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()
## )
```

```
clinic <- read_csv("clinical_data.csv")
```

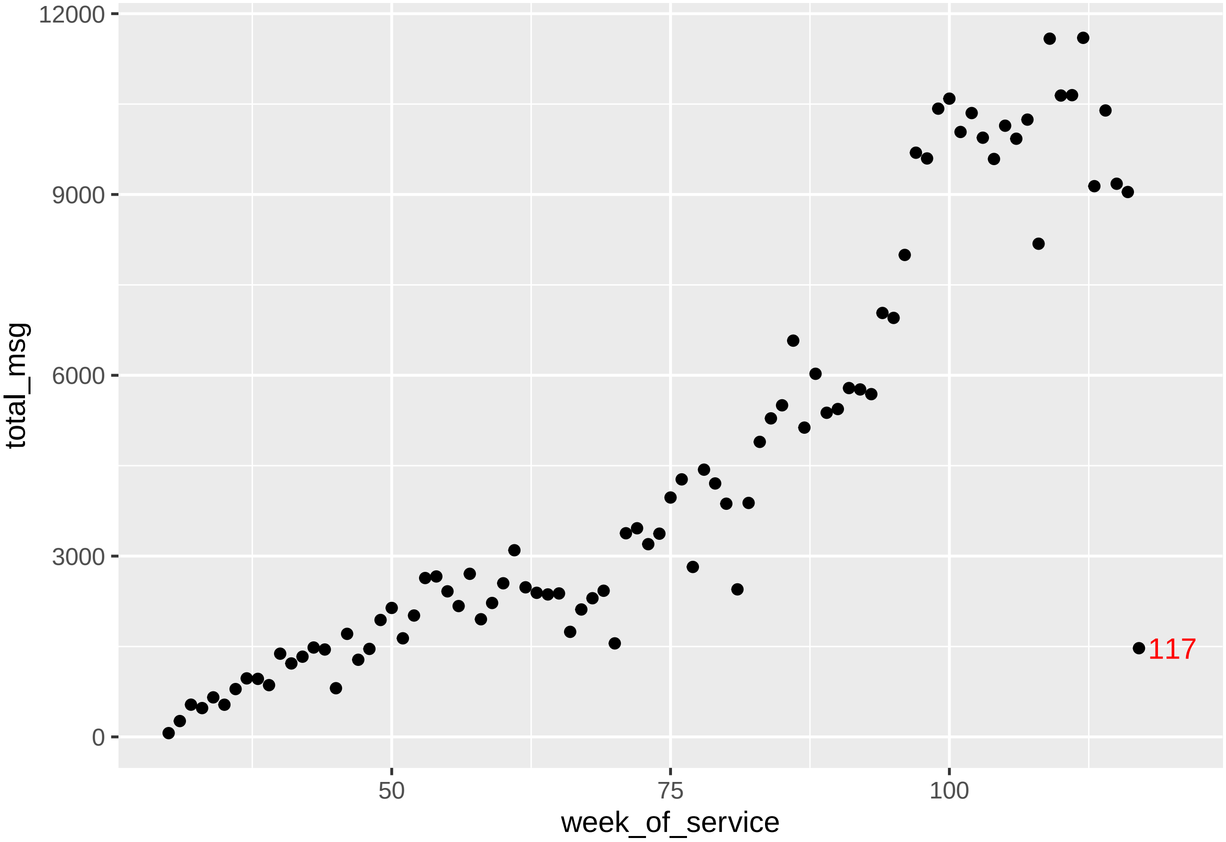
```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   hashed_member_id = col_character(),
##   hashed_clincian_id = col_character(),
##   week_of_service = col_double(),
##   provider_type = col_character(),
##   num_ginger_visits = col_double(),
##   icd_10_codes = col_character()
## )
```

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

