#_ Becoming a Data Scientist [the StudyPlan]

🌐 **Phase 1: Foundational Knowledge**

Duration: 2 months

1. **Mathematics**
    - ◯ **Linear Algebra** (15 hours)
        - ■ Study concepts like vectors, matrices, eigenvalues, and eigenvectors.
        - ■ Resources: Khan Academy's Linear Algebra course Khan Academy - Linear Algebra
    - ◯ **Calculus** (15 hours)
        - ■ Learn about differentiation, integration, limits, and derivatives.
        - ■ Resources: Khan Academy's Calculus courses Khan Academy - Calculus
    - ◯ **Probability and Statistics** (15 hours)
        - ■ Study probability theory, random variables, distributions, and basic statistics.
        - ■ Resources: Khan Academy's Probability and Statistics courses Khan Academy - Probability and Statistics
2. **Programming**
    - ◯ **Python** (60 hours)
        - ■ Syntax and Basic Concepts (10 hours)
        - ■ Data Structures (15 hours)
        - ■ Control Structures (10 hours)
        - ■ Functions (10 hours)
        - ■ Object-Oriented Programming (15 hours)
        - ■ Resources: Python.org's official tutorial Python.org Official Tutorial
    - ◯ **R (optional)** (20 hours)
        - ■ If you choose R as well, allocate time for syntax and basic concepts.
        - ■ Resources: "R for Data Science" by Hadley Wickham and Garrett Grolemund R for Data Science

Phase 2: Data Manipulation and Visualization

Duration: 2 months

1. **Data Manipulation**
   - ○ **Numpy (Python)** (20 hours)
     - ■ Learn how to work with arrays and matrices.
     - ■ Resources: Numpy documentation Numpy Documentation
   - ○ **Pandas (Python)** (30 hours)
     - ■ Study data structures like Series and DataFrames for data manipulation.
     - ■ Resources: "Python for Data Analysis" by Wes McKinney Python for Data Analysis
   - ○ **Dplyr (R)** (20 hours)
     - ■ If you chose R, learn data manipulation using dplyr.
     - ■ Resources: DataCamp's "Introduction to the Tidyverse" course Introduction to the Tidyverse
2. **Data Visualization**
   - ○ **Matplotlib (Python)** (20 hours)
     - ■ Start with basic plotting techniques.
     - ■ Resources: Matplotlib documentation Matplotlib Documentation
   - ○ **Seaborn (Python)** (20 hours)
     - ■ Explore more advanced and aesthetic visualizations.
     - ■ Resources: Seaborn documentation Seaborn Documentation
   - ○ **ggplot2 (R)** (20 hours)
     - ■ If you chose R, learn data visualization using ggplot2.
     - ■ Resources: "Data Visualization with ggplot2" by Hadley Wickham Data Visualization with ggplot2
   - ○ **Interactive Visualization Tools** (10 hours)
     - ■ Explore libraries like Plotly and Bokeh for interactive visualizations.
     - ■ Resources: Plotly documentation Plotly Documentation

📊 Phase 3: Exploratory Data Analysis and Preprocessing

Duration: 1 month

1. **Exploratory Data Analysis (EDA)** (20 hours)
    - ◯ Study techniques like histograms, scatter plots, box plots, and correlation matrices.
    - ◯ Resources: DataCamp's "Exploratory Data Analysis in Python" course Exploratory Data Analysis in Python
2. **Feature Engineering** (15 hours)
    - ◯ Understand techniques to create new features from existing data.
    - ◯ Resources: "Feature Engineering for Machine Learning" by Alice Zheng and Amanda Casari Feature Engineering for Machine Learning
3. **Data Cleaning** (10 hours)
    - ◯ Learn about identifying and handling missing values, duplicates, and inconsistencies.
    - ◯ Resources: DataCamp's "Cleaning Data in Python" course Cleaning Data in Python
4. **Handling Missing Data** (10 hours)
    - ◯ Study methods like imputation and understand the implications of missing data.
    - ◯ Resources: "Handling Missing Data in R" on DataCamp Handling Missing Data in R
5. **Data Scaling and Normalization** (5 hours)
    - ◯ Understand the importance of scaling and normalizing data for certain algorithms.
    - ◯ Resources: "Feature Scaling in Machine Learning" on Analytics Vidhya Feature Scaling in Machine Learning
6. **Outlier Detection and Treatment** (10 hours)
    - ◯ Learn how to identify and handle outliers in your data.
    - ◯ Resources: Techniques of Outlier Detection and Treatment

