https://github.com/UC-Berkeley-I-School/Project2_Franza_Pitts_Monaco

DATASCI200, MIDS
April 2023

Group 5:
Austin Pitts, austinpitts@berkeley.edu
Alejandro Franza Garcia, afranza@berkeley.edu
Philip Monaco, philip.monaco@bekeley.edu

Background

For this EDA project, we decided to use the pre-approved datasets at analytics.usa.gov/data/. These collections of datasets provide metrics around how users interact with government websites online. The data comes from a unified Google Analytics account for U.S. federal government agencies known as the Digital Analytics Program. Currently, the Digital Analytics Program collects web traffic from around 400 executive branch government domains, across about 5,700 total websites, including every cabinet department.

One of the key challenges of this project was acquiring the data. We prospected the original dataset available for download in the main page, which contained *Visits to all domains over 30 days* in CSV format, but found it to be corrupted and not containing the longitudinal data we expected.

For this report we would work under the assumption that this is a request from our manager, with the following prompt:

"We have recently been informed that the data from the Digital Analytics Program (DAP), a unified Google Analytics account for U.S. federal government agencies, has been corrupted. This is quite concerning, as the DAP data is a valuable resource for understanding website traffic and user behavior across various government websites, particularly during the COVID-19 pandemic.

The timeline of interest for this data is from the start of the COVID-19 pandemic until the present time. Before we proceed with any future projects that rely on this data, we need to determine whether it is still usable and reliable. Therefore, I would like to request that you conduct an exploratory data analysis (EDA) on the corrupted DAP data during the specified timeline. This analysis should focus on assessing data quality, identifying inconsistencies, and determining the extent of the corruption. In the Data collection section we describe in detail the challenges faced and how we ultimately were able to use the API."

Data collection

As mentioned briefly in the Background section, the data we were originally planning to use was corrupted by human error. The developers who maintain the dataset, allowed synthetic data to contaminate the downloadable `.csv` files they make available to the public. Over half of the observations contained website domains such as www.fakesite.com.

Instead of waiting for the data errors to be fixed and running the risk of having incomplete data for our project, we decided to proceed with acquiring data through the API. We did need to adjust our initial data questions as the API is still in BETA and did not have the same headers available as the `.csv` files. Where we did lose some variable information, we did gain the ability to acquire time-series data, which was not available in the original dataset.

The analytics.usa.gov has made many reports available in their API, and we provide a example selection below:

- Download: Refers to the number of times a file or asset has been downloaded from a website.
- Traffic-source**:** Refers to the channel or medium that a visitor used to reach a website, such as organic search, social media, or paid advertising.
- Device-model: Refers to the specific model of device that a visitor used to access a website, such as an iPhone X or Samsung Galaxy S21.
- Site: Refers to a specific subdomain or section of a website, such as blog.example.com.
- Second-level-domain: Refers to the part of a domain name that comes before the top-level domain (TLD), such as example in example.com.
- Language: Refers to the primary language of a website or the language preference of a visitor's browser.
- Browser: Refers to the specific web browser used by a visitor to access a website, such as Chrome, Firefox, or Safari.
- Windows-ie: Refers specifically to the Internet Explorer web browser running on the Windows operating system.
- OS: Refers to the specific operating system used by a visitor to access a website, such as Windows, macOS, iOS, or Android.
- Windows: Refers specifically to the Windows operating system.
- IE: Refers specifically to the Internet Explorer web browser.

Reports we are analyzing

We have decided to analyze the three following reports: ‘site’, ‘traffic-source’ and ‘language’. While there is some good information about the data at https://analytics.usa.gov/data/ and https://digital.gov/guides/dap/, there doesn't seem to be detailed documentation for each of the reports we pulled. This might be in part due to the fact that the API we used is in beta mode. However, the column names are mostly self-explanatory and through spot checking the data and with some experience having worked with Web traffic datasets in the past, we were able to come to the following conclusions:

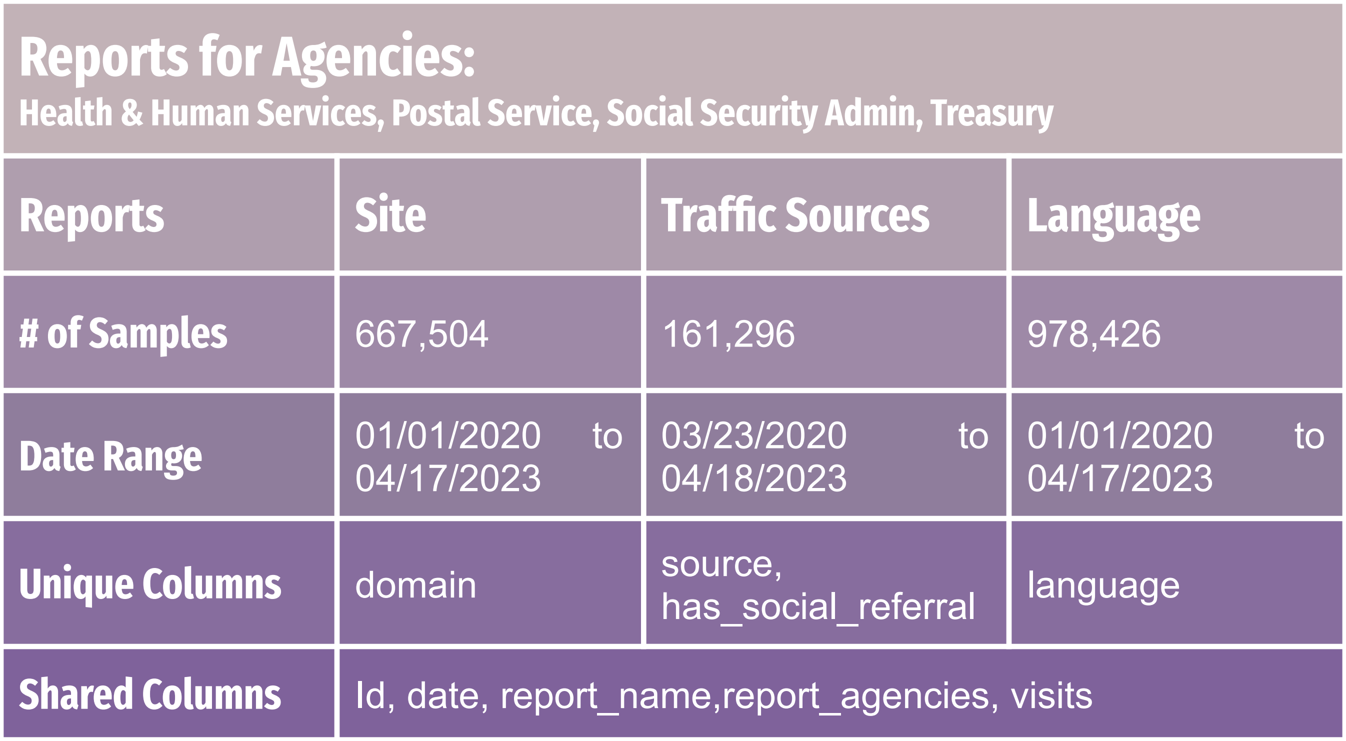
- `id`: random unique identifier for each row in each report. It can't be used to join across reports so it could potentially be dropped.
- `date`: datetime field
- `report_name`: name of each of the 3 reports pulled
- `report_agency`: name of each of the 4 agencies for which we have data
- `domain`: web domain. Each domain (eg. ssa.gov) aggregates data for all its sites (eg. ssa.gov/retirement)

- `visits`: number of visits to each individual domain. Typically the page_view metric counts the number of times a page is viewed, whereas visits counts the number of sessions for visitors
- `language`: we think this represents the inferred language of the user visiting the page
- `source`: the page where the visit originated from. For example, if someone reaches ssa.gov by clicking a link at google.com, then google.com would be the source for that visit

Agencies in scope

To narrow down our analysis, we have decided to focus on analyzing data for the four following government agencies:

- ‘health-human-services',
- 'postal-service',
- 'social-security-administration',
- 'Treasury'



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

To clean the data we took the following steps; checked for missing/NaN and duplicate values, time-range trimming, string formatting, and dropped unnecessary columns.

Here we are doing a side by side comparison of the 3 reports to check the sum of the number of `NaN` values. There are no `NaN` values in any of the reports. Next, we saw that our data is being duplicated, which could be due to the data curation errors. We needed to remove these duplicated values.

We would like to ensure that all three of our reports are consistent in the start and end dates. There was almost a 3 month difference in the data collection start time between the traffic source report and the site and language reports. There is also a 1 day difference in the end date between the traffic source report and the site and language reports. After filtering, all three reports now have a consistent start and end date and we can overwrite the original dataframes with the filtered outputs.

We suspected that there might be some unexpected string formatting within some of the reports. Because we are expecting to answer analytics questions on the traffic source and user language, we want to inspect the columns for any odd string formatting patterns.

