covid project

Claire Morrison

2/2/2022

Introduction:

I am interested in population density and a potential difference in COVID-19 cases/deaths. There are a few factors that could play in: in more densely populated areas, transmission seems more likely, thus expected an increase in cases from sparse to densely populated areas. However, in more urban places, you would expect people might be more likely to have vaccines (given the more urbran lean of cities towards the progressive side of the political spectrum), so potentially the severity of cases (deaths) would be less. Finally the number of hospital beds in rual areas might be fewer per 1000 people, and therefore if cases are severe people might not be able to fully get the help they need to recover, leading to perhaps more deaths.

Overall, I hypothesize more densely populated areas will have more cases but not more deaths than sparsely populated areas, scaled by population.

Data: The COVID-19 data came from the New York Times publicly available via Github. It included 1,932 US counties with cases per day and dearths per day recorded daily from Jan 21 2020 to Feb 02 2022. I merged this with a population denisty dataset I downloaded from the US Census https://covid19.census.gov/datasets/USCensus::average-household-size-and-population-density-county/explore?location=23.148887%2C0.315550%2C2.06&showTable=true. However, I could not figure out how to get the link from that website to download via URL with read_csv, so I downlaoded it and uploaded it to my github so you could knit this file without having to read one in manually. Finally, I merged the data with a look up table that contained population so I could control for deaths and cases per population, as that seems important to distinguish when looking at densely vs sparsely populated areas.

```
dat <-
  read_csv("https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv")

deaths_by_county <- dat %>% group_by(county) %>%
  mutate(total_deaths= sum(deaths),
         total_cases= sum(cases)) %>%
    filter(date == max(date)) %>%
  rename(GEOID= fips)
```

```
popdens<- read_csv("https://raw.githubusercontent.com/clairemo22/covid_project/main/Average_Household_S

popdens<- popdens %>% select(GEOID, B01001_calc_PopDensity) %>%
  rename(pop_density=B01001_calc_PopDensity) %>%
  mutate(urb_vs_rural= ifelse(pop_density<38, "rural", ifelse(pop_density>37, "urban", NA))) ### create
```

```
uid_lookup_url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/U


uid <- read_csv(uid_lookup_url) %>%
  select(-c(Lat, Long_, Combined_Key, code3, iso2, iso3, Admin2)) %>%
  rename(GEOID=FIPS)
```

```
US_pop<- deaths_by_county %>% left_join(popdens, x_names = GEOID, y_names=GEOID)
US_pop<- US_pop %>% left_join(uid, x_names = GEOID, y_names=GEOID)
```

```
