Project1\_JacksonDial

Jackson Dial

9/3/2021

### Load Libraries

#setwd("C:/Users/jacks/OneDrive/Desktop/DUKE\_FALL2021/721/BIOS721")  
#libraries  
library(here)  
library(readr)  
library(tidyverse)  
library(pander)  
library(lubridate)  
  
library(knitr)  
#Don't forget to do Lintr and Styler

## Read Data

#setwd("C:/Users/jacks/OneDrive/Desktop/DUKE\_FALL2021/721/BIOS721")  
#use here() function for better reproducibility  
patient <- read\_csv(here("Data", "Patient Level Data\_proj1.csv"))  
encounter <- read\_csv(here("Data", "Encounter Level Data\_proj1.csv"))  
  
#join the patient level and encounter level data frames into one on MRN (ID)  
dat <- encounter %>%  
 left\_join(patient, by = "MRN")

## Data Cleaning

#check for any wrong values in the categorical variables  
table(dat$enc\_type, dat$race)

##   
## Black Other White  
## Hospital Encounter 11 5 37  
## Office visit 101 34 362

#just using table because it does 2 at a time and is less code  
table(dat$financialclass, dat$ethnicity)

##   
## Hispanic non-Hispanic  
## Medicare 10 312  
## Private 29 199

table(dat$hypertension, dat$CHF)

##   
## N Y  
## N 284 38  
## Y 215 13

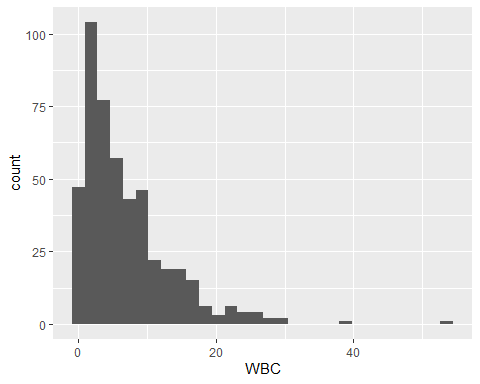
levels(as.factor(dat$diabetes))

## [1] "N" "Y"

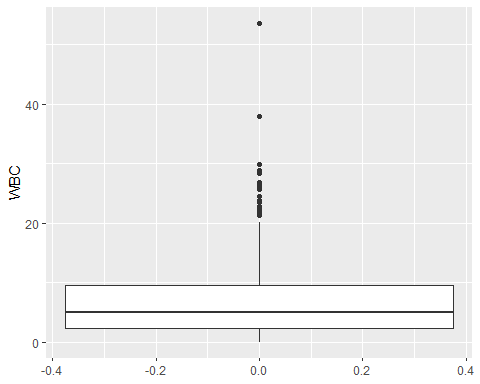
#they all look good so no imputation/removal must occur  
  
summary(dat)

## MRN contact\_date enc\_type temp   
## Length:550 Length:550 Length:550 Min. : 96.26   
## Class :character Class :character Class :character 1st Qu.: 97.92   
## Mode :character Mode :character Mode :character Median : 98.50   
## Mean : 98.54   
## 3rd Qu.: 99.07   
## Max. :103.20   
##   
## distress\_score WBC BMI.r DOB   
## Min. :0.000 Min. : 0.000 Min. :-999.00 Length:550   
## 1st Qu.:1.000 1st Qu.: 2.255 1st Qu.: 23.92 Class :character   
## Median :2.000 Median : 5.010 Median : 27.82 Mode :character   
## Mean :2.513 Mean : 6.856 Mean : 16.22   
## 3rd Qu.:4.000 3rd Qu.: 9.465 3rd Qu.: 32.50   
## Max. :7.000 Max. :53.600 Max. : 352.13   
## NA's :53 NA's :72 NA's :86   
## race financialclass ethnicity hypertension   
## Length:550 Length:550 Length:550 Length:550   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## CHF diabetes   
## Length:550 Length:550   
## Class :character Class :character   
## Mode :character Mode :character   
##   
##   
##   
##

#replace impossible values with NA  
dat["BMI.r"][which(dat$BMI.r > 70), ] <- NA  
dat["BMI.r"][which(dat$BMI.r < 0), ] <- NA  
  
# determine if 53 is an outlier for WBC  
  
ggplot(dat, aes(x = WBC)) +  
 geom\_histogram()



ggplot(dat, aes(y = WBC)) +  
 geom\_boxplot()



dat %>% select(WBC) %>% filter(WBC > 37)

## # A tibble: 2 x 1  
## WBC  
## <dbl>  
## 1 53.6  
## 2 38.0

#it does appear that there are two possible outliers, at 38.0 & 53.6  
#I am going to replace them with NA values for now but may change it later  
dat %>% select(WBC) %>% filter(WBC <= 0) %>% nrow()

