COVID-19 Research Group: EDA and Modeling

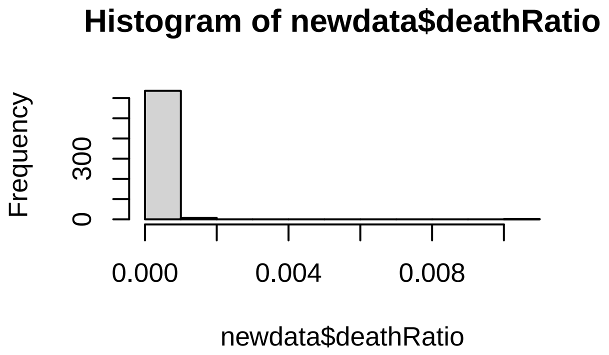
Eddy D. Varela, Chenyang Sun, Jackson Bibbens

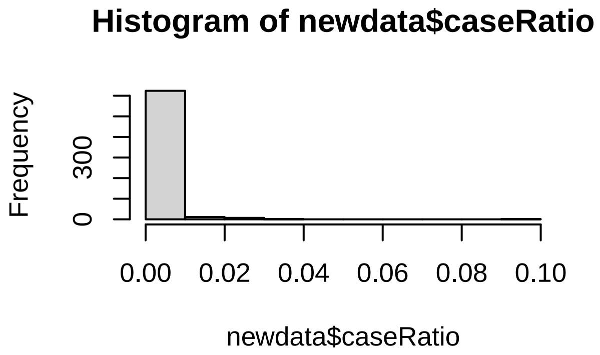
4/27/2020

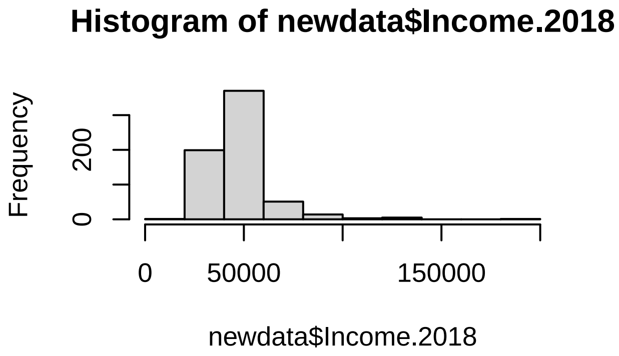
```
library(data.table)
countyData = read.csv("./data/countyData.csv")
countyData$deathRatio<-(countyData$totalDeaths)/(countyData$Pop.2019)
countyData$caseRatio<-(countyData$totalCases)/(countyData$Pop.2019)
countyData<-countyData[which(!is.na(countyData$Beds)),]
countyData<-countyData[which(!is.na(countyData$Income.2018)),]
newdata<-countyData[,c("deathRatio", "caseRatio", "Income.2018","Beds","StayHomeDate","Pop.2019","popDen
```

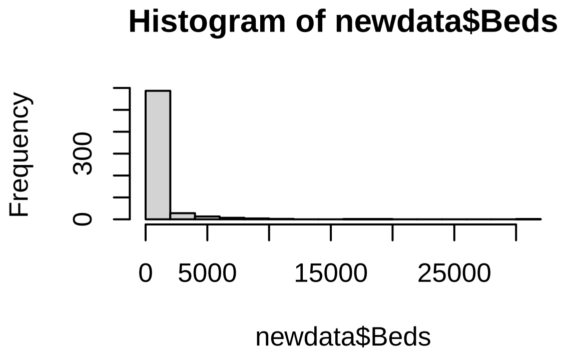
There are a small number of observations for which there are not Beds or Income measurements; those are discarded. For sake of continuity, we code date as an ordered categrical variable in the form of serial date. Examining the histograms of the variables, we see that every single one except for date displays strong right skew, which suggests a logarithmic transform for each of them. Note that the +0.00001 is added to include measurements of 0, since log(0) is undefined.

```
par(mfrow=c(2,2))
hist(newdata$deathRatio)
hist(newdata$caseRatio)
hist(newdata$Income.2018)
hist(newdata$Beds)
```



