# PhenoQC Quality Control Report

**Source file:** e2e_medium_input.csv

---

## Summary

| | |
|---|---|
| **Imputation Strategy** | Mean |
| **Schema Validation Score** | 0.00% |
| **Missing Data Score** | 92.98% |
| **Mapping Success Score** | 100.00% |
| **Overall Quality Score** | 64.33% |

### Imputation Settings

| | |
|---|---|
| **Global Strategy** | knn |
| **Global Params** | {'n_neighbors': 5, 'weights': 'uniform'} |
| **Tuning Enabled** | True |
| **Best Params** | {'n_neighbors': 7} |
| **Tuning Score** | 11.2896 (MAE) |

## Data Quality Scores:

**Schema Validation Score:** 0.00%
**Missing Data Score:** 92.98%
**Mapping Success Score:** 100.00%
**Overall Quality Score:** 64.33%

## Schema Validation Results

**Format Validation:** False
**Duplicate Records:** 2 issues found.
**Conflicting Records:** 2 issues found.
**Integrity Issues:** 1000 issues found.
**Referential Integrity Issues:** No issues found.
**Anomalies Detected:** 17 issues found.

**Invalid Mask:** 1000 issues found.

---

## Additional Quality Dimensions

**Accuracy Issues:** No issues found.
**Redundancy Issues:** No issues found.
**Traceability Issues:** 2 issues found.

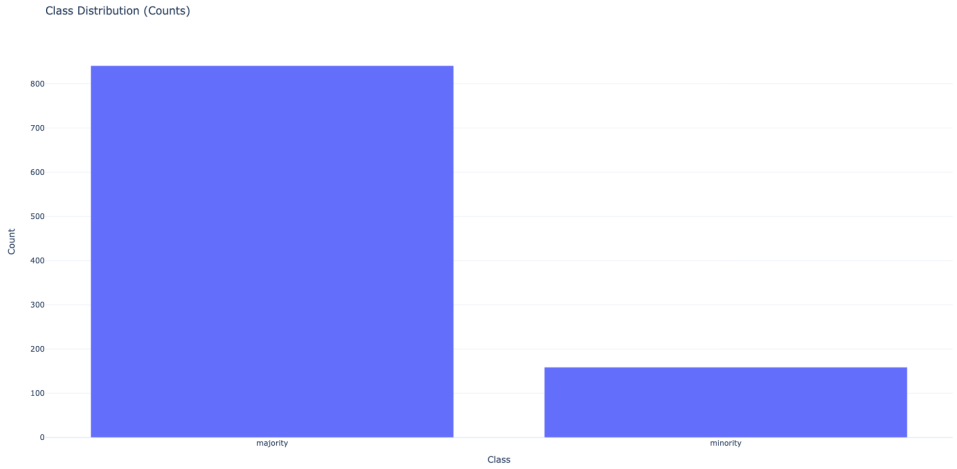
| row | issue |
|---|---|
| 0 | duplicate_identifier |
| 1 | duplicate_identifier |

**Timeliness Issues:** No issues found.

---

## Class Distribution

| Class | Count | Proportion |
|---|---|---|
| majority | 841 | 84.10% |
| minority | 159 | 15.90% |



# Missing Data Summary

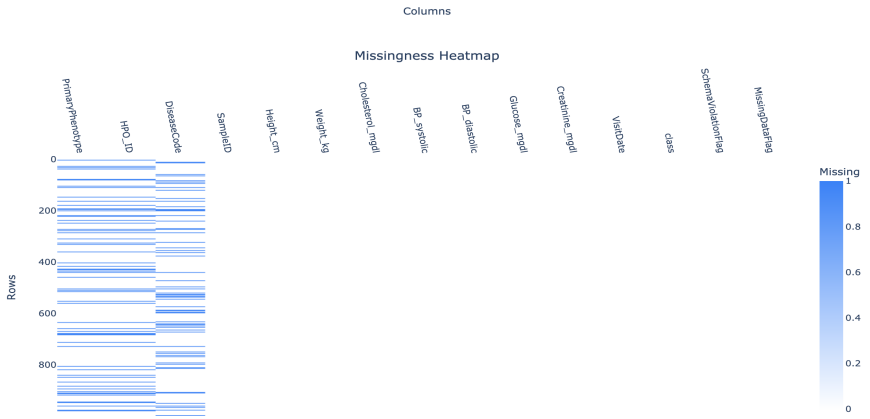
| Column | Missing Count |
|---|---|
| Creatinine_mgdl | 214 |
| PrimaryPhenotype | 159 |
| Cholesterol_mgdl | 152 |
| DiseaseCode | 148 |
| Glucose_mgdl | 123 |
| Height_cm | 89 |
| Weight_kg | 76 |
| BP_diastolic | 57 |
| BP_systolic | 35 |

