NYPD Shooting Incident Data (Historic) Analyzing

2023-09-15

Import Library

At first, we have to import some useful library of R, which can help to analyze the data in the following sections.

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- 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)
```

Import Data

According to the assignment, we import the NYPD Shooting Incident Data (Historic) from the U.S. Government's Open Data website. Additionally, we can look at the table displayed by R studio to understand the variables of the data, Furthermore, we can find a data dictionary on the same website ("https://data.cityofnewyork.us/api/views/833y-fsy8/files/f5f61d94-6961-47bd-8d3c-e57ebeb4cb55?download=true&filename=NYPD_Shootings_Historic_DataDictionary.xlsx") to gain a clearer understanding of the definition of each variable.

```
## 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.
```

```
## # A tibble: 27,312 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
##  7     85295722 06/17/2012 22:47      QUEENS   <NA>                   114
##  8     71662474 03/08/2010 19:41      BROOKLYN <NA>                    81
##  9     83002139 02/05/2012 05:45      QUEENS   <NA>                   105
## 10     86437261 08/26/2012 01:10      QUEENS   <NA>                   101
## # i 27,302 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>
```

Date formatting

To further analyze the relationship between time and the occurrence of cases, we adjusted the format of the OCCUR_DATE from "chr" to "date" .

```
NYPD_Data$OCCUR_DATE<- as.Date(NYPD_Data$OCCUR_DATE,"%m/%d/%Y")
glimpse(NYPD_Data)
```

```
## Rows: 27,312
## Columns: 21
## $ INCIDENT_KEY            <dbl> 228798151, 137471050, 147998800, 146837977, 58~
## $ OCCUR_DATE              <date> 2021-05-27, 2014-06-27, 2015-11-21, 2015-10-0~
## $ OCCUR_TIME              <time> 21:30:00, 17:40:00, 03:56:00, 18:30:00, 22:58~
## $ BORO                    <chr> "QUEENS", "BRONX", "QUEENS", "BRONX", "BRONX",~
## $ LOC_OF_OCCUR_DESC       <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ PRECINCT                <dbl> 105, 40, 108, 44, 47, 81, 114, 81, 105, 101, 2~
## $ JURISDICTION_CODE       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 2, 2~
## $ LOC_CLASSFCTN_DESC      <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ LOCATION_DESC           <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, "MULTI DWE~
## $ STATISTICAL_MURDER_FLAG <lgl> FALSE, FALSE, TRUE, FALSE, TRUE, TRUE, FALSE, ~
## $ PERP_AGE_GROUP          <chr> NA, NA, NA, NA, "25-44", NA, NA, NA, NA, "25-4~
## $ PERP_SEX                <chr> NA, NA, NA, NA, "M", NA, NA, NA, NA, "M", NA, ~
## $ PERP_RACE               <chr> NA, NA, NA, NA, "BLACK", NA, NA, NA, NA, "BLAC~
## $ VIC_AGE_GROUP           <chr> "18-24", "18-24", "25-44", "<18", "45-64", "25~
## $ VIC_SEX                 <chr> "M", "M", "M", "M", "M", "M", "M", "M", "M", "~
## $ VIC_RACE                <chr> "BLACK", "BLACK", "WHITE", "WHITE HISPANIC", "~
## $ X_COORD_CD              <dbl> 1058925.0, 1005028.0, 1007667.9, 1006537.4, 10~
## $ Y_COORD_CD              <dbl> 180924.0, 234516.0, 209836.5, 244511.1, 262189~
## $ Latitude                <dbl> 40.66296, 40.81035, 40.74261, 40.83778, 40.886~
## $ Longitude               <dbl> -73.73084, -73.92494, -73.91549, -73.91946, -7~
## $ Lon_Lat                 <chr> "POINT (-73.73083868899994 40.662964620000025)~
```

Dropping Variables Dropping the variables of the data which we do not need in analyzing at these sector.

