An Exploratory Analysis of King County's Jail Data

Bryce McManus

10/09/2019

Contents

Introduction

Incarceration data collection and analysis has generally focused on prison populations, rather than local jail populations. Consequently, most agency reports, policy studies, and academic research on incarceration concentrate on state and national level prison trends. While some jail-oriented studies do exist, these also tend to be analyzed at the state and national level, and rarely focus on local jail populations. However, local jails constitute a far greater proportion of the incarceration admissions than state and federal prisons, nearly 19 times higher as of 2015. They are also considered the "gateway" to the criminal justice system.[1]

Fewer studies have been done on local jails because finding data has generally been difficult and often involves the consent and cooperation of the jail, or the state department of correction. Recently, however, this trend has been diminishing with the advent of open data initiatives aimed at creating greater data transparency in states and local communities. King County in Washington state, for example, has gathered many types of data of interest to the public–including law enforcement response, housing, and transportation data. The King County Open Data Project[2] website allows anyone to access the data, and offers helpful data visualization tools that enable users to explore on their own.

This report makes use of this available data, providing a brief overview of data from King County adult correctional facilities from June 2018 to May 2019.[3] Part I explains how the data was collected and cleaned,

[1] https://www.vera.org/downloads/publications/incarcerations-front-door-report_02.pdf, accessed 10/02/2019

[2] https://data.kingcounty.gov/

[3] At the time of writing this report, only a single year of data was available. However, the data is updated monthly. One month after collecting the data used here, only data from June 2018 to July 2019 was available.

and also includes a description of the data. Part II covers summary statistics of key variables–such as total jail time, the number of bookings, and the frequency of charges. Part III briefly explores the relationship between recidivism, offense type, incarceration length, and release reason. Given the broad scope of the report and the limitations of the data, the last section focuses mainly on general patterns between the variables and is guided by the following questions:

1. What type of crimes tend to occur the most, and how do they vary by offender type?
2. How long do inmates tend to be incarcerated? Are repeat offenders incarcerated for longer than first-time offenders for the same offenses, and if so, by how much?
3. What are the most common reasons an inmate is released? How does the type of offense affect the reason an inmate is released?

The available data is comprised solely of adults held in custody and does not include juvenile detention facilities or adults under community supervision, such as probation. It also does not contain demographic information.

Regarding the measurement of recidivism, an inmate who has been booked on two or more separate occasions during the period of June 2018 to May 2019 is considered a “repeat” offender. Otherwise, they are categorized as a “first-time” offender. Given the limited time frame of the data available, this is not a perfect measure. It is quite possible that a “first-time” offender was arrested at some point before June 2018. Therefore, it is likely that this approach under-estimates the number of repeat offenders in the data set and should be considered a conservative approximation of recidivism.

The report concludes with a comparison between King County and national trends as reported by the Bureau of Justices Statistics, and points to potential areas of future analysis.

Preprocessing

Library

The analysis was done in R, version 3.5.0. The packages used are listed below.

```
library(tidyverse)
library(lubridate)
library(hms)
library(stringi)
library(quanteda)
library(tidytext)
library(wordcloud)
library(reshape2)
library(moments)
library(scales)
library(gridExtra)
library(kableExtra)
library(seriation)
```

Source

The data set used in this analysis can be found at the King County Open Data website.[4] The structure of the data is shown below.

[4] The most recent data can accessed at: https://data.kingcounty.gov/api/views/j56h-zgnm/rows.csv?accessType=DOWNLOAD.

```
jail_df <- read_csv("Adult_Jail_Booking_June_1__2018_to_May_31__2019_as_of_June_6__2019.csv")

jail_df %>% glimpse()
```

```
## Observations: 56,988
## Variables: 13
## $ `Book of Arrest Number`     <dbl> 218015189, 218015191, 218015194, 2...
## $ `Last Name`                 <chr> "BRICKSON", "CARNLEY", "ROGERS", "...
## $ `First Name`                <chr> "BEAU", "AUTUMN", "RAYMECKO", "RAY...
## $ `Middle Name`               <chr> "FINLEY", "HEATHER", "DELANI", "DE...
## $ JrSr                        <chr> NA, NA, NA, NA, NA, NA, NA, NA, "2...
## $ `Booking Date Time`         <chr> "06/01/2018 12:23:00 AM", "06/01/2...
## $ `Release Date Time`         <chr> "06/01/2018 05:32:00 PM", "06/01/2...
## $ `Current Facility`          <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ Charge                      <chr> "BURGLARY INV", "BURGLARY INV", "F...
## $ `Court Case / Cause Number` <chr> NA, NA, "615788", "171024628", "63...
## $ Court                       <chr> NA, NA, "Seattle Municipal Court",...
## $ `RCW / Ordinance Number`    <chr> "2299", "2299", "11.56.020", "46.6...
## $ `Release Reason`            <chr> "CONDITIONAL/COURT RELEASE", "COND...
```

The data consists of 13 columns and 56,988 rows, meaning there are 13 variables and 56,988 observations. These include the inmates' name, a unique booking number, and the RCW statute they were charged with. Inmates were often charged with more than one crime, meaning that some bookings include multiple rows.

King County Open Data does not give detailed descriptions of each variable in the data set. Most variables are self-evident or can be understood through context; however, others are ambiguous and harder to interpret.

The variable "Book of Arrest Number" is the unique number assigned to each booking. The next four variables include information about the names of the inmates booked. The booking and release columns contain the date and time the inmate entered and left the jail. "Current Facility" contains only NAs, or missing information, but likely refers to whether an inmate is housed at one of two of King County's adult correctional facilities: King County Correctional Facility and the Regional Justice Center. "Charge" is a short description of what the inmate was charged with. "Court Case / Court Number" appears to be for inmates that have a court case pending. "Court" refers to the name of the court the inmate has been assigned to. "RCW / Ordinance Number" refers to the Revised Code of Washington (RCW), the state's statutes, and/or the local ordinance the inmate was charged with violating. Lastly, "Release Reason" states why the inmate was released, if applicable.

Below is a sample of how the data are formatted.

Book of Arrest Number	Last Name	First Name	Middle Name	JrSr	Booking Date Time
218015189	BRICKSON	BEAU	FINLEY	NA	06/01/2018 12:23:00 AM
218015191	CARNLEY	AUTUMN	HEATHER	NA	06/01/2018 12:34:00 AM
218015194	ROGERS	RAYMECKO	DELANI	NA	06/01/2018 12:11:00 AM

Cleaning the Data

Cleaning the data started with renaming the variables so that they are easier to reference in-code.

```
jail_df <- jail_df %>% rename(boa_number = "Book of Arrest Number",
                              last_name = "Last Name",
                              first_name = "First Name",
                              middle_name = "Middle Name",
                              jrsr = "JrSr",
                              booking_dt = "Booking Date Time",
                              release_dt = "Release Date Time",
                              release_reason = "Release Reason",
                              current_facility = "Current Facility",
                              court_case = "Court Case / Cause Number",
                              charge = "Charge",
                              rcw = "RCW / Ordinance Number",
                              court = "Court")
```

Next, I change the format of the booking and release dates so that they can be used for calculations. Then, I take the difference to determine the total amount of time each person spent in jail for each booking. I also make some minor alterations to “charge” and “release_reason” variables to make them easier to work with later on.

```
jail_df <- jail_df %>% mutate(booking_dt = mdy_hms(booking_dt),
                              release_dt = mdy_hms(release_dt),
                              jail_diff = as.numeric(difftime(release_dt,
                                                              booking_dt,
                                                              units = "days")),
                              charge = tolower(charge),
                              release_reason = factor(tolower(release_reason)))
```

An important part of this analysis will look at repeat offenders, or individuals who cycle through jail. Because the inmate identities are known, it is possible to determine the number of times each inmate was booked during the time frame. To do this, I combine all the columns containing name information into one column.

```
# Remove NAs from the middle name and Jr/Sr columns
jail_df <- jail_df %>% mutate(middle_name = replace_na(middle_name, ""),
                              jrsr = replace_na(jrsr, "")) %>%
# Combine the name columns into one
                       unite(name_temp,
                         c("first_name", "middle_name", "jrsr"), sep = " ") %>%
                       unite(name, c("last_name", "name_temp"), sep = ", ") %>%
                       mutate(name = str_trim(name))
```

Then, by counting the number of times an individual has been booked, I determine whether an inmate has reactivated and therefore cycled back into the system. While this is far from a perfect measure of recidivism, it does provide a conservative estimate of the number of repeat offenders.

```
bookings <- jail_df %>% count(name, boa_number) %>%
                        group_by(name) %>%
                        summarize(bookings = n())

jail_df <- left_join(jail_df, bookings, by = "name")
```

I use a similar method to determine the total number of charges.

```
# Since each row represents one charge, counting by name
# will give the total # of charges someone has received
charges <- jail_df %>% count(name) %>%
                       rename(charges = n)
```

```r
jail_df <- left_join(jail_df, charges, by = "name")
```

Next, I create a new binary categorical variable that shows whether an inmate has been booked more than once.

```r
jail_df <- jail_df %>% mutate(offender_type = ifelse(bookings == 1, "First-Time", "Repeat"))
```

```r
bookings_df <- jail_df %>% count(boa_number, name, jail_diff, offender_type)

inmate_df <- jail_df %>% count(name, offender_type, bookings, charges) %>%
                          mutate(n = NULL)

inmate_df <- bookings_df %>% group_by(name) %>%
                             summarize(avg_time = mean(jail_diff, na.rm = TRUE)) %>%
                             right_join(inmate_df, by = "name")

inmate_df <- jail_df %>% count(name, boa_number, jail_diff) %>%
                         group_by(name) %>%
                         summarize(sum_time = sum(jail_diff, na.rm = TRUE)) %>%
                         ungroup() %>%
                         left_join(inmate_df, by = "name")
```

Missing Values

The following table shows the number of missing values, or NAs, in each column. While inspecting the first six rows of the data, we saw that there are several missing values, particularly for the variable "Current Facility."

boa_number	name	booking_dt	release_dt	current_facility	charge
0	0	0	4932	56988	0

court_case	court	rcw	release_reason	jail_diff	bookings	charges	offender_type
11545	5068	85	0	4932	0	0	0

Given that the Current Facility variable consists entirely of missing values it provides no information and can be removed from the data frame.

The rows with missing RCW entries have charge descriptions mostly containing "US Marshall hold" or some variant: these inmates are most likely state or federal prisoners being housed temporarily in local jails. This is supported by the fact that many of the release types are "transfer of custody." Most studies only remove these types of inmates when analyzing both jail and prison populations to avoid over-counting. Since that is not the case here, I keep them in.

release_reason	n
transfer of custody	67
not found	14
conditional/court release	3
case dismissed	1

name	charge	rcw	jail_diff	release_reason
TURNER, LEON	marshall hold	NA	2.778472	transfer of custody
HANSEN, RON ROCKWELL	marshall hold (fbi)	NA	1.780556	transfer of custody
BROOKS, JOSHUA DYLAN	marshall hold	NA	110.331944	transfer of custody
SHRECK, NICKOLAS JAY	federal hold	NA	132.795139	transfer of custody
HINOJOSA, DAVID	usm hold vufa	NA	123.075694	transfer of custody
PAZ-FAJARDO, VICTOR Y	marshal hold	NA	188.550694	transfer of custody

The remaining columns that have NAs, however, contain valuable insights into the data.

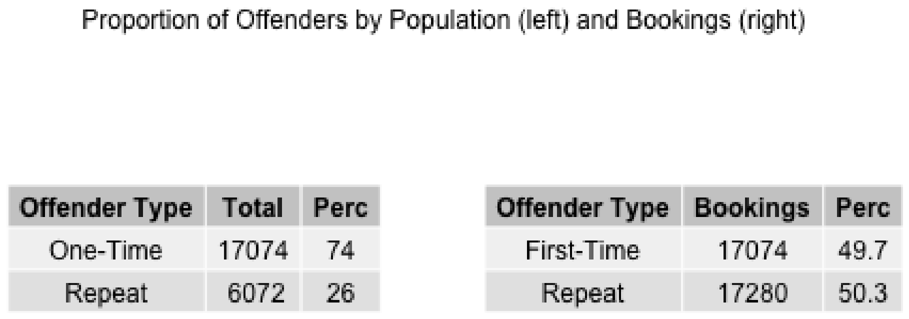
The release date/time column, which is central to this analysis, contain 4,932 while the booking column has none. This makes sense if the inmate was booked but has not been released from custody. There are several ways these missing values could be addressed. They could be removed from the data, replaced with an aggregate (e.g. the mean or median time), or the difference could be taken between the booking date and the last day in the data set. The solution that was employed in this study was to include them when calculating counts and proportion of the inmate population, and to remove them when calculating averages. Excluding inmates based on an arbitrary stopping point in the data seems unnecessary. On the other hand, using measures of central tendency to fill in the gaps or randomly sampling from the population is likely to lead to questionable results seeing as the data is extremely skewed.

The "Court" column contains 5,068 missing values, and "Court Case" contains 11,545. These columns won't be used for this analysis but could be used for an interesting project in the future.

Summary of Key Variables

There were a total of 23,146 individuals arrested and 34,354 bookings in King County from August 2018 to July 2019. In total, inmates spent 441,836 days in jail and were charged with 56,988 offenses.

First-time offenders–those who were booked only once during the time frame of the data set–make up 74% of the inmate population, with repeat offenders accounting for 26%. Repeat offenders, however, accounted for half of all bookings. Therefore, a disproportionate number of inmates who were arrested and booked have been incarcerated before.



Proportion of Offenders by Population (left) and Bookings (right)

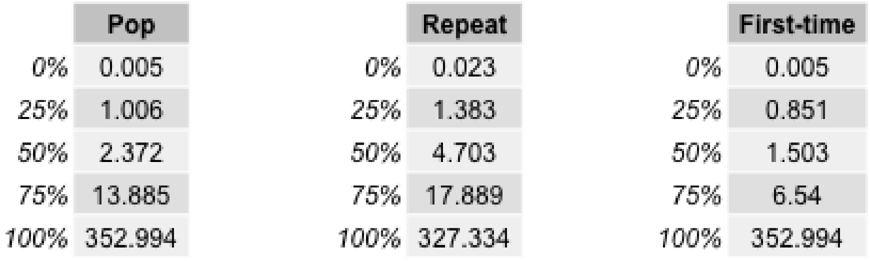
Offender Type	Total	Perc
One-Time	17074	74
Repeat	6072	26

Offender Type	Bookings	Perc
First-Time	17074	49.7
Repeat	17280	50.3

For each booking, inmates tended to be incarcerated for approximately 2.4 days. 50% were incarcerated between 1 and 14 days, a 12.8-day range.[5] The shortest amount of time spent in jail was 0.005 days, or 7 minutes, and the maximum was 353 days.

