## **Data Scientist RoadMap**

```
1. Foundational Knowledge:
      Mathematics:
      Linear Algebra

✓ Calculus

      Probability and Statistics
    Programming:
      Python:
          Syntax and Basic Concepts
          Data Structures
          Control Structures
          Object-Oriented Programming
      R (optional, based on preference)
2. Data Manipulation and Visualization:

→ 

■ Data Manipulation:

        Mumpy (Python)
       Pandas (Python)
        Dplyr (R)
      Data Visualization:
        Matplotlib (Python)
        Seaborn (Python)
        ggplot2 (R)
        Interactive Visualization Tools
```





$\vdash$	3. Ехр	loratory Data Analysis (EDA) and Preprocessing:
1	<b></b>	Exploratory Data Analysis Techniques
1	H- 88	Feature Engineering
1	<b>├</b>	Data Cleaning
1	H 0	Handling Missing Data
1	• • • • • • • • • • • • • • • • • •	Data Scaling and Normalization
1	L	Outlier Detection and Treatment
$\vdash$	4. Mac	thine Learning:
I	<b>⊢</b>	Supervised Learning:
1	I —	- Regression:
1	1 1	├─ ☑ Linear Regression
1	1 1	├─ ☑ Polynomial Regression
I	1 1	- Regularization Techniques
1	1 1	L— 🔟 Classification:
1	1 1	- Logistic Regression
1	1 1	├─ 🥑 k-Nearest Neighbors (k-NN)
1	1 1	- Support Vector Machines (SVM)
1	1 1	- Decision Trees
1	1 1	- Random Forest
1	1 1	└─ Ø 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

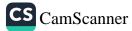
       Beinforcement Learning
       Model Evaluation and Validation:
              Cross-validation
       - W Hyperparameter Tuning
              Model Selection Techniques
             Evaluation Metrics
```

├ 5. Deep Learning:			
├─ ◯ Neural Networks:			
│			
│			
├── ☐ Convolutional Neural Networks (CNNs):			
│			
│			
│			
├── 🎖 Recurrent Neural Networks (RNNs):			
├─ ∑ Sequence-to-Sequence Models			
├─ 🗄 Text Classification			
U C Sentiment Analysis			
├─ ☑ Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU):			
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
    └─ Ø 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
    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:
      A Hadoop:
      HDFS
     MapReduce
   - Spark:
         III RDDs
          DataFrames
         MLlib
      NoSQL Databases:
      - MongoDB
         Cassandra
            HBase
         Couchbase
   Stream Processing Frameworks:
      - Apache Kafka
      - 4 Apache Flink
```

```
Dashboarding Tools:
          III 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
```

8. Data Visualization and Reporting:



$\vdash$	12.	Continuous Learning and Staying Updated:
1	-	nline Courses and Tutorials
I		Books and Research Papers
1	<u> </u>	Blogs and Podcasts
I	-	Conferences and Workshops
I	L	Networking and Community Engagement
∟	13.	Recommended Resources:
	$\vdash$	Online Courses:
		Coursera - Data Science Specialization
		<u>edX - Data Science MicroMasters Program</u>
		<u>Kaggle Courses</u>
	$\vdash$	☐ Books:
		- "Python for Data Analysis" by Wes McKinney
<u>by</u>	Auré]	— "Hands-On Machine Learning with Scikit-Learn and TensorFlow" ien Géron
Cou	rvill	<u>"Deep Learning" by Ian Goodfellow, Yoshua Bengio, and Aaron</u> e
	L	YouTube Channels:
		- Sentdex
		— Data School
		- 3Blue1Brown
		— PyData
		- StatQuest with Josh Starmer

