**🚀 Roadmap to Becoming a Data Analyst**

⏳ **Duration:** 3 - 6 months  
🔹 **Goal:** Learn Python, SQL, Excel, and data visualization tools while building real-world projects  
🔹 **Method:** Combine structured learning with hands-on projects

# Details on projects

**📌 Phase 1: Python for Data Analysis (March 3 - April 7)**

**🔹 Goal: Learn Python fundamentals and how to handle real-world data  
🔹 Tools: Python, Jupyter Notebook, Pandas, NumPy, Matplotlib  
🔹 Duration: 5 weeks**

|  |  |  |  |
| --- | --- | --- | --- |
| **Week** | **Learning Topic** | **Hands-on Practice** | **Project** |
| **March 3 - March 10** | **Python Basics (Variables, Loops, Functions)** | **Small coding exercises (loops, conditions, lists, functions)** | **Print "Hello, Data!" and do simple calculations** |
| **March 11 - March 17** | **Pandas & NumPy (DataFrames, Arrays)** | **Create a DataFrame with sample data** | **Project 1: Data Cleaning in Python (Netflix dataset)** |
| **March 18 - March 24** | **Data Visualization (Matplotlib, Seaborn)** | **Plot basic graphs (bar chart, scatterplot)** | **Project 2: Visualizing Sales Data (Retail sales dataset)** |
| **March 25 - March 31** | **Exploratory Data Analysis (EDA)** | **Find trends, outliers in a dataset** | **Project 3: Analyzing Employee Data** |
| **April 1 - April 7** | **Mini Capstone: Combine all skills** | **Work with a real dataset** | **Project 4: Analyzing COVID-19 Trends** |

**✅ By April 7: You will be comfortable using Python for data cleaning, visualization, and analysis.**

**📌 Phase 2: SQL for Data Analysis (April 8 - April 28)**

**🔹 Goal: Learn SQL for querying and managing databases  
🔹 Tools: MySQL or PostgreSQL  
🔹 Duration: 3 weeks**

|  |  |  |  |
| --- | --- | --- | --- |
| **Week** | **Learning Topic** | **Hands-on Practice** | **Project** |
| **April 8 - April 14** | **SQL Basics (SELECT, WHERE, ORDER BY)** | **Write simple queries** | **Project 5: Analyzing Customer Orders** |
| **April 15 - April 21** | **Joins & Aggregations (JOIN, GROUP BY, COUNT)** | **Write queries using multiple tables** | **Project 6: Employee Performance Analysis** |
| **April 22 - April 28** | **Subqueries & Window Functions** | **Use advanced SQL techniques** | **Project 7: Tracking Sales Trends Over Time** |

**✅ By April 28: You will be able to extract and analyze data from databases.**

**📌 Phase 3: Advanced Excel & Power BI (April 29 - May 19)**

**🔹 Goal: Learn Excel formulas, pivot tables, and Power BI for dashboards  
🔹 Tools: Microsoft Excel, Power BI  
🔹 Duration: 3 weeks**

|  |  |  |  |
| --- | --- | --- | --- |
| **Week** | **Learning Topic** | **Hands-on Practice** | **Project** |
| **April 29 - May 5** | **Excel Basics (Formulas, Pivot Tables)** | **Create reports and summaries** | **Project 8: HR Employee Dashboard** |
| **May 6 - May 12** | **Power BI Basics** | **Build simple dashboards** | **Project 9: Sales Insights Dashboard** |
| **May 13 - May 19** | **Interactive Dashboards** | **Combine Excel & Power BI** | **Project 10: Financial Performance Dashboard** |