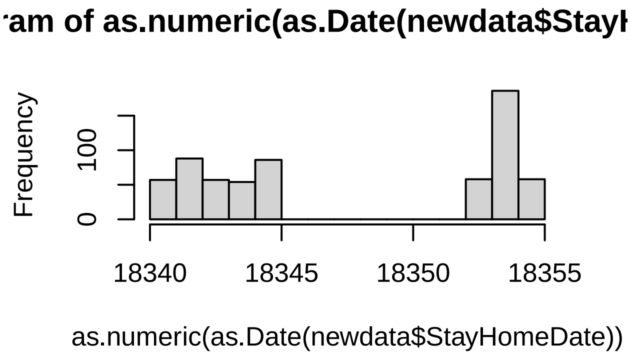


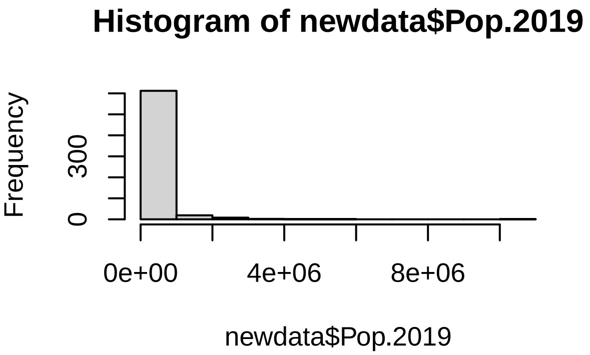


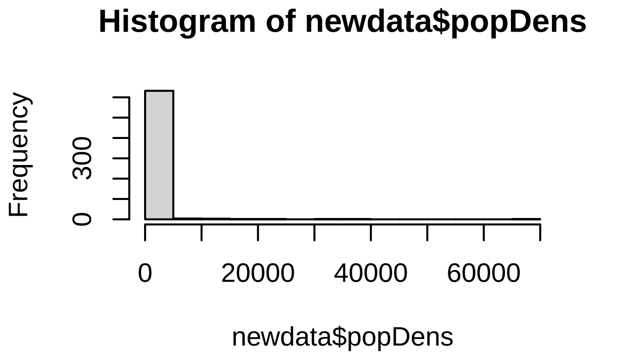


```r
hist(as.numeric(as.Date(newdata$StayHomeDate)))
hist(newdata$Pop.2019)
hist(newdata$popDens)


#Logging variables
newdata$Income<-log((newdata$Income))
newdata$date<-as.numeric(as.Date(newdata$StayHomeDate))
newdata$log.Beds<-log(newdata$Beds)
newdata$log.deathRatio<-log(newdata$deathRatio+0.00001)
newdata$log.caseRatio<-log(newdata$caseRatio+0.00001)
newdata$log.pop<-log(newdata$Pop.2019+0.00001)
newdata$log.popDens<-log(newdata$popDens)
newdata$log.Income<-log(newdata$Income.2018)
```

