# NYPD_Shooting_Data

Hector Santillan

2022-11-30

```
library(readr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v dplyr   1.0.10
## v tibble  3.1.8      v stringr 1.4.1
## v tidyr   1.2.1      v forcats 0.5.2
## v purrr   0.3.5
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

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

```
library(ggplot2)
```

## NYPD Shooting Data (historic)

This is an analysis of the NYPD shooting historical data. the data has been obtained from www.Data.gov.

```
x <- read_csv('https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD')
```

```
## Rows: 25596 Columns: 19
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (10): OCCUR_DATE, BORO, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_R...
## 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.
```

```
x
```

```
## # A tibble: 25,596 x 19
##    INCID~1 OCCUR~2 OCCUR~3 BORO  PRECI~4 JURIS~5 LOCAT~6 STATI~7 PERP_~8 PERP_~9
##      <dbl> <chr>   <time>  <chr>   <dbl>   <dbl> <chr>   <lgl>   <chr>   <chr>
##  1  2.36e8 11/11/~ 15:04   BROO~      79       0 <NA>    FALSE   <NA>    <NA>
##  2  2.31e8 07/16/~ 22:05   BROO~      72       0 <NA>    FALSE   45-64   M
##  3  2.31e8 07/11/~ 01:09   BROO~      79       0 <NA>    FALSE   <18     M
##  4  2.38e8 12/11/~ 13:42   BROO~      81       0 <NA>    FALSE   <NA>    <NA>
##  5  2.24e8 02/16/~ 20:00   QUEE~     113       0 <NA>    FALSE   <NA>    <NA>
##  6  2.28e8 05/15/~ 04:13   QUEE~     113       0 <NA>    TRUE    <NA>    <NA>
##  7  2.27e8 04/14/~ 21:08   BRONX      42       0 COMMER~ TRUE    <NA>    <NA>
##  8  2.38e8 12/10/~ 19:30   BRONX      52       0 <NA>    FALSE   <NA>    <NA>
##  9  2.25e8 02/22/~ 00:18   MANH~      34       0 <NA>    FALSE   <NA>    <NA>
## 10  2.25e8 03/07/~ 06:15   BROO~      75       0 <NA>    TRUE    25-44   M
## # ... with 25,586 more rows, 9 more variables: 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>, and
## #   abbreviated variable names 1: INCIDENT_KEY, 2: OCCUR_DATE, 3: OCCUR_TIME,
## #   4: PRECINCT, 5: JURISDICTION_CODE, 6: LOCATION_DESC,
## #   7: STATISTICAL_MURDER_FLAG, 8: PERP_AGE_GROUP, 9: PERP_SEX
```

```
summary(x)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:25596       Length:25596       Length:25596
##  1st Qu.: 61593633   Class :character   Class1:hms         Class :character
##  Median : 86437258   Mode  :character   Class2:difftime    Mode  :character
##  Mean   :112382648                      Mode  :numeric
##  3rd Qu.:166660833
##  Max.   :238490103
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:25596       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   FALSE:20668
##  Median : 69.00   Median :0.0000    Mode  :character   TRUE :4928
##  Mean   : 65.87   Mean   :0.3316
##  3rd Qu.: 81.00   3rd Qu.:0.0000
##  Max.   :123.00   Max.   :2.0000
##                   NA's   :2
##  PERP_AGE_GROUP       PERP_SEX          PERP_RACE         VIC_AGE_GROUP
##  Length:25596       Length:25596       Length:25596       Length:25596
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_SEX            VIC_RACE           X_COORD_CD        Y_COORD_CD
##  Length:25596       Length:25596       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.:1000011   1st Qu.:182782
##  Mode  :character   Mode  :character   Median :1007715   Median :194038
##                                        Mean   :1009455   Mean   :207894
```

```
##                                      3rd Qu.:1016838   3rd Qu.:239429
##                                      Max.   :1066815   Max.   :271128
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:25596
##  1st Qu.:40.67   1st Qu.:-73.94   Class :character
##  Median :40.70   Median :-73.92   Mode  :character
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##
```

## Transform/Clean

The data was transformed and cleaned to show only the variables that are important to our analysis.