[5] I use the median and interquartile range (IQR) to determine central tendency and spread. I avoided using the arithmetic

Quartiles of Days in Jail by Population, Repeat, and First-time Offenders



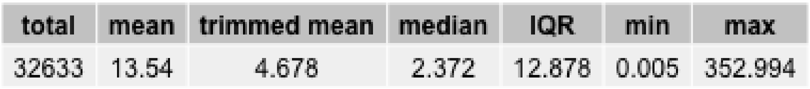
	Pop
0%	0.005
25%	1.006
50%	2.372
75%	13.885
100%	352.994

	Repeat
0%	0.023
25%	1.383
50%	4.703
75%	17.889
100%	327.334

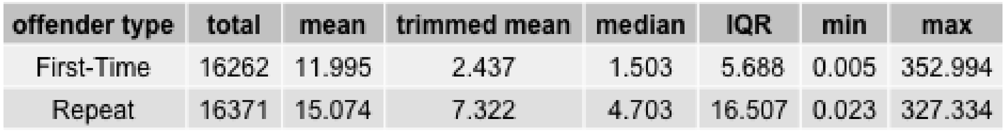
	First-time
0%	0.005
25%	0.851
50%	1.503
75%	6.54
100%	352.994

When aggregating based on offender type, first-time offenders tended to be incarcerated for 1.5 days while repeat offenders typically spent 4.8 days in jail.[6] The interquartile range (IQR) for these two groups is also quite disparate. First-time offenders had incarceration lengths from 0.8 to 6.7 days, a 5.9-day range. Repeat offenders had incarceration lengths from 1.3 to 18 days, a 16.6-day range. Repeat offenders tended to be incarcerated for 3.3 days longer than first-time offenders, and experienced a range of incarceration that was nearly three times as broad. This difference in incarceration length is tested for statistical significance in Part III.

Summary of Time in Jail by Population (top) and Offender Type (bottom)



total	mean	trimmed mean	median	IQR	min	max
32633	13.54	4.678	2.372	12.878	0.005	352.994



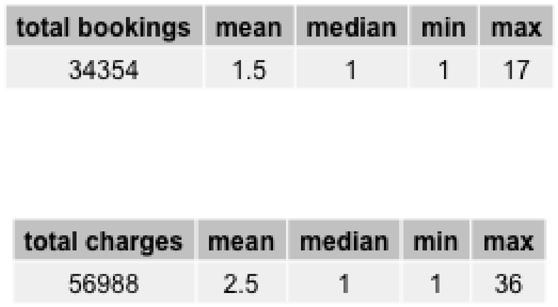
offender type	total	mean	trimmed mean	median	IQR	min	max
First-Time	16262	11.995	2.437	1.503	5.688	0.005	352.994
Repeat	16371	15.074	7.322	4.703	16.507	0.023	327.334

When looking at the total number of bookings and charges, inmates had an average of 2.4 charges with 11,147 inmates having more than one charge.

mean and standard deviation because the distribution of incarceration length is extremely positively skewed, and the mean is sensitive to extreme values. The median and IQR are far less sensitive to extreme values and give a more accurate representation of the typical length of time an inmate spent in King County jails.

[6] The measures of central tendency and spread of jail time do not include those inmates that were still in custody as of the last day of the data set (2019-05-31 23:55:00), which include 1721 bookings and 812 individuals.

Total Number of Charges and Bookings

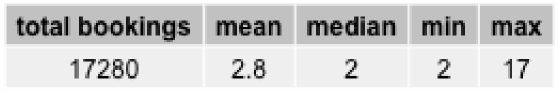


total bookings	mean	median	min	max
34354	1.5	1	1	17

total charges	mean	median	min	max
56988	2.5	1	1	36

In total, repeat offenders had 17,280 bookings, with 2.8 bookings on average.

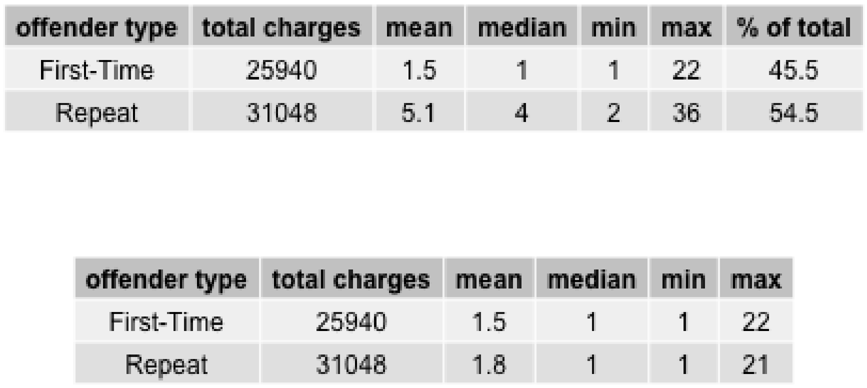
Repeat Offender Bookings



total bookings	mean	median	min	max
17280	2.8	2	2	17

When breaking down the total number of charges by offender type, repeaters make up 55% and have an average of 5.1 charges. When calculating the average number of charges per booking, repeat offenders have an average of 1.8 charges and first-timers have 1.5.

Total Charges (top) and Charges Per Booking (bottom) by Offender Type

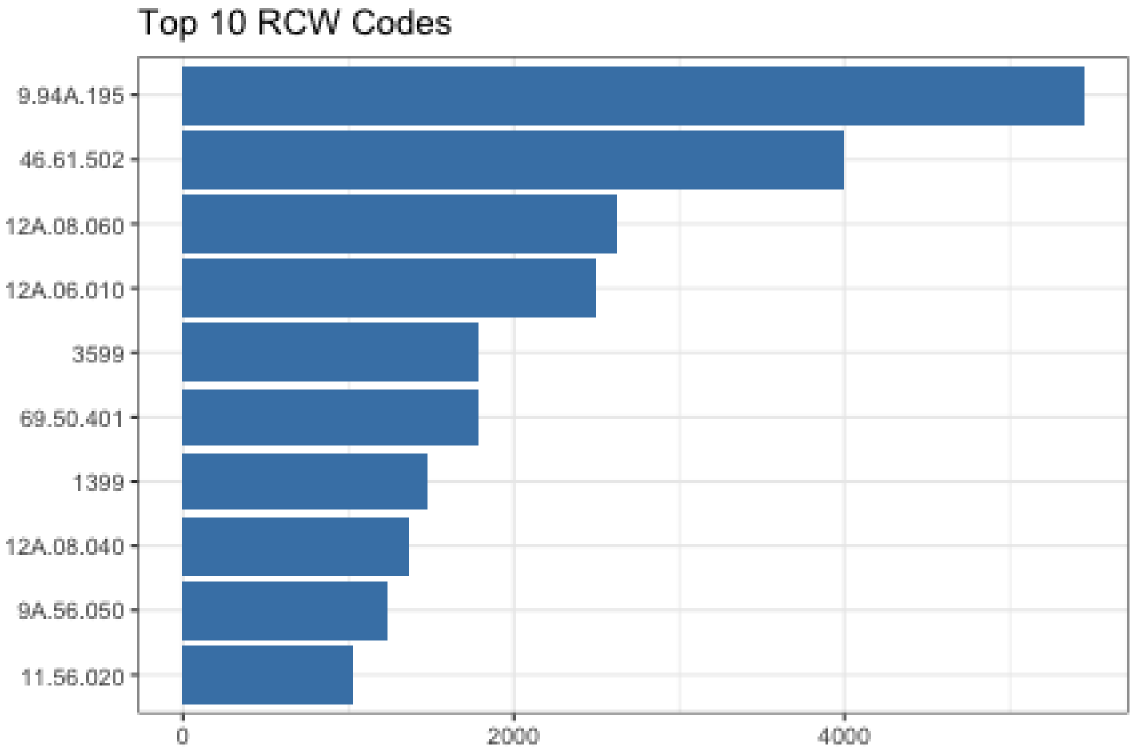


offender type	total charges	mean	median	min	max	% of total
First-Time	25940	1.5	1	1	22	45.5
Repeat	31048	5.1	4	2	36	54.5

offender type	total charges	mean	median	min	max
First-Time	25940	1.5	1	1	22
Repeat	31048	1.8	1	1	21

RCW Statutes

Nearly every inmate was charged with at least one Revised Code of Washington statute. There are 1,052 unique RCW codes in this data set. The top 10 most common codes are briefly described and listed in the graph below.[7] While RCW statutes give information about specific offenses, they are not easily sorted into more general categories of crime. In Part III, I use a method that attempts to capture these broader categories.



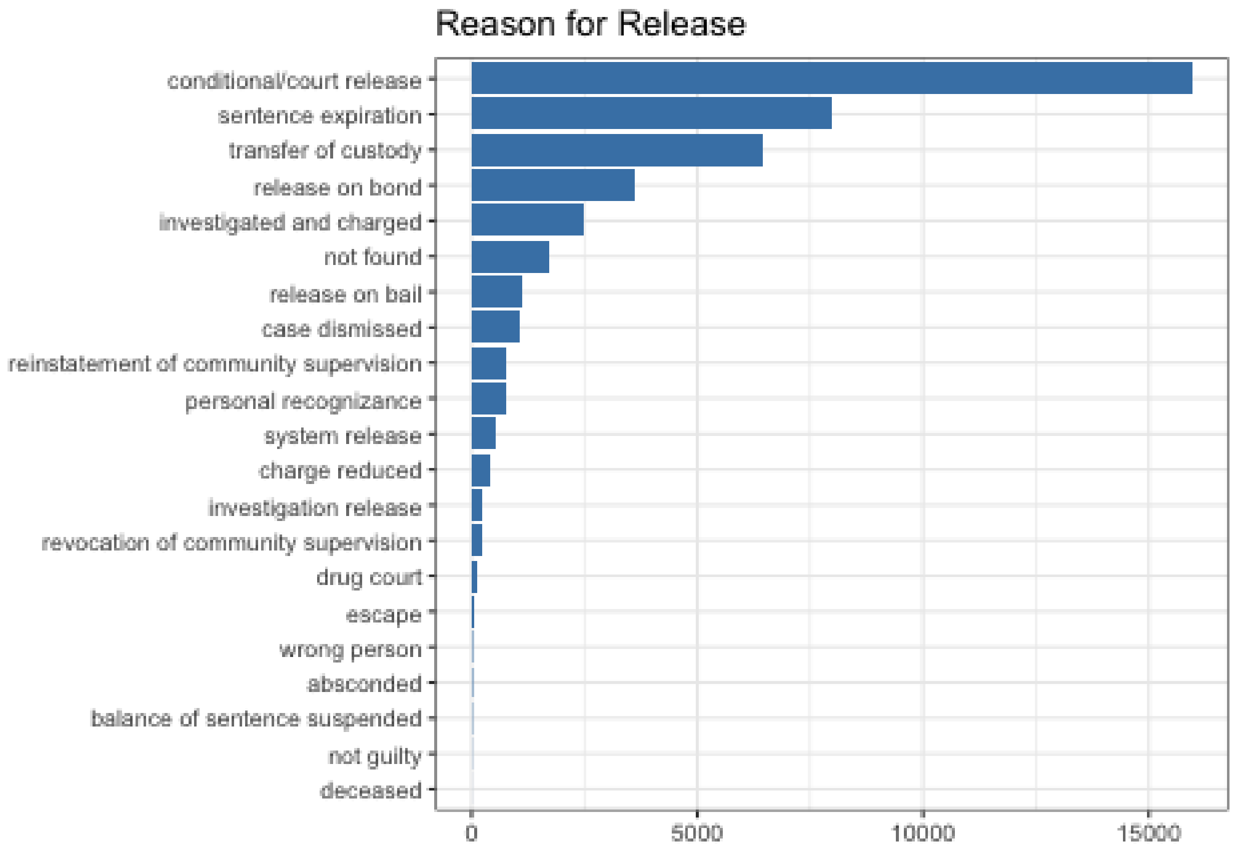
1. 9.94A.195: Probation violation
2. 46.61.502: Driving under the influence
3. 12A.08.060: Theft
4. 12A.06.010: Assault
5. 3599: Public intoxication
6. 69.50.401: Drug possession
7. 1399: Traffic-related offenses
8. 12A.08.040: Trespassing
9. 9A.56.050: Theft, 3rd degree (less than $750)
10. 11.56.020: DUI (Seattle Municipal Code)

Reason for Release

There are 21 different reasons an inmate may be released from custody. The most common is by conditional or court release, which is usually in the form of probation. 37% of all inmates were released in this way. The next three categories–sentence expiration, transfer of custody, and release on bond–make up 41% of the reasons inmates are released. The remaining 17 categories account for 22% of releases.

[7] Code descriptions provided from the Washington State Legislature's website: https://apps.leg.wa.gov/rcw/default.aspx accessed 09/14/2019.

Release Reason	Total	Percent
conditional/court release	15962	36.59
sentence expiration	7995	18.33
transfer of custody	6451	14.79
release on bond	3610	8.28
investigated and charged	2505	5.74
not found	1696	3.89
release on bail	1129	2.59
case dismissed	1043	2.39
reinstatement of community supervision	779	1.79
personal recognizance	754	1.73
system release	520	1.19
charge reduced	405	0.93
investigation release	237	0.54
revocation of community supervision	226	0.52
drug court	139	0.32
escape	89	0.20
absconded	25	0.06
wrong person	25	0.06
balance of sentence suspended	17	0.04
not guilty	9	0.02
deceased	3	0.01

