Dengue raw data cleaner

Table of contents

Load libraries

```
.pkgs <- c("tidyverse", "janitor", "fs", "readxl", "skimr", "stringi")
xfun::pkg_attach(.pkgs)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.2     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.4
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
Attaching package: 'janitor'
```

```
The following objects are masked from 'package:stats':

    chisq.test, fisher.test
```

```
theme_set(theme_bw())
```

Data ingestion

Read raw excel files and quickly clean column names with `janitor::clean_names()`

```
xlsx_raws <- dir_ls("incidence_data_2000_2022", regexp = "xlsx") %>%
  map(
    ~ map(excel_sheets(.x), \(sh) {
      suppressWarnings(read_excel(.x, sheet = sh)) %>% clean_names()
```

```
    })
  ) %>%
  flatten() %>%
  set_names(NULL)
```

Our columns of interest are: “sex”, “age”, “date of admission”, “district (of patient’s residential address)”, “commune (of patient’s residential address)”, “hospital”, “icd”, “in-/out-patient”

EDA and data cleaning

There are 2 excel files, first file contains cases from 2000-2016, a separate sheet for each year. Second find contains cases from 2017-2022, all in one sheet. The 2 files contains 2 different format of columns because they come from 2 different reporting systems.

Data from 2000-2016

First we will look at data from 2000-2016

We can look at mismatched columns for all the sheets

```
compare_df_cols(xlsx_raws[-length(xlsx_raws)], return = "mismatch")
```

```
  column_name xlsx_raws[-length(xlsx_raws)]_1 xlsx_raws[-length(xlsx_raws)]_2
1   cd_ravien                         logical                         logical
2       ng_bc                POSIXct, POSIXt                 POSIXct, POSIXt
3   ng_tuvong                         logical                         logical
  xlsx_raws[-length(xlsx_raws)]_3 xlsx_raws[-length(xlsx_raws)]_4
1                         logical                         logical
2                POSIXct, POSIXt                         logical
3                         logical                         logical
  xlsx_raws[-length(xlsx_raws)]_5 xlsx_raws[-length(xlsx_raws)]_6
1                         logical                         logical
2                POSIXct, POSIXt                 POSIXct, POSIXt
3                         logical                 POSIXct, POSIXt
  xlsx_raws[-length(xlsx_raws)]_7 xlsx_raws[-length(xlsx_raws)]_8
1                         logical                       character
2                POSIXct, POSIXt                 POSIXct, POSIXt
3                         logical                 POSIXct, POSIXt
  xlsx_raws[-length(xlsx_raws)]_9 xlsx_raws[-length(xlsx_raws)]_10
1                       character                        character
```

```
2                 POSIXct, POSIXt                  POSIXct, POSIXt
3                 POSIXct, POSIXt                  POSIXct, POSIXt
  xlsx_raws[-length(xlsx_raws)]_11 xlsx_raws[-length(xlsx_raws)]_12
1                        character                             <NA>
2                 POSIXct, POSIXt                  POSIXct, POSIXt
3                          logical                  POSIXct, POSIXt
  xlsx_raws[-length(xlsx_raws)]_13 xlsx_raws[-length(xlsx_raws)]_14
1                        character                        character
2                 POSIXct, POSIXt                  POSIXct, POSIXt
3                          logical                          logical
  xlsx_raws[-length(xlsx_raws)]_15 xlsx_raws[-length(xlsx_raws)]_16
1                        character                        character
2                 POSIXct, POSIXt                  POSIXct, POSIXt
3                 POSIXct, POSIXt                  POSIXct, POSIXt
  xlsx_raws[-length(xlsx_raws)]_17
1                        character
2                 POSIXct, POSIXt
3                 POSIXct, POSIXt
```

Mismatched columns `cd_ravien`, `ng_bc`, `ng_tuvong` are insignificant, so we will go ahead with row binding all the sheets together ::: {.callout-note} `xlsx_raws[-length(xlsx_raws)]` means look at all sheets from first file, ignore last sheet (i.e. second file) :::

```
raw_2000_2016 <- xlsx_raws[-length(xlsx_raws)] %>% bind_rows()
```

Skimming to see what's going on in the data

```
skim(raw_2000_2016)
```

Table 1: Data summary

Name	raw_2000_2016
Number of rows	147927
Number of columns	33
_____	
Column type frequency:	
character	12
logical	14
numeric	2
POSIXct	5
_____	

Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
maso	43969	0.70	1	15	0	88299	0
hoten	0	1.00	3	39	0	110896	0
gioi	6351	0.96	1	1	0	7	0
diachi	5557	0.96	1	92	0	133295	0
px	0	1.00	2	16	0	171	0
qh	0	1.00	2	10	0	24	0
ma_tinh	0	1.00	3	3	0	1	0
cd_ravien	96007	0.35	14	14	0	1	0
nguon_du_lieu	0	1.00	7	18	0	40	0
do_sxh	31125	0.79	1	1	0	15	0
m_icd	7495	0.95	3	6	0	37	0
naso	144843	0.02	4	10	0	3082	0

Variable type: logical

skim_variable	n_missing	complete_rate	mean	count
ma_moi_bc	147927	0	NaN	:
ap	147927	0	NaN	:
ten_cha	147927	0	NaN	:
laymauxetnghiem	147927	0	NaN	:
elisa	147927	0	NaN	:
plvr	147927	0	NaN	:
ns1	147927	0	NaN	:
odn	147927	0	NaN	:
ng_khoibenh	147927	0	NaN	:
cd_vaovien	147927	0	NaN	:
ly_do_tu_vong	147927	0	NaN	:
nv_nhap	147927	0	NaN	:
ng_hieuchinh	147927	0	NaN	:
ghi_chu	147927	0	NaN	:

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
tuoi	22100	0.85	16.33	240.8	-34669	8	13	22	2015	
ng_sinh	32149	0.78	1994.34	251.0	0	1988	1996	2003	36685	

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
ng_vaovien	0	1.00	2000-01-01	2016-12-31 22:14:28	2010-04-19 00:00:00	8103
ng_ravien	1411	0.99	1900-01-22	2017-03-16 00:00:00	2010-07-21 00:00:00	7109
ng_tuvong	147875	0.00	2000-09-09	2011-11-25 00:00:00	2008-10-27 12:00:00	52
ng_bc	56911	0.62	2000-01-07	2017-03-22 00:00:00	2007-11-29 00:00:00	4101
ng_nhap	364	1.00	2000-01-07	2017-03-24 00:00:00	2010-05-07 00:00:00	2033

The actual columns based on our columns of interest are (with `complete_rate`): - `sex` = `gioi` (0.957) - `age` = `tuoi` (0.851); `ng_sinh` (0.873); year of birth might be in `hoten` col - `date of admission` = `ng_vaovien` (1) - `district` = `qh` (1) - `commune` = `px` (1) - `hospital` = `nguon_du_lieu` (1) - `icd` = `m_icd` (0.949) - `in-patient` = this is all in-patient data

Selecting columns of interest

```
s1_2000_2016 <- raw_2000_2016 %>%
  select(hoten, gioi, tuoi, ng_sinh, ng_vaovien, qh, px, nguon_du_lieu, m_icd)
# s1_2000_2016
```

Keeping track of raw number of rows at start to see how much lost during data cleaning

```
start_nrow <- nrow(s1_2000_2016)
start_nrow
```

```
[1] 147927
```

Fix age if possible

Some year of births (YOBs) are stored in the name (`hoten`) column, let's see how much of the data is like this

```
s1_2000_2016 %>%
  select(hoten, tuoi, ng_sinh) %>%
  filter(is.na(tuoi)) %>%
  rowwise() %>%
  mutate(
    no_age = any(
      varhandle::check.numeric(hoten),
      !is.na(ng_sinh)
    )
  ) %>%
  tabyl(no_age)
```

```
 no_age     n     percent
  FALSE 21898 0.990859729
   TRUE   202 0.009140271
```

Less than 1% of the data has YOB in the name column and able to get age from YOB, so it's not worth going deeper.

For now, people with no age will have `NA` as their age

Rename columns

```
s2_2000_2016 <- s1_2000_2016 %>%
  select(-hoten, -ng_sinh) %>%
  rename(
    sex = gioi,
    age = tuoi,
    date = ng_vaovien,
    district = qh,
    commune = px,
    hospital = nguon_du_lieu,
    icd = m_icd
  ) %>%
  mutate(
```

```
    in_out_patient = "in-patient"
  )
# s2_2000_2016
```

Fix dates data

Date data is in `datetime`, convert to `date` only

```
s3_2000_2016 <- s2_2000_2016 %>%
  mutate(date = convert_to_date(date))
# s3_2000_2016
```

```
skim(s3_2000_2016)
```

Table 6: Data summary

Name	s3_2000_2016
Number of rows	147927
Number of columns	8
_____	
Column type frequency:	
character	6
Date	1
numeric	1
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
sex	6351	0.96	1	1	0	7	0
district	0	1.00	2	10	0	24	0
commune	0	1.00	2	16	0	171	0
hospital	0	1.00	7	18	0	40	0
icd	7495	0.95	3	6	0	37	0
in_out_patient	0	1.00	10	10	0	1	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
date	0	1	2000-01-01	2016-12-31	2010-04-19	6184

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	22100	0.85	16.33	240.8	-34669	8	13	22	2015	

We see that there are 7 unique sexes, 37 unique ICDs, negative age. Let's clean all of this

Fix sexes

Look into `sex` column first

```
s3_2000_2016 %>% tabyl(sex)
```

```
  sex     n      percent valid_percent
    1     1 6.760091e-06  7.063344e-06
    G 59814 4.043481e-01  4.224869e-01
    I  9073 6.133431e-02  6.408572e-02
    N   890 6.016481e-03  6.286376e-03
    T 71786 4.852799e-01  5.070492e-01
    g     4 2.704036e-05  2.825338e-05
    t     8 5.408073e-05  5.650675e-05
 <NA>  6351 4.293334e-02            NA
```

Impossible to know what "I" and "N" means skip this for now as it's not that important

Fix ICD

```
s3_2000_2016 %>%
  mutate(year = year(date), icd = tolower(icd)) %>%
  tabyl(icd, year)
```

icd	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
a90	0	0	0	0	0	0	0	0	0	0	140	937	15	1	0	0	0
a91	375	556	865	2680	3329	2526	2955	1180	3672	1378	38	146	9	20	21	258	6990
a91-1	0	30	31	55	64	52	78	129	83	137	225	30	4	0	0	0	0
a91-2	6	456	571	1300	1187	994	1807	2437	2278	2686	2241	480	11	0	0	0	0
a91-3	0	74	99	325	194	115	151	280	355	389	334	79	1	0	0	0	0
a91-4	0	7	10	15	11	4	16	17	13	35	27	9	0	0	0	0	0
a91-e	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0
a91.	0	0	0	0	0	0	0	0	0	0	9	5	0	0	0	0	0
a91.]	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
a91.`	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
a91.0	5	8	4	6	0	1	0	4	0	2	0	0	0	0	0	0	0
a91.1	946	979	657	1769	748	814	947	1368	1516	1319	62	223	0	0	0	0	0
a91.2	647	548	428	1791	1538	1469	2221	3424	5631	4723	200	542	0	0	0	0	0
a91.3	147	133	119	407	227	151	198	251	613	411	21	61	0	0	0	0	0
a91.4	10	11	6	29	13	8	8	11	66	31	1	0	0	0	0	0	0
a91.9	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0
a91.a	0	0	0	0	0	0	0	0	0	0	0	6113	7407	5310	0	0	0
a91.b	0	0	0	0	0	0	0	0	0	0	0	1149	1400	1081	0	0	0
a91.c	0	0	0	0	0	0	0	0	0	0	0	253	244	156	0	0	0
a91.c1	0	0	0	0	0	0	0	0	0	0	0	317	423	285	0	0	0
a91.c2	0	0	0	0	0	0	0	0	0	0	0	17	30	35	0	0	0
a91.c3	0	0	0	0	0	0	0	0	0	0	0	7	1	3	0	0	0
a91.c4	0	0	0	0	0	0	0	0	0	0	0	10	12	16	0	0	0
a99.1	0	0	0	0	0	0	0	0	0	0	3	30	0	0	0	0	0
a99.2	0	0	0	0	0	0	0	0	0	0	12	478	0	0	0	0	0
a99.3	0	0	0	0	0	0	0	0	0	0	2	41	0	0	0	0	0
a99.4	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0
<NA>	0	0	0	0	0	0	0	0	0	0	6373	888	233	1	0	0	0

```
   0    0    0
   0    0    0
   0    0    0
5051 8832 8767
1216 2346 2287
  64  151  113
 298  339  360
  27   34   16
   2    7   10
  30   58   66
   0    0    0
   0    0    0
   0    0    0
   0    0    0
   0    0    0
```

Needs to consult dengue experts on this, skip for now

Fix age, again

```
s3_2000_2016 %>% tabyl(age)
```

```
   age     n      percent valid_percent
-34669     1 6.760091e-06  7.947420e-06
-31868     1 6.760091e-06  7.947420e-06
-31450     1 6.760091e-06  7.947420e-06
-30781     1 6.760091e-06  7.947420e-06
-28917     1 6.760091e-06  7.947420e-06
-26725     1 6.760091e-06  7.947420e-06
-17994     1 6.760091e-06  7.947420e-06
-17993     2 1.352018e-05  1.589484e-05
 -4670     1 6.760091e-06  7.947420e-06
  -297     1 6.760091e-06  7.947420e-06
    -2     1 6.760091e-06  7.947420e-06
     0  2237 1.512232e-02  1.777838e-02
     1  4871 3.292840e-02  3.871188e-02
     2  2895 1.957046e-02  2.300778e-02
     3  2976 2.011803e-02  2.365152e-02
     4  3380 2.284911e-02  2.686228e-02
     5  3941 2.664152e-02  3.132078e-02
```

6	4782	3.232676e-02	3.800456e-02
7	5278	3.567976e-02	4.194648e-02
8	5456	3.688306e-02	4.336112e-02
9	5494	3.713994e-02	4.366312e-02
10	5964	4.031718e-02	4.739841e-02
11	5956	4.026310e-02	4.733483e-02
12	6053	4.091883e-02	4.810573e-02
13	6165	4.167596e-02	4.899584e-02
14	5786	3.911389e-02	4.598377e-02
15	3881	2.623591e-02	3.084394e-02
16	2565	1.733963e-02	2.038513e-02
17	2435	1.646082e-02	1.935197e-02
18	2850	1.926626e-02	2.265015e-02
19	2925	1.977327e-02	2.324620e-02
20	3090	2.088868e-02	2.455753e-02
21	2946	1.991523e-02	2.341310e-02
22	2863	1.935414e-02	2.275346e-02
23	2757	1.863757e-02	2.191104e-02
24	2519	1.702867e-02	2.001955e-02
25	2241	1.514936e-02	1.781017e-02
26	2141	1.447336e-02	1.701543e-02
27	1966	1.329034e-02	1.562463e-02
28	1894	1.280361e-02	1.505241e-02
29	1720	1.162736e-02	1.366956e-02
30	1608	1.087023e-02	1.277945e-02
31	1440	9.734531e-03	1.144428e-02
32	1368	9.247805e-03	1.087207e-02
33	1190	8.044508e-03	9.457430e-03
34	967	6.537008e-03	7.685155e-03
35	931	6.293645e-03	7.399048e-03
36	857	5.793398e-03	6.810939e-03
37	770	5.205270e-03	6.119513e-03
38	670	4.529261e-03	5.324771e-03
39	569	3.846492e-03	4.522082e-03
40	531	3.589608e-03	4.220080e-03
41	484	3.271884e-03	3.846551e-03
42	397	2.683756e-03	3.155126e-03
43	377	2.548554e-03	2.996177e-03
44	302	2.041548e-03	2.400121e-03
45	295	1.994227e-03	2.344489e-03
46	270	1.825225e-03	2.145803e-03
47	240	1.622422e-03	1.907381e-03
48	215	1.453420e-03	1.708695e-03

