Amy Fann
Renjun Zhu
Info W18

U.S. Government Websites Analytics

Background

For this project, we were interested in learning more about how people interact with government websites. To do this, we pulled several data sets from a U.S. Government analytics website (link:https://analytics.usa.gov/data/). While the data provided by this website does not cover every single governmental website, it is frequently updated and is currently tracking over 4500 websites.

Our primary focus is to see what people care about when visiting U.S. government websites, how they access that information, and who those people are. We hope that through this project, we can gain a deeper understanding regarding the relationship between various people and their usage of government websites.

Method

We splitted our investigation into three sub-categories and answered them individually :

1. Which websites and/or topics are paid most attention to?
2. How are the websites being accessed?
3. What are the visitor demographics?

For in-depth analysis supplemented with graphs, please see the following pages.

Conclusion

From our data sets, we concluded and summarized our insights into a few bullet points:

1. People from other countries also visit U.S. government websites for variety reasons, though it is most likely for either citizen-related issues (such as getting a passport, immigration, job searching, taxes) or to access search engines.
2. Someone who is using a Windows Operating System on a computer is more likely to access government websites.
3. While most people who access governmental websites reside in the USA, other countries over the world also access these sites as well--with the most prominent being countries who have a strong relationship with the USA.

Reflection

Through this project, we were challenged to think critically about each data set, what it contains and what we could potentially learn from each one of them. Then, based on the questions we asked, we also had to learn how to collect and gather the proper data that we wanted. For example, there was one data (visitors online for all websites) set that only offered

data in five minute segments with no accessible archive, which limited how it could be used. We wanted to see the difference in a 24 hour period, so we learned to code a system that would aggregate information for a set period of time instead. Another challenging aspect was creating good graphs to display the conclusions reached from our data sets. One example was the time series detail plots of the browser data. We learned to be creative in how we decided to show the data.

Overall, from this project, we not only learned a lot about how to more effectively use pandas library in diverse settings, but also learned how to consider all aspects of the data and fully utilize it to understand more about the world around us.

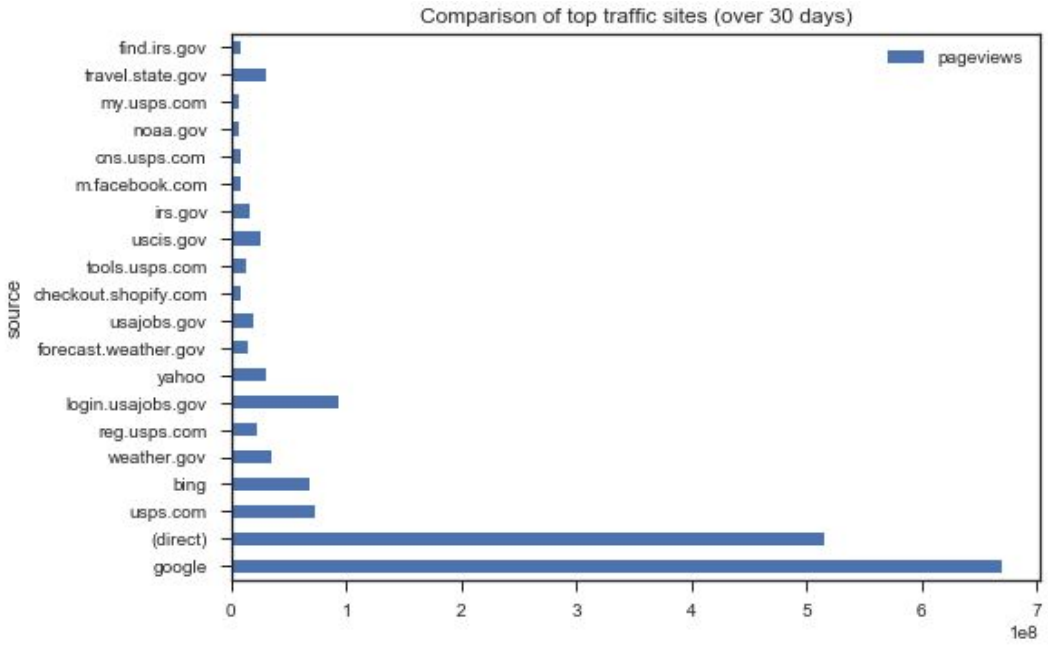
---

(1) Sub Question: Which websites and/or topics are paid most attention to?

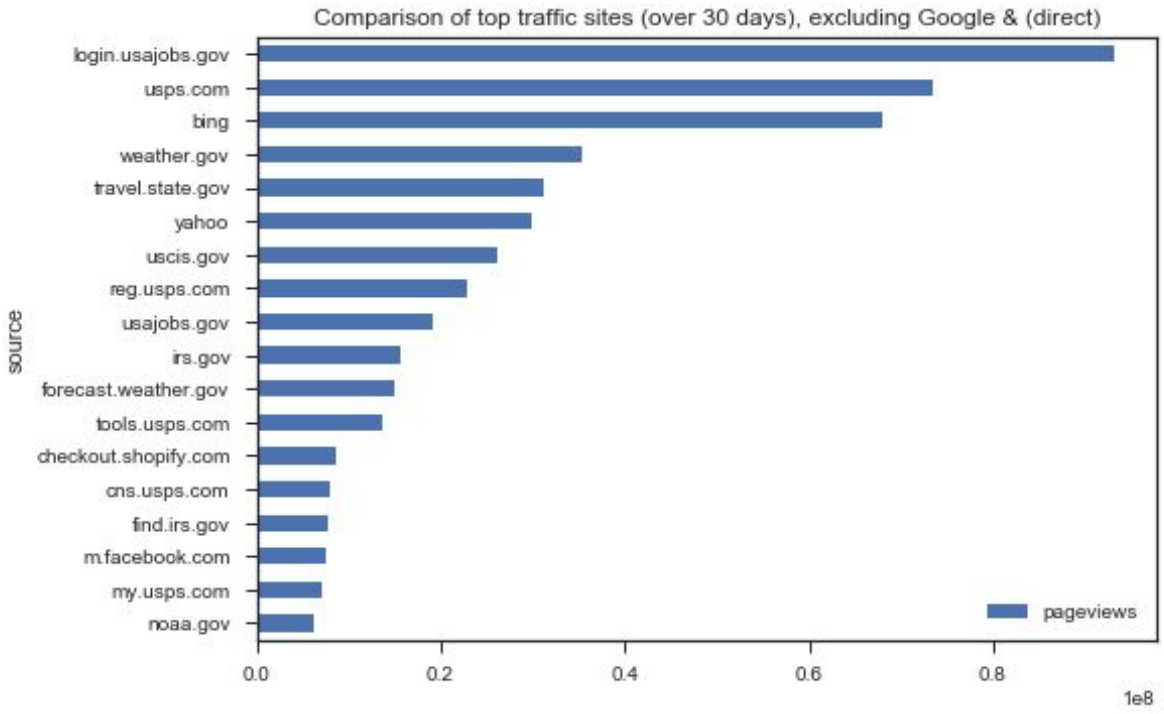
Traffic Sources Data Sets:

By tracking top traffic sources in the past 30 days and some of the top downloads in a single day, we wanted to analyze what type of websites are most visited and utilized.

First, we graphed the top 20 websites on June 25th, 2017, that had the highest total number of cumulative visits. We quickly see that “google” and “(direct)” are 5-6 times more viewed than the rest, which quickly skews the graph. This generally implies that people have two ways of getting to specific content on government websites, either through Google or a direct connection.



There may be several avenues that fall under the category “(direct)”, such as links directly from government PDF forms, legal referral websites, etc. However, if we exclude these two extreme cases, we can better see the relationship between all the other referral sources.



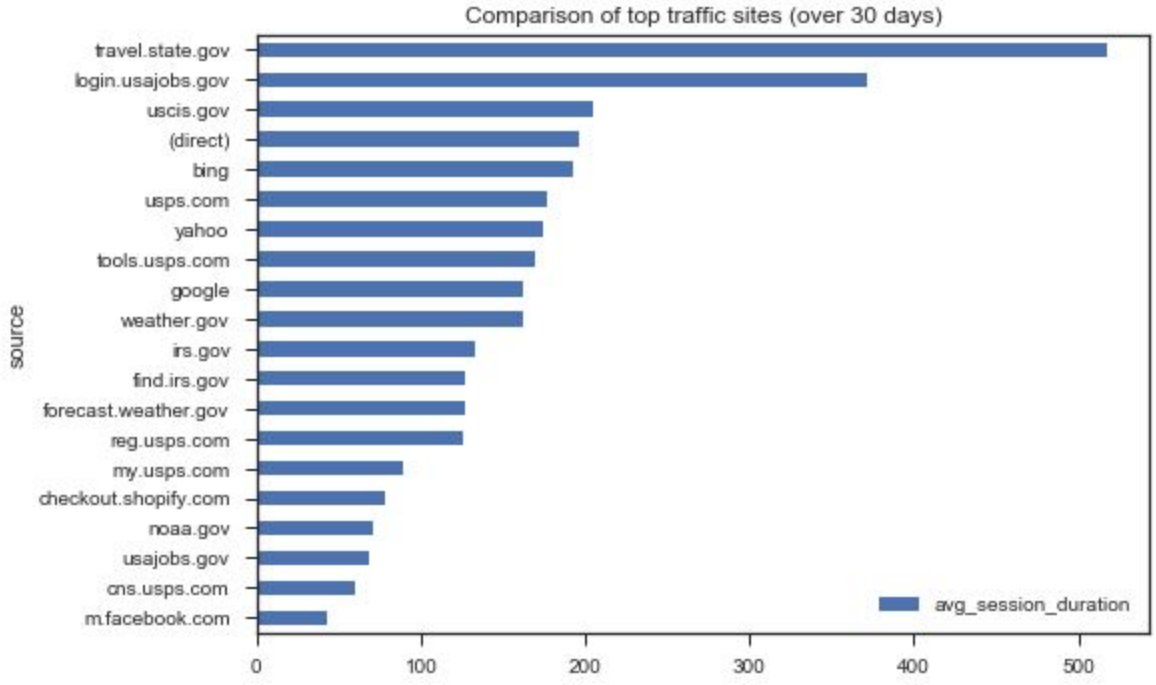
From an ordered graph of the remaining bars, we see that the number one referral for traffic is the portal to look for jobs. This shows that people are most interested in finding employment within the U.S. Government. It is surprising that the second largest traffic source is the website of the U.S. Postal Service, which show how widespread it is still utilized.

Without Google, it is interesting that search engines are not the majority of traffic referral services for Government websites. All of the sites make sense for accessing government websites for daily needs (i.e. immigration, taxes, etc.), except Shopify.com and NOAA.

We found Shopify particularly unexpected because, while we knew that potentially millions of transactions are being done each day, we expected online shopping to be more popular with websites such as Amazon or eBay. Prior to analyzing this data, we were not aware of the existence of Shopify.

On the other hand, NOAA (National Oceanic and Atmospheric Administration) was also equally surprising. On NOAA’s website, information regarding the environment is readily available in the format of individual articles. One reason we suspected for the rise of NOAA’s pageviews may be tied to USA’s recent departure from the Paris Accord as well as an increasing attention focused on environmental issues.

We also decided to look at the average session durations per website and compared the results with the previous pageviews graphs. We did not take out any websites as the distribution was more even across the board.



