W209 Homework assignment for Week 2

Name: Linfang Yang

Email: yanglinfang@ischool.berkeley.edu

Assignment 1: Getting Started in Tableau (credit/no credit)

Using one of the provided sample datasets already loaded within Tableau, create four or more views that highlight different features of the data. Experiment with varied views by selecting different dimensions and measures, several visualization types (e.g., scatter plot, bar chart, etc.), changing color and shape encodings, and applying filters. Provide screen captures of each of the views with clear labels describing the axes and variables. No additional text is needed. This is a credit/no credit assignment.

Tutorials to Get You Started in Tableau:
https://learn.datascience.berkeley.edu/mod/page/view.php?id=15006#/cardContent
http://www.tableausoftware.com/learn/training

Dataset source

In this assignment, I explored different diagrams that Tableau can provide using data set from kaggle challenge for Display Advertising click prediction.

https://www.kaggle.com/c/criteo-display-ad-challenge/data

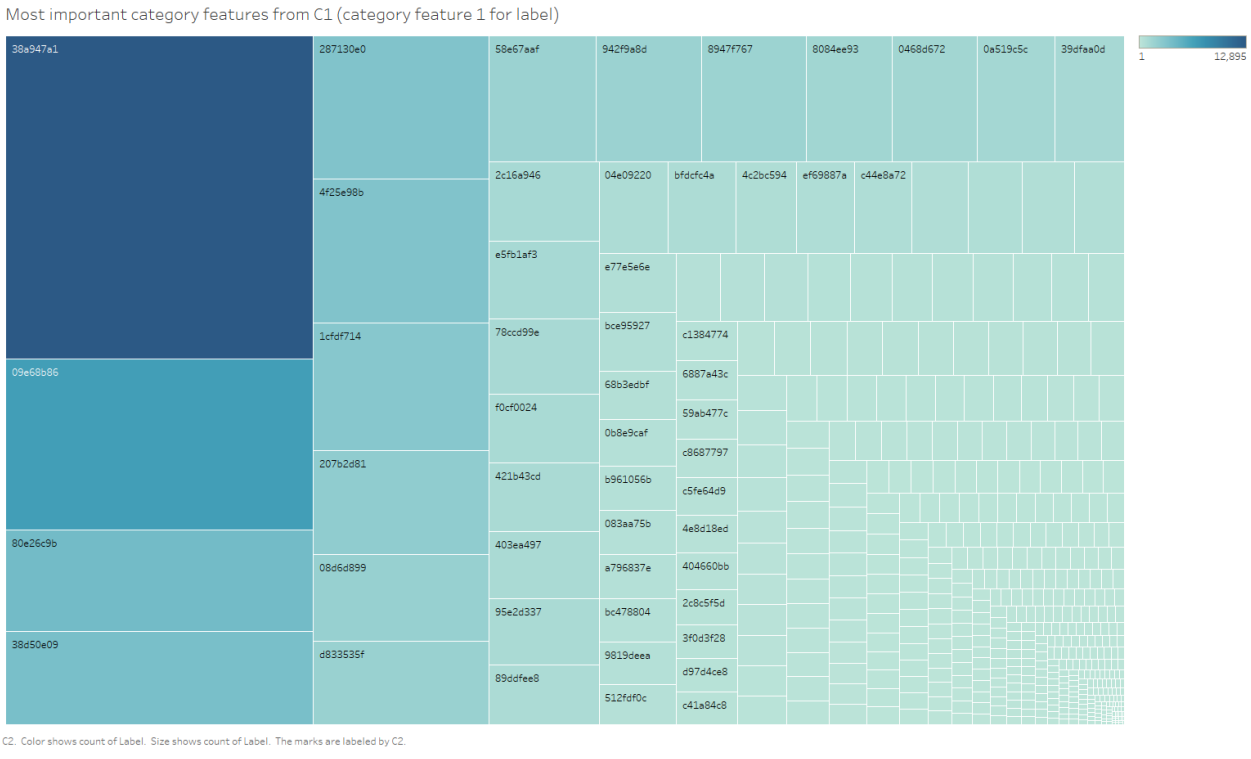
Data fields

Label - Target variable that indicates if an ad was clicked (1) or not (0).