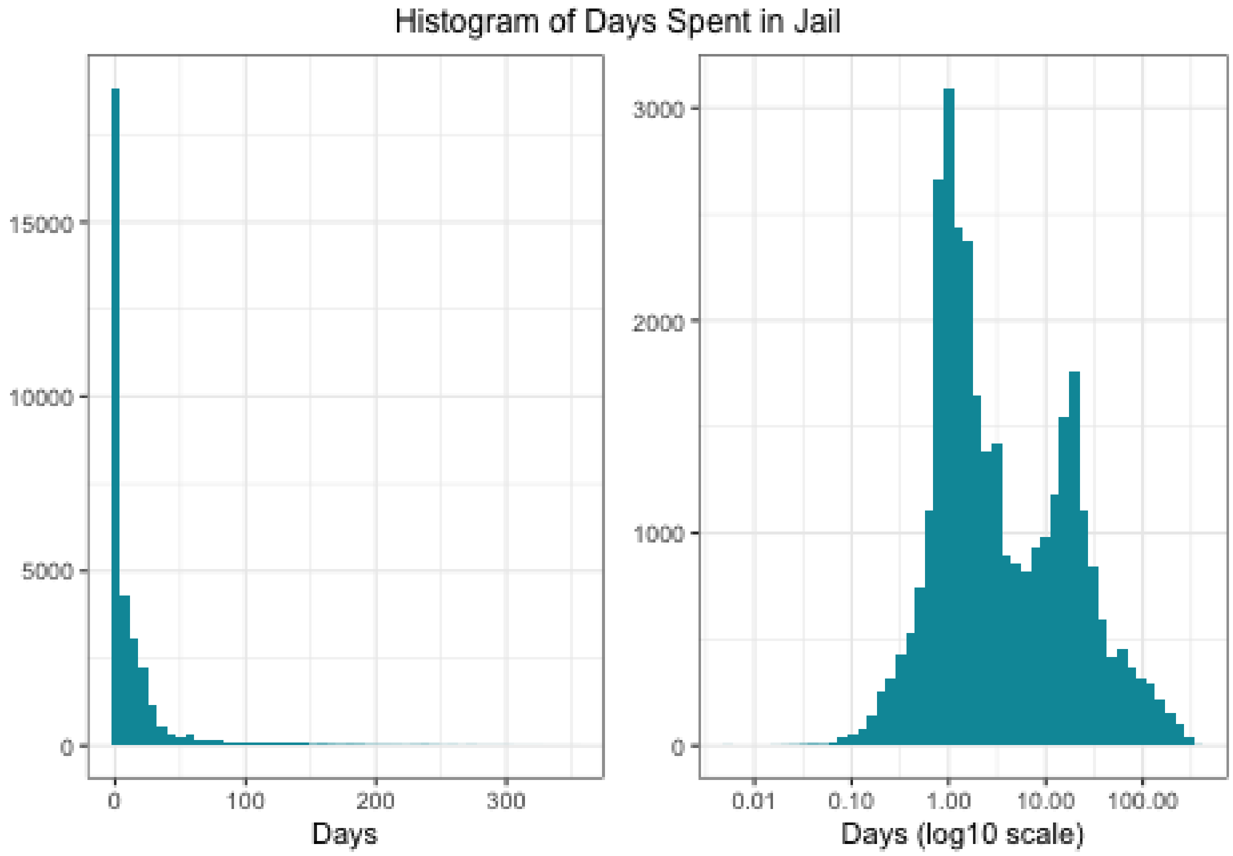


Distributions

Jail Time

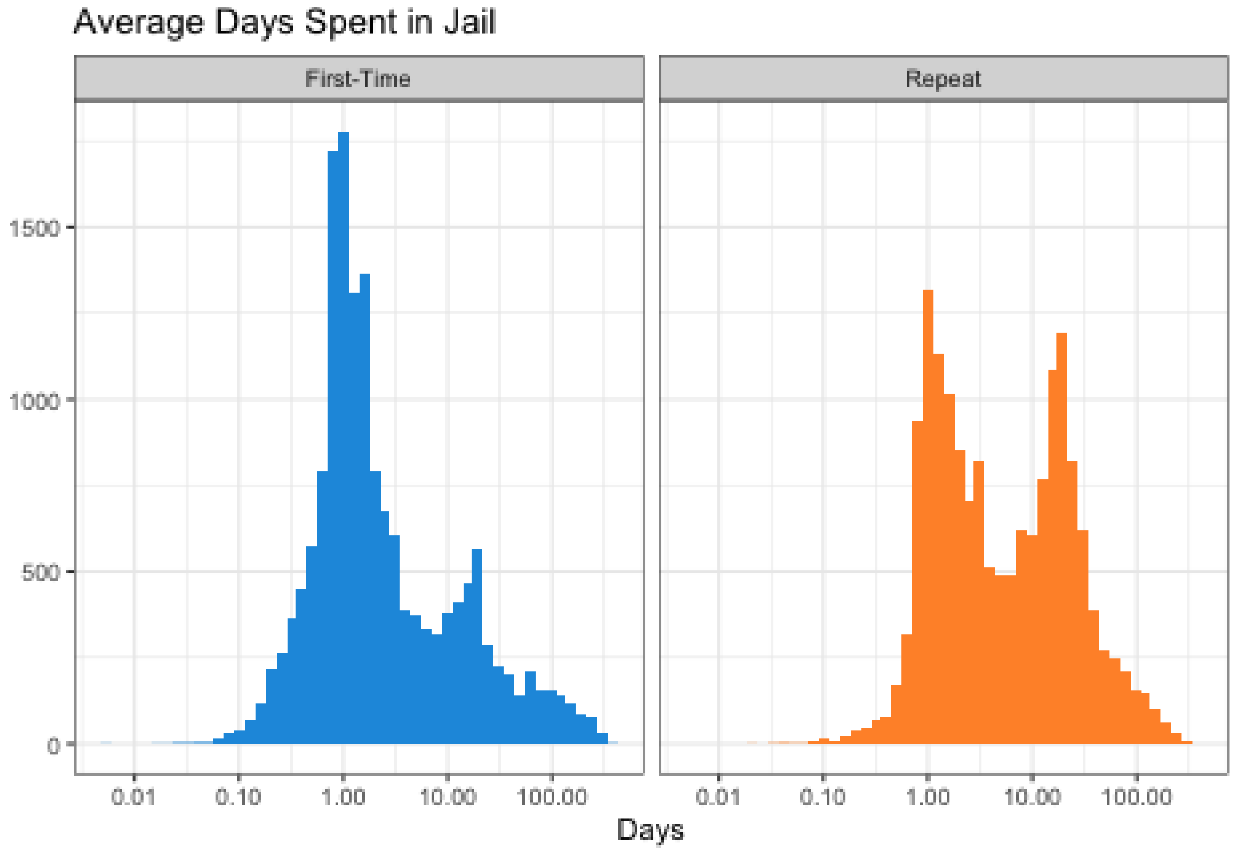
How much time are inmates incarcerated for in King County? As shown in the histogram below, the distribution of jail time is positively skewed. This means that most observations are condensed within a short range, which creates a “peak,” while increasingly extreme values stretch out in a long, thin “tail.” This makes it difficult to interpret variation within the packed areas of the distribution. By using a log10 transformation, the variation becomes clearer. The distribution is bimodal, meaning there are two peaks where values tend to cluster. The first peak occurs around 1-3 days and the second peak at approximately 14-18 days.

```r
p1 <- bookings_df %>% ggplot(aes(jail_diff)) +
                             geom_histogram(fill = "#0097A7", bins = 50) +
                             labs(y = NULL,
                                  x = "Days") +
                             theme_bw()

p2 <- bookings_df %>% ggplot(aes(jail_diff)) +
                             geom_histogram(fill = "#0097A7", bins = 50) +
                             labs(y = NULL,
                                  x = "Days (log10 scale)") +
                             scale_x_continuous(trans = "log10",
                                                labels = comma_format(accuracy = .01)) +
                             theme_bw()

grid.arrange(p1, p2, nrow = 1, top = "Histogram of Days Spent in Jail")
```



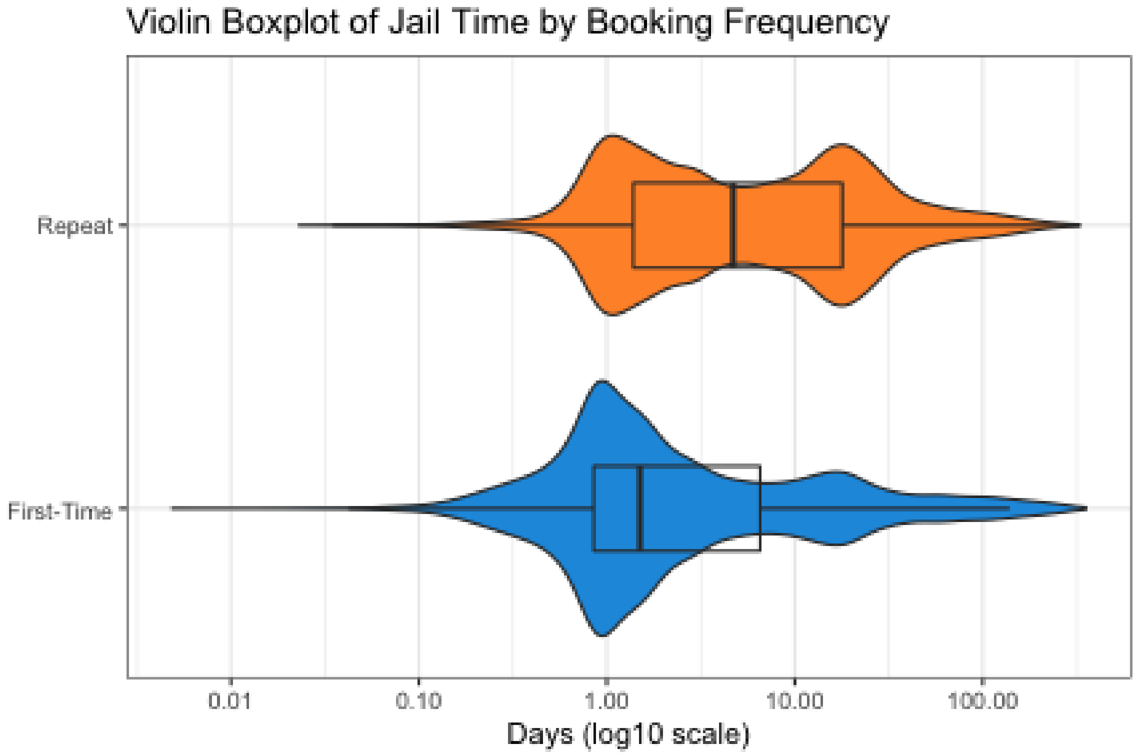
When the distribution is split between first-time and repeat offenders, an interesting shift occurs. While both distributions remain bimodal, their peaks vary in height.

```
palette3 <- c("#1e9adf", "#ff9333")

bookings_df %>% ggplot(aes(jail_diff, fill = offender_type)) +
                      geom_histogram(bins = 50) +
                      facet_wrap(~offender_type) +
                      scale_x_continuous(trans = "log10",
                                         labels = comma_format(accuracy = .01)) +
                      labs(title = "Average Days Spent in Jail",
                           y = NULL,
                           x = "Days") +
                      scale_fill_manual(values = palette3) +
                      theme_bw() +
                      theme(legend.position = "none")
```



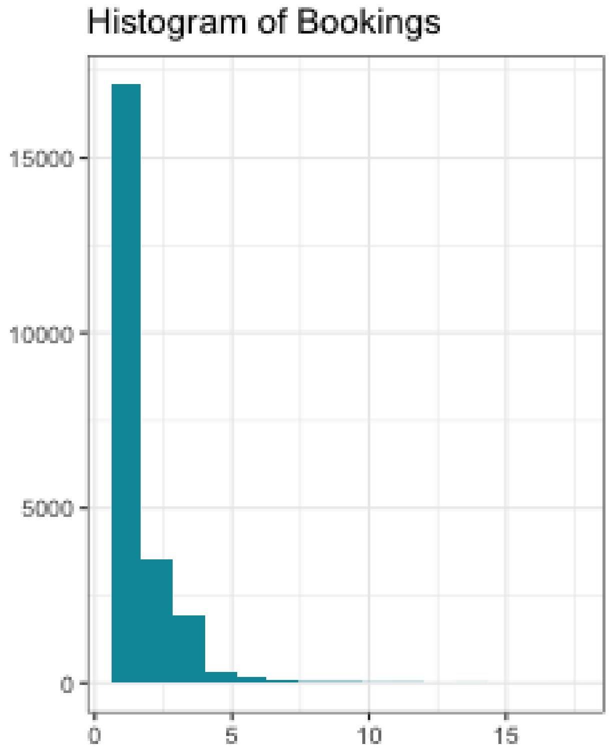
The distribution for first-time offenders becomes nearly unimodal while the peaks for the repeat-offender distribution become more balanced. The interquartile range (IQR) for first-time offenders is shorter with half of the observations occurring within a six-day range, while the IQR for repeat-offenders is spread across a seventeen-day range. The violin boxplot below shows both the shape and location of the IQR for each distribution.

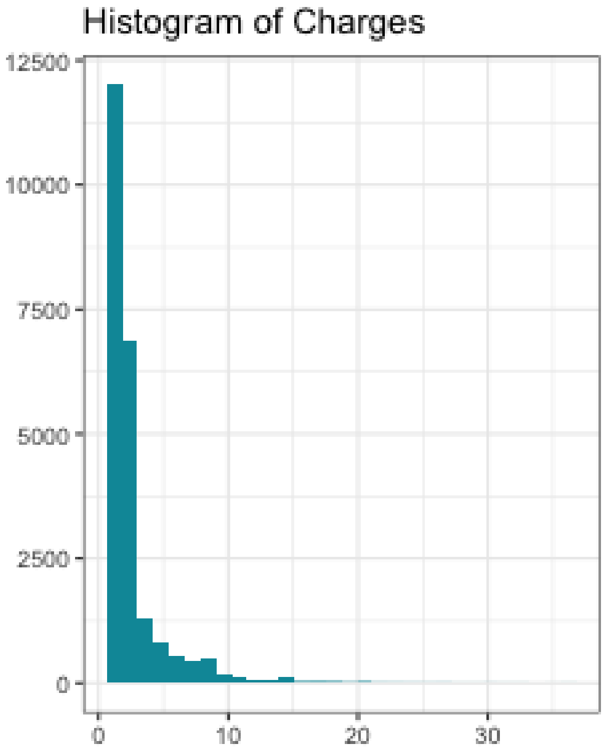
```
bookings_df %>% ggplot(aes(offender_type, jail_diff, fill = offender_type)) +
                       geom_violin() +
                       geom_boxplot(alpha = 0, width = .3) +
                       labs(title = "Violin Boxplot of Jail Time by Booking Frequency",
                             x = NULL,
                             y = "Days (log10 scale)") +
                       coord_flip() +
                       scale_y_continuous(trans = "log10",
                                          labels = comma_format(accuracy = .01)) +
                       scale_fill_manual(values = palette3) +
                       theme_bw() +
                       theme(legend.position = "none")
```



Bookings and Charges

As with the jail time distribution, the histograms of both bookings frequency and the number of charges skew right.





Frequency of Offense Type

Which types of crimes are most common and how do they vary by offender type? While the RCW codes provide specific information about the type of crime committed, they are difficult to sort into broad categories of offenses, such as violent, property, and drug crimes. In order to do this, the jail data would need to be merged with an existing RCW data set that has pre-determined crime-type by statute—which is not publicly available—or hand-code each RCW statute, which would be time-consuming and prohibitive.

One alternative is to perform a text analysis of the charge description for each inmate, which consists of a short string of words that can be tokenized (i.e. broken down into individual terms) and counted. These words can then be sorted into different categories using a dictionary-based method of categorization.[8]

Text Analysis of Charges

To perform the text analysis, I create a "tidy" data frame, which contains one token per row, and a Document Feature Matrix (DFM).

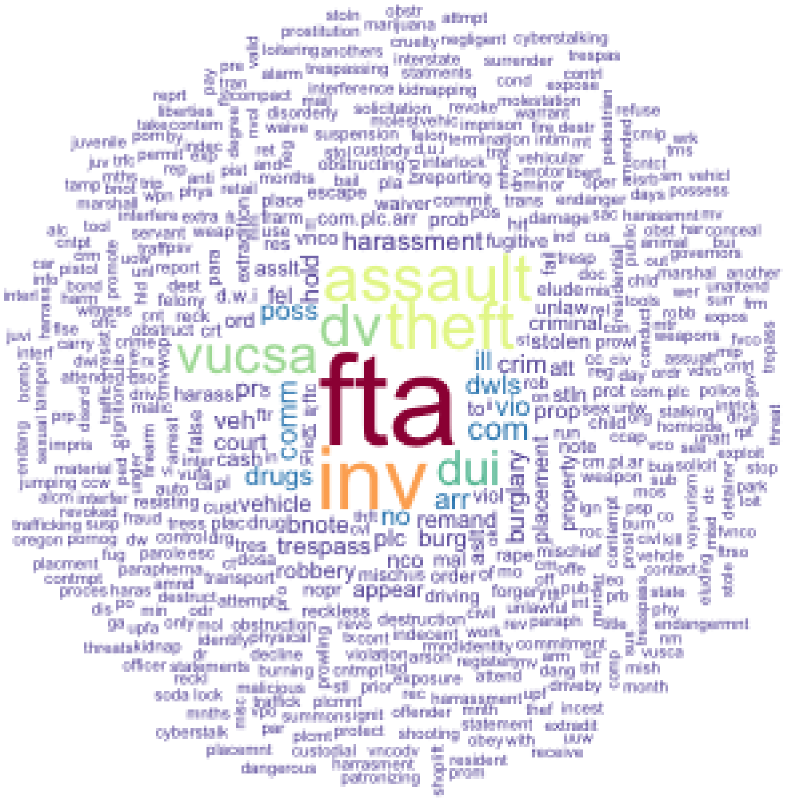
```r
# Tidy Approach
jail_tidy <- jail_df %>% unnest_tokens(word, charge, token = "words") %>%
                         filter(nchar(word) > 1 & !str_detect(word, "[[:digit:]]"))

# Corpus and DFM
jail_corp <- corpus(jail_df, text_field = "charge")
```

[8] Another option would have been to use an unsupervised k-means topic model to sort charges into "topics," but because the length of each text is so short, sparsity became an issue. These topics were mixed with several offense types and were not useful. Using a bi-term topic model, a method that attempts to circumvent this issue, also resulted in mixed topics.

```
jail_dfm <- dfm(jail_corp, remove_punct = TRUE, remove_numbers = TRUE,
                stem = TRUE, remove = stopwords("english"))
```

The word-cloud below gives an effective visualization of the range and frequency of terms. It shows that there is a considerable amount of variation in how charges are spelled and abbreviated. It also shows that there are about six terms that occur the most frequently: fta, inv, assault, theft, vucsa, dv, and dui, respectively. These terms will be defined in the next section. The variation in abbreviation is important because it obscures the exact frequency of a charge and may lead to mis-categorization. Also, there is no way to know for certain what each term means. Many terms can be determined through context but others are more difficult to interpret.[9] Therefore, the results of the text analysis should be interpreted as an approximation rather than an exact representation of the terms in the data set.

