Browsing History Analysis & Storytelling

**Introduction**

This project delves into the analysis of website user browsing behavior using a dataset containing visitId, referringVisitId, eventTimeUTC, and URL information. The primary goal of this analysis is to understand how users navigate through the website, identify patterns in their browsing activity, and gain insights into their engagement levels.

**Project Workflow**

**1. Understanding the Task**

Since this is the first time dealing with these terms, the initial step involved gaining knowledge about the dataset columns and their significance.

**2. Data Cleaning**

* Verified missing values in each column and handled them accordingly.

**3. Data Preprocessing**

* Added a new column URL\_type to classify URLs as valid, invalid, or local file accesses.

**4. Extracting Base URL**

* Processed the URL column to extract the base domain for further categorization.

**5. Adding Category Information**

* Manually extracted category names for each base URL, which was time-consuming.

**6. Using Kaggle for Additional Python Code**

* Utilized Kaggle resources to enhance the data processing workflow.

**7. Loading Clean Data**

* Re-loaded the dataset after cleaning and preprocessing.

**8. Sorting eventTimeUTC**

* Ensured the eventTimeUTC column was sorted chronologically for accurate time-based analysis.

**9. Initial Insights**

* Identified key insights such as:
  + **Top transitions**
  + **Top visited domains**
  + **Top categories**
  + **Browsing activity over 24 hours**
  + **Browsing activity across the month**

**10. Ensuring Unique Identifiers**

* Confirmed that each visitId was unique to maintain data integrity.

**11. Calculating Time Differences**

* Computed time differences between visitId and referringVisitId to analyze user navigation behavior.

**12. Identifying Highest Navigation Paths**

* Tracked the most frequent navigation paths using visitId and referringVisitId.

**Dashboards & Visualizations**

**13. Browsing History Dashboard**

* **Overview Dashboard with Key Metrics:**
  + Total visits
  + URL count
  + Top participants and organizations (OrgId)
  + Top categories and transitions
  + Browsing start and end time

**14. Time Series Analysis**

* Analyzed browsing trends on a daily and monthly basis for each category.

**15. Category Analysis**

* Identified:
  + The most typed category
  + The most clicked category
  + Number of visits per category

**16. Data Visualizations**

* **Pie Charts:**
  + Transition count distribution
  + URL type distribution
* **Most Frequently Used Titles:**
  + Determined the most accessed page titles.

**Conclusion**

This project provided deep insights into user browsing behavior by leveraging time-series analysis, categorical classification, and navigation path tracking. The dashboards and visualizations enabled a comprehensive understanding of user engagement and website interaction patterns.