HW1 - Problem 2

Naveen Nagarajan

2/15/2021

Problem 2

The datasets provided nyt1.csv, nyt2.csv, and nyt3.csv represents three (simulated) days of ads shown and clicks recorded on the New York Times homepage. Each row represents a single user. There are 5 columns: age, gender (0=female, 1=male), number impressions, number clicks, and logged-in. Use R to handle this data. Perform some exploratory data analysis:

- Create a new variable, age_group, that categorizes users as "<20", "20-29", "30-39", "40-49", "50-59", "60-69", and "70+".
- For each day:
  - Plot the distribution of number of impressions and click-through-rate (CTR = #clicks / #impressions) for these age categories
  - Define a new variable to segment or categorize users based on their click behavior.
  - Explore the data and make visual and quantitative comparisons across user segments/demographics (<20-year-old males versus <20-year-old females or logged-in versus not, for example).
- Extend your analysis across days. Visualize some metrics and distributions over time.

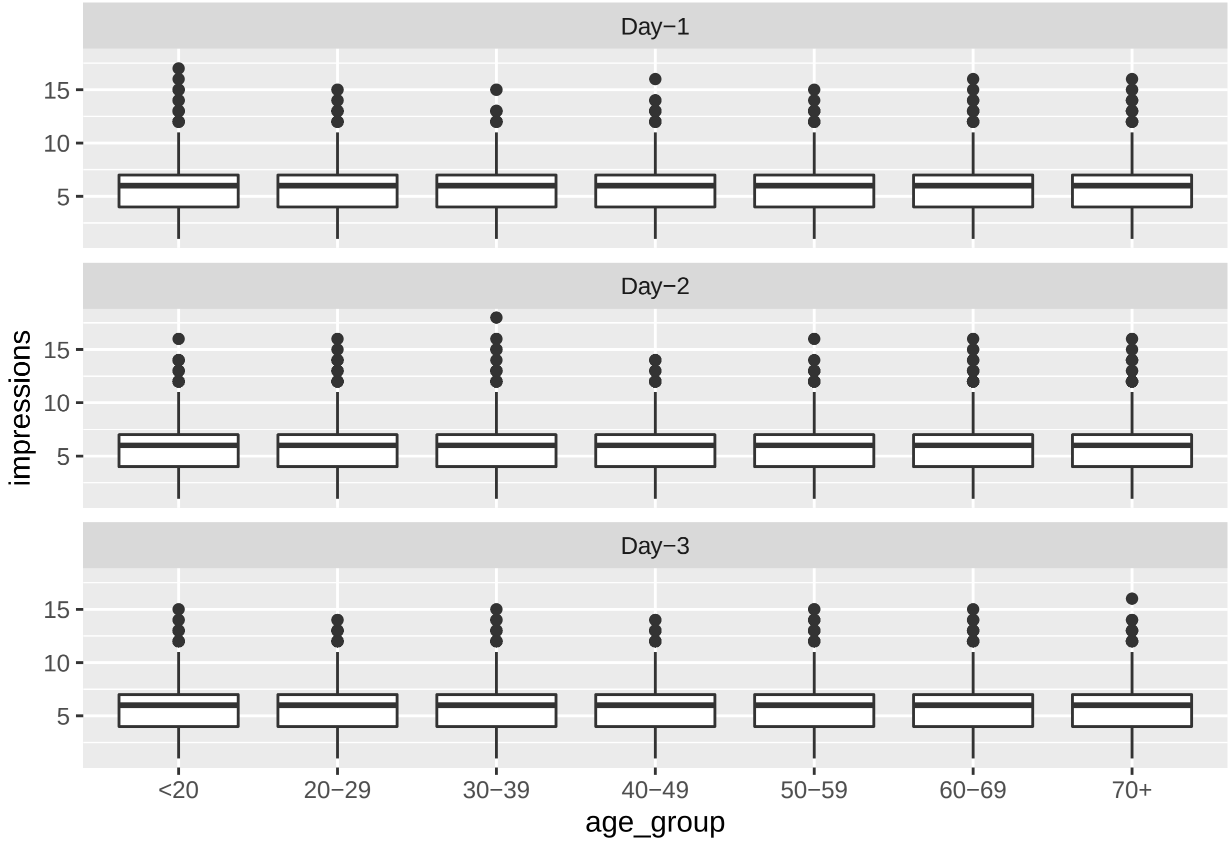
Solution

- Summary

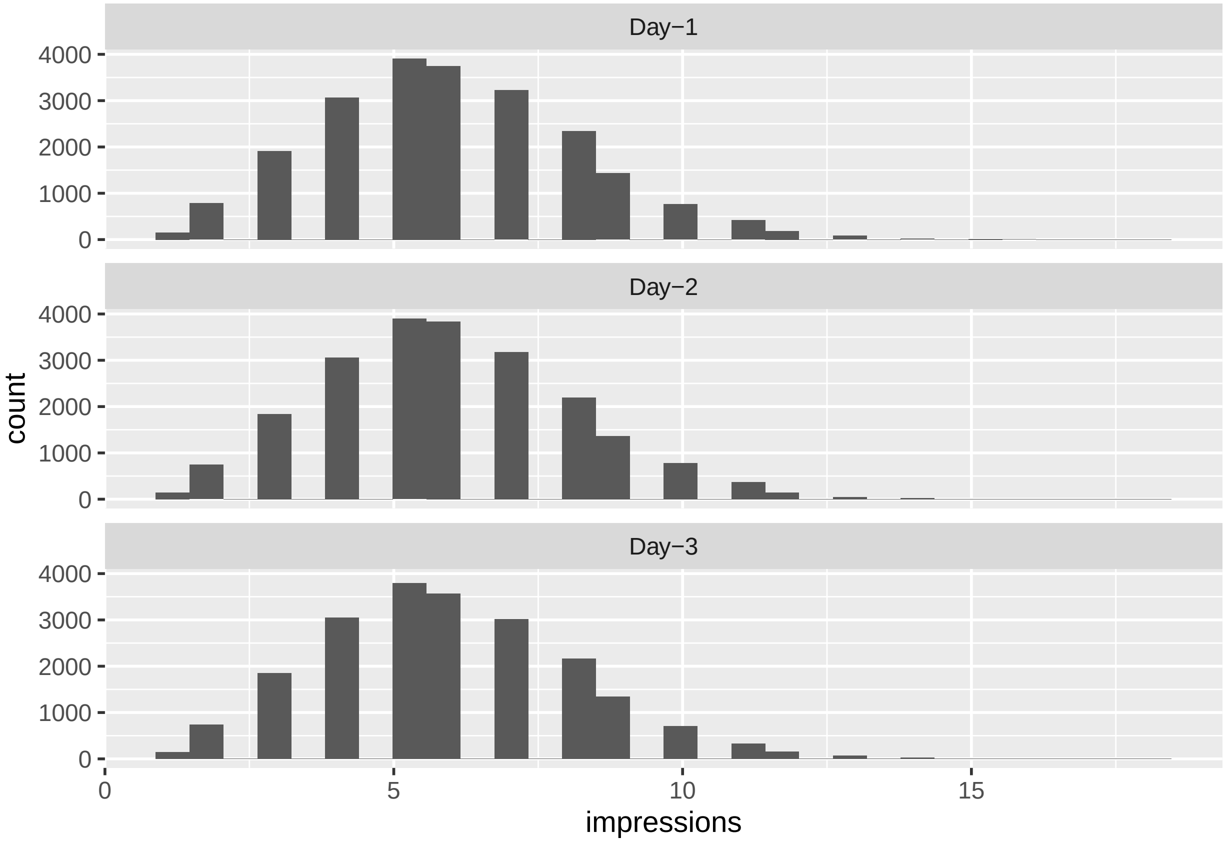
Day	Count
Day1	458441
Day2	449935
Day3	440370
Total	1348746

