NYPDData

Samuel Head

2023-08-18

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document.

Before we being the assignmnet we are first going to load the packages needed to perform our analysis of the NYPD Shooting Incident Data.

Step One: Start an RMD document

Start an RMD document that describes and imports the shooting project data set in a reproducible manner.

Uplodaing Data

The next step in our prcoess is to uplaod the NYPD Shooting Incident Data. The data is currenlty a CSV file that we donwloaded from the internet. Note: Having trouble downloading CSV to computer then upoading to R Studio. Solution, copy the direct link and have RStudio read csv from URL.

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

```
## 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.
```

After uploading the data, we run a test to make sure the data was properly retrieved from the website.

```
nypd_data
```

```
## # 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>
```

As of this step we have sucessfully created a RStudio Markdown file and have sucessfully uploaded the required data for this assignment.

Describe Data

From viewing the data. The files contains multiple columns labeled: INCIDENT_KEY, OCCUR_DAT, OCCUR_TIME, BORO, LOC_OF_OCCUR_DESC, PRESCINT, JURISDICTION_CODE, LOC_CLASSFCTN_DESC, LOCATION_DESC, STATISTICAL_MURDER_FLAG, PERP_AGE_GROUP, PERP_RACE, VIC_AGE_GROUP, VIC_SEX, VIC_RACE, X_COOR, Y_COOR, LATITUDE, LONGITUDE, LON_LAT. The dataset contain columns that are misssing data. For my analysis I will also not be utilizing all the columns. Thereore, I need to hihglihgt which columns I will be using for my analysis. The data also hold information regarding shooting incident in New York. I will be able to learn more about the data as I go through my analysis.

We can also summarize the data using the summary function.

```
summary(nypd_data)
```

```
##   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
```

Step 2: Tidy and Transform 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.

Our goal is to clean our data now. First we want to find which columns are unnecessary and which columns are the wrong data type. The columns that I want to keep for my analysis are: -Incident_Key -Occur_Date -Occur_Time -Boro -Precinct

I also noticed that the OCCUR_DATE is in the wrong format. I need to change the format from character to date. I would also like to create a more condensed data frame containing only the columns that I am interested in.

```
data_frame = nypd_data %>%
  select(c(INCIDENT_KEY, OCCUR_DATE, OCCUR_TIME, BORO, PRECINCT)) %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE))

data_frame
```

```
## # A tibble: 27,312 x 5
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO     PRECINCT
##           <dbl> <date>     <time>     <chr>       <dbl>
##  1    228798151 2021-05-27 21:30      QUEENS        105
##  2    137471050 2014-06-27 17:40      BRONX          40
##  3    147998800 2015-11-21 03:56      QUEENS        108
##  4    146837977 2015-10-09 18:30      BRONX          44
##  5     58921844 2009-02-19 22:58      BRONX          47
##  6    219559682 2020-10-21 21:36      BROOKLYN       81
##  7     85295722 2012-06-17 22:47      QUEENS        114
##  8     71662474 2010-03-08 19:41      BROOKLYN       81
##  9     83002139 2012-02-05 05:45      QUEENS        105
## 10     86437261 2012-08-26 01:10      QUEENS        101
## # i 27,302 more rows
```

Next, I am producing a summary of the new data frame.

```
summary(data_frame)
```

```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME          BORO
##  Min.   :  9953245   Min.   :2006-01-01   Length:27312      Length:27312
##  1st Qu.: 63860880   1st Qu.:2009-07-18   Class1:hms        Class :character
##  Median : 90372218   Median :2013-04-29   Class2:difftime   Mode  :character
##  Mean   :120860536   Mean   :2014-01-06   Mode  :numeric
##  3rd Qu.:188810230   3rd Qu.:2018-10-15
##  Max.   :261190187   Max.   :2022-12-31
##     PRECINCT
##  Min.   :  1.00
##  1st Qu.: 44.00
##  Median : 68.00
##  Mean   : 65.64
##  3rd Qu.: 81.00
##  Max.   :123.00
```

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?

From what I can tell, there seems to be no missing information. Therefore, I can continue with my analysis. For the next step of the project I have to add a minimum of two visualizations and some analyis to my R markdown.

For the first half of my visualization and analysis portion. I would like to see which areas have the highest crime. I can do this by analyzing the dataframe to see which BORO occurs most in the data. For the second half of my analysis, I will be determining which year did the most cases appear. These two analyses can help provide insight into which area has the most shootings and if crime has increased, decreased, or has not changed over the years. To begin the analysis portion, we need to find how many distinct BORO and which years the incodents occur.

