**Engineering Materials – A Data-Driven Approach to Mechanical, Physical & Chemical Properties**

**Dataset 1 – Detailed Properties**

Mechanical, physical, and environmental properties with structured metadata

|  |  |  |  |
| --- | --- | --- | --- |
| **Abbreviation** | **What It Means** | **Simple Description** | **Real-Life Example** |
| Std | Standard | A rule or guideline everyone follows for parts or materials. | Using ANSI bolts to fit engine parts correctly. |
| ID | Identification Number | Identification Number | Identification Number |
| Material | Material | The stuff a part is made of. | Aluminium used for lightweight car wheels. |
| Heat treatment | Heat Treatment | Heating and cooling metal to make it stronger or tougher. | Gears that are hardened to last longer. |
| Su | Ultimate Strength | The most pulling force a material can handle before breaking. | Strong chassis frames made to survive crashes. |
| Sy | Yield Strength | When a material starts to bend and doesn't return to shape. | Control arms built to resist bending during bumps. |
| A5 | Elongation at Break | How much a material can stretch before snapping. | Car bumpers made from metals that can stretch in accidents. |
| Bhn | Brinell Hardness Number | How hard the surface of a material is when pressed? | Truck axles tested for surface hardness. |
| E | Young’s Modulus | How stiff a material is when you pull or push it. | Steel car frames that don't easily bend. |
| G | Shear Modulus | How stiff a material is when you twist it? | Driveshafts designed to resist twisting forces. |
| mu (μ) | Coefficient of Friction | How slippery or sticky two surfaces are. | Tires designed with a high grip for dry roads. |
| Ro (ρ) | Density | How heavy something is for its size. | Aluminium panels are light but strong, helping fuel economy. |
| pH | Acidity/Alkalinity Level | Shows if a fluid is more acidic or basic. | Coolant tested to keep engine parts from corroding. |
| Desc | Description | A short explanation about a part. | Labelling a part as "Forged steel crankshaft" in specs. |
| HV | Vickers Hardness | A way to measure hardness using a sharp diamond point. | Checking hardness of a camshaft after heat treatment. |

**Std, Material, Desc** → Tell you what a part is.

**Su, Sy, A5, Bhn, E, G, HV** → Tell you how strong or tough the material is.

**Ro, mu, pH** → Tell you physical properties important for real-world performance.

**Dataset 2 – Simplified View + Use Flag**

Flattened format with “Use = True/False” label indicating suitability for engineering application

|  |  |  |  |
| --- | --- | --- | --- |
| **Abbreviation** | **What It Means** | **Simple Description** | **Real-Life Automotive Example** |
| Material | Material | The type of stuff a part is made from. | Steel used for car body panels. |
| Su | Ultimate Strength | The maximum force a material can take before it breaks. | Crash bars built with high Su steel to survive impacts. |
| Sy | Yield Strength | The point where a material starts to bend and stays bent. | Suspension arms designed with high Sy to avoid bending on rough roads. |
| E | Young’s Modulus | How stiff a material is when stretched or squeezed? | Chassis frames made stiff to keep the car stable. |
| G | Shear Modulus | How stiff a material is when twisted. | Driveshafts designed to resist twisting during acceleration. |
| mu (μ) | Coefficient of Friction | How much two surfaces resist sliding against each other. | Tires made with high μ for better grip on dry pavement. |
| Ro (ρ) | Density | How heavy something is compared to its size. | Using aluminium instead of steel to make car parts lighter. |
| Use Flag | Use Flag | A marker showing if a material or part is allowed for use. | A material marked as “Approved” for making door beams. |

**PART 1 – Single Dataset Engineering Tasks (Dataset 1)**

**Task 1: Initial Exploration & Summary**

1. Identify the total number of materials and heat treatment types.
   1. Total number of materials in df1: **1552**
   2. Total number of unique materials in df1: **1225**
   3. Total number of heat treatments in df1: **1552**
   4. Total number of unique heat treatments in df1: **44**
2. Check for any missing or inconsistent data values.
3. **Missing Values per Column:** Heat treatment 750, A5 206, Bhn 1089, pH 1359, Desc 571, HV 1387
4. No duplicates.
5. **Columns containing outliers:** ['Su', 'Sy', 'A5', 'Bhn', 'G', 'mu', 'Ro', 'pH', 'HV']
6. **Number of outliers in each column:** Su: 73 outliers Sy: 96 outliers A5: 151 outliers Bhn: 439 outliers G: 51 outliers mu: 97 outliers Ro: 311 outliers pH: 175 outliers HV: 147 outliers.
7. Summarize key statistics of mechanical properties like Su (ultimate tensile strength), Sy (yield strength), and A5 (elongation at break).
8. **Handle Missing Values:** Missing values are filled with mode (first occurrence in case of multiple modes)
9. **Handle Outliers:** Using IQR – Capped lower bound to values below lower bound and similarly upper bound to values above upper bound.

|  |  |  |  |
| --- | --- | --- | --- |
| **Statistic** | **Before Cleaning** | **After Cleaning** | **What Changed? (Interpretation)** |
| **Mean** | 572.75 | 559.05 | Slight decrease — outliers pulled the average down. |
| **Standard Deviation** | 326.83 | 285.05 | Less spread — data is now more consistent. |
| **Minimum** | 69.00 | 69.00 | No change — weakest material remains. |
| **Maximum** | 2220.00 | 1252.50 | Capped — extremely high values adjusted using IQR cap. |

