GM

```
df = read.csv("rote_test.csv")
head(df)
```

```
##                         session_id session_start_time cluster treat       test
## 1 0B7865D5822D8D87F75FB5BAFB2102DD       1.595626e+12       1 false   baseline
## 2 8080E2E61BA04074F123155741AC29DC       1.595626e+12       2 false   baseline
## 3 0B7865D5822D8D87F75FB5BAFB2102DD       1.595626e+12       1 false experiment
## 4 8080E2E61BA04074F123155741AC29DC       1.595626e+12       2 false experiment
## 5 BE7C20F73505C684DB5613B8702BD522       1.595639e+12       1 false   baseline
## 6 BE7C20F73505C684DB5613B8702BD522       1.595640e+12       2  true   baseline
##   test_submit_time item_id1 a11 c11 a12 c12 a13 c13 a14 c14 item_id2 a21 c21
## 1     1.595626e+12        3   3   5   4   4   5   5   5   5        4   2   2
## 2     1.595627e+12        8   3   3   1   1   5   5   2   2        5   2   1
## 3     1.595627e+12        5   1   1  NA   1   4   4   3   3        6   3   3
## 4     1.595627e+12        7   2   2   2   2   4   4   3   3        4   2   2
## 5     1.595640e+12        3   3   5   4   4   4   5   5   5        7   2   2
## 6     1.595640e+12        5   1   1   1   1   4   4   3   3        2   1   1
##   a22 c22 a23 c23 a24 c24 blank_column
## 1   2   2   1   1   3   3           NA
## 2   1   1   4   4   3   3           NA
## 3   4   4   1   1   2   2           NA
## 4   2   2   1   1   3   3           NA
## 5   2   2   4   4   3   3           NA
## 6   3   3   3   3   3   3           NA
```

```
df_cov = read.csv(("rote_cov.csv"))
head(df_cov)
```

```
##                         session_id session_start_time cluster treat
## 1 0B7865D5822D8D87F75FB5BAFB2102DD       1.595626e+12       1 false
## 2 8080E2E61BA04074F123155741AC29DC       1.595626e+12       2 false
## 3 BE7C20F73505C684DB5613B8702BD522       1.595639e+12       1 false
## 4 BE7C20F73505C684DB5613B8702BD522       1.595639e+12       1 false
## 5 BE7C20F73505C684DB5613B8702BD522       1.595640e+12       2  true
## 6 064D45ABDE08A0D54486ED13C1D68AF8       1.595643e+12       3 false
##   cov_submit_time age gender practice reading item_id1 knowledge1 item_id2
## 1    1.595626e+12  13   <NA>        4       5        3          5        4
## 2    1.595626e+12  45      M        2       3        8          2        5
## 3    1.595639e+12  28      F        2       2        3          1        7
## 4    1.595639e+12  28      F        2       2        3          1        7
## 5    1.595640e+12  28      F        2       2        5          1        2
## 6    1.595643e+12  16      F        5       4        6          1        8
##   knowledge2 item_id3 knowledge3 item_id4 knowledge4
## 1          1        5          2        6          1
## 2          2        7          2        4          3
## 3          1        2          1        5          1
## 4          1        2          1        5          1
```

```
## 5          1        7          1        1          1
## 6          3        3          3        2          1
```

```
df_b = df[df$test == 'baseline',]
df_e = df[df$test == 'experiment',]
```

```
sum(duplicated(df_cov$session_id))
```

```
## [1] 9
```

```
df_cov = df_cov[!duplicated(df_cov$session_id),]
nrow(df_cov)
```

```
## [1] 171
```

```
sum(duplicated(df_b$session_id))
```

```
## [1] 2
```

```
df_b = df_b[!duplicated(df_b$session_id),]
nrow(df_b)
```

```
## [1] 123
```

```
sum(duplicated(df_e$session_id))
```

```
## [1] 1
```

```
df_e = df_e[!duplicated(df_e$session_id),]
nrow(df_e)
```

```
## [1] 108
```

```
df2 = left_join(df_b, df_e, by='session_id')
nrow(df2)
```

```
## [1] 123
```

```
df3 = left_join(df_cov, df2, by='session_id')
nrow(df3)
```

```
## [1] 171
```

```
df3$completed = as.factor(ifelse(!is.na(df3$a11.y), "completed", ifelse(!is.na(df3$a11.x), "baseline", 
summary(df3$completed)
```

```
