Preppin Data 2022

Avishai M. Tsur, MD MHA

2023-07-21

Table of contents

Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

```
1 + 1
```

```
[1] 2
```

1 Week 1

1.1 Requirements

- Input the csv file (link above) (*help)*
- Form the pupil's name correctly for the records in the format *Last Name, First Name* (*help)*
- Form the parental contact's name in the same format as the pupil's
  - The Parental Contact Name 1 and 2 are the first names of each parent.
  - Use parental contact column to select which parent first name to use along with the pupil's last name
- Create the email address to contact the parent using the format *Parent First Name.Parent Last Name@Employer.com*
- Create the academic year the pupils are in (*help*)
  - Each academic year starts on 1st September.
  - Year 1 is anyone born after 1st Sept 2014
  - Year 2 is anyone born between 1st Sept 2013 and 31st Aug 2014 etc
- Remove any unnecessary columns of data (*help)*
- Output the data (*help*)

1.2 Setup

```
library(tidyverse)
raw <- googlesheets4::read_sheet("https://docs.google.com/spreadsheets/d/1SXZMY-kVx2Dz5q3D
```

1.3 EDA

1.4 Table

```
raw
```

```
# A tibble: 1,000 x 9
      id `pupil first name` `pupil last name` gender `Date of Birth`
   <dbl> <chr>              <chr>             <chr>  <dttm>
 1     1 Ronna              Nellies           Female 2013-12-21 00:00:00
 2     2 Rusty              Andriulis         Male   2012-07-21 00:00:00
 3     3 Roberta            Oakeshott         Female 2011-12-04 00:00:00
 4     4 Lola               Rubinfajn         Male   2012-06-29 00:00:00
 5     5 Kamila             Benedtti          Female 2012-07-10 00:00:00
 6     6 Avery              Colebourn         Female 2012-08-30 00:00:00
 7     7 Valentino          Klimko            Female 2014-12-23 00:00:00
 8     8 Cal                Shearwood         Male   2015-01-18 00:00:00
 9     9 King               Truswell          Female 2012-09-14 00:00:00
10    10 Towney             Stichall          Male   2015-06-04 00:00:00
# i 990 more rows
# i 4 more variables: `Parental Contact Name_1` <chr>,
#   `Parental Contact Name_2` <chr>, `Preferred Contact Employer` <chr>,
#   `Parental Contact` <dbl>
```

1.5 Summary

```
skimr::skim(raw)
```

Table 1.1: Data summary

Name	raw
Number of rows	1000
Number of columns	9
Column type frequency:	
character	6
numeric	2
POSIXct	1

Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
pupil first name	0	1	2	12	0	935	0
pupil last name	0	1	3	17	0	989	0
gender	0	1	4	11	0	8	0
Parental Contact Name_1	0	1	2	15	0	945	0
Parental Contact Name_2	0	1	2	14	0	940	0
Preferred Contact Employer	0	1	3	13	0	353	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
id	0	1	500.50	288.82	1	250.75	500.5	750.25	1000	
Parental Contact	0	1	1.49	0.50	1	1.00	1.0	2.00	2	

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
Date of Birth	0	1	2011-09-01	2015-08-30	2013-08-21 12:00:00	713

1.6 Preppin

```
calc_year <- \(date, ref = ymd("2014-09-01")){
  ceiling((interval(date, ref) / years(1)) + 1)
}
```

```r
output <- raw |>
  mutate(pc = if_else(
    `Parental Contact` == 1,
    `Parental Contact Name_1`,
    `Parental Contact Name_2`
    )
    ) |>
  transmute(
    "Academic Year" = calc_year(`Date of Birth`),
    "Pupil's Name" = paste0(`pupil last name`, ", ", `pupil first name`),
    "Parental Contact Full Name" = paste0(`pupil last name`, ", ", pc),
    "Parental Contact Email Address" = paste0(pc, ".", `pupil last name`, "@", `Preferred
    )
```

1.7 Output

1.8 Table

```r
output
```

```
# A tibble: 1,000 x 4
   `Academic Year` `Pupil's Name`  Parental Contact Ful~1 Parental Contact Ema~2
             <dbl> <chr>           <chr>                  <chr>
 1               2 Nellies, Ronna  Nellies, Purcell       Purcell.Nellies@Demiz~
 2               4 Andriulis, Rus~ Andriulis, Vassili     Vassili.Andriulis@Bra~
 3               4 Oakeshott, Rob~ Oakeshott, Haskell     Haskell.Oakeshott@Cen~
 4               4 Rubinfajn, Lola Rubinfajn, Tresa       Tresa.Rubinfajn@Edgeb~
 5               4 Benedtti, Kami~ Benedtti, Adela        Adela.Benedtti@Trudoo~
 6               4 Colebourn, Ave~ Colebourn, Dalenna     Dalenna.Colebourn@Lin~
 7               1 Klimko, Valent~ Klimko, Onofredo       Onofredo.Klimko@Thoug~
 8               1 Shearwood, Cal  Shearwood, Berne       Berne.Shearwood@Brows~
 9               3 Truswell, King  Truswell, Evvy         Evvy.Truswell@Photosp~
10               1 Stichall, Town~ Stichall, Joyann       Joyann.Stichall@Kwimb~
# i 990 more rows
# i abbreviated names: 1: `Parental Contact Full Name`,
#   2: `Parental Contact Email Address`
```

1.9 Summary

```
skimr::skim(output)
```

Table 1.5: Data summary

Name	output
Number of rows	1000
Number of columns	4
_____	
Column type frequency:	
character	3
numeric	1
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Pupil's Name	0	1	9	24	0	1000	0
Parental Contact Full Name	0	1	8	24	0	1000	0
Parental Contact Email Address	0	1	17	36	0	1000	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Academic Year	0	1	2.52	1.12	1	1.75	3	4	4	