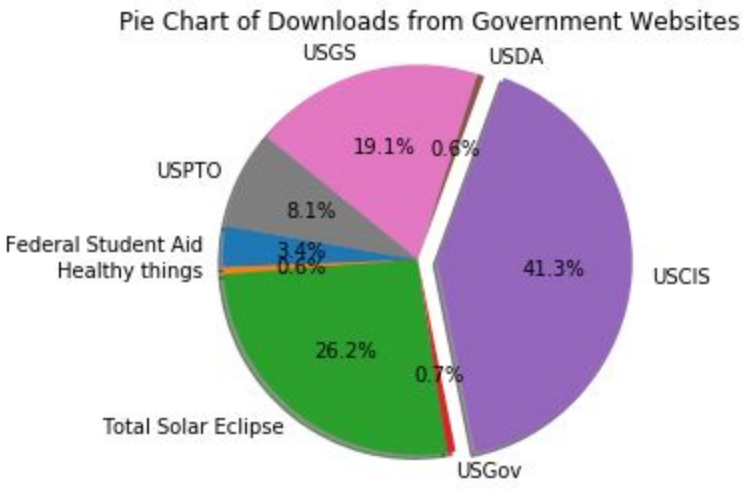
Through this, we see that although search engines had the most views, the websites that people spend the most time on are citizen-related tasks, such as passports or jobs. One thing we noted was how Facebook had the lowest average session duration. We speculate that since Facebook is a social media platform, many people spend a lot of time on social media in total but this does not necessarily mean they would spend a lot of time in one sitting.

Another interesting thing to note is that the USA Jobs portal for applicants (login.usajobs.gov) not only has one of the highest page views, but is also one of where people spend the most time per session. This is contrasted with the landing page “usajobs.gov”, which has a high amount of pageviews, but has a relatively shorter average time spent per visit. We think that because this website can act as a search engine or a brief entry point from which to log in to the portal for applicants. This difference makes sense, since job applications are time consuming, but looking at a static webpage is not.

Download Dataset:

To dig deeper into our exploration of this, we employed another set of data and analyzed the downloads from government websites on a given day, to see which sources are the most popular.

Here, we categorized each downloaded website by category, all organized in a pie chart, as shown below. This data was taken on July 30th, representing data from July 29th, from the same data sources.



As we looked at this pie chart, USCIS (United States Citizenship and Immigration Services) has been consistent in both data sets. This shows that USCIS is not only frequently visited, but that its website is also useful in offering resources from which people can download. Conversely, we would expect that there would not be a large portion attributed to something like the IRS, since this is not a typical tax filing season and therefore people would not need to download related forms.

We imagine that this data is potentially not generalizable, because of the "total solar eclipse" portion. This is an event that will take place in America, known to be a very rare sighting, which makes sense that many of the downloads would be surrounding this hot topic this year.

Setting that aside, we do believe that we can gleam some other observations from this graph, making it hard to verify or not whether this is because the data is skewed for one day or of other confounding factors. For example, we noticed that USGS (United States Geological Survey) and USPTO (United States Patent and Trademark Office) were very popular download sites and were not on the top traffic sites list.

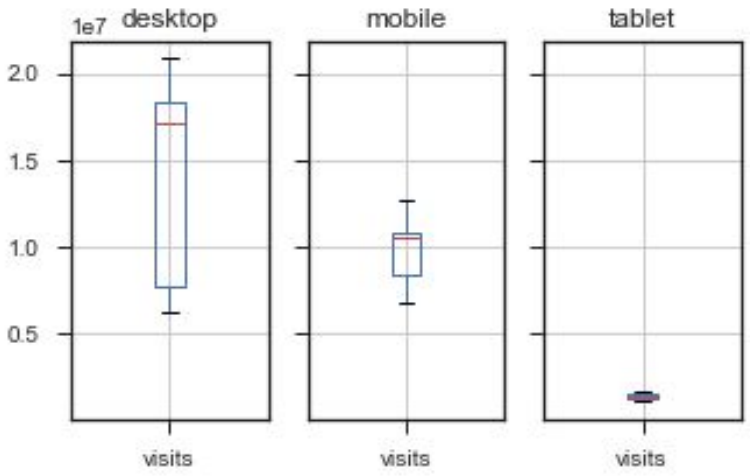
---

(2) Sub Question: How are the websites being accessed?

Devices DataSet:

The purpose of this sub question is to surmise the potential reason of why someone would visit a website. This is based on the speculation that often those who visit a website via a computer would differ in purpose from those who visit via a mobile phone. Thus, by taking a look at how websites are being accessed, we can either confirm or reject this idea.

We first compared and contrasted the total amount of visits across three devices: computers (or here, desktop), mobile cell phones, and tablets. Before we created a plot, we hypothesized that most of the visits would come from a computer.



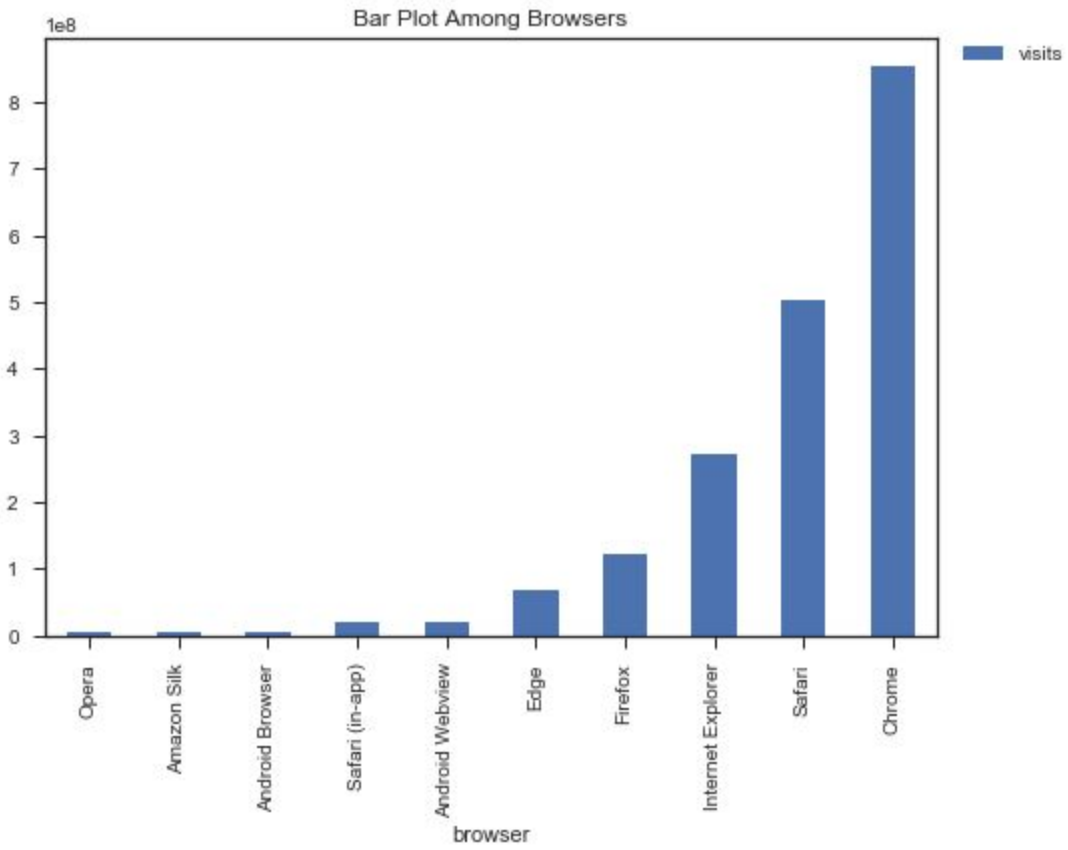
We collected data covering 90 days from the same web source, grouped by device type, and generated a box plot for comparison. This graph confirms our hypothesis. Computer access was most prominent because of its wide range and larger mean. Mobile phone access has a much smaller mean, around 10 million visits. However, the range of tablet is very concentrated below 2 million visits. The range and mean statistics informed us, people generally like to use desktop to access government websites. In order to understand the potential reasoning of high volumes of users using desktop, we summarized a few possibilities:

1. For work purposes, in the case of US government staff or research staff
2. To download a document
3. To browse frequently or for a long duration
4. Convenience (already using a computer and/or for speed)
5. Printing purposes

We hypothesize that the use of mobile devices is primarily due to immediate interest or convenience. As for the low usage of tablets for access, one potential factor may be because tablets are less commonly owned. Furthermore, we guessed that tablets, due to the bulkiness of structure, will be used mostly in stationary settings. This could either be in a work setting (probably less likely) or a home setting. In this case, when people access governmental websites from a tablet, it is probably more out of interest rather than necessity.

Browser-OS DataSet:

To verify consistency of the data regarding of how people access government websites, we analyzed the browser usage dataset.

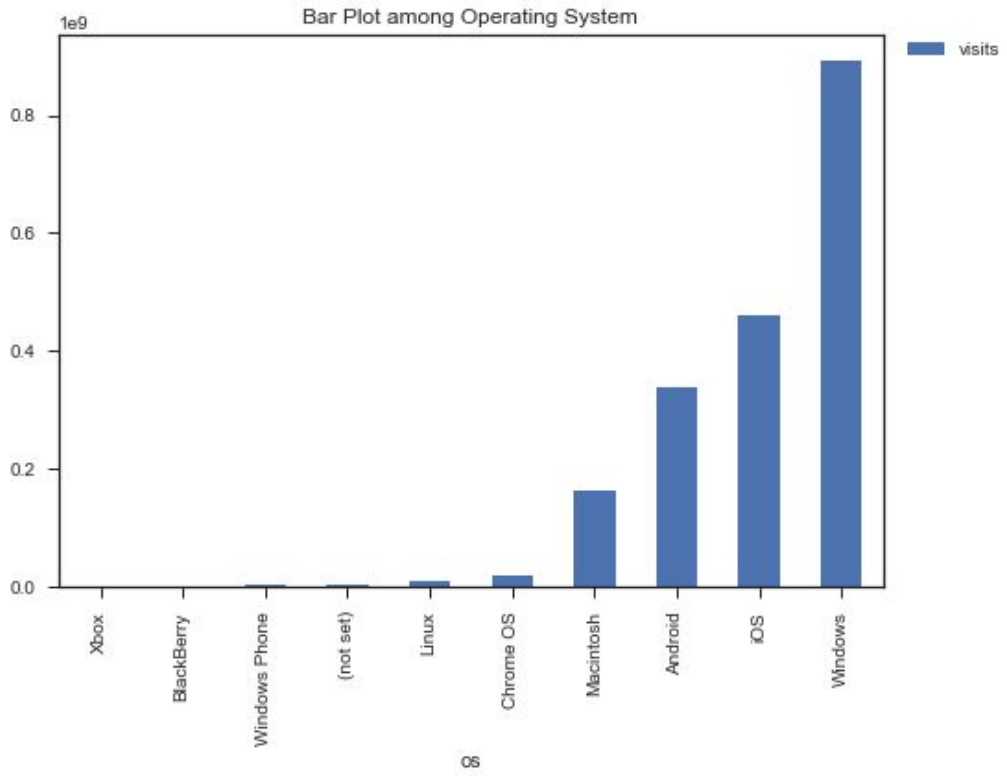


From the bar graph above, we see Chrome is the most often used browser. We can connect this information to the previous analysis in which Google ranked first as the traffic source to government websites. We know that Google produces Chrome and has its own search engine as the default for that browser. Likewise, we also know that all Apple products come with Safari as a default browser, which can skew the data results.

Another reason why Chrome and Safari might be so popular is because they are not only available on desktop, but they are also available in mobile and tablet versions as well. On the other hand, other types of browsers are much less common.

However, we still see Internet Explorer and Firefox as the next two most commonly used browsers, which are nearly exclusively used for desktops, which correlates with our previous findings regarding how desktop users are much more widespread.

