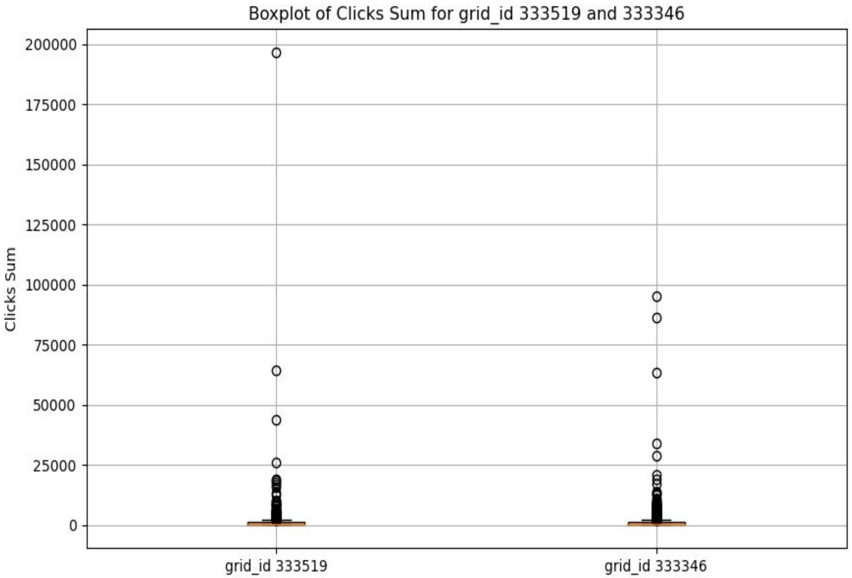
Capstone Project: Heatmap Anomaly Detection

Week 4 Progress Report

Agenda for Today

- Some EDA of two datasets
- Discuss baseline approaches and metrics
- Questions (technical and administrative)

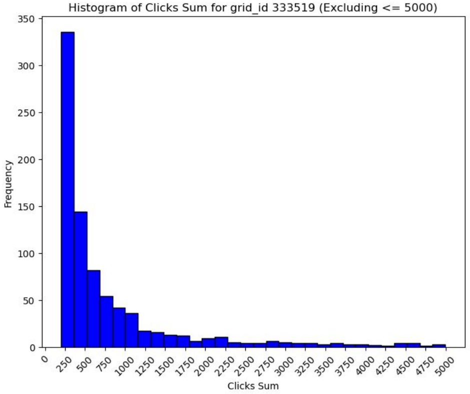
Heatmap Dataset: Total # of clicks

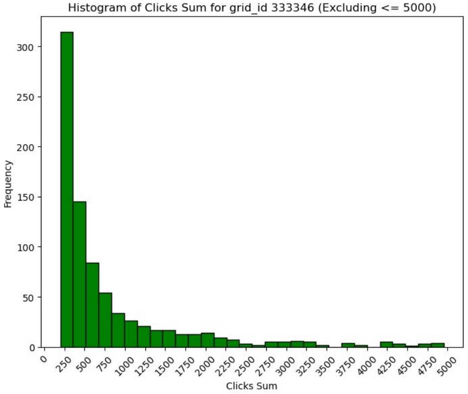


Notable features:

- Some extreme outliers → Zoom into domain-id's with less than 5000 total # of clicks.

