Homework 0: Give This A Name!

Olivier Binette and Rebecca C. Steorts, 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.

Introduction

Can you provide a bit more background on the pantry. Who is the pantry? Who does the pantry assist?

A Durham pantry provides food and basic necessities to 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 bag orders through a Qualtrics survey, and the bags are then delivered to them or ready for pick-up on the next Saturday.

As a volunteer statistician at the Pantry, you are tasked to:

1. describe the composition of pantry users through bag order records; and
2. 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 the upcoming Campus Food Insecurity Symposium.

Data

The raw Pantry records consists of weekly bag orders filled between June 24 and November 4 2020 using a Qualtrics survey. This is an online survey made available to all of Duke's graduate and professional students. The survey is mainly advertised over email in periodical newsletters and it records all bag orders for the given time period. Its content 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.

Protecting Privacy of Users I might elaborate here more regarding why it's improtant to protect the privacy and how you did this. You might want to build a small lecture on this where we can talk about the data set with the students in class + how you did the encryption. I think that this would be a great lecture to add into the class and would make it quite exciting. Up to you.

In order to protect the privacy of pantry users, the raw records have been anonymized and only a subset of this data is made available to you. 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 data is contained in the `order_data.rds` file and the anonymization pre-processing script can be found in `encrypt_responses.R`.

Methodology and Goals

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: Deterministic 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 goal is identify unique individuals to help the pantry's operations. To do this, you 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.

Should we put the helper sections in an appendix so they try it on their own and so it looks a bit more clean?

Where can students find the encrypted data? Do they need the real data? I would mentioned at the start of the assignment what data they will work with and where to find it in the homework folder:)

Parts (a)-(g) provide a guided solution. You can also submit a solution of your own. Any correct solution will be accepted.

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?

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`.

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.

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.

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`.

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?

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

Task 2

Construct a data frame where each **row represents a unique pantry user** and each **column is a representative answer to the survey question**. Remind the student where to find the survey questions.

Hint: This is easier than you might think. For example, you can choose the representative answer to be the first non NA answer for this individual.

Parts (a)-(b) provide a guided solution. You can also submit a solution of your own. Any correct solution will be accepted.

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

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.

Task 3

Given your answer in Task 2, answer the following questions:

1. How many individuals have only used the pantry **only once** in the given time period?
2. Plot the distribution of the number of visits per pantry user.
3. What is the composition of degree type and order type among pantry users? Hint: you can use pie charts for simplicity. Could you clarify what you mean about the composition of degree type? Do you mean the degree type such as undergrad, postdoc, PhD or something else?
4. 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? Make sure to keep in mind the importance of the need to protect the privacy of the pantry users.