Colorado COVID-19 Likelihood of Releases

Albert Sun

9/30/2020

```
library(tidyverse)
library(patchwork)
library(janitor)
library(broom)
library(knitr)
library(here)
library(rms)
library(skimr)
```

Background

We wish to assess the quality of data collection and efficacy of using existing open-source jail databases to understand how jail populations respond to outside legislation and large trends, particularly disasters like COVID-19. The Colorado Jail Database is currently the most comprehensive, open source published, and catalogued state-wide jail database in the United States. Thus, we seek to conduct various statistical analyses on Colorado's jail database to understand how similar data collections can motivate statistical conclusions in the future. If proven insightful, Colorado can be an effective model that other states can look towards to adopt a similar system of data collection.

Some stakeholders in Colorado, and across the country, have attempted to reduce jail populations at the beginning of the outbreak. Colorado Governor Jared Polis signed an executive order relaxing the standards for early release in March in fear of the effects of excessive overcrowding in jails [7]. Jails who initially thought of expanding have halted their plans due to general decreases in jail populations [8]. Overall, Colorado has had a reported net decrease in jail population during the COVID-19 criss, which can immediately be seen on the Colorado dashboard itself.

Research Question

What are the motivating factors or decisions that are correlated with a county jail decreasing its jail population at the outbreak of and during the COVID-19 pandemic? Which effect is greater, decreasing intake of new people or increasing outtake of new people? We seek to impute other demographic variables in a particular region (general population, political profile, etc) and to assess the nearby community's relationship with the likelihood for releases.

Data Preparation

```
#load colorado hb-19 1279
colorado <- read_csv(here("data", "HB19-1297Data.csv")) %>%
  clean_names()

#load colorado population data
```

```r
pop <- read_csv(here("data", "colorado-population.csv")) %>%
  clean_names() %>%
  filter(county != "Total") %>%
  mutate(county = str_sub(county, 2, -18))

#load colorado population demographics data
demo <- read_csv(here("data", "mit-demographics.csv")) %>%
  filter(state == "Colorado")

demo <- demo %>%
  mutate(liberal = factor(if_else(clinton16 - trump16 > 0, 1, 0))) %>%
  mutate(urbanicity = factor(ruralurban_cc)) %>%
  mutate(urbanicity = fct_collapse(urbanicity,
                    metro = c("1", "2", "3"),
                    urban = c("4", "5", "6", "7"),
                    rural = c("8", "9"))) %>%
  mutate(urbanicity = fct_relevel(urbanicity,
                                   'rural',
                                   'urban')) %>%
  select(county, lesscollege_whites_pct, black_pct, rural_pct, urbanicity, liberal)

demo
```

```
## # A tibble: 64 x 6
##    county     lesscollege_whites_pct black_pct rural_pct urbanicity liberal
##    <chr>                       <dbl>     <dbl>     <dbl> <fct>      <fct>
##   1 Adams                       70.2     3.00      3.62  metro      1
##   2 Alamosa                     66.0     1.68     36.9   urban      1
##   3 Arapahoe                    52.7    10.0       1.58  metro      1
##   4 Archuleta                   59.9     0.850    59.4   urban      0
##   5 Baca                        76.4     1.23    100     rural      0
##   6 Bent                        86.5     8.12     38.0   urban      0
##   7 Boulder                     35.9     0.849     8.91  metro      1
##   8 Broomfield                  44.5     0.932     0.583 metro      1
##   9 Chaffee                     62.1     1.28     37.4   urban      0
##  10 Cheyenne                    74.7     0.290   100     rural      0
## # ... with 54 more rows
```

Overall Comments

Overall, the HB 19-1297 data set is the only statewide jail database that offers an open source ".csv" file for the public to see. Even other states who do collect jail population data, like California and Texas, don't have their data published completely open source; instead, researchers are forced to scrape the data periodically from their website. As such, we find that the HB-1297 data set to be the most reproducible thus far.

The 23 columns represent variables: the quarter, the year, county jail, jail management system, etc.

The 2280 rows generally reflect specific jail information per each quarter; however, the reason why there are 2280 rows instead of 152 rows (the number of jails times three quarters in Colorado) is because each jail has 15 rows separated into different areas of measurement, i.e. "Number of inmates", "Sentenced", etc. In more-technical SQL terms, it seems like the measure column was cross joined with jail column.

1. Check Proportion Missing

Instead of leaving blank values in missing columns, the Colorado HB 19-1297 jail dataset uses the `not_available` column to annotate and comment on missingness. They add `0` to a datapoint that is missing. Thus, because we cannot use conventional functions like is.na() to detect missingness, we will take a look at the jail observations that contain missing data.

Here are the 10 most common NA messages.

```
colorado %>%
  count(not_available) %>%
  group_by(not_available) %>%
  arrange(-n) %>%
  ungroup %>%
  slice(1:10)
```

```
## # A tibble: 10 x 2
##    not_available                                                               n
##    <chr>                                                                   <int>
##  1 <NA>                                                                     2437
##  2 JMS does not calculate this                                                24
##  3 Eforce does not seprate this data                                          18
##  4 Not able to capture information                                            18
##  5 ESTIMATES                                                                  17
##  6 My JMS doesn't break unsentenced inmates by Gender or Race/Ethnicity.      16
##  7 We do not have the program to pull these stats from our JMS.               16
##  8 Population down due to COVID-19                                            15
##  9 Data not available                                                         14
## 10 The Eagle County Sheriff's Office made every effort to comply with thi~    14
```

Out of 2280 rows, there are 488 (2280-1792) rows with some sort of `not_available` message.

This means that 19.6491228% of the data has some sort of `not_available` message to it, which is relatively low.

Most of the data exists, and almost all jails at least provide some sort of ethnicity data. Most of the data that is missing is that for specific measures as aforementioned above, a jail's JMS (Jail Management System) might not break down types of sentences by gender, race, or ethnicity. When conducting data analysis on race and gender for some particular measures, it will be a good idea to remove these rows, or at least account for them.

