ANALYSIS REPORT

Nikhil Kumar Bharti

INTRODUCTION

- The given dataset consists of detailed records of websites visited, including timestamps, domains, and page interactions.
- This report covers behavioral patterns, tries to identify intent and interests as well as construct a data-driven story based on web activity.

BROWSING INFO

The images here shows some important information regarding the dataset:

- A total of 5104 entries
- 9 columns, of which 2 are int64 type
- There are 53 missing or null values in ‘title’ column. These can be dropped.
- Since the task is analysis, not predictive modelling, we have limited the number of pre-processing steps, to preserve data originality.

```
RangeIndex: 5104 entries, 0 to 5103
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   OrgId            5104 non-null   object
 1   ParticipantId    5104 non-null   object
 2   DeviceId         5104 non-null   object
 3   url              5104 non-null   object
 4   eventtimeutc     5104 non-null   object
 5   transition       5104 non-null   object
 6   title            5051 non-null   object
 7   visitId          5104 non-null   int64
 8   referringVisitId 5104 non-null   int64
 9   eventtime        5104 non-null   object
dtypes: int64(2), object(8)
memory usage: 398.9+ KB
```

```
OrgId              0
ParticipantId      0
DeviceId           0
url                0
eventtimeutc       0
transition         0
title             53
visitId            0
referringVisitId   0
eventtime          0
dtype: int64
```

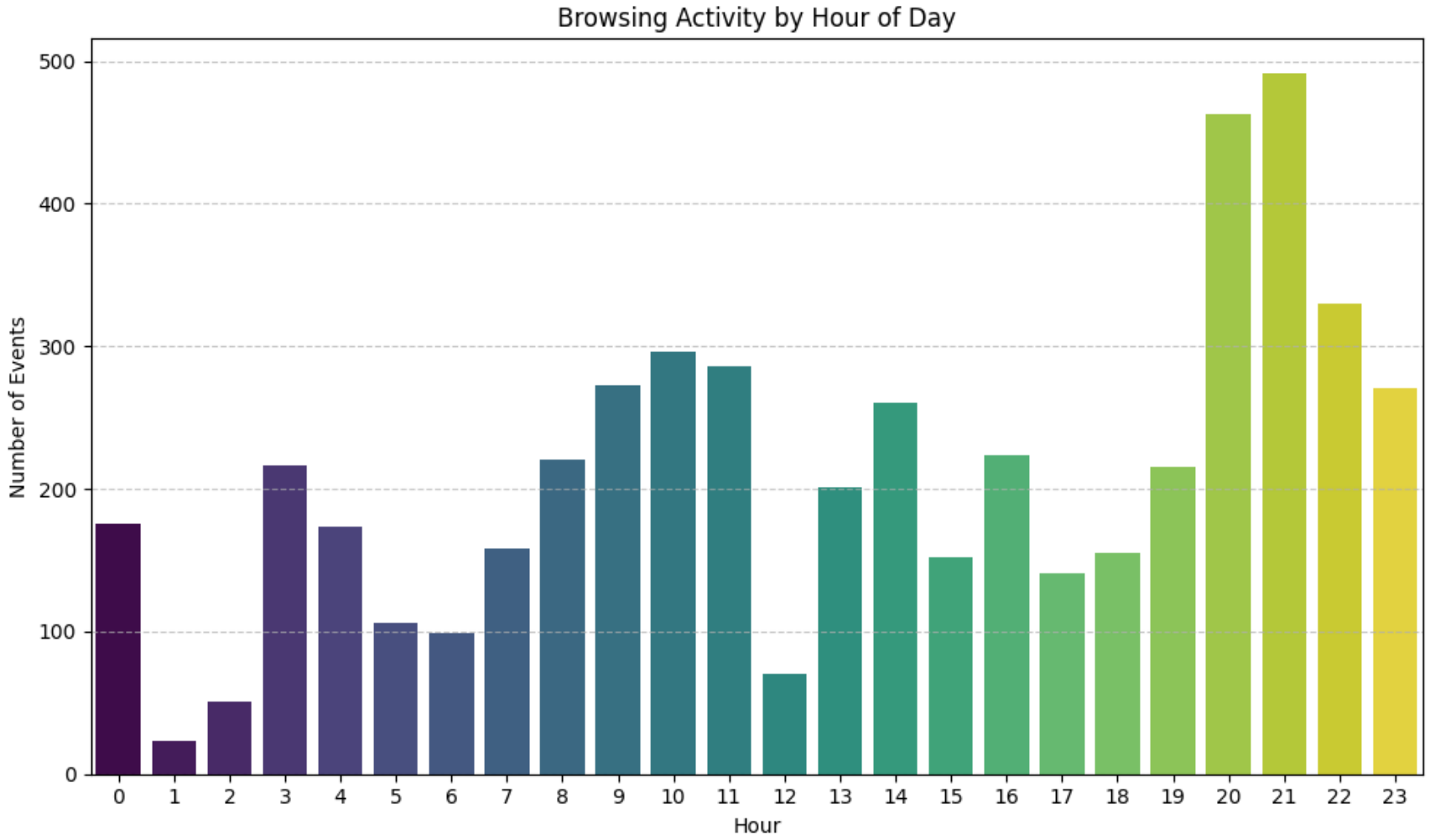
PRE-PROCESSING

- The only few pre-processing steps which have been performed here have been on the ‘eventtimeutc’ and ‘url’ columns of the original dataset.
- The ‘eventtimeutc’ has been converted into Python recognizable data format, and new columns each for ‘hour’, ‘day_of_week’ and ‘date’, have been created.
- A new column ‘domain’ was created store the domain names from the ‘url’ column.

