NYPD Shooting data analysis

Dataset: this dataset includes a breakdown of all recorded shooting incidents that occured in NYC's borough's. This data was published by the City of New York on the GSA's data.gov .

The dataset includes time information about the incident, location descriptions along with boroughs, information about the perps if known and the victims

With this dataset I am primarly intrested in a couple of items. I initially want to initially look at data distributions over time. Starting with the highest resolultion and the reducing it over time to see if there are any patterns we can identify. With that we'll look at month by month and year by year and finally seasonalilty. Next we'll look at any incident characteristics we can compare like the locations, perp and vic age group. Finally we'll see if we can predict any patterns based on the historical data we have.

```
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
## Warning: package 'tidyr' was built under R version 4.3.2
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
```

```
data_path <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv"

NYPD_data <- read_csv(data_path)
```

```
## 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.
```

```
NYPD_data
```

```
## # A tibble: 28,562 × 21
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO      LOC_OF_OCCUR_DESC PRECINCT
##           <dbl> <chr>      <time>     <chr>     <chr>                <dbl>
##  1    244608249 05/05/2022 00:10      MANHATTAN INSIDE                  14
##  2    247542571 07/04/2022 22:20      BRONX     OUTSIDE                 48
##  3     84967535 05/27/2012 19:35      QUEENS    <NA>                   103
##  4    202853370 09/24/2019 21:00      BRONX     <NA>                    42
##  5     27078636 02/25/2007 21:00      BROOKLYN  <NA>                    83
##  6    230311078 07/01/2021 23:07      MANHATTAN <NA>                    23
##  7    229224142 06/07/2021 19:55      QUEENS    <NA>                   113
##  8    231246224 07/22/2021 01:47      BROOKLYN  <NA>                    77
##  9    228559720 05/22/2021 18:39      BRONX     <NA>                    48
## 10    238210279 12/22/2021 23:17      BRONX     <NA>                    49
## # ℹ 28,552 more rows
## # 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>
```

```
glimpse(NYPD_data)
```

```
## Rows: 28,562
## Columns: 21
## $ INCIDENT_KEY           <dbl> 244608249, 247542571, 84967535, 202853370, 270…
## $ OCCUR_DATE             <chr> "05/05/2022", "07/04/2022", "05/27/2012", "09/…
## $ OCCUR_TIME             <time> 00:10:00, 22:20:00, 19:35:00, 21:00:00, 21:00…
## $ BORO                   <chr> "MANHATTAN", "BRONX", "QUEENS", "BRONX", "BROO…
## $ LOC_OF_OCCUR_DESC      <chr> "INSIDE", "OUTSIDE", NA, NA, NA, NA, NA, NA, N…
## $ PRECINCT               <dbl> 14, 48, 103, 42, 83, 23, 113, 77, 48, 49, 73, …
## $ JURISDICTION_CODE      <dbl> 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ LOC_CLASSFCTN_DESC     <chr> "COMMERCIAL", "STREET", NA, NA, NA, NA, NA, NA…
## $ LOCATION_DESC          <chr> "VIDEO STORE", "(null)", NA, NA, NA, "MULTI DW…
## $ STATISTICAL_MURDER_FLAG <lgl> TRUE, TRUE, FALSE, FALSE, FALSE, FALSE, TRUE, …
## $ PERP_AGE_GROUP         <chr> "25-44", "(null)", NA, "25-44", "25-44", NA, N…
## $ PERP_SEX               <chr> "M", "(null)", NA, "M", "M", NA, NA, NA, NA, "…
## $ PERP_RACE              <chr> "BLACK", "(null)", NA, "UNKNOWN", "BLACK", NA,…
## $ VIC_AGE_GROUP          <chr> "25-44", "18-24", "18-24", "25-44", "25-44", "…
## $ VIC_SEX                <chr> "M", "M", "M", "M", "M", "M", "M", "M", "M", "…
## $ VIC_RACE               <chr> "BLACK", "BLACK", "BLACK", "BLACK", "BLACK", "…
## $ X_COORD_CD             <dbl> 986050, 1016802, 1048632, 1014493, 1009149, 99…
## $ Y_COORD_CD             <dbl> 214231.0, 250581.0, 198262.0, 242565.0, 190104…
## $ Latitude               <dbl> 40.75469, 40.85440, 40.71063, 40.83242, 40.688…
## $ Longitude              <dbl> -73.99350, -73.88233, -73.76777, -73.89071, -7…
## $ Lon_Lat                <chr> "POINT (-73.9935 40.754692)", "POINT (-73.8823…
```

```
head(NYPD_data)
```

```
## # A tibble: 6 × 21
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO      LOC_OF_OCCUR_DESC PRECINCT
##          <dbl> <chr>      <time>     <chr>     <chr>                <dbl>
## 1    244608249 05/05/2022 00:10      MANHATTAN INSIDE                  14
## 2    247542571 07/04/2022 22:20      BRONX     OUTSIDE                 48
## 3     84967535 05/27/2012 19:35      QUEENS    <NA>                   103
## 4    202853370 09/24/2019 21:00      BRONX     <NA>                    42
## 5     27078636 02/25/2007 21:00      BROOKLYN  <NA>                    83
## 6    230311078 07/01/2021 23:07      MANHATTAN <NA>                    23
## # 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_data <- NYPD_data %>%
  mutate(
    OCCUR_DATE = as.Date(OCCUR_DATE, format = "%m/%d/%Y"),
    OCCUR_TIME = hms::as_hms(OCCUR_TIME)
  )
NYPD_data
```

```
## # A tibble: 28,562 × 21
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO      LOC_OF_OCCUR_DESC PRECINCT
##           <dbl> <date>     <time>     <chr>     <chr>                <dbl>
##  1    244608249 2022-05-05 00:10      MANHATTAN INSIDE                  14
##  2    247542571 2022-07-04 22:20      BRONX     OUTSIDE                 48
##  3     84967535 2012-05-27 19:35      QUEENS    <NA>                   103
##  4    202853370 2019-09-24 21:00      BRONX     <NA>                    42
##  5     27078636 2007-02-25 21:00      BROOKLYN  <NA>                    83
##  6    230311078 2021-07-01 23:07      MANHATTAN <NA>                    23
##  7    229224142 2021-06-07 19:55      QUEENS    <NA>                   113
##  8    231246224 2021-07-22 01:47      BROOKLYN  <NA>                    77
##  9    228559720 2021-05-22 18:39      BRONX     <NA>                    48
## 10    238210279 2021-12-22 23:17      BRONX     <NA>                    49
## # ℹ 28,552 more rows
## # 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>
```