B1-B13 - A total of 13 columns of integer features (mostly count features).

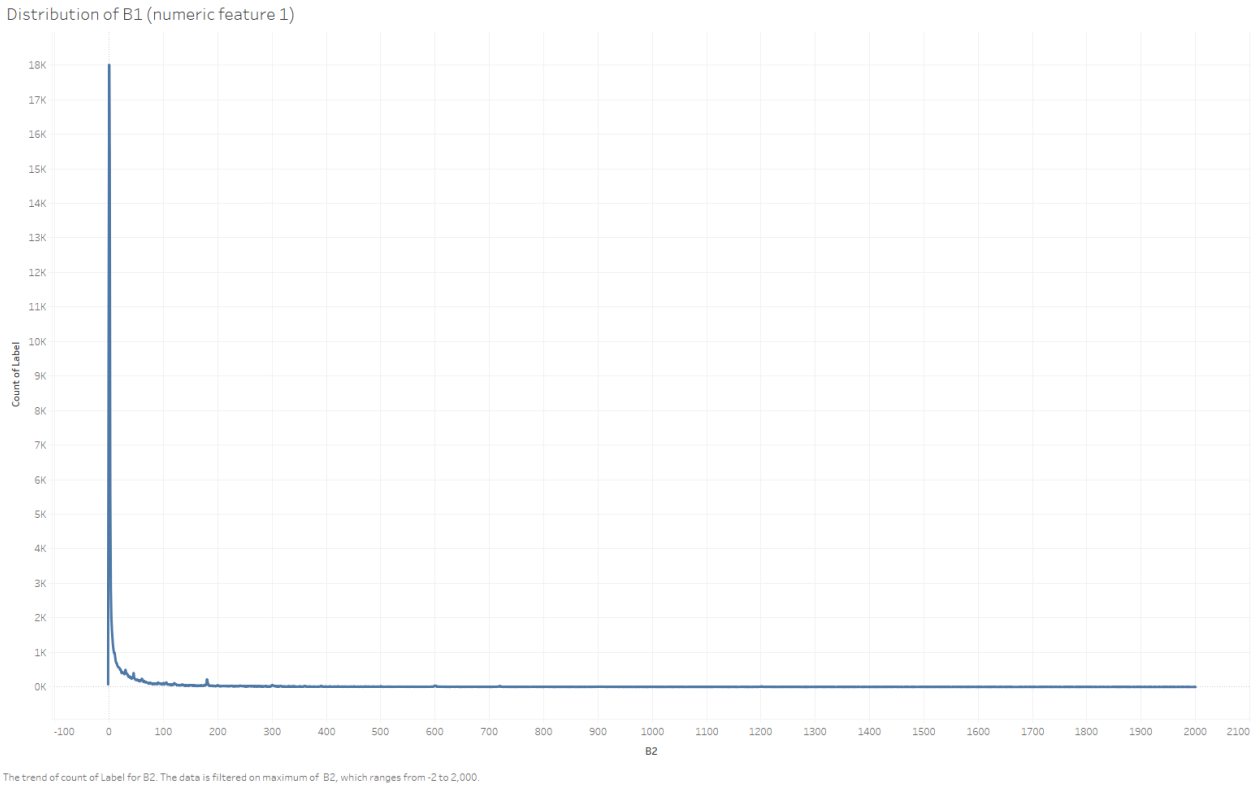
C1-C26 - A total of 26 columns of categorical features. The values of these features have been hashed onto 32 bits for anonymization purposes.

