# *Final Report: US Government Website Analytics - Exploratory Data Analysis*

Team Members: Courtney Mazzulla, Julian Rippert, Raymond Tang

https://github.com/UC-Berkeley-I-School/Project2_Mazzulla_Rippert_Tang

## Introduction

The purpose of this exploratory data analysis project was to identify key insights of US government website traffic data. The US government across its numerous departments hosts many websites which each day are generating traffic from within the US and abroad. This traffic data is being recorded and monitored through the US Government’s Digital Analytics Program (DAP) program, which seeks to help understand how people are interacting with the government online. Our analysis was two fold. Initially, we analyzed aggregate data sets which encompassed all US governments domains. Using this analysis, we were able to identify key departments which generated the most traffic, most downloads and most exits to explore our sub questions with. As such below is the breakout of research questions the team was focused on:

| ***All Participating Websites (Aggregate):*** | ***Department Specific Sub-Questions:*** |
|---|---|
| 1. What web pages have the most traffic? Least traffic?<br>2. Which web pages have the most users? Least amount of users?<br>3. Which web pages do users stay on the longest?<br>4. Which web pages keep users browsing the most?<br>5. What web pages have the most downloads? Least downloads?<br>6. Which web pages do users exit the most frequently? | 1. From what countries are the users interacting with the web pages?<br>2. From what cities are the users interacting with the web pages?<br>3. From what languages are the users interacting with the web pages?<br>4. How are visitors interacting with the web pages? |

## Automated Data Pull

The data comes from a unified Google Analytics account for U.S. federal government agencies known as the DAP. The dataset is not only huge, but has different types of data scattered across multiple web links inside a departmental pulldown menu. In order to avoid manually clicking each weblink, developing an automated data pull script is the best way to obtain all the different types of data in one execution. The other reason for this approach is to allow getting the entire dataset at a specific time so that US daytime and nighttime data could be analyzed and compared. The automated data pull mechanism is developed as a Python script launched by a cron job set to run on July 14 at 11pm and July 15 at 11am separately. Downloaded data are in CSV or JSON file format and saved in their corresponding department folders. The following table shows the dataset downloaded and the variables extracted from the dataset to answer the questions focused on all participating websites:

| US Government’s Web Traffic Logs | | | |
|---|---|---|---|
| **Data Source** | https://analytics.usa.gov/data/ | | |
| All Participating Websites (Aggregate) | | | |
| **Dataset Name** | **Time Aggregation** | **Variables** | **Data Type** |
| Visits to all domains | 30 days | Domain name | string |
| Top downloads yesterday | 1 day | Number of visits | integer |

| | |
|---|---|
| Top traffic sources | 30 days |
| Top exit pages | 30 days |
| All pages people are visiting | Every 5 min real-time |
| Total people online | Every 5 min real-time |

| | |
|---|---|
| Number of pageviews | integer |
| Number of users | integer |
| Number of pageviews per session | float |
| Average session duration | float (in min) |
| Number of exits | integer |
| Exit page name | string |
| Download page title | string |
| Download file URL | string |
| Download file path | string |
| Number of download events | integer |
| Traffic source name | string |
| Traffic source has social referral | boolean |

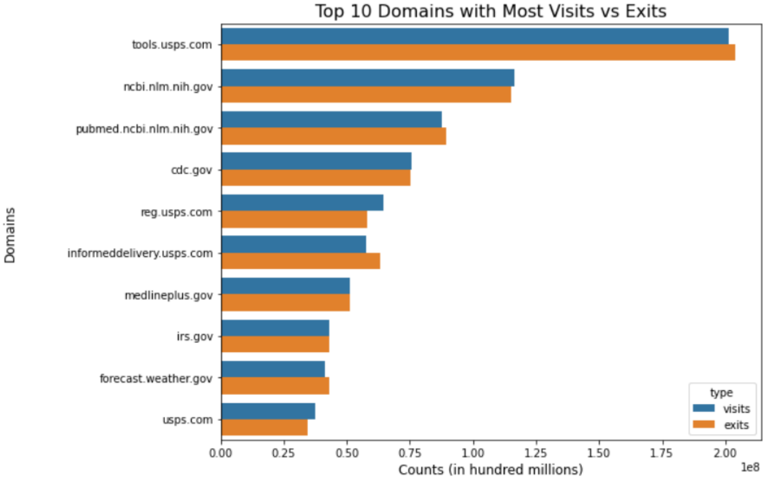
## Data Cleansing, Assumptions and Sanity Checks

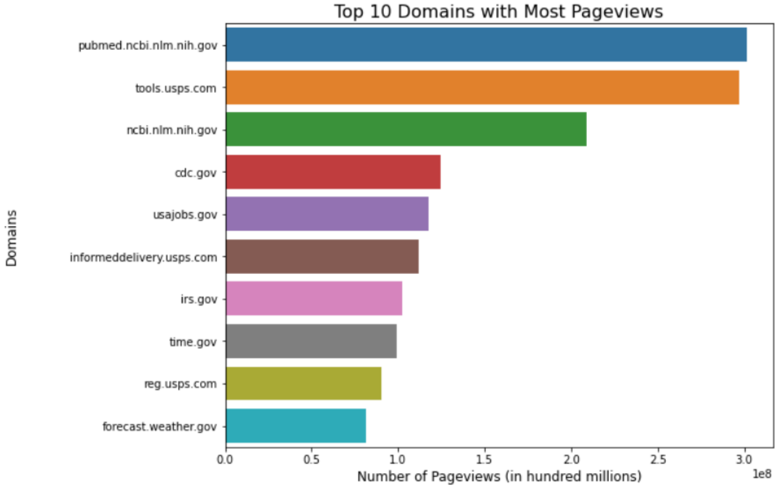
1. Assumptions on the dataset:
    a. Data sets with 30-day, 1-day, and real-time aggregated data are being analyzed separately and we do not try to validate data consistency between them. Some data sets are updated every day versus some are updated every 5 minutes by the government website. We don't know exactly what time data get refreshed by their server. We assume 12:01am PST is the moment all data get refreshed.
    b. Some variable exists in multiple data files and we assumed the data files with more correlated variables are more accurate. For example, "number of exits" variable exists in both "Visits to all domains" data file and the "Top exit pages" data file. We pick the exit data from the "Visits to all domains" because it has both visit and exit correlation in a single file, and we assumed it is more accurate.
2. Data Cleansed:
    a. Web domains with average session duration less than 1-minute are being dropped from the data set.
    b. Web domains with zero exit count are being dropped from the data set.
    c. Web page title equal to string "(not set)" cannot be identified as valid web domain access and is dropped from data set. (Appendix E.)
    d. Active visitors by city data - The top value for each data set was 'zz', which the team was not able to identify as a valid city, thus dropped from the analysis. Additionally, there was a city titled '(not set)' which was also dropped from our analysis.
    e. All data that spans the last 30 days will be filtered to the date ranges 04-04-2022 to 05-07-2022 in order to capture the most comparable data (Appendix F and section 3. What languages are users interacting with the most?).
3. Sanity Checks:
    a. Refer to appendix part D for Sanity Checks Conducted over the country and city active visitors data set.

## All Participating Websites (Aggregate)

1. ***What web pages have the most traffic? Least traffic? Which web pages do users exit the most frequently?***

The following two diagrams show top 10 web domains with most visits, most exits, and most page-views.

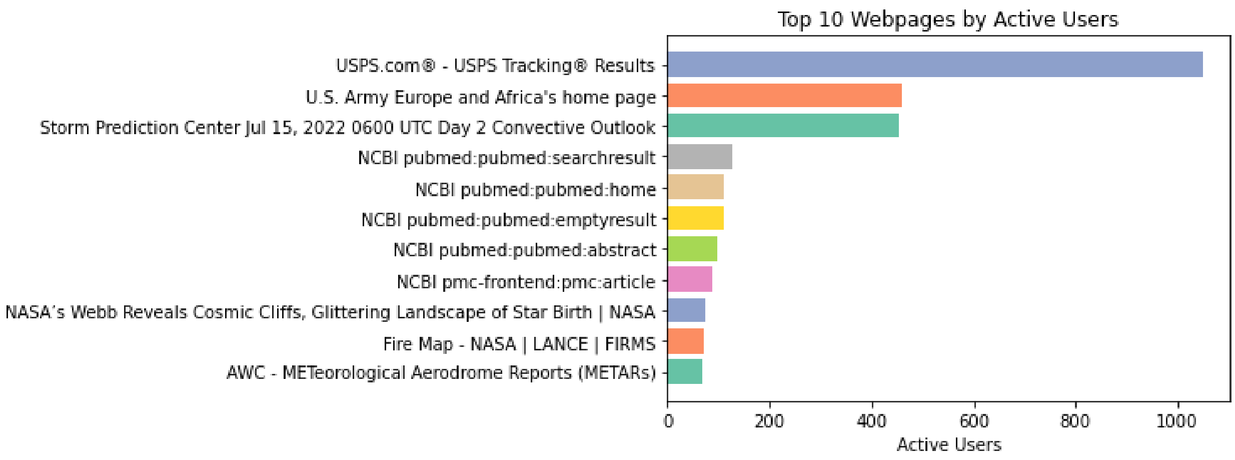




The tools.usps.com webpage has more than 200 million visits which has the most traffic. This webpage provides tracking tools from the US Postal Service to track mails and packages. The 2nd, 3rd and 4th places are nih.gov and cdc.gov webpages which belong to the Department of Health and Human Services. If we group by the number of pageviews, pubmed.ncbi.nlm.nih.gov webpage is the highest. The PubMed® webpage comprises more than 34 million citations for biomedical literature from MEDLINE, life science journals, and online books. It also includes the COVID-19 Information. The least traffic is the forecast.weather.gov webpage which provides weather forecast information from the Department of Commerce. Webpage being exited most frequently is the tools.usps.com webpage which we suspect users would exit immediately after using their online tracking tools.

2. ***Which web pages have the most users? Least amount of users?***

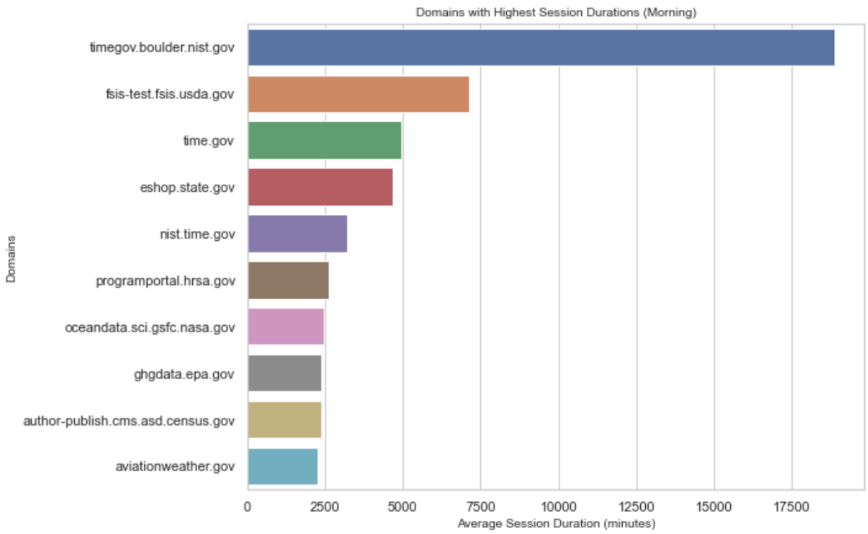
To start the analysis, the team started by sorting the pages dataframe by active users, allowing us to easily see the top  and bottom 10 web pages by active visitors. We compared these metrics with the day and nighttime dataset and found there to be no difference. With this, we decided to focus on the morning data.



