

Amini Graduate Fellowship Program Technical Assignment

By: Christopher Croney

Introduction

This technical assignment was based on the health_registry.csv file which was provided. Work done on this CSV was split into two sections, data profiling and data cleaning. The CSV contains 9 columns as shown below:

facility_id (*primary key*)

facility_name (*object*)

facility_type (*object*)

capacity (*object*)

region (*object*)

licence_issue_date (*object*)

inspection_date (*object*)

gps_location (*object*)

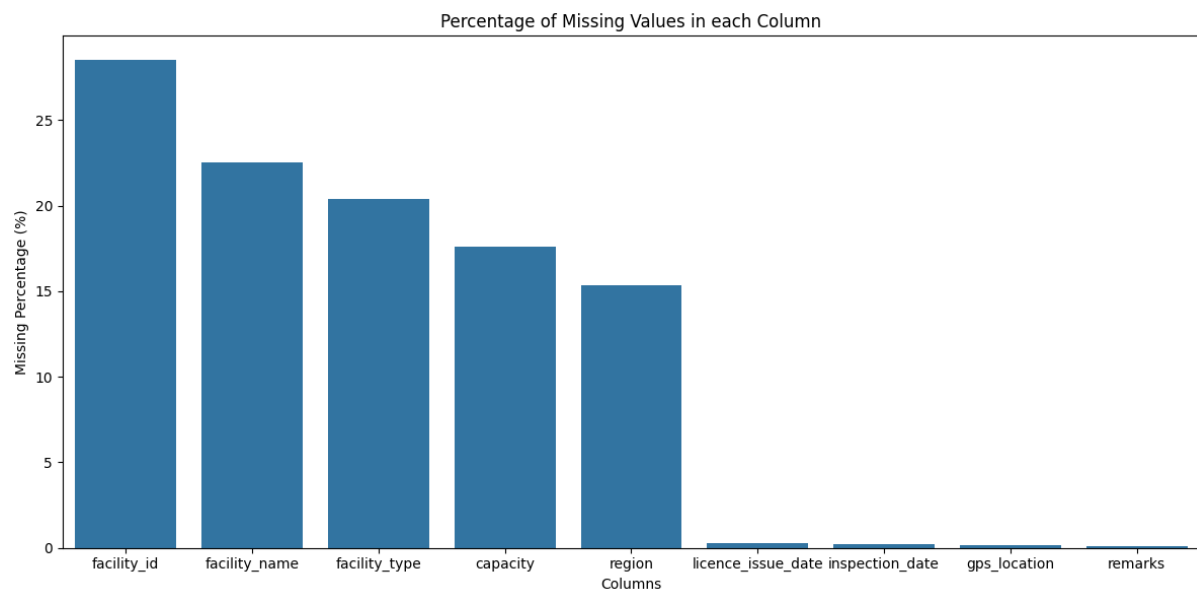
remarks (*object*)

Data Profiling

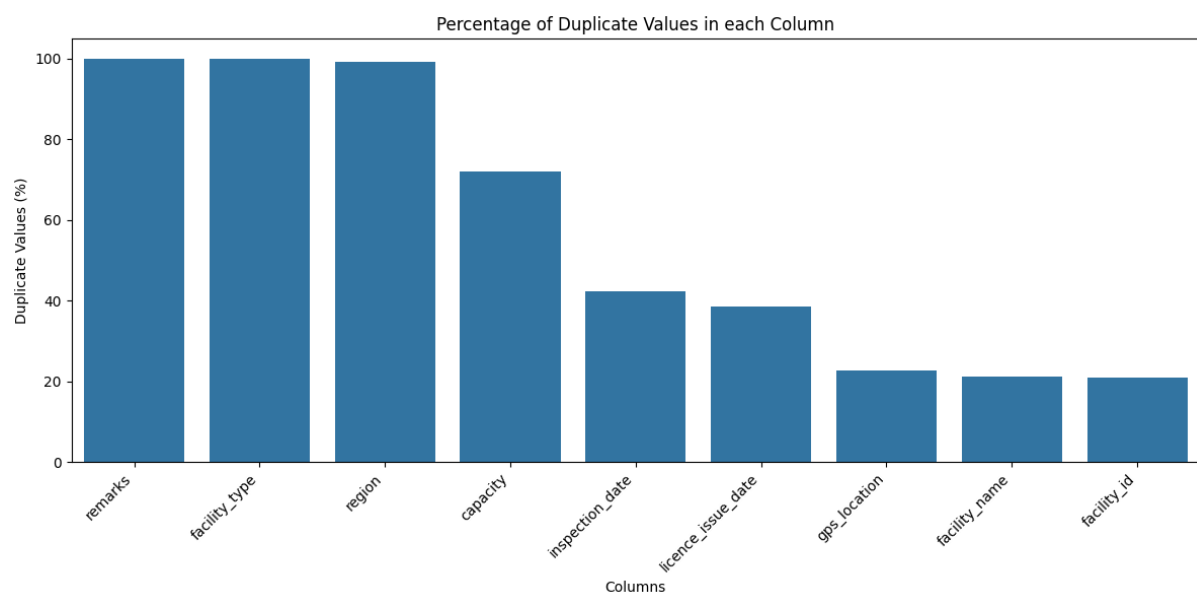
Data profiling began by importing the CSV into a Jupyter notebook hosted by Google Colab and was set to a dataframe labelled as 'df'. Upon viewing df with a head function, the following observations were made:

