This document explains the manual and excel work that I have done to transform the raw data into something that are processed later in R, including assumptions made for some special entries.

# California

* ***2.raw\_data\_CA\_Supp.xlsx*** is the raw data downloaded at the end of Jan 2022 on <https://oag.ca.gov/privacy/databreach/list>, after removing breaches with ‘Reported Date’ later than Dec 31, 2021.
* In the raw data, the field ‘Date(s) of Breach’ contains a set of dates. I use the excel function ‘Text to Columns’ to put different dates into separate columns. This is ***4.data\_put\_in\_R\_CA\_Supp.csv***.

# Indiana

* ***2.raw\_data\_IN\_Supp.pdf***  is created by combining a set of pdfs that contains the raw data downloaded at the end of Jan 2022 on <https://www.in.gov/attorneygeneral/2874.htm>.
* Date of breach is the start date of breach. See the correspondence below between me and the office of Indiana Attorney General.
* ***4.data\_put\_in\_R\_IN\_Supp.csv*** incorporated the following changes compared to the raw data.
* I turn the pdfs into excel spreadsheets and combine them into one.
* In the field ‘Total Affected’, millions are not in numbers (e.g., “10 mils”). I manually change them to numbers.
* There is one entry with the fields ‘IN Affected’and ‘Total Affected’ as “>6480” and “>600.00”. I manually change them to “6480” and “600”.
* Recording error – decimal places in the total number of people affected e.g., “22.1982”, keep it as it is as there is no way to crosscheck. Didn’t use this variable, so no issues.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

# Montana

* ***2.raw\_data\_MT\_Supp.xlsx*** is the raw data downloaded at the of end of Jan 2022 on <https://dojmt.gov/consumer/databreach/>, by copy-pasting all the data from the websites. The field ‘NOTIFICATION DOCUMENTS’ was deleted.
* ***4.data\_put\_in\_R\_MT\_Supp.csv*** incorporated the following changes compared to the raw data.
* I scan through the entire document and find entries that contain text. Some are supplementary notices, which record reported dates of both the original and supplementary notice, and the updated confirmed number of Montanan affected. I delete supplementary notices and update the original notice with the the correct number of Montanan affected.
* dates in the field ‘MONTANANS AFFECTED’ e.g., “06/18/2020” – change to unknown
* Fewer than 10 – change to 10
* 04/01/2020 (approx.) – change to 04/01/2020

# Maine

* ***2.raw\_data\_ME1\_Supp.xlsx*** and ***2.raw\_data\_ME2\_Supp.xlsx*** are raw data downloaded at the of end of Jan 2022 on <https://www.maine.gov/ag/consumer/identity_theft/index.shtml>

## Cleaning data between 2010-2018

***4.data\_put\_in\_R\_ME1\_Supp.csv*** incorporated the following changes compared to the raw data. I begin by establishing my objectives and then going over the steps I take to reach them.

1. In ***2.raw\_data\_ME1\_Supp.xlsx***, date formats are not consistent (See Note 1 below). I dealt with breaches with different date formats separately.

Note 1: Date formats for 2010-2018 Maine breaches. DoB stands for *Date of Breach* and DoN stands for *Date of Notification*.

* When the DoB column contains only one date and no text (e.g., 18/9/2018), the date format is dd/mm/yyyy. In other circumstances (e.g., the DoB column contains multiple dates or has text), the DoB format is mm/dd/yyyy.
* Date format of the DoN column before the breach which relates to the organisation named ‘Crystal & Company’ is mm/dd/yyyy; DoN after that is dd/mm/yyyy.

2. Update the original notice with information provided by supplementary notices and delete supplementary notices.

1. Manually scan through the columns ‘*Date of Breach’*, *‘Date of Notification’, and ‘Number of Maine Residents Affected’* to recognise supplementary notices. e.g., the DoB field says ‘Supplement to 10/18/18 Notice’, the DoN field says ‘8/1/18 and follow up to 3/27/18 breach notice’.
2. Identify the original notice by searching for the same organisation name and breach occurrence date.
3. Update the original notice with information provided by supplementary notices (e.g., add additional state residents affected indicated by supplementary notices to the original notice).
4. Delete supplementary notices.

If a breach is found to only contain dates when supplementary notices are submitted but not when the original notice is submitted (1 breach is found to satisfy this), the breach is deleted.