Summary Statistics and Metadata

```
summary(NYPD_data)
```

```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME          BORO
##  Min.   :  9953245   Min.   :2006-01-01   Length:28562      Length:28562
##  1st Qu.: 65439914   1st Qu.:2009-09-04   Class1:hms        Class :character
##  Median : 92711254   Median :2013-09-20   Class2:difftime   Mode  :character
##  Mean   :127405824   Mean   :2014-06-07   Mode  :numeric
##  3rd Qu.:203131993   3rd Qu.:2019-09-29
##  Max.   :279758069   Max.   :2023-12-29
##
##  LOC_OF_OCCUR_DESC     PRECINCT     JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:28562       Min.   :  1.0   Min.   :0.0000    Length:28562
##  Class :character   1st Qu.: 44.0   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 67.0   Median :0.0000    Mode  :character
##                     Mean   : 65.5   Mean   :0.3219
##                     3rd Qu.: 81.0   3rd Qu.:0.0000
##                     Max.   :123.0   Max.   :2.0000
##                                     NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:28562       Mode :logical           Length:28562
##  Class :character   FALSE:23036             Class :character
##  Mode  :character   TRUE :5526              Mode  :character
##
##
##
##
##    PERP_SEX           PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:28562       Length:28562       Length:28562       Length:28562
##  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:28562       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000068   1st Qu.:182912   1st Qu.:40.67
##  Mode  :character   Median :1007772   Median :194901   Median :40.70
##                     Mean   :1009424   Mean   :208380   Mean   :40.74
##                     3rd Qu.:1016807   3rd Qu.:239814   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :59
##    Longitude         Lon_Lat
##  Min.   :-74.25   Length:28562
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :59
```

By Borough stats

```
NYPD_data %>%
  count(BORO)
```

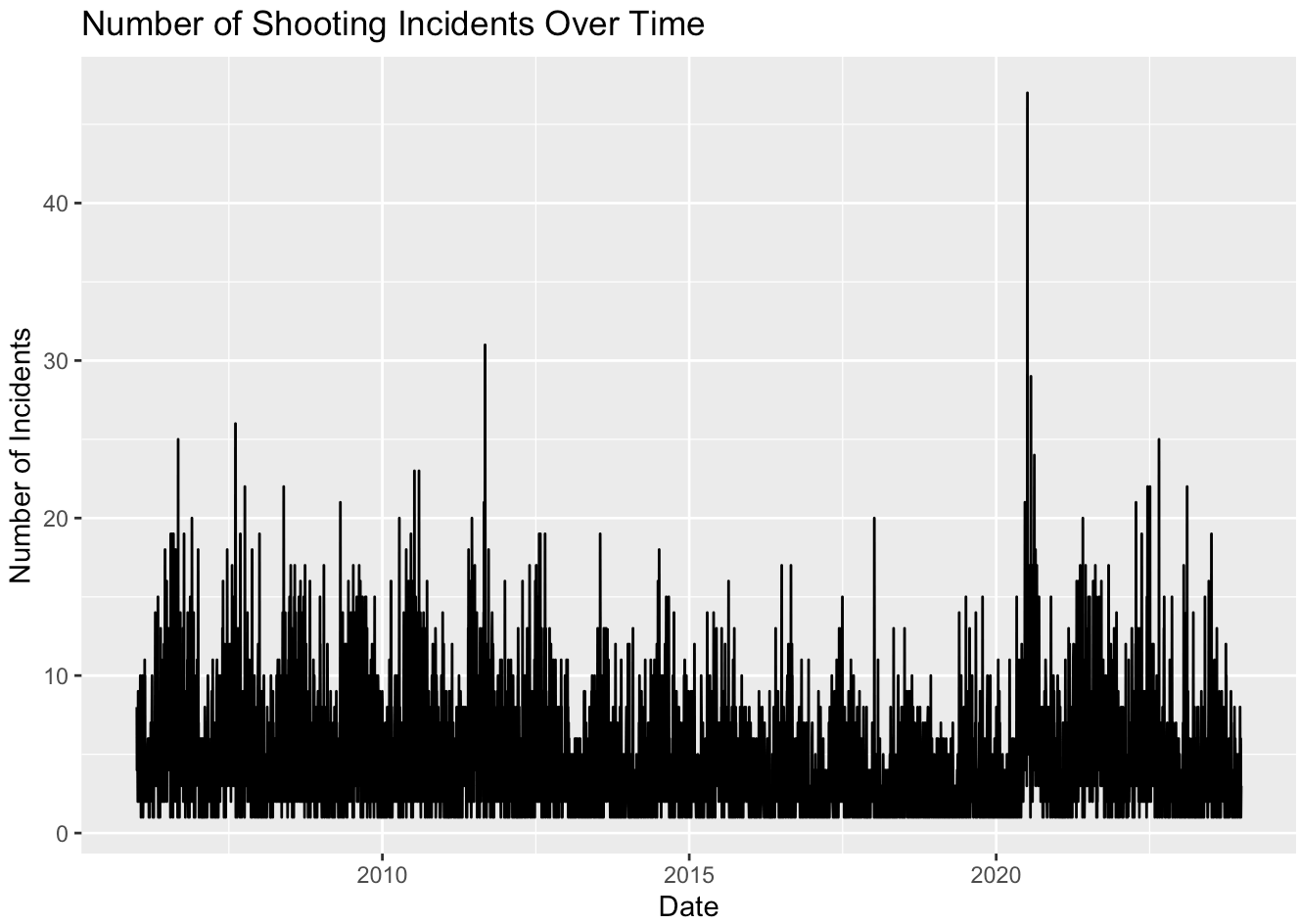
```
## # A tibble: 5 × 2
##   BORO              n
##   <chr>         <int>
## 1 BRONX          8376
## 2 BROOKLYN      11346
## 3 MANHATTAN      3762
## 4 QUEENS         4271
## 5 STATEN ISLAND   807
```

According to the table below sourced from here (https://www.citypopulation.de/en/usa/newyorkcity/) it seems pretty logical that the Borough with the most population (Queens) has the highest number of incidents

Name	Status	Population Census 1990-04-01	Population Census 2000-04-01	Population Census 2010-04-01	Population Census 2020-04-01	Population Estimate 2022-07-01
Bronx	Borough	1,203,789	1,332,650	1,385,108	1,472,653	1,356,476
Brooklyn (Kings County)	Borough	2,300,664	2,465,326	2,504,700	2,736,119	2,561,225
Manhattan (New York County)	Borough	1,487,536	1,537,195	1,585,873	1,694,250	1,597,451
Queens	Borough	1,951,598	2,229,379	2,230,722	2,405,425	2,252,196
Staten Island (Richmond County)	Borough	378,977	443,728	468,730	495,752	490,687
New York City	City	7,322,564	8,008,278	8,175,133	8,804,199	8,258,035