##   baseline  completed covariates
##         16        107         48
```

```
df3$treat = as.numeric(df3$treat)
str(df3)
```

```
## 'data.frame':    171 obs. of  66 variables:
##  $ session_id          : Factor w/ 171 levels "0035AF289E4C2D1138C7604D6E6F38DD",..: 9 108 138 6 118
##  $ session_start_time  : num  1.6e+12 1.6e+12 1.6e+12 1.6e+12 1.6e+12 ...
##  $ cluster             : int  1 2 1 3 4 4 4 5 6 7 ...
##  $ treat               : num  1 1 1 1 2 2 2 2 2 1 ...
##  $ cov_submit_time     : num  1.6e+12 1.6e+12 1.6e+12 1.6e+12 1.6e+12 ...
##  $ age                 : Factor w/ 58 levels "","?","106","11",..: 6 28 17 9 5 9 30 30 37 30 ...
##  $ gender              : Factor w/ 3 levels "F","M","other": NA 2 1 1 1 2 1 1 1 2 ...
##  $ practice            : int  4 2 2 5 3 3 2 1 3 2 ...
##  $ reading             : int  5 3 2 4 3 3 3 2 3 3 ...
```

```
##  $ item_id1            : int  3 8 3 6 6 4 4 3 7 7 ...
##  $ knowledge1          : int  5 2 1 1 1 1 1 1 3 3 ...
##  $ item_id2            : int  4 5 7 8 1 5 6 5 8 3 ...
##  $ knowledge2          : int  1 2 1 3 1 3 1 2 3 1 ...
##  $ item_id3            : int  5 7 2 3 8 1 5 1 1 6 ...
##  $ knowledge3          : int  2 2 1 3 1 2 2 2 3 2 ...
##  $ item_id4            : int  6 4 5 2 2 2 1 2 5 5 ...
##  $ knowledge4          : int  1 3 1 1 1 1 3 1 3 3 ...
##  $ session_start_time.x: num  1.6e+12 1.6e+12 1.6e+12 1.6e+12 1.6e+12 ...
##  $ cluster.x           : int  1 2 1 3 4 4 NA 5 NA NA ...
##  $ treat.x             : Factor w/ 2 levels "false","true": 1 1 1 1 2 2 NA 2 NA NA ...
##  $ test.x              : Factor w/ 2 levels "baseline","experiment": 1 1 1 1 1 1 NA 1 NA NA ...
##  $ test_submit_time.x  : num  1.6e+12 1.6e+12 1.6e+12 1.6e+12 1.6e+12 ...
##  $ item_id1.x          : int  3 8 3 6 6 4 NA 3 NA NA ...
##  $ a11.x               : int  3 3 3 6 3 2 NA 5 NA NA ...
##  $ c11.x               : int  5 3 5 3 3 2 NA 5 NA NA ...
##  $ a12.x               : int  4 1 4 4 6 2 NA 4 NA NA ...
##  $ c12.x               : int  4 1 4 4 4 2 NA 4 NA NA ...
##  $ a13.x               : int  5 5 4 1 1 5 NA 4 NA NA ...
##  $ c13.x               : int  5 5 5 1 1 1 NA 5 NA NA ...
##  $ a14.x               : int  5 2 5 6 2 3 NA 1 NA NA ...
##  $ c14.x               : int  5 2 5 2 2 3 NA 5 NA NA ...
##  $ item_id2.x          : int  4 5 7 8 1 5 NA 5 NA NA ...
##  $ a21.x               : int  2 2 2 3 4 1 NA 1 NA NA ...
##  $ c21.x               : int  2 1 2 3 5 1 NA 1 NA NA ...
##  $ a22.x               : int  2 1 2 1 2 1 NA 1 NA NA ...
##  $ c22.x               : int  2 1 2 1 1 1 NA 1 NA NA ...
##  $ a23.x               : int  1 4 4 5 6 4 NA 4 NA NA ...
##  $ c23.x               : int  1 4 4 5 3 4 NA 4 NA NA ...
##  $ a24.x               : int  3 3 3 2 6 3 NA 3 NA NA ...
##  $ c24.x               : int  3 3 3 2 2 3 NA 3 NA NA ...
##  $ blank_column.x      : logi  NA NA NA NA NA NA ...
##  $ session_start_time.y: num  1.6e+12 1.6e+12 1.6e+12 1.6e+12 1.6e+12 ...
##  $ cluster.y           : int  1 2 2 3 4 4 NA 5 NA NA ...
##  $ treat.y             : Factor w/ 2 levels "false","true": 1 1 2 1 2 2 NA 2 NA NA ...
##  $ test.y              : Factor w/ 2 levels "baseline","experiment": 2 2 2 2 2 2 NA 2 NA NA ...
##  $ test_submit_time.y  : num  1.6e+12 1.6e+12 1.6e+12 1.6e+12 1.6e+12 ...
##  $ item_id1.y          : int  5 7 7 3 8 1 NA 1 NA NA ...
##  $ a11.y               : int  1 2 2 3 3 2 NA 2 NA NA ...
##  $ c11.y               : int  1 2 2 5 3 5 NA 5 NA NA ...
##  $ a12.y               : int  NA 2 2 4 1 1 NA 1 NA NA ...
##  $ c12.y               : int  1 2 2 4 1 1 NA 1 NA NA ...
##  $ a13.y               : int  4 4 4 5 2 3 NA 3 NA NA ...
##  $ c13.y               : int  4 4 4 5 5 3 NA 3 NA NA ...
##  $ a14.y               : int  3 3 3 5 2 2 NA 2 NA NA ...
##  $ c14.y               : int  3 3 3 5 2 2 NA 2 NA NA ...
##  $ item_id2.y          : int  6 4 1 2 2 2 NA 2 NA NA ...
##  $ a21.y               : int  3 2 2 1 1 1 NA 1 NA NA ...
##  $ c21.y               : int  3 2 5 1 1 1 NA 1 NA NA ...
##  $ a22.y               : int  4 2 1 3 3 3 NA 3 NA NA ...
##  $ c22.y               : int  4 2 1 3 3 3 NA 3 NA NA ...
##  $ a23.y               : int  1 1 3 3 3 3 NA 3 NA NA ...
##  $ c23.y               : int  1 1 3 3 3 3 NA 3 NA NA ...
##  $ a24.y               : int  2 3 2 3 3 3 NA 3 NA NA ...
```

