Data Task

Arturo Bardales

2024-03-16

Description of the Task

The following is a coding task I have completed for a position's application process. The data is simulated information from a particular childcare center's work shifts and child visits. Each row in the data represents a child's stay at the center, including information on their caregiver, their check-in and pick-up times, and their caregiver's shift times. Here is a sample of the data:

Stay Num	Caregiver Name	Shift ID	Check-in Time	···
1	Andrew	17may1982 1 p.m. to 10 p.m.	1982-05-17 14:47:00	···
2	Andrew	14jul1982 1 p.m. to 10 p.m.	1982-07-14 17:49:00	···
3	Andrew	30jun1982 11 a.m. to 8 p.m.	1982-06-30 13:02:00	···
4	Andrew	21may1982 6 a.m. to 4 p.m.	1982-05-21 10:51:00	···

A few facts bear mentioning of the simulated data: if a child is not picked up by the end of their caregiver's shift, the caregiver must work overtime until that child is picked up. If a child arrives when no caregiver is available, the child must wait with an accompanying adult until a caregiver is available.

Section 1

In this section, I would like to find the percentage of children who wait for an available caregiver and the percentage who are picked up after their caregiver's shift ends.

First, we need to construct two variables for the beginning and end times of a caregiver's shift. These new variables will provide us a point of comparison to the pick up and check in times for each child.

With our new shift start and end time variables, we can detect when a child arrives before the start of someone's shift, which implies that this child waited for the caregiver. However, this does not take into account if for some reason caregivers had a limit on the number of children they can care for at any one time. If there were a limit, then children would have to wait, even if they arrive during a caregiver's shift. Hence, this estimate of the percent of children who had to wait would be a lower bound for the true percentage. If there were no limit on the number of children that a caregiver could care for, then this would be the correct percentage. I assume that if the check in time and shift start time is the same, that a child does not need to wait.

% of Visits with Child Waiting	% of Visits with Caregiver Overtime
7.36	18.59

Note that this is an estimate of the number of *children* who needed to wait only if each stay is from different children. However, we may have the same child on multiple stays, and we cannot distinguish between children in the data (besides visit number).

Section 2

In this section, I would like to calculate the number of children that arrive at the childcare center on a given day and the associated average booked stay hours.

First, we need a day variable from *check_in_time*. With this day variable, we can find the number of children seen by counting up the number of rows associated with that day and calculate the average booked hours across rows with that day.

Table 2: Number of Children Arriving at the Center on a Given Day

check_in_day	Number of Children	Mean Booked Hours	Median Booked Hours
1982-05-15	7	4.143	4
1982-05-16	146	5.164	5
1982-05-17	158	5.063	5
1982-05-18	160	5.006	5
1982-05-19	143	4.965	5
1982-05-20	146	5.027	5
1982-05-21	151	4.921	5
1982-05-22	145	5.041	5
1982-05-23	142	4.514	5
1982-05-24	171	4.702	5
1982-05-25	173	5.064	5
1982-05-26	131	4.718	5
1982-05-27	153	4.922	5
1982-05-28	106	4.491	5
1982-05-29	122	4.68	5
1982-05-30	183	4.869	5
1982-05-31	150	5.133	5
1982-06-01	130	5.092	5
1982-06-02	147	4.81	5
1982-06-03	153	4.758	5
1982-06-04	146	4.781	5
1982-06-05	143	4.811	5
1982-06-06	176	4.795	5