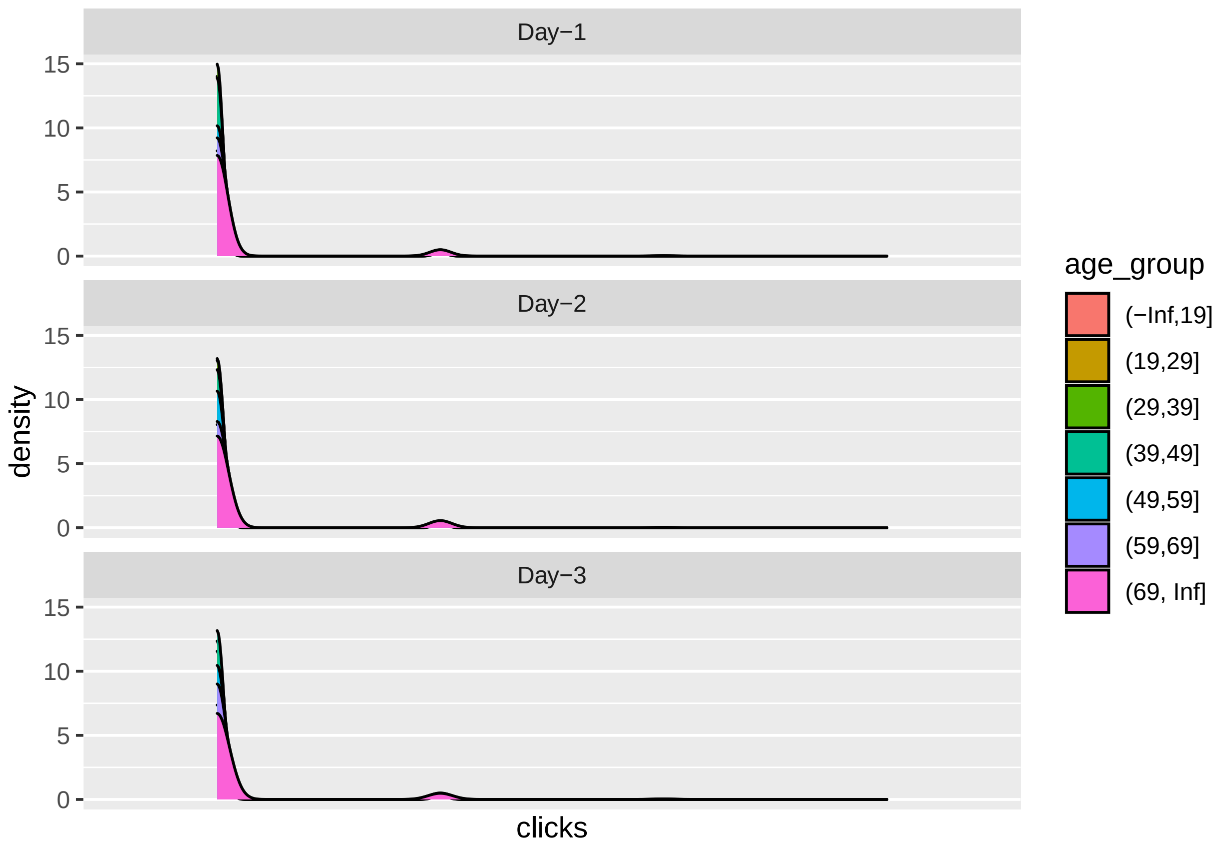
```
##       age             gender        impressions         clicks
##  Min.   :  0.00   Min.   :0.0000   Min.   : 0.000   Min.   :0.00000
##  1st Qu.:  0.00   1st Qu.:0.0000   1st Qu.: 3.000   1st Qu.:0.00000
##  Median : 31.00   Median :0.0000   Median : 5.000   Median :0.00000
##  Mean   : 29.49   Mean   :0.3694   Mean   : 5.001   Mean   :0.09255
##  3rd Qu.: 48.00   3rd Qu.:1.0000   3rd Qu.: 6.000   3rd Qu.:0.00000
##  Max.   :111.00   Max.   :1.0000   Max.   :20.000   Max.   :6.00000
##    signed_in          day
##  Min.   :0.0000   Length:1348746
##  1st Qu.:0.0000   Class :character
##  Median :1.0000   Mode  :character
##  Mean   :0.7006
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

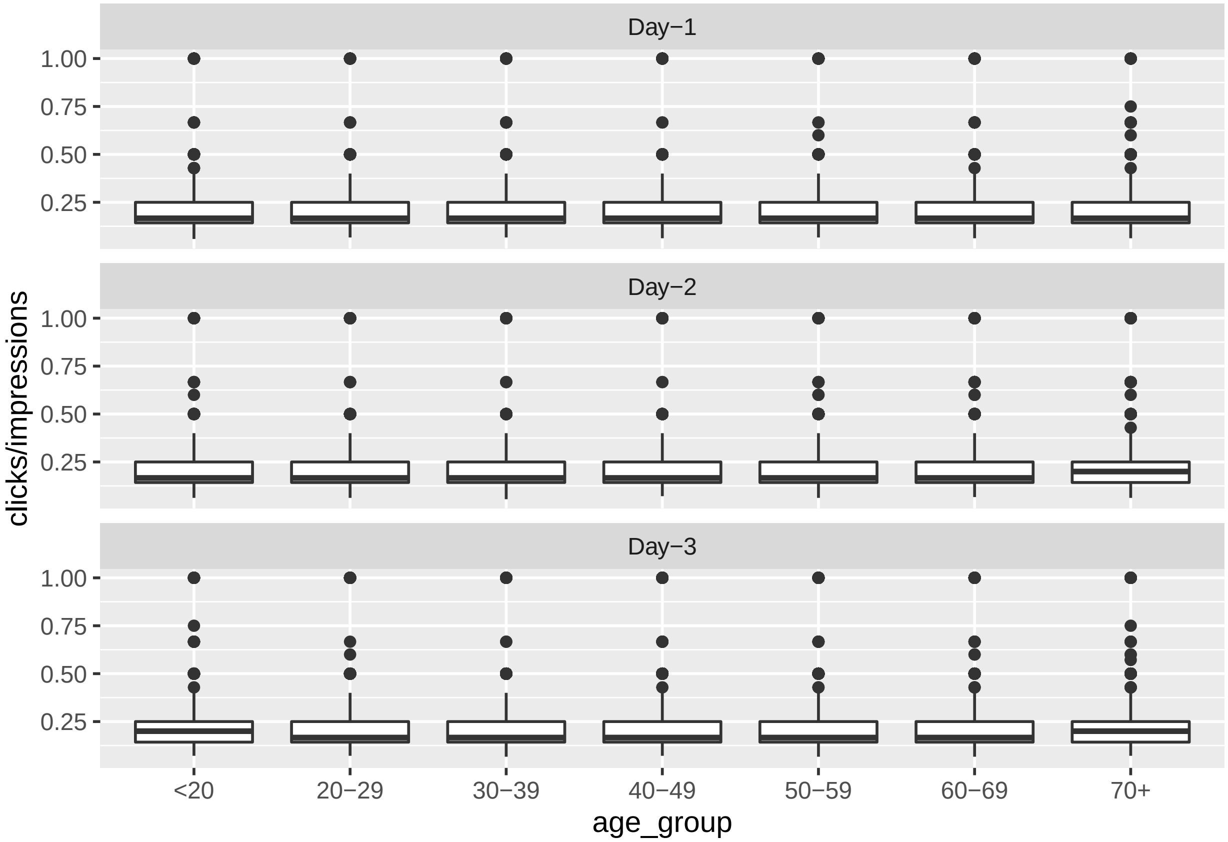
- Create category age group, factor days and rename gender
