Effect of Booster on Weekly New Cases at County Level in the U.S.

Ming Zhao (919024805)

Mar 18, 2022

Introduction

The omicron variant becomes the dominate strain of COVID-19 since December 2021, and it is more infectious than Delta among vaccinated and boosted people (ref1). Plotting the global data from WHO in Figure 1, the new cases significantly increased due to the omicron variant from then on (data1). From Figure 2, we can see that Europe encounters the greatest increases since the end of 2021, and the Americas has the second greatest increases. In the worldwide growth by country, the U.S takes up the most cumulative cases, as shown in Figure 3. By further looking at the U.S. in Figure 4, the growth pattern is similar with the global. As the Figure 5 shows, the data at the state level also show a similar pattern (data2).

As a consequence, I am curious about whether vaccinations, specifically boosters, are still effective against infection in the U.S. This project focuses on the effects of booster rate on weekly new COVID-19 cases at the county level. The goal is to examine whether the percent of people who are fully vaccinated and have received a booster dose is associated with the most recent cases within the last 7 days. Linear Regression and Negative Binomial Regression are the two methods proposed for studying the count data. Additionally, I include demographic features in the models, such as population density, race distribution, metro status. In the following, I first introduce the data I use, explore the data using descriptive statistics, fit models with diagnostics, and finally draw a conclusion.