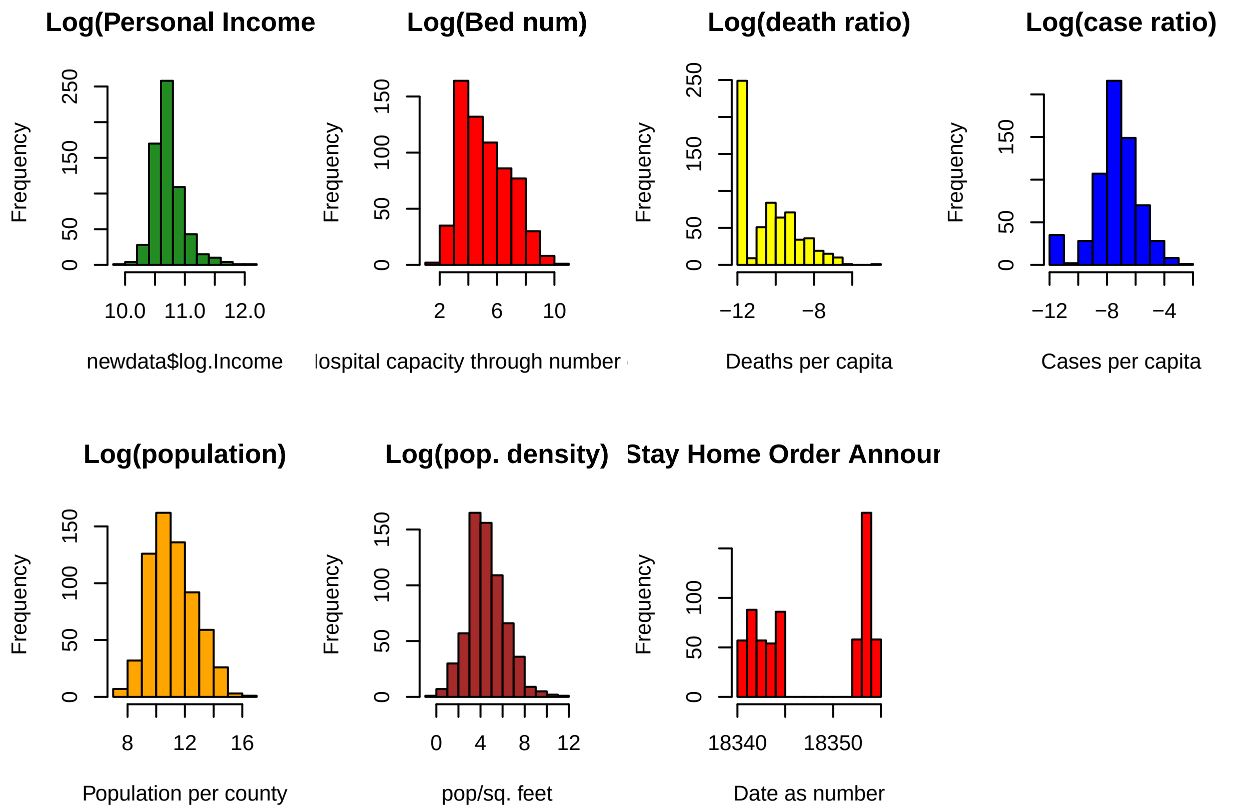




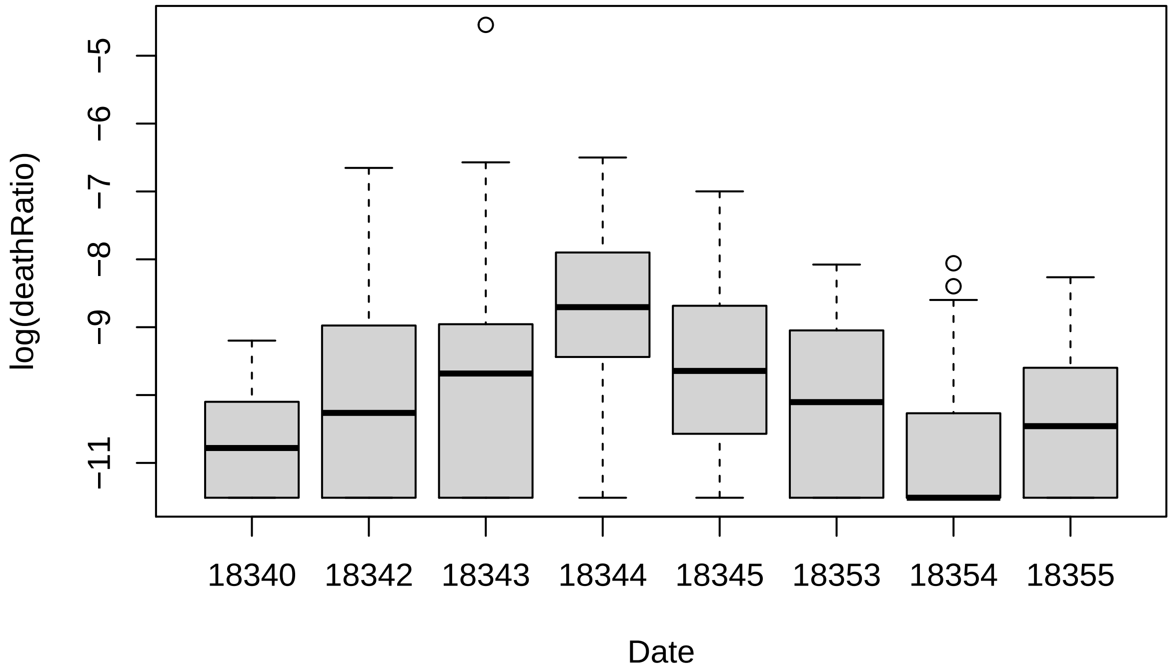


```r
library(xtable)
par(mfrow = c(2,4))
hist(newdata$log.Income, main = "Log(Personal Income)", col = 'forestgreen')
hist(newdata$log.Beds, main = "Log(Bed num)", col = "red", xlab = " Hospital capacity through number of
hist(newdata$log.deathRatio, main = "Log(death ratio)", xlab= "Deaths per capita",col = "yellow")
hist(newdata$log.caseRatio, main = "Log(case ratio)", xlab = "Cases per capita", col = "blue")
hist(newdata$log.pop, main = "Log(population)", xlab= "Population per county", col = "orange")
hist(newdata$log.popDens, main = "Log(pop. density)", xlab = "pop/sq. feet", col = "brown")
hist(as.numeric(as.Date(newdata$StayHomeDate)), main ="Stay Home Order Announced", xlab ="Date as numbe
table(newdata$StayHomeDate)
```

```
##
## 2020-03-19 2020-03-21 2020-03-22 2020-03-23 2020-03-24 2020-04-01 2020-04-02
##         57         88         57         54         86         58        186
## 2020-04-03
##         58
```

