NYPD_Shooting_Historical

Ji

2022/3/13

Import and Clean data

Firstly, we import the data and check whether there exists errors, inconsistencies, missing data or irrelevant values.
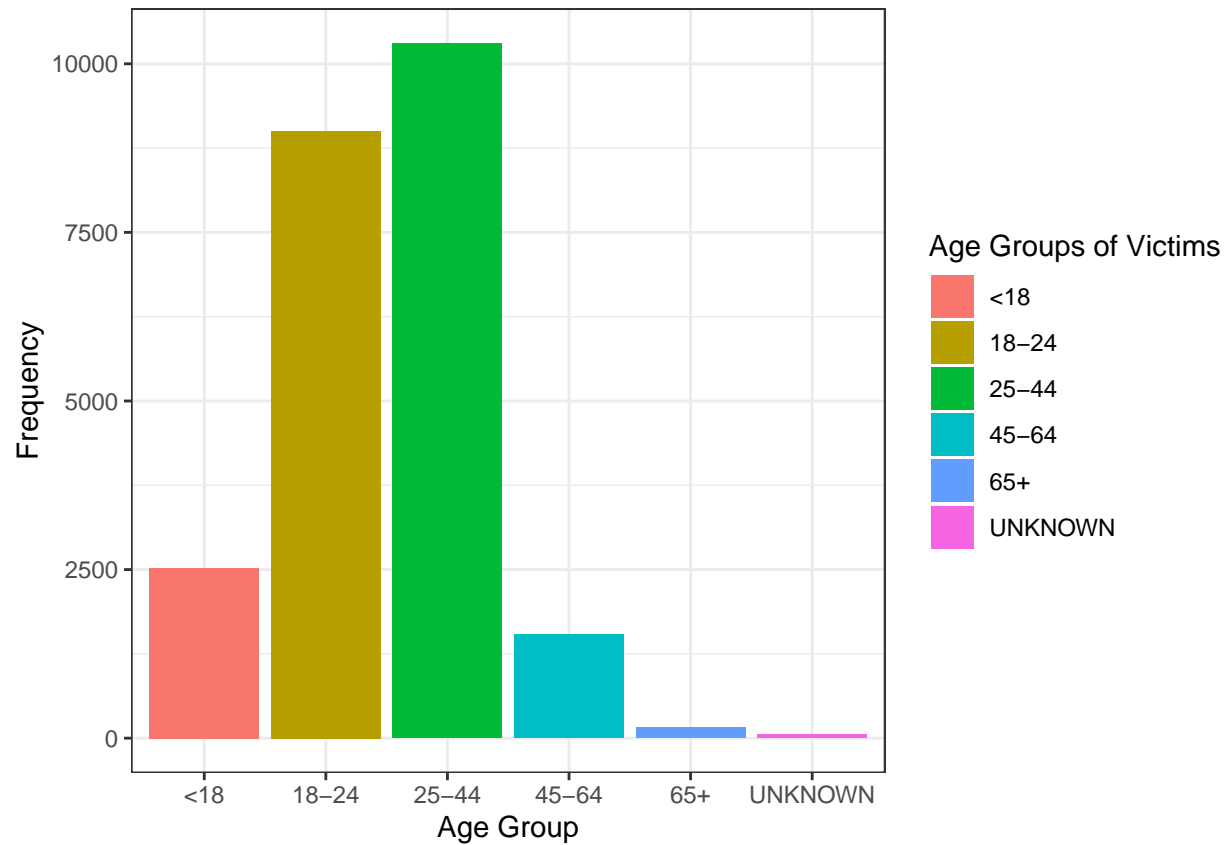
```r
df = read.csv("NYPD_Shooting_Incident_Data__Historic_.csv")
```
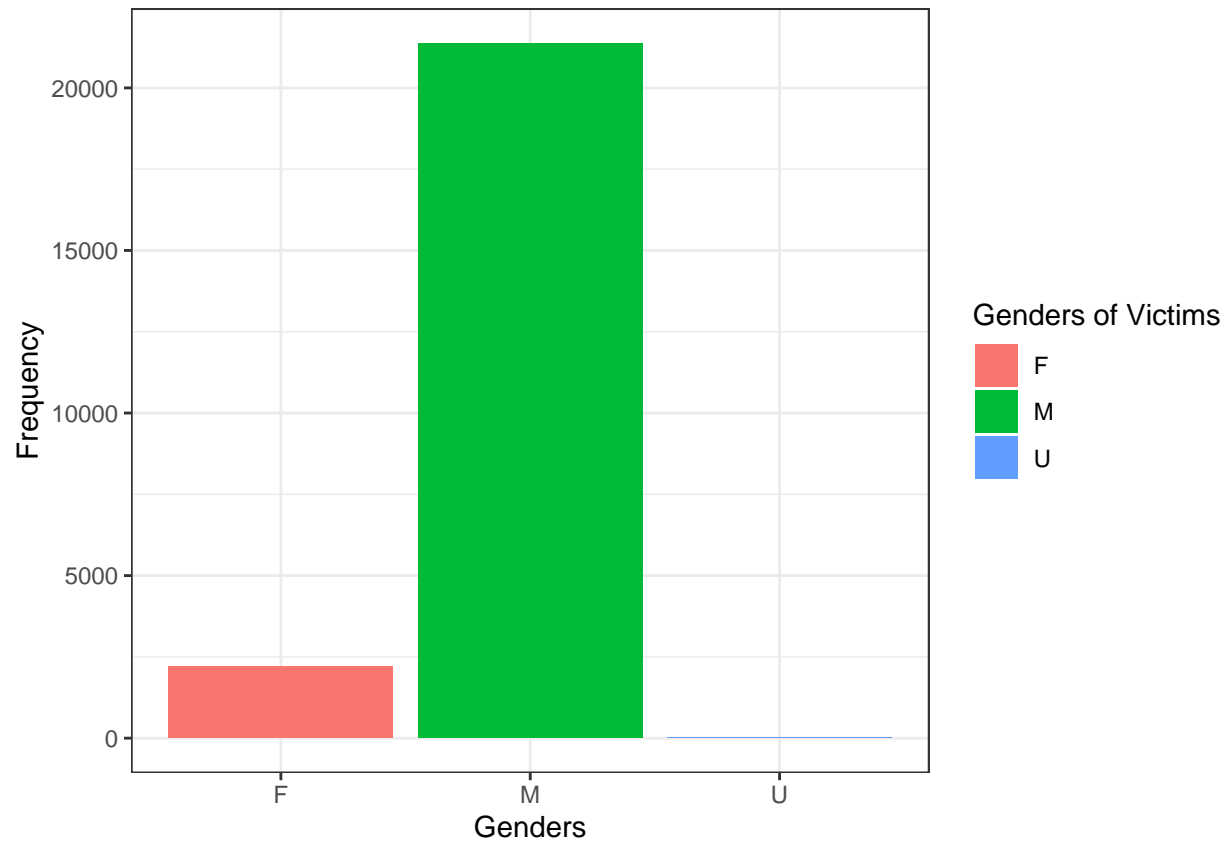
Analysis of data