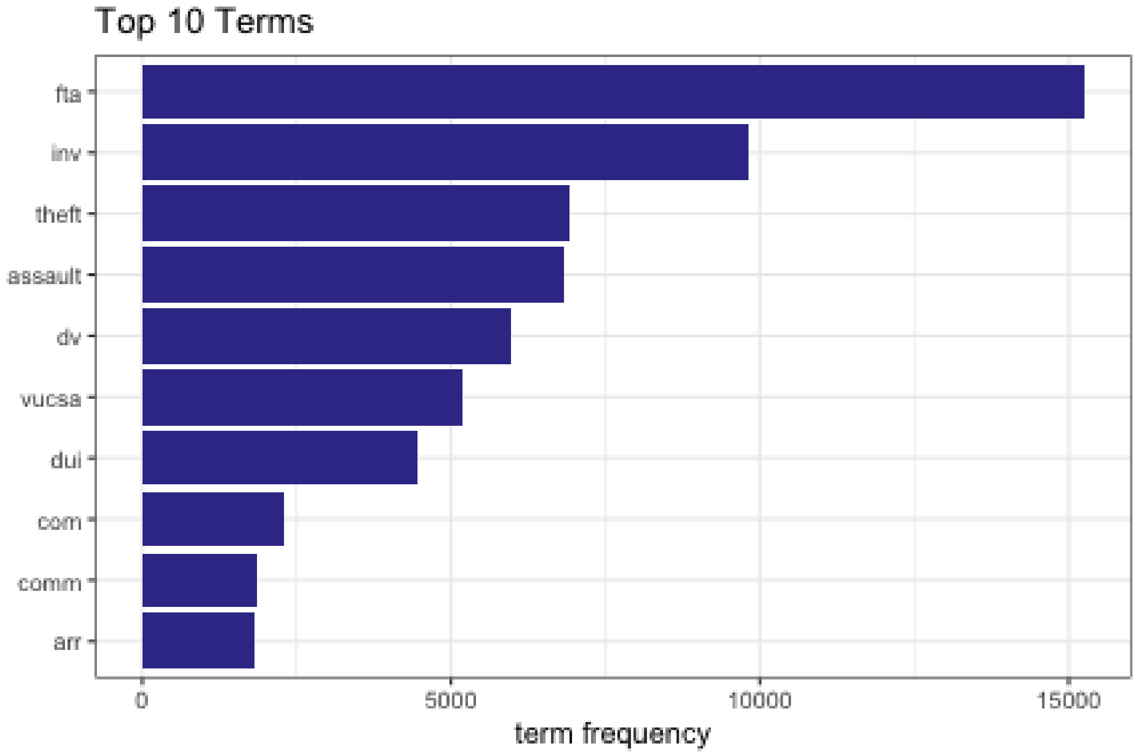
```
jail_tidy %>% count(word) %>%
              with(wordcloud(word, n,
                             min.freq = 10,
                             max.words = 500,
                             random.order = FALSE,
                             color = rev(RColorBrewer::brewer.pal(10, "Spectral"))))
```



The ten most common charge terms and their frequencies are shown below. "fta" and "inv" ("failure to appear" and "investigation") occur far more often than any other term. A "failure to appear" charge typically occurs when an offender fails to show in court, resulting in the judge issuing a bench warrant for

[9] I visited many legal forums where people asked attorneys to describe what a specific term, such as "inv" or "arr", meant. There was no consensus on the meaning of several terms.

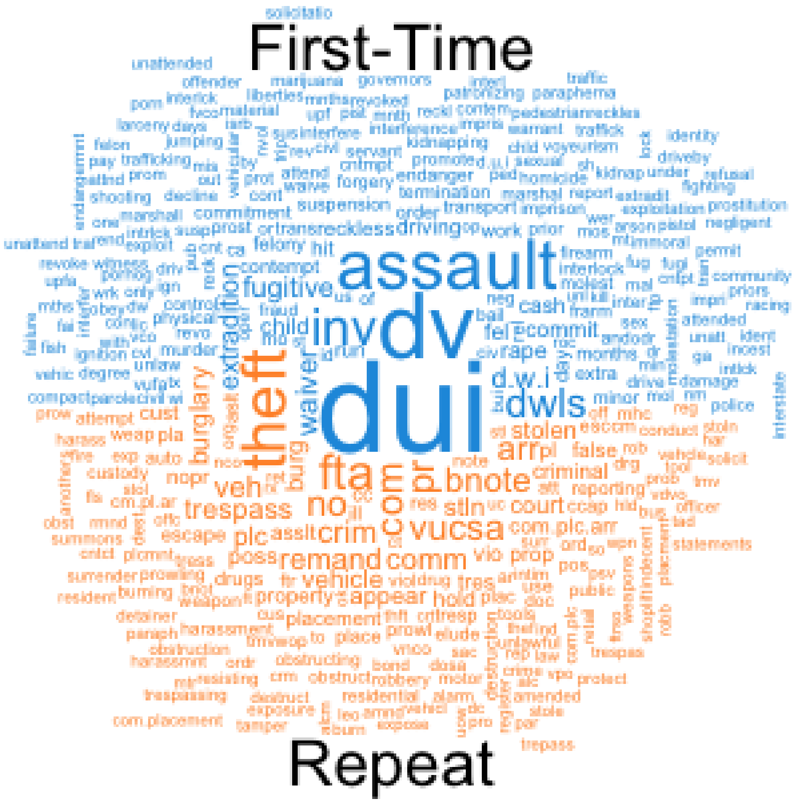
their arrest. These may be issued for a range of crimes, the most common of which tend to be traffic citations. The term “inv” is likely a generic term that means that an inmate was arrested during an investigation of a serious crime, like robbery or burglary, but has not been officially charged yet. The next four terms describe specific offenses: theft, assault, domestic violence, drug possession, and driving under the influence. The following two terms are variations of community placement, and most likely refer to a violation of probation. Lastly, “arr” likely means arraignment or arrest.



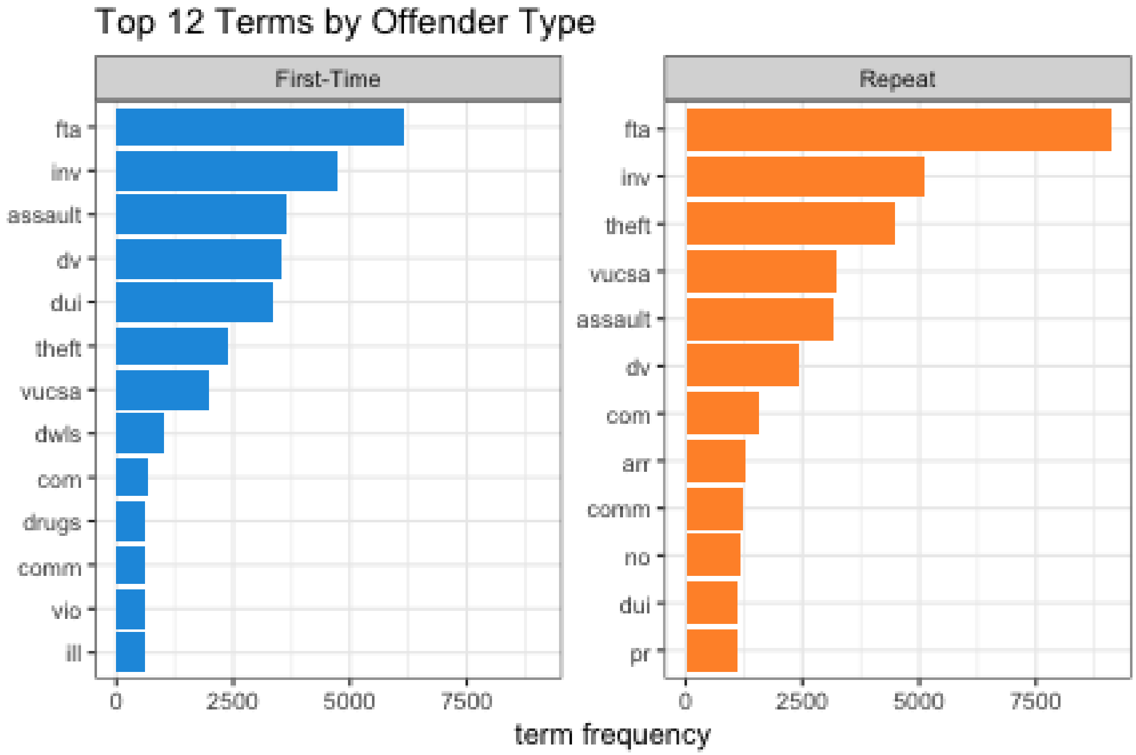
Offense Frequency by Offender Type

The following comparison cloud shows how charges vary by offender type. As with the previous word-cloud, several familiar offense types are present, however, they are distributed differently between offender types. DUIs, traffic offenses (e.g. “dwls” or “driving with a suspended license”), and domestic violence tend to be associated with first-time offenders, while drug and theft violations are more often associated with repeat offenders.

```
jail_tidy %>% count(word, offender_type) %>%
              acast(word ~ offender_type, value.var = "n", fill = 0) %>%
              comparison.cloud(random.order = FALSE,
                               colors = palette3,
                               max.words = 400,
                               title.size = 2,
                               title.bg.colors = "white")
```



The graph below shows in greater detail the frequency of the top 12 charge terms for first-time and repeat offenders. Both have "fta" and "inv" as the top two terms, though repeat offenders have a larger share. This make sense because repeat offenders are more likely to have court dates than first-time offenders, and therefore they have more opportunities to miss them.



While these terms provide insight into which crimes tend to occur, they do not give an accurate summary of general offenses. By using a dictionary method, terms can be sorted into general crime categories and their frequencies estimated. The advantage to this approach is that specific terms of interest can be targeted and grouped together while “filler” words can be ignored.

However, this method is sensitive to spelling. While this concern can be mitigated using regular expressions (regex), it can’t be eliminated entirely. Also, there is the possibility of over- and under-counting terms. If two or more terms assigned to a category appear in a text, then the corresponding offense category may be overrepresented for that booking. Conversely, if a term has not been included in the dictionary, it won’t be counted. Also, there is the problem of the sheer volume of charges, of which there are 2,123 unique terms. It is unrealistic to include all of them in the dictionary. Lastly, there is the issue of deciding which terms to include. Some crimes are easy to sort, while others can be harder to place and require the use of judgement. To address this, I attempted to use a relatively short and simple dictionary list that focuses on the most common offenses. Therefore, this approach should be considered a rough estimate, rather than a true representation, of offense frequency.

Below is a dictionary containing the key terms for eight common offense categories: property crime, violent crime, sexual assault, DUI, drug offenses, probation violation, traffic offenses, and failure to appear.

```
dict <- dictionary(list(property_crime = c("thef", "thft", "forg", "burg", "brg",
                                           "stol", "stl", "shoplift", "tmvwop"),
                        violent_crime = c("dv", "assault", "asslt", "aslt", "murd",
                                          "rob", "arson", "burn", "hom", "intim"),
                        sexual_assault = c("mol", "rape", "inde", "incest"),
                        dui = c("dui", "d.u.i", "d.w.i", "dwi"),
                        drug = c("vucsa", "drug", "mari", "drg"),
                        probation_parole = c("paro", "com", "cc"),
                        traffic_offense = c("dwl"),
                        fta = c("fta")))
```

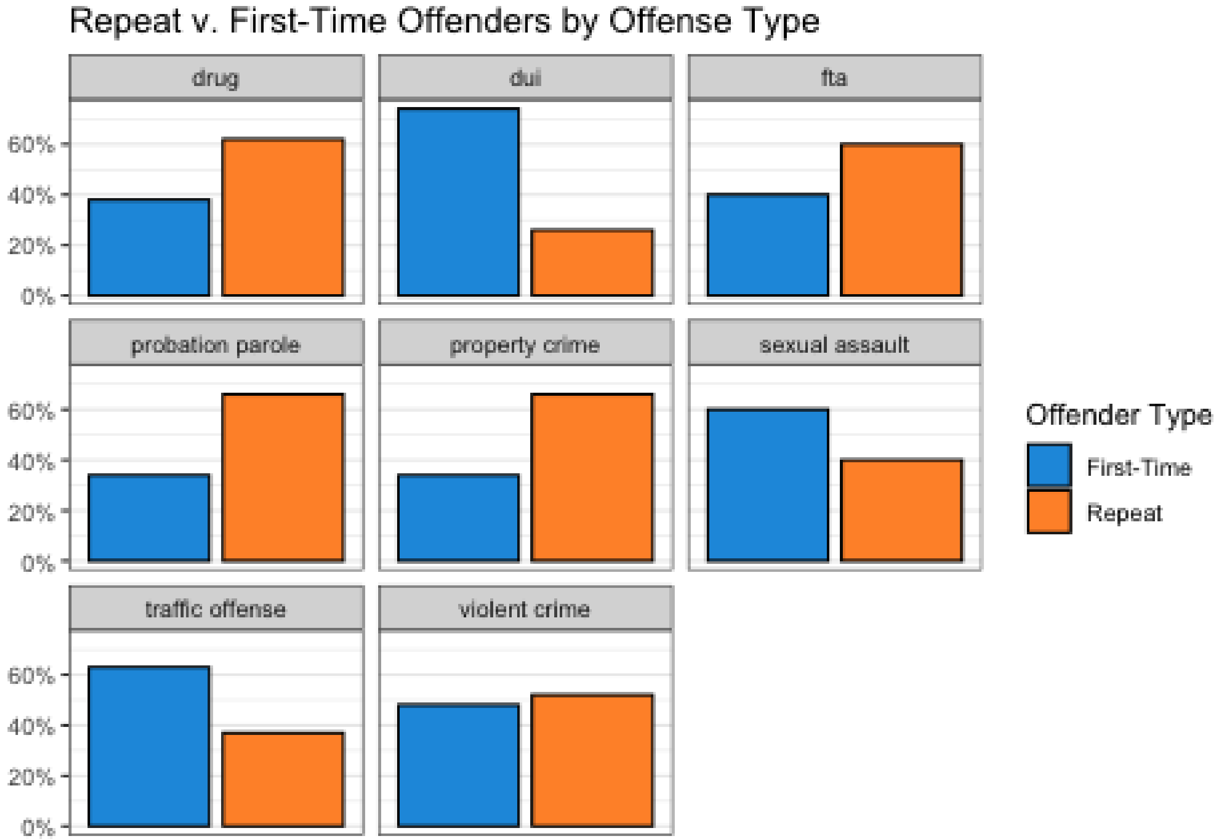
```
dict_df <- jail_dfm %>%
           dfm_lookup(dictionary = dict, valuetype = "regex") %>%
           convert(to = "data.frame")

jail_df <- jail_df %>% bind_cols(dict_df)

jail_df$document <- NULL
```

The table below shows the number and percentage of offenses for each category, and the percent share committed by offender type. Violent crime and failure to appear each make up a quarter of all offenses, with property crime close behind at 19%. Drug crime, probation violations, and DUIs combined account for approximately 27% of offenses, and sexual assault and traffic offenses each make up 1.2% and 2.5% of bookings.

