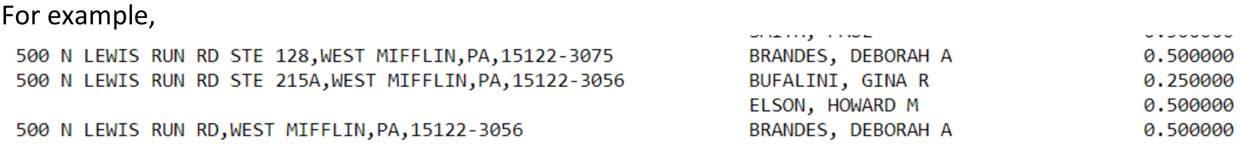
Pennsylvania Data Notes

The Issue of Address Variations

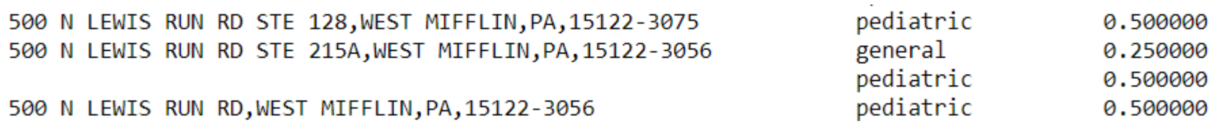
Similar to other states, Pennsylvania data has many inconsistencies in address input, such that each address can have more than one version if an address is associated with multiple providers.

For example,



```
500 N LEWIS RUN RD STE 128,WEST MIFFLIN,PA,15122-3075        BRANDES, DEBORAH A          0.500000
500 N LEWIS RUN RD STE 215A,WEST MIFFLIN,PA,15122-3056       BUFALINI, GINA R            0.250000
                                                             ELSON, HOWARD M             0.500000
500 N LEWIS RUN RD,WEST MIFFLIN,PA,15122-3056                BRANDES, DEBORAH A          0.500000
```

Looking at the same rows grouped by specialty:



```
500 N LEWIS RUN RD STE 128,WEST MIFFLIN,PA,15122-3075        pediatric        0.500000
500 N LEWIS RUN RD STE 215A,WEST MIFFLIN,PA,15122-3056       general          0.250000
                                                             pediatric        0.500000
500 N LEWIS RUN RD,WEST MIFFLIN,PA,15122-3056                pediatric        0.500000
```

Row 1 and 3 appear to belong to the same provider. However, due to the address variations being treated as distinct, the provider Brandes is listed as having .5 head in each place. If these two lines of addresses are to be combined, the head value should be summed. It would not be right to simply delete one of the two rows.

Row 2 has a different suite number and different providers, and thus should be treated as a separate practice.

There were suggestions regarding whether GPS, rather than addresses, is a better index for grouping. From my observations, I found that GPS values can also have variations for the same address if the latter is spelled differently. Therefore, using GPS as index does not solve the address variation problem.

Grouping Output Views

The users of the FTE output should consult the html file out put which gives various views of how data is grouped before it is reshaped into the final FTE format.

Please also consult data notes from other states I have worked on, especially Texas.

Python Pandas cannot handle data inconsistencies caused by inconsistent human input. It is up to the user of the FTE files to gain an understanding of how data is processed and find ways to adapt to the situation such that FTE data is used correctly and effectively.