18 March 2024

311 SR Data
How clean is it?

Presented by:

David Tussey & Dr. Jun Yan (Uconn)

Visit open-data.nyc to view the full program.

First, you should clean the data. It’s probably dirty.

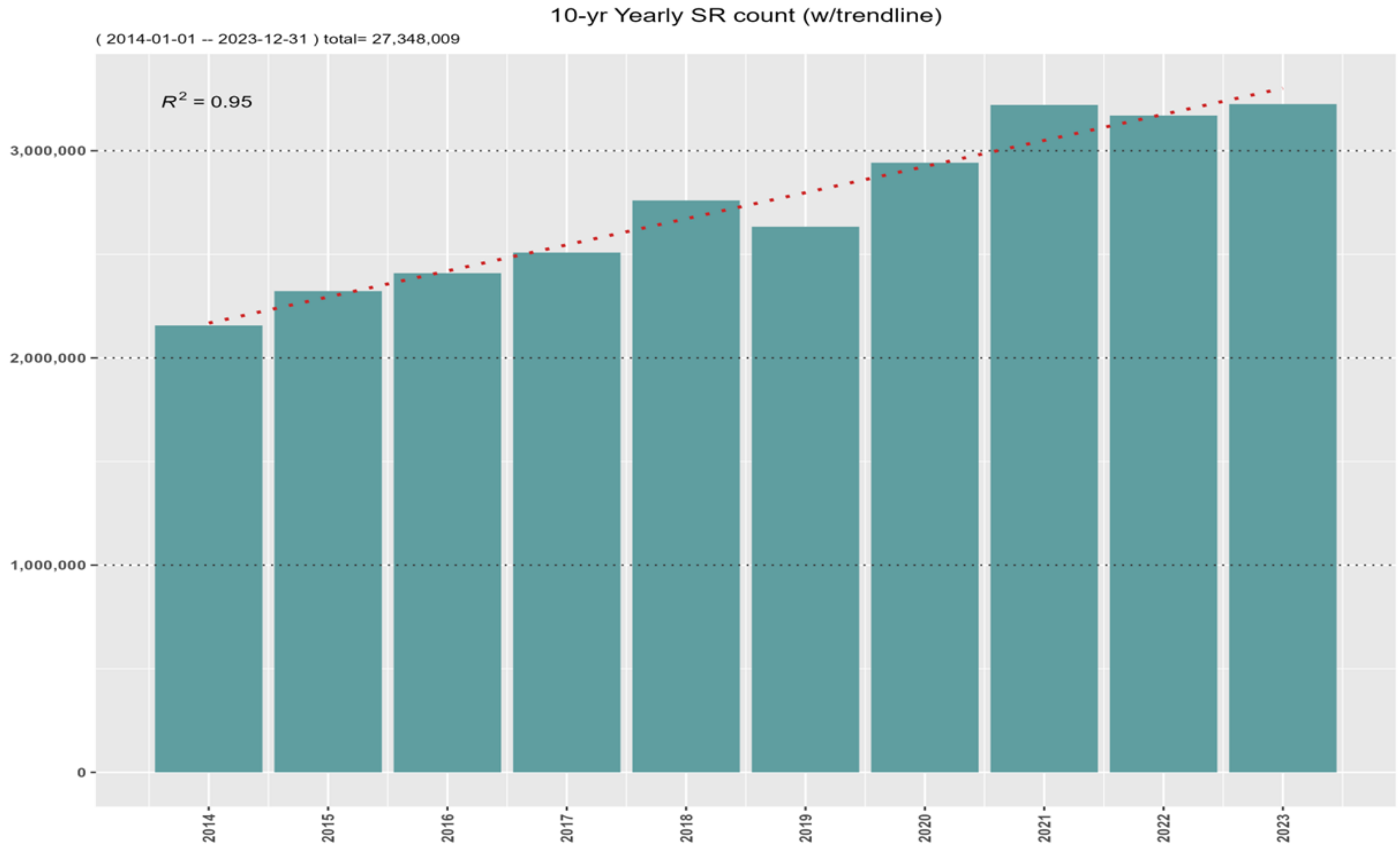
- Data cleansing is a necessary first step for data analysis.
- It is tedious, often taking longer than the analysis itself.
- Failing to do so can lead to invalid conclusions.
- Data cleansing is almost always unique to the dataset and not typically scalable.

Clean datasets are all alike;

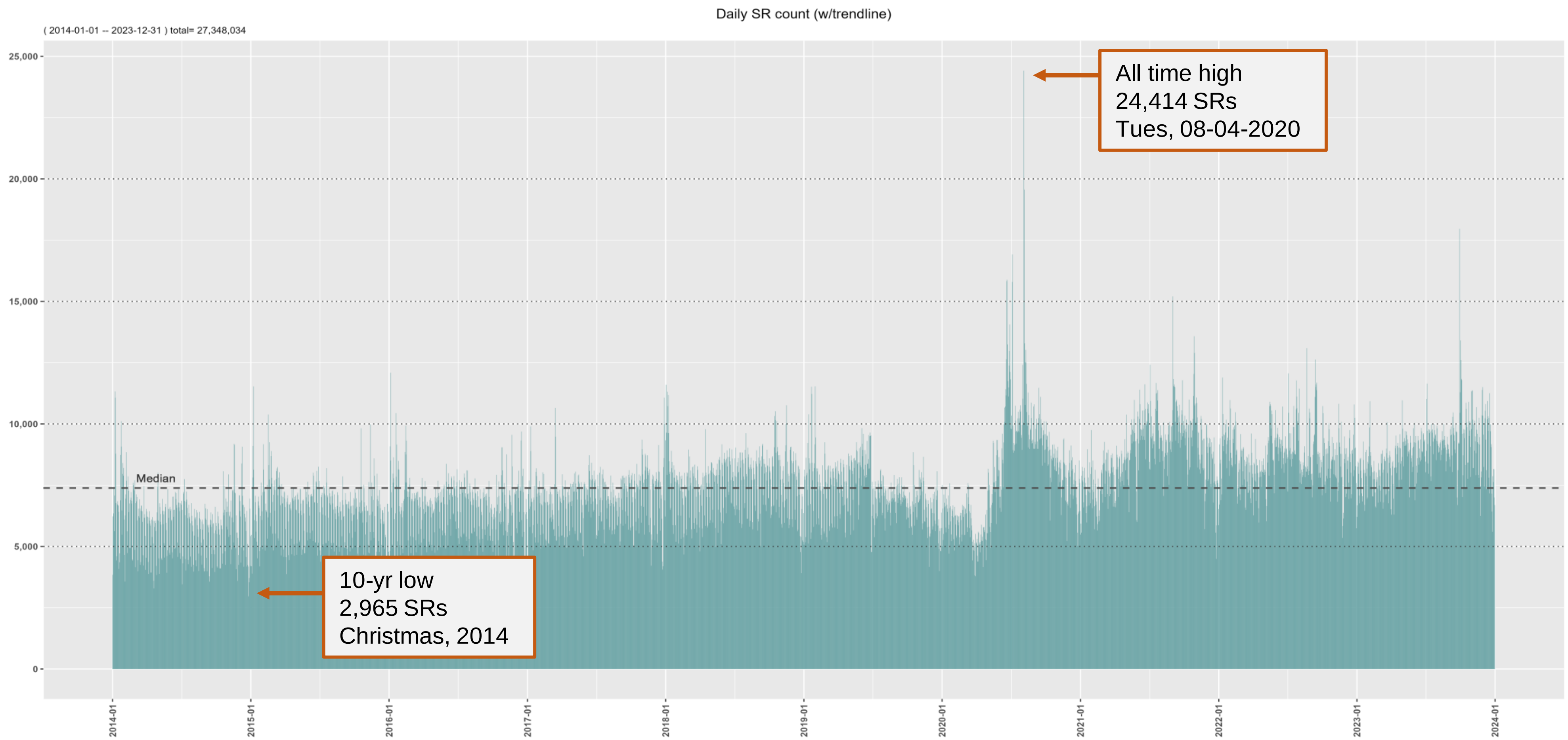
every dirty dataset is dirty in its own way.

–Hadley Wickham (cf. LeoTolstoy)

311 SR volume has grown 50% in last 10 years



10-yr Daily data. Noisy!

