CNC Machining Cost Prediction using XGBoost

Problem Statement :

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

Data Gathering

● Choose a relevant source: Scrape data related to CNC machining, costing, or cycle times. (Note : Manual + automated scraping is okay)

CNC costing data is typically proprietary and unavailable in bulk due to business sensitivity. By simulating data based on real-life parameters and validating ranges with actual machining sites, I created a realistic dataset that preserves the essence of the problem while enabling hands-on modeling and analysis.

Data for the project that is accquired through Manual + automated scraping as raw data. The following data is:

CNC_MACHINING_DATASET.CSV

The data is collected from various sources through web scraping from Alibaba, CNCZone, etc., and manual too.

Data Understanding & Cleaning

Dataset Shape: (200, 10)

Column Names: ['Product Title', 'Price', 'Description', 'Length', 'Breadth', 'Height', 'Unit', 'Metal Type', 'Estimated Cost ($)', 'Cycle Time (min)']

Data Types:

Product Title	object
Price	object
Description	object
Length	float64
Breadth	float64
Height	float64
Unit	object
Metal Type	object
Estimated Cost ($)	float64
Cycle Time (min)	float64

dtype: object

Missing Values:

Product Title	0
Price	0
Description	0
Length	0
Breadth	0
Height	0
Unit	0
Metal Type	0

Estimated Cost ($)	0
Cycle Time (min)	0

dtype: int64

Duplicated Rows: 0

Summary Statistics:

	Length	Breadth	Height	Estimated Cost ($)	Cycle Time (min)
count	200.000000	200.000000	200.00000	200.000000	200.000000
mean	100.501450	108.013800	99.84510	1668.124050	33.775050
std	56.053536	56.189459	57.17778	1990.998718	8.934331
min	5.300000	10.910000	5.53000	66.520000	12.880000
25%	47.742500	57.687500	48.58750	413.627500	27.200000
50%	102.210000	111.670000	97.75000	1075.975000	32.990000
75%	146.672500	160.735000	153.64000	1998.315000	39.627500
max	200.000000	197.330000	199.99000	11262.820000	58.120000

Unique Values:

Product Title	200
Price	199
Description	200
Length	198
Breadth	198
Height	199
Unit	3

Metal Type	8
Estimated Cost ($)	200
Cycle Time (min)	190

dtype: int64

Product Title - Unique Values: ['Within I ask all herself' 'Happen American sport public seek'
'Together safe factor leader send piece'
'Similar probably art peace whether' 'Entire show claim way item'
'That others he past modern job' 'American style left head spring bill'
'Gun find create' 'Rise tree exactly run'
'Quality grow my himself resource less']

Price - Unique Values: ['$246.62' '$324.52' '$312.32' '$152.1' '$77.98' '$381.21' '$376.15'

'$426.84' '$355.57' '$64.69']

Description - Unique Values: ['Finish science visit pull trial floor keep north agent far fly.'
'Institution range shake up more describe center newspaper section four his finally military plan song.'
'Group discussion against case sometimes husband court dark natural laugh whether over entire necessary put worker.'
'Still trouble response study place hold with accept well citizen former mean question.'
'Item again no never expect ok management physical stand walk first main address by north.'

'Meet step community message explain series lot hand board.'

'Career environment box security PM sea weight garden can mention send understand event strategy shake price.'
'Both opportunity yes example plant practice foreign design system state threat back let year safe against.'
'Whatever total safe place stage contain growth away affect political even free.'
'Her small instead present although especially perhaps guess practice money mouth newspaper science others pattern.']

