Final Report

Mayra Smith-Coronado

2024-06-26

Intro

The included datasets come from two tables within the DSD relational database: Incarceration and Person. The incarceration table's unit of analysis is one booking into the jail and includes booking-related data such as the times into and out of the jail. The person table's unit of analysis is one person and includes demographic information like age and race. The "Person_id" column is the common variable between them. (For this exercise, the real Person_id has been suppressed and replaced with a unique random number to protect identities.)

The point of this exercise is to demonstrate how you think analytically as much as it is to arrive at the "correct" answers. Please provide your best answers to the questions below, using the tools and methods you deem most effective. Please submit written answers in a clear and concise form by the deadline. **Please also share your code so we can review it.**

Exercise

Consider the year from July 1, 2021 to June 30, 2022 as the analysis period.

---

1. How many total bookings into the jail were there in that time period?

There were 21,842 total bookings into the jail from July 1st, 2021 to June 30, 2022.

---

2. How many unique people were booked into the jail?

There were 15,510 unique people who were booked into the jail from July 1st, 2021 to June 30, 2022.

---

3. How many people were in the jail at the moment of the data extraction?

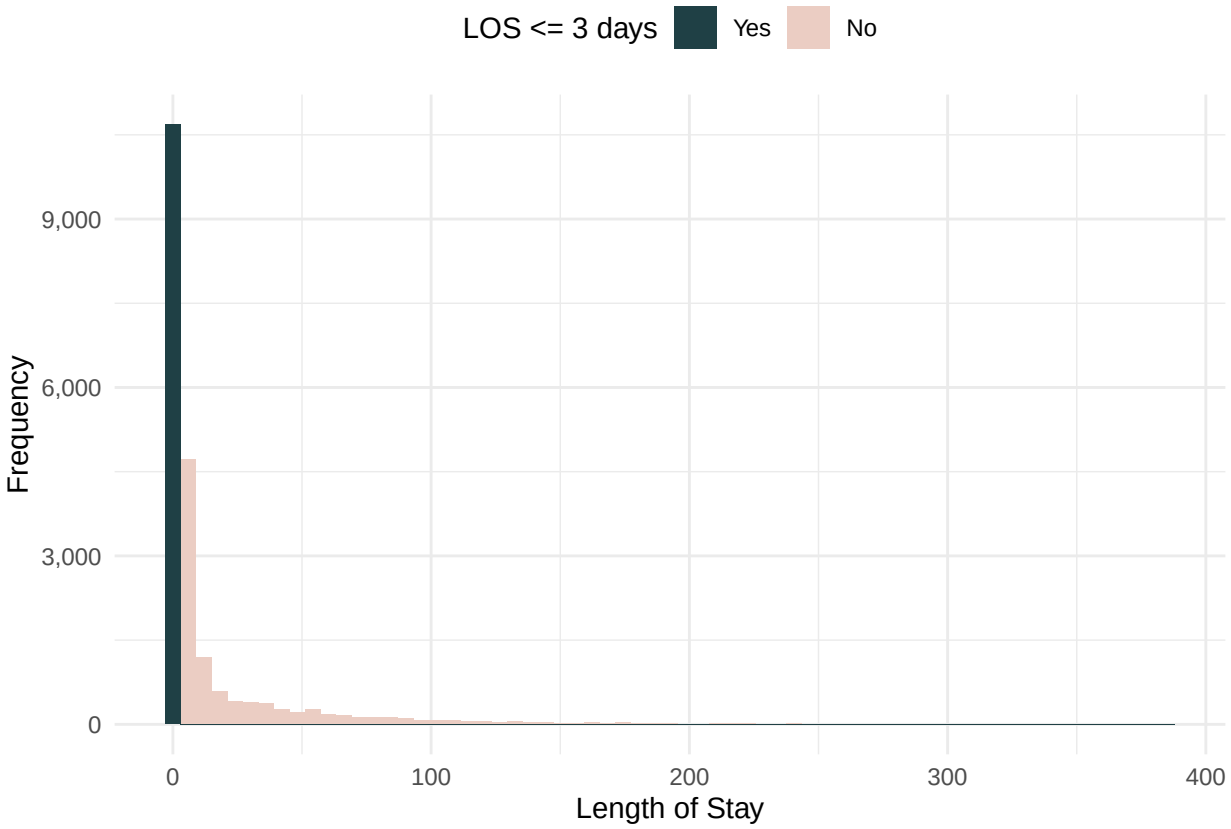
This would include not only the people who were booked during the analysis period, but the people who were booked before analysis period, but were not yet released. Based on this understanding of the question, I estimated approximately 16,456 people in the jail at the moment of the data extraction.

