```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

```
df = pd.read_csv("wagih_2020_data.csv")
```

This contains datas for Binary Alloys (that's why solvent and solute columns)

```
df
```

	Solvent	Solute	Potential_ID	Alpha	Mu	Sigma	R_Squared
0	Ag	Al	1	0.0	-3	3	0.98
1	Ag	Au	2	0.8	-2	2	1.00
2	Ag	Au	3	0.0	0	2	1.00
3	Ag	Au	1	2.6	-2	5	1.00
4	Ag	Co	1	-0.2	3	12	1.00
...	...	...	...	...	...	...	...
422	Zr	Pd	1	-1.2	4	16	1.00
423	Zr	Pt	1	-1.3	3	17	1.00
424	Zr	Ta	1	1.9	-7	13	1.00
425	Zr	Ti	1	0.0	-3	12	1.00
426	Zr	W	1	2.8	1	24	1.00

427 rows × 7 columns

Next steps: Generate code with df | New interactive sheet

Segregation energy is the energy difference between a solute atom residing at a grain boundary (GB) versus residing inside the "bulk" (the middle of a crystal grain). The μ value tells you the average tendency of that solute.

If μ is Negative (e.g., -50 kJ/mol): This means the solute atom is, on average, more stable (at a lower energy) at the grain boundary. This creates a strong driving force for the solute to accumulate at the GBs. This is favorable segregation.

Example: For the Mo (solvent) - Pb (solute) alloy, μ = -52. This indicates that Lead (Pb) atoms will strongly segregate to the grain boundaries in Molybdenum (Mo).

If μ is Positive (e.g., +50 kJ/mol): This means the solute atom is, on average, less stable (at a higher energy) at the grain boundary. The solute atoms will actively avoid the GBs and prefer to stay in the bulk. This is unfavorable segregation (or depletion).

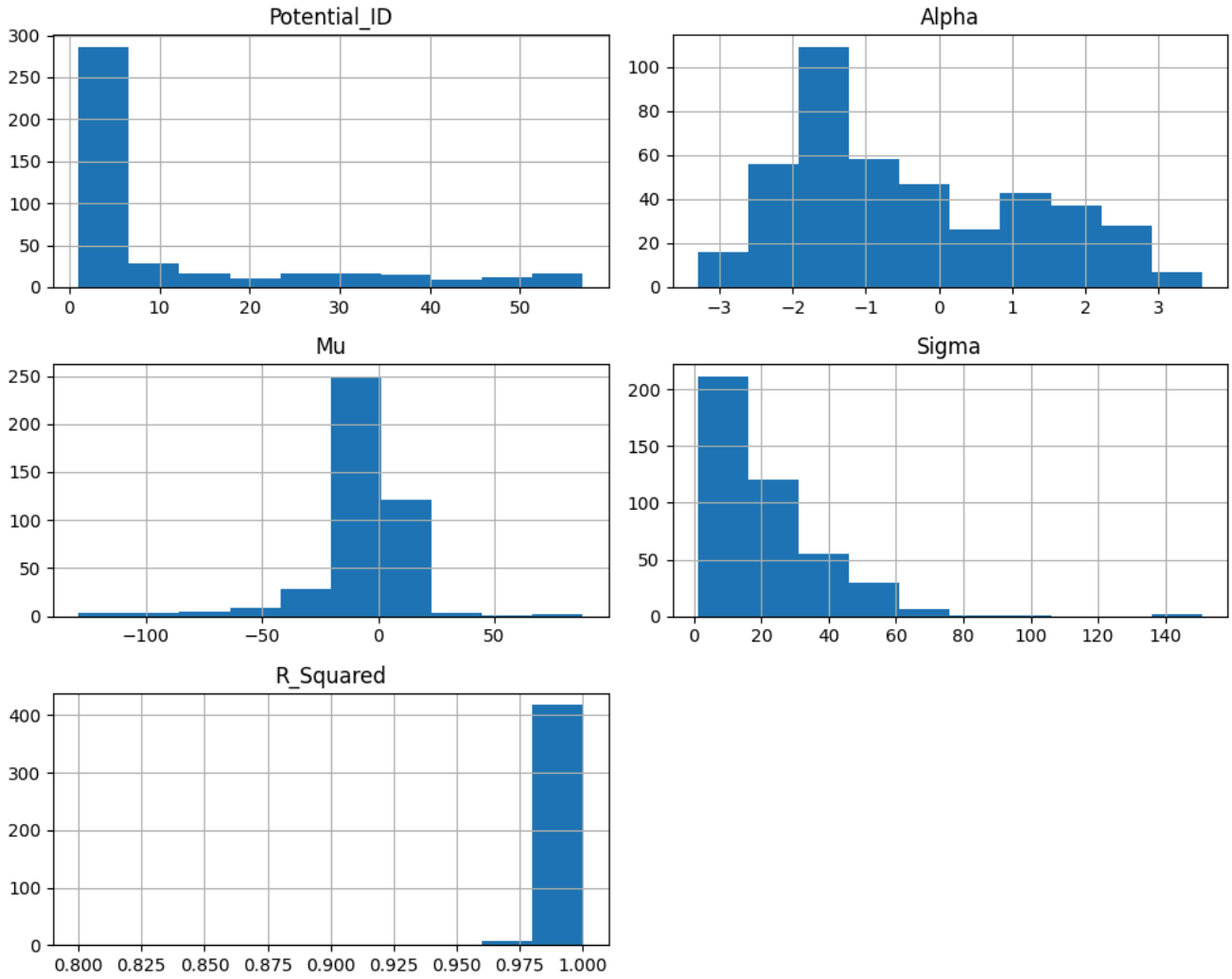
Example: For the Al (solvent) - Nb (solute) alloy, μ = 72. This indicates that Niobium (Nb) atoms will be repelled from the grain boundaries in Aluminum (Al).

If μ is Near Zero (e.g., -2 to +2 kJ/mol): This means there is, on average, no significant energy difference between the bulk and the grain boundary. There is no strong driving force for the solute to either accumulate at or avoid the GBs.

Example: For the Ag (solvent) - Au (solute) alloy, one potential gives μ = 0.

