# Temporal Analysis: A Timesaving Approach

By Jon Blue (jb7259) and Wilson Li (tl2894)

# Table of Contents

# Introduction

The temporal attributes of a dataset are those that relate to time. Their granularity can vary depending on the data's purpose - an electrical engineer might need nanosecond-precision to handle CPU-related information, while a geologist might only require millennial granularity when studying rock layers. What makes temporal data useful across disciplines is how it enables the data to tell a story across time.

## Problem Formulation

Fittingly, given our study of temporal data, the problem we are trying to solve dates back to the very beginning of the course: slides 59 and 61 of the course overview presentation. Slide 59 raises a fundamental Big Data question: How do we find data relevant to a specific information need? Slide 61 then offers up an example of an information need: How do we explain the three outliers in the 2011-2012 NYC Taxi Trip Data?

That specific information need has already been met (hurricanes and gas prices), but we want to help address that issue in a more general sense. If a data scientist is looking at a graph with time-related anomalies like the one on Slide 61, she might not know what data set(s) will help explain the outliers, but she will know one thing for certain. If the solution lies within a particular dataset, that dataset MUST include temporal data.

Given the massive number of datasets available to review, any pre-processing method that can eliminate datasets from consideration is incredibly useful. Our project examines existing techniques for identifying temporal data, assesses their strengths and weaknesses, and builds on their approaches. The ultimate goal is a tool that can be used by data scientists trying to answer temporal questions.

## Related Work

Temporal detection relies mainly on the pattern existing in data themselves or the metadata. Existing sophisticated open source dataset analytical tools like Datamart (VIDA-NYU / datamart / Auctus · GitLab) and IBM InfoSphere Information Analyzer (IBM InfoSphere Information Analyzer) will always prioritize information gained from metadata to help make judgment. Accurate and sufficient metadata provides much information about the datasets - sometimes it will reveal hidden assumptions and context intended by the author that's hard to detect from data themselves. When metadata is lacking, the tools above rely heavily on various regular expression analysis and type check on data cells for detection. Metadata needs to be manually incorporated into datasets, so many datasets with lack of metadata existing in the data pool further exacerbate the issue. Thus, we mostly focus on temporal detection techniques that derive results from cell data.

# Methods, Architecture & Design

## Datasets

Given that a tool to profile datasets is most useful if it works for the most popular ones, we used the top 100 most viewed datasets on NYC OpenData as the basis for our analysis. They cover a variety of contexts including government registry, health & clinical,public transportation and so on. Hence, they provide a good representation of how temporal data are managed and stored in our modern day setting. A complete breakdown of those sets, including names, temporal column attributes, and the various formats of temporal information can be found in Appendix I.

## Assembling the Data

The NYC OpenData collection is available for download via their website (https://data.cityofnewyork.us/). It's available in both .JSON and .CSV formats. We opted for .CSV to maintain consistency with the 135 other datasets that were loaded separately on HFS by the Teaching Assistants. All downloads default to 1000 lines. We downloaded both the default versions and 100K-line versions to determine if the larger sample sizes made a difference. We observed similar results across both.

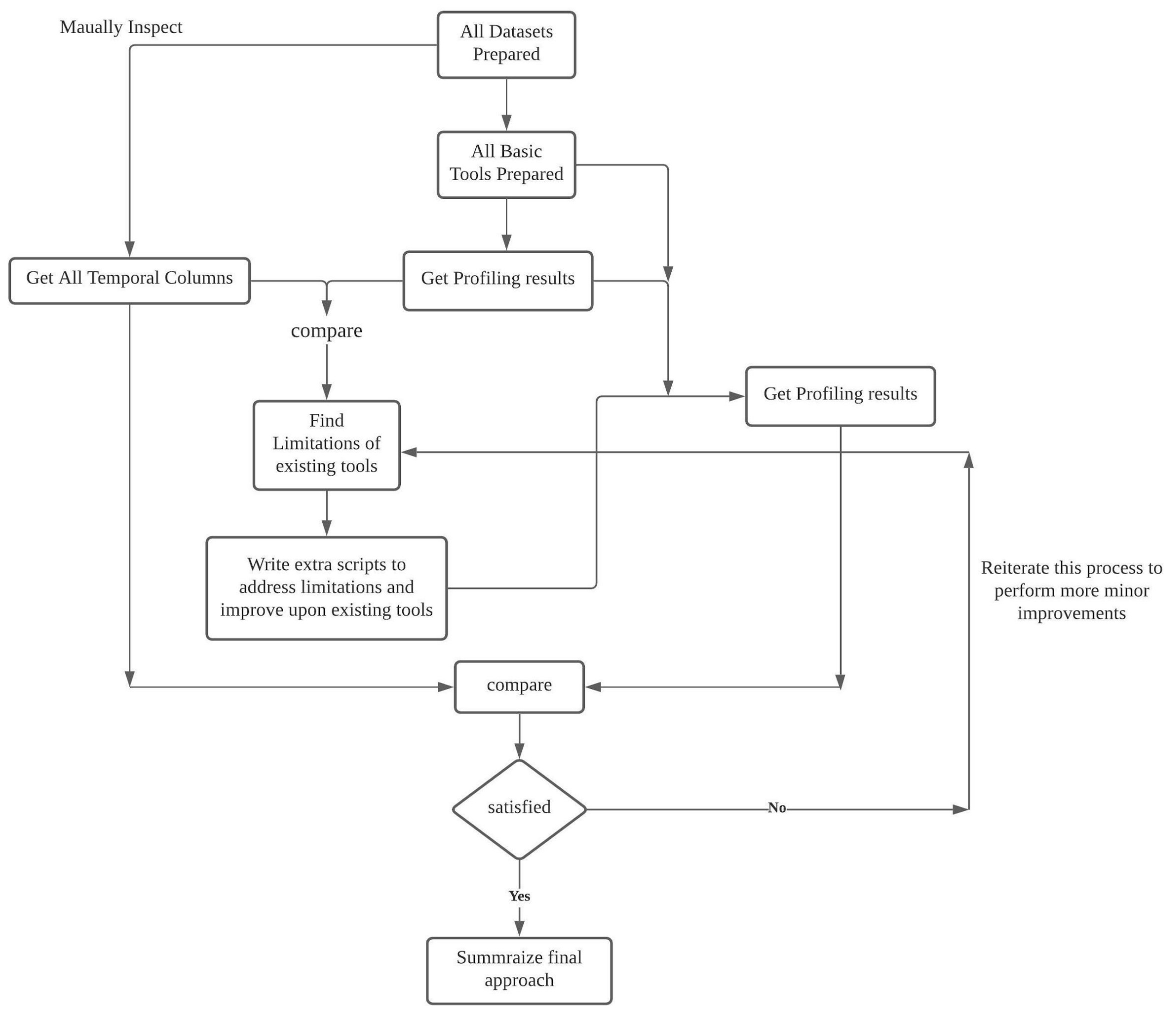
Assembling the data was a two-step process. The files were downloaded with our downloader.py script and saved to HFS in the directories - {/home/jb7259/top100} and {/home/jb7259/top100_small} - without issue. However, when trying to run our parsing scripts with them, we discovered that the spark csv reader could not handle the utf-8 characters that were present in nineteen of the sets. We resolved that problem by running them through a second script, converter.py, to convert them to ascii.

## Profiling Tools

We selected Datamart and Dateutil as our two baseline profiling tools. Datamart is a dataset search engine and data augmentation platform developed at New York University. Per the documentation, It can be used to index the content of datasets from a variety of sources, that can later be queried to find data that can be joined or appended to a user's data. The Dateutil module provides powerful extensions to python's standard datetime module, which can help detect temporal patterns in a more flexible manner.

# Architecture Design

We proposed to process these datasets according to the following flowchart. One more thing we are required to accomplish with our datasets is to tag every temporal column across all of them. Manual inspection is our only reliable way to derive the result. After comparing test results from different profilers, we hope to pinpoint the reason for which a profiler fails in every case,and construct new scripts to address the problems revealed. We intend this to be an iterative process because enhanced profilers have the potential to bring in new problems that are absent with basic profilers. Multiple rounds of minor adjustments are much required. In the end, we want the process to halt at states where further improvements cause our enhanced profilers to overfit our training datasets.



## Manual Review

Our first step was to manually review the datasets. This revealed the following:

**Summary Statistics**

| | |
|---|---|
| Sets containing some form of temporal data | 81 |
| Total number of temporal attributes | 274 |
| Total number of different temporal data formats | 20 |

*Table 1*

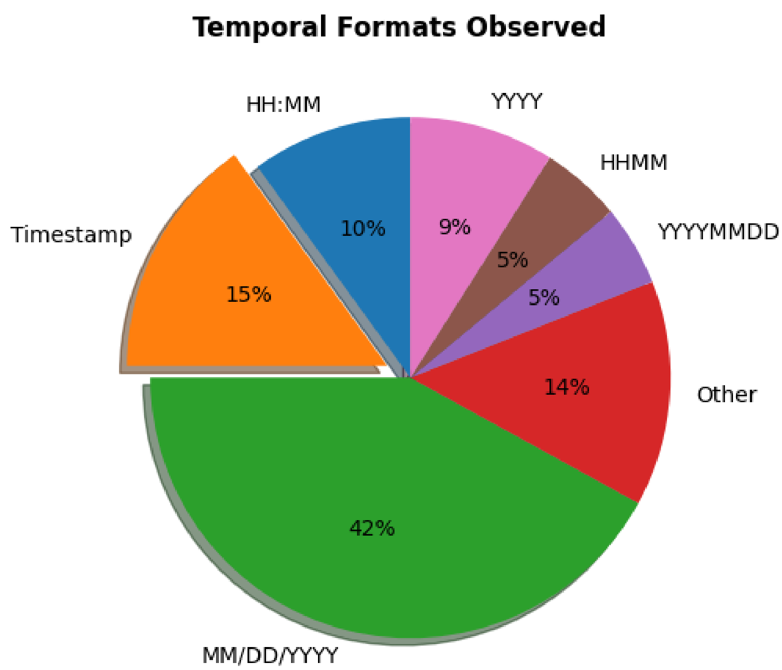



*Figure 1*

As noted in Table 1, nineteen of the sets contained no temporal data whatsoever. This reinforced for us the idea that a tool like ours could prove useful. If our end users were able to ignore nearly twenty percent of a given collection of data sets after preprocessing them with our approach, that would yield a massive gain in efficiency. Of course we cannot assume that the twenty percent number would hold across different collections, but it is a promising data point.

Table 1 also notes that there were a total of twenty different temporal formats observed within the collection. At a glance, this suggests that it could be extremely difficult to develop a general-purpose tool that catches all cases. However, Figure 1, shows that eighty-six percent of the temporal attributes within the collection were accounted for by merely six different formats.

This suggests that while a perfect tool might be difficult to achieve, a relatively simple tool could still deliver a significant impact.

The manual review also raised some questions. First and foremost was the question about what to categorize as a temporal value. Consider the “DOHMH New York City Inspection Results” (the twentieth most popular set) as an example. It includes date attributes for when the inspection took place and for when the grade was given. Both of those attributes tell a story through time and could provide additional helpful information for a scientist trying to solve a big data problem. Clearly they fit our mold.

However, this set also includes a third date attribute called record_date. This attribute is meant to reflect when the set was processed, and since it is processed every day, this attribute is always set to the current day’s date for every entry in the set. It matches the inspection and grade attributes in its format (MM/DD/YYYY), but it does not tell a story through time. So should it be considered a temporal attribute?