---

4. Consider the length of stay (LOS): the duration of each booking. Describe the LOS over the year analysis period. What insights (statistical and otherwise) does it provide you about variations in the jail population?

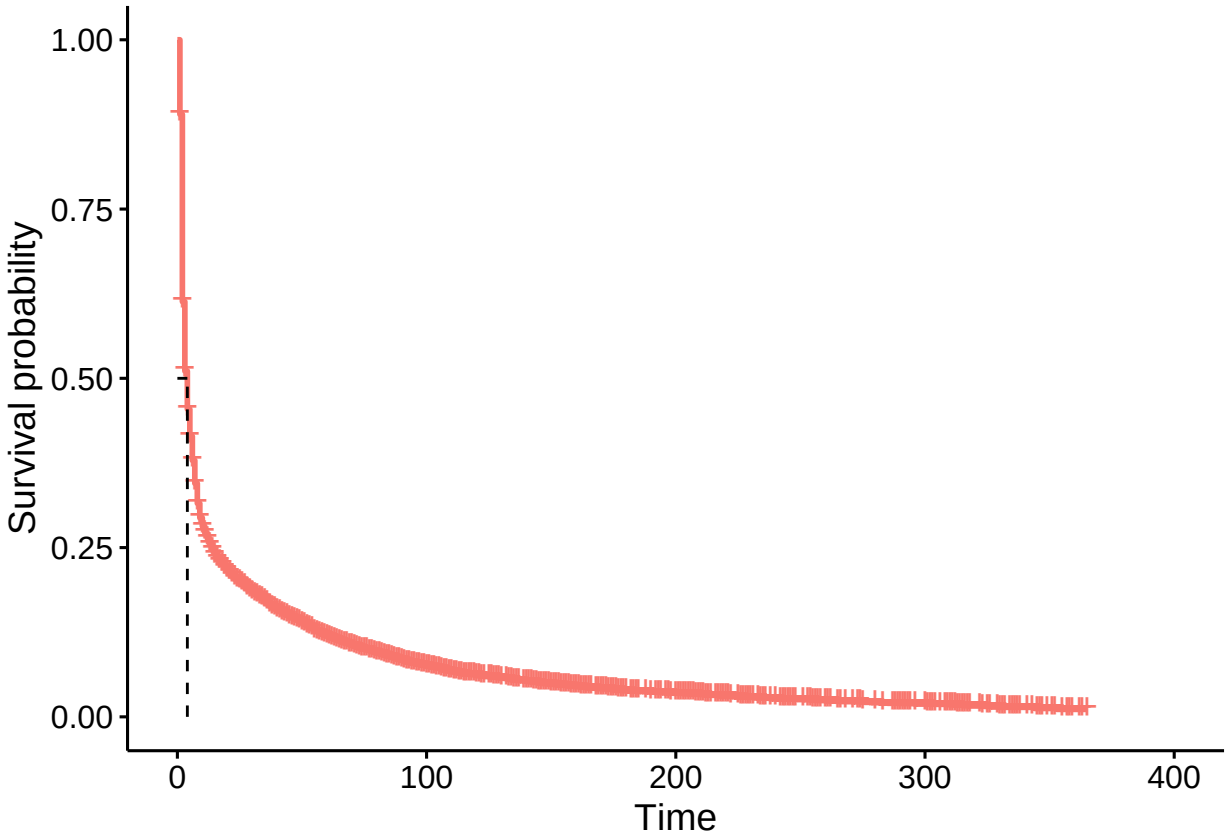
To look at length of stay during the analysis period, I think it is best to look at the bookings that were from July 1, 2021 to June 30, 2022. Bookings missing a release date were excluded from the visual review. Overall, the data seems to suggest that the people who are detained in jail tend to be changing frequently, if a little over half the jail population is only detained for 3 days or less.

