#4: Data Quality

by Diana Crowe

Issue: *does the provided data set have any significant data quality issues?*

Task: *check if there are any significant data quality issues that should be considered or fixed before someone starts using it for analytical or decision-making purposes.*

I. Introduction

In order for data to be reliable, it needs to be:

* Free of errors * Accurate * Complete

* Relevant * Accessible

Data sets picked up from disparate sources could have a number of issues, including:

- missing values,
- inaccuracies,
- duplicates,
- incorrect or missing delimiters,
- inconsistent records,
- insufficient parameters,
- Bias,
- Null values,
- Outliers, etc.

If data cannot be repaired, it must be removed.

II. Data cleaning Workflow

A data cleaning workflow includes:

1) Inspection
2) Cleaning
3) Verification

1) <u>Inspection</u>

Inspection includes:

- Detecting issues and errors,
- Validating against rules and constraints
- Profiling data to inspect source data

    - This helps you inspect the source data to understand the:
        - Structure
        - Content
        - Inter-relationships
    - It uncovers anomalies and data quality issues (blanks, Null values, duplicates, whether the value of a field falls within the expected range, …)
- Visualizing data using statistical methods
    - Can help you spot outliers

2) <u>Cleaning</u>

The technologies that are applied depend on the use case and on the type of issues that are encountered.

- **Missing values** – can cause unexpected or biased results. We can:
    - Filter out records with missing data
    - Source missing information
    - Imputate, that is, calculate the missing value based on statistical values
- **Duplicate data** – are data points that are repeated in the dataset.
    - Need to be removed
- **Irrelevant data** – is data that is not contextual to the use case.
- **Data type conversion** – is needed to ensure that values in a field are stored as the data type of that field.
- **Standardizing data** – is needed to ensure date-time formats and units of measurement are standard across the dataset.
- **Syntax errors** – such as white spaces, extra spaces, typos, and formats need to be fixed.
- **Outliers** – need to be examined for accuracy and inclusion in the dataset.

3) Verification

Verification includes:

- Inspecting results to establish the effectiveness and accuracy achieved as a result of the data cleaning.

-//-

It is important to document:

- The changes undertaken as part of the data cleaning operations.
- The reasons for undertaking those changes.
- The Quality of the currently stored data.

III. Looking at our data

When inspecting the data supplied, I found the following issues:

- **Missing data/ Blanks**:
    - Completely empty columns:
        - date_updated
        - energy_class
        - is_public_housing
    - Columns that have missing data:
        - municipality
        - postal_code
        - built_year
        - floor_area
        - number_of_rooms
        - floor_number
        - number_of_floors
        - room_type
        - building_purpose
        - building_id
        - national_building_id
        - property_type
- **Errors in the data**:
    - In the column "street_address" there are mistakes in the street names. These can be:
        - Random extra spaces, as in the case of "Pitk änkalliontie" and "V äin önkatu"
        - Missing spaces between the street name and number, as in the case of "Radanp ää6"
    - Mistakes in the column "municipality" – Municipality listed as "-".
    - Wrong values in the column "rent". There is something very fishy over monthly rents being "0", "1", "9.25" or "15" euros/month...
    - There are unexpected values in the column "built_year". Several buildings are listed as having built in the year 1000, for example in Pori (which is one of the oldest cities in Finland, established only in 1558).
    - There are unexpected values in the column "floor_area", for example areas of "0", "1", and "4.5".
    - In the column "floor_number" there are negative values (maybe below street level?) and a floor 43497 (which doesn't exist).
    - In the column "number_of_floors" there is an impossible value of 43497.
- **Duplicate data** – there is potentially duplicate data in the dataset. There are instances where the street_address, the municipality, postal_code, rent, floor_area, floor_number, number_of_rooms, number_of_floors, room_type, etc, are all the same (for example, Sirkanpolku 3 in 44200 Äänekoski, with rent 500, 75 sqm, on the 1st floor, with 3 rooms, etc).