```
nypd <- x %>%
rename(
date_full = OCCUR_DATE,
time = OCCUR_TIME,
borough = BORO,
precinct = PRECINCT,
jurisdiction_code = JURISDICTION_CODE,
statistical_murder = STATISTICAL_MURDER_FLAG,
vic_age = VIC_AGE_GROUP,
vic_sex = VIC_SEX,
vic_race = VIC_RACE) %>%
mutate(date = mdy(date_full)) %>%
separate(date, into =c("year", "month", "day")) %>%
select(-c(X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat, INCIDENT_KEY, LOCATION_DESC, PERP_AGE_G
nypd
```

```
## # A tibble: 25,596 x 12
##    date_full time  borough preci~1 juris~2 stati~3 vic_age vic_sex vic_r~4 year
##    <chr>     <tim> <chr>     <dbl>   <dbl> <lgl>   <chr>   <chr>   <chr>   <chr>
##  1 11/11/20~ 15:04 BROOKL~      79       0 FALSE   18-24   M       BLACK   2021
##  2 07/16/20~ 22:05 BROOKL~      72       0 FALSE   25-44   M       ASIAN ~ 2021
##  3 07/11/20~ 01:09 BROOKL~      79       0 FALSE   25-44   M       BLACK   2021
##  4 12/11/20~ 13:42 BROOKL~      81       0 FALSE   25-44   M       BLACK   2021
##  5 02/16/20~ 20:00 QUEENS      113       0 FALSE   25-44   M       BLACK   2021
##  6 05/15/20~ 04:13 QUEENS      113       0 TRUE    25-44   M       BLACK   2021
##  7 04/14/20~ 21:08 BRONX        42       0 TRUE    18-24   M       BLACK   2021
##  8 12/10/20~ 19:30 BRONX        52       0 FALSE   25-44   M       BLACK   2021
##  9 02/22/20~ 00:18 MANHAT~      34       0 FALSE   25-44   M       BLACK ~ 2021
## 10 03/07/20~ 06:15 BROOKL~      75       0 TRUE    25-44   M       WHITE ~ 2021
## # ... with 25,586 more rows, 2 more variables: month <chr>, day <chr>, and
## #   abbreviated variable names 1: precinct, 2: jurisdiction_code,
## #   3: statistical_murder, 4: vic_race
```