## [1] 2

# there are no values below 0, but 2 that are 0. Given the measurement  
#being in parts per thousand these will be replaced with NA values  
  
#here I am making the two outliers discovered above NA values  
dat["WBC"][which(dat$WBC > 30), ] <- NA  
  
#replacing 'impossible' values with NA  
dat["WBC"][which(dat$WBC <= 0), ] <- NA  
  
# I have defined acceptable values to be from 0-30

## Data Explanation

The data included in this analysis is comprised of two original data sets, one on patient data and one on encounter data. Patient data can be understood as information regarding a specific person, containing demographic information. This original dataset has 50 observations (patients) and 8 variables. The encounter dataset is a record of every time any of the 50 patients had an interaction with the Healthcare system being considered, and contains information regarding that visit such as reason for visit, date, and measurements regarding their current health status. This dataset contains 550 observations and 7 variables. A more precise description of each variable will be given below.

These two described data sets were joined together on the “MRN” column, which is an acronym meaning Medical Record Number. The dimension of this data frame is 550 observations with 14 variables.

### Data Dictionary

* **MRN**: Medical Record Number; an individual’s personal identificaiton code
* **contact\_date**: The date that the encounter was recorded
* **enc\_type** : The type of encounter, or visit, recorded
* **temp**: An individual’s temperature at arrival, recorded in Fahrenheit
* **distress\_score**: A score from 0-7 as reported by the patient
* **WBC**: White blood cell count, in parts per 1,000
* **BMI.r**: An individual’s BMI score
* **DOB**: Date of Birth
* **race**: An individual’s race with 3 levels:
  + White
  + Black
  + Other
* **financialclass**: An individual’s healthcare status with 2 levels:
  + Private
  + Medicare
* **ethnicity**: An individual’s ethnic background with 2 levels:
  + Hispanic
  + Non-Hispanic
* **hypertension**: A binary variable indicating whether or not an individual has hypertension
* **CHF**: A binary variable indicating whether or not an individual has congestive heart failure
* **diabetes**: A binary variable indicating whether or not an individual has diabetes

sum(is.na(dat)) #223

## [1] 223

#it looks like when enc\_type == "Hospital Encounter" the distress score and   
#wbc are NA values, let's take a better look:  
  
dat %>% filter(enc\_type == "Hospital Encounter") %>% nrow() #53 total hospital encounters

## [1] 53

sum(is.na(dat %>% filter(enc\_type == "Hospital Encounter"))) #158 = 53\*3 - 1

## [1] 159

#(there is one that is -999)  
  
  
#there is also just a bunch of random BMI.r and WBC NA values that don't   
#appear to have any patterns in them. We will just leave these as they are   
#and not bother with imputation.

#round temperature to 1 decimal  
dat <- dat %>% mutate(  
 temp\_rounded = round(temp, 1),  
 WBC\_Cat = case\_when(  
 WBC < 3.2 ~ "Low",  
 WBC < 9.8 & WBC >= 3.2 ~ "Normal",  
 WBC > 9.8 ~ "High",  
 T ~ "Not Taken"  
 )  
)

## Table of Categorical WBC

table(dat$WBC\_Cat) %>%   
 pander()

|  |  |  |  |
| --- | --- | --- | --- |
| High | Low | Normal | Not Taken |
| 111 | 167 | 195 | 77 |

## MRN Table

dat %>%   
 select(MRN, contact\_date, temp\_rounded) %>%   
 filter(temp\_rounded > 100) %>%   
 arrange(MRN) %>%   
 kable(caption = "Temperature > 100 Organized by MRN")

Temperature > 100 Organized by MRN

|  |  |  |
| --- | --- | --- |
| MRN | contact\_date | temp\_rounded |
| EG7985 | 7/7/2014 | 101.0 |
| EG7985 | 8/24/2017 | 100.2 |
| GE5166 | 3/13/2016 | 101.2 |
| HP5350 | 6/11/2019 | 100.2 |
| HQ8624 | 6/22/2015 | 100.6 |
| IO6623 | 6/22/2014 | 102.4 |
| IY1696 | 7/24/2015 | 103.1 |
| JV9469 | 11/19/2018 | 100.3 |
| NC5186 | 4/21/2019 | 100.4 |
| ON7427 | 8/28/2019 | 100.2 |
| RA3868 | 5/13/2014 | 103.2 |
| RA3868 | 12/10/2018 | 100.1 |
| TG7671 | 6/21/2018 | 100.1 |
| TJ3799 | 7/2/2018 | 100.9 |
| TO4566 | 11/13/2017 | 100.5 |
| XV9573 | 1/22/2019 | 100.1 |
| XV9573 | 4/19/2016 | 103.0 |

## Mean BMI

dat %>%   
 select(BMI.r, MRN) %>%   
 filter(MRN %in% c("CI6950", "IW9164", "HJ8458", "XE4615")) %>%   
 group\_by(MRN) %>%   
 summarise(avg\_bmi = mean(BMI.r, na.rm = T)) %>%   
 kable(digits = 1, caption = "Average BMI for Specified MRN's")