2. Check Class

Check class of data:

```
glimpse(colorado)
```

```
## Rows: 3,060
## Columns: 23
## $ qtr_year          <dbl> 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2...
## $ qtr               <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ county            <chr> "Clear Creek", "Clear Creek", "Clear Creek", "Cle...
## $ jms               <chr> "E-Force", "E-Force", "E-Force", "E-Force", "E-Fo...
## $ capacity          <dbl> 105, 105, 105, 105, 105, 105, 105, 105, 105, 105,...
## $ beds              <dbl> 105, 105, 105, 105, 105, 105, 105, 105, 105, 105,...
## $ deaths            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ bookings          <dbl> 253, 253, 253, 253, 253, 253, 253, 253, 253, 253,...
## $ releases          <dbl> 183, 183, 183, 183, 183, 183, 183, 183, 183, 183,...
## $ measure           <chr> "Number of inmates", "Sentenced", "Unsentenced - ...
## $ total             <dbl> 70.0, 2.0, 4.0, 64.0, 53.0, 11.0, 3.0, 5.0, 2.0, ...
## $ male              <dbl> 62, 2, 4, 56, 46, 10, 2, 5, 2, 78, 118, 194, 85, ...
## $ female            <dbl> 8, 0, 0, 8, 7, 1, 1, 0, 0, 14, 93, 46, 128, 115, ...
## $ other_gender      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ black             <dbl> 10, 1, 1, 8, 8, 0, 1, 2, 0, 12, 19, 23, 26, 27, 0...
## $ native_american   <dbl> 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.6,...
## $ other_race        <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0...
## $ white             <dbl> 58, 1, 2, 55, 44, 11, 2, 3, 2, 76, 0, 205, 13, 56...
## $ unknown_race      <dbl> 2.0, 0.0, 1.0, 1.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.3,...
## $ non_hispanic      <dbl> 43, 2, 2, 39, 31, 8, 2, 5, 2, 56, 178, 161, 16, 4...
## $ hispanic          <dbl> 20, 0, 2, 18, 16, 2, 1, 0, 0, 30, 227, 56, 6, 97,...
## $ unknown_ethnicity <dbl> 7, 0, 0, 7, 6, 1, 0, 0, 0, 6, 46, 23, 10, 85, 0, ...
## $ not_available     <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
```

Judging from the datatypes above, we will change Qtr to become a factor variable, because Qtr represents periodic stages of data collection, not a continuous value.

```
colorado <- colorado %>%
  mutate(qtr = as.factor(qtr))
```

Overall, the other variables seem to have the correct data type.

3. Investigate Missingness

```
colorado <- colorado %>%
  mutate(isNA = !is.na(not_available))

colorado %>%
  count(qtr, isNA) %>%
  filter(isNA == TRUE)
```

```
## # A tibble: 4 x 3
##   qtr   isNA      n
##   <fct> <lgl> <int>
## 1 1     TRUE    170
## 2 2     TRUE    162
## 3 3     TRUE    156
## 4 4     TRUE    135
```

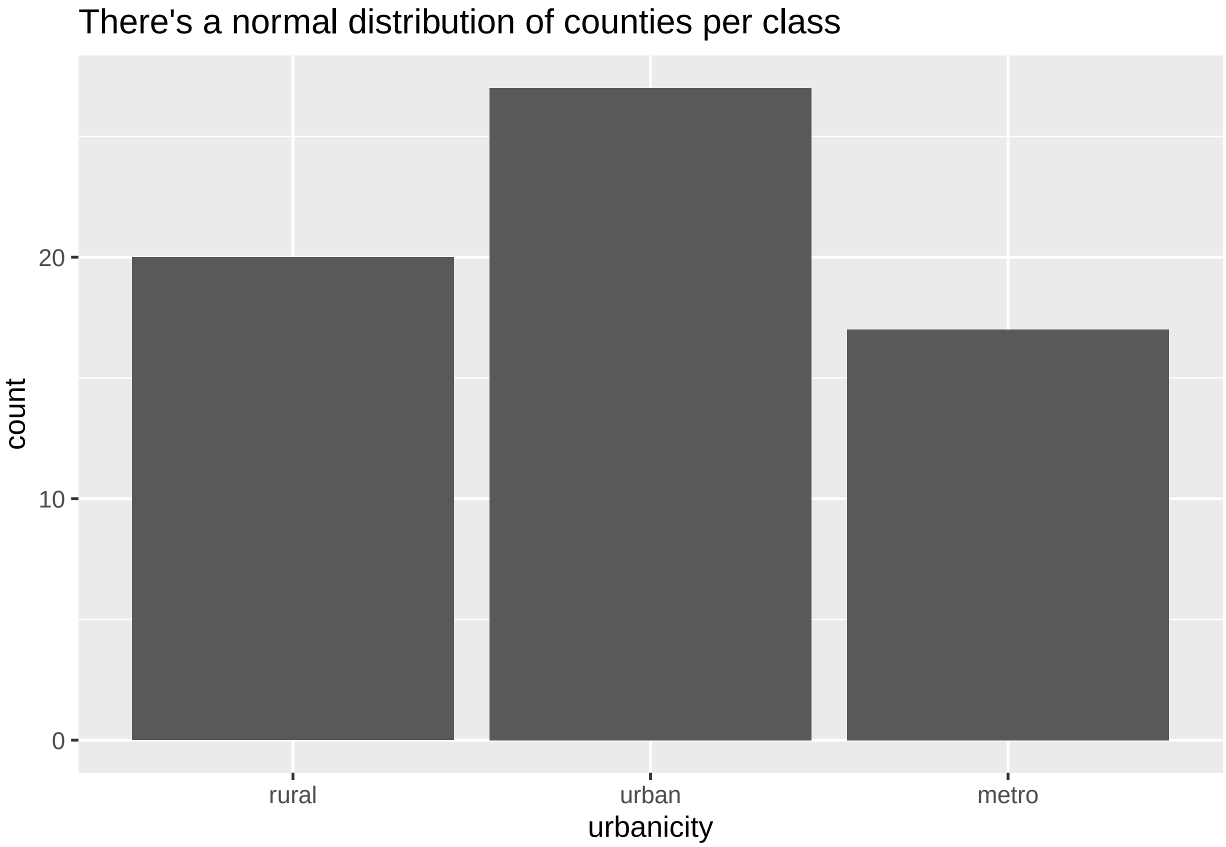
```
colorado
```

```
## # A tibble: 3,060 x 24
##    qtr_year qtr   county jms    capacity  beds deaths bookings releases measure
##       <dbl> <fct> <chr>  <chr>     <dbl> <dbl>  <dbl>    <dbl>    <dbl> <chr>
##  1     2020 1     Clear~ E-Fo~       105   105      0      253      183 Number~
##  2     2020 1     Clear~ E-Fo~       105   105      0      253      183 Senten~
##  3     2020 1     Clear~ E-Fo~       105   105      0      253      183 Unsent~
##  4     2020 1     Clear~ E-Fo~       105   105      0      253      183 Unsent~
##  5     2020 1     Clear~ E-Fo~       105   105      0      253      183 Unsent~
##  6     2020 1     Clear~ E-Fo~       105   105      0      253      183 Unsent~
##  7     2020 1     Clear~ E-Fo~       105   105      0      253      183 Munici~
##  8     2020 1     Clear~ E-Fo~       105   105      0      253      183 Admini~
##  9     2020 1     Clear~ E-Fo~       105   105      0      253      183 Compet~
## 10     2020 1     Clear~ E-Fo~      105   105      0      253      183 Averag~
## # ... with 3,050 more rows, and 14 more variables: total <dbl>, male <dbl>,
## #   female <dbl>, other_gender <dbl>, black <dbl>, native_american <dbl>,
## #   other_race <dbl>, white <dbl>, unknown_race <dbl>, non_hispanic <dbl>,
## #   hispanic <dbl>, unknown_ethnicity <dbl>, not_available <chr>, isNA <lgl>
```