Tree map



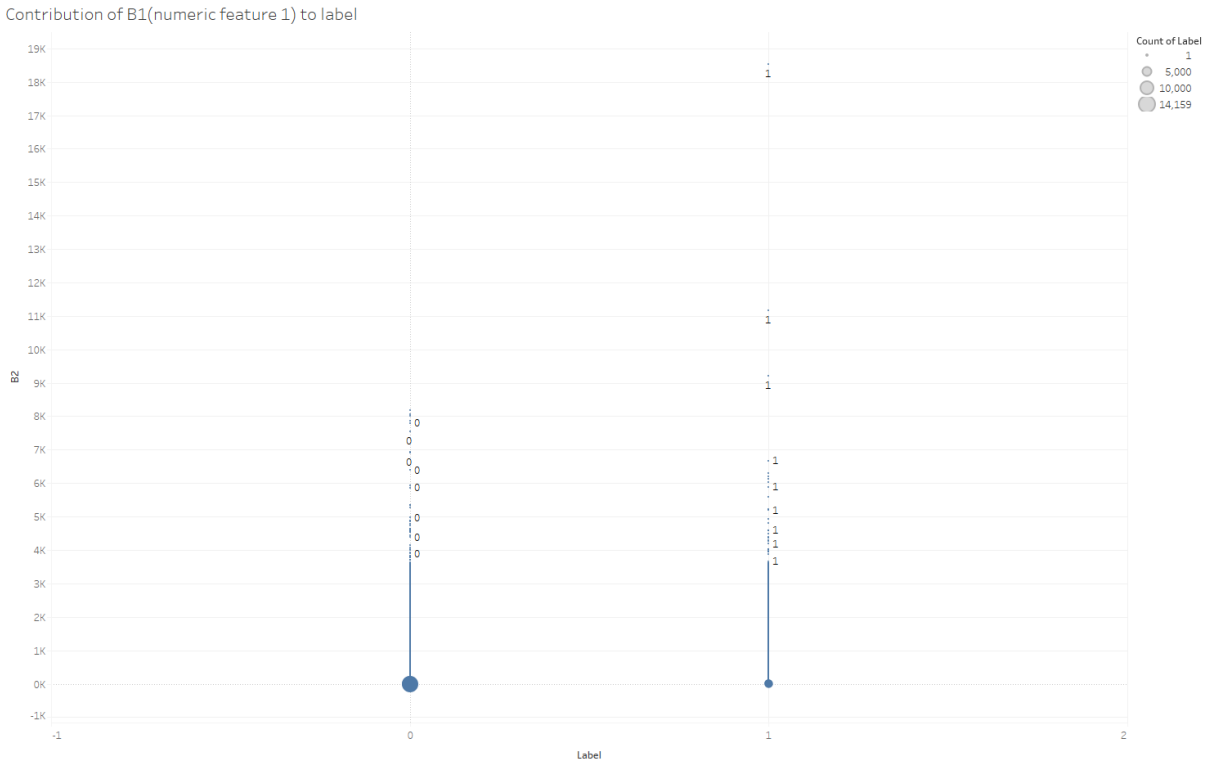
The above tree map explored relationship between label and C1 (categorical feature 1). And identified the most common values for C1 (the blocks with darker colors). Tree map is good at visualizing obvious patterns values among complicated values, and makes efficient use of space.

