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]Former Executive Director, NYC DoITT
[2]Department of Statistics, University of Connecticut

June 9, 2024

Contents

List of Figures

List of Tables

Abstract

In the early 21st century, the open data movement began to transform societies and governments through principles of transparency, innovation, and public engagement. New York City (NYC) has emerged as a leader in this movement with the enactment of the Open Data Law in 2012, leading to the creation of the NYC Open Data portal, which now hosts 2700 datasets from 80 city agencies. This resource has proven invaluable for research across various domains, including health, urban development, and transportation. The success of these initiatives underscores the importance of data curation, ensuring the utility and reliability of datasets.

This paper examines the data curation challenges using the NYC 311 Service Request (SR) Data as a case study, addressing issues of data validity, consistency, and curation efficiency. Based on insights from this case study, we propose a set of data curation principles tailored for government-released open data. These principles aim to enhance data management practices and ensure the ongoing utility of open data. The paper concludes with actionable suggestions for improving data curation and offers general principles for the release of open data.

Keywords: Data science; Open data; Data cleansing; Quality control; Open Data Movement; Data Democratization; Transparency; Data Curation; Government Data; NYC Open Data; 311 Service Request Data; Data Validity; Data Consistency; Data Efficiency; Smart Cities; Public Engagement; Research Data Management;

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, a pioneering platform in making government data widely accessible. Shortly after, the European Union followed suit, unveiling its Open Data Portal in 2012, further cementing the movement's global reach. Furthermore, the World Bank's Open Data initiative, initiated in 2010, stands out as a comprehensive repository for global development data, available at World Bank Open Data.

These initiatives represent significant strides in democratizing data, in breaking barriers that once kept valuable information on government performance in silos. Their collective impact is profound, extending beyond mere data sharing to fostering a culture of openness that benefits individuals, communities, governments, and economies worldwide (Barns, 2016; Wang and Lo, 2016).

New York City (NYC) has emerged as a forerunner in the open data movement, marked by the enactment of the Open Data Law in 2012 (Zuiderwijk and Janssen, 2014). This landmark legislation led to the creation of the NYC Open Data portal, which today hosts an impressive array of 2700 datasets from 80 different city agencies. This resource has become invaluable for researchers in various fields as well an enabling local government transparency

In health, datasets have enabled significant studies on facets of health and healthcare delivery (Cantor et al., 2018; Shankar et al., 2021). In the realm of urban development, data has been instrumental in advancing smart city initiatives (Neves et al., 2020). Additionally, transportation research has benefited greatly from this wealth of data, aiding in the understanding of urban mobility and infrastructure (Gerte et al., 2019). NYC's Open Data initiative not only exemplifies commitment to transparency and public engagement but also illustrates how open data can be a powerful tool in addressing complex urban challenges. Examples of some of the more popular NYC datasets include:

- NYC restaurant violations
- Popular baby names
- Mapping of car crashes involving pedestrians
- Mapping of sidewalk widths in NYC
- Visualization of NYC High School and College enrollment
- An interactive dashboard to filter and view charges for jail inmates
- Location of City-wide free Internet access points
- ...and 220 other complaint types.

Data curation is fundamental in the open data ecosystem, ensuring the utility and reliability of datasets for diverse applications. Among the earliest discussions, Witt et al. (2009) focus on the development of data curation profiles tailored to specific contexts, setting a precedent for targeted data management strategies. Addressing broader challenges in data sharing and management, Borgman (2012) highlights the complexities of research data distribution, emphasizing the need for robust strategies. This is complemented by the work of Hart et al. (2016), who outline essential principles for effective data management, particularly emphasizing the importance of meticulous curation practices.

In the realm of collaborative data management, Beheshti et al. (2019) underscore the significance of cooperative environments for managing and sharing social data effectively. This aspect of data curation gains further relevance in the research by McLure et al. (2014), which delves into the specific practices and needs within data curation communities. The practical implications of data curation are vividly illustrated in the context of public health and global challenges. Cantor et al. (2018) demonstrate the utility of curated open data in evaluating community health determinants. Furthermore, the COVID-19 pandemic serves as a real-world example, with Shankar et al. (2021) observing the critical role of collective data curation efforts in managing and responding to the crisis. Collectively, these studies not only highlight the multifaceted nature of data curation but also emphasize its indispensable role in enhancing the applicability and value of open data across various domains.

The contributions of this paper are twofold. First, we delve into the specifics of data curation challenges using the NYC 311 Service Request (SR) Data as a case study. This renowned and frequently viewed dataset serves as a prime example for examining key issues in data curation, including data validity, consistency, and curation efficiency. We illustrate these points with live examples drawn from our processing of the 311 SR data.

Secondly, building upon insights gained from this case study, we propose a set of data curation principles tailored for government-released open data. These principles are designed to address the unique challenges and requirements observed in the curation of such datasets.

The paper is organized as follows: Section 2 offers a brief review of the history of the 311 system. In Section 3 we take a look at long-term trends presented via a 10-year analysis. Section 4 offers a general discussion of data cleansing issues impacting data quality and curation efficiency. Section **??** examines the technical dimensions of the dataset, the structural issues. In Section 6 we examine the data fields for compliance with the stated data type as found in the Data Dictionary. Section 7 looks at the data fields as regards missing, blank, or N/A entries. Section 8 explores how data fields comply with a domain of legal or acceptable values. Section **??** deals with the important issues surrounding logical inconsistencies and concerning patterns in the data. Section 9 explores the age-old issue of precision versus accuracy. Section 10 identifies duplicate and redundant data fields. Section **??** provides actionable suggestions for mitigating or resolving identified issues. Following this, Section 13 outlines a series of general principles for the release of open data, drawing from our findings. The paper concludes with a discussion in Section 14, encapsulating the key insights and implications of our research.

A note on naming conventions: dataset field names are typically named using "snake_case" where each word in the variable name is written in lowercase, and words are separated by underscores, e.g created_date, complaint_type, zip_code, etc..

2 NYC 311 System History

The NYC 311 service, a critical component of New York City's public engagement and service response framework, serves as a centralized hub for non-emergency inquiries and requests. Introduced in 2003, the NYC 311 system was designed to streamline the city's response to non-emergency issues, ranging from noise complaints to street maintenance requests. The evolution of this system can be traced from its initial implementation as a simple inquiry channel via telephone to a comprehensive data management system that handles millions of requests annually. Each inquiry is logged as a complaint in the 311 system. Key milestones in the 311 system development include:

- 2003 - NYC 311 system goes live with a phone-based call center only. The "311" phone number replaces a myriad of numbers to call for each individual City Agency.
- 2009 - 311 online & mobile apps launched. Today 60% of all Service Requests come from mobile & online channels.
- 2019 - Major software upgrade of 311 system with enhanced capabilities, dynamic load-balancing, and cloud-based resilience. Only major software upgrade since 2004.
- 2020 - In August 2020, driven by the COVID pandemic, monthly 311 Service Requests hit an all-time high of 348,463. Seven of the Top 10 busiest days of all-time occur in 2020.
- 2021 - 311 System is expanded to into the MTA's city subway system, the largest

expansion since its inception.

- 2023 - Record high year for 311 with 3.23 million Service Requests.

Today, the NYC 311 data system is a robust platform that manages a vast array of urban living-related inquiries. The system handles over 3 million service requests (SRs) per year, encompassing such complaints as street noise, illegal parking, heat/hot water, abandoned vehicles, and unsanitary conditions. The data, managed through a tailored application built on the Microsoft Dynamics and Azure platform, is a valuable resource for city administration and policy-making. The 311 data is publicly accessible through the NYC Open Data Portal which provides access to all the data sets (approx. 2700) and provides a convenient method for querying, grouping, aggregating, geo-mapping, visualization, and exporting results. This open data initiative enables not only governmental transparency but also empowers researchers, civic developers, and the general public.

Additional Open Data projects can be found at NYC Open Data Project Gallery and at βetaNYC products and tools.

Despite its success, the system constantly faces challenges such as:

- Data timeliness, accuracy, and consistency
- Difficulties correlating data over long time periods, e.g. over a 10-year period, e.g. Agency name changes, new complaint types.
- Data anonymization and handling of Personally identifiable information (PII)
- Integration with stand-alone systems at selected NYC Agencies
- Managing API usage, authentication, and load for 3rd party users, of which there are many
- Incorporating ongoing technology changes and upgrades
- Understanding the role of chatbots and AI to provide online customer service, along

with other emergent technologies

These and other challenges are continually addressed by the various Agency open data managers as well as the NYC Office of Technology and Innovation (OTI) which provides the technology support for the open data system. Many of the underlying datasets are updated daily, usually running 1-2 days behind from creation. However, a number of datasets are updated monthly (such as restaurant inspections), and some are not updated for one or more years (such as community board boundaries). This poses problems when merging dataset with others, as would be expected. Typically, the data must be exported and manipulated in another, external program. The authors found this to be the case with the large 311 dataset employed.

Data accuracy and consistency is typically established between OTI and the other contributing City Agencies, such as NYPD, Housing Preservation and Development (HPD), Parks & Recreation, etc., via Service Level Agreements (SLAs). However, enforcement of those SLAs is often challenging. The result is that often data accuracy and logical consistency is occasionally compromised. However, the Open Data Team is quite responsive to inquires from users. The authors have provided many such suggestions during the course of this research effort, and the most current data shows improvements in many of the areas highlighted. Questions and suggestion can be provide here: Contact Us.

The NYC 311 organization, part of OTI, operates a highly tailored and sophisticated case management. However, one constraint was particularly challenging during the recent complete overhaul of the software and hardware in 2019. When the original 311system was developed in 2003, a decision was made to allow select NYC Agencies to maintain their own local computer systems, such as the mainframe systems at the NYC Department of Sanitation (DSNY) and others. And when the 311 system was completely upgraded in 2019, those same integrations were honored in order to reduce the scope of the project. Perhaps in retrospect, while a good budgetary and policy decision, it was perhaps not such a good technology decision.

At least nine NYC Agencies have local systems that are integrated directly to the 311 system either via an API or custom code. These include:

- Department of Social Services (DSS), one of the largest City Agencies
- Department of Sanitation, although recently part of DSNY has moved to the core 311 software.
- Department of Parks & Recreation (forestry complaints only)
- Housing, Preservation & Development (fully integrated)
- Department of Environmental Protection (DEP) (fully integrated)
- Department of Healthy & Mental Hygiene (fully integrated)
- Department of Consumer & Worker Protection (DCWP - formerly known as the Dept of Consumer Affairs)
- New York Police Department (NYPD) (integrated with their reporting system)
- Department of Business (DoB) (fully integrated)

One unfortunate side-effect from these integrations, and other Agencies that use the 311 APIs, is that some of the data consistency checks which are inherent in the 311 system are not applied uniformly, and in some cases not at all, to the integrated data streams. Errors often creep in from other Agency systems and are not easily corrected or detected in the 311 application itself, but rather require correction at the source Agency; challenging.

One area that will almost certainly remain an ongoing issue is the ongoing incorporation of new software, hardware, and cloud services which are constantly changing while new approaches emerge. The 311 telephonic integration recently was reconfigured due to a new telephonic vendor. Desktop software at the 311 Agency are subject to the same upgrades that home and business users experience. And while the 311 system was designed to be easily configured and updated with new features, some emerging business issues as well as technologies nonetheless require additional software engineering efforts to accommodate, along with the concomitant budgets.

One area in particular which is heavily influencing customer outreach for the 311 system is the appropriate use of online chatbots and assistants. Increasing these features are influenced by artificial intelligence (AI) advances. It is a given that a number of citizen's interactions with City government could be handled accurately and quickly by AI-driven applications thus improving timeliness and accuracy. But how to do that while preserving customer privacy and anonymization and user engagement are challenging issues.

The impact of the NYC 311 data extends beyond operational efficiency; it has become instrumental in shaping City governance and community engagement. The data has been pivotal in such areas as:

- Providing advice on shelters during hurricanes, heat emergencies, and winter blizzards
- Handling countless inquires during the COVID pandemic (testing centers, vaccine sites, mobile vaccine clinics, etc.)
- Enforcement of standards between landlords and tenants
- Re-allocation of NYC's yellow taxi routes based upon an insightful analysis conducted by the Taxi & Limousine Commission (TLC)
- Improved responsiveness for City Agencies with regards to direct client-facing responsibilities such as Parks & Rec, Dept of Transportation, DSNY, etc.
- Increased 311 SRs related to Homelessness resulted in the launch of the City's Homeless Outreach & Mobile Engagement Street Action Teams (HOME-STAT) which includes street canvassers and increases in homeless housing facilities.

- A significant increase in graffiti complains in 2017 led directly to the 'Graffiti-Free NYC" campaigns and increased funding for removal of graffiti in public spaces.

- Enhanced geospatial data regarding the location of incidents, the responding Agency, and actions taken across a myriad of complaint_types

- Establishment of Agency SLAs in regard to certain complaint types, such as residential noise, homelessness person assistance, street light outages, rat sightings, and other neighborhood quality of life issues.

The NYC 311 service exemplifies the dynamic nature of urban data management and its critical role in modern governance. In addition it provides an invaluable resource for data scientists to practice data science activities, including data curation.

We extracted 311 SR data for the period of CY 2022-2023, a dataset consisting of 6.4 million records. That dataset serves as the primary source for the analysis effort. This is a large dataset which enables observation of some rarely seen events, data inaccuracies, and logical inconsistencies. It provides enough data to examine trends and data outlier analysis. We queried the data directly from the NYC Open Data Portal and then exported the data in CSV format for analysis using custom written R programs.

Additionally, we extracted two longer-term datasets, a 5-year look (2019-2023) and a 10-year look (2014-2023) ; some 14 and 27 million rows respectively. These larger datasets afford an opportunity for long-term trend observations, confirming and supporting the analysis using the 2-year dataset. The 2022-2023 dataset was used exclusively for the detailed, field-by-field level analysis which comprises the majority of this effort.

3 10-year Analysis & Trends

Since the NYC 311 System went live in 2003, the usage has grown at an annual rate of 4.6%. As part of this study, we also looked at a 10-year period (2014-2023) to observe system performance over a longer time-frame . During that time-frame, the number of 311 SRs grew 50%.

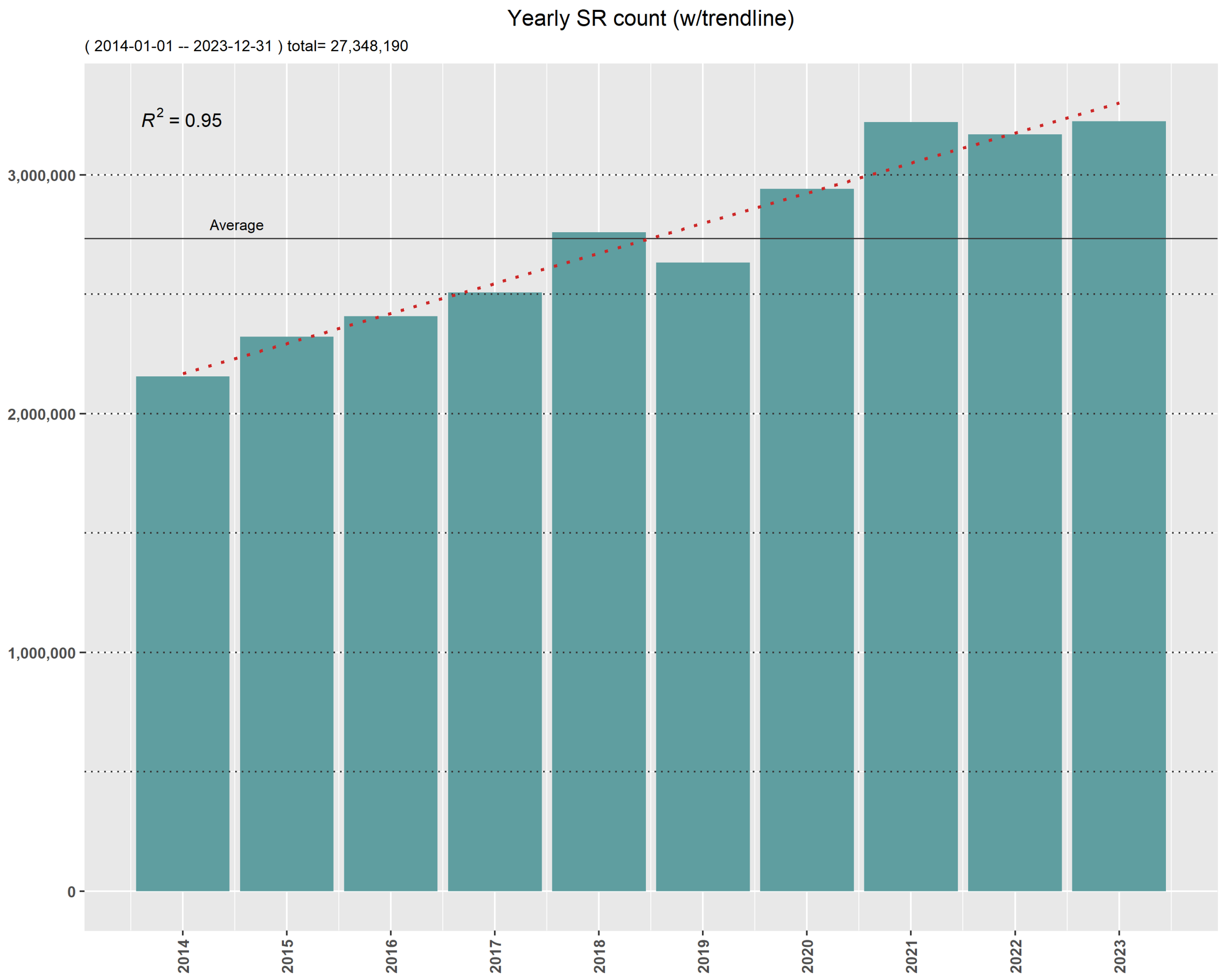


Figure 1: Yearly SR counts over 10-year period

This 10-year timeframe (2014-2023) includes the COVID pandemic. Note that the pandemic impact reached a peak in the July-September 2020 timeframe, accompanied by additional spikes in the 2ndhalf of 2021 as well. The all-time daily high for SRs occurred on Tuesday, 2020-08-04, with a spike of 24,415 SRs, a full 9.8 σ's from the mean.

This chart shows the 2-year period which is the basis for this analysis, by calendar month. Note that almost all the months before the COVID pandemic outbreak in Feb/Mar of 2020 are "below average", while most of the months after that period are "above average". That

343 level of higher usage of the 311 system has continued into the early months of 2024.

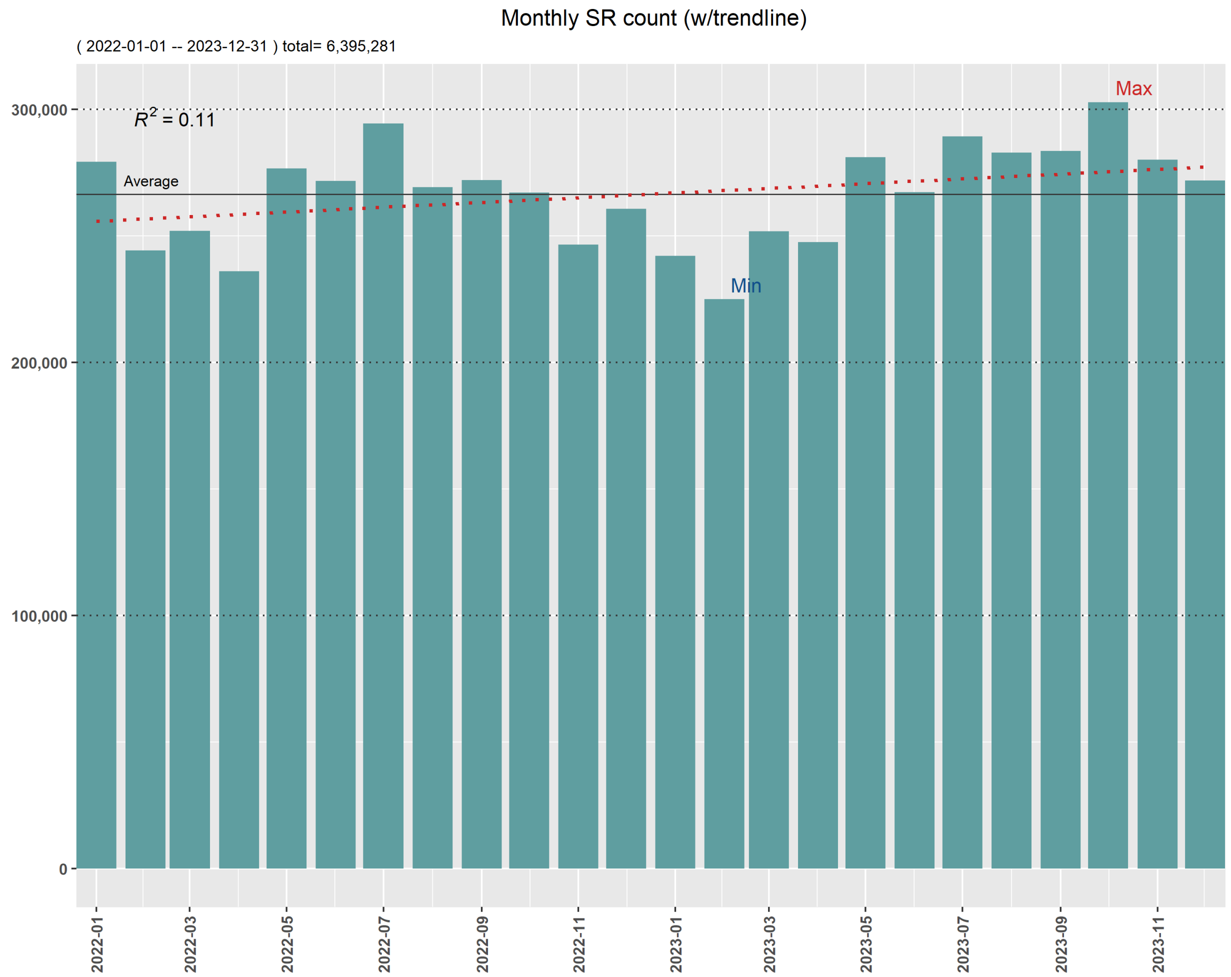


Figure 2: Monthly SR counts over 10-year period

344 A quick look at the corresponding daily tend. An average day is 8761 SRs.

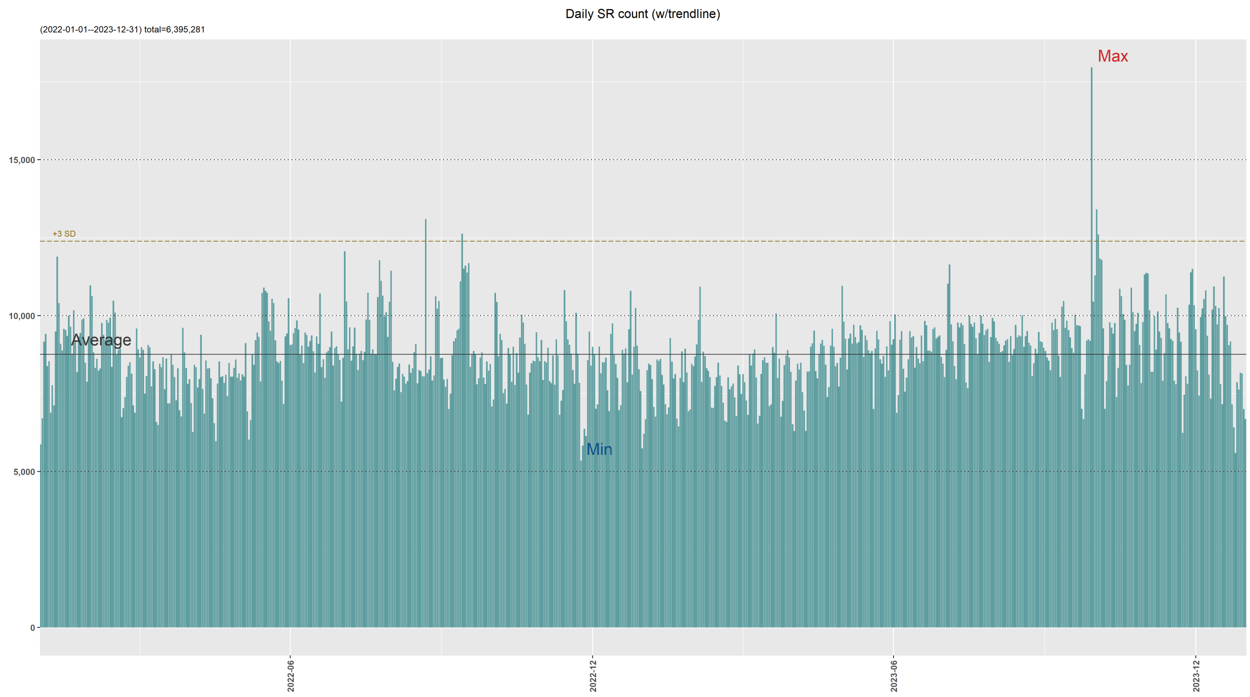


Figure 3: Daily SR counts over 2022-2023 period

Seasonal trends are also prevalent. The months of November-thru-April are “below average” months, while the months May-thru-October are “above average”. This seasonal trend is observed throughout the 10-year period as visualized here.