49	183	1.237097e-03	1.454378e-03
50	213	1.439899e-03	1.692800e-03
51	204	1.379059e-03	1.621274e-03
52	140	9.464128e-04	1.112639e-03
53	127	8.585316e-04	1.009322e-03
54	121	8.179710e-04	9.616378e-04
55	109	7.368499e-04	8.662688e-04
56	105	7.098096e-04	8.344791e-04
57	74	5.002467e-04	5.881091e-04
58	116	7.841706e-04	9.219007e-04
59	73	4.934867e-04	5.801617e-04
60	66	4.461660e-04	5.245297e-04
61	55	3.718050e-04	4.371081e-04
62	63	4.258857e-04	5.006875e-04
63	51	3.447646e-04	4.053184e-04
64	50	3.380046e-04	3.973710e-04
65	37	2.501234e-04	2.940545e-04
66	27	1.825225e-04	2.145803e-04
67	29	1.960426e-04	2.304752e-04
68	28	1.892826e-04	2.225278e-04
69	26	1.757624e-04	2.066329e-04
70	33	2.230830e-04	2.622649e-04
71	23	1.554821e-04	1.827907e-04
72	23	1.554821e-04	1.827907e-04
73	17	1.149215e-04	1.351061e-04
74	15	1.014014e-04	1.192113e-04
75	12	8.112109e-05	9.536904e-05
76	13	8.788118e-05	1.033165e-04
77	6	4.056055e-05	4.768452e-05
78	8	5.408073e-05	6.357936e-05
79	9	6.084082e-05	7.152678e-05
80	21	1.419619e-04	1.668958e-04
81	13	8.788118e-05	1.033165e-04
82	8	5.408073e-05	6.357936e-05
83	4	2.704036e-05	3.178968e-05
84	5	3.380046e-05	3.973710e-05
85	3	2.028027e-05	2.384226e-05
86	2	1.352018e-05	1.589484e-05
87	3	2.028027e-05	2.384226e-05
88	3	2.028027e-05	2.384226e-05
90	25	1.690023e-04	1.986855e-04
91	5	3.380046e-05	3.973710e-05
92	5	3.380046e-05	3.973710e-05

93	1	6.760091e-06	7.947420e-06
94	1	6.760091e-06	7.947420e-06
99	1	6.760091e-06	7.947420e-06
103	1	6.760091e-06	7.947420e-06
106	1	6.760091e-06	7.947420e-06
117	1	6.760091e-06	7.947420e-06
120	1	6.760091e-06	7.947420e-06
1816	1	6.760091e-06	7.947420e-06
1825	1	6.760091e-06	7.947420e-06
1952	1	6.760091e-06	7.947420e-06
1954	1	6.760091e-06	7.947420e-06
1957	1	6.760091e-06	7.947420e-06
1958	1	6.760091e-06	7.947420e-06
1959	1	6.760091e-06	7.947420e-06
1962	3	2.028027e-05	2.384226e-05
1964	2	1.352018e-05	1.589484e-05
1970	2	1.352018e-05	1.589484e-05
1972	4	2.704036e-05	3.178968e-05
1973	1	6.760091e-06	7.947420e-06
1974	2	1.352018e-05	1.589484e-05
1975	3	2.028027e-05	2.384226e-05
1977	1	6.760091e-06	7.947420e-06
1978	2	1.352018e-05	1.589484e-05
1981	1	6.760091e-06	7.947420e-06
1982	1	6.760091e-06	7.947420e-06
1983	3	2.028027e-05	2.384226e-05
1984	2	1.352018e-05	1.589484e-05
1985	4	2.704036e-05	3.178968e-05
1986	3	2.028027e-05	2.384226e-05
1987	3	2.028027e-05	2.384226e-05
1988	2	1.352018e-05	1.589484e-05
1989	4	2.704036e-05	3.178968e-05
1990	2	1.352018e-05	1.589484e-05
1991	6	4.056055e-05	4.768452e-05
1992	4	2.704036e-05	3.178968e-05
1993	7	4.732064e-05	5.563194e-05
1994	5	3.380046e-05	3.973710e-05
1995	4	2.704036e-05	3.178968e-05
1996	4	2.704036e-05	3.178968e-05
1997	7	4.732064e-05	5.563194e-05
1998	3	2.028027e-05	2.384226e-05
1999	3	2.028027e-05	2.384226e-05
2000	5	3.380046e-05	3.973710e-05

```
  2001     2 1.352018e-05  1.589484e-05
  2002     5 3.380046e-05  3.973710e-05
  2003     3 2.028027e-05  2.384226e-05
  2004     2 1.352018e-05  1.589484e-05
  2005     2 1.352018e-05  1.589484e-05
  2006     2 1.352018e-05  1.589484e-05
  2007     5 3.380046e-05  3.973710e-05
  2008     5 3.380046e-05  3.973710e-05
  2009     5 3.380046e-05  3.973710e-05
  2010     2 1.352018e-05  1.589484e-05
  2011     4 2.704036e-05  3.178968e-05
  2012     4 2.704036e-05  3.178968e-05
  2013     2 1.352018e-05  1.589484e-05
  2014     2 1.352018e-05  1.589484e-05
  2015     2 1.352018e-05  1.589484e-05
    NA 22100 1.493980e-01            NA
```

There are a lot of YOB that are put in as age, let's quickly fix that

```
s3_2000_2016 %>%
  mutate(age = if_else(age > 1000, year(date) - age, age)) %>%
  tabyl(age)
```

```
   age     n      percent valid_percent
 -34669     1 6.760091e-06  7.947420e-06
 -31868     1 6.760091e-06  7.947420e-06
 -31450     1 6.760091e-06  7.947420e-06
 -30781     1 6.760091e-06  7.947420e-06
 -28917     1 6.760091e-06  7.947420e-06
 -26725     1 6.760091e-06  7.947420e-06
 -17994     1 6.760091e-06  7.947420e-06
 -17993     2 1.352018e-05  1.589484e-05
  -4670     1 6.760091e-06  7.947420e-06
   -297     1 6.760091e-06  7.947420e-06
     -2     1 6.760091e-06  7.947420e-06
      0  2246 1.518316e-02  1.784991e-02
      1  4873 3.294192e-02  3.872778e-02
      2  2897 1.958398e-02  2.302368e-02
      3  2977 2.012479e-02  2.365947e-02
      4  3381 2.285587e-02  2.687023e-02
      5  3943 2.665504e-02  3.133668e-02
      6  4783 3.233352e-02  3.801251e-02
```

7	5283	3.571356e-02	4.198622e-02
8	5461	3.691686e-02	4.340086e-02
9	5499	3.717374e-02	4.370286e-02
10	5966	4.033070e-02	4.741431e-02
11	5958	4.027662e-02	4.735073e-02
12	6055	4.093235e-02	4.812163e-02
13	6166	4.168272e-02	4.900379e-02
14	5791	3.914769e-02	4.602351e-02
15	3883	2.624943e-02	3.085983e-02
16	2570	1.737343e-02	2.042487e-02
17	2438	1.648110e-02	1.937581e-02
18	2853	1.928654e-02	2.267399e-02
19	2932	1.982059e-02	2.330184e-02
20	3095	2.092248e-02	2.459726e-02
21	2950	1.994227e-02	2.344489e-02
22	2868	1.938794e-02	2.279320e-02
23	2764	1.868489e-02	2.196667e-02
24	2524	1.706247e-02	2.005929e-02
25	2247	1.518992e-02	1.785785e-02
26	2143	1.448688e-02	1.703132e-02
27	1970	1.331738e-02	1.565642e-02
28	1896	1.281713e-02	1.506831e-02
29	1723	1.164764e-02	1.369340e-02
30	1611	1.089051e-02	1.280329e-02
31	1443	9.754811e-03	1.146813e-02
32	1370	9.261325e-03	1.088797e-02
33	1193	8.064789e-03	9.481272e-03
34	968	6.543768e-03	7.693102e-03
35	931	6.293645e-03	7.399048e-03
36	857	5.793398e-03	6.810939e-03
37	770	5.205270e-03	6.119513e-03
38	672	4.542781e-03	5.340666e-03
39	570	3.853252e-03	4.530029e-03
40	531	3.589608e-03	4.220080e-03
41	487	3.292164e-03	3.870393e-03
42	399	2.697276e-03	3.171021e-03
43	378	2.555314e-03	3.004125e-03
44	306	2.068588e-03	2.431910e-03
45	295	1.994227e-03	2.344489e-03
46	272	1.838745e-03	2.161698e-03
47	240	1.622422e-03	1.907381e-03
48	215	1.453420e-03	1.708695e-03
49	183	1.237097e-03	1.454378e-03

50	213	1.439899e-03	1.692800e-03
51	204	1.379059e-03	1.621274e-03
52	142	9.599329e-04	1.128534e-03
53	127	8.585316e-04	1.009322e-03
54	124	8.382513e-04	9.854801e-04
55	109	7.368499e-04	8.662688e-04
56	105	7.098096e-04	8.344791e-04
57	75	5.070068e-04	5.960565e-04
58	117	7.909307e-04	9.298481e-04
59	74	5.002467e-04	5.881091e-04
60	66	4.461660e-04	5.245297e-04
61	55	3.718050e-04	4.371081e-04
62	64	4.326458e-04	5.086349e-04
63	51	3.447646e-04	4.053184e-04
64	51	3.447646e-04	4.053184e-04
65	37	2.501234e-04	2.940545e-04
66	27	1.825225e-04	2.145803e-04
67	29	1.960426e-04	2.304752e-04
68	28	1.892826e-04	2.225278e-04
69	26	1.757624e-04	2.066329e-04
70	33	2.230830e-04	2.622649e-04
71	23	1.554821e-04	1.827907e-04
72	23	1.554821e-04	1.827907e-04
73	17	1.149215e-04	1.351061e-04
74	15	1.014014e-04	1.192113e-04
75	12	8.112109e-05	9.536904e-05
76	13	8.788118e-05	1.033165e-04
77	6	4.056055e-05	4.768452e-05
78	8	5.408073e-05	6.357936e-05
79	9	6.084082e-05	7.152678e-05
80	21	1.419619e-04	1.668958e-04
81	13	8.788118e-05	1.033165e-04
82	8	5.408073e-05	6.357936e-05
83	4	2.704036e-05	3.178968e-05
84	5	3.380046e-05	3.973710e-05
85	3	2.028027e-05	2.384226e-05
86	2	1.352018e-05	1.589484e-05
87	3	2.028027e-05	2.384226e-05
88	3	2.028027e-05	2.384226e-05
90	25	1.690023e-04	1.986855e-04
91	5	3.380046e-05	3.973710e-05
92	5	3.380046e-05	3.973710e-05
93	1	6.760091e-06	7.947420e-06

```
 94     1 6.760091e-06  7.947420e-06
 99     1 6.760091e-06  7.947420e-06
103     1 6.760091e-06  7.947420e-06
106     1 6.760091e-06  7.947420e-06
117     1 6.760091e-06  7.947420e-06
120     1 6.760091e-06  7.947420e-06
191     1 6.760091e-06  7.947420e-06
200     1 6.760091e-06  7.947420e-06
 NA 22100 1.493980e-01            NA
```

Then let's drop negative age and age > 90

```
s4_2000_2016 <- s3_2000_2016 %>%
  mutate(age = if_else(age > 1000, year(date) - age, age)) %>%
  filter(age >= 0, age < 91)

s4_2000_2016 %>% skim()
```

Table 10: Data summary

Name	Piped data
Number of rows	125796
Number of columns	8
_____	
Column type frequency:	
character	6
Date	1
numeric	1
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
sex	2774	0.98	1	1	0	7	0
district	0	1.00	2	10	0	24	0
commune	0	1.00	2	16	0	171	0
hospital	0	1.00	7	18	0	37	0
icd	5280	0.96	3	6	0	35	0
in_out_patient	0	1.00	10	10	0	1	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
date	0	1	2000-01-01	2016-12-31	2011-01-29	6128

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	15.95	11.67	0	8	13	22	90	

Wrap up

Probably finished cleaning 2000-2016 data for now

Let's check final number of rows

```
nrow(s4_2000_2016)
```

```
[1] 125796
```

```
start_nrow - nrow(s4_2000_2016)
```

```
[1] 22131
```

```
(start_nrow - nrow(s4_2000_2016)) / start_nrow * 100
```

```
[1] 14.96076
```

Lost about 15% of rows

Data from 2017-2022

Extract the data

```
raw_2017_2022 <- xlsx_raws[length(xlsx_raws)][[1]]
```

Quick skim at the data

```
skim(raw_2017_2022)
```

Table 14: Data summary

Name	raw_2017_2022
Number of rows	268738
Number of columns	21
_____	
Column type frequency:	
character	15
numeric	2
POSIXct	4
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
gioi	0	1.00	2	3	0	11	0
tinh_noi_o	0	1.00	5	9	0	3	0
quan_huyen_noi_o	0	1.00	6	16	0	25	0
phuong_xa_noi_o	0	1.00	8	23	0	181	0
tinh_trang_hien_tai	46	1.00	2	170	0	180	0
don_vi_bao_cao	1	1.00	4	56	0	530	0
tinh_bao_cao	135	1.00	3	20	0	70	0
phan_do	201045	0.25	1	2	0	15	0
lay_mau	21313	0.92	2	25	0	128	0
ngay_lay_mau	194951	0.27	2	22	0	2498	0
loai_xet_nghiem	144446	0.46	3	54	0	34	0
ket_qua_xet_nghiem	147225	0.45	4	15	0	13	0
don_vi_xet_nghiem	235984	0.12	4	66	0	340	0
x1_mabtt	0	1.00	2	2	0	1	0
benh_kem	265104	0.01	1	255	0	1365	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
stt	0	1	134369.50	77578.12	1	67185.25	134369.5	201553.8	268738	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
tuoi	39	1	22.45	27.14	-7974	11.00	20.0	31.0	2021	

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
ngay_khoi_phat	8881	0.97	1974-10-29	2219-06-14	2019-10-08	2252
ngay_nhap_vien	0	1.00	2017-01-01	2022-12-31	2019-10-26	2191
ngay_xuat_vien	192154	0.28	1899-12-31	2023-02-21	2019-01-09	2155
ngay_bao_cao	0	1.00	2018-01-02	2023-02-22	2019-10-31	1087

The actual columns based on our columns of interest are (with `complete_rate`): - `sex` = `gioi` (1) - `age` = `tuoi` (1.00) - `date of admission` = `ngay_nhap_vien` (1) - `district` = `quan_huyen_noi_o` (1) - `commune` = `phuong_xa_noi_o` (1) - `hospital` = `don_vi_bao_cao` (1.00) – lots of cases out of HCMC -> filter with `tinh_bao_cao` (0.999) - `icd` = `phan_do` (0.252) - `in-patient` = `tinh_trang_hien_tai` (1.00) – very complex freetext

Selecting columns of interest

```
s1_2017_2022 <- raw_2017_2022 %>%
  select(
    gioi,
    tuoi,
    ngay_nhap_vien,
    quan_huyen_noi_o,
    phuong_xa_noi_o,
    don_vi_bao_cao,
    tinh_bao_cao,
    phan_do,
    tinh_trang_hien_tai
  )
s1_2017_2022 %>% skim()
```

Table 18: Data summary

Name	Piped data
Number of rows	268738
Number of columns	9
_____	
Column type frequency:	
character	7
numeric	1
POSIXct	1
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
gioi	0	1.00	2	3	0	11	0
quan_huyen_noi_o	0	1.00	6	16	0	25	0
phuong_xa_noi_o	0	1.00	8	23	0	181	0
don_vi_bao_cao	1	1.00	4	56	0	530	0
tinh_bao_cao	135	1.00	3	20	0	70	0
phan_do	201045	0.25	1	2	0	15	0
tinh_trang_hien_tai	46	1.00	2	170	0	180	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
tuoi	39	1	22.45	27.14	-7974	11	20	31	2021	

