Journal of Data Science 0 (0), 1–21
??? 2026
DOI: 10.6339/26-JDS1252
Data Science in Action

Principles for Open Data Curation: A Case Study with the New York City 311 Service Request Data

David Tussey[1,*] and Jun Yan[2]

[1]*NYC Department of Information Technology and Telecommunications, New York, NY, USA*
[2]*Department of Statistics, University of Connecticut, Storrs, CT, USA*

Abstract

In the early 21st century, the open data movement transformed societies and governments by promoting transparency, innovation, and public engagement. The City of New York (NYC) was at the forefront of this movement since the enactment of the 2012 Open Data Law, which led to the creation of the NYC Open Data portal. The portal currently hosts over 2,700 datasets, serving as a resource for research across domains such as health, urban development, and transportation. However, the effective use of open data depends critically on data quality and usability, both of which remain insufficiently addressed in the literature. This paper examines these issues through a case study of the NYC 311 Service Request dataset, identifying key concerns related to data validity, consistency, and curation efficiency. This paper proposes a set of data curation principles tailored for government-released open data to address these issues. Our findings highlight the importance of harmonized field definitions, streamlined storage, and automated quality checks, and offer practical guidelines for improving the reliability and utility of open datasets.

Keywords *data cleansing; data quality; data science; data validation; NYC open data*

1 Introduction

In the early 21st century, the open data movement began to take shape, driven by the fundamental belief that freely accessible data can transform both societies and governments. This movement champions the principles of transparency, innovation, and public engagement. A landmark in this journey was the launch of the United States' Data.gov portal in 2009 (General Services Administration, 2024), a pioneering platform in making government data widely accessible. Shortly afterward, the European Union followed suit, unveiling its Open Data portal (European Union Publications Office, 2024) in 2012, further cementing the movement's global reach. Another example is the World Bank's Open Data initiative, launched in 2010, which provides a comprehensive repository of global development data (World Bank, 2024).

These initiatives represent significant strides in democratizing data, removing barriers that once kept valuable information regarding government performance in silos. Their collective impact extends beyond mere data sharing to fostering a culture of openness that benefits individuals, communities, governments, and economies worldwide (Barns, 2016; Wang and Lo, 2016). Despite these global efforts, the effective use of open data depends critically on effective data curation, particularly with respect to quality, consistency, and usability. Indeed, poor data quality is a well-documented barrier to the use of open government data (Janssen et al., 2012).

*Corresponding author. Email: davidtussey@gmail.com.

© 2026 The Author(s). Published by the School of Statistics and the Center for Applied Statistics, Renmin University of China. Open access article under the CC BY license.
Received February 4, 2026; Accepted October 29, 2026

The City of New York (NYC) emerged as a leader in the open data movement, marked by the enactment of the 2012 Open Data Law (New York City Council, 2012; Zuiderwijk and Janssen, 2014). This landmark legislation led to the creation of the NYC Open Data portal (NYC Open Data Team, 2024), which today hosts an impressive array of 2,700 datasets across 80 different city agencies. This wealth of data serves as a powerful tool for researchers and policymakers, significantly enhancing local government transparency. Popular datasets include information on restaurant health inspection violations, car crashes, high school and college enrollment, jail inmate charges, and the location of citywide free Internet access points. These datasets have been applied in civil life in various ways, such as mapping car crashes involving pedestrians and visualizing high school and college enrollment trends. These initiatives have enabled research across multiple domains, including health (Cantor et al., 2018; Shankar et al., 2021), urban development (Neves et al., 2020), and transportation (Gerte et al., 2019), and have advanced efforts to address complex urban issues.

However, open data present substantial curation demands. Data curation—the process of organizing, maintaining, and ensuring dataset quality—is essential to maximizing their utility. Effective management ensures that datasets remain consistent, accurate, and useful across diverse applications. For example, well-curated data are essential for machine learning systems, which require high-quality data to produce reliable insights (Polyzotis et al., 2019; Jain et al., 2020). A critical component of data curation is data cleansing, which involves identifying and rectifying inconsistencies, errors, and inaccuracies in datasets. While open data initiatives have made vast amounts of data available, ensuring the reliability and utility of such data hinges on rigorous data cleansing processes. Efficient usage of software tools can significantly streamline this aspect of curation (e.g., Cody, 2017; Van der Loo and De Jonge, 2018). Poor curation may result in issues such as missing data, formatting errors, or inconsistent values, leading to biased or inaccurate outcomes (Geiger et al., 2020). This is particularly critical in domains where machine learning is applied to sensitive tasks, such as public health or policy (Rahm and Do, 2000).

Research on data curation has examined these issues extensively. Among the earliest contributions, Witt et al. (2009) proposed data curation profiles tailored to specific contexts, establishing a foundation for targeted data management strategies. In examining broader issues in data sharing and management, Borgman (2012) highlighted the complexities of research data distribution and emphasized the need for robust strategies. This is complemented by Hart et al. (2016), who outlined essential principles for effective data management, particularly emphasizing meticulous curation practices. The value of curated open data is evident in public health and global contexts. For example, Cantor et al. (2018) demonstrated its application in evaluating community health determinants, while Shankar et al. (2021) highlighted its critical role during the COVID-19 pandemic in managing collective responses. At a broader level, the FAIR Guiding Principles hold that data should be findable, accessible, interoperable, and reusable (Wilkinson et al., 2016). These principles, however, focus largely on how data are described and shared rather than on the quality of the data content itself. Despite these advances, there remains a gap in providing actionable frameworks for managing large-scale municipal datasets.

By addressing this gap, this study transitions from theory to practical insights, aiming to establish principles that enhance the reliability and utility of open government data. This paper makes two primary contributions. First, it examines key aspects of data curation through a case study of the NYC 311 Service Request (SR) dataset. This dataset serves as a prime example for examining key issues in data curation: data validity, data consistency, and curation efficiency. Examples drawn from the 311 SR data illustrate these points. Second, this study proposes a set

of data curation principles tailored for government-released open data. These principles address the unique requirements and constraints observed in the management of such datasets.

The paper is organized as follows. Section 2 provides an overview of the NYC 311 SR system and presents a summary of SR counts over a five-year period (2020–2024). Section 3 examines specific data cleansing issues affecting data quality, including structural inconsistencies, adherence to the data dictionary, and missing entries. This section also investigates data field compliance with reference or acceptable values, highlights logical inconsistencies, and examines notable patterns in the data. This study examines the balance between precision and accuracy, identifies duplicate or redundant fields, and provides observations on the 311 SR Data Dictionary. Section 4 offers practical recommendations for mitigating or resolving these issues, while Section 5 encapsulates key insights and discusses the broader implications of our findings.

2 NYC 311 Service Request (SR) Data

The NYC 311 service, a critical component of New York City's public engagement and service response framework, serves as a hub for non-emergency inquiries and requests. Introduced in 2003, the NYC 311 system streamlined responses to non-emergency issues, ranging from noise complaints to street maintenance requests. Initially a phone-based call center, the system evolved into a comprehensive data management platform handling millions of requests annually. Key milestones since its launch in 2003 include the addition of online and mobile channels in 2009, a record monthly high of 348,463 SRs in August 2020 during the COVID-19 pandemic, the 2021 expansion to include the Metropolitan Transit Authority (MTA) NYC subway system (the biggest expansion in 311 history), and a yearly record of 3.5 million SRs in 2024.