```
NYPD_Data_analyzing = select(NYPD_Data, -c(LOC_OF_OCCUR_DESC,
                                           PRECINCT,
                                           JURISDICTION_CODE,
                                           LOC_CLASSFCTN_DESC,
                                           LOCATION_DESC,
                                           X_COORD_CD,
                                           Y_COORD_CD,
                                           Latitude,
                                           Longitude,
                                           Lon_Lat))
```

Tidy and Transform Variables At this stage, we need to examine the columns in our data to check for outliers and missing values. Properly addressing these issues is essential to prevent potential errors in subsequent data visualization and model analysis.

```
lapply(NYPD_Data_analyzing, function(x) sum(is.na(x)))
```

```
## $INCIDENT_KEY
## [1] 0
##
## $OCCUR_DATE
## [1] 0
##
## $OCCUR_TIME
## [1] 0
##
## $BORO
## [1] 0
##
## $STATISTICAL_MURDER_FLAG
## [1] 0
##
## $PERP_AGE_GROUP
## [1] 9344
##
## $PERP_SEX
## [1] 9310
##
## $PERP_RACE
## [1] 9310
##
## $VIC_AGE_GROUP
## [1] 0
##
## $VIC_SEX
## [1] 0
##
## $VIC_RACE
## [1] 0
```

For non-numerical classes, use "Unknown" as a complement to avoid deleting other data that can be analyzed due to the presence of NA values in the field. In addition, we have transformed the variables into a categorical (factor) form using 'as.factor' to facilitate further analysis.

```
#Unknown
NYPD_Data_analyzing <- NYPD_Data_analyzing%>%
  replace_na(list(PERP_AGE_GROUP = 'UNKNOWN',
                  PERP_SEX = 'UNKNOWN',
                  PERP_RACE = 'UNKNOWN' ))

#Delete
NYPD_Data_analyzing <- subset(NYPD_Data_analyzing,PERP_AGE_GROUP!="1020" & PERP_AGE_GROUP!="224" &PERP_A

#date
NYPD_Data_analyzing$year <- year(NYPD_Data_analyzing$OCCUR_DATE)
NYPD_Data_analyzing$month <- month(NYPD_Data_analyzing$OCCUR_DATE)

#(null)
NYPD_Data_analyzing$PERP_AGE_GROUP = recode(NYPD_Data_analyzing$PERP_AGE_GROUP, "(null)" = 'UNKNOWN')
NYPD_Data_analyzing$PERP_SEX = recode(NYPD_Data_analyzing$PERP_SEX, U = 'UNKNOWN')
NYPD_Data_analyzing$PERP_SEX = recode(NYPD_Data_analyzing$PERP_SEX, "(null)" = 'UNKNOWN')
NYPD_Data_analyzing$PERP_RACE = recode(NYPD_Data_analyzing$PERP_RACE, "(null)" = 'UNKNOWN')
NYPD_Data_analyzing$PERP_RACE = recode(NYPD_Data_analyzing$PERP_RACE, "(Other)" = 'UNKNOWN')

#factor
NYPD_Data_analyzing$BORO = as.factor(NYPD_Data_analyzing$BORO)
NYPD_Data_analyzing$PERP_AGE_GROUP = as.factor(NYPD_Data_analyzing$PERP_AGE_GROUP)
NYPD_Data_analyzing$PERP_SEX = as.factor(NYPD_Data_analyzing$PERP_SEX)
NYPD_Data_analyzing$PERP_RACE = as.factor(NYPD_Data_analyzing$PERP_RACE)
NYPD_Data_analyzing$VIC_AGE_GROUP = as.factor(NYPD_Data_analyzing$VIC_AGE_GROUP)
NYPD_Data_analyzing$VIC_SEX = as.factor(NYPD_Data_analyzing$VIC_SEX)
NYPD_Data_analyzing$VIC_RACE = as.factor(NYPD_Data_analyzing$VIC_RACE)
```

We reviewed the processed data using the 'summary()' function, and the results are as follows

