An Exploratory Analysis of Mental Healthcare Data

BY: DREW WILIMITIS

Motivation

- **Anxiety Disorders are Very Common**
  - It is estimated that 1/3 of the population is affected by an anxiety disorder in their lifetime [1]
- **Anxiety and other Mental Health Disorders are Underreported**
  - According to a WHO study, only approximately half of the cases of anxiety disorders have been reported [1]
- **These Mental Health Disorders Impose Tremendous Healthcare and Societal Costs**
  - One study estimated a global cost of $2.5 trillion when also accounting for the indirect costs like disability, mortality, potential imprisonment, etc. [2]
- **Mental Healthcare is in its infancy, and could most likely gain significant clinical value from a data-driven approach**

Project Outline

In this exploratory data analysis, I will investigate the following questions and consider how they might improve clinical outcomes and drive value-based care:

1) **How do mental health diagnoses vary amongst different demographic groups?**

2) **Is there a correlation between mental health diagnoses and living alone?**

3) **What are some areas of further investigation?**

Preliminary Investigation

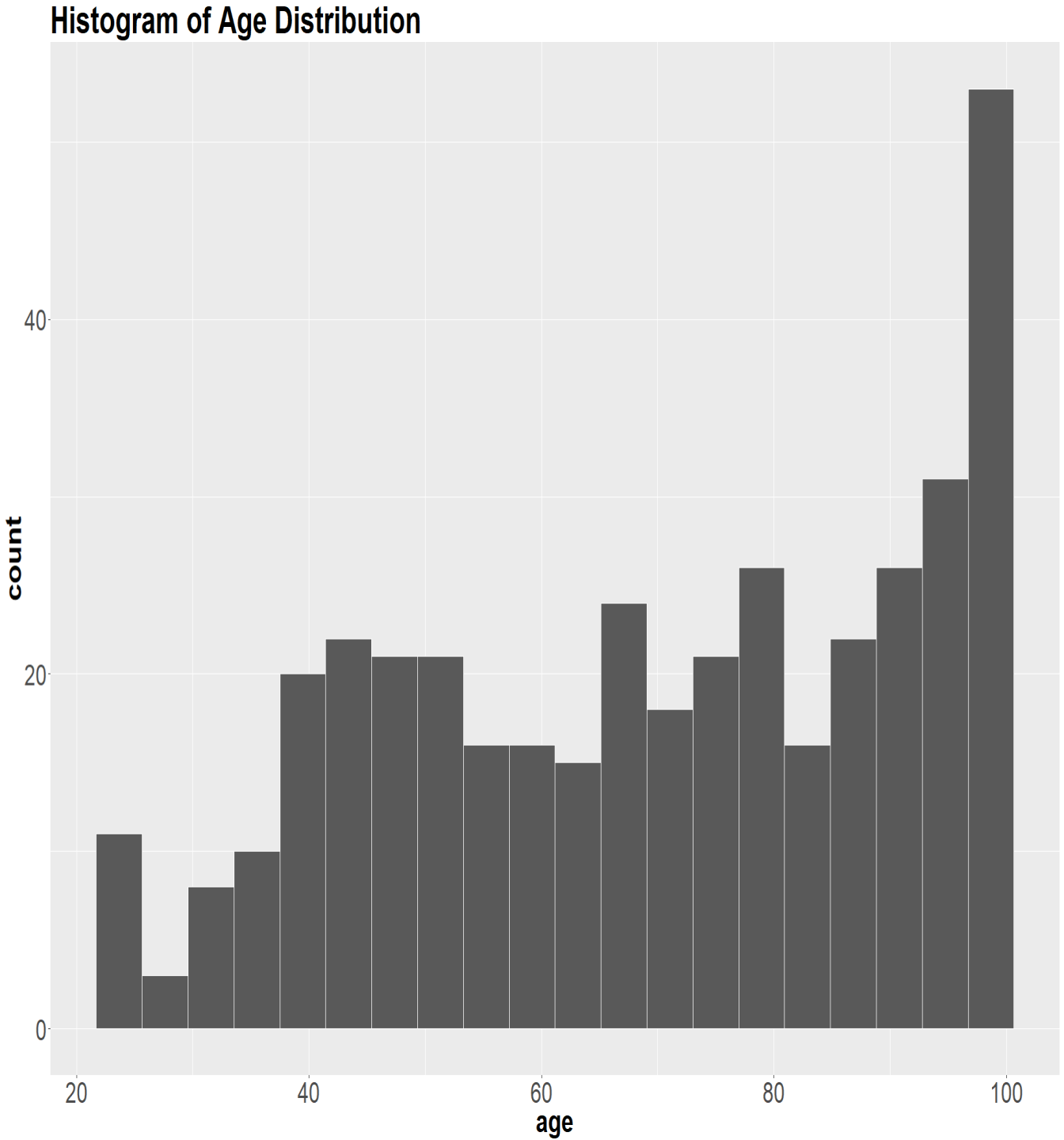
PATIENT DATA, DEMOGRAPHICS AND VISUALIZATIONS

