Homework 0: Record linkage of business records for a community pantry

Olivier Binette, STA 490/690

General instructions for homeworks: Please follow the uploading file instructions according to the syllabus. Your code must be completely reproducible and must compile.

Advice: Start early on the homeworks and it is advised that you not wait until the day of as these homeworks are meant to be longer and treated as case studies.

Commenting code Code should be commented. See the Google style guide for questions regarding commenting or how to write code https://google.github.io/styleguide/Rguide.xml. No late homework's will be accepted.

R Markdown Test

0. Open a new R Markdown file; set the output to HTML mode and "Knit". This should produce a web page with the knitting procedure executing your code blocks. You can edit this new file to produce your homework submission.

Total points on assignment: 5 (reproducibility) + 10 points for the assignment.

1. Introduction

A community pantry located in Durham[1] provides food and basic necessities to Duke graduate students in need of assistance. Since the beginning of the COVID-19 pandemic, pantry usage has mainly been restricted to the weekly bag program: students fill food bag orders through a Qualtrics survey, and the bags are then delivered to them or ready for pick-up on the next Saturday. The bags typically contain canned beans, canned fruits and vegetables, cereals, rice, flour, and other basic necessities such as diapers.

As a volunteer statistician at the Pantry, you are tasked to describe the composition of pantry users through bag order records (see section 2), where each bag order record represents a food bag that was given to a student. Furthermore, you will provide recommendations regarding how the pantry could improve its data collection efforts to better evaluate the need of students. Your analysis will assist the pantry in its negotiation with university stakeholders and will help them prepare for an upcoming symposium on the subject of food insecurity in U.S. colleges.

In order to describe the composition of pantry users, you will perform the following three tasks:

1. perform deterministic record linkage on anonymized data;
2. construct representative records for individual pantry users identified in (1); and
3. report summary statistics based on (2).

The tasks are outlined in more detail in section 3.

2. Data

The pantry's bag order records consists of bag orders filled between June 24 and November 4 2020 through an online Qualtrics form. Each record contains a date, name, phone number, email, physical address, the requested food order for the week and the answers to a few background questions. This is summarized in Table 1.

	Question no.	Question	Answer form
Identifiers	2	First name and last initial	Free form
	3	Duke email	Free form
	4	Phone number	Free form
Order	5	Delivery or Pickup?	Multiple choices
	$6-7$	Adress and delivery instructions	Free form
	8	Food allergies	Free form
	9	Number of members in household	1-2 or 3+
	10	Want baby bag?	Yes or no
	$11-29$	Order items	Multiple choices
Survey	30	Degree	Multiple choices or Other
	31	School	Multiple choices or Other
	32	Year in graduate school	Multiple choices
	33	Number of adults in household	Multiple choices
	34	Number of children in household	Multiple choices
	35	Main challenges accessing food	Multiple choices or Other
	36	Feedback	Free form

Table 1: Summary of the Qualtrics "Weekly Grocery Bag Request Survey" questions. Note that number of sub-questions which are not relevant to this analysis have been omitted.

[1]The pantry allowed the use of their anonymized data for this case study bu requested not to be named.

Protecting privacy of users. The data collected by the pantry about its users is kept private and is only used for the purpose of fulfilling orders, of reporting summaries of operations and for improving its services. For the purpose of this homework, the pantry has agreed to share an anonymized subset of the bag order records. The anonymization procedure (see file `encrypt_responses.R`) ensures that no private information is shared and that it is highly unlikely that any individual pantry user could be re-identified through a cryptographic or linkage attacks. Specifically, the personal identifiers `name`, `phone` and `email` have been encrypted using an MD5 hash function. Furthermore, only the non free-form survey answers (questions 30-34) have been provided to you for your analysis. This anonymized data is contained in the file `order_data.rds` which will be individual provided to you. Do not share the data.

3. Methodology

Many pantry users have filed more than one order in the considered time period. First, you will resolve individual pantry users using the pantry data set. Second, you will consolidate survey answers for each individual in the pantry. Finally, you will provide summary statistics and provide recommendations regarding the pantry's data collection efforts. In order to do this, you will perform three tasks below!

Task 1: record linkage

The features `name`, `phone` and `email` may have variations, such as variants in spelling of first and last name. For example, one user may enter the first name "Olivier" or "Oliver" by accident. As another example, this same user may have multiple email addresses such as olivier.binette@duke.edu and ob37@duke.edu. The user may decide to use both email addresses at random. Finally, the pantry user may have both a cell number or a landline, and thus, you may observe different `phone` numbers for the same user.

Your first task is to identify unique individuals who filed bag orders during the considered time period. To do this, perform **deterministic record linkage** using the three fields `name`, `phone` and `email` and define two records to be a match if they agree on at least one of these fields. Given your deterministic record linkage, assign a unique entity identifier to each pantry user.