Incidents over time

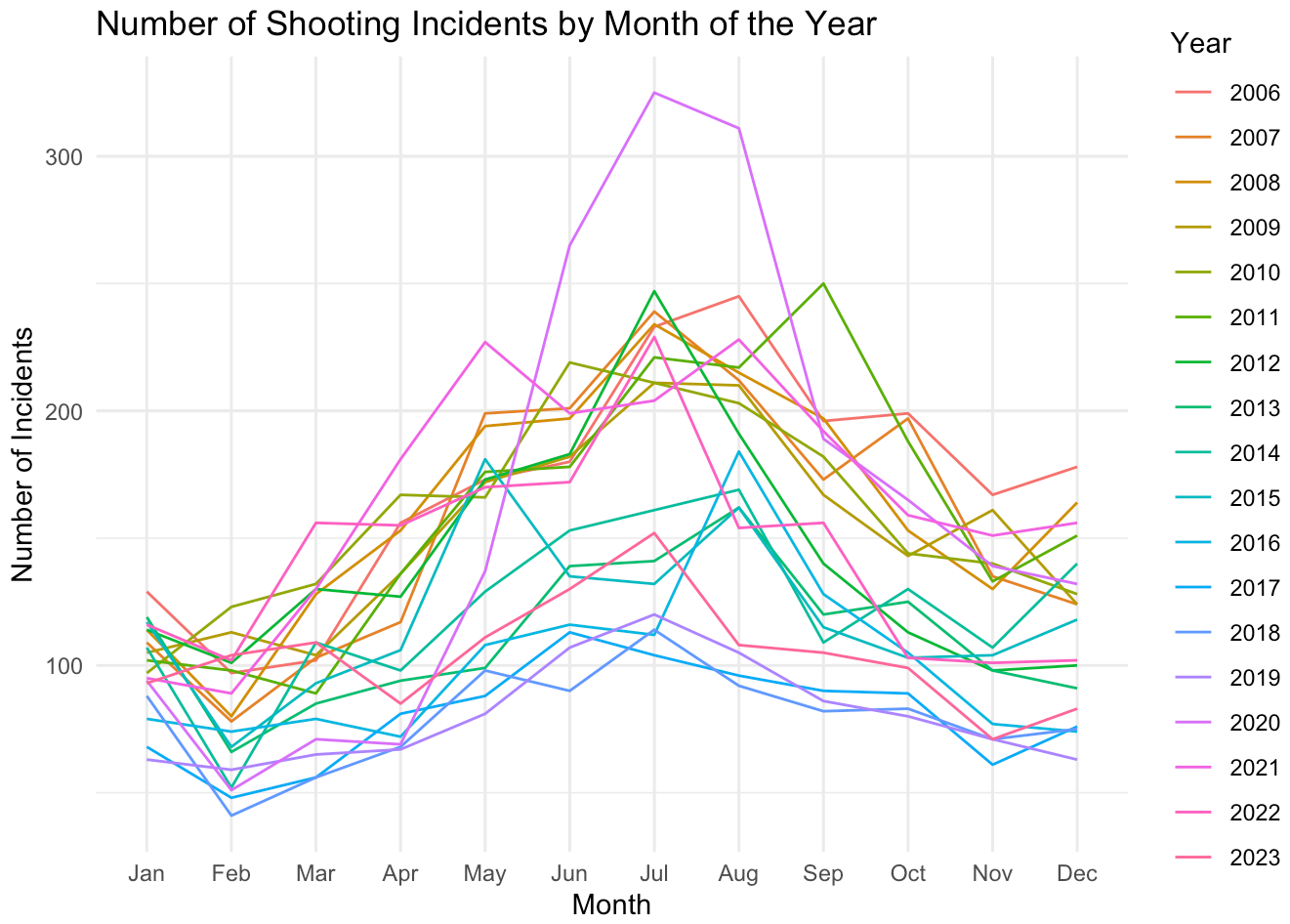
```
NYPD_data %>%
  count(OCCUR_DATE) %>%
  ggplot(aes(x = OCCUR_DATE, y = n)) +
  geom_line() +
  labs(title = "Number of Shooting Incidents Over Time", x = "Date", y = "Number of Inci
dents")
```



With this data it is very hard to really tell a pattern besides one of individual dates that are spikes. You can see a roughly seasonal chart though