Heatmap Dataset: Total # of clicks

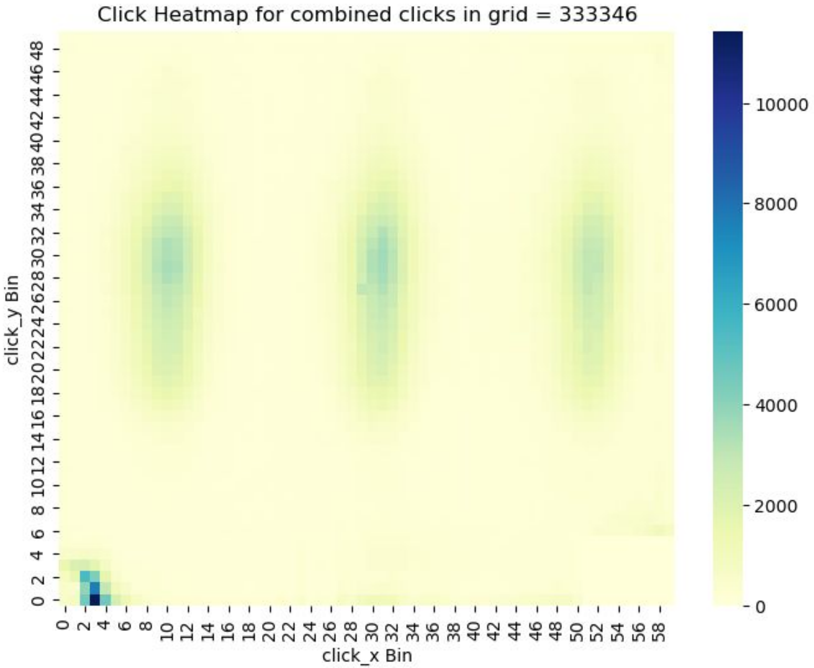




Notable features:

- Heatmaps constrained to have more than 200 total clicks.
- Bulk below 1000 total clicks
- Distributions of data for both types of grid-id’s very comparable.

Heatmap Dataset: Grid_ids



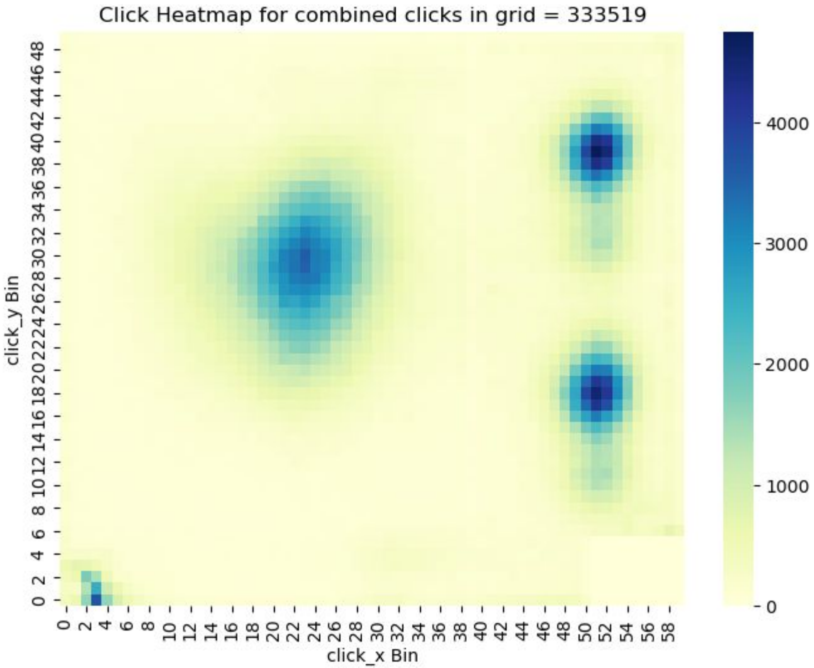
Notable features:

- There are 13.1M total clicks
- Three long stripes in center of image
- Bottom left corner concentration of clicks and (possibly) some missing region (non-empty bins though)
- Bottom right corner large area missing click data.

Questions:

- Why are clicks missing in bottom corners?
- Do you have an example image for this type of heatmap (for more intuition)?