```
##  $ c24.y               : int  2 3 2 3 3 3 NA 3 NA NA ...
##  $ blank_column.y      : logi  NA NA NA NA NA NA ...
##  $ completed           : Factor w/ 3 levels "baseline","completed",..: 2 2 2 2 2 2 3 2 3 3 ...
```

```
df3$completed = factor(df3$completed, levels=c("started", "covariates", "baseline", "completed"))
```

```
log = read.csv("log2.txt")
colnames(log) = c("session_id", "time")
log = log[!duplicated(log$session_id),]
nrow(log)
```

```
## [1] 218
```

```
df_a = data.frame(id=1:4,stage=c("started", "covariates", "baseline", "complete"), n=c(218, 171, 123, 1
df_a
```

```
##   id      stage   n
## 1  1    started 218
## 2  2 covariates 171
## 3  3   baseline 123
## 4  4   complete 108
```

```
str(df3)
```

```
## 'data.frame':    171 obs. of  66 variables:
##  $ session_id          : Factor w/ 171 levels "0035AF289E4C2D1138C7604D6E6F38DD",..: 9 108 138 6 118
##  $ session_start_time  : num  1.6e+12 1.6e+12 1.6e+12 1.6e+12 1.6e+12 ...
##  $ cluster             : int  1 2 1 3 4 4 4 5 6 7 ...
##  $ treat               : num  1 1 1 1 2 2 2 2 2 1 ...
##  $ cov_submit_time     : num  1.6e+12 1.6e+12 1.6e+12 1.6e+12 1.6e+12 ...
##  $ age                 : Factor w/ 58 levels "","?","106","11",..: 6 28 17 9 5 9 30 30 37 30 ...
##  $ gender              : Factor w/ 3 levels "F","M","other": NA 2 1 1 1 2 1 1 1 2 ...
##  $ practice            : int  4 2 2 5 3 3 2 1 3 2 ...
##  $ reading             : int  5 3 2 4 3 3 3 2 3 3 ...
##  $ item_id1            : int  3 8 3 6 6 4 4 3 7 7 ...
##  $ knowledge1          : int  5 2 1 1 1 1 1 1 3 3 ...
##  $ item_id2            : int  4 5 7 8 1 5 6 5 8 3 ...
##  $ knowledge2          : int  1 2 1 3 1 3 1 2 3 1 ...
##  $ item_id3            : int  5 7 2 3 8 1 5 1 1 6 ...
##  $ knowledge3          : int  2 2 1 3 1 2 2 2 3 2 ...
##  $ item_id4            : int  6 4 5 2 2 2 1 2 5 5 ...
##  $ knowledge4          : int  1 3 1 1 1 1 3 1 3 3 ...
##  $ session_start_time.x: num  1.6e+12 1.6e+12 1.6e+12 1.6e+12 1.6e+12 ...
##  $ cluster.x           : int  1 2 1 3 4 4 NA 5 NA NA ...
##  $ treat.x             : Factor w/ 2 levels "false","true": 1 1 1 1 2 2 NA 2 NA NA ...
##  $ test.x              : Factor w/ 2 levels "baseline","experiment": 1 1 1 1 1 1 NA 1 NA NA ...
##  $ test_submit_time.x  : num  1.6e+12 1.6e+12 1.6e+12 1.6e+12 1.6e+12 ...
##  $ item_id1.x          : int  3 8 3 6 6 4 NA 3 NA NA ...
##  $ a11.x               : int  3 3 3 6 3 2 NA 5 NA NA ...
##  $ c11.x               : int  5 3 5 3 3 2 NA 5 NA NA ...
##  $ a12.x               : int  4 1 4 4 6 2 NA 4 NA NA ...
##  $ c12.x               : int  4 1 4 4 4 2 NA 4 NA NA ...
##  $ a13.x               : int  5 5 4 1 1 5 NA 4 NA NA ...
##  $ c13.x               : int  5 5 5 1 1 1 NA 5 NA NA ...
##  $ a14.x               : int  5 2 5 6 2 3 NA 1 NA NA ...
##  $ c14.x               : int  5 2 5 2 2 3 NA 5 NA NA ...
```

```
##  $ item_id2.x          : int  4 5 7 8 1 5 NA 5 NA NA ...
##  $ a21.x               : int  2 2 2 3 4 1 NA 1 NA NA ...
##  $ c21.x               : int  2 1 2 3 5 1 NA 1 NA NA ...
##  $ a22.x               : int  2 1 2 1 2 1 NA 1 NA NA ...
##  $ c22.x               : int  2 1 2 1 1 1 NA 1 NA NA ...
##  $ a23.x               : int  1 4 4 5 6 4 NA 4 NA NA ...
##  $ c23.x               : int  1 4 4 5 3 4 NA 4 NA NA ...
##  $ a24.x               : int  3 3 3 2 6 3 NA 3 NA NA ...
##  $ c24.x               : int  3 3 3 2 2 3 NA 3 NA NA ...
##  $ blank_column.x      : logi  NA NA NA NA NA NA ...
