EDA

Reagan Gonzales & Abril Perez

Loading/cleaning Football Data Set

```
library(nflreadr)
library(tidyverse)
football_games <- load_schedules(2020:2023)
```

I will start by exploring the unique football stadiums in the data set.

```
# Checking unique football stadiums
unique(football_games$stadium_id)
```

```
 [1] "KAN00" "ATL97" "BAL00" "BUF00" "CAR00" "DET00" "JAX00" "MIN01" "BOS00"
[10] "WAS00" "CIN00" "NOR00" "SFO01" "LAX01" "NYC01" "DEN00" "CLE00" "CHI98"
[19] "DAL00" "GNB00" "IND00" "MIA00" "PHI00" "PIT00" "TAM00" "NAS00" "PHO00"
[28] "HOU00" "SEA00" "VEG00" "LON02" "LON00" "GER00" "MEX00" "FRA00"
```

Of these stadiums, I can see that the only stadium in LA is SoFi Stadium, so I will filter the data set to focus on games held at this stadium.

```
# Filter the dataset for games played in SoFi
la_games <- football_games |> filter(stadium == "SoFi Stadium")
```

```
# For my later join
game_days <- la_games |>
  distinct(gameday) |>
  mutate(game_day = 1)
```

Loading/cleaning DV LA Data Set

```
la_data <- read_csv("Domestic_Violence_Calls_from_2020_to_Present_20250415.csv")
la_data <- janitor::clean_names(la_data)
```

```
head(la_data)
```

```
# A tibble: 6 x 28
   dr_no date_rptd date_occ time_occ area  area_name rpt_dist_no part_1_2 crm_cd
   <dbl> <chr>     <chr>    <chr>    <chr> <chr>     <chr>          <dbl>  <dbl>
1 2.00e8 05/12/20~ 05/10/2~ 2200     01    Central   0111               2    626
2 2.01e8 12/07/20~ 12/07/2~ 1203     09    Van Nuys  0935               2    900
3 2.01e8 09/17/20~ 09/17/2~ 2255     05    Harbor    0558               2    901
4 2.00e8 08/28/20~ 05/01/2~ 0100     01    Central   0154               1    121
5 2.01e8 08/04/20~ 08/04/2~ 1000     12    77th Str~ 1249               2    626
6 2.01e8 10/04/20~ 10/04/2~ 1800     12    77th Str~ 1268               2    626
# i 19 more variables: crm_cd_desc <chr>, mocodes <chr>, vict_age <dbl>,
#   vict_sex <chr>, vict_descent <chr>, premis_cd <dbl>, premis_desc <chr>,
#   weapon_used_cd <dbl>, weapon_desc <chr>, status <chr>, status_desc <chr>,
#   crm_cd_1 <dbl>, crm_cd_2 <dbl>, crm_cd_3 <dbl>, crm_cd_4 <lgl>,
#   location <chr>, cross_street <chr>, lat <dbl>, lon <dbl>
```

```
unique(la_data$crm_cd_desc)
```

```
 [1] "INTIMATE PARTNER - SIMPLE ASSAULT"
 [2] "VIOLATION OF COURT ORDER"
 [3] "VIOLATION OF RESTRAINING ORDER"
 [4] "RAPE, FORCIBLE"
 [5] "CRIMINAL THREATS - NO WEAPON DISPLAYED"
 [6] "BRANDISH WEAPON"
 [7] "INTIMATE PARTNER - AGGRAVATED ASSAULT"
 [8] "LETTERS, LEWD  -  TELEPHONE CALLS, LEWD"
 [9] "OTHER ASSAULT"
[10] "ROBBERY"
[11] "THEFT PLAIN - PETTY ($950 & UNDER)"
[12] "VANDALISM - MISDEAMEANOR ($399 OR UNDER)"
[13] "DOCUMENT FORGERY / STOLEN FELONY"
[14] "CONTEMPT OF COURT"
[15] "VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)"
```

```
[16] "ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT"
[17] "BURGLARY"
[18] "OTHER MISCELLANEOUS CRIME"
[19] "THEFT PLAIN - ATTEMPT"
[20] "THEFT, PERSON"
[21] "ATTEMPTED ROBBERY"
[22] "FALSE IMPRISONMENT"
[23] "BATTERY - SIMPLE ASSAULT"
[24] "STALKING"
[25] "TRESPASSING"
[26] "SEXUAL PENETRATION W/FOREIGN OBJECT"
[27] "KIDNAPPING"
[28] "KIDNAPPING - GRAND ATTEMPT"
[29] "BATTERY WITH SEXUAL CONTACT"
[30] "THEFT-GRAND ($950.01 & OVER)EXCPT,GUNS,FOWL,LIVESTK,PROD"
[31] "VIOLATION OF TEMPORARY RESTRAINING ORDER"
[32] "THEFT FROM MOTOR VEHICLE - GRAND ($950.01 AND OVER)"
[33] "CHILD STEALING"
[34] "RAPE, ATTEMPTED"
[35] "VEHICLE - STOLEN"
[36] "SODOMY/SEXUAL CONTACT B/W PENIS OF ONE PERS TO ANUS OTH"
[37] "ORAL COPULATION"
[38] "SEX,UNLAWFUL(INC MUTUAL CONSENT, PENETRATION W/ FRGN OBJ"
[39] "EMBEZZLEMENT, GRAND THEFT ($950.01 & OVER)"
[40] "SEX OFFENDER REGISTRANT OUT OF COMPLIANCE"
[41] "CHILD NEGLECT (SEE 300 W.I.C.)"
[42] "THREATENING PHONE CALLS/LETTERS"
[43] "CRIMINAL HOMICIDE"
[44] "THEFT FROM PERSON - ATTEMPT"
[45] "UNAUTHORIZED COMPUTER ACCESS"
[46] "THEFT OF IDENTITY"
[47] "CHILD ABUSE (PHYSICAL) - SIMPLE ASSAULT"
[48] "PEEPING TOM"
[49] "CHILD ANNOYING (17YRS & UNDER)"
[50] "BURGLARY FROM VEHICLE"
[51] "BURGLARY, ATTEMPTED"
[52] "HUMAN TRAFFICKING - INVOLUNTARY SERVITUDE"
[53] "ARSON"
[54] "DISTURBING THE PEACE"
[55] "BATTERY POLICE (SIMPLE)"
[56] "THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)"
[57] "CONTRIBUTING"
[58] "EXTORTION"
```

```
[59] "PURSE SNATCHING"
[60] "THROWING OBJECT AT MOVING VEHICLE"
[61] "SHOPLIFTING - PETTY THEFT ($950 & UNDER)"
[62] "HUMAN TRAFFICKING - COMMERCIAL SEX ACTS"
[63] "SHOTS FIRED AT INHABITED DWELLING"
[64] "LEWD/LASCIVIOUS ACTS WITH CHILD"
[65] "CHILD ABUSE (PHYSICAL) - AGGRAVATED ASSAULT"
[66] "CRM AGNST CHLD (13 OR UNDER) (14-15 & SUSP 10 YRS OLDER)"
[67] "PIMPING"
[68] "SHOTS FIRED AT MOVING VEHICLE, TRAIN OR AIRCRAFT"
[69] "DISCHARGE FIREARMS/SHOTS FIRED"
[70] "CONSPIRACY"
[71] "CRUELTY TO ANIMALS"
[72] "BUNCO, GRAND THEFT"
[73] "PROWLER"
[74] "LEWD CONDUCT"
[75] "TELEPHONE PROPERTY - DAMAGE"
[76] "CHILD PORNOGRAPHY"
[77] "FAILURE TO YIELD"
[78] "FALSE POLICE REPORT"
```

I will now filter the data set by crime committed. I want to focus on cases of domestic violence, so I will be filtering by crimes that start with "INTIMATE".

```
domestic_violence <- la_data |> filter(grepl("INTIMATE", crm_cd_desc, ignore.case = TRUE))
```

Joining The Data Sets

I will now join the data sets by date. I will join by my domestic violence data set column "date occured (date_occ)" and football data set column "gameday". Before joining, I need to ensure the columns are the right data type.

```
# Observing the format of date occured column in domestic violence data set
head(domestic_violence$date_occ)
```

```
[1] "05/10/2020 12:00:00 AM" "08/04/2020 12:00:00 AM" "10/04/2020 12:00:00 AM"
[4] "02/11/2020 12:00:00 AM" "08/30/2020 12:00:00 AM" "11/10/2020 12:00:00 AM"
```

```
# Convert to Date by specifying format and removing time
domestic_violence$date_occ <-
  as.Date(domestic_violence$date_occ, format = "%m/%d/%Y %I:%M:%S %p")
```

```
# Observing format of gameday column in football data set
head(game_days$gameday)
```

```
[1] "2020-09-13" "2020-09-20" "2020-09-27" "2020-10-04" "2020-10-25"
[6] "2020-10-26"
```

```
class(game_days$gameday)
```

```
[1] "character"
```

```
# Convert to Date type
game_days$gameday <- as.Date(game_days$gameday)
class(game_days$gameday)
```

```
[1] "Date"
```

Now that my data types are adjusted, I will join them.

```
# Joining data sets
dv_with_games <- domestic_violence |>
  left_join(game_days, by = c("date_occ" = "gameday")) |>
  mutate(game_day = ifelse(is.na(game_day), 0, game_day))
```

EDA

Now I will conduct EDA on the joined data set.

Q1: What's in my data set?

```
str(dv_with_games)
```

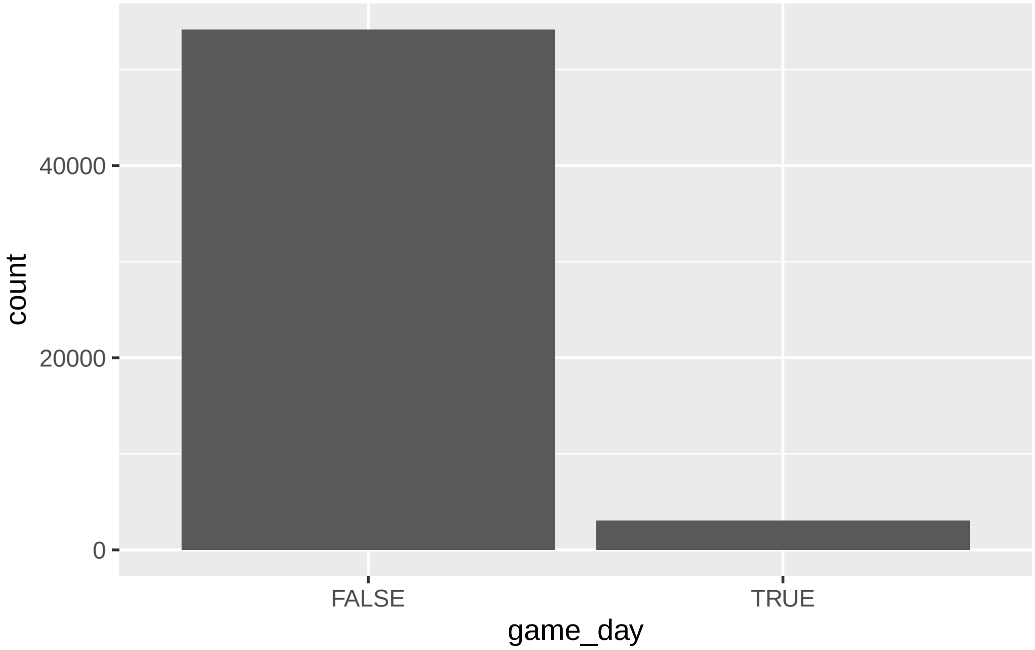
```
tibble [57,202 x 29] (S3: tbl_df/tbl/data.frame)
 $ dr_no         : num [1:57202] 2.00e+08 2.01e+08 2.01e+08 2.00e+08 2.01e+08 ...
 $ date_rptd     : chr [1:57202] "05/12/2020 12:00:00 AM" "08/04/2020 12:00:00 AM" "10/04/20
 $ date_occ      : Date[1:57202], format: "2020-05-10" "2020-08-04" ...
 $ time_occ      : chr [1:57202] "2200" "1000" "1800" "1915" ...
 $ area          : chr [1:57202] "01" "12" "12" "01" ...
 $ area_name     : chr [1:57202] "Central" "77th Street" "77th Street" "Central" ...
 $ rpt_dist_no   : chr [1:57202] "0111" "1249" "1268" "0154" ...
 $ part_1_2      : num [1:57202] 2 2 2 2 2 2 2 2 2 2 ...
 $ crm_cd        : num [1:57202] 626 626 626 626 626 626 626 626 626 626 ...
 $ crm_cd_desc   : chr [1:57202] "INTIMATE PARTNER - SIMPLE ASSAULT" "INTIMATE PARTNER - SIM
 $ mocodes       : chr [1:57202] "2000 0416 0913" "0400 0416 2000 1814 0913" "0913 0400 0416
 $ vict_age      : num [1:57202] 30 31 32 34 23 29 48 38 27 55 ...
 $ vict_sex      : chr [1:57202] "F" "F" "F" "M" ...
 $ vict_descent  : chr [1:57202] "W" "H" "H" "O" ...
 $ premis_cd     : num [1:57202] 502 501 101 502 501 122 101 501 501 517 ...
 $ premis_desc   : chr [1:57202] "MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC)" "SINGLE FAMI
 $ weapon_used_cd: num [1:57202] 400 400 400 400 400 400 400 400 400 400 ...
 $ weapon_desc   : chr [1:57202] "STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)" "STRONG-AR
 $ status        : chr [1:57202] "AA" "AO" "IC" "IC" ...
 $ status_desc   : chr [1:57202] "Adult Arrest" "Adult Other" "Invest Cont" "Invest Cont" ..
 $ crm_cd_1      : num [1:57202] 626 626 626 626 626 626 626 626 626 626 ...
 $ crm_cd_2      : num [1:57202] 998 NA NA NA NA 740 NA NA NA NA ...
 $ crm_cd_3      : num [1:57202] NA NA NA NA NA NA NA NA NA NA ...
 $ crm_cd_4      : logi [1:57202] NA NA NA NA NA NA ...
 $ location      : chr [1:57202] "700 N  HILL                         PL" "6700 S  FIGUEROA
 $ cross_street  : chr [1:57202] NA NA NA NA ...
 $ lat           : num [1:57202] 34.1 34 34 34 34.2 ...
 $ lon           : num [1:57202] -118 -118 -118 -118 -118 ...
 $ game_day      : num [1:57202] 0 0 1 0 0 0 0 0 0 0 ...
```

I have a joined data set with character, date, logical, and int data types.

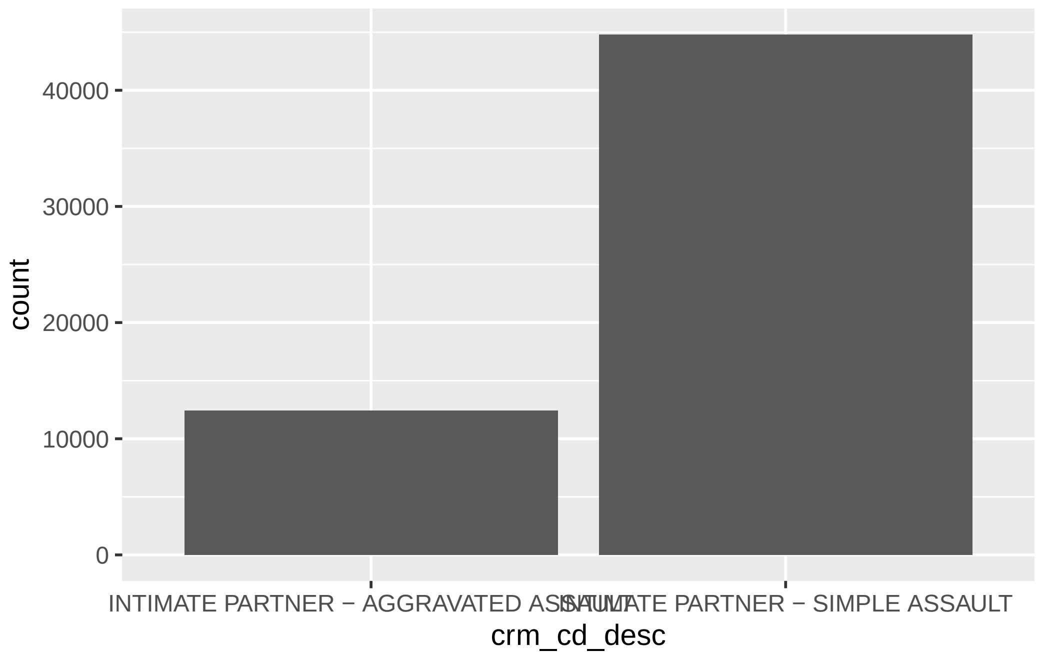
Q2: What type of variation occurs within my variables?

Here I'll visualize some key variables we are looking at

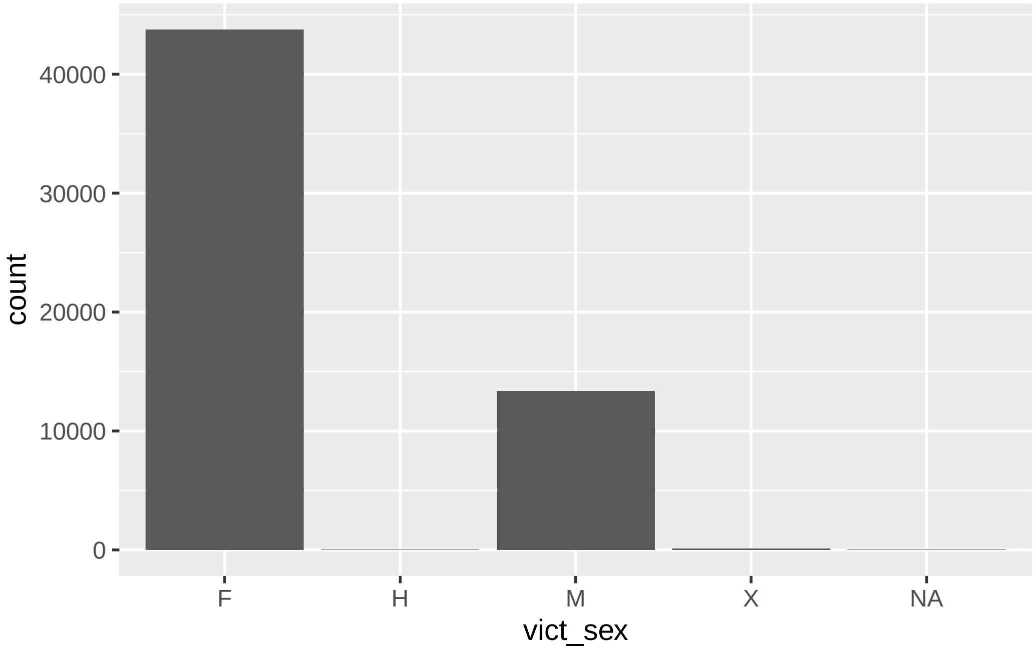
```
# Game_day variable
dv_with_games$game_day <- as.logical(dv_with_games$game_day)
ggplot(data = dv_with_games) +
  geom_bar(mapping = aes(x = game_day))
```



```r
# Crime Description variable
ggplot(data = dv_with_games) +
  geom_bar(mapping = aes(x = crm_cd_desc))
```



```
# Victim's sex variable
ggplot(data = dv_with_games) +
  geom_bar(mapping = aes(x = vict_sex))
```



In my next chunk I created another data frame to easily asses the amount of domestic violence incidents, the date, and whether or not it was a game day or not.

```
# Creating data set of count of dv reports, grouped by date
dv_daily <- domestic_violence |>
  group_by(date_occ) |>
  summarize(dv_reports = n())
# Joining data set into our original big data set dv_with_games
dv_with_games <- dv_with_games |>
  group_by(date_occ) |>
  mutate(dv_reports = n())
```

Q3: Am I missing any data?

```
summary(dv_with_games)
```

```
    dr_no             date_rptd            date_occ              time_occ
 Min.   :190101087   Length:57202       Min.   :2020-01-01   Length:57202
 1st Qu.:210217332   Class :character   1st Qu.:2021-02-01   Class :character
 Median :220401014   Mode  :character   Median :2022-03-01   Mode  :character
 Mean   :217964779                      Mean   :2022-02-22
 3rd Qu.:230506618                      3rd Qu.:2023-03-20
```

```
Max.   :252004209                      Max.   :2024-12-28

    area            area_name         rpt_dist_no         part_1_2
Length:57202       Length:57202       Length:57202       Min.   :1.000
Class :character   Class :character   Class :character   1st Qu.:2.000
Mode  :character   Mode  :character   Mode  :character   Median :2.000
                                                         Mean   :1.783
                                                         3rd Qu.:2.000
                                                         Max.   :2.000

    crm_cd      crm_cd_desc          mocodes             vict_age
Min.   :236.0   Length:57202       Length:57202       Min.   :-1.00
1st Qu.:626.0   Class :character   Class :character   1st Qu.:26.00
Median :626.0   Mode  :character   Mode  :character   Median :32.00
Mean   :541.4                                         Mean   :34.44
3rd Qu.:626.0                                         3rd Qu.:41.00
Max.   :626.0                                         Max.   :99.00

  vict_sex         vict_descent         premis_cd     premis_desc
Length:57202       Length:57202       Min.   :101.0   Length:57202
Class :character   Class :character   1st Qu.:158.0   Class :character
Mode  :character   Mode  :character   Median :501.0   Mode  :character
                                      Mean   :400.8
                                      3rd Qu.:502.0
                                      Max.   :971.0

weapon_used_cd weapon_desc          status          status_desc
Min.   :101    Length:57202       Length:57202       Length:57202
1st Qu.:400    Class :character   Class :character   Class :character
Median :400    Mode  :character   Mode  :character   Mode  :character
Mean   :393
3rd Qu.:400
Max.   :516
NA's   :100
   crm_cd_1        crm_cd_2        crm_cd_3     crm_cd_4
Min.   :236.0   Min.   :310.0   Min.   :626.0   Mode:logical
1st Qu.:626.0   1st Qu.:998.0   1st Qu.:981.0   NA's:57202
Median :626.0   Median :998.0   Median :998.0
Mean   :541.3   Mean   :964.7   Mean   :970.7
3rd Qu.:626.0   3rd Qu.:998.0   3rd Qu.:998.0
Max.   :626.0   Max.   :999.0   Max.   :999.0
                NA's   :53909   NA's   :57134
  location         cross_street            lat             lon
```

