OSEI Senior Data Analyst Exercise: Exploratory Analysis

Mayra Smith-Coronado

2024-06-24

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.**

```r
# Load in Incarceration Data ----------------------------------------------
# note time is not going to be read in for this data set, only date information
incarceration_data <- read.xlsx("~/osei_data_exercise/01 - data/Incarceration.xlsx") %>%
  # convert to tibble
  tibble() %>%
  # update column names to follow snake case naming convention
  janitor::clean_names() %>%
  # correctly read excel dates as dates instead of numeric values
  mutate(across(.cols = c("date_in", "release_out"),
                .fns = janitor::excel_numeric_to_date))

# check to make sure that there are no duplicate booking numbers and that
# each row represents one booking
incarceration_data %>%
  count(booking_number) %>%
  filter(n > 1)
```

```
## # A tibble: 0 x 2
## # i 2 variables: booking_number <chr>, n <int>
```

```r
# Load in Person Data -----------------------------------------------------
person_data <- read.xlsx("~/osei_data_exercise/01 - data/Person.xlsx") %>%
  # convert to tibble
  tibble() %>%
  # update column names to follow snake case naming convention
  janitor::clean_names() %>%
  # correctly read excel dates as dates instead of numeric values
  mutate(across(.cols = c("dob"),
                .fns = janitor::excel_numeric_to_date))

# check to make sure that there are no duplicate people and that
# each row represents one person
person_data %>%
  count(person_id) %>%
  filter(n > 1)
```

```
## # A tibble: 0 x 2
## # i 2 variables: person_id <dbl>, n <int>
```

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?

```
bookings_in_analysis_period <- incarceration_data %>%
  filter(date_in >= "2021-07-01" &
           date_in <= "2022-06-30")

nrow(bookings_in_analysis_period) %>%
  scales::comma()
```

```
## [1] "21,842"
```

---

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

```
bookings_in_analysis_period %>%
  select(person_id) %>%
  distinct() %>%
  count() %>%
  pull() %>%
  scales::comma()
```

```
## [1] "15,510"
```

---

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

```
# gather bookings where a person was booked before the analysis period, but
# has no release date
bookings_still_incarcerated <- incarceration_data %>%
  filter(date_in < "2021-07-01",
         is.na(release_out))

nrow(bookings_still_incarcerated) %>%
  scales::comma()
```

```
## [1] "130"
```

```
bookings_still_incarcerated %>%
  summarise(first_booking = min(date_in),
            last_booking = max(date_in))
```

```
## # A tibble: 1 x 2
##   first_booking last_booking
##   <date>        <date>
## 1 2016-09-07    2021-06-30
```

```
# gather bookings where a person was booked before the analysis period,
# but they were released within the analysis period
bookings_released_during_extraction <- incarceration_data %>%
  filter(date_in < "2021-07-01",
         release_out >= "2021-07-01" &
           release_out <= "2022-06-30")

nrow(bookings_released_during_extraction) %>%
  scales::comma()
```

```
## [1] "1,276"
```

```
bookings_released_during_extraction %>%
  summarise(first_release = min(release_out),
            last_release = max(release_out))
```

```
## # A tibble: 1 x 2
##   first_release last_release
##   <date>        <date>
## 1 2021-07-01    2022-06-29
```

```
# now create a table with all the bookings that we have identified to
# have occurred during the extraction and bookings that
# occurred before the extraction, but were not yet released
bookings_during_extraction <- bookings_in_analysis_period %>%
  bind_rows(bookings_still_incarcerated) %>%
  bind_rows(bookings_released_during_extraction)

# verify no duplicate bookings
bookings_during_extraction %>%
  count(booking_number) %>%
  filter(n > 1)
```

```
## # A tibble: 0 x 2
## # i 2 variables: booking_number <chr>, n <int>
```

```
# now count the number of people in jail at the moment of the
# data extraction
bookings_during_extraction %>%
  select(person_id) %>%
  distinct() %>%
  count() %>%
  pull() %>%
  scales::comma()
```

```
## [1] "16,456"
```

