Intro.

Where is everything?

This document reference files in different places within the project folder. The main project folder is where you can find the folders /data , /reports , /results, /src and README.txt.

README.txt is an important document that document every file in the project folder one by one. So if you have doubts regarding any folder or any document, please check out this file.

The files with sensitive information are only stored in the OneDrive folder. The Google Drive shared folder contain everything exactly the same, but NOT the patient data. It does contain all the documentation regarding variables names for example.

What is SHA1?

Let say you have a file with thousand of rows and thousand of columns, and I also have the same file. But we suspect that our files are different because we are not getting the same results.

Normally, we would have to check each of the million cells one by one until we find the the data that is wrong. But it might be that there is nothing wrong in the data and we are getting different result for another reason.

In order to avoid having to check all the data one by one every time we run a test, we can simply use what is call a hash function, which is this SHA1. This simply a function that takes all the bytes in the data, and generate a unique number based on those bytes. So if you input the same bytes you get the same bytes you will get the same unique number. If the file is exactly the same for both parties, the generated number must also be the same. If it is not, it means that indeed the error for getting different result is because we have different source files.

Original files

The file that I use to read all the data that we use in the scripts is located in:

"../data/aureus/csv/saureus_19022020.csv"

with SHA1:

c4d047e998dd5d3701f4ce416b4fbebcd2da37a0

This data is however still "dirty". Meaning that we have numbers instead of categories for each variable (what does 1 means? man or woman?), everything is mixed together, there are a lot of missing numbers that indicates that something is unknown, and so on.

In order to clean the data, I use the script XXXXXXXXXXX . Later on I will explain what this script does specifically in another section of this document.

Once the data is "clean", is stored in these files:

with SHA1:

Cleaning the data

What is each column?

The original column names are not descriptive at all, so I changed all the names to make then more human readable. Here is a table where I summarize the new column names, and where each name come from.

I also divided the original one file into three tables, joining together variables that makes sense to have with each other within the same topic.

You can find a verbose description for each original column in the metadata file where everything is explained with detail.

Phenotype table

Original Name	New Name	Description
pers_key_ff1	ID	Unique ID for each person.
AGE_FF1	Age	
SEX_FF1	Sex	
BMI_FF1	BMI	The BMI is already stored in the table. There is no transformation from any other variable to find out this value.
HIGH_SCHOOL_MAIN_PROGRAM_FF1	School	
SMOKE_FF1	Smoke	
SNUFF_FF1	Snuff	In Norwegian snus, the powder form tobacco.
PHYS_ACT_LEISURE_FF1	Sports	Self reported sport activity outside school.
PHYS_ACT_OUTSIDE_SCHOOL_FF1	Active	Self reported physical activity outside school.
ANTIBIOTICS_FF1	Antibiotics	Whether a person is taking antibiotics or not.
ANTIBIOTICS_BRAND1_FF1	AntiBrand	Which brand of antibiotic the person is taking.

Network Table

Original Name	New Name	Description
pers_key_ff1	ID	Unique ID for each person.
NETWORK_DATE_FF1	Created	When was this question filled up.
NETWORK_OVERVIEW_FF1	Overwiew	0 to 10, how good this network of friends describe your life.
FRIEND_1_FF1	Friend1	The ID of your selected Friend 1
FRIEND_2_FF1	Friend2	ID of your selected Friend 2
FRIEND_3_FF1	Friend3	ID of Friend 3 and so on
FRIEND_4_FF1	Friend4	
FRIEND_5_FF1	Friend5	
*		
FRIEND1_PHYSICAL_CONTACT_FFX	FriendXPhysical	Did you have physical contact with friend X
FRIEND1_CONTACT_SCHOOL_FFX	FriendXSchool	Did you have contact at the school
FRIEND1_CONTACT_SPORT_FFX	FriendXSport	Contact practicing sports
FRIEND1_CONTACT_HOME_FFX	FriendXHome	At home
FRIEND1_CONTACT_OTHER_FFX	FriendXOther	Somewhere else

* There are 5 friends, so there are also 5 of the same questions for each of the 5 friends

Aureus Table

Original Name	New Name	Description
pers_key_ff1	ID	Unique ID for each person.
DATE_CULTURE_DAY0_FF1	Date	Date of the culture
*		
CONTROL_NASAL_DAY2_FF1	NasalGrowth	Did we try to grow something from the nasal sample.
STAPH_NASAL_DAY2_FF1	NasalAureus	Did something grew from the nasal sample
STAPH_GROWTH_NASAL_DAY2_FF1	NasalPopulation	How much something grew
STAPH_NASAL_ENRICH_FF1	EnrichNasalAureus	Did we try the enrichment process
STAPH_GROWTH_NASAL_ENRICH_FF1	EnrichNasalPopulation	How much something grew after the enrichment
STAPH_COAGULASE_NASAL_FF1	CoagulaseNasal	The coagulase test
STAPH_COAGULASE_THROAT_FF1	CoagulaseEnrichNasal	The coagulase test for the enrichment.
SPA_THROAT1_FF1	SPAThroat1	The different SPA typing variables
SPA_THROAT2_FF1	SPAThroat2	
CC_THROAT1_FF1	SPAThroatClonning	
CCN_THROAT1_FF1	SPAThroatCount	
SPA_NASAL1_FF1	SPANasal1	
SPA_NASAL2_FF1	SPANasal2	

* We have the same variables for the throat.