Questionable attributes were common throughout the sets under consideration. A year value is clearly temporal when it reflects the birth year of a child in the “Popular Baby Names” dataset (forty-second most popular), but not so clearly temporal when it reflects the manufacture date of a ticketed car in the “Parking Violations Issued - Fiscal Year 2021” dataset (forty-third most popular). The question, then, was where to draw the line. Is it better to rule out potentially helpful datasets or to include extra datasets in the final population? In other words, which is it more useful to minimize: false negatives or false positives?

We opted to minimize false negatives. Our reasoning was that our tool should first and foremost do no harm to the data scientist’s process. Without our tool, she would need to go through every dataset in the collection. If our preprocessing step reduces that number even by a small percentage, it helps. If it doesn’t reduce that number at all, it does no harm. If, on the other hand, it rules out a dataset that would prove useful, it ends up as a net negative.

# Methods

## Method 1: Datamart Profiler

As discussed above, the Datamart Profiler tool is a fully featured profiler. It takes a data set in .CSV or .JSON format as input and produces a .JSON file with all of the temporal data it inferred. Ideally, this would be run on the HPC cluster, but some limitations with the required libraries made that impossible. Instead, we ran it locally.

## Method 2: Dateutil

As discussed above, the dateutil package includes a powerful parser. Our approach was to run every cell in each dataset through this parser and then categorize it based on whether or not it parsed successfully. We used a python script (main.py) and Apache spark on the HPC cluster. We made two passes through the data. The first pass parsed the cell using Dateutil, and stored a 1 if the parse was successful, and a 0 if it was not. The second pass folded the rows together, generating a row of column sums. The sums were divided by the total number of rows, producing a percentage. To account for data quality issues, we set a threshold of 98%. Any column that exceeded that threshold was deemed a temporal value.

## Method 3: Column Name RegEx

During our manual review, we noted that a significant number of the temporal column names included key words or abbreviations that signified temporal data. The main ones: date, time, hour, year, month, dt, tm, yr. Our approach, then, was to use string matching with the header file to identify these columns. We used a python script (main_re.py) and Apache spark on the HPC cluster. In a single pass through the data, we captured the header and compared each of its strings to our collection of words. Any match was deemed a temporal value.

## Method 4: Enhanced Datamart Profiler

Failed cases with the standard Datamart Profiler revealed the following limitations with its approach:

1) It only tags a temporal column if the data represents a specific moment in time. For instance, 10-22-2020 or other alternative format is regarded as a specific moment while just October 22nd doesn’t since the year of this timestamp isn’t specified. It’s a valid assumption to make but there are many cases where the original dataset separates a complete timestamp into multiple parts like year,month,day or date,time of day. In this case, the month, day parts or time of day part will be ignored by temporal detection even though it’s still a valid temporal attribute. We need to bypass this restraint and add another layer of regular expression check to re-tag pieces of a complete timestamp temporal again.

2) It prioritizes numeric value detection over temporal detection. If a column contains only year or month data, or a complete date is in numeric format(eg. 10222020 represents October 22nd,2020), it'll be tagged as "Integer" and temporal detection steps are automatically skipped . We need to make sure that Integer columns will be tested against one layer of regular expression check.
3) Some extreme edge cases that contain abnormal temporal format. Like a period of time will be represented as 2005-2007, time of day is accompanied by AM/PM or A/P for example. We need to tackle each edge case and write extra checkups for them in order for them to be detected in the future.

Our enhanced version of Datamart will add layers of regular expression checking on top of the original return values from Datamart. For each column in the dataset, we will perform the following extra checks:

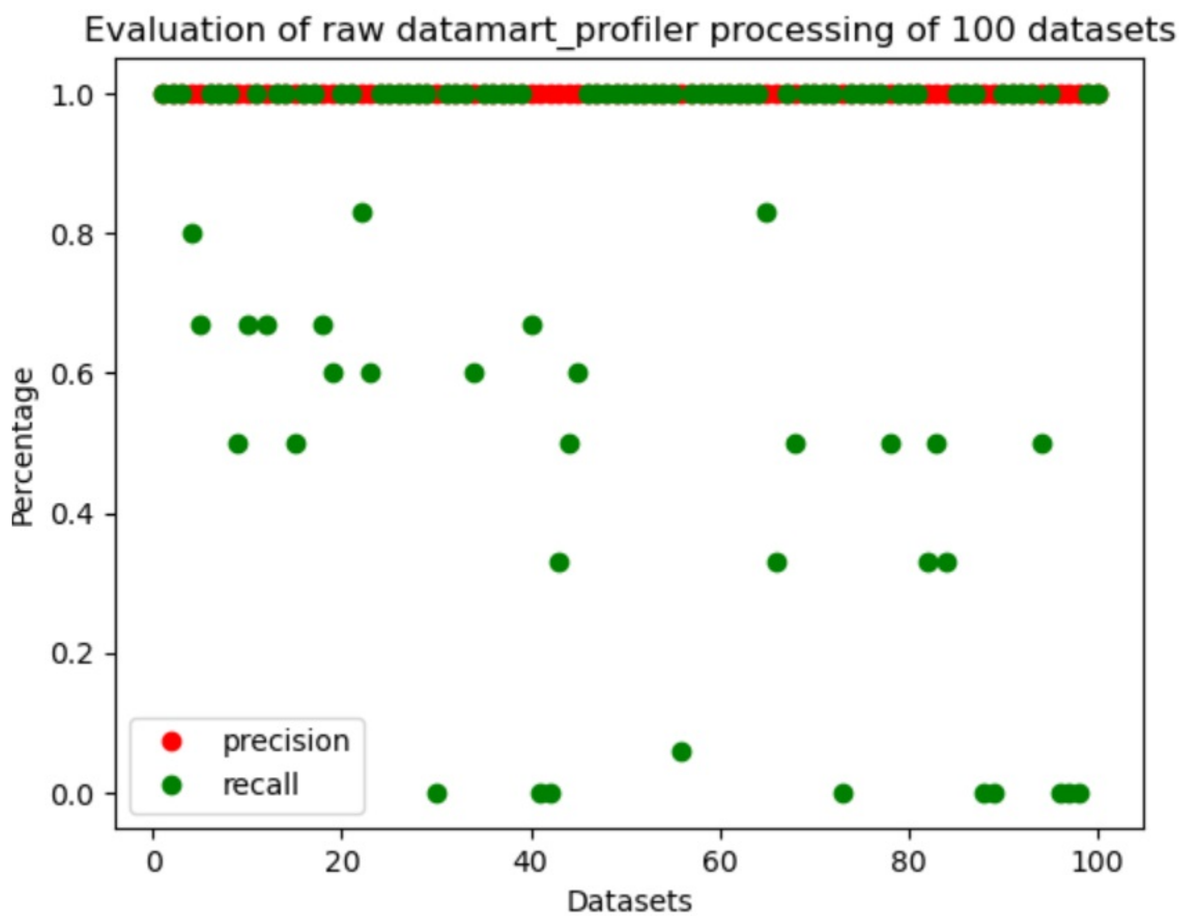
1) For each cell, we will check using a regular expression or time.strptime to detect whether this data is a month,year,time of day with AM/PM or without or a datetime category.
2) For the entire column, if a certain threshold of the columns fall in the above mentioned categories, we'll tag the column as temporal.

# Results

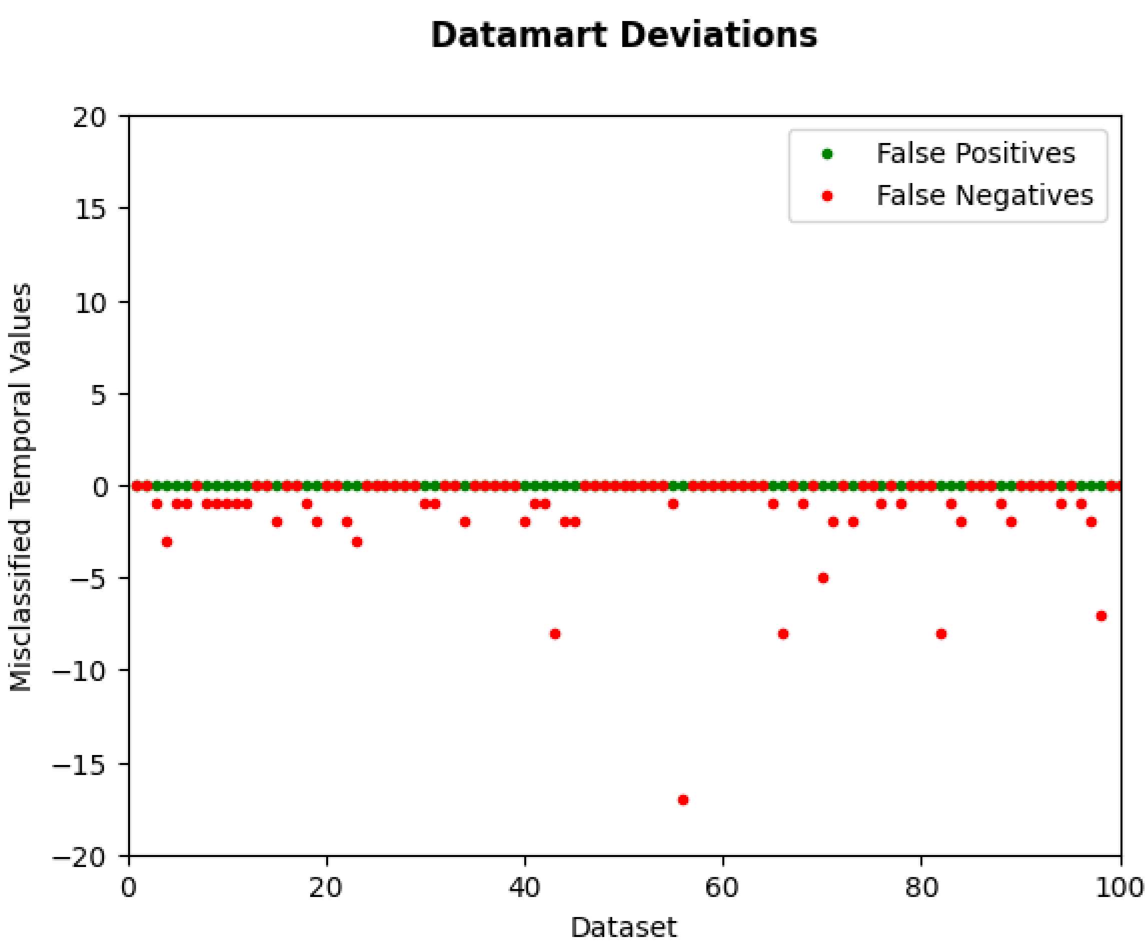
## Method 1: Datamart Profiler

For our first pass with Datamart, we ran the existing tool as is, and then compared its output to our manual analysis of the temporal data. The results are summarized in the charts below. Most notably, it achieved perfect precision - every column it identified as temporal was indeed temporal. However, the recall was not perfect.

Observation of this evaluation illustrates the fact that datamart is a relatively passive temporal attributes profiler. Precision values remain at one indicates a reliable avoidance of false positive cases. Datamart enforces various kind of constraints on detection algorithm like requiring a specific moment in history and numeric column escaping temporal checking. They're responsible for the majority of the false negative cases here. Further relaxation of the constraints are much required for improvements and datamart provide a solid lower bound for general temporal detection.



