BCM Datathon Team 17 Data Cleaning

Andrew Zimolzak, MD, MMSc

April, 2022

Contents

Patient visits (descriptive stats)

Encounter class

```
table(ed_encounter_tidy$Patient.Class) %>% kable()
```

Var1	Freq
Emergency	6983
Inpatient	6026
Observation	1103
Outpatient	22

ED disposition

```
table(ed_encounter_tidy$ED.Disposition) %>% kable()
```

Var1	Freq
Admit	5231
AMA	75
Discharge	5843
Eloped	133
Expired	8
Left After Medical Screening Exam	38
LWBS after Triage	96
LWBS before Triage	24
Observation	1816
Registration Error	2
Send to L&D	7
Transfer to Another Facility	858

Var1	Freq
Unknown	3

Skip this section (Andy's notes)

My observations:

- disposition=Discharge and class=Emergency: common but not perfectly?
- d=Admit c=Inpatient: common
- d=Observation c=Observation: somewhat common
- d=Observation c=Inpatient: even more common (surprisingly)
- d=Admit class=Observation: common
- c=Outpatient: uncommon in general
- disposition=Admit class=Emergency: rare

The main dispositions that seem to matter:

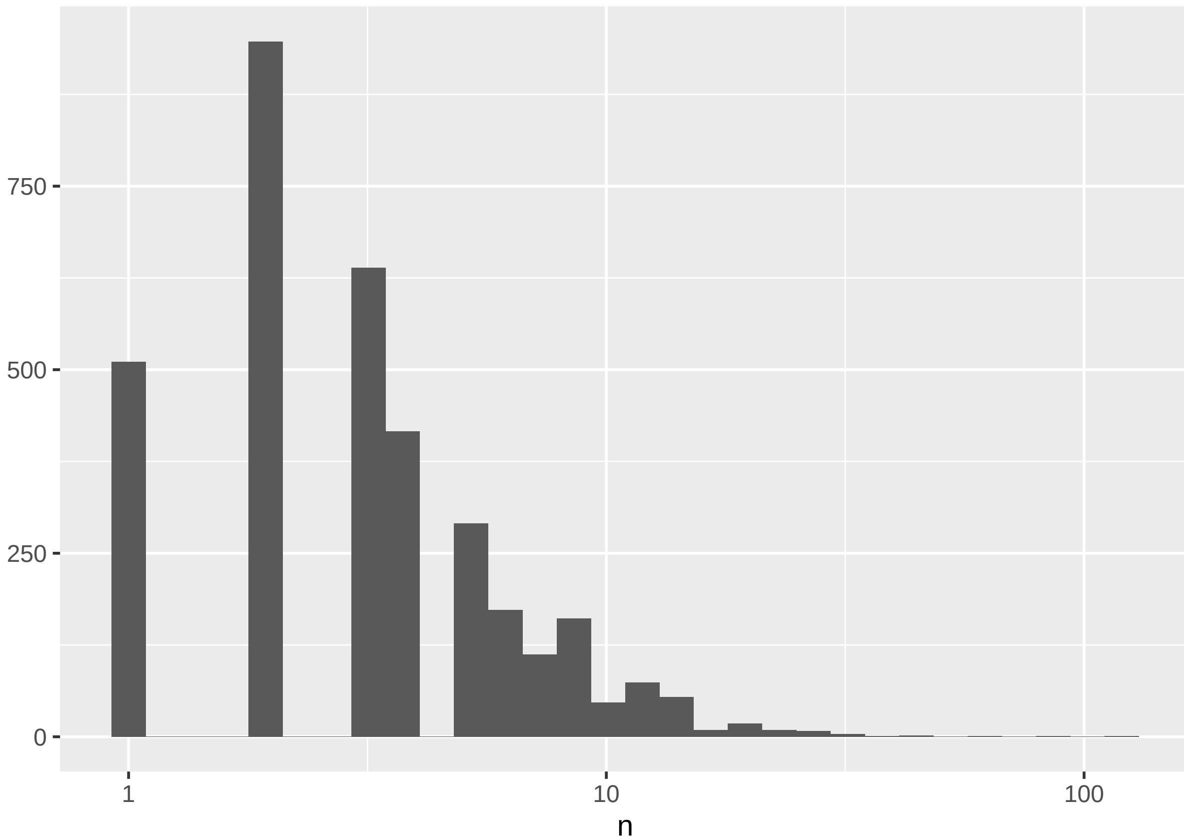
- Discharge
- Admit
- Observation
- Transfer