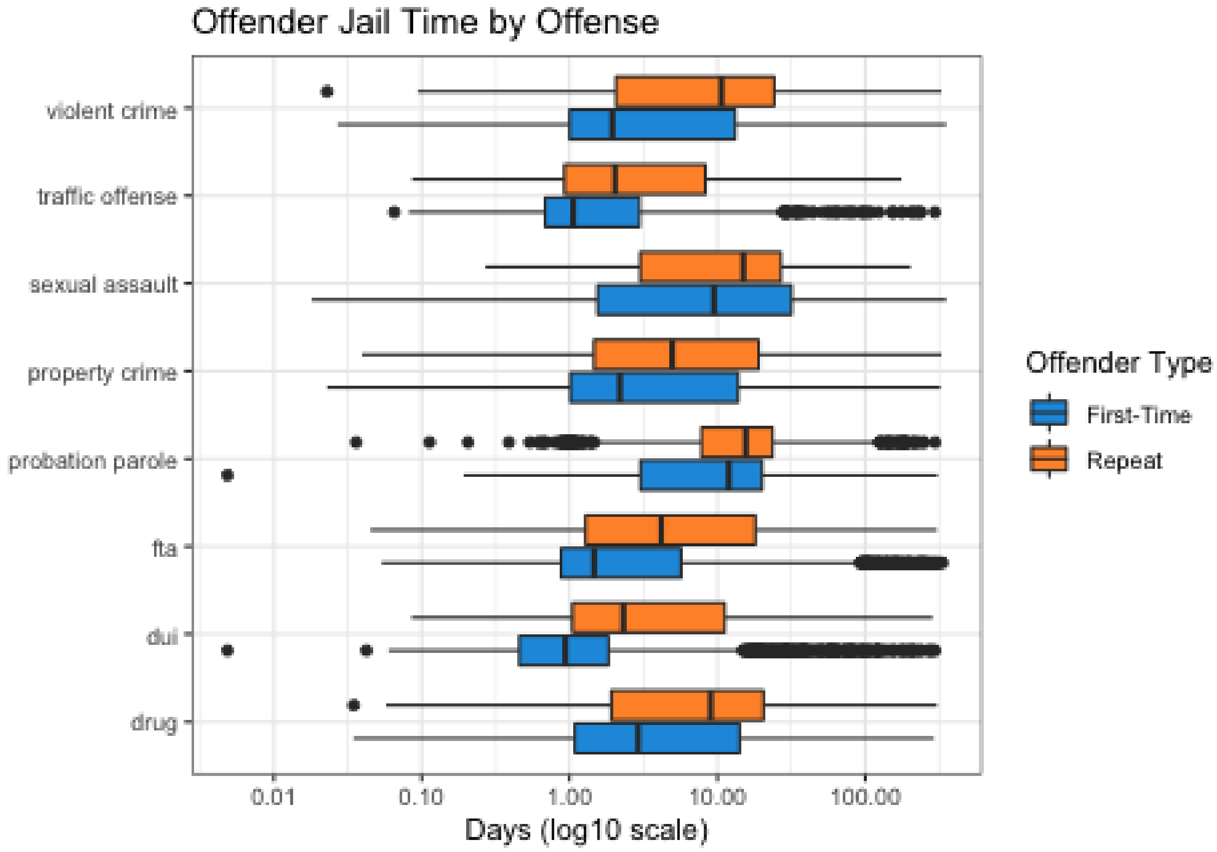
Repeat offenders disproportionately commit property and drug crime by 20% and 30%, respectively. Repeaters are also 1.7 times more likely to violate probation and 1.4 times more likely to fail to appear in court. First-time offenders, on the other hand, are nearly 3 times more likely to be booked for DUI and commit 28% more serious traffic offenses than repeat offenders.

Offense	First-Time	Repeat	Total	% of Total
violent crime	52%	48%	17098	26.2%
fta	40%	60%	15266	23.4%
property crime	34%	66%	12042	18.4%
drug	39%	61%	7507	11.5%
probation parole	34%	66%	5848	9%
dui	74%	26%	5127	7.9%
traffic offense	63%	37%	1621	2.5%
sexual assault	60%	40%	775	1.2%

```
offense_df <- jail_df %>% select(boa_number, name, jail_diff,
                                 offender_type, bookings, booking_dt,
                                 property_crime:fta) %>%
                          gather("offense", "value", property_crime:fta) %>%
                          filter(value > 0) %>%
                          mutate(offense = str_replace(offense, "_", " "))

offense_df %>% count(offense, offender_type) %>%
               group_by(offense) %>%
               mutate(total = sum(n)) %>%
               mutate(perc = round(n / total, 2)) %>%
               ggplot(aes(offender_type, perc, fill = offender_type)) +
                      geom_col(position = "dodge", color = "black") +
                      facet_wrap(~offense) +
                      scale_y_continuous(labels = percent_format(accuracy = 1)) +
                      labs(title = "Repeat v. First-Time Offenders by Offense Type",
                           x = NULL,
                           y = NULL) +
                      scale_fill_manual(name = "Offender Type",
                                        values = palette3) +
                      theme_bw() +
                      theme(panel.grid.major.x = element_blank(),
                            axis.text.x = element_blank(),
                            axis.ticks.x = element_blank())
```



When looking at the IQRs of each offense category by offender type, it is clear that the median time served in jail is consistently higher for repeat offenders than for first-timers, as shown in the boxplot below.



Difference in Median Incarceration Length

To test whether the observed difference in median jail time between offenders is statistically significant (i.e. not due to random chance) I ran one two-sample, two-sided Wilcoxon Rank Sum test for the entire inmate population and one for each offense type.[10] In general, repeat offenders spent an estimated 2.2 more days in jail than first time offenders (95% confidence interval between 2.06 and 2.33 days at the 0.05 significance level).

```
# convert offense_df to long format
offense_long <- jail_df %>%
                select(boa_number, jail_diff, offender_type, property_crime:fta) %>%
                gather("offense", "value", property_crime:fta) %>%
                filter(value > 0, !is.na(jail_diff)) %>%
                count(boa_number, jail_diff, offender_type, offense) %>%
                mutate(offense = str_replace(offense, "_", " "))

# Wilcoxon Rank Sum test
tidy(wilcox.test(jail_diff ~ fct_rev(offender_type),
            data = offense_long,
            alternative = "two.sided",
            mu = 0,
            paired = FALSE,
```

[10] I've included a normal t-test and log10 t-test in the Appendix, along with their point range plots. The Wilcoxon test was chosen over the t-test due to the extreme skew of the distribution, which violate assumptions of normality. A log10 transformation could not correct for the skewness. I've also included several qq-plots that illustrate this.

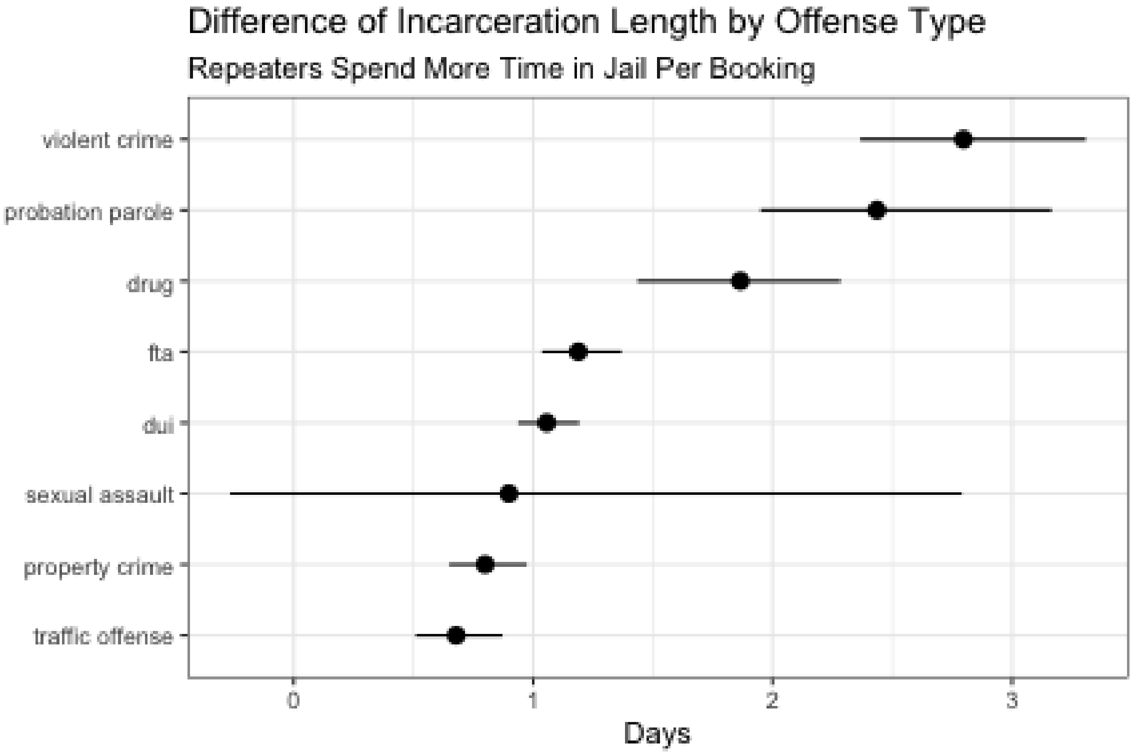
```
             conf.int = TRUE,
             conf.level = .95)) %>%
     mutate_at(vars(estimate, statistic, conf.low, conf.high), funs(round(., 2))) %>%
     mutate(p.value = scientific(p.value, digits = 2),
            statistic = scientific(statistic, digits = 2)) %>%
     select(estimate:conf.high) %>%
     kable() %>%
     kable_styling(bootstrap_options = c("bordered", "condensed"),
                   full_width = FALSE)
```

estimate	statistic	p.value	conf.low	conf.high
2.09	3.3e+08	0e+00	1.97	2.21

When looking at the difference in incarceration length by offense type, we also see that repeat offenders consistently serve more time, but that the magnitude and certainty of the estimates vary. All estimates are statistically significant, with the exception of sexual assault. The test shows that repeaters spend as much as 1 day longer in jail for sexual assault offenses, but the 95% confidence interval ranges from -0.26 to 2.79 days. While this could indicate that little difference exists between the treatment of repeat and first-time offenders in regard to sexual assault, it is more likely that the terms used to define the sexual assault dictionary are picking up less serious, or possibly unrelated, offenses which are thereby obscuring the results. The remaining offense types have considerably narrower confidence intervals that range from 1.4 to 0.35 days. Repeat offenders that commit violent crime tend to spend 3.2 more days in jail, and spend 2.7 and 1.7 more days for probation violation and drug offenses.

```
# nested df of Wilcoxon models
wilcox_nest <- offense_long %>% mutate(offender_type = fct_rev(offender_type)) %>%
                                nest(-offense) %>%
                                mutate(model =
                                        map(data, ~ wilcox.test(jail_diff ~ offender_type,
                                                                data = .,
                                                                alternative = "two.sided",
                                                                mu = 0,
                                                                paired = FALSE,
                                                                conf.int = TRUE,
                                                                conf.level = .95)))

wilcox_models <- bind_rows(lapply(wilcox_nest$model, tidy), .id = "offense")

wilcox_models <- wilcox_models %>% mutate(offense = wilcox_nest$offense)

# Wilcoxon table
wilcox_models %>% select(offense:conf.high) %>%
                  mutate_at(vars(estimate, statistic, conf.low, conf.high),
                            funs(round(., 2))) %>%
                  mutate(p.value = scientific(p.value, digits = 2),
                         statistic = scientific(statistic, digits = 2)) %>%
                  arrange(-estimate) %>%
                  kable() %>%
                  kable_styling(bootstrap_options = c("striped", "bordered", "condensed"),
                                full_width = FALSE)
```

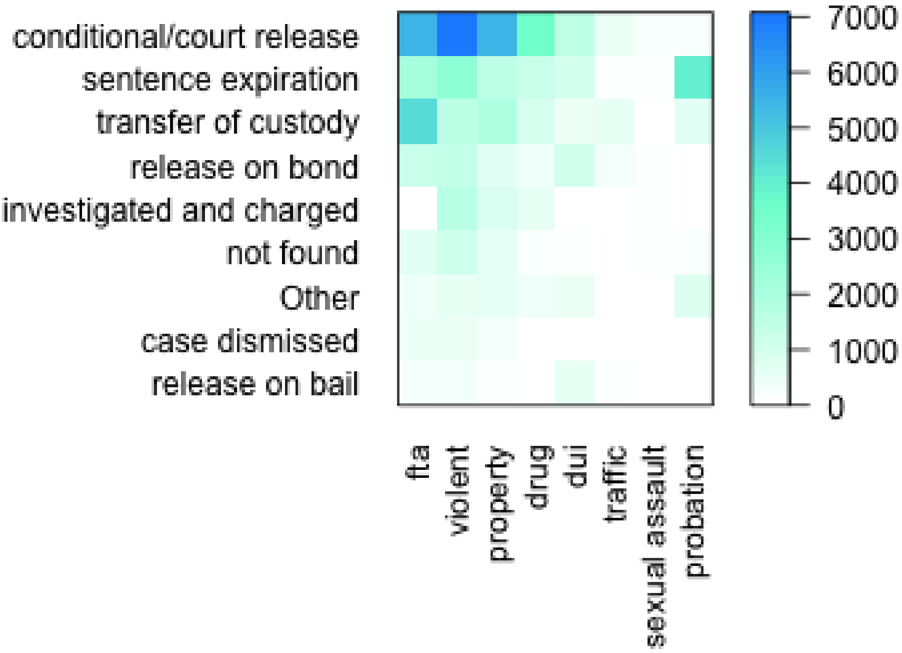
offense	estimate	statistic	p.value	conf.low	conf.high
violent crime	2.80	1.4e+07	9.0e-138	2.37	3.31
probation parole	2.44	3.6e+06	3.0e-20	1.95	3.16
drug	1.87	3.5e+06	9.5e-42	1.44	2.28
fta	1.19	1.8e+07	5.9e-139	1.04	1.37
dui	1.06	2.9e+06	7.4e-105	0.94	1.19
sexual assault	0.90	3.1e+04	1.6e-01	-0.26	2.79
property crime	0.80	8.7e+06	2.2e-37	0.65	0.97
traffic offense	0.68	3.0e+05	4.7e-17	0.51	0.87