1. There is no consistent naming convention used for the data values. (In the facility_id series we see values like "HF-0000", "#2", "2", then back to something like the first value, "HF-0003".)
2. There are rows that are almost duplicates, just different facility_id naming conventions. (index 2 and 3)
3. There are missing values. (capacity value in index 1)
4. Inconsistent date formats. (licence_issue_date and inspection_date)

First, missing values were counted and displayed as a percentage of their entire column.



From this graph we learned that facility_id had the highest percentage of missing values. Other columns had very high amounts of missing values as well. This was repeated for duplicate values next.



This graph shows that remarks had the highest percentage of duplicate values. This is fine, but columns like facility_id and facility_name should not have any duplicate values at all. This probably occurred because the CSV file was compiled from several different data sources which must have all had the same health care facility recorded. The amount of missing and duplicate values uncovered here will frame our thinking in the data cleaning section.

Data Cleaning

All exact duplicate rows were first removed, but after that, we still had some duplicates in the facility_id column. This has been because different health facilities were attributed the same facility_id value by accident. All rows containing duplicate facility_id values were then removed, keeping the first occurrence of them.

Afterwards, it was noted that facility_type only had 11 unique values, so we could count all equal values and display them to see what kinds of variables were in the column. This revealed that there really should only be 4 unique values in the column, the 11 values were just variations of the 4. Dictionaries were defined to replace these values into 4 unique values of "hospital", "community health centre", "clinic" and "polyclinic".

This technique was repeated for all the other columns, with some slight variations and additions as follows.

capacity – This column had alphanumeric values which we separated into numeric and letter columns. These columns were then named capacity_count and capacity_type respectively. For example: one data value was "300Beds", in this case, 300 was assigned to capacity_count while Beds was assigned to capacity_type. (Beds was also replaced by bed afterwards). We ended up with only 3 unique values: "unknown", "bed", "cot". Also, before this replacement occurred, a very frequent data value of "ten beds" was noticed. All rows with this value had to have their capacity_count column set to 10 prior to replacement, otherwise this data would have been lost. This would have been very detrimental to data quality as 10 also ended up being the mode of the column.

region – String values in this column had varying errors. Missing spaces, too many spaces, and even reversed values. Values had to be stripped, the word "Parish" was removed for redundancy, and another dictionary was used to do the rest of the cleaning. At the end, only 10 unique values remained, one being "unknown" so the remaining 9 represented the whole country. Barbados has 11 parishes, so this was an unexpected finding, it turned out that St. Philip and St. Thomas were missing from the dataset. The dataset is fictional however so perhaps these values simply were not generated. If this was a real dataset though we would have to consider whether we lost important data during the cleaning process or if our data collection method was insufficient, because it is safe to assume that all parishes would have at least one healthcare facility.

licence_issue_date and **inspection_date** – These columns were cleaned using the pandas function `to_datetime` and that seemed to work well.

remarks and **gps_location** – The remarks column was quite clean compared to the rest of the dataset. A dictionary handled the heavy lifting here. One funny finding was the number of remarks that were simply a “heart emoji”. These were replaced to “Loved it” instead as I think that would be easier to work with during an analysis of the data. **gps_location** just had its null values replaced by “unknown”.

In fact, after all this cleaning, I set all the remaining null values to “unknown”. I think this would make the data analysis section much cleaner, readable and more impactful. It would also show which variables still had too many unknown values, and thus where more data collection needs to occur. The `df_clean` dataframe was then exported as `cleaned_health_registry.csv`.