US_pop$fatality_rate<- (US_pop$deaths/US_pop$cases)*100 ## fatality rate
US_pop$cases_per_county<- (US_pop$cases/US_pop$Population)*100 ## cases per population
US_pop$deaths_per_county<- (US_pop$deaths/US_pop$Population)*100 ## deaths per population


glimpse(US_pop) ### we can only focus on US, so can filter out the rest
```

```
## Rows: 29,844
## Columns: 17
## Groups: county [1,932]
## $ date              <date> 2022-02-03, 2022-02-03, 2022-02-03, 2022-02-03, 2022-02-03, 2022-02-03,~
## $ county            <chr> "Autauga", "Baldwin", "Barbour", "Bibb", "Blount", "Bullock", "Butler", ~
## $ state             <chr> "Alabama", "Alabama", "Alabama", "Alabama", "Alabama", "Alabama", "Alaba~
## $ GEOID             <chr> "01001", "01003", "01005", "01007", "01009", "01011", "01013", "01015", ~
## $ cases             <dbl> 14826, 53083, 5297, 6158, 14158, 2245, 4830, 30342, 8239, 4869, 10040, 1~
## $ deaths            <dbl> 168, 616, 85, 96, 208, 48, 109, 558, 148, 69, 183, 33, 93, 72, 63, 208, ~
## $ total_deaths      <dbl> 53009, 239825, 41566, 260940, 157547, 19351, 598749, 395599, 88627, 4284~
## $ total_cases       <dbl> 3676315, 15018297, 2140314, 10550628, 11727421, 662940, 38084268, 204090~
## $ pop_density       <dbl> 35.853419, 50.541504, 11.247981, 13.973114, 34.515816, 6.417620, 9.95277~
## $ urb_vs_rural      <chr> "rural", "urban", "rural", "rural", "rural", "rural", "rural", "urban", ~
## $ UID               <dbl> 84001001, 84001003, 84001005, 84001007, 84001009, 84001011, 84001013, 84~
## $ Province_State    <chr> "Alabama", "Alabama", "Alabama", "Alabama", "Alabama", "Alabama", "Alaba~
## $ Country_Region    <chr> "US", "US", "US", "US", "US", "US", "US", "US", "US", "US", "US", "US", ~
## $ Population        <dbl> 55869, 223234, 24686, 22394, 57826, 10101, 19448, 113605, 33254, 26196, ~
## $ fatality_rate     <dbl> 1.133144, 1.160447, 1.604682, 1.558948, 1.469134, 2.138085, 2.256729, 1.~
## $ cases_per_county  <dbl> 26.53708, 23.77908, 21.45751, 27.49844, 24.48380, 22.22552, 24.83546, 26~
## $ deaths_per_county <dbl> 0.3007034, 0.2759436, 0.3443247, 0.4286863, 0.3596998, 0.4752005, 0.5604~
```

```
US_pop<- US_pop %>% filter(Country_Region=="US")
US_pop<- US_pop[complete.cases(US_pop),] ### there was a lot of missing data with pop density or popula

US_pop %>% select(county, state, fatality_rate, cases_per_county, deaths_per_county) %>%
  arrange(deaths_per_county) ### fewest deaths per county
```

```