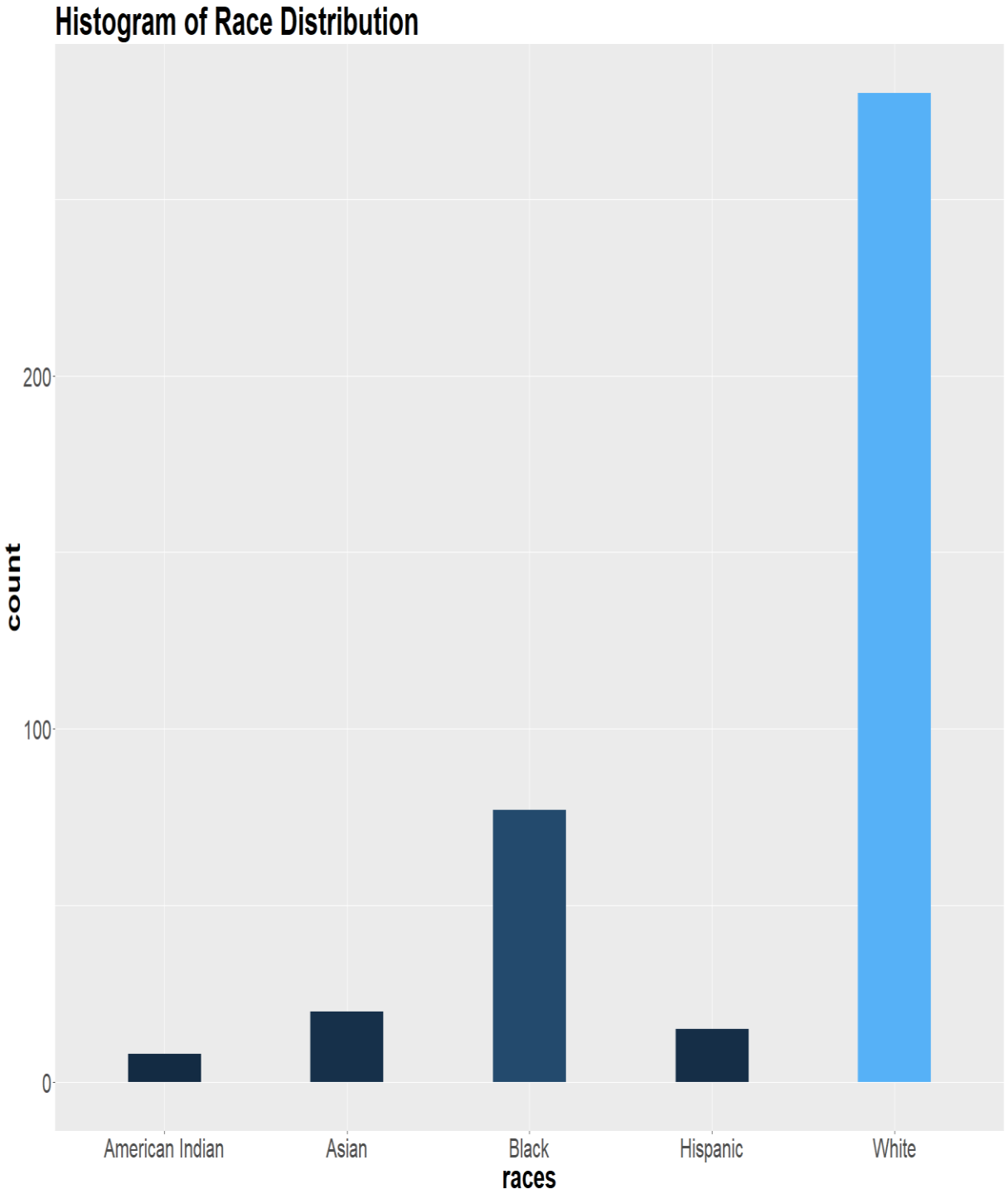
Data and the Patient Demographics

- 400 unique patients
- Datasets Explored:
  1.) Demographic Data
  2.) Diagnostic Data
  3.) ER visits data

Summary Statistics on the Patient Group

Age	Gender	Race
Min = 22 years old	200 Males	White = 70%
Max = 97 years old	200 Females	Black = 19.25%
Mean = 69 years		Hispanic = 3.75%
Median = 73 years		Asian = 5%
		American Indian = 2%


Histogram of Age Distribution
count
40
20
0
20
40
60
80
100
age


Histogram of Race Distribution
count
200
100
0
American Indian
Asian
Black
Hispanic
White
races

1: How do mental health diagnoses vary with patient demographics?

Mental Health Demographics

- Mental Health Diagnoses defined by ICD9 Codes
- 25.5% had at least one mental health diagnosis
- 102 patients total
- No bias in age or race
- Gender bias - 5:1 Female to Male Ratio

Patients with a mental health diagnosis

Age	Gender	Race
Min = 23 years old	16 Males	White = 71%
Max = 97 years old	86 Females	Black = 19%
Mean = 70 years		Hispanic = 3%
Median = 74.5 years		Asian = 5%
		American Indian = 2%

2: Can we determine any potential relationships between lifestyle factors and mental health?

Correlation between Mental Health and Lifestyle Factors

- Parsed whether patient “lives alone” or not from ER visit notes
- Among those living alone, 38% of patients have had a mental health diagnosis (higher than average)
- The severe, negative health effects of social isolation are recently being studied [3]
- Loneliness has been cited as an important predictor of mortality [3]
- Also found among patients that were “current smokers”, 38% of patients have had a mental health diagnosis (higher than average)

Limitations and Further Potential Insights

- This exploratory analysis suggests many complex relationships between patient variables
- Leads to Prediction and Modeling
- Build and test statistical/predictive models on other real datasets
- Necessary to drive the real value (prediction)

Clinical Impact and Additional Value

- Mental Health data can be easily obtained from the patient with standardized guidelines
- No expensive or painful testing needed for these diagnoses
- Mental Health data likely can be used to predict, diagnose, and treat other health conditions
- High comorbidity of these disorders might suggest different fundamental explanations of physical/mental health
- Focusing on mental health likely aligns with the transition to truly preventative healthcare

