Ginger EDA

Matthew Cui

10/8/2020

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
coach <- read_csv("coach_data.csv") %>%
  rename(num_msg = `Number of messages per week`)
```

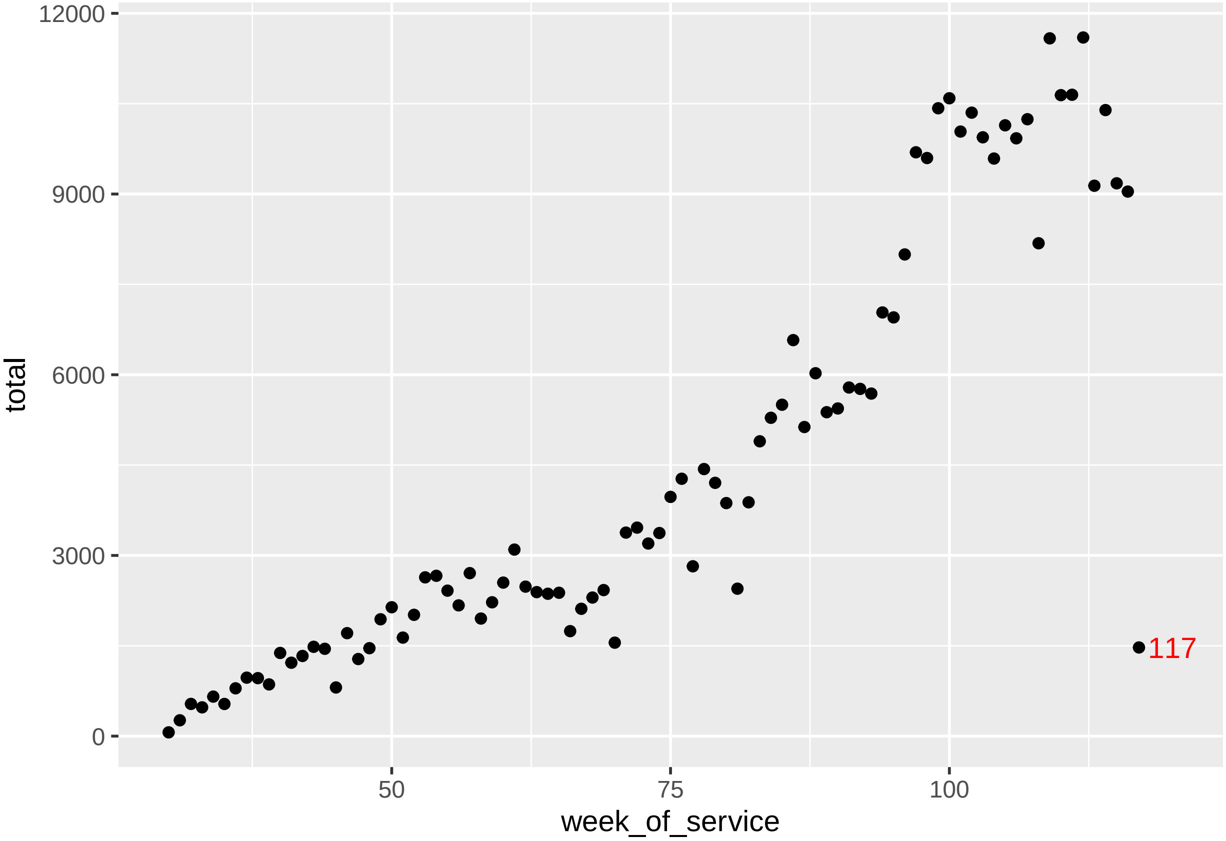
```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   hashed_member_id = col_character(),
##   week_of_service = col_double(),
##   `Number of messages per week` = col_double()
## )
```

```
counts <- coach %>%
  group_by(week_of_service) %>%
  count(num_msg) %>%
  mutate(total = sum(n * num_msg)) %>%
  distinct(total)
counts
```

```
## # A tibble: 88 x 2
## # Groups:   week_of_service [88]
##    week_of_service total
##              <dbl> <dbl>
##  1              30    63
##  2              31   263
##  3              32   535
##  4              33   478
##  5              34   656
##  6              35   534
##  7              36   794
##  8              37   971
##  9              38   963
## 10              39   859
## # ... with 78 more rows
```