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
ngay_nhap_vien	0	1	2017-01-01	2022-12-31	2019-10-26	2191

Keeping track of raw number of rows at start to see how much lost during data cleaning

```
start_nrow2 <- nrow(s1_2017_2022)
start_nrow2
```

```
[1] 268738
```

Filter cases outside of HCMC

First, let's filter out cases that are reported outside of HCMC.

```
s1_2017_2022 %>% tabyl(tinh_bao_cao)
```

```
         tinh_bao_cao      n      percent valid_percent
             An Giang    260 9.674851e-04  9.679713e-04
           BV An Sinh      6 2.232658e-05  2.233780e-05
         Bà Rịa-V.Tàu    126 4.688581e-04  4.690938e-04
           Bình Dương   2514 9.354836e-03  9.359538e-03
           Bình Phước     13 4.837425e-05  4.839857e-05
           Bình Thuận     44 1.637282e-04  1.638105e-04
            Bình Định    105 3.907151e-04  3.909115e-04
             Bạc Liêu     24 8.930631e-05  8.935120e-05
             Bắc Ninh      2 7.442193e-06  7.445933e-06
              Bến Tre    182 6.772395e-04  6.775799e-04
 Bệnh viện Nhi đồng 1    192 7.144505e-04  7.148096e-04
 Bệnh viện Nhi đồng 2     60 2.232658e-04  2.233780e-04
    Bệnh viện Quận 4      10 3.721096e-05  3.722967e-05
              Cà Mau      24 8.930631e-05  8.935120e-05
             Cần Thơ      91 3.386198e-04  3.387900e-04
             Gia Lai      14 5.209535e-05  5.212153e-05
                 HCM      12 4.465316e-05  4.467560e-05
              Hà Nam       1 3.721096e-06  3.722967e-06
              Hà Nội      34 1.265173e-04  1.265809e-04
             Hà Tĩnh       1 3.721096e-06  3.722967e-06
            Hưng Yên       2 7.442193e-06  7.445933e-06
            Hải Dương      5 1.860548e-05  1.861483e-05
            Hải Phòng      7 2.604767e-05  2.606077e-05
            Hậu Giang     20 7.442193e-05  7.445933e-05
          Hồ Chí Minh     16 5.953754e-05  5.956747e-05
            Khánh Hòa    113 4.204839e-04  4.206952e-04
           Kiên Giang     53 1.972181e-04  1.973172e-04
              Kon Tum      3 1.116329e-05  1.116890e-05
```

Long An	378	1.406574e-03	1.407281e-03
Lào Cai	1	3.721096e-06	3.722967e-06
Lâm Đồng	126	4.688581e-04	4.690938e-04
Nam Định	3	1.116329e-05	1.116890e-05
Nghệ An	3	1.116329e-05	1.116890e-05
Ninh Thuận	15	5.581645e-05	5.584450e-05
Phú Yên	70	2.604767e-04	2.606077e-04
Quảng Bình	6	2.232658e-05	2.233780e-05
Quảng Nam	50	1.860548e-04	1.861483e-04
Quảng Ngãi	119	4.428105e-04	4.430330e-04
Quảng Ninh	3	1.116329e-05	1.116890e-05
Quảng Trị	8	2.976877e-05	2.978373e-05
Sóc Trăng	54	2.009392e-04	2.010402e-04
TP . H.C.M	1	3.721096e-06	3.722967e-06
TP HCM	1	3.721096e-06	3.722967e-06
TP Hồ Chí Minh	8	2.976877e-05	2.978373e-05
TP. H.C.M	243423	9.058004e-01	9.062557e-01
TP. HCM	183	6.809606e-04	6.813029e-04
TP. HỒ CHÍ MINH	13	4.837425e-05	4.839857e-05
TP.C.M	37	1.376806e-04	1.377498e-04
TP.H.C.M	10293	3.830125e-02	3.832050e-02
TP.HCM	629	2.340570e-03	2.341746e-03
TP.Hồ Chí Minh	10	3.721096e-05	3.722967e-05
TPHCM	8191	3.047950e-02	3.049482e-02
TT- Huế	13	4.837425e-05	4.839857e-05
Thanh Hóa	9	3.348987e-05	3.350670e-05
Thái Bình	3	1.116329e-05	1.116890e-05
Thái Nguyên	1	3.721096e-06	3.722967e-06
Tiền Giang	171	6.363075e-04	6.366273e-04
Tp HCM	3	1.116329e-05	1.116890e-05
Tp.HCM	18	6.697973e-05	6.701340e-05
TpHCM	3	1.116329e-05	1.116890e-05
TpHcm	37	1.376806e-04	1.377498e-04
TpP.H.C.M	15	5.581645e-05	5.584450e-05
Trà Vinh	38	1.414017e-04	1.414727e-04
Tây Ninh	71	2.641978e-04	2.643306e-04
Vĩnh Long	66	2.455924e-04	2.457158e-04
Đà Nẵng	13	4.837425e-05	4.839857e-05
Đắk Lắk	48	1.786126e-04	1.787024e-04
Đắk Nông	33	1.227962e-04	1.228579e-04
Đồng Nai	327	1.216799e-03	1.217410e-03
Đồng Tháp	175	6.511919e-04	6.515192e-04
<NA>	135	5.023480e-04	NA

Filter based on reporting province

```
s2_2017_2022 <- s1_2017_2022 %>%
  mutate(
    cleaned_tinh_bao_cao = tolower(tinh_bao_cao) %>%
      stri_trans_general(id = "Latin-ASCII") %>%
      str_replace_all("[. -]", "")
  ) %>%
  filter(
    cleaned_tinh_bao_cao %in%
      c(
        "benhviennhidong1",
        "benhviennhidong2",
        "benhvienquan4",
        "bvansinh",
        "hcm",
        "hochiminh",
        "tpcm",
        "tphcm",
        "tphochiminh",
        "tpphcm"
      ) |
      is.na(cleaned_tinh_bao_cao)
  )

s2_2017_2022 %>% tabyl(cleaned_tinh_bao_cao)
```

```
 cleaned_tinh_bao_cao      n      percent valid_percent
     benhviennhidong1    192 7.292173e-04  7.295914e-04
     benhviennhidong2     60 2.278804e-04  2.279973e-04
        benhvienquan4     10 3.798007e-05  3.799955e-05
             bvansinh      6 2.278804e-05  2.279973e-05
                  hcm     12 4.557608e-05  4.559946e-05
            hochiminh     16 6.076811e-05  6.079928e-05
                 tpcm     37 1.405263e-04  1.405983e-04
                tphcm 262782 9.980478e-01  9.985598e-01
          tphochiminh     31 1.177382e-04  1.177986e-04
               tpphcm     15 5.697010e-05  5.699933e-05
                 <NA>    135 5.127309e-04            NA
```

```
s2_2017_2022 %>% skim()
```

Table 22: Data summary

Name	Piped data
Number of rows	263296
Number of columns	10
_____	
Column type frequency:	
character	8
numeric	1
POSIXct	1
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
gioi	0	1.00	2	3	0	11	0
quan_huyen_noi_o	0	1.00	6	16	0	25	0
phuong_xa_noi_o	0	1.00	8	23	0	181	0
don_vi_bao_cao	1	1.00	4	56	0	118	0
tinh_bao_cao	135	1.00	3	20	0	22	0
phan_do	197169	0.25	1	2	0	15	0
tinh_trang_hien_tai	46	1.00	2	170	0	174	0
cleaned_tinh_bao_cao	135	1.00	3	16	0	10	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
tuoi	39	1	22.41	27.34	-7974	11	20	31	2021	

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
ngay_nhap_vien	0	1	2017-01-01	2022-12-31	2019-10-26	2191

Filter based on reporting hospital

```
s2_2017_2022 %>% tabyl(don_vi_bao_cao)
```

don_vi_bao_cao	n	percent
Bênh viện Bệnh nhiệt đới TPHCM	72139	2.739844e-01
Bênh viện Pháp Việt	10	3.798007e-05
Bệnh Viện Columbia Gia Định	10	3.798007e-05
Bệnh Viện Hoàn Hảo	1	3.798007e-06
Bệnh Viện Đức Khang	50	1.899003e-04
Bệnh viện 175	19	7.216213e-05
Bệnh viện An Bình	281	1.067240e-03
Bệnh viện An Sinh	353	1.340696e-03
Bệnh viện Chợ Rẫy	502	1.906599e-03
Bệnh viện Gaya Việt Hàn	76	2.886485e-04
Bệnh viện Gia An 115	231	8.773396e-04
Bệnh viện Hoàn Mỹ	8037	3.052458e-02
Bệnh viện Huyện Bình Chánh	1700	6.456612e-03
Bệnh viện Huyện Cần Giờ	833	3.163740e-03
Bệnh viện Huyện Củ Chi	1141	4.333526e-03
Bệnh viện Huyện Nhà Bè	1728	6.562956e-03
Bệnh viện Hồng Đức III	1387	5.267835e-03
Bệnh viện Lê Lợi - Bà Rịa-V.Tàu	1	3.798007e-06
Bệnh viện Minh Anh	33	1.253342e-04
Bệnh viện Mỹ Đức	545	2.069914e-03
Bệnh viện Mỹ Đức Phú Nhuận	3	1.139402e-05
Bệnh viện Nguyễn Tri Phương	576	2.187652e-03
Bệnh viện Nguyễn Trãi	693	2.632019e-03
Bệnh viện Nhi Đồng 1	20	7.596014e-05
Bệnh viện Nhi đồng 1	21078	8.005439e-02
Bệnh viện Nhi đồng 2	21522	8.174070e-02
Bệnh viện Nhi đồng thành phố	5416	2.057000e-02
Bệnh viện Nhân Dân 115	1122	4.261364e-03
Bệnh viện Nhân Dân Gia Định	471	1.788861e-03
Bệnh viện Nhân dân Gia Định	4	1.519203e-05
Bệnh viện Pháp Việt	1855	7.045303e-03
Bệnh viện Phụ sản MêKông	26	9.874818e-05

Bệnh viện Phục hồi chức năng - Điều trị Bệnh nghề nghiệp	18	6.836412e-05
Bệnh viện Quân dân Y Miền Đông	1441	5.472928e-03
Bệnh viện Quận 1	3552	1.349052e-02
Bệnh viện Quận 10	419	1.591365e-03
Bệnh viện Quận 11	3623	1.376018e-02
Bệnh viện Quận 12	10260	3.896755e-02
Bệnh viện Quận 2	1583	6.012245e-03
Bệnh viện Quận 3	609	2.312986e-03
Bệnh viện Quận 4	1498	5.689414e-03
Bệnh viện Quận 5	474	1.800255e-03
Bệnh viện Quận 6	1202	4.565204e-03
Bệnh viện Quận 7	1402	5.324806e-03
Bệnh viện Quận 8	1090	4.139827e-03
Bệnh viện Quận 9	909	3.452388e-03
Bệnh viện Quận Bình Thạnh	10482	3.981071e-02
Bệnh viện Quận Bình Tân	8322	3.160701e-02
Bệnh viện Quận Gò Vấp	1582	6.008447e-03
Bệnh viện Quận Phú Nhuận	1041	3.953725e-03
Bệnh viện Quận Thủ Đức	3276	1.244227e-02
Bệnh viện Quận Tân Bình	5610	2.130682e-02
Bệnh viện Quận Tân Phú	26189	9.946600e-02
Bệnh viện Quốc Tế Mỹ	257	9.760877e-04
Bệnh viện Quốc tế Becamex	2	7.596014e-06
Bệnh viện Quốc tế City	488	1.853427e-03
Bệnh viện Quốc Ánh	808	3.068789e-03
Bệnh viện Thống Nhất	1530	5.810950e-03
Bệnh viện Triều An	514	1.952175e-03
Bệnh viện Trưng Vương	4398	1.670363e-02
Bệnh viện Tâm Trí Sài Gòn	665	2.525675e-03
Bệnh viện Tân Hưng	472	1.792659e-03
Bệnh viện Từ Dũ	31	1.177382e-04
Bệnh viện Vinmec	190	7.216213e-04
Bệnh viện Vạn Hạnh	38	1.443243e-04
Bệnh viện Xuyên Á	3799	1.442863e-02
Bệnh viện huyện Bình Chánh	11	4.177807e-05
Bệnh viện huyện Củ Chi	5	1.899003e-05
Bệnh viện quận Bình Tân	37	1.405263e-04
Bệnh viện quận Tân Bình	14	5.317210e-05
Bệnh viện ĐKKV Củ Chi	7486	2.843188e-02
Bệnh viện ĐKKV Hóc Môn	6654	2.527194e-02
Bệnh viện ĐKKV Thủ Đức	4873	1.850769e-02
Bệnh viện ĐKQT Nam Sài Gòn	15	5.697010e-05
Bệnh viện Đa khoa Bưu Điện-CS1	540	2.050924e-03

Bệnh viện Đa khoa Bưu Điện-CS3	1	3.798007e-06
Bệnh viện Đa khoa Hòan Hảo (Cs2)	2	7.596014e-06
Bệnh viện Đa khoa Quốc tế Nam Sài Gòn	1	3.798007e-06
Bệnh viện Đa khoa Sài Gòn	6	2.278804e-05
Bệnh viện Đa khoa Tâm Anh	1295	4.918419e-03
Bệnh viện Đa khoa khu vực Hậu Nghĩa	1	3.798007e-06
Bệnh viện Đại học Y Dược TPHCM	702	2.666201e-03
Bệnh viện đa khoa 30/4 Tp.HCM	713	2.707979e-03
Bệnh viện đa khoa Quốc Tế Hoàn Mỹ Thủ Đức	93	3.532146e-04
Bệnh viện đa khoa Sài Gòn	479	1.819245e-03
Bệnh viện đa khoa quốc tế Vũ Anh	438	1.663527e-03
Bệnh viện đa khoa tỉnh Bình Dương	1	3.798007e-06
Bệnh viện đa khoa tỉnh Đắk Lắk	1	3.798007e-06
HCDC	1	3.798007e-06
PK THANH CONG	21	7.975814e-05
PKĐK Phước Lợi - Long An	1	3.798007e-06
Phòng khám đa khoa Trần Diệp Khanh	87	3.304266e-04
TTYT Thị xã Trảng Bàng	1	3.798007e-06
Trung tâm Kiểm soát bệnh tật Tp.Hồ Chí Minh	2	7.596014e-06
Trung tâm Y tế Huyện Vĩnh Hưng	2	7.596014e-06
Trung tâm Y tế Huyện Xuyên Mộc	1	3.798007e-06
Trung tâm Y tế Quận 1	1	3.798007e-06
Trung tâm Y tế Quận 3	1	3.798007e-06
Trung tâm Y tế Quận 9	112	4.253768e-04
Trung tâm Y tế Quận Bình Thạnh	5	1.899003e-05
Trung tâm Y tế Quận Bình Tân	1	3.798007e-06
Trung tâm Y tế Quận Phú Nhuận	2	7.596014e-06
Trung tâm kiểm soát bệnh tật Long An	1	3.798007e-06
Trung tâm kiểm soát bệnh tật Thành phố HCM	1	3.798007e-06
Trung tâm y tế Dự phòng TP. H.C.M	1	3.798007e-06
Trung tâm y tế tỉnh Trà Vinh	1	3.798007e-06
Trạm Y tế Phuoc Tan Hung	2	7.596014e-06
Trạm Y tế Phường 03	1	3.798007e-06
Trạm Y tế Phường An Lợi Đông	2	7.596014e-06
Trạm Y tế Phường An Phú	5	1.899003e-05
Trạm Y tế Phường Bình An	2	7.596014e-06
Trạm Y tế Phường Bình Khánh	1	3.798007e-06
Trạm Y tế Phường Bình Trưng Đông	3	1.139402e-05
Trạm Y tế Phường Cát Lái	1	3.798007e-06
Trạm Y tế xã Hiệp Hòa	1	3.798007e-06
Trạm y tế Bình Trưng Tây	3	1.139402e-05
Viện Tim TPHCM	3	1.139402e-05
bv.nd2.khth@gmail.com	24	9.115216e-05

```
                                                    <NA>      1 3.798007e-06
valid_percent
 2.739855e-01
 3.798021e-05
 3.798021e-05
 3.798021e-06
 1.899011e-04
 7.216240e-05
 1.067244e-03
 1.340701e-03
 1.906607e-03
 2.886496e-04
 8.773429e-04
 3.052470e-02
 6.456636e-03
 3.163752e-03
 4.333542e-03
 6.562981e-03
 5.267855e-03
 3.798021e-06
 1.253347e-04
 2.069922e-03
 1.139406e-05
 2.187660e-03
 2.632029e-03
 7.596042e-05
 8.005469e-02
 8.174101e-02
 2.057008e-02
 4.261380e-03
 1.788868e-03
 1.519208e-05
 7.045329e-03
 9.874855e-05
 6.836438e-05
 5.472949e-03
 1.349057e-02
 1.591371e-03
 1.376023e-02
 3.896770e-02
 6.012268e-03
 2.312995e-03
 5.689436e-03
```

