Network Analysis of Clinical Interactions (NACI): Data Cleaning Log

Setup

```
# set options
# This is an example setup chunk from the N741 project
knitr::opts_chunk$set(root.dir = "~/Documents/1_Research/2_Data_Science/0_Projects/1_NACI/Data",
                      echo = TRUE,
                      message = FALSE,
                      warning = FALSE)
# options(na.action = na.warn)??
```

```
# Load packages
# library(igraph) # package for working with and visualizing network analysis objectve
library(haven) # package for importing SAS data files (i.e., ".sas7bdat")
library(tidyverse) # packages for data import, cleaning, transformation, and analysis
library(gt) # package for creating and formating latex tables
library(lubridate) # package for working with date data
# library(pander) # ????
# library(printr) # ????
# library(forcats) # package for making and working with factors
# library(modelr) # package for statistical modeling in r
library(readxl)
library(readr)
library(stringr)
library(labelled)
```

```
data_path <- paste0(getwd(), "/Data/")
# If you need to change the working directory, use `setwd(data_path)`
# Create a list of all items in the current working directory
files <- list.files(path = data_path)
# Print directory file list
writeLines(files)
```

```
## completepat.sas7bdat
## completestaff.sas7bdat
## Data_Files
## Data_Reference
## datafiles.numbers
## USB from George
```