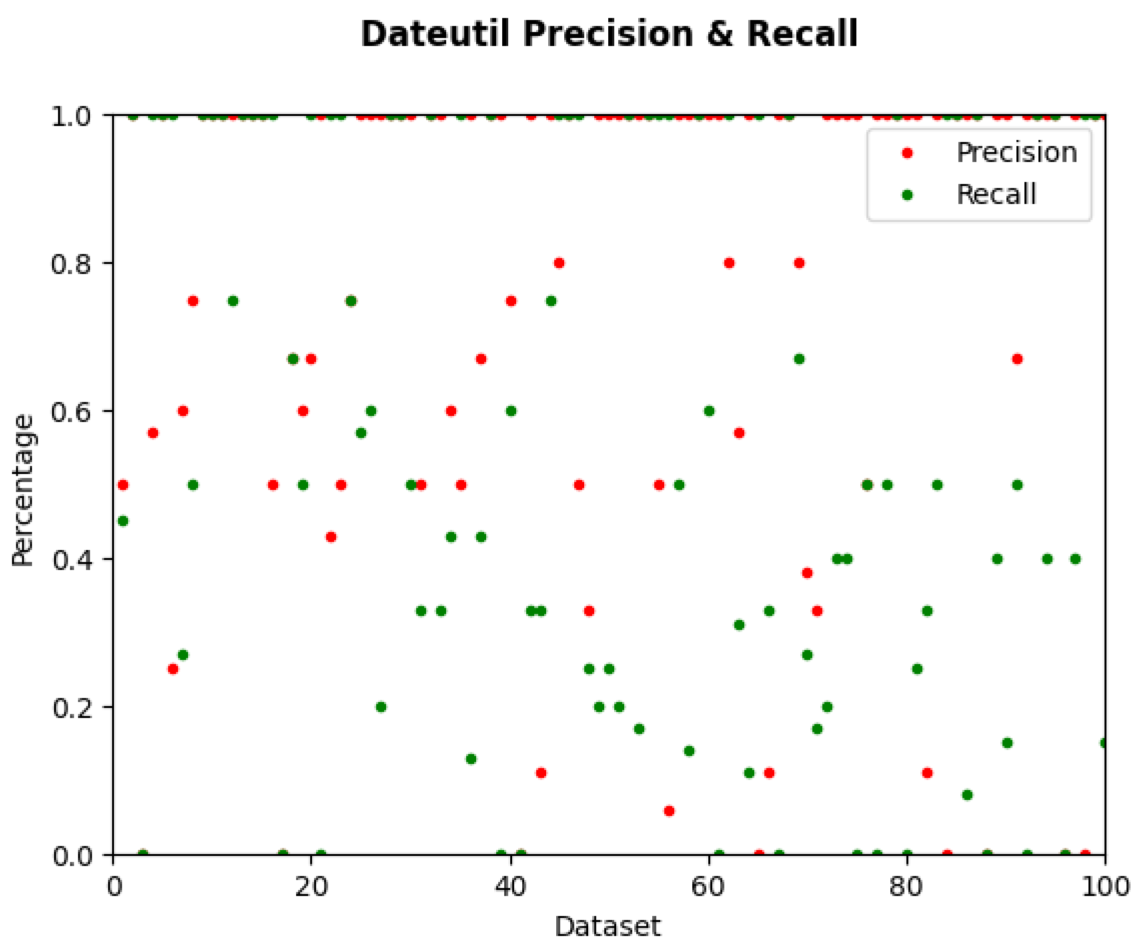
Displaying the data in terms of discrete false positive and false negative values as shown below illustrates more clearly the opportunities for improvement. Every bit of space between the red dots and the x axis represents a missed temporal attribute. With our other proposed methods, we will attempt to close those gaps.



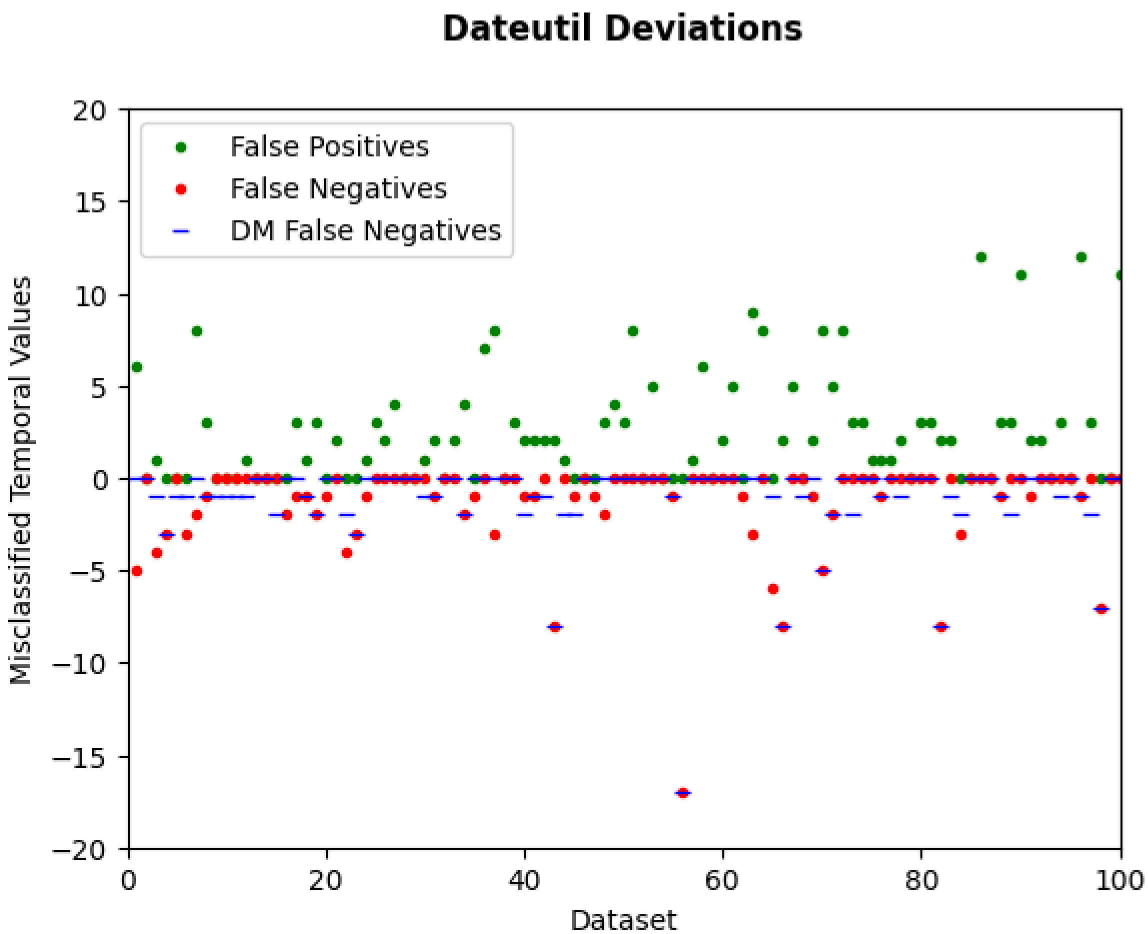
## Method 2: Dateutil

The goal of the Dateutil approach was to improve recall (reduce the number of false negatives) by keeping the parsing as general as possible. We expected this to lead to reduced precision (increased false positives), but the hope was that the tradeoff would be justified. Unfortunately, this did not work out particularly well. As can be seen in the figure below, the precision plummeted as compared to Datamart, and the recall was similar.

The biggest challenge for the Dateutil approach to overcome is the fact that numeric values between -9999 and 9999 parse successfully. They return dates using that number to determine the year and using the current month and day. It is possible to use filtering to overcome this limitation, but the risk then is overfitting the test datasets.



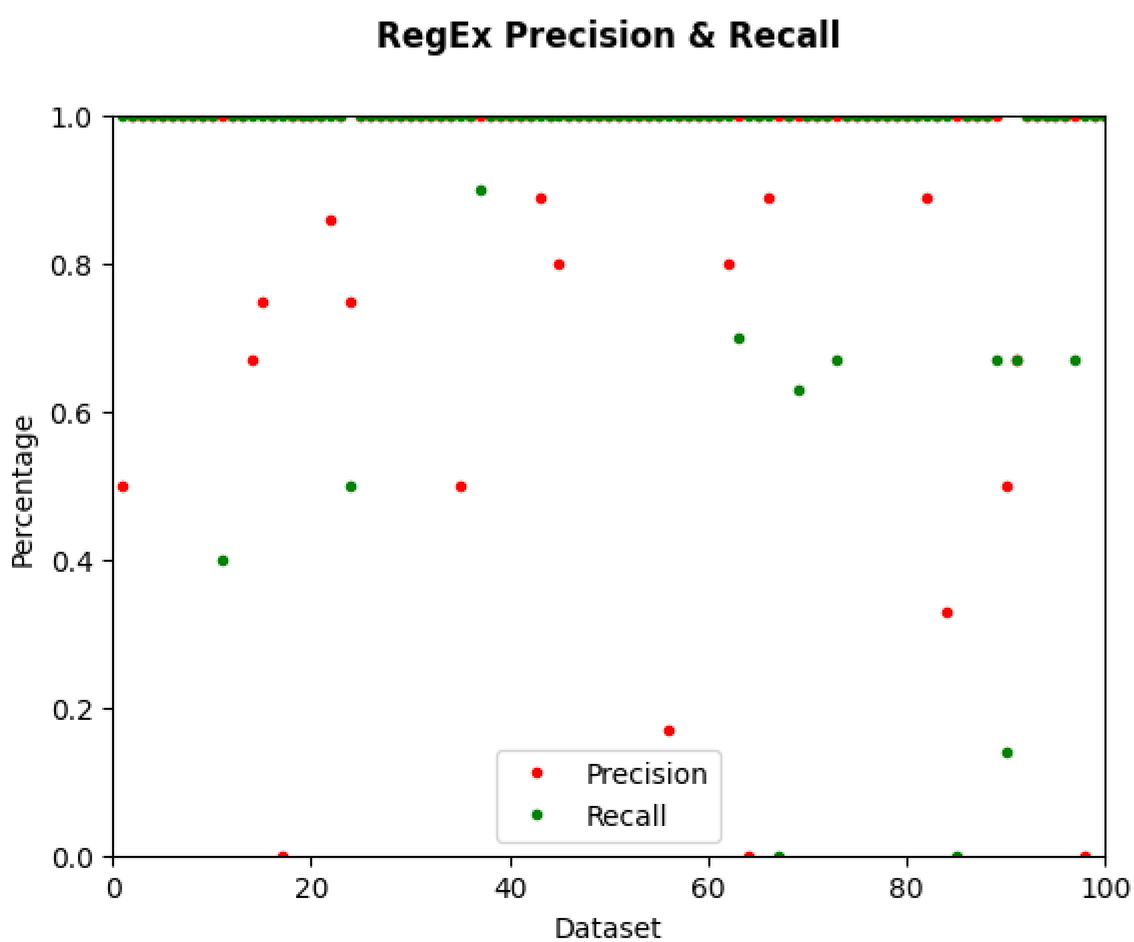
Dateutil successfully identified some attributes which Datamart did not, most notably times and years. This led to a reduced number of false negatives in a few specific cases, as can be seen in the figure below. In addition, the Dateutil method correctly characterized fourteen out of the nineteen temporal-free sets as such. It also produced only three false negatives on a set-wise basis, meaning all but three sets that included some temporal information were correctly identified as such despite the fact that some individual attributes were missed. Based on these results, this approach would provide some value to a scientist working with this particular collection. However, given the limitation with numeric values stated above, a pure Dateutil approach does not provide the optimal mix of complexity and performance.



