Data Profiling and Cleaning

1) **Spatial profiling:** Open data often comes with little or no metadata. This makes it difficult to search for and find datasets relevant for a given information need.
In this project, you will profile a collection of open data sets and *identify their spatial attributes*. Your goal is to understand the challenges involved in identifying spatial attributes and propose strategies for addressing these.

You can use existing tools as a starting point, including

- Datamart Geo (https://pypi.org/project/datamart-geo)
- Datamart Profiler (https://pypi.org/project/datamart-profiler, https://docs.auctus.vida-nyu.org/python/datamart-profiler.html)
- To visualize the profiler results, you can use the data-profile-viewer library.

You should also look for related work in the literature as well as tools that can be useful for this task.

Select a sample of datasets from NYC Open Data (https://opendata.cityofnewyork.us) large enough to enable you to make statistically significant observations (e.g., 100-200 datasets)

- Spatial information can be explicit in the form of attributes with information about latitude and longitude, but it can also be implicit in attributes containing addresses, borough names, neighborhood names, and even in the dataset name or description. Locations may also be specified over multiple columns, e.g., street address, city, state. How many of these datasets contain information about latitude and longitude? How many contain other types of spatial information? You will need to manually inspect the datasets to obtain this information.
- As a baseline, use existing tools/libraries to perform the detection of spatial attributes, and report the precision and recall.
- Based on the results, identify the limitations of the techniques, design and implement improvements or new techniques.
- Evaluate the new/improved techniques and report their precision and recall.
- Discuss when and why your approach fails.

Specific details about data sets and output formats will be provided in a separate document.

Extra credit:

1) You will use what you learned while profiling the data to identify (potential) quality issues, including, incorrect values (typos - brklyn; inconsistent zipcodes or city names), data missing for certain regions, and generate quality reports for the datasets used in your experiments.

2) For each dataset that does not have columns for latitude and longitude, you should create a new column with the latitude and longitude associated with each record in the dataset. You can use a library or service to geocode the location information you identify.

2) **Temporal profiling:** Open data often comes with little or no metadata. This makes it difficult to search for and find datasets relevant for a given information need.
In this project, you will profile a collection of open data sets and *identify their temporal attributes*. Your goal is to understand the challenges involved in identifying temporal attributes and propose strategies for addressing these.

You can use existing tools as a starting point, including

- Datamart Profiler (https://pypi.org/project/datamart-profiler, https://docs.auctus.vida-nyu.org/python/datamart-profiler.html)
- To visualize the profiler results, you can use the data-profile-viewer library.
- Dateutil (https://dateutil.readthedocs.io/en/stable)

You should also look for related work and tools that can be useful for this task.

Select a sample of datasets from NYC Open Data (https://opendata.cityofnewyork.us) large enough to enable you to make statistically significant observations (e.g., 100-200 datasets)

- How many of these datasets contain temporal information in a *timestamp* format? How many datasets contain temporal information in other formats? Note that temporal information can be explicit in the form of attributes with information about timestamps, but it can also come in different formats, or be split over multiple attributes, e.g., month, day, year; they can also be present in the dataset name or description Therefore, manual inspection will be required to obtain this information.
- As a baseline, use an existing tool/library to perform the detection of temporal attributes, and report its precision and recall.
- Based on the results, identify the limitations of the techniques and propose improvements or new techniques for detection, as well as to *normalize* the temporal information across the datasets you are working with. For each dataset, you should add a new column with the normalized value for the date/time information associated with each record.
- Evaluate the new/improved techniques.
- Discuss when and why your approach fails.

Specific details about data sets and output formats will be provided in a separate document.

Extra credit:

1) You will use what you learned while profiling the data to identify potential quality issues, including, e.g., incorrect values, data missing for certain time periods in the dataset (e.g.,

the dataset covers years 2010-2020, but has no information for 2012), and generate quality reports for the datasets used in your experiments.
2) Run your improved techniques over a larger number of datasets, e.g., all datasets in NYC open data, and evaluate their effectiveness. For example, you could obtain a random sample and verify the precision of your approach.

3) **Data quality assessment:** Open data often has quality issues. These can negatively impact the results of analyses or data-driven models based on the data. In this project, your goal is to identify quality issues in datasets from NYC Open Data (https://opendata.cityofnewyork.us). In particular, you will develop methods to automatically identify suspicious/anomalous values in tables, along with the reason why. These values may be explicit null values (e.g., N/A), disguised null values (e.g., 999-999-9999 in a phone number column), syntactic outliers (e.g., alpha-numeric strings in a column consisting of alphabetic last names), semantic outliers (e.g., a state name in a column of city names), misspellings (e.g., brokklyn vs brooklyn), etc.