1.800262e-03
4.565222e-03
5.324826e-03
4.139843e-03
3.452401e-03
3.981086e-02
3.160713e-02
6.008470e-03
3.953740e-03
1.244232e-02
2.130690e-02
9.946638e-02
9.760915e-04
7.596042e-06
1.853434e-03
3.068801e-03
5.810972e-03
1.952183e-03
1.670370e-02
2.525684e-03
1.792666e-03
1.177387e-04
7.216240e-04
1.443248e-04
1.442868e-02
4.177823e-05
1.899011e-05
1.405268e-04
5.317230e-05
2.843199e-02
2.527203e-02
1.850776e-02
5.697032e-05
2.050931e-03
3.798021e-06
7.596042e-06
3.798021e-06
2.278813e-05
4.918437e-03
3.798021e-06
2.666211e-03
2.707989e-03
3.532160e-04

```
  1.819252e-03
  1.663533e-03
  3.798021e-06
  3.798021e-06
  3.798021e-06
  7.975845e-05
  3.798021e-06
  3.304278e-04
  3.798021e-06
  7.596042e-06
  7.596042e-06
  3.798021e-06
  3.798021e-06
  3.798021e-06
  4.253784e-04
  1.899011e-05
  3.798021e-06
  7.596042e-06
  3.798021e-06
  3.798021e-06
  3.798021e-06
  3.798021e-06
  7.596042e-06
  3.798021e-06
  7.596042e-06
  1.899011e-05
  7.596042e-06
  3.798021e-06
  1.139406e-05
  3.798021e-06
  3.798021e-06
  1.139406e-05
  1.139406e-05
  9.115251e-05
           NA
```

Manual cleaning based on an "eye test" and some quick Googling

```
s2b_2017_2022 <- s2_2017_2022 %>%
  filter(
    !(don_vi_bao_cao %in%
      c(
```

```
        "Bệnh viện Lê Lợi - Bà Rịa-V.Tàu",
        "Bệnh viện Quốc tế Becamex",
        "Bệnh viện đa khoa tỉnh Bình Dương",
        "Bệnh viện đa khoa tỉnh Đắk Lắk",
        "PKĐK Phước Lợi - Long An",
        "TTYT Thị xã Trảng Bàng",
        "Trung tâm Y tế  Huyện Vĩnh Hưng",
        "Trung tâm Y tế  Huyện Xuyên Mộc",
        "Trung tâm kiểm soát bệnh tật Long An",
        "Trung tâm y tế tỉnh Trà Vinh"
      ))
  )
s2b_2017_2022 %>% skim()
```

Table 26: Data summary

Name	Piped data
Number of rows	263284
Number of columns	10
_____	
Column type frequency:	
character	8
numeric	1
POSIXct	1
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
gioi	0	1.00	2	3	0	11	0
quan_huyen_noi_o	0	1.00	6	16	0	25	0
phuong_xa_noi_o	0	1.00	8	23	0	181	0
don_vi_bao_cao	1	1.00	4	56	0	108	0
tinh_bao_cao	133	1.00	3	20	0	22	0
phan_do	197162	0.25	1	2	0	15	0
tinh_trang_hien_tai	46	1.00	2	170	0	174	0
cleaned_tinh_bao_cao	133	1.00	3	16	0	10	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
tuoi	39	1	22.41	27.34	-7974	11	20	31	2021	

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
ngay_nhap_vien	0	1	2017-01-01	2022-12-31	2019-10-26	2191

Filter out cases where the reporting hospital is null (only 1 case)

```
s2c_2017_2022 <- s2b_2017_2022 %>% drop_na(don_vi_bao_cao)
s2c_2017_2022 %>% skim()
```

Table 30: Data summary

Name	Piped data
Number of rows	263283
Number of columns	10
_____	
Column type frequency:	
character	8
numeric	1
POSIXct	1
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
gioi	0	1.00	2	3	0	11	0
quan_huyen_noi_o	0	1.00	6	16	0	25	0
phuong_xa_noi_o	0	1.00	8	23	0	181	0
don_vi_bao_cao	0	1.00	4	56	0	108	0
tinh_bao_cao	132	1.00	3	20	0	22	0
phan_do	197162	0.25	1	2	0	15	0
tinh_trang_hien_tai	46	1.00	2	170	0	174	0

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
cleaned_tinh_bao_cao	132	1.00	3	16	0	10	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
tuoi	39	1	22.41	27.34	-7974	11	20	31	2021	

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
ngay_nhap_vien	0	1	2017-01-01	2022-12-31	2019-10-26	2191

Wrapping this up

```
s3_2017_2022 <- s2c_2017_2022 %>% select(-tinh_bao_cao, -cleaned_tinh_bao_cao)
```

Rename columns

Let's rename columns before continuing

```
s4_2017_2022 <- s3_2017_2022 %>%
  rename(
    sex = gioi,
    age = tuoi,
    date = ngay_nhap_vien,
    district = quan_huyen_noi_o,
    commune = phuong_xa_noi_o,
    hospital = don_vi_bao_cao,
    icd = phan_do,
    in_out_patient = tinh_trang_hien_tai
  )

s4_2017_2022 %>% skim()
```

Table 34: Data summary

Name	Piped data
Number of rows	263283
Number of columns	8
_____	
Column type frequency:	
character	6
numeric	1
POSIXct	1
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
sex	0	1.00	2	3	0	11	0
district	0	1.00	6	16	0	25	0
commune	0	1.00	8	23	0	181	0
hospital	0	1.00	4	56	0	108	0
icd	197162	0.25	1	2	0	15	0
in_out_patient	46	1.00	2	170	0	174	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	39	1	22.41	27.34	-7974	11	20	31	2021	

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
date	0	1	2017-01-01	2022-12-31	2019-10-26	2191

Fix age

We see the same problem with some negative age and YOB put in as age

```r
s4_2017_2022 %>% tabyl(age)
```

```
   age    n      percent valid_percent
 -7974    1 3.798194e-06  3.798757e-06
 -7153    1 3.798194e-06  3.798757e-06
  -974    1 3.798194e-06  3.798757e-06
   -77    1 3.798194e-06  3.798757e-06
   -11    1 3.798194e-06  3.798757e-06
    -1    1 3.798194e-06  3.798757e-06
     0 1200 4.557833e-03  4.558508e-03
     1 7559 2.871055e-02  2.871480e-02
     2 4725 1.794647e-02  1.794913e-02
     3 4596 1.745650e-02  1.745909e-02
     4 4587 1.742232e-02  1.742490e-02
     5 5302 2.013803e-02  2.014101e-02
     6 6846 2.600244e-02  2.600629e-02
     7 7291 2.769263e-02  2.769674e-02
     8 7340 2.787875e-02  2.788288e-02
     9 7233 2.747234e-02  2.747641e-02
    10 8268 3.140347e-02  3.140812e-02
    11 8172 3.103884e-02  3.104344e-02
    12 7587 2.881690e-02  2.882117e-02
    13 8120 3.084134e-02  3.084591e-02
    14 7867 2.988039e-02  2.988482e-02
    15 7156 2.717988e-02  2.718391e-02
    16 6091 2.313480e-02  2.313823e-02
    17 5395 2.049126e-02  2.049429e-02
    18 5804 2.204472e-02  2.204799e-02
    19 6525 2.478322e-02  2.478689e-02
    20 5904 2.242454e-02  2.242786e-02
    21 5779 2.194977e-02  2.195302e-02
    22 5970 2.267522e-02  2.267858e-02
    23 5908 2.243973e-02  2.244306e-02
    24 5863 2.226881e-02  2.227211e-02
    25 6210 2.358679e-02  2.359028e-02
    26 6100 2.316899e-02  2.317242e-02
    27 6205 2.356780e-02  2.357129e-02
    28 5859 2.225362e-02  2.225692e-02
    29 5775 2.193457e-02  2.193782e-02
    30 5381 2.043808e-02  2.044111e-02
    31 5021 1.907073e-02  1.907356e-02
    32 4856 1.844403e-02  1.844676e-02
```

```
33 4565 1.733876e-02  1.734133e-02
34 4414 1.676523e-02  1.676771e-02
35 4188 1.590684e-02  1.590919e-02
36 3797 1.442174e-02  1.442388e-02
37 3621 1.375326e-02  1.375530e-02
38 3271 1.242389e-02  1.242573e-02
39 3155 1.198330e-02  1.198508e-02
40 2798 1.062735e-02  1.062892e-02
41 2515 9.552459e-03  9.553874e-03
42 2305 8.754838e-03  8.756135e-03
43 2138 8.120539e-03  8.121743e-03
44 1993 7.569801e-03  7.570923e-03
45 1760 6.684822e-03  6.685812e-03
46 1705 6.475921e-03  6.476881e-03
47 1503 5.708686e-03  5.709532e-03
48 1355 5.146553e-03  5.147316e-03
49 1277 4.850294e-03  4.851013e-03
50 1169 4.440089e-03  4.440747e-03
51 1083 4.113444e-03  4.114054e-03
52  991 3.764011e-03  3.764568e-03
53  955 3.627276e-03  3.627813e-03
54  887 3.368998e-03  3.369498e-03
55  802 3.046152e-03  3.046603e-03
56  754 2.863839e-03  2.864263e-03
57  765 2.905619e-03  2.906049e-03
58  693 2.632149e-03  2.632539e-03
59  607 2.305504e-03  2.305846e-03
60  608 2.309302e-03  2.309644e-03
61  541 2.054823e-03  2.055128e-03
62  527 2.001648e-03  2.001945e-03
63  455 1.728178e-03  1.728434e-03
64  432 1.640820e-03  1.641063e-03
65  345 1.310377e-03  1.310571e-03
66  350 1.329368e-03  1.329565e-03
67  279 1.059696e-03  1.059853e-03
68  279 1.059696e-03  1.059853e-03
69  224 8.507955e-04  8.509216e-04
70  197 7.482443e-04  7.483551e-04
71  162 6.153075e-04  6.153986e-04
72  136 5.165544e-04  5.166310e-04
73  137 5.203526e-04  5.204297e-04
74  119 4.519851e-04  4.520521e-04
75  108 4.102050e-04  4.102658e-04
```

76	87	3.304429e-04	3.304919e-04
77	75	2.848646e-04	2.849068e-04
78	65	2.468826e-04	2.469192e-04
79	63	2.392862e-04	2.393217e-04
80	55	2.089007e-04	2.089316e-04
81	63	2.392862e-04	2.393217e-04
82	52	1.975061e-04	1.975354e-04
83	36	1.367350e-04	1.367553e-04
84	28	1.063494e-04	1.063652e-04
85	31	1.177440e-04	1.177615e-04
86	27	1.025512e-04	1.025664e-04
87	22	8.356028e-05	8.357266e-05
88	26	9.875305e-05	9.876768e-05
89	24	9.115666e-05	9.117017e-05
90	17	6.456930e-05	6.457887e-05
91	8	3.038555e-05	3.039006e-05
92	12	4.557833e-05	4.558508e-05
93	3	1.139458e-05	1.139627e-05
94	6	2.278917e-05	2.279254e-05
95	3	1.139458e-05	1.139627e-05
96	4	1.519278e-05	1.519503e-05
98	2	7.596389e-06	7.597514e-06
99	1	3.798194e-06	3.798757e-06
101	2	7.596389e-06	7.597514e-06
110	1	3.798194e-06	3.798757e-06
112	8	3.038555e-05	3.039006e-05
113	3	1.139458e-05	1.139627e-05
114	6	2.278917e-05	2.279254e-05
117	3	1.139458e-05	1.139627e-05
122	4	1.519278e-05	1.519503e-05
147	1	3.798194e-06	3.798757e-06
195	1	3.798194e-06	3.798757e-06
197	2	7.596389e-06	7.597514e-06
198	6	2.278917e-05	2.279254e-05
199	4	1.519278e-05	1.519503e-05
200	1	3.798194e-06	3.798757e-06
223	1	3.798194e-06	3.798757e-06
232	2	7.596389e-06	7.597514e-06
345	1	3.798194e-06	3.798757e-06
427	1	3.798194e-06	3.798757e-06
440	1	3.798194e-06	3.798757e-06
551	1	3.798194e-06	3.798757e-06
819	1	3.798194e-06	3.798757e-06

```
  820    1 3.798194e-06  3.798757e-06
  821    4 1.519278e-05  1.519503e-05
  822    2 7.596389e-06  7.597514e-06
  823    1 3.798194e-06  3.798757e-06
  824    2 7.596389e-06  7.597514e-06
  825    2 7.596389e-06  7.597514e-06
  925    1 3.798194e-06  3.798757e-06
 1010    1 3.798194e-06  3.798757e-06
 1013    1 3.798194e-06  3.798757e-06
 1038    1 3.798194e-06  3.798757e-06
 2013    1 3.798194e-06  3.798757e-06
 2021    1 3.798194e-06  3.798757e-06
   NA   39 1.481296e-04             NA
```

```r
s5_2017_2022 <- s4_2017_2022 %>%
  mutate(age = if_else(age > 2000, year(date) - age, age)) %>%
  filter(age >= 0, age < 91)

s5_2017_2022 %>% tabyl(age)
```

```
 age    n      percent
   0 1200 4.560431e-03
   1 7560 2.873072e-02
   2 4725 1.795670e-02
   3 4596 1.746645e-02
   4 4587 1.743225e-02
   5 5302 2.014951e-02
   6 6846 2.601726e-02
   7 7291 2.770842e-02
   8 7340 2.789464e-02
   9 7234 2.749180e-02
  10 8268 3.142137e-02
  11 8172 3.105654e-02
  12 7587 2.883333e-02
  13 8120 3.085892e-02
  14 7867 2.989743e-02
  15 7156 2.719537e-02
  16 6091 2.314799e-02
  17 5395 2.050294e-02
  18 5804 2.205729e-02
  19 6525 2.479735e-02
  20 5904 2.243732e-02
```

```
21 5779 2.196228e-02
22 5970 2.268815e-02
23 5908 2.245252e-02
24 5863 2.228151e-02
25 6210 2.360023e-02
26 6100 2.318219e-02
27 6205 2.358123e-02
28 5859 2.226631e-02
29 5775 2.194708e-02
30 5381 2.044973e-02
31 5021 1.908161e-02
32 4856 1.845455e-02
33 4565 1.734864e-02
34 4414 1.677479e-02
35 4188 1.591591e-02
36 3797 1.442997e-02
37 3621 1.376110e-02
38 3271 1.243098e-02
39 3155 1.199013e-02
40 2798 1.063341e-02
41 2515 9.557904e-03
42 2305 8.759829e-03
43 2138 8.125169e-03
44 1993 7.574117e-03
45 1760 6.688633e-03
46 1705 6.479613e-03
47 1503 5.711940e-03
48 1355 5.149487e-03
49 1277 4.853059e-03
50 1169 4.442620e-03
51 1083 4.115789e-03
52  991 3.766156e-03
53  955 3.629343e-03
54  887 3.370919e-03
55  802 3.047888e-03
56  754 2.865471e-03
57  765 2.907275e-03
58  693 2.633649e-03
59  607 2.306818e-03
60  608 2.310619e-03
61  541 2.055994e-03
62  527 2.002789e-03
63  455 1.729164e-03
```

```
64  432 1.641755e-03
65  345 1.311124e-03
66  350 1.330126e-03
67  279 1.060300e-03
68  279 1.060300e-03
69  224 8.512805e-04
70  197 7.486708e-04
71  162 6.156582e-04
72  136 5.168489e-04
73  137 5.206493e-04
74  119 4.522428e-04
75  108 4.104388e-04
76   87 3.306313e-04
77   75 2.850270e-04
78   65 2.470234e-04
79   63 2.394226e-04
80   55 2.090198e-04
81   63 2.394226e-04
82   52 1.976187e-04
83   36 1.368129e-04
84   28 1.064101e-04
85   31 1.178111e-04
86   27 1.026097e-04
87   22 8.360791e-05
88   26 9.880935e-05
89   24 9.120863e-05
90   17 6.460611e-05
```

```
s5_2017_2022 %>% skim()
```