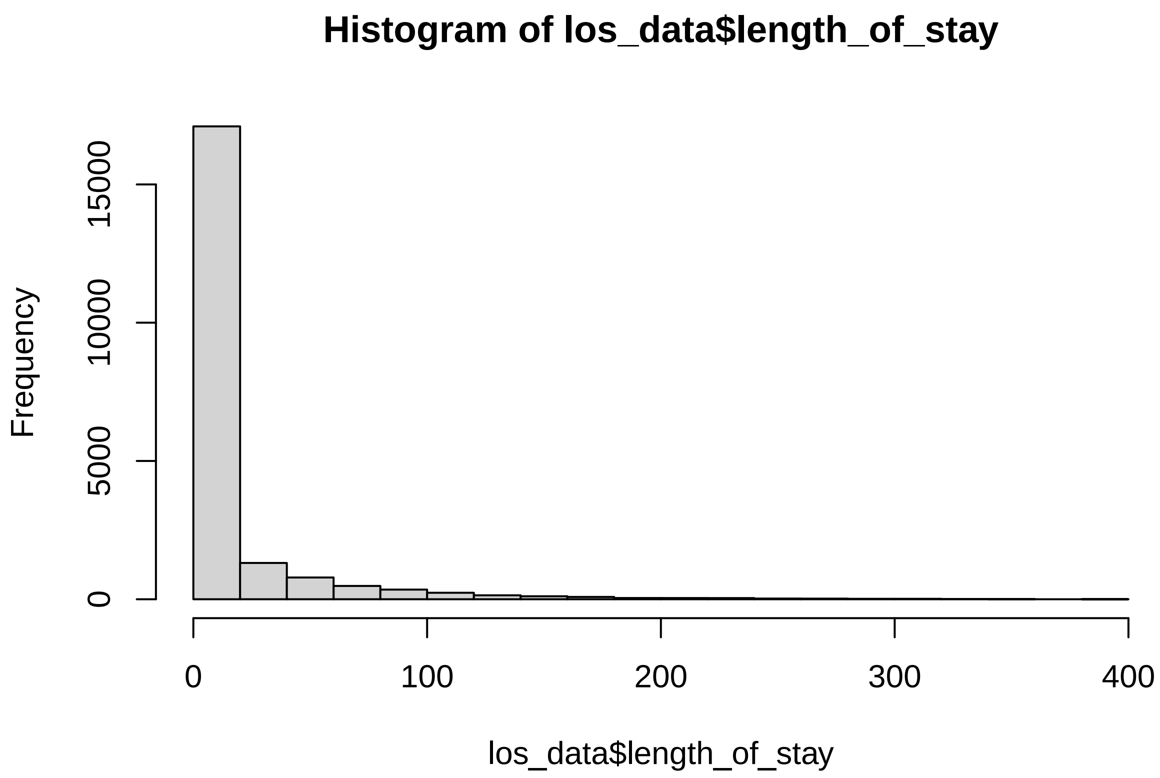
---

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.

- About 48.5% of bookings had a length of stay of at least 3 days.
- The most frequent length of stay being 2 day.
- There are about 25% of bookings that ranged from 10 days to 386.
- About 75.3% of people were only booked once

```
los_data <- bookings_in_analysis_period %>%
  # remove bookings with no release date
  filter(!is.na(release_out)) %>%
  # calculate length of stay
  mutate(length_of_stay = as.numeric(release_out - date_in) + 1)


# examine the distribution of the length of stay for this analysis period
hist(los_data$length_of_stay)
```



```
# get descriptive statistics to better understand the distribution
los_data %>%
  select(length_of_stay) %>%
  summary()
```

```
##  length_of_stay
##  Min.   :  1.00
##  1st Qu.:  2.00
##  Median :  3.00
##  Mean   : 16.54
##  3rd Qu.: 10.00
##  Max.   :386.00
```

```
los_data %>%
  tabyl(length_of_stay) %>%
  adorn_pct_formatting() %>%
  slice(1:21)
```

```
##  length_of_stay    n percent
##               1 2404   11.5%
##               2 6041   29.0%
##               3 2238   10.8%
##               4 1227    5.9%
##               5  867    4.2%
##               6  772    3.7%
##               7  734    3.5%
##               8  651    3.1%
##               9  477    2.3%
##              10  274    1.3%
##              11  220    1.1%
##              12  174    0.8%
##              13  181    0.9%
##              14  188    0.9%
##              15  155    0.7%
##              16  124    0.6%
##              17  114    0.5%
##              18   80    0.4%
##              19  114    0.5%
##              20   65    0.3%
##              21   90    0.4%
```

```r
los_data %>%
  mutate(grouped_los = ifelse(length_of_stay >= 10, "10+", length_of_stay),
         grouped_los = factor(grouped_los, levels = c(0:9, "10+"))) %>%
  tabyl(grouped_los) %>%
  adorn_pct_formatting()
```

```
##  grouped_los    n percent
##            0    0    0.0%
##            1 2404   11.5%
##            2 6041   29.0%
##            3 2238   10.8%
##            4 1227    5.9%
##            5  867    4.2%
##            6  772    3.7%
##            7  734    3.5%
##            8  651    3.1%
##            9  477    2.3%
##          10+ 5407   26.0%
```

```r
# how many times are people rebooked (booked at least one time)
rebooking <- los_data %>%
  arrange(person_id, date_in) %>%
  group_by(person_id) %>%
  mutate(frequency_booked = 1:n()) %>%
  ungroup() %>%
  arrange(person_id, desc(frequency_booked)) %>%
  distinct(person_id, .keep_all = T) %>%
  mutate(booked_multiple_times = ifelse(frequency_booked >= 2, "2+", "1"))

# review the total times people have been booked
rebooking %>%
  tabyl(frequency_booked) %>%
  adorn_pct_formatting()
```

```
##  frequency_booked     n percent
##                 1 11348   75.6%
##                 2  2391   15.9%
##                    3   774     5.2%
##                    4   298     2.0%
##                    5   109     0.7%
##                    6    54     0.4%
##                    7    23     0.2%
##                    8     7     0.0%
##                    9     5     0.0%
##                   13     1     0.0%
##                   14     1     0.0%
##                   16     1     0.0%
```

```r
# review how many people have been booked more than once
rebooking %>%
  tabyl(booked_multiple_times) %>%
  adorn_pct_formatting()
```

```
##  booked_multiple_times     n percent
##                      1 11348   75.6%
##                     2+  3664   24.4%
```

