An Analysis of COVID-19 Prevalence as a Function of Ethnic and Economic Characteristics and Proximity to Major Universities in San Diego County

[redacted]

Data Analysis Exam

Fall 2022

August 23, 2022

An Analysis of COVID-19 Prevalence as a Function of Ethnic and Economic Characteristics and Proximity to Major Universities in San Diego County

[redacted]

Data Analysis Exam

Fall 2022

August 23, 2022

Executive Summary

The novel coronavirus SARS-Cov-2 was the driving force behind a global pandemic of COVID-19, the disease associated with the virus; it began in March 2020 and may now be at a point where the disease is endemic. It is of interest to study the pandemic's progression from April 1, 2020 to June 29, 2021 in San Diego County in the state of California by gaining insight from data about the county's ZIP codes. OLS regression models, one for each season of the pandemic (with the exception of Summer 2021), were used to make inference about average COVID-19 prevalence, given predictors for various ethnicities and economic characteristics of the county's ZIP codes, as well as proximity to any of the five major universities in the county. This study finds that proportion of Hispanic residents in a ZIP code is often a predictor of increased average COVID prevalence metrics, and similarly behaved predictors of increased average COVID prevalence (yet with less frequency) are the proportions of Black and Native Hawaiian residents in a ZIP code. Proportion of unemployed residents predicted increased average COVID prevalence metrics in Fall 2020 and Spring 2021, and proportion of uninsured residents in a ZIP code was a predictor of decreased average COVID prevalence in Spring 2020, the earliest segment of the pandemic. Proximity to a major university had a significant effect in predicting increasing average COVID prevalence metrics in Summer 2020, but an opposite effect in Winter 2021. These results suggest that minority residents, especially Hispanic residents, ought to be prioritized in the implementation of public health strategies for prevention and containment of COVID-19 (including in the accessibility of healthcare facilities and understanding health insurance), and although proximity to a university has different effects depending on season, it is prudent for students at any time to implement COVID safety practices in their daily routine and to get vaccinated against COVID-19.

Introduction

N.B. "COVID" and "COVID-19" are utilized interchangeably in this paper.

COVID-19 is the disease caused by a novel coronavirus known as the SARS-CoV-2, and it is characterized by symptoms including (but not limited to) fever or chills, cough, shortness of breath or difficulty breathing, fatigue and aching, new loss of taste or smell, sore throat, and congestion (Centers for Disease Control and Prevention n.p.). In March 2020, as COVID-19 was becoming known to spread with facility, and as the World Health Organization labeled the spread of this virus and associated disease a global pandemic, the residents of many cities around the world were ordered to quarantine at home until scientists could better understand, and develop a vaccine for, this novel disease (Yale Medicine n.p.). After numerous "waves" in which COVID-19 would spike in certain regions, vaccines began to roll out in mid-December 2020 in an effort to increase the U.S. population's immunity to COVID-19 (Ibid. n.p.). A general loosening of masking requirements implemented earlier in the pandemic occurred in California on June 15, 2021 (California n.p.). The pandemic is transitioning to an endemic state (Locklear n.p.). Although this may be the case, it is meritorious to pose questions about the COVID-19 pandemic to learn from it and to prepare for a future pandemic.

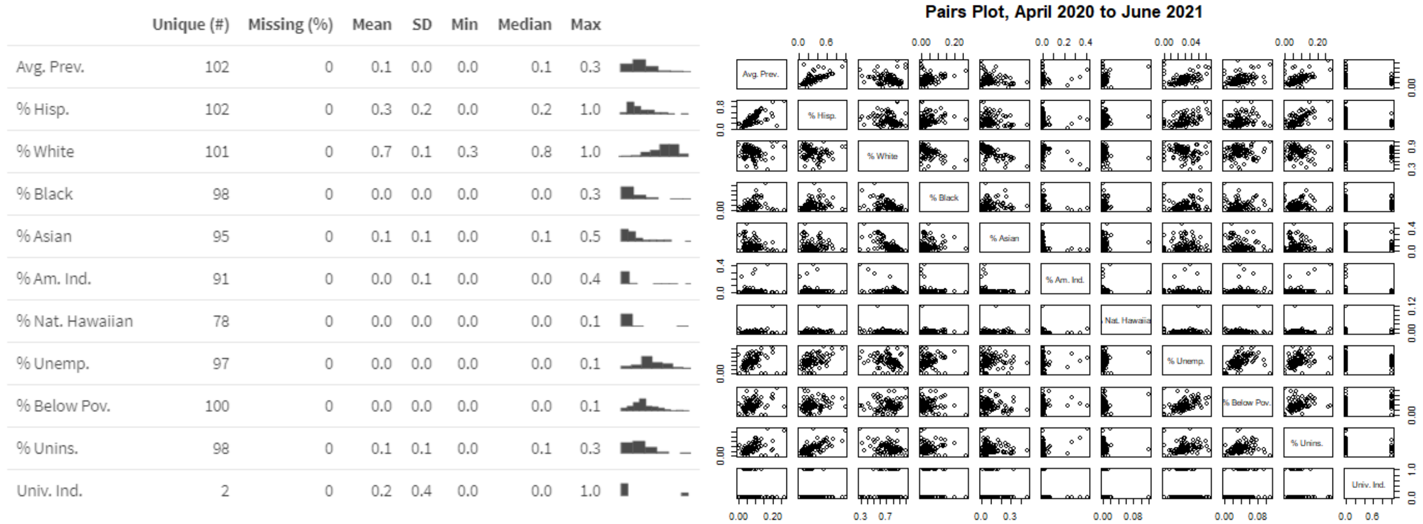
For each season of the pandemic accounted for in the data, how does average COVID prevalence tend to respond to different compositions of various ethnicities and economic characteristics in distinct ZIP codes, especially in certain seasons? Also, are certain characteristics with significant effects around a major university in San Diego County? People of lower socioeconomic standing tend to have limited access to healthcare and other resources, and certain ethnicities such as Hispanic and Black historically tend to be in low socioeconomic standing. I predict that primarily these two ethnicities will predict higher average COVID prevalence in their ZIP codes than other ethnicities, and I also predict that with the rise of proportions of unemployed, uninsured, or impoverished persons, average COVID prevalence will also rise since these factors may limit access to healthcare facilities. Despite university evacuations in the county during the start of the pandemic, if students stayed in residences off-campus, they would still tend to live in community and to be involved in activities outside of the home unless they are consistent in applying COVID-19 safety practices. My hypothesis is that, by means of students being more conscientious of COVID, prevalence around universities will be low, but this prevalence should rise in the early parts of the pandemic as well as in coincidence with the typical flu season in winter. I will approach this problem by building regression models predicting average COVID prevalence (or a transformation of it) for each season, and by including in those models an indicator for a ZIP code indicating proximity to a major university and selecting appropriate variables and models via best subsets regression. This method is appropriate because it will generate performance metrics for each model and allow me to determine, from all possible models with the predictors inputted into best subsets regression, which ones are best for each season after fitting those candidate variable sets to linear models.

Exploratory Data Analysis

For the 452 days spanning April 1, 2020 to June 29, 2021 (452 days), this study analyzes data for 102 ZIP codes in San Diego County including 16 variables containing demographic information, as well as COVID count data for 452 days. This data does not consist of individual residents and their information (i.e., all data analyzed here is group data unconditional of (the knowledge of) individual demographic variable values). The demographics file does not contain 11 ZIP codes that are listed in the COVID file (for a total of 113 ZIP codes; see Appendix A1 for deletion process). In total, 3,298,704 people are counted in the demographic file.

COVID counts are discrete and positive, so this analysis does not lend itself to OLS regression; Poisson or negative binomial regression ought to be considered. I took the mean and variance of COVID count for

each ZIP code across the 452 days and used this information to assess the equality of mean and variance (see Appendix A2 for an overdispersion assumption verification). The count data are underdispersed in each season, so I decided to look at average COVID prevalence, defined to be the difference between the COVID counts in a ZIP code on the first and last days of the season, divided by the ZIP code's population size, a transformation rendering COVID count a continuous variable (see Appendix A3 for code).
I considered the distributions of COVID counts and COVID prevalence in each ZIP code across the entire 14 months and in each season (see Appendix A3).



	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max
Avg. Prev.	102	0	0.1	0.0	0.0	0.1	0.3
% Hisp.	102	0	0.3	0.2	0.0	0.2	1.0
% White	101	0	0.7	0.1	0.3	0.8	1.0
% Black	98	0	0.0	0.0	0.0	0.0	0.3
% Asian	95	0	0.1	0.1	0.0	0.1	0.5
% Am. Ind.	91	0	0.0	0.1	0.0	0.0	0.4
% Nat. Hawaiian	78	0	0.0	0.0	0.0	0.0	0.1
% Unemp.	97	0	0.0	0.0	0.0	0.0	0.1
% Below Pov.	100	0	0.0	0.0	0.0	0.0	0.1
% Unins.	98	0	0.1	0.1	0.0	0.1	0.3
Univ. Ind.	2	0	0.2	0.4	0.0	0.0	1.0

Figure 1: Summary statistics and pairs plot for data on 102 San Diego County ZIP codes, over the 14 months of interest.

Before creating pairs plots to explore variable relationships, I created a "university (proximity) indicator" variable that evaluates to 1 when a ZIP code contains a major university and residences that are not university dormitories or if it is a ZIP code containing residences and is contiguous with a ZIP code containing only a major university; otherwise, the variable takes on the value of 0. I created a data summary and pairs plot to understand the relationship between various independent variables and average COVID prevalence, all as proportions in terms of population size rather than as discrete counts (see Figure 1 above for a plot for the entire study period and see Appendix A3 for code for seasonal pairs plots). It is apparent in considering seasons that proportion of Hispanics and proportion of uninsured residents share one of the strongest correlations present. It is also apparent there are outliers in some of the predictors, and they may need to be accounted for in final model building. Visual inspection of all plots shows that multicollinearity is not a major concern moving forward with regression. Relationships between predictors and certain observations are more erratic in Summer 2021 since only ten days of data are available for this season, so Summer 2020 will not have a model built for it. I also found that the distributions of COVID prevalence and predictors tend to become more normal under logarithmic and square root transformations, with consistently more normal appearance under logarithmic transformation than square root transformation (see Appendix A4). These same relationships might not hold under transformation of variables within the models to be developed.

Statistical Analysis

I took the variables for average COVID prevalence in a given season, proportion of Hispanic residents, proportion of White residents, proportion of Black residents, proportion of Asian residents, proportion of American Indian residents, proportion of Native Hawaiian residents, proportion of unemployment residents, proportion of residents living below poverty, proportion of uninsured residents, and the university indicator, and ran them through best subsets regression by considering season; a model was not created for Summer 2022 since this period lasts for only ten days (see Appendix A5). Within seasons, I compared models of the same number of terms by comparing R^2 and Mallow's C_p, with the "best" model having the highest R^2 and lowest Mallow's C_p. This method is appropriate because since I have ZIP code-level data, I could bring into consideration constituent ethnicities and other economic characteristics that would represent a broad range of predictors of COVID. This allows me to propose models that will be more suited to OLS regression. These candidate models could still require transformation and deletion of variables, however, depending on diagnostics.

The best subsets regression returned five sets of variable choices (and five more to consider as alternatives should the need arise). These variable choices were fitted to linear model objects in R and their diagnostic measures assessed for propriety for OLS regression. Variables and measures for diagnostic tests were assessed for statistical significance at the level of significance $\alpha = 0.05$. Different transformations and their reasons for inclusion or exclusion are in Appendices A6 and A7 (with final models and their summaries and diagnostics being in Appendix A7), though the most common transformation is the square root of average COVID prevalence since the square root function is strictly increasing on its domain and increases slower than a linear function does. All models were found to have zero-distributed residuals (assessed by residual plots and histograms), zero-covariance in residuals (by the Durbin-Watson Test), normally distributed errors (by visual inspection and histograms), and homoskedasticity of residuals (by the Breusch-Pagan Test), thus satisfying the usual assumptions for OLS regression.

Spring 2020: (Average COVID prevalence)$^{1/2}$ = 0.0150 + 0.1110(% Hispanic) + 0.1870(% Black) - 0.1498(% Uninsured).

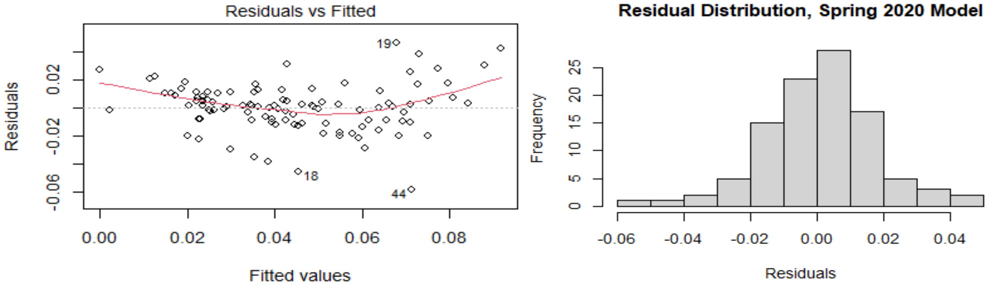


Figure 2: Diagnostics for Spring 2020 Model.

See Figure 2 above for diagnostics. All predictors β_i ($i = 0, 1, 2, 3$) were subjected to the hypothesis test H_0: $\beta_i = 0$ v. H_1: $\beta_i \neq 0$, with the above coefficient estimates for intercept ($t = 4.394$, $p < 0.0001$), % Hispanic ($t = 9.560$, $p < 0.0001$), % Black ($t = 5.032$, $p < 0.0001$) and % Uninsured ($t = -3.474$, $p < 0.0001$) being statistically significant at significance level $\alpha = 0.05$. This model suggests that in the spring of 2020, for every 1% increase in the proportion of Hispanic residents in a ZIP code, there was an average increase of 0.111 in the square root of average COVID prevalence in that ZIP code (all else constant). For every 1% increase in the proportion of Black residents in a ZIP code, there was an average increase of 0.187 in the square root of average COVID prevalence in that ZIP code (all else constant). For every 1% increase in the proportion of uninsured residents in a ZIP code, there was an average decrease of 0.1498 in the square root of average COVID prevalence in that ZIP code (all else constant). The intercept term suggests that when the predictors here are each at zero-value, the square root of average COVID prevalence in a ZIP code was 0.0150.

Summer 2020: (Average COVID prevalence)$^{1/2}$ = 0.0469 + 0.1252(% Hispanic) + 1.0204(% Native Hawaiian) + 0.0168(university indicator) + 0.1089(ZIP codes 92059, 92060, 92135).

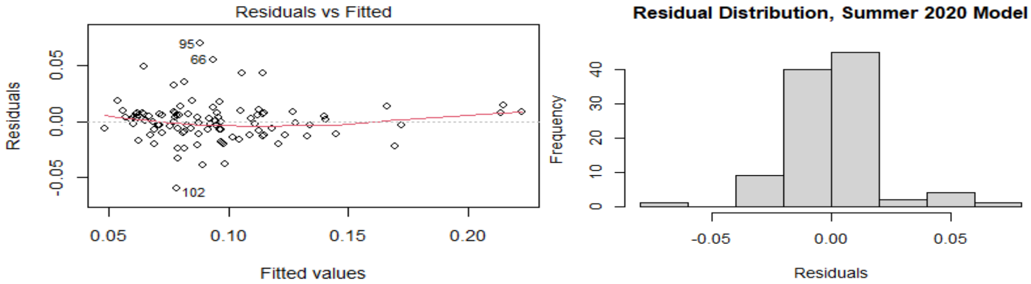


Figure 3: Diagnostics for Summer 2020 Model.

See Figure 3 above for diagnostics. All predictors β_i (i = 0, 1, 2, 3, 4) were subjected to the hypothesis test H_0: $\beta_i = 0$ v. H_1: $\beta_i \neq 0$, with the above coefficient estimates for intercept (t = 12.438, $p < 0.0001$), % Hispanic (t = 12.535, $p < 0.0001$), % Native Hawaiian (t = 6.590, $p < 0.0001$), university indicator (t = 3.441, $p \approx 0.0009$), and indicator for ZIP codes 92059, 92060, and 92135 (t = 9.823, $p < 0.0001$) being statistically significant at significance level $\alpha = 0.05$. This model suggests that in Summer 2020, for each 1% increase in the proportion of Hispanic residents in a ZIP code, there was an average 0.1252 increase in the square root of average COVID prevalence (all else constant). For each 1% increase in the proportion of Native Hawaiian residents in a ZIP code, there was an average 1.0204 increase in the square root of average COVID prevalence (all else constant). 0.0168 is a measure of the effect of proximity to a major university on the square root of the average COVID prevalence relative to not being in a ZIP code that is proximate to a major university. The second indicator term fits outlier observations belonging to ZIP codes 92059, 92060, and 92135. 0.1809 is a measure of the effect of being in these ZIP codes on the square root of average COVID prevalence relative to not being in these ZIP codes. The intercept term suggests that when the predictors here are each at zero-value, the square root of average COVID prevalence in a ZIP code was 0.0469.

Fall 2020: (Average COVID prevalence)$^{1/2}$ = 0.1242 – 0.1389(% Asians) + 1.0948(% Native Hawaiian) + 0.9825(% Unemployment).

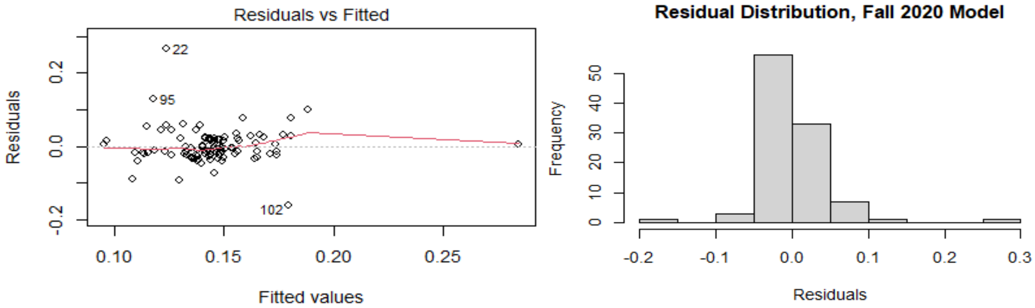


Figure 4: Diagnostics for Fall 2020 Model.

See Figure 4 above for diagnostics. All predictors β_i (i = 0, 1, 2, 3) were subjected to the hypothesis test H_0: $\beta_i = 0$ v. H_1: $\beta_i \neq 0$, with the above coefficient estimates for intercept (t = 10.359, $p < 0.0001$), % Asian (t = -2.708, $p \approx 0.0080$), % Native Hawaiian (t = 2.779, $p \approx 0.0065$) and % Unemployed (t = 2.801, $p \approx 0.0061$) being statistically significant at significance level $\alpha = 0.05$. This model suggests that in Fall 2020, for each 1% increase in the proportion of Asians in a ZIP code, there was an average 0.1389 decrease in the square root of average COVID prevalence (all else constant). For each 1% increase in the proportion of Native Hawaiian residents in a ZIP code, there was an average 1.0948 increase in the square root of average COVID prevalence in that ZIP code (all else constant). For each 1% increase in the proportion of unemployed residents in a ZIP code, there was an average 0.9825 increase in the square root of average COVID prevalence in that ZIP code (all else constant). The intercept term suggests that when each predictor here is at zero-value, the square root of average COVID prevalence in a ZIP code is 0.1242.

Winter 2021: (Average COVID prevalence) = 0.0129 + 0.0587(% Hispanic) + 0.1456(% Below Poverty) – 0.0058(university indicator) + 0.0953*I(ZIP = 92060).

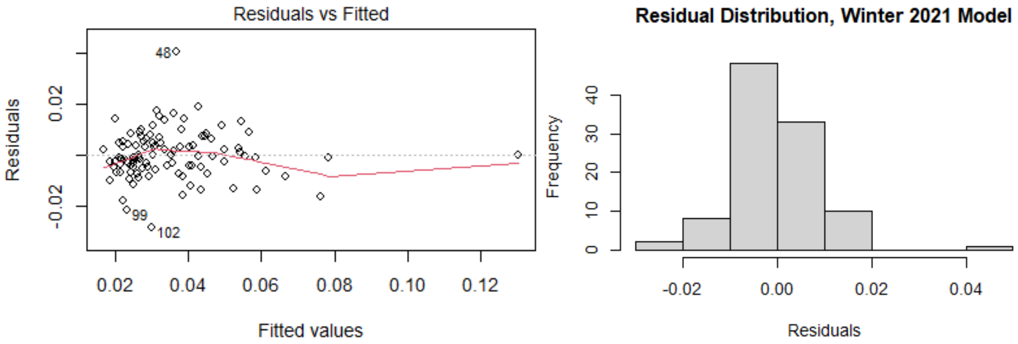


Figure 5: Diagnostics for Winter 2021 Model.

See Figure 5 above for diagnostics. All predictors β_i (i = 0, 1, 2, 3, 4) were subjected to the hypothesis test $H_0: \beta_i = 0$ v. $H_1: \beta_i \neq 0$, with the above coefficient estimates for intercept (t = 5.499, $p < 0.0001$), % Hispanic (t = 10.863, $p < 0.0001$), % Below Poverty (t = 2.751, $p \approx 0.0071$), university indicator (t = -2.240, $p \approx 0.0274$), and indicator for ZIP code 92060 (t = 8.927, $p < 0.0001$) being statistically significant at significance level $\alpha = 0.05$. This model suggests that for each 1% increase in the proportion of Hispanic residents in a ZIP code, there was an average increase in average COVID prevalence by 0.0587 (all else constant). For each 1% increase in the proportion of residents below poverty in a ZIP code, there was an average 0.1456 increase in average COVID prevalence in that ZIP code (all else constant). 0.0058 is a measure of the effect of proximity to a major university on the average COVID prevalence relative to not being in a ZIP code that is proximate to a major university. The second indicator term fits an outlier observation belonging to the ZIP code 92060, which encompasses Palomar Mountain. 0.0953 is a measure of the effect of being in ZIP code 92060 on average COVID prevalence relative to not being in ZIP code 92060. The outlier observation associated with this ZIP code has been fitted as another term in the model since the present study is most interested in COVID prevalence in the major urban areas of San Diego County. The intercept term suggests that when each predictor here is at zero-value, the average COVID prevalence in a ZIP code is 0.0129.

Spring 2021: (Average COVID Prevalence) = 0.0020 + 0.0047(% Hispanic) + 0.0196(% Black) + 0.0381(% Unemployed) + 0.0301*I(ZIP = 92060)

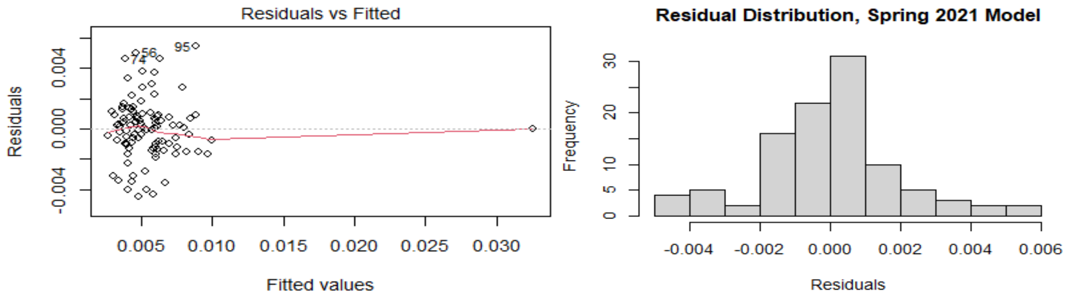


Figure 6: Diagnostics for Spring 2021 Model.

See Figure 6 above for diagnostics. All predictors β_i (i = 0, 1, 2, 3, 4) were subjected to the hypothesis test $H_0: \beta_i = 0$ v. $H_1: \beta_i \neq 0$, with the above coefficient estimates for intercept (t = 3.732, $p \approx 0.0003$), % Hispanic (t = 4.111, $p < 0.0001$), % Black (t = 4.374, $p < 0.0001$), % Unemployed (t = 2.362, $p \approx 0.0202$), and indicator for ZIP code 92060 (t = 14.715, $p < 0.0001$) being statistically significant at significance level $\alpha = 0.05$. This model suggests that for every 1% increase in the proportion of Hispanic residents in a ZIP code, there was an average 0.0047 increase in average COVID prevalence in that ZIP code (all else constant). For every 1% increase in the proportion Black residents in a ZIP code, there was an average 0.0196 increase in average COVID prevalence in that ZIP code (all else constant). For every 1% increase in the proportion of unemployed residents in a ZIP code, there was an average 0.0381 increase in average COVID prevalence in that ZIP code (all else constant). The indicator term fits an outlier observation belonging to the ZIP code 92060, which encompasses Palomar Mountain. 0.0301 is a measure of the effect of being in ZIP code 92060 (contains Palomar Mountain) on average COVID prevalence relative to not being in ZIP code 92060. This ZIP code has been fitted as another term in the model since the present study is most interested in COVID prevalence in the major urban areas of San Diego County. The intercept term suggests that when each predictor here is at zero-value, the average COVID prevalence in a ZIP code is 0.002.

Conclusions

This study finds that in looking at each separate season (with the exception of Summer 2021 due to paucity of data), Hispanic, Black, and Native Hawaiian residents are consistently among the predictors of increasing average COVID prevalence, with Hispanic resident proportion appearing to be a significant predictor in every season except Fall 2020, and Native Hawaiians predicting average COVID prevalence only in Summer and Fall 2020. Proportion below poverty was only in the Winter 2021 model and was found to predict increased average COVID prevalence, and proportion of unemployment was found to be significant in Fall 2020 and Spring 2021, both times predicting increase in average COVID prevalence metrics. The aforementioned findings align with my earlier predictions. Proximity to a university was only considered significant in Summer 2020 (possibly in contribution to “waves” of COVID during summer vacation) and Winter 2021 (possibly due to vaccination rollout and workers in education receiving priority for vaccination at the time (Lambert and Tadayon n.p.)); this finding both confirms and runs counter to my earlier predictions. One conclusion drawn that runs contrary to my earlier predictions was the effect of a ZIP code’s uninsured proportion in Spring 2020, with one explanation being that this season started about two weeks into initial lockdown, and with fear and uncertainty surrounding the then-novel COVID-19 disease, the uninsured population in the county would need to practice more caution and could be more likely than the insured population to quarantine and practice COVID safety measures more consistently and diligently.

Limitations of this study include the fact that the data provided is not subdivided into lower levels such as the demographic and health records of individual residents. Also, as this data is only ZIP-code level, it may have led to an underspecified quality of the models selected by best subsets regression (a procedure which could potentially exacerbate this under-specification). In addition, as we only know COVID counts for each day and only have one total population figure for each ZIP code, we do not know if the counts include or exclude people who have recovered from COVID, moved away from that ZIP code, or have died. In addition, ZIP codes were created with the organization of mail delivery routes as their primary intention (U.S. Census Bureau n.p.). These divisions are not suitable indicators of human patterns or trends such as the spread of disease and are therefore inappropriate proxies for geographical area.

Topics for future study include exploring the association between race, unemployment, and insurance status; this idea of health disparity between majority and minority races is well-explored in public health. In addition to these new topics, another approach to the same questions of the present study would involve a discrete time series analysis approach since this data set lends itself well to time series analysis, with one such implementation being a negative binomial model with a time series component (Davis and Wu 735-736). Spatial epidemiological methods would also be suited to the present data, as well as calculating Moran’s I and performing spatial regression (especially in the research question regarding proximity to universities, and this could be extended to other potential “hotspots” for COVID-19), but these phenomena should be studied with respect to areal regions, not to ZIP codes.

References

California, State of. “Safely Reopening California.” *Covid19.Ca.gov*, covid19.ca.gov/safely-reopening/.