- About 10,683 (51.3%) of bookings had a length of stay of at least 3 days.
- The most frequent length of stay being 2 days with a total of 6,041 (29.0%).
- There are about 25% of bookings that ranged from 10 days to 386.



From reviewing the distribution of the length of stay, it was determined that our vairable of interest was positively skewed. To determine what the median length of stay was for this population with 95% certainty, a survival analysis would make the most sense. In this scenario, we would be trying to understand the amount of time it takes a person who was booked to be released from jail during our analysis period.

From a survival analysis, the median release time for a person can be estimated to be 4 days, with a 95% CI [4,4]. One thing to note is that this median does differ from our review of the descriptive statistics because in this data set we kept bookings with a missing date where these bookings were previously excluded.


1.00
0.75
0.50
0.25
0.00
Survival probability
0
100
200
300
400
Time

5. What was the average daily population in the jail during that year? Daily population ought to include anyone who spent even one minute in the jail in a given day. Please describe the methods/approach you used to answer this question. What tool did you use? What functions or other capabilities within the tool? (That is, help another analyst replicate what you did. Sharing code with your answers is encouraged but by no means required.)

Use a for loop to create a data set where if a booking occurred from Monday to Friday (5 days) the booking will be spread across 5 rows. This will create a data set where we can look at who was in jail each day of the analysis period. In order to correctly capture all the individuals in jail during this time, we will use the data set `bookings_during_extraction` created to answer question 3. The for loop I used can be found below.

```
# initialize table
booking_long = NULL

for(ith_booking in 1:nrow(bookings_during_extraction)){

  # gather the current bookings information
  booking = bookings_during_extraction[ith_booking, ]

  # if release date is missing, set this to the last day of the analysis period
  # to capture each individual, even people who are still in jail after the analysis time constraint
  if(!is.na(booking$release_out)) {
    release_out <- booking$release_out
  } else {
     release_out = as.Date("2022-06-30")
  }

  # convert the booking row into a long table where each row represents a day
  # the person was in jail based on their current booking.
  dates_incarcerated = seq(booking$date_in, release_out, by = "1 day")
  ith_booking_long <- tibble(date = dates_incarcerated,
                             booking_number = booking$booking_number,
                             person_id = booking$person_id)

  # combine the current bookings table with the larger data set.
  booking_long <- booking_long %>%
    bind_rows(ith_booking_long)
}
```