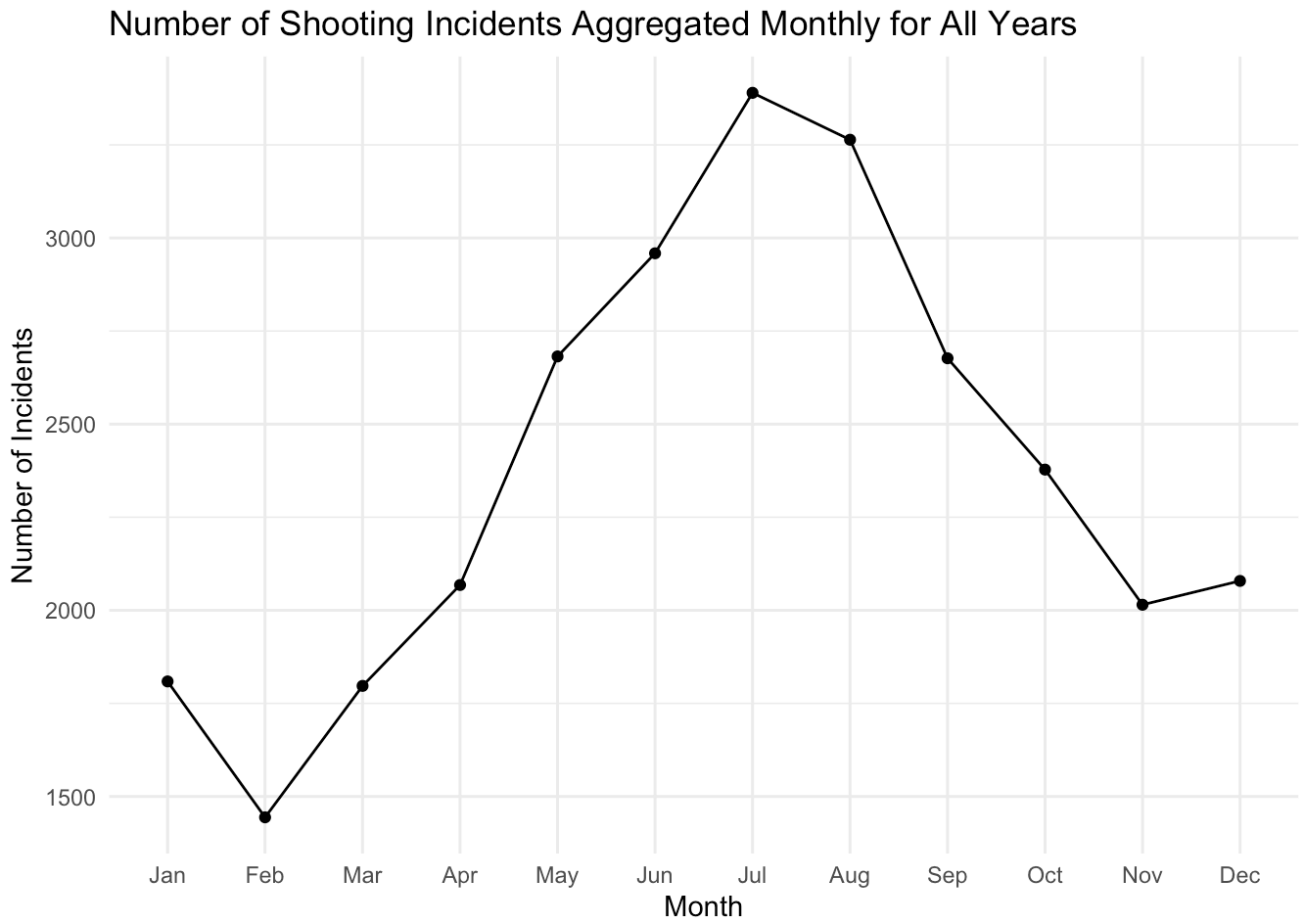
To look at it monthly we need to do the following:

```
NYPD_data <- NYPD_data %>%
  mutate(
    Year = lubridate::year(OCCUR_DATE),
    Month = lubridate::month(OCCUR_DATE, label = TRUE)
  )

monthly_data <- NYPD_data %>%
  count(Year, Month)

monthly_data %>%
  ggplot(aes(x = Month, y = n, group = Year, color = as.factor(Year))) +
  geom_line() +
  labs(title = "Number of Shooting Incidents by Month of the Year", x = "Month", y = "Nu
mber of Incidents", color = "Year") +
  theme_minimal()
```



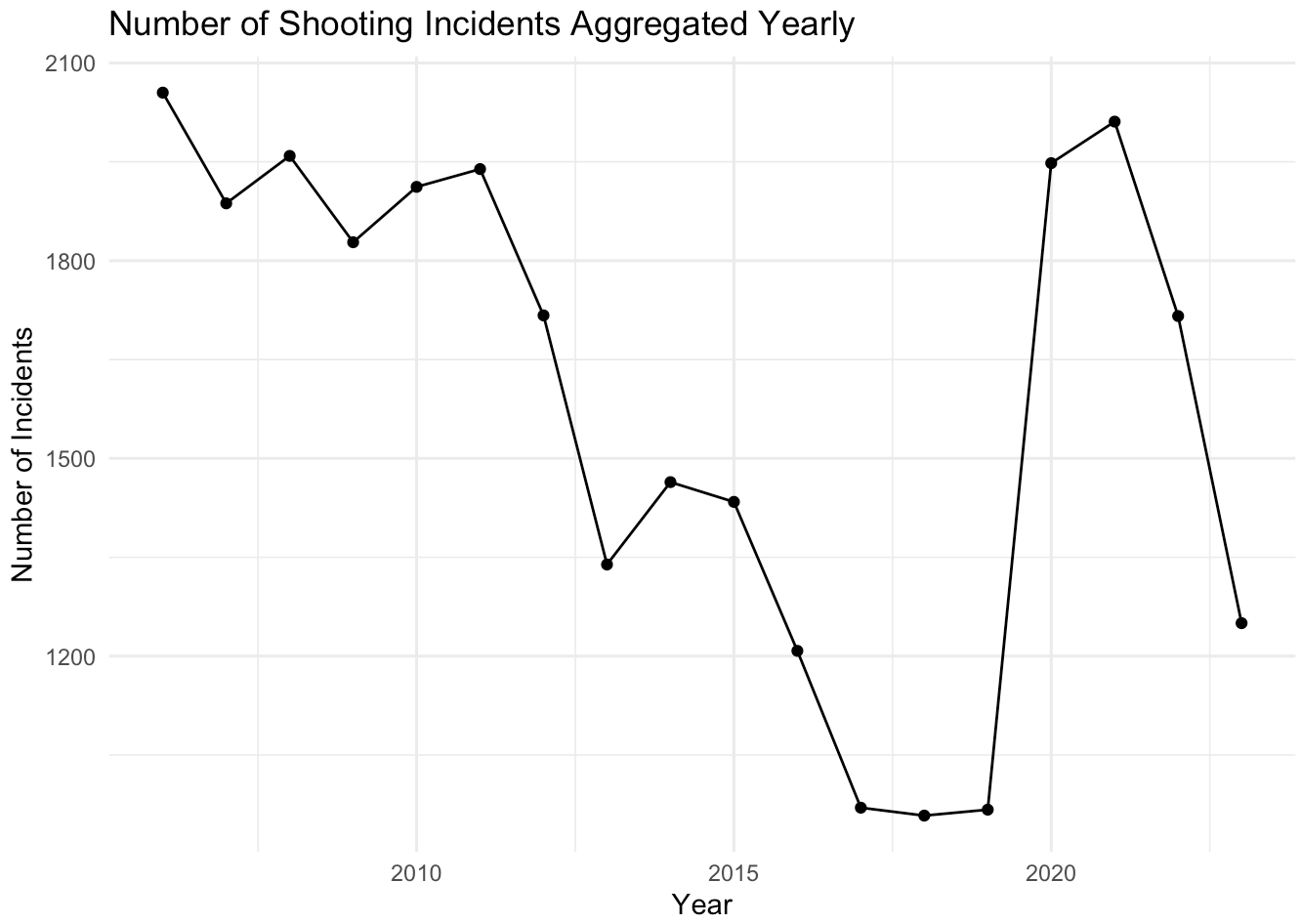
Not being super scientific about it but it seems like the years might be split between unimodal with a peak in the middle of the year and bimodal peaks around the ends of the summer with a very clear peak in july 2020 whule it seems that feb 2020 was a pretty clear drop being almost the third lowest month of all time (in the included range). If I were to guess NYC in Feb is cold and snowy and July is the warmest month the most people are outside vs inside and COVID year being lower (mostly) does seem to support that thought

```
monthly_aggregated_data <- NYPD_data %>%
  count(Month) %>%
  arrange(Month)


monthly_aggregated_data %>%
  ggplot(aes(x = Month, y = n)) +
  geom_line(group = 1) +
  geom_point() +
  labs(title = "Number of Shooting Incidents Aggregated Monthly for All Years", x = "Mon
th", y = "Number of Incidents") +
  theme_minimal()
```



