Apr 6, 2023

Austin Pitts, austinpitts@berkeley.edu
Alejandro Franza Garcia, afranza@berkeley.edu
Philip Monaco, philip.monaco@bekeley.edu

- **Name of your team's GitHub repository**
  - https://github.com/UC-Berkeley-I-School/Project2_Franza_Pitts_Monaco

- **A primary dataset you intend to analyze**
  - https://analytics.usa.gov/data/
    - Visits to all domains over 30 days table, downloaded in CSV format
    - Currently, the Digital Analytics Program collects web traffic from around 400 executive branch government domains, across about 5,700 total websites, including every cabinet department.

- **Initial plots, figures, or tables**
  - Python packages: pandas, seaborn, matplotlib, sklearn
  - Data Understanding:
    - We shall use pandas to do the initial exploration of the dataset. We can use pandas methods on dataframes to construct tables for descriptive statistics and generate a dictionary of the dataset.
    - We will also generate tables to understand where there are missing values, duplicate data, and intersections in additional datasets that we can combine to generate further insights.
  - Identifying Outliers:
    - We shall use visualization packages such as seaborn and matplotlib to identify outliers that may skew additional analysis.
  - Summary Statistics:
    - We shall generate summary statistics visualizations using seaborn and sklearn that will visualize an interpretation of each of the variables and allow us to explain why we chose to use them in our analysis.
  - Analysis/Modeling:
    - We shall use seaborn and sklearn to explore relationships between attributes via scatter plots, correlation, cross-tabulation, and group-wise averaging as appropriate.
- **Some of the variables (column names) you intend to explore and what kind of insights you expect to glean**
  - domain,visits,pageviews,users,pageviews_per_session,avg_session_duration,exits
  - Insights: what domains are most visited, what domains have the highest user retention, correlations between columns

- **Supplemental datasets, if any, to complement your primary dataset - this means links, columns that you'll join on, etc.**
    - Potentially, depending on the results of the early exploration

- **What you plan to cover in the final report and how you plan to organize it.**
    - Data exploration
        - Does the table contain information for all the 5,700 websites?
        - Any missing or null values?
        - Do aggregate values fall within expectations (eg. # users as % of US population seems realistic)
        - Are there any unexpected values? (eg. visits > page views)
        - Do we understand the meaning of all columns? For example “exits”.
    - Analysis
        - What were the most visited domains in the last 30 days? Any interesting trends? (maybe a vertical pattern or a common use case)
        - What are the domains / groups of domains with better engagement or user retention metrics (pageviews_per_session, avg_session_duration)
        - Is there a correlation between lower engagement and higher bounce rate (exits per visits)