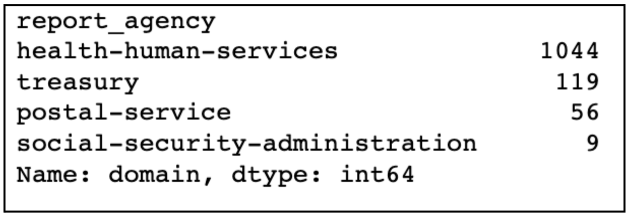
Finally we will want to see if we can drop any columns that aren't necessary for analysis. The only column we suspect we will need to consider dropping is the "id" column if the ids are not shared across all the reports. We are making the assumption that the ids are generated

In total we lost 209,015 samples in the site report, 8,678 in the traffic source report, and 491,373 samples from the language report. In time, we lost ~3 months from the site and language reports.

Data exploration

How many sites and agencies does the site report contain information about?

The site report contains information about 4 agencies (Health & Human Services, Treasury, Postal Service and Social Security Administration). Health & Human Services is the agency with the largest number of domains (1044), whereas Social Security Administration is the lower (9).

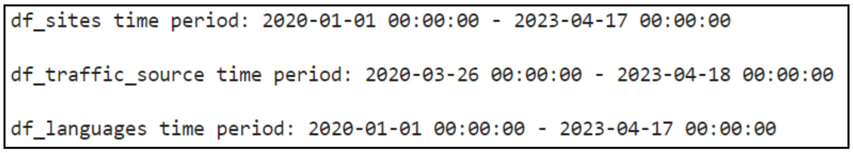


```
report_agency
health-human-services              1044
treasury                            119
postal-service                       56
social-security-administration        9
Name: domain, dtype: int64
```

# unique domains contained in each of the agencies' reports

Any missing or null values? Any other gaps in the data?

Upon inspection of the datasets we found no missing or null values that needed to be accounted for. However, we did find that the various datasets had different date ranges.To ensure consistency in our analysis, we compensated for this in our data cleaning by creating a new date range that would not leave any gaps and restricting our data to that range.

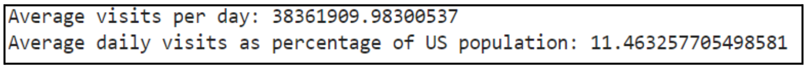


```
df_sites time period: 2020-01-01 00:00:00 - 2023-04-17 00:00:00

df_traffic_source time period: 2020-03-26 00:00:00 - 2023-04-18 00:00:00

df_languages time period: 2020-01-01 00:00:00 - 2023-04-17 00:00:00
```

Data ranges of our datasets upon initial inspection before cleaning.

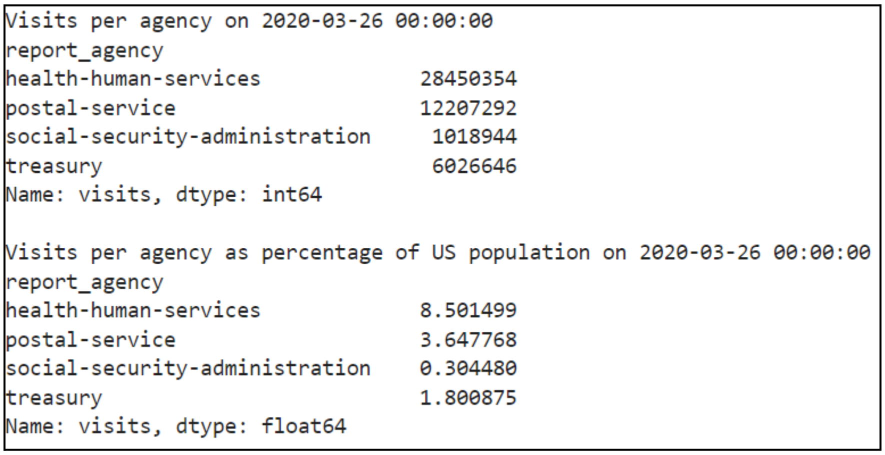
Do aggregate visit values fall within expectations (eg. # visits as % of US population seems realistic)?

We performed a sanity check on our data to ensure its validity by verifying if our daily aggregate visit values fell within expectations as a percentage of the entire US population. In the table below we can see that our data results are within reasonable bounds. This analysis also helps to give an idea of the scale of the data we are looking at.



```
Average visits per day: 38361909.98300537
Average daily visits as percentage of US population: 11.463257705498581
```

Average site visits per day across all agencies and also as a percentage of the US population.



```
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
```