Meanwhile, I get rid of unnecessary text in *‘Date of Notification’, and ‘Number of Maine Residents Affected’,* with the following assumptions made for texts appeared in *‘Number of Maine Residents Affected’*:

* Approx. 6 – change to 6
* 604 students (all ME residents?) – change to 604
* 50 (majority is ME residents) – change to 50
* 55 (addit'l 10 people listed as beneficiaries to a BFE acct) – change to 55
* 225 credit cards from maine residents – change to 225
* 27 (individuals and businesses) – change to unknown
* 2 stores affected – change to unknown
* 1 or more – change to 1
* 1(?) - change to 1
* 2600000 worldwide – change to 2600000
* 60842 in the United States – change to 60842
* Approximately 70,60 – change to unknown

3. Recognise breaches with the DoB column containing only one date and no text.

1. Delete entries with the DoB column looking like ‘8/3/2018 (discovered 10/2/18)’ by looking for the DoB column containing ‘(dis’ or ‘(Dis’.
2. Load all breaches in R, use ‘as.Date’ function to transform the DoB column, and removed all NAs.

4. Recognise the start date of breach of breaches with the DoB column NOT in the format in Step 3.

1. Delete the breaches recognised in Step 3.
2. Use the ‘Text to column’ function in excel to split the text in the DoB column into different columns by characters (e.g., space, hyphen, comma, left bracket).
3. Scan through the first column, which is supposed to contain the start date of the breach, to look for entries which are not in mm/dd/yyyy format. Manually enter the correct format by checking the information in later columns. The following assumptions are made for special cases.

* assumptions made for entries which limits the possible occurrence/discovery period to be less than a month
* May 2020 – change to May 1 2020
* early May 2020 – change to May 1 2020
* sometime in March 2015 – change to March 1 2015
* mid Dec. 2019 – change to December 15 2019
* Late April 2019 – change to April 30 2019
* Mid Dec 2015 (discovered 12/12/15) – change date of breach to 12/12/15
* ignore entries which limits the possible occurrence period to be greater than a month e.g., First half of 2016
* ignore entries with no clear start date of occurrence e.g., On or before April 16 2018, prior to April 16 2018, Discovered 7/6/16
* Date of Breach is shown as “Notified customers on 5/11/15” – change to unknown

## Cleaning data between 2018-2020

***4.data\_put\_in\_R\_ME2\_Supp.csv*** incorporated the following changes compared to the raw data.

1. Completed date is the Reported Date. Date of consumer notification is not the Reported Date.

2. Search “supplement” in the raw data to recognise supplementary notices, update the original notice, and delete supplementary notices.

3. Extract start date of breach from the column *‘Date(s) Breach Occurred’*.

1. Use the ‘Text to column’ function in excel to split the text in this column into different columns by characters (e.g., space, hyphen, comma).
2. Scan through the first column, which is supposed to contain the start date of the breach, to look for entries which are not in mm/dd/yyyy format.
3. Manually enter the correct format by checking the information in later columns.

4. Date format for Date of Breach: some are in “11/14/2018”, and others in “August 26 2019”. I recognise them in R separately.

# North Dakota

* Raw data is downloaded at the of end of Jan 2022 on <https://attorneygeneral.nd.gov/consumer-resources/data-breach-notices>
* Originally, Date of Breach is of the format ‘April 11, 2019 to July 22, 2019’. I use ‘Text to column’ to split text into multiple columns by the character ‘to’, which is the only difference between ***4.data\_put\_in\_R\_ND\_Supp.csv*** and the raw data.

# Washington

* Raw data is downloaded at the of end of Jan 2022 on <https://www.atg.wa.gov/data-breach-notifications>
* ***4.data\_put\_in\_R\_WA\_Supp.csv*** is the same as the raw data.

# Delaware

* Raw data is downloaded at the of end of Jan 2022 on <https://attorneygeneral.delaware.gov/fraud/cpu/securitybreachnotification/database/>
* ***4.data\_put\_in\_R\_DE\_Supp.csv*** is the same as the raw data.

# Oregon

* Raw data is downloaded at the of end of Jan 2022 on <https://justice.oregon.gov/consumer/databreach/>
* Originally, Dates of Breach is of the format ‘9/6/2021 - 9/11/2021, 11/7/2021 - 11/7/2021’. I use ‘Text to column’ to split text into multiple columns by comma and hyphen, which is the only difference between ***4.data\_put\_in\_R\_OR\_Supp.csv*** and the raw data.