Unconvicted Inmates in American Jails

Hubert Luo, Aniket Mandalik, Jessica Yu

Introduction

Our study examines the proportion of unconvicted inmates confined in United States jail jurisdictions in 2016. We utilized data from the Annual Survey of Jails, a survey conducted by the Department of Justice of the United States. This is an annual survey first started in 1982 that collects information about jail facilities and their inmates, including data on inmate demographics, jail occupancy statistics, and the supervision status of inmates. The data we used was collected during 2016, the last available year.

To examine the proportion of unconvicted inmates at jail jurisdictions in greater detail, we used data on race groups, gender, crime type, and citizenship status in our analyses to explore any patterns between the unconvicted proportion of inmates across these variables.

Our results from regression and chi-squared testing indicate that certain variables related to race, gender, and crime type are related to the proportion of unconvicted inmates confined in jail jurisdictions. For example, the proportion of Hispanic inmates seems to have a strong relationship with the proportion of unconvicted inmates. Meanwhile, citizenship status does not seem to be related to the proportion of unconvicted inmates. This initial analysis of the proportion of unconvicted inmates leads to interesting questions for future analyses, which may further examine the relationship between specific races as well as additional variables.

Survey Design

Survey Basics

The Survey Data we have with us is from the Annual Survey of Jails 2016 (ASJ). This is a survey run every year by the Bureau of Justice Statistics. Most of the statistics are collected from the web, but a small proportion of data was submitted through pre-paid postage/fax. The annual survey has been run by the Bureau of Justice Statistics since 1982, excluding the years 1983, 1988, 1993, 1999, and 2005. In the aforementioned five years, a census of the jails was taken instead of a survey. The purpose of the survey is to gather more information on the number and characteristics of inmates from year to year.

Survey Design Elements

The jail jurisdictions were grouped into ten strata. For two of the strata, all the jurisdictions in the strata were sampled. One of the strata included jurisdictions that manage jails operated by more than 1 jurisdiction (multi-county jail). The other included jurisdictions that were either in

California, had more than 750 prisoners daily on average, or had more than 500 prisoners daily on average and also held juveniles. The other strata were created based on the average daily population, and incarceration of juveniles. Figure 1 from the ASJ report[1] (page 11) provides more details on the various strata:

Table 1 - Sample Design for the 2016 Annual Survey of Jails (ASJ)

Stratum	Description		Number of jurisdictions in Census	Number of jurisdictions sampled	Number of ineligible jurisdictions	Number of respondent jurisdictions	Number of nonrespondent jurisdictions	Design Wt	Final Wt
1	Jurisdiction certainties based on ADP[a,c]		253	257[c]	0	244	13	1.000	See Table 2
1.1	California jail certainties[a,b]		65	65	1	63	1	1.000	See Table 2
2	Holding at least one juvenile on December 31, 2013	ADP between 264 and 499	95	35	1	34	0	2.714	2.714
3		ADP between 141 and 263	94	20	0	18	2	4.700	5.222
4		ADP between 69 and 140	72	8	0	8	0	9.000	9.000
5		ADP between 0 and 68	89	12	0	12	0	7.417	7.417
7	Holding adults only on December 31, 2013	ADP between 227 and 749	266	208	1	203	4	1.279	1.304
8		ADP between 103 and 226	407	83	1	78	4	4.904	5.155
9		ADP between 40 and 102	567	63	0	61	2	9.000	9.295
10		ADP between 0 and 39	894	60	1	58	1	14.900	15.157
12	Regional jail certainties[a,c]		69	70[c]	1	67	2	1.000	See Table 2
TOTALS			**2,871**	**881**	**6**	**846**	**29**		

Note: For this collection year, BJS implemented nonresponse weight adjustment procedures to account for missing data for respondents that did not participate. See Methodology for a description of nonresponse weight adjustment procedures. For the certainty stratum (1 and 12) weighting class adjustments were performed by jail size. See Table 2 for the final weights for stratum 1 and 12 by the weighting classes.

[a] A jurisdiction is a certainty if either (1) the nonregional jurisdiction held at least one juvenile on Census day and had an average daily population (ADP) of 500 or more, or (2) the nonregional jurisdiction held adults only on Census day and had an ADP of 750 or more. Stratum 1 also includes a small number of jails (5) identified as eligible for the ASJ following the 2013 Census.

[b] All California jails are included in the 2015 ASJ. (See Sampling Procedures).

[c] Includes new jurisdictions identified before the 2015 ASJ.

Figure 1: Detailed Strata Information

In the other eight strata, jurisdictions were sampled randomly (SRS), and data for all the jails under that jurisdiction were gathered. With this knowledge, we can say that the sampling method was a stratified one-stage cluster sample, where the stratum are those mentioned above, the PSUs are the jail jurisdictions, and the SSUs are the individual jails within those jurisdictions. Most jail jurisdictions have only one jail in each of them, but there are two-three incarceration centers in other jurisdictions. While the “unit of analysis” mentioned in the survey summary is the jail jurisdiction, that statement is contradicted by the fact that there is jail-level data, which leads us to believe the “unit of analysis” mentioned in the report is analogous to the sampling unit we learned about in class.

The public release data

There are a few variables that appear to have been changed from the raw data to the final report. The design weight (DESIGNWT) variable was used to give a weight to different jails based on the stratum they were a part of. A non-response rate of 3% on average caused the weights to be readjusted to the values in the final weight (FINALWT) to different values based on the non-response in each stratum. The survey recommends using the design weight for

[1] *Annual Survey of Jails, 2016: United States Department of Justice. Office of Justice Programs. Bureau of Justice Statistics*. Inter-University Consortium for Political and Social Research, 2016.

reference only, as the variable is only made for reference. Instead, the final weight should be used for computing national estimates.

Exploration of Design Elements

While the data collection was done in the vein of a stratified one-stage cluster sample, in the survey report (ASJ) it is mentioned that all the weights are done at the cluster level and that it would be ideal to aggregate all the statistics before doing analysis. Because of this, once all the data was aggregated, it was recommended that we treat the survey as if it was a stratified SRS of all the jurisdictions. FINALWT is the weight given to each jurisdiction based on the total population and response rates within each stratum so all samples in the same stratum have the same weight. Specifically, the final weight is the product of the sample weight and a non-response adjustment weight.

The weights have a fairly high spread, with the majority of jurisdiction samples having weight between 1.0 and 1.5, while those in other strata are spread between 2 and 15, as demonstrated in the histogram of the final weights below in Figure 2. None of the weights are high enough that they pose a big risk of bias for the final result.

