MSDS - Datascience as a Field - Week 3

Marco Tulio Teixeira

2023-08-30

Project Description

Import, tidy and analyze the NYPD Shooting Incident dataset obtained. Be sure your project is reproducible and contains some visualization and analysis. You may use the data to do any analysis that is of interest to you. You should include at least two visualizations and one model. Be sure to identify any bias possible in the data and in your analysis.

Project Goals

Validate the number of domestic incidents by year, sex and age range.

Data Engineering Processes

1. Collect
2. Clean
3. Aggregate
4. Visualize
5. Analyze
6. Model

Installing and adding required librarie

```
# Data Collecting
url_base = "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
nypd_shooting_data_raw = read.csv(url_base, sep=",")
```

```
nyc_map <- get_map(location = "New York", maptype = "roadmap", )
```

Project Step 2: Tidy and Transform Your Data

Add to your Rmd document a summary of the data and clean up your dataset by changing appropriate variables to factor and date types and getting rid of any columns not needed. Show the summary of your data to be sure there is no missing data. If there is missing data, describe how you plan to handle it.

```r
# Understanding the data
summary(nypd_shooting_data_raw)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :  9953245   Length:27312       Length:27312       Length:27312
##  1st Qu.: 63860880   Class :character   Class :character   Class :character
##  Median : 90372218   Mode  :character   Mode  :character   Mode  :character
##  Mean   :120860536
##  3rd Qu.:188810230
##  Max.   :261190187
##
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:27312       Min.   :  1.00   Min.   :0.0000    Length:27312
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 68.00   Median :0.0000    Mode  :character
##                     Mean   : 65.64   Mean   :0.3269
##                     3rd Qu.: 81.00   3rd Qu.:0.0000
##                     Max.   :123.00   Max.   :2.0000
##                                      NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:27312       Length:27312            Length:27312
##  Class :character   Class :character        Class :character
##  Mode  :character   Mode  :character        Mode  :character
##
##
##
##
##    PERP_SEX           PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:27312       Length:27312       Length:27312       Length:27312
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_RACE           X_COORD_CD        Y_COORD_CD        Latitude
##  Length:27312       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000028   1st Qu.:182834   1st Qu.:40.67
##  Mode  :character   Median :1007731   Median :194487   Median :40.70
##                     Mean   :1009449   Mean   :208127   Mean   :40.74
##                     3rd Qu.:1016838   3rd Qu.:239518   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :10
##    Longitude         Lon_Lat
##  Min.   :-74.25   Length:27312
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :10
```

```r
head(nypd_shooting_data_raw)
```

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME     BORO LOC_OF_OCCUR_DESC PRECINCT
## 1    228798151 05/27/2021   21:30:00   QUEENS                       105
## 2    137471050 06/27/2014   17:40:00    BRONX                        40
## 3    147998800 11/21/2015   03:56:00   QUEENS                       108
## 4    146837977 10/09/2015   18:30:00    BRONX                        44
## 5     58921844 02/19/2009   22:58:00    BRONX                        47
## 6    219559682 10/21/2020   21:36:00 BROOKLYN                        81
##   JURISDICTION_CODE LOC_CLASSFCTN_DESC LOCATION_DESC STATISTICAL_MURDER_FLAG
## 1                 0                                                   false
## 2                 0                                                   false
## 3                 0                                                    true
## 4                 0                                                   false
## 5                 0                                                    true
## 6                 0                                                    true
##   PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP VIC_SEX       VIC_RACE
## 1                                          18-24       M          BLACK
## 2                                          18-24       M          BLACK
## 3                                          25-44       M          WHITE
## 4                                            <18       M WHITE HISPANIC
## 5          25-44        M     BLACK         45-64       M          BLACK
## 6                                          25-44       M          BLACK
##   X_COORD_CD Y_COORD_CD Latitude Longitude
## 1    1058925   180924.0 40.66296 -73.73084
## 2    1005028   234516.0 40.81035 -73.92494
## 3    1007668   209836.5 40.74261 -73.91549
## 4    1006537   244511.1 40.83778 -73.91946
## 5    1024922   262189.4 40.88624 -73.85291
## 6    1004234   186461.7 40.67846 -73.92795
##                                     Lon_Lat
## 1 POINT (-73.73083868899994 40.662964620000025)
## 2  POINT (-73.92494232599995 40.81035186300006)
## 3  POINT (-73.91549174199997 40.74260663300004)
## 4  POINT (-73.91945661499994 40.83778200300003)
## 5  POINT (-73.85290950899997 40.88623791800006)
## 6 POINT (-73.92795224099996 40.678456718000064)
```

Cleaning data

- Filtering JURISDICTION_CODE where is equals to 2 (Housing);
- Filtering VIC_SEX different the U (UNKNOW);
- Filtering PERP_AGE_GROUP different the 1020 (ERROR DATA);
- Create OCURR_DATE_TIME combining OCCUR_DATE and OCCUR_TIME
- Convert STATISTICAL_MURDER_FLAG to integer (0, 1)
- Removing unuseful columns: INCIDENT_KEY, OCCUR_TIME, PRECINCT, JURISDICTION_CODE, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION_DESC, X_COORD_CD, Y_COORD_CD, Lon_Lat

