

# Government Digital Analytics Program

Web traffic analysis of government agencies during the time of COVID-19

Alejandro Franza, Analytics Lead  
Austin Pitts, Analytics Engineering  
Philip Monaco, Data Engineering

# Agenda

Background

Research Question

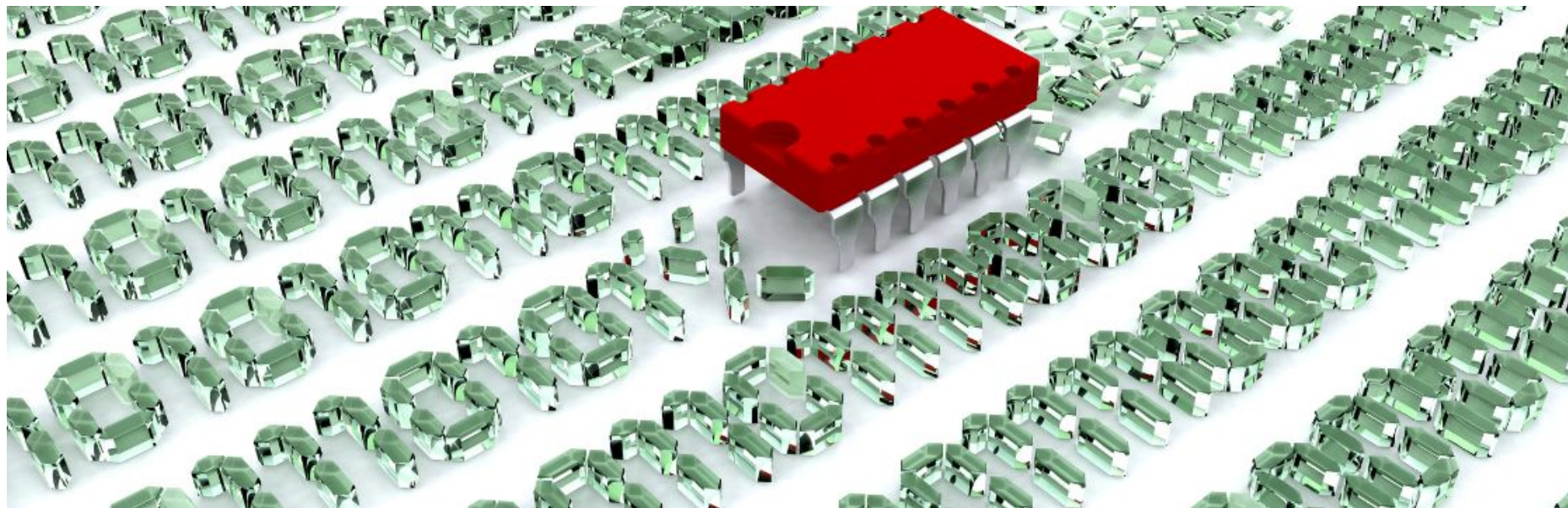
Data collection

Data cleansing

Data exploration

Visualization and analysis

# Background



# Research Question

- Our primary goal is to assess the veracity of the data
- Secondary goal is to get insights such as:
  - Exploring seasonal patterns in the data
  - Most visited domains for a specific timeframe
  - Most common sources of traffic and how they've changed over time
  - How many languages are supported by websites from different agencies

# Data Collection (1/2)

openGSA

[Search](#) [Data](#) [APIs](#) [Code](#) [Events](#)

## analytics.usa.gov API

Overview

Getting Started

OpenAPI Specification File

The Response

Querying reports

Filtering based on agencies

Filtering by domain

Query params

HTTP Response Codes

Contact Us



### This project is in BETA

This API is under active development, and breaking changes may be made without warning. Have feedback or questions? [Please let us know!](#) Please note we have recently updated to `v1.1`, please update your requests accordingly.

## Overview

In addition to being published and available for download, the data generated for analytics.usa.gov is also available via an API.

