Design of Experiment (DoE)

Introduction to Design of Experiments (DoE): Design of Experiments (DoE) is a systematic and rigorous approach to investigating processes or systems. Developed by Sir Ronald A. Fisher in the 1930s, DoE provides a framework for conducting experiments efficiently by varying multiple input variables simultaneously. Instead of using a trial-and-error approach, DoE facilitates understanding the effect of independent variables on a response variable and, potentially, the interactions between them. This methodology is not just limited to the field of agriculture, where it originated, but has vast applications across industries.

1. Fundamental Principles: Three core principles underpin DoE are:

- *Randomization:* It ensures that experiments are free from bias by randomly assigning experimental units to different treatments. This helps in managing unknown or unexpected sources of variability.

- *Replication:* It involves repeating the experiments under the same conditions to estimate experimental error and enhance the reliability of the results.

- *Blocking:* If an external source of variability is not of primary interest but might influence the response, the experiments are grouped (or "blocked") based on this source to account for its potential effect.

1. Types of Experimental Designs: Depending on the objective and constraints, various types of designs can be adopted:

- *Factorial Designs:* These are the most common, allowing for studying multiple factors and their interactions. In a full factorial design, all possible combinations of levels of all factors are investigated.

- *Fractional Factorial Designs:* When full factorial designs become impractical due to the number of experiments required, a fraction can be chosen systematically, allowing for insights into the main effects and some interactions.

- *Response Surface Methodology (RSM):* This technique is used for modeling and analysis when the main goal is optimization. It explores relationships between several explanatory variables and one or more response variables.

1. Analyzing Results and Model Building: Once the experimental data is obtained, statistical methods, especially regression analysis, are employed to develop empirical models. These models capture the relationship between independent factors and the response. The significance of the factors, the presence of interactions, and the model's goodness of fit are evaluated using various statistical tests and metrics. The resulting models are not just descriptive but also predictive, facilitating decision-making.

2. Applications of DoE: DoE has a broad spectrum of applications:

- In *manufacturing*, it can optimize processes, improve quality, and reduce costs.
- In *agriculture*, it assists in maximizing yields or optimizing the use of fertilizers.
- The *pharmaceutical industry* employs DoE for drug formulation and testing.
- In *marketing*, it aids in assessing the impact of different strategies on sales or customer engagement.

Recently, it's also seen applications in *machine learning* for hyperparameter tuning and model selection.

In essence, DoE offers a structured method to gain insights and make informed decisions based on empirical evidence, making it an indispensable tool across diverse domains.

Response Surface Methodology: Froth flotation example

As discussed above Response Surface Methodology (RSM) is a collection of mathematical and statistical techniques that help model and analyze problems in which the response of interest is influenced by several variables. Its main aim is to optimize the response.

We illustrate RSM using froth flotation data for copper sulfide ore. The inputs are grind size, pH, pulp density, and collector dosage, and the response variable is copper flotation recovery (Recovery_Cu**)**.

In [1]:

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
from statsmodels.formula.api import ols
import matplotlib.pyplot as plt

# The data
# Convert the data to a pandas DataFrame
data = {
    'Grind_P80': [105, 105, 157.5, 105, 210, 210, 210, 210, 105, 210, 210, 105, 105, 157.5, 210, 105, 157.5, 105, 210],
    'pH': [11.45, 7.25, 9.35, 7.25, 7.25, 7.25, 11.45, 7.25, 7.25, 7.25, 11.45, 11.45, 11.45, 9.35, 11.45, 11.45, 9.35, 7.25, 11
    'PulpDensity': [25, 25, 30, 25, 35, 25, 25, 35, 35, 25, 35, 25, 35, 30, 25, 35, 30, 35, 35],
    'Collector_gt': [20, 20, 15, 10, 20, 20, 10, 10, 20, 10, 10, 10, 20, 15, 20, 10, 15, 10, 20],
    'Recovery_Cu': [84.5, 82.4, 89.5, 74.4, 87.2, 76.7, 84.8, 86.2, 82.3, 77.4, 85.1, 85.8, 87.3, 88.5, 80.9, 88.7, 87.3, 89.3,
}

df = pd.DataFrame(data)

# Fit a quadratic model
formula = "Recovery_Cu ~ pH + Collector_gt + I(pH**2) + I(Collector_gt**2) + pH:Collector_gt"
model = ols(formula, data=df).fit()

# Display model statistics
print(model.summary())

# Predict for a grid of values to plot
x = np.linspace(df['pH'].min(), df['pH'].max(), 100)
y = np.linspace(df['Collector_gt'].min(), df['Collector_gt'].max(), 100)
x, y = np.meshgrid(x, y)
z = model.predict(pd.DataFrame({'pH': x.ravel(), 'Collector_gt': y.ravel()})).values.reshape(x.shape)

