R Markdown Final Part 1- NYSHOOTING data

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(lubridate)
library(hms)
```

```
##
## Attaching package: 'hms'
##
## The following object is masked from 'package:lubridate':
##
##     hms
```

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

This is a Notebook for Week 3 of Data Science as a Field.

Project Step 1: Start an Rmd Document

Start an Rmd document that describes and imports the shooting project dataset in a reproducible manner.

The first dataset is from catalog.data.gov and is called NYPD Shooting Incident Data (Historic). I pulled this data 12/18/2023 from here. According to the site, this dataset: “This is a breakdown of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year. This data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning before being posted on the NYPD website. Each record represents a shooting incident in NYC and includes information about the event, the location and time of occurrence. In addition, information related to suspect and victim demographics is also included.”

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

```
## Rows: 27312 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.
```

Project Step 2: Tidy and Transform Your Data

Step 2: 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.

One piece of this is that we haven't been told what analysis we are actually doing so determining what is unnecassary at this point is sort of impossible. Dropping Lat long because she did in the lecture.

```
summary(NYPD_shootings)
```

```
##   INCIDENT_KEY        OCCUR_DATE          OCCUR_TIME           BORO
##  Min.   :  9953245   Length:27312       Length:27312      Length:27312
##  1st Qu.: 63860880   Class :character   Class1:hms        Class :character
##  Median : 90372218   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :120860536                      Mode  :numeric
##  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       Mode :logical           Length:27312
##  Class :character   FALSE:22046             Class :character
##  Mode  :character   TRUE :5266              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
```

```
head(NYPD_shootings)
```

```
## # A tibble: 6 x 21
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO     LOC_OF_OCCUR_DESC PRECINCT
##          <dbl> <chr>      <time>     <chr>    <chr>                <dbl>
## 1    228798151 05/27/2021 21:30      QUEENS   <NA>                   105
## 2    137471050 06/27/2014 17:40      BRONX    <NA>                    40
## 3    147998800 11/21/2015 03:56      QUEENS   <NA>                   108
## 4    146837977 10/09/2015 18:30      BRONX    <NA>                    44
## 5     58921844 02/19/2009 22:58      BRONX    <NA>                    47
## 6    219559682 10/21/2020 21:36      BROOKLYN <NA>                    81
## # 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>
```

```
NYPD_shootings$OCCUR_DATE <- as.Date(NYPD_shootings$OCCUR_DATE, format="%m/%d/%Y")
NYPD_shootings$BORO <- as.factor(NYPD_shootings$BORO)
NYPD_shootings$PRECINCT <- as.factor(NYPD_shootings$PRECINCT)
NYPD_shootings$JURISDICTION_CODE <- as.factor(NYPD_shootings$JURISDICTION_CODE)
NYPD_shootings$LOC_CLASSFCTN_DESC <- as.factor(NYPD_shootings$LOC_CLASSFCTN_DESC)
NYPD_shootings$PERP_AGE_GROUP <- as.factor(NYPD_shootings$PERP_AGE_GROUP)
NYPD_shootings$PERP_SEX <- as.factor(NYPD_shootings$PERP_SEX)
NYPD_shootings$PERP_RACE <- as.factor(NYPD_shootings$PERP_RACE)
NYPD_shootings$VIC_AGE_GROUP <- as.factor(NYPD_shootings$VIC_AGE_GROUP)
NYPD_shootings$VIC_SEX <- as.factor(NYPD_shootings$VIC_SEX)
NYPD_shootings$VIC_RACE <- as.factor(NYPD_shootings$VIC_RACE)

NYPD_shootings$Lon_Lat <- NULL
NYPD_shootings$X_COORD_CD <- NULL
NYPD_shootings$Y_COORD_CD  <- NULL
NYPD_shootings$Latitude <- NULL
NYPD_shootings$Longitude <- NULL