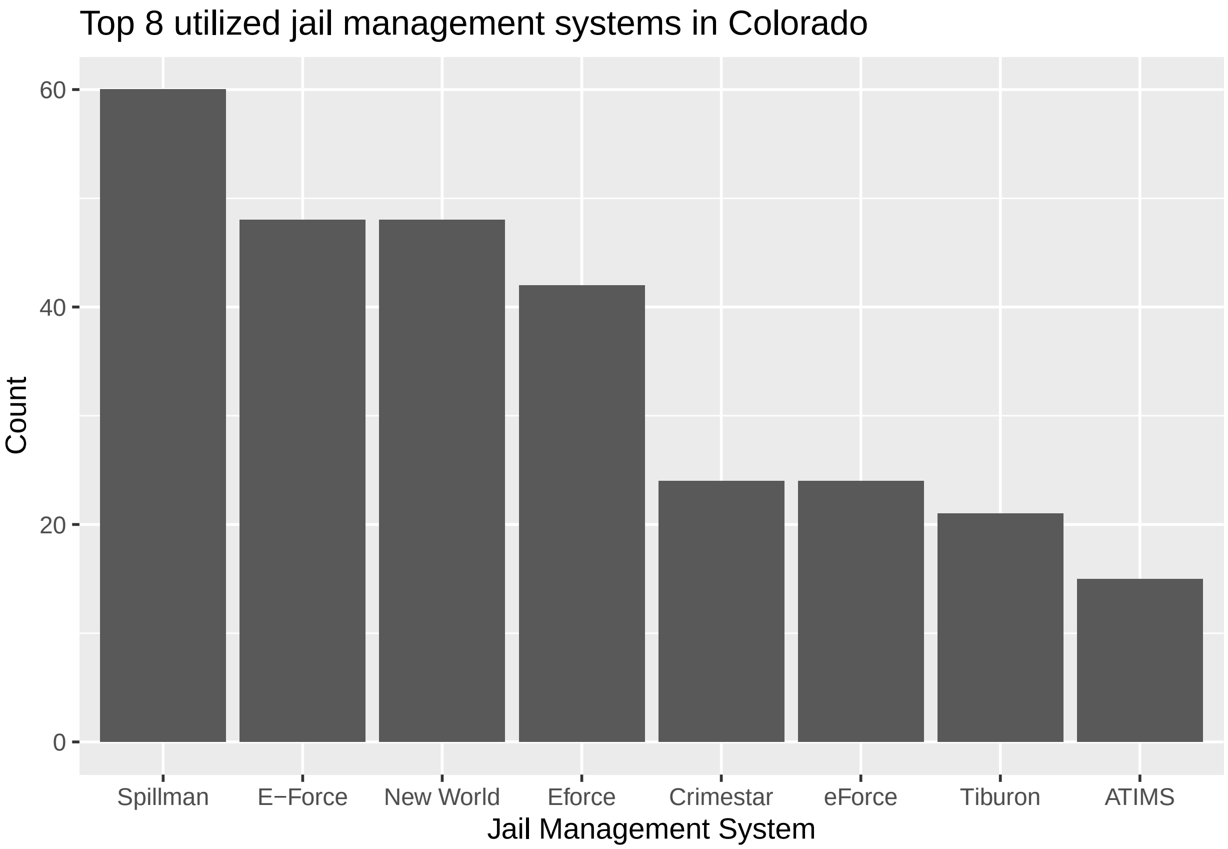
Missingness was generally reduced throughout the three quarters of jail data collection in 2020, possibly suggesting improvements in jail collection throughout this time period.