Heatmap Dataset: Grid_ids



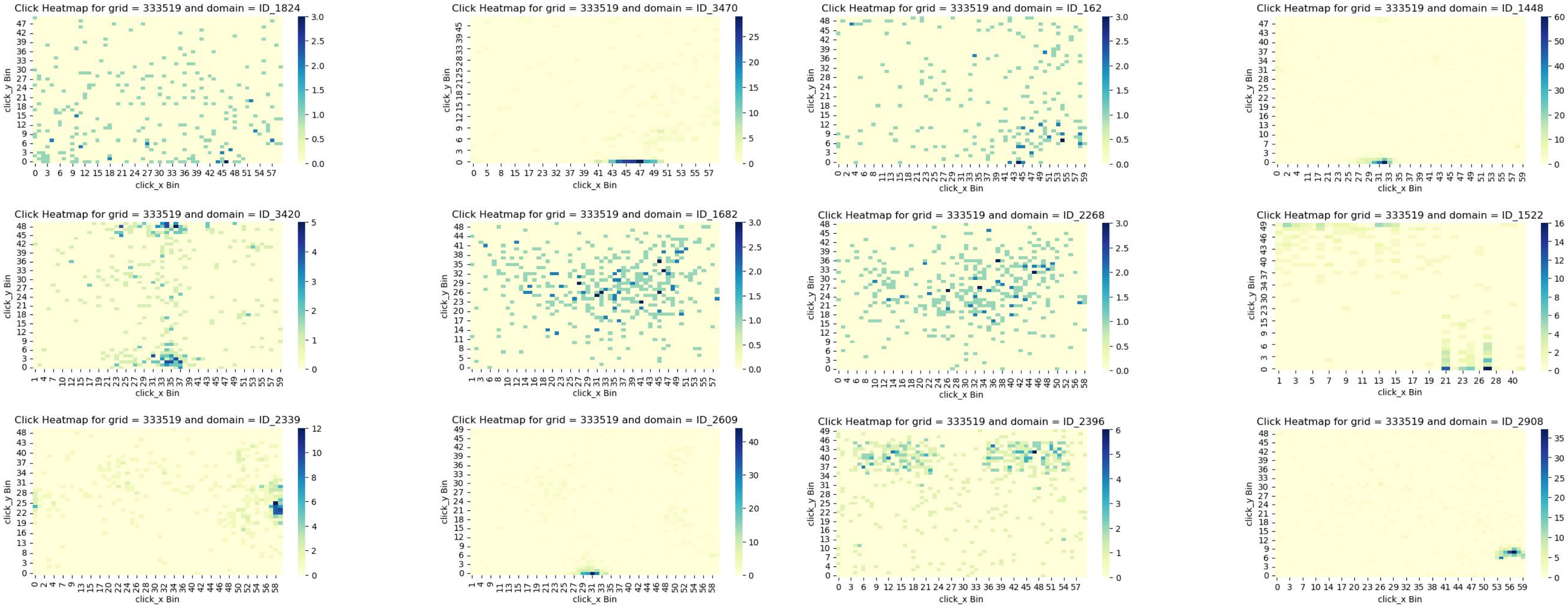
Notable features:

- There are 12.7M total clicks on this type of grid.
- Triangle type structure.
- Smaller peak below larger one on the right hand side.
- As before bottom corners missing.

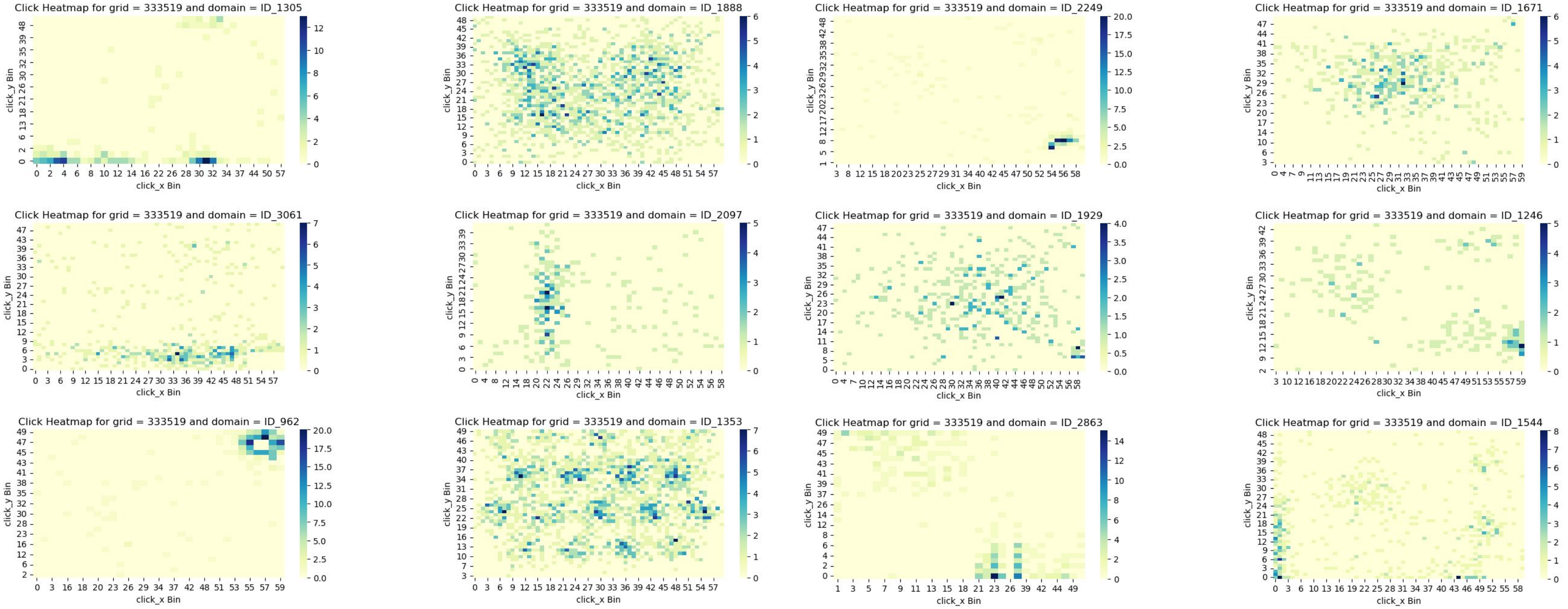
Questions:

- Are these clicks generated 1/user or can a single user generate many clicks? Is there a way to distinguish?
- Do you have an example image for this type of heatmap (for more intuition)?

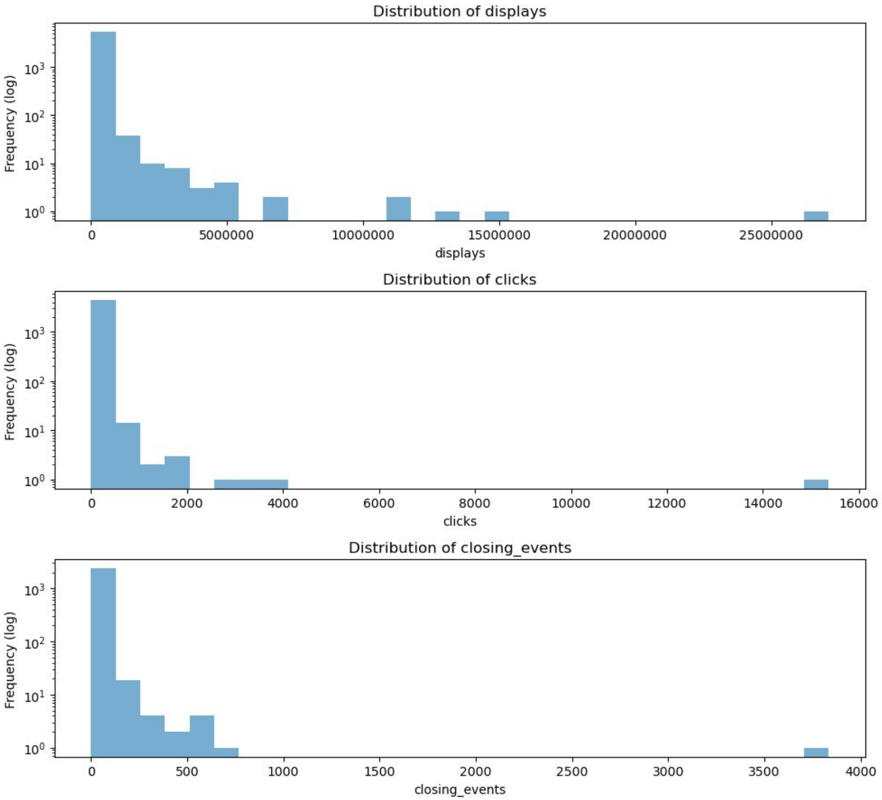
Heatmap Dataset: Example of Broken banners?