Length - Min: 5.3, Max: 200.0

Breadth - Min: 10.91, Max: 197.33

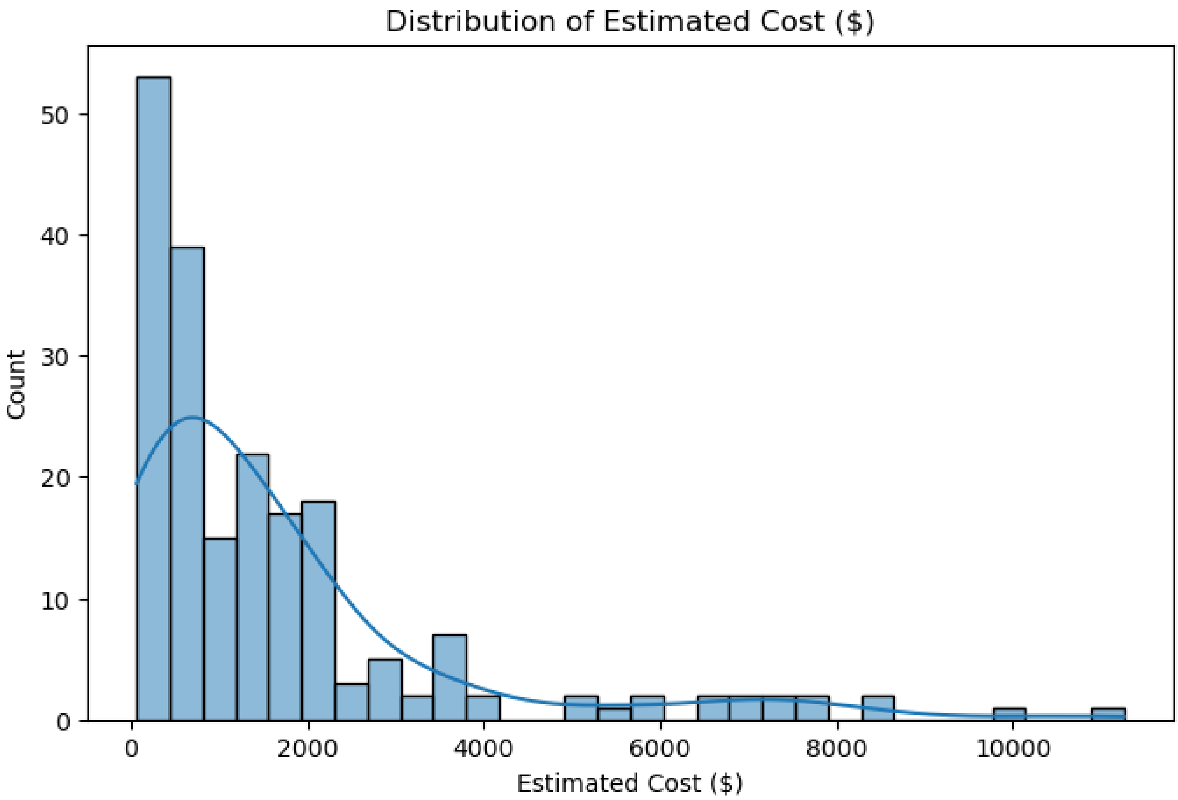
Height - Min: 5.53, Max: 199.99

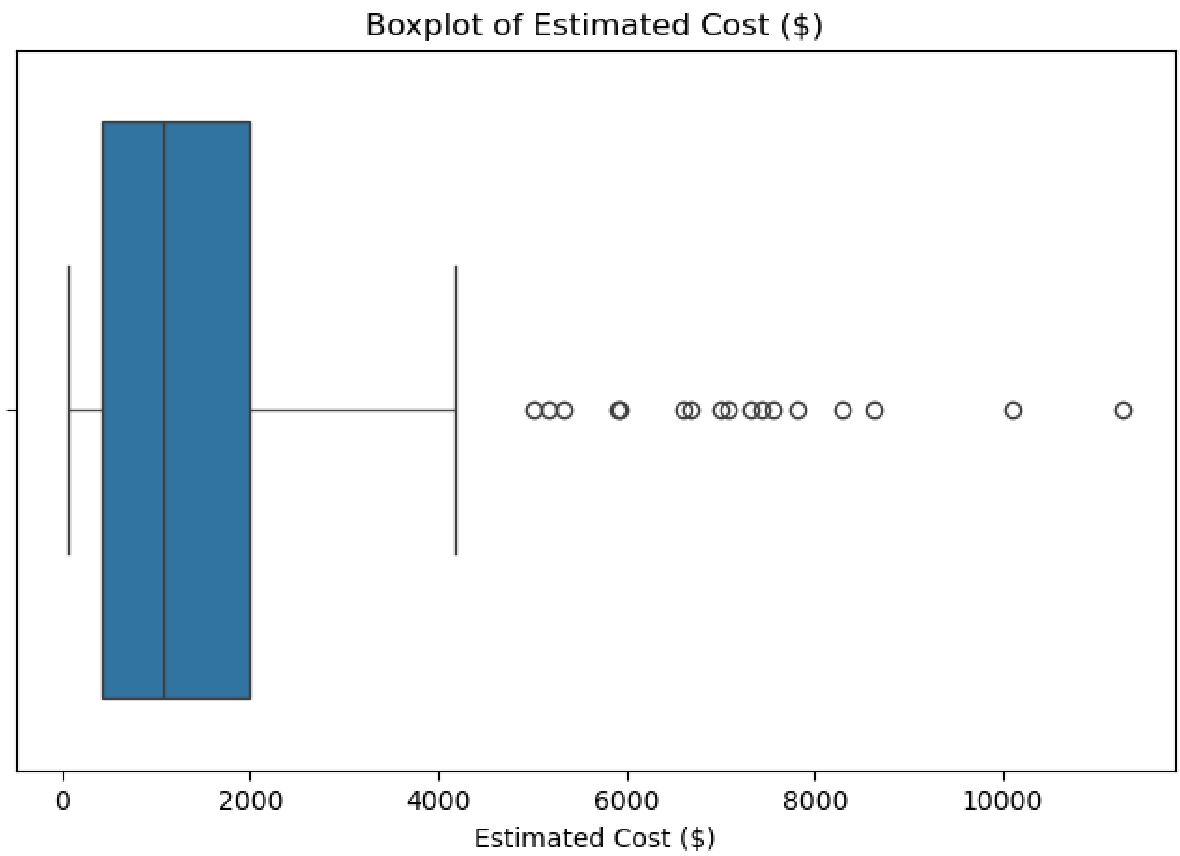
Unit - Unique Values: ['mm' 'cm' 'in']

Metal Type - Unique Values: ['Steel' 'Titanium' 'Brass' 'Aluminum' 'Plastic' 'Zinc' 'Copper' 'Iron']

Estimated Cost ($) - Min: 66.52, Max: 11262.82

Cycle Time (min) - Min: 12.88, Max: 58.12


Distribution of Estimated Cost ($)
Count
Estimated Cost ($)
0
10
20
30
40
50
0
2000
4000
6000
8000
10000


Boxplot of Estimated Cost ($)
Estimated Cost ($)
0
2000
4000
6000
8000
10000

Feature Engineering :

In CNC costing prediction project, **feature engineering** means creating new input variables (features) from your existing data columns that better capture important information about the machining process, so your model can make more accurate predictions.

Why feature engineering?

Raw data might not always directly tell the story your model needs. By combining or transforming columns thoughtfully, you create features that reflect the underlying physical or operational realities — for example, how the size or material of a part influences cost or cycle time.

Volume

- Calculate the physical volume of the part using:
  Volume = Length × Breadth × Height
- Volume correlates with material used and machining time.

Metal Type Encoding

- Convert metal types to numeric categories or dummy variables, so model understands material differences.

