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)
```

Table of contents:

Background

Motivation

The Colorado Jail Database, instituted by Colorado House Bill 19-1297, is currently the most comprehensive, open source published, and catalogued state-wide jail database in the United States. HB19-1297 was a bipartisan bill sponsored by State Representatives Michael Weissman (D) and Hugh McKean (R). Supported by ACLU-Colorado, HB19-1297 was signed May of 2019. HB19-1297 expanded the information that jail keepers must submit to the Division of Criminal Justice, mandating that unless data is submitted quarterly, jail keepers could be subject to a $500 fine.

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's HB 19-1297 Jail Data Collection legislation can be an effective model that other states can look towards to adopt a similar system of data collection.

Our goal is to focus on COVID-19 outcomes for jail population numbers. Public officials in Colorado have attempted to reduce jail populations at the beginning of the outbreak. 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. Jails that initially had expansion plans have halted their plans due to general COVID-19 decreases in jail populations. Overall, it is clear that Colorado has had a reported net decrease in jail population during the COVID-19 crisis, albeit slow and unsatisfactory.

Research Question

What demographic or jail information is correlated with a county jail decreasing its jail population at the outbreak of and during the COVID-19 pandemic?

Data Comments

HB 19-1297:

The public and published HB 19-1297 dataset represents 3 collection periods in 2020, split up into yearly quarters. Q1 was Jan-Mar, Q2 was Apr-Jun, Q3 was Jul-Aug.

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.

Source: https://ors.colorado.gov/ors-coll-jaildata

MIT Election Lab

The data file election-context-2018.csv contains demographic and past election data at the county level.

Source: https://github.com/MEDSL/2018-elections-unoffical/blob/master/election-context-2018.md

Data Preparation

Data Loading

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

#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')) %>%
  rename(population = total_population) %>%
  select(county, lesscollege_whites_pct, black_pct, rural_pct, urbanicity, liberal,
         population)
```

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
```

Most of the data that is missing is that 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.

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.

Fix Type

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

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
```

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

Merging Datasets

Let's focus on only quarter 1 and 3 as before and after end points for COVID-19. This will allow us to adjudicate the full response to the ongoing COVID-19 virus.

```
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>
```

Remove 4 jails that don't have both first and third quarter: Grand, Huerfano, Las Animas, Saguache. Remove unneeded variables. Merge unknown and other race.

```
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)
```

Pivot the dataset wider to treat each jail as its own observation, with the ability to create new columns representing the changes in population counts.

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

Join the Colorado HB 19-1297, population, and MIT Election Lab Aggregated Demographic datasets.

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

Create new dataset called colorado_num_percent, which transforms all the variables into percent changes in order to standardize the numbers across different sizes of jails.

