HM1

2025-02-03

Part 1: Data Cleaning and Exploration (Python & R)

1) Identify columns with missing values and their respective counts Drop columns where more than 50% of the data is missing (store this version as a new dataset).

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.4.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
df <- read.csv("C:\\Users\\Hovgr\\OneDrive\\Desktop\\DataViz\\HM1\\crime_data.csv")
head(df, 5)
```

```
##       DR_NO              Date.Rptd               DATE.OCC TIME.OCC AREA
## 1 241711715 08/01/2024 12:00:00 AM 08/01/2024 12:00:00 AM     1319   17
## 2 231014031 09/21/2023 12:00:00 AM 09/15/2023 12:00:00 AM     1930   10
## 3 231010808 06/27/2023 12:00:00 AM 06/26/2023 12:00:00 AM     1230   10
## 4 211410441 04/25/2021 12:00:00 AM 04/25/2021 12:00:00 AM     2330   14
## 5 211114569 10/25/2021 12:00:00 AM 10/25/2021 12:00:00 AM     1455   11
##     AREA.NAME Rpt.Dist.No Part.1.2 Crm.Cd                        Crm.Cd.Desc
## 1  Devonshire        1791        1    440 THEFT PLAIN - PETTY ($950 & UNDER)
## 2 West Valley        1011        2    354                  THEFT OF IDENTITY
## 3 West Valley        1015        2    354                  THEFT OF IDENTITY
## 4     Pacific        1488        2    626  INTIMATE PARTNER - SIMPLE ASSAULT
## 5   Northeast        1123        1    210                            ROBBERY
##               Mocodes Vict.Age Vict.Sex Vict.Descent Premis.Cd
## 1           0344 0394       25        M            O       501
## 2           1822 0930       23        F            W       501
## 3           1822 0928       37        F            O       501
## 4      0913 0400 0448       25        F            B       503
## 5 1309 0945 0334 0325        0        X            X       412
##                                Premis.Desc Weapon.Used.Cd
## 1                   SINGLE FAMILY DWELLING            NA
## 2                   SINGLE FAMILY DWELLING            NA
## 3                   SINGLE FAMILY DWELLING            NA
## 4                                    HOTEL           400
## 5 ELECTRONICS STORE (IE:RADIO SHACK, ETC.)           200
```

```
##                                   Weapon.Desc Status Status.Desc Crm.Cd.1
## 1                                                IC Invest Cont      440
## 2                                                IC Invest Cont      354
## 3                                                IC Invest Cont      354
## 4 STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)     IC Invest Cont      626
## 5               KNIFE WITH BLADE 6INCHES OR LESS     IC Invest Cont      210
##   Crm.Cd.2 Crm.Cd.3 Crm.Cd.4                                LOCATION
## 1       NA       NA       NA  8300    KELVIN                     AV
## 2       NA       NA       NA 18900    CANTLAY                    ST
## 3       NA       NA       NA  7300    ENFIELD                    AV
## 4       NA       NA       NA  5800 W  CENTURY                    BL
## 5       NA       NA       NA  2900    LOS FELIZ                  BL
##   Cross.Street     LAT       LON
## 1              34.2200 -118.5863
## 2              34.2023 -118.5458
## 3              34.2033 -118.5241
## 4              33.9456 -118.3835
## 5               0.0000    0.0000
```

2) Identify columns with missing values and their respective counts. Drop columns where more than 50% of the data is missing (store this version as a new dataset).