As you can see, of the top 10 webpages, 6 are a part of the Department of Health and Human Services. This leads us to believe that people are still coming to government websites to keep up with uncertain health conditions in the world, presumably lingering effects and the current state of the COVID-19 pandemic. The webpage with the most active visitors is the government login page. This would make sense, considering government employees are required to login every day.

### *3. Which web pages do users stay on the longest?*

We sorted aggregate data set for the past 30 days by the 'avg_session_duration' in descending order for morning and evening datasets and elected to analyze the top 10 domains. As it is aggregate data from the past 30 days, there was no difference in the top 10 domains between the morning and evening datasets. As such, for questions (3 &4) we will display the morning sessions.

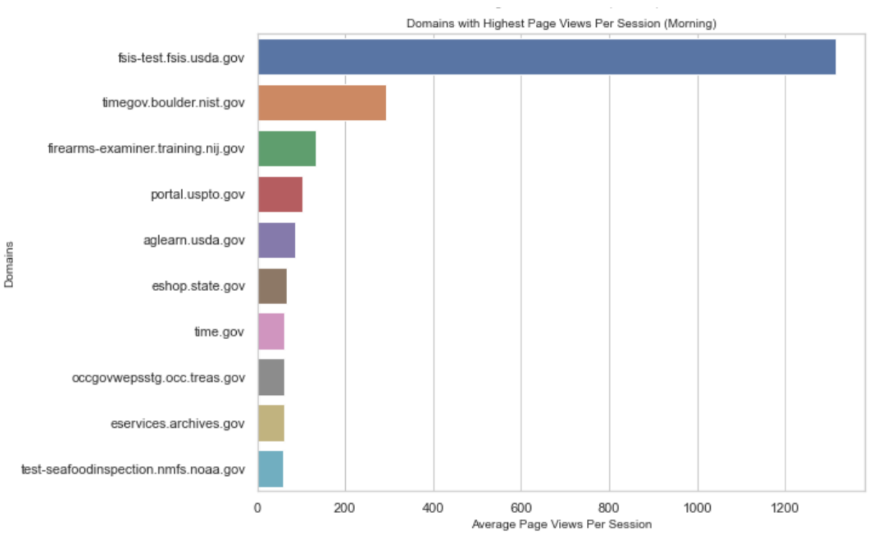


Interestingly 3 of the top 5 domains in terms of session duration were related to an official US government time website provided by the department of Commerce. The 2nd highest domain was related to Food Safety and Inspection Services, which led us to believe that lengthy food safety acts and regulations especially during COVID times might have contributed to users spending so much time on this site (leading to higher session duration).

### *4. Which web pages keep users browsing the most?*

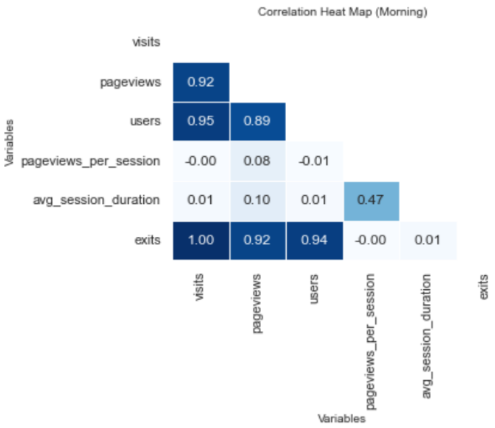
To determine which web pages have kept users browsing the most, we considered page views per session to be an appropriate metric. We sorted by page views per session in descending order and elected to analyze the top 10 domains respectively. The top value by far was the domain related to Food

Safety and Inspection Services which was also a top domain in question 3. After inspection of the website, we believe this site has the highest average page views because of the complex food regulations, numerous sub-chapters and sub-paragraphs within. (See Appendix Part B)



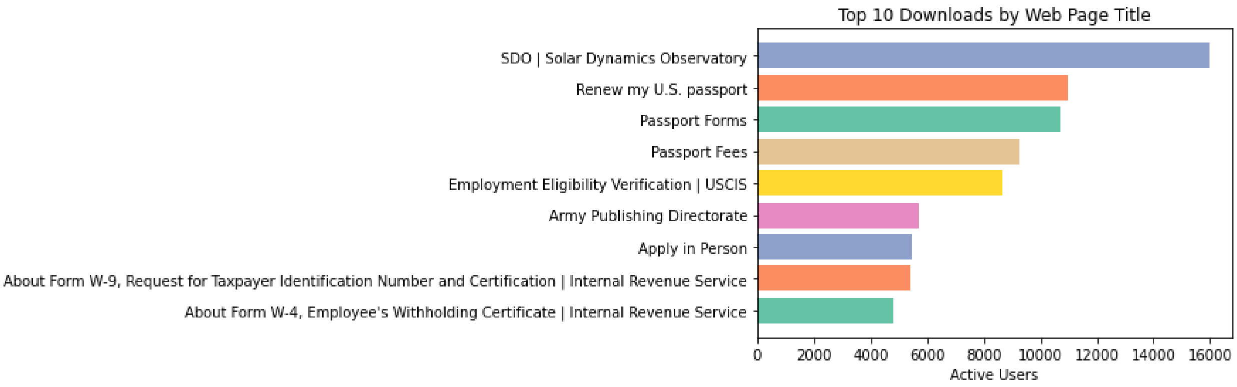
After noticing similar domains for questions 3 and 4, we decided to prepare a correlation heat map, to see if indeed these variables are correlated with each other.

The correlation between average session duration and pageviews per session was not as strong as expected with a value of 0.47. However, there were other stronger correlations of interest such as users and pageviews which were correlated at 0.89. This also inadvertently acted as a sanity check as visits and exits were perfectly correlated. Which is to be expected as ending a visit would correspond to one exit.



***5. What web pages have the most downloads? Least downloads?***

To start the analysis, the team started by sorting the downloads data frame by total events (total downloads), allowing us to easily see the top and bottom 10 web pages by downloads. We compared these metrics with the day and nighttime dataset and found there to be no difference. With this, we decided to focus on the morning data.



As we can see above, the majority of downloads has to do with renewing a passport, or handling items that relate to job specifics such as an W-9, W-4 or applications. Interestingly enough, during this time, the solar dynamics observatory web page contained the most downloads.

6. ***Which web pages do users exit the most frequently?***

This question has a close correlation with Question 1; therefore, this analysis was done in conjunction with Question 1 above.

## Department Specific Sub-Questions:

Based on the analysis conducted above, the team judgmentally decided to focus the sub-question on the following three departments:

1. Department of Health and Human Services
2. Department of Commerce
3. Postal Service (USPS)

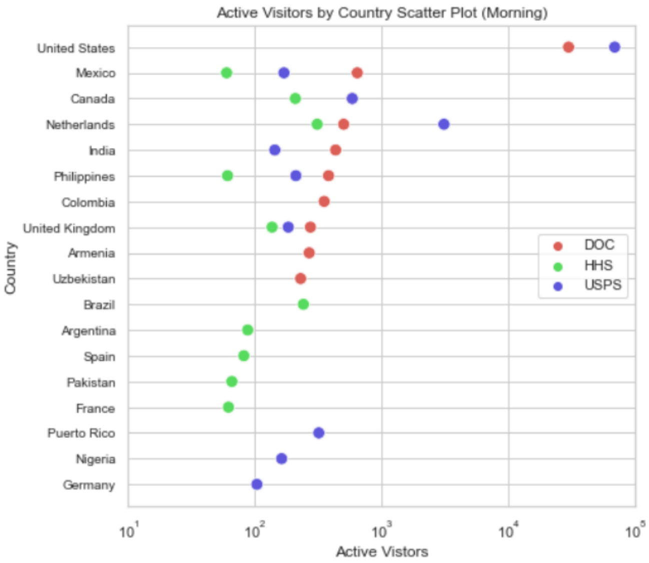
These departments were selected as they were frequently represented in the top domains in the questions related to our analysis of the aggregate datasets. The following table shows the dataset downloaded from individual departments, and the variables extracted from the dataset to answer the Department Specific Sub-Questions:

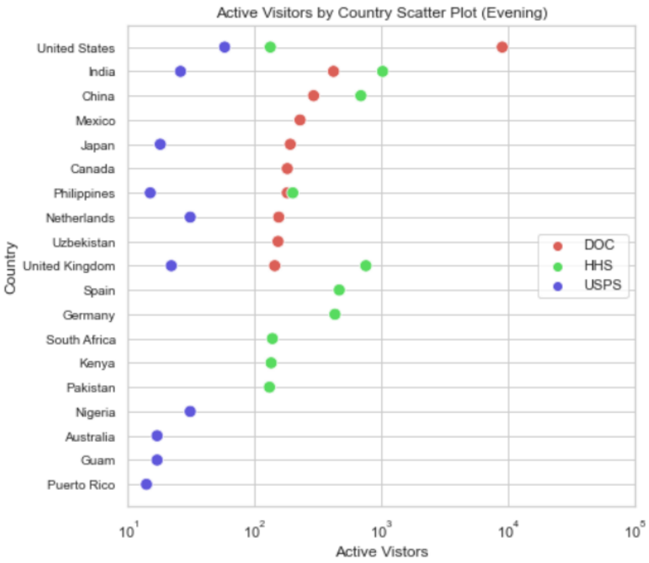
| Individual Departmental Websites: | | | |
|---|---|---|---|
| **Dataset Name** | **Time Aggregation** | **Variables** | **Data Type** |
| Language | 30 days | Date | datetime64[ns] |
| Visitors per country | Every 5 min real-time | Number of active visitors | integer |
| Visitors per city | Every 5 min real-time | Language | string |
| Devices: Desktop/mobile/tablet | 30 days | Country name | string |
| Web browsers | 30 days | City name | string |
| Operating systems | 30 days | Device type | string |
| OS & browser (combined) | 30 days | Browser type | string |
| Screen sizes | 30 days | OS type | string |
| | | OS version | string |
| | | Screen resolution | string |

1. ***From what countries are the users interacting with the web pages?***

The team used the visitors per country dataset for each of the above 3 departments, both for morning and evening data sets. Refer to the sanity checks over this data set in appendix part D.