## # A tibble: 3,128 x 5
## # Groups:   county [1,842]
##    county               state         fatality_rate cases_per_county deaths_per_county
##    <chr>                <chr>                 <dbl>            <dbl>              <dbl>
##  1 Skagway Municipality Alaska                    0             9.30                  0
##  2 Alpine               California                0            11.0                   0
##  3 Sierra               California                0            10.2                   0
##  4 Hinsdale             Colorado                  0            14.3                   0
##  5 Jackson              Colorado                  0            10.6                   0
##  6 Kalawao              Hawaii                    0             1.16                  0
##  7 Clark                Idaho                     0            12.4                   0
##  8 Dukes                Massachusetts             0            19.1                   0
##  9 Nantucket            Massachusetts             0            28.7                   0
## 10 Hayes                Nebraska                  0            10.8                   0
## # ... with 3,118 more rows
```

```
US_pop %>% select(county, state, fatality_rate, cases_per_county, deaths_per_county) %>%
  arrange(desc(deaths_per_county)) ### most deaths per county
```

```
## # A tibble: 3,128 x 5
## # Groups:   county [1,842]
##    county       state        fatality_rate cases_per_county deaths_per_county
##    <chr>        <chr>                <dbl>            <dbl>             <dbl>
##  1 Galax city   Virginia              3.10             36.6             1.13
##  2 McMullen     Texas                 5.13             21.0             1.08
##  3 Hancock      Georgia               5.65             18.0             1.02
##  4 Robertson    Kentucky              3.68             27.0             0.996
##  5 Foard        Texas                 5.67             16.8             0.952
##  6 Jerauld      South Dakota          4.81             19.6             0.944
##  7 Motley       Texas                 4.20             21.8             0.917
##  8 Emporia city Virginia              4.51             19.5             0.879
##  9 Gove         Kansas                3.03             28.8             0.873
## 10 Buffalo      South Dakota          2.79             31.0             0.866
## # ... with 3,118 more rows
```

```
US_pop %>% select(county, state, fatality_rate, cases_per_county, deaths_per_county) %>%
  arrange(fatality_rate) ### lowest fatality rate
```

```
## # A tibble: 3,128 x 5
## # Groups:   county [1,842]
##    county               state         fatality_rate cases_per_county deaths_per_county
##    <chr>                <chr>                 <dbl>            <dbl>             <dbl>
##  1 Skagway Municipality Alaska                    0             9.30                 0
##  2 Alpine               California                0            11.0                  0
##  3 Sierra               California                0            10.2                  0
##  4 Hinsdale             Colorado                  0            14.3                  0