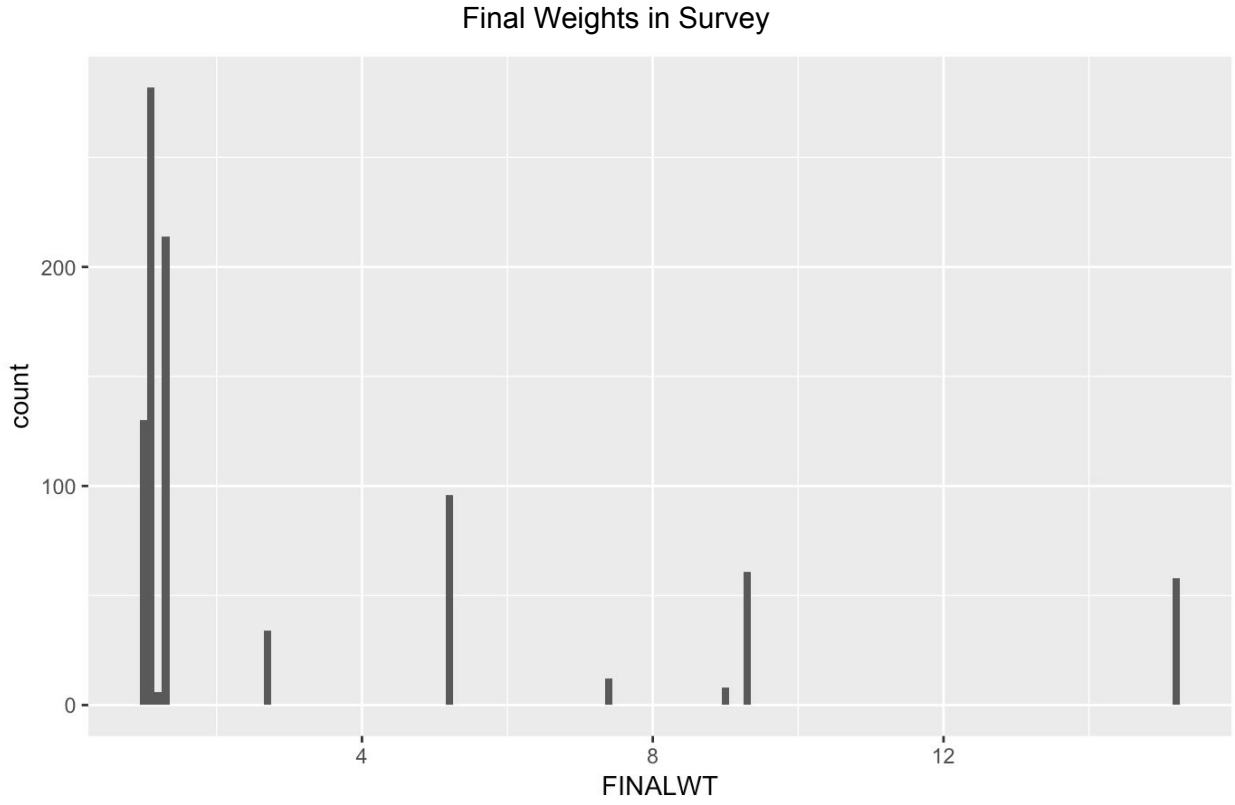

Figure 2. Histogram of Final Weights in Survey

Figure 3 below consists of a boxplot of the final weights for each strata plotted individually. It is clear that across the different strata, the box plots look fairly consistent, and all the interquartile regions intersect. Based on the spread of the mean, however, we cannot conclude that the means of the strata are the same. This reinforces the correctness of the decision to stratify, since we can now account for differences between the strata in our analysis.

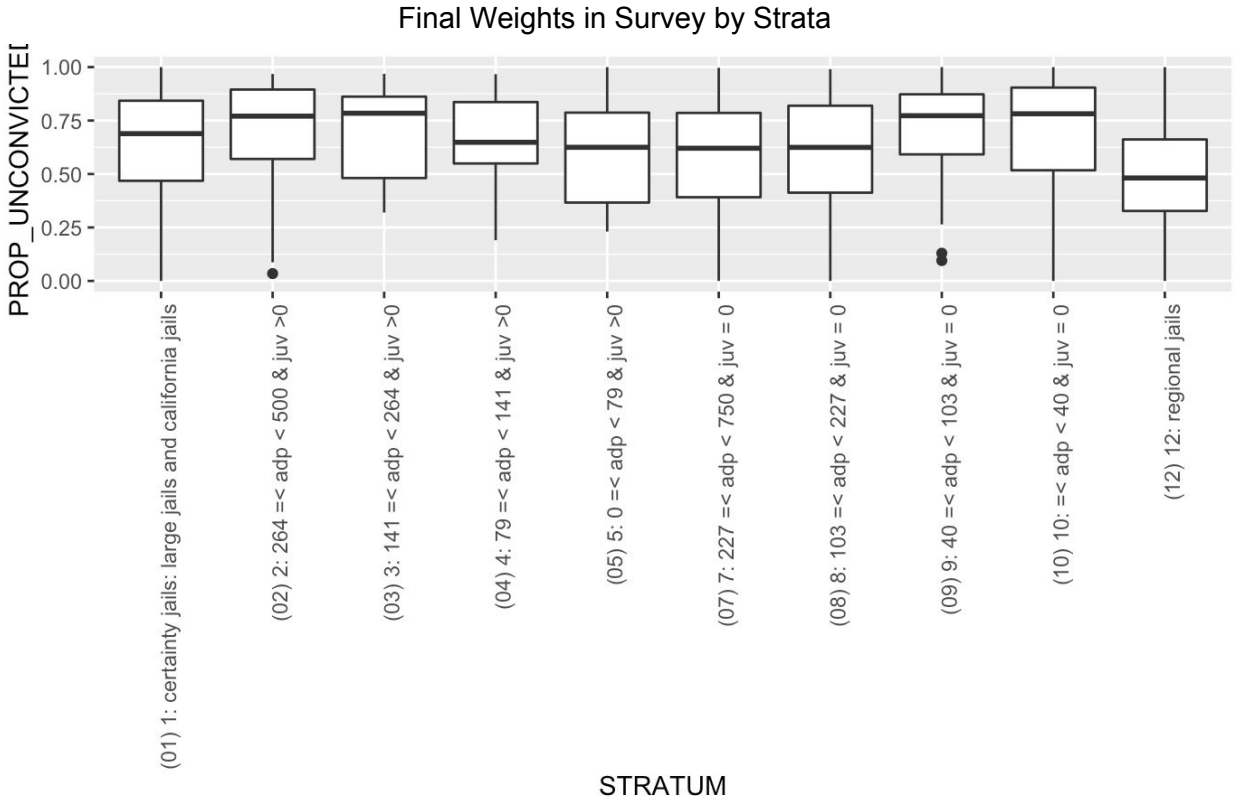

Figure 3. Boxplot of Final Weights in Survey for each Strata

Methodology

The data collected was from the Annual Survey of Jails, found at this URL link: https://www.icpsr.umich.edu/icpsrweb/NACJD/studies/37135/summary. In order to prepare the data for analysis, the first step necessary was to aggregate the original jail-level data into the jail jurisdiction-level data, as recommended by the Bureau of Justice Statistics (ASJ, Page 6). This aggregation was necessary since the final weights are the same for all the jail facilities in each jail jurisdiction, meaning that the weights are only relevant for a jail jurisdiction overall and not for an individual jail facility. Therefore, the data must first be aggregated by jail jurisdiction in order to provide meaningful analysis.

In order to aggregate the data, each jail facility was first grouped by jail jurisdiction (JURISID) and then aggregated totals were calculated for each relevant column, such as the year-end number of unconvicted inmates, correctional staff, confined population, ADP, rated capacity,

and inmates by offense type/race/gender/citizenship. The jail jurisdiction id (JURISID), stratum, and final weight were the same for each jail facility in a jal jurisdiction.

Variables and Feature Engineering

A number of new variables were created since the variables being examined were all proportions, while the data was primarily in totals. First, the response variable for the proportion of unconvicted inmates in a jail jurisdiction (PROP_UNCONVICTED) was created by dividing the year-end number of unconvicted inmates on December 31, 2016 (UNCONV) by the total year-end number of inmates of a jail jurisdiction, including both convicted and unconvicted, on December 31, 2016 (CONFPOP).

Then, explanatory variables were created, such as the percentage capacity of a jail jurisdiction on average (PERCENT_CAPACITY). This was calculated by dividing the average daily population of a jail jurisdiction between January 1, 2016 and December 31, 2016 (ADP) by the rated maximum capacity of jail jurisdictions as determined by the maximum number of beds or inmates (RATED).

For the other six explanatory variables which deal with race, type of crime, gender, and citizenship status, the original totals in the dataset were divided by the year-end total number of inmates of a jail jurisdiction (CONFPOP). The decision was made to divide them by CONFPOP rather than the average daily population of a jail jurisdiction (ADP) as the original total variables were in terms of the number of such inmates of a jail jurisdiction at the end of the year rather than the average daily population of such inmates. Using ADP would have led to numerous instances where the proportion would be found to be much greater than 100% depending on the difference between the number of inmates of a jail jurisdiction on December 31, 2016 compared to the annual average.