**Please note we have recently updated to v1.1, please update your requests accordingly.**

The URL for the API is <https://api.gsa.gov/analytics/dap/v1.1>, and it exposes 3 routes to query data:

- `/reports/<report name>/data`
- `/agencies/<agency name>/reports/<report name>/data`
- `/domain/<domain>/reports/<report name>/data`

## Response Query

```
{
  "id": 60716,
  "report_name": "today",
  "report_agency": "justice",
  "date_time":
    "2017-04-07T14:00:00.000Z",
  "data": {
    "visits": "4240"
  },
  "created_at":
    "2017-04-07T04:23:55.792Z",
  "updated_at":
    "2017-04-07T04:23:55.792Z"
}
```

# Data Collection (2/2) - Response

Reports for Agencies: Health & Human Services, Postal Service, Social Security Admin, Treasury			
Reports	Site	Traffic Sources	Language
# of Samples	667,504	161,296	978,426
Date Range	01/01/2020 to 04/17/2023	03/23/2020 to 04/18/2023	01/01/2020 to 04/17/2023
Unique Columns	domain	source, has_social_referral	language
Shared Columns	Id, date, report_name,report_agencies, visits		

# Data Cleansing

- ✗ Missing & NaN values
- ✓ Duplicate values
- ✓ Trimmed time-range
- ✓ Checked for string formatting
- ✓ Dropped unnecessary columns

# Cleansing Stats and Deltas

Reports for Agencies: Health & Human Services, Postal Service, Social Security Admin, Treasury			
Reports	Site	Traffic Sources	Language
# of Samples Rem	458,489	152,618	487,053
# Sample Loss	209,015	8,678	491,373
Time Lost	3 months	1 day	3 months



# Data Exploration

- How many sites and agencies do the reports contain information about?
- Any important gaps in the data?
- Do we understand the meaning of all columns?
- Do aggregate visit values fall within expectations?
- Are there any unexpected values?
- Outliers analysis

## Example: do we understand the meaning of all columns?

- **Visits:** number of visits to each individual domain.
  - ! visits across sites and days (and even within the same day) are not de-duplicated for users
- **Source:** page where the visit originated from.
- **Language:** likely a combination of site and user generated signals

