NYPD Shooting Incident Data

5/28/2021

Tidyverse was installed.

Import data

Data source and description: Data was retrieved from Data.gov. The name of the dataset is NYPD Shooting Incident Data (Historic). It is a list of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year. Each record represents a shooting incident in NYC and includes information about the event, the location and time of occurrence. In addition, information related to suspect and victim demographics is also included.

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

```
## [1] "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
```

```
nypd_shooting <- read.csv(url[1])
```

Tidy and Transform data

```
nypd_shooting <- nypd_shooting %>%
  select(-c(JURISDICTION_CODE,LOCATION_DESC,X_COORD_CD,Y_COORD_CD,Latitude,Longitude,Lon_Lat))

nypd_shooting$YEAR <- substr(nypd_shooting$OCCUR_DATE, nchar(nypd_shooting$OCCUR_DATE)-3, nchar(nypd_sh

nypd_shooting_by_year <- nypd_shooting %>%
  group_by(YEAR) %>%
  tally(name = "INCIDENTS_COUNT")

nypd_shooting_by_boro <- nypd_shooting %>%
  group_by(BORO)%>%
  tally(name = "INCIDENTS_COUNT")
```

Visualize Data

Plot 1: NYPD Shooting Incident by Year

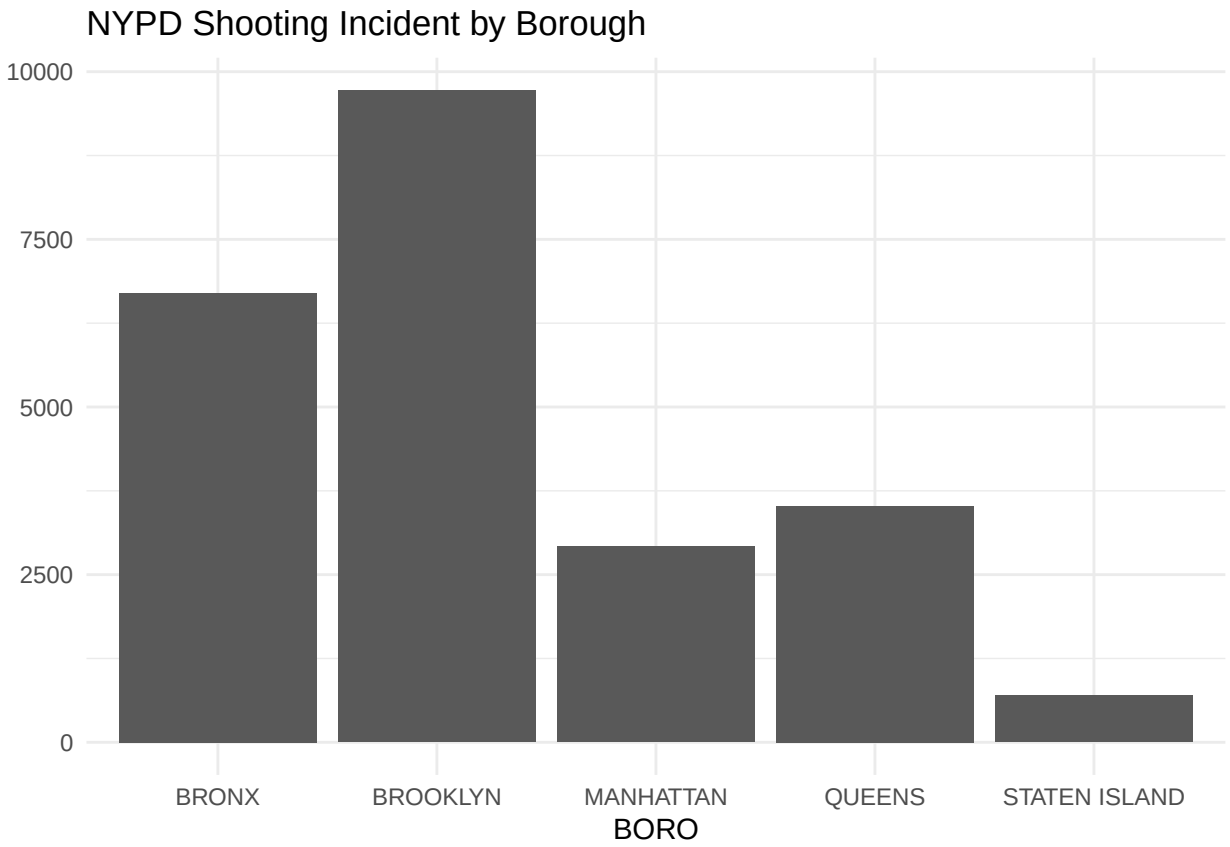
Visualize the number of shooting incidents by year to identify a potential trend

```
plot_nypd_shotting_by_year <- ggplot() + geom_line(aes(y = INCIDENTS_COUNT, x = YEAR, group = 1), data =
theme(legend.position = "bottom", axis.text.x = element_text(angle = 90)) +
labs(title = "NYPD Shooting Incident by Year", y = NULL)
plot_nypd_shotting_by_year
```

