Analysis of NYPD Shooting Incident Data (Historic)

Jingheng C

11/17/2021

A quick analysis of the NYPD Shooting Incident Data. We want to see whether there is a temporal trend with the incident count.

Data Publisher: data.cityofnewyork.us Data Maintainer: NYC OpenData

Importing data and libraries

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- 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
```

```
NYPD_Shooting_data = read.csv("NYPD_Shooting_Incident_Data__Historic_.csv")

summary(NYPD_Shooting_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME             BORO
##  Min.   :  9953245   Length:23568       Length:23568       Length:23568
##  1st Qu.: 55317014   Class :character   Class :character   Class :character
##  Median : 83365370   Mode  :character   Mode  :character   Mode  :character
```

```
##  Mean   :102218616
##  3rd Qu.:150772442
##  Max.   :222473262
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:23568       Length:23568
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   Class :character
##  Median : 69.00   Median :0.0000    Mode  :character   Mode  :character
##  Mean   : 66.21   Mean   :0.3323
##  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:23568       Length:23568       Length:23568       Length:23568
##  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:23568       Length:23568       Length:23568       Length:23568
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:23568
##  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
##
```

Data Cleaning & Transformation

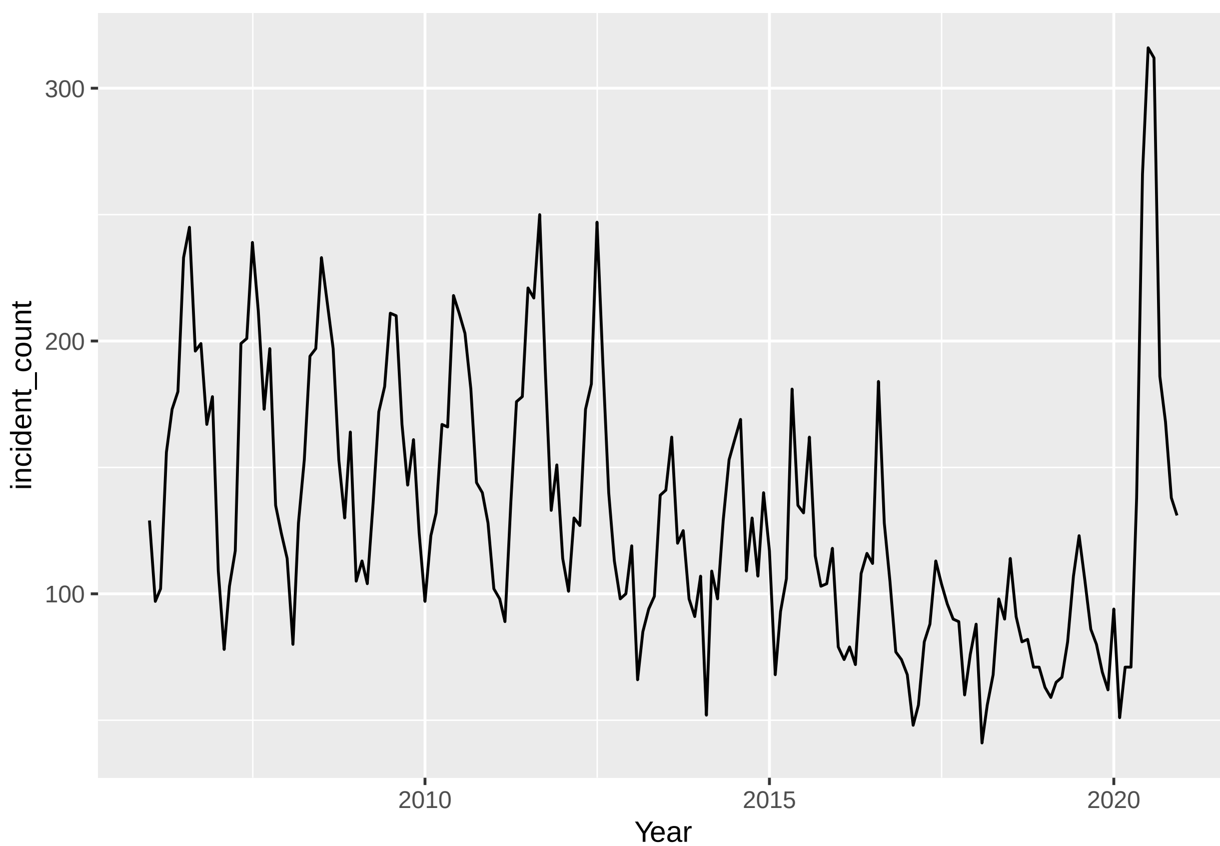
```r
# Extract year&month from OCCUR_DATE because we need to group the incident
# by month and year for this analysis
NYPD_Shooting_data = NYPD_Shooting_data %>%
  mutate(OCCUR_DATE_Month_Year = format(mdy(OCCUR_DATE),"%Y-%m"))