Table 38: Data summary

Name	Piped data
Number of rows	263133
Number of columns	8
_____	
Column type frequency:	
character	6
numeric	1
POSIXct	1
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
sex	0	1.00	2	3	0	11	0
district	0	1.00	6	16	0	25	0
commune	0	1.00	8	23	0	181	0
hospital	0	1.00	4	56	0	108	0
icd	197058	0.25	1	2	0	15	0
in_out_patient	46	1.00	2	170	0	174	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	22.36	14.92	0	11	20	31	90	

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
date	0	1	2017-01-01	2022-12-31	2019-10-26	2191

Fix dates

Date data is in `datetime`, convert to `date` only

```
s6_2017_2022 <- s5_2017_2022 %>%
  mutate(date = convert_to_date(date))

s6_2017_2022 %>% skim()
```

Table 42: Data summary

Name	Piped data
Number of rows	263133
Number of columns	8

Column type frequency:	
character	6
Date	1
numeric	1
———	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
sex	0	1.00	2	3	0	11	0
district	0	1.00	6	16	0	25	0
commune	0	1.00	8	23	0	181	0
hospital	0	1.00	4	56	0	108	0
icd	197058	0.25	1	2	0	15	0
in_out_patient	46	1.00	2	170	0	174	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
date	0	1	2017-01-01	2022-12-31	2019-10-26	2191

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	22.36	14.92	0	11	20	31	90	

Fix ICD

```
s6_2017_2022 %>%
  mutate(year = year(date), icd = tolower(icd)) %>%
  tabyl(icd, year)
```

```
  icd  2017  2018  2019  2020 2021  2022
    1    42    12     0     0    0     1
```

```
    2     5     2     0     0    0     0
   2a     9     5     8     0    1     0
   2b     0     0     0     0    0     1
    a  2871  7589 12965  5555 3230 18627
    b   689   581   650   576  524  9366
    c   260   130   112   102   93  2048
   c1     0    18     0     0    0     0
   c2     0     2     0     0    0     0
    v     0     0     1     0    0     0
 <NA> 29297 36673 53135 19190 8603 50160
```

Mostly inconsistent letter casing, the classes are rather consistent

```
s7_2017_2022 <- s6_2017_2022 %>%
  mutate(icd = tolower(icd))

s7_2017_2022 %>%
  mutate(year = year(date)) %>%
  tabyl(icd, year)
```

```
  icd  2017  2018  2019  2020 2021  2022
    1    42    12     0     0    0     1
    2     5     2     0     0    0     0
   2a     9     5     8     0    1     0
   2b     0     0     0     0    0     1
    a  2871  7589 12965  5555 3230 18627
    b   689   581   650   576  524  9366
    c   260   130   112   102   93  2048
   c1     0    18     0     0    0     0
   c2     0     2     0     0    0     0
    v     0     0     1     0    0     0
 <NA> 29297 36673 53135 19190 8603 50160
```

Fix sexes

```
s7_2017_2022 %>% tabyl(sex)
```

```
 sex      n      percent
 NAM   1171 4.450221e-03
```

```
 NAm      2 7.600719e-06
  NU      2 7.600719e-06
 Nam 141773 5.387884e-01
  Nữ     62 2.356223e-04
  NỮ   1092 4.149993e-03
  Nữ 118434 4.500918e-01
 nAM      1 3.800360e-06
 nam    304 1.155309e-03
  nỮ      3 1.140108e-05
  nữ    289 1.098304e-03
```

Very easy fix, just remove diacritics, normalise letter casing and recode into english

```
s8_2017_2022 <- s7_2017_2022 %>%
  mutate(sex = stri_trans_general(sex, id = "Latin-ASCII") %>% tolower()) %>%
  mutate(sex = case_when(sex == "nam" ~ "male", sex == "nu" ~ "female"))

s8_2017_2022 %>% tabyl(sex)
```

```
    sex      n   percent
 female 119882 0.4555947
   male 143251 0.5444053
```

```
s8_2017_2022 %>% skim()
```

Table 46: Data summary

Name	Piped data
Number of rows	263133
Number of columns	8
_____	
Column type frequency:	
character	6
Date	1
numeric	1
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
sex	0	1.00	4	6	0	2	0
district	0	1.00	6	16	0	25	0
commune	0	1.00	8	23	0	181	0
hospital	0	1.00	4	56	0	108	0
icd	197058	0.25	1	2	0	10	0
in_out_patient	46	1.00	2	170	0	174	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
date	0	1	2017-01-01	2022-12-31	2019-10-26	2191

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	22.36	14.92	0	11	20	31	90	

See that there are 25 districts, instead of 24, let's see what's wrong

```
s8_2017_2022 %>% tabyl(district)
```

```
         district     n      percent
 Huyện Bình Chánh 22955 0.0872372526
    Huyện Cần Giờ  1754 0.0066658306
     Huyện Củ Chi 12295 0.0467254202
    Huyện Hóc Môn 15597 0.0592742073
     Huyện Nhà Bè  5739 0.0218102633
           Quận 1  6923 0.0263098889
          Quận 10  5559 0.0211261985
          Quận 11  6063 0.0230415797
          Quận 12 19205 0.0729859045
           Quận 2  4867 0.0184963498
           Quận 3  5392 0.0204915385
           Quận 4  4625 0.0175766628
           Quận 5  4543 0.0172650333
           Quận 6  6563 0.0249417595
           Quận 7  9305 0.0353623453
```

```
          Quận 8 12411 0.0471662619
          Quận 9  9321 0.0354231510
  Quận Bình Thạnh 14128 0.0536914792
    Quận Bình Tân 31456 0.1195441089
      Quận Gò Vấp    52 0.0001976187
      Quận Gò vấp  9854 0.0374487427
   Quận Phú Nhuận  4428 0.0168279919
     Quận Thủ Đức 12428 0.0472308680
    Quận Tân Bình 15722 0.0597492523
     Quận Tân Phú 21948 0.0834102906
```

Just letter casing issue, let's normalise that

Fix districts

```
s9_2017_2022 <- s8_2017_2022 %>% mutate(district = tolower(district))

s9_2017_2022 %>% skim()
```

Table 50: Data summary

Name	Piped data
Number of rows	263133
Number of columns	8
_____	
Column type frequency:	
character	6
Date	1
numeric	1
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
sex	0	1.00	4	6	0	2	0
district	0	1.00	6	16	0	24	0
commune	0	1.00	8	23	0	181	0
hospital	0	1.00	4	56	0	108	0

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
icd	197058	0.25	1	2	0	10	0
in_out_patient	46	1.00	2	170	0	174	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
date	0	1	2017-01-01	2022-12-31	2019-10-26	2191

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	22.36	14.92	0	11	20	31	90	

Fix in-/out-patient

Now let's do the hardest part, fixing in-patient and out-patient classification

```
s9_2017_2022 %>% tabyl(in_out_patient)
```

Nặng xin về ngày 29/08/2022 tại Hậu phẫu phòng mổ. Sốc mất máu; Rối loạn đông máu không đặc

Đang điều trị xin về n

n	percent	valid_percent
20	7.600719e-05	7.602048e-05
9	3.420324e-05	3.420922e-05
1	3.800360e-06	3.801024e-06
1	3.800360e-06	3.801024e-06
1	3.800360e-06	3.801024e-06
1	3.800360e-06	3.801024e-06
1	3.800360e-06	3.801024e-06
1	3.800360e-06	3.801024e-06
1	3.800360e-06	3.801024e-06
1	3.800360e-06	3.801024e-06
1	3.800360e-06	3.801024e-06
4	1.520144e-05	1.520410e-05
1	3.800360e-06	3.801024e-06

```
  10 3.800360e-05  3.801024e-05
   1 3.800360e-06  3.801024e-06
   1 3.800360e-06  3.801024e-06
   1 3.800360e-06  3.801024e-06
   4 1.520144e-05  1.520410e-05
   2 7.600719e-06  7.602048e-06
   1 3.800360e-06  3.801024e-06
   6 2.280216e-05  2.280614e-05
   1 3.800360e-06  3.801024e-06
   1 3.800360e-06  3.801024e-06
   1 3.800360e-06  3.801024e-06
   4 1.520144e-05  1.520410e-05
   1 3.800360e-06  3.801024e-06
   1 3.800360e-06  3.801024e-06
   1 3.800360e-06  3.801024e-06
   1 3.800360e-06  3.801024e-06
   1 3.800360e-06  3.801024e-06
   1 3.800360e-06  3.801024e-06
  47 1.786169e-04  1.786481e-04
   1 3.800360e-06  3.801024e-06
 581 2.208009e-03  2.208395e-03
   1 3.800360e-06  3.801024e-06
   1 3.800360e-06  3.801024e-06
   1 3.800360e-06  3.801024e-06
   1 3.800360e-06  3.801024e-06
   7 2.660252e-05  2.660717e-05
 129 4.902464e-04  4.903321e-04
   2 7.600719e-06  7.602048e-06
   1 3.800360e-06  3.801024e-06
   4 1.520144e-05  1.520410e-05
   7 2.660252e-05  2.660717e-05
   1 3.800360e-06  3.801024e-06
   1 3.800360e-06  3.801024e-06
   1 3.800360e-06  3.801024e-06
  37 1.406133e-04  1.406379e-04
   1 3.800360e-06  3.801024e-06
  26 9.880935e-05  9.882662e-05
  11 4.180395e-05  4.181126e-05
   1 3.800360e-06  3.801024e-06
1014 3.853565e-03  3.854238e-03
   1 3.800360e-06  3.801024e-06
   1 3.800360e-06  3.801024e-06
 318 1.208514e-03  1.208726e-03
```

```
    2 7.600719e-06  7.602048e-06
    1 3.800360e-06  3.801024e-06
   14 5.320503e-05  5.321434e-05
    3 1.140108e-05  1.140307e-05
    1 3.800360e-06  3.801024e-06
    1 3.800360e-06  3.801024e-06
    1 3.800360e-06  3.801024e-06
    1 3.800360e-06  3.801024e-06
  174 6.612626e-04  6.613782e-04
    8 3.040288e-05  3.040819e-05
  196 7.448705e-04  7.450007e-04
    7 2.660252e-05  2.660717e-05
   10 3.800360e-05  3.801024e-05
   54 2.052194e-04  2.052553e-04
    5 1.900180e-05  1.900512e-05
36794 1.398304e-01  1.398549e-01
    1 3.800360e-06  3.801024e-06
    1 3.800360e-06  3.801024e-06
    1 3.800360e-06  3.801024e-06
    1 3.800360e-06  3.801024e-06
    1 3.800360e-06  3.801024e-06
    1 3.800360e-06  3.801024e-06
    5 1.900180e-05  1.900512e-05
    1 3.800360e-06  3.801024e-06
    2 7.600719e-06  7.602048e-06
   13 4.940467e-05  4.941331e-05
    1 3.800360e-06  3.801024e-06
   35 1.330126e-04  1.330358e-04
    1 3.800360e-06  3.801024e-06
    1 3.800360e-06  3.801024e-06
    2 7.600719e-06  7.602048e-06
    3 1.140108e-05  1.140307e-05
    1 3.800360e-06  3.801024e-06
    1 3.800360e-06  3.801024e-06
   14 5.320503e-05  5.321434e-05
    3 1.140108e-05  1.140307e-05
   20 7.600719e-05  7.602048e-05
    3 1.140108e-05  1.140307e-05
    1 3.800360e-06  3.801024e-06
   68 2.584244e-04  2.584696e-04
    1 3.800360e-06  3.801024e-06
    1 3.800360e-06  3.801024e-06
    1 3.800360e-06  3.801024e-06
```

```
   1 3.800360e-06  3.801024e-06
   1 3.800360e-06  3.801024e-06
   1 3.800360e-06  3.801024e-06
  46 1.748165e-04  1.748471e-04
   6 2.280216e-05  2.280614e-05
 159 6.042572e-04  6.043628e-04
   1 3.800360e-06  3.801024e-06
   1 3.800360e-06  3.801024e-06
   1 3.800360e-06  3.801024e-06
   1 3.800360e-06  3.801024e-06
   1 3.800360e-06  3.801024e-06
   1 3.800360e-06  3.801024e-06
   3 1.140108e-05  1.140307e-05
  70 2.660252e-04  2.660717e-04
   1 3.800360e-06  3.801024e-06
   2 7.600719e-06  7.602048e-06
  10 3.800360e-05  3.801024e-05
   3 1.140108e-05  1.140307e-05
  15 5.700539e-05  5.701536e-05
   4 1.520144e-05  1.520410e-05
   2 7.600719e-06  7.602048e-06
   1 3.800360e-06  3.801024e-06
  95 3.610342e-04  3.610973e-04
   2 7.600719e-06  7.602048e-06
 275 1.045099e-03  1.045282e-03
  78 2.964280e-04  2.964799e-04
   1 3.800360e-06  3.801024e-06
   2 7.600719e-06  7.602048e-06
   1 3.800360e-06  3.801024e-06
  15 5.700539e-05  5.701536e-05
   1 3.800360e-06  3.801024e-06
   1 3.800360e-06  3.801024e-06
   3 1.140108e-05  1.140307e-05
   2 7.600719e-06  7.602048e-06
   1 3.800360e-06  3.801024e-06
  46 1.748165e-04  1.748471e-04
2390 9.082859e-03  9.084447e-03
   1 3.800360e-06  3.801024e-06
   4 1.520144e-05  1.520410e-05
   2 7.600719e-06  7.602048e-06
   1 3.800360e-06  3.801024e-06
   2 7.600719e-06  7.602048e-06
  12 4.560431e-05  4.561229e-05
```

```
    19 7.220683e-05  7.221946e-05
     5 1.900180e-05  1.900512e-05
    11 4.180395e-05  4.181126e-05
    12 4.560431e-05  4.561229e-05
    10 3.800360e-05  3.801024e-05
119151 4.528166e-01  4.528958e-01
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     2 7.600719e-06  7.602048e-06
     8 3.040288e-05  3.040819e-05
     1 3.800360e-06  3.801024e-06
100584 3.822554e-01  3.823222e-01
     1 3.800360e-06  3.801024e-06
     8 3.040288e-05  3.040819e-05
     4 1.520144e-05  1.520410e-05
     1 3.800360e-06  3.801024e-06
     5 1.900180e-05  1.900512e-05
    60 2.280216e-04  2.280614e-04
    43 1.634155e-04  1.634440e-04
     3 1.140108e-05  1.140307e-05
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     2 7.600719e-06  7.602048e-06
     2 7.600719e-06  7.602048e-06
    84 3.192302e-04  3.192860e-04
     1 3.800360e-06  3.801024e-06
    59 2.242212e-04  2.242604e-04
     3 1.140108e-05  1.140307e-05
    46 1.748165e-04            NA
```