Census Bureau. United States. “ZIP Code Tabulation Areas (ZCTAs).” https://www.census.gov/programs-surveys/geography/guidance/geo-areas/zctas.html

Centers for Disease Control and Prevention. “Symptoms of COVID-10.” 11 August 2022. https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html

Davis, Richard A., and Rongning Wu. “Negative Binomial Model for Time Series of Counts.” *OUP Academic*, Oxford University Press, 1 July 2009, https://doi.org/10.1093/biomet/asp029.

https://blink.ucsd.edu/technology/help-desk/directory/address.html

https://tools.usps.com/

https://www.calstate.edu/attend/campuses/san-marcos

https://www.sdsu.edu/contact
https://www.timeanddate.com/calendar/seasons.html

https://www.yalemedicine.org/news/covid-timeline
https://www.zipdatamaps.com/92093
https://www.zipdatamaps.com/92136

https://www.zipdatamaps.com/92155

https://www.zipdatamaps.com/92161

https://www.zipdatamaps.com/92672

Lambert, Diana and Ali Tadayon. “California teachers could get Covid-19 vaccinations starting in February.” EdSource. 26 January 2021. https://edsource.org/2021/california-teachers-could-get-covid-19-vaccinations-starting-in-february/647643

Locklear, Mallory. “For Covid-19, Endemic Stage Could Be Two Years Away,” *YaleNews*, 5 July 2022, https://news.yale.edu/2022/07/05/covid-19-endemic-stage-could-be-two-years-away.

Appendix

A1: Reading in Data

```
#The COVID and demographic files have a discrepancy in the number of ZIP codes represented. The code below isolates the ZIP codes comprising this discrepancy:

# Use as a check for ZIP code discrepancies before deleting rows of sdCovid to get covidData

setdiff(sort(sdCovid[,1]), sort(sdDemo[,1])) # Which zipcode values are not in the demographics file?
```

The ZIP codes returned by this code (before altering the original data files) correspond to the following cities and code type recognized by the United States Postal Service on its website, with further clarification/commentary being found by means of other websites:

91931: Guatay, CA (rural/desert)

91934: Jacumba, CA (rural/desert)

91948: Mount Laguna, CA (rural/desert) (all zeroes)

92093: La Jolla, CA (contains the University of California, San Diego campus, the northern half and primary ZIP code for correspondence)

92096: San Marcos, CA (contains the California State University, San Marcos campus)

92136: San Diego, CA (contains National City, CA and the Naval Base San Diego)

92155: San Diego, CA (contains the Naval Amphibious Base in Coronado)

92161: San Diego, CA (contains the University of California, San Diego campus, the southern half)

92182: San Diego, CA (contains the San Diego State University campus)

92259: Ocotillo, CA (rural/desert) (all zeroes)

92672: San Clemente, CA (this is in Orange County, not San Diego County). This ZIP code straddles San Diego and Orange Counties, despite officially being in Orange County. Unique to this ZIP code is its containment of San Onofre Beach and the northernmost gate to Camp Pendleton, a military base primary in San Diego County. Yet, if these two locations are in San Clemente, then the city straddles Orange and San Diego Counties.

I could carry out the COVID analysis here, but I would need to pull in demographic information from other sources, or I could simply consider COVID counts without other covariates. The focus of this study is the transmission of COVID-19 in the major metropolitan area of San Diego, so ZIP codes corresponding to rural or sparse desert areas of the county will be deleted. Therefore, delete ZIP codes 91931, 91934, 91948, 92259. Also, San Clemente will be treated as being part of Orange County, and since the focus of this study is San Diego County, ZIP code 92672 will be deleted. Zip code 92182 was deleted because it is the ZIP code for San Diego State University, which only houses students during the academic year, rendering this ZIP code's population as a number being in annual flux. However, shelter-in-place orders forced students to return to their homes away from campus, so little to no data could be collected here during the period of data collection. I made similar decisions for 92093 and 92161 (UCSD), 92096 (CSUSM), Once I decided on ZIP codes to remove, I created a copy of the COVID data that does not contain those ZIP codes.

```
covidData <- sdCovid[-c(12, 14, 19, 69, 70, 102, 107, 108, 110, 111, 113),] # delete 11 of 113 zip codes (about 9.7% of ZIP codes); see Appendix XX for this rationale
```

```
sdDemo <- read.csv("demographic_SD_ZIP_4_DAE_F22.csv", header = T)
head(sdCovid[,c(1:5)])

##   Zipcode X4.1.2020 X4.2.2020 X4.3.2020 X4.4.2020
## 1   91901         1         1         1         1
## 2   91902         9        10        10        11
## 3   91905         0         0         0         0
## 4   91906         0         0         0         0
## 5   91910        23        28        28        30
## 6   91911        21        24        26        29

head(sdDemo[,c(1:5)])

##   zipcode Total_pop  male female white
## 1   91901     18558  9257   9301 16558
## 2   91902     20138  9869  10269 14388
## 3   91905      1478   789    689  1206
## 4   91906      4351  2399   1952  3079
## 5   91910     74297 36775  37522 54470
## 6   91911     85365 42535  42830 61838

# Create data table with indicators of proximity to university.
sdDemoUnivInd <- read.csv("demographic_SD_ZIP_4_DAE_F22_univInd.csv", header = T)
head(sdDemoUnivInd)

##   zipcode Total_pop  male female white black asian american_Indian
## 1   91901     18558  9257   9301 16558   257   459             437
## 2   91902     20138  9869  10269 14388   551  2630              51
## 3   91905      1478   789    689  1206    83     0              85
## 4   91906      4351  2399   1952  3079    33    22             267
## 5   91910     74297 36775  37522 54470  3057  7483             192
## 6   91911     85365 42535  42830 61838  2943  7601             219
##   native_Hawaiian other_race_one Total_hispanic non_hisp_white non_hisp_black
## 1               0            158           2530          14606            246
## 2             190            706           9279           6746            520
## 3               0              3            185           1074             83
## 4               0            709           1799           2163             31
## 5             272           5156          45307          17132           2610
## 6             692           8876          61189          13013           2317
##   non_hisp_asian unemployedCvilian popBelowpoverty popUninsured univInd
## 1            459               392             432         1083       0
## 2           2500               768             498         1243       0
## 3              0                30             141           50       0
## 4             22               223             246          364       0
## 5           7241              3657            3499         6510       0
## 6           7216              5204            3652         8793       0

# Seasons in the covidData table:
# Analysis will be divided into calendar seasons determined by dates of equinoxes and solstices in the Northern Hemisphere.

# Spring 2020 (April 1, 2020 - June 19, 2020) (col. 2-81)
# Summer 2020 (June 20, 2020 - September 21, 2020) (col. 82-175)
# Fall 2020 (September 22, 2020 - December 20, 2020) (col. 176-265)
# Winter 2020-2021 (December 21, 2020 - March 19, 2021) (col. 266-352)
# Spring 2021 (March 20, 2020 - June 19, 2021) (col. 353-443)
# Summer 2021 (June 20, 2021 - June 29, 2021) (col. 444-453)

sp20CovidData <- covidData[,c(2:81)]
su20CovidData <- covidData[,c(82:175)]
f20CovidData <- covidData[,c(176:265)]
w21CovidData <- covidData[,c(266:352)]
sp21CovidData <- covidData[,c(353:443)]
```

```r
su21CovidData <- covidData[,c(444:453)]
# original was read in, convert to matrix here
covidData <- as.matrix(covidData) # I can access COVID counts alone now that I've turned a list object into a matrix object.

# Variable names for plots
listPairs <- list("Avg. Prev.", "% Hisp.", "% White", "% Black", "% Asian", "% Am. Ind.", "% Nat. Hawaiian", "% Unemp.", "% Below Pov.", "% Unins.", "Univ. Ind.")

# COVID Prevalence, unconditional of season (looks at the entire pandemic)
avgPrevWhole <- (covidData[,length(covidData[1,])]-covidData[,2])/sdDemoUnivInd[,2]
newWhole <- data.frame(avgPrevWhole, sdDemoUnivInd$Total_hispanic/sdDemoUnivInd$Total_pop, sdDemoUnivInd$white/sdDemoUnivInd$Total_pop, sdDemoUnivInd$black/sdDemoUnivInd$Total_pop, sdDemoUnivInd$asian/sdDemoUnivInd$Total_pop, sdDemoUnivInd$american_Indian/sdDemoUnivInd$Total_pop, sdDemoUnivInd$native_Hawaiian/sdDemoUnivInd$Total_pop, sdDemoUnivInd$unemployedCvilian/sdDemoUnivInd$Total_pop, sdDemoUnivInd$popBelowpoverty/sdDemoUnivInd$Total_pop, sdDemoUnivInd$popUninsured/sdDemoUnivInd$Total_pop, sdDemoUnivInd$univInd)
names(newWhole) <- c(listPairs)


# COVID Prevalence, conditioning on season

# look at explanatory variables (ethnicity, unemployed, uninsured, etc.)

# Spring 2020
avgPrevSp20 <- (sp20CovidData[,length(sp20CovidData[1,])]-sp20CovidData[,2])/sdDemoUnivInd[,2]
newSp20 <- data.frame(avgPrevSp20, sdDemoUnivInd$Total_hispanic/sdDemoUnivInd$Total_pop, sdDemoUnivInd$white/sdDemoUnivInd$Total_pop, sdDemoUnivInd$black/sdDemoUnivInd$Total_pop, sdDemoUnivInd$asian/sdDemoUnivInd$Total_pop, sdDemoUnivInd$american_Indian/sdDemoUnivInd$Total_pop, sdDemoUnivInd$native_Hawaiian/sdDemoUnivInd$Total_pop, sdDemoUnivInd$unemployedCvilian/sdDemoUnivInd$Total_pop, sdDemoUnivInd$popBelowpoverty/sdDemoUnivInd$Total_pop, sdDemoUnivInd$popUninsured/sdDemoUnivInd$Total_pop, sdDemoUnivInd$univInd)
names(newSp20) <- c(listPairs)

# Summer 2020
avgPrevSu20 <- (su20CovidData[,length(su20CovidData[1,])]-su20CovidData[,2])/sdDemoUnivInd[,2]
newSu20 <- data.frame(avgPrevSu20, sdDemoUnivInd$Total_hispanic/sdDemoUnivInd$Total_pop, sdDemoUnivInd$white/sdDemoUnivInd$Total_pop, sdDemoUnivInd$black/sdDemoUnivInd$Total_pop, sdDemoUnivInd$asian/sdDemoUnivInd$Total_pop, sdDemoUnivInd$american_Indian/sdDemoUnivInd$Total_pop, sdDemoUnivInd$native_Hawaiian/sdDemoUnivInd$Total_pop, sdDemoUnivInd$unemployedCvilian/sdDemoUnivInd$Total_pop, sdDemoUnivInd$popBelowpoverty/sdDemoUnivInd$Total_pop, sdDemoUnivInd$popUninsured/sdDemoUnivInd$Total_pop, sdDemoUnivInd$univInd)
names(newSu20) <- c(listPairs)


# Fall 2020
avgPrevf20 <- (f20CovidData[,length(f20CovidData[1,])]-f20CovidData[,2])/sdDemoUnivInd[,2]
newf20 <- data.frame(avgPrevf20, sdDemoUnivInd$Total_hispanic/sdDemoUnivInd$Total_pop, sdDemoUnivInd$white/sdDemoUnivInd$Total_pop, sdDemoUnivInd$black/sdDemoUnivInd$Total_pop, sdDemoUnivInd$asian/sdDemoUnivInd$Total_pop, sdDemoUnivInd$american_Indian/sdDemoUnivInd$Total_pop, sdDemoUnivInd$native_Hawaiian/sdDemoUnivInd$Total_pop, sdDemoUnivInd$unemployedCvilian/sdDemoUnivInd$Total_pop, sdDemoUnivInd$popBelowpoverty/sdDemoUnivInd$Total_pop, sdDemoUnivInd$popUninsured/sdDemoUnivInd$Total_pop, sdDemoUnivInd$univInd)
names(newf20) <- c(listPairs)

# Winter 2021
avgPrevw21 <- (w21CovidData[,length(w21CovidData[1,])]-w21CovidData[,2])/sdDemoUnivInd[,2]
neww21 <- data.frame(avgPrevw21, sdDemoUnivInd$Total_hispanic/sdDemoUnivInd$Total_pop, sdDemoUnivInd$white/sdDemoUnivInd$Total_pop, sdDemoUnivInd$black/sdDemoUnivInd$Total_pop, sdDemoUnivInd$asian/sdDemoUnivInd$Total_pop, sdDemoUnivInd$american_Indian/sdDemoUnivInd$Total_pop, sdDemoUnivInd$native_Hawaiian/sdDemoUnivInd$Total_pop, sdDemoUnivInd$unemployedCvilian/sdDemoUnivInd$Total_pop, sdDemoUnivInd$popBelowpoverty/sdDemoUnivInd$Total_pop, sdDemoUnivInd$popUninsured/sdDemoUnivInd$Total_pop, sdDemoUnivInd$univInd)
names(neww21) <- c(listPairs)

# Spring 2021
avgPrevsp21 <- (sp21CovidData[,length(sp21CovidData[1,])]-sp21CovidData[,2])/sdDemoUnivInd[,2]
newsp21 <- data.frame(avgPrevsp21, sdDemoUnivInd$Total_hispanic/sdDemoUnivInd$Total_pop, sdDemoUnivInd$white/sD
emoUnivInd$Total_pop, sdDemoUnivInd$black/sdDemoUnivInd$Total_pop, sdDemoUnivInd$asian/sdDemoUnivInd$Total_po
p, sdDemoUnivInd$american_Indian/sdDemoUnivInd$Total_pop, sdDemoUnivInd$native_Hawaiian/sdDemoUnivInd$Total_po
p, sdDemoUnivInd$unemployedCvilian/sdDemoUnivInd$Total_pop, sdDemoUnivInd$popBelowpoverty/sdDemoUnivInd$Total
_pop, sdDemoUnivInd$popUninsured/sdDemoUnivInd$Total_pop, sdDemoUnivInd$univInd)
names(newsp21) <- c(listPairs)

# Summer 2021
avgPrevsu21 <- (su21CovidData[,length(su21CovidData[1,])]-su21CovidData[,2])/sdDemoUnivInd[,2]
newsu21 <- data.frame(avgPrevsu21, sdDemoUnivInd$Total_hispanic/sdDemoUnivInd$Total_pop, sdDemoUnivInd$white/sdD
emoUnivInd$Total_pop, sdDemoUnivInd$black/sdDemoUnivInd$Total_pop, sdDemoUnivInd$asian/sdDemoUnivInd$Total_po
p, sdDemoUnivInd$american_Indian/sdDemoUnivInd$Total_pop, sdDemoUnivInd$native_Hawaiian/sdDemoUnivInd$Total_po
p, sdDemoUnivInd$unemployedCvilian/sdDemoUnivInd$Total_pop, sdDemoUnivInd$popBelowpoverty/sdDemoUnivInd$Total
_pop, sdDemoUnivInd$popUninsured/sdDemoUnivInd$Total_pop, sdDemoUnivInd$univInd)
names(newsu21) <- c(listPairs)
```

A2: Assessing Dispersion

```
# Unconditional mean and variance for all 102 ZIP codes

means <- rep(0, 102)
variances <- rep(0, 102)

for (i in c(1:102)){
  means[i] <- mean(covidData[i,-c(1)])
  variances[i] <- var(covidData[i,-c(1)])
}
poi_verify <- cbind(means, variances)
head(poi_verify) # The results rule out Poisson regression at its simplest. Perhaps a more complex version of it will work?
sum((poi_verify[,1] > poi_verify[,2]), na.rm=TRUE) # 0/102 cases where mean is greater than variance; evidence of overdispers
ion. But what happens when I look at each season?

# Conditional mean and variance for all 102 ZIP codes, conditioning on season

# SPRING 2020
cond_means_sp20 <- rep(0, 102)
cond_var_sp20 <- rep(0, 102)

for (i in c(1:102)){
  cond_means_sp20[i] <- mean(sp20CovidData[i,-c(1)])
  cond_var_sp20[i] <- var(sp20CovidData[i,-c(1)])
}
poi_verify_sp20 <- as.matrix(cbind(cond_means_sp20, cond_var_sp20))
sum((poi_verify_sp20[,1] > poi_verify_sp20[,2]), na.rm=TRUE) # 17/102 cases where mean is greater than variance; evidence
of underdispersion

# SUMMER 2020
cond_means_su20 <- rep(0, 102)
cond_var_su20 <- rep(0, 102)

for (i in c(1:102)){
  cond_means_su20[i] <- mean(su20CovidData[i,-c(1)])
  cond_var_su20[i] <- var(su20CovidData[i,-c(1)])
}
poi_verify_su20 <- as.matrix(cbind(cond_means_su20, cond_var_su20))
sum((poi_verify_su20[,1] > poi_verify_su20[,2]), na.rm=TRUE) # 10/102 cases where mean is greater than variance; evidence
of underdispersion

# FALL 2020
cond_means_f20 <- rep(0, 102)
cond_var_f20 <- rep(0, 102)

for (i in c(1:102)){
  cond_means_f20[i] <- mean(f20CovidData[i,-c(1)])
  cond_var_f20[i] <- var(f20CovidData[i,-c(1)])
}
poi_verify_f20 <- cbind(cond_means_f20, cond_var_f20)
sum((poi_verify_f20[,1] > poi_verify_f20[,2]), na.rm=TRUE) # 7/102 cases where mean is greater than variance; evidence of underdispersion

# WINTER 2021
cond_means_w21 <- rep(0, 102)
cond_var_w21 <- rep(0, 102)

for (i in c(1:102)){
  cond_means_w21[i] <- mean(w21CovidData[i,-c(1)])
  cond_var_w21[i] <- var(w21CovidData[i,-c(1)])
}
poi_verify_w21 <- cbind(cond_means_w21, cond_var_w21)
sum((poi_verify_w21[,1] > poi_verify_w21[,2]), na.rm=TRUE) # 7/102 cases where mean is greater than variance; evidence of underdispersion

# SPRING 2021
cond_means_sp21 <- rep(0, 102)
cond_var_sp21 <- rep(0, 102)

for (i in c(1:102)){
  cond_means_sp21[i] <- mean(sp21CovidData[i,-c(1)])
  cond_var_sp21[i] <- var(sp21CovidData[i,-c(1)])
}
poi_verify_sp21 <- cbind(cond_means_sp21, cond_var_sp21)
sum((poi_verify_sp21[,1] > poi_verify_sp21[,2]), na.rm=TRUE) # 41/102 cases where mean is greater than variance; evidence of underdispersion

# SUMMER 2021
cond_means_su21 <- rep(0, 102)
cond_var_su21 <- rep(0, 102)

for (i in c(1:102)){
  cond_means_su21[i] <- mean(su21CovidData[i,-c(1)])
  cond_var_su21[i] <- var(su21CovidData[i,-c(1)])
}
poi_verify_su21 <- cbind(cond_means_su21, cond_var_su21)
sum((poi_verify_su21[,1] > poi_verify_su21[,2]), na.rm=TRUE) # 101/102 cases where mean is greater than variance; evidence of overdispersion
```

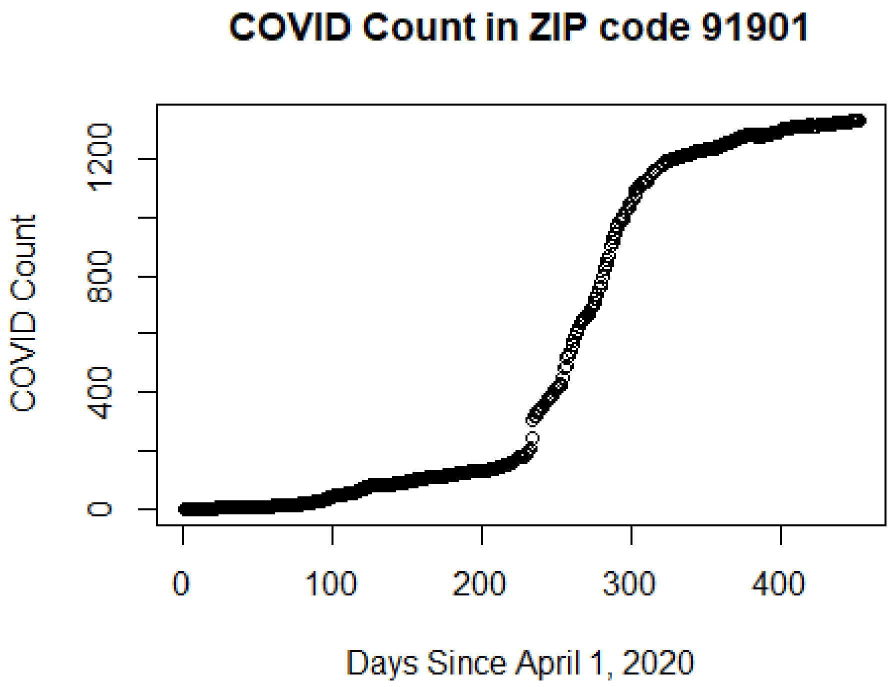
A3: Code for COVID Graphs and Other EDA

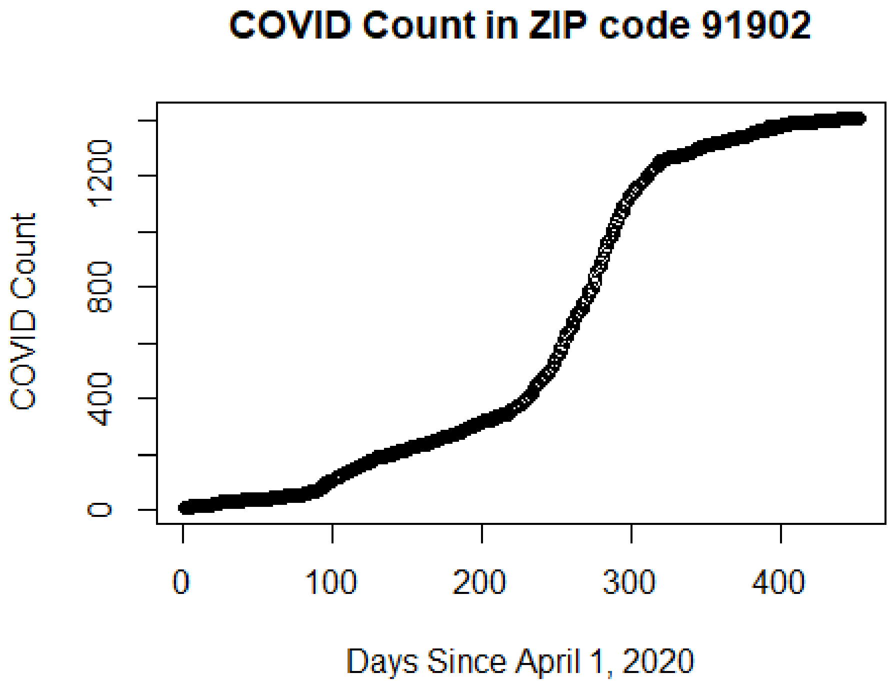
```
library(flextable)
library(modelsummary)
modelsummary::datasummary_skim(newWhole)
```

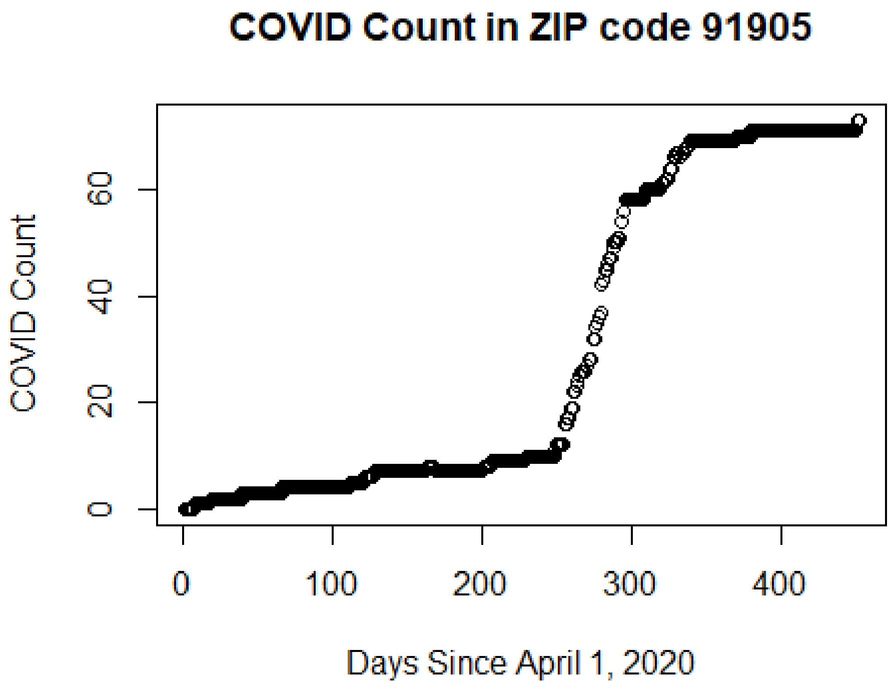
```
## Warning in datasummary_skim_numeric(data, output = output, fmt = fmt, histogram
## = histogram, : The histogram argument is only supported for (a) output types
## "default", "html", or "kableExtra"; (b) writing to file paths with extensions
## ".html", ".jpg", or ".png"; and (c) Rmarkdown or knitr documents compiled to PDF
## or HTML. Use `histogram=FALSE` to silence this warning.
```

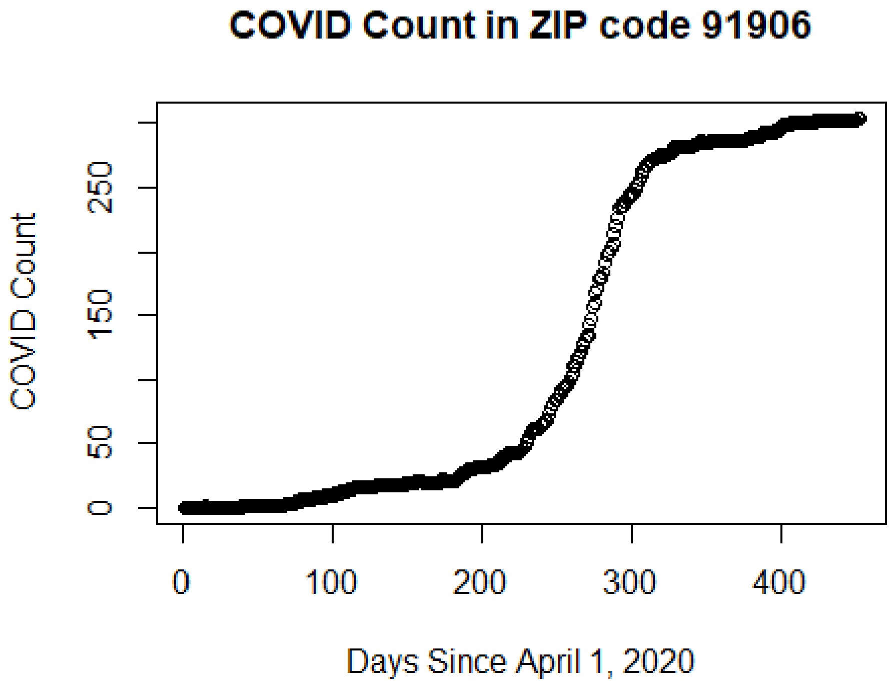
	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max
Avg. Prev.	102	0	0.1	0.0	0.0	0.1	0.3
% Hisp.	102	0	0.3	0.2	0.0	0.2	1.0
% White	101	0	0.7	0.1	0.3	0.8	1.0
% Black	98	0	0.0	0.0	0.0	0.0	0.3
% Asian	95	0	0.1	0.1	0.0	0.1	0.5
% Am. Ind.	91	0	0.0	0.1	0.0	0.0	0.4
% Nat. Hawaiian	78	0	0.0	0.0	0.0	0.0	0.1
% Unemp.	97	0	0.0	0.0	0.0	0.0	0.1
% Below Pov.	100	0	0.0	0.0	0.0	0.0	0.1
% Unins.	98	0	0.1	0.1	0.0	0.1	0.3
Univ. Ind.	2	0	0.2	0.4	0.0	0.0	1.0