Specifically when working with the eight original variables related to race, we first combined five of them into a single new variable denoted as OTHERRACE_NEW. These five variables had lower individual counts per jail jurisdiction (all less than 1%), so the OTHERRACE_NEW variable was created by summing the counts of the five race categories, which were AIAN (American Indian or Alaska Native), ASIAN, NHOPI (Native Hawaiian or Pacific Islander), OTHERRACE, and TWORACE. The decision was made to combine these smaller individual totals into one overall total was so that the analysis would be more robust and less prone to severe fluctuations caused by outlier values in the smaller categories. Although there are clearly differences between the races included in this new variable, the advantages from making the analysis more robust to outliers outweighed concerns about over-generalizing the data, especially as the main variables of focus with regards to race were not included in this new cumulative variable.

As a result, we had four total variables for race that contained the year-end number of inmates at jails belonging to each group, specifically OTHERRACE_NEW, BLACK, WHITE, and HISP.

These groups were sufficiently large were not modified as they had larger counts per jail jurisdiction and represented larger groups within the data. They were all then divided by the year-end total number of inmates (CONFPOP) to get the proportion of inmates at the end of 2016 which belonged to each racial category.

In addition, another variable of interest was to examine the offense type an inmate had, which had three possible values of felony, misdemeanour, or other. Specifically, two new variables were created, the proportion of inmates who had an offense type of felony (PROP_FELONY), which was created by dividing the year-end number of inmates at a jail jurisdiction with an offense type of felony (FELONY) by the number of inmates at the end of 2016 (CONFPOP). In addition, the proportion of inmates who had an offense type of misdemeanor (PROP_MISD) by likewise dividing the year-end number of inmates at a jail (CONFPOP). The decision was made to include only two out of the three offense types because knowing these two proportions provides all the information about the third as they all sum up to 100%. Specifically felony and misdemeanour were chosen over other due to the fact that these offense types were more frequent than other, accounting for 95.41% of the offense types for inmates at the end of 2016 on average, while also providing more descriptive and detailed information than simply “other.”

The impact of gender was also examined by creating a variable for the proportion of female inmates (PROP_FEMALE), as calculated by dividing the year-end number of female inmates (ADULTF) by the number of inmates in a jail jurisdiction at the end of 2016 (CONFPOP). Finally, the impact of an inmate’s citizenship was investigated by creating a variable for the proportion of inmates at a jail jurisdiction who were not American citizens (PROP_NONCITZ). This was created by dividing the number of inmates in a jail jurisdiction who were non-American citizens (NONCITZ) by the year-end number of inmates (CONFPOP). This resulted in a total of one response variable and nine explanatory variables, all based on proportions.

Non-Response in Variables

There were five instances of item non-response, as five jail jurisdictions (0.55%) had no reported year-end number of inmates in a jail on December 31, 2016. Diving deeper into the specific details for each of these 5 jurisdictions, they were all missing information on not only the number of inmates at the end of 2016, but also the conviction status, race, offense type, and gender of inmates.

There were particular characteristics that distinguished the five non-responding jail jurisdictions from the other responding jails, specifically with regards to the number of correctional staff, the rated capacity, and the average daily population. These differences are explored in more detail below.

The five non-responding jail jurisdictions had much fewer correctional staff (median of 10) compared to the vast majority of other jail jurisdictions, which had a median of 66 for all jail jurisdictions. This is demonstrated graphically in Figure 4 below, a density plot which compares

the distribution for the number of correctional staff in a jail jurisdiction based on the response status of a jail jurisdictions. Note the plot excludes jail jurisdictions with more than 500 correctional staff in order to more clearly demonstrate the distribution of the non-responding jail jurisdictions.

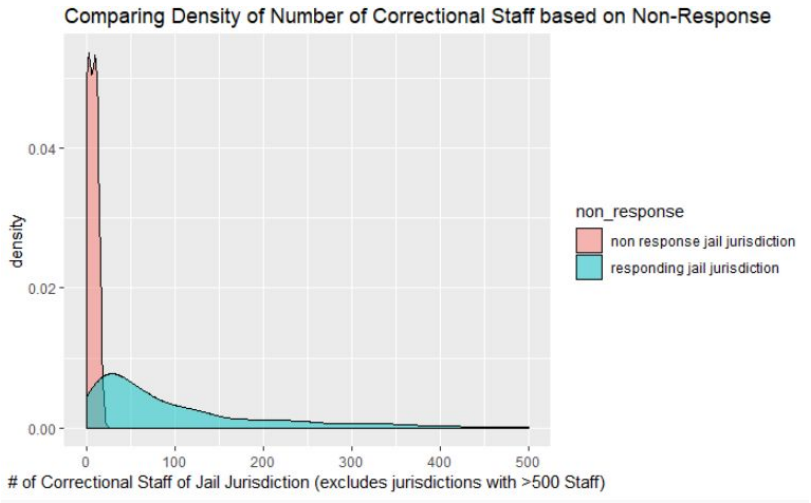

Figure 4: Comparing Density of Number of Correctional Staff based on Non-Response

In addition, these non-responding jail jurisdictions had much lower maximum capacities than almost all the other jails jurisdictions, with a median capacity of 32 compared to the median capacity for the whole dataset of 424. This is demonstrated graphically in Figure 5 below, a density plot which compares the distribution for the rated capacity of a jail jurisdiction based on its response status. Note the plot excludes jail jurisdictions with a rated capacity greater than 500 in order to more clearly demonstrate the distribution of the non-responding jail jurisdictions.

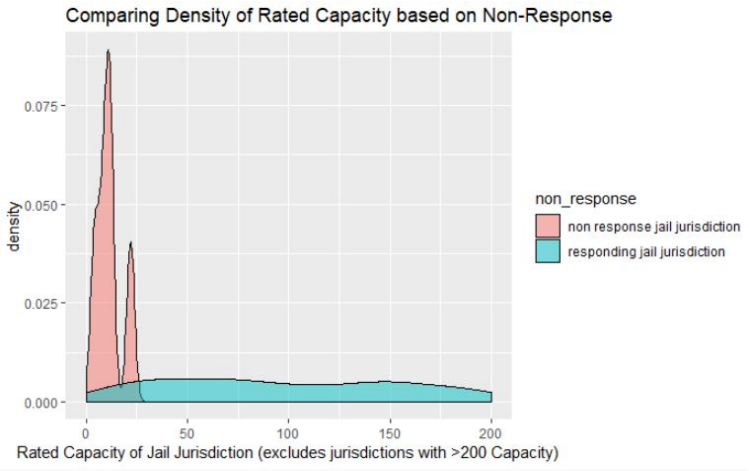

Figure 5: Comparing Density of Rated Capacity based on Non-Response

Furthermore, these five non-responding jail jurisdictions had an average daily population with a median of 11, much less than the median ADP for all jails in the survey of 350. This is

demonstrated graphically in Figure 6 below, which compares the distribution for the average daily population of a jail jurisdiction based on its response status. Note the plot excludes jail jurisdictions with an ADP of greater than 500 in order to more clearly demonstrate the distribution of the non-responding jail jurisdictions.

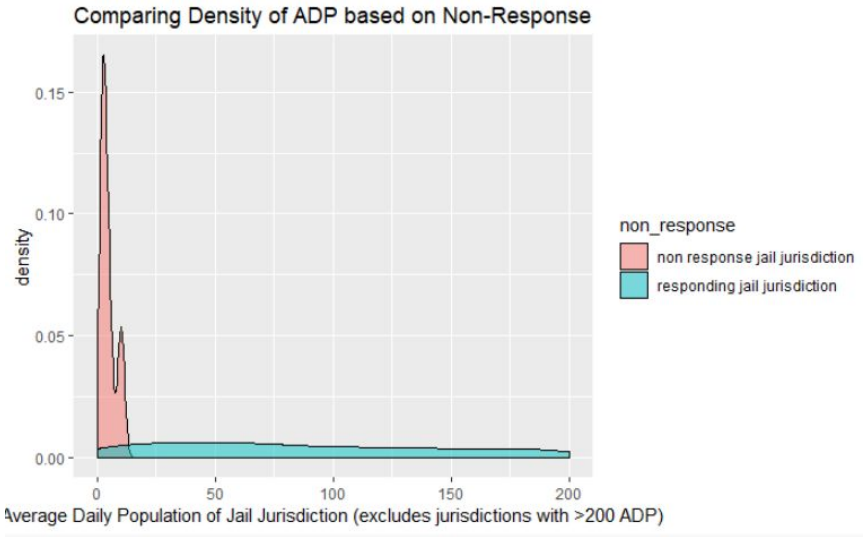

Figure 6: Comparing Average Daily Population based on Non-Response

Therefore, there are clear differences evident between non-responding jail jurisdictions and the other jail jurisdictions in the dataset in the known variables of the number of correctional staff members, the rated capacity, and the average daily population.

