lab1-bigdata

Wanni Lei

2023-01-13

2.1

File icd10cm_codes_2020.txt contains ICD-10-CM codes that are to be used from October 1, 2019 through September 30, 2020.

Using this file create a data frame that has two columns: "ICD10" and "description".

```
setwd("/Users/wanni/Desktop/big-data/lecture/lab/lab1")
icd10cm <- read.delim(file ="icd10cm_codes_2020.txt")

icd10cm2 <- str_split_fixed(icd10cm$A000....Cholera.due.to.Vibrio.cholerae.01..biovar.cholerae, " ",1)

string <- function(df){
icd10cm22 <- strsplit(df, " ")[[1]][1] ## results of strsplit is list
icd10cm22
}

ICD10 <- apply(icd10cm2, 1, string)
n <- str_length (ICD10)

description <- c()
for (i in 1:length(icd10cm2)){
a <- str_trim(substring(icd10cm2[i],(n[i]+1),last=1000000L))
description[i] <- a
}

ICD10<- as.vector(ICD10)
description <- as.vector(description)

df <-data.frame(ICD10, description)
head(df)
```

Solution:

```
##   ICD10                                    description
## 1  A001 Cholera due to Vibrio cholerae 01, biovar eltor
## 2  A009                              Cholera, unspecified
## 3 A0100                       Typhoid fever, unspecified
## 4 A0101                                 Typhoid meningitis
## 5 A0102            Typhoid fever with heart involvement
## 6 A0103                                 Typhoid pneumonia
```

2.2

From the created data frame find a number of different diagnoses for the first chapter "Certain infectious and parasitic diseases" with codes start at "A00" and end at "B99".

```
# find a number of different diagnoses for the first chapter
df[1:10,2]
```

Solution:

```
##  [1] "Cholera due to Vibrio cholerae 01, biovar eltor"
##  [2] "Cholera, unspecified"
##  [3] "Typhoid fever, unspecified"
##  [4] "Typhoid meningitis"
##  [5] "Typhoid fever with heart involvement"
##  [6] "Typhoid pneumonia"
##  [7] "Typhoid arthritis"
##  [8] "Typhoid osteomyelitis"
##  [9] "Typhoid fever with other complications"
## [10] "Paratyphoid fever A"
```

3.1

Select only first admission for each patient

```
setwd("/Users/wanni/Desktop/big-data/lecture/lab/lab1")
Claims <- read.csv("DE1_0_2008_to_2010_Inpatient_Claims_Sample_1.csv")

Summary <- read.csv("DE1_0_2008_Beneficiary_Summary_File_Sample_1.csv")


# Remove duplicates based on DESYNPUF_ID
Claims_nodup <- Claims[!duplicated(Claims$DESYNPUF_ID), ]
head(Claims_nodup)
```