Simplest things to do now are remove diacritics, normalise casing, normaling spacing

```
s9b_2017_2022 <- s9_2017_2022 %>%
  mutate(
    in_out_patient = stri_trans_general(in_out_patient, id = "Latin-ASCII") %>%
      tolower()
  )
```

```r
s9b_2017_2022 %>%
  tabyl(in_out_patient) %>%
  arrange(desc(n))
```

dang dieu tri xin ve ng

nang xin ve ngay 29/08/2022 tai hau phau phong mo. soc mat mau; roi loan dong mau khong dac

n	percent	valid_percent
121979	4.635641e-01	4.636451e-01
100865	3.833233e-01	3.833903e-01
36878	1.401497e-01	1.401742e-01
1110	4.218399e-03	4.219137e-03
588	2.234611e-03	2.235002e-03

```
390 1.482140e-03  1.482399e-03
331 1.257919e-03  1.258139e-03
329 1.250318e-03  1.250537e-03
129 4.902464e-04  4.903321e-04
 85 3.230306e-04  3.230870e-04
 47 1.786169e-04  1.786481e-04
 46 1.748165e-04            NA
 37 1.406133e-04  1.406379e-04
 35 1.330126e-04  1.330358e-04
 20 7.600719e-05  7.602048e-05
 15 5.700539e-05  5.701536e-05
 14 5.320503e-05  5.321434e-05
 14 5.320503e-05  5.321434e-05
 13 4.940467e-05  4.941331e-05
 12 4.560431e-05  4.561229e-05
 10 3.800360e-05  3.801024e-05
  9 3.420324e-05  3.420922e-05
  9 3.420324e-05  3.420922e-05
  8 3.040288e-05  3.040819e-05
  8 3.040288e-05  3.040819e-05
  7 2.660252e-05  2.660717e-05
  7 2.660252e-05  2.660717e-05
  7 2.660252e-05  2.660717e-05
  6 2.280216e-05  2.280614e-05
  5 1.900180e-05  1.900512e-05
  5 1.900180e-05  1.900512e-05
  4 1.520144e-05  1.520410e-05
  4 1.520144e-05  1.520410e-05
  4 1.520144e-05  1.520410e-05
  3 1.140108e-05  1.140307e-05
  3 1.140108e-05  1.140307e-05
  3 1.140108e-05  1.140307e-05
  3 1.140108e-05  1.140307e-05
  3 1.140108e-05  1.140307e-05
  3 1.140108e-05  1.140307e-05
  2 7.600719e-06  7.602048e-06
  2 7.600719e-06  7.602048e-06
  2 7.600719e-06  7.602048e-06
  2 7.600719e-06  7.602048e-06
  2 7.600719e-06  7.602048e-06
  2 7.600719e-06  7.602048e-06
  2 7.600719e-06  7.602048e-06
  2 7.600719e-06  7.602048e-06
```

```
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
1 3.800360e-06  3.801024e-06
```

```
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
     1 3.800360e-06  3.801024e-06
```

Lots of random stuff, but major are the in-, out-patient labels and some others (e.g. referred, discharged). The only way to fix this seems to be by hand with regex

```
s10_2017_2022 <- s9b_2017_2022 %>%
  mutate(
    in_out_patient = case_when(
      str_detect(in_out_patient, "(dieu tri)?.*n?g[opa]{2}[ij]{1,2}.*tr?u") ~
        "out-patient",
      str_detect(
        in_out_patient,
        "((di?eu t[tr][ij]{1})?.*no[it]{1}.*tr?[ui])|(n{1,2}hap vien)"
      ) ~
        "in-patient",
      str_detect(
        in_out_patient,
```

```
        "((ra|xuat|tron|bo).*vien)|((xin|bo|cho).*(ve|xv))|(ve nha)|(bo kham)"
      ) ~
        "discharged",
      str_detect(in_out_patient, "chuyen.*(benh|bv|vie[nb]|tuyen|khoa)") ~
        "referred",
      .default = "miscellanous"
    )
  )
```

This is as best as I can do…

```
s10_2017_2022 %>%
  tabyl(in_out_patient)
```

```
 in_out_patient      n     percent
     discharged  37485 0.142456476
     in-patient 101655 0.386325546
   miscellanous    264 0.001003295
    out-patient 123112 0.467869860
       referred    617 0.002344822
```

```
s10_2017_2022 %>% skim()
```

Table 54: Data summary

Name	Piped data
Number of rows	263133
Number of columns	8
_____	
Column type frequency:	
character	6
Date	1
numeric	1
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
sex	0	1.00	4	6	0	2	0
district	0	1.00	6	16	0	24	0
commune	0	1.00	8	23	0	181	0
hospital	0	1.00	4	56	0	108	0
icd	197058	0.25	1	2	0	10	0
in_out_patient	0	1.00	8	12	0	5	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
date	0	1	2017-01-01	2022-12-31	2019-10-26	2191

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	22.36	14.92	0	11	20	31	90	

Wrap up

Probably finished cleaning 2017-2022 data

Let's check final number of rows

```
nrow(s10_2017_2022)
```

```
[1] 263133
```

```
start_nrow2 - nrow(s10_2017_2022)
```

```
[1] 5605
```

```
(start_nrow2 - nrow(s10_2017_2022)) / start_nrow2 * 100
```

```
[1] 2.085675
```

Lost about 2% of rows

Data joining

Join the 2 data tibles

```
cleaned_incidence_dat <- s4_2000_2016 %>% bind_rows(s10_2017_2022)

cleaned_incidence_dat %>% skim()
```

Table 58: Data summary

Name	Piped data
Number of rows	388929
Number of columns	8
_____	
Column type frequency:	
character	6
Date	1
numeric	1
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
sex	2774	0.99	1	6	0	9	0
district	0	1.00	2	16	0	48	0
commune	0	1.00	2	23	0	352	0
hospital	0	1.00	4	56	0	145	0
icd	202338	0.48	1	6	0	45	0
in_out_patient	0	1.00	8	12	0	5	0

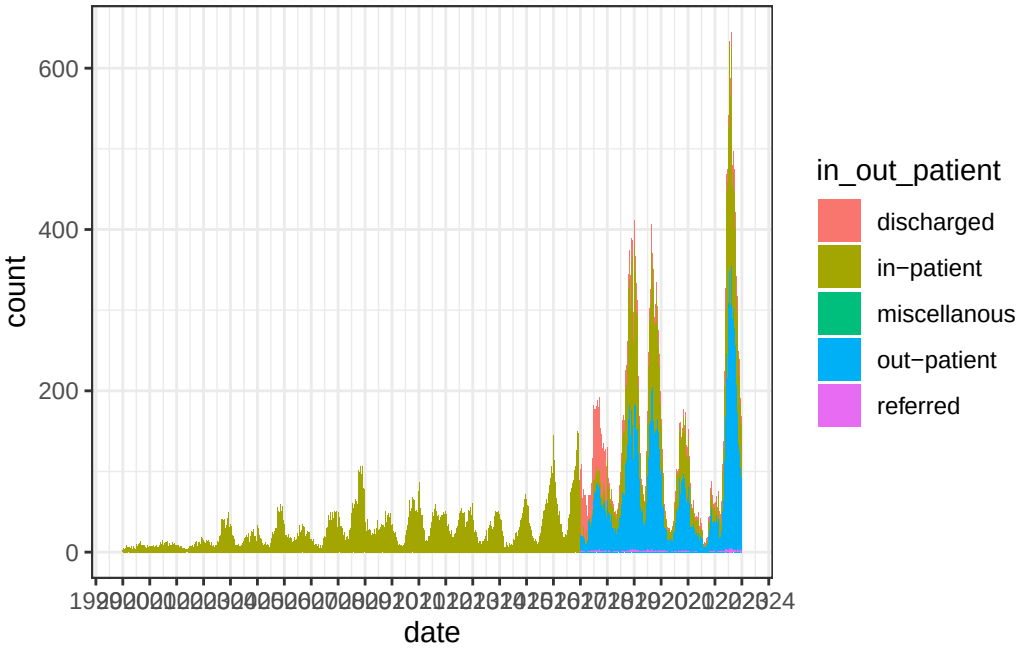
Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
date	0	1	2000-01-01	2022-12-31	2018-11-23	8319

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	20.29	14.27	0	9	17	29	90	

```
cleaned_incidence_dat %>%
  ggplot() +
  geom_bar(aes(x = date, group = in_out_patient, fill = in_out_patient)) +
  scale_x_date(date_breaks = "1 year", date_labels = "%Y")
```



Hospital names

```
cleaned_incidence_dat %>% tabyl(hospital)
```

```
                hospital     n      percent
                 AN BINH    32 8.227723e-05
             BENH VIEN 115     5 1.285582e-05
BV BENH NHIET DOI 63000 1.619833e-01
             BV BINH CHANH   167 4.293843e-04
               BV BINH TAN   822 2.113496e-03
             BV BINH THANH   259 6.659313e-04
                   BV DHYD     8 2.056931e-05
```

BV HOC MON	116	2.982550e-04
BV NHA BE	102	2.622587e-04
BV PHU NHUAN	159	4.088150e-04
BV QUAN 1	292	7.507797e-04
BV QUAN 11	333	8.561974e-04
BV QUAN 12	337	8.664821e-04
BV QUAN 3	4	1.028465e-05
BV QUAN 4	115	2.956838e-04
BV QUAN 5	13	3.342512e-05
BV QUAN 6	454	1.167308e-03
BV QUAN 7	174	4.473824e-04
BV QUAN 8	18	4.628094e-05
BV QUAN 9	21	5.399443e-05
BV QUAN THU DUC	557	1.432138e-03
BV TAN BINH	801	2.059502e-03
BV TAN PHU	1157	2.974836e-03
BVDK CU CHI	596	1.532413e-03
BVDK THU DUC	2061	5.299168e-03
Bênh viện Bệnh nhiệt đới TPHCM	72136	1.854734e-01
Bênh viện Pháp Việt	10	2.571163e-05
Bệnh Viện Columbia Gia Định	10	2.571163e-05
Bệnh Viện Hoàn Hảo	1	2.571163e-06
Bệnh Viện Đức Khang	50	1.285582e-04
Bệnh viện 175	19	4.885210e-05
Bệnh viện An Bình	281	7.224969e-04
Bệnh viện An Sinh	352	9.050495e-04
Bệnh viện Chợ Rẫy	501	1.288153e-03
Bệnh viện Gaya Việt Hàn	76	1.954084e-04
Bệnh viện Gia An 115	231	5.939387e-04
Bệnh viện Hoàn Mỹ	8035	2.065930e-02
Bệnh viện Huyện Bình Chánh	1698	4.365835e-03
Bệnh viện Huyện Cần Giờ	832	2.139208e-03
Bệnh viện Huyện Củ Chi	1135	2.918270e-03
Bệnh viện Huyện Nhà Bè	1727	4.440399e-03
Bệnh viện Hồng Đức III	1387	3.566204e-03
Bệnh viện Minh Anh	33	8.484839e-05
Bệnh viện Mỹ Đức	545	1.401284e-03
Bệnh viện Mỹ Đức Phú Nhuận	3	7.713490e-06
Bệnh viện Nguyễn Tri Phương	576	1.480990e-03
Bệnh viện Nguyễn Trãi	691	1.776674e-03
Bệnh viện Nhi Đồng 1	20	5.142327e-05
Bệnh viện Nhi đồng 1	21077	5.419241e-02
Bệnh viện Nhi đồng 2	21521	5.533401e-02

Bệnh viện Nhi đồng thành phố	5416	1.392542e-02
Bệnh viện Nhân Dân 115	1121	2.882274e-03
Bệnh viện Nhân Dân Gia Định	471	1.211018e-03
Bệnh viện Nhân dân Gia Định	4	1.028465e-05
Bệnh viện Pháp Việt	1855	4.769508e-03
Bệnh viện Phụ sản MêKông	26	6.685025e-05
Bệnh viện Phục hồi chức năng - Điều trị Bệnh nghề nghiệp	18	4.628094e-05
Bệnh viện Quân dân Y Miền Đông	1440	3.702475e-03
Bệnh viện Quận 1	3547	9.119916e-03
Bệnh viện Quận 10	419	1.077317e-03
Bệnh viện Quận 11	3623	9.315325e-03
Bệnh viện Quận 12	10259	2.637757e-02
Bệnh viện Quận 2	1582	4.067580e-03
Bệnh viện Quận 3	606	1.558125e-03
Bệnh viện Quận 4	1495	3.843889e-03
Bệnh viện Quận 5	474	1.218731e-03
Bệnh viện Quận 6	1202	3.090538e-03
Bệnh viện Quận 7	1402	3.604771e-03
Bệnh viện Quận 8	1089	2.799997e-03
Bệnh viện Quận 9	904	2.324332e-03
Bệnh viện Quận Bình Thạnh	10473	2.692779e-02
Bệnh viện Quận Bình Tân	8313	2.137408e-02
Bệnh viện Quận Gò Vấp	1582	4.067580e-03
Bệnh viện Quận Phú Nhuận	1036	2.663725e-03
Bệnh viện Quận Thủ Đức	3276	8.423131e-03
Bệnh viện Quận Tân Bình	5607	1.441651e-02
Bệnh viện Quận Tân Phú	26162	6.726678e-02
Bệnh viện Quốc Tế Mỹ	257	6.607890e-04
Bệnh viện Quốc tế City	488	1.254728e-03
Bệnh viện Quốc Ánh	808	2.077500e-03
Bệnh viện Thống Nhất	1530	3.933880e-03
Bệnh viện Triều An	513	1.319007e-03
Bệnh viện Trưng Vương	4381	1.126427e-02
Bệnh viện Tâm Trí Sài Gòn	665	1.709824e-03
Bệnh viện Tân Hưng	472	1.213589e-03
Bệnh viện Từ Dũ	31	7.970606e-05
Bệnh viện Vinmec	190	4.885210e-04
Bệnh viện Vạn Hạnh	38	9.770421e-05
Bệnh viện Xuyên Á	3796	9.760136e-03
Bệnh viện huyện Bình Chánh	11	2.828280e-05
Bệnh viện huyện Củ Chi	5	1.285582e-05
Bệnh viện quận Bình Tân	37	9.513304e-05
Bệnh viện quận Tân Bình	14	3.599629e-05

Bệnh viện ĐKKV Củ Chi	7483	1.924002e-02
Bệnh viện ĐKKV Hóc Môn	6633	1.705453e-02
Bệnh viện ĐKKV Thủ Đức	4869	1.251899e-02
Bệnh viện ĐKQT Nam Sài Gòn	15	3.856745e-05
Bệnh viện Đa khoa Bưu Điện-CS1	540	1.388428e-03
Bệnh viện Đa khoa Bưu Điện-CS3	1	2.571163e-06
Bệnh viện Đa khoa Hòan Hảo (Cs2)	2	5.142327e-06
Bệnh viện Đa khoa Quốc tế Nam Sài Gòn	1	2.571163e-06
Bệnh viện Đa khoa Sài Gòn	6	1.542698e-05
Bệnh viện Đa khoa Tâm Anh	1294	3.327085e-03
Bệnh viện Đa khoa khu vực Hậu Nghĩa	1	2.571163e-06
Bệnh viện Đại học Y Dược TPHCM	699	1.797243e-03
Bệnh viện đa khoa 30/4 Tp.HCM	713	1.833239e-03
Bệnh viện đa khoa Quốc Tế Hoàn Mỹ Thủ Đức	93	2.391182e-04
Bệnh viện đa khoa Sài Gòn	478	1.229016e-03
Bệnh viện đa khoa quốc tế Vũ Anh	437	1.123598e-03
CHO RAY	11	2.828280e-05
DK 30/4	53	1.362717e-04
DK SAI GON	28	7.199257e-05
HCDC	1	2.571163e-06
NGUYEN TRAI	94	2.416894e-04
NGUYEN TRI PHUONG	110	2.828280e-04
NHAN DAN GIA DINH	2329	5.988239e-03
NHI DONG 1	31033	7.979091e-02
NHI DONG 2	19117	4.915293e-02
PHAP VIET	20	5.142327e-05
PK THANH CONG	21	5.399443e-05
Phòng khám đa khoa Trần Diệp Khanh	87	2.236912e-04
QUAN DAN MIEN DONG	24	6.170792e-05
THONG NHAT	555	1.426996e-03
TRUNG VUONG	819	2.105783e-03
Trung tâm Kiểm soát bệnh tật Tp.Hồ Chí Minh	2	5.142327e-06
Trung tâm Y tế Quận 1	1	2.571163e-06
Trung tâm Y tế Quận 3	1	2.571163e-06
Trung tâm Y tế Quận 9	112	2.879703e-04
Trung tâm Y tế Quận Bình Thạnh	5	1.285582e-05
Trung tâm Y tế Quận Bình Tân	1	2.571163e-06
Trung tâm Y tế Quận Phú Nhuận	2	5.142327e-06
Trung tâm kiểm soát bệnh tật Thành phố HCM	1	2.571163e-06
Trung tâm y tế Dự phòng TP. H.C.M	1	2.571163e-06
Trạm Y tế Phuoc Tan Hung	2	5.142327e-06
Trạm Y tế Phường 03	1	2.571163e-06
Trạm Y tế Phường An Lợi Đông	2	5.142327e-06