The data is missing variables for not only the year-end number of inmates but also the total number for each category based on conviction status, race, offense type, and gender. Most importantly, however, this data was not missing at random. It is clear the non-response jail jurisdictions have distinctive characteristics that differentiate them from other jail jurisdictions. Due to the fact the non-response jail jurisdictions were so different from those that responded, it would be unwise to simply ignore them as they have characteristics which are distinct from the other jails. Just removing the non-response jail jurisdictions from the analysis would bias the results by having a blind spot for smaller jail jurisdictions in this particular case.

Data was imputed based on known variables such as the average daily population (ADP), number of correctional staff (CORRSTAFF), and the rated capacity of the jail (RATED). The presence of these other known elements alleviates concerns that there would not be enough data to impute on.

In general, non-response might lead to misleading results by not including certain jail jurisdictions with characteristics that are different from those that respond. For example, if the non-responding jail jurisdictions are almost all located in Minnesota, not including these jail

jurisdictions in the analysis would lead to conclusions which do not properly generalize to all jail jurisdictions, which obviously would include jails in Minnesota.

Survey Design Object

To create the survey design object, the PSU used was the jail jurisdiction (JURISID), the strata was one of the ten strata previously defined (STRATUM), and the weights was the variable called FINALWT. Thus, the function call to create the survey design object using the survey package was: svydesign(ids = ~1, strata = ~STRATUM, weights = ~FINALWT, data = jails_imputed, fpc = ~fpc, nest = TRUE).

Although the data was collected using a stratified 1-stage cluster sample (ids = ~JURISID), the data was aggregated by jail jurisdiction above as recommended by the Bureau of Justice Statistics. As a result, there are no longer any clusters and the survey design is just that of a stratified sample. The strata in the sample are clearly given by the variable STRATUM, the weights are simply the weights for the survey (FINALWT), and the data argument refers to the dataset used, which is the jails dataset with imputed values (jails_imputed). In addition, the finite population correction factor was set to 2,871 as there were 2,871 jail jurisdictions in the United States according to the survey description (ASJ, 2016). Finally, the nest argument refers to whether the PSU are nested within a strata, which is the case for this dataset as a jail is nested within a group of jail jurisdictions (strata), so the argument nest is set to TRUE.

Results

Overview

We performed both simple and multiple regression involving our eight variables using svyglm() from the survey package. We also utilized chi-squared tests and created contingency tables. While the variables are not all normally distributed, as seen from the following figures 7-14 below of the distribution and Q-Q plots for each explanatory variable, we can still get some insights into factors that predict unconvicted inmate rates.

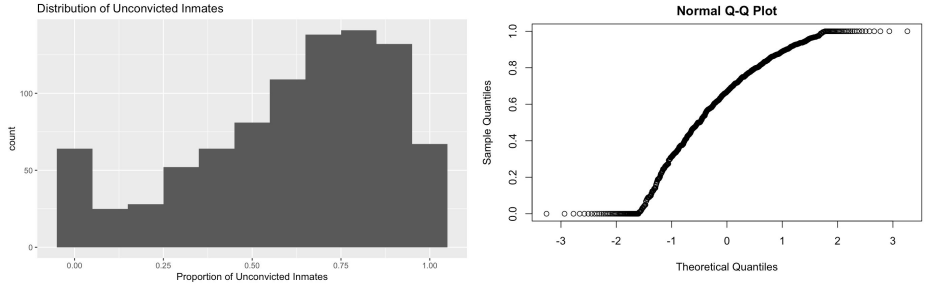

Figure 7. Distribution and Normal Q-Q Plot of Unconvicted Inmates

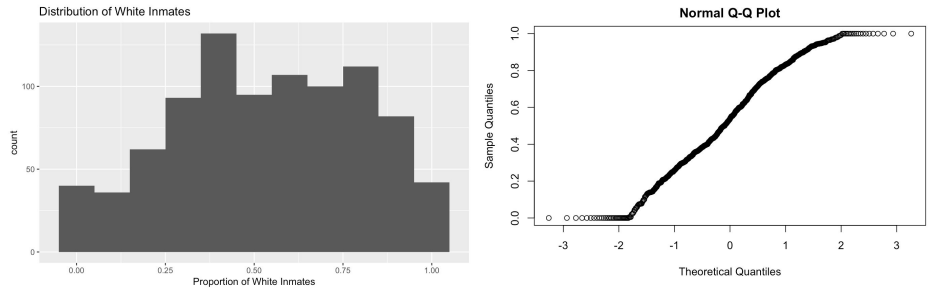

Figure 8. Distribution and Normal Q-Q Plot of Proportion of White Inmates

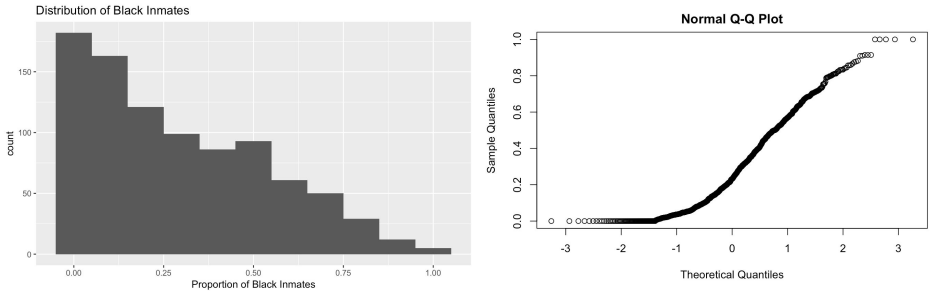

Figure 9. Distribution and Normal Q-Q Plot of Proportion of Black Inmates

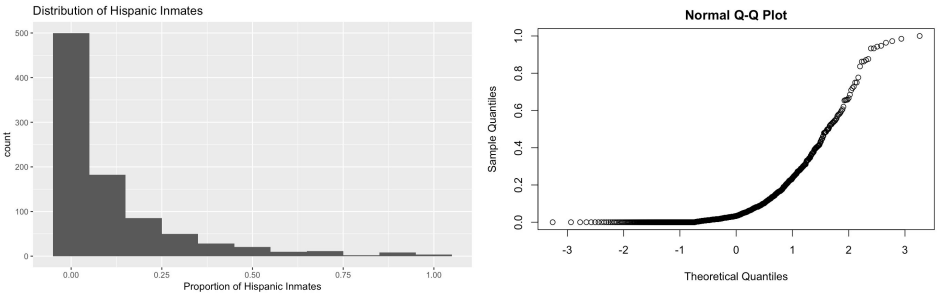

Figure 10. Distribution and Normal Q-Q Plot of Proportion of Hispanic Inmates

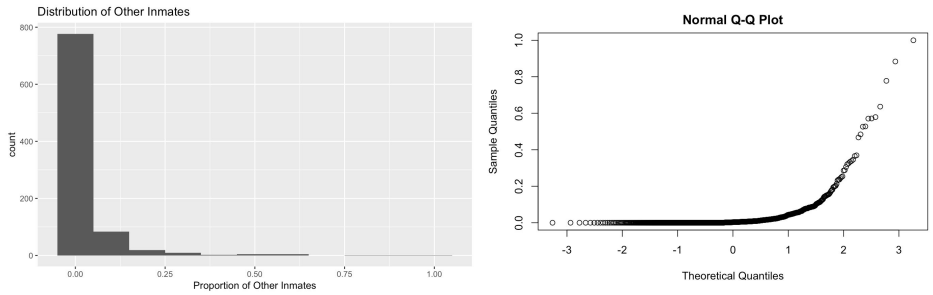

Figure 11. Distribution and Normal Q-Q Plot of Proportion of Inmates of Other Races