summary(NYPD_shootings)
```

```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Length:27312
##  1st Qu.: 63860880   1st Qu.:2009-07-18   Class1:hms
##  Median : 90372218   Median :2013-04-29   Class2:difftime
##  Mean   :120860536   Mean   :2014-01-06   Mode  :numeric
##  3rd Qu.:188810230   3rd Qu.:2018-10-15
##  Max.   :261190187   Max.   :2022-12-31
##
##             BORO       LOC_OF_OCCUR_DESC     PRECINCT     JURISDICTION_CODE
##  BRONX        : 7937   Length:27312       75     : 1557   0   :22809
##  BROOKLYN     :10933   Class :character   73     : 1452   1   :   74
##  MANHATTAN    : 3572   Mode  :character   67     : 1216   2   : 4427
##  QUEENS       : 4094                      44     : 1020   NA's:    2
##  STATEN ISLAND:  776                      79     : 1012
##                                           47     :  953
##                                           (Other):20102
##   LOC_CLASSFCTN_DESC LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  STREET    : 1103    Length:27312       Mode :logical           18-24  :6222
##  HOUSING   :  280    Class :character   FALSE:22046             25-44  :5687
##  DWELLING  :  127    Mode  :character   TRUE :5266              UNKNOWN:3148
##  COMMERCIAL:  100                                               <18    :1591
##  OTHER     :   31                                               (null) : 640
##  (Other)   :   75                                               (Other): 680
##  NA's      :25596                                               NA's   :9344
##    PERP_SEX                 PERP_RACE     VIC_AGE_GROUP   VIC_SEX
##  (null):  640   BLACK         :11432   <18    : 2839   F: 2615
##  F     :  424   WHITE HISPANIC: 2341   1022   :    1   M:24686
##  M     :15439   UNKNOWN       : 1836   18-24  :10086   U:   11
##  U     : 1499   BLACK HISPANIC: 1314   25-44  :12281
##  NA's  : 9310   (null)        :  640   45-64  : 1863
##                 (Other)       :  439   65+    :  181
##                 NA's          : 9310   UNKNOWN:   61
##                             VIC_RACE
##  AMERICAN INDIAN/ALASKAN NATIVE:   10
##  ASIAN / PACIFIC ISLANDER      :  404
##  BLACK                         :19439
##  BLACK HISPANIC                : 2646
##  UNKNOWN                       :   66
##  WHITE                         :  698
##  WHITE HISPANIC                : 4049
```

There is a fair amount of missing data, and there is also a lot of 'UNKNOWN' data. Without knowing what question I'm trying to answer, I will probably leave the missing data in. I expect missing in this case is not randomly missing, so there could be an important insights here that would be missed otherwise.

One thing that does immediately need to be fixed is that in the PERP_RACE column we have both (null) and NA data. Before continuing I will make all the (null) into NAs.

```
NYPD_shootings$PERP_RACE[NYPD_shootings$PERP_RACE == "(null)"] <- NA
summary(NYPD_shootings$PERP_RACE)
```