4. EDA

```
demo %>%
  ggplot(aes(urbanicity)) +
  geom_bar() +
  labs(title = "There's a normal distribution of counties per class")
```



```
colorado %>%
  count(jms) %>%
  arrange(-n) %>%
  mutate(n = n / 5) %>%
  slice(1:8) %>%
  ggplot(aes(x = reorder(jms, -n), y = n)) +
  geom_bar(stat = "identity") +
  labs(x = "Jail Management System",
       y = "Count",
       title = "Top 8 utilized jail management systems in Colorado")
```

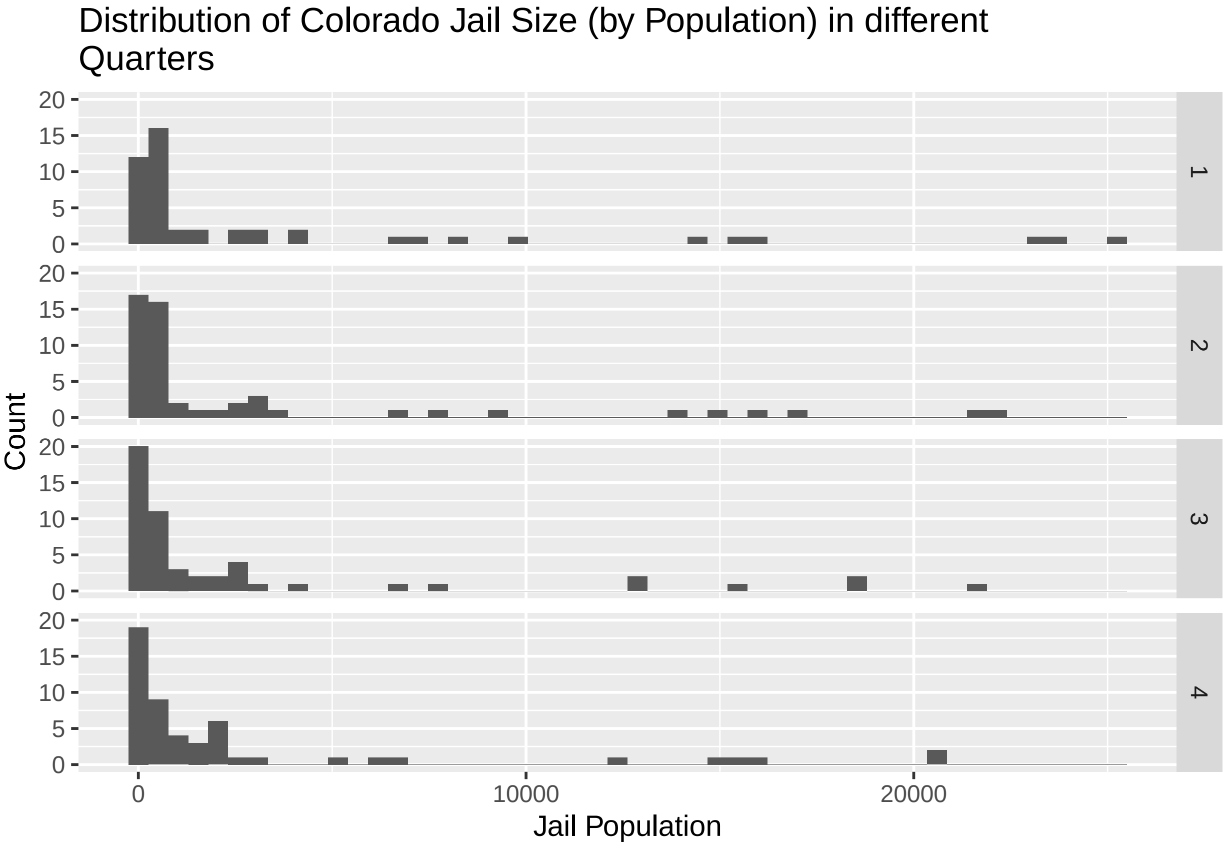


```
colorado_population <- colorado %>%
  arrange(qtr) %>%
  group_by(qtr) %>%
  summarise(total = sum(total))

colorado_population
```

```
## # A tibble: 4 x 2
##   qtr     total
## * <fct>   <dbl>
## 1 1     182537.
## 2 2     161983.
## 3 3     149324.
## 4 4     151404.
```

As you can see, Colorado jail population has significantly reduced during COVID-19.

```
# facet version

colorado_dist <- colorado %>%
  group_by(county, qtr) %>%
  summarise(total = sum(total))

colorado_dist %>%
  ggplot(aes(x = total)) +
  geom_histogram(bins = 50) +
  facet_grid(qtr ~ .) +
  labs(title = "Distribution of Colorado Jail Size (by Population) in different
Quarters",
       y = "Count",
       x = "Jail Population")
```



The population per jail throughout Colorado is Unimodal, right-skewed distribution with significant outliers on the right of the graph.

```
colorado_dist %>%
  filter(total > 20000) %>%
  distinct(county)
```

```
## # A tibble: 4 x 1
## # Groups:   county [4]
##   county
##   <chr>
## 1 Adams
## 2 Arapahoe
## 3 El Paso
## 4 Weld
```

The largest jails in Colorado are the Adams, Arapahoe, and El Paso County Jails.

```
total_summary <- colorado_dist %>%
  summarise(mean_total = mean(total),
    sd_total = sd(total),
    median_total = median(total),
    IQR_total = IQR(total))
```

```
total_summary
```

```
## # A tibble: 53 x 5
##    county      mean_total sd_total median_total IQR_total
##   * <chr>           <dbl>    <dbl>        <dbl>     <dbl>
##  1 Adams           19966    3300.        20081     4386.
##  2 Alamosa           239      58.2         256       66
##  3 Arapahoe        19734.   3737.        20351     5064
##  4 Baca              195.     40.1         201.      44.7
##  5 Bent              560.    174.          512.     161
##  6 Boulder          8334.   1298.         8468.    1658.
##  7 Broomfield        509     121.          493      141
##  8 Chaffee           348.     68.5         381       34.8
##  9 Clear Creek      1134.    303.         1045.     274.
## 10 Conejos           177.     64.8         158       68.8
## # ... with 43 more rows
```