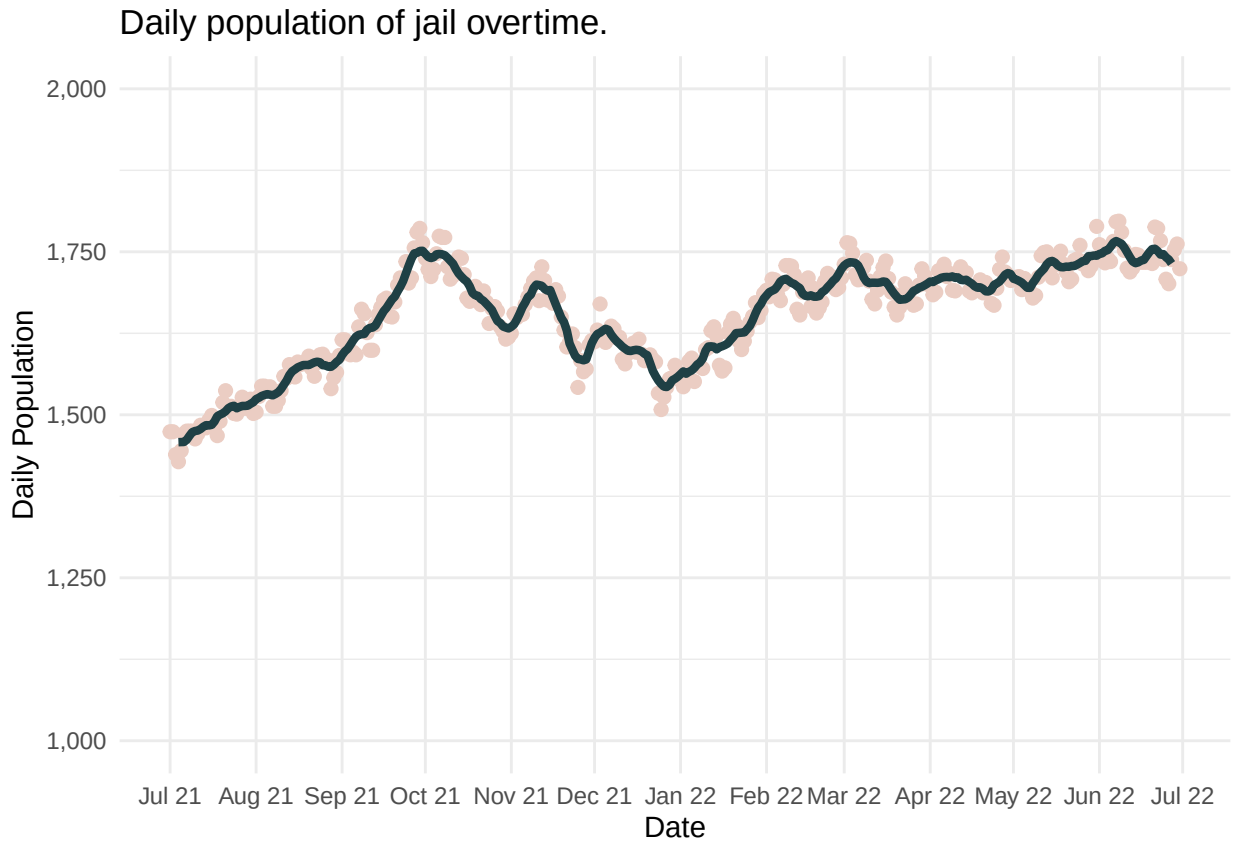
Now that we have a data set where a bookings length of stay is spread out by date, we can look at the daily population. We can do this by counting the number of bookings by date. This will provide us the daily population for our analysis period.