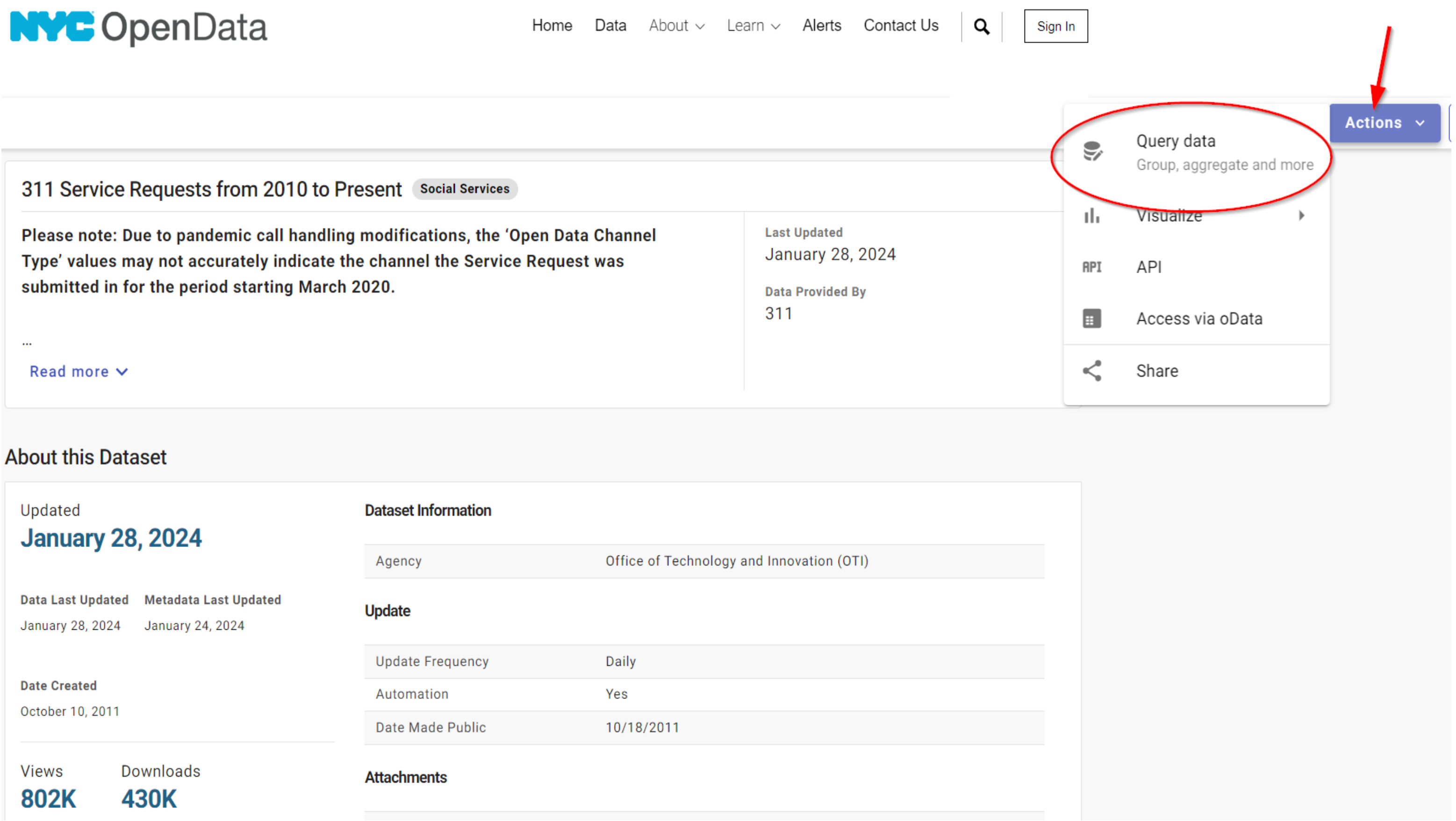


Using 311 Service Requests (SRs) for 2022-2023

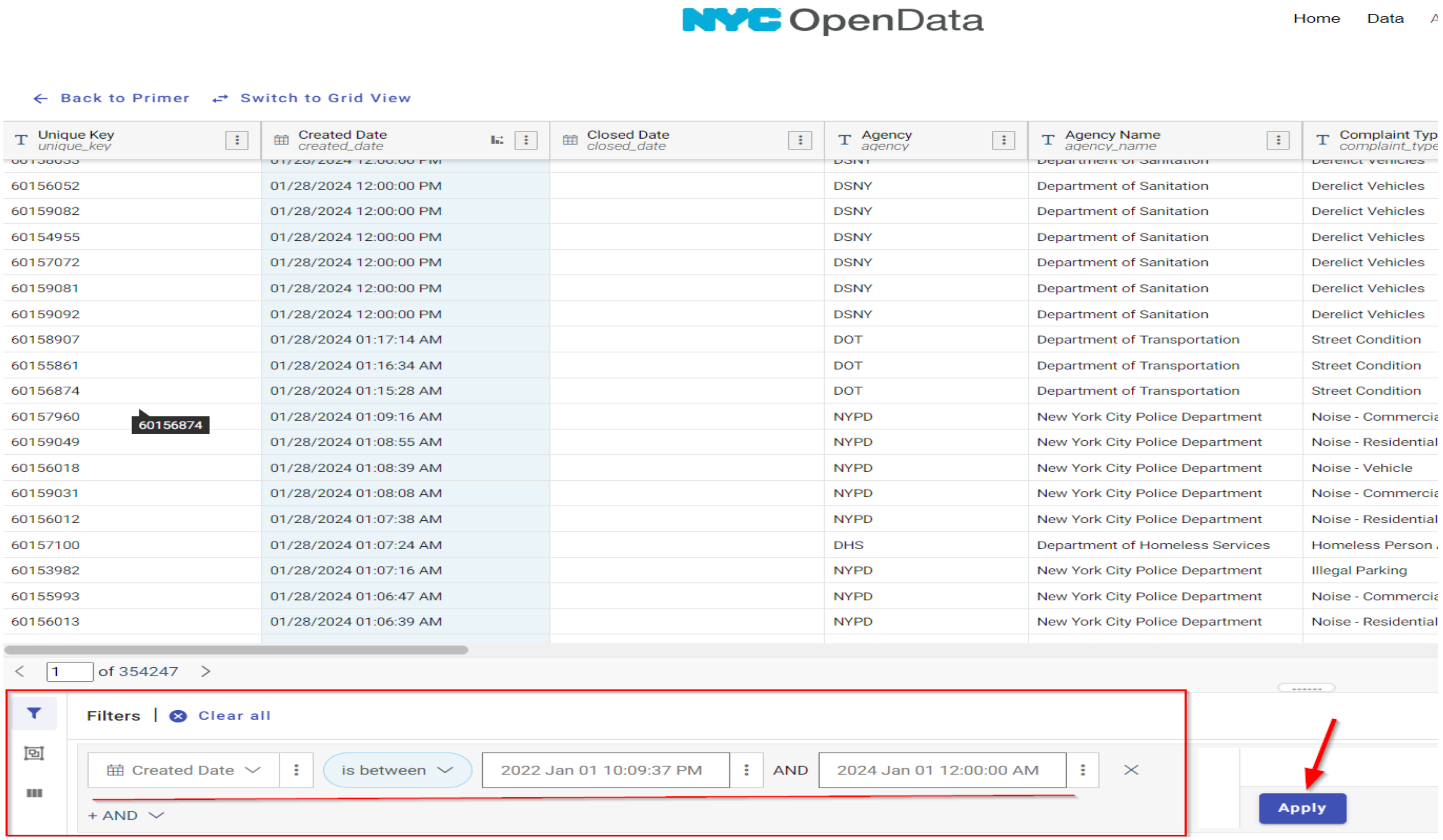
- This analysis uses CY 2022-2023 data:
    - 47 columns per row
    - Representing 16 Agencies
    - 210 different Complaint Types
    - Each row is a single SR **~6.4 million rows**

- Service Requests (SRs) come from three channels:
    - Online submission (website) – 44%
    - Phone calls – 27%
    - Mobile app – 20%
    - Other – 9%

How to get the data: *Query Data* from the *Actions* drop down



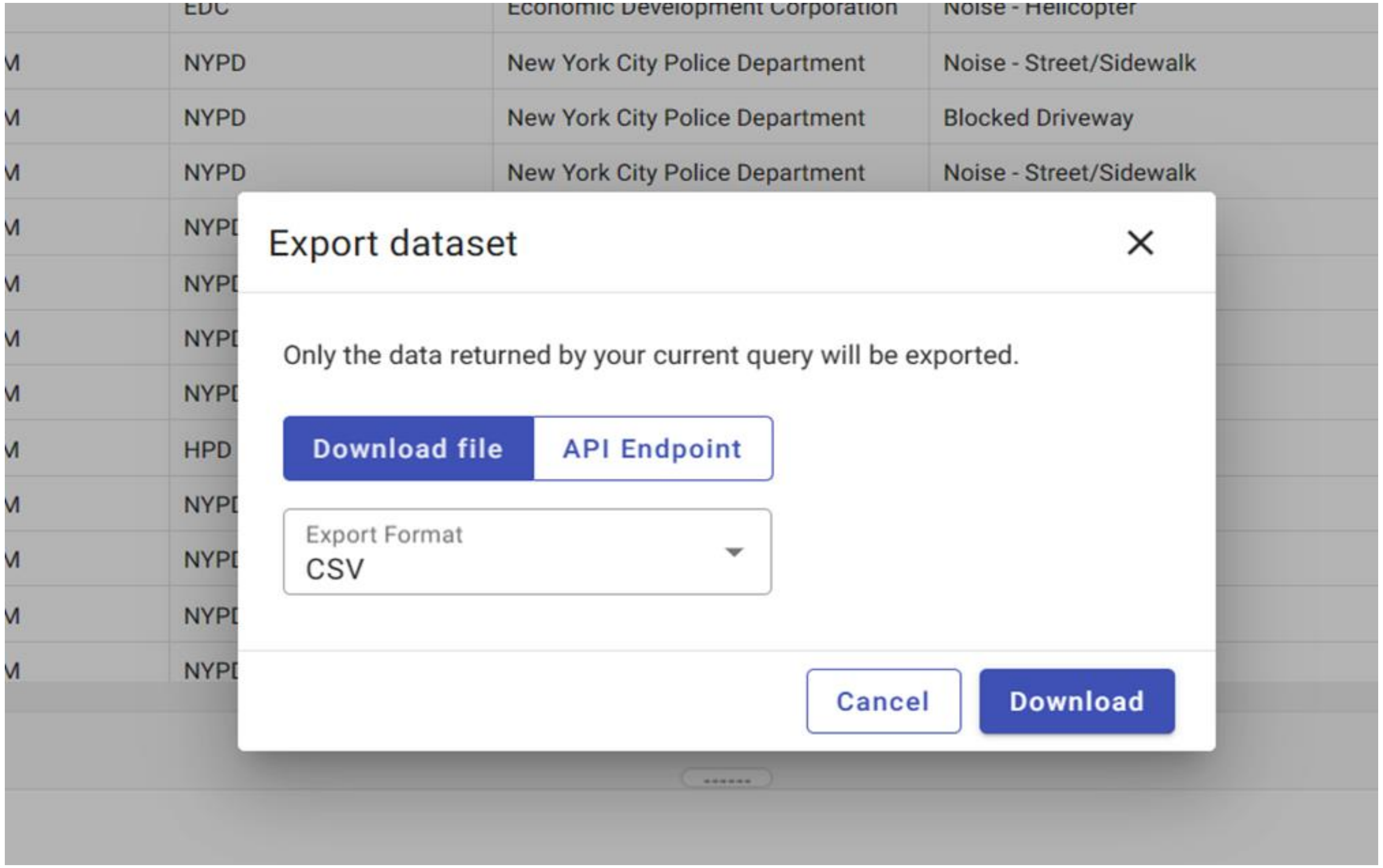
Filter by *Created Date* and “Apply”