```
summary(nypd)
```

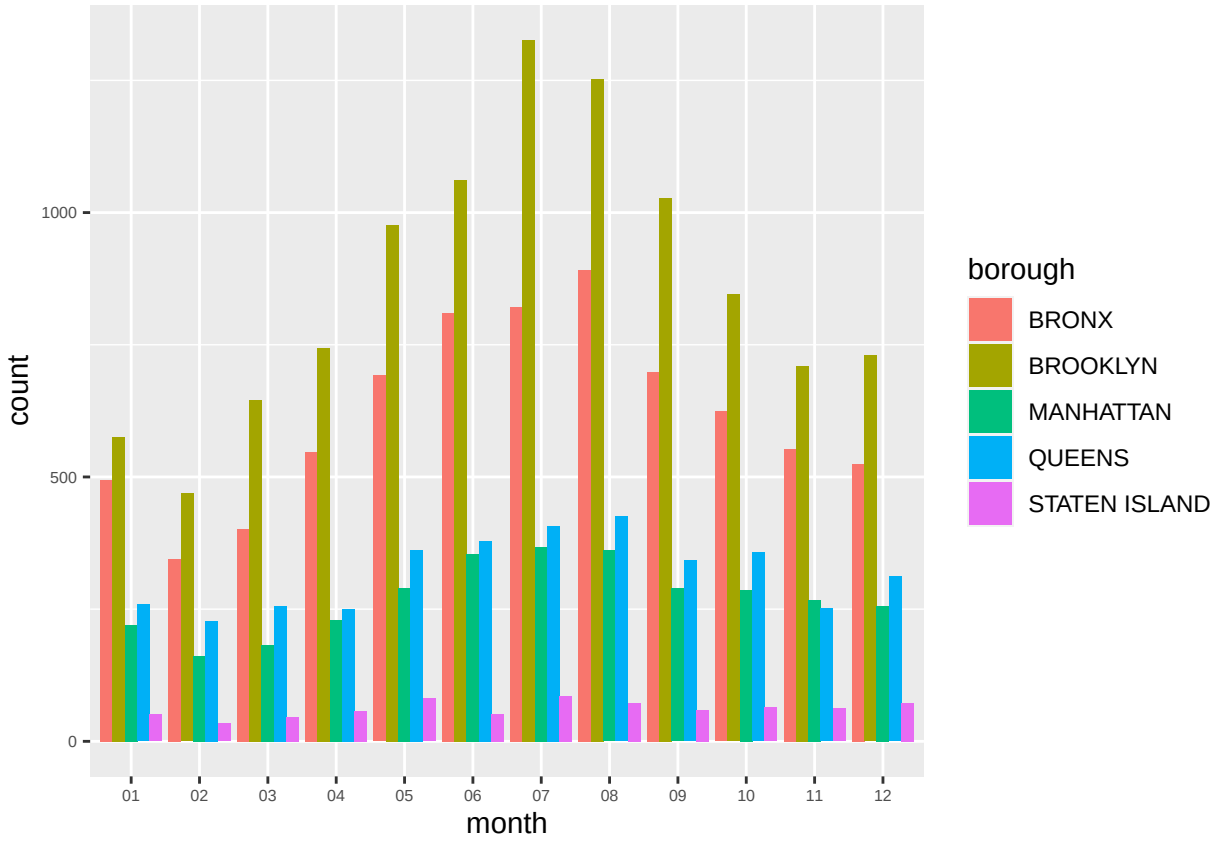
```
##   date_full             time            borough             precinct
```

