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

Berkeley
UNIVERSITY OF CALIFORNIA

Presenter: TBD

- Title of the project and group members

Agenda

Background

Research Question

Data collection

Data cleansing

Data exploration

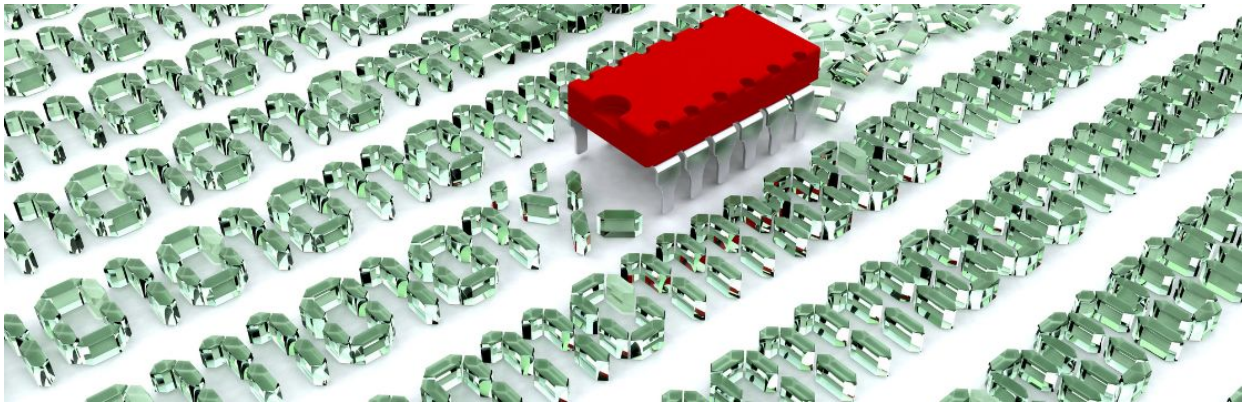
Visualization and analysis

Berkeley
UNIVERSITY OF CALIFORNIA

Presenter:

- Agenda for today

Background



Berkeley
UNIVERSITY OF CALIFORNIA

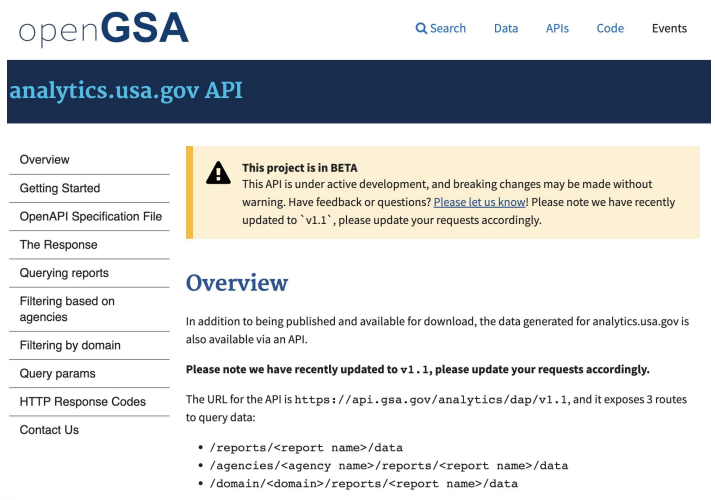
Presenter: Philip

- We are assuming that our manager requests us to explore the veracity of analytics.usa.gov/data/ data and consider whether the data is good to use for future analytics projects.
- These datasets provide metrics on how users interact with government websites
- The data comes from a unified Google Analytics account for U.S. federal government agencies called the Digital Analytics Program
- Acquiring the data was a key challenge, as the original dataset was corrupted and did not contain the expected longitudinal data
- Later on in the data collection section we will describe the challenges faced and how we used the API to obtain the data.

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)