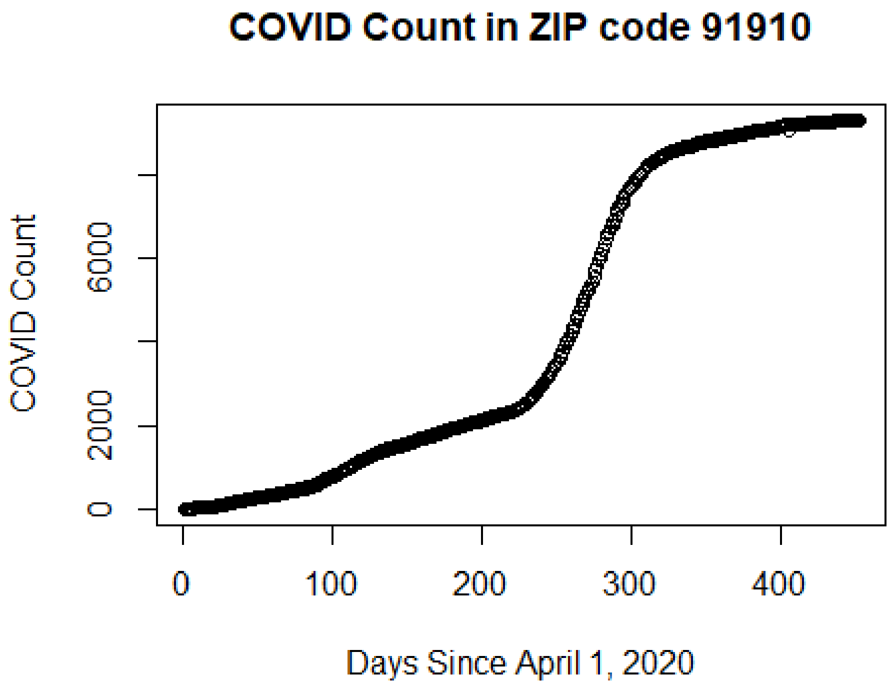
```
# COVID counts over elapsed time
for (z in 1:6){ # 1:length(covidData[,1])
  plot(covidData[z,2:453], main = paste("COVID Count in ZIP code", covidData[z,1]), xlab = "Days Since April 1, 2020", ylab =
"COVID Count")
}
```

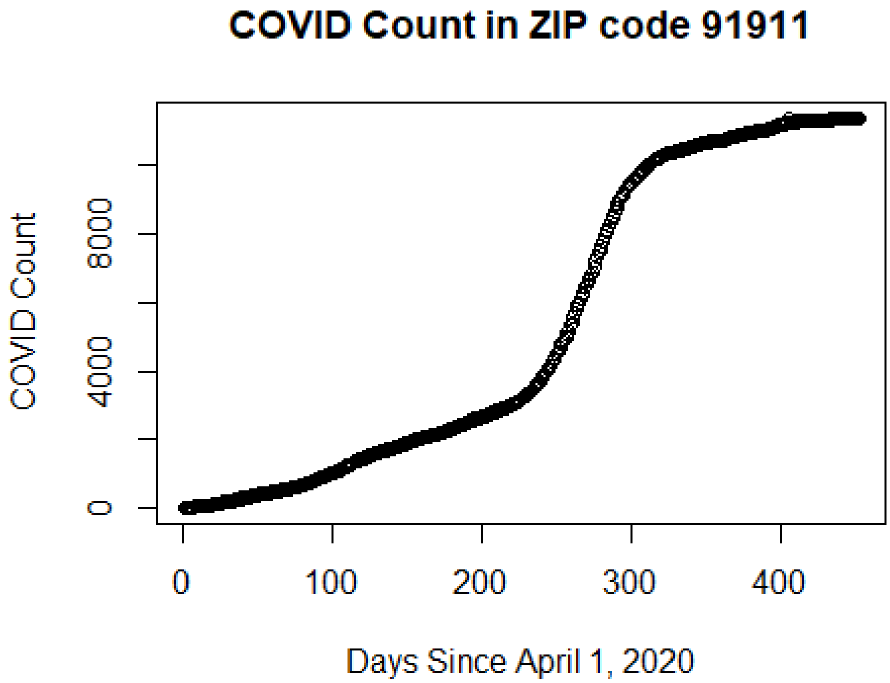

COVID Count in ZIP code 91901
COVID Count
0
400
800
1200
0
100
200
300
400
Days Since April 1, 2020


COVID Count in ZIP code 91902
COVID Count
0
400
800
1200
0
100
200
300
400
Days Since April 1, 2020

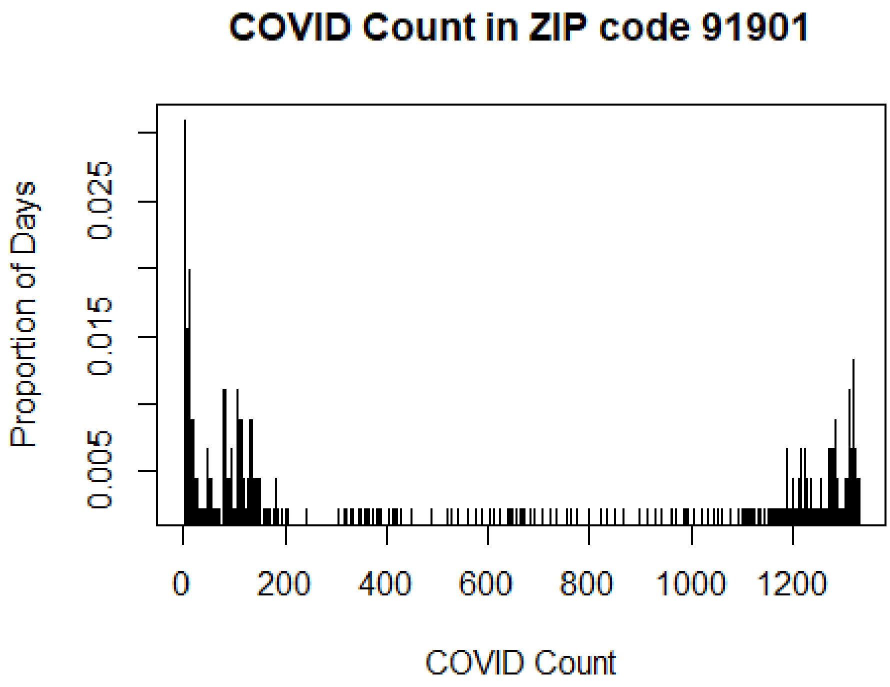

COVID Count in ZIP code 91905
COVID Count
0
20
40
60
0
100
200
300
400
Days Since April 1, 2020

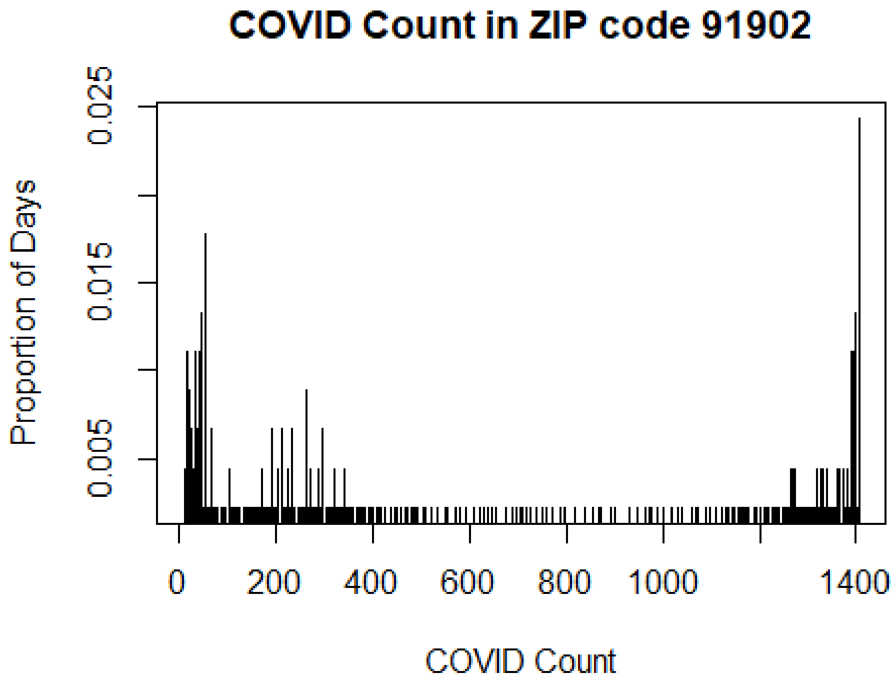

COVID Count in ZIP code 91906
COVID Count
0
50
150
250
0
100
200
300
400
Days Since April 1, 2020

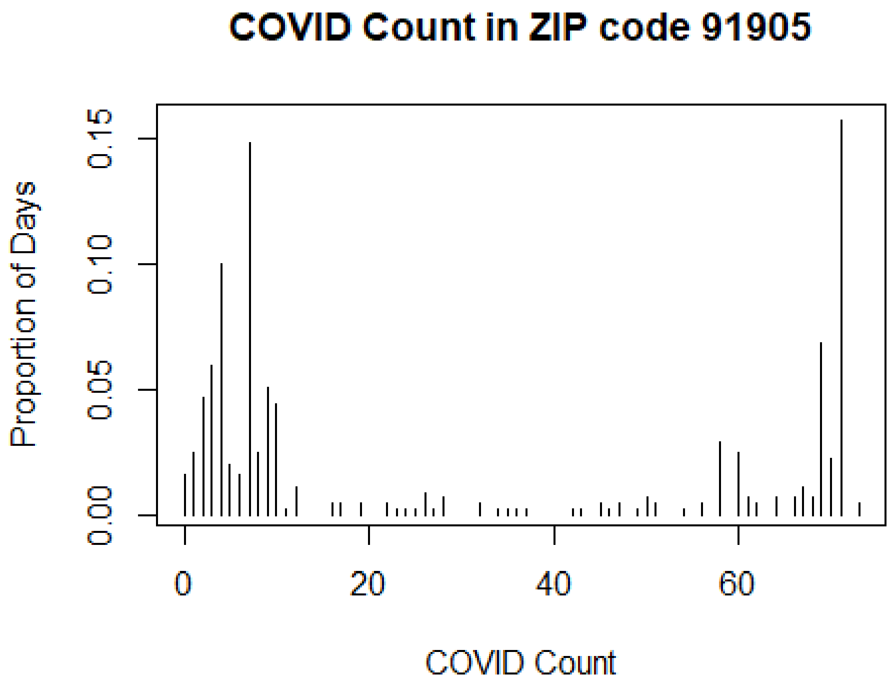


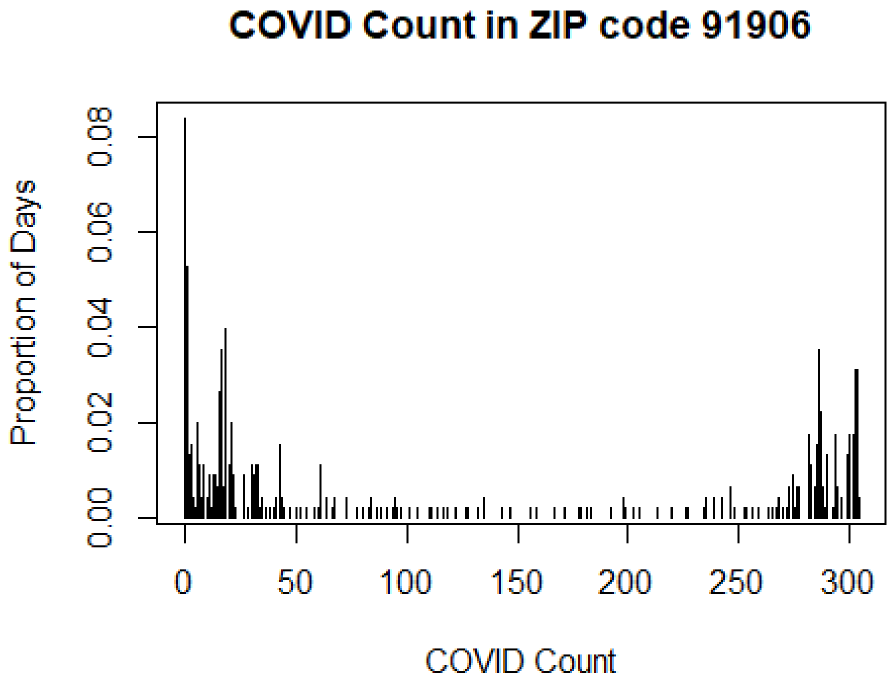


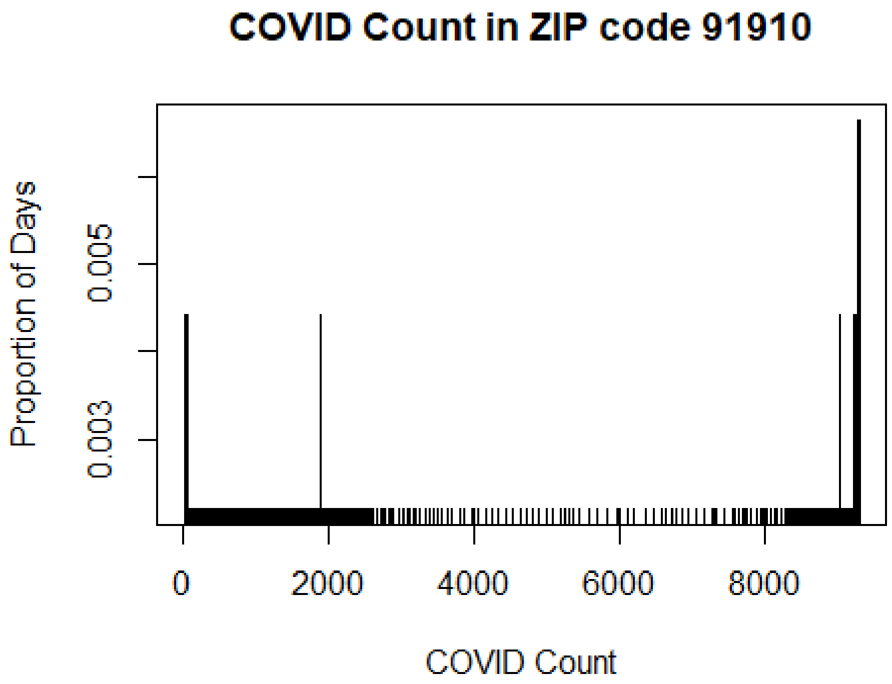
```
# COVID, no conditioning on season, counts
# need VGAM package for spike plot
library(VGAM)

## Loading required package: stats4
## Loading required package: splines

##
## Attaching package: 'VGAM'

## The following object is masked from 'package:modelsummary':
##
##     Max

for (z in 1:6){ # 1:length(covidData[,1])
  spikeplot(covidData[z,2:453], main = paste("COVID Count in ZIP code", covidData[z,1]), xlab = "COVID Count", ylab = "Proportion of Days")
}
```

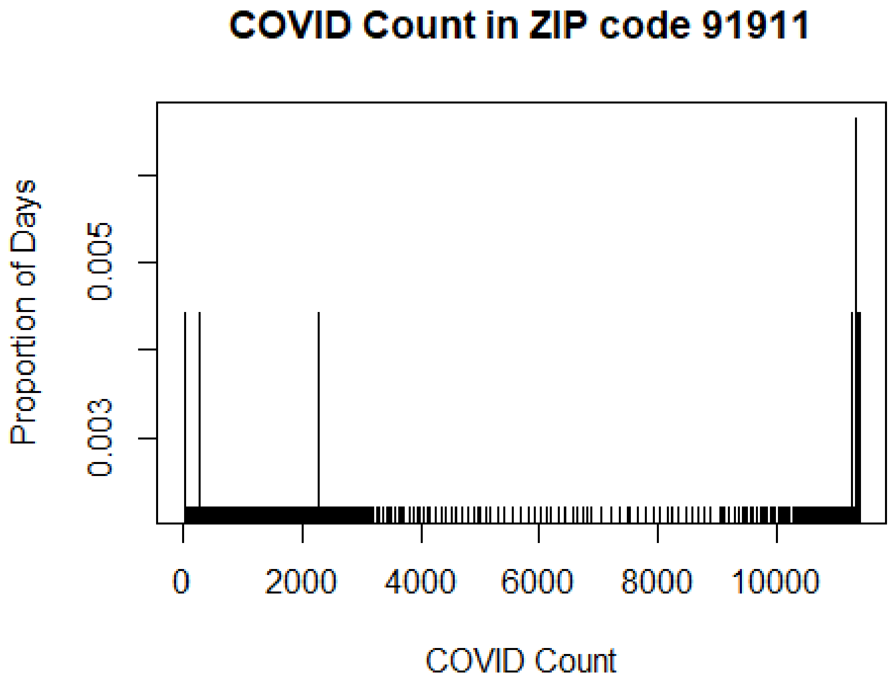

COVID Count in ZIP code 91901
Proportion of Days
0.005
0.015
0.025
0
200
400
600
800
1000
1200
COVID Count


COVID Count in ZIP code 91902
Proportion of Days
0.005
0.015
0.025
0
200
400
600
800
1000
1400
COVID Count

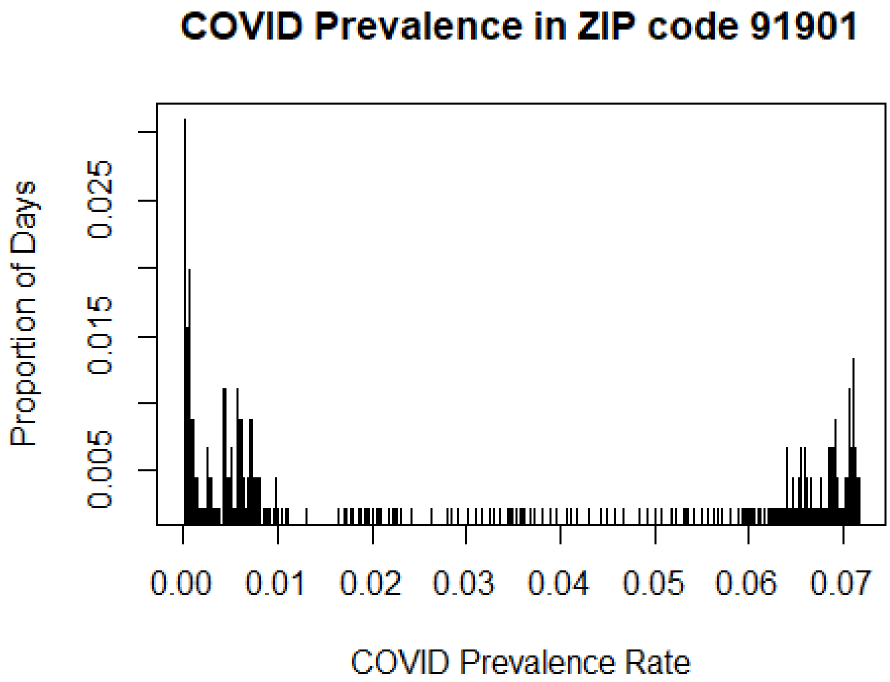

COVID Count in ZIP code 91905
Proportion of Days
0.00
0.05
0.10
0.15
0
20
40
60
COVID Count

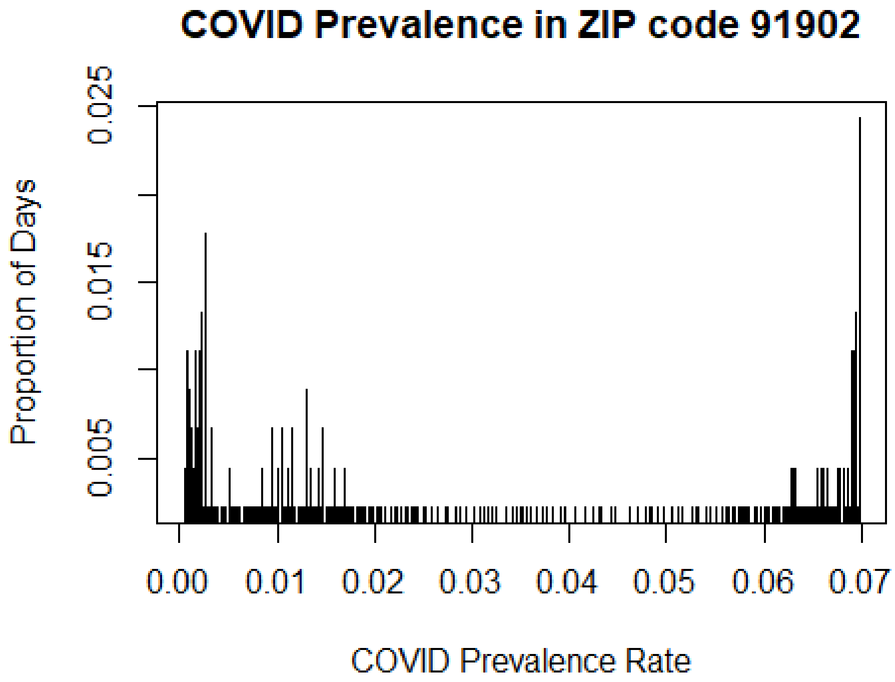

COVID Count in ZIP code 91906
Proportion of Days
0.00
0.02
0.04
0.06
0.08
0
50
100
150
200
250
300
COVID Count

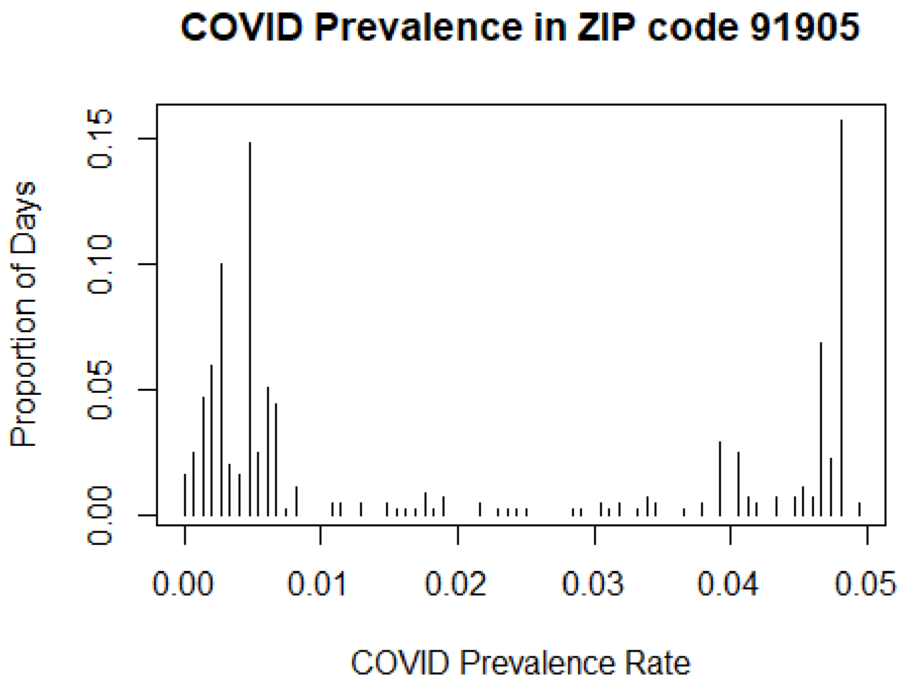


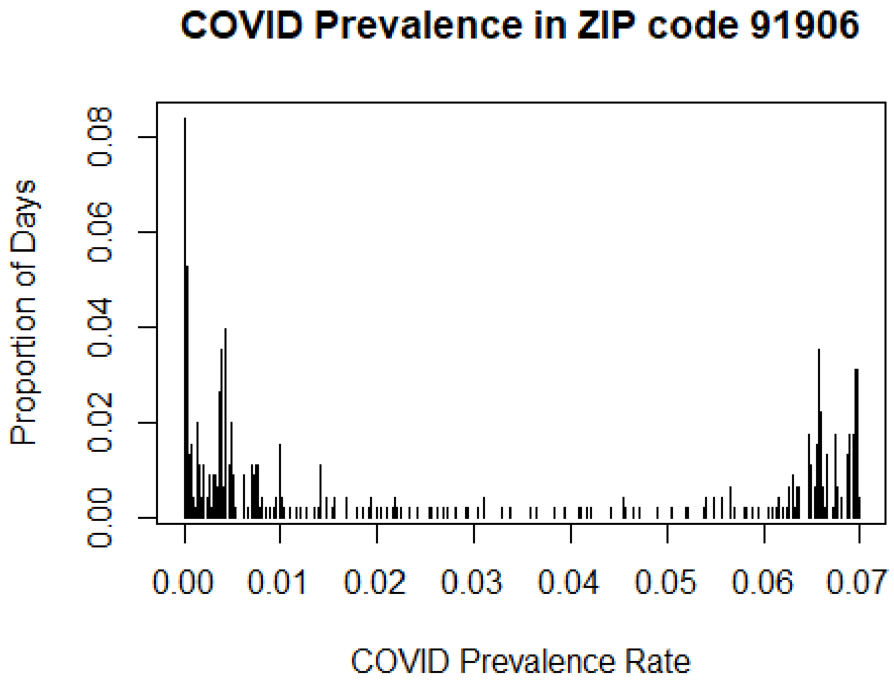


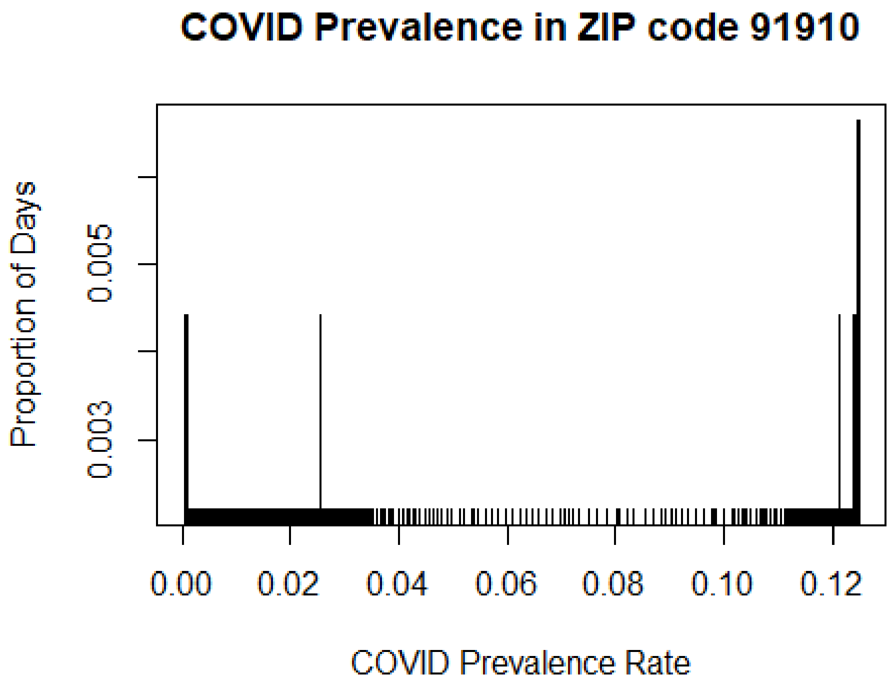
```
# Same, but as rate of COVID, or prevalence

for (z in 1:6){ # 1:length(covidData[,1])
  spikeplot(covidData[z,2:453]/sdDemo[z,2], main = paste("COVID Prevalence in ZIP code", covidData[z,1]), xlab = "COVID P
revalence Rate", ylab = "Proportion of Days")
}
```