It does seem like feb generally is a historically lower month compared to all the months so it is a good sample to check for with one off years like 2020

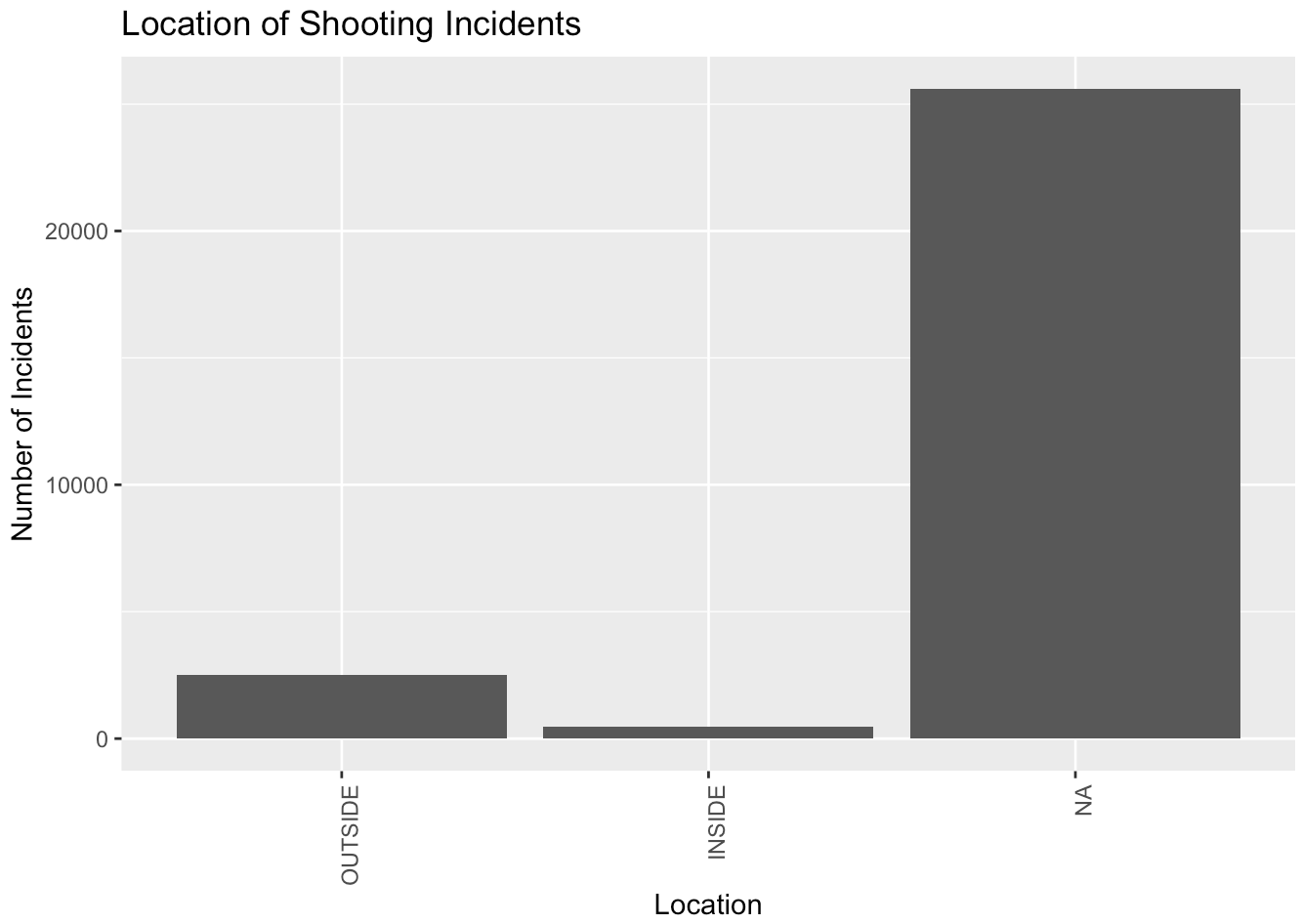
```
yearly_data <- NYPD_data %>%
  count(Year)

yearly_data %>%
  ggplot(aes(x = Year, y = n)) +
  geom_line(group = 1) +
  geom_point() +
  labs(title = "Number of Shooting Incidents Aggregated Yearly", x = "Year", y = "Number
of Incidents") +
  theme_minimal()
```



