**Filter using query**

A data frames columns can be queried with a boolean expression. Every frame has the module query() as one of its objects members.

We start by importing pandas, numpy and creating a dataframe:

|  |
| --- |
| **import** pandas **as** pd  **import** numpy **as** np    data = {'name': ['Alice', 'Bob', 'Charles', 'David', 'Eric'],  'year': [2017, 2017, 2017, 2017, 2017],  'salary': [40000, 24000, 31000, 20000, 30000]}    df = pd.DataFrame(data, index = ['Acme', 'Acme', 'Bilbao', 'Bilbao', 'Bilbao'])    **print**(df) |

After creation of the Data Frame, we call the query method with a boolean expression. This expression is based on the column names that we defined as ‘ABCD’. The query method will return a new filtered data frame.

|  |
| --- |
| df\_filtered = df.query('salary>30000')  **print**(df\_filtered) |

Total code of data frame creation and filter using boolean expression:

|  |
| --- |
| **import** pandas **as** pd  **import** numpy **as** np    data = {'name': ['Alice', 'Bob', 'Charles', 'David', 'Eric'],  'year': [2017, 2017, 2017, 2017, 2017],  'salary': [40000, 24000, 31000, 20000, 30000]}    df = pd.DataFrame(data, index = ['Acme', 'Acme', 'Bilbao', 'Bilbao', 'Bilbao'])    **print**(df)  **print**('----------')    df\_filtered = df.query('salary>30000')  **print**(df\_filtered) |

**Filter by indexing, chain methods**  
Instead of queries, we can use in-dices.  
We do that by using an array index with boolean expressions:

|  |
| --- |
| df\_filtered = df[(df.salary >= 30000) & (df.year == 2017)]  **print**(df\_filtered) |

Pandas groupby

|  |
| --- |
| **import** pandas **as** pd  **import** numpy **as** np    df1 = pd.DataFrame( {  "Name" : ["Alice", "Ada", "Mallory", "Mallory", "Billy" , "Mallory"] ,  "City" : ["Sydney", "Sydney", "Paris", "Sydney", "Sydney", "Paris"]} ) |

We can then summarize the data using the groupby method:

|  |
| --- |
| **print** df1.groupby(["City"])[['Name']].count() |
| **import pandas as pd**  **import numpy as np**    **df1 = pd.DataFrame( {**  **"Name" : ["Alice", "Ada", "Mallory", "Mallory", "Billy" , "Mallory"] ,**  **"City" : ["Sydney", "Sydney", "Paris", "Sydney", "Sydney", "Paris"]} )**      **df2 = df1.groupby(["City"])[['Name']].count()**  **print(df2)** |

# Read xls with Pandas

**Read excel with Pandas**  
The code below reads excel data into a Python dataset (the dataset can be saved below).

|  |
| --- |
| **from** pandas **import** DataFrame, read\_csv  **import** matplotlib.pyplot **as** plt  **import** pandas **as** pd    file = r'data/Presidents.xls'  df = pd.read\_excel(file)  **print**(df['Occupation']) |

The dataframe can be used, as shown in the example below:

|  |
| --- |
| **from** pandas **import** DataFrame, read\_csv  **import** matplotlib.pyplot **as** plt  **import** pandas **as** pd    file = r'data/Presidents.xls'  df = pd.read\_excel(file)    *# remove messy data*  df = df[df['Years in office'] != 'n/a']    *# show data*  **print**('Min: ', df['Years in office'].min())  **print**('Max: ', df['Years in office'].max())  **print**('Sum: ', df['Years in office'].sum()) |

# Visualize data with Pandas

<http://qrc.depaul.edu/Excel_Files/Presidents.xls>

**Beautiful Plots with Pandas**  
We can plot data of this large excel file with a few lines of code. We select the column “Occupation” for this demonstration using:

|  |
| --- |
| df['Occupation']. |

Full code:

|  |
| --- |
| **from** pandas **import** DataFrame, read\_csv  **import** matplotlib.pyplot **as** plt  **import** pandas **as** pd    file = r'data/Presidents.xls'  df = pd.read\_excel(file)    *# plot data*  colors = ['yellowgreen', 'gold', 'lightskyblue', 'lightcoral','red','green','blue','orange','white','brown']  df['Occupation'].value\_counts().plot(kind='pie',title='Occupation by President',colors=colors)  plt.show() |

|  |
| --- |
| *#df = df.replace('NA()',0)*  df = df[df['% popular'] != 'NA()'] |

To plot the popularity:

|  |
| --- |
| **from** pandas **import** DataFrame, read\_csv  **import** matplotlib.pyplot **as** plt  **import** pandas **as** pd    file = r'data/Presidents.xls'  df = pd.read\_excel(file)    *# plot data*  *#df = df.replace('NA()',0)*  df = df[df['% popular'] != 'NA()']    **print**( df['% popular'] )  df['% popular'].plot(kind='hist', bins=8, title='Popularity by President', facecolor='blue', alpha=0.5, normed=1)  plt.show() |