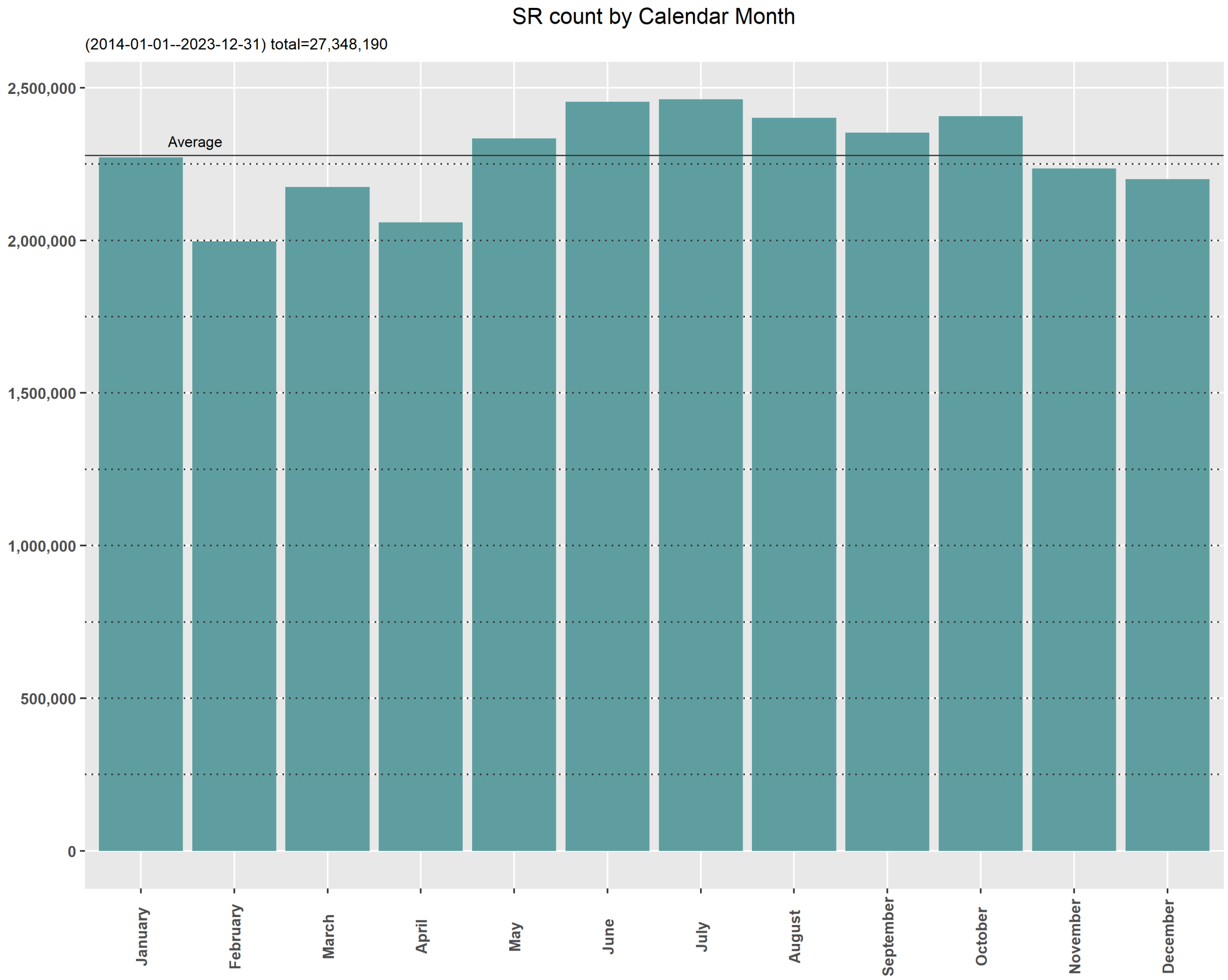


Figure 4: Calendar Month SR counts over a 10-year period

It is not surprising that the lowest SR counts-per-day occur during the Christmas holiday period. The summer months and early fall months make up all of the Top 5 SR counts by day. (As a note, the all-time daily high for SRs occurred on Tuesday, 2020-08-04 while the 10-year low count occurs on Christmas Day in 2016, a Sunday. Note also that the difference between the peak day (24,415) and the minimum day (2965) varies by a factor of 8X; all of which creates a challenging system capacity and performance issue.

Top 5 Days			Bottom 5 Days		
Date	Count	Rank	Date	Count	Rank
2020-08-04	24,415	1	2014-12-25	2965	1
2020-08-05	19,560	2	2015-12-25	3247	2
2023-09-29	17,962	3	2014-04-20	3287	3
2020-07-05	16,916	4	2015-12-26	3405	4
2020-06-21	15,883	5	2016-11-06	3419	5

Table 1: Top 5 and Bottom 5 Days

Two important fields in this analysis effort are the responsible NYC Agency, and the type of complaint. Given the size of the City of New York government, it is necessary to identify the responsible Agency as well as the type of complaint in order to identify the responsible party and assist in troubleshooting discrepancies.

When we discovered a data error, we typically observe one of two trends; either a single (or a few) Agencies are responsible (such as Dept of Transportation), or it is a system-wide issue that follows the general distribution of SRs across the system. Thus we can often identify if it is an Agency specific problem, or a systemic issue. (Note: a few key NYC Agencies are not represented in the 311 dataset. These include the New York City Housing Authority (NYCHA) and the Department of Corrections (DOC).)

Although the 311 SRs involve 16 different agencies, there are the “big six” Agencies that comprise 90% of the complaints. These six are (in order of % of complaints):

- New York Police Department (NYPD)
- Housing Preservation & Development (HPD
- New York City Department of Sanitation (DSNY)
- Department of Transportation (DOT)

- Department of Environmental Protection (DEP)
- Department of Parks & Recreation (DPR)

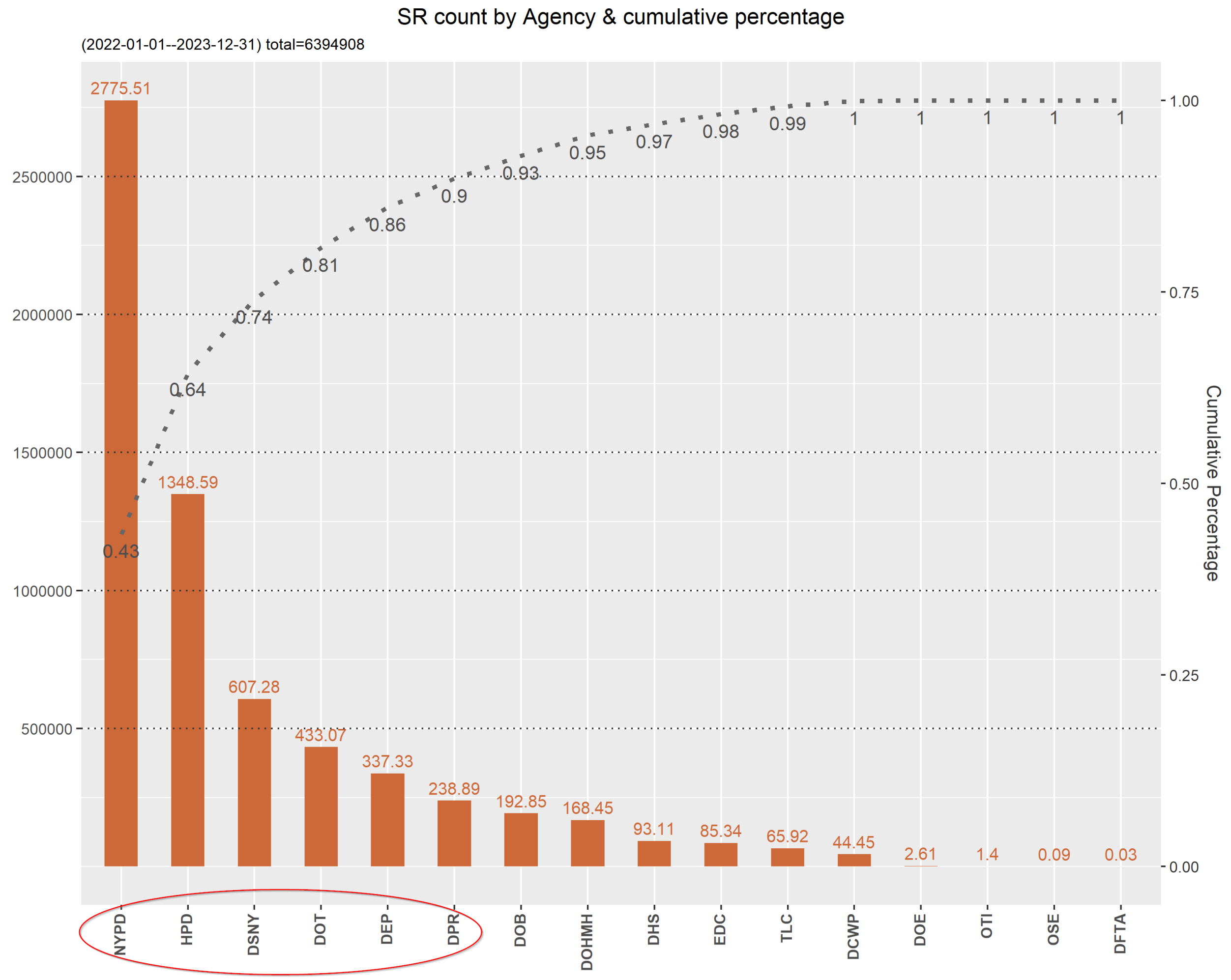


Figure 5: SR counts by Agency and Cumulative Percentage

The distribution of complaint_type skews to a few select complaints accompanied by a long tail. This graph shows the top 20 complaint_type(s) comprising 70% out of a total of 220.

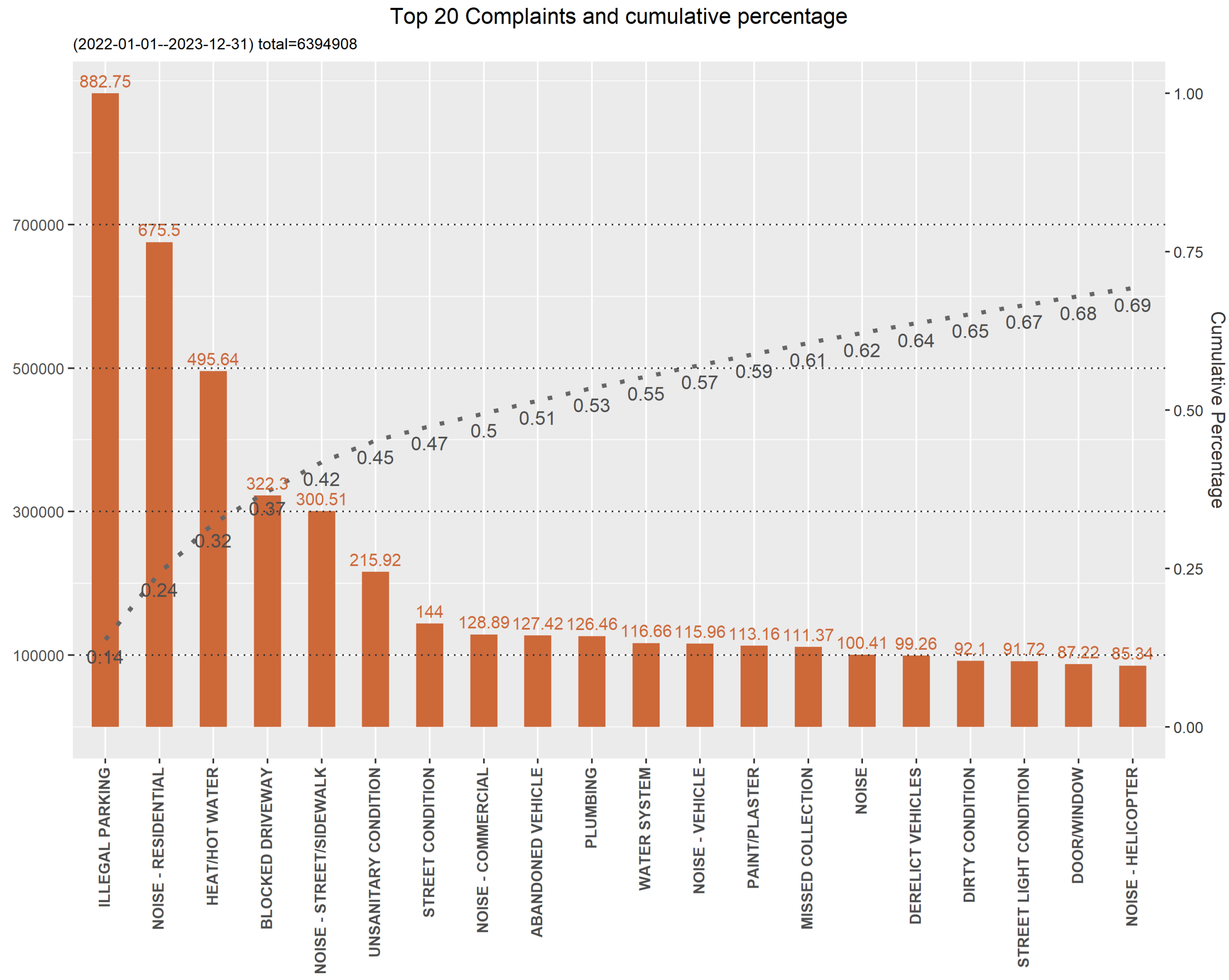


Figure 6: Top 20 SR complaint_type and Cumulative Percentage

The top 20 complaints contain several of the eight different types of "Noise" complaints (Residential, Commercial, Street, Helicopter, etc.). These Noise-related complaints total 1.435 million in the 2022-2023 dataset, a full 22% of all complaint types, which, when consolidated, is the most frequently occurring complaint_type. Ilegal Parking is second, followed by Heat/Hot Water. Noise complaints are handled by NYPD, as is Illegal Parking and Blocked Driveway complaints, which contributes significantly to the top ranking of the NYPD as the lead responsible Agency, receiving a full 43% of all SRs. Housing, Preservation, & Development are 2nd, with 21%

For non-NYC residents, HPD is the City agency that manages the 177,569 NYC public housing units as well as monitoring all City rental and leased apartment units. HPD collectively serves 528,105 people, a population larger than Atlanta or Miami. Hence the large number of Heat/Hot Water and Unsanitary complaints handled by HPD.

complaint_type	**Count**	**Percentage**	**Agency**
NOISE - RESIDENTIAL	676	11	NYPD
NOISE - STREET/SIDEWALK	301	5	NYPD
NOISE - COMMERCIAL	129	2	NYPD
NOISE - VEHICLE	116	2	NYPD
NOISE	100	2	DEP
NOISE - HELICOPTER	85	1	EDC
NOISE - PARK	17	0	NYPD
NOISE - HOUSE OF WORSHIP	2757	0	NYPD

Table 2: Noise-related complaints_type(s) by count with Agency

There are also some curious and unusual Service Requests in the 311 data: tanning, tattooing, trans fat, unsanitary pigeon condition, illegal animal kept as pet, harboring bees/wasps, and the authors' favorite radioactive material. New Yorkers certainly live interesting lives.

4 Data Cleansing Issues

What is data cleansing? Wikipedia Data Cleansing offers a good definition.

"Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data."

Many quality criteria are required in order to process high-quality data. These include:

- Data validation - this effort can span a number of criteria

  - Mandatory fields: Certain data fields cannot be empty.
  - Data-types: Certain fields must be of the correct type, e.g. numeric, character, date, 5 numeric digits, etc. Typically these are identified in a Data Dictionary.
  - Domain adherence: Many data fields must adhere to a specific domain of values, e.g. statuses, state names, zip codes, gender. This includes data fields that are restricted to certain ranges, such as longitudes and latitudes.
- Structural errors to include naming conventions, extra fields not in the Data Dictionary, or inconsistent data entry. For example intermixing blanks, spaces, NA, N/A, and <NA> all to indicate the absence of data.
- Redundant, unnecessary, or irrelevant fields
- Logical inconsistencies such as related fields that violate the nature of that relationship, e.g. a "due date" that is before the "created date"
- Data standardization. Often free-form entry fields can suffer from inconsistent data structures. Efforts to standardize these fields can greatly assist analysis.
- Accuracy and precision, which of course are not the same thing

This analysis effort is to identify the presence of such errors; not to correct them. That effort would be undertaken only after an investigation as to the why and how such errors came about, and a discussion as to whether or not it is even an "error". Typically this requires discussions with subject matter experts (SMEs) and liaison within the various NYC Agency open data coordinators who can help identify the underlying cause for the error. For example, an invalid date field that is provided to the NYC Open Data data lake via an Agency integration could be the result of selecting the wrong field from the source Agency's system, e.g. due_date instead of closed_date. In other cases, without a subject matter expert, it is not possible to actually determine if the data is inaccurate or not. There are many empty values in the taxi_company_borough field. Are those entries missing or does the entry in that field only apply to a very small number of complaint_type(s) or circumstances. Only a SME from the Taxi & Limo Commission (TLC) can provide that answer. The R code used here to analyze the 311 Service Request dataset identifies both obvious inaccuracies as well as potential errors requiring further investigation. It is not a conclusive list of errors, but rather mostly identifies "a place to look" for further evaluation and potential action.

This paper will focus on a set of specific data cleansing areas:

- Examining structural issues within the data
- Validating data type correctness
- Identifying missing, blank, or N/A data
- Identifying invalid values
- Exposing logical inconsistencies & inconsistent or unusual patterns in the data
- Accuracy and precision issues
- Identifying potential redundant or irrelevant data

5 Structural Issues

Structural issues in the context of data cleansing refers to issues related to the how data is organized, formatted, or structured within a dataset. Structural issues can make it difficult to analyze the data effectively. Some common structural concerns in this 311 SR 2022-2023 dataset include:

- Data structure (data fields, columns, formats, data types, etc.) not corresponding to the Data Dictionary
- Verification of correct data types (numeric, character, geospactial, dates, etc.)
- Embedded or combined values: data elements that contain multiple pieces of informa-

tion

Here are some characteristics of the 311 SR data set:

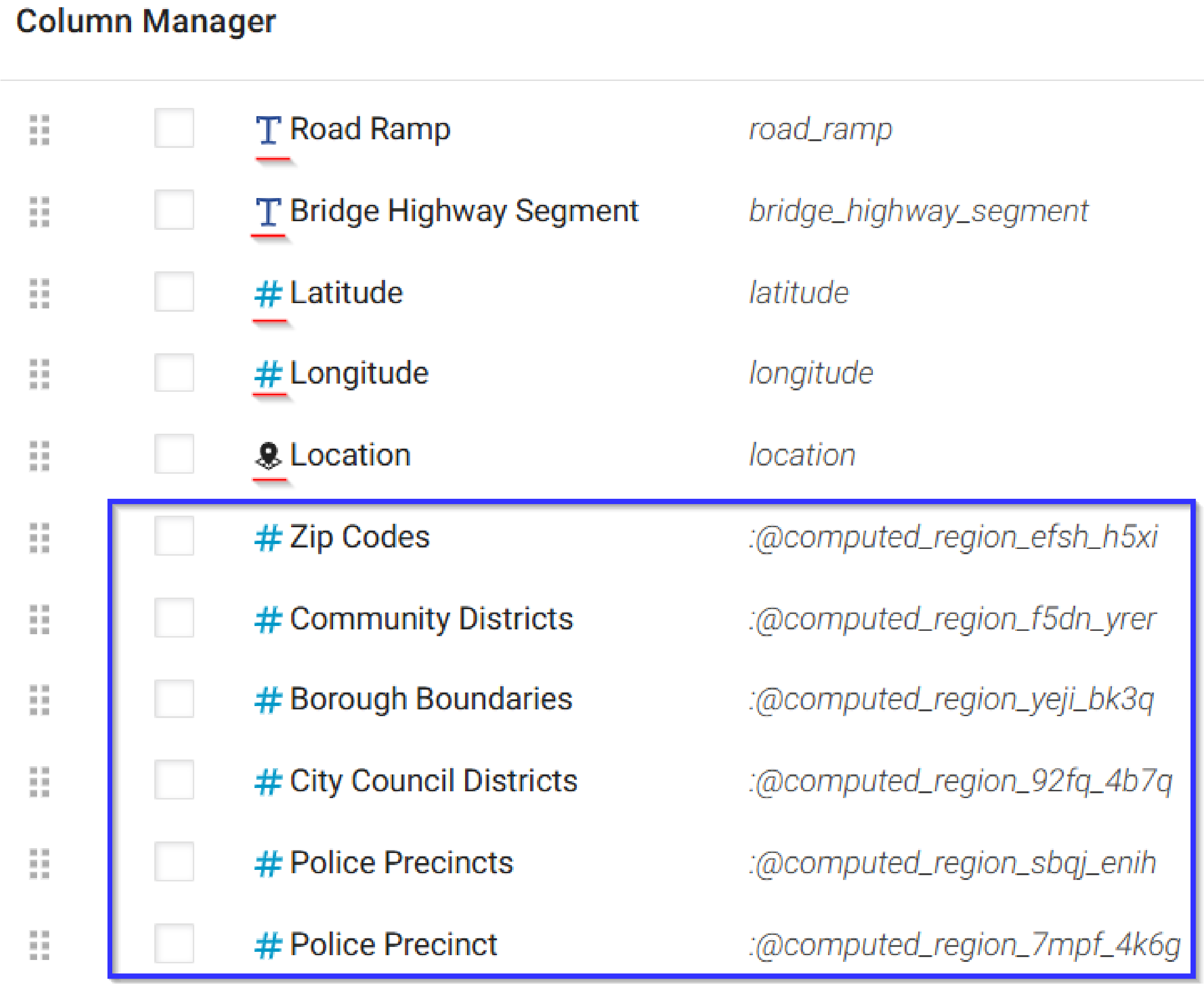
- There are 47 columns of data for each row, exportable as a CSV file.
- There are four date fields (created, closed, updated, due).
- There are three borough fields; two of which we believe to be duplicates.
- Two zip code fields, but not duplicates
- Seven street fields; one pair of which we believe to be duplicates
- Agency name and Agency abbreviation
- Two Police Precinct fields; not duplicates
- In addition to street/city, there are three other location fields: lat/long, NY State plane, and Block #
- incident_address and street_name, e.g. incident_address, 25 Grymes Hill Road and street_name (Grymes Hill Road)
- One free-form text field, resolution_description, which supports 934 characters of input, including commas and special characters

Issue: Fields not in the Data Dictionary The 311 SR Data Dictionary identifies 41 data columns (fields) along with related information for each column. However, when you download the data or explore it via the online portal, you immediately note that there are 47 columns (fields). he additional six fields are not addressed in the Data Dictionary, but do and show up on the Column Manager widget as "@computed_region_xxxx_xxxx". To some extent one can infer what the these @computed_fields are, but it is not possible to know for certain. These six fields are:

- zip_codes

- community_districts
- borough_boundaries
- city_council_districts
- police_precincts
- police_precinct

Here is a screenshot of the NYC Open Data Portal "Column Manager" page showing those columns:



Column Manager

□	T	Road Ramp	*road_ramp*
□	T	Bridge Highway Segment	*bridge_highway_segment*
□	#	Latitude	*latitude*
□	#	Longitude	*longitude*
□		Location	*location*
□	#	Zip Codes	*:@computed_region_efsh_h5xi*
□	#	Community Districts	*:@computed_region_f5dn_yrer*
□	#	Borough Boundaries	*:@computed_region_yeji_bk3q*
□	#	City Council Districts	*:@computed_region_92fq_4b7q*
□	#	Police Precincts	*:@computed_region_sbqj_enih*
□	#	Police Precinct	*:@computed_region_7mpf_4k6g*

Figure 7: Screenshot of data portal showing the "@computed" columns

This, of course, raises several questions:

- How are these fields computed? Using what source data? Using what computational method?
- How can these fields be verified or cross-checked?
- Why are there two fields that appear redundant "police_precinct' and "police_precincts"? Are these redundant?
- What are "borough boundaries"? They appear to be borough names. Why are they referred to as "boundaries"?
- What are "community districts"? Is this the same as the "community_board(s)" that are part of the NYC government?
- Why are not these fields included in the official 311 SR Data Dictionary?
- Can these fields be reliably used for analytical purposes?

