PhenoQC Quality Control Report

Source file: synthetic_multi2_data.csv

Summary

Imputation Strategy	Mean
Schema Validation Score	0.00%
Missing Data Score	92.94%
Mapping Success Score	99.75%
Overall Quality Score	64.23%

Imputation Settings

Global Strategy	knn
Global Params	{'n_neighbors': 5, 'weights': 'uniform'}
Tuning Enabled	True
Best Params	{'n_neighbors': 7}
Tuning Score	198.6406 (MAE)

Data Quality Scores:

Schema Validation Score: 0.00%
Missing Data Score: 92.94%
Mapping Success Score: 99.75%
Overall Quality Score: 64.23%

Schema Validation Results

Format Validation: False
Duplicate Records: 300 issues found.
Conflicting Records: 300 issues found.
Integrity Issues: 3150 issues found.
Referential Integrity Issues: No issues found.
Anomalies Detected: 44 issues found.

Invalid Mask: 3150 issues found.

Missing Data Summary

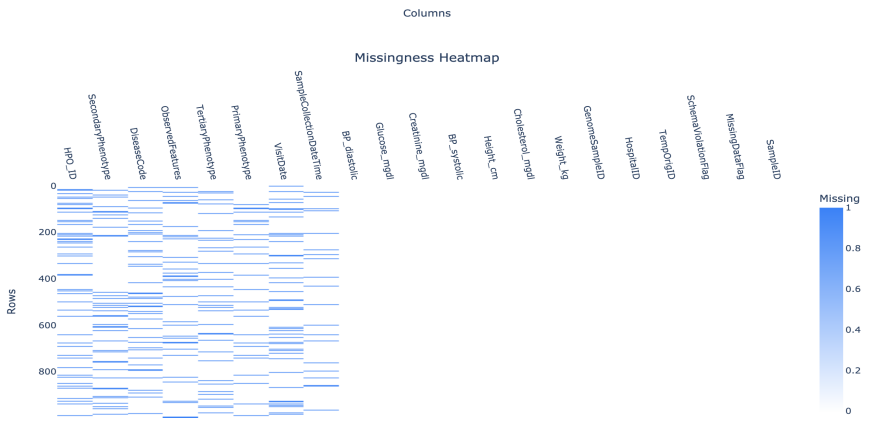
Column	Missing Count
SecondaryPhenotype	399
DiseaseCode	377
TertiaryPhenotype	367
PrimaryPhenotype	364
ObservedFeatures	350
Glucose_mgdl	322
BP_systolic	321
Height_cm	320
BP_diastolic	318
VisitDate	316
Weight_kg	315
Creatinine_mgdl	307
Cholesterol_mgdl	299
SampleCollectionDateTime	295