```
summary(NYPD_Data_analyzing)
```

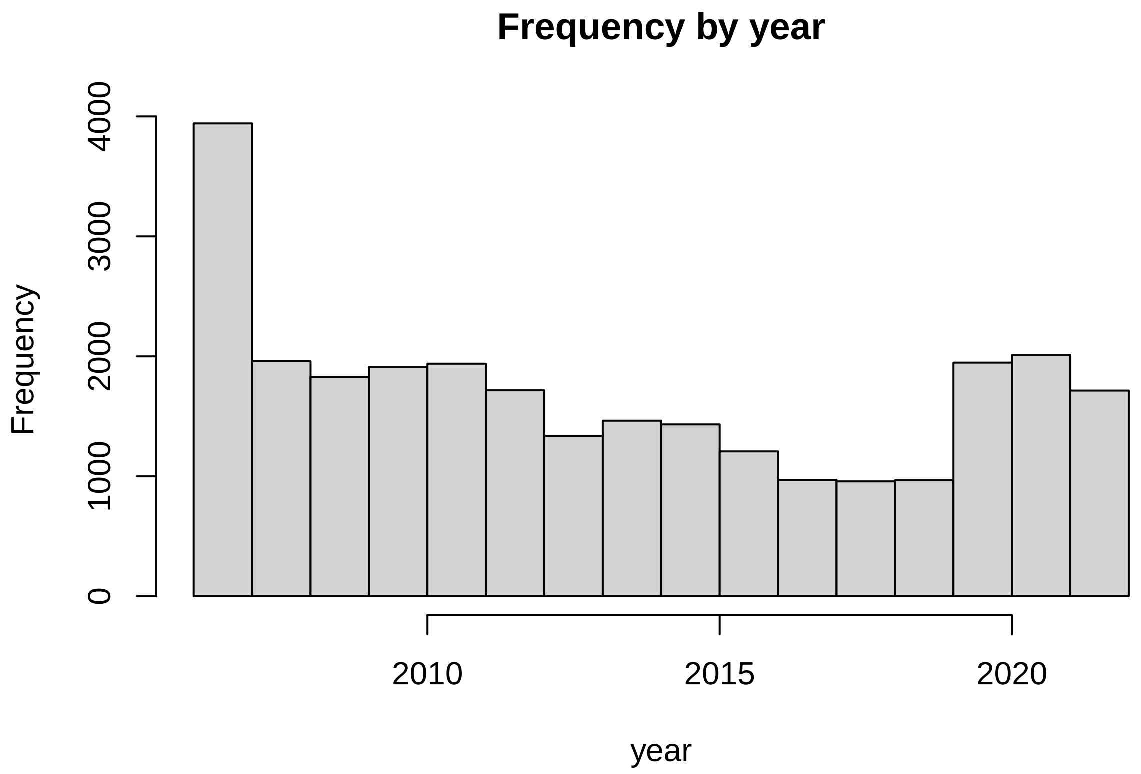
```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Length:27308
##  1st Qu.: 63859975   1st Qu.:2009-07-18   Class1:hms
##  Median : 90372218   Median :2013-04-29   Class2:difftime
##  Mean   :120858027   Mean   :2014-01-06   Mode  :numeric
##  3rd Qu.:188810230   3rd Qu.:2018-10-15
##  Max.   :261190187   Max.   :2022-12-31
##
##             BORO       STATISTICAL_MURDER_FLAG PERP_AGE_GROUP     PERP_SEX
##  BRONX        : 7935   Mode :logical           <18    : 1591   F      :  424
##  BROOKLYN     :10932   FALSE:22042             18-24  : 6221   M      :15435
##  MANHATTAN    : 3571   TRUE :5266              25-44  : 5687   UNKNOWN:11449
##  QUEENS       : 4094                           45-64  :  617
##  STATEN ISLAND:  776                           65+    :   60
##                                                UNKNOWN:13132
##
##                            PERP_RACE     VIC_AGE_GROUP   VIC_SEX
##  AMERICAN INDIAN/ALASKAN NATIVE:    2   <18    : 2839   F: 2615
##  ASIAN / PACIFIC ISLANDER      :  154   18-24  :10085   M:24682
##  