5. Linear Regression

Let's focus on only quarter 1 and 3 as before and after end points for COVID-19.

```
colorado_num_inmates <- colorado %>%
  filter(measure == "Number of inmates") %>%
  filter(qtr == 1 | qtr == 3)

colorado_num_inmates
```

```
## # A tibble: 100 x 24
##    qtr_year qtr   county jms    capacity  beds deaths bookings releases measure
##       <dbl> <fct> <chr>  <chr>    <dbl> <dbl>  <dbl>    <dbl>    <dbl> <chr>
##  1     2020 1     Clear~ E-Fo~      105   105      0      253      183 Number~
##  2     2020 1     Park   Jail~      255   200      0      199      191 Number~
##  3     2020 1     Eagle  Inte~      112   112      0      366      377 Number~
##  4     2020 1     El Pa~ Beac~     1837  1837      1     5161     5356 Number~
##  5     2020 1     Logan  New ~      120   120      0     1748     1731 Number~
##  6     2020 1     Baca   None        26    26      0       34       30 Number~
##  7     2020 1     San M~ Spil~       32    32      0       55       66 Number~
##  8     2020 1     Gunni~ Omni~       85    85      0      647      655 Number~
##  9     2020 1     Monte~ Efor~      104   104      0      517      495 Number~
## 10     2020 1     Pueblo Spil~      780   509      0     1901     1843 Number~
## # ... with 90 more rows, and 14 more variables: total <dbl>, male <dbl>,
## #   female <dbl>, other_gender <dbl>, black <dbl>, native_american <dbl>,
## #   other_race <dbl>, white <dbl>, unknown_race <dbl>, non_hispanic <dbl>,
## #   hispanic <dbl>, unknown_ethnicity <dbl>, not_available <chr>, isNA <lgl>
```

```
colorado_num_inmates %>%
  count(county) %>%
  filter(n == 1)
```

```
## # A tibble: 4 x 2
##   county          n
##   <chr>       <int>
## 1 Grand           1
## 2 Huerfano        1
## 3 Las Animas      1
## 4 Saguache       1
```

Remove 4 jails that don't have both first and third quarter: Grand, Huerfano, Las Animas, Saguache

```
colorado_num_inmates <- colorado_num_inmates %>%
  filter(county != "Grand" &
           county != "Huerfano" &
           county != "Las Animas" &
           county != "Saguache") %>%
  select(-c(not_available, isNA, jms, qtr_year, measure, deaths,
            other_gender,
            bookings, releases)) %>%
  mutate(other_race = unknown_race + other_race) %>%
  select(-c(unknown_race)) %>%
  arrange(county)

colorado_num_inmates
```

```
## # A tibble: 96 x 14
##    qtr   county capacity  beds total  male female black native_american
##    <fct> <chr>     <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl>           <dbl>
##  1 1     Adams      1271  1678   956   787    169   124               2
##  2 3     Adams      1271  1678   600   536     64    76               0
##  3 3     Alamo~      163    36    36    31      5     0               0
##  4 1     Alamo~      163    72    72    58     14     0               0
##  5 1     Arapa~     1174  1468  1115   948    167   354               4
##  6 3     Arapa~     1174  1468   628   574     53   197               5
##  7 1     Baca         26    26     7     6      1     2               0
##  8 3     Baca         26    26     6     6      0     1               0
##  9 1     Bent         62    58    47    37     10     2               0
## 10 3     Bent         96    96    29    25      4     2               0
## # ... with 86 more rows, and 5 more variables: other_race <dbl>, white <dbl>,
## #   non_hispanic <dbl>, hispanic <dbl>, unknown_ethnicity <dbl>
```

Pivot_wider;

```
colorado_num_inmates <- colorado_num_inmates %>%
  pivot_wider(names_from = qtr, values_from = capacity:unknown_ethnicity)
```

```
colorado_num_inmates <- colorado_num_inmates %>%
  inner_join(pop, by = "county") %>%
  inner_join(demo, by = "county")
```

```
colorado_num_inmates <- colorado_num_inmates %>%
  mutate(difference = (total_3 - total_1)/total_1)

colorado_num_inmates
```

```
## # A tibble: 48 x 32
##    county capacity_1 capacity_3 beds_1 beds_3 total_1 total_3 male_1 male_3
##    <chr>       <dbl>      <dbl>  <dbl>  <dbl>   <dbl>   <dbl>  <dbl>  <dbl>
##  1 Adams        1271       1271   1678   1678     956     600    787    536
##  2 Alamo~        163        163     72     36      72      36     58     31
##  3 Arapa~       1174       1174   1468   1468    1115     628    948    574
##  4 Baca           26         26     26     26       7       6      6      6
##  5 Bent           62         96     58     96      47      29     37     25
##  6 Bould~        519        543    543    543     400     223    348    197
##  7 Broom~        218         218     218     218      122       55     100      44
##  8 Chaff~        105         105     105     105       66       30      52      22
##  9 Clear~        105         105     105     105       70       57      62      48
## 10 Conej~         82          86      82      86       65       11      47       9
## # ... with 38 more rows, and 23 more variables: female_1 <dbl>, female_3 <dbl>,
## #   black_1 <dbl>, black_3 <dbl>, native_american_1 <dbl>,
## #   native_american_3 <dbl>, other_race_1 <dbl>, other_race_3 <dbl>,
## #   white_1 <dbl>, white_3 <dbl>, non_hispanic_1 <dbl>, non_hispanic_3 <dbl>,
## #   hispanic_1 <dbl>, hispanic_3 <dbl>, unknown_ethnicity_1 <dbl>,
## #   unknown_ethnicity_3 <dbl>, population <dbl>, lesscollege_whites_pct <dbl>,
## #   black_pct <dbl>, rural_pct <dbl>, urbanicity <fct>, liberal <fct>,
## #   difference <dbl>
```

```
colorado_num_percent <- colorado_num_inmates %>%
  mutate(jail_male_pct = male_1 / total_1) %>%
  mutate(jail_black_pct = black_1 / total_1) %>%
  mutate(jail_hispanic_pct = hispanic_1 / total_1) %>%
  select(county, difference, lesscollege_whites_pct, jail_male_pct, jail_black_pct, jail_hispanic_pct,
         population, rural_pct, black_pct, liberal, urbanicity)

colorado_num_percent
```

```
## # A tibble: 48 x 11
##    county difference lesscollege_whi~ jail_male_pct jail_black_pct
##    <chr>       <dbl>            <dbl>         <dbl>          <dbl>
##  1 Adams      -0.372             70.2         0.823         0.130
##  2 Alamo~     -0.5               66.0         0.806         0
##  3 Arapa~     -0.437             52.7         0.850         0.317
##  4 Baca       -0.143             76.4         0.857         0.286
##  5 Bent       -0.383             86.5         0.787         0.0426
##  6 Bould~     -0.442             35.9         0.87          0.0775
##  7 Broom~     -0.549             44.5         0.820         0.0492
##  8 Chaff~     -0.545             62.1         0.788         0.0303
##  9 Clear~     -0.186             54.2         0.886         0.143
## 10 Conej~     -0.831             76.7         0.723         0.0154
## # ... with 38 more rows, and 6 more variables: jail_hispanic_pct <dbl>,
## #   population <dbl>, rural_pct <dbl>, black_pct <dbl>, liberal <fct>,
## #   urbanicity <fct>
```