1. **Su (ultimate tensile strength) –** Before cleaning the data, the average material looked very strong but some values were very high like 2220. After handling the outliers, the average and standard deviation dropped providing a consistent and reliable value. On handling the extreme values, it helped to make a realistic decision.
2. **Sy (yield strength) -** Yield Strength is the stress level at which a material **starts to deform permanently.**

*The source data had few anomalies 280 max, 240 max, 210 max, 250 max, 210 max, 280 max, 240 max, 25 max and the datatype were also not integer. So first these were first cleaned and converted to integer before running the stats.*

**Before cleaning,** the data included a max yield strength of **2048** — extremely high, likely **exotic** or **erroneous.** This could mislead an engineer into thinking super-strong materials are standard. After cleaning, the max was capped at **867.5**, which is more realistic.

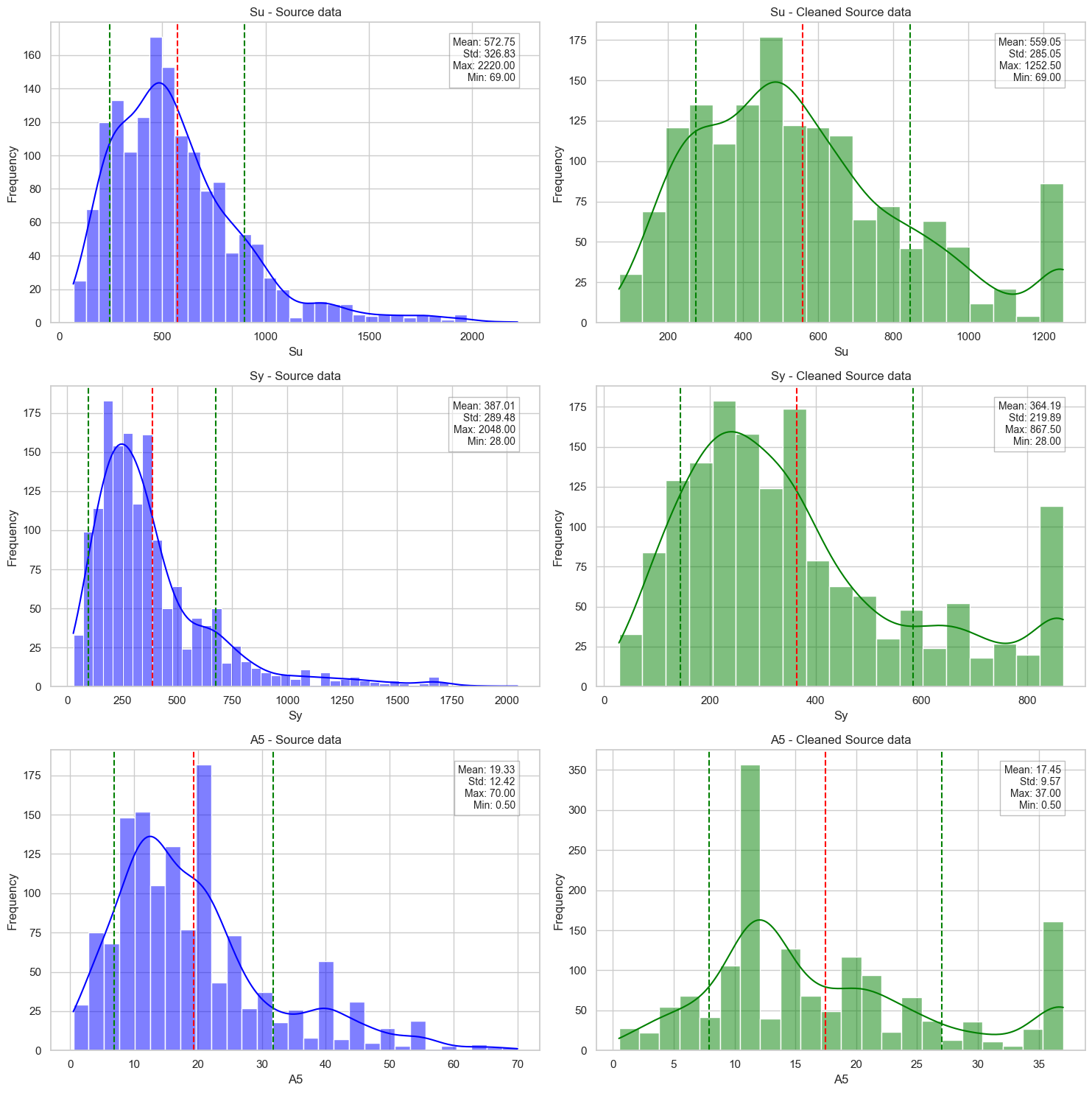
|  |  |  |  |
| --- | --- | --- | --- |
| **Statistic** | **Before Cleaning** | **After Cleaning** | **What Changed?** |
| **Mean** | 387.01 | 364.19 | Average lowered slightly after removing extreme highs. |
| **Standard Deviation** | 289.48 | 219.89 | Data spread reduced — results more consistent. |
| **Minimum** | 28.00 | 28.00 | Weakest material retained. |
| **Maximum** | 2048.00 | 867.50 | High outliers capped using IQR — unrealistic values removed. |

1. **A5 (elongation at break) -** It tells us **how much a material can stretch before it breaks,** expressed as a **percentage of its original length**. A higher A5 means the material is more **ductile** (can deform more before failing), while a lower value means it's **brittle**.