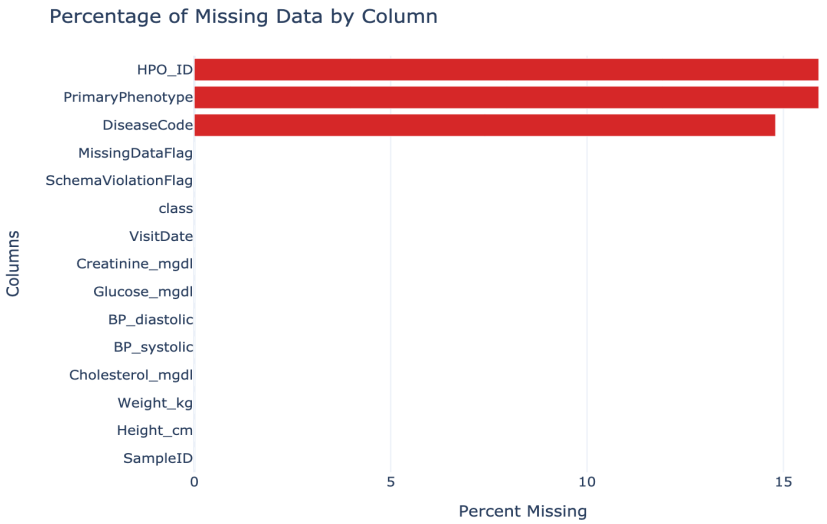
**Records Flagged for Missing Data:** 672