The last bullet point is the key here. Are these fields reliable and accurate, and can they be used for subsequent analysis? Unfortunately, as of this date, those questions have not been fully addressed by the NYC Open Data Team. This paper will further address that issue by exploring the accuracy, validity, and candidacy for subsequent analytical usage.

6 Validating data types

Fortunately, both the Data Dictionary and the portal column manager do a good job of indicating the data type. As indicated by the red underline in Figure 7, there is a small icon next to the field name which indicates the data type of the field. The stylized capital "T" indicates "text", the "#" sign indicates "numeric", a calendar icon indicates a "date" field, and the pin on a map symbol indicates a "geospatial" field.

The following fields were checked for their specified data types:

- created_date, closed_date, due_date, and resolution_action_updated_date: All are all in proper date format.
- zip_codes and incident_zip: All are 5 numeric digits except for 2 non-numeric entries for incident_zip (“na” and “N/A”).
- x_coordinate_state_plane & y_coordinate_state_plane (a NY State geo-location system): All are all numeric.
- latitude & longitude: Both are numeric.
- community_districts, borough_boundaries, city_council_district, police_precinct, and police_precincts: All are numeric.
- Free-form text fields that contain a “,” (e.g. incident_address & resolution_description) are properly enclosed in quotation marks.

All data fields were subjected to a test of “correctness” according to their “type” as specified in the Data Dictionary. Of the nearly 300 million data fields to check (6.5 million rows, 47 fields per row), there were only two values that failed to test the test for appropriate data type; those being two text entries in the incident_zip field (“na”, and “N/A”). All date fields were dates. All numbers were numeric. All text fields were character. Data typing appears to not be an issue with this dataset.

7 Blank & N/A data

Understanding the absence of data by field is an important factor when undertaking analysis. For example, if you wanted to see if the SRs were closed before or after their due_date, you would be challenged as 99.57% of the due_date field is blank. (Although with such a large dataset that still leaves 24,411 values, all of which are “N/A’. Which calls into question...are these missing values, or is the concept of a due_date simply not applicable for the vast majority of SRs?

When counting the various fields for blank or N/A values, the fields appear to divide into three groups: **Mostly Empty**, **Partially Empty**, and **Few/None Empty**. he Mostly Empty category ranges from 93-99.9% blank. It includes such fields as taxi_company due_date, pickup_location, and landmark. The Partially Empty includes such fields as location_type, borough, and cross_street. And the Few/None Empty includes created_date, complaint_type, agency, and status. In some cases it may make sense to inquiry as to why some fields are frequently blank. Is the data difficult to capture? Does it pertain to only a small set of complaint_type(s) or Agencies? And if the data is truly that sparse, does it even make sense to collect it? Is it legacy data? For example, one would expect that every complaint_type that has a geographical component, such as an address, would also have a corresponding police_precinct, community_board, and community_district since these are all-inclusive throughout the five Boroughs of NYC. Addtionally, that it would have a lat/long. And generally that is the case, but these fields do have small number, but nonetheless significant, of empty entries.

During the analysis, it was determined that the following fields are mandatory, and any row missing these fields should be removed from the dataset (which fortunately very rarely occurred): created_date, agency, complaint_type, and unique_key. The absence of any one of these four fields prohibits further analysis as currently conducted. Again, it is very rare occurrence, happening only 1-2 times in 6.4 million records. Here is a breakdown of bank and N/A entries by field.

Field	Total Empty	Percent Empty	N/A Count
taxi_company_borough	6391341	99.94	0
road_ramp	6379080	99.75	0
vehicle_type	6376266	99.71	0
due_date	6370497	99.62	6370497
bridge_highway_direction	6367626	99.57	0
bridge_highway_name	6340197	99.14	0
bridge_highway_segment	6340190	99.14	0
taxi_pick_up_location	6329954	98.98	0
facility_type	5966588	93.30	0
landmark	2635736	41.22	0
intersection_street_1	2143847	33.52	0
intersection_street_2	2140707	33.48	0
cross_street_1	1846594	28.88	0
cross_street_2	1846008	28.87	0
location_type	798775	12.49	0
bbl	768109	12.01	0
city	332086	5.19	0
street_name	280463	4.39	0
incident_address	280268	4.38	0
closed_date	248839	3.89	248839
resolution_description	132333	2.07	0
zip_codes	123167	1.93	0
borough_boundaries	101453	1.59	0
community_districts	101445	1.59	0
city_council_districts	101445	1.59	0
police_precincts	101445	1.59	0
police_precinct	101416	1.59	0
latitude	99778	1.56	0
longitude	99778	1.56	0
location	99778	1.56	0
x_coordinate_state_plane	99668	1.56	0
y_coordinate_state_plane	98822	1.55	0
resolution_action_updated_date	91399	1.43	91399
incident_zip	83220	1.30	2
descriptor	74645	1.17	0
address_type	38655	0.60	0
unique_key	0	0.00	0
agency	0	0.00	0
agency_name	0	0.00	0
complaint_type	0	0.00	0
status	0	0.00	0
community_board	0	0.00	0
borough	0	0.00	0
open_data_channel_type	0	0.00	0
park_facility_name	0	0.00	0
park_borough	0	0.00	0
created_date	0	0.00	0

Table 3: Blank and N/A entries by Field

Here is a graphic depiction of total empty (blank & N/As) for each fields. You can see the natural grouping into the Most, Some, and Few categories in the stair-type visualization.

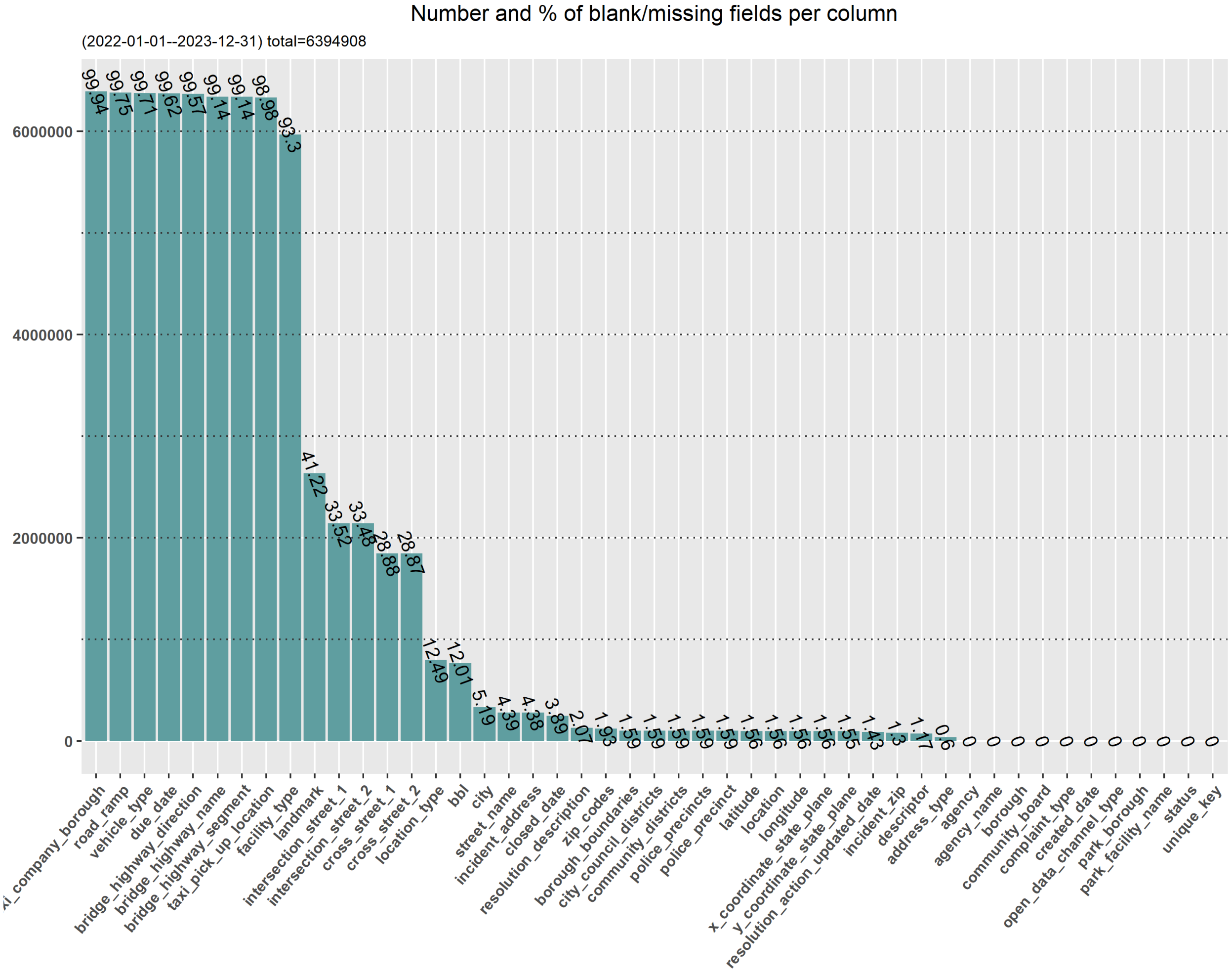


Figure 8: Number and Percentage of Empty Entries

8 Validating Data for Acceptable Values

While the dataset may have fields populated with data that appears correct, it is necessary to inspect those fields to determine if the data is indeed valid. For example a date field that contains the value "February 30, 2024" is clearly not valid. Nor is a latitude of +95 deg north.

In this study, testing for invalid values revealed several serious issues associated with select fields. Accordingly, an analyst would need to be extra diligent if using these fields for analysis. In all likelihood, rows with invalid field values would need to be removed from the dataset before conduction further analysis. Here are some results.

- Latitude and Longitude fields were tested to ensure all fell within the geographic boundaries of the City of New York. All did.
- The unique_key field was in fact unique.
- Many fields have a domain of acceptable values. Often these values are determined from common usage or by examining larger historical datasets. Unfortunately, these domains of acceptable values are not specified in the Data Dictionary. These fields all tested as compliant within their domain of acceptable values as determined by examining a 10-year dataset:
  - address_type
  - status
  - borough, borough_boundaries, & park_borough
  - data_channel
  - vehicle_type
  - city_council_district

8.1 Issues with Zip Codes

Unfortunately some fields proved to be problematic when comparing the values to a domain of legal values. For example, all zip codes (two fields: zip_codes and incident_zip) should be valid as defined by the USPS database which contains 37,946 valid zip codes.

The computed field zip_codes proved especially problematic with 58% (3.6 million) of the field entries being invalid. That high number indicates that the computation of this field is highly inaccurate. We recommend dropping this field for future analytic efforts.

Furthermore, the field incident_zip while having only .07% invalid entries, still is a problem with 4163 such errors. As the zip code is a primary measure of many NYC city services, and with a freely available database for which to validate entries, it seems an oversight that any such errors could creep into the 311 system. Some invalid entries can clearly be identified as incorrect just by observation, e.g. 10000, 12345, 11111, etc.

The breakdown of the invalid entries in the zip_code, sorted by Agency shows that the distribution by percentage almost precisely mirrors the overall breakdown of SRs by Agency as shown in . This indicates a systemic problem, and since the zip_code fields is one of the computed fields, it appears these errors are caused by incorrect computations rather than by any Agency mistake.

The errors in the incident_zip field are more troubling, even though they are, percentage-wise, small. Here is a graphic illustrating how these errors occur by Agency. Here we can see that the majority of these errors lie in the Dept of Transportation, NYPD, Taxi & Limo Commission (TLC), and the Economic Development Council (EDC) among others. This representation does not follow the full SR distribution indicating that these errors are likely generated by an incorrect process or application residing at those particular Agencies; it is not a systemic problem.

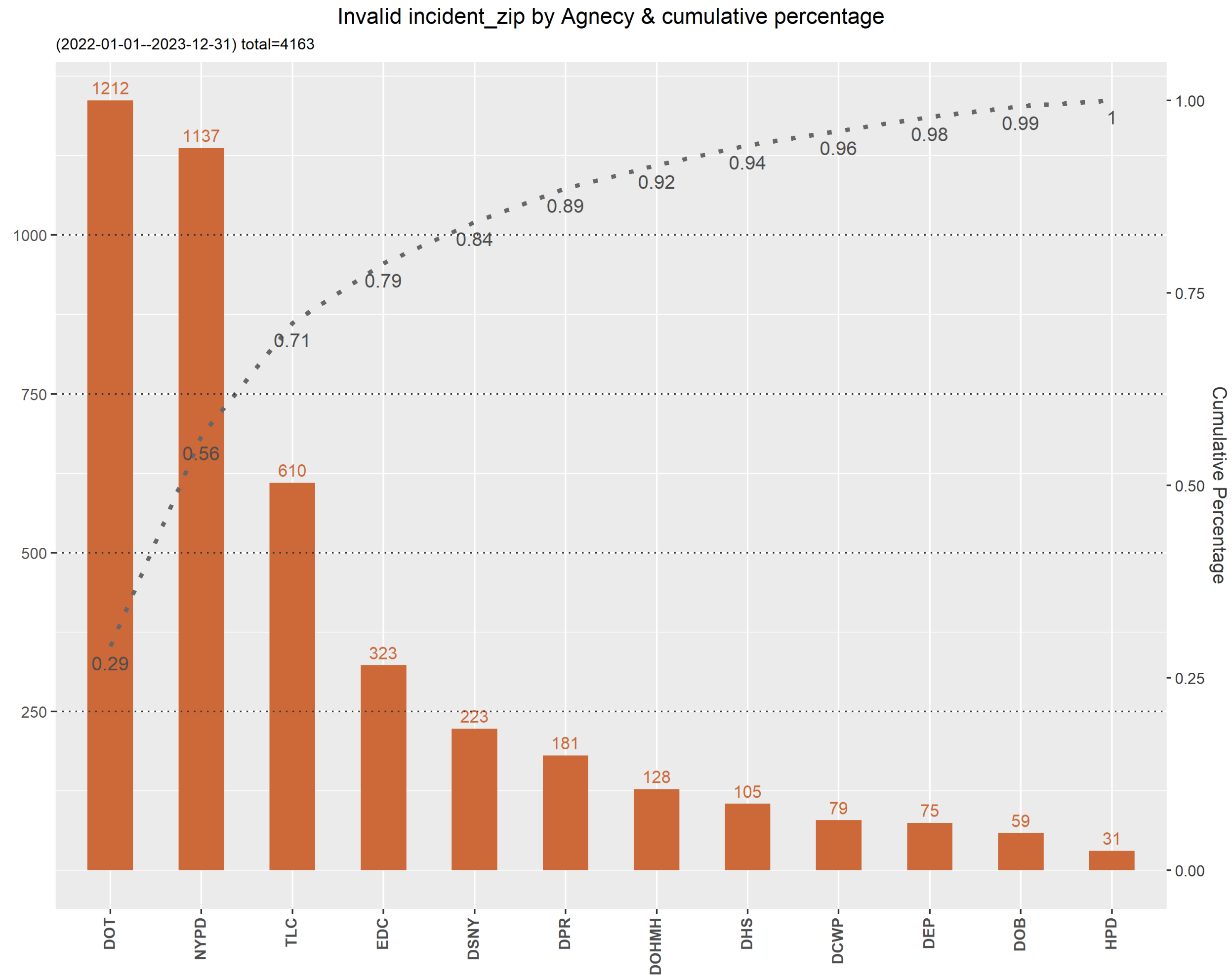


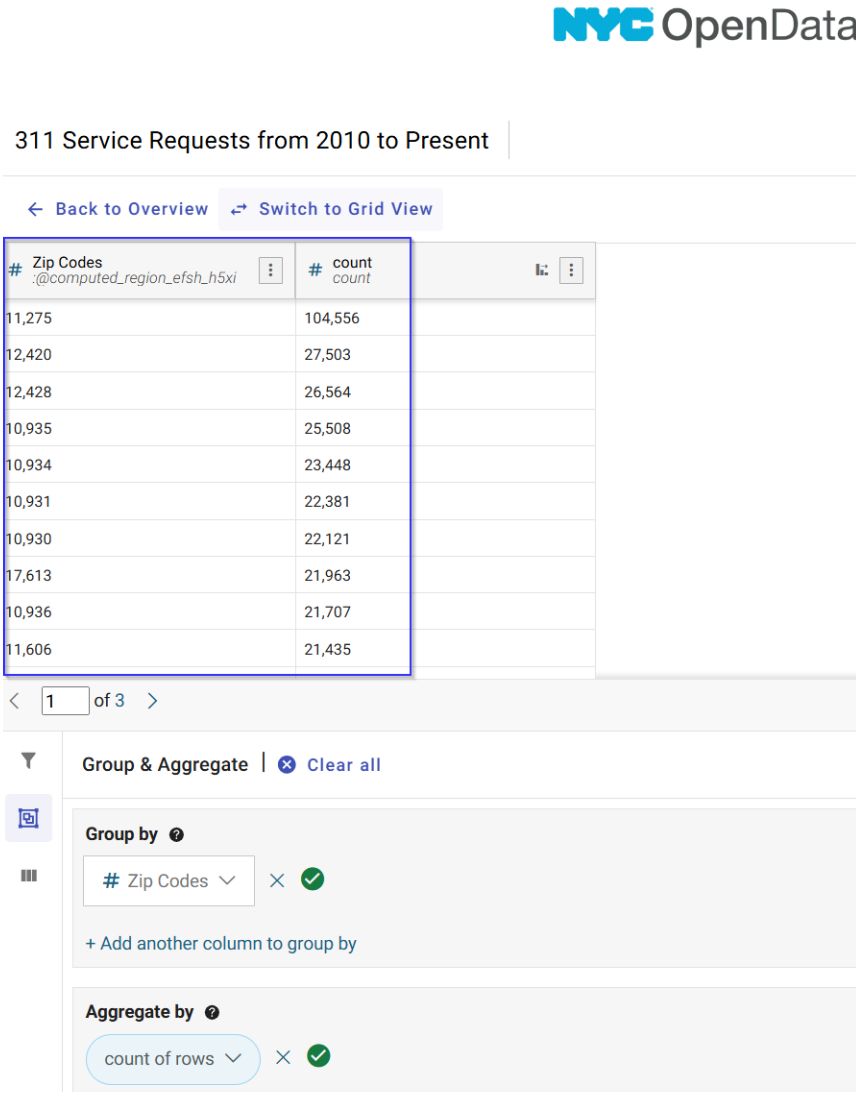
Figure 9: Invalid incident-zip by Agency

8.1.1 Case Study: Noise Complaints by Zip Code

labelsec:case-study-zip-codes **Scenario:** The NYC Office of Nightlife (ONL) wants to know "What are the top 10 zip codes for Noise Complaints (all 8 types) over the last two years?" The goal is to assess the impact of the recent NYC effort looking to promote a safe and vibrant nightlife scene in NYC while seeking to ease strained relations between bar and club owners.

It may come as a surprise to non-NYC residents that dancing at bars and restaurants has been illegal since 1926 (during prohibition), and was only just repealed in 2018.

On the surface this would seem like a simple effort. The NYC Open Data Portal allows for selection of a timeframe (2022-2023), complaint-type begins with "Noise", and group by Zip code, sorted descending by count. You can do this analysis and grouping right inside the NYC Open Data Portal. *Voila!*



311 Service Requests from 2010 to Present

# Zip Codes :@computed_region_efsh_h5xi	# count count
11,275	104,556
12,420	27,503
12,428	26,564
10,935	25,508
10,934	23,448
10,931	22,381
10,930	22,121
17,613	21,963
10,936	21,707
11,606	21,435

1 of 3

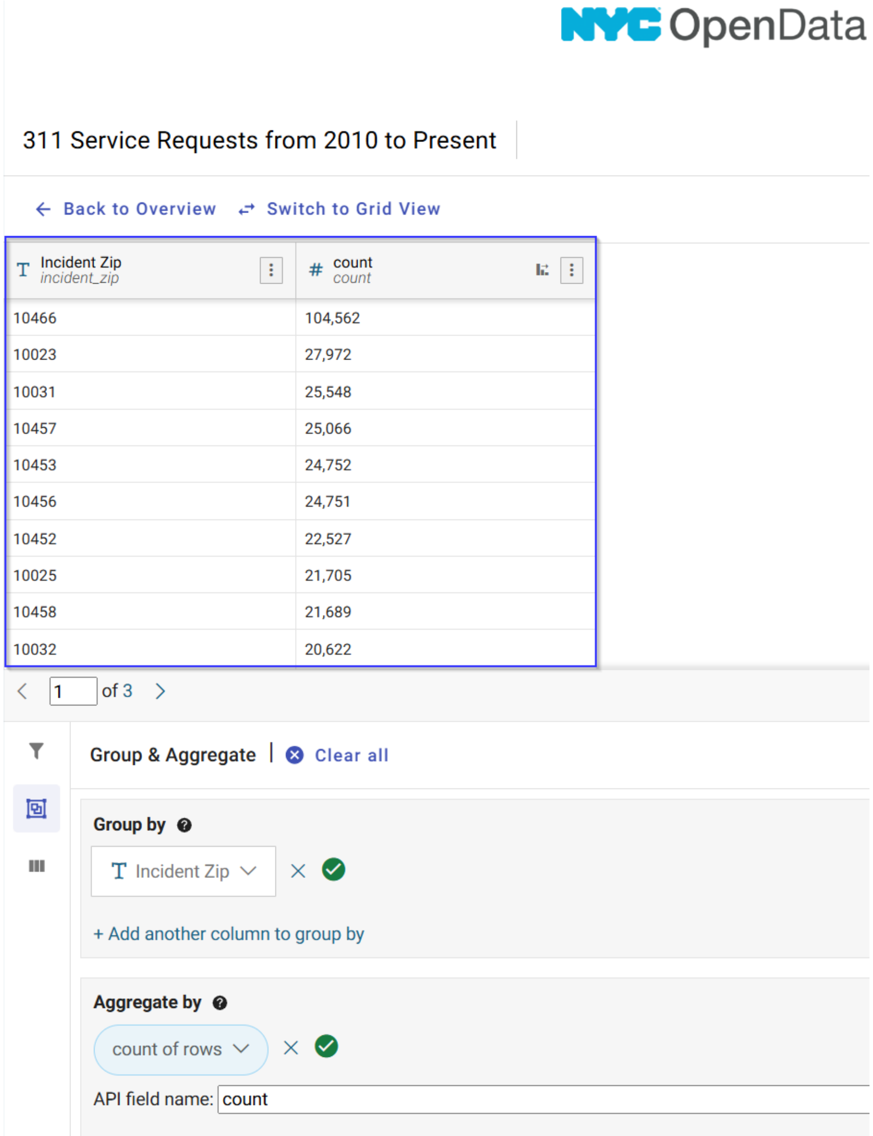
Group & Aggregate

Group by: # Zip Codes

Aggregate by: count of rows

Figure 10: Top 10 zip codes for Noise complaints (Zip Codes)

However, this analysis uses the zip_codes field, one of the (six) computed fields that has shown to have validity problems. If we repeat the analysis with the incident_zip field instead of the zip_codes field, we obtain a different set of zip code results:



NYC OpenData

311 Service Requests from 2010 to Present

Incident Zip incident_zip	count count
10466	104,562
10023	27,972
10031	25,548
10457	25,066
10453	24,752
10456	24,751
10452	22,527
10025	21,705
10458	21,689
10032	20,622

Group by: Incident Zip

Aggregate by: count of rows

API field name: count

Figure 11: Top 10 zip codes for Noise complaints(Incident Zip)

Let's subject these two data sets to validation against the US Postal Service zip code database.

zip_codes			**incident_zip**		
Zip Code	**Count**	**Valid?**	**Zip Code**	**Count**	**Valid?**
11275	104,556	FALSE	10466	104,562	TRUE
12420	27,503	TRUE	10023	27,972	TRUE
12428	26,564	TRUE	10031	25,548	TRUE
10935	25,508	FALSE	10457	25,066	TRUE
10934	23,448	FALSE	10453	24,752	TRUE
10931	22,381	TRUE	10456	24,751	TRUE
10930	22,121	TRUE	10452	22,527	TRUE
17613	21,963	FALSE	10025	21,705	TRUE
10936	21,707	FALSE	10458	21,689	TRUE
11606	21,435	FALSE	10032	20,622	TRUE

Table 4: Comparison of Top Ten Zip Codes Lists

As indicated, six out of ten zip_codes are invalid, which corresponds closely with what is observed in the overall dataset (58%). Whereas the incident_zip field is completely valid, again in-line with the overall incident_zip validation percentage (99.04%). If the analysis performed using the zip_codes field were to be presented to the Director, ONL, the conclusion would be in error. Perhaps more curious than the differing zip codes on the two lists, is the fact that the numerical counts are nearly identical; the overall counts are quite close. This is curious since one field is computed while the other is via data entry; the computed protocol is again the suspect. The algorithm appears to be counting corrrectly (or nearly so) but mis-identifying the corresponding zip code.

