Statistical Consulting &
Research Services

Montana State University

Statistical Consulting and Research Services

Missing Persons Reports 2017 - 2019 Data Analysis

Prepared for: Tina Chamberlain, MT DOJ

Greta Linse, Assistant Director

Dr. Mark Greenwood, Director

July 31, 2020

Author Note:

This material is provided to communicate advice from SCRS statisticians based on our best understanding of your research needs. We encourage you to use this report in discussions with colleagues. Please do not publish any portion of this material without permission.

This report is prepared to support a report prepared for the Montana Legislature. The work completed by the Statistical Consulting and Research Services at Montana State University was funded by the Montana Department of Justice. The source code is available in an R-markdown that documents the data analysis steps, which also contains comments within the code that can further explain the results contained here. The main purpose of this report is to generate the tables and figures for use in the main report and provide some additional details and results in a few areas. This work should be considered only as a supplement to the main report.

Data Analysis and Validation of Missing Persons Data from 2017-2019

Overall Summary

After detection of individuals with alternative spellings of their names the data from 2017-2019 consisted of 5,570 entries, and 3,254 unique individuals. Of those individuals who were reported missing, 957 were reported missing more than once and 2,297 were reported missing only once.

Table 1: Counts of individuals as in the database only once or more than once.

NUMBER OF TIMES IN DATABASE	FREQUENCY
Multiple Entries	957
Single Entry	2297

Time missing summary statistics, graphics, and analysis

This section explores the reported times missing and the impacts of the censoring of the time missing for those that have not been found in the time frame of observation on the estimated median and mean time missing.

Censoring of timing of events can lead to biased estimates of summary statistics since the true length of missingness is unknown for some observations, but, in this case, the censoring is of the time of resolution of the case, so its minimum length of time is known. This analysis explores that result for all records in the data set from January 1, 2017 to December 31, 2019. Further analyses could explore just the first or last time of missing for each unique subject but are left for future possible work. For events where the report and resolution were on the same day, the length of the event was recorded as 0. All other records record the whole number for the number of days missing, except where the event was not resolved in the time of observation (up to December 31, 2019). Note that some events were resolved in early 2020 but that information is not included in these analyses.

Also, ten records contained event report dates after resolution dates. For intervals of less than one day (seven records), the resolution date was set to the following day. The remaining three records are being investigated further and currently are removed from the following analyses.

In the first subsection, we explore the impacts of treating the right censored (unresolved) event end points as missing or “imputing” (filling in) the values with the longest time possible to be observed (up to three years but often much less than that) and using conventional median and mean estimators (“naive method”). These are biased estimates, especially for the mean, because the missing times (which can be long) are either deleted or filled in with possibly large but still smaller than they should be.

In the next subsection, we take a more sophisticated approach to estimate the probability of being missing as a function of time which then allow us to estimate the mean and median missing times in an unbiased fashion, and results are compared to the previous approaches. The techniques used here fall under the domain of “Survival Analysis” and have been developed for modeling times to events where sometimes the event is not observed in the time that is available to wait to observe it. They go by many names in different fields and possibly “duration analysis” is the best term for the analysis being performed here given the results.

In the final subsection on these analyses, we attempt to better estimate the probability of missing for the short duration events by using the record starting time of day. Because of the outcome times were not specific by time of day in the database, these are "interval censored" to have had their outcome within the day of that report (if an outcome was obtained). The interval censored statistical methods are much more complicated but yield similar results to those from the results where only the day of events was used, so are not pursued in as much detail.

Simple estimate of summary of time missings

Table 2: Table of summary statistics for missingness durations using simple estimates.

	MIN	Q1	MEDIAN	Q3	MAX	MEAN	SD	N	MISSING	SE
Observed intervals	0	0	1	5	486	10.40	30.88	5422	145	0.42
Filled with last date intervals	0	0	1	6	1050	18.36	71.98	5567	0	0.96

The table of results shows summary statistics for the time missing for just the intervals where the endpoints are known ("Observed intervals") and treats the censored end point observations as completely missing records (so are not included in the results). The estimated median is 1 and the mean 10.4 days (standard deviation (SD) 30.9, standard error (SE) 0.42) from the 5,420 non-missing records (147 (2.6%) were missing). Filling in the missing endpoints with December 31, 2019 provides an estimated median of 1 day and mean of 18.36 days (SD 72.3, SE 0.96), with no missing values (aside from the three removed for negative intervals). The second set of results is likely better but still does not fully account for the impacts of the missing end points on the mean estimator.

Another option is just to categorize the time missing into categories of lengths of time. The censored endpoints still create issues as to which category is correct for those observations unless the minimum time definitely places the observations in the highest time length category. So two approaches are considered again, treating the endpoints as missing and filling them in with the last observation date. Categories are created of 0, 1-7, 8-30, 31-365, over 365 days, and (when not imputed) missing ("NA").

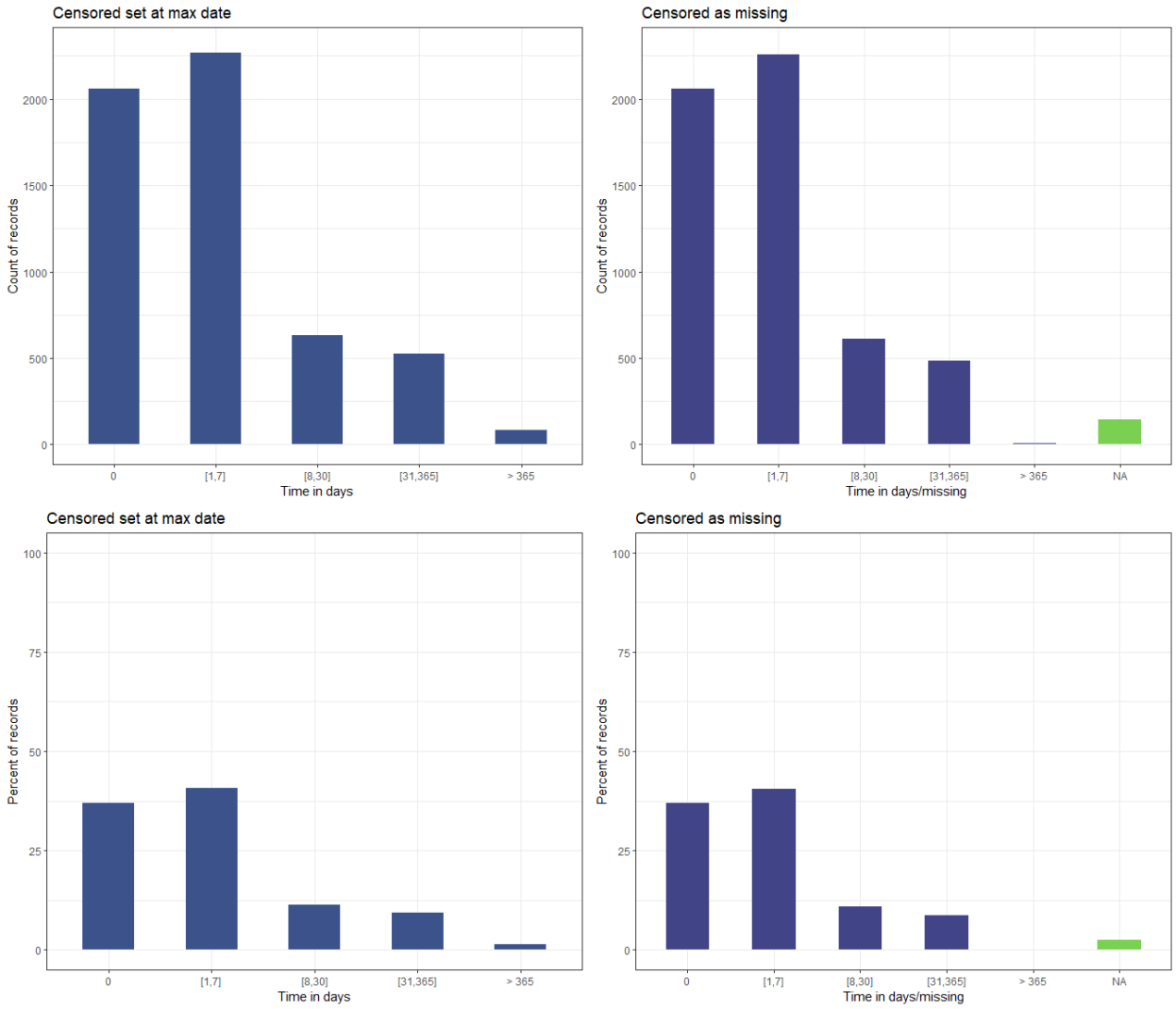

Figure 1: Display of duration missing counts (top row) and percentages (bottow row) based on two approaches to handling censored endpoints of duration times. NAs correspond to observations with no observed endpoint before 12/31/2019.

Table 3: Table of counts and percentages of total based on different methods for handling censored endpoints for time intervals.

	0	[1,7]	[8,30]	[31,365]	> 365	NA
Counts, Censored set at max date	2061	2265.0	633.0	523.0	85.0	0.0
Counts, Censored as missing	2061	2258.0	613.0	484.0	6.0	145.0
Percents, Censored set at max date	37	40.7	11.4	9.4	1.5	0.0
Percents, Censored as missing	37	40.6	11.0	8.7	0.1	2.6

In the figures and the table that contains the same information, we can see that the highest percentage of cases are resolved within either the same day or within seven days. The way of handling the censored endpoints has the greatest impact on the higher time length observations.

This is reinforced in looking at the plots of the distributions of the responses (tick marks) in the following modified density curve plots.

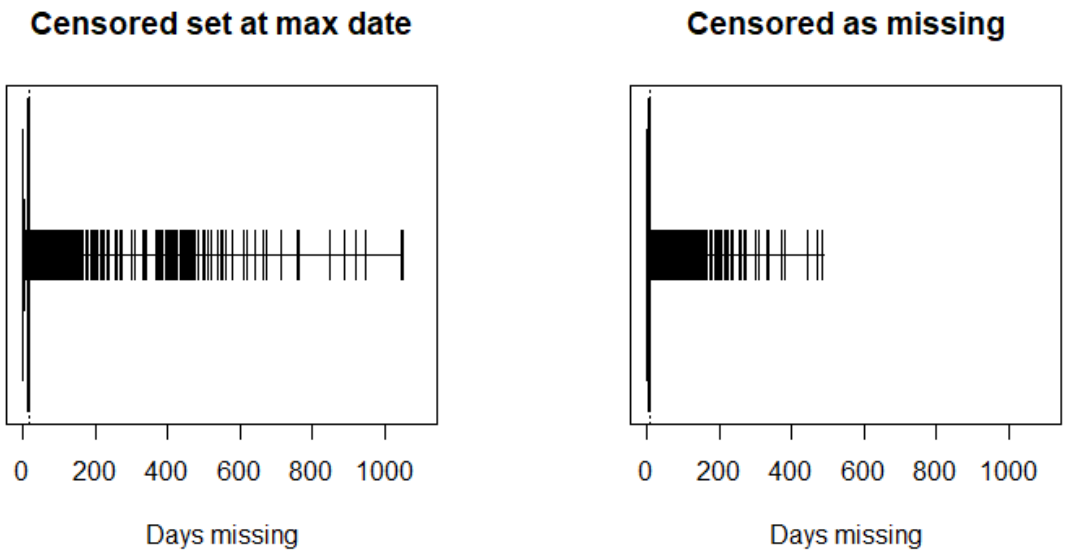

Figure 2: Plots of the observed times of duration missing using two methods of handling unobserved endpoints for duration times. In the right panel, unobserved endpoint observations are not displayed. Bold lines close to 0 days are the means of the displayed observations in each plot.

Duration time for missingness analysis

As noted above, there are statistical methods that directly incorporate the known information (non-censored) and the partial information (censored) to better estimate the probability of a subject being missing as a function of the time missing. The estimated distribution can then be used to estimate the median and mean times. This method uses Kaplan-Meier curves (Kaplan and Meier, 1958) to start treating any observations with unobserved resolutions by 12/31/2019 as right censored. Additionally, because only the day of the resolution is known in the database and not the detailed time of day, all events for this analysis are rounded to the nearest day (with any records with length less than 24 coded as 0 days). These methods also provide the beginning of a framework for analyzing and comparing time to event across different groups, such as by gender, as is demonstrated below. These analyses are performed using the R packages “survival” and “survminer”.