For example, the following two records should be linked since they agree on email and phone number.

date	name	email	phone	how	n_household	babybag	degree	school	year	n_children
2020-07-19	d7ec7...	02410...	7b37f...	8.0	3+ People	Yes	NA	NA	NA	1
2020-08-30	ab5cd...	02410...	7b37f...	Pickup	3+ People	Yes	NA	NA	NA	NA

You may use the guide provided in the appendix to help with this task. Any correct solution will be accepted.

Task 2: canonicalization

Construct a data frame where each **row represents a unique pantry user** and with **columns `degree` and `school`**, where each entry is a representative answer for this individual. For example, you can choose the representative answer to be the first non NA answer for this individual.

The first three records of this data frame could look like this:

uniqueID	degree	school
1	Master's	Fuqua School of Business
2	Master's	Graduate School
3	Master's	Fuqua School of Business

You may use the guide provided in the appendix to help with this task. Any correct solution will be accepted.

Task 3: summary statistics

Given the above, answer the following questions:

1. How many distinct individuals used the pantry? How many orders have they placed on average? How many individuals used the pantry only once?
2. Plot the distribution of the number of visits per pantry user.
3. What is the composition of degree type (Masters, PhD, etc) and school (Graduate School, Nicholas School, etc) among pantry users? Report the proportions using pie charts and use your best judgment if the categories need to be cleaned up.

Furthermore, discuss the following: How could the pantry improve its data collection efforts in order to gain more meaningful insights into its users and their needs? Keep in mind the need to protect the privacy of the pantry users.

Appendix

Guide to task 1

Part a) Read-in the encrypted data `order_data.rds` and view it using the function `View()`. How many non-empty orders were placed in the considered time period?

Solution.

```
library(dplyr)
data = readRDS("order_data.rds")
#View(data)
sum(!is.na(data$name))
```

```
## [1] 549
```

Part b) Construct a function `linkage_rule(A, B)` which returns `TRUE` if the record number `A` matches record number `B` on any of the fields `name`, `phone` or `email`.

Solution.

```
linkage_rule <- function(A, B) {
  recordA = data[A, c("name", "phone", "email"), drop=TRUE]
  recordB = data[B, c("name", "phone", "email"), drop=TRUE]
  return(any(recordA == recordB, na.rm=TRUE))
}
```

Part c) Enumerate all possible record pairs using `pairs = t(combn(1:nrow(data), 2))`. Apply the function `linkage_rule` to each pair in order to determine if they match and store the result in a vector named `matches`. This might take a few minutes to run.

Solution.

```
pairs = t(combn(1:nrow(data), 2))
matches = sapply(1:nrow(pairs), function(i) linkage_rule(pairs[i,1], pairs[i,2]))
```

Part d) Construct a matrix with two columns, where each row is a pair of two matching record numbers. Make sure that each record is matching itself and that you remove NA values. This is called an edge list.

Solution.

```
matching_pairs = pairs[matches, ]
edge_list = rbind(matching_pairs,
                  cbind(1:nrow(data), 1:nrow(data)))
```

Part e) Use the function `igraph::graph_from_edgelist()` to construct a linkage graph based on the edge list. Then use `igraph::components()` to find connected components and assign unique entity identifiers. Hint: use `igraph::components(g)$membership`.

Solution.

```
g = igraph::graph_from_edgelist(edge_list)
IDs = igraph::components(g)$membership
```

Part f) Add a column to the data with the unique entity identifiers. How many distinct individuals visited the pantry in the considered time period?

Solution.

```
data = data %>% mutate(uniqueID = IDs)
length(unique(IDs))
```

```
## [1] 178
```

Part g) (Optional challenge) Can you perform the record linkage much more efficiently, without enumerating all record pairs? Hint: use sorting.

Solution.

```
links <- function(field) {
  o = order(data[[field]])
  sorted = (1:nrow(data))[o]
  edges = t(sapply(1:(length(sorted)-1), function(i) {
    if (linkage_rule(sorted[i], sorted[i+1])) {
      return(c(sorted[i], sorted[i+1]))
    }
    return(c(NA, NA))
  }))

  edges[complete.cases(edges), ]
}

edge_list = rbind(links("name"),
                  links("phone"),
                  links("email"))

g = igraph::graph_from_edgelist(edge_list)
IDs = igraph::components(g)$membership
```

Guide to task 2

Part a) Write a function `first_non_na()` which returns the first non-NA entry of a vector or otherwise returns "NA".

Solution.

```
first_non_na <- function(x) {
  if (all(is.na(x))) return("NA")
  x[!is.na(x)][[1]]
}
```

Part b) Construct a data frame where each row represent a unique pantry user and each column is the first non-NA data entry for this person. Hint: group `data` by unique ID and use the function `dplyr::summarize_all()` to summarize each column using the `first_non_na()` function.

Solution.

```
representers = data %>%
  group_by(uniqueID) %>%
  summarize_all(first_non_na)
```

Solution to task 3

1. How many distinct individuals used the pantry? How many orders have they placed on average? How many individuals used the pantry only once?