check_in_day	Number of Children	Mean Booked Hours	Median Booked Hours
1982-06-07	137	4.715	5
1982-06-08	119	4.857	5
1982-06-09	128	5	5
1982-06-10	136	5.147	5
1982-06-11	149	4.792	5
1982-06-12	154	4.721	5
1982-06-13	92	4.978	5
1982-06-14	163	5.049	5
1982-06-15	164	4.933	5
1982-06-16	146	4.973	5
1982-06-17	156	4.75	5
1982-06-18	160	5.112	5
1982-06-19	145	4.69	5
1982-06-20	149	4.993	5
1982-06-21	145	4.586	5
1982-06-22	127	4.937	5
1982-06-23	151	5.079	5
1982-06-24	141	5.163	5
1982-06-25	142	5.028	5
1982-06-26	130	4.985	5
1982-06-27	163	4.761	5
1982-06-28	116	5.086	5
1982-06-29	155	4.69	5
1982-06-30	131	4.725	5
1982-07-01	136	4.89	5
1982-07-02	153	4.778	5
1982-07-03	138	4.826	5
1982-07-04	126	4.849	5
1982-07-05	140	4.814	5
1982-07-06	136	4.985	5
1982-07-07	126	5.302	5
1982-07-08	162	5.185	5
1982-07-09	161	4.727	5
1982-07-10	150	4.707	5
1982-07-11	149	4.946	5
1982-07-12	145	4.71	5
1982-07-13	145	4.966	5
1982-07-14	155	4.735	5
1982-07-15	128	4.883	5

time	children	caregiver	day
1982-05-17 13:00:00	3	Andrew	1982-05-17
1982-05-17 13:10:00	4	Andrew	1982-05-17
1982-05-17 13:20:00	5	Andrew	1982-05-17
1982-05-17 13:30:00	5	Andrew	1982-05-17

Section 3

In this section, I want to calculate the number of children under a caregiver's supervision at any moment in time during the caregiver's shift.

I will create a time series dataframe containing a "census" count of the number of children under supervision at any given time for every 10 minutes in a caregiver's shift. To do this, I will create a list of dataframes containing each unique combination of caregiver and shift.

I will then create a 10-minute time sequence for each caregiver and shift. For each sequence, I will also record the number of children being cared for by that caregiver at each 10-minute point. I will store this information, along with the caregiver and shift start day of the sequence, in a nested list.

Afterwards, I rearrange and condense the nested list. Now, rather than a nested list that corresponds to each unique combination of shift and caregiver, I will make the list elements unique to each caregiver, and each list element containing a list of information pertaining to their shifts.

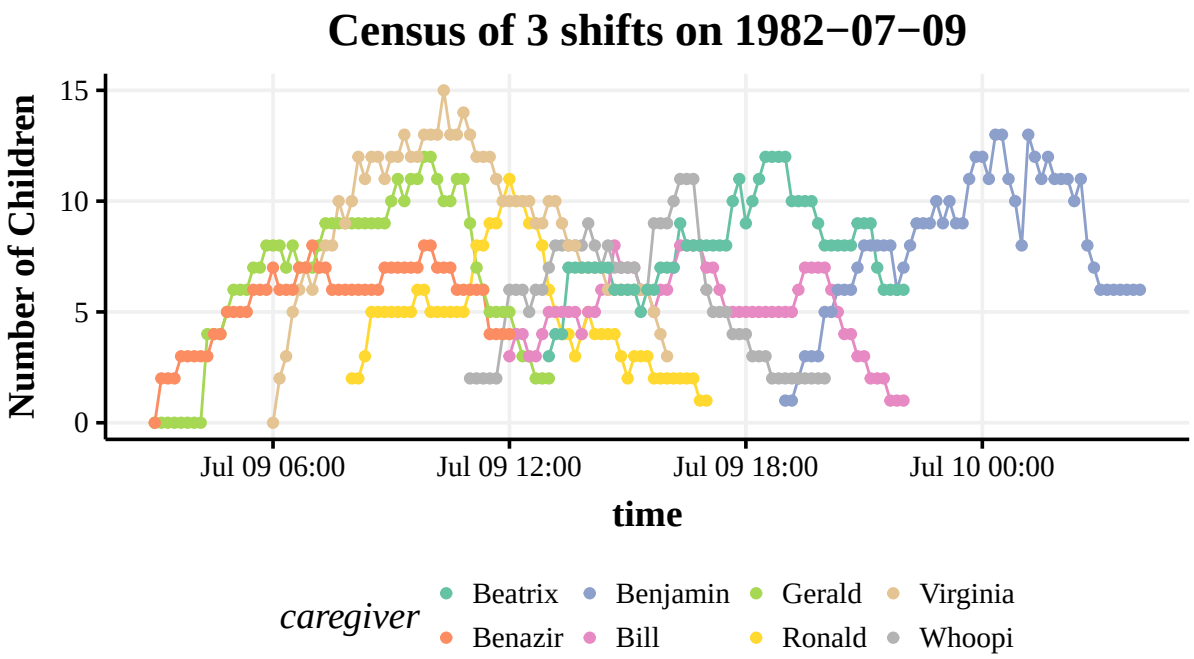
With this rearranged list, I now convert each list element in the nested list to a dataframe of time sequence, children census, caregiver, and shift start day.