8.2 Issues with Police Precincts

A curious case also exists when examining the two nearly identical fields - police_precincts and police_precinct. Both of those fields are among the "computed" fields in the dataset. Using NYPD Precinct listings it's possible to determine the valid police precincts; they are generally numeric or have numeric representations in the dataset.

However, what we find is that both fields police_precinct and police_precincts have 35% invalid entries. Unfortunately, they're not the same invalid entries.

For the police_precincts field, there are 2,171,864 invalid entries (35%)., a consequential number of errors. Similarly for the police_precinct field, there are 2,171,778 invalid entries (also 35%), however they are not the same counts for the two similar fields.

The top ten (by count) of invalid precincts for each field are:

Invalid Precinct	**police_precinct count**	**police_precincts count**	**Delta**
62	151808	151720	88
72	148881	148796	85
67	133171	133115	56
64	111236	111221	15
60	105252	105317	-65
65	103334	103306	28
53	100097	100074	23
73	97746	97755	-9
68	94204	94254	-50
66	93654	93848	-194

Table 5: Comparison of the fields police_precinct and police_precincts

A graphic representation of the invalid police_precincts plotted by Agency show a near perfect match to the distribution of the entire 6.4 million SRs as seen in Figure **??**. Note that the "big six" Agencies account for 90% of these invalid precinct errors, identical to the overall distribution of all SRs. (The same is true for a similar graph of by the police_precint field.) This strongly suggests that this is a systemic problem, almost certainly caused by the computation method applied for these two fields.

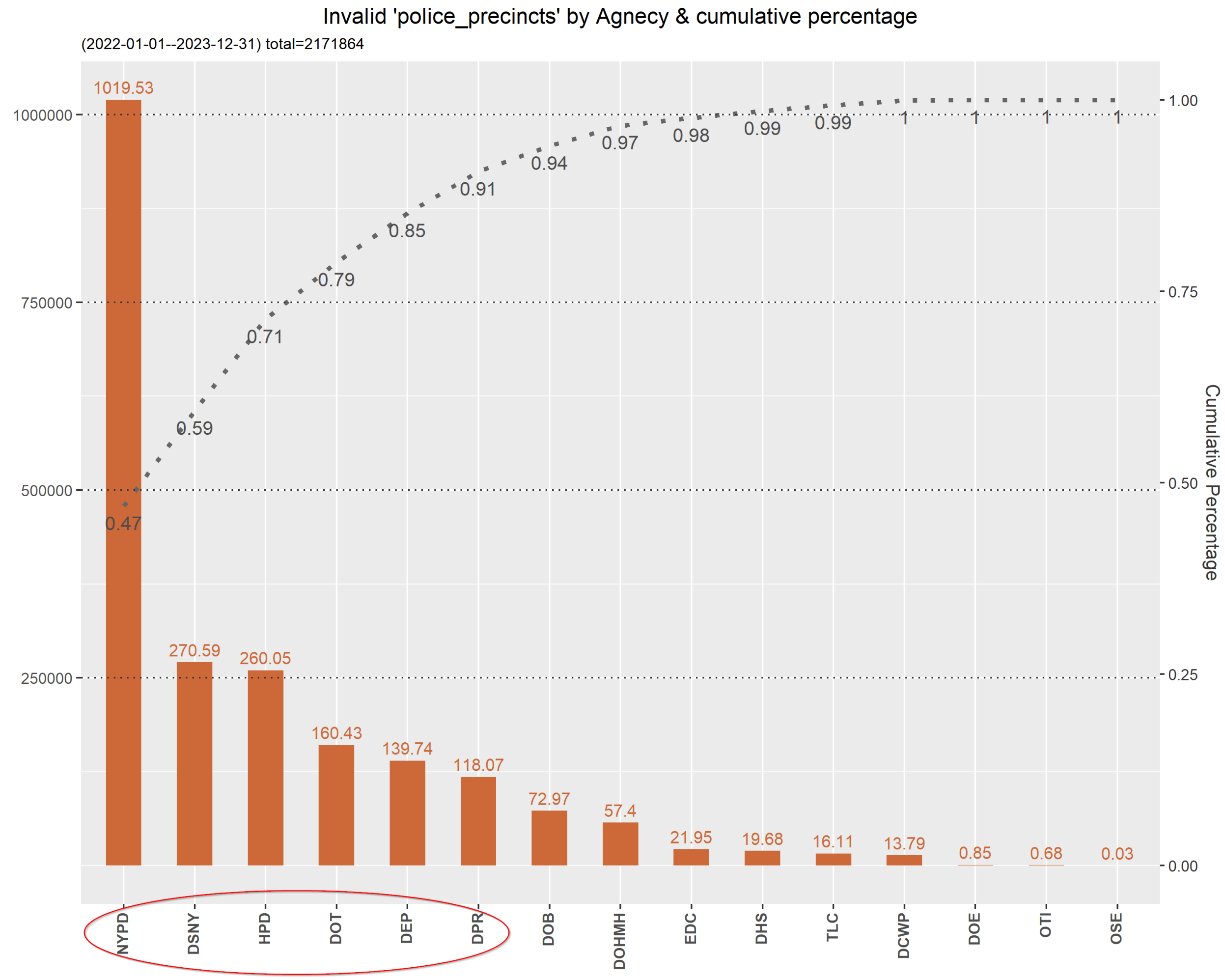


Figure 12: Invalid NYPD precincts by Agency

8.3 Issues with Community Boards

Community Boards are an important aspect of NYC government. They are the most local, grassroots form of City government, and a vital connection between communities and elected officials and City agencies. They are organized by the five Boroughs of NYC and represented in this dataset as “##-Borough”, e.g. “10 Bronx” and report to the various Borough Presidents in NYC. As such, they play an important role in the quality of life for all New Yorkers. There are 59 community boards throughout the City: 12 Community Districts in the Bronx, 18 in Brooklyn, 12 in Manhattan, 14 in Queens and 3 in Staten Island. The Community Boards are used frequently as a way to measure distribution of City services throughout the five boroughs, so correctness is important. NYC Borough Presidents in particular are focused on the services offered at the Community Board level within their respective Borough.

In the 2022-2023 dataset there are 27,276 invalid community_board entries which represents 0.43% of non-blank data. There are a total of 12 different invalid community boards, to include:

Rank	**Invalid CB**	**Count**
1	64 MANHATTAN	7144
2	83 QUEENS	5133
3	55 BROOKLYN	3327
4	81 QUEENS	3180
5	80 QUEENS	2407
6	26 BRONX	2364
7	28 BRONX	1128
8	82 QUEENS	1054
9	95 STATEN ISLAND	515
10	27 BRONX	514

Table 6: Top Ten Invalid 'community_board' Values

The distribution of invalid Community Boards by Agency is not consistent with the overall SR Agency distribution. This indicates that there are likely specific issues at some key Agencies, in this case Taxi & Limo Commission (TLC), Parks & Recreation, etc. It's not a large error, but the division of services (and complaints) by Community Board is a well-tracked measure so accuracy is again important.

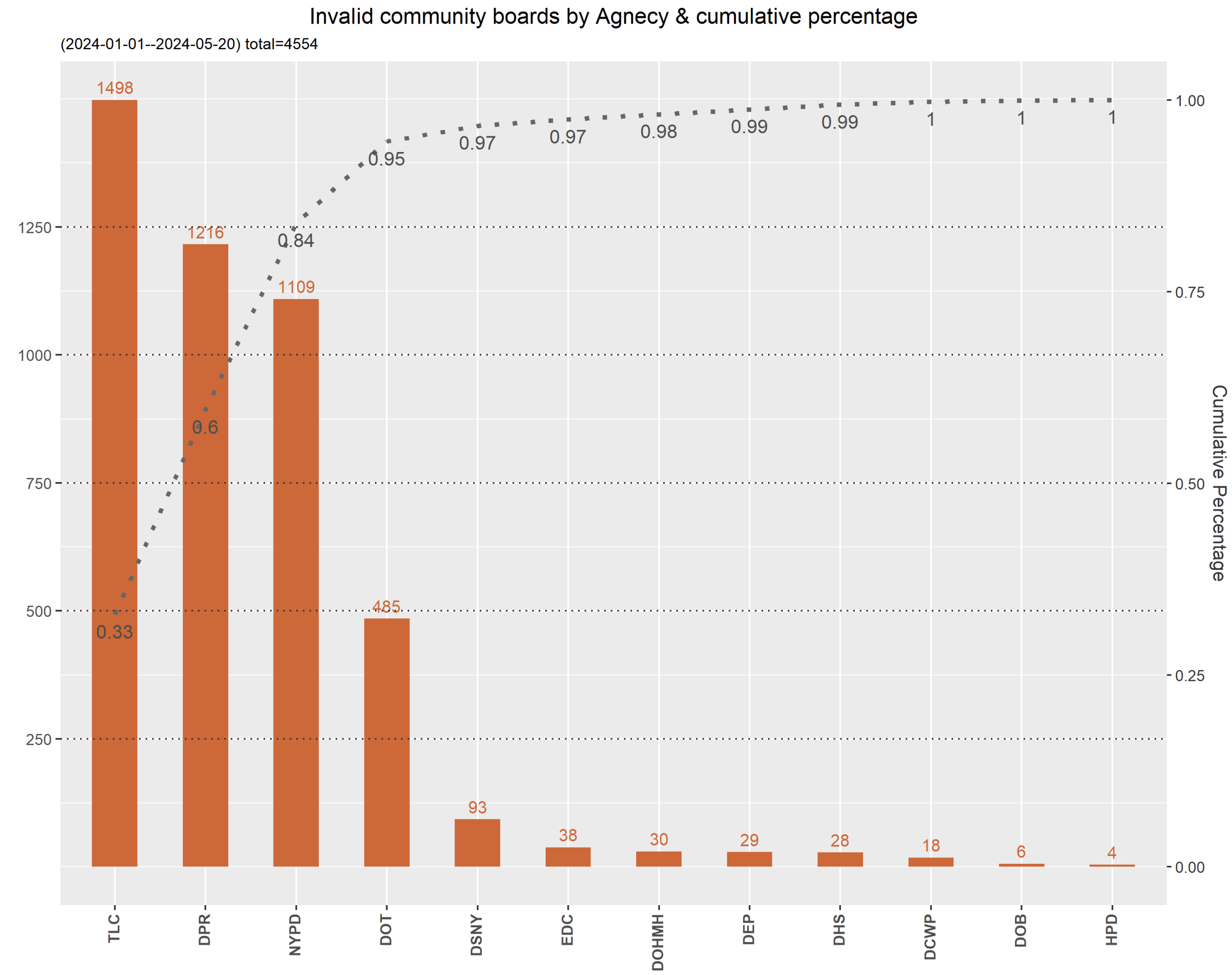


Figure 13: Invalid Community Board by Agency

8.4 Issues with Community Districts

Another of the computed fields is community_districts. Community Districts are the boundaries for the Community Boards, but unlike how Community Boards are an arm of the government of New York City, the Community District is used by the Department of City Planning (DCP) for purposes of environmental, socio-economic, and demographic purposes. DCP is the NYC"s primary land use agency and is instrumental in the City's physical framework. It is a geographical division rather than a local government division. And like Community Boards, the Community Districts are frequently cited as a means to track zoning, housing, community facilities, and waterfront/open/public spaces. It is another of the NYC well-tracked measures.

Due to how the community_district data is formatted, it is not possible to establish validity. However, it is possible to determine that the dataset contains 72 unique entries, while there are only 59 valid Community Districts. So it would appear that there are 13 invalid community_board(s) contained in the dataset.

As previously mentioned, the dataset has four data fields: created_date, closed_date, due_date, and resolution_action_update_date. The due_date field has few entries to analyze; 99.62% of due_date entries are "N/A". It is not known if the due_date(s) are truely "not applicable" or that the due_date field is rarely used. The resolution_action_update_date field serves as a date-time marker showing when the SR record was last updated. All the date fields are recorded to the second, e/g/ "2022-02-11 13:43:05".

8.5 created_date and closed_date(s) – Negative Duration

Let's look first at the created_date and closed_date fields. You might expect that these two fields are automatically populated by the SR application software, either the software used at the 311 organization or that used at the Agencies that are integrated with the 311 system; and typically that is indeed the case. In such a case, the act of creating an SR would automatically trigger an entry in created_date based on a system-wide clock. Thus the created_date is associated with setting an SR status to "new" or "open". And similarly when an SR's status is changed to "closed" then the closed_date would be populated.

Unfortunately, our analysis found that that assumption is not always the case. When you analyze the various date fields, several anomalies become quickly apparent:

- Some SRs have a closed_date that occurs before the created_date.
- Some created_date(s) and closed_date(s) appear either in the future or in the far distant past.

- There are a large number of SRs that are closed and created exactly at midnight or exactly at noon.

Why are these date fields important? They're important because citizens, NYC Government Officials, and NYC Agencies use these dates to measure the responsiveness for providing the services. How long does it take to replace an out-of-order street light? How rapid is the response to a report of domestic violence? Does the NYPD response to a noise complaint in Staten Island take longer than in Manhattan? How long does it take to repair a street pothole in Queens?

The measure of "duration", of the response time of a 311 call is a closely watched, carefully scrutinized metric, both in terms of overall performance and with a geographical area. And it can have broad political, organizational, and budgetary consequences. Duration is not a field in the dataset, but it is easily computed by closed_date - created_date. Let's look at some anomalies in this "duration" field.

Closed-before-Created: There are 12,251 SRs that are closed before they were created, thereby generating a nonsensical "negative duration". While this is a small percentage overall (0.2%) it can have significant impact on response time analysis. In most cases, these "negative durations" appear to just be mistakes. Here is a sample of some of the largest and the smallest errors:

Largest errors (days) excluding extreme negative values

created_date	closed_date	duration	agency
2023-01-27 14:40:00	2022-01-14 14:40:00	-378.000000	DOT
2023-01-18 10:06:00	2022-01-12 10:06:00	-371.000000	DOT
2023-01-27 14:36:00	2022-01-22 14:35:00	-370.000694	DOT
2023-01-11 11:10:00	2022-01-09 11:10:00	-367.000000	DOT
2023-12-18 03:13:00	2023-01-16 13:10:00	-335.585417	DOT

Smallest errors (days)

created_date	closed_date	duration	agency
2023-06-28 07:07:31	2023-06-28 07:07:00	-0.000359	DOT
2023-06-29 09:10:20	2023-06-29 09:10:00	-0.000231	DOT
2023-01-12 06:50:13	2023-01-12 06:50:00	-0.000150	DOT
2023-06-26 08:09:07	2023-06-26 08:09:00	-0.000081	DOT
2023-01-12 06:51:01	2023-01-12 06:51:00	-0.000012	DOT

Table 7: Largest and Smallest errors (days)

The largest errors are shown "excluding extreme negative values". We found eight SRs with extremely large negative durations (< -730), all containing an entry of "1900-01-01" as the closed_date, which generates extremely large negative durations exceeding 44,601 days (122 yrs). That large of an anomaly can skew statistical results, even though the number of occurrences is small. This is especially curious as the date "1900-01-01" is also the earliest date that Microsoft Excel can accommodate as well as how Excel treats a blank cell that is designated as a date. All raising suspicions that perhaps these dates were imported into the 311 system from Excel.

According, these SR rows are removed from the box & whiskers plot. Note that these extremely large negative duration SRs all are from the Department of Homeland Services

(DHS) Sample:

created_date	closed_date	duration	agency
2022-02-11 13:43:05	1900-01-01	-44602	DHS
2022-02-11 11:03:38	1900-01-01	-44602	DHS
2022-02-11 08:29:49	1900-01-01	-44601	DHS
2022-02-11 14:08:08	1900-01-01	-44602	DHS
2022-02-11 14:05:16	1900-01-01	-44602	DHS

Table 8: Sample of SRs with extremely large negative durations

Here is a graphic visualization of the negative-duration SRs by Agency. It is obvious that this negative-duration issue is a problem at the Dept. of Transportation (DOT) where 95% of these types of errors occur.

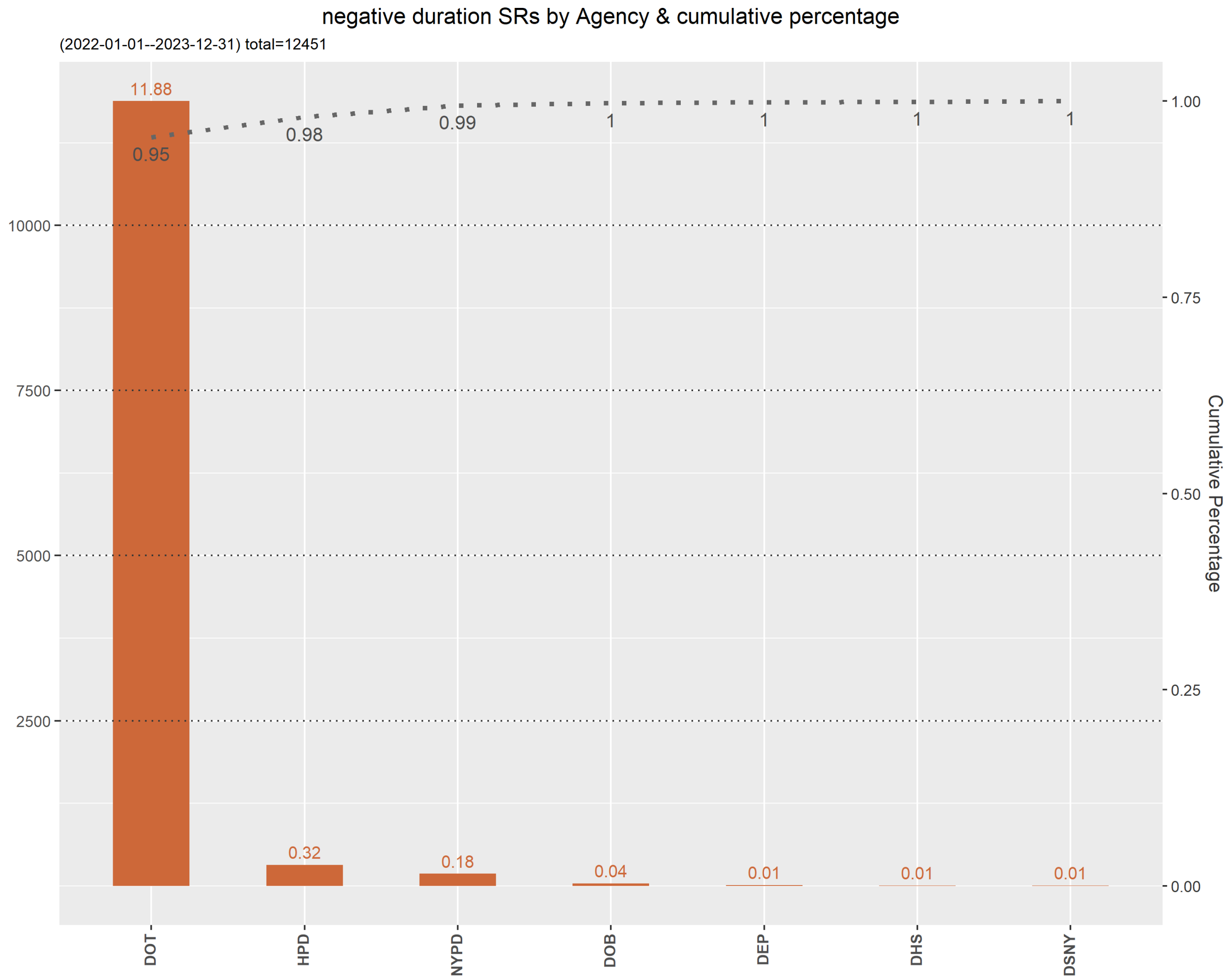


Figure 14: SRs with a Negative Duration by Agency

This violin chart shows the broad spread of negative-duration SRs, albeit with the extremely large negative-duration SR removed. While there are few outlier points, the magnitude of the negative-durations is troubling and can produce bizarre analytical results.

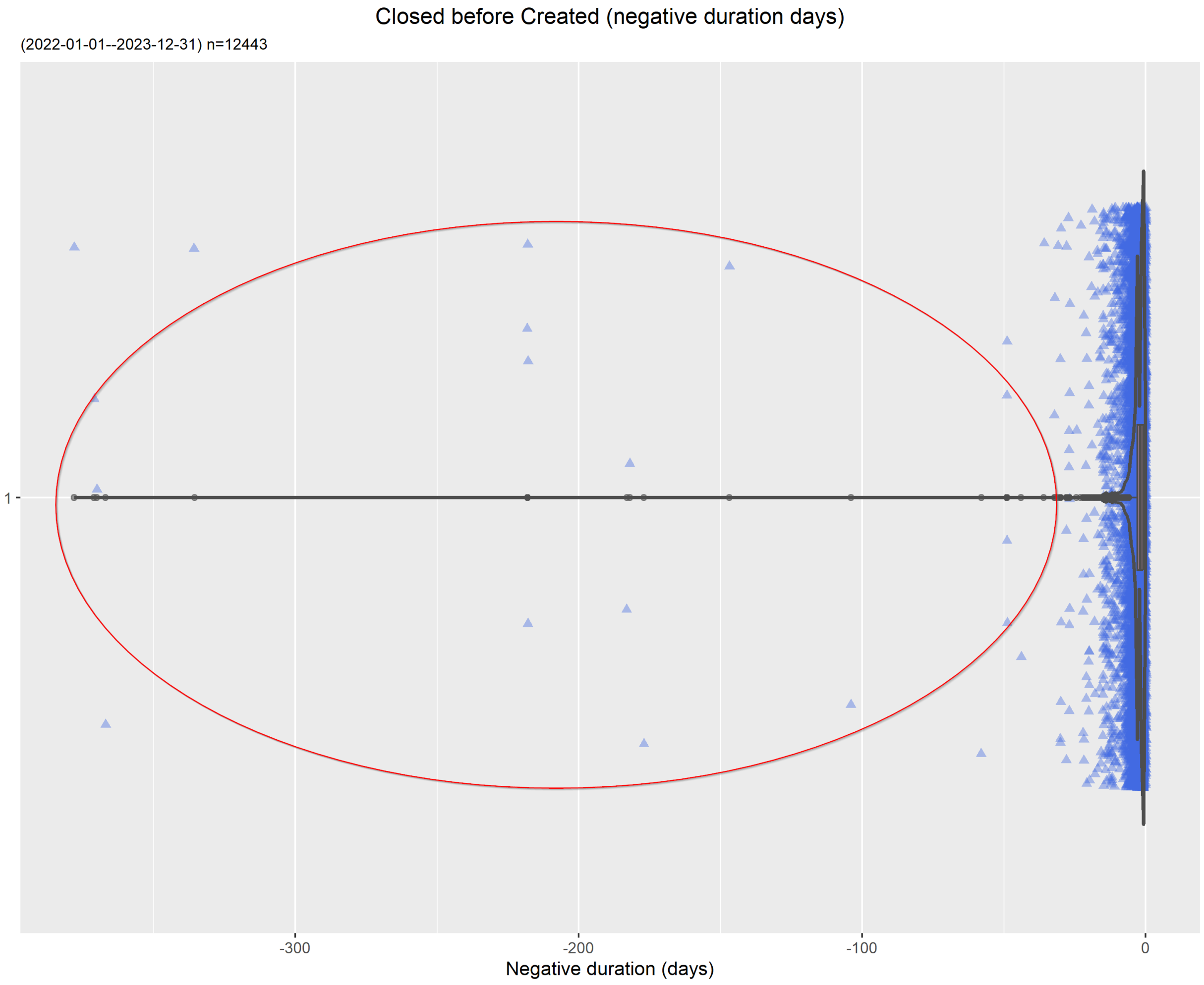


Figure 15: Negative duration distribution

8.5.1 Case Study: Homeless Person Assistance

Scenario: The Dept. of Homeless Services wants to know "How quickly are 311 calls for "Homeless Person Assistance over the past two years?" resolved. This is a typical request made by both the public and City government. It's a performance metric along the lines of "How quickly is my Agency responding to "critical" requests, and does that performance vary by Borough, Zip Code, etc.