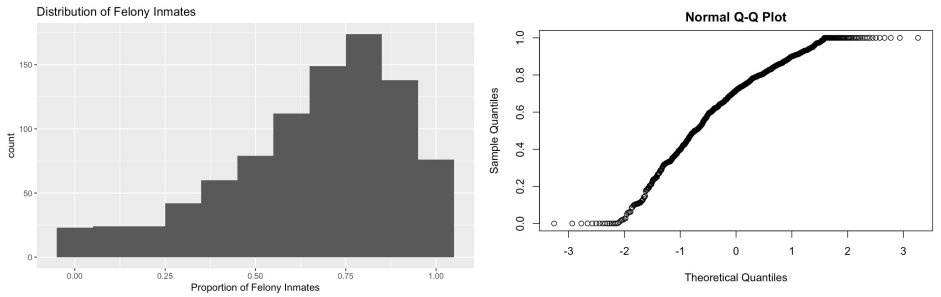

Figure 12. Distribution and Normal Q-Q Plot of Proportion of Inmates with Felony Offense Type

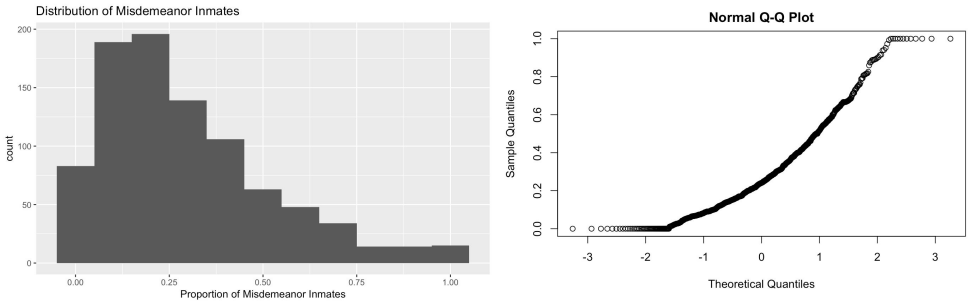

Figure 13. Distribution and Normal Q-Q Plot of Proportion of Inmates with Misdemeanour Offense Type

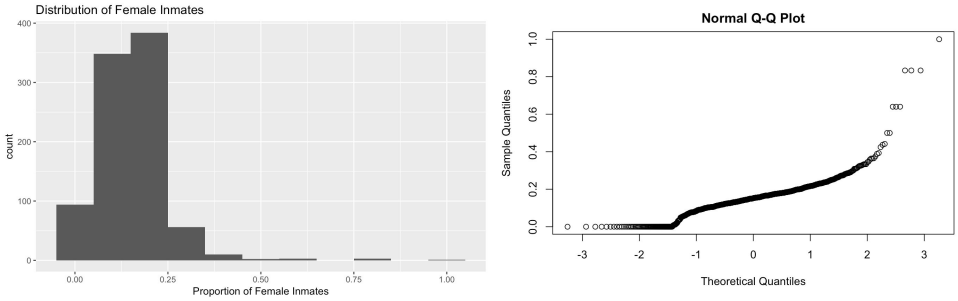

Figure 14. Distribution and Normal Q-Q Plot of Proportion of Female Inmates

Simple Regression and Diagnostics

We ran simple regressions using each of the eight explanatory variables. By extracting p-values from each of the simple regression outputs, we determined that the proportion of hispanic inmates, PROP_HISP, is the only significant variable at a significance level of .05. We created diagnostic plots of PROP_HISP to assess the appropriateness of using a linear model. In the Residuals vs. Fitted Values plot, we see that residuals for PROP_HISP are centered about 0 and generally follow a straight line. Residuals tend to be more spread out at low predicted values, indicating heteroskedasticity. The Quantile-Quantile plot shows that residuals generally fall on the straight line referencing normality, indicating that errors are roughly normally distributed. In the Scale-Location plot, residuals are more widely spread at low predicted values,

which confirm our concerns regarding heteroskedasticity. Nonetheless, the line passing through the residuals is horizontal, which indicates no upwards or downwards trend in residual values. In the last plot on Residuals vs Leverage, we see that no points lie outside of the Cook's distance lines. This indicates that there are no influential outliers in PROP_HISP for which removal would affect regression results. Overall, our regression diagnostic plots in Figure 16 below suggest that although it is not ideal, applying a linear model to PROP_HISP is appropriate.

Variable	P-Value
Black	0.283411
Hispanic	0.000049
White	0.108028
Other Race	0.587531
Felony	0.055401
Misdemeanour	0.089390
Female	0.063349
Non-Citizen	0.442232

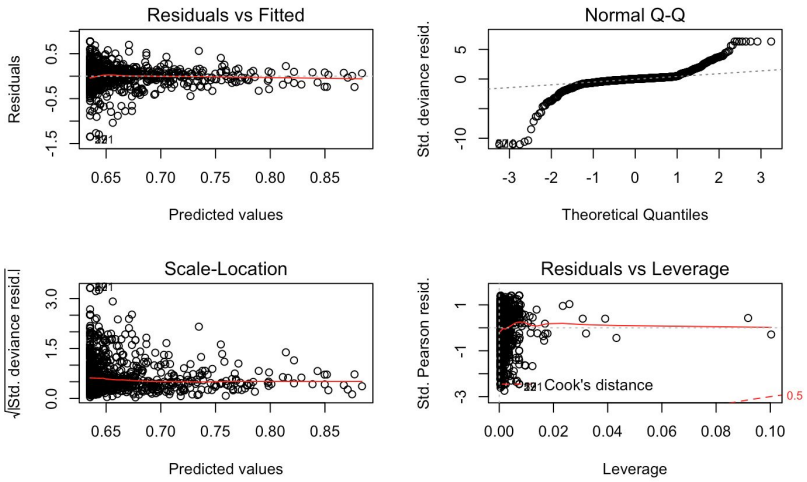

Figure 15. P-Values from Simple Regression. Figure 16. Simple Regression Diagnostics Plots

Variable	Estimate	Standard Error	t-value
Black	0.06872899	0.05459000	1.259003
Hispanic	0.2485114	0.06094416	4.077689
White	-0.09137526	0.05680026	-1.608712
Other Race	-0.07983853	0.14714271	-0.5425925
Felony	0.1459909	0.07610617	1.918253
Misdemeanour	-0.1383826	0.08137914	-1.700468
Female	0.3071114	0.16518818	1.859161
Non-Citizen	0.06039461	0.07856443	0.7687271

Figure 17: Simple Regression Coefficients, Standard Error, and t-Value for each Variable

Intuitively, impoverished groups of people, dangerous criminals, and those denied due process rights are more likely to be unconvicted but in prison nevertheless.

From Figures 15 and 17 above, we can see that PROP_HISP has a coefficient of .2485, and a p-value of $5x10^{-5}$, which is significant even with the Bonferroni-corrected p-value .05/9=.00556. This shows us that the proportion of Hispanic inmates has a significant positive relationship with the proportion of unconvicted inmates in a jail, i.e., if a jail's Hispanic inmate proportion increases by one, the unconvicted inmate proportion would be expected to increase by .2485. The correlation between the proportion of Hispanic and the proportion of non-citizens is 0.38, which is a moderate positive correlation. This leads us to believe that there is a positive relationship whereby the proportion of Hispanic inmates at jail jurisdictions tends to increase as the proportion of non-citizen inmates increases.

While the p-value is not significant enough to make conclusions about the other variables, a trend similar to the previous statement appears when jails have high proportions of felonies and misdemeanors. Judges are less likely to grant bail to those accused of felonies, while those accused of misdemeanors are more likely to have a lower bail and be released. Another non-significant but surprising result is that more female inmates is indicative of higher rates of unconvicted inmates. This could be because women accused of crimes are less likely to be able to afford bail than men.