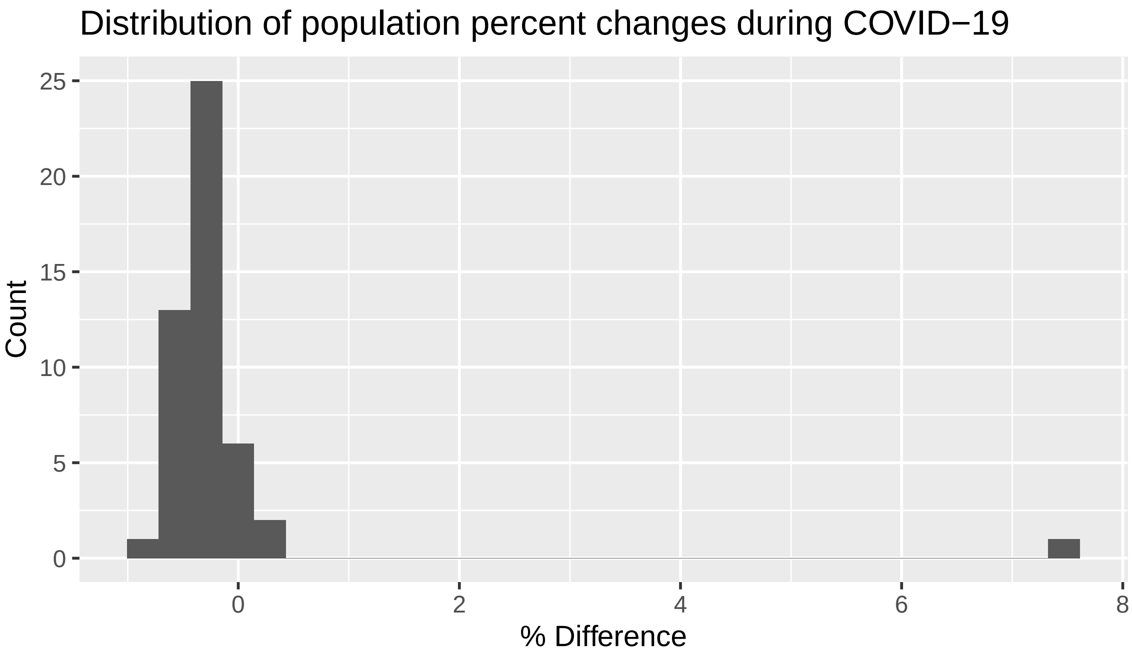
```
colorado_num_percent <- colorado_num_inmates %>%
  mutate(difference = (total_3 - total_1)/total_1) %>%
  mutate(jail_male_pct = male_1 / total_1) %>%
  mutate(jail_black_pct = black_1 / total_1) %>%
  mutate(jail_hispanic_pct = hispanic_1 / total_1) %>%
  select(county, difference, lesscollege_whites_pct, 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>
```

Exploratory Data Analysis

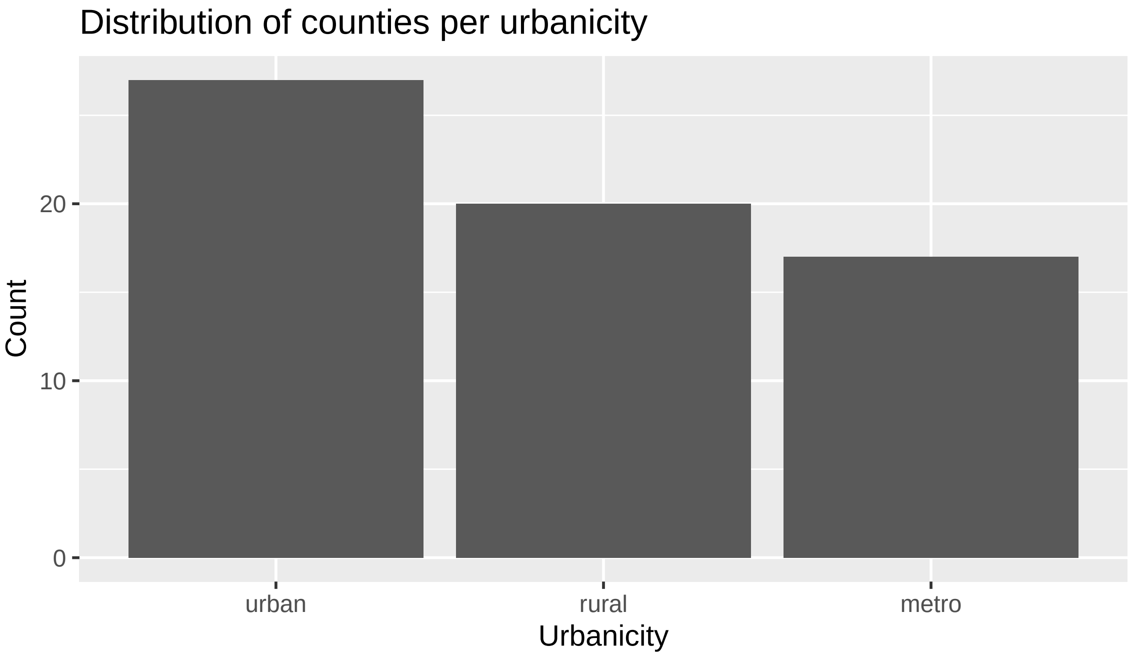
General Summary Statistics:

```
colorado_num_percent %>%
  ggplot(aes(x = difference)) +
  geom_histogram(bins = 30) +
  labs(title = "Distribution of population percent changes during COVID-19",
       x = "% Difference",
       y = "Count"
         )
```