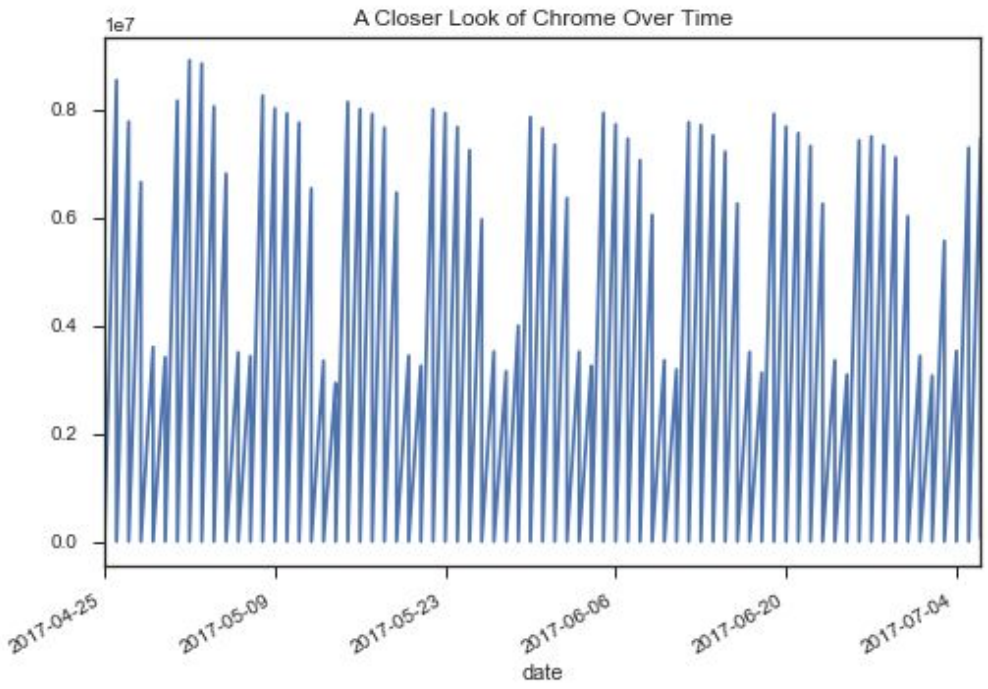
We can further verify this by taking a closer look at the operating system:



Here we see something interesting. While Windows, which is the desktop version of Microsoft, is the most popular operating system, the second most operating system is not Macintosh, but iOS. In fact, Macintosh comes in fourth place. This shows that among those who access governmental websites, most of the users use Windows. We suspect that the reason is because it is more likely for government staffs and researchers to use Windows due its availability and affordability.

On the other hand, we think that iOS had more recorded usage because of the widespread of apple products. Most mobile users have iPhone products and many commonly seen tablets today are iPads, iPad minis as well.

For a microanalysis of a specific browser, Chrome, we produced the periodic usage as a graph below.



Due to the regular peaks, we can guess that probably most of the people accessing the websites are doing so during a normal work day (i.e. most peaks are made of a group of 4 or 5 smaller peaks which probably indicate the days of a work week). Since the peaks occur during the same frame of a week, they are also likely to come from the similar time zones. Later we will discuss the demographics of the regional visits.

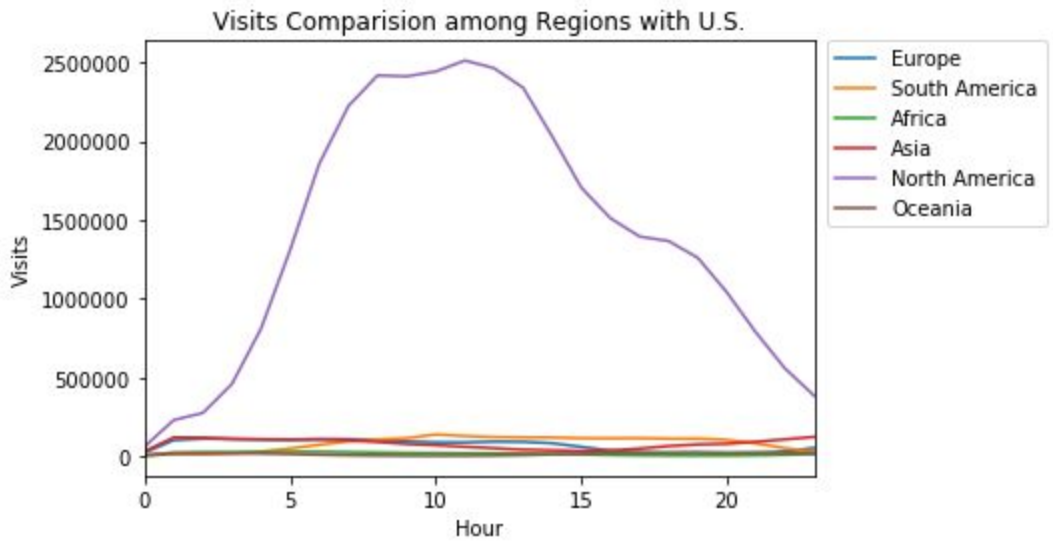
---

(3) Sub Question: What are the visitor demographics?

We were interested in learning whether or not most of the people who access these websites were government staff. In other words, does the average citizen or person from another country have interest in the US government?

To take a closer look, we took data of active visitors in 5 minute intervals over the span of 24 hours in one day. Then we divided the countries by region (using the UN guidelines for regions as represented in http://www.iucnredlist.org/technical-documents/data-organization/countries-by-regions). We plotted the data as a line graph, and compared which region had the most active visitors for a given time of day.

We hypothesized that North America, which contains the U.S.A., will have the most visitors during the daytime across all U.S. timezones.



