Cheatography

Data Analysis Cheat Sheet

by wilecoyote via cheatography.com/178221/cs/37165/

Fixing dates

Date string	Timestamp to string	string to Timestamp
'Feb-2023'	pd.Timestamp('2023-02-25').strftime('%b-%Y')	pd.to_datetime('Feb-2023', format='%b-%Y')
'02-2023'	pd.Timestamp('2023-02-25').strftime('%m-%Y')	pd.to_datetime('Feb-2023', format='%m-%Y')

df['ts'] = pd.to_datetime(df[['Year', 'Month']].apply(lambda x: '{} {} {}'.format(x[1], 15, int(x[0])), axis=1)

df['period'] = df['ts'].apply(lambda x: x.to_period('M'))

Importing data

df = pd.read_csv('path/filename.csv')

df = pd.read_csv('https://example.com/page')

df = pd.read_excel('path/filename.xlxs',sheet='sheet1')

CSV options: index_col='A', header=2, parse_dates=['D1', 'D2'], thousands=","

Excel options:

Cleaning data

df.dropna(inplace=True)

df.drop(inplace=True, columns=["A","B","C"]

New columns with apply

```
s1 = s.apply(function)
df['B'] = df.app ly( fun ction, args=())
df['B'] = df['A' ].a ppl y(f unc tion, axis=0|1, args=())
```

axis=0 is index, applies function to each column (e.g. sum down columns)

axis=1 is columns, applies function to each row (e.g. sum across rows)

each row or column in DataFrame or value in Series is passed to function

args are passed as additional positional parameters to function

Selecting data

Return series	df['A']
Return dataset	df[['A']]
Return series of booleans	df['A'] < 20
Return filtered DataFrame	df[df['A']<20]
Return filtered Series	df[df['A']<20,'Column']
Return filtered DataFrame	df[df['A']<20,['Column']]

Combining data sets

ds1.merge(ds2, on='field', how='inner')

pd.concat([ds1,ds2,ds3], join='inner')

pd.merge_ordered(df1,df2)

pd.merge_asof(df1,df2,on='field', direction='nearest')

how/join='inner','outer','left','right'

left_on, right_on; suffixes=('_left','_right')

fill_method='ffill'

When forward filling data on multi-index, generally put date last to use the correct fill.

Semi-joins

```
ds1.merge(ds2, on='id', how='inner')
```