To conduct the analysis, we concatenated the data into morning and evening data sets and selected the top 10 countries per department to visualize in a scatter plot. (*Note a log scale c-axis was used in order to provide a cleaner visualization)*

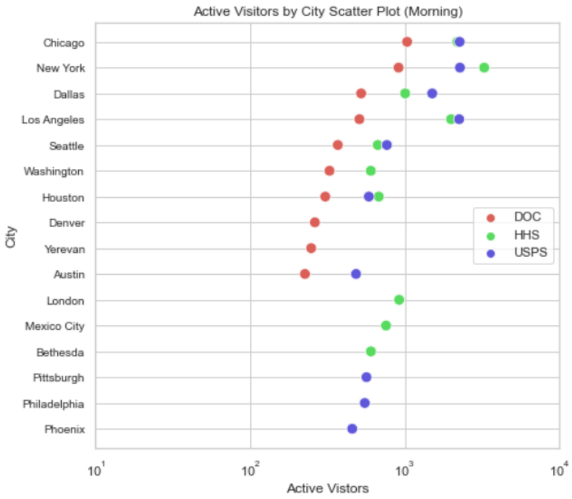


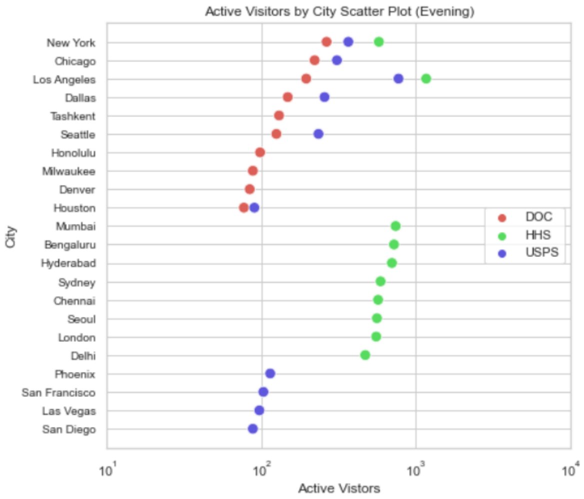


The United States (as we expected for US government websites) makes the bulk of the traffic for all departments. However, what is interesting to notice is that during the evening, foreign countries increase their traffic to these US government domains as their time zones are different. We can see that India and China represent big jumps in the evening along with other European countries such as Germany and Spain.

### 2. *From what cities are the users interacting with the web pages?*

To analyze this question we used the active visitors by country data set, for morning and evening times. The team had to filter out two rows from the filtered data sets, see appendix part D for details. Similarly to subquestion 1, we concatenated the data into morning and evening data sets and selected the top 10 cities per department to visualize in a scatter plot.

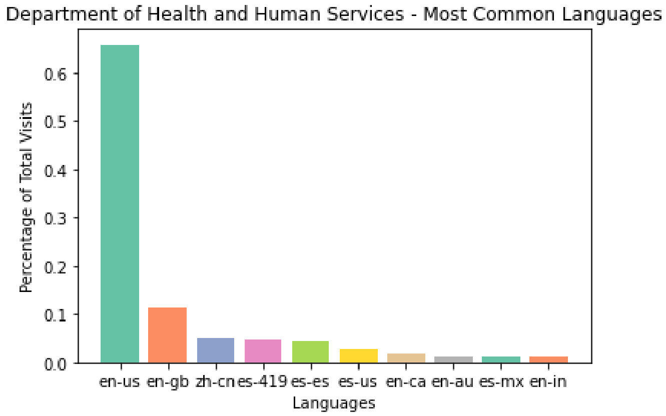




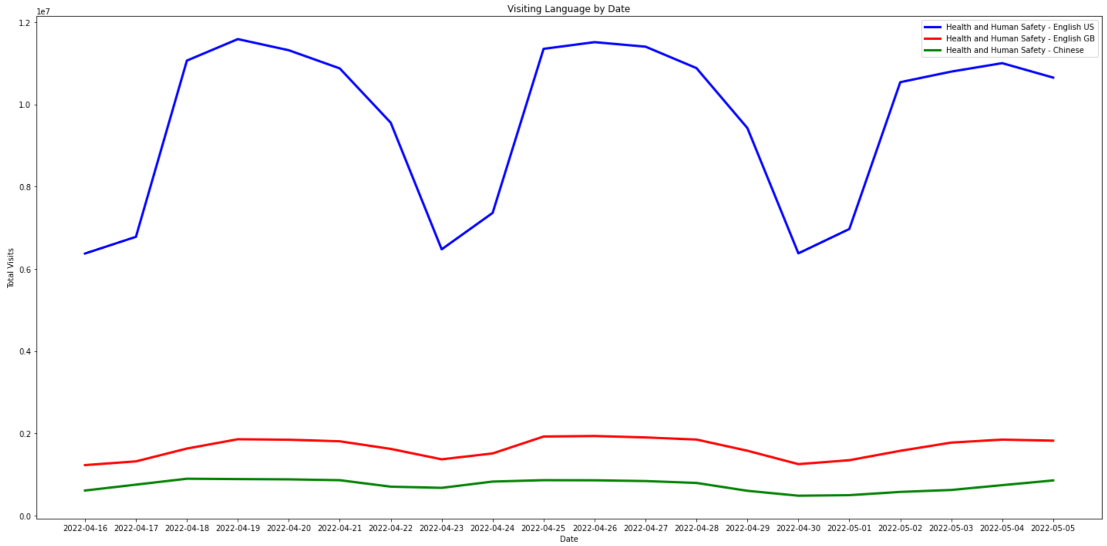
Across the three departments, US cities make up a good majority of the active visitor traffic. However, in the evening we do see several foreign cities appear, especially Indian cities such as Mumbai, Bengaluru, Hyderabad and Chennai. This jump in Indian cities in the evening data set is to be expected as India was second in traffic as seen in our analysis of the active visitors by country analysis.

### 3. *From what languages are the users interacting with the web pages?*

To handle this data, we decided to use and analyze as many dates as possible between all three groups, meaning all data from 04-04-2022 to 05-07-2022 (Appendix F). From here, We began looking into what language users were using to browse each department's web page. Both the department of commerce and USPS has more than 90% of total visitors using eng-us, but interestingly enough, the department of health and human services was more dispersed.



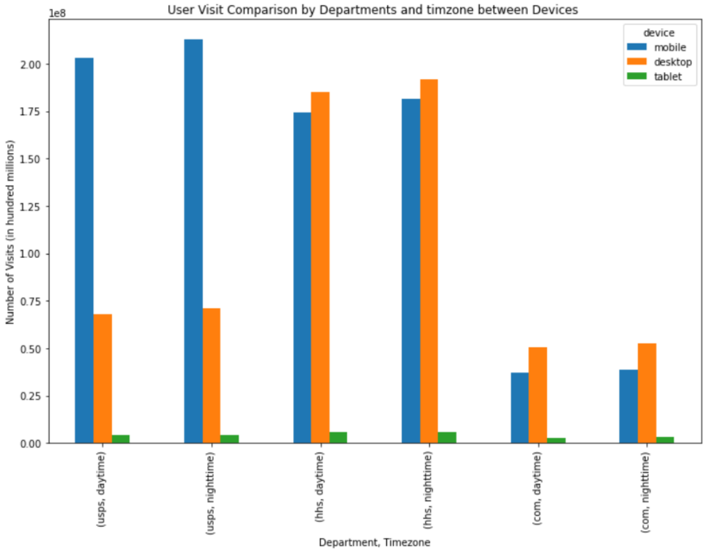
With this finding, we decided to investigate if the popularity of english-us has changed over time for this department. We found that over time, active visitors using english-us has ebbed and flowed. Furthermore, looking into the second most popular language, english-gb we see a slightly similar trend.Lastly, looking at the third most popular, chinese. For closer looks individually, see (Appendix G).

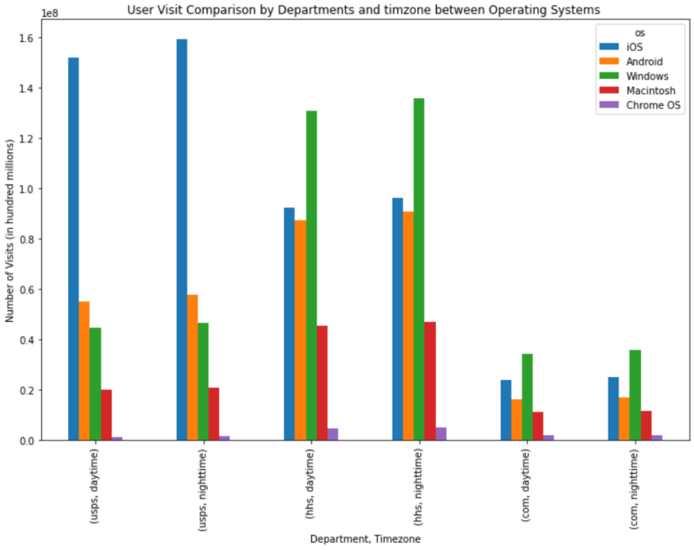


We can assume that some of the peaks could be around instances where there were more COVID-19 scares or heightened fear. We can also conclude, even though more people are browsing in English, it appears that users browsing in Chinese have less of a decline between the earlier months.

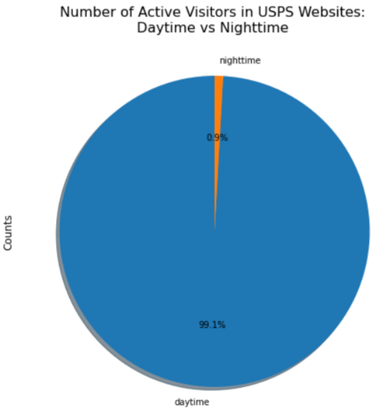
### 4. *How are visitors interacting with the web pages?*

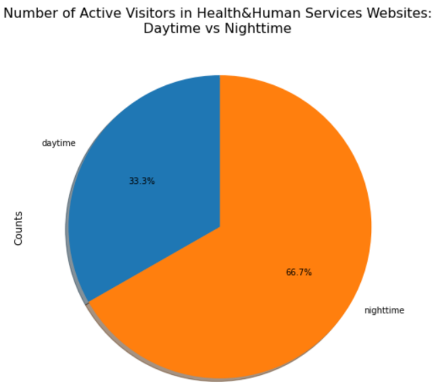
The following two diagrams compare what devices and operating systems used to access websites across the three departments between daytime and nighttime in the last 30 days.

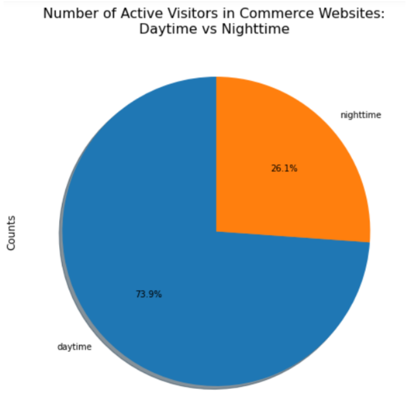


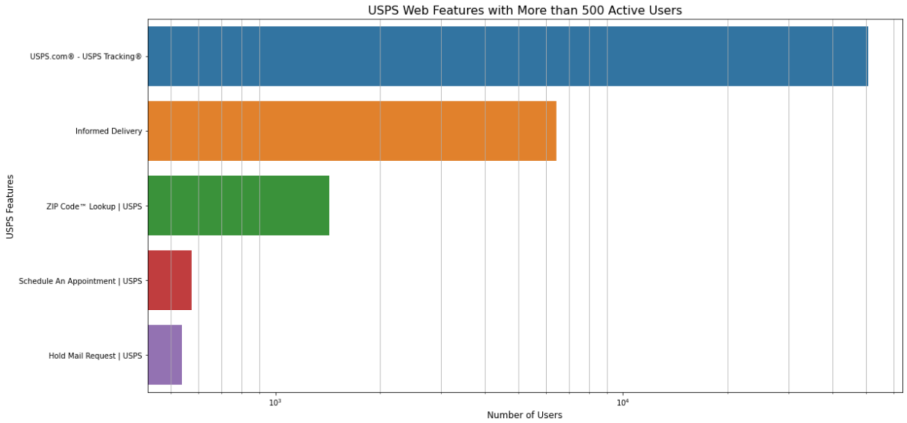


Mobile users have twice as many visits to US Postal Service(USPS) websites compared to desktop users. Health and Human Services(HHS) websites, on the other hand, has similar number of visits between mobile and desktop devices. Desktop users have just a slightly higher visit count compared to mobile users. The commerce(COM) websites have far less total number of visits compared to both USPS and HHS, and desktop users have a slightly higher visit count compared to mobile users. Although the total number of visits by tablet users are far less compared to both mobile and desktop, they are evenly distributed among the three departments. In addition, all three departments show similar daytime and nighttime counts with nighttime counts slightly higher. If we compare operating systems, using iPhoneOS(iOS) to access USPS websites has far higher visits compared to Android and Windows (more than twice as many). In contrast, Windows OS is higher in both HHS and COM. Again, daytime and nighttime visits show similar patterns among the three departments with nighttime visits slightly higher. The following three pie charts show snapshots of the number of visitors between daytime and nighttime. The snapshots are taken at 11am and 11pm on the same day. USPS has 99% daytime visitors and Commerce has 74% daytime visitors in contrast to the HHS which has 66.7% nighttime visitors.