```
##  Length:25596       Length:25596      Length:25596       Min.   :  1.00
##  Class :character   Class1:hms        Class :character   1st Qu.: 44.00
##  Mode  :character   Class2:difftime   Mode  :character   Median : 69.00
##                     Mode  :numeric                       Mean   : 65.87
##                                                          3rd Qu.: 81.00
##                                                          Max.   :123.00
##
##  jurisdiction_code statistical_murder   vic_age            vic_sex
##  Min.   :0.0000    Mode :logical      Length:25596       Length:25596
##  1st Qu.:0.0000    FALSE:20668        Class :character   Class :character
##  Median :0.0000    TRUE :4928         Mode  :character   Mode  :character
##  Mean   :0.3316
##  3rd Qu.:0.0000
##  Max.   :2.0000
##  NA's   :2
##    vic_race             year              month              day
##  Length:25596       Length:25596       Length:25596       Length:25596
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
```

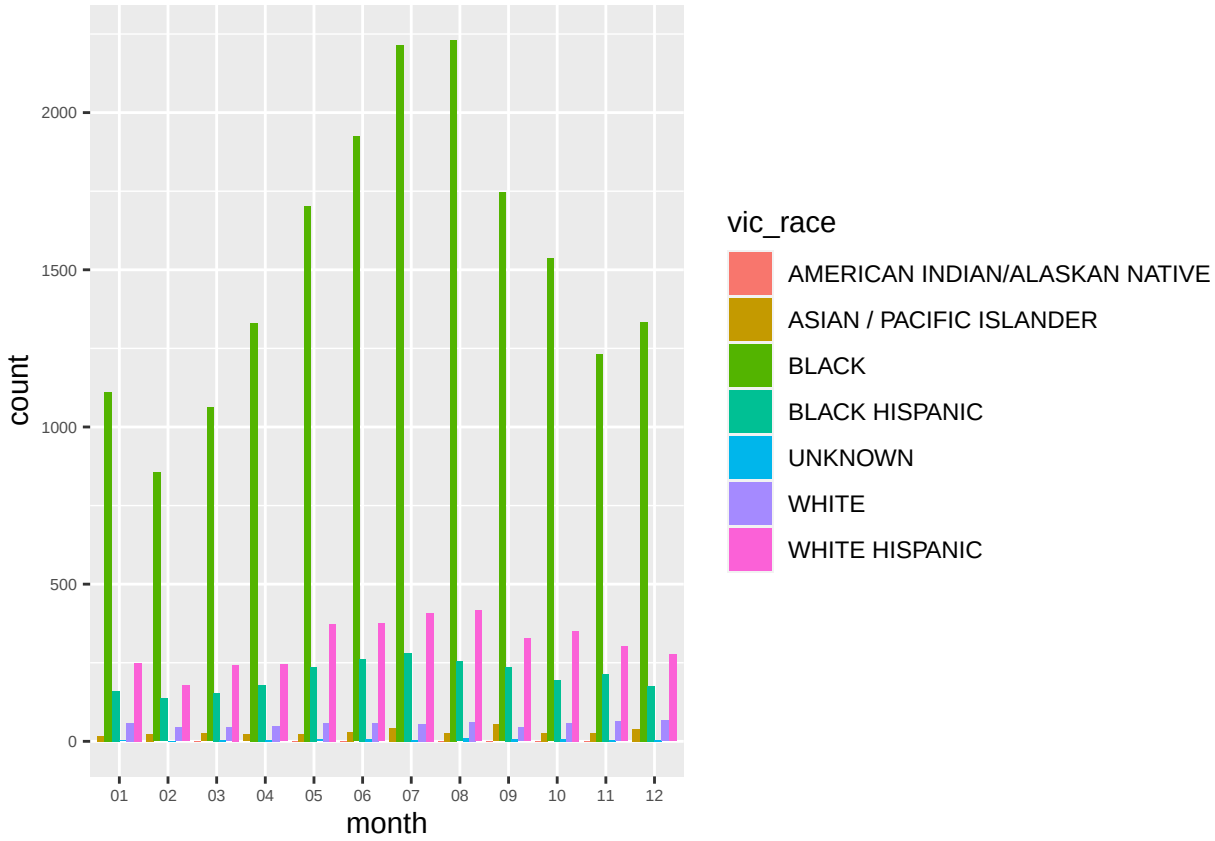
## Plots

I am very interested to see the number of shootings by month. Are there months where shootings are greater? statistically significantly greater?

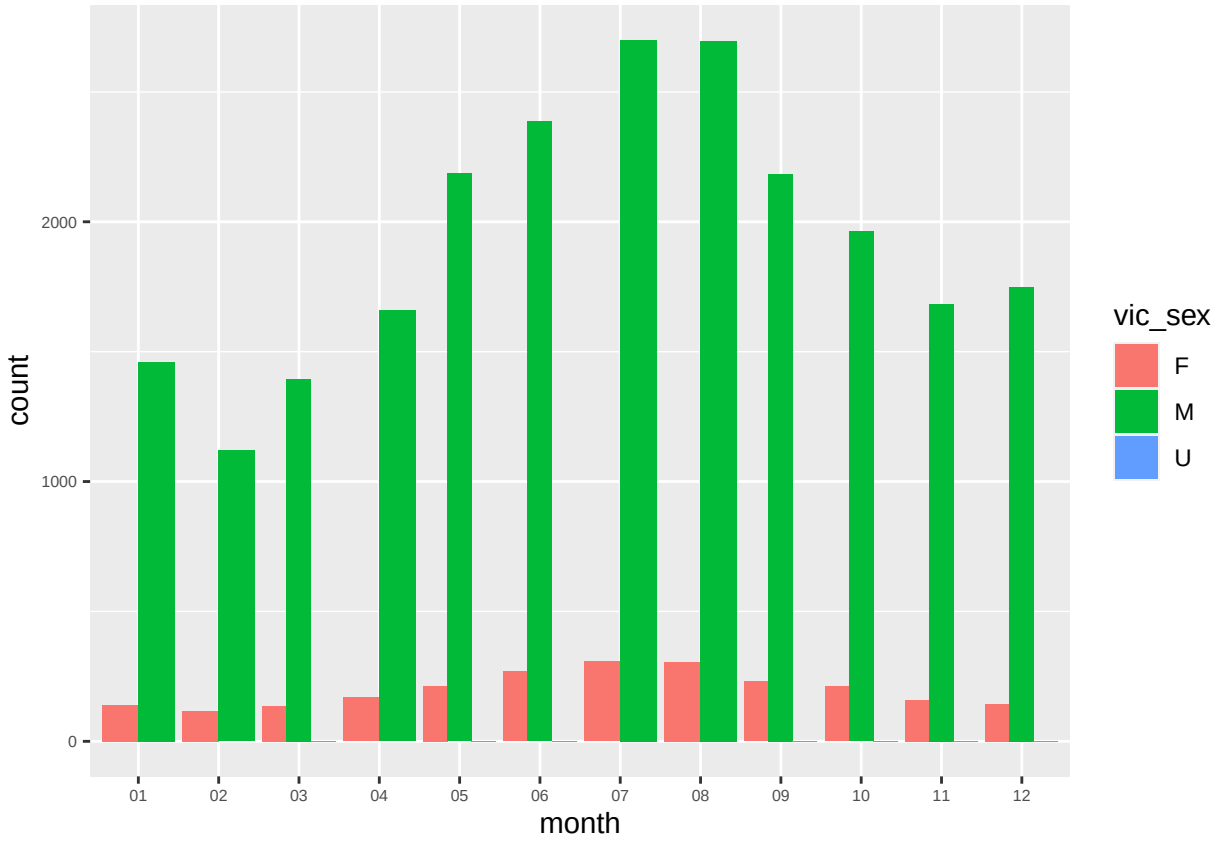
```
plot <- ggplot(data = nypd) +
geom_bar(mapping = aes(x = month, fill = borough), position = "dodge")
plot + theme(axis.text = element_text(size = rel(0.5)))
```



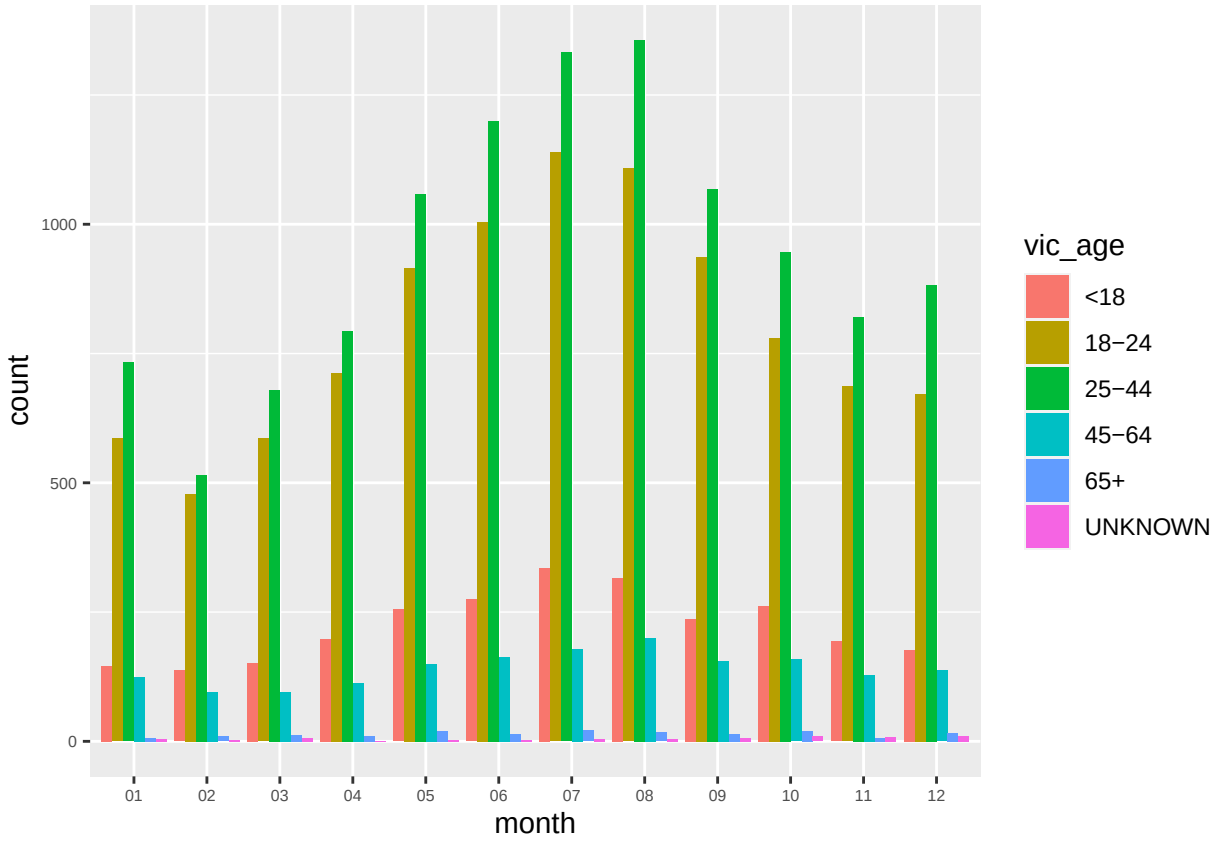