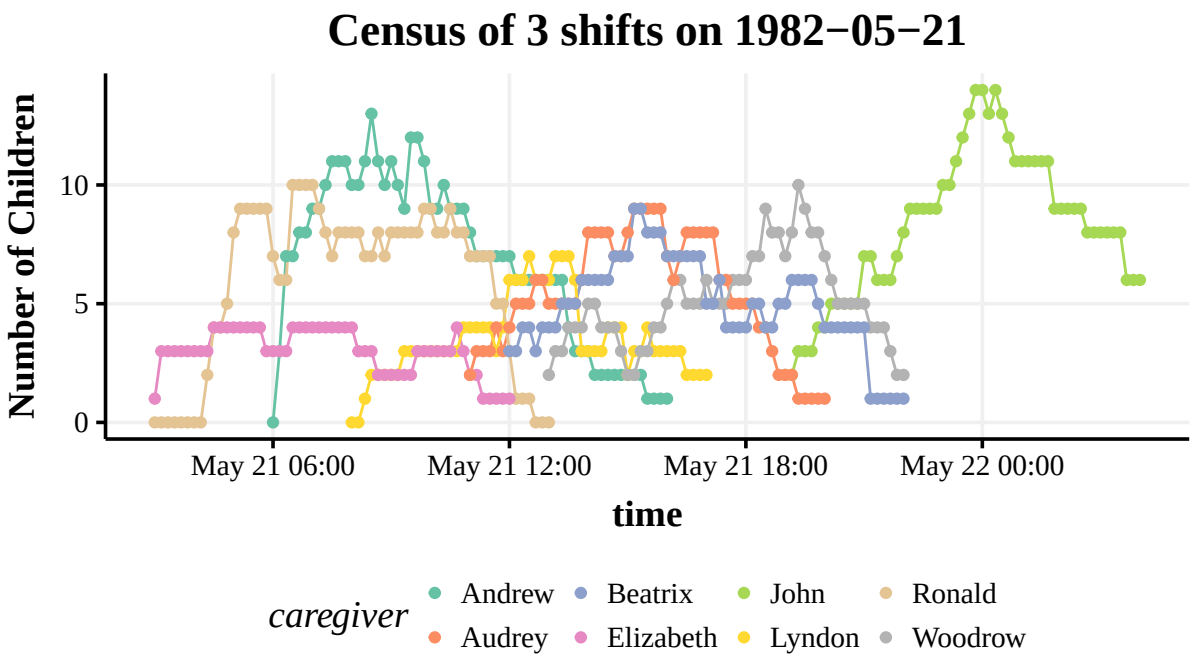
Finally, we can simply rowbind all the dataframes to generate our final census dataset.