From the review of the distribution, 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.

- release dates after study period are considered censored event.
- release dates that are missing are considered a censored event.

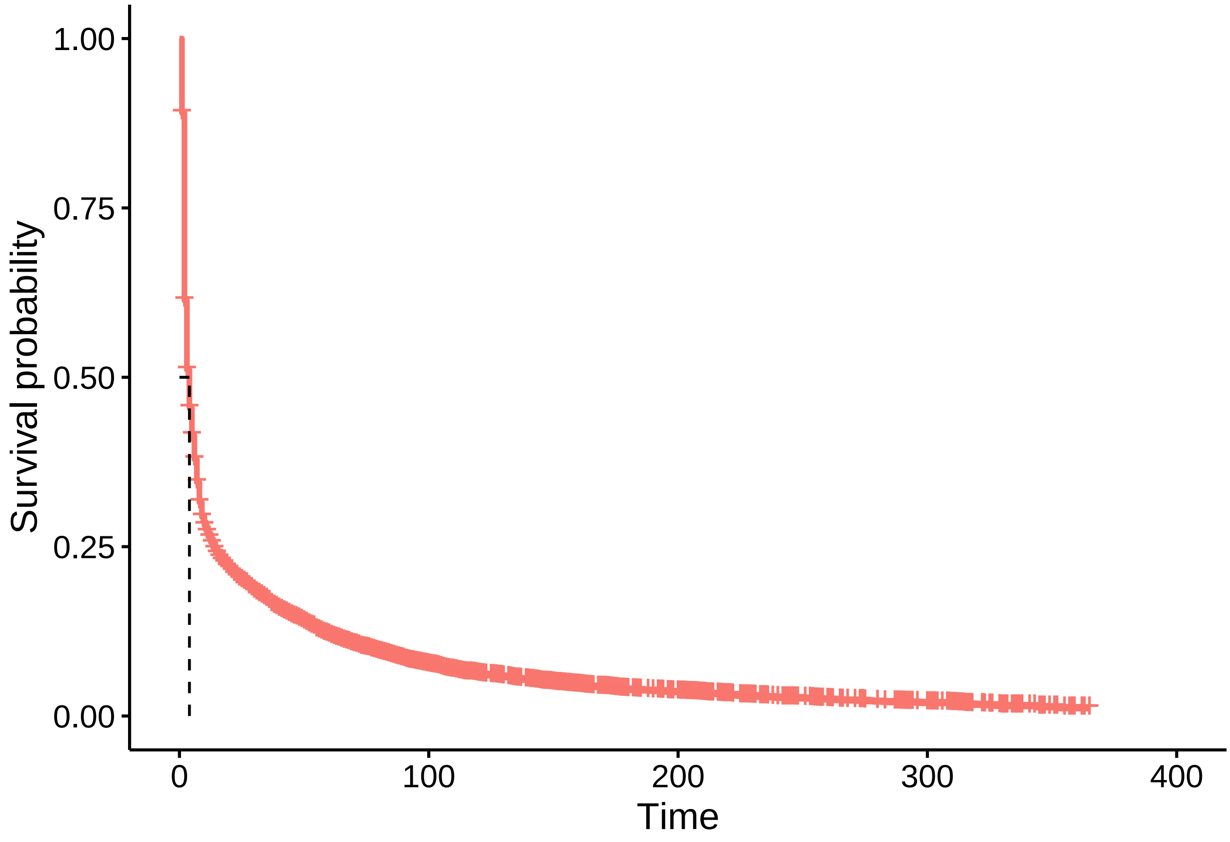
```r
# create the event flag and create a time flag that represents length of stay
# remembering that this length of stay is until the end of the study period
survival_data <- bookings_in_analysis_period %>%
  mutate(event = case_when(
    is.na(release_out) ~ 0,
    release_out > "2022-06-30" ~ 0,
    T ~ 1)) %>%
  mutate(time = case_when(
    is.na(release_out) ~ as.numeric(ymd("2022-06-30") - date_in) + 1,
    release_out > "2022-06-30" ~ as.numeric(ymd("2022-06-30") - date_in) + 1,
    T ~ as.numeric(release_out - date_in) + 1
  )) %>%
  select(booking_number, time, event)

# check the number of events (releases) that occurred during the analysis period
survival_data %>%
  tabyl(event) %>%
  adorn_pct_formatting()
```

```
##  event     n percent
##      0  1536    7.0%
##      1 20306   93.0%
```

```r
# calculate the Kaplan-Meier estimate
km <- survival::survfit(Surv(time, event) ~ 1,
  data = survival_data
)

km
```

```
## Call: survfit(formula = Surv(time, event) ~ 1, data = survival_data)
##
##          n events median 0.95LCL 0.95UCL
## [1,] 21842  20306      4       4       4
```

```r
# examine a graph of the probablity of release
survminer::ggsurvplot(km,
  conf.int = FALSE,
  surv.median.line = "hv",
  legend = "none"
)
```