```
## # A tibble: 88 x 2
## # Groups:   week_of_service [88]
##    week_of_service total_msg
##               <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
```

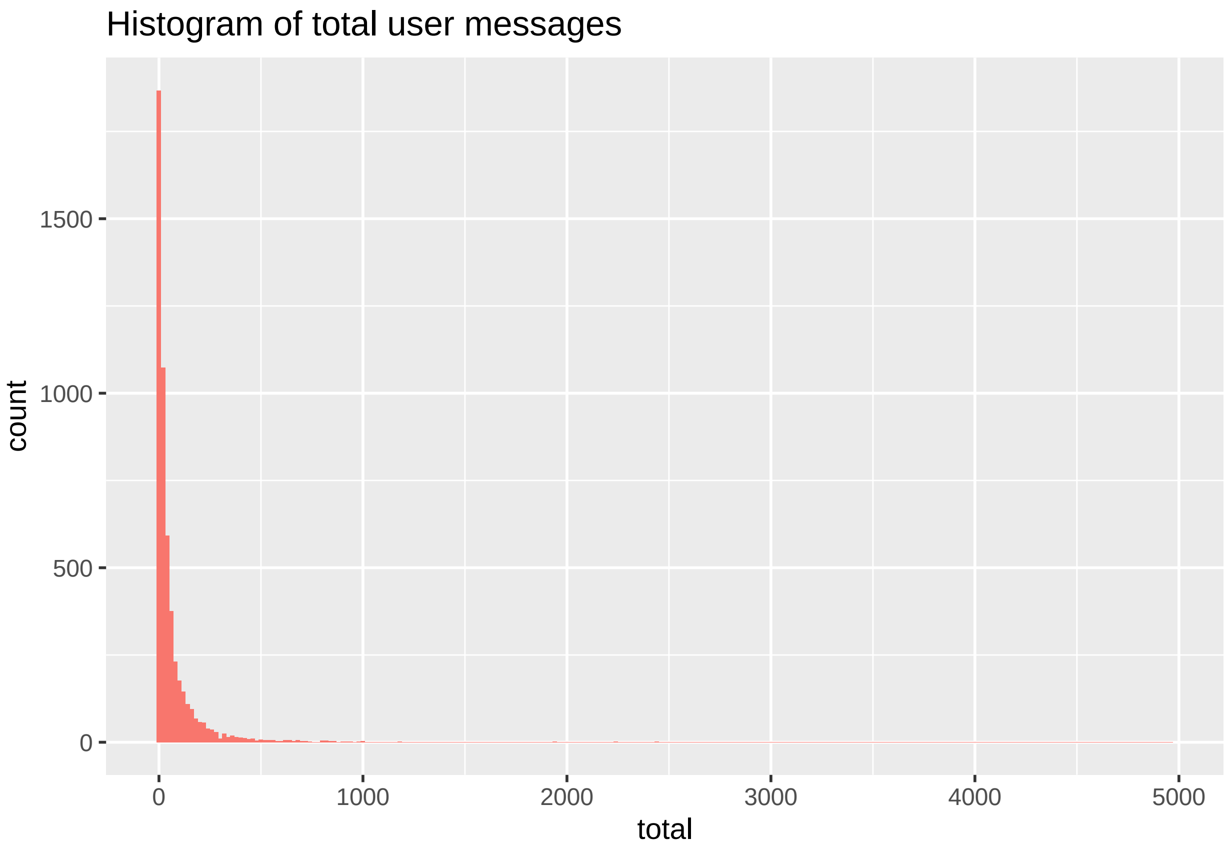
```
ggplot(counts, aes(x = week_of_service, y = total_msg)) +
  geom_point() +
  geom_text(aes(label= ifelse(week_of_service == 117,
                              as.character(week_of_service), "")),
            nudge_x = 3,
            color = "red")
```



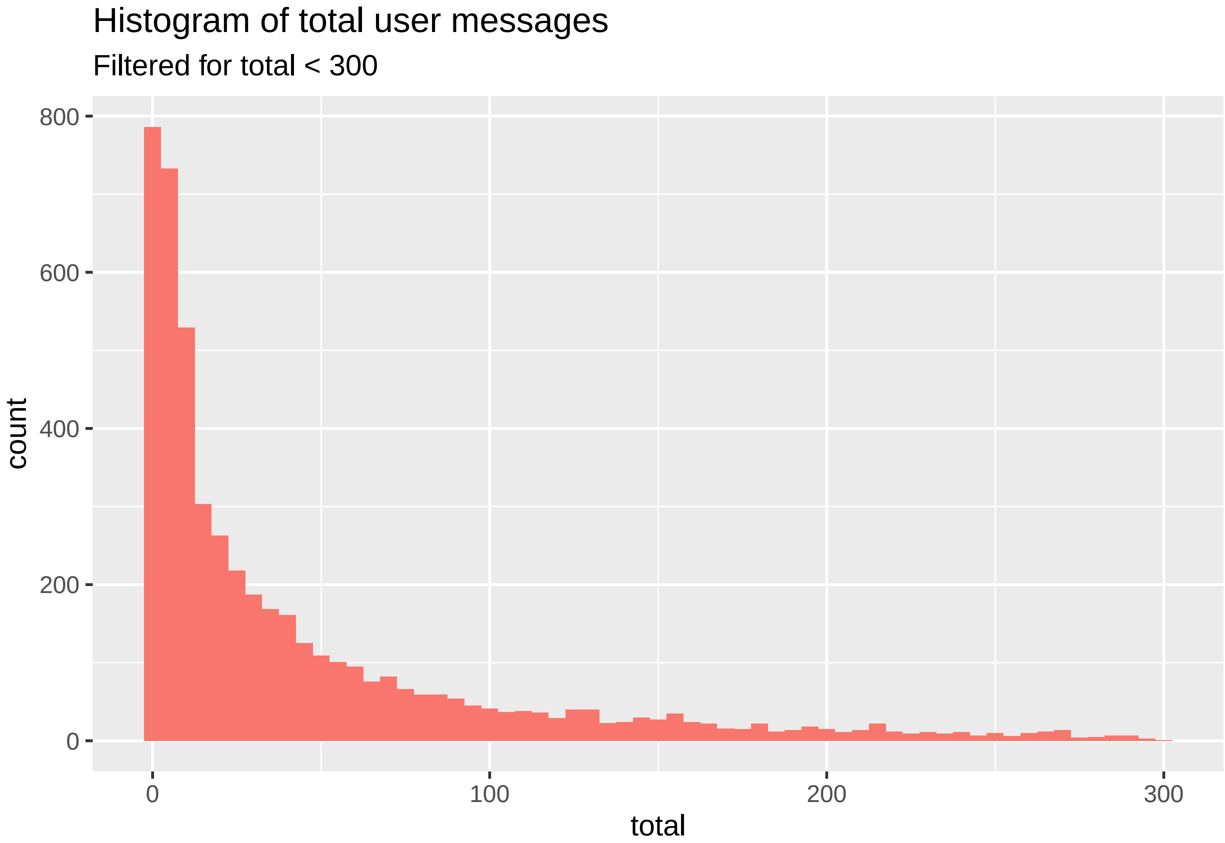
```
actives <- coach %>%
  group_by(hashed_member_id) %>%
  count(num_msg) %>%
  mutate(total = sum(n * num_msg)) %>%
  distinct(total) %>%
  arrange(desc(total))
actives
```

```
## # 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
```

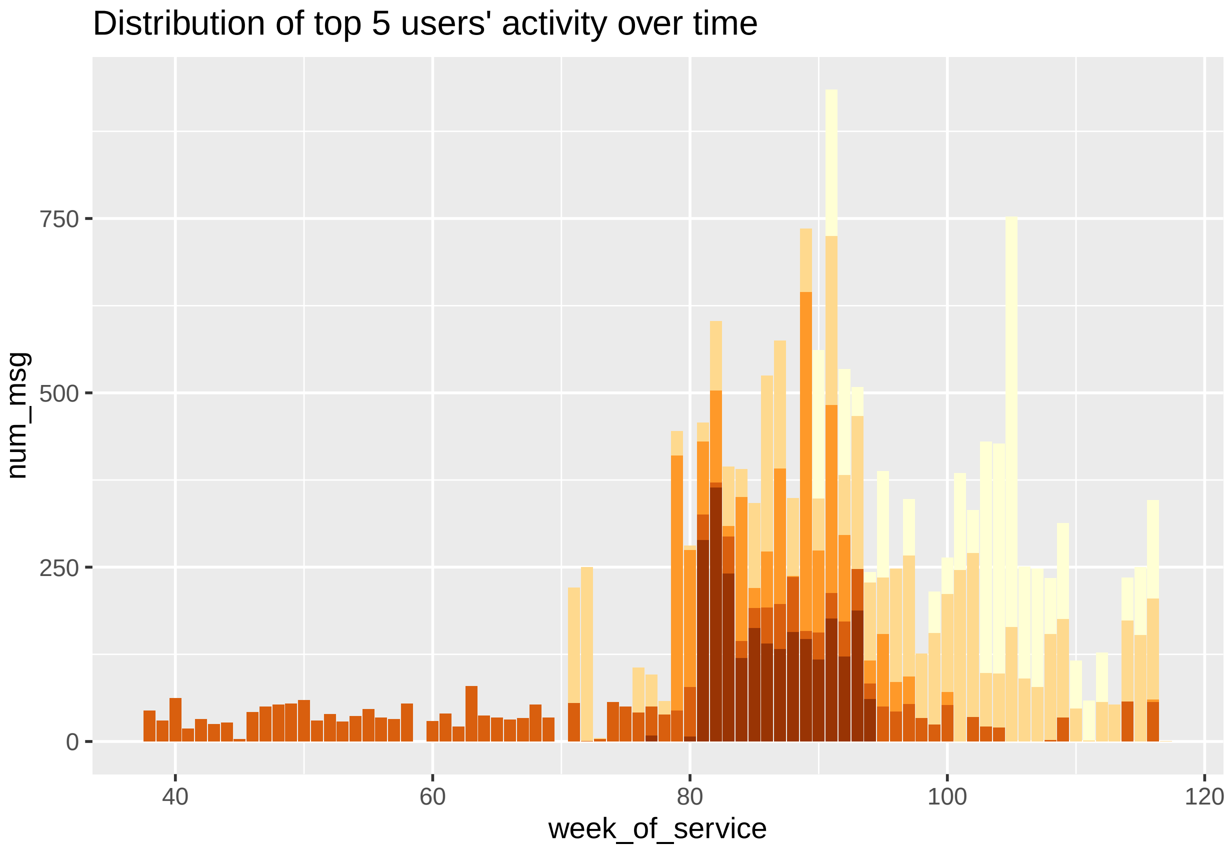
```r
top5_member <- actives %>%
  head(5) %>%
  pull(hashed_member_id)

