Cobb County, GA Seasonal Analysis

Sarah Tillman

```
# function to check if a character variable is actually a number
# (in the context of this dataset)
check_if_number <- function(str){
  if(is.na(str) | str == '-'){
    return(FALSE)
  }
  else{
    return(TRUE)
  }
}

# function to reformat character variable to numeric variable when , or . was
# entered to mark the thousands place
reformat_number <- function(str){
  if(str_detect(str, "[,.]")){
    str <- str_c(substr(str, 1, 1), substr(str, 3, 5), sep='')
  }
  return(str)
}
```

Data Import

All data was forked and cloned from the GitHub repo except the Jun-Oct data zip file which was downloaded separately.

```
raw_oct_june_data <-
  read_csv(unzip("../Race-and-Ethnicity-Data-1.1/Health_equity_data.csv.zip",
                                      "Health_equity_data.csv"))
raw_nov_data <- read_csv(unzip("../Health_equity_data.zip",
                               "Health_equity_data.csv"), col_names = F)
raw_dec_data <- read_csv("../pub_equity_thru_122020.csv")
raw_jan_data <- read_csv("../pub_equity_thru_012021.csv")
raw_feb_data <- read_csv("../pub_equity_thru_022021.csv")
```

Data was then filtered so it only included Cobb County.

```
# renaming columns
colnames(raw_oct_june_data) <- c("county", "date", "group", "cases")
colnames(raw_nov_data) <- c("county", "date", "group", "cases")
colnames(raw_dec_data) <- c("county", "date", "group", "cases")
colnames(raw_jan_data) <- c("county", "date", "group", "cases")
colnames(raw_feb_data) <- c("county", "date", "group", "cases")
```

```
# filtering monthy data to only include Cobb
cobb_jun_oct_long <- raw_oct_june_data[ which(raw_oct_june_data$county=="Cobb County, Georgia"), ]
cobb_nov_long <- raw_nov_data[ which(raw_nov_data$county=="Cobb County, Georgia"), ]
cobb_dec_long <- raw_dec_data[ which(raw_dec_data$county=="Cobb County, Georgia"), ]
cobb_jan_long <- raw_jan_data[ which(raw_jan_data$county=="Cobb County, Georgia"), ]
cobb_feb_long <- raw_feb_data[ which(raw_feb_data$county=="Cobb County, Georgia"), ]
```

Cobb County did not report any data for 2 or more races, and the data for cumulative cases was inaccurate, so both of these groups were dropped from the dataset. The "county" column was also dropped from the data frame since it was no longer necessary.

```
cobb_jun_oct_long <- cobb_jun_oct_long %>%
  select(-c("county")) %>%
  filter(group != "2+_races" & group != "cumulative_cases")
```

The data was then converted from long form to short form to make identifying data entry errors & formatting issues easier. The Jun-Oct data was converted first since it contained several errors and data was reported inconsistently.

```
# convert char to numeric variables
for(n in 1:nrow(cobb_jun_oct_long)){
  val <- cobb_jun_oct_long$cases[n]
  if(check_if_number(val)){
    cobb_jun_oct_long$cases[n] <- reformat_number(val)
  }
  else{
    cobb_jun_oct_long$cases[n] <- NA
  }
}

# list of dates in data set
date <- unique(cobb_jun_oct_long$date)
# initializing new data frame for short form
cobb_jun_oct_short <- data.frame(date)

# list of all racial/ethnic groups reported
race_eth_categories <- unique(cobb_jun_oct_long$group)

# add each group as its own column w each row representing a single date
for(cat in race_eth_categories){
  single_group <- cobb_jun_oct_long %>%
    filter(group == cat) %>%
    select(-c(group))
  colnames(single_group)[2] <- cat

  cobb_jun_oct_short <- merge(cobb_jun_oct_short, single_group, by="date")
}
```

Due to inconsistent reporting of the unknown and other race categories and the discontinuation of data on American Indian/Alaska Native and Native Hawaiian/Other Pacific Islander cases, these 4 groups were added together into one category in the table.

```
other_race <- rep(0, nrow(cobb_jun_oct_short))
for(d in 1:nrow(cobb_jun_oct_short)){

  # add american_indian_alaska_native + native_hawaiian_pacific_islander +
  # other + unknown into "other_race"

  aian <- cobb_jun_oct_short$american_indian_alaska_native[d]
  if(check_if_number(aian)){
    other_race[d] <- other_race[d] + as.numeric(aian)
  }

  nhpi <- cobb_jun_oct_short$native_hawaiian_pacific_islander[d]
  if(check_if_number(nhpi)){
    other_race[d] <- other_race[d] + as.numeric(nhpi)
  }

  other <- cobb_jun_oct_short$other[d]
  if(check_if_number(other)){
    other_race[d] <- other_race[d] + as.numeric(other)
  }

  unknown <- cobb_jun_oct_short$unknown[d]
  if(check_if_number(unknown)){
    other_race[d] <- other_race[d] + as.numeric(unknown)
  }
}
```

After adding those 4 groups together, the original columns were dropped from the table. Additionally, dates which contained any NA values were dropped from the dataset.

```
cobb_jun_oct_short <- cobb_jun_oct_short %>%
  select(-c("native_hawaiian_pacific_islander", "american_indian_alaska_native",
            "other", "unknown")) %>%
  cbind(other_race) %>%
  drop_na()
```

The data from Nov-Feb was formatted more consistently and only required non-reported racial/ethnic groups to be dropped (as well as the "county" column).

```
cobb_nov_long <- cobb_nov_long %>%
  select(-c(county)) %>%
  filter(group != 'native_hawaiian_pacific_islander' &
           group != "american_indian_alaska_native" &
           group != "2+_races" & group != "cumulative_cases")

# converting remaining counts to numeric
cobb_nov_long$cases <- as.numeric(cobb_nov_long$cases)

# Dec, Feb & Jan data no longer had rows for Native Hawaiian, American Indian &
# 2+ races

cobb_dec_long <- cobb_dec_long %>%
  select(-c(county)) %>%
```

```
  filter(group != "cumulative_cases")

cobb_jan_long <- cobb_jan_long %>%
  select(-c(county)) %>%
  filter(group != "cumulative_cases")

cobb_feb_long <- cobb_feb_long %>%
  select(-c(county)) %>%
  filter(group != "cumulative_cases")
```

The data Nov-Feb was then combined into a single data frame and converted from long form to short form.

```
# combining Nov-Feb data frames
cobb_nov_feb_long <- rbind(cobb_nov_long, cobb_dec_long, cobb_jan_long,
                           cobb_feb_long)

# list of dates in data frame
date <- unique(cobb_nov_feb_long$date)
# initializing new data frame for short form
cobb_nov_feb_short <- data.frame(date)

# list of racial/ethnic groups in dataset
race_eth_categories <- unique(cobb_nov_feb_long$group)

# add each group as its own column w each row representing a single date
for(cat in race_eth_categories){
  single_group <-  cobb_nov_feb_long %>%
    filter(group == cat) %>%
    select(-c(group))
  colnames(single_group)[2] <- cat

  cobb_nov_feb_short <- merge(cobb_nov_feb_short, single_group, by="date")
}
```

To keep formatting consistent with the Jun-Oct data, the other & unknown race categories in the Nov-Feb dataset were combined into a single column.

```
# create other_race column
cobb_nov_feb_short$other_race <- cobb_nov_feb_short$other + cobb_nov_feb_short$unknown

# drop other and unknown columns
cobb_nov_feb_short <- cobb_nov_feb_short %>%
  select(-c(other, unknown))
```

The Jun-Oct & Nov-Feb data frames were then combined. After combining the datasets, there were 21 dates for which data was not reported. Most of these dates with the exception of Thanksgiving were in the Jun-Oct dataset. The first date was which data was reported was 6/15/2020.

```
# convert columns into numeric variables
cobb_jun_oct_short$asian <- as.numeric(cobb_jun_oct_short$asian)
cobb_jun_oct_short$black_african_american <- as.numeric(cobb_jun_oct_short$black_african_american)
cobb_jun_oct_short$`hispanic_(all_races)` <- as.numeric(cobb_jun_oct_short$`hispanic_(all_races)`)
```

```
cobb_jun_oct_short$`non-hispanic` <- as.numeric(cobb_jun_oct_short$`non-hispanic`)
cobb_jun_oct_short$not_specified <- as.numeric(cobb_jun_oct_short$not_specified)
cobb_jun_oct_short$white <- as.numeric(cobb_jun_oct_short$white)

# combine both short df into single df
cobb_all_months_short <- rbind(cobb_jun_oct_short, cobb_nov_feb_short)
```

To make analysis more meaningful and visualizations more informative, major trend breaks were either removed from the data or corrected when possible. There were several dates where it looked like the numbers for the Hispanic count and the Non-Hispanic count was swapped, so these cases were fixed. There was also an error in the other_case column on 7/31 which resulted in a significant decrease in cases, likely due to a number being left out during data entry. The value from the day before was used to replace this value.

For the trend breaks not due to data entry mistakes, most were simply removed from the data. Most of the major decreases in cases were immediately preceded by a large jump in the number of reported cases, so usually the date of the large increase was removed, allowing the case number to remain relatively constant.

There were several trend breaks in December which were slightly more complicated that were not removed at this time.

```
cobb_all_months_short <- cobb_all_months_short %>%
  filter(date < '2021-02-25' & date != '2020-09-12' & date != '2020-07-18'
         & date != '2020-09-18' & date != '2020-11-12')

# switching incorrect data entry

cobb_sept_nine <- cobb_all_months_short %>%
  filter(date == "2020-09-09")
cobb_all_months_short$`hispanic_(all_races)`[69] <- cobb_sept_nine$`non-hispanic`[1]
cobb_all_months_short$`non-hispanic`[69] <- cobb_sept_nine$`hispanic_(all_races)`[1]

cobb_aug_sixteen_seventeen <- cobb_all_months_short %>%
  filter(date == "2020-08-16" | date == "2020-08-17")
# fixing error on 8/16
cobb_all_months_short$`hispanic_(all_races)`[52] <- cobb_aug_sixteen_seventeen$`non-hispanic`[1]
cobb_all_months_short$`non-hispanic`[52] <- cobb_aug_sixteen_seventeen$`hispanic_(all_races)`[1]
# fixing error on 8/17
cobb_all_months_short$`hispanic_(all_races)`[53] <- cobb_aug_sixteen_seventeen$`non-hispanic`[2]
cobb_all_months_short$`non-hispanic`[53] <- cobb_aug_sixteen_seventeen$`hispanic_(all_races)`[2]

cobb_july_errors <- cobb_all_months_short %>%
  filter(date == '2020-06-29' | date == '2020-07-02' | date == '2020-07-03')
# fixing error on 6/29
cobb_all_months_short$`hispanic_(all_races)`[14] <- cobb_july_errors$`non-hispanic`[1]
cobb_all_months_short$`non-hispanic`[14] <- cobb_july_errors$`hispanic_(all_races)`[1]
# fixing error on 7/2
cobb_all_months_short$`hispanic_(all_races)`[16] <- cobb_july_errors$`non-hispanic`[2]
cobb_all_months_short$`non-hispanic`[16] <- cobb_july_errors$`hispanic_(all_races)`[2]
# fixing error on 7/3
cobb_all_months_short$`hispanic_(all_races)`[17] <- cobb_july_errors$`non-hispanic`[3]
cobb_all_months_short$`non-hispanic`[17] <- cobb_july_errors$`hispanic_(all_races)`[3]
# fixing error in other race cases on 7/31
cobb_all_months_short$other_race[36] <- cobb_all_months_short$other_race[35]
```

After the trend breaks were removed, the overall case count for all races (including the other_race category) and the overall case count for ethnicity (including the not_specified category) were added as columns in the data set. The counts for each racial/ethnic group were also converted to case rates based on the population numbers reported in the 2019 ACS.

```
# population counts
cobb_pop_all <- 751218
cobb_white_pop <- 440462
cobb_black_pop <- 207059
cobb_asian_pop <- 40271
cobb_hispanic_pop <- 97481
cobb_non_hispanic_pop <- 653737

# get cumulative case counts
cobb_all_months_short$all_cases_race <- cobb_all_months_short$asian +
  cobb_all_months_short$black_african_american + cobb_all_months_short$white +
  cobb_all_months_short$other_race

cobb_all_months_short$all_cases_eth <-
  cobb_all_months_short$`hispanic_(all_races)` +
  cobb_all_months_short$`non-hispanic` + cobb_all_months_short$not_specified

cobb_all_months_short$asian_rate <-
  cobb_all_months_short$asian / cobb_asian_pop

cobb_all_months_short$black_rate <-
  cobb_all_months_short$black_african_american / cobb_black_pop

cobb_all_months_short$white_rate <-
  cobb_all_months_short$white / cobb_white_pop

cobb_all_months_short$hispanic_rate <-
  cobb_all_months_short$`hispanic_(all_races)` / cobb_hispanic_pop

cobb_all_months_short$non_hispanic_rate <-
  cobb_all_months_short$`non-hispanic` / cobb_non_hispanic_pop

cobb_all_months_short$all_race_rate <-
  cobb_all_months_short$all_cases_race / cobb_pop_all

cobb_all_months_short$all_eth_rate <-
  cobb_all_months_short$all_cases_eth / cobb_pop_all
```

