EICT ACADEMY
IIT GUWAHATI

DIYguru
E-Mobility Academy

Engineering Materials

A Data-Driven Approach to Mechanical, Physical & Chemical Properties

Comprehensive Analysis Report - Parts 1 & 2

Initial Exploration

Materials: **1,552**

Properties: **44**

Su Range: **179-1226 MPa**

Design Ratio Analysis

Top Su/Ro: **0.212**

Top Su×A5: **34,110**

Top Su/Bhn: **9.42**

Environmental Compatibility

Acidic: **~400 MPa**

Neutral: **~900 MPa**

Basic: **~1100 MPa**

Outlier Materials

Identified: **29**

Su Max: **2220 MPa**

A5 Min: **1%**

Explore Key Findings

Executive PG Certification Program in Data Science & AI/ML
for R&D Engineering Applications

May 8, 2025
Submitted by Jai Kumar Gupta

Materiial Selection Insights for Electric Vehicles

Material Selection Insights

- High Strength
- Balanced Strength-Ductility
- High Ductility

ENGINEERING MATERIAL INSIGHTS FOR ELECTRIC VEHICLES

ADVANCED SELECTION THROUGH DATA-DRIVEN ANALYSIS

Jai Kumar Gupta
Executive PG Program in Data Science & AI/ML
Mobile: +91 8953947619
Email: jaiku7867@gmail.com

May 8, 2025

The Program Director
DIYguru E-Mobility Academy

Subject: Submission of Project Report - “Engineering Materials: A Data-Driven Approach to Mechanical, Physical & Chemical Properties”

Respected Sir/Madam,

I am pleased to submit my comprehensive project report titled “Engineering Materials: A Data-Driven Approach to Mechanical, Physical & Chemical Properties” as part of the course requirements for the Executive PG Certification Program in Data Science & AI/ML for R&D Engineering Applications.

This report presents an in-depth analysis of engineering materials across twelve specialized tasks, divided into two parts:

Part 1: Focuses on single dataset engineering tasks, including initial exploration, groupwise comparison, design ratio analysis, hardness scale correlation, elasticity insights, and environmental compatibility.

Part 2: Examines cross-dataset engineering tasks, covering material identifier matching, discrepancy audit, use-case suitability mapping, multi-criteria ranking, outlier identification, and material descriptor analysis.

The findings highlight fundamental material selection trade-offs and provide comprehensive guidelines for engineering applications. The analysis demonstrates the application of data science techniques to extract meaningful insights from complex engineering datasets.

I would like to express my sincere gratitude to the faculty especially Vandana Mam and mentors at EICT Academy, IIT Guwahati and DIYguru E-Mobility Academy for their invaluable guidance throughout this project.

I look forward to your feedback and evaluation.

Sincerely,

Jai Kumar Gupta

Engineering Materials Analysis

Comprehensive Summary Report

PART 1 – Single Dataset Engineering Tasks

Task 1: Initial Exploration & Summary
Task 2: Groupwise Comparison
Task 3: Design Ratio Analysis
Task 4: Hardness Scale Correlation
Task 5: Elasticity and Deformability Insight
Task 6: Environmental Compatibility

May 8, 2025

Contents

1 Introduction

This report synthesizes findings from six engineering materials analysis tasks, providing comprehensive insights into material properties, performance metrics, and selection criteria for engineering applications. The analysis spans from basic mechanical properties to specialized aspects such as environmental compatibility and elastic behavior.

The materials dataset contains 1,552 entries covering 1,225 unique materials subjected to 44 different heat treatment processes. Each analysis task focuses on specific aspects of material performance to guide engineering decision-making across various applications.

Key Concept

Material selection is a critical engineering decision process requiring the systematic understanding of mechanical, physical, and chemical properties. The six analysis tasks presented in this report provide complementary perspectives on how to evaluate and select materials based on different criteria and application requirements.

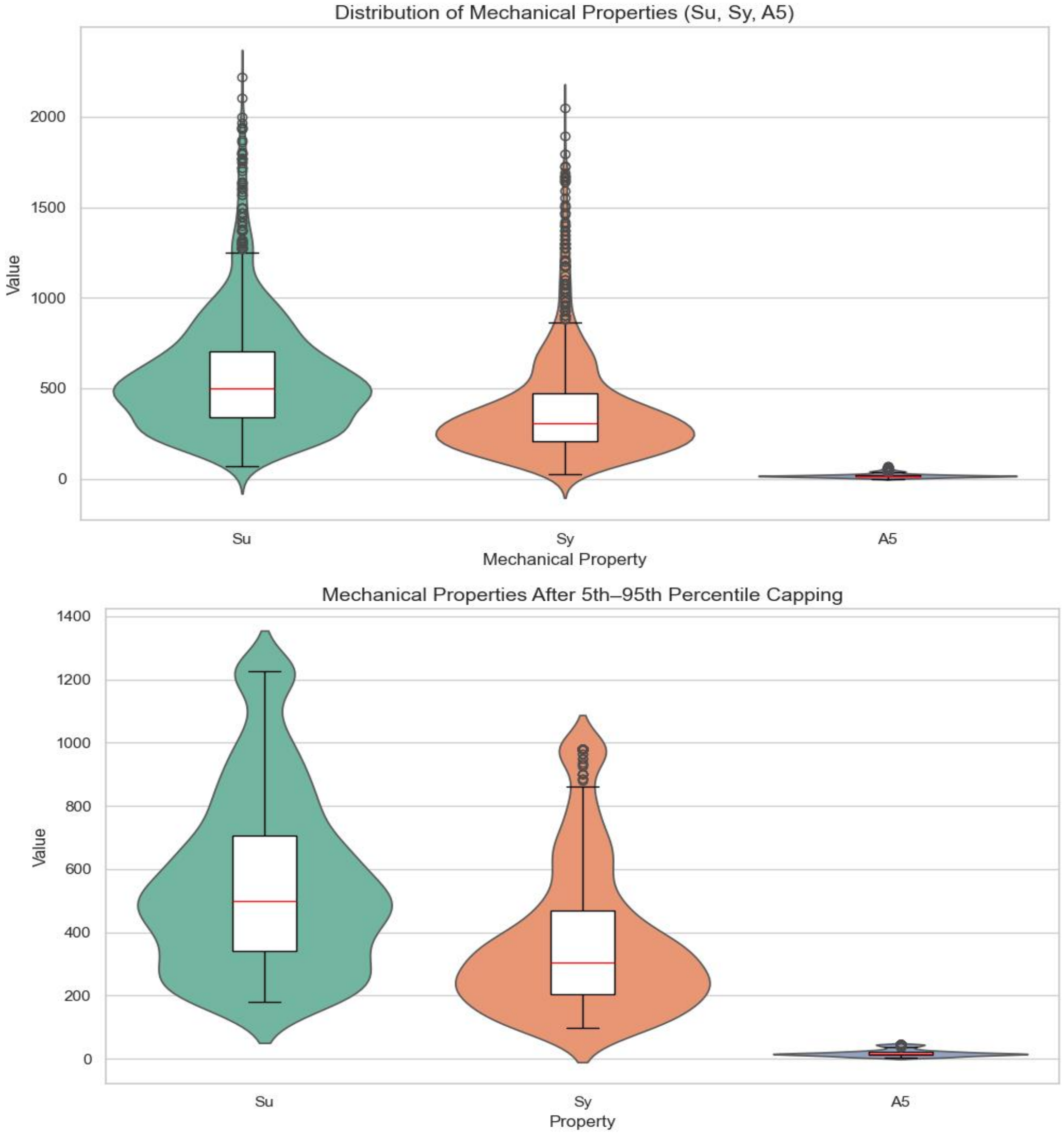
2 Task 1: Initial Exploration & Summary

Material Selection Insights

- The dataset reveals significant trade-offs between strength and ductility that engineers must consider when selecting materials.
- High-strength materials (Su $>$ 1000 MPa) typically exhibit lower elongation values (A5 $<$ 10%), making them suitable for high-stress applications where deformation must be minimized.
- High-ductility materials (A5 $>$ 30%) generally have moderate strength values (Su between 300-500 MPa), making them appropriate for applications requiring formability and energy absorption.
- Balanced materials (Su = 500 MPa, Sy = 305 MPa, A5 = 16%) represent a compromise between strength and ductility suitable for general-purpose applications.
- Heat treatment significantly modifies material properties, as demonstrated by the example of Steel SAE 1030, which shows dramatic increases in strength when tempered at the expense of reduced ductility.

Data Quality Observations

- Several properties showed significant missing value rates: Heat treatment (48.3%), Elongation (13.3%), Brinell hardness (70.2%), pH (87.6%), Description (36.8%), and Vickers hardness (89.4%).
- Yield strength (Sy) values contained string entries with qualifiers (e.g., "280 max", "240 max") that required preprocessing.
- Some material property values showed extreme outliers, potentially indicating specialized materials or measurement errors.
- After applying 5th-95th percentile capping to manage outliers, more typical value ranges were established: Su (179-1226 MPa), Sy (97-980 MPa), and A5 (4.0-45.0%).



Su (MPa): 69–2220 MPa (original), 179–1226 MPa (capped)
Sy (MPa): 28–2048 MPa (original), 97–980 MPa (capped)
A5 (%): 0.5–70.0% (original), 4.0–45.0% (capped)

✓Key Observation:
The distribution reveals significant trade-offs between strength and ductility. Materials with higher strength values typically exhibit lower elongation percentages, creating a fundamental engineering selection challenge.

Figure 1: Distribution of Mechanical Properties

3 Task 2: Groupwise Comparison

Material Selection Insights

- Materials with identical ultimate tensile strength (Su = 1226 MPa) showed varying ductility levels, creating important selection trade-offs.
- Materials like CSN 16640 offer an optimal balance of high strength (1226 MPa) and good ductility (16%), making them suitable for applications requiring both properties.
- Heat treatment methods significantly alter material properties, creating distinct performance profiles.
- Surface engineering methods like nitro-case-hardening deliver exceptional surface hardness (HV = 630) while preserving reasonable ductility (A5 = 12.5%), making them optimal for components requiring wear resistance.
- Specific material-treatment combinations are recommended for different applications:
  - Wear-resistant components: SAE 8640/8660 with nitro-case-hardening
  - High-load transmission: CSN 16640 tempered at 800°F
  - Static structural parts: BS grades with full-hard treatment
  - Forming operations: Lower-strength steels with annealed/normalized treatment

Data Quality Observations

- Missing hardness data across multiple material groups limited hardness-based comparisons.
- Full-hard and 3/4-hard treatments lack hardness values, preventing evaluation of surface effects.
- Inconsistent use of Brinell versus Vickers hardness scales created challenges in direct comparisons.
- Multiple materials show identical Su values (1226 MPa), suggesting potential standardization, rounding, or measurement limitations.
- Properties critical for comprehensive material evaluation (like impact resistance or fatigue behavior) are not included in this groupwise analysis.

Table 1: Material Properties by Material Type (Top 5 by Strength)

Material	Su (MPa)	A5 (%)	Hardness	Application Suitability
BS 525A60	1226	6	N/A	High-stress, limited deformation
BS 735A51	1226	6	N/A	Static loads, minimal ductility required
CSN 16640	1226	16	HV: 510	Balanced strength-ductility profile
Steel SAE 8660	1226	13	BHN: 460	Components needing toughness
Steel SAE 8640	1226	10	BHN: 505	Wear-resistant applications

Table 2: Material Properties by Heat Treatment Method (Top 5 by Strength)

Heat Treatment	Su (MPa)	A5 (%)	Hardness	Application Suitability
Full-hard	1226	6	N/A	Maximum strength requirements
Nitro-case-hard.	1226	12.5	HV: 630	Wear resistance with ductility
Tempered at 800°F	1226	10.5	BHN: 465	Balanced mechanical properties
3/4-hard	1207	8.5	N/A	High strength, formability
Tempered at 400°F	1173	10.5	BHN: 463	General mechanical components

Table 3: Recommended Material-Treatment Combinations by Application

Application	Material Treatment	Key Properties
Wear-resistant components	SAE 8640/8660 Nitro-case-hardened	High Su, HV = 630
High-load transmission	CSN 16640 Tempered at 800°F	Su = 1226 MPa, A5 = 10.5%
Static structural parts	BS grades Full-hard	Maximum strength, minimal ductility
Forming operations	Lower-strength steels Annealed/Normalized	Moderate Su, high A5
General mechanical parts	SAE grades Tempered at 400°F	Balanced strength-ductility

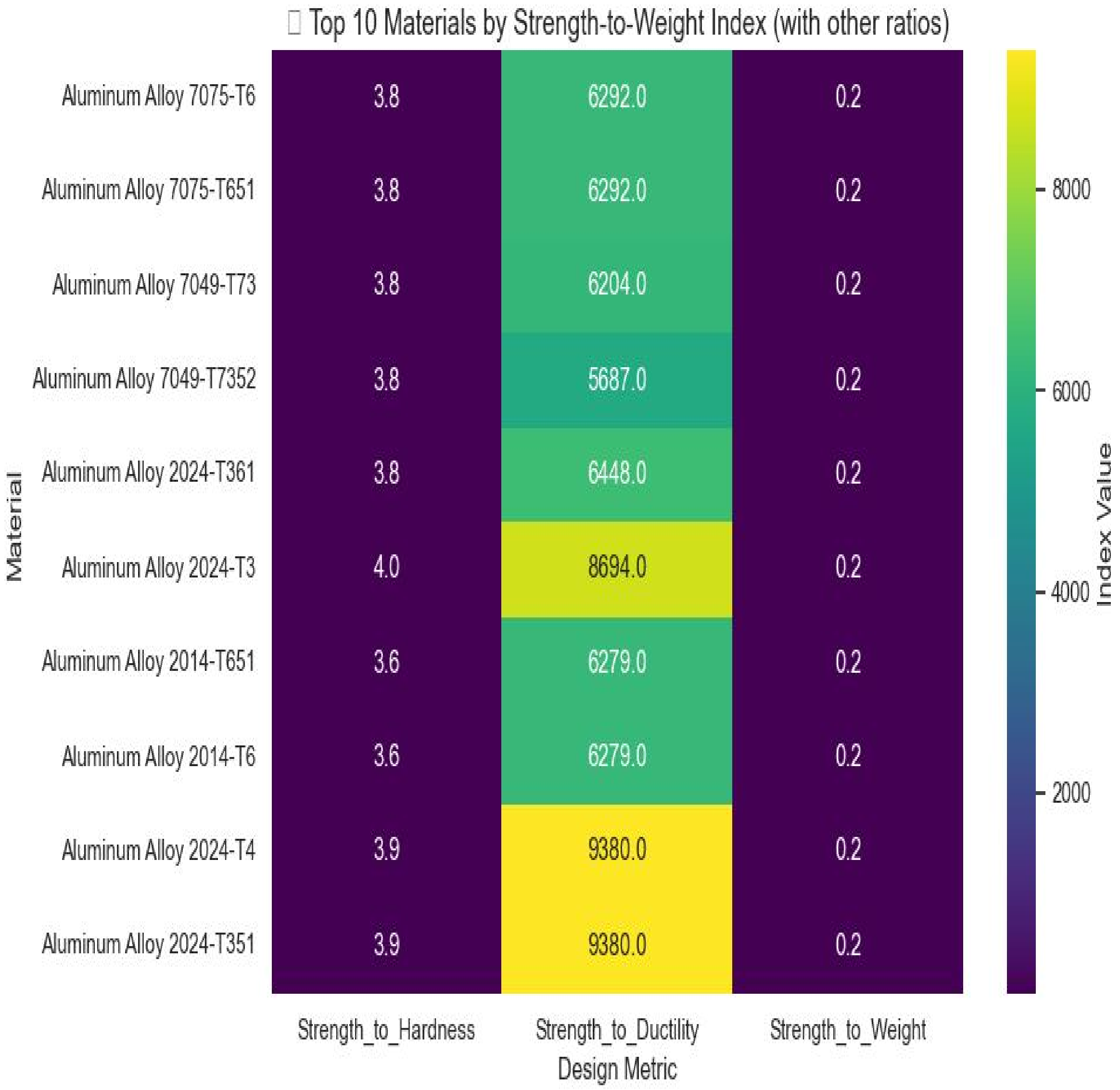
4 Task 3: Design Ratio Analysis

Material Selection Insights

- Three custom strength metrics were analyzed to reveal material performance patterns:
  - Strength-to-Hardness ratio (Su/Bhn): Wrought aluminum alloys outperform other materials, with values reaching 9.42 for Aluminum Alloy 1060-O.
  - Strength-to-Ductility index (Su×A5): SAE 303 series stainless steels in annealed condition dominate with exceptionally high values reaching 34,110 for SAE 30301.
  - Strength-to-Weight ratio (Su/Ro): Aluminum alloys in the 7000 series (particularly 7075-T6/T651) demonstrate superior performance with values reaching 0.212.
- The analysis reveals fundamental material selection trade-offs—excellence in one metric typically comes at the expense of another.
- Different material families show distinct ratio strengths—aluminum alloys excel in Strength-to-Weight and Strength-to-Hardness, while stainless steels dominate Strength-to-Ductility performance.
- Application-specific material recommendations:
  - Lightweight structures: Aluminum Alloy 7075-T6/T651
  - Energy absorption: Steel SAE 30301 (annealed)
  - Complex machined components: Aluminum Alloy 1060-O
  - Balanced performance: Aluminum Alloy 7049-T73