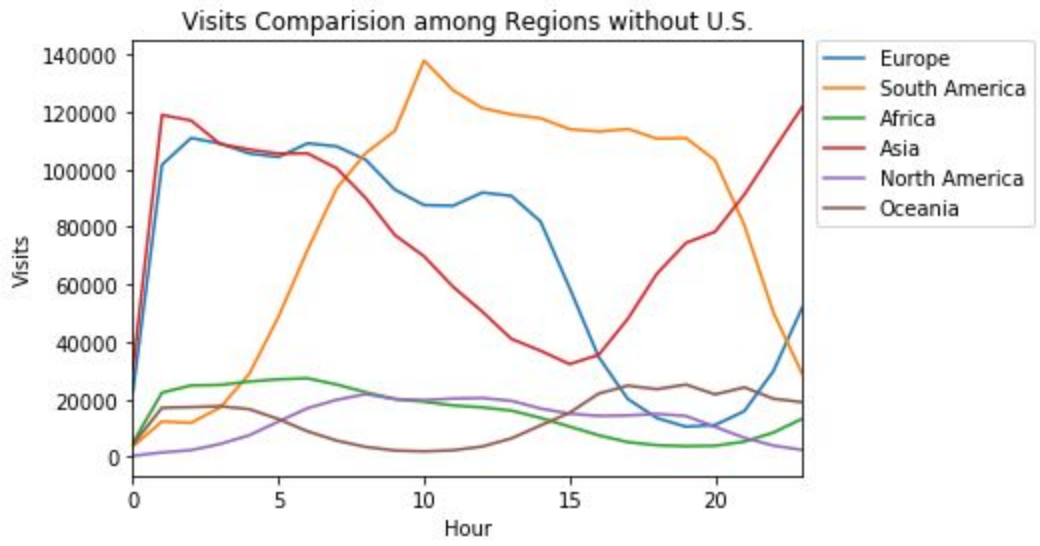
Not surprisingly, we see that the North America had the most visitors among all the regions. The ratio is quite overwhelming.

We suspected that a large amount of visits are done for one of the following reasons:

1. Governmental staff need access to certain data or reports, and thus constitute the majority of visits to these websites

2. People in the U.S.A. mostly visit these websites because they are most relevant to those living the U.S.

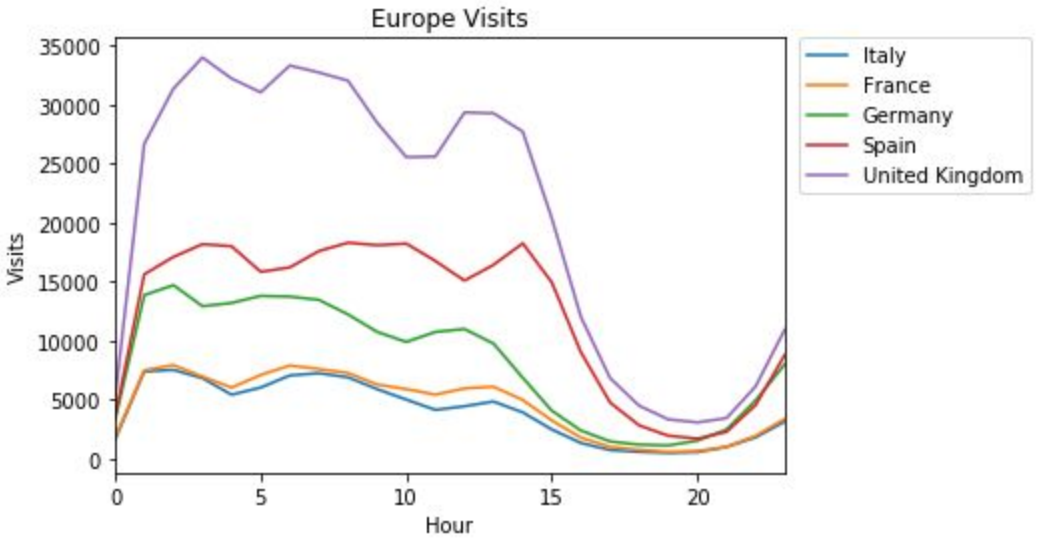
In order to get a better gauge of how much other countries access these websites, we decided to remove the U.S.A. from the North American region and compare the relative levels of other countries. Due to varying time zones, we hypothesized that each region would peak around noon or afternoon of their own respective time zones. Data collection was done in the Pacific Daylight Savings (PDS) time zone. This is done so that even though each time zone within each region will not be the same, there will be less variation within a region than between regions for comparison.



Due to the shifted time zones, we can see that the peaks from each region are different. However, from the data we can clearly see that certain peaks are higher than others. In the graph, we see that the top regions that access these websites include Europe, South America, and Asia.

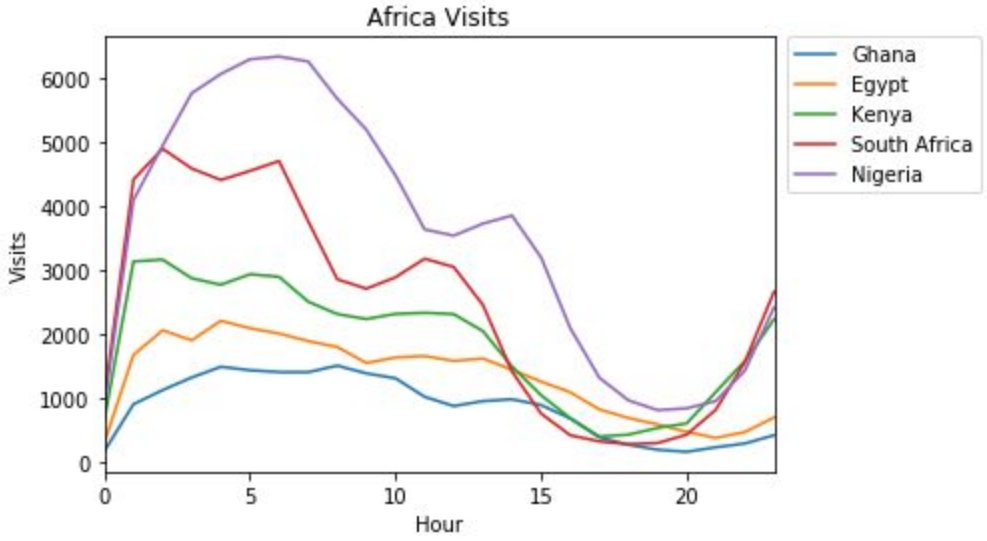
For each of the top three regions, we performed a further analysis and present a line graph for each of the top 5 countries that visited these websites to see if we could see any patterns or trends.

Europe:



We assumed that since U.K. was also an English-speaking country, it would make the most sense to have the European region represent one of the top regions. In the above graph, the detailed analysis shows that, indeed, it was the U.K. which heavily influenced the traffic coming to U.S. government sites from Europe. We wonder, if U.S. governmental websites were featured in more languages, would this impact how much other countries would visit the websites.