- Distribution of impressions and CTR for age categories



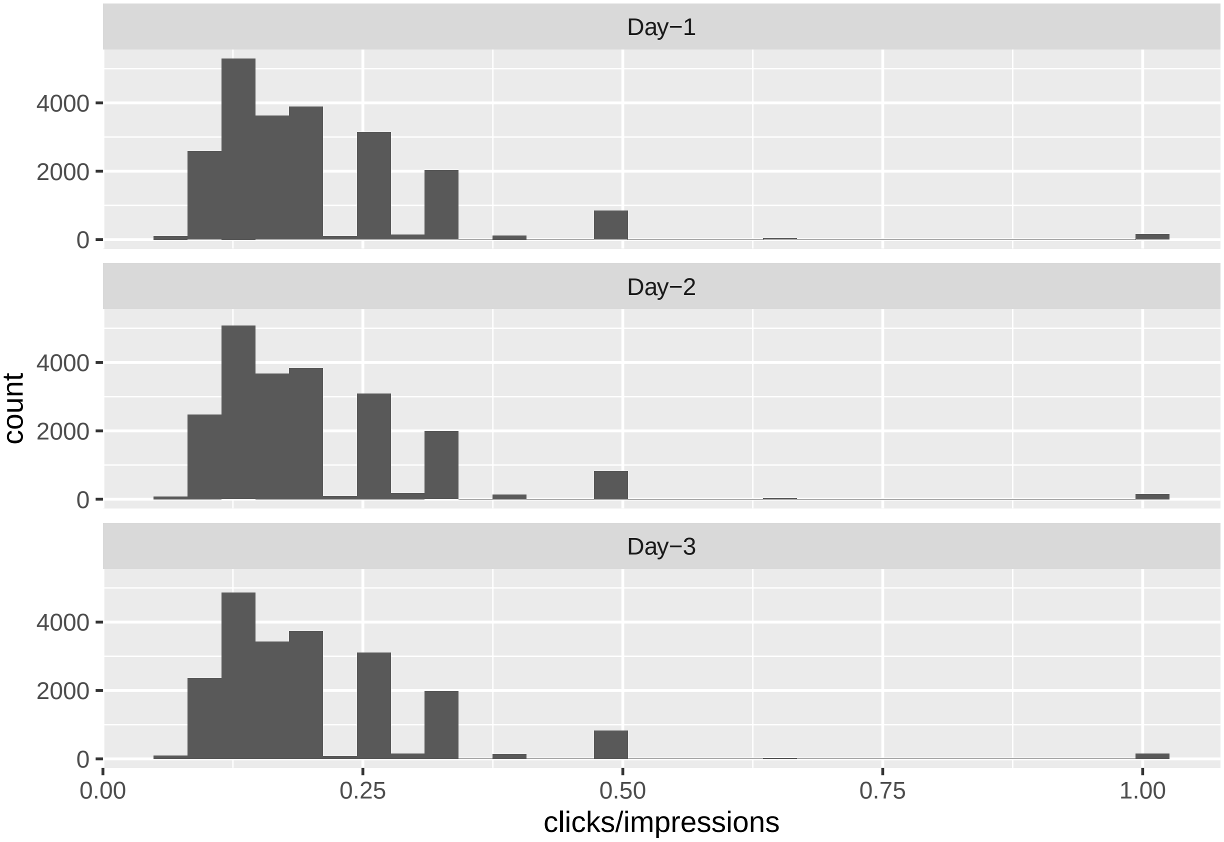
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

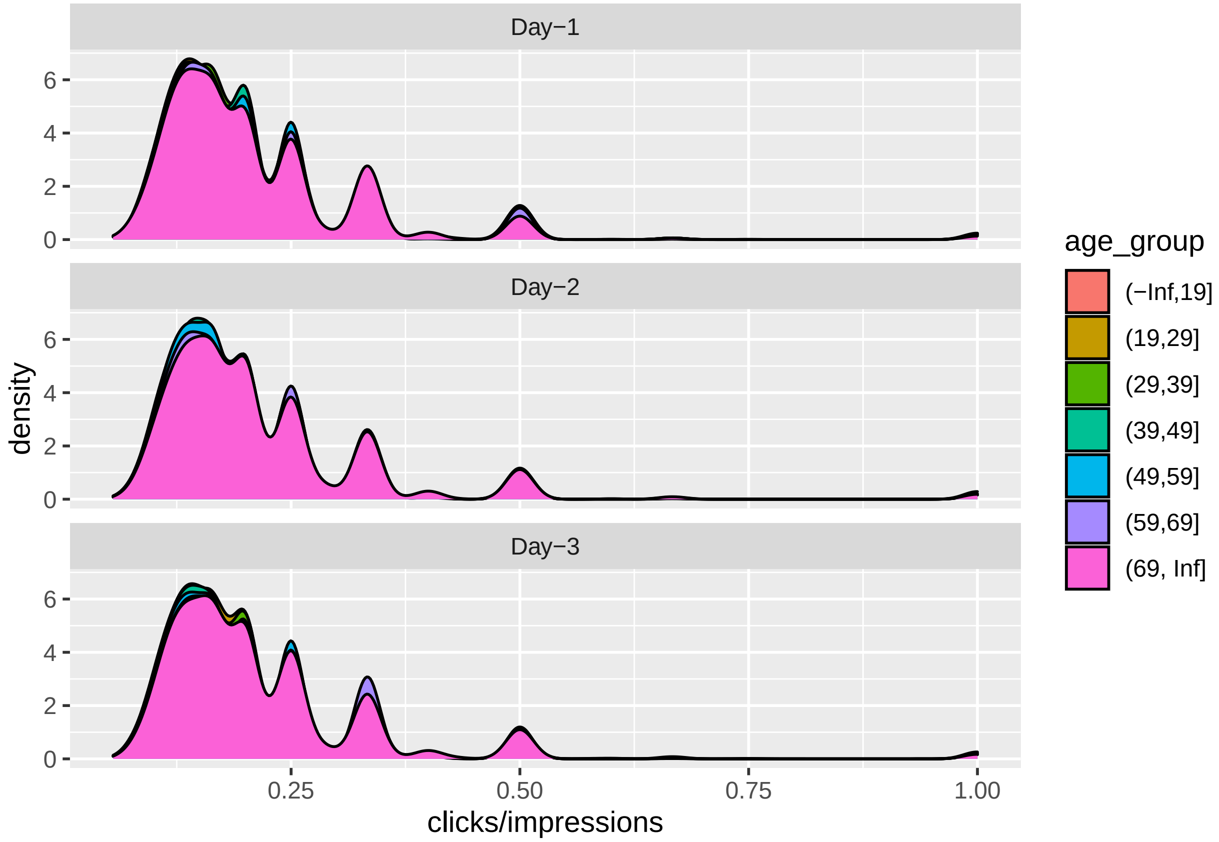

Day−1
Day−2
Day−3
count
4000
3000
2000
1000
0
0
5
10
15
impressions


Day−1
Day−2
Day−3
15
10
5
0
density
clicks
age_group
(−Inf,19]
(19,29]
(29,39]
(39,49]
(49,59]
(59,69]
(69, Inf]

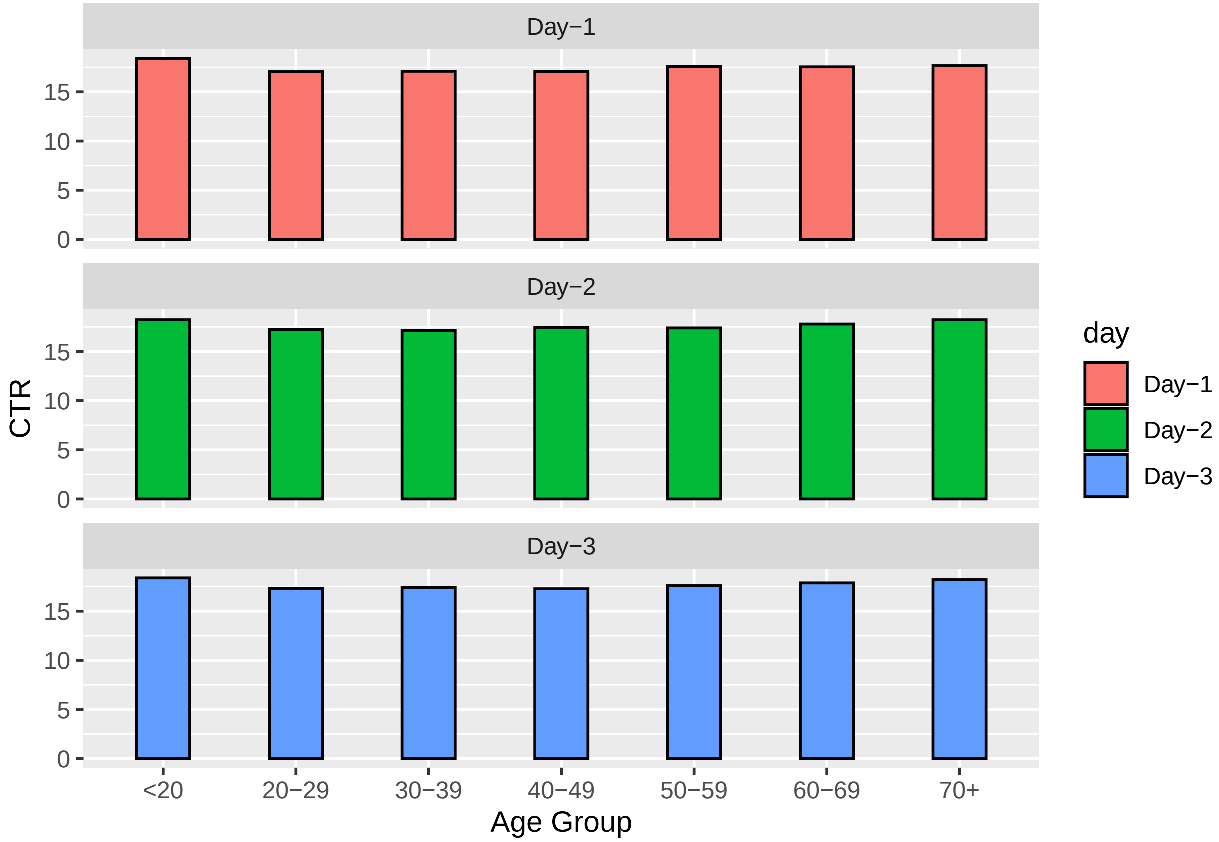


