Problem Set 6

NAME HERE

10/1/2020

Question

There's a few inconsistencies with how NAs have been recorded in the gender and orienation column. We have blanks, -999, -1, and NA. Fix these inconsistencies by changing the value to NA.

How many NAs do we have in the entire data set?

```
# check unique gender & orientation entries
unique(df$gender)
```

```
## [1] "-999"   "female" "-1"     "male"   NA
```

```
unique(df$orientation)
```

```
## [1] "heterosexual"      "lesbian/gay woman" "gay"
## [4] "-999"              "-1"                NA
## [7] "other"
```

```
df <- df %>%
  mutate(gender = if_else(gender %in% c("female", "male"), # if female/male
                          gender, # keep same
                          NA_character_), # otherwise, NA
         orientation = if_else(orientation %in% # check for:
                                 c("heterosexual", "lesbian/gay woman", "gay", "other"),
                               orientation, # keep same
                               NA_character_)) # otherwise NA

# 25 NAs
sum(is.na(df))
```

```
## [1] 34
```

Question

At a glance, we can already see errors with city and state names. Let's first fix these entries to have uniform naming where cities are properly capitalized and states are capitalized. For example, we want to see "San Antonio" and "TX" rather than "san Antonio" and "tx". We want you to use distinct(), pull(), and case_when() for this question.

```
 # pull/look at unique city names
df %>%
  select(city) %>%
  distinct() %>%
  pull()
```

```
##  [1] "atlanta"     "Atlanta"     "atlAnTa"     "San Antonio" "austin"
##  [6] "oakland"     "Hayward"     "hayward"     "san Antonio" "iakland"
## [11] "Haywarf"
```

```
# pull/look at unique states
df %>%
  select(state) %>%
  distinct() %>%
  pull()
```

```
##  [1] "GA"  "gA"  "TX"  "tX"  "ca"  "CA"  NA    "ga"  "tx"  "C A" "G A" "CA_"
```

```
# fix city and state using case_when()
df <- df %>%
  mutate(
    city = case_when(
      city %in% c("Atlanta", "atlanta", "atlAnTa") ~ "Atlanta",
      city %in% c("Austin", "austin") ~ "Austin",
      city %in% c("San Antonio", "san Antonio") ~ "San Antonio",
      city %in% c("Oakland", "oakland", "iakland") ~ "Oakland",
      city %in% c("Hayward", "hayward", "Haywarf") ~ "Hayward"),
    state = case_when(
      state %in% c("GA", "gA", "ga", "G A") ~ "GA",
      state %in% c("TX", "tX", "tx") ~ "TX",
      state %in% c("CA", "ca", "C A", "CA_") ~ "CA"))
```

Question

Format the date column into a date format. Ominously, these interventions all occurred on the 25th day of the month.

```
df$date <- dmy(df$date)
```

Question

More errors! We see that some cities and states do not match appropriately. We can assume there were no errors with the date data. Use the following information to fix the state information. We know the interventions occurred in these cities during the following dates:

- 03/2018 - Oakland
- 03/2018 - Hayward
- 05/2018 - Atlanta
- 02/2019 - San Antonio
- 02/2019 - Austin

```
df <- df %>%
  mutate(state = case_when(
    date == as.Date("2018-03-25") ~ "CA",
    date == as.Date("2018-05-25") ~ "GA",
    date == as.Date("2019-02-25") ~ "TX"
  ))
```

Question

Now we want to fix the city information, but you may realize that we have two cities in California during the same date. We can't, at least from our data, distinguish the difference. Let's drop those rows with this inconsistency. One suggestion is to create a variable indicating whether to drop the row. If you performed this step correctly you should have 45 rows.

```
df <- df %>%
  # create drop variable to indicate which rows to drop
  mutate(drop = case_when(
    state == "CA" & city %in% c("Oakland", "Hayward") ~ "keep",
    state == "GA" & city == "Atlanta" ~ "keep",
    state == "TX" & city %in% c("San Antonio", "Austin") ~ "keep",
    TRUE ~ NA_character_)) %>%
  drop_na(drop)
```

Question

We have one last issue: our interventions column has missing data. We have two interventions that occurred in these locations:

- Intervention 1: Hayward, Atlanta, San Antonio
- Intervention 2: Oakland, Atlanta, Austin

For all of the cities except Atlanta it's clear what intervention took place. Fix these clear instances. As for Atlanta, we are forced to throw out these observations since we cannot reliably determine which intervention occurred. If you performed this step correctly you should have 44 rows.

```
df <- df %>%
  mutate(intervention = case_when(
    city %in% c("Hayward", "San Antonio") ~ 1,
    city %in% c("Oakland", "Austin") ~ 2,
    TRUE ~ intervention)) %>%
  drop_na(intervention)
```

Challenge

Create a box plot comparing the two interventions and their outcome. The outcome is a continuous variable from 0 to 10.

```
# make intervention a factor
df <- df %>% mutate(intervention = as.factor(intervention))

ggplot(df, aes(x=intervention, y=outcome)) +
  geom_boxplot() +
  theme_minimal()
```