Data Quality Observations

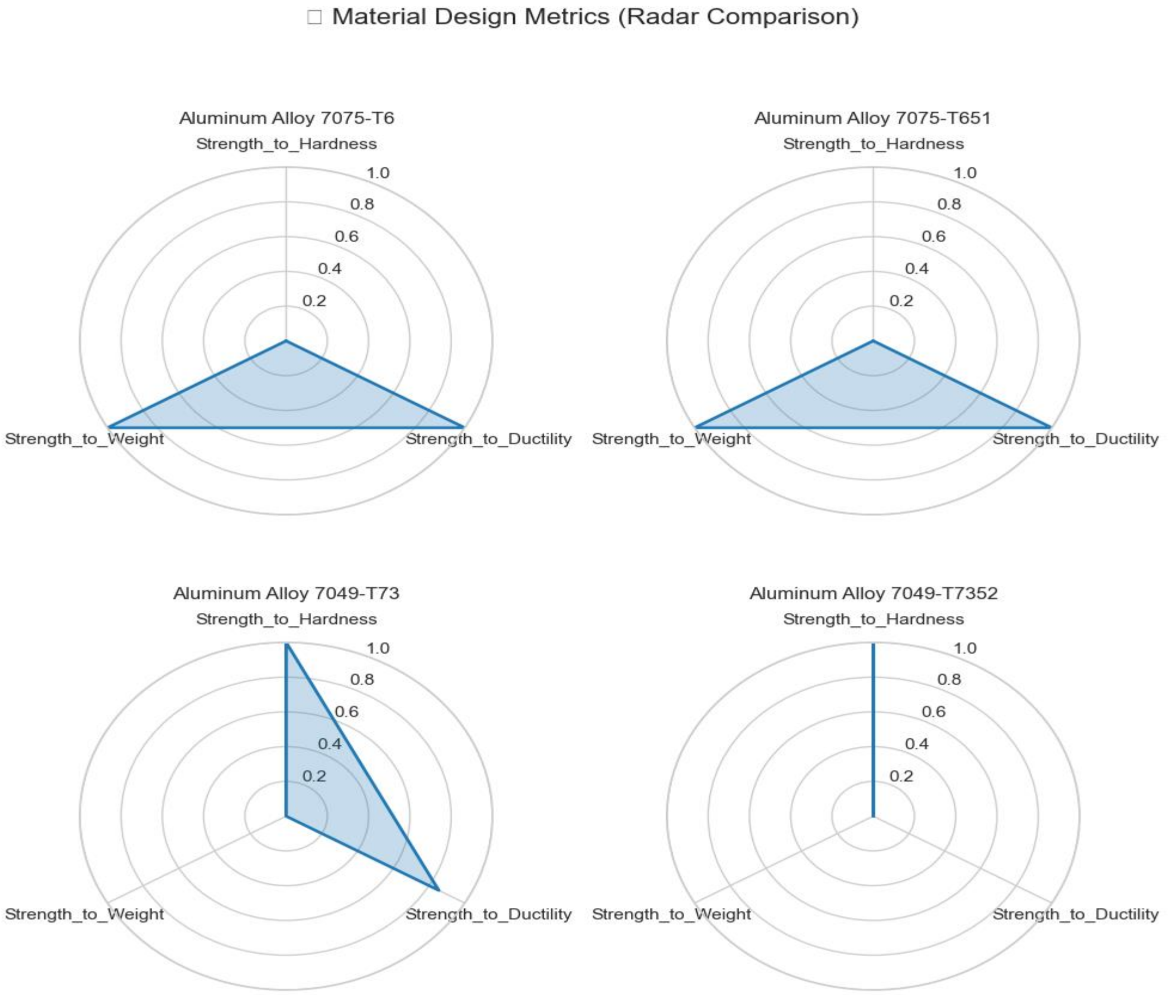
- Numerous materials had to be excluded from the analysis due to missing values for one or more properties required for ratio calculations (Su, Bhn, A5, or Ro).
- The remarkably consistent Strength-to-Weight values (approximately 0.2) across aluminum alloys suggests possible data standardization, rounding effects, or measurement limitations.
- The exceptionally high Strength-to-Ductility values for annealed stainless steels (exceeding 30,000) warrant verification to ensure measurement accuracy.
- Limited information about precise processing history (cold work percentage, exact aging parameters, etc.) restricts full understanding of why certain materials perform better in specific metrics.



Strength-to-Weight: Consistent values 0.2 across aluminum alloys
Strength-to-Ductility: Ranges from 5600 to 9380, with 2024-T4 and 2024-T351 superior
Strength-to-Hardness: Moderate differentiation (3.6-4.0), 2024-T3 slightly better

✓Key Observation:
Aluminum alloys show excellent structural efficiency (similar Su/Ro), but vary significantly in toughness (Su×A5). This enables selection optimization based on secondary property requirements while maintaining consistent weight efficiency.

Figure 2: Key Performance Indexes for Aluminum Alloys



7075 Alloys: High in Strength-to-Weight and Strength-to-Ductility, but low in Strength-to-Hardness
7049-T73: Low-to-moderate across all three ratios
7049-T7352: Maximizes Strength-to-Hardness at significant expense to other ratios

✓Key Observation:
The inverse relationship between ratio properties demonstrates that simultaneous optimization of all three metrics is impossible. Material selection must prioritize the most critical performance ratio for the specific application.

Figure 3: Trade-offs Between Performance Ratios

5 Task 4: Hardness Scale Correlation

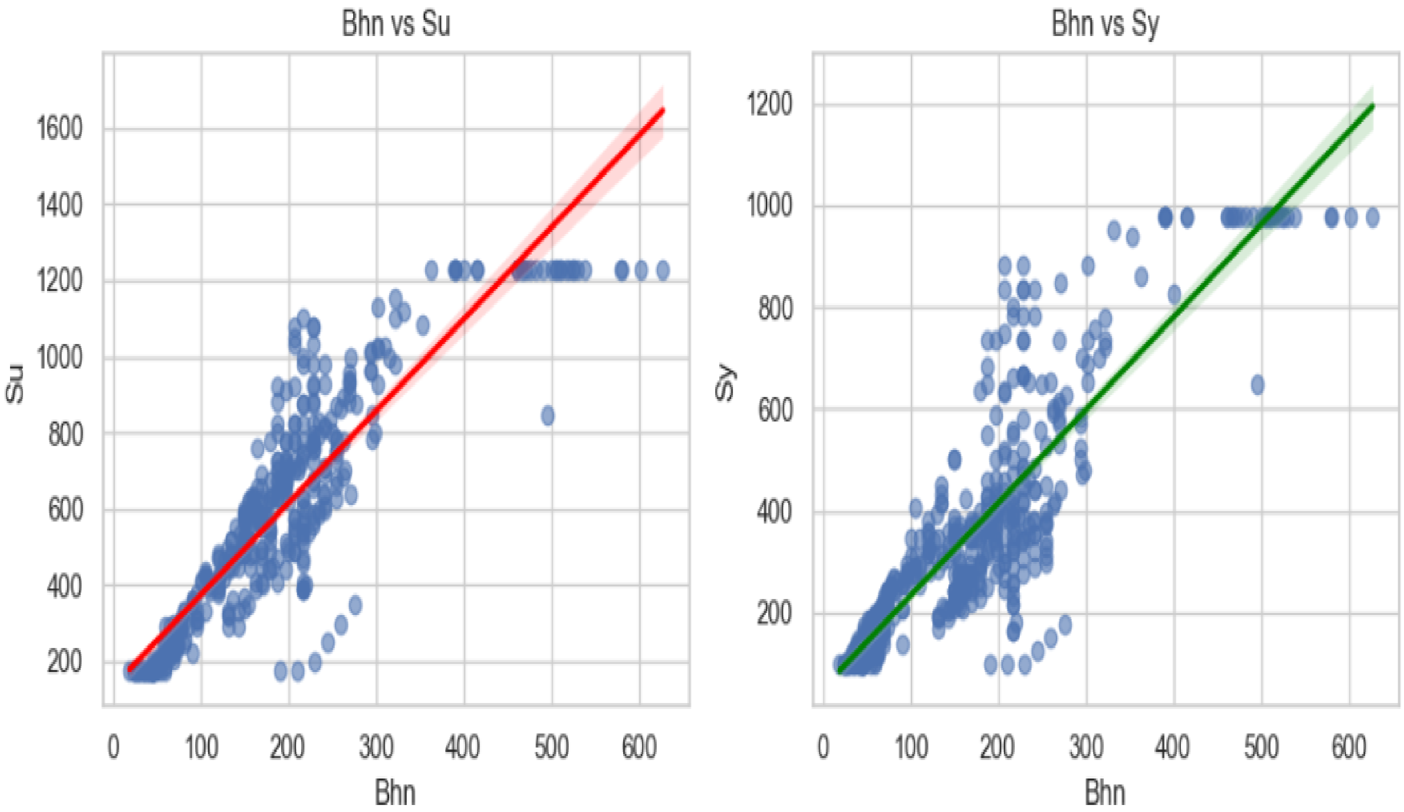
Material Selection Insights

- Brinell hardness shows strong correlations with mechanical properties:
  - Bhn vs Su: Very strong positive linear relationship (r = 0.900)
  - Bhn vs Sy: Strong positive linear relationship (r = 0.866)
  - Bhn vs A5: Weak negative linear relationship (r = -0.109)
- Vickers hardness shows different correlation patterns:
  - HV vs Su: Moderate-to-strong monotonic increase (Spearman = 0.559)
  - HV vs Sy: Moderate-to-strong monotonic increase (Spearman = 0.617)
  - HV vs A5: Moderate monotonic decrease (Spearman = -0.461)
- The difference between hardness measurement techniques reveals the importance of matching the test method to application requirements:
  - Vickers (with smaller indenter and lighter loads) better captures surface and local hardness variations
  - Brinell (with larger indenter and heavier test load) provides better bulk property evaluation
- Application-specific guidelines:
  - Structural Components: Brinell hardness serves as an excellent screening tool (Bhn 150-300)
  - Wear-Critical Applications: Vickers hardness better predicts performance (HV $>$ 500)
  - Impact-Resistant Components: Keep hardness values below HV 400
  - Fatigue-Critical Applications: Target moderate Bhn (150-250) with elongation values above 10%

Data Quality Observations

- A critical observation is the complete absence of samples with both Brinell and Vickers hardness values simultaneously present:
  - 463 samples had Brinell hardness values
  - 165 samples had Vickers hardness values
  - 0 samples had both values present
- This significant data quality issue prevents direct correlation between the two hardness scales, forcing reliance on indirect correlations through other mechanical properties.
- Correlation strengths vary between properties and testing methods:
  - Brinell hardness shows stronger linear correlations with strength metrics (Pearson $>$ 0.86)
  - Vickers hardness shows moderate linear correlations with strength (0.55-0.62)

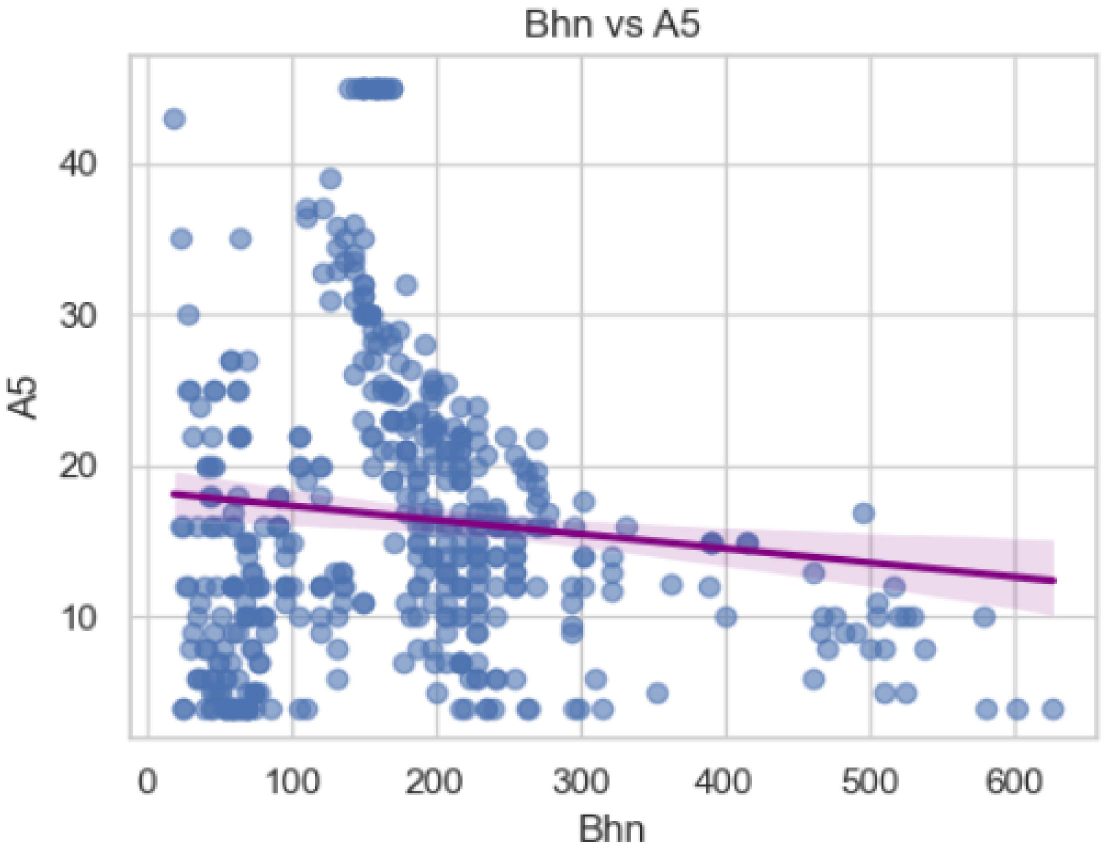
  - The ductility correlation is better captured by Spearman (rank-based) than Pearson (linear) analysis



Bhn vs Su: Pearson correlation coefficient = 0.900 (very strong positive relationship)
Bhn vs Sy: Pearson correlation coefficient = 0.866 (strong positive relationship)
Trend Lines: Clear positive slopes indicate proportional increase of strength with hardness

✓**Key Observation:**
The remarkably strong correlations confirm that Brinell hardness testing can serve as a reliable proxy for strength estimation when direct tensile testing is impractical, with hardness explaining over 80% of the variance in tensile and yield strength.