top10_member <- actives %>%
  head(10)

ggplot(actives, aes(x = total)) +
  geom_histogram(binwidth = 20, aes(fill = "e3a42c")) +
  theme(legend.position = "none") +
  labs(title = "Histogram of total user messages")
```



```r
actives %>%
  filter(total < 300) %>%
  ggplot(aes(x = total)) +
    geom_histogram(binwidth = 5, aes(fill = "e3a42c")) +
    theme(legend.position = "none")  +
    labs(title = "Histogram of total user messages",
         subtitle = "Filtered for total < 300")
```



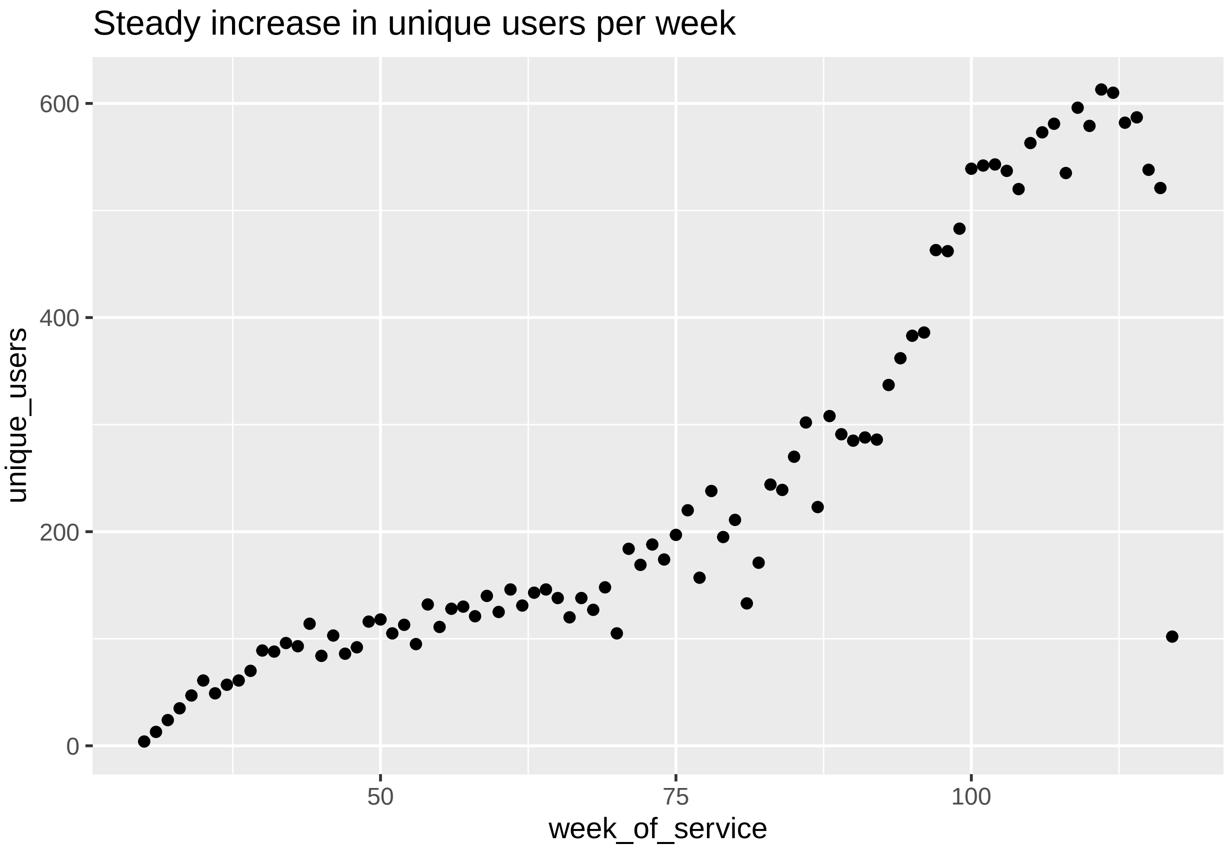
```r
top5_activity <- coach %>%
  filter(hashed_member_id %in% top5_member) %>%
  arrange(desc(num_msg))

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") +
  labs(title = "Distribution of top 5 users' activity over time")
```