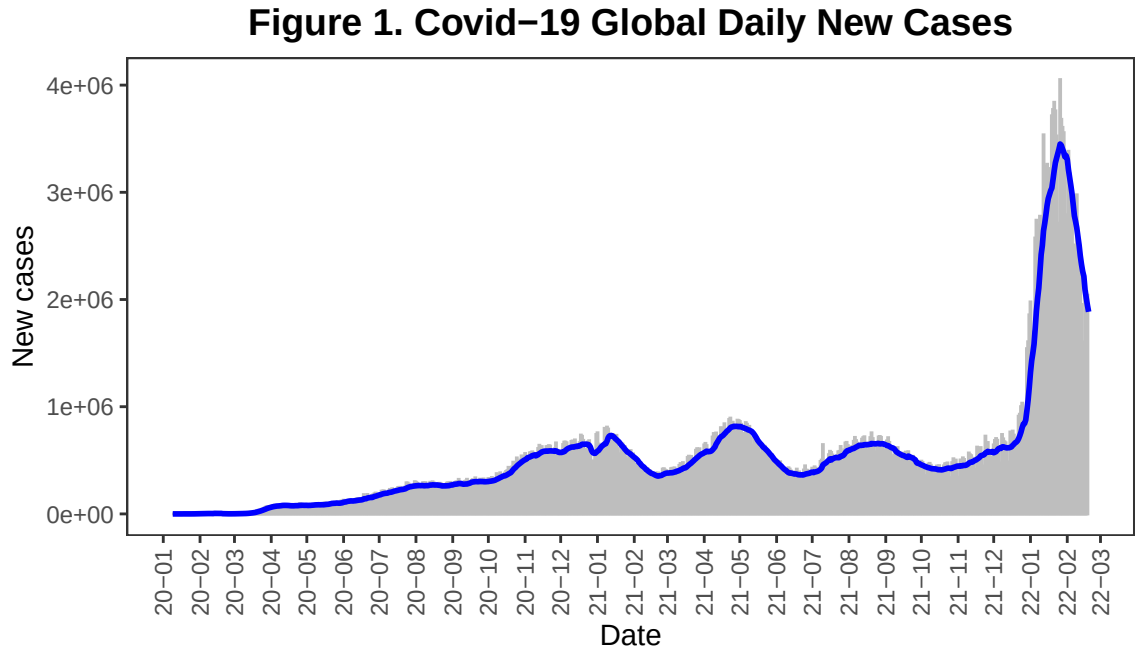


Figure 2. Covid−19 Global Daily New Cases by Region

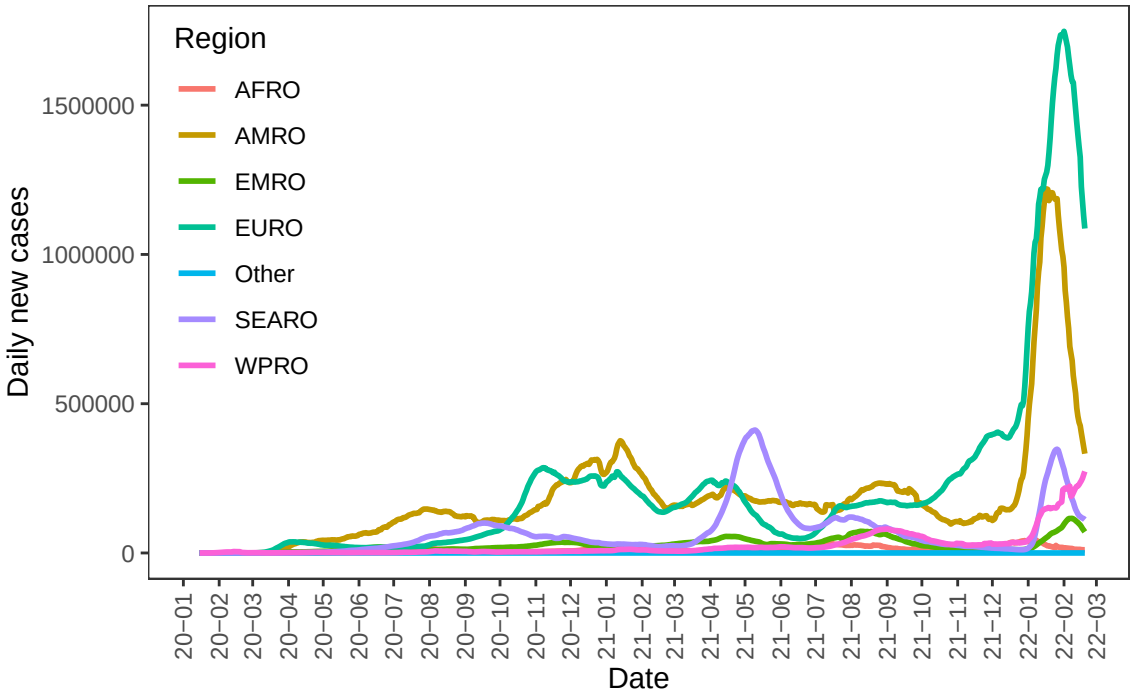


Figure 3. World Map of Cumulative Cases

Total Cases Per 100K People on Feb 19, 2022

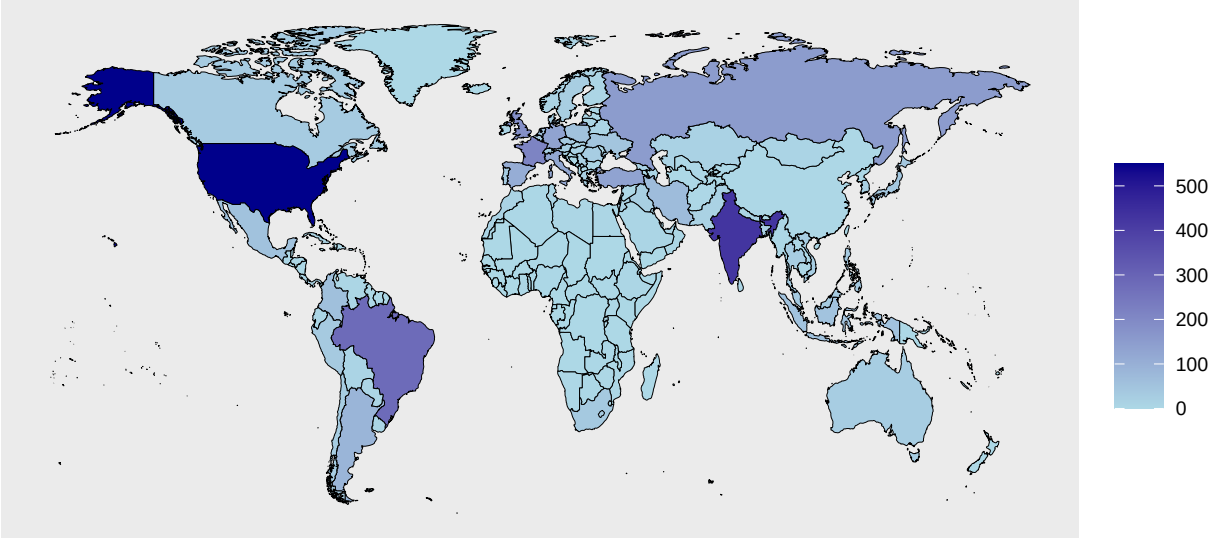


Figure 4. Covid−19 Daily New Cases in USA

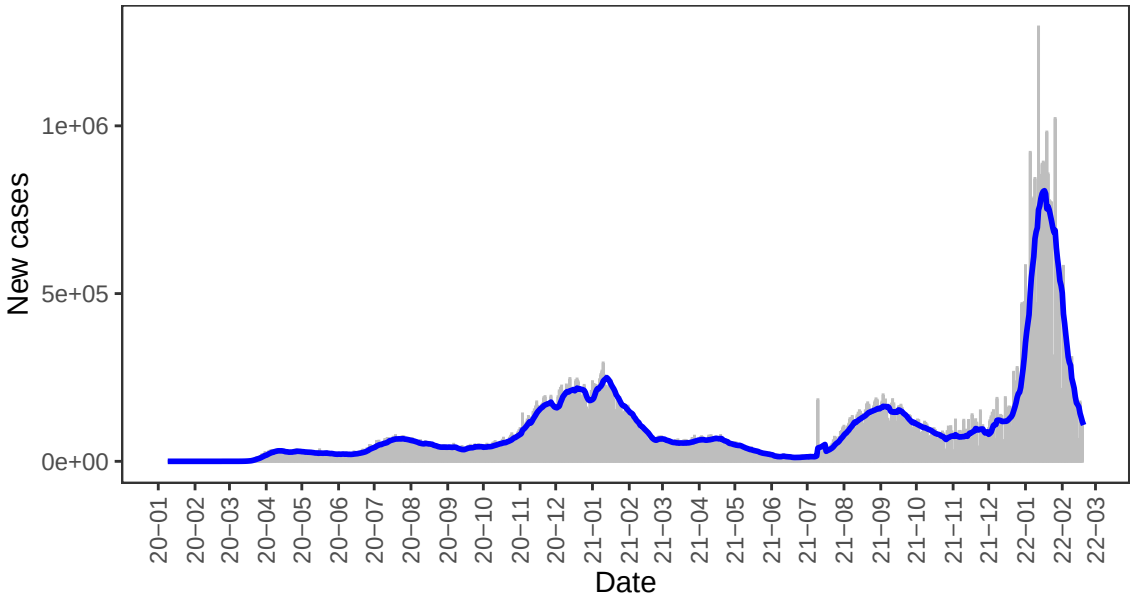
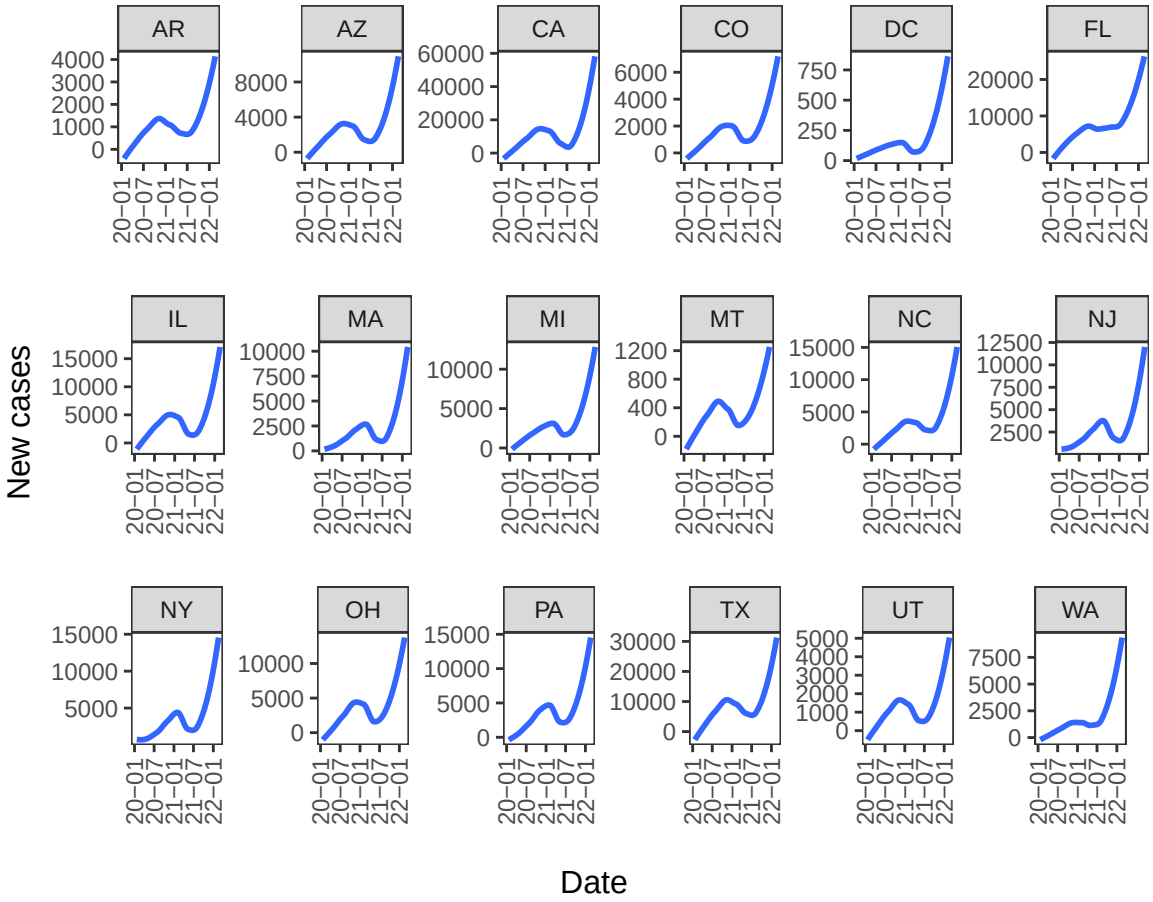


Figure 5. Covid−19 Daily New Cases by State (Selected)



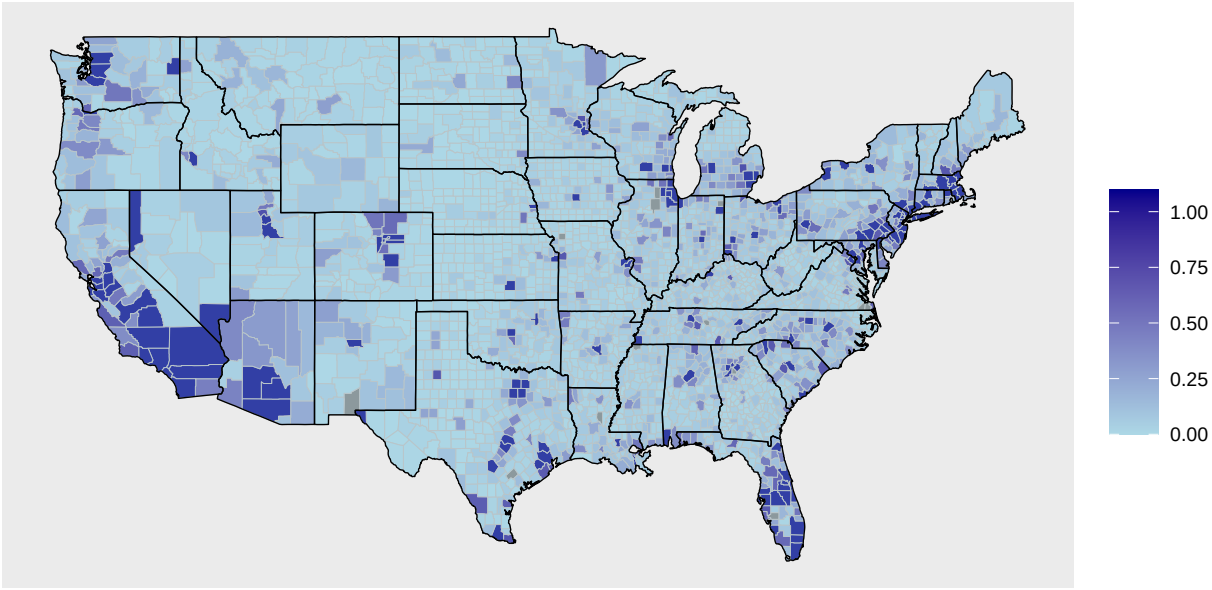
Background

Figure 6 displays the cumulative cases by county and well shows the variation of infection among the counties (data2). In this project, I use the the U.S data at the county level. For the purpose of conducting the analysis, I first combine the case data (data3) and vaccination data (data4) from CDC. Then I collect demographic features about the counties: population density from Census (data5), race distribution, including Black percent, Hispanic percent, and Asian percent from Health Data (data6). State and metro status are also added in the model and CDC provide such information in the vaccination data. Specifically, the outcome variable is weekly new cases defined as the total new cases from "2022-02-12" to "2022-02-19" per 100K people. The variable of interest is booster rate defined as the total number of people who are fully vaccinated and have received a booster shot divided by the total number of people who are fully vaccinated (have second dose of a two-dose vaccine or one dose of a single-dose vaccine), according to CDC (ref2).

