Colorado: Exploratory Data Analysis and Linear Regression

Albert Sun

9/30/2020

The Dataset

Open-source: 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 California and Texas don't have their data published completely open source; instead, researchers are forced to scrape the data periodically from the state website. As such, we find that the HB-1297 data set to be the most friendly for public use thus far.

Collection periods: This dataset represents 3 collection periods in 2020, split up into quarters. Q1 was in January, Q2 was in April, Q3 was in July. A frequent periodic collection of data, like quarterly, lends the dataset to really good data analysis, because we can better see the effects of policies or huge events (hint hint COVID-19) over time.

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

Rows: 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.

```
## Rows: 2,280
## 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...
```

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.

```
## # A tibble: 10 x 2
##    not_available                                                          n
##    <chr>                                                              <int>
##  1  <NA>                                                               1792
##  2 "Eforce does not seprate this data"                                   18
##  3 "JMS does not calculate this"                                         18
##  4 "ESTIMATES"                                                           13
##  5 "This is not tracked in our system"                                   13
##  6 "My JMS doesn't break unsentenced inmates by Gender or Race/Ethnicity."    12
##  7 "Not able to capture information"                                     11
##  8 "JMS does not currently break down \"sentenced\" by gender/race/ethnic~    10
##  9 "Not able to caputre information"                                     10
## 10 "Population down due to COVID-19"                                     10
```

