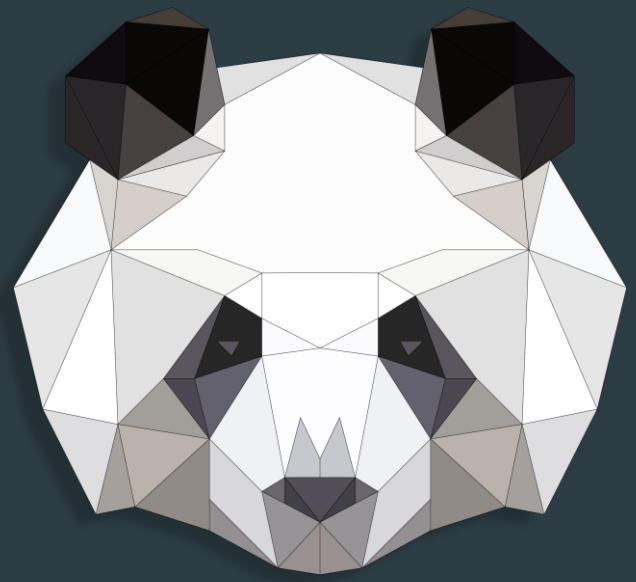
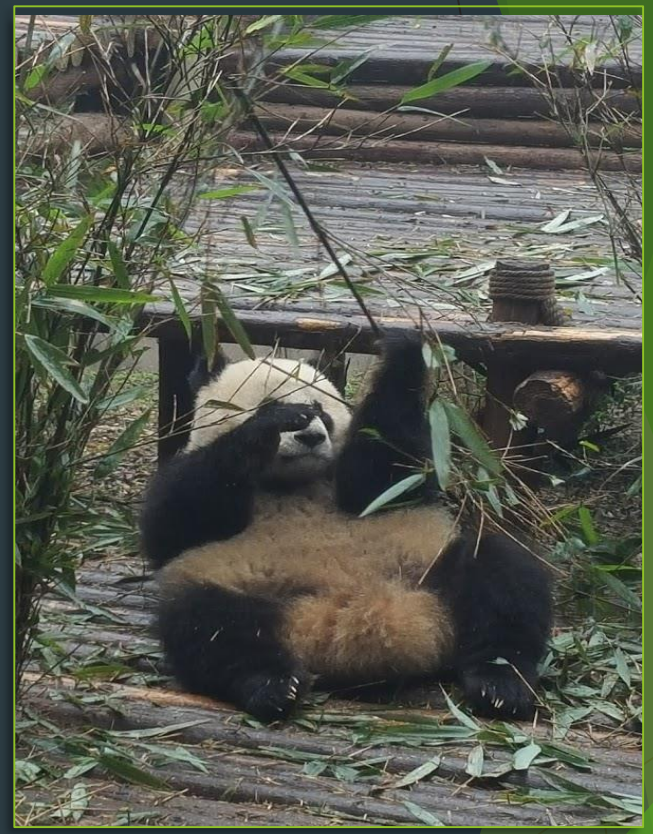
Pandas



An intro to functionality and code

Thomas Bacas





Pandas

- Pandas is a software library used for data analysis and manipulation
- Pandas makes it easy to organize and manipulate data in Python by organizing data into easy to read Data Frames
- Along side its basic functionality Pandas includes a variety of tools to understand your data further

Getting Started

- import pandas as pd
- df = pd.DataFrame({'A':[1,2,3,4,5],
  'B':[1,2,3,4,5],
  'C':[1,2,3,4,5]})

X	y	A	B	C
d	1	1	1	1
	2	2	2	2
	3	3	3	3
e	4	4	4	4
f	5	5	5	5

- Index = [1,2,3,4,5]
- pd.MultiIndex.from_tuples([(‘d’,1),(‘d’,2),(‘d’,3)(‘e’,4),(‘f’,5)],names =[‘x’,’y’])
- pd.read_csv(‘./file/csv_file.csv’)
  - The read function can take a variety of file types: Excel, html, sql, etc..