```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


Day−1
Day−2
Day−3
count
4000
2000
0
0.00
0.25
0.50
0.75
1.00
clicks/impressions



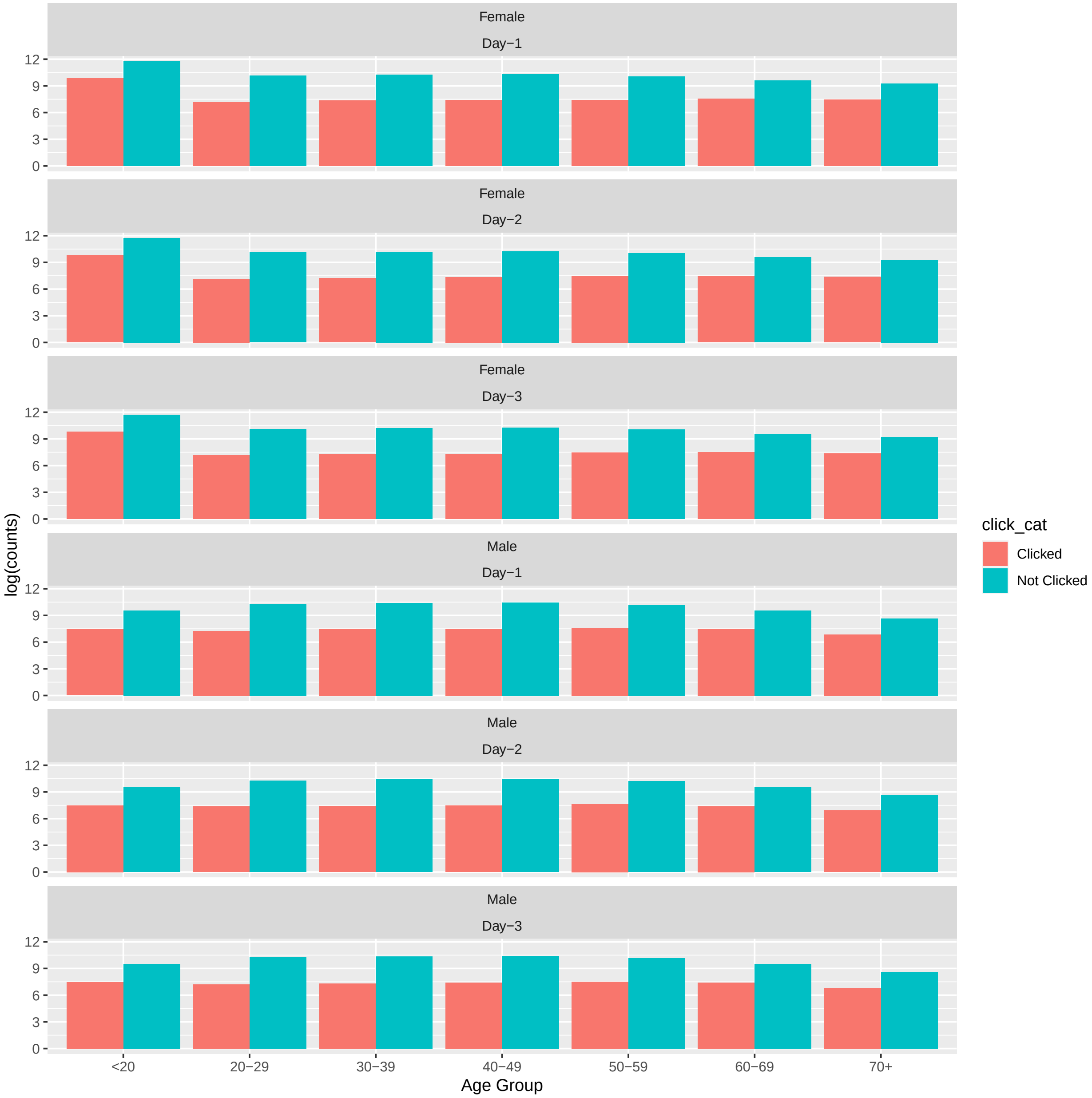
```
## ‘summarise()‘ has grouped output by ’day’. You can override using the ‘.groups‘ argument.
```



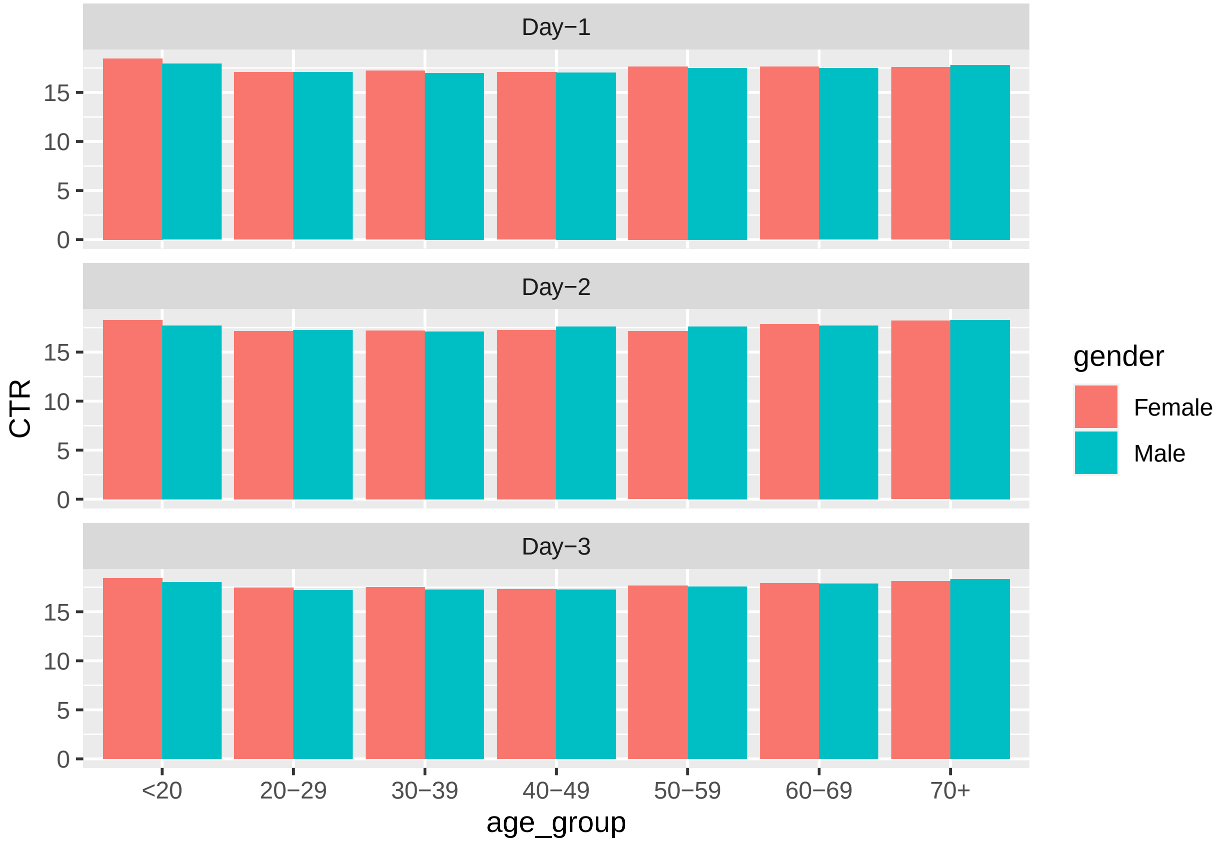
- Categorize based on Clicked, Not Clicked

```
##     Clicked Not Clicked
##      117143     1231603
```

