Wikipedia User Behavior Analysis

Introduction

In today's digital era, Wikipedia stands as a crucial source of crowd-sourced information, significantly influencing global knowledge access and engagement. This report presents a focused analysis of Wikipedia's user interaction data, shedding light on trending search topics, daily traffic patterns, language diversity, and device preferences. Key dates such as New Year's Day and the US Election Day in 2016 are examined to decipher user interests, alongside broader trends revealing weekly usage patterns and device utilization.

Utilizing a comprehensive dataset, this analysis aims to uncover user behavior and preferences, providing insights critical for enhancing content strategy and user experience. The insights drawn are particularly valuable for Wikipedia's strategic planning, offering a guide for content development, platform design, and audience engagement.

By dissecting user interaction trends, this report seeks to empower Wikipedia's decision-makers with actionable data, aiding in the platform's continuous evolution to meet the diverse needs of its global user base.

Dataset description

The dataset contains the daily number of webpage visits for several Wikipedia webpages. If you are not familiar with Wikipedia, you can check it out here: https://en.wikipedia.org/wiki/Main_Page. The dataset includes daily page visit counts for 1,500 Wikipedia pages starting on 2016-01-01 until 2016-12-31.

The audience for this analysis are senior executives at Wikipedia who require my analysis to understand the trends in the data.

Dataset source: Kaggle.com

Project Questions:
Uncovering answers to the following questions will help senior executives understand the trends in the data:

1. What were some of the most trending search topics on Wikipedia on the following days?
    - New year's day
    - November 8, 2016

2. Which day(s) of the week is/are the most/least popular for visiting wikipedia?

3. How many languages are represented in this dataset? What proportion of the pages does each language represent?

4. Which device type is used more frequently for visiting wikipedia i.e. desktop or mobile devices?

Data Preparation and Processing
The following steps were taken to prepare and process the dataset for analysis:

1. The raw dataset was downloaded from Kaggle as a pivot table and was unflattened to make it easier for manipulation and analysis.
2. A function was defined with a Python script to transform the dataset from a wide to long format.

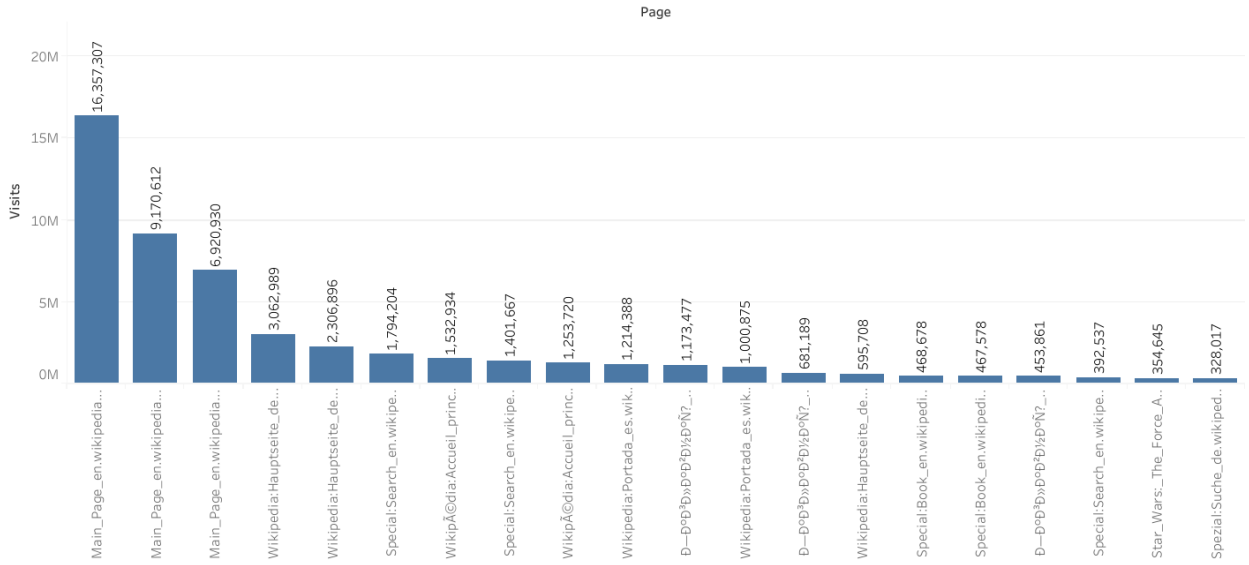
3. The Excel workbook was read into a Pandas data frame for cleaning before transforming it into a SQLite database for querying.
4. The language codes and devices were extracted from the strings in the page column using Pandas to be able to answer questions 5 and 6.