##  $ session_start_time.y: num  1.6e+12 1.6e+12 1.6e+12 1.6e+12 1.6e+12 ...
##  $ cluster.y           : int  1 2 2 3 4 4 NA 5 NA NA ...
##  $ treat.y             : Factor w/ 2 levels "false","true": 1 1 2 1 2 2 NA 2 NA NA ...
##  $ test.y              : Factor w/ 2 levels "baseline","experiment": 2 2 2 2 2 2 NA 2 NA NA ...
##  $ test_submit_time.y  : num  1.6e+12 1.6e+12 1.6e+12 1.6e+12 1.6e+12 ...
##  $ item_id1.y          : int  5 7 7 3 8 1 NA 1 NA NA ...
##  $ a11.y               : int  1 2 2 3 3 2 NA 2 NA NA ...
##  $ c11.y               : int  1 2 2 5 3 5 NA 5 NA NA ...
##  $ a12.y               : int  NA 2 2 4 1 1 NA 1 NA NA ...
##  $ c12.y               : int  1 2 2 4 1 1 NA 1 NA NA ...
##  $ a13.y               : int  4 4 4 5 2 3 NA 3 NA NA ...
##  $ c13.y               : int  4 4 4 5 5 3 NA 3 NA NA ...
##  $ a14.y               : int  3 3 3 5 2 2 NA 2 NA NA ...
##  $ c14.y               : int  3 3 3 5 2 2 NA 2 NA NA ...
##  $ item_id2.y          : int  6 4 1 2 2 2 NA 2 NA NA ...
##  $ a21.y               : int  3 2 2 1 1 1 NA 1 NA NA ...
##  $ c21.y               : int  3 2 5 1 1 1 NA 1 NA NA ...
##  $ a22.y               : int  4 2 1 3 3 3 NA 3 NA NA ...
##  $ c22.y               : int  4 2 1 3 3 3 NA 3 NA NA ...
##  $ a23.y               : int  1 1 3 3 3 3 NA 3 NA NA ...
##  $ c23.y               : int  1 1 3 3 3 3 NA 3 NA NA ...
##  $ a24.y               : int  2 3 2 3 3 3 NA 3 NA NA ...
##  $ c24.y               : int  2 3 2 3 3 3 NA 3 NA NA ...
##  $ blank_column.y      : logi  NA NA NA NA NA NA ...
##  $ completed           : Factor w/ 4 levels "started","covariates",..: 4 4 4 4 4 4 2 4 2 2 ...
```

```
n=218-171
df_started = data.frame(session_id = as.factor(1:n), treat=3, completed=factor("started"))
str(df_started)
```

```
## 'data.frame':    47 obs. of  3 variables:
##  $ session_id: Factor w/ 47 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ treat     : num  3 3 3 3 3 3 3 3 3 3 ...
##  $ completed : Factor w/ 1 level "started": 1 1 1 1 1 1 1 1 1 1 ...
```

```
df4 = full_join(df3, df_started, by="session_id")
str(df4)
```

```
## 'data.frame':    218 obs. of  68 variables:
##  $ session_id          : Factor w/ 218 levels "0035AF289E4C2D1138C7604D6E6F38DD",..: 9 108 138 6 118
##  $ session_start_time  : num  1.6e+12 1.6e+12 1.6e+12 1.6e+12 1.6e+12 ...
##  $ cluster             : int  1 2 1 3 4 4 4 5 6 7 ...
##  $ treat.x             : num  1 1 1 1 2 2 2 2 2 1 ...
##  $ cov_submit_time     : num  1.6e+12 1.6e+12 1.6e+12 1.6e+12 1.6e+12 ...
##  $ age                 : Factor w/ 58 levels "","?","106","11",..: 6 28 17 9 5 9 30 30 37 30 ...
##  $ gender              : Factor w/ 3 levels "F","M","other": NA 2 1 1 1 2 1 1 1 2 ...
```

```
##  $ practice            : int  4 2 2 5 3 3 2 1 3 2 ...
##  $ reading             : int  5 3 2 4 3 3 3 2 3 3 ...
##  $ item_id1            : int  3 8 3 6 6 4 4 3 7 7 ...
##  $ knowledge1          : int  5 2 1 1 1 1 1 1 3 3 ...
##  $ item_id2            : int  4 5 7 8 1 5 6 5 8 3 ...
##  $ knowledge2          : int  1 2 1 3 1 3 1 2 3 1 ...
##  $ item_id3            : int  5 7 2 3 8 1 5 1 1 6 ...
##  $ knowledge3          : int  2 2 1 3 1 2 2 2 3 2 ...
##  $ item_id4            : int  6 4 5 2 2 2 1 2 5 5 ...
##  $ knowledge4          : int  1 3 1 1 1 1 3 1 3 3 ...
##  $ session_start_time.x: num  1.6e+12 1.6e+12 1.6e+12 1.6e+12 1.6e+12 ...
##  $ cluster.x           : int  1 2 1 3 4 4 NA 5 NA NA ...
##  $ treat.x.x           : Factor w/ 2 levels "false","true": 1 1 1 1 2 2 NA 2 NA NA ...
##  $ test.x              : Factor w/ 2 levels "baseline","experiment": 1 1 1 1 1 1 NA 1 NA NA ...
##  $ test_submit_time.x  : num  1.6e+12 1.6e+12 1.6e+12 1.6e+12 1.6e+12 ...
##  $ item_id1.x          : int  3 8 3 6 6 4 NA 3 NA NA ...
##  $ a11.x               : int  3 3 3 6 3 2 NA 5 NA NA ...
##  $ c11.x               : int  5 3 5 3 3 2 NA 5 NA NA ...