```
unique <- coach %>%
  group_by(week_of_service) %>%
  summarise(n_distinct(hashed_member_id)) %>%
  rename(unique_users = "n_distinct(hashed_member_id)")
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
ggplot(unique, aes(x = week_of_service, y = unique_users)) +
  geom_point() +
  labs(title = "Steady increase in unique users per week")
```

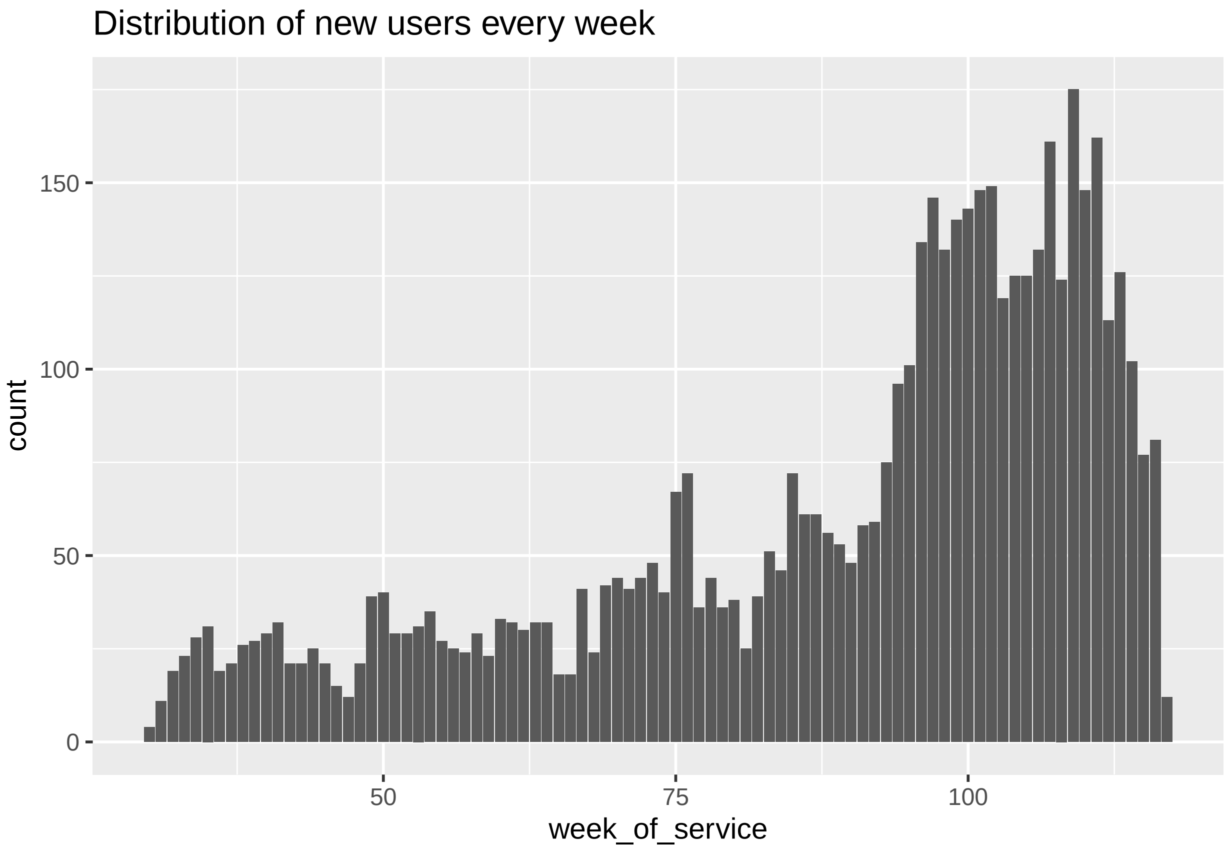


Meeting on Sunday, October 25th

```
new_users <- coach %>%
  group_by(hashed_member_id) %>%
  filter(week_of_service == min(week_of_service)) %>%
  slice(1) %>%
  ungroup()
new_users
```

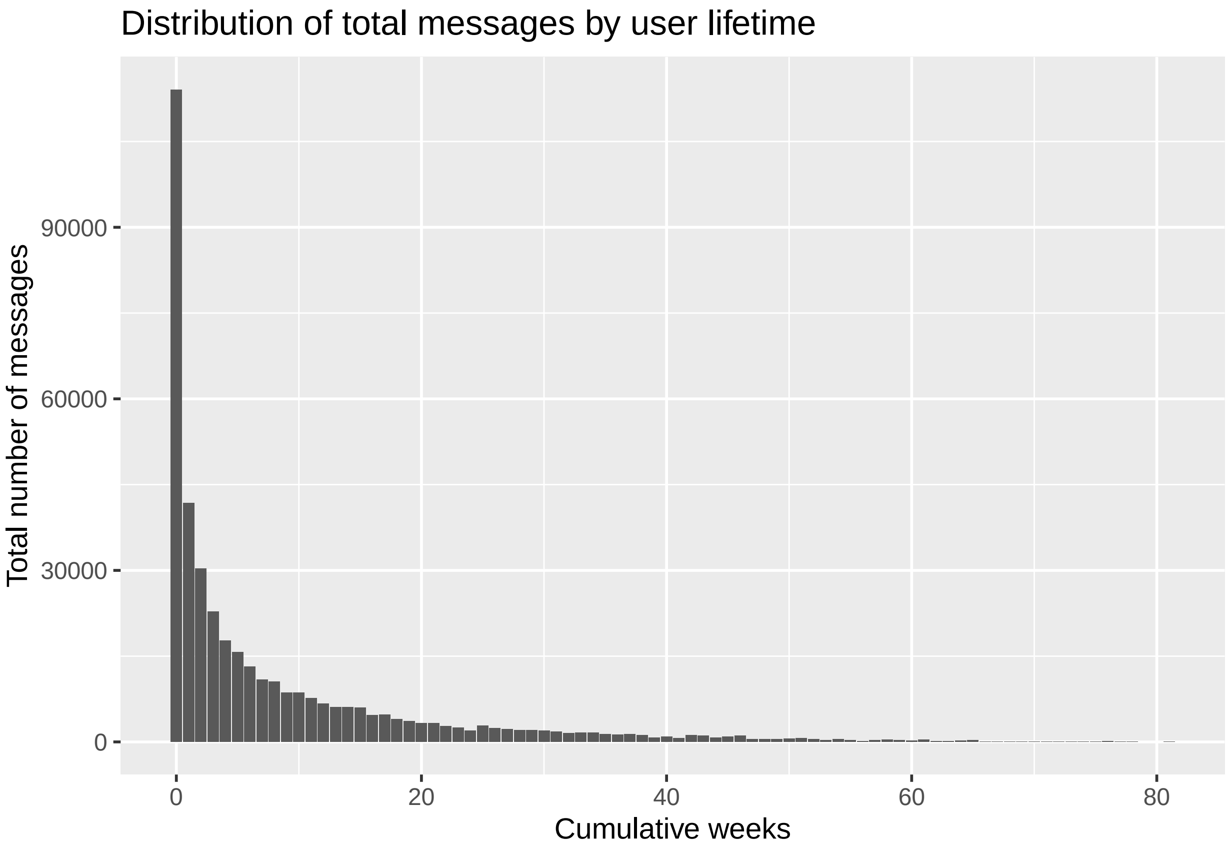
```
## # A tibble: 5,224 x 4
##       X1 hashed_member_id                                week_of_service num_msg
##    <dbl> <chr>                                                     <dbl>   <dbl>
##  1 17292 0017aaf65be7d4c48b697b8dad15d9789a072326b19f46~              72       0
##  2  9627 001dcc2b865bf05616efdbe157b1bc900bc0cc9c7a7744~             100       1
##  3  7123 0021c1c7e6639a7b22a81fba1b6e9cca239c91178c91dc~             104      37
##  4 21442 005e3c7cabd16d1e5e3f70b7570cfdc95607442d4ed4f1~              34       7
##  5 19189 00617cf30e0072acd329acd8a86f22e621a7e706122d32~              58      20
##  6 16450 00795baa13e165f42931957173092c3afcf0065cfd15a8~              76      26
##  7 18384 007af113a7c50ff4f1c2ca0506b8586f5d4515981db673~              64       0
##  8 11213 00ac8b530db42e04a3759f8b16d2f76ec5d886b6d275dd~              96      17
##  9 12305 00be9d433d0a81adf15cb9ff437be7c5b9955257fa76a1~              93      17
## 10  3263 00bf9e9e3397b79ad7edf154df2f35e4290624a713c485~             111      31
## # ... with 5,214 more rows
```