Citations/Code

- 1. Bandelow B, Michaelis S. Epidemiology of anxiety disorders in the 21st century. *Dialogues in Clinical Neuroscience*. 2015;17(3):327-335.
- 2. Trautmann S, Rehm J, Wittchen H. The economic costs of mental disorders: Do our societies react appropriately to the burden of mental disorders? *EMBO Reports*. 2016;17(9):1245-1249. doi:10.15252/embr.201642951.
- 3. Luo Y, Hawkley LC, Waite LJ, Cacioppo JT. Loneliness, Health, and Mortality in Old Age: A National Longitudinal Study. *Social science & medicine (1982)*. 2012;74(6):907-914. doi:10.1016/j.socscimed.2011.11.028

```
# Loading Dependencies
library(readr)
library(lubridate)
library(dplyr)
library(stringr)
library(ggplot2)
library(plotly)
library(MASS)
library(GGally)
library(openintro)
library(mosaic)
library(knitr)
library(tidyverse)
library(ggformula)
library(gridExtra)
library(broom)

# Loading data files
dem_data <- read_csv('C:/Users/Drew/Desktop/Ongoing Applications/Harvard Systems Medicine Data
Analyst/dem.csv')
diag_data <- read_csv('C:/Users/Drew/Desktop/Ongoing Applications/Harvard Systems Medicine Data
Analyst/dia.csv')
ed_visits <- read.delim2('C:/Users/Drew/Desktop/Ongoing Applications/Harvard Systems Medicine Data
Analyst/ed_visits.txt', sep = "$")

# Joining diag_data and dem_data
patient_data <- dem_data %>%
  full_join(diag_data, by = c('empi' = 'empi'))

patient_data_full <- patient_data %>%
  full_join(ed_visits, by = c('empi' = 'empi'))


### EXPLORING DEM DATA  (400 Unique Patients)
length(unique(dem_data$empi))

head(dem_data)
glimpse(dem_data)
summary(dem_data)
sum(is.na(dem_data))

summary(dem_data$age)
sum(is.na(dem_data$age))
sum(is.na(dem_data$gender))
sum(is.na(dem_data$date_of_death))



#####################################################
#_____________________DATA CLEANING_____________________#
#####################################################
```

```
# Cleaning dem_data
split_race <- str_split_fixed(dem_data$race, "-", 2)
dem_data$race <- split_race[, 1]
dem_data$race <- gsub("BLACK OR AFRICAN AMERICAN", "Black", dem_data$race)

# Cleaning the ed_visits.txt file
split_notes <- str_split_fixed(ed_visits$note_text, "SOCIAL HISTORY", 2)
split_2 <- str_split_fixed(split_notes[,2], "PHYSICAL EXAMINATION", 2)
ed_visits <- ed_visits_rm
ed_visits$lives_alone <- 0
ed_visits$current_smoker <- 0
ed_visits$clean_notes <- split_2[,1]

ed_visits <- ed_visits %>%
  mutate(lives_alone = as.numeric(grepl("lives alone", ed_visits$clean_notes)),
         current_smoker = as.numeric(grepl("current smoker", ed_visits$clean_notes)))

names(ed_visits)[10] <- "lifestyle_factors"

# Dropping rest of the ER Notes column
ed_visits_tmp <- ed_visits[ , !names(ed_visits) %in% c('note_text')]
ed_visits <- ed_visits_tmp

# Lifestyle Factors (27% live alone, 37% smoke, only 7.8% smoke and live alone)
# Relationship with Age/Race/Anxiety etc.?? ##
sum(ed_visits$lives_alone) / length(unique(ed_visits$empi))
sum(ed_visits$current_smoker) / length(unique(ed_visits$empi))
sum(ed_visits$lives_alone & ed_visits$current_smoker) / length(unique(ed_visits$empi))

# Joining three datasets after cleaning
patient_data <- dem_data %>%
  full_join(diag_data, by = c('empi' = 'empi'))

patient_data_full <- patient_data %>%
  full_join(ed_visits, by = c('empi' = 'empi'))


#####################################################
#_________________INITIAL EXPLORATION_________________#
#####################################################


# GRAPHING DEM DATA PLOTS

#_______ AGE HISTOGRAM (GG PLOT) _______#
gf_histogram(~ age, data = dem_data, bins = 20, color = "white")

#_______ RACIAL PROPORTION BARPLOT (BASE R) _______#
summary(dem_data$race)

barplot(prop.table(table(dem_data$race)),
        names.arg = c("American Indian", "Asian", "Black", "Hispanic", "White"),