The daily change in case numbers for each group was then calculated and added to the table. The daily changes in cases were also scaled by the population size of each group.

```
daily_asian <- rep(0, nrow(cobb_all_months_short))
daily_all_race <- rep(0, nrow(cobb_all_months_short))
daily_all_eth <- rep(0, nrow(cobb_all_months_short))
daily_black_aa <- rep(0, nrow(cobb_all_months_short))
daily_hispanic <- rep(0, nrow(cobb_all_months_short))
daily_non_hispanic <- rep(0, nrow(cobb_all_months_short))
daily_white <- rep(0, nrow(cobb_all_months_short))
daily_other_race <- rep(0, nrow(cobb_all_months_short))
```

```
daily_not_specified <- rep(0, nrow(cobb_all_months_short))

for(d in 2:nrow(cobb_all_months_short)){
  daily_asian[d] <- cobb_all_months_short$asian[d] -
                    cobb_all_months_short$asian[d-1]
  daily_all_race[d] <- cobb_all_months_short$all_cases_race[d] -
                    cobb_all_months_short$all_cases_race[d-1]
  daily_all_eth[d] <- cobb_all_months_short$all_cases_eth[d] -
                    cobb_all_months_short$all_cases_eth[d-1]
  daily_black_aa[d] <- cobb_all_months_short$black_african_american[d] -
                    cobb_all_months_short$black_african_american[d-1]
  daily_hispanic[d] <- cobb_all_months_short$`hispanic_(all_races)`[d] -
                    cobb_all_months_short$`hispanic_(all_races)`[d-1]
  daily_non_hispanic[d] <- cobb_all_months_short$`non-hispanic`[d] -
                      cobb_all_months_short$`non-hispanic`[d-1]
  daily_white[d] <- cobb_all_months_short$white[d] -
                  cobb_all_months_short$white[d-1]
  daily_other_race[d] <- cobb_all_months_short$other_race[d] -
                       cobb_all_months_short$other_race[d-1]
  daily_not_specified[d] <- cobb_all_months_short$not_specified[d] -
                      cobb_all_months_short$not_specified[d-1]
}
cobb_all_months_short$daily_asian <- daily_asian
cobb_all_months_short$daily_all_race <- daily_all_race
cobb_all_months_short$daily_all_eth <- daily_all_eth
cobb_all_months_short$daily_black_aa <- daily_black_aa
cobb_all_months_short$daily_hispanic <- daily_hispanic
cobb_all_months_short$daily_non_hispanic <- daily_non_hispanic
cobb_all_months_short$daily_white <- daily_white
cobb_all_months_short$daily_other_race <- daily_other_race
cobb_all_months_short$daily_not_specified <- daily_not_specified

cobb_all_months_short$daily_asian_rate <- cobb_all_months_short$daily_asian /
                                          cobb_asian_pop
cobb_all_months_short$daily_black_rate <- cobb_all_months_short$daily_black_aa /
                                          cobb_black_pop
cobb_all_months_short$daily_hispanic_rate <- cobb_all_months_short$daily_hispanic /
                                             cobb_hispanic_pop
cobb_all_months_short$daily_non_hispanic_rate <-
  cobb_all_months_short$daily_non_hispanic / cobb_non_hispanic_pop
cobb_all_months_short$daily_white_rate <- cobb_all_months_short$daily_white /
                                          cobb_white_pop

cobb_all_months_short$daily_hispanic_rate[is.nan(cobb_all_months_short$daily_hispanic_rate)] <- 0
cobb_all_months_short$daily_non_hispanic_rate[is.nan(cobb_all_months_short$daily_non_hispanic_rate)] <-
cobb_all_months_short$daily_asian_rate[is.nan(cobb_all_months_short$daily_asian_rate)] <- 0
cobb_all_months_short$daily_black_rate[is.nan(cobb_all_months_short$daily_black_rate)] <- 0
cobb_all_months_short$daily_white_rate[is.nan(cobb_all_months_short$daily_white_rate)] <- 0
```

Data was split into seasonal datasets for initial analysis and to identify any remaining trend breaks.

```
cobb_summer <- cobb_all_months_short %>%
  filter(date < '2020-09-01')
```

```
cobb_fall <- cobb_all_months_short %>%
  filter(date >= '2020-09-01' & date < '2020-12-01')
cobb_winter <- cobb_all_months_short %>%
  filter(date >= '2020-12-01')
```

Data was then recombined into a long format using only the cumulative and daily case rates calculated above. This was done to allow the racial/ethnic group to be used as a factor variable which made graphing easier using the ggplot package.

```
# data frame w only cumulative case rates by race
cobb_all_short_race_rates <- cobb_all_months_short %>%
  select(c(1, 11, 12, 13, 16))
# data frame w only cumulative case rates by ethnicity
cobb_all_short_eth_rates <- cobb_all_months_short %>%
  select(c(1, 14, 15, 17))
# data frame w only daily case rates by race
cobb_all_short_daily_race_rates <- cobb_all_months_short %>%
  select(c(1, 27, 28, 31))
# data frame w only daily case rates by ethnicity
cobb_all_short_daily_eth_rates <- cobb_all_months_short %>%
  select(c(1, 29, 30))


# new data frame for cumulative race rates
cobb_all_long_race_rates <- data.frame()

# converting short form back into long form
for(d in 1:nrow(cobb_all_short_race_rates)){
  date <- cobb_all_short_race_rates$date[d]
  for(c in 2:ncol(cobb_all_short_race_rates)){
    cat <- colnames(cobb_all_short_race_rates)[c]
    rate <- cobb_all_short_race_rates[d,c]
    single_row <- data.frame(rate = rate, category = cat, date = date)
    cobb_all_long_race_rates <- rbind(cobb_all_long_race_rates, single_row)
  }
}

# converting group to a factor variable
cobb_all_long_race_rates$category <- as.factor(cobb_all_long_race_rates$category)

# spliting data up by season
cobb_summer_long_rr <- cobb_all_long_race_rates %>%
  filter(date < '2020-09-01')
cobb_fall_long_rr <- cobb_all_long_race_rates %>%
  filter(date >= '2020-09-01' & date < '2020-12-01')
cobb_winter_long_rr <- cobb_all_long_race_rates %>%
  filter(date >= '2020-12-01')


 # new data frame for cumulative ethnicity rates
cobb_all_long_eth_rates <- data.frame()

# converting short form back into long form
```

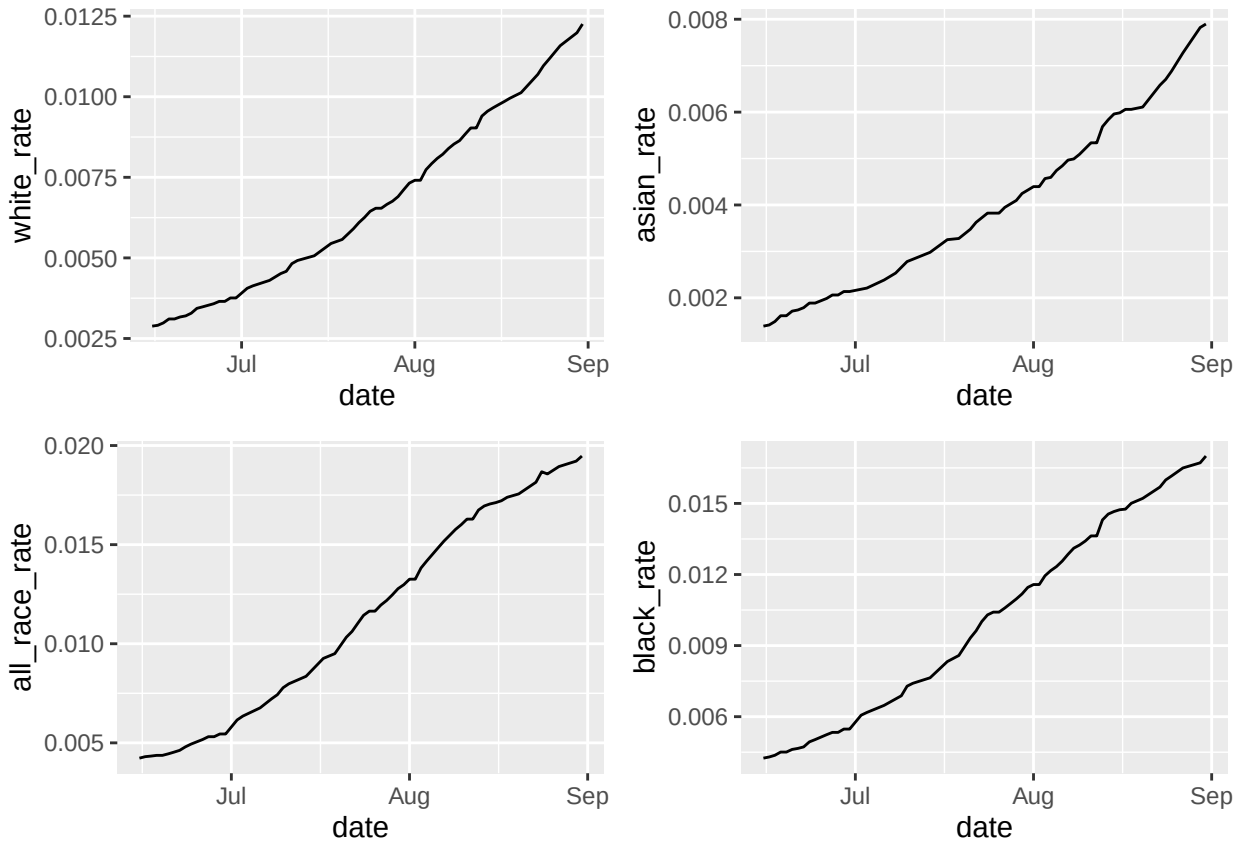
```r
for(d in 1:nrow(cobb_all_short_eth_rates)){
  date <- cobb_all_short_eth_rates$date[d]
  for(c in 2:ncol(cobb_all_short_eth_rates)){
    cat <- colnames(cobb_all_short_eth_rates)[c]
    rate <- cobb_all_short_eth_rates[d,c]
    single_row <- data.frame(rate = rate, category = cat, date = date)
    cobb_all_long_eth_rates <- rbind(cobb_all_long_eth_rates, single_row)
  }
}

# converting group to a factor variable
cobb_all_long_eth_rates$category <- as.factor(cobb_all_long_eth_rates$category)

# splitting data up by season
cobb_summer_long_er <- cobb_all_long_eth_rates %>%
  filter(date < '2020-09-01')
cobb_fall_long_er <- cobb_all_long_eth_rates %>%
  filter(date >= '2020-09-01' & date < '2020-12-01')
cobb_winter_long_er <- cobb_all_long_eth_rates %>%
  filter(date >= '2020-12-01')



# new data frame for daily race rates
cobb_all_long_daily_race_rates <- data.frame()

# converting short form back into long
for(d in 1:nrow(cobb_all_short_daily_race_rates)){
  date <- cobb_all_short_daily_race_rates$date[d]
  for(c in 2:ncol(cobb_all_short_daily_race_rates)){
    cat <- colnames(cobb_all_short_daily_race_rates)[c]
    rate <- cobb_all_short_daily_race_rates[d,c]
    single_row <- data.frame(rate = rate, category = cat, date = date)
    cobb_all_long_daily_race_rates <- rbind(cobb_all_long_daily_race_rates, single_row)
  }
}

# converting group into a factor variable
cobb_all_long_daily_race_rates$category <- as.factor(cobb_all_long_daily_race_rates$category)

# splitting data up by season
cobb_summer_long_drr <- cobb_all_long_daily_race_rates %>%
  filter(date < '2020-09-01')
cobb_fall_long_drr <- cobb_all_long_daily_race_rates %>%
  filter(date >= '2020-09-01' & date < '2020-12-01')
cobb_winter_long_drr <- cobb_all_long_daily_race_rates %>%
  filter(date >= '2020-12-01')



# new data frame for daily ethnicity rates
cobb_all_long_daily_eth_rates <- data.frame()
```

```r
# converting short form back into long form
for(d in 1:nrow(cobb_all_short_daily_eth_rates)){
  date <- cobb_all_short_daily_eth_rates$date[d]
  for(c in 2:ncol(cobb_all_short_daily_eth_rates)){
    cat <- colnames(cobb_all_short_daily_eth_rates)[c]
    rate <- cobb_all_short_daily_eth_rates[d,c]
    single_row <- data.frame(rate = rate, category = cat, date = date)
    cobb_all_long_daily_eth_rates <- rbind(cobb_all_long_daily_eth_rates, single_row)
  }
}

# converting group into factor variable
cobb_all_long_daily_eth_rates$category <- as.factor(cobb_all_long_daily_eth_rates$category)

# splitting data up by season
cobb_summer_long_der <- cobb_all_long_daily_eth_rates %>%
  filter(date < '2020-09-01')
cobb_fall_long_der <- cobb_all_long_daily_eth_rates %>%
  filter(date >= '2020-09-01' & date < '2020-12-01')
cobb_winter_long_der <- cobb_all_long_daily_eth_rates %>%
  filter(date >= '2020-12-01')
```

Data Visualization

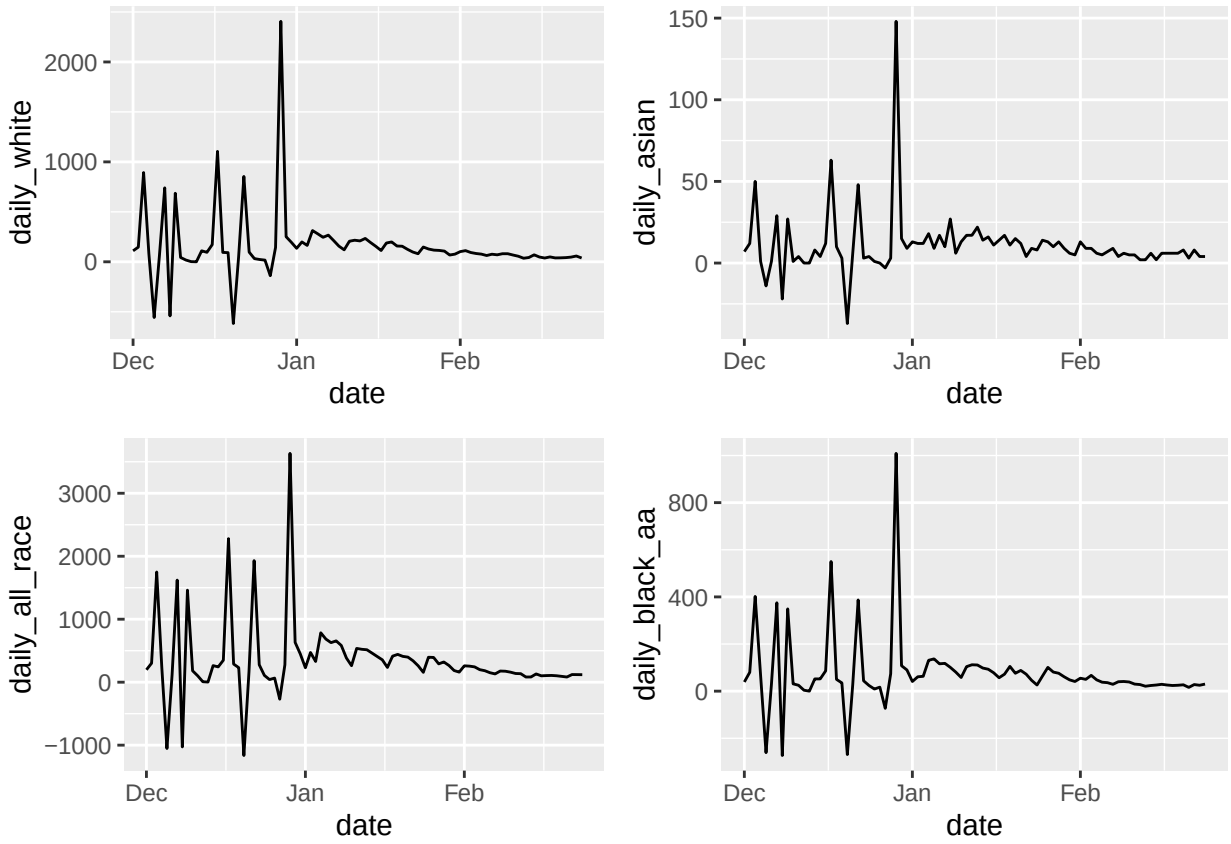
Basic plots from initial short form summer data, split by race.

```r
p1 <- cobb_summer %>%
  ggplot(aes(y= white_rate, x=date)) + geom_line()
p2 <- cobb_summer %>%
  ggplot(aes(y= asian_rate, x=date)) + geom_line()
p3 <- cobb_summer %>%
  ggplot(aes(y=all_race_rate, x=date)) + geom_line()
p4 <- cobb_summer %>%
  ggplot(aes(y=black_rate, x=date)) + geom_line()

grid.arrange(p1, p2, p3, p4)
```