The date of booster rate I use in the model is two week before the date of new cases, which is "2022-02-05" because it takes two weeks for booster to be effective (ref3). My hypothesis is that although the omicron variant becomes dominant, the booster still works for its function. According to Hill and Artiga, ethnicity minority is associated with more proportion of COVID-19 cases and deaths, so my hypothesis is that the more percent of Black, Hispanic, and Asian a county has, the more new cases the county has. For the population density, based on the study by Wong and Li, population density affects the COVID-19 cumulative cases at the county level to a certain extent, so my hypothesis is that, higher population density is positively associated with more new cases because I assume more contacts indicate more chance of infection, and it is similar for metro status.

Figure 6. USA Map of Cumulative Cases

Total Cases Per 100K People on Feb 19, 2022



Methodology

Descriptive Analysis

After removing all the missing values, the final sample size for analysis is 2761. Table 1 shows summaries of all quantitative variables. The minimum new cases from Feb 12, 2022 to Feb 19, 2022 are 0 for some counties, while the maximum is 5188.68 per 100K people. The mean is 333.95 and the standard deviation is 324.51 with the same unit, per 100K. The lowest booster rate is 0 while the highest is 70%. The mean of booster rate is 40.26% with standard deviation equal to 19.42. For Black, Hispanic, and Asian percent, respectively, their means are 9%, 10%, and 2%, their maximums are 86%, 96%, and 43%, and their minimums are similar. Population density has a large range from 0 to 27819.80 and thus a large standard deviation. Table 2 shows qualitative variables, metro status, and state. There are 2 levels in metro status and non-metro takes up more proportions. In this, the total number of states is 48.

Figure 7 displays the distribution of new cases and its relationship with booster rate, state, and metro status. The count of new cases has a right skewed distribution with a long tail. The scatter plot of news cases and booster rate fails to show any clear trend. For the box plot of state, some states have significantly more new cases than the others; for the boxplot of metro status, it seems that non-metro tends to have more new cases than metro although the difference is not very clear. Figure 8 displays a scatter plot matrix for all quantitative variables. Among all the explanatory variables, no significant correlation is being detected. Looking at the plot diagonal, it is important to note that after taking log transformation on population density, its distribution becomes normal, otherwise the standard deviation is possibly influential.

Table 1: Descriptive Statistics ($n = 2761$)

Quantitative Measures	Min	Mean	SD	Max
New Cases	0.00	333.95	324.51	5188.68
Booster Rate	0.00	40.26	10.42	70.00
Black Percent	0.00	0.09	0.14	0.86
Hispanic Percent	0.01	0.10	0.14	0.96
Asian Percent	0.00	0.02	0.03	0.43
Pop. Density	0.01	114.74	739.43	27819.80

Table 2: Continued Descriptive Statistics ($n = 2761$)

Qualitative Measures	Level	N	%
Metro status	Metro	1098	39.77
	Non-metro	1663	60.23
Selected States	AK	21	0.76
	NH	10	0.36
	UT	18	0.65
	SD	39	1.41
	WA	37	1.34
	TX	204	7.39
	OH	88	3.19
	MO	112	4.06
	MN	86	3.11
	CT	8	0.29

Figure 7. Exploratory Plots (I)

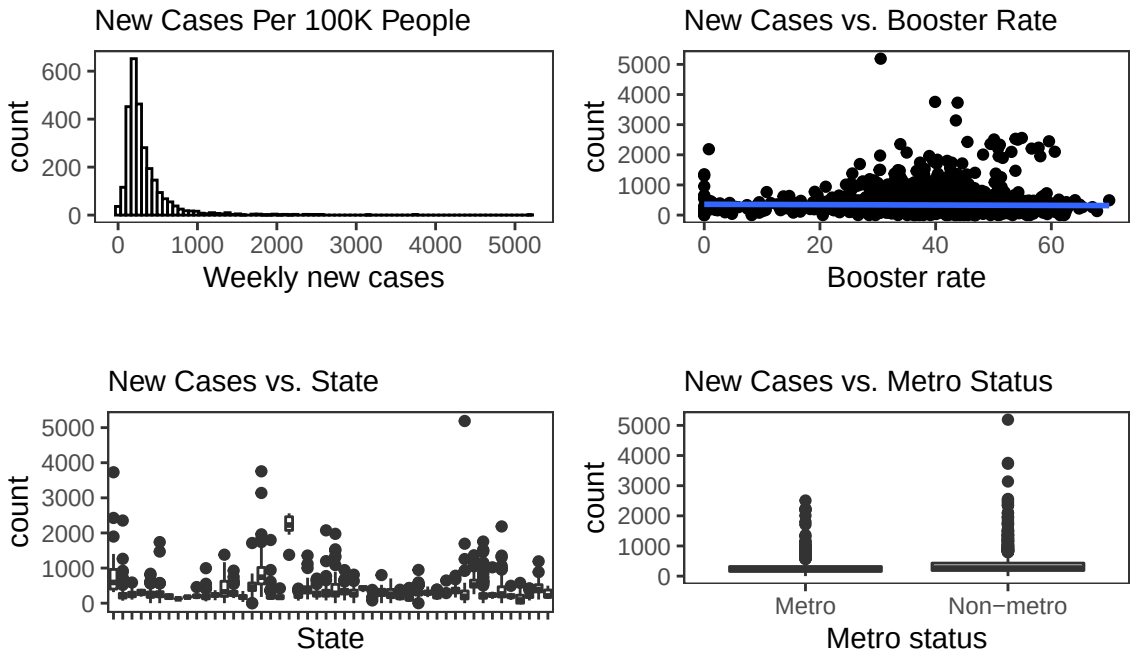
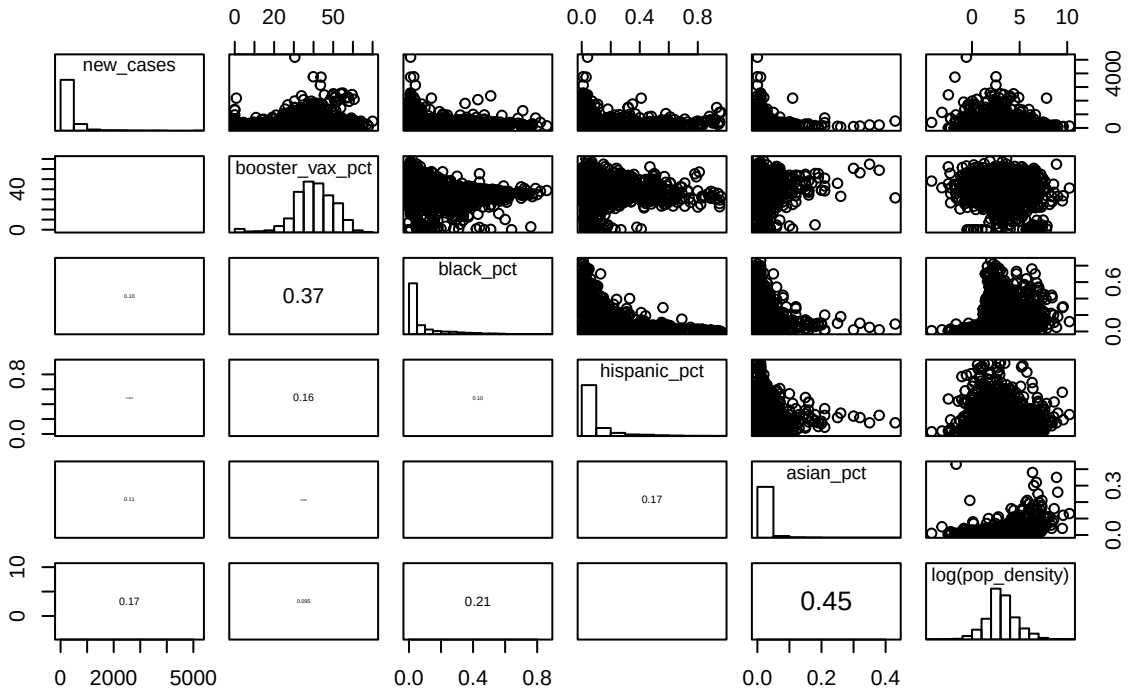


Figure 8. Exploratory Plots (II)



Statistical Models

Linear Regression

The question of interest in this project is to examine whether booster rate is related to weekly new cases. The proposed model for this question is multiple Linear Regression, which defined as:

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i$, i=1,...,n

where Y_i is the value of response variable in the ith case, $X_{i1}, \ldots, X_{ip}$ are the values of the explanatory variables in the ith case, $\beta_0, \beta_1, \ldots, \beta_p$ are regression coefficients, and ε_i are random errors which are i.i.d. $N(0, \sigma^2)$. Here, p is the number of explanatory variables, and n is the sample size.