We could make a more limited table of 4 dispo X 3 class. This is all possibly because we requested ER discharge to home and not admit?

Looks like if dispo = discharge, then ER departure time = Hospital discharge time.

Distribution of encouters per patient ID

```
ed_encounter_tidy %>%  count(PAT_ID) -> pat_id_counts
qplot(x=n, data=pat_id_counts) +  scale_x_log10()
```



Skip this section (Andy's notes)

Why do some have only 1 encounter?

Inspect singles by themselves.

2022-04-13 meeting notes.

ED admiss and ED discharge is literally when they came/went from ED. Whereas, inpatient admit/dis date and time are about when they *first showed up to hospital* no matter what part of hospital. Many of the obs people will expect to have inp discharge time which is *after* the ED disch time. But not *all.* Occasional data error or whatever.

Also some things come thru diff in Slicer vs Caboodle.

I had noted that: A few (singletons) have an obs admit with an inpatient LOS > ER los (but not all).

Very likely valid approach: ignore the ~500 patients who are singletons (only one row in the ED Admissions table). Because so few have actual inpatient-looking encounters anyway.

`Pat Enc CSN` will be needed eventually to join up *notes* to *encounters.*

Demographics (Table 1)

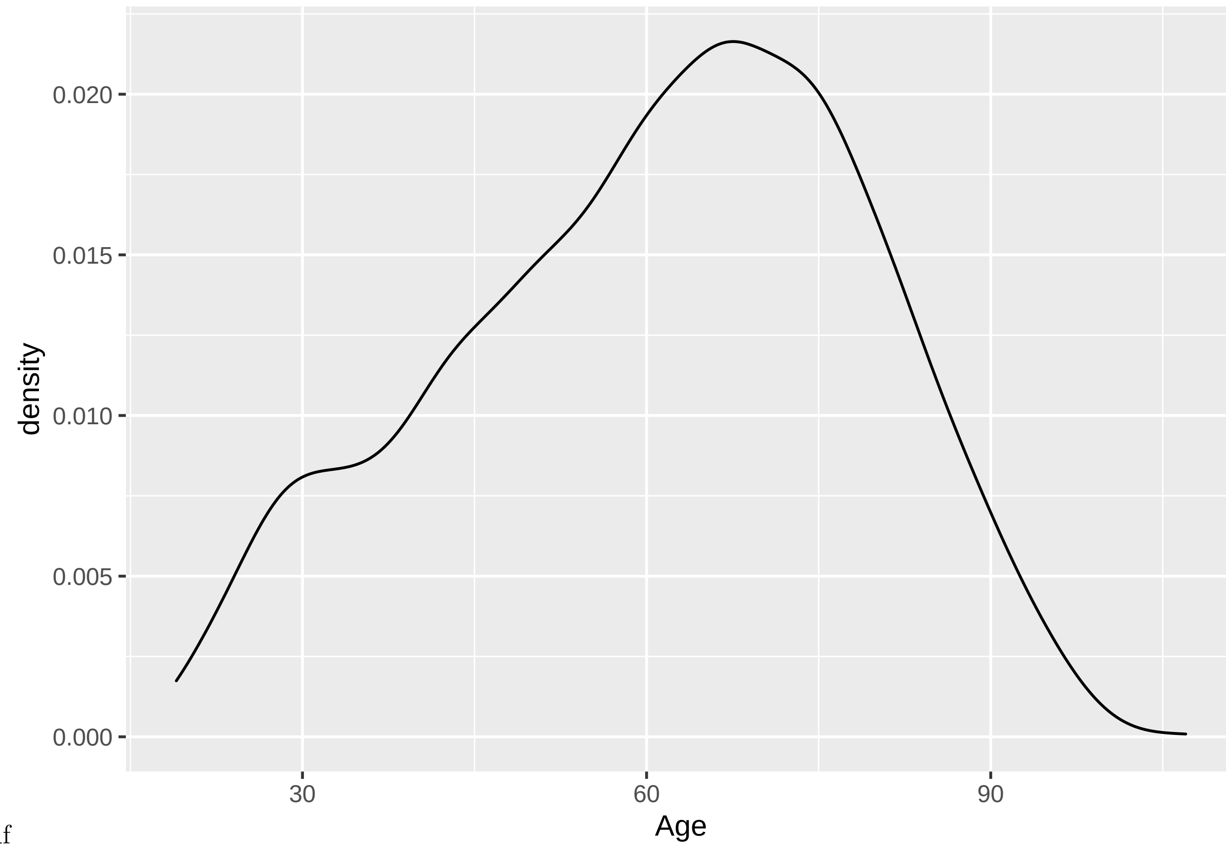
Observations in brief: Number of patients is 3494. 2:1 white:black. Most common employment is retired. Most are married. 2:1 nonmedicare vs yes. Most common financial status is: managed care, or Medicare managed care.

```
demographics_tidy_rm_dups %>% select(-pat_id) -> tabulate_me
CreateTableOne(data=tabulate_me) -> t1
kableone(t1)
```