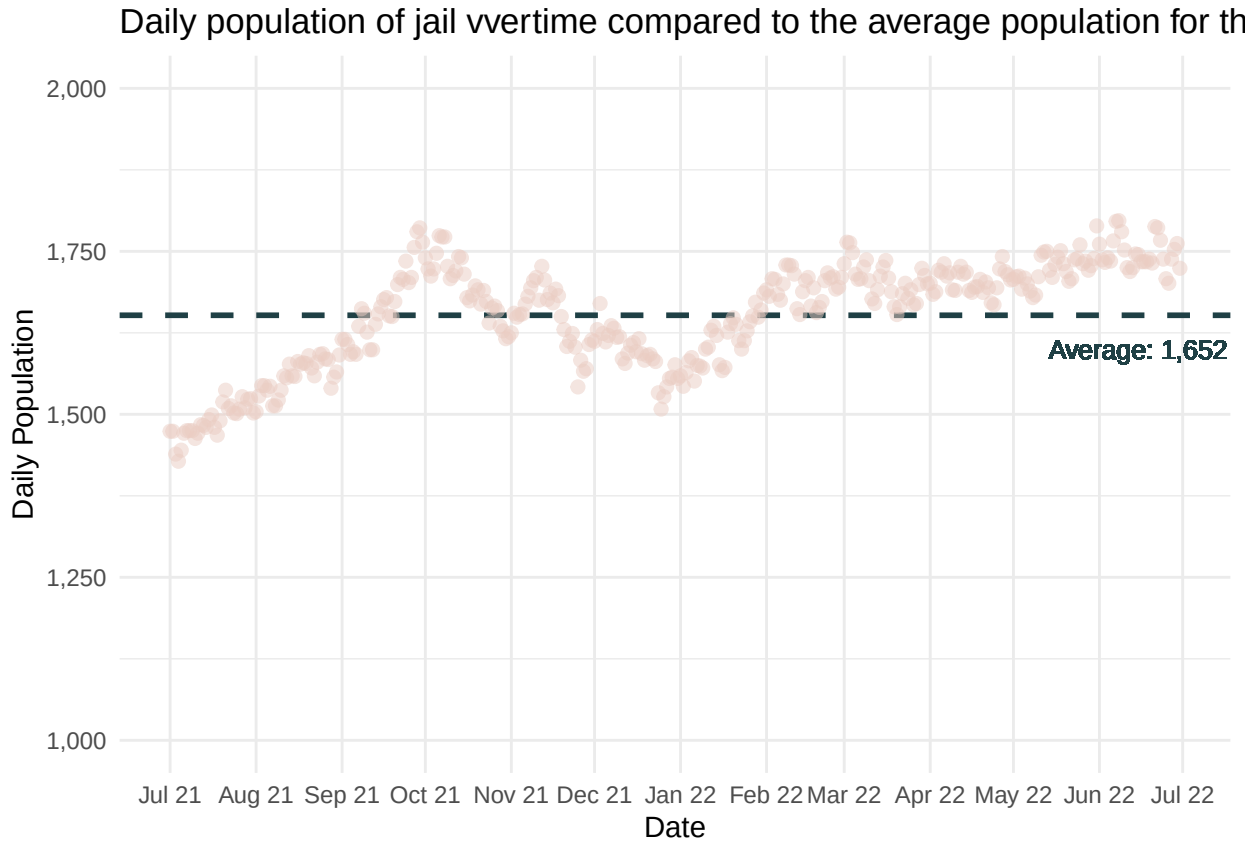
```
# count the number of people in jail each day
bookings_by_day <- booking_long %>%
  filter(date >= "2021-07-01" & date <= "2022-06-30") %>%
  arrange(date) %>%
  group_by(date) %>%
  count() %>%
  ungroup() %>%
  rename(daily_population = n)

# What is the average daily population for the analysis period?
(average_daily_population <- mean(bookings_by_day$daily_population))
```

```
## [1] 1651.923
```

From the calculations above, we find that the average daily population in the jail during our analysis period is 1,652. By creating a data set in this way, we can also examine the trend of the daily population overtime. Examples of some graphics are shown below.





---

6. Which day during that year had the lowest daily population? What was the population that day? Which day had the highest daily population? What was the population that day?

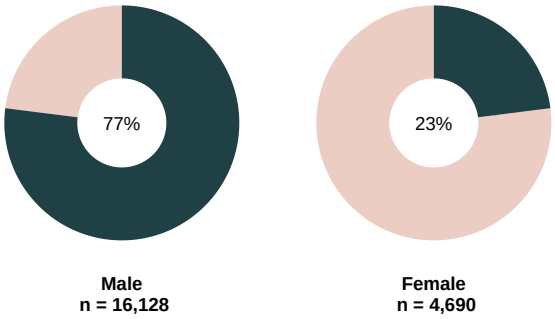
July 04, 2021 was the day with the lowest daily population of 1,428. The day with the highest daily population was June 08, 2022 with a population of 1,797.