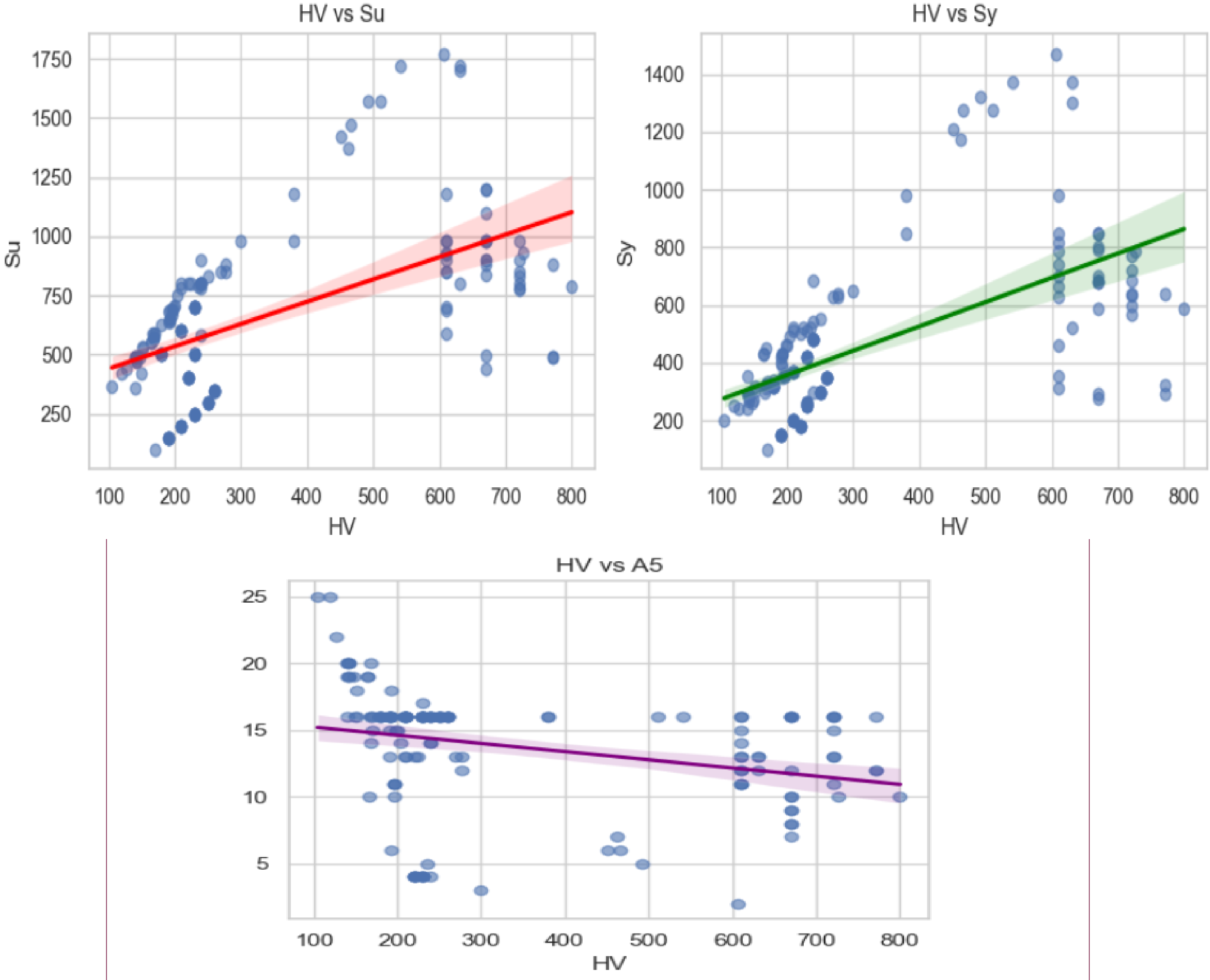
Figure 4: Brinell Hardness vs. Strength Properties



Bhn vs A5: Pearson correlation coefficient = -0.109 (weak negative linear relationship)
Trend Pattern: As Bhn increases, A5 decreases, illustrating the hardness-ductility trade-off
Data Spread: Significant scatter at lower hardness values (Bhn < 300) shows material variability

✓Key Observation:
While the linear correlation is weak, the overall negative trend confirms the fundamental materials science principle that increasing hardness compromises ductility. The Spearman (rank) correlation better captures this relationship than the Pearson (linear) coefficient.

Figure 5: Brinell Hardness vs. Ductility



HV vs Su: Spearman correlation = 0.559 (moderate-to-strong monotonic increase)
HV vs Sy: Spearman correlation = 0.617 (moderate-to-strong monotonic increase)
HV vs A5: Spearman correlation = -0.461 (moderate monotonic decrease)

✓Key Observation:
Vickers hardness shows stronger correlation with ductility (A5) than Brinell, capturing the hardness-ductility trade-off more effectively. Materials with HV > 500 typically show significant ductility reduction, crucial for applications requiring deformability such as sheet metal forming.

Figure 6: Vickers Hardness Relationships with Mechanical Properties

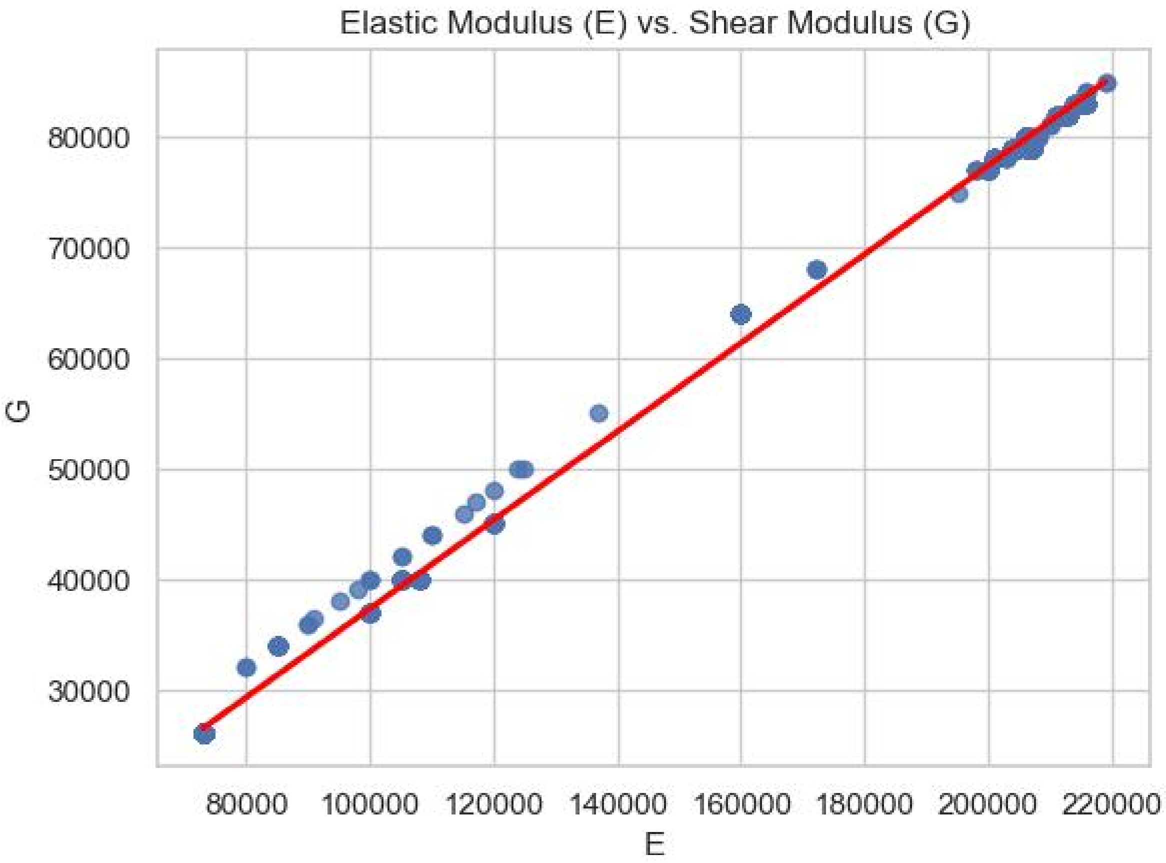
6 Task 5: Elasticity and Deformability Insight

Material Selection Insights

- The remarkably strong correlation between Elastic modulus (E) and Shear modulus (G) (r = 0.999) confirms that real engineering materials closely follow the theoretical isotropic relationship $G = \frac{E}{2(1+\mu)}$.
- The moderate negative correlation between E and Poisson's ratio (μ) (r = -0.579) indicates that as materials become stiffer (higher E), they generally exhibit slightly lower Poisson's ratios.
- Most engineering metals maintain Poisson's ratios within a narrow expected range (0.26-0.34), regardless of their elastic modulus values.
- Different material classes cluster at specific E-G value ranges:
  - Aluminum alloys: E 73 GPa, G 26-27 GPa
  - Steels: E 200 GPa, G 80 GPa
- The majority of materials show excellent agreement between measured G and calculated G expected, validating that isotropic material models are appropriate for most engineering applications.
- Application-specific material selection guidelines:
  - Structural Rigidity: Select high-E materials (carbon steels, E 200+ GPa)
  - Weight-Critical: Choose aluminum alloys (E 73 GPa) with lower density
  - Deformation Control: Consider materials with lower Poisson's ratios (closer to 0.25) such as cast irons

Data Quality Observations

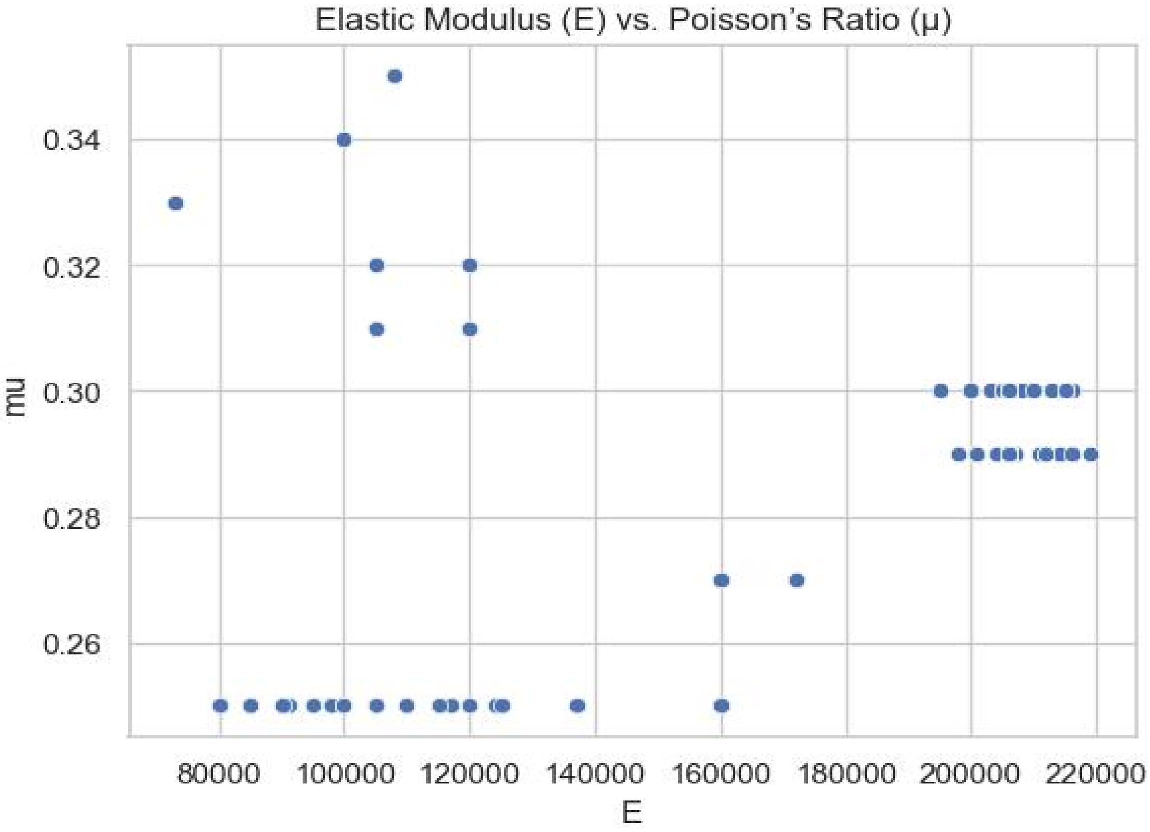
- The presence of identical values across different material grades suggests some data may represent standardized values rather than individual measurements.
- Perfect zero deviations between measured G and theoretical G (calculated from E and μ) strongly suggest that some G values were derived rather than independently measured.
- The largest deviations (approximately -1444 MPa) occur systematically in aluminum alloys, representing about 5.3% of the expected value.
- Values outside the typical Poisson's ratio range ($\mu < 0.25$ or $\mu > 0.35$) may indicate specialized materials (ceramics, polymers) or potential measurement errors.



E-G Correlation: Pearson coefficient = 0.999 (near-perfect linear relationship)
Theoretical Slope: 0.38 (matches observed data)
Clustering Patterns: Aluminum alloys (E 73 GPa, G 26-27 GPa); Steels (E 200 GPa, G 80 GPa)

✓**Key Observation:**
The extraordinarily strong correlation confirms that materials closely follow the theoretical isotropic relationship. Engineers can confidently predict shear modulus from elastic modulus, simplifying material selection when stiffness is critical and supporting the use of isotropic material models in engineering analysis.

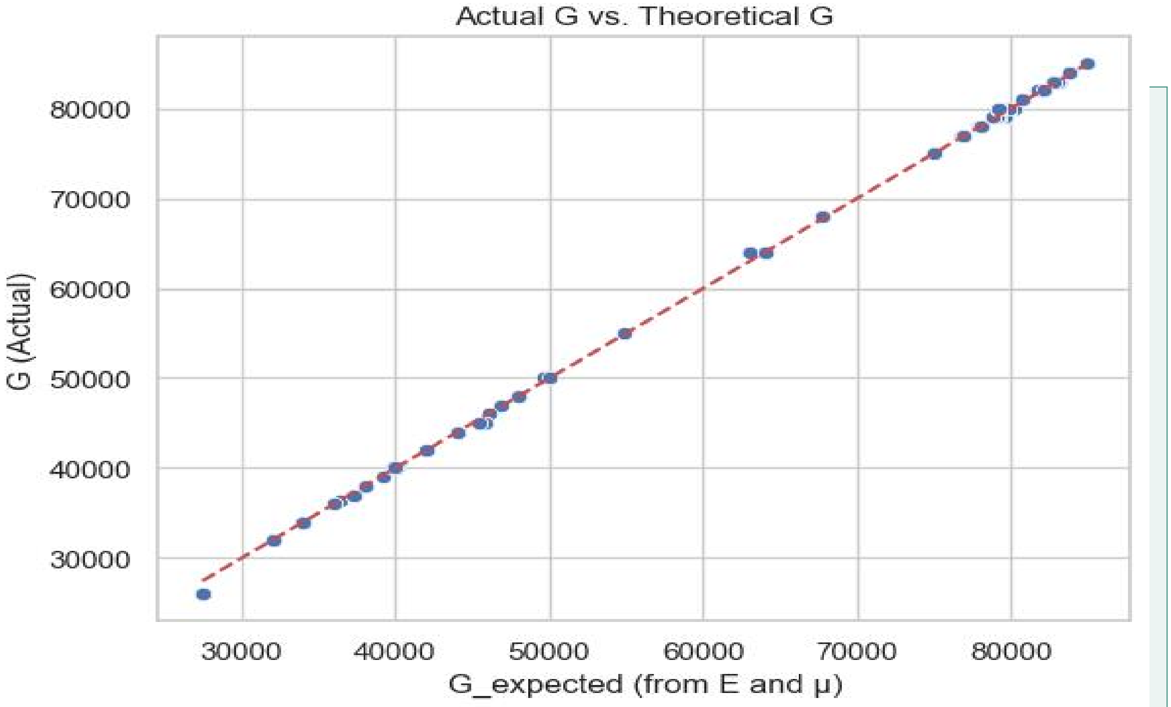
Figure 7: Relationship Between Elastic and Shear Moduli



E-μ Correlation: Pearson coefficient = -0.579 (moderate negative relationship)
Poisson's Ratio Range: 0.26-0.34 (typical for most metals)
Distribution Pattern: As E increases, μ slightly decreases, contrary to intuitive expectations

✓Key Observation:
Most engineering metals maintain Poisson's ratios within a narrow expected range (0.26-0.34), regardless of their elastic modulus values. Values outside this range ($\mu < 0.25$ or $\mu > 0.35$) may indicate specialized materials (ceramics, polymers) or potential measurement errors requiring further investigation.