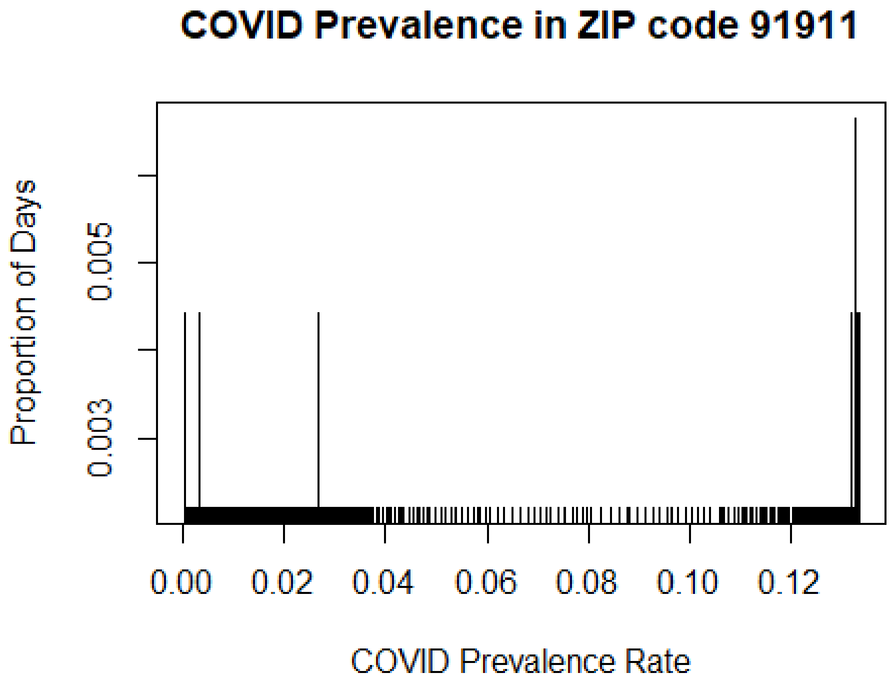




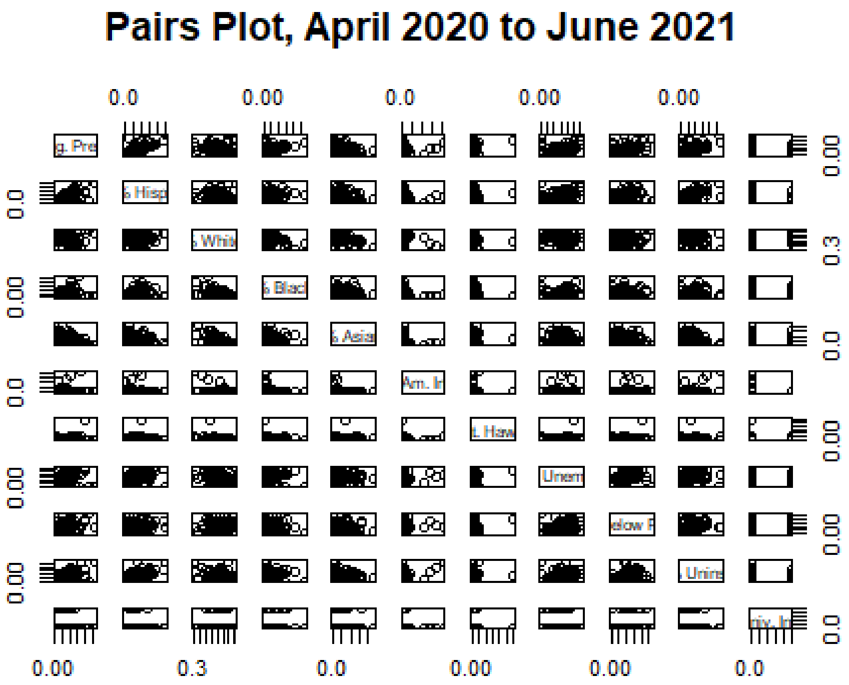

COVID Prevalence in ZIP code 91905
Proportion of Days
0.00
0.05
0.10
0.15
0.00
0.01
0.02
0.03
0.04
0.05
COVID Prevalence Rate


COVID Prevalence in ZIP code 91906
Proportion of Days
0.00
0.02
0.04
0.06
0.08
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
COVID Prevalence Rate

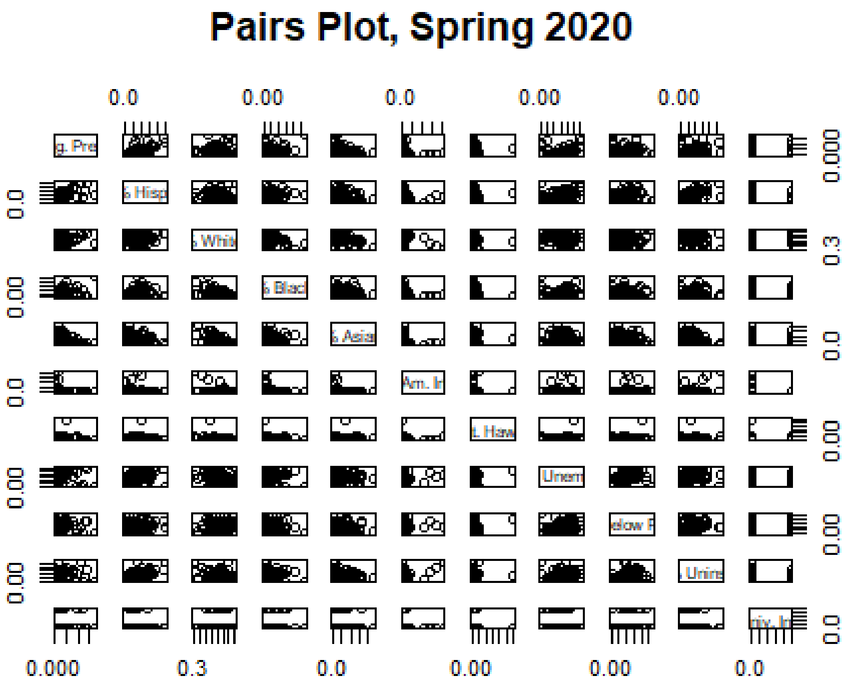




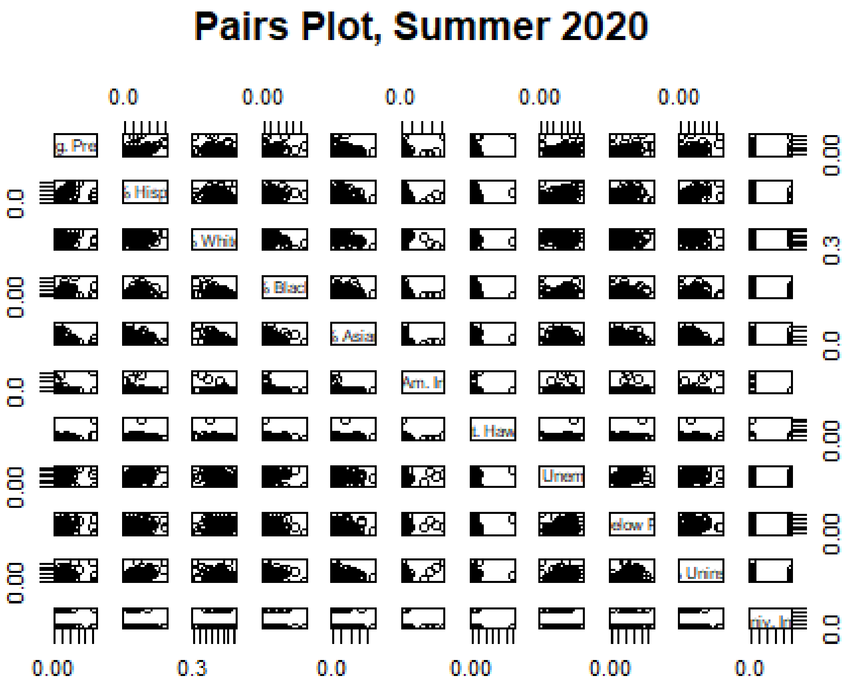
```
# Show 1:6 for Appendix.

# COVID Prevalence, unconditional of season (looks at the entire pandemic)
pairs(newWhole, main = "Pairs Plot, April 2020 to June 2021")
```



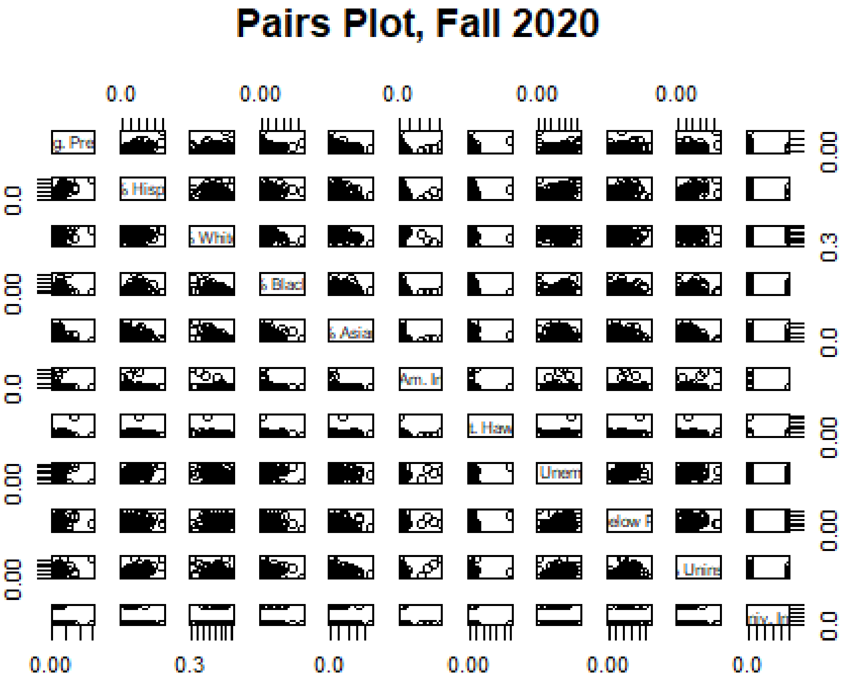
```
# COVID Prevalence, conditioning on season
# look at explanatory variables (ethnicity, unemployed, uninsured, etc.)

# Spring 2020
pairs(newSp20, main = "Pairs Plot, Spring 2020")
```


Pairs Plot, Spring 2020

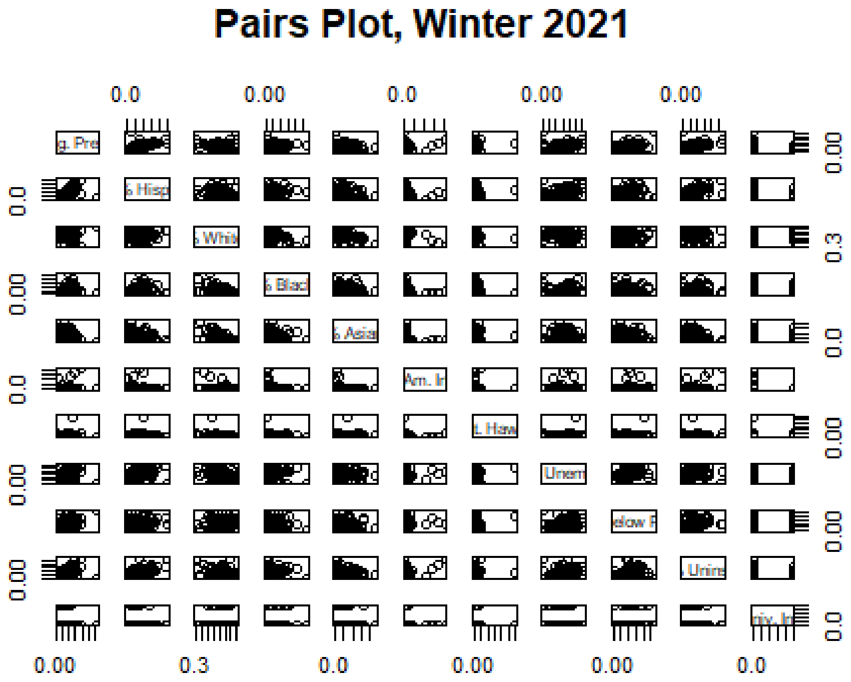
```
# Summer 2020
pairs(newSu20, main = "Pairs Plot, Summer 2020")
```


Pairs Plot, Summer 2020

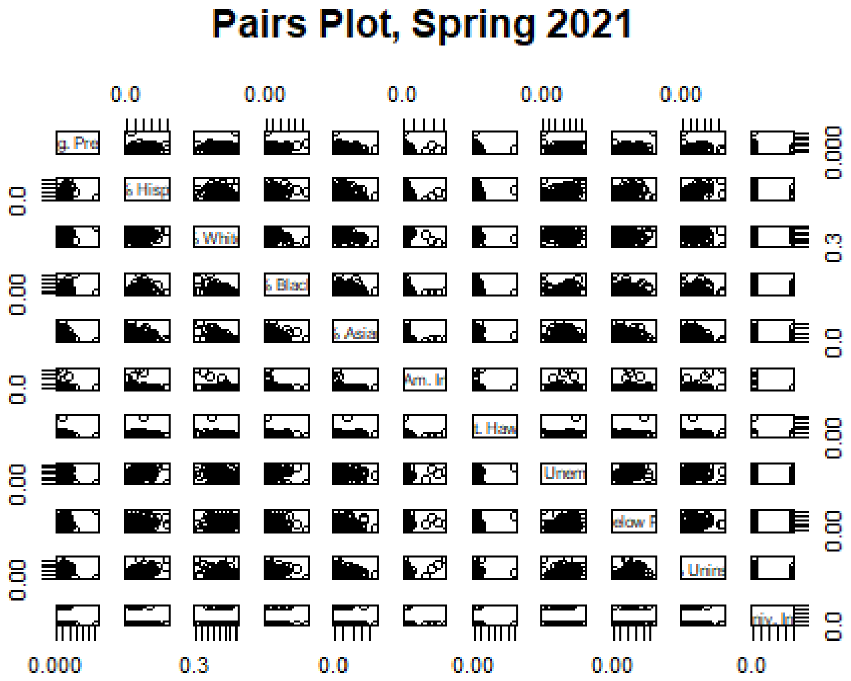
```
# Fall 2020
pairs(newf20, main = "Pairs Plot, Fall 2020")
```



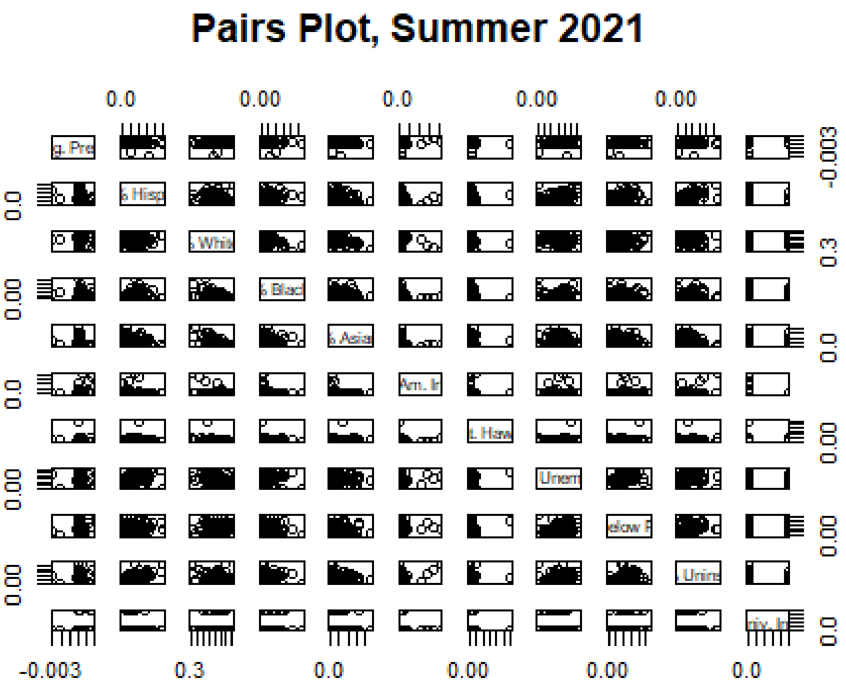
```
# Winter 2021
pairs(neww21, main = "Pairs Plot, Winter 2021")
```



```
# Spring 2021
pairs(newsp21, main = "Pairs Plot, Spring 2021")
```

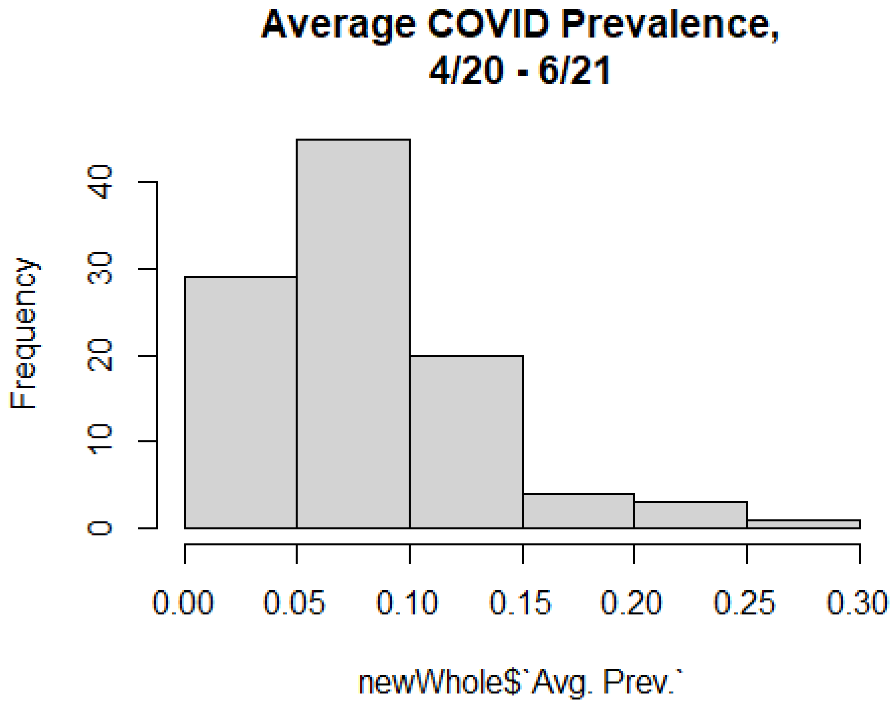


```
# Summer 2021
pairs(newsu21, main = "Pairs Plot, Summer 2021")
```

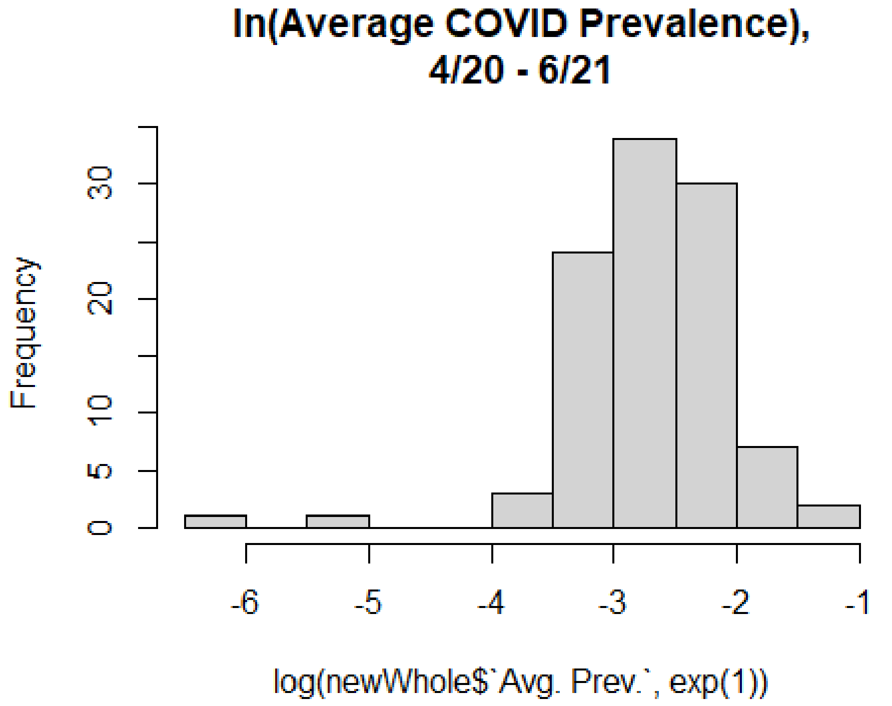


A4: EDA Involving Transformation

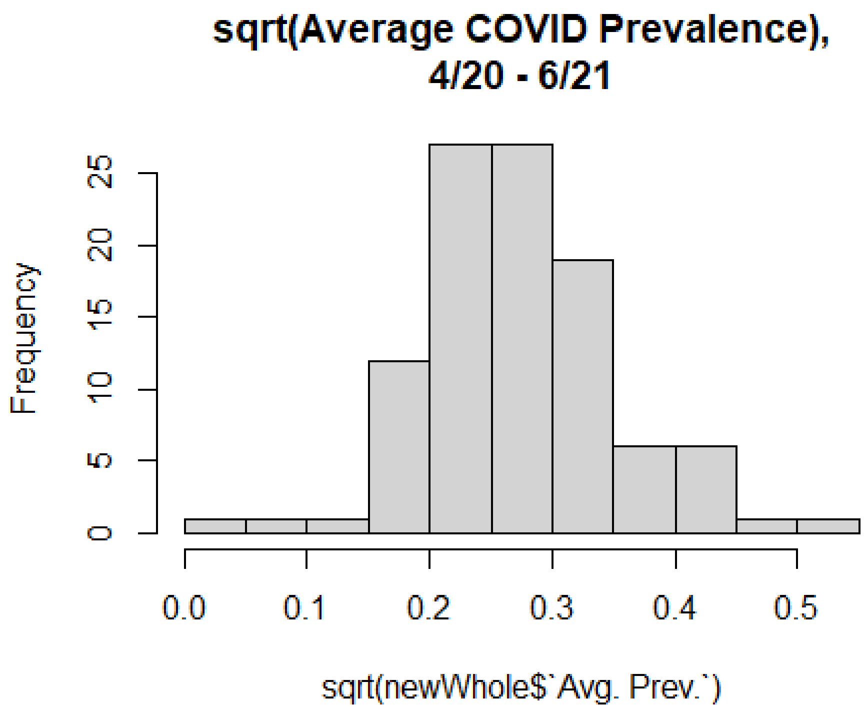
```
# EDA for log-transformed and square root-transformed variables, including average COVID prevalence

# Unconditional of season
hist(newWhole$`Avg. Prev.`, main = "Average COVID Prevalence,\n4/20 - 6/21")
```



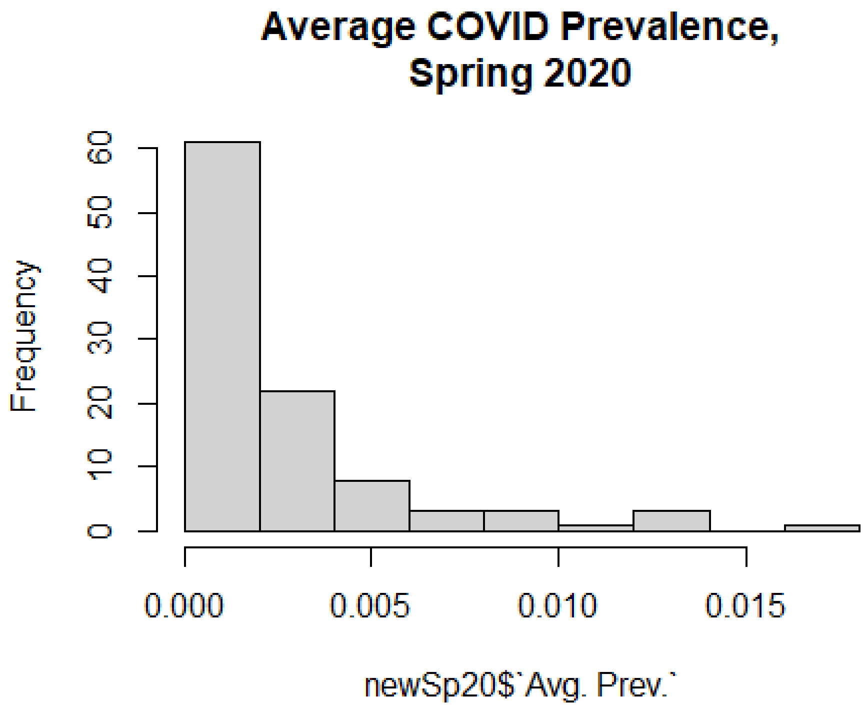
```
hist(log(newWhole$`Avg. Prev.`, exp(1)), main = "ln(Average COVID Prevalence),\n4/20 - 6/21")
```



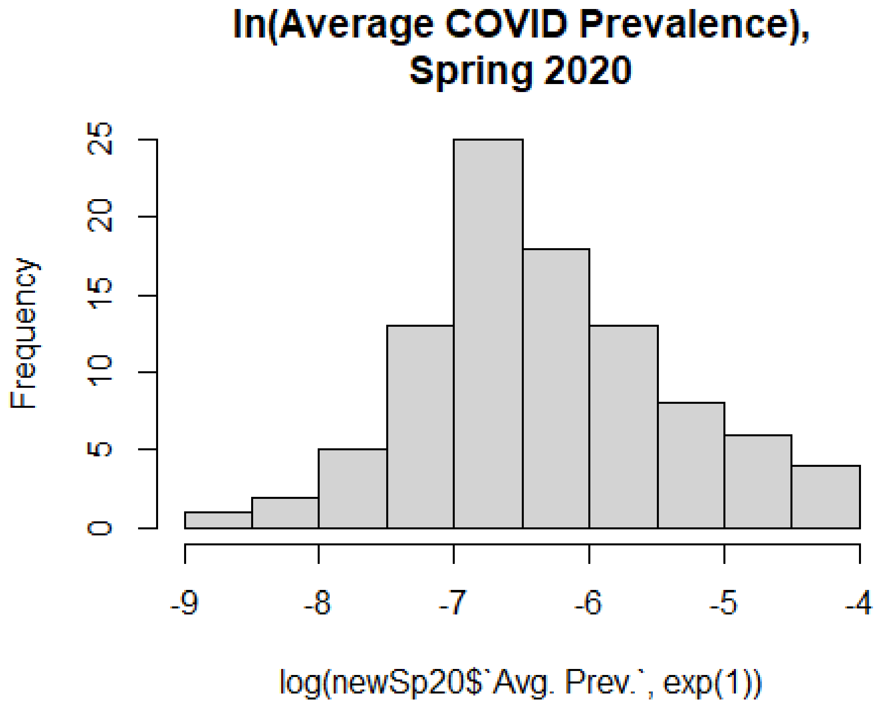
```
hist(sqrt(newWhole$`Avg. Prev.`), main = "sqrt(Average COVID Prevalence),\n4/20 - 6/21")
```



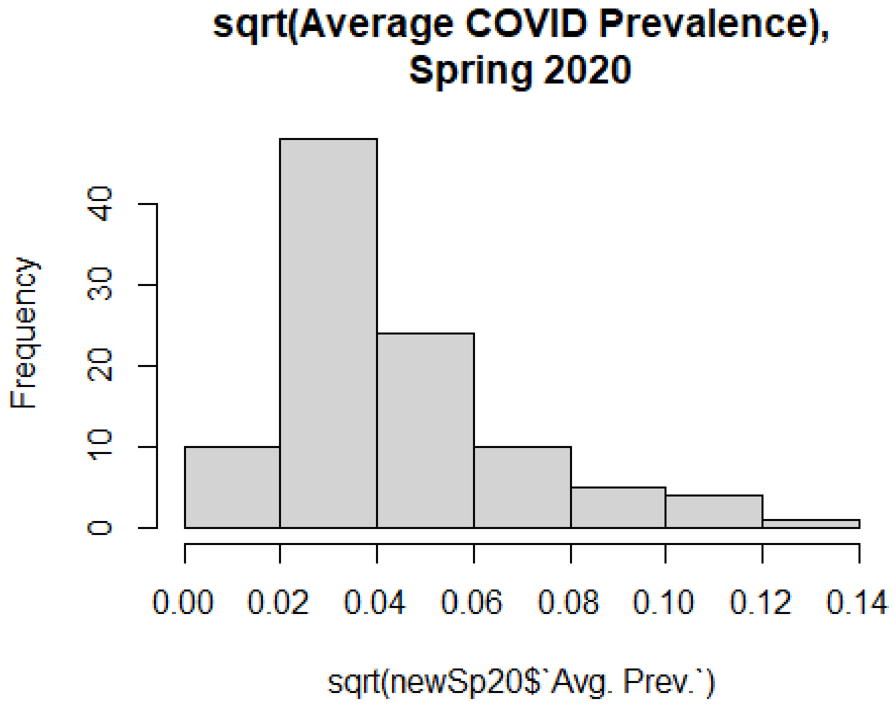
```
# Sp20
hist(newSp20$`Avg. Prev.`, main = "Average COVID Prevalence,\nSpring 2020")
```