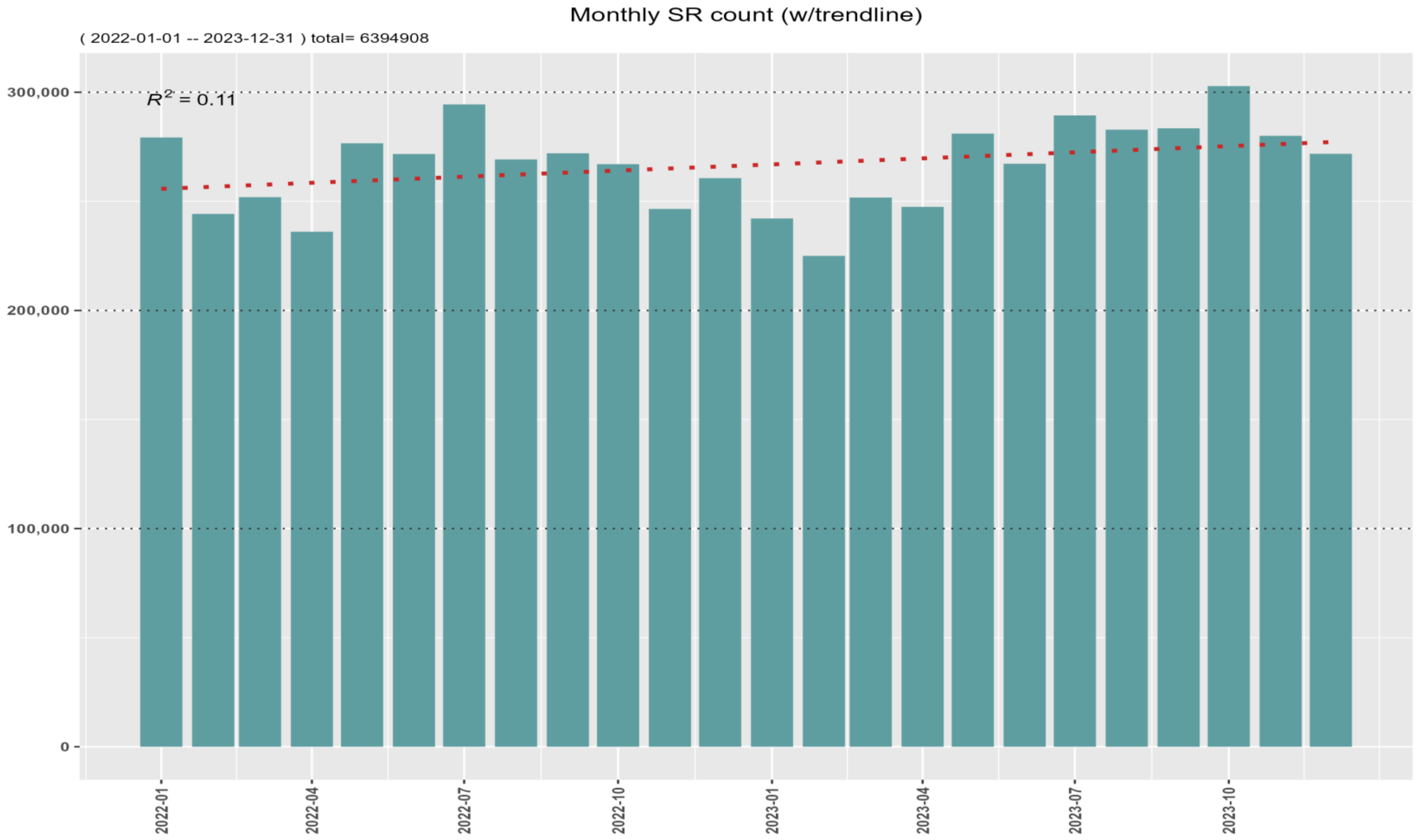
Unique Key (unique_key)	Created Date (created_date)	Closed Date (closed_date)	Agency (agency)	Agency Name (agency_name)	Complaint Typ (complaint_type)
60156052	01/28/2024 12:00:00 PM		DSNY	Department of Sanitation	Derelict Vehicles
60159082	01/28/2024 12:00:00 PM		DSNY	Department of Sanitation	Derelict Vehicles
60154955	01/28/2024 12:00:00 PM		DSNY	Department of Sanitation	Derelict Vehicles
60157072	01/28/2024 12:00:00 PM		DSNY	Department of Sanitation	Derelict Vehicles
60159081	01/28/2024 12:00:00 PM		DSNY	Department of Sanitation	Derelict Vehicles
60159092	01/28/2024 12:00:00 PM		DSNY	Department of Sanitation	Derelict Vehicles
60158907	01/28/2024 01:17:14 AM		DOT	Department of Transportation	Street Condition
60155861	01/28/2024 01:16:34 AM		DOT	Department of Transportation	Street Condition
60156874	01/28/2024 01:15:28 AM		DOT	Department of Transportation	Street Condition
60157960	01/28/2024 01:09:16 AM		NYPD	New York City Police Department	Noise - Commercia
60159049	01/28/2024 01:08:55 AM		NYPD	New York City Police Department	Noise - Residential
60156018	01/28/2024 01:08:39 AM		NYPD	New York City Police Department	Noise - Vehicle
60159031	01/28/2024 01:08:08 AM		NYPD	New York City Police Department	Noise - Commercia
60156012	01/28/2024 01:07:38 AM		NYPD	New York City Police Department	Noise - Residential
60157100	01/28/2024 01:07:24 AM		DHS	Department of Homeless Services	Homeless Person .
60153982	01/28/2024 01:07:16 AM		NYPD	New York City Police Department	Illegal Parking
60155993	01/28/2024 01:06:47 AM		NYPD	New York City Police Department	Noise - Commercia
60156013	01/28/2024 01:06:39 AM		NYPD	New York City Police Department	Noise - Residential

Powered by NYC OpenData

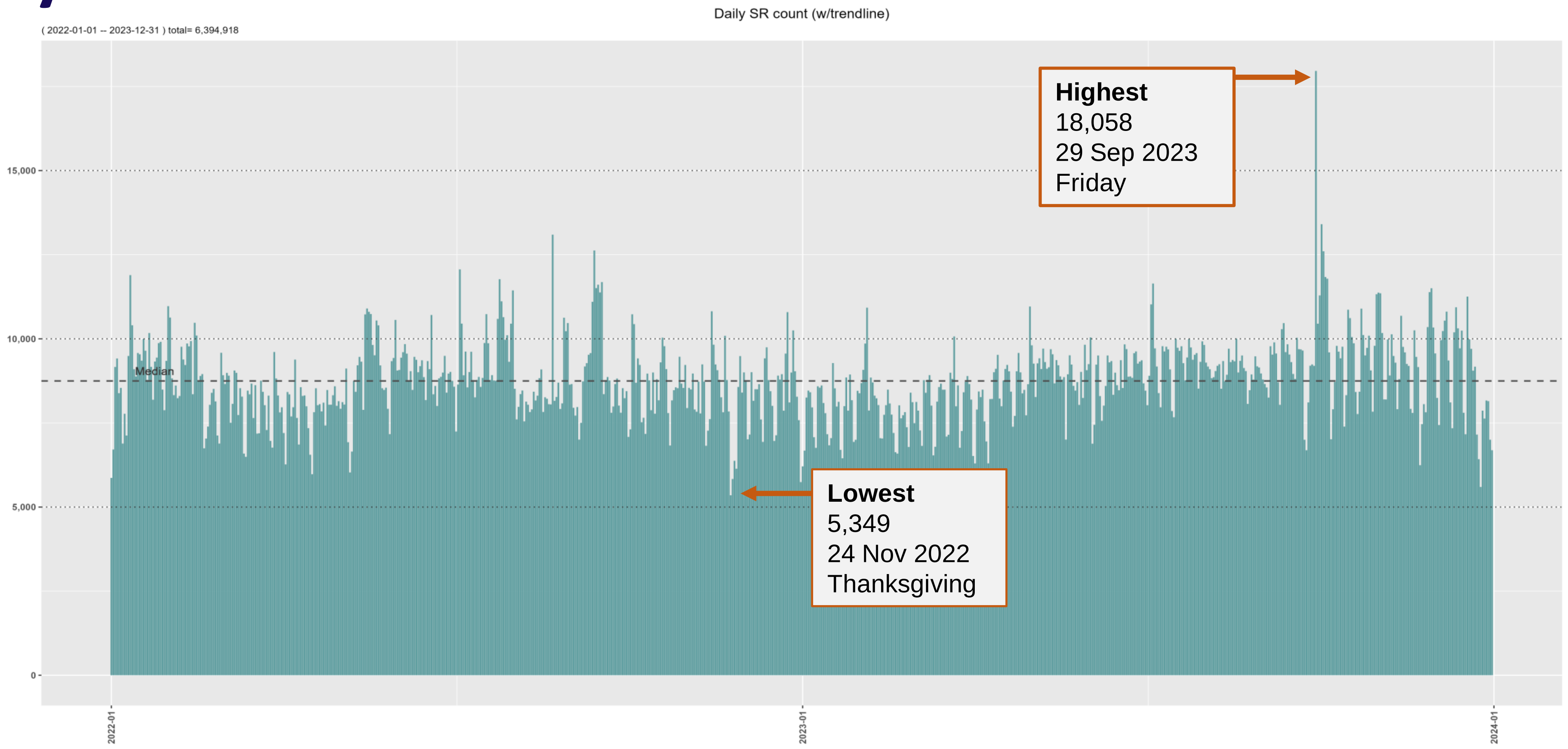
Export in CSV format. Analyze in a custom R program.