Before cleaning, we saw a max stretch of 70% — which is very high and might not reflect realistic metal behaviour. It could be a soft polymer or even an error. After capping outliers, the max value became 37%, which is more typical for high-ductility metals used in crash-absorbing components.

|  |  |  |  |
| --- | --- | --- | --- |
| **Statistic** | **Before Cleaning** | **After Cleaning** | **What Changed? (Simplified)** |
| **Mean** | 19.33 | 17.45 | Average slightly lowered — extreme values removed. |
| **Standard Deviation** | 12.42 | 9.57 | Variation reduced — values more consistent. |
| **Minimum** | 0.50 | 0.50 | No change — still includes very brittle materials. |
| **Maximum** | 70.00 | 37 | Unrealistically high stretch values capped using IQR. |

**Original Source data vs Cleaned Source data**

****

**Task 2: Groupwise Comparison**

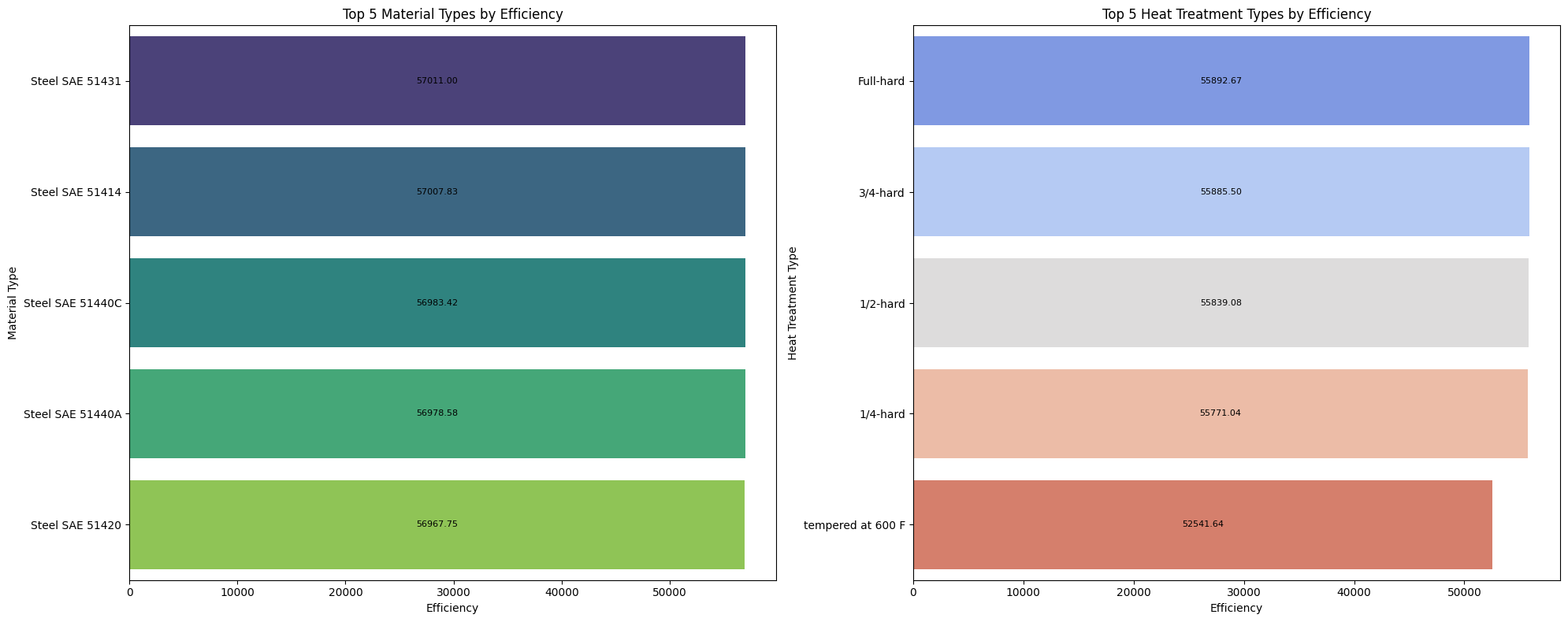
Compare average strength, ductility, and hardness values grouped by:

1. Material type
2. Heat treatment method

On grouping the Material Types and Heat Treatment Types **separately** with respect to Su, Sy, A5, E, G and HV,

**Material - Steel SAE 51431** has the highest efficiency of 57011.00 and

**Heat treatment - Full-hard** has the highest efficiency of 55892.67



But when calculating the efficiency of combined grouping of Material Types and Heat Treatment Types with respect to Strength, Ductility and Hardness, below are the insights

Top Material and Heat Treatment Types by Strength Efficiency (Su, Sy, E, G):