From this analysis, the median release time for a person can be estimated to be 4 days, with a 95% CI [4,4]. A next step would be to examine if an individual being booked again during the study period would impact their length of stay.

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 dataset where if a booking occurred from monday to friday (5 days) the booking will be spread across 5 rows. This will create a dataset where we can look at who was in the jail each day of the analysis period. In order to correctly capture all the individuals in the jail during this time, we will use the dataset `bookings_during_extraction` created in question 3.

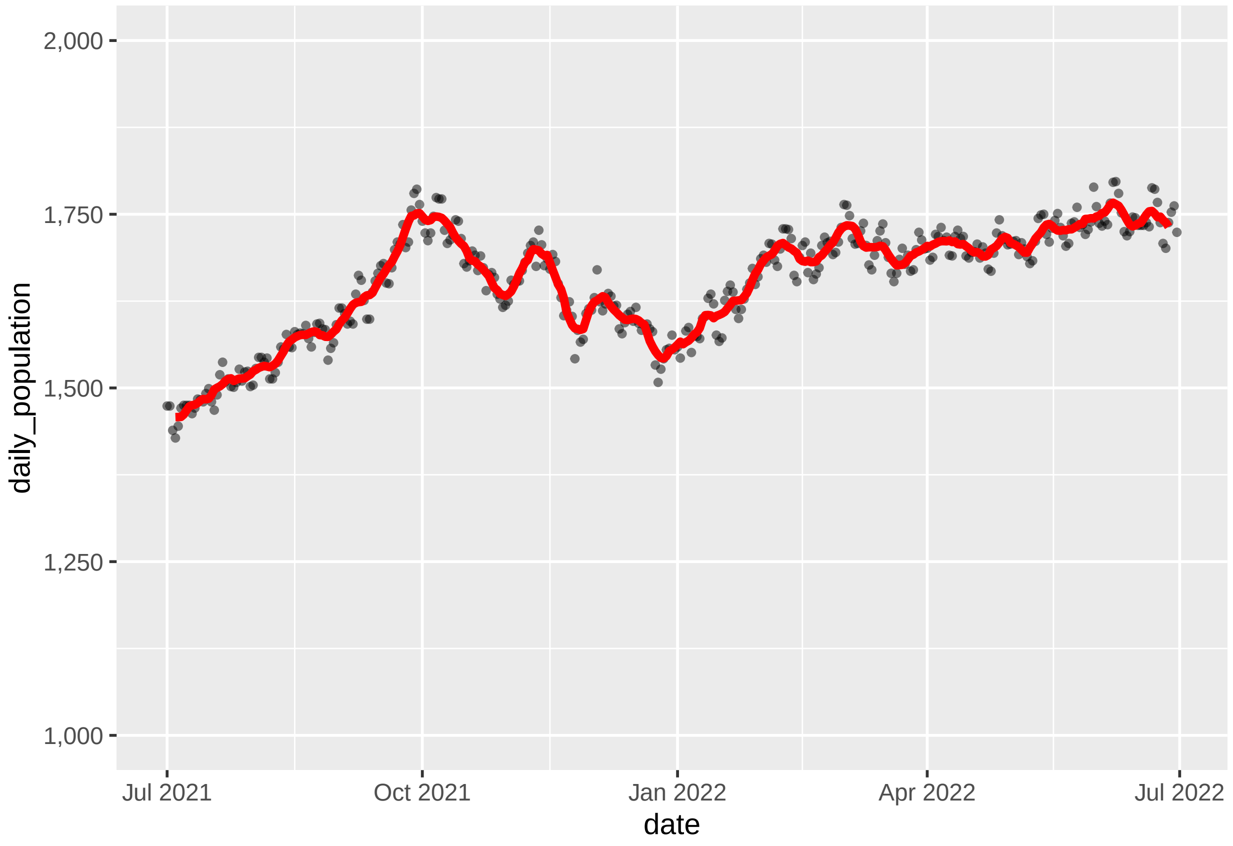
```
# initialize table
booking_long = NULL

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 dataset.
  booking_long <- booking_long %>%
    bind_rows(ith_booking_long)
}
```