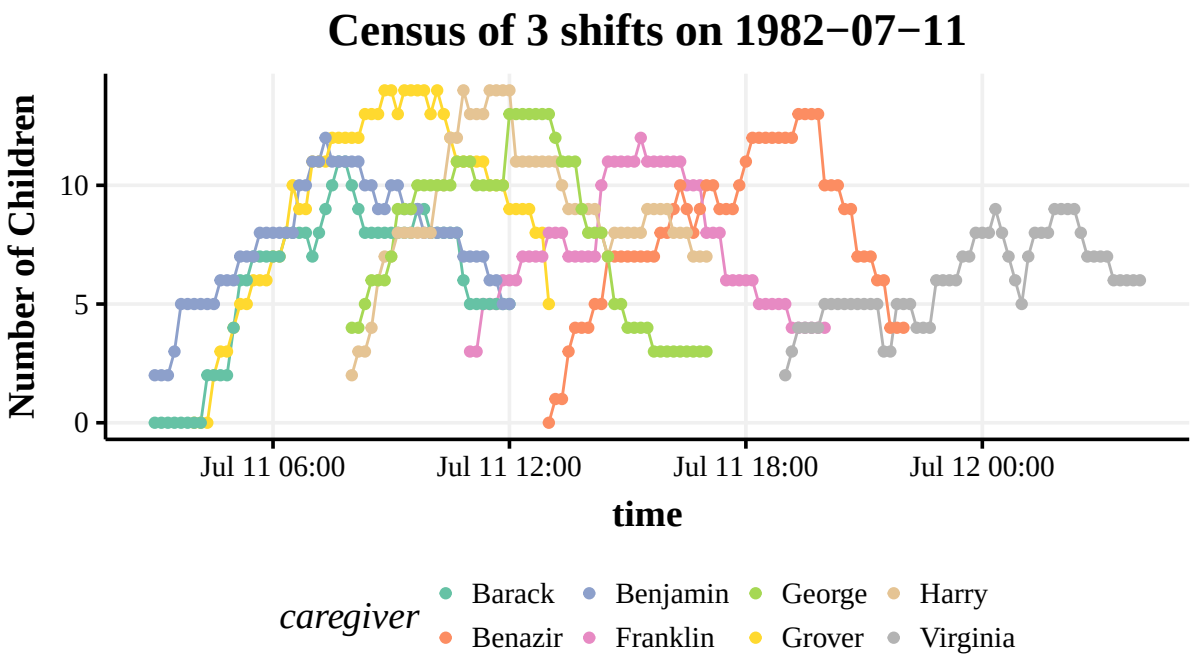
The dataset is of the following form:

where *time* represents a time value, *children* represents the number of children being cared for by a given caregiver at the *time* in the same row, and *day* is the shift start day of the observation. *day* and *caregiver* uniquely identify all shifts. Now, we can plot a small graph of a few caregivers on (a) certain day(s). Let's graph 3 random days as an example.

The following line plots represent the census count of children at any given 10-minute mark within a caregiver's shift. Each line corresponds to a different caregiver. Each plot corresponds to a different day. The lines simply connect the observation scatters and do not represent a model fit.


Census of 3 shifts on 1982−07−09
Number of Children
15
10
5
0
Jul 09 06:00
Jul 09 12:00
Jul 09 18:00
Jul 10 00:00
time
caregiver
Beatrix
Benjamin
Gerald
Virginia
Benazir
Bill
Ronald
Whoopi


Census of 3 shifts on 1982−05−21
Number of Children
10
5
0
May 21 06:00
May 21 12:00
May 21 18:00
May 22 00:00
time
caregiver
Andrew
Beatrix
John
Ronald
Audrey
Elizabeth
Lyndon
Woodrow



As the shift closes, on average for any given caregiver, the census number increases up to a certain point and then decreases until the end of the shift. This trend seems almost universal, and no one seems to have more than one general peak census area or a particularly large number of children at the end of their shift (compared to the rest of their shift) – most children do seem to be getting picked up before the end of a caregiver's shift. However, it seems interesting that caregivers must sometimes stay after their shift to care for a child who has not been picked up. The graphs show tremendous overlap between caregiver shifts. So why not shift responsibility for that child to the other caregiver? There does not appear to be a rule for how many children a caregiver can take care of at any given moment. At least, it is not clear from three graphs. Perhaps there is a limit around the maximum observed in the data.

At any 10-minute mark (including start and end points) in a caregiver's shift, the maximum number of children observed under a caretakers supervision in the data is 17.

Alternatively, the maximum possible number of children that a caregiver can see throughout their entire shift is 30.

Section 4

In this section, I want to explore whether parents are more likely to be late to pick up their child on certain weekdays and whether this difference is statistically significant.

We need to create a variable that shows the amount of time elapsed between when a child was left with the caregiver and their pickup time. Importantly, the possibility of children

arriving early, before a shift starts, needs to be considered, otherwise we overcount the time since the time elapsed between the early arrival and shift start would signify time where the child spent waiting with their parent, guardian, adult, etc.

To determine if someone was late to pick up their child, I will define late as picking up the child any time past the number of booked hours (for example, 1 minute late counts).

I will create two different days of the week variables: one will signify the day of the week when the child was dropped off, and the other will signify the day of the week when the child was picked up. These two variables should have tremendous overlap, but perhaps the subset of people that pick up their child on a different day shows an interesting pattern.