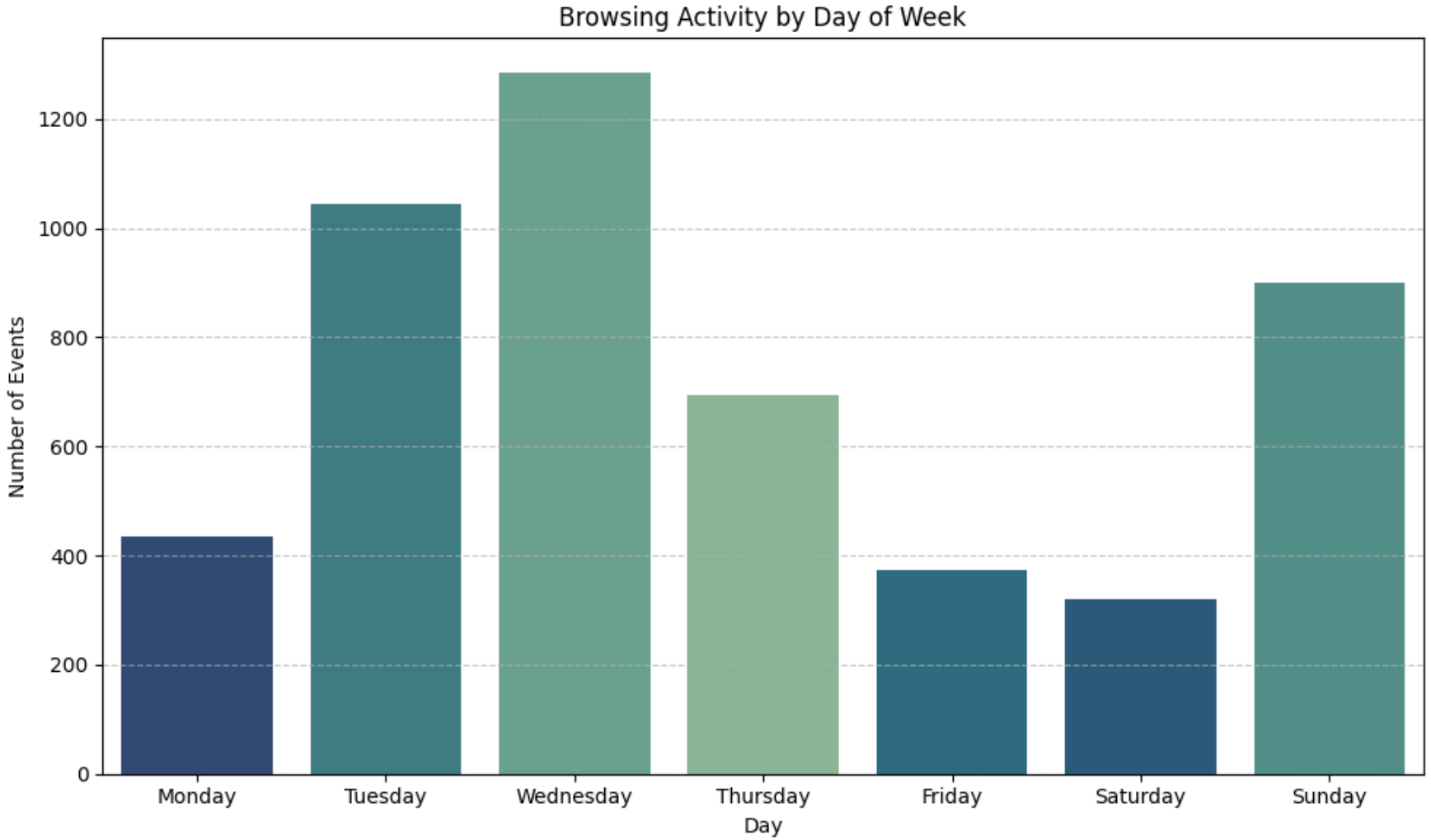
VISUALIZATIONS

The following slides contain some of the important visualizations.

