**10-3-23 Notes**

Exploring Pandas

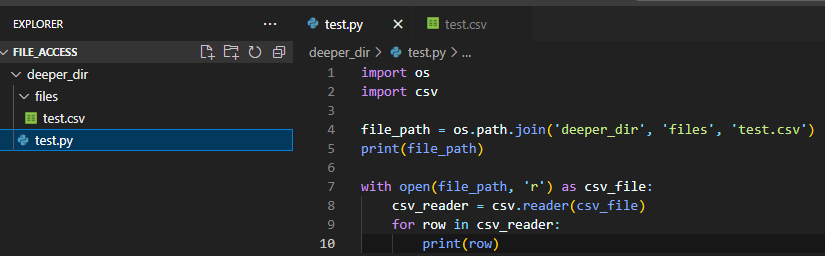
Goals:

By the end of this lesson, you will be able to:

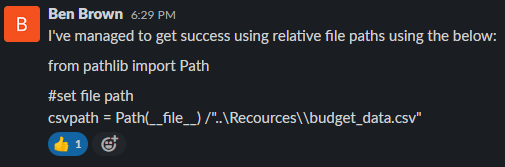
* Navigate through DataFrames with loc and iloc.
* Filter and slice Pandas DataFrames.
* Create and access Pandas groupby objects.
* Sort DataFrames.

Python- os.join- THIS IS HELPFUL FOR PYTHON CHALLENGE





Alt method: Import path from pathlib



Loc is short for location.

Iloc is Index location.

In the example below, the street name column is unique.



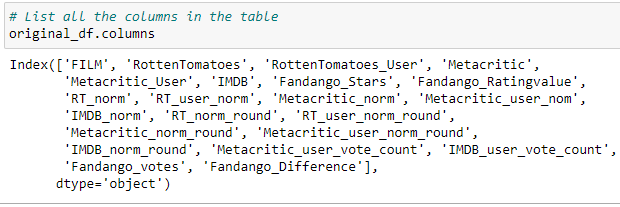
If no series of unique identifier, you can use  to have pandas make a series in the working area made of just numbers for the index

Loc & Ilock examples:

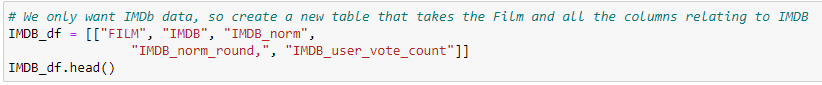


**Good Movies Activity solution:**

List column names.

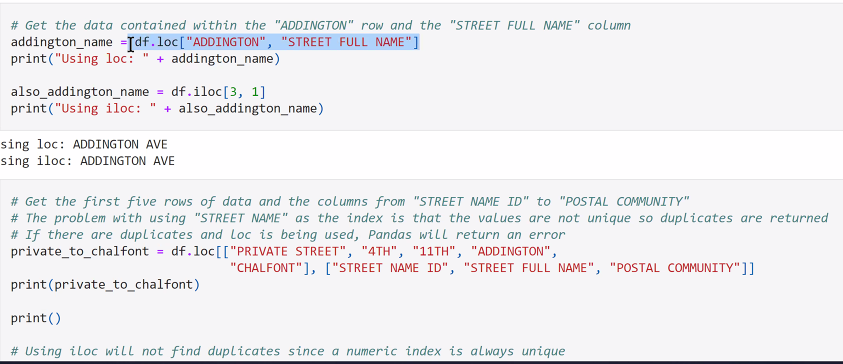


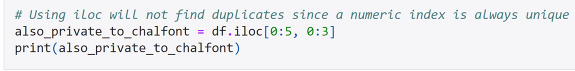
Make a new data frame based on columns of parent dataframe





^ original\_df = pd.read\_csv(movie\_file) works, but not original\_df = pd.read\_csv(file)





^ =df.iloc[0:5, 0:3] is a Slice

* returns rows 0 through 4, columns 0 through 2

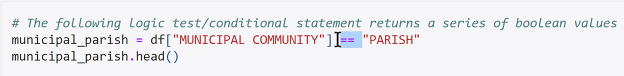


^ above is a chain of steps (6:54pm in class recording)

^ highlighted colon is pandas symbol for All. In this case, it’s in the Rows place of the code

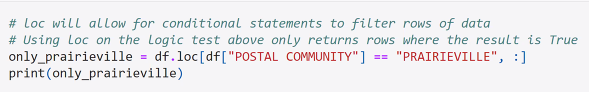


^ 1:3 means pulling columns 1 and 2, but not including column 3

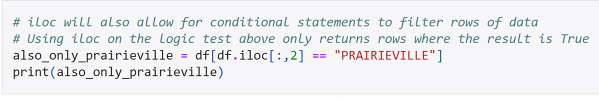


Two equal signs means Compare. One would be to assign

In case above, it’ll indicate for each record if the argument is true or false

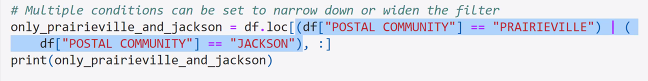


^ This example is returning values where the argument is true



^ This example is a way of writing the same procedure using iloc instead of loc.

^ iloc is much more common in the working world

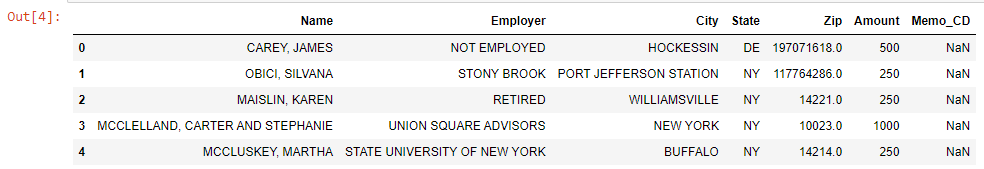


^ The vertical purple line (called a Pipe) stands for the word or

If you type in a ~ in the df area, it’ll do the inverse of the command. Ie: if you’re selecting one series, it’ll select all series except that one.