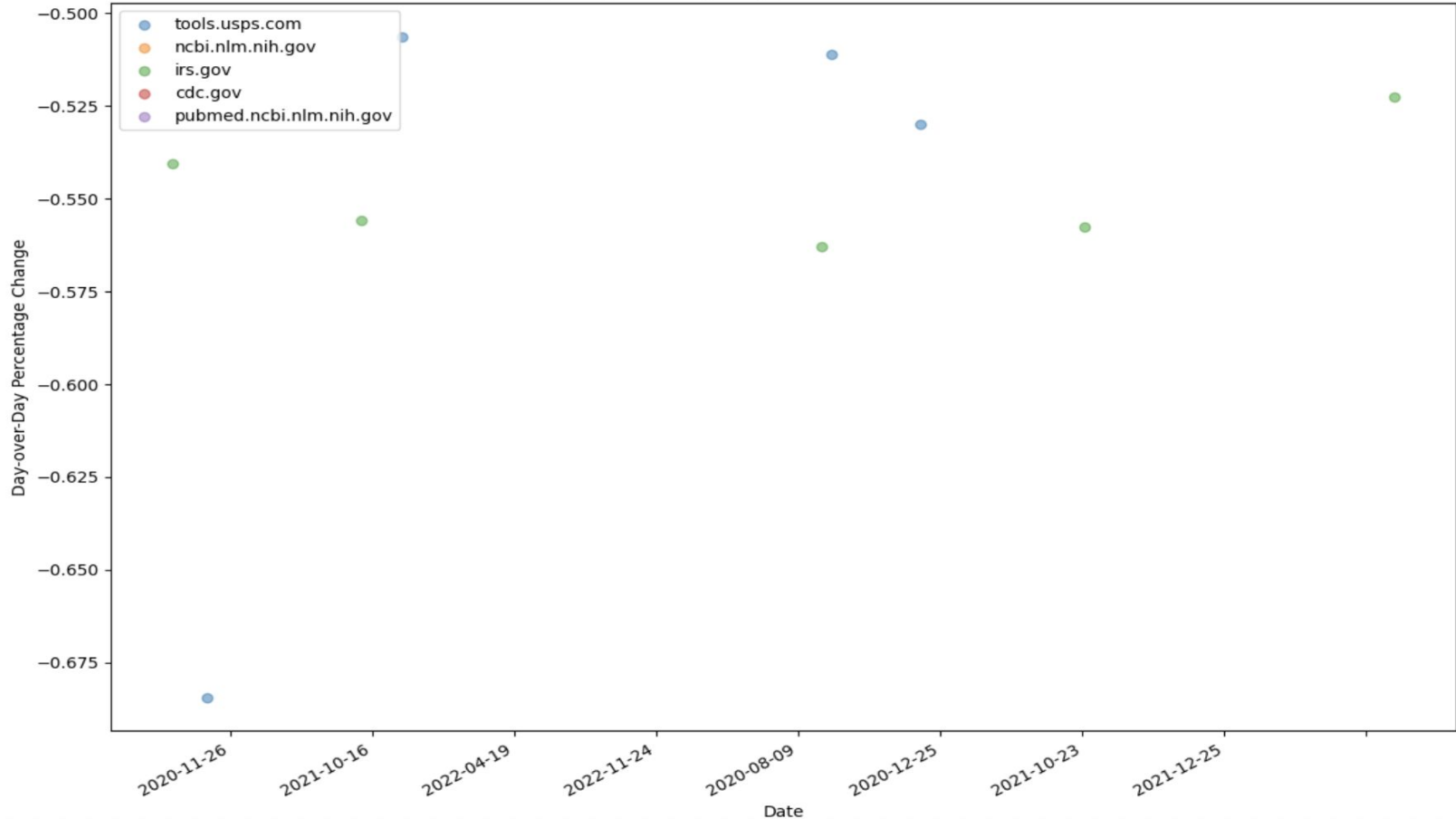
## Example: do aggregate values fall within expectations?

```
Visits per agency on 2020-03-26 00:00:00
report_agency
health-human-services      28450354
postal-service             12207292
social-security-administration 1018944
treasury                   6026646
Name: visits, dtype: int64
```