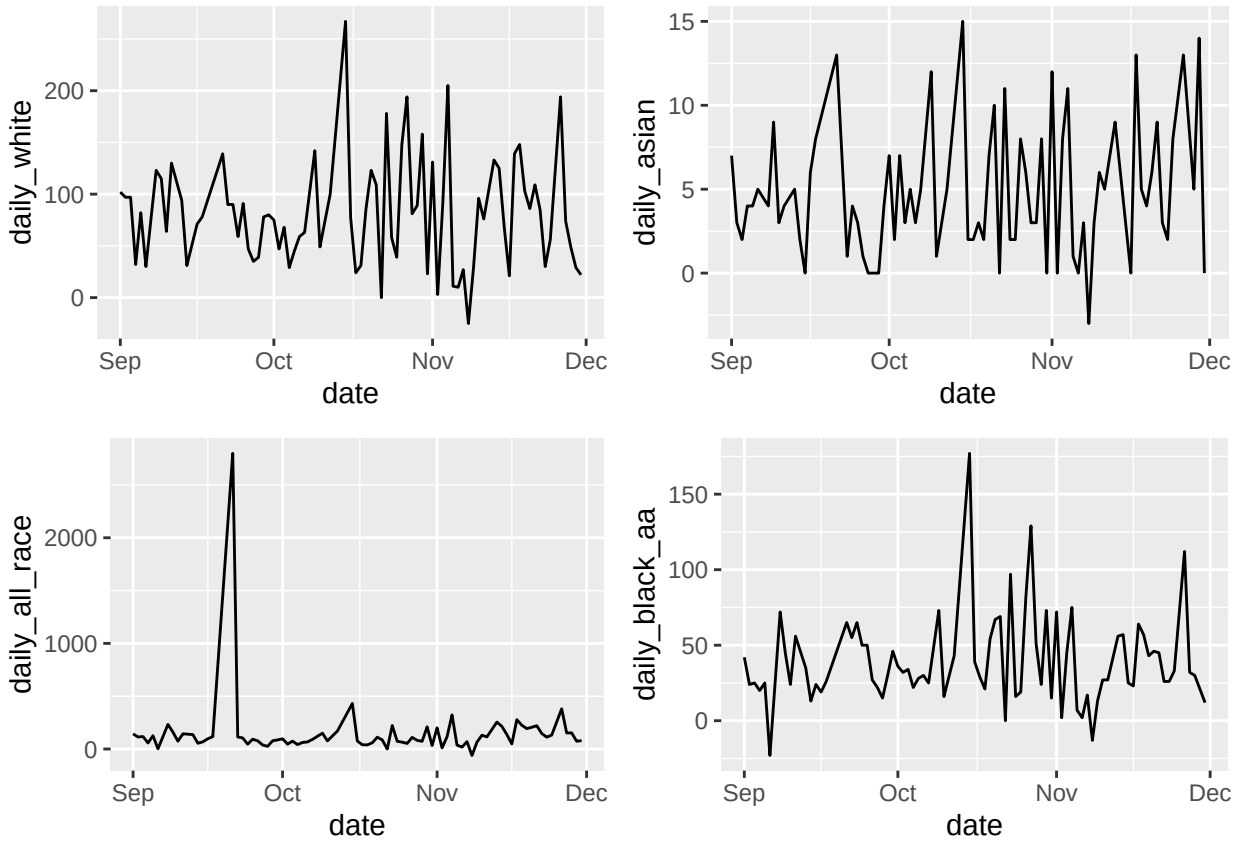
Basic plots from initial short form winter data, split by race.

```
p1 <- cobb_winter %>%
  ggplot(aes(y=daily_white, x=date)) + geom_line()
p2 <- cobb_winter %>%
  ggplot(aes(y=daily_asian, x=date)) + geom_line()
p3 <- cobb_winter %>%
  ggplot(aes(y=daily_all_race, x=date)) + geom_line()
p4 <- cobb_winter %>%
  ggplot(aes(y=daily_black_aa, x=date)) + geom_line()

grid.arrange(p1, p2, p3, p4, ncol=2)
```



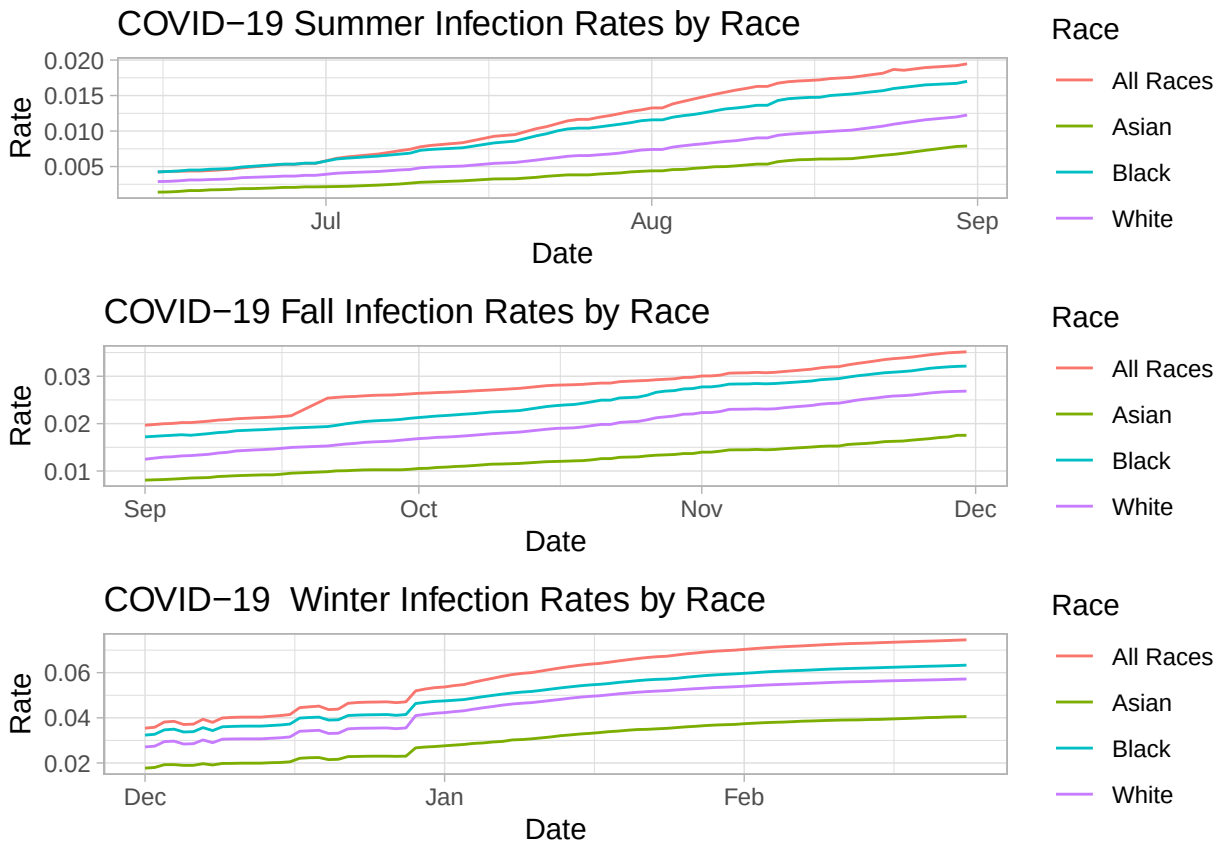
Basic plots from initial short form fall data, split by race.

```
p1 <- cobb_fall %>%
  ggplot(aes(y=daily_white, x=date)) + geom_line()
p2 <- cobb_fall %>%
  ggplot(aes(y=daily_asian, x=date)) + geom_line()
p3 <- cobb_fall %>%
  ggplot(aes(y=daily_all_race, x=date)) + geom_line()
p4 <- cobb_fall %>%
  ggplot(aes(y=daily_black_aa, x=date)) + geom_line()

grid.arrange(p1, p2, p3, p4, ncol=2)
```



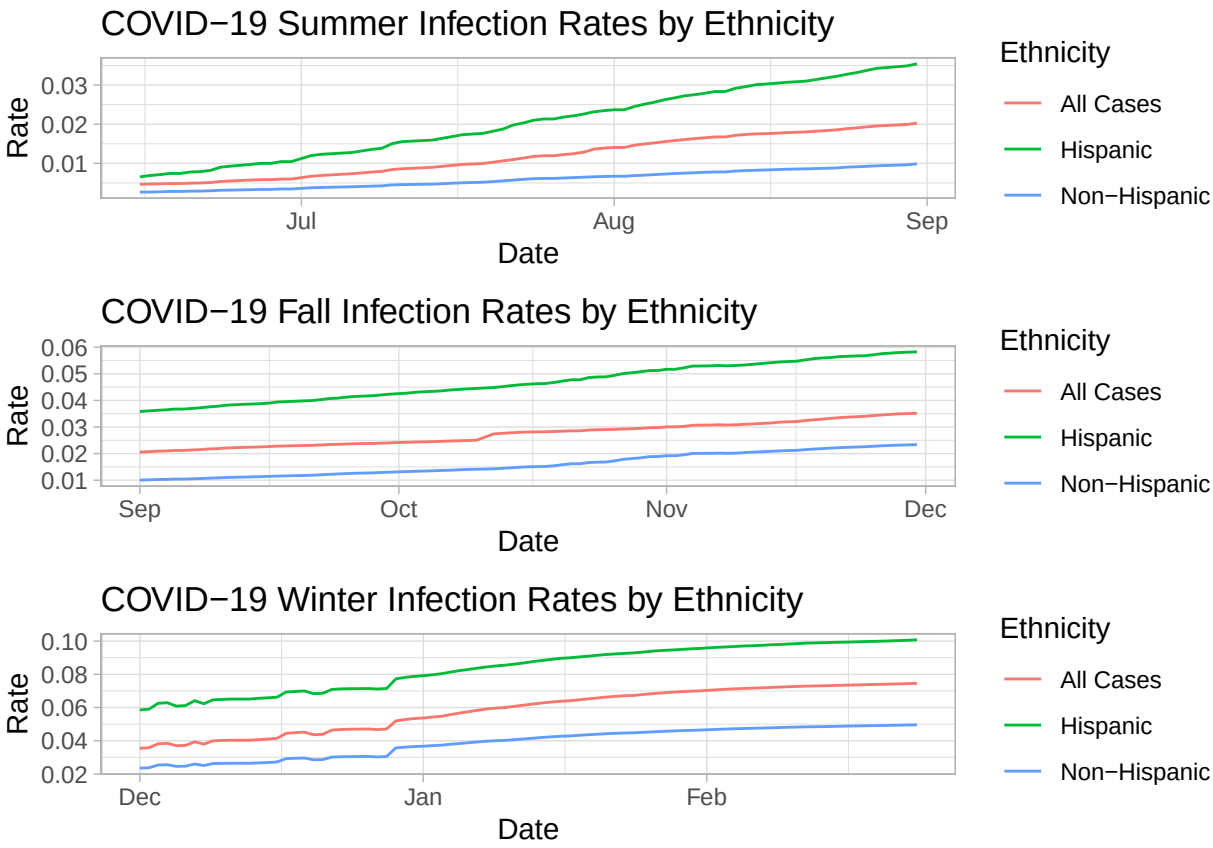
Graphs using long form data sets to show cumulative rates by each race for each season.

```
p1 <- cobb_summer_long_rr %>%
  ggplot(aes(group= category, colour=category, x=date, y=rate)) + geom_line() +
  labs(title="COVID-19 Summer Infection Rates by Race", x="Date", y="Rate") +
  theme_light() +
  scale_colour_discrete(name="Race",
                        labels= c("All Races", "Asian", "Black", "White")) +
  scale_x_date(date_breaks = "1 month", date_labels = "%b")

p2 <- cobb_fall_long_rr %>%
  ggplot(aes(group= category, colour=category, x=date, y=rate)) + geom_line() +
  labs(title="COVID-19 Fall Infection Rates by Race", x="Date", y="Rate") +
  theme_light() +
  scale_colour_discrete(name="Race",
                        labels= c("All Races", "Asian", "Black", "White")) +
  scale_x_date(date_breaks = "1 month", date_labels = "%b")

p3 <- cobb_winter_long_rr %>%
  ggplot(aes(group= category, colour=category, x=date, y=rate)) + geom_line() +
  labs(title="COVID-19  Winter Infection Rates by Race", x="Date", y="Rate") +
  theme_light() +
  scale_colour_discrete(name="Race",
                        labels= c("All Races", "Asian", "Black", "White")) +
  scale_x_date(date_breaks = "1 month", date_labels = "%b")

grid.arrange(p1, p2, p3)
```



