data analysis

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Load Data

The original data frame has 439248 observations, 247 variables.

```
dim(data)
```

```
## [1] 439248    247
```

Target Variables

mortality_90 is a 0/1 flag indicating whether the patient died within 90 days post index discharge. For the mortality, we have 380213 patients who are alive.

```
mean(data$mortality_90)
```

```
## [1] 0.1344002
```

```
length(data$mortality_90)-sum(data$mortality_90)
```

```
## [1] 380213
```

readm_flag is a 0/1 flag indicating whether the patient was readmitted within 90 days post index discharge. For the readmission, we have 342935 patients who are not readmitted.

```
mean(data$readm_flag)
```

```
## [1] 0.2192679
```

```r
length(data$readm_flag)-sum(data$readm_flag)
```

```
## [1] 342935
```

er_90days is a 0/1 flag indicating ER visit within 90 days post index discharge.
For the ER visit, we have 353680 patients who are not visited ER.

```r
mean(data$er_90days)
```

```
## [1] 0.1948057
```

```r
length(data$er_90days)-sum(data$er_90days)
```

```
## [1] 353680
```

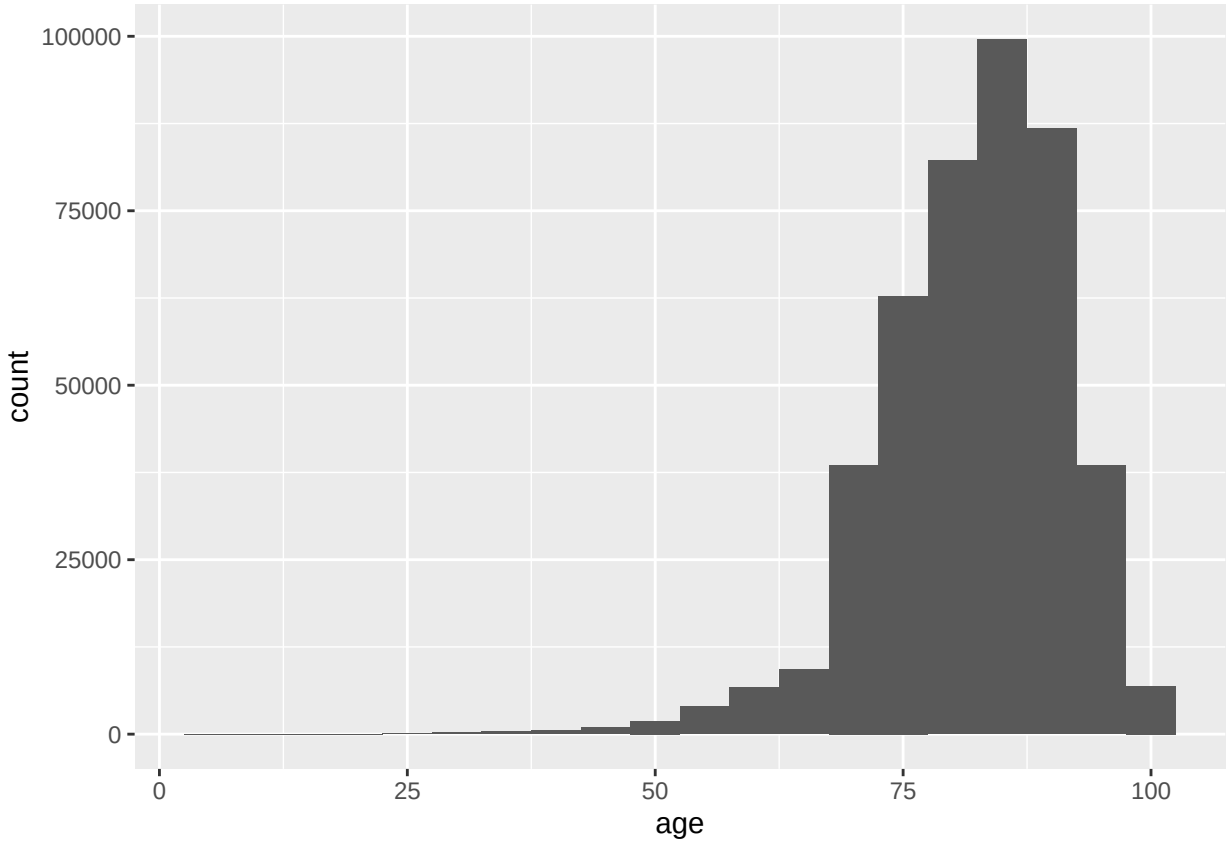
Data About Patients

Age:

```r
summary(data$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00   76.00   83.00   81.96   89.00   98.00
```

```r
ggplot(data, aes(x=age)) +
  geom_histogram(binwidth = 5)
```

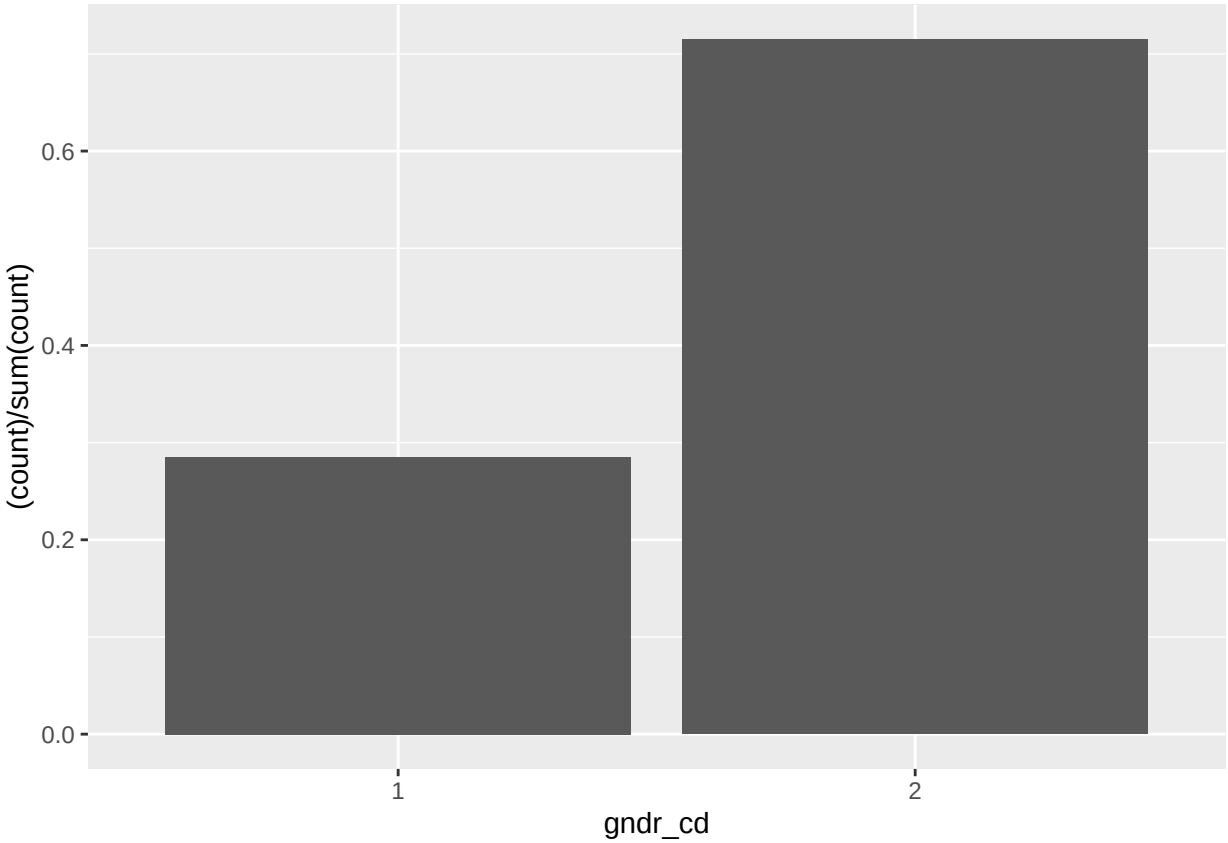


gender: There is no unknown data in gender. 1 is male, 2 is female.

```
sum(data$gndr_cd != 0)
```