Now that we have a dataset where a bookings length of stay is spreadout by date, we can look at the daily population.

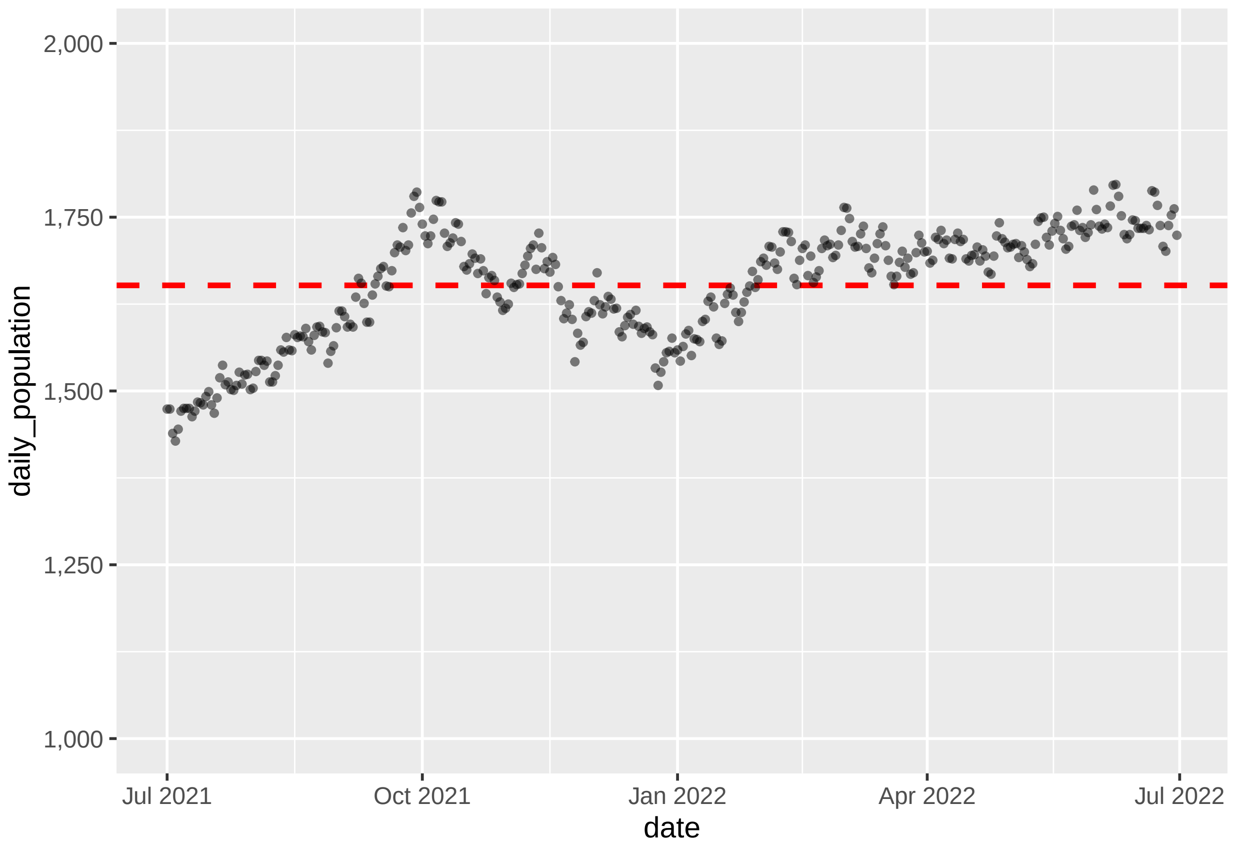
```
# count the number of people in jail each day
bookings_by_day <- booking_long %>%
  filter(date >= "2021-07-01" & date <= "2022-06-30") %>%
  arrange(date) %>%
  group_by(date) %>%
  count() %>%
  ungroup() %>%
  rename(daily_population = n)

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

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