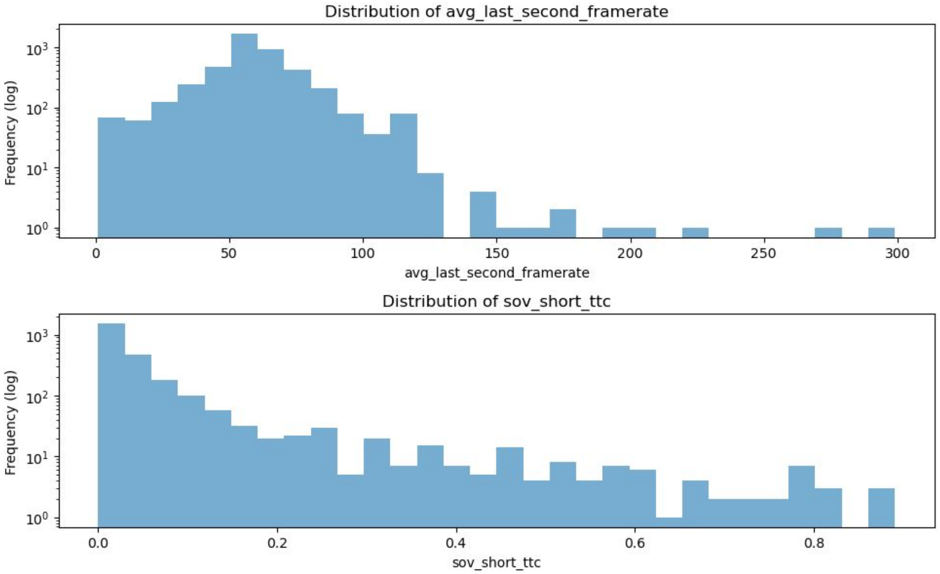


Heatmap Dataset: Example of Broken banners?



Performance Metric Dataset: Major Metric Distribution





Metric Dataset: Interpretation

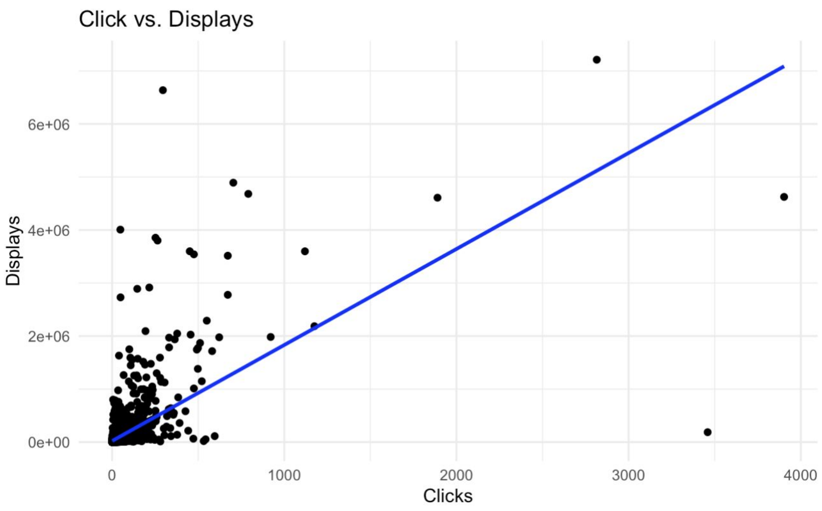
- domain: identifier for domain
- grid_id: identifier for grid, related to specific layout
- webview_height, webview_width: dimensions of the banner
- displays: # times the ad was displayed
- clicks: # times users clicked (why are these numbers different from the heatmap data? Because they have different time frame)
- landed_clicks: # clicks that successfully led to a landing page
- non_bounced_clicks: # clicks where the user did not immediately leave the site
- closing_events: # times users closed the ad?
- avg_last_second_framerate: average framerate in the last second?
- sov_short_ttc:
- sov_short_ttc_global:
- sov_short_ttc_score:

Metric Dataset: Data Cleaning

- Counts of missing values in each column :
    - Same rows missing for ‘clicks’, ‘landed_clicks’, ‘non_bounced_clicks’ and for ‘sov_short_ttc’, ‘sov_short_ttc_global’, sov_short_ttc_score
- Missing data:
    - webview_height, webview_width: autofill in 250 and 300
    - clicks, closing_event, landed_clicks, non_bounced_clicks: : input with 0 for absence of these actions