```
##                         (null) AMERICAN INDIAN/ALASKAN NATIVE
##                              0                              2
##       ASIAN / PACIFIC ISLANDER                          BLACK
##                            154                          11432
##                 BLACK HISPANIC                        UNKNOWN
##                           1314                           1836
##                          WHITE                 WHITE HISPANIC
##                             283                           2341
##                            NA's
##                            9950
```

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?

Temporal Analysis of NYPD Shooting Incident Data

This section focuses on the temporal analysis of the NYPD Shooting Incident Data to uncover patterns and insights related to the timing of shooting incidentsI'm going to do two graphs, one investigating frequency over time and one frequency of time of day.

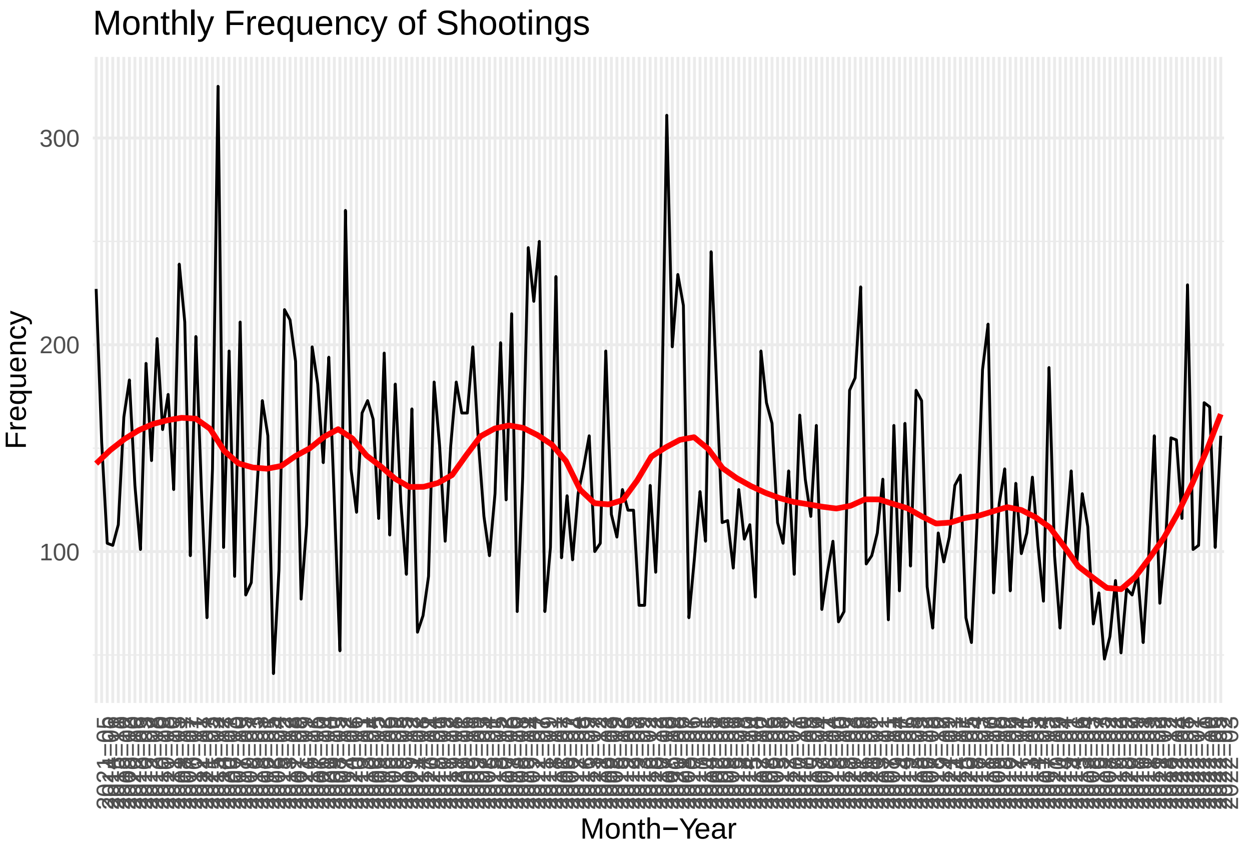
```r
# Convert MonthYear to an ordered factor
NYPD_shootings$MonthYear <- format(NYPD_shootings$OCCUR_DATE, "%Y-%m")

NYPD_shootings$MonthYear <- factor(NYPD_shootings$MonthYear, levels = unique(NYPD_shootings$MonthYear))


# Group and summarize data
monthly_counts <- NYPD_shootings %>%
  group_by(MonthYear) %>%
  summarise(Frequency = n())

# Plotting
ggplot(monthly_counts, aes(x = MonthYear, y = Frequency, group = 1)) +
  geom_line() +
  geom_smooth(method = "loess", span = 0.2, se = FALSE, color = "red") +
  labs(title = "Monthly Frequency of Shootings",
       x = "Month-Year",
       y = "Frequency") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