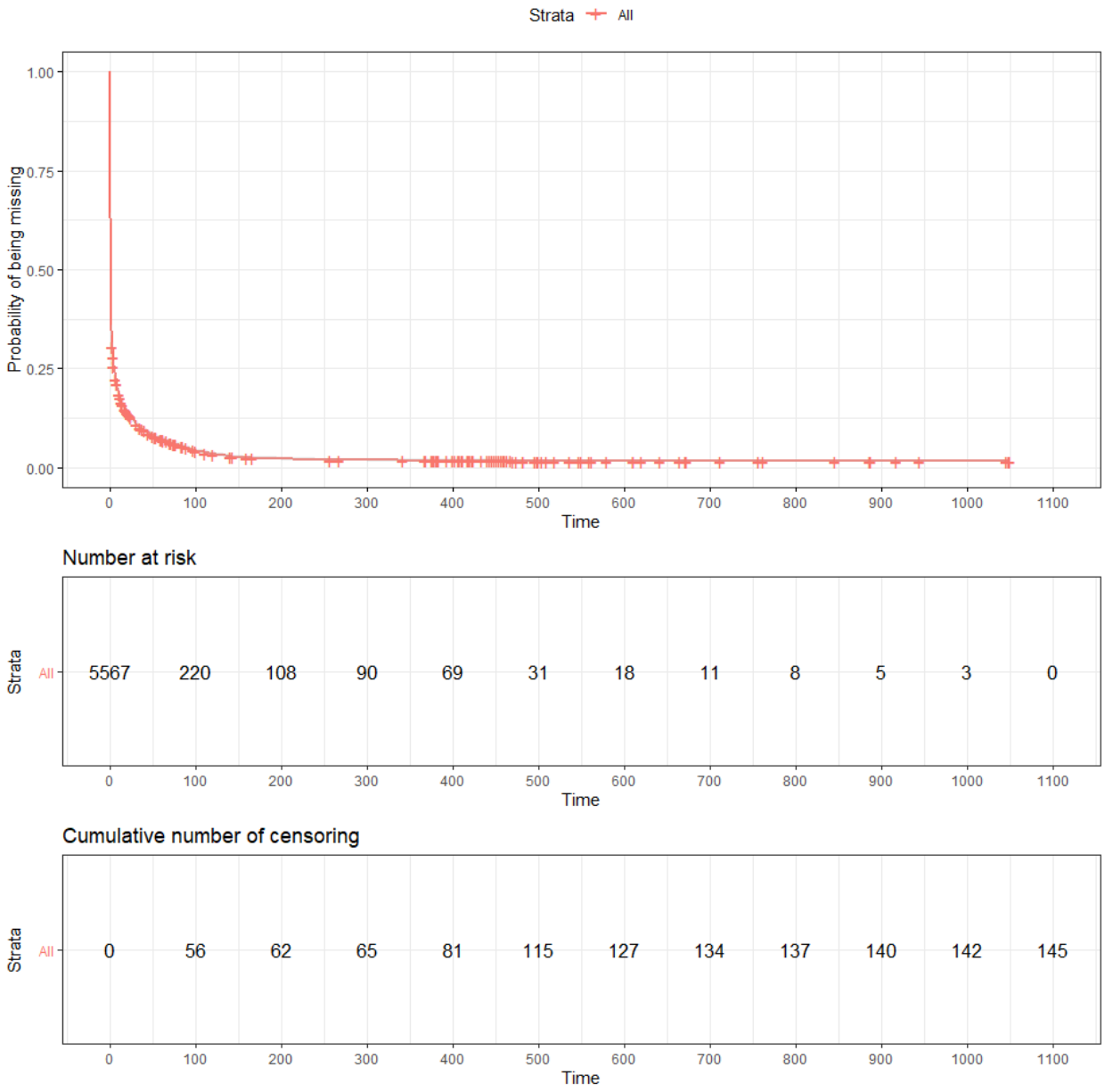

Figure 3: Probability of missing estimates as a function of days missing. Small pluses are censored observation times. The probability of being missing drops rapidly in the first days. Display also contains information on the number of observations available and censored in the analysis.

Table 4: Table of summary statistics from the Kaplan-Meier curve.

SUMMARY STATISTIC	VALUE
records	5567.000000
n.max	5567.000000
n.start	5567.000000
events	5422.000000
*rmean	28.545016
*se(rmean)	1.934788
median	1.000000
0.95LCL	1.000000
0.95UCL	1.000000

The median using the duration analysis properly handling the censoring of times to events provides a median of 1 day missing. The 95% confidence interval goes from 1 to 1, which is a bit of an odd result to have an interval with width of 0. This happened partially because this is such a large data set and partially because the time is rounded to whole days in this analysis (an alternate method is discussed below to avoid this rounding). This matches the median in the naive estimator using either only the complete or filled in times for the censored observations. Because it is just a measure of the middle observation and the censoring mainly impacts the long length records, this result is not surprising and also agrees with the exploration of percentages in binned categories of days performed previously.

The mean time missing is estimated to be 18.36 days using the conventional mean estimator with the last observed day to create durations for censored observations. Using time to event with censoring incorporated with the Kaplan-Meier curve, the estimated mean is 28.55 days (SE of 1.93), which is a dramatic increase in the estimated mean length of missing durations when handling the censoring correctly. Also note that the SE is also much larger at nearly 2 to reflect a less precise estimate when incorporating this information than when having complete case information used.

Missing time duration analysis by gender

It is also possible to use similar methods to compare groups for different estimated duration curves. For example, the comparison of estimated probability of missing as a function of time between male and females does not suggest clear differences in the plot. It is also possible to perform a hypothesis test for a difference between the groups and there is limited evidence against the null hypothesis of no difference in the true curves for the two groups with a p-value of 0.42. The medians for both groups are 1 day and the estimated mean time missing is 27.21 days in the female group and 29.95 days in the male group. Future research can explore these sorts of differences in reported missing times across various demographics of subjects.

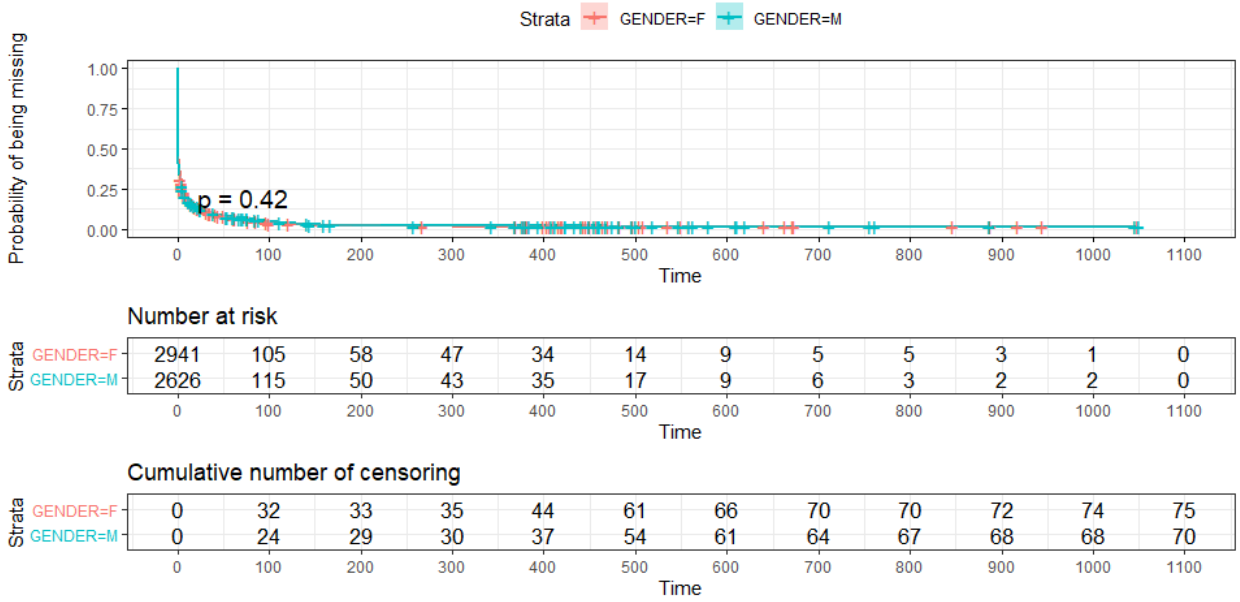

Figure 4: Kaplan-Meier curves for male and females. Differences are barely noticeable in the plot.

Table 5: Table of summary results for comparison of male and female reported cases.

GENDER	RECORDS	N.MAX	N.START	EVENTS	*RMEAN	*SE(RMEAN)	MEDIAN	0.95LCL	0.95UCL
GENDER=F	2941	2941	2941	2866	27.20904	2.573479	1	1	1
GENDER=M	2626	2626	2626	2556	29.94818	2.909762	1	1	1

Using time of reporting and interval censored endpoints

Because many of these events are of short duration, it is possible that incorporating the exact reporting time would improve the estimation of the duration probabilities. This would be most impactful on the estimates for a small number of days. There is a possible concern that the initial recorded time is only a proxy for the actual time of missing starting and so the extra work for this approach (and its limited gains/changes in results) is not necessarily suggested. This methods uses an EM-algorithm based on the interval censoring of knowing that the even resolved itself sometime during the day where the even was cleared, if it is known. It is able to more precisely work with the short time interval responses and so do better at estimating percentiles of missing proportions within the first day or two. It is much more sophisticated computationally and results on the "Icens" R package. Results are commented out in this file because of the length of time required to estimate the model. The provided plot shows that results are not much different from the previous simpler methods. For more information on interval-censored data analysis see Bogaerts, Komarek, and Lesaffre (2017).

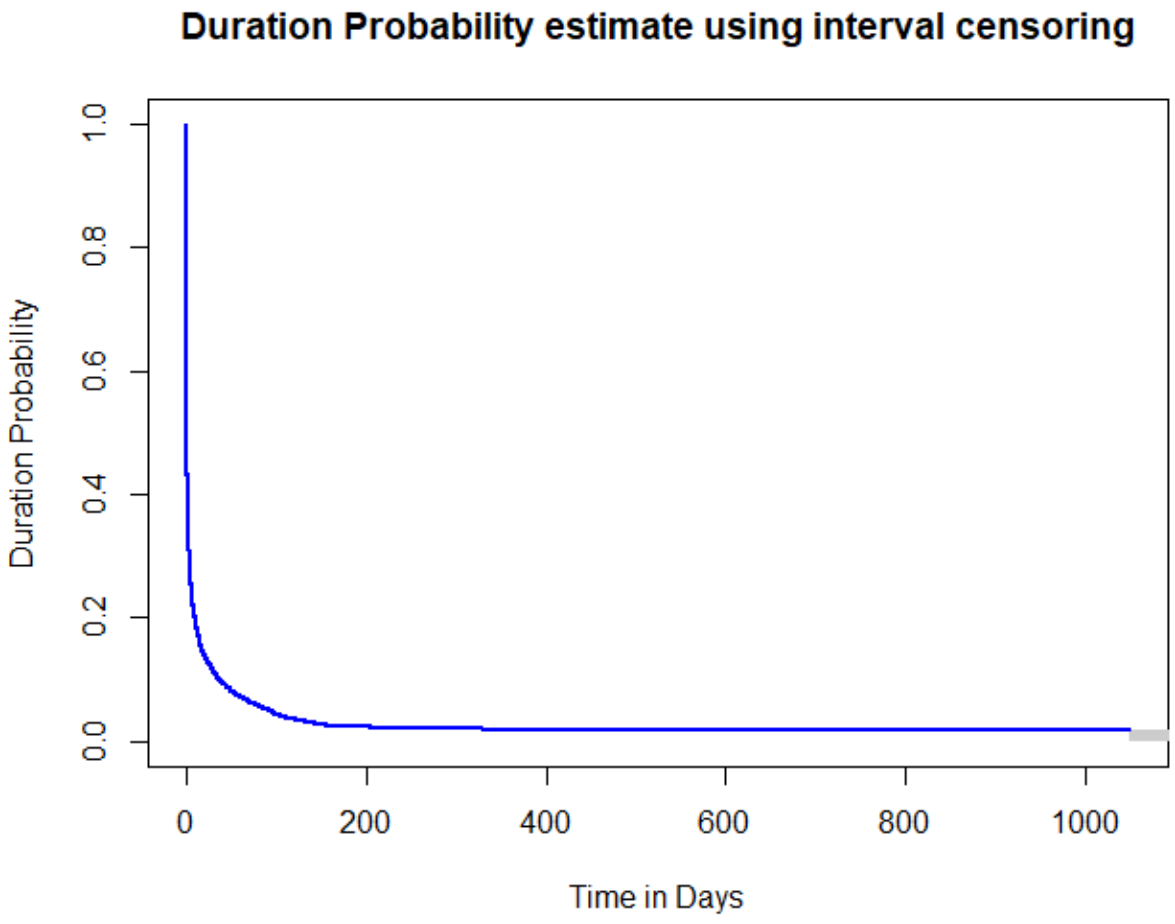

Figure 5: Interval and right censored EM-algorithm estimate of duration.