Today, the NYC 311 data system manages over 3 million Service Requests (SRs) per year. The data are publicly accessible through the NYC Open Data portal, which provides tools for querying, aggregating, visualizing, and exporting results. The NYC Office of Technology and Innovation (OTI) provides technical support for the 311 application, the open data infrastructure, mobile, and web applications, and the public-facing data portal. Input data for the 311 system are sourced from 16 different NYC agencies. Some of these agencies use the core 311 software, while others transfer data from their in-house systems to the 311 Open Data Warehouse.

The impact of NYC 311 data extends beyond operational efficiency; it has become instrumental in shaping city governance and community engagement. Open data not only ensure governmental transparency but also empower civic developers, the general public, and policy-makers (Minkoff, 2016; O'Brien et al., 2017; Kontokosta and Hong, 2021). The data have been pivotal in providing advice on shelters during emergencies, handling inquiries during the COVID-19 pandemic, enforcing standards between landlords and tenants, reallocating taxi routes based on analyses by the Taxi and Limousine Commission (TLC), and improving responsiveness across city agencies. It also supports a wide range of studies in resource allocation and emergency response strategies (Zha and Veloso, 2014; Raj et al., 2021), social equity in service provision (White and Trump, 2018; Kontokosta and Hong, 2021), and urban issues such as noise pollution (Dove et al., 2022) and street flooding (Agonafir et al., 2022).

Our investigation uses a 311 SR dataset covering a five-year period (2020–2024), collected on 23 September 2025. This dataset is used to conduct analysis in data consistency, data validity, redundant fields, and other data quality metrics. Because the portal restructured the dataset in December 2025 (NYC Open Data, 2025), the extract used here is archived on Figshare (NYC Office of Technology and Innovation, 2025); the supplementary material describes the changes

and how to reproduce the analysis. Characteristics include:

- The dataset is approximately 9 GB in size and contains over 16 million rows, each row representing a single SR. The dataset exports as a CSV file. The analysis was conducted in R (R Core Team, 2025) using the `data.table` package (Barrett et al., 2025); for efficiency, the file was converted to RDS format.
- Each record has 41 columns of data, representing a single SR (complaint, issue, request).
- There are 252 distinct `complaint_type` categories. The top 20 account for 68 % of all SRs (illegal parking, noise, heat/hot water, trash, etc.)
- Each SR has four date fields in the MM/DD/YYYY HH:MM:SS AM/PM format.
  - `created_date`
  - `closed_date`
  - `resolution_action_updated_date`
  - `due_date`
- There are two borough fields, `borough` and `park_borough`, which appear to be duplicates.
- The dataset contains seven street-related fields, including two likely redundant pairs, e.g., cross_street_1 & intersection_street_1 and cross_street_2 & intersection_street_2.
  - `cross_street_1`
  - `cross_street_2`
  - `intersection_street_1`
  - `intersection_street_2`
  - `landmark`
  - `street_name`
  - `incident_address`
- In addition to `incident_address`, each SR has six additional geographic fields:
  - `latitude` & `longitude`
  - `location`
  - `street_name`
  - `landmark`
  - `block`
  - `x & y_coordinate_state_plane`

Noise-related complaints are especially prominent, represented by eight separate categories (e.g., vehicle, residential, helicopter, street/sidewalk). Collectively, these noise complaints account for nearly one-quarter (23.4 %) of all complaints, making *noise* by far the most common complaint. Other major categories include illegal parking, heat/hot water, and blocked driveway.

Data anomalies are associated with specific agencies, facilitating targeted corrective actions. SR volume is concentrated among six agencies which together handle 90 % of all SRs, while the remaining 10 % are distributed across 10 additional agencies. The concentration of SRs among these six agencies underscores the necessity for optimal data curation processes within.

3 Data Cleansing Issues

Data cleansing refers to the process of identifying and rectifying errors, inconsistencies, and inaccuracies within datasets to ensure they are of high quality and reliable for analysis (Maletic and Marcus, 2005; Hosseinzadeh et al., 2023). The process typically involves removing duplicate records, handling missing or incomplete data, correcting mislabeled or inaccurate values, and standardizing data formats (e.g., Cody, 2017; Van der Loo and De Jonge, 2018). In the context of open data, cleansing is especially important as open datasets often come from diverse, often uncoordinated sources, leading to variations in data quality, completeness, and consistency.

Table 1: Top six agencies by SR volume. These 'big six' constitute 90% of all SRs.

Agency	Abbreviation	Percentage	Cumulative %
New York Police Department	NYPD	43.12 %	43.12 %
Dept of Housing Preservation & Development	HPD	19.15 %	62.27 %
Dept of Sanitation	DSNY	12.01 %	74.28 %
Dept of Transportation	DOT	6.73 %	81.01 %
Dept of Environmental Protection	DEP	5.21 %	86.22 %
Dept of Parks & Recreation	DPR	3.97 %	90.19 %

Without thorough data cleansing, the utility of open data can be limited and perhaps untrustworthy, affecting its reliability for research, policy-making, and innovation. The main purpose of cleansing open data is to ensure that they are accurate, consistent, and usable across multiple analytical platforms and by various stakeholders with myriad purposes. Data cleansing improves the trustworthiness of the data and enables more accurate analysis, better decision-making, and the improved integration of data into machine learning models or other systems.

Many quality criteria are employed to ensure high-quality data processing. One of the primary efforts is data validation, which includes several critical checks. For instance, mandatory fields must not be left empty, ensuring that key information is always captured. Additionally, certain fields must conform to specific data types, such as numeric, character, or date formats, as typically outlined in a Data Dictionary. An essential aspect of validation is domain compliance, where fields must adhere to a predefined set of values such as statuses, state names, or zip codes. Structural errors can also play a significant role in data quality, particularly when naming conventions or data entries are inconsistent. Common issues include fields that do not appear in the Data Dictionary and inconsistent usage of data fields. Redundant or irrelevant fields can clutter datasets, reducing efficiency, creating confusion, and increasing the risk of errors. Logical inconsistencies are another consideration, such as related fields that violate expected relationships, for example, a `due_date` that precedes `created_date`. Lastly, maintaining an appropriate balance between accuracy and precision is essential, as both must be carefully managed to preserve the interpretability and reliability of the data.

This study identifies issues in the 311 SR dataset, not to resolve them, but to highlight their scope, magnitude, and potential impact. Corrective action should be considered only after further investigation into the underlying cause of the issue, including discussion with the originating agency to determine whether it represents an error, an intended value, or another status.

3.1 Data Preprocessing and Standardization

Prior to analysis, the raw dataset undergoes a series of preparatory modifications to facilitate efficient processing. These steps include both structural adjustments and content standardization:

- Standardization of structure: Field names are normalized for ease of reference, and the raw `.csv` file is converted into the more efficient RDS format.
- Date handling: All date fields are converted to the `POSIXct` class and interpreted in the `America/New_York` time zone, to support analysis of Daylight Saving Time (DST) effects (discussed in Section 3.4).

- Agency harmonization: Several agencies were renamed during the study period; records were consolidated under their current agency designations.
- Validation of completeness: Missing values—recorded in multiple formats—are standardized to `NA` to facilitate processing.
- Field formatting: Text fields are converted to uppercase for consistent matching, and numeric fields are explicitly cast to the appropriate type.