	Overall
n	3494
demog_entries (mean (SD))	1.07 (0.28)
Sex = Male (%)	1600 (45.8)
Age (mean (SD))	61.05 (17.82)
Ethnic.Group (%)	
	1 (0.0)
Declined	3 (0.1)
Hispanic or Latino	605 (17.3)
Not Hispanic or Latino	2862 (81.9)
Unable to Determine	23 (0.7)
Race (%)	
	4 (0.1)
American Indian or Alaska Native	12 (0.3)
Asian	97 (2.8)
Black or African American	1068 (30.6)
Declined	42 (1.2)
Native Hawaiian or Other Pacific Islander	12 (0.3)
Other	53 (1.5)
Unable to Determine	81 (2.3)
White or Caucasian	2125 (60.8)
Employment.Status (%)	
Disabled	398 (11.4)
Full Time	752 (21.5)
Not Employed	817 (23.4)
Not listed	17 (0.5)
Part Time	59 (1.7)
Retired	1280 (36.6)
Self Employed	113 (3.2)
Student - Full time	25 (0.7)
Unknown	33 (0.9)
interpreter (%)	
	1 (0.0)
N	3353 (96.0)
Y	140 (4.0)
Language (%)	
Arabic	2 (0.1)
Chinese (Mandarin)	2 (0.1)
Chinese, Cantonese (Inc Toishanese)	1 (0.0)
English	3301 (94.5)
Farsi, Persian	2 (0.1)
Laotian	1 (0.0)
Other	12 (0.3)
Russian	4 (0.1)
Spanish	155 (4.4)
Unknown	2 (0.1)
Vietnamese	12 (0.3)
Marital.Status (%)	
Divorced	283 (8.1)

	Overall
Legally Separated	34 (1.0)
Life Partner	6 (0.2)
Married	1659 (47.5)
No Answer	1 (0.0)
Significant Other	1 (0.0)
Single	1079 (30.9)
Unknown	41 (1.2)
Widow/Widower	390 (11.2)
mcaid = TRUE (%)	270 (7.7)
mcare = TRUE (%)	1247 (35.7)
Financial.Class (%)	
Champus/Tricare	17 (0.5)
Commercial	66 (1.9)
Institutional	9 (0.3)
International	3 (0.1)
Managed Care	963 (27.6)
Medicaid	33 (0.9)
Medicaid Mgd Care	326 (9.3)
Medicare	632 (18.1)
Medicare Mgd Care	1100 (31.5)
Other	1 (0.0)
Pending Charity	3 (0.1)
Pending Eligibility	5 (0.1)
Self-Pay	332 (9.5)
Special Handling	2 (0.1)
Workers Comp	2 (0.1)

```
ggplot(demographics_tidy_rm_dups, aes(Age)) +   geom_density()
```



density-1.pdf

Skip this section (Andy's notes)

Var1	Freq
1	3278
2	206
3	2
4	8

Table. People with multiple demographic entries. Seems it is often for multiple financial classes. Sometimes for multiple races, though.