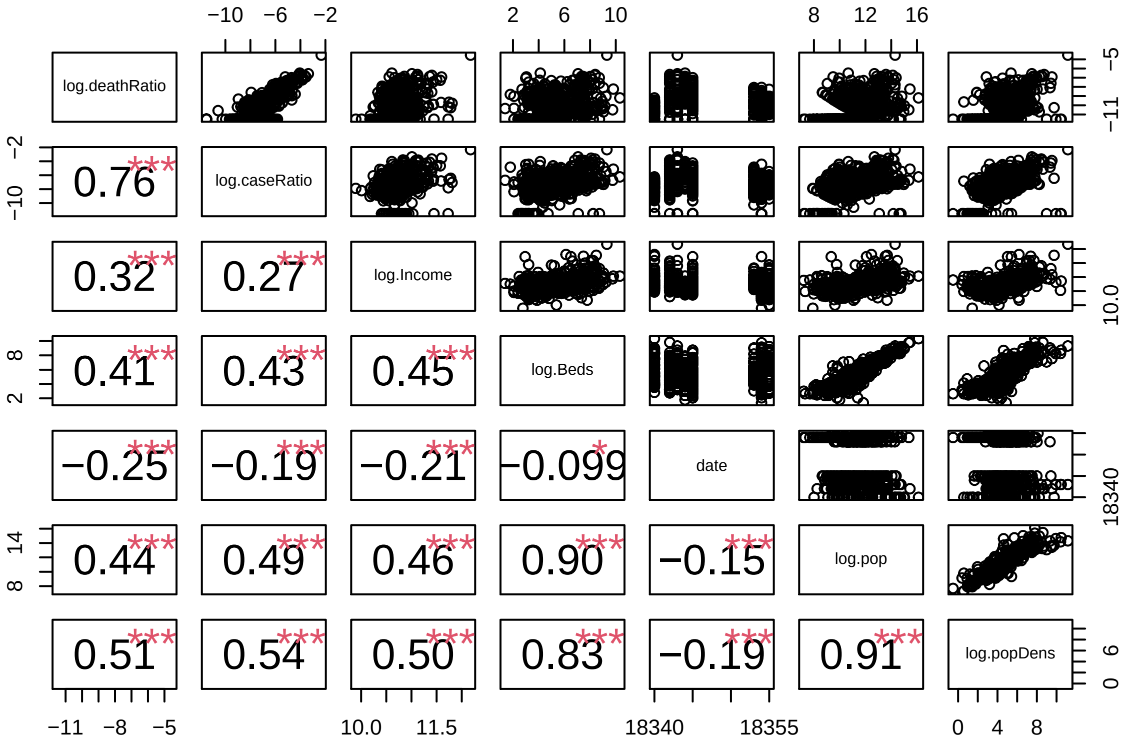


```r
boxplot(newdata$log.deathRatio~newdata$date,xlab="Date",ylab="log(deathRatio)")
```



Now we examine a paired scatterplot. We see quite a bit of multicollinearity between log(Beds) and population metrics, which makes sense since a large population mandates the construction of more beds. Also, some scatterplots seem to have a triangular formation; this may be attributed to the rarity of observations with extremely large values, which is made more evident by the logarithmic transform.

```r
vars<-with(newdata, cbind(log.deathRatio,log.caseRatio,log.Income,log.Beds,date,log.pop,log.popDens))
source("./dependencies/panelfxns.R")
pairs(vars,lower.panel=panel.cor)
```