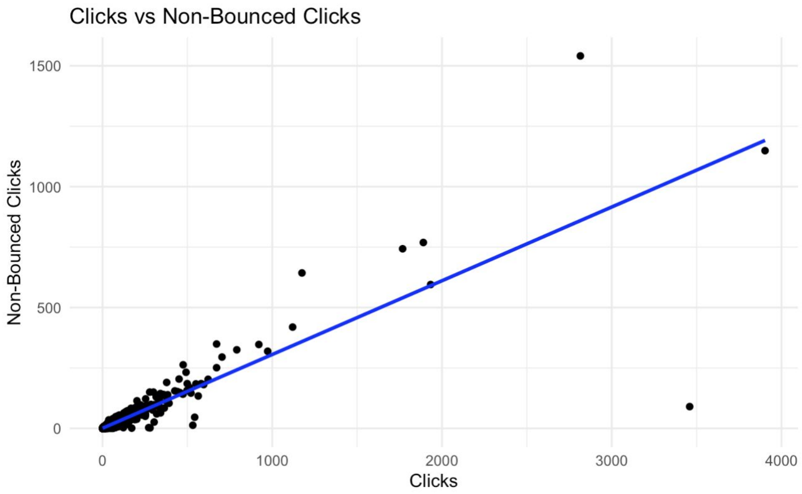
```
                   X                 domain              grid_id
                   0                      0                    0
      webview_height          webview_width             displays
                1076                   1076                    0
              clicks          landed_clicks   non_bounced_clicks
                1047                   1047                 1047
      closing_events avg_last_second_framerate     sov_short_ttc
                3092                   1076                 2953
 sov_short_ttc_global    sov_short_ttc_score
                2953                   2953
```

Metric Dataset: Click Analysis



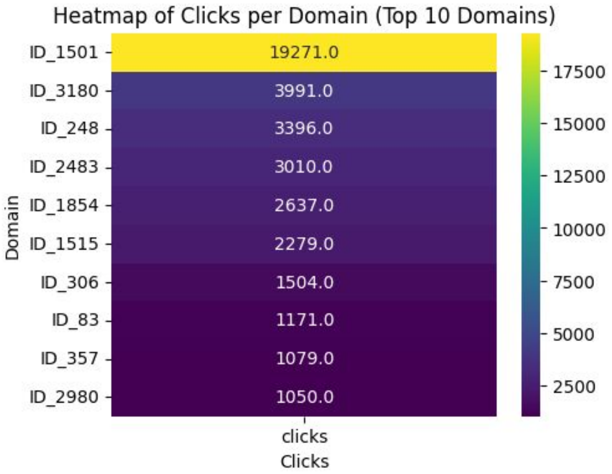
- Relationship between the number of clicks and displays.
- Filtered data of clicks < 5,000
- Related to displays, there is only relatively low number of clicks.

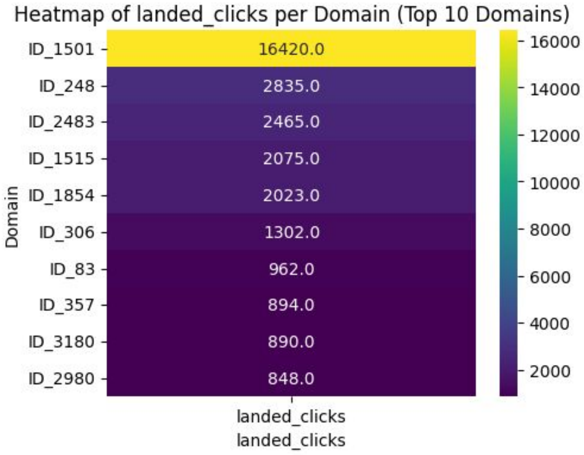
Metric Dataset: Click Analysis

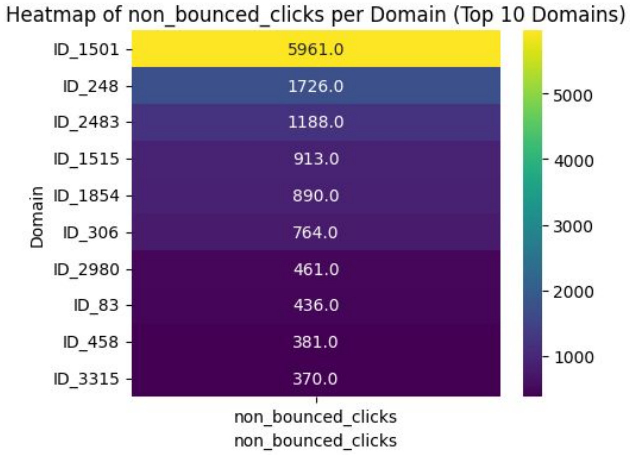


- Relationship between the number of clicks and the number of non bounced clicks.
- Filtered data of clicks < 5,000, displays < 10,000,000
- Around ⅓ of clicks are non-bounced clicks