Specifically, in this project, Y represents weekly new cases, $X_1, \ldots, X_p$ are booster rate, Black percent, Hispanic percent, Asian percent, log-transformed population density, metro status, and state, respectively. For this method, I have two models:

Model (1): $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_7 X_{i7} + \varepsilon_i$, i=1,...,2761

Model (2): $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \beta_7 X_{i7} + \varepsilon_i$, i=1,...,2761

where Model (1) is the reduced model and Model (2) is the full model. First, I test whether X_2, X_3, X_4, X_5, X_6 may be dropped out from the full model. The null hypothesis, H_0 is that all $\beta_2, \beta_3, \beta_4, \beta_5, \beta_6$ are equal to 0 and the alternative is that at least one of them is not 0. Based on the result shown in Table 3, p-value is 0, which is < 0.05, so H_0 is rejected and conclude that the 5 variables should not be dropped from the full model. In addition, I do not test the interaction terms as I assume the effect of booster rate is not varied by state.

Next, I obtain ANOVA table for Model (2). The results in Table 4 show that the effect of booster rate were not significant if no variables were controlled. It is identical to the scatter plot in Figure 6 that there is no trend identified. However, when adding other explanatory variables, booster rate becomes significant.

Table 3: ANOVA Test for Model (1) and (2)

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
2712	162361389				
2707	157704504	5	4656884	15.987	0

Table 4: ANOVA Table for Model (2)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Booster Rate	1	110367.0	110367.04	1.894	0.169
State	47	128173295.1	2727091.38	46.811	0.000
Black Percent	1	914003.4	914003.40	15.689	0.000
Hispanic Percent	1	281061.4	281061.42	4.824	0.028
Asian Percent	1	1314261.2	1314261.16	22.559	0.000
log(Pop. Density)	1	1765715.6	1765715.60	30.309	0.000
Metro Status	1	381842.6	381842.61	6.554	0.011
Residuals	2707	157704504.5	58258.04		

The Table 5 displays the coefficients from Model (1) and Model (2) excluding the states' as there are 48 states and they are not the variables of interest. The primary variable of interest is booster rate. In Model (1), its coefficient is -2.093 and it is significant, controlling for states. In Model (2), with more variables controlled, the coefficient is still significant, but becomes -1.632. We can interpret it as with one unit increase in booster rate, new cases decrease by 1.632 per 100K on average, when all other predictors are help constant. For other variables, Black percent and log-transformed population density are negatively associated with new cases, Hispanic percent is negatively associated with the outcome, and Asian percent is not statistically significant. Compared to metro, non-metro is related to more new cases.

Table 5: Coefficient Estimates for Model (2)

Predictor	Coefficient	SE	p-value
Intercept	947.064	63.710	<0.001
Booster Rate	-1.632	0.747	0.029
Black Percent	-136.986	48.219	0.005
Hispanic Percent	160.975	50.894	0.002
Asian Percent	-290.385	226.111	0.199
log(Pop. Density)	-16.876	4.988	<0.001
Non Metro	30.992	12.106	0.011

Looking at the diagnostic plots for Model (2) in Figure 9, the violation of model assumptions is observed. First, from the residuals vs fitted plot, the relationship between outcome and predictors is not linear, and the residuals are not strictly constant. From the Q-Q plot, the normality of residuals are also violated. For these concerns, the results no longer hold based on the parametric test. I then perform a non-parametric permutation test for Model (2), which requires no distribution assumptions. The permutation test shows a consistent result in Table 6 that booster rate and other variables except Asian percent are significant. Therefore, the violation of normality can be reasonably ignored as the permutation test gives a consistent result. However, the violation of linearity and unequal variance is still a problem.

Figure 9, Diagnostic Plots for Model (2)

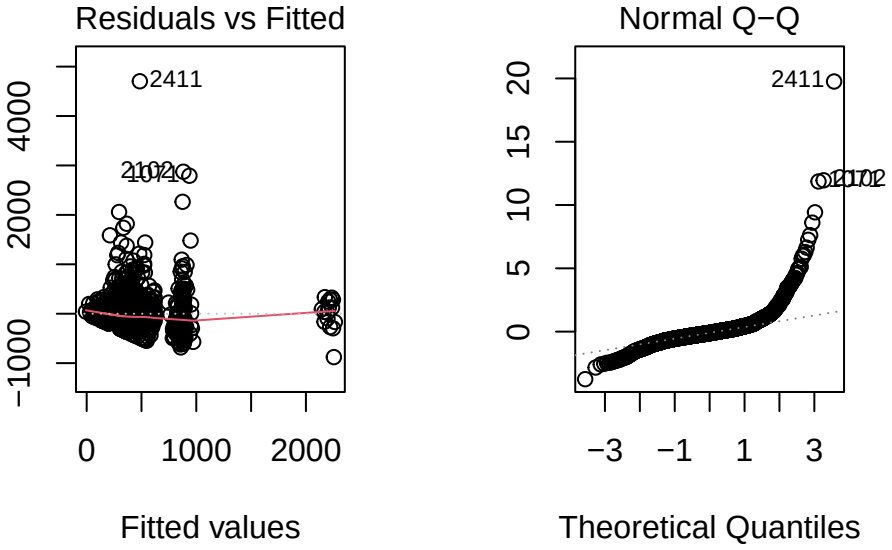


Table 6: Permutation Test for Model (2)

	Df	R Sum Sq	R Mean Sq	Iter	Pr(Prob)
State	47	118994173.17	2531790.92	5000	0.000
Black Percent	1	470184.44	470184.44	5000	0.000
Hispanic Percent	1	582815.83	582815.83	5000	0.000
Asian Percent	1	96085.69	96085.69	5000	0.000
log(Pop. Density)	1	666911.97	666911.97	5000	0.000
Metro Status	1	381842.61	381842.61	4657	0.021
Booster Rate	1	278076.21	278076.21	5000	0.000
Residuals	2707	157704504.46	58258.04		

Negative Binomial Regression

Poisson and Negative Binomial regression are two types of regression models for discrete count data. If the variance and mean are equal, Poisson regression is appropriate; if the variance is significantly larger than the mean, Negative Binomial regression is appropriate. First, I need to decide which method I should use. The overdispersion test shows that overdispersion is detected in Table 7 and therefore suggests Negative Binomial regression.

Table 7: Overdispersion Test

Dispersion ratio	=	143.746
Pearson's Chi-Squared	=	389838.435
P-value	=	< 0.001
		Overdispersion detected

With similar process as for Linear Regression, I first construct the Negative Binomial Regression as below:

$Y_i = exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip})$

I then define Model (3) with booster rate and state and Model (4) with all explanatory variables in the data set. Specifically, they are:

Model (3): $Y_i = exp(\beta_0 + \beta_1 X_{i1} + \beta_7 X_{i7})$

Model (4): $Y_i = exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \beta_7 X_{i7})$

As Table 8 shows, the full model is better than the reduced model as the p-value is 0. ANOVA table for the full model, Model (4), in Table 9, shows that without controlling for other variables, booster rate is only marginally significant. Table 10 displays the coefficients of Negative Binomial Regression. The overall results are consistent that Booster rate, Black percent, and log-transformed population density are negatively related to the new cases, Hispanic percent is positively, and Asian percent is not significantly. Also, non-metro is associated with more new cases than metro. Specifically, I am interested in the relationship between new cases and booster rate. The coefficient of booster rate is -0.003, so we can interpret it as the percent change in the incident rate of new cases is a 0.3% decrease for every unit increase in booster rate.

Table 8: ANOVA Test for Model (3) and (4)

Resid. df	2 x log-lik.	Test	df	LR stat.	Pr(Chi)
2712	-35591.11				
2707	-35421.23	1 vs 2	5	169.8849	0

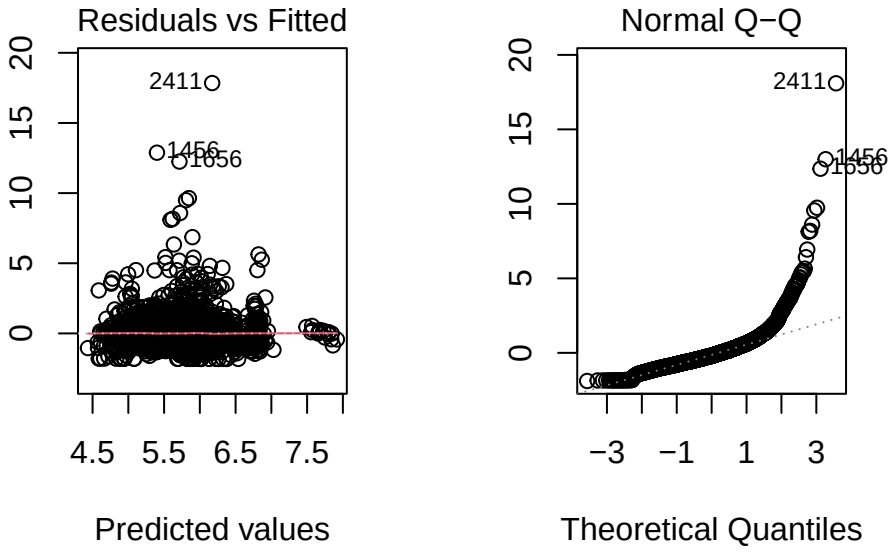
Table 9: ANOVA Table for Model (4)