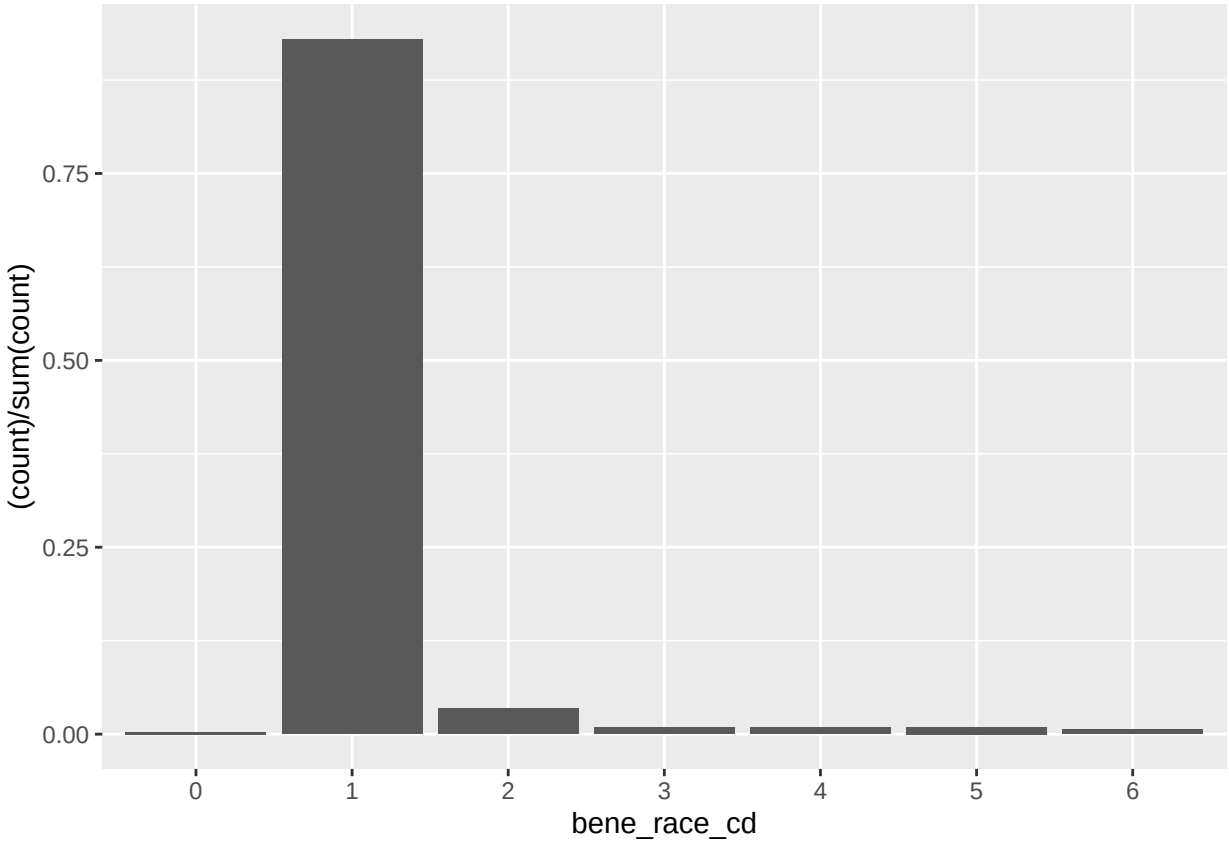
```
## [1] 439248
```

```
ggplot(data, aes(x = gndr_cd)) +
  geom_bar(aes(y = (..count..)/sum(..count..)))
```



race: 1 is white.

```
ggplot(data, aes(x = bene_race_cd)) +
  geom_bar(aes(y = (..count..)/sum(..count..)))
```