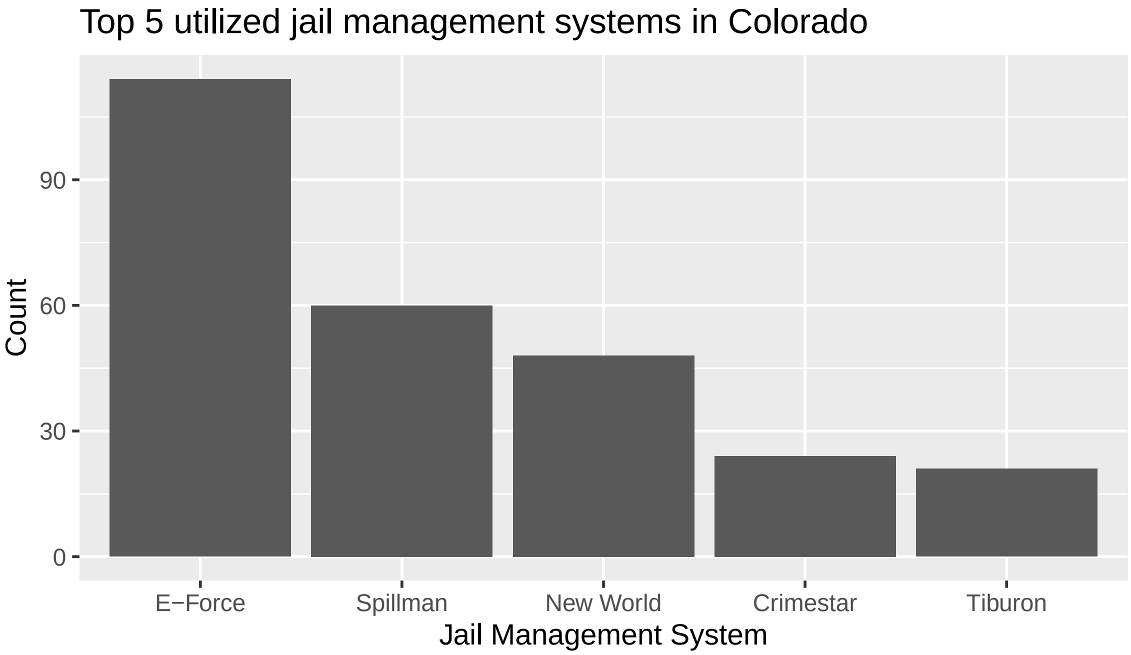


The percent change in population numbers has a unimodal normal distribution. It has a slight right-skew with one significant outlier at +7.5% difference.

```
demo %>%
  mutate(urbanicity = fct_relevel(urbanicity, "urban")) %>%
  ggplot(aes(urbanicity)) +
  geom_bar() +
  labs(title = "Distribution of counties per urbanicity",
       y = "Count",
       x = "Urbanicity")
```