The time of the median duration is around 0.67 of a day, so slightly lower in this approach than the rounded day version of the durations. Estimation of a confidence interval for the median or an estimate

of the mean is not readily available in the software being used and so is not reported. It looks like the estimated curve is similar to the previous results where those additional results are readily available.

Scope of Inference/Limitations

It is important to recognize some constraints on these results. First, the methods used for duration of missingness assume that all records are independent and there are multiple observations for many subjects in the database. Further analyses could explore the first or last response for each subject to avoid this issue, but that would modify the group inferences are being done for slightly and could introduce a different sort of bias (shorter or longer lengths if the first or last of multiple is chosen). Future work could engage more sophisticated statistical models that account for repeated measures of duration events across subjects, as these were not considered or explored in detail for suitability with these data.

The results are only as good as the data going into the analysis. There are two potential issues with the data quality that can't be addressed here. There may be records that should have been included and were not incorporated, missing reports completely. The records here could be inaccurately reported, especially for the detailed times of events since these are times of reporting and might not match the time of the actual events (both in the start of a missing event being at or before the reported time and the end being at or possibly before the reported clearing of the event), and the information could be inaccurately entered/transferred across databases. One of the main tasks here was to assess data quality and some minor coding errors were encountered across all fields including the time of events, but other issues and errors with reporting could easily have been missed.

The estimates provided then only apply to the records that were included in the database, so would not generalize to any events that were not in the database as these could be systematically different from these. And these only pertain to events in the three year period being assessed and do not extend to other time periods or locations that were not part of the reporting process to generate this database.

A multi-phase approach was taken to identify individuals in the database that had entries with an alternate spelling of their name. Special characters and spaces were removed before identifying similar names. The first phase split individuals by reported gender and then binned by birth date with (-Inf, 1990], (1985, 1998],(1998, 2000], (2000, 2002], (2002,2004], (2004, 2006], (2006, 2008], (2008,2010], (2010, 2020]. Within these groups, all pairwise comparisons on the names were formed and string comparisons of Jaccard and Jaro-Winkler distances were used to identify if a name was likely to be similar. We then went through the lists by hand and identified whether or not the names were likely misspellings of each other and thus the same individual. These alternate spellings were recorded in a separate spreadsheet.

To account for data entry errors the same process was repeated with different splits for the birth years. The years were split based on (-Inf, 1985], (1985, 1995],(1995, 1999], (1999, 2001], (2001,2003], (2003, 2005], (2005, 2007], (2007,2009], (2009, 2020]. In addition, names were checked to see if they were exact substrings of one another, *i.e.*, "SMITH,BOB" is an exact substring of "SMITH,BOB M.".

Again, these lists were checked by hand to confirm identification of duplicate/alternate spellings of names.

Finally the list we we came up with was checked against the list provided by Tina Chamberlain and further names were identified at this pass. In the end, there were 3254 unique names.

As birth dates were misspelled, at this time, only the birth date for the first entry (earliest entry) was recorded as the birth date. Further work can take the consensus birth date if there are more than two entries, we can select the birth date recorded most often. Most of the time, it is the month and day that was entered wrong and only a couple instances had different birth years.

Other cross-tabulation validations from the unique subjects data set:

The following tables and figures were created based on the summary statistics reported in the draft report.

Gender

Table 6: Counts of Unique Individuals by Gender

GENDER	COUNT
F	1673
M	1581
Total	3254

Table 7: Proportions of Unique Individuals by Gender

GENDER	PROPORTION
F	0.5141364
M	0.4858636
Total	1.0000000

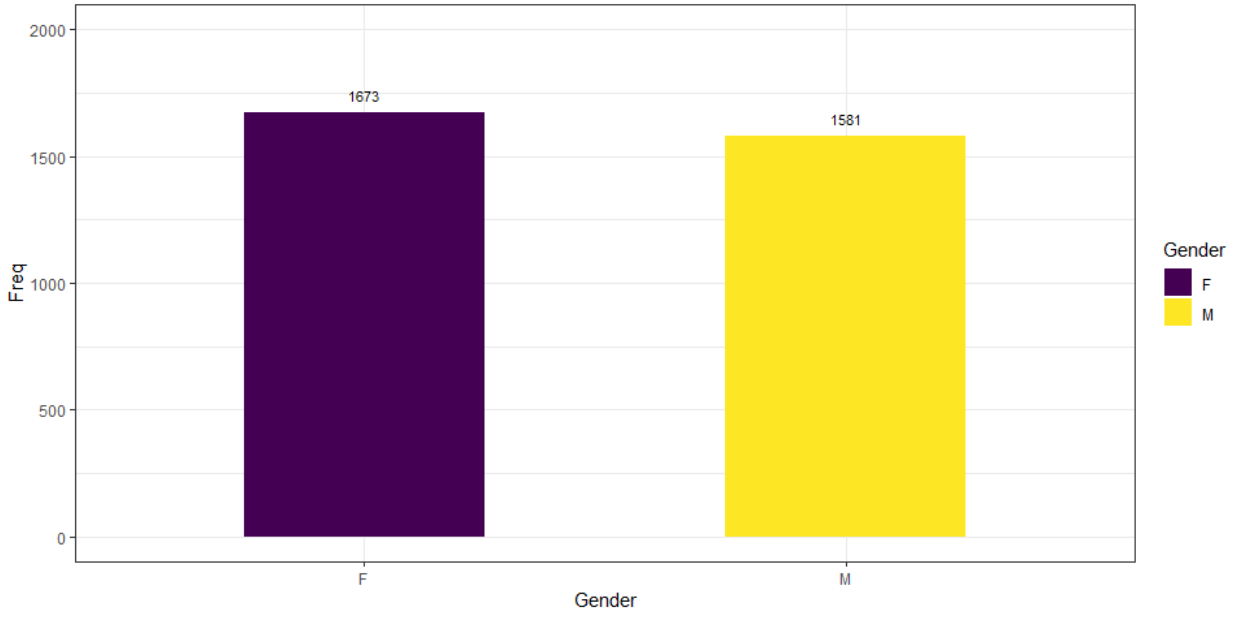

Figure 6: Number of missing persons by gender.

Table 8: Counts of unique individuals by gender and the year of the first time the individual was reported missing.

	2017	2018	2019
F	647	523	503
M	565	510	506
Total	1212	1033	1009

Table 9: Proportions of unique individuals by gender and the year of the first time the individual was reported missing.

	2017	2018	2019
F	0.5338284	0.5062924	0.4985134
M	0.4661716	0.4937076	0.5014866
Total	1.0000000	1.0000000	1.0000000

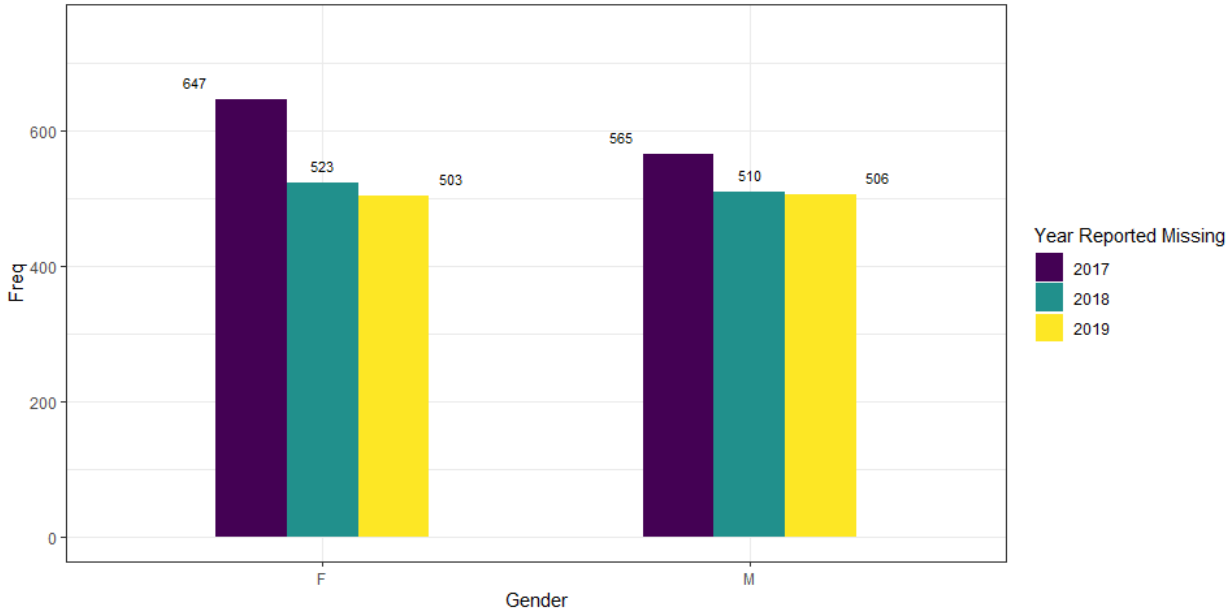

Figure 7: Number of missing persons by gender by year of first time reported missing.

Table 10: Counts of Indigenous missing individuals by gender and the year of the first time the individual was reported missing.

	2017	2018	2019
F	185	142	164
M	107	103	129
Total	292	245	293

Table 11: Proportions of Indigenous missing individuals by gender and the year of the first time the individual was reported missing.

	2017	2018	2019
F	0.6335616	0.5795918	0.559727
M	0.3664384	0.4204082	0.440273
Total	1.0000000	1.0000000	1.000000

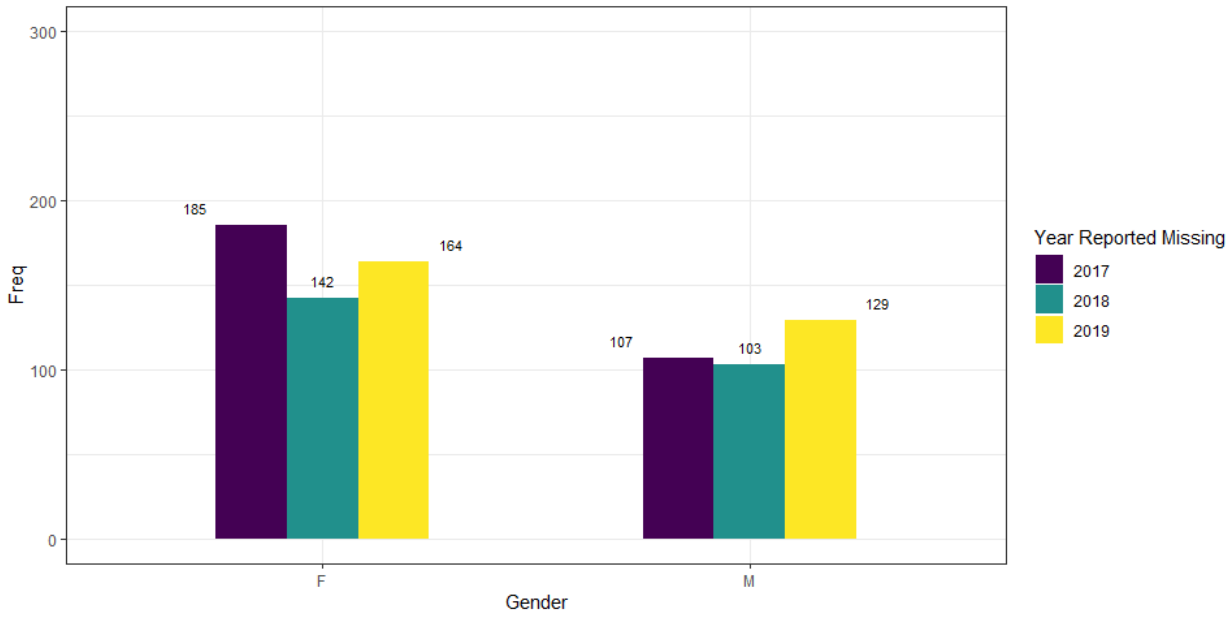

Figure 8: Number of Indigenous missing persons by gender by year of first time reported missing.