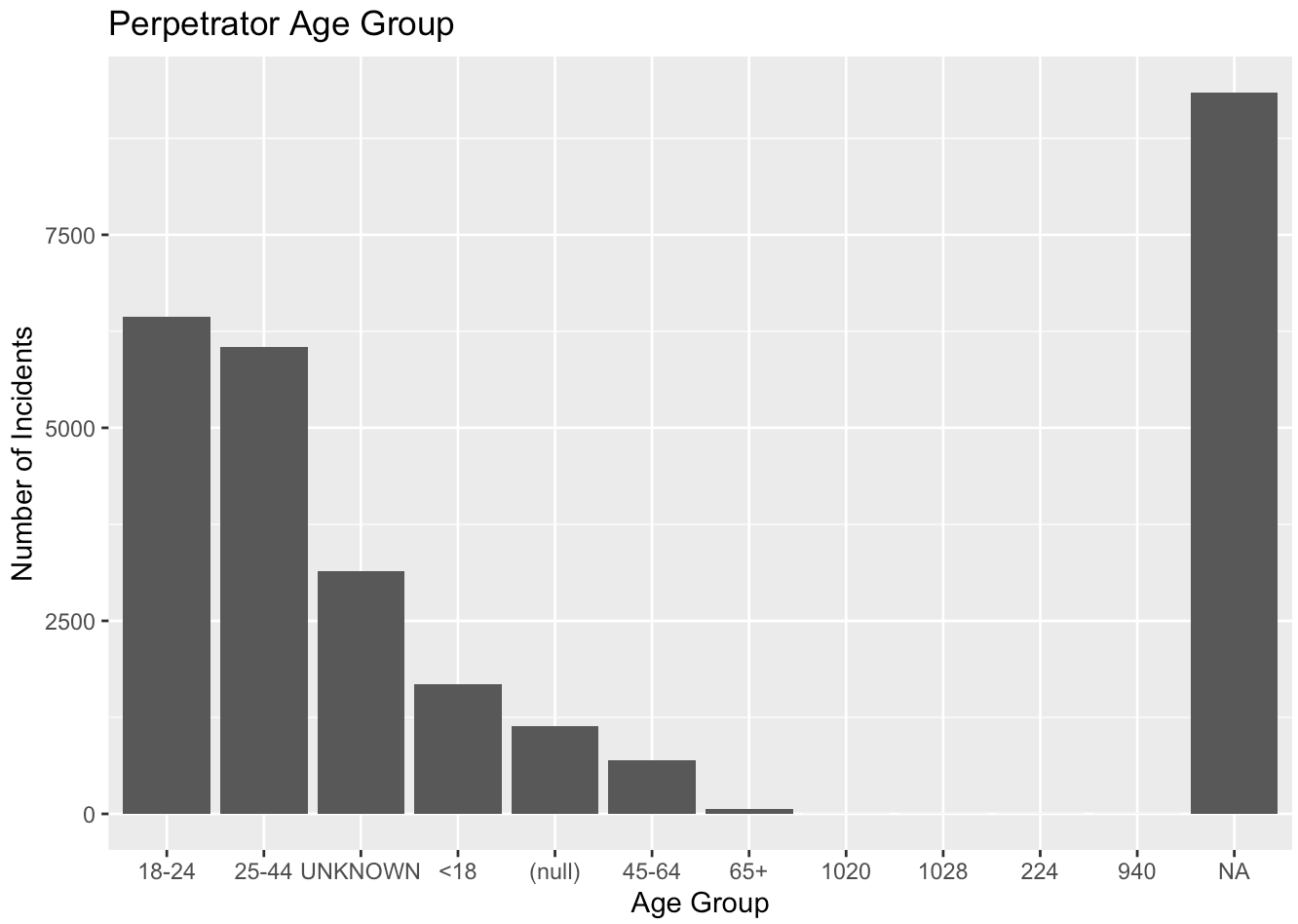
And when we do check our hypothesis against the annual aggregate it does seems like this stands true, that the rate was dropping YOY but increased dramatically during COVID.

Incident characteristics analysis

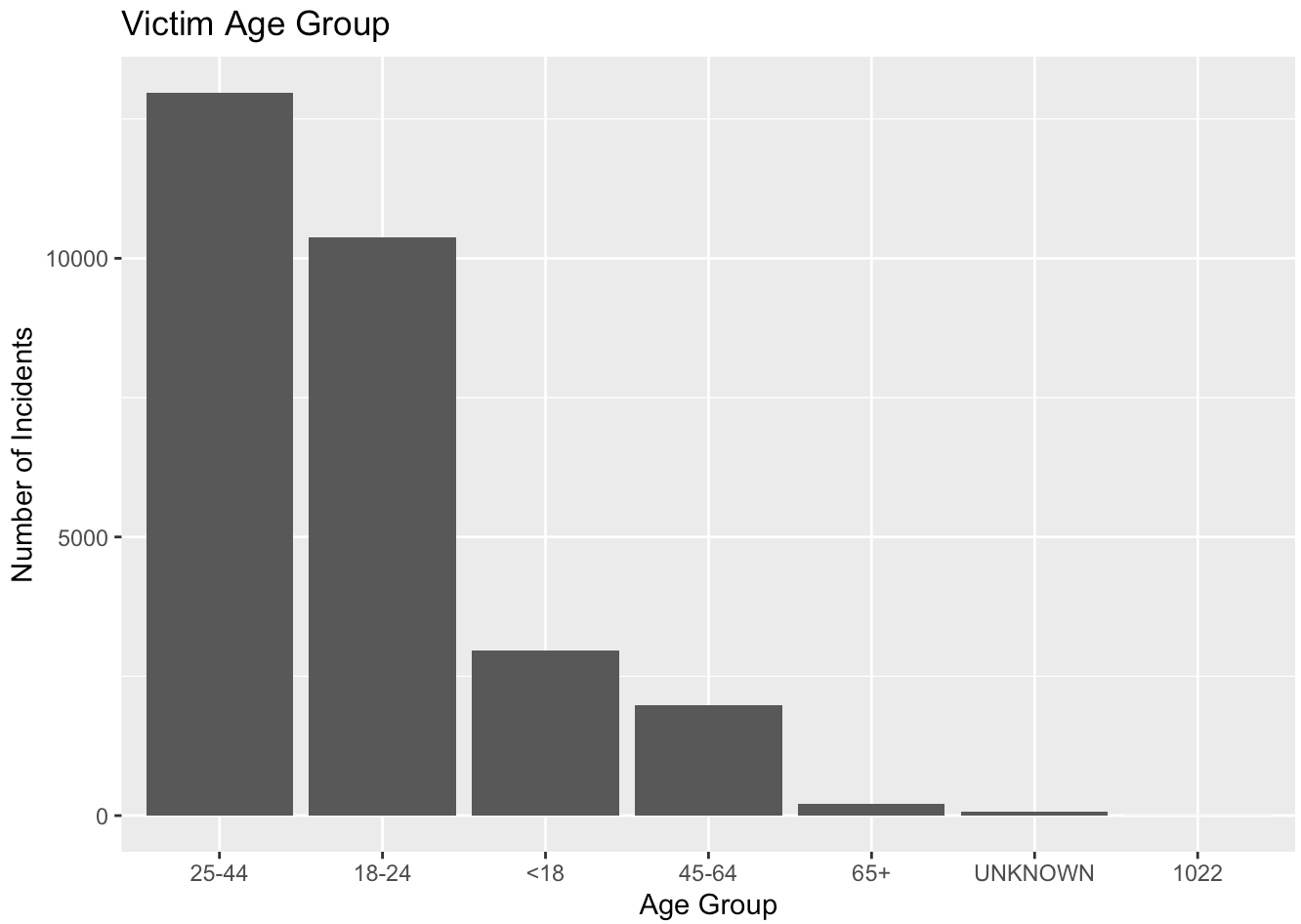
```
NYPD_data %>%
  count(LOC_OF_OCCUR_DESC) %>%
  ggplot(aes(x = reorder(LOC_OF_OCCUR_DESC, -n), y = n)) +
  geom_bar(stat = "identity") +
  labs(title = "Location of Shooting Incidents", x = "Location", y = "Number of Incident
s") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```