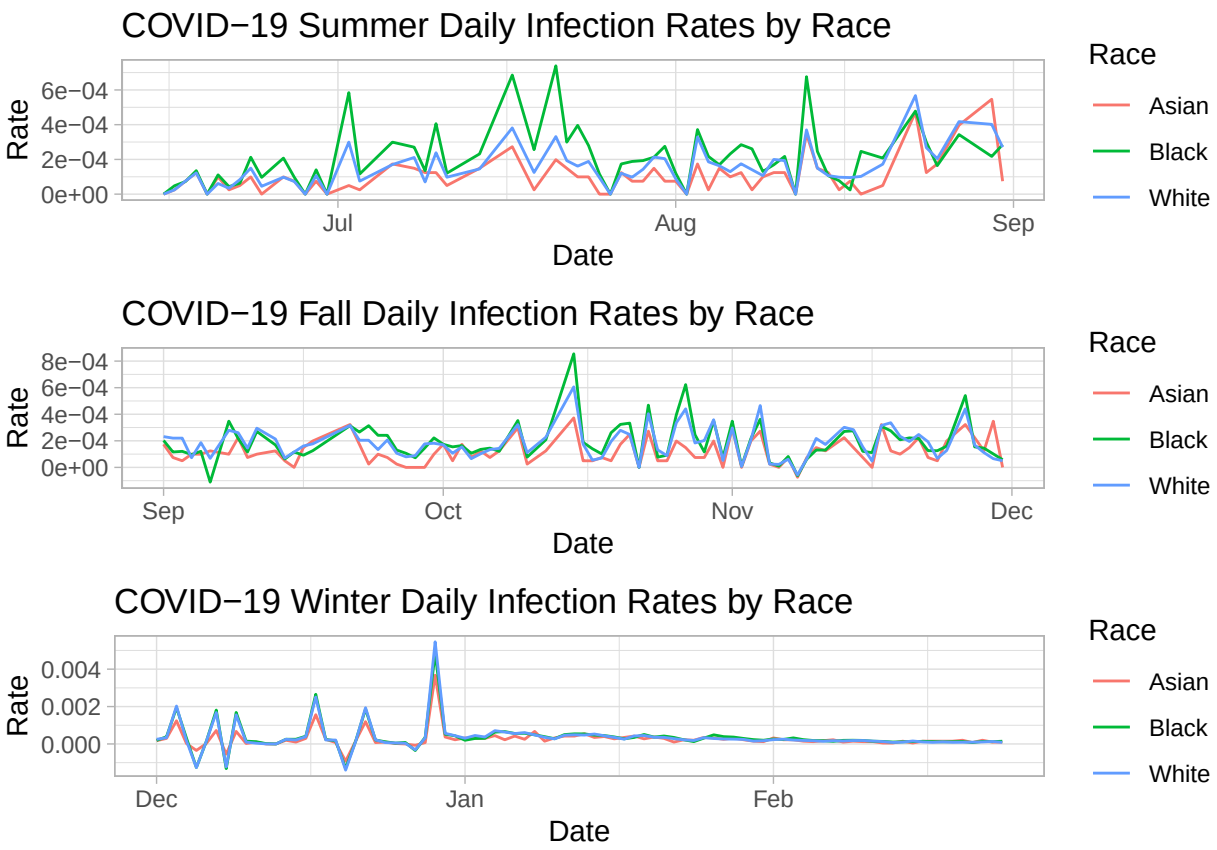
Graphs using long form data sets to show cumulative rates by ethnicity for each season.

```
p1 <- cobb_summer_long_er %>%
  ggplot(aes(group= category, colour=category, x=date, y=rate)) + geom_line() +
  labs(title="COVID-19 Summer Infection Rates by Ethnicity",
       x = "Date", y="Rate") + theme_light() +
  scale_colour_discrete(name="Ethnicity",
                        labels=c("All Cases", "Hispanic", "Non-Hispanic")) +
  scale_x_date(date_breaks = "1 month", date_labels = "%b")

p2 <- cobb_fall_long_er %>%
  ggplot(aes(group= category, colour=category, x=date, y=rate)) + geom_line() +
  labs(title="COVID-19 Fall Infection Rates by Ethnicity",
       x = "Date", y="Rate") + theme_light() +
  scale_colour_discrete(name="Ethnicity",
                        labels=c("All Cases", "Hispanic", "Non-Hispanic")) +
  scale_x_date(date_breaks = "1 month", date_labels = "%b")

p3 <- cobb_winter_long_er %>%
  ggplot(aes(group= category, colour=category, x=date, y=rate)) + geom_line() +
  labs(title="COVID-19 Winter Infection Rates by Ethnicity",
       x = "Date", y="Rate") + theme_light() +
  scale_colour_discrete(name="Ethnicity",
                        labels=c("All Cases", "Hispanic", "Non-Hispanic")) +
  scale_x_date(date_breaks = "1 month", date_labels = "%b")

grid.arrange(p1, p2, p3)
```



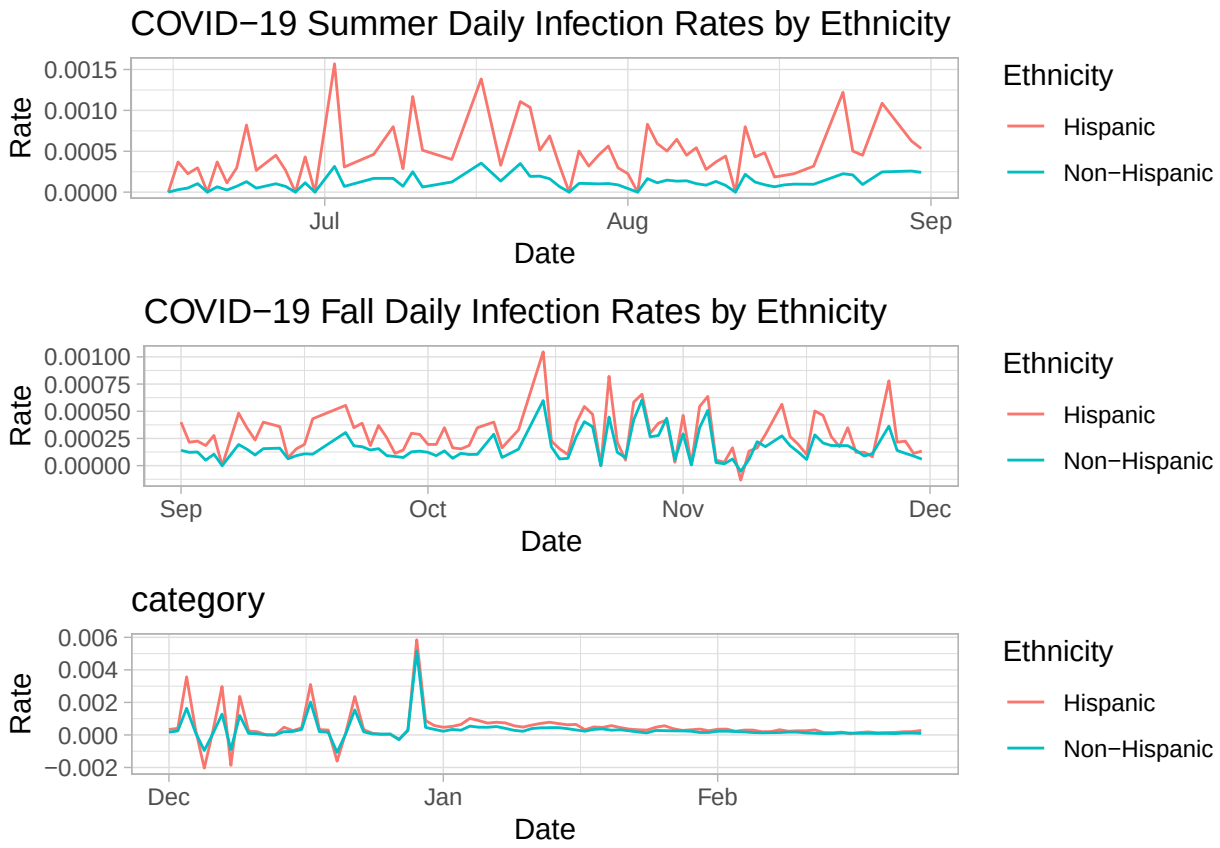
Graphs using long form data sets to show daily rates by race for each season.

```
p1 <- cobb_summer_long_drr %>%
  ggplot(aes(group= category, colour=category, x=date, y=rate)) + geom_line() +
  labs(title="COVID-19 Summer Daily Infection Rates by Race",
       x="Date", y="Rate") + theme_light() +
  scale_colour_discrete(name="Race", labels= c("Asian", "Black", "White")) +
  scale_x_date(date_breaks = "1 month", date_labels = "%b")

p2 <- cobb_fall_long_drr %>%
  ggplot(aes(title = category, colour=category, x=date, y=rate)) + geom_line() +
  labs(title="COVID-19 Fall Daily Infection Rates by Race", x="Date", y="Rate") +
  theme_light() +
  scale_colour_discrete(name="Race", labels= c("Asian", "Black", "White")) +
  scale_x_date(date_breaks = "1 month", date_labels = "%b")

p3 <- cobb_winter_long_drr %>%
  ggplot(aes(group= category, colour=category, x=date, y=rate)) + geom_line() +
  labs(title="COVID-19 Winter Daily Infection Rates by Race",
       x="Date", y="Rate") + theme_light() +
  scale_colour_discrete(name="Race", labels= c("Asian", "Black", "White")) +
  scale_x_date(date_breaks = "1 month", date_labels = "%b")

grid.arrange(p1, p2, p3)
```



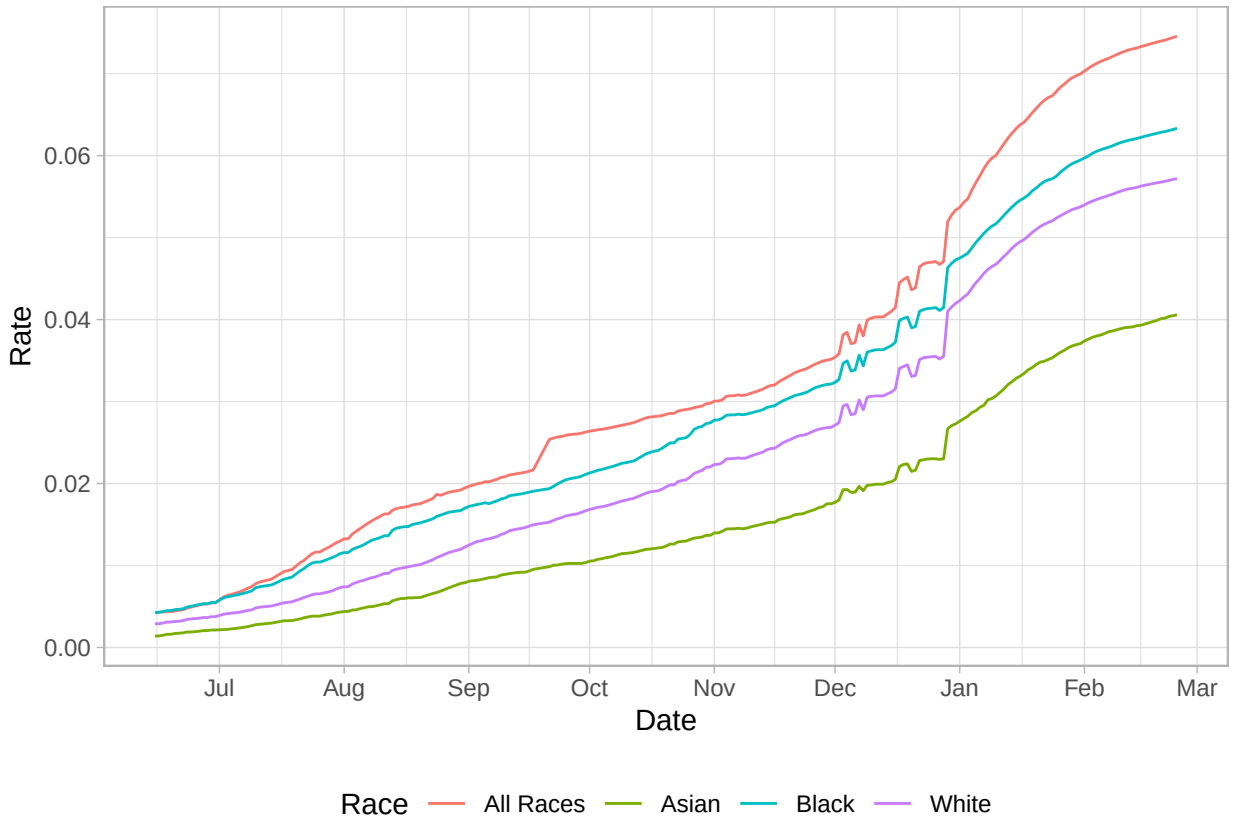
Graphs using long form data sets to show daily rates by ethnicity for each season.

```
p1 <- cobb_summer_long_der %>%
  ggplot(aes(group= category, colour=category, x=date, y=rate)) + geom_line() +
  labs(title="COVID-19 Summer Daily Infection Rates by Ethnicity",
       x = "Date", y="Rate") + theme_light() +
  scale_colour_discrete(name="Ethnicity", labels=c("Hispanic", "Non-Hispanic")) +
  scale_x_date(date_breaks = "1 month", date_labels = "%b")

p2 <- cobb_fall_long_der %>%
  ggplot(aes(group= category, colour=category, x=date, y=rate)) + geom_line() +
  labs(title="COVID-19 Fall Daily Infection Rates by Ethnicity",
       x = "Date", y="Rate") + theme_light() +
  scale_colour_discrete(name="Ethnicity", labels=c("Hispanic", "Non-Hispanic")) +
  scale_x_date(date_breaks = "1 month", date_labels = "%b")

p3 <- cobb_winter_long_der %>%
  ggplot(aes(title= category, colour=category, x=date, y=rate)) + geom_line() +
  labs(main="COVID-19 Winter Daily Infection Rates by Ethnicity",
       x = "Date", y="Rate") + theme_light() +
  scale_colour_discrete(name="Ethnicity", labels=c("Hispanic", "Non-Hispanic")) +
  scale_x_date(date_breaks = "1 month", date_labels = "%b")

grid.arrange(p1, p2, p3)
```

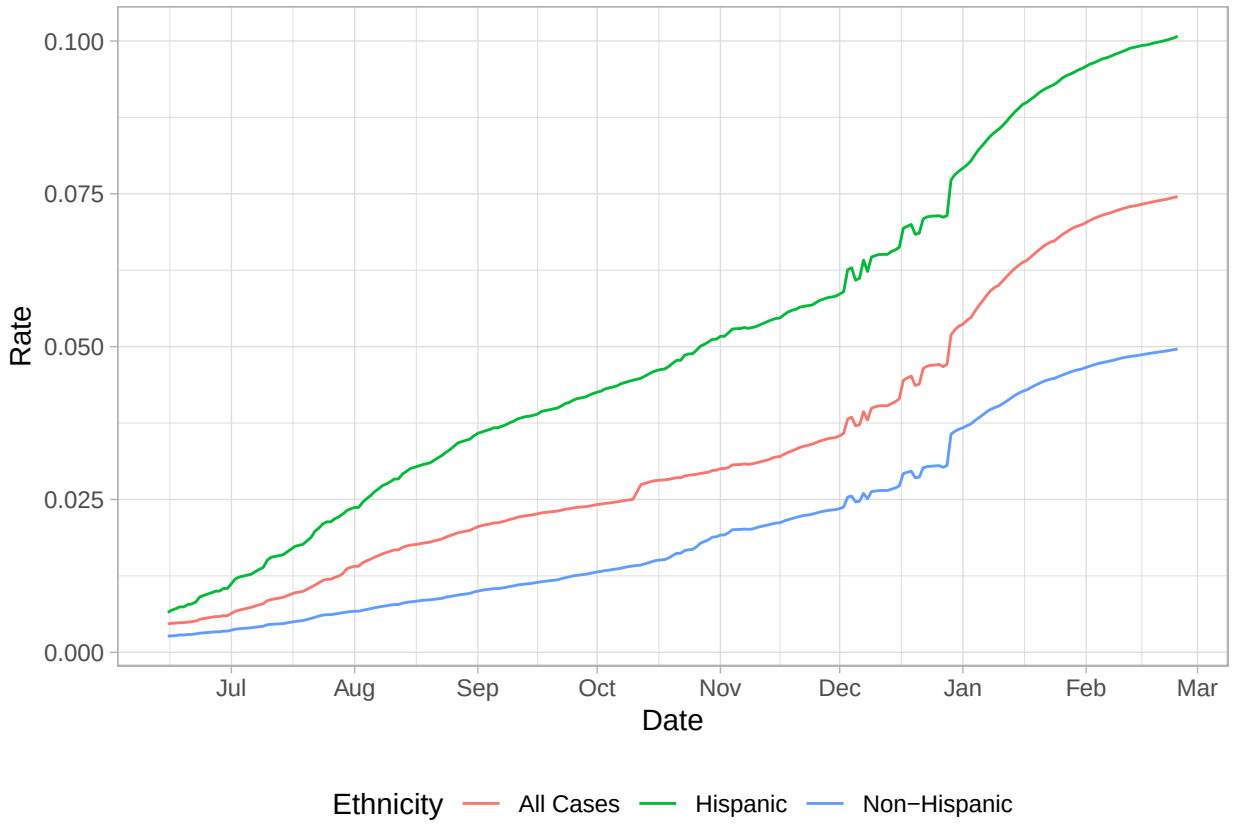


Graphs for all 3 seasons showing the same 4 things as above. Graphs 1 & 2 were used in the presentation

