Covid Time Series

Cleaning the data set (which it desperately needs)

```
cnd <- read.csv("United_States_COVID-19_Cases_and_Deaths_by_State_over_Time.csv", header = T)
clt <- read.csv("United_States_COVID-19_County_Level_of_Community_Transmission_as_Originally_Posted.csv

names(clt)[1] <- "state_name"

# Check for duplicated observations
sum(duplicated(cnd))
```

```
## [1] 0
```

```
sum(duplicated(clt))
```

```
## [1] 0
```

```
# Check for empty rows
sum(apply(cnd, 1, function(x) all(is.na(x))))
```

```
## [1] 0
```

```
sum(apply(clt, 1, function(x) all(is.na(x))))
```

```
## [1] 0
```

```
# Dimensions of Data Sets
dim(cnd) # Dim of cases and deaths by states
```

```
## [1] 53400    15
```

```
dim(clt) # Dim of county level transmission
```

```
## [1] 1021114       7
```

```
# Summary statistics of data sets
summary(cnd)
```

```
##  ï..submission_date    state            tot_cases          conf_cases
##  Length:53400       Length:53400       Length:53400       Length:53400
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##   prob_cases          new_case          pnew_case          tot_death
##  Length:53400       Length:53400       Length:53400       Length:53400
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##   conf_death          prob_death         new_death          pnew_death
##  Length:53400       Length:53400       Length:53400       Length:53400
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##   created_at         consent_cases      consent_deaths
##  Length:53400       Length:53400       Length:53400
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
```

```
summary(clt)
```

```
##   state_name         county_name          fips_code     report_date
##  Length:1021114     Length:1021114     Min.   : 1001   Length:1021114
##  Class :character   Class :character   1st Qu.:19031   Class :character
##  Mode  :character   Mode  :character   Median :30023   Mode  :character
##                                        Mean   :31383
##                                        3rd Qu.:46105
##                                        Max.   :72153
##
##  cases_per_100K_7_day_count_change
##  Length:1021114
##  Class :character
##  Mode  :character
##
##
##
##
##  percent_test_results_reported_positive_last_7_days
##  Min.   :  0.00
##  1st Qu.:  4.84
##  Median : 10.26
##  Mean   : 13.11
##  3rd Qu.: 17.92
##  Max.   :100.00
##  NA's   :125582
##  community_transmission_level
##  Length:1021114
##  Class :character
##  Mode  :character
##
##
##
##
```

```r
# Something is wrong. Despite the majority of variables in cnd being numerical by inspecting
# them, they are labeled as character variables. I'll change that quickly.
cnd[, c(3:12)] <- sapply(cnd[, c(3:12)], function(x) as.numeric(gsub(",","",x)))

# Similarly, clt seems to have an issue with one of the variables being primarily being numeric
# but being counted as a character variable due to "suppressed" being a recurring element.
# So, I'll change the instances of "suppressed" to be -1 instead.
clt[, 5] <- gsub("suppressed", -1, clt[, 5])
clt[, 5] <- as.numeric(gsub(",", "", clt[, 5]))

# Finally, we just need the dates to register as the date class
cnd$ï..submission_date <- as.Date(cnd$ï..submission_date, format = "%m/%d/%Y")
clt$report_date <- as.Date(clt$report_date, format = "%Y/%m/%d")
```

EDA

I primarily wanted to do this with Los Angeles and California in mind but I saw the amount of counties in the county transmission data set (clt) with "suppressed" in their cases variable and thought I would do a generalized exploratory data analysis just to see what is going on there. Then it's back to what I was originally going to do.

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

```r
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 4.1.3
```

```
##
## Attaching package: 'lubridate'
```

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

```r
suppressed <- clt %>%
  select(state_name, county_name, community_transmission_level,
         cases_per_100K_7_day_count_change) %>%
  filter(cases_per_100K_7_day_count_change == -1)
head(suppressed, 10)
```

```
##      state_name      county_name community_transmission_level
## 1      Michigan Ontonagon County                  substantial
## 2     Minnesota      Lake County                  substantial
## 3       Montana  Chouteau County                  substantial
## 4          Iowa    Monona County                  substantial
## 5       Montana  Big Horn County                     moderate
## 6  North Dakota    Oliver County                         high
## 7  North Dakota     Burke County                     moderate
## 8      Colorado   Crowley County                     moderate
## 9       Montana    Wibaux County                         high
## 10     Oklahoma     Grant County                         high
##    cases_per_100K_7_day_count_change
## 1                                 -1
## 2                                 -1
## 3                                 -1
## 4                                 -1
## 5                                 -1
## 6                                 -1
## 7                                 -1
## 8                                 -1
## 9                                 -1
## 10                                -1
```

```r
d <- which(duplicated(suppressed[, c(1, 2)]) == T)
suppressed <- suppressed[-d, ]
paste(dim(suppressed)[1], "unique counties that have chosen to suppress their own case counts.")