To explain why USPS has such extreme differences between mobile and desktop, and also between daytime and nighttime. The bar chart above shows the USPS web features with more than 500 simultaneous users. The top three features are ‘USPS Tracking’, ‘Informed Delivery’, and ‘Zip Code Lookup’. These features are more conveniently used in mobile devices and usually done in daytime.

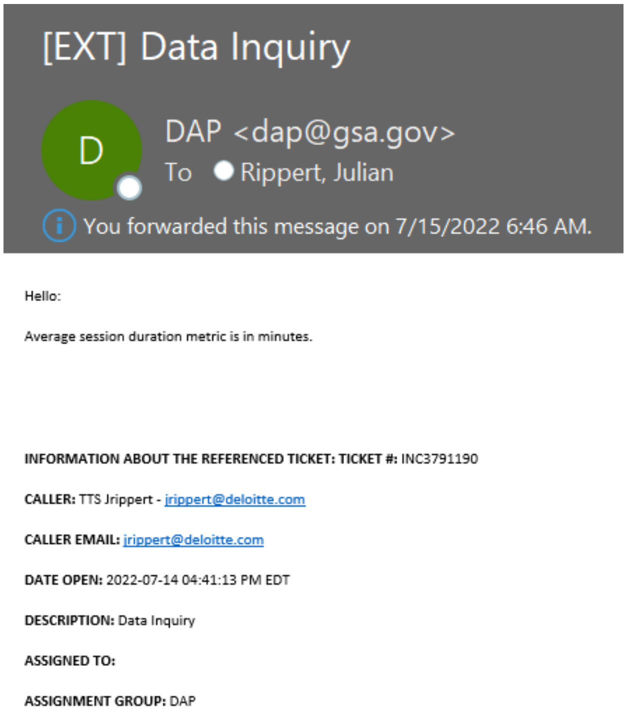
## Challenges & Conclusions:

It is important to note that according to the data source that not every government website is represented in this data. These domains are from the government’s Digital Analytics Program (DAP), which encompasses around 400 executive branch government domains, across about 5,700 total websites. Due to the domains not covering every website in the US government , it is a challenge to extrapolate our insights from this project to all US government websites.

Throughout our exploratory research project, we noticed how frequently there were domains related to COVID represented in the top 10’s of our data sorts. It would be interesting to analyze these data sets prior to the onset of COVID to inspect the before and after. As COVID continues, and regulations and notices for a wide variety of industries including education, medicine, retail, we expect the pages will continue to make up a bulk of the traffic to US government websites.

## Appendix:

A. Confirmation of Session Duration measurement: The team validated per an IT ticket with the DAP Program that the avg session duration column data is measured in minutes.



[EXT] Data Inquiry

DAP <dap@gsa.gov>
To Rippert, Julian

You forwarded this message on 7/15/2022 6:46 AM.

Hello:

Average session duration metric is in minutes.

**INFORMATION ABOUT THE REFERENCED TICKET: TICKET #:** INC3791190

**CALLER:** TTS Jrippert - jrippert@deloitte.com

**CALLER EMAIL:** jrippert@deloitte.com

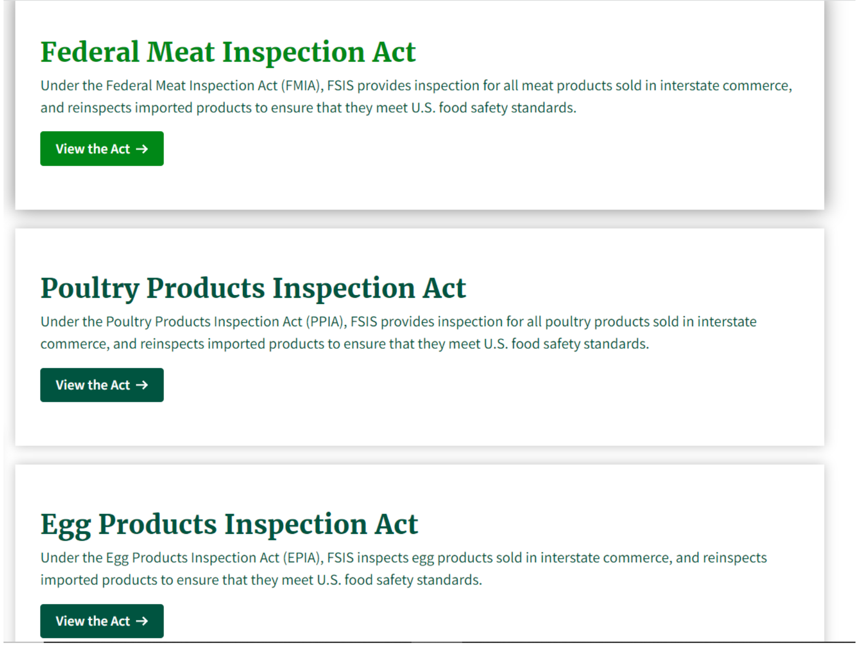
**DATE OPEN:** 2022-07-14 04:41:13 PM EDT

**DESCRIPTION:** Data Inquiry

**ASSIGNED TO:**

**ASSIGNMENT GROUP:** DAP

B. Inspection of the Food Safety and Inspection Services website. We noted this site is incredibly regulatory heavy, where consumers and businesses frequent to inspect the latest safety acts, guidelines and notices. Within each of these, there are incredible amounts of sub-pages and hyperlinks, where users can dive deeper into individual paragraphs of regulations for instance.



Federal Meat Inspection Act

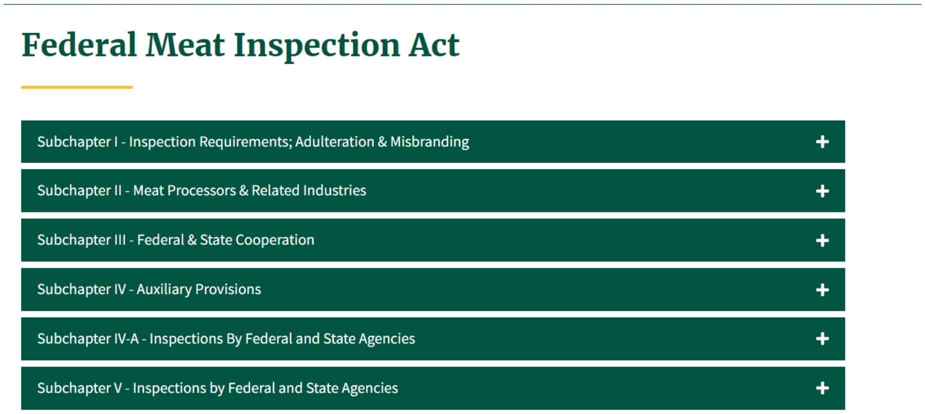
Under the Federal Meat Inspection Act (FMIA), FSIS provides inspection for all meat products sold in interstate commerce, and reinspects imported products to ensure that they meet U.S. food safety standards.

Poultry Products Inspection Act

Under the Poultry Products Inspection Act (PPIA), FSIS provides inspection for all poultry products sold in interstate commerce, and reinspects imported products to ensure that they meet U.S. food safety standards.

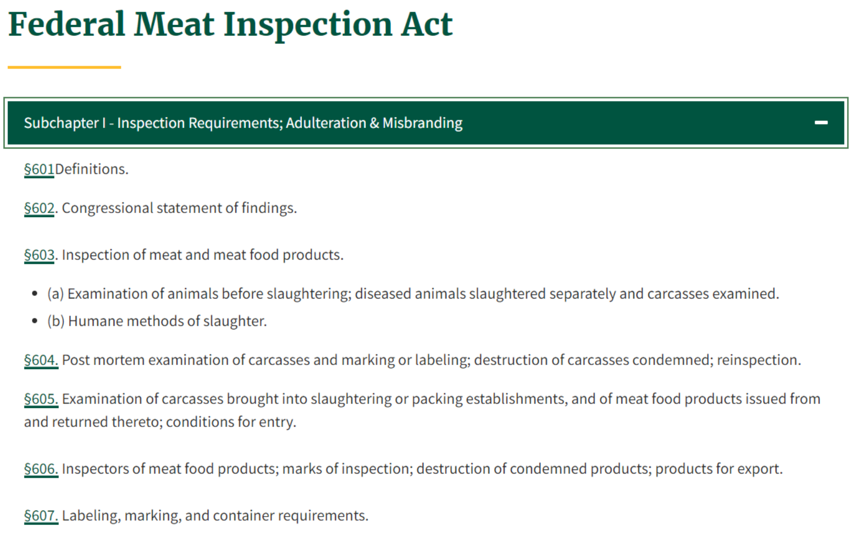
Egg Products Inspection Act

Under the Egg Products Inspection Act (EPIA), FSIS inspects egg products sold in interstate commerce, and reinspects imported products to ensure that they meet U.S. food safety standards.



## Federal Meat Inspection Act

- Subchapter I - Inspection Requirements; Adulteration & Misbranding +
- Subchapter II - Meat Processors & Related Industries +
- Subchapter III - Federal & State Cooperation +
- Subchapter IV - Auxiliary Provisions +
- Subchapter IV-A - Inspections By Federal and State Agencies +
- Subchapter V - Inspections by Federal and State Agencies +



## Federal Meat Inspection Act

Subchapter I - Inspection Requirements; Adulteration & Misbranding −

§601Definitions.

§602. Congressional statement of findings.

§603. Inspection of meat and meat food products.

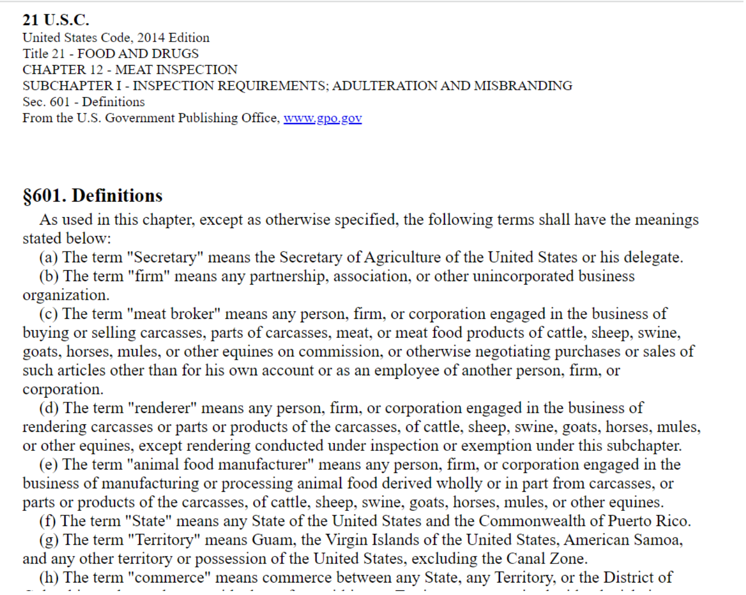
- (a) Examination of animals before slaughtering; diseased animals slaughtered separately and carcasses examined.
- (b) Humane methods of slaughter.