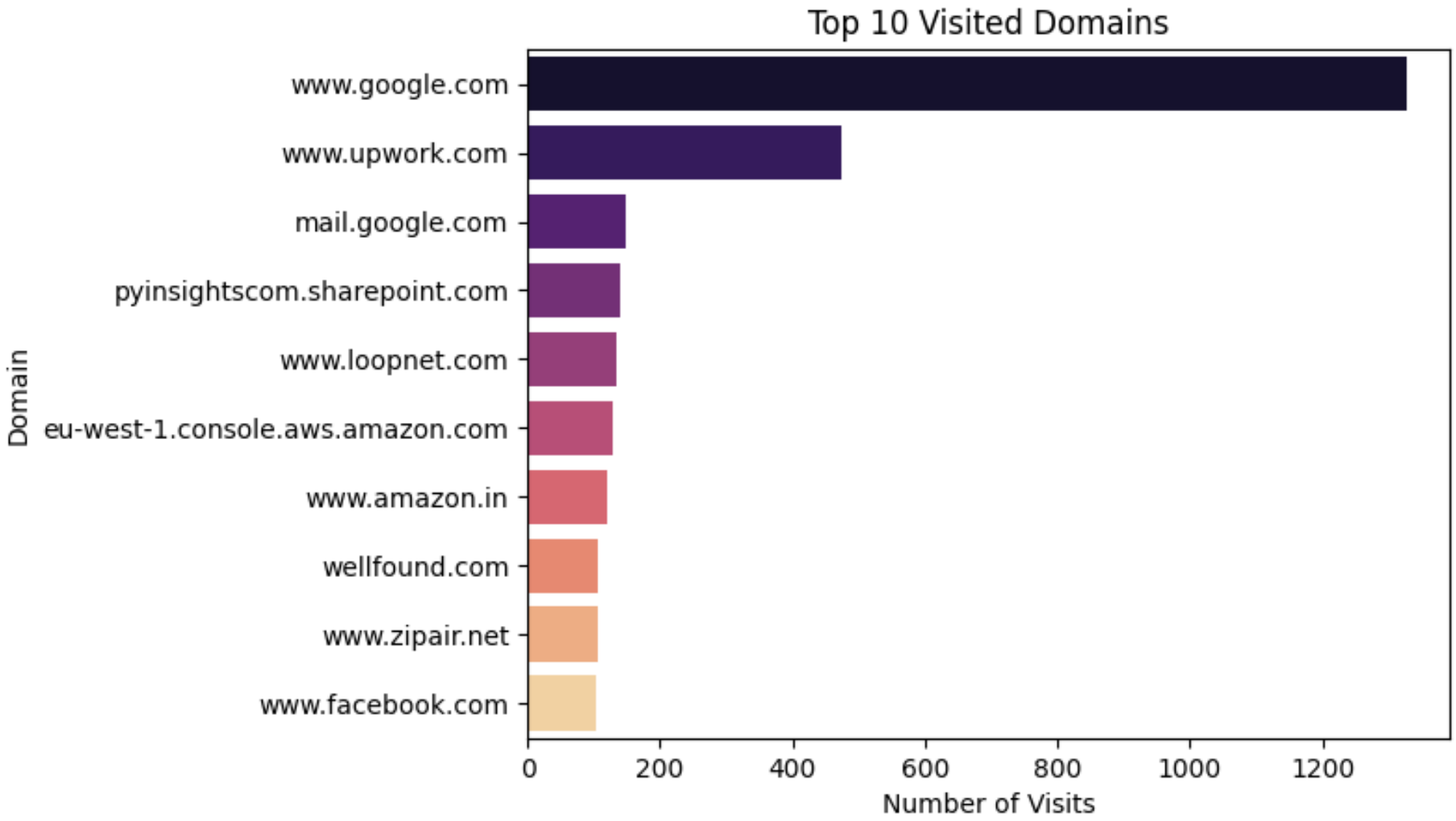
VISUALIZATIONS



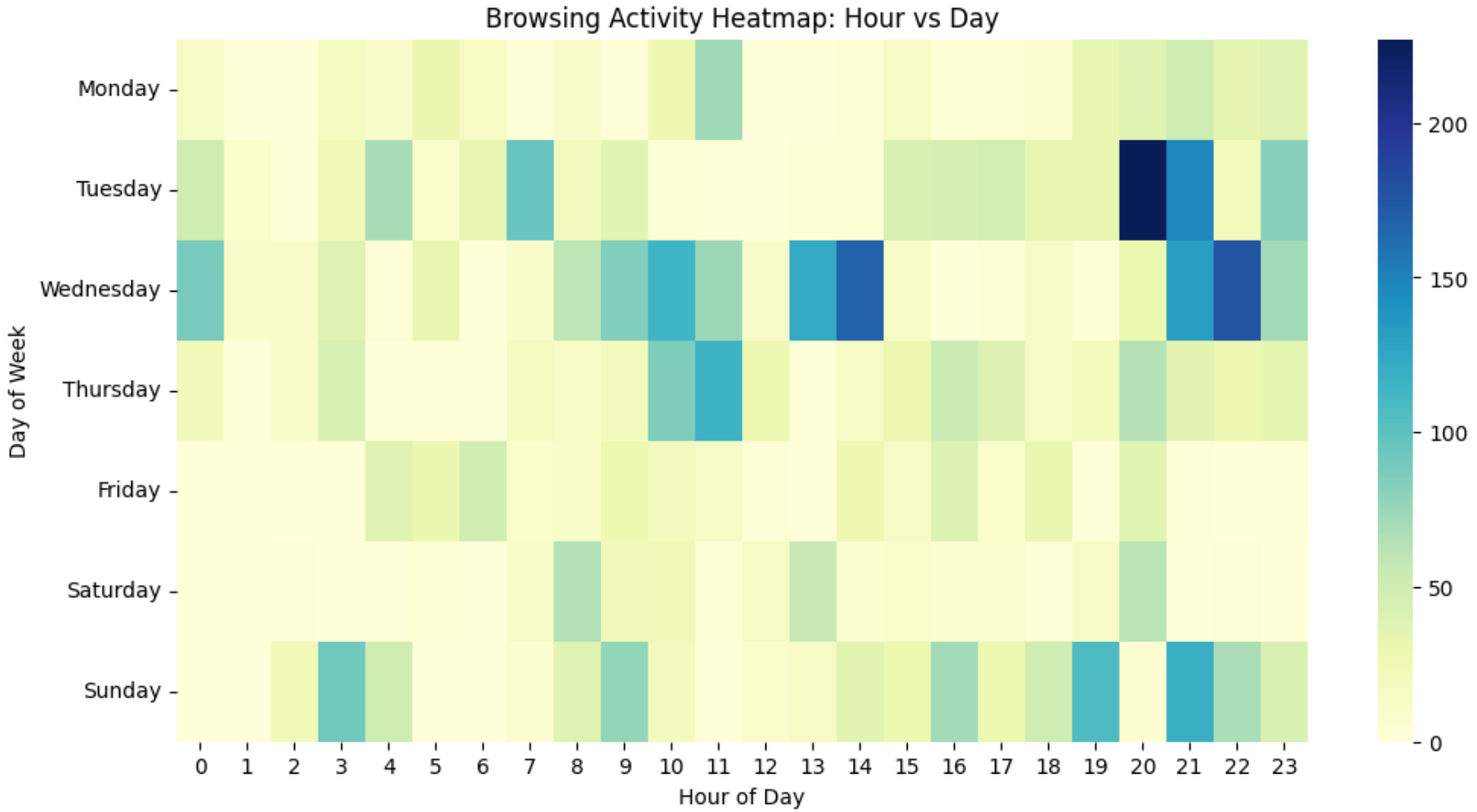
VISUALIZATIONS



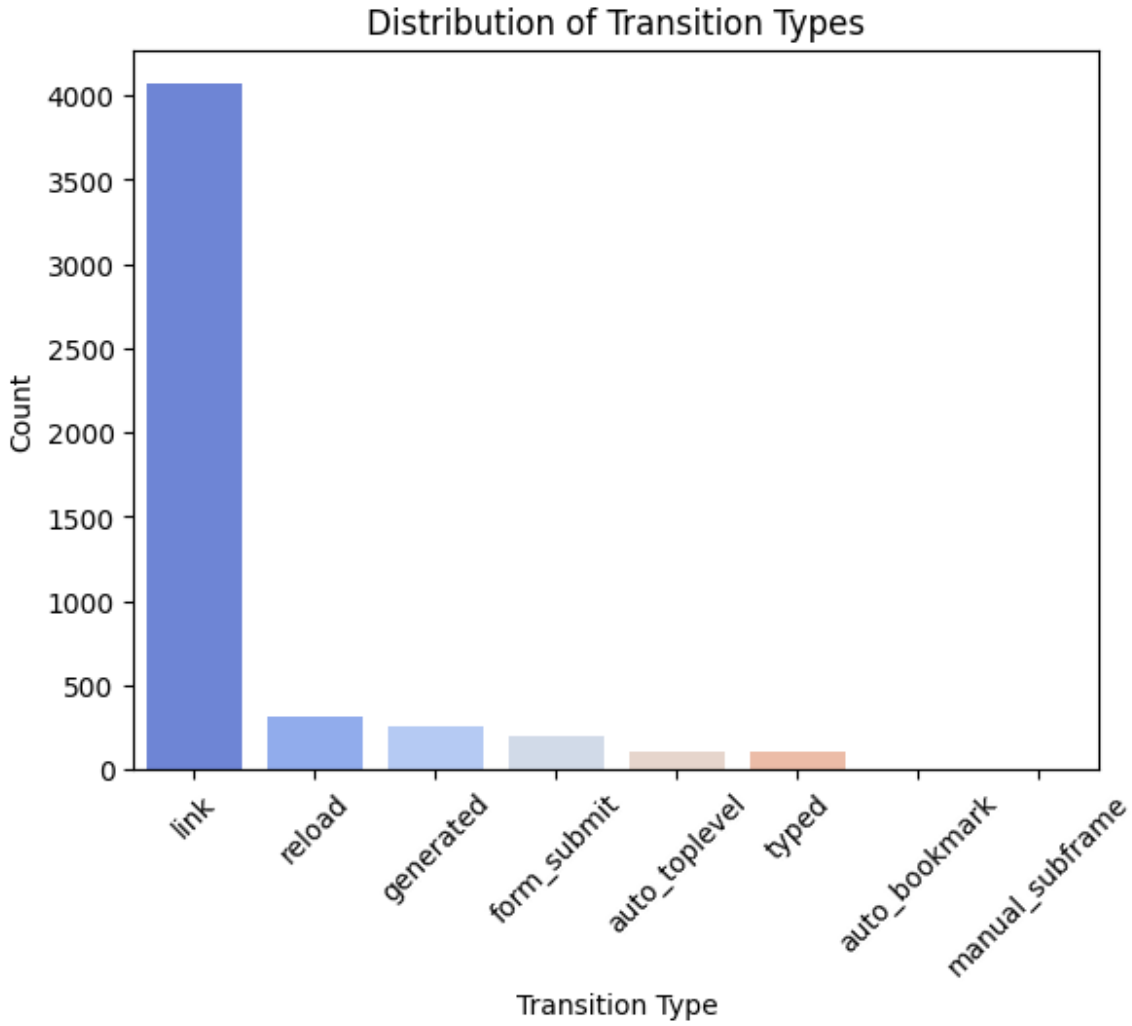
VISUALIZATIONS



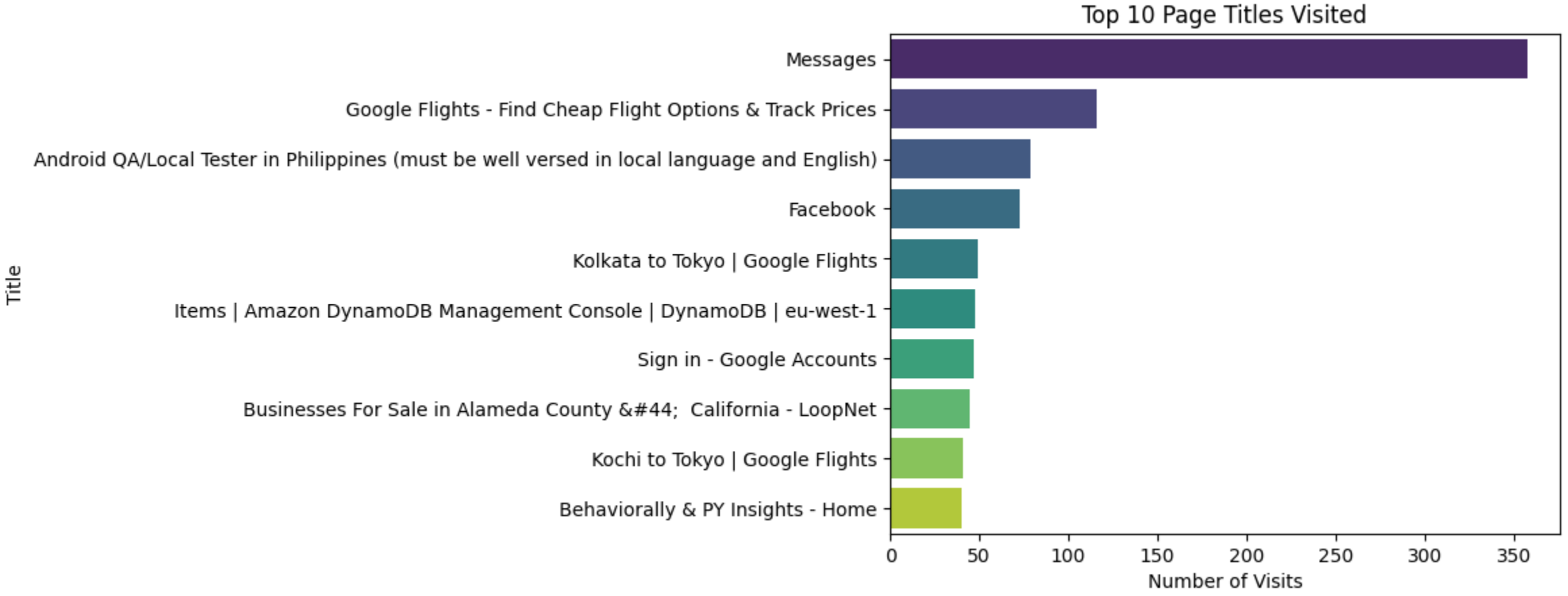
VISUALIZATIONS



VISUALIZATIONS



VISUALIZATIONS



INTERPRETATION & ANALYSIS

From the frequent visits to job search platforms, especially LinkedIn and Indeed, and several career-related domains, there's a clear pattern of active job-seeking behavior. The user shows recurring interest in Quality Assurance Tester roles — we infer this based on:

- Repeated access to QA-related job listings.
- Short bursts of high activity tied to such job posts.
- Deep scrolling within QA job search pages in certain sessions.
- This suggests targeted exploration, rather than passive browsing.