Trạm Y tế Phường An Phú	5	1.285582e-05
Trạm Y tế Phường Bình An	2	5.142327e-06
Trạm Y tế Phường Bình Khánh	1	2.571163e-06
Trạm Y tế Phường Bình Trưng Đông	3	7.713490e-06
Trạm Y tế Phường Cát Lái	1	2.571163e-06
Trạm Y tế xã Hiệp Hòa	1	2.571163e-06
Trạm y tế Bình Trưng Tây	3	7.713490e-06
Viện Tim TPHCM	3	7.713490e-06
bv.nd2.khth@gmail.com	24	6.170792e-05

Normalise hospital names

```
s1_cleaned_incidence_dat <- cleaned_incidence_dat %>%
  mutate(
    hospital = stri_trans_general(hospital, id = "Latin-ASCII") %>% tolower(),
    hospital = gsub("\\s+", " ", hospital) %>% str_replace("bv", "benh vien")
  )

s1_cleaned_incidence_dat %>%
  tabyl(hospital) %>%
  arrange(desc(n))
```

hospital	n	percent
benh vien benh nhiet doi tphcm	72136	1.854734e-01
benh vien benh nhiet doi	63000	1.619833e-01
nhi dong 1	31033	7.979091e-02
benh vien quan tan phu	26162	6.726678e-02
benh vien nhi dong 2	21521	5.533401e-02
benh vien nhi dong 1	21097	5.424383e-02
nhi dong 2	19117	4.915293e-02
benh vien quan 12	10596	2.724405e-02
benh vien quan binh thanh	10473	2.692779e-02
benh vien quan binh tan	8350	2.146921e-02
benh vien hoan my	8035	2.065930e-02
benh vien dkkv cu chi	7483	1.924002e-02
benh vien dkkv hoc mon	6633	1.705453e-02
benh vien quan tan binh	5621	1.445251e-02
benh vien nhi dong thanh pho	5416	1.392542e-02
benh vien dkkv thu duc	4869	1.251899e-02
benh vien trung vuong	4381	1.126427e-02
benh vien quan 11	3956	1.017152e-02
benh vien quan 1	3839	9.870696e-03

benh vien quan thu duc	3833	9.855269e-03
benh vien xuyen a	3796	9.760136e-03
nhan dan gia dinh	2329	5.988239e-03
benh viendk thu duc	2061	5.299168e-03
benh vien phap viet	1865	4.795220e-03
benh vien huyen nha be	1727	4.440399e-03
benh vien huyen binh chanh	1709	4.394118e-03
benh vien quan 6	1656	4.257847e-03
benh vien quan 4	1610	4.139573e-03
benh vien quan 2	1582	4.067580e-03
benh vien quan go vap	1582	4.067580e-03
benh vien quan 7	1576	4.052153e-03
benh vien thong nhat	1530	3.933880e-03
benh vien quan dan y mien dong	1440	3.702475e-03
benh vien hong duc iii	1387	3.566204e-03
benh vien da khoa tam anh	1294	3.327085e-03
benh vien tan phu	1157	2.974836e-03
benh vien huyen cu chi	1140	2.931126e-03
benh vien nhan dan 115	1121	2.882274e-03
benh vien quan 8	1107	2.846278e-03
benh vien quan phu nhuan	1036	2.663725e-03
benh vien quan 9	925	2.378326e-03
benh vien huyen can gio	832	2.139208e-03
benh vien binh tan	822	2.113496e-03
trung vuong	819	2.105783e-03
benh vien quoc anh	808	2.077500e-03
benh vien tan binh	801	2.059502e-03
benh vien da khoa 30/4 tp.hcm	713	1.833239e-03
benh vien dai hoc y duoc tphcm	699	1.797243e-03
benh vien nguyen trai	691	1.776674e-03
benh vien tam tri sai gon	665	1.709824e-03
benh vien quan 3	610	1.568410e-03
benh viendk cu chi	596	1.532413e-03
benh vien nguyen tri phuong	576	1.480990e-03
thong nhat	555	1.426996e-03
benh vien my duc	545	1.401284e-03
benh vien da khoa buu dien-cs1	540	1.388428e-03
benh vien trieu an	513	1.319007e-03
benh vien cho ray	501	1.288153e-03
benh vien quoc te city	488	1.254728e-03
benh vien quan 5	487	1.252157e-03
benh vien da khoa sai gon	484	1.244443e-03
benh vien nhan dan gia dinh	475	1.221303e-03

benh vien tan hung	472	1.213589e-03
benh vien da khoa quoc te vu anh	437	1.123598e-03
benh vien quan 10	419	1.077317e-03
benh vien an sinh	352	9.050495e-04
benh vien an binh	281	7.224969e-04
benh vien binh thanh	259	6.659313e-04
benh vien quoc te my	257	6.607890e-04
benh vien gia an 115	231	5.939387e-04
benh vien vinmec	190	4.885210e-04
benh vien binh chanh	167	4.293843e-04
benh vien phu nhuan	159	4.088150e-04
benh vien hoc mon	116	2.982550e-04
trung tam y te quan 9	112	2.879703e-04
nguyen tri phuong	110	2.828280e-04
benh vien nha be	102	2.622587e-04
nguyen trai	94	2.416894e-04
benh vien da khoa quoc te hoan my thu duc	93	2.391182e-04
phong kham da khoa tran diep khanh	87	2.236912e-04
benh vien gaya viet han	76	1.954084e-04
dk 30/4	53	1.362717e-04
benh vien duc khang	50	1.285582e-04
benh vien van hanh	38	9.770421e-05
benh vien minh anh	33	8.484839e-05
an binh	32	8.227723e-05
benh vien tu du	31	7.970606e-05
dk sai gon	28	7.199257e-05
benh vien phu san mekong	26	6.685025e-05
benh vien.nd2.khth@gmail.com	24	6.170792e-05
quan dan mien dong	24	6.170792e-05
pk thanh cong	21	5.399443e-05
phap viet	20	5.142327e-05
benh vien 175	19	4.885210e-05
benh vien phuc hoi chuc nang - dieu tri benh nghe nghiep	18	4.628094e-05
benh vien dkqt nam sai gon	15	3.856745e-05
cho ray	11	2.828280e-05
benh vien columbia gia dinh	10	2.571163e-05
benh vien dhyd	8	2.056931e-05
benh vien 115	5	1.285582e-05
tram y te phuong an phu	5	1.285582e-05
trung tam y te quan binh thanh	5	1.285582e-05
benh vien my duc phu nhuan	3	7.713490e-06
tram y te binh trung tay	3	7.713490e-06
tram y te phuong binh trung dong	3	7.713490e-06

```
                                  vien tim tphcm     3 7.713490e-06
                       benh vien da khoa hoan hao (cs2)     2 5.142327e-06
                              tram y te phuoc tan hung     2 5.142327e-06
                          tram y te phuong an loi dong     2 5.142327e-06
                              tram y te phuong binh an     2 5.142327e-06
                  trung tam kiem soat benh tat tp.ho chi minh     2 5.142327e-06
                         trung tam y te quan phu nhuan     2 5.142327e-06
                        benh vien da khoa buu dien-cs3     1 2.571163e-06
                  benh vien da khoa khu vuc hau nghia     1 2.571163e-06
                benh vien da khoa quoc te nam sai gon     1 2.571163e-06
                               benh vien hoan hao     1 2.571163e-06
                                           hcdc     1 2.571163e-06
                                tram y te phuong 03     1 2.571163e-06
                           tram y te phuong binh khanh     1 2.571163e-06
                              tram y te phuong cat lai     1 2.571163e-06
                                 tram y te xa hiep hoa     1 2.571163e-06
               trung tam kiem soat benh tat thanh pho hcm     1 2.571163e-06
                    trung tam y te du phong tp. h.c.m     1 2.571163e-06
                                trung tam y te quan 1     1 2.571163e-06
                                trung tam y te quan 3     1 2.571163e-06
                           trung tam y te quan binh tan     1 2.571163e-06
```

Coding the names of the most busy hospitals

```
s2_cleaned_incidence_dat <- s1_cleaned_incidence_dat %>%
  mutate(
    hospital = case_when(
      str_detect(hospital, "benh vien benh nhiet doi") ~ "HTD",
      str_detect(hospital, "nhi dong 1") ~ "CH1",
      str_detect(hospital, "(nhi dong 2)|nd2") ~ "CH2",
      str_detect(hospital, "nhi dong thanh pho") ~ "CHC",
      str_detect(hospital, "tan phu") ~ "TPH",
      .default = hospital
    )
  )

s2_cleaned_incidence_dat %>%
  tabyl(hospital) %>%
  arrange(desc(n))
```

```
 hospital      n      percent
      HTD 135136 3.474567e-01
```

CH1	52130	1.340347e-01
CH2	40662	1.045486e-01
TPH	27319	7.024161e-02
benh vien quan 12	10596	2.724405e-02
benh vien quan binh thanh	10473	2.692779e-02
benh vien quan binh tan	8350	2.146921e-02
benh vien hoan my	8035	2.065930e-02
benh vien dkkv cu chi	7483	1.924002e-02
benh vien dkkv hoc mon	6633	1.705453e-02
benh vien quan tan binh	5621	1.445251e-02
CHC	5416	1.392542e-02
benh vien dkkv thu duc	4869	1.251899e-02
benh vien trung vuong	4381	1.126427e-02
benh vien quan 11	3956	1.017152e-02
benh vien quan 1	3839	9.870696e-03
benh vien quan thu duc	3833	9.855269e-03
benh vien xuyen a	3796	9.760136e-03
nhan dan gia dinh	2329	5.988239e-03
benh viendk thu duc	2061	5.299168e-03
benh vien phap viet	1865	4.795220e-03
benh vien huyen nha be	1727	4.440399e-03
benh vien huyen binh chanh	1709	4.394118e-03
benh vien quan 6	1656	4.257847e-03
benh vien quan 4	1610	4.139573e-03
benh vien quan 2	1582	4.067580e-03
benh vien quan go vap	1582	4.067580e-03
benh vien quan 7	1576	4.052153e-03
benh vien thong nhat	1530	3.933880e-03
benh vien quan dan y mien dong	1440	3.702475e-03
benh vien hong duc iii	1387	3.566204e-03
benh vien da khoa tam anh	1294	3.327085e-03
benh vien huyen cu chi	1140	2.931126e-03
benh vien nhan dan 115	1121	2.882274e-03
benh vien quan 8	1107	2.846278e-03
benh vien quan phu nhuan	1036	2.663725e-03
benh vien quan 9	925	2.378326e-03
benh vien huyen can gio	832	2.139208e-03
benh vien binh tan	822	2.113496e-03
trung vuong	819	2.105783e-03
benh vien quoc anh	808	2.077500e-03
benh vien tan binh	801	2.059502e-03
benh vien da khoa 30/4 tp.hcm	713	1.833239e-03
benh vien dai hoc y duoc tphcm	699	1.797243e-03

benh vien nguyen trai	691	1.776674e-03
benh vien tam tri sai gon	665	1.709824e-03
benh vien quan 3	610	1.568410e-03
benh viendk cu chi	596	1.532413e-03
benh vien nguyen tri phuong	576	1.480990e-03
thong nhat	555	1.426996e-03
benh vien my duc	545	1.401284e-03
benh vien da khoa buu dien-cs1	540	1.388428e-03
benh vien trieu an	513	1.319007e-03
benh vien cho ray	501	1.288153e-03
benh vien quoc te city	488	1.254728e-03
benh vien quan 5	487	1.252157e-03
benh vien da khoa sai gon	484	1.244443e-03
benh vien nhan dan gia dinh	475	1.221303e-03
benh vien tan hung	472	1.213589e-03
benh vien da khoa quoc te vu anh	437	1.123598e-03
benh vien quan 10	419	1.077317e-03
benh vien an sinh	352	9.050495e-04
benh vien an binh	281	7.224969e-04
benh vien binh thanh	259	6.659313e-04
benh vien quoc te my	257	6.607890e-04
benh vien gia an 115	231	5.939387e-04
benh vien vinmec	190	4.885210e-04
benh vien binh chanh	167	4.293843e-04
benh vien phu nhuan	159	4.088150e-04
benh vien hoc mon	116	2.982550e-04
trung tam y te quan 9	112	2.879703e-04
nguyen tri phuong	110	2.828280e-04
benh vien nha be	102	2.622587e-04
nguyen trai	94	2.416894e-04
benh vien da khoa quoc te hoan my thu duc	93	2.391182e-04
phong kham da khoa tran diep khanh	87	2.236912e-04
benh vien gaya viet han	76	1.954084e-04
dk 30/4	53	1.362717e-04
benh vien duc khang	50	1.285582e-04
benh vien van hanh	38	9.770421e-05
benh vien minh anh	33	8.484839e-05
an binh	32	8.227723e-05
benh vien tu du	31	7.970606e-05
dk sai gon	28	7.199257e-05
benh vien phu san mekong	26	6.685025e-05
quan dan mien dong	24	6.170792e-05
pk thanh cong	21	5.399443e-05

phap viet	20	5.142327e-05
benh vien 175	19	4.885210e-05
benh vien phuc hoi chuc nang - dieu tri benh nghe nghiep	18	4.628094e-05
benh vien dkqt nam sai gon	15	3.856745e-05
cho ray	11	2.828280e-05
benh vien columbia gia dinh	10	2.571163e-05
benh vien dhyd	8	2.056931e-05
benh vien 115	5	1.285582e-05
tram y te phuong an phu	5	1.285582e-05
trung tam y te quan binh thanh	5	1.285582e-05
benh vien my duc phu nhuan	3	7.713490e-06
tram y te binh trung tay	3	7.713490e-06
tram y te phuong binh trung dong	3	7.713490e-06
vien tim tphcm	3	7.713490e-06
benh vien da khoa hoan hao (cs2)	2	5.142327e-06
tram y te phuoc tan hung	2	5.142327e-06
tram y te phuong an loi dong	2	5.142327e-06
tram y te phuong binh an	2	5.142327e-06
trung tam kiem soat benh tat tp.ho chi minh	2	5.142327e-06
trung tam y te quan phu nhuan	2	5.142327e-06
benh vien da khoa buu dien-cs3	1	2.571163e-06
benh vien da khoa khu vuc hau nghia	1	2.571163e-06
benh vien da khoa quoc te nam sai gon	1	2.571163e-06
benh vien hoan hao	1	2.571163e-06
hcdc	1	2.571163e-06
tram y te phuong 03	1	2.571163e-06
tram y te phuong binh khanh	1	2.571163e-06
tram y te phuong cat lai	1	2.571163e-06
tram y te xa hiep hoa	1	2.571163e-06
trung tam kiem soat benh tat thanh pho hcm	1	2.571163e-06
trung tam y te du phong tp. h.c.m	1	2.571163e-06
trung tam y te quan 1	1	2.571163e-06
trung tam y te quan 3	1	2.571163e-06
trung tam y te quan binh tan	1	2.571163e-06