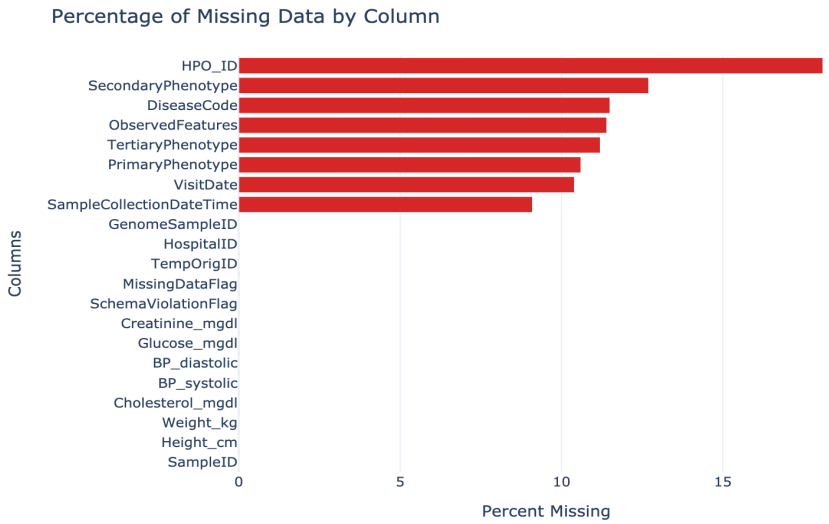
Records Flagged for Missing Data: 2485

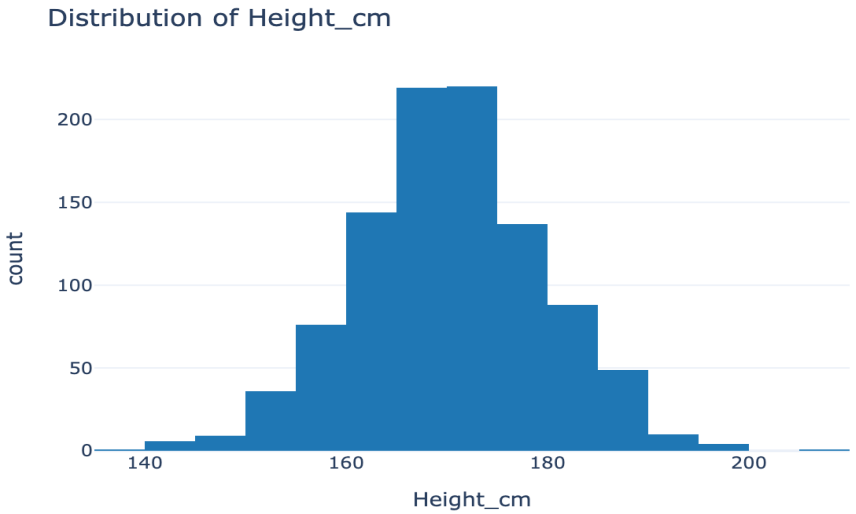
Ontology Mapping Summary

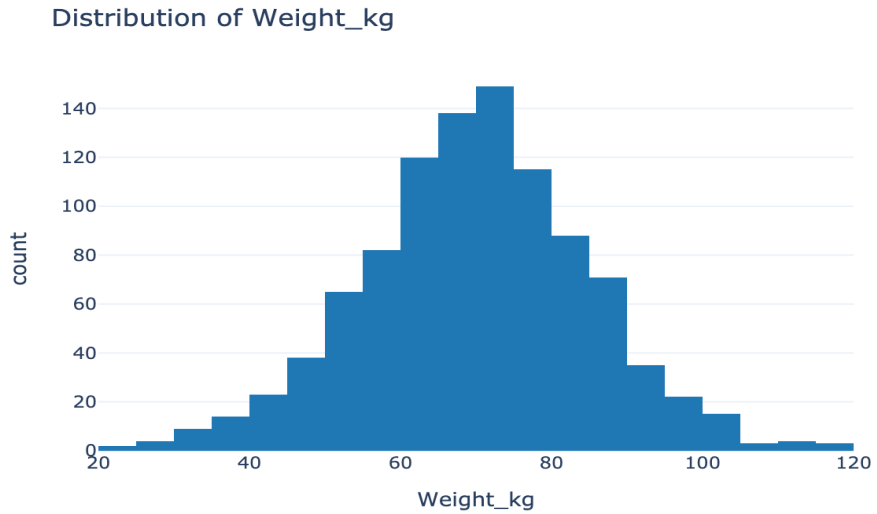
Ontology	Total Terms	Mapped	Success Rate
HPO	1985	1980	99.75%