**✅ By May 19: You will be able to create interactive dashboards for business insights.**

**📌 Phase 4: Machine Learning & Portfolio Projects (May 20 - June 30)**

**🔹 Goal: Learn basic machine learning and build a strong portfolio  
🔹 Tools: Python (Scikit-learn), Tableau, GitHub  
🔹 Duration: 6 weeks**

|  |  |  |  |
| --- | --- | --- | --- |
| **Week** | **Learning Topic** | **Hands-on Practice** | **Project** |
| **May 20 - May 26** | **Intro to Machine Learning (Linear Regression)** | **Train a basic model** | **Project 11: Predicting House Prices** |
| **May 27 - June 2** | **Classification (Logistic Regression, Decision Trees)** | **Work with classification models** | **Project 12: Credit Card Fraud Detection** |
| **June 3 - June 9** | **Clustering & Recommendation Systems** | **Segment customers using ML** | **Project 13: Movie Recommendation System** |
| **June 10 - June 16** | **Tableau & Data Storytelling** | **Present data visually** | **Project 14: Interactive Business Report in Tableau** |
| **June 17 - June 30** | **Portfolio & Resume Building** | **Upload projects to GitHub** | **Final Capstone: End-to-End Data Analysis Project** |

**✅ By June 30: You will have a strong portfolio with 15+ projects to showcase your skills.**

**🔹 Summary of Your Roadmap**

**✅ March 3 - April 7: Python for Data Analysis (Pandas, NumPy, Visualization)  
✅ April 8 - April 28: SQL for Data Analysis (Querying Databases)  
✅ April 29 - May 19: Advanced Excel & Power BI for Reporting  
✅ May 20 - June 30: Machine Learning + Portfolio Building**

# Datasets I could use

**📌 Phase 1: Python for Data Analysis (March 3 - April 7)**

**1️⃣ Project: Data Cleaning in Python**

* **Dataset:** Netflix Movies & TV Shows
* **Description:** Contains details of TV shows and movies on Netflix (title, release year, country, genre, etc.).
* **What to Do?** Clean missing values, format dates, remove duplicates.

**2️⃣ Project: Visualizing Sales Data**

* **Dataset:** Superstore Sales Data
* **Description:** Sales records from an office supplies store, including sales, profit, and category details.
* **What to Do?** Create bar charts, scatter plots, and histograms to analyze trends.

**3️⃣ Project: Analyzing Employee Data**

* **Dataset:** HR Employee Attrition Dataset
* **Description:** Employee data with factors like salary, job satisfaction, promotions, etc.
* **What to Do?** Analyze trends and patterns in employee attrition.

**4️⃣ Project: COVID-19 Trends Analysis**

* **Dataset:** COVID-19 Dataset
* **Description:** Global COVID-19 cases, deaths, and recoveries by country.
* **What to Do?** Analyze trends in cases and visualize key patterns.

**📌 Phase 2: SQL for Data Analysis (April 8 - April 28)**

**5️⃣ Project: Analyzing Customer Orders**

* **Dataset:** Chinook Database (Music Store)
* **Description:** SQL database containing music store sales, customers, and employees.
* **What to Do?** Write SQL queries to find top-selling artists, most loyal customers, etc.

**6️⃣ Project: Employee Performance Analysis**

* **Dataset:** HR Database
* **Description:** SQL dataset with employee records, salary, and department data.
* **What to Do?** Use SQL to analyze employee performance and promotions.

**7️⃣ Project: Tracking Sales Trends Over Time**

* **Dataset:** Retail Sales SQL Dataset
* **Description:** Sales transactions, product categories, and customer demographics.
* **What to Do?** Find trends, monthly sales growth, and customer behavior.

**📌 Phase 3: Advanced Excel & Power BI (April 29 - May 19)**

**8️⃣ Project: HR Employee Dashboard**

* **Dataset:** HR Analytics Dataset
* **Description:** Employee salaries, satisfaction, and department performance.
* **What to Do?** Create pivot tables, charts, and an interactive Excel dashboard.

**9️⃣ Project: Sales Insights Dashboard**

* **Dataset:** Global Superstore Sales
* **Description:** Sales data with categories, regions, and customer information.
* **What to Do?** Build a Power BI dashboard showing sales trends and customer behavior.

**🔟 Project: Financial Performance Dashboard**

* **Dataset:** Stock Market Data
* **Description:** Historical stock prices for major companies.
* **What to Do?** Analyze stock price trends and create a Power BI report.