Solution:

```
##         DESYNPUF_ID       CLM_ID SEGMENT CLM_FROM_DT CLM_THRU_DT PRVDR_NUM
## 1  00013D2EFD8E45D1 1.966612e+14       1    20100312    20100313    2600GD
## 2  00016F745862898F 1.962012e+14       1    20090412    20090418    3900MB
## 6  00052705243EA128 1.969912e+14       1    20080912    20080912    1401HG
## 7  0007F12A492FD25D 1.966612e+14       1    20080919    20080922    3400WD
## 11 000B97BA2314E971 1.962312e+14       1    20091209    20091213    2200GD
## 12 000C7486B11E7030 1.966412e+14       1    20081015    20081021    4400MM
##    CLM_PMT_AMT NCH_PRMRY_PYR_CLM_PD_AMT AT_PHYSN_NPI OP_PHYSN_NPI OT_PHYSN_NPI
## 1         4000                       0   3139083564           NA           NA
```

```
## 2        26000                       0   6476809087          NA          NA
## 6        14000                       0   6132010904          NA          NA
## 7         5000                       0   8956735757  6551008003          NA
## 11        2000                       0   3115083157          NA          NA
## 12       30000                       0   5520894646  6142535054  6786236779
##    CLM_ADMSN_DT ADMTNG_ICD9_DGNS_CD CLM_PASS_THRU_PER_DIEM_AMT
## 1      20100312               4580                         0
## 2      20090412               7866                         0
## 6      20080912              78079                         0
## 7      20080919              78097                         0
## 11     20091209               4019                         0
## 12     20081015               4260                         0
##    NCH_BENE_IP_DDCTBL_AMT NCH_BENE_PTA_COINSRNC_LBLTY_AM
## 1                    1100                             0
## 2                    1068                             0
## 6                    1024                             0
## 7                    1024                             0
## 11                   1068                             0
## 12                   1024                             0
##    NCH_BENE_BLOOD_DDCTBL_LBLTY_AM CLM_UTLZTN_DAY_CNT NCH_BENE_DSCHRG_DT
## 1                               0                  1           20100313
## 2                               0                  6           20090418
## 6                               0                  1           20080912
## 7                               0                  3           20080922
## 11                              0                  4           20091213
## 12                              0                  6           20081021
##    CLM_DRG_CD ICD9_DGNS_CD_1 ICD9_DGNS_CD_2 ICD9_DGNS_CD_3 ICD9_DGNS_CD_4
## 1         217           7802          78820          V4501           4280
## 2         201           1970           4019           5853           7843
## 6         201            486           3004          42731          42830
## 7         951          33811          53550          42820            496
## 11        241           4010          78791          60000          41401
## 12        308          42781          41400           4260           5849
##    ICD9_DGNS_CD_5 ICD9_DGNS_CD_6 ICD9_DGNS_CD_7 ICD9_DGNS_CD_8 ICD9_DGNS_CD_9
## 1            2720           4019          V4502          73300          E9330
## 2            2768          71590           2724          19889           5849
## 6            2724          V4581           4019
## 7           V1259          42731          78729           V103          34290
## 11          V1254           4372          78650           7813           4254
## 12           2720          78551          99591           4019          25000
##    ICD9_DGNS_CD_10 ICD9_PRCDR_CD_1 ICD9_PRCDR_CD_2 ICD9_PRCDR_CD_3
## 1                               NA
## 2                               NA
## 6                               NA
## 7                             8659
## 11                              NA
## 12                              50            4280           40390
##    ICD9_PRCDR_CD_4 ICD9_PRCDR_CD_5 ICD9_PRCDR_CD_6 HCPCS_CD_1 HCPCS_CD_2
## 1                                                          NA         NA
## 2                                                          NA         NA
## 6                                                          NA         NA
## 7                                                          NA         NA
## 11                                                         NA         NA
## 12             5180                                        NA         NA
```

```
##    HCPCS_CD_3 HCPCS_CD_4 HCPCS_CD_5 HCPCS_CD_6 HCPCS_CD_7 HCPCS_CD_8 HCPCS_CD_9
## 1          NA         NA         NA         NA         NA         NA         NA
## 2          NA         NA         NA         NA         NA         NA         NA
## 6          NA         NA         NA         NA         NA         NA         NA
## 7          NA         NA         NA         NA         NA         NA         NA
## 11         NA         NA         NA         NA         NA         NA         NA
## 12         NA         NA         NA         NA         NA         NA         NA
##    HCPCS_CD_10 HCPCS_CD_11 HCPCS_CD_12 HCPCS_CD_13 HCPCS_CD_14 HCPCS_CD_15
## 1           NA          NA          NA          NA          NA          NA
## 2           NA          NA          NA          NA          NA          NA
## 6           NA          NA          NA          NA          NA          NA
## 7           NA          NA          NA          NA          NA          NA
## 11          NA          NA          NA          NA          NA          NA
## 12          NA          NA          NA          NA          NA          NA
##    HCPCS_CD_16 HCPCS_CD_17 HCPCS_CD_18 HCPCS_CD_19 HCPCS_CD_20 HCPCS_CD_21
## 1           NA          NA          NA          NA          NA          NA
## 2           NA          NA          NA          NA          NA          NA
## 6           NA          NA          NA          NA          NA          NA
## 7           NA          NA          NA          NA          NA          NA
## 11          NA          NA          NA          NA          NA          NA
## 12          NA          NA          NA          NA          NA          NA
##    HCPCS_CD_22 HCPCS_CD_23 HCPCS_CD_24 HCPCS_CD_25 HCPCS_CD_26 HCPCS_CD_27
## 1           NA          NA          NA          NA          NA          NA
## 2           NA          NA          NA          NA          NA          NA
## 6           NA          NA          NA          NA          NA          NA
## 7           NA          NA          NA          NA          NA          NA
## 11          NA          NA          NA          NA          NA          NA
## 12          NA          NA          NA          NA          NA          NA
##    HCPCS_CD_28 HCPCS_CD_29 HCPCS_CD_30 HCPCS_CD_31 HCPCS_CD_32 HCPCS_CD_33
## 1           NA          NA          NA          NA          NA          NA
## 2           NA          NA          NA          NA          NA          NA
## 6           NA          NA          NA          NA          NA          NA
## 7           NA          NA          NA          NA          NA          NA
## 11          NA          NA          NA          NA          NA          NA
## 12          NA          NA          NA          NA          NA          NA
##    HCPCS_CD_34 HCPCS_CD_35 HCPCS_CD_36 HCPCS_CD_37 HCPCS_CD_38 HCPCS_CD_39
## 1           NA          NA          NA          NA          NA          NA
## 2           NA          NA          NA          NA          NA          NA
## 6           NA          NA          NA          NA          NA          NA
## 7           NA          NA          NA          NA          NA          NA
## 11          NA          NA          NA          NA          NA          NA
## 12          NA          NA          NA          NA          NA          NA
##    HCPCS_CD_40 HCPCS_CD_41 HCPCS_CD_42 HCPCS_CD_43 HCPCS_CD_44 HCPCS_CD_45
## 1           NA          NA          NA          NA          NA          NA
## 2           NA          NA          NA          NA          NA          NA
## 6           NA          NA          NA          NA          NA          NA
## 7           NA          NA          NA          NA          NA          NA
## 11          NA          NA          NA          NA          NA          NA
## 12          NA          NA          NA          NA          NA          NA
```