3.2 Structural Issues

Structural issues refer to how data are organized, formatted, or presented within a dataset. Structural issues can complicate analyzing the data effectively, and often require extensive data cleansing efforts. Our analysis identified three structural issues:

1. **Precision vs. Accuracy**: discrepancies in the `latitude`, `longitude`, and `location` fields;
2. **Missing data representation**: inconsistent representations of missing values; and
3. **Daylight Saving Time anomalies**: systemic irregularities tied to DST transitions.

Precision vs. Accuracy in Latitude and Longitude Values The `latitude`, `longitude`, and duplicative `location` fields in the raw dataset exhibit sizable variation in numeric precision. The vast majority (90 %) of values are recorded with thirteen decimal places (e.g., `latitude` = 40.7022221884127, `longitude` = −73.9126848719299).

However, precision ranges from eight to fifteen decimal places for `latitude` and from zero to fifteen for `longitude`. A small subset of 216 records contains integer longitude values with no decimal component. Moreover, positional precision at thirteen decimal places implies a spatial resolution at the molecular scale—certainly implausible. Thirteen decimal places correspond to sub-millimeter theoretical precision, which exceeds realistic GPS accuracy. This is a classic case of excessive precision rather than meaningful accuracy. This span of precision could suggest inconsistent data capture. These coordinate fields are most plausibly the result of geocoding the SR `incident_address`, particularly given that 90% of the values share the same level of precision. It would be useful to evaluate the functionality of the geocoding software.

For reference, the official New York City Borough Boundary–Water Included (NYBBWI) dataset, published by the Department of City Planning (DCP), provides coordinates to six decimal places. Decimal precision alone does not establish positional accuracy. Original coordinates were retained without rounding and checked against a bounding box derived from the NYBBWI metadata, with a tolerance of 100 m. All nonmissing coordinates tested met this criterion except one DHS record with coordinates of approximately (0, 0), evidently a placeholder; this check does not establish that every point lies within the city polygon.

Inconsistent Representation of Missing Data The 311 SR data contains inconsistent representations of missing data, including nulls, spaces, `NA`, `N/A`, and `<NA>`. This variety of representations complicates programming efforts, increasing the likelihood of errors when searching for or excluding missing fields. A more consistent approach to treating missing data would improve usability. To address this issue, all missing values were standardized to `NA` during pre-processing, resulting in 27.86 % of the dataset being classified as missing values at the field level.

Anomalies Associated with Daylight Saving Time (DST) The 311 SR dataset captures local NYC times (e.g., `America/New_York`). Accordingly, local time changes twice each year due

to the observance of Daylight Saving Time (DST): clocks move forward one hour in the spring and backward one hour in the fall. This can introduce some unexpected results. For example, on both March 13, 2022, and March 12, 2023, there are no times between 2:00 AM and 2:59 AM. The 2:00 AM hour on those dates is simply missing owing to the clocks moving forward; at 2:00 AM it instantly becomes 3:00 AM. This creates a gap in the hourly distribution of SRs on those days. Subtracting local clock times without accounting for the transition can overstate elapsed time by one hour.

During preprocessing, several date fields were found to violate the DST spring-forward rule, with timestamps recorded during the nonexistent hour between 02:00:00 and 02:59:59 (e.g., `2024-03-10 02:00:46`, `2020-03-08 02:10:00`). Such values are invalid, as local time advances directly from 01:59:59 to 03:00:00 on those dates. These anomalies were adjusted by shifting the affected timestamps forward one hour to valid times (e.g., `2024-03-10 03:00:46`, `2020-03-08 03:10:00`). This adjustment assumes a one-hour offset; the original event times cannot be established from these values alone.

The fall DST adjustment repeats the 01:00 hour, so local timestamps without UTC offsets can appear to yield negative durations even when closure follows creation. Such ambiguity can distort performance measures, particularly for time-sensitive SRs.

The following example illustrates the problem:

- A residential noise complaint is received by 311 at 01:50 on Sunday, 2024-11-03, the date of the DST fallback transition. An SR is created.
- The NYPD responds promptly, resolving the issue in forty minutes. Without DST adjustment, the resolution would be recorded at 02:30.
- However, during this interval the local clock is set back one hour, so the resolution is time-stamped in the 311 system as 01:30 local time.
- As a result, the SR is recorded as having closed 20 min *before* it was created, yielding an apparent wall-clock difference of −20 min.

On average, 86 SRs per year exhibit this systemic issue, with a mean negative duration of −42.24 min. Although rare, these anomalies can disproportionately affect time-sensitive complaint categories such as `drug_activity`, `homeless_assistance`, and `disorderly_youth`, and they can distort aggregate performance metrics. Even a small number of implausible durations can shift summary statistics and obscure true operational patterns. Additional analysis of DST effects are found in Section 3.4.

Missing Data by Field Understanding the absence of data by field greatly aids analysis. For example, determining whether SRs were closed before or after their `due_date` is effectively impossible, as 99.67 % of the field is empty. Field completeness divides into three groups, as summarized in Table 2, which presents select fields exhibiting moderate or low completeness rates.

- Complete: 94 % to 100 % filled.
- Moderately Complete: 57 % to 88 % filled.
- Sparse/Incomplete: 0.05 % to 1.5 % filled.

For some analysis efforts, it is suggested to investigate as to why some fields are almost always blank. Additionally, field usage varies considerably across agencies. For instance, the `landmark` field is widely used across agencies, whereas `facility_type` is used exclusively by DSNY. Not surprisingly, `taxi_pick_up_location` is used exclusively by TLC. Table 3 highlights

Table 2: Fields with moderate or sparse data completeness (2020–2024).

Field	Filled count	Completeness (%)
taxi_company_borough	8,564	0.05
due_date	52,345	0.33
vehicle_type	181,533	1.13
facility_type	211,240	1.32
landmark	9,058,627	56.56
intersection_street_1	10,251,223	64.00
intersection_street_2	10,257,370	64.04
cross_street_2	11,693,684	73.01
cross_street_1	11,696,471	73.03
address_type	13,246,924	82.71
location_type	13,823,312	86.31
bbl	14,128,215	88.21

Table 3: Sample field usage containing non-null values by select agencies (2020–2024).

Field	DEP	DHS	DOB	DOT	DPR	DSNY	TLC
landmark	2,935	142,890	282	339,427	595,474	656,882	113,242
facility_type	0	0	0	0	0	211,240	0
due_date	0	0	0	0	16	52,329	0
vehicle_type	0	0	0	0	0	0	3,689
taxi_company_borough	0	0	0	0	0	0	8,564
taxi_pick_up_location	0	0	0	0	0	0	136,863
bridge_highway_name	15	45,896	0	19,234	0	894	1,564
road_ramp	15	4,614	0	19,213	0	891	1,546

the varied usage of certain fields across NYC agencies, revealing both specialization and broad usage. Fortunately, most of the fields used in this analysis are well populated.

3.3 Validating Data for Acceptable Values

Analytic efforts must ensure that fields contain valid values as identified by a reference source. Invalid values should be identified and isolated from the analysis. For some attributes, the Data Dictionary specifies domains; for others the documented domain was incomplete or inaccurate. Valid values had to be inferred from operational usage, supplementary documentation, or external datasets. Additionally, the NYC Open Data portal supplements the Data Dictionary by indicating field types such as text, numeric, date, and geospatial. The analysis identified several anomalies in which observed legal values did not align with published specifications.