# Aggregation to get the monthly incident count

Monthly_incident_count = NYPD_Shooting_data %>% group_by(OCCUR_DATE_Month_Year) %>%
  summarise(incident_count = n())

Monthly_incident_count$date = ym(Monthly_incident_count$OCCUR_DATE_Month_Year)

summary(Monthly_incident_count)
```

```
##  OCCUR_DATE_Month_Year incident_count       date
##  Length:180            Min.   : 41.0   Min.   :2006-01-01
##  Class :character      1st Qu.: 92.5   1st Qu.:2009-09-23
##  Mode  :character      Median :119.5   Median :2013-06-16
##                        Mean   :130.9   Mean   :2013-06-16
##                        3rd Qu.:167.0   3rd Qu.:2017-03-08
##                        Max.   :316.0   Max.   :2020-12-01
```

Data Visualization

We can easily spot a huge spike in incident count around mid-2020, this spike is likely caused by the COVID-19 recession. There seems to be a downward trend, further quantitative analysis is needed to conclude whether there is a trend.

```
timeseries_plot = ggplot(Monthly_incident_count,aes(date,incident_count))+
  geom_line()+
  xlab('Year')
timeseries_plot
```



Regression analysis of the trend

We will use the index of row to acts as a variable of the linear trend

```
# Adding index column
Monthly_incident_count$Trend = seq.int(nrow(Monthly_incident_count))

# Linear Regression
model = lm(incident_count ~ Trend,data = Monthly_incident_count)
summary(model)
```

```
##
## Call:
## lm(formula = incident_count ~ Trend, data = Monthly_incident_count)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -85.909 -32.869  -8.484  27.752 221.491
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 169.94358    7.17906  23.672  < 2e-16 ***
## Trend        -0.43105    0.06879  -6.266 2.71e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47.96 on 178 degrees of freedom
## Multiple R-squared:  0.1807, Adjusted R-squared:  0.1761
## F-statistic: 39.26 on 1 and 178 DF,  p-value: 2.712e-09
```

p = 2.3e-15 < 0.05 We conclude that there is significant evidence to suggest the presence of a linear trend in the incident count.

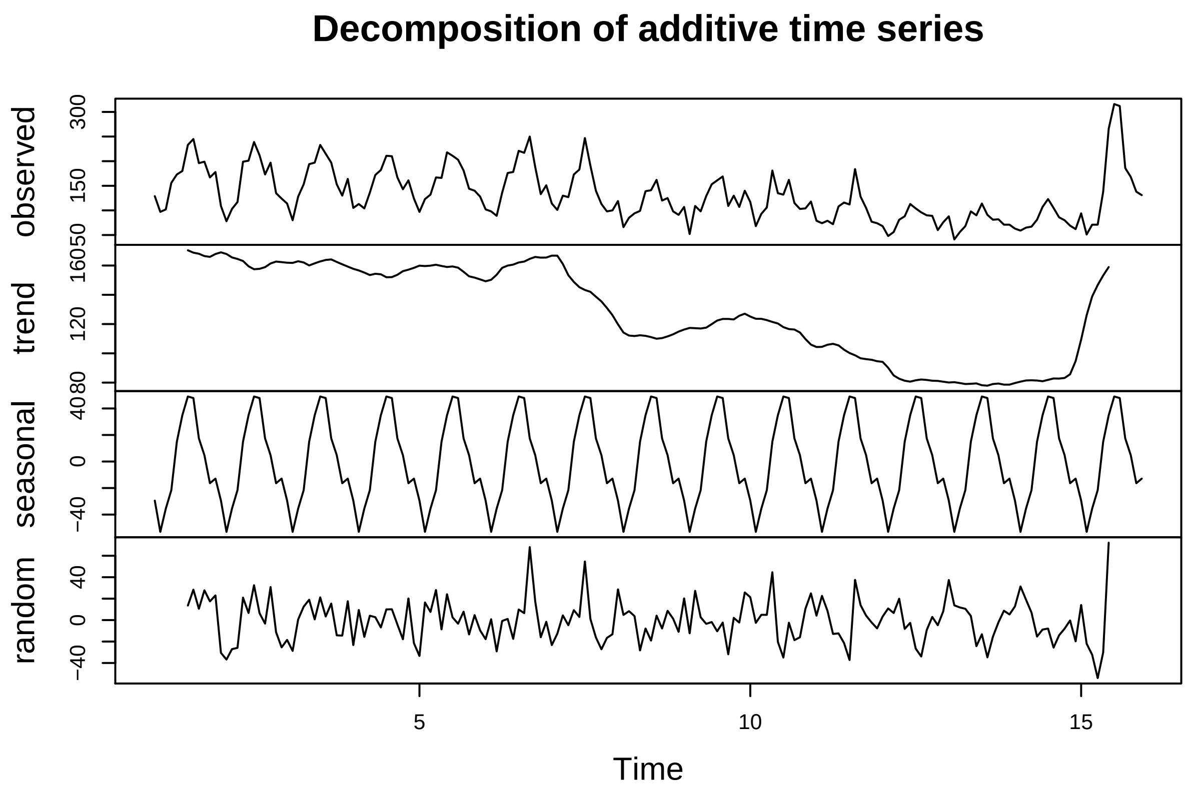
Adding more independent variable

In the previous regression analysis, R-squared is only 0.29. Around 70% of the variance is unexplained by our model. Seasonality seems to be present in the data,

Data Visualization

Decompose the data to get a better visual representation of the seasonality.