Distribution using line chart



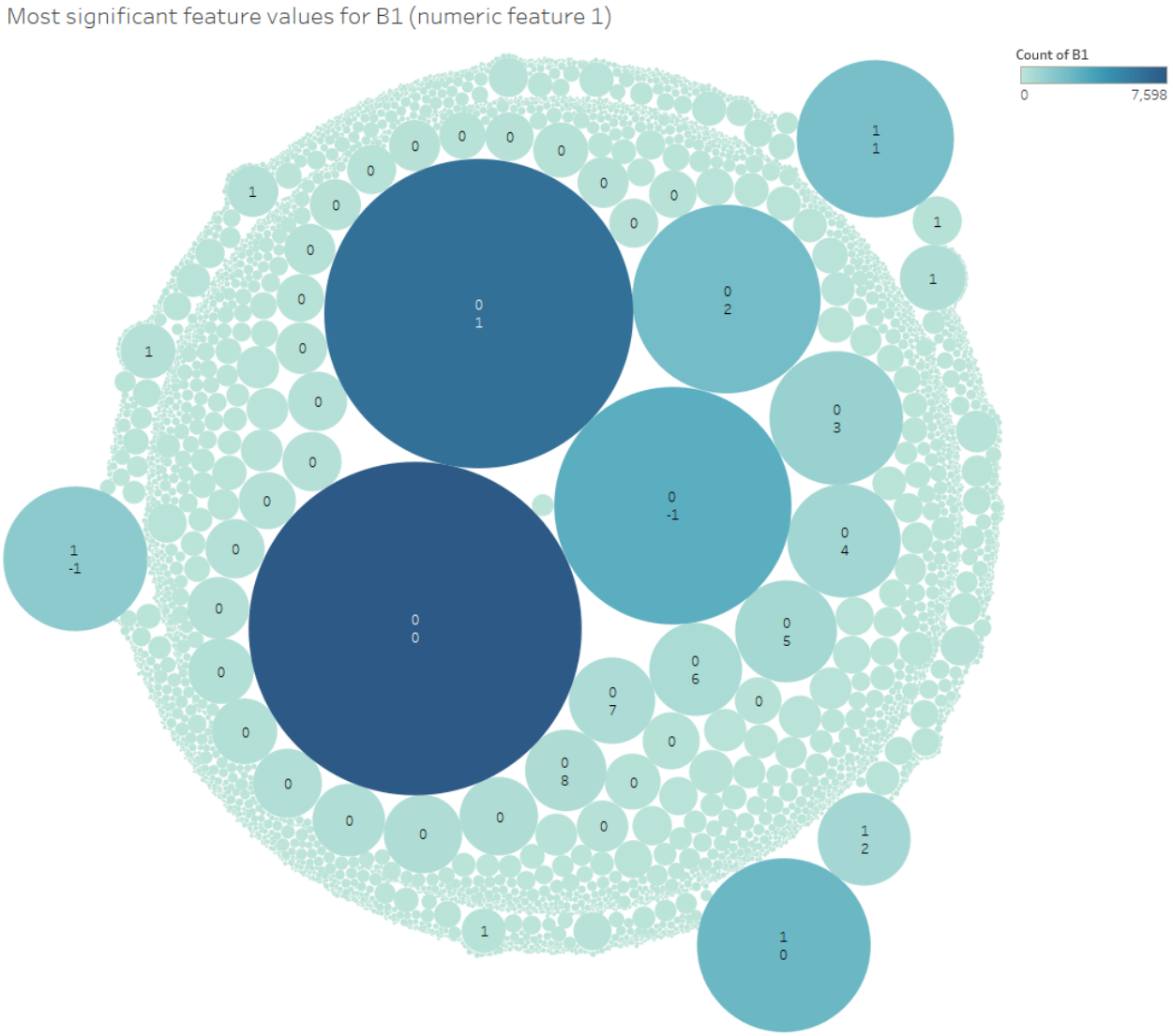
The above line chart explored the relationship between B1 (numeric feature B1). And identified the significant skew of B1 values, positive skew elongated tail at the right (more data at the right tail than would be expected in a normal distribution)

Scatter plot



The above scatter plot of B1 (numeric feature 1) over label identified some outlier of label = 1 and B1 value above 9K. Scatter plot is good at detecting outliers.

Packed bubbles chart



The above packed bubbles identified the most common value pairs between B1 (numeric feature 1 and label). Packed bubbles are sort of like tree map, but it is displaying more dimensions of data via value pairs, color (count of the pair). You can use bubble chart instead of scatter chart if your data has three data series that each contain a set of values.