First analysis and visualization

```r
unique_cities <- unique(data_frame$BORO)

unique_cities
```

```
## [1] "QUEENS"        "BRONX"         "BROOKLYN"      "MANHATTAN"
## [5] "STATEN ISLAND"
```

Since, we have now identified the unique cities in the data we can use them to further our analysis. Create a loop to count how often each city occurs in the dataframe. We can create a loop to do though the BORO column and count the number of time each city name occurred.

```r
Queens = 0
Bronx = 0
Brooklyn = 0
Manhattan = 0
StatenIsland = 0

for (i in 1:length(data_frame$BORO)){
  if (data_frame$BORO[i] == "QUEENS")
  {
    Queens = Queens +1
  }
  else if (data_frame$BORO[i]== "BRONX")
  {
    Bronx = Bronx +1
  }
  else if (data_frame$BORO[i] == "BROOKLYN")
  {
    Brooklyn = Brooklyn +1
  }
  else if (data_frame$BORO[i] == "MANHATTAN" )
  {
    Manhattan = Manhattan +1

  }
  else if (data_frame$BORO[i]== "STATEN ISLAND")
  {
    StatenIsland = StatenIsland +1
  }
}



total = Queens + Bronx+ Brooklyn + Manhattan+ StatenIsland
print(paste("There was a total of ", total, " cases throughout Queens, Bronx, Brooklyn, Manhattan, and S
```

```
## [1] "There was a total of  27312  cases throughout Queens, Bronx, Brooklyn, Manhattan, and Staten Isl
```

```r
print(paste("There was a total of", Bronx," cases in the Bronx."))
```

```
## [1] "There was a total of 7937  cases in the Bronx."
```

```r
print(paste("There was a total of", Queens," cases in the Queens."))
```

```
## [1] "There was a total of 4094  cases in the Queens."
```

```r
print(paste("There was a total of", Brooklyn," cases in the Brooklyn."))
```

```
## [1] "There was a total of 10933  cases in the Brooklyn."
```

```r
print(paste("There was a total of", Manhattan," cases in the Manhattan."))
```

```
## [1] "There was a total of 3572  cases in the Manhattan."
```

```r
print(paste("There was a total of", StatenIsland," cases in the Staten Island."))
```

```
## [1] "There was a total of 776  cases in the Staten Island."
```

Since we know which cities are involved and how many casses occur in each city. We can now create a bar graph to represent the data given.

From observing the above graph we can tell that the city that has the most incidents is Brooklyn while the city with the least amount of incidents is Staten Island.

Second analysis and visualization

For the next portion of my analysis I will determine which year had the most crime. I will begin by editing the the data frame. I am going to seperate OCCUR_DATE into three sections: Year, Month, and Day. By doing this I can create a loop to go thru each year and tally how many cases occured in each year.

```r
data_frame2 <- separate(data_frame, col = OCCUR_DATE, into = c("Year", "Month", "Day"), sep = "-")
```

I am going to find the max and min of the years that the incidents occured to determine which years the data set covers. I am also going to determine which days the data set being and ends to determine if the data cover each year completely.

```r
begin_year <- min(data_frame2$Year)
end_year <-max(data_frame2$Year)

begin_day <- min(data_frame$OCCUR_DATE)
end_day <- max(data_frame$OCCUR_DATE)

print(paste("The data set begins recording data in", begin_year, "until", end_year, "."))
```

```
## [1] "The data set begins recording data in 2006 until 2022 ."
```

```r
print(paste("The data set begins recording data on", begin_day, "until", end_day, "."))
```

```
## [1] "The data set begins recording data on 2006-01-01 until 2022-12-31 ."
```