new model

```
full <- lm(difference ~ liberal +
     lesscollege_whites_pct +
     population +
     jail_male_pct +
     jail_black_pct +
     jail_hispanic_pct+
     urbanicity +
     black_pct,
   data = colorado_num_percent)

full %>%tidy
```

```
## # A tibble: 10 x 5
##    term                       estimate  std.error statistic p.value
##    <chr>                         <dbl>      <dbl>     <dbl>   <dbl>
##  1 (Intercept)             4.05        1.80          2.25   0.0302
##  2 liberal1                0.00774     0.500         0.0155 0.988
##  3 lesscollege_whites_pct  0.00702     0.0179        0.393  0.697
##  4 population              0.000000388 0.00000126    0.307  0.761
##  5 jail_male_pct          -4.34        1.37         -3.17   0.00302
##  6 jail_black_pct         -1.21        2.58         -0.469  0.642
##  7 jail_hispanic_pct      -1.07        0.794        -1.35   0.184
##  8 urbanicityurban        -0.964       0.432        -2.23   0.0314
##  9 urbanicitymetro        -0.923       0.544        -1.70   0.0979
## 10 black_pct              -0.0239      0.0781       -0.306  0.761
```

```
int_only_model <- lm(difference ~ 1, data = colorado_num_percent)

covid_model <- step(full, scope = formula(int_only_model), direction = "backward")
```

```
## Start:  AIC=10.93
## difference ~ liberal + lesscollege_whites_pct + population +
##     jail_male_pct + jail_black_pct + jail_hispanic_pct + urbanicity +
##     black_pct
##
##                           Df Sum of Sq    RSS     AIC
## - liberal                  1    0.0003 39.732  8.9261
## - black_pct                1    0.0980 39.830  9.0440
## - population               1    0.0985 39.830  9.0446
## - lesscollege_whites_pct   1    0.1613 39.893  9.1202
## - jail_black_pct           1    0.2300 39.962  9.2028
## <none>                                39.732 10.9258
## - jail_hispanic_pct        1    1.9133 41.645 11.1833
## - urbanicity               2    5.4682 45.200 13.1152
## - jail_male_pct            1   10.5011 50.233 20.1826
##
## Step:  AIC=8.93
## difference ~ lesscollege_whites_pct + population + jail_male_pct +
##     jail_black_pct + jail_hispanic_pct + urbanicity + black_pct
##
##                           Df Sum of Sq    RSS     AIC
## - population               1    0.0983 39.830  7.0447
## - black_pct                1    0.0989 39.831  7.0455
## - jail_black_pct           1    0.2343 39.966  7.2083
## - lesscollege_whites_pct   1    0.3410 40.073  7.3363
## <none>                                39.732  8.9261
## - jail_hispanic_pct        1    1.9206 41.653  9.1920
## - urbanicity               2    5.4688 45.201 11.1160
## - jail_male_pct            1   10.5355 50.268 18.2158
##
## Step:  AIC=7.04
## difference ~ lesscollege_whites_pct + jail_male_pct + jail_black_pct +
##     jail_hispanic_pct + urbanicity + black_pct
##
##                           Df Sum of Sq    RSS     AIC
## - black_pct                1    0.0351 39.866  5.0870
## - jail_black_pct           1    0.1745 40.005  5.2545
## - lesscollege_whites_pct   1    0.2688 40.099  5.3676
## <none>                                39.830  7.0447
## - jail_hispanic_pct       1    1.8812 41.712  7.2598
## - urbanicity              2    5.4294 45.260  9.1786
## - jail_male_pct           1   11.2241 51.055 16.9613
##
## Step:  AIC=5.09
## difference ~ lesscollege_whites_pct + jail_male_pct + jail_black_pct +
##     jail_hispanic_pct + urbanicity
##
##                          Df Sum of Sq    RSS     AIC
## - lesscollege_whites_pct  1    0.2421 40.108  3.3776
## - jail_black_pct          1    0.3179 40.183  3.4682
## <none>                                39.866  5.0870
## - jail_hispanic_pct       1    1.9015 41.767  5.3235
## - urbanicity              2    5.4886 45.354  7.2785
## - jail_male_pct           1   11.4549 51.320 15.2106
##
## Step:  AIC=3.38
## difference ~ jail_male_pct + jail_black_pct + jail_hispanic_pct +
##     urbanicity
##
##                     Df Sum of Sq    RSS     AIC
## - jail_black_pct     1    0.4212 40.529  1.8791
## <none>                           40.108  3.3776
## - jail_hispanic_pct  1    1.7640 41.872  3.4437
## - urbanicity         2    6.8278 46.935  6.9235
## - jail_male_pct      1   11.2964 51.404 13.2888
##
## Step:  AIC=1.88
## difference ~ jail_male_pct + jail_hispanic_pct + urbanicity
##
##                     Df Sum of Sq    RSS     AIC
## - jail_hispanic_pct  1    1.6725 42.201  1.8202
## <none>                           40.529  1.8791
## - urbanicity         2    6.7288 47.258  5.2519
## - jail_male_pct      1   11.7954 52.324 12.1405
##
## Step:  AIC=1.82
## difference ~ jail_male_pct + urbanicity
##
##                 Df Sum of Sq    RSS     AIC
## <none>                       42.201  1.8202
## - urbanicity     2    6.6391 48.841  4.8333
## - jail_male_pct  1   14.3411 56.543 13.8621
```

