outliers
- Data quality assessment
- Using pandas-profiling for automated EDA
- Creating data dictionaries

Univariate Analysis

- Analyzing single variable distributions
- Measures of central tendency
- Measures of spread and variability
- Skewness and kurtosis
- Visualizing single variables
- Identifying patterns in individual features

Bivariate & Multivariate Analysis

- Correlation analysis
- Covariance matrices
- Scatter plot matrices
- Cross-tabulation
- Chi-square tests for categorical variables
- Feature relationships and dependencies
- Interaction effects
- Dimensionality reduction for visualization

Data Cleaning & Preprocessing

- Handling missing data strategies
- Imputation techniques (mean, median, mode, forward/backward fill)
- Outlier detection and treatment
- Data normalization techniques
- Data standardization (z-score)
- Min-Max scaling
- Robust scaling
- Log transformation
- Box-Cox transformation
- Handling skewed data
- Encoding categorical variables (One-Hot, Label, Ordinal)
- Binary encoding
- Target encoding
- Feature binning and discretization
- Text data cleaning
- Date/time parsing and formatting
- Handling inconsistent data entries

Phase 6: Feature Engineering

Feature Creation

- Domain knowledge application
- Creating interaction features
- Polynomial features
- Aggregation features
- Ratio and proportion features
- Date/time feature extraction (day, month, year, weekday)
- Cyclical feature encoding (sin/cos for time)
- Text feature extraction (length, word count)
- Mathematical transformations
- Binning continuous variables

Feature Selection

- Filter methods (correlation, chi-square, ANOVA)
- Wrapper methods (forward/backward selection, RFE)
- Embedded methods (Lasso, Ridge, Tree-based)
- Variance threshold
- Mutual information
- Feature importance from models
- Principal Component Analysis (PCA)
- Handling multicollinearity
- Dimensionality reduction techniques
- t-SNE for visualization
- UMAP for dimensionality reduction
- Feature extraction vs feature selection

Advanced Feature Engineering

- Automated feature engineering (Featuretools)
- Deep feature synthesis
- Target encoding for high cardinality
- Embeddings for categorical variables

- Feature crosses
- Geospatial feature engineering
- Time series feature engineering
- Domain-specific feature engineering

Phase 7: Machine Learning Fundamentals

ML Concepts

- Supervised vs unsupervised learning
- Regression vs classification
- Training, validation, and test sets
- Overfitting and underfitting
- Bias-variance tradeoff
- Cross-validation techniques
- K-fold cross-validation
- Stratified sampling
- Model evaluation metrics
- Confusion matrix
- Accuracy, precision, recall, F1-score
- ROC curves and AUC
- PR curves
- MSE, RMSE, MAE for regression
- R-squared and adjusted R-squared

Regression Algorithms

- Simple linear regression
- Multiple linear regression
- Polynomial regression
- Ridge regression (L2 regularization)
- Lasso regression (L1 regularization)
- Elastic Net regression
- Support Vector Regression (SVR)
- Decision tree regression
- Random forest regression
- Gradient boosting regression
- XGBoost for regression
- LightGBM for regression
- CatBoost for regression

Classification Algorithms

- Logistic regression
- K-Nearest Neighbors (KNN)
- Naive Bayes classifiers
- Decision trees
- Random forests
- Gradient boosting classifiers
- XGBoost for classification
- LightGBM for classification
- CatBoost for classification
- Support Vector Machines (SVM)
- Multi-class classification strategies
- Handling imbalanced datasets (SMOTE, undersampling, oversampling)

Unsupervised Learning

- K-means clustering
- Hierarchical clustering
- DBSCAN clustering
- Gaussian Mixture Models (GMM)
- Anomaly detection techniques
- Isolation Forest
- One-Class SVM
- Principal Component Analysis (PCA)
- Independent Component Analysis (ICA)
- Factor analysis
- Association rule learning (Apriori, FP-Growth)

Model Selection & Tuning

- Hyperparameter tuning concepts
- Grid search
- Random search
- Bayesian optimization
- Hyperopt and Optuna
- Learning curves analysis
- Validation curves
- Model comparison strategies
- Ensemble methods (bagging, boosting, stacking)
- Voting classifiers
- Pipeline creation with Scikit-learn

Phase 8: Advanced Machine Learning

Time Series Analysis

- Time series components (trend, seasonality, noise)
- Stationarity and differencing
- Autocorrelation (ACF) and Partial Autocorrelation (PACF)
- Moving averages
- Exponential smoothing
- ARIMA models
- SARIMA for seasonal data
- Prophet for forecasting
- Time series cross-validation
- Feature engineering for time series
- LSTM for time series forecasting
- Multivariate time series analysis

Natural Language Processing (NLP)

- Text preprocessing (tokenization, stemming, lemmatization)
- Stop word removal
- Bag of Words (BoW)
- TF-IDF (Term Frequency-Inverse Document Frequency)
- N-grams
- Word embeddings (Word2Vec, GloVe)
- Sentiment analysis
- Text classification
- Named Entity Recognition (NER)
- Topic modeling (LDA, NMF)
- Text similarity measures
- NLTK and spaCy libraries
- Regular expressions for text
- Introduction to transformers and BERT

