Houzz data challenge

Instructions:

* You have **3** hours to complete this exercise as much as possible
* Please work on this challenge independently and do not share with others
* This is purely a coding challenge, so please don’t spend extra time doing any statistical tests/models for any questions
* Feel free to use any open source tools to solve this – R, Python. You could also use SQL within these tools to solve this, using packages like sqldf or pandasql (or any one of your choice)
* Make assumptions wherever necessary, and clearly call them out
* Create graphs where necessary
* Your final output can be a doc file or a presentation with 1 – 2 slides for each question. We do include the raw code as well in a separate file. Rmd or Jupyter outputs are also appreciated!

Data:

Houzz sends email newsletters to our email subscribers promoting current sales. Here is the metadata:

* newsletters.csv
  + ts – timestamp (epoch time in seconds)
  + user\_id – Identifier for each user
  + nl\_id – This is the id of the houzz email newsletters. It is unique at the newsletter batch level, not user level.
  + dt – date
  + hr – hour
  + event\_type – there are 3 types:
    - nlsent: a newsletter email was sent
    - nlpv: a newsletter email was viewed
    - nllc: newsletter link clicks (user clicked on clickable content in the newsletter and was directed to relevant content on the Houzz website)
  + event\_type\_param – extra parameters for nllc/newsletter link positions. Each newsletter contains multiple sale banners.
    - g0: clicks on the first sale banner in the newsletter
    - g1: clicks on the the 2nd sale banner in the newsletter
    - g2: clicks on the the 3rd sale banner in the newsletter
    - …
    - Please ignore event\_type\_param for nlsent/newsletter sends and nlpv/newsletter opens

Questions:

* Define open\_rate as: # unique users who opened a newsletter / # unique users who received a newsletter.
* Define CTR as: # clicks at a position / # opens. Example:

User 1: opened once, clicked on position 1 and 3

User 2: opened twice, clicked once on position 1

User 3: didn’t open

Based on these 3 users,

CTR\_position\_1 = 2/3, CTR\_position\_2 = 0, CTR\_position\_3 = 1/3

1. How many newsletters were sent vs. opened for nl\_id 2885 and 2912? What’s the overall open rate for each newsletter?
2. What % of users opened the email within 1, 2, 3, 4, 5, 6, 7 days? Visualize the results for each newsletter. Which newsletter has the best open\_rate?
3. Make a graph of the CTRs by link position for nl\_id 2873 and 2885. Visualize the results for each newsletter.
4. What other metrics we can use to measure newsletter performance/quality? For each metric please state why it is important in one sentence.