NYC Shooting Incident Report Analysis

In this report, we'll being analyzing the NYC Shooting Incidence data. We'll begin with tidying up and transforming our data, then visualizing it and doing some analysis, and finally we'll discuss potential biases from the analysis and summarize our findings. The question we will address in this analysis is: what can we infer about the relationship between the number of incidents and the time and place?

Project 1: Use R Markdown to create document

Load the packages needed for the analysis

```r
##We will be using the tidyverse package for this analysis

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   2.1.1     v forcats 0.5.1

## Warning: package 'tibble' was built under R version 4.1.2

## Warning: package 'tidyr' was built under R version 4.1.2

## Warning: package 'readr' was built under R version 4.1.2

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 4.1.2

##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

Project 2: Tidy and Transform your data

We will start by reading the the public data and substituting any blank or missing values in the datset with na's.

For the values with NA, we need to consider the unknown data and not omit it as because we don't know whether the data points are important later in the analysis

```
dat<-read.csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD",
              na.strings = c(""," ","na","NA"))
head(dat)
```

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME     BORO PRECINCT JURISDICTION_CODE
## 1     24050482 08/27/2006   05:35:00    BRONX       52                 0
## 2     77673979 03/11/2011   12:03:00   QUEENS      106                 0
## 3    203350417 10/06/2019   01:09:00 BROOKLYN       77                 0
## 4     80584527 09/04/2011   03:35:00    BRONX       40                 0
## 5     90843766 05/27/2013   21:16:00   QUEENS      100                 0
## 6     92393427 09/01/2013   04:17:00 BROOKLYN       67                 0
##   LOCATION_DESC STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX PERP_RACE
## 1          <NA>                    true           <NA>     <NA>      <NA>
## 2          <NA>                   false           <NA>     <NA>      <NA>
## 3          <NA>                   false           <NA>     <NA>      <NA>
## 4          <NA>                   false           <NA>     <NA>      <NA>
## 5          <NA>                   false           <NA>     <NA>      <NA>
## 6          <NA>                   false           <NA>     <NA>      <NA>
##   VIC_AGE_GROUP VIC_SEX       VIC_RACE X_COORD_CD Y_COORD_CD Latitude Longitude
## 1         25-44       F BLACK HISPANIC    1017542   255918.9 40.86906 -73.87963
## 2           65+       M          WHITE    1027543   186095.0 40.67737 -73.84392
## 3         18-24       F          BLACK     995325   185155.0 40.67489 -73.96008
## 4           <18       M          BLACK    1007453   233952.0 40.80880 -73.91618
## 5         18-24       M          BLACK    1041267   157133.5 40.59780 -73.79469
## 6           <18       M          BLACK    1001694   170112.9 40.63359 -73.93715
##                                          Lon_Lat
## 1  POINT (-73.87963173099996 40.86905819000003)
## 2 POINT (-73.84392019199998 40.677366895000034)
## 3 POINT (-73.96007501899999 40.674885741000026)
## 4  POINT (-73.91618413199996 40.80879780500004)
## 5 POINT (-73.79468553799995 40.597796249000055)
## 6  POINT (-73.93715330699996 40.63358818100005)
```

```
summary(dat)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :  9953245   Length:23585       Length:23585       Length:23585
##  1st Qu.: 55322804   Class :character   Class :character   Class :character
##  Median : 83435362   Mode  :character   Mode  :character   Mode  :character
##  Mean   :102280741
##  3rd Qu.:150911774
##  Max.   :230611229
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.000     Length:23585       Length:23585
```

```
##  1st Qu.: 44.00   1st Qu.:0.000     Class :character   Class :character
##  Median : 69.00   Median :0.000     Mode  :character   Mode  :character
##  Mean   : 66.21   Mean   :0.333
##  3rd Qu.: 81.00   3rd Qu.:0.000
##  Max.   :123.00   Max.   :2.000
##                   NA's   :2
##  PERP_AGE_GROUP      PERP_SEX          PERP_RACE         VIC_AGE_GROUP
##  Length:23585       Length:23585       Length:23585       Length:23585
##  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:23585       Length:23585       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.: 999925   1st Qu.:182539
##  Mode  :character   Mode  :character   Median :1007654   Median :193470
##                                        Mean   :1009379   Mean   :207300
##                                        3rd Qu.:1016782   3rd Qu.:239163
##                                        Max.   :1066815   Max.   :271128
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:23585
##  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
##
```

For the purooses of this analysis, all values of "NA" will be labeled as "UNKNOWN" and will later be omitted. This will help us focus on data that are known and make it simplier to draw conclusions

```
dat[is.na(dat)]<- "UNKNOWN"
```