§604. Post mortem examination of carcasses and marking or labeling; destruction of carcasses condemned; reinspection.

§605. Examination of carcasses brought into slaughtering or packing establishments, and of meat food products issued from and returned thereto; conditions for entry.

§606. Inspectors of meat food products; marks of inspection; destruction of condemned products; products for export.

§607. Labeling, marking, and container requirements.



**21 U.S.C.**
United States Code, 2014 Edition
Title 21 - FOOD AND DRUGS
CHAPTER 12 - MEAT INSPECTION
SUBCHAPTER I - INSPECTION REQUIREMENTS; ADULTERATION AND MISBRANDING
Sec. 601 - Definitions
From the U.S. Government Publishing Office, www.gpo.gov

### §601. Definitions

As used in this chapter, except as otherwise specified, the following terms shall have the meanings stated below:

(a) The term "Secretary" means the Secretary of Agriculture of the United States or his delegate.

(b) The term "firm" means any partnership, association, or other unincorporated business organization.

(c) The term "meat broker" means any person, firm, or corporation engaged in the business of buying or selling carcasses, parts of carcasses, meat, or meat food products of cattle, sheep, swine, goats, horses, mules, or other equines on commission, or otherwise negotiating purchases or sales of such articles other than for his own account or as an employee of another person, firm, or corporation.

(d) The term "renderer" means any person, firm, or corporation engaged in the business of rendering carcasses or parts or products of the carcasses, of cattle, sheep, swine, goats, horses, mules, or other equines, except rendering conducted under inspection or exemption under this subchapter.

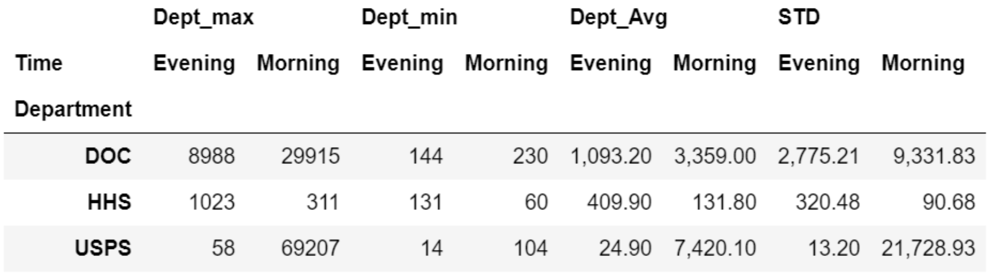
(e) The term "animal food manufacturer" means any person, firm, or corporation engaged in the business of manufacturing or processing animal food derived wholly or in part from carcasses, or parts or products of the carcasses, of cattle, sheep, swine, goats, horses, mules, or other equines.

(f) The term "State" means any State of the United States and the Commonwealth of Puerto Rico.

(g) The term "Territory" means Guam, the Virgin Islands of the United States, American Samoa, and any other territory or possession of the United States, excluding the Canal Zone.

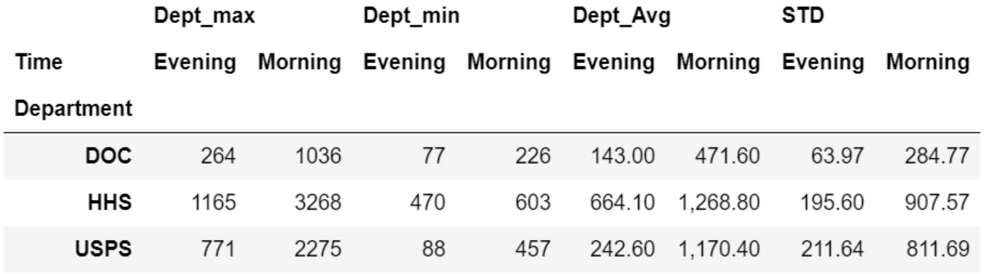
(h) The term "commerce" means commerce between any State, any Territory, or the District of

C. We conducted groupby operations for countries and cities across departments and inspected some descriptive statistics for the active visitors variable.
By Country:



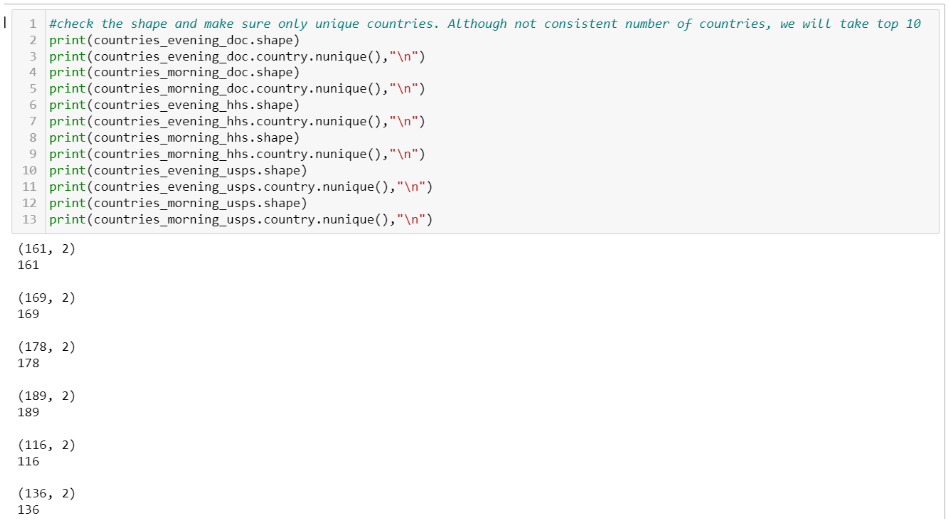
| | Dept_max | | Dept_min | | Dept_Avg | | STD | |
|---|---|---|---|---|---|---|---|---|
| Time | Evening | Morning | Evening | Morning | Evening | Morning | Evening | Morning |
| Department | | | | | | | | |
| **DOC** | 8988 | 29915 | 144 | 230 | 1,093.20 | 3,359.00 | 2,775.21 | 9,331.83 |
| **HHS** | 1023 | 311 | 131 | 60 | 409.90 | 131.80 | 320.48 | 90.68 |
| **USPS** | 58 | 69207 | 14 | 104 | 24.90 | 7,420.10 | 13.20 | 21,728.93 |

By City:



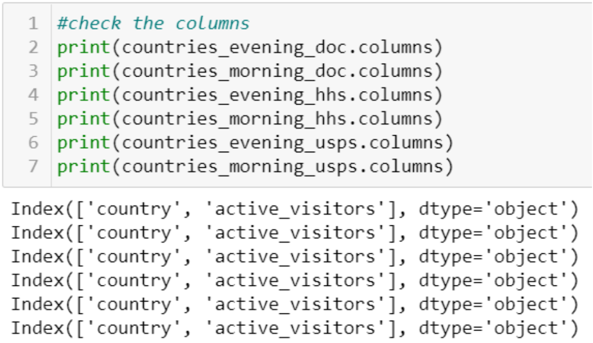
| | Dept_max | | Dept_min | | Dept_Avg | | STD | |
|---|---|---|---|---|---|---|---|---|
| Time | Evening | Morning | Evening | Morning | Evening | Morning | Evening | Morning |
| Department | | | | | | | | |
| **DOC** | 264 | 1036 | 77 | 226 | 143.00 | 471.60 | 63.97 | 284.77 |
| **HHS** | 1165 | 3268 | 470 | 603 | 664.10 | 1,268.80 | 195.60 | 907.57 |
| **USPS** | 771 | 2275 | 88 | 457 | 242.60 | 1,170.40 | 211.64 | 811.69 |

D. The team conducted the following sanity checks for the country and city data.
Country Data Set:



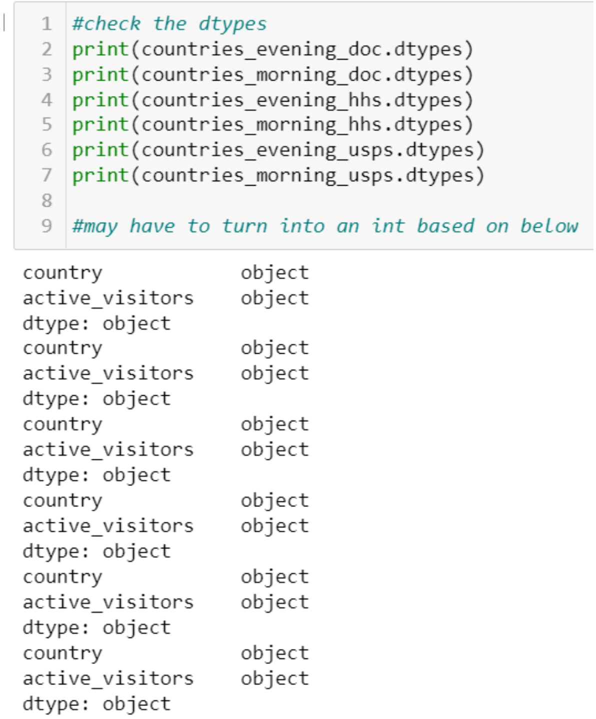
```
#check the shape and make sure only unique countries. Although not consistent number of countries, we will take top 10
print(countries_evening_doc.shape)
print(countries_evening_doc.country.nunique(),"\n")
print(countries_morning_doc.shape)
print(countries_morning_doc.country.nunique(),"\n")
print(countries_evening_hhs.shape)
print(countries_evening_hhs.country.nunique(),"\n")
print(countries_morning_hhs.shape)
print(countries_morning_hhs.country.nunique(),"\n")
print(countries_evening_usps.shape)
print(countries_evening_usps.country.nunique(),"\n")
print(countries_morning_usps.shape)
print(countries_morning_usps.country.nunique(),"\n")
```

```
(161, 2)
161

(169, 2)
169

(178, 2)
178

(189, 2)
189

(116, 2)
116

(136, 2)
136
```



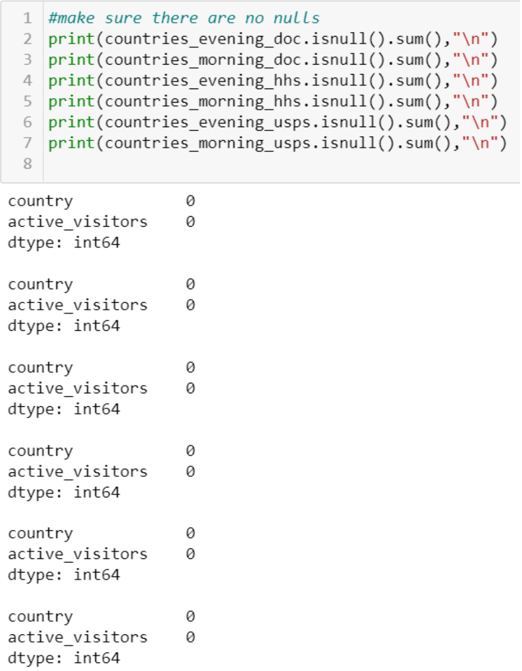
```
#check the columns
print(countries_evening_doc.columns)
print(countries_morning_doc.columns)
print(countries_evening_hhs.columns)
print(countries_morning_hhs.columns)
print(countries_evening_usps.columns)
print(countries_morning_usps.columns)
```

```
Index(['country', 'active_visitors'], dtype='object')
Index(['country', 'active_visitors'], dtype='object')
Index(['country', 'active_visitors'], dtype='object')
Index(['country', 'active_visitors'], dtype='object')
Index(['country', 'active_visitors'], dtype='object')
Index(['country', 'active_visitors'], dtype='object')
```