```
ggplot(counts, aes(x = week_of_service, y = total)) +
  geom_point() +
  geom_text(aes(label= ifelse(week_of_service == 117,
                             as.character(week_of_service), "")),
                nudge_x = 3,
                color = "red")
```

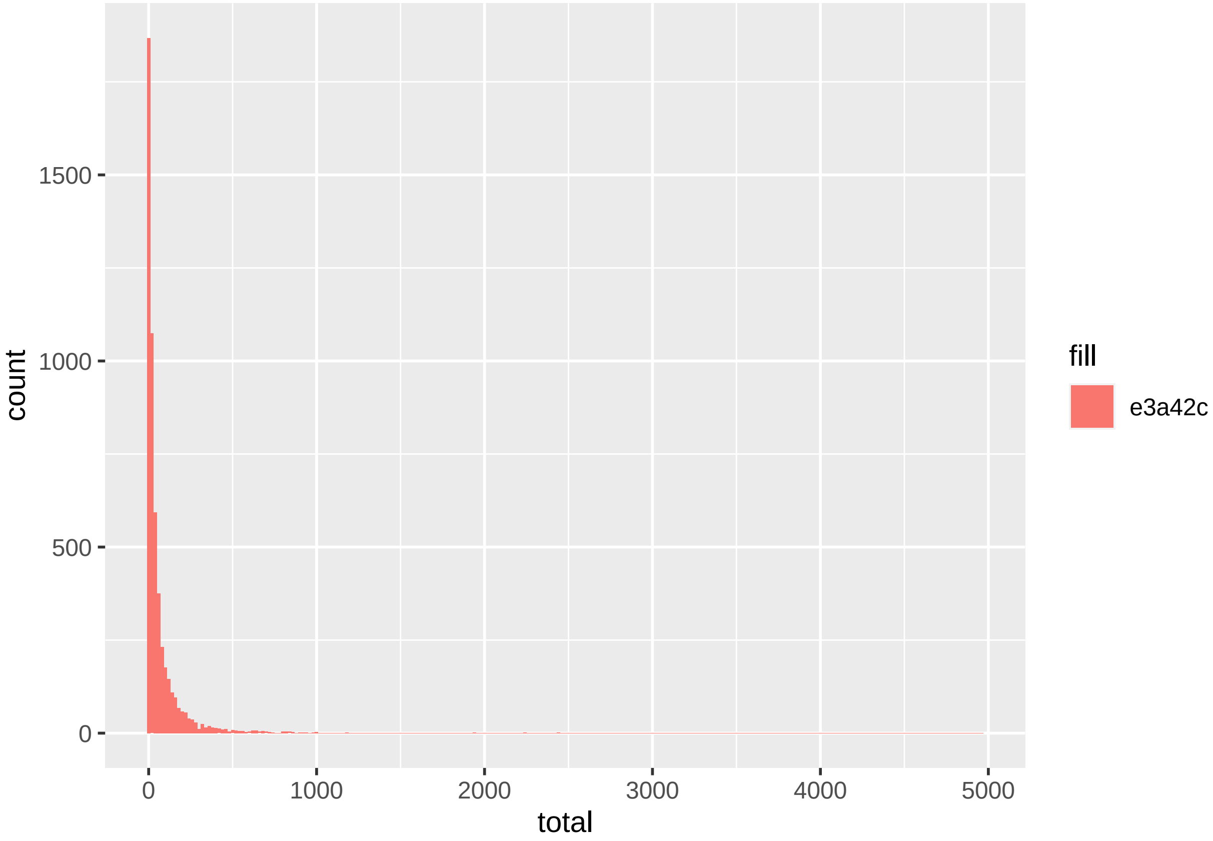