Create new dataframe and select only important columns for analysis. Convert them to appropriate data types

```
##colnames(dat)
dat2<-dat %>%
  select(-c(INCIDENT_KEY,OCCUR_DATE,PRECINCT,
            JURISDICTION_CODE,STATISTICAL_MURDER_FLAG,
            X_COORD_CD,Y_COORD_CD,
            Latitude,Longitude,Lon_Lat)) %>%
  mutate(time=as.factor(hms(OCCUR_TIME)@hour))

colnamesvec<- colnames(dat2)
colnamesvec

dat3<- lapply(select_if(dat2[colnamesvec],is.character), factor)
datmer<- merge(dat3,dat2)
```

```r
##check data structure
str(datmer)
```

Let's create a long data file so that we can view all counts of each group. Create a function to summarize each column and then recombine to form a long data format. Then view the long data frame

```r
funsum<-function(dat,newcolval){
  df<-as.data.frame(summary(dat))
new_df<-cbind(variable=row.names(df),df)
new_df<-rename(new_df, count="summary(dat)")
row.names(new_df)<-NULL
new_df<-cbind(group=newcolval,new_df)
return(new_df)}

datlong<-rbind(funsum(datmer$BORO,"boro"),
funsum(datmer$time,"time"),
funsum(datmer$LOCATION_DESC,"location"),
funsum(datmer$PERP_AGE_GROUP,"perp age"),
funsum(datmer$PERP_SEX,"perp sex"),
funsum(datmer$PERP_RACE,"perp race"),
funsum(datmer$VIC_AGE_GROUP,"vic age"),
funsum(datmer$VIC_SEX,"vic sex"),
funsum(datmer$VIC_RACE,"vic race"))
datlong$group<-as.factor(datlong$group)
datlong$variable<-as.factor(datlong$variable)

tail(datlong,20)
```

Now filter out the "UNKNOWN" values from the rows and check to see that there are no more rows with missing values

```r
datlong<-datlong %>% filter(variable!="UNKNOWN") %>% filter(variable!="U")
tail(datlong,20)
```

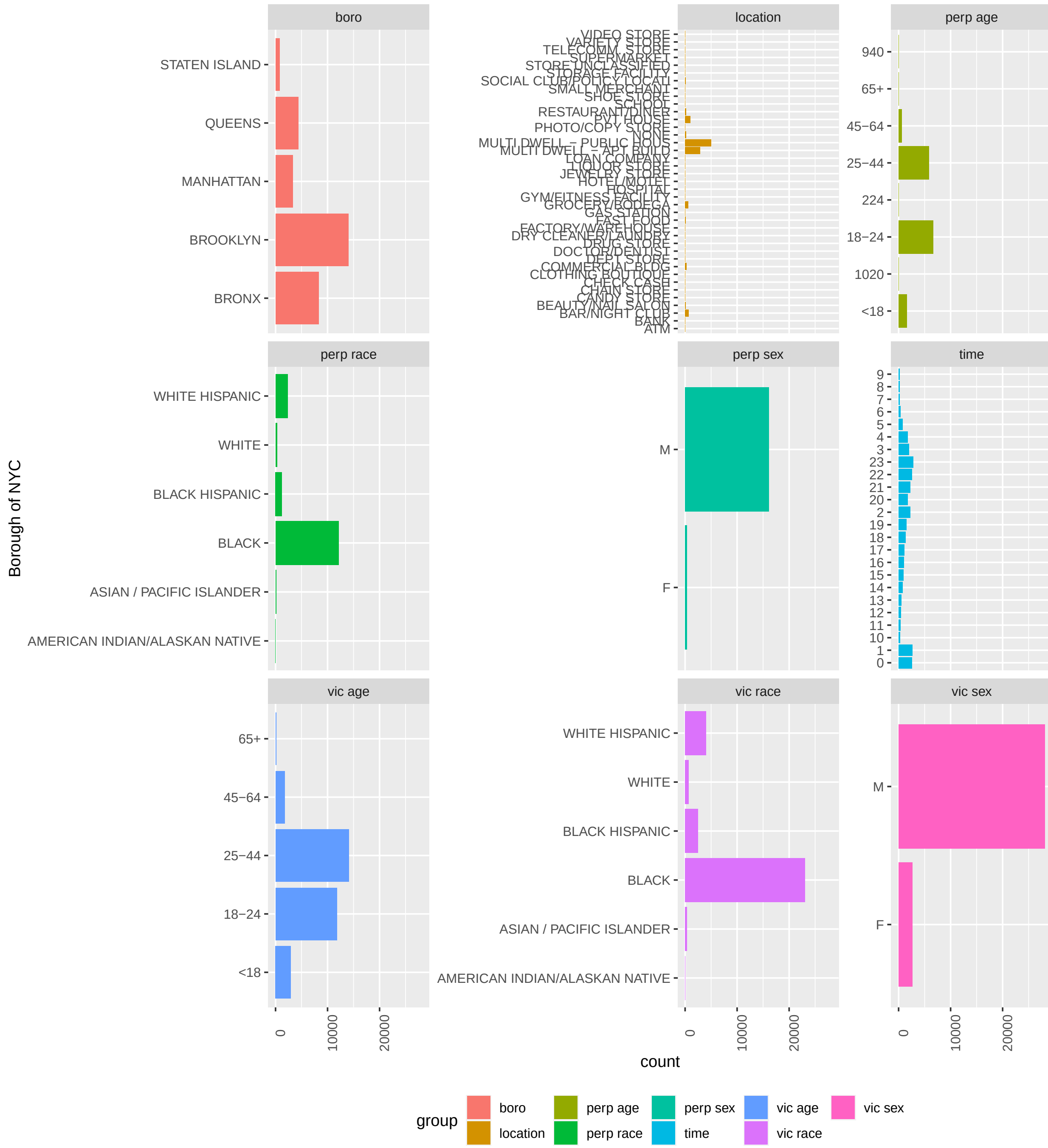
```
##        group                       variable count
## 78  perp sex                              M 16102
## 79 perp race AMERICAN INDIAN/ALASKAN NATIVE     2
## 80 perp race       ASIAN / PACIFIC ISLANDER   152
## 81 perp race                          BLACK 12125
## 82 perp race                 BLACK HISPANIC  1228
## 83 perp race                          WHITE   291
## 84 perp race                 WHITE HISPANIC  2324
## 85   vic age                            <18  2947
## 86   vic age                          18-24 11785
## 87   vic age                          25-44 14089
## 88   vic age                          45-64  1763
## 89   vic age                            65+   162
## 90   vic sex                              F  2666
## 91   vic sex                              M 28138
## 92  vic race AMERICAN INDIAN/ALASKAN NATIVE     9
## 93  vic race       ASIAN / PACIFIC ISLANDER   359
## 94  vic race                          BLACK 23079
```

```
## 95   vic race                  BLACK HISPANIC  2525
## 96   vic race                           WHITE   692
## 97   vic race                  WHITE HISPANIC  4080
```