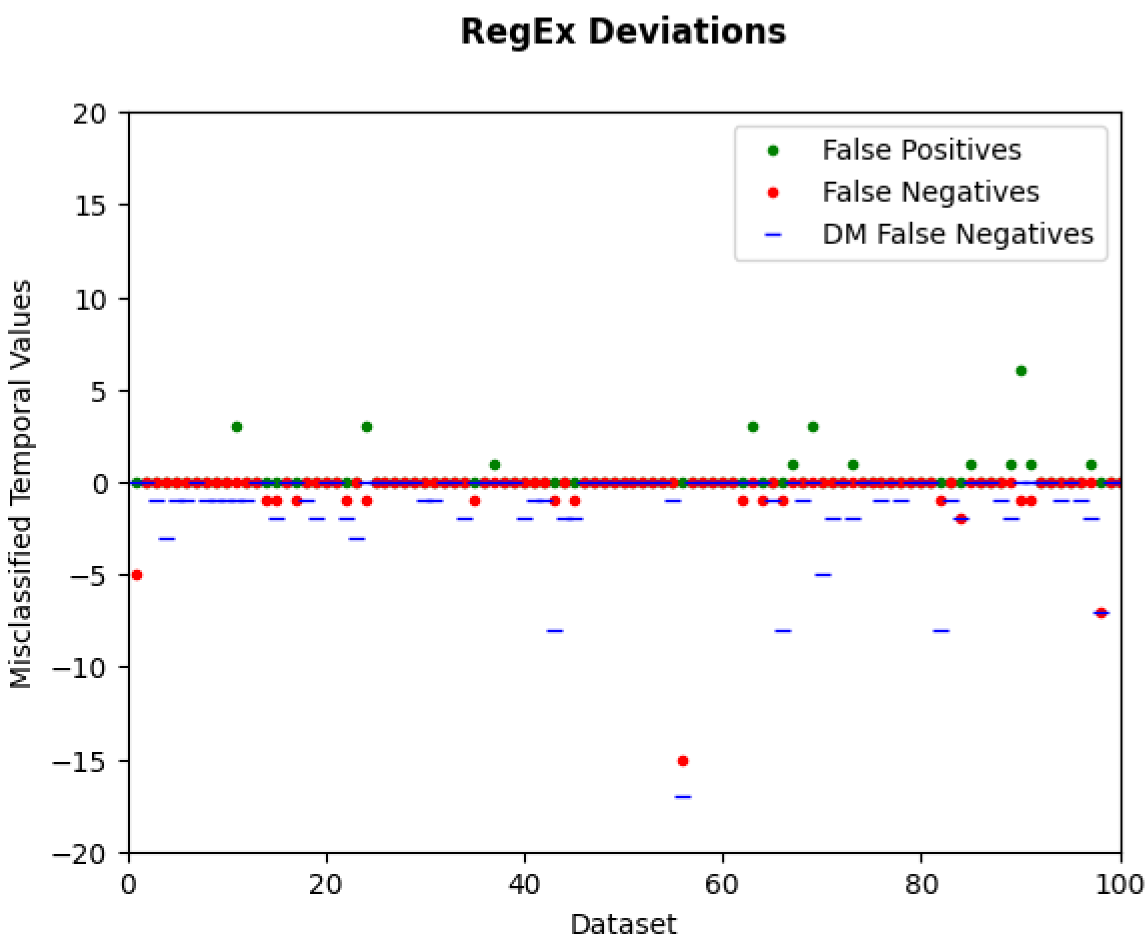
## Method 3: Column Name RegEx

The goal of the RegEx approach was to improve recall (reduce the number of false negatives) by taking advantage of the commonality of different strings in temporal column names. We expected this to lead to reduced precision (increased false positives) as compared to Datamart due to the possibility of our target strings being found in non-temporal column names as well. As can be seen in the figure below, the precision could not match that of Datamart, but the recall was better.

The biggest challenge for a string-matching approach to overcome is the prevalence of abbreviations in attribute names. Full words like “date” and “time” are uncommon outside of their singular use cases, but abbreviations like “dt” and “tm” can be found more readily in other places. For instance, “addtl_info” was an attribute name in the 2018 DOE High School Directory set (97th most popular), which led to a false positive. Increasing the number of abbreviations considered helps maximize recall, but the tradeoff is a reduction in precision.



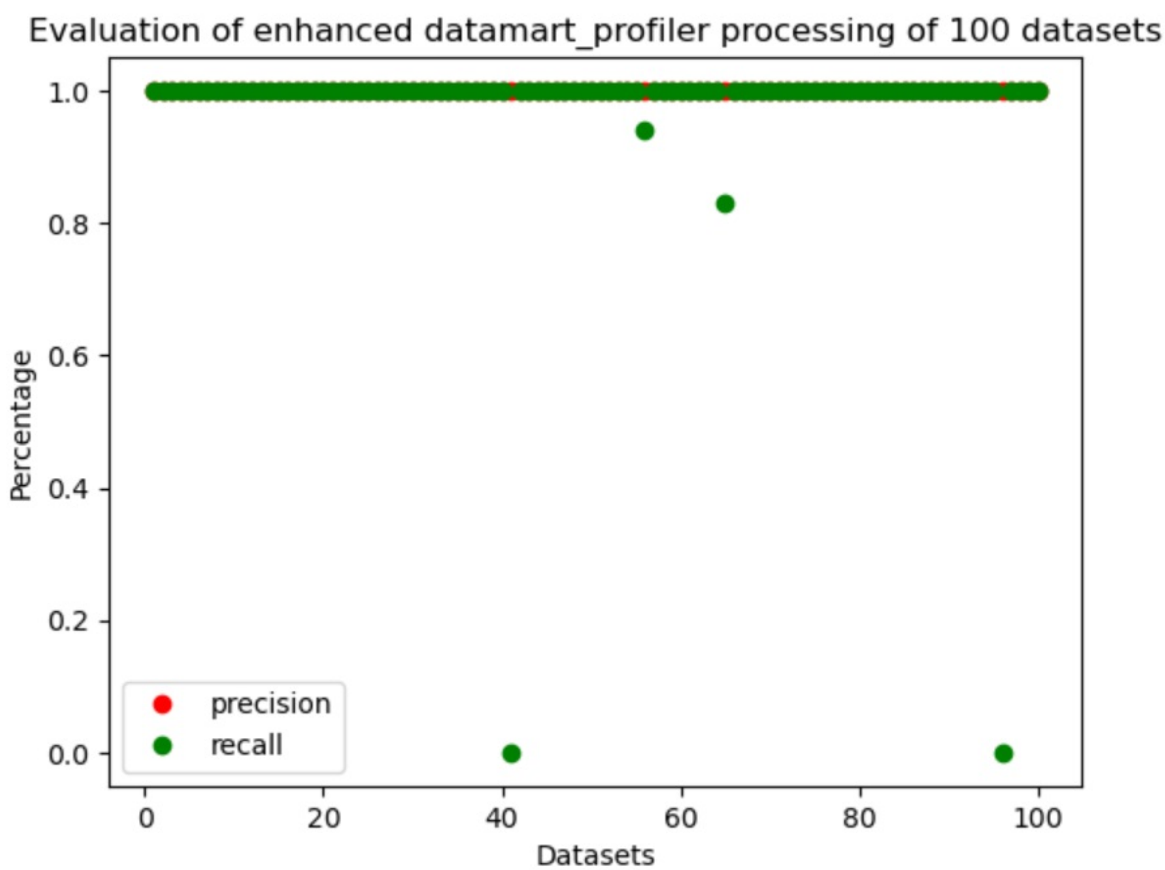
The RegEx approach successfully identified many temporal attributes that Datamart missed. As can be seen in the figure below, it matched or bettered Datamart in virtually every set. False positives are virtually inevitable with an approach like this, but given that our goal is to minimize false negatives, the tradeoff seems to justify them.

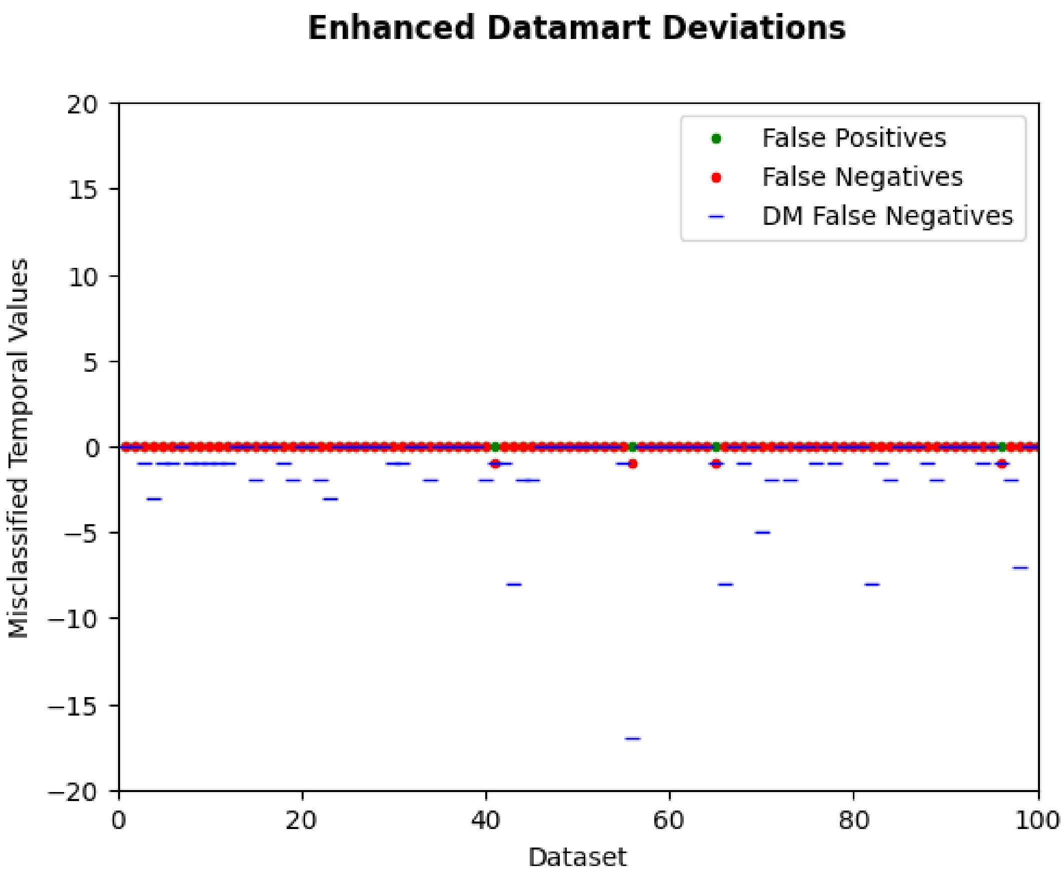


## Method 4: Enhanced Datamart Profiler

After making the adjustments discussed in the previous datamart section, we saw a massive improvement in the recall with no loss of precision. As illustrated in the charts below, this approach resolved all but a handful of false negatives. The combination of extra layers of checking with the base datamart profiler brings in no false positive cases, which proved to be quite robust.

The remaining false negatives are all special edge cases that require specific regular expression to handle. This is the point where the enhanced datamart profiler starts to reach its limit and pend towards overfitting the data. Every step of enhancement remains passive and is derived from a new edge case. The challenge of improving upon it with a more generalized and random sets of datasets lie in infinite number of possible forms of expression in temporal attributes and noises in data. But with our high emphasis on low false positives, enhanced datamart profiler is tailored to our needs.