The following fields were tested and found to comply with their expected domains, as determined by public usage and as evidenced in other historical datasets:

- `address_type`
- `status`
- `borough`
- `borough_boundaries`
- `park_borough`
- `data_channel`
- `vehicle_type`
- `city_council_district`

However, some fields proved problematic as to their compliance with a domain of allowable values. These include

ZIP Codes All `incident_zip` values should validate against a ZIP code database compiled from United States Postal Service (USPS) data (UnitedStatesZipCodes.org, 2025); the version used here contains 42,735 ZIP codes. The `incident_zip` field contains 0.02 % invalid values, corresponding to 3,274 errors. This includes ZIP codes such as "11111", "00000", and "00083". The presence of invalid ZIP codes is notable because this field is among the easiest to validate against an authoritative reference.

Community Boards The `community_board` field captures one of New York City's local governance units. Validation is expected to be straightforward, given the publicly available list of community boards. However, this analysis identified 56,914 (0.36 %) invalid values.

3.4 Date Field Issues

In addition to the DST-related structural anomalies discussed in Section 3.2, the dataset exhibits further temporal issues. The NYC 311 dataset includes four such fields: `created_date`, `due_date`, `resolution_action_updated_date`, and `closed_date`. These date fields often exhibit implausible or extreme relationships with one another, complicating both analysis and interpretation.

Created Date All `created_date` values were checked to ensure that none were set in the future or before the inception of the 311 system (circa 2003); all records were compliant. However, a clustering of timestamps occurs at exactly midnight (14,409 instances) and exactly noon (31,872 instances), a pattern examined in greater detail later in this section.

Due Date Anomalies The `due_date` field is used almost exclusively by DSNY and is populated in only about 0.33 % of cases. All values passed validation, and no unusual patterns were observed. However, sixteen cases, all from DPR, were found where the `due_date` precedes the `created_date`—a nonsensical condition. These ranged from 452 to 727 days early (roughly 1.2 to 2 years).

Closed Date Most nonmissing closing dates were within the expected temporal bounds. Exceptions include dates in 1899 and 1900 and four dates later than the extraction date, including one in 2033. However, there is a clustering of timestamps at exactly midnight (1,083,760; 6.87 % of `closed_date` values) and exactly noon (294,815; 1.87 %), representing an important artifact discussed further in Section 3.4. Additionally, 45,296 records show `closed_date` values preceding `created_date` values; these negative durations are examined in detail in Section 3.5.

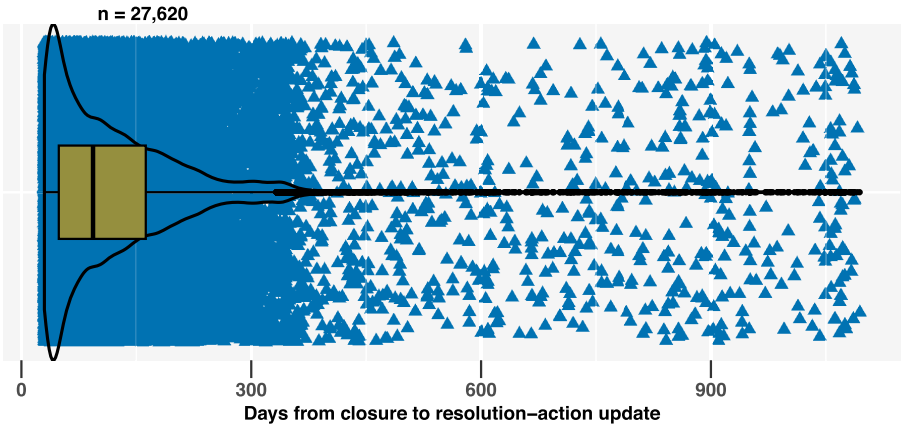


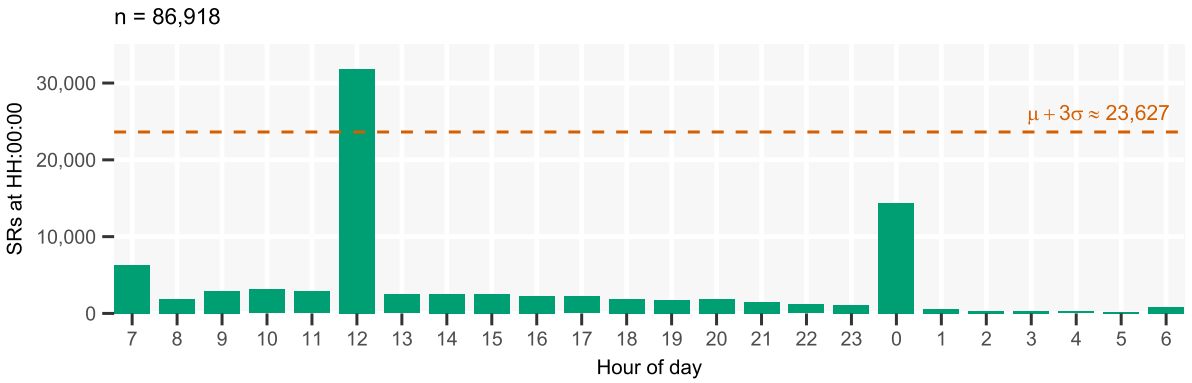
Figure 1: Days from closure to resolution-action update for 27,620 SRs with intervals greater than 30 and at most 1,095 days. Violin width represents estimated density, the boxplot summarizes the distribution, and jittered points represent individual SRs. Most intervals are concentrated within 30–200 days, with a long right tail.

Resolution Action Updated Date The `resolution_action_updated_date` field records modifications or follow-up actions related to an SR. No values were found in implausible years; however, 375,116 entries (2.34 %) occur before the corresponding `created_date`. Nearly 80 % of these premature dates occur within one day (often within seconds) of the `created_date`, suggesting that the field is populated automatically during early processing. A more consequential anomaly concerns post-closure updates.

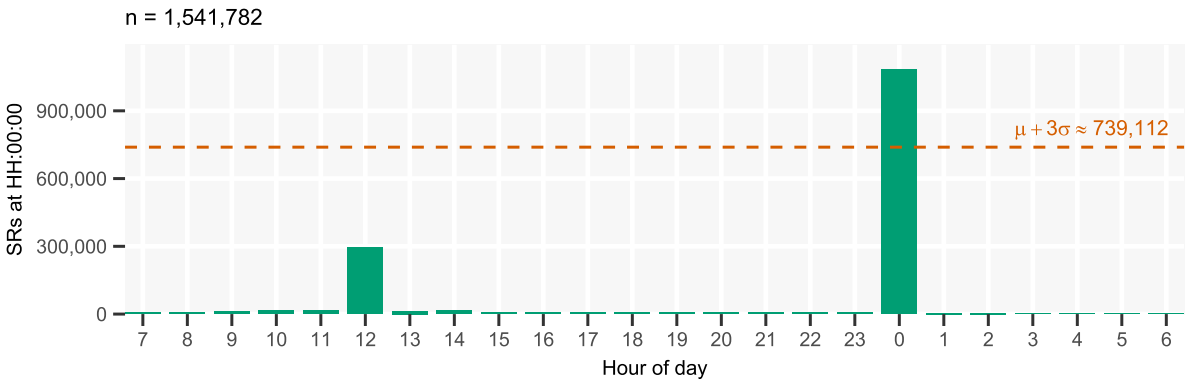
While it is certainly possible—and even routine—to update an SR after it is closed, 54 % of updates do occur after the `closed_date`. Figure 1 summarizes 27,620 SRs with updates more than 30 and at most 1,095 days after closure. This subset is concentrated within the 30–200 day range. TLC, DPR, and DOHMH account for approximately 36 %, 29 %, and 20 % of the subset, respectively. Longer intervals are excluded from the figure; some involve implausible historical closing dates. Agency consultation is needed to distinguish normal follow-up activity from timestamp errors.