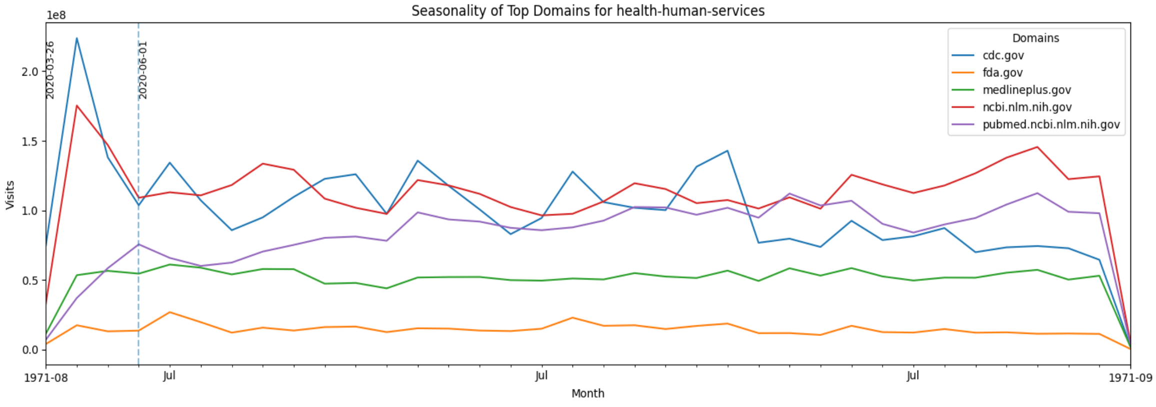
Taking any date in our range as input, 3-26-2020 in this example, we get the numerical site visits per agency that day and also as a percentage of the US population.

Are there any unexpected values?

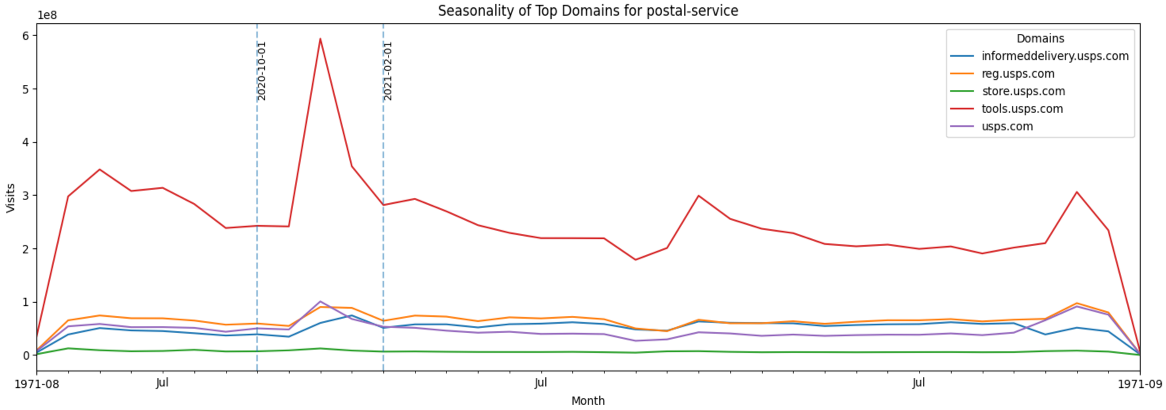
We would consider unexpected values as any of the following: negative domain visits, large day over day variations on traffic on a per domain basis. **Drops in Visits:** the largest drops in percent change to traffic were for the domains irs.gov and tools.ups.com. The drops are likely due to the end of tax season and the end of the US Presidential Election.

Seasonality patterns

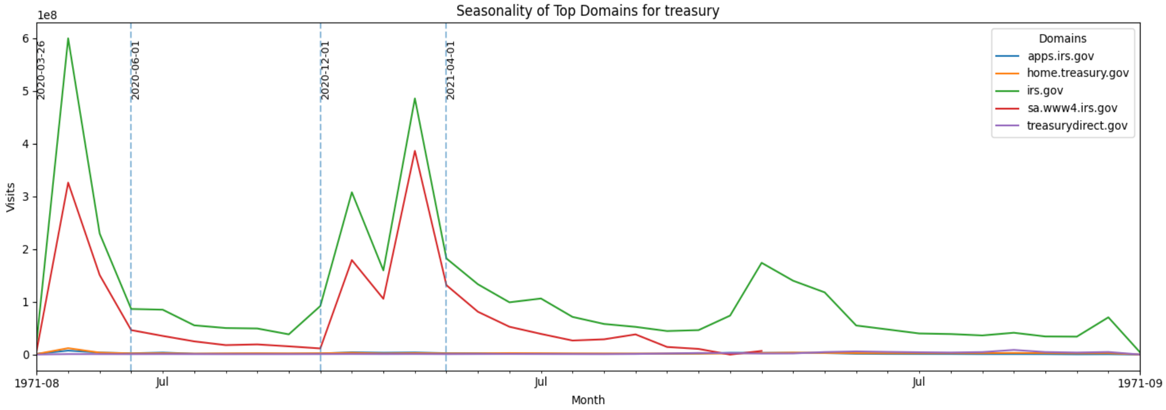
We plotted the daily visits data for each of the for agencies in scope for the whole available time range as seen in the chart below:



Health-human-services - Between March 2020 and June 2020, there was a noticeable spike in site visits to cdc.gov and ncbi.nlm.nih.gov. This increase in traffic can be largely attributed to the COVID-19 pandemic, which had a significant impact on public health and the global economy during this time. The Centers for Disease Control and Prevention (CDC) website, cdc.gov, is a primary source of information on public health and safety, including guidelines and recommendations to prevent the spread of infectious diseases. The public's increased reliance on the CDC website for critical information likely contributed to the spike in visits during this period.



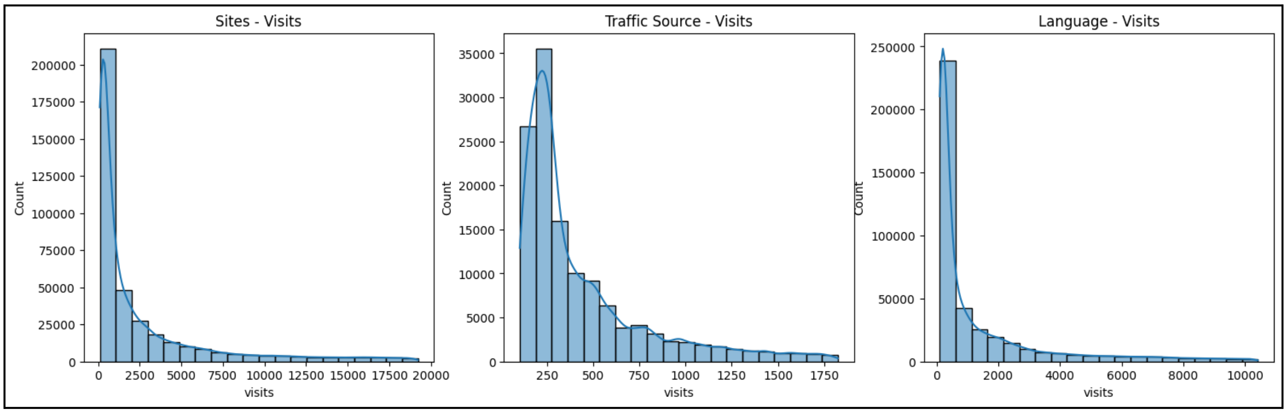
postal-service - Between October 2020 and February 2021, there was a noticeable spike in site visits to tools.usps.com, the United States Postal Service (USPS) website. This increase in traffic can be primarily attributed to the 2020 United States Presidential Election. The 2020 Presidential Election took place on November 3, 2020. In the lead-up to the election, there was a heightened emphasis on mail-in voting due to concerns about the spread of COVID-19 at in-person polling places.



treasury - Between March 2020 and June 2020, and then again between December 2020 and April 2021, there were notable spikes in site visits to irs.gov and sa.www4.irs.gov, the official websites of the United States Internal Revenue Service (IRS). These increases in traffic can be mainly attributed to two key factors: the impact of the COVID-19 pandemic on federal tax filing and the distribution of stimulus payments. The COVID-19 pandemic motivated more people to file their taxes online instead of in person. The federal government approved multiple rounds of economic stimulus payments to provide financial relief to millions of Americans during the COVID-19 pandemic. The IRS was responsible for distributing these payments. The sa.www4.irs.gov website, became a primary resource for individuals tracking their stimulus payments.

Are there any outliers in the data

To analyze the presence of outliers we looked at whether there are any domains with too high or too low # visits? After blox-potting the visits data for each of the three reports analyzed, we found this to be the case. We then decided to remove observations with visits outside of the interquartile range of the middle 50%. After removing outliers we created a histogram for each report - as depicted below - and found the distribution to be smoother and within expectations, as it is typical for website traffic data to have a positive-skew distribution with a long right tail.