```

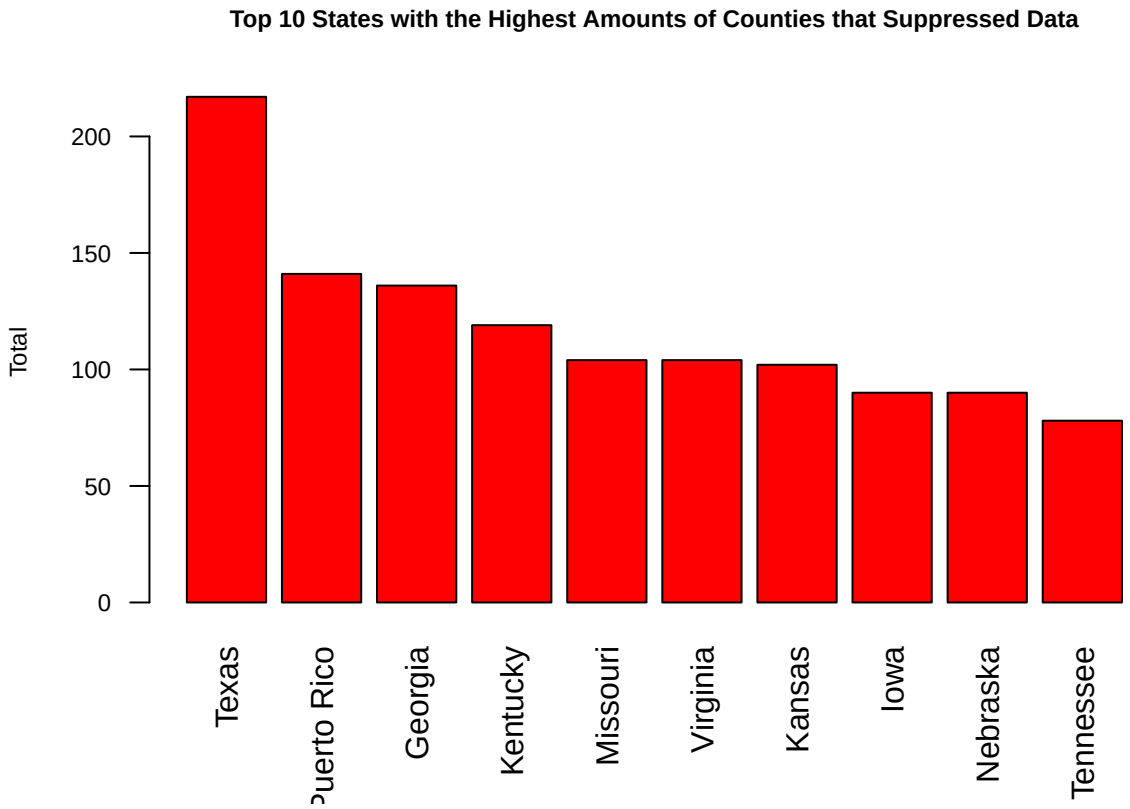
```
## [1] "2545 unique counties that have chosen to suppress their own case counts."
```

```r
table_states <- sort(table(suppressed$state_name), decreasing = T) # Amount of counties suppressing dat
table_states
```

```
##
##          Texas    Puerto Rico        Georgia       Kentucky       Missouri
##            217            141            136            119            104
##       Virginia         Kansas           Iowa       Nebraska      Tennessee
##            104            102             90             90             78
##        Indiana    Mississippi North Carolina       Illinois       Oklahoma
##             76             76             76             70             69
##       Arkansas   South Dakota      Minnesota        Alabama       Michigan
##             68             64             63             58             56
##           Ohio      Louisiana        Montana   North Dakota       Colorado
##             56             55             55             51             49
##  West Virginia          Idaho      Wisconsin South Carolina        Florida
##             44             42             41             32             31
##     California     New Mexico   Pennsylvania     Washington        Wyoming
```

```
##             26             25             25             25             22
##         Alaska         Oregon           Utah         Nevada       Maryland
##             21             21             16             14             10
##       New York        Arizona  New Hampshire        Vermont          Maine
##              7              4              4              4              3
##         Hawaii  Massachusetts     New Jersey
##              2              2              1
```

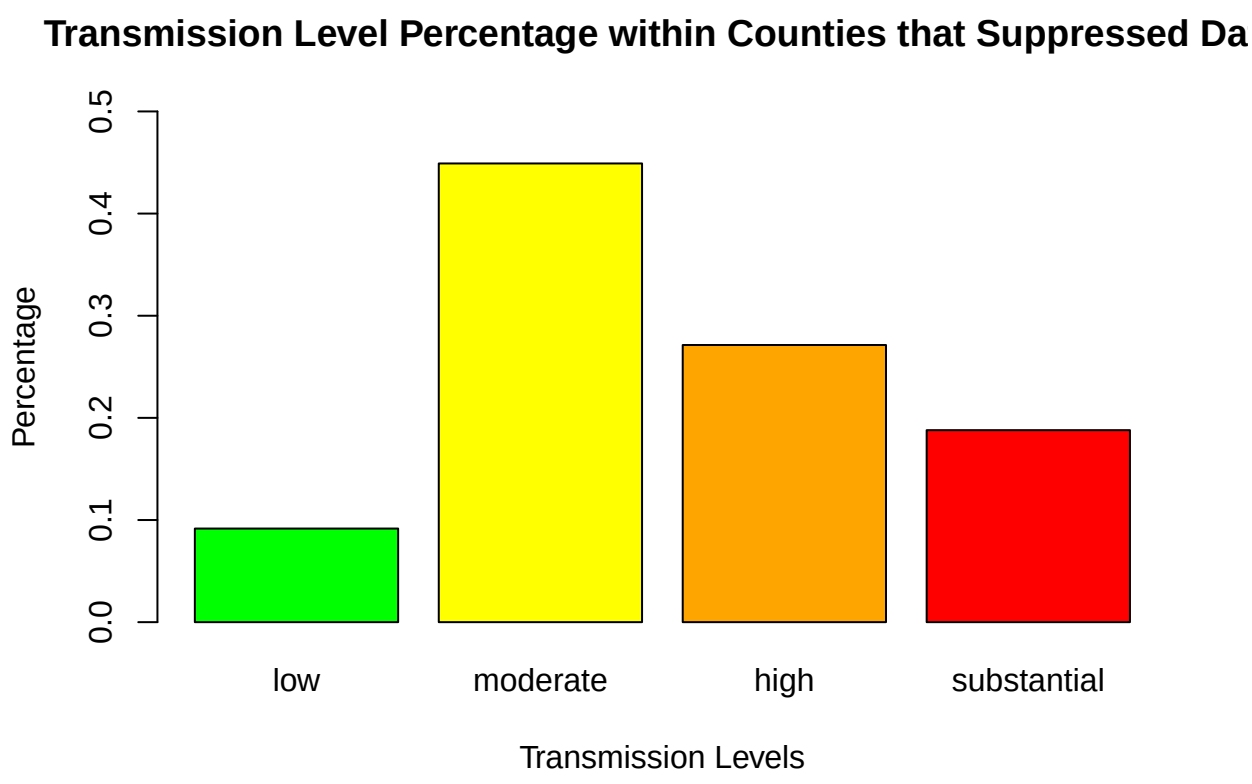
```