The incidents are recorded from 2006 to 2022. Therefore, we can create a loop to tally the number of incidents for each year. The incidents cover a span of 16 years.

```
total_2006 = 0
total_2007 = 0
total_2008 = 0
total_2009 = 0
total_2010 = 0
total_2011 = 0
total_2012 = 0
total_2013 = 0
total_2014 = 0
total_2015 = 0
total_2016 = 0
total_2017 = 0
total_2018 = 0
total_2019 = 0
total_2020 = 0
total_2021 = 0
total_2022 = 0


for (i in 1:length(data_frame2$Year))
{
  if (data_frame2$Year[i] == 2006)
  {
    total_2006 = total_2006 + 1
  }
  else if (data_frame2$Year[i] == 2007)
  {
    total_2007 = total_2007 + 1
  }
    else if (data_frame2$Year[i] == 2008)
  {
    total_2008 = total_2008 + 1
  }
    else if (data_frame2$Year[i] == 2009)
  {
    total_2009 = total_2009 + 1
  }
    else if (data_frame2$Year[i] == 2010)
  {
    total_2010 = total_2010 + 1
  }
    else if (data_frame2$Year[i] == 2011)
  {
    total_2011 = total_2011 + 1
  }
    else if (data_frame2$Year[i] == 2012)
  {
    total_2012 = total_2012 + 1
  }
    else if (data_frame2$Year[i] == 2013)
  {
    total_2013 = total_2013 + 1
  }
```

```
    else if (data_frame2$Year[i] == 2014)
  {
    total_2014 = total_2014 + 1
  }
    else if (data_frame2$Year[i] == 2015)
  {
    total_2015 = total_2015 + 1
  }
    else if (data_frame2$Year[i] == 2016)
  {
    total_2016 = total_2016 + 1
  }
    else if (data_frame2$Year[i] == 2017)
  {
    total_2017 = total_2017 + 1
  }
  else if (data_frame2$Year[i] == 2018)
  {
    total_2018 = total_2018 + 1
  }
  else if (data_frame2$Year[i] == 2019)
  {
    total_2019 = total_2019 + 1
  }
  else if (data_frame2$Year[i] == 2020)
  {
    total_2020 = total_2020 + 1
      }
  else if (data_frame2$Year[i] == 2021)
  {
    total_2021 = total_2021 + 1
  }
  else if (data_frame2$Year[i] == 2022)
  {
    total_2022 = total_2022 + 1
  }
}


print(paste("The was a total of", total_2006, "shooting incidents in 2006."))
```

```
## [1] "The was a total of 2055 shooting incidents in 2006."
```

```
print(paste("The was a total of", total_2007, "shooting incidents in 2007."))
```

```
## [1] "The was a total of 1887 shooting incidents in 2007."
```

```
print(paste("The was a total of", total_2008, "shooting incidents in 2008."))
```

```
## [1] "The was a total of 1959 shooting incidents in 2008."
```

```r
print(paste("The was a total of", total_2009, "shooting incidents in 2009."))
```

```
## [1] "The was a total of 1828 shooting incidents in 2009."
```

```r
print(paste("The was a total of", total_2010, "shooting incidents in 2010."))
```

```
## [1] "The was a total of 1912 shooting incidents in 2010."
```

```r
print(paste("The was a total of", total_2011, "shooting incidents in 2011."))
```

```
## [1] "The was a total of 1939 shooting incidents in 2011."
```

```r
print(paste("The was a total of", total_2012, "shooting incidents in 2012."))
```

```
## [1] "The was a total of 1717 shooting incidents in 2012."
```

```r
print(paste("The was a total of", total_2013, "shooting incidents in 2013."))
```

```
## [1] "The was a total of 1339 shooting incidents in 2013."
```

```r
print(paste("The was a total of", total_2014, "shooting incidents in 2014."))
```

```
## [1] "The was a total of 1464 shooting incidents in 2014."
```

```r
print(paste("The was a total of", total_2015, "shooting incidents in 2015."))
```

```
## [1] "The was a total of 1434 shooting incidents in 2015."
```

```r
print(paste("The was a total of", total_2016, "shooting incidents in 2016."))
```

```
## [1] "The was a total of 1208 shooting incidents in 2016."
```

```r
print(paste("The was a total of", total_2017, "shooting incidents in 2017."))
```

```
## [1] "The was a total of 970 shooting incidents in 2017."
```

```r
print(paste("The was a total of", total_2018, "shooting incidents in 2018."))
```

```
## [1] "The was a total of 958 shooting incidents in 2018."
```

```r
print(paste("The was a total of", total_2019, "shooting incidents in 2019."))
```

```
## [1] "The was a total of 967 shooting incidents in 2019."
```

```
print(paste("The was a total of", total_2020, "shooting incidents in 2020."))
```

```
## [1] "The was a total of 1948 shooting incidents in 2020."
```

```
print(paste("The was a total of", total_2021, "shooting incidents in 2021."))
```

```
## [1] "The was a total of 2011 shooting incidents in 2021."
```

```
print(paste("The was a total of", total_2022, "shooting incidents in 2022."))
```

```
## [1] "The was a total of 1716 shooting incidents in 2022."
```

Now that I have determined how many incidents occur in each year. I am going to create a visualization to help display the information I found.

Model

Lastly, I need to include a model to go along with my above diagrams. I am going to model a predictive line of the total number of cases throughout the years over the above bar graph containig the total number of cases each year.

Questions that arose from analysis

During my analysis there were additionally questions that arose. While analysing the total cases per year and total cases per city there were multiple question that arose that could use further analysis. 1) How does crime overlap with the time of the year? For instace, is there more or less crime in the summer versus the winter? 2) how do individual cities crime rates differ throughout the year? We analyzed the overall state but how have individual cities total incidents change throuhghout the year? Are cities reporting more or less incidents. 3) How many incidents occured that have not been documented? 4) How does age and race relate to incident rate?

All of these questions arose and could use further analysis.

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.

To conclude my report, I discovered that there was a total of 27,312 shooting incident cases that spread across five distinct districts in New York over the course of 16 year (2006-2022). The city that expereinced the most incidents over these year was Brooklyn while the least being Staten Island. Additionally, over the years New York expereinced the most incidents in 2006. With the least amount of incidents being in 2018. From the analysis, the overall number of shooting incidents seemed to be decreasing until recent years where a large increase occurred in 2020. Overall, from the model that overlaps the bar graph containing the number of total incidents the trend shows a decreasing amount of incidents over the years.

During the analysis I came across selection bias. I wanted to focus on specific years. For instance, specifically studying the number of incidents during a designated timeline such as the most recent years (2020-2022). However, I furthered my analysis to cover the whole timeline that occurred in the the data set (2006-2022). Expanding my analysis to cover the whole data set helped to mitigate bias because during the most recent years there was a spike in incidents. Without acknowledging the previous decrease in incidents my analysis would have greatly differed.