# 3D Surface plot
fig = plt.figure(figsize=(10, 6))
ax = fig.add_subplot(111, projection='3d')
ax.plot_surface(x, y, z, cmap='viridis')
ax.set_xlabel('pH')
ax.set_ylabel('Collector_gt')
ax.set_zlabel('Recovery_Cu')
plt.title('3D Surface Plot')
plt.show()

# 2D Contour plot
fig, ax = plt.subplots(figsize=(8, 6))
cp = ax.contourf(x, y, z, cmap='viridis')
fig.colorbar(cp, label='Recovery_Cu')
ax.set_xlabel('pH')
ax.set_ylabel('Collector_gt')
plt.title('2D Contour Plot')
plt.show()
```

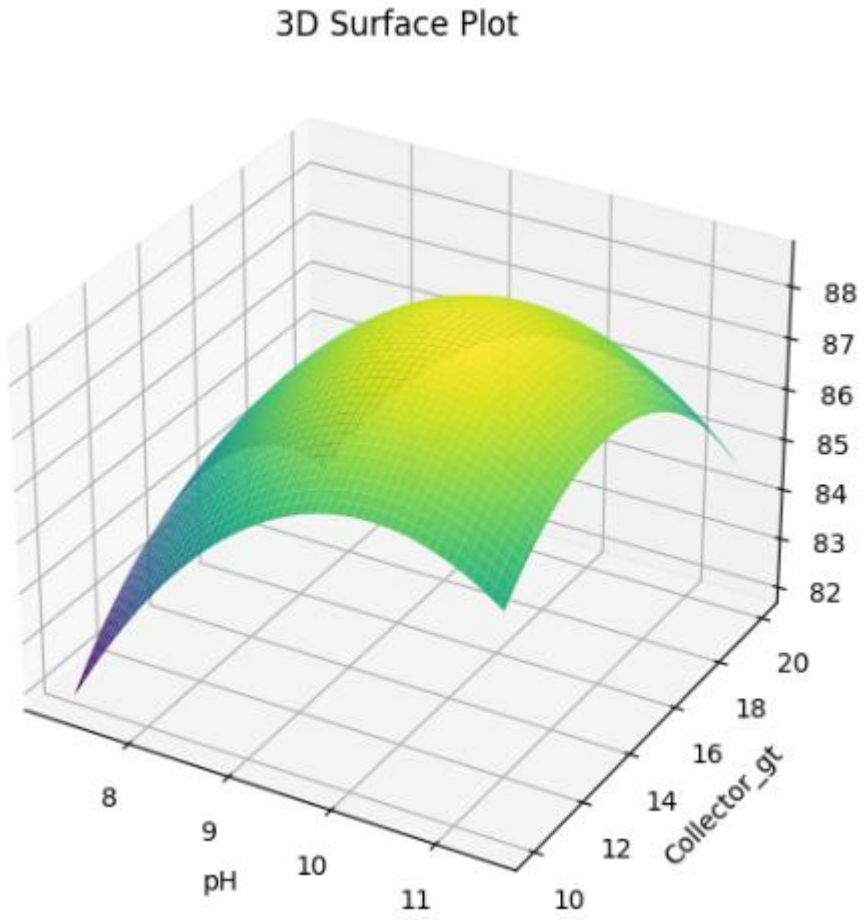
OLS Regression Results

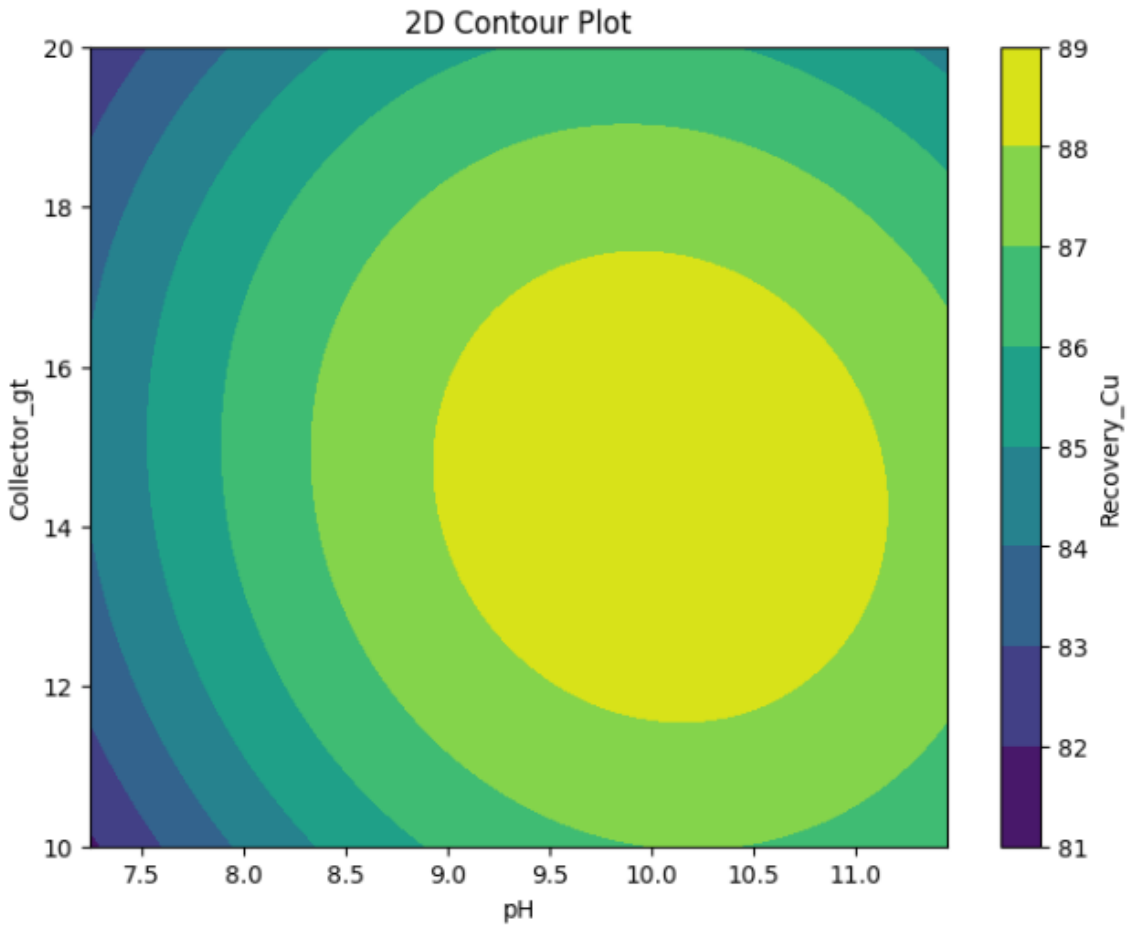
Dep. Variable:	Recovery_Cu	R-squared:	0.309
Model:	OLS	Adj. R-squared:	0.112
Method:	Least Squares	F-statistic:	1.568
Date:	Mon, 07 Aug 2023	Prob (F-statistic):	0.237
Time:	12:37:21	Log-Likelihood:	-51.040
No. Observations:	19	AIC:	112.1
Df Residuals:	14	BIC:	116.8
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.6182	3.736	1.236	0.237	-3.394	12.631
pH	12.5751	7.919	1.588	0.135	-4.410	29.560
Collector_gt	2.8904	5.637	0.513	0.616	-9.199	14.979
I(pH ** 2)	-0.5954	0.466	-1.277	0.222	-1.596	0.405
I(Collector_gt ** 2)	-0.0850	0.177	-0.480	0.639	-0.465	0.295
pH:Collector_gt	-0.0423	0.099	-0.429	0.674	-0.254	0.169

Omnibus:	0.255	Durbin-Watson:	2.655
Prob(Omnibus):	0.880	Jarque-Bera (JB):	0.018
Skew:	-0.054	Prob(JB):	0.991
Kurtosis:	2.893	Cond. No.	1.28e+18

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.26e-30. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.


3D Surface Plot
88
87
86
85
84
83
82
8
9
10
11
pH
10
12
14
16
18
20
Collector_gt



Explanation:

- The formula used in the regression model includes main effects (pH and Collector_gt), interaction terms (pH: Collector_gt), and squared terms (I(pH**2) and I(Collector_gt**2)) to capture curvature.
- We predicted the response (Recovery_Cu) for a grid of pH and Collector_gt values used for plotting.
- The 3D surface plot visually represents how the response changes as a function of the two factors.
- The 2D contour plot is another way to represent the same information, where contour lines represent places of equal response.

The plots give you a sense of the regions where the Recovery_Cu is maximized or minimized. The model summary will provide you with detailed statistics for each term in the model.

Reference

https://francisdakubo.com/blogs/

https://bookdown.org/apicellapv/rsm_tutorial/rsm_tutorial.html

https://www.jmp.com/en_ph/statistics-knowledge-portal/what-is-design-of-experiments.html