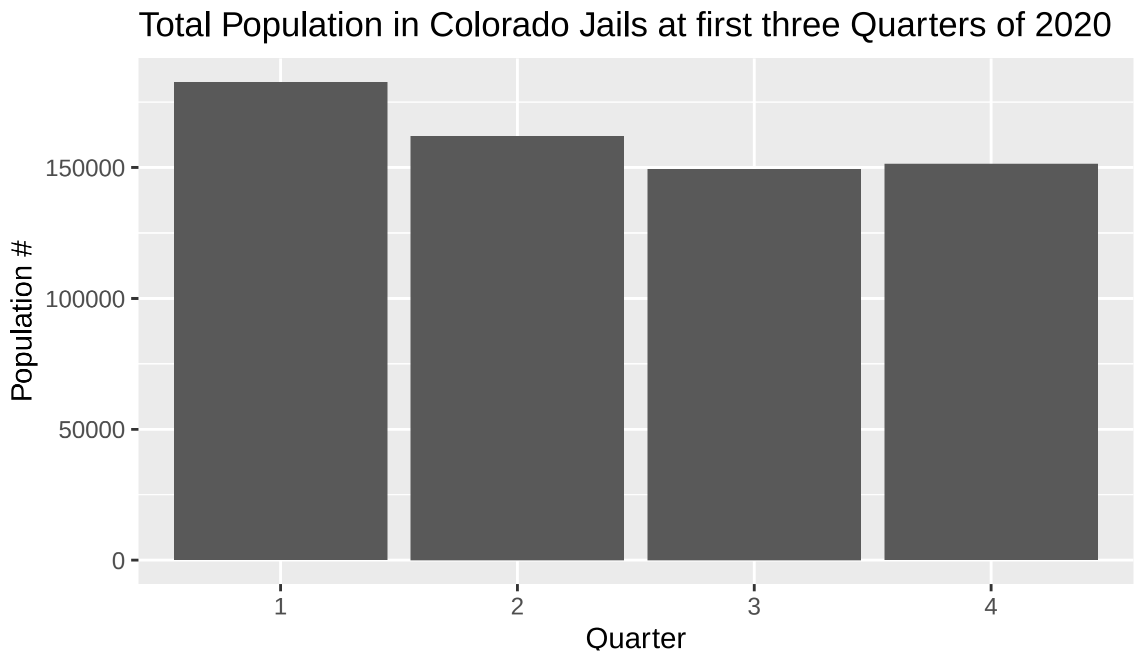


```
colorado %>%
  mutate(jms = str_replace(jms, "Eforce", "E-Force")) %>%
  mutate(jms = str_replace(jms, "eForce", "E-Force")) %>%
  count(jms) %>%
  arrange(-n) %>%
  mutate(n = n / 5) %>%
  slice(1:5) %>%
  ggplot(aes(x = reorder(jms, -n), y = n)) +
  geom_bar(stat = "identity") +
  labs(x = "Jail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



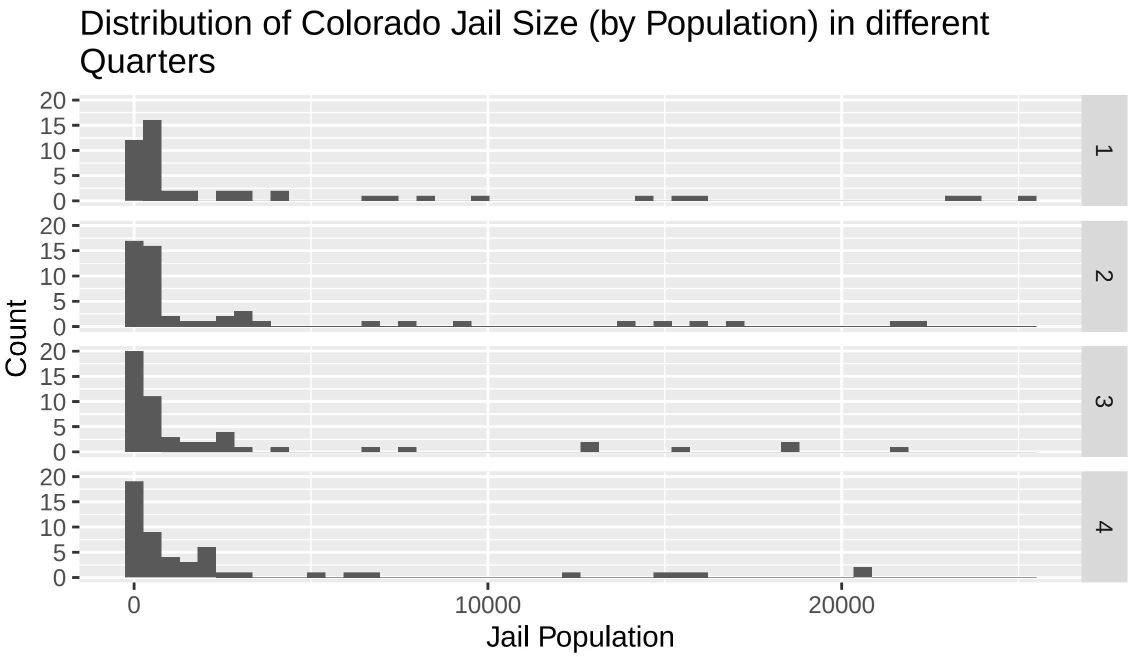
The top utilized jail management system in Colorado are E-Force, Spillman, New World, Crimestar, and Tiburon.