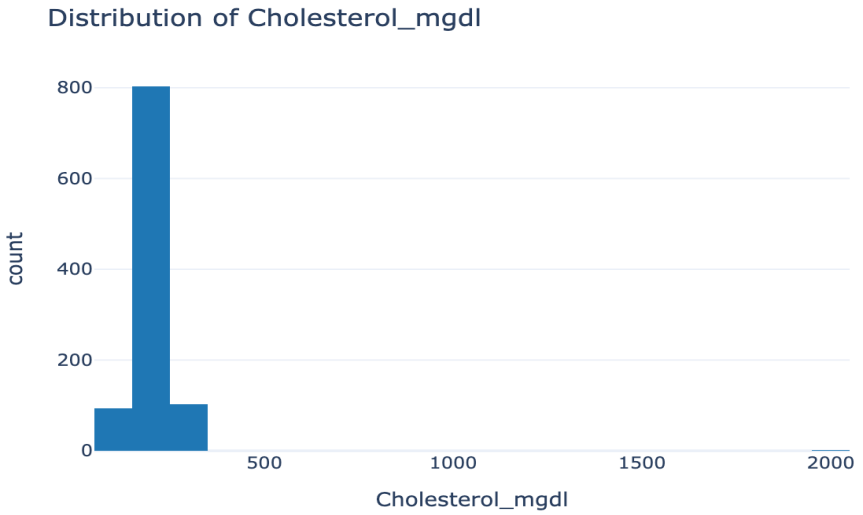
Visualizations

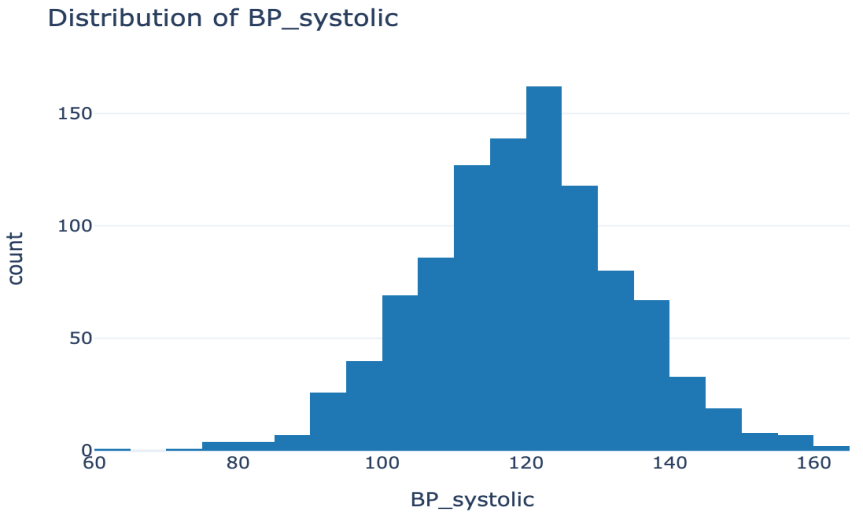


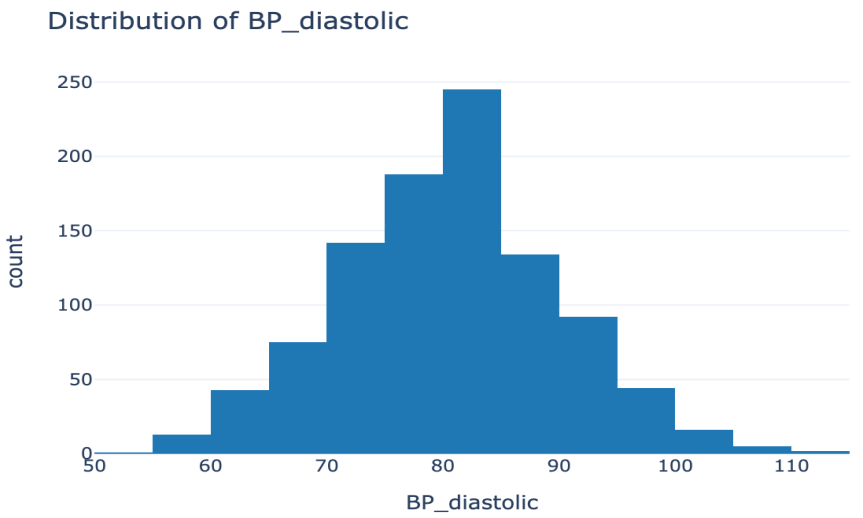


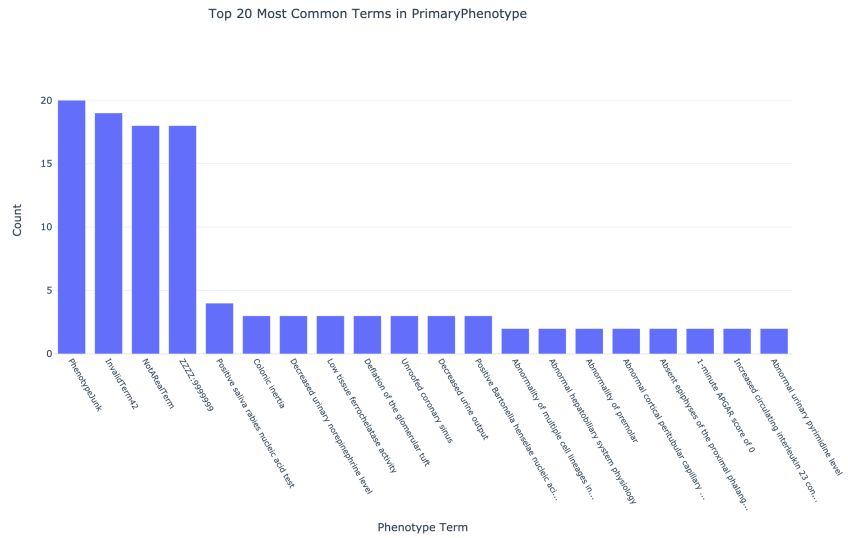


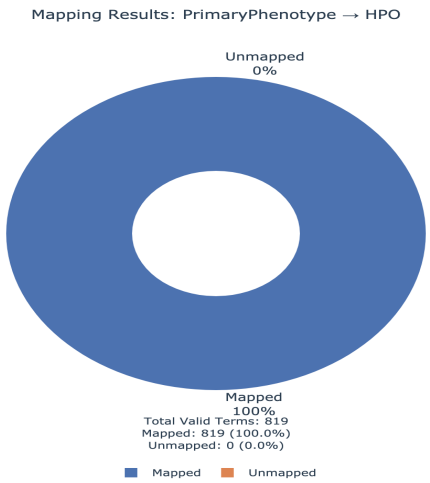



Distribution of Cholesterol_mgdl
count
Cholesterol_mgdl
0
200
400
600
800
500
1000
1500
2000


Distribution of BP_systolic
count
BP_systolic
0
50
100
150
60
80
100
120
140
160


Distribution of BP_diastolic
count
BP_diastolic
0
50
100
150
200
250
50
60
70
80
90
100
110


Top 20 Most Common Terms in PrimaryPhenotype
Count
Phenotype Term
0
5
10
15
20
PhenotypeJunk
InvalidTerm42
NotAtAllaTerm
ZZZZ:9999999
Positive saliva rabies nucleic acid test
Colonic inertia
Decreased urinary norepinephrine level
Low tissue ferrochelatase activity
Deflation of the glomerular tuft
Unmodified coronary sinus
Decreased urine output
Positive Bartonella henselae nucleic ac...
Abnormality of multiple cell lineages in...
Abnormal hepatobiliary system physiology
Abnormality of premolar
Abnormal cortical pericellular capillary ...
Absent epiphyses of the proximal phalang...
1-minute APGAR score of 0
Increased circulating interleukin 23 con...
Abnormal urinary pyrimidine level


Mapping Results: PrimaryPhenotype → HPO
Unmapped
0%
Mapped
100%
Total Valid Terms: 819
Mapped: 819 (100.0%)
Unmapped: 0 (0.0%)
Mapped
Unmapped