Normalise district and commune

Let's normalise district names from the 2 different reporting systems

```
s2_cleaned_incidence_dat %>%
  tabyl(district)
```

district	n	percent
01	4479	0.011516241
02	2201	0.005659131
03	3905	0.010040393
04	3333	0.008569688
05	4435	0.011403110
06	7584	0.019499703
07	6352	0.016332030
08	12223	0.031427330
09	3205	0.008240579
10	5344	0.013740297
11	5848	0.015036163
12	5191	0.013346909
BINH CHANH	7801	0.020057645
BINH TAN	8820	0.022677661
BINH THANH	6418	0.016501727
CAN GIO	994	0.002555736
CU CHI	2030	0.005219462
GO VAP	4916	0.012639839
HOC MON	3994	0.010269227
NHA BE	2608	0.006705594
PHU NHUAN	2182	0.005610278
TAN BINH	9393	0.024150938
TAN PHU	6949	0.017867014
THU DUC	5591	0.014375374
huyện bình chánh	22955	0.059021055
huyện cần giờ	1754	0.004509821
huyện củ chi	12295	0.031612454
huyện hóc môn	15597	0.040102435
huyện nhà bè	5739	0.014755907
quận 1	6923	0.017800164
quận 10	5559	0.014293097
quận 11	6063	0.015588964
quận 12	19205	0.049379193
quận 2	4867	0.012513852
quận 3	5392	0.013863713
quận 4	4625	0.011891631
quận 5	4543	0.011680795
quận 6	6563	0.016874545
quận 7	9305	0.023924675
quận 8	12411	0.031910709
quận 9	9321	0.023965814
quận bình thạnh	14128	0.036325396

```
    quận bình tân 31456 0.080878515
      quận gò vấp  9906 0.025469944
   quận phú nhuận  4428 0.011385111
     quận thủ đức 12428 0.031954418
    quận tân bình 15722 0.040423831
     quận tân phú 21948 0.056431894
```

```
s3_cleaned_incidence_dat <- s2_cleaned_incidence_dat %>%
  mutate(
    district = stri_trans_general(district, id = "Latin-ASCII") %>% tolower()
  ) %>%
  mutate(district = str_remove(district, "huyen|quan") %>% trimws()) %>%
  mutate(district = trimws(district, which = "left", whitespace = "0"))

s3_cleaned_incidence_dat %>%
  tabyl(district)
```

```
   district     n     percent
          1 11402 0.029316405
         10 10903 0.028033394
         11 11911 0.030625127
         12 24396 0.062726102
          2  7068 0.018172983
          3  9297 0.023904106
          4  7958 0.020461318
          5  8978 0.023083905
          6 14147 0.036374248
          7 15657 0.040256705
          8 24634 0.063338039
          9 12526 0.032206392
 binh chanh 30756 0.079078701
   binh tan 40276 0.103556176
 binh thanh 20546 0.052827123
    can gio  2748 0.007065557
     cu chi 14325 0.036831915
     go vap 14822 0.038109784
    hoc mon 19591 0.050371662
     nha be  8347 0.021461501
  phu nhuan  6610 0.016995390
   tan binh 25115 0.064574768
    tan phu 28897 0.074298908
    thu duc 18019 0.046329793
```

Now normalise commune names

```
s3_cleaned_incidence_dat %>%
  tabyl(district, commune) %>%
  View()
```

```
s3_cleaned_incidence_dat %>%
  filter(between(date, as.Date("2017-01-01"), as.Date("2020-01-01"))) %>%
  tabyl(in_out_patient)
```

```
 in_out_patient     n      percent
     discharged 28642 0.1972956404
     in-patient 52981 0.3649507829
   miscellanous    89 0.0006130617
    out-patient 63152 0.4350120201
       referred   309 0.0021284950
```

Export to CSV

```
s3_cleaned_incidence_dat %>% write_csv("incidence_full.csv")
```

Data viz

Raw time series

```
raw_ts_plot_dat <- s3_cleaned_incidence_dat %>%
  group_by(district, date) %>%
  tally() %>%
  complete(date, fill = list(n = 0)) %>%
  mutate(
    year = year(date),
    isoweek = isoweek(date),
    # isoweek = ifelse(isoweek == 53, 52, isoweek)
  ) %>%
  group_by(year, isoweek) %>%
  summarise(date = min(date), n = sum(n))
```

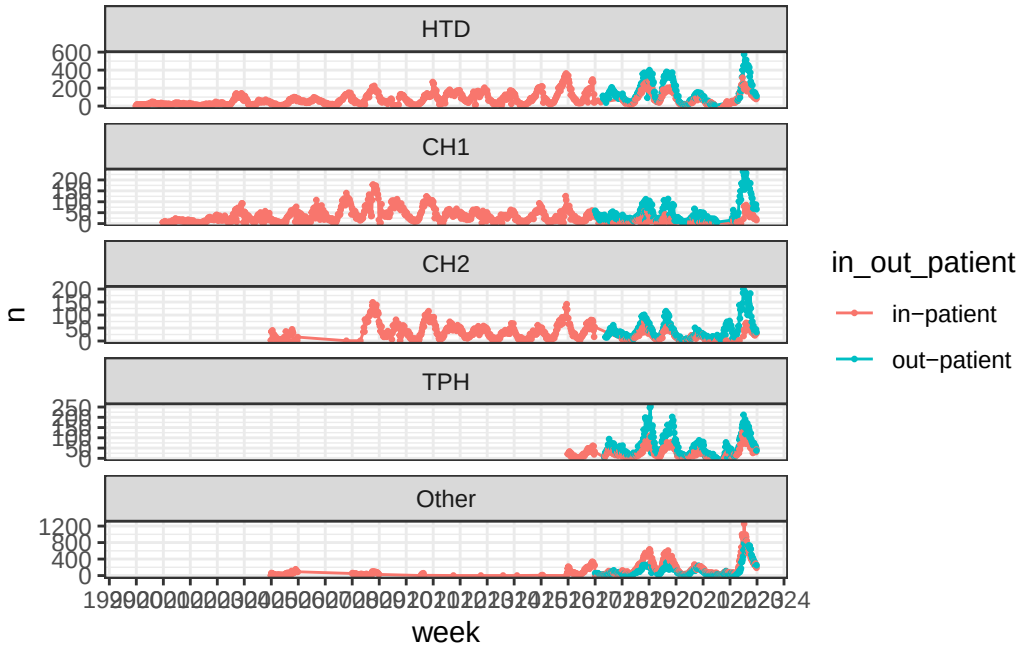
```r
raw_ts_plot_dat %>%
  ggplot(aes(x = date, y = n)) +
  geom_step() +
  facet_wrap(~district, scales = "free_y", ncol = 4) +
  scale_y_continuous("Weekly reported dengue incidence") +
  scale_x_date(
    "Date",
    breaks = seq.Date(
      as.Date("2000/01/01"),
      as.Date("2023/01/01"),
      by = "4 year"
    ),
    # date_breaks = "1 year",
    date_labels = "%Y"
  )

raw_ts_plot_dat %>%
  ggplot(aes(x = date, y = n)) +
  annotate(
    "rect",
    xmin = as.Date("2020/01/01"),
    xmax = as.Date("2022/01/01"),
    ymin = 0,
    ymax = max(raw_ts_plot_dat$n),
    fill = "red",
    alpha = 0.5
  ) +
  annotate(
    "label",
    x = as.Date("2018/08/15"),
    y = 3400,
    label = "COVID-19\nlockdowns",
    color = "#ff4946"
  ) +
  geom_step() +
  annotate(
    "segment",
    x = as.Date("2017/01/01"),
    y = 0,
    yend = max(raw_ts_plot_dat$n),
    linetype = 2,
    color = "blue"
```

```
  ) +
  annotate(
    "label",
    x = as.Date("2015/01/01"),
    y = 3400,
    label = "Circular\n54/2015/TT-BYT",
    color = "blue"
  ) +
  scale_y_continuous("Weekly reported dengue incidence") +
  scale_x_date(
    "Date",
    breaks = seq.Date(
      as.Date("2000/01/01"),
      as.Date("2023/01/01"),
      by = "1 year"
    ),
    # date_breaks = "1 year",
    date_labels = "%Y"
  )

ggsave("plots/raw_ts_plot.jpg", width = 10)
```

Total number of cases per hospital

```
hospital_order <- s3_cleaned_incidence_dat %>%
  mutate(week = lubridate::floor_date(date, "week")) %>%
  group_by(hospital) %>%
  tally(name = "total_n") %>%
  arrange(desc(total_n)) %>%
  pull(hospital)


s3_cleaned_incidence_dat %>%
  filter(in_out_patient %in% c("in-patient", "out-patient")) %>%
  mutate(
    week = lubridate::floor_date(date, "week"),
    # hospitals take make up less the 5% of total cases will be put in "Other"
    hospital = factor(hospital, levels = hospital_order) %>% fct_lump_prop(0.05)
  ) %>%
  group_by(week, in_out_patient, hospital) %>%
```

```
  tally() %>%
  ggplot(aes(x = week, y = n, color = in_out_patient)) +
  geom_line() +
  geom_point(size = 0.5) +
  scale_x_date(
    date_breaks = "1 year",
    date_labels = "%Y",
    minor_breaks = NULL
  ) +
  #   facet_wrap(~hospital, ncol = 1)
  facet_wrap(~hospital, ncol = 1, scales = "free_y")
```




```
#   theme(legend.position = "none")
```

We can see some weird period where there are so little in-patient data from big hospitals like in 2017.

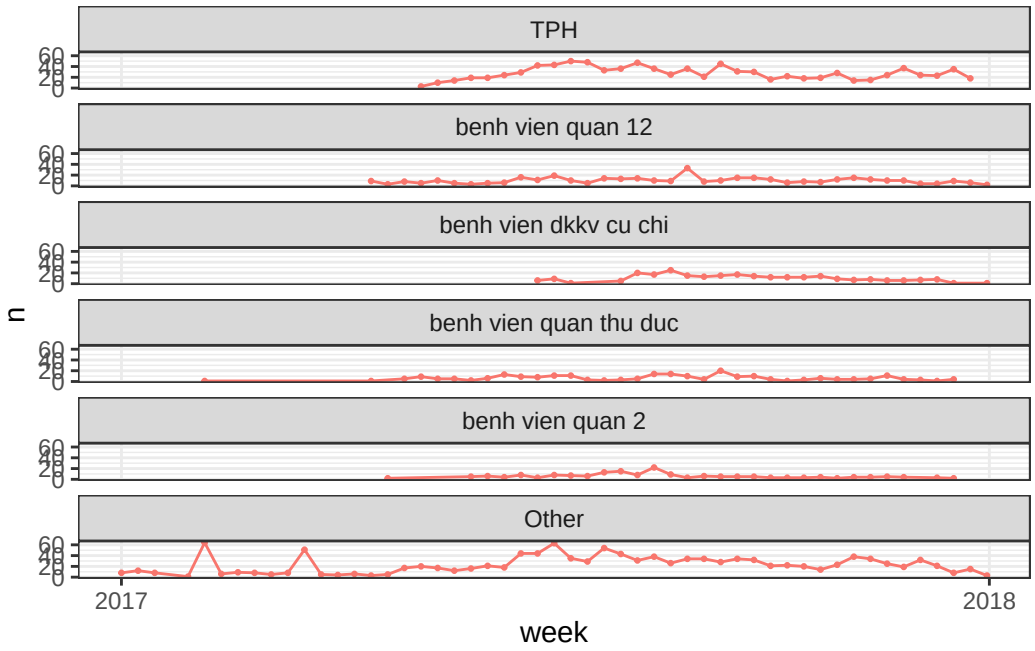
Let's zoom into that

```
s3_cleaned_incidence_dat %>%
  filter(year(date) == 2017, in_out_patient == "in-patient") %>%
  tabyl(hospital) %>%
  arrange(desc(n))
```

hospital	n	percent
TPH	934	0.2953826692
benh vien quan 12	373	0.1179633144
benh vien dkkv cu chi	260	0.0822264390
benh vien quan thu duc	230	0.0727387729
benh vien quan 2	177	0.0559772296
benh vien quan dan y mien dong	139	0.0439595193
HTD	102	0.0322580645
benh vien quoc anh	81	0.0256166983
benh vien tam tri sai gon	70	0.0221378874
benh vien quan 9	69	0.0218216319
benh vien quan 5	67	0.0211891208
benh vien quan 1	65	0.0205566097
benh vien quan 6	63	0.0199240987
trung tam y te quan 9	52	0.0164452878
benh vien trieu an	47	0.0148640101
benh vien an binh	46	0.0145477546
benh vien da khoa 30/4 tp.hcm	41	0.0129664769
benh vien dkkv hoc mon	37	0.0117014548
benh vien huyen binh chanh	27	0.0085388994
benh vien quan 4	24	0.0075901328
benh vien quan 8	23	0.0072738773
benh vien xuyen a	22	0.0069576218
benh vien da khoa quoc te vu anh	20	0.0063251107
benh vien quan 11	19	0.0060088552
benh vien da khoa buu dien-cs1	16	0.0050600886
benh vien nhan dan 115	16	0.0050600886
benh vien quan phu nhuan	15	0.0047438330
benh vien vinmec	14	0.0044275775
benh vien 175	13	0.0041113219
benh vien huyen can gio	11	0.0034788109
benh vien nguyen trai	11	0.0034788109
benh vien dkkv thu duc	10	0.0031625553
CH1	9	0.0028462998
benh vien da khoa sai gon	9	0.0028462998
benh vien huyen nha be	9	0.0028462998
benh vien quan 3	8	0.0025300443
benh vien quan go vap	7	0.0022137887
benh vien phap viet	4	0.0012650221
benh vien quan 7	4	0.0012650221
benh vien quan binh tan	4	0.0012650221
benh vien tan hung	4	0.0012650221
benh vien trung vuong	4	0.0012650221

```
benh vien quan tan binh   2 0.0006325111
   benh vien thong nhat   2 0.0006325111
      benh vien cho ray   1 0.0003162555
  trung tam y te quan 1   1 0.0003162555
```

```r
s3_cleaned_incidence_dat %>%
  filter(year(date) == 2017, in_out_patient == "in-patient") %>%
  mutate(
    week = lubridate::floor_date(date, "week"),
    hospital = factor(hospital, levels = hospital_order) %>% fct_lump_prop(0.05)
  ) %>%
  group_by(week, in_out_patient, hospital) %>%
  tally() %>%
  ggplot(aes(x = week, y = n, color = in_out_patient)) +
  geom_line() +
  geom_point(size = 0.5) +
  scale_x_date(
    date_breaks = "1 year",
    date_labels = "%Y",
    minor_breaks = NULL
  ) +
  facet_wrap(~hospital, ncol = 1) +
  #   facet_wrap(~hospital, ncol = 1, scales = "free_y")
  theme(legend.position = "none")
```



Data availability map

```
s3_cleaned_incidence_dat %>%
  mutate(week = floor_date(date, "week")) %>%
  group_by(week, in_out_patient, hospital) %>%
  tally() %>%
  ungroup() %>%
  complete(week, in_out_patient, hospital) %>%
  filter(in_out_patient %in% c("in-patient", "out-patient")) %>%
  group_by(week, in_out_patient, hospital) %>%
  mutate(total_n = sum(n, na.rm = TRUE)) %>%
  ungroup() %>%
  mutate(hospital = fct_lump_prop(hospital, prop = 0.01, w = total_n)) %>%
  ggplot(aes(x = week, y = fct_reorder(hospital, total_n), fill = n)) +
  geom_raster() +
  facet_wrap(~in_out_patient, ncol = 1) +
  theme(legend.position = "none") +
  scale_x_date(
    date_breaks = "1 year",
    date_labels = "%Y",
    minor_breaks = NULL
  ) +
  scale_fill_viridis_c(na.value = "transparent")
```