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:

```
## Rows: 2,280
## 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.

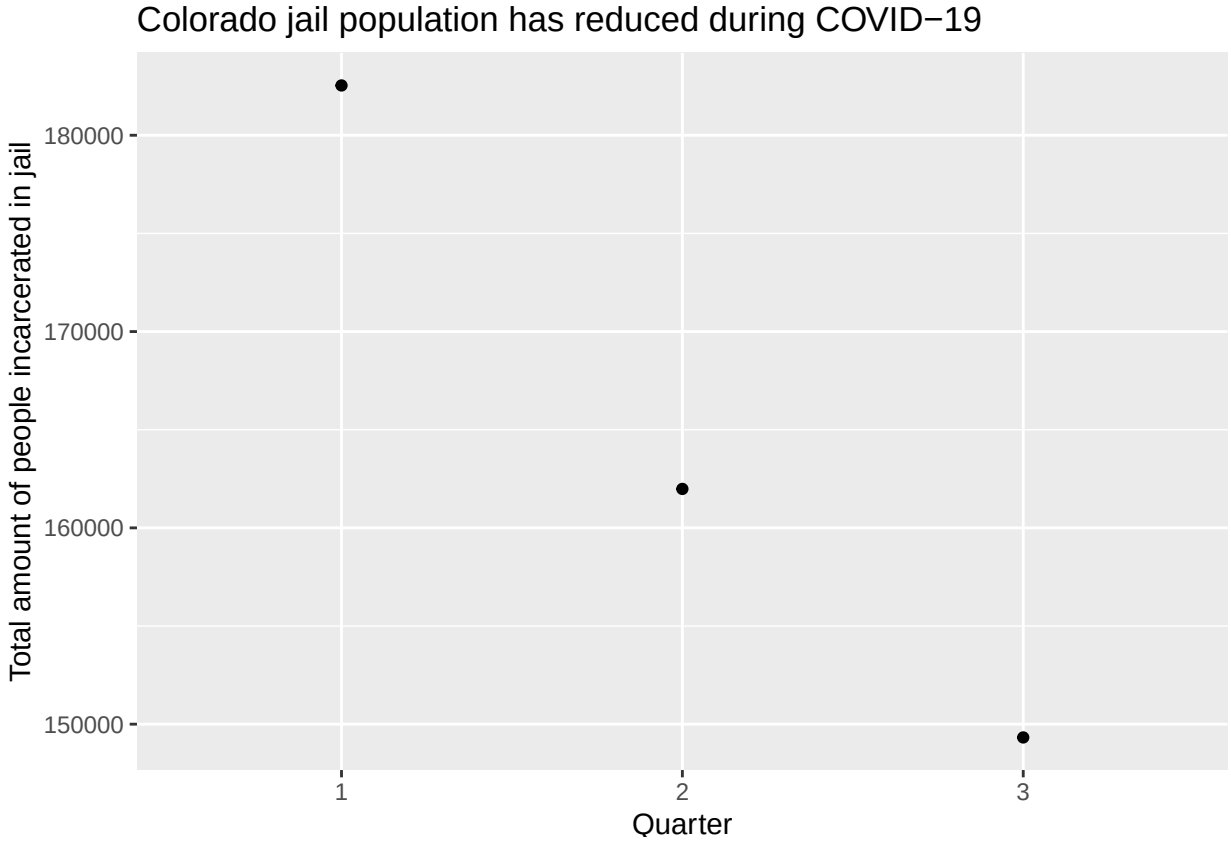
Overall, the other variables seem to have the correct data type.

3. Investigate Missingness

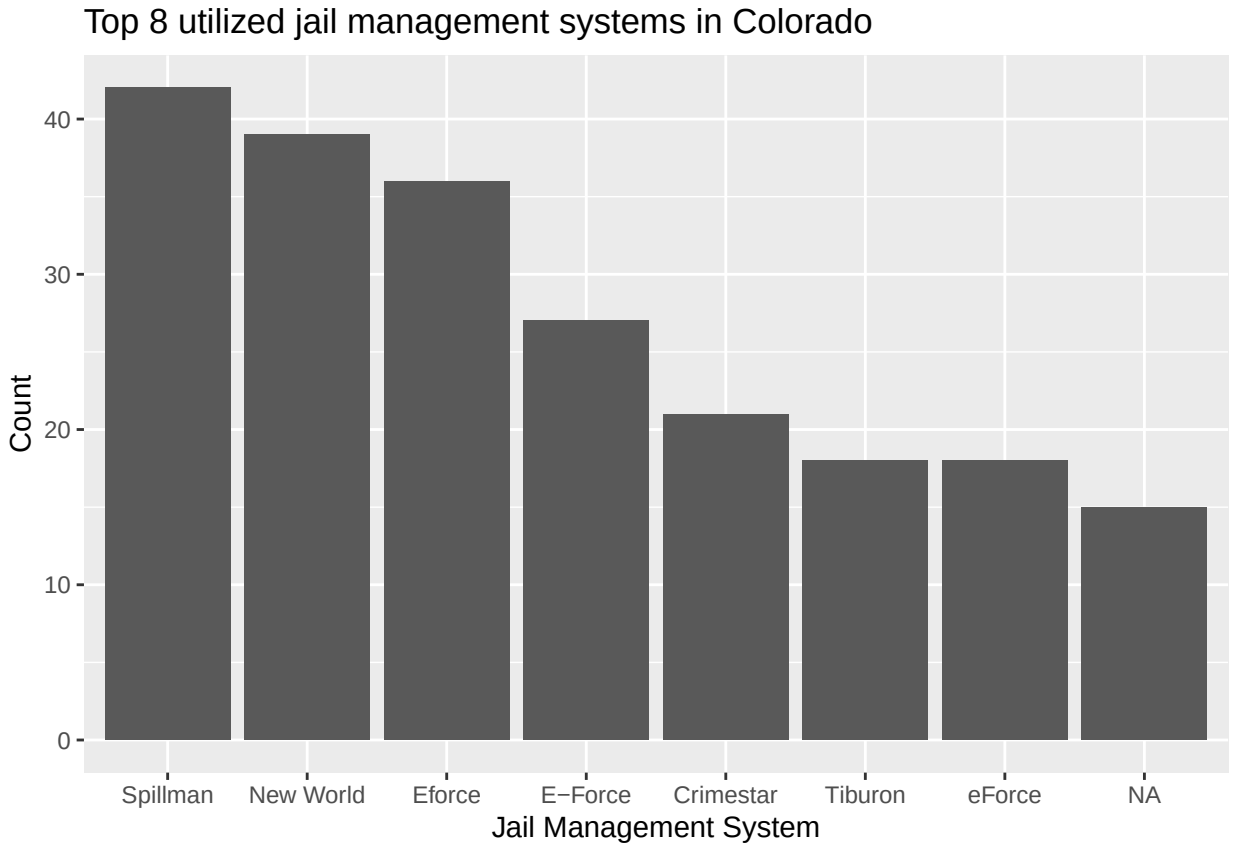
```
## # A tibble: 3 x 3
##   qtr   isNA      n
##   <fct> <lgl> <int>
## 1 1     TRUE    170
## 2 2     TRUE    162
## 3 3     TRUE    156
```

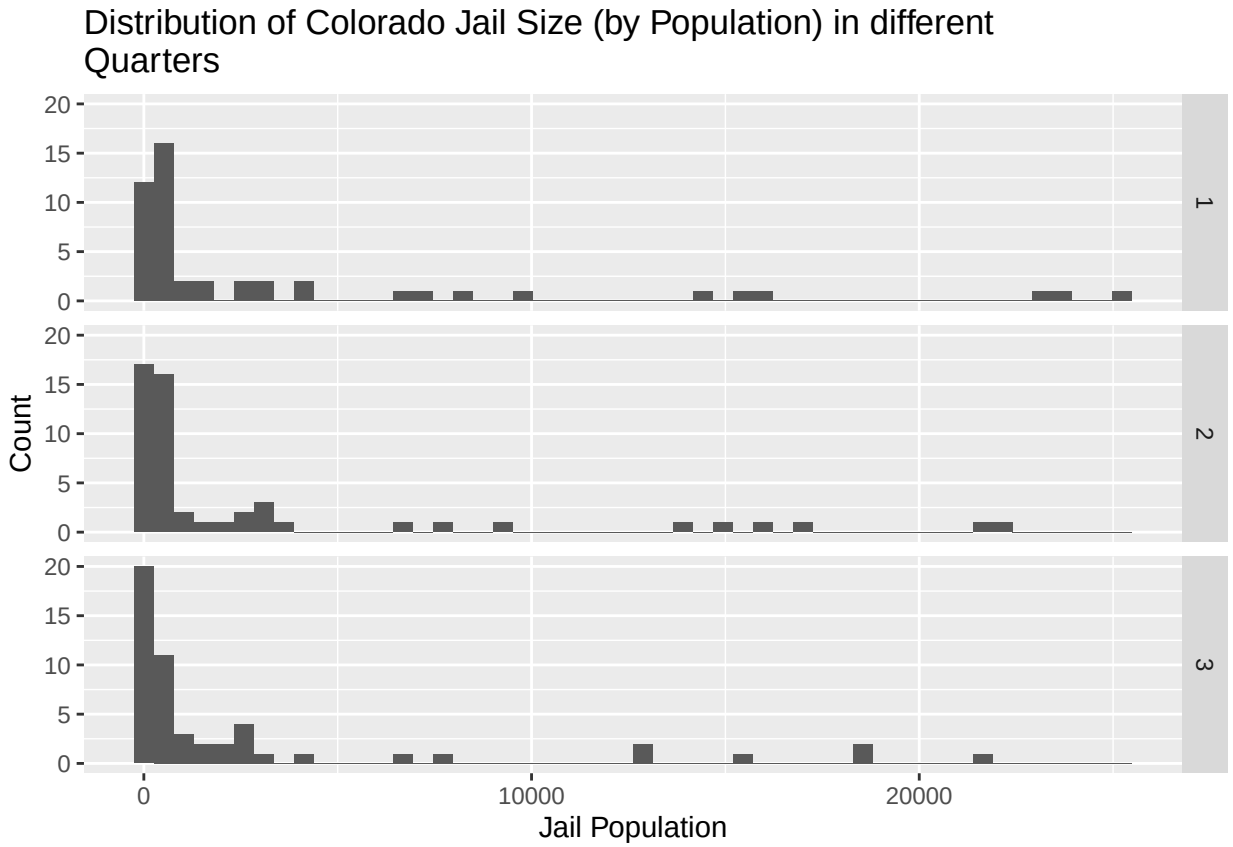
Missingness was generally reduced throughout the three quarters of jail data collection in 2020, possibly suggesting improvements in jail collection throughout this time period.

4. EDA



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


Top 8 utilized jail management systems in Colorado
Count
40
30
20
10
0
Spillman
New World
Eforce
E−Force
Crimestar
Tiburon
eForce
NA
Jail Management System


Distribution of Colorado Jail Size (by Population) in different Quarters
Count
20
15
10
5
0
1
2
3
0
10000
20000
Jail Population

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

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

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