Analysis and Findings

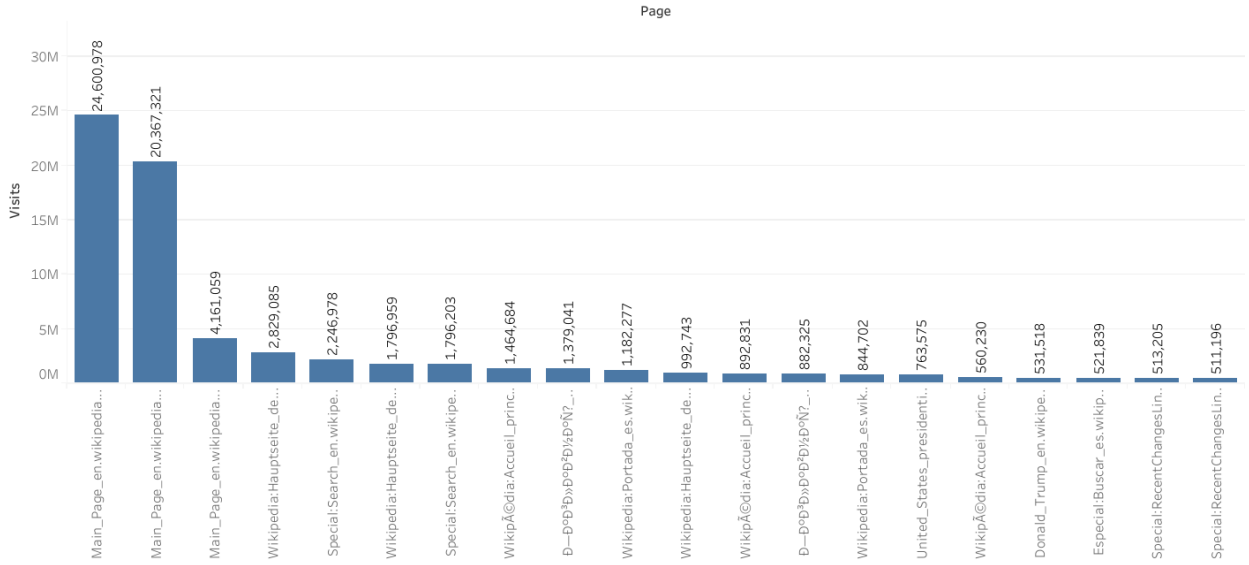
Most Trending Search Topics:

- Identifying trending topics on January 1st and November 8th helps us understand Wikipedia users' search behavior on New Year's Day and the US Election Day.
- This was done by filtering for the specific dates and analyzing page visit counts.
- **The analysis and visualizations revealed that on New Year's Day the highest page visits was the wikipedia main page however the newly released Star Wars movie "The Force Awakens" made it to the top 20 most visited pages. This showed clearly that most people were really interested in the trending new movie released December 18th, 2015.**
- **The Wikipedia page visits on November 8, 2016 showed a similar trend as the New Year's day with the most visits going to the Wikipedia Main Page. However, the United States Presidential Election and Donald Trump appeared in the top 20 most visited pages on that day. Considering that this was an election day, it shows the clear interest of the Wikipedia users in the 2016 presidential election and the candidate that eventually won.**