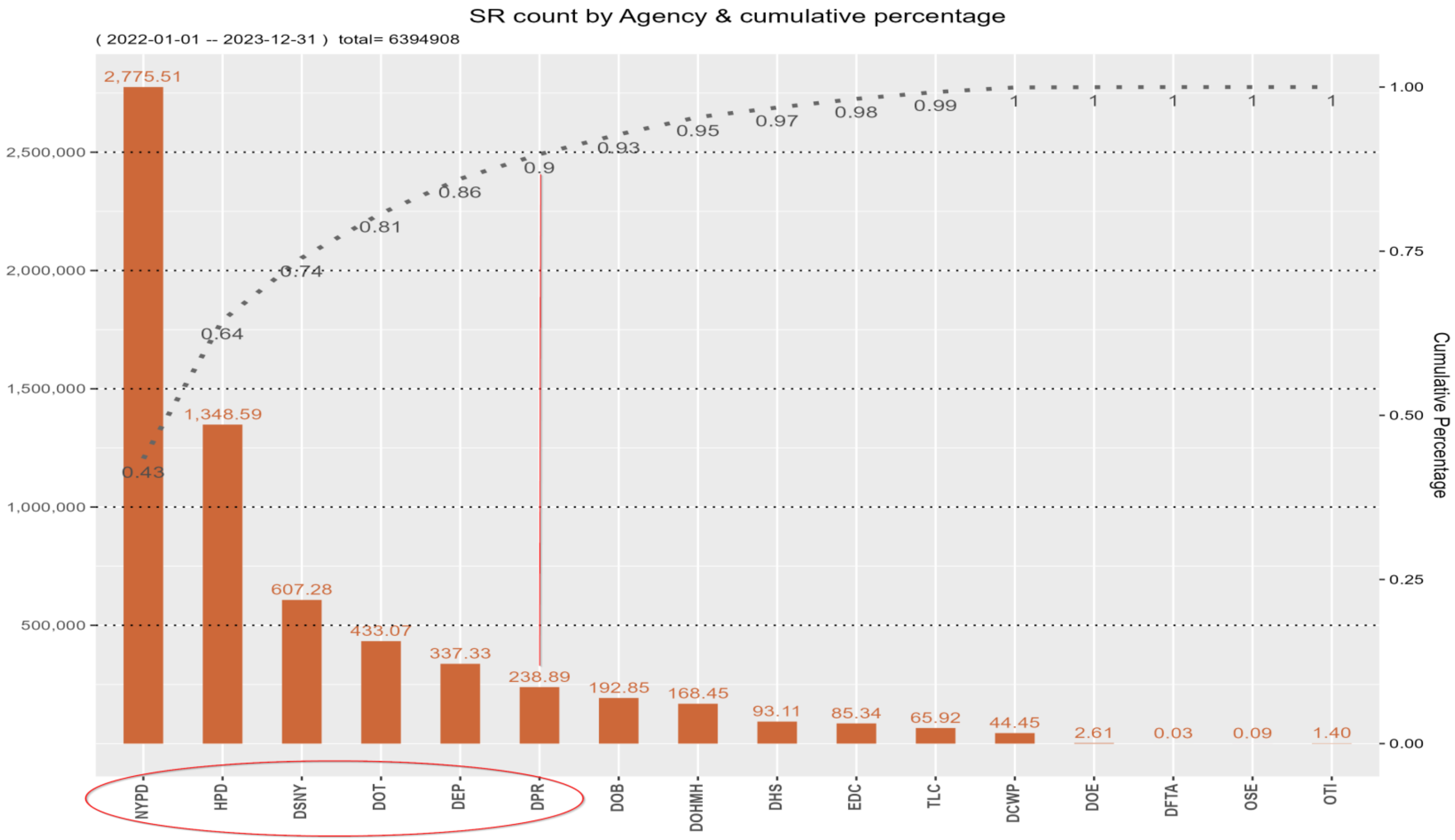
Monthly data: 2022-2023



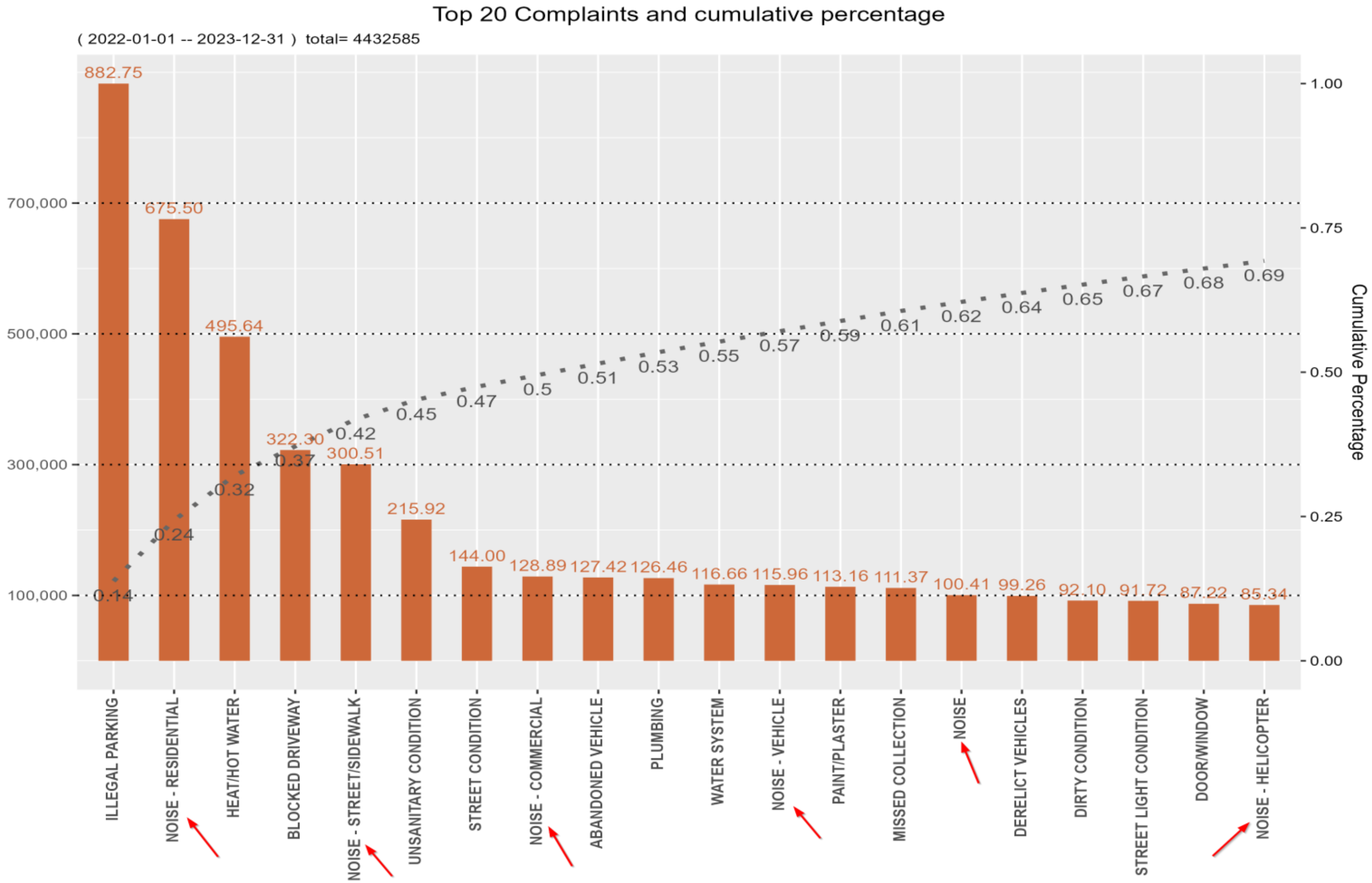
Daily Data: 2022-2023



Responsible Agencies: 'Big 6' comprise 90% of SRs



210 Complaint types



Top Five Complaints comprise 50%

- **Noise – 22% in-total (8 different types)**
  1. Noise – Residential (~ ½ of the noise complaints)
  2. Noise - Street/Sidewalk
  3. Noise – Commercial
  4. Noise – Vehicle
  5. Noise
  6. Noise – Helicopter
  7. Noise – Park
  8. Noise – House of Worship

- Illegal Parking – 14%
- Heat/Hot Water – 8%
- Blocked Driveway – 5%
- Unsanitary Condition – 3%

Fewest Complaints

- Trans Fat
- Tanning
- Tattooing
- Unlicensed Dog (1)
- Quality of Life
- Taxi Compliment (3)
- Radioactive material (!)

Data Cleansing: What should we check for?

Identifying dirty data. What to include/exclude?

Six areas to investigate:

1. **Structural issues** with the data. Is it in the expected format?
2. **Missing**/blank data
3. Do data fields contain the correct **data types**?
4. **Invalid** values?
5. **Logical inconsistencies**? Concerning patterns?
6. **Redundant** columns?