After importing the data, we plan to analyze the data to check the distribution of age groups and genders (F: female, M: male, U: unknown) amongst victims and to find the correlation between races of victims and locations (BORO).
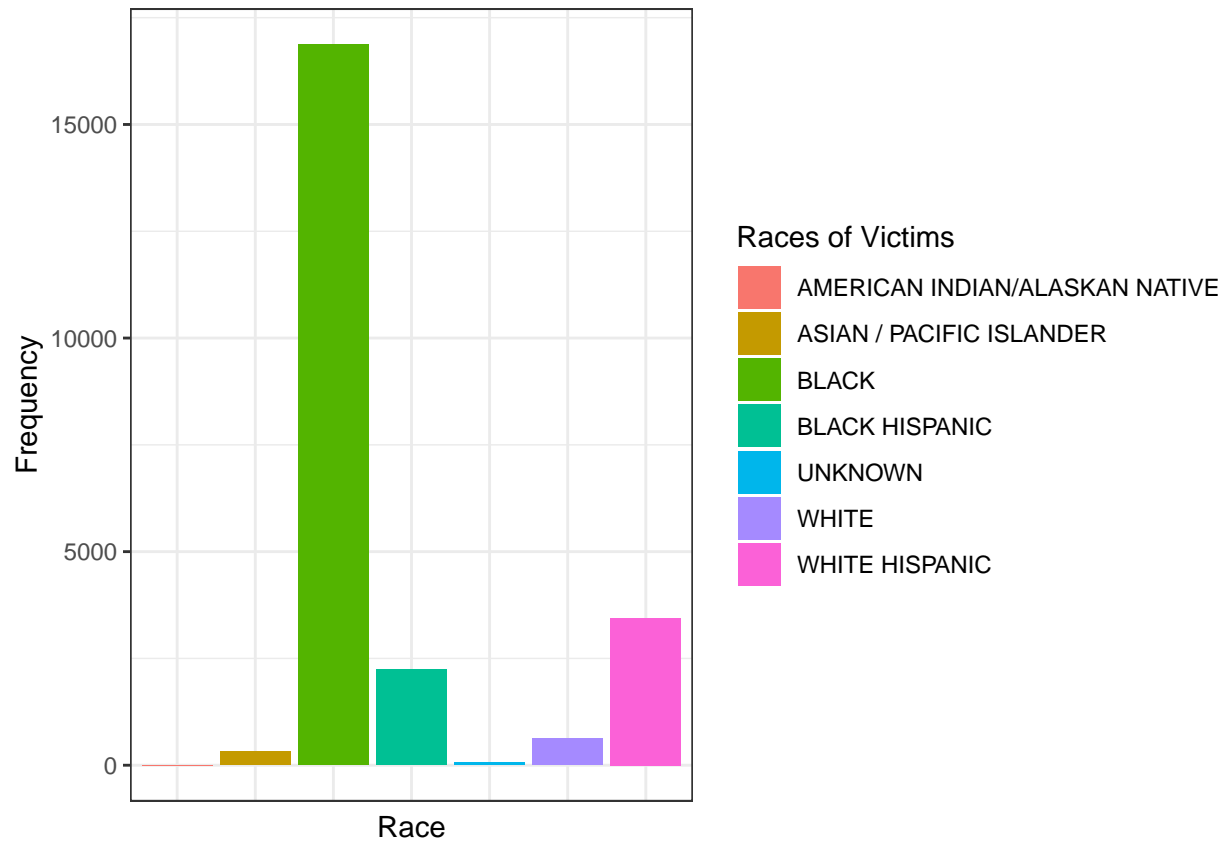
```r
library(ggplot2)
library(ggpubr)

ggplot(df, aes(x=VIC_AGE_GROUP, fill = VIC_AGE_GROUP))+
  geom_bar(stat = "count")+
  theme_bw()+
  xlab("Age Group")+
  ylab("Frequency")+
  guides(fill=guide_legend(title="Age Groups of Victims"))
```
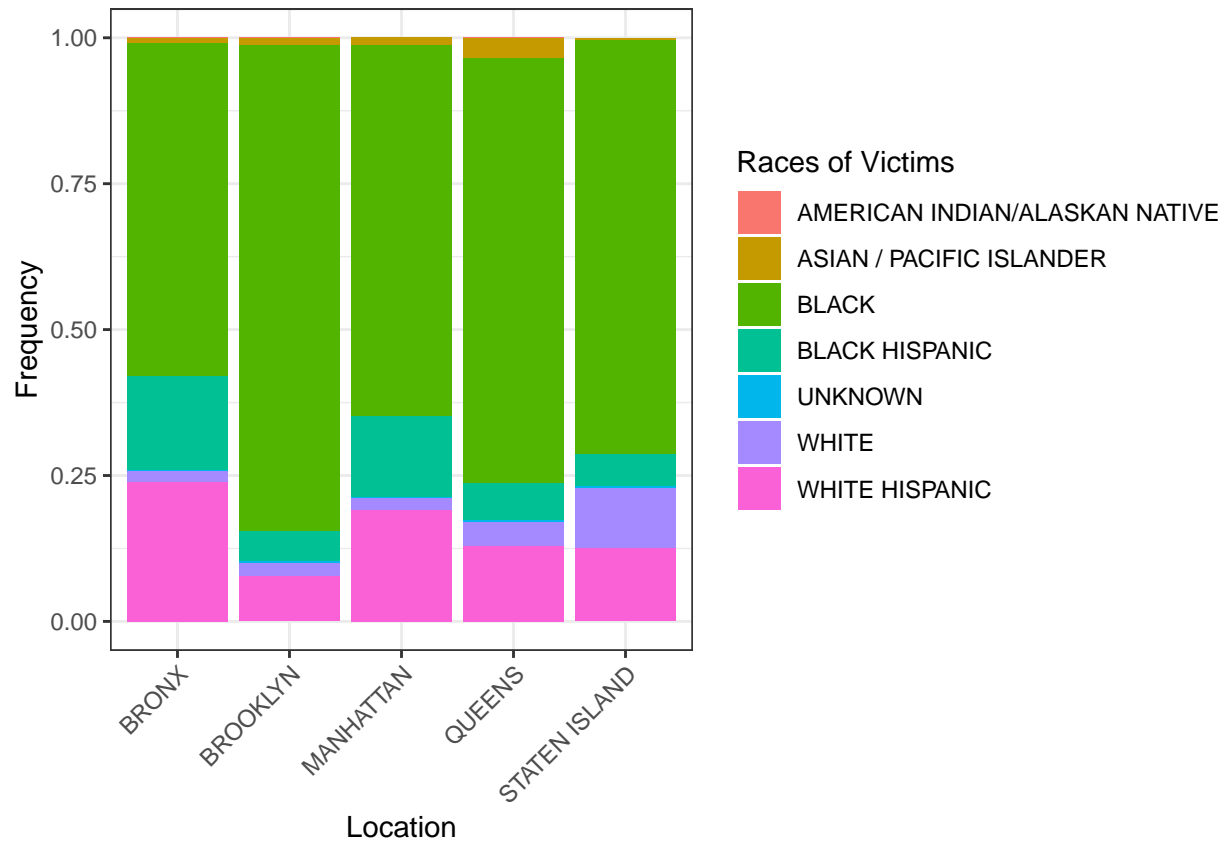
```
ggplot(df, aes(x=VIC_SEX, fill = VIC_SEX))+
  geom_bar(stat = "count")+
  theme_bw()+
  xlab("Genders")+
  ylab("Frequency")+
  guides(fill=guide_legend(title="Genders of Victims"))
```

```
ggplot(df, aes(x=VIC_RACE, fill = VIC_RACE))+
  geom_bar(stat = "count")+
  theme_bw()+
  xlab("Race")+
  ylab("Frequency")+
  theme(axis.ticks.x = element_blank(),
        axis.text.x = element_blank())+
  guides(fill=guide_legend(title="Races of Victims"))
```

```
ggplot(df, aes(x=BORO, fill = VIC_RACE))+
  geom_bar(position = "fill")+
  theme_bw()+
  xlab("Location")+
  ylab("Frequency")+
  guides(fill=guide_legend(title="Races of Victims"))+theme(axis.text.x = element_text(angle = 45, hjus
```