🛠️ Phase 4: Machine Learning

Duration: 3 months

1. **Supervised Learning: Regression** (25 hours)
    - ◯ **Linear Regression** (10 hours)
    - ◯ **Polynomial Regression** (5 hours)
    - ◯ **Regularization Techniques** (10 hours)
    - ◯ Resources: "Introduction to Machine Learning with Python" by Andreas C. Müller and Sarah Guido Introduction to Machine Learning with Python
2. **Supervised Learning: Classification** (35 hours)
    - ◯ **Logistic Regression** (10 hours)
    - ◯ **k-Nearest Neighbors (k-NN)** (5 hours)
    - ◯ **Support Vector Machines (SVM)** (10 hours)
    - ◯ **Decision Trees** (5 hours)
    - ◯ **Random Forest** (5 hours)
    - ◯ **Gradient Boosting** (10 hours)
    - ◯ Resources: Coursera's "Machine Learning" by Andrew Ng Machine Learning
3. **Unsupervised Learning: Clustering** (15 hours)
    - ◯ **K-means** (5 hours)
    - ◯ **DBSCAN** (5 hours)
    - ◯ **Hierarchical Clustering** (5 hours)
    - ◯ Resources: "Introduction to Unsupervised Learning" on DataCamp Introduction to Unsupervised Learning
4. **Unsupervised Learning: Dimensionality Reduction** (15 hours)
    - ◯ **Principal Component Analysis (PCA)** (5 hours)
    - ◯ **t-Distributed Stochastic Neighbor Embedding (t-SNE)** (5 hours)
    - ◯ **Linear Discriminant Analysis (LDA)** (5 hours)
    - ◯ **Association Rule Learning** (5 hours)
    - ◯ Resources: Introduction to Unsupervised Learning
5. **Model Evaluation and Validation** (20 hours)
    - ◯ **Cross-validation** (5 hours)
    - ◯ **Hyperparameter Tuning** (5 hours)
    - ◯ **Model Selection Techniques** (5 hours)

- ◯ **Evaluation Metrics** (5 hours)
- ◯ Resources: scikit-learn documentation on Model Selection and Evaluation [scikit-learn Model Selection and Evaluation]

⚙️ Phase 5: Deep Learning

Duration: 3 months

1. **Neural Networks** (20 hours)
   - ◯ **Perceptron** (5 hours)
   - ◯ **Multi-Layer Perceptron (MLP)** (15 hours)
   - ◯ Resources: Coursera's "Neural Networks and Deep Learning" by Andrew Ng [Neural Networks and Deep Learning]
2. **Convolutional Neural Networks (CNNs)** (25 hours)
   - ◯ **Image Classification** (10 hours)
   - ◯ **Object Detection** (10 hours)
   - ◯ **Image Segmentation** (5 hours)
   - ◯ Resources: Deep Learning Specialization on Coursera by Andrew Ng [Deep Learning Specialization]
3. **Recurrent Neural Networks (RNNs)** (20 hours)
   - ◯ **Sequence-to-Sequence Models** (10 hours)
   - ◯ **Text Classification** (5 hours)
   - ◯ **Sentiment Analysis** (5 hours)
   - ◯ Resources: "Natural Language Processing Specialization" on Coursera by deeplearning.ai [Natural Language Processing Specialization]
4. **Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU)** (15 hours)
   - ◯ **Time Series Forecasting** (10 hours)
   - ◯ **Language Modeling** (5 hours)
   - ◯ Resources: "Sequence Models" course on Coursera by deeplearning.ai [Sequence Models]
5. **Generative Adversarial Networks (GANs)** (15 hours)
   - ◯ **Image Synthesis** (5 hours)
   - ◯ **Style Transfer** (5 hours)
   - ◯ **Data Augmentation** (5 hours)

- ◯ Resources: "Generative Adversarial Networks (GANs) Specialization" on Coursera by deeplearning.ai Generative Adversarial Networks (GANs) Specialization

📈 **Phase 6: Advanced Topics**

Duration: 3 months

1. **Natural Language Processing (NLP)** (30 hours)
   - ◯ **Text Preprocessing** (10 hours)
   - ◯ **Word Embeddings** (10 hours)
   - ◯ **Recurrent Neural Networks for NLP** (5 hours)
   - ◯ **Transformer Models (e.g., BERT, GPT)** (5 hours)
   - ◯ Resources: "Natural Language Processing in Action" by Hobson Lane, Cole Howard, and Hannes Hapke Natural Language Processing in Action
2. **Time Series Analysis** (20 hours)
   - ◯ **Time Series Decomposition** (5 hours)
   - ◯ **Autoregressive Integrated Moving Average (ARIMA)** (5 hours)
   - ◯ **Seasonal ARIMA (SARIMA)** (5 hours)
   - ◯ **Exponential Smoothing Methods** (5 hours)
   - ◯ **Prophet** (5 hours)
   - ◯ Resources: "Time Series Analysis and Its Applications" by Robert H. Shumway and David S. Stoffer Time Series Analysis and Its Applications

🌟 **Phase 6: Advanced Topics (Continued)**

Duration: 3 months

3. **Recommender Systems** (15 hours)
   - ◯ **Collaborative Filtering** (5 hours)
   - ◯ **Content-Based Filtering** (5 hours)
   - ◯ **Matrix Factorization** (5 hours)
   - ◯ **Hybrid Methods** (5 hours)
   - ◯ Resources: "Recommender Systems Handbook" by Francesco Ricci, Lior Rokach, and Bracha Shapira Recommender Systems Handbook
   - ○

