Vanilla Steel

Case Assessment

Junior Data Scientist

Vanilla Steel is a digital platform revolutionizing metal trading. We process large volumes of supplier and buyer data to facilitate material trading. As we scale, we need data-driven insights to optimize our marketplace and improve analytics for our internal products.

Link to repository for download: https://shorturl.at/wvMYz

Assessment Objective

You will solve **two complementary data challenges**:

1. **Supplier Data Cleaning (Excel)** — build a unified inventory dataset from two messy supplier files.

2. **RFQ Similarity (RFQ list + Reference properties)** — enrich RFQs with grade properties and compute similarity between RFQs.

Vanilla Steel

Assessment Requirements

Scenario A — Supplier Data Cleaning

Interpretation of Data Content:

supplier_data1.xlsx

- **Quality/Choice:** Indicates the quality level (e.g., "1st", "2nd","3rd").
- **Grade**: The grade of the material (e.g., "C100S", "C200S").
- **Finish**: Surface treatment of the material (e.g., "ungebeizt","gebeizt").
- **Thickness (mm)**: Thickness of the material in millimeters.
- **Width (mm)**: Width of the material in millimeters.
- **Description**: Additional information or defects (e.g., "Sollmasse(Gewicht) unterschritten").
- **Gross weight (kg)**: The weight of the material in kilograms.
- **Quantity:** Number of units.

supplier_data2.xlsx

- **Material**: Describes the type of material (e.g., "DX51D +Z140").
- **Description**: Additional details (e.g., "Material is Oiled").
- **Article ID**: A unique identifier for the material.
- **Weight (kg)**: The weight of the material in kilograms.
- **Quantity**: Number of units.
- **Reserved**: Indicates whether the material is reserved or not.

Task A.1 — Clean & Join

Your goal:

- Clean and normalize both datasets (e.g., unify thickness/width formats, standardize names).
- Handle missing/inconsistent values.
- Join into a single table called **inventory_dataset**.
- Document your assumptions.

Deliverable: inventory_dataset.csv

Vanilla Steel

Scenario B — RFQ Similarity

Interpretation of Data Content

rfq.csv

- **id**: Unique identifier for each RFQ line.
- **grade**: Steel grade requested (e.g., “S235JR”).
- **grade_suffix**: Extra suffix or variant of the grade.
- **coating**: Requested coating (e.g., galvanized, oiled).
- **finish**: Requested surface finish.
- **surface_type**: Specific surface requirement.
- **surface_protection**: Whether protective coating is required.
  form: Shape/form (e.g., coil, sheet, bar, pipe).
  thickness_min / thickness_max: Requested thickness range.
- **width_min / width_max**: Requested width range.
- **length_min**: Minimum requested length.
- **height_min / height_max**: Requested height range.
- **weight_min / weight_max**: Weight range.
- **inner_diameter_min / inner_diameter_max**: Inner diameter range.
- **outer_diameter_min / outer_diameter_max**: Outer diameter range.
- **yield_strength_min / yield_strength_max**: Yield strength range.
- **tensile_strength_min / tensile_strength_max**: Tensile strength range.

reference_properties.tsv

A grade-level lookup table containing chemical, mechanical, and contextual properties. Columns include:

- **Grade/Material**: Steel grade name (e.g., *S235JR*).
- **UNS_No**: Unified Numbering System identifier (if available).
- **Steel_No**: European steel number (if available).
- **Standards**: Relevant specification standards (e.g., *EN 10025-2:2019*).
- **Carbon (C)**: Carbon content or maximum range.
- **Manganese (Mn)**: Manganese content or range.
- **Silicon (Si)**: Silicon content or range.
- **Sulfur (S)**: Maximum sulfur content.
- **Phosphorus (P)**: Maximum phosphorus content.
- **Chromium (Cr)**: Chromium content or range.
- **Nickel (Ni)**: Nickel content or range.
- **Molybdenum (Mo)**: Molybdenum content or range.