You can use existing profiling tools to help you better understand the data, e.g., Datamart Profiler (https://pypi.org/project/datamart-profiler, https://docs.auctus.vida-nyu.org/python/datamart-profiler.html). To visualize the profiler results, you can use the data-profile-viewer library.

You should also look for related work in the literature as well as tools that can be useful for this task.

You will work with a sample of datasets from NYC Open Data (https://opendata.cityofnewyork.us).

Different features can be used to identify outliers within a list of column values. Example features are value length, frequency, character composition, special characters, etc.. What combinations of these features are most suitable to discover the different types of quality issues? A second approach is to use similarity (e.g., edit distance, n-grams, phonetic similarity) between values. This approach works best when given a controlled vocabulary of valid terms. Such vocabularies can either be extracted from online data sources (e.g., list of US cities) or from the data at hand itself (e.g., the most frequent names across many city name columns can be used as a seed for a curated list of city names). A third source of information is the co-occurrence with values in other columns (e.g., two city names that are similar in edit distance and that have the same ZIP code are likely to be the same).

Your task is to develop (semi-)automated techniques that classify column values into one of the following four categories:

1) Valid value
2) Misspelling/Abbreviation of a valid value

3) Invalid value
4) NULL value

For values in class 2 list the valid value. For invalid values, flag them as semantic outliers (e.g., values that are valid in a different column/domain) if appropriate.

You should also look for related work in the literature as well as tools that can be useful for this task.

Note that NULL values can be represented in many different ways (e.g., NULL, n/a, 999, 999-999-9999) and there are different types of outliers, therefore, manual inspection will be required to obtain this information. You can use existing profiling tools to help you better understand the data, e.g., the Datamart profiler (https://docs.auctus.vida-nyu.org/python/datamart-profiler.html) and data-profile-viewer, both available as Python packages:
https://pypi.org/project/datamart-profiler
https://github.com/soniacq/DataProfileVis

You should also try to answer the following questions:

- Are there patterns for how NULL values/outliers are represented?
- What is the precision and recall of the techniques you designed/implemented?
- When and why does your approach fail?

Specific details about data sets and output formats will be provided in a separate document.

Some potentially useful references:
Efficient Algorithms for Mining Outliers from Large Data Sets. Ramaswamy et al., SIGMOD 2000.
(ftp://ftp10.us.freebsd.org/users/azhang/disc/disc01/cd1/out/papers/sigmod/efficientalgorisrrak.pdf);
Ming Hua, Jian Pei: Cleaning disguised missing data: a heuristic approach. KDD 2007: 950-958

Data Analysis

4) **Understanding the Impact of the COVID-19 Pandemic:** The COVID-19 pandemic has touched many different aspects of our lives, at an individual level (e.g., how residents move in cities, the use of masks) and at a macro level (e.g., economic implications). Some of these effects can be captured from data, by comparing indicators before and during the pandemic, and also by analyzing how these effects evolve over space and time.

Different cities and states have adopted different strategies to contain the pandemic. For example, in New York, the governor ordered all non-essential business to be closed. This in

turn, led to the decrease of use of public transportation and taxis and the number of people on the streets. The disease spread was eventually contained, businesses re-opened, and (some) life returned to the city. There are many open data sets in NYC that serve as sensors for City life, e.g., subway turnstile data, taxi and for-hire-vehicles, bus data, 311 complaints, crime. There are also other sensors, including the number of flights that arrive in a city and their origin, level of pollution, sounds. What story does city data tell about the pandemic? How has the city changed? Can we find a relationship between the trends in these data compared to how the number of cases and deaths change over time? Do trends differ based on region (e.g., different neighborhoods with different demographics)?

See an example of what changed in the sounds of NYC in
https://www.nytimes.com/interactive/2020/08/20/nyregion/nyc-sights-sounds-coronavirus.html

In the project, you will:

- Select an aspect of New York City and analyze how it has changed (e.g., transportation, noise) -- what it was like before and how it changed over the different stages of the pandemic. You should articulate a set of questions you want to answer or hypotheses you want to test.
- Do a literature search to find relevant work on the aspect/topic you choose.
- Identify and obtain the datasets needed to answer your questions (you must use more than one dataset).
- Clean and integrate the data: describe the steps you performed to clean and integrate the datasets, and discuss the challenges you faced.
- Analyze the data and report findings: describe the methodology you used to answer your research questions, the challenges you faced, and your findings.