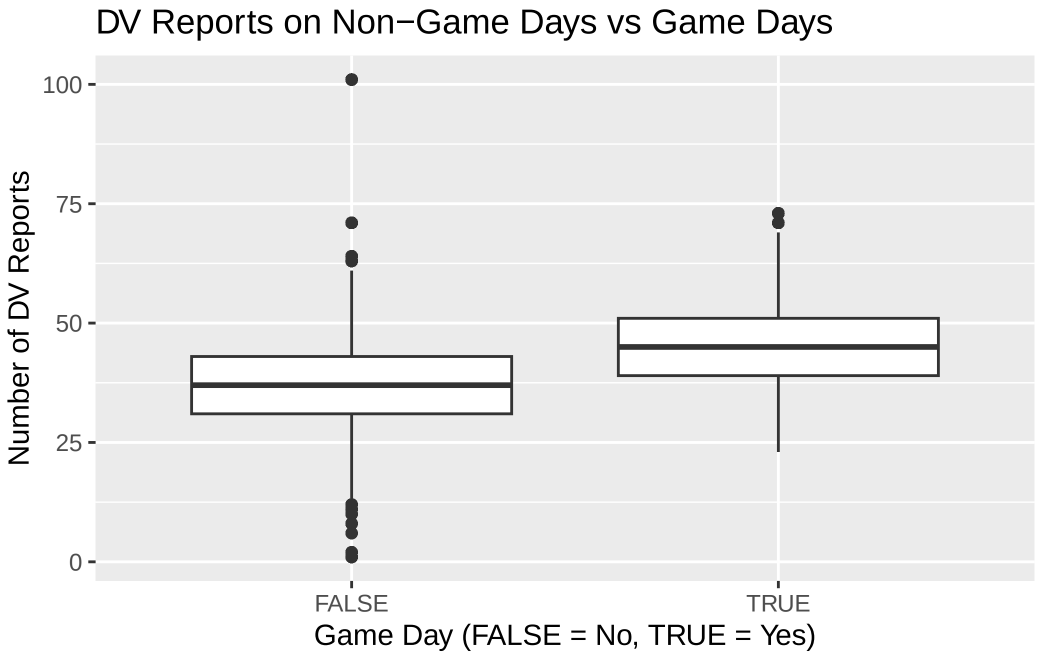
```
Length:57202       Length:57202       Min.   : 0.00   Min.   :-118.7  
Class :character   Class :character   1st Qu.:33.99   1st Qu.:-118.4  
Mode  :character   Mode  :character   Median :34.05   Median :-118.3  
                                      Mean   :33.94   Mean   :-117.9  
                                      3rd Qu.:34.18   3rd Qu.:-118.3  
                                      Max.   :34.33   Max.   :   0.0  

 game_day         dv_reports    
Mode :logical   Min.   :  1.00  
FALSE:54153     1st Qu.: 31.00  
TRUE :3049      Median : 37.00  
                Mean   : 38.24  
                3rd Qu.: 44.00  
                Max.   :101.00  
```

I am missing data in columns weapon_used_cd, crm_cd_2, crm_cd_3, crm_cd_4

Q4: What type of covariation occurs within my variables?

```r
ggplot(data = dv_with_games, mapping= aes(x = game_day, y = dv_reports)) +
  geom_boxplot() +
  labs(x = "Game Day (FALSE = No, TRUE = Yes)", y = "Number of DV Reports",
       title = "DV Reports on Non-Game Days vs Game Days")
```



Big Questions for our Project

What question is the project is trying to answer?

Do NFL games at SoFi Stadium have a measurable effect on the frequency of reported domestic violence incidents in Los Angeles? The project is trying to explore whether the is a statistically significant and potentially predictable relationship between spikes in domestic violence cases and NFL game days.

How have people answered it / gotten around it before?

Previously many have relied on using summary stats, like comparing average DV reports on game days vs. non-game days. Most studies seem to want to answer the question "does the outcome of NFL games correlate with spikes in domestic abuse?", using methods such as regression. However they often don't use cross-validation and don't explore unsupervised techniques such as clustering

What new idea does this project offer that improves on the old way of doing things?

This project will use regression, but with further use cross-validation and model selection to identify the most predictive and robust models. Time permitted we will also consider unsupervised learning to detect patterns in the data (DV patterns on game days). Also we are narrowing the geographic focus to Los Angeles across multiple seasons (2020-2023), instead of the entire US or world

What are the (major) building blocks the project will need to be successful?

The major building blocks I see are completing the cross-validation and model selection. If we're able to possibly clustering.

Which ones are in place already, and which ones are still under construction or TBD?

Joining is already taking place, and the methods that we will create to finish our project are still under construction