Figure 8: Relationship Between Elastic Modulus and Poisson's Ratio



Alignment Quality: Points cluster tightly along the identity line
Maximum Deviation: 5.3% in aluminum alloys (-1444 MPa)
Perfect Matches: Multiple cast iron varieties show zero deviation

✓**Key Observation:**
The excellent agreement between measured and calculated G values validates isotropic material assumptions for most engineering applications. Perfect zero deviations in some materials suggest G values may have been calculated rather than independently measured, raising data source considerations.

Figure 9: Validation of Isotropic Material Behavior

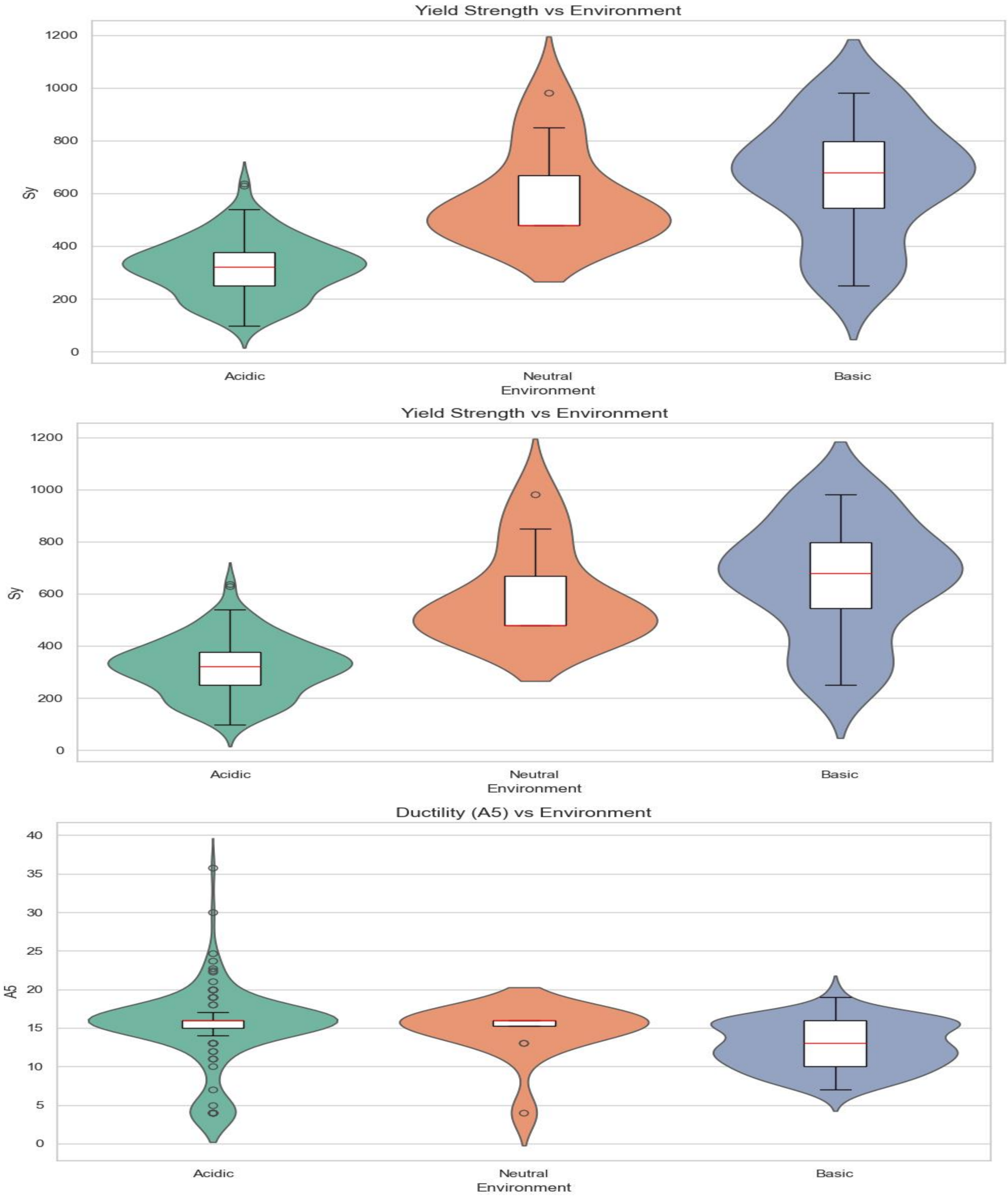
7 Task 6: Environmental Compatibility

Material Selection Insights

- Materials show distinct mechanical property patterns across different pH environments:
  - Acidic (<6): Lower strength (Su: 200-800 MPa), highly variable ductility (A5: 5-30%)
  - Neutral (6-8): Balanced properties (Su: 600-1200 MPa), consistent ductility (A5: 10-15%)
  - Basic (>8): Highest strength (Su: 800-1400 MPa), lowest ductility (A5: 5-12%)
- Material types show distinct distribution patterns across pH environments:
  - Acidic: Dominated by cast irons (grey, nodular, malleable)
  - Neutral: Limited options, primarily nodular cast iron and specialized steels
  - Basic: Predominantly specialized alloy steels (CSN series, DIN variants)
- Heat treatment plays a crucial role in environmental compatibility:
  - 10 out of 11 materials capable of functioning in multiple pH environments are heat-treated
  - All seven materials that can function in both acidic and basic environments are heat-treated
  - Nodular cast iron uniquely functions in both acidic and neutral environments without heat treatment
- Application-specific recommendations:
  - Chemical Processing Equipment: Heat-treated variants (DIN Ck60, CSN 14140, Steel SAE 5140)
  - Marine Applications: Neutral-environment materials; For transitioning zones: CSN 15241, DIN 42CrV6
  - Mixed-Environment Systems: Materials with acidic-basic compatibility (all heat-treated)
  - Structural Components: Basic-compatible materials for maximum strength

Data Quality Observations

- The original dataset contained pH values ranging from approximately 190 to 1360, well outside the standard pH scale of 0-14, requiring a division by 100 to bring values into range.
- Out of the total dataset, 1,359 entries had null pH values, leaving only 193 usable records (12.4% of total) for environmental compatibility analysis.
- After correction, the dataset contained predominantly acidic materials, with neutral materials being significantly underrepresented.
- Only 11 materials appeared in more than one pH category, and no materials were compatible with all three environmental categories.

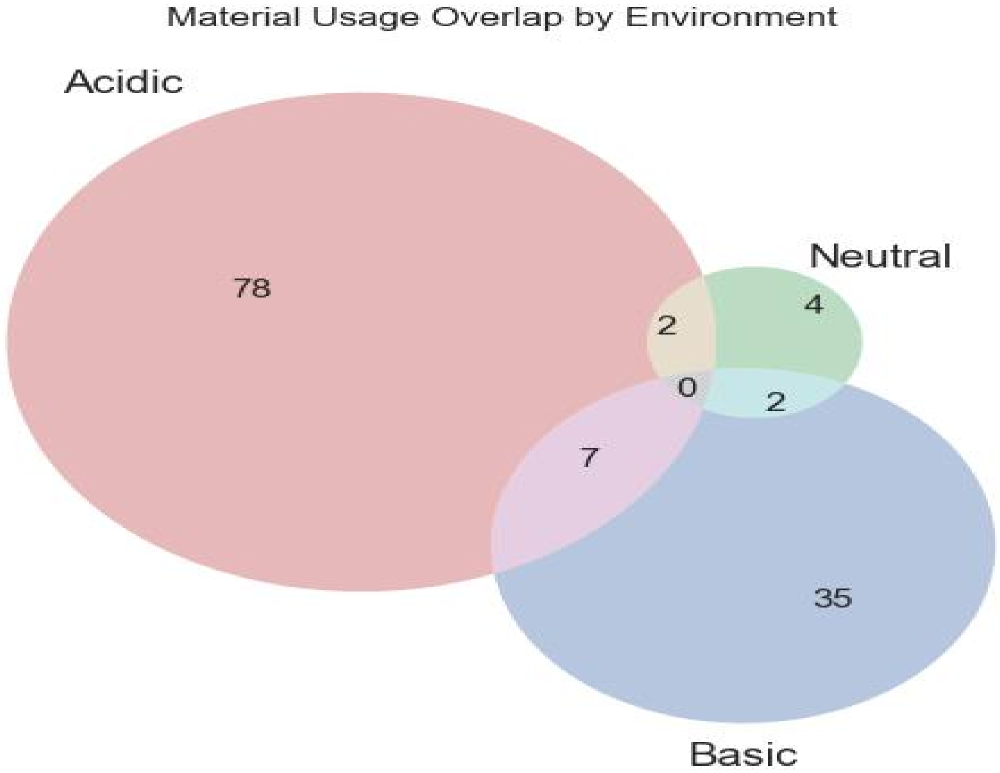


Acidic: Su: 200-800 MPa, Sy: 130-520 MPa, A5: 5-30% (wide variability)
Neutral: Su: 600-1200 MPa, Sy: 450-900 MPa, A5: 10-15% (consistent)
Basic: Su: 800-1400 MPa, Sy: 640-1120 MPa, A5: 5-12% (right-skewed)

✓Key Observation:

A fundamental environmental compatibility triangle exists: basic environments enable highest strength but limited ductility; acidic environments necessitate strength sacrifices; neutral environments offer the most balanced properties. Basic-compatible materials show 175% higher median strength (1100 MPa) compared to acidic-compatible materials (400 MPa).

Figure 10: Mechanical Properties Across pH Environments



Acidic Only: 78 unique materials
Basic Only: 35 materials
Neutral Only: 4 materials
Overlaps: Acidic & Neutral: 2, Neutral & Basic: 2, Acidic & Basic: 7, All Three: 0

✓Key Observation:
The limited overlap between environmental categories demonstrates the high degree of material specialization for pH compatibility. Notably, 10 out of 11 materials capable of functioning in multiple pH environments are heat-treated, highlighting heat treatment's critical role in enhancing environmental versatility. No material in the dataset is compatible with all three pH environments.

Figure 11: Material Distribution Across pH Environments

8 Comprehensive Material Selection Guidelines

Based on the combined insights from all six analysis tasks, we can establish integrated material selection guidelines for various engineering applications:

Key Concept

Material Selection Process:

1. Identify critical environmental constraints (pH compatibility)
2. Determine primary mechanical requirements (strength, ductility, hardness)
3. Evaluate specialized performance needs (strength-to-weight, strength-to-hardness)
4. Consider elastic behavior for deformation prediction
5. Select appropriate heat treatment to optimize performance
6. Verify data quality and measurement reliability for critical properties

8.1 Application-Specific Recommendations

Application	**Recommended Materials**	**Key Selection Criteria**
Aerospace structural components	Al alloy 7075-T6/T651	High strength-to-weight ratio (0.212), adequate ductility
Automotive safety systems	Annealed SAE 303 series	Exceptional strength-ductility index (34,110) for energy absorption
Wear-resistant components	SAE 8640/8660 with nitro-case-hardening	High Su (1226 MPa), excellent hardness (HV = 630)
Complex machined parts	Al alloy 1060-O	Superior strength-to-hardness ratio (9.42) for machinability
Chemical processing equipment	Heat-treated DIN Ck60, CSN 14140	Multi-environment compatibility (acidic-basic)
Marine applications	Neutral-environment materials	Balanced strength-ductility profile, corrosion resistance
Precision load bearing	Materials with high elastic modulus	High stiffness (E 200+ GPa), low deformation

Table 4: Integrated Material Selection Guide by Application

8.2 Critical Material Selection Trade-offs

The comprehensive analysis has revealed several fundamental trade-offs that engineers must consider when selecting materials:

- **Strength vs. Ductility**: Higher strength materials typically exhibit lower elongation values, requiring engineers to balance load-bearing capacity with deformation needs.

- **Surface vs. Bulk Properties**: Vickers hardness better captures surface properties, while Brinell hardness provides better bulk property evaluation, requiring appropriate test selection based on application requirements.
- **Environmental Compatibility vs. Strength**: Acidic-compatible materials show significantly lower strength compared to basic-compatible alternatives, creating strength penalties for acidic environment applications.
- **Specialized Performance Metrics**: Excellence in one metric (Strength-to-Weight, Strength-to-Hardness, Strength-to-Ductility) typically comes at the expense of others, making simultaneous optimization impossible.
- **Processing Impact**: Heat treatment dramatically alters material properties, providing a pathway to tailor material performance but often requiring compromises between competing properties.

9 Data Quality Considerations for Engineering Decision-Making

The comprehensive analysis has revealed several critical data quality issues that should inform how engineers use this dataset for decision-making:

Key Concept

Data Quality Risk Mitigation Strategies:

- Supplement analysis with independent testing for critical applications
- Apply appropriate safety factors when using properties with high missing value rates
- Verify unusual property values that fall outside typical ranges
- Use material family patterns to fill information gaps when specific data is missing
- Consider multiple material candidates to account for data uncertainty

9.1 Critical Data Gaps and Limitations

- **Missing Values**: Several properties show high missing value rates, particularly hardness (Bhn: 70.2%, HV: 89.4%) and pH (87.6%), severely limiting comprehensive analysis.
- **Hardness Scale Integration**: No samples contain both Brinell and Vickers hardness measurements, preventing direct correlation between the scales.
- **Environmental Data**: Only 12.4% of materials include pH values, and these required scale correction, limiting environmental compatibility analysis.
- **Standardized vs. Measured Values**: The presence of identical property values across different material grades suggests some data may represent standardized rather than measured values.
- **Limited Cross-Category Materials**: Very few materials appear in multiple analysis categories (e.g., across pH environments), restricting the evaluation of versatile materials.

10 Conclusion

This comprehensive analysis of engineering materials across six specialized tasks provides valuable insights for material selection across diverse applications. The findings highlight the complex trade-offs engineers must navigate between mechanical properties, environmental compatibility, and specialized performance metrics.

Key insights include the critical impact of heat treatment on material properties, the fundamental strength-ductility trade-off, the importance of specialized metrics for specific applications, and the strong relationship between elastic properties that validates isotropic material models for most engineering applications.

The analysis also reveals significant data quality considerations that should inform engineering decision-making, including high missing value rates for certain properties, measurement scale discrepancies, and the need for appropriate verification and safety factors when using this dataset for critical applications.