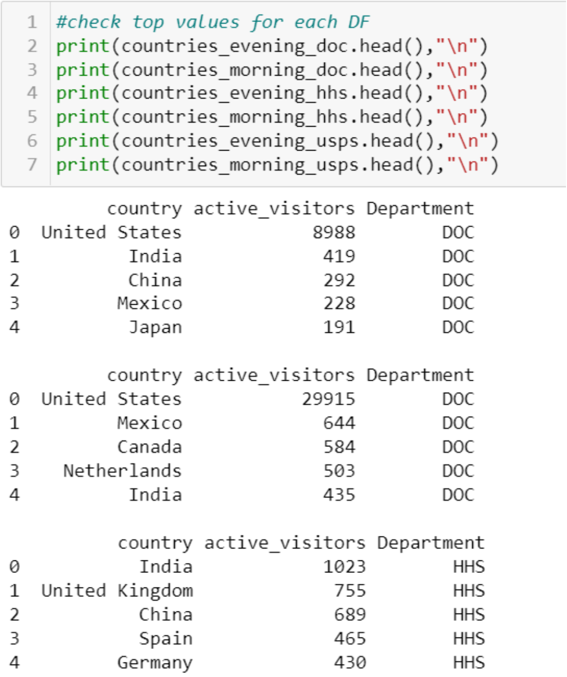
```
#check the dtypes
print(countries_evening_doc.dtypes)
print(countries_morning_doc.dtypes)
print(countries_evening_hhs.dtypes)
print(countries_morning_hhs.dtypes)
print(countries_evening_usps.dtypes)
print(countries_morning_usps.dtypes)

#may have to turn into an int based on below
```

```
country            object
active_visitors    object
dtype: object
country            object
active_visitors    object
dtype: object
country            object
active_visitors    object
dtype: object
country            object
active_visitors    object
dtype: object
country            object
active_visitors    object
dtype: object
country            object
active_visitors    object
dtype: object
```



```
#make sure there are no nulls
print(countries_evening_doc.isnull().sum(),"\n")
print(countries_morning_doc.isnull().sum(),"\n")
print(countries_evening_hhs.isnull().sum(),"\n")
print(countries_morning_hhs.isnull().sum(),"\n")
print(countries_evening_usps.isnull().sum(),"\n")
print(countries_morning_usps.isnull().sum(),"\n")

```

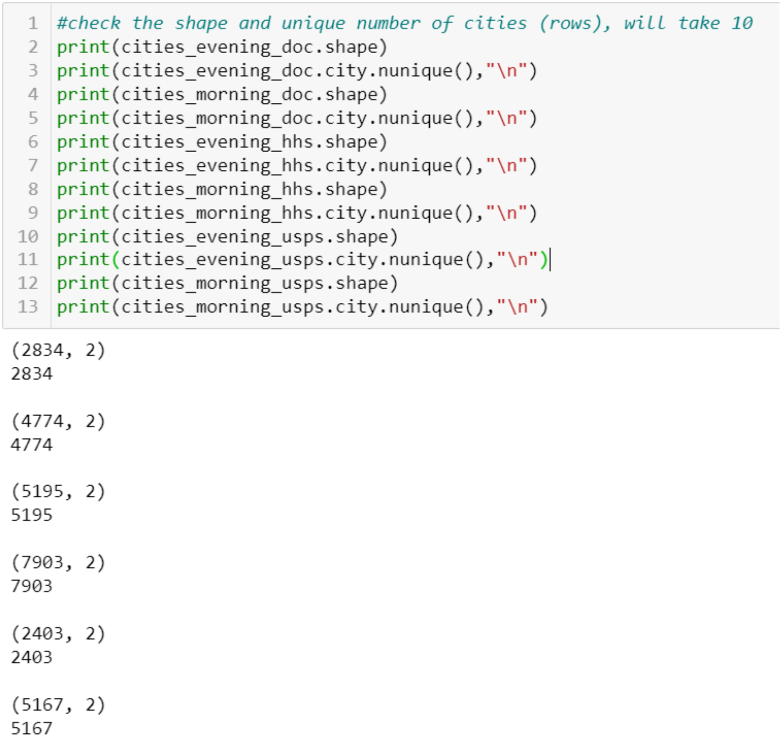
```
country            0
active_visitors    0
dtype: int64

country            0
active_visitors    0
dtype: int64

country            0
active_visitors    0
dtype: int64

country            0
active_visitors    0
dtype: int64

country            0
active_visitors    0
dtype: int64

country            0
active_visitors    0
dtype: int64
```



```
#check top values for each DF
print(countries_evening_doc.head(),"\n")
print(countries_morning_doc.head(),"\n")
print(countries_evening_hhs.head(),"\n")
print(countries_morning_hhs.head(),"\n")
print(countries_evening_usps.head(),"\n")
print(countries_morning_usps.head(),"\n")
```

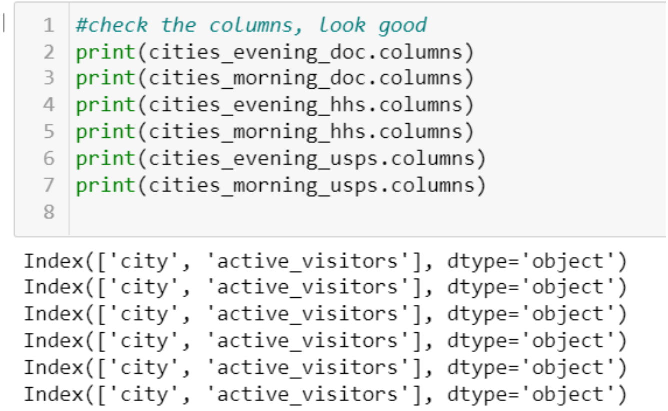
```
         country  active_visitors Department
0  United States             8988        DOC
1          India              419        DOC
2          China              292        DOC
3         Mexico              228        DOC
4          Japan              191        DOC

         country  active_visitors Department
0  United States            29915        DOC
1         Mexico              644        DOC
2         Canada              584        DOC
3    Netherlands              503        DOC
4          India              435        DOC

          country  active_visitors Department
0           India             1023        HHS
1  United Kingdom              755        HHS
2           China              689        HHS
3           Spain              465        HHS
4         Germany              430        HHS
```

City Data Set:



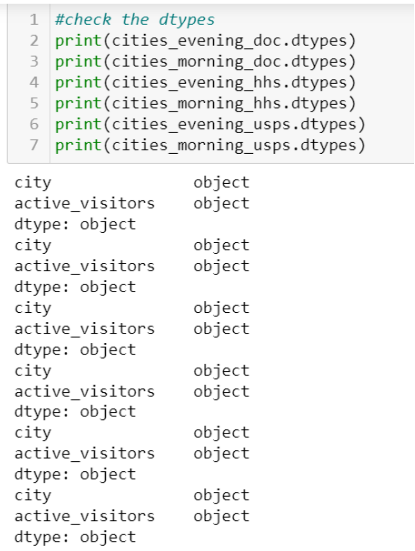
```
#check the shape and unique number of cities (rows), will take 10
print(cities_evening_doc.shape)
print(cities_evening_doc.city.nunique(),"\n")
print(cities_morning_doc.shape)
print(cities_morning_doc.city.nunique(),"\n")
print(cities_evening_hhs.shape)
print(cities_evening_hhs.city.nunique(),"\n")
print(cities_morning_hhs.shape)
print(cities_morning_hhs.city.nunique(),"\n")
print(cities_evening_usps.shape)
print(cities_evening_usps.city.nunique(),"\n")
print(cities_morning_usps.shape)
print(cities_morning_usps.city.nunique(),"\n")
```

```
(2834, 2)
2834

(4774, 2)
4774

(5195, 2)
5195

(7903, 2)
7903

(2403, 2)
2403

(5167, 2)
5167
```



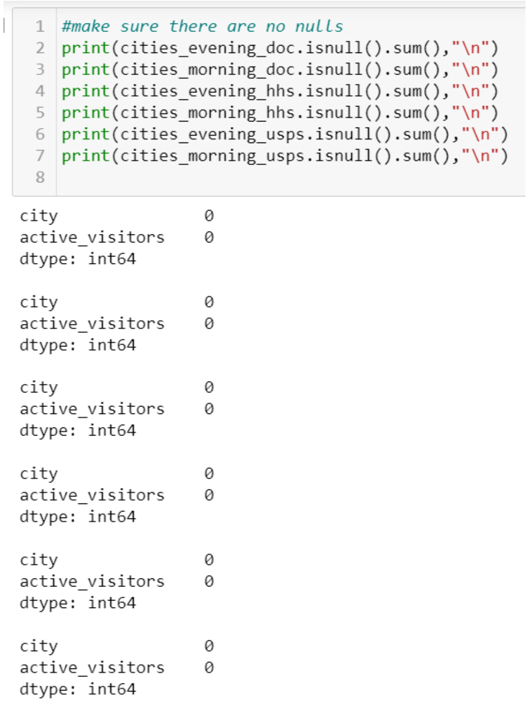
```
#check the columns, look good
print(cities_evening_doc.columns)
print(cities_morning_doc.columns)
print(cities_evening_hhs.columns)
print(cities_morning_hhs.columns)
print(cities_evening_usps.columns)
print(cities_morning_usps.columns)

```

```
Index(['city', 'active_visitors'], dtype='object')
Index(['city', 'active_visitors'], dtype='object')
Index(['city', 'active_visitors'], dtype='object')
Index(['city', 'active_visitors'], dtype='object')
Index(['city', 'active_visitors'], dtype='object')
Index(['city', 'active_visitors'], dtype='object')
```



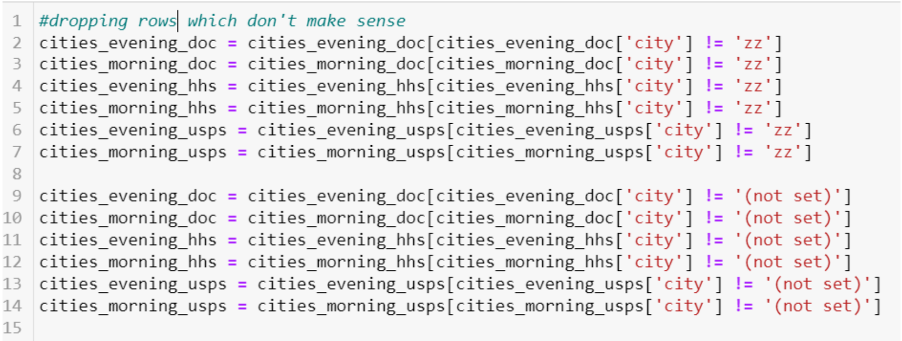
```
#check the dtypes
print(cities_evening_doc.dtypes)
print(cities_morning_doc.dtypes)
print(cities_evening_hhs.dtypes)
print(cities_morning_hhs.dtypes)
print(cities_evening_usps.dtypes)
print(cities_morning_usps.dtypes)
```

```
city               object
active_visitors    object
dtype: object
city               object
active_visitors    object
dtype: object
city               object
active_visitors    object
dtype: object
city               object
active_visitors    object
dtype: object
city               object
active_visitors    object
dtype: object
city               object
active_visitors    object
dtype: object
```