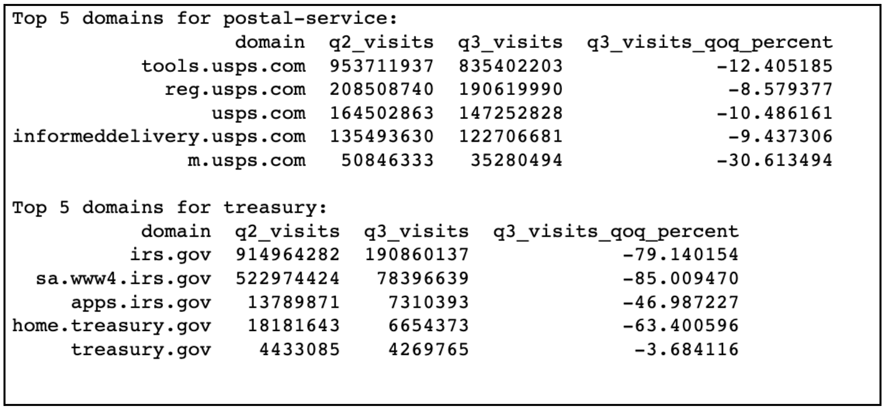
Histogram showing count of domains for different visit count for each of the three reports analyzed

Visualizing and Analyzing Results

Question 1

What domains were the most visited (top 5) for the Postal Service and Treasury agencies during Q2 '20 (chose this period to investigate any potential impact from the COVID-19 pandemic in visits) and compared to Q1'2020. Any interesting observations?

All sites across those two agencies saw meaningful traffic declines ranging from -5% to -80% quarter-over-quarter. Particularly, all sites across the **treasury** agency saw large declines in quarter-over-quarter growth, which is aligned with expectations as tax season begins in Mar-Apr.



```
Top 5 domains for postal-service:
                   domain  q2_visits  q3_visits  q3_visits_qoq_percent
          tools.usps.com  953711937  835402203             -12.405185
            reg.usps.com  208508740  190619990              -8.579377
                usps.com  164502863  147252828             -10.486161
informeddelivery.usps.com  135493630  122706681              -9.437306
              m.usps.com   50846333   35280494             -30.613494

Top 5 domains for treasury:
          domain  q2_visits  q3_visits  q3_visits_qoq_percent
         irs.gov  914964282  190860137             -79.140154
  sa.www4.irs.gov  522974424   78396639             -85.009470
    apps.irs.gov   13789871    7310393             -46.987227
home.treasury.gov   18181643    6654373             -63.400596
    treasury.gov    4433085    4269765              -3.684116
```

Table showing top 5 domains per agency by Q3’20 visits and differences with Q2’20

Question 2

For a given year (for instance, 2021) can we spot any seasonal patterns in # visits across each of the main agencies?

Plotted below are the daily visits to each of the 4 agencies' sites for the year of 2021. Every agency's visits numbers have **weekly seasonality patterns** (weekdays have higher visits than weekends) and that is why we see a saw-toothed shape for all the plots. Other notable trends by agency:

- The clearest annual seasonal trend appears in **Treasury** domains, where tax season (March-April) sees a large spike of visits to then flatten out for the rest of the year.
- **Postal service** visits seem to have a seasonal pattern in Q4 where visits slowly ramp towards Black Friday and the holiday season.
- **Social security administration** and **health and human services** have more stable patterns throughout the year with no notable seasonality.

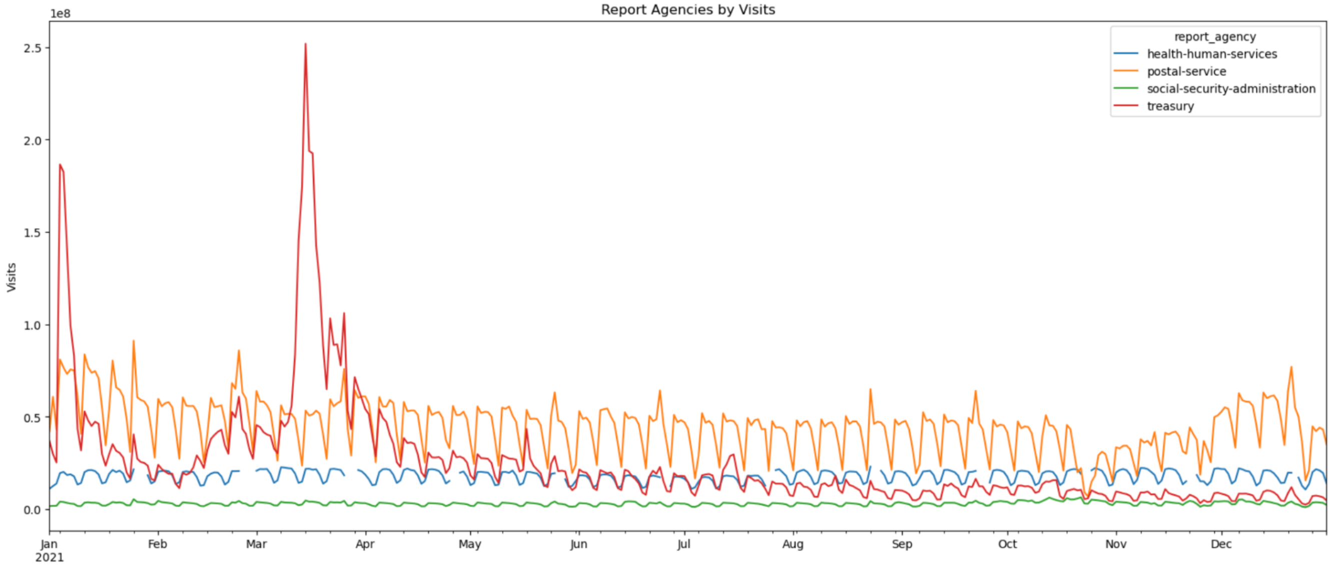
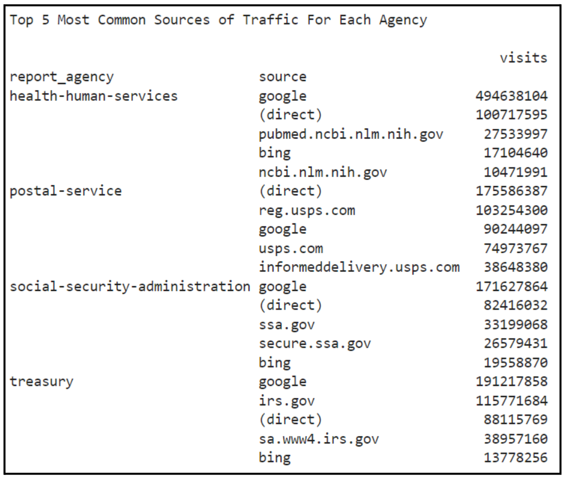