Project 3: Add Visualizations and Analysis

Now let's visualize the dataset to get a better understanding of what's in the data

```r
ggplot(data = datlong, aes(x=as.factor(variable),y=count, fill=group))+
  ##geom_bar(stat="identity")+
  geom_bar(stat="identity")+
  xlab("Borough of NYC")+
  coord_flip()+
  facet_wrap(~group, scales = "free_y")+
  theme(legend.position="bottom",
        axis.text.x=element_text(angle = 90))
```

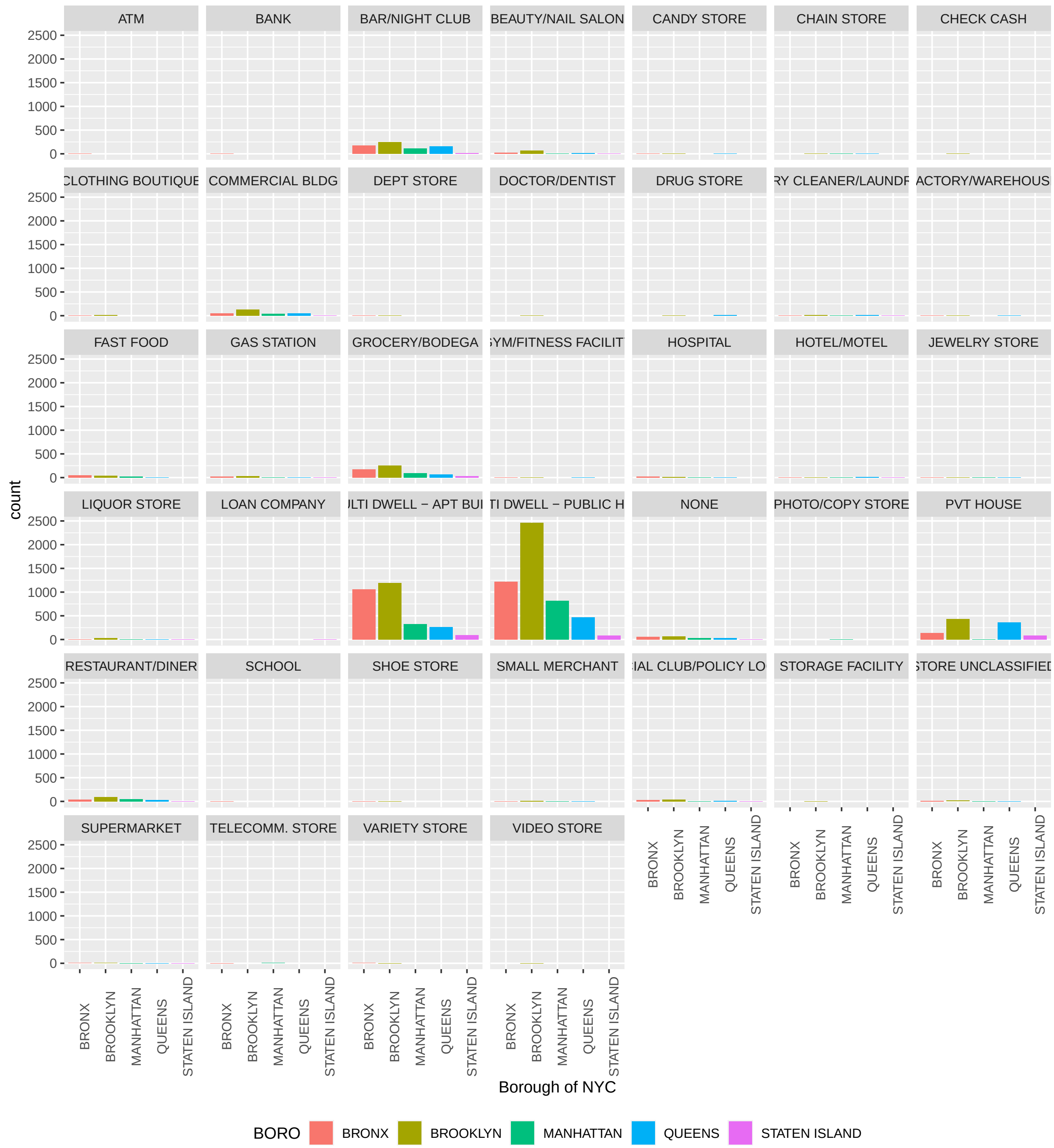


*We see some interesting results. For this analysis we will focus on the location (Bororugh) in which the incidents occured and the time***