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"
}
```


Berkeley UNIVERSITY OF CALIFORNIA

Presenter: Philip

- The data we were originally planning to use was corrupted by human error.
- The CSV file contained 30 day site traffic of all of the government domains.
- The CSV contained errors such as test-site.com type artifacts possible left by developers.
- We were asked by our manager to test the data provided by the openGSA analytics.usa.gov API which is still in BETA.
- The api queries can be filtered by 27 Agency, 15 Reports, and time.
- The response from the API is an array of json files with all of the headers and contains a data array of the unique variables for each report.
- Once we figured out how to use the API, we narrowed the scope of the data to 4 main agencies and 3 key reports asked by our manager in the timeframe of the beginning of the COVID-19 pandemic approximately March 2020.
- Little to no upfront information could be found on what each of the values meant or how they were generated.
- Go through json example on slide

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		

Berkeley
UNIVERSITY OF CALIFORNIA

We received data based on reports and agencies.
The reports we were asked to query were the site, traffic sources, and language.
The Agencies we were ask to look at are..

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

Berkeley
UNIVERSITY OF CALIFORNIA

- How many sites and agencies do the reports contain information about?
  - Health and human services more than 1k sites. Social security administration just 9 sites
- Any important gaps in the data?
  - No gaps in the data were found other than inconsistencies in time ranges available for different reports and agencies
  - 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.
- Do we understand the meaning of all columns?
- Do aggregate visit values fall within expectations?
  - Sanity-check of volume of visits. From the site report, for a given day calculate the number of visits per agency. Compare against each other and see the scale of the values and whether it makes sense
- Are there any unexpected values?
  - We could choose a large domain and for a given month, show a table with d/d differences %s.
- Do we understand the meaning of all columns?
  - Source: page where user is coming from (direct, search referral (google.com, , social referral facebook, ig, twitter)
- Outliers analysis

- are there any domains with too high or too low # visits?
- 2-3 std diff
- box-plots

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

Berkeley
UNIVERSITY OF CALIFORNIA

- We didn't find detailed documentation for each of the reports we pulled, but most of the fields were self-explanatory.
- Two fields seem critically important to understand
  - 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
    - De-duplication
  - Source: For example, if someone reaches ssa.gov by clicking a link at google.com, then google.com would be the source for that visit
  - Language: we think this represents the inferred language of the user visiting the page

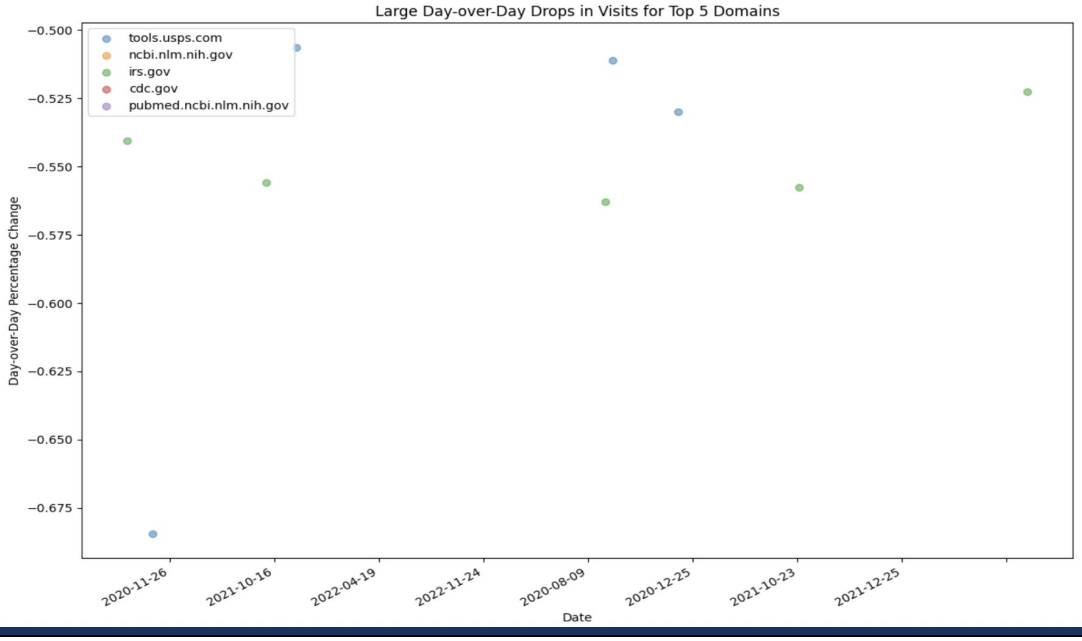
Example: do aggregate values fall within expectations?