BLACK                         :11430   25-44  :12279   U:   11
##  BLACK HISPANIC                 : 1314   45-64  : 1863
##  UNKNOWN                        :11786   65+    :  181
##  WHITE                          :  283   UNKNOWN:   61
##  WHITE HISPANIC                 : 2339
##                             VIC_RACE          year          month
##  AMERICAN INDIAN/ALASKAN NATIVE:   10   Min.   :2006   Min.   : 1.000
##  ASIAN / PACIFIC ISLANDER      :  404   1st Qu.:2009   1st Qu.: 5.000
##  BLACK                         :19437   Median :2013   Median : 7.000
##  BLACK HISPANIC                : 2646   Mean   :2013   Mean   : 6.826
##  UNKNOWN                       :   66   3rd Qu.:2018   3rd Qu.: 9.000
##  WHITE                         :  698   Max.   :2022   Max.   :12.000
##  WHITE HISPANIC                : 4047
```

Visualization

1.(Date)

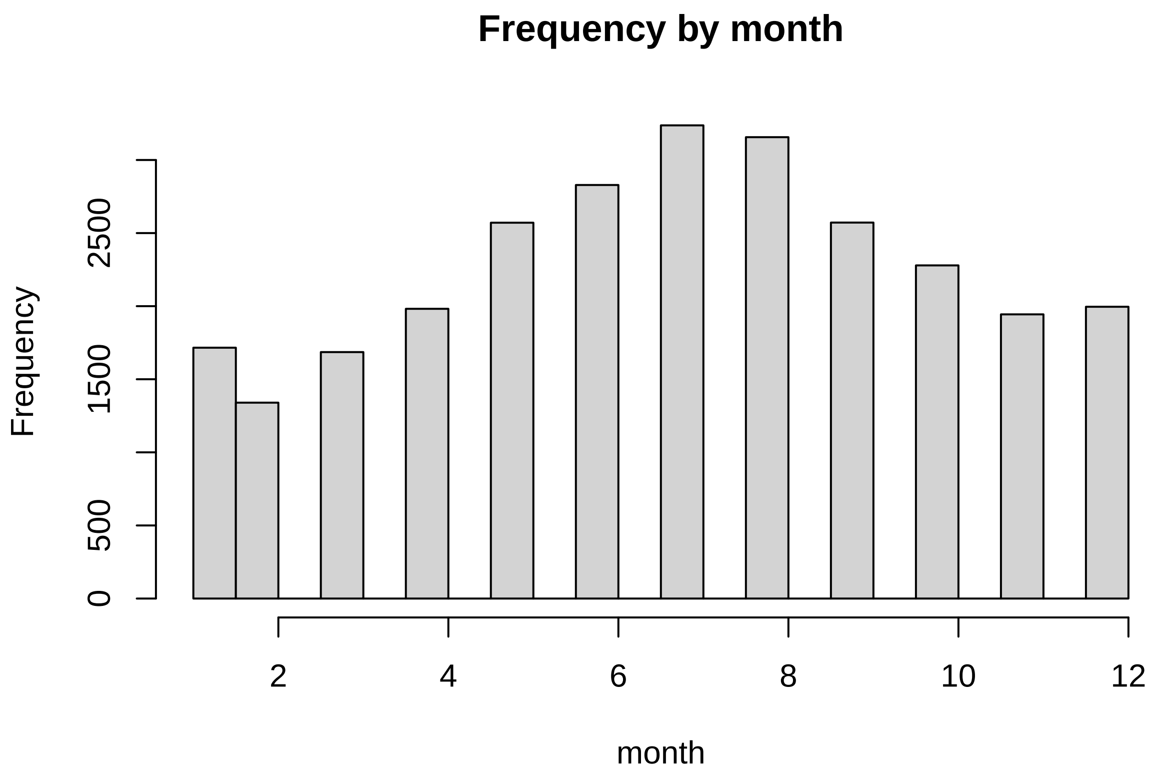
In addition to using the 'summary()' method to identify missing values or other categories, data visualization can enhance our understanding of the data. Here, we attempted to visualize the data on an annual and monthly basis, and the results are as follows:

```
#
hist(NYPD_Data_analyzing$year,xlab = "year",main = "Frequency by year")
```



Firstly, visualizing the data on an annual basis reveals that the frequency of occurrences was higher in earlier years, with a gradual decline in frequency until around 2010. However, there has been a gradual increase in frequency since 2020.

```
hist(NYPD_Data_analyzing$month,xlab = "month",main = "Frequency by month")
```



Next, when we visualized the data on a monthly basis, the results showed a bell-shaped distribution of events, with a concentration in the months of June to August. July had the highest frequency of occurrences. The data accumulated over several years mostly clustered around these months. This pattern may be related to recurring events such as festivals or parades, but due to limitations in the data available to us, we cannot conduct further analysis in this regard.

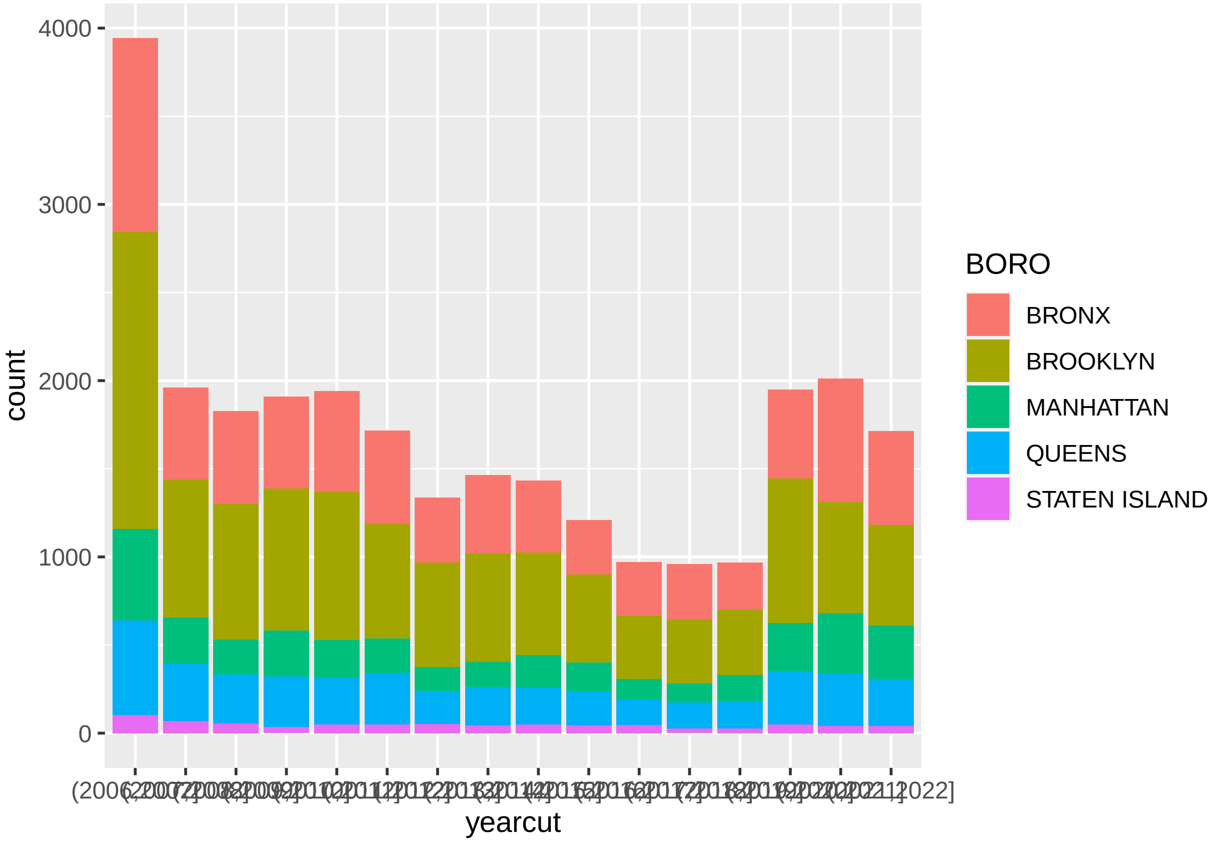
2.(Boro)

In addition to the annual and monthly data, this study aims to focus on the regional aspect to examine whether there have been changes or any notable variations in the frequency of events over the years. We present the data using two types of visualizations: the first is a stacked bar chart displaying the actual counts, and the second is a percentage breakdown of events by region for each year. Here are the specific details:

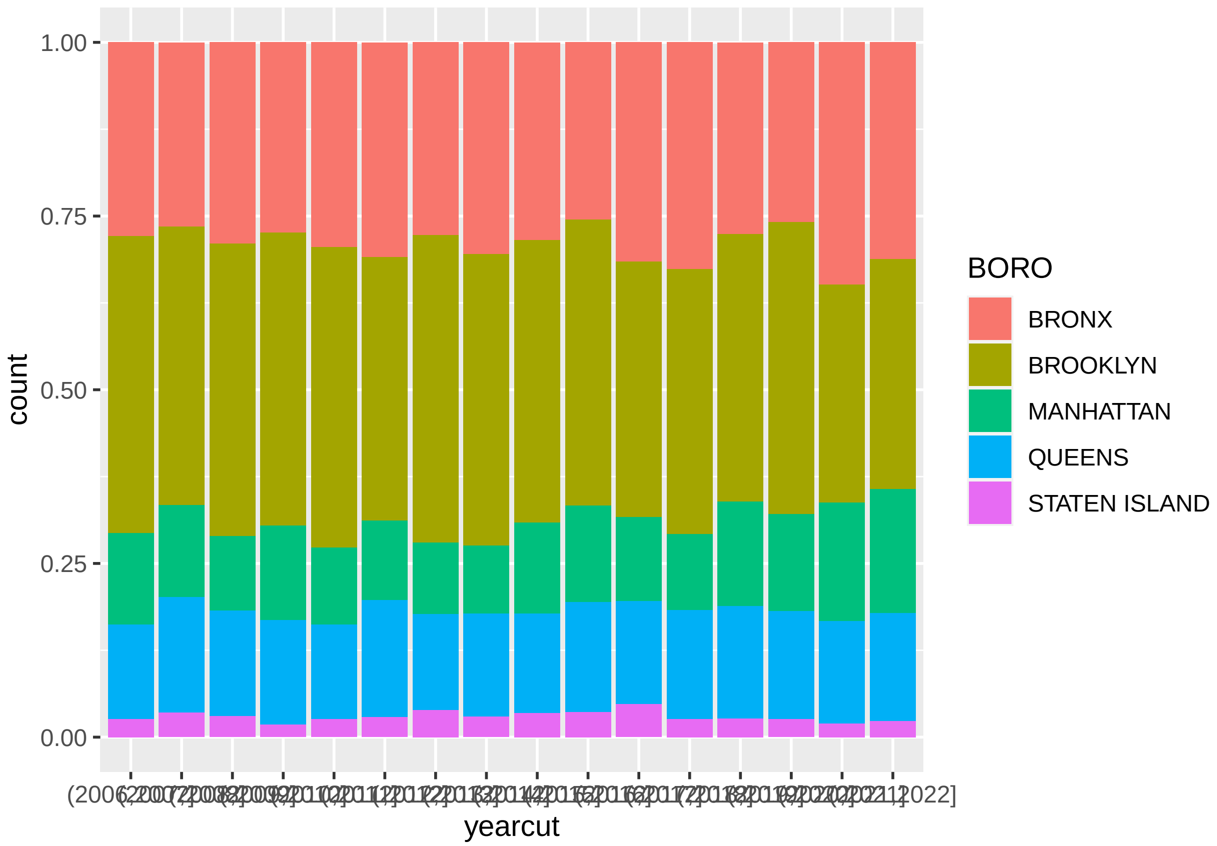
```
NYPD_Data_analyzing$count <- 1