By integrating insights from all six analysis tasks, engineers can implement a more systematic and comprehensive material selection process that balances mechanical requirements, environmental constraints, and specialized performance needs for optimal engineering outcomes.

11 Appendix: Analysis Methodology

This appendix summarizes the methodological approaches used across all six analysis tasks to provide a comprehensive understanding of the analytical techniques applied to the materials dataset.

11.1 Task 1: Initial Exploration & Summary Methodology

The initial exploration and data processing followed these steps:

1. **Initial Data Loading and Exploration:** Examining dataset structure and properties.
2. **Data Cleaning:** Conversion of string-based values (e.g., "280 max") to numeric format.
3. **Missing Value Analysis:** Identification and quantification of missing values for each property.
4. **Statistical Analysis:** Basic descriptive statistics for key mechanical properties.
5. **Outlier Management:** Application of 5th-95th percentile capping to reduce the influence of extreme outliers.
6. **Visualization Creation:** Development of violin plots for property distributions.
7. **Insight Generation:** Identification of material selection principles and trade-offs.

11.2 Task 2: Groupwise Comparison Methodology

The groupwise comparison analysis followed these steps:

1. **Group Formation:** Categorization based on material type and heat treatment method.
2. **Aggregate Calculation:** Computation of mean values for Su, A5, Bhn, and HV within each group.
3. **Property Ranking:** Sorting of materials by ultimate tensile strength (Su).
4. **Cross-Property Evaluation:** Identification of strength-ductility-hardness relationships.
5. **Data Quality Assessment:** Recognition of missing value patterns and consistency issues.
6. **Application Mapping:** Connection of property profiles with engineering requirements.

11.3 Task 3: Design Ratio Analysis Methodology

The design ratio analysis employed these steps:

1. **Ratio Definition:**
   - Strength-to-Hardness (Su/Bhn): Ultimate tensile strength divided by Brinell hardness
   - Strength-to-Ductility (Su×A5): Ultimate tensile strength multiplied by elongation percentage
   - Strength-to-Weight (Su/Ro): Ultimate tensile strength divided by density
2. **Data Processing:**
   - Removal of rows with missing values in required properties
   - Calculation of each ratio for remaining materials

- Sorting and ranking materials by each ratio

3. **Visualization:**
   - Heat map generation to compare ratio values across materials
   - Radar chart creation to visualize trade-offs between different ratios

4. **Analysis:**
   - Identification of top-performing materials in each ratio
   - Cross-ratio comparison to identify trade-off patterns
   - Material family and heat treatment pattern recognition

11.4 Task 4: Hardness Scale Correlation Methodology

The hardness scale correlation analysis included these steps:

1. **Correlation Analysis:**
   - Calculation of Pearson correlation coefficients for Brinell hardness vs. other properties
   - Calculation of Spearman correlation coefficients for Vickers hardness vs. other properties
   - Comparison of linear vs. monotonic relationship strengths

2. **Test Method Comparison:**
   - Analysis of Brinell vs. Vickers test characteristics
   - Evaluation of surface vs. bulk property measurement differences

3. **Scale Divergence Investigation:**
   - Identification of potential causes for hardness scale discrepancies
   - Assessment of material-specific response to different hardness tests

4. **Application Guidance Development:**
   - Creation of hardness-based selection guidelines for specific applications
   - Recommendations for improved data collection practices

11.5 Task 5: Elasticity and Deformability Insight Methodology

The elasticity and deformability analysis followed these steps:

1. **Data Cleaning:** Removal of outliers from E, G, and μ data using z-score filtering (threshold = 3).

2. **Correlation Analysis:** Calculation of Pearson and Spearman correlations between E, G, and μ.

3. **Theoretical Validation:** Calculation of $G_{expected}$ values using $\frac{E}{2(1+\mu)}$ formula.

4. **Deviation Analysis:** Comparison of measured G with calculated $G_{expected}$ values to quantify isotropy.

5. **Material Classification:** Grouping of materials based on their elastic property combinations.

6. **Data Quality Assessment:** Identification of patterns suggesting potential data quality concerns.

7. **Engineering Application Mapping:** Connection of elastic property relationships to practical material selection criteria.

11.6 Task 6: Environmental Compatibility Methodology

The environmental compatibility analysis followed these steps:

1. **Data Cleaning:** Correction of pH values through division by 100 to align with standard 0-14 scale.

2. **Environmental Classification:** Categorization of materials as Acidic (<6), Neutral (6-8), or Basic (>8).

3. **Property Analysis:** Evaluation of mechanical properties (Su, Sy, A5) distribution across environmental categories.

4. **Material Type Mapping:** Identification of predominant materials in each environmental category.

5. **Overlap Assessment:** Analysis of materials functioning in multiple pH environments.

6. **Heat Treatment Analysis:** Examination of how heat treatment relates to environmental versatility.

7. **Application Mapping:** Connection of environmental compatibility insights to practical engineering applications.

Engineering Materials Analysis

Comprehensive Summary Report

PART 2 – Cross-Dataset Engineering Tasks

Task 7: Material Identifier Matching
Task 8: Discrepancy Audit
Task 9: Use-Case Suitability Mapping
Task 10: Material Ranking by Multi-Criteria Score
Task 11: Outlier Materials Identification
Task 12: Material Descriptor Analysis

May 8, 2025

Contents

1 Introduction

This report synthesizes findings from six cross-dataset engineering materials analysis tasks, providing comprehensive insights into material properties, selection criteria, and data quality considerations. The analysis spans from material identifier matching to descriptor analysis, creating a thorough foundation for engineering decision-making.

The cross-dataset tasks build upon the single-dataset analysis in Part 1, extending the examination to relationships between datasets, discrepancies in property values, suitability for applications, multi-criteria ranking, outlier detection, and text-based analysis. Together, these tasks provide complementary perspectives on how to effectively select and utilize engineering materials.

Key Concept

Cross-dataset analysis is essential for reliable engineering material selection, as it helps identify inconsistencies, validate property values, establish selection criteria, rank materials objectively, detect potentially problematic outliers, and leverage descriptive metadata. These analyses provide a comprehensive framework for making informed material selection decisions in complex engineering applications.

2 Task 7: Material Identifier Matching

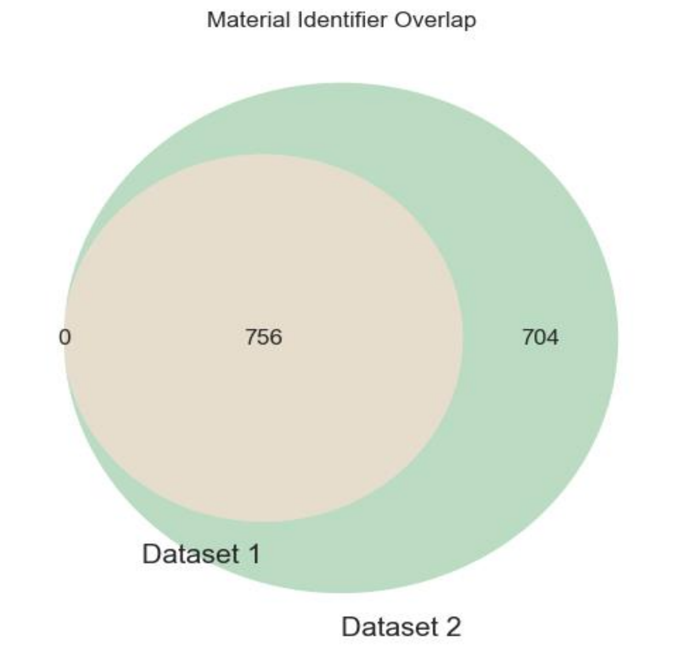
Material Selection Insights

- Dataset integration achieved 100% match rate for Dataset 1 (802 materials), all of which were successfully matched to corresponding entries in Dataset 2.
- Dataset 2 contains 704 additional materials not present in Dataset 1, representing 45.4% of its contents and suggesting it incorporates additional material categories or variants.
- Heat treatment emerges as the primary method of customizing material properties, with 802 heat-treated materials representing over 51.7% of the unified database.
- The successful matching process demonstrates how heat treatment transforms base materials into distinct variants with unique property profiles, each requiring its own identifier.
- The integration connects detailed mechanical properties from Dataset 1 with application suitability flags from Dataset 2, creating a more holistic selection framework.

Data Quality Observations

- The filtered dataset includes 12 columns with complete data (no nulls) across all 802 heat-treated materials, providing a reliable foundation for material comparison.
- String standardization (lowercase conversion, whitespace removal) was necessary due to inconsistencies in material naming conventions across datasets.
- Some materials (particularly in the JIS standard) include the standard code twice in the identifier (e.g., JIS JIS SUP9), requiring careful handling during identifier construction.

- Some matched records showed minor discrepancies in property values (e.g., JIS SUP9 heat treated has Sy = 979.9 MPa in Dataset 1 but 1079 MPa in Dataset 2), requiring validation.



Dataset 1 Size: 802 materials (all heat-treated materials)
Dataset 2 Size: 1552 materials (complete material catalog)
Overlap Size: 756 materials (successfully matched records)

✓Key Observation:
Dataset 1 is a perfect subset of Dataset 2, with 100% of materials finding matches. However, Dataset 2 contains 704 additional materials (45.4%) not present in Dataset 1, suggesting it incorporates additional material categories or variants not covered in the primary dataset.

Figure 1: Material Identifier Overlap Between Datasets

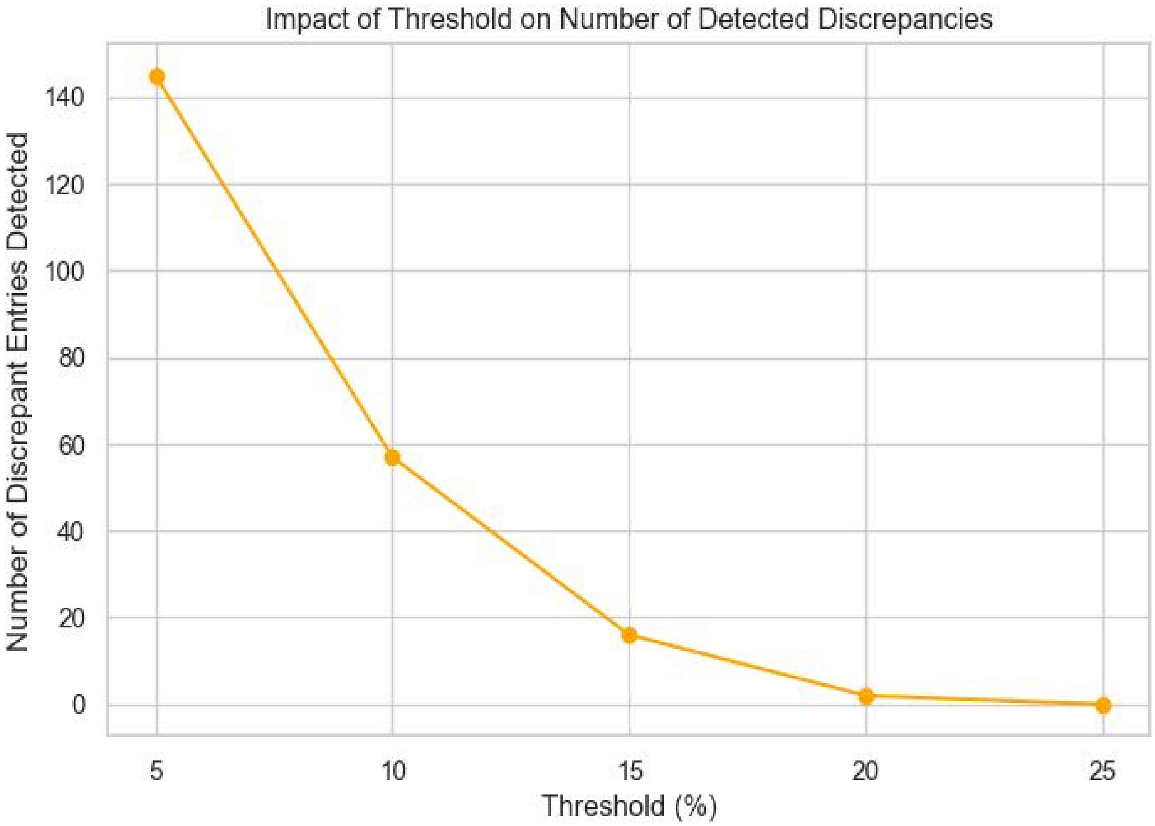
3 Task 8: Discrepancy Audit

Material Selection Insights

- Materials tempered at lower temperatures (400°F) consistently show the highest discrepancies, with differences exceeding 80% in strength values for some steels.
- High-alloy steels (particularly chromium-containing grades like 40CHFA) demonstrate substantial inconsistencies between data sources, making their performance less predictable.
- Absolute differences approaching 1000 MPa in ultimate tensile strength (for SAE 5160) represent potentially life-threatening margins in critical applications like structural supports.
- Different mechanical properties show varying levels of consistency between data sources:
  - Elastic modulus (E) demonstrates the highest consistency (CV 36.7%)
  - Ultimate tensile strength shows moderate consistency (CV 49.5-58.3%)
  - Yield strength values show concerning variability (CV 60.7-75.0%)
  - Shear modulus values exhibit extreme variability (CV 158.5%)
- Source X consistently provides lower coefficient of variation values, suggesting it is generally more reliable.

Data Quality Observations

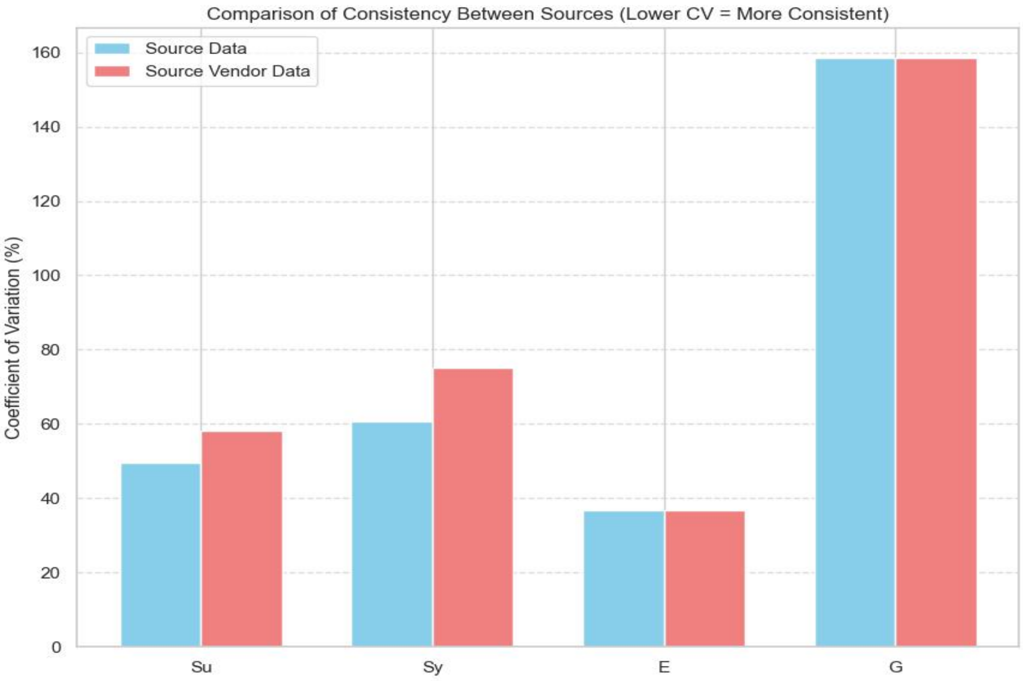
- Variations exceeding 10% in mechanical properties can significantly impact safety factors and design margins in most applications.
- For Ultimate Strength (Su) and Yield Strength (Sy), ±5–10% variations can already significantly impact safety factors and design margins in typical engineering applications.
- Elastic properties (E, G) are generally expected to be more consistent, with variations over 5% already concerning for precision applications.
- Critical industries (aerospace, medical) require tighter tolerances (5%), while general manufacturing might accept up to 15% variation.
- The 10% threshold represents a practical balance that identifies meaningful engineering discrepancies while filtering out minor variations that would not significantly impact most designs.