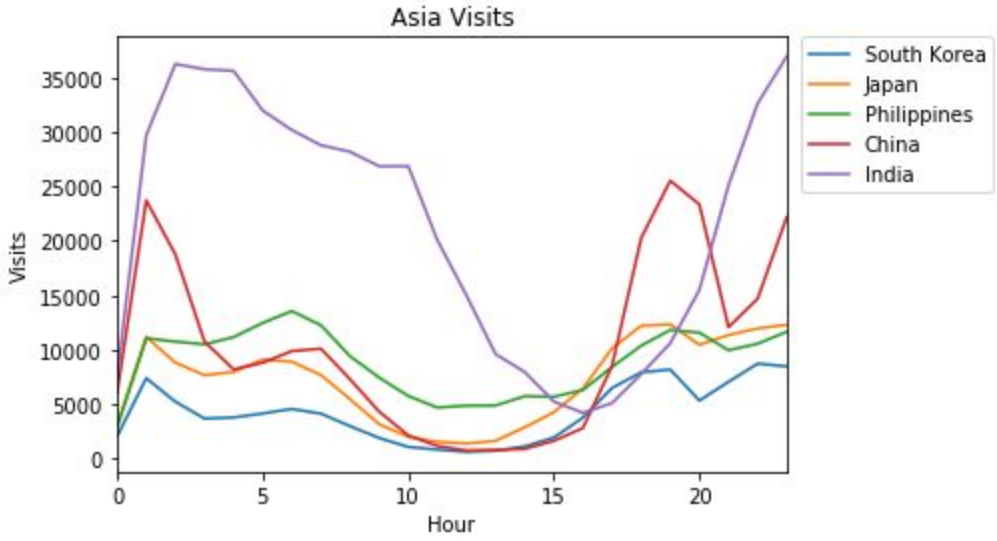
Africa:



Although the total proportion of visits from Africa is very small compared to other regions, the highest tick mark for visits from any country in the region is around 6000. Note that this, compared with the Europe graph, the lowest country in the Europe graph is still higher than

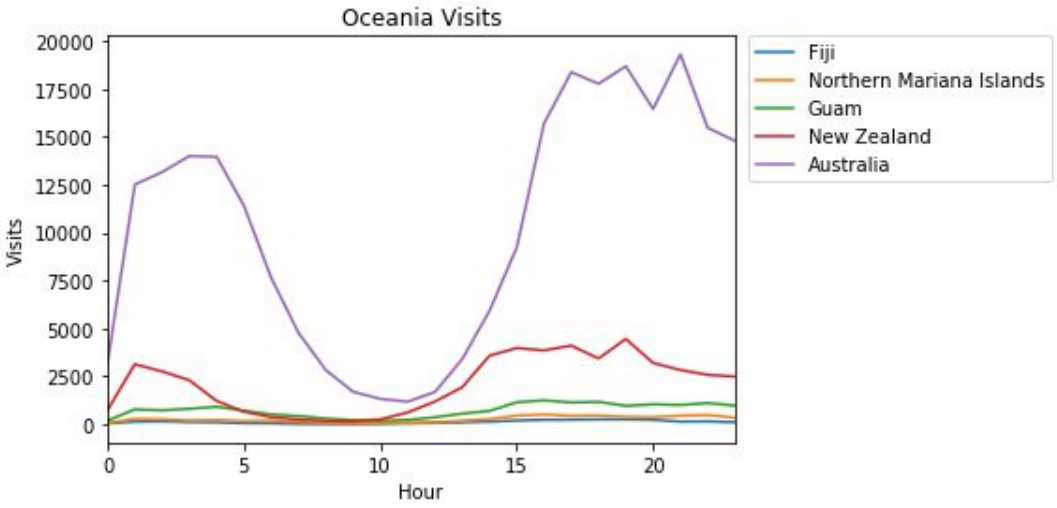
the highest country in Africa. However, the gap between each country is relatively evenly spread out with a difference in their peaks averaging around 1000 visits.

Asia:



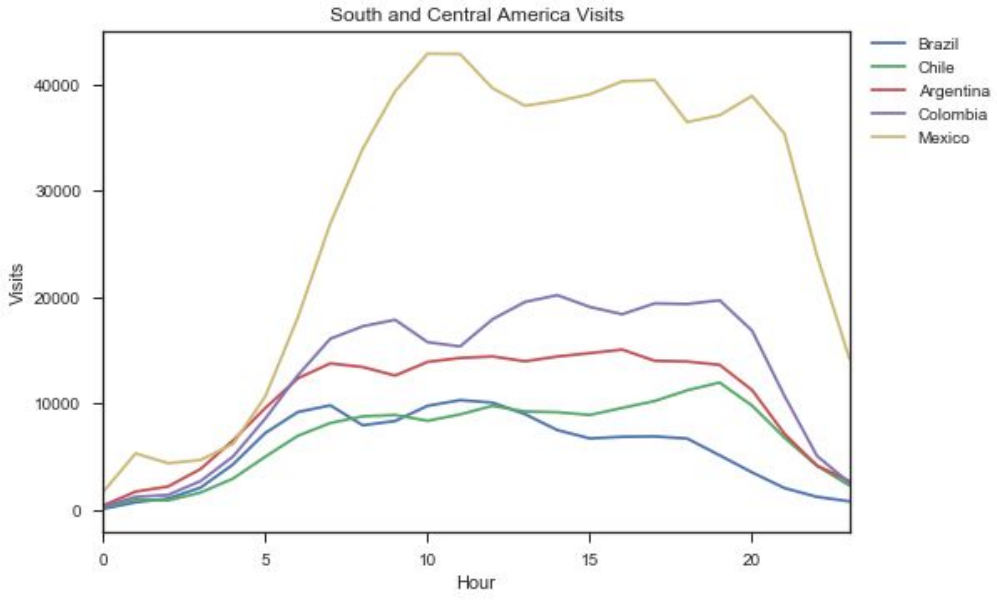
We can see that India is dominant in Asian visits, even though China has a slightly larger population. The difference might be related to several factors. For instance, policies regarding viewing U.S. government websites is very different between India and China; China holds a much more restrictive policy on viewing other governments' websites than India. Additionally, according to the Migration Policy Institute (http://www.migrationpolicy.org/programs/data-hub/us-immigration-trends) there has been a larger immigrant population in the U.S. from India than from China since 2000. One would assume that with a larger number of immigrants, one would expect a larger population interested in immigration.