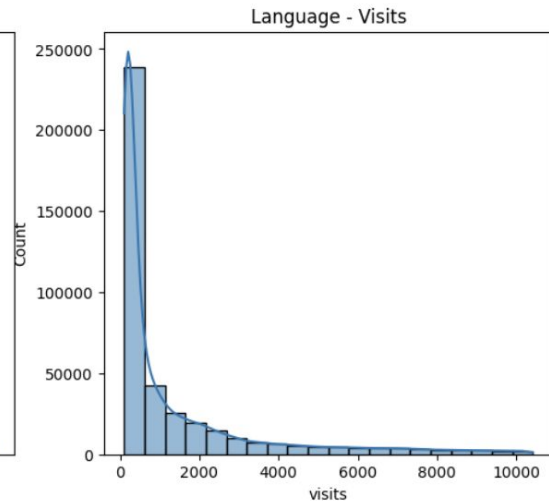
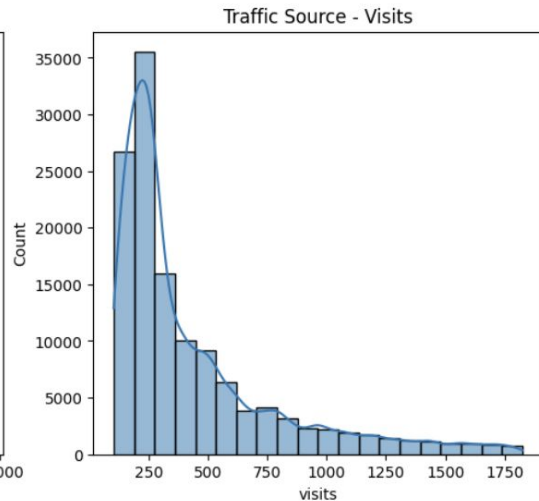
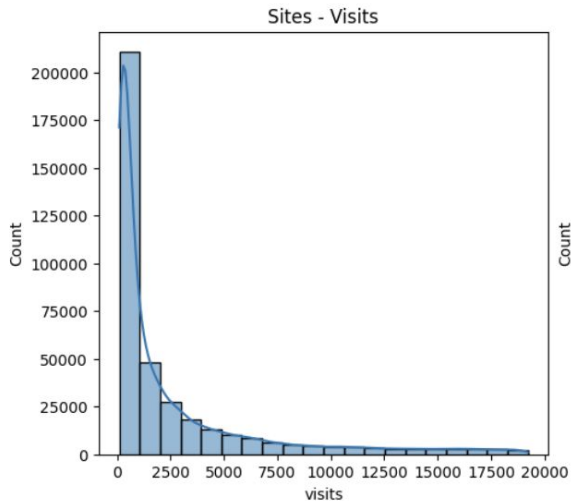
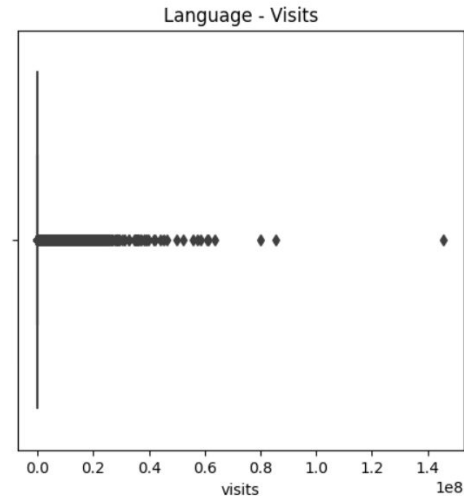
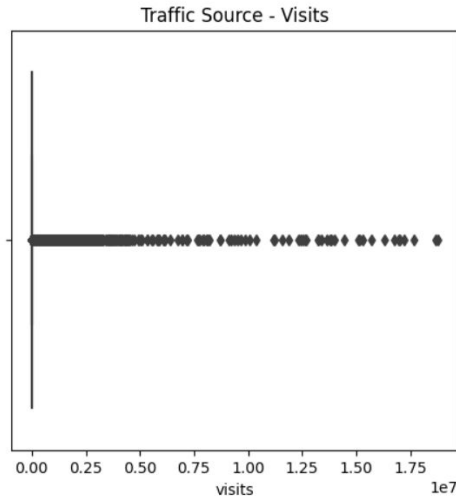
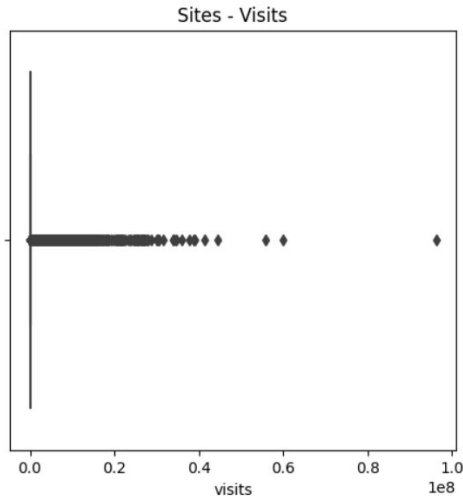
```
Visits per agency as percentage of US population on 2020-03-26 00:00:00
report_agency
health-human-services      8.501499
postal-service             3.647768
social-security-administration 0.304480
treasury                   1.800875
Name: visits, dtype: float64
```