Where do we see the most incidents?

```
##Where do we see the most incidents?
ggplot(data = datmer %>% filter(LOCATION_DESC!="UNKNOWN") %>% filter(LOCATION_DESC!="U"),
       aes(x=factor(BORO), fill= BORO))+
  geom_bar()+
  facet_wrap(~LOCATION_DESC)+
  xlab("Borough of NYC")+
```

```r
  theme(legend.position="bottom",
        axis.text.x=element_text(angle = 90))
```



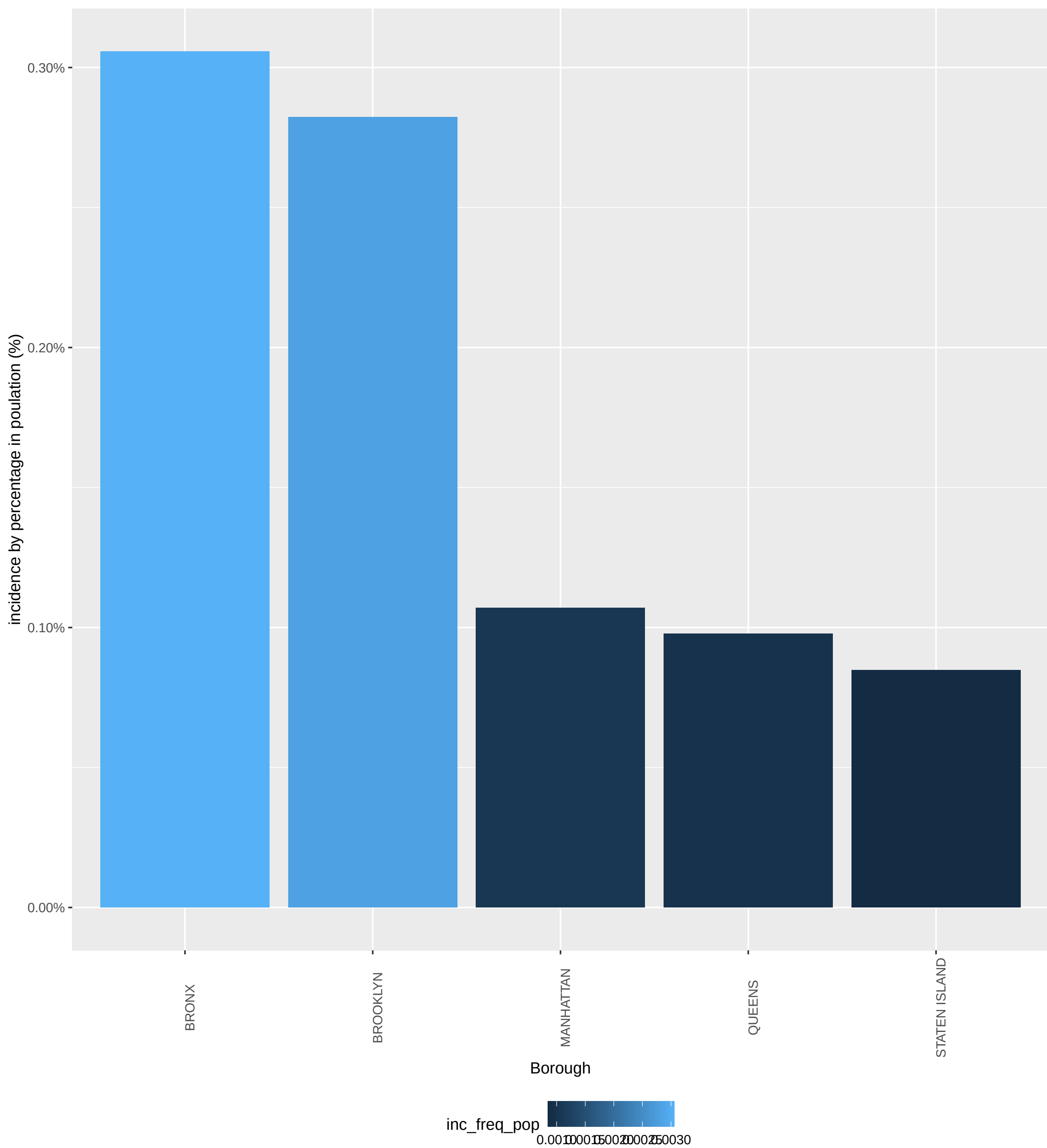
It looks like multi-dwelling groups have the most reported incidents

