Predicting Internet Traffic Across Websites Categories

The Data Cavaliers
Group Leader: John Hope
Adina Kugler, John Hope and Adam Kippenhan
DS 4002
October 12, 2022

Restate Hypothesis

1. E-commerce sites will peak each year in page visits around December, whereas sports sites will peak around major sports' championships and news sites around November.
2. Furthermore, our time-series forecasting model will correctly predict the number of page visits within a certain range at least 50% of the time.

Executive Summary

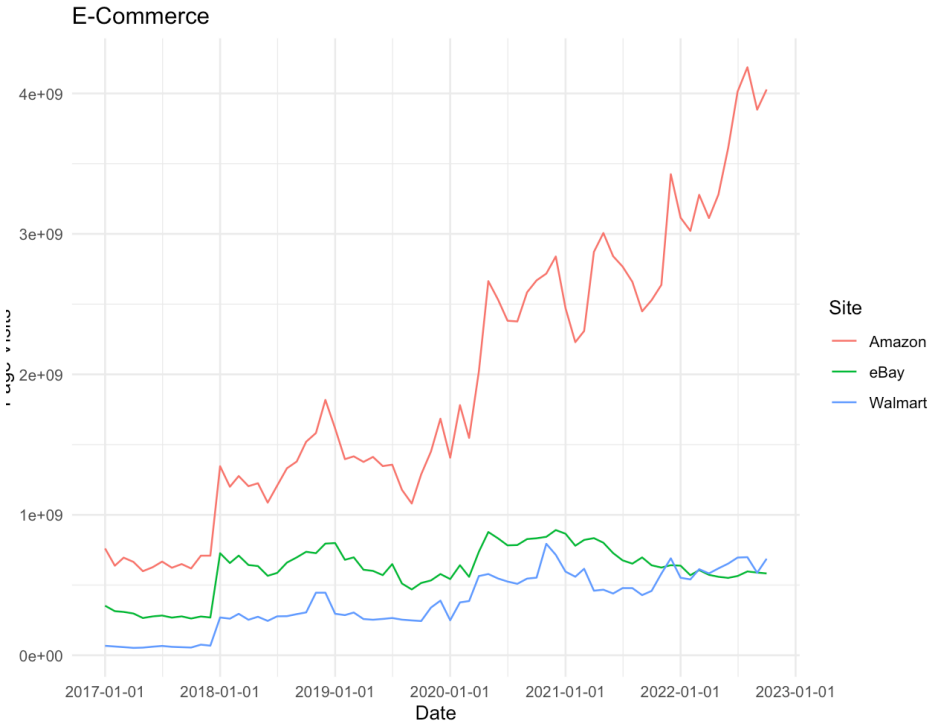
This document contains our data summary and plan for how we will analyze the data. We discuss the finding and establishment of our data with some summary statistics and visualizations and lay out a plan for how we will analyze that data using time series machine learning methods. This plan will dictate how we will move forward in the manipulation and analysis of our data and the basis on which we will build our conclusions and create additional visualizations.