Data Import & Cleaning

```
# ---- `pt_complete`
# 1a. read "completepat.sas7bdat",
pt_complete <-
  read_sas(paste0(data_path, "completepat.sas7bdat"))
```

```
require(lubridate)
# 2a. subset first 10 observations for data transformation code preparation
pt_head <- head(pt_complete) %>%
  # 3a. Pivot the data.frame from wide to long by placing all column names that start with "floc" into
  pivot_longer(cols = starts_with("floc"), names_to = "seconds", values_to = "location_num") %>%
  # 4a. Remove the prefix "floc" from `time_seconds` and keep the digits as `seconds`
  mutate(seconds = as.integer(str_replace(seconds, "floc", "")),
         shift_num_ampm = str_trim(shift_num_ampm),
         shift_num = as.integer(str_extract(shift_num_ampm, "[:digit:]+")),
         am_pm = str_extract(shift_num_ampm, "am|pm"),
         date = make_date(year = year, month = mon, day = day)) # %>%
  # 5. Filter out all rows for which no location was recorded
  # filter(!is.na(location)) %>%

# View data frame structure
# glimpse(pt_head)
pt_head
```

```
## # A tibble: 259,212 x 12
##    sid      shift_num_ampm    d8   day   mon  year firstday seconds location_num
##    <chr>    <chr>          <dbl> <dbl> <dbl> <dbl>    <dbl>   <int>        <dbl>
##  1 0002d045 1pm            18087     9     7  2009    18087       1           NA
##  2 0002d045 1pm            18087     9     7  2009    18087       2           NA
##  3 0002d045 1pm            18087     9     7  2009    18087       3           NA
##  4 0002d045 1pm            18087     9     7  2009    18087       4           NA
##  5 0002d045 1pm            18087     9     7  2009    18087       5           NA
##  6 0002d045 1pm            18087     9     7  2009    18087       6           NA
##  7 0002d045 1pm            18087     9     7  2009    18087       7           NA
##  8 0002d045 1pm            18087     9     7  2009    18087       8           NA
##  9 0002d045 1pm            18087     9     7  2009    18087       9           NA
## 10 0002d045 1pm            18087     9     7  2009    18087      10           NA
## # ... with 259,202 more rows, and 3 more variables: shift_num <int>,
## #   am_pm <chr>, date <date>
```

```
# ---- `staff_complete
# 1b. read "completestaff.sas7bdat"
staff_complete <-
  read_sas(paste0(data_path, "completestaff.sas7bdat"))
```

```
# 2a. subset first 10 observations for data transformation code preparation
staff_head <- head(staff_complete) %>%
  # 3a. Pivot the data.frame from wide to long by placing all column names that start with "floc" into
  pivot_longer(cols = starts_with("floc"), names_to = "seconds", values_to = "location_num") %>%
  # 4a. Remove the prefix "floc" from `time_seconds` and keep the digits as `seconds`
  mutate(seconds = as.integer(str_replace(seconds, "floc", "")),
         shift_num_ampm = str_trim(shift_num_ampm),
         shift_num = as.integer(str_extract(shift_num_ampm, "[:digit:]+")),
         am_pm = str_extract(shift_num_ampm, "am|pm"),
         date = make_date(year = year, month = mon, day = day)) # %>%
  # 5. Filter out all rows for which no location was recorded
  # filter(!is.na(location)) %>%
# View data frame structure
# str(staff_head)
# glimpse(staff_head)
staff_head
```

```
## # A tibble: 259,212 x 12
##    sid      d8           day  year shift_num_ampm   mon firstday seconds
##    <chr>    <date>     <dbl> <dbl> <chr>          <dbl>    <dbl>   <int>
##  1 0002f4e2 2009-07-09     9  2009 1pm                7    18087       1
##  2 0002f4e2 2009-07-09     9  2009 1pm                7    18087       2
##  3 0002f4e2 2009-07-09     9  2009 1pm                7    18087       3
##  4 0002f4e2 2009-07-09     9  2009 1pm                7    18087       4
##  5 0002f4e2 2009-07-09     9  2009 1pm                7    18087       5
##  6 0002f4e2 2009-07-09     9  2009 1pm                7    18087       6
##  7 0002f4e2 2009-07-09     9  2009 1pm                7    18087       7
##  8 0002f4e2 2009-07-09     9  2009 1pm                7    18087       8
##  9 0002f4e2 2009-07-09     9  2009 1pm                7    18087       9
## 10 0002f4e2 2009-07-09     9  2009 1pm                7    18087      10
## # ... with 259,202 more rows, and 4 more variables: location_num <dbl>,
## #   shift_num <int>, am_pm <chr>, date <date>
```