Additionally, the Borough Brookyln also has a higher count of reported incidents. Let's include the population data to see if the incident rate as a function of population is different

```r
loc_dat<-as.data.frame(summary(datmer$BORO))
new_dat<-cbind(boro=row.names(loc_dat),
```

```r
               total=loc_dat[1])
new_dat<-rename(new_dat, count="summary(datmer$BORO)")
row.names(new_dat)<-NULL
new_dat<-data.frame(new_dat,
                          population= c(2717758,
                                        4970026,
                                        3123068,
                                        4460101,
                                        912458)) %>% mutate(inc_freq_pop= count/population)

ggplot(data = new_dat, aes(x=factor(boro), y=inc_freq_pop, fill= inc_freq_pop))+
  geom_bar(stat="identity")+
  ##facet_wrap(~LOCATION_DESC)+
  theme(legend.position="bottom",
        axis.text.x=element_text(angle = 90))+
  scale_y_continuous(labels = scales::percent_format())+
  labs(x="Borough", y="incidence by percentage in poulation (%)")
```

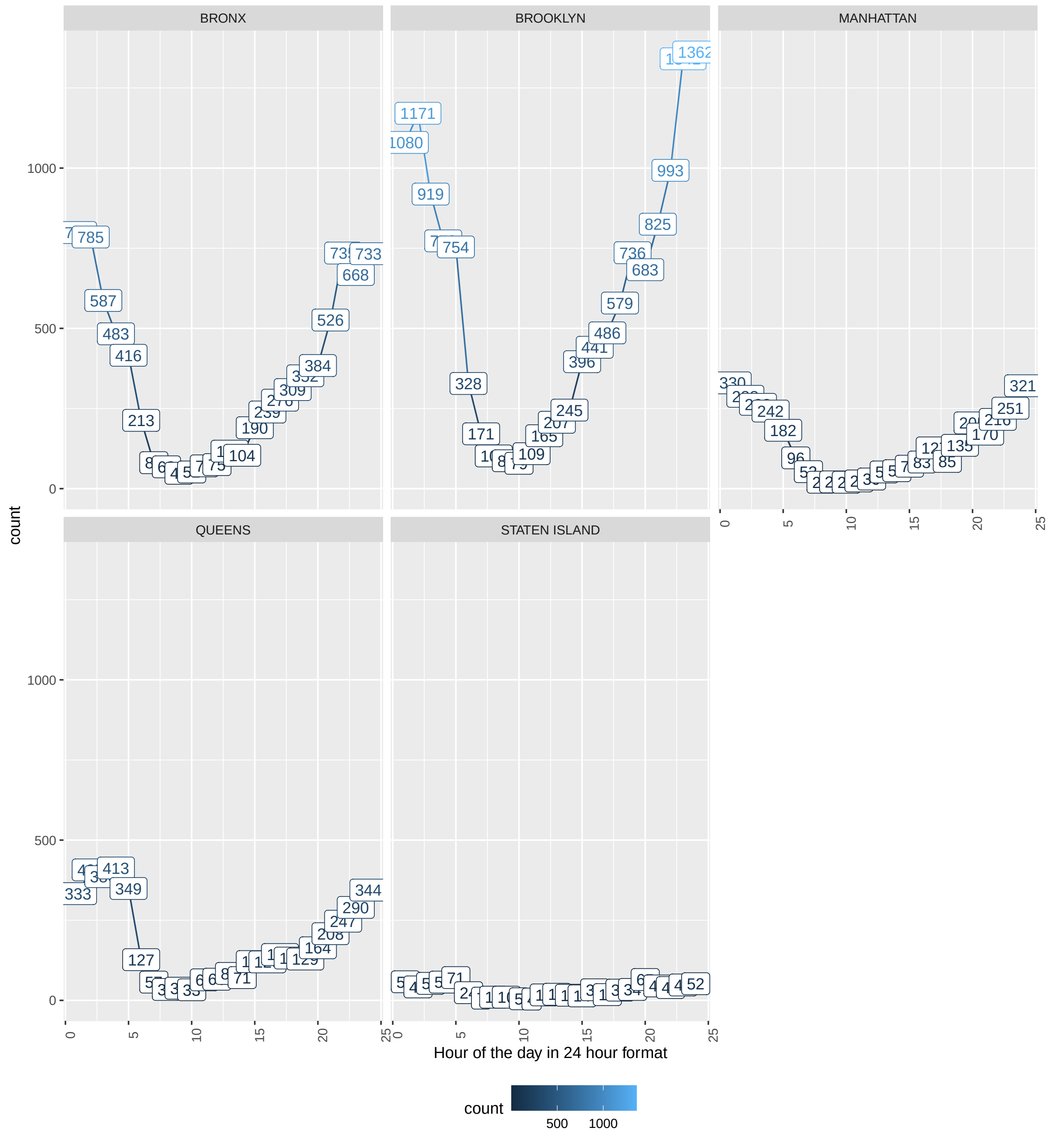


Interesting. Now let's check when and where these incidents occured

```
time_dat<-rename(count(datmer,time,BORO),count=n)

##plot number of incidence reported tp the count
ggplot(data = time_dat, aes(x=as.numeric(time),y=count, color=count))+
  geom_point()+
  geom_line()+
  facet_wrap(~BORO)+
  xlab("Hour of the day in 24 hour format")+
```

```
theme(legend.position="bottom",
      axis.text.x=element_text(angle = 90))+
geom_label(aes(label=count))
```



We can see that the highest reported shooting incidents are around mignight (values= 0,1,23,24) and they occur most frequently in Bronx and Brookyn. Could it be that these areas are very dangerous around those hours?

Now let's do some analysis and predict the incidents by Borough and Time

```r
mod<-lm(count~BORO, data=time_dat)
summary(mod)$adj.r.squared
```

```
## [1] 0.4016487
```

```r
mod<-lm(count~time, data=time_dat)
summary(mod)$adj.r.squared
```

```
## [1] 0.1999399
```

```r
mod<-lm(count~BORO+time, data=time_dat)
summary(mod)$adj.r.squared
```

```