Based on backwards AIC selection, the two significant predictors for are the percent of males in a jail and whether a jail is in a rural, urban, or metropolitan area.

Interaction Term

```
reduced_model <- covid_model
full_model <- lm(difference ~
     jail_male_pct +
     urbanicity +
     jail_male_pct * urbanicity,
   data = colorado_num_percent)

anova(reduced_model, full_model) %>%
  tidy() %>%
  kable(digits = 3)
```

res.df	rss	df	sumsq	statistic	p.value
44	42.201	NA	NA	NA	NA
42	19.223	2	22.978	25.102	0

Since F-statistic is high and p-value is close to 0, the interaction effect between jail_male_pct * ruralurban_cc exists.

Model and Interpretations:

Model:

```
full_model %>%
  tidy(conf.int = TRUE) %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	12.347	1.334	9.254	0	9.654	15.039
jail_male_pct	-14.067	1.566	-8.984	0	-17.227	-10.907
urbanicityurban	-11.840	1.602	-7.392	0	-15.073	-8.608
urbanicitymetro	-13.398	3.158	-4.242	0	-19.771	-7.025
jail_male_pct:urbanicityurban	13.107	1.906	6.878	0	9.261	16.953
jail_male_pct:urbanicitymetro	14.841	3.749	3.959	0	7.276	22.407

Urbanicity:

Rural jails have failed to slow to decreasing their jail population during COVID-19 in comparison to urban/metropolitan areas:

- A jail in an urban area is expected to decrease its population 12 percent more than a jail in a rural area, on average.
- A jail in an metropolitan area is expected to decrease its population 13 percent more than a jail in a rural area, on average.

Male Population Percentage and its Interaction with Urbanicity:

Jails in rural and urban areas with a higher male population have a higher chance to decreasing their jail population. Jails with higher male populations in metropolitan areas have a higher chance of increasing their jail population. Specifically,

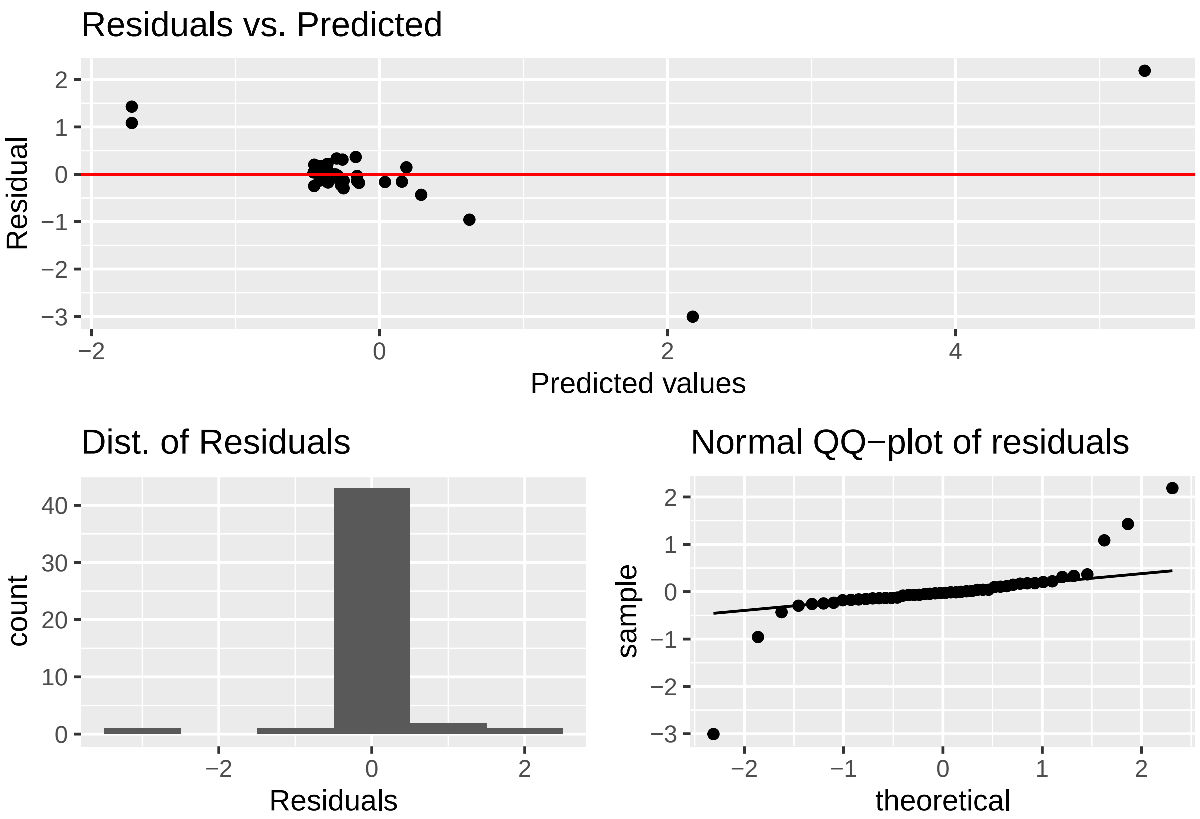
- For rural jails, for every one percent increase in male inmates, there is expected to be a 14 percent decrease in jail population between Jan to Sept 2020, on average.