##Important Features

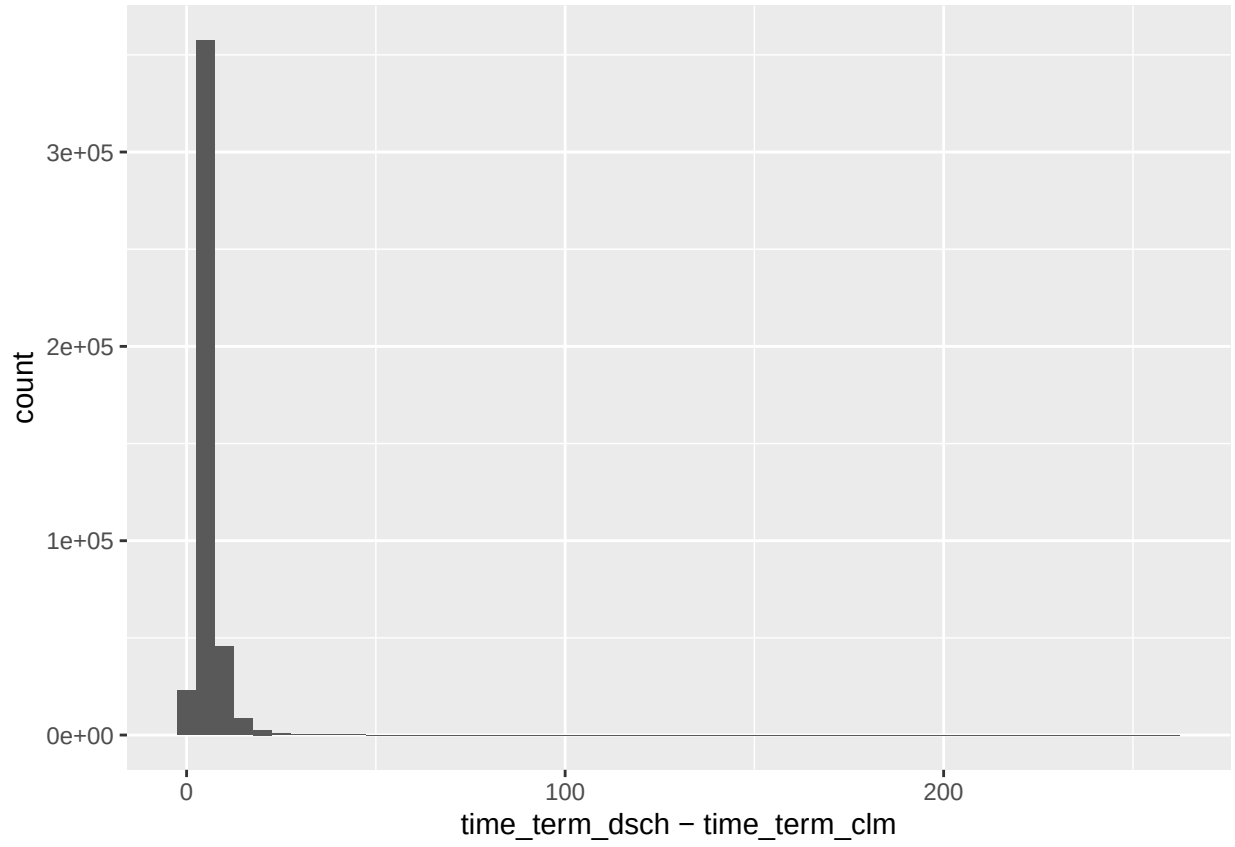
```
#creating variables
data_new <- data %>% mutate(#interval for age. Broken into 5 year intervals and under 65
                            # correlation between age and age_interval?????
                            age_interval = cut(age, breaks = c(0,65,70,75,80,85,90,95,100)),

                            #creating a simpler version of diagnosis type
                            diagnosis_type = case_when(substr(prncpal_dgns_cd,1,1) == "S" ~ "S
                                                       substr(prncpal_dgns_cd,1,1) == "M" ~ "M
                                                       TRUE ~ "Other"),

                            #creating interval variable for length of stay


                            los_interval = cut(index_los, breaks = c(0,1,2,5,10,20,40,80,160,3

                            #making time series terms from dates

                            #Index admission date
                            month_clm = as.numeric(substr(clm_admsn_dt,6,7)),

                            #Index discharge date
                            month_dsch = as.numeric(substr(nch_bene_dschrg_dt,6,7)),


                            season_clm = as.factor(case_when(month_clm <= 3 ~ "Winter",
```

```
                                                        month_clm >= 10 ~ "Fall",
                                                        month_clm > 3 & month_clm < 7 ~ "
                                                        TRUE ~ "Summer")),


                                    season_dsch = as.factor(case_when(month_dsch <= 3 ~ "Winter",
                                                                      month_dsch >= 10 ~ "Fall",
                                                                      month_dsch > 3 & month_dsch < 7
                                                                      TRUE ~ "Summer")),

                                    time_term_clm = as.numeric(substr(clm_admsn_dt,9,10)) +
                                      month_clm * 30 + 365 * (yr_adm - min(yr_adm)),

                                    time_term_dsch = as.numeric(substr(nch_bene_dschrg_dt,9,10)) +
                                      month_dsch * 30 + 365 * (yr_disch - min(yr_disch))
)
```

```
summary(data_new$time_term_dsch  - data_new$time_term_clm)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.000   4.000   5.182   6.000 260.000
```

