

PhenoQC Quality Control Report

Source file: quality_metrics_input.csv

Summary

| Imputation Strategy | Mean |
|-------------------------|--------|
| Schema Validation Score | 33.33% |
| Missing Data Score | 97.44% |
| Mapping Success Score | 0.00% |
| Overall Quality Score | 43.59% |

Imputation Settings

| | |
|-----------------|--|
| Global Strategy | knn |
| Global Params | {'n_neighbors': 5, 'weights': 'uniform'} |
| Tuning Enabled | True |
| Best Params | {'n_neighbors': 3} |
| Tuning Score | 90.0750 (MAE) |

Data Quality Scores:

Schema Validation Score: 33.33%
Missing Data Score: 97.44%
Mapping Success Score: 0.00%
Overall Quality Score: 43.59%

Schema Validation Results

Format Validation: False
Duplicate Records: 2 issues found.
Conflicting Records: 2 issues found.
Integrity Issues: 2 issues found.
Referential Integrity Issues: No issues found.
Anomalies Detected: No issues found.

Invalid Mask: 3 issues found.
Accuracy Issues: 1 issues found.
Redundancy Issues: 4 issues found.
Traceability Issues: 3 issues found.
Timeliness Issues: 2 issues found.

Additional Quality Dimensions

Accuracy Issues: 1 issues found.

| row | column | value | minimum | maximum |
|-----|-----------|-------|---------|---------|
| 1 | Height_cm | -5 | 0 | None |

Redundancy Issues: 4 issues found.

| column_1 | column_2 | metric | value |
|-------------|------------------|-------------|--------------------|
| Height_cm | Cholesterol_mgdl | correlation | 1.0 |
| Weight_kg | Creatinine_mgdl | correlation | 0.9999999999999994 |
| BP_systolic | BP_diastolic | correlation | 1.0 |
| Height_cm | Cholesterol_mgdl | identical | 1.0 |

Traceability Issues: 3 issues found.

| row | issue |
|-----|----------------------|
| 0 | duplicate_identifier |
| 1 | duplicate_identifier |
| 2 | missing_identifier |

Timeliness Issues: 2 issues found.

| SampleID | Height_cm | Weight_kg | Cholesterol_mgdl | BP_systolic | BP_diastolic | Glucose_mgdl | Creatinine_mgdl | PrimaryPhenotype | VisitDate | SchemaViolationFlag | issue |
|----------|-----------|-----------|------------------|-------------|--------------|--------------|-----------------|------------------|------------|---------------------|-------------------------|
| 1.0 | 170 | 70 | 170 | 120 | 60 | 90 | 1.0 | HP:0001250 | 2025-07-30 | False | lag_exceeded |
| nan | 180 | 75 | 180 | 140 | 70 | 100 | 1.1 | HP:0000001 | not_a_date | True | missing_or_invalid_date |

Missing Data Summary

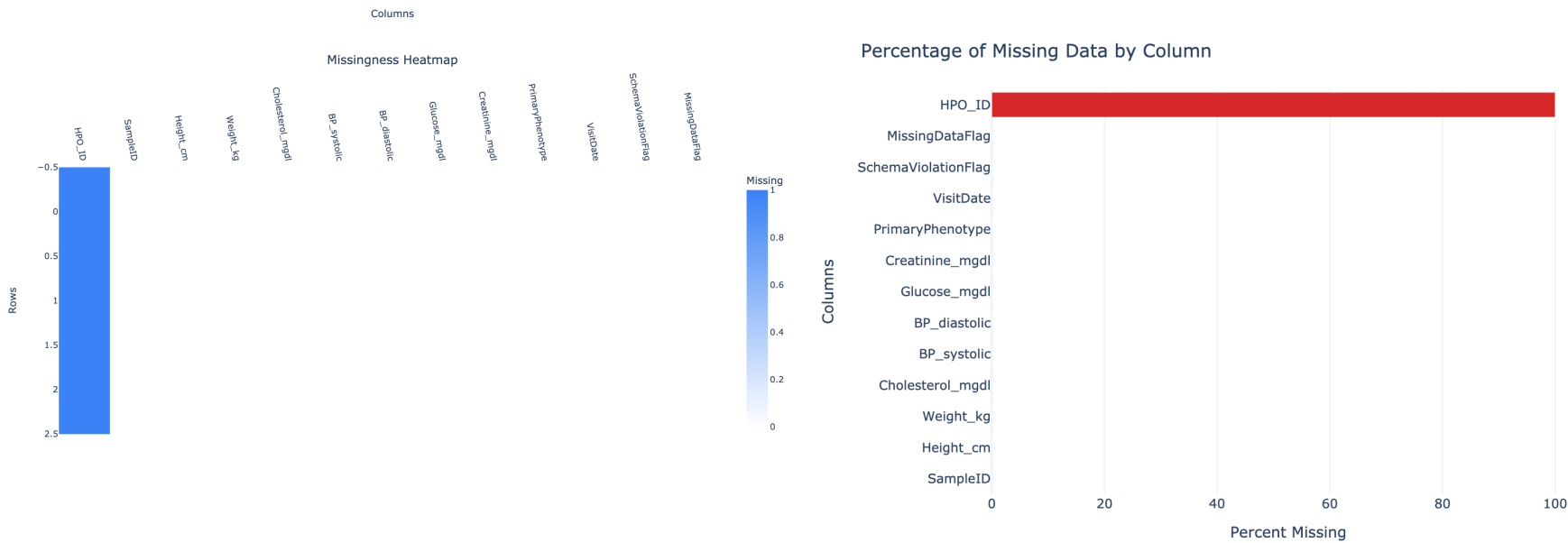
| Column | Missing Count |
|----------|---------------|
| SampleID | 1 |