Enhanced Datamart Deviations
False Positives
False Negatives
DM False Negatives
Misclassified Temporal Values
Dataset
20
15
10
5
0
−5
−10
−15
−20
0
20
40
60
80
100

# Conclusions

Temporal analysis is challenging. For example, the number 1999 could represent a year, a price, or almost any other numeric value. Without metadata, there is no way to know for certain what type of output was intended.

However, if the dataset is large enough, there are techniques that can give us a good idea of what we are dealing with. By parsing individual cells, string matching column names, and running broader regular expressions and statistical analysis across the columns as a whole, we can deduce with a high degree of accuracy whether or not an attribute is temporal. Armed with this knowledge, a data scientist can make an informed decision about whether or not to pursue further analysis on a given dataset, making for a more productive use of her time.

# References

Datamart Profiler Homepage: https://pypi.org/project/datamart-profiler/
Dateutil Documentation: https://dateutil.readthedocs.io/en/stable/
NYU Big Data Fall 2020 Week 1 Class Presentation

# APPENDIX

# Appendix I: Manual Review Results

| RANK | Dataset Name | API ID | MM/DD/YYYY | T/S | HH:MM | Year | HHMMA | YYYYMMDD | HH:MM:SS | Empty | DD-MON-YYYY | MM | Month | Month/Year | Date DoW | Month day, year | YYYY-YYYY | xx,xxx,xxx | DAYS | Seq/rg of times | MMDDYYYY | YYYY-MM-DD | Temp Cols Obs. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | DOB Job Application Filings | "ic3t-wcy2" | 10 | | | | | | | | | | | | | | | | | | | | 10 |
| 2 | TLC New Driver Application Status | "dpec-ucu7" | 1 | 1 | | | | | | | | | | | | | | | | | | | 2 |
| 3 | Civil Service List (Active) | "vx8i-nprf" | 4 | | | | | | | | | | | | | | | | | | | | 4 |
| 4 | For Hire Vehicles (FHV) - Active | "8wbx-tsch" | 4 | | 1 | 1 | | | | 1 | | | | | | | | | | | | | 7 |
| 5 | For Hire Vehicles (FHV) - Active Drivers | "xjfq-wh2d" | 2 | | 1 | | | | | | | | | | | | | | | | | | 3 |
| 6 | 311 Service Requests from 2010 to Present | "erm2-nwe9" | | 4 | | | | | | | | | | | | | | | | | | | 4 |
| 7 | DOB Permit Issuance | "ipu4-2q9a" | 5 | | | | | | | | | | | | | | | | | | | | 5 |
| 8 | Civil Service List Certification | "a9md-ynri" | 3 | | | | | | | 1 | | | | | | | | | | | | | 4 |
| 9 | Motor Vehicle Collisions - Crashes | "h9gi-nx95" | 1 | | 1 | | | | | | | | | | | | | | | | | | 2 |
| 10 | Medallion Drivers - Active | "jb3k-j3gp" | 2 | | 1 | | | | | | | | | | | | | | | | | | 3 |
| 11 | Citywide Payroll Data (Fiscal Year) | "k397-673e" | 1 | | | 1 | | | | | | | | | | | | | | | | | 2 |
| 12 | Street Hail Livery (SHL) Drivers - Active | "5tub-eh45" | 2 | | 1 | | | | | | | | | | | | | | | | | | 3 |
| 13 | New York City Leading Causes of Death | "jb7j-dtam" | | | | 1 | | | | | | | | | | | | | | | | | 1 |
| 14 | Film Permits | "tg4x-b46p" | | 3 | | | | | | | | | | | | | | | | | | | 3 |
| 15 | Medallion Vehicles - Authorized | "rhe8-mgbb" | 2 | | 1 | 1 | | | | | | | | | | | | | | | | | 4 |
| 16 | City Record Online | "dg92-zbpx" | 2 | 2 | | | | | | | | | | | | | | | | | | | 4 |
| 17 | Active Projects Under Construction | "8586-3zfm" | 1 | | | | | | | | | | | | | | | | | | | | 1 |
| 18 | Open Parking and Camera Violations | "nc67-uf89" | 2 | | 1 | | | | | | | | | | | | | | | | | | 3 |
| 19 | NYPD Complaint Data Historic | "qgea-i56i" | 3 | | | | | | 2 | | | | | | | | | | | | | | 5 |
| 20 | DOHMH New York City Restaurant Inspection Results | "43nn-pn8j" | 3 | | | | | | | | | | | | | | | | | | | | 3 |
| 21 | NYC Civil Service Titles | "nzjr-3966" | - | | | | | | | | | | | | | | | | | | | | 0 |
| 22 | Street Hail Livery (SHL) Permits | "yhuu-4pt3" | 5 | | 1 | 1 | | | | | | | | | | | | | | | | | 7 |

| | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 23 | OATH Hearings Division Case Status | "jz4z-kudi" | 4 | | | | | | 2 | | | | | | | | | | | | | | 6 |
| 24 | NYC Jobs | "kpav-sd4t" | 3 | | | | | | | | 1 | | | | | | | | | | | | 4 |
| 25 | DOB Complaints Received | "eabe-havv" | 3 | 1 | | | | | | | | | | | | | | | | | | | 4 |
| 26 | DOB NOW: Build – Approved Permits | "rbx6-tga4" | | 3 | | | | | | | | | | | | | | | | | | | 3 |
| 27 | NYPD Arrest Data (Year to Date) | "uip8-fykc" | 1 | | | | | | | | | | | | | | | | | | | | 1 |
| 28 | Demographic Statistics By Zip Code | "kku6-nxdu" | | | | | | | | | | | | | | | | | | | | | 0 |
| 29 | FDNY Firehouse Listing | "hc8x-tcnd" | | | | | | | | | | | | | | | | | | | | | 0 |
| 30 | Civil List | "ye3c-m4ga" | | | | 1 | | | | | | | | | | | | | | | | | 1 |
| 31 | DOB Violations | "3h2n-5cm9" | | | | | | 2 | | | | | | | | | | | | | | | 2 |
| 32 | 2012 SAT Results | "f9bf-2cp4" | | | | | | | | | | | | | | | | | | | | | 0 |
| 33 | Water Consumption In The New York City | "ia2d-e54m" | | | | 1 | | | | | | | | | | | | | | | | | 1 |
| 34 | NYPD Complaint Data Current (Year To Date) | "5uac-w243" | 3 | | | | | | 2 | | | | | | | | | | | | | | 5 |
| 35 | Legally Operating Businesses | "w7w3-xahh" | 2 | | | | | | | | | | | | | | | | | | | | 2 |
| 36 | COVID-19 Daily Counts of Cases, Hospitalizations, and Deaths | "rc75-m7u3" | 1 | | | | | | | | | | | | | | | | | | | | 1 |
| 37 | Housing Maintenance Code Violations | "wvxf-dwi5" | 9 | | | | | | | | | | | | | | | | | | | | 9 |
| 38 | New York City Population by Borough, 1950 - 2040 | "xywu-7bv9" | | | | | | | | | | | | | | | | | | | | | 0 |
| 39 | DOF Parking Violation Codes | "ncbg-6agr" | | | | | | | | | | | | | | | | | | | | | 0 |
| 40 | DOB ECB Violations | "6bgk-3dad" | | | | | 1 | 3 | | | | | | | | | | | | | | | 4 |
| 41 | Air Quality | "c3uy-2p5r" | | | | 1 | | | | | | | | | | | | | | | | | 1 |
| 42 | Popular Baby Names | "25th-nujf" | | | | 1 | | | | | | | | | | | | | | | | | 1 |
| 43 | Parking Violations Issued - Fiscal Year 2021 | "pvqr-7yc4" | 1 | | | 1 | 4 | 1 | | | | | | | | | | 1 | 1 | | | | 9 |
| 44 | FHV Base Aggregate Report | "2v9c-2k7f" | | | | 1 | | | | | | 1 | 1 | | | | | | | | | | 3 |
| 45 | Medallion Vehicles - Inactive | "jgtb-hmpg" | 3 | | 1 | 1 | | | | | | | | | | | | | | | | | 5 |
| 46 | Evictions | "6z8x-wfk4" | 1 | | | | | | | | | | | | | | | | | | | | 1 |

| | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 47 | Emergency Response Incidents | "pasr-j7fb" | | 2 | | | | | | | | | | | | | | | | | | | 2 |
| 48 | Housing New York Units by Building | "hg8x-zxpr" | 3 | | | | | | | | | | | | | | | | | | | | 3 |
| 49 | NYPD Arrests Data (Historic) | "8h9b-rp9u" | 1 | | | | | | | | | | | | | | | | | | | | 1 |
| 50 | DHS Daily Report | "k46n-sa2m" | 1 | | | | | | | | | | | | | | | | | | | | 1 |
| 51 | 2018 Yellow Taxi Trip Data | "t29m-gskq" | | 2 | | | | | | | | | | | | | | | | | | | 2 |
| 52 | NYC Health + Hospitals patient care locations - 2011 | "f7b6-v6v3" | | | | | | | | | | | | | | | | | | | | | 0 |
| 53 | Tax Lien Sale Lists | "9rz4-mjek" | | | | | | | | | | | | 1 | | | | | | | | | 1 |
| 54 | Greenbook | "mdcw-n682" | | | | | | | | | | | | | | | | | | | | | 0 |
| 55 | TLC Approved LabCorp Patient Services Drug Test Locations (Dataset) | "pe54-wf39" | 1 | | 1 | | | | | | | | | | | | | | | | | | 2 |
| 56 | Open Streets Locations | "uiay-nctu" | | 1 | 14 | | | | | | | | | | | | | | 1 | 2 | | | 18 |
| 57 | New York City Population By Neighborhood Tabulation Areas | "swpk-hqdp" | | | | 1 | | | | | | | | | | | | | | | | | 1 |
| 58 | DOB Certificate Of Occupancy | "bs8b-p36w" | 1 | | | | | | | | | | | | | | | | | | | | 1 |
| 59 | SAT (College Board) 2010 School Level Results | "zt9s-n5aj" | | | | | | | | | | | | | | | | | | | | | 0 |
| 60 | Property Data (Buildings Information System) | "e98g-f8hy" | | 3 | | | | | | | | | | | | | | | | | | | 3 |
| 61 | DOB NOW: Build – Job Application Filings | "w9ak-ipjd" | | | | | | | | | | | | | | | | | | | | | 0 |
| 62 | Trade Waste Hauler Licensees | "867j-5pgi" | 2 | | | | | | | | | | | | | | | | | | | 3 | 5 |
| 63 | EMS Incident Dispatch Data | "76xm-jjuj" | | 7 | | | | | | | | | | | | | | | | | | | 7 |
| 64 | 2015 Street Tree Census - Tree Data | "uvpi-gqnh" | 1 | | | | | | | | | | | | | | | | | | | | 1 |
| 65 | M/WBE, LBE, and EBE Certified Business List | "ci93-uc8s" | 6 | | | | | | | | | | | | | | | | | | | | 6 |
| 66 | Parking Violations Issued - Fiscal Year 2014 | "jt7v-77mi" | 1 | | | 1 | 4 | 2 | | | | | | | | | | | 1 | | | | 9 |
| 67 | IPIS (Integrated Property Information System) | "n5mv-nfpy" | | | | | | | | | | | | | | | | | | | | | 0 |
| 68 | CURRENT BASES | "eccv-9dzr" | 1 | | | | | | 1 | | | | | | | | | | | | | | 2 |
| 69 | Fire Incident Dispatch Data | "8m42-w767" | | 5 | | | | | | | | | | | | | | | | | | | 5 |

| | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 70 | Property Exemption Detail | "muvi-b6kx" | 1 | | | 5 | | 2 | | | | | | | | | | | | | | | 8 |
| 71 | Suitability of City-Owned and Leased Property for Urban Agriculture (LL 48 of 2011) | "4e2n-s75z" | 1 | | | 2 | | | | | | | | | | | | | | | | | 3 |
| 72 | 2017 Yellow Taxi Trip Data | "biws-g3hs" | | 2 | | | | | | | | | | | | | | | | | | | 2 |
| 73 | J-51 Exemption and Abatement | "y7az-s7wc" | | | | 2 | | | | | | | | | | | | | | | | | 2 |
| 74 | Multiple Dwelling Registrations | "tesw-yqqr" | 2 | | | | | | | | | | | | | | | | | | | | 2 |
| 75 | New York City Population By Community Districts | "xi7c-iu2" | | | | | | | | | | | | | | | | | | | | | 0 |
| 76 | Daily Inmates In Custody | "7479-ugqb" | | 1 | | | | | | 1 | | | | | | | | | | | | | 2 |
| 77 | 2010 - 2011 School Attendance and Enrollment Statistics by District | "7z8d-msnt" | | | | | | | | | | | | | | | | | | | | | 0 |
| 78 | NYPD Shooting Incident Data (Historic) | "833y-fsy8" | 1 | | | | | | 1 | | | | | | | | | | | | | | 2 |
| 79 | Traffic Volume Counts (2012-2013) | "p424-amsu" | 1 | | | | | | | | | | | | | | | | | | | | 1 |
| 80 | Directory Of Homebase Locations | "ntcm-2w4k" | | | | | | | | | | | | | | | | | | | | | 0 |
| 81 | DSNY Monthly Tonnage Data | "ebb7-mvp5" | | | | | | | | | | | | 1 | | | | | | | | | 1 |
| 82 | Parking Violations Issued - Fiscal Year 2016 | "kiv2-tbus" | 1 | | | 1 | 4 | 2 | | | | | | | | | | | 1 | | | | 9 |
| 83 | NYPD Shooting Incident Data (Year To Date) | "5ucz-vwe8" | 1 | | | | | | 1 | | | | | | | | | | | | | | 2 |
| 84 | Workforce1 Recruitment Events | "kf2b-aeh5" | | 2 | | | | | | | | | | | 1 | | | | | | | | 3 |
| 85 | Directory Of DHS Contacts | "cete-9g3v" | | | | | | | | | | | | | | | | | | | | | 0 |
| 86 | Current Reservoir Levels | "zkky-n5j3" | 1 | | | | | | | | | | | | | | | | | | | | 1 |
| 87 | FDNY Line Of Duty Deaths | "32y8-s55c" | | | | | | | | | | | | | | 1 | | | | | | | 1 |
| 88 | 2018 Central Park Squirrel Census - Squirrel Data | "vfnx-vebw" | | | | | | | | | | | | | | | | | | | 1 | | 1 |
| 89 | 2019 DOE High School Directory | "uq7m-95z8" | | | 2 | | | | | | | | | | | | | | | | | | 2 |
| 90 | NYCHA Development Data Book | "evjd-dqpz" | 2 | | | | | | | | | | | | | | | | | | | | 2 |
| 91 | DOHMH Childcare Center Inspections | "dsg6-ifza" | 3 | | | | | | | | | | | | | | | | | | | | 3 |

| | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 92 | Voting/Poll Sites | "mifw-tguq" | | | | | | | | | | | | | | | | | | | | | 0 |
| 93 | DYCD after-school programs: Beacon Programs | "35sw-rdxj" | | | | | | | | | | | | | | | | | | | | | 0 |
| 94 | Expense Budget | "mwzb-yiwb" | | | | 1 | | 1 | | | | | | | | | | | | | | | 2 |
| 95 | Public Recycling Bins | "sxx4-xhzg" | | | | | | | | | | | | | | | | | | | | | 0 |
| 96 | 2013 - 2018 Demographic Snapshot School | "s52a-8aq6" | | | | | | | | | | | | | | | 1 | | | | | | 1 |
| 97 | 2018 DOE High School Directory | "vw9i-7mzq" | | | 2 | | | | | | | | | | | | | | | | | | 2 |
| 98 | Queens Library Branches | "kh3d-xhq7" | | | | | | | | | | | | | | | | | | 7 | | | 7 |
| 99 | NYCHA Application Priority Codes | "2ei9-vg68" | | | | | | | | | | | | | | | | | | | | | 0 |
| 100 | 2016 Green Taxi Trip Data | "hvrh-b6nb" | | 2 | | | | | | | | | | | | | | | | | | | 2 |
| | **TOTALS** | | 116 | 41 | 28 | 25 | 13 | 13 | 9 | 3 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 4 | 9 | 1 | 3 | 274 |