Most Trending Search Topic - New Year

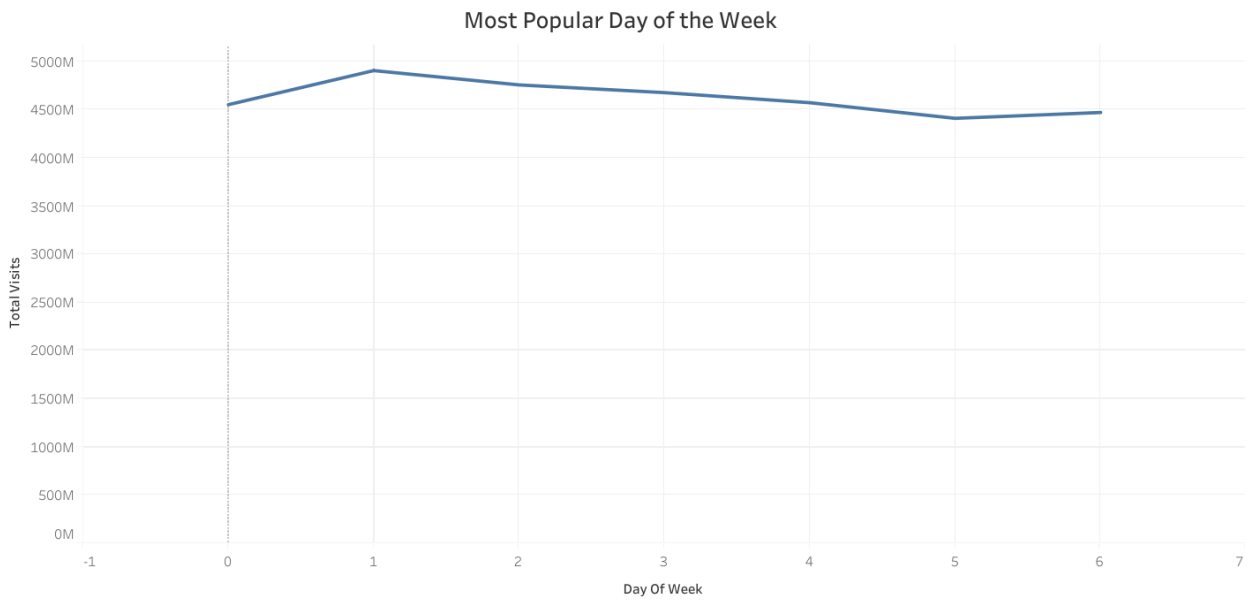


Most Trending Search Topic on Election Day



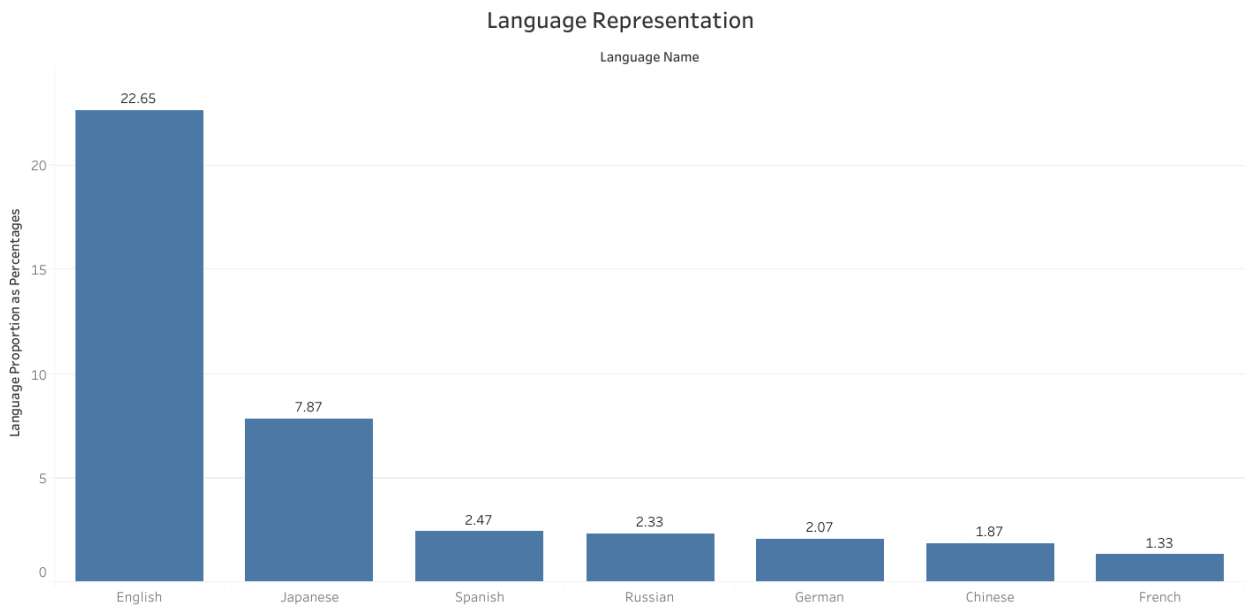
Popularity by Day of the Week:

- Determining which days attract more visits helps us understand the popularity of the site on different days of the week. For instance, higher visits on weekdays might indicate heavy usage by researchers or students, while weekends might suggest leisure reading.
- This was done by aggregating visits by day of the week using SQL and plotting a bar chart in Tableau.
- **The analysis and visualization showed that Mondays are the most popular for visiting Wikipedia with 4.9m visits while Fridays are the least popular with 4.4m visits.**



