⚙️ Phase 1: Python Refresher (Week 1)

Python basics (variables, loops, functions, data types)

Practice file I/O, error handling, basic OOP (optional)

📦 Project:

Title: “Python-Based Data Profiler”

Goal: Build a Python script that takes a CSV file and prints insights like:

Row/column count

Column types

Null counts

Sample rows

🔢 Phase 2: NumPy & Pandas Mastery (Week 2–3)

Learn NumPy arrays, slicing, broadcasting

Learn Pandas Series, DataFrames, filtering, joins, groupby

📦 Project:

Title: “Titanic Data Explorer”

Goal: Clean, analyze, and visualize Titanic dataset with Pandas

Stretch: Save insights and plots to a report-style Jupyter notebook

📊 Phase 3: Data Visualization (Week 4)

Learn matplotlib, seaborn, and styling plots

Visualize distributions, correlations, trends

📦 Project:

Title: “Spotify Songs”

Goal: Visualize Spotify Song datasets with scatter plots, histograms, box plots, etc.

🌐 Phase 4: Data Collection & Wrangling (Week 5)

Web scraping with requests, BeautifulSoup

JSON/XML parsing, CSV handling

📦 Project:

Title: “Top GitHub Repos Analyzer”

Goal: Scrape trending GitHub repos (or use GitHub API) and analyze stars, forks, languages

🗃️ Phase 5: SQL for Data Science (Week 6)

SELECT, JOIN, GROUP BY, HAVING, subqueries

Practice on SQLite / Mode / Kaggle

Practice on SQLite, but also create a free-tier cloud database (e.g., AWS RDS) and learn to connect to it.

📦 Project:

Title: “Sales Insights Dashboard (SQL + Pandas)”

Goal: Use SQL to extract data, then analyze with Pandas

NEW - Phase 6.1: Data Engineering & Automation (Week 7)

- **New:** Learn the concepts of ETL/ELT pipelines.
- **New:** Install and run Apache Airflow (using Docker is the industry standard).
- **New Project:** "Automated GitHub Trends Pipeline" - Convert your Phase 4 project into an Airflow DAG that runs daily, scrapes the data, and loads it into your SQLite/AWS RDS database.

📐 Phase 6.2: Probability & Statistics (Week 7–8)

Probability theory, distributions, CLT

Descriptive stats, z-scores, outliers

📦 Project:

Title: “Stats Report Generator”

Goal: Write a Python script that computes key stats + generates a PDF/HTML report for any CSV

🤖 Phase 7: Machine Learning Intro (Week 9–10)

Understand supervised vs unsupervised learning

Use Scikit-learn for regression, classification

** *Try running your model training script inside an AWS SageMaker notebook.*

📦 Project:

Title: “House Price Predictor”

Goal: Train a linear regression model on housing data

Stretch: Include train/test split, evaluation metrics

🧠 Phase 8: Neural Networks & Deep Learning (Week 11)

Perceptrons, activation, forward/backpropagation

Intro to TensorFlow/Keras /Pytorch

📦 Project:

Title: “Digit Classifier (MNIST)”

Goal: Train a basic NN to classify handwritten digits

🔍 Phase 9: LLMs & GenAI Concepts (Week 12)

Learn how LLMs work (tokens, embeddings, context)

Intro to RAG systems and prompt engineering

📦 Project:

Title: “Your Custom GPT Bot”

Goal: Build a Q&A bot using GPT + CSV-based RAG-style search (conceptual at this stage)

🌍 Phase 10: Web Apps for Data Science (Week 13)

Learn Flask or Streamlit

Connect models/data to web interfaces

ML-Ops Introduction

Deploy your web app to a cloud service (e.g., Heroku, AWS Elastic Beanstalk, or Streamlit Cloud). This is a HUGE win.

📦 Project:

Title: “DS Web Dashboard”

Goal: Turn your earlier project (Titanic, IMDb, or Sales) into an interactive web app

🔁 Bonus Projects (Anytime)

EDA on a Kaggle dataset (you pick the domain)

Your GitHub Portfolio – centralize all notebooks/projects

Write a blog post explaining one project

Peer review – ask someone to give you feedback