```
print(colSums(is.na(df))[colSums(is.na(df)) > 0])
```

```
## Weapon.Used.Cd       Crm.Cd.1       Crm.Cd.2       Crm.Cd.3       Crm.Cd.4
##          33654              2          46448          49885          49995
```

```
df_cleaned <- df %>% select(-all_of(names(colSums(is.na(df))[colSums(is.na(df)) > nrow(df) * 0.5])))
head(df_cleaned, 5)
```

```
##       DR_NO              Date.Rptd               DATE.OCC TIME.OCC AREA
## 1 241711715 08/01/2024 12:00:00 AM 08/01/2024 12:00:00 AM     1319   17
## 2 231014031 09/21/2023 12:00:00 AM 09/15/2023 12:00:00 AM     1930   10
## 3 231010808 06/27/2023 12:00:00 AM 06/26/2023 12:00:00 AM     1230   10
## 4 211410441 04/25/2021 12:00:00 AM 04/25/2021 12:00:00 AM     2330   14
## 5 211114569 10/25/2021 12:00:00 AM 10/25/2021 12:00:00 AM     1455   11
##     AREA.NAME Rpt.Dist.No Part.1.2 Crm.Cd                       Crm.Cd.Desc
## 1  Devonshire        1791        1    440 THEFT PLAIN - PETTY ($950 & UNDER)
## 2 West Valley        1011        2    354                  THEFT OF IDENTITY
## 3 West Valley        1015        2    354                  THEFT OF IDENTITY
## 4     Pacific        1488        2    626  INTIMATE PARTNER - SIMPLE ASSAULT
## 5   Northeast        1123        1    210                            ROBBERY
##               Mocodes Vict.Age Vict.Sex Vict.Descent Premis.Cd
## 1           0344 0394       25        M            O       501
## 2           1822 0930       23        F            W       501
## 3           1822 0928       37        F            O       501
## 4      0913 0400 0448       25        F            B       503
## 5 1309 0945 0334 0325        0        X            X       412
##                              Premis.Desc
## 1                  SINGLE FAMILY DWELLING
## 2                  SINGLE FAMILY DWELLING
## 3                  SINGLE FAMILY DWELLING
## 4                                   HOTEL
## 5 ELECTRONICS STORE (IE:RADIO SHACK, ETC.)
```

```
##                                     Weapon.Desc Status Status.Desc Crm.Cd.1
## 1                                                   IC Invest Cont      440
## 2                                                   IC Invest Cont      354
## 3                                                   IC Invest Cont      354
## 4 STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)    IC Invest Cont      626
## 5               KNIFE WITH BLADE 6INCHES OR LESS    IC Invest Cont      210
##                                  LOCATION Cross.Street     LAT       LON
## 1  8300    KELVIN                      AV              34.2200 -118.5863
## 2 18900    CANTLAY                     ST              34.2023 -118.5458
## 3  7300    ENFIELD                     AV              34.2033 -118.5241
## 4  5800 W  CENTURY                     BL              33.9456 -118.3835
## 5  2900    LOS FELIZ                   BL               0.0000    0.0000
```

3) Convert the DATE OCC column to a datetime format. Extract the year, month, and day into separate columns. Create a new column for the hour using the TIME OCC column.

```
df_cleaned <- df_cleaned %>% mutate(
 DATE.OCC = as.Date(DATE.OCC, format = "%m/%d/%Y"),
 Year = format(DATE.OCC, "%Y"),
 Month = format(DATE.OCC, "%m"),
 Day = format(DATE.OCC, "%d"),
 Hour = as.integer(TIME.OCC / 100)
)