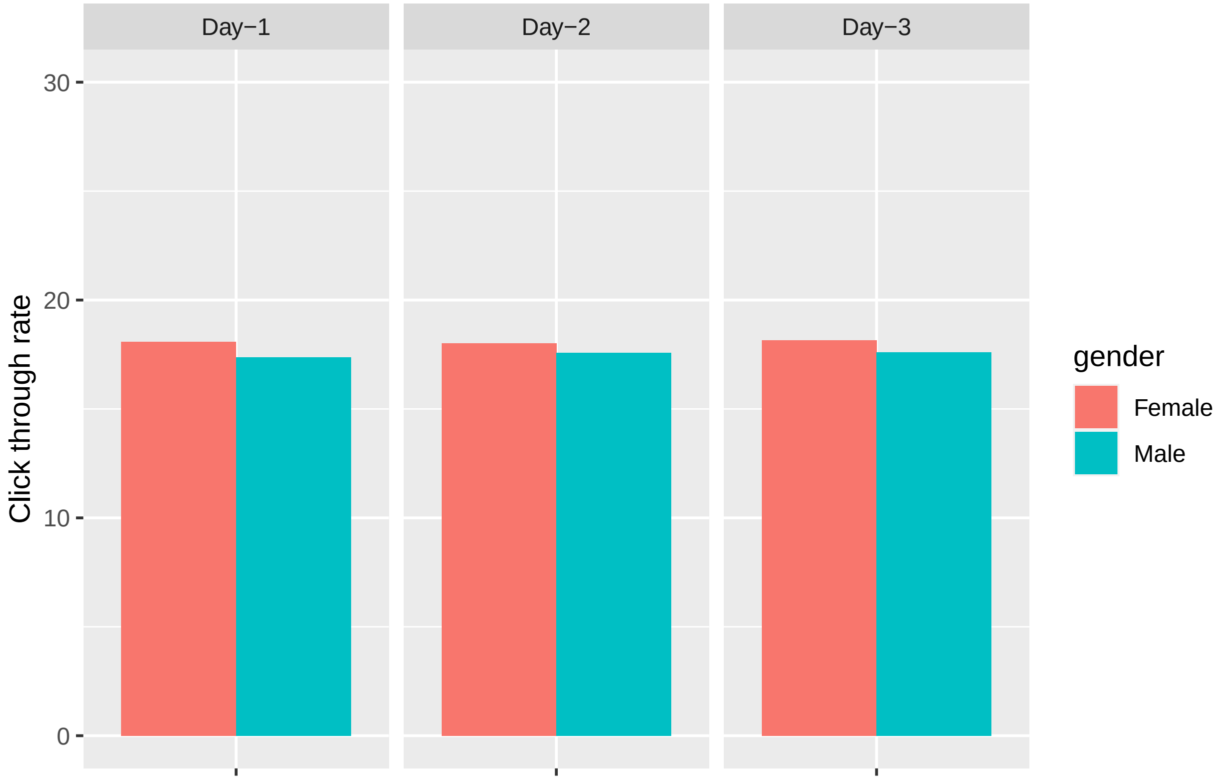
- Quantitative comparison across segments/demo



```
## ‘summarise()‘ has grouped output by ’day’, ’age_group’. You can override using the ‘.groups‘ argumen
```

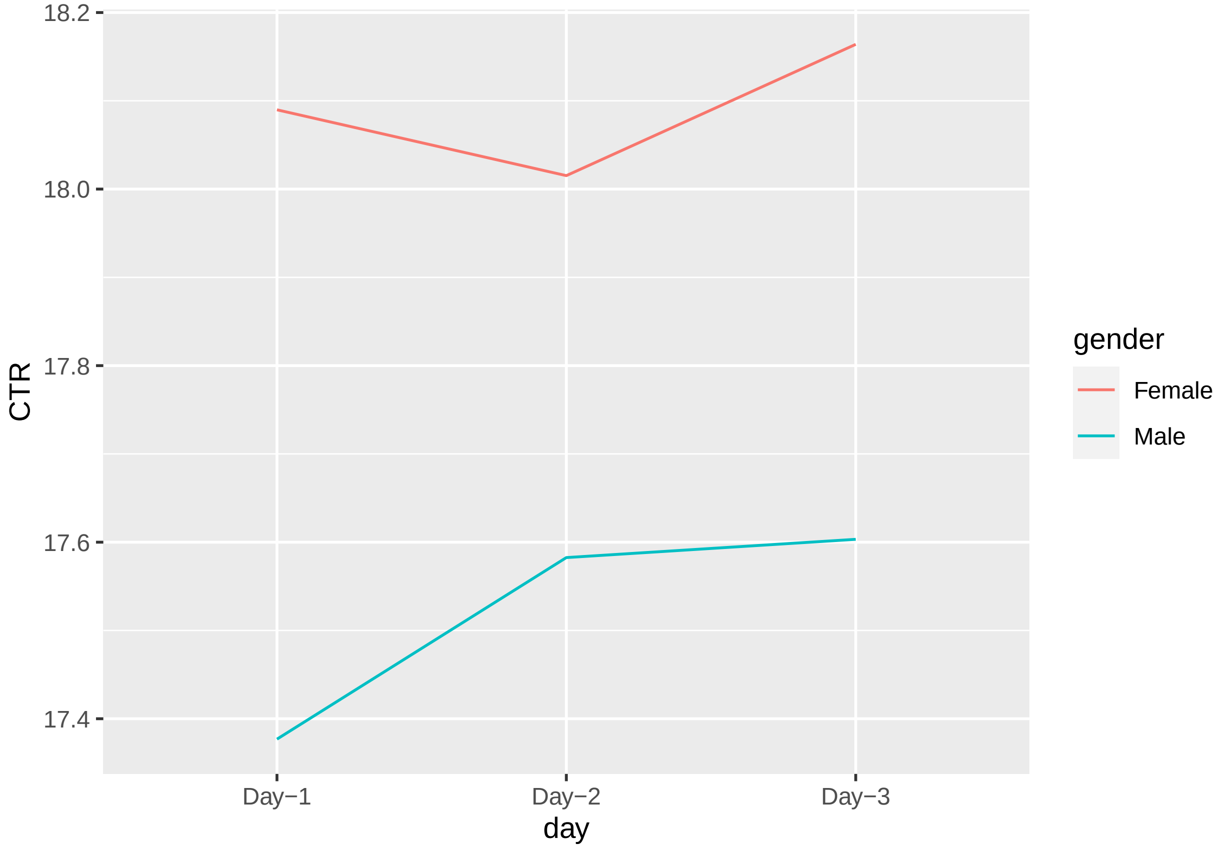


```
## `summarise()` has grouped output by 'day'. You can override using the `.groups` argument.
```