Multiple Regression and Diagnostics

Next, we performed a multiple regression involving all eight variables. From the diagnostic plots, it appears that the residuals are generally normally distributed and centered about 0, with slight heteroskedasticity. It seems that residuals have lower variance at higher predicted values. Overall, there is no trend to the residuals and no influential points were identified. Although the fit is not precisely linear, we believe that regression analysis is still appropriate from the diagnostic plots. Multiple regression selected PROP_BLACK, PROP_HISP, and PROP_FEMALE as the significant variables, with the p-values as provided in Figure 18. We see that in both simple and multiple regression, PROP_HISP is selected as a significant variable.

Variable	P-Value
Black	0.009374
Hispanic	2.33×10^{-7}
White	0.377911
Other Race	0.984039
Felony	0.169034
Misdemeanour	0.953392
Female	0.020541
Non-Citizen	0.543324

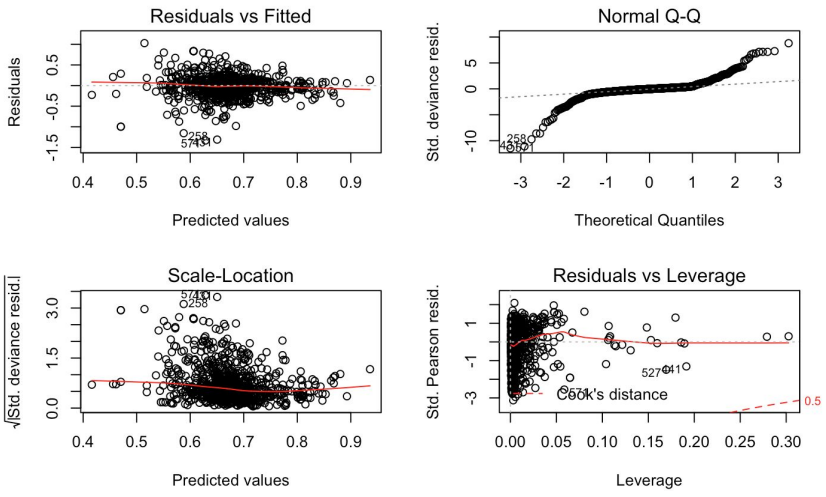

Figure 18. P-Values from Multiple Regression. Figure 19. Multiple Regression Diagnostics Plots

Variable	Estimate	Standard Error	t-value
Black	0.138741	0.053282	2.604
Hispanic	0.345201	0.066208	5.214
White	0.054126	0.061359	0.882
Other Race	0.002951	0.147498	0.020
Felony	0.129575	0.094142	1.376
Misdemeanour	-0.006359	0.108784	-0.058
Female	0.389822	0.167984	2.321
Non-Citizen	-0.044359	0.072960	0.608

Figure 20: Multiple Regression Coefficients, Standard Error, and t-Value for each Variable

Similar results were seen from multiple regression compared to the results from the simple regression performed earlier, with the proportion of hispanic inmates in a jail jurisdiction having the smallest p-value by far. Therefore, like earlier, the analysis suggests that the proportion of hispanic inmates at a jail is associated with the proportion of unconvicted inmates at that jail.

Important Variables Using Forward Selection

To narrow down our variables, we performed variable selection through stepwise regression. We used forward selection at an alpha level of .05 to choose the most significant variables in our regression. The three selected variables are PROP_HISP, PROP_FEMALE, and PROP_FELONY. These variables correspond to the unconvicted proportion of Hispanic inmates, female inmates, and inmates charged with a felony. This matches our simple and multiple regression where PROP_HISP is selected as a significant variable.

Chi-Squared Test

We took our eight variables and transformed the proportions they contained into categorical data involving three levels: “L”, “M”, and “H”, referring to low proportion, medium proportion, and high proportion. For example, for BLACK_PROP, we assigned jail jurisdictions below the lowest quartile (0-25th percentile) of BLACK_PROP values to low, values within the interquartile range (25-75th percentile) to medium, and values within the top quartile (75-100th percentile) to high. After applying this process to each of our variables, we obtained eight new categorical variables that we next used for our chi-squared tests.

We produced the contingency tables below in Figure 21. The 25th percentile of the proportion of inmates in a jail jurisdictions who were in the other racial group (PROP_OTHER) and the 25th percentile of the proportion of inmates in a jail jurisdiction who were non-citizens (PROP_NONCITZ) were zero. Therefore, there were no inmates assigned to the low categorical group, as all the jail jurisdictions which had zeroes, the lowest possible value, for these two variables were assigned to the medium categorical group instead.

```
                 PROP_BLACK
PROP_UNCONVICTED   H   L   M
               H  58  52 100
               L  62  57  93
               M  91 102 230
                PROP_HISP
PROP_UNCONVICTED   H   L   M
               H  58  49 103
               L  36  72 104
               M 118  90 215
                PROP_WHITE
PROP_UNCONVICTED   H   L   M
               H  46  64 100
               L  63  50  99
               M 103  97 223
                PROP_OTHER
PROP_UNCONVICTED   H   M
               H  50 160
               L  42 170
               M 119 304
                PROP_FELONY
PROP_UNCONVICTED   H   L   M
               H  49  55 106
               L  64  51  97
               M  98 105 220
                PROP_MISD
PROP_UNCONVICTED   H   L   M
               H  51  57 102
               L  54  66  92
               M 107  89 227
                PROP_FEMALE
PROP_UNCONVICTED   H   L   M
               H  52  55 103
               L  56  71  85
               M 102  85 236
                PROP_NONCITZ
PROP_UNCONVICTED   H   M
               H  55 155
               L  41 171
               M 116 307
```

Figure 21. Contingency Tables for Chi-Squared Analysis

After running chi-squared tests involving PROP_UNCONVICTED against each of the variables, we extracted the p-values of the test, as seen in Figure 22 below. The one variable with a significant p-value is PROP_MISD.

For the Chi-Square tests, our null hypothesis is that all categorical variables are independent from the categorical unconvicted incarceration rate. From Figure 22 below, we can see, using quantile-based classes, that because none of the p-values were statistically significant (accounting for the Bonferroni correction), we fail to reject the null hypothesis. While the Misdemeanor variable appears to be significantly dependent, this is not true if we use the Bonferroni correction p-value .05/9=0.00556. Therefore, we are unable to conclude that any of the categorical variables have an effect on the categorical unconvicted incarceration rate.

Variable	p-value
Black	0.4405
Hispanic	0.1219
White	0.1356
Other Race	0.1942
Felony	0.1596
Misdemeanour	0.0287
Female	0.2263
Non-Citizen	0.3414

Figure 22. P-values from Chi-Squared Tests

Estimates of the Total Number of Unconvicted Inmates in Local Jails

Estimates were also conducted of the total number of unconvicted inmates and the total number of inmates confined in local jails in the United States. Using the survey design outlined earlier in this report, the estimated total number of unconvicted inmates is 458,533 with a standard error of 19,858. The estimated total number of inmates confined in local American jails is 704,021 with a standard error of 26,411. This is extremely close to the total number of inmates confined in local American jails reported by the US Department of Justice, which estimates that 704,500 inmates are in local jails[2]. Therefore, the results are in line with what is expected and supports the validity of the survey design used in this report.

Discussion/Conclusion

Our results indicate that in examining the unconvicted proportion of inmates held in jail jurisdictions in the United States, certain variables related to race, gender, and crime type may

[2] "United States of America." *The World Prison Brief*, http://www.prisonstudies.org/country/united-states-america. Accessed 9 May 2019.

be useful. We used simple regression, multiple regression, and chi-squared tests along with diagnostic plots to assess the relationship of these variables with the unconvicted proportion of inmates. In performing regression, we examined diagnostic plots to see the appropriateness of fitting linear models to our data. Through residual plots, we observed that the residuals have some heteroskedasticity, which may be problematic in fitting linear models. However, we also found that the residuals are generally clustered around zero and do not exhibit any upwards or downwards trend, and that the quantile-quantile plots indicate approximate normality. Thus, we included linear models as part of our analysis.