```r
nypd_shooting_cleaned = nypd_shooting_data_raw %>%
    filter(JURISDICTION_CODE == 2, VIC_SEX != "U", PERP_AGE_GROUP != "1020") %>%
    mutate(OCCUR_DATE = mdy(OCCUR_DATE)) %>%
```

```
    unite("OCCUR_DATE_TIME", c(OCCUR_DATE, OCCUR_TIME), sep = " ", na.rm = TRUE, remove = FALSE) %>%
    mutate(OCCUR_DATE_TIME = as_datetime(OCCUR_DATE_TIME), OCCUR_YEAR = year(OCCUR_DATE)) %>%
    mutate(STATISTICAL_MURDER_FLAG = as.integer(as.logical(STATISTICAL_MURDER_FLAG))) %>%
    select(-c(INCIDENT_KEY, OCCUR_TIME, PRECINCT, JURISDICTION_CODE, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DE

summary(nypd_shooting_cleaned)
```

```
##  OCCUR_DATE_TIME                   OCCUR_DATE            BORO
##  Min.   :2006-01-02 00:49:00.00   Min.   :2006-01-02   Length:4424
##  1st Qu.:2010-02-21 04:53:00.00   1st Qu.:2010-02-21   Class :character
##  Median :2013-11-01 09:00:00.00   Median :2013-10-31   Mode  :character
##  Mean   :2014-04-10 04:10:37.75   Mean   :2014-04-09
##  3rd Qu.:2019-01-01 04:44:30.00   3rd Qu.:2019-01-01
##  Max.   :2022-12-22 18:26:00.00   Max.   :2022-12-22
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP       PERP_SEX
##  Min.   :0.0000          Length:4424        Length:4424
##  1st Qu.:0.0000          Class :character   Class :character
##  Median :0.0000          Mode  :character   Mode  :character
##  Mean   :0.1612
##  3rd Qu.:0.0000
##  Max.   :1.0000
##   PERP_RACE          VIC_AGE_GROUP        VIC_SEX             VIC_RACE
##  Length:4424        Length:4424        Length:4424        Length:4424
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     Latitude       Longitude        OCCUR_YEAR
##  Min.   :40.57   Min.   :-74.17   Min.   :2006
##  1st Qu.:40.67   1st Qu.:-73.95   1st Qu.:2010
##  Median :40.70   Median :-73.93   Median :2013
##  Mean   :40.73   Mean   :-73.92   Mean   :2014
##  3rd Qu.:40.81   3rd Qu.:-73.90   3rd Qu.:2019
##  Max.   :40.89   Max.   :-73.75   Max.   :2022
```