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			2760	5373.251	
Booster Rate	1	3.312	2759	5369.939	0.069
State	47	2134.524	2712	3235.415	0.000
Black Percent	1	25.303	2711	3210.112	0.000
Hispanic Percent	1	17.759	2710	3192.353	0.000
Asian Percent	1	43.626	2709	3148.728	0.000
log(Pop. Density)	1	81.127	2708	3067.600	0.000
Metro Status	1	7.288	2707	3060.312	0.007

Table 10: Coefficient Estimates for Model (4)

Predictor	Coefficient	SE	p-value
Intercept	6.805	0.145	<0.001
Booster Rate	-0.003	0.002	0.041
Black Percent	-0.394	0.110	<0.001
Hispanic Percent	0.598	0.116	<0.001
Asian Percent	-0.574	0.518	0.268
log(Pop. Density)	-0.073	0.011	<0.001
Non Metro	0.075	0.028	0.007

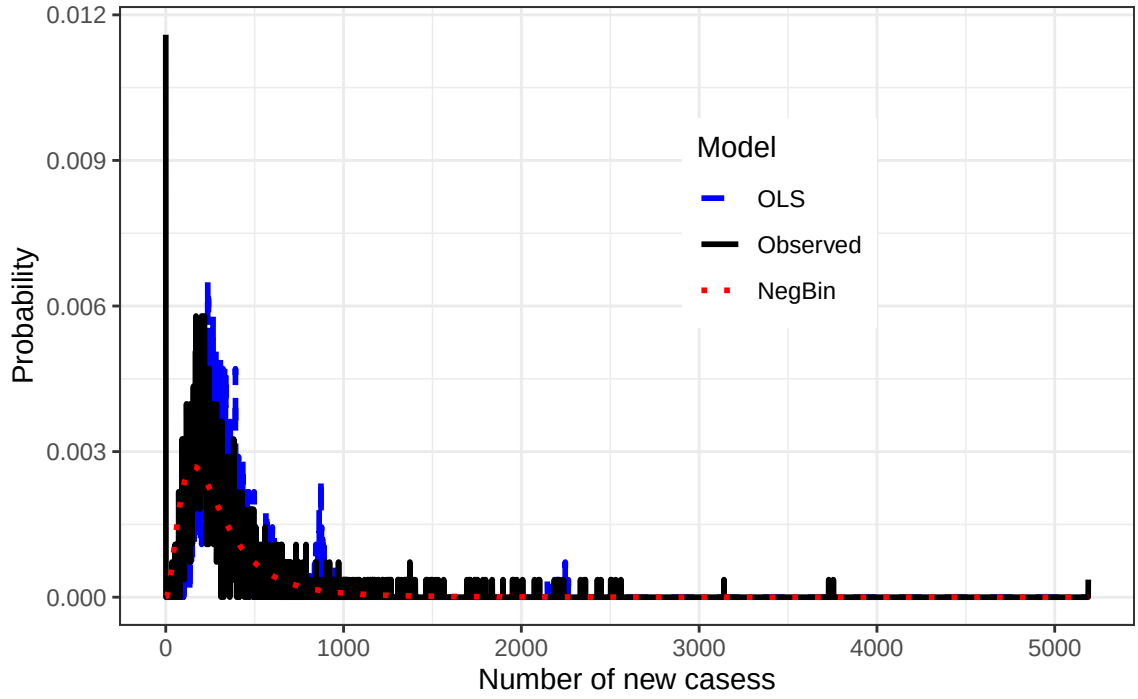
Figure 10. Diagnostic Plots for Model (4)



Discussion and Conclusion

For assistance in understanding the Linear Regression and Negative Binomial Regression models, we can look at the probabilities of predicted new cases in contrast to those of observed new cases in Figure 11. The observed values and predicted values by Linear Regression have more fluctuations, while the predicted values by Negative Binomial Regression are smoother. Compared to Linear Regression, Negative Binomial Regression is more conservative and its predictions are closer to the overall pattern of observation. Thus, Negative Binomial Regression is preferable. Nevertheless, both methods show that there is a significant negative effect of booster rate on weekly new cases at the county level, when holding state, Black percent, Hispanic percent, log-transformed population density, and metro status constant. That is, the booster is effective against COVID-19 even for the omicron variant, but its effect size is assumed to be reduced due to the more infectious variant. Research on this topic can be conducted in the future.