Recommender Systems

- Collaborative filtering (user-based, item-based)
- Content-based filtering
- Matrix factorization (SVD, NMF)
- Evaluation metrics (precision@k, recall@k, NDCG)
- Cold start problem solutions
- Hybrid recommendation systems
- Deep learning for recommendations

Dimensionality Reduction

- PCA (Principal Component Analysis)
- t-SNE (t-Distributed Stochastic Neighbor Embedding)
- UMAP (Uniform Manifold Approximation and Projection)
- LDA (Linear Discriminant Analysis)
- Autoencoders for feature learning
- Kernel PCA
- Incremental PCA for large datasets

Phase 9: Deep Learning Basics

Neural Networks Fundamentals

- Perceptron and neuron model
- Activation functions (ReLU, sigmoid, tanh, softmax)
- Feedforward neural networks
- Backpropagation algorithm
- Loss functions (MSE, cross-entropy)
- Optimization algorithms (SGD, Adam, RMSprop)
- Learning rate and scheduling
- Batch, mini-batch, and stochastic gradient descent
- Weight initialization strategies
- Regularization techniques (L1, L2, dropout)
- Batch normalization

Deep Learning Frameworks

- TensorFlow basics
- Keras API
- PyTorch fundamentals
- Building neural networks from scratch
- Model training and evaluation
- Model saving and loading
- Transfer learning concepts
- GPU acceleration basics

Computer Vision Basics

- Image data preprocessing
- Convolutional Neural Networks (CNN)
- Convolutional and pooling layers
- Image classification
- Data augmentation for images
- Transfer learning with pre-trained models (VGG, ResNet)
- Object detection fundamentals

- Image segmentation basics

Sequence Models

- Recurrent Neural Networks (RNN)
- Long Short-Term Memory (LSTM)
- Gated Recurrent Units (GRU)
- Sequence-to-sequence models
- Attention mechanisms
- Transformer architecture basics
- BERT and GPT fundamentals

Phase 10: Big Data & Cloud Computing

Big Data Technologies

- Hadoop ecosystem overview
- HDFS (Hadoop Distributed File System)
- MapReduce fundamentals
- Apache Spark basics
- PySpark for data processing
- Spark DataFrames and SQL
- Spark MLlib for machine learning
- Distributed computing concepts
- Apache Kafka for streaming
- Apache Hive for data warehousing
- Data lakes vs data warehouses

Cloud Platforms

- AWS (Amazon Web Services) basics
- AWS S3 for storage
- AWS EC2 for compute
- AWS SageMaker for ML
- AWS Lambda for serverless
- Google Cloud Platform (GCP) basics
- Google BigQuery
- Google Cloud Storage
- Google Vertex AI
- Azure fundamentals
- Azure Machine Learning
- Cloud cost optimization
- Serverless architectures

Data Engineering Basics

- ETL (Extract, Transform, Load) processes
- Data pipelines design

- Apache Airflow for workflow orchestration
- Data quality and validation
- Data versioning
- Batch vs stream processing
- API development and integration
- Docker for containerization
- CI/CD basics

Phase 11: Model Deployment & MLOps

Model Deployment

- Model serialization (pickle, joblib)
- Creating REST APIs with Flask
- Creating REST APIs with FastAPI
- Model serving with TensorFlow Serving
- Model serving with TorchServe
- Docker containerization for models
- Kubernetes basics
- Model versioning strategies
- A/B testing for models
- Canary deployments
- Blue-green deployments

MLOps Practices

- Experiment tracking with MLflow
- Weights & Biases (wandb) for tracking
- Model registry and management
- Feature stores
- Data versioning with DVC
- Model monitoring in production
- Data drift detection
- Model drift detection
- Automated retraining pipelines
- Model performance monitoring
- Logging and alerting
- Model explainability (SHAP, LIME)
- Model interpretability techniques

Production Best Practices

- Scalability considerations
- Latency optimization

- Error handling and fallbacks
- Load balancing
- Caching strategies
- Security best practices
- Model governance
- Compliance and regulations (GDPR, etc.)
- Documentation and communication

Phase 12: Specialized Domains

Business Intelligence & Analytics

- KPI definition and tracking
- Dashboard design principles
- Data storytelling
- Executive reporting
- Cohort analysis
- Funnel analysis
- RFM (Recency, Frequency, Monetary) analysis
- Customer segmentation
- Churn prediction
- Customer lifetime value (CLV)
- Marketing mix modeling
- Attribution modeling

A/B Testing & Experimentation

- Experimental design
- Sample size calculation
- Statistical significance testing
- P-values and confidence intervals
- Multiple testing correction
- Sequential testing
- Multi-armed bandit algorithms
- Causal inference basics

Financial Analytics

- Risk modeling
- Credit scoring
- Fraud detection
- Algorithmic trading basics
- Portfolio optimization
- Time series forecasting for finance

- Survival analysis

Healthcare Analytics

- Clinical data analysis
- Medical image analysis basics
- Predictive modeling for patient outcomes
- Drug discovery fundamentals
- HIPAA compliance
- Electronic Health Records (EHR) analysis