Daylight Saving Time (DST) Anomalies As discussed in Section 3.2, local timestamps can be ambiguous during DST transitions. The fall-back diagnostic identified approximately 86 SRs per year with apparent negative wall-clock differences averaging −42.24 min. The spring-forward diagnostic identified approximately 98,500 SRs annually (492,652 over five years) spanning the transition. This describes exposure to the clock change, rather than confirmed errors in elapsed duration: correctly interpreted timezone-aware timestamps account for the missing hour. Reliable calculations require consistent timezone handling and, for the repeated fall-back hour, the applicable UTC offset.

Midnight and Noon: Creation and Closure Issues SR activity generally follows a typical workday pattern, with the majority of SRs created and closed during daytime hours. Closer examination of the `created_date` and `closed_date` fields reveals spikes at midnight (00:00:00) and noon (12:00:00). These spikes suggest automated batch-processing, perhaps assigning timestamps of midnight (00:00:00) to large batches of SRs prior to sending the data to the 311



(a) SRs created at the exact top of the hour. Note the spikes at noon and midnight.



(b) SRs closed at the exact top of the hour. Note the spikes at noon and midnight.

Figure 2: 311 SRs created and closed exactly at the top of the hour (HH:00:00), showing spikes at exactly midnight (00:00:00) and exactly noon (12:00:00). The orange dashed line marks $\mu + 3\sigma$, where μ and σ represent the mean and standard deviation of the hourly counts.

database. If timestamps are assigned this way, they may distort calculated durations and the apparent timing of system usage. Distribution by agency indicates that 97.6 % of the "closed-exactly-at-midnight" SRs originate from just two agencies: DSNY (60 %) and the Department of Buildings (DOB) (38 %).

As shown in Figure 2, both `created_date` and `closed_date` timestamps display peaks at noon and midnight. In the creation panel, the noon peak exceeds the $\mu + 3\sigma$ reference line, whereas the midnight peak remains below it. This descriptive screening threshold is motivated by statistical quality control (Grant and Leavenworth, 1996); it does not account for expected daily variation or establish that a peak is an error.

3.5 Duration Issues

SR duration is one of the most frequently analyzed 311 performance metrics, informing assessments of equity in service across neighborhoods and boroughs, while also serving as a proxy for agency responsiveness. While duration is not directly present in the dataset, it is easily computed as the difference between `closed_date` and `created_date`. Given the prevalence of automated workflows, one might expect these timestamps to be generated by the system itself: when an SR's `status` changes to `Closed`, for example, the `closed_date` should be recorded automati-

cally, with no manual entry. However, this appears not to always be the case, as the dataset includes anomalies in these fields that would ordinarily be prevented by robust automation.

Negative Durations A total of 45,296 SRs (0.29 %) have a `closed_date` that precedes the `created_date`, producing a nonsensical negative duration. Across all affected SRs, negative durations average −118.7 days, a substantial temporal distortion. Within this group, 115 SRs have extreme negative durations exceeding five years in magnitude. Most involve closing dates in 1899 or 1900, suggesting placeholder values or date-conversion artifacts. Their origin cannot be established from the export alone. DOT accounts for 44,143 negative-duration SRs overall; this concentration should be distinguished from the extreme subset.

Zero and One-Second Durations A more substantial issue arises when `closed_date` and `created_date` share identical timestamps to the second, creating a zero-duration value that warrants investigation. There are 387,426 (2.5 %) such zero-duration SRs.

In addition, 3,460 SRs are closed within one second of their creation. While not logically impossible, these near-instantaneous SR resolutions are statistically improbable. 92 % of these cases originate from just three agencies: the Department of Health and Mental Hygiene (DOHMH) (54 %), the Economic Development Corporation (EDC) (24 %), and DPR (14 %).

Near-Zero Durations Service request (SR) durations are right-skewed, with some cases extending for months or years. The analysis determined appropriate lower-bound thresholds for identifying implausibly small duration values. Durations used to estimate thresholds were restricted to values of at least 2 s (excluding zero- and one-second artifacts) and less than or equal to 172,800 s (48 h).

The analysis evaluated five complementary outlier screening methods: standard deviation (SD), median absolute deviation (MAD), log-normal thresholds, percentile cutoffs (1 % to 10 %), and interquartile range (IQR). The SD, MAD, and IQR methods did not flag additional anomalies, whereas the log-normal and percentile approaches produced operationally reasonable limits. The log-normal approach is used here as a screening model, rather than as an assertion that SR durations follow a log-normal population distribution.

The log-normal method uses the lower-tail cutoff $\exp(\mu_{\log} - 3\sigma_{\log})$, where $\mu_{\log}$ and $\sigma_{\log}$ are the mean and standard deviation of log-transformed durations in the restricted sample. The cutoff of 27.95 s identifies 10,231 durations of at least 2 s below this boundary. The duration-category summary instead uses a rounded boundary of at most 28 s, yielding 14,734 positive durations, including the 3,460 one-second cases.

Combining 387,426 zero-duration cases with these 14,734 short positive durations gives 402,160 SRs (2.55 % of the 15,785,257 SRs with a closing date). These are screening flags for investigation, rather than confirmed errors.

Positive Durations Some SRs exhibit very long lifespans with a total of 139,483 (0.88 %) SRs in the *large-duration* category (greater than 730 and at most 1,825 days, approximately 2–5 years). DPR, DOHMH, and EDC together account for approximately 94 % of these cases.

Additionally, 1,251 SRs have *extreme* positive durations (greater than 5 years), not counting one invalid SR whose `closed_date` is 2033. Of these, 1,176 (94 %) are DPR SRs, and 1,149 of them were closed on a single day, 22 April 2025, suggesting a bulk administrative closure. Many of these extreme DPR durations reflect tree planting, removal, trimming, or uprooting.

Likewise, DOB handles plumbing, building use, and construction issues that can involve lengthy operational or administrative processes. These complaint types imply that long durations are plausible, but do not establish that individual multi-year intervals are accurate.

3.6 Duplicate and Redundant Fields

This analysis revealed several redundant fields, indicating a potential need for consolidation:

- `latitude/longitude` & `location`: The `location` field is a pure concatenation of the `latitude` and `longitude` fields, with a comma and parentheses added (e.g., `latitude` = 40.6044676912084 and `longitude` = −73.9667785228311 combine to form the `location` field (40.6044676912084, −73.9667785228311). This representation reduces the usability of the `location` field relative to the individual coordinate fields.
- `borough` & `park_borough`: This represents 100 % matches.
- `borough` & `taxi_company_borough`: Despite similar names, these fields exhibit only a 0.29 % match indicating that these fields appear to serve different purposes for two different agencies. The `taxi_company_borough` field is exclusive to the TLC.
- `agency` & `agency_name`: The `agency` field contains the abbreviations of City agencies (e.g., NYPD, DOT, DEP), while `agency_name` contains the full names (e.g., Department of Environmental Protection). Including both fields seems redundant, especially given that NYC agency abbreviations are widely recognized and understood.
- `landmark` & `street_name`: Per the 311 SR Data Dictionary, the `landmark` field is intended to refer to noteworthy locations such as parks, hospitals, airports, sports facilities, or performance spaces. However, most entries are actually street names, with a 59.7% match to the `street_name` field. Many non-matches also appear to be logical matches differing only in spelling or formatting (e.g., "NINTH AVE" vs. "9 AVE"), suggesting misuse of the `landmark` field and highlighting redundancy.
- `cross_street_1 & 2` & `intersection_street_1 & 2`: These two street pairs are used to help identify the `incident address`. Approximately 83.6 % of each corresponding street pair are duplicates. Given this high level of duplication, it raises the question of which field-pair should be trusted and whether they serve distinct purposes. It is possible that these originated from different agencies using different variable names to represent the same information, but no documentation exists to clarify this relationship.

