Big Data-Lab1

Haotian Liu-hal4021

2023-01-15

Problem 2

```
# Read in data, header = FALSE
icd10 <- read.delim("icd10cm_codes_2020.txt",header = FALSE)
# Split by white space, make new columns
icd10 <- icd10 %>%
  mutate(col = str_replace(V1, "\\s", "|")) %>%
  separate(col, into = c("ICD10", "description"), sep = "\\|")
# Remove first column
icd10 <- as.data.frame(icd10) %>%
  select(-V1)

chapter1 <- icd10 %>%
    filter(substr(ICD10, 1, 1) == "A" | substr(ICD10, 1, 1) == "B")
# First ten observations
chapter1[1:10,]
```

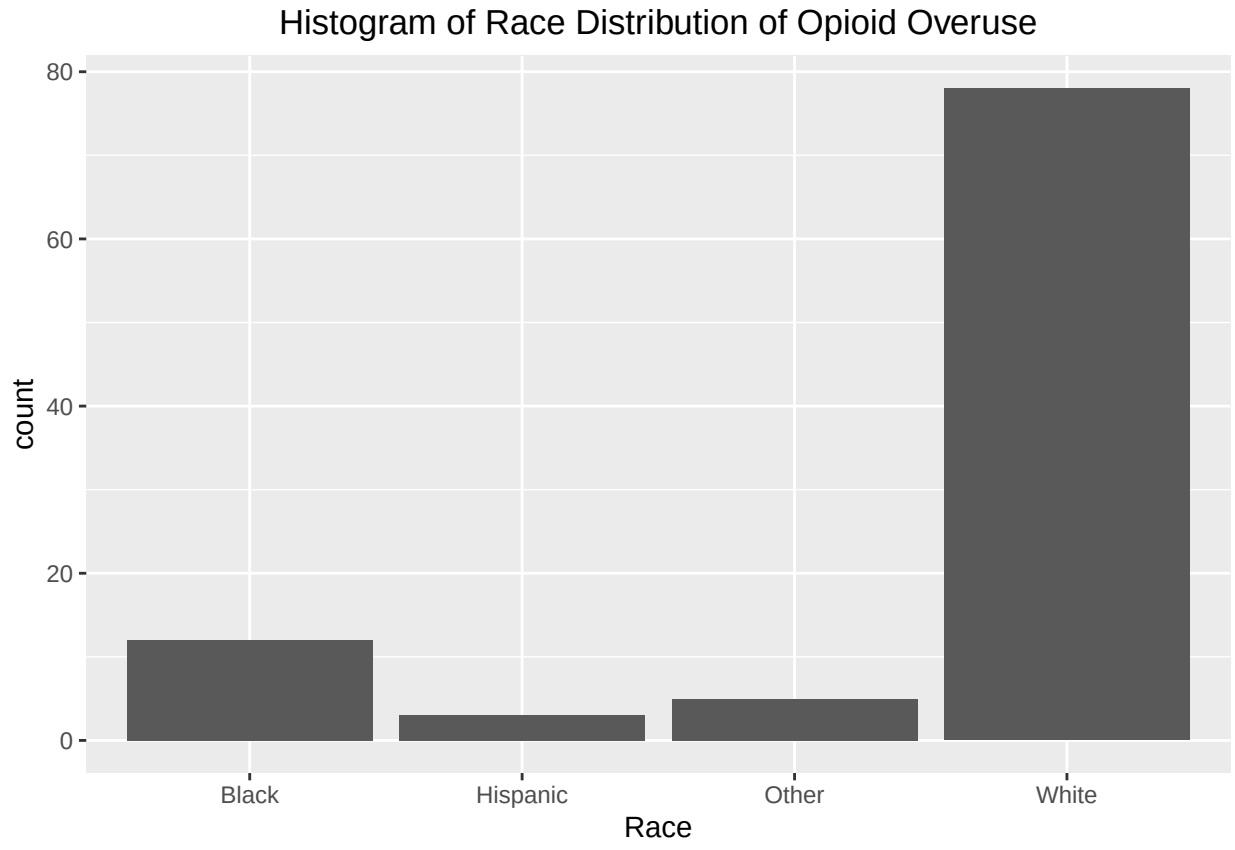
```
##    ICD10                                          description
## 1   A000    Cholera due to Vibrio cholerae 01, biovar cholerae
## 2   A001       Cholera due to Vibrio cholerae 01, biovar eltor
## 3   A009                                    Cholera, unspecified
## 4  A0100                               Typhoid fever, unspecified
## 5  A0101                                      Typhoid meningitis
## 6  A0102                  Typhoid fever with heart involvement
## 7  A0103                                       Typhoid pneumonia
## 8  A0104                                       Typhoid arthritis
## 9  A0105                                   Typhoid osteomyelitis
## 10 A0109                Typhoid fever with other complications
```

There are 1058 different diagnoses for the first chapter.

Problem 3

```
data <- read.csv("DE1_0_2008_to_2010_Inpatient_Claims_Sample_1.csv")
# Keep only the first admission of a patient
data <- data[!duplicated(data$DESYNPUF_ID),]
data2 <- read.csv("DE1_0_2008_Beneficiary_Summary_File_Sample_1.csv")
```

```r
# Combine data frames together
data3 <- merge(data, data2, by = "DESYNPUF_ID")
# Subset data to opioid overuse
opioid_overuse <- data3 %>%
  select(1,20,85) %>%
  group_by(DESYNPUF_ID, BENE_RACE_CD) %>%
# Filter data to opioid abuse: code 305
  filter(CLM_DRG_CD == 305)
opioid_overuse <- opioid_overuse %>%
  mutate(BENE_RACE_CD = case_when(BENE_RACE_CD == 1 ~ "White",
                                  BENE_RACE_CD == 2 ~ "Black",
                                  BENE_RACE_CD == 3 ~ "Other",
                                  BENE_RACE_CD == 4 ~ "Asian",
                                  BENE_RACE_CD == 5 ~ "Hispanic")
           )
# Plot distribution of overuse
opioid_overuse %>%
    ggplot(aes(x = BENE_RACE_CD)) +
    geom_bar() +
    labs(x = "Race", title = "Histogram of Race Distribution of Opioid Overuse") +
    theme(plot.title = element_text(hjust = 0.5))
```



From the plot, we observe that most opioid overuse inpatients are White, and least opioid overuse inpatients are Asian(There is 0 asian inpatient here). Black, Hispanic, and other race have few opioid overuse inpatients.