Geospatial Analytics

- Working with geographic data
- Mapping and visualization with Folium
- GeoPandas for spatial analysis
- Location-based analysis
- Spatial clustering
- Route optimization

Phase 13: Soft Skills & Communication

Data Storytelling

- Understanding your audience
- Structuring data narratives
- Choosing the right visualizations
- Presenting insights effectively
- Creating compelling presentations
- Executive communication
- Technical writing
- Report writing

Business Acumen

- Understanding business metrics
- ROI analysis for data projects
- Stakeholder management
- Translating business problems to data problems
- Project scoping
- Resource estimation
- Cost-benefit analysis
- Product thinking
- Domain knowledge acquisition

Collaboration & Teamwork

- Cross-functional collaboration
- Working with engineers
- Working with product managers
- Working with business stakeholders
- Code review practices
- Knowledge sharing
- Mentoring and teaching
- Agile methodologies
- Project management basics

Ethics & Responsible AI

- Bias in machine learning models
- Fairness and equity considerations
- Privacy and data protection
- Explainable AI principles
- Model transparency
- Ethical data collection
- Responsible AI frameworks
- Legal and regulatory compliance

Phase 14: Tools & Development Environment

Development Tools

- Jupyter Notebooks and JupyterLab
- Google Colab
- VS Code for data science
- PyCharm IDE
- Anaconda distribution
- Virtual environments (venv, conda)
- Git and GitHub
- Git workflows and collaboration
- Code documentation best practices
- Markdown for documentation

Version Control & Collaboration

- Git fundamentals (commit, push, pull)
- Branching and merging
- Pull requests and code review
- Resolving merge conflicts
- GitHub/GitLab/Bitbucket
- Collaborative workflows
- Issue tracking
- Project management tools (Jira, Trello)

Testing & Quality Assurance

- Unit testing in Python (pytest, unittest)
- Test-driven development (TDD)
- Data validation testing
- Model testing strategies
- Integration testing
- Continuous integration (CI)
- Code quality tools (pylint, flake8)
- Code formatting (black, autopep8)

Phase 15: Continuous Learning & Growth

Learning Resources

- Following data science blogs and publications
- Reading research papers
- Participating in Kaggle competitions
- Contributing to open-source projects
- Attending meetups and conferences
- Online courses and MOOCs
- YouTube channels and podcasts
- Data science communities (Reddit, Stack Overflow)
- Building a personal portfolio
- Writing technical blog posts
- Creating tutorials and sharing knowledge

Key Books

- "Python for Data Analysis" by Wes McKinney
- "Hands-On Machine Learning" by Aurélien Géron
- "The Elements of Statistical Learning" by Hastie, Tibshirani, Friedman
- "Pattern Recognition and Machine Learning" by Christopher Bishop
- "Introduction to Statistical Learning" by James, Witten, Hastie, Tibshirani
- "Storytelling with Data" by Cole Nussbaumer Knaflic
- "Data Science for Business" by Foster Provost and Tom Fawcett
- "Designing Data-Intensive Applications" by Martin Kleppmann

Online Courses

- Andrew Ng's Machine Learning (Coursera)
- IBM Data Science Professional Certificate
- Google Data Analytics Certificate
- Fast.ai courses
- DataCamp career tracks
- Udacity Data Science Nanodegree
- edX MicroMasters in Statistics and Data Science

Practice Platforms

- Kaggle (competitions and datasets)
- LeetCode (programming practice)
- HackerRank (coding challenges)
- DataCamp (interactive learning)
- StrataScratch (SQL and Python interview prep)
- Mode Analytics (SQL practice)
- Google Dataset Search

Project Portfolio Ideas

Beginner Projects

- Exploratory data analysis on a public dataset
- Sales forecasting with time series
- Customer segmentation with clustering
- Sentiment analysis on social media data
- House price prediction with regression
- Credit card fraud detection
- Movie recommendation system
- COVID-19 data visualization dashboard

Intermediate Projects

- A/B test analysis framework
- Customer churn prediction system
- Product recommendation engine
- Real-time stock price prediction
- Natural language processing for text classification
- Image classification with deep learning
- Web scraping and analysis pipeline
- Interactive dashboard with Streamlit or Dash
- SQL database design and analysis project

Advanced Projects

- End-to-end ML pipeline with deployment
- Multi-model ensemble system
- Real-time data processing with Spark
- Cloud-based ML solution (AWS/GCP/Azure)
- Automated ML (AutoML) system
- Deep learning model for computer vision
- NLP chatbot with transformers
- Big data analytics on large-scale datasets
- MLOps pipeline with monitoring and retraining

- ■ Causal inference analysis project

Total Topics: 500+

Estimated Timeline: 12-18 months of dedicated study

This roadmap covers the complete data science journey from fundamentals to advanced specializations. Focus on building a strong foundation before moving to advanced topics.

Remember: Practice is key! Work on real projects, participate in competitions, and continuously learn. Track your progress by checking off completed items. Good luck on your data science journey!