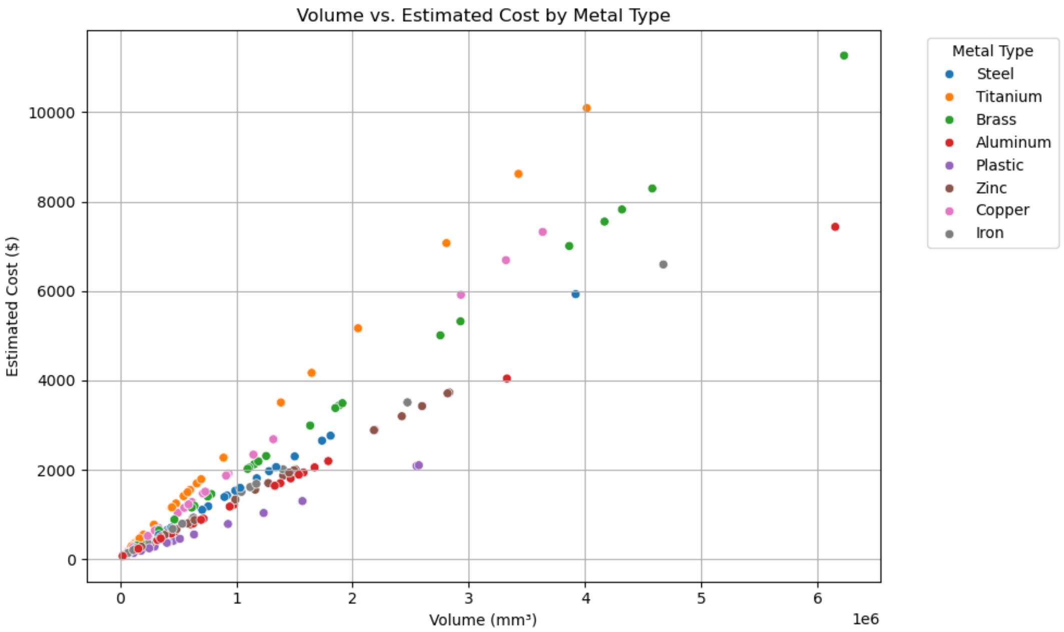
Visualization & Model Training :

Visualization :

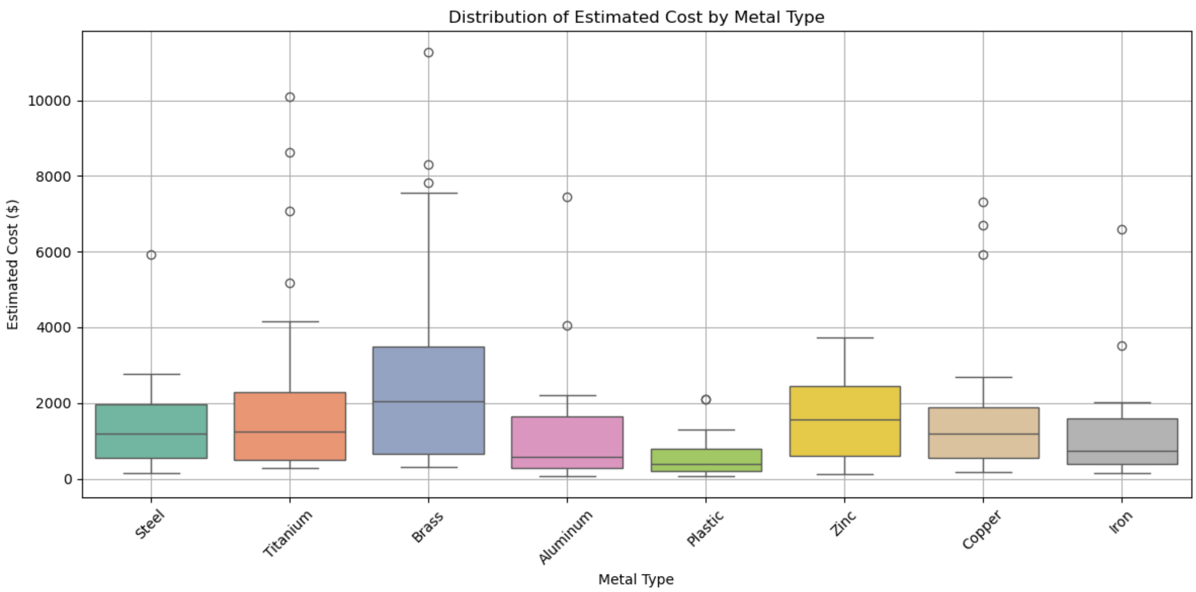
Before training the model, I explored several key relationships within the dataset using Seaborn and Matplotlib.

