NYPD Data

Andrew

1/29/2022

Include needed packages

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.2     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
```

Import:

Import the NYPD Data from the source. The data included all shooting incident Data from 2006-2021 and was manually extracted and reviewed by Office of Management Analysis and Planning before posting. Here is the link to read more about it. https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic

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

Tidy:

Convert to tibble and remove data that isn't required for analysis

```
NYPD_Data <- tibble(NYPD_Data)
NYPD_Data <- select(NYPD_Data,OCCUR_DATE,VIC_AGE_GROUP)
```

Transform:

Transform data into format that will be easy for visualization and analysis. Total shootings are added to be able to plot total shootings over time.

```
NYPD_Data <- NYPD_Data %>%
  mutate(OCCUR_DATE=mdy(OCCUR_DATE))%>%arrange(OCCUR_DATE)
NYPD_Data <- NYPD_Data %>%
  mutate(total_shootings=(1:23585))
#Output Summary to Check for Missing Data
summary(NYPD_Data)
```

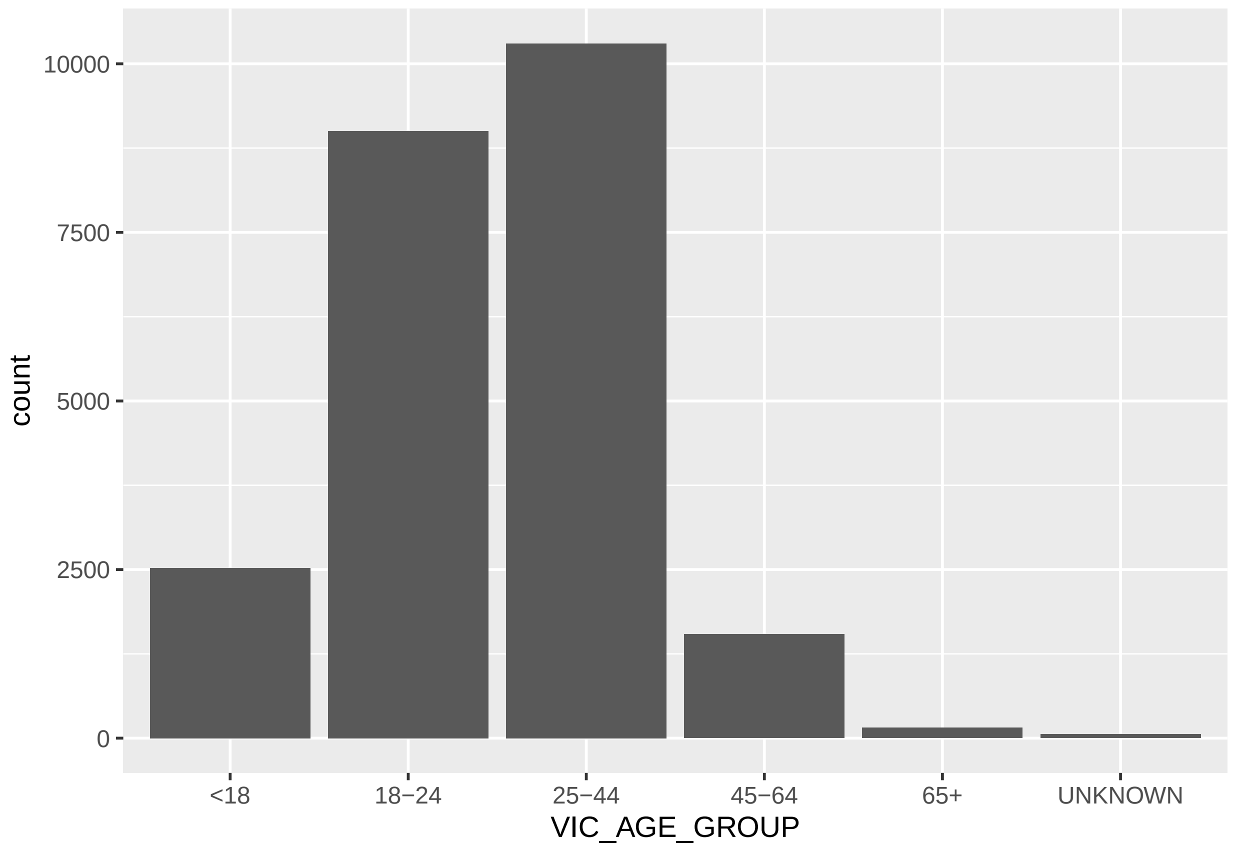
```
##    OCCUR_DATE         VIC_AGE_GROUP      total_shootings
##  Min.   :2006-01-01   Length:23585       Min.   :    1
##  1st Qu.:2008-12-31   Class :character   1st Qu.: 5897
##  Median :2012-02-27   Mode  :character   Median :11793
##  Mean   :2012-10-05                      Mean   :11793
##  3rd Qu.:2016-03-02                      3rd Qu.:17689
##  Max.   :2020-12-31                      Max.   :23585
```

Note: The data being analyzed does not have any NA values, but there is a portion of the victims who age group is unknown. However, it is a small amount compared to most categories other than 65+.

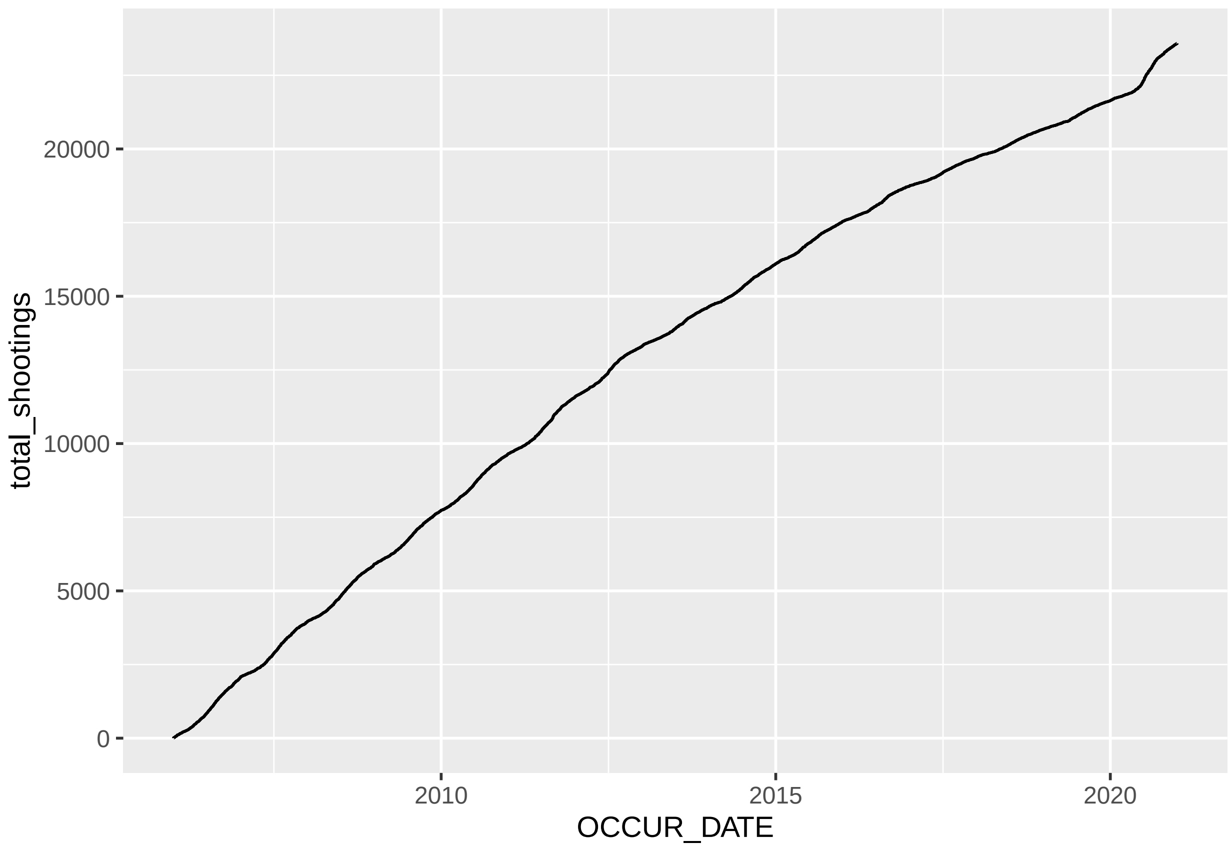
Visualize:

Here we will look at a histogram occurrences by victim age group and a plot of occurrences over time of day

