NYPD Shooting Incidents: A cleaning project.

J. Valencia

2024-11-06

Objective

This project provides an overview of the fundamental stages of a data science process, offering insight into the initial steps, which include:

- Data summarisation to gain a deeper understanding of the data type in question
- Data transformation to ensure compatibility with the desired data type
- Data cleansing and transformation to facilitate the generation of meaningful insights.

Process

Import data

The initial step is to accurately set the data into a variable that can be used at a later stage. The data was sourced from https://data.gov and is publicly accessible and free to use.

```
nypd_shooting <- read_csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv")
```

```
## Rows: 28562 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD...
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Be aware of you data.

Once the data has been imported, it is important to be aware of the data types in order to perform the initial transformations. Let us now review the structure of the data.

```
str(nypd_shooting)
```

```
## spc_tbl_ [28,562 x 21] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ INCIDENT_KEY           : num [1:28562] 2.32e+08 1.78e+08 2.55e+08 2.54e+07 7.26e+07 ...
```

```
##  $ OCCUR_DATE             : chr [1:28562] "08/09/2021" "04/07/2018" "12/02/2022" "11/19/2006" ...
##  $ OCCUR_TIME             : 'hms' num [1:28562] 01:06:00 19:48:00 22:57:00 01:50:00 ...
##   ..- attr(*, "units")= chr "secs"
##  $ BORO                   : chr [1:28562] "BRONX" "BROOKLYN" "BRONX" "BROOKLYN" ...
##  $ LOC_OF_OCCUR_DESC      : chr [1:28562] NA NA "OUTSIDE" NA ...
##  $ PRECINCT               : num [1:28562] 40 79 47 66 46 42 71 69 75 69 ...
##  $ JURISDICTION_CODE      : num [1:28562] 0 0 0 0 0 2 0 2 0 0 ...
##  $ LOC_CLASSFCTN_DESC     : chr [1:28562] NA NA "STREET" NA ...
##  $ LOCATION_DESC          : chr [1:28562] NA NA "GROCERY/BODEGA" "PVT HOUSE" ...
##  $ STATISTICAL_MURDER_FLAG: logi [1:28562] FALSE TRUE FALSE TRUE TRUE FALSE ...
##  $ PERP_AGE_GROUP         : chr [1:28562] NA "25-44" "(null)" "UNKNOWN" ...
##  $ PERP_SEX               : chr [1:28562] NA "M" "(null)" "U" ...
##  $ PERP_RACE              : chr [1:28562] NA "WHITE HISPANIC" "(null)" "UNKNOWN" ...
##  $ VIC_AGE_GROUP          : chr [1:28562] "18-24" "25-44" "25-44" "18-24" ...
##  $ VIC_SEX                : chr [1:28562] "M" "M" "M" "M" ...
##  $ VIC_RACE               : chr [1:28562] "BLACK" "BLACK" "BLACK" "BLACK" ...
##  $ X_COORD_CD             : num [1:28562] 1006343 1000083 1020691 985107 1009854 ...
##  $ Y_COORD_CD             : num [1:28562] 234270 189065 257125 173350 247503 ...
##  $ Latitude               : num [1:28562] 40.8 40.7 40.9 40.6 40.8 ...
##  $ Longitude              : num [1:28562] -73.9 -73.9 -73.9 -74 -73.9 ...
##  $ Lon_Lat                : chr [1:28562] "POINT (-73.92019278899994 40.80967347200004)" "POINT (-73
##  - attr(*, "spec")=
##   .. cols(
##   ..   INCIDENT_KEY = col_double(),
##   ..   OCCUR_DATE = col_character(),
##   ..   OCCUR_TIME = col_time(format = ""),
##   ..   BORO = col_character(),
##   ..   LOC_OF_OCCUR_DESC = col_character(),
##   ..   PRECINCT = col_double(),
##   ..   JURISDICTION_CODE = col_double(),
##   ..   LOC_CLASSFCTN_DESC = col_character(),
##   ..   LOCATION_DESC = col_character(),
##   ..   STATISTICAL_MURDER_FLAG = col_logical(),
##   ..   PERP_AGE_GROUP = col_character(),
##   ..   PERP_SEX = col_character(),
##   ..   PERP_RACE = col_character(),
##   ..   VIC_AGE_GROUP = col_character(),
##   ..   VIC_SEX = col_character(),
##   ..   VIC_RACE = col_character(),
##   ..   X_COORD_CD = col_double(),
##   ..   Y_COORD_CD = col_double(),
##   ..   Latitude = col_double(),
##   ..   Longitude = col_double(),
##   ..   Lon_Lat = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

In relation to the aforementioned structure, we have identified the most appropriate data type for analysis, as outlined below:

- Factor: INCIDENT_KEY, BORO, LOC_OF_OCCUR_DESC, PRECINCT, JURISDICTION_CODE, LOC_CLASSFCTN_DESC, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_RACE, VIC_AGE_GROUP, VIC_SEX, VIC_RACE, Lon_Lat
- Date: OCCUR_DATE

- Time: OCCUR_TIME
- Logic: STATISTICAL_MURDER_FLAG
- Numeric: X_COORD_CD, Y_COORD_CD, Latitude, Longitud,

The results indicate that the majority of columns are of the factor type.

Then, we proceed to inspect a few observations

```
head(nypd_shooting)
```

```
## # A tibble: 6 x 21
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO     LOC_OF_OCCUR_DESC PRECINCT
##          <dbl> <chr>      <time>     <chr>    <chr>                <dbl>
## 1    231974218 08/09/2021 01:06      BRONX    <NA>                    40
## 2    177934247 04/07/2018 19:48      BROOKLYN <NA>                    79
## 3    255028563 12/02/2022 22:57      BRONX    OUTSIDE                 47
## 4     25384540 11/19/2006 01:50      BROOKLYN <NA>                    66
## 5     72616285 05/09/2010 01:58      BRONX    <NA>                    46
## 6     85875439 07/22/2012 21:35      BRONX    <NA>                    42
## # i 15 more variables: JURISDICTION_CODE <dbl>, LOC_CLASSFCTN_DESC <chr>,
## #   LOCATION_DESC <chr>, STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>,
## #   PERP_SEX <chr>, PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>,
## #   VIC_RACE <chr>, X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>,
## #   Longitude <dbl>, Lon_Lat <chr>
```

It is important to note that the data frame handles missing data differently. For some columns, the value is "NA". For others, it is "(null)", and for a few, a combination of both is used.

Following the initial insight, we then proceeded to data type conversion and consistent handling of missing values.

```
#Change to factors
factor_cols <- c("INCIDENT_KEY", "BORO", "LOC_OF_OCCUR_DESC", "PRECINCT", "JURISDICTION_CODE", "LOC_CLAS
nypd_shooting[, factor_cols] <- lapply(nypd_shooting[, factor_cols], factor)

#Change to dates
nypd_shooting[["OCCUR_DATE"]] <- mdy(nypd_shooting[["OCCUR_DATE"]])

#Change to time
nypd_shooting[["OCCUR_TIME"]] <- hms(nypd_shooting[["OCCUR_TIME"]])

#Numeric data are already in proper format.

summary(nypd_shooting)
```

```
##     INCIDENT_KEY     OCCUR_DATE           OCCUR_TIME
##  173354054:   18   Min.   :2006-01-01   Min.   :0S
##  263503175:   16   1st Qu.:2009-09-04   1st Qu.:3H 30M 0S
##  23749375 :   12   Median :2013-09-20   Median :15H 15M 0S
##  24717013 :   12   Mean   :2014-06-07   Mean   :12H 44M 16.7131153281152S
##  33478089 :   12   3rd Qu.:2019-09-29   3rd Qu.:20H 45M 0S
##  33706902 :   12   Max.   :2023-12-29   Max.   :23H 59M 0S
##  (Other)  :28480
```

```
##             BORO       LOC_OF_OCCUR_DESC    PRECINCT     JURISDICTION_CODE
##  BRONX        : 8376   INSIDE :  460     75     : 1628   0   :23923
##  BROOKLYN     :11346   OUTSIDE: 2506     73     : 1500   1   :   81
##  MANHATTAN    : 3762   NA's   :25596     67     : 1259   2   : 4556
##  QUEENS       : 4271                     44     : 1076   NA's:    2
##  STATEN ISLAND:  807                     79     : 1045
##                                          47     : 1006
##                                          (Other):21048
##   LOC_CLASSFCTN_DESC                  LOCATION_DESC   STATISTICAL_MURDER_FLAG
##  STREET    : 1886    MULTI DWELL - PUBLIC HOUS: 5007   Mode :logical
##  HOUSING   :  460    MULTI DWELL - APT BUILD  : 2964   FALSE:23036
##  DWELLING  :  243    (null)                   : 1711   TRUE :5526
##  COMMERCIAL:  208    PVT HOUSE                :  983
##  OTHER     :   59    GROCERY/BODEGA           :  750
##  (Other)   :  110    (Other)                  : 2170
##  NA's      :25596    NA's                     :14977
##  PERP_AGE_GROUP   PERP_SEX              PERP_RACE     VIC_AGE_GROUP   VIC_SEX
##  18-24  :6438   (null): 1141   BLACK         :11903   <18    : 2954   F: 2760
##  25-44  :6041   F     :  444   WHITE HISPANIC: 2510   1022   :    1   M:25790
##  UNKNOWN:3148   M     :16168   UNKNOWN       : 1837   18-24  :10384   U:   12
##  <18    :1682   U     : 1499   BLACK HISPANIC: 1392   25-44  :12973
##  (null) :1141   NA's  : 9310   (null)        : 1141   45-64  : 1981
##  (Other): 768                  (Other)       :  469   65+    :  205
##  NA's   :9344                  NA's          : 9310   UNKNOWN:   64
##                            VIC_RACE       X_COORD_CD        Y_COORD_CD
##  AMERICAN INDIAN/ALASKAN NATIVE:   11   Min.   : 914928   Min.   :125757
##  ASIAN / PACIFIC ISLANDER      :  440   1st Qu.:1000068   1st Qu.:182912
##  BLACK                         :20235   Median :1007772   Median :194901
##  BLACK HISPANIC                : 2795   Mean   :1009424   Mean   :208380
##  UNKNOWN                       :   70   3rd Qu.:1016807   3rd Qu.:239814
##  WHITE                         :  728   Max.   :1066815   Max.   :271128
##  WHITE HISPANIC                : 4283
##     Latitude       Longitude
##  Min.   :40.51   Min.   :-74.25
##  1st Qu.:40.67   1st Qu.:-73.94
##  Median :40.70   Median :-73.92
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##  NA's   :59      NA's   :59
##                                          Lon_Lat
##  POINT (-73.88151014499994 40.67141260500006) :   66
##  POINT (-73.84760778699996 40.88745131300004) :   47
##  POINT (-73.91339091999998 40.670655072000045):   47
##  POINT (-73.88143295699996 40.67110691100004) :   44
##  POINT (-74.17125343299995 40.63898537500006) :   44
##  (Other)                                      :28255
##  NA's                                         :   59
```

```
str(nypd_shooting)
```

```
## spc_tbl_ [28,562 x 21] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ INCIDENT_KEY           : Factor w/ 22394 levels "9953245","9953246",..: 19506 15647 21338 1381 63
##  $ OCCUR_DATE             : Date[1:28562], format: "2021-08-09" "2018-04-07" ...
```

```
##  $ OCCUR_TIME             :Formal class 'Period' [package "lubridate"] with 6 slots
##   .. ..@ .Data : num [1:28562] 0 0 0 0 0 0 0 0 0 0 ...
##   .. ..@ year  : num [1:28562] 0 0 0 0 0 0 0 0 0 0 ...
##   .. ..@ month : num [1:28562] 0 0 0 0 0 0 0 0 0 0 ...
##   .. ..@ day   : num [1:28562] 0 0 0 0 0 0 0 0 0 0 ...
##   .. ..@ hour  : num [1:28562] 1 19 22 1 1 21 22 23 15 15 ...
##   .. ..@ minute: num [1:28562] 6 48 57 50 58 35 26 45 36 23 ...
##  $ BORO                   : Factor w/ 5 levels "BRONX","BROOKLYN",..: 1 2 1 2 1 1 2 2 2 2 ...
##  $ LOC_OF_OCCUR_DESC      : Factor w/ 2 levels "INSIDE","OUTSIDE": NA NA 2 NA NA NA NA NA NA NA ...
##  $ PRECINCT               : Factor w/ 77 levels "1","5","6","7",..: 23 51 30 39 29 25 44 42 47 42 ..
##  $ JURISDICTION_CODE      : Factor w/ 3 levels "0","1","2": 1 1 1 1 1 3 1 3 1 1 ...
##  $ LOC_CLASSFCTN_DESC     : Factor w/ 10 levels "(null)","COMMERCIAL",..: NA NA 8 NA NA NA NA NA NA N
##  $ LOCATION_DESC          : Factor w/ 40 levels "(null)","ATM",..: NA NA 18 29 25 26 NA 26 25 NA ...
##  $ STATISTICAL_MURDER_FLAG: logi [1:28562] FALSE TRUE FALSE TRUE TRUE FALSE ...
##  $ PERP_AGE_GROUP         : Factor w/ 11 levels "(null)","<18",..: NA 7 1 11 7 5 NA NA 7 5 ...
##  $ PERP_SEX               : Factor w/ 4 levels "(null)","F","M",..: NA 3 1 4 3 3 NA NA 3 3 ...
##  $ PERP_RACE              : Factor w/ 8 levels "(null)","AMERICAN INDIAN/ALASKAN NATIVE",..: NA 8 1 6
##  $ VIC_AGE_GROUP          : Factor w/ 7 levels "<18","1022","18-24",..: 3 4 4 3 1 3 4 4 4 3 ...
##  $ VIC_SEX                : Factor w/ 3 levels "F","M","U": 2 2 2 2 1 2 2 2 2 2 ...
##  $ VIC_RACE               : Factor w/ 7 levels "AMERICAN INDIAN/ALASKAN NATIVE",..: 3 3 3 3 3 3 3 7 3
##  $ X_COORD_CD             : num [1:28562] 1006343 1000083 1020691 985107 1009854 ...
##  $ Y_COORD_CD             : num [1:28562] 234270 189065 257125 173350 247503 ...
##  $ Latitude               : num [1:28562] 40.8 40.7 40.9 40.6 40.8 ...
##  $ Longitude              : num [1:28562] -73.9 -73.9 -73.9 -74 -73.9 ...
##  $ Lon_Lat                : Factor w/ 13403 levels "POINT (-73.70204616699993 40.74174860900007)",..
##  - attr(*, "spec")=
##   .. cols(
##   ..   INCIDENT_KEY = col_double(),
##   ..   OCCUR_DATE = col_character(),
##   ..   OCCUR_TIME = col_time(format = ""),
##   ..   BORO = col_character(),
##   ..   LOC_OF_OCCUR_DESC = col_character(),
##   ..   PRECINCT = col_double(),
##   ..   JURISDICTION_CODE = col_double(),
##   ..   LOC_CLASSFCTN_DESC = col_character(),
##   ..   LOCATION_DESC = col_character(),
##   ..   STATISTICAL_MURDER_FLAG = col_logical(),
##   ..   PERP_AGE_GROUP = col_character(),
##   ..   PERP_SEX = col_character(),
##   ..   PERP_RACE = col_character(),
##   ..   VIC_AGE_GROUP = col_character(),
##   ..   VIC_SEX = col_character(),
##   ..   VIC_RACE = col_character(),
##   ..   X_COORD_CD = col_double(),
##   ..   Y_COORD_CD = col_double(),
##   ..   Latitude = col_double(),
##   ..   Longitude = col_double(),
##   ..   Lon_Lat = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Once the factors have been identified, we will proceed to inspect the levels of the column in order to search for any unintended values.

```r
#Inspect factor levels
factor_cols_w_lev <- c("BORO", "LOC_OF_OCCUR_DESC", "PRECINCT", "JURISDICTION_CODE", "LOC_CLASSFCTN_DESC
lapply(nypd_shooting[, factor_cols_w_lev], levels)
```

```
## $BORO
## [1] "BRONX"         "BROOKLYN"      "MANHATTAN"     "QUEENS"
## [5] "STATEN ISLAND"
##
## $LOC_OF_OCCUR_DESC
## [1] "INSIDE"  "OUTSIDE"
##
## $PRECINCT
##  [1] "1"   "5"   "6"   "7"   "9"   "10"  "13"  "14"  "17"  "18"  "19"  "20"
## [13] "22"  "23"  "24"  "25"  "26"  "28"  "30"  "32"  "33"  "34"  "40"  "41"
## [25] "42"  "43"  "44"  "45"  "46"  "47"  "48"  "49"  "50"  "52"  "60"  "61"
## [37] "62"  "63"  "66"  "67"  "68"  "69"  "70"  "71"  "72"  "73"  "75"  "76"
## [49] "77"  "78"  "79"  "81"  "83"  "84"  "88"  "90"  "94"  "100" "101" "102"
## [61] "103" "104" "105" "106" "107" "108" "109" "110" "111" "112" "113" "114"
## [73] "115" "120" "121" "122" "123"
##
## $JURISDICTION_CODE
## [1] "0" "1" "2"
##
## $LOC_CLASSFCTN_DESC
##  [1] "(null)"      "COMMERCIAL"  "DWELLING"    "HOUSING"     "OTHER"
##  [6] "PARKING LOT" "PLAYGROUND"  "STREET"      "TRANSIT"     "VEHICLE"
##
## $LOCATION_DESC
##  [1] "(null)"                    "ATM"
##  [3] "BANK"                      "BAR/NIGHT CLUB"
##  [5] "BEAUTY/NAIL SALON"         "CANDY STORE"
##  [7] "CHAIN STORE"               "CHECK CASH"
##  [9] "CLOTHING BOUTIQUE"         "COMMERCIAL BLDG"
## [11] "DEPT STORE"                "DOCTOR/DENTIST"
## [13] "DRUG STORE"                "DRY CLEANER/LAUNDRY"
## [15] "FACTORY/WAREHOUSE"         "FAST FOOD"
## [17] "GAS STATION"               "GROCERY/BODEGA"
## [19] "GYM/FITNESS FACILITY"      "HOSPITAL"
## [21] "HOTEL/MOTEL"               "JEWELRY STORE"
## [23] "LIQUOR STORE"              "LOAN COMPANY"
## [25] "MULTI DWELL - APT BUILD"   "MULTI DWELL - PUBLIC HOUS"
## [27] "NONE"                      "PHOTO/COPY STORE"
## [29] "PVT HOUSE"                 "RESTAURANT/DINER"
## [31] "SCHOOL"                    "SHOE STORE"
## [33] "SMALL MERCHANT"            "SOCIAL CLUB/POLICY LOCATI"
## [35] "STORAGE FACILITY"          "STORE UNCLASSIFIED"
## [37] "SUPERMARKET"               "TELECOMM. STORE"
## [39] "VARIETY STORE"             "VIDEO STORE"
##
## $PERP_AGE_GROUP
##  [1] "(null)"  "<18"     "1020"    "1028"    "18-24"   "224"     "25-44"
##  [8] "45-64"   "65+"     "940"     "UNKNOWN"
##
```

```
## $PERP_SEX
## [1] "(null)" "F"      "M"      "U"
##
## $PERP_RACE
## [1] "(null)"                          "AMERICAN INDIAN/ALASKAN NATIVE"
## [3] "ASIAN / PACIFIC ISLANDER"        "BLACK"
## [5] "BLACK HISPANIC"                  "UNKNOWN"
## [7] "WHITE"                           "WHITE HISPANIC"
##
## $VIC_AGE_GROUP
## [1] "<18"     "1022"    "18-24"   "25-44"   "45-64"   "65+"     "UNKNOWN"
##
## $VIC_SEX
## [1] "F" "M" "U"
##
## $VIC_RACE
## [1] "AMERICAN INDIAN/ALASKAN NATIVE" "ASIAN / PACIFIC ISLANDER"
## [3] "BLACK"                          "BLACK HISPANIC"
## [5] "UNKNOWN"                        "WHITE"
## [7] "WHITE HISPANIC"
```

As a result, LOC_CLASSFCTN_DESC, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX and PERP_RACE uses ("null") can be used as a method to handle missing values, even in instances where the columns themselves already contain NA values. Also, AGE groups have strange values like "1020", "1028", " "224", "940", "UNKNOWN", and SEX and RACE groups also have "U" and "UNKNOWN" as a way to also handle null valueS. Let's change ("null"), "UNKNOWN" and "U" to a consistent format

```
#Inspect factor levels
factor_cols_null <- c("LOC_CLASSFCTN_DESC", "LOCATION_DESC", "PERP_AGE_GROUP", "PERP_SEX", "PERP_RACE",

for (col in factor_cols_null) {
  nypd_shooting[[col]] <- fct_na_level_to_value(nypd_shooting[[col]], c("(null)", "U", "UNKNOWN"))
}

lapply(nypd_shooting[, factor_cols_w_lev], levels)
```

```
## $BORO
## [1] "BRONX"         "BROOKLYN"      "MANHATTAN"     "QUEENS"
## [5] "STATEN ISLAND"
##
## $LOC_OF_OCCUR_DESC
## [1] "INSIDE"  "OUTSIDE"
##
## $PRECINCT
##  [1] "1"   "5"   "6"   "7"   "9"   "10"  "13"  "14"  "17"  "18"  "19"  "20"
## [13] "22"  "23"  "24"  "25"  "26"  "28"  "30"  "32"  "33"  "34"  "40"  "41"
## [25] "42"  "43"  "44"  "45"  "46"  "47"  "48"  "49"  "50"  "52"  "60"  "61"
## [37] "62"  "63"  "66"  "67"  "68"  "69"  "70"  "71"  "72"  "73"  "75"  "76"
## [49] "77"  "78"  "79"  "81"  "83"  "84"  "88"  "90"  "94"  "100" "101" "102"
## [61] "103" "104" "105" "106" "107" "108" "109" "110" "111" "112" "113" "114"
## [73] "115" "120" "121" "122" "123"
##
## $JURISDICTION_CODE
```

```
## [1] "0" "1" "2"
##
## $LOC_CLASSFCTN_DESC
## [1] "COMMERCIAL"  "DWELLING"    "HOUSING"     "OTHER"       "PARKING LOT"
## [6] "PLAYGROUND"  "STREET"      "TRANSIT"     "VEHICLE"
##
## $LOCATION_DESC
##  [1] "ATM"                        "BANK"
##  [3] "BAR/NIGHT CLUB"             "BEAUTY/NAIL SALON"
##  [5] "CANDY STORE"                "CHAIN STORE"
##  [7] "CHECK CASH"                 "CLOTHING BOUTIQUE"
##  [9] "COMMERCIAL BLDG"            "DEPT STORE"
## [11] "DOCTOR/DENTIST"             "DRUG STORE"
## [13] "DRY CLEANER/LAUNDRY"        "FACTORY/WAREHOUSE"
## [15] "FAST FOOD"                  "GAS STATION"
## [17] "GROCERY/BODEGA"             "GYM/FITNESS FACILITY"
## [19] "HOSPITAL"                   "HOTEL/MOTEL"
## [21] "JEWELRY STORE"              "LIQUOR STORE"
## [23] "LOAN COMPANY"               "MULTI DWELL - APT BUILD"
## [25] "MULTI DWELL - PUBLIC HOUS" "NONE"
## [27] "PHOTO/COPY STORE"           "PVT HOUSE"
## [29] "RESTAURANT/DINER"           "SCHOOL"
## [31] "SHOE STORE"                 "SMALL MERCHANT"
## [33] "SOCIAL CLUB/POLICY LOCATI" "STORAGE FACILITY"
## [35] "STORE UNCLASSIFIED"         "SUPERMARKET"
## [37] "TELECOMM. STORE"            "VARIETY STORE"
## [39] "VIDEO STORE"
##
## $PERP_AGE_GROUP
## [1] "<18"   "1020"  "1028"  "18-24" "224"   "25-44" "45-64" "65+"   "940"
##
## $PERP_SEX
## [1] "F" "M"
##
## $PERP_RACE
## [1] "AMERICAN INDIAN/ALASKAN NATIVE" "ASIAN / PACIFIC ISLANDER"
## [3] "BLACK"                          "BLACK HISPANIC"
## [5] "WHITE"                          "WHITE HISPANIC"
##
## $VIC_AGE_GROUP
## [1] "<18"   "1022"  "18-24" "25-44" "45-64" "65+"
##
## $VIC_SEX
## [1] "F" "M"
##
## $VIC_RACE
## [1] "AMERICAN INDIAN/ALASKAN NATIVE" "ASIAN / PACIFIC ISLANDER"
## [3] "BLACK"                          "BLACK HISPANIC"
## [5] "WHITE"                          "WHITE HISPANIC"
```

Subsequently, we will examine the extent to which data within specified age groups exhibit anomalous values.

```r
to_exclude <- nypd_shooting %>%
filter(PERP_AGE_GROUP %in% c("1020", "1025", "224", "940") | VIC_AGE_GROUP == "1022")

to_exclude
```

```
## # A tibble: 4 x 21
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO      LOC_OF_OCCUR_DESC PRECINCT
##   <fct>        <date>     <Period>   <fct>     <fct>             <fct>
## 1 248480012    2022-07-23 16H 48M 0S MANHATTAN OUTSIDE           13
## 2 142247967    2015-04-19 2H 5M 0S   BRONX     <NA>              47
## 3 89595619     2013-03-12 20H 28M 0S BROOKLYN  <NA>              90
## 4 71625599     2010-03-06 4H 14M 0S  BRONX     <NA>              41
## # i 15 more variables: JURISDICTION_CODE <fct>, LOC_CLASSFCTN_DESC <fct>,
## #   LOCATION_DESC <fct>, STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <fct>,
## #   PERP_SEX <fct>, PERP_RACE <fct>, VIC_AGE_GROUP <fct>, VIC_SEX <fct>,
## #   VIC_RACE <fct>, X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>,
## #   Longitude <dbl>, Lon_Lat <fct>
```

Given the limited number of columns with anomalous values and the inability to infer which data should be replaced, the four observations in question will be deleted.

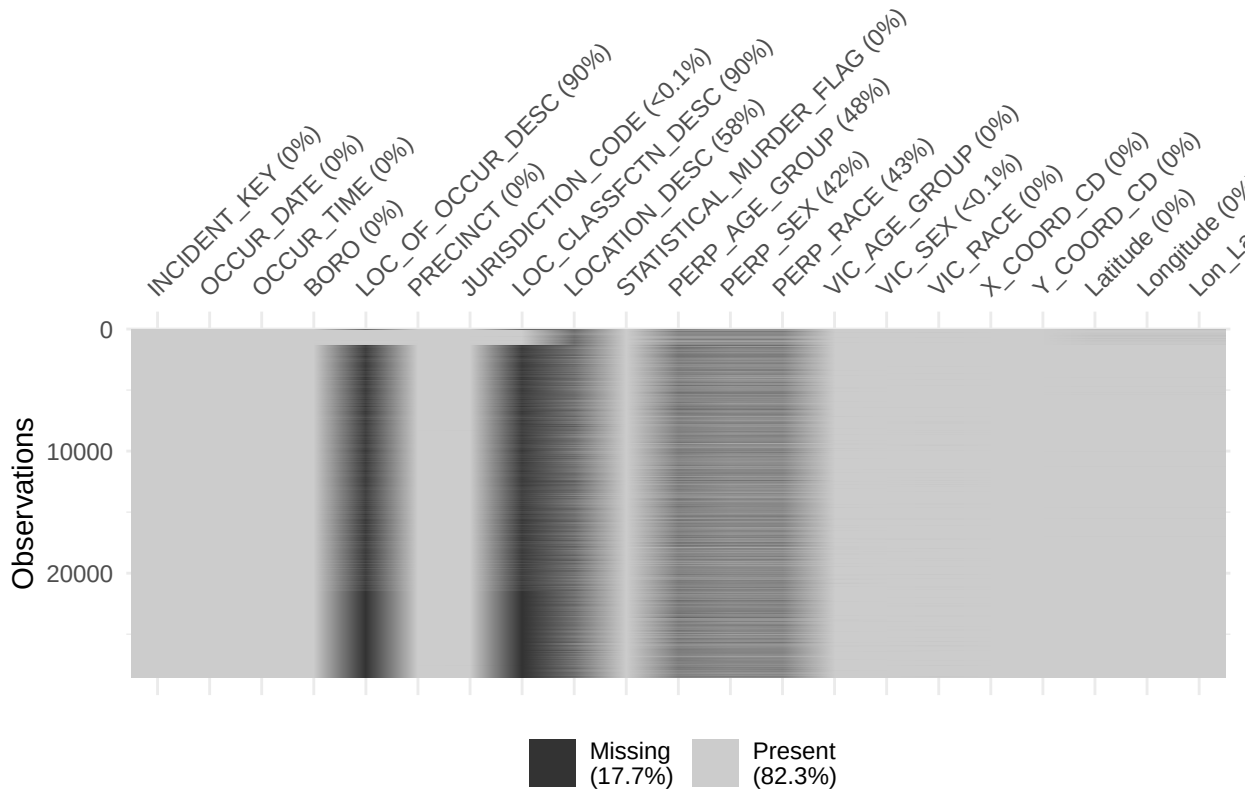
```r
to_exclude <- nypd_shooting %>%
  filter(PERP_AGE_GROUP %in% c("1020", "1025", "224", "940") | VIC_AGE_GROUP == "1022")

nypd_shooting <- nypd_shooting %>%
  anti_join(to_exclude, by = c("PERP_AGE_GROUP", "VIC_AGE_GROUP"))

summary(nypd_shooting)
```

```
##     INCIDENT_KEY     OCCUR_DATE           OCCUR_TIME
##  173354054:   18   Min.   :2006-01-01   Min.   :0S
##  263503175:   16   1st Qu.:2009-09-03   1st Qu.:3H 30M 0S
##  23749375 :   12   Median :2013-09-20   Median :15H 15M 0S
##  24717013 :   12   Mean   :2014-06-07   Mean   :12H 44M 17.64199173622915
##  33478089 :   12   3rd Qu.:2019-09-29   3rd Qu.:20H 45M 0S
##  33706902 :   12   Max.   :2023-12-29   Max.   :23H 59M 0S
##  (Other)  :28476
##             BORO       LOC_OF_OCCUR_DESC    PRECINCT     JURISDICTION_CODE
##  BRONX        : 8374   INSIDE :  460     75     : 1628   0   :23920
##  BROOKLYN     :11345   OUTSIDE: 2505     73     : 1500   1   :   81
##  MANHATTAN    : 3761   NA's   :25593     67     : 1259   2   : 4555
##  QUEENS       : 4271                     44     : 1076   NA's:    2
##  STATEN ISLAND:  807                     79     : 1045
##                                          47     : 1005
##                                          (Other):21045
##   LOC_CLASSFCTN_DESC                   LOCATION_DESC   STATISTICAL_MURDER_FLAG
##  STREET    : 1885    MULTI DWELL - PUBLIC HOUS: 5006   Mode :logical
##  HOUSING   :  460    MULTI DWELL - APT BUILD  : 2964   FALSE:23032
##  DWELLING  :  243    PVT HOUSE                :  983   TRUE :5526
##  COMMERCIAL:  208    GROCERY/BODEGA           :  750
##  OTHER     :   59    BAR/NIGHT CLUB           :  667
##  (Other)   :  108    (Other)                  : 1501
```

```
##  NA's      :25595    NA's                     :16687
##  PERP_AGE_GROUP  PERP_SEX                              PERP_RACE
##  18-24  : 6437   F   :  444   AMERICAN INDIAN/ALASKAN NATIVE:    2
##  25-44  : 6041   M   :16164   ASIAN / PACIFIC ISLANDER      :  169
##  <18    : 1682   NA's:11950   BLACK                         :11901
##  45-64  :  699                BLACK HISPANIC                : 1392
##  65+    :   65                WHITE                         :  298
##  (Other):    1                WHITE HISPANIC                : 2508
##  NA's   :13633                NA's                          :12288
##  VIC_AGE_GROUP VIC_SEX                                VIC_RACE
##  <18  : 2954   F   : 2760   AMERICAN INDIAN/ALASKAN NATIVE:   11
##  1022 :    0   M   :25786   ASIAN / PACIFIC ISLANDER      :  440
##  18-24:10383   NA's:   12   BLACK                         :20233
##  25-44:12971                BLACK HISPANIC                : 2795
##  45-64: 1981                WHITE                         :  728
##  65+  :  205                WHITE HISPANIC                : 4281
##  NA's :   64                NA's                          :   70
##    X_COORD_CD        Y_COORD_CD        Latitude       Longitude
##  Min.   : 914928   Min.   :125757   Min.   :40.51   Min.   :-74.25
##  1st Qu.:1000068   1st Qu.:182907   1st Qu.:40.67   1st Qu.:-73.94
##  Median :1007772   Median :194887   Median :40.70   Median :-73.92
##  Mean   :1009425   Mean   :208378   Mean   :40.74   Mean   :-73.91
##  3rd Qu.:1016807   3rd Qu.:239814   3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :1066815   Max.   :271128   Max.   :40.91   Max.   :-73.70
##                                     NA's   :59      NA's   :59
##                                           Lon_Lat
##  POINT (-73.88151014499994 40.67141260500006) :   66
##  POINT (-73.84760778699996 40.88745131300004) :   47
##  POINT (-73.91339091999998 40.670655072000045):   47
##  POINT (-73.88143295699996 40.67110691100004) :   44
##  POINT (-74.17125343299995 40.63898537500006) :   44
##  (Other)                                      :28251
##  NA's                                         :   59
```

Finally, lets inspect the ammount of "NA" values which can bias our data.

```
vis_miss(nypd_shooting)
```



As can be seen, the following variables contain over 5% missing values: LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, and PERP_RACE. While it is important to gain insights from the data we have, it is not advisable to make decisions based solely on these variables.

It would be beneficial to investigate further the missing data across the different columns to see if we can gain some insight.

```
na_loc_occur <- nypd_shooting %>%
                  mutate(loc_occur_missing = is.na(LOC_OF_OCCUR_DESC)) %>%
                  group_by(loc_occur_missing) %>%
                  summarize(across(everything(),
                   ~ if(is.factor(.)) {
                     # For factors, return the most frequent level (mode)
                     as.character(names(sort(table(.), decreasing = TRUE))[1])
                   } else if(is.numeric(.)) {
                     # For numeric variables, return the median
                     median(., na.rm = TRUE)
                   } else {
                     # For other types (e.g., character), return the first value
                     first(.)
                   }))

na_loc_occur
```

```
## # A tibble: 2 x 22
```

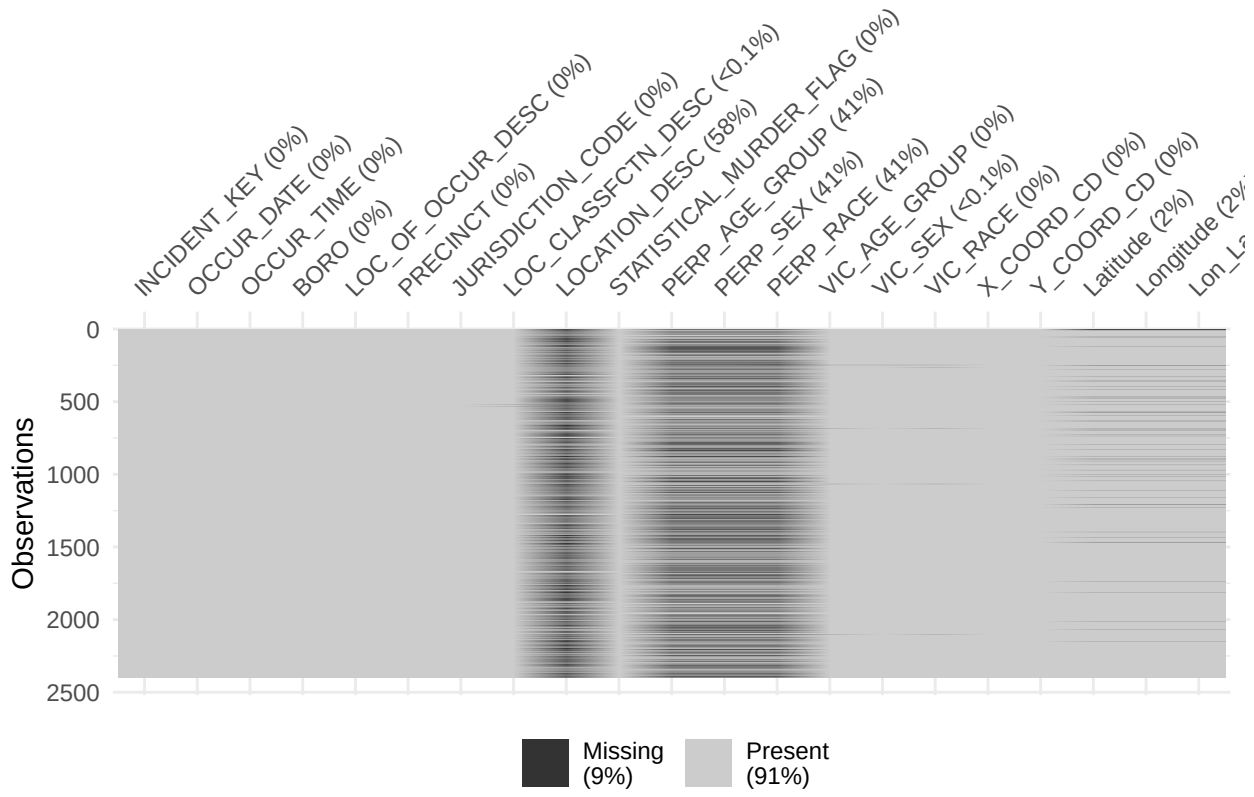
```
##   loc_occur_missing INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO    LOC_OF_OCCUR_DESC
##   <lgl>             <chr>        <date>     <Period>   <chr>   <chr>
## 1 FALSE             263503175    2022-12-02 15H 41M 0S BROOKL~ OUTSIDE
## 2 TRUE              173354054    2021-08-09 15H 10M 0S BROOKL~ INSIDE
## # i 16 more variables: PRECINCT <chr>, JURISDICTION_CODE <chr>,
## #   LOC_CLASSFCTN_DESC <chr>, LOCATION_DESC <chr>,
## #   STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>, PERP_SEX <chr>,
## #   PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>, VIC_RACE <chr>,
## #   X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>, Longitude <dbl>,
## #   Lon_Lat <chr>
```

By setting the first variable, we can gain insight:

- Date difference between the FALSE and TRUE categories. This is significant because it allows us to identify categories that may have been introduced later in the data frame. Consequently, older observations may have more missing values.
- The frequency of the Statistical Murder Flag being TRUE is higher than that of FALSE. This can be interpreted as more data being collected about the incident in murder cases.

To verify the first insight, we will apply a filter based on the date and then review the *vis_miss* again.

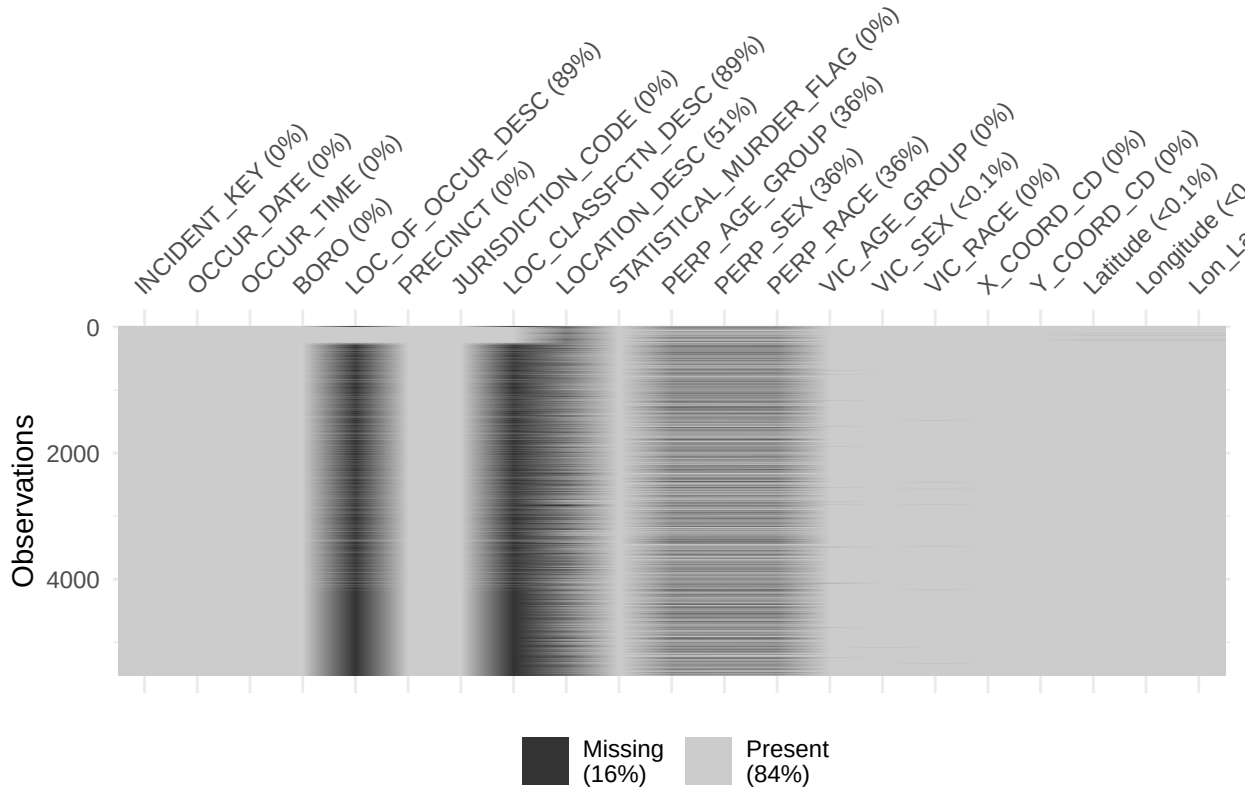
```
check_na_by_date <- nypd_shooting %>%
  filter(OCCUR_DATE > as.Date("2022-05-05"))

  vis_miss(check_na_by_date)
```



We have ascertained that the LOC_OF_OCCUR_DESC and LOC_CLASSFCTN_DESC columns no longer contain any missing values. However, there are no differences in other columns missing values.

Lets inspect the second hypothesis and run the *vis_miss* again.

```
check_na_by_murder <- nypd_shooting %>%
                      filter(STATISTICAL_MURDER_FLAG == TRUE)


vis_miss(check_na_by_murder)
```



Here is no significant distinction between the missing values. We could conduct a similar analysis on the "Perpetrator" columns to ascertain whether we would gain more insight.

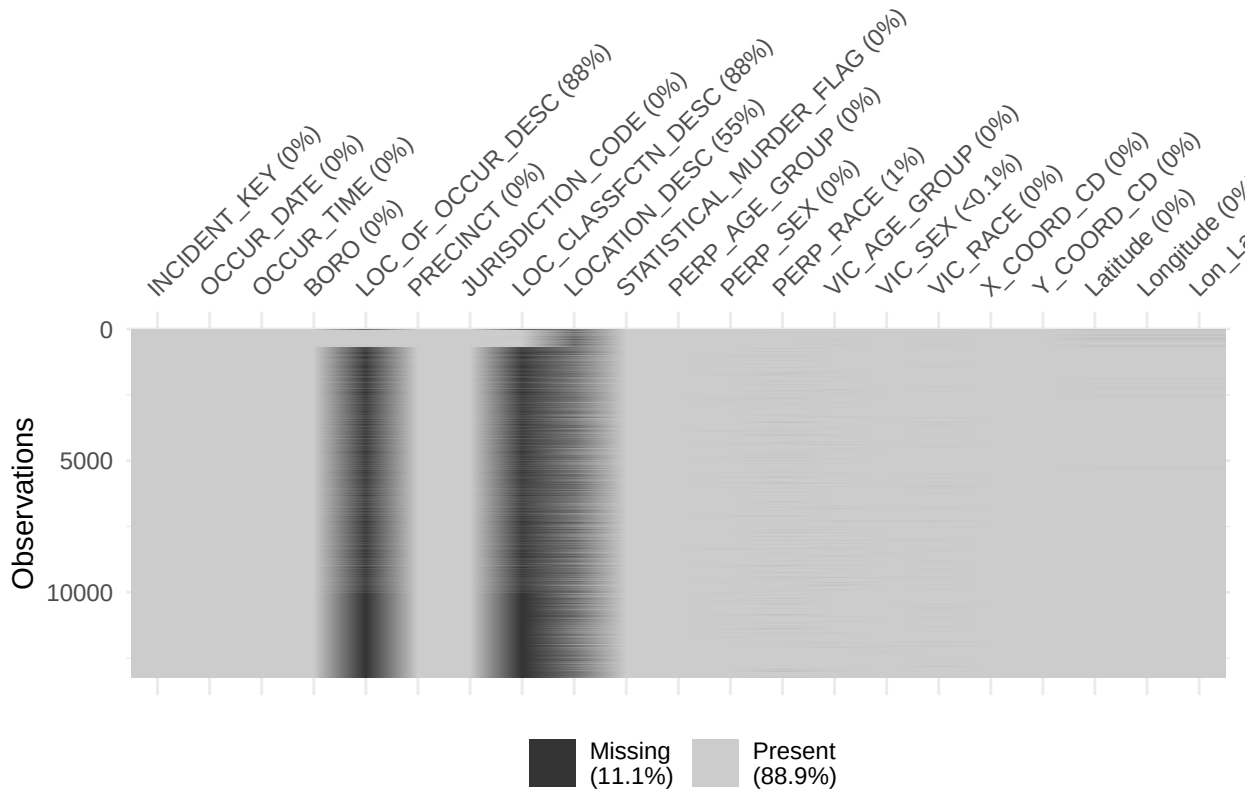
```
na_perp <- nypd_shooting %>%
                  mutate(perp_age_missing = is.na(PERP_AGE_GROUP)) %>%
                  group_by(perp_age_missing) %>%
                  summarize(across(everything(),
                   ~ if(is.factor(.)) {
                     # For factors, return the most frequent level (mode)
                     as.character(names(sort(table(.), decreasing = TRUE))[1])
                   } else if(is.numeric(.)) {
                     # For numeric variables, return the median
                     median(., na.rm = TRUE)
                   } else {
                     # For other types (e.g., character), return the first value
```

```
                    first(.)
                  }))

na_perp
```

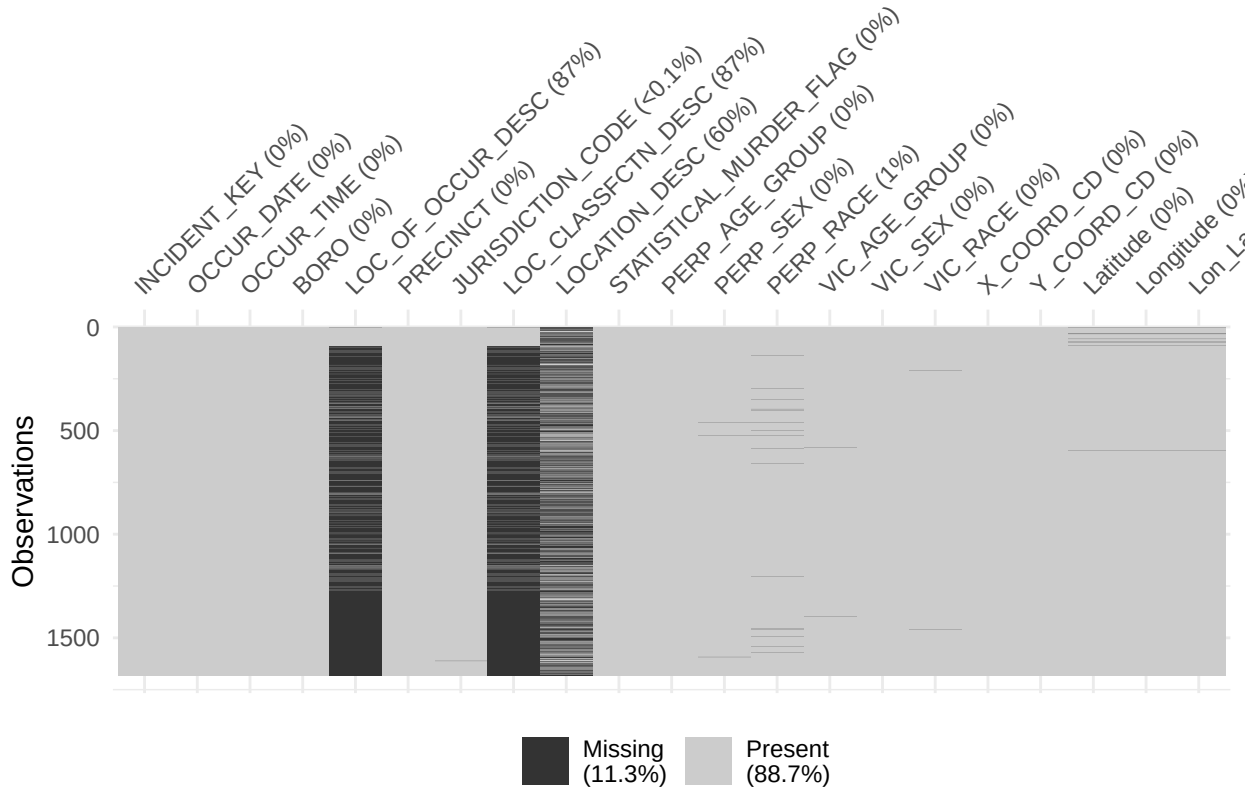
```
## # A tibble: 2 x 22
##   perp_age_missing INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO     LOC_OF_OCCUR_DESC
##   <lgl>            <chr>        <date>     <Period>   <chr>    <chr>
## 1 FALSE            173354054    2018-04-07 15H 39M 0S BROOKLYN OUTSIDE
## 2 TRUE             246884942    2021-08-09 14H 30M 0S BROOKLYN OUTSIDE
## # i 16 more variables: PRECINCT <chr>, JURISDICTION_CODE <chr>,
## #   LOC_CLASSFCTN_DESC <chr>, LOCATION_DESC <chr>,
## #   STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>, PERP_SEX <chr>,
## #   PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>, VIC_RACE <chr>,
## #   X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>, Longitude <dbl>,
## #   Lon_Lat <chr>
```

The significant disparities appear to be largely concentrated in the other "Perpetraror" column. It is noteworthy that PERP_AGE_GROUP for missing values is most common <18. Consequently, we can postulate that when the "Perp" is less than 18, the information is either not collected or not available. Let us examine this further.

```
check_na_by_perp <- nypd_shooting %>%
                    filter(PERP_AGE_GROUP != "<18")

check_na_by_perp_child <- nypd_shooting %>%
                    filter(PERP_AGE_GROUP == "<18")

vis_miss(check_na_by_perp)
```




```
vis_miss(check_na_by_perp_child)
```



As can be seen in both visualisations, the PERP missing values remain consistent. This may indicates that the data is missing due to a lack of information from the perpetrator itself.

Finally, let's investigete the column LOCATION_DESC to see if we can gain some insight

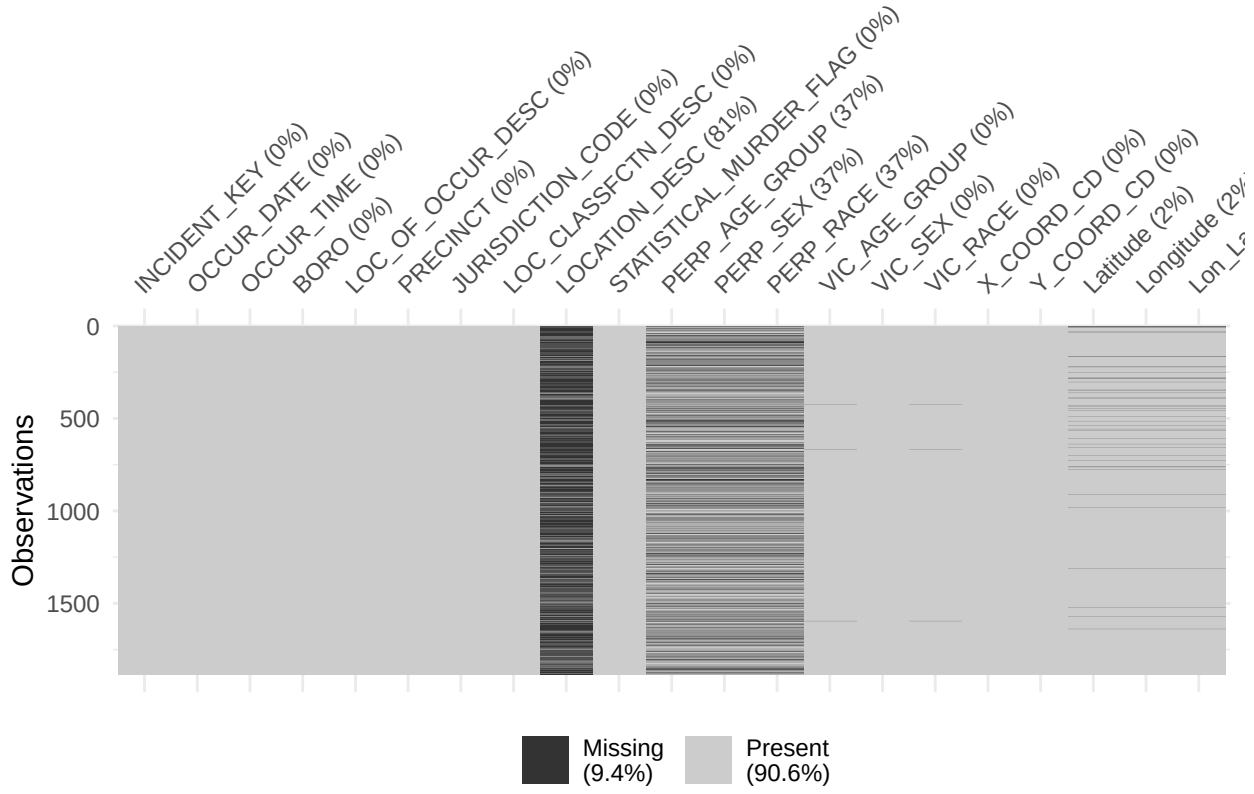
```
na_loc_desc <- nypd_shooting %>%
                  mutate(loc_des_miss = is.na(LOCATION_DESC)) %>%
                  group_by(loc_des_miss) %>%
                  summarize(across(everything(),
                   ~ if(is.factor(.)) {
                     # For factors, return the most frequent level (mode)
                     as.character(names(sort(table(.), decreasing = TRUE))[1])
                   } else if(is.numeric(.)) {
                     # For numeric variables, return the median
                     median(., na.rm = TRUE)
                   } else {
                     # For other types (e.g., character), return the first value
                     first(.)
                   }))

na_loc_desc
```

```
## # A tibble: 2 x 22
##   loc_des_miss INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO     LOC_OF_OCCUR_DESC
##   <lgl>        <chr>        <date>     <Period>   <chr>    <chr>
## 1 FALSE        173354054    2022-12-02 14H 54M 0S BROOKLYN OUTSIDE
```

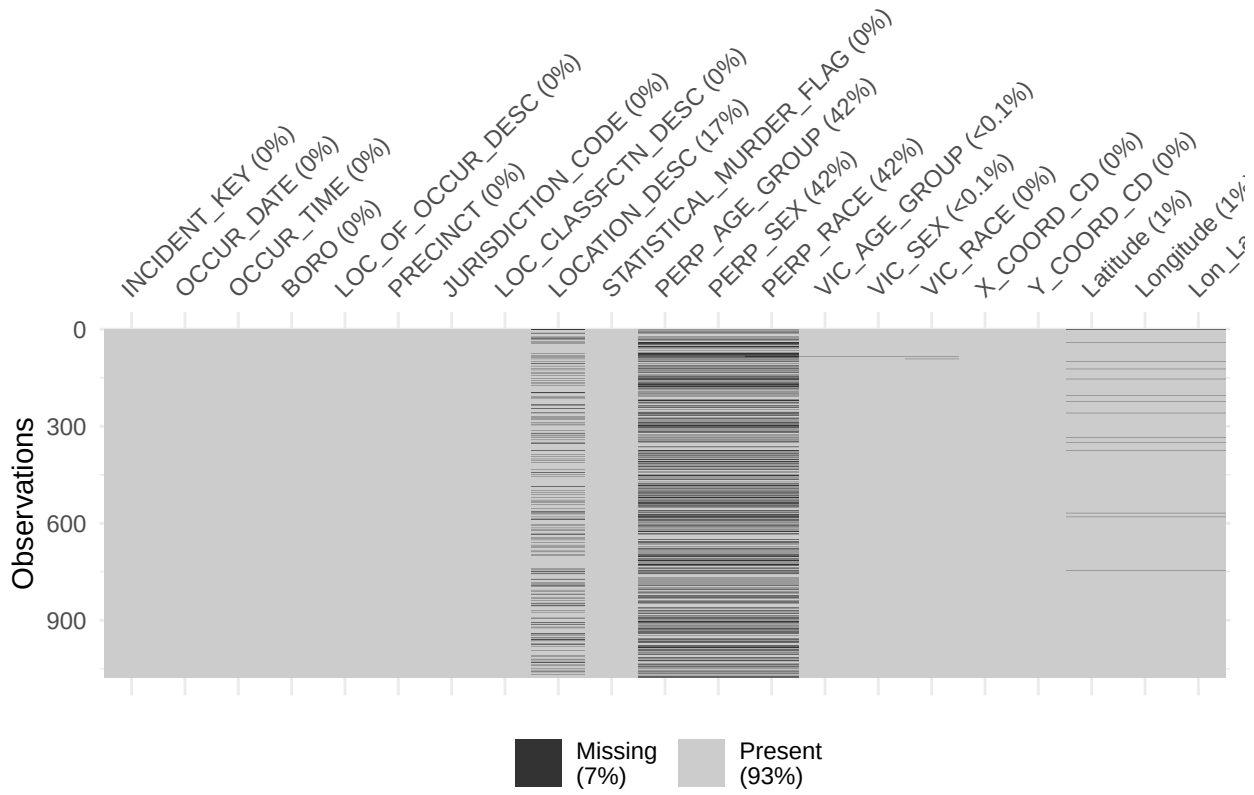
```
## 2 TRUE         33478089     2021-08-09 15H 30M 0S BROOKLYN OUTSIDE
## # i 16 more variables: PRECINCT <chr>, JURISDICTION_CODE <chr>,
## #   LOC_CLASSFCTN_DESC <chr>, LOCATION_DESC <chr>,
## #   STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>, PERP_SEX <chr>,
## #   PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>, VIC_RACE <chr>,
## #   X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>, Longitude <dbl>,
## #   Lon_Lat <chr>
```

The primary difference appears to be in LOC_CLASSFCTN_DESC. This could be attributed to Street having a greater number of observations, but let's conduct a similar analysis.

```
check_na_street <- nypd_shooting %>%
                    filter(LOC_CLASSFCTN_DESC == "STREET")

check_na_other_than_street <- nypd_shooting %>%
                    filter(LOC_CLASSFCTN_DESC != "STREET")

vis_miss(check_na_street)
```



```
vis_miss(check_na_other_than_street)
```



The number of proportional missing values has decreased significantly, indicating that we have less location information in outdoor environments. It should be noted that LOC_CLASSFCTN_DESC was introduced later in the data collection process.

Start data transformation

Since this is a High-Level analysis and we are not using other tables to merge the data, we're going to get rid of INCIDENT_KEY, X_COORD_CD, Y_COORD_CD, Latitude, Longitude and Lon_Lat

```
nypd_shooting <- nypd_shooting %>%
                 select(-c("INCIDENT_KEY", "X_COORD_CD", "Y_COORD_CD", "Latitude", "Longitude", "Lon_La

head(nypd_shooting)
```

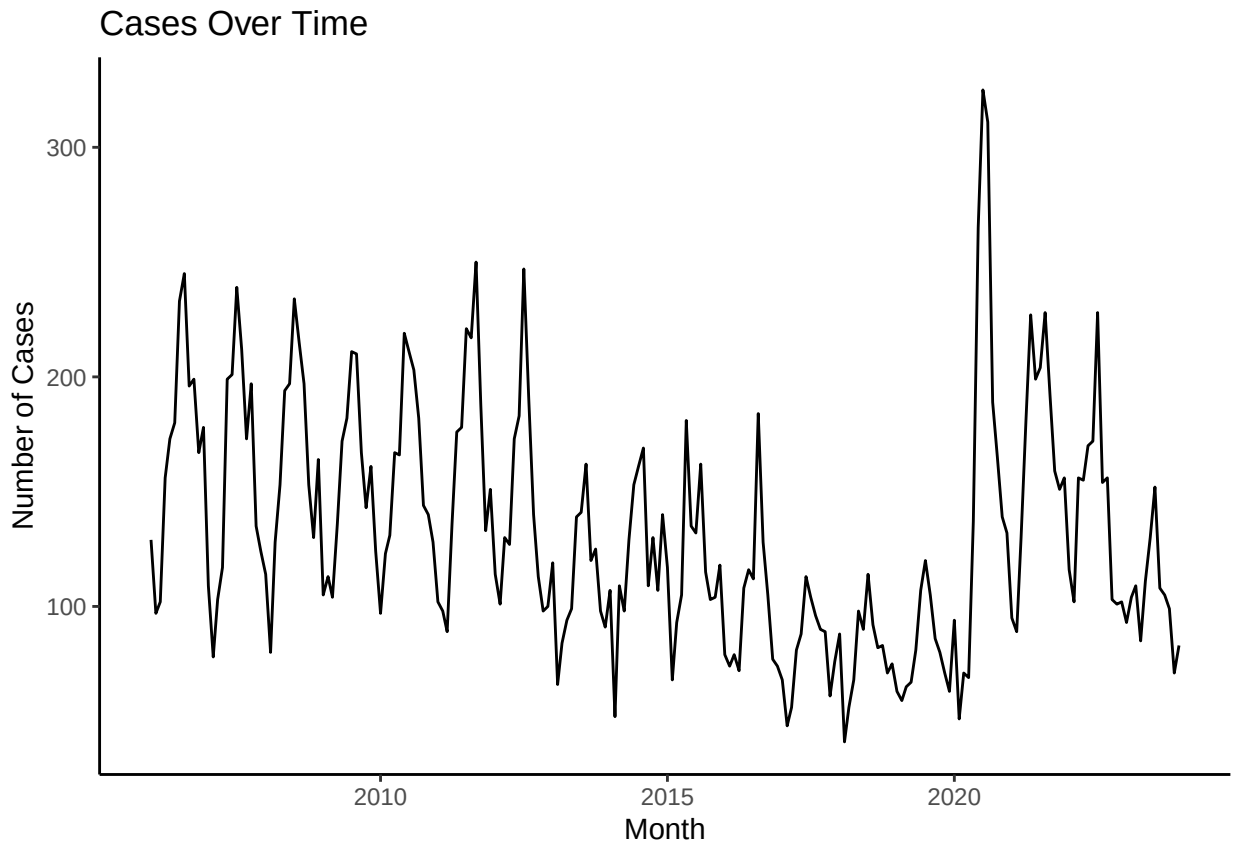
```
## # A tibble: 6 x 15
##   OCCUR_DATE OCCUR_TIME BORO     LOC_OF_OCCUR_DESC PRECINCT JURISDICTION_CODE
##   <date>     <Period>   <fct>    <fct>             <fct>    <fct>
## 1 2021-08-09 1H 6M 0S   BRONX    <NA>              40       0
## 2 2018-04-07 19H 48M 0S BROOKLYN <NA>              79       0
## 3 2022-12-02 22H 57M 0S BRONX    OUTSIDE           47       0
## 4 2006-11-19 1H 50M 0S  BROOKLYN <NA>              66       0
## 5 2010-05-09 1H 58M 0S  BRONX    <NA>              46       0
```

```
## 6 2012-07-22 21H 35M 0S BRONX    <NA>              42       2
## # i 9 more variables: LOC_CLASSFCTN_DESC <fct>, LOCATION_DESC <fct>,
## #   STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <fct>, PERP_SEX <fct>,
## #   PERP_RACE <fct>, VIC_AGE_GROUP <fct>, VIC_SEX <fct>, VIC_RACE <fct>
```

Analizing data.

At first, it would be beneficial to understand the evolution of cases over time grouped by month

```
shoot_cases_over_time <- nypd_shooting %>%
                           mutate(YearMonth = floor_date(OCCUR_DATE, "month")) %>%
                           group_by(YearMonth) %>%
                           summarize(COUNT = n())


plot_cases_over_time = ggplot(shoot_cases_over_time, aes(YearMonth, COUNT)) +
                       geom_line(color = "black") +
                       labs(title = "Cases Over Time",
                            x = "Month",
                            y = "Number of Cases") +
                       theme_classic()


plot_cases_over_time
```

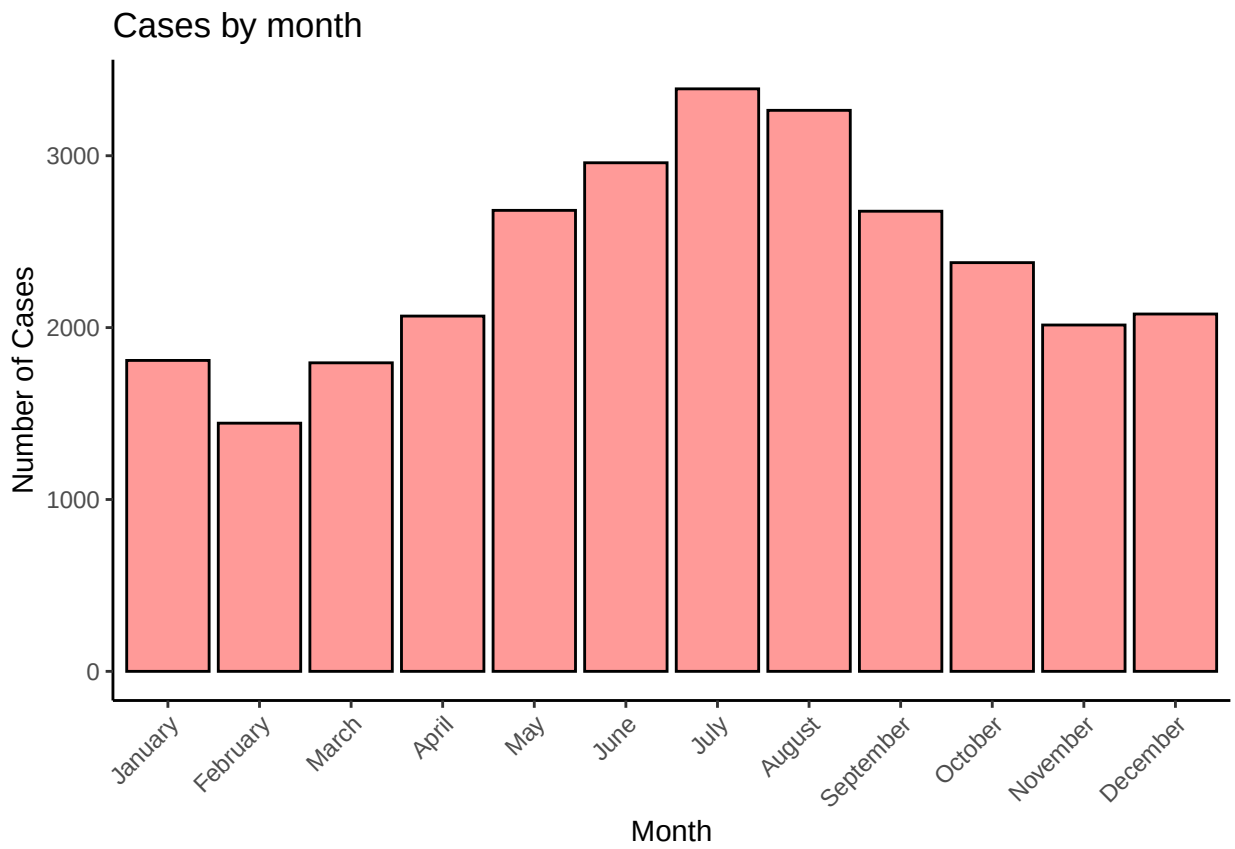


```
shoot_cases_over_time
```

```
## # A tibble: 216 x 2
##    YearMonth  COUNT
##    <date>     <int>
##  1 2006-01-01   129
##  2 2006-02-01    97
##  3 2006-03-01   102
##  4 2006-04-01   156
##  5 2006-05-01   173
##  6 2006-06-01   180
##  7 2006-07-01   233
##  8 2006-08-01   245
##  9 2006-09-01   196
## 10 2006-10-01   199
## # i 206 more rows
```

It is notable that there appear to be two distinct patterns: - There seems to be a seasonal variation in the number of cases, that lead us to the question: which month tend to have more cases ? - There was a decline in the number of cases towards the end of the 2010s, followed by an increase in the early 2020s. A significant occurrence took place in early 2020 that led to an increase in reported cases?

```
shoot_cases_month <- shoot_cases_over_time %>%
                        mutate(MONTH = month(YearMonth, label = TRUE, abbr=FALSE)) %>%
                        group_by(MONTH) %>%
                        arrange(MONTH) %>%
                        summarize(CASES_BY_MONTH = sum(COUNT))

plot_cases_by_month <- ggplot(shoot_cases_month, aes(MONTH, CASES_BY_MONTH)) +
                        geom_bar(stat = "identity", fill = "#FF9a98", color = "Black") +
                        labs(title = "Cases by month",
                                x = "Month",
                                y = "Number of Cases") +
                        theme_classic() +
                        theme(axis.text.x = element_text(angle = 45, hjust = 1))


plot_cases_by_month
```



It is interesting to note that cases occurring over a year seems a normal distribution, with a higher concentration occurring midway through the year.

It is reasonable to conclude that an increase in the number of shootings may result in a corresponding increase in the number of total murders. But that lead us to the next question: have shootings become more lethal?. The next area for analysis is the percentages of shootings that result in a murder over time. The analysis is going to be set by quarter to reduce noise

```
pct_murder_cases_over_time <- nypd_shooting %>%
                              mutate(YearMonth = floor_date(OCCUR_DATE, "quarter")) %>%
                              group_by(YearMonth, STATISTICAL_MURDER_FLAG) %>%
                              summarize(MURDER_COUNT = n()) %>%
                              group_by(YearMonth) %>%
                              mutate(TOTAL = sum(MURDER_COUNT), Percentage_True = round(ifelse(STATISTI
                              filter(STATISTICAL_MURDER_FLAG == TRUE) %>%
                              ungroup()
```

```
## `summarise()` has grouped output by 'YearMonth'. You can override using the
## `.groups` argument.
```

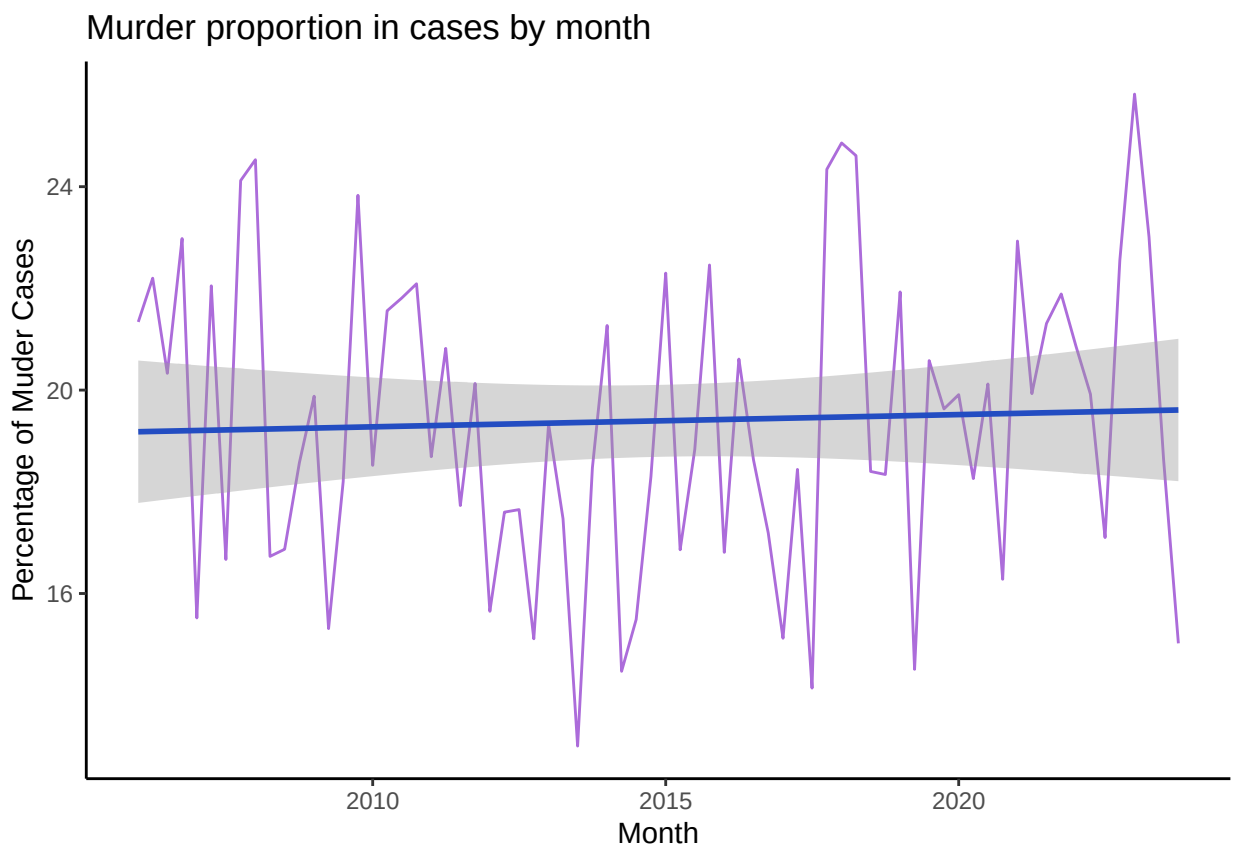
```
pct_murder_cases_over_time
```

