COMP8430 Assignment 2 Report

uid: *u5976992*
word count: *637 words*

Part 1 - Merging the data sets

1. Choose attributes

The first thing we need to do is to decide which attributes we need. Since we want to find the links between an individual's education, employment history, and their health, some irrelevant personal information is useless, such as name, address, phone, email. We can remove those attributes.

For *emp-edu.csv*, I choose attributes as follow:

- ssn
- gender
- birth_date
- education
- occupation
- salary
- year_of_experience
- employment_timestamp

For medical.csv, I choose attributes as follow:

- ssn
- height
- weight
- blood_pressure
- cholesterol_level
- deceased_status
- consultation_timestamp

Here explain why I choose those attributes:

1. `ssn`: as the key to merge the data sets.
2. `gender`, `birth_date`: I think when we find the links between education, employment history, and health, we need to consider the gender and age. For instance, men and women maybe have different weight standard for healthy. Age will be related to education, employment history, and health.
3. `education`: education information
4. `occupation`, `salary`, `year_of_experience`: employment history
5. `height`, `weight`, `blood_pressure`, `cholesterol_level`: health information
6. `employment_timestamp`, `consultation_timestamp`: I will explain it in part 3
   Notice that I don't choose `bmi` since it can be calculated by `height` and `weight`. Besides, I don't think `deceased_status` can reflect health well without detail. For example, maybe someone just died by the car accident. Therefore, I also don't choose `deceased_status`.

2. Union `ssn`

Since we want to find the links between education, employment history and health, those records that only occurred in a single data set are not helpful. We can remove those records. In other words, we can union `ssn` from the data sets and use it as the keys to merge the data sets.

3. Remove duplicate records

After selecting attributes, I remove duplicate records for each data set.

4. Inconsistencies

The original data sets have many inconsistencies. For instance, some person has different values of `middle_name` or `address` in two data sets. However, after removing some personal attributes, there is no inconsistency at all.

Part 2 - Missing and incorrect values

1. Missing values

According to my experiment, most of the records that contain the missing value are

duplicated records. Therefore, I just remove all records that contain the missing value. It turns out that it basically won't change the number of record of the result.

2. Incorrect values

In the `birth_date`, most of the values use the data format DD/MM/YYYY, while some data use the data format MM/DD/YYYY. I change those data that use the data format MM/DD/YYYY to use the data format DD/MM/YYYY.
There are some values are negative in `height` and `weight`, which is impossible. I remove those records since most of those are duplicated records. Notice that there are 83 records that are not duplicated records. Since the whole number of records is more than 40000, I don't think changing those illegal values to mean or median is a good idea.

Part 3 - Other data cleaning

1. Data quality dimensions

- Timeliness/Usability: I think if two timestamps are too far apart, merging them together will no longer make sense. For instance, for the person whose `ssn` is `d107285434`, we have the employment history of 2018 but the health information of 1997. Using those two data together to find the link makes no sense.
- Validity: In the attribute `phone`, most of the values are 10 digits, while some values are 8 digits.

2. Data reduction

I compare two timestamps for each record, if the time interval is larger than 5 years, I will remove this record. I choose 5 years just for intuition and it will reduce about 2000 records.

3. Values in the wrong attribute

In *emp-edu.csv*, some email values fill in the attribute `phone`.