# Example: are there any unexpected values?

Large Day-over-Day Drops in Visits for Top 5 Domains



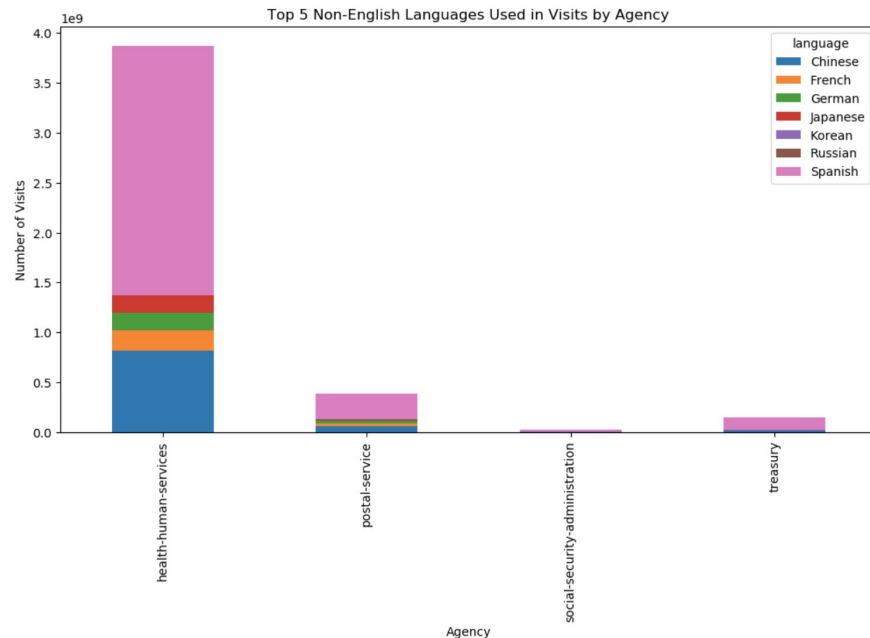
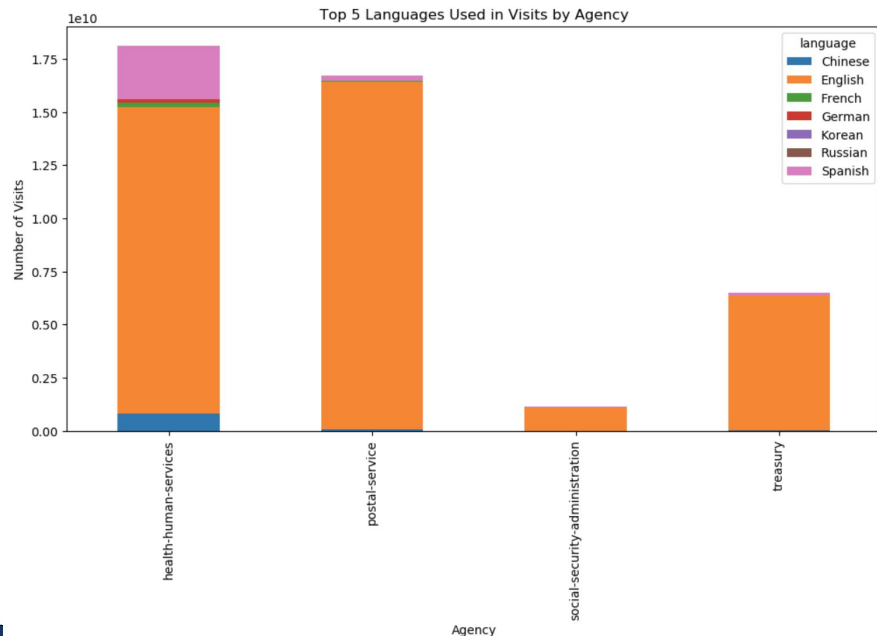
# Example: outlier analysis