Chart: Daily visits per agency for the year of 2021

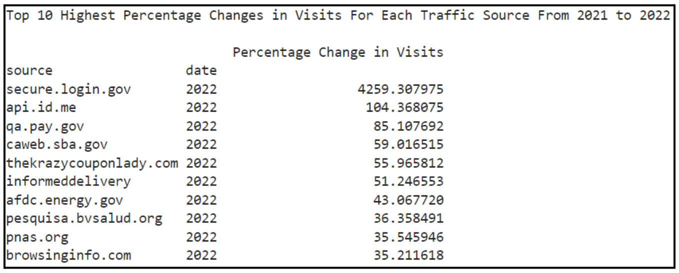
Question 3

What are the most common sources of traffic to sites of each of the different agencies? What are the sources that grew the most y/y in 2022 (vs 2021)?

The tables below show us the most common sources of traffic for each agency and the highest percentage changes in visits for each traffic source from 2021 to 2022.



```
Top 5 Most Common Sources of Traffic For Each Agency

                                                           visits
report_agency                   source
health-human-services           google                  494638104
                                (direct)                100717595
                                pubmed.ncbi.nlm.nih.gov  27533997
                                bing                     17104640
                                ncbi.nlm.nih.gov         10471991
postal-service                  (direct)                175586387
                                reg.usps.com            103254300
                                google                   90244097
                                usps.com                 74973767
                                informeddelivery.usps.com 38648380
social-security-administration  google                  171627864
                                (direct)                 82416032
                                ssa.gov                  33199068
                                secure.ssa.gov           26579431
                                bing                     19558870
treasury                        google                  191217858
                                irs.gov                 115771684
                                (direct)                 88115769
                                sa.www4.irs.gov          38957160
                                bing                     13778256
```



```
Top 10 Highest Percentage Changes in Visits For Each Traffic Source From 2021 to 2022

                                  Percentage Change in Visits
source                 date
secure.login.gov       2022                      4259.307975
api.id.me              2022                       104.368075
qa.pay.gov             2022                        85.107692
caweb.sba.gov          2022                        59.016515
thekrazycouponlady.com 2022                        55.965812
informeddelivery       2022                        51.246553
afdc.energy.gov        2022                        43.067720
pesquisa.bvsalud.org   2022                        36.358491
pnas.org               2022                        35.545946
browsinginfo.com       2022                        35.211618
```

Tables of the top 5 most common sources of traffic for each agency and the top 10 highest percentage changes in traffic sources that grew the most y/y in 2022.

Question 4

How many languages do sites from different agencies support? Are there any notable differences between language distribution across the 4 agencies selected? (for example: social-security might be x% ES language whereas 'postal-service' might be y%)?