Reducing File Size Beyond analytical concerns, removing duplicate and near-duplicate fields can reduce file size by 18.5 %. A smaller file size means faster downloads, reduced storage, and often simpler data analysis. The following list identifies duplicate and near-duplicate fields that could be consolidated after verifying that distinct information is preserved.

- `agency_name`: redundant with `agency`.
- `park_borough`: identical to `borough`.
- `location`: a concatenation of `latitude` and `longitude` coordinates.
- `cross_street_1/2` and `intersection_street_1/2`: approximately 83.6 % agreement; reconcile nonmatching values before retaining a single pair.

Additionally, certain fields may not be useful for broad analysis due to their sparse data population. This applies to the fields below, whose use is limited or agency-specific. For example, `due_date` is 0.33 % complete and `facility_type` is 1.32 % complete.

- `taxi_company_borough`
- `taxi_pick_up_location`
- `vehicle_type`
- `facility_type`
- `due_date`
- `bridge_highway_direction`
- `bridge_highway_segment`
- `road_ramp`

It's worth noting that while such sparsity may limit their use in broad analyses, these fields can still contain valuable insights for specific agency-level studies, given that even a 1% non-blank rate can be meaningful given a 16 million-row dataset. Proper handling of sparse fields, such as storing them in separate tables or using them in targeted analyses, could improve data accessibility and efficiency without compromising the dataset's broader utility.

Field Encoding Further size reduction is possible by encoding selected categorical variables. Using `complaint_type` as an example, instead of repeatedly storing the full text of each possible value, one could use a numeric representation. For example, rather than storing verbose strings such as the `complaint_type` "`REQUEST LARGE BULKY ITEM COLLECTION`" 632,148 times, numeric or symbolic codes could be used instead. Encoding not only reduces file size but also improves processing efficiency by using numeric values instead of text strings. Moreover, standardized categorical inputs often prevent misspellings and invalid values, ensuring cleaner, more reliable data. This approach is particularly useful in large datasets, where frequently repeated text-based categorical variables significantly increase storage requirements.

Although publishing a flat CSV file supports lightweight analysis, it becomes inefficient for analytics involving millions of records. One promising alternative for reducing storage requirements is the cross-platform Apache Arrow framework. Apache Arrow (arrow.apache.org) is an innovative, cross-language development platform designed for in-memory data. It defines a standard for representing columnar data, enabling efficient real-time analytics and data processing. As such, it provides reductions in file storage size, along with efficient analysis and cross-platform support for R, Python, and Julia. Substantial storage reduction is possible (Bates and Yan, 2024).

3.7 Data Dictionary

The 311 SR Data Dictionary would benefit from an update. The analysis identified several discrepancies between the Data Dictionary and the actual data that may mislead analytic efforts. Examples are provided below.

Data Typing As previously noted, while the NYC Open Data portal displays the data types of various fields, the Data Dictionary does not. For example, the Data Dictionary specifies `incident_zip` as text. This may not be the most suitable data definition. Perhaps a better dictionary description would be "categorical" with a note indicating that the data must contain only numeric values and is not subjected to arithmetic operations.

Data Domains Additionally, the Data Dictionary often shows incomplete or out-of-date domains of legal values. For example, the `status` field is noted as having values of *assigned, canceled, closed*, and *pending*. However, the dataset also contains additional values, including *in progress, started*, and *unspecified*. Similarly, the `address_type` field shows acceptable values of *address, blockface, intersection, latlong*, and *placename*. However, the dataset also contains additional values, including *bbl* and *unrecognized*. Other fields, such as `facility_type`,

`vehicle_type`, `taxi_pick_up_location`, `road_ramp`, and `city`, exhibit similar inaccuracies in their specified domain values.

4 Recommendations for Government Open Data Curation

Building on these insights, we propose a set of data curation principles tailored for government open data. These principles address the unique requirements observed in the curation of such data, ensuring reliability, consistency, and utility for public services, research, and related applications. To facilitate implementation, these principles are prioritized based on their impact and feasibility, offering a roadmap for improving data curation practices.

4.1 Principles

Principle 1: Maintain Cross-Organizational Consistency A key challenge in government-released datasets is harmonizing data fields across agencies and maintaining consistency over time. Inconsistent field definitions, data formats, or value domains can lead to erroneous analyses and hinder cross-agency collaboration. Agencies should standardize naming conventions, value domains, usage, and formats across all datasets. Implement regular audits and data harmonization processes to ensure long-term consistency across datasets, particularly as agency structures and reporting standards evolve. For instance, applying the same field naming conventions to historical datasets prevents inaccurate analyses across years.

Principle 2: Ensure Data Entry Accuracy Data accuracy and validity are fundamental to the utility of open datasets. In the NYC 311 data, issues such as invalid ZIP codes and nonsensical dates highlight the need for rigorous validation processes. Governments should establish protocols for data entry, ensuring that fields are populated with valid, legally acceptable values. To the maximum extent possible, the software should ensure that entries are accurate and valid, for example:

- Having software populate certain date fields, triggered by a user action such as changing status.
- Subjecting select fields to automated checks that prevent logical errors, such as a closing date before the creation date.
- Validating each field that has a defined set of permissible values against a reference dataset at point of entry.

Principle 3: Optimize Storage Efficiency Efficient storage matters for large datasets. Removing duplicate and near-duplicate fields identified in Section 3.6 would reduce the dataset size by 18.5%, provided that distinct information in nonmatching records is preserved. Encoding categorical variables in a standardized format could further reduce file sizes.

Principle 4: Maintain Up-to-Date Data Dictionaries Clear, accurate, and up-to-date documentation is essential for effective use of government datasets. Our study found several instances where the current Data Dictionary did not accurately reflect the underlying data. Document allowable values explicitly. Conduct regular reviews, especially when there are changes in data structure, usage, or context.

Principle 5: Automate Quality Assurance Automated data validation and quality assurance processes can significantly improve the reliability of government datasets. Such automation can identify trends and flag irregularities. In the NYC 311 case, large spikes in SR creation and closure exactly at midnight and noon suggest systematic or automated timestamp-generation processes that warrant investigation. Real-time validation systems could help detect and prevent such errors, ensuring data integrity at the point of entry and maintaining data quality.

Principle 6: Ensure Data Transparency Transparency is a core tenet of open data, and governments must ensure that datasets are easily accessible and readily understandable. This includes publishing clear metadata, enabling users to understand the dataset's structure, limitations, and intended use. Most important, agencies should be open about data quality issues, for example by notifying the public when errors in SR fields (e.g., erroneous closed dates) are discovered and explaining how they will be resolved.