Weekday of Checkin	Percent Late
Sunday	30.09
Thursday	24.03
Wednesday	22.26
Monday	22.11
Tuesday	22.08
Saturday	21.16
Friday	20.46

(a) Check in

Weekday of Checkin	Percent Late
Sunday	30.77
Thursday	23.15
Wednesday	22.67
Monday	21.83
Friday	21.76
Tuesday	21.68
Saturday	20.53

(b) Pick up

Table 3: Likelihood that a child is picked up late

Regardless of whether we use the weekday of pickup and the weekday of checkin, we obtain similar results for the most likely days for a child to be picked up late. Specifically, the top four most likely days for a child to be picked up late, from largest to smallest likelihood, is Sunday (~30%), Thursday(~24%), Wednesday(~22%), and Monday. Sunday has a large lead: there is a difference of roughly $6 - 10\%$ between the Sunday and the other weekdays. For Thursday, there is a notable but smaller difference: $1 - 4\%$. Where the last 3 days of the week rank in the likelihood of a child being picked up late depends on the weekday variable.

Table 4: Relationship between Weekday and being picked up late at the Childcare center

	(1)
(Intercept)	−1.358***
	(0.073)
check_in_wdayMonday	0.099
	(0.098)
check_in_wdaySaturday	0.043
	(0.103)
check_in_wdaySunday	0.515***
	(0.094)
check_in_wdayThursday	0.206*
	(0.097)
check_in_wdayTuesday	0.097
	(0.098)
check_in_wdayWednesday	0.107
	(0.099)
Num.Obs.	8831
AIC	9552.1
BIC	9601.7
Log.Lik.	−4769.046
F	7.544
RMSE	0.42

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Each variable in the regression represents a binary variable for whether a given child stay happened on the corresponding weekday. For example, ' '$Monday$ represents a binary variable for whether a stay occurred on Monday. The dependent variable is a binary variable for whether the child was picked up late, relative to their booked hours.

Whether the difference between the most likely weekdays and less likely weekdays is statistically significant can be ascertained through a regression relating weekday and being late. The following is a logistic regression with uncorrected standard errors. Notice that the coefficient estimate for the binary Sunday variable is statistically significant at the 0.1 level, and the coefficient for Thursday is statistically significant at the 0.05 level. This indicates that the difference in likelihood of a child being picked up late on these days compared to other days is statistically significant. There is stronger evidence this is the case for Sunday.

Section 5

I would like to analyze the simulated data and compare its patterns to expectations we might have of real data.

The amount of activity the childcare center receives throughout all hours of the day seems unrealistic. While a small number of hours near the start of the 24-hour day seem to not garner any child drop-offs, in general the rest of the day appears more busy than it would normally during late or early hours. In question 3 plots, the people working around midnight seem to have almost as much, if not more, activity than some people during daylight hours.

In question 3 line plots, each person seems to have their own fair share of the childcare responsibility, which might also be unreasonable to assume. Maybe one person at the daycare center takes on much more responsibility on a given day, or there are different tasks that each caregiver does which would also affect their workload and census (a caregiver for disabled children, for example). In a similar sense, it seems strange that a caregiver may work past the end of their shift because of a child who has not been picked up. Judging from the graphs, many caregivers work overlapping shifts, and often there is another caregiver with a manageable number of children (compared to the rest of their shift). At least, one would expect that caregivers would offload children to each other.

Each person's census for a given shift appears to have a unimodal or near-unimodal distribution, and each person's peak typically appears around the middle of their shift. In real life, there may very well be more than one peak in the data and at more varied times. For example, one peak in the morning and another peak later in the day could be possible, and the peaks would not seem to depend on the amount of time elapsed in a person's shift (peaks in middle) as it seems to do in the data. For example, seasonal variation and parent work hours would affect these peaks. Along those lines, the average booked hours seem to not vary much across days of the week and month, as we saw in question 2. In real data, there likely would be more busy and less busy days across the month and week.

Finally, there seems to be a lot of children who are being picked up late and forcing the caregiver to work overtime. The frequency (18.6% of visits) perhaps is larger than it would be in a real dataset.