Average BMI for Specified MRN’s

|  |  |
| --- | --- |
| MRN | avg\_bmi |
| CI6950 | 25.8 |
| HJ8458 | 28.9 |
| IW9164 | 29.4 |
| XE4615 | 29.8 |

## Hospital Encounters per year

dat %>%   
 select(contact\_date, enc\_type) %>%   
 filter(enc\_type == "Hospital Encounter") %>%   
 transmute(Year = year(mdy(contact\_date))) %>%   
 group\_by(Year) %>%  
 summarise(Visits = n()) %>%   
 kable()

|  |  |
| --- | --- |
| Year | Visits |
| 2014 | 12 |
| 2015 | 9 |
| 2016 | 9 |
| 2017 | 7 |
| 2018 | 8 |
| 2019 | 8 |

## Age

age\_dat <- dat %>%   
 select(DOB) %>%   
 mutate(AgeAtStartOfCohort =   
 floor(time\_length(interval(mdy(DOB),   
 ymd("2019-09-01")), "years")))   
age\_dat %>%   
summary(AgeAtStartOfCohort) %>%   
 pander()

|  |  |
| --- | --- |
| DOB | AgeAtStartOfCohort |
| Length:550 | Min. : 20.0 |
| Class :character | 1st Qu.: 31.0 |
| Mode :character | Median : 43.0 |
| NA | Mean : 45.2 |
| NA | 3rd Qu.: 54.0 |
| NA | Max. :130.0 |

sd(age\_dat$AgeAtStartOfCohort)

## [1] 21.46259

The above table shows a summary for the age of participants. What sticks out the most is the maximum age being 130. This is unlikely and should be treated as an NA value during further analysis. The standard deviation is 21.46 years, with a mean of 45.2 and median 43.

## Patient Level Table

tab1 <- patient %>%  
 group\_by(race) %>%   
 summarise(count = n()) %>%   
 mutate(percent = count / sum(count) \* 100) %>%   
 as.data.frame() %>%   
 rename(Level = race)  
  
tab2 <- patient %>%  
 group\_by(ethnicity) %>%   
 summarise(count = n()) %>%   
 mutate(percent = count / sum(count) \* 100) %>%   
 as.data.frame() %>%   
 rename(Level = ethnicity)  
  
tab3 <- patient %>%   
 group\_by(financialclass) %>%   
 summarise(count = n()) %>%   
 mutate(percent = count / sum(count) \* 100) %>%   
 as.data.frame() %>%   
 rename(Level = financialclass)  
  
tab4 <- patient %>%   
 group\_by(hypertension) %>%   
 summarise(count = n()) %>%   
 mutate(percent = count / sum(count) \* 100) %>%   
 as.data.frame() %>%   
 rename(Level = hypertension)  
  
tab5 <- patient %>%   
 group\_by(CHF) %>%   
 summarise(count = n()) %>%   
 mutate(percent = count / sum(count) \* 100) %>%   
 as.data.frame() %>%   
 rename(Level = CHF)  
  
tab6 <- patient %>%   
 group\_by(diabetes) %>%   
 summarise(count = n()) %>%   
 mutate(percent = count / sum(count) \* 100) %>%   
 as.data.frame() %>%   
 rename(Level = diabetes)  
  
  
Variable <- c("Race", "Race", "Race", "Ethnicity","Ethnicity",   
 "Financial Class", "Financial Class", "Hypertension",   
 "Hypertension", "CHF", "CHF", "Diabetes", "Diabetes")  
  
  
cbind(Variable,(rbind(tab1, tab2, tab3, tab4, tab5, tab6))) %>% pander()

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Level | count | percent |
| Race | Black | 10 | 20 |
| Race | Other | 3 | 6 |
| Race | White | 37 | 74 |
| Ethnicity | Hispanic | 3 | 6 |
| Ethnicity | non-Hispanic | 47 | 94 |
| Financial Class | Medicare | 29 | 58 |
| Financial Class | Private | 21 | 42 |
| Hypertension | N | 30 | 60 |
| Hypertension | Y | 20 | 40 |
| CHF | N | 45 | 90 |
| CHF | Y | 5 | 10 |
| Diabetes | N | 48 | 96 |
| Diabetes | Y | 2 | 4 |

## Histogram of distress score

ggplot(dat, aes(x = as.factor(distress\_score))) +  
 geom\_histogram(stat = "count", fill = "#00539B") +  
 theme(panel.grid.minor = element\_blank(),  
 panel.grid.major.x = element\_blank()) +  
 labs(x = "Distress Score",  
 y = "Count",  
 title = "Histogram of Distress Score",  
 subtitle = "Shown in Duke Blue")