We can clearly find that more black victims are identified in Brooklyn and Queen regions compared to others. Interestingly, male is the most vulnerable gender and so is 25-44 age group.

There may exist bias when it comes to occurring time for the shooting. It is very difficult for people to still remember it in danger. Moreover, unknown information will also have effects on final conclusions.

Packages

```
## R version 4.1.1 (2021-08-10)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19042)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=Chinese (Simplified)_China.936
## [2] LC_CTYPE=Chinese (Simplified)_China.936
## [3] LC_MONETARY=Chinese (Simplified)_China.936
## [4] LC_NUMERIC=C
## [5] LC_TIME=Chinese (Simplified)_China.936
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] ggpubr_0.4.0  ggplot2_3.3.5
```

```
## 
## loaded via a namespace (and not attached):
##  [1] highr_0.9        pillar_1.7.0     compiler_4.1.1   tools_4.1.1
##  [5] digest_0.6.29    evaluate_0.14    lifecycle_1.0.1  tibble_3.1.6
##  [9] gtable_0.3.0     pkgconfig_2.0.3  rlang_0.4.12     DBI_1.1.2
## [13] yaml_2.2.2       xfun_0.29        fastmap_1.1.0    withr_2.4.3
## [17] stringr_1.4.0    dplyr_1.0.7      knitr_1.37       generics_0.1.2
## [21] vctrs_0.3.8      grid_4.1.1       tidyselect_1.1.1 glue_1.6.1
## [25] R6_2.5.1         rstatix_0.7.0    fansi_1.0.2      rmarkdown_2.11
## [29] carData_3.0-5    farver_2.1.0     car_3.0-12       tidyr_1.1.4
## [33] purrr_0.3.4      magrittr_2.0.2   backports_1.4.1  scales_1.1.1
## [37] ellipsis_0.3.2   htmltools_0.5.2  abind_1.4-5      assertthat_0.2.1
## [41] colorspace_2.0-2 ggsignif_0.6.3   labeling_0.4.2   utf8_1.2.2
## [45] stringi_1.7.6    munsell_0.5.0    broom_0.7.12     crayon_1.5.0
```