1. Structural: What does the 311 SR data look like?

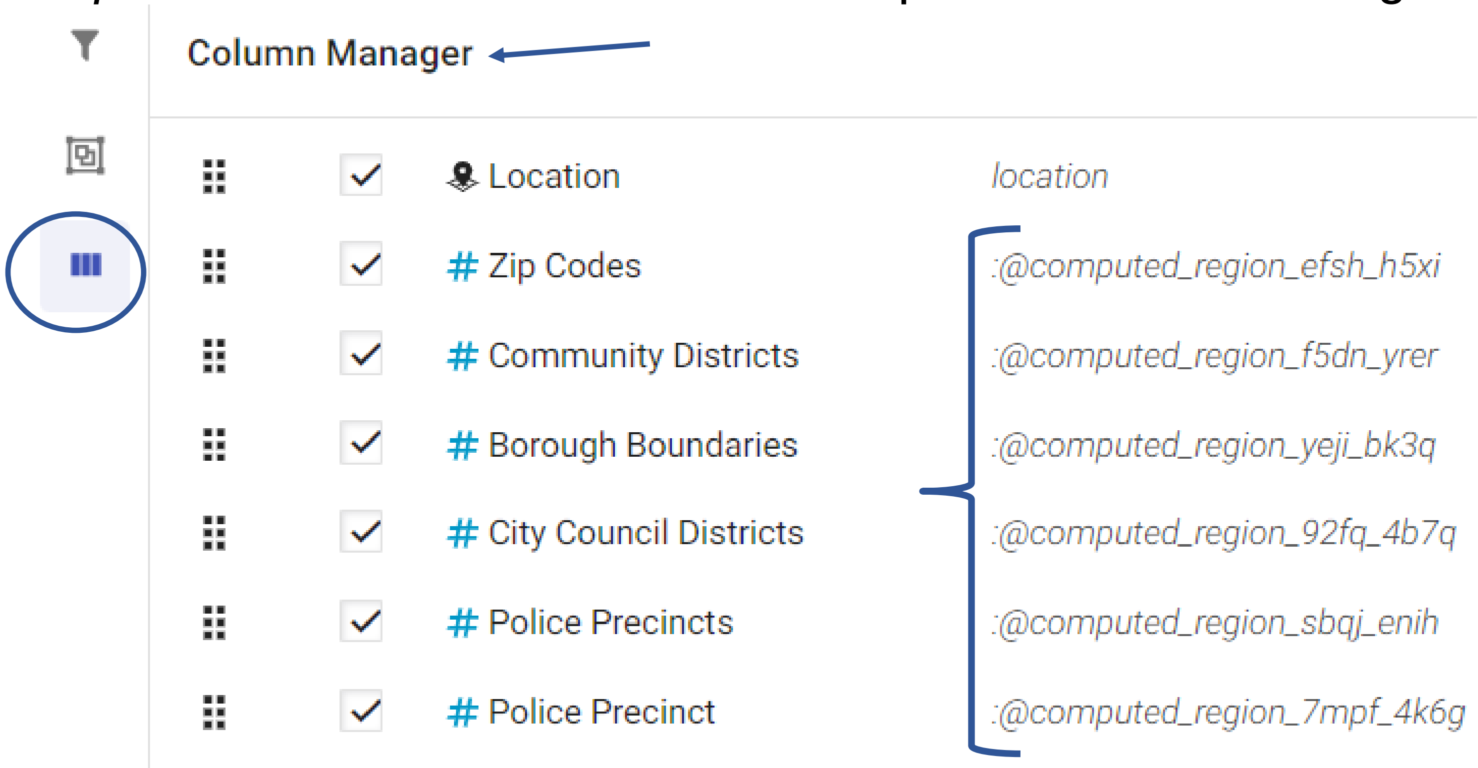
- Four date fields (created, closed, updated date, due date)
- Three borough fields; two of which are duplicates
- Two zip code fields; one has many invalid codes
- Seven street fields; two pair of which are (near) duplicates (cross_street, intersection_st)
- Agency abbreviation and Agency formal name
- Two Police Precinct fields (precinct, precincts); not 100% duplicates
- Six *computed* fields that have validity issues.
- Three *location* fields in addition to street: lat/long, X/Y state plane, Block # (BBL)

1. Structural Issues: *Computed* fields not in the Data Dictionary

- Data Dictionary defines 41 columns
- However, data extracts contain 47 columns
- The extra columns are 6 *computed* fields. (Displayed as such in the Open Data Portal, but not in the Dictionary.)
  - *zip_codes*
  - *community_districts*
  - *borough_boundaries*
  - *city_council_districts*
  - *police_precincts*
  - *police_precinct*
- What is the validity of these fields?

1. Structural: *computed* fields not in Data Dictionary

Computed fields are shown as such in portal Column Manager



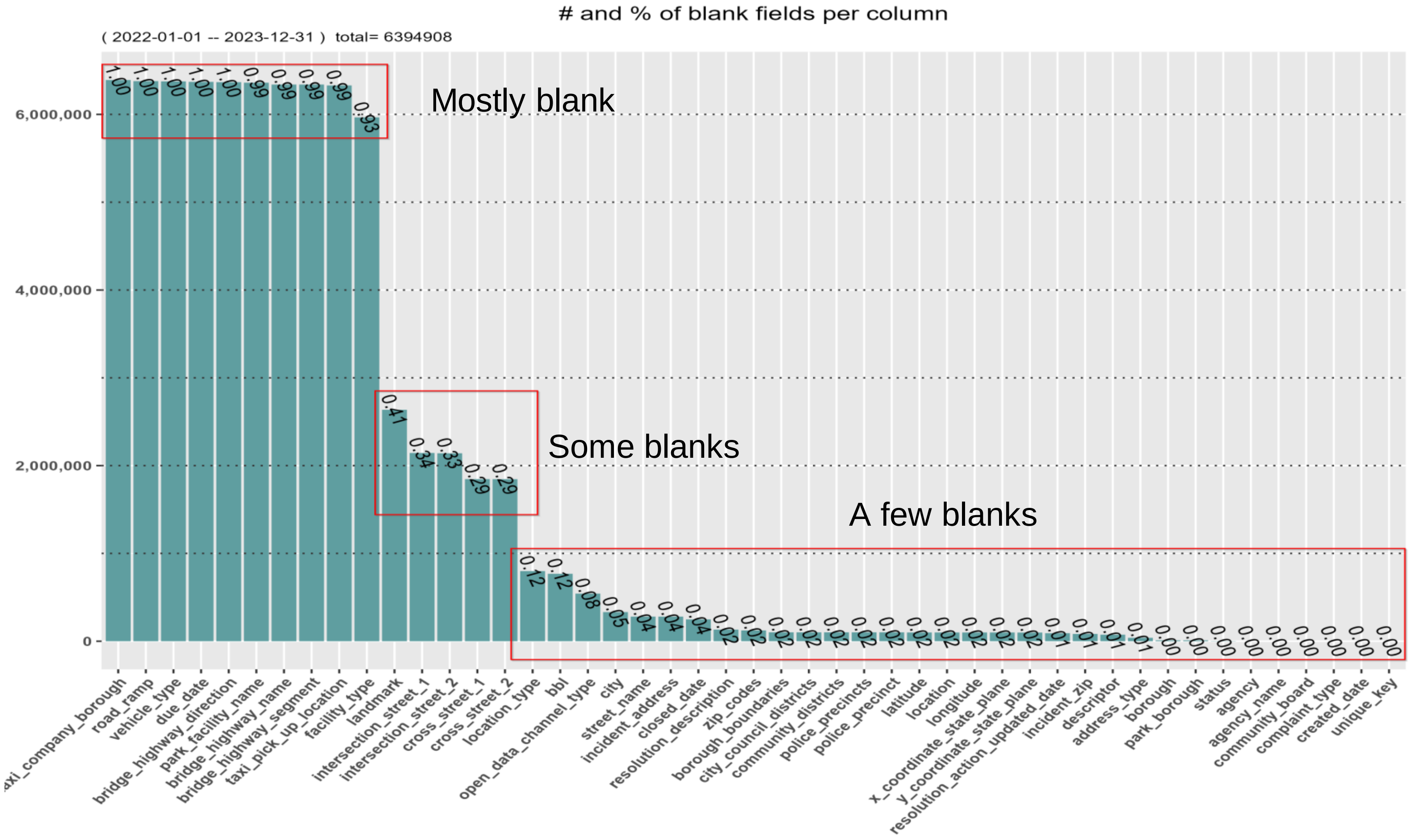
2. Missing/blank data: three groups (mostly, some, none)

- **Mostly blanks** (93–99.9% blank) – 10 fields
  - Taxi & Limo fields? (*pickup_location, taxi_company*)
  - *due_date*, highway fields (*ramp, bridge, name*)
  - *Facility, park_facility_name, landmark*

- **Some blanks** (29-41% blank) – 5 fields
  - *Intersection_street(s), cross_street,(s), location_type, borough*

- **No/few blanks** (0-12% blank) – 32 fields
  - *created_date, closed_date, complaint_type, agency, status, community_board , zip, descriptor, borough*, etc.

2. Missing data: mostly blank, some blanks, almost none



3. Data types: Do columns contain the correct data type? <u>Yes</u>

- Subjected all columns to data “type” validation (numeric, character, date)
    - Almost all were compliant ☺

- All four date fields are valid dates
    - All missing dates are represented as “NA”