Age

Table 12: Counts of unique individuals by gender and their age for the first time the individual was reported missing.

	0-17	18-21	>21
F	1391	38	244
M	1230	38	313
Total	2621	76	557

Table 13: Proportions of unique individuals by gender and their age for the first time the individual was reported missing.

	0-17	18-21	>21
F	0.5307135	0.5	0.438061
M	0.4692865	0.5	0.561939
Total	1.0000000	1.0	1.000000

Table 14: Counts of unique individuals by the year of the first time the individual was reported missing by age groups.

	2017	2018	2019
0-17	1006	844	771
18-21	26	23	27
>21	180	166	211
Total	1212	1033	1009

Table 15: Proportions of unique individuals by the year of the first time the individual was reported missing by age groups.

	2017	2018	2019
0-17	0.8300330	0.8170378	0.7641229
18-21	0.0214521	0.0222652	0.0267592
>21	0.1485149	0.1606970	0.2091179
Total	1.0000000	1.0000000	1.0000000

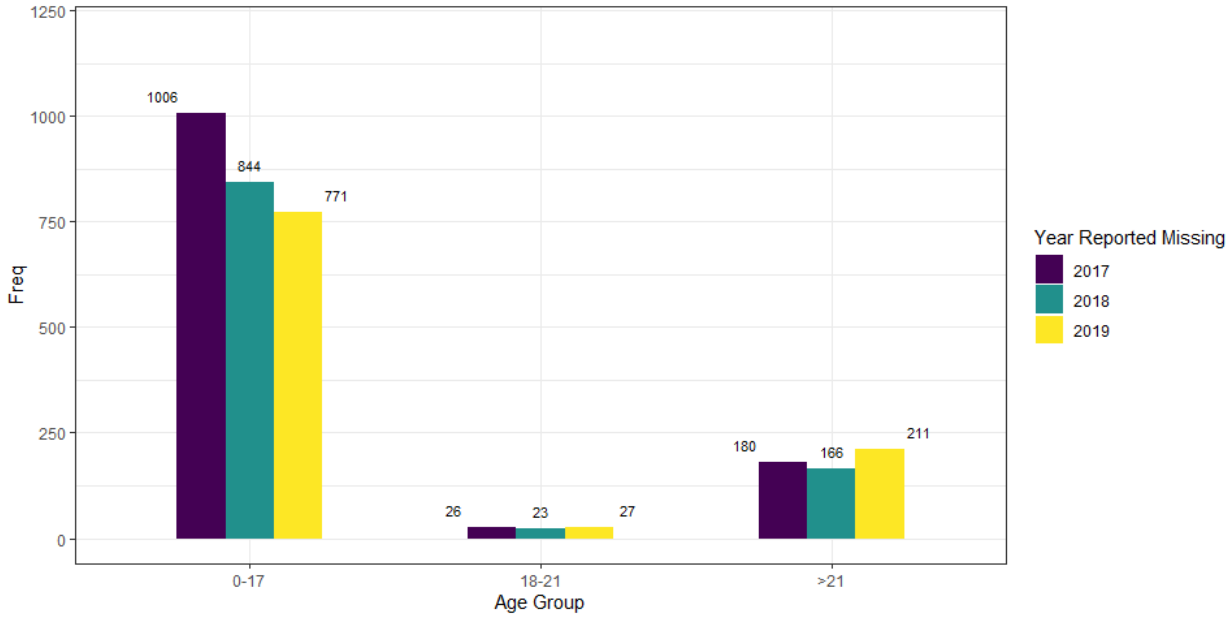

Figure 9: Number of unique individuals by the year of the first time the individual was reported missing by age groups.

Table 16: Counts of unique individuals by gender and the year of the first time the individual was reported missing by age groups.

AGECAT	YEAR.FIRSTDATE.	GENDER	FREQ
0-17	2017	F	549
18-21	2017	F	14
>21	2017	F	84
Total	2017	F	647
0-17	2018	F	447
18-21	2018	F	11
>21	2018	F	65
Total	2018	F	523
0-17	2019	F	395
18-21	2019	F	13
>21	2019	F	95
Total	2019	F	503
0-17	Total	F	1391
18-21	Total	F	38
>21	Total	F	244
Total	Total	F	1673
0-17	2017	M	457
18-21	2017	M	12
>21	2017	M	96
Total	2017	M	565
0-17	2018	M	397
18-21	2018	M	12
>21	2018	M	101
Total	2018	M	510
0-17	2019	M	376
18-21	2019	M	14
>21	2019	M	116
Total	2019	M	506
0-17	Total	M	1230
18-21	Total	M	38
>21	Total	M	313
Total	Total	M	1581

Table 17: Proportions of unique individuals by gender and the year of the first time the individual was reported missing by age groups.

AGECAT	YEAR.FIRSTDATE.	GENDER	FREQ
0-17	2017	F	0.3281530
18-21	2017	F	0.0083682
>21	2017	F	0.0502092
Total	2017	F	0.3867304
0-17	2018	F	0.2671847
18-21	2018	F	0.0065750
>21	2018	F	0.0388524
Total	2018	F	0.3126121
0-17	2019	F	0.2361028
18-21	2019	F	0.0077705
>21	2019	F	0.0567842
Total	2019	F	0.3006575
0-17	Total	F	0.8314405
18-21	Total	F	0.0227137
>21	Total	F	0.1458458
Total	Total	F	1.0000000
0-17	2017	M	0.2890576
18-21	2017	M	0.0075901
>21	2017	M	0.0607211
Total	2017	M	0.3573688
0-17	2018	M	0.2511069
18-21	2018	M	0.0075901
>21	2018	M	0.0638836
Total	2018	M	0.3225806
0-17	2019	M	0.2378242
18-21	2019	M	0.0088552
>21	2019	M	0.0733713
Total	2019	M	0.3200506
0-17	Total	M	0.7779886
18-21	Total	M	0.0240354
>21	Total	M	0.1979760
Total	Total	M	1.0000000

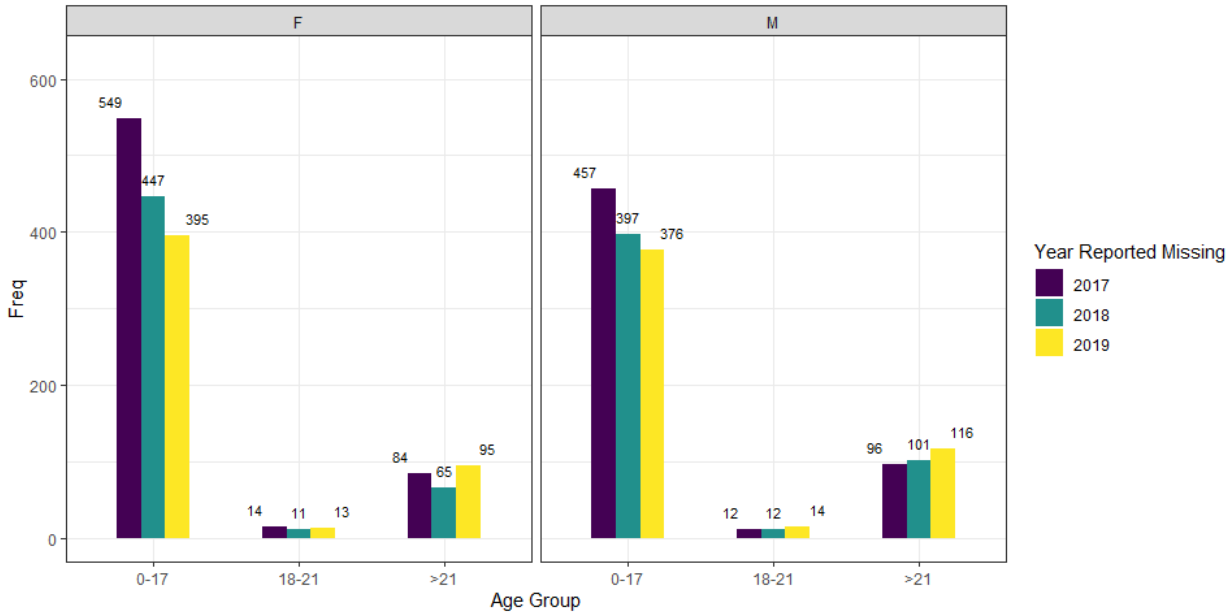

Figure 10: Number of unique individuals by the year of the first time the individual was reported missing by age groups and Gender.

Table 18: Counts of individuals who only went missing once by age group.

	AGE CATEGORIES
0-17	1684
18-21	70
>21	543
Total	2297

Table 19: Counts of individuals who only went missing once by gender and their age.

	0-17	18-21	>21
F	889	34	237
M	795	36	306
Total	1684	70	543

Table 20: Proportions of individuals who only went missing once by gender and their age.

	0-17	18-21	>21
F	0.5279097	0.4857143	0.4364641
M	0.4720903	0.5142857	0.5635359
Total	1.0000000	1.0000000	1.0000000

Table 21: Counts of individuals who only went missing once by their age by year they went missing.

	2017	2018	2019
0-17	575	534	575
18-21	25	20	25
>21	176	162	205
Total	776	716	805

Table 22: Proportions of individuals who only went missing once by their age by year they went missing.

	2017	2018	2019
0-17	0.7409794	0.7458101	0.7142857
18-21	0.0322165	0.0279330	0.0310559
>21	0.2268041	0.2262570	0.2546584
Total	1.0000000	1.0000000	1.0000000

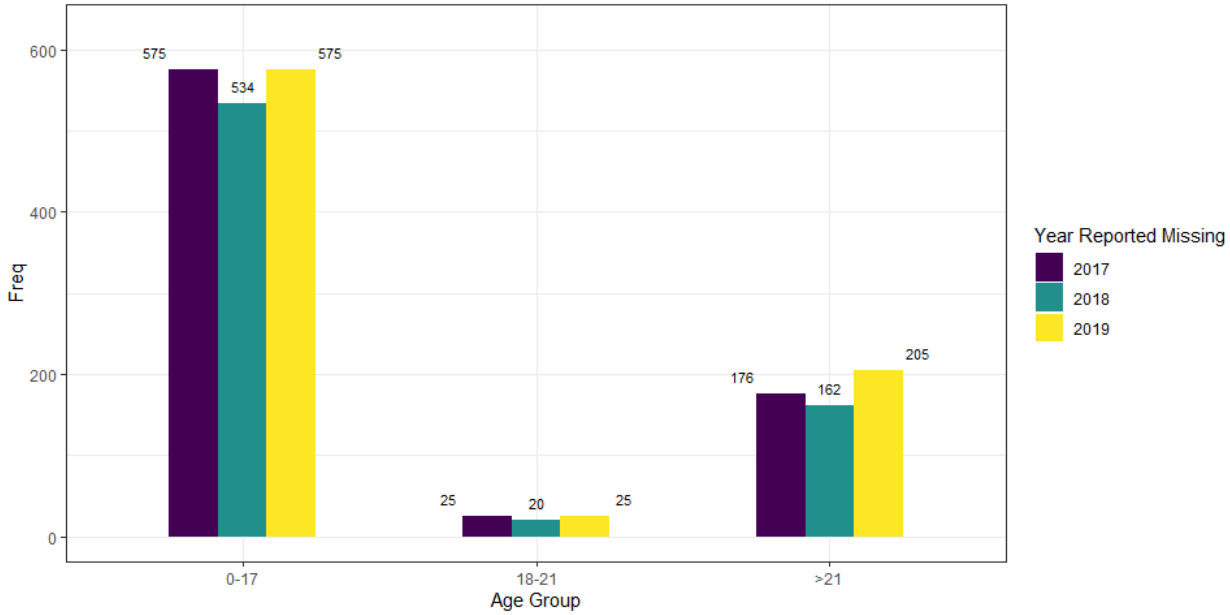

Figure 11: Number of individuals who only went missing once by their age by year they went missing.

