Question1

Part(b): This operation introduces truncation error because we are dropping from infinite terms to 100 terms in our approximation.

Part(c): f(0.00000001) is 0 in this case since Python uses double precision floating point numbers. We have only 16 digits available. When x is 0.00000001, tan(x) will be rounded to x. So we have 0 as the numerator, which results 0 regardless of the denominator. f(0.00000001) should be really close to -0.333.

Part(f): For x values around multiple of pi, evaluating f(x) would be highly sensitive since Kf(x) is much larger than 1 when taking these x.

Question2

Part(c): Our implementation is associative. With chopping as the rounding method, the machine precision for F(β=4,p=6,L=−4,U=4) is 4^{-5}, which is not representable in this system since the exponent is less than L. Therefore it is impossible to break associativity using some float number less than machine precision in this implementation.

Part(d): The length of `all_floats` should be 55297. We have 2 choices for the sign; 3 choices for leading digit; the rest of the mantissa have $4^5 = 1024$ possible combinations; the exponent has 4-(-4)+1=9 choices; finally we add one for number 0. We have 2*3*1024*9 + 1 = 55297 numbers.

Question3

Part(a): Catastrophic cancellation occurs when computing x1. Since $4ac$ is really small compare to b^2, so $\sqrt{b^2 - 4ac}$ is approximately b. Therefore the numerator has catastrophic cancellation as it is a subtraction with 2 similar magnitude numbers. On the other hand, no catastrophic cancellation occurs when computing x2 since the numerator is a subtraction of numbers with different signs.

Part(d): The only positive integer for which the value of `z(n)` results non-zero is 28. After the pow operation of `z(n)`, the value stored become 4^{28}. If we add 9.0 to 4^{28}, (2^{56})the result takes 53 bits of mantissa(1+52). According to IEEE Double Precision for floating point arithmetic, it will round to even, which produces a rounding error.

Question4

Part(b):

$$A = \begin{bmatrix} 1 & 3 & 4 \\ 2 & 7 & 9 \\ 4 & 11 & 17 \end{bmatrix} \quad \vec{b} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

①: $R_2 \leftarrow R_2 - 2R_1$
$R_3 \leftarrow R_3 - 4R_1$

$$A = \begin{bmatrix} 1 & 3 & 4 \\ 0 & 1 & 1 \\ 0 & -1 & 1 \end{bmatrix} \quad \vec{b} = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

②: $R_3 \leftarrow R_3 + R_2$

$$A = \begin{bmatrix} 1 & 3 & 4 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{bmatrix} \quad \vec{b} = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

Backward Substitution

$2x_3 = -1$

$x_3 = -\frac{1}{2}$

$x_2 + x_3 = 0$

$x_2 - \frac{1}{2} = 0$

$x_2 = \frac{1}{2}$

$x_1 + 3x_2 + 4x_3 = 1$

$x_1 + \frac{3}{2} - 2 = 1$

$x_1 = \frac{3}{2}$

$$X = \begin{bmatrix} \frac{3}{2} \\ \frac{1}{2} \\ -\frac{1}{2} \end{bmatrix}$$