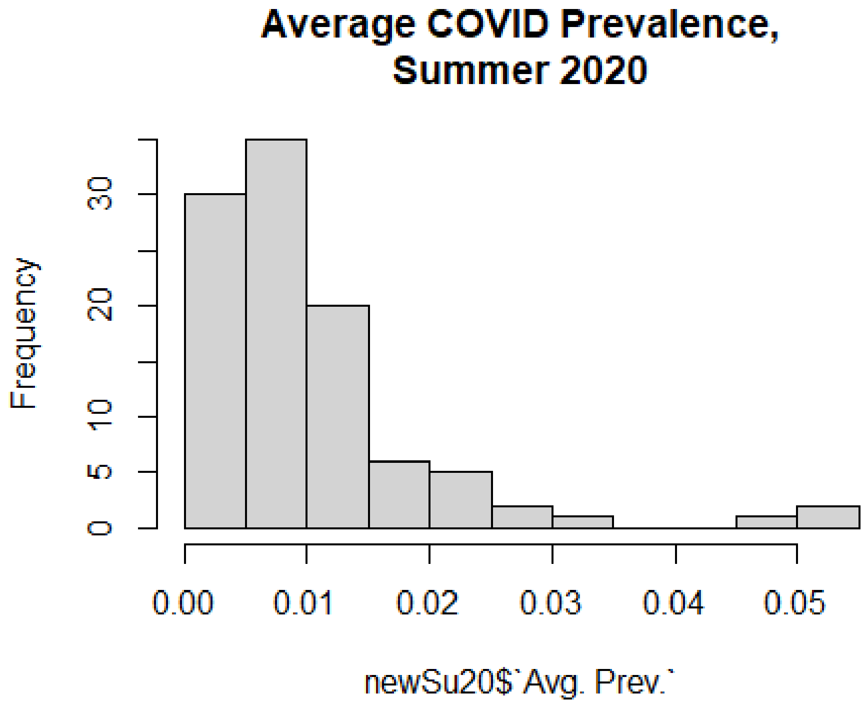
```
hist(log(newSp20$`Avg. Prev.`, exp(1)), main = "ln(Average COVID Prevalence),\nSpring 2020")
```



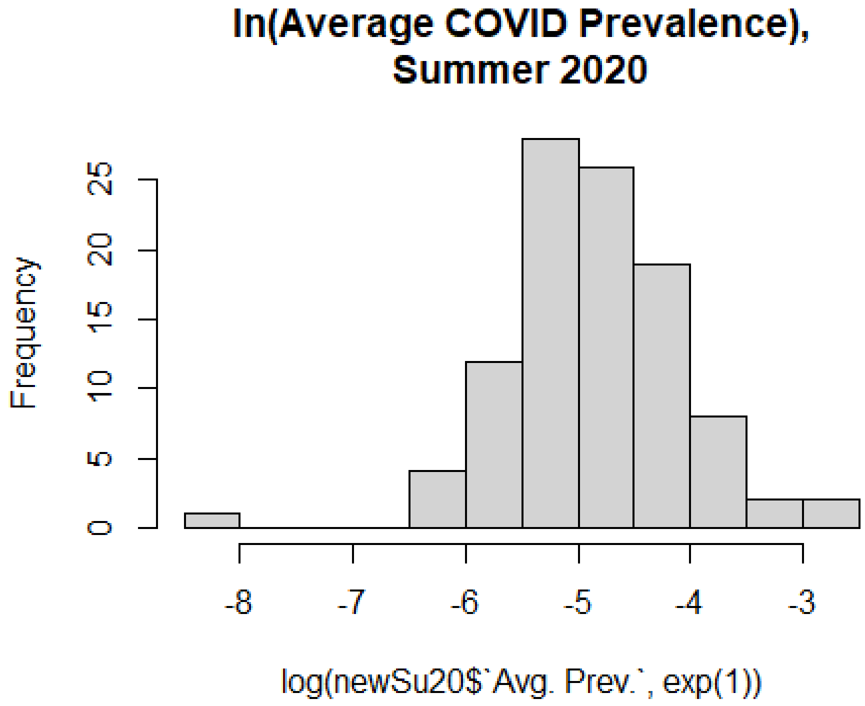
```
hist(sqrt(newSp20$`Avg. Prev.`), main = "sqrt(Average COVID Prevalence),\nSpring 2020")
```



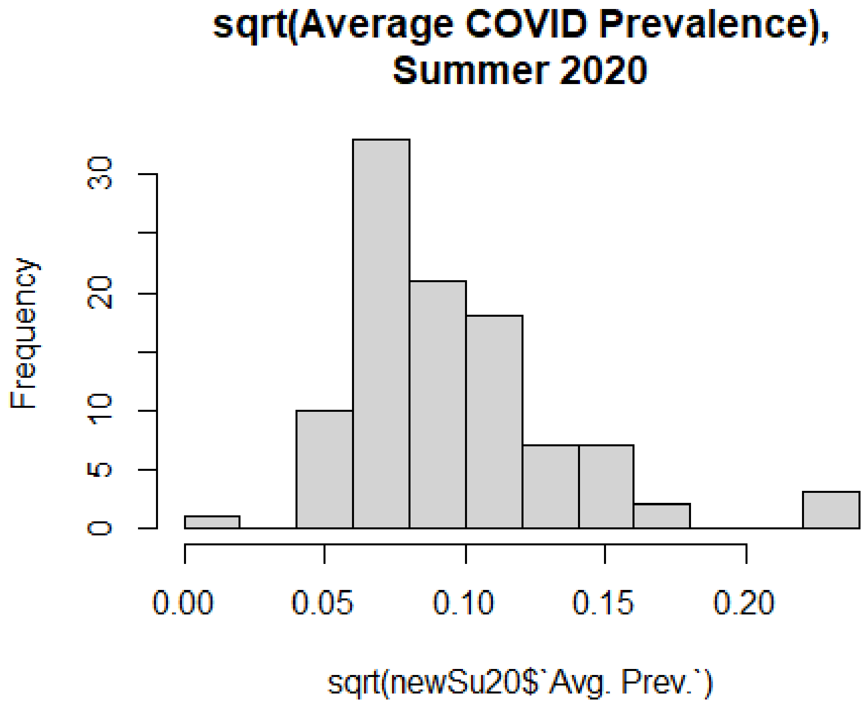
```
# Su20
hist(newSu20$`Avg. Prev.`, main = "Average COVID Prevalence,\nSummer 2020")
```



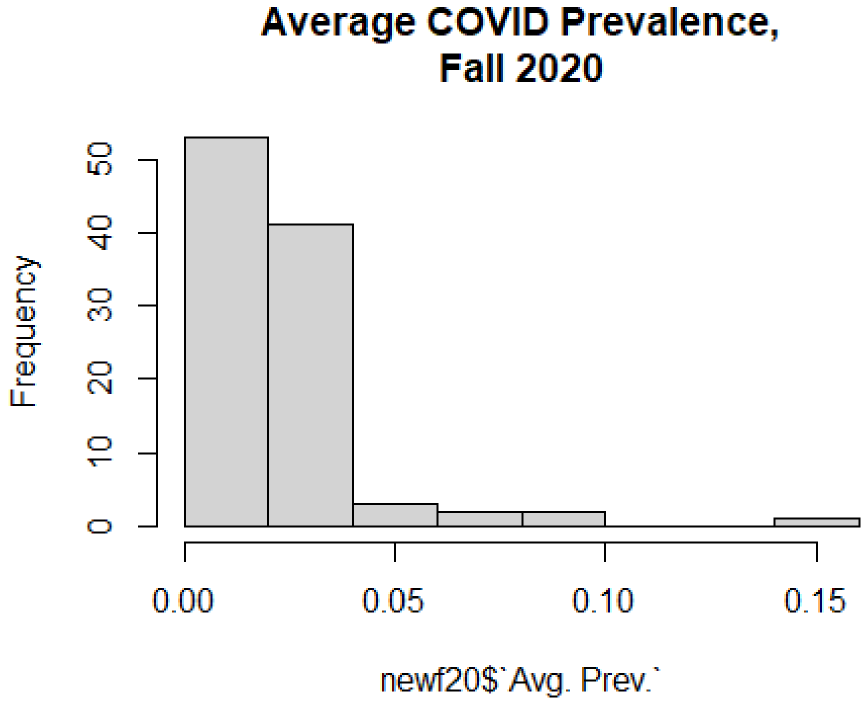
```
hist(log(newSu20$`Avg. Prev.`, exp(1)), main = "ln(Average COVID Prevalence),\nSummer 2020")
```



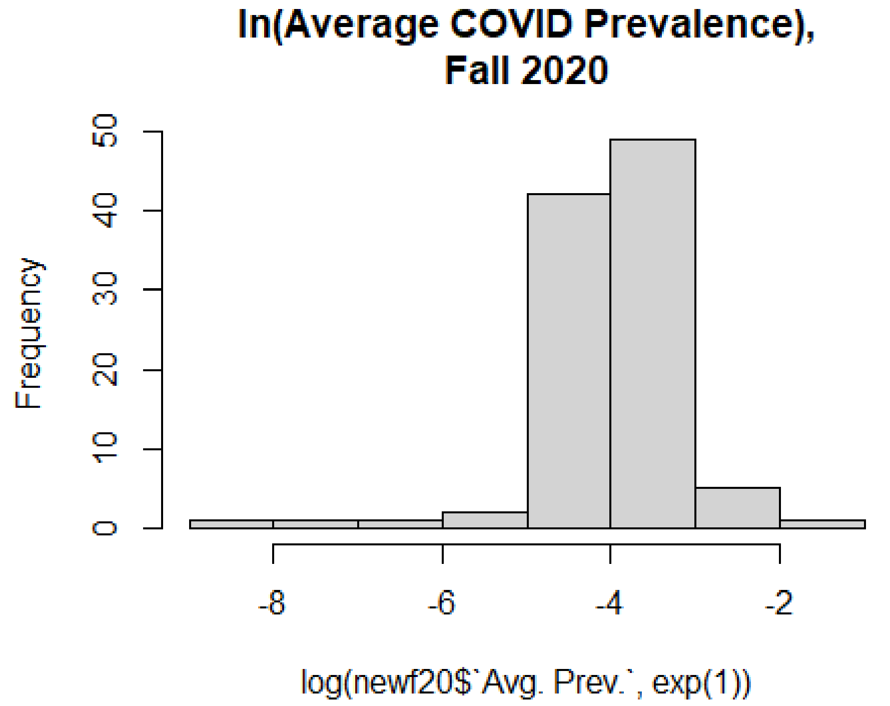
```
hist(sqrt(newSu20$`Avg. Prev.`), main = "sqrt(Average COVID Prevalence),\nSummer 2020")
```



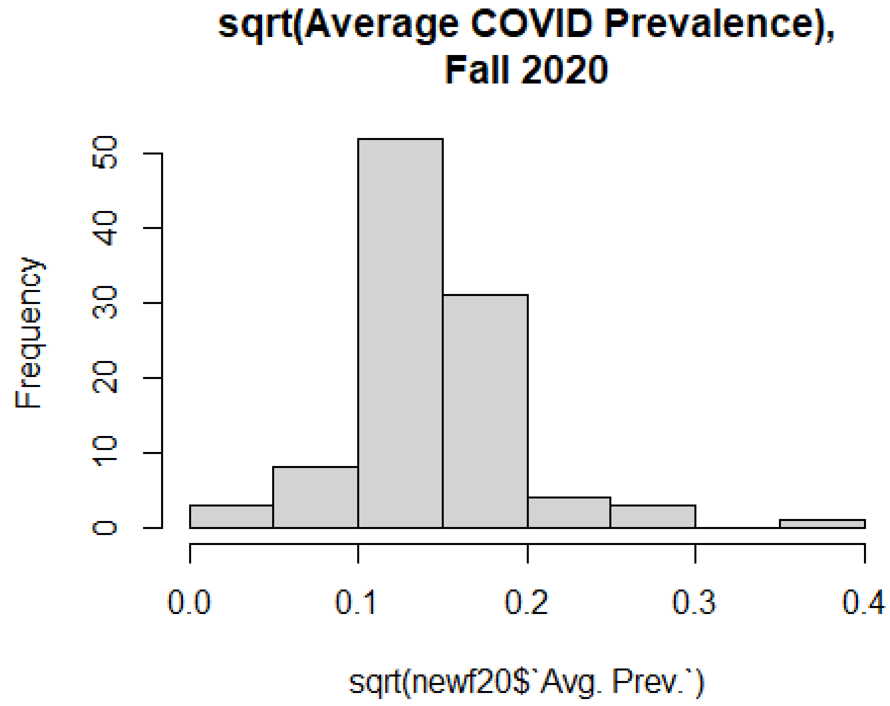
```r
# F20
hist(newf20$`Avg. Prev.`, main = "Average COVID Prevalence,\nFall 2020")
```



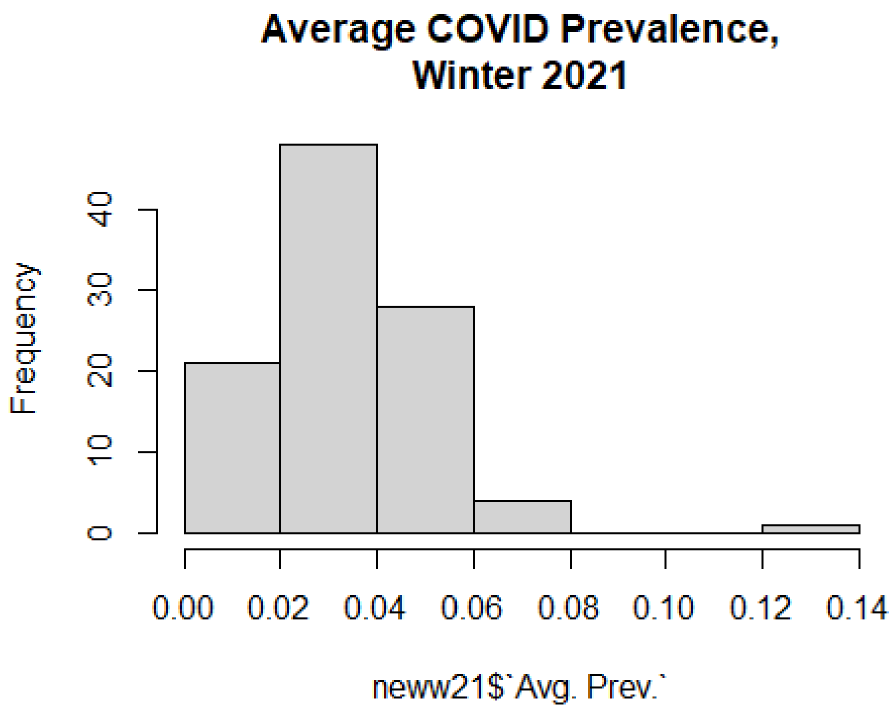
```r
hist(log(newf20$`Avg. Prev.`, exp(1)), main = "ln(Average COVID Prevalence),\nFall 2020")
```



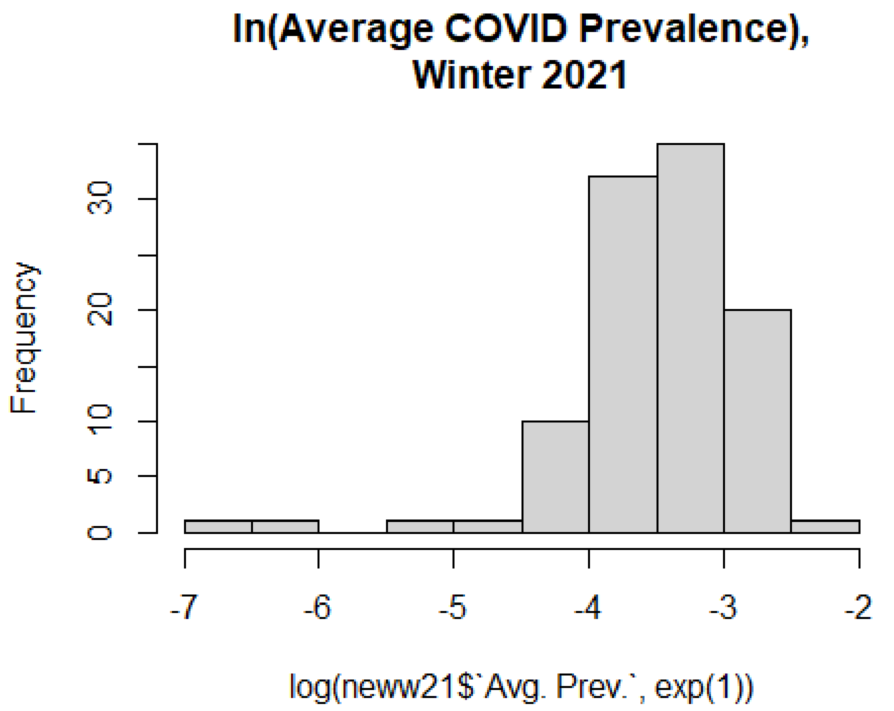
```
hist(sqrt(newf20$`Avg. Prev.`), main = "sqrt(Average COVID Prevalence),\nFall 2020")
```



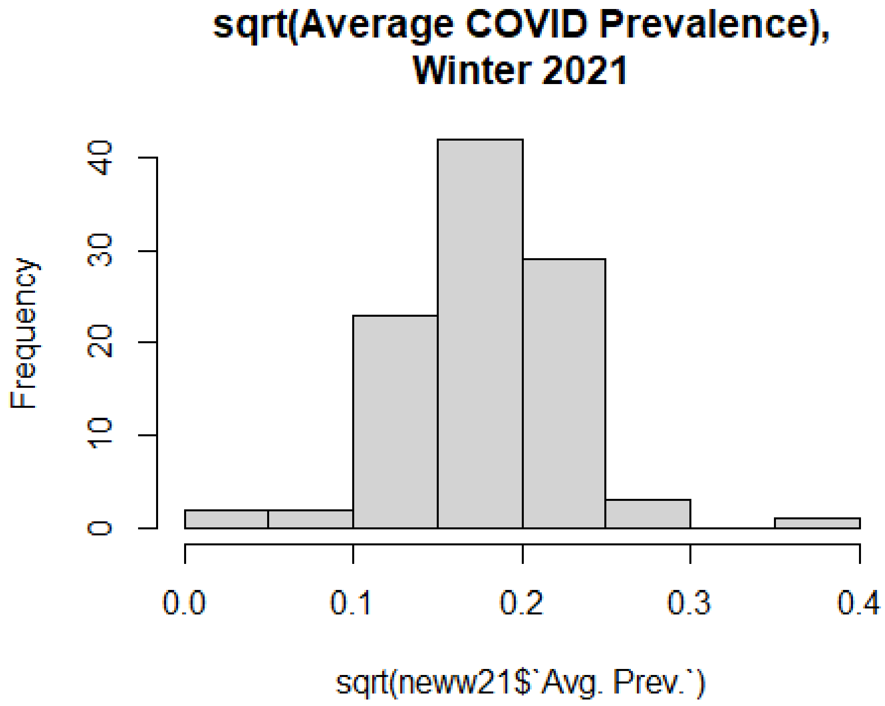
```
# W21
hist(neww21$`Avg. Prev.`, main = "Average COVID Prevalence,\nWinter 2021")
```



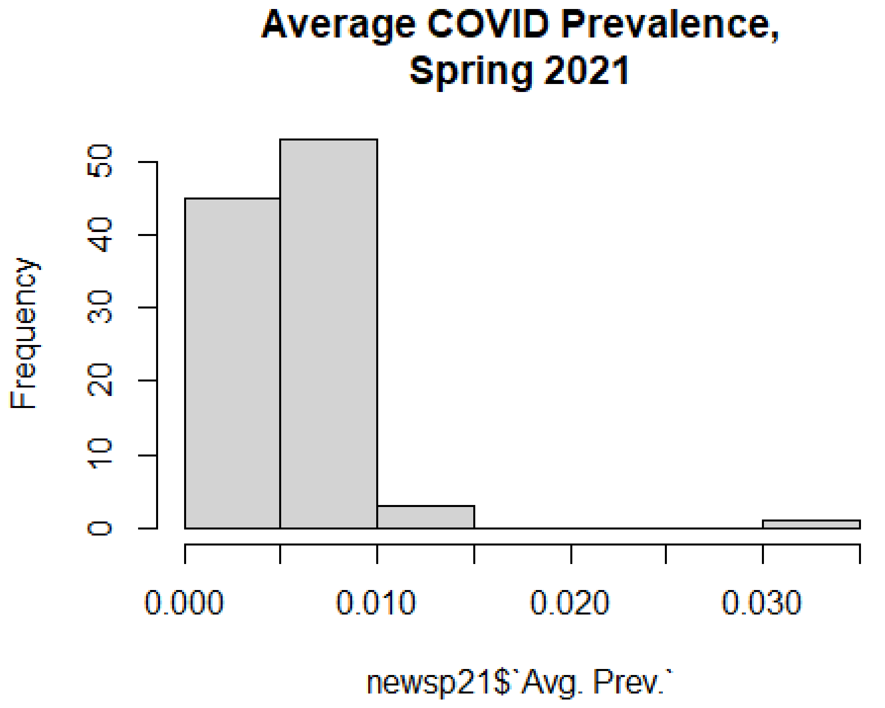
```
hist(log(neww21$`Avg. Prev.`, exp(1)), main = "ln(Average COVID Prevalence),\nWinter 2021")
```



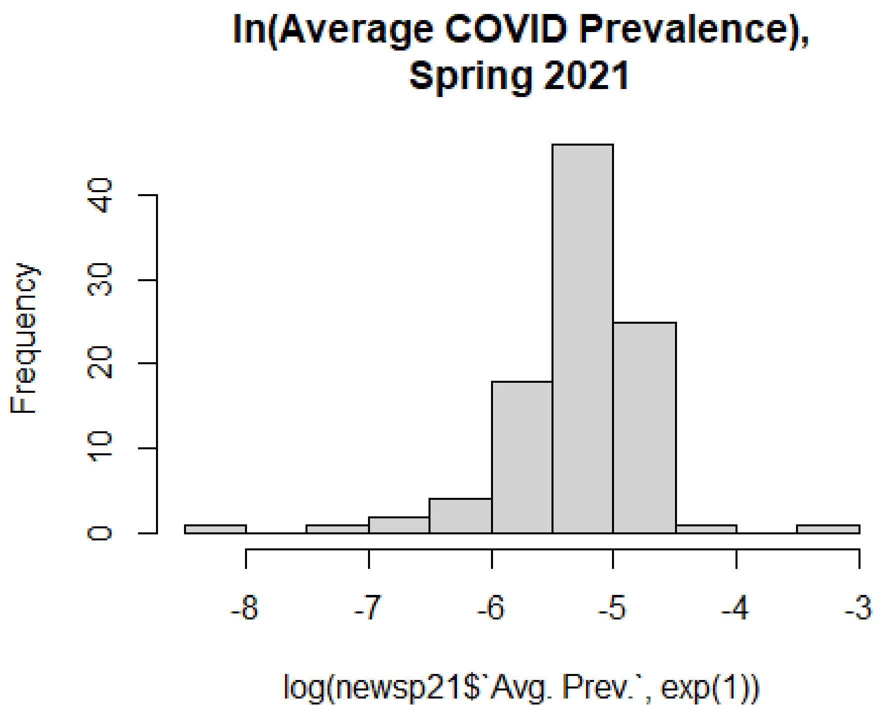
```
hist(sqrt(neww21$`Avg. Prev.`), main = "sqrt(Average COVID Prevalence),\nWinter 2021")
```



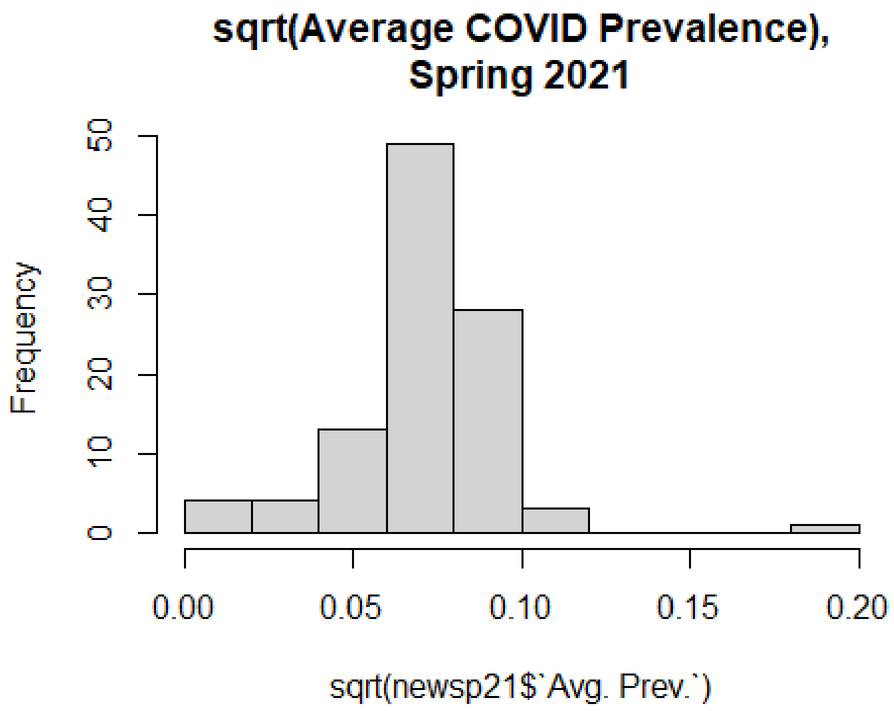
```
# Sp21
hist(newsp21$`Avg. Prev.`, main = "Average COVID Prevalence,\nSpring 2021")
```



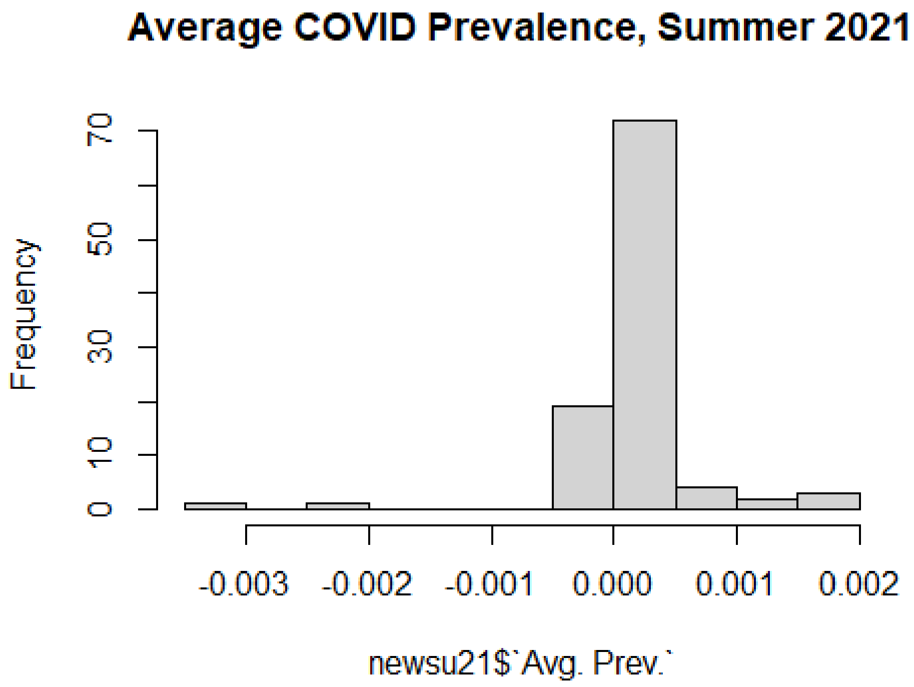
```
hist(log(newsp21$`Avg. Prev.`, exp(1)), main = "ln(Average COVID Prevalence),\nSpring 2021")
```