```
actives <- coach %>%
  group_by(hashed_member_id) %>%
  count(num_msg) %>%
  mutate(total = sum(n * num_msg)) %>%
  distinct(total) %>%
  arrange(desc(total))
actives
```

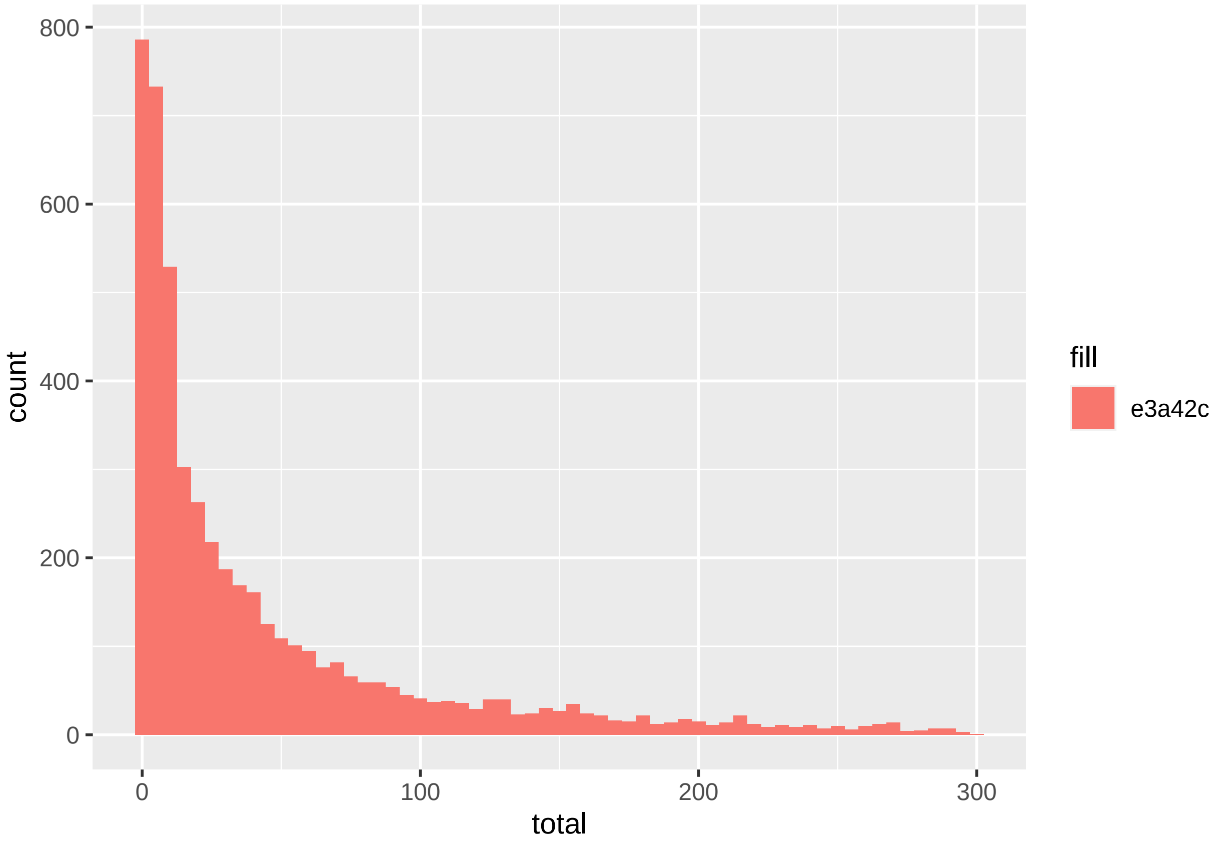
```
## # A tibble: 5,224 x 2
## # Groups:   hashed_member_id [5,224]
##    hashed_member_id                                                 total
##    <chr>                                                            <dbl>
##  1 59aa0fd91f8b1360dc0b2c0d6c0f318871d9841a52f95a4cd60ddff7022c5acb  4952
##  2 3a43343c99f7da36168915a92f100157045e553cb63039987fe3714302b3e5c2  3472
##  3 958e6c7babfcbfd60631dcb5cde72d447e1bb270937bccb517fbd6ea48bc8325  2633
##  4 74fc94c43f1a69b6a674b797b0d96bf1591fedd18a6eb6ce4bf9c30056dfec53  2565
##  5 cab986efaaaf5d2593c8b79c22d2fb1e9767f36588b40d6abf2cb242997a2bc1  2436
##  6 682026a92521ef5d017500cbdb67b7f0f30f1a6c831104e578c8c3e8e7e00f38  2434
##  7 3cf2e4e402cde10ce2a7bf0645859a788a3cb7af21b397612b2bb8ceac83bee0  2321
##  8 923ebea5206a91229ceda996cee3d7a2603d5200669ce4a9fb1c5ad07358c08d  2247
##  9 cb78d540ea4ca173ef14ca101d7b4b19960604517eba60bc3a9dbc9ce3d7fd18  2245
## 10 2bbb8cdeaafb6e491a605351c17916f4dce13ecf261c26c34f93a24b716c22fa  2182
## # ... with 5,214 more rows
```