##  5 Jackson              Colorado                  0            10.6                  0
##  6 Kalawao              Hawaii                    0             1.16                 0
##  7 Clark                Idaho                     0            12.4                  0
##  8 Dukes                Massachusetts             0            19.1                  0
##  9 Nantucket            Massachusetts             0            28.7                  0
## 10 Hayes                Nebraska                  0            10.8                  0
## # ... with 3,118 more rows
```

```
US_pop %>% select(county, state, fatality_rate, cases_per_county, deaths_per_county) %>%
  arrange(desc(fatality_rate)) ### highest fatality rate
```

```
## # A tibble: 3,128 x 5
## # Groups:   county [1,842]
##    county   state        fatality_rate cases_per_county deaths_per_county
##    <chr>    <chr>                <dbl>            <dbl>             <dbl>
##  1 Sabine   Texas                 6.54             11.3             0.740
##  2 Foard    Texas                 5.67             16.8             0.952
##  3 Hancock  Georgia               5.65             18.0             1.02
##  4 Harding  New Mexico            5.56             11.5             0.64
##  5 McMullen Texas                 5.13             21.0             1.08
##  6 Grant    Nebraska              4.95             16.2             0.803
##  7 Blaine   Nebraska              4.92             13.1             0.645
##  8 Jerauld  South Dakota          4.81             19.6             0.944
##  9 Knox     Texas                 4.63             13.0             0.600
## 10 Hooker   Nebraska              4.63             15.8             0.733
## # ... with 3,118 more rows
```

```
by_state <- US_pop %>% group_by(state)
avgs <- summarise(by_state,
```

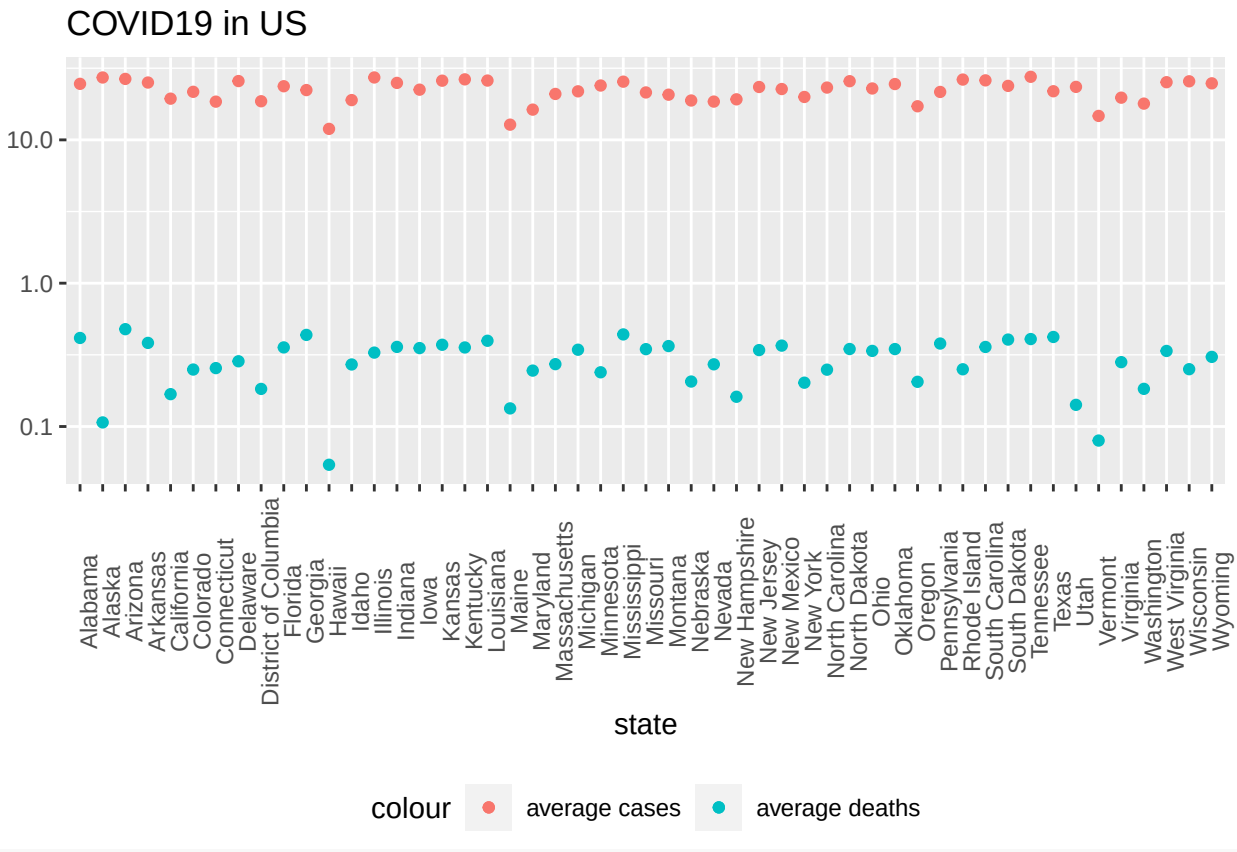
```
  num_counties = n(),
  avg_cases = mean(cases_per_county, na.rm = TRUE),
  avg_deaths = mean(deaths_per_county, na.rm = TRUE),
  avg_fatality_rate = mean(fatality_rate, na.rm = TRUE))

avgs %>% arrange(desc(avg_fatality_rate)) ## highest fatality rates by state
```

```
## # A tibble: 51 x 5
##    state        num_counties avg_cases avg_deaths avg_fatality_rate
##    <chr>               <int>     <dbl>      <dbl>             <dbl>
##  1 Texas                 254      21.8      0.422              2.06
##  2 Georgia               159      22.3      0.436              2.00
##  3 Arizona                15      26.6      0.478              1.80
##  4 Montana                56      20.7      0.364              1.78
##  5 Pennsylvania           67      21.6      0.379              1.77
##  6 Mississippi            82      25.5      0.439              1.74
##  7 South Dakota           66      23.8      0.404              1.72
##  8 New Mexico             32      22.6      0.367              1.71
##  9 Alabama                67      24.6      0.415              1.69
## 10 Missouri              115      21.4      0.347              1.65
## # ... with 41 more rows
```