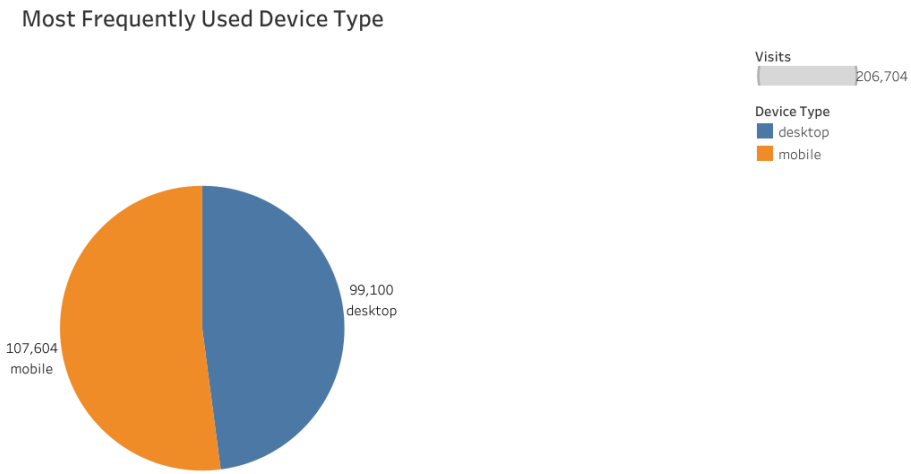
Language Representation:

- Understanding the linguistic diversity of the pages offers insight into the cultural diversity of Wikipedia users.
- This was done by extracting the language codes of the Wikipedia users and aggregating the visits by each code. The proportion of each language was calculated and a barchart was plotted showing the proportion for each language represented.
- **There are 7 languages represented in the dataset including Chinese, English, French, German, Japanese, Russian and Spanish.**
- **English had the highest proportion for languages represented in the dataset at 22.65%, followed by Japanese at 7.87% with French coming a distant 7th at 1.33%.**



Device Type Usage:

- Understanding preferred devices of users for accessing Wikipedia helps in optimizing the website's design and functionality. If most users visit via mobile devices, prioritizing a mobile-first design approach ensures a better user experience.
- **The device types were extracted from the page name string in the data set and aggregated by device.**
- **In 2016, Wikipedia had more visits from mobile devices at 107.6k visits than desktop devices at 99.1k visits.**



Discussion

Most Trending Search Topics:

My analysis on the most trending search topics on specific significant days—New Year's Day and the US Election Day—provides intriguing insights into user interests and external events influencing online behavior. The popularity of the Wikipedia main page indicates a general pattern of starting broad before delving into specific topics. The interest in "The Force Awakens" aligns with global entertainment trends, while the heightened focus on the US Presidential Election and Donald Trump on November 8, 2016, reflects Wikipedia's role as a key information source during major global events. These findings underscore Wikipedia’s relevance in disseminating timely information and its reflection of current societal interests.

Popularity by Day of the Week:

The analysis revealing Mondays as the peak of Wikipedia visits and Fridays as the least suggests a potential correlation between Wikipedia use and the typical work or academic week. This pattern might indicate a high reliance on Wikipedia for research and educational purposes during the beginning of the workweek, with a gradual decrease as the weekend approaches. This insight can guide content and resource allocation strategies to cater to these usage patterns effectively.

Language Representation:

The linguistic diversity in the dataset, with English and Japanese leading, points to a broader cultural and demographic reach of Wikipedia. The dominance of English could be attributed to the widespread use of the language globally, while the significant presence of Japanese indicates a strong user base in Japan. The relatively lower representation of other languages like French suggests an opportunity to further diversify Wikipedia's content to

cater to a more varied global audience. Enhancing content in underrepresented languages could improve accessibility and inclusivity, fostering a more diverse user base.

Device Type Usage:
The predominance of mobile device usage over desktop in accessing Wikipedia highlights the ongoing shift towards mobile internet browsing. This trend underscores the importance of optimizing Wikipedia for mobile devices, ensuring that the mobile user experience is seamless, responsive, and accessible. As mobile usage continues to grow, focusing on mobile-friendly content and design will be crucial for maintaining and expanding Wikipedia's user base.

Conclusion
The analysis of Wikipedia's dataset reveals valuable insights into user behavior, preferences, and the platform's global reach. The study highlights key areas where Wikipedia plays a significant role, such as providing timely information on global events and serving as a primary resource for research and education. The findings also indicate the importance of mobile optimization and the potential for expanding content in various languages to enhance user experience and accessibility.

Moving forward, it is recommended that Wikipedia continues to monitor these trends to adapt its strategies for content development, user experience design, and resource allocation. Embracing the shift towards mobile usage, prioritizing content diversification, and aligning resources with user traffic patterns will be essential in maintaining Wikipedia's position as a leading global information resource. Additionally, continued analysis of user behavior and preferences can provide ongoing insights to guide future enhancements and strategic decisions.