```
# Perpetrator age group
NYPD_data %>%
  count(PERP_AGE_GROUP) %>%
  ggplot(aes(x = reorder(PERP_AGE_GROUP, -n), y = n)) +
  geom_bar(stat = "identity") +
  labs(title = "Perpetrator Age Group", x = "Age Group", y = "Number of Incidents")
```



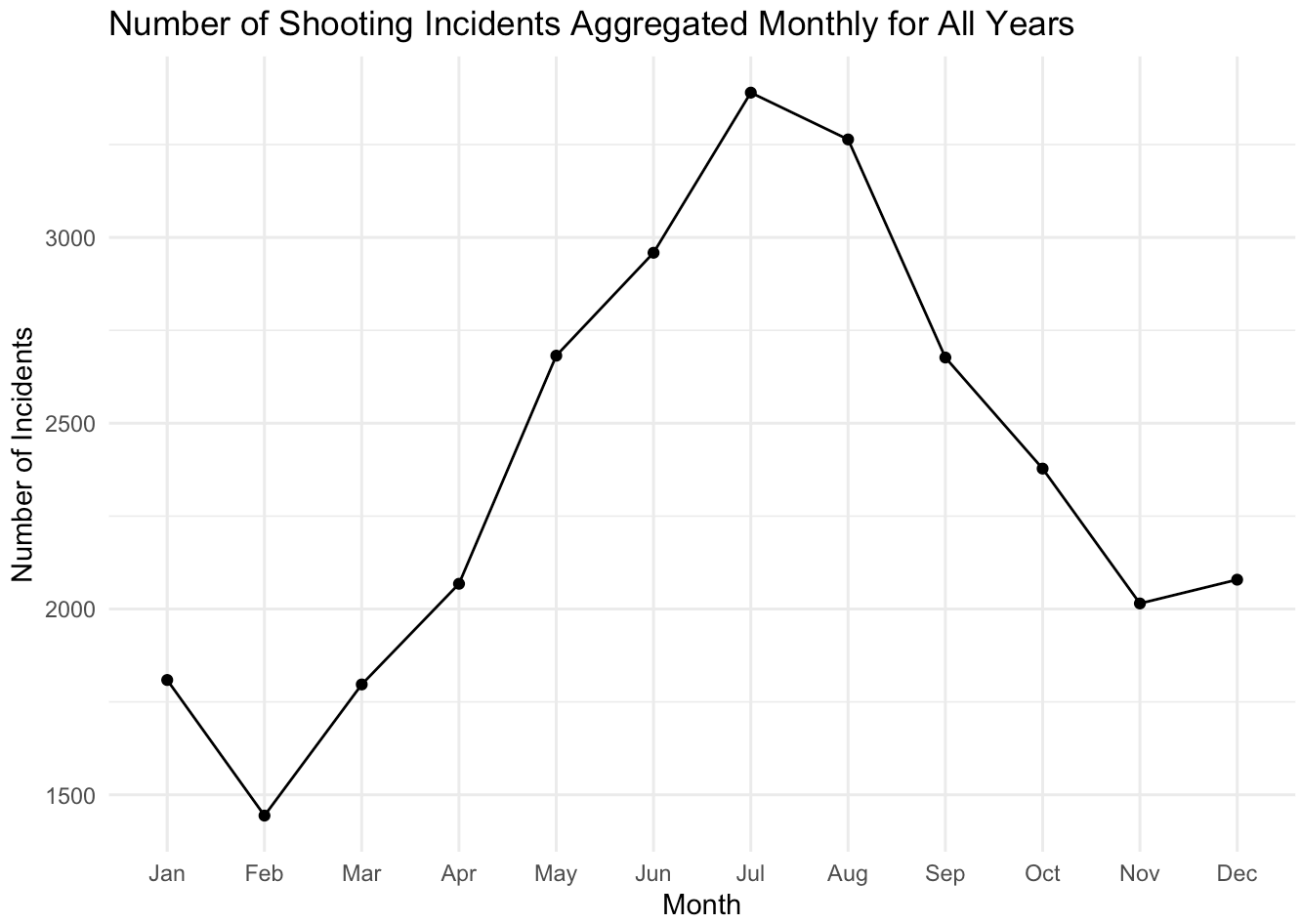

```r
# Victim age group
NYPD_data %>%
  count(VIC_AGE_GROUP) %>%
  ggplot(aes(x = reorder(VIC_AGE_GROUP, -n), y = n)) +
  geom_bar(stat = "identity") +
  labs(title = "Victim Age Group", x = "Age Group", y = "Number of Incidents")
```



Based on this data I do think there is a clear bias in the age catigory that would lead to some issues if we try to model using it. Given the fact that we have almost all the victims ages but a lot of the perp age is missing indicate 1. that we don't know who the perp that comitted it is and 2. So it would be misguided to try to corelate heavily incomplete data with close to complete data.

I do think though that it would be useful and likely easier to try to predict the number of incidents by month

```
NYPD_data <- NYPD_data %>%
  mutate(
    Year = lubridate::year(OCCUR_DATE),
    Month = lubridate::month(OCCUR_DATE, label = TRUE),
    Month_num = lubridate::month(OCCUR_DATE)
  )

monthly_aggregated_data <- NYPD_data %>%
  count(Month, Month_num) %>%
  arrange(Month_num)

