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",
                      results = "asis",
                      echo = TRUE,
                      message = FALSE,
                      warning = FALSE,
                      background = "#F7F7F7",
                      tidy = TRUE,
                      tidy.opts = list(width.cutoff = 60))
# 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(knitr)
# 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

RFID badge location room number for every second of every shift, patients only.

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

Large data.frame, using the first six observations to code for data cleaning. The table is extremely wide (>4300 columns), I used `pivot_longer()` to reshape it by collapsing all location-by-second columns into two columns, names to `seconds` and values to `location`. This process causes there to be **many repeated SIDs**.

```
# 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 a new
    # column, 'seconds,' and placing respective
    # observations for each 'floc' variable into a
    # 'location_num' column
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)
kable(head(pt_head))
```

sid	shift_num_ampm	d8	day	mon	year	firstday	seconds	location_num	shift_num	am_pm	date
0002d045	1pm	18087	9	7	2009	18087	1	NA	1	pm	2009-07-09
0002d045	1pm	18087	9	7	2009	18087	2	NA	1	pm	2009-07-09
0002d045	1pm	18087	9	7	2009	18087	3	NA	1	pm	2009-07-09
0002d045	1pm	18087	9	7	2009	18087	4	NA	1	pm	2009-07-09
0002d045	1pm	18087	9	7	2009	18087	5	NA	1	pm	2009-07-09
0002d045	1pm	18087	9	7	2009	18087	6	NA	1	pm	2009-07-09

Need to figure out how to filter out redundency without cutting data.

staff_complete

RFID badge location room number for every second of every shift, patients only

```
# ---- `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 a new
    # column, 'seconds,' and placing respective
    # observations for each 'floc' variable into a
    # 'location_num' column
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)
kable(head(staff_head))
```

sid	d8	day	year	shift_num_ampm	mon	firstday	seconds	location_num	shift_num	am_pm	date
0002f4e2	2009-07-09	9	2009	1pm	7	18087	1	NA	1	pm	2009-07-09
0002f4e2	2009-07-09	9	2009	1pm	7	18087	2	NA	1	pm	2009-07-09
0002f4e2	2009-07-09	9	2009	1pm	7	18087	3	NA	1	pm	2009-07-09
0002f4e2	2009-07-09	9	2009	1pm	7	18087	4	NA	1	pm	2009-07-09
0002f4e2	2009-07-09	9	2009	1pm	7	18087	5	NA	1	pm	2009-07-09
0002f4e2	2009-07-09	9	2009	1pm	7	18087	6	NA	1	pm	2009-07-09

edge_list

Read & print data from “allshifts_edges.sas7bdat” and “edges2.sas7bdat.”

```
# ---- '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"))
```

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

i	any	staffi	idi	d8	H1N1	quarter	shiftampm	d9	edgeweight	j	staffj	combo	idj	comboc
1	1	1	7920091	2009-07-09	0	1	2	18087	0.5247222	2	1	2	7920092	2 staff-staff
1	1	1	7920091	2009-07-09	0	1	2	18087	3.7672222	3	1	2	7920093	2 staff-staff
1	1	1	7920091	2009-07-09	0	1	2	18087	1.1116667	4	1	2	7920094	2 staff-staff
1	1	1	7920091	2009-07-09	0	1	2	18087	0.4872222	5	1	2	7920095	2 staff-staff
1	1	1	7920091	2009-07-09	0	1	2	18087	0.7936111	6	1	2	7920096	2 staff-staff
1	1	1	7920091	2009-07-09	0	1	2	18087	0.5130556	7	1	2	7920097	2 staff-staff

```r
# Print out the variable labels for all columns of
# edge_list (object varbles1)
varbles1 <- var_label(edge_list)
paste(names(varbles1), varbles1, sep = ": ")
```

[1] "i: one member of contact pair (find real id using id__sid__matchuplist)"
[2] "any: any contact 1yes 0no"
[3] "staffi: i is a staff member 1yes 0no"
[4] "idi: id for i made of d8 and i"
[5] "d8: 1st d8 in the shift"
[6] "H1N1: in H1N1 season 1yes"
[7] "quarter: study qtr, July-Sept09 is first qtr"
[8] "shiftampm: time of shift (1day, 2night)"
[9] "d9: day of week that shift started"
[10] "edgeweight: hours of contact"
[11] "j: second member of contact pair (find real id using id__sid__matchuplist)" [12] "staffj: j is a staff member 1 yes 0no"
[13] "combo: type of contact 0(pp) 1(ps) 2(ss)"
[14] "idj: id for j made of d8 and j"
[15] "comboc: patient-staff combinations"