```
head(nypd_shooting_cleaned)
```

```
##       OCCUR_DATE_TIME OCCUR_DATE      BORO STATISTICAL_MURDER_FLAG
## 1 2010-10-10 03:21:00 2010-10-10 MANHATTAN                       0
## 2 2008-11-09 20:13:00 2008-11-09  BROOKLYN                       0
## 3 2007-07-05 01:27:00 2007-07-05     BRONX                       0
## 4 2009-07-26 03:47:00 2009-07-26  BROOKLYN                       1
## 5 2012-05-13 12:34:00 2012-05-13  BROOKLYN                       0
## 6 2021-09-18 19:41:00 2021-09-18 MANHATTAN                       0
##   PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP VIC_SEX       VIC_RACE
## 1                                          25-44       M          BLACK
## 2        UNKNOWN        U   UNKNOWN         25-44       M BLACK HISPANIC
## 3        UNKNOWN        M   UNKNOWN         18-24       M          BLACK
## 4                                          18-24       M          BLACK
## 5          25-44        M     BLACK         18-24       M          BLACK
## 6          18-24        M     BLACK         25-44       M WHITE HISPANIC
```

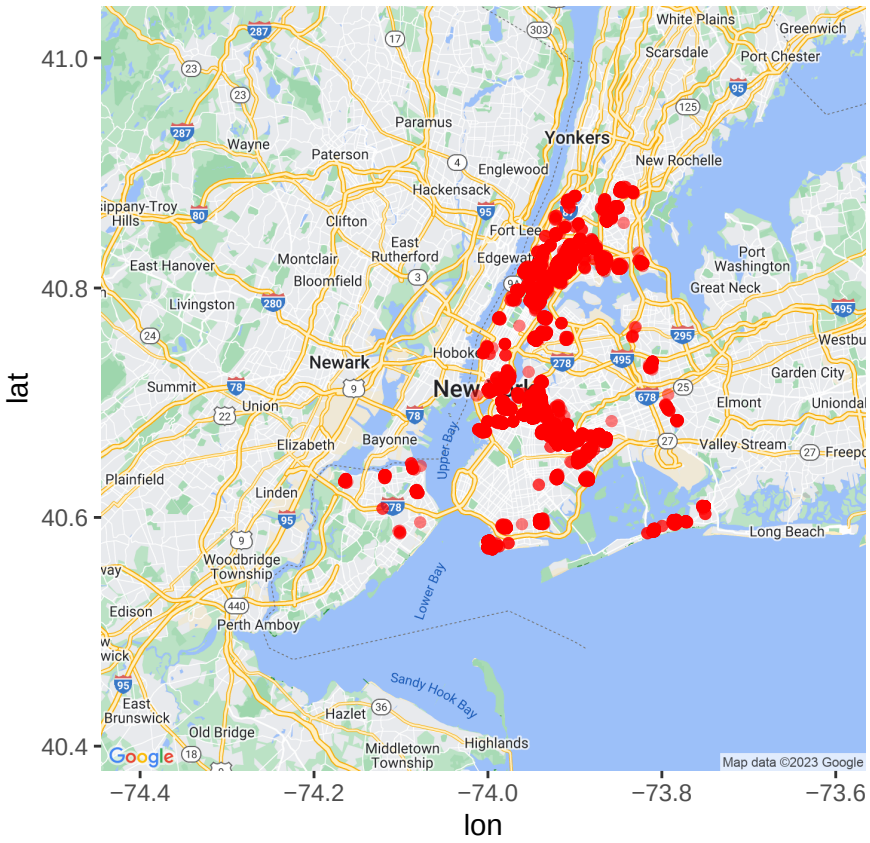
```
##   Latitude Longitude OCCUR_YEAR
## 1 40.79773 -73.94651       2010
## 2 40.68257 -73.98504       2008
## 3 40.88412 -73.84897       2007
## 4 40.57284 -73.99543       2009
## 5 40.66052 -73.88345       2012
## 6 40.79101 -73.94930       2021
```

Project Step 3: Add Visualizations and Analysis

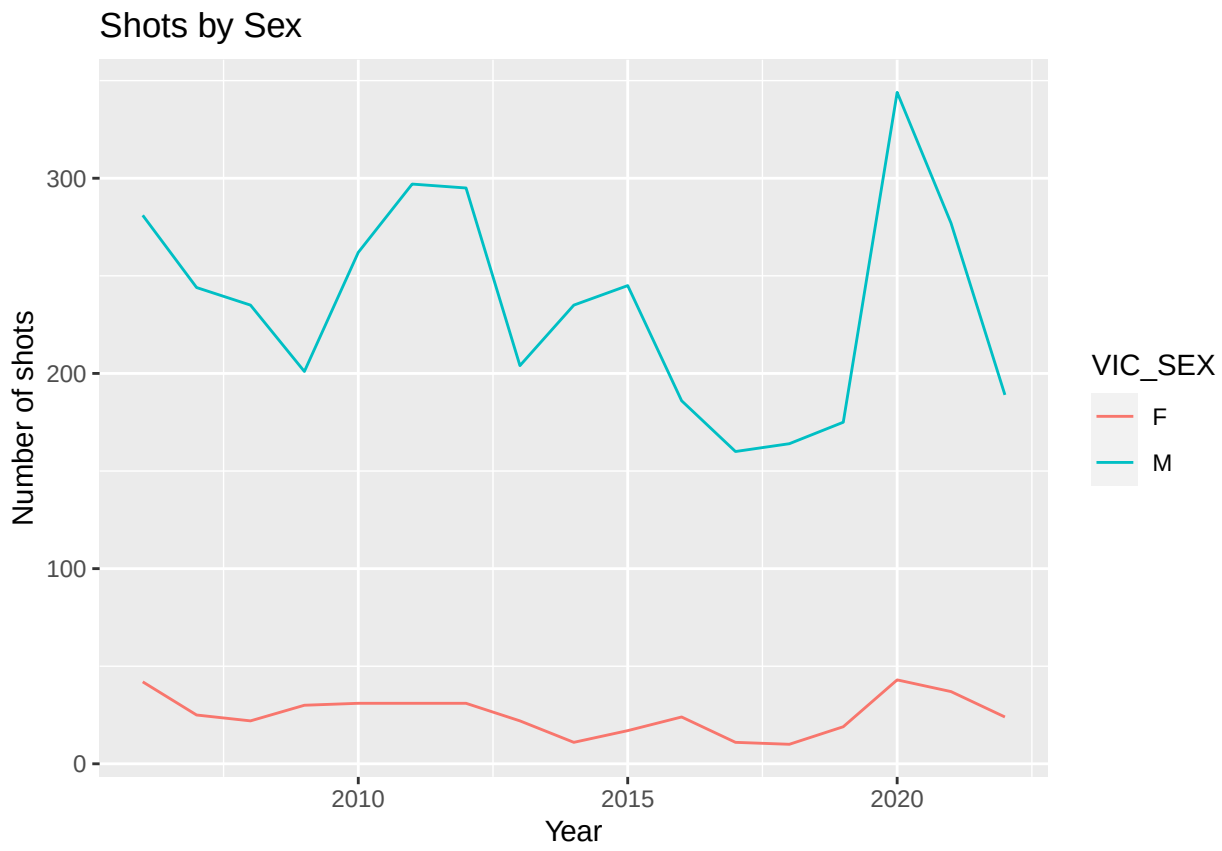
Add at least two different visualizations & some analysis to your Rmd. Does this raise additional questions that you should investigate?

