Amitree Data Science Business Intelligence Challenge

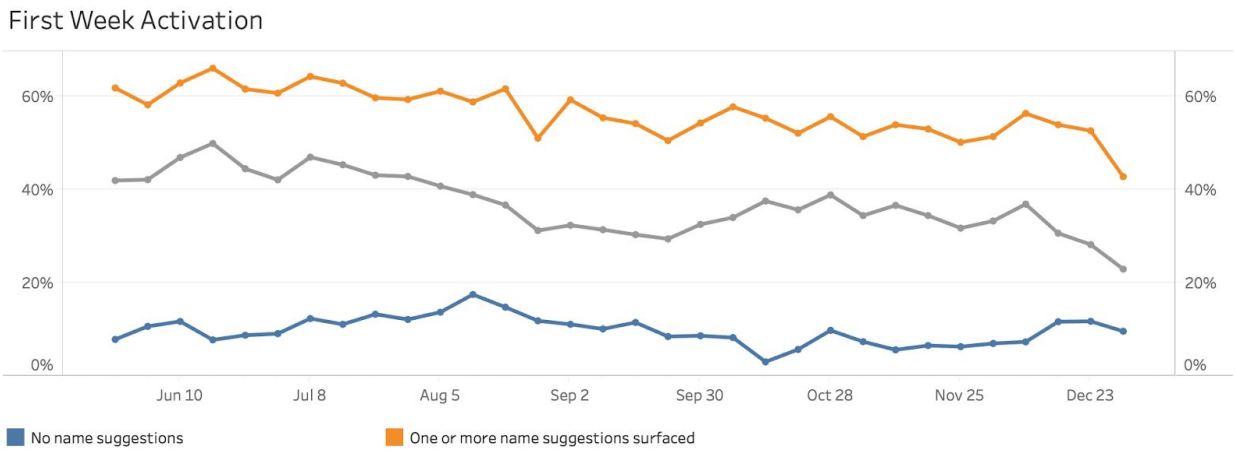
Welcome to the Amitree Data Science Business Intelligence take home challenge. This challenge is intended to allow you to show your skills in two primary areas of data science practice: 1) data management and 2) visualization construction. For this challenge you will be presented with a data table of event data for a hypothetical product. You will be asked to transform the data and build visualizations which describe specific aspects of the data.

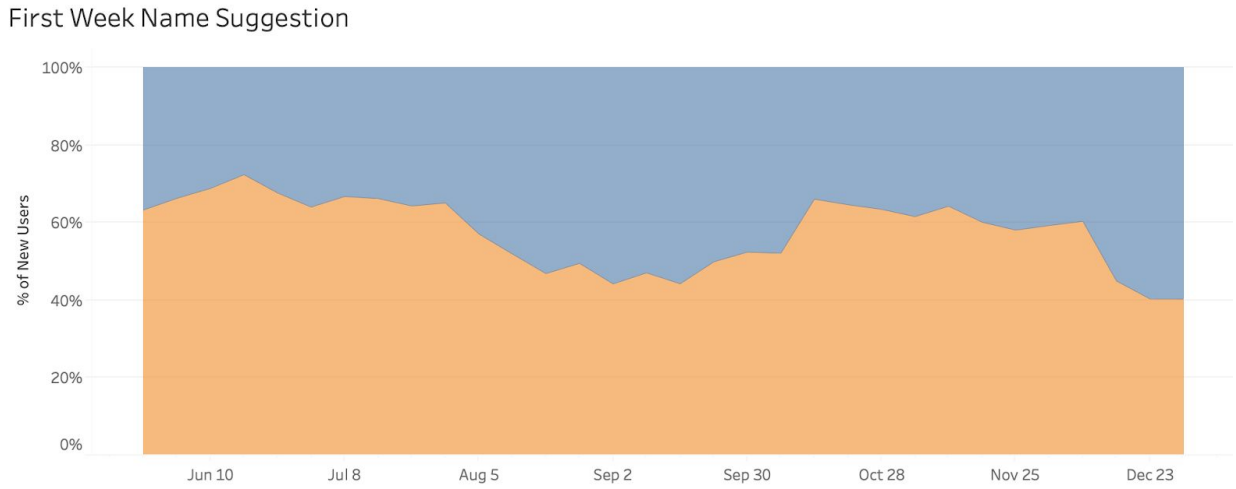
This exercise should take about 4 hours. We recognize that this exercise is somewhat involved and it may not be possible for everyone to demonstrate the above skill sets to the best of their abilities in the time given. This challenge is intended to give us insight into how you think about solving data science problems including your approach to data management, ability to describe data, as well as your coding style. In this spirit, it is more important for the purposes of the challenge to contribute something in each of the key sections rather than doing extensive work on any one particular area.