Records Flagged for Missing Data: 1

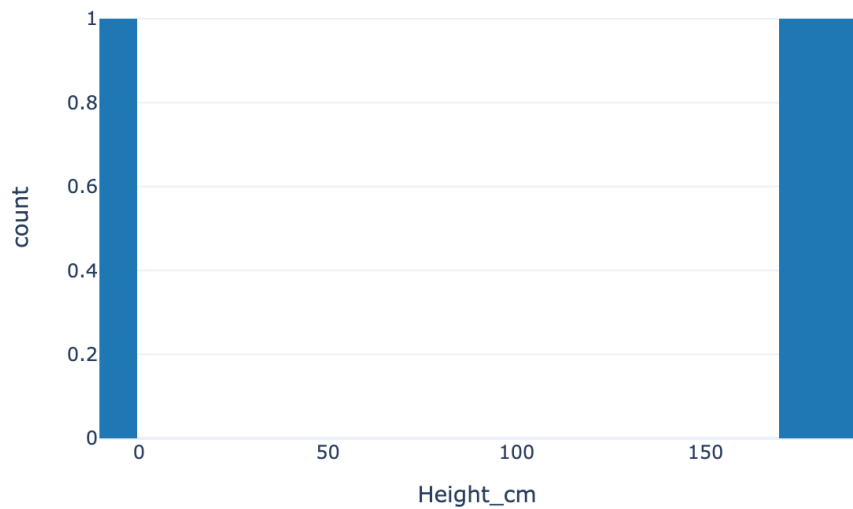
Ontology Mapping Summary

| Ontology | Total Terms | Mapped | Success Rate |
|----------|-------------|--------|--------------|
| HPO | 3 | 0 | 0.00% |