```
# ---- 'edge_list` ----
# 1c. read "allshifts_edges.sas7bdat"
edge_list <- read_sas(paste0(data_path, "Data_Files/allshifts_edges.sas7bdat"))
edge_list2 <- read_sas(paste0(data_path, "Data_Files/edges2.sas7bdat"))
```

```
# Print the first 6 observations of edge_list
head(edge_list)
```

```
## # A tibble: 6 x 15
##       i   any staffi idi     d8          H1N1 quarter shiftampm    d9 edgeweight
##   <dbl> <dbl>  <dbl> <chr>   <date>     <dbl>   <dbl>     <dbl> <dbl>      <dbl>
## 1     1     1      1 7920091 2009-07-09     0       1         2 18087      0.525
## 2     1     1      1 7920091 2009-07-09     0       1         2 18087      3.77
## 3     1     1      1 7920091 2009-07-09     0       1         2 18087      1.11
## 4     1     1      1 7920091 2009-07-09     0       1         2 18087      0.487
## 5     1     1      1 7920091 2009-07-09     0       1         2 18087      0.794
## 6     1     1      1 7920091 2009-07-09     0       1         2 18087      0.513
## # ... with 5 more variables: j <dbl>, staffj <dbl>, combo <dbl>, idj <chr>,
## #   comboc <chr>
```

```
# Print out the variable labels for all columns of edge_list
var_label(edge_list)
```

```
## $i
## [1] "one member of contact pair (find real id using id_sid_matchuplist)"
##
## $any
## [1] "any contact 1yes 0no"
##
## $staffi
## [1] "i is a staff member 1yes 0no"
##
## $idi
## [1] "id for i made of d8 and i"
##
## $d8
## [1] "1st d8 in the shift"
##
## $H1N1
## [1] "in H1N1 season 1yes"
##
## $quarter
## [1] "study qtr, July-Sept09 is first qtr"
##
## $shiftampm
## [1] "time of shift (1day, 2night)"
##
## $d9
## [1] "day of week that shift started"
##
## $edgeweight
## [1] "hours of contact"
##
## $j
## [1] "second member of contact pair (find real id using id_sid_matchuplist)"
##
## $staffj
## [1] "j is a staff member 1 yes 0no"
##
## $combo
## [1] "type of contact 0(pp) 1(ps) 2(ss)"
##
## $idj
## [1] "id for j made of d8 and j"
##
## $comboc
## [1] "patient-staff combinations"
```

```
# Print the first 6 observations of edge_list2
head(edge_list2)
```

```
## # A tibble: 6 x 29
##   numshift shiftampm D8            d9  H1N1 quarter sidi     sidj         i     j
##      <dbl>     <dbl> <date>     <dbl> <dbl>   <dbl> <chr>    <chr>    <dbl> <dbl>
## 1        1         2 2009-07-09 18087     0       1 0002f35c 0002f4~     1     2
## 2        1         2 2009-07-09 18087     0       1 0002f35c 0002f4~     1     3
## 3        1         2 2009-07-09 18087     0       1 0002f35c 0002f4~     1     4
## 4        1         2 2009-07-09 18087     0       1 0002f35c 0002f4~     1     5
## 5        1         2 2009-07-09 18087     0       1 0002f35c 0002f4~     1     6
## 6        1         2 2009-07-09 18087     0       1 0002f35c 0002f4~     1     7
## # ... with 19 more variables: idi <chr>, idj <chr>, i_participant_type <chr>,
## #   j_participant_type <chr>, staffi <dbl>, staffj <dbl>, anycontact <dbl>,
## #   combo <dbl>, comboc <chr>, combo4 <chr>, MD_CONTACTS <dbl>,
## #   RN_CONTACTS <dbl>, STAFF_CONTACTS <dbl>, PAT_CONTACTS <dbl>,
## #   MD_WITHWHOM <chr>, RN_WITHWHOM <chr>, STAFF_WITHWHOM <chr>,
## #   PAT_WITHWHOM <chr>, edgeweight <dbl>
```

```
# Print out the variable labels for all columns of edge_list2
var_label(edge_list2)
```

```
## $numshift
## [1] "shift number"
##
## $shiftampm
## [1] "time of shift (1day, 2night)"
##
## $D8
## [1] "first date in shift"
##
## $d9
## [1] "day of week that shift started"
##
## $H1N1
## [1] "in H1N1 season 1yes"
##
## $quarter
## [1] "study qtr, July-Sept09 is first qtr"
##
## $sidi
## [1] "SID OF NODE I"
##
## $sidj
## [1] "SID OF NODE J"
##
## $i
## [1] "one member of contact pair (find real id using id_sid_matchuplist)"
##
## $j
## [1] "arbitrary sid for this d8"
##
## $idi
## [1] "id for i made of d8 and i"
##
## $idj
## [1] "id for j made of d8 and j"
##
## $i_participant_type
## [1] "participant type"
##
## $j_participant_type
## [1] "participant type"
##
```

```