As an analyst, this appears to be a routine effort:

1. Select data from 2022-2023 and filter the data by complaint_type = "Homeless Person

Assistance” (yields 55,000 SRs)

2. Compute a new field “duration” (closed_date – created_date)
3. Take an average of the “duration” field == **Answer: -4.8 days**

Clearly that answer is nonsensical. How did such a simple task result in an absurd answer? The answer lies in the computation of the “duration” field. As it turns out, there are eight DHS SRs that have a closed_date of “1900-01-01”. Each of those SRs creates a negative duration of -44,602 days (-122 years). So just those eight SRs with extreme negative duration values is enough to drive the average of the 75,000 Homeless Assistance SRs to a negative value, clearly incorrect.

As it turns out the distribution of the “Homeless Person Assistance” is quite skewed due to a small number of long duration SRs. In this case, the median is a better measure of central tendency than the average. And the median of the duration field is 0.2 days (approx 5 hrs). Here is a look at the distribution of durations for the Homeless Person Assistance SRs showing the large outliers.

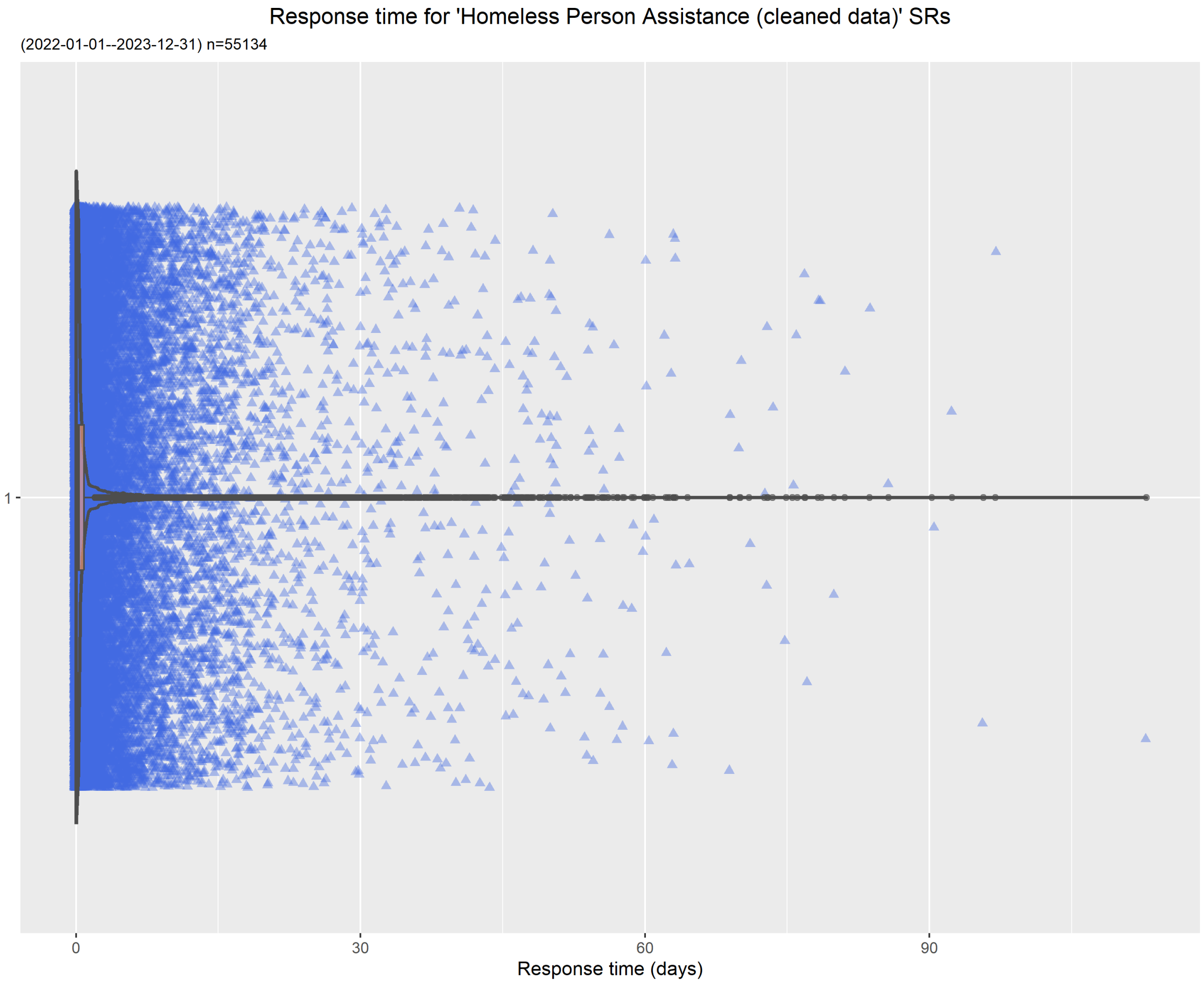


Figure 16: Homeless Assistance SR Durations

8.6 created_date and closed_date(s) – Zero Duration

An even more serious, and more prevalent problem with closed_date(s) occurs when the closed_date and created_date are exactly the same – to the second – which accordingly creates a **zero duration'**. Again this is nonsensical, but the presence of such zero durations can again severely distort analytics. There are 191,141 such SRs representing 3.1% of all non-blank data; not an insignificant number. As shown in the chart below, 99% of the zero duration SRs appear to predominately occur in five Agencies (Dept of Mental Health &Hygiene, Dept of Transportation, Dept of Business, Dept of Sanitation, and Dept of Environmental Protection.) This is not in-line with the overall Agency distribution of SRs and indicates an Agency-specific issue, and the solution likely lies at those Agencies.

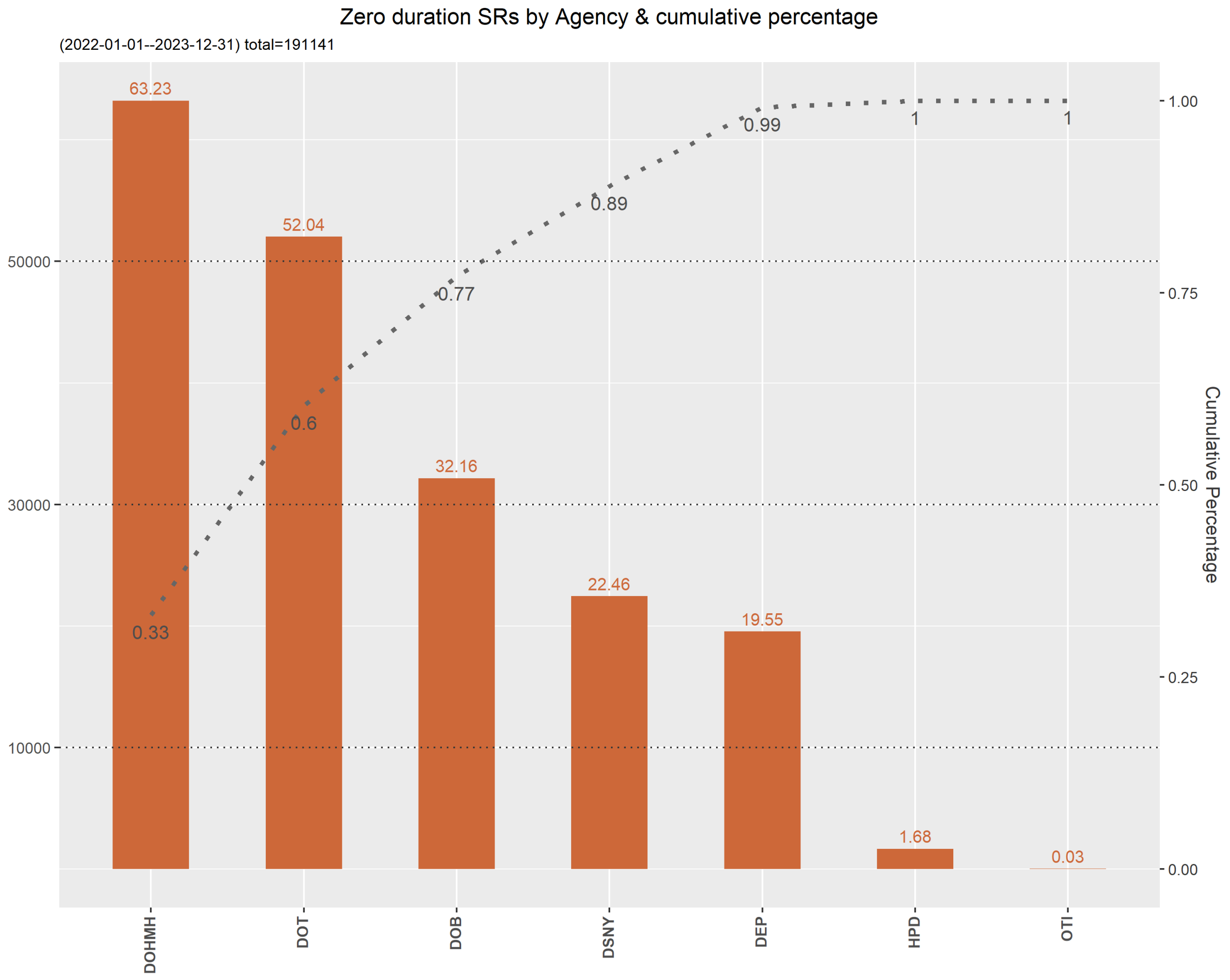


Figure 17: SRs with Zero Durations by Agency

8.7 created_date and closed_date(s) – Midnight &Noon

An additional problem discovered with the created_date and closed_date fields; there appears to be an unusually large number of SRs created or closed at exactly midnight (00:00:00) and exactly noon (12:00:00), to the second. The distribution of SR creation and closure largely follows the work-day clock with many SRs created during day-light hours, and fewer SRs created at night and in the early hours of the morning. However, here there appears to be significantly greater numbers of SR closed exactly at midnight and exactly at noon, as well as a significant number of SRs created exactly at noon. The reason behind this anomaly is worth investigating as it affects the SR "duration" question. If SR durations are used to measure responsiveness for the delivery of services, and if those times are not an accurate reflection of the start/finish of the Service, then the use of duration as a service quality metric is called into question. Observations during this 2-year period include:

- There were 99,779 SRs created exactly at noon (12:00:00)
- There were 235,347 SRs closed exactly at midnight (00:00:00)
- There were 105,505 SRs closed exactly at noon.

There is, of course, a broad distribution of the time that SRs are created by day and hour, by night and day, by weekday and weekend. One would think that that distribution would be (mostly) evenly distributed by the minute and second. That is, if there are normally 100 SRs created between 1-2pm, that the distribution to the minute and second would be close to random. For example, there is no reason to suspect that a larger number of SRs would be created at 6 minutes past the hour rather than at 5 minutes past the hour. And certainly, if you look at the second that these SRs are created you would again expect a (more-or-less) random distribution; that there would be no reason to believe that a higher (or lower) number of SRs would arrive at 18 seconds past the minute versus at 39 seconds past the minute. Unfortunately, that turned out not to be true.

Let's look at the SR creation distribution. We discovered that there are a significant number of SRs created exactly at noon. To reveal this trend, we looked at the exact date-time stamp of the created_date(s). We began by selected only SRs that were created exactly on the minute; at 09:00:00 or 13:00:00 for example. Here is a visualization of the number of SRs created and closed by hour-of-the-day:

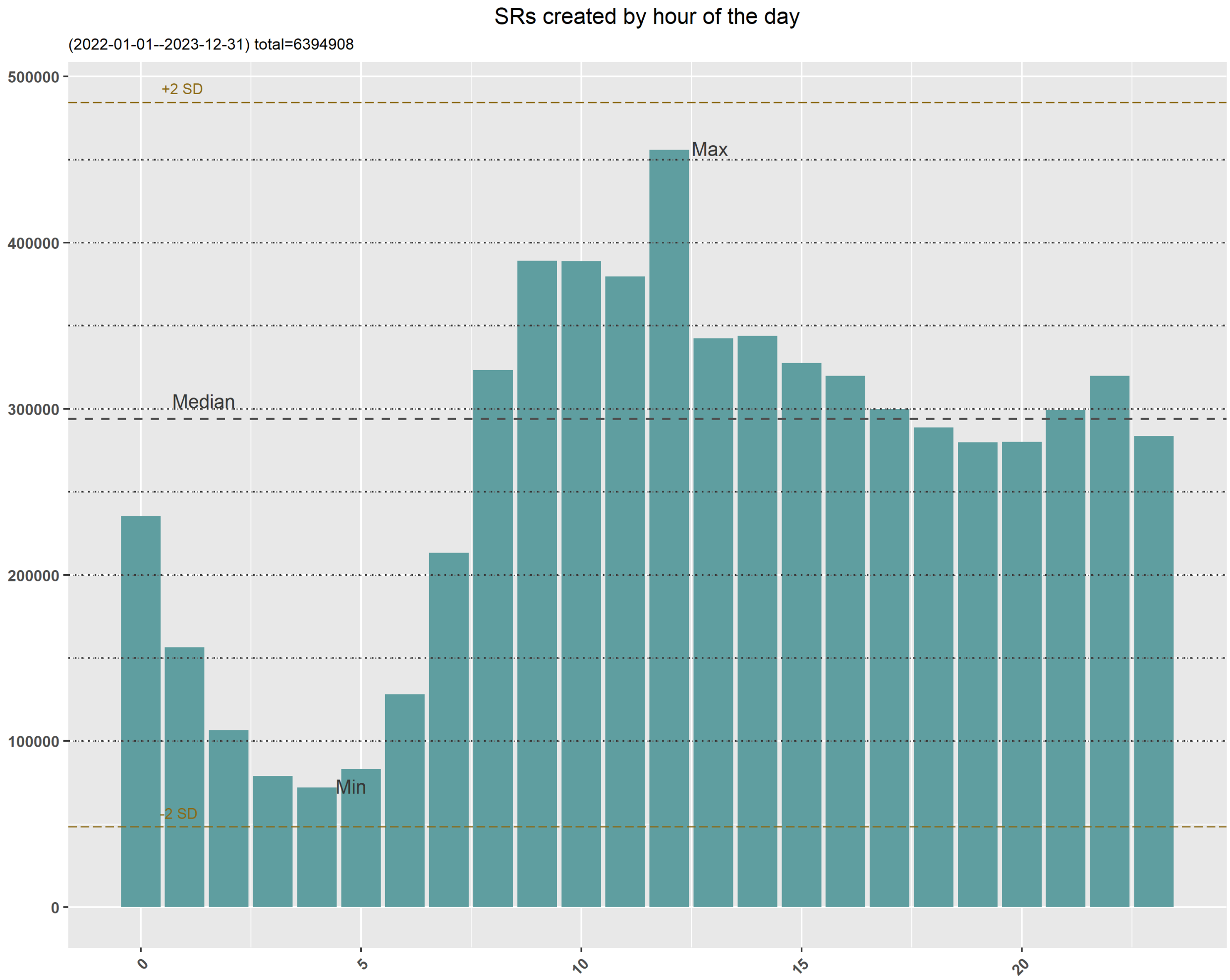


Figure 18: SRs by Hourly created_date(s)

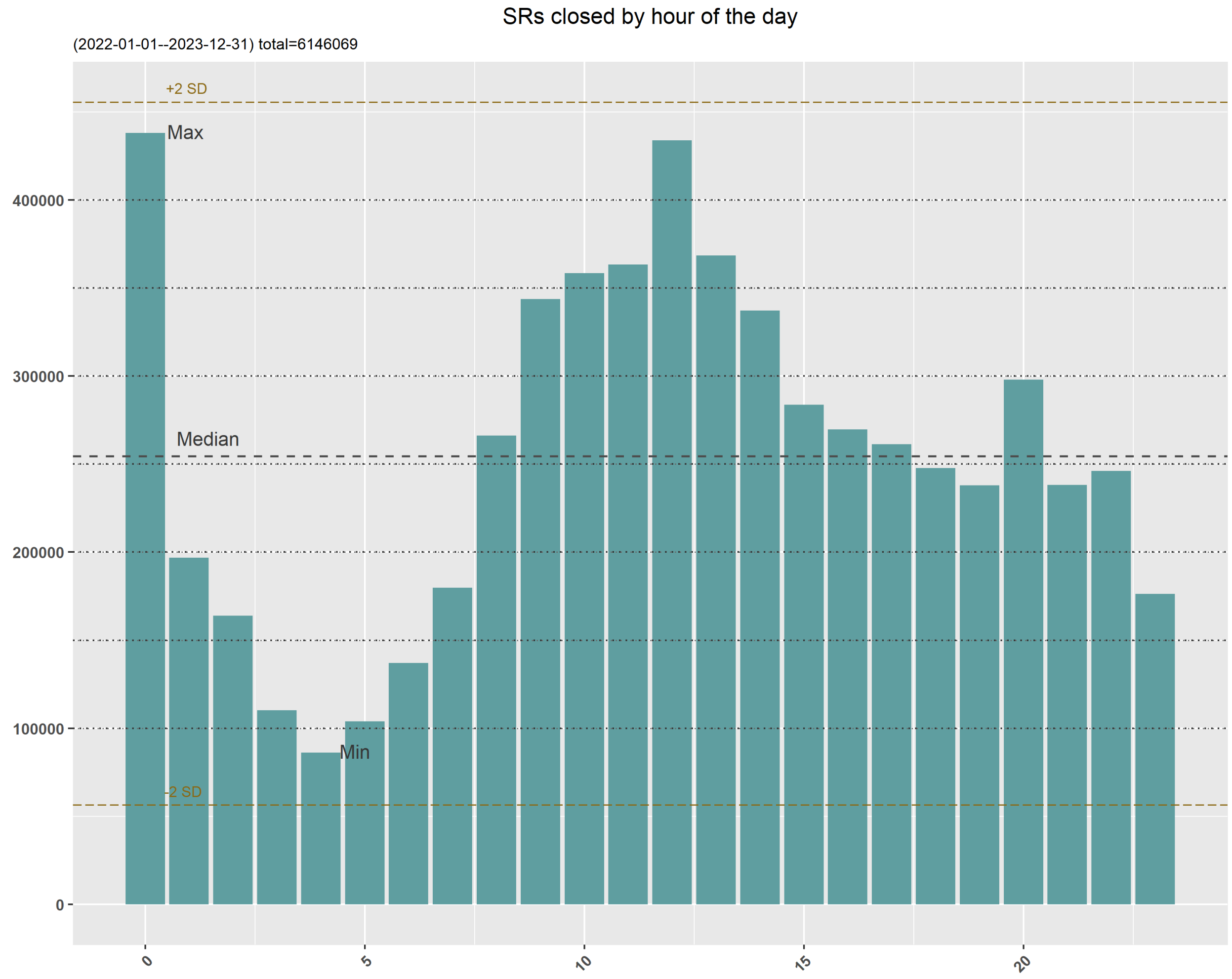


Figure 19: SRs by Hourly closed_date(s)

The created_date shows a maximum value at 12:00. The closed_date chart show significant spikes at both midnight and noon. The midnight (00:00) spike appears to be an anomaly especially as compared to the hours just before and after midnight where there is normally a pronounced slump in SR closures. And both the closed-at-noon and closed-at-midnight are approaching 2σ's from the hourly average.

Let's drill down on each of these outlier points. First, we aggregate the created_date over the 2-year period that fall exactly on the top-of-the-hour with the minute and second values equal to 0, e.g. 09:00:00, 14:00:00. This visualization reveals a clearer picture of where these spikes are:

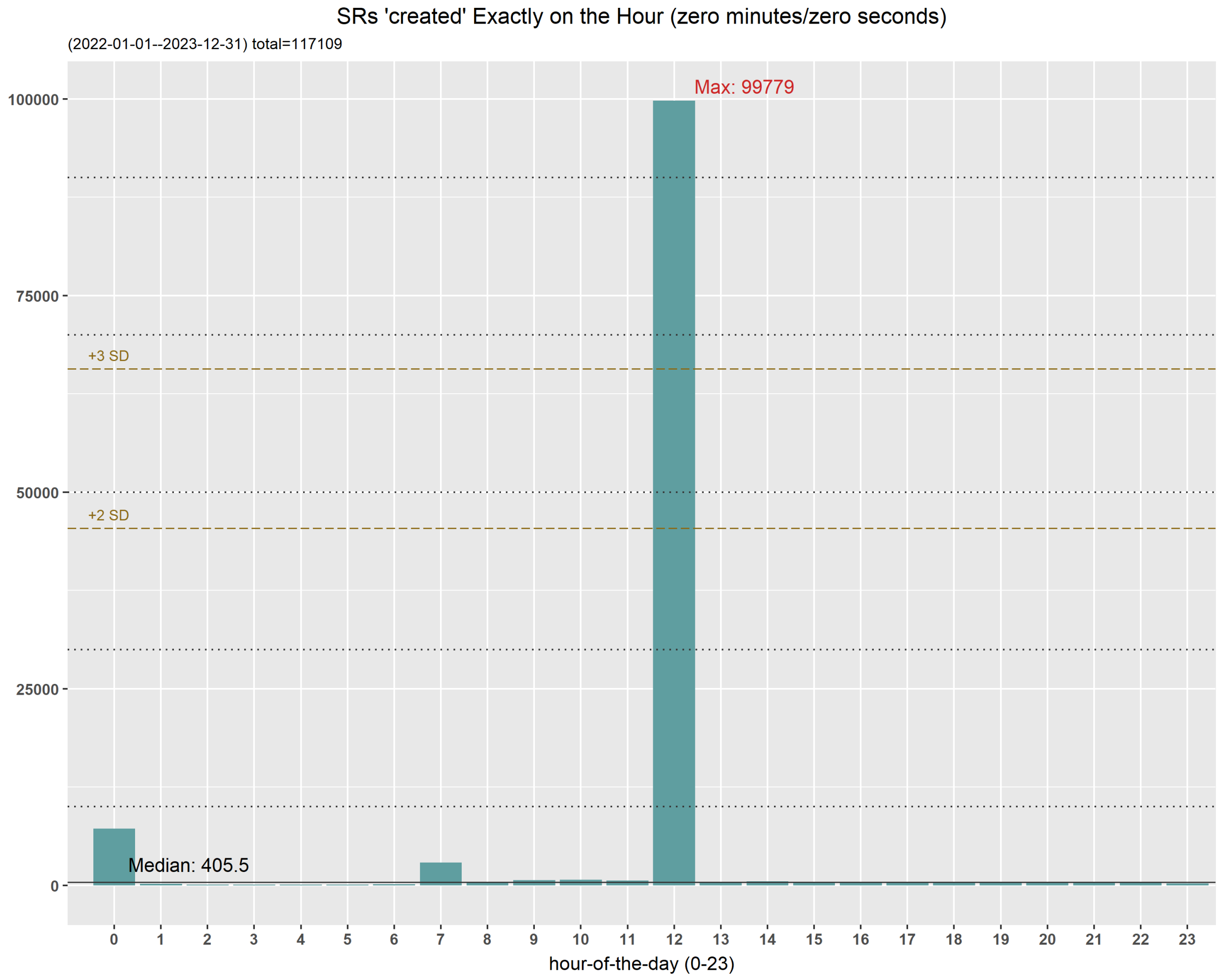


Figure 20: SRs Created at Top of the Hour

One other way to visualize this anomaly is to look at the busiest day during this 2-year period (Friday, 2023-09-29). Here we will aggregated the created_date(s) by minute (with seconds equal to zero). This provides a minute-by-minute look at SR creation on the busiest day of this study. Here we see a clear spike at exactly noon (12:00), well beyond the 3σline, and well above any other hour of the day.

