**Big Data Science**

For each of the following projects, you should make a new GitHub repository with an appropriate name (relevant to the dataset you chose to analyze). The repositories should all contain a “README.md” file, the required Python/Jupyter/Power BI files, and the dataset files.

**Resources:**

* <https://mlcourse.ai/book/index.html>
* <https://www.kaggle.com/learn/data-cleaning>
* <https://www.kaggle.com/learn/pandas>
* <https://www.kaggle.com/learn/data-visualization>
* <https://www.kaggle.com/learn/intro-to-sql>
* <https://jakevdp.github.io/PythonDataScienceHandbook/04.00-introduction-to-matplotlib.html>
* <https://www.datacamp.com/cheat-sheet/descriptive-statistics-cheat-sheet>
* <https://www.datacamp.com/cheat-sheet/data-viz-cheat-sheet>
* <https://www.datacamp.com/cheat-sheet/power-bi-cheat-sheet>
* <https://www.datacamp.com/cheat-sheet/sql-basics-cheat-sheet>
* <https://www.datacamp.com/cheat-sheet/pandas-cheat-sheet-for-data-science-in-python>
* <https://pynative.com/python-sqlite/>
* [Data Science Lecture Notes](https://drive.google.com/file/d/1wXutrlH2xtFQj-dTPO-Sb2OroJk31jlo/view?usp=drive_link)
* [AI Lecture Notes](https://drive.google.com/file/d/1rKTIdNfJCTALl21tmRRtuGwUbOxOvKuT/view?usp=drive_link)
* [Statistics Lecture Notes](https://drive.google.com/file/d/1fykH1FS1IreFsG9vkOblrITxwhWGFbnD/view?usp=drive_link)
* <https://deepnote.com/>
* <https://readme.so/>

**Project 1: Introduction to Pandas and Data Manipulation**

* **Objective**: Learn basic data manipulation using Pandas.
* **Key Skills**: Reading data, data cleaning, basic operations.
* **Python Files**:
  + data\_cleaner.py: Clean and preprocess data.
    - Save as DATASETNAME\_clean.EXT; e.g., covid19\_clean.csv.
  + main.py: Load data from CSV and perform simple statistical analysis.
* **Data Set**: Any public dataset like [COVID-19 data](https://www.kaggle.com/datasets/imdevskp/corona-virus-report), [world population data](https://www.kaggle.com/datasets/iamsouravbanerjee/world-population-dataset), etc.
* **Tasks**:
  + Read data from a CSV file.
  + Perform basic data cleaning (handling missing values, incorrect types, etc.).
  + Compute basic summary statistics (mean, median, mode, standard deviation).
* **Outcome**: A report summarizing key statistics of the dataset ([Markdown](https://enterprise.github.com/downloads/en/markdown-cheatsheet.pdf), README.md).

**Project 2: Data Visualization with Matplotlib**

* **Objective**: Learn to create basic plots and visualizations.
* **Key Skills**: Data visualization, basic plotting.
* **Python Files**:
  + main.ipynb: [Jupyter Notebook](https://jupyter.org/) for generating plots.
* **Data Set**: [Stock market data](https://www.kaggle.com/datasets/camnugent/sandp500) or any [time-series data](https://data.world/datasets/time-series).
* **Tasks**:
  + Create line plots to visualize stock trends.
  + Generate bar charts to compare different stocks.
  + Use scatter plots to identify correlations.
* **Outcome**: A collection of plots/statistics that tells a story about the stock market trends.