```
ggplot(new_users, aes(x = week_of_service)) +
  geom_bar() +
  labs(title = "Distribution of new users every week")
```



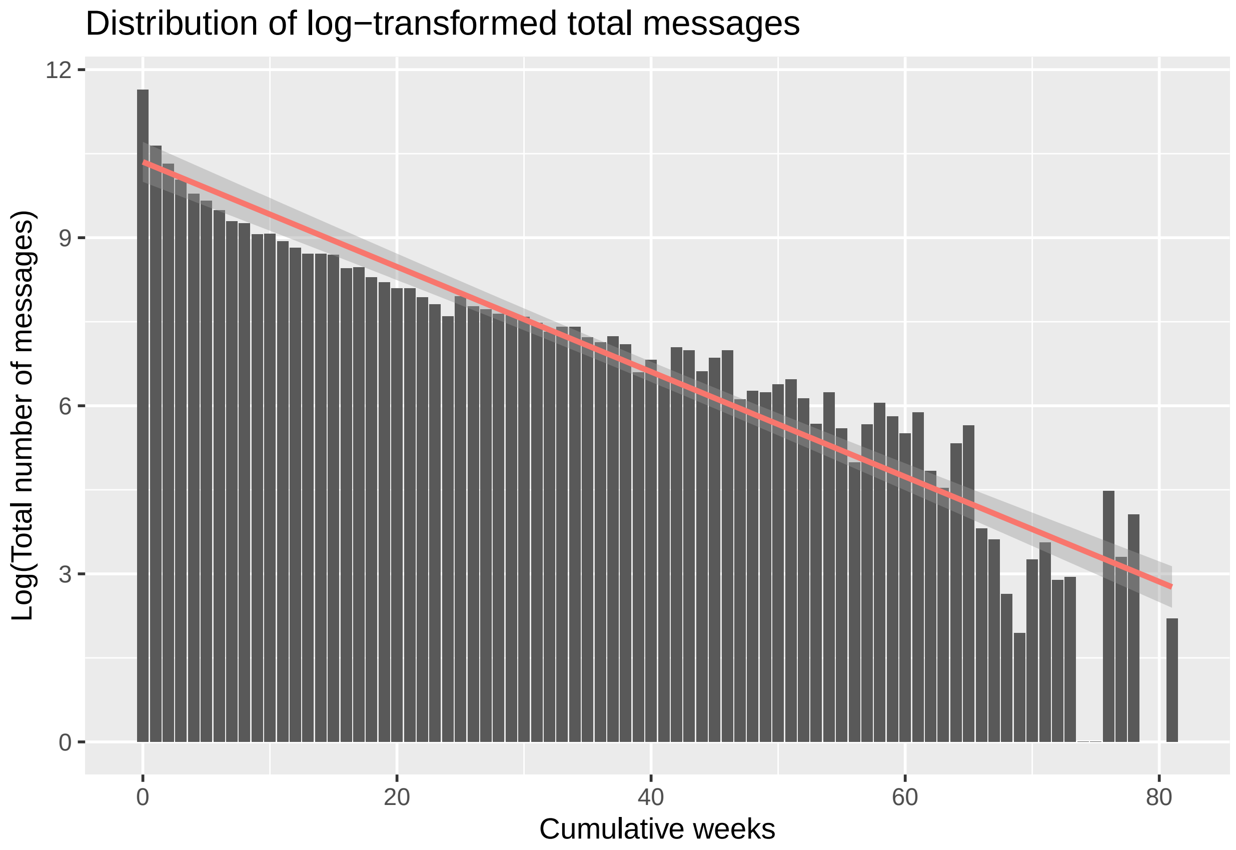
```
week_joined <- new_users %>%
  select(hashed_member_id, week_of_service) %>%
  rename(week_joined = week_of_service)

lifetime <- coach %>%
  inner_join(week_joined, by = "hashed_member_id") %>%
  mutate(cum_weeks = week_of_service - week_joined)

lifetime_count <- lifetime %>%
  group_by(cum_weeks) %>%
  count(num_msg) %>%
  mutate(total_msg = sum(n * num_msg)) %>%
  distinct(total_msg)
```

```
ggplot(lifetime_count, aes(x = cum_weeks, y = total_msg)) +
  geom_col() +
  labs(title = "Distribution of total messages by user lifetime",
       y = "Total number of messages",
       x = "Cumulative weeks")
```