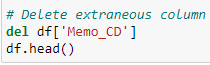
**NO FANCY CODING (LOOPS) IN DATA FRAMES. It’s all simple.**

**Cleaning Data example:**



^ Pandas will put NaN in a blank cell if there’s not any data in the cell. See Memo\_CD column on right.

How to Delete a series:





^ This will drop ANY row if there’s a single blank/NaN/Null cell

^This will also even out the count for the number of records for each column to the same number.

^ If you use how=’’ instead of how=’any’ it’ll only delete rows that are completely blank



^Find and Replace

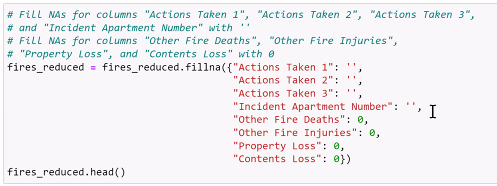
**Hong Kong LPG solution:**



In example above- you can put the ~ after the loc[ before the word no to pull all records in the Place of Manufacture column that are NOT Malaysia

Alternately, you can us != instead of ==. The exclaimation point means “Not”

**Pandas Recap Solution:**



Do not use a space between the single quotes above. Just have two single quotes next to each other.

* This will make an empty string, rather than a space as the contents of the cell

In [15]: changes the format of the data from days, hours, and minutes to seconds.

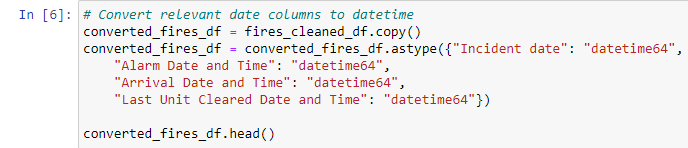
^ This is also shown in GroupBy solution, IN [7]

**GroupBy solution:**

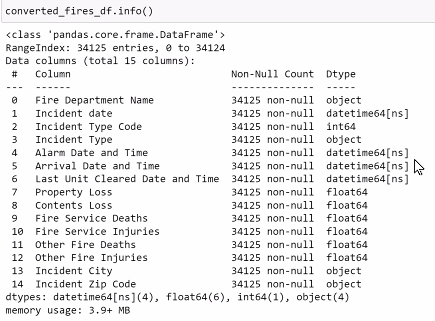
Is effectively implementing a pivot table.

In [6]: if you don’t want original data frame to be changed, make a copy

* Don’t make too many copies of all the rows. It’s draining on the resources for the computer.
  + You can make a copy of a subsection of rows. Like 1,000 rows instead of 75k rows.



Converted\_fires\_df.info() shows the data type for each column



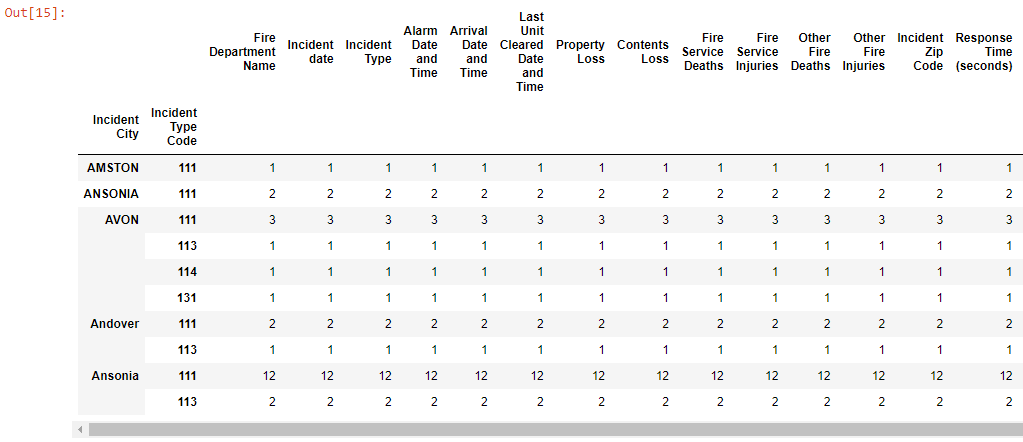
REVIEW GroupBy IN [11] (8:31pm in video recording)

IN [16] (8:38pm in video recording)

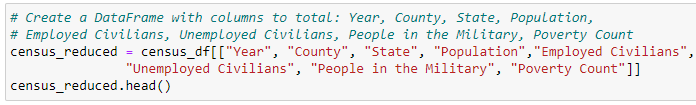
IN [17] (8:36pm in video recording)

If you GroupBy multiple columns, it may remove any lines that aren’t the first one for a column.

In example below, it would remove out the rows for Avon that aren’t Incident Type Code 111



**Census GroupBy solution:**

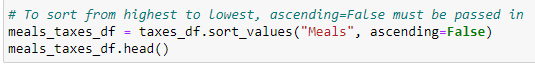


Makes a new Data Frame based on the columns in census\_df.

If you omit the census\_df after the equals sign, it will not throw an error, but it will not be able to display the .head()

**Sorting data solution:**

Groupby uses square brackets [].  
Sort can have square brackets [] within parentheses (), but the parentheses () are required.



^ If sorting by a column, you can add an order within the parentheses

^ ascending=False basically means Descending order

Each Pandas action has a set of arguments in a certain order. Some are optional, some required.

Pandas.pydata.org is the official pandas website, and it has pages with the arguments for each action.

9:15pm in video recording