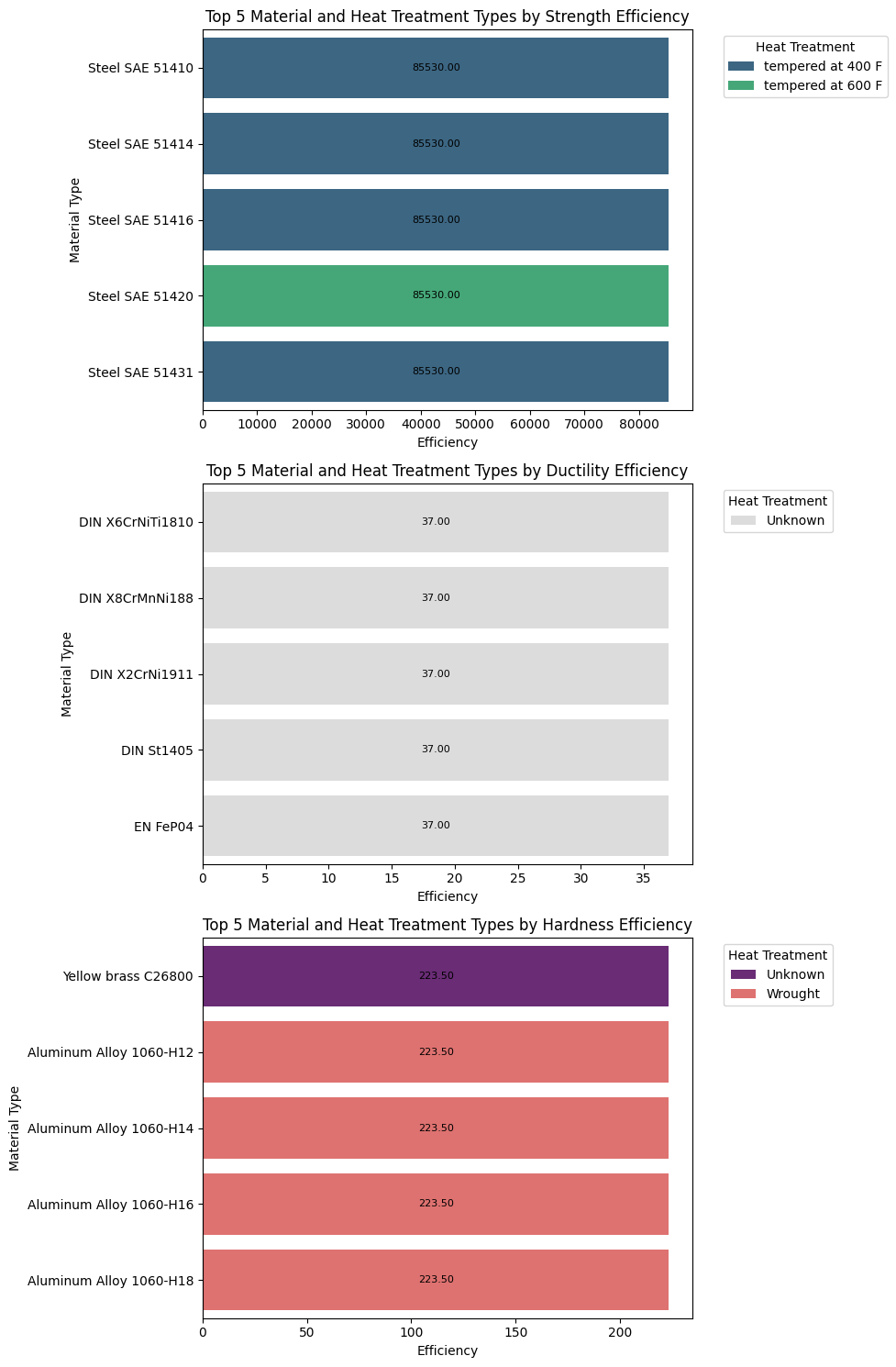
**Steel SAE 51410 - tempered at 400 has the highest efficiency of 85530.0**

Top Material and Heat Treatment Types by Ductility Efficiency (A5):

**DIN X6CrNiTi1810 – Unknown has the highest efficiency of 37.0**

Top Material and Heat Treatment Types by Hardness Efficiency (Bhn, HV):

**Yellow brass C26800 – Unknown has the highest efficiency of 223.5**

****

**Task 3: Design Ratio Analysis**

|  |  |  |
| --- | --- | --- |
| **Metric** | **Formula** | **Purpose** |
| **1. Strength-to-Hardness Ratio** | Su / Bhn | For comparing strength vs wear resistance (machinability, forming). |
| **2. Strength-to-Ductility Index** | Su × A5 | Measures energy absorption or formability under stress. |
| **3. Strength-to-Weight Proxy** | Su / Ro | For lightweight yet strong material choices. |

**Top 5 Materials by Strength-to-Hardness Ratio:**

|  |  |  |
| --- | --- | --- |
| **Material** | **Strength\_Hardness** | **Strength\_Hardness\_Rank** |
| **Steel SAE 1141** | 5.771889 | 1 |
| **DIN 55Cr3** | 5.771889 | 1 |
| **DIN 37Cr4** | 5.771889 | 1 |
| **DIN 50CrV4** | 5.771889 | 1 |
| **Steel SAE 1340** | 5.771889 | 1 |

The above materials can be used when choosing **chassis parts** or **suspension arms** that must be strong but still easy to machine or shape.