Release Reasons, Offense, and Offender Type

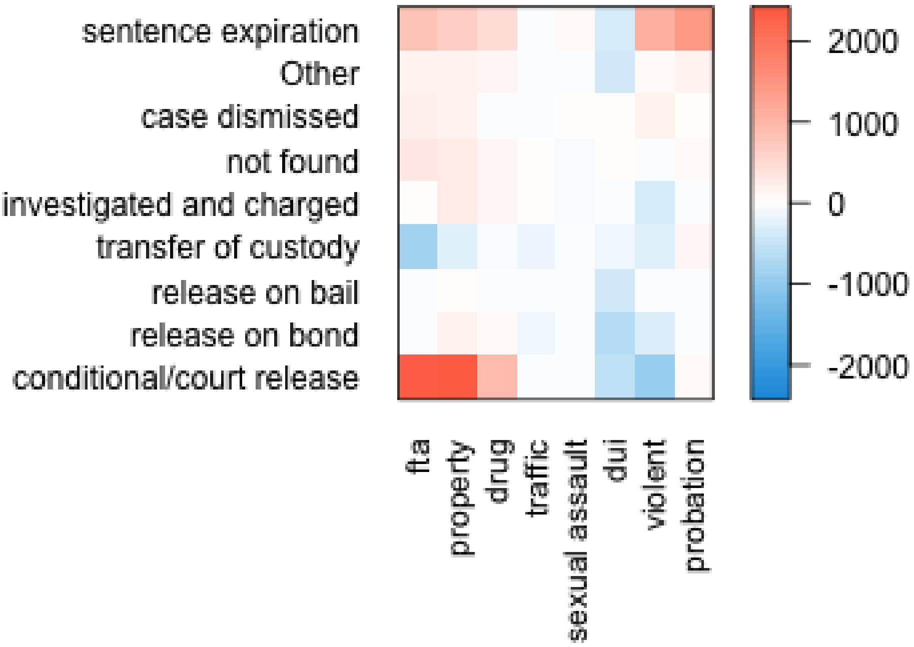
How do release reasons vary by offense type? In the heat map below, the vertical axis shows the release reason and the offense committed on the horizontal axis. The color of each cell corresponds to the number of cases, with white being low and dark blue being high. The map clusters similar groups based on frequency, demonstrating that most releases (in this case conditional court releases) are related to DUIs, FTAs, drug, property, and violent crime. The release types with the fewest cases include the investigated and charged or released on bail for probation, traffic, and DUI offenses.

```
release_df <- jail_df %>% mutate(release_reason = fct_lump(release_reason, n = 8)) %>%
                          group_by(release_reason) %>%
                          summarize_at(vars(property_crime:fta),
                                       funs(sum)) %>%
                          rename(`release reason` = release_reason,
                                 property = property_crime,
                                 violent = violent_crime,
                                 `sexual assault` = sexual_assault,
                                         probation = probation_parole,
                                         traffic = traffic_offense)

release_matrix <- as.matrix(release_df[2:9], dimnames = list(names(release_df)[2:9]))

rownames(release_matrix) <- as.vector(release_df$`release reason`)

hmap(release_matrix, method = "TSP",
     col = colorRampPalette(c("white", "aquamarine", "dodgerblue")) (100))
```



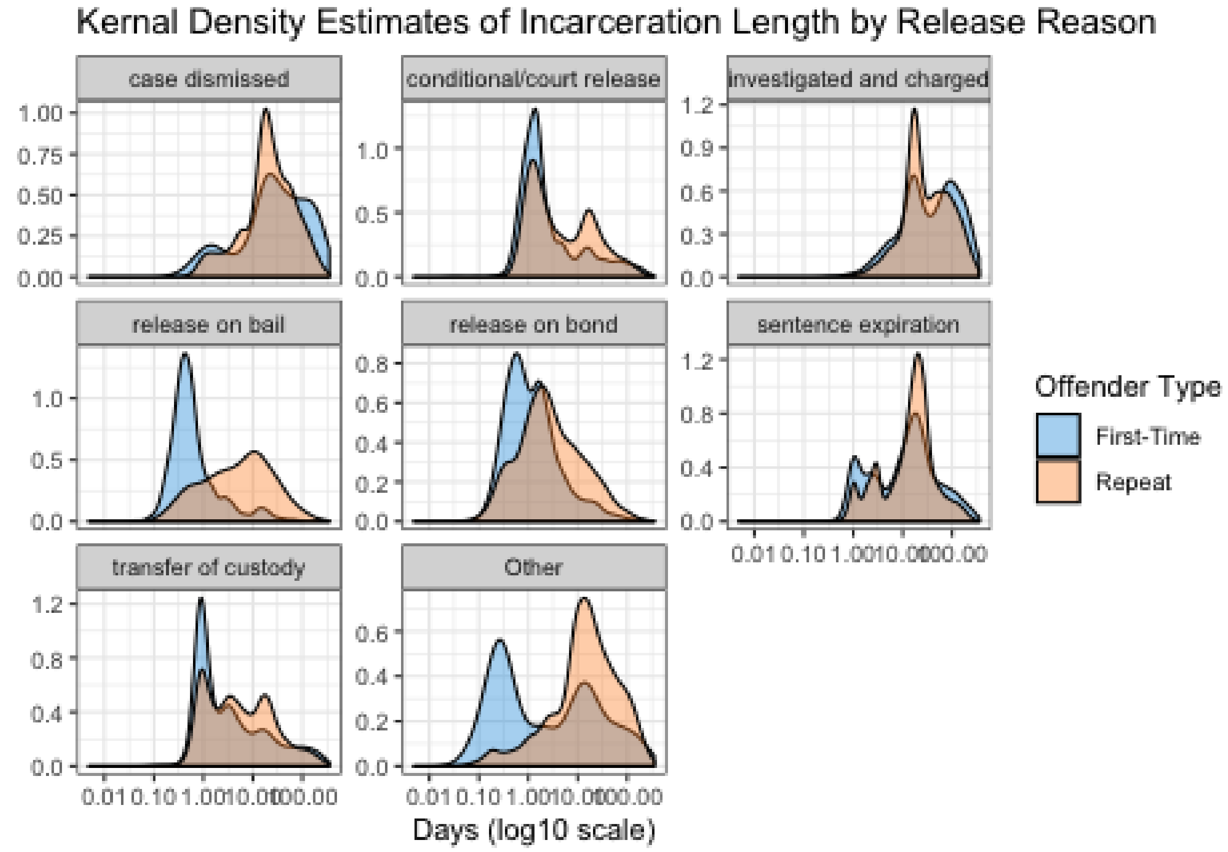
The next heat map visualizes the differences between repeat and first-time offenders. Orange corresponds to more cases being attributed to repeat offenders, while blue correspond to first-time offenders. The white areas show areas of little to no difference between offenders. A higher number of first-time offenders who committed violent and DUI offenses are given a conditional release while repeat offenders tend to serve out their sentences for violent offenses. Being released on bail for a DUI offense is also more common for first-time offenders. Repeat offenders tended to receive probation more often for property, drug, and failure to appear offenses.

```
hmap(release_matrix_repeat - release_matrix_first, method = "TSP",
     col = colorRampPalette(c("#1e9adf", "white", "#ff7251"))(100))
```



How does incarceration length vary by release? The kernel density plots below give a picture of the difference between repeat and first-time offenders.[11]

[11] The "not found" category–totaling 1,700 individuals–was omitted because each inmate did not have a release date, therefore the length of incarceration couldn't be calculated. These individuals, however, all had booking dates. It is not clear whether these inmates were in custody, or what exactly the "not found" category means since no codebook was publicly available.



For the most part, both offender types tend to spend similar amounts of time in jail before having their case dismissed, being released on probation, being investigated and charged, completing their sentence, and being transferred out of facility. Repeat offenders, however, usually spend more time in jail before bail release and "other" release reasons (a category made of 13 additional reasons). A table of average incarceration length for bail release is shown below. A complete table of release reasons is provided in the Appendix.

offender type	mean	median	IQR
First-Time	2.5	0.5	0.6
Repeat	13.7	5.9	13.4

This could be due to the fact that first-timers commit the majority of DUIs, which may have lower bail amounts than other offenses. Bond release appears to fall somewhere in-between. While there is considerable overlap, there is also a noticeable trend affecting repeaters. Exploring the dynamics between offender type, offense, and release reason could be addressed in a later analysis. Additional graphs are included in the Appendix.

National Comparison

It can be difficult to measure how different regions compare to one another or to national trends because carceral policies and practices can vary greatly from state-to-state. Per capita jail incarceration rates[12]

[12] This should not be confused with overall incarceration rates, which include all forms of carceral custody, such as state and federal prisons.

are considered one important measure for comparing different regions. To calculate King County's jail incarceration rate, I determine the average daily population from June 1st, 2018 to May 31st, 2019.

```
jail_int <- jail_df %>% count(boa_number, booking_dt, release_dt) %>%
                        mutate(interval = interval(booking_dt, release_dt))

timeframe <- interval(min(jail_df$booking_dt, na.rm = TRUE),
                      max(jail_df$release_dt, na.rm = TRUE))

days <- int_shift(
          interval(
            floor_date(min(jail_df$booking_dt, na.rm = TRUE), "days"),
            floor_date(min(jail_df$booking_dt, na.rm = TRUE), "days") + hms(59, 59, 23)),
                      days(1:(timeframe/days(1))))

daily_pop <- map_int(days, function(x) sum(int_overlaps(x, jail_int$interval), na.rm = TRUE))
```

On any given day, King County adult correctional facilities house around 1,424 inmates. King County has a jail incarceration rate of approximately 64 per 100,000,[13] which is nearly a quarter of the 2017 national average of 229 per 100,000, according to the Bureau of Justice Satistics.[14] The average length of incarceration is considerably lower as well: The 2017 national average was 26 days, while King County averages 14, a difference of 12 days.

Conclusion

Inmates tended to be incarcerated for an average of 14 days (and a median of 2.4 days), however, there is a significant gap between offender types. Repeat offenders tend to stay in jail for 15 days (median 4.7) and first-time offenders stay on average 12 days (median 1.5). Repeat offenders make up only 26% of the inmate population, but half of all bookings.

The most common RCW statute used to charge inmates is related to probation violation, and the most common form of release was also probation. When counting the frequency of crime using a dictionary-based method, the most common crimes tend to be violent (26%), bench warrants issued for "failure to appear" in court (23%), property crime (18%), and drug offenses (11.5%). Of these, repeat offenders made up 60% of violent crime, 66% of property crime, 61% of drug crime, and 66% of probation violations. First-time offenders, on the other hand, commit the majority of DUIs (74%) and traffic-related offenses (63%). For each of these offense categories, repeat offenders consistently served more time per booking than first-timers.

Conditional releases (probation) are most closely linked with failure to appear, drug, property, and violent offenses. First-timers are more likely to receive probation after committing a DUI or violent offense, while repeaters tend to receive probation for property, drug, and failure to appear offenses.

At a glance, incarceration lengths across release reasons do not greatly differ by offender type, except for bail release. However, further analysis is needed.

The jail incarceration rate for King County is roughly a quarter of the national rate. The average incarceration length is also lower–nearly half of the national average.

Lastly, there are many avenues for future analysis. As mentioned above, determining the effect that offense and offender type have on incarceration length and release reason could be useful for risk assessment. Offense

[13] The rate lowers to 57 per 100,000 when using the mean and rises to 62 when using a 20% trimmed mean. I use the median because the counting method creates a negative skew of the daily population. The skew is due to the limited time frame of the data set. Population is based on the 2018 Census estimate for King County, which is 2,233,163 according to the Census website: https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk, accessed 10/03/2019

[14] https://www.bjs.gov/content/pub/pdf/ji17.pdf, accessed 10/03/2019.

and offender type variables could be used to predict the likelihood of receiving a particular release. Also, analyzing the relationship between courts and the amount of time an offender spends in jail could shed light on how long inmates wait before their hearings, or how punitive a court is. And finally, a time series analysis of King County's jail data would be useful in understanding how offense and offender trends have, or have not, changed.

Appendix

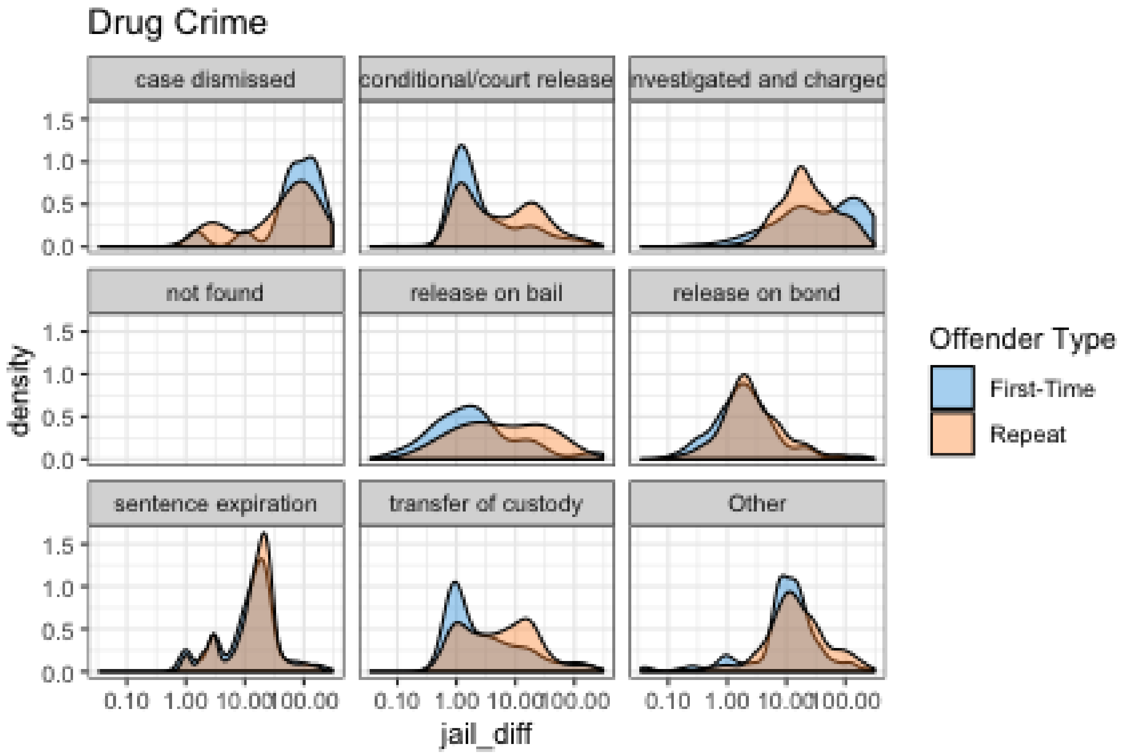
Table of Difference in Incarceration Length by Release Reason

```r
jail_df %>% mutate(release_reason = fct_lump(release_reason, n = 8)) %>%
            count(boa_number, name, jail_diff, offender_type, release_reason) %>%
            filter(!is.na(jail_diff)) %>%
            group_by(release_reason, offender_type) %>%
            summarize(mean = mean(jail_diff),
                      median = median(jail_diff),
                      IQR = IQR(jail_diff)) %>%
            gather("measure", "value", mean:IQR) %>%
            spread(offender_type, value) %>%
            mutate_if(is.numeric, round, 1) %>%
            mutate(diff = `Repeat` - `First-Time`) %>%
            kable() %>%
            kable_styling(bootstrap_options = c("condensed", "striped", "bordered"),
                              full_width = FALSE)
```