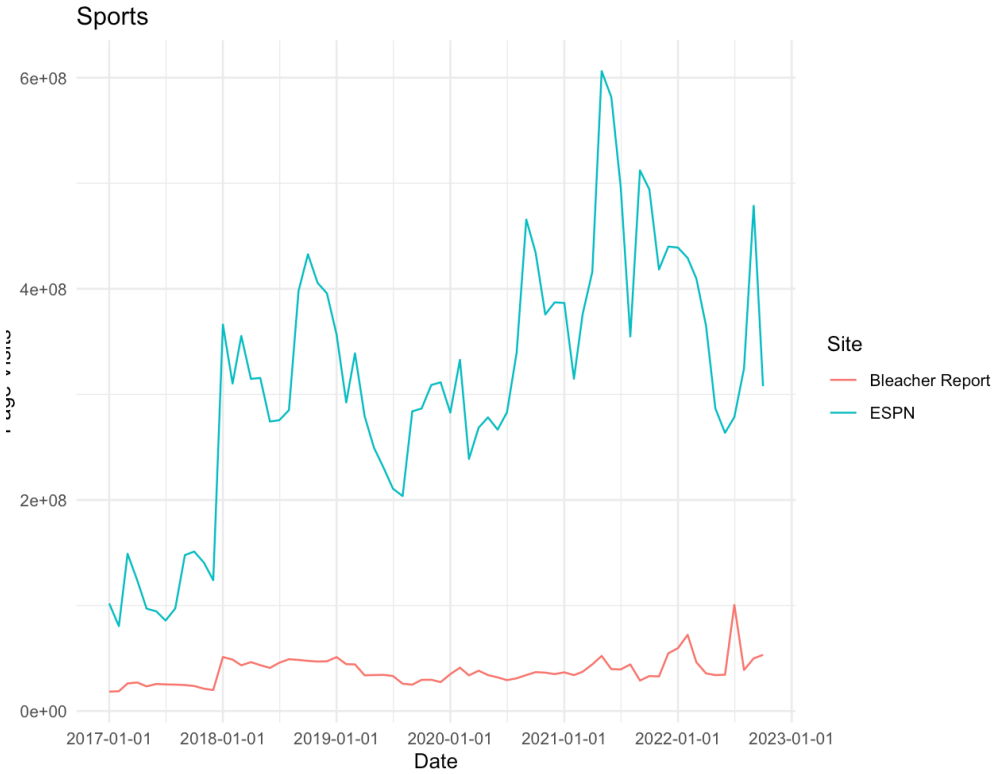
Data Discovery Findings

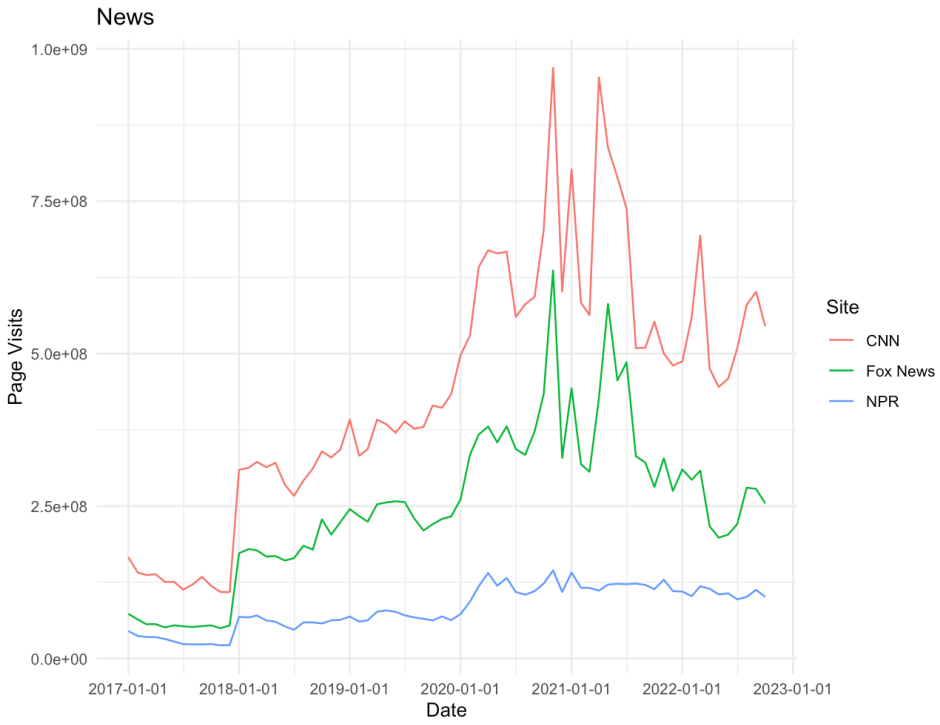
The dataset to be used for this project contains page visit data for different websites over time. The data was collected from Semrush, a website that provides website traffic information for different domains. Data for our different websites of interest were downloaded on Semrush, and then all compiled in Excel. The dataset contains page visits for each month from January 2017 to October 2022 for our 8 separate websites, so there are 560 total records. The dataset, in CSV format, and an accompanying data dictionary can be found in the following GitHub repository: https://github.com/john-hope/DS4002-Project2