Summarize Data

- df['column'].value_counts()
  - Count number of unique values
- len(df)
  - Number of rows in DataFrame
- df.nunique()
  - Number of distinct observations
- df.describe()
  - Provides a wide array of statistics across the DataFrame
- df.isnull().sum()
- df.info()
- df.shape

Reshaping your Data

- pd.melt(df) or df.melt()

- df2= pd.DataFrame({'Foo':['one','one','one','two', 'two'],

  'First' ':['a','b','c','a','b'],

  'Second:[1,2,3,4,5],

  'Third':['t','w','x','y','z']})

- df.pivot(index = 'Foo', columns = 'First', values ='Second')

	Foo	First	Second	Third
0	one	a	1	t
1	one	b	2	w
2	one	c	3	x
3	two	a	4	y
4	two	b	5	z

Second	a	b	C
Foo			
one	1	2	3
two	4	5	NaN

Reshaping continued

- df1 =pd.DataFrame({'A':[6,7,8,9,10],
  'B':[6,7,8,9,10],
  'C':[6,7,8,9,10]})
- pd.concat([df,df1])
- pd.concat([df,df2])
- pd.concat([df,df2], axis = 1)

	A	B	C	First	Foo	Second	Third
0	1.0	1.0	1.0	1.0	one	a	t
1	2.0	2.0	2.0	2.0	one	b	w
2	3.0	3.0	3.0	3.0	one	c	x
3	4.0	4.0	4.0	4.0	two	a	y
4	5.0	5.0	5.0	5.0	two	b	z
0	NaN	NaN	NaN	1.0	one	a	t
1	NaN	NaN	NaN	2.0	one	b	w
2	NaN	NaN	NaN	3.0	one	c	x
3	NaN	NaN	NaN	4.0	two	a	y
4	NaN	NaN	NaN	5.0	two	b	z

Merging

- pd.merge(A,B, how = ‘left’, on = ‘x1’)
- pd.merge(A,B, how = ‘right’, on =‘x1’)
- pd.merge(A,B, how = ‘inner’, on =‘x1’)
- pd.merge(A,B, how = ‘outer’, on =‘x1’)

A	
x1	x2
A	1
B	2
C	3

B	
x1	x3
A	4
B	5
D	6

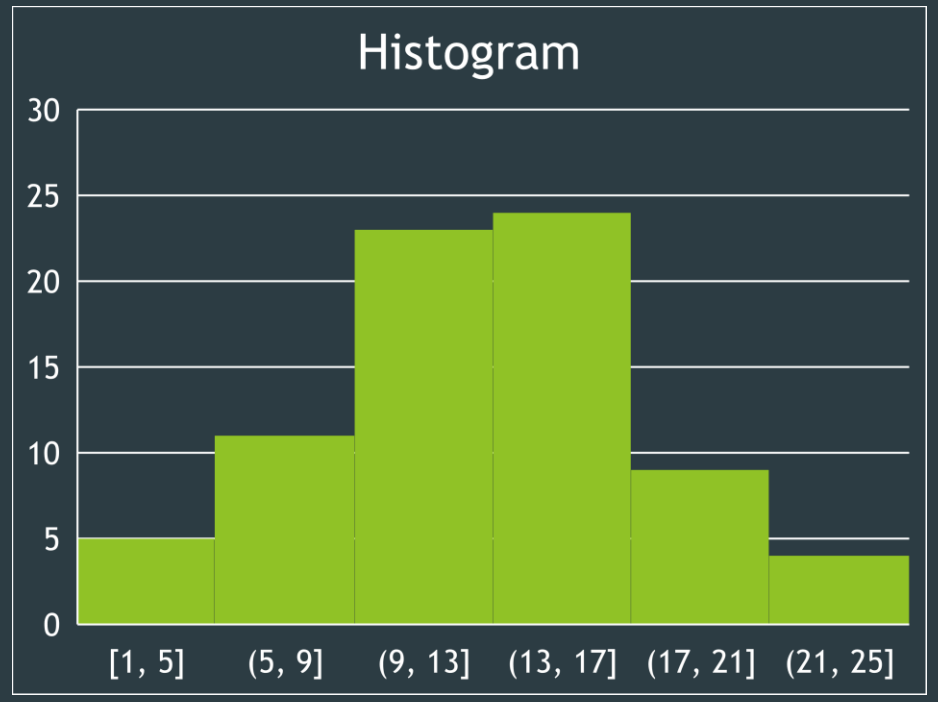
C		
x1	x2	X3
A	1	4
B	2	5
C	3	NaN
D	NaN	6

