Q1

1.1

S1 = ‘1236773’

S2 = ‘1233547’

Q1 = [‘1’, ’2’, ’3’, ’6’, ’7’]

Q2 = [‘1’, ’2’, ’3’, ‘5’, ’4’, ’7’]

Intersection (Q1, Q2) = [‘1’, ’2’, ’3’, ’7’] = 4

Union (Q1, Q2) = [‘1’, ’2’, ’3’, ’4’, ’5’, ’6’, ’7’] = 7

Sim-Jacc (S1, S2) = 4 / 7 = 0.57

1.2

S1 = ‘1236773’

S2 = ‘1233547’

Q1 = [‘12’, ’23’, ’36’, ’67’, ’77’, ’73’] = 6

Q2 = [‘12’, ’23’, ’33’, ’35’, ’54’, ’47’] = 6

Intersection (Q1, Q2) = [‘12’, ’23’] = 2

Sim-Dice (S1, S2) = 2 * 2 / (6 + 6) = 1 / 3 = 0.33

Q2

2.1

The number of SSN occurred in common in both data sets is 15991, because there are 785 duplicate records in education data set, there are only 3224 values in the education data set and there are 4009 only in the medical data set.

2.2

I merged two data sets by inner and SSN as key. If there were records (except SSN) that only occurred in a single data set, they will merge to one row (with the same SSN), but for any SSN that only occurred in a single data set, the row will be deleted, because it is meaningless to the row combine from two datasets but one of datasets is empty.

2.3

There are 785 duplicate records with same SSN in the education data set, I dropped duplicate records and keep the newest record. People maybe duplicate registration, therefore, choose the newest record is the best way.

2.4

HINT: attribute name(number of inconsistent records)

There are 12 same attributes in two datasets, include rec id(15991), first name(0), middle name(305 5), last name(42), gender(1710), birth date (0), street address(6803), suburb(6709), postcode(8554), state(2902),phone(9334),email(7519).

For type of object, there are rec id, first name, middle name, last name, gender, birth date, street ad dress, suburb, state, phone, email. For type float, there is postcode.

For rec id, I kept both, they are two datasets so that it is necessary that keep their independent rec i d. For other data, my approach is based on consultation timestamp. If there were inconsistent recor ds with the same SSN, I will replace old records with new records. For example, people maybe move to a new place and change their phone number, using newest records is a useful way.

Q3

3.1

The combination of three attributes with the highest number of missing values are:

marital status, occupation, and credit card number, which has 1346 missing values.

marital status, occupation, and phone, which has 212 missing values.

occupation, credit card number, and phone, which has 212 missing values.

3.2

For the two attributes with the highest number of missing values (individually) are salary and phone, which have 2390 and 2220 missing values. For salary, my approach is using “nan” to replace them so that it will not affect the accuracy when calculating the average salary, etc.

3.3

I found there are 1589 records of weight are negative numbers. Absolutely, human weight should be a positive number.

3.4

I used the BMI formula, and the formula is “BMI = kg/m^2”. For each incorrect weight record, I used BMI and height to calculate the new weight.

Q4

4.1

Yes, there are some “-9999” in salary and there are some postcodes only have three digitals. My approach is using “nan” to replace them so that it will not affect the accuracy when calculating the average salary, etc.

4.2

Yes, for cholesterol level, it is difficult to make accurate judgments due to complex digital data. Ther efore, my approach is that transfer them to two different types: normal and abnormal. When choles terol level lower than 200, it is a normal cholesterol level, otherwise, it is a abnormal cholesterol leve l.

4.3

Yes, I think Medicare number is personal privacy and it does not helpful for analysis, because it is a random string of numbers. Hiding Medicare number is an effective method to privacy protection. My approach is that delete the column of Medicare number.

4.4

Yes, there are some postcodes only have three digitals, for example, there are 9 results of “800” from Australia Post checker, my approach is using “nan” to replace to any uncertain postcode.

Yes, there are 1624 gender records that are contradictory, for example, with the same SSN number, it is male in the education dataset, but it is female in the medical dataset. Considering the transgender people, I used the newest record as their gender.