# Brief:

The dataset part of the njson file consists of 1000 records and 5 fields. To get a feel of the data issues, pandas\_profiling library is used that provides out of box stats for the dataset.

Based on the observations made in the data profile, the steps are taken to cleanse and transform data.

This document provides the summary of data issues and lays out of the steps that were taken to cleanse data.

The actual script and dependencies are included in the project.

# Data Issue Summary:

Here is the summary of the data issues that are flagged by the pandas\_profiling library:

Graphical user interface, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

# Common Issues:

The common issues encountered in the data set are:

* Missing data – null/NA/empty
* Invalid entries – special characters
* Data type issues – floating, integers, strings, date
* Edge cases – either values are too high or too low

From data perspective, the missing values could mean either the value is NULL/NaT/NA or there is no entry. This issue is common across all the fields.

Missing values –

* All fields have issues that relate to missing values that is either entries are NULL/Na or empty

Invalid entries –

* In case of ***Value*** field, some of the records have currency symbols is included
* In case of ***Name*** field, some of the names are embedded with dict values that is first name and the last name
* In case of ***Description*** field, some of the records have either single- or double-digit numeric values

Edge Cases:

* There is huge variation in the min and max values for ***Value*** and ***Weight*** fields

Data Type issues:

* In case of ***date*** field, some of the records are null/empty and the data field selected is as an object

# Approach – Data cleansing and transformation

## Name field:

Graphical user interface, text, application

Description automatically generated

* Some of the entries have embedded dictionary pairs of first and last name. These records are filtered and based on the entries converted in the full name
* Records are check for missing of invalid entries and accordingly updated

## Value field:

Text

Description automatically generated

* Some of the fields have currency symbols. This is being fixed by parsing the digits and merging them to the original data frame
* Null and na are replaced with 0
* Field is cast to int data type

## Weight field:

Graphical user interface, application

Description automatically generated

* Some of the na records are replaced with 0
* Field is cast to int data type

## Date field:

Graphical user interface, text, application

Description automatically generated

* Some of the entries in the date field have precision to microseconds. These records are normalized to the midnight entries
* Null dates are replaced with outlier date so that they can omitted in a later steps

## Description field:

Text

Description automatically generated

* The empty entries are replaced with a text that identifies the record as either having missing data or invalid entry
* Some of the entries have numeric entries (5, 1, 20 etc) and they appear to be invalid records and hence replaced with a text

## Data types:

Graphical user interface, text, application

Description automatically generated

The fields are cast to appropriate data types

# Outcome:

Here is the outcome of the data cleansing exercise:

Graphical user interface

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated