Homework 1

2025-02-06

Part 1

1.1

load libraries

```
library(dplyr)
library(lubridate)
```

- load dataset
- print first 5 rows

```
crime_data <- read.csv("crime_data.csv", stringsAsFactors = FALSE)

head(crime_data, 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
```

1.2

- get number of missing values for columns
- delete columns which miss more than 50% of data

```
missing_values <- colSums(is.na(crime_data))

missing_values[missing_values > 0]
```

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

```
threshold <- 0.5 * nrow(crime_data)
columns_to_drop <- names(missing_values[missing_values > threshold])

crime_data_cleaned <- crime_data %>% select(-one_of(columns_to_drop))

names(crime_data_cleaned)
```

```
##  [1] "DR_NO"        "Date.Rptd"    "DATE.OCC"     "TIME.OCC"     "AREA"
##  [6] "AREA.NAME"    "Rpt.Dist.No"  "Part.1.2"     "Crm.Cd"       "Crm.Cd.Desc"
## [11] "Mocodes"      "Vict.Age"     "Vict.Sex"     "Vict.Descent" "Premis.Cd"
## [16] "Premis.Desc"  "Weapon.Desc"  "Status"       "Status.Desc"  "Crm.Cd.1"
## [21] "LOCATION"     "Cross.Street" "LAT"          "LON"
```

1.3

- Convert Date.OCC to date format
- Extract Year, Month, Day to new columns
- Calculate Hour from TIME.OCC

```
crime_data_cleaned$DATE.OCC <- mdy_hms(crime_data_cleaned$DATE.OCC)

crime_data_cleaned$Year <- year(crime_data_cleaned$DATE.OCC)
crime_data_cleaned$Month <- month(crime_data_cleaned$DATE.OCC)
```

```r
crime_data_cleaned$Day <- day(crime_data_cleaned$DATE.OCC)

crime_data_cleaned$Hour <- as.integer(crime_data_cleaned$TIME.OCC / 100)

head(crime_data_cleaned[, c("DATE.OCC", "TIME.OCC", "Year", "Month", "Day", "Hour")])
```

```
##     DATE.OCC TIME.OCC Year Month Day Hour
## 1 2024-08-01     1319 2024     8   1   13
## 2 2023-09-15     1930 2023     9  15   19
## 3 2023-06-26     1230 2023     6  26   12
## 4 2021-04-25     2330 2021     4  25   23
## 5 2021-10-25     1455 2021    10  25   14
## 6 2022-04-28     2239 2022     4  28   22
```

1.4

- Filter for 2023
- Filter for burglaries
- check if size changed

```r
crime_data_2023 <- crime_data_cleaned %>% filter(Year == 2023)

crime_burglary_2023 <- crime_data_2023 %>%
  filter(grepl("BURGLARY", crime_data_2023$Crm.Cd.Desc, ignore.case = TRUE))

cat("unfiltered data:", dim(crime_data_cleaned))
```

```
## unfiltered data: 50000 28
```

```r
cat("data for 2023:", dim(crime_data_2023))
```

```
## data for 2023: 11665 28
```

```r
cat("data for burglary + 2023:", dim(crime_burglary_2023))
```

```
## data for burglary + 2023: 1404 28
```

1.5

- Group by AREA.NAME
- Calculate total crimes and avg victim age
- Display results

```r
crime_summary <- crime_burglary_2023 %>%
  group_by(AREA.NAME) %>%
  summarise(
    Total_Crimes = n(),
    Avg_Victim_Age = mean(Vict.Age, na.rm = TRUE)
  ) %>%
```

```r
  arrange(desc(Total_Crimes))

print(crime_summary, n = Inf)
```

```
## # A tibble: 21 x 3
##    AREA.NAME   Total_Crimes Avg_Victim_Age
##    <chr>              <int>          <dbl>
##  1 Central              146           32.8
##  2 West LA              107           40.2
##  3 Olympic               96           32.7
##  4 Wilshire              90           38.2
##  5 Devonshire            88           42.4
##  6 West Valley           87           36.1
##  7 N Hollywood           84           33.4
##  8 Pacific               77           29.2
##  9 Van Nuys              69           40.1
## 10 Northeast             65           30.5
## 11 Southwest             61           35.6
## 12 Newton                59           29.2
## 13 Hollywood             55           30.4
## 14 Rampart               54           28.9
## 15 Topanga               52           42.1
## 16 77th Street           51           36.7
## 17 Harbor                39           28.5
## 18 Foothill              33           36.6
## 19 Southeast             33           40.8
## 20 Hollenbeck            29           20.2
## 21 Mission               29           38.4
```

Part 3

3.1

- Group by Month
- summaries total crimes for each month
- Display results

```r
crimes_by_month <- crime_data_cleaned %>%
  group_by(Month) %>%
  summarise(Total_Crimes = n())

print(crimes_by_month)
```

```
## # A tibble: 12 x 2
##    Month Total_Crimes
##    <dbl>        <int>
##  1     1         4578
##  2     2         4290
##  3     3         4361
##  4     4         4189
##  5     5         4088
```

```
##  6     6         4058
##  7     7         4179
##  8     8         4147
##  9     9         4054
## 10    10         4226
## 11    11         3948
## 12    12         3882
```

3.2

- Filter crimes where weapon was not used
- Get number of such crimes
- Display results

NOTE: Using original crime_data because crime_data_cleaned does not have the Weapon.Used.Cd column

```
crimes_with_weapon <- crime_data %>%
  filter(!is.na(Weapon.Used.Cd)) %>%
  summarise(Weapon_Crimes = n())

print(crimes_with_weapon)
```

```
##   Weapon_Crimes
## 1         16346
```

3.3

- Group by premis.desc
- Get number of crimes for each premis.desc
- Display results

```
crime_by_premis_desc <- crime_data_cleaned %>%
  group_by(Premis.Desc) %>%
  summarise(Total_Crimes = n())

print(crime_by_premis_desc)
```

```
## # A tibble: 267 x 2
##    Premis.Desc                                                Total_Crimes
##    <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
##  6 "ALLEY"                                                            336
##  7 "APARTMENT/CONDO COMMON LAUNDRY ROOM"                                17
##  8 "ARCADE,GAME ROOM/VIDEO GAMES (EXAMPLE CHUCKIE CHEESE)*"              5
##  9 "AUTO DEALERSHIP (CHEVY, FORD, BMW, MERCEDES, ETC.)"                 18
## 10 "AUTO REPAIR SHOP"                                                   82
## # i 257 more rows
```

Part 4

- Add a severity.score column which will be based on the rows data
- Group by area and get sum of severity scores for each area
- Display results

NOTE: Using original crime_data because crime_data_cleaned does not have the Weapon.Used.Cd column

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

severity_by_area <- crime_data %>%
  group_by(AREA.NAME) %>%
  summarise(Total_Severity_Score = sum(Severity.Score))

print(severity_by_area)
```

```
## # A tibble: 21 x 2
##    AREA.NAME   Total_Severity_Score
##    <chr>                      <dbl>
##  1 77th Street                 9439
##  2 Central                     9513
##  3 Devonshire                  4703
##  4 Foothill                    3969
##  5 Harbor                      5096
##  6 Hollenbeck                  4615
##  7 Hollywood                   6950
##  8 Mission                     4665
##  9 N Hollywood                 5789
## 10 Newton                      7047
## # i 11 more rows
```