**Project 3: Database Interaction with SQLite**

* **Objective**: Learn to interact with databases using [SQLite](https://docs.python.org/3/library/sqlite3.html).
* **Key Skills**: SQL queries, database connection.
* **Python Files**:
  + database\_connector.py: Connect to [SQLite database](https://www.sqlitetutorial.net/sqlite-sample-database/), any preprocessing needed.
  + main.py: Execute SQL queries and fetch results, perform analysis.
* **Data Set**: Use a sample [e-commerce database](https://www.sqlitetutorial.net/wp-content/uploads/2018/03/chinook.zip).
* **Tasks**:
  + Create and populate tables with sample data.
  + Perform basic SQL [CRUD](https://medium.com/analytics-vidhya/sqlite-database-crud-operations-using-python-3774929eb799) queries (SELECT, JOIN, GROUP BY, INSERT, UPDATE, DELETE).
  + Fetch data into Pandas for further analysis.
* **Outcome**: A report of key insights derived from the database, like sales trends.

**Project 4: Advanced Data Analysis with Pandas**

* **Objective**: Perform advanced data analysis.
* **Key Skills**: Data aggregation, merging datasets, complex queries.
* **Python Files**:
  + main.ipynb: Perform complex data operations.
* **Data Set**: Use datasets from Kaggle, like a dataset on [global education statistics](https://www.kaggle.com/datasets/andrewmvd/global-education-statistics).
* **Tasks**:
  + Merge multiple datasets.
  + Perform complex data transformations.
  + Aggregate data to find trends and patterns.
* **Outcome**: An in-depth analysis report on global education trends.

**Project 5: Integrating with Power BI**

* **Objective**: Learn to use Power BI for advanced data visualization.
* **Key Skills**: [Power BI](https://powerbi.microsoft.com/en-us/desktop/) [online version [here](https://app.powerbi.com/)], data storytelling.
* **Tasks**:
  + Connect Power BI to a dataset (use an advanced dataset like [world economic indicators](https://www.kaggle.com/datasets/mittvin/world-economic-indicators-1960-2022-dataset)).
  + Create interactive dashboards in Power BI.
  + Tell a story through data using various visual elements in Power BI.
* **Outcome**: An interactive dashboard that provides insightful visualizations of the data.

**Data Story Example Suggestions**

**Project 1: Pandas and Basic Data Manipulation - Story: The Global Impact of COVID-19**

* **Description**: Analyze COVID-19 data to understand its global impact.
* **Tasks**:
  + Track the spread of the virus across different countries.
  + Compare infection rates and recovery rates over time.
  + Identify the most affected regions and times.
* **Storytelling Goal**: Highlight how the pandemic evolved globally, showcasing the power of data manipulation in understanding real-world issues.

**Project 2: Data Visualization with Matplotlib - Story: Analyzing Market Trends**

* **Description**: Visualize stock market trends to understand market behaviors.
* **Tasks**:
  + Show historical trends of different stocks.
  + Compare the performance of tech vs. non-tech stocks.
  + Highlight periods of significant volatility.
* **Storytelling Goal**: Demonstrate how visual data representation can reveal insights into market dynamics and investor behavior.

**Project 3: Database Interaction with SQLite - Story: E-commerce Sales Analysis**

* **Description**: Analyze an e-commerce database to uncover sales patterns.
* **Tasks**:
  + Identify the best-selling products.
  + Determine peak sales periods.
  + Analyze customer buying patterns.
* **Storytelling Goal**: Use database querying to reveal insights into consumer behavior and sales strategies in e-commerce.

**Project 4: Advanced Data Analysis with Pandas - Story: Global Education Insights**

* **Description**: Dive into global education statistics to identify trends and disparities.
* **Tasks**:
  + Compare literacy rates across different countries and over time.
  + Analyze the impact of GDP on education.
  + Identify gender disparities in education.
* **Storytelling Goal**: Showcase the ability to draw meaningful conclusions from complex, multi-dimensional data, emphasizing global education issues.

**Project 5: Integrating with Power BI - Story: World Economic Indicators Dashboard**

* **Description**: Create a dashboard to explore and present global economic indicators.
* **Tasks**:
  + Visualize GDP growth, unemployment rates, and inflation across countries.
  + Create interactive elements to explore different time periods and regions.
  + Highlight correlations between different economic indicators.
* **Storytelling Goal**: Demonstrate the power of interactive dashboards in making complex economic data accessible and understandable.