We now fit the initial model with all predictors. Suspecting an interaction between population density and case ratio, we also add the interaction term.

```
lm.init<-lm(log.deathRatio~log.caseRatio+log.Income+log.Beds+date+log.pop+log.popDens+log.popDens*log.ca
summary(lm.init)
```

```
##
## Call:
## lm(formula = log.deathRatio ~ log.caseRatio + log.Income + log.Beds +
##     date + log.pop + log.popDens + log.popDens * log.caseRatio,
##     data = newdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.12267 -0.57358 -0.05056  0.59344  2.50930
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                398.499434 105.261000   3.786 0.000168 ***
## log.caseRatio                0.252035   0.045173   5.579 3.58e-08 ***
## log.Income                  -0.013042   0.153195  -0.085 0.932180
## log.Beds                     0.002997   0.044045   0.068 0.945768
## date                        -0.022162   0.005724  -3.872 0.000119 ***
## log.pop                     -0.096375   0.064794  -1.487 0.137405
## log.popDens                  0.705394   0.080874   8.722  < 2e-16 ***
## log.caseRatio:log.popDens    0.069972   0.008289   8.441  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.7943 on 636 degrees of freedom
## Multiple R-squared:  0.6503, Adjusted R-squared:  0.6464
## F-statistic: 168.9 on 7 and 636 DF,  p-value: < 2.2e-16
```

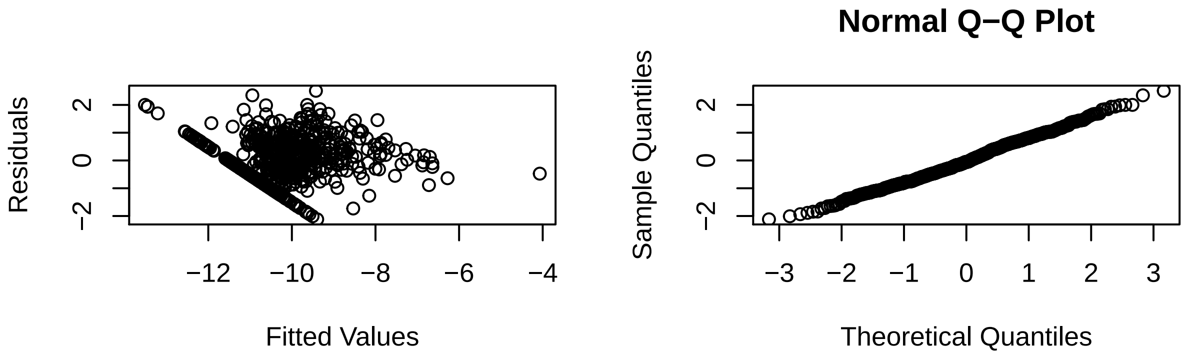
Looking at the diagnostics, we see that the residual plot looks patternless and uniform in variance except for the line corresponding to zero deaths. The normal quantile plot looks very straight with no deviations even at the tails, normality is satisfied.

We imediately see that Income and Beds are each not significant given the other predictors, with p-values above 0.9; in fact, after removing one of them, the resulting model allows the removal of the other. We fit a model with both removed:

```
lm.final<-lm(log.deathRatio~log.caseRatio+date+log.pop+log.popDens+log.popDens*log.caseRatio, data=newda
summary(lm.final)
```

```
##
## Call:
## lm(formula = log.deathRatio ~ log.caseRatio + date + log.pop +
##     log.popDens + log.popDens * log.caseRatio, data = newdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.11845 -0.57277 -0.05284  0.59032  2.50833
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               396.398682 103.239667   3.840 0.000136 ***
## log.caseRatio               0.252532   0.042433   5.951 4.39e-09 ***
## date                       -0.022056   0.005628  -3.919 9.85e-05 ***
## log.pop                    -0.093653   0.049193  -1.904 0.057386 .
## log.popDens                 0.703855   0.073638   9.558  < 2e-16 ***
## log.caseRatio:log.popDens   0.069859   0.007657   9.124  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.7931 on 638 degrees of freedom
## Multiple R-squared:  0.6503, Adjusted R-squared:  0.6475
## F-statistic: 237.2 on 5 and 638 DF,  p-value: < 2.2e-16
```

We have arrived at a good candidate for a final model; all predictors and interactions also happen to be significant. We re-evaluate the diagnostics.



Again, aside from the line corresponding to the huge number of observations with zero deaths, the residual plot looks patternless and uniform in variance. The normal quantile plot looks very straight with no deviations even at the tails; normality is well satisfied. The final model is settled.