##  $ a12.x               : int  4 1 4 4 6 2 NA 4 NA NA ...
##  $ c12.x               : int  4 1 4 4 4 2 NA 4 NA NA ...
##  $ a13.x               : int  5 5 4 1 1 5 NA 4 NA NA ...
##  $ c13.x               : int  5 5 5 1 1 1 NA 5 NA NA ...
##  $ a14.x               : int  5 2 5 6 2 3 NA 1 NA NA ...
##  $ c14.x               : int  5 2 5 2 2 3 NA 5 NA NA ...
##  $ item_id2.x          : int  4 5 7 8 1 5 NA 5 NA NA ...
##  $ a21.x               : int  2 2 2 3 4 1 NA 1 NA NA ...
##  $ c21.x               : int  2 1 2 3 5 1 NA 1 NA NA ...
##  $ a22.x               : int  2 1 2 1 2 1 NA 1 NA NA ...
##  $ c22.x               : int  2 1 2 1 1 1 NA 1 NA NA ...
##  $ a23.x               : int  1 4 4 5 6 4 NA 4 NA NA ...
##  $ c23.x               : int  1 4 4 5 3 4 NA 4 NA NA ...
##  $ a24.x               : int  3 3 3 2 6 3 NA 3 NA NA ...
##  $ c24.x               : int  3 3 3 2 2 3 NA 3 NA NA ...
##  $ blank_column.x      : logi  NA NA NA NA NA NA ...
##  $ session_start_time.y: num  1.6e+12 1.6e+12 1.6e+12 1.6e+12 1.6e+12 ...
##  $ cluster.y           : int  1 2 2 3 4 4 NA 5 NA NA ...
##  $ treat.y             : Factor w/ 2 levels "false","true": 1 1 2 1 2 2 NA 2 NA NA ...
##  $ test.y              : Factor w/ 2 levels "baseline","experiment": 2 2 2 2 2 2 NA 2 NA NA ...
##  $ test_submit_time.y  : num  1.6e+12 1.6e+12 1.6e+12 1.6e+12 1.6e+12 ...
##  $ item_id1.y          : int  5 7 7 3 8 1 NA 1 NA NA ...
##  $ a11.y               : int  1 2 2 3 3 2 NA 2 NA NA ...
##  $ c11.y               : int  1 2 2 5 3 5 NA 5 NA NA ...
##  $ a12.y               : int  NA 2 2 4 1 1 NA 1 NA NA ...
##  $ c12.y               : int  1 2 2 4 1 1 NA 1 NA NA ...
##  $ a13.y               : int  4 4 4 5 2 3 NA 3 NA NA ...
##  $ c13.y               : int  4 4 4 5 5 3 NA 3 NA NA ...
##  $ a14.y               : int  3 3 3 5 2 2 NA 2 NA NA ...
##  $ c14.y               : int  3 3 3 5 2 2 NA 2 NA NA ...
##  $ item_id2.y          : int  6 4 1 2 2 2 NA 2 NA NA ...
##  $ a21.y               : int  3 2 2 1 1 1 NA 1 NA NA ...
##  $ c21.y               : int  3 2 5 1 1 1 NA 1 NA NA ...
##  $ a22.y               : int  4 2 1 3 3 3 NA 3 NA NA ...
##  $ c22.y               : int  4 2 1 3 3 3 NA 3 NA NA ...
##  $ a23.y               : int  1 1 3 3 3 3 NA 3 NA NA ...
```

```
##  $ c23.y               : int  1 1 3 3 3 3 NA 3 NA NA ...
##  $ a24.y               : int  2 3 2 3 3 3 NA 3 NA NA ...
##  $ c24.y               : int  2 3 2 3 3 3 NA 3 NA NA ...
##  $ blank_column.y      : logi  NA NA NA NA NA NA ...
##  $ completed.x         : Factor w/ 4 levels "started","covariates",..: 4 4 4 4 4 4 2 4 2 2 ...
##  $ treat.y.y           : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ completed.y         : Factor w/ 1 level "started": NA NA NA NA NA NA NA NA NA NA ...
```

```
df4$treat = ifelse(is.na(df4$treat.x), df4$treat.y.y, df4$treat.x)
df4$completed = as.factor(ifelse(is.na(df4$completed.x), df4$completed.y, df4$completed.x))
```

```
summary(df4$completed)
```

```
##   1   2   3   4
##  47  48  16 107
```

```
str(df4)
```

```
## 'data.frame':    218 obs. of  70 variables:
##  $ session_id          : Factor w/ 218 levels "0035AF289E4C2D1138C7604D6E6F38DD",..: 9 108 138 6 118
##  $ session_start_time  : num  1.6e+12 1.6e+12 1.6e+12 1.6e+12 1.6e+12 ...
##  $ cluster             : int  1 2 1 3 4 4 4 5 6 7 ...
##  $ treat.x             : num  1 1 1 1 2 2 2 2 2 1 ...
##  $ cov_submit_time     : num  1.6e+12 1.6e+12 1.6e+12 1.6e+12 1.6e+12 ...
##  $ age                 : Factor w/ 58 levels "","?","106","11",..: 6 28 17 9 5 9 30 30 37 30 ...
##  $ gender              : Factor w/ 3 levels "F","M","other": NA 2 1 1 1 2 1 1 1 2 ...
##  $ practice            : int  4 2 2 5 3 3 2 1 3 2 ...
##  $ reading             : int  5 3 2 4 3 3 3 2 3 3 ...
##  $ item_id1            : int  3 8 3 6 6 4 4 3 7 7 ...
##  $ knowledge1          : int  5 2 1 1 1 1 1 1 3 3 ...