**📌 Phase 4: Machine Learning & Portfolio Projects (May 20 - June 30)**

**1️⃣1️⃣ Project: Predicting House Prices**

* **Dataset:** House Prices Dataset
* **Description:** Home prices with features like location, number of rooms, and square footage.
* **What to Do?** Use regression models to predict house prices.

**1️⃣2️⃣ Project: Credit Card Fraud Detection**

* **Dataset:** Credit Card Fraud Dataset
* **Description:** Transactions labeled as fraudulent or not.
* **What to Do?** Build a classification model to detect fraud.

**1️⃣3️⃣ Project: Movie Recommendation System**

* **Dataset:** MovieLens Dataset
* **Description:** Movie ratings by users.
* **What to Do?** Use machine learning to recommend movies to users.

**1️⃣4️⃣ Project: Interactive Business Report in Tableau**

* **Dataset:** E-commerce Sales Data
* **Description:** Online store sales transactions.
* **What to Do?** Build an interactive Tableau dashboard.

**1️⃣5️⃣ Final Capstone Project: End-to-End Data Analysis**

* **Dataset:** Choose from above or find your own on Kaggle
* **What to Do?** Combine Python, SQL, and visualization to tell a full data story.

# Way to check if data is analysed properly

**✅ 1. Data Cleaning & Preprocessing**

✔ **Check for missing values**

* Have you identified and handled missing data (e.g., filled, removed, or flagged)?  
  ✔ **Check for duplicates**
* Have you removed or accounted for duplicate entries?  
  ✔ **Check for inconsistencies**
* Are date formats, text fields, and numerical values standardized?  
  ✔ **Check for outliers**
* Have you used visualization (box plots, histograms) to detect outliers?

🔹 **Test Yourself**

* If you run .info() or .describe() on a Pandas DataFrame, do the columns match your expectations?
* Does df.isnull().sum() confirm that missing values are properly handled?

**✅ 2. Data Exploration & Analysis**

✔ **Ask meaningful questions**

* Have you defined specific questions you want to answer with the data?  
  ✔ **Summarize statistics**
* Do average, median, min, and max values make sense?  
  ✔ **Group & filter the data**
* Have you broken down data by categories (e.g., sales per region, age groups)?

🔹 **Test Yourself**

* If analyzing sales data, does grouping by month show seasonal trends?
* If working with employee data, does filtering by department give logical results?

**✅ 3. Data Visualization & Insights**

✔ **Choose the right chart types**

* Line chart for trends
* Bar chart for comparisons
* Scatter plot for relationships  
  ✔ **Interpret patterns correctly**
* Are the trends and relationships meaningful?  
  ✔ **Check for misleading visuals**
* Are the scales consistent? Do the axis labels make sense?

🔹 **Test Yourself**

* If you remove outliers, does the trend still hold?
* Does your visualization answer a question clearly?

**✅ 4. Data Validation & Accuracy**

✔ **Cross-check results**

* If SQL returns a total sum, does it match Pandas calculations?  
  ✔ **Test with a small dataset first**
* If you manually check 10 rows, do the results match your analysis?  
  ✔ **Compare with external sources**
* Do your findings align with real-world data (e.g., stock market trends)?

🔹 **Test Yourself**

* Run a SQL query like SELECT COUNT(\*) FROM table; to confirm row counts match Pandas.
* Do multiple aggregation methods (sum, mean, median) give consistent results?

**✅ 5. Business/Application Perspective**

✔ **Can your insights drive decisions?**

* If working with sales data, can you identify top products?  
  ✔ **Is your analysis reproducible?**
* Can someone else follow your code and get the same results?  
  ✔ **Document your findings**
* Have you summarized key insights in a report or dashboard?

🔹 **Test Yourself**

* If a manager asked, "What should we do with this data?" could you answer?
* If your code runs again on new data, does it still work correctly?

**🚀 Final Checklist**

☑ **Run sanity checks** (Are values logical?)  
☑ **Use visualizations to confirm patterns**  
☑ **Compare SQL & Python results for consistency**  
☑ **Make sure insights answer a clear question**  
☑ **Write a short report summarizing findings**

If you can confidently answer **YES** to these checks, you've properly analyzed and manipulated the data! 🎯