Figure 11. Comparison of Models



Reference

data1 data2 data3 data4 data5 data6 ref1 ref2 ref3 ref4 ref5 ref6

Appendix

```
knitr::opts_chunk$set(warning = FALSE, message = FALSE)
options(tinytex.verbose = TRUE)

library(tidyverse)
library(dplyr)
library(ggplot2)
library(zoo)
library(lubridate)
library(tidyquant)
library(grkmisc)
library(maps)
library(usmap)
library(gridExtra)
library(ggpubr)
library(lme4)
library(car)
library(lmerTest)
library(lmPerm)
library(MASS)
library(performance)
library(knitr)
library(kableExtra)
library(tidyverse)
library(broom)
library(pscl)
library(reshape2)


# Data Loading

setwd("/Users/mingzhao/Desktop/Covid-19")

covid <- read_csv("https://covid19.who.int/WHO-COVID-19-global-data.csv")
covid_states <- read_csv("COVID-19_Cases_State.csv")
vaccine <- read_csv("COVID-19_Vaccinations.csv")
cases <- read_csv("COVID-19_Cases.csv")
cum_cases <- read_csv("COVID-19_Cumulatives.csv")
density <- read_csv("Density.csv")
income <- read_csv("Income.csv")
demographics <- read_csv("Demographics.csv")


# vaccine data
glimpse(vaccine)

## convert date format
vaccine$Date <- as.Date(vaccine$Date, format = "%m/%d/%Y")
class(vaccine$Date)

## rename variables
names(vaccine)[c(2:5)] <- c("fips", "week", "county", "state")
```

```r
## filter data
sum(vaccine$fips=="UNK")
vaccine <- subset(vaccine, fips != "UNK")

## generate longitudinal data
vaccines <- subset(vaccine, Date == "2022-02-05",
                   select = c(Date, fips, week, county, state, Metro_status,
                              Census2019, Series_Complete_Pop_Pct,
                              Booster_Doses,Booster_Doses_Vax_Pct,
                              Booster_Doses_12Plus, Booster_Doses_18Plus,
                              Booster_Doses_50Plus, Booster_Doses_65Plus,
                              Booster_Doses_12Plus_Vax_Pct, Booster_Doses_18Plus_Vax_Pct,
                              Booster_Doses_50Plus_Vax_Pct, Booster_Doses_65Plus_Vax_Pct))


# case data
glimpse(cases)

## convert date format
cases$date <- as.Date(cases$date, format = "%m/%d/%Y")
class(cases$date)

## filter data
cases <- subset(cases, date == "2022-02-19")

## create new variables
cases$cases_per_100K_7_day_count_change <- gsub("\\,", "", as.character(cases$cases_per_100K_7_day_coun
cases$cases_per_100K_7_day_count_change <- as.numeric(cases$cases_per_100K_7_day_count_change)
cases$fips <- cases$fips_code

## check outliers
cases[order(cases$cases_per_100K_7_day_count_change, decreasing =TRUE)[c(1:10)],] #fips=46041, 46137, 2
vaccines[vaccines$fips == "46041",]$Census2019  #5892
vaccines[vaccines$fips == "46137",]$Census2019  #2756
vaccines[vaccines$fips == "21129",]$Census2019  #7403
vaccines[vaccines$fips == "02180",]$Census2019  #10004
vaccines[vaccines$fips == "21063",]$Census2019  #7517

## filter data
cases <- subset(cases, select = c("fips", "cases_per_100K_7_day_count_change"))

sum(is.na(cases$cases_per_100K_7_day_count_change))
dim(cases)


# demographic data

density$fips <- density$GEOID
demo1 <- subset(density, select = c(fips, B25010_001E, B01001_calc_PopDensity))
names(demo1)[c(2:3)] <- c("hh_size", "pop_density")  #average household size, Population Density (peopl

glimpse(income)
income$fips <- income$FIPS_Code
```

```r
demo2 <- subset(income, select = c(fips, Median_Household_Income_2019))
names(demo2)[2] <- c("hh_income")

demographics$fips <- sprintf("%05d", demographics$`FIPS code`)
demo3 <- subset(demographics, select = c("fips", "% In Poverty", "% Over Age 65", "% Non-Hispanic Black
                                         "% Hispanic", "% Non-Hispanic Native American / Alaskan Native
                                         "% Non-Hispanic Asian"))
names(demo3)[c(2:7)] <- c("poverty_pct", "over_65_pct", "black_pct", "hispanic_pct", "native_pct", "asi


# final data

data0 <- vaccines %>% left_join(demo1, by=c("fips"))
data0 <- data0 %>% left_join(demo2, by=c("fips"))
data0 <- data0 %>% left_join(demo3, by=c("fips"))

data <- data0 %>% left_join(cases, by=c("fips"))

data$booster_pop_pct_2 <- data$Booster_Doses/data$Census2019
data$complete_pop_pct <- data$Series_Complete_Pop_Pct
data$booster_vax_pct <- data$Booster_Doses_Vax_Pct

df <- subset(data, select = c(cases_per_100K_7_day_count_change,
                              state,
                              booster_pop_pct_2,
                              complete_pop_pct, booster_vax_pct,
                              hh_size, pop_density, hh_income, poverty_pct, over_65_pct,
                              black_pct, hispanic_pct, native_pct, asian_pct, Metro_status))
names(df)[1] <- c("new_cases")
dim(df) #3224

df_reg <- subset(df, select = c(new_cases, booster_vax_pct,
                                black_pct, hispanic_pct, asian_pct, pop_density, Metro_status, state))

df_reg <- df_reg[complete.cases(df_reg),]
dim(df_reg) #2762

df_reg$log_new_cases <- log(df_reg$new_cases + 1)


# WHO data
glimpse(covid)

## new cases
covid.daily <- covid %>%
                    group_by(Date_reported) %>%
                    summarise(daily_newcases = sum(New_cases),
                              daily_cumcases = sum(Cumulative_cases))
covid.daily$avg_daily_newcases <- stats::filter(covid.daily$daily_newcases, rep(1/7, 7), sides = 1)
covid.daily <- covid.daily %>% filter(Date_reported < "2022-02-20")

## new cases by region
covid.region <- covid %>%
```

```
                      group_by(Date_reported, WHO_region) %>%
                      summarize(daily_newcases = sum(New_cases),
                                daily_cumcases = sum(Cumulative_cases))
covid.region$daily_newcases[covid.region$daily_newcases < 0] <- 0

covid.region <- covid.region %>% filter(Date_reported < "2022-02-20")

## cumulative cases by country
covid.today<- covid %>%
  filter(Date_reported == "2022-02-19") %>%
  mutate(Country.Region=Country)

world <- map_data("world")

covid.today$Country.Region[!covid.today$Country.Region %in% world$region]

list <- which(!covid.today$Country.Region %in% world$region)

covid.today$country <- as.character(covid.today$Country.Region)

covid.today$country[list] <-
  c("Antigua", "Bolivia", "Virgin Islands",
    "Brunei", "Cape Verde", "Democratic Republic of the Congo",
    "Ivory Coast", "Curacao", "Czech Republic",
    "North Korea", "Swaziland", "Falkland Islands",
    "Gibraltar", "Holy See", "Iran",
    "Kosovo", "Laos", "Micronesia",
    "Northern Mariana Islands", "Palestine", "Other",
    "South Korea", "Moldova", "Reunion",
    "Russia", "Saint Barthelemy", "Saint Kitts",
    "Saint Vincent", "Syria", "UK",
    "Tokelau", "Trinidad", "Tuvalu",
    "Tanzania", "USA", "Virgin Islands",
    "Venezuela", "Vietnam")

covid.today$Country.Region[!covid.today$country %in% world$region]

world$country <- world$region

worldmap <- left_join(world, covid.today, by="country")

worldmap$Cumulative_cases[is.na(worldmap$Cumulative_cases)] <- 0

summary(covid.today$Cumulative_cases)

table(covid.today[covid.today$Cumulative_cases > 50000000,]$Country)
covid.today$Cumulative_cases[covid.today$Cumulative_cases > 50000000] <- 55000000


worldmap$Cumulative_case[worldmap$Cumulative_case > 50000000] <- 55000000

worldmap$cum_case_100k <- worldmap$Cumulative_case/100000
```

```r
worldmap <- subset(worldmap, group < 25 | group >132)


# USA data
glimpse(covid)
glimpse(covid_states)

## new cases
covid.us <- subset(covid, Country == "United States of America" & Date_reported < "2022-02-20")
covid.us$avg_newcases <- stats::filter(covid.us$New_cases, rep(1/7, 7), sides = 1)

## new cases by state
covid_states$date <- as.Date(covid_states$submission_date, format = "%m/%d/%Y")
class(covid_states$date)
covid.state <- subset(covid_states, date < "2022-02-20",
                      select = c(date, state, new_case) )

covid.state <- covid.state[covid.state$state %in% unique(data$state),]

state <- sort(unique(df_reg$state))

cum_cases <- read_csv("COVID-19_Cumulatives.csv")

## cumulative cases by county
glimpse(cum_cases)

class(cum_cases$date)
cum.cases <- subset(cum_cases, date == "2022-02-19", select = c(fips, state, county, cases))

sum(is.na(cum.cases$cases))
sum(is.na(cum.cases$fips))
sum(is.na(cum.cases$state))
sum(is.na(cum.cases$county))

cum.cases$state <- tolower(cum.cases$state)
cum.cases$county <- tolower(cum.cases$county)

mainstates <- map_data("state")
allcounty <- map_data("county")

glimpse(allcounty)

names(allcounty)[5:6] <- c("state", "county")

unique(allcounty$county[!allcounty$county %in% cum.cases$county])

cum.cases$county <- gsub("\\.", "", as.character(cum.cases$county))

cum.cases$fips[cum.cases$county=="new york city"] <- "99999"
cum.cases$county[cum.cases$county=="new york city"] <- "new york"

cum.cases$county <-recode_if(cum.cases$county, cum.cases$fips == "01049", "dekalb" = "de kalb")
cum.cases$county <-recode_if(cum.cases$county, cum.cases$fips == "17043", "dupage" = "du page")
```

```
cum.cases$county <-recode_if(cum.cases$county, cum.cases$fips == "18091", "laporte" = "la porte")