With the dataset being established, we can begin the exploratory analysis phase of the project. Our exploratory data analysis consisted of generating different visuals to get an initial understanding of the overall trends of page visits across different sites. The first question we considered is what is the general trend of each category of sites over time? Answering this question gives us the general understanding of trends to keep in mind while

conducting our analysis. The three plots below are line plots showing the total page visits each month for each site, divided into the three categories of interest: e-commerce, sports and news.

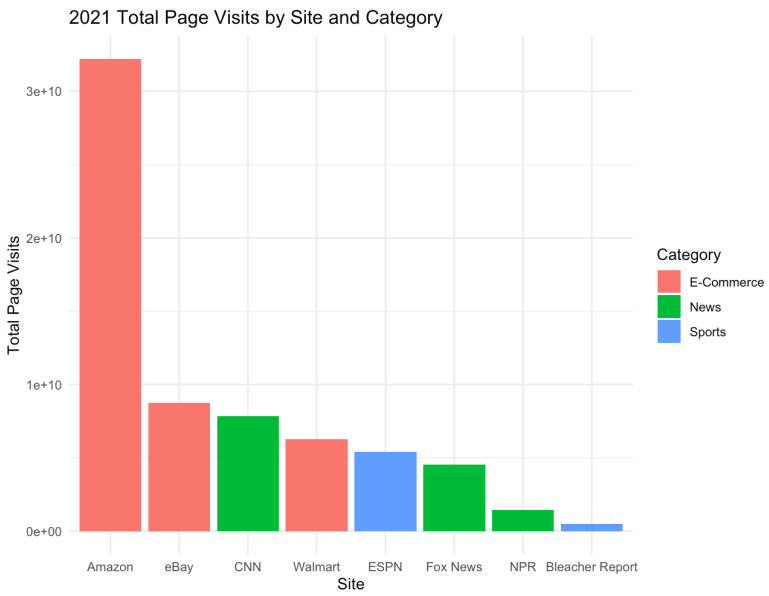






From the plots above we note that the general trend for most sites is that page visits are increasing over time. For the e-commerce sites, Amazon and Walmart appear to be increasing over time, with patterns of seasonality with spikes around the end of the year, while eBay has appeared to plateau and actually decrease in visits over time. The sports sites, especially ESPN, appear to be more volatile, with a lot more peaks throughout the year, but still maintain the increasing trend over time. For the news sites, we see a general increasing trend from 2017 to mid-2021, and decreases since. In addition, we see very severe spikes in CNN and Fox News around late 2020 and early 2021.

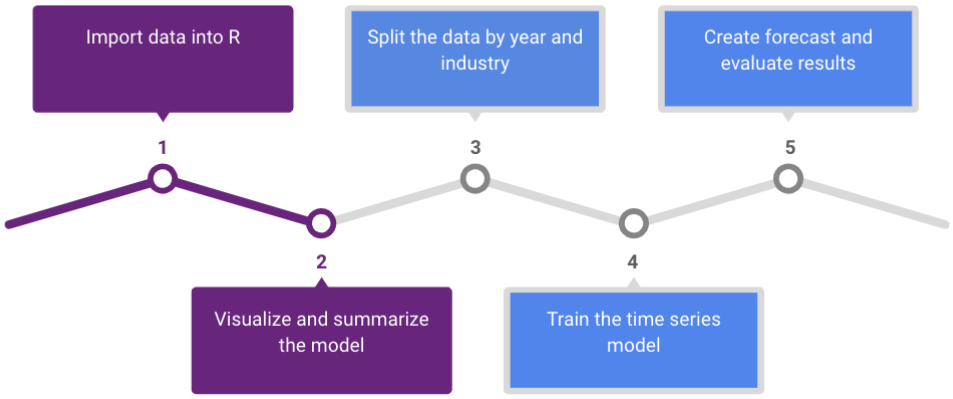
The next question we wanted to explore was “What sites were the most popular in 2021?” Since 2021 is our last full year of data, we are interested in seeing how popular each site and each category is relative to each other. The following bar plot displays the total number of page visits each of the websites experienced in the year 2021, and are color-coded by category.



