HealthFirst Care Data Cleaning Manipulation and Analysis

Capstone Project M04L01

Adwoa Boama Acheampong | IBM Business Analyst learning project

Method and corrected data treatment

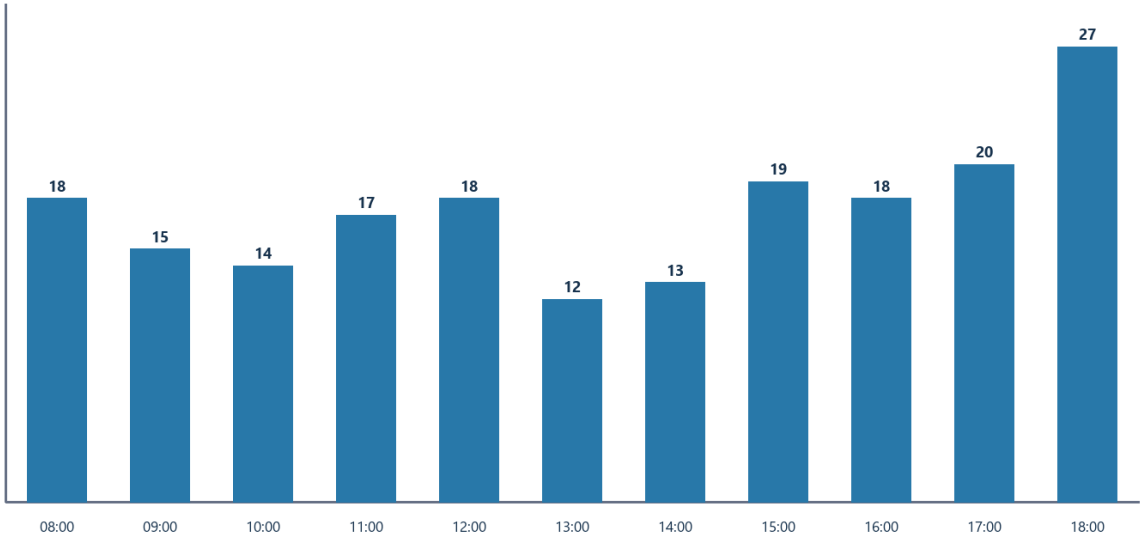
The analysis uses the Module 4 lab files. Original CSVs remain unchanged. Eighteen entirely empty feedback records were quarantined before duplicate removal or imputation. The 198 substantive feedback rows have no exact duplicates. Six missing scores were imputed with the valid-score mean of 6.2708, as the lab directs; three missing wait times remain missing. Thirteen exact resource duplicates were removed. Repeated feedback records were aggregated by PatientID before appointment lookup, preventing an arbitrary first-match result.

Dataset	Raw rows	Empty / duplicates	Clean rows	Material missingness
Appointments	200	0 / 0	200	11 AppointmentID; 2 date; 9 time; 9 status
Feedback	216	18 / 0	198	6 score; 3 wait; 7 department
Resources	213	0 / 13	200	3 ResourceID; 2 department; 4 type; 6 date

Appointment trends

The 200 appointment records include 56 completed, 46 no-show, 45 rescheduled, 44 cancelled and 9 missing-status records. Completion is 28.0% of all rows and no-show is 23.0%; analyses using known status should state that denominator separately. The busiest recorded hour is 18:00 with 27 appointments. Sunday has the highest day count at 35 among 198 parseable dates.