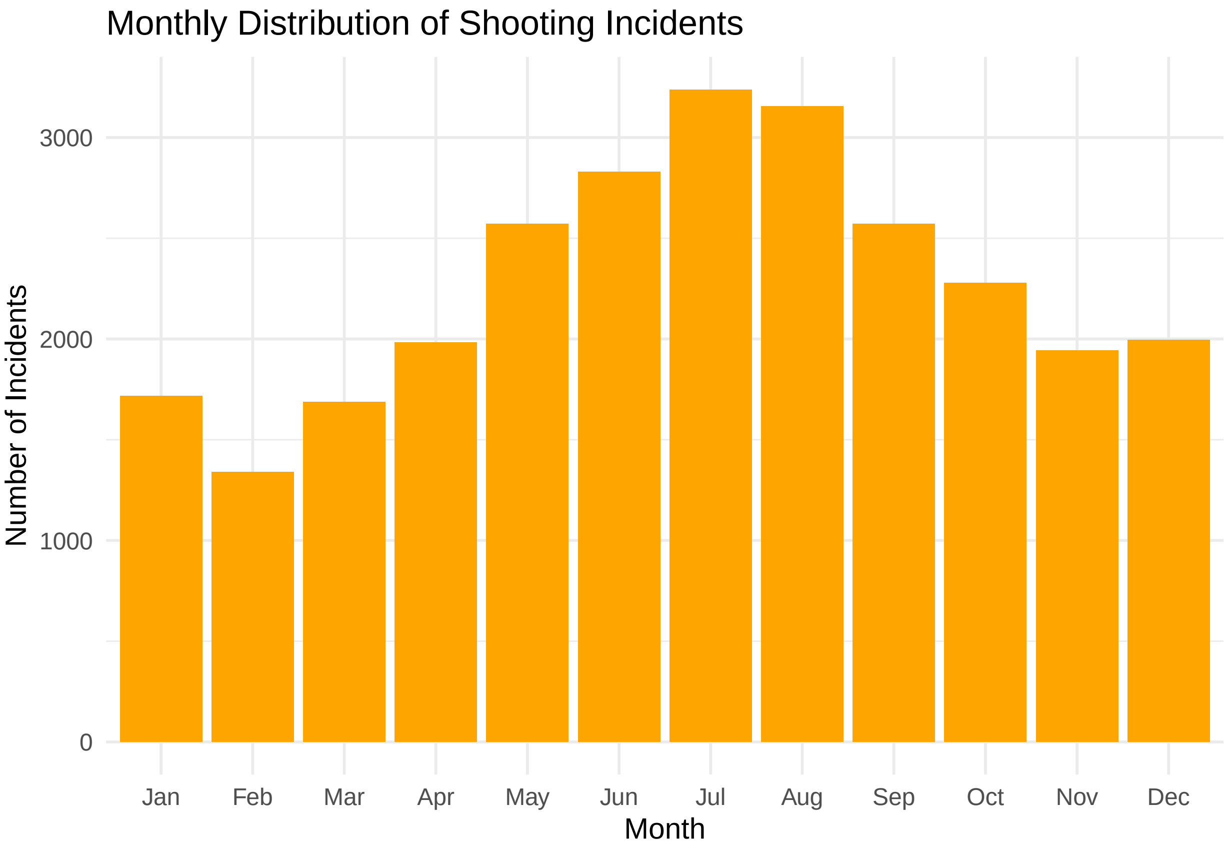
```
## `geom_smooth()` using formula = 'y ~ x'
```



Looking at this graph I have a few follow ups, what happened in 2021 and 2020? 2012 had a sharp decrease over the year in shooting, 2020 a sharp increase. I'd want to go investigate how the data was collected and if anything changed data wise before assuming both of these trends reflect real-world trends.

Now let's look at periodicity, both by month of year and time of day.

```
# Monthly Seasonality
NYPD_shootings %>%
    count(month = month(OCCUR_DATE, label = TRUE)) %>%
    ggplot(aes(x = month, y = n)) +
    geom_bar(stat = "identity", fill='orange') +
    theme_minimal() +
    labs(title = "Monthly Distribution of Shooting Incidents", x = "Month", y = "Number of Incidents")
```



There is a definite trend that more shootings happen in the summer months. I believe this is a well researched and established trend.