```
ggplot(lifetime_count, aes(x = cum_weeks, y = log(total_msg))) +
  geom_col() +
  geom_smooth(method = "lm", aes(colour = "grey70")) +
  labs(title = "Distribution of log-transformed total messages",
       y = "Log(Total number of messages)",
       x = "Cumulative weeks") +
  theme(legend.position = "none")
```

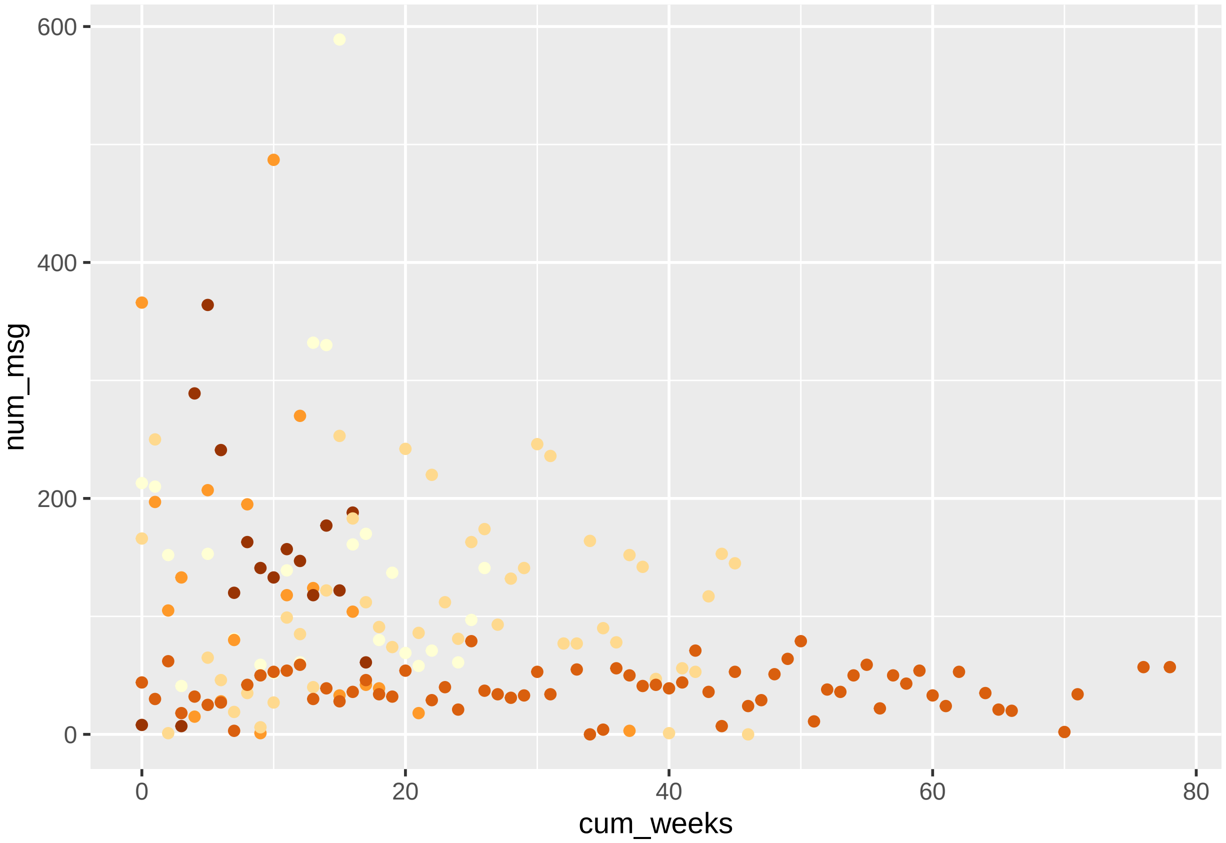
```
## `geom_smooth()` using formula 'y ~ x'
```



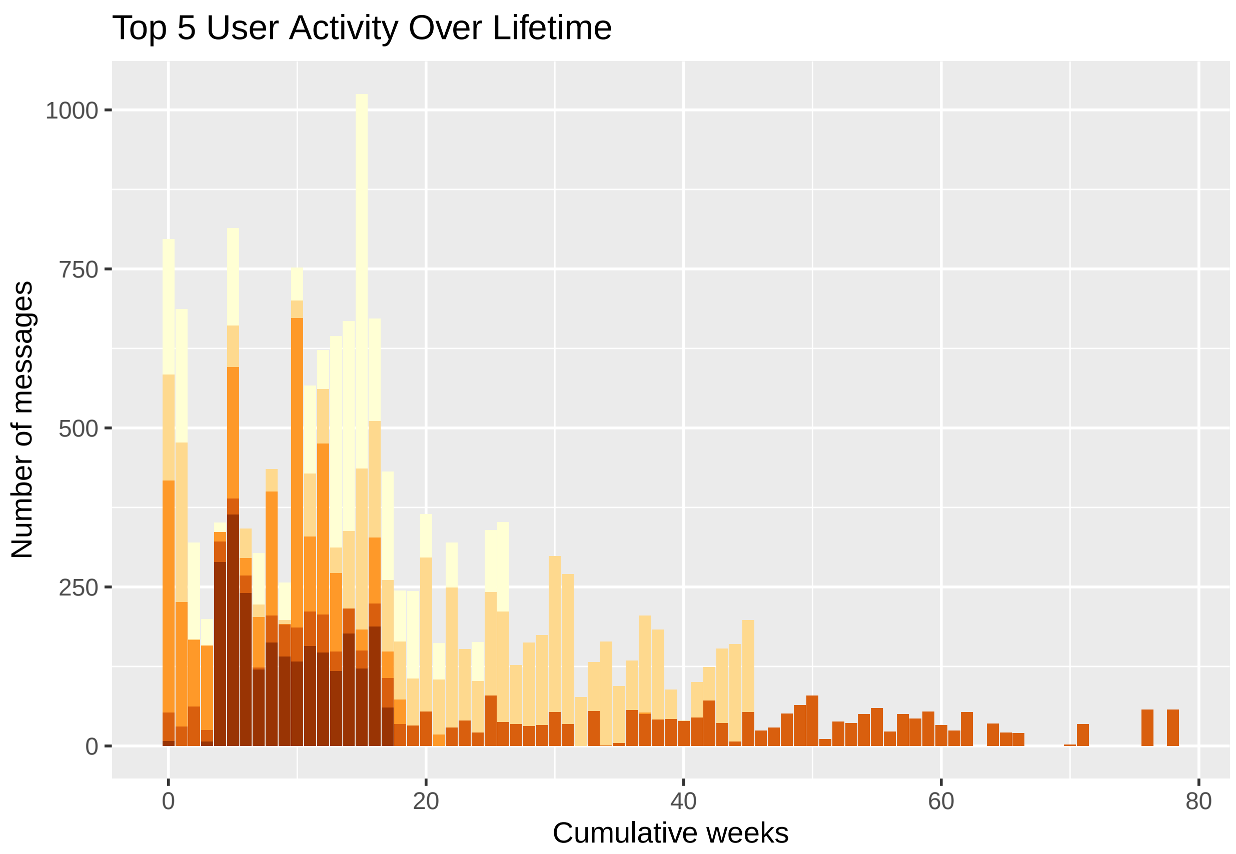
```
lm_model <- lm(data = lifetime_count, log(total_msg) ~ cum_weeks)
tidy(lm_model)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  10.4      0.179        57.7 9.31e-66
## 2 cum_weeks    -0.0937   0.00392     -23.9 1.15e-37
```

```
top5_lifetime <- lifetime %>%
  filter(hashed_member_id %in% top5_member)

ggplot(top5_lifetime, aes(x = cum_weeks, y = num_msg)) +
  geom_point(aes(color = hashed_member_id)) +
  theme(legend.position = "none") +
  scale_color_brewer(palette = "YlOrBr")
```



```
ggplot(top5_lifetime, aes(x = cum_weeks, y = num_msg)) +
  geom_col(aes(fill = hashed_member_id)) +
  scale_fill_brewer(palette = "YlOrBr") +
  theme(legend.position = "none") +
  labs(title = "Top 5 User Activity Over Lifetime",
       y = "Number of messages",
       x = "Cumulative weeks")
```