```
## # A tibble: 72 x 5
##    YearMonth  STATISTICAL_MURDER_FLAG MURDER_COUNT TOTAL Percentage_True
##    <date>     <lgl>                          <int> <int>           <dbl>
##  1 2006-01-01 TRUE                              70   328            21.3
```

```
##  2 2006-04-01 TRUE                          113   509          22.2
##  3 2006-07-01 TRUE                          137   674          20.3
##  4 2006-10-01 TRUE                          125   544          23.0
##  5 2007-01-01 TRUE                           45   290          15.5
##  6 2007-04-01 TRUE                          114   517          22.0
##  7 2007-07-01 TRUE                          104   624          16.7
##  8 2007-10-01 TRUE                          110   456          24.1
##  9 2008-01-01 TRUE                           79   322          24.5
## 10 2008-04-01 TRUE                           91   544          16.7
## # i 62 more rows
```

```
plot_murder_cases_over_time <- ggplot(pct_murder_cases_over_time, aes(YearMonth, Percentage_True)) +
                               geom_line(color = "#AC6CDA") +
                               geom_smooth(method = "lm", color = "#234dc2", se = TRUE) +
                               labs(title = "Murder proportion in cases by month",
                                     x = "Month",
                                     y = "Percentage of Muder Cases") +
                               theme_classic()

plot_murder_cases_over_time
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



As the data illustrates, the fatality rate associated with shootings remains consistent across all reported observations. While the overall tendency is similar. Are some races at higher risk of being involved in a shooting incident? How's the mortality rate?

```
shooting_by_race <- nypd_shooting %>%
                    group_by(STATISTICAL_MURDER_FLAG, VIC_RACE) %>%
                    summarize(COUNT = n()) %>%
                    arrange(STATISTICAL_MURDER_FLAG, ascending = TRUE) %>%
                    group_by(VIC_RACE) %>%
                    mutate(TOTAL = sum(COUNT), PCT_MURDER = round(ifelse(STATISTICAL_MURDER_FLAG == TRU
                    ungroup()
```

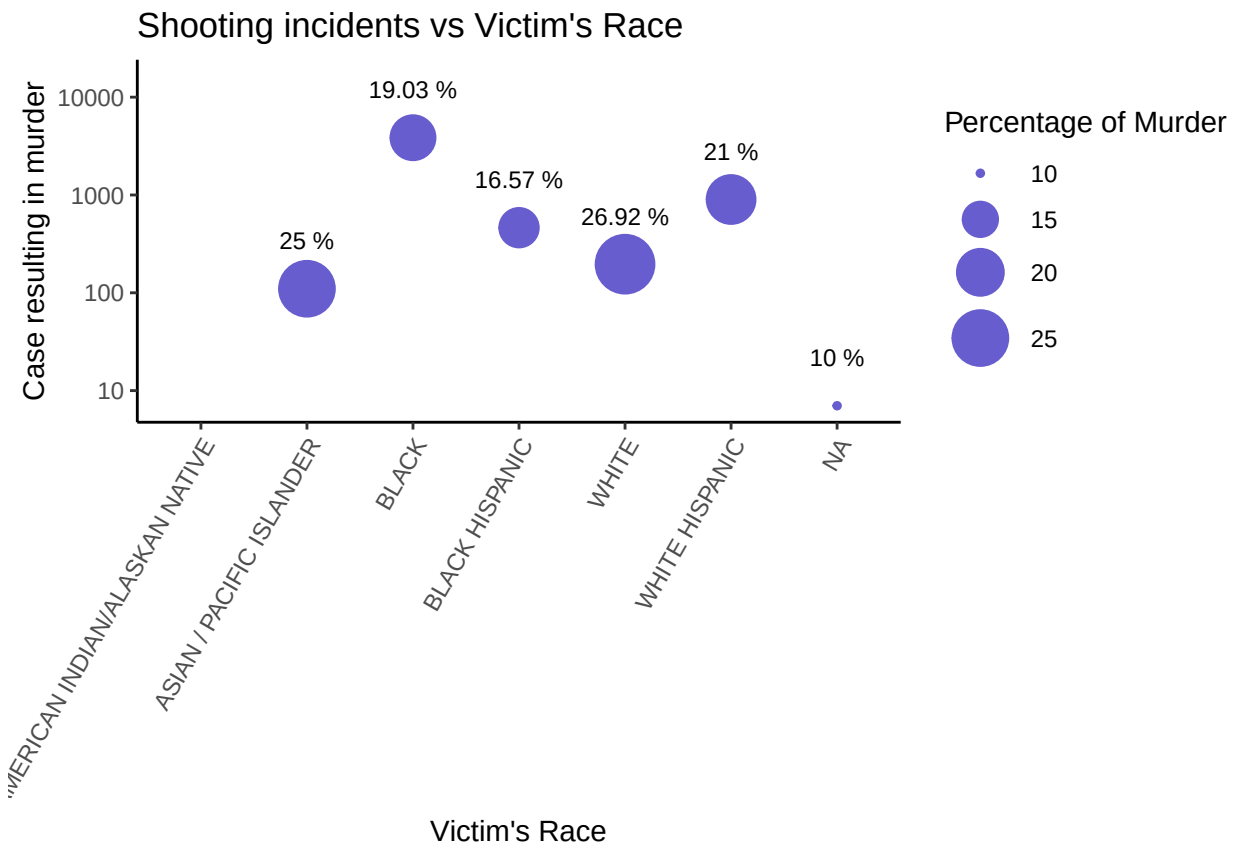
```
## `summarise()` has grouped output by 'STATISTICAL_MURDER_FLAG'. You can override
## using the `.groups` argument.
```

```
shooting_by_race
```

```
## # A tibble: 13 x 5
##    STATISTICAL_MURDER_FLAG VIC_RACE                       COUNT TOTAL PCT_MURDER
##    <lgl>                   <fct>                          <int> <int>      <dbl>
##  1 FALSE                   AMERICAN INDIAN/ALASKAN NATIVE    11    11       NA
##  2 FALSE                   ASIAN / PACIFIC ISLANDER         330   440       NA
##  3 FALSE                   BLACK                          16382 20233       NA
##  4 FALSE                   BLACK HISPANIC                  2332  2795       NA
##  5 FALSE                   WHITE                            532   728       NA
##  6 FALSE                   WHITE HISPANIC                  3382  4281       NA
##  7 FALSE                   <NA>                              63    70       NA
##  8 TRUE                    ASIAN / PACIFIC ISLANDER         110   440       25
##  9 TRUE                    BLACK                           3851 20233       19.0
## 10 TRUE                    BLACK HISPANIC                   463  2795       16.6
## 11 TRUE                    WHITE                            196   728       26.9
## 12 TRUE                    WHITE HISPANIC                   899  4281       21
## 13 TRUE                    <NA>                               7    70       10
```

```
plot_shot_race <- ggplot(shooting_by_race, aes(VIC_RACE, COUNT, size = PCT_MURDER)) +
                    geom_point(color = "#685DCE") +
                    geom_text(data = subset(shooting_by_race, !is.na(PCT_MURDER)), aes(label = paste(PC
                              vjust = -2.3,size = 3) +
                    scale_y_continuous(trans="log10") +
                    scale_size_continuous(range = c(1, 10)) +
                    labs(title = "Shooting incidents vs Victim's Race",
                          x = "Victim's Race",
                          y = "Case resulting in murder",
                          size = "Percentage of Murder") +
                    theme_classic() +
                    theme(axis.text.x = element_text(angle = 60, hjust = 1))


plot_shot_race
```

```
## Warning: Removed 7 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



As the graph illustrates, the y-axis is presented on a logarithmic scale to accommodate discrepancies between ethnic groups. Individuals of Black ethnicity are significantly more likely to be involved in incidents, whereas other ethnicities exhibit a lower incidence of such occurrences. It is noteworthy that while White individuals experience a relatively low number of incidents, their mortality rate is marginally higher. A further investigation could provide insights into the underlying reasons for this phenomenon.

Conclusions

The "New York Police Department Shooting Incidents" Datasets offer valuable insight into behavioral patterns, victims, and tendencies over time, which may ultimately influence policy and procedure changes.

It is crucial to identify potential limitations, such as the presence of bias, at the initial stages of the investigation. This includes the identification of common sources of bias, such as selection bias, information bias, and contrast effect. Also, understanding the reasons behind the absence of a significant amount of information is also essential. During the course of this work, we will examine the underlying causes of this phenomenon.

In terms of personal bias, it is possible that I may exhibit confirmation bias and affinity bias, particularly in relation to recent cases of excessive police force. One method of addressing this issue is to challenge one's own assumptions, be mindful of disconfirming evidence, and utilize objective measures to maintain as much structure as possible.

```
sessionInfo()
```

```
## R version 4.4.2 (2024-10-31)
## Platform: aarch64-apple-darwin20
## Running under: macOS Sequoia 15.1.1
```

```
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib;  LAPACK v
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: America/Bogota
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] visdat_0.6.0    lubridate_1.9.3 forcats_1.0.0   stringr_1.5.1
##  [5] dplyr_1.1.4     purrr_1.0.2     readr_2.1.5     tidyr_1.3.1
##  [9] tibble_3.2.1    ggplot2_3.5.1   tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] utf8_1.2.4        generics_0.1.3    lattice_0.22-6    stringi_1.8.4
##  [5] hms_1.1.3         digest_0.6.37     magrittr_2.0.3    evaluate_1.0.0
##  [9] grid_4.4.2        timechange_0.3.0  fastmap_1.2.0     Matrix_1.7-1
## [13] mgcv_1.9-1        fansi_1.0.6       scales_1.3.0      cli_3.6.3
## [17] rlang_1.1.4       crayon_1.5.3      splines_4.4.2     bit64_4.5.2
## [21] munsell_0.5.1     withr_3.0.1       yaml_2.3.10       tools_4.4.2
## [25] parallel_4.4.2    tzdb_0.4.0        colorspace_2.1-1  curl_5.2.3
## [29] vctrs_0.6.5       R6_2.5.1          lifecycle_1.0.4   bit_4.5.0
## [33] vroom_1.6.5       pkgconfig_2.0.3   pillar_1.9.0      gtable_0.3.5
## [37] glue_1.7.0        xfun_0.47         tidyselect_1.2.1  highr_0.11
## [41] rstudioapi_0.16.0 knitr_1.48        farver_2.1.2      nlme_3.1-166
## [45] htmltools_0.5.8.1 rmarkdown_2.28    labeling_0.4.3    compiler_4.4.2
```