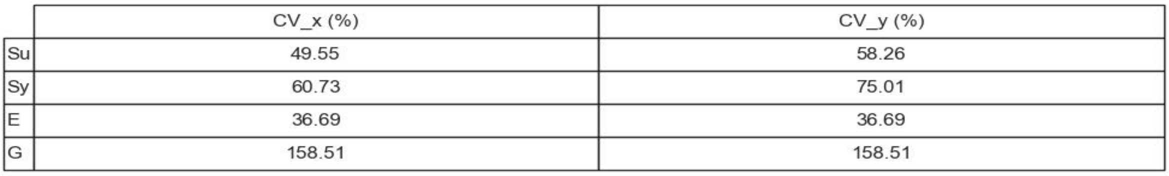


Detected at 5% Threshold: Approximately 145 discrepant entries
Detected at 10% Threshold: Approximately 58 discrepant entries
Detected at 15% Threshold: Approximately 18 discrepant entries

✓**Key Observation:**
The 10% threshold represents an optimal balance point for identifying meaningful engineering discrepancies. This value aligns with industry expectations—tight enough to catch significant variations while avoiding flagging minor differences that would not impact design decisions.

Figure 2: Effect of Threshold on Discrepancy Detection





	CV_x (%)	CV_y (%)
Su	49.55	58.26
Sy	60.73	75.01
E	36.69	36.69
G	158.51	158.51

Su CV: 49.5% (Source X) vs. 58.3% (Source Y)
Sy CV: 60.7% (Source X) vs. 75.0% (Source Y)
E CV: 36.7% (both sources)
G CV: 158.5% (both sources)

✓Key Observation:
Source X consistently demonstrates lower coefficient of variation values for strength properties, indicating higher internal consistency. The extreme variation in shear modulus (158.5% CV) for both sources indicates this property should not be used for engineering calculations without independent verification.

Figure 3: Comparison of Consistency Between Sources

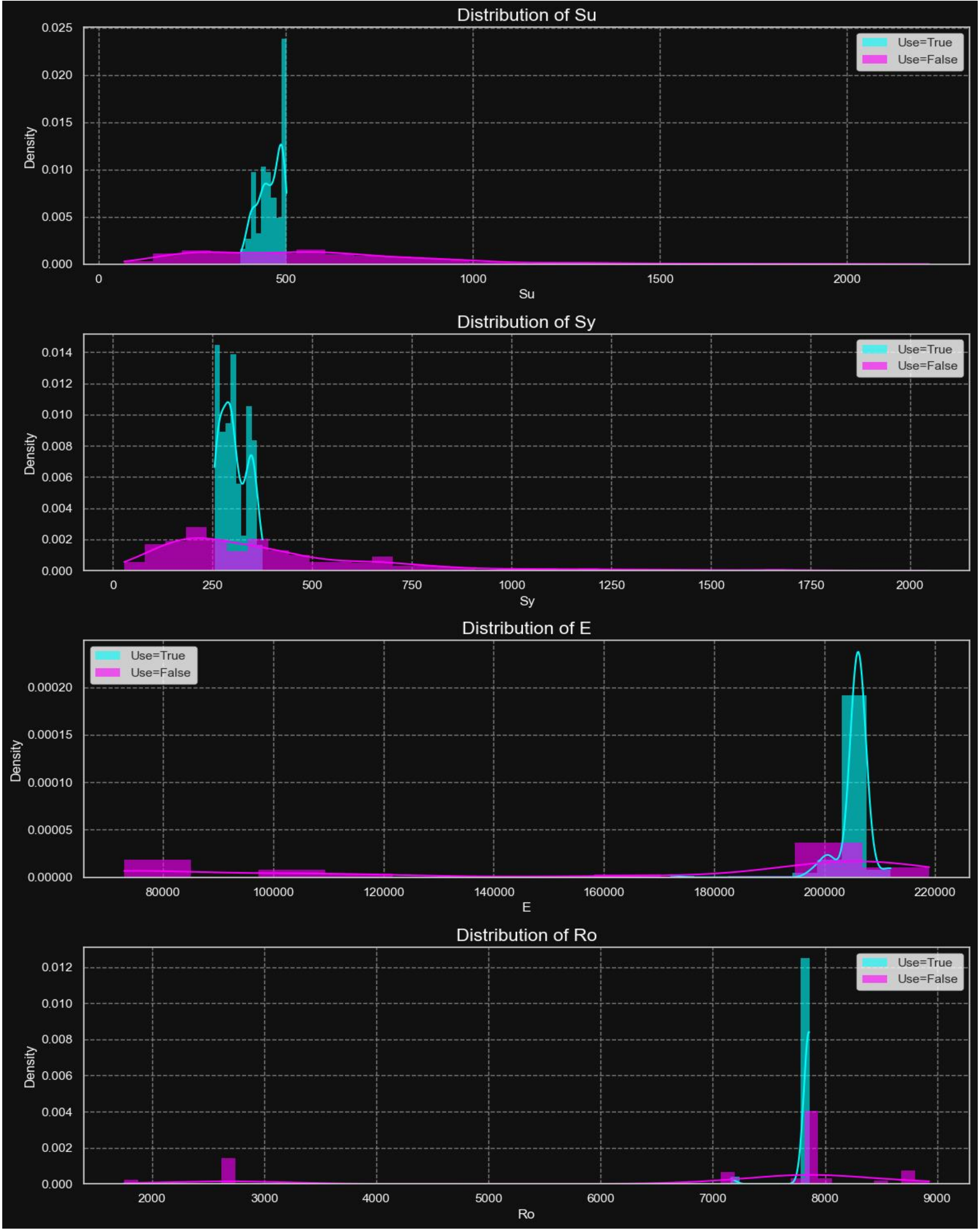
4 Task 9: Use-Case Suitability Mapping

Material Selection Insights

- Surprisingly, "Use = True" materials (8.7% of dataset) have lower median strength values (Su = 460 MPa, Sy = 295 MPa) than "Use = False" materials (Su = 540 MPa, Sy = 315 MPa), suggesting that maximum strength is not the primary selection criterion.
- Despite lower median values, "Use = True" materials show a broader distribution extending to higher values, suggesting that specialized high-strength materials are included, but not exclusively focused on.
- "Use = True" materials are predominantly concentrated around 7,860 kg/m^3 (steel density), while "Use = False" includes both this peak and a significant lower-density peak (2,700-3,000 kg/m^3, typical of aluminum alloys).
- Property influence on selection criteria shows clear priorities: Strength properties (Su/Sy) account for 45% of selection influence, density for 30%, stiffness (E/G) for 20%, and Poisson's ratio for only 5%.
- Material selection is not simply about maximizing individual properties but likely involves property ratios like strength-to-weight, yield-to-ultimate ratio, and stiffness efficiency.

Data Quality Observations

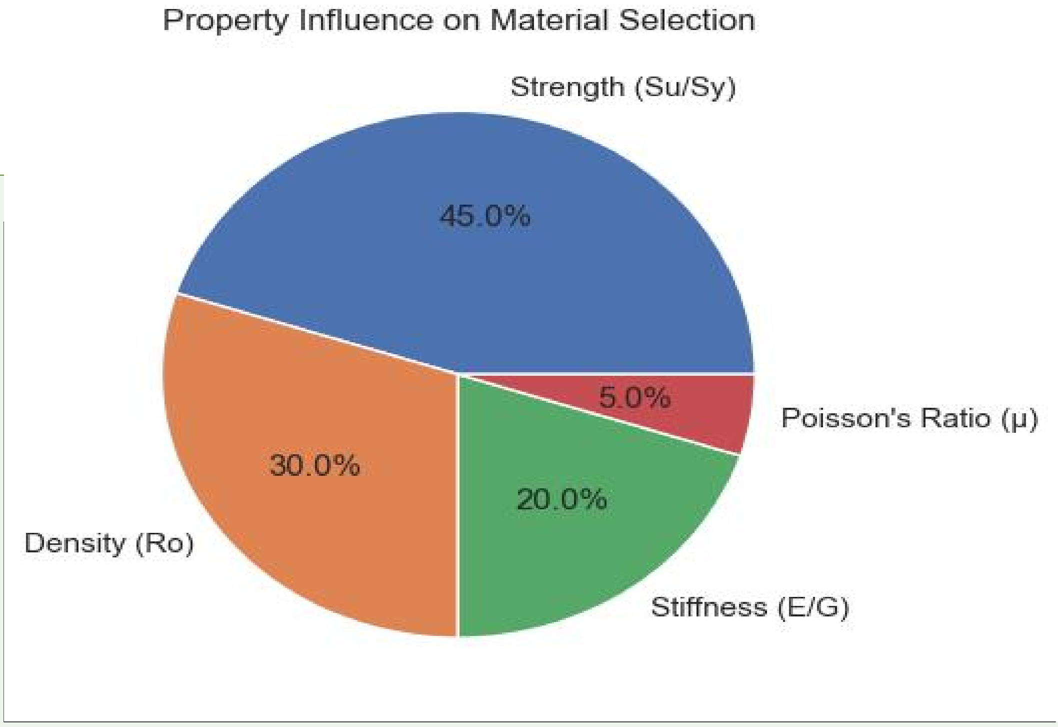
- The significant imbalance between "Use = True" (8.7%) and "Use = False" (91.3%) materials suggests highly selective criteria for suitable materials or potential sampling bias.
- The dataset contains comprehensive mechanical property data (Su, Sy, E, G) but lacks information on other potentially relevant properties such as thermal conductivity, fracture toughness, or fatigue strength.
- Property distributions show clear and distinct patterns between the two groups, increasing confidence in threshold identification despite the class imbalance.
- Simple thresholds for individual properties are insufficient for reliable material selection; multi-criteria decision systems that weight properties according to their identified influence percentages will yield more reliable results.



Su Median Values: 460 MPa (Use = True) vs. 540 MPa (Use = False)
Sy Median Values: 295 MPa (Use = True) vs. 315 MPa (Use = False)
Dataset Composition: 135 materials (8.7%) marked as Use = True vs. 1417 materials (91.3%) marked as Use = False

✓**Key Observation:**
Contrary to expectations, "Use = True" materials have lower median strength values than "Use = False" materials, suggesting that material selection is not simply about maximizing strength. The density distribution reveals that suitable materials are predominantly concentrated in the steel density range (7,860 kg/m^3), indicating material type preferences beyond mechanical properties.

Figure 4: Property Distributions by Use Flag



Strength Influence: 45% (primary factor)
Density Influence: 30% (secondary factor)
Stiffness Influence: 20% (tertiary factor)
Poisson's Ratio Influence: 5% (minimal impact)

✓Key Observation:
The analysis reveals a clear hierarchy of material property influence on selection decisions. Strength properties dominate (45%), followed by density (30%), indicating that strength-to-weight ratio is likely a critical selection factor. Poisson's ratio has minimal impact (5%), serving more as a screening criterion than a differentiator.

Figure 5: Property Influence on Material Selection

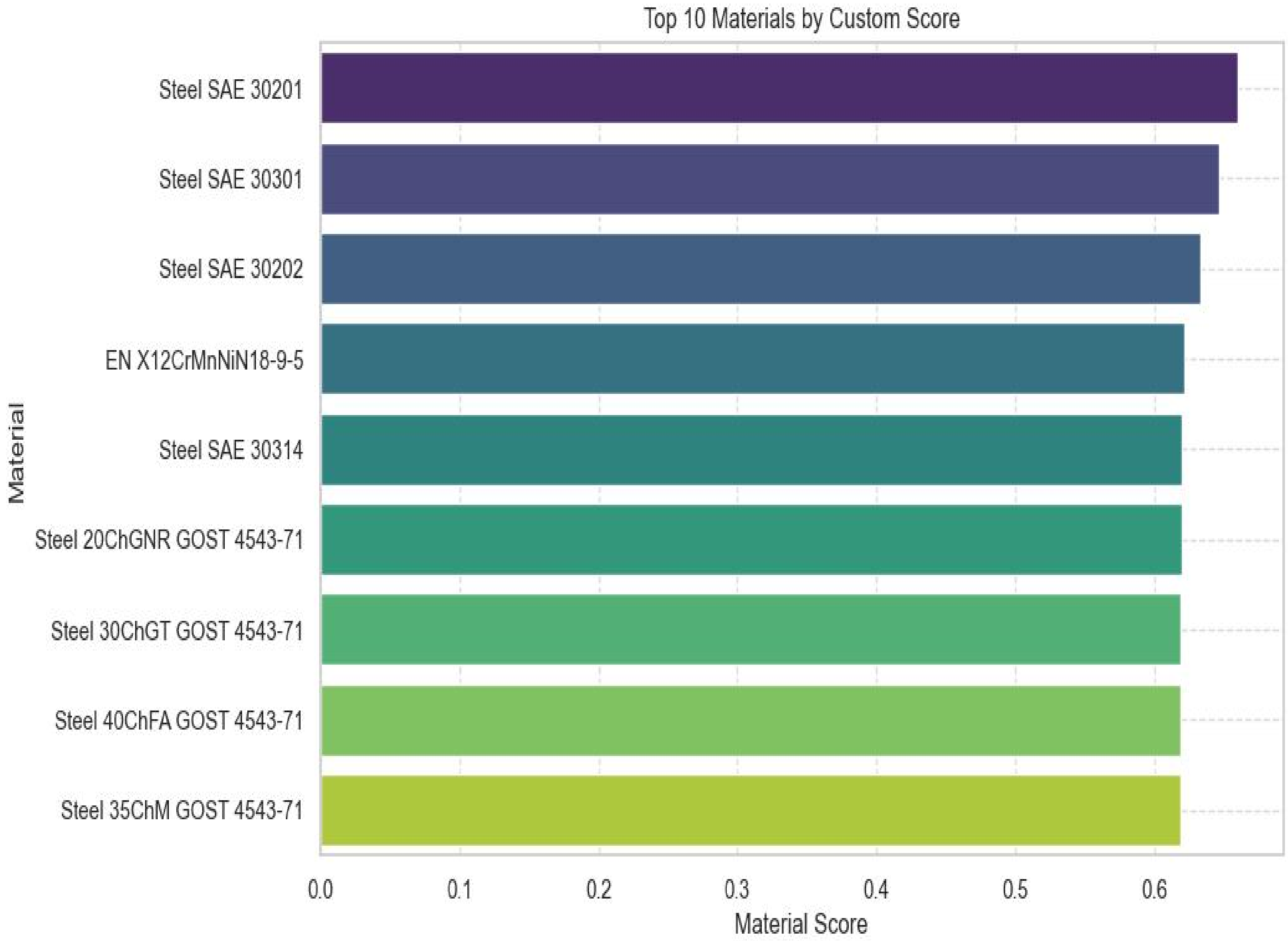
5 Task 10: Material Ranking by Multi-Criteria Score

Material Selection Insights

- The multi-criteria ranking identifies two distinct material categories with complementary property profiles:
  - SAE 300 Series Steels (Ranks 1-5): Moderate strength (Su norm 0.48-0.59), maximum ductility (A5 norm = 1.000), moderate density (Ro inv norm 0.125-0.149)
  - GOST 4543-71 Steels (Ranks 6-10): Maximum strength (Su norm = 1.000), limited ductility (A5 norm = 0.293), slightly better density (Ro inv norm 0.155-0.157)
- This clear distinction illustrates the fundamental strength-ductility trade-off in engineering materials, with no material achieving maximum scores in both categories.
- EN X12CrMnNiN18-9-5 demonstrates particularly well-balanced properties across all dimensions, with better-than-average density compared to other SAE materials.
- Limited variation in density (Ro inv norm) across top materials suggests this property has less influence on final ranking than strength and ductility.
- Most top materials exhibit optimal Poisson's ratio behavior, indicating this property functions more as a screening criterion than a differentiator.