Plot 2: NYPD Shooting Incident by Borough

Visualize the number of shooting incidents by borough to see which borough had the most/least incidents

```
plot_nypd_shotting_by_boro <- ggplot(data = nypd_shooting_by_boro, aes(x = BORO, y = INCIDENTS_COUNT)) +
geom_bar(stat = "identity") +
theme_minimal() +
labs(title = "NYPD Shooting Incident by Borough", y = NULL)
plot_nypd_shotting_by_boro
```



Model data

Question of interest: Does the number of shooting incidents show a trend over the years? To answer this question, a linear regression model is first built, using year to predict number of incidents.

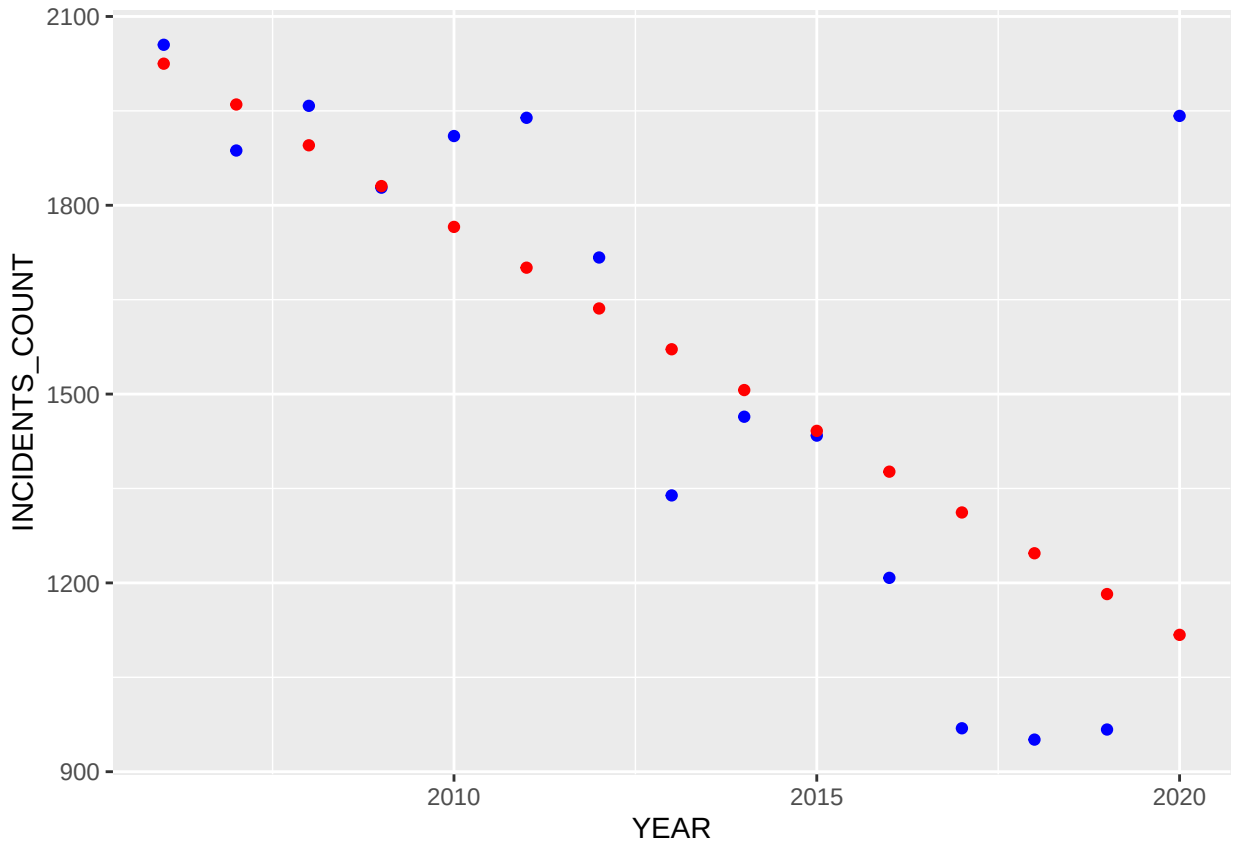
Fit a Linear Model

```
nypd_shooting_by_year$YEAR <- as.numeric(nypd_shooting_by_year$YEAR)
linearmod <- lm(INCIDENTS_COUNT ~ YEAR, data = nypd_shooting_by_year)
summary(linearmod)
```

```
##
## Call:
## lm(formula = INCIDENTS_COUNT ~ YEAR, data = nypd_shooting_by_year)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -342.90 -191.99   -7.55   71.82  824.58
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 132063.93   35152.83   3.757  0.00240 **
## YEAR           -64.83      17.46  -3.712  0.00261 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 292.2 on 13 degrees of freedom
## Multiple R-squared:  0.5146, Adjusted R-squared:  0.4772
## F-statistic: 13.78 on 1 and 13 DF,  p-value: 0.002609
```

```
nypd_shooting_by_year_pred <- nypd_shooting_by_year %>% mutate(pred = predict(linearmod))
nypd_shooting_by_year_pred %>% ggplot() +
geom_point(aes(x = YEAR, y = INCIDENTS_COUNT), color = "blue") +
geom_point(aes(x = YEAR, y = pred), color = "red")
```