Scenario & Data

In this hypothetical scenario, you work for a company which stores and organizes digital photographers for users. One of the features of the product is image recognition software, which surfaces suggestions to users for what to name their photographs. Prior research has revealed that users who name their photographs are much more likely to continue using the product, so the success of the feature is considered critical the health of the company.

Currently, the company pays a third party provider to track the use, success, and impact of this feature, but now your company would like to be able to do so in house. The first priority is to generate an internal version of this dashboard, which is used throughout the company to track progress towards goals:





The dashboard shows two visualizations. The visualization on top reflects the proportion of users who are active in their first week on the product, with the date on the x-axis. A user is considered ‘active’ if they have named a photograph, either by doing so manually or accepting a suggestion. This metric is shown for all users (in gray) and is also broken out for users for whom at least one name suggestion was surfaced (orange) and for whom none were (blue). The visual on the bottom shows the proportion of users for whom at least one name suggestion was surfaced (orange for at least one, blue for none) in their first week on the product. Again, the date is on x-axis.

The data you have available to reproduce the dashboard is a user events table, which you can access here. The table contains approximately 1.2M rows, one row for every event in the product, including actions taken by the user and name suggestions surfaced to a user. Each event contains an event_id, user_id, event_type, and timestamp.

Goals & Deliverables

The first goal of this challenge is to transform the data from its current state into a table or tables which will provide the information needed to provide the information contained the dashboard above. It is up to you to determine appropriate table structure and content, including selection of appropriate metadata and annotations. Deliverables for this goal include structured data outputs (eg pg_dump file, csv(s)), and commented code used in data processing. We recommend spending at most 2 hours on this portion of the challenge.

The second goal is to build a dashboard or series of visualizations which describe the data and provide similar metrics to those contained in the dashboard above. Again, it is up to you how to approach these visualizations: you can create a similar set of visuals as the dashboard or take a different approach, but the information conveyed should be the same. Deliverables for this goal include the dashboard or visualizations, as well as any code or workbooks used to create them. We recommend spending at most 1 hour on this portion of the challenge.

The third goal is to briefly describe your data visualizations and examine any discrepancies between your output and the dashboard above. Imagine your audience to be the person(s) who are accustomed to referencing the dashboard above and will now be using your set of visualizations instead. What should they be aware of? How will you pre-empt any questions or concerns they might have about the differences between what they are used to and what you created? The deliverable for this goal is a paragraph or two of text and anything additional you might add to help explain your visualizations. We recommend spending at most 1 hour on this portion of the challenge.

Tools & Methods

You may use any tools that you find appropriate to accomplish the exercise. We recommend using Python or R to perform operations on the data and build models/perform statistical tests, PostgreSQL to store the data in a table, or series of tables, and Tableau, D3.js or ggplot2 (R) for descriptive analysis and visualization, but you may use any tools you like as long as you are able to show your work.

Good Luck!

We recognize that this challenge is difficult and that finding time to complete it may not be easy for all applicants. Please read this document thoroughly and give yourself an opportunity to understand the basic structure of the data before beginning the challenge. Finally, don't hesitate to reach out to us with any concerns around timing, any questions about the data, or for clarification on the content of the challenge.