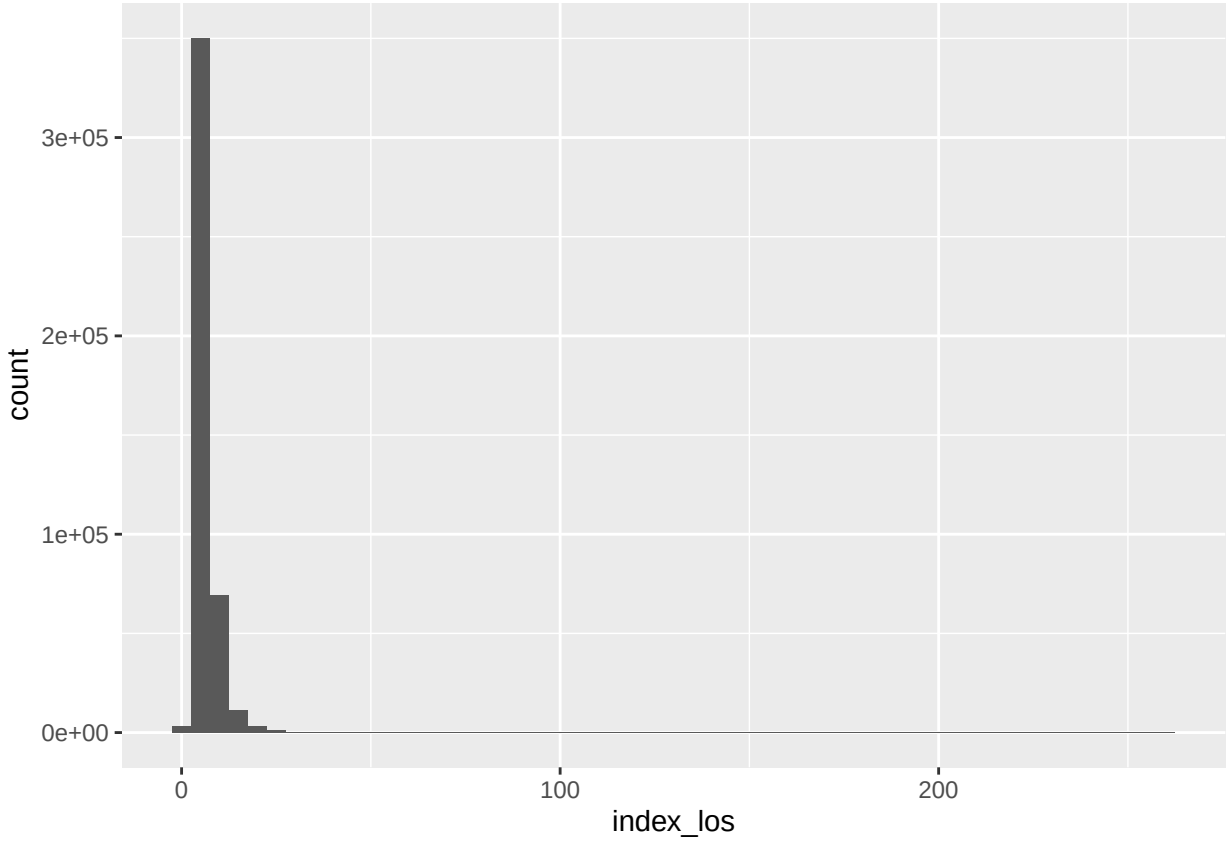
```
ggplot(data_new, aes(x=time_term_dsch  - time_term_clm)) +
  geom_histogram(binwidth = 5)
```



```r
summary(data$index_los)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   4.000   5.000   6.203   7.000 259.000
```

```r
ggplot(data, aes(x=index_los)) +
  geom_histogram(binwidth = 5)
```



