Retention analysis

Tianpei Qian

2023-02-07

Step 1: read data

```
subscription_file <- 'subscription.csv'
read_lines(subscription_file, n_max = 5)
```

```
## [1] "\"user_id\",\"subscription_signup_date\",\"subscription_monthly_cost\",\"country\",\"source\",\
## [2] "1459,\"January, 2015\",29,\"Spain\",\"ads\",4,0"
## [3] "12474,\"January, 2015\",49,\"France\",\"ads\",5,0"
## [4] "12294,\"January, 2015\",49,\"Germany\",\"ads\",2,0"
## [5] "3878,\"January, 2015\",49,\"China\",\"ads\",1,0"
```

```
subscription_raw <-
  read_csv(
    subscription_file,
    col_types =
      list(
        col_integer(),
        col_character(),
        col_integer(),
        col_character(),
        col_character(),
        col_integer(),
        col_logical()
      )
  )
```

Step 2: data cleaning

Join tables, process factors, etc.

```
subscription <-
  subscription_raw %>%
  mutate(
    sku = factor(subscription_monthly_cost, levels = c(29, 49, 99)),
    country = factor(country, levels = unique(country)),
    source = factor(source, levels = c("ads", "friend_referral", "seo"))
  ) %>%
  select(-subscription_monthly_cost)
```

Step 3: exploratory analysis

```
subscription %>% summary()
```

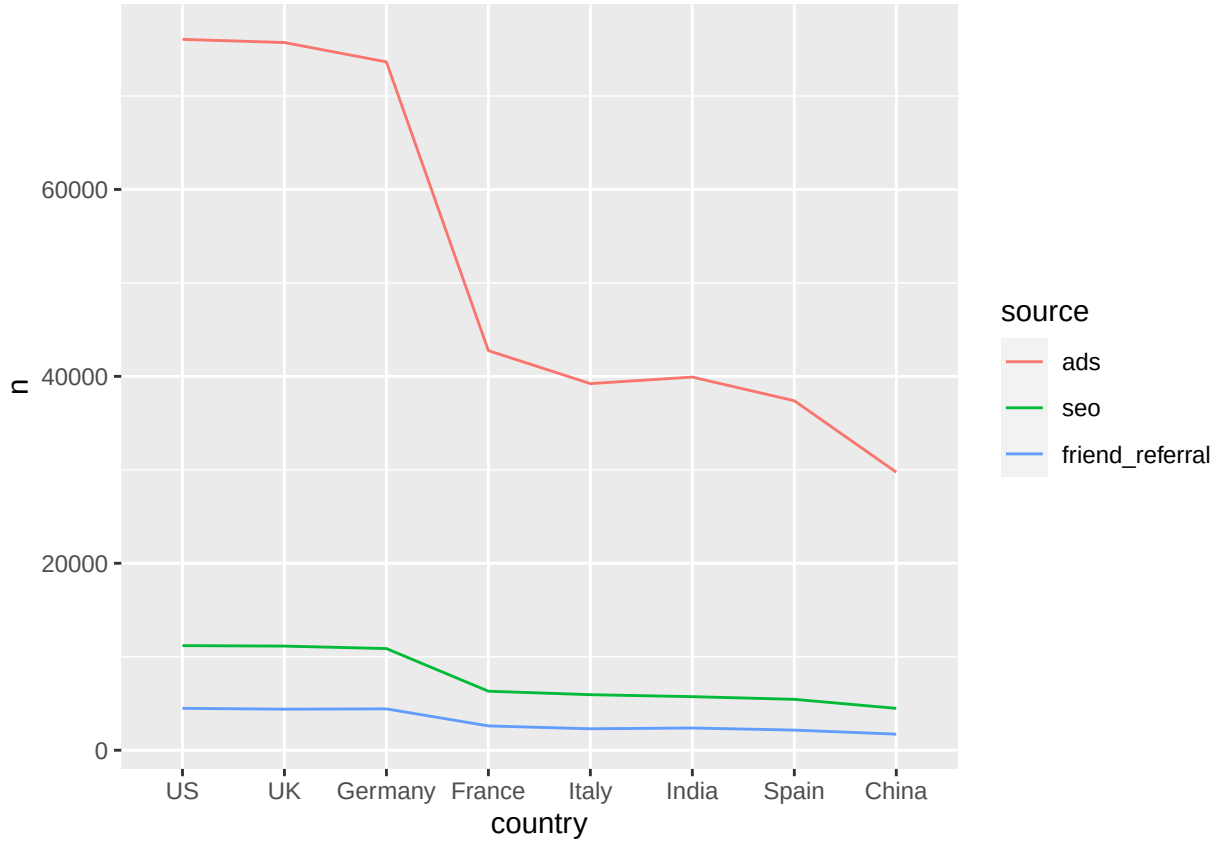
```
##     user_id       subscription_signup_date    country
```

```
##  Min.   :     1   Length:500000           US     :91731
##  1st Qu.:125001   Class :character        UK     :91252
##  Median :250000   Mode  :character        Germany:88944
##  Mean   :250000                           France :51662
##  3rd Qu.:375000                           India  :48027
##  Max.   :500000                           Italy  :47459
##                                           (Other):80925
##              source       billing_cycles  is_active       sku
##  ads            :414469   Min.   :1.000   Mode :logical   29:146362
##  friend_referral: 24428   1st Qu.:1.000   FALSE:450001    49:300397
##  seo            : 61103   Median :1.000   TRUE :49999     99: 53241
##                           Mean   :2.385
##                           3rd Qu.:3.000
##                           Max.   :8.000
##
```