```

```
      cex.names = 0.9)

#_______ RACIAL BARPLOT (GGPLOT) _______#
races <- dem_data$race
race_ggplot <- ggplot(data.frame(races), aes(x = races)) + geom_bar()
race_ggplot

#_______ PROPORTIONAL RACIAL BARPLOT (GGPLOT) _______#
race_prop_table <- table(dem_data$race)
summary(race_prop_table)
dem_data <- dem_data[, 1:8]
dem_data$race_prop <- 0
dem_data$race_prop[dem_data$race == "White"] <- 0.7000
dem_data$race_prop[dem_data$race == "Hispanic"] <- 0.0375
dem_data$race_prop[dem_data$race == "Black"] <- 0.1925
dem_data$race_prop[dem_data$race == "Asian"] <- 0.0500
dem_data$race_prop[dem_data$race == "American Indian"] <- 0.0200

race_df <- data.frame(races)
race_df$race_prop <- dem_data$race_prop
prop_race_ggplot <- ggplot(race_df, aes(x = races, fill = race_prop)) + geom_bar()
prop_race_ggplot

### EXPLORING DIAG DATA  (400 Unique Patients)
length(unique(dem_data$empi))

head(diag_data)
glimpse(diag_data)
summary(diag_data)
sum(is.na(diag_data))

summary(diag_data$race)
plot(table(diag_data$race))

summary(diag_data$age)
sum(is.na(diag_data$age))
sum(is.na(diag_data$gender))
sum(is.na(diag_data$date_of_death))

### EXPLORING ED_VISIT DATA (346 unique patients). NOT EVERY PATIENT HAS AN ER VISIT
RECORD
length(unique(ed_visits$empi))

head(ed_visits)
glimpse(ed_visits)
summary(ed_visits)
sum(is.na(ed_visits))

##################################################
## EXPLORING DATA WITH ANXIETY/MENTAL HEALTH ##
##################################################
```

```
# More diagnoses in the diag_data set than ER visits (makes sense).
# 688 Unique Diagnoses in diag_data, 308 unique primary diagnoses in ER_visits,
# 560 diagnoses listed as additional in ER_visits
er_diag_codes <- ed_visits[, 6:8]
diag_data_codes <- diag_data[, 3:4]
er_diag_codes$empi <- ed_visits$empi
diag_data_codes$empi <- diag_data$empi

length(unique(diag_data_codes$dia_name))
length(unique(er_diag_codes$principal_dia_name))
length(unique(diag_data_codes$dia_code))
length(unique(er_diag_codes$principal_dia_code))

# Splitting lists of additional diagnosis codes into their own columns
er_diag_codes$principal_dia_code <- as.character(er_diag_codes$principal_dia_code)
er_diag_codes$additional_dia_code <- as.character(er_diag_codes$additional_dia_code)

split_diag_codes <- str_split_fixed(er_diag_codes$additional_dia_code, "," , 4)
er_diag_codes <- er_diag_codes %>%
  mutate(add_code_1 = split_diag_codes[,1],
         add_code_2 = split_diag_codes[,2],
         add_code_3 = split_diag_codes[,3],
         add_code_4 = split_diag_codes[,4])


# Should be 1277 unique diagnosis codes total (including primary and additional from ER_visits and from
diag_data)
add1 <- er_diag_codes$add_code_1
add2 <- er_diag_codes$add_code_2
add3 <- er_diag_codes$add_code_3
add4 <- er_diag_codes$add_code_4
primary_codes <- as.character(er_diag_codes$principal_dia_code)
all_ER_codes <- c(add1, add2, add3, add4, primary_codes)
length(unique(all_ER_codes))

more_codes <- unique(diag_data$dia_code)
our_diag_codes <- c(all_ER_codes, more_codes)
our_diag_codes <- unique(our_diag_codes)

# Approach to getting all anxiety/mental health related codes/names into a table for lookup
our_codes_df <- as.data.frame(our_diag_codes, stringsAsFactors = FALSE)
relevant_codes <- sort(our_codes_df$our_diag_codes)[162:238]
relevant_codes_df <- as.data.frame(relevant_codes, stringsAsFactors = FALSE)

er_names <- as.character(ed_visits$principal_dia_name)
diag_data_names <- diag_data$dia_name
our_diag_names <- c(er_names, diag_data_names)
our_diag_names <- as.character(our_diag_names)
our_diag_names_df <- as.data.frame(our_diag_names)
our_diag_names_df <- as.data.frame(unique(our_diag_names_df$our_diag_names))
names(our_diag_names_df) <- c("unique_diag_names")