Appointments by hour

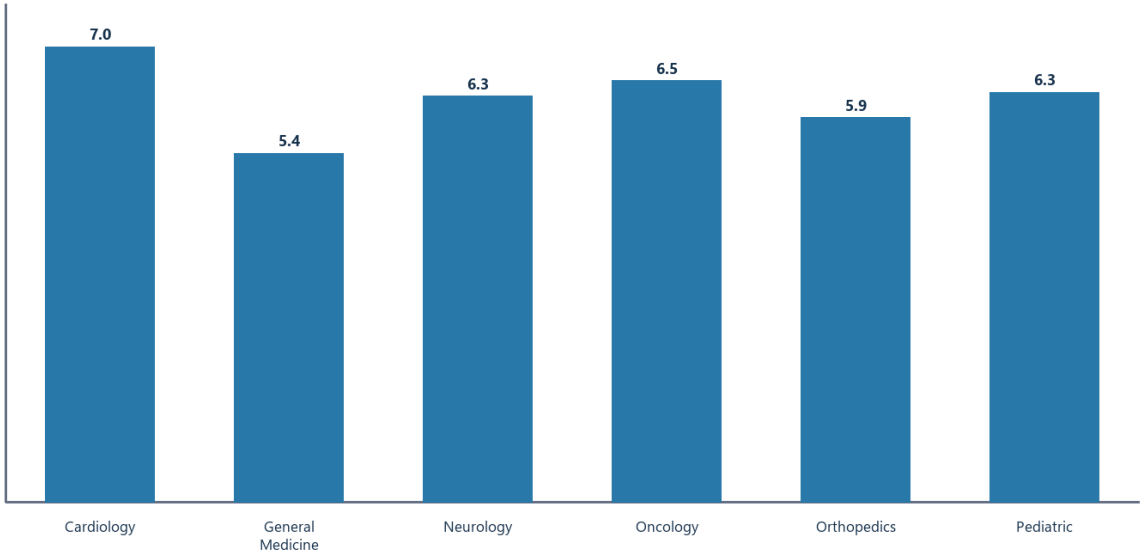


Appointments by parsed hour. Nine rows have missing time and are excluded.

Patient satisfaction

Department	Responses	Average score	Average wait minutes	High	Low
Cardiology	33	7.03	40.00	31	2
General Medicine	26	5.38	43.85	21	5
Neurology	34	6.26	42.79	33	1
Oncology	32	6.50	39.38	26	6
Orthopedics	29	5.93	42.41	21	8
Pediatric	37	6.32	41.35	32	5

Average feedback score by department



Feedback score uses the lab threshold of 4 for High versus Low satisfaction. The source score range is 1 to 10, so the classification is retained as instructed and interpreted cautiously.

General Medicine has the lowest average feedback score among named departments (5.38). Orthopedics has the most low-score responses (8). General Medicine has the longest average recorded wait (43.85 minutes). These are department-level associations, not patient-level causal effects.

Dissatisfaction at the peak appointment hour

Appointment time	Appointments	Patient-score matches	Mean linked score	Below 4 among matches
18:00 peak	27	26	5.77	7/26 (26.9%)
Other recorded hours	164	158	6.34	19/158 (12.0%)

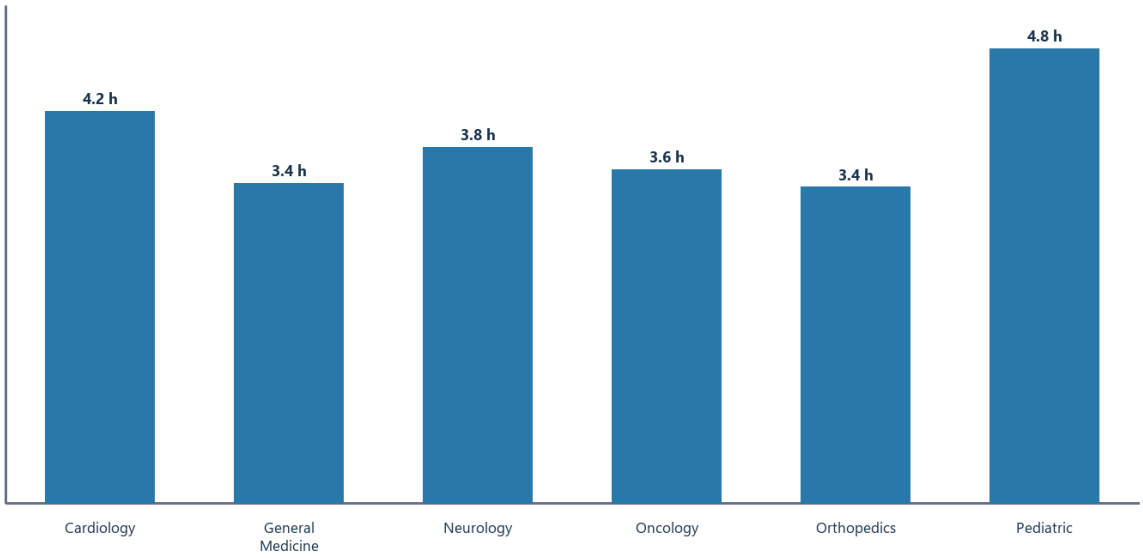
The 18:00 hour has 27 appointments (26 patient-score matches), versus 164 appointments at other recorded hours (158 matches). Nine appointments with missing time are excluded. Repeated feedback was first averaged by PatientID so the join does not multiply appointment rows. A patient's mean feedback is not linked to a specific appointment event; this is a descriptive contrast, not evidence that the peak caused dissatisfaction. The below-4 threshold follows the lab's classification rule.