cum.cases$county <-recode_if(cum.cases$county, cum.cases$fips == "19141", "o'brien" = "obrien")
cum.cases$county <-recode_if(cum.cases$county, cum.cases$fips == "24033", "prince george's" = "prince g
cum.cases$county <-recode_if(cum.cases$county, cum.cases$fips == "24035", "queen anne's" = "queen annes
cum.cases$county <-recode_if(cum.cases$county, cum.cases$fips == "24037", "st mary's" = "st marys")
cum.cases$county <-recode_if(cum.cases$county, cum.cases$fips == "38045", "lamoure" = "la moure")
cum.cases$county <-recode_if(cum.cases$county, cum.cases$fips == "46102", "oglala lakota" = "oglala dak
cum.cases$county <-recode_if(cum.cases$county, cum.cases$fips == "51700", "newport news city" = "newpor
cum.cases$county <-recode_if(cum.cases$county, cum.cases$fips == "51810", "virginia beach city" = "virg

unique(allcounty$county[!allcounty$county %in% cum.cases$county])

all.county <- allcounty %>% left_join(cum.cases, by=c("state", "county"))

summary(cum.cases$cases)
cum.cases$cases[cum.cases$cases > 100000] <- 100000

sum(is.na(all.county$cases))

all.county$cases[all.county$cases > 100000] <- 110000

all.county$cases_100k <- all.county$cases/100000


# World plot 1
ggplot(covid.daily, aes(x=Date_reported, y=daily_newcases)) +
  geom_bar(stat="identity", width=0.8, color = "grey") +
  geom_line(aes(y=avg_daily_newcases), size = 1, color = "blue") +
  theme_classic() +
  theme_bw() +
  labs(title = "Figure 1. Covid-19 Global Daily New Cases", x= "Date", y= "New cases") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1),
        panel.grid.major=element_blank(),
        panel.grid.minor=element_blank(),
        plot.title = element_text(hjust = 0.5, size = 14, face = "bold")) +
  scale_x_date(limits = as.Date(c("2020-01-09", "2022-02-19")), date_breaks = "1 month", date_labels =


# World plot 2
ggplot(covid.region, aes(x=Date_reported, y=daily_newcases, group=WHO_region)) +
  coord_cartesian() +
  geom_ma(ma_fun = SMA, n=7, size=1, aes(color=WHO_region), linetype=1) +
  theme_classic() +
  theme_bw() +
  labs(title = "Figure 2. Covid-19 Global Daily New Cases by Region", x= "Date", y= "Daily new cases",
  theme(plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1),
        panel.grid.major=element_blank(),
        panel.grid.minor=element_blank(),
        legend.position=c(0.1, 0.65)) +
  scale_x_date(limits = as.Date(c("2020-01-09", "2022-02-19")), date_breaks = "1 month", date_labels =
```

```r
# World plot 3
fig.map  <- ggplot() +
  geom_polygon(data = worldmap, aes(x=long, y = lat, group = group, fill=cum_case_100k),
               color="black", size = 0.1) +
  scale_fill_continuous(name="",
                        low = "lightblue",
                        high = "darkblue",
                        limits = c(0,550),
                        breaks = seq(0, 500, 100)) +
  ggtitle("Figure 3. World Map of Cumulative Cases", subtitle="Total Cases Per 100K People on Feb 19, 20
  coord_fixed(1.3) +
  theme(plot.title = element_text(size = 14, face = "bold"),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.title.x = element_blank(),
        axis.title.y = element_blank(),
        axis.text.x = element_blank(),
        axis.text.y = element_blank(),
        axis.ticks.x = element_blank(),
        axis.ticks.y = element_blank(),
        legend.position="right",
        legend.key.size = unit(0.8, "cm"))
fig.map


# US plot 1
ggplot(covid.us, aes(x=Date_reported, y=New_cases)) +
  geom_bar(stat="identity", width=0.1, color = "grey") +
  geom_line(aes(y=avg_newcases), size = 1, color = "blue") +
  theme_classic() +
  theme_bw() +
  labs(title = "Figure 4. Covid-19 Daily New Cases in USA", x= "Date", y= "New cases") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1),
        panel.grid.major=element_blank(),
        panel.grid.minor=element_blank(),
        plot.title = element_text(hjust = 0.5, size = 14, face = "bold")) +
  scale_x_date(limits = as.Date(c("2020-01-09", "2022-02-19")), date_breaks = "1 month", date_labels =


# US plot 2
state1 <-
ggplot(covid.state[covid.state$state %in% state[c(3:6, 8, 10)],],
       aes(x=date, y=new_case)) +
  coord_cartesian() +
  geom_smooth(method="loess", se=F) +
  facet_wrap(~state, scales = "free", ncol = 6) +
  theme_classic() +
  theme_bw() +
  labs(title = "", x= "", y= "") +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 0.5),
        panel.grid.major=element_blank(),
        panel.grid.minor=element_blank(),
```

```
        plot.margin=unit(c(-0.4,0,0,-0.3), "cm")) +
  scale_x_date(limits = as.Date(c("2020-01-09", "2022-02-19")), date_labels = "%y-%m")

state2 <-
  ggplot(covid.state[covid.state$state %in% state[c(14, 19, 21, 25, 26, 29)],],
         aes(x=date, y=new_case)) +
  coord_cartesian() +
  geom_smooth(method="loess", se=F) +
  facet_wrap(~state, scales = "free", ncol = 6) +
  theme_classic() +
  theme_bw() +
  labs(title = "", x= "", y= "") +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 0.5),
        panel.grid.major=element_blank(),
        panel.grid.minor=element_blank(),
        plot.margin=unit(c(-0.4,0,0,-0.3), "cm")) +
  scale_x_date(limits = as.Date(c("2020-01-09", "2022-02-19")), date_labels = "%y-%m")

state3 <-
  ggplot(covid.state[covid.state$state %in% state[c(32, 33, 36, 41, 42, 45)],],
         aes(x=date, y=new_case)) +
  coord_cartesian() +
  geom_smooth(method="loess", se=F) +
  facet_wrap(~state, scales = "free", ncol = 6) +
  theme_classic() +
  theme_bw() +
  labs(title = "", x= "", y= "") +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 0.5),
        panel.grid.major=element_blank(),
        panel.grid.minor=element_blank(),
        plot.margin=unit(c(-0.4,0.3,0,-0.3), "cm")) +
  scale_x_date(limits = as.Date(c("2020-01-09", "2022-02-19")), date_labels = "%y-%m")

title=text_grob("Figure 5. Covid-19 Daily New Cases by State (Selected)", size = 14, face = "bold")
grid.arrange(state1, state2, state3, ncol=1, top = title, left = "New cases", bottom = "Date")


# US plot 3
fig.map  <- ggplot() +
  geom_polygon(data = all.county, aes(x=long, y = lat, group = group, fill=cases_100k),
               color="grey", size = 0.2) +
  geom_polygon( data=mainstates, aes(x=long, y=lat, group=group),
                color="black", fill="lightblue", size = 0.3, alpha = 0.3)+
  scale_fill_continuous(name="",
                        low = "lightblue",
                        high = "darkblue",
                        limits = c(0, 1.1),
                        breaks = seq(0, 1, 0.25)) +
  ggtitle("Figure 6. USA Map of Cumulative Cases", subtitle="Total Cases Per 100K People on Feb 19, 202
  coord_fixed(1.3) +
  theme(plot.title = element_text(size = 14, face = "bold"),
```

```
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.title.x = element_blank(),
        axis.title.y = element_blank(),
        axis.text.x = element_blank(),
        axis.text.y = element_blank(),
        axis.ticks.x = element_blank(),
        axis.ticks.y = element_blank(),
        legend.position="right",
        legend.key.size = unit(0.8, "cm"))
fig.map


# remove outliers
df_reg <- df_reg[-order(df_reg$new_cases, decreasing =TRUE)[1],]
max(df_reg$new_cases)


# summary
tab_01 = data.frame(
  Variables = c("New Cases", "Booster Rate", "Black Percent", "Hispanic Percent" , "Asian Percent", "Po
  Min  = c(min(df_reg$new_cases),  min(df_reg$booster_vax_pct),  min(df_reg$black_pct),  min(df_reg$his
  Mean = c(mean(df_reg$new_cases), mean(df_reg$booster_vax_pct), mean(df_reg$black_pct), mean(df_reg$his
  SD   = c(sd(df_reg$new_cases),   sd(df_reg$booster_vax_pct),   sd(df_reg$black_pct),   sd(df_reg$hispa
  Max  = c(max(df_reg$new_cases),  max(df_reg$booster_vax_pct),  max(df_reg$black_pct),  max(df_reg$his
)

tab_01 %>%
  kbl(booktabs=T, linesep = "",
      align = c("l", "r", "r", "r", "r"),
      col.names = c("Quantitative Measures", "Min", "Mean", "SD", "Max"),
      digits = 2,
      caption = "Descriptive Statistics ($n=2761$)") %>%
  kable_classic(full_width = F) %>%
  kable_styling(latex_options = "hold_position")

tab_02 <- data.frame(table(df_reg$Metro_status))
tab_02$Prop <- round(tab_02$Freq/dim(df_reg)[1]*100,2)
tab_02$names <- c("Metro status", "")
tab_02 <- subset(tab_02, select=c(names, Var1, Freq, Prop))

tab_03 <- data.frame(table(df_reg$state))
tab_03$Prop <- round(tab_03$Freq/dim(df_reg)[1]*100,2)
tab_03$names <- c("Selected States", rep("", 47))
tab_03 <- subset(tab_03, select=c(names, Var1, Freq, Prop))

tab_04 <- rbind(tab_02, tab_03[c(1, sample(2:48, 9)),])
rownames(tab_04) <- NULL

tab_04 %>%
  kbl(booktabs=T, linesep = "",
      align = c("l", "r", "r","r"),
      col.names = c("Qualitative Measures", "Level", "N", "%"),
```

```
      caption = "Continued Descriptive Statistics ($n=2761$)") %>%
  kable_classic(full_width = F) %>%
  scroll_box(height = "200px") %>%
  kable_styling(latex_options = "hold_position")


g1 <- ggplot(df_reg, aes(x = new_cases)) +
         geom_histogram(color="black", fill="white", bins=80) +
         xlab("Weekly new cases") +
         ggtitle("", subtitle="New Cases Per 100K People") +