Customer distribution Firstly, let's understand how our customers are distributed by country, by sku and by source.

```
subscription %>%
  group_by(country, source) %>%
  summarise(
    n = n()
  ) %>%
  ungroup() %>%
  mutate(
    country = fct_reorder(country, -n),
    source = fct_reorder2(source, country, n)
  ) %>%
  ggplot(aes(country, n)) +
  geom_line(aes(color = source, group = source))
```

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

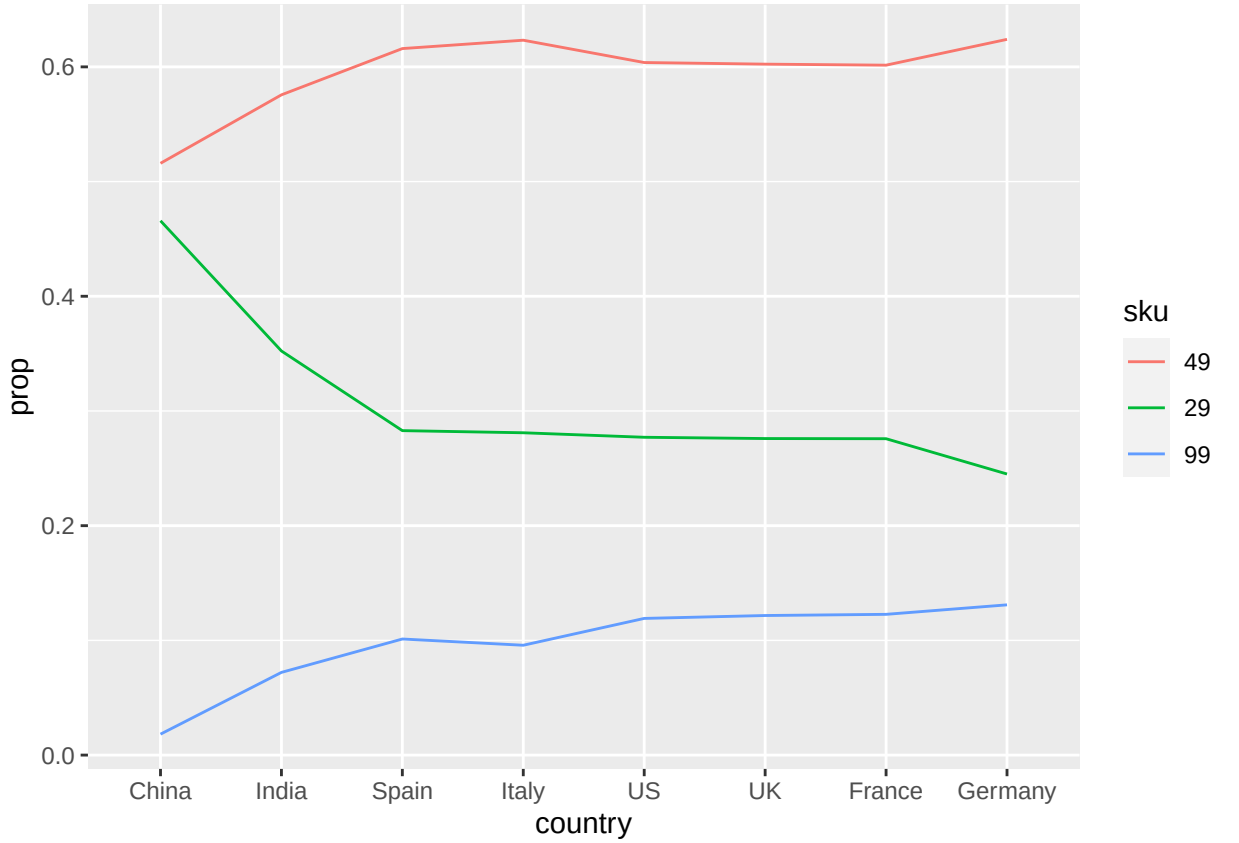


Most of our customers come come ads. US, UK and Germany and the Top 3 countries our product sell to.