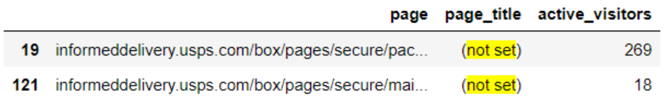
```
#make sure there are no nulls
print(cities_evening_doc.isnull().sum(),"\n")
print(cities_morning_doc.isnull().sum(),"\n")
print(cities_evening_hhs.isnull().sum(),"\n")
print(cities_morning_hhs.isnull().sum(),"\n")
print(cities_evening_usps.isnull().sum(),"\n")
print(cities_morning_usps.isnull().sum(),"\n")

```

```
city               0
active_visitors    0
dtype: int64

city               0
active_visitors    0
dtype: int64

city               0
active_visitors    0
dtype: int64

city               0
active_visitors    0
dtype: int64

city               0
active_visitors    0
dtype: int64

city               0
active_visitors    0
dtype: int64
```



```
#dropping rows which don't make sense
cities_evening_doc = cities_evening_doc[cities_evening_doc['city'] != 'zz']
cities_morning_doc = cities_morning_doc[cities_morning_doc['city'] != 'zz']
cities_evening_hhs = cities_evening_hhs[cities_evening_hhs['city'] != 'zz']
cities_morning_hhs = cities_morning_hhs[cities_morning_hhs['city'] != 'zz']
cities_evening_usps = cities_evening_usps[cities_evening_usps['city'] != 'zz']
cities_morning_usps = cities_morning_usps[cities_morning_usps['city'] != 'zz']

cities_evening_doc = cities_evening_doc[cities_evening_doc['city'] != '(not set)']
cities_morning_doc = cities_morning_doc[cities_morning_doc['city'] != '(not set)']
cities_evening_hhs = cities_evening_hhs[cities_evening_hhs['city'] != '(not set)']
cities_morning_hhs = cities_morning_hhs[cities_morning_hhs['city'] != '(not set)']
cities_evening_usps = cities_evening_usps[cities_evening_usps['city'] != '(not set)']
cities_morning_usps = cities_morning_usps[cities_morning_usps['city'] != '(not set)']

```

E. Unidentified Page Title is removed from our data set such as the following:



| | page | page_title | active_visitors |
|---|---|---|---|
| 19 | informeddelivery.usps.com/box/pages/secure/pac... | (not set) | 269 |
| 121 | informeddelivery.usps.com/box/pages/secure/mai... | (not set) | 18 |

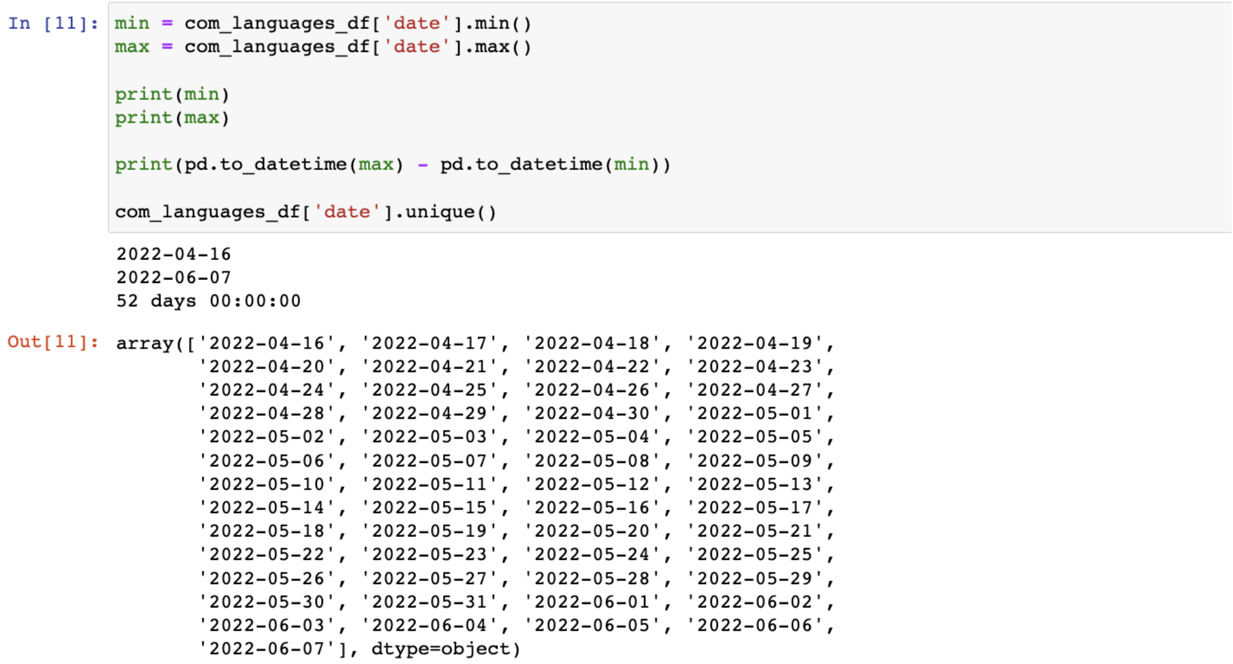
F. Date range differences

To handle this data, we decided to use and analyze as many dates as possible between all three groups, meaning all data from 04-04-2022 to 05-07-2022.

To analyze this data we started by looking at the language data for each chosen department.

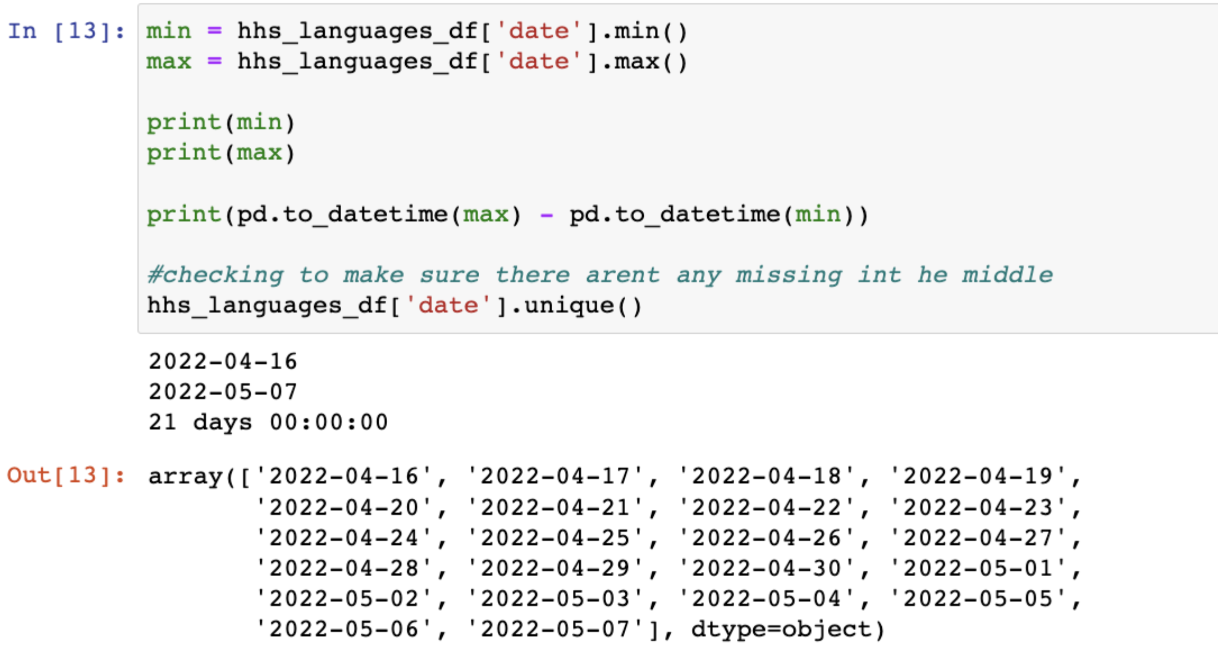
Within the language data, we found that each department has a different range of dates that it is tracking, however, all datasets contain the same number of rows (10,000). For example, the Department of Commerce contains data about the visits per language ranging from date 04-04-2022 through 06-07-2022, while the Department of Health and Human Safety ranges dates from 04-16-2022 through 05-07-2022, and the United States Postal Service ranges dates from 04-16-2022 through 06-03-2022. After filtering the data sets to these dates, we made sure to look at all the unique dates checking to see if there were any dates missing between the ranges in any of the datasets. There were not.

Department of Commerce



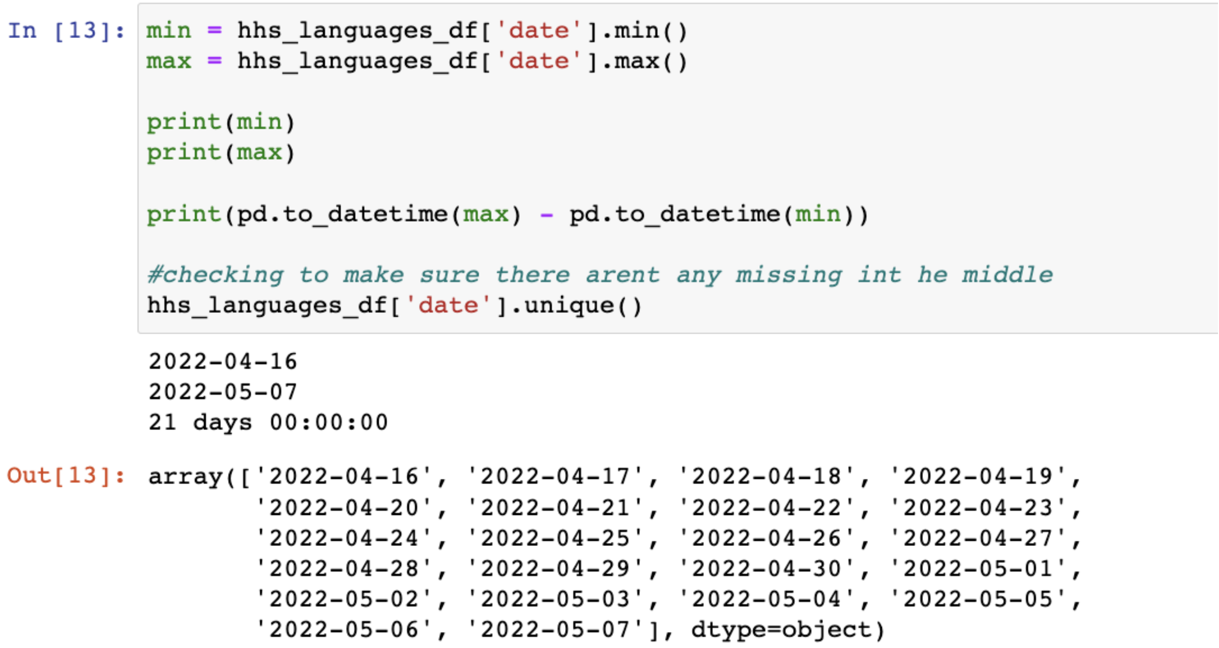
```
In [11]: min = com_languages_df['date'].min()
         max = com_languages_df['date'].max()

         print(min)
         print(max)

         print(pd.to_datetime(max) - pd.to_datetime(min))

         com_languages_df['date'].unique()

         2022-04-16
         2022-06-07
         52 days 00:00:00

Out[11]: array(['2022-04-16', '2022-04-17', '2022-04-18', '2022-04-19',
                '2022-04-20', '2022-04-21', '2022-04-22', '2022-04-23',
                '2022-04-24', '2022-04-25', '2022-04-26', '2022-04-27',
                '2022-04-28', '2022-04-29', '2022-04-30', '2022-05-01',
                '2022-05-02', '2022-05-03', '2022-05-04', '2022-05-05',
                '2022-05-06', '2022-05-07', '2022-05-08', '2022-05-09',
                '2022-05-10', '2022-05-11', '2022-05-12', '2022-05-13',
                '2022-05-14', '2022-05-15', '2022-05-16', '2022-05-17',
                '2022-05-18', '2022-05-19', '2022-05-20', '2022-05-21',
                '2022-05-22', '2022-05-23', '2022-05-24', '2022-05-25',
                '2022-05-26', '2022-05-27', '2022-05-28', '2022-05-29',
                '2022-05-30', '2022-05-31', '2022-06-01', '2022-06-02',
                '2022-06-03', '2022-06-04', '2022-06-05', '2022-06-06',
                '2022-06-07'], dtype=object)
```