```
colorado %>%
  arrange(qtr) %>%
  group_by(qtr) %>%
  summarise(total = sum(total)) %>%
  ggplot(aes(x = qtr, y = total)) +
  geom_bar(stat = "identity") +
  labs(title = "Total Population in Colorado Jails at first three Quarters of 2020",
       x = "Quarter",
       y = "Population #")
```



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

```r
# 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.

Model

Model Selection

Variables:

`liberal` (Factor): 1 if county majority voted for Clinton in 2016, 0 if county majority voted for Trump

`lesscollege_whites_pct` (Double): white population with an education of less than a bachelor's degree as a percentage of total population

`population` (Double): Total population of a county

`jail_male_pct` (Double): Percent male of a jail

`jail_black_pct` (Double): Percent Black of a jail

`jail_hispanic_pct` (Double): Percent Hispanic of a jail

`urbanicity` (Factor): metro - Counties in Metropolitan Areas. urban - Counties in Urban Areas. rural - Counties in Rural Areas.

Definitions of Metropolitan, Urban, and Rural are designated by the US Office of Management and Budget (OMB) delineation as of February 2013 black_pct' (Double): Black population as a percentage of total population

Initial Model Fitting

First, put as many predictor variables as possible to create a full 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)
```

Backwards selection with AIC as selection criterion to choose the model with the best AIC.

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

Now, let's use nested F-test to check if the addition of an interaction term between the two variables above is statistically significant:

```
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 is statistically significant.

Model Interpretations:

```
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.

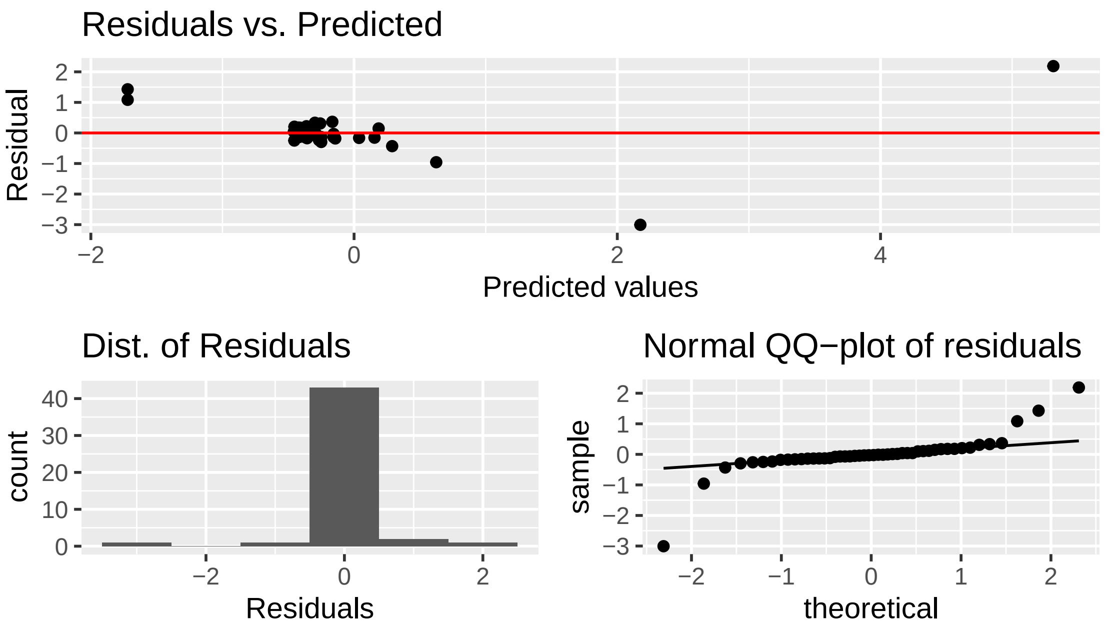
Male Population Percentage and its Interaction with Urbanicity:

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:

```
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
```



4 Conditions for Linear Regression:

1. Linearity: Partially Satisfied - Although there is no distinguishable pattern in the residuals vs. predicted plot, the points are generally scattered. There seem to be some high leverage points that are worth investigating. However, lee-way will be given given the small number of observations. Overall, the residuals are randomly scattered.

2. Normality: Satisfied - the points genearlly fall along a straight diagonal line on the normal quantile plot.

3. Constant Variance: Satisfied - the vertical spread of the residuals remains relatively constant across the plot.

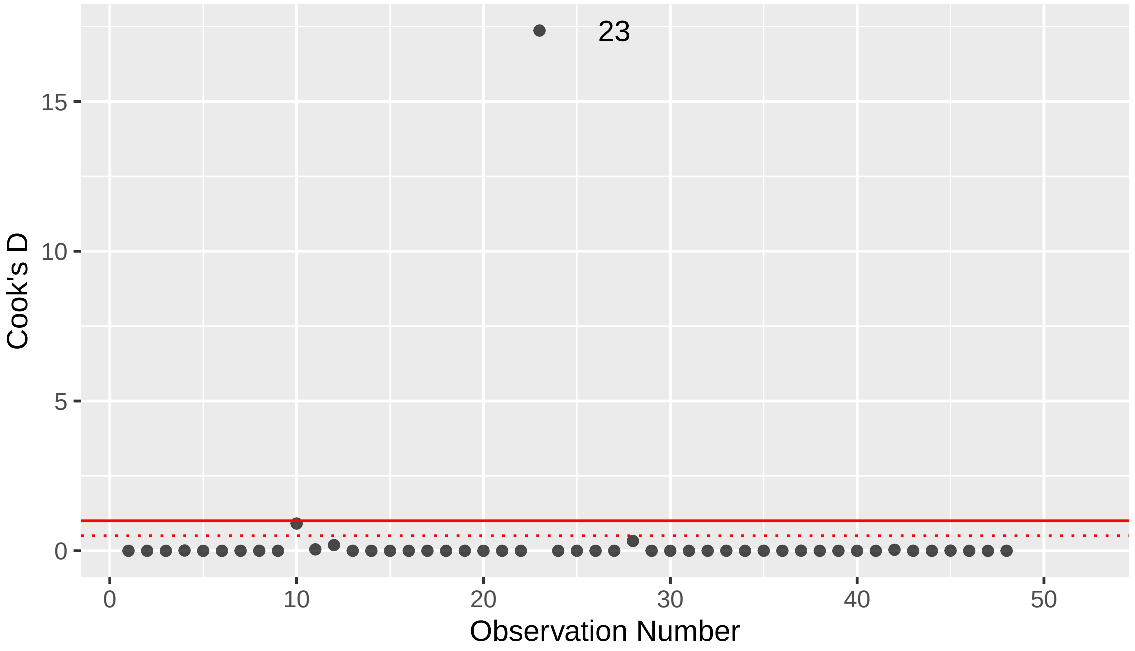
4. Independence: Satisfied - The error for one county does not tell us anything about the error for another county. We also put urbanicity of the counties as one of the variables, which avoids some of the problems of spatial collinarity/correlation. However, in order to improve this model, we can add a "region" variable to account for the locations of the counties in the state of Colorado.

Model Diagnostics

Cook's distance

Let's investigate those possible high-leverage points:

```
#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.