Insights: Since there are great product oppotunities in the remaining countries, especially China and India given the large population base, it would be interesting to understand whether this is a marketing problem (not enough publicity) or a product problem (e.g. bad localization, better alternatives, etc.)

```
subscription %>%
  group_by(country, sku) %>%
  summarise(
    n = n()
  ) %>%
  mutate(
    prop = n / sum(n)
  ) %>%
  ungroup() %>%
  mutate(
    country = fct_reorder(country, -prop),
    sku = fct_reorder2(sku, country, prop)
  ) %>%
  ggplot(aes(country, prop)) +
  geom_line(aes(color = sku, group = sku))
```

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

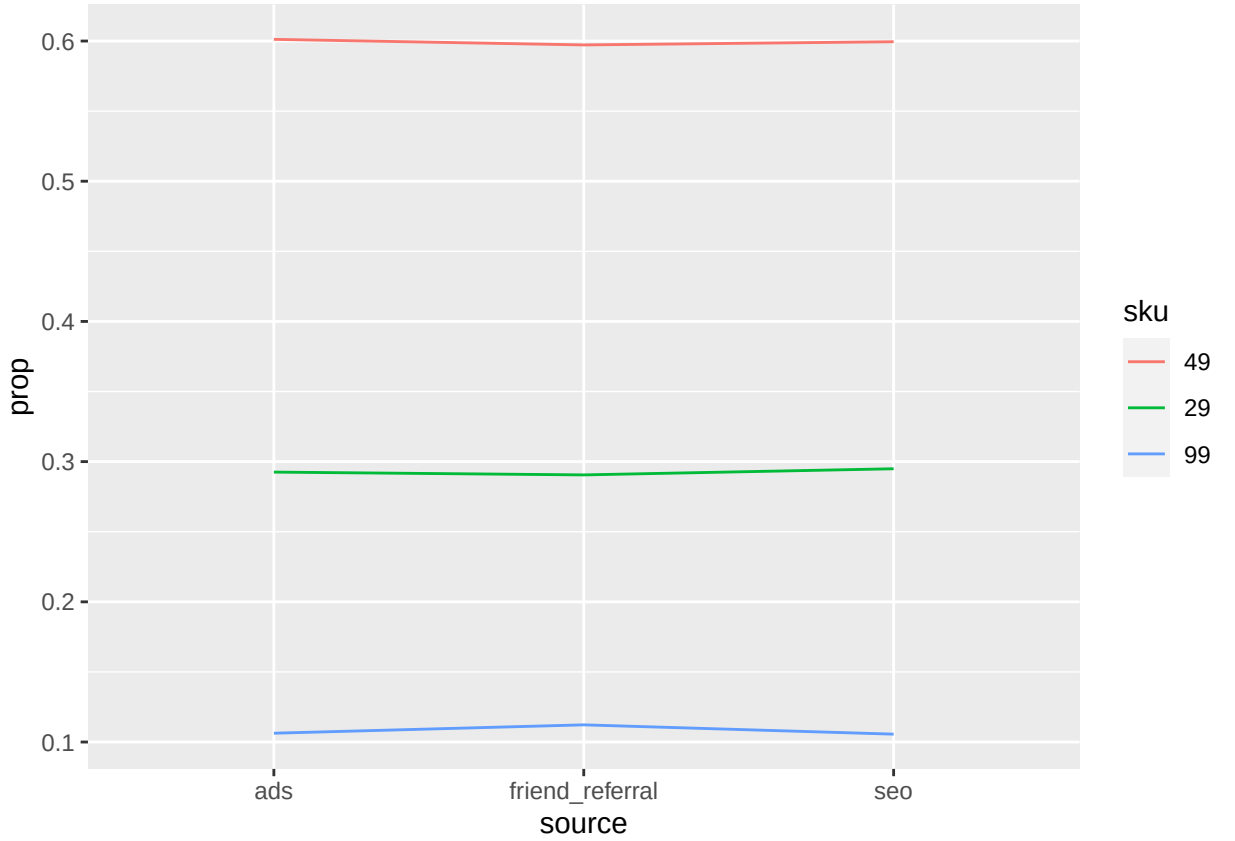


Overall, the $49/month product is the most popular across all countries. However, Chinese and Indian customers are much more likely to purphase the $29/month product.

Insights: There are 2 explanations for the China and India data points. Firstly, it's possible that the $29 product already satisfies the needs of most customers in the 2 countries. It would then be interesting to explore breaking the $29 into more SKUs so that we can attract more customers with a finer SKU lineup. Secondly, our product is overpriced in China and India. The low absolute number of customers from the 2 countries also support this hypothesis. We may be able to get more customers by adjusting down the price levels. Hopefully the larger customer base and better customer retention can lead to more revenue in the end. It is not possible to have regional pricing, then we may also consider sending out coupons to customers in China and India.

```
subscription %>%
  group_by(source, sku) %>%
  summarise(n = n()) %>%
  mutate(prop = n / sum(n)) %>%
  ungroup() %>%
  mutate(sku = fct_reorder(sku, -prop) ) %>%
  ggplot(aes(source, prop)) +
  geom_line(aes(color = sku, group = sku))
```

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



The relationpship between source and SKU is not very interesting. Customers from the 3 sources have the same distributions across the 3 SKUs.