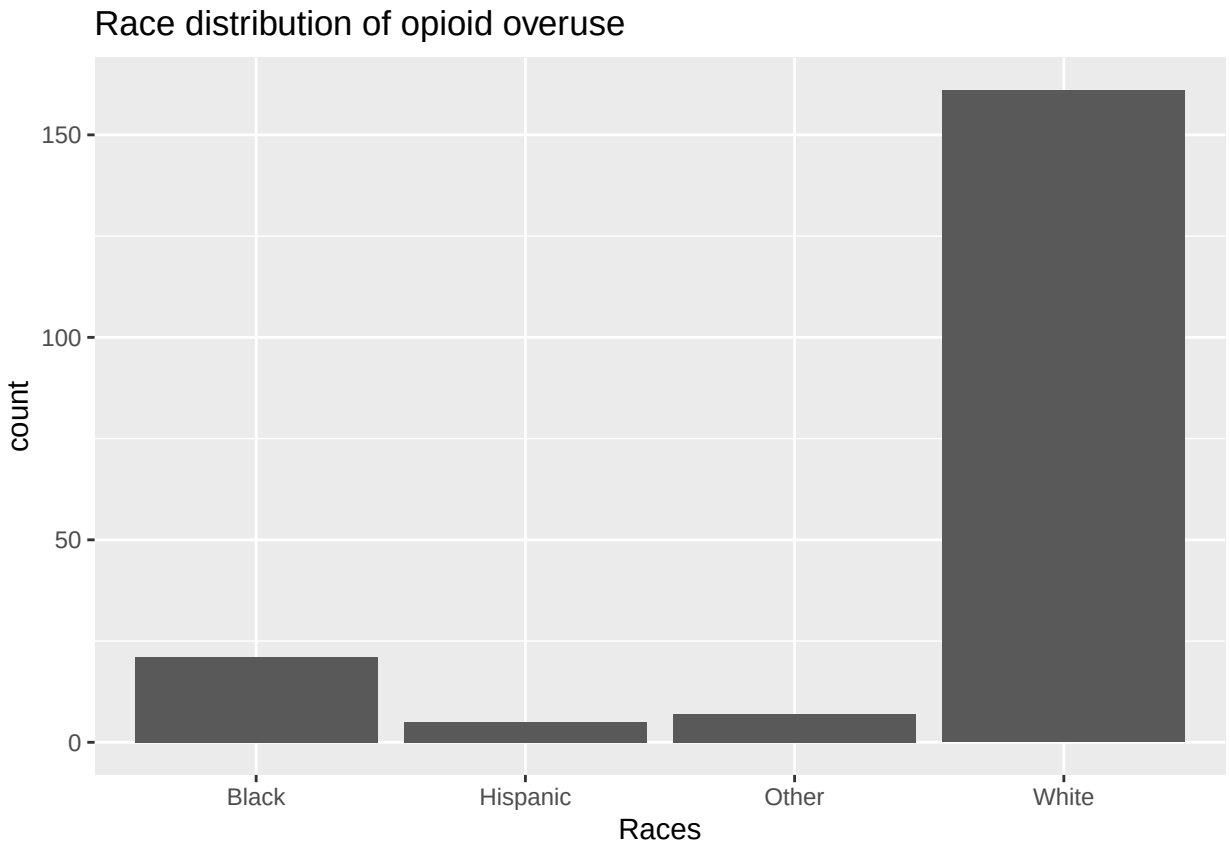
3.2

Using both files, find the race distribution of opioid overuse.

```r
## filter code 304 and 305
Claims_nodup1 <- Claims_nodup %>% filter(CLM_DRG_CD=="304"|CLM_DRG_CD=="305")

data_merge <- merge(Claims_nodup1, Summary, by = "DESYNPUF_ID")
grouped <- data_merge %>%
  group_by(BENE_RACE_CD) %>%
  summarise(frequency=n())

grouped
```

```
## # A tibble: 4 x 2
##   BENE_RACE_CD frequency
##          <int>     <int>
## 1            1       161
## 2            2        21
## 3            3         7
## 4            5         5
```

```r
data_new <- data_merge %>% mutate(BENE_RACE_CD=
  case_when(BENE_RACE_CD==1~"White",
            BENE_RACE_CD ==2~ "Black",
            BENE_RACE_CD ==3~ "Other",
             BENE_RACE_CD ==5~ "Hispanic"))

data_new %>% ggplot()+
  geom_bar(aes(x=BENE_RACE_CD))+
  xlab("Races")+
  ggtitle("Race distribution of opioid overuse")
```



The distribution of race distribution of opioid overuse is shown as the plots.

3.3

Comment on your results

Solution: White has the largest proportion of opioid overuse, Hispanic has the smallest proportion of opioid overuse, black has a medium proportion of opioid overuse, and Other races have a small proportion of opioid overuse.