Oceania:



Among all the oceania countries and districts, Australia has the highest population, so we are not surprised by the huge difference within this region. In addition, Australia is also an English-speaking country, which excludes the idea of any language barriers. However, we notice that visits in this regions are relatively smaller than other regions. The maximum vertical tick mark in this graph shows half the height of the South and Central America plot.

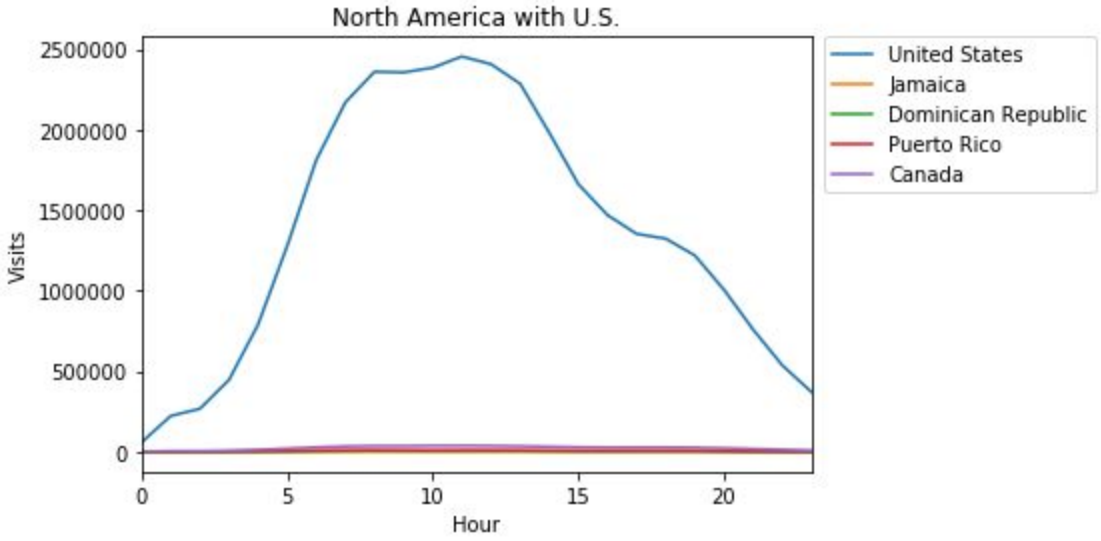
South America:

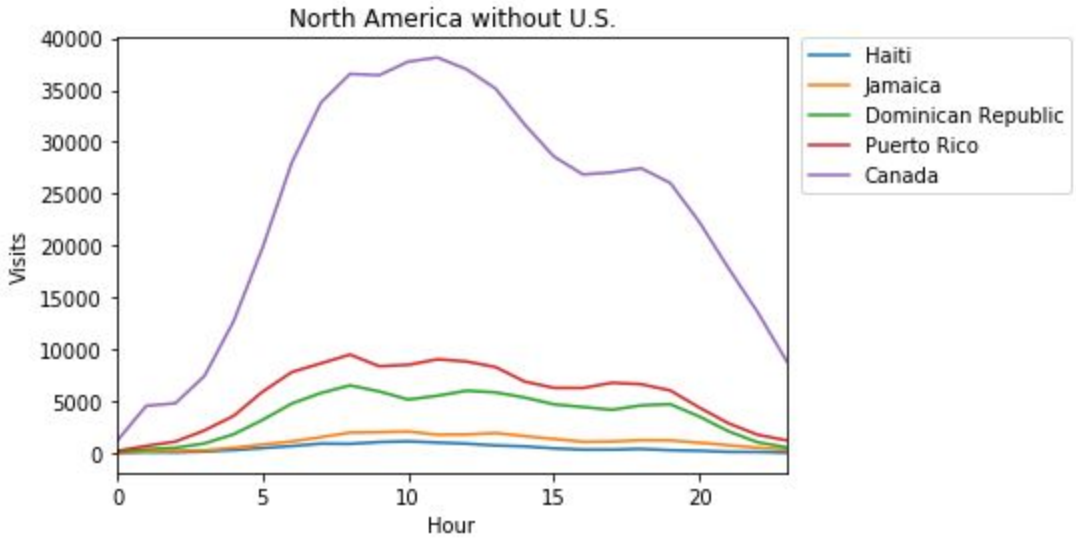


South America has been a large source of immigrants into the U.S. for the last few decades. Even though Mexico is a North American country, its pattern from the line graph is more consistent with the peaks from the South and Central American countries. From this graph, we think a main reason why Mexico has the largest rate of interest in U.S. government websites is because of its close proximity to the U.S. As the U.S. and Mexico are large trading partners, one would expect that there is a large amount of interest in things like trade policies and tax information.

North America:

Finally, we compared North America both with and without the U.S.A.





The graph with U.S.A will dominate the whole region, and makes the visits from other countries insignificant. We decided to remove U.S.A, and compare the nontrivial other countries. Here, we can see populational-wise that Canada leads the regions with a huge gap over other districts. This is expected, as in the South and Central America plot, because Canada is the second largest trading partner with the U.S. The frequency of transit between Canada and the U.S. is quite high, even though there is not a large proportion of immigration from Canada to the U.S.

Among Citites:

On a very microscopic level, we wanted to examine if some major cities contributed a high number of visits to the country and to the regions we analyzed. The ranked pie graph below describes the visits among all the cities in the world. We can see that most of the cities are from U.S.A. New York is ranked first, probably because of its leading economic position, Washington D.C.'s second ranking must relate to its traditional position as a government territory. It is not surprising that other major cities in the U.S.A have similar proportion of the pie. We can see the pie become smaller and smaller until it becomes a stairway which fades away in importance.