# Appendix II: Script Outputs

| RANK | Dataset Name | T_Cols Observed | # T_Cols Obs | T_Cols Datamart | # T_Cols DM | T_Cols Dateutil | # T_Cols DU | T_Cols RegEx | # T_Cols RE |
|---|---|---|---|---|---|---|---|---|---|
| 1 | DOB Job Application Filings | latest_action_date, pre__filing_date, paid, fully_paid, assigned, approved, fully_permitted, dobrundate, signoff_date, special_action_date | 10 | latest_action_date, pre__filing_date, paid, fully_paid, assigned, approved, fully_permitted, dobrundate, signoff_date, special_action_date | 10 | ['doc__', 'lot', 'latest_action_date', 'community___board', 'pre__filing_date', 'paid', 'fully_paid', 'dobrundate', 'gis_latitude', 'gis_longitude', 'gis_council_district'] | 11 | ['latest_action_date', 'pre__filing_date', 'dobrundate', 'signoff_date', 'special_action_date'] | 5 |
| 2 | TLC New Driver Application Status | app_date, lastupdate | 2 | app_date, lastupdate | 2 | ['app_date', 'lastupdate'] | 2 | ['app_date', 'lastupdate'] | 2 |
| 3 | Civil Service List (Active) | extension_date, published_date, established_date, anniversary_date | 4 | published_date, established_date, anniversary_date | 3 | ['exam_no'] | 1 | ['published_date', 'established_date', 'anniversary_date', 'extension_date'] | 4 |
| 4 | For Hire Vehicles (FHV) - Active | license_type, vehicle_year, last_time_updated, expiration_date, certification_date, hack_up_date, last_date_updated | 7 | expiration_date, certification_date, hack_up_date, last_date_updated | 4 | ['expiration_date', 'vehicle_year', 'last_date_updated', 'last_time_updated'] | 4 | ['expiration_date', 'certification_date', 'hack_up_date', 'vehicle_year', 'order_date', 'last_date_updated', 'last_time_updated'] | 7 |
| 5 | For Hire Vehicles (FHV) - Active Drivers | last_time_updated, expiration_date, last_date_updated | 3 | expiration_date, last_date_updated | 2 | ['expiration_date', 'last_date_updated', 'last_time_updated'] | 3 | ['expiration_date', 'last_date_updated', 'last_time_updated'] | 3 |
| 6 | 311 Service Requests from 2010 to Present | due_date, created_date, closed_date, resolution_action_updated_date | 4 | created_date, closed_date, resolution_action_updated_date | 3 | ['created_date'] | 1 | ['created_date', 'closed_date', 'due_date', 'resolution_action_updated_date'] | 4 |
| 7 | DOB Permit Issuance | filing_date, issuance_date, expiration_date, job_start_date, dobrundate | 5 | filing_date, issuance_date, expiration_date, job_start_date, dobrundate | 5 | ['job_doc___', 'lot', 'community_board', 'bldg_type', 'permit_sequence__', 'filing_date', 'job_start_date', 'dobrundate', 'gis_latitude', 'gis_longitude', 'gis_council_district'] | 11 | ['filing_date', 'issuance_date', 'expiration_date', 'job_start_date', 'dobrundate'] | 5 |
| 8 | Civil Service List Certification | reissue_date, request_date, cert_date, cert_expiration_date | 4 | request_date, cert_date, cert_expiration_date | 3 | ['exam_no', 'list_agency_code', 'request_date', 'cert_date', 'cert_expiration_date', 'no_vacancies'] | 6 | ['request_date', 'cert_date', 'reissue_date', 'cert_expiration_date'] | 4 |
| 9 | Motor Vehicle Collisions - Crashes | crash_time, crash_date | 2 | crash_date | 1 | ['crash_date', 'crash_time'] | 2 | ['crash_date', 'crash_time'] | 2 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 10 | Medallion Drivers - Active | last_updated_time, expiration_date, last_updated_date | 3 | expiration_date, last_updated_date | 2 | ['expiration_date', 'last_updated_date', 'last_updated_time'] | 3 | ['expiration_date', 'last_updated_date', 'last_updated_time'] | 3 |
| 11 | Citywide Payroll Data (Fiscal Year) | fiscal_year, agency_start_date | 2 | agency_start_date | 1 | ['fiscal_year', 'agency_start_date'] | 2 | ['fiscal_year', 'payroll_number', 'agency_start_date', 'regular_hours', 'ot_hours'] | 5 |
| 12 | Street Hail Livery (SHL) Drivers - Active | last_update_time, expiration_date, last_update_date | 3 | expiration_date, last_update_date | 2 | ['status_code', 'expiration_date', 'last_update_date', 'last_update_time'] | 4 | ['expiration_date', 'last_update_date', 'last_update_time'] | 3 |
| 13 | New York City Leading Causes of Death | year | 1 | year | 1 | ['year'] | 1 | ['year'] | 1 |
| 14 | Film Permits | startdatetime, enddatetime, enteredon | 3 | startdatetime, enddatetime, enteredon | 3 | ['startdatetime', 'enddatetime', 'enteredon'] | 3 | ['startdatetime', 'enddatetime'] | 2 |
| 15 | Medallion Vehicles - Authorized | model_year, last_updated_time, type, last_updated_date | 4 | type, last_updated_date | 2 | ['type', 'model_year', 'last_updated_date', 'last_updated_time'] | 4 | ['model_year', 'last_updated_date', 'last_updated_time'] | 3 |
| 16 | City Record Online | start_date, end_date, due_date, event_date | 4 | start_date, end_date, due_date, event_date | 4 | ['start_date', 'end_date'] | 2 | ['start_date', 'end_date', 'due_date', 'event_date'] | 4 |
| 17 | Active Projects Under Construction | data_as_of | 1 | data_as_of | 1 | ['geo_dist', 'community_board', 'community_council'] | 3 | [] | 0 |
| 18 | Open Parking and Camera Violations | violation_time, issue_date, judgment_entry_date | 3 | issue_date, judgment_entry_date | 2 | ['issue_date', 'violation_time', 'fine_amount'] | 3 | ['issue_date', 'violation_time', 'judgment_entry_date'] | 3 |
| 19 | NYPD Complaint Data Historic | cmplnt_fr_tm, cmplnt_to_tm, cmplnt_fr_dt, cmplnt_to_dt, rpt_dt | 5 | cmplnt_fr_dt, cmplnt_to_dt, rpt_dt | 3 | ['cmplnt_fr_dt', 'cmplnt_fr_tm', 'addr_pct_cd', 'rpt_dt', 'ky_cd', 'pd_cd'] | 6 | ['cmplnt_fr_dt', 'cmplnt_fr_tm', 'cmplnt_to_dt', 'cmplnt_to_tm', 'rpt_dt'] | 5 |
| 20 | DOHMH New York City Restaurant Inspection Results | inspection_date, grade_date, record_date | 3 | inspection_date, grade_date, record_date | 3 | ['inspection_date', 'record_date'] | 2 | ['inspection_date', 'grade_date', 'record_date'] | 3 |
| 21 | NYC Civil Service Titles | | 0 | | 0 | ['std_hrs', 'union_cd'] | 2 | [] | 0 |
| 22 | Street Hail Livery (SHL) Permits | license_type, certification_date, hack_up_date, vehicle_year, date_updated | 7 | license_type, certification_date, hack_up_date, vehicle_year, date_updated | 5 | ['license_type', 'date_updated', 'time_updated'] | 3 | ['certification_date', 'hack_up_date', 'vehicle_year', 'suspension_date', 'date_updated', 'time_updated'] | 6 |
| 23 | OATH Hearings Division Case Status | violation_date, hearing_date, decision_date | 6 | violation_date, hearing_date, decision_date | 3 | ['violation_date', 'hearing_date', 'hearing_time'] | 3 | ['violation_date', 'violation_time', 'hearing_date', 'hearing_time', 'decision_date', 'date_judgment_docketed'] | 6 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 24 | NYC Jobs | posting_date, post_until, posting_updated, process_date | 4 | posting_date, post_until, posting_updated, process_date | 4 | ['number_of_positions', 'posting_date', 'posting_updated', 'process_date'] | 4 | ['full_time_part_time_indicator', 'hours_shift', 'recruitment_contact', 'posting_date', 'posting_updated', 'process_date'] | 6 |
| 25 | DOB Complaints Received | date_entered, disposition_date, inspection_date, dobrundate | 4 | date_entered, disposition_date, inspection_date, dobrundate | 4 | ['date_entered', 'house_number', 'community_board', 'complaint_category', 'disposition_date', 'inspection_date', 'dobrundate'] | 7 | ['date_entered', 'disposition_date', 'inspection_date', 'dobrundate'] | 4 |
| 26 | DOB NOW: Build – Approved Permits | approved_date, issued_date, expired_date | 3 | approved_date, issued_date, expired_date | 3 | ['lot', 'c_b_no', 'approved_date', 'issued_date', 'expired_date'] | 5 | ['approved_date', 'issued_date', 'expired_date'] | 3 |
| 27 | NYPD Arrest Data (Year to Date) | arrest_date | 1 | arrest_date | 1 | ['arrest_date', 'pd_cd', 'ky_cd', 'arrest_precinct', 'latitude'] | 5 | ['arrest_date'] | 1 |
| 28 | Demographic Statistics By Zip Code | | 0 | | 0 | [] | 0 | [] | 0 |
| 29 | FDNY Firehouse Listing | | 0 | | 0 | [] | 0 | [] | 0 |
| 30 | Civil List | | 1 | | 0 | ['calendar_year', 'dpt'] | 2 | ['calendar_year'] | 1 |
| 31 | DOB Violations | issue_date | 2 | issue_date | 1 | ['boro', 'lot', 'issue_date'] | 3 | ['issue_date', 'disposition_date'] | 2 |
| 32 | 2012 SAT Results | | 0 | | 0 | [] | 0 | [] | 0 |
| 33 | Water Consumption In The New York City | year | 1 | year | 1 | ['year', 'nyc_consumption_million_gallons_per_day', 'per_capita_gallons_per_person_per_day'] | 3 | ['year'] | 1 |
| 34 | NYPD Complaint Data Current (Year To Date) | cmplnt_fr_dt, cmplnt_to_dt, rpt_dt | 5 | cmplnt_fr_dt, cmplnt_to_dt, rpt_dt | 3 | ['addr_pct_cd', 'cmplnt_fr_dt', 'cmplnt_fr_tm', 'ky_cd', 'pd_cd', 'rpt_dt', 'latitude'] | 7 | ['cmplnt_fr_dt', 'cmplnt_fr_tm', 'cmplnt_to_dt', 'cmplnt_to_tm', 'rpt_dt'] | 5 |
| 35 | Legally Operating Businesses | lic_expir_dd, license_creation_date | 2 | lic_expir_dd, license_creation_date | 2 | ['license_creation_date'] | 1 | ['license_creation_date'] | 1 |
| 36 | COVID-19 Daily Counts of Cases, Hospitalizations, and Deaths | date_of_interest | 1 | date_of_interest | 1 | ['date_of_interest', 'case_count', 'hospitalized_count', 'bk_case_count', 'bk_hospitalized_count', 'mn_case_count', 'qn_case_count', 'qn_hospitalized_count'] | 8 | ['date_of_interest'] | 1 |
| 37 | Housing Maintenance Code Violations | inspectiondate, approveddate, originalcertifybydate, originalcorrectbydate, newcertifybydate, newcorrectbydate, certifieddate, novissueddate, currentstatusdate | 9 | inspectiondate, approveddate, originalcertifybydate, originalcorrectbydate, newcertifybydate, newcorrectbydate, certifieddate, novissueddate, currentstatusdate | 9 | ['boroid', 'lot', 'inspectiondate', 'approveddate', 'originalcertifybydate', 'originalcorrectbydate', 'ordernumber', 'novissueddate', 'currentstatusid', 'currentstatusdate', 'latitude', 'longitude', 'communityboard', 'councildistrict'] | 14 | ['apartment', 'inspectiondate', 'approveddate', 'originalcertifybydate', 'originalcorrectbydate', 'newcertifybydate', 'newcorrectbydate', 'certifieddate', 'novissueddate', | 10 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | 'currentstatusdate'] | |
| 38 | New York City Population by Borough, 1950 - 2040 | | 0 | | 0 | [] | 0 | [] | 0 |
| 39 | DOF Parking Violation Codes | | 0 | | 0 | ['code', 'manhattan_96th_st_below', 'all_other_areas'] | 3 | [] | 0 |
| 40 | DOB ECB Violations | hearing_date, issue_date | 4 | hearing_date, issue_date | 2 | ['boro', 'lot', 'hearing_date', 'hearing_time', 'issue_date'] | 5 | ['hearing_date', 'hearing_time', 'served_date', 'issue_date'] | 4 |
| 41 | Air Quality | | 1 | | 0 | ['indicator_id', 'geo_entity_id'] | 2 | ['year_description'] | 1 |
| 42 | Popular Baby Names | | 1 | | 0 | ['brth_yr', 'cnt', 'rnk'] | 3 | ['brth_yr'] | 1 |
| 43 | Parking Violations Issued - Fiscal Year 2021 | issue_date | 9 | issue_date | 1 | ['issue_date', 'violation_code', 'law_section'] | 3 | ['issue_date', 'vehicle_expiration_date', 'violation_time', 'time_first_observed', 'date_first_observed', 'from_hours_in_effect', 'to_hours_in_effect', 'vehicle_year'] | 8 |
| 44 | FHV Base Aggregate Report | year | 3 | year | 1 | ['year', 'month', 'month_name', 'unique_dispatched_vehicles'] | 4 | ['year', 'month', 'month_name'] | 3 |
| 45 | Medallion Vehicles - Inactive | type, suspension_date, last_updated_date | 5 | type, suspension_date, last_updated_date | 3 | ['type', 'suspension_date', 'last_updated_date', 'last_updated_time'] | 4 | ['model_year', 'suspension_date', 'last_updated_date', 'last_updated_time'] | 4 |
| 46 | Evictions | executed_date | 1 | executed_date | 1 | ['executed_date'] | 1 | ['executed_date'] | 1 |
| 47 | Emergency Response Incidents | creation_date, closed_date | 2 | creation_date, closed_date | 2 | ['creation_date'] | 1 | ['creation_date', 'closed_date'] | 2 |
| 48 | Housing New York Units by Building | project_start_date, project_completion_date, building_completion_date | 3 | project_start_date, project_completion_date, building_completion_date | 3 | ['project_start_date', 'council_district', 'all_counted_units', 'total_units'] | 4 | ['project_start_date', 'project_completion_date', 'building_completion_date'] | 3 |