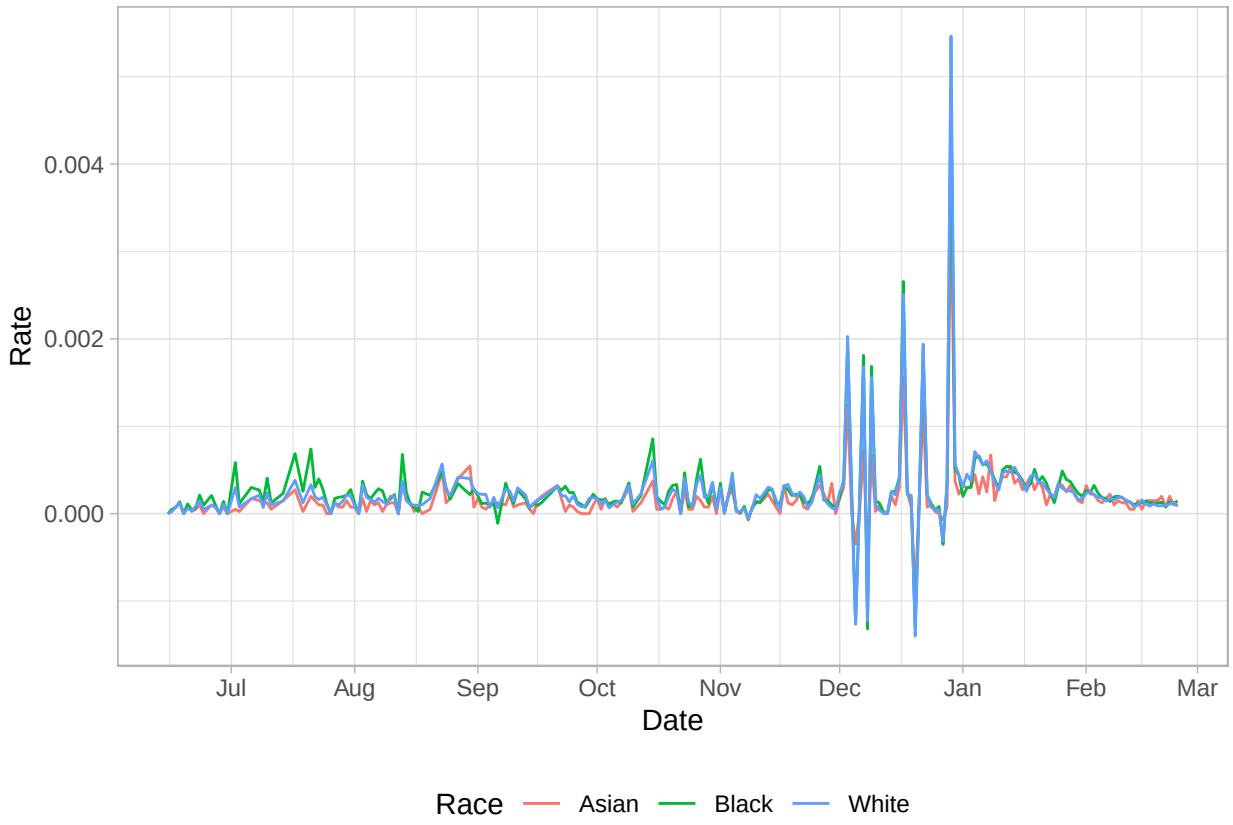
```
cobb_all_long_race_rates %>%
  ggplot(aes(group= category, colour=category, x=date, y=rate)) + geom_line() +
  labs(x="Date", y="Rate") +
  theme_light() +
  scale_colour_discrete(name="Race",
                        labels= c("All Races", "Asian", "Black", "White")) +
  scale_x_date(date_breaks = "1 month", date_labels = "%b") +
  theme(legend.position = "bottom")
```



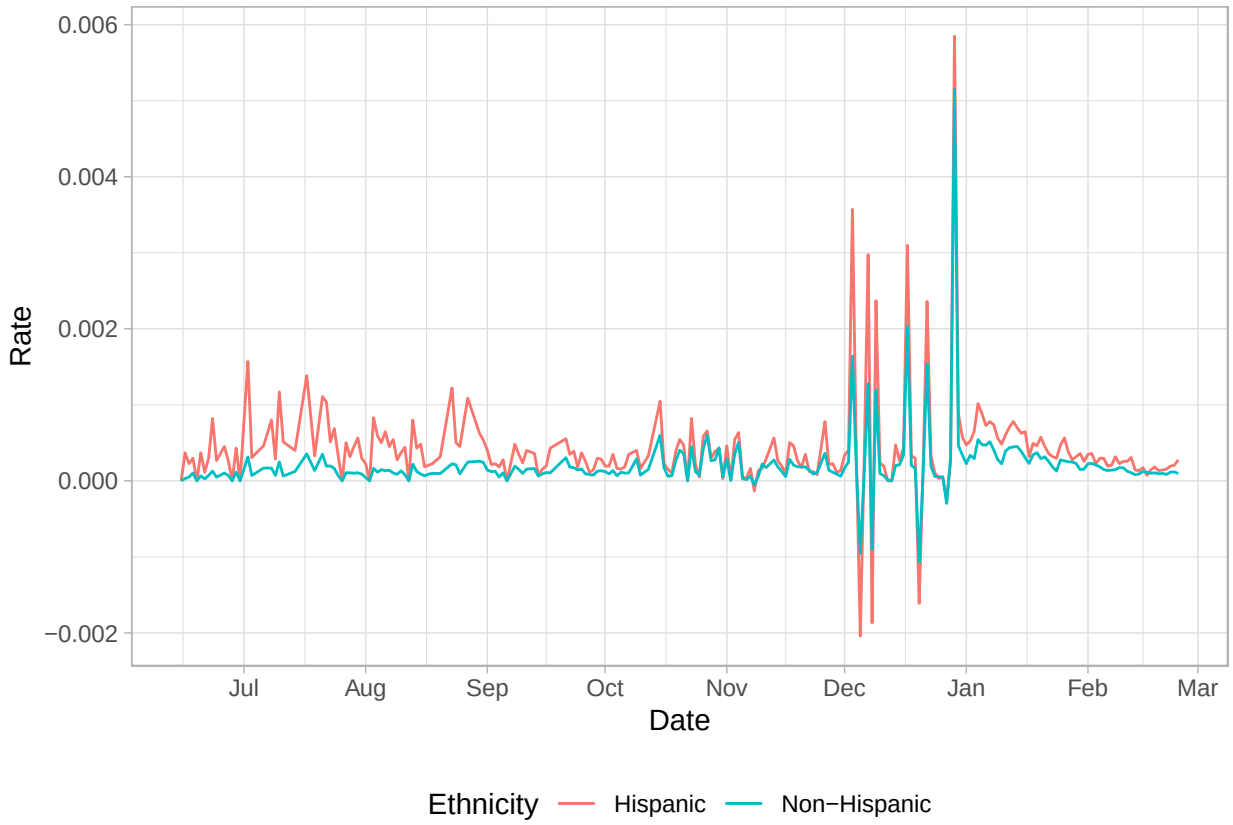

```
cobb_all_long_eth_rates %>%
  ggplot(aes(group= category, colour=category, x=date, y=rate)) + geom_line() +
  labs(x = "Date", y="Rate") +
  theme_light() +
  scale_colour_discrete(name="Ethnicity",
                        labels=c("All Cases", "Hispanic", "Non-Hispanic")) +
  scale_x_date(date_breaks = "1 month", date_labels = "%b") +
  theme(legend.position = "bottom")
```




```
cobb_all_long_daily_race_rates %>%
  ggplot(aes(group= category, colour=category, x=date, y=rate)) + geom_line() +
  labs(x="Date", y="Rate") +
  theme_light() +
  scale_colour_discrete(name="Race", labels= c("Asian", "Black", "White")) +
  scale_x_date(date_breaks = "1 month", date_labels = "%b") +
  theme(legend.position = "bottom")
```




```
cobb_all_long_daily_eth_rates %>%
  ggplot(aes(group= category, colour=category, x=date, y=rate)) + geom_line() +
  labs(x = "Date", y="Rate") +
  theme_light() +
  scale_colour_discrete(name="Ethnicity", labels=c("Hispanic", "Non-Hispanic")) +
  scale_x_date(date_breaks = "1 month", date_labels = "%b") +
  theme(legend.position = "bottom")
```



Removing December Outliers

After doing the initial graphs, I decided to address the trend breaks in December to make the daily case graphs more informative. I removed several of the dates from December to keep the overall case number increasing by a reasonable amount from day-to-day. Similar to the earlier trend breaks, a lot of these dates would show a large increase in case number for a day or two and then a sharp decrease. Since the cumulative case number preceding the large increase and the cumulative case number after the sharp decrease often made sense being consecutive in the data set (in terms of the difference between the 2 numbers), no additional adjustments were made and the daily case counts were recalculated as if those 2 numbers were consecutive days.

Some of the trend breaks present were large increases not followed by subsequent decreases. I assumed that these days were data dumps and included them in the initial data frame used to calculate the daily rates, but dropped that day afterwards to minimize its impact on the daily case rates graphs. Based on the other daily increases in the data set, I considered a daily increase of larger than 1000 cases to be an abnormally high number since the daily rate varied between 200 and 800 usually.

The first date (6/15) was also removed post-daily number calculations since there was no preceding date to calculate the change in numbers for this date.

```
# reselecting only the case counts for each group
all_short_no_dec <- cobb_all_months_short %>%
  select(1:8)

# removing trend break days
all_short_no_dec <- all_short_no_dec %>%
  filter(date != '2020-12-03' & date != '2020-12-04' & date != '2020-12-08' & date != '2020-12-20'
```

```
         & date != '2020-12-21' & date != '2020-12-27')

# get cumulative case counts for all dates
all_short_no_dec$all_cases_race <- all_short_no_dec$asian + all_short_no_dec$black_african_american +
  all_short_no_dec$white + all_short_no_dec$other_race
all_short_no_dec$all_cases_eth <- all_short_no_dec$`hispanic_(all_races)` + all_short_no_dec$`non-hispa
  all_short_no_dec$not_specified

# calculate case rates
all_short_no_dec$asian_rate <- all_short_no_dec$asian / cobb_asian_pop
all_short_no_dec$black_rate <- all_short_no_dec$black_african_american / cobb_black_pop
all_short_no_dec$white_rate <- all_short_no_dec$white / cobb_white_pop
all_short_no_dec$hispanic_rate <- all_short_no_dec$`hispanic_(all_races)` / cobb_hispanic_pop
all_short_no_dec$non_hispanic_rate <- all_short_no_dec$`non-hispanic` / cobb_non_hispanic_pop
all_short_no_dec$all_race_rate <- all_short_no_dec$all_cases_race / cobb_pop_all
all_short_no_dec$all_eth_rate <- all_short_no_dec$all_cases_eth / cobb_pop_all

# calculating daily case counts
daily_asian <- rep(0, nrow(all_short_no_dec))
daily_all_race <- rep(0, nrow(all_short_no_dec))
daily_all_eth <- rep(0, nrow(all_short_no_dec))
daily_black_aa <- rep(0, nrow(all_short_no_dec))
daily_hispanic <- rep(0, nrow(all_short_no_dec))
daily_non_hispanic <- rep(0, nrow(all_short_no_dec))
daily_white <- rep(0, nrow(all_short_no_dec))
daily_other_race <- rep(0, nrow(all_short_no_dec))
daily_not_specified <- rep(0, nrow(all_short_no_dec))

for(d in 2:nrow(all_short_no_dec)){
  daily_asian[d] <- all_short_no_dec$asian[d] -
                    all_short_no_dec$asian[d-1]
  daily_all_race[d] <- all_short_no_dec$all_cases_race[d] -
                       all_short_no_dec$all_cases_race[d-1]
  daily_all_eth[d] <- all_short_no_dec$all_cases_eth[d] -
                      all_short_no_dec$all_cases_eth[d-1]
  daily_black_aa[d] <- all_short_no_dec$black_african_american[d] -
                       all_short_no_dec$black_african_american[d-1]
  daily_hispanic[d] <- all_short_no_dec$`hispanic_(all_races)`[d] -
                         all_short_no_dec$`hispanic_(all_races)`[d-1]
  daily_non_hispanic[d] <- all_short_no_dec$`non-hispanic`[d] -
                           all_short_no_dec$`non-hispanic`[d-1]
  daily_white[d] <- all_short_no_dec$white[d] - all_short_no_dec$white[d-1]
  daily_other_race[d] <- all_short_no_dec$other_race[d] -
                          all_short_no_dec$other_race[d-1]
  daily_not_specified[d] <- all_short_no_dec$not_specified[d] -
                            all_short_no_dec$not_specified[d-1]
}
all_short_no_dec$daily_asian <- daily_asian
all_short_no_dec$daily_all_race <- daily_all_race
all_short_no_dec$daily_all_eth <- daily_all_eth
all_short_no_dec$daily_black_aa <- daily_black_aa
all_short_no_dec$daily_hispanic <- daily_hispanic
all_short_no_dec$daily_non_hispanic <- daily_non_hispanic
```

```r
all_short_no_dec$daily_white <- daily_white
all_short_no_dec$daily_other_race <- daily_other_race
all_short_no_dec$daily_not_specified <- daily_not_specified

# converting daily case counts to case rates
all_short_no_dec$daily_asian_rate <- all_short_no_dec$daily_asian /
                                     cobb_asian_pop
all_short_no_dec$daily_black_rate <- all_short_no_dec$daily_black_aa /
                                     cobb_black_pop
all_short_no_dec$daily_hispanic_rate <- all_short_no_dec$daily_hispanic /
                                       cobb_hispanic_pop
all_short_no_dec$daily_non_hispanic_rate <- all_short_no_dec$daily_non_hispanic /
                                           cobb_non_hispanic_pop
all_short_no_dec$daily_white_rate <- all_short_no_dec$daily_white /
                                     cobb_white_pop

# replacing all NaN case rates w 0
all_short_no_dec$daily_hispanic_rate[is.nan(all_short_no_dec$daily_hispanic_rate)] <- 0
all_short_no_dec$daily_non_hispanic_rate[is.nan(all_short_no_dec$daily_non_hispanic_rate)] <- 0
all_short_no_dec$daily_asian_rate[is.nan(all_short_no_dec$daily_asian_rate)] <- 0
all_short_no_dec$daily_black_rate[is.nan(all_short_no_dec$daily_black_rate)] <- 0
all_short_no_dec$daily_white_rate[is.nan(all_short_no_dec$daily_white_rate)] <- 0

# removing date of data dump on 12/29
all_short_no_dec <- all_short_no_dec %>%
  filter(date != '2020-12-29' & date != '2020-12-17' & date != '2020-12-07')
# removing first date -- 6/15
all_short_no_dec <- all_short_no_dec %>%
  filter(date != '2020-06-15')
```