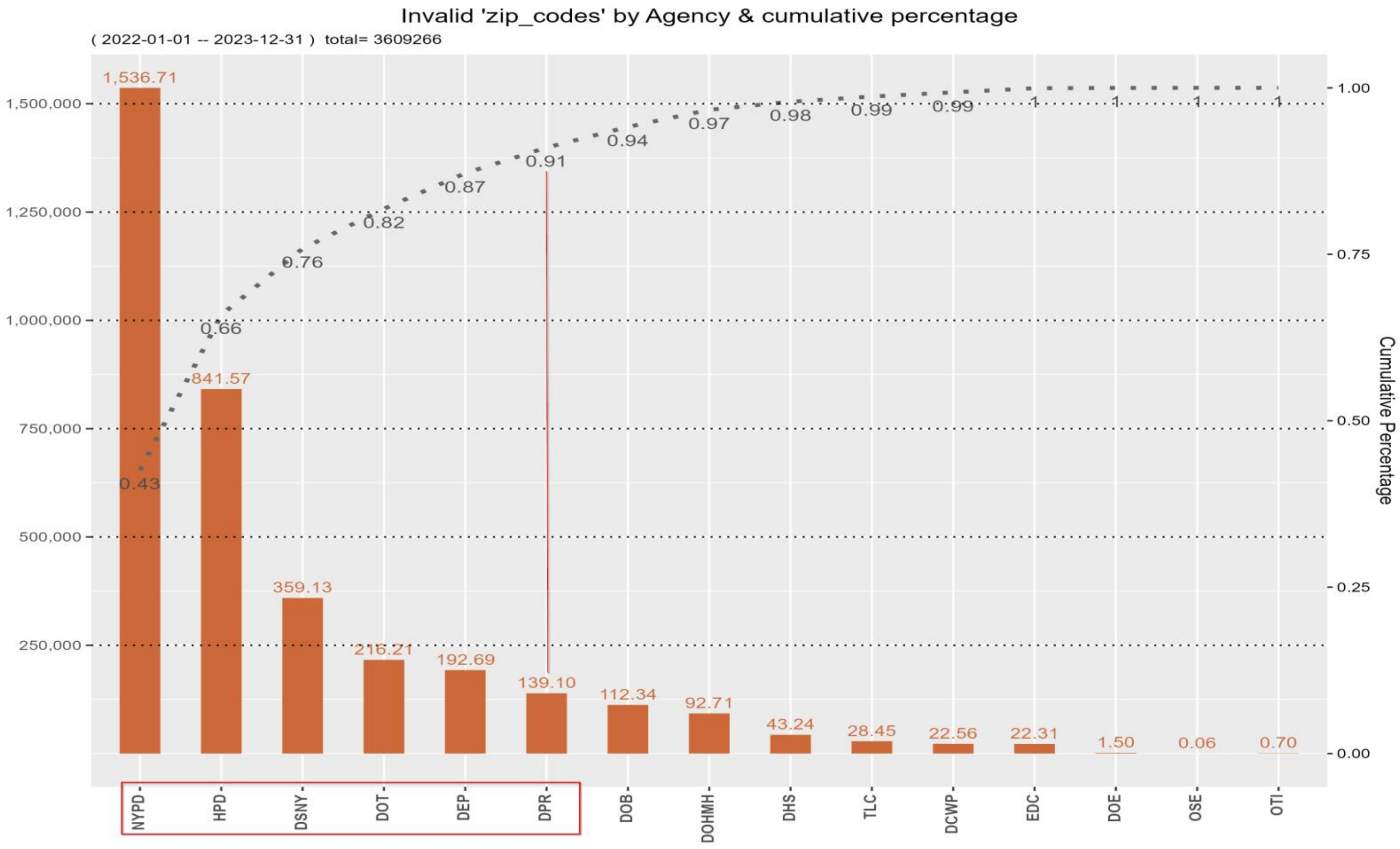
4. Valid data: Allowable values

- *Lat/Long's* are all within boundaries of NYC
- All *unique_key* values are in fact unique
- Most columns have a domain of legal values:
    - *address_type*
    - *status*
    - *borough & borough_boundaries & park_borough*
    - *data_channel*
    - *vehicle_type*
    - *city_council_district*

4. Invalid data: Non-allowable values

- 58% (3.6 million) of computed *zip_codes* are invalid
- 35% (2 million) of computed *police_precincts* are invalid
- 0.4% (27,000) of *community_board's* are invalid
- 0.1% (4,000) of *incident_zip* are invalid values
- Data representation issues prevent evaluating computed *community_district*. However, there are 72 unique entries, and only 59 valid Community Districts.
- *Latitude* & *Longitude* are captured as a 14-decimal field (**atomic level)**
    - Ex: Lat: 40.86769186022511 : **(1.1nm or ~3 atoms wide)**

4. Invalid data: *zip_codes* by Agency (3.6 million/58%)



Case Study: Where are Noise Complaints by Zipcode?

- NYC Office of Nightlife wants to know: *What are the top 10 zip codes for Noise Complaints (all 8 types) over the last two years?*
- Select 2022-23 SRs where *complaint_type* begins with 'Noise', select the ***zip_codes*** column and aggregate by count. *Voila!*

Rank	zip_codes	Counts
1	11275	104,556
2	12420	27,503
3	12428	26,564
4	10935	25,508
5	10934	23,448
6	10931	22,381
7	10930	22,121
8	17613	21,963
9	10936	21,707
10	11606	21,435

Case study: What went wrong?

- Unfortunately, 58% of the computed *zip_codes* are invalid.
- Luckily, the *incident_zip* field is more accurate (99.93%).
- This is a typical issue when faced with duplicate fields; which one is right?

zip_codes	Count	Valid?	incident_zip	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

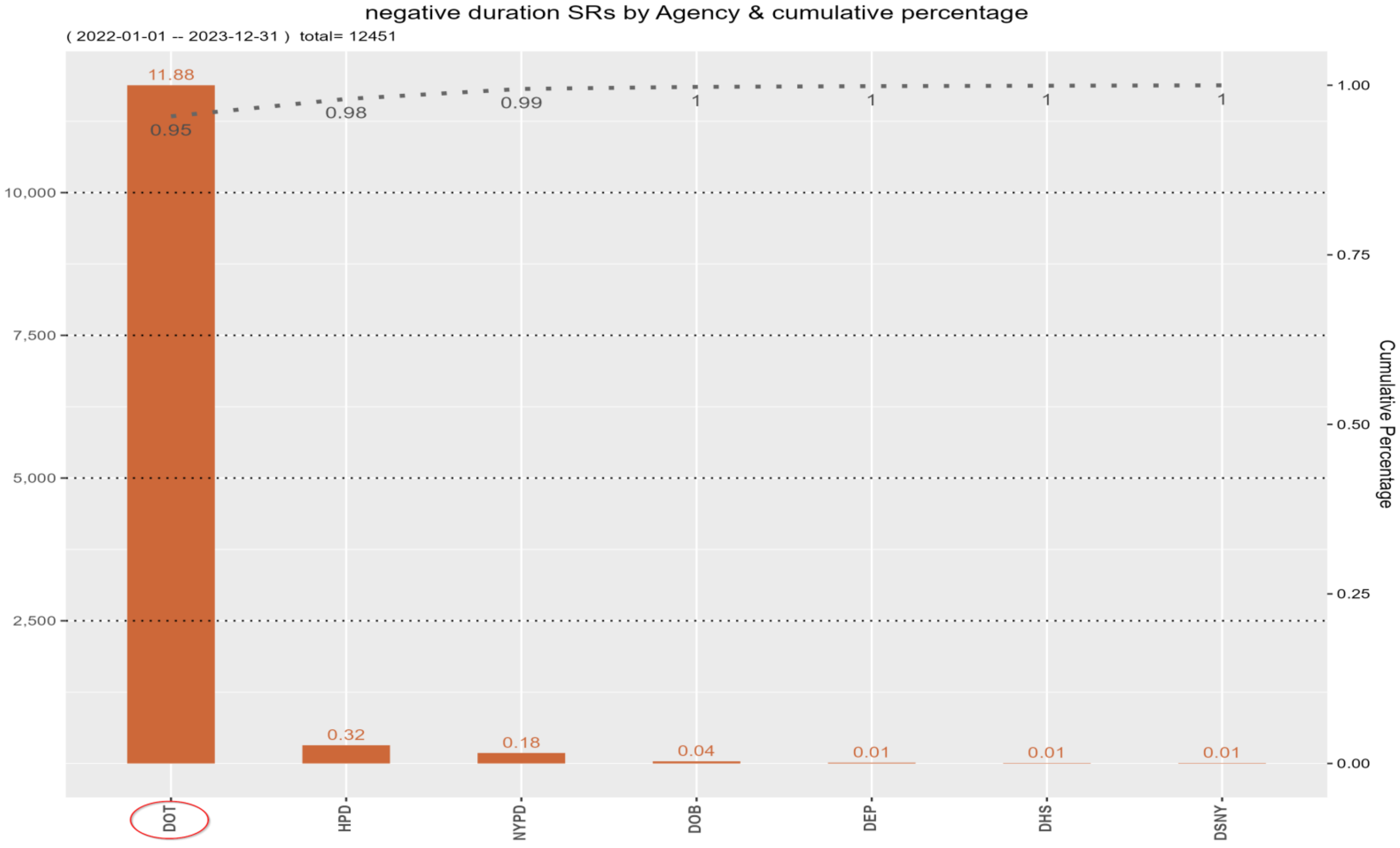
5. Inconsistent & Unusual patterns with dates

- 12,000 (0.2%) SRs are **“closed” before they are “created”** creating a negative duration/response time
  - max: -44,602 days (122 yrs) Some have *closed_date* of 01/01/1900 – Excel?

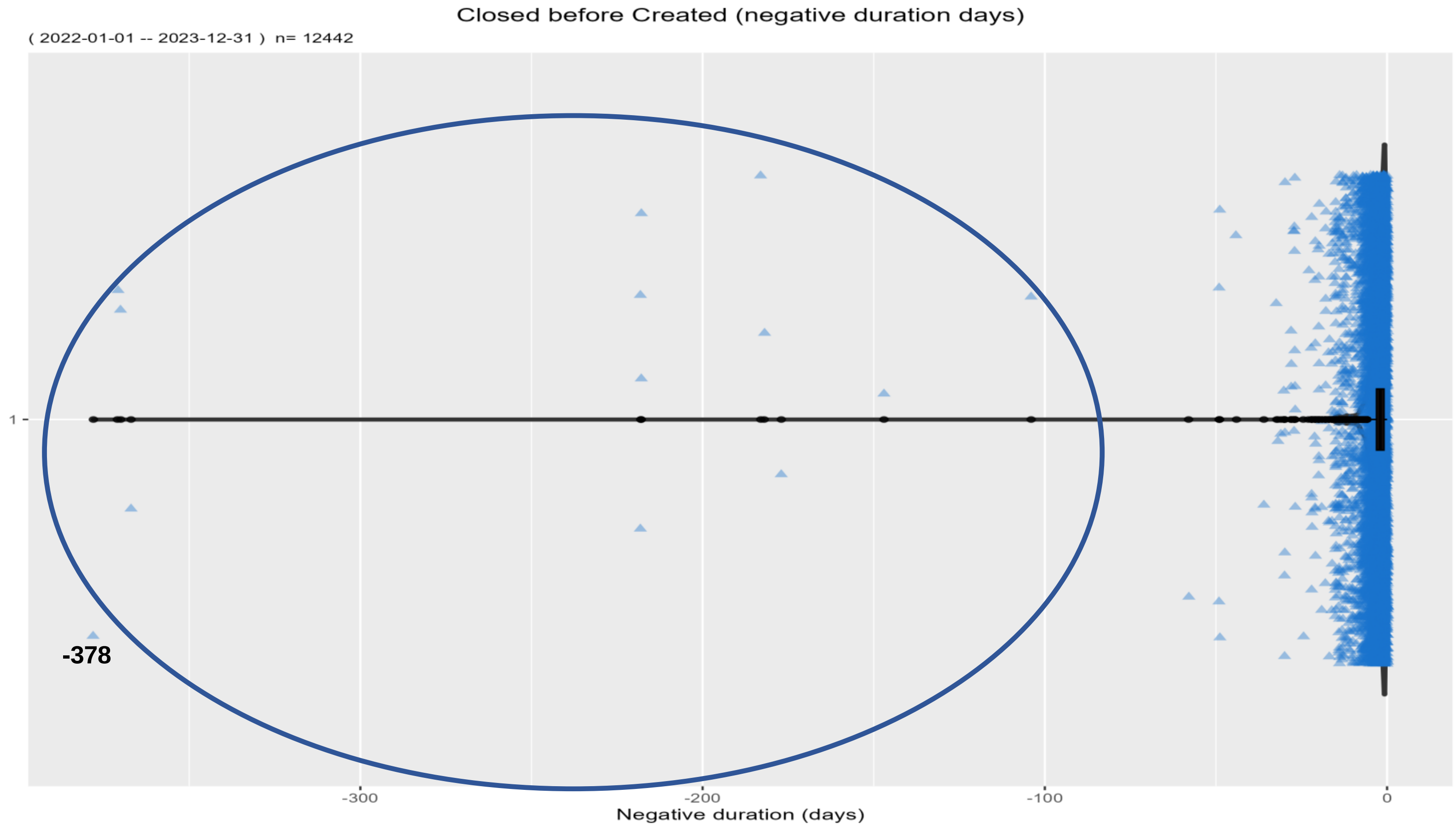
- 193,000 SRs (3%) **“created” and “closed” at the exact same time** (*to the second*) creating a zero duration