```r
# what does the daily population look like overtime? note using a moving average
# to smooth the estimates a little
bookings_by_day %>%
  mutate(seven_day_moving_average = zoo::rollmean(daily_population, k = 7, fill = NA)) %>%
  ggplot(aes(x = date, y = daily_population)) +
  geom_point(alpha = 0.5, size = 1) +
  geom_line(aes(x = date, y = seven_day_moving_average), color = "red", size = 1.5) +
  scale_y_continuous(limit = c(1000, 2000), labels = comma)
```



```r
# View the daily population against the average daily population. Around
# what months is the daily population greater than the average?
bookings_by_day %>%
  ggplot(aes(x = date, y = daily_population)) +
  geom_hline(aes(yintercept = average_daily_population), color = "red", size = 1, linetype = "dashed") +
  geom_point(alpha = 0.5, size = 1) +
  scale_y_continuous(limit = c(1000, 2000), labels = comma)
```



---

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?

```r
# date with the lowest daily population
bookings_by_day %>%
  filter(daily_population == min(daily_population))
```

```
## # A tibble: 1 x 2
##   date       daily_population
##   <date>                <int>
## 1 2021-07-04             1428
```

```r
# date with the highest daily  population
bookings_by_day %>%
  filter(daily_population == max(daily_population))
```

```
## # A tibble: 1 x 2
##   date       daily_population
##   <date>                <int>
## 1 2022-06-08             1797
```

---

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?

```
booking_demographics <- bookings_in_analysis_period %>%
  # remove people with a missing release date
  filter(!is.na(release_out)) %>%
  # add in demographic data
  left_join(person_data) %>%
  mutate(age_at_booking = floor(interval(dob, date_in) / years(1))) %>%
  # calculate length of stay
  mutate(length_of_stay = as.numeric(release_out - date_in) + 1)
```

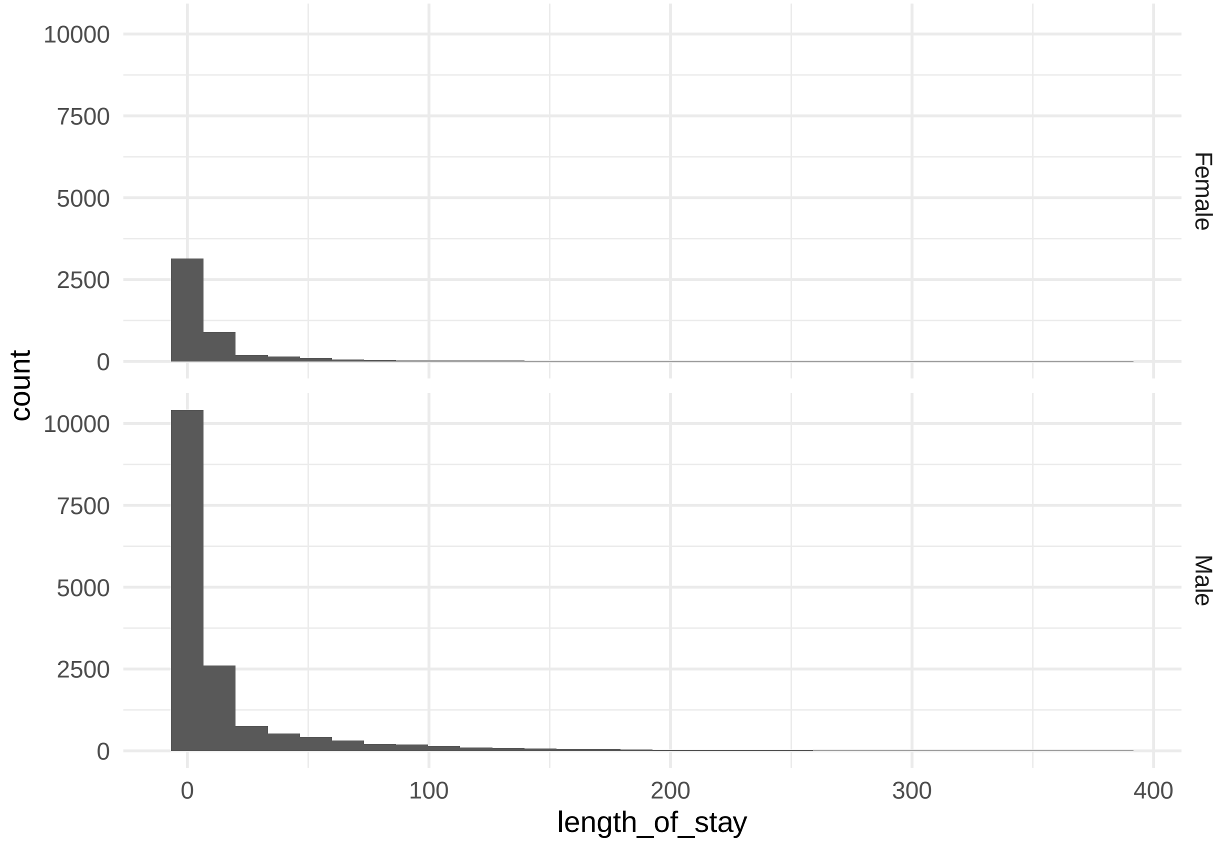
```
## Joining with ‘by = join_by(person_id)‘
```

