Language Assignment #4: Awk

Issued: Wednesday, November 13
Due: Wednesday, November 20

Purpose

This assignment asks you to begin using a scripting language named Awk, which is part of every Unix distribution. Awk was developed by, and named after, Al Aho, Peter Weinberger, and Brian Kernighan, from Bell Labs, in 1977. Awk is procedural and imperative, but based on a stream-editor model of computation.

The GNU distribution of Awk is called Gawk, it is actively maintained by Arnold Robbins.

Documentation

Awk lecture slides are at:

```
1 pub/slides/slides-awk.pdf
```

Awk is briefly described in Section 14.2.2 of our textbook.

The onyx cluster has an Awk interpreter, the documentation of which can be viewed by:

```
1 man awk
2 info awk
3 info gawk
```

and demonstrated by:

```
1 pub/sum/awk
```

Assignment

Suppose you work for a realtor (my condolences) and your employer wants to put Ada County building-permit information on the company web page.

The following filename extensions are relevant:

`.xlsx`	Microsoft Excel Spreadsheet
`.html`	HyperText Markup Language
`.csv`	Comma-Separated Values

The building-permit data is public, but, of course, only as a `.xlsx` file. You want to process the data, eventually producing a `.html` file. You decide to use the LibreOffice program `unoconv` to batch-convert the `.xlsx` file to a `.csv` file, and then process it with an Awk script to produce the `.html` file.

Ada County provides the `.xlsx` file, from:

```
1 https://adacounty.id.gov/Development-Services/
2       Building-Division/
```

I provide the corresponding `.csv` file, at:

```
1 pub/la4
```

You need to write the Awk script to produce a simple `.html` file (see below). You can view your result with Firefox.

Hints and Advice

- Use patterns/actions.
- Do not overlook Section 4.7 of (Edition 4) the Gawk manual (Defining Fields By Content). It contains the ominous "The most notorious such case is so-called 'comma separated value' (CSV) data."
- Your script should read from `stdin` and write to `stdout`. Use no other files.
- Keep your HTML simple. There is a sample skeleton at:

  ```
  1 pub/la4
  ```

Simple headings are nice, but don't get carried away.

- Your employer only cares about single-family dwellings, but watch out for scruffy human-generated data. Case conversion and regular expressions can help. Furthermore, only date, subdivision name, lot, block, and value are important.