```r
spanning <- lifetime %>%
  group_by(hashed_member_id) %>%
  filter(cum_weeks == max(cum_weeks)) %>%
  slice(1) %>%
  ungroup() %>%
  rename(total_weeks = cum_weeks)

spanning
```

```
## # A tibble: 5,224 x 6
##       X1 hashed_member_id        week_of_service num_msg week_joined total_weeks
##    <dbl> <chr>                             <dbl>   <dbl>       <dbl>       <dbl>
##  1 17292 0017aaf65be7d4c48b697b~              72       0          72           0
##  2  8973 001dcc2b865bf05616efdb~             101       2         100           1
##  3  2156 0021c1c7e6639a7b22a81f~             113       0         104           9
##  4 21442 005e3c7cabd16d1e5e3f70~              34       7          34           0
##  5 19049 00617cf30e0072acd329ac~              59      18          58           1
##  6 16450 00795baa13e165f4293195~              76      26          76           0
##  7 18384 007af113a7c50ff4f1c2ca~              64       0          64           0
##  8  9558 00ac8b530db42e04a3759f~             100       0          96           4
##  9 11837 00be9d433d0a81adf15cb9~              95       0          93           2
## 10  3263 00bf9e9e3397b79ad7edf1~             111      31         111           0
## # ... with 5,214 more rows
```

```r
spanning %>%
  summarise(mean = mean(total_weeks))
```

```
## # A tibble: 1 x 1
##    mean
##   <dbl>
## 1  7.70
```

```
spanning %>%
  filter(hashed_member_id %in% top5_member) %>%
  summarise(mean = mean(total_weeks))
```

```
## # A tibble: 1 x 1
##    mean
##   <dbl>
## 1  40.8
```

```
counts %>%
  arrange(desc(total_msg)) %>%
  head(5)
```

```
## # A tibble: 5 x 2
## # Groups:   week_of_service [5]
##   week_of_service total_msg
##             <dbl>     <dbl>
## 1             112     11599
## 2             109     11585
## 3             111     10648
## 4             110     10642
## 5             100     10590
```

```
lifetime %>%
  filter(week_of_service == 112) %>% #610 users present in this week
  summarise(span = mean(cum_weeks))
```

```
## # A tibble: 1 x 1
##    span
##   <dbl>
## 1  10.2
```

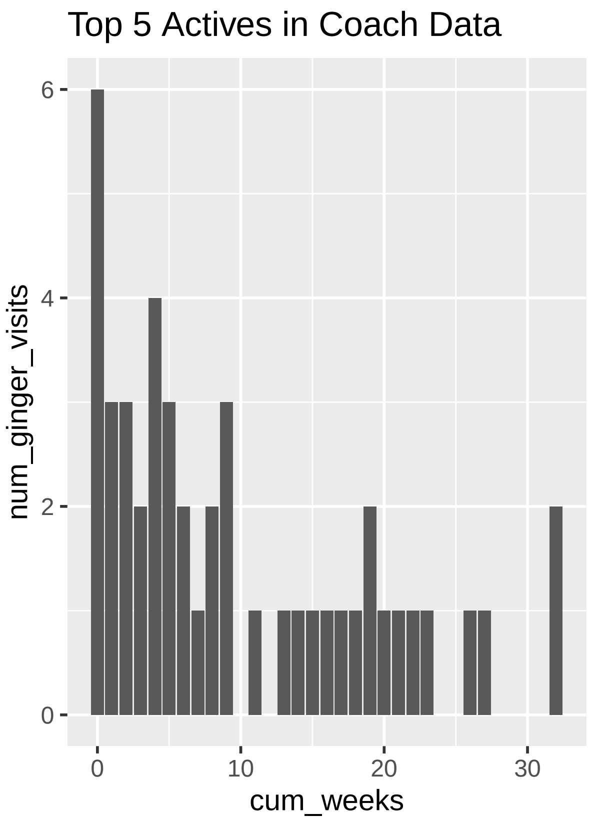
```
new_users_clinic <- clinic %>%
  group_by(hashed_member_id) %>%
  filter(week_of_service == min(week_of_service)) %>%
  slice(1) %>%
  ungroup()

week_joined_clinic <- new_users_clinic %>%
  select(hashed_member_id, week_of_service) %>%
  rename(week_joined = week_of_service)


clinic_lifetime <- clinic %>%
  inner_join(week_joined_clinic, by = "hashed_member_id") %>%
  mutate(cum_weeks = week_of_service - week_joined)
```

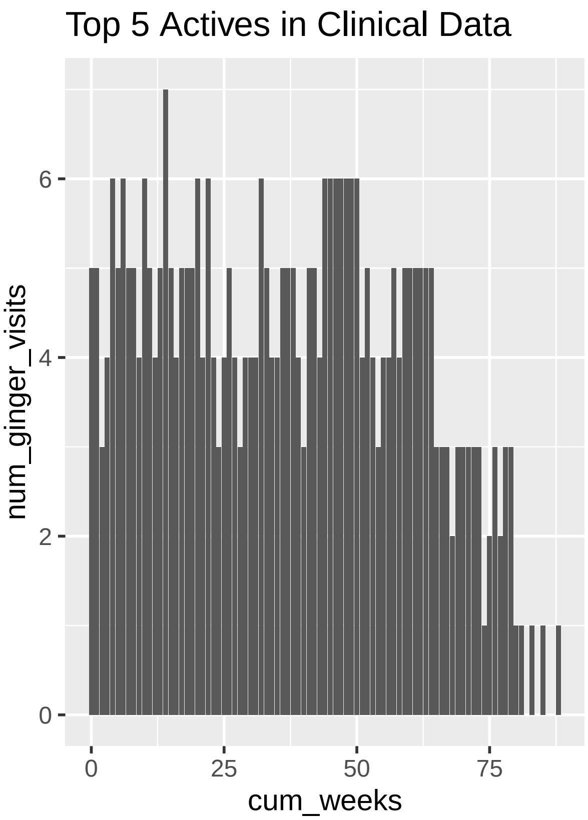
```
clinic_actives <- clinic %>%
  group_by(hashed_member_id) %>%
  count(num_ginger_visits) %>%
  mutate(total = sum(n * num_ginger_visits)) %>%
  distinct(total) %>%
  arrange(desc(total))
```