```
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
```

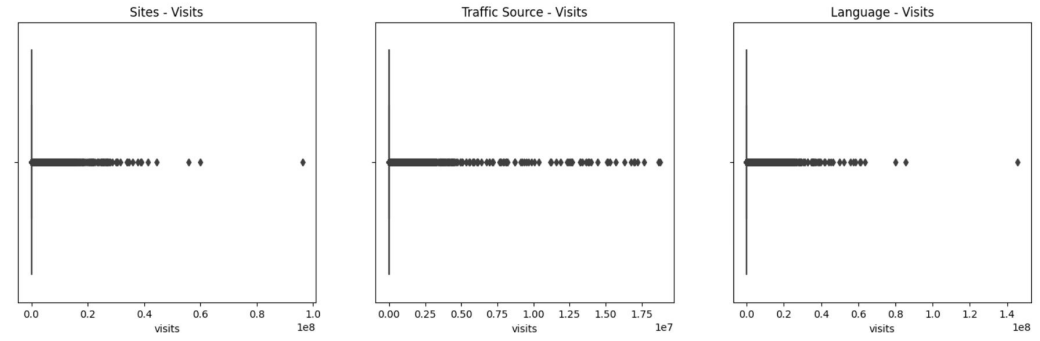

- Sanity check on our data to ensure its validity by verifying if our daily aggregate visit values fell within expectations as a percentage of the entire US population.
- In the table below we can see that our data results are within reasonable bounds.
- The reason % looks high is because visits are not deduplicated across time or sites so numbers are inflated as expected

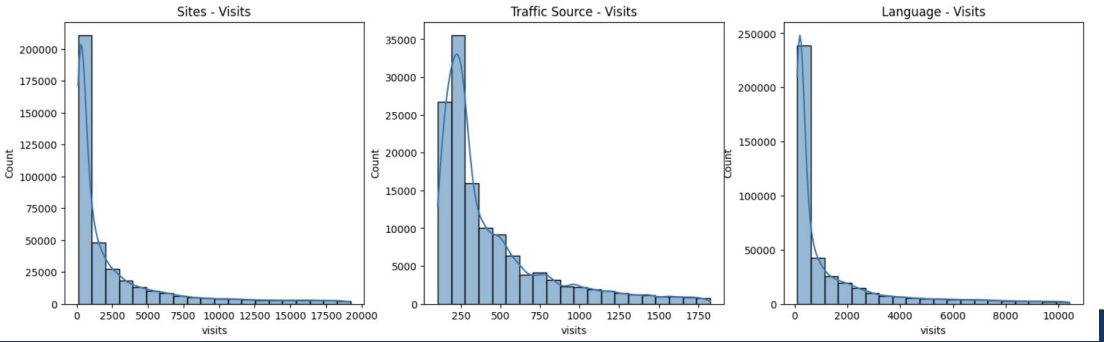
Example: are there any unexpected values?



1. For the top 5 domains, created a data frame with the percentage of daily change in visits for each day
2. Filtered the data to keep only the largest > 50% drops and plotted
3. There are very few such drops which seem to be aligned with expectations

Example: outlier analysis



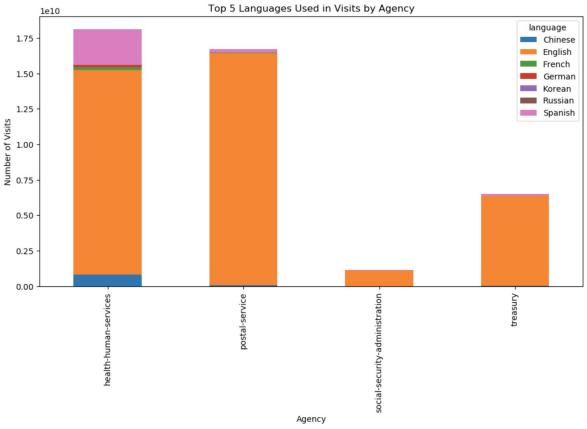