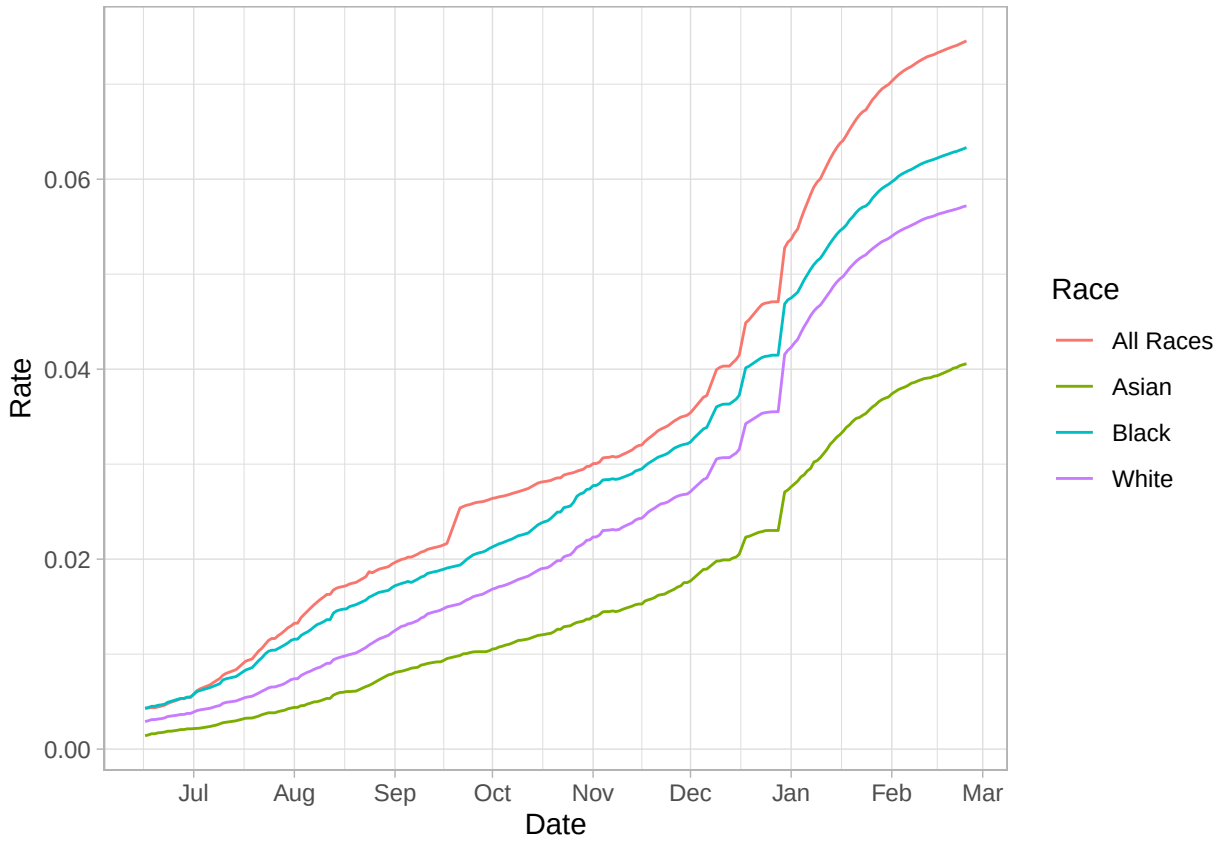
```r
# This code chunk contains essentially the same code as before, just on the
# updated data set
all_short_rr_no_dec <- all_short_no_dec %>%
  select(c(1, 11, 12, 13, 16))
all_short_er_no_dec <- all_short_no_dec %>%
  select(c(1, 14, 15, 17))
all_short_drr_no_dec <- all_short_no_dec %>%
  select(c(1, 27, 28, 31))
all_short_der_no_dec <- all_short_no_dec %>%
  select(c(1, 29, 30))

all_long_rr_no_dec <- data.frame()

for(d in 1:nrow(all_short_rr_no_dec)){
  date <- all_short_rr_no_dec$date[d]
  for(c in 2:ncol(all_short_rr_no_dec)){
    cat <- colnames(all_short_rr_no_dec)[c]
    rate <- all_short_rr_no_dec[d,c]
    single_row <- data.frame(rate = rate, category = cat, date = date)
    all_long_rr_no_dec <- rbind(all_long_rr_no_dec, single_row)
  }
}
```

```r
all_long_rr_no_dec$date <- as.Date(all_long_rr_no_dec$date, format="%Y-%m-%d")
all_long_rr_no_dec$category <- as.factor(all_long_rr_no_dec$category)


all_long_er_no_dec <- data.frame()

for(d in 1:nrow(all_short_er_no_dec)){
  date <- all_short_er_no_dec$date[d]
  for(c in 2:ncol(all_short_er_no_dec)){
    cat <- colnames(all_short_er_no_dec)[c]
    rate <- all_short_er_no_dec[d,c]
    single_row <- data.frame(rate = rate, category = cat, date = date)
    all_long_er_no_dec <- rbind(all_long_er_no_dec, single_row)
  }
}

all_long_er_no_dec$date <- as.Date(all_long_er_no_dec$date)
all_long_er_no_dec$category <- as.factor(all_long_er_no_dec$category)

all_long_drr_no_dec <- data.frame()

for(d in 1:nrow(all_short_drr_no_dec)){
  date <- all_short_drr_no_dec$date[d]
  for(c in 2:ncol(all_short_drr_no_dec)){
    cat <- colnames(all_short_drr_no_dec)[c]
    rate <- all_short_drr_no_dec[d,c]
    single_row <- data.frame(rate = rate, category = cat, date = date)
    all_long_drr_no_dec <- rbind(all_long_drr_no_dec, single_row)
  }
}

all_long_drr_no_dec$date <- as.Date(all_long_drr_no_dec$date)
all_long_drr_no_dec$category <- as.factor(all_long_drr_no_dec$category)

all_long_der_no_dec <- data.frame()

for(d in 1:nrow(all_short_der_no_dec)){
  date <- all_short_der_no_dec$date[d]
  for(c in 2:ncol(all_short_der_no_dec)){
    cat <- colnames(all_short_der_no_dec)[c]
    rate <- all_short_der_no_dec[d,c]
    single_row <- data.frame(rate = rate, category = cat, date = date)
    all_long_der_no_dec <- rbind(all_long_der_no_dec, single_row)
  }
}

all_long_der_no_dec$date <- as.Date(all_long_der_no_dec$date)
all_long_der_no_dec$category <- as.factor(all_long_der_no_dec$category)
```

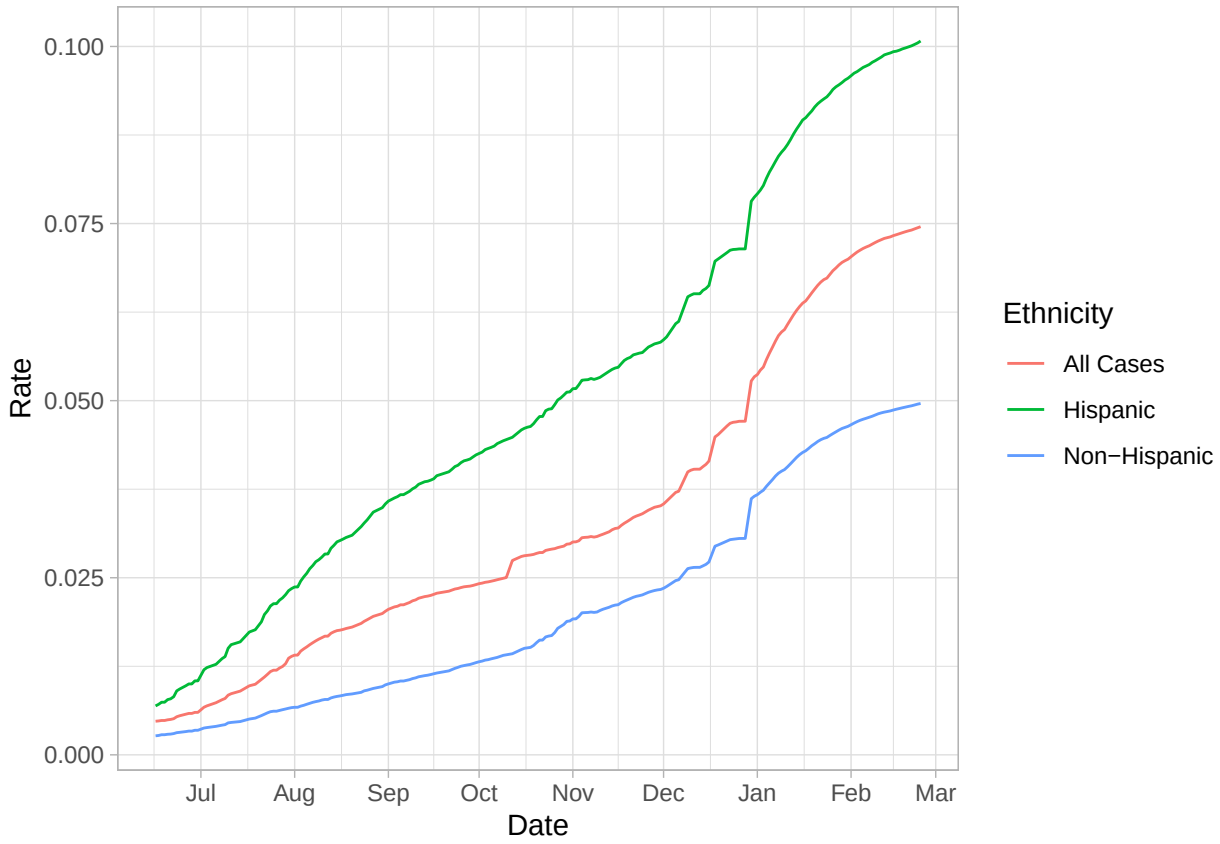
Data Visualization Without December Outliers

Graphs for all 3 seasons showing the same 4 different things as all of the above graphs. Graphs 3 & 4 were used in the presentation

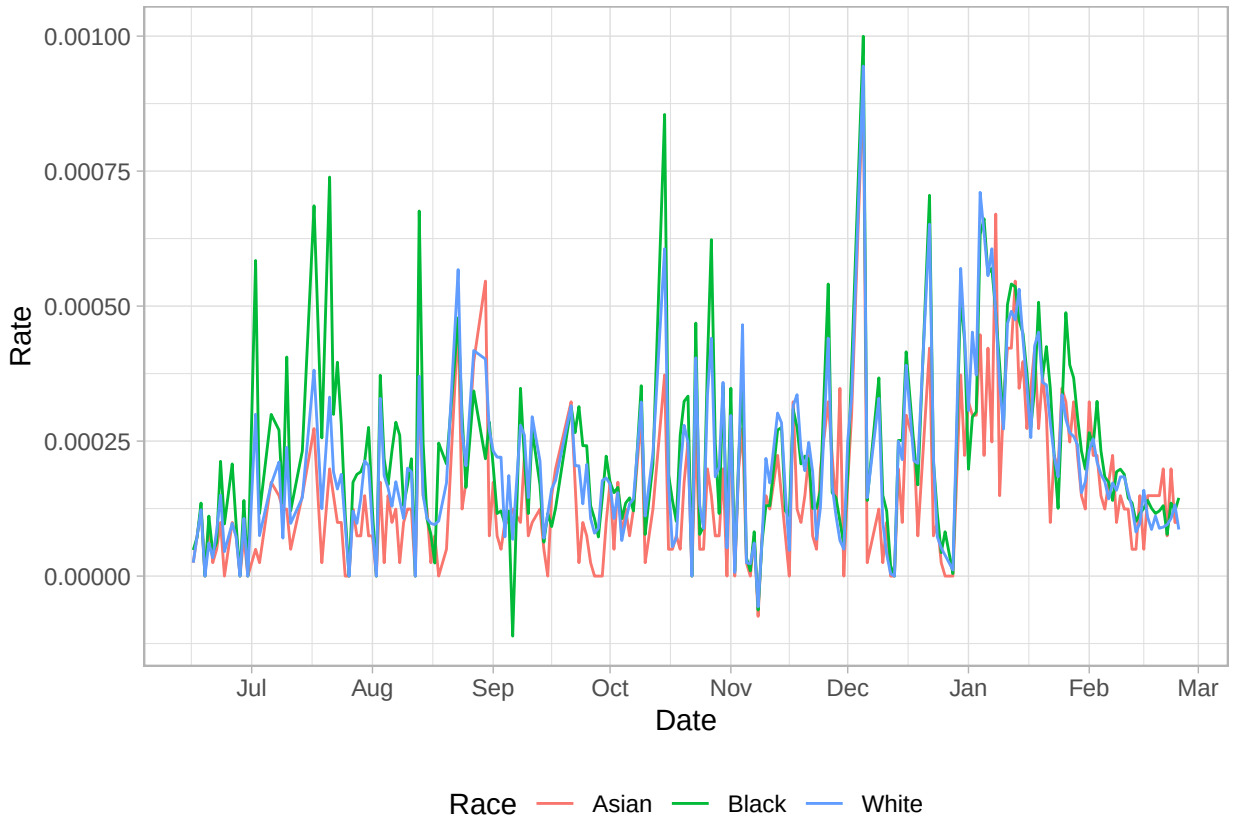
```
all_long_rr_no_dec %>%
  ggplot(aes(group= category, colour=category, x=date, y=rate)) + geom_line() +
  labs(x="Date", y="Rate") + theme_light() +
  scale_colour_discrete(name="Race",
                        labels= c("All Races", "Asian", "Black", "White")) +
  scale_x_date(date_breaks = "1 month", date_labels = "%b")
```