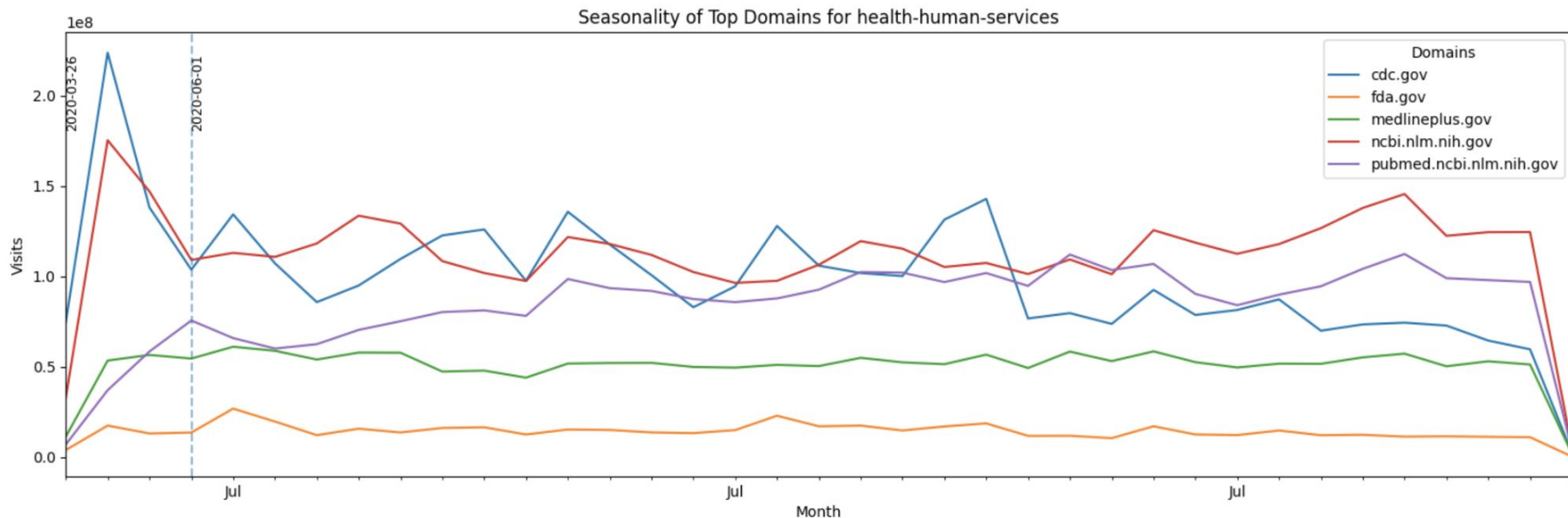
# Visualization and Analysis

- How many languages are attributed to visits to different websites?
- Exploring seasonal patterns in the data
- Most visited domains for a specific timeframe
- Most common sources of traffic and how they've changed over time