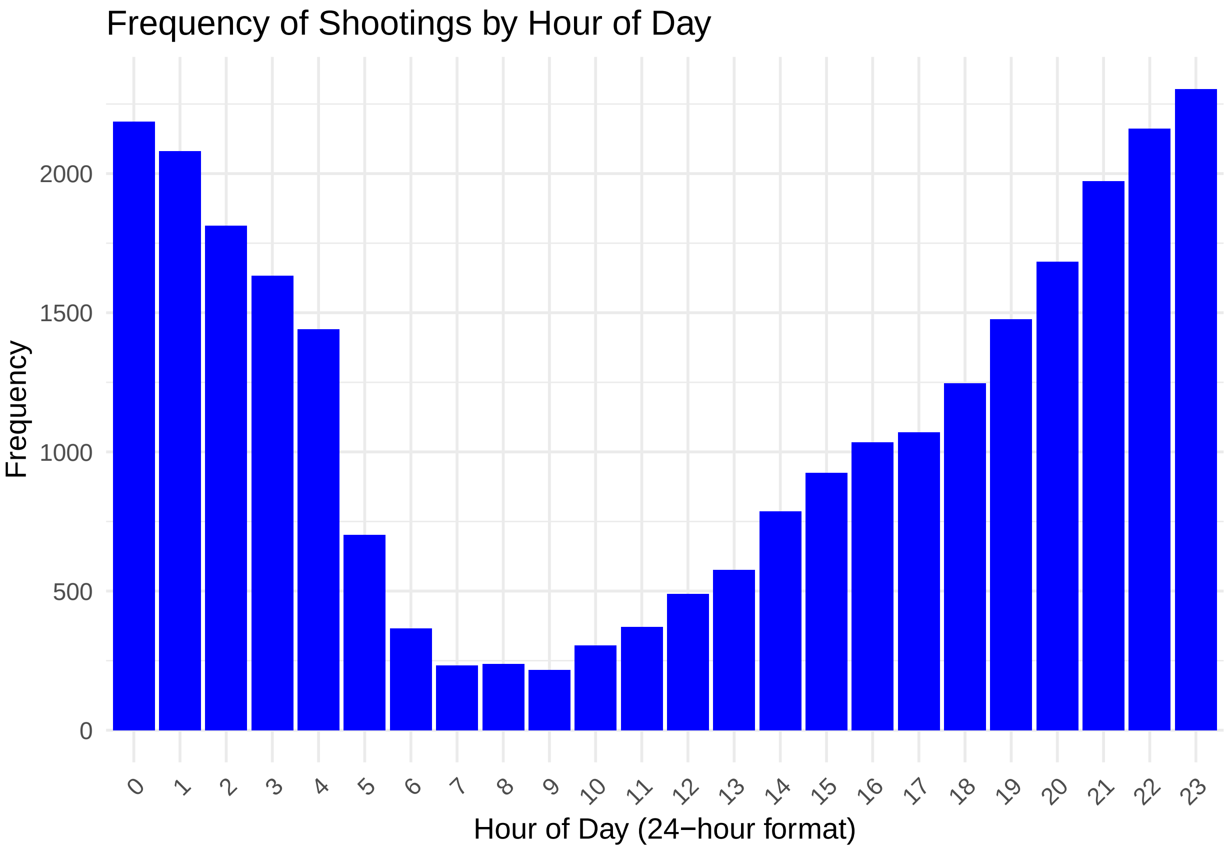
```r
# Convert hms to period
NYPD_shootings$TimePeriod <- as.period(NYPD_shootings$OCCUR_TIME)

# Extract the hour component
NYPD_shootings$HourOfDay <- hour(NYPD_shootings$TimePeriod)

# Count the number of occurrences by hour
hourly_counts <- table(NYPD_shootings$HourOfDay)

# Convert the frequency table to a data frame for plotting
hourly_counts_df <- as.data.frame(hourly_counts)

# Plotting with ggplot2
library(ggplot2)
ggplot(hourly_counts_df, aes(x = Var1, y = Freq)) +
  geom_bar(stat = "identity", fill = "blue") +
  labs(title = "Frequency of Shootings by Hour of Day",
       x = "Hour of Day (24-hour format)",
       y = "Frequency") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) # Adjusting x-axis labels for better readabi
```



There are also definite time of day trends, between 7 and 10 AM in the morning there are the fewest shootings, which then peak around midnight.

Predictive Power of Time

All three of these variables- year, month, and hour- seem to have strong relationships with out data. I'm going to build a regression model with these variables utilizing stepwise regression to ensure each is important.

```
# Extracting year, month, and hour
shooting_data <- NYPD_shootings %>%
                mutate(year = as.factor(year(OCCUR_DATE)),
                       month = as.factor(month(OCCUR_DATE, label = TRUE)),
                       hour = as.factor(hour(OCCUR_TIME)))

# Count incidents per year, month, and hour
shooting_data_grouped <- shooting_data %>%
                        group_by(year, month, hour) %>%
                        summarise(n = n())
```

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

```
# Initial model with all predictors
initial_model <- lm(n ~ year + month + hour, data = shooting_data_grouped)

# Stepwise model selection
stepwise_model <- stepAIC(initial_model, direction = "both")
```

```
## Start:  AIC=12282.69
## n ~ year + month + hour
##
##         Df Sum of Sq    RSS   AIC
## <none>                71724 12283
## - year  16      8272  79996 12725
## - month 11      9977  81701 12826
## - hour  23     44737 116461 14342
```