| 49 | NYPD Arrests Data (Historic) | arrest_date | 1 | arrest_date | 1 | ['arrest_date', 'pd_cd', 'ky_cd', 'arrest_precinct', 'latitude'] | 5 | ['arrest_date'] | 1 |
| 50 | DHS Daily Report | date_of_census | 1 | date_of_census | 1 | ['date_of_census', 'single_adult_women_in_shelter', 'adult_families_in_shelter', 'individuals_in_adult_families_in_shelter'] | 4 | ['date_of_census'] | 1 |
| 51 | 2018 Yellow Taxi Trip Data | tpep_pickup_datetime, tpep_dropoff_datetime | 2 | tpep_pickup_datetime, tpep_dropoff_datetime | 2 | ['vendorid', 'tpep_pickup_datetime', 'tpep_dropoff_datetime', 'passenger_count', 'ratecodeid', 'pulocationid', 'dolocationid', | 10 | ['tpep_pickup_datetime', 'tpep_dropoff_datetime'] | 2 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 'payment_type', 'fare_amount', 'total_amount'] | | | |
| 52 | NYC Health + Hospitals patient care locations - 2011 | | 0 | | 0 | [] | 0 | [] | 0 |
| 53 | Tax Lien Sale Lists | month | 1 | month | 1 | ['month', 'borough', 'lot', 'tax_class_code', 'community_board', 'council_district'] | 6 | ['month'] | 1 |
| 54 | Greenbook | | 0 | | 0 | [] | 0 | [] | 0 |
| 55 | TLC Approved LabCorp Patient Services Drug Test Locations (Dataset) | last_updated_date | 2 | last_updated_date | 1 | ['last_updated_date'] | 1 | ['last_updated_date', 'last_updated_time'] | 2 |
| 56 | Open Streets Locations | open_date | 18 | open_date | 1 | ['open_date'] | 1 | ['open_date', 'start_time', 'end_time'] | 3 |
| 57 | New York City Population By Neighborhood Tabulation Areas | year | 1 | year | 1 | ['year', 'fips_county_code'] | 2 | ['year'] | 1 |
| 58 | DOB Certificate Of Occupancy | c_o_issue_date | 1 | c_o_issue_date | 1 | ['c_o_issue_date', 'lot', 'item_number', 'latitude', 'longitude', 'community_board', 'council_district'] | 7 | ['c_o_issue_date'] | 1 |
| 59 | SAT (College Board) 2010 School Level Results | | 0 | | 0 | [] | 0 | [] | 0 |
| 60 | Property Data (Buildings Information System) | permit_status_date, permit_issuance_date, permit_experation_date | 3 | permit_status_date, permit_issuance_date, permit_experation_date | 3 | ['permit_application_document_number', 'permit_sequence_number', 'permit_status_date', 'permit_issuance_date', 'permit_experation_date'] | 5 | ['permit_status_date', 'permit_issuance_date', 'permit_experation_date'] | 3 |
| 61 | DOB NOW: Build – Job Application Filings | | 0 | | 0 | ['lot', 'commmunity_board', 'latitude', 'longitude', 'council_district'] | 5 | [] | 0 |
| 62 | Trade Waste Hauler Licensees | created, disposition_date, effective_date, expiration_date, export_date | 5 | created, disposition_date, effective_date, expiration_date, export_date | 5 | ['created', 'effective_date', 'expiration_date', 'export_date'] | 4 | ['disposition_date', 'effective_date', 'expiration_date', 'export_date'] | 4 |
| 63 | EMS Incident Dispatch Data | incident_datetime, first_assignment_datetime, first_activation_datetime, first_on_scene_datetime, first_to_hosp_datetime, first_hosp_arrival_datetime, incident_close_datetime | 7 | incident_datetime, first_assignment_datetime, first_activation_datetime, first_on_scene_datetime, first_to_hosp_datetime, first_hosp_arrival_datetime, incident_close_datetime | 7 | ['incident_datetime', 'initial_severity_level_code', 'final_severity_level_code', 'first_assignment_datetime', 'dispatch_response_seconds_qy', 'first_activation_datetime', 'incident_close_datetime', 'incident_disposition_code', 'policeprecinct', 'citycouncildistrict', 'communitydistrict', 'communityschooldistrict', 'congressionaldistrict'] | 13 | ['incident_datetime', 'first_assignment_datetime', 'valid_dispatch_rspns_time_indc', 'first_activation_datetime', 'first_on_scene_datetime', 'valid_incident_rspns_time_indc', 'incident_travel_tm_seconds_qy', 'first_to_hosp_datetime', 'first_hosp_arrival_datetime', 'incident_close_datetime'] | 10 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 64 | 2015 Street Tree Census - Tree Data | created_at | 1 | created_at | 1 | ['created_at', 'cb_num', 'borocode', 'cncldist', 'st_assem', 'st_senate', 'latitude', 'longitude', 'council_district'] | 9 | [] | 0 |
| 65 | M/WBE, LBE, and EBE Certified Business List | date_of_establishment, dateofjob1, dateofjob2, dateofjob3, dateofjob4 | 6 | date_of_establishment, dateofjob1, dateofjob2, dateofjob3, dateofjob4 | 5 | [] | 0 | ['cert_renewal_date', 'date_of_establishment', 'dateofjob1', 'dateofjob2', 'dateofjob3', 'dateofjob4'] | 6 |
| 66 | Parking Violations Issued - Fiscal Year 2014 | issue_date | 9 | issue_date | 1 | ['issue_date', 'violation_code', 'law_section'] | 3 | ['issue_date', 'vehicle_expiration_date', 'violation_time', 'time_first_observed', 'date_first_observed', 'from_hours_in_effect', 'to_hours_in_effect', 'vehicle_year'] | 8 |
| 67 | IPIS (Integrated Property Information System) | | 0 | | 0 | ['boro', 'lot', 'prop_front', 'prop_depth', 'cd'] | 5 | ['final_commitment_text'] | 1 |
| 68 | CURRENT BASES | date | 2 | date | 1 | ['date', 'time'] | 2 | ['date', 'time'] | 2 |
| 69 | Fire Incident Dispatch Data | incident_datetime, first_assignment_datetime, first_activation_datetime, first_on_scene_datetime, incident_close_datetime | 5 | incident_datetime, first_assignment_datetime, first_activation_datetime, first_on_scene_datetime, incident_close_datetime | 5 | ['incident_datetime', 'alarm_box_number', 'dispatch_response_seconds_qy', 'first_assignment_datetime', 'first_activation_datetime', 'incident_close_datetime'] | 6 | ['incident_datetime', 'first_assignment_datetime', 'first_activation_datetime', 'first_on_scene_datetime', 'incident_close_datetime', 'valid_dispatch_rspns_time_indc', 'valid_incident_rspns_time_indc', 'incident_travel_tm_seconds_qy'] | 8 |
| 70 | Property Exemption Detail | year, create_date, extractdt | 8 | year, create_date, extractdt | 3 | ['boro', 'block', 'lot', 'rectype', 'year', 'period', 'exmp_code', 'exmp_seq', 'create_date', 'lline', 'extractdt'] | 11 | ['year', 'create_date', 'enter_date', 'no_years', 'baseyr', 'eff_date', 'prelimdate', 'extractdt'] | 8 |
| 71 | Suitability of City-Owned and Leased Property for Urban Agriculture (LL 48 of 2011) | date_created | 3 | date_created | 1 | ['date_created', 'boro', 'lot', 'land_use_category', 'community_district', 'lot_type_code'] | 6 | ['date_created', 'year_alter_1', 'year_alter_2'] | 3 |
| 72 | 2017 Yellow Taxi Trip Data | tpep_pickup_datetime, tpep_dropoff_datetime | 2 | tpep_pickup_datetime, tpep_dropoff_datetime | 2 | ['vendorid', 'tpep_pickup_datetime', 'tpep_dropoff_datetime', 'passenger_count', 'ratecodeid', 'pulocationid', 'dolocationid', 'payment_type', | 10 | ['tpep_pickup_datetime', 'tpep_dropoff_datetime'] | 2 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 'fare_amount', 'total_amount'] | | | |
| 73 | J-51 Exemption and Abatement | | 2 | | 0 | ['b', 'block', 'lot', 'init_year', 'tax_year'] | 5 | ['init_year', 'ex_years', 'tax_year'] | 3 |
| 74 | Multiple Dwelling Registrations | lastregistrationdate, registrationenddate | 2 | lastregistrationdate, registrationenddate | 2 | ['boroid', 'lot', 'communityboard', 'lastregistrationdate', 'registrationenddate'] | 5 | ['lastregistration date', 'registrationend date'] | 2 |
| 75 | New York City Population By Community Districts | | 0 | | 0 | ['cd_number'] | 1 | [] | 0 |
| 76 | Daily Inmates In Custody | admitted_dt | 2 | admitted_dt | 1 | ['admitted_dt', 'age'] | 2 | ['admitted_dt', 'discharged_dt'] | 2 |
| 77 | 2010 - 2011 School Attendance and Enrollment Statistics by District | | 0 | | 0 | ['ytd_attendance_avg_'] | 1 | [] | 0 |
| 78 | NYPD Shooting Incident Data (Historic) | occur_date | 2 | occur_date | 1 | ['occur_date', 'occur_time', 'precinct', 'latitude'] | 4 | ['occur_date', 'occur_time'] | 2 |
| 79 | Traffic Volume Counts (2012-2013) | date | 1 | date | 1 | ['date'] | 1 | ['date'] | 1 |
| 80 | Directory Of Homebase Locations | | 0 | | 0 | ['latitude', 'community_board', 'council_district'] | 3 | [] | 0 |
| 81 | DSNY Monthly Tonnage Data | month | 1 | month | 1 | ['month', 'communitydistrict', 'refusetonscollected', 'borough_id'] | 4 | ['month'] | 1 |
| 82 | Parking Violations Issued - Fiscal Year 2016 | issue_date | 9 | issue_date | 1 | ['issue_date', 'violation_code', 'law_section'] | 3 | ['issue_date', 'vehicle_expirati on_date', 'violation_time', 'time_first_obse rved', 'date_first_obse rved', 'from_hours_in _effect', 'to_hours_in_eff ect', 'vehicle_year'] | 8 |
| 83 | NYPD Shooting Incident Data (Year To Date) | occur_date | 2 | occur_date | 1 | ['occur_date', 'occur_time', 'precinct', 'latitude'] | 4 | ['occur_date', 'occur_time'] | 2 |
| 84 | Workforce1 Recruitment Events | event_date | 3 | event_date | 1 | [] | 0 | ['event_date'] | 1 |
| 85 | Directory Of DHS Contacts | | 0 | | 0 | [] | 0 | ['department_n ame'] | 1 |
| 86 | Current Reservoir Levels | neversink_date | 1 | neversink_date | 1 | ['neversink_date', 'ashokan_east_storag e', 'ashokan_west_storag e', 'ashokan_release', 'schoharie_storage', 'schoharie_elevation', 'rondout_storage', 'rondout_release', 'neversink_storage', 'neversink_elevation', 'pepacton_storage', 'pepacton_conservatio n_flow_release', 'cannonsville_storage'] | 13 | ['neversink_dat e'] | 1 |
| 87 | FDNY Line Of Duty Deaths | date | 1 | date | 1 | ['date'] | 1 | ['date'] | 1 |
| 88 | 2018 Central Park Squirrel Census - Squirrel Data | | 1 | | 0 | ['x', 'y', 'hectare_squirrel_number'] | 3 | ['date'] | 1 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 89 | 2019 DOE High School Directory | | 2 | | 0 | ['total_students', 'start_time', 'end_time', 'community_board', 'council_district'] | 5 | ['start_time', 'end_time', 'addtl_info1'] | 3 |
| 90 | NYCHA Development Data Book | data_as_of, completion_date | 2 | data_as_of, completion_date | 2 | ['data_as_of', 'tds_', 'consolidated_tds_', 'development_edp_', 'operating_edp_', 'number_of_current_apartments', 'total_number_of_apartments', 'number_of_rental_rooms', 'avg_no_r_r_per_apartment', 'total_population', 'number_of_residential_bldgs', 'density', 'completion_date'] | 13 | ['consolidated_tds_', 'number_of_section_8_transition_apartments', 'number_of_current_apartments', 'total_number_of_apartments', 'avg_no_r_r_per_apartment', 'avg_monthly_gross_rent', 'completion_date'] | 7 |
| 91 | DOHMH Childcare Center Inspections | permitexp, datepermitted, inspectiondate | 3 | permitexp, datepermitted, inspectiondate | 3 | ['permitexp', 'violationavgratepercent', 'avgcriticalviolationrate', 'inspectiondate'] | 4 | ['datepermitted', 'inspectiondate', 'inspectionsummaryresult'] | 3 |
| 92 | Voting/Poll Sites | | 0 | | 0 | ['community_board', 'council_district'] | 2 | [] | 0 |
| 93 | DYCD after-school programs: Beacon Programs | | 0 | | 0 | [] | 0 | [] | 0 |
| 94 | Expense Budget | publication_date | 2 | publication_date | 1 | ['publication_date', 'fiscal_year', 'agency_number', 'unit_appropriation_number', 'object_class_number'] | 5 | ['publication_date', 'fiscal_year'] | 2 |
| 95 | Public Recycling Bins | | 0 | | 0 | [] | 0 | [] | 0 |
| 96 | 2013 - 2018 Demographic Snapshot School | | 1 | | 0 | ['total_enrollment', 'female_1', 'female_2', 'male_1', 'male_2', 'black_1', 'hispanic_1', 'hispanic_2', 'students_with_disabilities_1', 'students_with_disabilities_2', 'poverty_1', 'poverty_2'] | 12 | ['year'] | 1 |
| 97 | 2018 DOE High School Directory | | 2 | | 0 | ['total_students', 'start_time', 'end_time', 'community_board', 'council_district'] | 5 | ['start_time', 'end_time', 'addtl_info1'] | 3 |
| 98 | Queens Library Branches | | 7 | | 0 | [] | 0 | [] | 0 |
| 99 | NYCHA Application Priority Codes | | 0 | | 0 | [] | 0 | [] | 0 |
| 100 | 2016 Green Taxi Trip Data | lpep_pickup_datetime, lpep_dropoff_datetime | 2 | lpep_pickup_datetime, lpep_dropoff_datetime | 2 | ['vendorid', 'lpep_pickup_datetime', 'lpep_dropoff_datetime', 'ratecodeid', 'pickup_longitude', 'pickup_latitude', 'dropoff_longitude', 'dropoff_latitude', 'passenger_count', 'fare_amount', 'total_amount', 'payment_type', 'trip_type'] | 13 | ['lpep_pickup_datetime', 'lpep_dropoff_datetime'] | 2 |