##  $ item_id2            : int  4 5 7 8 1 5 6 5 8 3 ...
##  $ knowledge2          : int  1 2 1 3 1 3 1 2 3 1 ...
##  $ item_id3            : int  5 7 2 3 8 1 5 1 1 6 ...
##  $ knowledge3          : int  2 2 1 3 1 2 2 2 3 2 ...
##  $ item_id4            : int  6 4 5 2 2 2 1 2 5 5 ...
##  $ knowledge4          : int  1 3 1 1 1 1 3 1 3 3 ...
##  $ session_start_time.x: num  1.6e+12 1.6e+12 1.6e+12 1.6e+12 1.6e+12 ...
##  $ cluster.x           : int  1 2 1 3 4 4 NA 5 NA NA ...
##  $ treat.x.x           : Factor w/ 2 levels "false","true": 1 1 1 1 2 2 NA 2 NA NA ...
##  $ test.x              : Factor w/ 2 levels "baseline","experiment": 1 1 1 1 1 1 NA 1 NA NA ...
##  $ test_submit_time.x  : num  1.6e+12 1.6e+12 1.6e+12 1.6e+12 1.6e+12 ...
##  $ item_id1.x          : int  3 8 3 6 6 4 NA 3 NA NA ...
##  $ a11.x               : int  3 3 3 6 3 2 NA 5 NA NA ...
##  $ c11.x               : int  5 3 5 3 3 2 NA 5 NA NA ...
##  $ a12.x               : int  4 1 4 4 6 2 NA 4 NA NA ...
##  $ c12.x               : int  4 1 4 4 4 2 NA 4 NA NA ...
##  $ a13.x               : int  5 5 4 1 1 5 NA 4 NA NA ...
##  $ c13.x               : int  5 5 5 1 1 1 NA 5 NA NA ...
##  $ a14.x               : int  5 2 5 6 2 3 NA 1 NA NA ...
##  $ c14.x               : int  5 2 5 2 2 3 NA 5 NA NA ...
##  $ item_id2.x          : int  4 5 7 8 1 5 NA 5 NA NA ...
##  $ a21.x               : int  2 2 2 3 4 1 NA 1 NA NA ...
##  $ c21.x               : int  2 1 2 3 5 1 NA 1 NA NA ...
##  $ a22.x               : int  2 1 2 1 2 1 NA 1 NA NA ...
##  $ c22.x               : int  2 1 2 1 1 1 NA 1 NA NA ...
##  $ a23.x               : int  1 4 4 5 6 4 NA 4 NA NA ...
```

```
##  $ c23.x               : int  1 4 4 5 3 4 NA 4 NA NA ...
##  $ a24.x               : int  3 3 3 2 6 3 NA 3 NA NA ...
##  $ c24.x               : int  3 3 3 2 2 3 NA 3 NA NA ...
##  $ blank_column.x      : logi  NA NA NA NA NA NA ...
##  $ session_start_time.y: num  1.6e+12 1.6e+12 1.6e+12 1.6e+12 1.6e+12 ...
##  $ cluster.y           : int  1 2 2 3 4 4 NA 5 NA NA ...
##  $ treat.y             : Factor w/ 2 levels "false","true": 1 1 2 1 2 2 NA 2 NA NA ...
##  $ test.y              : Factor w/ 2 levels "baseline","experiment": 2 2 2 2 2 2 NA 2 NA NA ...
##  $ test_submit_time.y  : num  1.6e+12 1.6e+12 1.6e+12 1.6e+12 1.6e+12 ...
##  $ item_id1.y          : int  5 7 7 3 8 1 NA 1 NA NA ...
##  $ a11.y               : int  1 2 2 3 3 2 NA 2 NA NA ...
##  $ c11.y               : int  1 2 2 5 3 5 NA 5 NA NA ...
##  $ a12.y               : int  NA 2 2 4 1 1 NA 1 NA NA ...
##  $ c12.y               : int  1 2 2 4 1 1 NA 1 NA NA ...
##  $ a13.y               : int  4 4 4 5 2 3 NA 3 NA NA ...
##  $ c13.y               : int  4 4 4 5 5 3 NA 3 NA NA ...
##  $ a14.y               : int  3 3 3 5 2 2 NA 2 NA NA ...
##  $ c14.y               : int  3 3 3 5 2 2 NA 2 NA NA ...
##  $ item_id2.y          : int  6 4 1 2 2 2 NA 2 NA NA ...
##  $ a21.y               : int  3 2 2 1 1 1 NA 1 NA NA ...
##  $ c21.y               : int  3 2 5 1 1 1 NA 1 NA NA ...
##  $ a22.y               : int  4 2 1 3 3 3 NA 3 NA NA ...
##  $ c22.y               : int  4 2 1 3 3 3 NA 3 NA NA ...
##  $ a23.y               : int  1 1 3 3 3 3 NA 3 NA NA ...
##  $ c23.y               : int  1 1 3 3 3 3 NA 3 NA NA ...
##  $ a24.y               : int  2 3 2 3 3 3 NA 3 NA NA ...
##  $ c24.y               : int  2 3 2 3 3 3 NA 3 NA NA ...
##  $ blank_column.y      : logi  NA NA NA NA NA NA ...
##  $ completed.x         : Factor w/ 4 levels "started","covariates",..: 4 4 4 4 4 4 2 4 2 2 ...
##  $ treat.y.y           : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ completed.y         : Factor w/ 1 level "started": NA NA NA NA NA NA NA NA NA NA ...
##  $ treat               : num  1 1 1 1 2 2 2 2 2 1 ...
##  $ completed           : Factor w/ 4 levels "1","2","3","4": 4 4 4 4 4 4 2 4 2 2 ...
```