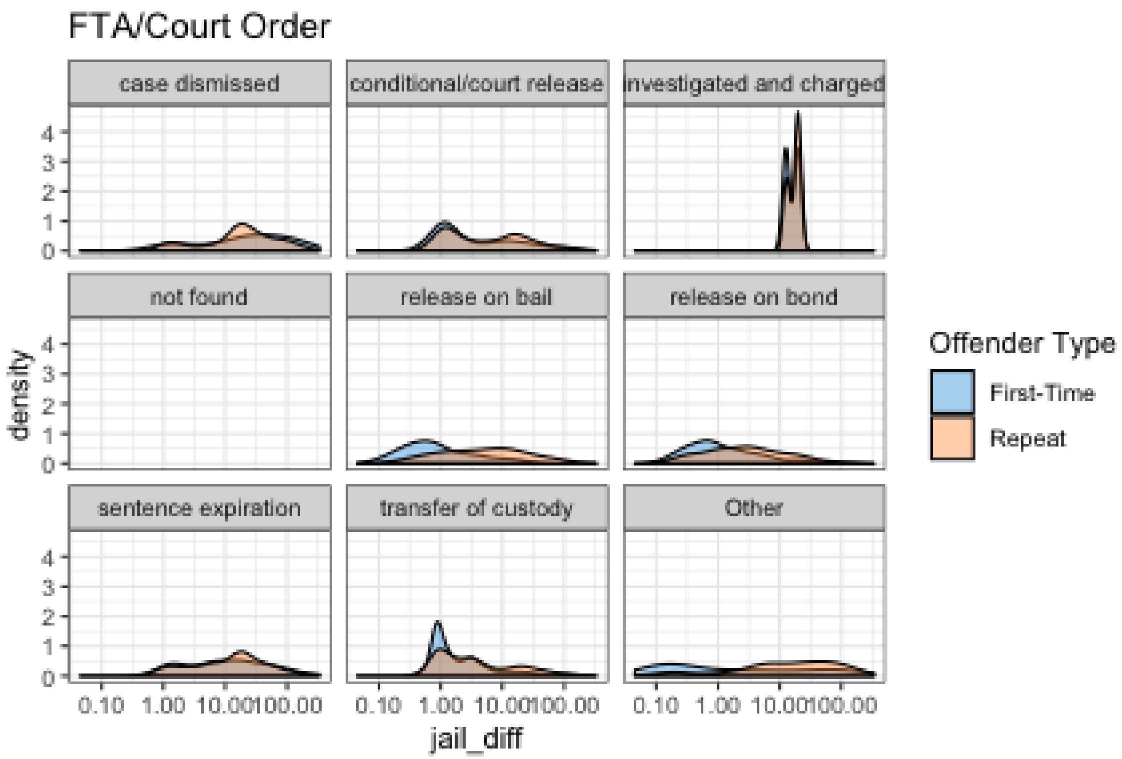
release_reason	measure	First-Time	Repeat	diff
case dismissed	IQR	83.4	35.2	-48.2
case dismissed	mean	63.1	37.7	-25.4
case dismissed	median	28.5	20.2	-8.3
conditional/court release	IQR	3.6	15.0	11.4
conditional/court release	mean	12.4	14.9	2.5
conditional/court release	median	1.6	3.0	1.4
investigated and charged	IQR	88.2	55.2	-33.0
investigated and charged	mean	68.9	52.2	-16.7
investigated and charged	median	37.9	25.8	-12.1
release on bail	IQR	0.6	13.4	12.8
release on bail	mean	2.5	13.7	11.2
release on bail	median	0.5	5.9	5.4
release on bond	IQR	1.9	6.6	4.7
release on bond	mean	4.1	8.5	4.4
release on bond	median	1.0	2.4	1.4
sentence expiration	IQR	24.4	20.1	-4.3
sentence expiration	mean	29.0	24.6	-4.4
sentence expiration	median	13.6	17.2	3.6
transfer of custody	IQR	9.7	15.9	6.2
transfer of custody	mean	18.3	17.8	-0.5
transfer of custody	median	2.0	4.6	2.6
Other	IQR	15.6	32.8	17.2
Other	mean	21.1	33.4	12.3
Other	median	1.7	15.3	13.6

KDE Plots of Incarceration Length, Release Reason, Offense and Offender Type

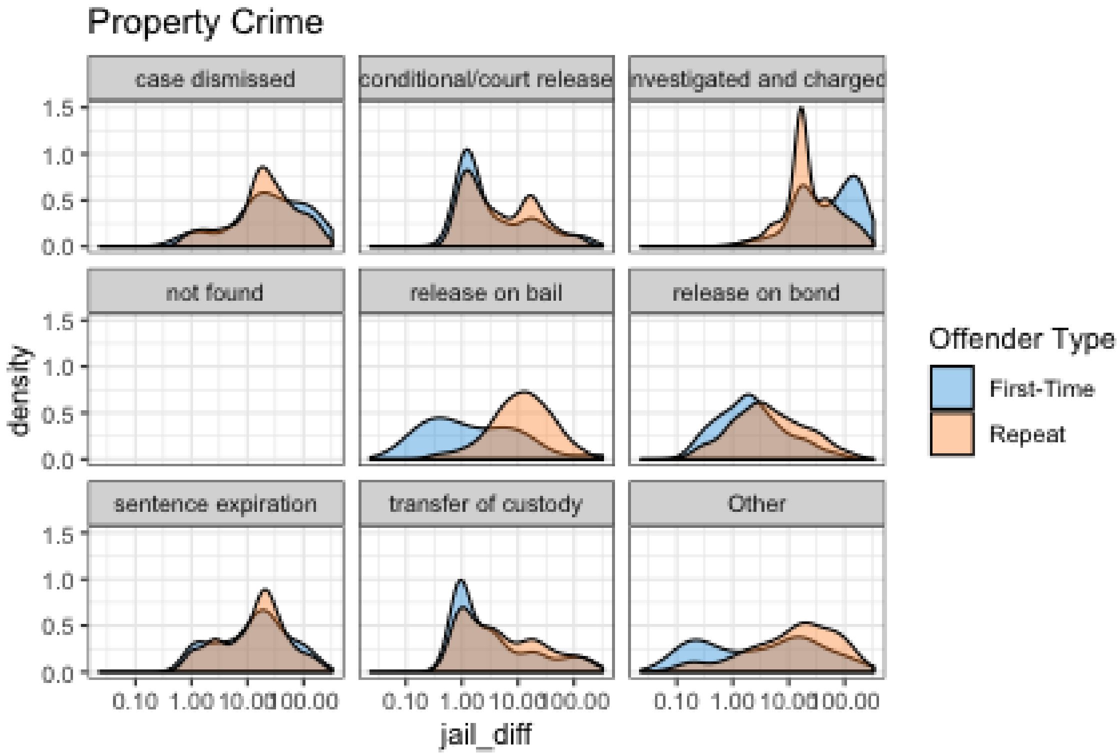
```
jail_df %>% mutate(release_reason = fct_lump(release_reason, n = 8)) %>%
            filter(drug > 0) %>%
            count(boa_number, offender_type, jail_diff, release_reason) %>%
            ggplot(aes(jail_diff, fill = offender_type)) +
                   geom_density(alpha = .4) +
                   scale_x_continuous(trans = "log10",
                                      labels = comma_format(accuracy = .01)) +
                   scale_fill_manual(values = palette3,
                                     name = "Offender Type") +
                   facet_wrap(~release_reason) +
                   labs(title = "Drug Crime") +
                   theme_bw()
```



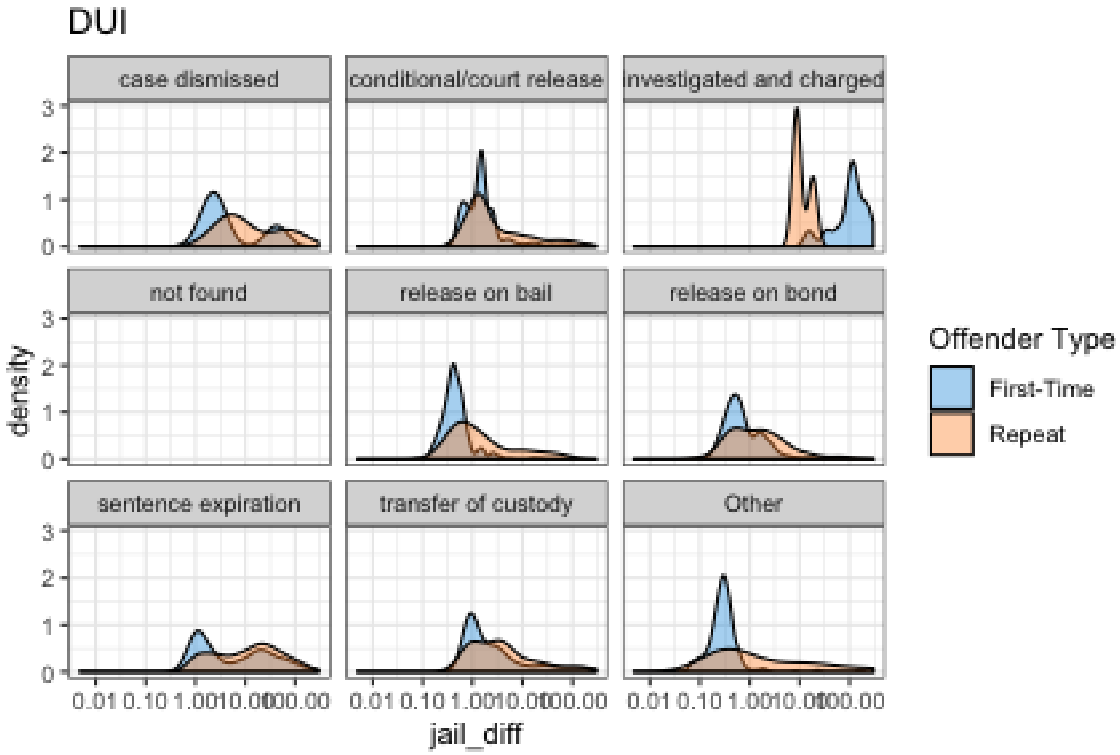
```
jail_df %>% mutate(release_reason = fct_lump(release_reason, n = 8)) %>%
            filter(fta > 0) %>%
            count(boa_number, offender_type, jail_diff, release_reason) %>%
            ggplot(aes(jail_diff, fill = offender_type)) +
                   geom_density(alpha = .4) +
                   scale_x_continuous(trans = "log10",
                                      labels = comma_format(accuracy = .01)) +
                   scale_fill_manual(values = palette3,
                                     name = "Offender Type") +
                   facet_wrap(~release_reason) +
                   labs(title = "FTA/Court Order") +
                   theme_bw()
```



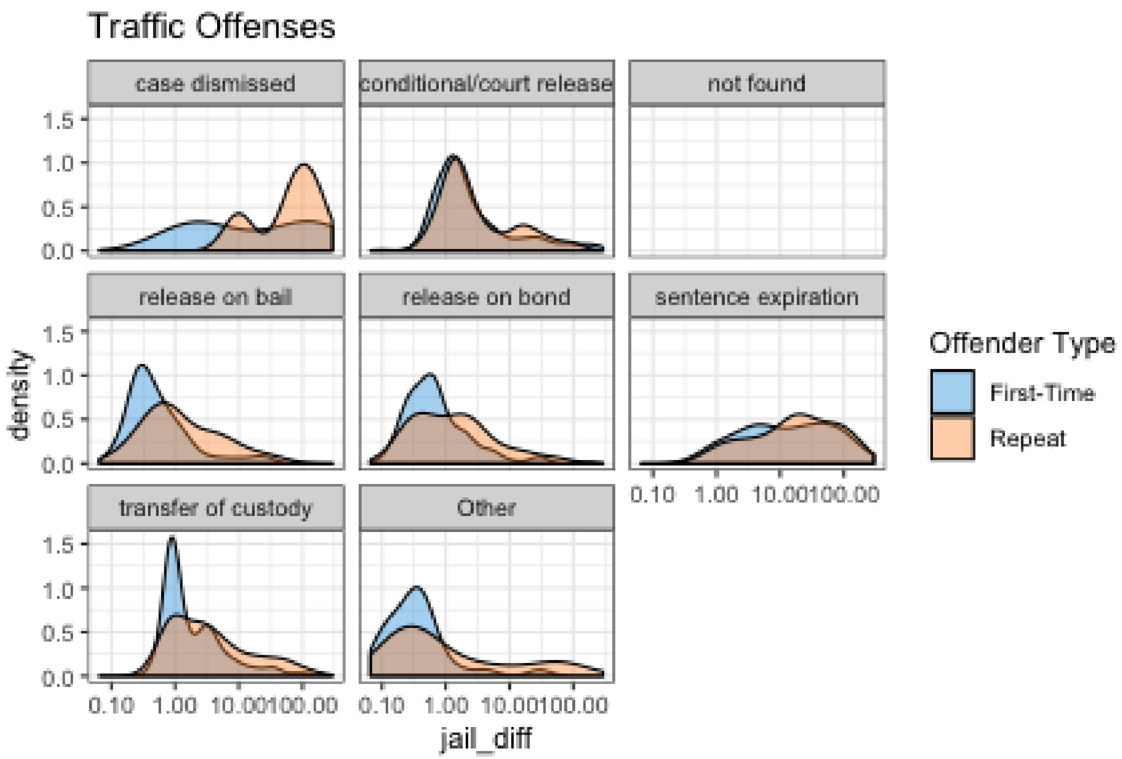
```
jail_df %>% mutate(release_reason = fct_lump(release_reason, n = 8)) %>%
            filter(property_crime > 0) %>%
            count(boa_number, offender_type, jail_diff, release_reason) %>%
            ggplot(aes(jail_diff, fill = offender_type)) +
                   geom_density(alpha = .4) +
                   scale_x_continuous(trans = "log10",
                                      labels = comma_format(accuracy = .01)) +
                   scale_fill_manual(values = palette3,
                                     name = "Offender Type") +
                   facet_wrap(~release_reason) +
                   labs(title = "Property Crime") +
                   theme_bw()
```



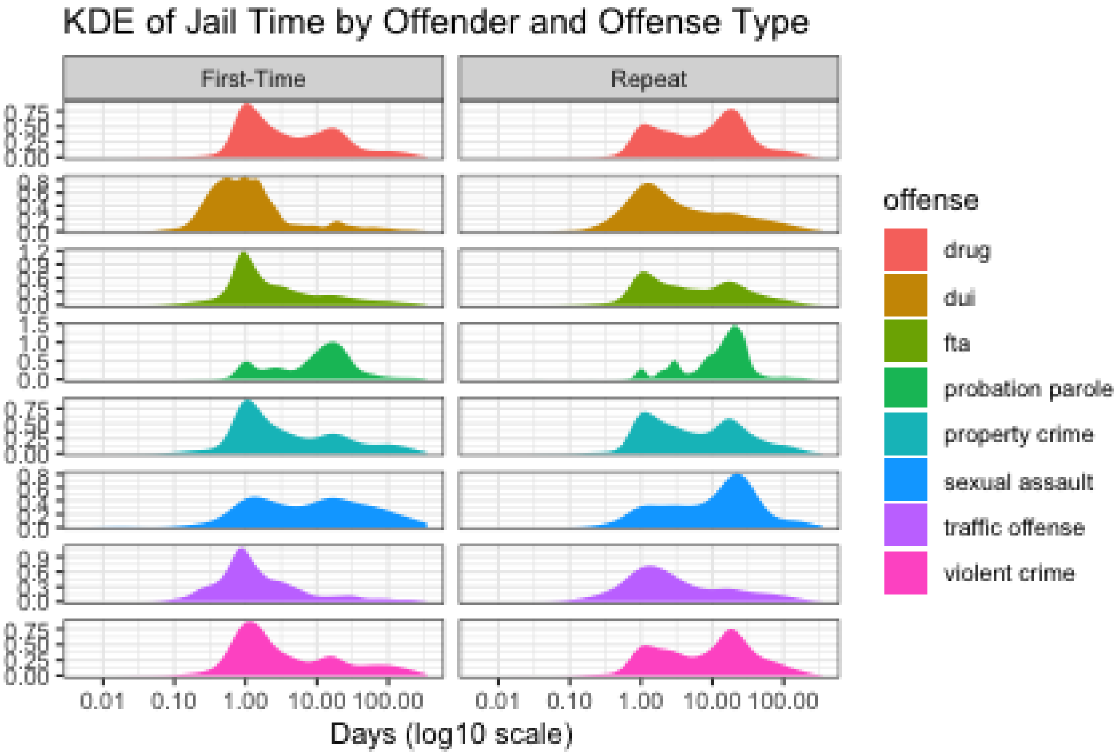
```
jail_df %>% mutate(release_reason = fct_lump(release_reason, n = 8)) %>%
            filter(dui > 0) %>%
            count(boa_number, offender_type, jail_diff, release_reason) %>%
            ggplot(aes(jail_diff, fill = offender_type)) +
                   geom_density(alpha = .4) +
                   scale_x_continuous(trans = "log10",
                                      labels = comma_format(accuracy = .01)) +
                   scale_fill_manual(values = palette3,
                                     name = "Offender Type") +
                   facet_wrap(~release_reason) +
                   labs(title = "DUI") +
                   theme_bw()
```



```
jail_df %>% mutate(release_reason = fct_lump(release_reason, n = 8)) %>%
            filter(traffic_offense > 0) %>%
            count(boa_number, offender_type, jail_diff, release_reason) %>%
            ggplot(aes(jail_diff, fill = offender_type)) +
                   geom_density(alpha = .4) +
                   scale_x_continuous(trans = "log10",
                                      labels = comma_format(accuracy = .01)) +
                   scale_fill_manual(values = palette3,
                                     name = "Offender Type") +
                   facet_wrap(~release_reason) +
                   labs(title = "Traffic Offenses") +
                   theme_bw()
```



```
jail_df %>% select(boa_number, jail_diff, offender_type, property_crime:fta) %>%
            gather("offense", "value", property_crime:fta) %>%
            filter(value > 0) %>%
            count(boa_number, jail_diff, offender_type, offense) %>%
            mutate(offense = str_replace(offense, "_", " ")) %>%
            ggplot(aes(jail_diff, fill = offense)) +
                   stat_density() +
                   facet_grid(offense ~ offender_type,
                              scales = "free_y") +
                   labs(title = "KDE of Jail Time by Offender and Offense Type",
                        x = "Days (log10 scale)",
                        y = NULL) +
                   scale_x_continuous(trans = "log10",
                                      labels = comma_format(accuracy = .01)) +
                   theme_bw() +
                   theme(strip.background.y = element_blank(),
                         strip.text.y = element_blank())
```