barplot(head(table_states, 10), col = "red", las = 2, cex.main = 0.75, cex.lab = 0.75, cex.axis = 0.75,
        main = "Top 10 States with the Highest Amounts of Counties that Suppressed Data", ylab = "Total")
```




```
table_tlevel <- prop.table(table(suppressed$community_transmission_level)[-1])
table_tlevel <- table_tlevel[c(2, 3, 1, 4)]
table_tlevel
```

```
##
##         low    moderate        high substantial
##  0.09162407  0.44907589  0.27133307  0.18796697
```

```
barplot(table_tlevel, col = c("green", "yellow", "orange", "red"), ylim = c(0, 0.5),
        xlab = "Transmission Levels", ylab = "Percentage",
        main = "Transmission Level Percentage within Counties that Suppressed Data")
```




```
# Now we move onto the actual work
cali_clt <- clt %>%
  filter(state_name == "California")
cali_cnd <- cnd %>%
  filter(state == "CA")
cali_cnd[, 6] <- gsub("-", "", cali_cnd[, 6])
cali_cnd <- cali_cnd %>% arrange(ï..submission_date)
head(cali_cnd, 10)
```

```
##    ï..submission_date state tot_cases conf_cases prob_cases new_case pnew_case
## 1          2020-01-22    CA         0          0          0        0         0
## 2          2020-01-23    CA         0          0          0        0         0
## 3          2020-01-24    CA         0          0          0        0         0
## 4          2020-01-25    CA         0          0          0        0         0
## 5          2020-01-26    CA         0          0          0        0         0
## 6          2020-01-27    CA         0          0          0        0         0
## 7          2020-01-28    CA         0          0          0        0         0
## 8          2020-01-29    CA         0          0          0        0         0
## 9          2020-01-30    CA         0          0          0        0         0
## 10         2020-01-31    CA         0          