Assignment # 4 Data quality

Issue: does this data set has 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. You can choose the format in which you want to share those insights (anything from plain .txt to a slideshow is acceptable).

Resources: link to data (data set for analysis and field definitions)

Answer:

Incomplete information: there is a lot of fields that aren't filled in fully or are left completely blank.

- Columns "built_year", "number_of_rooms", "floor_number", "number_of_floors", "room_type", "building_purpose", "building_id" and "national_building_id" have a lot of fields that are not filled.
- All the fields are left completely blank in columns "date_updated", "energy_class" and "is_public_housing".

Duplicated data:

- There is a column called "street_address" and then two separate columns called "street" and "address_number"
- In columns "date_created" and "date_removed" the data inside those fields is the same. The time when the data of the advertisement was scraped for the first time and the timestamp when the advertisement was removed seems to have always happened at the same second at midnight.
- The data in "row_update_date" is the same in all of the fields inside that column. Not sure if that information adds any important value on those rows.
- There are two columns one named "building_type" and the other called "property_type". Both are stating the building type. Data stored in multiple languages can also create difficulties later on.

Different languages:

Special characters like umlauts can be a challenge later on if a system isn't configured to handle them. I'm sure you are very used to Finnish addresses, but the table seems to have a space in a middle of a street name every time after ä or ö and might slower down someone trying to use this table for copying addresses of it. The same problem might also be with the leading zeroes on postal codes which are currently not shown.