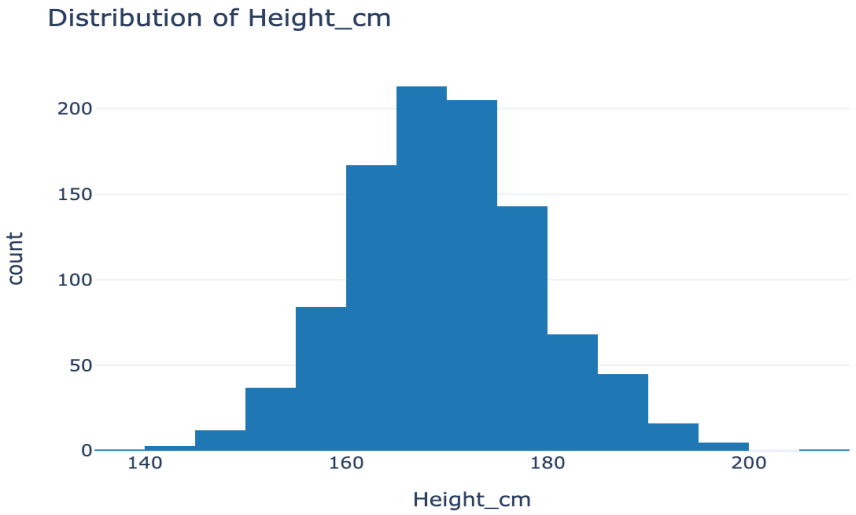
# Ontology Mapping Summary

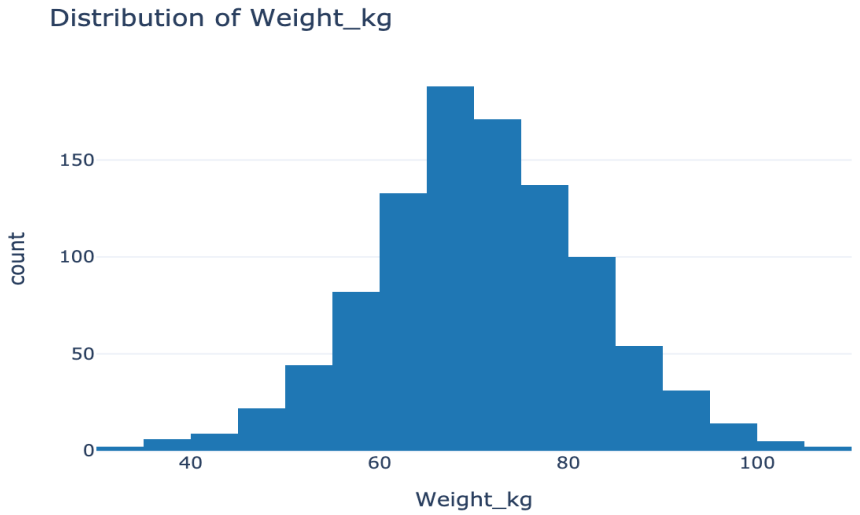
| Ontology | Total Terms | Mapped | Success Rate |
|---|---|---|---|
| HPO | 4 | 4 | 100.00% |