Clinical Notes (descriptive stats)

```
notes %>%
  group_by(PAT_ID) %>%
  summarise(note_count = n()) -> counted_notes
dim(counted_notes)
```

```
## [1] 100   2
```

There are 100 patients for whom we have notes.

Note chunks per patient

```
ggplot(counted_notes, aes(note_count)) + geom_histogram() + scale_x_log10()
```

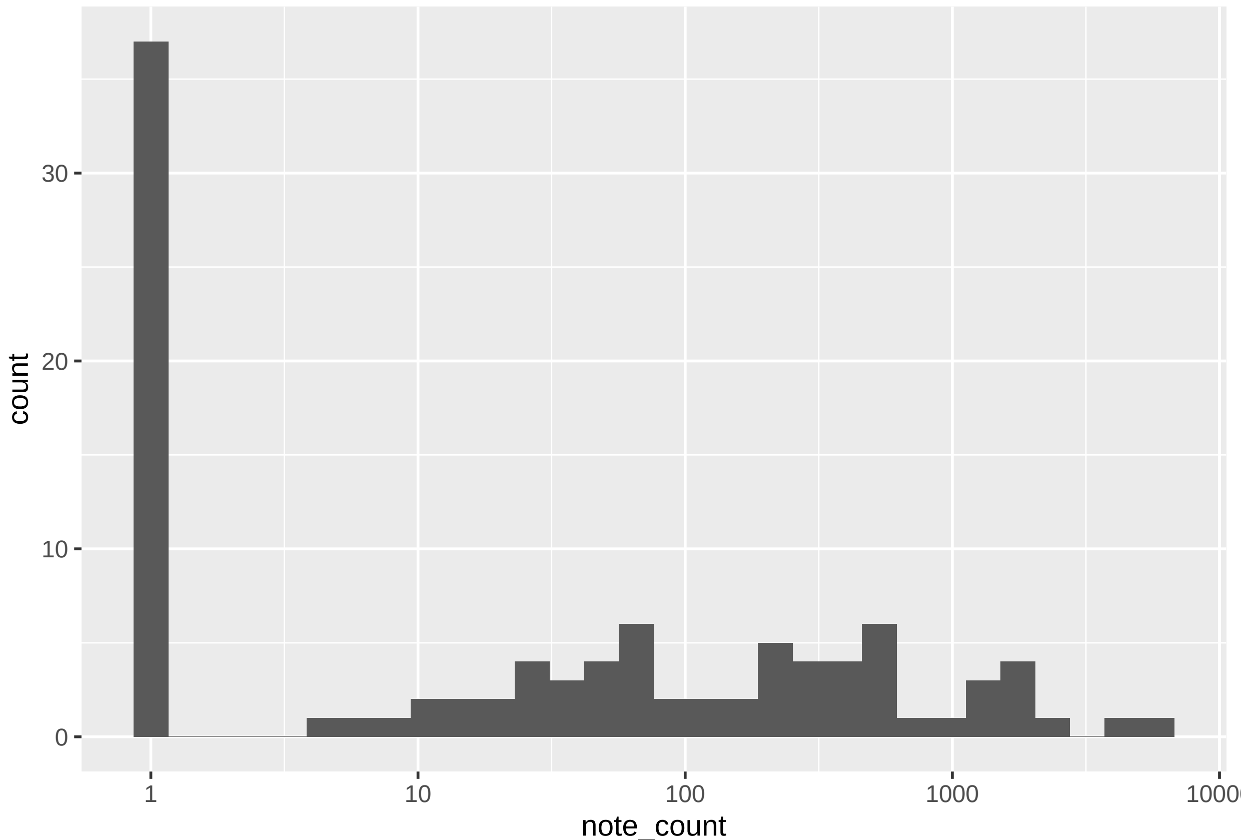


Figure. Notes per patient. About 35 patients have only one note chunk.