```
avgs %>% arrange(avg_fatality_rate) ## lowest fatality rates by state
```

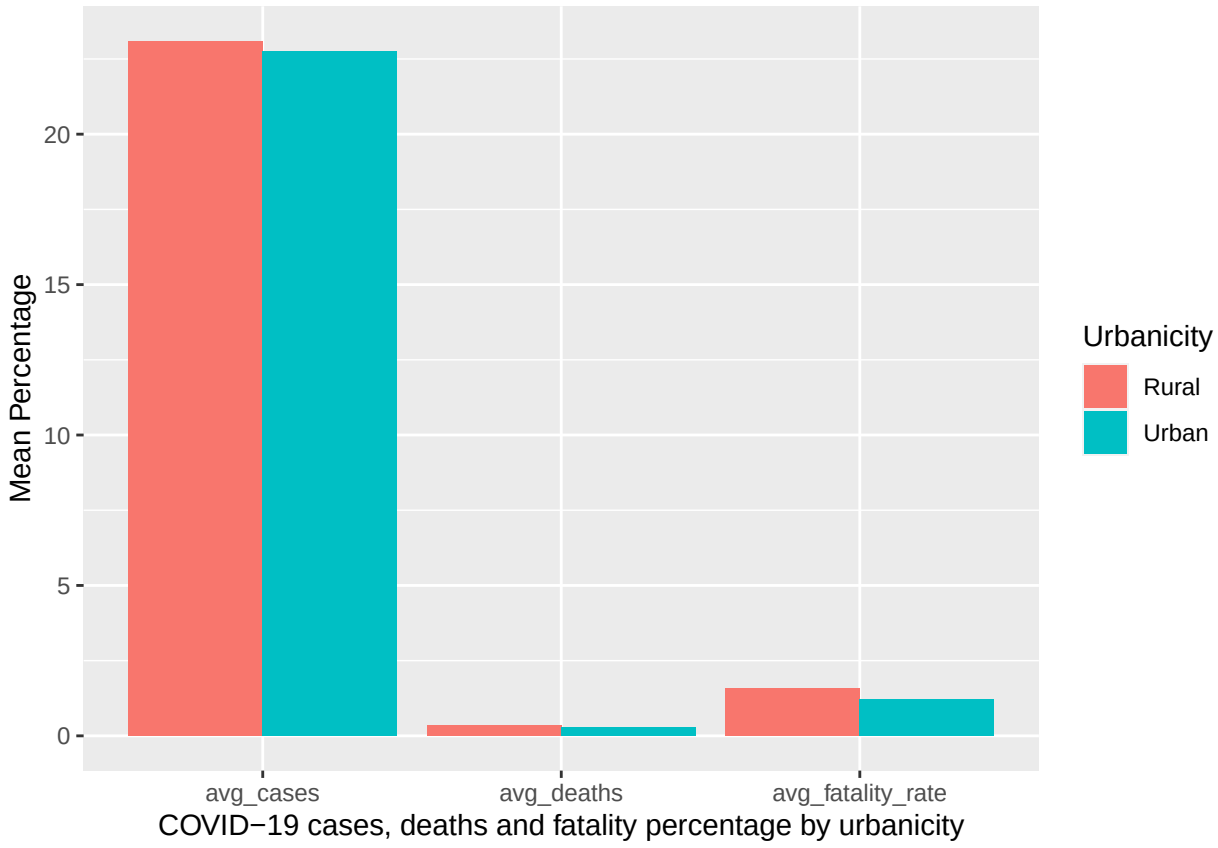
```
## # A tibble: 51 x 5
##    state                num_counties avg_cases avg_deaths avg_fatality_rate
##    <chr>                       <int>     <dbl>      <dbl>             <dbl>
##  1 Hawaii                          5      11.9     0.0541             0.376
##  2 Alaska                         25      27.2     0.107              0.391
##  3 Vermont                        14      14.7     0.0798             0.529
##  4 Utah                           28      23.4     0.142              0.615
##  5 New Hampshire                  10      19.2     0.161              0.827
##  6 California                     58      19.4     0.168              0.837
##  7 Rhode Island                    5      26.4     0.251              0.921
##  8 District of Columbia            1      18.6     0.183              0.983
##  9 Wisconsin                      72      25.6     0.251              0.990
## 10 Minnesota                      87      23.9     0.239              1.00
## # ... with 41 more rows
```

```
avgs %>%
  filter(avg_cases > 0) %>%
  ggplot(aes(x = state, y = avg_cases)) +
  geom_point(aes(color = "average cases")) +
  geom_point(aes(y = avg_deaths, color = "average deaths")) +
    scale_y_log10() +
    theme(legend.position="bottom",
          axis.text.x = element_text(angle = 90)) +
  labs(title = "COVID19 in US", y= NULL)
```




```
by_urb <- US_pop %>% group_by(urb_vs_rural)
avgs_urb <- summarise(by_urb,
  num_counties = n(),
  avg_cases = mean(cases_per_county, na.rm = TRUE),
  avg_deaths = mean(deaths_per_county, na.rm = TRUE),
  avg_fatality_rate = mean(fatality_rate, na.rm = TRUE))

avgs_urb<- avgs_urb %>% pivot_longer(
  cols = starts_with("avg"),
  names_to = "var",
  values_to = "rate")
```

```
ggplot(avgs_urb,aes(x=var,y=rate,fill=factor(urb_vs_rural)))+
  geom_bar(stat="identity",position="dodge")+
  xlab("COVID-19 cases, deaths and fatality percentage by urbanicity")+ylab("Mean Percentage") + scale_
```



Cases per county

```
summary(lmer(cases_per_county~pop_density+ (1|state), US_pop))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: cases_per_county ~ pop_density + (1 | state)
##    Data: US_pop
##
## REML criterion at convergence: 18164.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.1236 -0.5911 -0.0346  0.5223 13.4041
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  state    (Intercept) 12.28    3.505