```
# Create datasets based on multiple assumptions

# Shooting by Year and Sex
nypd_shooting_by_sex_year = nypd_shooting_cleaned %>%
  group_by(OCCUR_YEAR, VIC_SEX) %>%
  summarise(count_shoots = n(), .groups = 'drop')

nypd_shooting_by_murder_sex_year = nypd_shooting_cleaned %>%
  filter(STATISTICAL_MURDER_FLAG == 1) %>%
  group_by(OCCUR_YEAR, VIC_SEX) %>%
  summarise(count_shoots = n(), .groups = 'drop')

nypd_shooting_by_no_murder_sex_year = nypd_shooting_cleaned %>%
  filter(STATISTICAL_MURDER_FLAG == 0) %>%
  group_by(OCCUR_YEAR, VIC_SEX) %>%
  summarise(count_shoots = n(), .groups = 'drop')

nypd_shooting_by_male_age_range = nypd_shooting_cleaned %>%
  filter(!is.na(PERP_AGE_GROUP), PERP_AGE_GROUP != "", PERP_AGE_GROUP != "(null)" ) %>%
  filter(VIC_SEX == "M") %>%
  group_by(OCCUR_YEAR, PERP_AGE_GROUP) %>%
  summarise(count_shoots = n(), .groups = 'drop')