```
plot <- ggplot(data = nypd) +
geom_bar(mapping = aes(x = month, fill = vic_race), position = "dodge")
plot + theme(axis.text = element_text(size = rel(0.5)))
```




```
plot <- ggplot(data = nypd) +
geom_bar(mapping = aes(x = month, fill = vic_sex), position = "dodge")
plot + theme(axis.text = element_text(size = rel(0.5)))
```

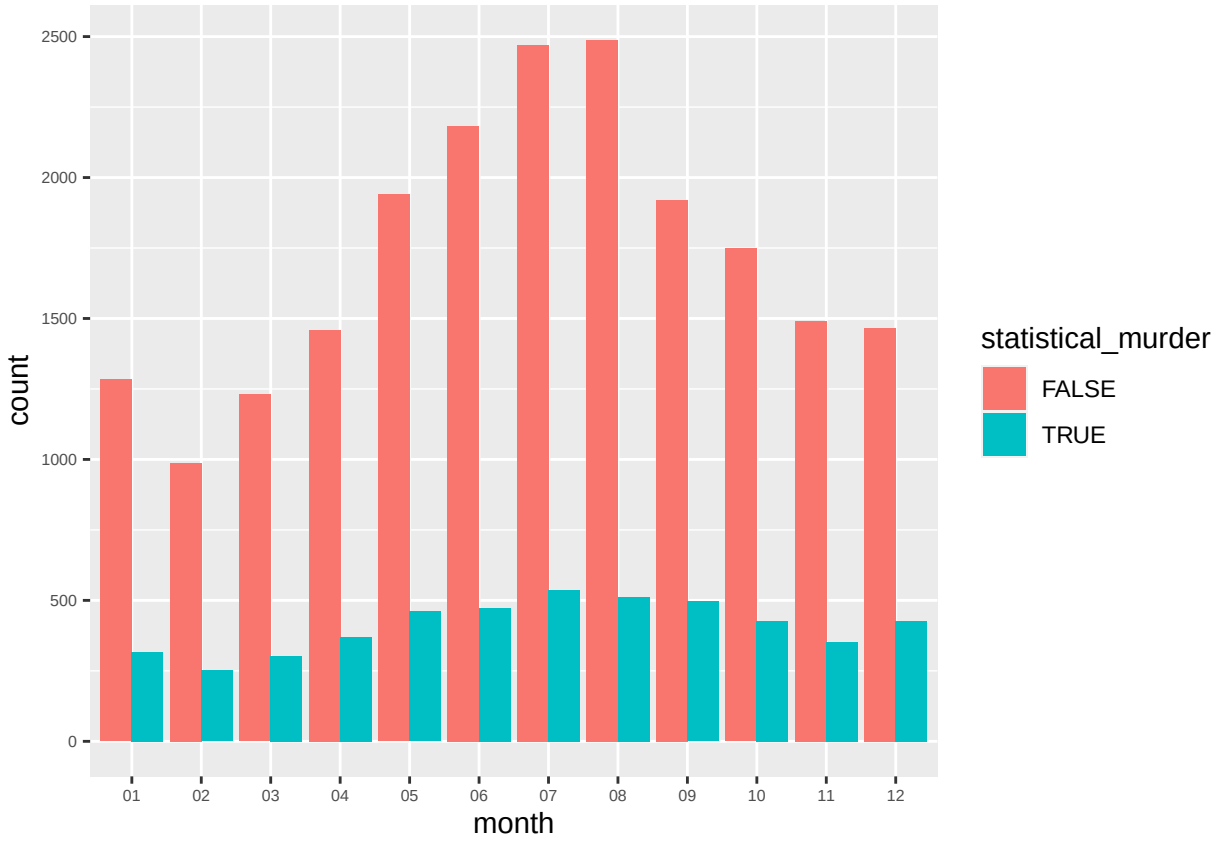