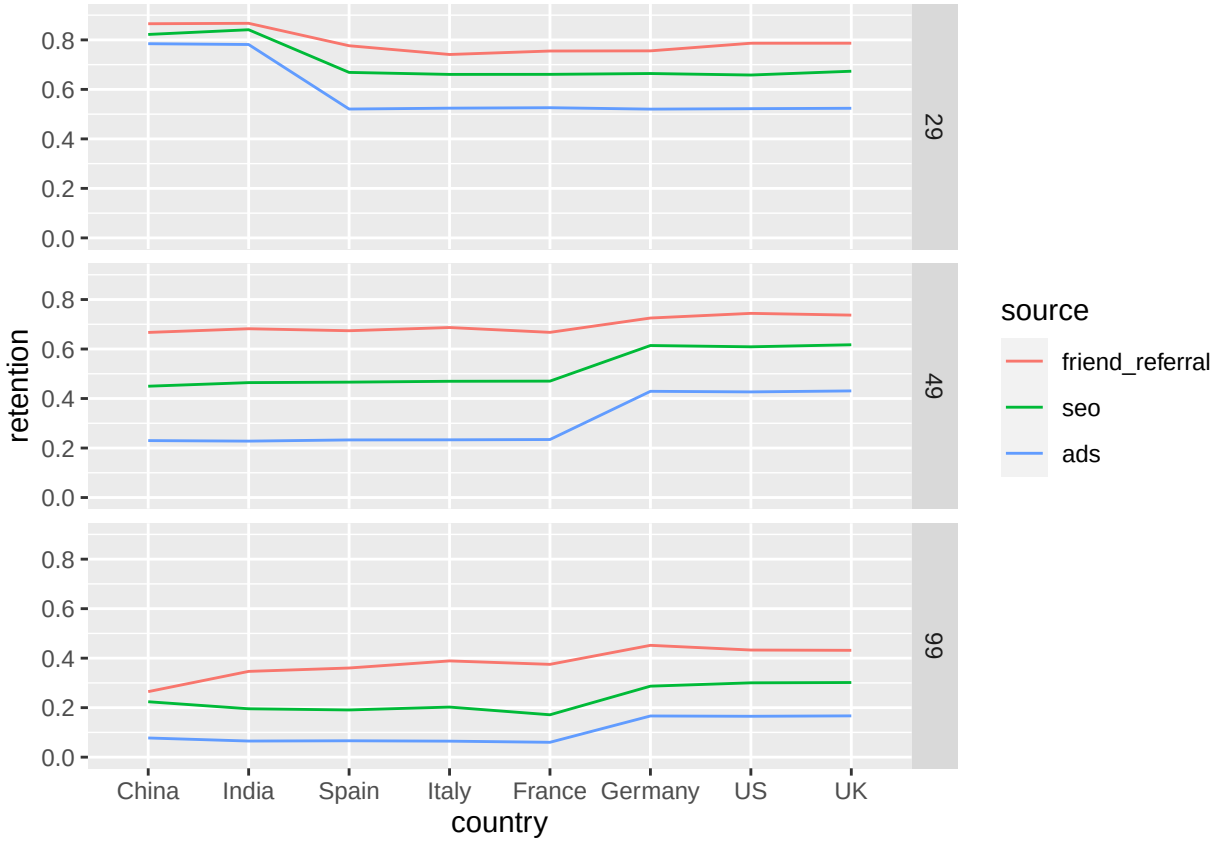
Customer retention Let's look at both short-term retention (1-month retention) and long-term retention (7-month retention) They are both important because short-term retention tells us how fast we can make the product value clear to the customers while long-term retention tells us whether we can consistently and constantly deliver the value to the customers.

```
short_term_retention <-
  subscription %>%
  group_by(sku, country, source) %>%
  summarize(
    n = n(),
    retention = sum(billing_cycles > 1) / n()
  ) %>%
  ungroup() %>%
  mutate(
    country = country %>% fct_reorder(retention),
    source = source %>% fct_reorder2(country, retention)
  )
```

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

```
short_term_retention %>%
  ggplot(aes(country, retention)) +
  geom_line(aes(color = source, group = source)) +
```

