1)
[in the following, don’t insert return characters into the strings being hashed, otherwise you may get a different value from the online SHA-256 calculator]

Question mark in row1:
=SHA-256(COMP20008)

=23a602232c74b3e00a31ed1eddda091669f77acf24f2db04591e74259047e6ba

n.b. One can demonstrate that

SHA-256(000+1205170931+
23a602232c74b3e00a31ed1eddda091669f77acf24f2db04591e74259047e6ba)=block ID of first block

----

3rd Question mark in row2=

SHA-256(COMP20009)=

611bf18429335a9a544f5734b0c8b3081a1c304247d7b61226bad8777551456c

2nd question mark in row2=block id of parent block=

a2488feb8593c533d55481d1a0026f563801b63778a36b93d0b4f1763393b0c2

1st question mark in row2=

SHA-256(a2488feb8593c533d55481d1a0026f563801b63778a36b93d0b4f1763393b0c2+
1305171033+611bf18429335a9a544f5734b0c8b3081a1c304247d7b61226bad8777551456c)

=e8e801a51cb7683d81b5ebb5275b4aa12b24a2bc39655bebf4543b3bc25bd2a2

3rd question mark in row3=

SHA-256(COMP20010)= cd76ca0b4c2691071a40c5b1c4a21486757c3f32ac5b4c5c4ffbf6d1a8d28bc0

2nd question mark in row3=block id of parent block=

e8e801a51cb7683d81b5ebb5275b4aa12b24a2bc39655bebf4543b3bc25bd2a2

1st question mark in row3=

SHA-256(e8e801a51cb7683d81b5ebb5275b4aa12b24a2bc39655bebf4543b3bc25bd2a2+1405172333+cd76ca0b4c2691071a40c5b1c4a21486757c3f32ac5b4c5c4ffbf6d1a8d28bc0)=

d6f384aa4bfc506d7a0c345b5c2535e9ac8cc09a25c89b6be617bc0a2b0641ef

------

2)i) This modification will require a change to the header of first block (because that includes a hash of the data).   A change to the header of first block, will require a change to its block id (which is a hash of the header). A change to the block ID of first block will require a change to the header of second block (which records its parent’s id). A change to the header of second block will require a change to its block id (which is a hash of its header).   A change to id of second block will require a change to the header of third block (which records its parents id). A change to the header of the third will require a change to its block id (which is a hash of its header) ……

The bottom line: a cascading series of changes to headers of all the blocks after the first block.

ii) if timestamp for second block was modified, this would cause a change to its block id (which is a hash of its header).   A change to id of second block will require a change to the header of third block (which records its parents id). A change to the header of the third will require a change to its block id (which is a hash of its header) ……

[so similar behavior to part i), except that cascade begins at second block]

So the general conclusion is that if an adversary wants to change information recorded a long time ago (in early blocks), it is very hard, because it requires complete recomputation of all subsequent blocks. In practice, this will be infeasible, since the blockchain is replicated across computers in a P2P network, and it would require a node to convince all other nodes in the network to substantially revise their list of blocks.   There are mechanisms in place for the blockchain that make it demanding to convince other nodes to update their most recent block, let alone revise all previous blocks[these consensus mechanisms are beyond the scope of the subject]

3) This question is about understanding how digital signatures work. See the digital signature diagram from the lecture notes (slide from the blockchain lecture titled "Digital signature (from wikimedia commons)".