Solution.

```
nrow(representers)
```

```
## [1] 178
```

```
nrow(data)/nrow(representers)
```

```
## [1] 3.129213
```

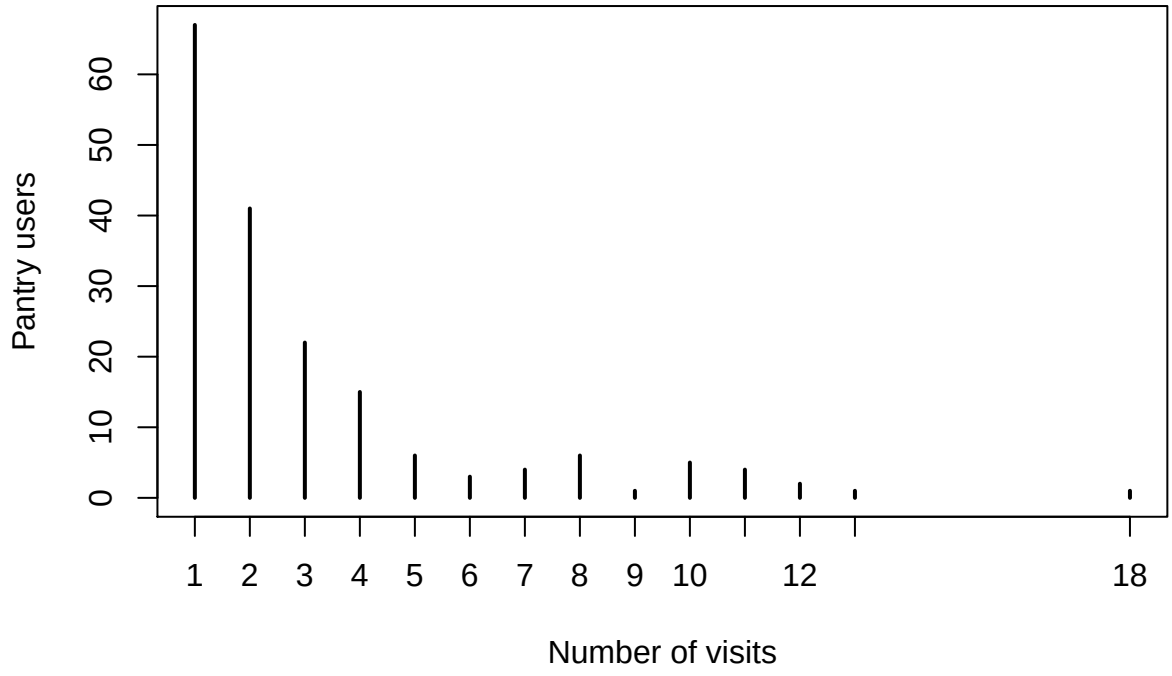
```
table(table(data$uniqueID))[1]
```

```
##  1
## 67
```

2. Plot the distribution of the number of visits per pantry user.

Solution.

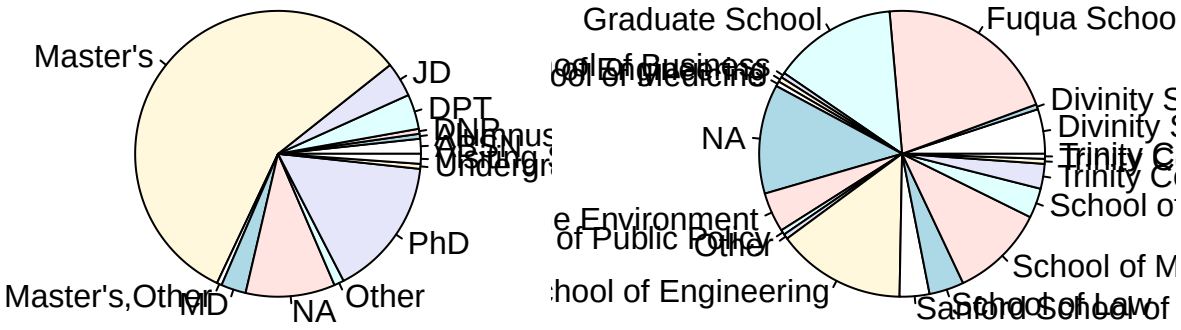
```
plot(table(table(data$uniqueID)),
     ylab="Pantry users",
     xlab="Number of visits")
```



3. What is the composition of degree type (Masters, PhD, etc) and school (Graduate School, Nicholas School, etc) among pantry users? Report the proportions using pie charts and use your best judgment if the categories need to be cleaned up.

Solution.

```
# Note: this is not cleaned up. I have code for nicer doughnut charts.
par(mfrow=c(1,2))
pie(table(representers$degree))
pie(table(representers$school))
```



Furthermore, discuss the following: How could the pantry improve its data collection efforts in order to gain more meaningful insights into its users and their needs? Keep in mind the need to protect the privacy of the pantry users.