```
top5_member <- actives %>%
  head(5) %>%
  pull(hashed_member_id)

ggplot(actives, aes(x = total)) +
  geom_histogram(binwidth = 20, aes(fill = "e3a42c"))
```

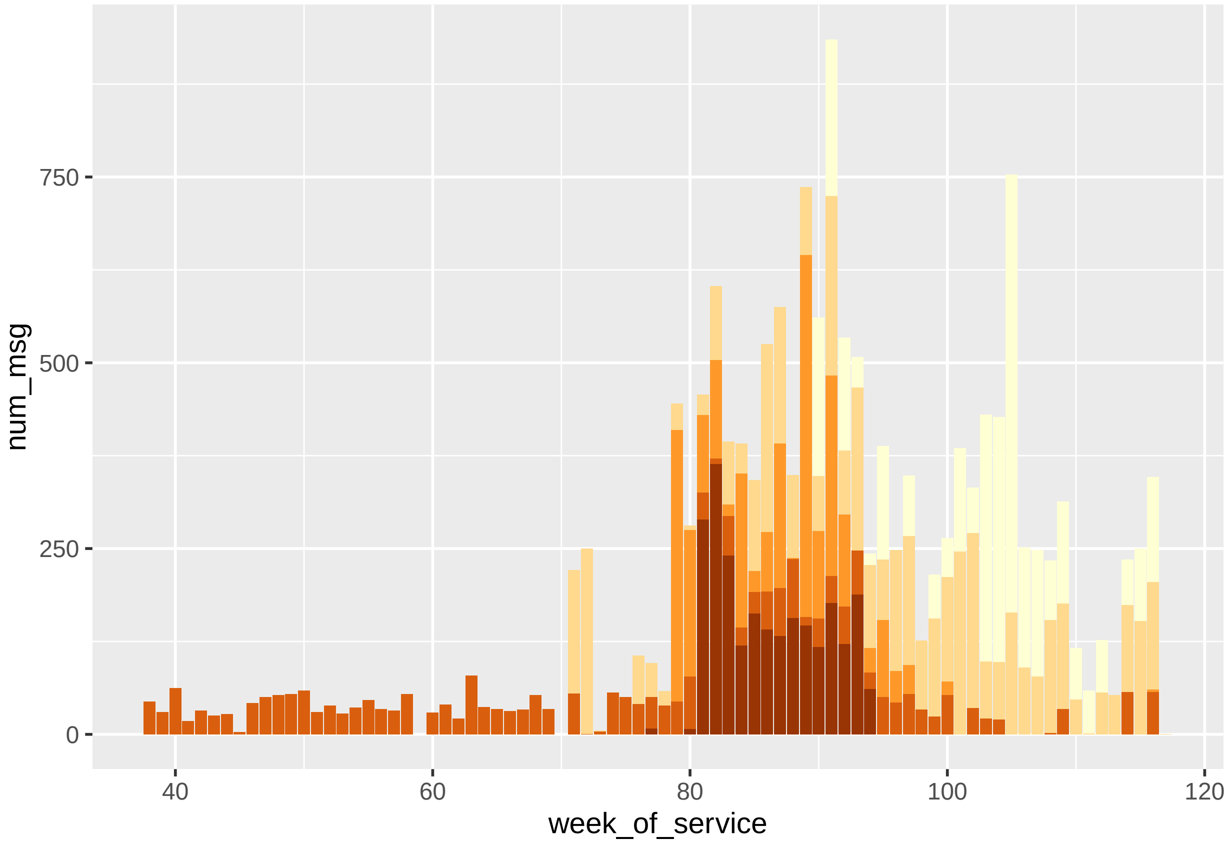


```
actives %>%
  filter(total < 300) %>%
  ggplot(aes(x = total)) +
    geom_histogram(binwidth = 5, aes(fill = "e3a42c"))
```



```r
top5_activity <- coach %>%
  filter(hashed_member_id %in% top5_member) %>%
  arrange(desc(num_msg))

ggplot(top5_activity, aes(x = week_of_service, y = num_msg)) +
  geom_col(aes(fill = hashed_member_id)) +
  theme(legend.position = "none") +
  scale_fill_brewer(palette = "YlOrBr")
```


num_msg
750
500
250
0
40
60
80
100
120
week_of_service