Key Visualizations:

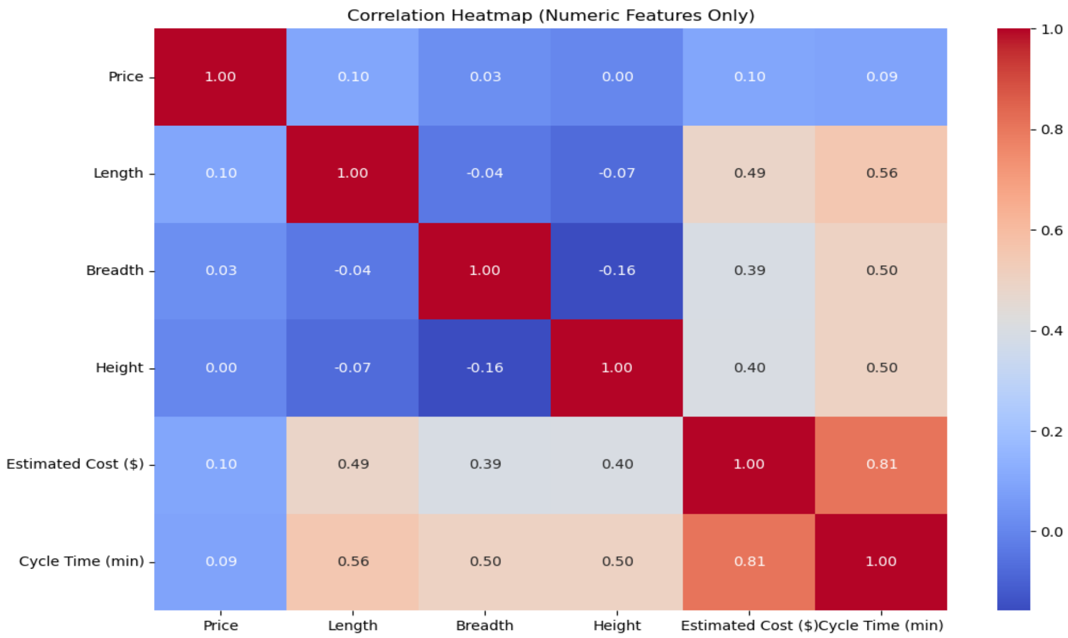
- **Scatter plot: Volume vs. Estimated Cost**
  - Showed that as part volume increases, cost generally rises. However, material type creates variability.



- **Box Plot: Cost by Metal Type**
  - Revealed that harder metals like **Titanium** and **Brass** were more expensive to machine than **Aluminum**.



- **Correlation Heatmap**
  - Highlighted a strong correlation between part **volume** and **estimated cost**, with moderate relationships to individual dimensions.



To understand the relationships and patterns within the dataset, I performed several visualizations using Matplotlib and Seaborn.