nypd_shooting_by_female_age_range = nypd_shooting_cleaned %>%
  filter(!is.na(PERP_AGE_GROUP), PERP_AGE_GROUP != "", PERP_AGE_GROUP != "(null)" ) %>%
  filter(VIC_SEX == "F") %>%
  group_by(OCCUR_YEAR, PERP_AGE_GROUP) %>%
  summarise(count_shoots = n(), .groups = 'drop')
```

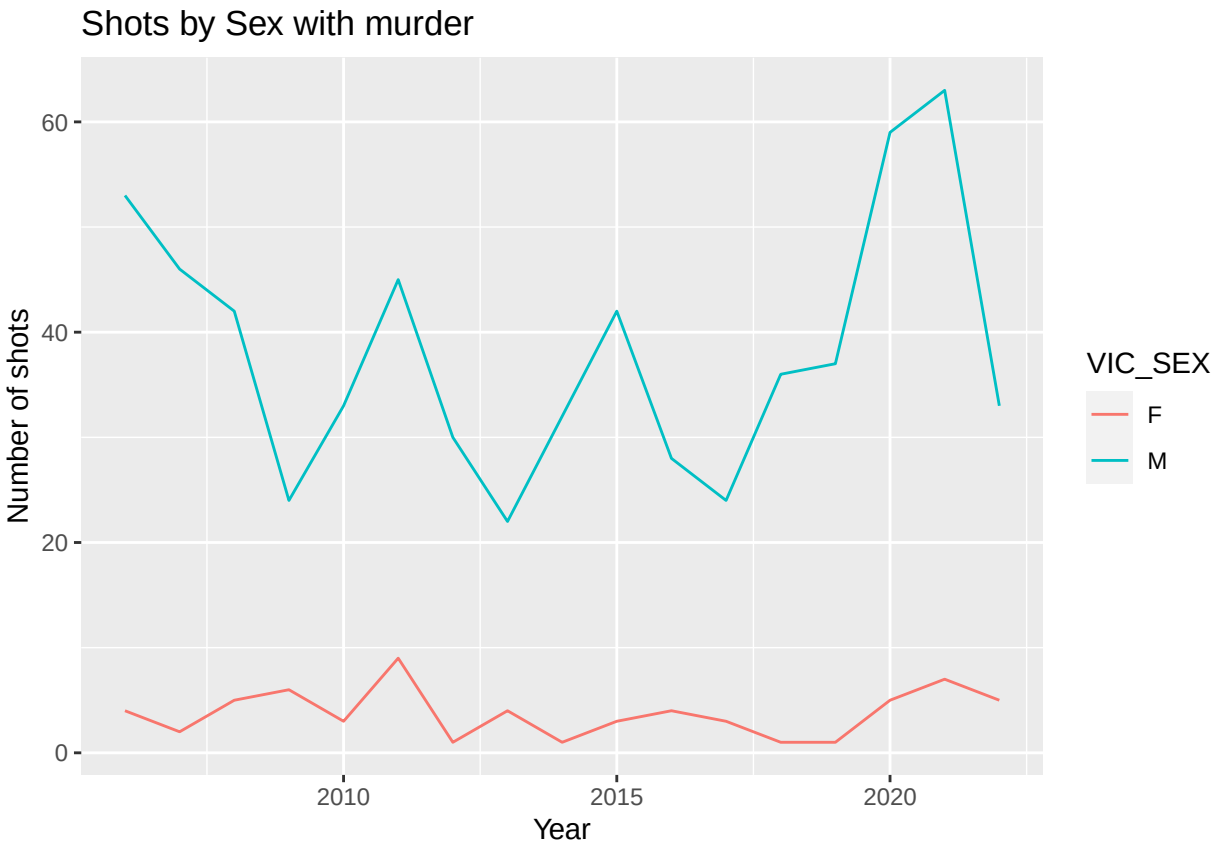
```
# Show occurrencies on a map
p = ggmap(nyc_map)
p + geom_point(data=nypd_shooting_cleaned, aes(x=Longitude, y=Latitude), color="red", alpha=0.5)
```



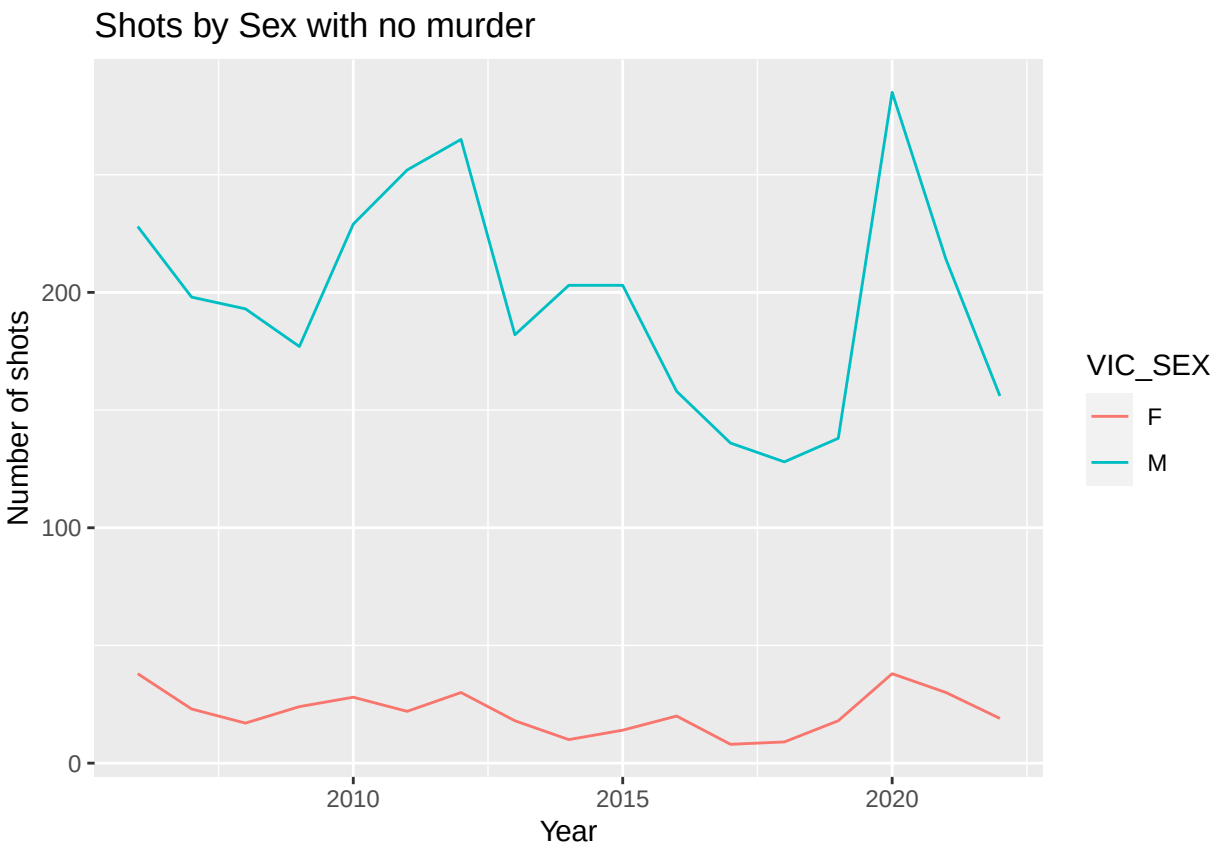