From the bar plot, we see that of our categories, e-commerce sites tended to be the most popular. Amazon was overwhelmingly more popular than any of our other sites, with over 30 billion page views in 2021, where the next closest site, eBay only had just under 10 billion page views. There also seems to be a mix of popularity in news and sports sites, with CNN and ESPN dominating their categories, and Fox News, NPR, and Bleacher report falling behind.

With these gained insights in mind, we can move onto our analysis plan, which will outline our process of testing our hypothesis.

Analysis Plan



Our main goal in our analysis will be to compare our findings to our hypothesis to see if website traffic to specific sites actually correlates with our predicted peaks with e-commerce

sites in December, sports sites around major sports' championships and news sites around November. Additionally, we will create a machine learning model which we will use to attempt to predict traffic to specific websites based on the data we have collected.

Import Data into R

Our first step to begin the analysis will be to import our data into our analysis tool of choice: the data science-oriented programming language R. We chose to use R for our analysis because of our familiarity with the language as well as its ability to easily create visualizations that will effectively convey our findings. Using our data which is in a CSV file format, we will import the data into a dataframe in R and from there we can use various libraries found in R to create visualizations that coincide with our analysis.

Visualize and Summarize Data

To evaluate our initial hypothesis regarding the peak time of year for each industry, we need to utilize visualizations. Each industry will be plotted on a series of plots by month, per year. Differences in peak times of these charts will be analyzed for differences with peaks being evaluated further. These graphs will be overlaid to provide differences in peaks amongst industries to provide information about how the industries differ in both amounts of internet traffic and peak times. The findings will be confirmed with numeric summaries to further understand the scales of the peaks. The visual and numeric analysis will be done using ggplot2 and other visualization for time series line graphs in R.

Split the Data Appropriately

Once our data has been visualized and summarized, our next step will be to split it up to make our future analysis and machine learning model creation and training easier. R will make it easy to accomplish this and put the data into separate data frames. We will split the data into different years between the range present in the original data from 2017 and 2022, which we will use later to correspond to each stage of the machine learning model creation and training.

Create and Train our Machine Learning Model

With our data split into separate years, we can divide each stage of the machine learning model creation and training process by the different data sets for each year. Our time-series machine learning model will be created using R. Once we have created the initial model, we can train it using a selection of our separated data of time.

Create Forecast and Assess Results

The model created from the trained data will be tested against the remainder of the data to assess its accuracy and further improve the model. This process consists of making predictions, in the form of a forecast, of page visits for a certain month, and assessing if our prediction falls within a certain range of error from the observed result. The model will further be evaluated

through forecasting methods in R [2]. This includes using the test data to predict trends and evaluate the accuracy, receiver operating characteristic, area under the curve, error rate, sensitivity and specificity. This will allow for testing of the second portion of the hypothesis. The accuracy will be the primary metric of evaluating that the model forecasts the test data at 50% accuracy. Further visualization will be performed at this stage to evaluate the strength of the predictions for each industry.

References

[1] R. Jogi, "How to Handle Heavy Internet Traffic on Your Website?," Cloud Minister Technologies. Sept. 28, 2021. [Online]. Available: https://cloudminister.com. [Accessed Oct. 5, 2022].

[2] A. Coghlan, "Using R for Time Series Analysis," Little Book of R for Time Series. 2010. [Online]. Available: https://a-little-book-of-r-for-time-series.readthedocs.io. [Accessed Oct. 12, 2022].