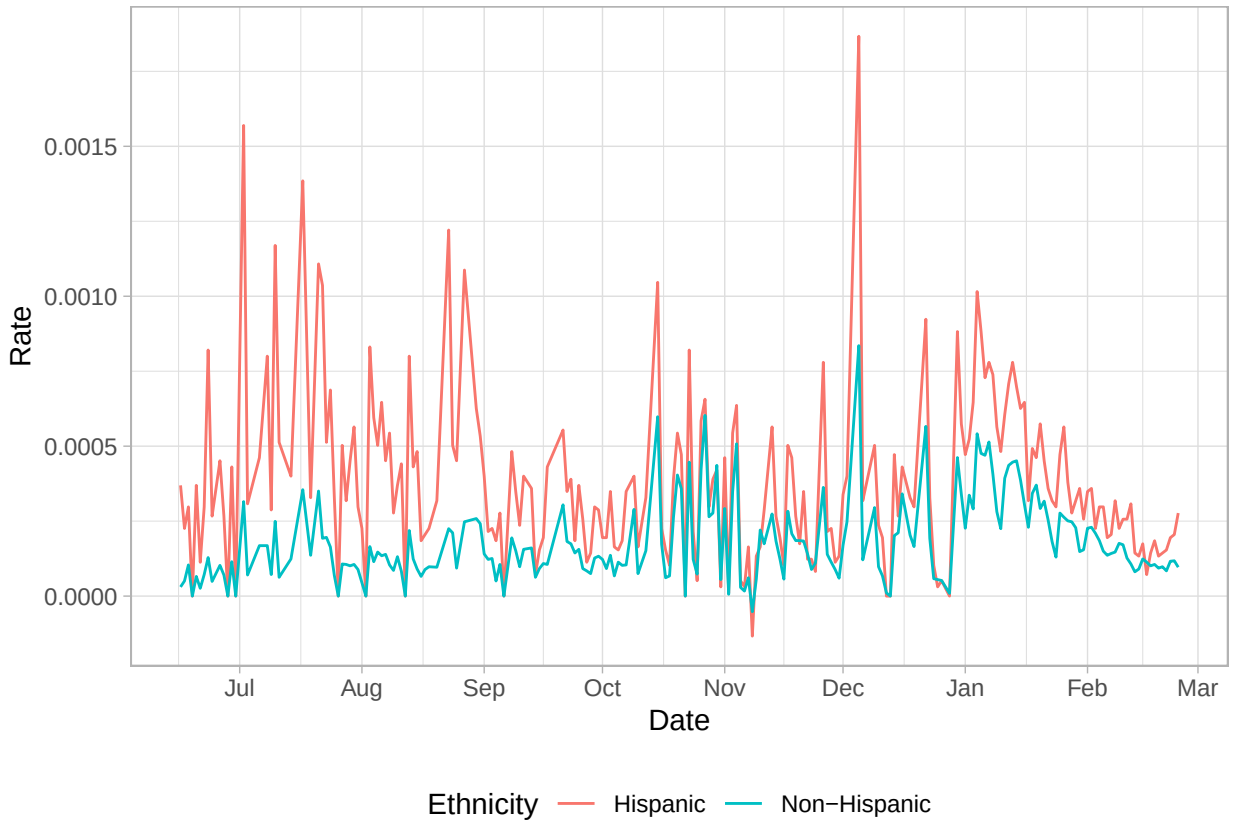

```
all_long_er_no_dec %>%
  ggplot(aes(group= category, colour=category, x=date, y=rate)) + geom_line() +
  labs(x = "Date", y="Rate") + theme_light() +
  scale_colour_discrete(name="Ethnicity",
                        labels=c("All Cases", "Hispanic", "Non-Hispanic")) +
  scale_x_date(date_breaks = "1 month", date_labels = "%b")
```




```
all_long_drr_no_dec %>%
  ggplot(aes(group= category, colour=category, x=date, y=rate)) + geom_line() +
  labs(x="Date", y="Rate") + theme_light() +
  scale_colour_discrete(name="Race", labels= c("Asian", "Black", "White")) +
  scale_x_date(date_breaks = "1 month", date_labels = "%b")+
  theme(legend.position = "bottom")
```




```
all_long_der_no_dec %>%
  ggplot(aes(group= category, colour=category, x=date, y=rate)) + geom_line() +
  labs(x = "Date", y="Rate") + theme_light() +
  scale_colour_discrete(name="Ethnicity", labels=c("Hispanic", "Non-Hispanic")) +
  scale_x_date(date_breaks = "1 month", date_labels = "%b")+
  theme(legend.position = "bottom")
```



After looking at the differences on the graphs, I decided to split the new data without the trend breaks in December up by season and find the average daily increase in cases. Due to the large number of dates which were removed, I added a column into each data frame which indicated how many days that date was after the previous date in the data. I then divided the daily case rate by this new column to help adjust for missing days in the data.

```
# splitting non-trend break data up by seasons
summer_nd <- all_short_no_dec %>%
  filter(date < '2020-09-01')
fall_nd <- all_short_no_dec %>%
  filter(date >= '2020-09-01' & date < '2020-12-01')
winter_nd <- all_short_no_dec %>%
  filter(date >= '2020-12-01')

# add column to indicate how many days from previous data point
summer_nd$num_days <- c(1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 2, 1, 3, 2, 1,
                        1, 1, 3, 3, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                        1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 3, 1, 1,
                        2, 3, 1)
fall_nd$num_days <- c(1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 2, 1, 1, 1, 1, 4, 1, 1, 1, 1,
                      1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 2, 3, 1, 1, 1, 1,
                      1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                      1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1)
winter_nd$num_days <-c(1, 1, 3, 1, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1, 1, 1, 1,
                       2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)
```

```r
# calculating mean number of daily cases for each group, adjusted by number of
# preceding days
asian_summer <- mean(summer_nd$daily_asian / summer_nd$num_days)
asian_fall <- mean(fall_nd$daily_asian / fall_nd$num_days)
asian_winter <- mean(winter_nd$daily_asian / winter_nd$num_days)
asian_avg <- c(asian_summer, asian_fall, asian_winter)

black_summer <- mean(summer_nd$daily_black_aa / summer_nd$num_days)
black_fall <- mean(fall_nd$daily_black_aa / fall_nd$num_days)
black_winter <- mean(winter_nd$daily_black_aa / winter_nd$num_days)
black_avg <- c(black_summer, black_fall, black_winter)

white_summer <- mean(summer_nd$daily_white / summer_nd$num_days)
white_fall <- mean(fall_nd$daily_white / fall_nd$num_days)
white_winter <- mean(winter_nd$daily_white / winter_nd$num_days)
white_avg <- c(white_summer, white_fall, white_winter)

all_race_summer <- mean(summer_nd$daily_all_race / summer_nd$num_days)
all_race_fall <- mean(fall_nd$daily_all_race / fall_nd$num_days)
all_race_winter <- mean(winter_nd$daily_all_race / winter_nd$num_days)
all_race_avg <- c(all_race_summer, all_race_fall, all_race_winter)

hispanic_summer <- mean(summer_nd$daily_hispanic / summer_nd$num_days)
hispanic_fall <- mean(fall_nd$daily_hispanic / fall_nd$num_days)
hispanic_winter <- mean(winter_nd$daily_hispanic / winter_nd$num_days)
hispanic_avg <- c(hispanic_summer, hispanic_fall, hispanic_winter)

non_hispanic_summer <- mean(summer_nd$daily_non_hispanic / summer_nd$num_days)
non_hispanic_fall <- mean(fall_nd$daily_non_hispanic / fall_nd$num_days)
non_hispanic_winter <- mean(winter_nd$daily_non_hispanic / winter_nd$num_days)
non_hispanic_avg <- c(non_hispanic_summer, non_hispanic_fall, non_hispanic_winter)

all_eth_summer <- mean(summer_nd$daily_all_eth / summer_nd$num_days)
all_eth_fall <- mean(fall_nd$daily_all_eth / fall_nd$num_days)
all_eth_winter <- mean(winter_nd$daily_all_eth / winter_nd$num_days)
all_eth_avg <- c(all_eth_summer, all_eth_fall, all_eth_winter)

# creating table of numbers for races
col_names <- c("Summer", "Fall", "Winter")
row_names <- c("All Races", "Asian", "Black", "White")
avg_df <- rbind(all_race_avg, asian_avg, black_avg, white_avg)
rownames(avg_df) <- row_names
kable(avg_df,  digits=2, format="markdown", col.names= col_names, align='c')
```

	Summer	Fall	Winter
All Races	150.64	109.16	266.72
Asian	3.32	4.30	8.42
Black	35.10	34.66	55.03
White	54.57	71.96	110.92

```r
# creating table of numbers for ethnicities
row_names <- c("All Ethnicities", "Hispanic", "Non-Hispanic")
avg_df <- rbind(all_eth_avg, hispanic_avg, non_hispanic_avg)
rownames(avg_df) <- row_names
kable(avg_df,  digits=2, format="markdown", col.names= col_names, align='c')
```

	Summer	Fall	Winter
All Ethnicities	156.29	115.34	266.72
Hispanic	37.90	24.69	35.95
Non-Hispanic	62.76	99.09	141.52

Comparing General Broadstreet Reported Case Data to Reported Equity Data

To verify the reported total number of cases, I downloaded the general data set from Broadstreet and compared the total cumulative and daily totals for all races and all ethnicities (including the not specified and unknown race categories) to the confirmed positive cases in Cobb County as reported in the general data set. I wanted to check the output from above and see if the general trends matched, especially since the daily numbers Broadstreet data set was checked by QA to reduce trend breaks. I only used the confirmed positive cases and not the probable cases in the comparison, since the Health Equity data was only reported for confirmed cases.

```r
# uploading latest release of Broadstreet daily numbers data set
cummulative_counts_raw <- read_csv("../Coronavirus by County.csv")
```

```
## Warning: Duplicated column names deduplicated: 'mort_020220' =>
## 'mort_020220_1' [60], 'pbmort_020220' => 'pbmort_020220_1' [61]
```

```
## Warning: 4878 parsing failures.
##  row           col expected actual                           file
## 1233 tstpos_050120 a number  #REF! '../Coronavirus by County.csv'
## 1233 pbpos_050120  a double  #REF! '../Coronavirus by County.csv'
## 1233 mort_050120   a number  #REF! '../Coronavirus by County.csv'
## 1233 pbmort_050120 a double  #REF! '../Coronavirus by County.csv'
## 1233 tstpos_050220 a number  #REF! '../Coronavirus by County.csv'
## .... ............. ........ ...... ..............................
## See problems(...) for more details.
```

```r
# filtering to only include Cobb County, using FIPS code
cobb_only <- cummulative_counts_raw %>%
  filter(fips == "05000US13067")

# drop data prior to 6/15/20 and after 2/28/21
cobb_cases_deaths <- cobb_only %>%
  select(-c(1:593, 1631:length(cobb_only)))

# only keep confirmed cases from reported numbers

reported_cases <- data.frame(date = as.Date(character()), all_cases = integer())
```

```r
for(c in 1:ncol(cobb_cases_deaths)){
  col <- colnames(cobb_cases_deaths)[c]
  if(str_detect(col, "tstpos")){
    year <- str_c('20', substring(col, 12, 13), sep ='')
    day <- substring(col, 10, 11)
    month <- substring(col, 8, 9)
    date <- as.Date(str_c(year, month, day, sep = '-'))
    new_row <- data.frame(date=date,
                          reported_cases=as.data.frame(cobb_cases_deaths)[1,col])
    reported_cases <- rbind(reported_cases, new_row)
  }
}

# adding confrimed positive counts to the seasonal data frames
cobb_summer <- inner_join(cobb_summer, reported_cases, by = "date")
cobb_winter <- inner_join(cobb_winter, reported_cases, by = "date")
cobb_fall <- inner_join(cobb_fall, reported_cases, by = "date")

# calculating the number of non-reported cases
cobb_summer$non_reported_race <- cobb_summer$reported_cases -
                                 cobb_summer$all_cases_race
cobb_summer$non_reported_eth <- cobb_summer$reported_cases -
                                 cobb_summer$all_cases_eth
cobb_fall$non_reported_race <- cobb_fall$reported_cases -
                                 cobb_fall$all_cases_race
cobb_fall$non_reported_eth <- cobb_fall$reported_cases -
                                 cobb_fall$all_cases_eth
cobb_winter$non_reported_race <- cobb_winter$reported_cases -
                                 cobb_winter$all_cases_race
cobb_winter$non_reported_eth <- cobb_winter$reported_cases -
                                 cobb_winter$all_cases_eth
```

```r
# printing mean proportion of cases reported in the general Broadstreet data set
# that were not included in the Health Equity Data set, split by season and
# race v. ethnicity totals
mean(cobb_summer$non_reported_race/ cobb_summer$reported_cases)
```

```
## [1] 0.3205592
```

```r
mean(cobb_fall$non_reported_race/ cobb_fall$reported_cases)
```

```
## [1] 0.07067608
```

```r
mean(cobb_winter$non_reported_race/ cobb_winter$reported_cases)
```

```
## [1] 0.01408401
```

```r
mean(cobb_summer$non_reported_eth/ cobb_summer$reported_cases)
```

```
## [1] 0.2808581
```

```
mean(cobb_fall$non_reported_eth/ cobb_fall$reported_cases)
```

```
## [1] 0.08240225
```

```
mean(cobb_winter$non_reported_eth/ cobb_winter$reported_cases)
```

```
## [1] 0.01408401
```