t-tests and qqplots

two-sample t-test (population)

```r
offense_long %>% mutate(offender_type = fct_rev(offender_type)) %>%
                 t.test(jail_diff ~ offender_type,
                        data = .,
                        alternative = "two.sided",
                        mu = 0,
                        paired = FALSE,
                        conf.int = TRUE,
                        conf.level = .95) %>%
                 tidy() %>%
                 mutate_if(is.numeric, round, 2) %>%
                 select(estimate:method) %>%
                 kable() %>%
                 kable_styling(bootstrap_options = c("condensed", "bordered"),
                               full_width = FALSE)
```

estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method
3.67	17.91	14.25	11.6	0	40037.52	3.05	4.29	Welch Two Sample t-test

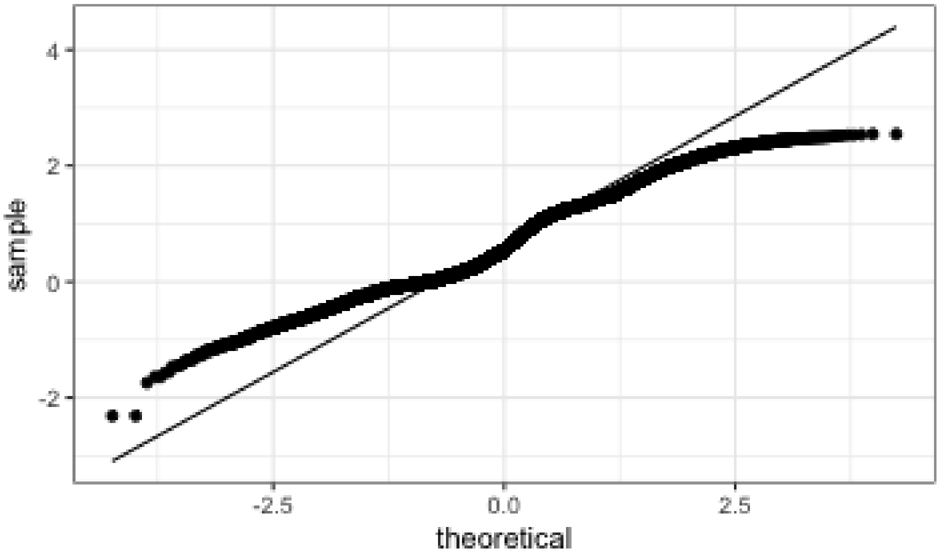
log10 two-sample t-test (population)

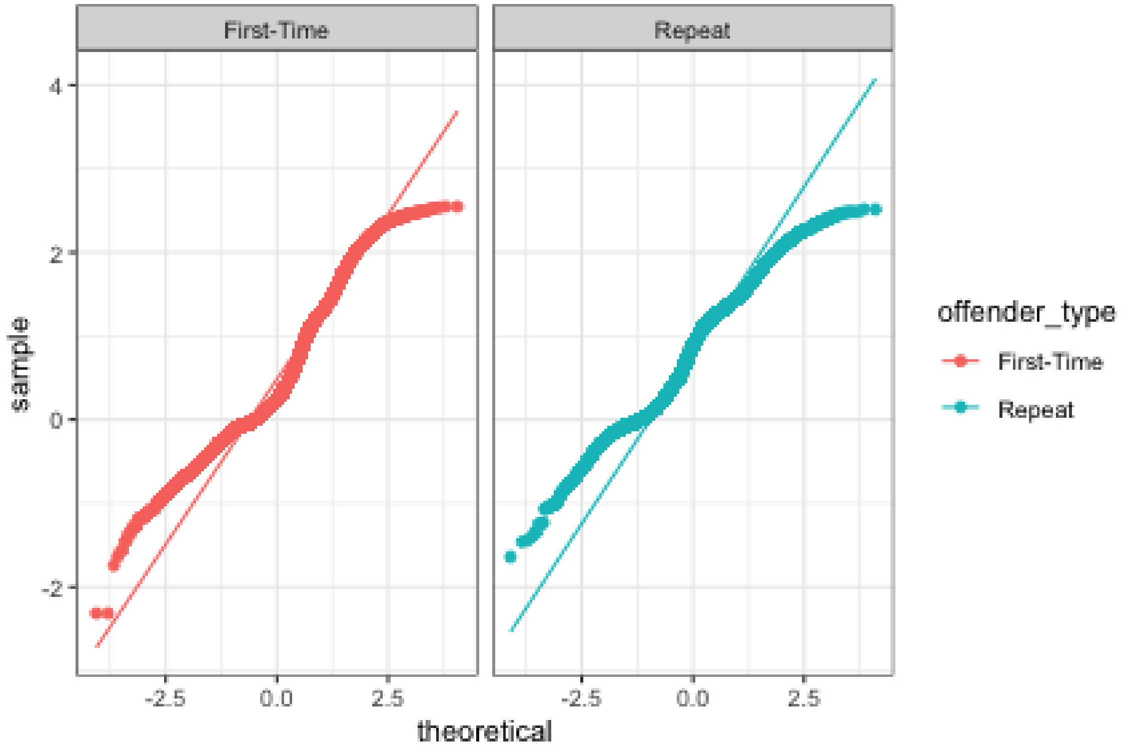
```r
ten_power <- function(x) {
  10^x
}

offense_long %>% mutate(offender_type = fct_rev(offender_type)) %>%
                 t.test(log10(jail_diff) ~ offender_type,
                         data = .,
                         alternative = "two.sided",
                         mu = 0,
                         paired = FALSE,
                         conf.int = TRUE,
                         conf.level = .95) %>%
                 tidy() %>%
                 mutate_at(vars(estimate:estimate2, conf.low, conf.high),
                           funs(ten_power)) %>%
                 mutate_if(is.numeric, round, 2) %>%
                 select(estimate:method) %>%
                 kable() %>%
                 kable_styling(bootstrap_options = c("condensed", "bordered"),
                               full_width = FALSE)
```

estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method
2.24	6.55	2.93	53.36	0	41559.05	2.17	2.31	Welch Two Sample t-test

qqplot of log10(jail_diff) population

```
offense_long %>% mutate(jail_diff = log10(jail_diff)) %>%
                 ggplot(aes(sample = jail_diff)) +
                       stat_qq() +
                       stat_qq_line() +
                       theme_bw()
```

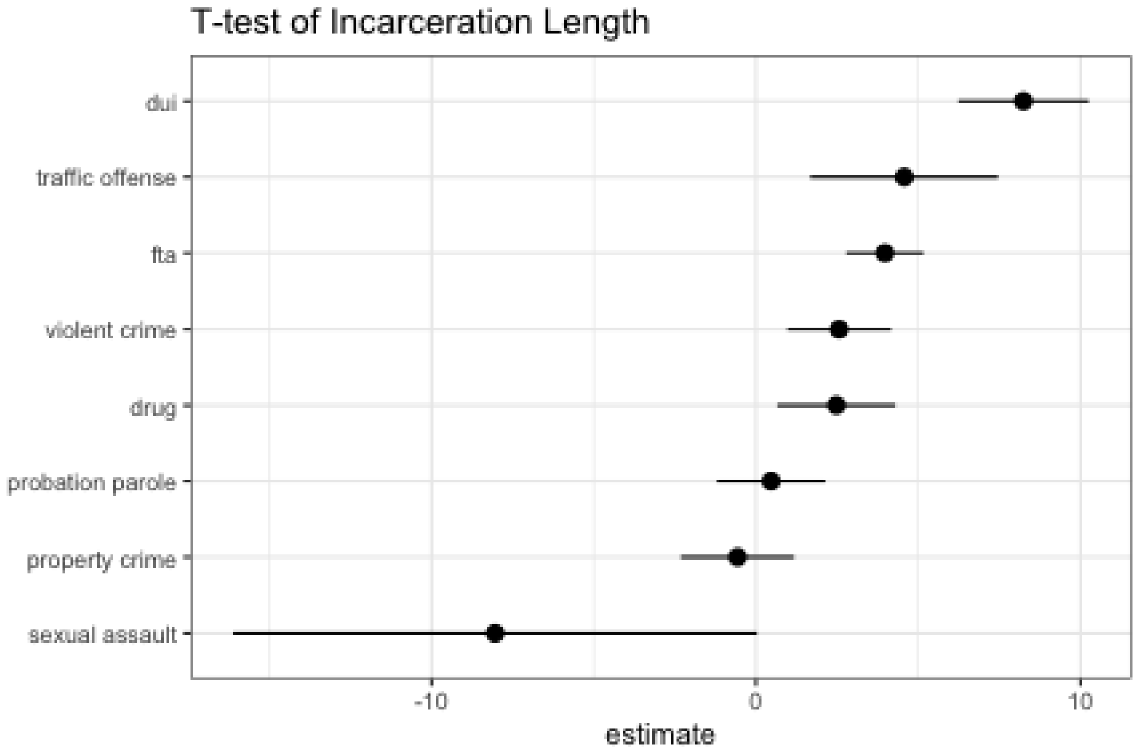




t-test df

```
t_nest <- offense_long %>% mutate(offender_type = fct_rev(offender_type)) %>%
                           nest(-offense) %>%
                           mutate(model = map(data, ~ t.test(jail_diff ~ offender_type,
                                                              data = .,
                                                              alternative = "two.sided",
                                                              mu = 0,
                                                              paired = FALSE,
                                                              conf.int = TRUE,
                                                              conf.level = .95)))

t_models <- bind_rows(lapply(t_nest$model, tidy), .id = "offense")

t_models <- t_models %>% mutate(offense = t_nest$offense)
```

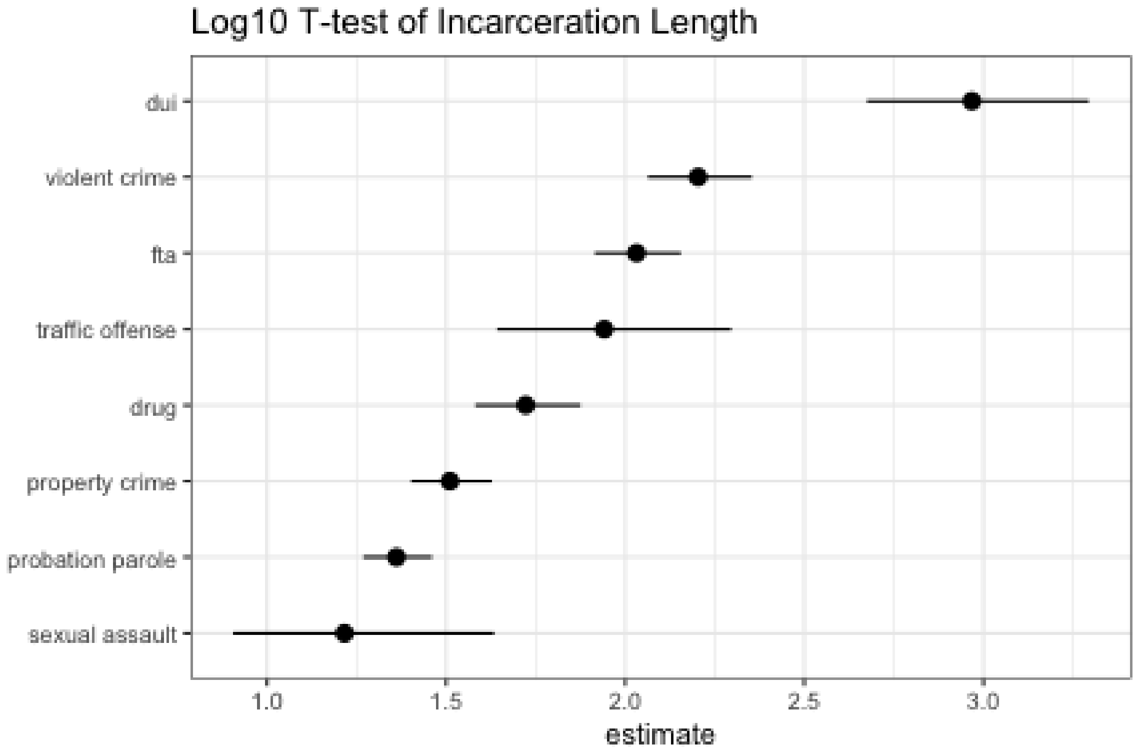
```
t_models %>% mutate(offense = fct_reorder(offense, estimate, last)) %>%
                              ggplot(aes(offense, estimate)) +
                                     geom_pointrange(aes(ymin = conf.low,
                                                         ymax = conf.high)) +
                                     coord_flip() +
                                     labs(title = "T-test of Incarceration Length",
                                          x = NULL) +
                                     theme_bw()
```



log10 t-test df

```
tlog_nest <- offense_long %>% mutate(jail_diff = log10(jail_diff),
                                     offender_type = fct_rev(offender_type)) %>%
                              nest(-offense) %>%
                              mutate(model = map(data, ~ t.test(jail_diff ~ offender_type,
                                                                 data = .,
                                                                 alternative = "two.sided",
                                                                 mu = 0,
                                                                 paired = FALSE,
                                                                 conf.int = TRUE,
                                                                 conf.level = .95)))

tlog_models <- bind_rows(lapply(tlog_nest$model, tidy), .id = "offense")

tlog_models <- tlog_models %>% mutate(offense = tlog_nest$offense)
```

```
tlog_models %>% mutate(offense = fct_reorder(offense, estimate, last)) %>%
                                 mutate_at(vars(estimate:estimate2, conf.low:conf.high),
                                           funs(ten_power)) %>%
                                 ggplot(aes(offense, estimate)) +
                                        geom_pointrange(aes(ymin = conf.low,
                                                            ymax = conf.high)) +
                                        coord_flip() +
                                        labs(title = "Log10 T-test of Incarceration Length",
                                             x = NULL) +
                                        theme_bw()
```


Log10 T-test of Incarceration Length
dui
violent crime
fta
traffic offense
drug
property crime
probation parole
sexual assault
1.0
1.5
2.0
2.5
3.0
estimate