## [1] 0.7106939
```

```r
summary(mod)
```

```
##
## Call:
## lm(formula = count ~ BORO + time, data = time_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -286.66  -77.03    4.03   77.87  493.54
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          609.48      77.36   7.878 6.45e-12 ***
## BOROBROOKLYN         238.38      46.23   5.156 1.44e-06 ***
## BOROMANHATTAN       -207.12      46.23  -4.480 2.14e-05 ***
## BOROQUEENS          -164.62      46.23  -3.561 0.000588 ***
## BOROSTATEN ISLAND  -314.04      46.23  -6.793 1.07e-09 ***
## time1                18.60     101.29   0.184 0.854710
## time2               -78.40     101.29  -0.774 0.440915
## time3              -126.40     101.29  -1.248 0.215242
## time4              -165.60     101.29  -1.635 0.105492
## time5              -362.40     101.29  -3.578 0.000555 ***
## time6              -446.20     101.29  -4.405 2.85e-05 ***
## time7              -473.20     101.29  -4.672 1.02e-05 ***
## time8              -479.20     101.29  -4.731 8.05e-06 ***
## time9              -482.20     101.29  -4.760 7.16e-06 ***
## time10             -465.60     101.29  -4.597 1.36e-05 ***
## time11             -449.40     101.29  -4.437 2.53e-05 ***
## time12             -424.20     101.29  -4.188 6.45e-05 ***
## time13             -421.60     101.29  -4.162 7.09e-05 ***
## time14             -361.80     101.29  -3.572 0.000566 ***
## time15             -336.80     101.29  -3.325 0.001271 **
## time16             -310.00     101.29  -3.060 0.002896 **
## time17             -292.60     101.29  -2.889 0.004823 **
## time18             -242.80     101.29  -2.397 0.018550 *
## time19             -219.80     101.29  -2.170 0.032586 *
```

```
## time20             -165.20     101.29  -1.631 0.106324
## time21              -73.80     101.29  -0.729 0.468105
## time22               -0.40     101.29  -0.004 0.996858
## time23               42.40     101.29   0.419 0.676490
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 160.2 on 92 degrees of freedom
## Multiple R-squared:  0.7763, Adjusted R-squared:  0.7107
## F-statistic: 11.83 on 27 and 92 DF,  p-value: < 2.2e-16
```