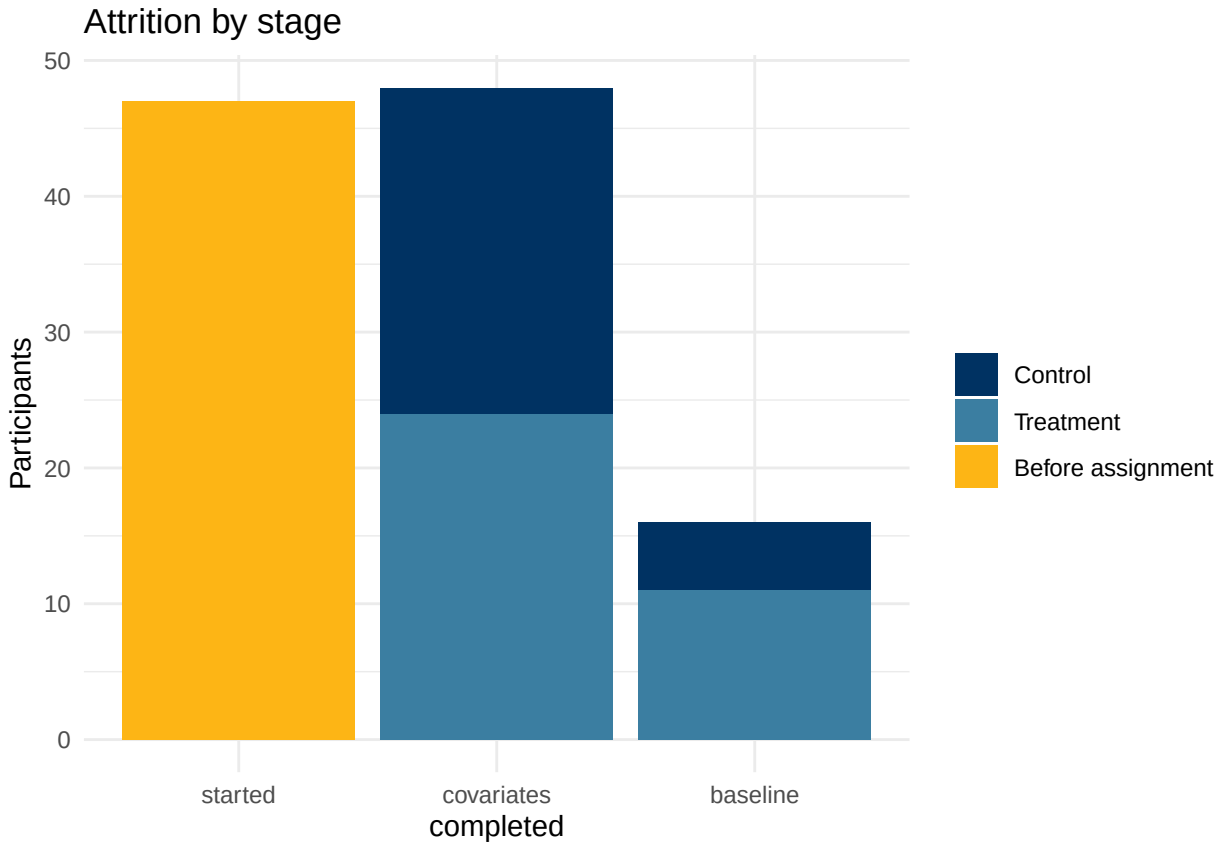
```
is.na(df4$treat)
```

```
##   [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [97] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [109] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [121] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [145] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [157] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [169] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [181] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [193] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [205] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [217] FALSE FALSE
```