**Top 5 Materials by Strength-to-Ductility Index:**

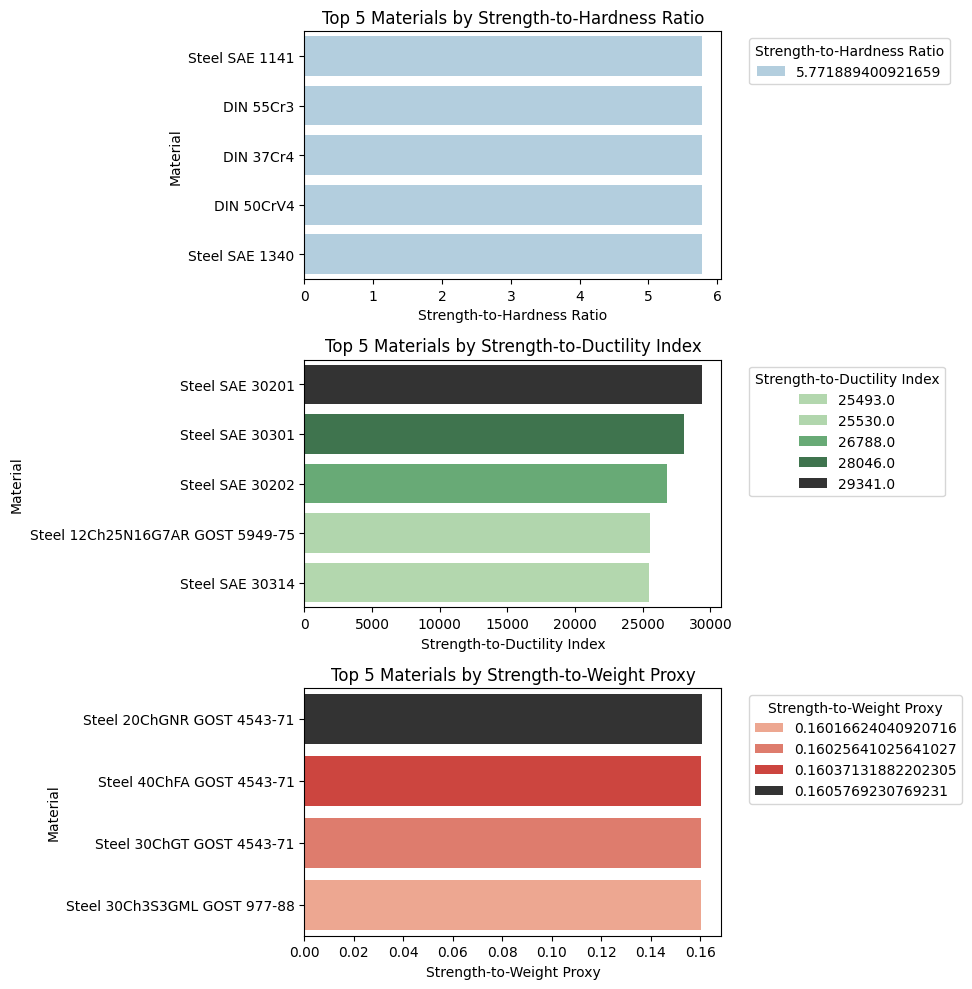
|  |  |  |
| --- | --- | --- |
| **Material** | **Strength\_Ductility** | **Strength\_Ductility\_Rank** |
| **Steel SAE 30201** | 29341 | 1 |
| **Steel SAE 30301** | 28046 | 2 |
| **Steel SAE 30202** | 26788 | 3 |
| **Steel 12Ch25N16G7AR GOST 5949-75** | 25530 | 4 |
| **Steel SAE 30314** | 25493 | 5 |

The above materials are ideal for **crumple zones** or **bumper reinforcements**, where materials must stretch before failure in crashes.

**Top 5 Materials by Strength-to-Weight Proxy:**

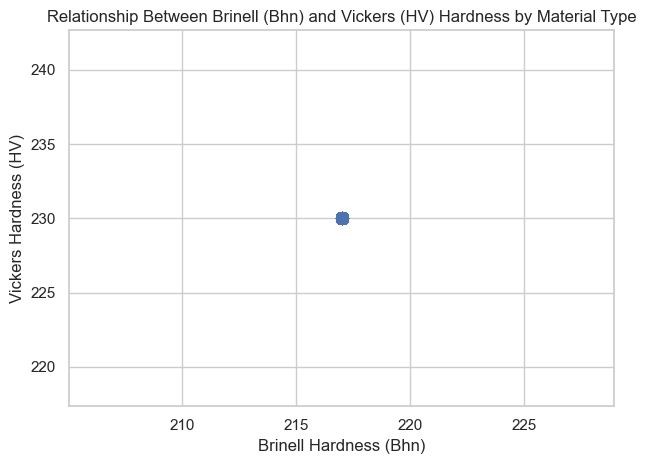
|  |  |  |
| --- | --- | --- |
| **Material** | **Strength\_Weight** | **Strength\_Weight\_Rank** |
| **Steel 20ChGNR GOST 4543-71** | 0.160576923 | 1 |
| **Steel 20ChGNR GOST 4543-71** | 0.160576923 | 1 |
| **Steel 40ChFA GOST 4543-71** | 0.160371319 | 2 |
| **Steel 30ChGT GOST 4543-71** | 0.16025641 | 3 |
| **Steel 30Ch3S3GML GOST 977-88** | 0.16016624 | 4 |

The above materials are **strong but lightweight** — critical for fuel efficiency especially in aerospace.