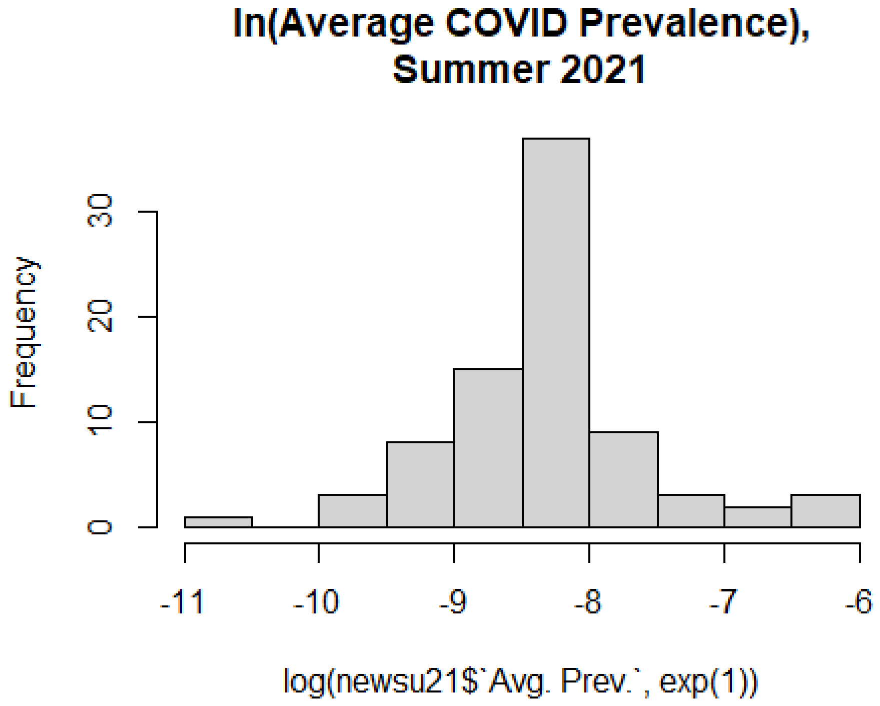
```
hist(sqrt(newsp21$`Avg. Prev.`), main = "sqrt(Average COVID Prevalence),\nSpring 2021")
```



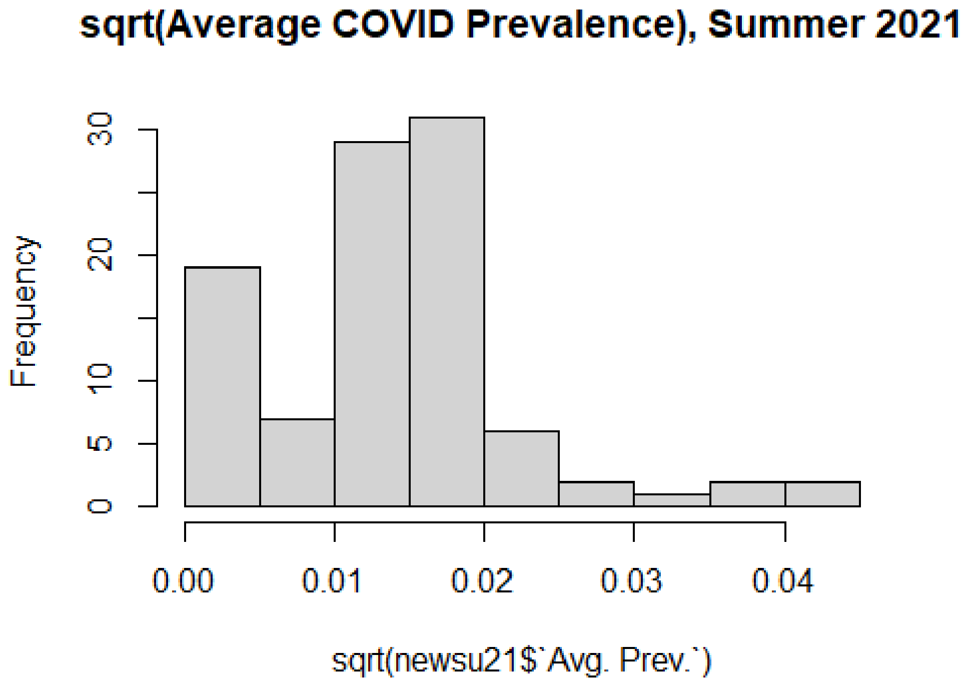
```
# Su21
hist(newsu21$`Avg. Prev.`, main = "Average COVID Prevalence, Summer 2021")
```



```
hist(log(newsu21$`Avg. Prev.`, exp(1)), main = "ln(Average COVID Prevalence),\nSummer 2021")

## Warning in hist(log(newsu21$`Avg. Prev.`, exp(1)), main = "ln(Average COVID
## Prevalence),\nSummer 2021"): NaNs produced
```



```
hist(sqrt(newsu21$`Avg. Prev.`), main = "sqrt(Average COVID Prevalence), Summer 2021")

## Warning in sqrt(newsu21$`Avg. Prev.`): NaNs produced
```



A5: Variable and Model Selection

```
# Use leaps package for best subsets procedure.
library(leaps)

# Consider the set of models with up to four variables, and search for the best three.
allWhole <- regsubsets(newWhole$`Avg. Prev.` ~ newWhole$`% Hisp.` + newWhole$`% White` + newWhole$`% Black` + new
Whole$`% Asian` + newWhole$`% Am. Ind.` + newWhole$`% Nat. Hawaiian` + newWhole$`% Unemp.` + newWhole$`% Belo
w Pov.` + newWhole$`% Unins.` + newWhole$`Univ. Ind.`, data = newWhole, nbest = 3, nvmax = 4)

summary(allWhole)$which # all models, with a logical matrix of variables included (TRUE = yes, FALSE = no)
summary(allWhole)$rsq # 10, then 11, then 12 (higher is better)
summary(allWhole)$adjr2 # 10, then 11, then 12 (higher is better)
summary(allWhole)$cp # 10, then 11, then 12 (smaller is better)
summary(allWhole)$bic # 10, then 7, then 11 (lower is better)
summary(allWhole)$rss # 10, then 11, then 12 (lower is better).

# Models 10 and 11 seem to be the best models here.
# Model 10 includes intercept, % Hispanic, % Native Hawaiian, % Unemployed, and % Below Poverty.
# Model 11 includes intercept, % Hispanic, % Asian, % Native Hawaiian, % Below Poverty.
# No possible models include the university indicator.

# SPRING 2020

# Consider the set of models with up to four variables, and search for the best three.
allSp20 <- regsubsets(newSp20$`Avg. Prev.` ~ newSp20$`% Hisp.` + newSp20$`% White` + newSp20$`% Black` + newSp20$`
% Asian` + newSp20$`% Am. Ind.` + newSp20$`% Nat. Hawaiian` + newSp20$`% Unemp.` + newSp20$`% Below Pov.` + new
Sp20$`% Unins.` + newSp20$`Univ. Ind.`, data = newSp20, nbest = 3, nvmax = 4)
```

```
summary(allSp20)$which # all models, with a logical matrix of variables included (TRUE = yes, FALSE = no)
summary(allSp20)$rsq #  10, then 11, then 12 (higher is better)
summary(allSp20)$adjr2 #  10, then 11, then 12 (higher is better)
summary(allSp20)$cp # 10, then 11, then 12 (smaller is better)
summary(allSp20)$bic # 7, then 10, then 11 (lower is better)
summary(allSp20)$rss # 10, then 11, then 12 (lower is better)

# Models 10 and 11 seem to be the best models here.
# Model 10 includes intercept, % Hispanic, % Black, % Asian, % Uninsured.
# Model 11 includes intercept, % Hispanic, % Black, % American-Indian, % Uninsured
# Only Model 6 includes the university indicator.

# SUMMER 2020

# Consider the set of models with up to four variables, and search for the best three.
allSu20 <- regsubsets(newSu20$`Avg. Prev.` ~ newSu20$`% Hisp.` + newSu20$`% White` + newSu20$`% Black` + newSu20$`% Asian` + newSu20$`% Am. Ind.` + newSu20$`% Nat. Hawaiian` + newSu20$`% Unemp.` + newSu20$`% Below Pov.` + newSu20$`% Unins.` + newSu20$`Univ. Ind.`, data = newSu20, nbest = 3, nvmax = 4)

summary(allSu20)
summary(allSu20)$which # (TRUE = yes, FALSE = no)
summary(allSu20)$rsq # 10, then 11, then 12 (higher is better)
summary(allSu20)$adjr2 # 10, then 11, then 12 (higher is better)
summary(allSu20)$cp # 10, then 11, then 12 (smaller is better)
summary(allSu20)$bic # 10, then 11, then 12 (lower is better)
summary(allSu20)$rss # 10, then 11, then 12 (lower is better)

# Models 10 and 11 tended to do well here.
# Model 10 includes intercept, % Hispanic, % Black, % American-Indian, % Native Hawaiian.
# Model 11 includes intercept, % Hispanic, % American-Indian, % Native Hawaiian, University Indicator.
# Only Models 9 and 11 included the university indicator.

# FALL 2020

# Consider the set of models with up to four variables, and search for the best three.
allf20 <- regsubsets(newf20$`Avg. Prev.` ~ newf20$`% Hisp.` + newf20$`% White` + newf20$`% Black` + newf20$`% Asian` + newf20$`% Am. Ind.` + newf20$`% Nat. Hawaiian` + newf20$`% Unemp.` + newf20$`% Below Pov.` + newf20$`% Unins.` + newf20$`Univ. Ind.`, data = newf20, nbest = 3, nvmax = 4)


summary(allf20)$which # (TRUE = yes, FALSE = no)
summary(allf20)$rsq # 10, then 11, then 12 (higher is better)
summary(allf20)$adjr2 # 10, then 11, then 12 (higher is better)
summary(allf20)$cp #  10, then 11, then 7 (smaller is better)
summary(allf20)$bic # 7, then 8, then 4 (lower is better)
summary(allf20)$rss # 10, then 11, then 12 (lower is better)

# Models 10 and 11 seem to be best here.
# Model 10 includes intercept, % Hispanic, % Asian, % Native Hawaiian, % Unemployed.
# Model 11 includes intercept, % Hispanic, % White, % American-Indian, % Native Hawaiian.

# WINTER 2021

# Consider the set of models with up to four variables, and search for the best three.
allw21 <- regsubsets(neww21$`Avg. Prev.` ~ neww21$`% Hisp.` + neww21$`% White` + neww21$`% Black` + neww21$`% Asian` + neww21$`% Am. Ind.` + neww21$`% Nat. Hawaiian` + neww21$`% Unemp.` + neww21$`% Below Pov.` + neww21$`% Unins.` + neww21$`Univ. Ind.`, data = neww21, nbest = 3, nvmax = 4)


summary(allw21)$which # (TRUE = yes, FALSE = no)
summary(allw21)$rsq # 10, then 11, then 12  (higher is better)
summary(allw21)$adjr2 # 10, then 11, then 12 (higher is better)
summary(allw21)$cp # 10, then 11, then 12 (smaller is better)
summary(allw21)$bic # 10, then 7, then 8 (lower is better)
summary(allw21)$rss # 10, then 11, then 12 (lower is better)

# Models 10 and 11 seem to be the best models here.
# Model 10 includes intercept, % Hispanic, % American-Indian, % Below Poverty, University Indicator
# Model 11 includes intercept, % Hispanic, % Unemployment, % Below Poverty, University Indicator
# 3 of 12 possible models used the university indicator.

# SPRING 2021

# Consider the set of models with up to four variables, and search for the best three.
allsp21 <- regsubsets(newsp21$`Avg. Prev.` ~ newsp21$`% Hisp.` + newsp21$`% White` + newsp21$`% Black` + newsp21$`% Asian` + newsp21$`% Am. Ind.` + newsp21$`% Nat. Hawaiian` + newsp21$`% Unemp.` + newsp21$`% Below Pov.` + newsp21$`% Unins.` + newsp21$`Univ. Ind.`, data = newsp21, nbest = 3, nvmax = 4)

summary(allsp21)$which # (TRUE = yes, FALSE = no)
summary(allsp21)$rsq #  10, then 11, then 12 (higher is better)
summary(allsp21)$adjr2 # 10, then 11, then 12 (higher is better)
summary(allsp21)$cp # 10, then 11, then 12 (smaller is better)
summary(allsp21)$bic # 4, then 10, then 11 (lower is better)
summary(allsp21)$rss # 10, then 11, then 12 (lower is better)

# Models 10 and 11 seem to do the best here.
# Model 10 includes intercept, % Hispanic, % Unemployed, % Below Poverty, % Uninsured.
# Model 11 includes intercept, % Hispanic, % Black, % Unemployed, % Below Poverty.
# No possible models use the university indicator.
```

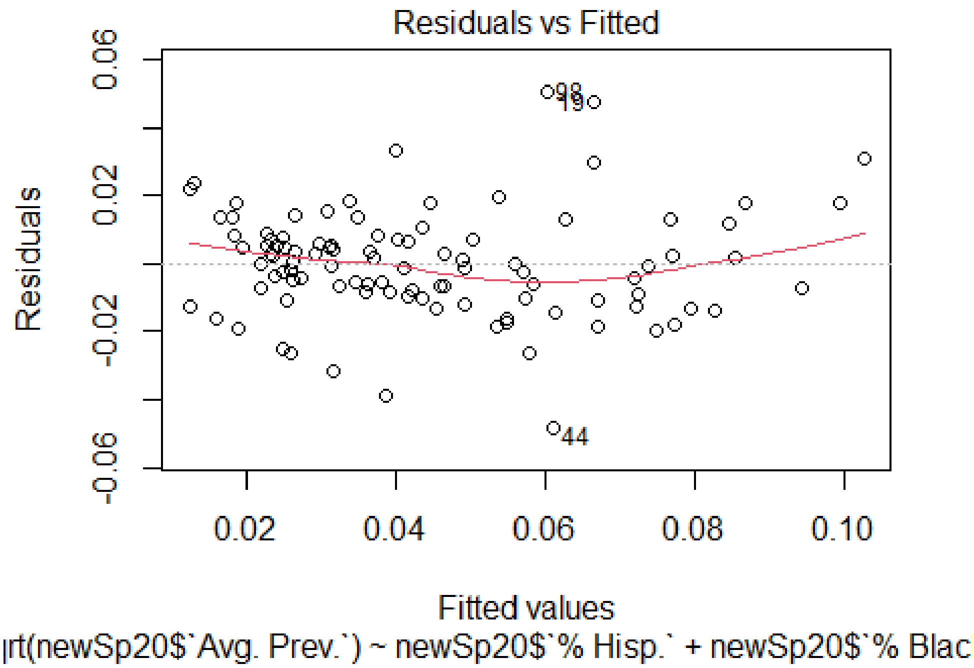
A6: Implementing Proposed Models and Assessing Diagnostics

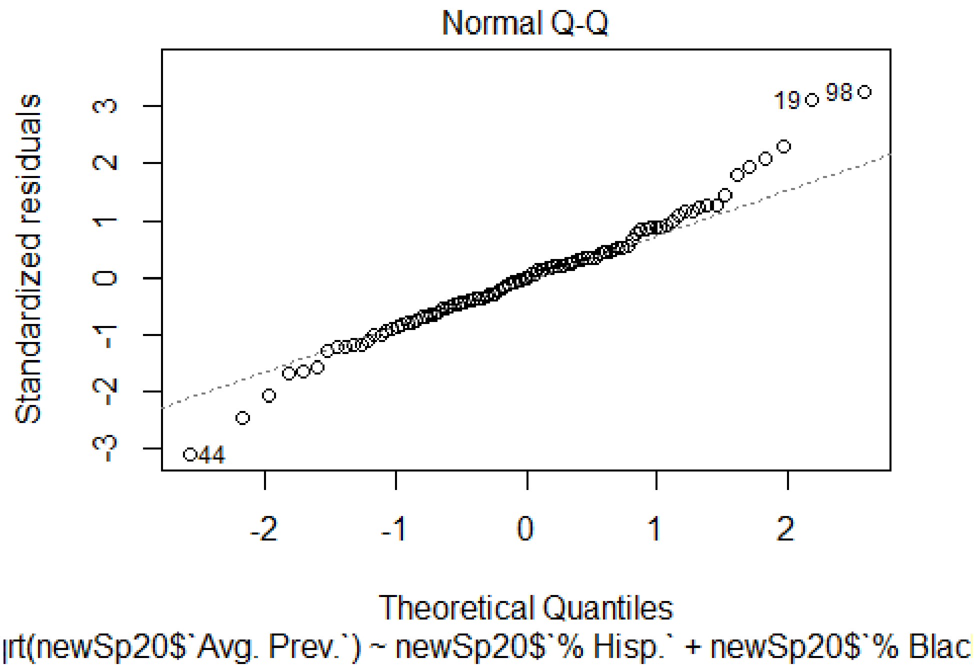
```
# Unconditional of season, 14 months
fitWhole <- lm(newWhole$`Avg. Prev.` ~ newWhole$`% Hisp.` + newWhole$`% Nat. Hawaiian` + newWhole$`% Unemp.` + newWhole$`% Below Pov.`)
summary(fitWhole)
plot(fitWhole)

# 19, 48, 49 are potential outliers (22 just in res v leverage) with no transformations
# 49, 99, 102 are potential outliers when just the response is log-transformed
# 49, 99, 102 (11 appears in res v leverage) when just the response is square-rooted

# SPRING 2020

fitSp20 <- lm(newSp20$`Avg. Prev.` ~ newSp20$`% Hisp.` + newSp20$`% Black` + newSp20$`% Asian` + newSp20$`% Unins.`)
summary(fitSp20)
plot(fitSp20)

# 19, 95, 98 are potential outliers (48 only in res v leverage) with no transformation
# 19, 44, 98 are potential outliers (95 only in res v leverage) under square root of response

# SUMMER 2020

fitSu20 <- lm(newSu20$`Avg. Prev.` ~ newSu20$`% Hisp.` + newSu20$`% Black` + newSu20$`% Am. Ind.` + newSu20$`% Nat. Hawaiian`)
summary(fitSu20)
plot(fitSu20)

# FALL 2020

fitf20 <- lm(newf20$`Avg. Prev.` ~ newf20$`% Hisp.` + newf20$`% Asian` + newf20$`% Nat. Hawaiian` + newf20$`% Unemp.`)
```

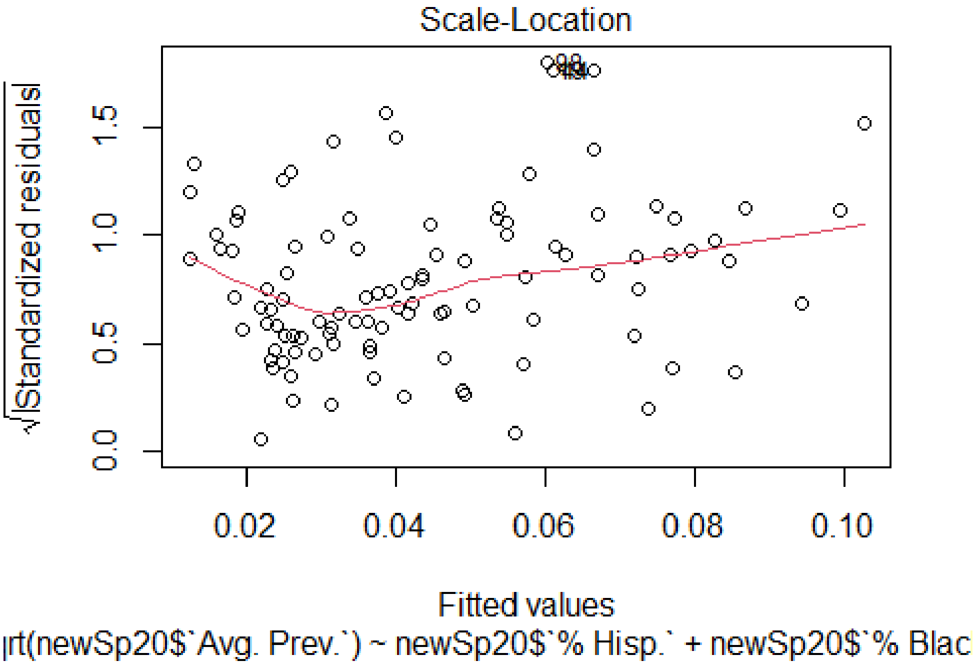
```
summary(fitf20)
plot(fitf20)

# WINTER 2021

fitw21 <- lm(neww21$`Avg. Prev.` ~ neww21$`% Hisp.` + neww21$`% Am. Ind.` + neww21$`% Below Pov.` + neww21$`Univ. Ind.`)
summary(fitw21)
plot(fitw21)

# SPRING 2021

fitsp21 <- lm(newsp21$`Avg. Prev.` ~ newsp21$`% Hisp.` + newsp21$`% Unemp.` + newsp21$`% Below Pov.` + newsp21$`% Unins.`)
summary(fitsp21)
plot(fitsp21)
```

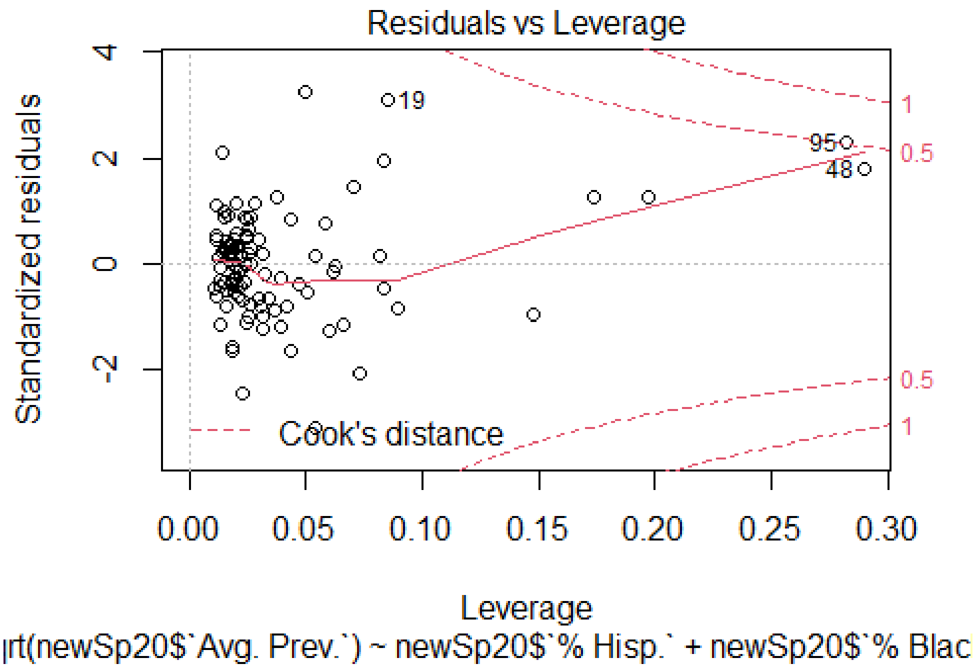
A7: Model Diagnostics for Proposed Models and Their Transformations

```
# TEMPLATE FOR INDICATOR VARIABLE FOR ZIP CODES
#
#
# as.numeric(sdDemo[,1]%in%c(91963, 92059, 92060, 91980))


# SPRING 2020
# 19, 95, 98 are potential outliers (48 only in res v leverage) with no transformation (ZIPs 91963, 92134, 92140, with 92059 in res v leverage)
# 19, 44, 98 are potential outliers (95 only in res v leverage) under square root of response (ZIPs 91963, 92055, 92140, with 92134 in res v leverage)

# Model 10 includes % Hispanic, % Black, % Asian, and % Uninsured.
fitSp20 <- lm(newSp20$`Avg. Prev.` ~ newSp20$`% Hisp.` + newSp20$`% Black` + newSp20$`% Asian` + newSp20$`% Unins.`) # This model alone has departure from normality in residuals and there is trend in the residuals. Try something else.
summary(fitSp20)
plot(fitSp20)

modfitSp20 <- lm(sqrt(newSp20$`Avg. Prev.`) ~ sqrt(newSp20$`% Hisp.`) + sqrt(newSp20$`% Black`) + sqrt(newSp20$`% Asian`) + sqrt(newSp20$`% Unins.`))
summary(modfitSp20)
plot(modfitSp20) # the residuals seem to have a trend
hist(modfitSp20$residuals) # residuals are approximately normal in distribution
dwtest(modfitSp20) # residuals not likely to be autocorrelated
bptest(modfitSp20) # suggests that error variance is not constant

modfitSp20a <- lm(sqrt(newSp20$`Avg. Prev.`) ~ newSp20$`% Hisp.` + newSp20$`% Black` + newSp20$`% Asian` + newSp20$`% Unins.`)
summary(modfitSp20a)
plot(modfitSp20a) # the residuals seem to have a trend
hist(modfitSp20a$residuals) # residuals are approximately normal in distribution
dwtest(modfitSp20a) # residuals not likely to be autocorrelated
bptest(modfitSp20a) # suggests that error variance is not constant
# Possible keep this model. ^ Next, remove % Asian.

# Need lmtest package for the Durbin-Watson Test.
library(lmtest)

## Loading required package: zoo

##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##
## Attaching package: 'lmtest'

## The following object is masked from 'package:VGAM':
##
##     lrtest

# FINAL MODEL, SPRING 2020
modfitSp20b <- lm(sqrt(newSp20$`Avg. Prev.`) ~ newSp20$`% Hisp.` + newSp20$`% Black` + newSp20$`% Unins.`)
summary(modfitSp20b)

##
## Call:
## lm(formula = sqrt(newSp20$`Avg. Prev.`) ~ newSp20$`% Hisp.` +
##     newSp20$`% Black` + newSp20$`% Unins.`)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.048124 -0.009392  0.000074  0.007579  0.050552
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.014006   0.003361   4.167 6.65e-05 ***
## newSp20$`% Hisp.`   0.110930   0.011468   9.673 6.20e-16 ***
## newSp20$`% Black`   0.192027   0.036701   5.232 9.53e-07 ***
## newSp20$`% Unins.` -0.144273   0.042583  -3.388  0.00102 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.016 on 98 degrees of freedom
## Multiple R-squared:  0.6518, Adjusted R-squared:  0.6411
## F-statistic: 61.14 on 3 and 98 DF,  p-value: < 2.2e-16

plot(modfitSp20b) # the residuals seem to have a trend
```

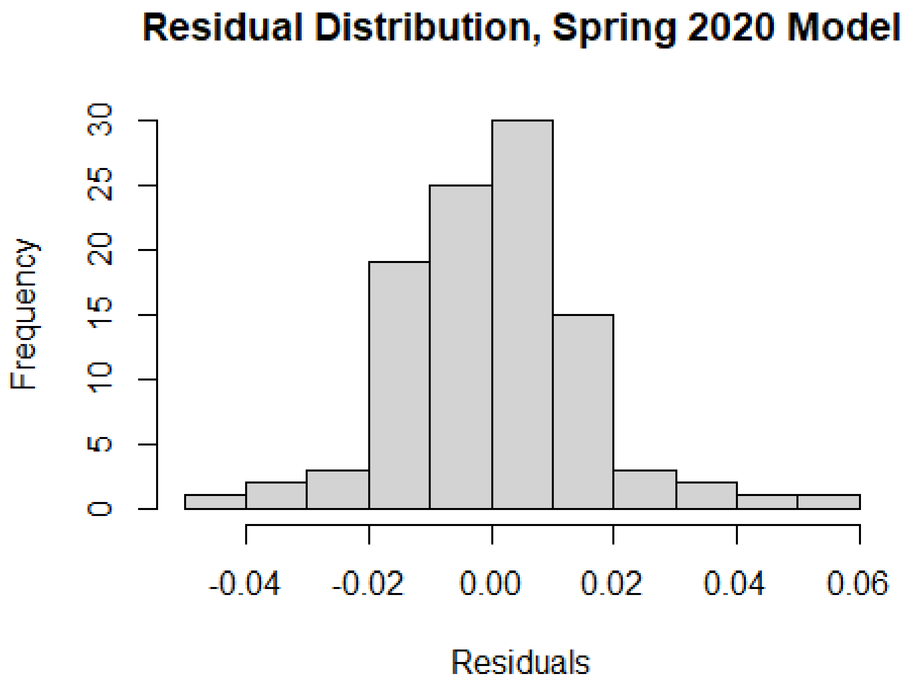

Residuals vs Fitted
Residuals
0.06
0.02
-0.02
-0.06
98
19
44
0.02
0.04
0.06
0.08
0.10
Fitted values
ırt(newSp20$`Avg. Prev.`) ~ newSp20$`% Hisp.` + newSp20$`% Blac


Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
-3
19
98
44
-2
-1
0
1
2
Theoretical Quantiles
ırt(newSp20$`Avg. Prev.`) ~ newSp20$`% Hisp.` + newSp20$`% Blac

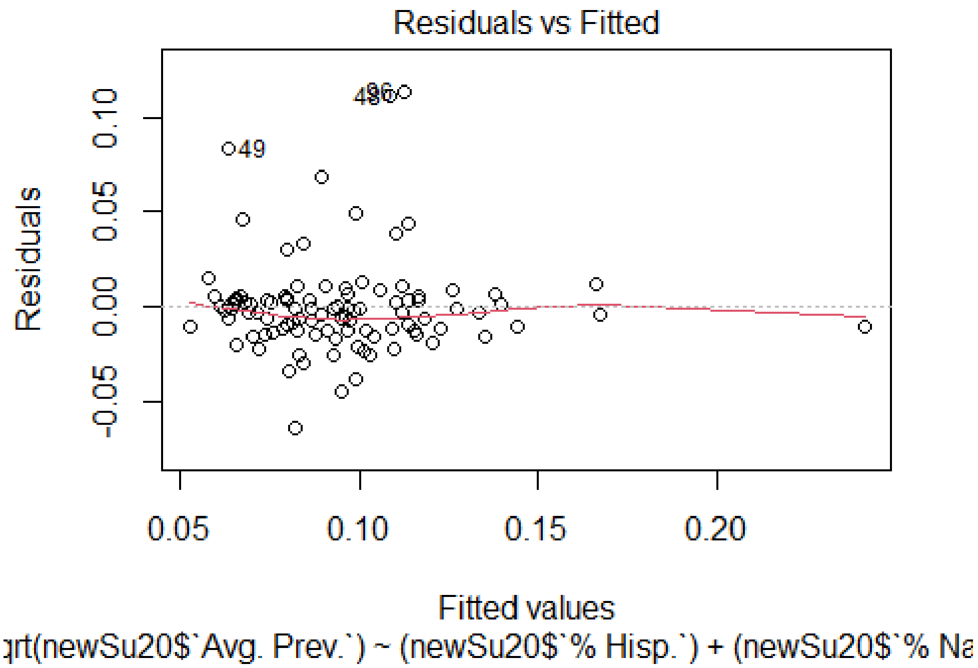


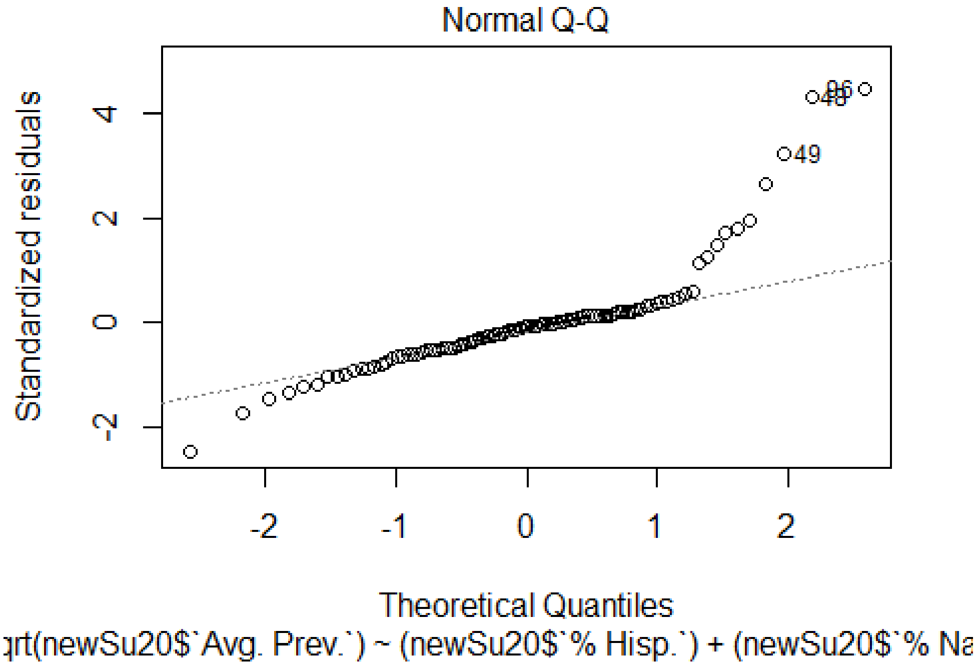


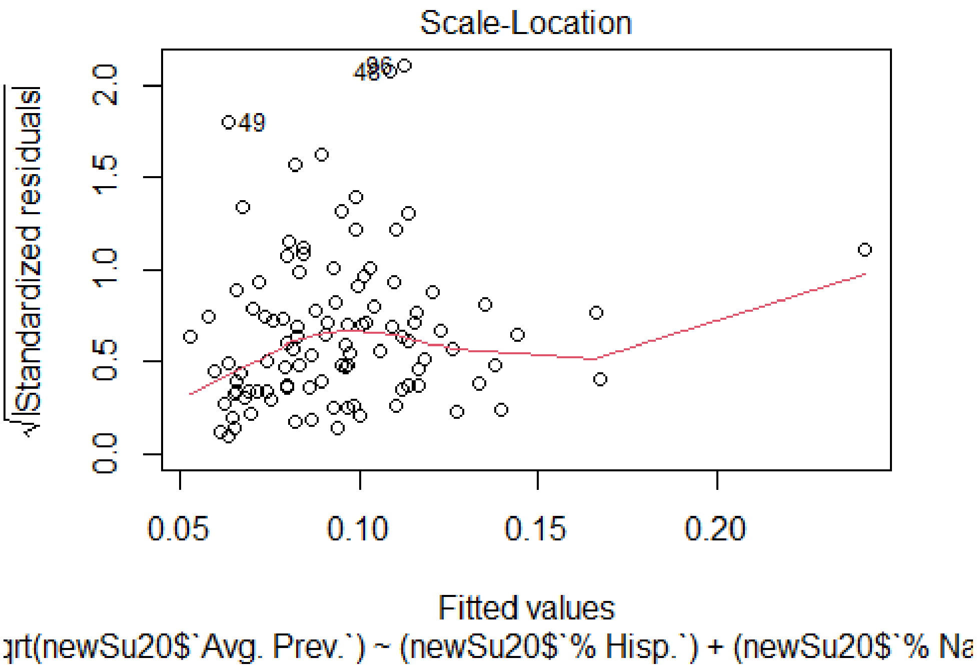
```
hist(modfitSp20b$residuals, main = "Residual Distribution, Spring 2020 Model", xlab = "Residuals") # residuals are approximat
ely normal in distribution
```

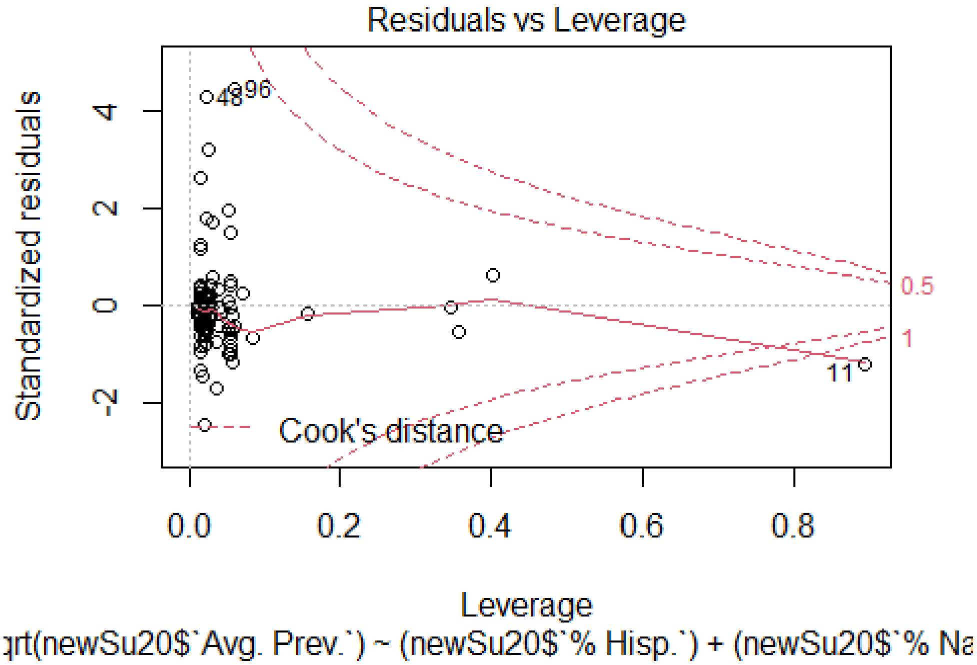


```
dwtest(modfitSp20b) # residuals not likely to be autocorrelated

##
##  Durbin-Watson test
##
## data:  modfitSp20b
## DW = 1.9759, p-value = 0.423
## alternative hypothesis: true autocorrelation is greater than 0

bptest(modfitSp20b) # suggests that error variance is more likely to be constant

##
##  studentized Breusch-Pagan test
##
## data:  modfitSp20b
## BP = 7.7717, df = 3, p-value = 0.05097

# SUMMER 2020
# Untransformed: 48, 49, 96 (50 in res v leverage) may be potential outliers. (ZIPs 92059, 92060, 92135, with 92061 in res v leve
rage)
# Log of response: 49, 96, 102 (48 in loc-scale, 56 in res v leverage) may be potential outliers. (ZIPs 92060, 92135, 92536, with
92059 in loc-scale, 92070 in res v leverage)
# Square root of response: 48, 49, 96 (50 in res v leverage) may be potential outliers. (ZIPs 92059, 92060, 92135, with 92061 in
res v leverage)


fitSu20 <- lm(newSu20$`Avg. Prev.` ~ newSu20$`% Hisp.` + newSu20$`% Black` + newSu20$`% Am. Ind.` + newSu20$`% Na
t. Hawaiian`)
#summary(fitSu20) # esentially perfect fit: summary may be unreliable
#plot(fitSu20)

# Try the second candidate model from best subsets:
fitSu20alt <- lm(newSu20$`Avg. Prev.` ~ newSu20$`% Hisp.` + newSu20$`% Am. Ind.` + newSu20$`% Nat. Hawaiian` + newS
u20$`Univ. Ind.`)
summary(fitSu20alt)
plot(fitSu20alt) # This model on its own has strange, inpermissible diagnostics.

modfitSu20alt <- lm(newSu20$`Avg. Prev.` ~ newSu20$`% Hisp.` + newSu20$`% Am. Ind.` + newSu20$`% Nat. Hawaiian` + n
ewSu20$`Univ. Ind.`)
#summary(modfitSu20alt)
#plot(modfitSu20alt)



fitSu20altBC <- lm((newSu20$`Avg. Prev.`)^(106/99) ~ newSu20$`% Hisp.` + newSu20$`% Am. Ind.` + newSu20$`% Nat. Ha
waiian` + newSu20$`Univ. Ind.`)
#summary(fitSu20altBC)
#plot(fitSu20altBC) # homoskedasticity is violated
# Do not use Boxcox.

# FINAL MODEL, SUMMER 2020
modfitSu20alt2 <- lm(sqrt(newSu20$`Avg. Prev.`) ~ (newSu20$`% Hisp.`)  + (newSu20$`% Nat. Hawaiian`) + newSu20$`Univ.
Ind.` + as.numeric(rownames(newSu20)%in%c("51", "52", "101"))) # 53 95 78, 50 under each variable and response square-roo
ted
#  + as.numeric(sdDemo[,1]%in%c(92066, 92114, 92134, 92061))        + (newSu20$`% Am. Ind.`)
summary(modfitSu20alt2)

##
## Call:
## lm(formula = sqrt(newSu20$`Avg. Prev.`) ~ (newSu20$`% Hisp.`) +
##     (newSu20$`% Nat. Hawaiian`) + newSu20$`Univ. Ind.` + as.numeric(rownames(newSu20) %in%
##     c("51", "52", "101")))
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.064110 -0.012617 -0.001850  0.003947  0.113532
##
## Coefficients:
##                                                     Estimate Std. Error
## (Intercept)                                         0.050972   0.005272
## newSu20$`% Hisp.`                                   0.116428   0.014154
## newSu20$`% Nat. Hawaiian`                           1.174222   0.217267
## newSu20$`Univ. Ind.`                                0.019381   0.006853
## as.numeric(rownames(newSu20) %in% c("51", "52", "101")) 0.004709   0.015705
##                                                     t value Pr(>|t|)
## (Intercept)                                           9.669 6.94e-16 ***
## newSu20$`% Hisp.`                                     8.226 8.87e-13 ***
## newSu20$`% Nat. Hawaiian`                             5.405 4.65e-07 ***
## newSu20$`Univ. Ind.`                                  2.828  0.00569 **
## as.numeric(rownames(newSu20) %in% c("51", "52", "101"))   0.300  0.76495
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02637 on 97 degrees of freedom
## Multiple R-squared:  0.5302, Adjusted R-squared:  0.5108
## F-statistic: 27.37 on 4 and 97 DF,  p-value: 3.269e-15

plot(modfitSu20alt2)
```

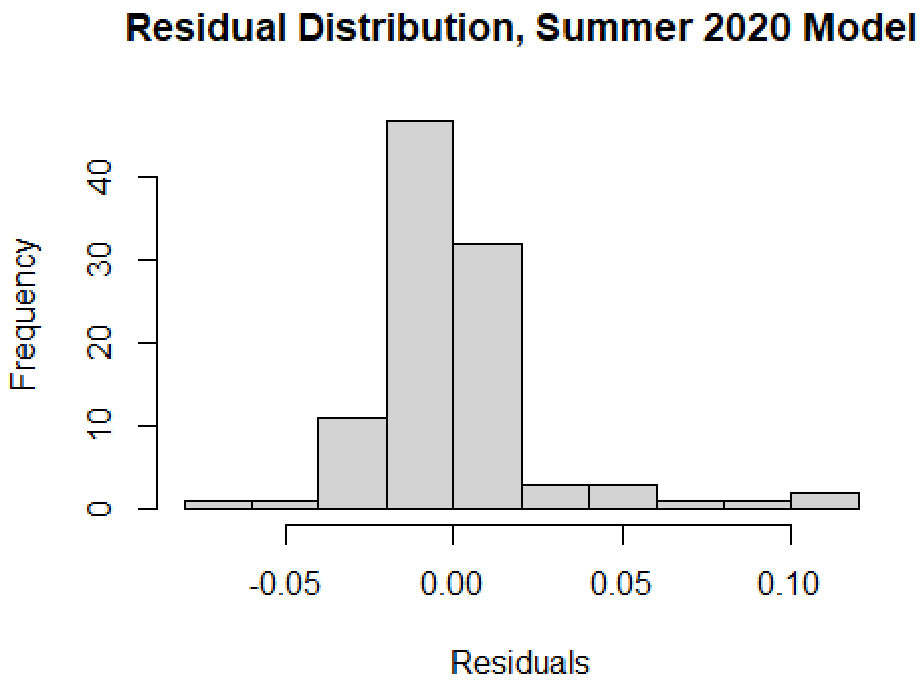

Residuals vs Fitted
Residuals
-0.05
0.00
0.05
0.10
96
48
49
0.05
0.10
0.15
0.20
Fitted values
qrt(newSu20$`Avg. Prev.`) ~ (newSu20$`% Hisp.`) + (newSu20$`% Na


Normal Q-Q
Standardized residuals
-2
0
2
4
96
48
49
-2
-1
0
1
2
Theoretical Quantiles
qrt(newSu20$`Avg. Prev.`) ~ (newSu20$`% Hisp.`) + (newSu20$`% Na

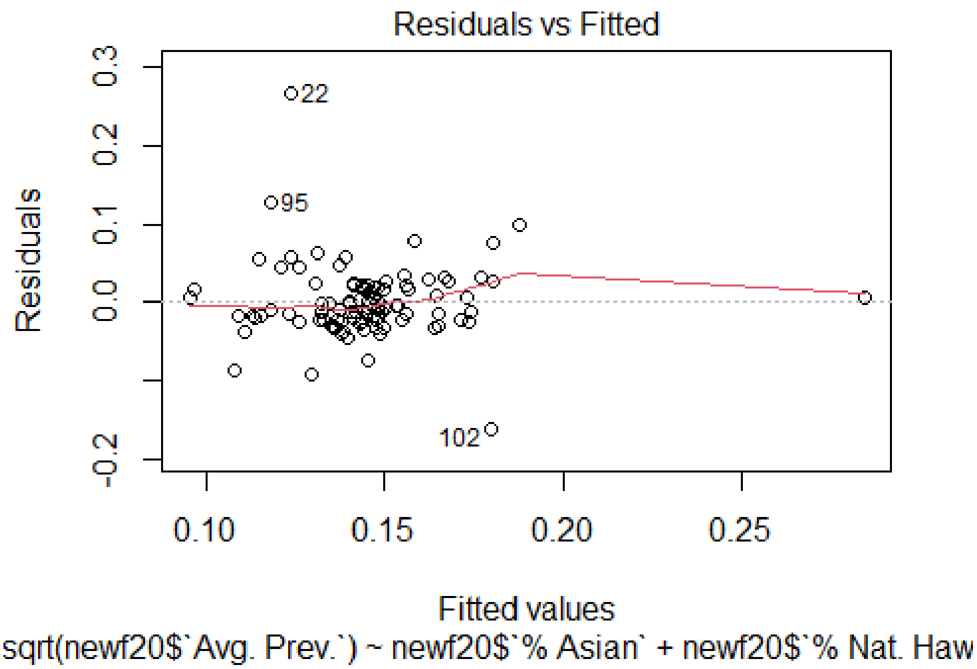


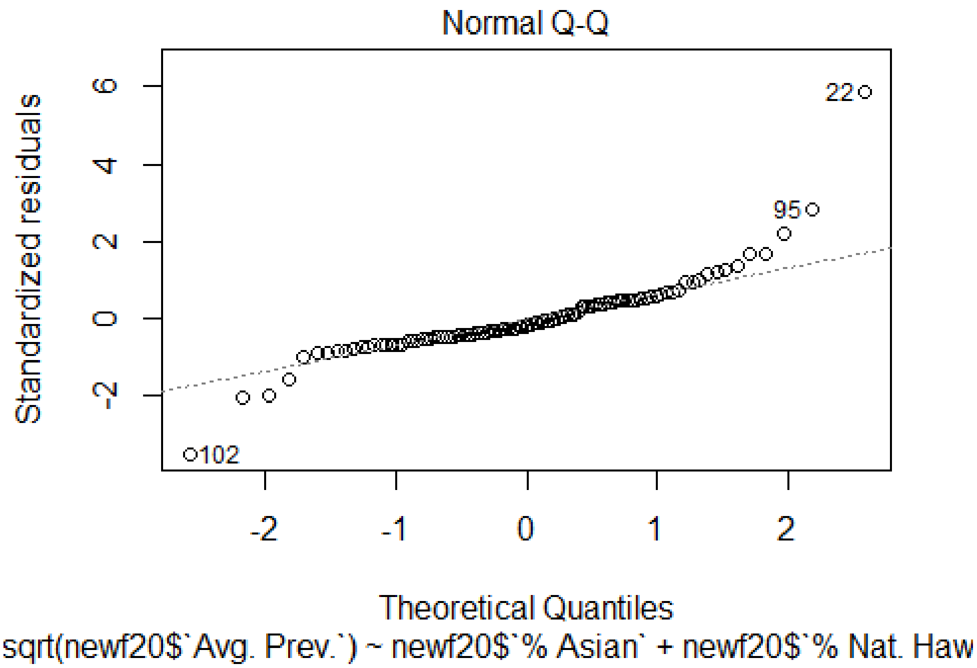


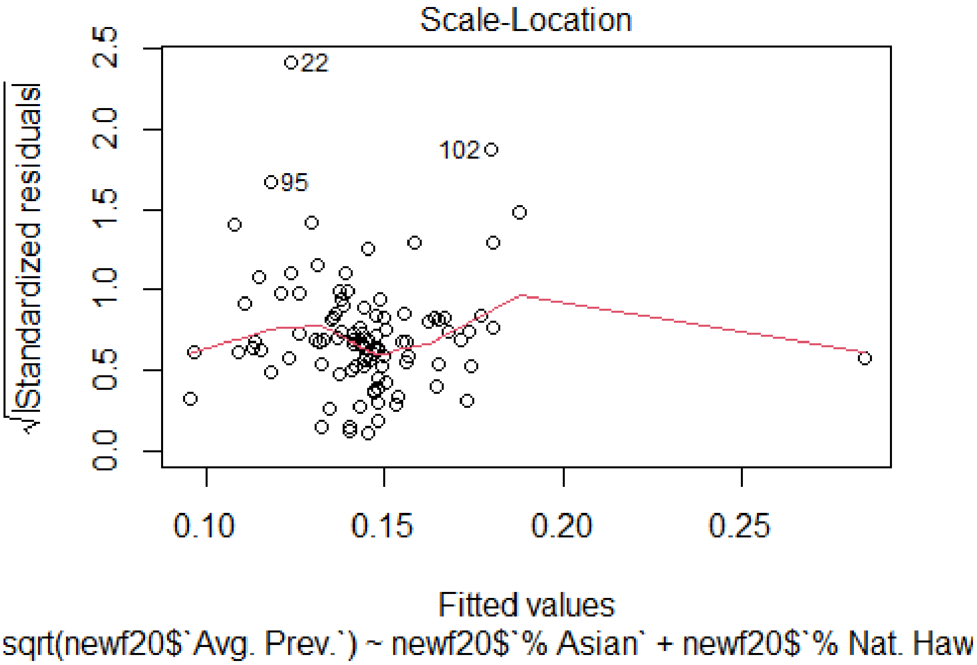
```
hist(modfitSu20alt2$residuals, main = "Residual Distribution, Summer 2020 Model", xlab = "Residuals")
```

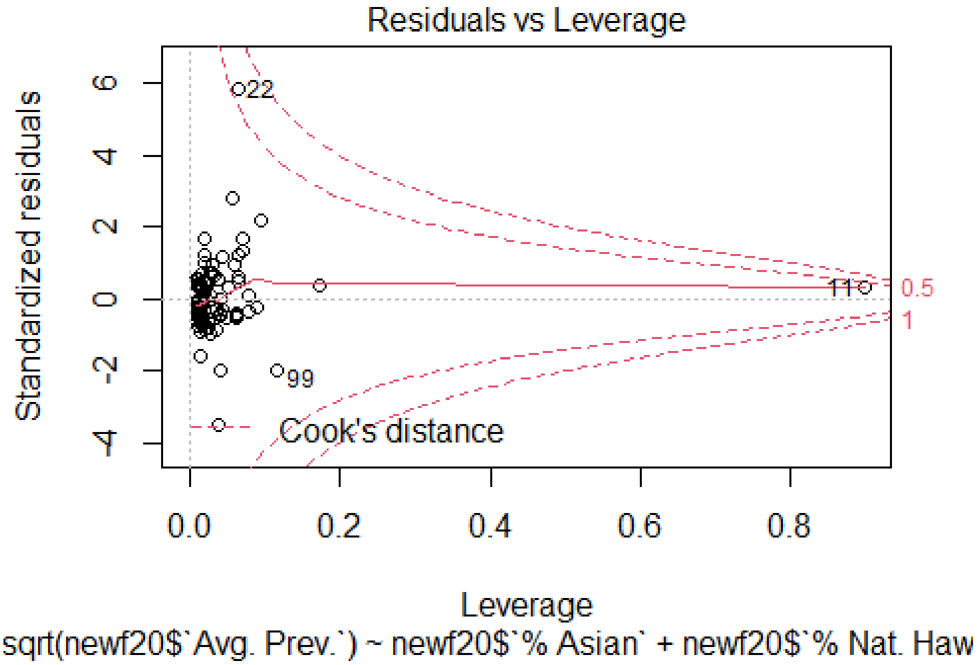


```
dwtest(modfitSu20alt2)

##
##  Durbin-Watson test
##
## data:  modfitSu20alt2
## DW = 1.4492, p-value = 0.001848
## alternative hypothesis: true autocorrelation is greater than 0

bptest(modfitSu20alt2)

##
##  studentized Breusch-Pagan test
##
## data:  modfitSu20alt2
## BP = 2.9483, df = 4, p-value = 0.5665

# FINAL MODEL, FALL 2020
# 19, 22, 95 possible outliers (11 in res v leverage)
# 98, 99, 102 possible outliers (11 in res v leverage) when response is log-transformed
# 22, 98, 102 possible outliers (11 in res v leverage) when response is square-rooted

# % Hispanic, % Asian, % Native Hawaiian, and % Unemployed.
fitF20 <- lm(sqrt(newf20$`Avg. Prev.`) ~ newf20$`% Asian` + newf20$`% Nat. Hawaiian` + newf20$`% Unemp.`)
summary(fitF20)

##
## Call:
## lm(formula = sqrt(newf20$`Avg. Prev.`) ~ newf20$`% Asian` + newf20$`% Nat. Hawaiian` +
##     newf20$`% Unemp.`)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.161434 -0.022291 -0.007743  0.019964  0.266377
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.12416    0.01199  10.359  < 2e-16 ***
## newf20$`% Asian`         -0.13938    0.05128  -2.718  0.00777 **
## newf20$`% Nat. Hawaiian`  1.09611    0.39384   2.783  0.00646 **
## newf20$`% Unemp.`         0.98094    0.35067   2.797  0.00620 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04714 on 98 degrees of freedom
## Multiple R-squared:  0.1921, Adjusted R-squared:  0.1673
## F-statistic: 7.765 on 3 and 98 DF,  p-value: 0.0001051

plot(fitF20)
```

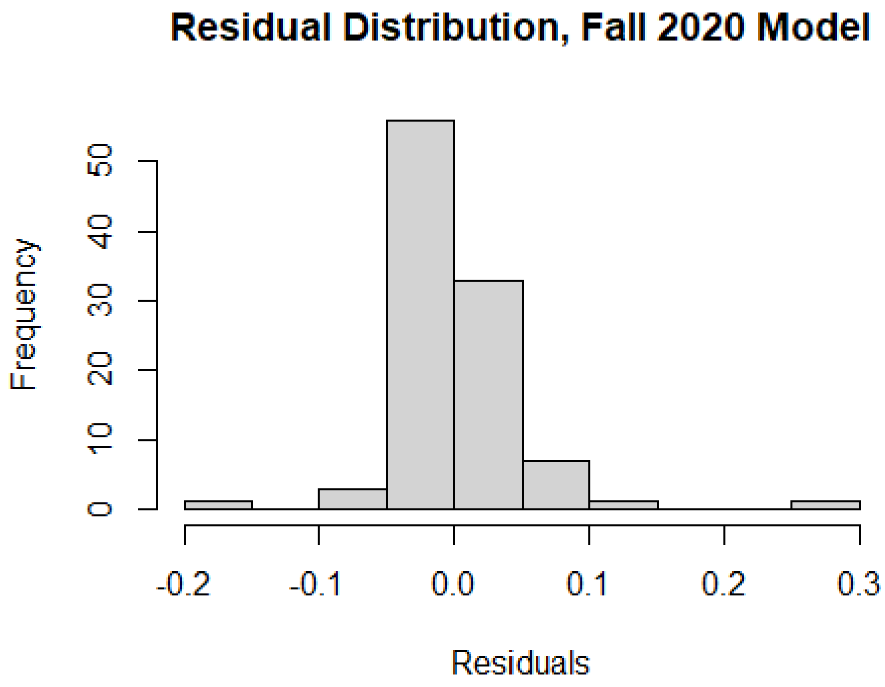

Residuals vs Fitted
Residuals
-0.2
0.0
0.1
0.2
0.3
0.10
0.15
0.20
0.25
22
95
102
Fitted values
sqrt(newf20$`Avg. Prev.`) ~ newf20$`% Asian` + newf20$`% Nat. Haw