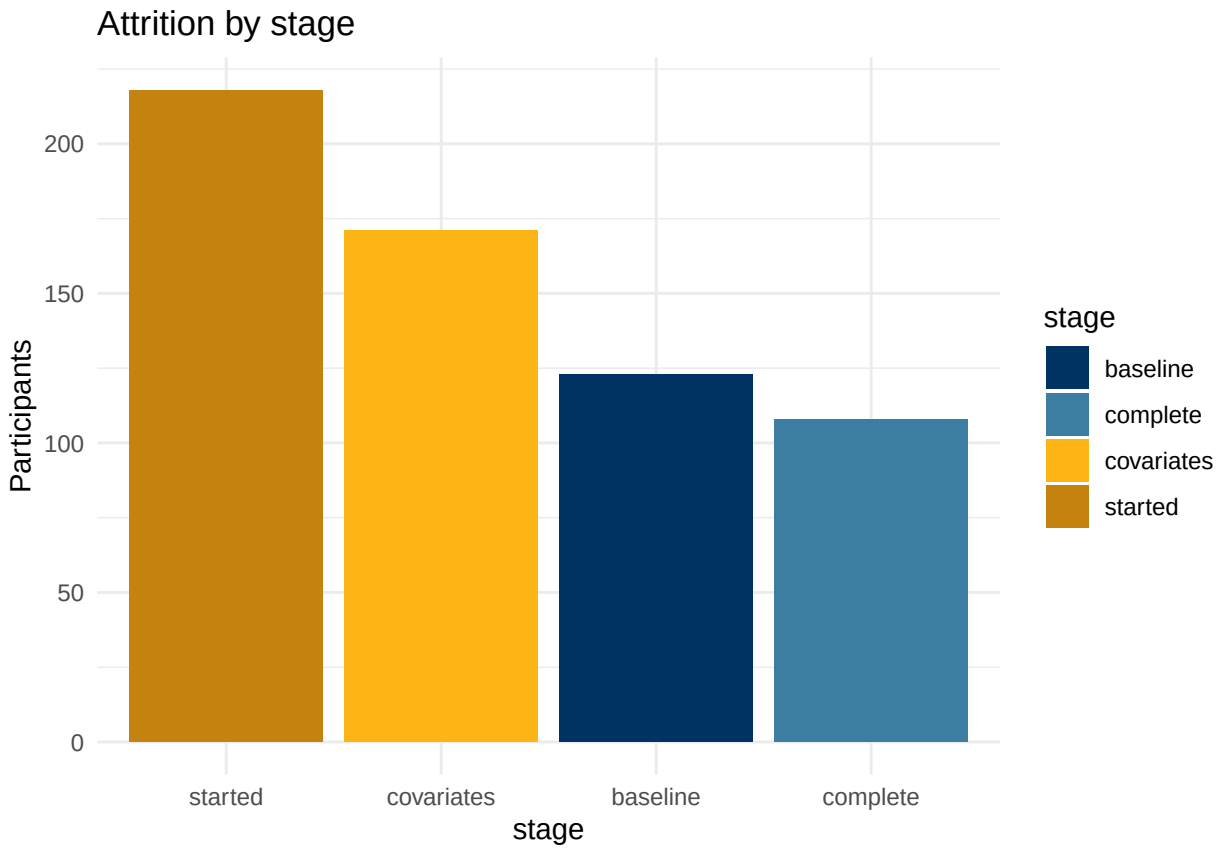
```
library(ggplot2)
df4 %>% filter(completed != 4) %>%
  ggplot(aes(x=completed, fill=as.factor(treat)))+
  geom_histogram(stat="count") +
  theme_minimal() +
  scale_x_discrete(labels=c("started", "covariates", "baseline", "complete")) +
  scale_fill_brewer(palette="Dark2") +
  scale_fill_manual(values=c("#003262", "#3B7EA1", "#FDB515", "#C4820E"),
                    labels=c("Control", "Treatment", "Before assignment"))+
  ylab("Participants") +
  ggtitle("Attrition by stage")  +
  theme(legend.title=element_blank())
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
## Scale for 'fill' is already present. Adding another scale for 'fill', which
## will replace the existing scale.
```




```
library(ggplot2)
df_a %>% ggplot(aes(x=stage, y=n, fill=stage))+
  geom_bar(stat="identity") +
  theme_minimal() +
  scale_x_discrete(limits=c("started", "covariates", "baseline", "complete")) +
  #scale_fill_brewer(palette="Dark2") +
  scale_fill_manual(values=c("#003262", "#3B7EA1", "#FDB515", "#C4820E"))+
  ylab("Participants") +
  ggtitle("Attrition by stage")
```




```
#str(df_cov)
library(stargazer)
```

```
##
## Please cite as:
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
df_cov$total_prior_knowledge = df_cov$knowledge1+df_cov$knowledge2+df_cov$knowledge3+df_cov$knowledge4
model = glm(treat~as.numeric(age)+gender+practice+reading+total_prior_knowledge, data=df_cov, family=bi
summary(model)
```

```
##
## Call:
## glm(formula = treat ~ as.numeric(age) + gender + practice + reading +
##     total_prior_knowledge, family = binomial(link = "logit"),
##     data = df_cov)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3466  -1.1913   0.9992   1.1545   1.3134
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -0.597085   0.819745  -0.728    0.466
## as.numeric(age)       -0.001914   0.010377  -0.184    0.854
## genderM               -0.226372   0.366925  -0.617    0.537
```

```
## genderother           -0.092063   1.430418  -0.064    0.949
## practice               0.029549   0.182874   0.162    0.872
## reading                0.121791   0.185597   0.656    0.512
## total_prior_knowledge  0.038501   0.078087   0.493    0.622
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 224.48  on 161  degrees of freedom
## Residual deviance: 223.09  on 155  degrees of freedom
##   (9 observations deleted due to missingness)
## AIC: 237.09
##
## Number of Fisher Scoring iterations: 3
```

```
stargazer(model, type="html")
```

```
##
## <table style="text-align:center"><tr><td colspan="2" style="border-bottom: 1px solid black"></td></t
## <tr><td></td><td colspan="1" style="border-bottom: 1px solid black"></td></tr>
## <tr><td style="text-align:left"></td><td>treat</td></tr>
## <tr><td colspan="2" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left"
## <tr><td style="text-align:left"></td><td>(0.010)</td></tr>
## <tr><td style="text-align:left"></td><td></td></tr>
## <tr><td style="text-align:left">genderM</td><td>-0.226</td></tr>
## <tr><td style="text-align:left"></td><td>(0.367)</td></tr>
## <tr><td style="text-align:left"></td><td></td></tr>
## <tr><td style="text-align:left">genderother</td><td>-0.092</td></tr>
## <tr><td style="text-align:left"></td><td>(1.430)</td></tr>
## <tr><td style="text-align:left"></td><td></td></tr>
## <tr><td style="text-align:left">practice</td><td>0.030</td></tr>
## <tr><td style="text-align:left"></td><td>(0.183)</td></tr>
## <tr><td style="text-align:left"></td><td></td></tr>
## <tr><td style="text-align:left">reading</td><td>0.122</td></tr>
## <tr><td style="text-align:left"></td><td>(0.186)</td></tr>
## <tr><td style="text-align:left"></td><td></td></tr>
## <tr><td style="text-align:left">total_prior_knowledge</td><td>0.039</td></tr>
## <tr><td style="text-align:left"></td><td>(0.078)</td></tr>
## <tr><td style="text-align:left"></td><td></td></tr>
## <tr><td style="text-align:left">Constant</td><td>-0.597</td></tr>
## <tr><td style="text-align:left"></td><td>(0.820)</td></tr>
## <tr><td style="text-align:left"></td><td></td></tr>
## <tr><td colspan="2" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left"
## <tr><td style="text-align:left">Log Likelihood</td><td>-111.547</td></tr>
## <tr><td style="text-align:left">Akaike Inf. Crit.</td><td>237.093</td></tr>
## <tr><td colspan="2" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left"
## </table>
```

Dependent variable:

treat

as.numeric(age)

-0.002

(0.010)

genderM

-0.226

(0.367)

genderother

-0.092

(1.430)

practice

0.030

(0.183)

reading

0.122

(0.186)

total_prior_knowledge

0.039

(0.078)

Constant

-0.597

(0.820)

Observations

162

Log Likelihood

-111.547

Akaike Inf. Crit.

237.093

Note:

p<0.1; ***p<0.05;*** p<0.01