Table 23: Counts of individuals who only went missing once by gender and the year the individual was reported missing by age groups.

AGECAT	YEAR.FIRSTDATE.	GENDER	FREQ
0-17	2017	F	308
18-21	2017	F	14
>21	2017	F	82
Total	2017	F	404
0-17	2018	F	290
18-21	2018	F	9
>21	2018	F	63
Total	2018	F	362
0-17	2019	F	291
18-21	2019	F	11
>21	2019	F	92
Total	2019	F	394
0-17	Total	F	889
18-21	Total	F	34
>21	Total	F	237
Total	Total	F	1160
0-17	2017	M	267
18-21	2017	M	11
>21	2017	M	94
Total	2017	M	372
0-17	2018	M	244
18-21	2018	M	11
>21	2018	M	99
Total	2018	M	354
0-17	2019	M	284
18-21	2019	M	14
>21	2019	M	113
Total	2019	M	411
0-17	Total	M	795
18-21	Total	M	36
>21	Total	M	306
Total	Total	M	1137

Table 24: Proportions of individuals who only went missing once by gender and the year the individual was reported missing by age groups.

AGECAT	YEAR.FIRSTDATE.	GENDER	FREQ
0-17	2017	F	0.2655172
18-21	2017	F	0.0120690
>21	2017	F	0.0706897
Total	2017	F	0.3482759
0-17	2018	F	0.2500000
18-21	2018	F	0.0077586
>21	2018	F	0.0543103
Total	2018	F	0.3120690
0-17	2019	F	0.2508621
18-21	2019	F	0.0094828
>21	2019	F	0.0793103
Total	2019	F	0.3396552
0-17	Total	F	0.7663793
18-21	Total	F	0.0293103
>21	Total	F	0.2043103
Total	Total	F	1.0000000
0-17	2017	M	0.2348285
18-21	2017	M	0.0096746
>21	2017	M	0.0826737
Total	2017	M	0.3271768
0-17	2018	M	0.2145998
18-21	2018	M	0.0096746
>21	2018	M	0.0870712
Total	2018	M	0.3113456
0-17	2019	M	0.2497801
18-21	2019	M	0.0123131
>21	2019	M	0.0993843
Total	2019	M	0.3614776
0-17	Total	M	0.6992084
18-21	Total	M	0.0316623
>21	Total	M	0.2691293
Total	Total	M	1.0000000

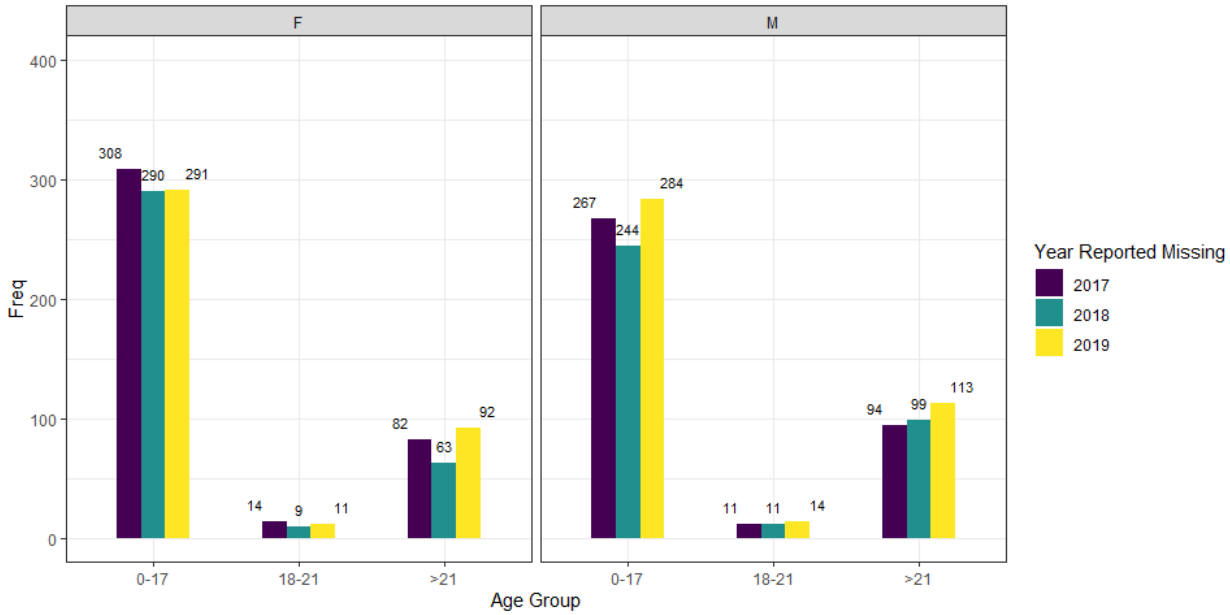

Figure 12: Number of individuals who only went missing once by gender and the year the individual was reported missing by age groups.

Race

Table 25: Counts of unique individuals by race.

	COUNT
Asian	24
Black	93
Indigenous	830
Unknown	104
White	2203
Total	3254

Table 26: Proportions of unique individuals by race.

	PROPORTION
Asian	0.0073755
Black	0.0285802
Indigenous	0.2550707
Unknown	0.0319607
White	0.6770129
Total	1.0000000

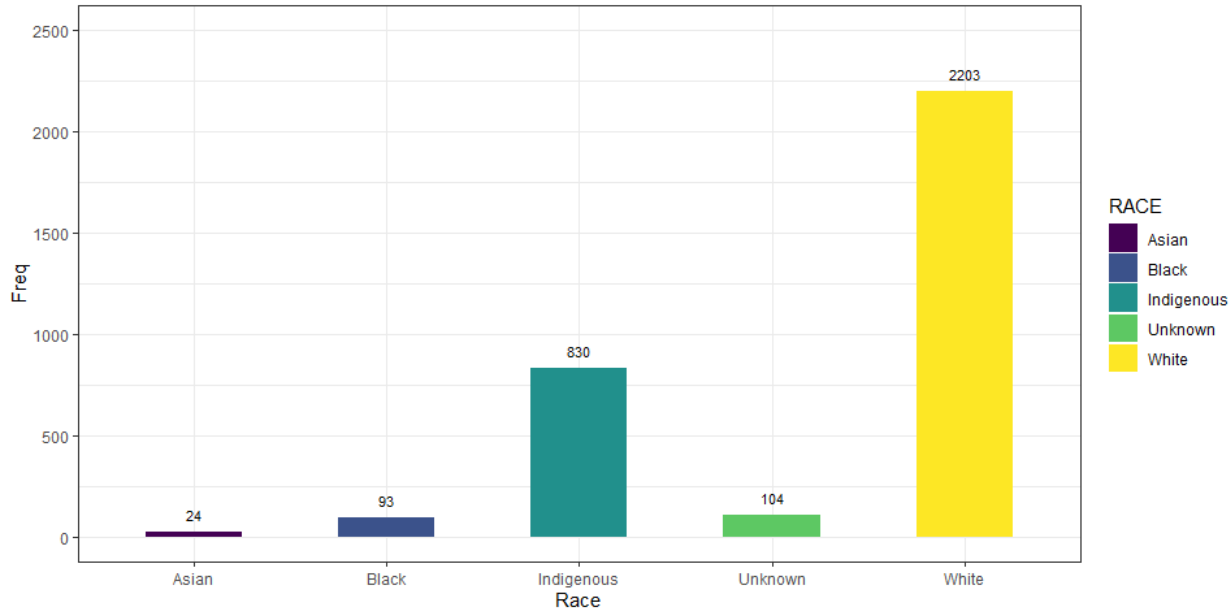

Figure 13: Number of unique individuals by race.

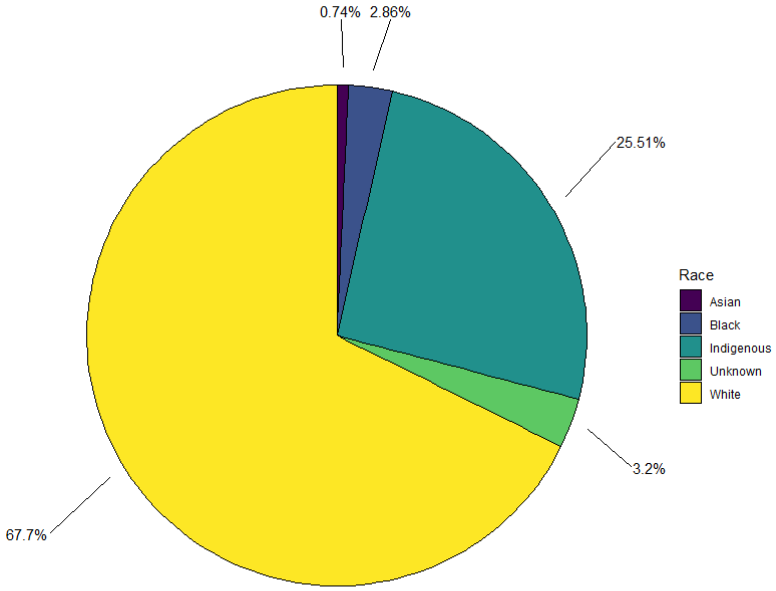

Figure 14: Percent of unique individuals by race.

Table 27: Counts of active cases by race.

	COUNT
Asian	2

	COUNT
Black	0
Indigenous	27
Unknown	3
White	51
Total	83

Table 28: Proportions of active cases by race.

	PROPORTION
Asian	0.0240964
Black	0.0000000
Indigenous	0.3253012
Unknown	0.0361446
White	0.6144578
Total	1.0000000

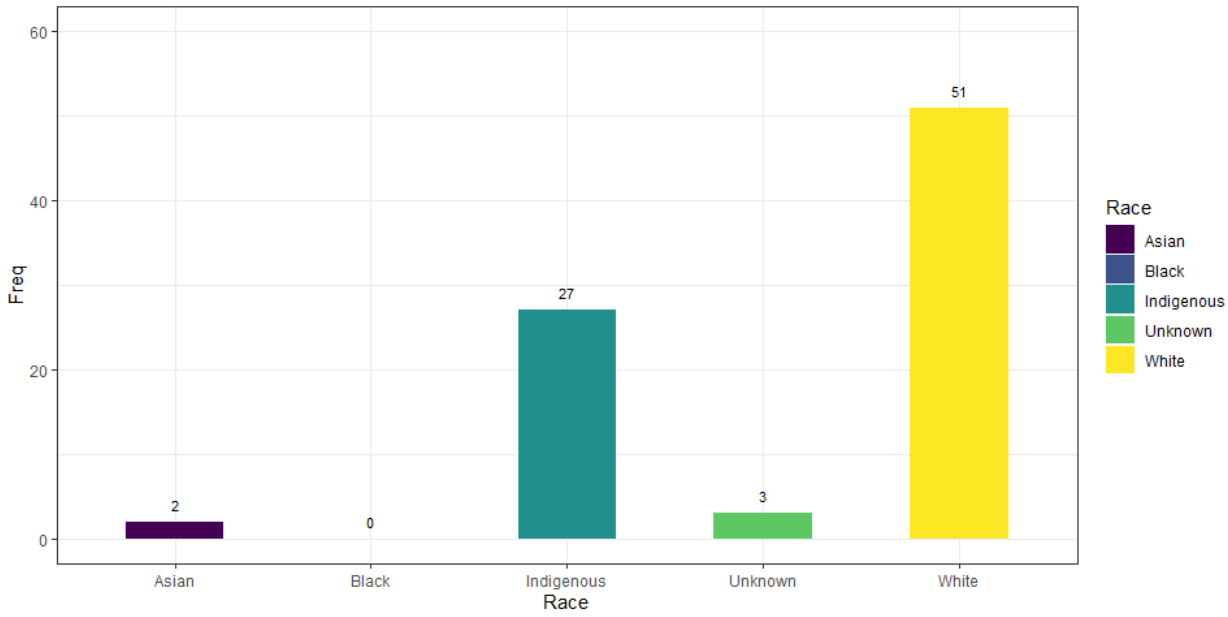

Figure 15: Number of active cases by race.