Transforming the data

The original data is now loaded in our tables. But it has a lot of values as integers that should be transform into categorical variables. In the following paragraph, we explain each transformation. Notice that the 1s and 0s are not constants in between variables, and sometimes 0 means no, and others 1 means no, for example.

Phenotype table

Variable	Original	Transformed
Sex	0	“Woman
	1	“Man”
	Other	“Unknown”
School	1	“Specialization in General Studies”
	2	“Sports and Physical Education”
	3	“Vocational Program”
	Other	“Unknown”
Smoke	1	“Never”
	2	“Sometimes”
	3	“Daily”
	Other	“Unknown”
Snuff	1	“Never”
	2	“Sometimes”
	3	“Daily”
	Other	“Unknown”
Sports	1	“None”
	2	“Light”
	3	“Medium”
	4	“Hard"
	Other	“Unknown”
Active	1	“Yes”
	0	“No”
	Other	“Unknown”
Antibiotics	1	“Yes”
	0	“No”
	Other	“Unknown”

Network table

There are no categorical changes in here for the moment. Technically, there are a bunch of 1s and 0s that represent “yes” or “no”, but I'm going to leave then as they are because is faster to do math with integers rather than strings like the "yes", "no", types. But notice that we still have a bunch of NAs values here which will be dealt with later on.

Aureus table

We describe here only the variables for the nasal samples, but the variables for the throat samples have the same transformation.

Variable	Original	Transformed
NasalGrowth	1	“Yes”
	0	“No”
	9	“Non-applicable”
	Other	“Unknown”
NasalAureus	1	“Yes”
	0	“No”
	9	“Non-applicable”
	Other	“Unknown”
NasalPopulation	0	“Light”
	1	“Moderate”
	2	“Rich”
	NA	“None”
	9	“Non-applicable”
EnrichNasalAureus	1	“Yes”
	0	“No”
	9	“Non-applicable”
	Other	“Unknown”
EnrichNasalPopulation	0	“Light”
	1	“Moderate”
	2	“Rich”
	Other	“None”
	9	“Non-applicable”
CoagulaseNasal	1	“Positive”
	0	“Negative”
	9	“Non-applicable”
	Other	“Unknown”
CoagulaseEnrichNasal	1	“Positive”
	0	“Negative”
	9	“Non-applicable”
	Other	“Unknown”

IDs

The IDs have 8 numbers as identifiers. To avoid visual cluttering and to for the anonymity the data, we substitute the IDs with a integer number that goes from 1 to approximately 1000.

We however save two special IDs. An ID equal to 0 means that a person have a friend that is not in our ID table, for example could be a distant friend from another country that doesn't participate in the study. An ID equal -1 means no friend. This way all the variable for each of the 5 friends have an integer as is easier and faster to do math and filtering later on.

Adding new data

Based on the columns or information that we already have, we are going to add some extra columns in our tables. This is could be new information, like calculating if a person is pain tolerant or not. Or could be redundant information so we can avoid calculating the same value twice in the future and optimize the running time of the script.

Some of the new columns are not in use yet because we don't have the data. As for example, all that reference the pain variables. But they are prepare anyway for later when we have it.

Aureus table

We added a new variable for the nasal, and the throat, that represent if a person is a carrier or S. Aureus or not. For both cases, this is the criteria to decided who is carrier:

If we tried a test on the subject which show growing ("NasalGrowth" == "Yes") and the coagulase test was ALSO positive ("CoagulaseNasal" == "Positive"), then we have a carrier.

There are some minor inconsistencies in the data with this definition, which can be found in the heatmaps done in the control script. But in the vast majority of the cases, all the control variables check-out correctly.

New Variable	Explanation
NasalCarrier	If this person is a carrier in the nose (Positive or Negative).
ThroatCarrier	Same for the throat.

Phenotype table

New Variable	Explanation
*	
OverallFollowingPainAverage	How many friends I nominate that are pain tolerant.
OverallPopularityPainAverage	How many people follow me that are pain tolerant.
OverallConnections	How many undirected friends I have.
OverallPopularity	How many people follow me.
OverallFollowing	How many people I nominated as friends.
OverallReciprocity	How many relationships are reciprocal.
OverallFriendsWithBugNasal	How many undirected friends have S.Aureus in the nose.
OverallFriendsWithBugThroat	How many undirected friends have S.Aureus in the thoat.

* We have 6 networks in total. The overall network, the physical, school, sports, home, and other network. This variables repeat for each of the network (but changing the name to reference each obviously).

Network table

We are not adding anything new to the network table. But the current form is not useful to do matrix operation. So we separate it into 6 different matrices of N x N, where N is the number of people we have in the phenotype table. The matrix has a 0 in each cell by default.

Each row correspond with the people that row ID nominate as a friend.

Each column correspond with the people which that column ID is popular.

So for example, if matrix[3,7] = 1 means that person number 3 likes person number 7. If matrix[7,3] = 0 it means that 7 does not reciprocate the relationships.