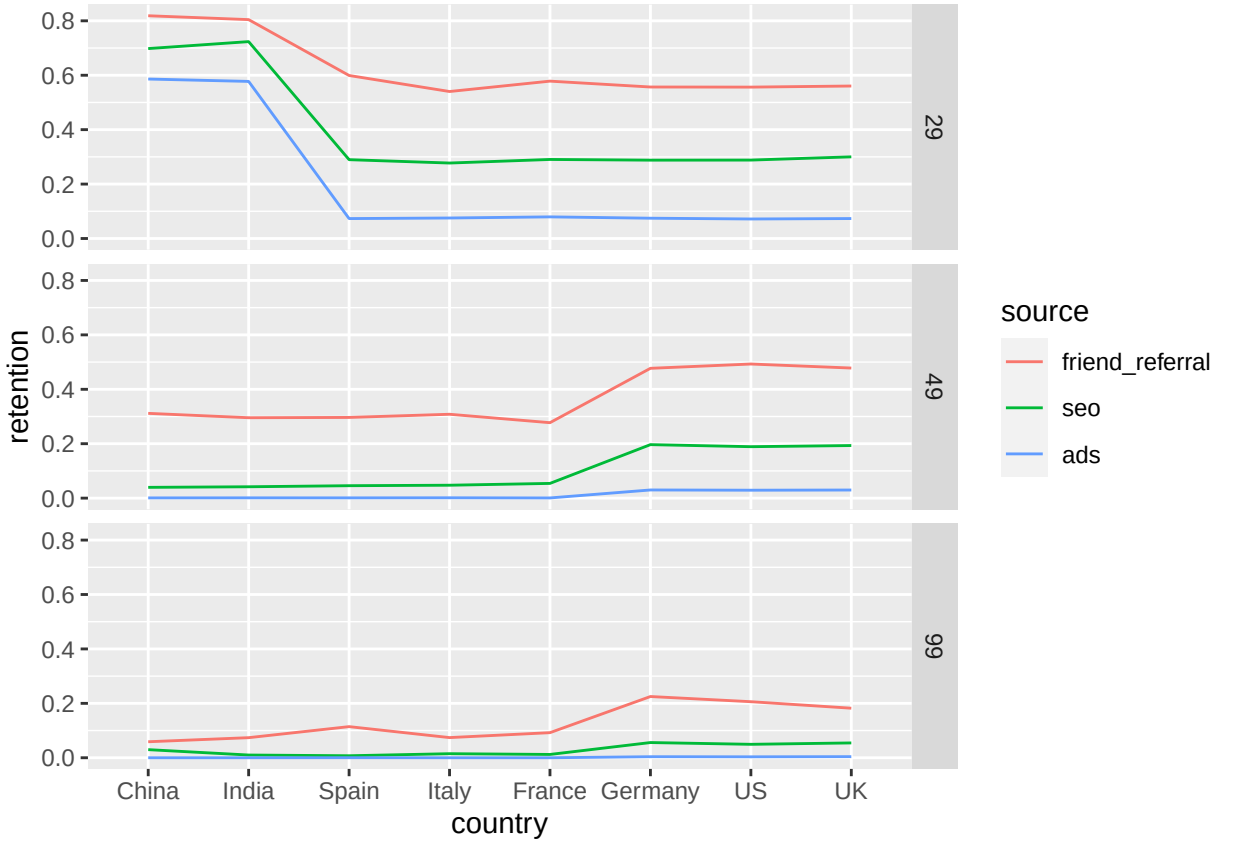
```
  scale_y_continuous(breaks = seq(0, 0.8, 0.2), limits = c(0, 0.9)) +
  facet_grid(rows = vars(sku))
```




```
long_term_retention <-
  subscription %>%
  group_by(sku, country, source) %>%
  summarize(
    n = n(),
    retention = sum(billing_cycles > 7) / n()
  ) %>%
  ungroup() %>%
  mutate(
    country = factor(country, levels = short_term_retention$country %>% levels()),
    source = source %>% fct_reorder2(country, retention)
  )
```

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

```
long_term_retention %>%
  ggplot(aes(country, retention)) +
  geom_line(aes(color = source, group = source)) +
  scale_y_continuous(breaks = seq(0, 0.8, 0.2)) +
  facet_grid(rows = vars(sku))
```



There are multiple insights from the 2 graphs

- Friend referral acquires the most committed customers. On the other hand, it is also the smallest channel according to our previous analysis. This suggests we should encourage more users to refer their friends by providing incentives like coupons.
- The $29 product has the most committed customers in China and India. Note that the product is also relatively more popular in the 2 countries than others. In particular, they start with high short-term retention and end with very little drop in long-term retention. There's almost no drop for friend referral. We need to understand the success and see if it can be replicated in other countries.
- The long-term retention for the $49 and $99 product is way too low, especially for customers from ads and seo. We still have moderate number of customers remaining after the first month, which indicates we do attract customers with some intentsion. Nonetheless, almost no customers from ads and seo remain after 7 months. It is very likely that we have some serious product problems.

Note: We may also approach this analysis with modelling, but it can easily get over-complicated with the interaction effects and all the categorical variables (it is important to pick the right base levels).

Lastly, let's work on predicting 1-year customer retention for each SKU. The ideal method is to find a linear relationship between cycle and retention and then fit a linear regression model.