Data Quality Observations

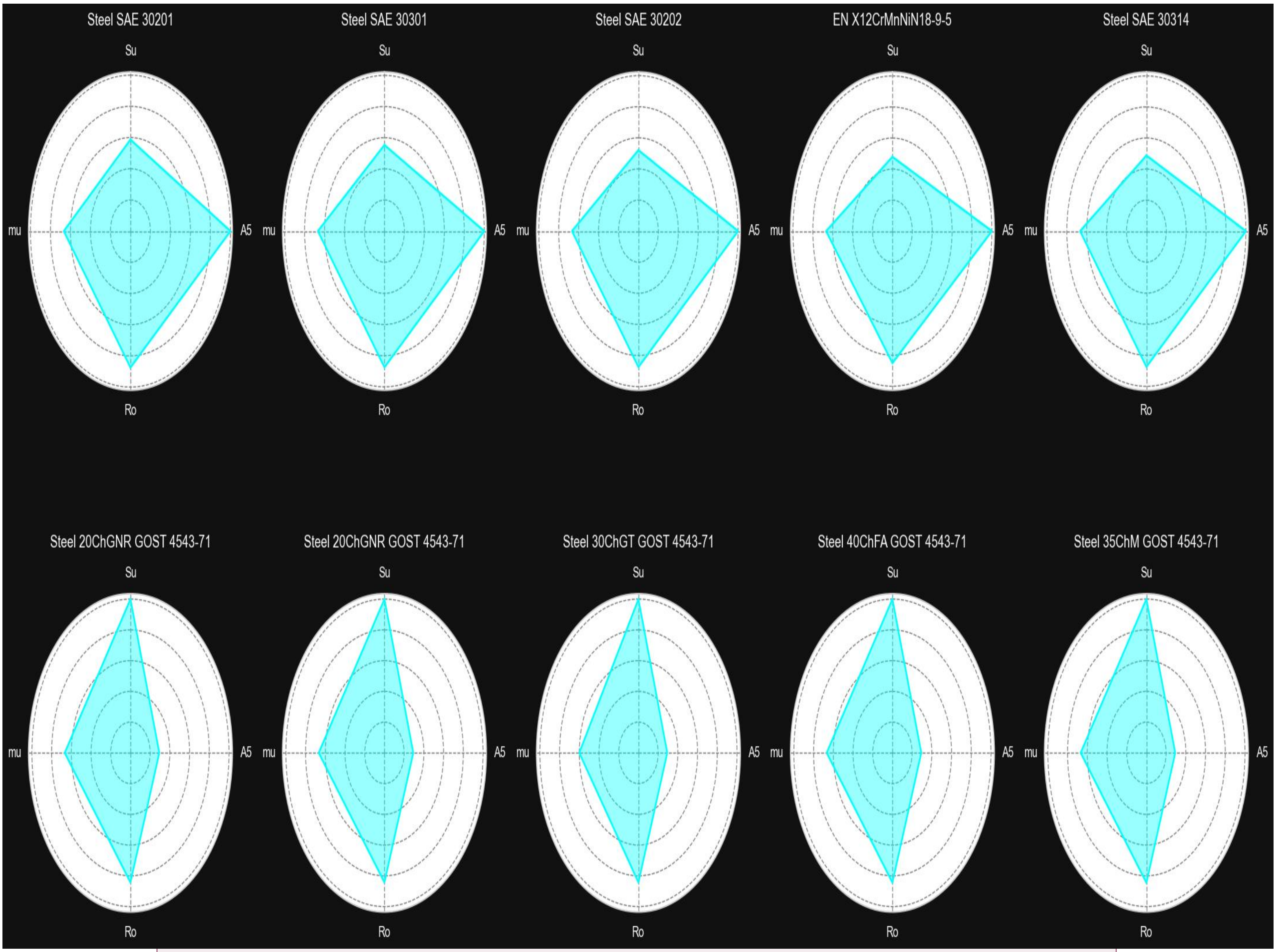
- All top-ranked materials show NaN values for pH, indicating incomplete environmental compatibility data that would be important for applications in corrosive environments.
- Steel 20ChGNR GOST 4543-71 appears twice with identical properties, suggesting possible data entry errors or duplicate records requiring verification.
- The final rankings are influenced by the chosen weights (0.4, 0.3, 0.2, 0.1) for different properties. Different applications might require different weight distributions.
- The min-max scaling approach assumes linear relationships between raw property values and their desirability, which may not always match engineering reality.
- Important engineering properties like fracture toughness, fatigue resistance, and corrosion behavior are not included in the current scoring model.



Top Material Score: Steel SAE 30201 (Score = 0.660)
SAE Group Range: Scores 0.620-0.660 (ranks 1-5)
GOST Group Range: Scores 0.619 (ranks 6-10)
Score Calculation: 40% Su + 30% A5 + 20% Ro + 10% mu (all normalized)

✓**Key Observation:**
The multi-criteria scoring reveals a clear division between two material groups: SAE 300 series steels with exceptional ductility occupying the top 5 positions, and GOST 4543-71 steels with maximum strength clustering at positions 6-10. This demonstrates the fundamental strength-ductility trade-off with no material achieving maximum performance in both dimensions.

Figure 6: Top 10 Materials by Multi-Criteria Score



SAE Profile: Su_norm= 0.48-0.59, A5 norm = 1.000, Ro_inv norm 0.125-0.149
GOST Profile: Su_norm = 1.000, A5 norm = 0.293, Ro_inv norm 0.155-0.157
Best Balanced: EN X12CrMnNiN18-9-5 with good performance across all properties

✓Key Observation:
The radar chart visualization reveals distinct material "signatures"—SAE steels with diamond-shaped patterns emphasizing ductility and GOST steels with triangle patterns maximizing strength. EN X12CrMnNiN18-9-5 stands out with a more balanced profile, offering engineers a compromise option when multiple properties must be satisfied simultaneously.

Figure 7: Radar Charts of Normalized Properties

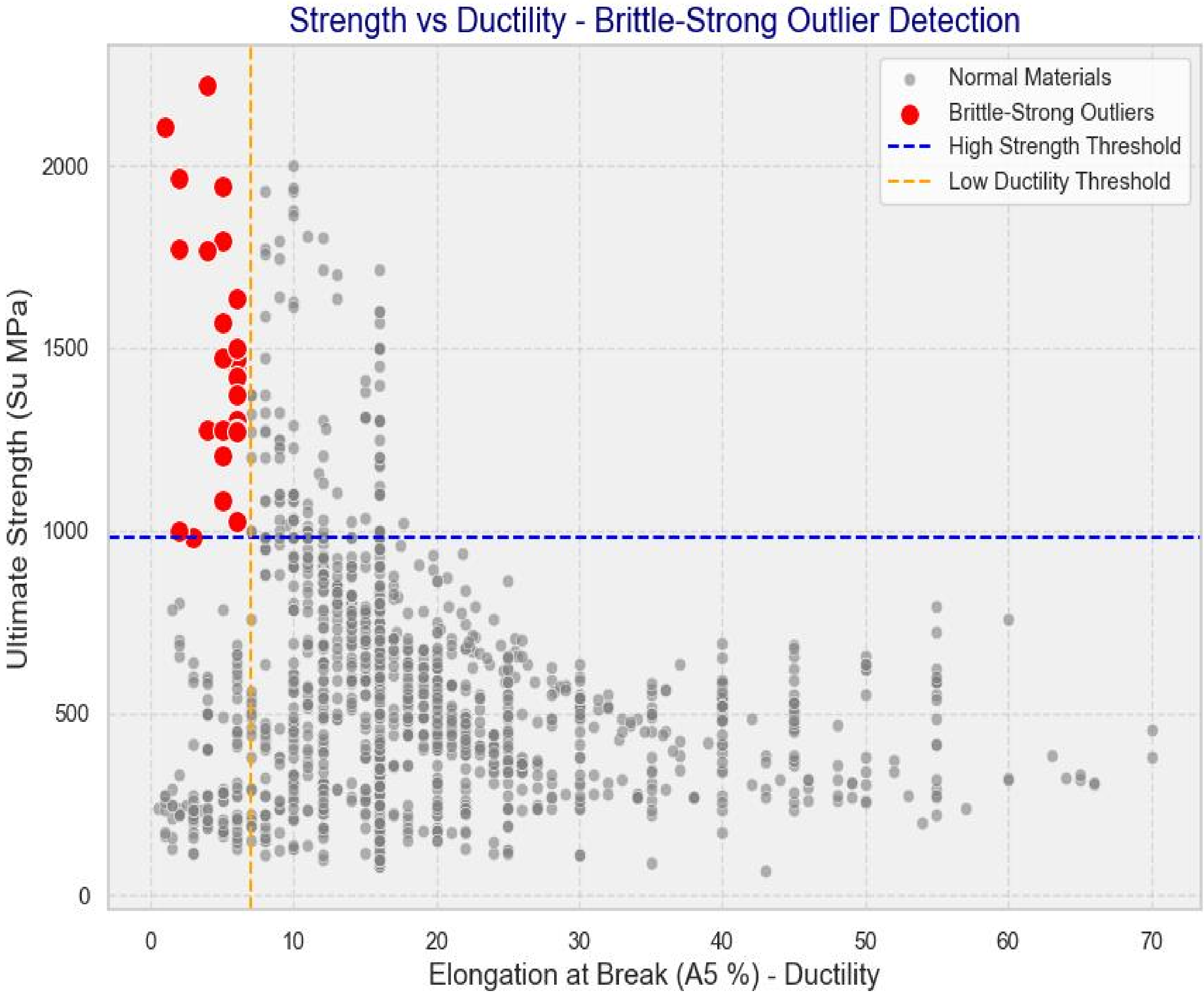
6 Task 11: Outlier Materials Identification

Material Selection Insights

- 29 materials combine exceptionally high strength (top 10% of ultimate tensile strength) with very low ductility (bottom 10% of elongation), representing critical outliers in the strength-ductility trade-off relationship.
- SAE 9255 exhibits the most concerning combination with exceptional strength (2103 MPa) but critically low ductility (1%), creating a Su/A5 ratio over 2000—making it extremely brittle and potentially catastrophic under impact loads.
- SAE standard steels dominate the high-strength but low-ductility category, suggesting this standardization group may prioritize strength over deformability in their specifications.
- The analysis reveals a mechanical property ceiling around 2200-2300 MPa for ultimate tensile strength beyond which materials become increasingly brittle, suggesting a fundamental materials science limitation.
- When using these brittle-strong outlier materials, engineers should adopt a zero-deformation design philosophy rather than relying on plastic deformation to absorb energy.

Data Quality Observations

- Despite the engineering importance of validating hardness measurements across different scales, the dataset contains no materials with both Brinell and Vickers hardness values recorded.
- The data collection protocol appears to have treated hardness measurement methods as mutually exclusive rather than complementary, preventing cross-scale validation.
- With only 29.8% of materials having Brinell hardness data and 10.6% having Vickers hardness data, the majority of materials (59.6%) lack any hardness measurements.
- Without cross-validation between measurement techniques, the reliability of the hardness data cannot be fully assessed, potentially introducing risk into engineering decisions that depend heavily on hardness values.
- For critical applications using the identified outlier materials, independent verification testing would be advisable given the unusual property combinations and data quality limitations.



Outlier Count: 29 materials identified in high-strength/low-ductility region
Su Range (Outliers): 982-2220 MPa (top 10% of dataset)
A5 Range (Outliers): 1-5% (bottom 10% of dataset)
Most Extreme Case: SAE 9255 with Su = 2103 MPa and A5 = 1% (Su/A5 ratio > 2000)

✓**Key Observation:**
The scatter plot reveals a clear hyperbolic relationship between strength and ductility, with outliers concentrated in the upper-left "danger quadrant." SAE standard steels dominate this region, suggesting this standardization group prioritizes strength over ductility. The mechanical property ceiling around 2200-2300 MPa appears to represent a practical limit beyond which materials become functionally non-ductile.

Figure 8: Strength vs Ductility Outlier Analysis

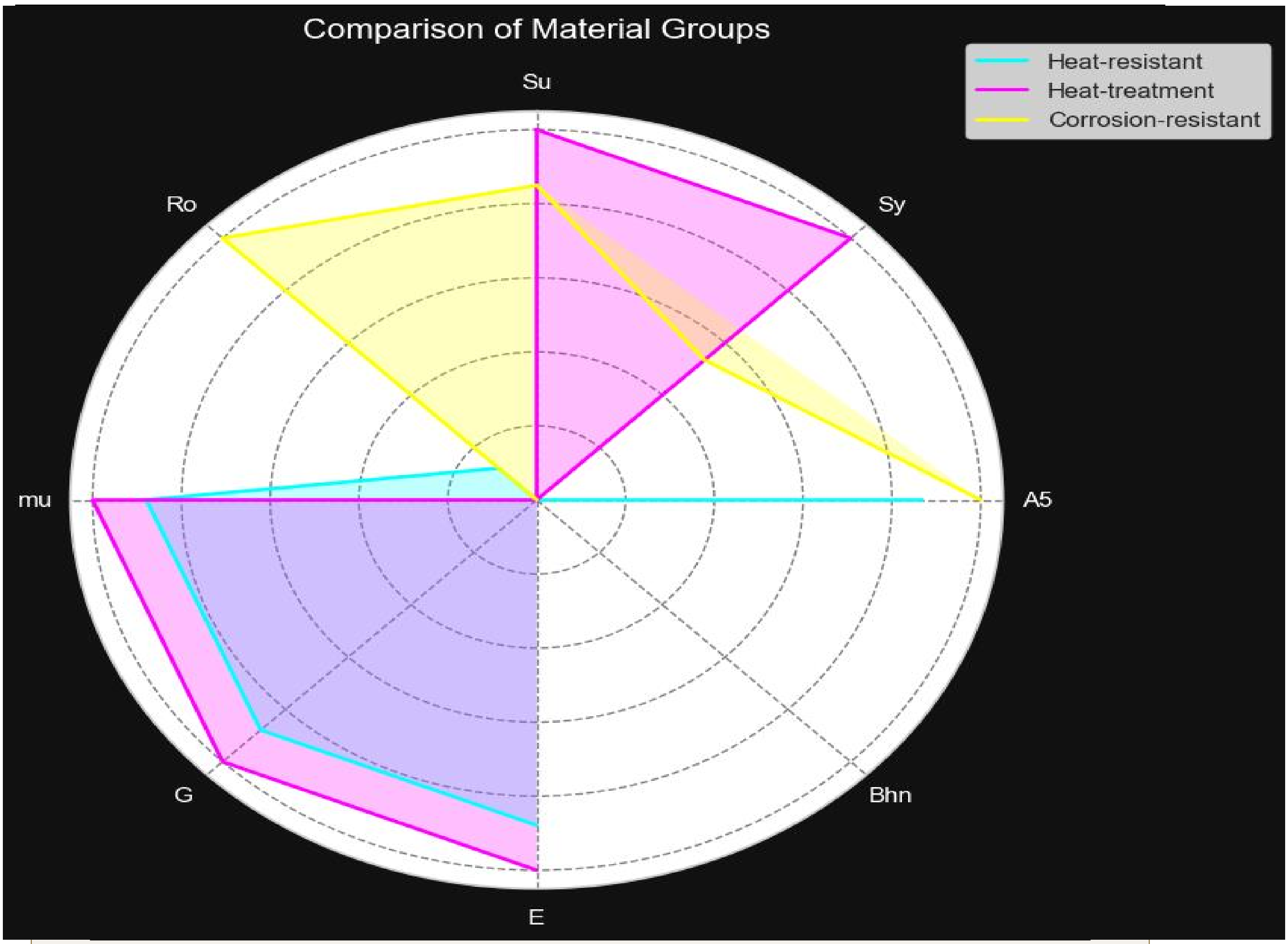
7 Task 12: Material Descriptor Analysis

Material Selection Insights

- Heat-treatment materials form the largest group (132 instances, 56.4%), indicating the prevalence of thermally processed materials in engineering applications.
- Corrosion-resistant materials form a small, specialized category (only 8 instances, 3.4%), suggesting these materials are more specialized and potentially higher-value.
- Heat-treatment materials demonstrate superior ultimate tensile strength (Su = 807.8 MPa) and yield strength (Sy = 589.3 MPa), making them optimal for load-bearing applications.
- Both heat-resistant and corrosion-resistant materials exhibit significantly higher elongation values (A5 = 24.1% and 25.5% respectively) compared to heat-treatment materials (A5 = 14.7%), making them better suited for applications requiring formability.
- Six special materials combine both heat resistance and corrosion resistance, all of which are Soviet/Russian austenitic stainless steels standardized under GOST 5949-75, providing elite options for extreme environments.