- For urban jails, for every one percent increase in male inmates, there is expected to be a 1 percent decrease in jail population between Jan to Sept 2020, on average.

- For metropolitan jails, for every one percent increase in male inmates, there is expected to be a 1 percent increase in jail population between Jan to Sept 2020, on average.

Model Conditions

Check Conditions

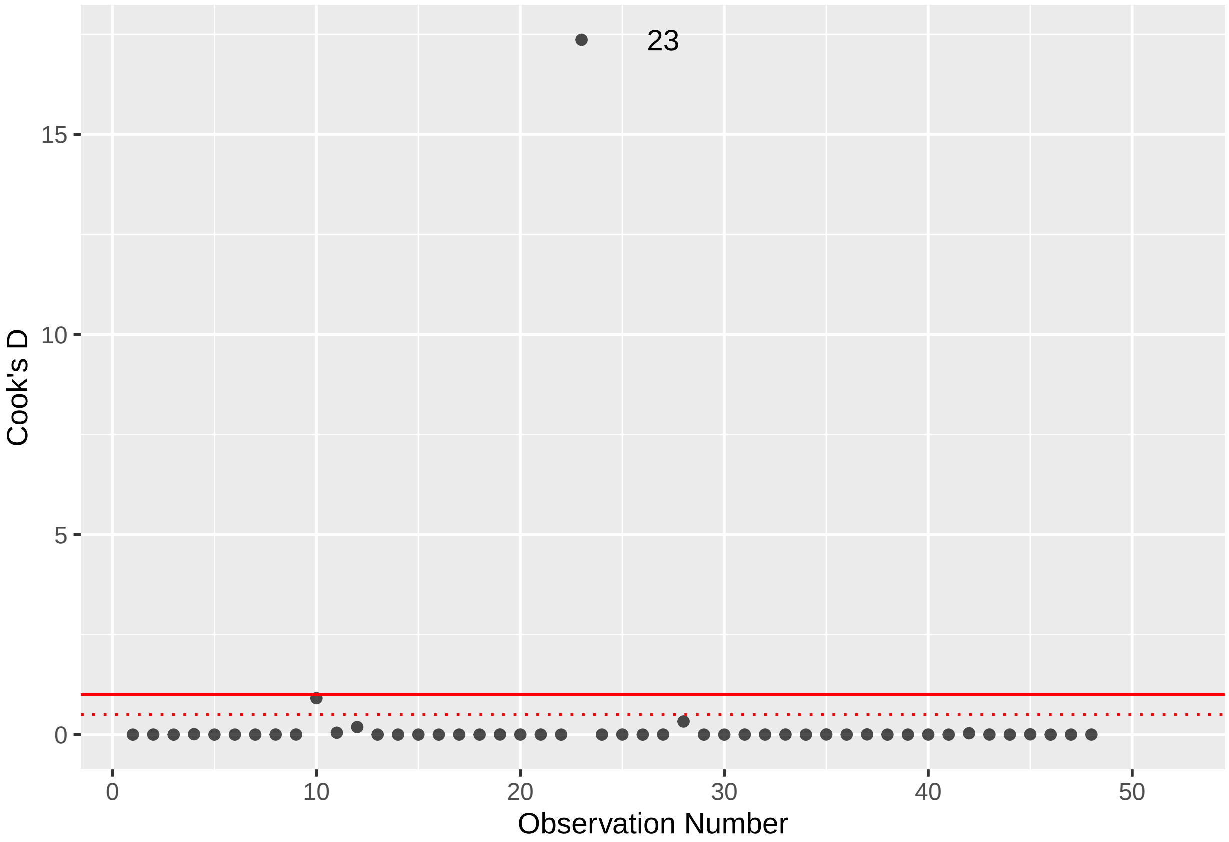
```r
model_aug <- augment(full_model) %>%
   mutate(obs_num = row_number()) #add row number to help with graphing

resid_fitted <- ggplot(data = model_aug, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red") +
  labs(x = "Predicted values",
    y = "Residual",
    title = "Residuals vs. Predicted")

resid_hist <- ggplot(data = model_aug, aes(x = .resid)) +
  geom_histogram(binwidth = 1) +
  labs(x = "Residuals", title = "Dist. of Residuals")

resid_qq <- ggplot(data = model_aug, aes(sample = .resid)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Normal QQ-plot of residuals")

conditions_plot <- resid_fitted / (resid_hist + resid_qq)

conditions_plot
```



Model Diagnostics

Cook's distance

```
#scatterplot of cook's d vs obs num
ggplot(data = model_aug, aes(x = obs_num, y = .cooksd)) +
  geom_point(alpha = 0.7) +
  geom_hline(yintercept = 1, color = "red") +
  geom_hline(linetype = "dotted", yintercept = 0.5, color = "red") +
  labs(x = "Observation Number", y = "Cook's D") +
  geom_text(aes(label=ifelse(.cooksd > 1,
                             as.character(obs_num), "")), nudge_x = 4)
```



Jackson County (Observation 23), which has a super small county jail, is a high leverage county. This is because it increased from having 2 people to 17 people in its jail over COVID-19. It is an influential point, meaning that it has a large impact on the coefficients and standard errors used for inference.

Because the goal of the model is explanation as opposed to prediction, it is worth keeping this point in the model.