4.2 Actionable Steps

These principles provide a conceptual foundation. However, agencies need concrete steps to put them into practice. Clear priorities help agencies focus efforts where they will have the greatest impact. The following actions represent a structured path toward improving data curation, validation, and long-term dataset stewardship.

Create Real-Time Validation Tools Begin by investing in real-time validation software, regularly updating data dictionaries, and building tools for encoding and standardizing data. Automate rule-based checks, such as those formalized in statistical data-validation frameworks (Van der Loo and De Jonge, 2018), to significantly improve data quality at the point of data entry. Use automated systems to proactively detect and correct common issues, reducing the downstream manual workload.

Improve Collaboration Among Stakeholders Effective data curation relies on collaboration among city agencies and engagement with the broader open-data community. Regular discussions regarding data standards, formats, and metadata ensure that NYC datasets support the needs of policymakers, researchers, and the public. Clear communication surfaces recurring issues, clarifies user requirements, and guides improvements to dataset documentation and structure.

Establish Clear Responsibility for Data Quality Collaboration requires clear ownership. NYC already has a governance structure for open data: the Technical Standards Manual implementing Local Law 11 of 2012 (New York City Council, 2012; City of New York, 2026). This law requires each agency to designate an Open Data Coordinator (ODC), who works with a citywide Chief Open Platform Officer in OTI. The ODC's responsibilities center on publication and compliance: delivering datasets to the portal, contributing to the annual compliance report, liaising with the Open Data team, responding to public feedback, and conducting outreach. None of these roles, however, explicitly assigns data quality or curation responsibilities.

We recommend that agencies formalize responsibility for data curation and quality, either by extending the ODC role or by appointing a dedicated Data Curation Officer. Give this role defined duties, training, and dedicated time. A named curator in each agency would give the

collaboration described above a consistent point of contact and make data quality an explicit, accountable responsibility rather than an incidental one.

Develop Automated Quality Assurance Systems Automated QA systems could monitor adherence to established validation standards, providing a real-time snapshot of data quality and highlighting deviations for corrective actions.

Ensure Expert Engagement Engagement with data scientists and statisticians is key to improving data quality. Experts can help design curation processes, such as developing algorithms for automatic data cleansing, identifying inconsistent/invalid data values, and ensuring curation rules are maintained over time. Their involvement also helps to establish best practices for validating data and maintaining consistency across data sources.

Maintain Robust Feedback Channels Ongoing feedback from dataset users—including researchers, civic groups, and the public—is essential for identifying issues and refining data practices. Establishing feedback loops for public engagement ensures that recommendations for improving data curation and validation are not only heard and discussed, but also implemented.

Encourage Public Engagement Through Community Initiatives Events such as NYC Open Data Week provide a valuable platform for collaboration, learning, and innovation. This project originated from such an event. Governments' support of such initiatives not only promotes data literacy but also encourages community-driven improvements, strengthens public engagement, and demonstrates the societal value of open data.

5 Conclusion

Our study highlights the critical role of robust data curation in ensuring the reliability and utility of open datasets, as demonstrated through an examination of NYC 311 SR data. When researchers apply these datasets in fields such as public services and academic research, inconsistencies, invalid values, and data errors undermine their findings. Quality concerns are especially consequential when the data serve as inputs to automated analytical systems. Trustworthy machine learning systems, which rely heavily on high-quality data, are particularly vulnerable to such deficiencies. Errors in the underlying data, including biases or inconsistencies, can compromise both the accuracy and fairness of predictive models, thereby affecting real-world decisions in urban planning and policy (Rahm and Do, 2000; Geiger et al., 2020).

While the findings here are specific to NYC 311 data, the proposed principles have broader applicability. Datasets in healthcare, transportation, and environmental monitoring can similarly benefit from these curation strategies, enabling improved outcomes across domains. This research further emphasizes the necessity of harmonizing data across city agencies, particularly resolving discrepancies in field names, formats, usage, and definitions. Consistency in data is also essential for trustworthy machine learning, where inconsistencies across sources can lead to distorted outcomes. This issue is even more pronounced in datasets spanning long time periods, as changes in agency structures, reporting standards, or data usage can introduce subtle biases and distortions. Long-term data consistency is essential for longitudinal studies and predictive modeling, as even minor data shifts can lead to significant deviations in model outcomes, as noted by Rahm and Do (2000) and Borgman (2012).

High-quality data enables machine learning models to better predict trends, allocate resources, and address service delivery issues. The COVID-19 pandemic underscored the importance of real-time, trustworthy data in crisis management, where the accuracy and timeliness of data-driven insights were critical for public health responses. Without proper data curation, machine learning models used during the pandemic would have been compromised, affecting decisions on resource allocation and service provision (Worby and Chang, 2020; Khemasuwan and Colt, 2021). Future efforts could focus on automating data curation processes, particularly in real-time data pipelines, to ensure that the data used remains accurate, clean, and reliable (Chu et al., 2016; Hurbean et al., 2021).

Supplementary Material

The Supplementary Materials contain extended results, supporting visualizations, and instructions for reproducing the analysis.

- R source code: https://github.com/tusseyd/nyc_311_data_cleaning
- NYC 311 SR dataset (Figshare): https://doi.org/10.6084/m9.figshare.30699800
- ZIP code database: https://www.unitedstateszipcodes.org/zip-code-database/

The supplementary material for this article is available on the official website of the Journal of Data Science at https://jds-online.org/. It may include additional results, appendices, tables, figures, code, and data.

AI Use Disclosure

The authors used Claude (Anthropic; model Claude Opus 5.5) and ChatGPT (OpenAI; model GPT-5.6 Sol), accessed September 2026, for coding assistance with R programming, validation of reported figures against program output, and LaTeX editing of the manuscript, supplementary materials, and response to the referee. All AI-assisted material was reviewed and validated by the authors, who take full responsibility for the manuscript's content, accuracy, originality, and integrity.

References

Agonafir C, Pabon AR, Lakhankar T, Khanbilvardi R, Devineni N (2022). Understanding New York City street flooding through 311 complaints. *Journal of Hydrology*, 605: 127300. https://doi.org/10.1016/j.jhydrol.2021.127300

Barns S (2016). Mine your data: Open data, digital strategies and entrepreneurial governance by code. *Urban Geography*, 37(4): 554–571. https://doi.org/10.1080/02723638.2016.1139876

Barrett T, Dowle M, Srinivasan A, Gorecki J, Chirico M, ..., Krylov I (2025). *data.table: Extension of 'data.frame'*. R package.

Bates DM, Yan J (2024). From CSV to Arrow: Creating a unified data set for efficient cross-platform analysis. *Chance*, 37(4): 48–51. https://doi.org/10.1080/09332480.2024.2434443

Borgman CL (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63(6): 1059–1078. https://doi.org/10.1002/asi.22634

Cantor MN, Chandras R, Pulgarin C (2018). FACETS: Using open data to measure community social determinants of health. *Journal of the American Medical Informatics Association*, 25(4): 419–422. https://doi.org/10.1093/jamia/ocx117

Chu X, Ilyas IF, Krishnan S, Wang J (2016). Data cleaning: Overview and emerging challenges. In: *Proceedings of the 2016 International Conference on Management of Data* (Özcan F, Koutrika G, Madden S, eds.), 2201–2206. ACM, San Francisco, CA.