For each jail jurisdiction, we transformed the variables, which originally involved proportions, to be categorized as either a low, medium, or high proportion so that we could perform chi-squared testing. Chi-squared testing identified the crime type (specifically as related to the proportion of inmates charged with misdemeanors).

Overall, it appears that the proportion of hispanic inmates came up frequently in our analyses as a useful variable. The crime type may also be useful. Some interesting follow-up analyses may involve looking further into the proportion of hispanic inmates held in jail jurisdictions. Given more detailed data on inmates rather than on jail facilities and jail jurisdictions, we may also want to study the interaction between race and crime type.

Appendix

```
###############
#Libraries
###############
library(knitr)
library(dplyr)
library(ggplot2)
library(survey)
library(mice)

###############
#Variables and Feature Engineering
###############

load("152proj_data.rda")
jails = da37135.0001

# Aggregate jails by jail jurisdiction. Stratum, FINALWT, and JURISID are the same for each jail in a jail jurisdiction
jails = jails %>% group_by(JURISID) %>% summarise(UNCONV = sum(UNCONV), CONFPOP = sum(CONFPOP), ADP = sum(ADP), RATED = sum(RATED), OTHERRACE_NEW = sum(AIAN + ASIAN + NHOPI + OTHERRACE + TWORACE), BLACK = sum(BLACK), HISP = sum(HISP), WHITE = sum(WHITE), FELONY = sum(FELONY), MISD = sum(MISD), ADULTF = sum(ADULTF), NONCITZ = sum(NONCITZ), STRATUM = first(STRATUM), FINALWT = mean(FINALWT), CORRSTAFF = sum(CORRSTAFF))

# Response variable is PROP_UNCONVICTED
jails$PROP_UNCONVICTED = jails$UNCONV/jails$CONFPOP

# Percentage capacity of a jail (Average daily population divided by rated capacity)
jails$PROP_CAPACITY = jails$ADP/jails$RATED

# Convert the number of inmates of each race into a proporiton
```

```
jails$PROP_BLACK = jails$BLACK/jails$CONFPOP
jails$PROP_HISP = jails$HISP/jails$CONFPOP
jails$PROP_WHITE = jails$WHITE/jails$CONFPOP
jails$PROP_OTHER = jails$OTHERRACE_NEW/jails$CONFPOP

# Get the proportion of inmates who were charged with a felony/misdemeanor and the proportion who were female
jails$PROP_FELONY = jails$FELONY/jails$CONFPOP
jails$PROP_MISD = jails$MISD/jails$CONFPOP
jails$PROP_FEMALE = jails$ADULTF/jails$CONFPOP
jails$PROP_NONCITZ = jails$NONCITZ/jails$CONFPOP

# All columns of jails dataset aggregated by jurisdiction
jails_by_jurisdiction_orig = jails

variables_of_interest = c('PROP_UNCONVICTED', 'PROP_CAPACITY', 'PROP_BLACK', 'PROP_HISP', 'PROP_WHITE', 'PROP_OTHER', 'PROP_FELONY', 'PROP_MISD', 'PROP_FEMALE','PROP_NONCITZ','JURISID','STRATUM','FINALWT')

jails = jails %>% select(variables_of_interest)

# Remove outlier of jail at supposedly 786% capacity
jails_by_jurisdiction_orig = jails_by_jurisdiction_orig[-(jails$PROP_CAPACITY > 7),]
jails = jails[-(jails$PROP_CAPACITY > 7),]

###############
#Missing Values
###############

missing_jails = jails_by_jurisdiction_orig[is.na(jails$PROP_UNCONVICTED),]

summary(jails_by_jurisdiction_orig$ADP)
summary(jails_by_jurisdiction_orig$RATED)
summary(jails_by_jurisdiction_orig$CORRSTAFF)

jails_by_jurisdiction_orig$non_response = 'responding jail jurisdiction'
jails_by_jurisdiction_orig[is.na(jails$PROP_UNCONVICTED), 'non_response'] = 'non response jail jurisdiction'

ggplot(data = jails_by_jurisdiction_orig) +
    geom_density(aes(x = ADP, fill = non_response), alpha = 0.5) +
  ggtitle("Comparing Density of ADP based on Non-Response") +
 xlim(0,200) +
   labs(x = "Average Daily Population of Jail Jurisdiction (excludes jurisdictions with >200 ADP)")

ggplot(data = jails_by_jurisdiction_orig) +
     geom_density(aes(x = RATED, fill = non_response), alpha = 0.5) +
  ggtitle("Comparing Density of Rated Capacity based on Non-Response") +
 xlim(0,200) +
   labs(x = "Rated Capacity of Jail Jurisdiction (excludes jurisdictions with >200 Capacity)")

ggplot(data = jails_by_jurisdiction_orig) +
   geom_density(aes(x = CORRSTAFF, fill = non_response), alpha = 0.5) +
    ggtitle("Comparing Density of Number of Correctional Staff based on Non-Response") +
 xlim(0,500) +
    labs(x = "# of Correctional Staff of Jail Jurisdiction (excludes jurisdictions with >500 Staff)")

jails[is.na(jails$PROP_UNCONVICTED),]

#jails = jails[-is.na(jails$PROP_UNCONVICTED),]
summary(jails)

###############
#Imputing Missing Data
```

```
###############

# Add columns for ADP, CORRSTAFF, and RATED as these are known variables for imputing the missing values
jails_full = jails_by_jurisdiction_orig %>% select(ADP, CORRSTAFF, RATED) %>% cbind(jails)
# Warning: this line below that creates the mids (mice) object takes a couple minutes to run
jails_mice = mice(jails_full, m = 1)

# Compare the imputed values with original missing values
complete(jails_mice)[is.na(jails$PROP_UNCONVICTED),-(1:3)]
jails[is.na(jails$PROP_UNCONVICTED),]

# Create new dataset called jails_imputed that has the imputed data instead of missing data
jails_imputed = complete(jails_mice)[-(1:3)]

jails_imputed$fpc = 2871

###############
#Survey Design
###############

# survey design object
svydesign <- svydesign(ids = ~1 , strata = ~STRATUM, weights = ~FINALWT, data = jails_imputed, nest = TRUE, fpc = ~fpc)

###############
#Simple Regression
###############

# First Evaluate Normality of Explanatory Variables

ggplot(data) +
                    geom_histogram(aes(x = PROP_UNCONVICTED), binwidth = 0.1) +
   labs(x = "Proportion of Unconvicted Inmates") +
  ggtitle("Distribution of Unconvicted Inmates")
qqnorm(data$PROP_UNCONVICTED)

ggplot(data) +
     geom_histogram(aes(x = PROP_WHITE), binwidth = 0.1) +
  labs(x = "Proportion of White Inmates") +
  ggtitle("Distribution of White Inmates")
qqnorm(data$PROP_WHITE)

ggplot(data) +
      geom_histogram(aes(x = PROP_HISP), binwidth = 0.1) +
  labs(x = "Proportion of Hispanic Inmates") +
  ggtitle("Distribution of Hispanic Inmates")
qqnorm(data$PROP_HISP)

ggplot(data) +
     geom_histogram(aes(x = PROP_OTHER), binwidth = 0.1) +
  labs(x = "Proportion of Other Inmates") +
  ggtitle("Distribution of Other Inmates")
qqnorm(data$PROP_OTHER)

ggplot(data) +
    geom_histogram(aes(x = PROP_FEMALE), binwidth = 0.1) +
  labs(x = "Proportion of Female Inmates") +
  ggtitle("Distribution of Female Inmates")
qqnorm(data$PROP_FEMALE)

ggplot(data) +
    geom_histogram(aes(x = PROP_FELONY), binwidth = 0.1) +
  labs(x = "Proportion of Felony Inmates") +
  ggtitle("Distribution of Felony Inmates")
qqnorm(data$PROP_FELONY)

ggplot(data) +
      geom_histogram(aes(x = PROP_MISD), binwidth = 0.1) +
   labs(x = "Proportion of Misdemeanor Inmates") +
  ggtitle("Distribution of Misdemeanor Inmates")
qqnorm(data$PROP_MISD)

ggplot(data) +
    geom_histogram(aes(x = PROP_NONCITZ), binwidth = 0.1) +
  labs(x = "Proportion of Non-Citizen Inmates") +
  ggtitle("Distribution of Non-Citizen Inmates")
```

```