```
# plot
options(repr.plot.width = 20, repr.plot.height = 8)

nypd_shooting_by_sex_year %>%
  ggplot(aes(x=OCCUR_YEAR, y=count_shoots, group=VIC_SEX, color=VIC_SEX)) +
    geom_line() +
    ggtitle("Shots by Sex") +
    ylab("Number of shots") +
    xlab("Year")
```




```
nypd_shooting_by_murder_sex_year %>%
  ggplot(aes(x=OCCUR_YEAR, y=count_shoots, group=VIC_SEX, color=VIC_SEX)) +
    geom_line() +
    ggtitle("Shots by Sex with murder") +
    ylab("Number of shots") +
    xlab("Year")
```

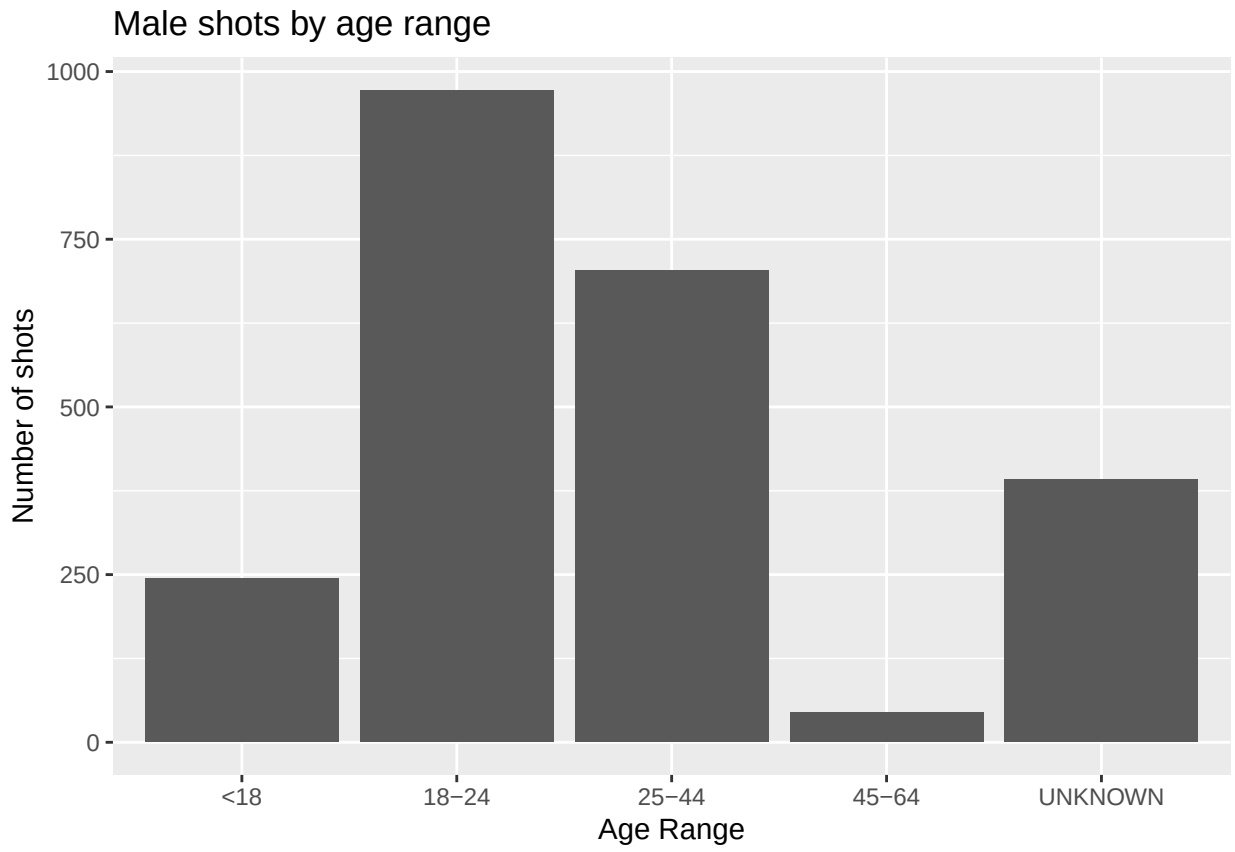