Zoomed in note chunks per patient

```
counted_notes %>%
  filter(note_count > 1) -> multi_notes
ggplot(multi_notes, aes(note_count)) + geom_histogram() + scale_x_log10()
```

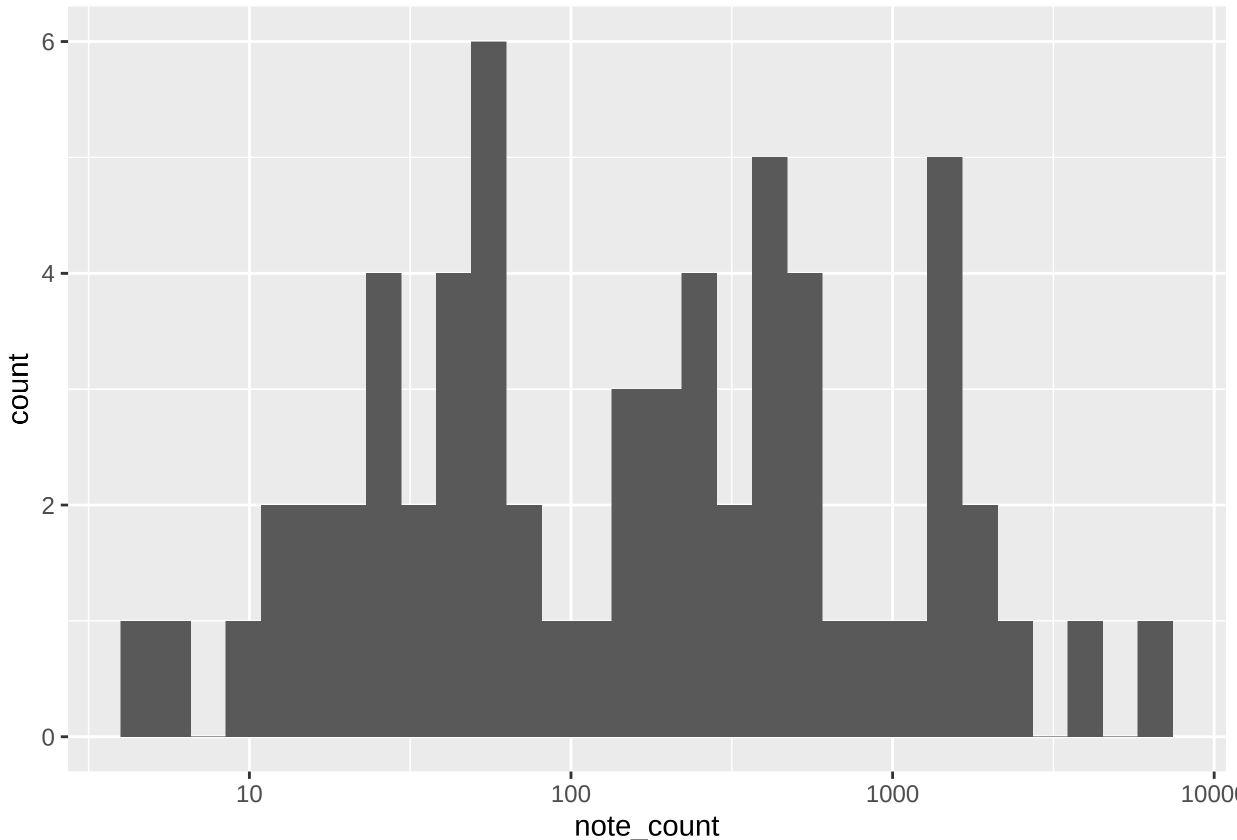


Figure. Zoomed in note chunks per patient. Many people have dozens of note chunks. Plenty have hundreds. A few have thousands(!)

