How did the acceptance of Social Distancing evolve during the Covid-19 pandemic in New York City?

Exam project for Management of Scientific Data

Anna Sterzik

Friedrich Schiller Universität Jena

August 11, 2020

Table of Contents

Project Information

Project name:	How did the acceptance of Social Distancing evolve during the Covid-19 pandemic in New York City?
Creator:	Anna Sterzik
Affiliation:	Friedrich Schiller University Jena
Template:	DCC Template Last modified: 10-08-2020 The tool DMPonline was used

Preexisting Data

- Pre-existing data from 311 Service Requests from 2010 to Present.
- Initial Data Filtering:

Dates:	020/03/01 - 2020/08/11
Description:	Including ”Social Distancing”

- raw data volume: 12.7 GB
- Data Format: saved as .csv
- Licence: Open Data Law(?)
- Accessed/Downloaded 2020/08/10 22:14:39 CEST

Generated Data

- Data will be filtered, analyzed and visualized using jupyter notebooks, such that there is no need to modify the raw data. Everything will be done by the skripts.

- Required additional storage space for processed data and secondary outputs therefore negligible.

- Formats: ipynb and pdf and tex

- Versioning will be done with git.

Documentation and Metadata

- Software versions used for this project:
- Documentation will be provided as a README
- Provenance will be handled by ProvBook

Not applicable

- Ethics and Legal Compliance

Storage and Backup

- Data will be stored with URZ, on github and on a DVD/USB stick (3-2-1 Backup rule) I will be responsible for backing up onto USB stick, URZ and github are managed(?)

- Data will be freely available for everyone at all times via github. (Therefore no security concerns)

Selection and Preservation

- Critical part of this project is the created software only, not the third party data which is already taken care of by 311 Service Calls data base. Therefore this needs to be preserved

- The data will be preserved in a github repository,usb and urz.

Data Sharing

- How will you share the data? github weil es software project ist.

- Are any restrictions on data sharing required? No restrictions required.

Responsibilities and Resources

- Who will be responsible for data management? I will be responsible for all that stuff because it is only a very small project.

- What resources will you require to deliver your plan? Only ressources required is storage capacity from URZ.

311 Service Requests in New York City from 2010 to present

- Non-emergency Social Service Requestss
- Provider: DoITT Department of Information Technology & Telecommunications
- Owner: NYC OpenData
- 41 Columns in the Dataset include Unique Key, Information about Dates (Opening, closing), Agency, Complaint Type, Location Information
- Each row is a service request

Quality Control

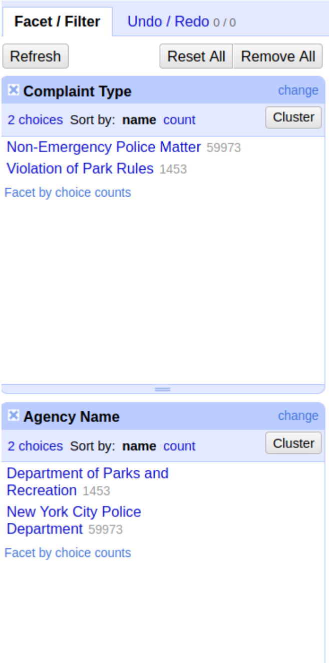
- Quality control will be done using OpenRefine

- The database states:
  “NOTE: This data does not present a full picture of 311 calls or service requests, in part because of operational and system complexities associated with remote call taking necessitated by the unprecedented volume 311 is handling during the Covid-19 crisis. The City is working to address this issue.”

- One can also see at first glance that there are several missing values

Facets

- Facets can be used to get a better overview over the data in specific columns.
- The Complaint Types and Agency Names seem to be reasonible.



Complaint Type

2 choices Sort by: **name** count

Non-Emergency Police Matter 59973
Violation of Park Rules 1453

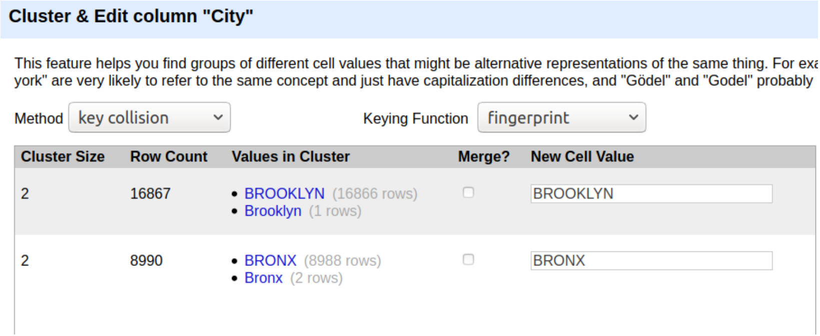
Agency Name

2 choices Sort by: **name** count

Department of Parks and Recreation 1453
New York City Police Department 59973

Clustering

Clustering is another option to identify erroneous data, especially spelling mistakes.