Metric Dataset: Clicks Analysis

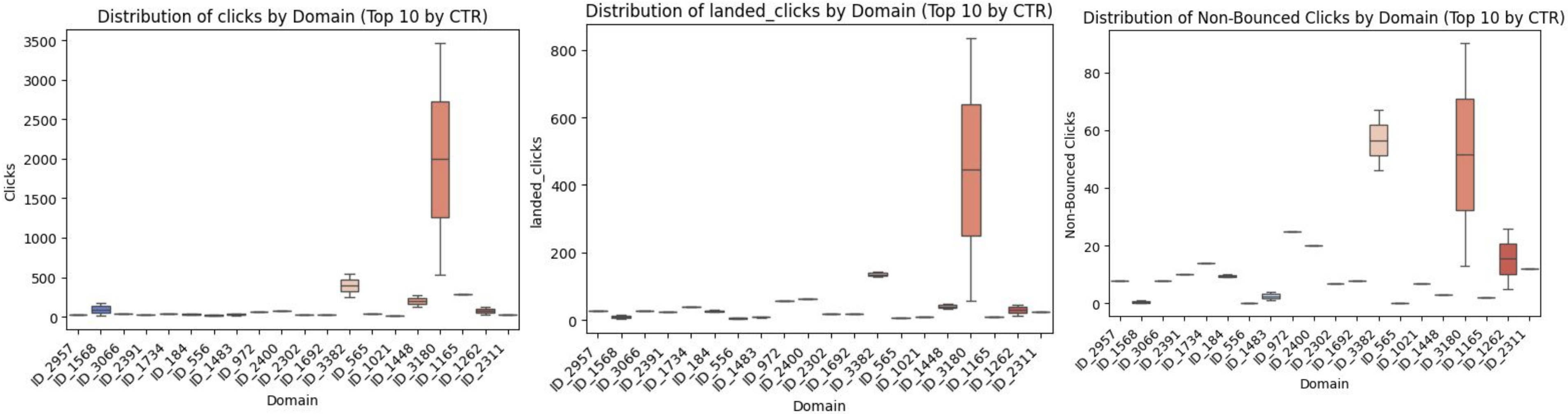






- ID_1501 consistently appears the most across all three categories
- ID_3180 is notable for its presence in the top 2 for clicks, yet it barely makes it into the top 10 for the other two categories.

Metric Dataset: Clicks analysis



Calculate the click-through rate: clicks/display*100
From the distribution of clicks by domain (box plot), we can see the domain ID_3180 has a notable distribution that the clicks range from 3500 to 500 compare to other domain

Next steps:

- Create baselines:
    - Bootstrap approach (WIP):
        - Create empirical distribution, p(x,y), function over x-y-grid for grid_id
        - Compare domain sample with p(x,y) using chi-squared (same distribution?) → reduction of ½
        - Compare probability of creating domain sample → anomalies have small probability → reduction to ⅓
        - Data enhancement using bootstrapping + noise → to do.
    - K-NN
    - K-Means
    - Isolation forest

- Understand Performance metric dataset and engineer features → Combine with Heatmap dataset for comprehensive analysis.
- Understand broken banners better → create comprehensive set of broken banners “by hand” (872, 861 different domains per grids).
- Research into SOTA methods/fancier methods: hyperdimensional computing, diffusion(?), other self-supervised approaches?

Technical questions

- Description of datasets, explanation of attributes:
  - displays, clicks, landed_clicks, non_bounced_clicks, closing_events, avg_last_second_framerate, sov_short_ttc, sov_short_ttc_global, sov_short_ttc_score? -> already discussed
  - How would these data affect on the anomaly detection, specifically on sov_short_ttc related data?
- <u>Exact problem statement</u>:
  - How should we generate the ground truth labels?
  - What is the argument against rules-based for classification?
  - What is the baseline upon which we should improve?
- Hardware constraints?
  - How will this model be applied in practice?

Administrative questions:

- Would it be useful to create joint Slack channel?
- Zoom account for weekly meetings?