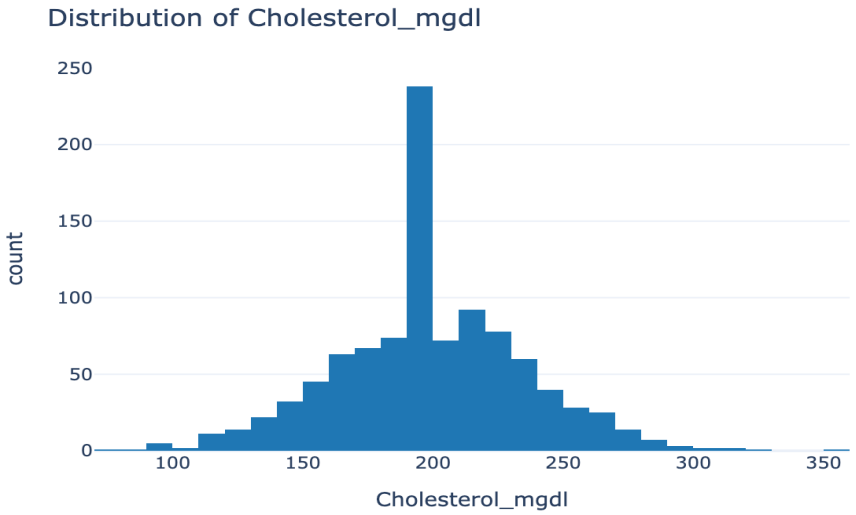
# Visualizations

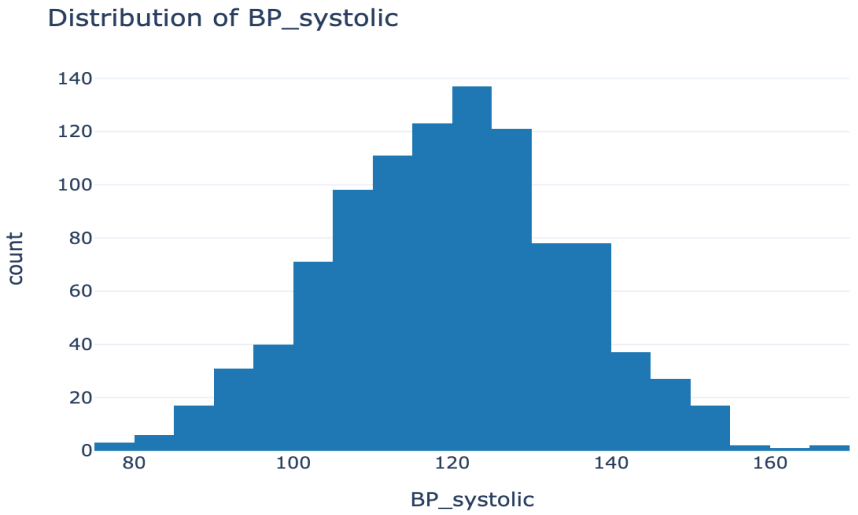


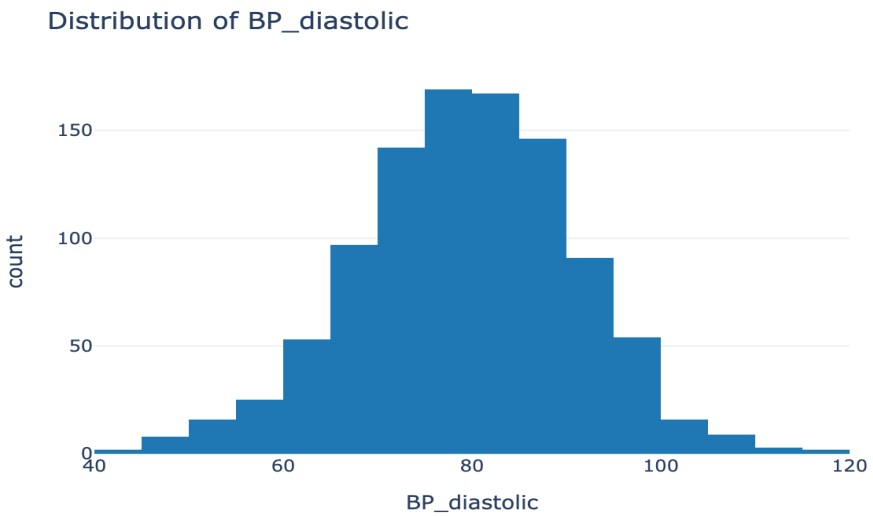



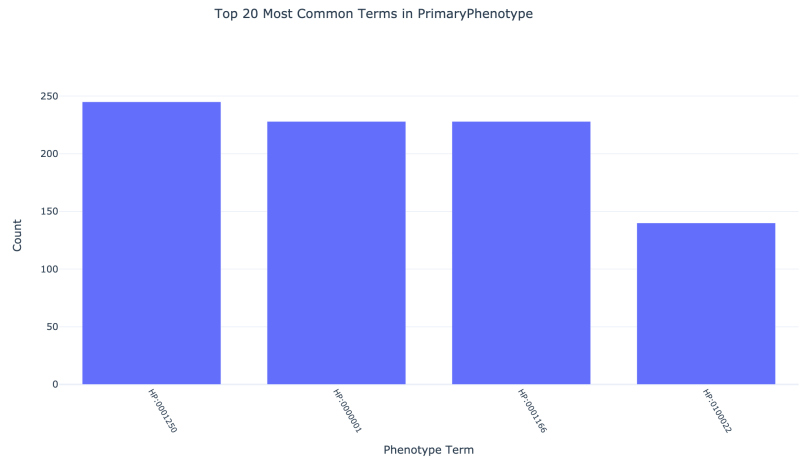



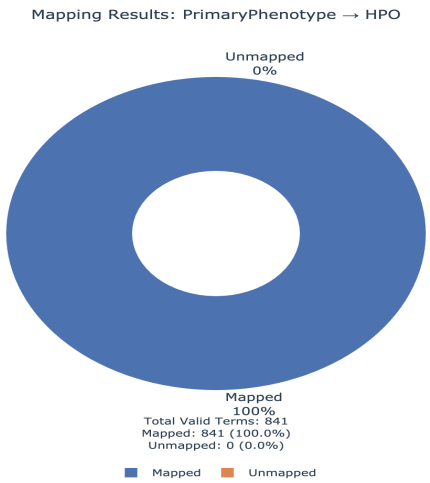




Distribution of Cholesterol_mgdl
count
Cholesterol_mgdl
0
50
100
150
200
250
100
150
200
250
300
350


Distribution of BP_systolic
count
BP_systolic
0
20
40
60
80
100
120
140
80
100
120
140
160


Distribution of BP_diastolic
count
BP_diastolic
0
50
100
150
40
60
80
100
120


Top 20 Most Common Terms in PrimaryPhenotype
Count
Phenotype Term
0
50
100
150
200
250
HP:0001250
HP:0000001
HP:0001166
HP:0100022


Mapping Results: PrimaryPhenotype → HPO
Unmapped
0%
Mapped
100%
Total Valid Terms: 841
Mapped: 841 (100.0%)
Unmapped: 0 (0.0%)
Mapped
Unmapped