Department of Health and Human Services



```
In [13]: min = hhs_languages_df['date'].min()
         max = hhs_languages_df['date'].max()

         print(min)
         print(max)

         print(pd.to_datetime(max) - pd.to_datetime(min))

         #checking to make sure there arent any missing int he middle
         hhs_languages_df['date'].unique()

         2022-04-16
         2022-05-07
         21 days 00:00:00

Out[13]: array(['2022-04-16', '2022-04-17', '2022-04-18', '2022-04-19',
                '2022-04-20', '2022-04-21', '2022-04-22', '2022-04-23',
                '2022-04-24', '2022-04-25', '2022-04-26', '2022-04-27',
                '2022-04-28', '2022-04-29', '2022-04-30', '2022-05-01',
                '2022-05-02', '2022-05-03', '2022-05-04', '2022-05-05',
                '2022-05-06', '2022-05-07'], dtype=object)
```

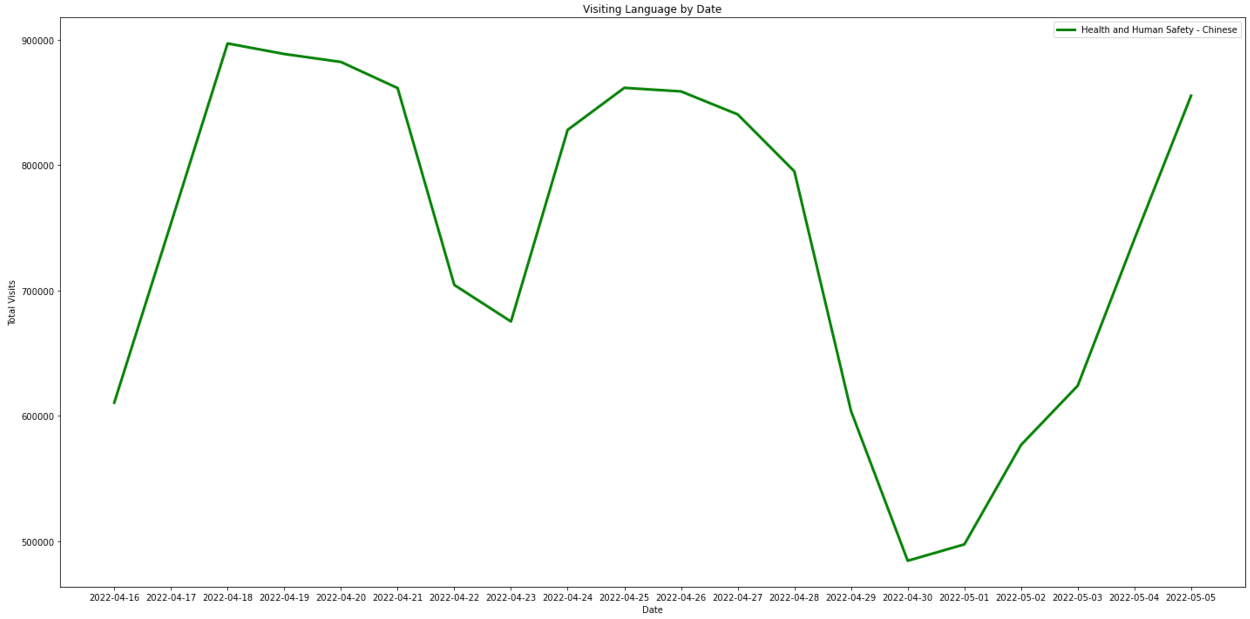
United States Postal Service



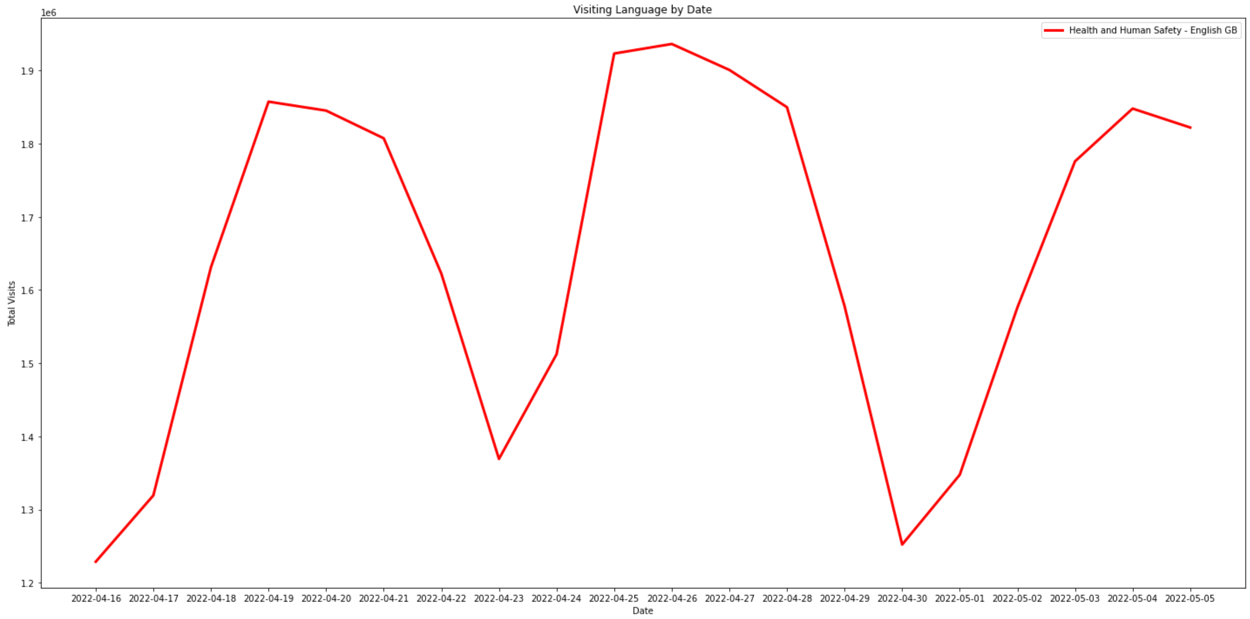
```
In [13]: min = hhs_languages_df['date'].min()
         max = hhs_languages_df['date'].max()

         print(min)
         print(max)

         print(pd.to_datetime(max) - pd.to_datetime(min))

         #checking to make sure there arent any missing int he middle
         hhs_languages_df['date'].unique()

         2022-04-16
         2022-05-07
         21 days 00:00:00

Out[13]: array(['2022-04-16', '2022-04-17', '2022-04-18', '2022-04-19',
                '2022-04-20', '2022-04-21', '2022-04-22', '2022-04-23',
                '2022-04-24', '2022-04-25', '2022-04-26', '2022-04-27',
                '2022-04-28', '2022-04-29', '2022-04-30', '2022-05-01',
                '2022-05-02', '2022-05-03', '2022-05-04', '2022-05-05',
                '2022-05-06', '2022-05-07'], dtype=object)
```

G. Visits by Langage

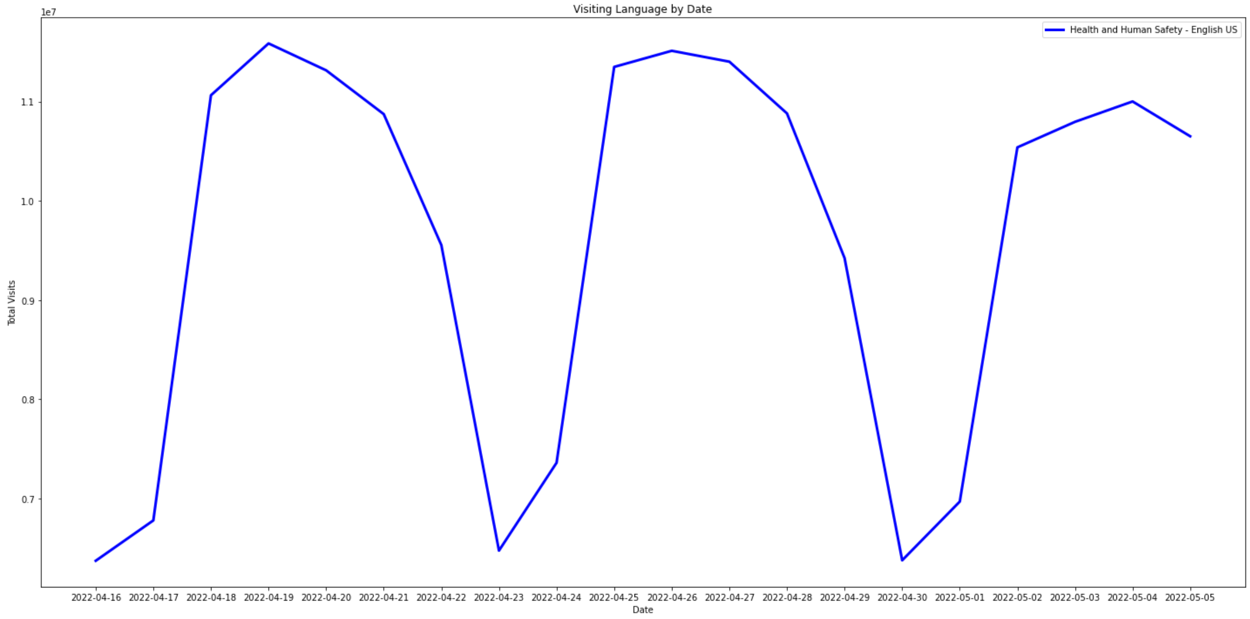
Chinese:



English GB



English US


Visiting Language by Date
1e7
Health and Human Safety - English US
Total Visits
1.1
1.0
0.9
0.8
0.7
2022-04-16
2022-04-17
2022-04-18
2022-04-19
2022-04-20
2022-04-21
2022-04-22
2022-04-23
2022-04-24
2022-04-25
2022-04-26
2022-04-27
2022-04-28
2022-04-29
2022-04-30
2022-05-01
2022-05-02
2022-05-03
2022-05-04
2022-05-05
Date