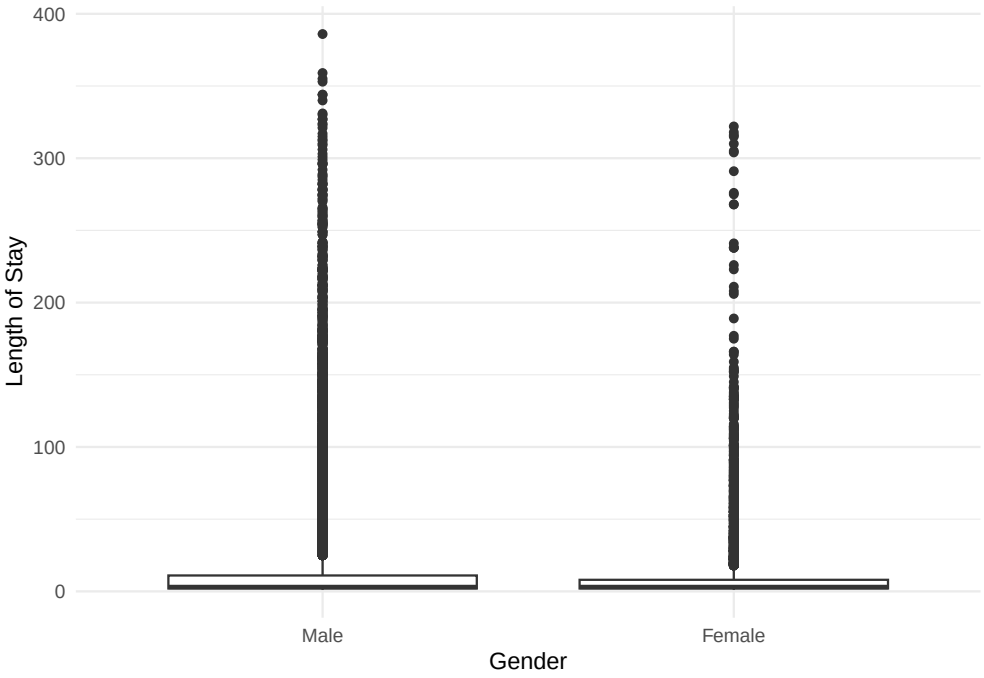
---

7. Please provide a basic analysis of jail demographics during that year period. Are there meaningful statistical relationships among the different groups in the jail?

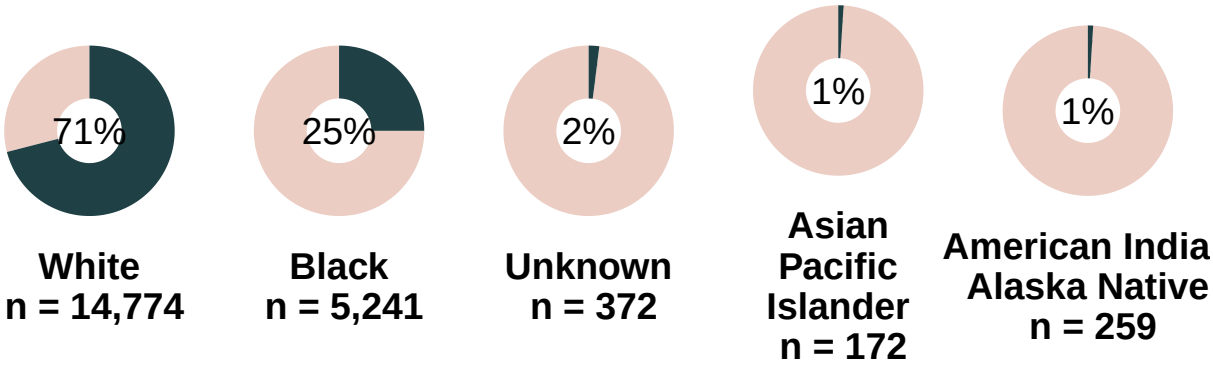
For this analysis, we are assuming that each booking is a new person, where people who have been booked more than once will be double counted.