*Borough and Time are very good predictors of count and fit the model better together than as indivudal predictors as can be seen from the adjusted r-squared values. The r-squared value for for the multiple regression model is 0.7107 which is very good***

Let's do some more analysis on the count values with repect to time and the predicted counts.How does the predicted values compare with the reported values? What's the correlation statistic?

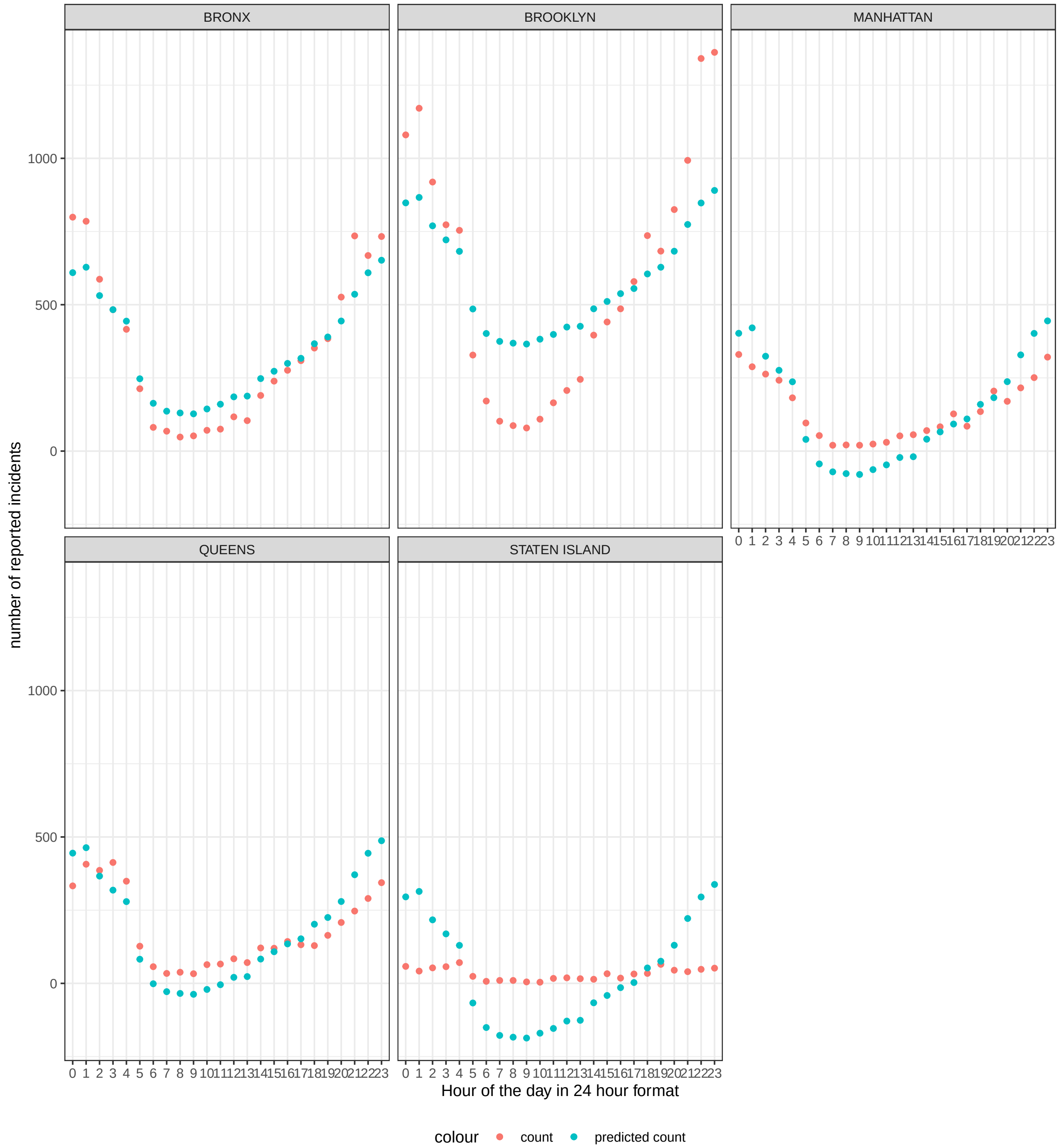
```
##check for statical differences between time
time_dat2<-time_dat%>% mutate(pred_vals=predict(mod))
##View(head(time_dat2))
correlationtest<-cor.test(time_dat2$count,time_dat2$pred_vals)
correlationtest
```

```
##
##  Pearson's product-moment correlation
##
## data:  time_dat2$count and time_dat2$pred_vals
## t = 20.238, df = 118, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8334897 0.9157208
## sample estimates:
##       cor
## 0.8810986
```