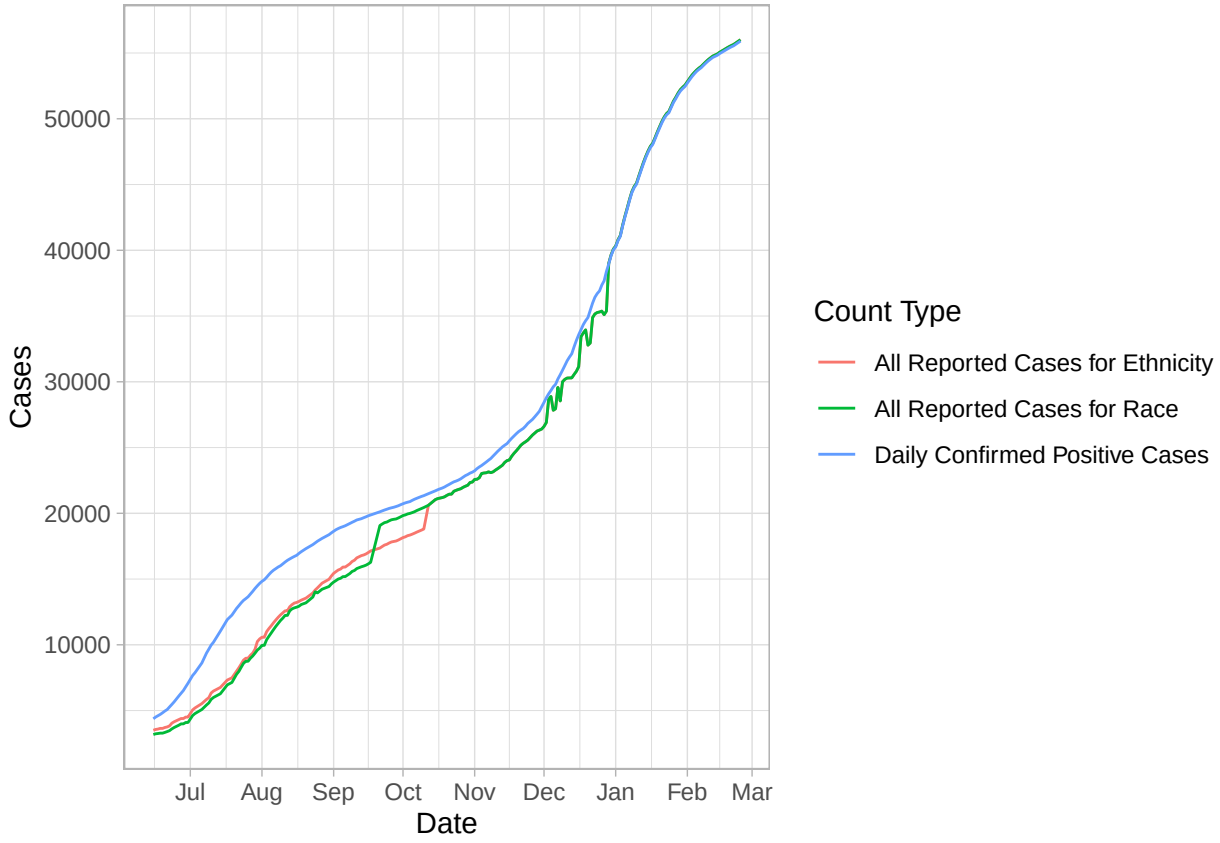
Similar to above, I combined the Broadstreet overall confirmed case totals, the total number of cases for all races and the total number of cases for ethnicity into a long form dataset, so that the 3 cumulative counts could be compared easily using ggplot.

```
# combined cumulative case count data frame
all_case_totals <- rbind(cobb_summer, cobb_winter, cobb_fall) %>%
  select(reported_cases, all_cases_race, all_cases_eth, date)

all_long_case_counts <- data.frame()

for(d in 1:nrow(all_case_totals)){
  date <- all_case_totals$date[d]
  for(c in 1:(ncol(all_case_totals)-1)){
    cat <- colnames(all_case_totals)[c]
    count <- all_case_totals[d,c]
    single_row <- data.frame(count = count, category = cat, date = date)
    all_long_case_counts<- rbind(all_long_case_counts, single_row)
  }
}

all_long_case_counts$category <- as.factor(all_long_case_counts$category)


# add in daily number differences for Broadstreet Overall Data
# using original data set which includes dates that were deleted from the
# health equity data set

daily_cases <- rep(0, nrow(reported_cases))

for(d in 2:nrow(reported_cases)){
  daily_cases[d] <- reported_cases$reported_cases[d] -
                    reported_cases$reported_cases[d-1]
}
reported_cases$daily_cases <- daily_cases

# removing 6/15 from daily numbers count
reported_cases <- reported_cases %>%
  filter(date != '2020-06-15')
```

Graph displaying the 3 different cumulative counts. Included in presentation

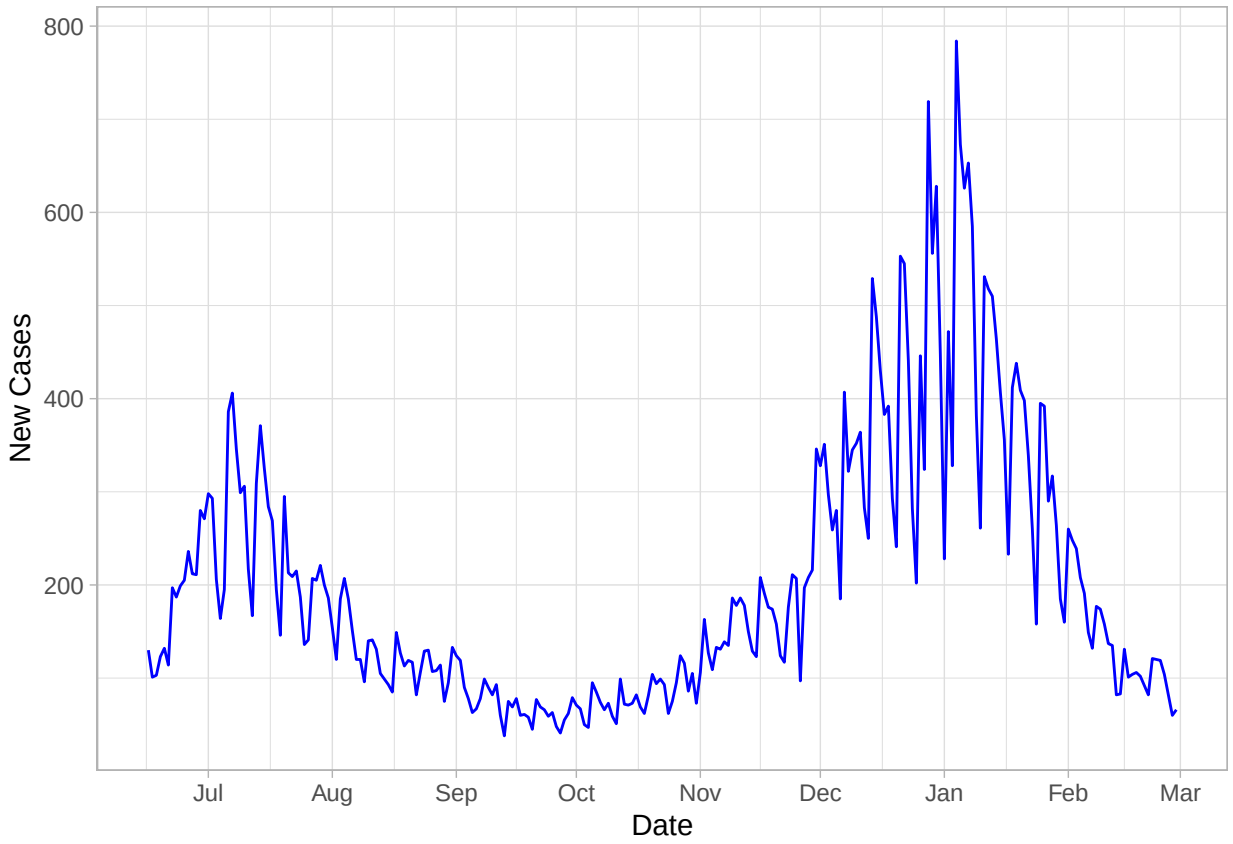
```
all_long_case_counts %>%
  ggplot(aes(group= category, colour=category, x=date, y=count)) + geom_line() +
```

```
  labs(x ="Date", y="Cases") + theme_light() +
  scale_colour_discrete(name="Count Type",
                        labels=c("All Reported Cases for Ethnicity",
                                 "All Reported Cases for Race",
                                 "Daily Confirmed Positive Cases")) +
  scale_x_date(date_breaks = "1 month", date_labels = "%b")
```



Graph showing the daily changes in the Broadstreet overall data.

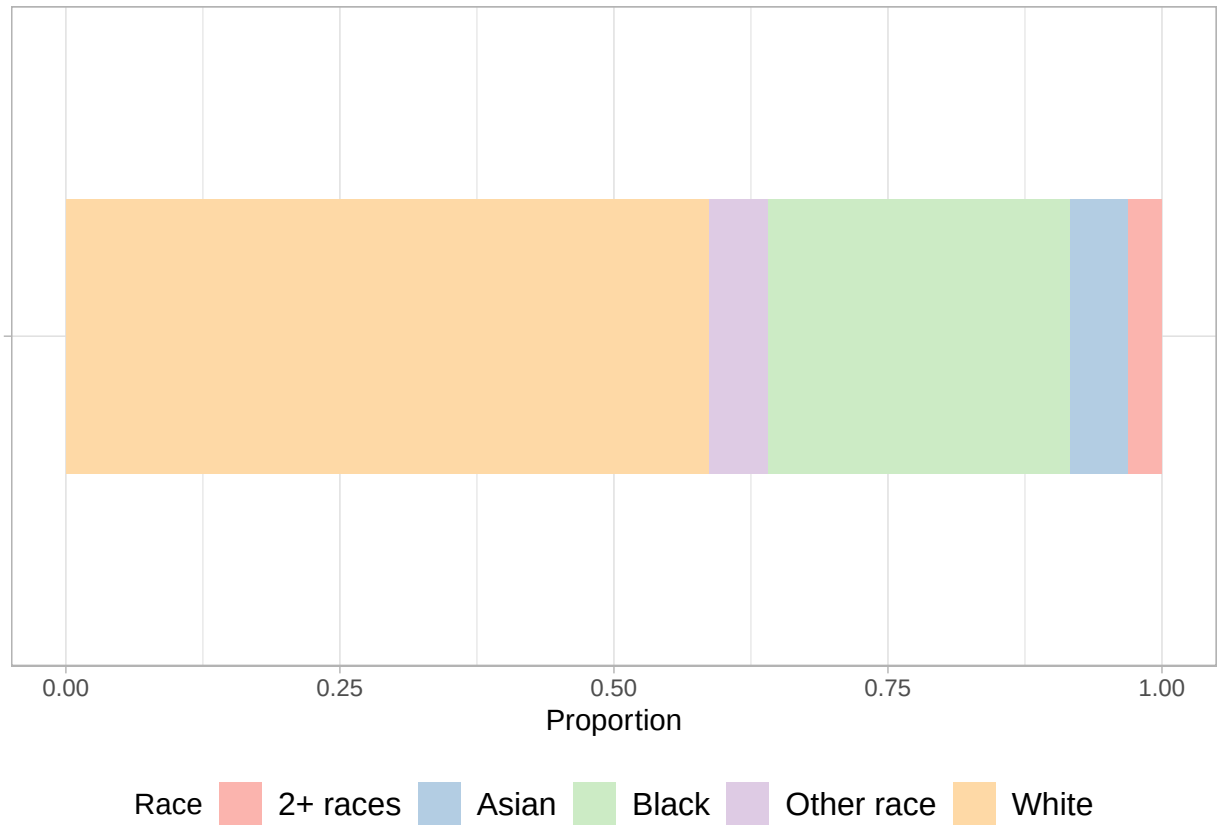
```
reported_cases %>%
    ggplot(aes(x=date, y=daily_cases)) + geom_line(color="blue") +
  labs(x ="Date", y="New Cases") + theme_light() +
  scale_x_date(date_breaks = "1 month", date_labels = "%b")
```



Charts for Presentation

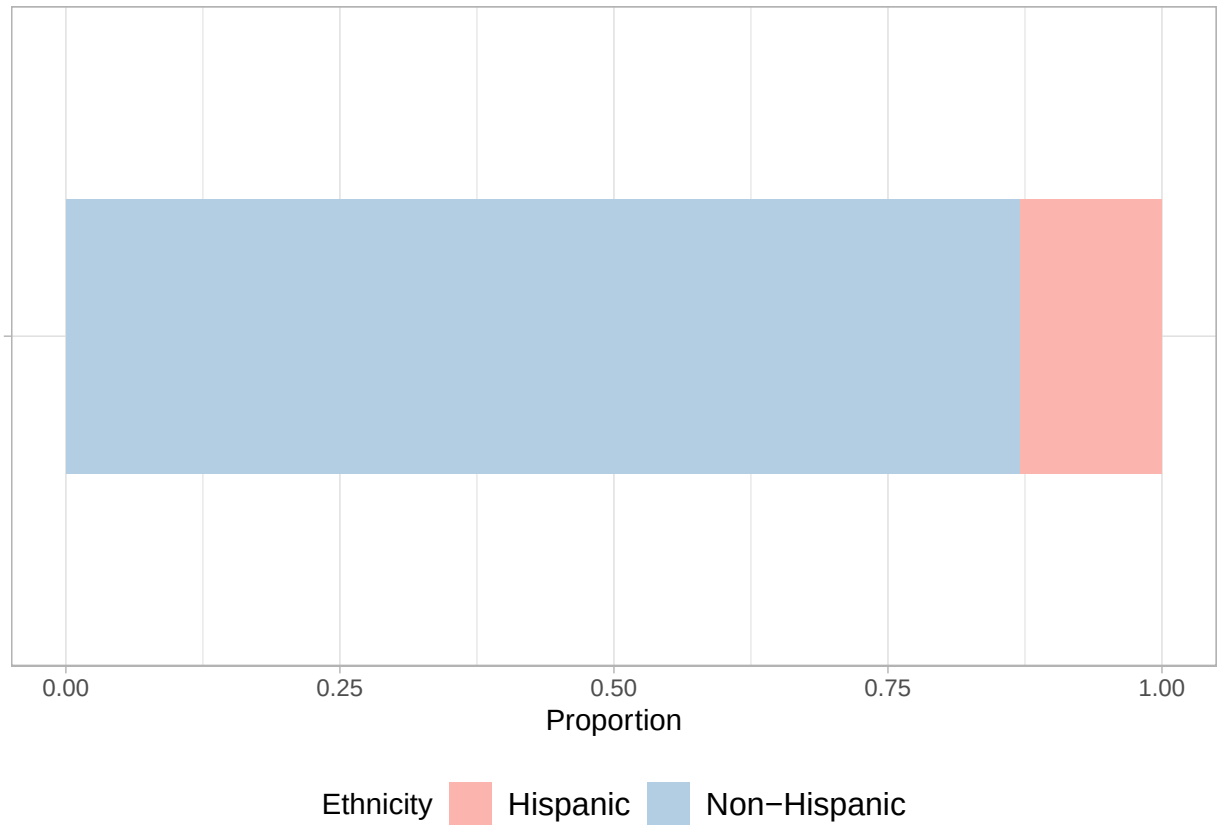
```
cobb_aian_pop <- 1941
cobb_nhpi_pop <- 409
cobb_two_plus_pop <- 22988
cobb_other_race_pop <- 38088 + cobb_aian_pop + cobb_nhpi_pop
```

```
pops <- c(cobb_asian_pop, cobb_black_pop, cobb_white_pop, cobb_two_plus_pop,
          cobb_other_race_pop)
races <- c("Asian", "Black", "White","2+ races", "Other race")
race.proportions.df <- data.frame(Race=as.factor(races),
                                  Proportion=pops/cobb_pop_all)

race.proportions.df %>%
  ggplot(aes(x="", y=Proportion, fill=Race)) +
  geom_bar(stat="identity", width=.5) +
  theme_light() + coord_flip() +
  theme(legend.position = "bottom", legend.text = element_text(size=12)) +
  scale_fill_brewer(palette="Pastel1") + labs(x=NULL)
```




```
ggsave("Race.png", height=5, width=10)
```

```
pops <- c(cobb_hispanic_pop, cobb_non_hispanic_pop)
ethnicity <- c("Hispanic", "Non-Hispanic")
eth.proportions.df <- data.frame(Ethnicity=as.factor(ethnicity),
                                 Proportion=pops/cobb_pop_all)
eth.proportions.df %>%
  ggplot(aes(x="", y=Proportion, fill=Ethnicity)) +
  geom_bar(stat="identity", width=.5) +
  theme_light() + coord_flip() +
  theme(legend.position = "bottom", legend.text = element_text(size=12)) +
  scale_fill_brewer(palette="Pastel1") + labs(x=NULL)
```




```
ggsave("Ethnicity.png", height=5, width=10)
```