Cheatography

Data Wrangling (Python) Cheat Sheet

by mitcht via cheatography.com/50563/cs/13934/

Read / Write .csv

```
# csv
df = pd.read_csv('file.csv', nrows = 5)
pd.to_csv('file.csv')
# excel
df = pd.read_excel('file.xlsx')
df = pd.read_excel(pd.ExcelFile('file.xlsx'),
'Sheet1')
pd.to_excel('file.xslx', sheet_name='Sheet1')
```

Meta Data

df.info()	df.columns.values
df.shape()	df.index.values
len(df)	len(df.columns)

Arrange

```
df.rename(columns = {'col1': 'rename1',
                     'col2': 'rename2'})
df[['col1', 'col2', ... ]] # order cols
df.sort_values([col1', 'col2'],
               ascending = [True, False])
```

Filter

```
df[(df.col1 > 1000) | (df.col2 != "A")]
df[collist] # collist = ['col1', ...]
df.iloc[0:5, :] / df.head(5) # by position
df.loc[(df.col1 > 5) & (df.col2 == "A"),
       ['col1', 'col2']] # by label
df.drop(['col1', ...], axis = 1)
df.drop_duplicates()
df.sample(frac=0.5 / n=10)
df.filter(regex = 'regex')
```

Useful Functions

.count() # non-NA	.min() / .max()
.sum()	.describe()
.cumsum()	.mean() / .median()
.quantile([0.25, 0.75])	.var() / .std()
.apply(function)	df['col'].nunique()
df['col'].value_counts()	np.nan
pd.isnull(obj)	pd.notnull(obj)
pd.to_datetime(var) # others similar	.abs()
.clip(lower=-10, upper=10)	df.colname.isin(list)
.corr()	

Write Functions

```
def function_name(var1, var2, ...):
        for i in 1:10
                if line1:
                        do this
                elif line2:
                        do that
                else:
                        do none
        return(result)
```

Applying functions

```
f = lambda x: x*2
df.apply(f)
df.applymap(f) # element wise
```

Summarise

```
df.groupby(by='col').function() # function = sum() ...
df.groupby(by='col').agg(function) # function = sum()
...
df.groupby(by='col').size()
```

Cheatography

Data Wrangling (Python) Cheat Sheet

by mitcht via cheatography.com/50563/cs/13934/

Join

```
pd.merge(df1, df2, how = 'left/right/inner/outer',
on='col')
df[~df['col'].isin(df2['col'])] # anti-join
```

Method Chaining

```
df = (pd.melt(df)
        .rename(columns={
                    'variable' : 'var',
                    'value' : 'val'})
        .query('val >= 200')
       )
```

New variable / column

```
df['Volume'] = df.Length *df.Height* df.Depth
pd.qcut(df.col, n, labels=False) # binning
```

Reshape Data

```
pd.melt(df) # cols into rows
pd.pivot(columns='var', values='val') # rows into cols
pd.concat([df1, df2]) # stack two dfs
```

By **mitcht**
cheatography.com/mitcht/

Not published yet.
Last updated 28th December, 2017.