- 7,500 SRs (0.1%) are **updated >30 days after they were “closed”** (*resolution_action_update_date*). Is this an error?
  - max: 44,602 days -- max (excluding extreme outliers): 636 days

5. Logical inconsistencies: *closed* before *created*



Problem: *closed* before *created* (negative duration)



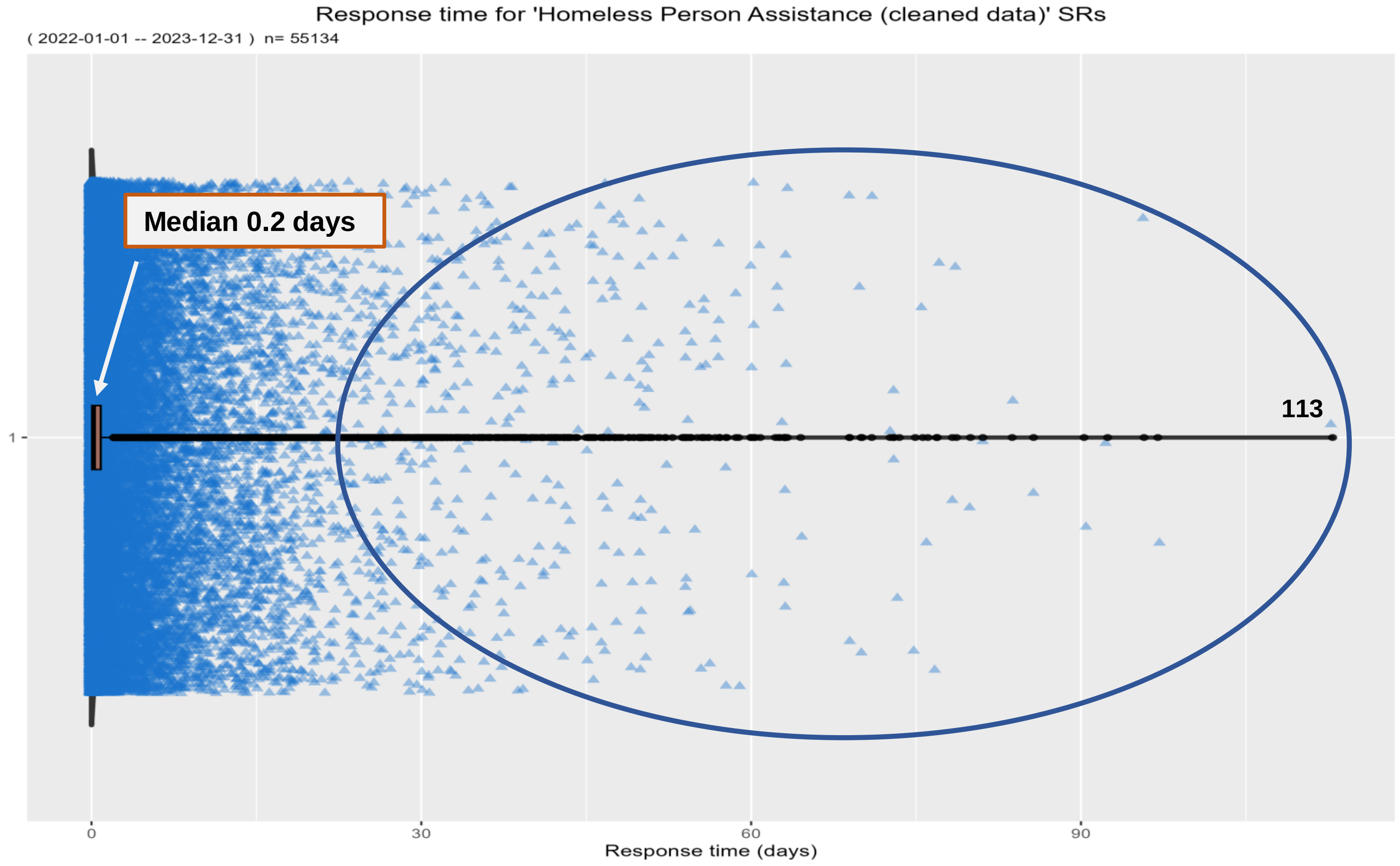
Case Study: Homeless Person Assistance

- DHS wants to know: *How quickly are 311 calls for “Homeless Person Assistance” resolved?*
- Filter data by *complaint_type* = “Homeless Person Assistance”
  - ~75K SRs
- Compute the *duration* (*closed_date – created_date*)
- Take an average of the “duration” field. *Voila!*
  - **Answer: -4.8 days (!)** What? How can that be?

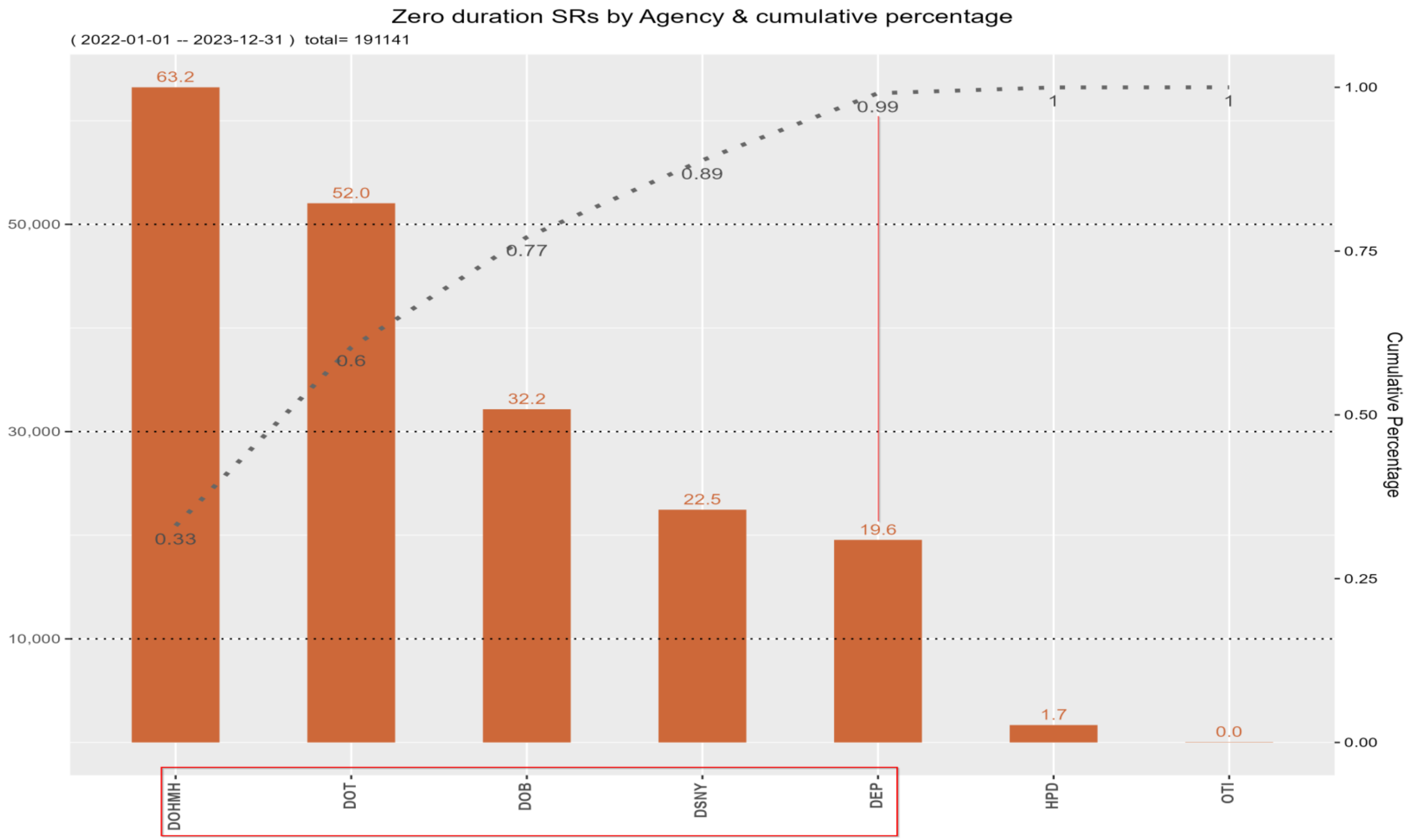
Case study: What went wrong?

- As it turns out, there are 8 DHS SRs with a *closed_date* of **01/01/1900** (Excel?)
  - Each of these SRs creates a **negative *duration* of -44,602 days**
- As a result, Average duration: -4.8 days. (Median 0.2 days)
- If you remove the obviously incorrect *closed_date*'s
  - Now the Average is 1.7 days (Median 0.2 days)
- **BUT** note that the median is 0.2 days (~5 hrs), meaning half of the Homeless Assistance requests are solved quite rapidly.
- Given the outliers, **MEDIAN is the better descriptive** statistic in this case.