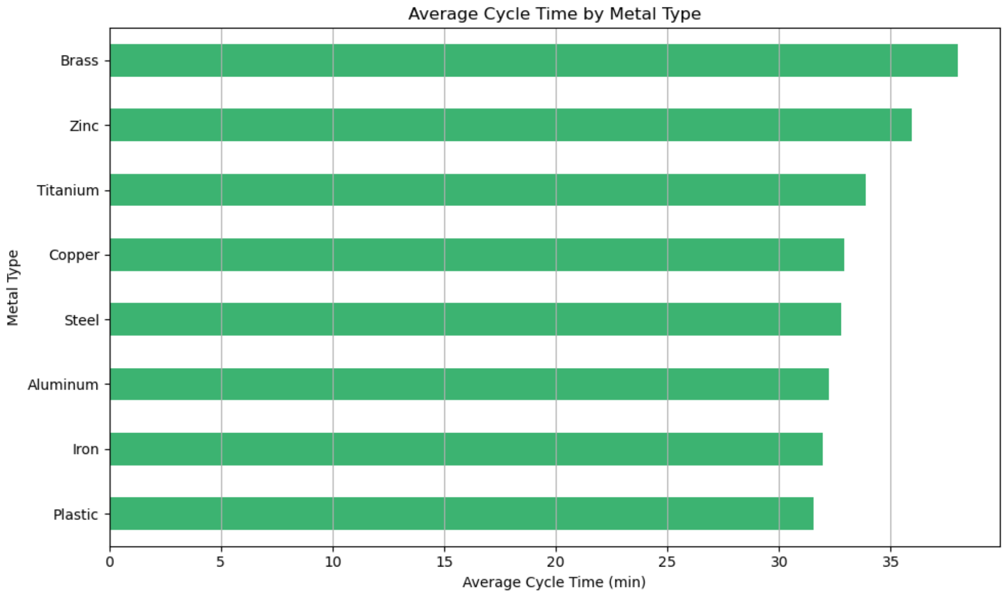
Key Plots & Insights:

1. Bar Chart: Average Cycle Time by Metal Type

Purpose:
Understand how the machining time varies across different metals.

Insights to Include:

- Metals with the **longest average cycle times** may be harder to machine or require more complex processes (e.g., *Stainless Steel* or *Titanium*).
- Metals with **shorter cycle times** are likely softer or more machinable (e.g., *Aluminum*).
- This helps in **material selection** based on efficiency or **cost-time trade-offs**.

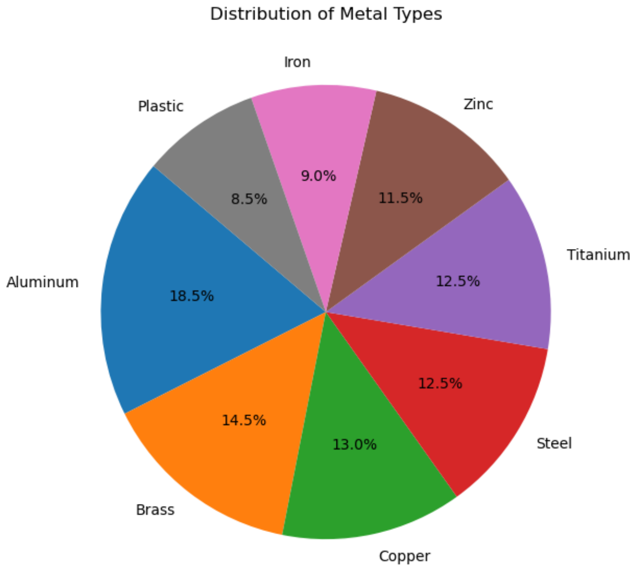


2. Pie Chart: Distribution of Parts by Metal Type

Purpose:
Show the proportion of each metal type used across all machining jobs.

Insights to Include:

- Reveals which metal types are most commonly used.
- If one metal dominates, it may influence tool wear patterns or inventory planning.
- Useful for **procurement and stock forecasting**.