clm_utlztn_day_cnt:
On an institutional claim, the number of covered days of care that are chargeable to Medicare facility utilization that includes full days, coinsurance days, and lifetime reserve days. It excludes any days classified as non-covered, leave of absence days, and the day of discharge or death.

```r
summary(data$clm_utlztn_day_cnt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.000   4.000   5.137   6.000 150.000
```

clm_pps_cptl_drg_wt_num:
DRG weight

```r
unique(data$clm_pps_cptl_drg_wt_num)
```

```
##  [1] 2.0036 3.2962 2.0501 1.6344 3.2010 2.0543 0.0000 2.0671 2.0623 3.0199
## [11] 1.9790 1.9898 2.9990 1.6692 1.6645 3.0014 1.6228 3.2906 2.0816 3.0304
## [21] 3.1742 1.7328 1.1531 1.6769 1.0000 1.4076 1.1287 1.0973 1.1001 1.3946
## [31] 1.1821 0.9109 1.6921 1.3888 1.1913 1.5031 0.6958 1.6357 1.6240
```

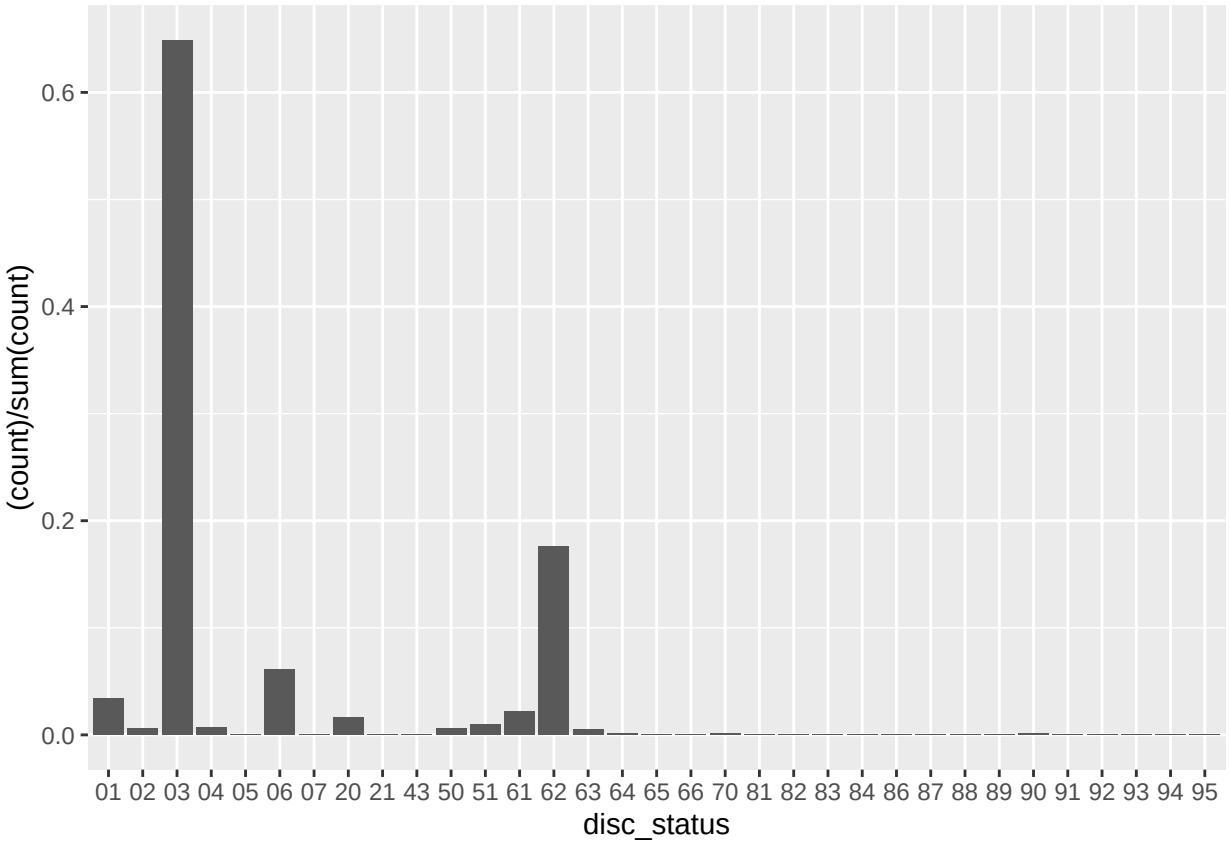
re_clm_pmt_amt:
Claim payment amount of the 1st readmission episode

```
summary(data$re_clm_pmt_amt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -24039    6394    9410   11471   13055  706277  342935
```