head(df_cleaned, 5)
```

```
##       DR_NO              Date.Rptd   DATE.OCC TIME.OCC AREA   AREA.NAME
## 1 241711715 08/01/2024 12:00:00 AM 2024-08-01     1319   17  Devonshire
## 2 231014031 09/21/2023 12:00:00 AM 2023-09-15     1930   10 West Valley
## 3 231010808 06/27/2023 12:00:00 AM 2023-06-26     1230   10 West Valley
## 4 211410441 04/25/2021 12:00:00 AM 2021-04-25     2330   14     Pacific
## 5 211114569 10/25/2021 12:00:00 AM 2021-10-25     1455   11   Northeast
##   Rpt.Dist.No Part.1.2 Crm.Cd                        Crm.Cd.Desc
## 1        1791        1    440 THEFT PLAIN - PETTY ($950 & UNDER)
## 2        1011        2    354                  THEFT OF IDENTITY
## 3        1015        2    354                  THEFT OF IDENTITY
## 4        1488        2    626  INTIMATE PARTNER - SIMPLE ASSAULT
## 5        1123        1    210                            ROBBERY
##              Mocodes Vict.Age Vict.Sex Vict.Descent Premis.Cd
## 1           0344 0394       25        M            O       501
## 2           1822 0930       23        F            W       501
## 3           1822 0928       37        F            O       501
## 4      0913 0400 0448       25        F            B       503
## 5 1309 0945 0334 0325        0        X            X       412
##                              Premis.Desc
## 1                  SINGLE FAMILY DWELLING
## 2                  SINGLE FAMILY DWELLING
## 3                  SINGLE FAMILY DWELLING
## 4                                   HOTEL
## 5 ELECTRONICS STORE (IE:RADIO SHACK, ETC.)
##                                     Weapon.Desc Status Status.Desc Crm.Cd.1
## 1                                                   IC Invest Cont      440
## 2                                                   IC Invest Cont      354
```

```
## 3                                                 IC Invest Cont      354
## 4 STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)  IC Invest Cont      626
## 5              KNIFE WITH BLADE 6INCHES OR LESS   IC Invest Cont      210
##                                  LOCATION Cross.Street     LAT       LON Year
## 1  8300    KELVIN                      AV              34.2200 -118.5863 2024
## 2 18900    CANTLAY                     ST              34.2023 -118.5458 2023
## 3  7300    ENFIELD                     AV              34.2033 -118.5241 2023
## 4  5800 W  CENTURY                     BL              33.9456 -118.3835 2021
## 5  2900    LOS FELIZ                   BL               0.0000    0.0000 2021
##   Month Day Hour
## 1    08  01   13
## 2    09  15   19
## 3    06  26   12
## 4    04  25   23
## 5    10  25   14
```

4) Filter the dataset for crimes that occurred in 2023. Further filter crimes with the description BURGLARY in the Crm Cd Desc column.

```
df_cleaned <- df_cleaned %>% filter(Year == 2023, Crm.Cd.Desc == "BURGLARY")
head(df_cleaned, 5)
```

```
##       DR_NO              Date.Rptd   DATE.OCC TIME.OCC AREA   AREA.NAME
## 1 231107877 04/15/2023 12:00:00 AM 2023-01-15      500   11   Northeast
## 2 231912840 08/15/2023 12:00:00 AM 2023-08-14     2200   19     Mission
## 3 230813484 08/19/2023 12:00:00 AM 2023-08-19      510    8     West LA
## 4 230126836 12/20/2023 12:00:00 AM 2023-12-10     1200    1     Central
## 5 231506351 02/26/2023 12:00:00 AM 2023-02-22     1230   15 N Hollywood
##   Rpt.Dist.No Part.1.2 Crm.Cd Crm.Cd.Desc             Mocodes Vict.Age Vict.Sex
## 1        1151        1    310    BURGLARY      0531 1822 0451       53        F
## 2        1962        1    310    BURGLARY      0314 0344 1402       45        F
## 3         857        1    310    BURGLARY 0358 1609 0321 1822        0        X
## 4         154        1    310    BURGLARY               0344       34        M
## 5        1562        1    310    BURGLARY           1609 0344       60        F
##   Vict.Descent Premis.Cd           Premis.Desc Weapon.Desc Status Status.Desc
## 1            H       504        OTHER RESIDENCE                  IC Invest Cont
## 2            B       507  CONDOMINIUM/TOWNHOUSE                  IC Invest Cont
## 3            X       210   RESTAURANT/FAST FOOD                  IC Invest Cont
## 4            B       221         PUBLIC STORAGE                  IC Invest Cont
## 5            W       501 SINGLE FAMILY DWELLING                  IC Invest Cont
##   Crm.Cd.1                                LOCATION Cross.Street     LAT
## 1      310  5000 W  SUNSET                      BL              34.0981
## 2      310 15000    CORE                       LN              34.2424
## 3      310  9400 W  PICO                       BL              34.0553
## 4      310   100 E  6TH                        ST              34.0460
## 5      310  4400    BABCOCK                    AV              34.1504
##         LON Year Month Day Hour
## 1 -118.2983 2023    01  15    5
## 2 -118.4596 2023    08  14   22
## 3 -118.3943 2023    08  19    5
## 4 -118.2493 2023    12  10   12
## 5 -118.4063 2023    02  22   12
```

5) Group the data by AREA NAME and calculate the total number of crimes and the average victim age. Sort the results by total crimes in descending order.