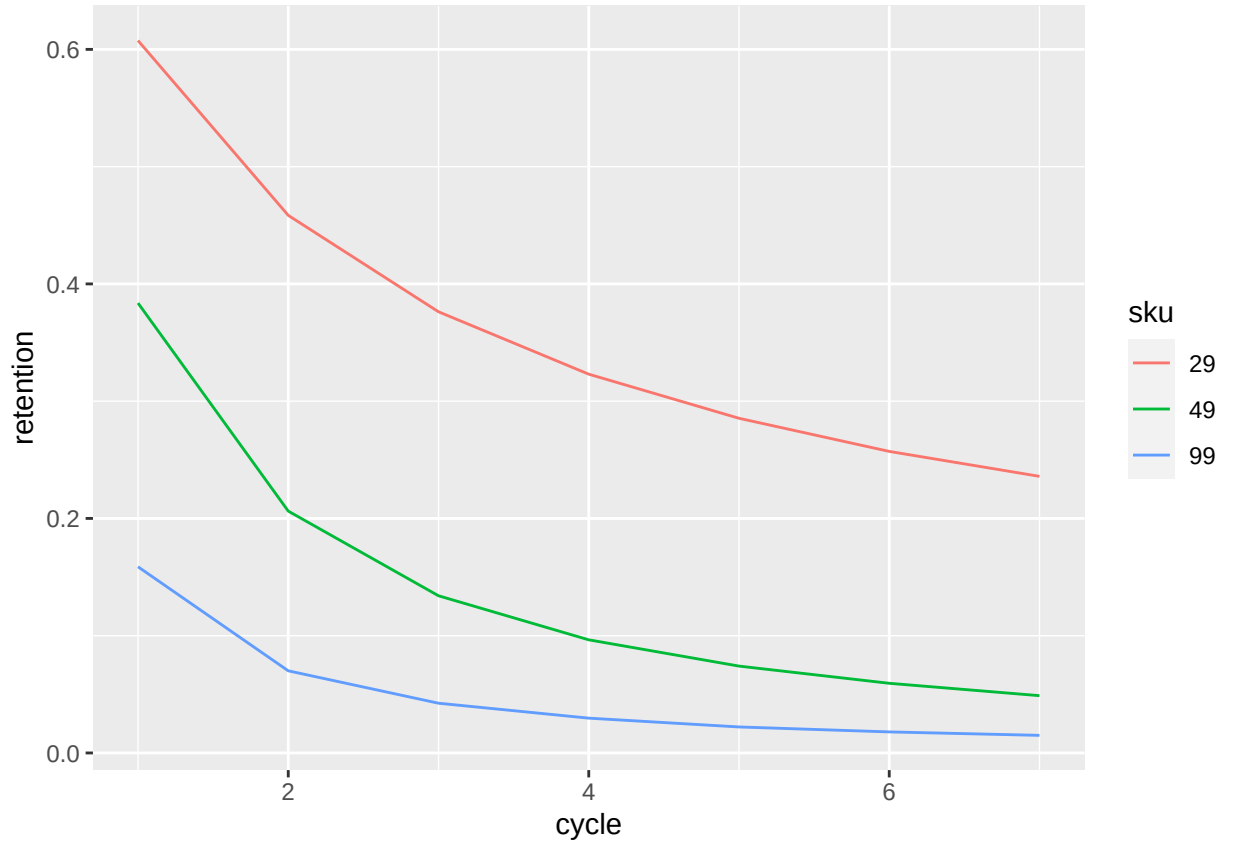
```
retention <-
  subscription %>%
  group_by(sku, billing_cycles) %>%
  summarize(
    n = n()
  ) %>%
  arrange(billing_cycles) %>%
```

```
  mutate(
    retention = 1 - cumsum(n) / sum(n),
    cycle = billing_cycles
  ) %>%
  ungroup() %>%
  filter(cycle < 8)
```

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

Note that we may also calculate retention for each cycle in other ways. For example, instead of looking at retention of all customers, we may also look at retention of customers that retain from the last cycle. It is harder to get a linear relationship from this definition so we do not proceed with it.

```
retention %>%
  ggplot() +
  geom_line(aes(cycle, retention, color = sku))
```



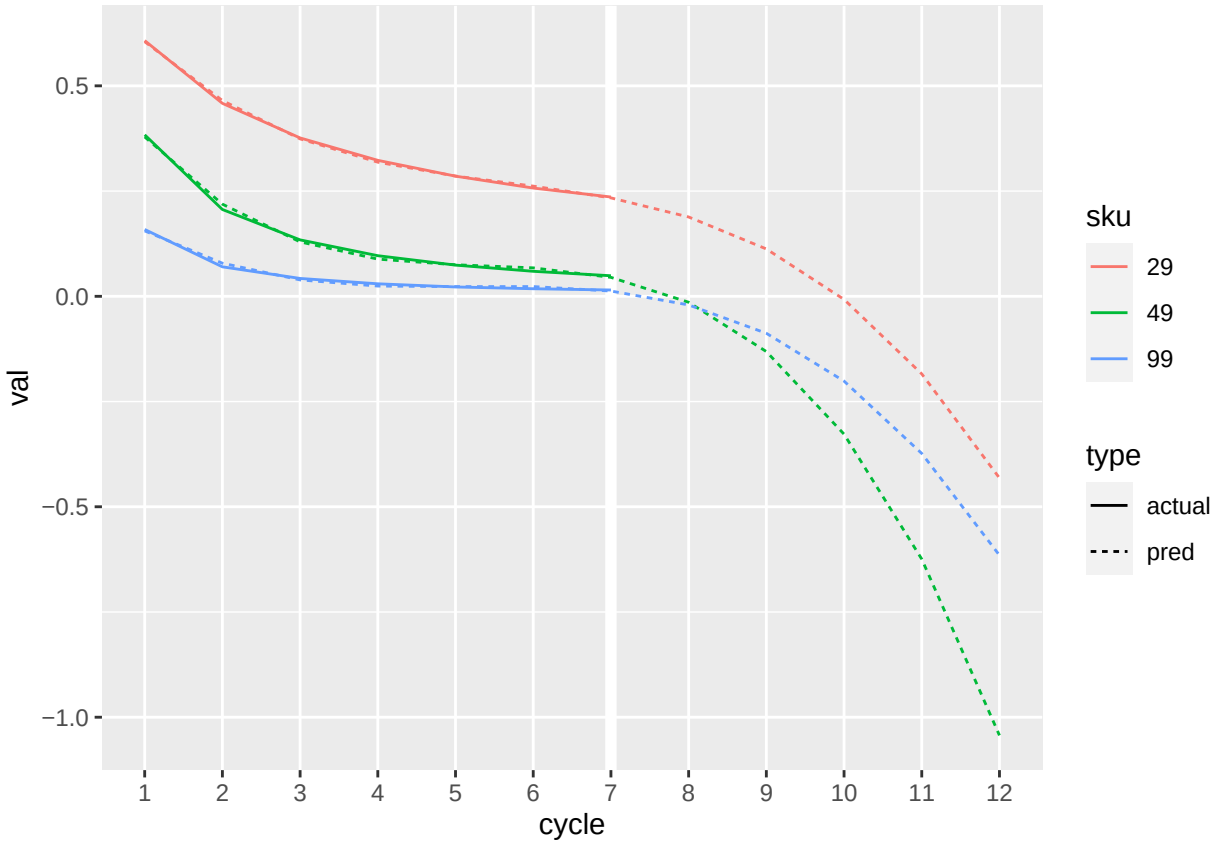
The relationship is obviously non-linear. It is tempting to consider polynomial regression but it can be very dangerous!

```
retention %>%
  transmute(sku, cycle, retention) %>%
  rbind(data.frame(sku = "29", cycle = 7:12, retention = NA_real_)) %>%
  rbind(data.frame(sku = "49", cycle = 7:12, retention = NA_real_)) %>%
  rbind(data.frame(sku = "99", cycle = 7:12, retention = NA_real_)) %>%
  add_predictions(
    lm(retention ~ poly(cycle, 3) * sku, data = retention)
```

```
  ) %>%
  rename(actual = retention) %>%
  pivot_longer(actual:pred, names_to = "type", values_to = "val") %>%
  ggplot() +
  geom_vline(xintercept = 7, color = "white", size = 2) +
  geom_line(aes(cycle, val, color = sku, linetype = type)) +
  scale_x_continuous(breaks = seq(1, 12, 1), minor_breaks = NULL)
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
```