Actual text processing

Review this: https://pubmed.ncbi.nlm.nih.gov/35044842/ . Sun M, Oliwa T, Peek ME, Tung EL. Negative Patient Descriptors: Documenting Racial Bias In The Electronic Health Record. *Health Aff (Millwood).* 2022;41(2):203-211. doi:10.1377/hlthaff.2021.01423

> Fifteen descriptors were selected for inclusion in the analysis: (non-)adherent, aggressive, agitated, angry, challenging, combative, (non-)compliant, confront, (non-)cooperative, defensive, exaggerate, hysterical, (un-)pleasant, refuse, and resist. We adjusted the descriptors to permit identification of alternative grammatical forms (for example, "adher" for "adherent," "adhere," or "adhered)
>
> From all sentences in the data set, we selected a random sample of sentences containing one or more of the fifteen selected patient descriptors for manual review .... We categorized the use of each descriptor in one of three possible ways: negative, positive, or out of context.
>
> A total of 6,818 sentences were classified.

```
corpus <- VCorpus(VectorSource(notes$NOTE_TEXT))  # fixme - consider SimpleCorpus?
corpus <- tm_map(corpus, stripWhitespace)
corpus <- tm_map(corpus, content_transformer(tolower))

# lapply(corpus[2], as.character)
# corpus <- tm_map(corpus, stemDocument)  # not convinced of fidelity
```

```r
dtm <- DocumentTermMatrix(corpus, list(dictionary = descrip))
# `descrip` comes from functions.R.
```

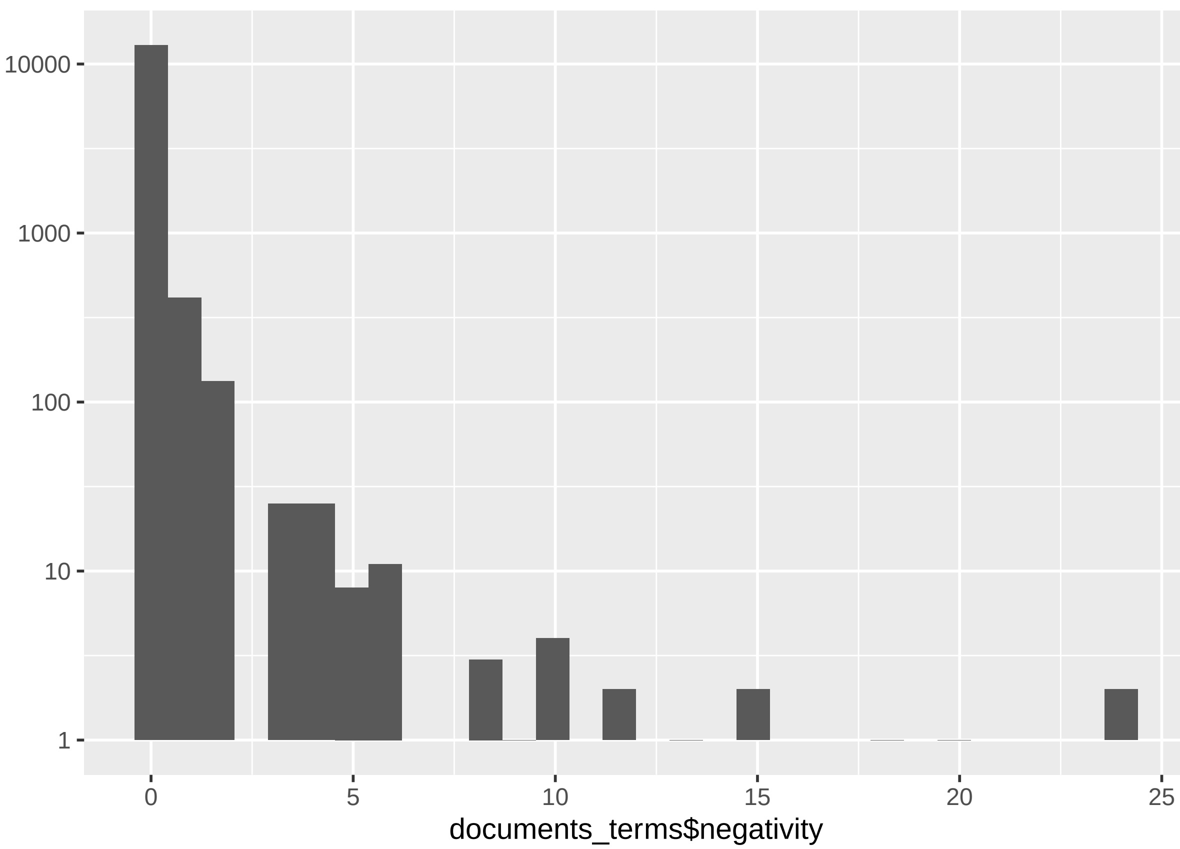
```r
notes %>% select(PAT_ID, NOTE_ID, LINE) -> notes_metadata
cbind(notes_metadata, as.matrix(dtm)) %>%
  arrange(PAT_ID, NOTE_ID, LINE) -> lines_terms

lines_terms %>%
  group_by(PAT_ID, NOTE_ID) %>%
  summarise_all(
    # vars(-LINE, LINE),
    list(~ sum(.), ~ n())
  ) %>%
  rename(n = LINE_n, pat_id = PAT_ID) %>%
  select(- ends_with("_n"), -LINE_sum) %>%
  ungroup() %>%
  rowwise() %>%
  mutate(negativity = sum(c_across(ends_with("_sum")))) %>%
  ungroup() -> documents_terms
```