Very nice. The p-value is less than 2.2e-16 and The correlation is 0.88, indicating significantly postive correlation

Now let's see how the predicted values compare with the reported values in a chart

```
##Now lets visualize
time_dat2 %>% ggplot()+
  geom_point(aes(x=time,
                 y=count,
             color="count"))+
  geom_point(aes(x=time,
                 y=pred_vals,
             color="predicted count"))+
  ##coord_flip()+
  theme_bw()+
  facet_wrap(~BORO)+
  ylab("number of reported incidents")+
  xlab("Hour of the day in 24 hour format")+
  theme(legend.position="bottom")
```




```
##axis.text.x=element_text(angle = 90))
```

A good fit!

Project Step 4: Add Bias Identification

Potential biases from the dataset includes how the data may be collected, the quality of the data collection and the frequency in which the data is collected in each part of NYC. To mitigate potential biases for myself, I explored the data of all relevant variables in the dataset. To avoid ethical issues that could arise with the *reporting of the data, I avoided exploring indepthly race, age, or sex. Doing the analyis based on time and location could be useful for the areas in NYC, because they can use the information and take action to try and reduce incidents without targeting specific groups of people. From the analyis, we see that most incidents occured around midnight. It may not be feasible for the Boroughs to enforce a curfew after 11pm due to the massive populations in each area but people could be made aware of of the higher than usual incident rate at night so that people can avoid being there. We also observe that the incidents occur at multi-dwelling units usch as apartments buildings. As a resident, it would be hard to avoid being near the incidents at the time but it is good to know when to stay indoors to avoid becoming a victim. We observe in our model that both Borough and time predict the incidents counts very well. The adjusted r-squared value is highest at above 0.7 when both factors are incorporated in the model. Nonetheless we have to be aware that these relationships to not imply causation and there may be other important factors that are not captured in the dataset.*

```r
sessionInfo()
```

```
## R version 4.1.0 (2021-05-18)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19043)
##
## 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] lubridate_1.8.0 forcats_0.5.1   stringr_1.4.0   dplyr_1.0.7
##  [5] purrr_0.3.4     readr_2.1.1     tidyr_1.1.4     tibble_3.1.6
##  [9] ggplot2_3.3.5   tidyverse_1.3.1
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.1.1 xfun_0.29        haven_2.4.3      colorspace_2.0-2
##  [5] vctrs_0.3.8      generics_0.1.1   htmltools_0.5.2  yaml_2.2.2
##  [9] utf8_1.2.2       rlang_0.4.12     pillar_1.6.5     glue_1.6.0
## [13] withr_2.4.3      DBI_1.1.2        dbplyr_2.1.1     modelr_0.1.8
## [17] readxl_1.3.1     lifecycle_1.0.1  munsell_0.5.0    gtable_0.3.0
## [21] cellranger_1.1.0 rvest_1.0.2      evaluate_0.14    labeling_0.4.2
## [25] knitr_1.37       tzdb_0.2.0       fastmap_1.1.0    fansi_1.0.2
## [29] highr_0.9        broom_0.7.12     Rcpp_1.0.7       scales_1.1.1
## [33] backports_1.4.1  jsonlite_1.7.3   farver_2.1.0     fs_1.5.2
## [37] hms_1.1.1        digest_0.6.29    stringi_1.7.6    grid_4.1.0
## [41] cli_3.1.1        tools_4.1.0      magrittr_2.0.1   crayon_1.4.2
## [45] pkgconfig_2.0.3  ellipsis_0.3.2   xml2_1.3.3       reprex_2.0.1
## [49] rstudioapi_0.13  assertthat_0.2.1 rmarkdown_2.11   httr_1.4.2
## [53] R6_2.5.1         compiler_4.1.0
```