If the document is modified by an adversary, then this will change its hash value. This hash value will then not match the contents of the decrypted digital signature. The way public key cryptography works, no-one else knows Bob’s private key. Using Bob’s public key to decrypt a message, will only work if the message was encrypted using Bob’s private key, but not one-else should know this private key, so we can trust that the message which has been decrypted, was originally created by Bob. [you should be able to understand the diagram and the high level notions of public/private keys, but it’s not necessary to know the details of the encryption/decryption mathematics, which is based on use of elliptic curves]

4a)

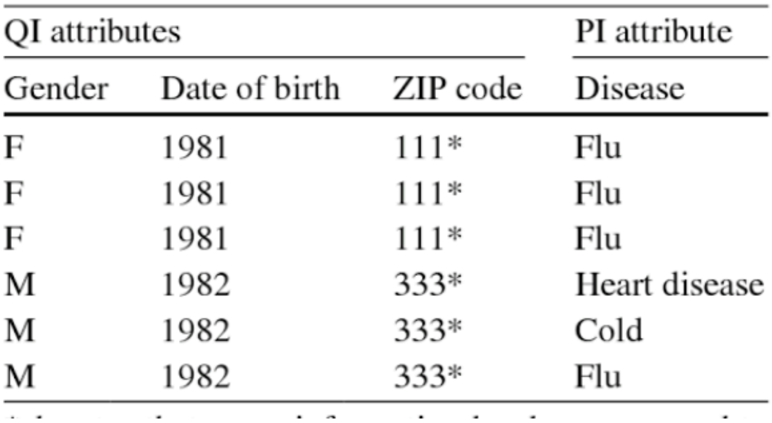
1 anonymous:yes

2 anonymous: yes

3 anonymous: no

4 anonymous:no

b) One possible answer. We would first get rid of name attribute before doing any further processing, we then just work with the quasi idnetifiers and sensitive (attribute)



QI attributes			PI attribute
Gender	Date of birth	ZIP code	Disease
F	1981	111*	Flu
F	1981	111*	Flu
F	1981	111*	Flu
M	1982	333*	Heart disease
M	1982	333*	Cold
M	1982	333*	Flu

c)

this example is taken from

https://www.rand.org/content/dam/rand/pubs/working_papers/WR1100/WR1161/RAND_WR1161.pdf

1 anonymous:Yes

2 anonymous:Yes

3 anonymous:Yes

4 anonymous:No

5-anonymous:No

1-diverse:Yes

2-diverse:Yes

3-diverse:No

4-diverse:No

5-diverse:No

5)

See the slides. Global sensitivity is evaluating the maximum possible change in query output due to a presence of a single record. The privacy budget determines how close the query result for a database with the record is expected to be compared to query result for a database without the record. For smaller k, or larger global sensitivity, more noise will be added to the query result.

6)

In the first case

Adding or removing any individual can affect the count of each column by maximum 1.
The maximum difference a single record can make query is 1+1=2

In the second case

Adding or removing any individual can affect the count of each column by maximum 1. If it affects the count of some column by one, then it will affect the counts of other columns by 0 (since the columns are mutually exclusive, you can't have a 1 in both columns)
The maximum difference a single record can make To F is thus 1+0+0+0=1

7)

Instance id	1	2	3
1	0	Sqrt(200)	Sqrt(250)
2	Sqrt(200)	0	Sqrt(650)
3	Sqrt(250)	Sqrt(650)	0

Visualising as a reordered heat map using algorithm such as VAT – identify the cluster structure of the data. i.e. How many clusters there are and which objects are in each cluster. Might also help identify anomalies – which objects are not similar to other objects.

Little apparent benefit in applying PCA – the dataset is only 2 dimensions and already easy to visualize.

8) i)

Use a metric that finds aggregate deviation from the true answers. E.g. something like Mean squared error=(1/100)* SUM_{i=1}^{i=100} (true_value(x_i)-imputed_value(x_i))^2

where i is an index that ranges over the 100 missing values.

Could also use the mean absolute error as well (average of the absolute values of the deviations)

ii) Reasons it might be better to discard

-we already have a large dataset, that contains sufficient information even when examples are discarded.

-if imputation method is likely to be computationally expensive (e.g. matrix factorization), then might choose discard instance if efficiency is important

-if we believe imputation is likely to cause problems or contaminate later analysis (due to its unreliability)

-scenarios where each instance is either complete (has nothing missing), or has mostly missing values.