Table 29: Counts of unique individuals by race and age at the time of the first time reported missing.

	0-17	18-21	>21
Asian	21	0	3
Black	80	5	8
Indigenous	708	23	99
Unknown	80	2	22
White	1732	46	425
Total	2621	76	557

Table 30: Proportions of unique individuals by race and age at the time of the first time reported missing.

	0-17	18-21	>21
Asian	0.0080122	0.0000000	0.0053860
Black	0.0305227	0.0657895	0.0143627
Indigenous	0.2701259	0.3026316	0.1777379
Unknown	0.0305227	0.0263158	0.0394973
White	0.6608165	0.6052632	0.7630162
Total	1.0000000	1.0000000	1.0000000

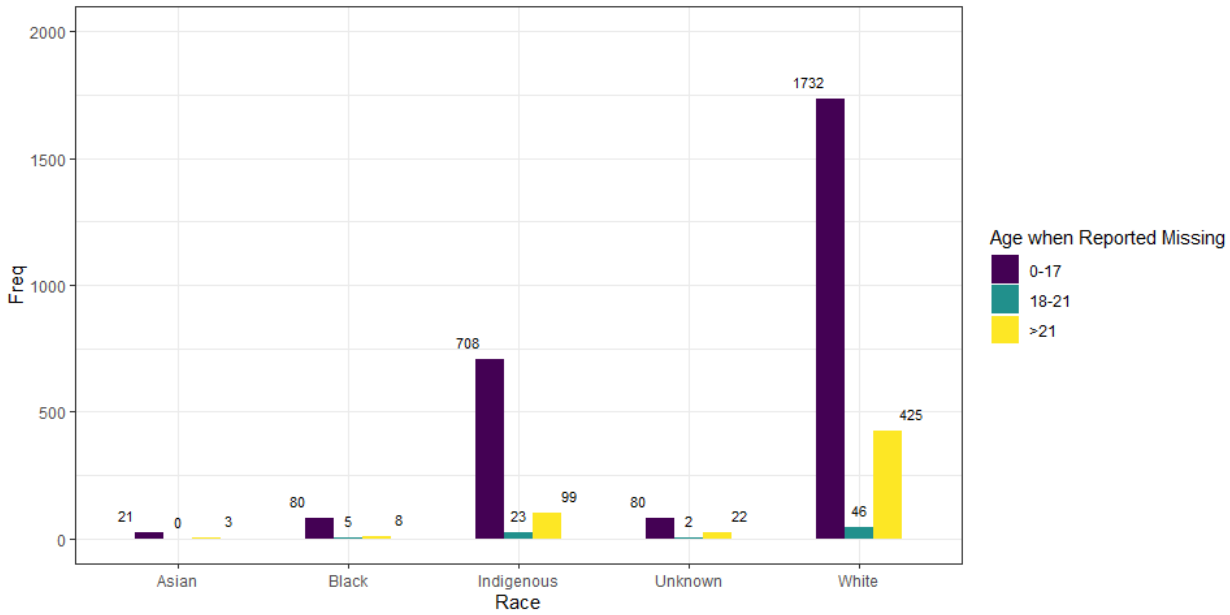

Figure 16: Number of unique individuals by by race and age at the time of the first time reported missing.

Table 31: Counts of unique individuals by race and age at the time of the first time reported missing.

RACE	YEAR.FIRSTDATE.	GENDER	FREQ
Asian	2017	F	5
Black	2017	F	18
Indigenous	2017	F	185
Unknown	2017	F	12
White	2017	F	427
Total	2017	F	647
Asian	2018	F	2
Black	2018	F	9
Indigenous	2018	F	142
Unknown	2018	F	20
White	2018	F	350
Total	2018	F	523
Asian	2019	F	4
Black	2019	F	8
Indigenous	2019	F	164
Unknown	2019	F	25
White	2019	F	302
Total	2019	F	503
Asian	Total	F	11
Black	Total	F	35
Indigenous	Total	F	491
Unknown	Total	F	57
White	Total	F	1079
Total	Total	F	1673
Asian	2017	M	4
Black	2017	M	34
Indigenous	2017	M	107
Unknown	2017	M	12
White	2017	M	408
Total	2017	M	565
Asian	2018	M	2
Black	2018	M	11
Indigenous	2018	M	103
Unknown	2018	M	18
White	2018	M	376

RACE	YEAR.FIRSTDATE.	GENDER	FREQ
Total	2018	M	510
Asian	2019	M	7
Black	2019	M	13
Indigenous	2019	M	129
Unknown	2019	M	17
White	2019	M	340
Total	2019	M	506
Asian	Total	M	13
Black	Total	M	58
Indigenous	Total	M	339
Unknown	Total	M	47
White	Total	M	1124
Total	Total	M	1581

Table 32: Proportions of unique individuals by race and age at the time of the first time reported missing.

RACE	YEAR.FIRSTDATE.	GENDER	FREQ
Asian	2017	F	0.0029886
Black	2017	F	0.0107591
Indigenous	2017	F	0.1105798
Unknown	2017	F	0.0071727
White	2017	F	0.2552301
Total	2017	F	0.3867304
Asian	2018	F	0.0011955
Black	2018	F	0.0053796
Indigenous	2018	F	0.0848775
Unknown	2018	F	0.0119546
White	2018	F	0.2092050
Total	2018	F	0.3126121
Asian	2019	F	0.0023909
Black	2019	F	0.0047818
Indigenous	2019	F	0.0980275
Unknown	2019	F	0.0149432
White	2019	F	0.1805140
Total	2019	F	0.3006575
Asian	Total	F	0.0065750

RACE	YEAR.FIRSTDATE.	GENDER	FREQ
Black	Total	F	0.0209205
Indigenous	Total	F	0.2934848
Unknown	Total	F	0.0340705
White	Total	F	0.6449492
Total	Total	F	1.0000000
Asian	2017	M	0.0025300
Black	2017	M	0.0215054
Indigenous	2017	M	0.0676787
Unknown	2017	M	0.0075901
White	2017	M	0.2580645
Total	2017	M	0.3573688
Asian	2018	M	0.0012650
Black	2018	M	0.0069576
Indigenous	2018	M	0.0651486
Unknown	2018	M	0.0113852
White	2018	M	0.2378242
Total	2018	M	0.3225806
Asian	2019	M	0.0044276
Black	2019	M	0.0082226
Indigenous	2019	M	0.0815939
Unknown	2019	M	0.0107527
White	2019	M	0.2150538
Total	2019	M	0.3200506
Asian	Total	M	0.0082226
Black	Total	M	0.0366856
Indigenous	Total	M	0.2144213
Unknown	Total	M	0.0297280
White	Total	M	0.7109424
Total	Total	M	1.0000000

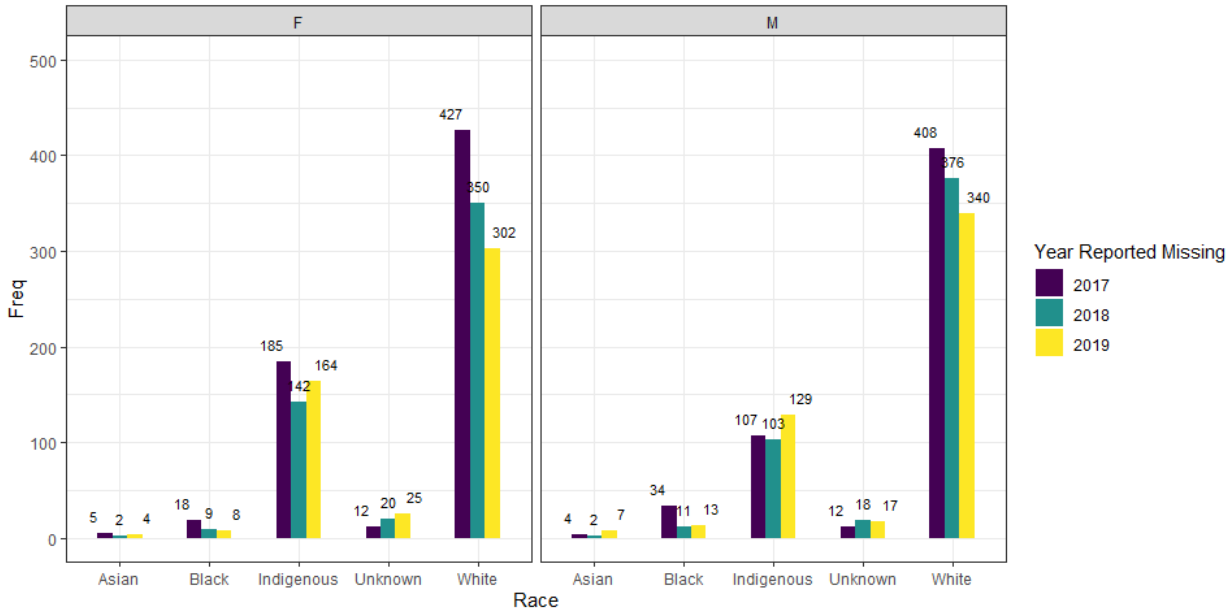

Figure 17: Number of unique individuals by the year of the first time the individual was reported missing by race and Gender.

Table 33: Counts of unique individuals by race and age at the time of the first time reported missing and year reported.

AGECAT	RACE	YEAR.FIRSTDATE.	FREQ
0-17	Asian	2017	8
18-21	Asian	2017	0
>21	Asian	2017	1
Total	Asian	2017	9
0-17	Black	2017	47
18-21	Black	2017	2
>21	Black	2017	3
Total	Black	2017	52
0-17	Indigenous	2017	269
18-21	Indigenous	2017	5
>21	Indigenous	2017	18
Total	Indigenous	2017	292
0-17	Unknown	2017	16
18-21	Unknown	2017	1
>21	Unknown	2017	7
Total	Unknown	2017	24
0-17	White	2017	666
18-21	White	2017	18

AGECAT	RACE	YEAR.FIRSTDATE.	FREQ
>21	White	2017	151
Total	White	2017	835
0-17	Total	2017	1006
18-21	Total	2017	26
>21	Total	2017	180
Total	Total	2017	1212
0-17	Asian	2018	4
18-21	Asian	2018	0
>21	Asian	2018	0
Total	Asian	2018	4
0-17	Black	2018	16
18-21	Black	2018	2
>21	Black	2018	2
Total	Black	2018	20
0-17	Indigenous	2018	215
18-21	Indigenous	2018	8
>21	Indigenous	2018	22
Total	Indigenous	2018	245
0-17	Unknown	2018	32
18-21	Unknown	2018	1
>21	Unknown	2018	5
Total	Unknown	2018	38
0-17	White	2018	577
18-21	White	2018	12
>21	White	2018	137
Total	White	2018	726
0-17	Total	2018	844
18-21	Total	2018	23
>21	Total	2018	166
Total	Total	2018	1033
0-17	Asian	2019	9
18-21	Asian	2019	0
>21	Asian	2019	2
Total	Asian	2019	11
0-17	Black	2019	17
18-21	Black	2019	1
>21	Black	2019	3

AGECAT	RACE	YEAR.FIRSTDATE.	FREQ
Total	Black	2019	21
0-17	Indigenous	2019	224
18-21	Indigenous	2019	10
>21	Indigenous	2019	59
Total	Indigenous	2019	293
0-17	Unknown	2019	32
18-21	Unknown	2019	0
>21	Unknown	2019	10
Total	Unknown	2019	42
0-17	White	2019	489
18-21	White	2019	16
>21	White	2019	137
Total	White	2019	642
0-17	Total	2019	771
18-21	Total	2019	27
>21	Total	2019	211
Total	Total	2019	1009

Table 34: Proportions of unique individuals by race and age at the time of the first time reported missing and year reported.