According to the model output, the slope is -64.83 ($p < .01$), which indicates that year is an indicator of the number of shooting incidents. In other words, shooting incidents significantly decreased over the years. However, if we look at the plot that shows the predictions (red) and the actual numbers (blue), we can see that the prediction of 2020 is far away from the true value. It may suggest that linear model is not the best model to be used here.

Fit a Quadratic Regression Model

We can take a step further to see if a quadratic regression model (i.e. a second order model) would fit better.

```
nypd_shooting_by_year$YEAR2 <- nypd_shooting_by_year$YEAR^2
quadraticmod <- lm(INCIDENTS_COUNT ~ YEAR + YEAR2, data = nypd_shooting_by_year)
summary(quadraticmod)
```

```
##
## Call:
## lm(formula = INCIDENTS_COUNT ~ YEAR + YEAR2, data = nypd_shooting_by_year)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -329.41 -149.30   10.99  118.51  671.10
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.063e+07  1.825e+07   1.131    0.280
## YEAR        -2.044e+04  1.813e+04  -1.127    0.282
## YEAR2        5.060e+00  4.504e+00   1.123    0.283
##
## Residual standard error: 289.3 on 12 degrees of freedom
## Multiple R-squared:  0.5608, Adjusted R-squared:  0.4875
## F-statistic:  7.66 on 2 and 12 DF,  p-value: 0.007182
```

Conclusion: The R-squared for the linear regression model is 0.5146, which means the total variance in the number of shooting incidents explained by the model is 51.46%. The R-squared for the quadratic regression model is 0.5608. Therefore, the quadratic regression model seems to perform better.

Possible bias: Although the results may suggest that quadratic regression model fits better here, we didn't consider the huge impact of Covid-19 on the society in 2020. Therefore, the linear model may work better under normal circumstance. Bias could have been introduced here if we adopt the quadratic model without considering the uniqueness of 2020.

Version Information about R

```
sessionInfo()
```

```
## R version 4.1.0 (2021-05-18)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19041)
##
## 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
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] ggplot2_3.3.3  knitr_1.33     magrittr_2.0.1 dplyr_1.0.6
##
## loaded via a namespace (and not attached):
##  [1] munsell_0.5.0     tidyselect_1.1.1  colorspace_2.0-1  R6_2.5.0
##  [5] rlang_0.4.11      fansi_0.4.2       highr_0.9         stringr_1.4.0
##  [9] tools_4.1.0       grid_4.1.0        gtable_0.3.0      xfun_0.23
```

```
## [13] utf8_1.2.1        DBI_1.1.1         withr_2.4.2       htmltools_0.5.1.1
## [17] ellipsis_0.3.2    assertthat_0.2.1  yaml_2.2.1        digest_0.6.27
## [21] tibble_3.1.2      lifecycle_1.0.0   crayon_1.4.1      farver_2.1.0
## [25] purrr_0.3.4       vctrs_0.3.8       glue_1.4.2        evaluate_0.14
## [29] rmarkdown_2.8     labeling_0.4.2    stringi_1.6.1     compiler_4.1.0
## [33] pillar_1.6.1      scales_1.1.1      generics_0.1.0    pkgconfig_2.0.3
```