qqnorm(data$PROP_NONCITZ)

ggplot(data) +
  geom_histogram(aes(x = PROP_CAPACITY), binwidth = 0.1) +
  labs(x = "Proportion Occupied") +
  ggtitle("Distribution of Proportion Occupied")
qqnorm(data$PROP_CAPACITY)

# explanatory variables
variables <- c("PROP_BLACK", "PROP_HISP", "PROP_WHITE", "PROP_OTHER", "PROP_FELONY", "PROP_MISD", "PROP_FEMALE", "PROP_NONCITZ")

# collect p-values from simple regression on each of the eight variables
p_values <- c()

for (i in 1:length(variables)) {
    formula <- paste("PROP_UNCONVICTED", "~", variables[i])
      glm_obj <- svyglm(formula, design = svydesign)
  summary <- summary(glm_obj)$coefficients
  p_values[i] <- summary[2,4]
  print(summary)
}

# create a matrix of p-values from simple regression
matrix(p_values, dimnames = list(variables, "p-values"))

# identify significant variables at level of .05
variables[which(p_values <= .05)]

# diagnostic plots for PROP_HISP
svyglm_hisp <- svyglm(PROP_UNCONVICTED~PROP_HISP, design = svydesign)
par(mfrow = c(2, 2))
plot(svyglm_hisp)

###############
#Multiple Regression and Forward Selection
###############

formula <- paste("PROP_UNCONVICTED", paste(variables, collapse = "+"), sep = "~")
multiple_regression <- svyglm(formula, design = svydesign)
multiple_regression

# multiple regression output
summary(multiple_regression)

# extract p-value from multiple regression
mult_reg_p <- summary(multiple_regression)$coefficients[,4]
mult_reg_p_df <- as.data.frame(matrix(mult_reg_p))
names(mult_reg_p_df) <- "p"
mult_reg_p_df

# diagnostic plots of multiple regression
par(mfrow = c(2, 2))
plot(multiple_regression)

# stepwise regression for forward selection
# pick first variable to include in the model, i.e. the one with the lowest p value <= 0.05
p <- c()
selected_variables <- c()

for (i in 1:length(variables)) {

  formula <- paste0('PROP_UNCONVICTED ~ ', paste(selected_variables), " + ", variables[i])
  svyglm <- svyglm(formula, design = svydesign)
    p[i] <- summary(svyglm)$coefficients[2,4] # picks out p value
}

var_min_p <- variables[which.min(p)]
selected_variables <- c(selected_variables, var_min_p)
selected_variables

# select the rest of the variables using forward selection
variables <- variables[-which.min(p)]
while (any(p <= .05))  {
 p <- c()
 for (i in 1:length(variables)) {
```

```
 formula <- paste0('PROP_UNCONVICTED ~ ', paste(selected_variables, collapse = " + "), " + ", variables[i])
  svyglm <- svyglm(formula, design = svydesign)
                                    p[i] <- summary(svyglm)$coefficients[nrow(summary(svyglm)$coefficients),4]
  names(p)[i]
}

  var_min_p <- variables[which.min(p)]
    selected_variables <- c(selected_variables, var_min_p)
  variables <- variables[-which.min(p)] # removes the selected variable

}

################
#Chi Squared
################

# make sure all 8 variables are listed since some were removed in forward selection
variables <- c("PROP_UNCONVICTED", "PROP_BLACK", "PROP_HISP", "PROP_WHITE", "PROP_OTHER", "PROP_FELONY", "PROP_MISD", "PROP_FEMALE", "PROP_NONCITZ")

# extract first and third quartile values to be used in creating low/med/high proportion categories later

first_quartile <- c()
third_quartile <- c()

for (i in 1:length(variables)) {

                        variable_col_index <- which(names(jails_imputed)== variables[i])
        summary <- summary(jails_imputed[ ,variable_col_index])
  first_quartile[i] <- summary[2]
  third_quartile[i] <- summary[5]
  print(summary)
}

# create a matrix of quartile values for each of the variables
quartiles_matrix <- matrix(c(first_quartile, third_quartile), byrow = T, nrow = 2, dimnames = list(c("Q1", "Q3"), variables))
quartiles_matrix

# set up a new data frame to contain the categories (low/med/high proportions) for each variable
categories <- as.data.frame(matrix(0, nrow = nrow(jails_imputed), ncol = length(variables)))
names(categories) <- variables

for (i in 1:length(variables)) {

   # extract the column corresponding to the variable from jails_imputed
                        variable_col_index <- which(names(jails_imputed) == variables[i])
  values <- jails_imputed[ , variable_col_index]

  # extract quartile 1 and quartile 3 values
  q1 <- quartiles_matrix[1, variables[i]]
  q3 <- quartiles_matrix[2, variables[i]]

   # find row indexes for assigning low/med/high in the next step
  low_index <- which(values < q1)
   med_index <- which((q1 <= values) & (values <= q3))
  high_index <- which(q3 < values)

  # fill the categories data frame with the labels
  categories[low_index, variables[i]] <- "L"
  categories[med_index, variables[i]] <- "M"
  categories[high_index, variables[i]] <- "H"

}

head(categories)

# column bind the categories table with STRATUM and FINALWT so that a new survey design object can be created for chi squared testing
```

```
jails_imputed2 <- cbind(jails_imputed$STRATUM, jails_imputed$FINALWT, categories)
names(jails_imputed2)[1:2] <- c("STRATUM", "FINALWT")
head(jails_imputed2)

# create survey design object
svydesign2 <- svydesign(ids = ~1, strata = ~STRATUM, weights = ~FINALWT, data = jails_imputed2, nest = TRUE, fpc = rep(2871,nrow(jails_imputed2)))
svydesign2

# create contingency tables
variables <- variables[-1] # drop PROP_UNCONVICTED
for (i in 1:length(variables)) {

  formula <- paste("~PROP_UNCONVICTED +", variables[i])
  table <- xtabs(formula, data = jails_imputed2)
  print(table)

}

# extract p-values for chi-squared tests

chisq_pvalues <- c(svychisq(~PROP_UNCONVICTED + PROP_BLACK, design = svydesign2, statistic = "Chisq")$p.value,
svychisq(~PROP_UNCONVICTED + PROP_HISP, design = svydesign2, statistic = "Chisq")$p.value,
svychisq(~PROP_UNCONVICTED + PROP_WHITE, design = svydesign2, statistic = "Chisq")$p.value,
svychisq(~PROP_UNCONVICTED + PROP_OTHER, design = svydesign2, statistic = "Chisq")$p.value,
svychisq(~PROP_UNCONVICTED + PROP_FELONY, design = svydesign2, statistic = "Chisq")$p.value,
svychisq(~PROP_UNCONVICTED + PROP_MISD, design = svydesign2, statistic = "Chisq")$p.value,
svychisq(~PROP_UNCONVICTED + PROP_FEMALE, design = svydesign2, statistic = "Chisq")$p.value,
svychisq(~PROP_UNCONVICTED + PROP_NONCITZ, design = svydesign2, statistic = "Chisq")$p.value)

# create data frame of chi-squared p-values
chisq_p <- as.data.frame(matrix(chisq_pvalues))
names(chisq_p) <- "p"
chisq_p

# pick out significant variables at alpha = .05 for chi-squared p-values

variables[which(chisq_p$p <= .05)]

# examine correlation between the selected variables
cor(jails_imputed$PROP_FELONY, jails_imputed$PROP_MISD)

################
#Estimate of Total Number of Unconvicted Inmates
################

```{r}
total_svy_dsgn = svydesign(id = ~1, strata = ~STRATUM, weights = ~FINALWT, data = jails_by_jurisdiction_orig, fpc = rep(2871,nrow(jails_by_jurisdiction_orig)))
svytotal(~UNCONV, total_svy_dsgn)
svytotal(~CONFPOP, total_svy_dsgn)
```
```