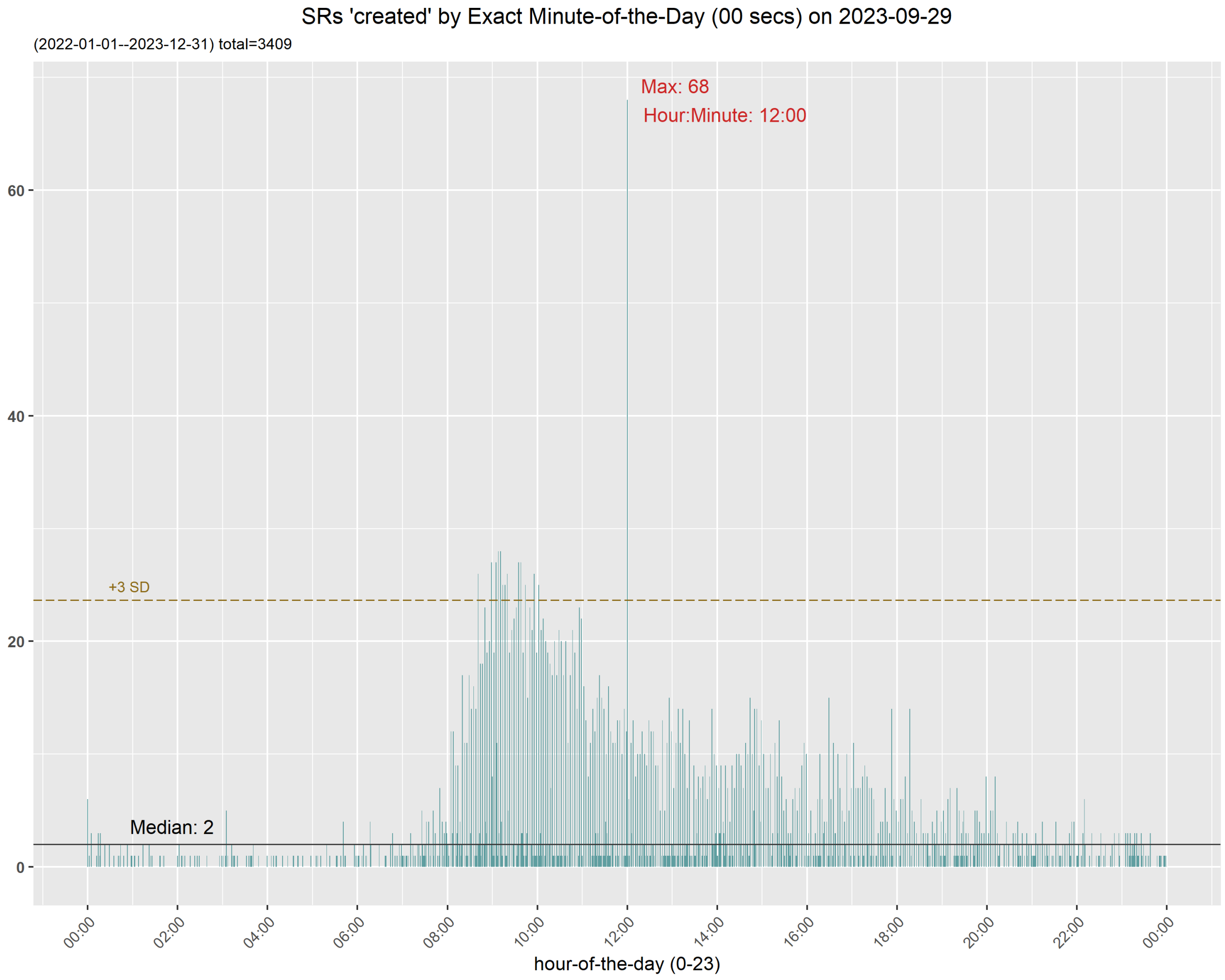


Figure 21: SRs created minute-by-minute on busiest day

Let's take a similar approach to the closed_date distribution over the 2-year period. We aggregate SRs by close_date to find the top-of-the-hour, where minute and second values equal to 0, e.g. 20:00:00. Here we see spikes at both midnight (00:00) and at noon (12:00), with the midnight spike being well beyond the 3σline. Also note how much lower (and nearly equal) all the other top-of-the-hour values are, relative to the midnight and noon spikes.

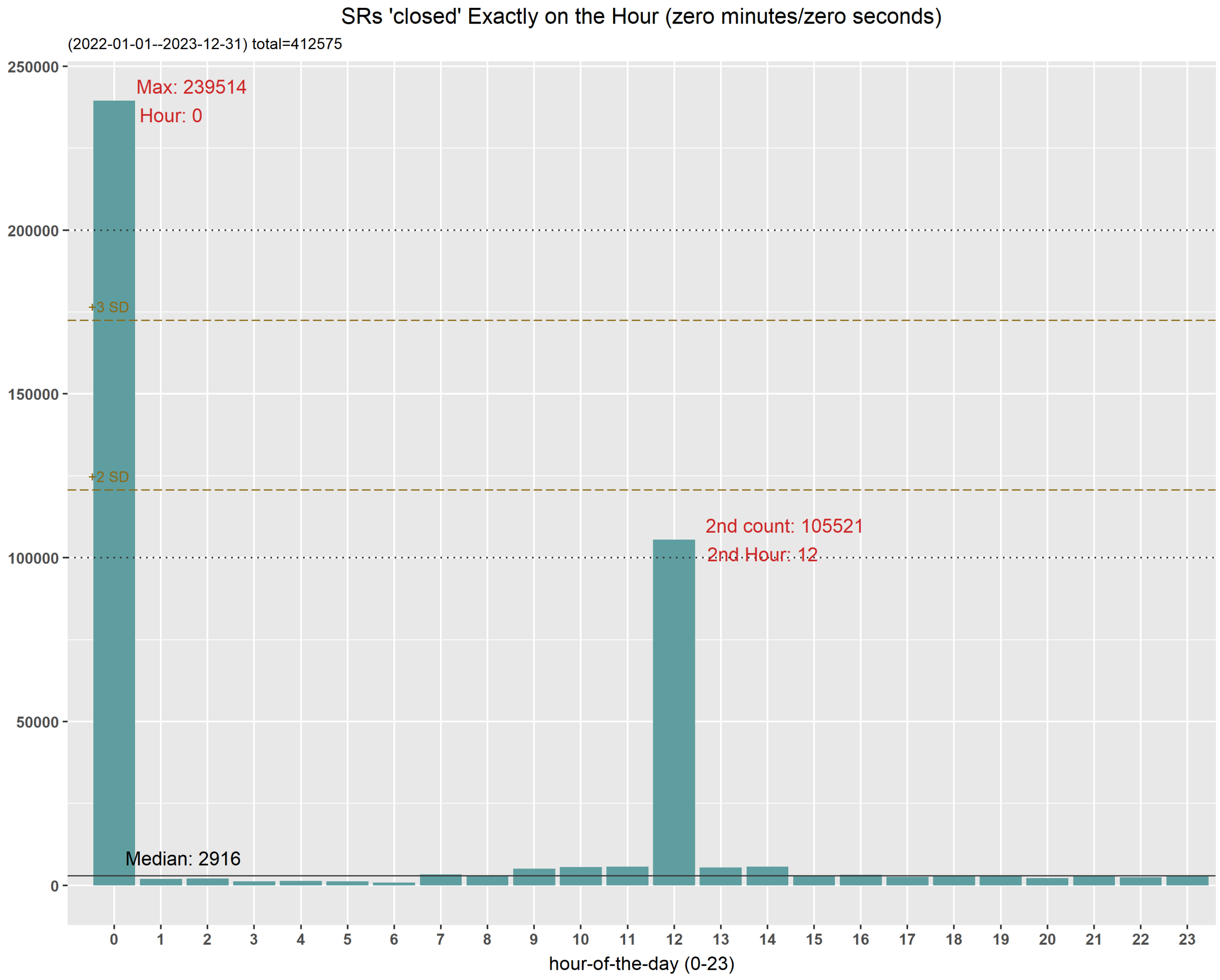


Figure 22: SRs Closed at Top of the Hour

As before, let's visualize this anomaly by looking at the closed_date(s) on the busiest day during this 2-year study period (Friday, 2023-09-29). We aggregate closed_date(s) by the minute (with seconds equal to zero). This gives a minute-by-minute view of SR closure on the busiest day of this study. Clearly visual is the very large closure spikes occurring at both midnight and noon. The chart highlights both of these anomalies with spikes of 286 and 120 at 00:00:00 and 12:00:00 respectively. Again, note how these spikes are so very much above the median values for all other minutes.

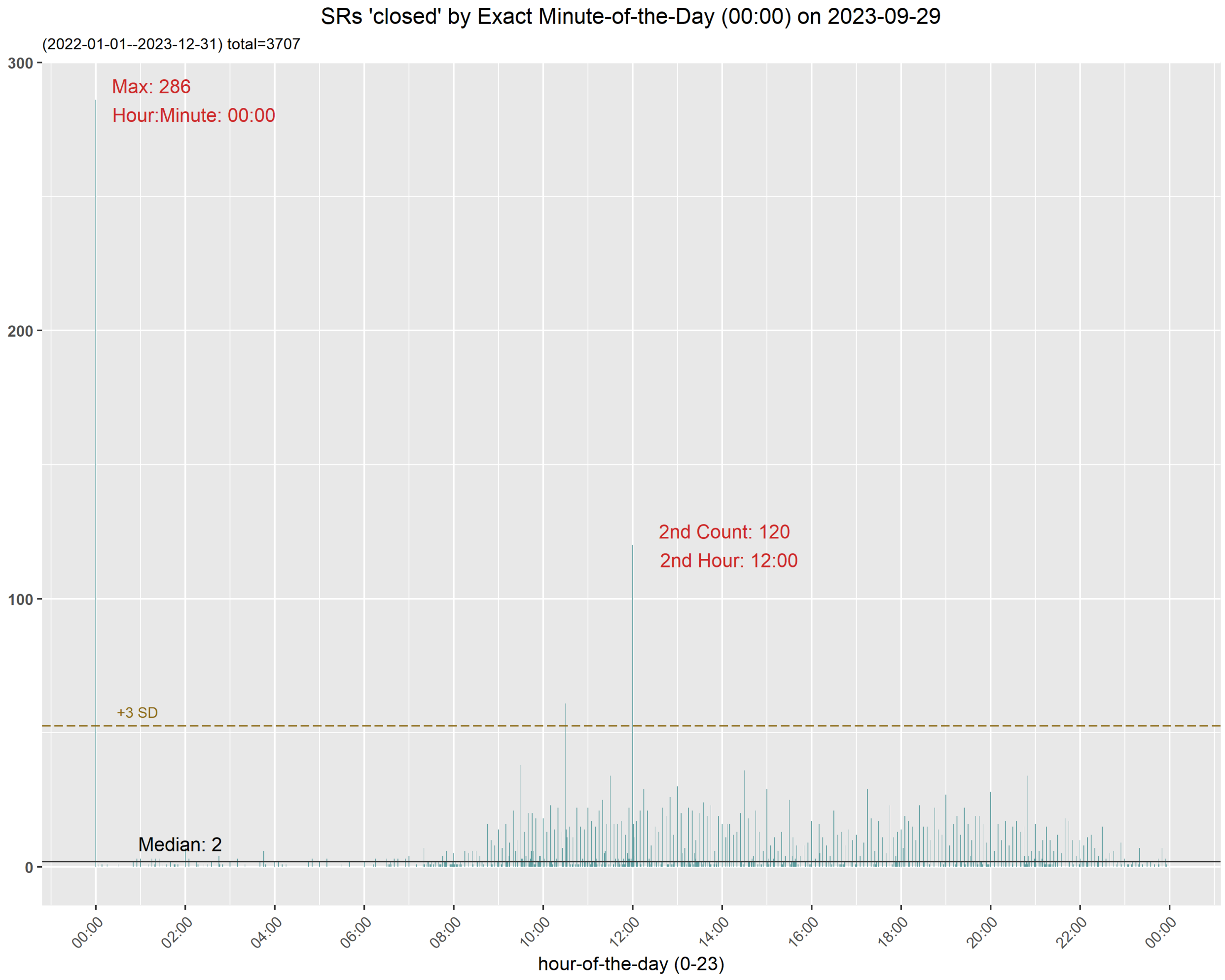


Figure 23: SRs closed minute-by-minute on busiest day

These unusual patterns of created & closed SRs at exactly the hours of midnight and noon likely indicates the presence of a bulk create/close software process; a process that perhaps automatically "closes" (or "creates") a large number of SRs with a provided time-stamp of midnight (00:00:00) or noon (12:00:00), and does so in bulk. If so, then addressing this issue likely requires a closer look at those Agencies providing these suspect SR closed/created times. The desirable state is that the created_date and closed_date fields correctly reflect the true opening and closing of the SRs, such that an accurate "duration" can be computed for the SR life-cycle. Otherwise, we are likely seeing durations that are artificially manipulated by a software process performing bulk uploads to the 311 system.

Looking at the distribution by Agency for the “closed-exactly-at-midnight” shows that >90% of these suspect SRs come from just two Agencies - Dept. of Buildings (DOB) and Dept. of Sanitation (DSNY). If you include HPD and DOT, that accounts for 99% of the suspect SR closures. This Agency distribution does not follow the overall SR distribution and we can assume that it is Agency specific.

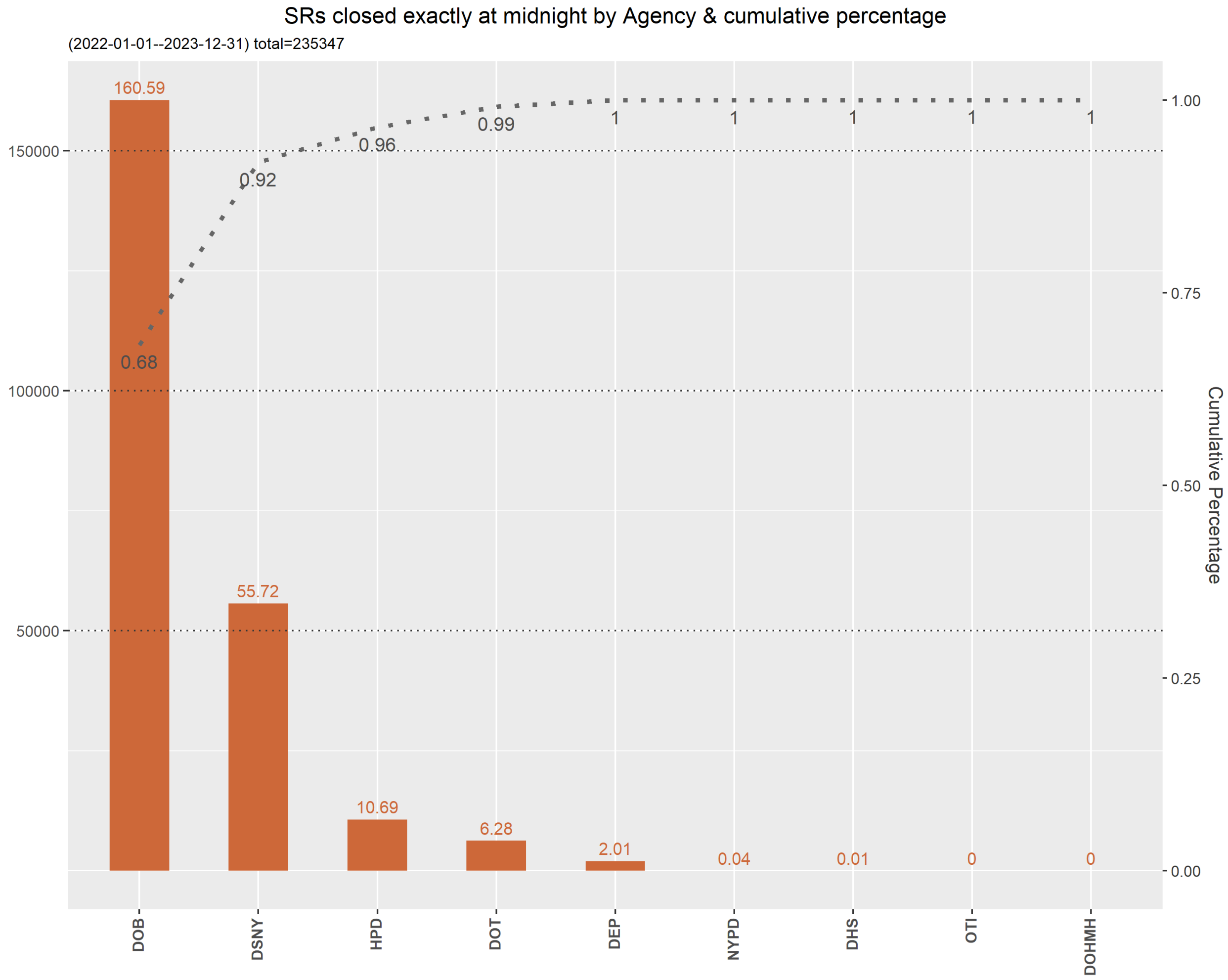


Figure 24: SRs Closed Exactly at Midnight by Agency

However, the “SR closed at noon” visualization shows that a single Agency, DSNY, is responsible for >99% of the SRs “closed-exactly-at-noon”. Accordingly, any effort to resolve this issue necessitates working with DSNY to fully understand how this is occurring.

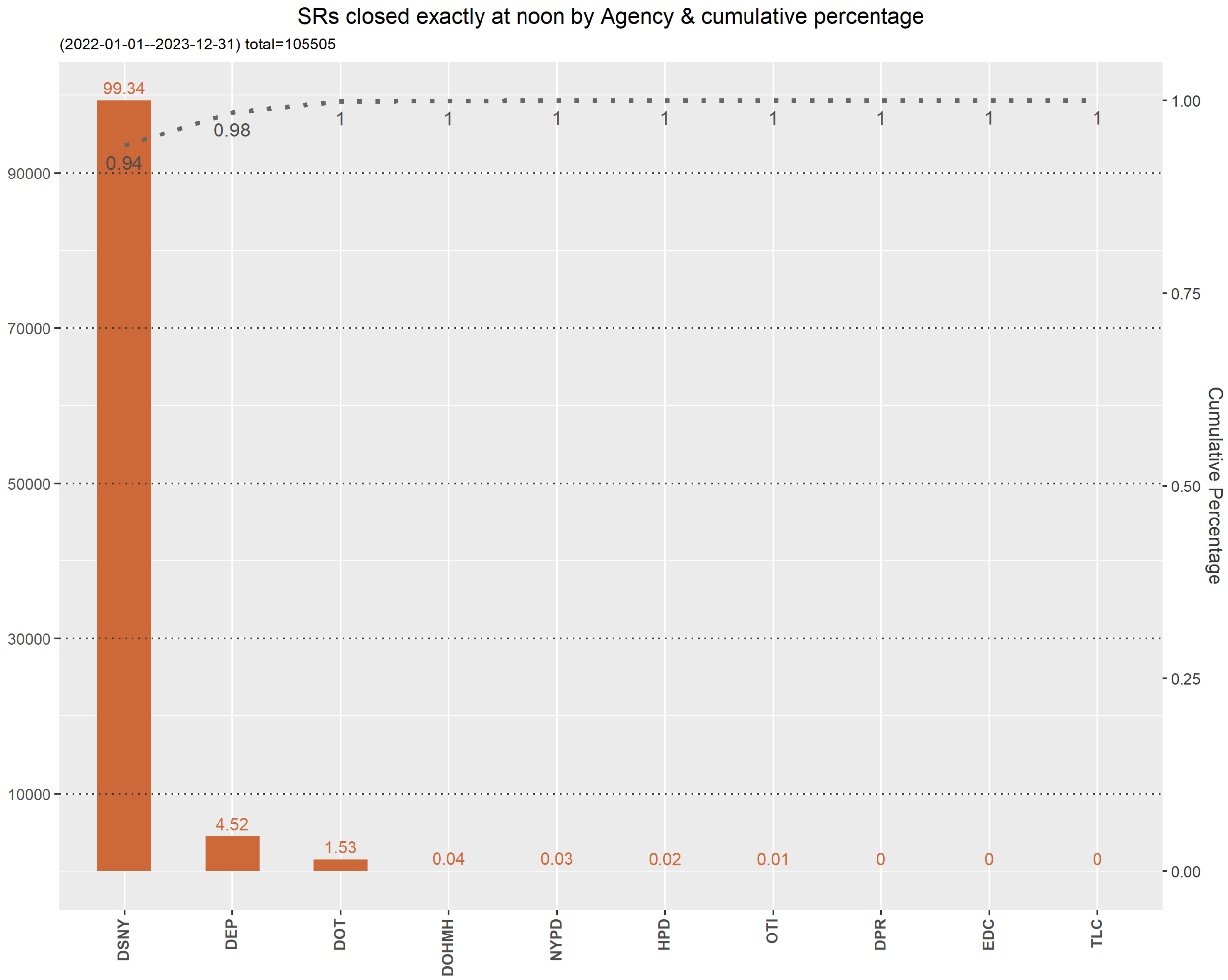


Figure 25: SRs Closed Exactly at Noon by Agency

The 99,779 SRs created-exactly-at-noon also appears to originate with the Dept of Sanitation (DSNY) which is responsible for >99% of the suspect SRs.

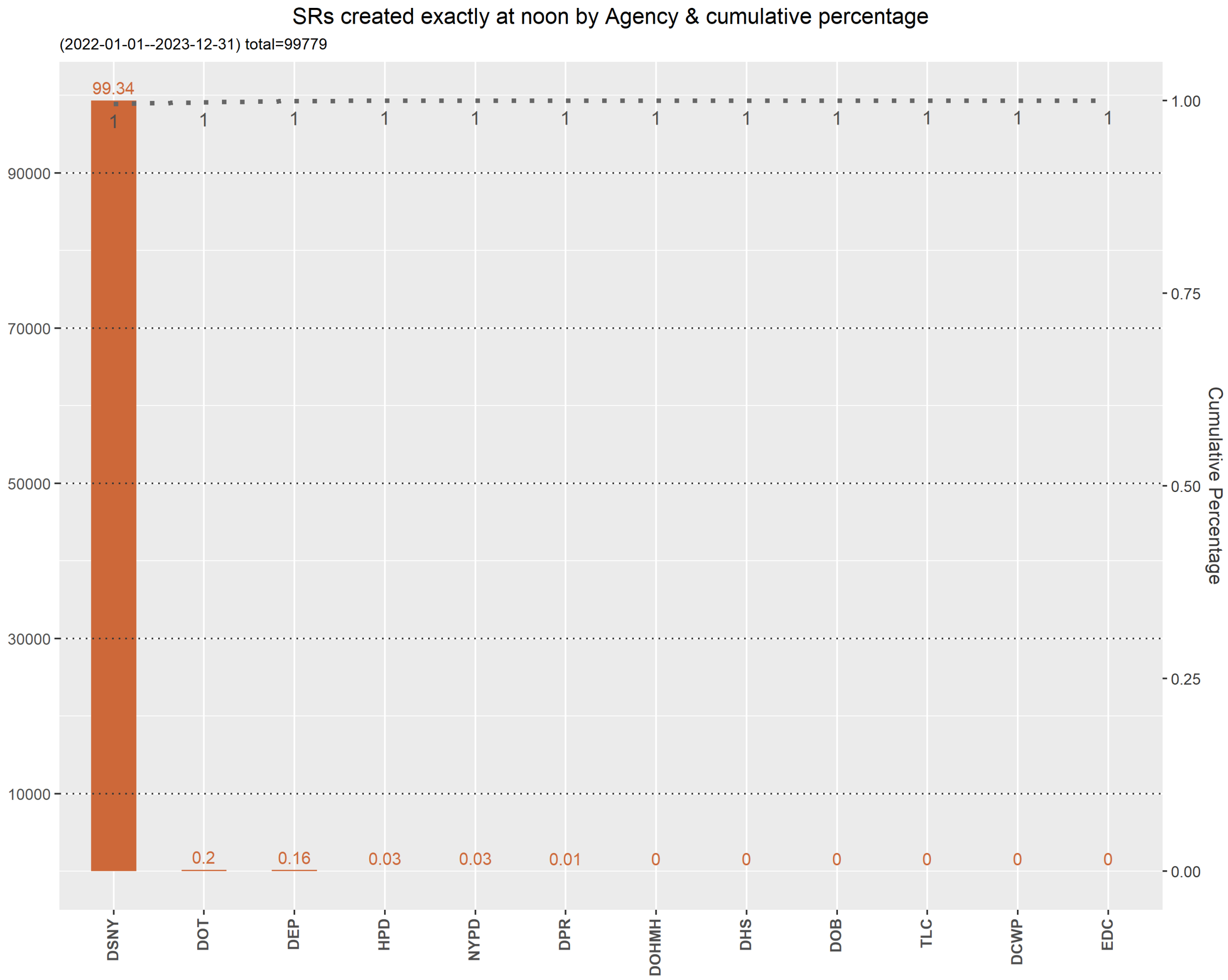


Figure 26: SRs Created Exactly at Noon by Agency

While it is difficult to ultimately know if these observations are indeed anomalies; they could be the result of normal, expected 311 SR activity. But at least by observation that outcome appears very unlikely.

