data-mining-homework-1

This README acts as the log addressed in the assignment.

Authors

Evan Conrad
Ryan Rozema

Log

Resolving Cases

1. `audi 100ls` was changed to `audi 100 ls` and then changed.
2. Misspellings were corrected to their nearest counter part.
3. When given duplicate rows, we googled the answer and took the closest result to what we found

Hiccups

On Monday night, we thought we had finished when in reality our join was broken. This caused a whole bunch of other problems and eventually required a rewrite of much of the program. But it works now.

Steps

Step 1

We downloaded the files.

Step 2

We wrote functions to count the number of instances in `hw1.py`. It may have been better to focus on maintainability than efficiency. While we could do everything in one for loop, it became quite the code nightmare after a bit.

When we found duplicates, we removed them by doing a Google search for the nearest attributes. Example duplicates:

- `Ford Pinto`
- `AMC Gremlin`
- `audi 100 ls` (NOTE: this was actually found in step 4)

Step 3

We wrote a full outer join. Our first join did not take into account that one file might be longer than the other. It also did not correctly write to CSV and would write everything as a python print of a list would show. We also weren't correctly converting the elements to their respective types when we were reading the array.

Step 4

At this step, we chose to ignore cases where there were prices but no mpg data but by taking the first choice in Step 6: removing all instances with missing values. We felt like we were making up information we did not know when we computed averages. In this section, we found another duplicate/misspelling we had not caught in Step 2: `audi 100ls`.

Step 5

We computer summary statistics, but they were initially incorrect due to our reading of the files incorrectly. We also misread the tip on using the tabulate module, so we wrote our own formatting system. It works by using pythons `ljust` function that adds whitespace padding to the right side of a string.

Step 6

For the third step, we chose to compute averages based on the rows year. We figured this would give us the most accurate description of the car's statistical atmosphere.

Files

hw1

The main file. Does the homework. Requires running and cleaning of the files.

util

A collection of general functions that are used across the project.

Join

Performs a full outer join.

Summary

Handles calculating and printing summary statistics

clean

Handles Step 6. Can remove all instance, compute averages, or compute averages by a year.

filesystem

A utility file that can be used to pull in a file dictionary/obj.
Cannot be run, must be imported.