Normal Q-Q
Standardized residuals
-2
0
2
4
6
-2
-1
0
1
2
22
95
102
Theoretical Quantiles
sqrt(newf20$`Avg. Prev.`) ~ newf20$`% Asian` + newf20$`% Nat. Haw





```
hist(fitF20$residuals, main = "Residual Distribution, Fall 2020 Model", xlab = "Residuals")
```

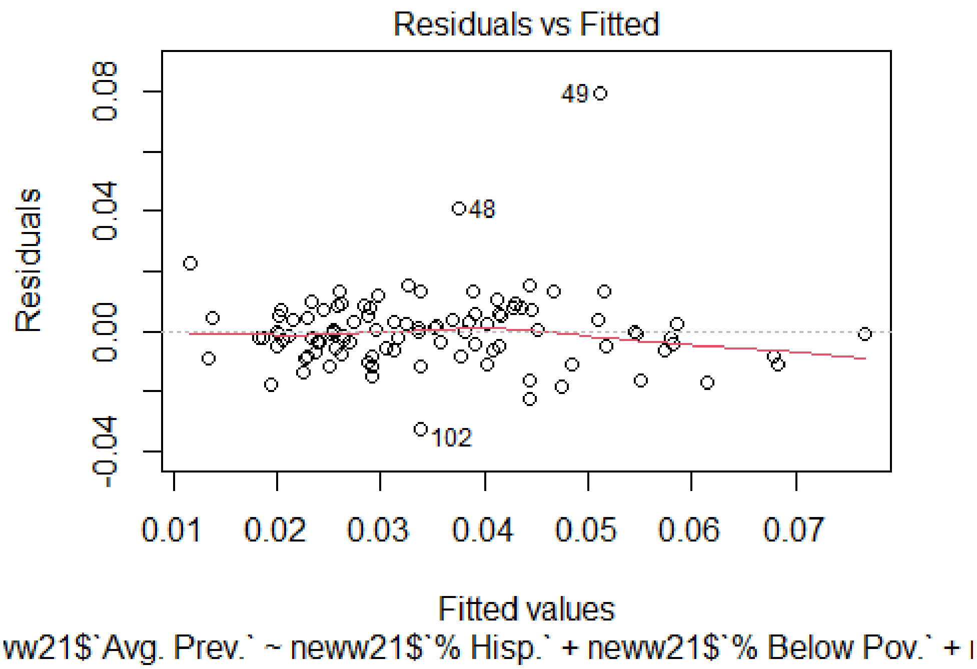


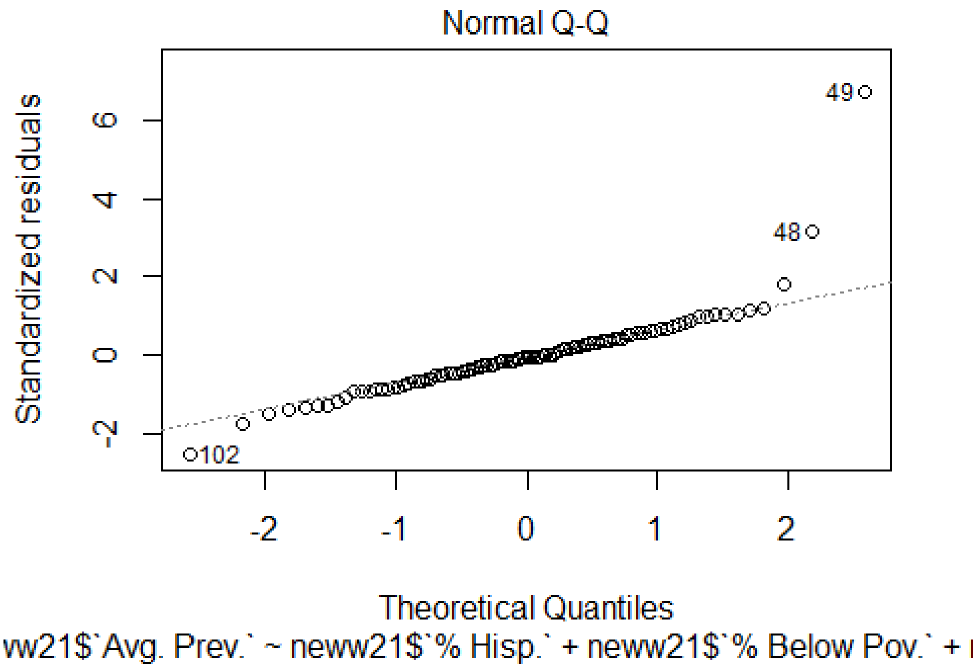
```
dwtest(fitF20) # residuals not likely to be autocorrelated

##
##  Durbin-Watson test
##
## data:  fitF20
## DW = 1.748, p-value = 0.08832
## alternative hypothesis: true autocorrelation is greater than 0

bptest(fitF20) # homoskedastic residuals, most likely

##
##  studentized Breusch-Pagan test
##
## data:  fitF20
## BP = 6.8651, df = 3, p-value = 0.07632

# FINAL MODEL, WINTER 2021
# This one actually looks somewhat well-behaved, but there are still some outliers. (49, 53, 102 for res; 48, 49, 102 for qq and scale-loc; 48, 49, 53 for res v leverage)
# When response is log-transformed, 98, 99, and 102 (49 in res v leverage) are potential outliers.
# When response is square-rooted, 49, 99, and 102 (53 in res v leverage) are potential outliers.

 # % Hispanic, % American-Indian, % Below Poverty, and the University Indicator.

fitW21 <- lm(neww21$`Avg. Prev.` ~ neww21$`% Hisp.` + neww21$`% Below Pov.` + neww21$`Univ. Ind.` + as.numeric(rownames(neww21)=="52"))
summary(fitW21)

##
## Call:
## lm(formula = neww21$`Avg. Prev.` ~ neww21$`% Hisp.` + neww21$`% Below Pov.` +
##     neww21$`Univ. Ind.` + as.numeric(rownames(neww21) == "52"))
```

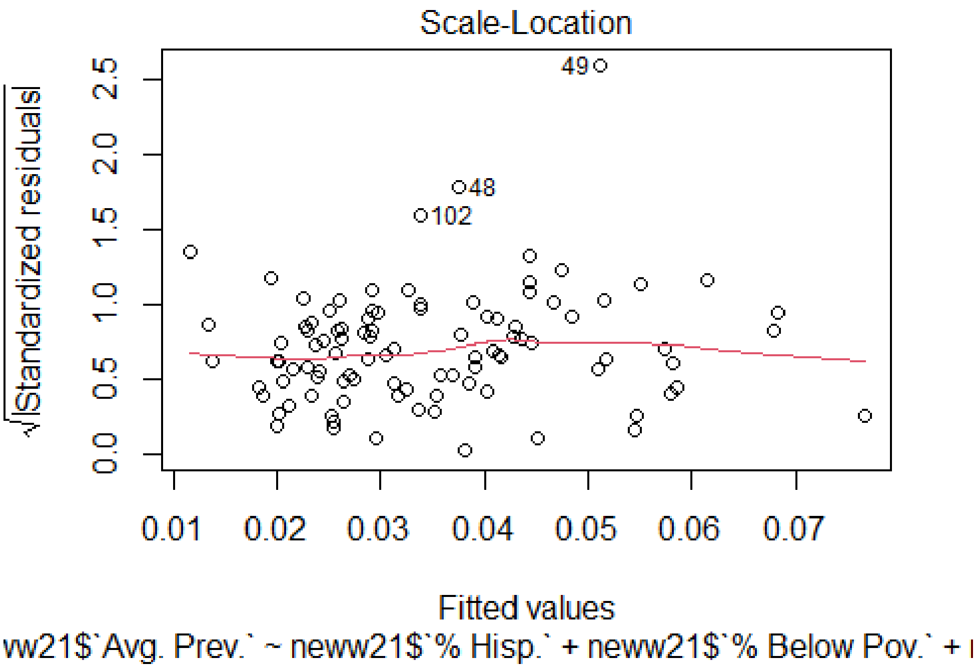
```
## 
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.032477 -0.006326 -0.000811  0.005286  0.079363 
## 
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                           0.010503   0.003168   3.315  0.00129 ** 
## neww21$`% Hisp.`                      0.047108   0.007112   6.624 1.96e-09 ***
## neww21$`% Below Pov.`                 0.326126   0.065580   4.973 2.85e-06 ***
## neww21$`Univ. Ind.`                  -0.009422   0.003464  -2.720  0.00774 ** 
## as.numeric(rownames(neww21) == "52")  0.004224   0.013157   0.321  0.74883    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.01304 on 97 degrees of freedom
## Multiple R-squared:  0.5108, Adjusted R-squared:  0.4907 
## F-statistic: 25.33 on 4 and 97 DF,  p-value: 2.235e-14
```

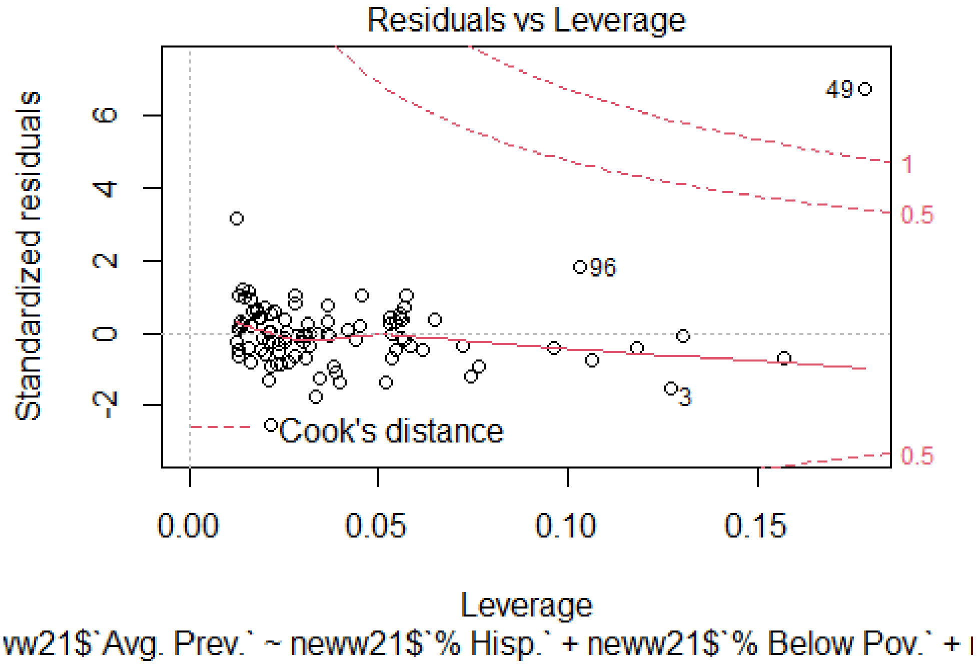
```
plot(fitW21)
```

```
## Warning: not plotting observations with leverage one:
##   52
```

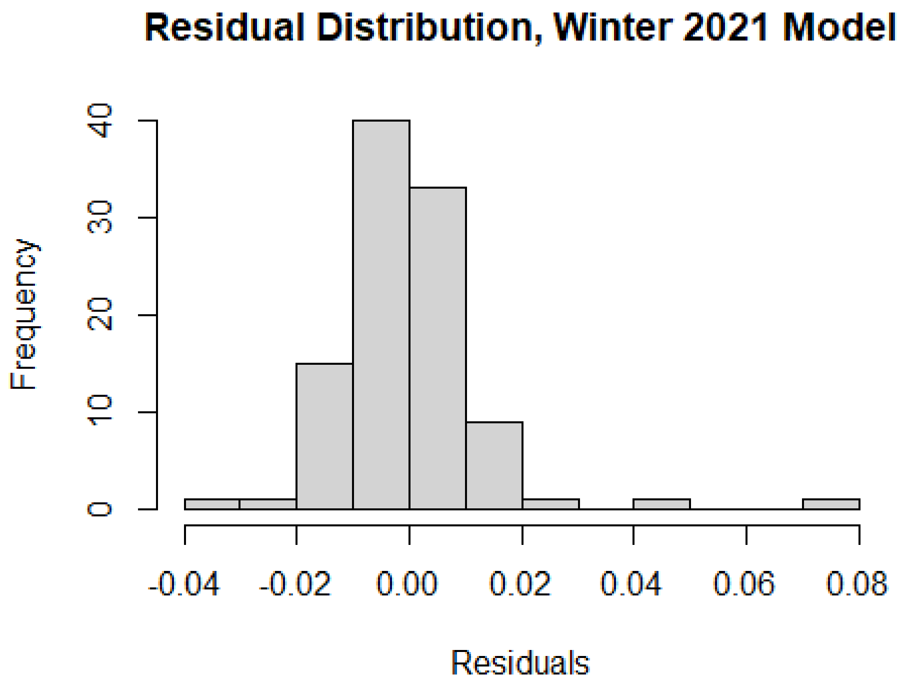

Residuals vs Fitted
Residuals
0.08
0.04
0.00
-0.04
49
48
102
0.01
0.02
0.03
0.04
0.05
0.06
0.07
Fitted values
vw21$`Avg. Prev.` ~ neww21$`% Hisp.` + neww21$`% Below Pov.` +


Normal Q-Q
Standardized residuals
6
4
2
0
-2
49
48
102
-2
-1
0
1
2
Theoretical Quantiles
vw21$`Avg. Prev.` ~ neww21$`% Hisp.` + neww21$`% Below Pov.` +





```
hist(fitW21$residuals, main = "Residual Distribution, Winter 2021 Model", xlab = "Residuals")
```

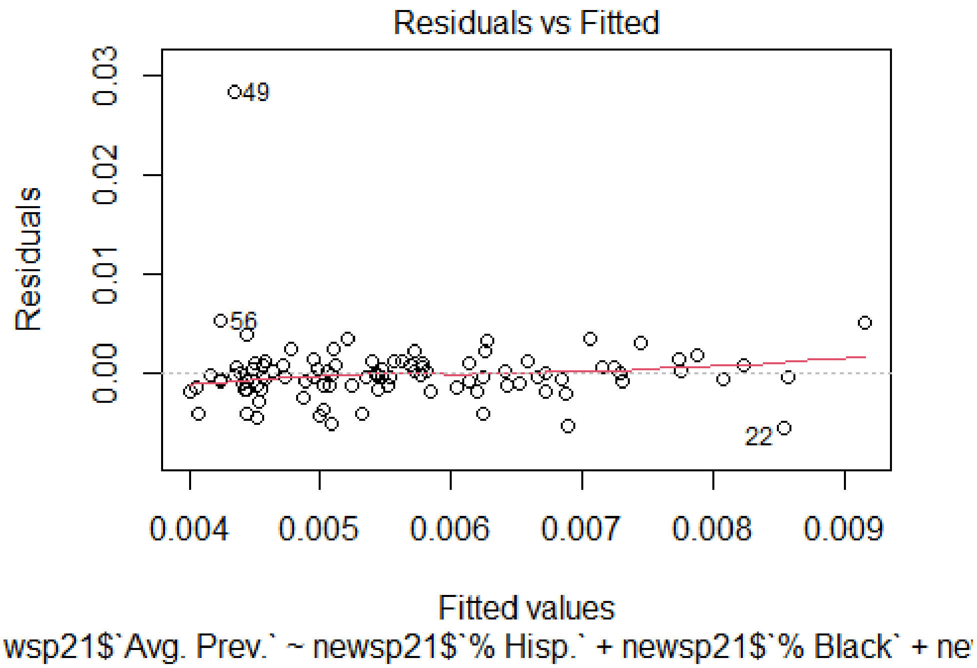


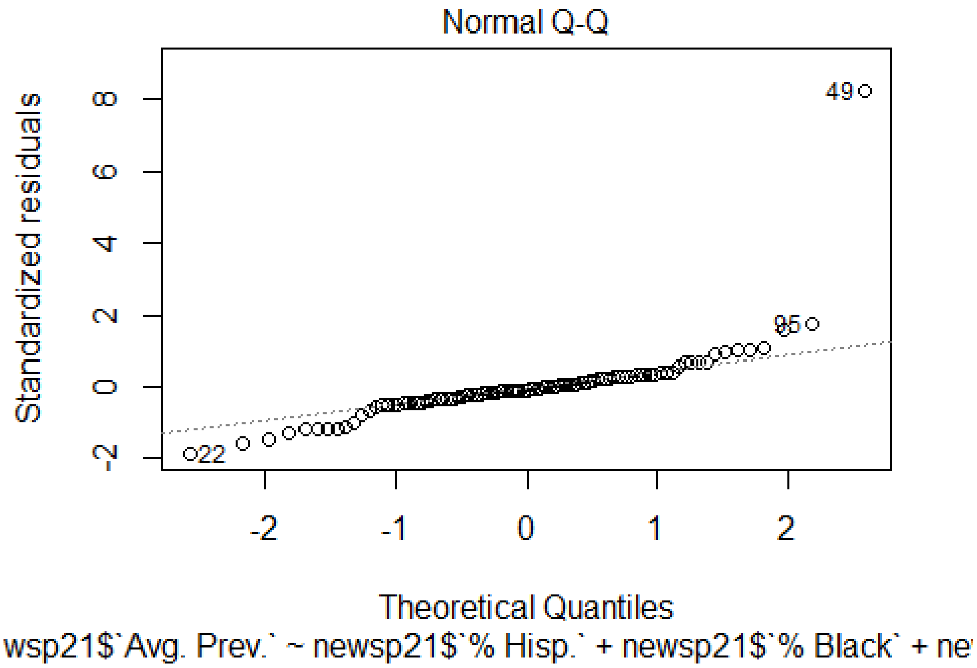
```
dwtest(fitW21)

##
##  Durbin-Watson test
##
## data:  fitW21
## DW = 1.2278, p-value = 2.66e-05
## alternative hypothesis: true autocorrelation is greater than 0

bptest(fitW21)

##
##  studentized Breusch-Pagan test
##
## data:  fitW21
## BP = 18.885, df = 4, p-value = 0.0008277

# SPRING 2021
# No transformation: 3, 49, and 95 (22 in res v leverage) are potential outliers.
# Square root transformation: 49, 53, 64 (22 in res v leverage) are potential outliers.

# % Hispanic, % Unemployed, % Below Poverty, and % Uninsured
fitSp21 <- lm(newsp21$`Avg. Prev.` ~ newsp21$`% Hisp.`  + newsp21$`% Below Pov.` + newsp21$`% Unins.`)
summary(fitSp21) #  + newsp21$`% Unemp.`
plot(fitSp21)
dwtest(fitSp21)
bptest(fitSp21) # homoskedasticity violated

# FINAL MODEL, FALL SPRING 2021
# CONSIDER ALTERNATIVE FROM BEST SUBSETS
mod2fitSp21 <- lm(newsp21$`Avg. Prev.` ~ newsp21$`% Hisp.` + newsp21$`% Black` + newsp21$`% Unemp.` + as.numeric(rownames(newsp21)=="52"))
summary(mod2fitSp21)
```

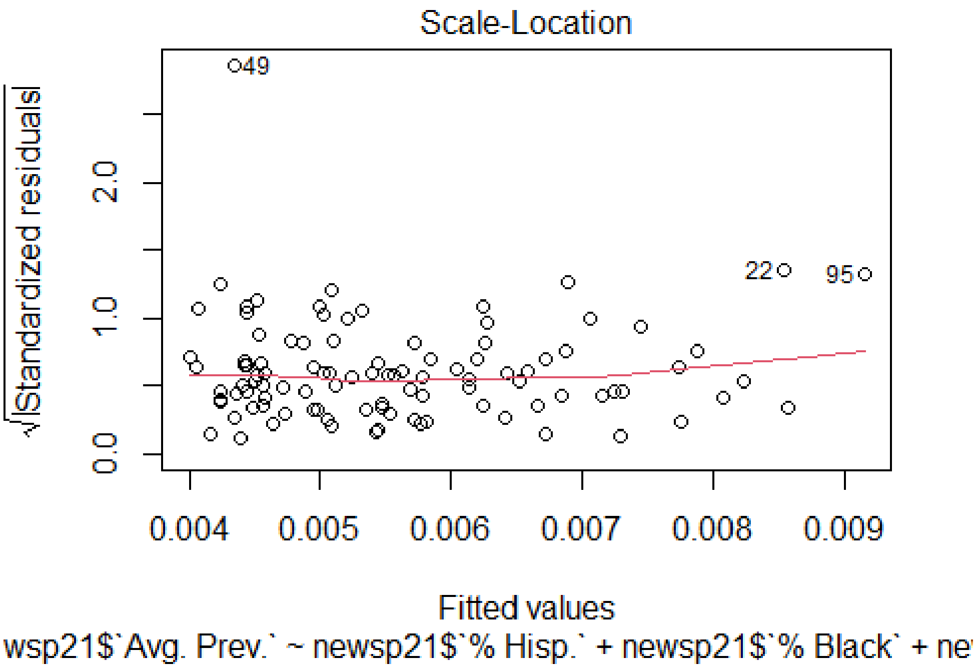
```
##
## Call:
## lm(formula = newsp21$`Avg. Prev.` ~ newsp21$`% Hisp.` + newsp21$`% Black` +
##     newsp21$`% Unemp.` + as.numeric(rownames(newsp21) == "52"))
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0053917 -0.0012447 -0.0002870  0.0008781  0.0282711
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           0.0038255  0.0009178   4.168 6.68e-05
## newsp21$`% Hisp.`                     0.0047109  0.0020522   2.295   0.0239
## newsp21$`% Black`                     0.0139335  0.0080258   1.736   0.0857
## newsp21$`% Unemp.`                   -0.0086478  0.0283443  -0.305   0.7609
## as.numeric(rownames(newsp21) == "52")  0.0005202  0.0035825   0.145   0.8848
##
## (Intercept)                           ***
## newsp21$`% Hisp.`                     *
## newsp21$`% Black`                     .
## newsp21$`% Unemp.`
## as.numeric(rownames(newsp21) == "52")
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.003554 on 97 degrees of freedom
## Multiple R-squared:  0.1034, Adjusted R-squared:  0.06643
## F-statistic: 2.797 on 4 and 97 DF,  p-value: 0.03022
```

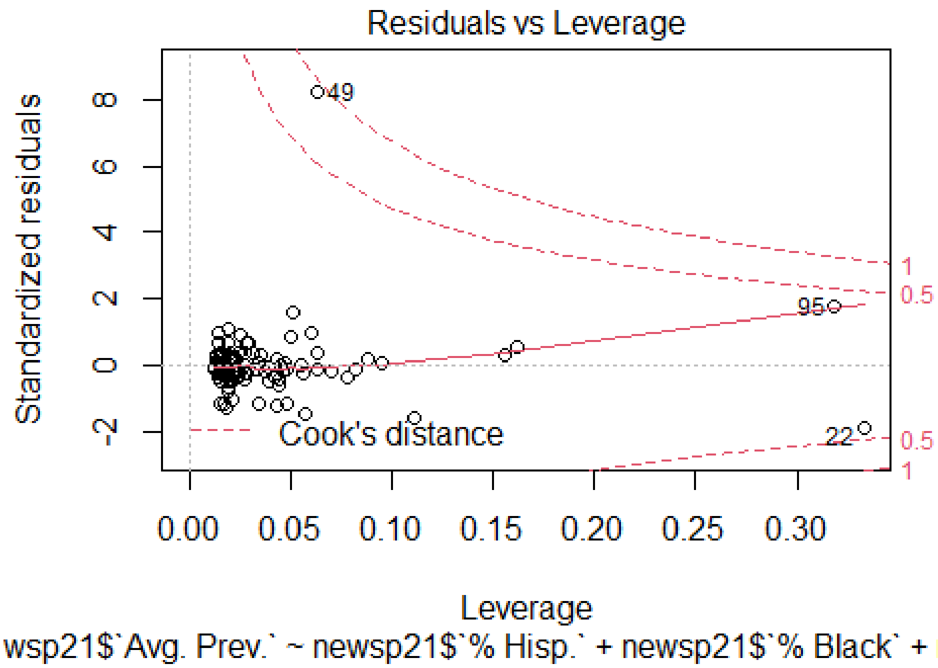
```
plot(mod2fitSp21)
```

```
## Warning: not plotting observations with leverage one:
##   52
```

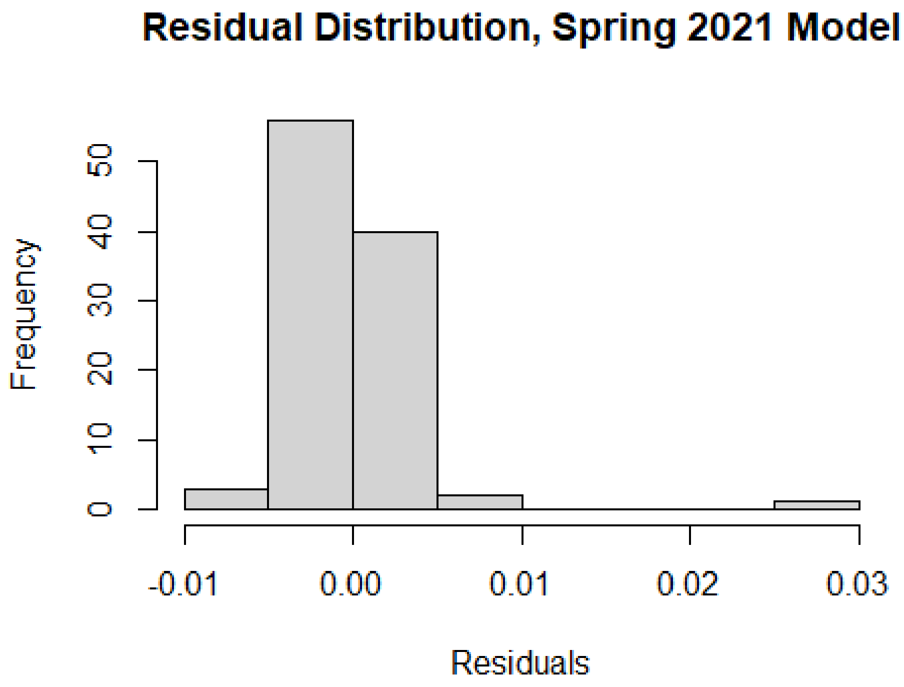

Residuals vs Fitted
Residuals
0.00
0.01
0.02
0.03
49
56
22
0.004
0.005
0.006
0.007
0.008
0.009
Fitted values
wsp21$`Avg. Prev.` ~ newsp21$`% Hisp.` + newsp21$`% Black` + ne


Normal Q-Q
Standardized residuals
-2
0
2
4
6
8
49
95
22
-2
-1
0
1
2
Theoretical Quantiles
wsp21$`Avg. Prev.` ~ newsp21$`% Hisp.` + newsp21$`% Black` + ne





```
hist(mod2fitSp21$residuals, main = "Residual Distribution, Spring 2021 Model", xlab = "Residuals")
```



```
dwtest(mod2fitSp21) # serial autocorrelation not likely

##
##  Durbin-Watson test
##
## data:  mod2fitSp21
## DW = 2.0818, p-value = 0.6324
## alternative hypothesis: true autocorrelation is greater than 0

bptest(mod2fitSp21) # we want less than significance

##
##  studentized Breusch-Pagan test
##
## data:  mod2fitSp21
## BP = 6.7138, df = 4, p-value = 0.1518
```