City of New York (2026). NYC Open Data Technical Standards Manual: City policies. https://cityofnewyork.github.io/opendatatsm/citypolicies.html. Accessed: 2026-09-24.

Cody R (2017). *Cody's Data Cleaning Techniques Using SAS*. SAS Institute.

Dove G, Mydlarz C, Bello JP, Nov O (2022). Sounds of New York City. *Interactions*, 29(3): 32–35.

European Union Publications Office (2024). European data. https://data.europa.eu/euodp. Accessed: 2024-11-22.

Geiger RS, Yu K, Yang Y, Dai M, Qiu J, ..., Huang J (2020). Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Hildebrandt M, Castillo C, Celis LE, Ruggieri S, Taylor L, Zanfir-Fortuna G, eds.), 325–336. ACM, Barcelona, Spain.

General Services Administration (2024). The home of the U.S. government's open data. https://data.gov/. Accessed: 2024-11-21.

Gerte R, Konduri KC, Ravishanker N, Mondal A, Eluru N (2019). Understanding the relationships between demand for shared ride modes: Case study using open data from New York City. *Transportation Research Record*, 2673(12): 30–39. https://doi.org/10.1177/0361198119849584

Grant EL, Leavenworth RS (1996). *Statistical Quality Control*. McGraw-Hill, New York, 6th edition.

Hart EM, Barmby P, LeBauer D, Michonneau F, Mount S, ..., Hollister JW (2016). Ten simple rules for digital data storage. *PLoS Computational Biology*, 12(10): e1005097. https://doi.org/10.1371/journal.pcbi.1005097

Hosseinzadeh M, Azhir E, Ahmed OH, Ghafour MY, Ahmed SH, ..., Vo B (2023). Data cleansing mechanisms and approaches for big data analytics: A systematic study. *Journal of Ambient Intelligence and Humanized Computing*, 14: 99–111. https://doi.org/10.1007/s12652-021-03590-2

Hurbean L, Danaiata D, Militaru F, Dodea AM, Negovan AM (2021). Open data based machine learning applications in smart cities: A systematic literature review. *Electronics*, 10(23): 2997. https://doi.org/10.3390/electronics10232997

Jain A, Patel H, Nagalapatti L, Gupta N, Mehta S, ..., Munigala V (2020). Overview and importance of data quality for machine learning tasks. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Gupta R, Liu Y, Tang J, Prakash BA, eds.), 3561–3562. ACM, Virtual Event, CA, USA.

Janssen M, Charalabidis Y, Zuiderwijk A (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, 29(4): 258–268. https://doi.org/10.1080/10580530.2012.716740

Khemasuwan D, Colt HG (2021). Applications and challenges of AI-based algorithms in the COVID-19 pandemic. *BMJ Innovations*, 7: 387–398. https://doi.org/10.1136/bmjinnov-2020-000648

Kontokosta CE, Hong B (2021). Bias in smart city governance: How socio-spatial disparities in 311 complaint behavior impact the fairness of data-driven decisions. *Sustainable Cities and Society*, 64: 102503. https://doi.org/10.1016/j.scs.2020.102503

Maletic JI, Marcus A (2005). Data cleansing. In: *Data Mining and Knowledge Discovery Hand-*

book (O Maimon, L Rokach, eds.), 21–36. Springer.

Minkoff SL (2016). NYC 311: A tract-level analysis of citizen–government contacting in New York City. *Urban Affairs Review*, 52(2): 211–246. https://doi.org/10.1177/1078087415577796

Neves FT, de Castro Neto M, Aparicio M (2020). The impacts of open data initiatives on smart cities: A framework for evaluation and monitoring. *Cities*, 106: 102860.

New York City Council (2012). Local Law 11 of 2012 (NYC Open Data Law). https://cityofnewyork.github.io/opendatatsm/LocalLaw11of2012.html. Accessed: 2026-09-24.

NYC Office of Technology and Innovation (2025). NYC 311 service requests, 2020–2024: Export as of 23 September 2025. Figshare archive, version 2, https://doi.org/10.6084/m9.figshare.30699800.

NYC Open Data (2025). 311 Service Requests updates. https://www.nyc.gov/opendata/news/all-news/311-Service-Requests-Updates. Describes the December 2025 changes to the 311 Service Requests dataset. Accessed: 2026-09-27.

NYC Open Data Team (2024). NYC Open Data. https://opendata.cityofnewyork.us/. Accessed: 2024-11-22.

O'Brien DT, Offenhuber D, Baldwin-Philippi J, Sands M, Gordon E (2017). Uncharted territoriality in coproduction: The motivations for 311 reporting. *Journal of Public Administration Research and Theory*, 27(2): 320–335.

Polyzotis N, Zinkevich M, Roy S, Breck E, Whang S (2019). Data validation for machine learning. In: *Proceedings of Machine Learning and Systems* (A Talwalkar, V Smith, M Zaharia, eds.), 334–347.

R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rahm E, Do HH (2000). Data cleaning: Problems and current approaches. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 23(4): 3–13.

Raj R, Ramesh A, Seetharam A, DeFazio D (2021). SWIFT: A non-emergency response prediction system using sparse Gaussian conditional random fields. *Pervasive and Mobile Computing*, 71: 101317. https://doi.org/10.1016/j.pmcj.2020.101317

Shankar K, Jeng W, Thomer A, Weber N, Yoon A (2021). Data curation as collective action during COVID-19. *The Journal of the Association for Information Science and Technology*, 72(3): 280–284. https://doi.org/10.1002/asi.24406

UnitedStatesZipCodes.org (2025). Free ZIP code database. https://www.unitedstateszipcodes.org/zip-code-database/. Free version, compiled from USPS and other sources. The version used here has 42,735 records (1,040 of them decommissioned ZIP codes).

Van der Loo M, De Jonge E (2018). *Statistical Data Cleaning with Applications in R*. John Wiley & Sons.

Wang HJ, Lo J (2016). Adoption of open government data among government agencies. *Government Information Quarterly*, 33(1): 80–88. https://doi.org/10.1016/j.giq.2015.11.004

White A, Trump KS (2018). The promises and pitfalls of 311 data. *Urban Affairs Review*, 54(4): 794–823. https://doi.org/10.1177/1078087416673202

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, . . . , Mons B (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3: 160018. https://doi.org/10.1038/sdata.2016.18

Witt M, Carlson J, Brandt DS, Cragin MH (2009). Constructing data curation profiles. *International Journal of Digital Curation*, 4(3): 93–103. https://doi.org/10.2218/ijdc.v4i3.117

Worby CJ, Chang HH (2020). Face mask use in the general population and optimal re-

source allocation during the COVID-19 pandemic. *Nature Communications*, 11(1): 4049. https://doi.org/10.1038/s41467-020-17922-x

World Bank (2024). World Bank Open Data. https://data.worldbank.org/. Accessed: 2024-11-22.

Zha Y, Veloso M (2014). Profiling and prediction of non-emergency calls in New York City. In: *Proceedings of the Workshop on Semantic Cities: Beyond Open Data to Models, Standards and Reasoning* (Fox M, Lécué F, McIlraith S, Srivastava B, Uceda-Sosa R, eds.), 41–47. AAAI Press, Technical Report WS-14-11.

Zuiderwijk A, Janssen M (2014). Open data policies, their implementation and impact: A framework for comparison. *Government Information Quarterly*, 31(1): 17–29. https://doi.org/10.1016/j.giq.2013.04.003