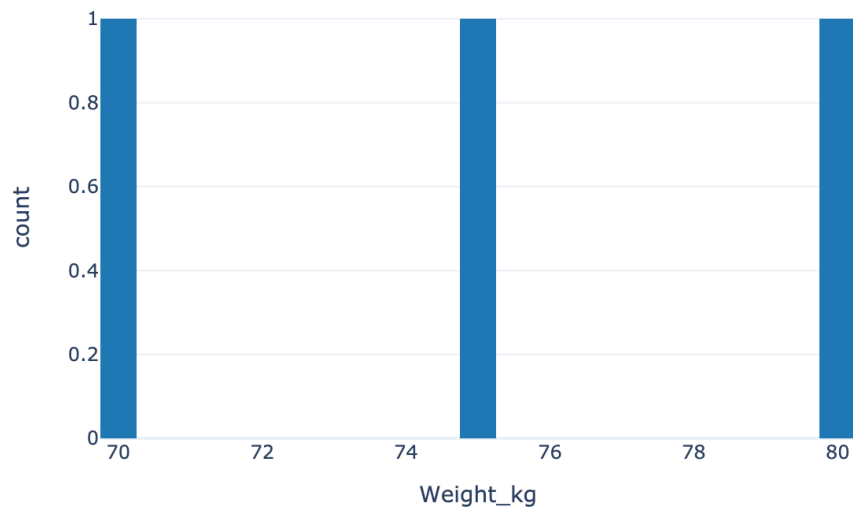
Visualizations



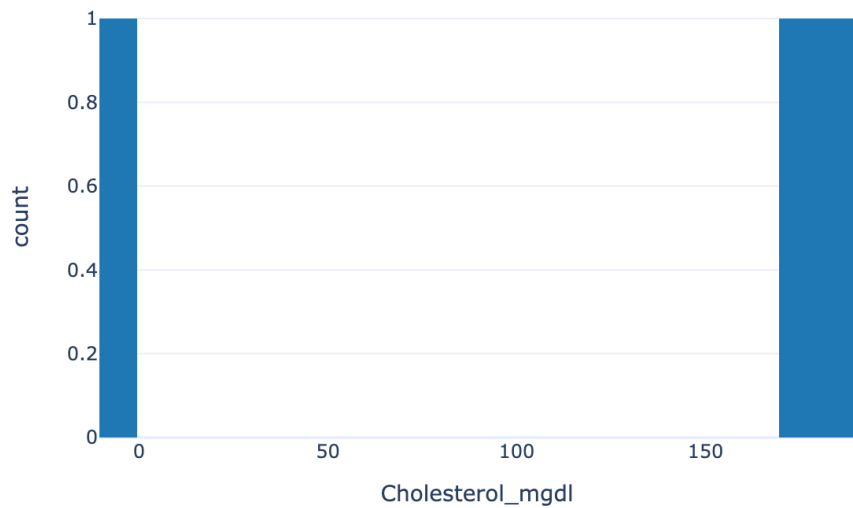
Distribution of Height_cm



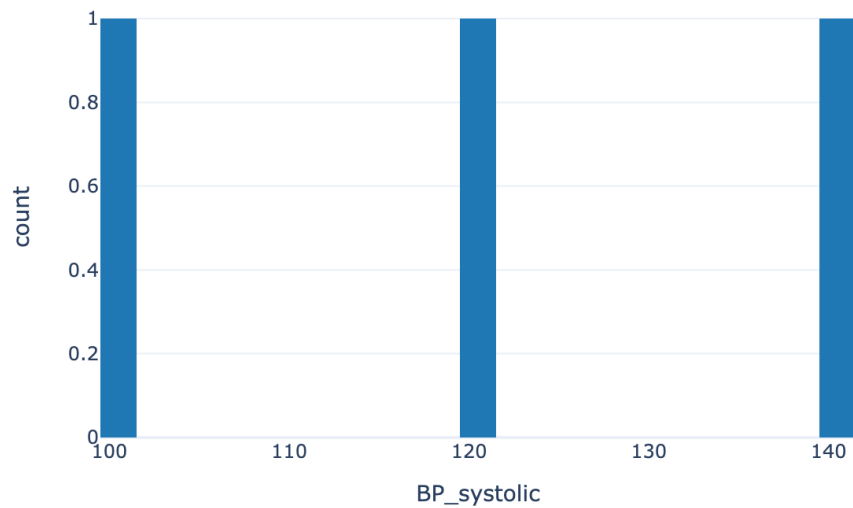
Distribution of Weight_kg



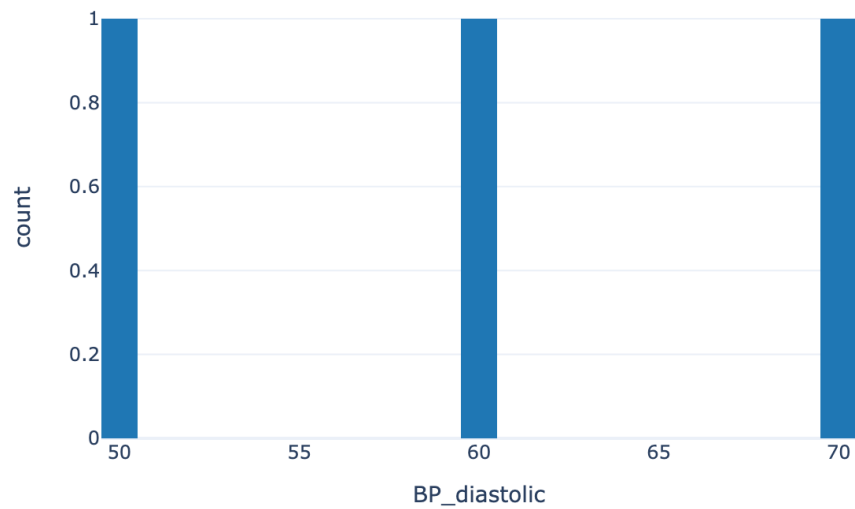
Distribution of Cholesterol_mgdl



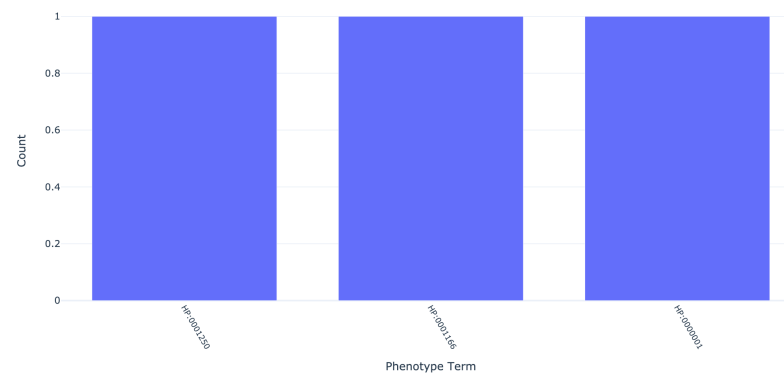
Distribution of BP_systolic



Distribution of BP_diastolic



Top 20 Most Common Terms in PrimaryPhenotype



Mapping Results: PrimaryPhenotype → HPO