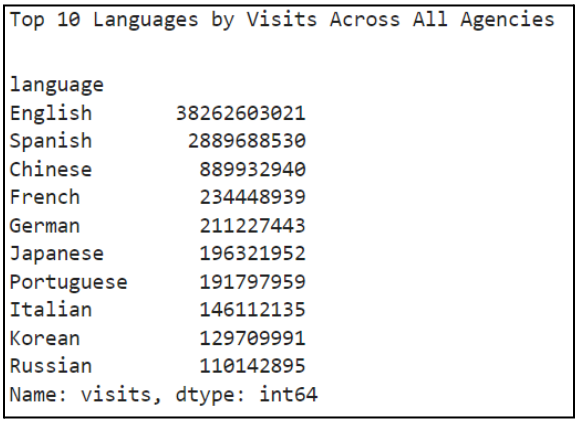
The language section of our data was initially represented in IEFT language tags with regional subtags. For example we would have "en-us" and "en-gb" as separate languages, but both represent English, just different regional versions of English. In order to get a more accurate representation of the visits associated with each language, we cleaned the data by removing these subtags and then grouping them by language tag. This will allow better comparison between languages, because otherwise the top 10 would be mostly all variations of English. Finally for cleaning, we turned the IEFT tag into its english translation of the language it represented. In our final resulting dataset, we found that the different agencies support 93 unique languages.

In the following figures, we can see the top 10 languages by visits across all agencies. Then, in order to get a better understanding of differences in language distribution across different agencies, we created the two stacked bar charts seen below. In the first plot we have top 5 languages used in visits by agency, while the second plot uses the same concept but only focusing on non-english languages. As expected, English is the most used language across all agencies. But when comparing the two graphs we can draw a couple interesting conclusions.

1. While postal service is a highly visited agency, it drops significantly when only non-english languages are looked at. This makes intuitive sense because foreigners who may be traveling or looking into the US would not need to use the postal service as much as people living in the US.
2. This same logic can also be applied to why health and human services have the most visits of any agency of non-english languages. Of the agencies represented, this is likely

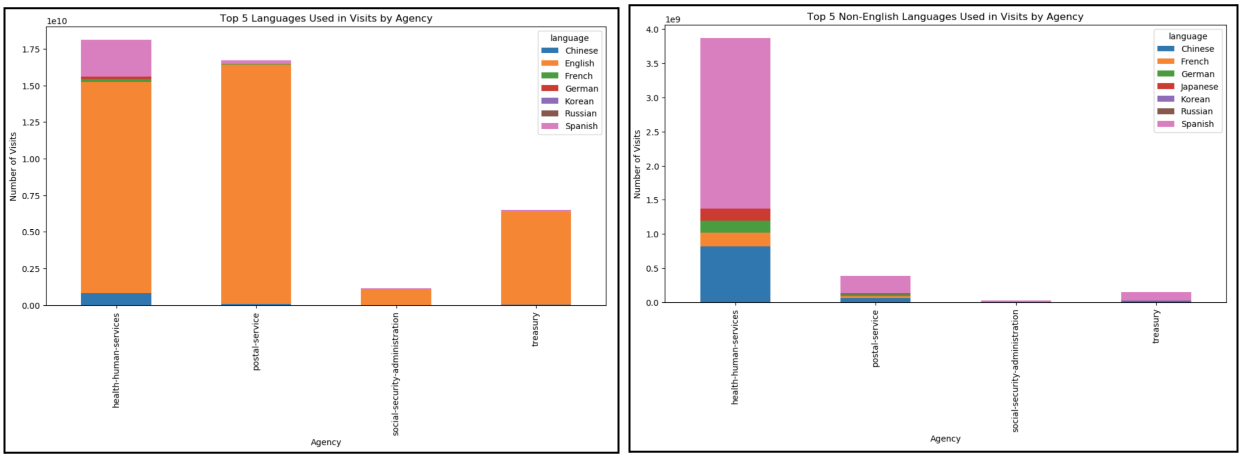
to be the one most useful to a foreigner and our data supports that assumption. 3) Japanese is one of the top 5 languages for the health and human services agency, but it is not one of the top 5 for the postal service agency.

Again, this provides support to the logic behind the previous claims made when analyzing the data.



```
Top 10 Languages by Visits Across All Agencies

language
English       38262603021
Spanish        2889688530
Chinese         889932940
French          234448939
German          211227443
Japanese        196321952
Portuguese      191797959
Italian         146112135
Korean          129709991
Russian         110142895
Name: visits, dtype: int64
```

The top 10 languages used in site visits across all agencies



Stacked bar charts of the top 5 languages used for visits by agency, the right bar chart only focusing on non-english languages.

Conclusion

In conclusion, we would report to our managers that the data proved to be reliable enough and usable to generate insights, like the ones we presented in this report. There are some caveats that we'd like to point out:

- Extensive data cleansing is required
- Preferably wait until API moves to a stable release
- Ideally we'd be provided with documentation