- **Vanadium (V)**: Vanadium content or range
- **Tungsten (W)**: Tungsten content or range.
- **Cobalt (Co)**: Cobalt content or range.
- **Copper (Cu)**: Copper content or range.
- **Aluminum (Al)**: Aluminum content or range.
- **Titanium (Ti)**: Titanium content or range.
- **Niobium (Nb)**: Niobium content or range.
- **Boron (B)**: Boron content or range.
- **Nitrogen (N)**: Nitrogen content or range.
- **Tensile strength (Rm)**: Range or typical tensile strength in MPa.
- **Yield strength (Re or Rp0.2)**: Yield strength in MPa.
- **Elongation (A%)**: Minimum elongation percentage at fracture.
- **Reduction of area (Z%)**: Optional measure of ductility (if provided).
- **Hardness (HB, HV, HRC)**: Hardness values in Brinell, Vickers, or Rockwell.
- **Impact toughness (Charpy V-notch)**: Impact energy at given temperature.
- **Fatigue limit**: Stress amplitude endurance limit (if available).
- **Creep resistance**: High-temperature creep resistance (if available).
- **Source_Pages**: Reference sources/pages.
- **Application**: Typical end-use applications (e.g., structural, high-strength).
- **Category**: Category grouping (e.g., *Structural Steel*, *High Strength Steel*).
- **Nb + V + Ti (Others)**: Summed micro-alloying element presence (if applicable).
- **Coating**: Any coating requirement specified at the grade level.

⚠️ **Notes for candidates**:

- Many values are **upper-bound (≤)**, lower-bound (≥), or **ranges (min–max)**.
- Some fields may be missing or inconsistent across grades.
- These are **not individual items** but **grade-level reference properties** to enrich RFQs.

Tasks

Task B.1 — Reference join & missing values (25 pts)

- Normalize grade keys (case, suffixes, aliases).

- Parse range strings into numeric min/max (and optionally mid).
- Join RFQs with reference. Handle:
    - RFQ grades missing in reference if any.
    - Missing values (choose keep-null, impute, or flag).

Task B.2 — Feature engineering (20 pts)

- **Dimensions**: Represent each dimension as an interval. For singletons, set min=max. Suggest one overlap metric (IoU, overlap ratio).
- **Categorical**: Define similarity as exact match (1/0) for coating, finish, form, surface_type.
- **Grade properties**: Use numeric midpoints of ranges. Ignore very sparse features if needed.
- Document briefly.

Task B.3 — Similarity calculation (30 pts)

- Define an aggregate similarity score between two RFQs. (e.g., weighted average of dimension overlap, categorical matches, grade similarity).
  Output **top-3 most similar RFQs per line (excluding self and exact matching)**.

Deliverable: top3.csv with columns [rfq_id, match_id, similarity_score].

Task B.4 — Pipeline & documentation (15 pts)

- Implement as a reproducible pipeline (script or notebook).
- Provide:
    - README.md with explanation.
    - top3.csv results.
    - A simple run.py (or notebook) to execute the flow.

Bonus / Stretch Goals (+20 pts)

- **Ablation analysis**: Compare similarity when dropping feature groups (dimensions only vs grade only) or adjusting weights.
- **Alternative metrics**: Try weighted cosine+jaccardi similarity vs IoU or other similarity measurement approaches. Describe your solution.
- **Clustering**: Group RFQs into families and provide a short interpretation.

Deliverables

- GitHub repository containing:
    - Scripts/notebooks
    - inventory_dataset.csv (Scenario A)

- top3.csv (Scenario B)

Submission Requirements

- Submission email: Please send your results to klaudia.pluta@vanillasteel.com
- Code Repository: Please submit your code via a Git repository link (e.g., GitHub, GitLab). Ensure that the repository is well-organized and includes a README file with instructions on how to set up and run the pipeline.
- Documentation: Include documentation that explains the tools, process, process clearly.

Evaluation Criteria

- supplier data cleaning & join (Task A.1) — 20
- Reference join & missing values (Task B.1) — 25
- Feature engineering (Task B.2) — 20
- Similarity definition (Task B.3) — 30
- Pipeline & reproducibility (Task B.4) — 15
- Reporting & clarity — 10

Additional Notes

- The assessment is designed to reflect real-world challenges that you might encounter at Vanilla Steel.
- **We do not require a 100% perfect solution.** While a fully functioning pipeline is ideal, the focus will also be on your problem-solving approach, optimization techniques, and how well you can communicate your solutions.
- If you encounter any issues or have questions during the assessment, please feel free to reach out.