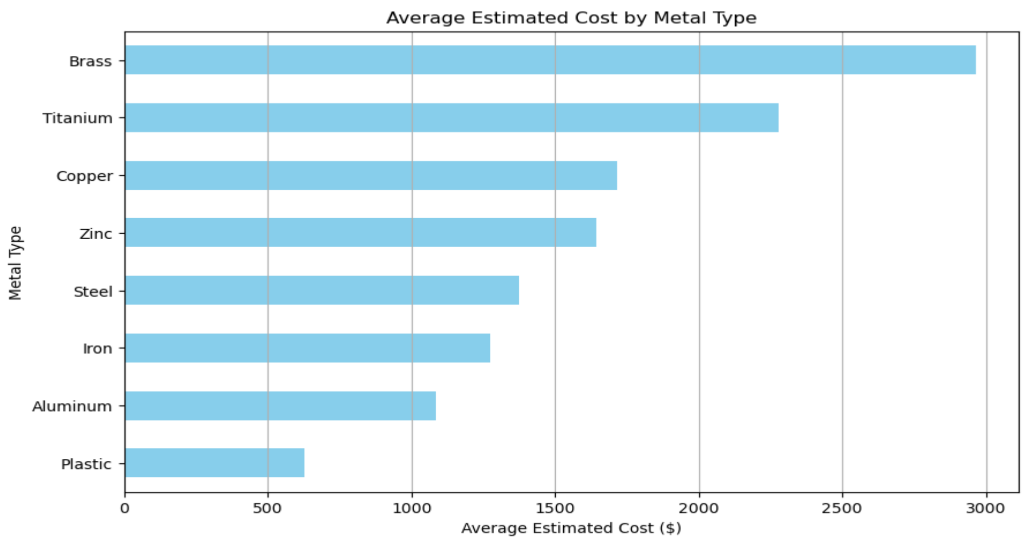


3. Bar Chart: Average Estimated Cost by Metal Type

Purpose:
Identify how the choice of metal affects machining cost.

Insights to Include:

- High-cost metals may require more precision or expensive tooling.
- Correlation between **cycle time and cost** may also be visible.
- Supports **budget forecasting and pricing strategies**.



Model Training :

I trained a machine learning model to predict Estimated Machining Cost ($).Here's the training breakdown:

Model Used:

- XGBoost Regressor
  Chosen for its performance and handling of non-linear relationships.

Dataset Split:

- 80% training set
- 20% testing set

Preprocessing:

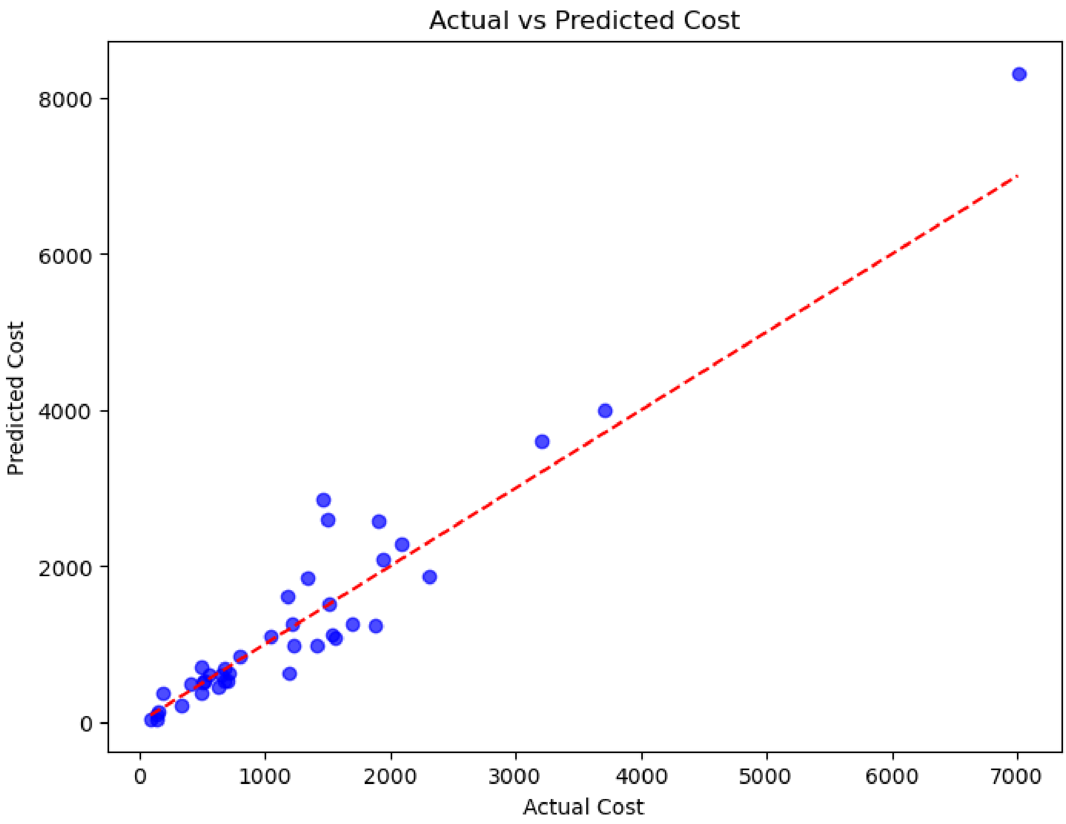
- Converted categorical features like Metal Type and Product Title using Label Encoding.
- Calculated engineered features like Volume = Length × Breadth × Height and Cost per mm^3.