0          0        0         0
##    tot_death conf_death prob_death new_death pnew_death             created_at
## 1          0         NA         NA         0          0 01/24/2020 12:00:00 AM
## 2          0         NA         NA         0          0 01/25/2020 12:00:00 AM
## 3          0         NA         NA         0          0 01/26/2020 12:00:00 AM
## 4          0         NA         NA         0          0 01/27/2020 12:00:00 AM
## 5          0         NA         NA         0          0 01/28/2020 12:00:00 AM
```

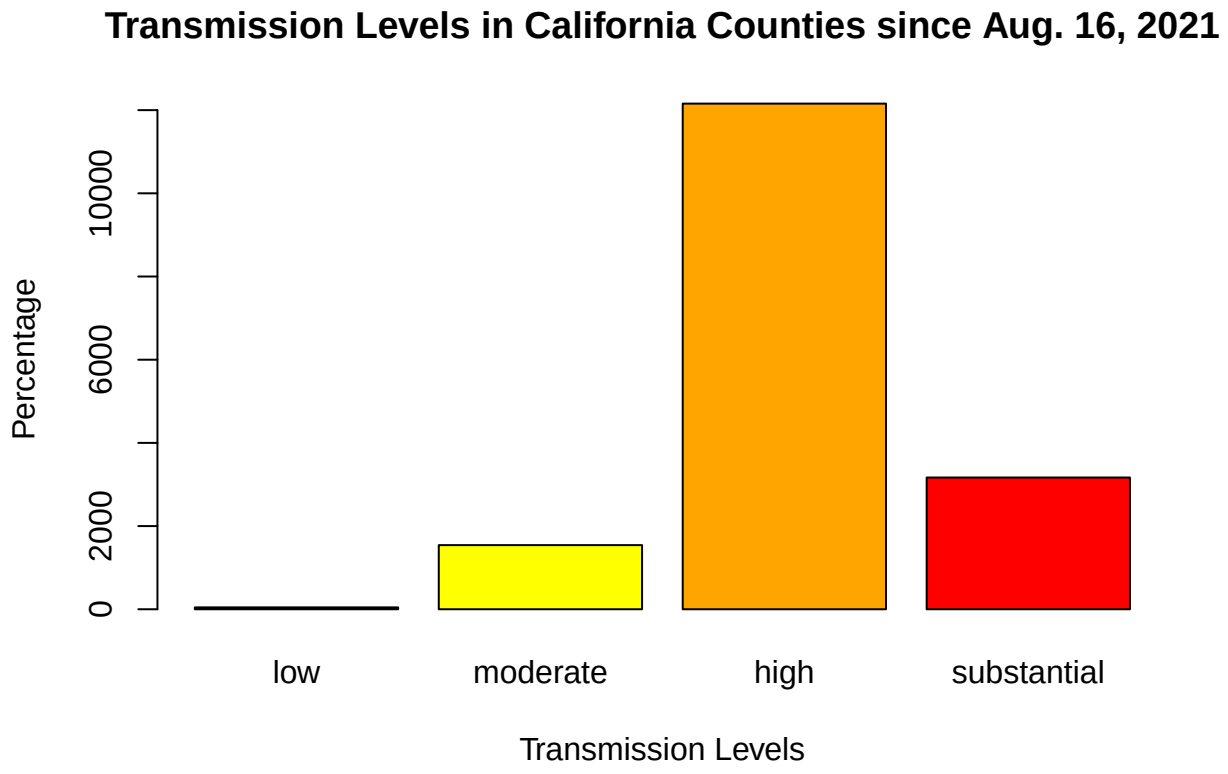
```
## 6          0         NA         NA         0          0 01/29/2020 12:00:00 AM
## 7          0         NA         NA         0          0 01/30/2020 12:00:00 AM
## 8          0         NA         NA         0          0 01/31/2020 12:00:00 AM
## 9          0         NA         NA         0          0 02/01/2020 12:00:00 AM
## 10         0         NA         NA         0          0 02/02/2020 12:00:00 AM
##    consent_cases consent_deaths
## 1          Agree      Not agree
## 2          Agree      Not agree
## 3          Agree      Not agree
## 4          Agree      Not agree
## 5          Agree      Not agree
## 6          Agree      Not agree
## 7          Agree      Not agree
## 8          Agree      Not agree
## 9          Agree      Not agree
## 10         Agree      Not agree
```

```
cali_clt <- cali_clt[-which(cali_clt$cases_per_100K_7_day_count_change <= 0), ]
cali_clt <- cali_clt %>% arrange(report_date)
head(cali_clt, 10)
```

```
##    state_name            county_name fips_code report_date
## 1  California           Kings County      6031  2021-08-16
## 2  California       Del Norte County      6015  2021-08-16
## 3  California           Butte County      6007  2021-08-16
## 4  California        Monterey County      6053  2021-08-16
## 5  California        Humboldt County      6023  2021-08-16
## 6  California         Ventura County      6111  2021-08-16
## 7  California        Tuolumne County      6109  2021-08-16
## 8  California San Luis Obispo County      6079  2021-08-16
## 9  California            Inyo County      6027  2021-08-16
## 10 California          Shasta County      6089  2021-08-16
##    cases_per_100K_7_day_count_change
## 1                             419.77
## 2                             798.22
## 3                             218.08
## 4                             105.05
## 5                             359.99
## 6                             158.04
## 7                             433.20
## 8                             275.51
## 9                              60.98
## 10                            239.34
##    percent_test_results_reported_positive_last_7_days
## 1                                                8.36
## 2                                                  NA
## 3                                               10.05
## 4                                                4.83
## 5                                               11.19
## 6                                                6.96
## 7                                                8.52
## 8                                                7.79
## 9                                                  NA
## 10                                               8.45
```

```
##    community_transmission_level
## 1                          high
## 2                          high
## 3                          high
## 4                          high
## 5                          high
## 6                          high
## 7                          high
## 8                          high
## 9                   substantial
## 10                         high
```