The user’s visits to country-code domains (e.g., .ca, .co.uk, .in, etc.) and job boards from various nationalities — suggesting an interest or inquiry about foreign opportunities or moving abroad. Coupled with geographically dispersed surfing habits, this indicates a user who could be either:

- Open to remote roles globally
- Actively investigating travel/emigration options concurrent to job search.

INTERPRETATION & ANALYSIS

Browsing Activity by Hour of Day:

- Peak activity is seen between 8 PM and 11 PM (20:00–23:00), indicating the user is most active online during late evenings.
- There's low activity during early morning hours (1 AM–6 AM), which is typical and shows sleep/rest periods.

Browsing Activity by Day of Week

- Highest activity occurs on Wednesday, followed by Tuesday and Sunday.
- Lowest activity is observed on Friday and Saturday, suggesting reduced browsing, possibly due to social/personal time offline.

INTERPRETATION & ANALYSIS

Top 10 Visited Domains

- www.google.com dominates with the highest number of visits (~1300+), followed by www.upwork.com.
- Other popular domains include Google services, SharePoint (likely work-related), and job platforms like LoopNet and Wellfound.
- The browsing activity leans heavily toward productivity/job-related and utility websites.
- The user's work involves the use of Amazon's DynamoDB tool, active during the usual work hours (9AM-6PM)

Browsing Activity Heatmap (Hour vs Day)

- Peak activity: Tuesday evenings (especially around 8–10 PM) are the most active.
- Other active periods include Wednesday mornings, Thursday late mornings, and Sunday evenings.
- Low activity: Fridays and early mornings across the week.
- The user's schedule seems flexible, with a mix of early and late sessions, and higher intensity mid-week. Sundays are somewhat active as well.

INTERPRETATION & ANALYSIS

Distribution of Transition Types

- The majority of browsing (~4000+) happens through link clicks, indicating the user is mostly navigating by clicking links rather than typing URLs or bookmarks.
- Other transitions like reload, generated, and form_submit are also present in smaller volumes.
- This suggests a typical passive browsing behavior, often following links rather than initiating direct navigation.

Top 10 Page Titles Visited

- The Messages page tops the list, followed by:
- Google Flights and job descriptions (e.g., Android QA Tester).
- DynamoDB Console, LoopNet, and Google Accounts login.
- User activity suggests active job or project searching, possibly travel planning, and interaction with cloud tools like AWS DynamoDB.

SESSION DETECTION

For this analysis, we have defined a session as a sequence of browsing events where the time gap between consecutive events does not exceed 30 minutes. If the time difference between two visits exceeds this threshold, we consider it the beginning of a new session.

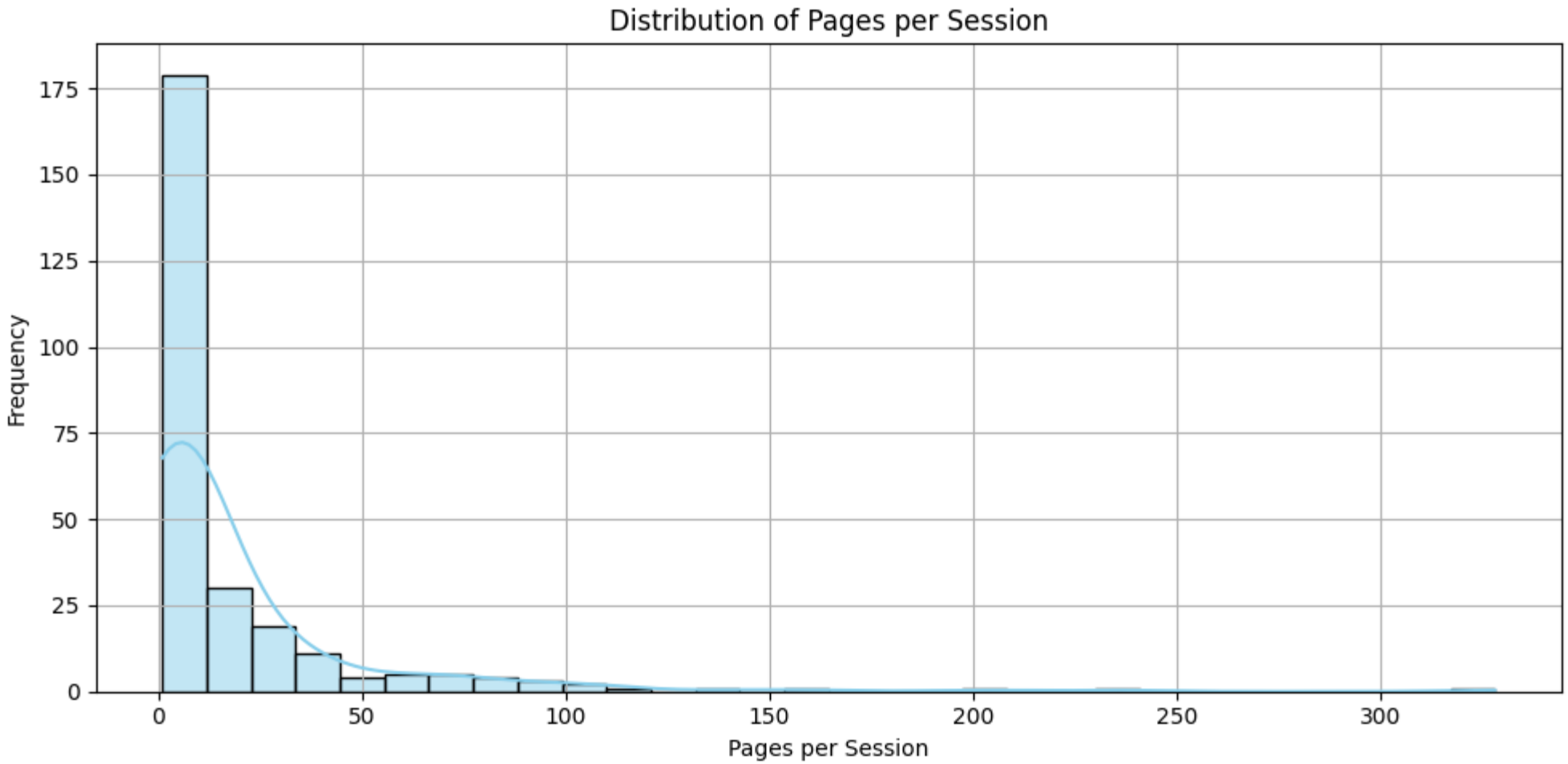
We performed session detection using the following approach:

- Converted timestamps to datetime objects for accurate time-based computation.
- Sorted events chronologically.
- Calculated time differences (time_diff) between consecutive events.
- Assigned a new session ID whenever a time gap exceeded 30 minutes.

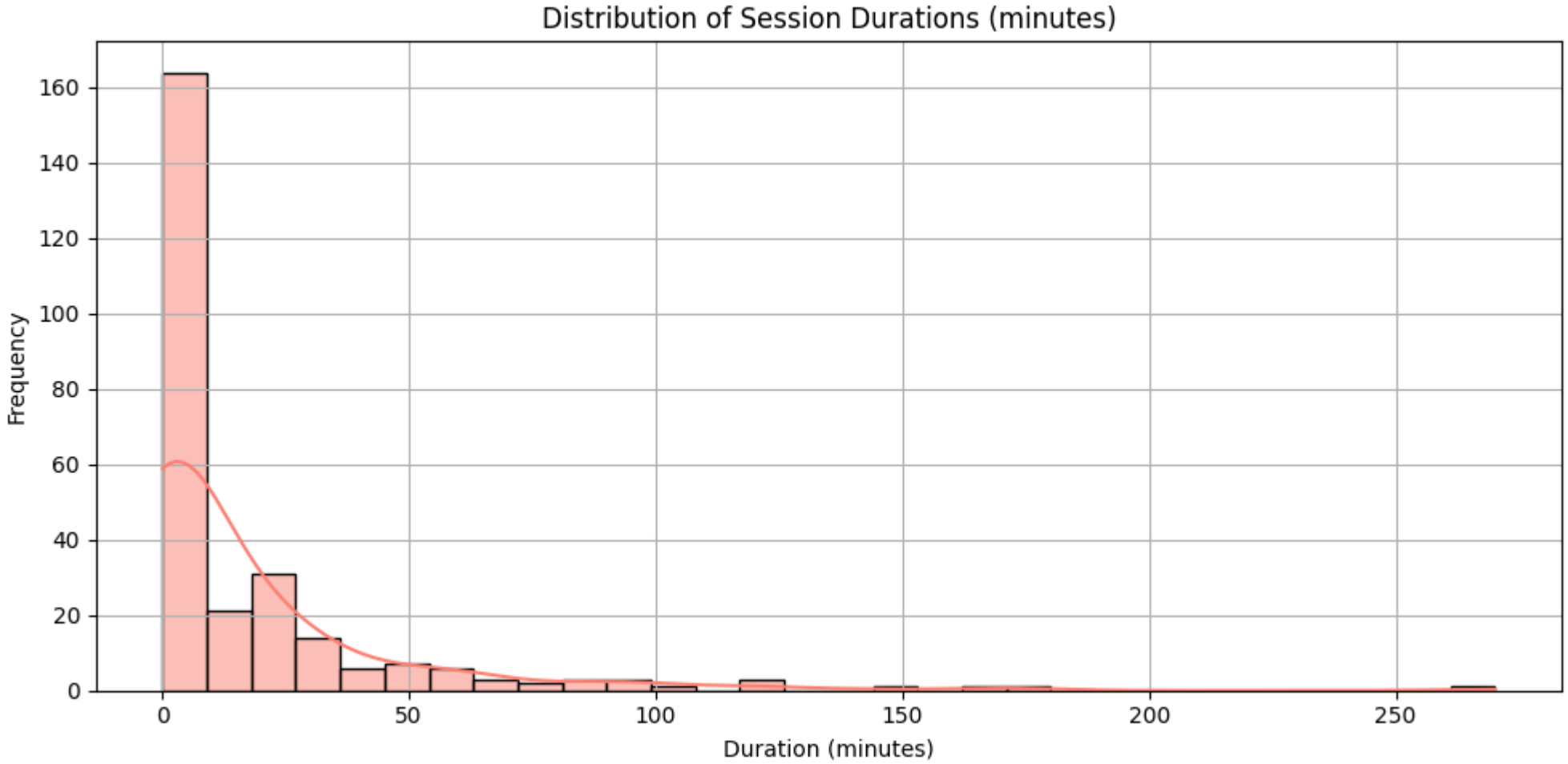
This logic allows us to group browsing activities meaningfully and analyze patterns like:

- How long users typically stay active in a session.
- How many pages they visit per session.
- When sessions occur most frequently during the day or week.
- This layer of insight enhances behavioral understanding beyond individual page visits.

