

# 311 Social Distancing NYC

Exam project for Management of Scientific Data

Anna Sterzik

Friedrich Schiller Universität Jena

August 11, 2020

# Table of Contents

Data Management Plan

Description of the Dataset

Quality Control

Data analysis

Preservation and Publishing

# Project Information

Project name: 311 Social Distancing NYC

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:  
[Description:](#) Including "Social Distancing"
- ▶ raw data volume: 32.8 MB
- ▶ Data Format: CSV
- ▶ Open Data <https://opendata.cityofnewyork.us/faq/>
- ▶ Accessed/Downloaded 2020-08-11

# Generated Data

- ▶ Data Quality will be monitored using [OpenRefine](#). For every version of refined data the OpenRefine project will be saved together with a version number.
- ▶ Data will be analyzed and visualized using jupyter notebooks.
- ▶ Formats: TXT, JSON, PDF, PNG, TEX, IPYNB
- ▶ Everything apart from raw data will be put under version control by using git.

# Documentation and Metadata

- ▶ Software versions used for this project:

OpenRefine: 3.3

Python: 3.7.4

Pandas: 0.25.1

Jupyter: 1.0.0

Matplotlib: 3.1.1

Numpy: 1.17.4

- ▶ Documentation will be provided as a README
- ▶ Provenance for Data Cleansing by usage of OpenRefine
- ▶ Provenance for Jupyter Notebooks will be handled by [ProvBook](#)

# Storage and Backup

- ▶ Project will be hosted on github and additional backup will be with URZ and on a USB stick
- ▶ Data will be available for everyone at all times via github.

# Selection, Preservation and Sharing

- ▶ The created software for analysis as well as the steps during data cleaning are essential part. The third party data is already preserved.
- ▶ The project will be hosted on github and will be available under a MIT licence.



# Resources

- ▶ The only resources required are storage capacity from URZ.

# Description of the Dataset

311 Service Requests in New York City from 2010 to present

- ▶ Non-emergency social service requests
- ▶ Provider: DoITT Department of Information Technology & Telecommunications
- ▶ Owner: NYC OpenData
- ▶ There are 41 columns in the dataset, they include but are not limited to unique key, information about time, 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 reasonable.

Facet / Filter

Undo / Redo 0 / 0

Refresh

Reset All

Remove All

✕ Complaint Type

change

2 choices Sort by: name count

Cluster

Non-Emergency Police Matter 59973

Violation of Park Rules 1453

Facet by choice counts

✕ Agency Name

change

2 choices Sort by: name count

Cluster

Department of Parks and Recreation 1453

New York City Police Department 59973

Facet by choice counts

# 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 example, "New York" and "NYC" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably

Method key collision

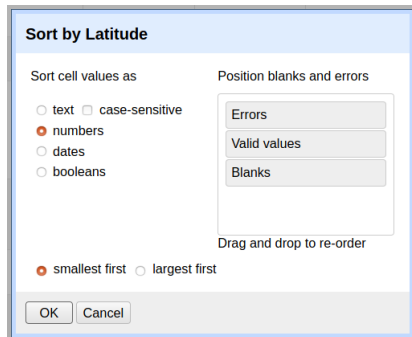
Keying Function fingerprint

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	16867	<ul style="list-style-type: none"><li>BROOKLYN (16866 rows)</li><li>Brooklyn (1 rows)</li></ul>	<input type="checkbox"/>	<input type="text" value="BROOKLYN"/>
2	8990	<ul style="list-style-type: none"><li>BRONX (8988 rows)</li><li>Bronx (2 rows)</li></ul>	<input type="checkbox"/>	<input type="text" value="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 2020-03-28 and end with 2020-08-10. This seems to be right as well, because PAUSE started at 2020-03-22.



The screenshot shows the 'Sort by Latitude' dialog box in OpenRefine. The dialog has a light blue header with the title 'Sort by Latitude'. Below the header, there are two main sections. The first section, 'Sort cell values as', contains five radio button options: 'text', 'case-sensitive', 'numbers' (which is selected with an orange dot), 'dates', and 'booleans'. The second section, 'Position blanks and errors', contains three stacked buttons: 'Errors', 'Valid values', and 'Blanks'. Below these buttons is the text 'Drag and drop to re-order'. At the bottom of the dialog, there are two radio button options: 'smallest first' (selected with an orange dot) and 'largest first'. At the very bottom, there are two buttons: 'OK' and 'Cancel'.

# Saving

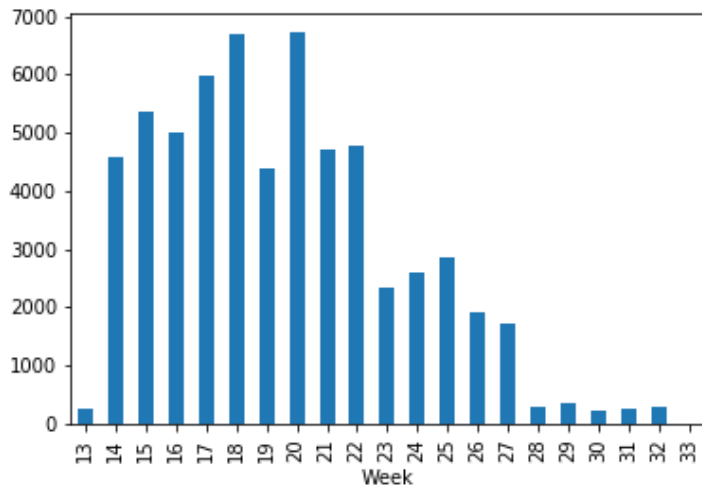
OpenRefine projects can be exported. The resulting files do only contain TXT files and JSON files. These files describe all changes made with the data.

# Data Analysis

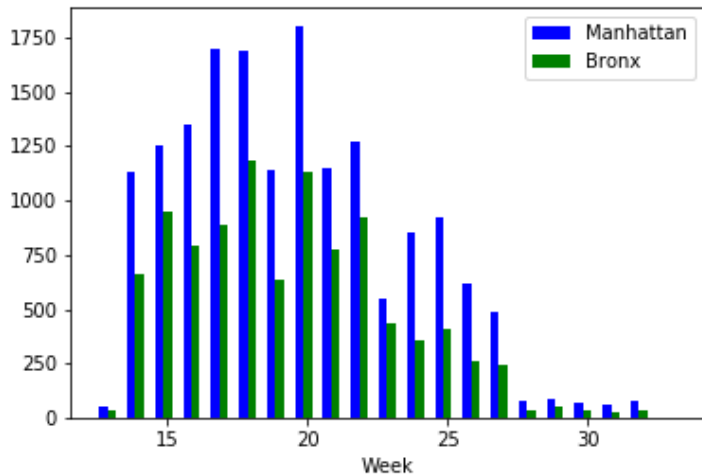
Data analysis will be done using pandas library in a jupyter notebook environment.



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



# Comparison of 'Social Distancing' Service Calls in Bronx and Mannhattans



# Preservation and Publishing

- ▶ Publishing on Github:  
[github.com/azuki-monster/311-Service-Calls-NYC](https://github.com/azuki-monster/311-Service-Calls-NYC)
- ▶ Backup copies with the URZ and a USB drive as well
- ▶ Material available on Github under a MIT Licence