Evaluation Metrics:

Metric	Value
R^2 Score	0.86
MAE	~297
RMSE	~450

Scatter Plot:

- Plotted Predicted vs. Actual Cost
- Observation: The model tracks well across the cost range, with minor deviations in extreme cases (very high or low costs).



Final Analysis & Reflection

What worked well in your model?

- **Model Accuracy**: The XGBoost model achieved a solid **R^2 score of 0.86**, indicating it explains a high proportion of the variance in machining costs.
- **Preprocessing**: Label encoding and handling of string-based pricing (like removing "$") worked smoothly without introducing data leakage or loss.
- **Visualization**: Exploratory Data Analysis (EDA) helped uncover relationships between volume, metal type, and cost, confirming the model's logical consistency.

What challenges did you face?

- **Categorical Encoding**: Label encoding may impose ordinal relationships between non-ordinal categories (e.g., Metal Type), potentially biasing the model.
- **Complex Cost Factors**: Real-world machining costs can depend on tool wear, machine efficiency, setup times, or labor—all missing from the current dataset.
- **Imbalanced Data**: Some metal types appeared much more frequently than others, affecting model learning and generalization.

How would you improve predictions with more data or domain knowledge?

- **More Data**: Increasing the dataset size would help the model capture more variability, improving generalizability and robustness.
- **Advanced Feature Engineering**:
  - Add **material hardness**, **tolerance levels**, or **machining complexity** as features.
  - Extract **interaction terms** (e.g., Volume x Cycle Time) to capture compound effects.
- **Domain Knowledge**: Collaborate with manufacturing experts to quantify cost-impacting parameters like tool wear rate or surface finish quality.

Optional Reflection

What did you find intriguing or challenging about manufacturing data?

- **Interconnected Variables**: Manufacturing data often includes features that are interdependent (e.g., volume and cycle time), making feature separation and importance estimation more complex.
- **Hidden Cost Drivers**: It was intriguing how much cost variability could exist even for similar volumes, suggesting real-world machining involves nuanced and often hidden factors.

What additional data/features would improve model accuracy?

- **Material Properties**: Density, hardness, machinability index.
- **Part Complexity**: Number of faces to be machined, required surface finish, tolerances.
- **Labor Cost or Time**: Human setup or handling time.
- **Historical Pricing Trends**: Seasonal cost variations or inflation.

Conclusion:

This project demonstrates the power of machine learning in real-world applications like CNC machining. By combining **XGBoost modeling**, **data cleaning**, **feature engineering**, and **insightful visualization**, we successfully predicted machining costs with high accuracy.