         theme_classic() +
         theme_bw() +
         theme(panel.grid.major=element_blank(),
               panel.grid.minor=element_blank())

g2 <- ggplot(df_reg, aes(x = booster_vax_pct, y = new_cases)) +
            geom_point() +
            geom_smooth(method=lm, se=FALSE) +
            xlab("Booster rate")+
            ylab("count") +
            ggtitle("", subtitle="New Cases vs. Booster Rate") +
            theme_classic() +
            theme_bw() +
            theme(panel.grid.major=element_blank(),
                  panel.grid.minor=element_blank())

g3 <- ggplot(df_reg, aes(x = state, y = new_cases)) +
            geom_boxplot() +
            xlab("State")+
            ylab("count") +
            ggtitle("", subtitle="New Cases vs. State") +
            theme_classic() +
            theme_bw() +
            theme(panel.grid.major=element_blank(),
                  panel.grid.minor=element_blank(),
                  axis.text.x = element_blank())

g4 <- ggplot(df_reg, aes(x = Metro_status, y = new_cases)) +
            geom_boxplot() +
            xlab("Metro status")+
            ylab("count") +
            ggtitle("", subtitle="New Cases vs. Metro Status") +
            theme_classic() +
            theme_bw() +
            theme(panel.grid.major=element_blank(),
                  panel.grid.minor=element_blank())

title=text_grob("Figure 7. Exploratory Plots (I)", size = 14, face = "bold")
grid.arrange(g1, g2, g3, g4, ncol=2, top = title)


panel.hist <- function(x, ...) {
    usr <- par("usr"); on.exit(par(usr))
```

```
    par(usr = c(usr[1:2], 0, 1.5) )
    h <- hist(x, plot = FALSE)
    breaks <- h$breaks; nB <- length(breaks)
    y <- h$counts; y <- y/max(y)
    rect(breaks[-nB], 0, breaks[-1], y)
}
panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...) {
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- abs(cor(x, y,use="pairwise.complete.obs"))
    txt <- format(c(r, 0.123456789), digits = digits)[1]
    txt <- paste0(prefix, txt)
    if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
    text(0.5, 0.5, txt, cex = cex.cor * r)
}
pairs(~ + new_cases + booster_vax_pct + black_pct + hispanic_pct + asian_pct + log(pop_density), data =


df_reg$count <- round(df_reg$new_cases, 0)
df_reg$Metro_status <- relevel(as.factor(df_reg$Metro_status),"Metro")

# Linear regression on new cases
fit1 <- lm(count ~ booster_vax_pct + state, data = df_reg)
fit1.1 <- lm(count ~ booster_vax_pct + state +
                     black_pct + hispanic_pct + asian_pct +
                     log(pop_density) + Metro_status, data = df_reg)

anova(fit1)
anova(fit1.1)

perms1 <- aovp(count ~ booster_vax_pct + state, data = df_reg)
perms1.1 <- aovp(count ~ state + black_pct + hispanic_pct + asian_pct +
                     log(pop_density) + Metro_status + booster_vax_pct, data = df_reg)

perms <- lmp(count ~ booster_vax_pct + state +
                     black_pct + hispanic_pct + asian_pct +
                     log(pop_density) + Metro_status, data = df_reg)
summary(perms)

summary(perms1)
summary(perms1.1)

summary(fit1)
summary(fit1.1)


# ANOVA Test
anova1 <- anova(fit1, fit1.1)

options(knitr.kable.NA = '')

test1 <- round(anova1, 3)
```

```r
test1 %>%
  kbl(booktabs=T, linesep = "",
      align = c("l", "r", "r", "r", "r", "r"),
      caption = "ANOVA Test for Model (1) and (2)") %>%
  kable_classic(full_width = F) %>%
  kable_styling(latex_options = "hold_position")

anova1.1 <- anova(fit1.1)

test2 <- round(anova1.1, 3)
rownames(test2) <- c("Booster Rate", "State", "Black Percent", "Hispanic Percent", "Asian Percent", "log

test2 %>%
  kbl(booktabs = T, linesep = "",
      align = c("r", "r", "r", "r", "r", "r"),
      caption = "ANOVA Table for Model (2)") %>%
  kable_classic(full_width = F) %>%
  kable_styling(latex_options = "hold_position")


tidy(fit1.1)[-c(3:49),-4]%>%
  mutate(
    p.value = scales::pvalue(p.value),
    term = c("Intercept", "Booster Rate", "Black Percent", "Hispanic Percent", "Asian Percent",
             "log(Pop. Density)", "Non Metro")) %>%
    kable(booktabs=T, linesep = "",
          caption = "Coefficient Estimates for Model (2)",
          col.names = c("Predictor", "Coefficient", "SE", "p-value"),
          digits = c(0, 3, 3, 3, 3),
          align = c("l", "r", "r", "r")) %>%
    kable_classic(full_width = F) %>%
    kable_styling(latex_options = "hold_position")


# Diagnostics
par(mfrow=c(1,2),oma=c(2,2,2,2),mar=c(4,3,3,3))
plot(fit1.1,1, sub.caption = "")
plot(fit1.1,2, sub.caption = "")
mtext("Figure 9, Diagnostic Plots for Model (2)", side = 3, font=2, line=-1, outer=TRUE)


perm1.1 <- summary(perms1.1)

test2 <- round(perm1.1[[1]], 3)
rownames(test2) <- c("State", "Black Percent", "Hispanic Percent", "Asian Percent", "log(Pop. Density)"

test2 %>%
  kbl(booktabs=T, linesep = "",
      align = c("r", "r", "r", "r", "r", "r"),
      caption = "Permutation Test for Model (2)") %>%
  kable_classic(full_width = F) %>%
  kable_styling(latex_options = "hold_position")
```

```r
## Overdispersion test
fit0 <- glm(count ~ booster_vax_pct + state, family = "poisson", data = df_reg)
dispertest <- check_overdispersion(fit0)
dispertest

fit2 <- glm.nb(count ~ booster_vax_pct + state, data = df_reg)
pvalue <- pchisq(2 * (logLik(fit2) - logLik(fit0)), df = 1, lower.tail = FALSE)


dispertable <- data.frame(v1=c("Dispersion ratio", "Pearson's Chi-Squared", "P-value", NA),
                          v2=c("  ", "  ", "  ", "  "),
                          v3=c(" = ", " = ", " = ", NA),
                          v4=c(round(dispertest[[2]],3), round(dispertest[[1]],3), "< 0.001", "Overdispe

options(knitr.kable.NA = '')
dispertable %>%
  kbl(booktabs=T, linesep = "",
      col.names = NULL,
      align = c("r", "r", "c","r"),
      caption = "Overdispersion Test") %>%
  kable_classic(full_width = F) %>%
  kable_styling(latex_options = "hold_position")


# Negative binomial regression on new cases
fit2 <- glm.nb(count ~ booster_vax_pct + state, data = df_reg)
fit2.1 <- glm.nb(count ~ booster_vax_pct + state +
                         black_pct + hispanic_pct + asian_pct +
                         log(pop_density) + Metro_status, data = df_reg)
anova(fit2)
summary(fit2)

anova(fit2.1)
summary(fit2.1)


anova2 <- anova(fit2, fit2.1)

options(knitr.kable.NA = '')

test3 <- cbind(anova2[3], anova2[4], anova2[5], anova2[6], anova2[7], anova2[8])

test3 %>%
  kbl(booktabs=T, linesep = "",
      align = c("l", "c", "c","c", "c", "c"),
      caption = "ANOVA Test for Model (3) and (4)") %>%
  kable_classic(full_width = F) %>%
  kable_styling(latex_options = "hold_position")

anova2.1 <- anova(fit2.1)

test4 <- round(anova2.1, 3)
rownames(test4) <- c("NULL", "Booster Rate", "State", "Black Percent", "Hispanic Percent", "Asian Percer
```

```
test4 %>%
  kbl(booktabs=T, linesep = "",
      align = c("r", "r", "r", "r", "r", "r"),
      caption = "ANOVA Table for Model (4)") %>%
  kable_classic(full_width = F) %>%
  kable_styling(latex_options = "hold_position")


tidy(fit2.1)[-c(3:49),-4] %>%
  mutate(
    p.value = scales::pvalue(p.value),
    term = c("Intercept", "Booster Rate", "Black Percent", "Hispanic Percent", "Asian Percent",
             "log(Pop. Density)", "Non Metro")) %>%
    kable(booktabs=T, linesep = "",
          caption = "Coefficient Estimates for Model (4)",
          col.names = c("Predictor", "Coefficient", "SE", "p-value"),
          digits = c(0, 3, 3, 3, 3),
          align = c("l", "r", "r", "r")) %>%
    kable_classic(full_width = F) %>%
    kable_styling(latex_options = "hold_position")


par(mfrow=c(1,2),oma=c(2,2,2,2),mar=c(4,3,3,3))
plot(fit2.1,1, sub.caption = "")
plot(fit2.1,2, sub.caption = "")
mtext("Figure 10. Diagnostic Plots for Model (4)", side = 3, font=2, line=-1, outer=TRUE)


data <- subset(df_reg, select = -c(new_cases, log_new_cases))
data$pop_density <- log(data$pop_density)

m1 <- glm(count ~ . , data = data)
m4 <- glm.nb(count ~ ., data = data)

po.nb <- predprob(m4) %>% colMeans

df <- data.frame(x = 0:max(data$count), NegBin = po.nb)

obs <- table(data$count) %>% prop.table() %>% data.frame
names(obs) <- c("x", 'Observed')

p1 <- predict(m1) %>% round() %>% table %>% prop.table %>% data.frame
names(p1) <- c('x', 'OLS')
p1 <- p1[-1, ] #there is a negative value

tmp <- merge(p1, obs, by = 'x', all = T)
tmp$x <- as.numeric(as.character(tmp$x))

comb <- merge(tmp, df, by = 'x', all = T)
comb[is.na(comb)] <- 0
mm <- melt(comb, id.vars = 'x', value.name = 'prob', variable.name = 'Model')

ggplot(mm, aes(x = x, y = prob, group = Model, col = Model)) +
```

```
  geom_line(aes(lty = Model), lwd = 1) +
  theme_bw() +
  labs(x = "Number of new casess", y = 'Probability',
  title = "Figure 11. Comparison of Models") +
  scale_color_manual(values = c('blue','black', 'red')) +
  scale_linetype_manual(values = c('dashed', 'solid', 'dotted')) +
  theme(legend.position=c(.65, .65),
        legend.background = element_rect(),
        plot.title = element_text(hjust = 0.5, size = 14, face = "bold"))
```