# How many languages are attributed to visits to different websites?

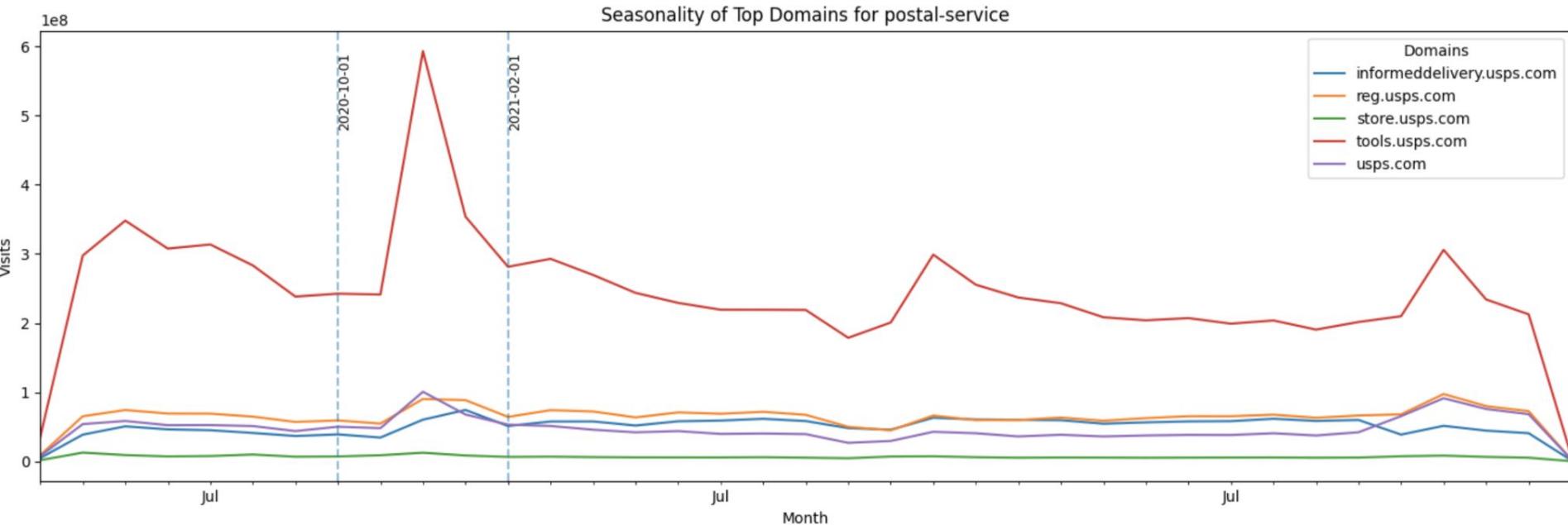


# Exploring seasonal patterns in the data

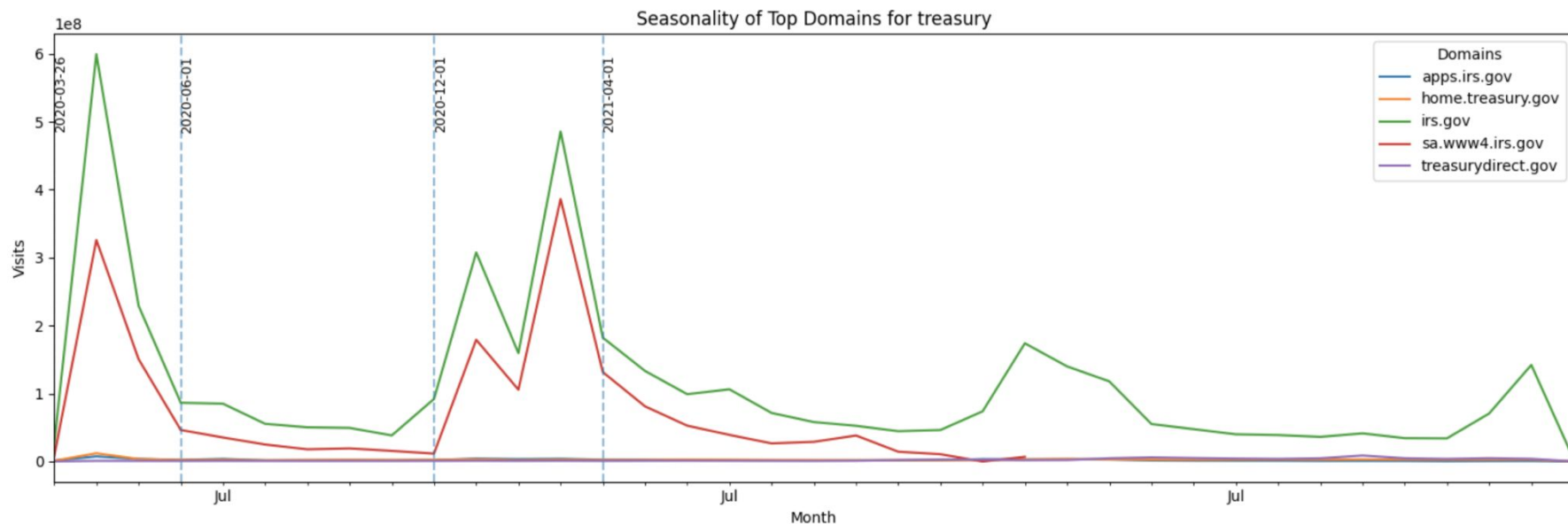




# Exploring seasonal patterns in the data



# Exploring seasonal patterns in the data



# Conclusion

- Data proved to be reliable enough and usable to generate insights
- With some caveats:
  - Extensive data cleansing is required
  - Preferably wait until API moves to alpha
  - Ideally we'd be provided with documentation