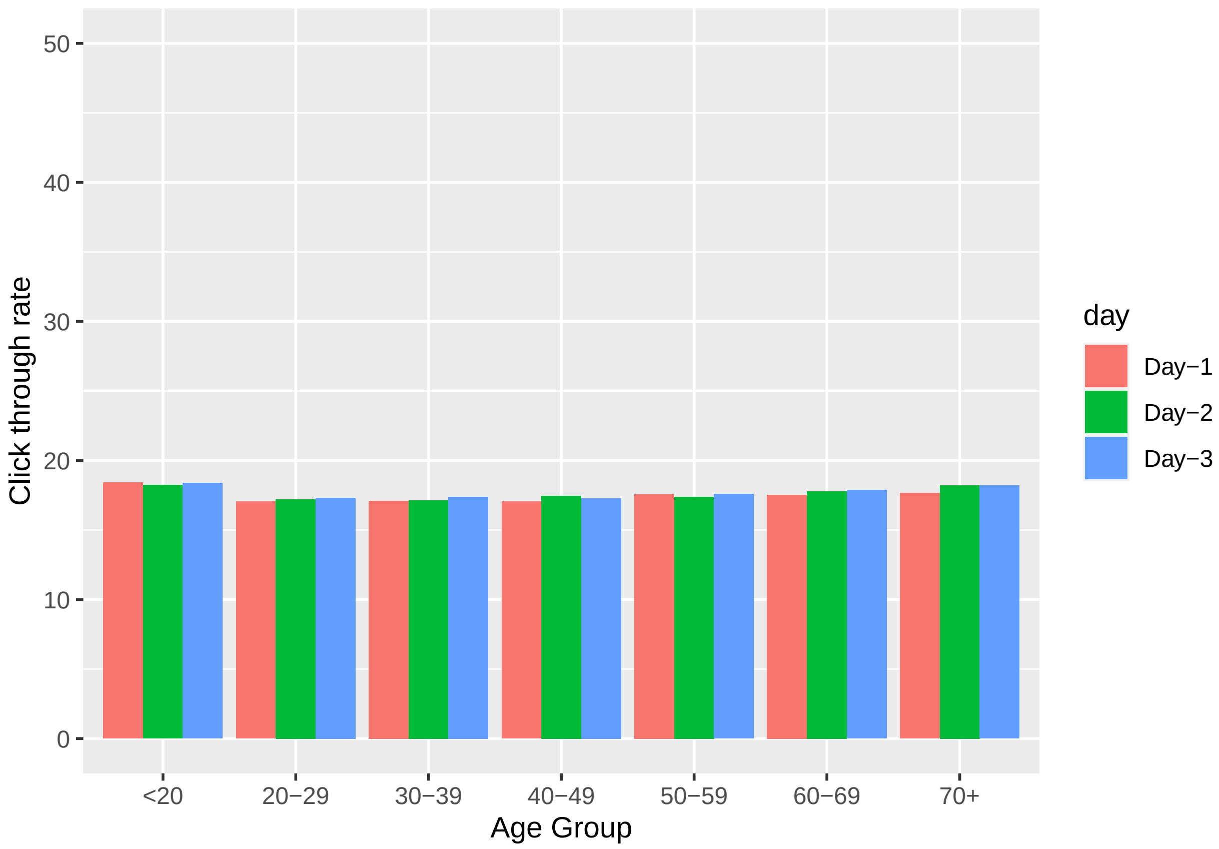


- Extend analysis across days

```
## ‘summarise()‘ has grouped output by ’day’. You can override using the ‘.groups‘ argument.
```



```
## ‘summarise()‘ has grouped output by ’day’. You can override using the ‘.groups‘ argument.
```


Click through rate
50
40
30
20
10
0
<20
20−29
30−39
40−49
50−59
60−69
70+
Age Group
day
Day−1
Day−2
Day−3