#
NYPD_Data_analyzing$yearcut <- cut(year(NYPD_Data_analyzing$OCCUR_DATE),breaks = 16 )
NYPD_Data_analyzing_year<- select(NYPD_Data_analyzing,yearcut,BORO,count)

#
ggplot(NYPD_Data_analyzing,aes(yearcut,fill=BORO))+geom_bar()
```



```
ggplot(NYPD_Data_analyzing,aes(yearcut,fill=BORO))+geom_bar(position = "fill")
```



Specifically, the percentages by region have remained relatively stable over the years. However, starting in 2020, a noticeable increase in events can be observed in the BRONX region, leading to a significant contribution to the overall growth in event counts in that area.

Modeling logistic refression.

According to the data dictionary we have obtained, events can be categorized as causing victim fatalities using the 'MURDER_FLAG' indicator. This is a significant and severe scenario that we wish to focus on. We conducted a logistic regression analysis, incorporating factors such as region, time (month), and victim-related data to identify significant factors associated with victim fatalities. The results of this analysis are presented using the 'summary()' function.

```
NYPD_Data_analyzing$month_factor = NYPD_Data_analyzing$month
NYPD_Data_analyzing$month_factor = as.factor(NYPD_Data_analyzing$month_factor)

glm.fit = glm(NYPD_Data_analyzing$STATISTICAL_MURDER_FLAG ~ NYPD_Data_analyzing$BORO
                                                          + NYPD_Data_analyzing$VIC_AGE_GROUP
                                                          + NYPD_Data_analyzing$VIC_SEX
                                                          + NYPD_Data_analyzing$VIC_RACE
                                                          + NYPD_Data_analyzing$month_factor )
summary(glm.fit)
```

