# Data Preprocessing (Step 1) Documentation

## Overview

This document outlines the data preprocessing steps performed in Step 1 of the Polytrauma Analysis project. The preprocessing was implemented in the ingestion\_preprocess.py script, which cleaned and transformed the raw data to prepare it for further analysis.

## Input Data

* **Source File:** C:\Users\Mohammad\PycharmProjects\Polytrauma Analysis\data\input\Polytrauma Analysis.xlsx
* **Initial Shape:** 153 rows, 60 columns

## Preprocessing Steps

### 1. Column Name Cleaning

All column headers were stripped of leading and trailing whitespaces to ensure consistency.

**Example transformations:** - Kopf → Kopf - Arm rechts → Arm rechts - Besuchsdatum → Besuchsdatum

### 2. Empty Column Removal

Only the completely empty ‘index’ column was removed, as it contained no useful information.

### 3. Date Column Conversion

The following columns were converted to proper datetime format: - Unfalldatum (Accident date) - Besuchsdatum (Visit date) - Gebursdatum (Birth date)

### 4. Age Calculation

A new column Age\_At\_Accident was created by calculating the difference between the patient’s birth date and accident date, converted to years.

### 5. Time Interval Assignment

A new column Time\_Interval was created by grouping the Monat nach Unfall (Month after accident) into 3-month intervals: - Interval 1: 0-3 months - Interval 2: 3-6 months - Interval 3: 6-9 months - Etc.

The formula used was: ((Monat nach Unfall // 3) + 1)

### 6. Days Since Accident Calculation

A new column Days\_Since\_Accident was added to track the exact number of days between the accident date and each visit date.

## Output Data

* **Output File:** C:\Users\Mohammad\PycharmProjects\Polytrauma Analysis\data\output\step1\Polytrauma\_Analysis\_Processed.xlsx
* **Final Shape:** 153 rows, 62 columns

## Column Details

### Original Columns

The dataset contains the following types of columns:

1. **Patient Identifiers and Basic Information:**
   * Schadennummer (Case number)
   * Unfalldatum (Accident date)
   * ICD-10 (Diagnosis code)
   * Geschlecht (Gender)
   * Alter in Dekaden (Age in decades)
   * Gebursdatum (Birth date)
   * Ort der Beratung (Location of consultation)
2. **Visit Information:**
   * Vor Ort-Besuch Nummer (On-site visit number)
   * Besuchsdatum (Visit date)
   * Monat nach Unfall (Month after accident)
3. **Injury Pattern (Body Parts):**
   * Kopf (Head)
   * Hals (Neck)
   * Thorax
   * Abdomen
   * Arm links (Left arm)
   * Arm rechts (Right arm)
   * Becken (Pelvis)
   * Wirbelsaeule (Spine)
   * Bein rechts (Right leg)
   * Bein links (Left leg)
4. **Medical and Social Factors:**
   * Various columns for somatic issues, activity status, environmental factors, personal factors, and different types of case management (medical, social, technical, vocational)

### Added Columns

Three new columns were added during preprocessing:

1. **Age\_At\_Accident:** Patient’s age (in years) at the time of the accident
2. **Time\_Interval:** Grouping of months after accident into 3-month intervals
3. **Days\_Since\_Accident:** Exact number of days between accident and each visit

## Implications for Analysis

The preprocessing performed in Step 1 accomplishes several important tasks:

1. **Data Cleaning:** Standardizes column names and formats
2. **Temporal Context:** Adds timing information relative to the accident
3. **Age Information:** Provides patient age at time of accident for demographic analysis
4. **Consistency:** Ensures date fields are in proper datetime format for accurate duration calculations

This processed dataset is now ready for quality assessment and further analysis to determine factors affecting healing duration.