```
# Convert the data into time series
ts = ts(Monthly_incident_count$incident_count,frequency=12)

# Time series decomposition to visualize the seasonality

ts = decompose(ts)

# Visualize the decomposed time series

plot(ts)
```



Regression Analysis

```
# Adding variable for Month
Monthly_incident_count$Month = as.factor(month(Monthly_incident_count$date))


model_se = lm(incident_count ~ Trend + Month ,data = Monthly_incident_count)

summary(model_se)
```

```
##
## Call:
## lm(formula = incident_count ~ Trend + Month, data = Monthly_incident_count)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.325 -20.025  -5.405  14.925 167.281
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 138.59809    9.71333  14.269  < 2e-16 ***
## Trend        -0.45017    0.04888  -9.209  < 2e-16 ***
## Month2      -23.28316   12.41645  -1.875 0.062513 .
## Month3       -5.96632   12.41674  -0.481 0.631496
```

```
## Month4       10.95052   12.41722   0.882 0.379108
## Month5       46.46736   12.41790   3.742 0.000251 ***
## Month6       65.78420   12.41876   5.297 3.67e-07 ***
## Month7       88.90104   12.41982   7.158 2.47e-11 ***
## Month8       87.75122   12.42107   7.065 4.15e-11 ***
## Month9       51.20139   12.42251   4.122 5.91e-05 ***
## Month10      38.31823   12.42415   3.084 0.002389 **
## Month11      16.70174   12.42598   1.344 0.180740
## Month12      20.08524   12.42799   1.616 0.107954
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34 on 167 degrees of freedom
## Multiple R-squared:  0.6136, Adjusted R-squared:  0.5858
## F-statistic:  22.1 on 12 and 167 DF,  p-value: < 2.2e-16
```

The p value of many seasonal dummy variables is less than 0.05. We can conclude that seasonality is present in the data. By adding months to our model, Adjusted R-squared increased to 0.5858. A majority of the variance is now explained by our model

Conclusion & Biases

The monthly shooting incidents count has a decreasing trend over time. Seasonal patterns can also be observed, the number of shooting incidents is significantly higher from May to October compared to other months.

One potential source of biases is with the data collection process, the data was collected and published by the government of New York City. There could be a political incentive for the government to publish data that suggest the number of shooting incidents is decreasing over the years. Ideally a third party NGO should verify the data is authenticate

```
sessionInfo()
```

```
## R version 4.0.4 (2021-02-15)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 22000)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
## system code page: 932
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lubridate_1.7.9.2 forcats_0.5.1     stringr_1.4.0     dplyr_1.0.4
```

```
##  [5] purrr_0.3.4       readr_1.4.0       tidyr_1.1.2       tibble_3.0.6
##  [9] ggplot2_3.3.3     tidyverse_1.3.0
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.1.0  xfun_0.28         haven_2.3.1       colorspace_2.0-0
##  [5] vctrs_0.3.6       generics_0.1.0    htmltools_0.5.1.1 yaml_2.2.1
##  [9] rlang_0.4.10      pillar_1.4.7      glue_1.4.2        withr_2.4.1
## [13] DBI_1.1.1         dbplyr_2.1.0      modelr_0.1.8      readxl_1.3.1
## [17] lifecycle_1.0.0   munsell_0.5.0     gtable_0.3.0      cellranger_1.1.0
## [21] rvest_0.3.6       evaluate_0.14     labeling_0.4.2    knitr_1.31
## [25] highr_0.8         broom_0.7.4       Rcpp_1.0.6        scales_1.1.1
## [29] backports_1.2.1   jsonlite_1.7.2    farver_2.0.3      fs_1.5.0
## [33] hms_1.0.0         digest_0.6.27     stringi_1.5.3     grid_4.0.4
## [37] cli_2.3.0         tools_4.0.4       magrittr_2.0.1    crayon_1.4.1
## [41] pkgconfig_2.0.3   ellipsis_0.3.1    xml2_1.3.2        reprex_1.0.0
## [45] assertthat_0.2.1  rmarkdown_2.6     httr_1.4.2        rstudioapi_0.13
## [49] R6_2.5.0          compiler_4.0.4
```