# Polytrauma Analysis Project Documentation

## Overview

This document provides a comprehensive overview of the Polytrauma Analysis Project, including data processing, quality checks, and analyses performed to date. The project aims to analyze polytrauma patient data to understand recovery patterns, medical interventions, and factors affecting healing duration.

## Project Structure

The project is organized into sequential steps:

1. **Step 1: Data Ingestion and Preprocessing**
2. **Step 2: Data Quality Check and Cleaning**
3. **Step 3: Advanced Analysis** (planned)

## Data Description

The dataset contains information about polytrauma patients, including:

* **Case Information**: Schadennummer (case ID), accident dates, visit information
* **Injury Locations**: Kopf (head), Arm links/rechts (left/right arm), Bein links/rechts (left/right leg), Wirbelsaeule (spine), Becken (pelvis), etc.
* **Medical Conditions**: Somatisch (somatic conditions), Personenbezogen (personal factors)
* **Activity Status**: Taetigkeit (employment status)
* **Environmental Factors**: Umwelt (environmental conditions)
* **Case Management Categories**: Med RM, Soziales RM, Technisches RM, Berufliches RM

The dataset consists of 30 unique cases with multiple visits per patient, for a total of 152 records (after removing duplicates) and 62 columns.

## Step 1: Data Ingestion and Preprocessing

### Process Overview

1. **Data Loading**: Original Excel file imported with appropriate data types (Schadennummer as string)
2. **Column Cleaning**: Whitespace trimmed from column headers
3. **Date Conversion**: Date columns (Unfalldatum, Besuchsdatum, Gebursdatum) converted to datetime format
4. **Derived Columns**:
   * Age\_At\_Accident: Age at time of accident calculated from Gebursdatum and Unfalldatum
   * Time\_Interval: Time periods categorized into 3-month intervals
   * Days\_Since\_Accident: Days between accident and each visit

### Key Metrics After Preprocessing

* Total records: 153 (including duplicates)
* Unique cases: 30
* Total columns: 62

## Step 2: Data Quality Check and Cleaning

### Missing Value Analysis

* 52 out of 62 columns have missing values
* Several columns have 100% missing values:
  + Personenbezogen–Suchtverhalten
  + Berufliches RM–Wiedereingliederung geförderter Arbeitsmarkt
  + Ehrenamt
  + Zuverdienst
* Most categorical columns have high missingness rates (>80%)
* A total of 6,831 missing values were identified across the dataset
* Highest missing percentage: 100%

### Data Type Validation

* All columns now have correct data types
* “Alter in Dekaden” properly recognized as categorical string
* Date columns confirmed as datetime format
* Categorical columns (Ja/Nein) stored as appropriate object types

### Duplicate Detection

* 1 completely duplicate row was detected and removed
* The duplicate was associated with Schadennummer 30.20.550132.089
* 3 instances of duplicate visits (same Schadennummer and Besuchsdatum) were found:
  + Case 2020.11.08.02772.1 had two different visits on 2020-11-30 (legitimate)
  + Case 2020.11.08.02772.1 had two different visits on 2021-03-09 (legitimate)
  + Case 30.20.550132.089 had two identical visits on 2020-05-13 (data error)

### Categorical Value Validation

* All categorical columns contain only expected values (“Ja”/“Nein”)
* No unexpected values were detected

### Data Cleaning Process

1. **Missing Value Replacement**:
   * All missing values in categorical columns replaced with “Nein”
   * A total of 6,580 categorical missing values were replaced
   * Non-categorical missing values (like dates) were preserved
2. **Duplicate Removal**:
   * Only 100% identical duplicate rows were removed
   * Kept first occurrence of each duplicate
   * Dataset reduced from 153 to 152 records
3. **Output Files**:
   * Cleaned dataset saved as “Polytrauma\_Analysis\_Clean.xlsx”
   * Detailed quality reports saved: missing\_values\_analysis.xlsx, data\_type\_validation.xlsx
   * Duplicate information saved for inspection

## Key Insights

1. **High Missingness Pattern**:
   * The high rate of missing values (89.5% for Kopf injury, for example) likely indicates that most patients do not have these specific conditions or interventions
   * Missing values in categorical columns represent absence of condition (correctly replaced with “Nein”)
2. **Visit Patterns**:
   * Patients have varied number of visits (2-21 visits per patient, average of 5.1)
   * First visit timing ranges from 20 to 659 days after accident
   * Last visit timing ranges from 182 to 1233 days after accident
3. **Head Injury Impact**:
   * Head injuries show significant impact on healing duration (t-statistic=-2.83, p-value=0.009)
   * Cohen’s d effect size (-1.03) indicates large effect
   * Patients with head injuries have longer recovery periods
4. **Age Distribution**:
   * Patient ages span several decades
   * Each decade has representation in the dataset

## Dataset Challenges and Limitations

1. **Sparse Data**:
   * Many categories have very few positive cases (marked as “Ja”)
   * Some conditions like “Personenbezogen–Suchtverhalten” have no recorded cases
2. **Sample Size**:
   * While there are 152 total records, they represent only 30 unique cases
   * Some subgroup analyses may be limited by small sample sizes
3. **Data Collection Variations**:
   * Some duplicate visits on the same day suggest potential data collection inconsistencies
   * Possible data entry errors identified and corrected

## Next Steps (Step 3)

1. **Advanced Statistical Analysis**:
   * Correlation analysis between injury types and recovery duration
   * Multivariate analysis of factors affecting recovery
   * Time-series analysis of patient progress
2. **Predictive Modeling**:
   * Develop models to predict recovery time based on injury profiles
   * Identify key predictors of extended recovery periods
3. **Visualization and Reporting**:
   * Create comprehensive dashboards of patient recovery trajectories
   * Generate summary reports for different injury categories

## Technical Implementation

The project is implemented in Python using the following key libraries: - pandas: Data manipulation and analysis - matplotlib/seaborn: Data visualization - logging: Comprehensive logging of all processing steps - os/dotenv: Environment configuration and file handling

All scripts are organized in a modular structure to facilitate maintenance and extension.

## Conclusion

The polytrauma dataset has been successfully preprocessed and cleaned, with comprehensive quality checks implemented. The dataset now has consistent handling of missing values, proper data types, and no duplicates. Despite high missingness in many columns, the data provides valuable insights into polytrauma recovery patterns, particularly the significant impact of head injuries on recovery duration.

The project is now ready for advanced analytical methods in Step 3 to extract deeper insights from this rich clinical dataset.