pandas

Sunday, November 15, 2020 7:56 PM

Pandas Dataframe is an alternative to excel spreadsheets

Many useful methods including:
.describe() which helps describe the data in the dataframe generally with things like the count, mean, std, min, 25%, 50%, 75%, and max
.drop() throws out data which we don't need from our dataframe
.value_counts() allows us count groups quickly and this can be combined with .plot.bar() to make quick bar plots of counts
.groupby() lets you group by different categories and then show the mean, max etc

- huge insight into data with very little coding

.loc vs iloc and using logic within loc such as 'latitude' >42
sort_values() with ascending descending on multiple columns
operations to combine columns together in some way
rearranging order of columns using

- cols = list(df.columns.values)
- df = df[cols[0:4]+[cols[-1]]+cols[4:12]] #this would bring final column into 5th position

.reset_index() if you saved a new df with smaller subset of data you will still have the old indices, so you can reset them. Note, this saves old index, but you can drop that column by passing drop=True and inplace= True as arguments
.contains you can modify individual cells in a column based on whether they contain a certain text. For example, df.loc[df['comments'].str.contains('UFO', na=False),'comments'] = 'XXXX redacted XXX' If you want the opposite, you don't use ! like you'd think, you use ~
You can modify based on logic. For example, df3.loc[df['Utah']==True,'country']='us'
chunksize = # if you want to load smaller chunks at a time so you don't run into memory issues.