```
# descriptive stats by gender & length of stay distribution
booking_demographics %>%
  group_by(gender) %>%
  summarise(n = n(),
            median_los = median(length_of_stay),
            mean_los = mean(length_of_stay)) %>%
  ungroup() %>%
  mutate(total = sum(n),
         pct_of_total = scales::percent(n/total)) %>%
  select(gender, n, pct_of_total, median_los, mean_los)
```

```
## # A tibble: 2 x 5
##   gender     n pct_of_total median_los mean_los
##   <chr>  <int> <chr>             <dbl>    <dbl>
## 1 Female  4690 23%                   3     12.4
## 2 Male   16128 77%                   3     17.7
```

```
booking_demographics %>%
  ggplot(aes(x = length_of_stay)) +
  geom_histogram() +
  facet_grid(gender ~ .) +
  theme_minimal()
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```

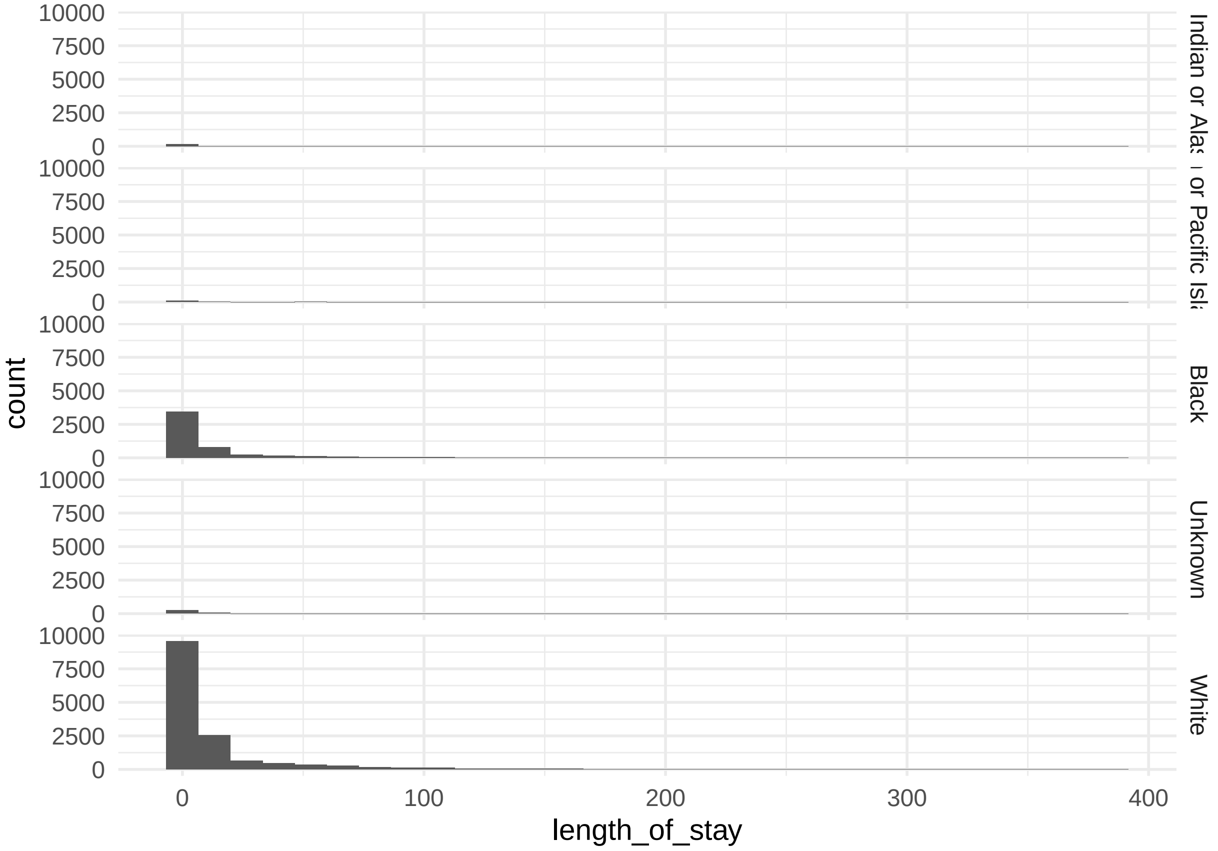


```
# descriptive stats by race & length of stay distribution
booking_demographics %>%
  group_by(race) %>%
  summarise(n = n(),
            median_los = median(length_of_stay),
            mean_los = mean(length_of_stay)) %>%
  ungroup() %>%
  mutate(total = sum(n),
         pct_of_total = scales::percent(n/total)) %>%
  select(race, n, pct_of_total, median_los, mean_los)
```

```
## # A tibble: 5 x 5
##   race                                  n pct_of_total median_los mean_los
##   <chr>                             <int> <chr>             <dbl>    <dbl>
## 1 American Indian or Alaska Native    259 1.24%               3       14.5
## 2 Asian or Pacific Islander           172 0.83%               3.5     18.8
## 3 Black                              5241 25.18%              3       17.4
## 4 Unknown                             372 1.79%               3       15.0
## 5 White                             14774 70.97%              3       16.3
```

```
booking_demographics %>%
  ggplot(aes(x = length_of_stay)) +
  geom_histogram() +
  facet_grid(race ~ .) +
  theme_minimal()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```r
# descriptive stats by age
booking_demographics %>%
  filter(!is.na(age_at_booking)) %>%
  group_by(age_at_booking) %>%