```r
# Due to the strangeness of cali_clt's variable, particularly with
# cases_per_100k_7_day_count_change, we'll only be using this data set to see
# the community transmission levels of counties in California that have recorded their
# cases and see how that looks like.
t_clt <- table(cali_clt$community_transmission_level)[c(2, 3, 1, 4)]
barplot(t_clt, col = c("green", "yellow", "orange", "red"),
        xlab = "Transmission Levels", ylab = "Percentage",
        main = "Transmission Levels in California Counties since Aug. 16, 2021")
```



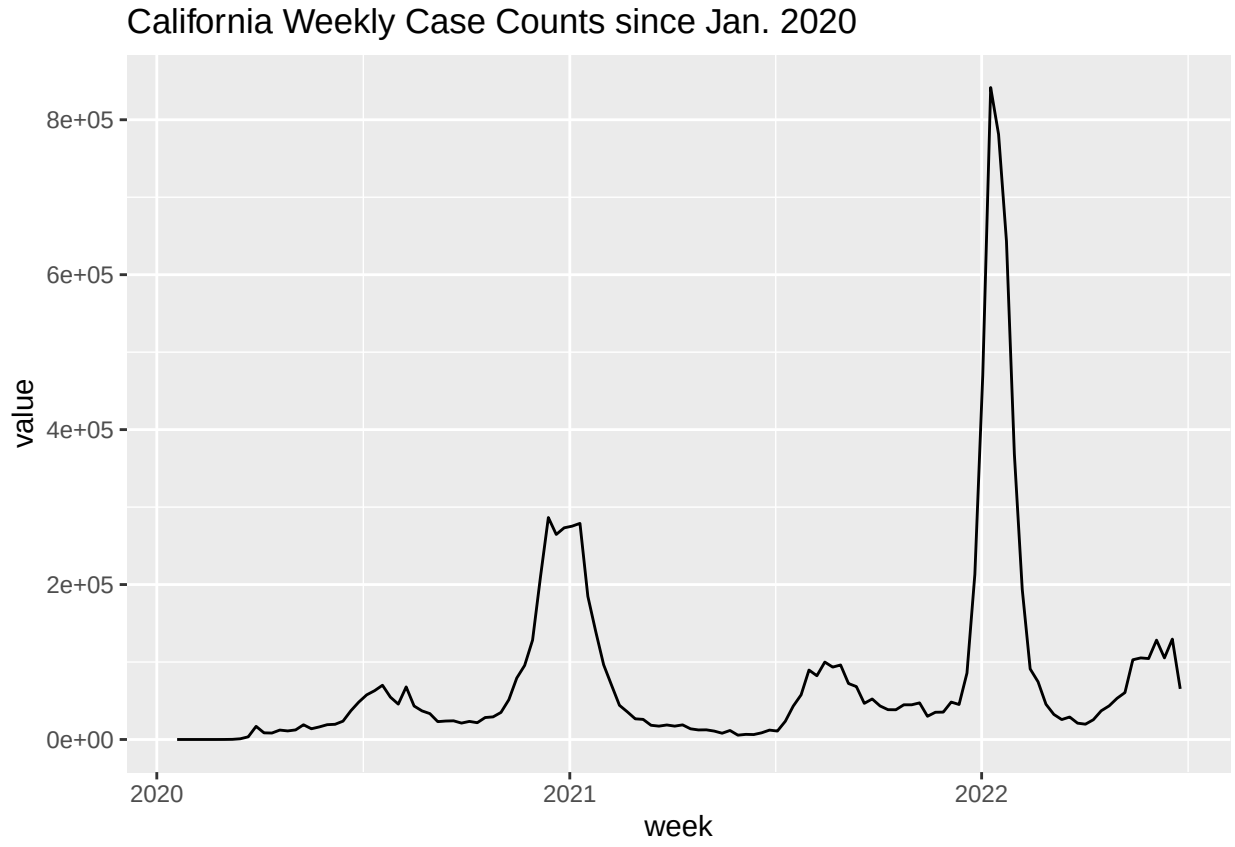
The Forecast

```
library(forecast)
```

```
## Warning: package 'forecast' was built under R version 4.1.3
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
cali_cnd$new_case <- as.numeric(cali_cnd$new_case)
county_weekly <- cali_cnd %>% group_by(week = cut(ï..submission_date, "week", start.on.monday = FALSE))
county_weekly$week <- as.Date(county_weekly$week)

ggplot(data = county_weekly, aes(x = week, y = value)) +
  geom_line() +
  ggtitle("California Weekly Case Counts since Jan. 2020")
```



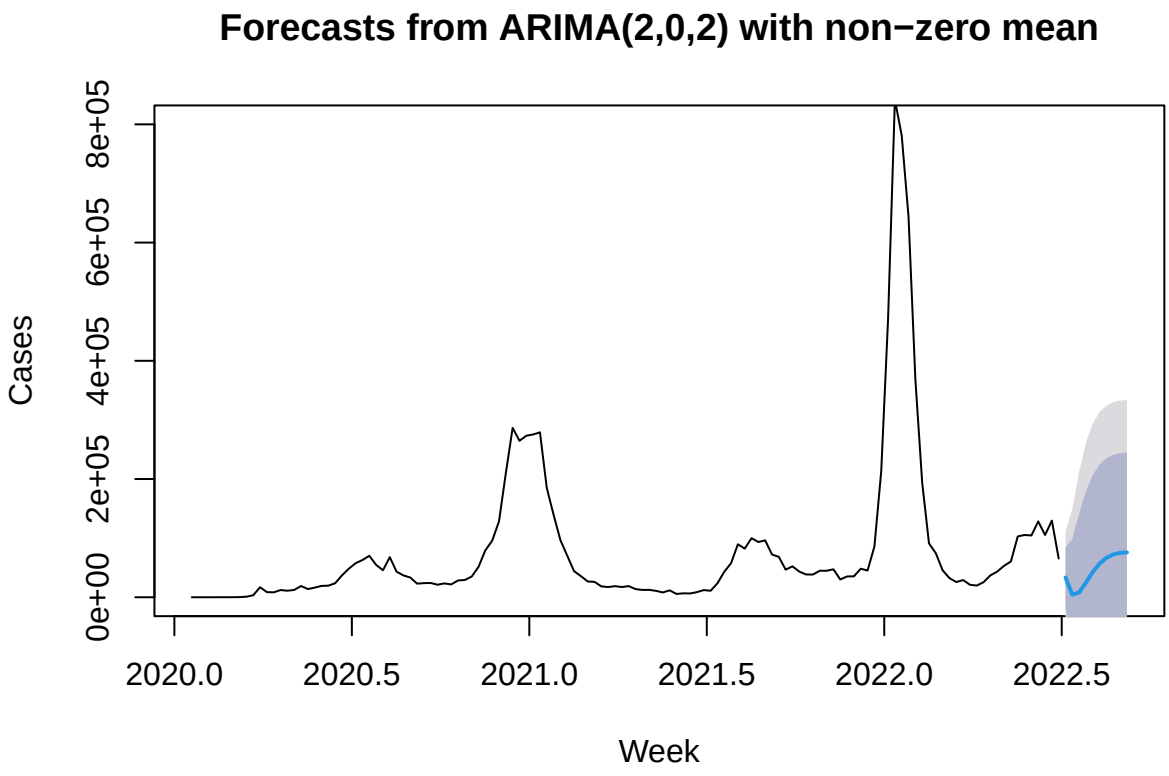

```
ts_weekly_cases <- ts(county_weekly[,2], start = decimal_date(ymd('2020-01-19')),
                      frequency = 52)

# Run ARIMA and create summary.
arima_model <- auto.arima(ts_weekly_cases)
summary(arima_model)
```

```