Cluster & Edit column "City"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For exa
york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably

Method: key collision

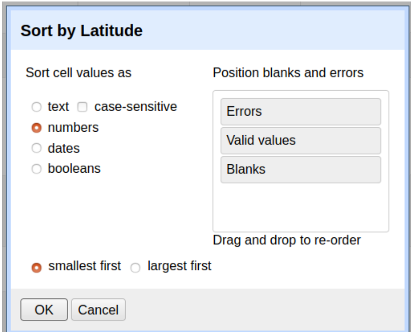
Keying Function: fingerprint

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	16867	• BROOKLYN (16866 rows) • Brooklyn (1 rows)	□	BROOKLYN
2	8990	• BRONX (8988 rows) • Bronx (2 rows)	□	BRONX

Sorting

Using OpenRefine one can also sort the values by certain columns. That way one can e.g. determine if the given longitudes and latitudes are reasonable. Here the latitudes and longitudes seem to be valid for NYC.

The same can be done for the dates. The creating dates for example start with 03/28/2020 and end with 08/10/2020. This seems to be right as well, because PAUSE started at 03/22/2020 (https://www.governor.ny.gov/news/governor-cuomo-signs-new-york-state-pause-executive-order)



Sort by Latitude

Sort cell values as

○ text □ case-sensitive
● numbers
○ dates
○ booleans

Position blanks and errors

- Errors
- Valid values
- Blanks

Drag and drop to re-order

● smallest first ○ largest first

There is more

Open Refine offers many more tools for Data Cleaning such as:

- Trimming of Leading and Trailing Whitespaces

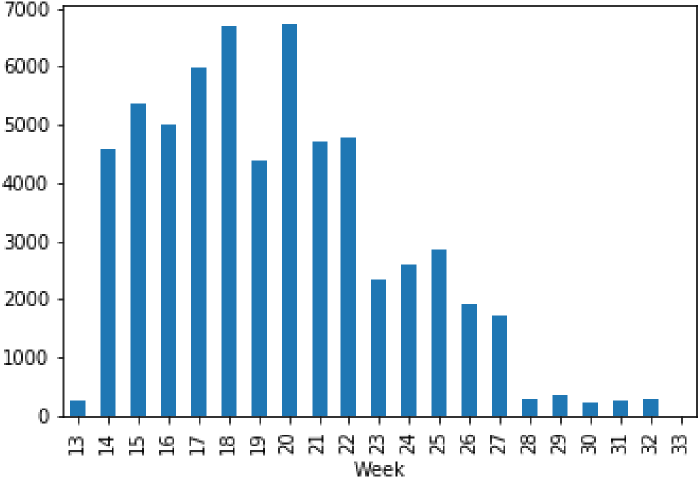
Saving

OpenRefine projects can be exported. The resulting files do only contain .txt files and .json files. These files describe all changes made with the data -¿ provenance is automatically included.

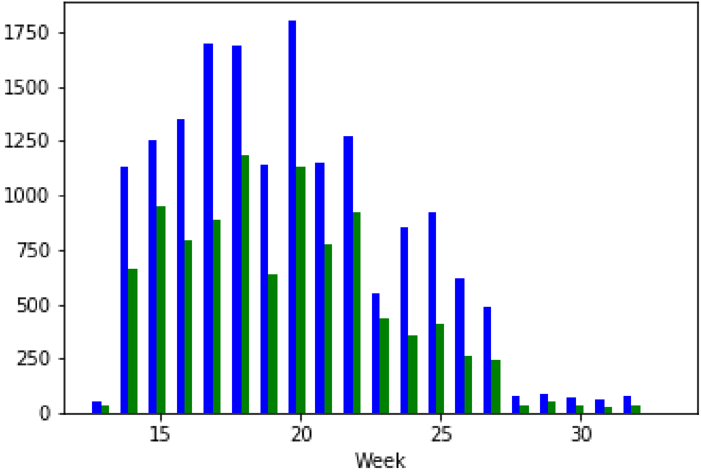
Data Analysis

Data analysis will be done using pandas library in a jupyter notebook environment.

Number of Service Calls about 'Social Distancing' in calendar weeks



Comparison of 'Social Distancing' Service Calls in Bronx and Manhattans



Preservation and Publishing

- Publishing on Github
- Backup copies with the URZ and a USB drive as well
- Material available on Github under a MIT Licence