Distribution of neg. descriptors per patient

Figure. Negative descriptors per note. There are 10k notes with 0 negative descriptors, 1k with 1, 100 with 3–4, etc.

```r
qplot(documents_terms$negativity) + scale_y_log10()
```



Interesting observation. The prior method using `count_descriptors` function appears to double the negativity relative to `tm` package.

Final analytic dataset

```
## Joining, by = "pat_id"
```

```
write.csv(joined_tm, here('analytic_dataset_tm.csv'))
names(joined_tm)
```

```
##  [1] "pat_id"               "NOTE_ID"              "adhere_sum"
##  [4] "adherence_sum"        "adherent_sum"         "adheres_sum"
##  [7] "adhering_sum"         "aggressive_sum"       "agitated_sum"
## [10] "angry_sum"            "challenging_sum"      "combative_sum"
## [13] "compliance_sum"       "compliant_sum"        "complies_sum"
## [16] "comply_sum"           "complying_sum"        "confront_sum"
## [19] "confrontational_sum" "cooperate_sum"        "cooperating_sum"
## [22] "defensive_sum"        "exaggerate_sum"       "exaggerated_sum"
## [25] "exaggerates_sum"      "exaggerating_sum"     "hysterical_sum"
## [28] "non-adherent_sum"     "non-compliance_sum"   "non-compliant_sum"
## [31] "non-cooperative_sum" "nonadherent_sum"      "noncompliance_sum"
## [34] "noncompliant_sum"     "noncooperative_sum"   "refuse_sum"
## [37] "refused_sum"          "refuses_sum"          "refusing_sum"
## [40] "resist_sum"           "resisted_sum"         "resisting_sum"
## [43] "resists_sum"          "uncooperative_sum"    "unpleasant_sum"
## [46] "n"                    "negativity"           "demog_entries"
## [49] "Sex"                  "Age"                  "Ethnic.Group"
## [52] "Race"                 "Employment.Status"    "interpreter"
## [55] "Language"             "Marital.Status"       "mcaid"
## [58] "mcare"                "Financial.Class"      "negativity_any"
## [61] "race_ethn"            "negativity_binned"
```

Table. Example analytic data set. I printed the outcome variable `negativity`, but only a selection of covariates, for simplicity. There is also a binary outcome variable `negativity_any`. All covariates are retained in the output CSV files.