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

num	shiftampm	D8	d9	H1N1	quarter	sidi	sidj	i	j	idi	idj	i_participant_type	j_participant_type	staffi	staffj	contact	combo	combo4	MD_RN...	CONTACT...	STAFF...	...	...
1	2	2009-07-09	18087	1	0002f352	2035f44	2	5	7920091	7920092	STAFF	STAFF	1	1	2	2 staff-staff	STAFF-STAFF	0	1	0		STAFF	0.5247222
1	2	2009-07-09	18087	1	0002f352	2035f46	3	8	7920091	7920093	STAFF	STAFF	1	1	2	2 staff-staff	STAFF-STAFF	0	1	0		STAFF	3.7672222

numshift	shiftampm	D8	d9	H1N1	quarter	sidi	sidj	i	j	idi	idj	i_participant_type	j_participant_type	staffi	staffj	anycontact	combo	comboc	combo4	MD_CONTACTS	RN_CONTACTS	STAFF_CONTACTS	PAT_CONTACTS	MD_WITHWHOM	RN_WITHWHOM	STAFF_WITHWHOM	PAT_WITHWHOM	edgeweight
1	2	2009-07-09	1-8087	1	0002f052f469	792092019	PATSTAFF	1	1	2	2 staff-staff	STAFF-0 STAFF	1	0	STAFF	1.1116667												
1	2	2009-07-09	1-8087	1	0002f052f46c	792092015	PATSTAFF	1	1	2	2 staff-staff	STAFF-0 STAFF	1	0	STAFF	0.4872222												
1	2	2009-07-09	1-8087	1	0002f052f472	792092016	PATSTAFF	1	1	2	2 staff-staff	STAFF-0 STAFF	1	0	STAFF	0.7936111												
1	2	2009-07-09	1-8087	1	0002f052f495	792092017	PATSTAFF	1	1	2	2 staff-staff	STAFF-0 STAFF	1	0	STAFF	0.5130556												

```r
# Print out the variable labels for all columns of
# edge_list2 (object varbles2)
varbles2 <- var_label(edge_list2)
paste(names(varbles2), varbles2, sep = ": ")
```

[1] "numshift: shift number"
[2] "shiftampm: time of shift (1day, 2night)"
[3] "D8: first date in shift"
[4] "d9: day of week that shift started"
[5] "H1N1: in H1N1 season 1yes"
[6] "quarter: study qtr, July-Sept09 is first qtr"
[7] "sidi: SID OF NODE I"
[8] "sidj: SID OF NODE J"
[9] "i: one member of contact pair (find real id using id_sid_matchuplist)" [10] "j: arbitrary sid for this d8"
[11] "idi: id for i made of d8 and i"
[12] "idj: id for j made of d8 and j"
[13] "i_participant_type: participant type"
[14] "j_participant_type: participant type"
[15] "staffi: i is a staff member 1yes 0no"
[16] "staffj: j is a staff member 1 yes 0no"
[17] "anycontact: any contact 1yes 0no"
[18] "combo: type of contact 0(pp) 1(ps) 2(ss)"
[19] "comboc: patient-staff combinations"
[20] "combo4: DETAILED CONTACT DESCRIPTION (PARTICIPANT TYPE COMBINATIONS)" [21] "MD_CONTACTS: the edge has at least one MD node"
[22] "RN_CONTACTS: the edge has at least one RN node"
[23] "STAFF_CONTACTS: the edge has at least one STAFF node"
[24] "PAT_CONTACTS: the edge has at least one PATIENT node"
[25] "MD_WITHWHOM: TYPE OF CONTACT PARTNER (MD)"
[26] "RN_WITHWHOM: TYPE OF CONTACT PARTNER (RN)"
[27] "STAFF_WITHWHOM: TYPE OF CONTACT PARTNER (STAFF)"
[28] "PAT_WITHWHOM: TYPE OF CONTACT PARTNER (PAT)"
[29] "edgeweight: hours of contact"

id_sid

Read & print data from "id_sid_matchup.sas7bdat" & "id_sid_matchup.sas7bdat."

```
# 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
kable(head(id_sid))
```

sid	day	mon	staff	newsid
0002f35c	9	7	1	1
0002f445	9	7	1	2
0002f468	9	7	1	3
0002f469	9	7	1	4
0002f46c	9	7	1	5
0002f472	9	7	1	6