```
nypd_shooting_by_no_murder_sex_year %>%
  ggplot(aes(x=OCCUR_YEAR, y=count_shoots, group=VIC_SEX, color=VIC_SEX)) +
    geom_line() +
    ggtitle("Shots by Sex with no murder") +
    ylab("Number of shots") +
    xlab("Year")
```

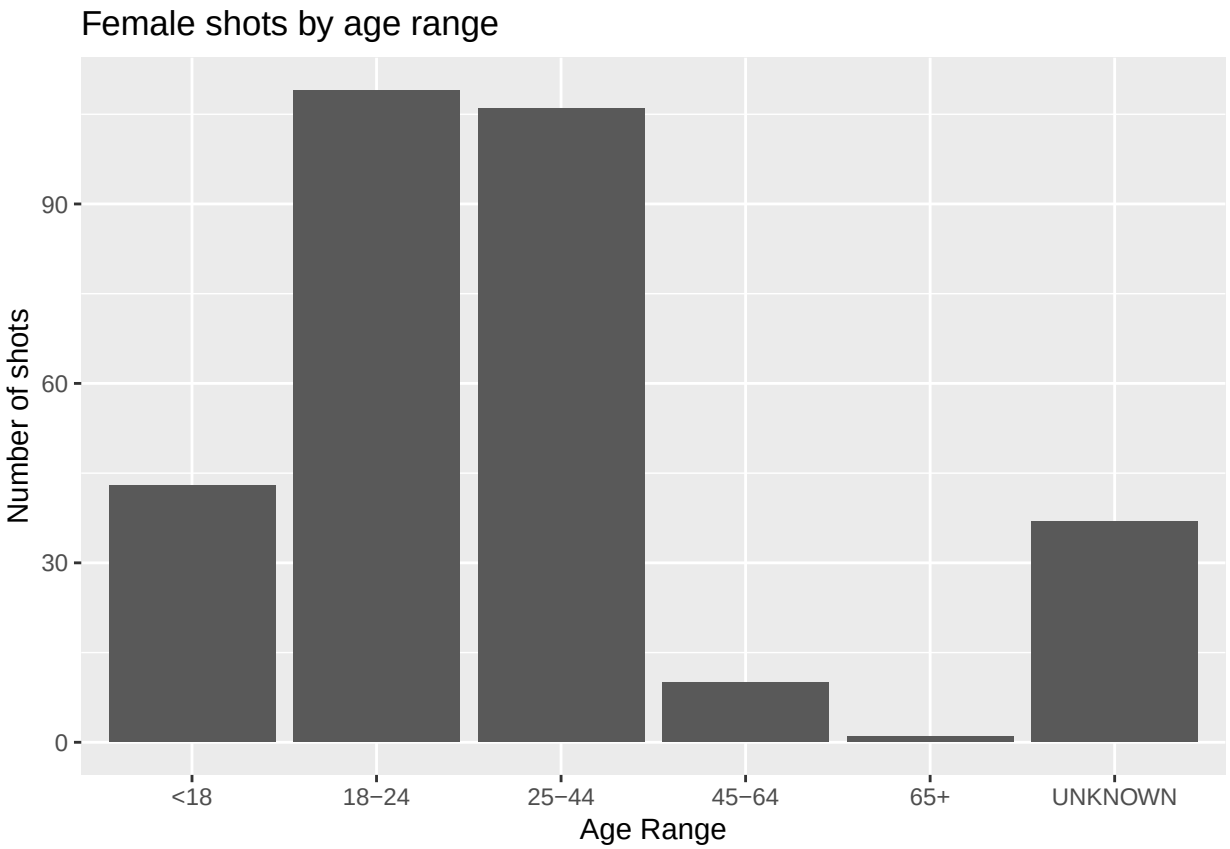


Incidents by age range

```
nypd_shooting_by_male_age_range %>%
  ggplot(aes(x=PERP_AGE_GROUP, y=count_shoots)) +
    geom_bar(stat="identity") +
    ggtitle("Male shots by age range") +
    ylab("Number of shots") +
    xlab("Age Range")
```




```
nypd_shooting_by_female_age_range %>%
  ggplot(aes(x=PERP_AGE_GROUP, y=count_shoots)) +
    geom_bar(stat="identity") +
    ggtitle("Female shots by age range") +
    ylab("Number of shots") +
    xlab("Age Range")
```



Project Step 4: Add Bias Identification

Write the conclusion to your project report and include any possible sources of bias. Be sure to identify what your personal bias might be and how you have mitigated that.

Answers and Conclusion

- The initial year provided by the dataset is 2006;
- The number of incidents is higher from 2020;
- This dataset doesn't have information about the reasons or motive of the incidents;
- At the first sigh we can conclude that the number of domestic shots incidents are greater for victims categorized as the sex M (Male);
- The number of incidents are higher in the age ranges of 18-24 and 25-44 for both Female and Male;
- As bias indentification:
  - Looking into the raw dataset we can assume that newest incidents are well reported and the information provided at this period is complete.
  - Using only this data can lead to misinformation and should be interesting to aggregate this data with other data sources like social classes, poverty and so.