```

```
# Other approaches (Left join our relevant codes with diag_data, and then with ed_visits)
info_from_diag_data <- diag_data[ , names(diag_data) %in% c("dia_name", "dia_code")]
str(info_from_diag_data)

relevant_codes_df_copy <- relevant_codes_df
relevant_codes_df_copy <- relevant_codes_df_copy %>%
  left_join(info_from_diag_data, by = c("relevant_codes" = "dia_code"))

current_anxiety_info <- relevant_codes_df_copy

# Coerce principal_dia_code to a character
info_from_ed_visits <- ed_visits[ , names(ed_visits) %in% c("principal_dia_name",
                                                            "principal_dia_code")]

info_from_ed_visits <- as.data.frame(info_from_ed_visits, stringsAsFactors = FALSE)
info_from_ed_visits$principal_dia_code <- as.character(info_from_ed_visits$principal_dia_code)
info_from_ed_visits$principal_dia_name <- as.character(info_from_ed_visits$principal_dia_name)
class(info_from_ed_visits$principal_dia_code)
class(info_from_ed_visits$principal_dia_name)

# Anxiety/Mental Health Diagnosis Codes with Names (17 missing names)
# Filled in with available online ICD9 data sets
ICD9_lookup <- c("Drug-induced delirium",
      "Delirium due to conditions classified elsewhere",
      "Dementia, unspecified, without behavioral disturbance",
      "Other persistent mental disorders due to conditions classified elsewhere",
      "Major depressive affective disorder, single episode, moderate",
      "Major depressive affective disorder, recurrent episode, unspecified",
      "Bipolar I disorder, most recent episode (or current) manic, severe, without mention of psychotic behavior",
      "Unspecified episodic mood disorder",
      "Alcohol Dependence Syndrome",
      "Opioid type dependence, unspecified",
      "Unspecified drug dependence, unspecified",
      "Sedative, hypnotic or anxiolytic abuse, unspecified",
      "Opioid abuse, unspecified",
      "Cocaine abuse, unspecified",
      "Amphetamine or related acting sympathomimetic abuse, unspecified",
      "Predominant disturbance of emotions",
      "Adjustment disorder with disturbance of conduct")

na_val <- sort(current_anxiety_info$dia_name)[1096:1112]
na_val <- ICD9_lookup
na_rows <- c(81,82,98,99,118,119,131,134,702,705,708,712,713,714,715,841,1106)
current_anxiety_info[na_rows, ]$dia_name <- ICD9_lookup

current_anxiety_info$relevant_codes <- unique(current_anxiety_info$relevant_codes)
current_anxiety_info$dia_name <- unique(current_anxiety_info$dia_name)

anxiety_mental_health_table <- current_anxiety_info %>%
  group_by(relevant_codes) %>%
  distinct()
```

```
anxiety_only_table <- anxiety_mental_health_table[grepl("anxiety",
                                        anxiety_mental_health_table$dia_name,
                                        ignore.case = TRUE), ]

#####################################################
## VALUABLE INSIGHTS FROM ANXIETY DIAGNOSES DATA
#####################################################

# What percentage of people have recieved an anxiety/mental health diagnoses and what additional
# variables are most commonly correlated with anxiety/mental health?

# Joining all cleaned, prepared datasets
ed_visits$add_code_1 <- er_diag_code$add_code_1
ed_visits$add_code_2 <- er_diag_code$add_code_2
ed_visits$add_code_3 <- er_diag_code$add_code_3
ed_visits$add_code_4 <- er_diag_code$add_code_4

patient_data <- dem_data %>%
  full_join(diag_data, by = c('empi' = 'empi'))

patient_data_full <- patient_data %>%
  full_join(ed_visits, by = c('empi' = 'empi'))

# Dropping unnecessary columns
patient_data_full <- patient_data_full[ , !names(patient_data_full) %in% c('provider',
                                                    'dia_flag',
                                                    'inpatient_outpatient',