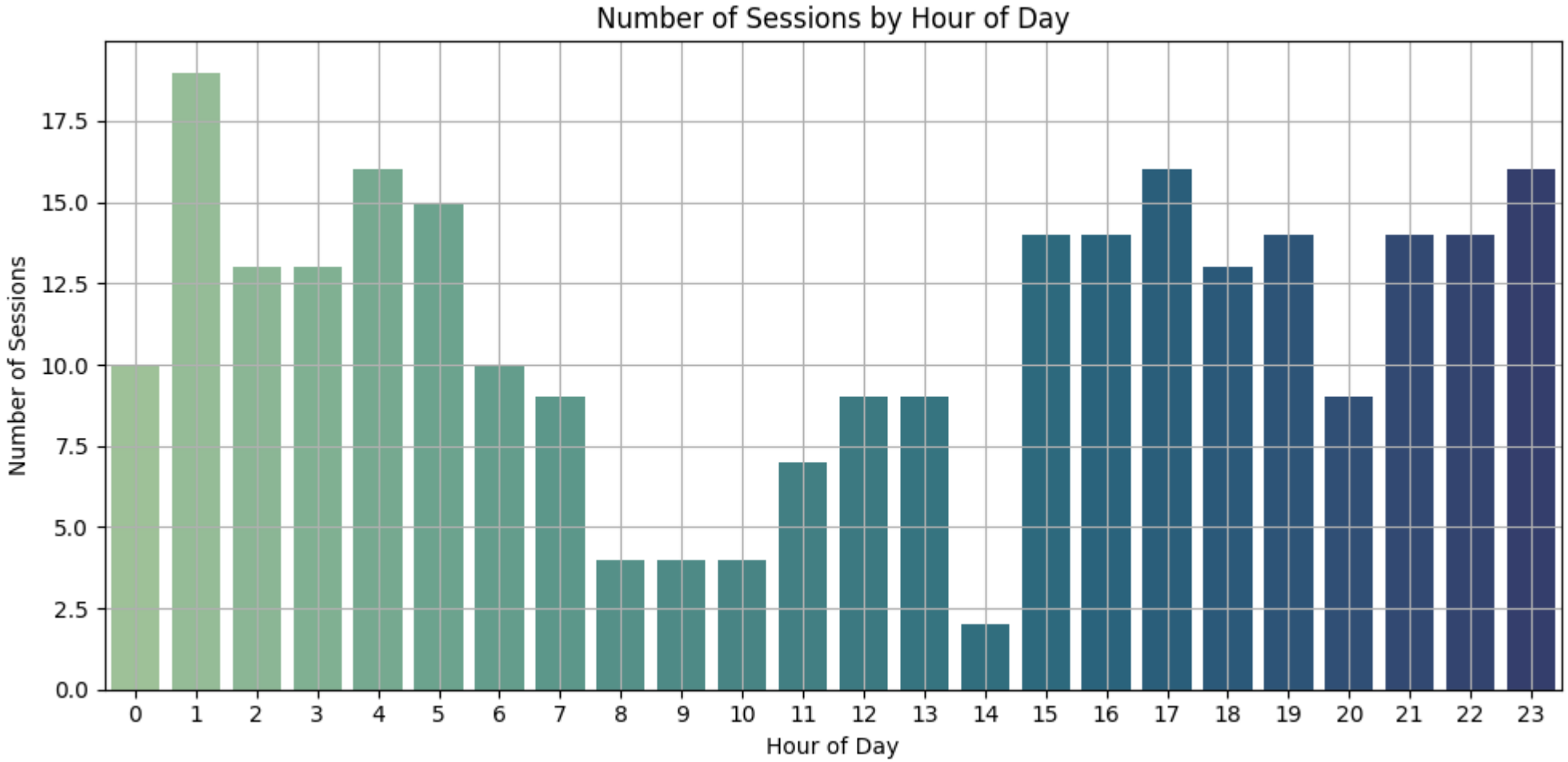
VISUALIZATIONS



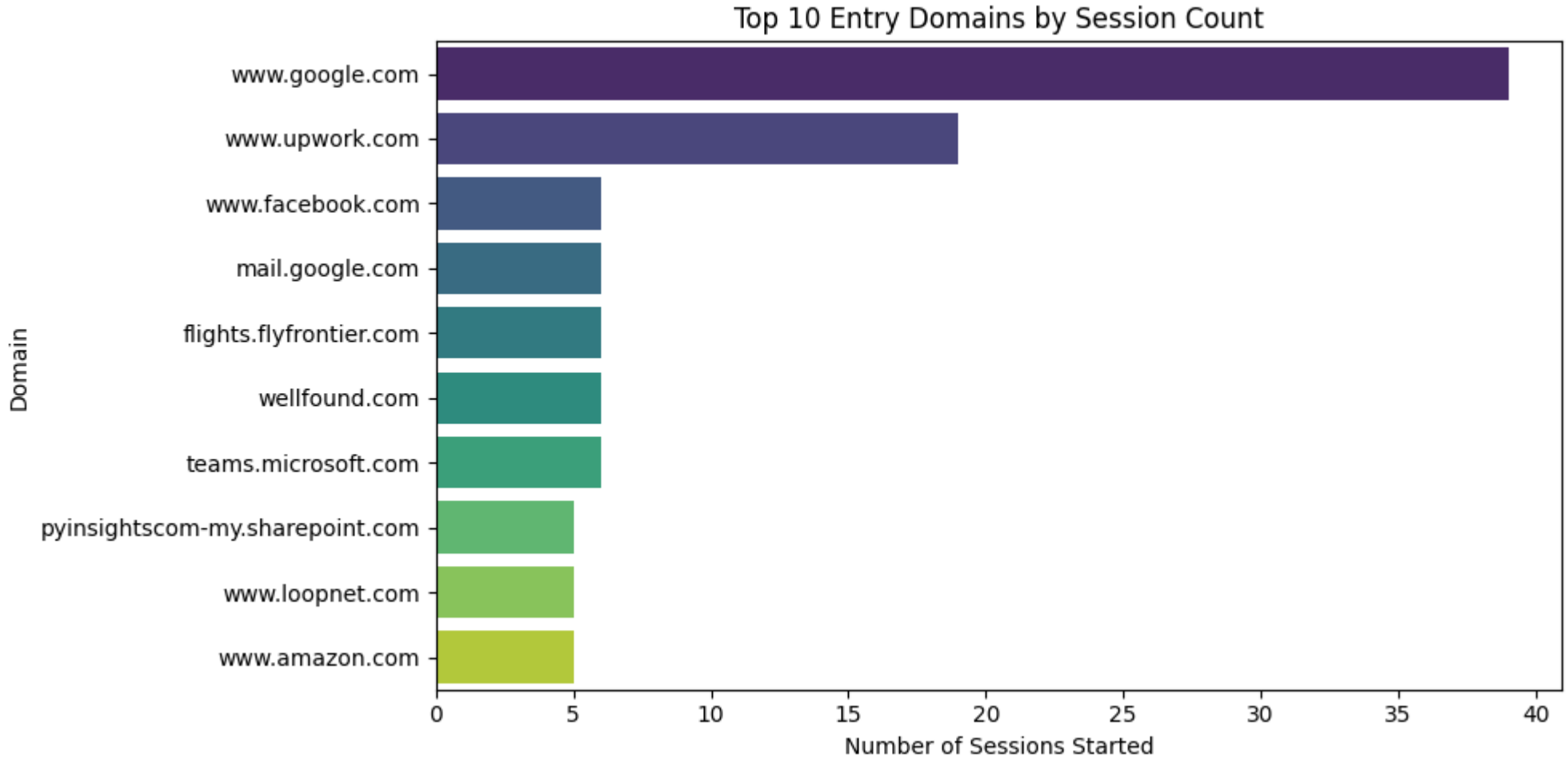
VISUALIZATIONS



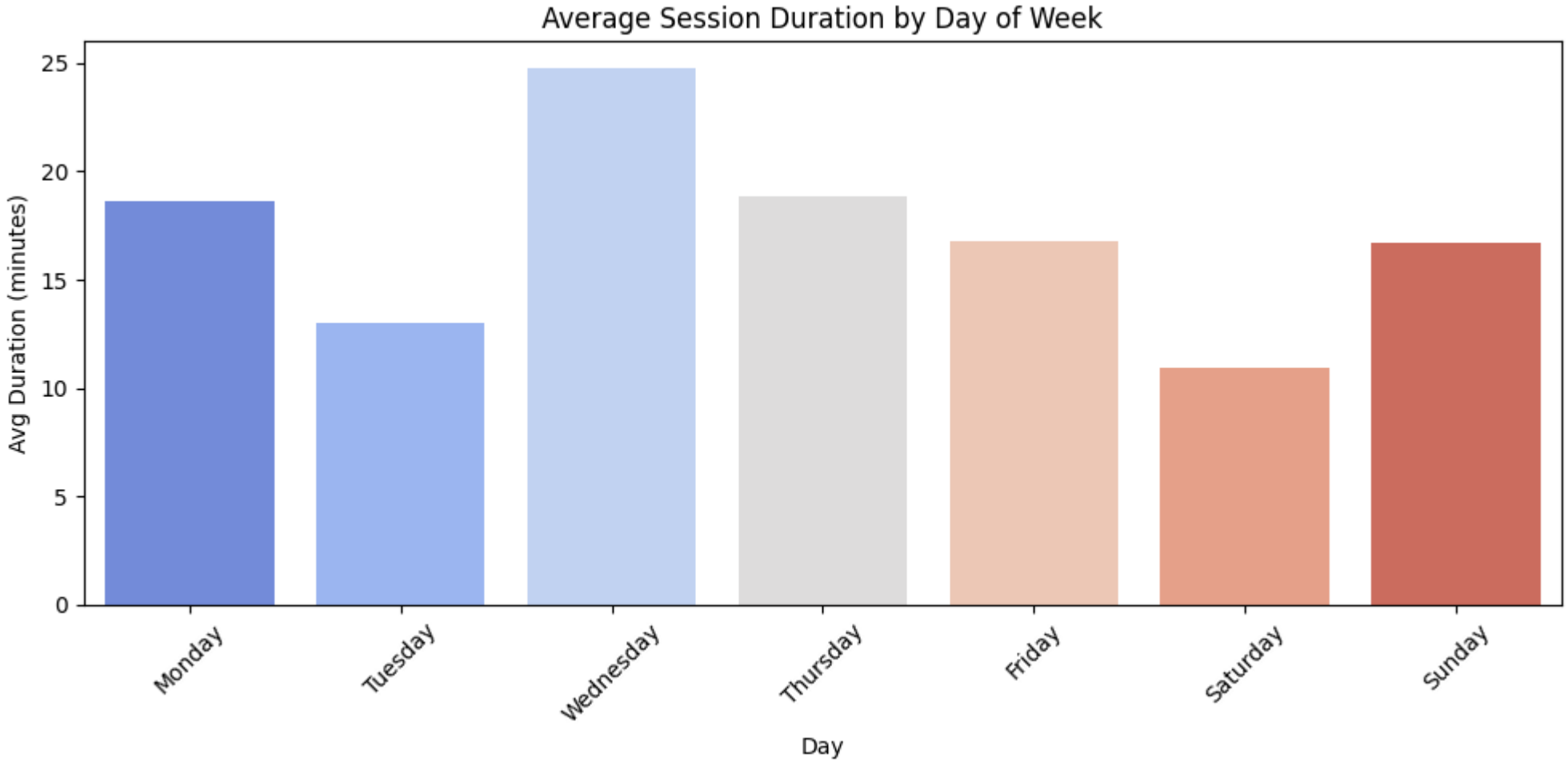
VISUALIZATIONS



VISUALIZATIONS



VISUALIZATIONS



OVERALL NARRATIVE

The user seems to be engaged in an early to mid-level QA tester or software testing-related position, as seen through repeated visits to pages and tools about QA testing, bug tracking, browser extensions, and documentation. Their interest in QA tester job listings and skill-related pages also supports this observation. , the user seems particularly interested in the QA Tester role based in the Phillippines.

Their browsing activity is largely focused within regular working hours (9 AM to 6 PM) on weekdays, which involved the use of AWS DynamoDB tool, with a peak in the middle of the week, corresponding to regular professional work habits. Session durations are mixed, since not much difference can be seen between session analysis and overall analysis, yet most are brief, indicating goal-directed activity or focused tasks. A couple of extended sessions may point to deep work or research activity. There is evidence of travel planning to Tokyo and California, or curiosity, possibly overseas, with repeated visits to travel websites and travel content. This could mean future personal or business travel or an interest in moving.

THANK YOU