4. **Causal Inference** (15 hours)
    - ◯ **Experimental Design** (5 hours)
    - ◯ **Observational Studies** (5 hours)
    - ◯ **Propensity Score Matching** (5 hours)
    - ◯ **Instrumental Variable Analysis** (5 hours)
    - ◯ Resources: "Causal Inference: What If" by Miguel A. Hernán and James M. Robins Causal Inference: What If
5. **Advanced Deep Learning** (25 hours)
    - ◯ **Advanced Architectures** (10 hours)
    - ◯ **Generative Models** (10 hours)
    - ◯ **Advanced Techniques for NLP and Computer Vision** (5 hours)
    - ◯ Resources: "Dive into Deep Learning" by Aston Zhang, Zachary C. Lipton, and Mu Li Dive into Deep Learning
6. **Bayesian Statistics and Probabilistic Programming** (20 hours)
    - ◯ **Bayesian Inference** (5 hours)
    - ◯ **Markov Chain Monte Carlo (MCMC)** (5 hours)
    - ◯ **Probabilistic Graphical Models** (5 hours)
    - ◯ **Stan, PyMC3, or Edward** (5 hours)
    - ◯ Resources: "Probabilistic Programming & Bayesian Methods for Hackers" by Cam Davidson-Pilon Probabilistic Programming & Bayesian Methods for Hackers

📄 Phase 7: Big Data Technologies

Duration: 2 months

1. **Cloud Services** (15 hours)
    - ◯ **Cloud Providers** (5 hours)
    - ◯ **AWS Services (Optional)** (10 hours)
    - ◯ Resources: AWS Documentation AWS Documentation
2. **Spark** (20 hours)
    - ◯ **Understanding RDDs** (5 hours)
    - ◯ **DataFrames** (5 hours)
    - ◯ **MLlib** (10 hours)
    - ◯ Resources: "Learning Spark" by Holden Karau, Andy Konwinski, Patrick Wendell, and Matei Zaharia Learning Spark
    - ○

3. **NoSQL Databases** (15 hours)
    - ◯ **MongoDB** (5 hours)
    - ◯ **Cassandra** (5 hours)
    - ◯ **HBase and Couchbase** (5 hours)
    - ◯ Resources: MongoDB Documentation MongoDB Documentation
4. **Stream Processing Frameworks** (10 hours)
    - ◯ **Apache Kafka** (5 hours) ◯ **Apache Flink** (2.5 hours)
    - ◯ **Apache Storm** (2.5 hours) ◯ Resources: Apache Kafka Documentation Apache Kafka

Documentation

📊 Phase 8: Data Visualization and Reporting

Duration: 1 month

1. **Dashboarding Tools** (15 hours)
    - ◯ **Tableau** (5 hours)
    - ◯ **Power BI** (5 hours)
    - ◯ **Dash (Python)** (2.5 hours)
    - ◯ **Shiny (R)** (2.5 hours)
    - ◯ Resources: Tableau Public Gallery
    - ◯ Resources: Power BI Learning Resources Power BI Learning Resources
    - ◯ Resources: Plotly Dash Documentation Plotly Dash Documentation
    - ◯ Resources: Shiny Gallery Shiny Gallery
2. **Storytelling with Data** (10 hours)
    - ◯ **"Storytelling with Data"** by Cole Nussbaumer Knaflic (Book)
    - ◯ Resources: "Storytelling with Data" by Cole Nussbaumer Knaflic Storytelling with Data
3. **Effective Communication** (5 hours)
    - ◯ **"Communicating Data Science Results"** on Coursera
    - ◯ Resources: "Communicating Data Science Results" on Coursera by the University of Washington Communicating Data Science Results

📙 Phase 9: Domain Knowledge and Soft Skills

Duration: Ongoing

1. **Industry-specific Knowledge** (Ongoing)
    - ◯ Stay updated with industry trends, use cases, and challenges.
2. **Problem-solving** (Ongoing)
    - ◯ Regularly solve coding challenges and participate in data science competitions.
3. **Communication Skills** (Ongoing)
    - ◯ Engage in discussions, write blog posts, and present your findings.
4. **Time Management** (Ongoing)
    - ◯ Continuously adjust your schedule based on your progress and goals.
5. **Teamwork** (Ongoing)
    - ◯ Collaborate on projects, join data science communities, and attend meetups.

🔒 Phase 10: Ethical Considerations and Bias in Data Science

Duration: Ongoing

1. **Fairness in Machine Learning** (Ongoing)
    - ◯ Follow recent research and guidelines on bias and fairness in AI.
2. **Bias Detection and Mitigation** (Ongoing)
    - ◯ Stay informed about techniques and tools for detecting and mitigating bias.
3. **Privacy and Data Security** (Ongoing)
    - ◯ Keep up with best practices and regulations related to data privacy.

💻 Phase 11: Deployment and Productionisation

Duration: Ongoing

1. **Model Deployment Techniques** (Ongoing)
    - ◯ Explore various deployment platforms and techniques.
2. **Containerization (e.g., Docker)** (Ongoing)
    - ◯ Learn how to package and deploy models using containers.
3. **Model Serving and APIs** (Ongoing)
    - ◯ Experiment with building APIs for serving models.
4. **Scalability and Performance Optimization** (Ongoing)
    - ◯ Study strategies to optimize model performance and scalability.