Week 2 GCD Assignments week 2

Kevin Jetten

i318843

# Favorite Chocolate

Validity - Most of the data is quite valid

Accuracy – A lot is a type of chocolate

Completeness - Only very few records are empty.

Consistency - There is a lot of consistency

Uniformity - There is barely any uniformity

# Cleaning strategies

## Gender

Make everything lowercase

Transform things like "m" or "f" to "Male" or "Female"

Remove unconventional answers like "Helicopter"

Transform the different languages to english

Check for typo's (Meel to Male)

## Chocolate

Make everything lowercase

Translate all the languages to english

Grouping of types of chocolate

Remove unconvitional answers like "Stay away from my children"

Sort by brand

## Reasoning

Lowercasing everything: This way you create a lot more consistency without actually changing anything that’s value to the data.

“m” to “M”: This creates a lot more uniformity throughout the data again without changing the value that the data holds.

Translate to one language: The chance that something is lost in translation is very little because the answers are all fairly small. Having everything in one language creates a lot more consistency and makes it easier to use the data in a later point.

Removing unconventional answers: This might be dangerous especially when you take a look at gender. Some genders you might see as unconventional but are real genders for other people.

# Which Missing Data Mechanisms (MCAR, MAR, NMAR) do you encounter? Explain.

There are two kinds of missing data that we do encounter: Missing completely at random and Missing at random. We actually don’t encounter Not missing at random.

**MCAR:**

There are some entries that have empty field while the other fields are filled in correctly, it looks like they have just forgot to fill in the answer.

The strategy to still use these records would be to not take in these fields but still use the others fields where they are filled in.

**MAR:**

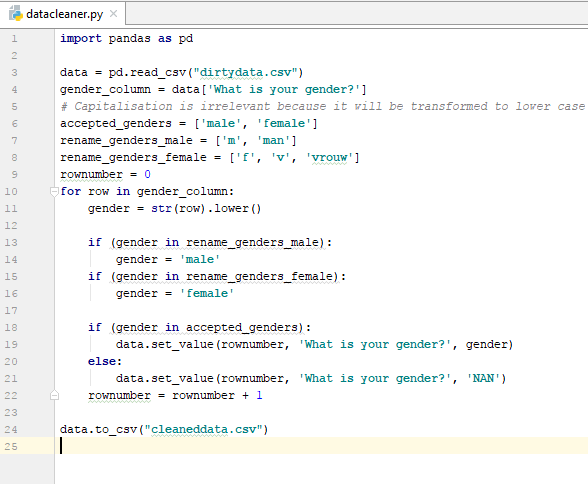
There are some rows that have, for the most part, correct data in them but they fill in the gender field with entries like “Attack helicopter” which is clearly a joke. This could have been easily avoided by having a dropdown list with genders instead of a text field.

The fields which are related to income are often also empty. I think this is not completely random but that the target audience that filled in the survey just don’t know what they can expect when it comes to their earnings in the future. It might be best for the data to make an extra column “I don’t know” based on if these fields are empty. This way you can still have valid numbers when it comes to what people think they will earn and you get more insight on how many people haven’t have a clue

**NMAS**

I don’t think there are any NMAR except maybe when you count the records that are not present at all. The NMAR part is that it is not random that people are too lazy to fill in a survey. If you’d know how many people were supposed to fill in the survey you could use this information to calculate how many data is missing because people just didn’t fill it in.

## Script



In this case I will accept only the genders “male” and “female”. First of all I get the column data and make it lower case, this way it’s easier to check instead of having to check for example “Male” an “male”.

The second step is to check for abbreviations and different languages (in this case its Dutch). The accepted values are stored in arrays declared at the start of the script. I change the original values to the uniform.

If the data is accepted its value will be set and when all the rows are done everything will be saved to a new file named ‘cleaneddata.csv’.

I think its save to disregard other genders because they will most likely be joke answers and therefor unreliable data.