Files

a4.py: Program to create bitmap index and perform WAH compression
README.txt: How to use a4.py

File	Size (in Bytes)	Runs	Literals	Compression Ratio
animals	1,800,000			
animals_WAH_8	1,558,008	76,429	152,147	0.8656
animals_WAH_16	1,661,760	14,025	92,647	0.9232
animals_WAH_32	1,650,016	1,271	50,345	0.9167
animals_WAH_64	1,626,144	26	25,382	0.9034
animals_sorted	1,800,000			
animals_sorted_WAH_8	51,744	226,996	1,580	0.02875
animals_sorted_WAH_16	56,736	104,962	1,710	0.03152
animals_sorted_WAH_32	115,616	49,838	1,778	0.06423
animals_sorted_WAH_64	227,232	23,604	1,804	0.1262

The bitmaps that are sorted and compressed are significantly smaller in size compared to the unsorted maps (as low as 2% of the original file). For unsorted bitmaps, the sizes only slightly varied between word sizes (about 7% difference at max, comparing sizes 8 and 16). While for sorted bitmaps, certain word sizes had huge differences from other sizes. For example, **sorted_WAH_8** was 80% smaller than **sorted_WAH_64**.

Sorting most definitely helped with compression. Sorted bitmaps (excluding the uncompressed one), totals up to 451,328 bytes, while unsorted bitmaps (also excluding the uncompressed one) totals up to 6,495,928 bytes. That makes sorted take up only about 6.95% of the storage that unsorted does. Sorting helps reduce size during compression because it groups similar data together, meaning runs of similar values will be grouped up moreso than unsorted. In other words, sorting maximizes runs while minimizing literals.

Different word sizes had different compression ratios (word size determines the number of runs/literals that can be stored in a single word). Having a smaller word size might create an issue where the compression will need to use multiple words to create one run

or literal, which leads to a larger file size. On the contrary, a larger word size might run into the issue of wasting space if the runs or literals are too small, which also leads to a larger file size. In the case of this project, the data used for compression was small enough for word size 8 to be the most effective size.