Data Cleaning Project - 1

The data is in shit condition.

The issues:

1. None of them columns are in consistent format
2. There are a lot of null values in all the columns
3. The columns “Age” consists of number (45), character(“forty”) and alnum, (85 years) data’s.
4. The metadata are not in consistent cases. Some are in lower case and some are in upper case.
5. There are 3 date columns and none of them are in datetime datatype. There are mixtures of dates, excel date formats, inconsistent date formats, and some other type of format is included.
6. The “Diagnosis” column consists of null values and case problems
7. The “Heart\_rate” column includes mixture of int, floats, word numbers and null cells.

My approach towards Data Cleaning

The first is AGE column. there are null values and numeric age numbers and words number(eg: forty) and mixture of number and words(eg: 85 years). so my plan is to convert these into proper numbers and then fill the unknown values with imputation. then convert it to int since its in object type now.

Wrote code to clean such noises

The age column is cleaned and since the issue was same with the “Heart Beat” column I applied the same code to that column as well and now 2 columns are pre processed.

Since the gender column consisted of upper and lower cases words I decided to comvert those into lowet case. There was null values as well so I decided to fill those with “unknown”. After converting it to lower case filling the null values with “unknown” I notice that nan is till there. The uniqueness of that column was ['female' 'male' 'nan']. Then I realised it was not Null values, it was the word “nan” itself. Now I need to replace nan with NaN and then fill it “unknown”. Or can I directly fill it with unknown!. Since I know it’s a string I can just replace it with the required word.

Next comes the diagnosis column. When I looked at the uniqueness of this column I got the below

['Elevated Pressure' nan '??' 'Hypertension' 'H.B.P' 'HBP'

'High blood pressure' 'high BP' 'HTN' '1234']

But if u see 'H.B.P', 'HBP', 'Elevated Pressure' , 'High blood pressure', 'high BP' are High blood pressure itself. And 'Hypertension' and 'HTN' are hypertension itself. Hence I need to replace these.

Now now, u see theres a sneaky “nan”. But is it actually NaN!! From our previous lesson we need check the nulls in this column as well.

Ok, looks like those are null values itself. I just re ran the codes from beginning for the, turns out the gender columns nan were also null values. I had run the replacement code once that’s why it was showing it as a string. Now I need to convert it to NaN and then handle it.

Ok so the conversion from nan to NaN and the replacements are done.

Now comes the most messed up columns------the date columns.

Enough cleaning for today. I will handle these columns tomorrow.

Ok so tomorrow is today. Lets continue with our data cleaning.

Lets deal with the date columns.

The 3 date columns had these type of values 'Saturday, October 21, 2017', 44819, Z04, Unknown, 14/12/2025, 2018-08-21abc,20211227. From my previous projects I know that 44819 is a excel date and is easily solvable. And Z04 is garbage. Hence I need to convert it to NaN, but wait NaN means Not a Numeric. Since this is date it should be NaT. Not a Time. Wow, this is so cool.

Ok somehow handled all the inconsistencies, but I did not handle the missing values.

I just checked that the max date is from 2025 and the minimum date is from 2015. I cannot just impute random date values. So why not proceed with median, that is a good option I guess. Lets do it.

But wait, there are 3 columns check in date, first consultant date and next visit. i cannot impute median for all the 3 columns. what if the patient dint visit in next visit, imputing median would be a false information. so i believe we have to handle each column separately so its accurate. right?

This was my approach and how I saw the problem:

if check in date exists, and if it does, check if the first consultant date is after the check in date and also check if next visit is after the first consultant date then keep it, and if its before the first consultant date then fill it with No return. if the check in is null but there is first consultant date and next visit date in proper order then fill it with unknown. check if the check in date is before the first consultant date then drop the row.

So the condition would be

print("Any bad consults left? ", ((df['first\_consult'] < df['check\_in']) & df['check\_in'].notna()).any())

yes, I told chatgpt to write the code for it.

And finally we have a very valid datetime columns which represents the data accurately.

Lets look at the null values again  
Patient\_ID 0

Name 0

Age 31

Gender 0

Check\_in\_Date 34

First\_Consultation 31

Next\_Visit 45

Diagnosis 0

Heart\_Rate 45

Ohhhhhh…a lot of null values. We have to handle this first. I would like to approach it by looking at the datatypes first. Because if the datatype of that column is int and I fill a string,,,,,uhmm …..not so professional is it. So I also want to see if there are any outliers so that I can choose the imputation method.

I thought that I am close to cleaning the data. But as deeper u dig the more problems u find.

Ok so I checked for outliers from age and heart rate columns and there were outliers. When I observed deeply there were inconsistencies in the datetime columns that I did not handle. For ex: first\_consult date is after next\_visit

Check in date is after next\_visit

Which are meaningless. So I dropped those rows.

The outliers reduced and now there are 1 outlier each in both columns. One has heart rate 0.0 😭. I had to drop this row too.

After testing mean and median I decided to impute the age column with mean.

I then imputed the heart rate with mode.

And finally comes our favourite datetime columns. I tried to fill the NaT with strings hoping that the datatype doesn’t change. But python doesn’t know hopes. It changed it to object. Now I have to find a way to fill it without changing the datatype. I think I will leave it as NaT itself.

I think we complete our Data cleaning here. But lets look at it again after lunch.

Data Cleaning Project - 2

This project consists of fetching data from API. The first step was to choose an API with messy data. I used twitter api but I was not satisfied by the quality of the data. In simple words, the data dint seem strong. Hence I continued with News API. I used Qwen 2.5 to produce a script to fetch data from news api. There were issues in the beginning to fetch the data as it used to fetch one page and I used to get error. I tried again but I got the same error. So I decided to fetch data only 1 page per topics like science, tech, business, environment. So totally I fetched 5 pages 2 from science. I got the individual csv files. Merged it to one using the command line. Ok so the data is available now to clean. I am going to use polars as we used pandas for the previous data. As I was exploring the data I found myself reading the news article and all or most of the news I could read of was of trump 😂.

**Problem Statement**

"Analyzing News Trends and Content Quality from Diverse Sources"  
The dataset contains news articles from multiple sources (e.g., The Verge, Wired, BBC) with inconsistent formatting, missing values, and unstructured text. The goal is to clean and standardize the data to:

* Identify trends in news topics (e.g., AI, tech, health).
* Compare content quality across sources.
* Analyze publication patterns over time.

My first experience with polars was not good. Tried from 2pm to 8.39pm. I got frustrated and decided to use pandas. Polars is a piece of shit as for now