****

**Task 4: Hardness Scale Correlation**

After handling the missing values and outliers, **all materials** have the **same Brinell Hardness (Bhn = 217)** and **same Vickers Hardness (HV = 230)**

****

Since all hardness values are the same across materials, we cannot compare or find any pattern

**Correlation between Bhn and HV: nan**

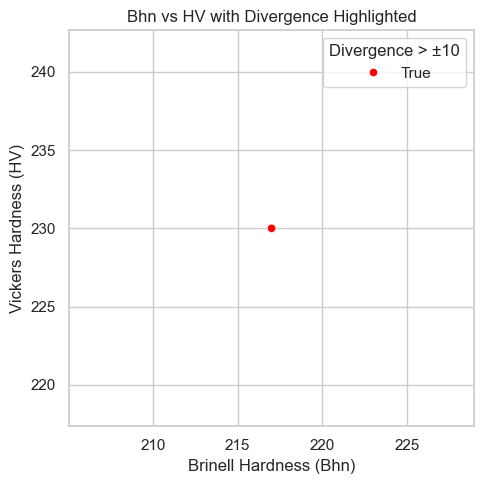
|  |  |  |  |
| --- | --- | --- | --- |
| **Observation** | **Statement** | **Real-World Example** | **Hypothesis** |
| All Bhn and HV values are the same | The dataset shows no variation in hardness values. | All car parts, including engine and brake components, have Bhn = 217 and HV = 230. | Default or placeholder values are used instead of real measurements. |
| Scatterplot shows only one dot | The plot repeats a single point with no visible spread. | A single dot represents all materials in the hardness comparison chart. | Test data lacks variation or was copied from a standard reference. |
| Correlation returns NaN or error | The correlation calculation fails due to zero variation. | Pearson correlation cannot compare identical values across materials. | Standard deviation is zero, making correlation undefined. |
| Trends and comparisons are not possible | Patterns, relationships, or insights cannot be derived. | Material selection cannot be justified based on hardness if all values are identical. | Dataset may be incomplete, synthetic, or over-normalized. |

**Detect & Plot Divergence Between Bhn and HV**

Bhn Unique Values: 217.0

HV Unique Values: 230.0

Divergence: 13.0

****

**Vickers Hardness (HV) is higher than Brinell Hardness (Bhn)** — especially by a noticeable amount (e.g., >10 units) — it often **suggests that the material's surface is harder than its core**.

A **deviation of +10 to +500 units** in favour of **HV** over **Bhn** suggests a **good fit** for applications requiring **wear resistance** (e.g., gears, pistons, etc.) where the **surface** needs to be harder than the **core**.

**Task 5: Elasticity and Deformability Insight**

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Full Name** | **What It Tells Us** | **Real-World Example (Automotive)** |
| E | Elastic Modulus | How stiff a material is — higher E means it resists stretching | Used in suspension springs to control how much they bend |
| G | Shear Modulus | How resistant a material is to shape changes (shear) | Important in drive shafts where twisting is involved |
| μ (mu) | Poisson’s Ratio | How much a material narrows when stretched | Critical in engine gaskets, which must stretch but stay sealed |

**Elastic Modulus (E) vs Shear Modulus (G)**

The data points are distributed linearly, indicating a proportional relationship between E and G. (One outlier in E–G plot)

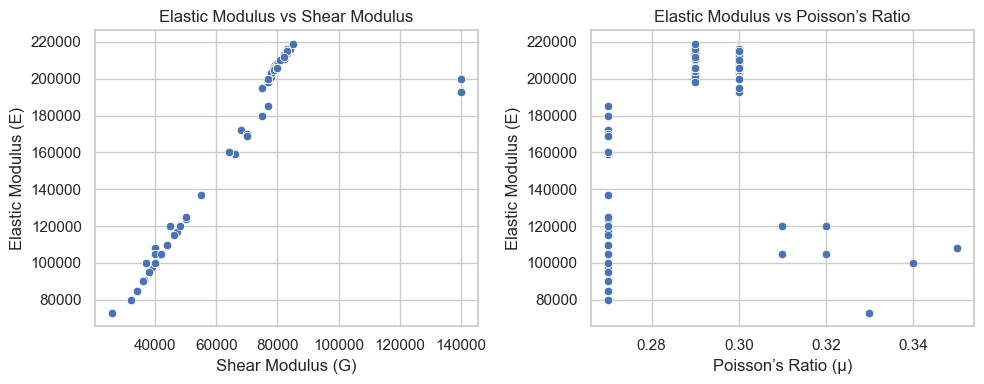
Most materials in the dataset behave isotropically (uniform stiffness in all directions)

For isotropic materials, E and G are mathematically related by: E=2G(1+μ)

**Elastic Modulus (E) vs Poisson’s Ratio (μ)**

The data points are scattered, showing no clear linear relationship between E and μ.

Poisson’s ratio doesn’t vary much — or E is more influenced by G than μ here.

****

**Task 6: Environmental Compatibility**

**All the values in the cleaned dataset are of pH\_Category: ['Basic']**

They are most suitable for environments where **alkaline exposure is common** and **corrosion resistance** is critical.