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

sid	day	mon	staff	newsid	year	d8	ShiftS	Shift	Shift_	Reason	shortS	Shift_	Shift_	Shift_	Shift_	Shift_	Shift_	Shift_	Shift_	Shift_	Shift_	Shift_	Shift_	Shift_
0002f35c	7	1	1	2009	2009-07-09	20:00:00	23:59:59	pm			2009-07-09 20:00:00	200979	pm	1	1	0	1	1
0002f445	7	1	2	2009	2009-07-09	20:00:00	23:59:59	pm			2009-07-09 20:00:00	200979	pm	1	1	0	1	1
0002f468	7	1	3	2009	2009-07-09	20:00:00	23:59:59	pm			2009-07-09 20:00:00	200979	pm	1	1	0	1	1
0002f469	7	1	4	2009	2009-07-09	20:00:00	23:59:59	pm			2009-07-09 20:00:00	200979	pm	1	1	0	1	1
0002f46c	7	1	5	2009	2009-07-09	20:00:00	23:59:59	pm			2009-07-09 20:00:00	200979	pm	1	1	0	1	1
0002f472	7	1	6	2009	2009-07-09	20:00:00	23:59:59	pm			2009-07-09 20:00:00	200979	pm	1	1	0	1	1

```
# Print variable labels
idsid_varbles <- var_label(id_sid)
idsid2_varbles <- var_label(id_sid2)

idsid_labeltable <- paste(names(idsid_varbles), idsid_varbles,
    sep = ": ")
tibble(idsid_labeltable)
```

A tibble: 5 x 1

idsid_labeltable
1 sid: NULL
2 day: NULL
3 mon: NULL
4 staff: NULL
5 newsid: NULL

```
idsid2_labeltable <- paste(names(idsid2_varbles), idsid2_varbles,
    sep = ": ")
tibble(idsid2_labeltable)
```

A tibble: 19 x 1

idsid2_labeltable

1 sid: NULL
2 day: NULL
3 mon: NULL
4 staff: NULL
5 newsid: NULL
6 year: NULL
7 d8: first date in shift
8 ShiftStart: ShiftStart
9 ShiftEnd: ShiftEnd
10 shift_ampm: shift_ampm
11 Reason_shortShift: Reason_shortShift
12 startd8time: NULL
13 shift_d8_ampm: NULL
14 shift_num_ampm: NULL
15 quarter: 1summer 2fall **correct definition of weekend for this study i~ 16 weekday: NULL
17 H1N1: NULL
18 SevenToTwelve: NULL
19 numshift: NULL

pt_acuity

Read & print patient acuity data in "ACUITY-patients.xlsx," which is an Excel workbook with 3 sheets.

```
# 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, range = "A2:H83")
# str(pt_acuity_s2)
pt_acuity_s3 <- read_xlsx(paste0(data_path, "Data_Files/ACUITY-patients.xlsx"),
    sheet = 3, range = "A2:G83")
# str(pt_acuity_s3)
```

The first sheet lists the number of patients in each ESI acuity level (columns) by shift (rows). The other two sheets appear to be variations of the first.

```
# Patient acuity (Emergency Severity Index; ESI) counts by
# shift
kable(head(pt_acuity))
```

Shift	Acuity
1	3 Urgent
1	3 Urgent
1	4 Stable
1	4 Stable
1	2 Emergent
1	3 Urgent

```
# 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() %>%
    kable()
```

Shift	1 Immediate	2 Emergent	3 Urgent	4 Stable	5 Non Urgent	Not Recorded
1	1	18	45	5	2	3
38	1	23	48	15	2	2
56	1	27	41	18	NA	2
63	6	19	49	15	NA	4
98	1	22	37	8	NA	NA
102	1	17	34	12	NA	NA

```
# Print the first 6 rows of the other two sheets in the
# xlsx file
kable(head(pt_acuity_s2))
```

Shift	1 Immediate	2 Emergent	3 Urgent	4 Stable	5 Non Urgent	Not Recorded	Grand Total
1	1	18	45	5	2	3	74
8	0	30	34	12	1	5	82
10	0	19	30	13	2	0	64
17	0	24	50	19	1	1	95
19	0	22	48	17	0	2	89
23	0	14	43	13	1	1	72

```
kable(head(pt_acuity_s3))
```

Shift	1 Immediate	2 Emergent	3 Urgent	4 Stable	5 Non Urgent	Not Recorded
1	0.0135135	0.2432432	0.6081081	0.0675676	0.0270270	0.0405405

Shift	1 Immediate	2 Emergent	3 Urgent	4 Stable	5 Non Urgent	Not Recorded
8	0.0000000	0.3658537	0.4146341	0.1463415	0.0121951	0.0609756
10	0.0000000	0.2968750	0.4687500	0.2031250	0.0312500	0.0000000
17	0.0000000	0.2526316	0.5263158	0.2000000	0.0105263	0.0105263
19	0.0000000	0.2471910	0.5393258	0.1910112	0.0000000	0.0224719
23	0.0000000	0.1944444	0.5972222	0.1805556	0.0138889	0.0138889

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