## Series: ts_weekly_cases
## ARIMA(2,0,2) with non-zero mean
##
## Coefficients:
##          ar1      ar2     ma1     ma2      mean
##       1.1786  -0.4057  0.3842  0.4145  74065.31
## s.e.  0.1419   0.1331  0.1404  0.0980  26832.47
##
## sigma^2 = 1.582e+09:  log likelihood = -1536.34
## AIC=3084.68   AICc=3085.37   BIC=3101.79
##
## Training set error measures:
##                    ME     RMSE      MAE  MPE MAPE     MASE        ACF1
## Training set 247.139 38985.66 18747.92 -Inf  Inf 0.261641 0.007890735
```

```
# Forecast the number of points required
data_forecast <- forecast(arima_model, 10)
print(data_forecast)
```

```
##          Point Forecast      Lo 80     Hi 80      Lo 95    Hi 95
## 2022.511      33311.130  -17656.38  84278.64  -44636.95 111259.2
## 2022.530       4174.278  -90389.62  98738.18 -140448.73 148797.3
## 2022.549       8223.508 -125346.81 141793.83 -196054.67 212501.7
## 2022.568      24815.750 -130301.61 179933.11 -212415.79 262047.3
## 2022.588      42728.914 -121746.77 207204.60 -208814.94 294272.8
## 2022.607      57110.681 -110537.91 224759.27 -199285.71 313507.1
## 2022.626      66794.457 -101623.05 235211.97 -190777.90 324366.8
## 2022.645      72373.690  -96142.89 240890.27 -185350.18 330097.6
## 2022.665      75021.072  -93495.75 243537.90 -182703.17 332745.3
## 2022.684      75878.011  -92650.73 244406.75 -181864.46 333620.5
```

```
# Plot the forecast data.
plot(data_forecast, xlab = 'Week', ylab = 'Cases', ylim = c(0, 8e+05))
```




```
# Looks like there will be a bit more cases in the near future