AGECAT	RACE	YEAR.FIRSTDATE.	FREQ
0-17	Asian	2017	0.0066007
18-21	Asian	2017	0.0000000
>21	Asian	2017	0.0008251
Total	Asian	2017	0.0074257
0-17	Black	2017	0.0387789
18-21	Black	2017	0.0016502
>21	Black	2017	0.0024752
Total	Black	2017	0.0429043
0-17	Indigenous	2017	0.2219472
18-21	Indigenous	2017	0.0041254
>21	Indigenous	2017	0.0148515
Total	Indigenous	2017	0.2409241
0-17	Unknown	2017	0.0132013
18-21	Unknown	2017	0.0008251
>21	Unknown	2017	0.0057756

AGECAT	RACE	YEAR.FIRSTDATE.	FREQ
Total	Unknown	2017	0.0198020
0-17	White	2017	0.5495050
18-21	White	2017	0.0148515
>21	White	2017	0.1245875
Total	White	2017	0.6889439
0-17	Total	2017	0.8300330
18-21	Total	2017	0.0214521
>21	Total	2017	0.1485149
Total	Total	2017	1.0000000
0-17	Asian	2018	0.0038722
18-21	Asian	2018	0.0000000
>21	Asian	2018	0.0000000
Total	Asian	2018	0.0038722
0-17	Black	2018	0.0154889
18-21	Black	2018	0.0019361
>21	Black	2018	0.0019361
Total	Black	2018	0.0193611
0-17	Indigenous	2018	0.2081317
18-21	Indigenous	2018	0.0077444
>21	Indigenous	2018	0.0212972
Total	Indigenous	2018	0.2371733
0-17	Unknown	2018	0.0309777
18-21	Unknown	2018	0.0009681
>21	Unknown	2018	0.0048403
Total	Unknown	2018	0.0367861
0-17	White	2018	0.5585673
18-21	White	2018	0.0116167
>21	White	2018	0.1326234
Total	White	2018	0.7028074
0-17	Total	2018	0.8170378
18-21	Total	2018	0.0222652
>21	Total	2018	0.1606970
Total	Total	2018	1.0000000
0-17	Asian	2019	0.0089197
18-21	Asian	2019	0.0000000
>21	Asian	2019	0.0019822
Total	Asian	2019	0.0109019

AGECAT	RACE	YEAR.FIRSTDATE.	FREQ
0-17	Black	2019	0.0168484
18-21	Black	2019	0.0009911
>21	Black	2019	0.0029732
Total	Black	2019	0.0208127
0-17	Indigenous	2019	0.2220020
18-21	Indigenous	2019	0.0099108
>21	Indigenous	2019	0.0584737
Total	Indigenous	2019	0.2903865
0-17	Unknown	2019	0.0317146
18-21	Unknown	2019	0.0000000
>21	Unknown	2019	0.0099108
Total	Unknown	2019	0.0416254
0-17	White	2019	0.4846383
18-21	White	2019	0.0158573
>21	White	2019	0.1357780
Total	White	2019	0.6362735
0-17	Total	2019	0.7641229
18-21	Total	2019	0.0267592
>21	Total	2019	0.2091179
Total	Total	2019	1.0000000

Missing more than once

Table 35: Summary statistics of number of times missing for all subjects.

MIN	Q1	MEDIAN	Q3	MAX	MEAN	SD	N	MISSING
1	1	1	2	20	1.710818	1.632825	3254	0

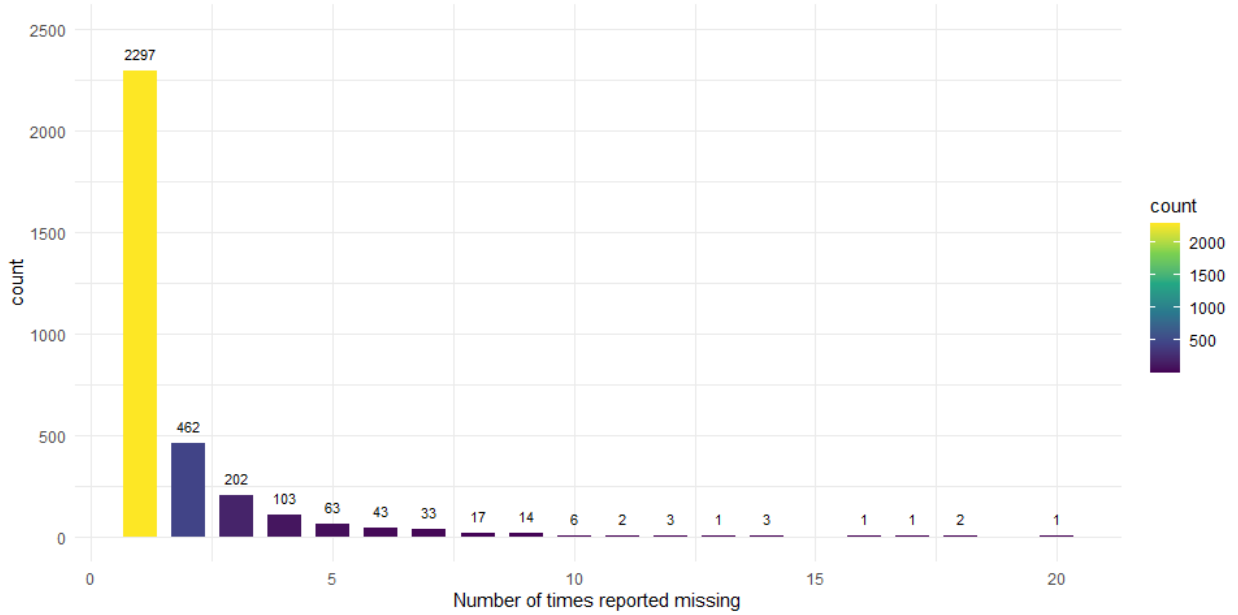

Figure 18: Plot of the number of subjects in each count of number of times reported missing in the database.

Tribal Reporting

Table 36: Counts of missing person reports by reservation and year at the time of the first time reported missing.

	2017	2018	2019
BLACKFEET	5	5	20
CROW	26	24	37
FLATHEAD	37	21	13
FORT PECK	1	8	11
NORTHERN CHEYENNE	5	5	37
ROCKY BOY	0	4	2
Total	74	67	120

Table 37: Proportions of unique individuals by reservation and year at the time of the first time reported missing.

	2017	2018	2019
BLACKFEET	0.0675676	0.0746269	0.1666667
CROW	0.3513514	0.3582090	0.3083333
FLATHEAD	0.5000000	0.3134328	0.1083333
FORT PECK	0.0135135	0.1194030	0.0916667
NORTHERN CHEYENNE	0.0675676	0.0746269	0.3083333
ROCKY BOY	0.0000000	0.0597015	0.0166667
Total	1.0000000	1.0000000	1.0000000

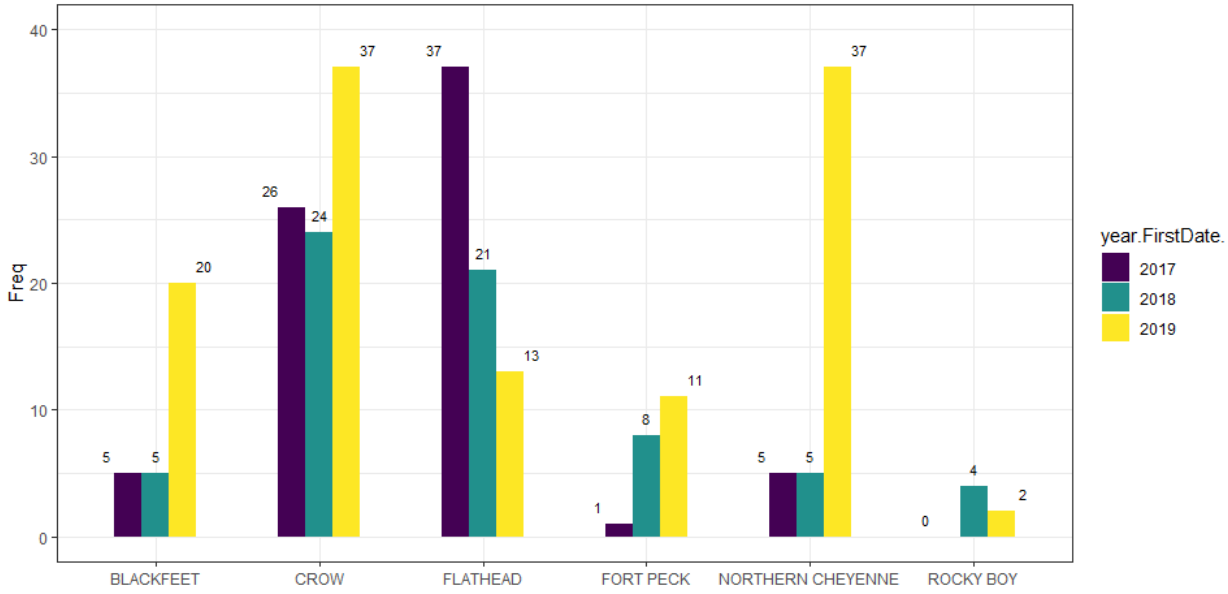

Figure 19: Number of unique individuals by reservation and year at the time of the first time reported missing.

Missing Person Reports by County 2017-2019

Using the estimated county populations from the 2018 American Community Survey we can create reports on the number of missing person reports by county for the state of Montana. Both the total number of missing person reports and the number of missing persons based on the first time reported missing is summarized in the graphics below.

The following code below was used to fix county names and add the information to the dataset.

The `fullDataR_withIntBins_withFixedCounties.csv` file was manually copied into the `../../MTDOJ All Missing Persons Reports 2017-2019 SCRS Edits.xlsx` file on the "AllCasesPosIntervals" tab.

Top 10 Counties for Missing Person Reports 2017-2019

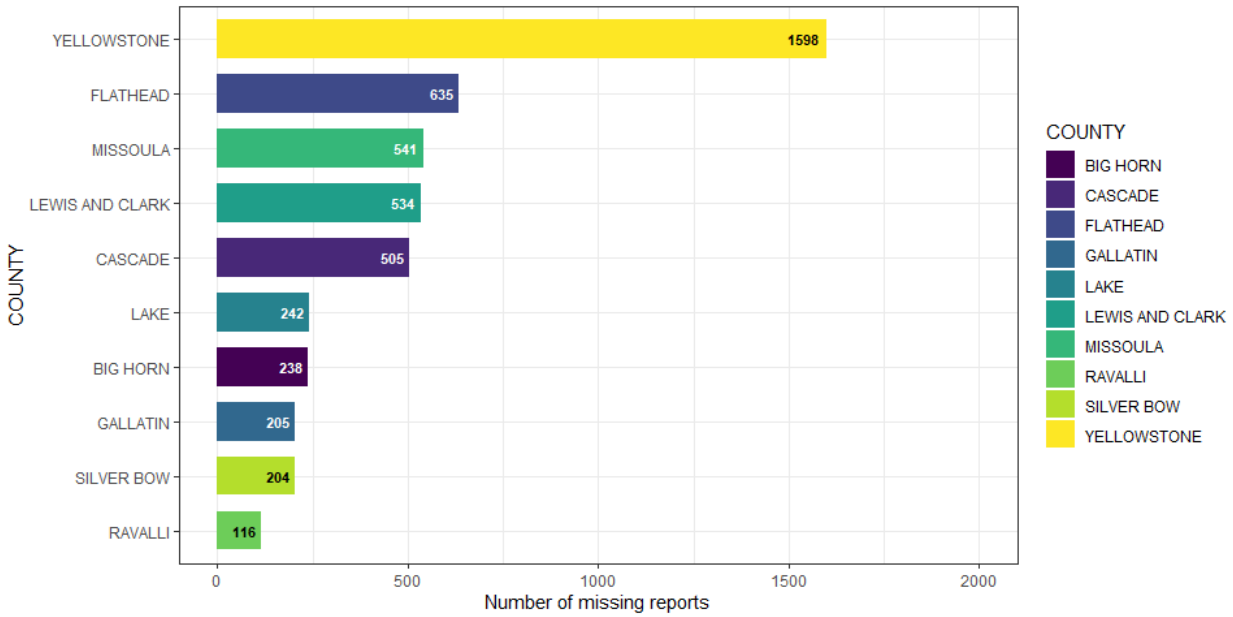

Figure 20: Plot of the counts of all missing person reports by county, for the top 10 counties.