8.8 resolution_action_update_date

When an SR is updated, by any means on any field, the software automatically populates the resolution_action_update_date. Analyzing that field it is apparent that some of the SR updates are happening long, long after the SR is closed. In this dataset there are eight SRs that have resolution_action_update_date values of 44,601 days after the closed_date. The problem lies not with the resolution_action_update_date, but rather with the closed_date values for these 8 SRs, all of which are in the format of “1900-01-01”. It has previously been discussed that these inaccurate dates may possibly be coming from an Excel spreadsheet as that is the minimum date which Excel can handle. Here is a sample of those SRs:

Agency	Closed Date	Resolution Action Updated Date	Post Closed Update Duration (days)
DHS	1900-01-01	2022-02-12 00:58:43	44602
DHS	1900-01-01	2022-02-11 14:28:59	44601
DHS	1900-01-01	2022-02-11 14:17:41	44601
DHS	1900-01-01	2022-02-11 13:45:59	44601
DHS	1900-01-01	2022-02-11 13:22:12	44601

Table 9: Extremely long Resolution Updates

These obvious inaccurate date values are removed from the analysis. But even so, there are some very late updates to SRs, with some updates occurring almost two years after the SR is closed. There are a total of 7460 SRs that are updated >30 days and <730 days after the closed_date. The median of these late (<30 days) post-closed resolution_action_update_date(s) is 84 days. However, the mean resolution_action_update_date for all of the SRs is only 0.39 days (9.3 hours) after the SR is closed. It is not known if this is normal behavior or an area that may require further investigation.

Agency	Closed Date	Resolution Action Updated Date	Post Closed Update Duration (days)
TLC	2023-05-25 14:21:12	2023-10-19 10:57:02	147
TLC	2023-07-12 11:03:46	2023-12-07 09:00:05	148
TLC	2023-08-03 10:34:23	2024-01-29 10:00:37	179
TLC	2022-04-04 10:54:39	2022-06-02 12:14:11	59
TLC	2022-12-05 15:23:45	2023-02-02 09:06:45	59

Table 10: Long post-closed Resolution Updates

Here is a look at the spread of the post-closed resolution_action_update_date values those

SRs that are >30 days and <730 days. There are some significant outliers

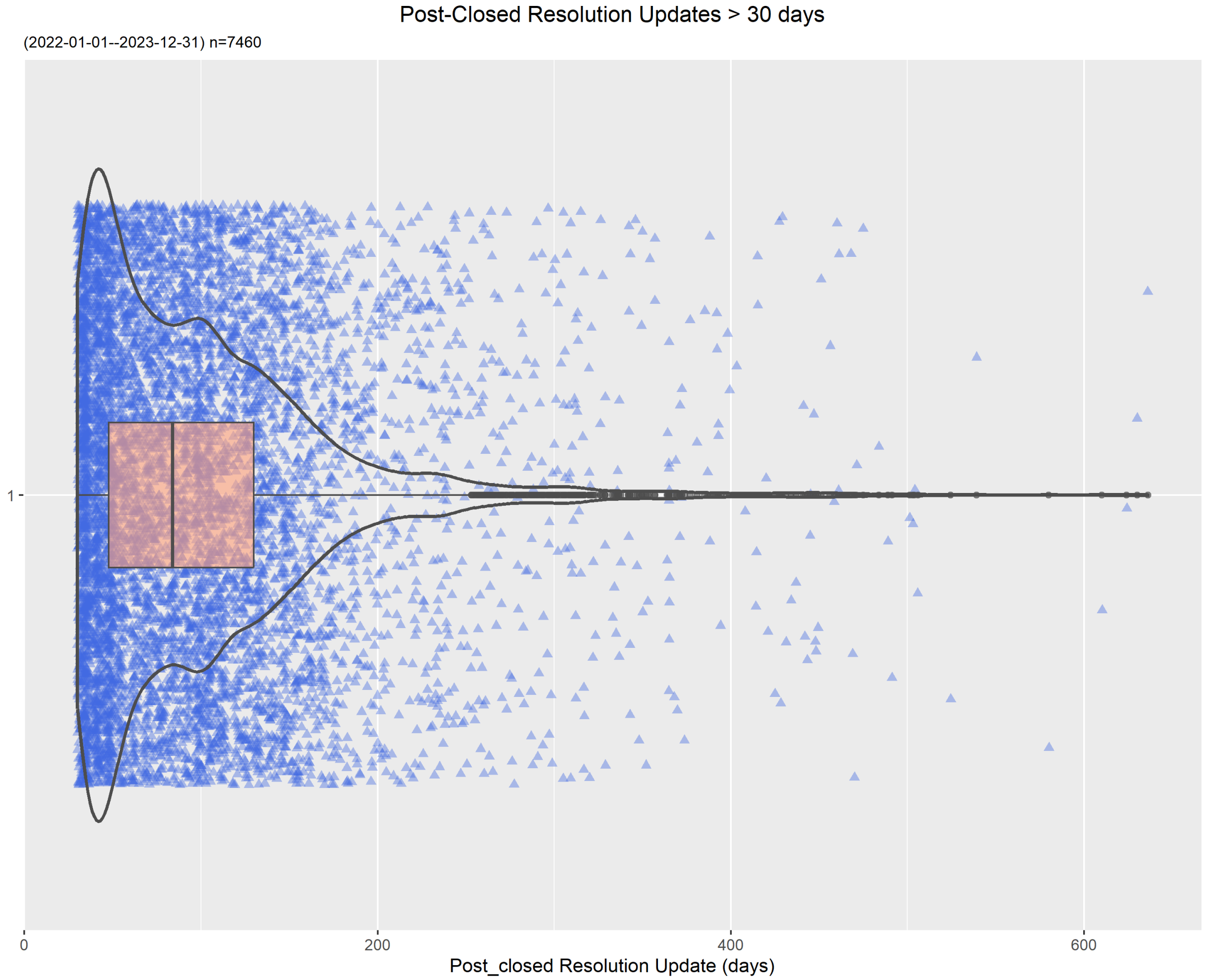


Figure 27: Post-Closed resolution_action_update_date ¿30 days

If we look at a breakout of these (late?) update dates by Agency, we can see that the vast majority li with two Agencies: Taxi & Limo Commission (TLC) and Dept of Sanitation (DSNY) which account for 96% of the SRs in question. Is there a valid reason that TLC and DSNY would be updating an SR over month (or months) after the SR has been closed? Perhaps, but it seems to merit closer investigation.

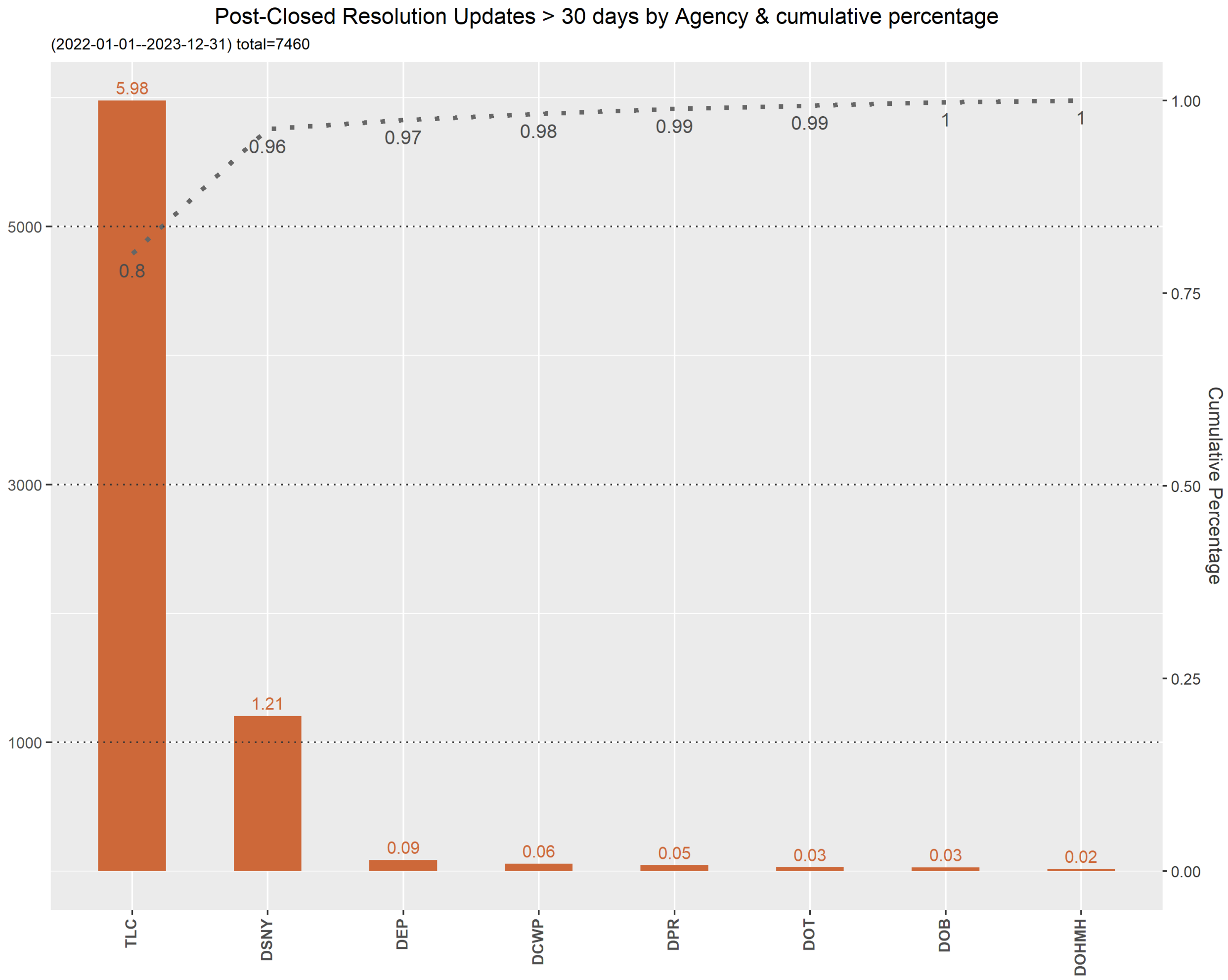


Figure 28: Post-Closed resolution_action_update_date ¿30 days by Agency

9 Accuracy and precision

There is one area in particular where the question of precision vs. accuracy immediately arises, and that is with the Latitude and Longitude fields. Both the Longitude and Latitude fields are expresses as a 14-decimal number, e.g. 40.86769186022511 (also the Location field which is a straight concatenation of latitude and longitude). Given that 1 degree of latitude at the equator is equal to 111.044736 kilometers, the "1" at the end of that number represents approximate 1.1104 nanometer or 1/1,000,000,000 of a meter. For reference a DNA molecule is approximately 2nm in width.

Clearly the representation of the Latitude and Longitude fields are a classic case of 14-digit precision, but limited accuracy. It is more likely that the Lat/Long values are accurate to the 5th decimal place, about 3.64 feet; even that might be overstating things.

10 Redundant & Duplicate fields

During this analysis, several redundant fields were observed. Listed below is a discussion for each of a pair of data fields that are redundant or near-redundant. These fields should be examined further for possible consolidation in the overall dataset.

10.1 latitude & longitude and the location fields

The location field is a pure concatenation of the latitude and longitude fields, with a comma an parenthesis added. Example:

- latitude: 40.768456429488
- longitude: -73.9575661888774
- location: (40.768456429488, -73.95756618887745)

The inclusion of the location field seems questionable as the data is arguably more difficult to extract than the two specific fields, especially for software programs.

10.2 borough and park_borough fields

These two fields are 100% matches; fully redundant.

10.3 borough and borough_boundaries fields

The borough field is a text field for the five New York City boroughs (Manhattan, Queens, Brooklyn, etc.). The borough_boundary field is numeric with values from 1-5 corresponding to the five boroughs (Staten Island, Brooklyn, Queens, Manhattan, Bronx). When the borough_boundary field is translated and compared to the borough field, there is a 98.3% match. Given the size of the dataset, that still leaves 110,715 non-matching values, not an insignificant number.

And that, in a nutshell, is the problem with two "near-duplicate" fields. It is like the story of a man wearing two watches; he's never quite sure what time it is. And with "near-duplicate" fields, which field is correct when they disagree? If one field is blank, can you use the other one? Which field can be used as the reference value. And as we will see, there are several such examples in this dataset.

We can tell where the non-matches occur. It is mostly in DOT, NYPD, and DSNY which represent 90% of the non-matching values.

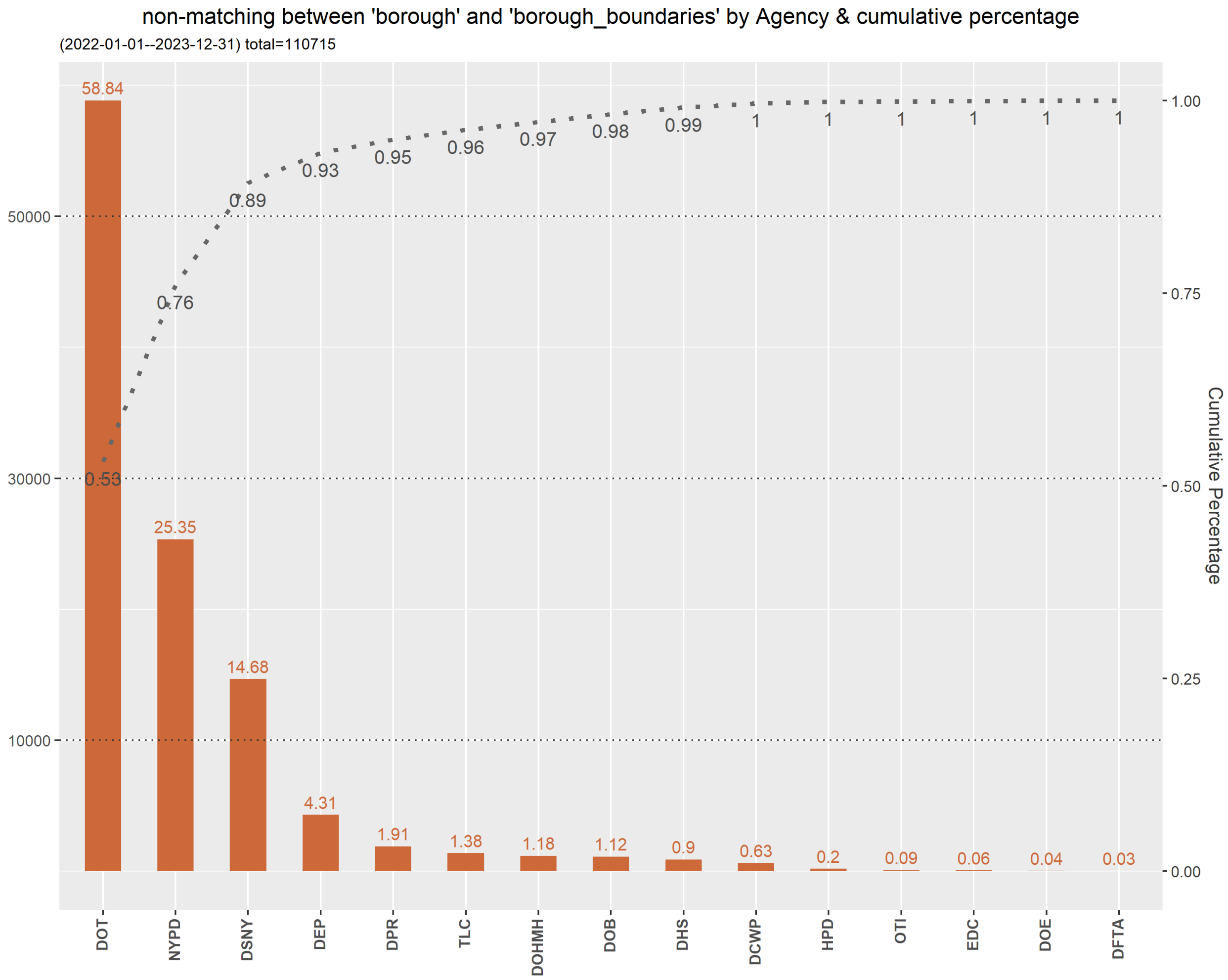


Figure 29: Non-matching borough and borough_boundary by Agency

10.4 borough and taxi_company_borough fields

The two fields borough and taxi_company_borough might appear to be related. In fact, despite the names, the two data fields are nearly completely different with only 0.05% matching. The mismatch is so high that it suggests the two fields are used very differently. This field is used exclusively by the Taxi & Limo Commission which governs taxis and other cars for hire. There are a total of 3567 non-blank entries in the taxi_company_borough field, with 99.94% blank.

10.5 incident_zip and zip_codes fields

The zip_code field is one of the six computed fields and its validity has been explored in Section 8.1 and found to be lacking with 55.55% of the entries invalid according to the USPS, while the companion zipcode field incident_zip field had an accuracy rate of 99.93%. It is not surprising that only 0.58% of the two fields match (37,292 matches out of 6.4 million SRs.). While an Agency breakdown can be determined, given the questionable computation of the zip_code field, such an analysis would seem to provide any useful information.

10.6 police_precinct and police_precincts fields

Both the police_precinct and police_precincts are among the computed fields not cited in the Data Dictionary. Their usage however, is easy to discern. Are these two fields duplicates? Nearly so: 99.94% of the entries match. Since the source of the fields and the underlying computational process are not know, it remains a challenge to determine which field is the "right" field to use. (The validity of these fields is covered in Section 8.2.)

10.7 agency and agency_name fields

While not duplicates, these fields have a 1:1 correspondence. The Agency field contains abbreviations for the City agencies such as NYPD, DOT, HPD, DEP, etc. The agency_name field contains the full name of the various organizations: New York Police Department, Department of Transportation, Department of Housing Preservation and Development, Department of Environmental Protection, and so on. Given how well known the NYC Agency abbreviations are it seems redundant to include both in the dataset.

10.8 landmark and street_name

The landmark field is listed in the Data Dictionary as "Can refer to any noteworthy location, including but not limited to, parks, hospitals, airports, sports facilities, performance spaces, etc." We found that was not the case. To be sure many of the entries in the landmark field do contain landmark names, e.g."Pennsylvania Station", "La Guardia Airport", "Gramercy Park", the vast majority of the entries are street names, e.g. "Fenton Avenue", "Steinway Street", "Broadway". These entries appear to be very similar to those in the street_name field. And in fact, the values of street_name and landmark exactly match 62% of the time, and while this is not a full duplicate, it is a precentage of matching that is certainly indicative of duplicate usage. Even the non-matches (excluding blanks) would appear to be matches except for minor spelling and nomenclature changes, e.g. "NINTH AVE" & "9 AVE".

street_name	**landmark**
MACDOUGAL ST	MAC DOUGAL ST
NINTH AVE	9 AVE
NORTH SIXTH ST	NORTH 6 ST
SOUTH FOURTH ST	SOUTH 4 ST
SIXTH AVE	6 AVE
THIRD AVE	3 AVE
EAST FIRST ST	EAST 1 ST
FOURTH AVE	4 AVE
SAINT LAWRENCE AVE	ST LAWRENCE AVE
SOUTH FIRST ST	SOUTH 1 ST
BRIGHTON SEVENTH ST	BRIGHTON 7 ST
WEST FIFTH ST	WEST 5 ST
MT HOPE PL	MOUNT HOPE PL
PENN STA	PENNSYLVANIA STA

Table 11: Non-matches between 'street_name' and 'landmark' fields

The distribution by Agency for the non-matches somewhat follows the overall SR distribution (NYPD, DPR, DSNY), but with some notable exceptions such as Taxi & Limo Commission, Dept of Homeless Services (DHS), and Dept of Health & Mental Hygiene (DOHMH).

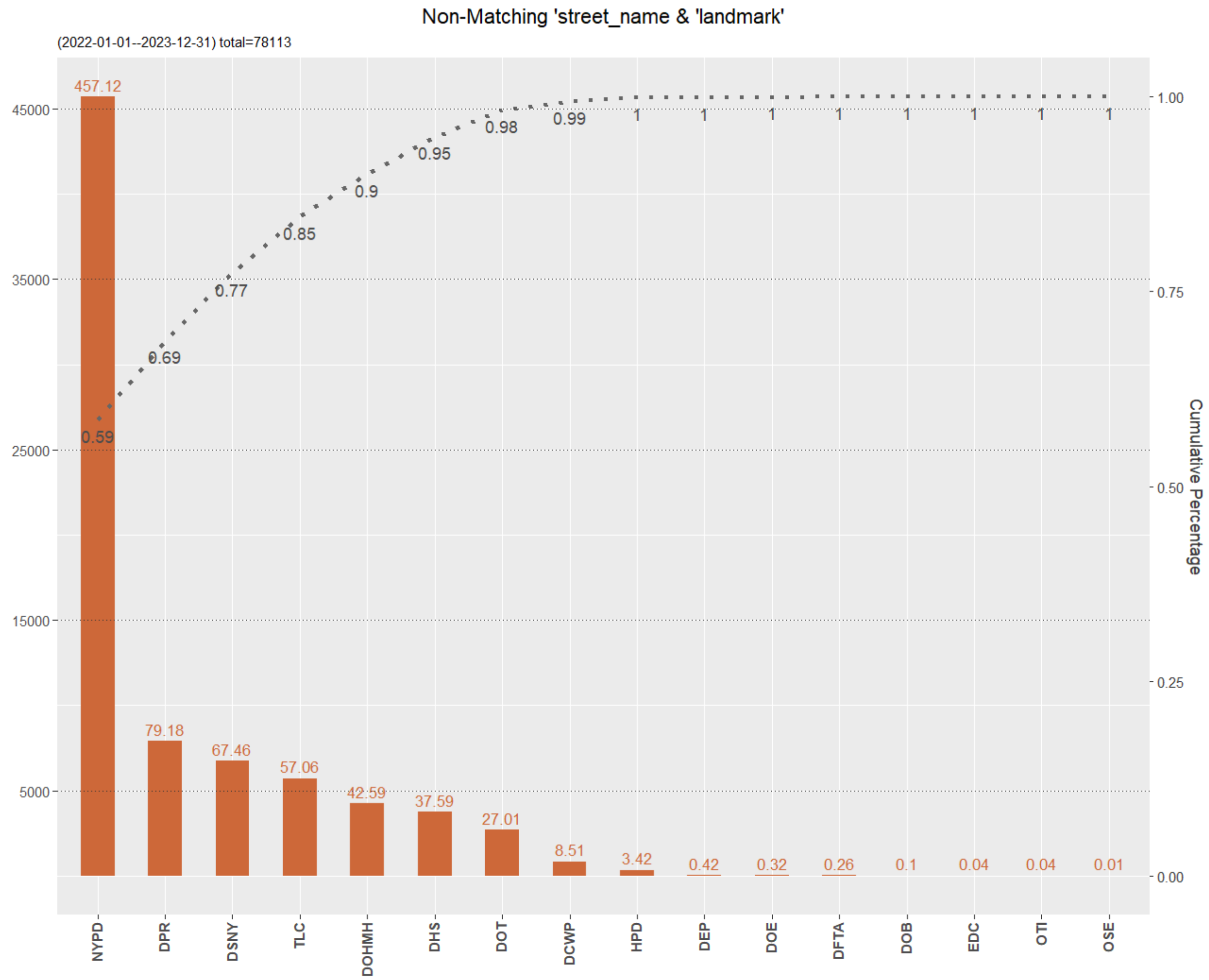


Figure 30: Non-matching street_name & landmark by Agency

10.9 cross_street_1 & intersection_street_1 and cross_street_2 & intersection_street_2

There are two sets of street pairs in the dataset in addition to the complaint address:

- cross_street_1
- cross_street_2
- intersection_street_1
- intersection_street_2

These two pairs of streets are used to help identify the location of the reported incident, as it is common in New York City to provide the street address and the cross street, as in "24 W 90th street between Columbus and Amsterdam". The Data Dictionary states that these streets are "based on the geo validated incident location". It then adds "If the service request pertains to a location in the middle of a block, cross street will be provided by geocoding...If the service request pertains to a street corner, intersection will be provided by geocoding." One might think that the cross_street(s) would apply for "middle of a block" incidents, while the intersection_street(s) would apply to those incidents that happen on a street corner; that those two sets of fields would be exclusive to one another; and either/or situation.