monthly_aggregated_data %>%
  ggplot(aes(x = Month, y = n)) +
  geom_line(group = 1) +
  geom_point() +
  labs(title = "Number of Shooting Incidents Aggregated Monthly for All Years", x = "Mon
th", y = "Number of Incidents") +
  theme_minimal()
```



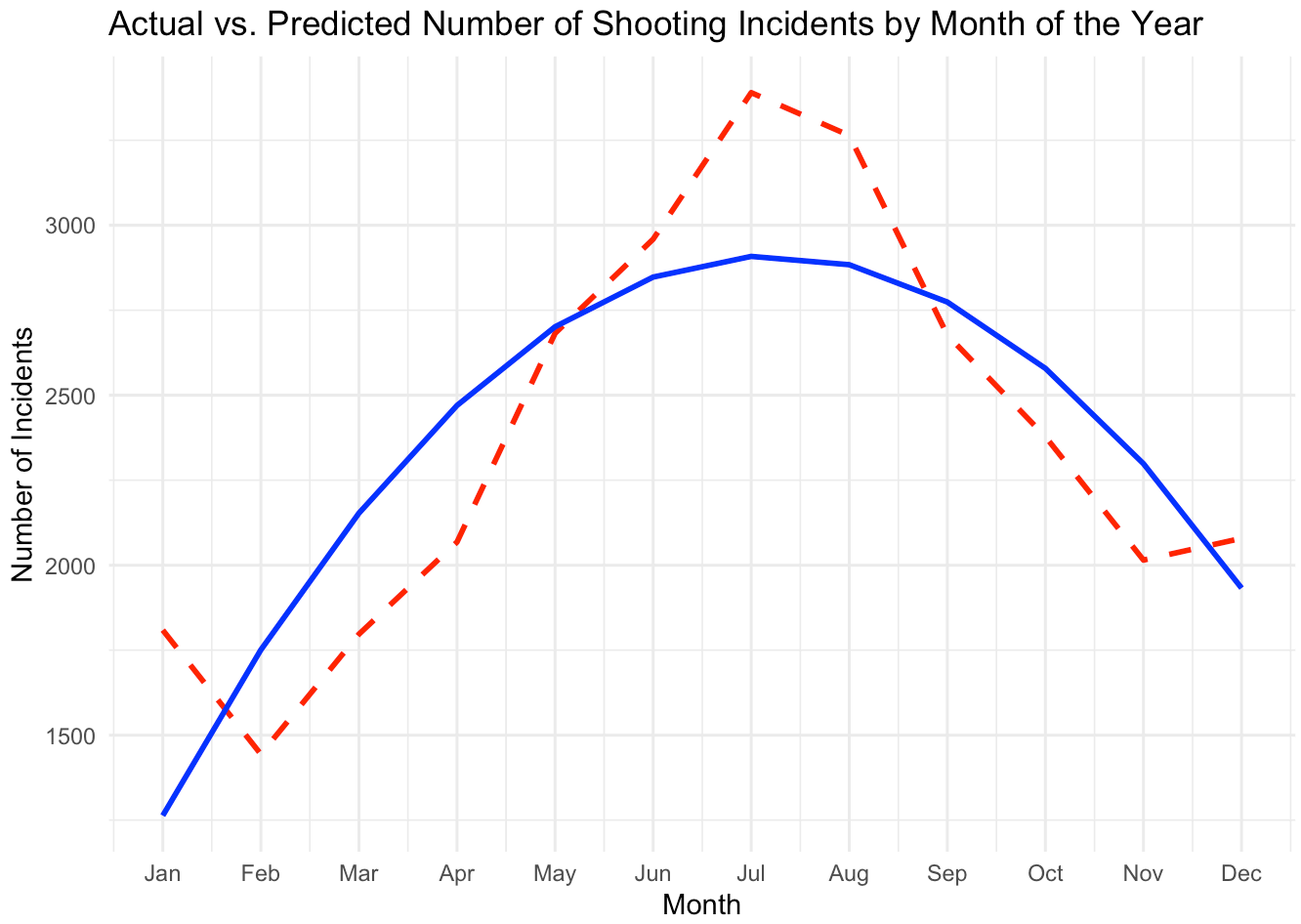

```
# Fit a polynomial model to predict the number of incidents by month
poly_model <- lm(n ~ poly(Month_num, 2), data = monthly_aggregated_data)

# Summary of the model
summary(poly_model)
```

```
##
## Call:
## lm(formula = n ~ poly(Month_num, 2), data = monthly_aggregated_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -401.85 -289.26  -58.17  204.86  545.57
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2380.2      106.3  22.386 3.35e-09 ***
## poly(Month_num, 2)1    727.4      368.3   1.975   0.0797 .
## poly(Month_num, 2)2  -1558.6      368.3  -4.232   0.0022 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 368.3 on 9 degrees of freedom
## Multiple R-squared:  0.7079, Adjusted R-squared:  0.643
## F-statistic:  10.9 on 2 and 9 DF,  p-value: 0.003936
```

```
# Predict values using the polynomial model
monthly_aggregated_data <- monthly_aggregated_data %>%
  mutate(predicted = predict(poly_model, newdata = .))

# Plot actual vs. predicted values
monthly_aggregated_data %>%
  ggplot(aes(x = Month_num)) +
  geom_line(aes(y = n), color = "red", size = 1, linetype = "dashed") +
  geom_line(aes(y = predicted), color = "blue", size = 1) +
  scale_x_continuous(breaks = 1:12, labels = levels(monthly_aggregated_data$Month)) +
  labs(title = "Actual vs. Predicted Number of Shooting Incidents by Month of the Year",
x = "Month", y = "Number of Incidents", color = "Legend") +
  theme_minimal() +
  scale_color_manual(values = c("red" = "Actual", "blue" = "Predicted"))
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: No shared levels found between `names(values)` of the manual scale and the
## data's colour values.
```




```
summary(poly_model)
```

```
##
## Call:
## lm(formula = n ~ poly(Month_num, 2), data = monthly_aggregated_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -401.85 -289.26  -58.17  204.86  545.57
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2380.2      106.3  22.386 3.35e-09 ***
## poly(Month_num, 2)1    727.4      368.3   1.975   0.0797 .
## poly(Month_num, 2)2  -1558.6      368.3  -4.232   0.0022 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 368.3 on 9 degrees of freedom
## Multiple R-squared:  0.7079, Adjusted R-squared:  0.643
## F-statistic:  10.9 on 2 and 9 DF,  p-value: 0.003936
```

With this we can see that we can likely predict the monthly crimes using a polynomial regression with a not so terrible 0.708 r squared - Not bad for a first pass

With this analysis so far we can make a few guesses.

1. There is a fairly clear seasonal trend where there is a peak in the summer and very low number of incidents in the winter
    1. this can be due to an actual seasonal trend or a reporting bias where there are less people outside to report the incident due to cold weather
2. There is a decently clear concentration of age groups in the vitims around the 25-44 and 18-24 age range. I would potentially like to see this adjusted for population to see if there is indeed a higher concentration of incidents per capita in any of the groups or if this is sampling bias of sorts
3. There was a fairly sizable YOY decrease in incidents until covid, so I'd like to see some added borough stats on potentially socio economic variables such as food insecurity over time to see if that has some correclation with increased incidents

Potential Biases:

- Reporting biases with time of day/seasonal and when people are actually outside
- Potential lack of reporting in neighborhoods or boroughs that have historically has issues with the police
- The dataset might only include incidents that meet a certain criteria for severity or involvement (e.g., only shootings with injuries or fatalities). Less serious incidents might not be reported.
- Police officers might be more likely to report shootings in certain neighborhoods or involving certain demographics due to implicit bias or focus on specific areas.

Other types of break downs I would try to do given I had enough time to do each in depth:

- A seasonalilty breakdown given that we saw some intresting trends between months
- A brough racial breakdown to see if there is a racial correclation in any of the boroughs with race of perp and vic
- An economic breakdown to see if average income ties into brough shootings per capita
- Long/Lat clusters
- Other possible correlations in the geography vs other variables