Top 10 Counties for Unique Missing Person Reports 2017-2019

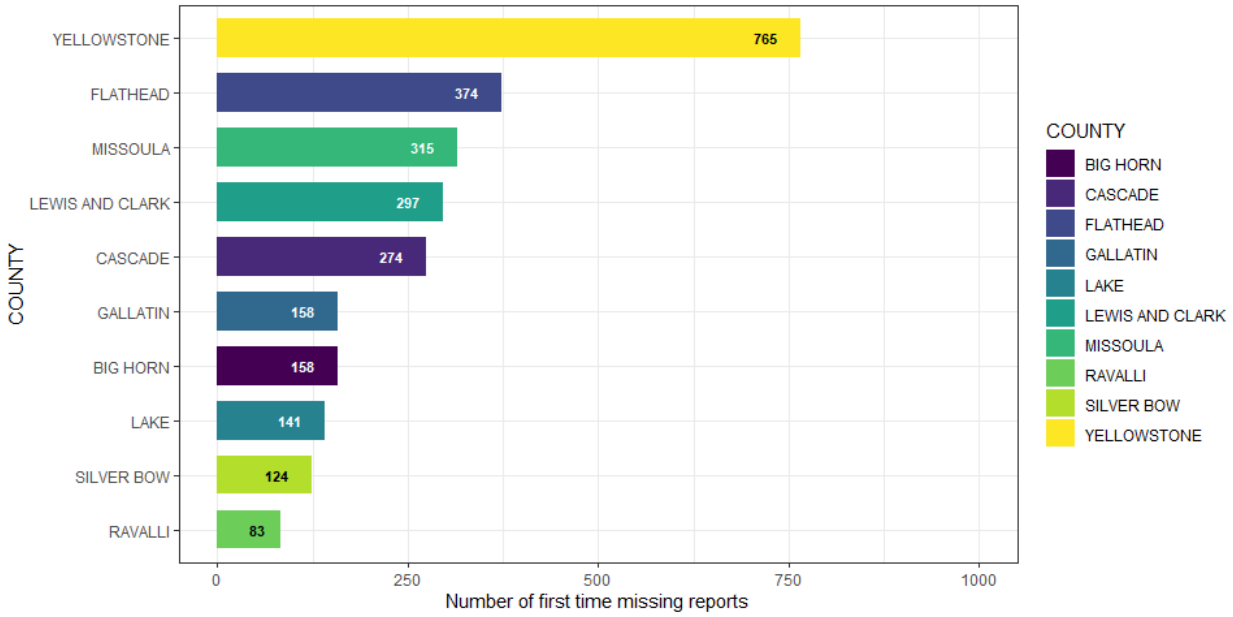

Figure 21: Plot of the counts of unique individual's first time missing reports by county, for the top 10 counties.

Map of Counties for Missing Person Reports 2017-2019

Across the state of Montana, several counties have higher instances of missing person reports than other counties. To account for the varying population sizes of counties across Montana, the number of first time missing reports per county is divided by the county's population and then scaled to be equivalent to the number of first time missing person reports per 1000 people per county.

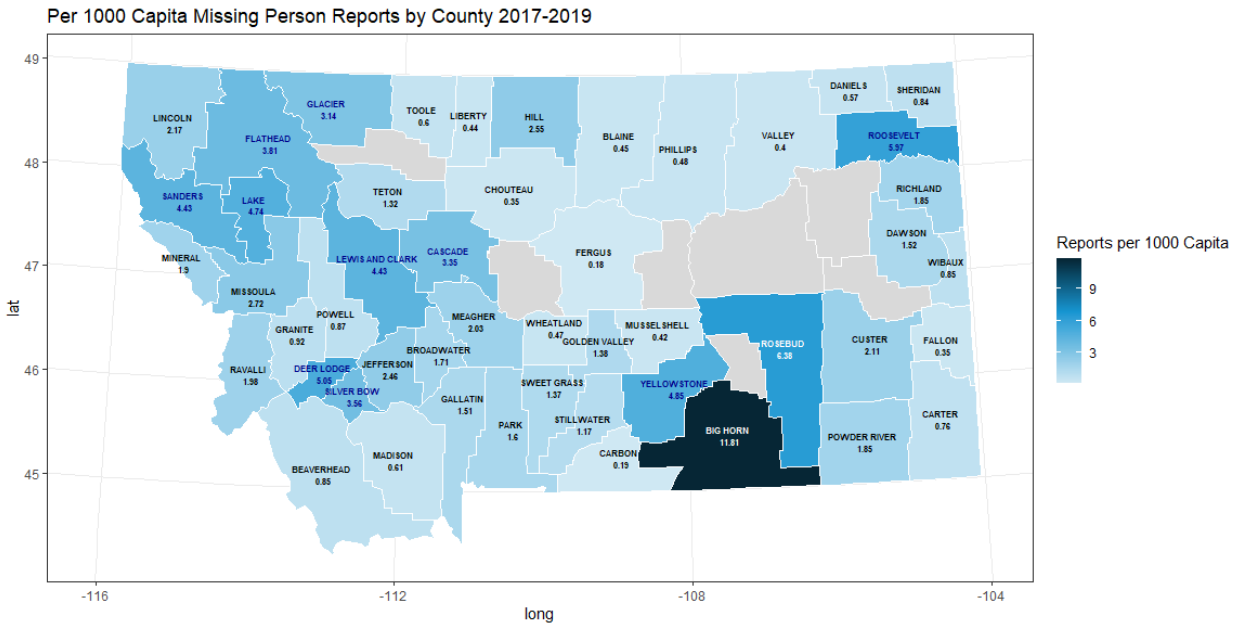

Figure 22: Number of first time missing person reports per 1000 capita per county in Montana.

If we take all missing person reports from 2017-2019, including reports for individuals who went missing more than once, we can also divide these number of reports by each county's population size and then scale to 1000 capita. When we do so we get the following map of the state of Montana.

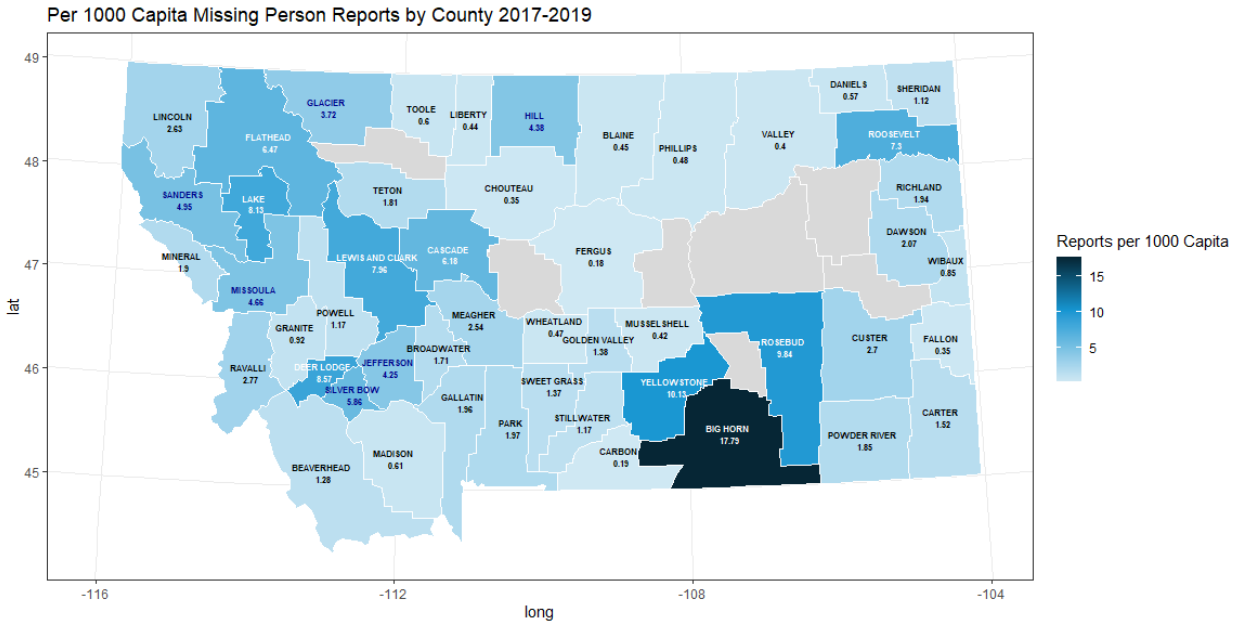

Figure 23: Number of missing person reports per 1000 capita per county in Montana.

Packages used

The R package bookdown (Xie 2020a) was used to create this report document using the R language (R Core Team 2020). In addition, the following were packages used for the analysis and/or formatting of this document:

- `knitr` (Xie 2020b)
- `rmarkdown` (Allaire et al. 2020)
- `readr` (Wickham, Hester, and Francois 2018)
- `readxl` (Wickham and Bryan 2019)
- `tibble` (Müller and Wickham 2020)
- `dplyr` (Wickham, François, et al. 2020)
- `mosaic` (Pruim, Kaplan, and Horton 2020)
- `ggplot2` (Wickham, Chang, et al. 2020)
- `lubridate` (Spinu, Grolemund, and Wickham 2020)
- `beanplot` (Kampstra 2014)
- `ggpubr` (Kassambara 2020)
- `survminer` (Kassambara, Kosinski, and Biecek 2020)
- `survival` (Therneau 2020)

- `ggrepel` (Slowikowski 2020)
- `urbnmapr` (Strochak, Ueyama, and Williams 2020)
- `urbnthemes` (Williams, Ueyama, and Chartoff 2020)

References

- Bogaerts, K., Komarek, A., and Lesaffre, E. (2017) Survival Analysis with Interval-Censored Data: A Practical Approach with Examples in R, SAS, and BUGS. CRC Press.
- Kaplan, E. L. and Meier, P. (1958). "Nonparametric estimation from incomplete observations". J. Amer. Statist. Assoc. 53 (282): 457–481. doi:10.2307/2281868

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2020. *Rmarkdown: Dynamic Documents for R*. https://CRAN.R-project.org/package=rmarkdown.

Kampstra, Peter. 2014. *Beanplot: Visualization via Beanplots (Like Boxplot/Stripchart/Violin Plot)*. https://CRAN.R-project.org/package=beanplot.

Kassambara, Alboukadel. 2020. *Ggpubr: 'Ggplot2' Based Publication Ready Plots*. https://CRAN.R-project.org/package=ggpubr.

Kassambara, Alboukadel, Marcin Kosinski, and Przemyslaw Biecek. 2020. *Survminer: Drawing Survival Curves Using 'Ggplot2'*. https://CRAN.R-project.org/package=survminer.

Müller, Kirill, and Hadley Wickham. 2020. *Tibble: Simple Data Frames*. https://CRAN.R-project.org/package=tibble.

Pruim, Randall, Daniel T. Kaplan, and Nicholas J. Horton. 2020. *Mosaic: Project Mosaic Statistics and Mathematics Teaching Utilities*. https://CRAN.R-project.org/package=mosaic.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Slowikowski, Kamil. 2020. *Ggrepel: Automatically Position Non-Overlapping Text Labels with 'Ggplot2'*. https://CRAN.R-project.org/package=ggrepel.

Spinu, Vitalie, Garrett Grolemund, and Hadley Wickham. 2020. *Lubridate: Make Dealing with Dates a Little Easier*. https://CRAN.R-project.org/package=lubridate.

Strochak, Sarah, Kyle Ueyama, and Aaron Williams. 2020. *Urbnmapr: State and County Shapefiles in Sf and Tibble Format*. https://github.com/UrbanInstitute/urbnmapr.

Therneau, Terry M. 2020. *Survival: Survival Analysis*. https://CRAN.R-project.org/package=survival.

Wickham, Hadley, and Jennifer Bryan. 2019. *Readxl: Read Excel Files*. https://CRAN.R-project.org/package=readxl.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. 2020. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2020. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Romain Francois. 2018. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Williams, Aaron R., Kyle Ueyama, and Ben Chartoff. 2020. *Urbnthemes: Additional Theme and Utilities for "Ggplot2" in the Urban Institute Style*. https://github.com/UrbanInstitute/urbnthemes.

Xie, Yihui. 2020a. *Bookdown: Authoring Books and Technical Documents with R Markdown*. https://CRAN.R-project.org/package=bookdown.

———. 2020b. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://CRAN.R-project.org/package=knitr.