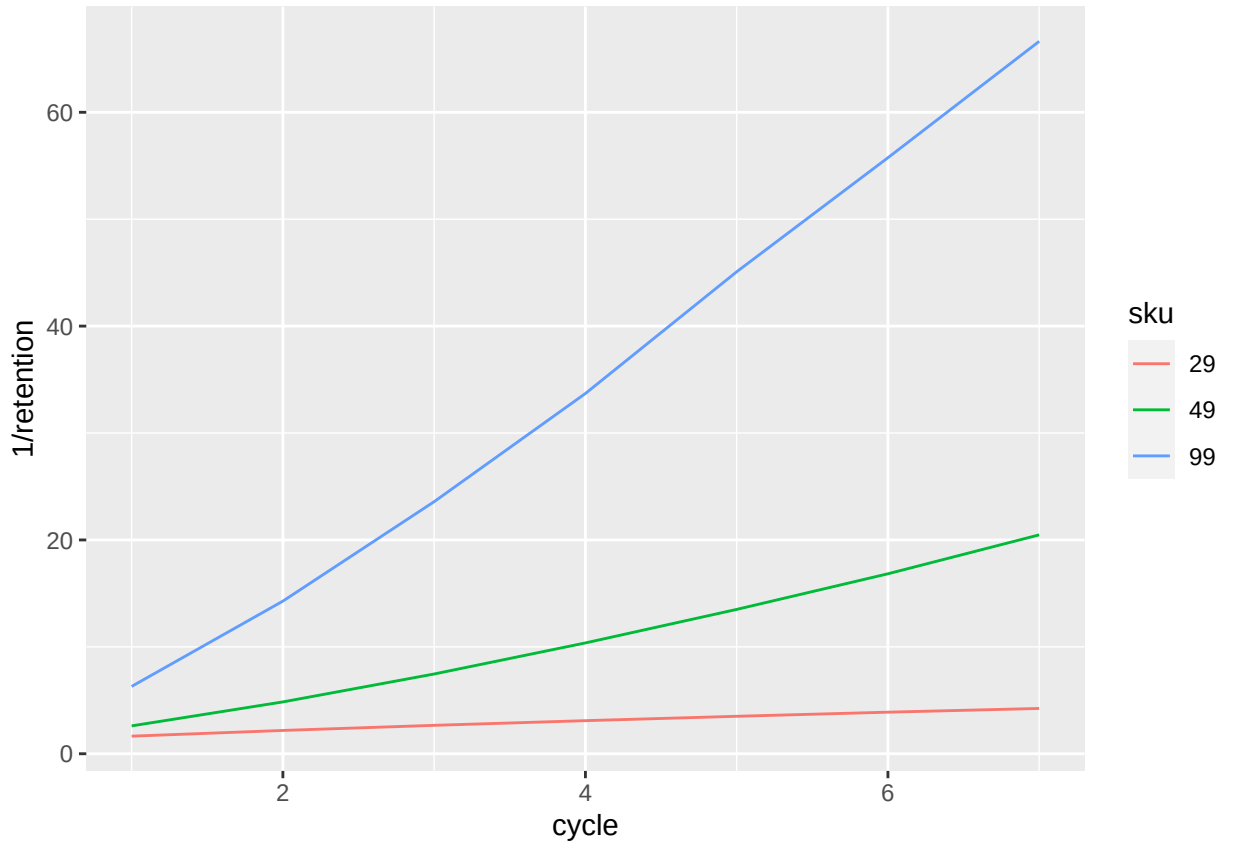
```
## Warning: Removed 18 rows containing missing values (`geom_line()`).
```



In our example, although the polynomial fits almost perfectly with our data, the prediction goes wild and nonsensical.