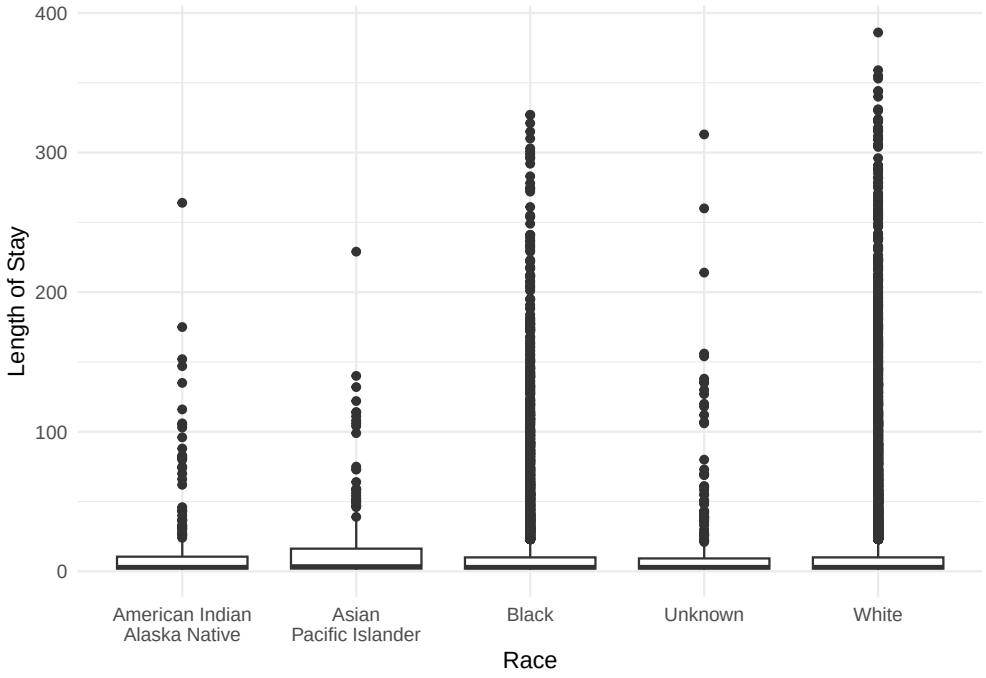
- The majority of the jail population was men (77%) where their length of stay had a larger variation (IQR = 6, 9) compared to women.



