# Manipulating Pandas DataFrames

- **Set index for easier filtering of dates**
  df.*set_index*('Col name',inplace=True)
  df.*sort_index*()
  df.*reset_index*(inplace=True)
- **Convert column to datetime if not already**
  df['Col name']=*pd.to_datetime*(df['Col name'])
- **Subset dataframe or series**
  series=df*['Col name']*
  df_new=df*[['Col name 1', 'Col name 2']]*
  df_new=df.*loc*['Index name 1':'Index name 2',:]
  (grp['Col name'].nunique()).*loc*['Index name']
  df_new=df.*iloc*[:,0:3]
- **Filter data by multiple conditions**
  Use *>*, *<*, *>=*, *<=*, *==*, *!=*
  filter = df['Col name']>10
  string_filter=df['Col name'].*str.contains*('text')
  df_filtered=df[filter]
  Combine multiple conditions with (filter1) *&* (filter2), (filter1) *|* (filter2)
  df_filtered=df[(df['Col name 1']>=number)|(df['Col name 2']=='value')]
- **Count values in a series**
  df[''Col name'].*value_counts*()
- **Group data and do aggregate calculations**
  grp= df.*groupby*('Col name')
  grp['Col name 1', 'Col name 2'].*count*()
  .*nunique*(),.*sum*(),.*mean*(),.*max*(),.*min*()
  grp['Col name 1', 'Col name 2'].*agg*(['min','max','mean'])
  grp.*apply* and grp.*transform*
- **Join data**
  df3=pd.*merge*(df1,df2,how='outer',on='Col name') #outer/inner/left/right
  any columns with same name get appended with _x (left) and _y (right)
- **Self-join data**
  when- you want to get a value from another row of same table
  df.*join*(df.drop('m_ids',1).set_index('e_ids'),on='m_ids',rsuffix='e_names')
- **Sort data**
  df.*sort_values*(by=['Col name 1', 'Col name 2'],ascending=True, inplace=True)
  *sorted*("series") #sort series
- **Handle missing/incomplete data**
  df[df.*isna*().any(axis=1)] #find missing values
  df.*dropna*(axis=0,subset=['Col name'],inplace=True) #drop rows w/ cols na
  df['Col name'].*fillna*(df['Col name'].mean(),inplace=True)
- **Handle duplicated data**
  df[df['Col name'].*duplicated*(keep=False)==True] #identify duplicates
  df.*drop*([row#,row#],axis=0,inplace=True) #drop specific rows
  df.*drop_duplicates*(subset='Col name',keep='first',inplace=True) #keep first/last
- **Create a unique key**
  df['key name']=df['Col name 1'].*astype(str)*+'-'+df['Col name 2'].*astype(str)*