

Data Dictionary: Employee Absenteeism Data

The dataset contains records of absenteeism at work from July 2007 to July 2010 at a courier company in Brazil.

Data Structure of the Employee absenteeism dataset:

1. ID: Unique identifier for each employee (Integer)
2. Reason for absence: Absence reason code (Integer)

Reason for absence (ICD):

Absences attested by the International Code of Diseases (ICD) stratified into 21 categories (I to XXI) as follows:

I Certain infectious and parasitic diseases

II Neoplasms

III Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism

IV Endocrine, nutritional and metabolic diseases

V Mental and behavioural disorders

VI Diseases of the nervous system

VII Diseases of the eye and adnexa

VIII Diseases of the ear and mastoid process

IX Diseases of the circulatory system

X Diseases of the respiratory system

XI Diseases of the digestive system

XII Diseases of the skin and subcutaneous tissue

XIII Diseases of the musculoskeletal system and connective tissue

XIV Diseases of the genitourinary system

XV Pregnancy, childbirth and the puerperium

XVI Certain conditions originating in the perinatal period

XVII Congenital malformations, deformations and chromosomal abnormalities

XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified

XIX Injury, poisoning and certain other consequences of external causes

XX External causes of morbidity and mortality

XXI Factors influencing health status and contact with health services.

And 7 categories without (CID) patient follow-up (22), medical consultation (23), blood donation (24), laboratory examination (25), unjustified absence (26), physiotherapy (27), dental consultation (28).

3. Month of absence: Month of absence (Integer, likely 1-12)
4. Day of the week: Day of the week (Integer, likely 1-7)
Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6))
5. Seasons: Season code (Integer)
Seasons (summer (1), autumn (2), winter (3), spring (4))
6. Transportation expense: Transportation cost (Float) in a specific format (e.g., removing commas might be needed for analysis)
7. Distance from Residence to Work: Distance from home to work (Integer)
8. Service time: Employee's tenure (Integer, likely years)
9. Age: Employee's age (Integer)
10. Work load Average/day: Average daily workload (Integer)
11. Hit target: Whether the employee met their target (Integer, likely 0 or 1)
12. Disciplinary failure: Any disciplinary actions (Integer, likely 0=no or 1=yes)
13. Education: Employee's education level code (Integer)
Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))
14. Son: Whether the employee has a son (Integer, likely 0 or 1)
15. Social drinker: Social drinking habits code (Integer: yes=1; no=0)
16. Social smoker: Social smoking habits code (Integer: yes=1; no=0)
17. Pet: Whether the employee has a pet (Integer, likely 0 or 1)
18. Weight: Employee's weight (Integer)
19. Height: Employee's height (Integer)
20. Body mass index: Calculated BMI (Float)
21. Absenteeism time in hours: Total absence duration in hours (Integer)

Link of the data source: <https://www.kaggle.com/code/faressayah/predict-employee-absenteeism-from-work/notebook>