```
## # A tibble: 53 x 5
##    county      mean_total sd_total median_total IQR_total
##    <chr>            <dbl>    <dbl>        <dbl>     <dbl>
##  1 Adams            21225   2613.         21784     2568.
##  2 Alamosa            241     71.1          279       63
##  3 Arapahoe         21274   2593.         22393     2406.
##  4 Baca               212.    23.7          216.      23.5
##  5 Bent               590.   200.           554      198.
##  6 Boulder           8862.   924.          9089.     903.
##  7 Broomfield         550.   109.           541      109
##  8 Chaffee            382      1.73         381        1.5
##  9 Clear Creek       1219.   308.          1114      294.
## 10 Conejos            194.    66.7          180       65.5
## # ... with 43 more rows
```

5. Linear Regression

In statistics, linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables).

One question we are trying to ask is to see which predictors are significant to whether a jail would on net release people during COVID-19. Let's fit a linear model about that.

```
## # A tibble: 6 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     Pueblo Spil~      780   509      0     1901     1843 Number~
## 2     2020 1     Pitkin SPIL~       30    32      1      173        0 Number~
## 3     2020 2     Pueblo Spil~      780   509      0     1889     2143 Number~
## 4     2020 2     Pitkin SPIL~       30    32      0       99       98 Number~
## 5     2020 3     Pitkin spil~       30    32      0       43       37 Number~
## 6     2020 3     Pueblo Spil~      780   509      0     1157     1213 Number~
## # ... with 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>
```

Since only two jails (Pitkin and Pueblo) don't offer information on ethnicity or race for the measure of number of total inmates in only one of their quarters, this data, focused on ethnicity counts per county, is mostly complete for modelling. We'll remove these numbers

Let's focus on only quarter 1 and 3:

```
## # A tibble: 4 x 2
```

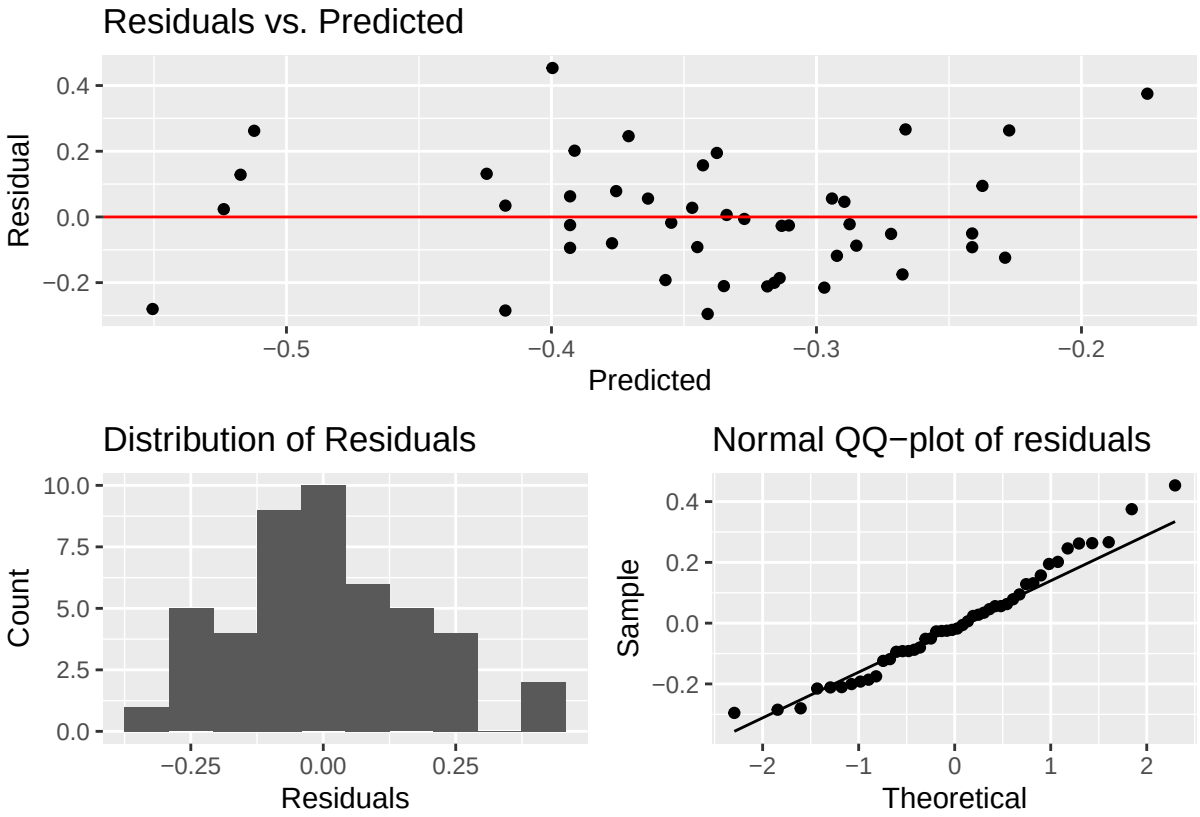
```
##   county         n
##   <chr>      <int>
## 1 Grand          1
## 2 Huerfano       1
## 3 Las Animas     1
## 4 Saguache       1
```

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

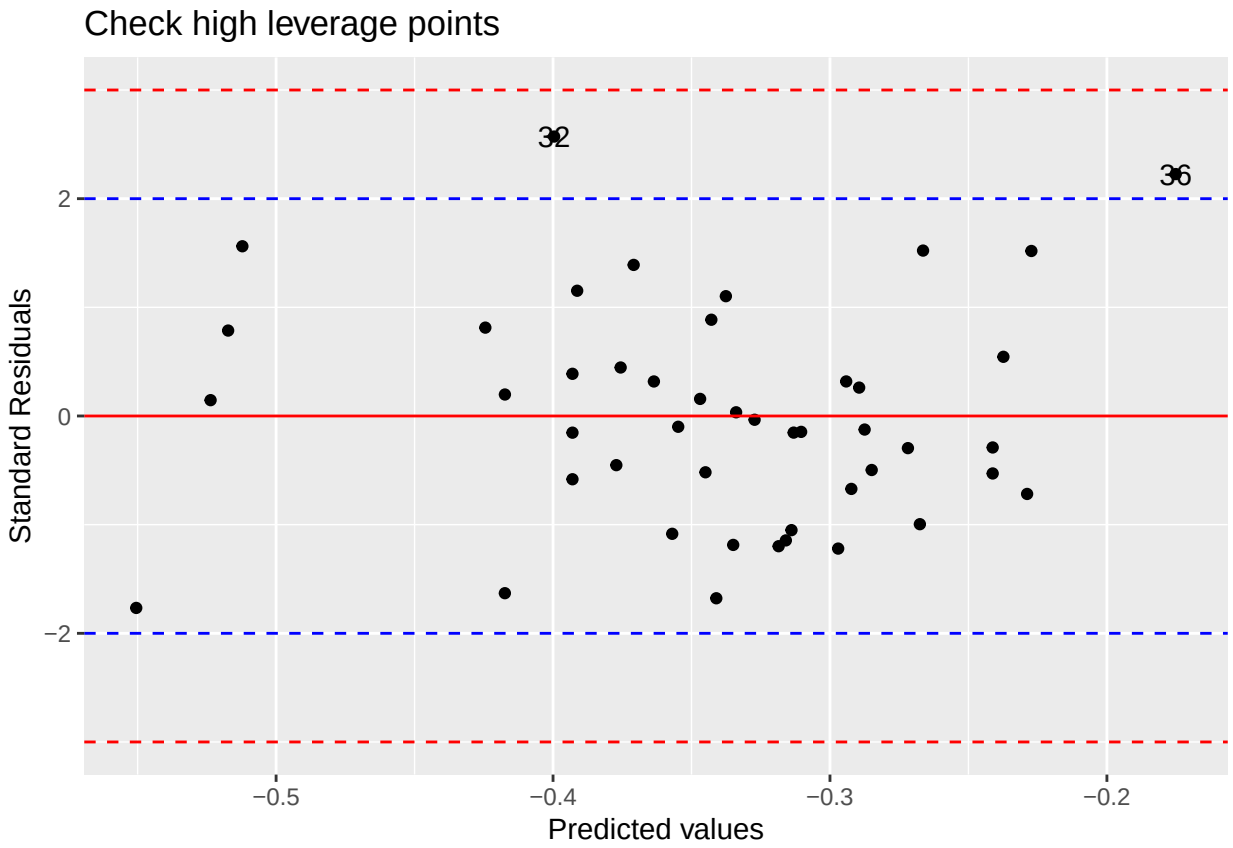
Pivot_wider:

Standardize total, race, gender, ethnicity to percentages:

Based on backwards AIC selection, we will use the model with white_percent, hispanic_percent.



Linearity, Constant Variance, Normality, and Independence are satisfied.



Removed Jackson and Routt earlier b/c super small population totals/outliers resulted in skewed numbers and skewed model.

term	estimate	std.error	statistic	p.value
(Intercept)	-0.393	0.078	-5.061	0.000
white_percent	0.218	0.107	2.032	0.048
hispanic_percent	-0.456	0.156	-2.920	0.006

$$percentchange - hat = 0.218 * whitepercent - 0.456 * hispanicpercent - 0.393$$

Interpretation:

1. There is a positive relationship between the % of hispanic people in a jail and the jail increasing its net population. All else held constant, with every percentage increase in the amount of hispanic people in a jail, we expect there to be a 0.5% decrease in jail population from 2020 Q1 to Q3.
2. There is a positive relationship between the percentage of white people in a jail and the jail increasing its net population. All else held constant, with every percentage increase in the amount of white people in a jail, we expect there to be a 0.2% increase in jail population from 2020 Q1 to Q3.

Future:

- Collect and add data on political leanings of counties (judiciary, national, county)
- ggmap() to create a spatial visualization to represent the counties that decreased the jail population counts the most

- Look at the race percentage changes themselves to see whether there are disparities in the people being released during COVID-19
- Look at gender
- Look at other predictor variables: beds, capacity, etc.
- Look at intake and outtake #'s themselves
- See which jails increased capacity and beds over time
- Analyze death counts possibly

Comments:

- This doesn't look at the "after" numbers for any of the predictor variables (race, gender, ethnicity). Another thing we can do in the future is analyze the difference between races over time.
- are the race variables independent enough?
- try looking at which race percentage decreases the most
- look at bed increases or decreases
- look at capacity changes
- make helper functions to streamline the creation of other regression models
- ANova means to asess whether there's been a significant change due to COVID-19
- Look at Legalization, COVID-19, COVID-19 on Race
- How do I research whether having a larger percentage of Black people made jails more/less likely to release people?