# So is our forecast accurate? Not really, no forecast is 100% accurate, but we want
# to see if ARIMA can at least capture a trend, and to do that, we'll see if it
# predicts the peak around December 2021 and January 2022.

cw <- county_weekly[-c((dim(county_weekly[1]) - 24):dim(county_weekly)[1]), ]
```

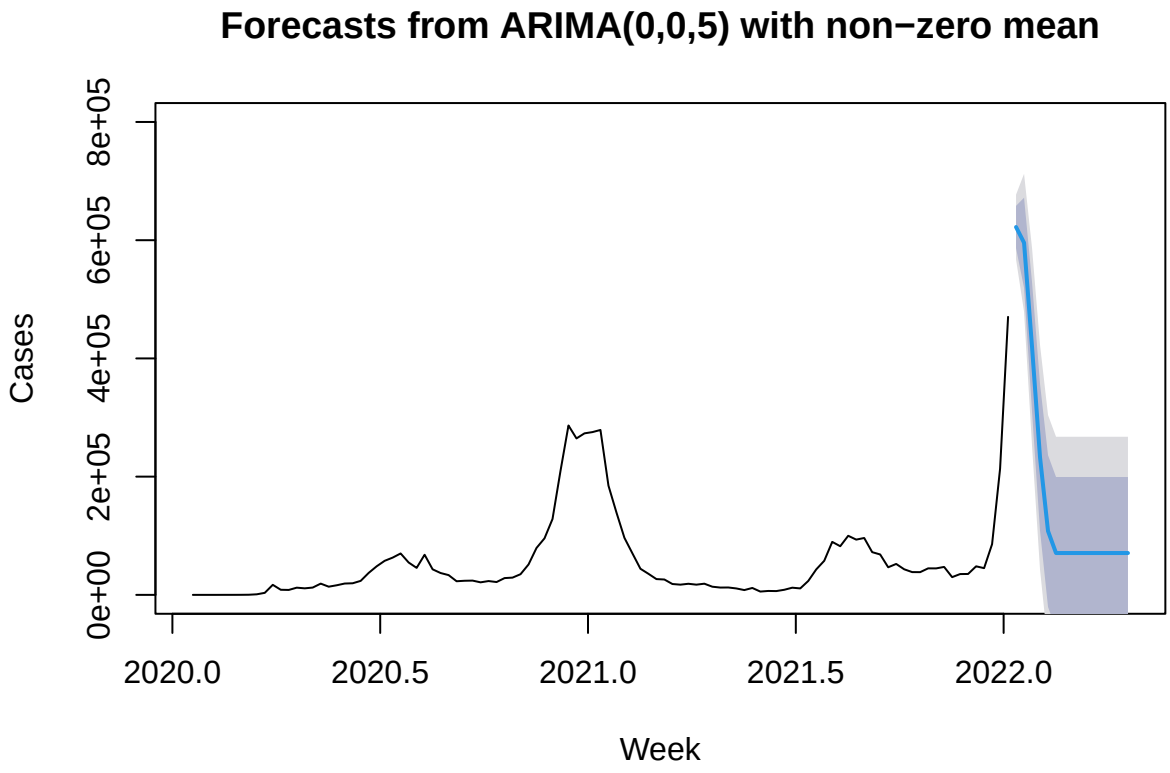
```
## Warning in (dim(county_weekly[1]) - 24):dim(county_weekly)[1]: numerical
## expression has 2 elements: only the first used
```

```
ts_weekly_cases2 <- ts(cw[,2], start = decimal_date(ymd('2020-01-19')),
                       frequency = 52)
arima_model2 <- auto.arima(ts_weekly_cases2)
summary(arima_model2)
```

```
## Series: ts_weekly_cases2
## ARIMA(0,0,5) with non-zero mean
##
## Coefficients:
##          ma1     ma2     ma3     ma4     ma5      mean
##       1.8773  2.1871  1.6453  0.9026  0.2411  70769.80
## s.e.  0.1083  0.1957  0.2290  0.2077  0.1069  20632.93
##
## sigma^2 = 781573086:  log likelihood = -1199.75
```

```
## AIC=2413.51   AICc=2414.69   BIC=2431.95
##
## Training set error measures:
##                     ME     RMSE      MAE  MPE MAPE     MASE     ACF1
## Training set 242.1152 27130.14 16318.45 -Inf  Inf 0.344048 0.080245
```

```
data_forecast2 <- forecast(arima_model2, 15)
plot(data_forecast2, xlab = 'Week', ylab = 'Cases', ylim = c(0, 8e+05))
```




```
# While the values aren't the same, the pattern is correct for the trend we
# see in rising covid cases in January 2022. Like previously mentioned, you
# can never have a model that 100% accurately predict the future but it can sure
# try to pick up the patterns at the very least.
```