```r
clinic_actives
```

```
## # A tibble: 5,928 x 2
## # Groups:   hashed_member_id [5,928]
##    hashed_member_id                                                 total
##    <chr>                                                            <dbl>
##  1 9a6c4bb4752f99c542060bcf8e08775ff742577bdd62407617e6eaa9783b11c0    79
##  2 dc3ed07355aa2228a08eed8a078c47a8903a6f40f555bdad7eba0aa80f8aaff6    73
##  3 0a58e1af2044304822c39f8ca63262d4876637f25e52adbe8dac291057fe7ce1    69
##  4 5d1b58ad0bf56e2b0d277c7bed664065a2410196186ddb19bddf5353741d1d4b    69
##  5 87b76a4573876af0a1bd4f348af26e5407df04c972262c4ba74513a37ffc144f    65
##  6 478b3b9f820d7fd0693972c69974cc88571133a877ee4b63a44ae02d9b013529    64
##  7 8a3c2d005c2987c3aa3df03c0b47f6b32e92bbb38bfe8fd290bfe2d629b70b07    64
##  8 cde577be3bb9c602f8d4498faffd82d2e7436ac176af627de60fc8172c06350a    64
##  9 2b70e9543194fe2024d5ee63029356ee96bfe3de3cab83041e10d4e446ad60f1    63
## 10 4a6c283c40d7fd7748b85b8bae0497c289be45e312e97a04e02df05adff6edb4    62
## # ... with 5,918 more rows
```

```r
top5_clinic_active <- clinic_actives %>%
  head(5) %>%
  pull(hashed_member_id)

top5_clinic_lifetime_coach <- clinic_lifetime %>%
  filter(hashed_member_id %in% top5_member)

top5_clinic_lifetime_clinic <- clinic_lifetime %>%
  filter(hashed_member_id %in% top5_clinic_active)


a <- ggplot(top5_clinic_lifetime_coach, aes(x = cum_weeks, y = num_ginger_visits)) +
  geom_col() +
  theme(legend.position = "none") +
  labs(title = "Top 5 Actives in Coach Data")

b <- ggplot(top5_clinic_lifetime_clinic, aes(x = cum_weeks, y = num_ginger_visits)) +
  geom_col() +
  theme(legend.position = "none") +
  labs(title = "Top 5 Actives in Clinical Data")

a + b
```





```
top5_clinic_active <- clinic_actives %>%
  head(5) %>%
  pull(hashed_member_id)
top5_clinic_active
```

```
## [1] "9a6c4bb4752f99c542060bcf8e08775ff742577bdd62407617e6eaa9783b11c0"
## [2] "dc3ed07355aa2228a08eed8a078c47a8903a6f40f555bdad7eba0aa80f8aaff6"
## [3] "0a58e1af2044304822c39f8ca63262d4876637f25e52adbe8dac291057fe7ce1"
## [4] "5d1b58ad0bf56e2b0d277c7bed664065a2410196186ddb19bddf5353741d1d4b"
## [5] "87b76a4573876af0a1bd4f348af26e5407df04c972262c4ba74513a37ffc144f"
```

```
top_both_coach <- lifetime %>%
  filter(hashed_member_id %in% top5_clinic_active | hashed_member_id %in% top5_member) %>%
  select(hashed_member_id, cum_weeks, num_msg, week_joined)

top_both_clinic <- clinic_lifetime %>%
  filter(hashed_member_id %in% top5_clinic_active | hashed_member_id %in% top5_member) %>%
  select(hashed_member_id, cum_weeks, num_ginger_visits, week_joined)

comparison <- top_both_clinic %>%
  inner_join(top_both_coach, by = "hashed_member_id")

all_join <- week_joined_clinic %>%
  inner_join(week_joined, by = "hashed_member_id") %>%
  mutate(time_taken = week_joined.x - week_joined.y) %>%#clinic
  #select(hashed_member_id, time_taken) %>%
  mutate(which_first = case_when(time_taken > 0 ~ "Coaching",
                                 time_taken < 0 ~ "Clinic")) %>%
  replace_na(list(which_first = "Same time"))
```

```
all_join
```

```
## # A tibble: 5,224 x 5
##    hashed_member_id           week_joined.x week_joined.y time_taken which_first
##    <chr>                              <dbl>         <dbl>      <dbl> <chr>
##  1 0017aaf65be7d4c48b697b8da~            76            72          4 Coaching
##  2 001dcc2b865bf05616efdbe15~           103           100          3 Coaching
##  3 0021c1c7e6639a7b22a81fba1~           109           104          5 Coaching
##  4 005e3c7cabd16d1e5e3f70b75~            34            34          0 Same time
##  5 00617cf30e0072acd329acd8a~            55            58         -3 Clinic
##  6 00795baa13e165f4293195717~            83            76          7 Coaching
##  7 007af113a7c50ff4f1c2ca050~            62            64         -2 Clinic
##  8 00ac8b530db42e04a3759f8b1~            98            96          2 Coaching
##  9 00be9d433d0a81adf15cb9ff4~            95            93          2 Coaching
## 10 00bf9e9e3397b79ad7edf154d~           112           111          1 Coaching
## # ... with 5,214 more rows
```

```
comparison #what's the question to be asked here?
```

```
## # A tibble: 4,043 x 7
##    hashed_member_id cum_weeks.x num_ginger_visi~ week_joined.x cum_weeks.y
##    <chr>                  <dbl>            <dbl>         <dbl>       <dbl>
##  1 5d1b58ad0bf56e2~          14                2            38           4
##  2 5d1b58ad0bf56e2~          14                2            38           3
##  3 5d1b58ad0bf56e2~          14                2            38           2
##  4 5d1b58ad0bf56e2~          14                2            38           1
##  5 5d1b58ad0bf56e2~          14                2            38           0
##  6 dc3ed07355aa222~          33                2            55          57
##  7 dc3ed07355aa222~          33                2            55          55
##  8 dc3ed07355aa222~          33                2            55          51
##  9 dc3ed07355aa222~          33                2            55          48
## 10 dc3ed07355aa222~          33                2            55          47
## # ... with 4,033 more rows, and 2 more variables: num_msg <dbl>,
## #   week_joined.y <dbl>
```

```
#how long does it take for coached patients to join clinic? (every member id)
```

```
all_join %>%
  group_by(which_first) %>%
  summarise(mean_weeksToSwtich = round(mean(abs(time_taken)), digits = 2)) %>%
  filter(which_first != "Same time") %>%
  kable()
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

which_first	mean_weeksToSwtich
Clinic	7.09
Coaching	4.30

```
all_join %>%
  group_by(which_first) %>%
  count(which_first) %>%
  rename(count = n) %>%
  kable()
```

which_first	count
Clinic	288
Coaching	3995
Same time	941

```
spanning_clinic <- clinic_lifetime %>%
  group_by(hashed_member_id) %>%
  filter(cum_weeks == max(cum_weeks)) %>%
  slice(1) %>%
  ungroup() %>%
  rename(total_weeks = cum_weeks)

spanning_clinic %>%
  summarise(mean = mean(total_weeks))
```

```
## # A tibble: 1 x 1
##    mean
##   <dbl>
## 1  13.3
```

```
spanning_clinic %>%
  filter(hashed_member_id %in% top5_clinic_active) %>%
  summarise(mean = mean(total_weeks))
```

```
## # A tibble: 1 x 1
##    mean
##   <dbl>
## 1    75
```

```
ggplot()
```