Resource utilisation

Department	Records	Average usage hours	Available	Availability rate
Cardiology	22	4.18	13	59.1%
General Medicine	39	3.41	22	56.4%
Neurology	34	3.79	22	64.7%
Oncology	36	3.56	21	58.3%

Department	Records	Average usage hours	Available	Availability rate
Orthopedics	27	3.37	14	51.9%
Pediatric	40	4.85	24	60.0%

Average resource usage hours by department



Average recorded usage hours by department. A record is an operational observation, not a full capacity denominator.

Join and cardinality validation

Appointments contain 200 unique PatientID values. Feedback contains 184 nonblank unique PatientID values; 10 patients have multiple substantive feedback records and 16 appointment patients have no feedback. The workbook aggregates feedback by patient before VLOOKUP and exposes the contributing record count. Resource data is compared only at department level because it has no PatientID.

Operational implications

- Prioritize Orthopedics and General Medicine for process review because no-show, feedback and wait signals indicate different operational problems.
- Treat the 18:00 peak as a scheduling and resource-planning hypothesis to validate by shift, clinician and capacity data.
- Keep missing and stale records visible. A dashboard should not transform unknown values into zero.
- Use the native PivotTables and slicers in the workbook for exploratory review; preserve metric definitions beside decisions.

Limitations

The files are simulated course data. No clinical outcomes, costs, approved targets or implementation benefits are supplied. Dates mix ISO and slash formats; slash values are treated as day-first because the source contains dates such as 14/01/24. Findings are descriptive and require operational validation.

Open the standalone analysis workbook with native PivotTables and slicers

Embedded cleaned workbook

The complete standalone Excel workbook is embedded below as an editable object as well as supplied separately. Double-click the icon in a compatible Word/Excel installation to inspect the cleaned datasets, formulas and native PivotTables.



Workbook inspection map

Open these sheets	What to inspect
Source Manifest; Raw_Appointment; Raw_Feedback; Raw_Resource	Source filenames, unchanged extracts and the cleaning decisions applied downstream.
Cleaned_Appointment; Cleaned_Feedback; Cleaned_Resource	200 appointment, 198 substantive feedback and 200 deduplicated resource rows; typed fields and preserved missing values.
Patient Feedback Summary; Combined Analysis	Patient-level aggregation before the join; contributor count and formulas that prevent repeat feedback from multiplying appointments.
Data Quality Log; Department KPIs	Reconciliation exceptions, explicit denominators and department-level measures.
Pivot Appointments; Pivot Satisfaction; Pivot Resources	Four refreshable native PivotTables, four PivotCharts and six slicers for exploratory review.

The standalone .xlsx is the primary interactive deliverable. The embedded copy is for compatible desktop Word and Excel; PDF viewers cannot operate its formulas or slicers. The separate dashboard workbook is a different deliverable and is not embedded here.