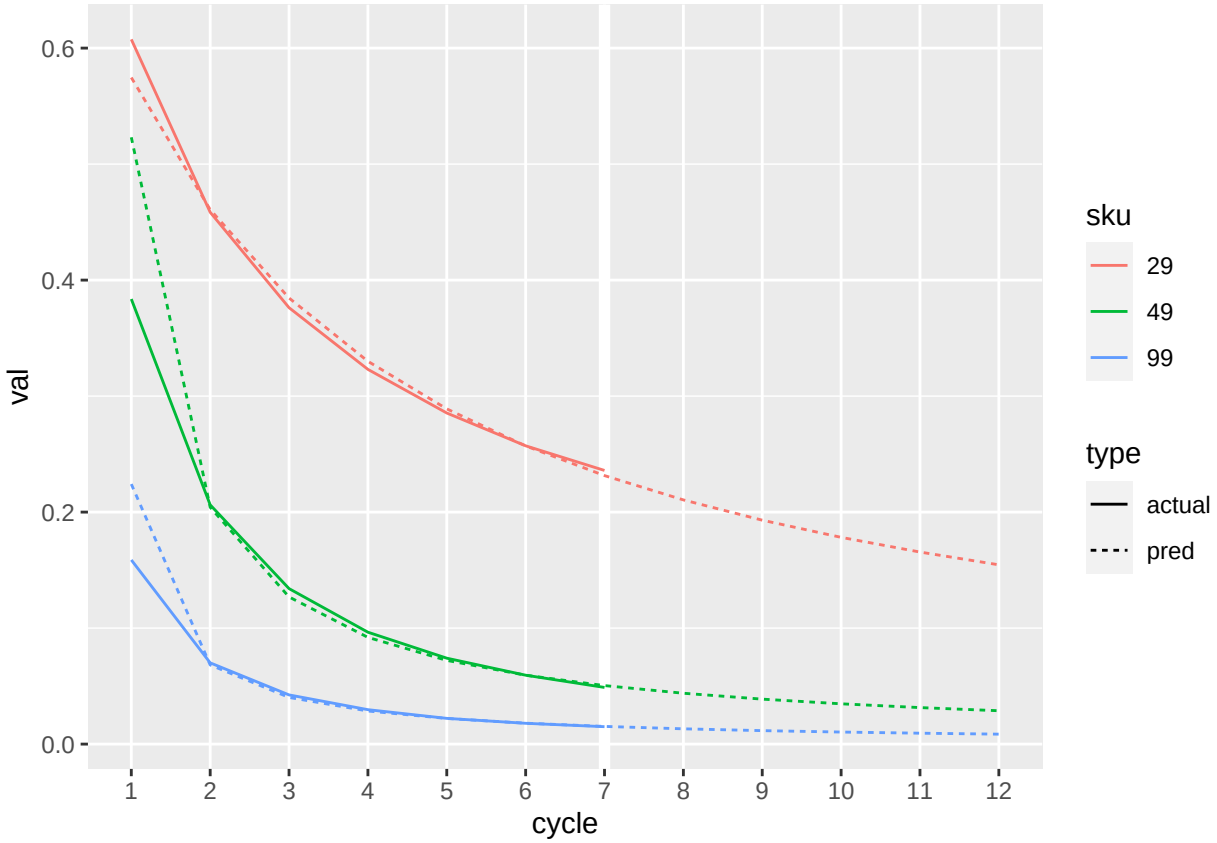
A better approach is to try different transformations for x and y and see if we can get a linear relationship. For this data, it turns that we only need to take the reciprocal of retnetion and we get a fairly linear relationship.

```
retention %>%
  ggplot(aes(cycle, 1 / retention)) +
  geom_line(aes(color = sku))
```




```
retention %>%
  transmute(sku, cycle, retention) %>%
  rbind(data.frame(sku = "29", cycle = 7:12, retention = NA_real_)) %>%
  rbind(data.frame(sku = "49", cycle = 7:12, retention = NA_real_)) %>%
  rbind(data.frame(sku = "99", cycle = 7:12, retention = NA_real_)) %>%
  add_predictions(
    lm(1 / retention ~ cycle * sku, data = retention)
  ) %>%
  mutate(pred = 1 / pred) %>%
  rename(actual = retention) %>%
  pivot_longer(actual:pred, names_to = "type", values_to = "val") %>%
  ggplot() +
  geom_vline(xintercept = 7, color = "white", size = 2) +
  geom_line(aes(cycle, val, color = sku, linetype = type)) +
  scale_x_continuous(breaks = seq(1, 12, 1), minor_breaks = NULL)
```

```
## Warning: Removed 18 rows containing missing values (`geom_line()`).
```



Both the fitting and prediction look reasonable with this approach so we should feel confident about the results from this model.