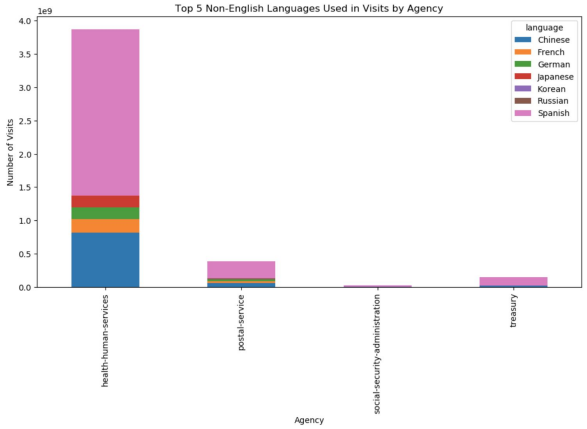
1. Box-plotted visits from each of the available reports
2. Removed visit outliers with the interquartile range of the middle 50%.
3. Created a histogram for each report without the outliers
4. Typical long-tail distribution for web traffic was observed

Visualization and Analysis

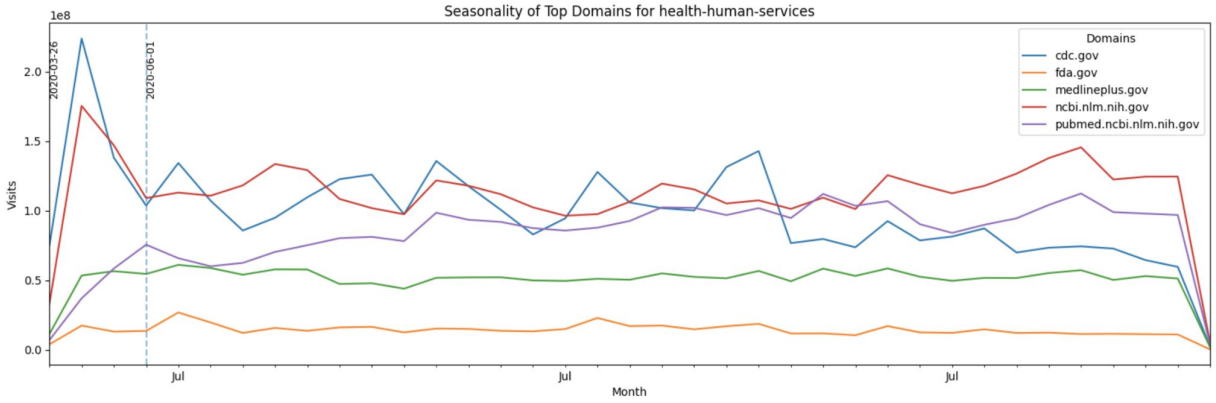
- How many languages are attributed to visits to different websites?
- Exploring seasonal patterns in the data
- Most visited domains for a specific timeframe
- Most common sources of traffic and how they've changed over time

How many languages are attributed to visits to different websites?

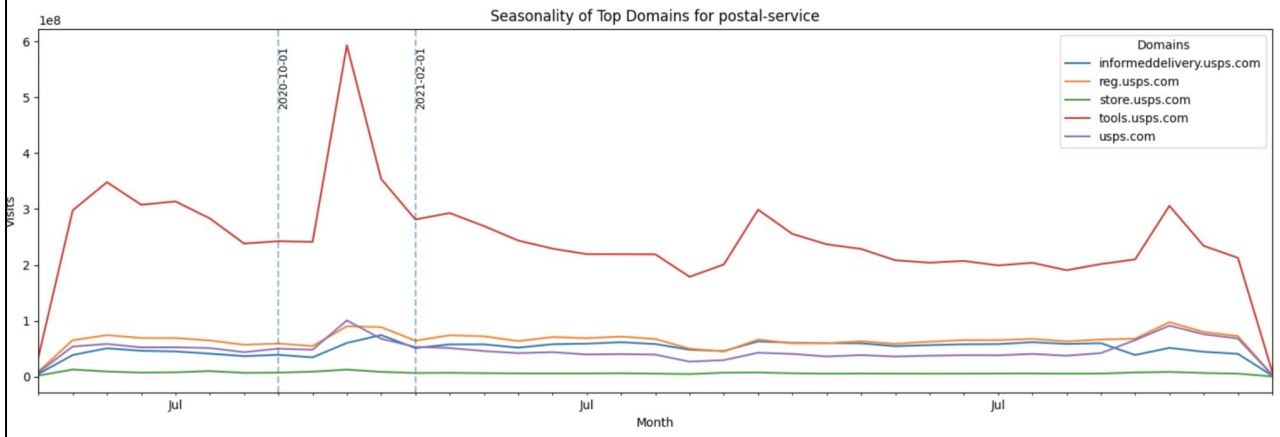




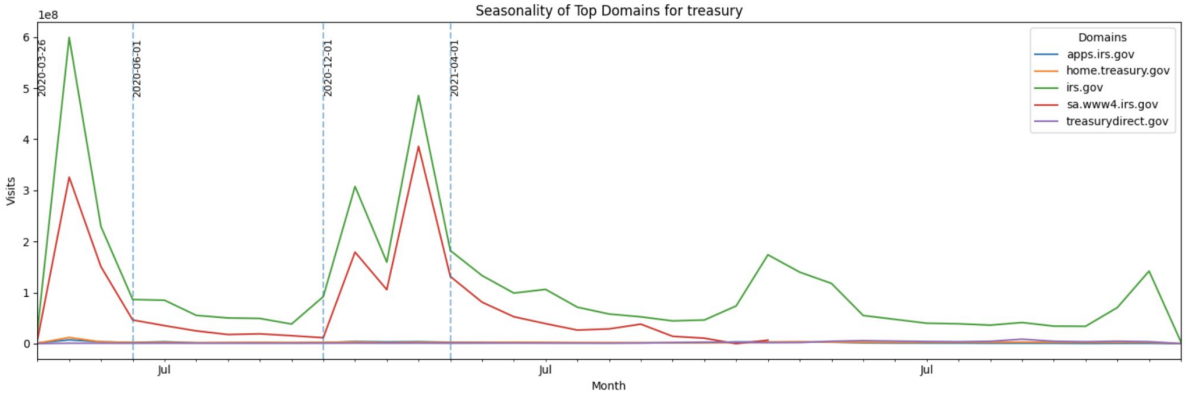
Exploring seasonal patterns in the data



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 stable release
  - Ideally we'd be provided with documentation