```
df_cleaned %>% group_by(AREA.NAME) %>% summarise(
    count = n(),
    mean = mean(Vict.Age, na.rm = TRUE) ) %>% arrange(desc(count)) %>% head(5)
```

```
## # A tibble: 5 x 3
##   AREA.NAME   count  mean
##   <chr>       <int> <dbl>
## 1 Devonshire     58  43.9
## 2 West LA        58  40.9
## 3 Olympic        53  30.3
## 4 West Valley    52  35.2
## 5 N Hollywood    50  29.7
```

Part 3: Further Exploration

##1. Group the data by Month and count the number of crimes.

```
df_cleaned %>% group_by(Month) %>% summarise(count = n())
```

```
## # A tibble: 12 x 2
##    Month count
##    <chr> <int>
##  1 01       61
##  2 02       59
##  3 03       54
##  4 04       59
##  5 05       36
##  6 06       59
##  7 07       67
##  8 08       67
##  9 09       70
## 10 10       67
## 11 11       58
## 12 12       68
```

2) Count the number of crimes where a weapon was used (Weapon Used Cd is not null).

```
df %>% filter(!is.na(Weapon.Used.Cd)) %>% summarise(count = n())
```

```
##   count
## 1 16346
```

3) Group the data by Premis Desc and count the number of crimes.

```
df %>% group_by(Premis.Desc) %>% summarise(count = n()) %>% head(5)
```

```
## # A tibble: 5 x 2
##   Premis.Desc                                 count
##   <chr>                                       <int>
## 1 ""                                             29
```

```
## 2 "7TH AND METRO CENTER (NOT LINE SPECIFIC)"    13
## 3 "ABANDONED BUILDING ABANDONED HOUSE"          45
## 4 "ABORTION CLINIC/ABORTION FACILITY*"           1
## 5 "AIRCRAFT"                                     1
```

Part 4: Advanced Analysis

1. Create a new column, Severity Score, based on the following rules:

• Assign 5 points if a weapon was used (Weapon Used Cd is not null). • Assign 3 points for crimes under BURGLARY. • Assign 1 point for all other crimes. • Group by AREA NAME and find the total severity score for each area.

```
df <- df %>% mutate(
    Severity.Score = case_when(
      !is.na(Weapon.Used.Cd) & Crm.Cd.Desc == "BURGLARY" ~ 8,
      !is.na(Weapon.Used.Cd) ~ 5,
      Crm.Cd.Desc == "BURGLARY" ~ 3,
      TRUE ~ 1
    )
)

df %>% group_by(AREA.NAME) %>%
  summarise(total = sum(Severity.Score)) %>%
  arrange(desc(total))
```

```
## # A tibble: 21 x 2
##    AREA.NAME   total
##    <chr>       <dbl>
##  1 77th Street  9298
##  2 Central      8625
##  3 Southeast    7162
##  4 Southwest    7029
##  5 Newton       6815
##  6 Hollywood    6707
##  7 Pacific      6652
##  8 Olympic      6275
##  9 Rampart      6267
## 10 N Hollywood  5476
## # i 11 more rows
```