For our stepwise model, we start with all three time predictors and then try removing each one. As you can see the AIC when each variable is dropped increaces, indicating the best model includes all 3 variables.

```
# Summary of the final model
summary(stepwise_model)
```

```
##
## Call:
## lm(formula = n ~ year + month + hour, data = shooting_data_grouped)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2353  -2.6015  -0.4652   1.9040  28.3965
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.33493    0.37654  32.759  < 2e-16 ***
## year2007    -0.50670    0.35688  -1.420 0.155730
## year2008    -0.30686    0.35453  -0.866 0.386789
## year2009    -0.79684    0.35654  -2.235 0.025473 *
## year2010    -0.34945    0.35982  -0.971 0.331504
## year2011    -0.32626    0.35693  -0.914 0.360721
## year2012    -1.18319    0.35654  -3.319 0.000912 ***
## year2013    -2.54118    0.36023  -7.054 2.01e-12 ***
## year2014    -2.08226    0.35876  -5.804 6.94e-09 ***
## year2015    -2.27086    0.35620  -6.375 2.02e-10 ***
## year2016    -3.06243    0.36053  -8.494  < 2e-16 ***
## year2017    -3.95903    0.36696 -10.789  < 2e-16 ***
## year2018    -4.01328    0.36689 -10.939  < 2e-16 ***
## year2019    -3.90781    0.36649 -10.663  < 2e-16 ***
## year2020    -0.30264    0.35722  -0.847 0.396925
## year2021    -0.21909    0.35289  -0.621 0.534738
## year2022    -1.20007    0.35488  -3.382 0.000727 ***
## month.L      1.83674    0.21545   8.525  < 2e-16 ***
## month.Q     -4.01737    0.21420 -18.755  < 2e-16 ***
## month.C     -1.26675    0.21448  -5.906 3.77e-09 ***
## month^4      2.44666    0.21483  11.389  < 2e-16 ***
## month^5      0.37706    0.21607   1.745 0.081048 .
## month^6     -0.07147    0.21582  -0.331 0.740549
## month^7     -0.45942    0.21571  -2.130 0.033246 *
## month^8      0.07185    0.21515   0.334 0.738429
## month^9      0.20206    0.21463   0.941 0.346533
## month^10     0.52988    0.21492   2.465 0.013723 *
## month^11    -0.16070    0.21361  -0.752 0.451913
## hour1       -0.40914    0.40725  -1.005 0.315120
## hour2       -1.76721    0.40673  -4.345 1.43e-05 ***
## hour3       -2.58355    0.40777  -6.336 2.60e-10 ***
## hour4       -3.54195    0.40724  -8.697  < 2e-16 ***
## hour5       -7.10376    0.41494 -17.120  < 2e-16 ***
## hour6       -8.54597    0.43692 -19.559  < 2e-16 ***
## hour7       -9.19605    0.46734 -19.677  < 2e-16 ***
## hour8       -8.94158    0.46855 -19.084  < 2e-16 ***
## hour9       -9.19683    0.47133 -19.512  < 2e-16 ***
## hour10      -8.87110    0.44129 -20.103  < 2e-16 ***
## hour11      -8.44571    0.43787 -19.288  < 2e-16 ***
## hour12      -7.98910    0.42507 -18.795  < 2e-16 ***
## hour13      -7.51759    0.42371 -17.742  < 2e-16 ***
## hour14      -6.69394    0.41153 -16.266  < 2e-16 ***
## hour15      -5.88116    0.41265 -14.252  < 2e-16 ***
## hour16      -5.57613    0.40830 -13.657  < 2e-16 ***
## hour17      -5.44860    0.40673 -13.396  < 2e-16 ***
## hour18      -4.54279    0.40673 -11.169  < 2e-16 ***
## hour19      -3.47604    0.40622  -8.557  < 2e-16 ***
## hour20      -2.45994    0.40572  -6.063 1.45e-09 ***
## hour21      -0.98500    0.40673  -2.422 0.015486 *
## hour22      -0.10033    0.40572  -0.247 0.804693
## hour23       0.68215    0.40673   1.677 0.093585 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.087 on 4293 degrees of freedom
## Multiple R-squared:  0.4625, Adjusted R-squared:  0.4563
## F-statistic: 73.89 on 50 and 4293 DF,  p-value: < 2.2e-16
```

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.

My project centered around when police shootings by the NYPD occur. I noticed two major trend in the years 2012 and 2020. One of the first things that I think of is that those are election years, I probably have some bias that I believe politics and policy can influence violence. Those are assumptions I would have to research and try to disprove before drawing any conclusions. In the furture I would take actions to mitigate this by looking into the subject matter and try and replicate the patterns here in other areas.