CSCI 2330 – Floating Point Exercises

1. Using an 8-bit IEEE floating point representation (with k=4 exponent bits and 3 fractional bits), convert **00110100** into a decimal value.

2. Using the same 8-bit representation, convert **10000101** into a decimal value (working with a fraction here is advisable).

3. If **d** is a double, does (d < 0.0) imply that ((d * 2) < 0.0)? Remember that this property is <u>not</u> guaranteed for integer types.

4. Excluding infinity, write down an expression giving the exact decimal value of the largest 32-bit IEEE floating point number (no need to simplify the expression).

5. IEEE 754 encodes the exponent value **E** using the **exp** bits as an unsigned value from which **bias** is subtracted. An alternative approach would be to just make the **exp** bits encode a signed value and get rid of the **bias** term. Is there a reason to prefer the **unsigned - bias** approach?

Hint: This is related to how floating point values can be compared. As an example in the above 8-bit format, consider 01000000 and 00100000 in either the signed or unsigned **exp** format. Which bit pattern encodes the larger value in each format?