```
plot <- ggplot(data = nypd) +
geom_bar(mapping = aes(x = month, fill = vic_age), position = "dodge")
plot + theme(axis.text = element_text(size = rel(0.5)))
```




```
shootings_vs_deaths <- nypd %>%
group_by(borough, month) %>%
mutate(shootings = n(), deaths = sum(statistical_murder)) %>%
select(borough, shootings, statistical_murder, deaths, month, date_full) %>%
ungroup() %>%
summarize(borough, month, shootings, statistical_murder, deaths, date_full)

plot <- ggplot(data = shootings_vs_deaths) +
geom_bar(mapping = aes(x = month, fill = statistical_murder), position = "dodge")
plot + theme(axis.text = element_text(size = rel(0.5)))
```




## Analysis

Questions raised by analysis: 1. Why do the months of July and August have the most shootings? Deaths? 2. Is there a reason why the summer months tend to have the highest number of shootings? 3. Are shootings statistically significantly higher during certain months?

## Model

```
mod <- lm(shootings ~ month, data = shootings_vs_deaths)
summary(mod)
```

```
##
## Call:
## lm(formula = shootings ~ month, data = shootings_vs_deaths)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -825.74 -213.68   18.88  239.88  415.26
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  433.684      7.065  61.381   <2e-16 ***
## month02      -95.744     10.698  -8.950   <2e-16 ***
## month03        9.982     10.102   0.988    0.323
## month04       97.713      9.675  10.099   <2e-16 ***
```

```
## month05      254.840      9.121  27.941   <2e-16 ***
## month06      339.973      8.943  38.015   <2e-16 ***
## month07      478.061      8.744  54.670   <2e-16 ***
## month08      459.061      8.748  52.476   <2e-16 ***
## month09      289.535      9.109  31.784   <2e-16 ***
## month10      171.437      9.307  18.420   <2e-16 ***
## month11       80.718      9.657   8.358   <2e-16 ***
## month12       81.385      9.596   8.481   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 282.6 on 25584 degrees of freedom
## Multiple R-squared:  0.2806, Adjusted R-squared:  0.2803
## F-statistic: 907.1 on 11 and 25584 DF,  p-value: < 2.2e-16
```

## Conclusion

Bias & Conclusion:

Bias: I don't think that there is any significant bias from myself to have affected in any significant manner to this analysis. To be honest, I think if there is any bias it might be in the data itself. There might have been some bias by the individuals putting together or collecting the data or even in the reporting of the incidents, in such factors as gender and race. There could have been any number of actions or decisions that could have effected the categorical data in this data set.

Conclusion: In conclusion, based on the data set, the months of July and August had the highest number of shootings. This raised the question of whether months were statistically significant predictors of shootings. After my analysis and modeling, months are a statistically significant predictor of shootings within the five boroughs of New York.