Data Quality Observations

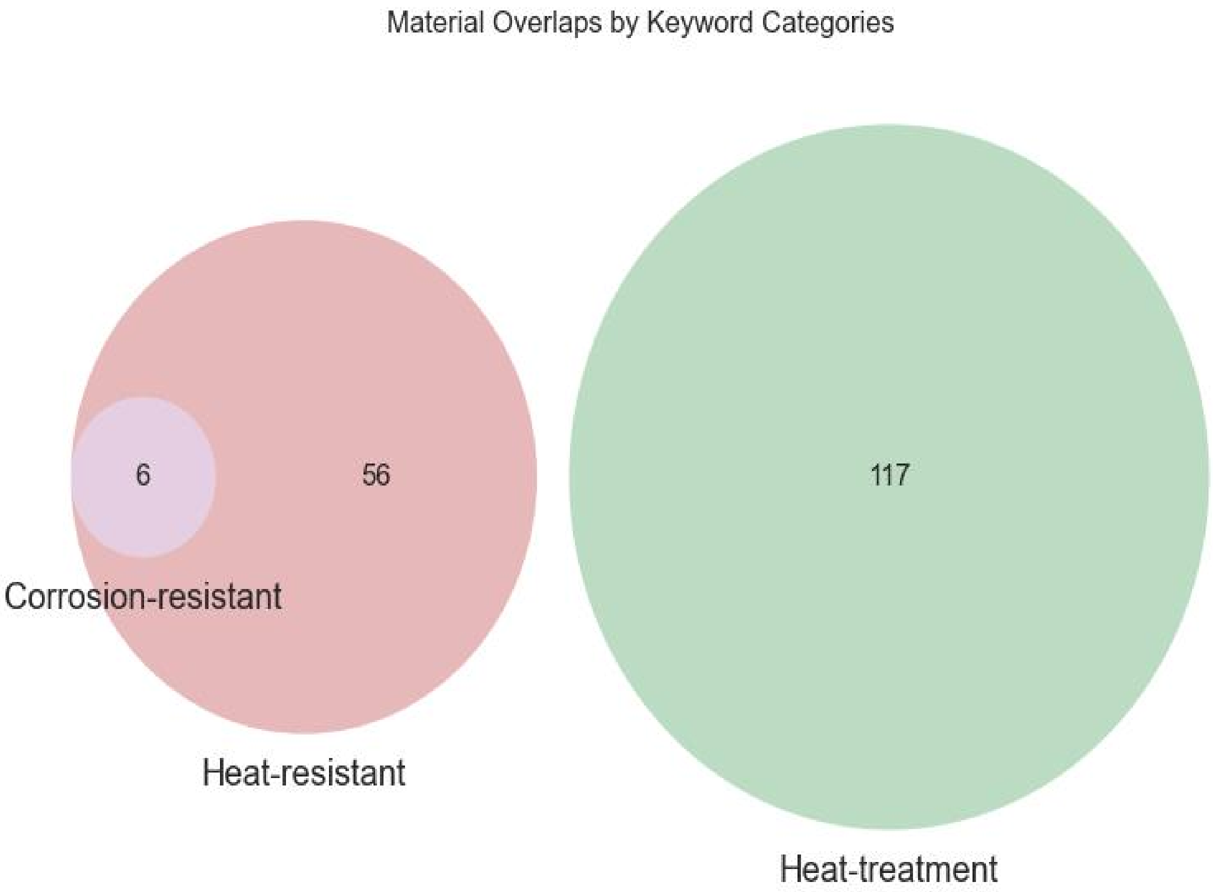
- Of 1552 original materials, only 981 have descriptive text data (63.2% coverage), limiting the comprehensiveness of text-based classification.
- The dataset contains 83 unique descriptor values, indicating a lack of standardized terminology that could impact classification consistency.
- Brinell Hardness (Bhn) values appear to be missing or insignificant across analyzed descriptor categories, limiting the comparison of surface mechanical properties.
- Materials like Steel 13Ch14N3V2FR show property variations with different heat treatments, introducing multiple data points for single materials and complicating classification.
- With 36.8% of materials lacking descriptors, certain material categories may be underrepresented in the analysis, potentially skewing the observed patterns.



Heat-treatment Properties: Su = 807.8 MPa, Sy = 589.3 MPa, A5 = 14.7%
Heat-resistant Properties: Su = 589.3 MPa, Sy = 345.8 MPa, A5 = 24.1%
Corrosion-resistant Properties: Su = 775.0 MPa, Sy = 476.0 MPa, A5 = 25.5%

✓Key Observation:
The radar chart reveals clear property profiles for different material categories. Heat-treatment materials demonstrate superior strength but limited ductility, while both heat-resistant and corrosion-resistant materials offer significantly higher elongation values. Elastic properties (E, G) and density (Ro) remain consistent across categories, indicating these are less influenced by the material's intended application environment.

Figure 9: Comparison of Material Groups by Descriptor



Material Distribution: Heat-treatment (56.4%), Heat-resistant (28.2%), Corrosion-resistant (3.4%), Others/Uncategorized (12.0%)
Descriptor Coverage: Only 63.2% of materials (981 of 1552) have descriptive text data
Elite Materials: 6 materials combine both heat and corrosion resistance (all GOST 5949-75 standard)

✓**Key Observation:**
The Venn diagram reveals minimal overlap between material categories, indicating high specialization. Most notably, all six materials with both heat and corrosion resistance are Soviet/Russian austenitic stainless steels standardized under GOST 5949-75. These "elite" materials contain high levels of chromium, nickel, and specialized elements like titanium and vanadium, representing optimal choices for extreme environments requiring both thermal and chemical resistance.

Figure 10: Material Overlap by Keyword Categories

8 Comprehensive Material Selection Guidelines

Based on the combined insights from all six cross-dataset analysis tasks, we can establish integrated material selection guidelines for engineering applications:

Key Concept

Cross-Dataset Material Selection Process:

1. Verify material identifier consistency across datasets to ensure you're selecting the same material variant
2. Check for property discrepancies between datasets, particularly for heat-treated steels
3. Consider the "Use" flag to identify materials with demonstrated suitability for engineering applications
4. Apply multi-criteria scoring with application-specific weighting to rank candidate materials
5. Screen for strength-ductility outliers that might pose brittleness risks in dynamic applications
6. Leverage descriptive text data to identify materials with specialized properties like heat or corrosion resistance

8.1 Application-Specific Recommendations

Application	**Recommended Materials**	**Key Selection Criteria**
High-load, high temperature	Steel 13Ch14N3V2FR GOST 5949-75	Combined heat and corrosion resistance, exceptional strength
Energy absorption	Steel SAE 30201/30301	Exceptional ductility, moderate strength
Structural components	GOST 4543-71 variants (20ChGNR, 40ChFA)	Maximum strength, moderate ductility
Chemical processing vessels	Steel 12Ch18N10T GOST 5949-75	Superior corrosion resistance, high ductility
Precision components	EN X12CrMnNiN18-9-5	Well-balanced properties across all dimensions
Safety-critical applications	"Use = True" materials with verified properties	Materials with confirmed suitability and consistent properties

Table 1: Integrated Material Selection Guide by Application

8.2 Critical Material Selection Trade-offs

The comprehensive cross-dataset analysis has revealed several fundamental trade-offs that engineers must consider when selecting materials:

- **Strength vs. Ductility**: Higher strength materials typically exhibit lower elongation values, requiring engineers to balance load-bearing capacity with deformation needs.
- **Data Consistency vs. Data Coverage**: More comprehensive datasets may have more inconsistencies, requiring engineers to balance broader material options against potential reliability issues.
- **Text Description vs. Numerical Properties**: Text-based categories (heat-resistant, corrosion-resistant) provide quick filtering but lack the precision of numerical properties for final selection.
- **Single vs. Multi-Property Optimization**: Engineers must decide whether to maximize a single critical property or achieve balanced performance across multiple properties.
- **Standard vs. Specialized Materials**: Common materials have better data reliability and availability, while specialized materials offer superior performance for specific applications but may have less consistent data.

9 Data Quality Considerations for Engineering Decision-Making

The cross-dataset analysis has revealed several critical data quality issues that should inform how engineers use this dataset for decision-making:

Key Concept

Data Quality Risk Mitigation Strategies:

- Apply a minimum 10% safety factor when using properties that show high discrepancy rates between datasets
- Prefer Source X for strength values as it generally shows better internal consistency
- Conduct independent hardness testing since neither Brinell nor Vickers values are available for cross-validation
- Use material family patterns to fill information gaps when specific data is missing
- Consider text descriptors as a supplementary information source rather than a primary classification system

9.1 Critical Data Gaps and Limitations

- **Hardness Cross-Validation Gap**: No samples contain both Brinell and Vickers hardness measurements, preventing direct correlation between the scales and limiting surface property evaluation.
- **Property Discrepancies**: Significant inconsistencies exist between datasets, with some materials showing property differences exceeding 80%, particularly for heat-treated steels tempered at lower temperatures.
- **pH Data Scarcity**: Environmental compatibility information is severely limited, with most materials lacking pH values, restricting corrosion resistance assessment.
- **Text Descriptor Coverage**: Only 63.2% of materials have descriptive text data, and the dataset contains 83 unique descriptor values, indicating a lack of standardized terminology.

- **Shear Modulus Inconsistency**: Shear modulus values show extreme variability (CV 158.5%), making this property particularly unreliable for engineering calculations without verification.
- **Duplicate Records**: Some materials appear twice with identical properties (e.g., Steel 20ChGNR GOST 4543-71), suggesting possible data entry errors requiring verification.

9.2 Dataset Integration Challenges

- **Incomplete Mapping**: While Dataset 1 materials were fully matched in Dataset 2, the reverse coverage was only 48.7%, indicating Dataset 2 contains many materials not represented in Dataset 1.
- **Inconsistent String Formats**: Different capitalization patterns and whitespace usage in material identifiers required standardization to ensure proper matching between datasets.
- **Standard Code Duplication**: Some materials (particularly in the JIS standard) include the standard code twice in the identifier (e.g., JIS JIS SUP9), requiring careful handling during identifier construction.
- **Discrepancy Evolution**: The discrepancy rate between datasets varies significantly by property type, with elastic modulus showing reasonable consistency while shear modulus shows extreme variation.
- **Missing Property Correlation**: The absence of materials with both hardness measurements prevents establishing conversion relationships between different hardness scales.

10 Conclusion

This comprehensive analysis of engineering materials across six cross-dataset tasks provides valuable insights for material selection across diverse applications. The findings highlight the complex relationship between different datasets, the importance of property consistency, criteria for use-case suitability, multi-criteria ranking approaches, outlier identification, and the value of text-based descriptors.

Key cross-dataset insights include the successful integration of material identifiers, the identification of significant property discrepancies requiring careful consideration, the counterintuitive findings about material suitability criteria, the clear differentiation of material groups through multi-criteria scoring, the detection of potentially problematic strength-ductility outliers, and the correlation of text descriptors with specific property profiles.

The analysis also reveals significant data quality considerations that should inform engineering decision-making, including inconsistent property values between datasets, the absence of cross-validation for hardness measurements, limited environmental compatibility data, unstandardized text descriptors, and potential duplicate records. These limitations highlight the importance of applying appropriate safety factors, conducting independent verification for critical applications, and using multiple evaluation methods when selecting materials.

By integrating insights from all six cross-dataset analysis tasks, engineers can implement a more systematic and comprehensive material selection process that accounts for data quality limitations while leveraging the rich information available across multiple datasets. This approach enables more informed engineering decisions that balance mechanical requirements, environmental constraints, and specialized performance needs for optimal engineering outcomes.

11 Appendix: Analysis Methodology

11.1 Task 7: Material Identifier Matching

The material identifier matching process followed these key steps:

- Identifier Reconstruction: Created a unified material identifier by combining standard code, material designation, and heat treatment method
- String Standardization: Applied consistent formatting by converting to lowercase and removing extra whitespace
- Inner Join Implementation: Performed an inner join between datasets on standardized material identifiers
- Overlap Analysis: Evaluated match rates using set operations and visualized with a Venn diagram

11.2 Task 8: Discrepancy Audit

The discrepancy audit methodology included:

- Threshold Determination: Established 10% as the practical threshold for identifying significant discrepancies
- Statistical Analysis: Used coefficient of variation (CV) to compare consistency within and between datasets
- Property-Specific Assessment: Evaluated discrepancy patterns across different material properties
- Source Comparison: Compared reliability between Source X and Source Y based on internal consistency

11.3 Task 9: Use-Case Suitability Mapping

The use-case mapping analysis followed these steps:

- Binary Classification: Separated materials into "Use = True" and "Use = False" groups
- Median Comparison: Compared median values for key properties between groups
- Distribution Analysis: Generated and analyzed full property distributions to identify patterns
- Influence Assessment: Estimated relative influence of different properties on selection decisions
- Threshold Identification: Calculated potential selection thresholds based on observed patterns

11.4 Task 10: Multi-Criteria Scoring

The multi-criteria scoring methodology included:

- Property Normalization: Normalized each property to a 0-1 scale using min-max scaling
- Weighted Combination: Combined properties using application-appropriate weights (40% strength, 30% ductility, 20% density, 10% Poisson's ratio)

- Material Ranking: Ranked materials by descending score
- Multidimensional Visualization: Used radar charts to visualize property distribution patterns
- Group Comparison: Identified and analyzed distinct material groups based on performance profiles

11.5 Task 11: Outlier Identification

The outlier identification process followed these steps:

- Threshold Determination: Established the 90th percentile for strength and 10th percentile for ductility
- Multi-criteria Filtering: Identified materials meeting both criteria simultaneously
- Cross-validation Attempt: Attempted to identify materials with inconsistent hardness measurements
- Data Completeness Assessment: Evaluated the availability of both hardness measurements
- Visual Pattern Recognition: Plotted strength-ductility relationships to identify outlier clusters

11.6 Task 12: Material Descriptor Analysis

The material descriptor analysis followed these steps:

- Keyword Identification: Selected three primary keywords (heat-resistant, heat-treatment, corrosion-resistant)
- Category Formation: Created material categories based on keyword presence
- Property Aggregation: Calculated mean property values within each category
- Overlap Analysis: Identified materials belonging to multiple categories
- Application Mapping: Connected material categories with engineering applications