                                                    'visit_date',
                                                    'admit_date',
                                                    'discharge date',
                                                    'name',
                                                    'date_of_birth',
                                                    'discharge_date')]
patient_data_full$dia_code <- as.character(patient_data_full$dia_code)
patient_data_full$principal_dia_code <- as.character(patient_data_full$principal_dia_code)

patient_data_tmp <- patient_data_full %>%
  mutate(mental_health_diagnosis = ifelse(dia_code %in% anxiety_mental_health_table$relevant_codes, 1,
                                ifelse(principal_dia_code %in% anxiety_mental_health_table$relevant_codes, 1, 0)))

colnames(patient_data_tmp)[23] <- "lives_alone"

percent_mental_health <- patient_data_tmp %>%
  group_by(empi) %>%
  summarise(mental_health_diagnosis = sum(mental_health_diagnosis))

patient_mental_health_data <- percent_mental_health[, 1] %>%
  inner_join(patient_data_tmp[, c("empi", "gender", "age", "race", "dia_code", "dia_name",
                            "inpatient_outpatient",
                            "lives_alone",
                            "mental_health_diagnosis")], by = c("empi" = "empi"))
```

```
patient_mental_health_data <- patient_mental_health_data[!duplicated.data.frame(patient_mental_health_data), ]


#_______ 1.) DEMOGRAPHICS OF PATIENTS WITH MENTAL HEALTH DIAGNOSES_______#

# Gives 291 patients with at least one mental health diagnosis among diagnostic data
# and ER records. => 72.25% of the whole patient cohort with a mental health diagnosis
pct_diagnosed_ <- sum(percent_mental_health$mental_health_diagnosis != 0) / 400

patient_mental_health_race <- patient_mental_health_data <- percent_mental_health[, 1] %>%
  inner_join(patient_data_tmp[, c("empi", "gender")], by = c("empi" = "empi"))

mental_health_dem <- percent_mental_health %>%
  left_join(dem_data, by = c("empi", "empi"))

table(mental_health_dem[mental_health_dem$mental_health_diagnosis != 0, ]$race)
table(mental_health_dem[mental_health_dem$mental_health_diagnosis != 0, ]$gender)
table(mental_health_dem[mental_health_dem$mental_health_diagnosis != 0, ]$age)

mental_health_dem_2 <- mental_health_dem[mental_health_dem$mental_health_diagnosis != 0, c("gender",
"age", "race")]
mental_health_dem_2$gender <- ifelse(mental_health_dem_2$gender == "male", 1, 2)

hist(mental_health_dem_2$race)
hist(mental_health_dem_2$age)



#_______ 2.) LONELINESS, SMOKING, AND MENTAL HEALTH_______#
living_alone <- patient_data_tmp[, c("empi", "name", "lives_alone")]
living_alone <- living_alone %>%
  group_by(empi) %>%
  summarise(lives_alone = sum(lives_alone))

living_alone[is.na(living_alone$lives_alone), ]$lives_alone <- 0
living_alone[living_alone$lives_alone != 0, ]$lives_alone <- 1

living_alone_data <- living_alone[living_alone$lives_alone != 0, ] %>%
  inner_join(percent_mental_health, by = c("empi", "empi"))

living_alone_data_tmp <- living_alone_data[living_alone_data$mental_health_diagnosis != 0, ]

current_smokers <- patient_data_tmp[, c("empi", "name", "current_smoker")]
current_smokers <- current_smokers %>%
  group_by(empi) %>%
  summarise(current_smokers = sum(current_smoker))

current_smokers[is.na(current_smokers$current_smokers), ]$current_smokers <- 0
current_smokers[current_smokers$current_smokers != 0, ]$current_smokers <- 1

current_smokers_data <- current_smokers[current_smokers$current_smokers != 0, ] %>%
  inner_join(percent_mental_health, by = c("empi", "empi"))
```

```
current_smokers_data_tmp <- current_smokers_data[current_smokers_data$mental_health_diagnosis != 0, ]
```