```
#create a histogram for age group
ggplot(data=NYPD_Data)+geom_bar(mapping = aes(x=VIC_AGE_GROUP))
```



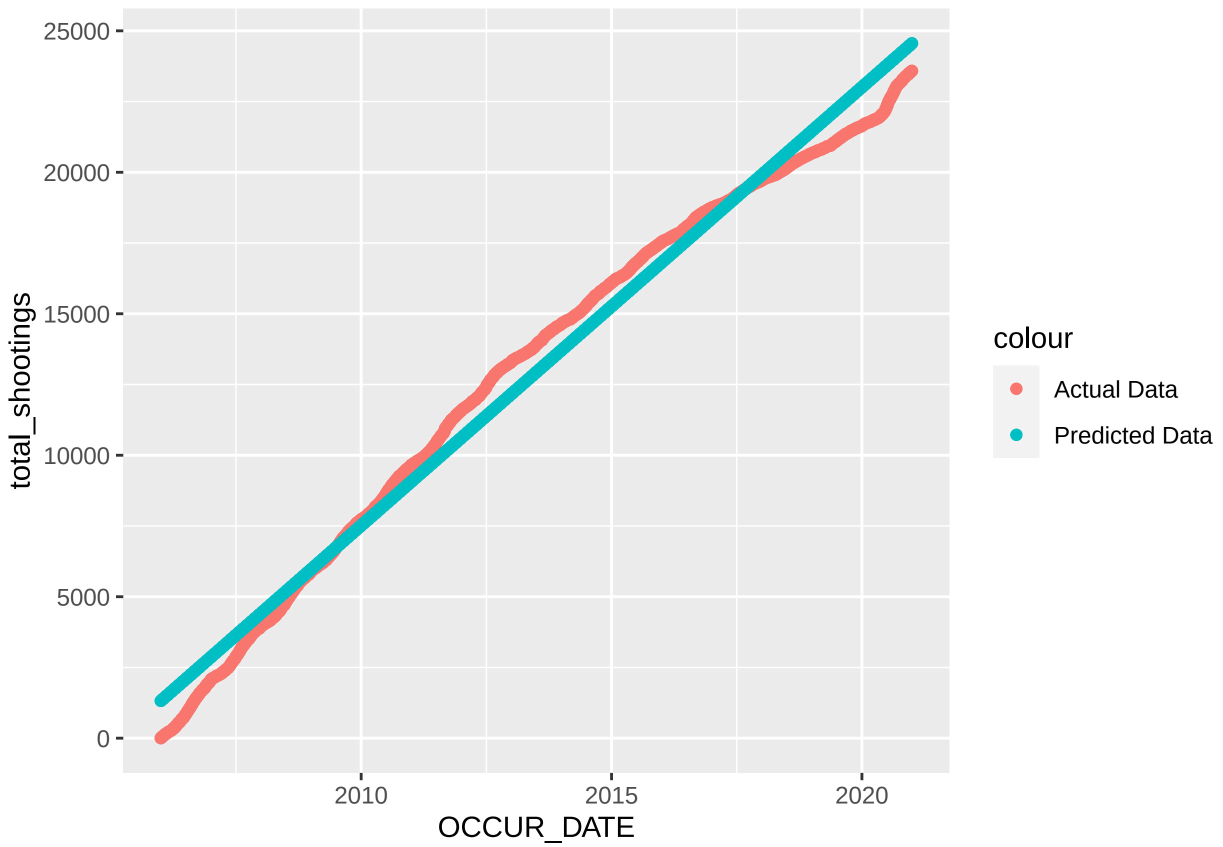
```
#plot the total shooting incendents vs. date
ggplot(data=NYPD_Data) + geom_line(mapping = aes(x=OCCUR_DATE,y=total_shootings))
```



Model:

As a basic model we will see if overall shootings have a linear relationship to the elapsed time. (i.e. the amount of shootings per day is relatively constant)