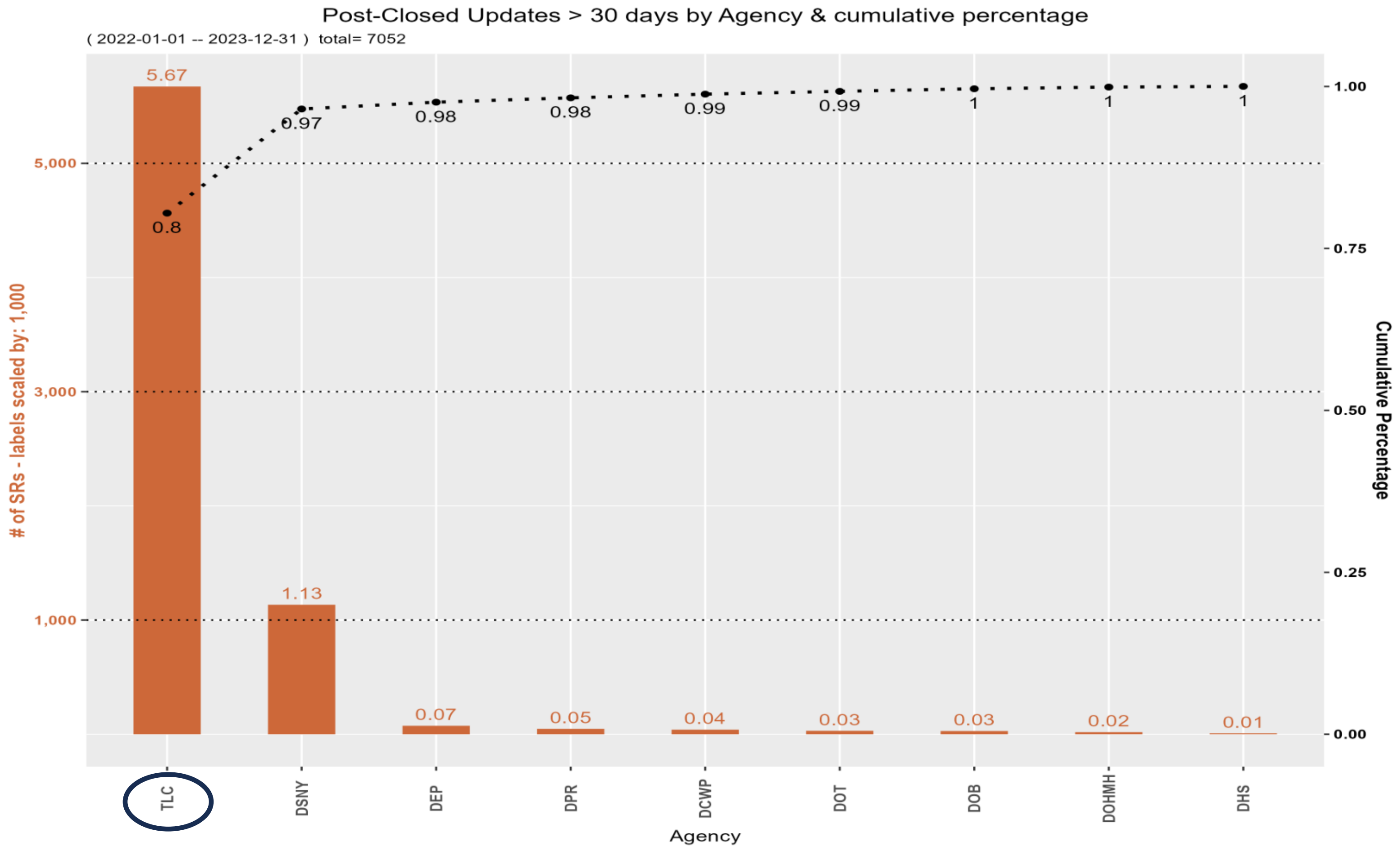
Homeless Response times: Outliers distort average



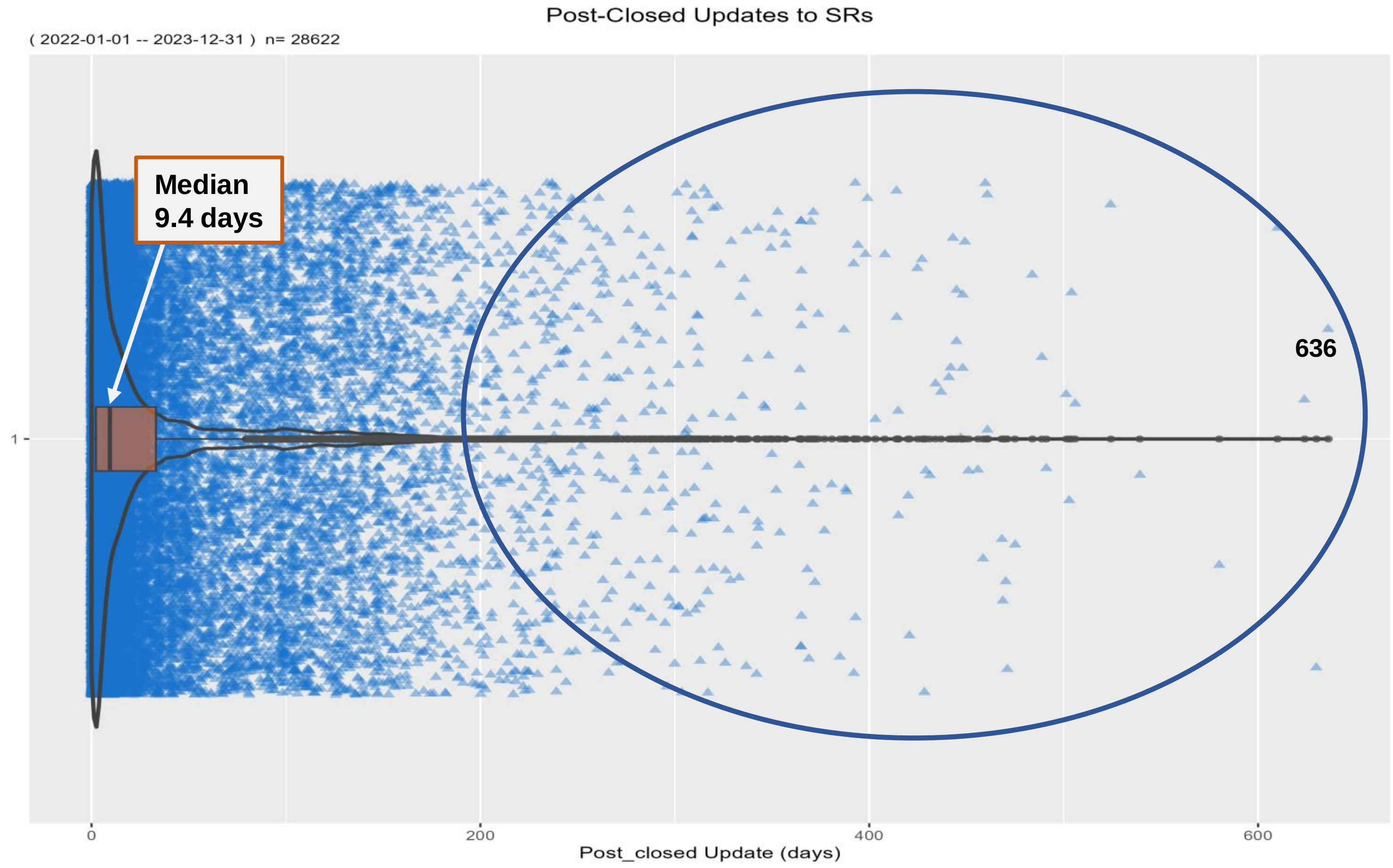
5. Logical inconsistencies: ~200K *closed* & *created* at same time (zero duration)



5. Logical inconsistencies: Post-closed resolution updates



Problem?: post-closed updates (lots of outliers)



6. Redundant/Duplicate columns: Which is right?

- 100% match of *borough* & *park_borough* (redundant)
- 98% match of *borough* & *borough_boundaries* (redundant)
- 0.05% match of *borough* & *taxi_company_borough* (mismatched)
- 99.9% match of *police_precincts* & *police_precinct* (redundant)
- *Location* is a pure concatenation of *Lat* & *Long* (redundant)
  - *Latitude*: 40.627308819544446 *Longitude*: -74.00862444250549
  - *Location*: (40.627308819544446, -74.00862444250549)

6. Redundant or Duplicate columns: Which is right?

- 88% match of *cross_street_1* & *intersection_street_1*
- 88% match of *cross_street_2* & *intersection_street_2*

Which field do you trust when one of the (almost) duplicate fields differ or one is blank and the other is not?

Overall problems in the data

- Blank/missing fields. **Which fields are useful for analysis?**
- Duplicate fields make analysis challenging and increase file size. **Which field is the correct one?**
- Invalid values distorts analysis especially *computed* fields *zipcodes, police precincts,* and *community board.*
- Incorrect *Closed* and *Created* dates can create negative/zero durations which distorts response time analysis.

Recommendations

- Many of the fixes lie at the Agency source, not 311 *per se*.
- Update Data Dictionary to provide more clarity and value domains (can provide draft)
- Review fields with high percentage of blank/unknown values.
- **Evaluate eliminating invalid values;** incorporate drop-downs, pick-lists, field validation, etc. Improve *computed* fields.
- **Create logical controls** on selected fields, especially dates
    - *created_date, closed_date, resolution_action_update_date*?
- **Eliminate/consolidate duplicate fields**
    - Intersection_street(s)/cross_street(s)
    - *borough, borough_boundaries, taxi_company_borough, park_borough*
    - *location* and *latitude, longitude*
- Correct excessive precision in Lat/Long fields (14 digits)

Final thoughts

- Data cleansing is a critical step before beginning analysis.
    - Data Cleansing determines which data fields can be trusted.
    - Data Cleansing indicates which data elements should be evaluated, possibly removing invalid, unusual, and illogical values.
- A full report is available for review including more examples, graphs, and R code.
- NYC ODW/School of Data might consider a basic course in data cleansing.
- Recommend data cleansing be included as part of a Data Science academic courses

311 SR Data:
How Clean is It?

For further information contact:

David Tussey
davidtussey@gmail.com

Dr Jun Yan (Uconn)
jun.yan@uconn.edu

Join us at more events through Sunday, March 24. Program at open-data.nyc.