Problem 1

a) From Binary to Decimal:

i) Each 0 or 1 is multiplied by its base (2 in this case) to the power of its position with 0 being the units' position and decimals go forth from -1.

256
128
0
32
16
8
0
0
1
0.5
0
0.125
0
0.03125
0.015625
0.007813
0
0.001953

Thus, the total is $(441.681640625)_{10}$.

ii) The same is done with the next binary number, noting the fraction point position in each case (the Excel sheet for such calculations is provided). The computed decimal number is $(613.40625)_{10}$.

b) From Decimal to Binary:

i) First, the integer part:

100	0
50	0
25	1
12	0
6	0
3	1
1	1

Then, the fraction part:

0.02	0
0.04	0
0.08	0
0.16	0

0.32	0
0.64	0
1.28	1
0.56	0
1.12	1
0.24	0

Therefore, the binary number is $(1100100.0000001010)_2$.

ii) Same as well was done for the second number, noting that after the third fraction part multiplication by 2 would result in exactly one, leaving all the following terms in the 10 fractional points to be zeroes. The determined binary number $(1000000.1010000000)_2$.

c) According to

$$(-1)^s \times 2^{c-127} \times (1.f)_2$$

The single-precision IEEE standard floating-point representation is (steps are evident in the formula used in the Excel sheet)

0	10001001	00000010100000000
s (sign)	c (biased exponent)	f (mantissa)

Problem 2

$$\because z = xy$$

$$\text{fl}(z) = \text{fl}\big(\text{fl}(x) \cdot \text{fl}(y)\big)$$

$$= \Big(x(1+\delta_x) \cdot y\big(1+\delta_y\big)\Big)(1+\delta_z)$$

$$= xy + xy\delta_y + xy\delta_x + xy\delta_x\delta_y + xy\delta_z + xy\delta_y\delta_z + xy\delta_x\delta_z + xy\delta_x\delta_y\delta_z$$

Absolute error:

$$\text{fl(z)} - z = xy\big(1 + \delta_y + \delta_x + \delta_x\delta_y + \delta_z + \delta_y\delta_z + \delta_x\delta_z + \delta_x\delta_y\delta_z\big) - xy$$

Removing $O \geq O\big(\delta_x\delta_y\big)$ terms where $x \neq y$

$$\text{Absolute error} = xy(\delta_x + \delta_y + \delta_z)$$

Relative error:

$$\frac{\text{fl(z)} - (z)}{z} = \delta_y + \delta_x + \delta_x\delta_y + \delta_z + \delta_y\delta_z + \delta_x\delta_z + \delta_x\delta_y\delta_z$$

Again, removing $O \geq O\big(\delta_x\delta_y\big)$ terms where $x \neq y$

$$\text{Relative error} = \delta_x + \delta_y + \delta_z$$