```
#Create a linear model for total shootings based on occurrence date
mod<-lm(total_shootings~OCCUR_DATE,NYPD_Data)
#Add predition to tibble
NYPD_Data <- NYPD_Data %>% mutate(pred=as.integer(predict(mod)))
#Plot the predicted and actual data
NYPD_Data %>% ggplot()+geom_point(aes(x=OCCUR_DATE,y=total_shootings, color="Actual Data"))+
  geom_point(aes(x=OCCUR_DATE,y=pred, color="Predicted Data"))
```



Analysis:

1. It appears that the majority of victims fall between the ages of 18 and 44. If I were to look at the data in more detail I would compare the number of victims to the proportion of the population in that age category to see if a certain age groups is more prone to being a victim of a shooting incident or if the victim age group is proportionate to general population trends. There is also some uncertainty in the data due to the "unknown" category. While the impact of the number of unknown is negligible for most age groups,it could have make a significant difference if all or most of the unknowns were 65+.

2. Total shootings over time appears to look generally linear for the time frame given. However, there is a noticeable uptick for the last half of 2020 and 2021. If I were to look at the data further I would explore:

- Did COVID-19 have an impact on the rate of shootings?
- Were the majority of shootings reported during COVID at someones residence or outside of their home?

Bias:

- There could be bias in the age data if some victims in the older age demographic go unreported due to having no family members or close connections remaining to identify them.
- There could be bias in the occurrence data if the person collecting the data had a certain political affiliation and wanted to prove that rules and regulations had, or didn't have, an effect on shootings.