





Data Scientist RoadMap

└─ 1. Foundational Knowledge:

| └─ Mathematics:


| | └─  Linear Algebra

| | └─  Calculus

| | └─  Probability and Statistics


| └─ Programming:


| └─ Python:

| | └─  Syntax and Basic Concepts

| | └─  Data Structures

| | └─  Control Structures

| | └─  Functions

| | └─  Object-Oriented Programming

| └─  R (optional, based on preference)

└─ 2. Data Manipulation and Visualization:

| └─ Data Manipulation:


| | └─  Numpy (Python)

| | └─  Pandas (Python)


| | └─  Dplyr (R)

| └─ Data Visualization:







| | └─  Matplotlib (Python)

| | └─  Seaborn (Python)













| | └─  ggplot2 (R)

| └─  Interactive Visualization Tools

└ 3. Exploratory Data Analysis (EDA) and Preprocessing:

- | └  Exploratory Data Analysis Techniques
- | └  Feature Engineering
- | └  Data Cleaning
- | └  Handling Missing Data
- | └  Data Scaling and Normalization
- | └  Outlier Detection and Treatment

└ 4. Machine Learning:


- | └  **Supervised Learning:**
 - | └  **Regression:**
 - | └  Linear Regression
 - | └  Polynomial Regression
 - | └  Regularization Techniques
 - | └  **Classification:**
 - | └  Logistic Regression
 - | └  k-Nearest Neighbors (k-NN)
 - | └  Support Vector Machines (SVM)
 - | └  Decision Trees
 - | └  Random Forest
 - | └  Gradient Boosting

- | | └─ 🤖 **Unsupervised Learning:**
 - | | └─ 📁 Clustering:
 - | | └─ 📁 K-means
 - | | └─ 🔵 DBSCAN
 - | | └─ 🌳 Hierarchical Clustering
 - | | └─ 📊 Dimensionality Reduction:
 - | | └─ 📊 Principal Component Analysis (PCA)
 - | | └─ 🔵 t-Distributed Stochastic Neighbor Embedding (t-SNE)
 - | | └─ 🏠 Linear Discriminant Analysis (LDA)
 - | | └─ ↔ Association Rule Learning
- | └─ 🤖 **Reinforcement Learning**
- | └─ 🏆 **Model Evaluation and Validation:**
 - | └─ 📊 Cross-validation
 - | └─ ⚙️ Hyperparameter Tuning
 - | └─ 👤 Model Selection Techniques
 - | └─ 📊 Evaluation Metrics

└ 5. Deep Learning:


| └ Neural Networks:


| | └  Perceptron

| | └  Multi-Layer Perceptron (MLP)


| └ Convolutional Neural Networks (CNNs):

| | └  Image Classification

| | └  Object Detection

| | └  Image Segmentation


| └ Recurrent Neural Networks (RNNs):

| | └  Sequence-to-Sequence Models

| | └  Text Classification


| | └  Sentiment Analysis

| └ Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU):

| | └  Time Series Forecasting

| | └  Language Modeling

| └ Generative Adversarial Networks (GANs):

| | └  Image Synthesis

| | └  Style Transfer

| | └  Data Augmentation

└ 6. Advanced Topics:

| └ 📖 Natural Language Processing (NLP):

- | | └ 📄 Text Preprocessing
- | | └ 📁 Word Embeddings (e.g., Word2Vec, GloVe)
- | | └ ⌚ Recurrent Neural Networks for NLP
- | | └ 🚀 Transformer Models (e.g., BERT, GPT)

| └ ⌚ Time Series Analysis:

- | | └ 📈 Time Series Decomposition
- | | └ ⌚ Autoregressive Integrated Moving Average (ARIMA)
- | | └ ☁ Seasonal ARIMA (SARIMA)
- | | └ 🌡 Exponential Smoothing Methods
- | | └ 📊 Prophet

| └ 🎯 Recommender Systems:






- | | └ 💛 Collaborative Filtering
- | | └ 📖 Content-Based Filtering
- | | └ ✂ Matrix Factorization
- | | └ 🔁 Hybrid Methods

| └ 🖋 Causal Inference:

















- | | └ 🖋 Experimental Design
- | | └ 👁 Observational Studies
- | | └ 🎯 Propensity Score Matching
- | | └ 🖋 Instrumental Variable Analysis

| └ 🧠 Advanced Deep Learning:








- | | └ 🏠 Advanced Architectures (e.g., Transformers, GPT models)
- | | └ 🎨 Generative Models (e.g., VAEs, flow-based models)
- | | └ 💡 Advanced Techniques for NLP and Computer Vision

- | └─  **Bayesian Statistics and Probabilistic Programming:**
- | └─  Bayesian Inference
- | └─  Markov Chain Monte Carlo (MCMC)
- | └─  Probabilistic Graphical Models
- | └─  Stan, PyMC3, or Edward for Probabilistic Programming






| 7. Big Data Technologies:

- | └─  Hadoop:
- | └─  HDFS
- | └─  MapReduce
- | └─  **Spark:**
- | └─  RDDs
- | └─  DataFrames
- | └─  MLlib
- | └─  **NoSQL Databases:**
- | └─  MongoDB
- | └─  Cassandra
- | └─  HBase
- | └─  Couchbase
- | └─  **Stream Processing Frameworks:**
- | └─  Apache Kafka
- | └─  Apache Flink
- | └─  Apache Storm




└ 8. Data Visualization and Reporting:

- | └  Dashboarding Tools:
 - | | └  Tableau
 - | | └  Power BI
 - | | └  Dash (Python)
 - | | └  Shiny (R)
- | └  Storytelling with Data
- | └  Effective Communication





└ 9. Domain Knowledge and Soft Skills:

- | └  Industry-specific Knowledge
- | └  Problem-solving
- | └  Communication Skills
- | └  Time Management
- | └  Teamwork

└ 10. Ethical Considerations and Bias in Data Science:

- | └  Fairness in Machine Learning
- | └  Bias Detection and Mitigation
- | └  Privacy and Data Security

└ 11. Deployment and Productionization:

- | └  Model Deployment Techniques
- | └  Containerization (e.g., Docker)
- | └  Model Serving and APIs
- | └  Scalability and Performance Optimization

└─ **12. Continuous Learning and Staying Updated:**

- | └─ 🎓 Online Courses and Tutorials
- | └─ 📖 Books and Research Papers
- | └─ 📄 Blogs and Podcasts
- | └─ 🏢 Conferences and Workshops
- | └─ 👤 Networking and Community Engagement

└─ **13. Recommended Resources:**

└─ 📖 Online Courses:

- └─ [Coursera - Data Science Specialization](#)
- └─ [edX - Data Science MicroMasters Program](#)
- └─ [Kaggle Courses](#)

└─ 📖 Books:

- └─ ["Python for Data Analysis" by Wes McKinney](#)
- └─ ["Hands-On Machine Learning with Scikit-Learn and TensorFlow" by Aurélien Géron](#)
- └─ ["Deep Learning" by Ian Goodfellow, Yoshua Bengio, and Aaron Courville](#)

└─ 📺 YouTube Channels:

- └─ [Sentdex](#)
- └─ [Data School](#)
- └─ [3Blue1Brown](#)
- └─ [PyData](#)
- └─ [StatQuest with Josh Starmer](#)