Unfortunately, that is not the case. For the majority of data rows, both pair of fields are populated and are duplicates. In fact, 88% of the time, cross_street_1 exactly matches intersection_street_1 and cross_street_2 exactly matches intersection_street_2, Only 11.9% are non-matches.

(Note: Address standardization was applied to the street addresses using the R campfin *package. This was done so as to prevent the software from generating a non-match between streets such as "240 E 69th ST" and "240 E 69th Street" which are clearly the same address, just slight formatting and spelling differences.)*

The problem occurs when the two pairs of streets do not match, such as when one is blank and the other is not. Which street is correct? If one is blank, say cross_street_1, but intersection_street_1 is not blank, should you use the non-blank one? Let's look at some examples:

Matching cross_street_1 and intersection_street_1		
cross_street_1	intersection_street_1	agency
FORT HAMILTON PKWY	FORT HAMILTON PKWY	NYPD
87 ST	87 ST	NYPD
EAST 169 ST	EAST 169 ST	NYPD
EAST 74 ST	EAST 74 ST	DOT
CHERRY AVE	CHERRY AVE	NYPD
Non-matching cross_street_1 and intersection_street_1		
cross_street_1	intersection_street_1	agency
KAPPOCK ST	NETHERLAND AVE	DOT
GREENE AVE	CLINTON AVE	DOT
LEWIS AVE	FULTON ST	DOT
GERARD AVE	EAST 161 ST	DOT
8 AVE	WEST 136 ST	DOT

Table 12: Matching/Non-Matching cross_street_1 & intersection_street_1

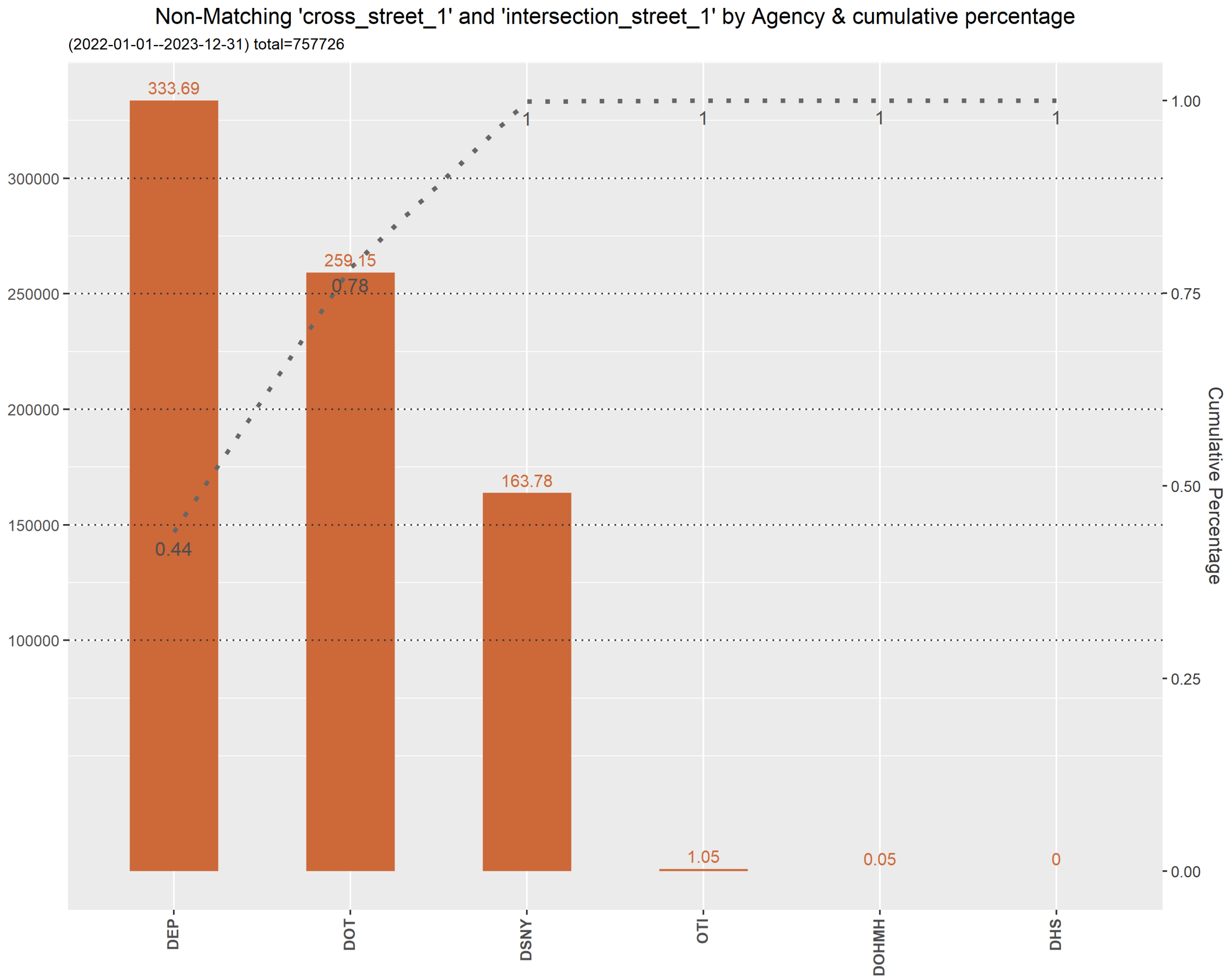


Figure 31: Non-matching X_Street_1 & Intersection_Street_1

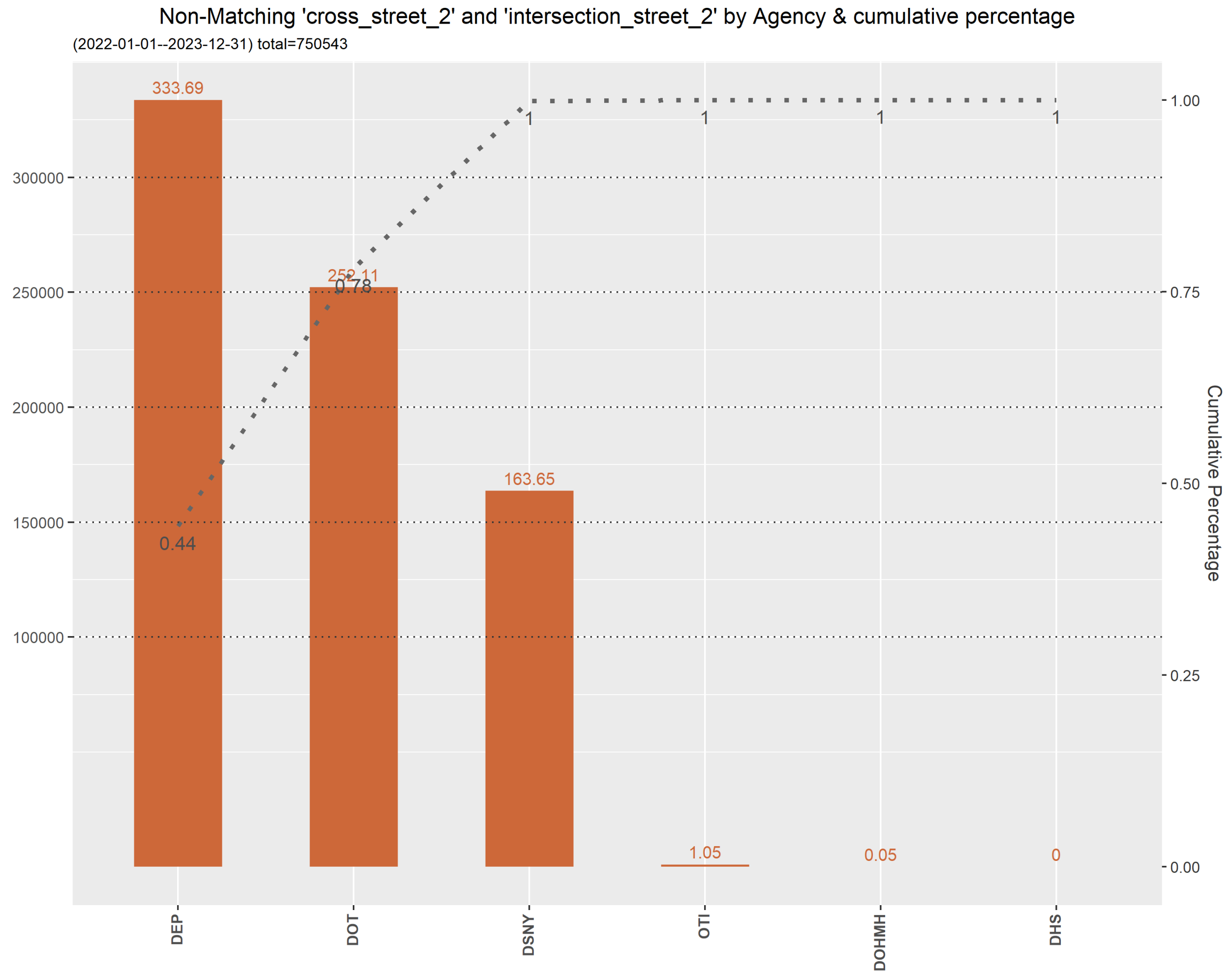


Figure 32: Non-matching X_Street_2 & Intersection_Street_2

In addition to the matching/non-matching values, we encountered a number of what you might call *near-matches*. There were numerous cases where the cross_street and intersection_street would be close to identical, e.g. HARMON DR and HARMON RD. To identify these near-matches we applied a common method of determining matches by measuring the Hamming Distance between the two fields. The Hamming Distance is a measure of how many letters have to be changed in order for the two fields to match. For this analysis, we chose a Hamming Distance of 2. Here are the results for both cross_street and intersection_street.

cross_street_1	**intersection_street_1**	**agency**	**hamming_distance**
WEST 168 ST	WEST 167 ST	DOT	1
115 AVE	120 AVE	DOT	2
105 AVE	107 AVE	DOT	1
71 ST	80 ST	DOT	2
145 ST	150 ST	DOT	2
138 ST	139 ST	DOT	1
19 AVE	20 AVE	DOT	2
25 AVE	30 AVE	DOT	2
76 ST	79 ST	DOT	1
67 ST	68 ST	DOT	1

Table 13: Near-matches for cross_street_1, intersection_street_1

category	**count**	**percentage**
Matching	5,637,182	88.15
Both blank – Matching	1,624,499	25.4
Non-matching	757,726	11.85
cross_street_1_blank	222,232	N/A
intersection_street_1_blank	519,348	N/A
Near-match	128	0.002002

Table 14: Summary of cross_street_1 and intersection_street_1

category	**count**	**percentage**
Matching – non-blank	4,012,683	62.75
Matching – both blank	1,623,276	25.38
Non-matching	750,543	11.74
cross_street_2_blank	222,788	N/A
intersection_street_2_blank	517,431	N/A
Near-match	1,500	0.023456

Table 15: Summary of cross_street_2 and intersection_street_2

category	**count**	**percentage**
Matching – non-blank	4,021,089	62.88
Matching – both blank	1,623,276	25.38
Non-matching	750,543	11.74
cross_street_2_blank	222,788	N/A
intersection_street_2_blank	517,431	N/A
Near-match	1,500	0.023456

Table 16: Summary of cross_street_2 and intersection_street_2

cross_street_2	intersection_street_2	agency	hamming_distance
19 AVE	18 AVE	DOT	1
227 ST	225 ST	DOT	1
65 ST	65 PL	DOT	2
18 AVE	17 AVE	DOT	1
BEACH 110 ST	BEACH 109 ST	DOT	2
3 AVE	2 AVE	DOT	1
88 ST	87 ST	DOT	1
WEST 114 ST	WEST 113 ST	DOT	1
192 ST	189 ST	DOT	2
5 AVE	4 AVE	DOT	1

Table 17: Sample of near-matching cross_street_2 and intersection_street_2 (both non-blank)

10.10 Shrinking file size by removing duplicates

By removing duplicate and "near-duplicate" fields, it is possible to shrink the file size by 12.3% which for this dataset equates to a reduction of 395 Mb. A smaller file size means faster downloads, less storage impact, as well as simplifying data analysis efforts.

Below is a list of duplicate and near-duplicate fields. There are some challenges with these proposed deletions, mostly with the "near duplicate" fields. So while the park_borough is a complete duplicate of the borough field, the borough_boundary field is only 98.3% duplicate. Admittedly, loosing 1.7% of the data in those two fields is quite small, for this large dataset that amounts to 110, 715 non-matching occurrences. Similarly with police_precincts and police_precincts where there is a 99.9% match; close to 100% but not a complete match.

Ultimately, it will require a deeper dive into the relationship between these duplicate and near-duplicate fields to determine if they can be removed or not. Pending that, the authors would propose deleting the following fields from the overall data set, which explanation provided.

- agency_name: Each row of data contains the agency field, which is an abbreviation of the Agency's name. The agency_name abbreviations are clear, simple, and well understood. The authors believe the agency_name field could be eliminated without loss of data quality.

- park_borough: This field is a 100% match with the borough field. Removal of this field will not result in any loss of data or data quality.

- location: The location field is simply a concatenation of the latitude and longitude fields, with a comma and parenthesis added. The authors feel this field actually hinders data analysis as the field must be re-split into the two components (latitude and longitude) to conduct geographic analysis. The data is a 100% duplication of the latitude and longitude fields. Removal of this field will not result in any loss of data.

- police_precinct: This field has a 99.95% match with the police_precincts fields. We are unable to determine which field is more correct, but feel that removal of one of the precinct fields is prudent and would result in very minimal data loss. (We should note that both police_precinct and police_precincts fields each have 34.5% invalid data.)

- borough_boundaries (computed field): This field is one of the six computed fields. It has a 98.3 match with the borough field. We recommend deleting this fields despite the somewhat limited loss of data.

- cross_street_1 & 2 and intersection_street_1 & 2: These two pairs of fields each have an 88% match. We would recommend deleting the two intersection_street fields while acknowledging some loss of data in doing so.

- zip_codes(computed field): This field is one of the six computed fields. However, this field has error rate of 58% and as shown in the **??** can lead to dangerous mistakes when used for analytical purposes. We would recommend deleting this field.

Additionally, it would be worth investigating if certain data fields are in-fact useful, as the population of these fields is very, very scarce. For example, the taxi_company_borough field is 99.94% blank, and the remaining entries match the borough field only 0.05% field. A similar situation exists for the road_ramp field (99.75% blank), the vehicle_type field (99.71%blank), and possibly due_date which is 99.62% blank. There are several others: bridge_highway_direction, bridge_highway_name, bridge_highway_segment, and taxi_pick_up_location...all of which have >99% blanks. Most of those fields appear to be almost exclusively used by the Taxi & Limo Commission (TLC), which should be consulted to determine if those fields are used and properly populated.

11 Data Dictionary Observations

During this analysis, it became apparent that the published Data Dictionary could use an update. We found many discrepancies between the Data Dictionary and the actual data. While not actually, incorrect or dirty data, it nonetheless can lead analytic efforts astray. In data analysis, it's essential to accurately describe the nature and purpose of each field to ensure proper handling and interpretation. Here are some of the noted discrepancies:

- While the Open Data portal lists the basic data types of the various fields, the Data Dictionary does not. Nor does it provide meaningful information on the specifics of various fields, such as constraints or domain of legal values. For example, the incident_zip field is specified as "text". That's probably not the best definition, perhaps "categorical" would be more appropriate with a description indicating that the data can contain only numeric characters and is not subjected to arithmetic operations. We found two instances with the zip_code field contained character strings ("na", "N/A").

- Domain of legal values as contained in the Data Dictionary are incomplete. For example, the Data Dictionary indicates that the status field has values of assigned, canceled, closed, or pending. However, we found additional status values: in progress, started, and unspecified. Additionally, we discovered no SRs with a status of canceled. Similarly, the address_type field is indicated to have values address, blockface, intersection, latlong, and placename. We discovered additional values of bbl and unrecognized. No values of latlong were observed. Other fields suffer the same inaccuracies (facility_type, vehicle_type, taxi_pick_up_location, road_ramp, city).

12 Recommendations

Having identified a number of data cleanliness related issues with the 311 Service Request (SR) dataset, what follows are a series of recommendations as to how to resolve these issues. Below are those recommendations;

- First, and most importantly, the Open Data Team will need to undertake a special effort to identify these data anomalies, confirm the veracity of the issue, and engage with the appropriate Agency Open Data Representatives to resolve these issues. That effort will not be easy, fast, or cheap as in some software changes will likely be required. Perhaps hardest will be working with the external City Agencies that feed data to the 311 SR dataset, but are not using the core 311 system. Integrations and API usage will need to be investigated, and in many cases the changes required to correct an issue will be on the source Agency side. Coordination and mutual approval of any changes will be key. Strong support from senior leadership will be a critical success factor.

- Secondly, standards will be to be identified, agreed to, and commitment to support those standards must be obtained. For example, what are the domain of legal values for certain fields, especially those with identified issues?

- Obviously, many of the data integrity issues can be resolved through rule enforcement by software. For example, the software could easily prevent invalid values, such as zip codes, from entering the system. Similarly for an entire range of data issues. One key area is the created_date and closed_date which are frequently cited when determining responsiveness of various Agency services. The authors would argue that such dates should be driven by change in “status”, not manually entered. For example, the closed_date is captured when the SR “status” is changed to “closed”, and not manually entered. Such a change, along with many others, could prevent errors such nras negative and zero durations which significantly affect measurements of responsiveness.

- The large spikes in SR closure and creation occurring at midnight and noon, almost certainly point to an underlying automated process from an external-to-311 Agency. Such processes again distort the accurate capture of an SR duration. Data rule enforcement at the software integration or API level should be investigated.

- One possible way to have software enforce certain data standards, would be to move more non-311 Agencies to the core 311 software system. That system was completely upgraded in 2019, and should be (relatively) easy to expand to other Agencies and usages.

- As noted, there are six “computed” fields in the data export that are not identified in the Data Dictionary. While these fields may be “experimental”, they are nonetheless present in any data portal presentation or export. As noted in this study, a number of those computed fields suffer from data inaccuracies. It is difficult to know what to recommend for these six field without further understanding of their origin, computational method, or intended usage. We would recommend updating the Data Dictionary at a minimum and possibly hiding those fields from public display until such time as their usage and accuracy can be ascertained.

- Removal of duplicate fields should be examined. In some cases, such a the borough and park_borough fields, the duplication is 100In other cases, the duplication is in the 80-90% range. These fields and their usage/utility should be carefully examined to determine if their inclusion in the 311 SR dataset is meaningful, adds value not found in other fields, and should be removed or continued to be collected. Initial analysis indicates that as much as 12% size reductions are easily achievable with the removal of duplicate fields. A more thorough review and understanding of the fields would no doubt yield more savings.

- Fields with extremely low data presence, such as taxi_company_borough which is populated only 0.06% of the time. Similarly for several other fields evaluated in this study which have a high percentage of blank/unknown values. If the presence of a small percentage of SRs with these fields being populated is deemed useful and necessary, then so be it. But if the usage of these fields is so diminished, then perhaps its inclusion in the 311 SR dataset is not warranted.

- The Data Dictionary is in need of an upgrade and update. The last revision is indicated at 6 March 2023, well over a year ago. As noted in \nameref: **??** there are many problems with the Data Dictionary, including domain values, absence of the six "computed" fields, and inaccurate description of the usage of several fields. Some of the Data Dictionary entries can be undertaken immediately, which further revision should be addressed as part of a deeper data clean-up initiative.

- The excessive precision in the latitude, longitude, and location fields is simply nonsensical. We would recommend no more than four decimal places (0.000X) which represents a distance of approximately 36.4 feet. Five decimals (0.0000X) represents 3.6 feet, which although possible, would be an extremely sophisticated and accurate geo-coding system. A 14 decimal number is simply a case of a lot of precision with limited accuracy, and should be avoided.

- The 10 address-related address fields (incident_address, street_name, city, intersection_street_1 & 2, cross_street_1 & 2, landmark, road_ramp, bridge_highway_segment, and taxi_pick_up_location should be in standard USPS format. There are several software packages which can standardize addresses to these standards. We believe such an effort should be automatically applied to these fields at data entry. This step would also greatly improve any underlying geo-coding routines which may be used to determine data values such as BBL, latitude, longitude, and New York State X/Y-plane coordinates.

- The use of the value “UNKNOWN” , “N/A”, “NA”, and “na” when entered directly as text into a field should be very carefully examined, and in all likelihood prohibited. There are some indeed a few complaint_type(s) where an “N/A” value is appropriate, such as a consumer_complaint about certain non-geo centric issues, such as a travel agency or tax preparation service . But even in those cases a value of “N/A” should be entered only via a drop-down menu selection. During this analysis, there were far too many data representations of missing data including: ¡N/A¿, “na”, “ “, UNKNOWN, and “N/A”. It’s challenging for analyst to have to account for so the many different methods of representing empty or null data fields.

13 Protocol Suggestions

14 Discussion

References

Barns, S. (2016), “Mine your data: Open data, digital strategies and entrepreneurial governance by code,” *Urban Geography*, 37, 554–571.

Beheshti, A., Benatallah, B., Tabebordbar, A., Motahari-Nezhad, H. R., Barukh, M. C., and Nouri, R. (2019), “Datasynapse: A social data curation foundry,” *Distributed and Parallel Databases*, 37, 351–384.

Borgman, C. L. (2012), “The conundrum of sharing research data,” *Journal of the American Society for Information Science and Technology*, 63, 1059–1078.

Cantor, M. N., Chandras, R., and Pulgarin, C. (2018), "FACETS: using open data to measure community social determinants of health," *Journal of the American Medical Informatics Association*, 25, 419–422.

Gerte, R., Konduri, K. C., Ravishanker, N., Mondal, A., and 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, 30–39.

Hart, E. M., Barmby, P., LeBauer, D., Michonneau, F., Mount, S., Mulrooney, P., Poisot, T., Woo, K. H., Zimmerman, N. B., and Hollister, J. W. (2016), "Ten simple rules for digital data storage," .

McLure, M., Level, A. V., Cranston, C. L., Oehlerts, B., and Culbertson, M. (2014), "Data curation: A study of researcher practices and needs," *portal: Libraries and the Academy*, 14, 139–164.

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

Shankar, K., Jeng, W., Thomer, A., Weber, N., and Yoon, A. (2021), "Data curation as collective action during COVID-19," *Journal of the Association for Information Science and Technology*, 72, 280–284.

Wang, H.-J. and Lo, J. (2016), "Adoption of open government data among government agencies," *Government Information Quarterly*, 33, 80–88.

Witt, M., Carlson, J., Brandt, D. S., and Cragin, M. H. (2009), "Constructing data curation profiles," *International Journal of Digital Curation*, 4, 93–103.

Zuiderwijk, A. and Janssen, M. (2014), "Open data policies, their implementation and impact: A framework for comparison," *Government Information Quarterly*, 31, 17–29.