```
# Select only numerical columns for plotting
numerical_cols = df.select_dtypes(include=np.number).columns

# Plot histograms for numerical columns
df[numerical_cols].hist(figsize=(10, 8))
plt.tight_layout()
plt.show()
```




```
round((df.isnull().sum() / df.shape[0]) * 100, 2)
```

	0
Solvent	0.0
Solute	0.0
Potential_ID	0.0
Alpha	0.0
Mu	0.0
Sigma	0.0
R_Squared	0.0

dtype: float64

No missing values, most of them are unique. R^2 values for most of them is pretty high like 0.98 and 1 (A lot of them fit the data completely). So, a suggested preprocessing technique would be to remove rows with R^2 values less than 0.9

```
cat_col = ['Solvent', 'Solute']
df[cat_col].nunique()
```

	0
Solvent	19
Solute	25

dtype: int64

```
print(df.info)
```

```
<bound method DataFrame.info of     Solvent Solute  Potential_ID  Alpha  Mu  Sigma  R_Squared
0        Ag     Al             1    0.0  -3      3       0.98
```

```
1        Ag     Au            2    0.8  -2      2       1.00
2        Ag     Au            3    0.0   0      2       1.00
3        Ag     Au            1    2.6  -2      5       1.00
4        Ag     Co            1   -0.2   3     12       1.00
..      ...    ...          ...    ...  ..    ...        ...
422      Zr     Pd            1   -1.2   4     16       1.00
423      Zr     Pt            1   -1.3   3     17       1.00
424      Zr     Ta            1    1.9  -7     13       1.00
425      Zr     Ti            1    0.0  -3     12       1.00
426      Zr      W            1    2.8   1     24       1.00

[427 rows x 7 columns]>
```

```
df.isnull().sum()
```

	0
Solvent	0
Solute	0
Potential_ID	0
Alpha	0
Mu	0
Sigma	0
R_Squared	0

dtype: int64

```
df.duplicated().sum()
```

```
np.int64(0)
```

```
df.sample(100)
```

1 to 100 of 100 entries Filter

index	Solvent	Solute	Potential_ID	Alpha	Mu	Sigma	R_Squared
359	Ta	Pt	1	-2.2	-2	18	1.0
150	Cu	Ni	28	2.2	0	10	1.0
270	Ni	Fe	35	3.2	-4	12	1.0
35	Al	Cu	11	1.0	-22	22	1.0
368	Ti	Co	1	-1.3	3	25	1.0
301	Pd	Cu	2	0.8	-12	12	1.0
253	Ni	Au	2	-1.8	-8	39	1.0
149	Cu	Ni	37	1.9	1	2	1.0
393	W	Fe	34	-0.7	-7	40	1.0
299	Pd	Au	1	0.0	-4	3	1.0
74	Au	Cu	1	0.7	-1	7	1.0
100	Co	Cu	1	-3.1	1	6	1.0
198	Fe	W	34	-1.2	6	47	0.99
402	W	Re	57	-1.6	-2	28	1.0
113	Co	Ti	1	-1.8	10	33	1.0
249	Ni	Al	20	1.4	-5	11	1.0
147	Cu	Ni	3	3.6	0	4	1.0
211	Mg	Cu	1	-1.2	2	15	1.0