## $staffi
## [1] "i is a staff member 1yes 0no"
##
## $staffj
## [1] "j is a staff member 1 yes 0no"
##
## $anycontact
## [1] "any contact 1yes 0no"
##
## $combo
## [1] "type of contact 0(pp) 1(ps) 2(ss)"
##
## $comboc
## [1] "patient-staff combinations"
##
## $combo4
## [1] "DETAILED CONTACT DESCRIPTION (PARTICIPANT TYPE COMBINATIONS)"
##
## $MD_CONTACTS
## [1] "the edge has at least one MD node"
##
## $RN_CONTACTS
## [1] "the edge has at least one RN node"
##
## $STAFF_CONTACTS
## [1] "the edge has at least one STAFF node"
##
## $PAT_CONTACTS
## [1] "the edge has at least one PATIENT node"
##
## $MD_WITHWHOM
## [1] "TYPE OF CONTACT PARTNER (MD)"
##
## $RN_WITHWHOM
## [1] "TYPE OF CONTACT PARTNER (RN)"
##
## $STAFF_WITHWHOM
## [1] "TYPE OF CONTACT PARTNER (STAFF)"
##
## $PAT_WITHWHOM
## [1] "TYPE OF CONTACT PARTNER (PAT)"
##
## $edgeweight
## [1] "hours of contact"
```

```
# 1d. read "id_sid_matchup.sas7bdat" into id_sid and "id_sid_matchup2.sas7bdat" into id_sid2
id_sid <- read_sas(paste0(data_path, "Data_Files/id_sid_matchup.sas7bdat"))
id_sid2 <- read_sas(paste0(data_path, "Data_Files/id_sid_matchup2.sas7bdat"))
```

```
# Print the first 6 rows of id_sid
head(id_sid)
```

```
## # A tibble: 6 x 5
##   sid        day   mon staff newsid
##   <chr>    <dbl> <dbl> <dbl>  <dbl>
## 1 0002f35c     9     7     1      1
## 2 0002f445     9     7     1      2
## 3 0002f468     9     7     1      3
## 4 0002f469     9     7     1      4
## 5 0002f46c     9     7     1      5
## 6 0002f472     9     7     1      6
```

```
# Print the first 6 rows of id_sid2
head(id_sid2)
```

```
## # A tibble: 6 x 19
##   sid     day   mon staff newsid  year d8         ShiftStart ShiftEnd shift_ampm
##   <chr> <dbl> <dbl> <dbl>  <dbl> <dbl> <date>     <time>     <time>   <chr>
## 1 0002~     9     7     1      1  2009 2009-07-09 20:00      23:59:59 pm
## 2 0002~     9     7     1      2  2009 2009-07-09 20:00      23:59:59 pm
## 3 0002~     9     7     1      3  2009 2009-07-09 20:00      23:59:59 pm
## 4 0002~     9     7     1      4  2009 2009-07-09 20:00      23:59:59 pm
## 5 0002~     9     7     1      5  2009 2009-07-09 20:00      23:59:59 pm
## 6 0002~     9     7     1      6  2009 2009-07-09 20:00      23:59:59 pm
## # ... with 9 more variables: Reason_shortShift <chr>, startd8time <dttm>,
## #   shift_d8_ampm <chr>, shift_num_ampm <chr>, quarter <dbl>, weekday <dbl>,
## #   H1N1 <dbl>, SevenToTwelve <dbl>, numshift <dbl>
```

```
# ----- `pt_acuity` -----
# 1e. read "ACUITY-patients.xlsx
pt_acuity <- read_xlsx(paste0(data_path, "Data_Files/ACUITY-patients.xlsx"))
# str(pt_acuity)
pt_acuity_s2 <- read_xlsx(paste0(data_path, "Data_Files/ACUITY-patients.xlsx"), sheet = 2)
# str(pt_acuity_s2)
pt_acuity_s3 <- read_xlsx(paste0(data_path, "Data_Files/ACUITY-patients.xlsx"), sheet = 3)
# str(pt_acuity_s3)
```

```
# Patient acuity (Emergency Severity Index; ESI) counts by shift
head(pt_acuity)
# Pivot_wider to view number of patients in each ESI category by shift
pt_acuity %>%
  group_by(Acuity) %>%
  count(Shift) %>%
  pivot_wider(names_from = Acuity, values_from = n) %>%
  head()
# Print the first 6 rows of the other two sheets in the xlsx file
head(pt_acuity_s2)
head(pt_acuity_s3)
```

RFID Badge & Location Data

"Cpat.zip" and "Cstaff.zip" contain "completepat.sas7bdat" and "completestaff.sas7bdat," respectively, that contain location information for patients and staff from all observed shifts, respectively. * Both completeXXX.sas7bdat tables have columns for every second of the day, named with the prefix "floc" followed by the second * Each row contains the locations (numeric values) for the respective SID and date combination +

Some patient SIDs repeat in the data because RFID tags were used by more than one patient per shift +
Staff had permanent tags, so SID numbers were not duplicated
* Room locations with square footage are in an Excel file, which links location numbers to location names

Columns (i.e., variables), variable classes, and variables definitions in 'completepat.sas7bdat'

Variable	Class	Definition
SID	char	individual RFID badge number
date	char	date observation started

Columns (i.e., variables), variable classes, and variables definitions in 'completestaff.sas7bdat'

Variable	Class	Definition
SID	char	individual RFID badge number
date	char	date observation started