  summarise(n = n(),
            median_los = median(length_of_stay),
            mean_los = mean(length_of_stay)) %>%
  ungroup() %>%
  mutate(total = sum(n),
         pct_of_total = scales::percent(n/total)) %>%
  select(age_at_booking, n, pct_of_total, median_los, mean_los)
```

```
## # A tibble: 63 x 5
##    age_at_booking     n pct_of_total median_los mean_los
##             <dbl> <int> <chr>             <dbl>    <dbl>
##  1             18   241 1.1578%               3     14.6
##  2             19   325 1.5614%               3     11.2
##  3             20   390 1.8736%               3     11.3
##  4             21   494 2.3733%               3     13.9
##  5             22   499 2.3973%               3     16.2
##  6             23   576 2.7672%               3     13.1
##  7             24   563 2.7048%               3     14.4
##  8             25   713 3.4254%               3     15.5
##  9              26   703 3.3774%               3     14.0
## 10              27   698 3.3534%               3     15.9
## # i 53 more rows
```

```
max_los = max(booking_demographics$length_of_stay)
# by gender there is a difference in median length of stay
coin::median_test(inverse_los ~ factor(gender),
                  data = booking_demographics %>%
                    mutate(inverse_los = (max_los + 1) - length_of_stay )
                    )
```

```
##
##  Asymptotic Two-Sample Brown-Mood Median Test
##
## data:  inverse_los by factor(gender) (Female, Male)
## Z = 2.6509, p-value = 0.008028
## alternative hypothesis: true mu is not equal to 0
```

```
# by race there is a difference in median length of stay
coin::median_test(inverse_los ~ factor(race),
                  data = booking_demographics %>%
                    mutate(inverse_los = (max_los + 1) - length_of_stay )
                    )
```

```
##
##  Asymptotic K-Sample Brown-Mood Median Test
##
## data:  inverse_los by
##   factor(race) (American Indian or Alaska Native, Asian or Pacific Islander, Black, Unknown, White)
## chi-squared = 10.311, df = 4, p-value = 0.03551
```

---

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