##  Residual             18.36    4.285
## Number of obs: 3128, groups:  state, 51
##
## Fixed effects:
##               Estimate Std. Error t value
## (Intercept) 22.3795069  0.5081428  44.042
## pop_density -0.0002501  0.0002669  -0.937
##
## Correlation of Fixed Effects:
##             (Intr)
## pop_density -0.075
```

Deaths per county

```
summary(lmer(deaths_per_county~pop_density + (1|state), data = US_pop))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: deaths_per_county ~ pop_density + (1 | state)
##    Data: US_pop
##
## REML criterion at convergence: -3959.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.3929 -0.6158 -0.0765  0.5471  6.8312
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  state    (Intercept) 0.008721 0.09338
##  Residual             0.015534 0.12463
## Number of obs: 3128, groups:  state, 51
##
## Fixed effects:
##               Estimate Std. Error t value
## (Intercept)  3.057e-01  1.361e-02  22.464
## pop_density -4.885e-05  7.747e-06  -6.306
##
## Correlation of Fixed Effects:
##             (Intr)
## pop_density -0.080
```

Case fatality rate

```
summary(lmer(fatality_rate~pop_density + (1|state), US_pop))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: fatality_rate ~ pop_density + (1 | state)
##    Data: US_pop
##
## REML criterion at convergence: 5763.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.4758 -0.5869 -0.1071  0.4460  7.5661
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  state    (Intercept) 0.1234   0.3512
##  Residual             0.3508   0.5923
## Number of obs: 3128, groups:  state, 51
##
## Fixed effects:
##               Estimate Std. Error t value
## (Intercept)  1.373e+00  5.208e-02  26.368
## pop_density -2.252e-04  3.661e-05  -6.151
##
## Correlation of Fixed Effects:
##             (Intr)
```

```
## pop_density -0.091
```

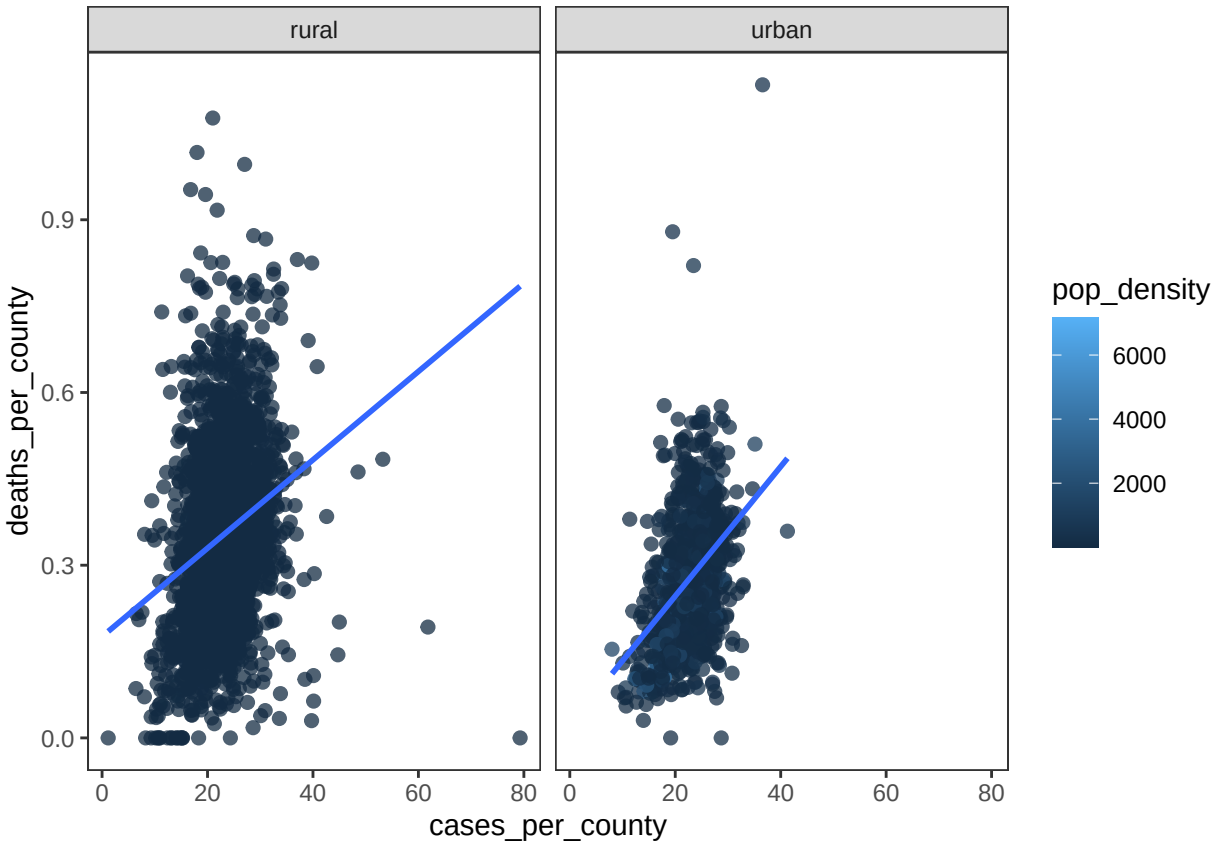
```
summary(lmer(deaths_per_county~cases_per_county+pop_density+cases_per_county*pop_density + (1|state), da
```

```
## Warning: Some predictor variables are on very different scales: consider rescaling
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: deaths_per_county ~ cases_per_county + pop_density + cases_per_county *
##     pop_density + (1 | state)
##    Data: US_pop
##
## REML criterion at convergence: -4086.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -6.5801 -0.6223 -0.0911  0.5340  6.0712
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  state    (Intercept) 0.006576 0.08109
##  Residual             0.014796 0.12164
## Number of obs: 3128, groups:  state, 51
##
## Fixed effects:
##                                Estimate Std. Error t value
## (Intercept)                   1.587e-01  1.655e-02   9.587
## cases_per_county              6.577e-03  5.162e-04  12.741
## pop_density                  -5.531e-05  3.162e-05  -1.749
## cases_per_county:pop_density  4.210e-07  1.569e-06   0.268
##
## Correlation of Fixed Effects:
##             (Intr) css_p_ pp_dns
## css_pr_cnty -0.694
## pop_density -0.129  0.198
## css_pr_cn:_  0.115 -0.199 -0.971
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