IV. Conclusion / Suggested Fixes

- Missing data/ Blanks:
    - Completely empty columns – depending on the case use, we have to decide whether we actually need this data or not. If we do need the data, we need to source it from somewhere. If we don't, we can just remove the columns.
    - Columns that have missing data:
        - municipality – if this data is relevant to our case, we can find the missing municipality through the postal code (which is not missing in any of the cases where the municipality data is missing).
        - postal_code – if this data is relevant to our case, it should be sourced and added. It's possible to get the postal code information from knowing the municipality and street_address. If irrelevant, then there is no point in wasting time finding and adding the missing data.
        - built_year – replace all the blanks with the value Zero.
        - floor_area - replace all the blanks with the value Zero.
        - number_of_rooms - replace all the blanks with the value Zero.
        - floor_number – we can't use Zero as a replacement value to indicate a missing value in this instance as there is an actual value "0" in the data already. We could replace the missing values for example with "999" to indicate missing data as we are sure there aren't usually 999 floors in an apartment building.
        - number_of_floors – suggested replacing blanks with "999" (see above).
        - room_type - replace all the blanks with the value Zero.
        - building_purpose - replace all the blanks with the value Zero.
        - building_id - replace all the blanks with the value Zero.
        - national_building_id - replace all the blanks with the value Zero.
        - property_type - replace all the blanks with the value Zero.
- Errors in the data:
    - Mistakes in the column "street_address":
        - Random extra spaces
        - Missing spaces between the street name and number
        - Luckily, all these street names are in Finnish or Swedish. This means that they are all agglomerated into one word and we would not have any spaces naturally occurring in a street name.
        - Also luckily, in this instance there are no extra letters in the house numbers (in this data we see for example building 6 instead of 6A).
        - One way to fix: in each string element of "street_address" we start at the end.
            - We determine how many characters from the end are digits. We could use, for example, the Isalpha() function which returns true if the character in a string is a letter and otherwise returns false.
            - On the left of the digits there should be a space. If there isn't a space, we add a space.

        - On the left of the space there should now only be letters and no spaces. Search the remaining characters and, if a space if found, remove it.
      - Alternative fix: the column “street” doesn’t have any issues with random spaces. We can generate a new issue-free “street_address” column to replace our current one by getting the data from “street”, then adding a space then adding the data from “address_number” (row per row).
  - Mistakes in the column “municipality” – Municipality listed as “-”. - If this data is relevant to our case, we can find the missing municipality through the postal code (which is not missing in any of the cases where the municipality data is missing).
  - Wrong values in the column “rent”: needs to be investigated.
    - The “15” €/month rental is actually a car parking spot (“autopaikka”) so it is probably correct. We can ignore those 15 € then and leave this row be.
    - There are several instances of storage space (“varastotila”) at 0 €/month... with a kitchenette??? These rows should be flagged and potentially deleted.
    - One of the places with rent 0 has floor area “1”. They manage to fit 1h+k/2h+k into that 1sqm (mini apartment for gerbils?). Flagged for deletion.
    - The remaining rents of “0”, “1”, and “9.25” €/month all look like actual properties. If possible, we should correct the rent value. If not possible, then potentially delete these rows.
  - Unexpected values in the column “built_year” - it feels safe to assume that building dates after 1500 are probably correct and building dates before 1500 are suspicious need to be examined for accuracy and inclusion in the dataset.
  - Unexpected values in the column “floor_area” - the zero values are incorrect and we either need to obtain the missing correct data or remove the column if not necessary for our purposes. The other small values of area should be examined for accuracy and inclusion in the dataset.
  - In the column “floor_number” there are negative values (maybe below street level?) and a floor 43497 (which doesn’t exist). - examine the negative values for accuracy and inclusion in the dataset, replace the value 43497 with a correct value or remove from the dataset.
  - In the column “number_of_floors” - replace the value 43497 with a correct value or remove from the dataset.

- Duplicate data – potentially duplicate data in the dataset.
  - The instances of potentially duplicate data should be investigated. The information date_created and date_removed can help with this. For example, in the case of Sirkanpolku 3 mentioned above, it appears to be the same house that was on offer in 2019, then was removed from the market, and placed back on offer in 2020.