Management of Scientific Data

~ Exam ~

Farin Lippmann

Dataset and Research Question

311 Service Requests in New York City

“Does Brooklyn get more rodent reports than usual after thanksgiving?”

DMP - Data Description

- 311 Service Requests in New York
- 2010 to 2022
- from the City of New York Open Data Portal
- tabular data, UTF-8 encoded .csv files
- extract relevant data and analyse using plots

DMP - Data Description

Raw Data

- 400 thousand rows
- 41 columns
- 260MB
- not reproducible

Processed Data

- 395 thousand rows
- 1 column representing dates
- 13MB
- reproducible given raw data

DMP - Data Quality

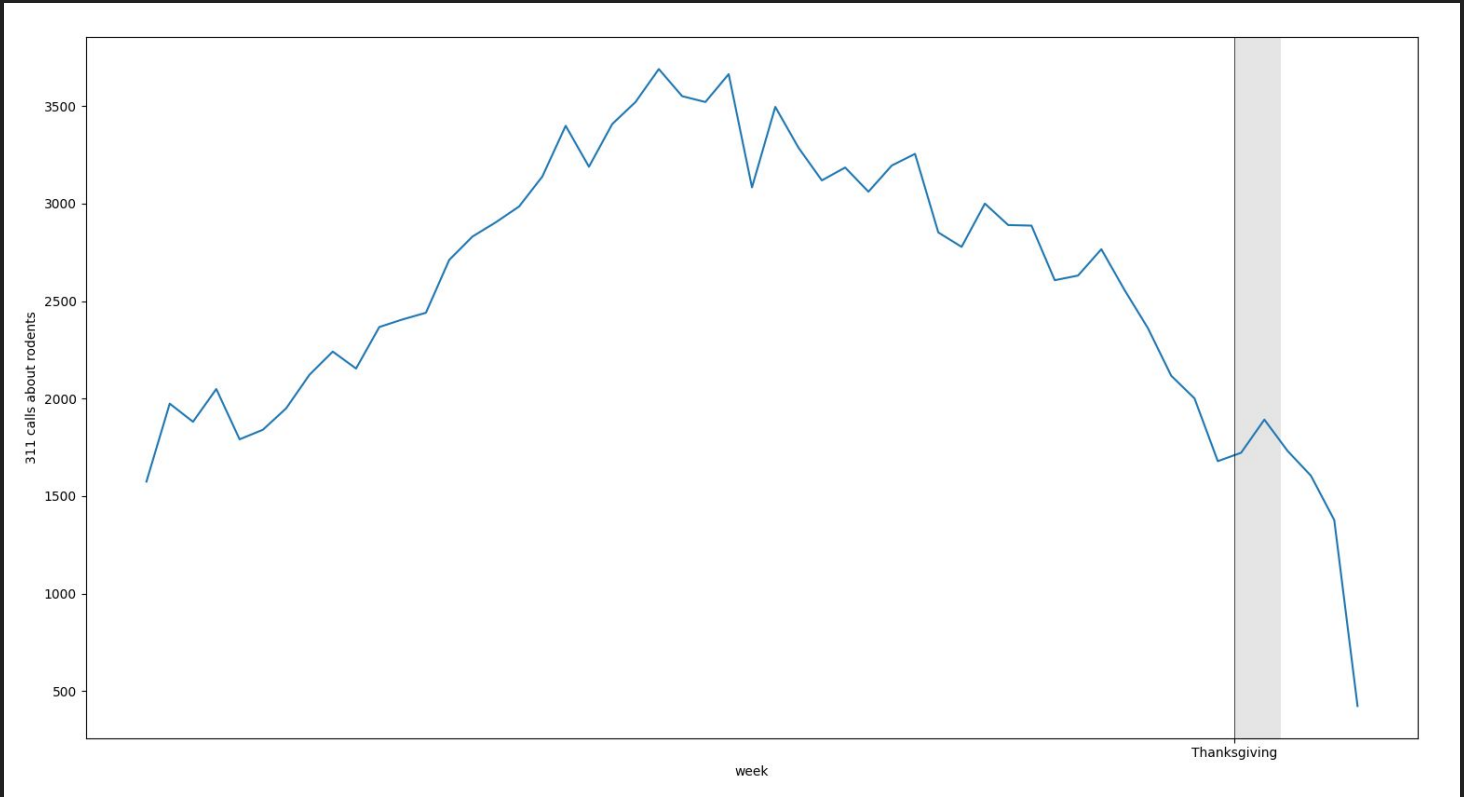
Problems

- missing values
- non-uniform format (“`brooklyn`”, “`Brooklyn`”, “`BROOKLYN`”)

Solutions

- remove rows with missing values `df = df.dropna()`
- trim whitespace `df['City'] = df['City'].apply(lambda x: x.strip())`
- turn all text lowercase `df['City'] = df['City'].apply(lambda x: x.lower())`
- tools used: python, pandas

Analysis


3500
3000
2500
2000
1500
1000
500
311 calls about rodents
Thanksgiving
week

DMP - Metadata

Raw Data

- metadata taken from the Open Data Portal

Processed data

- each row is creation date of a service request in the format “DD.MM.YY HH:MM”

Formalization

- for example using W3C tabular metadata standard

DMP - Provenance

Data Set Source (URL)

Query

- URL
- Description:
  - "Created Date" earlier than 06.07.2023 13:00
  - "Complaint Type" is Rodent

Code

DMP - Storage

Storage

- data is relatively small
- personal computer, laptop and university cloud (Nextcloud) for sharing and as a backup

Security

- all storage locations are password-protected
- university cloud is secured using two-factor authentication

DMP - Legal Conditions

Copyright

- according to the terms of use of the City of New York Open Data platform, all data is freely usable

Personal Data

- as the data has already been published by the City of New York, I expect no problems relating to personal data

DMP - Preservation and Publishing

What to publish/preserve

- the raw data
- metadata and provenance

Where to publish/preserve

- data repository zenodo.org
- tape storage of the FSU
- CC0 license

DMP - Resources and Responsibilities

Resources

- 300MB cloud storage for a week (supplied by FSU)
- 300MB long-term storage for 10 years (supplied by FSU)
- ~4h for data cleaning, uploading, writing DMP by me

Responsibility for Data Management

- me

DMP - FAIR Principles

Findability

- DOI, metadata, keywords on zenodo.org

Accessibility

- open access on zenodo.org

Interoperability

- human and machine readable metadata

Reusability

- detailed metadata and provenance, CC0 license

~ question time ~