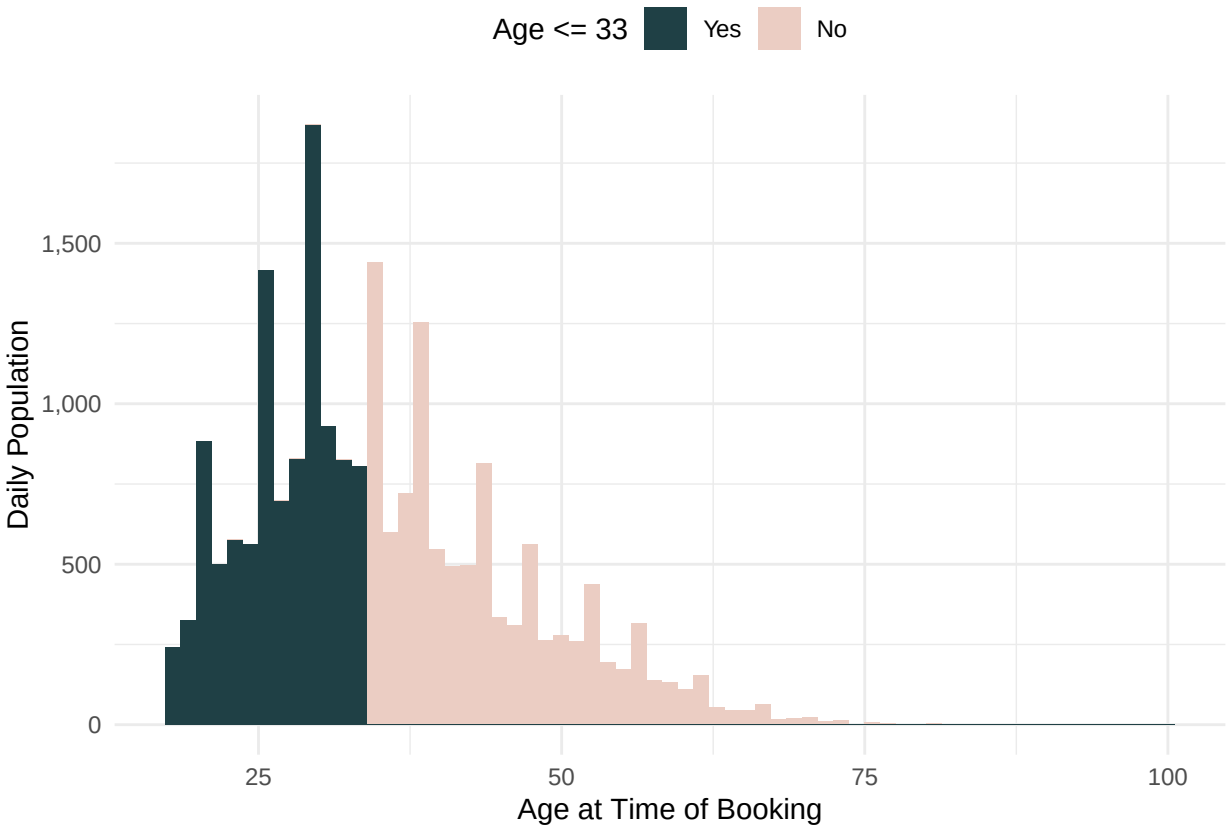


- A large proportion of the jail population was white (71%) and the variance of length of stay among the other race's seemed consistent, with the exception of the Asian or Pacific Islander group.





- The age of the people in jail during this time period was also positively skewed where the jail population tends to be younger. At least 50% of the population was 33 years old and younger.



Because of how skewed the data is for length of stay among each demographic, I believe that Wilcoxon rank sum test would be the best way to examine the median by demographic and see if there is a difference between groups. To examine the median for race, people who are white are going to be the control group and compared to the other races.

Reference Group	Comparison Group	P Value
Men	Women	<0.0001
White	Asian or Pacific Islander	0.28
White	American Indian or Alaska Native	0.03
White	Black	0.43
White	Unknown	0.54

The Wilcoxon test showed that the medians were significantly different ($p < 0.0001$) between men and women in our jail population where the median length of stay for women was 3 (IRQ = 6) and the median length of stay among men was 3 (IQR = 9). An additional Wilcoxon test showed that the length of stay median was significantly different ($p < 0.03$) between white jail detainees who had a median length of stay of 3 (IQR = 8) and American Indian or Alaska Natives who had a median length of stay of (IQR = 8.5).

8. What other insights, useful facts, or questions/concerns did you uncover, if any, in the data?

- While working with this data, I was curious about what others have noticed when it comes to length of stay among people who are detained. One similarity that I noticed between this data and the peer reviewed papers that I found published was how young people represented the majority of the jail population. In the paper that I found, a relationship the researchers noticed among their cohort was that the length of stay was longer for people who were younger. I would think it would be valuable to investigate this further and see if this relationship is seen in this study population, since in the demographic section it was visible that the jail population tended to be younger.
- An additional research question I found while getting familiar with this topic was the relationship between length of stay and being detained more than once. Among the bookings in this analysis period, 3,664 (24.4%) of people were detained more than once so I wonder how this impacted an individuals second stay. This could be examined using a survival analysis.
- Due to the large population of bookings with length of stays that were 10 days or longer (25%), I would like to examine the demographic differences between people with a short, medium, and long length of stay. An analysis that looked at these 3 categories might give us more insight to the relationship between demographics and length of stay.
- Lastly, I would like to make a visual that compares the number of people detained by race in our analysis year to the general ppopulation and maybe narrow this down to the state or county the jail is located to see if the population in the jail mirrors or is different to the general public. An example of the graphic I am thinking about can be found here: https://stephanieevergreen.com/proportion-plots/