```

```
ggplot(US_pop,aes(cases_per_county, deaths_per_county, color=pop_density, na.rm = T)) +
  facet_wrap(~ urb_vs_rural)+
  geom_point(size = 2, alpha = .75, position = "jitter", na.rm = T) +
  geom_smooth(na.rm = T, method = "lm", se = F, linetype = 1)+
  theme_bw()+
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Conclusion:

The three main statistical findings are: population density does not significantly predict COVID cases, but it does significantly predit COVID deaths and fatality rates. We controlled for any non-independence by state that could arise due to to state-wide laws such as mask mandates, shut downs, etc.

I found that as as population density increases, COVID case percentage per population *decreases* by a small and insignificant amount. This means that case percentage is actually slightly smaller for more densely populated areas, but again, insignificantly. As population density increases, death percentage per population also decreases, but by a more significant amount. Lastly, and as expected given the first two results, as population density increases, fatality rate also significantly decreases. As somewhat of a check on the data, I ran and plotted a model where percent of COVID cases per county and population density predict percent of COVID deaths per county. COVID cases per county do significantly predict deaths per county, but there was no significant interaction of population density on that relationship, perhaps because I had already controlled for population size in the variables.

These results are somewhat contrary to what I hypothesized. Here, worse COVID outcomes (deaths) seem overall better in more densely populated areas, but population density did not make a difference on cases per county. Better COVID outcomes in densley populated areas could be due to a number of factors not modelled here, such as socio economic status, better hospital systems, political beliefs or vaccine status.

Limitations/ bias:

Limitations of this analysis come from using publicly available data. It is a huge benefit that NYT publishes data like this for our own use and transparency, but since we aren't collecting it we cannot always model the ideal variables. Further, sources of bias can come from my own beliefs, such as thinking more rural places do have less progressive views on the pandemic overall. We can also get bias from the data by not knowing how many tests were being administered each day in each county, which is something that can obviously highly skew the data.