```
##
## Call:
## glm(formula = NYPD_Data_analyzing$STATISTICAL_MURDER_FLAG ~ NYPD_Data_analyzing$BORO +
##     NYPD_Data_analyzing$VIC_AGE_GROUP + NYPD_Data_analyzing$VIC_SEX + 
##     NYPD_Data_analyzing$VIC_RACE + NYPD_Data_analyzing$month_factor)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -0.3701  -0.2178  -0.1739  -0.1313   0.9528  
## 
## Coefficients:
##                                                       Estimate Std. Error
## (Intercept)                                          -0.030156   0.124912
## NYPD_Data_analyzing$BOROBROOKLYN                     -0.003117   0.005994
## NYPD_Data_analyzing$BOROMANHATTAN                    -0.020971   0.007926
## NYPD_Data_analyzing$BOROQUEENS                       -0.004466   0.007673
## NYPD_Data_analyzing$BOROSTATEN ISLAND                 0.003440   0.014881
## NYPD_Data_analyzing$VIC_AGE_GROUP18-24                0.035624   0.008363
## NYPD_Data_analyzing$VIC_AGE_GROUP25-44                0.086364   0.008206
## NYPD_Data_analyzing$VIC_AGE_GROUP45-64                0.112326   0.011778
## NYPD_Data_analyzing$VIC_AGE_GROUP65+                  0.164757   0.030223
## NYPD_Data_analyzing$VIC_AGE_GROUPUNKNOWN              0.124914   0.053088
## NYPD_Data_analyzing$VIC_SEXM                         -0.008647   0.008147
## NYPD_Data_analyzing$VIC_SEXU                         -0.076756   0.123986
## NYPD_Data_analyzing$VIC_RACEASIAN / PACIFIC ISLANDER  0.221262   0.125703
## NYPD_Data_analyzing$VIC_RACEBLACK                     0.175470   0.124188
## NYPD_Data_analyzing$VIC_RACEBLACK HISPANIC            0.149389   0.124387
## NYPD_Data_analyzing$VIC_RACEUNKNOWN                   0.082037   0.134302
## NYPD_Data_analyzing$VIC_RACEWHITE                     0.234352   0.125088
## NYPD_Data_analyzing$VIC_RACEWHITE HISPANIC            0.194306   0.124304
## NYPD_Data_analyzing$month_factor2                     0.012996   0.014312
## NYPD_Data_analyzing$month_factor3                    -0.002218   0.013467
## NYPD_Data_analyzing$month_factor4                     0.004438   0.012947
## NYPD_Data_analyzing$month_factor5                     0.001197   0.012240
## NYPD_Data_analyzing$month_factor6                    -0.018341   0.012015
## NYPD_Data_analyzing$month_factor7                    -0.017325   0.011732
## NYPD_Data_analyzing$month_factor8                    -0.024271   0.011781
## NYPD_Data_analyzing$month_factor9                     0.007114   0.012243
## NYPD_Data_analyzing$month_factor10                    0.002333   0.012551
## NYPD_Data_analyzing$month_factor11                   -0.002127   0.013003
## NYPD_Data_analyzing$month_factor12                    0.025459   0.012925
##                                                      t value Pr(>|t|)    
## (Intercept)                                           -0.241  0.80923    
## NYPD_Data_analyzing$BOROBROOKLYN                      -0.520  0.60305    
## NYPD_Data_analyzing$BOROMANHATTAN                     -2.646  0.00815 ** 
## NYPD_Data_analyzing$BOROQUEENS                        -0.582  0.56057    
## NYPD_Data_analyzing$BOROSTATEN ISLAND                  0.231  0.81720    
## NYPD_Data_analyzing$VIC_AGE_GROUP18-24                 4.260 2.05e-05 ***
## NYPD_Data_analyzing$VIC_AGE_GROUP25-44                10.524  < 2e-16 ***
## NYPD_Data_analyzing$VIC_AGE_GROUP45-64                 9.537  < 2e-16 ***
## NYPD_Data_analyzing$VIC_AGE_GROUP65+                   5.451 5.04e-08 ***
## NYPD_Data_analyzing$VIC_AGE_GROUPUNKNOWN               2.353  0.01863 *  
## NYPD_Data_analyzing$VIC_SEXM                          -1.061  0.28856    
## NYPD_Data_analyzing$VIC_SEXU                          -0.619  0.53588    
## NYPD_Data_analyzing$VIC_RACEASIAN / PACIFIC ISLANDER   1.760  0.07839 .  
## NYPD_Data_analyzing$VIC_RACEBLACK                      1.413  0.15769    
## NYPD_Data_analyzing$VIC_RACEBLACK HISPANIC             1.201  0.22976    
## NYPD_Data_analyzing$VIC_RACEUNKNOWN                    0.611  0.54131
## NYPD_Data_analyzing$VIC_RACEWHITE                      1.874  0.06101 .
## NYPD_Data_analyzing$VIC_RACEWHITE HISPANIC             1.563  0.11803
## NYPD_Data_analyzing$month_factor2                      0.908  0.36386
## NYPD_Data_analyzing$month_factor3                     -0.165  0.86919
## NYPD_Data_analyzing$month_factor4                      0.343  0.73177
## NYPD_Data_analyzing$month_factor5                      0.098  0.92207
## NYPD_Data_analyzing$month_factor6                     -1.526  0.12690
## NYPD_Data_analyzing$month_factor7                     -1.477  0.13975
## NYPD_Data_analyzing$month_factor8                     -2.060  0.03939 *
## NYPD_Data_analyzing$month_factor9                      0.581  0.56119
## NYPD_Data_analyzing$month_factor10                     0.186  0.85255
## NYPD_Data_analyzing$month_factor11                    -0.164  0.87005
## NYPD_Data_analyzing$month_factor12                     1.970  0.04888 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1540259)
##
##     Null deviance: 4250.5  on 27307  degrees of freedom
## Residual deviance: 4201.7  on 27279  degrees of freedom
## AIC: 26444
##
## Number of Fisher Scoring iterations: 2
```

The results indicate that certain factors, such as the victim's age, are significantly correlated. Region, on the other hand, is only significantly correlated with MANHATTAN. The factor of month shows significance at a confidence level of 0.05. These findings are based on our model analysis using the data we have obtained.

Bias Identification and conclusion

The data obtained allows for analysis, and statistical tests provide insights at the data level. However, it's essential to remember that such analytically results are heavily dependent on the data source. For instance, when reviewing the initial data, we observed a significant number of missing values (NA), which can be a limitation. Despite our imputation efforts, these missing values could still introduce bias into the analysis compared to the true data.

When making judgments, it's crucial to have a clear understanding of the data limitations and base your analysis on the information available. The most important aspect, in my opinion, is to remain humble and avoid making dogmatic conclusions based solely on your analysis.