disc_status:
Discharge status from index episode

```
ggplot(data, aes(x = disc_status)) +
  geom_bar(aes(y = (..count..)/sum(..count..)))
```



Entries of scores

cci_score_1825_days_b is **CCI score**: quantifies an individual's burden of disease and corresponding 1-year mortality risk.

```
length(unique(data$cci_score_1825_days_b))
```

```
## [1] 23
```

```
unique(data$cci_score_1825_days_b)
```

```
##  [1]  1  5  2  0  3  6 10  7 12  4  8  9 16 11 13 17 14 15 18 19 20 22 21
```

elix_score_1825_days_b is **Elixhauser Score**: A method for measuring patient comorbidity based on ICD-9-CM and ICD-10 diagnosis codes found in administrative data.

```
length(unique(data$elix_score_1825_days_b))
```

```
## [1] 25
```

```
unique(data$elix_score_1825_days_b)
```

```
##  [1]  5 10  6  4  9  8  0  3 11  2  1 12  7 13 18 16 14 15 19 17 21 20 22 23 24
```

fci_score_1825_days_b is **FCI Score**: is calculated by looking at the ratio of the required renewal cost of the current year to the current building replacement value

```
length(unique(data$fci_score_1825_days_b))
```

```
## [1] 19
```

```
unique(data$fci_score_1825_days_b)
```

```
##  [1]  5  6  4  3  7  0  8  9  1  2 12 11 10 13 15 14 17 16 18
```

Empty Characters

at_physn_upin: NPIs replaced UPINs as the standard provider identifiers beginning in 2007. The UPIN is almost never populated after 2009.

NPI: On an institutional claim, the national provider identifier (NPI) number assigned to uniquely identify the physician who has overall responsibility for the beneficiary's care and treatment.

```
unique(data$at_physn_upin)
```

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

```
unique(data$op_physn_upin)
```

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

Intresting Variables

cont_enroll_flag_1825b_89f is a 0/1 flag indicating that the patient is continuously enrolled for 1825 days in baseline and 90 days in follow-up. All entries is '1'.

```r
mean(data$cont_enroll_flag_1825b_89f)
```

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

hmo_enroll_flag_1825b_89f is a 0/1 flag indicating that the patient is continuously enrolled with HMO for 1825 days in baseline and 90 days in follow-up. All entries is '0'.

```r
mean(data$hmo_enroll_flag_1825b_89f)
```

```
## [1] 0
```

These variables has variance 0.

```r
summary(data$re_clm_pmt_amt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -24039    6394    9410   11471   13055  706277  342935
```