Surface Treatment can be still applied to increase the pH-resistant.

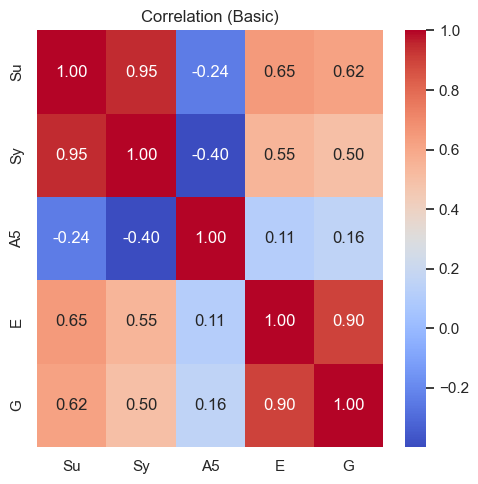
**Real-World Uses for Materials in Basic Environments**

Chemical Industry: Prevents corrosion in basic solutions (Storage tanks)

Marine: Sea water can shift toward alkaline due to salts and algae activity (Ship components)

Medical: Body fluids are slightly basic (pH ~7.4), requiring biocompatibility (Surgical tools)

**Correlation between various mechanical properties of materials.**



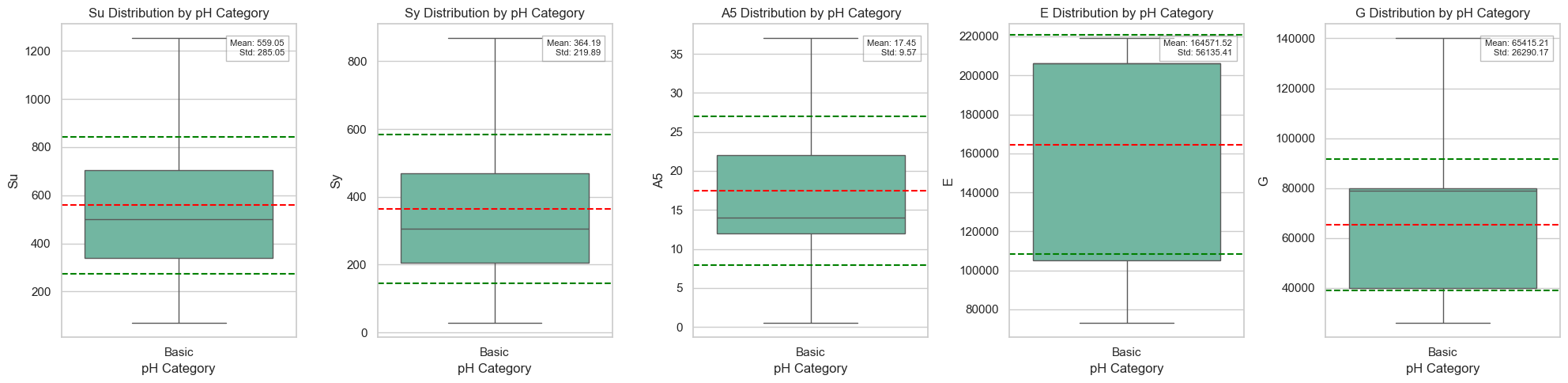
**Positive Correlation (Red):** Properties like tensile strength (Su) and yield strength (Sy) are strongly positively correlated, indicating that as one increases, the other tends to increase as well.

**Negative Correlation (Blue):** Some properties, such as elongation (A5) and strength metrics, show a negative correlation, suggesting an inverse relationship.

