Data Science Intern - Assignment

As part of our recruitment process for the **Data Science Intern – Automated CNC Costing Engine (ACE)** role, we invite you to complete the following assignment. This is designed to evaluate your ability to work with real-world data, build machine learning models, and draw meaningful insights from manufacturing-oriented datasets.

We welcome applicants from both Computer Science (related streams) and Mechanical/Manufacturing backgrounds. The goal is to test core data science skills while offering opportunities to explore mechanical reasoning for those interested.

Objective:

Using the provided dataset of CNC-machined parts, your task is to:

- Clean and preprocess the data
- Engineer relevant features
- Train a simple ML model to predict machining **cost** or **cycle time**
- Visualize insights and interpret model performance

Assignment Tasks :

Part 1: Data Gathering

- **Choose a relevant source**: Scrape data related to **CNC machining**, **costing**, or **cycle times**.
    - Use **Python libraries** like `BeautifulSoup`, `Scrapy`, or `Selenium` to scrape data from tables, charts, or other structured information.
    - Collect at least 100–200 data points relevant to machining costs, part dimensions, material types, or cycle times.
    - Get data like, material, dimensions, estimated time, material, estimated cost.

    - Sites like CNCZone, Practical Machinist, GrabCAD, TraceParts, Fictiv, Xometry, Alibaba - CNC listings, can be used for data gathering. Out of the box thinking is appreciated.

- **Important Notes**:
    - Ensure you follow ethical scraping practices. Always check if a website has any restrictions in place (robots.txt). Avoid scraping login- gated sites.
    - Your scraped data should contain multiple variables related to the machining process (e.g., volume, feature count, material, cost, cycle time).
    - Manual + automated scraping is okay

Part 2: Data Understanding & Cleaning

- Load the dataset and explore the columns
- Identify and handle missing values, duplicates, and data type.
- Visualize the data too, with the understanding of what kind of pattern does the data follow.
- Explain your cleaning choices clearly

Part 3: Feature Engineering

- Create **at least two new features** based on the existing columns
  (e.g., cost per mm^3, surface-area-to-volume ratio, tool density)
- Explain the reasoning behind each feature
- Suggest any machining-related insights (e.g., which features may affect cost)

Part 4: Visualization & Model Training

- Use `matplotlib`, `seaborn`, or `plotly` to visualize:
    - Cost/cycle time vs. material, volume, or feature count

    - Distributions and correlations between numeric features
    - Key trends or insights (2–3 observations)
- Train a model (e.g., Linear, Random Forest) to predict:
    - Option A: `Quoted_Cost` or Option B: `Cycle_Time_min`
- Split data into train/test sets and evaluate using:
    - MAE, RMSE
    - Scatter plot of predicted vs. actual
    - Feature importance (basic interpretation)

Part 5: Final Analysis & Reflection

- What worked well in your model?
- What challenges did you face?
- How would you improve predictions with more data or domain knowledge?

Optional Reflection:

- Share what you found intriguing or challenging about manufacturing data.
- Share what additional data/features would improve model accuracy.

Submission Guidelines

Please submit:

- A **Jupyter Notebook** named `CNC_Cost_Estimation.ipynb`
- A **short README** (`README.md` or `.pdf`) explaining:

    - Your approach
    - Summary of results
    - Key insights

Submit your work as a ZIP file or via a GitHub repository link.

Allowed Tools

- Python (NumPy, pandas, scikit-learn, matplotlib/seaborn)
- Jupyter Notebook
- Any IDE or notebook environment of your choice

All the best 👍