Sorting

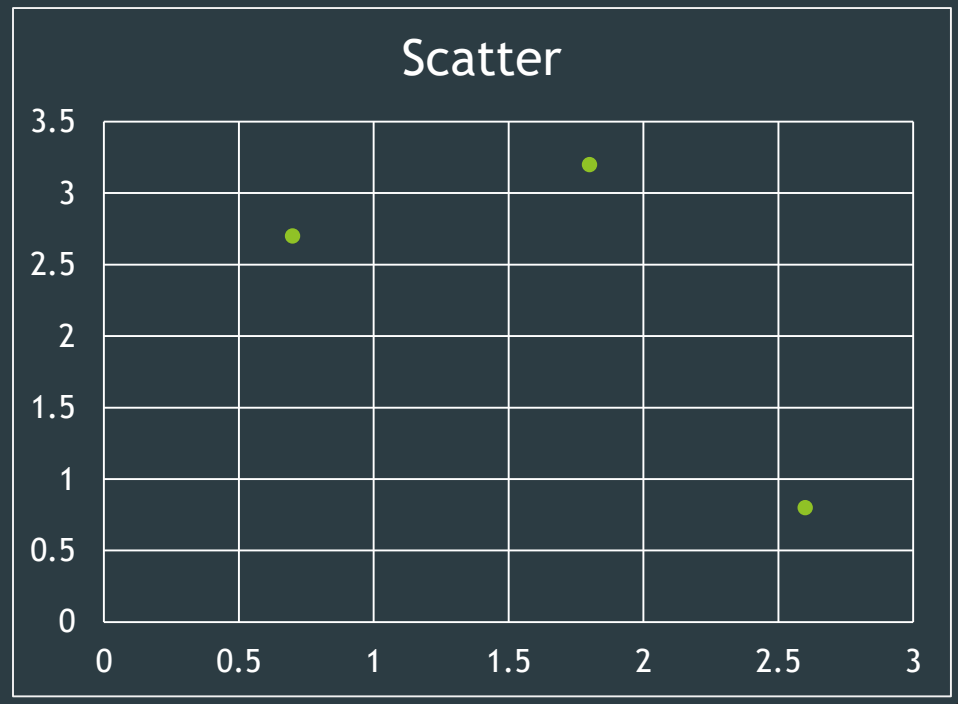
- df.sort_values('column_name')
  - Sort rows by values, low-high
- df.sort_values('column_name', ascending = False)
  - Sort row by values, high – low
- df.rename(columns = {'column_original' : 'column_new'})
  - You can also write this as a variable to simplify your code
  - mask = {'name1':'rename1', 'name2':'rename2', 'name3':'rename3'}
  - df.rename(columns = mask)

Plotting

- df.plot.hist()



- df.plot.scatter(x='w',y='h')



- https://towardsdatascience.com/a-quick-introduction-to-the-pandas-python-library-f1b678f34673
- https://pandas.pydata.org/
- https://github.com/pandas-dev/pandas/blob/master/doc/cheatsheet/Pandas_Cheat_Sheet.pdf