**Neutral Correlation (White):** Some properties have little to no correlation, indicating independence.

**Mechanical properties across the pH categories.**

**Basic** *(Since all the materials are of basic pH, we don’t have other categories to compare)*



**PART 2 – Cross-Dataset Engineering Tasks (Dataset 1 + Dataset 2)**

**Task 7: Material Identifier Matching**

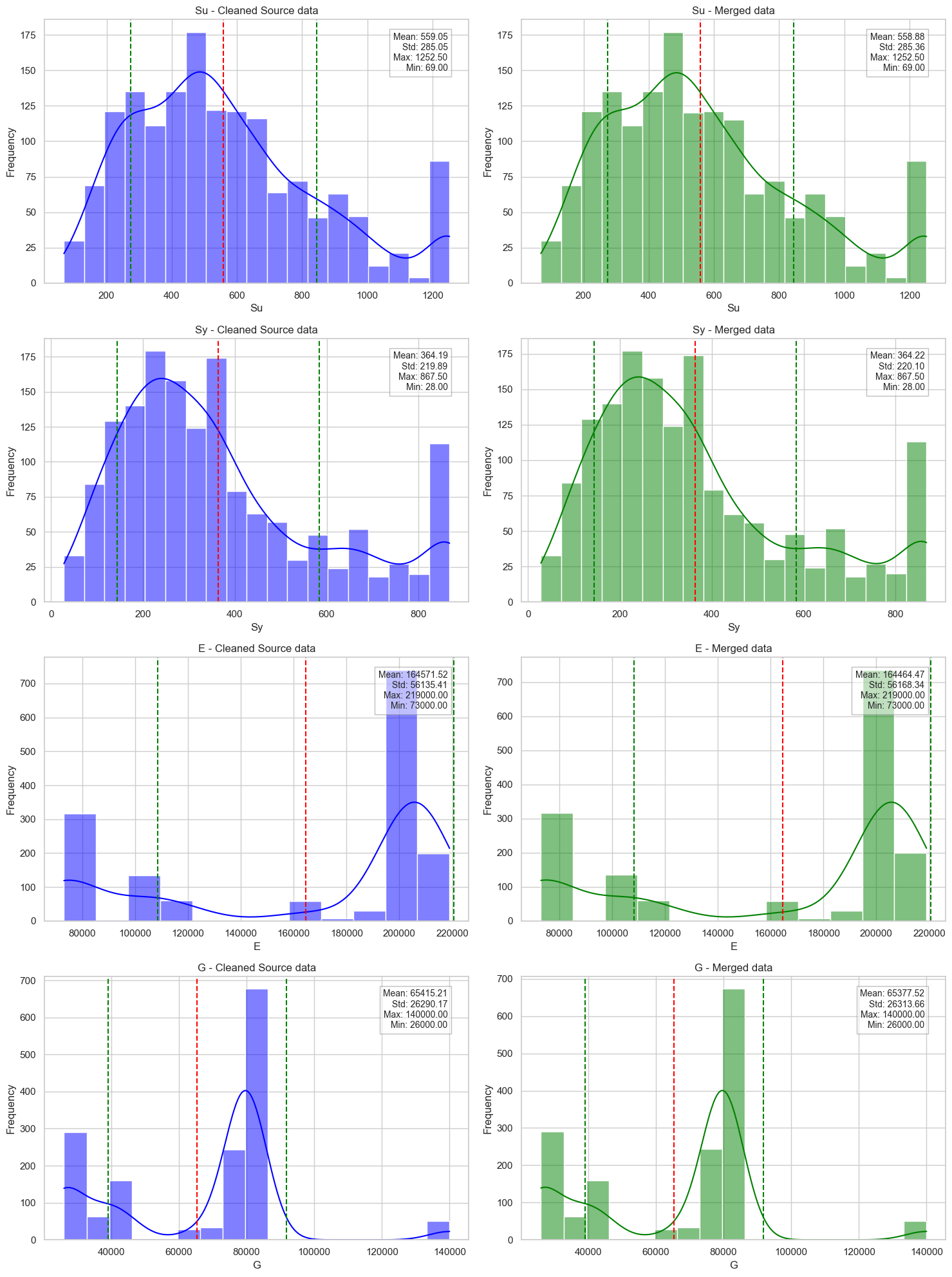
* Read the Dataset 2
* Dropped 4 duplicates materials: NF NF XC42TS, NF NF XC55H1, NF NF Z15CN2413, NF NF Z12CNS2520
* Identify and handle outliers:
  + Columns containing outliers: ['Su', 'Sy', 'G', 'mu', 'Ro']
  + Number of outliers in each column: Su: 73 outliers Sy: 96 outliers G: 51 outliers mu: 97 outliers Ro: 311 outliers
* Stripped leading and trailing spaces from all string columns in df2
* Created a new column called ‘Material\_Alias’ by concatenating ‘Std’, ‘Material’, ‘Heat treatment’

(For Heat treatment, already filled ‘Unknown’ value when the missing values are handled is replaced with blanks to match with the Dataset 2 values.)

* Since D2 doesn’t have ID column, a new ID is created by mapping the values from D1 on ‘Material\_Alias’ column to avoid duplicates.
* Records are merged based on ‘Material\_Alias’,’ID’ from D1 and ‘Material’,’ID’ from D2 using inner join.
* Initial counts of D1: 1552, D2: 1548, Merged\_Data: 1548

**Task 8: Discrepancy Audit**

* Among the 1548 merged records, 1461 materials have the same Su, Sy, E, G.
* After handling the outliers, D1 and D2 seems to have almost consistent data
* Above testimony is confirmed when Su, Sy, E, G values are compared from D1 and D2 values from Merged\_Data. **The mean, standard deviation, max and min value doesn’t have significant change.**



**Task 9: Use-Case Suitability Mapping**

Analyse patterns in the Use flag (True/False)

**Strength (Su, Sy)**

True: Mean values for Su (Ultimate Strength) and Sy (Yield Strength) are lower compared to the False Use Flag.

False: Higher variability in Su and Sy compared to True Use Flag.

**Ductility (A5)**

True: Higher ductility values with a mean of 20.80 and Std: 4.42.

False: Lower ductility values with a mean of 17.12 and Std: 7.41.

**Resistivity (Elastic Modulus - E, Shear Modulus - G)**

True: Higher mean values for both E (Elastic Modulus) and G (Shear Modulus).

False: Lower mean values with higher variability.

Common properties when for Use = True materials

* Lower Strength
* Higher Ductility
* Higher Elastic and Shear Modulus
* Less Standard Deviation

Threshold Summary

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Strength | | Ductility | Resistivity | |
| Use | Su | Sy | A5 | G | E |
| FALSE | **568.6** | **370** | **17.14** | **64020.8** | **160551.3** |
| TRUE | **457.7** | **303.5** | **20.6** | **79577.8** | **205422.2** |

"Use = True" Materials: Prioritize higher ductility and resistivity but have lower strength.

"Use = False" Materials: Favor higher strength but have lower ductility and resistivity.

