Python for Data Analysis: Pandas

Andreas Weller, PhD

WTCHG - NHS

23.4.2014

Pandas

DataFrame

- New object introduced by Pandas
- Inspired by R data.frame

DF creation

- pd.read_csv(filename, sep=",")
- pd.DataFrame(data)

Basic methods I

DF summary

Method	Effect
df.describe()	summary stats
df.head()	head
df.tail(10)	select rows+cols by index

Selection

df[["A","B"]]	select cols
df.loc[:5, "A"]	select rows+cols by name
df.iloc[:5, :5]	select rows+cols by index

Basic methods II

Boolean indexing

Method	Effect
df.value >12	Query all rows
df[df.value >12]	Select True rows
df[cond1 & cond2]	Select rows
df.shape	Count rows and cols

Column creation

df["new"] = 1	New col, same values
df["new"] = df.old * 2	New col from old col
df["new"] = df.old.str.upper()	New col from string method

Groupby

df.groupby(df.value2).func()

df.value1.groupby([df.value2, df.value3]).func()

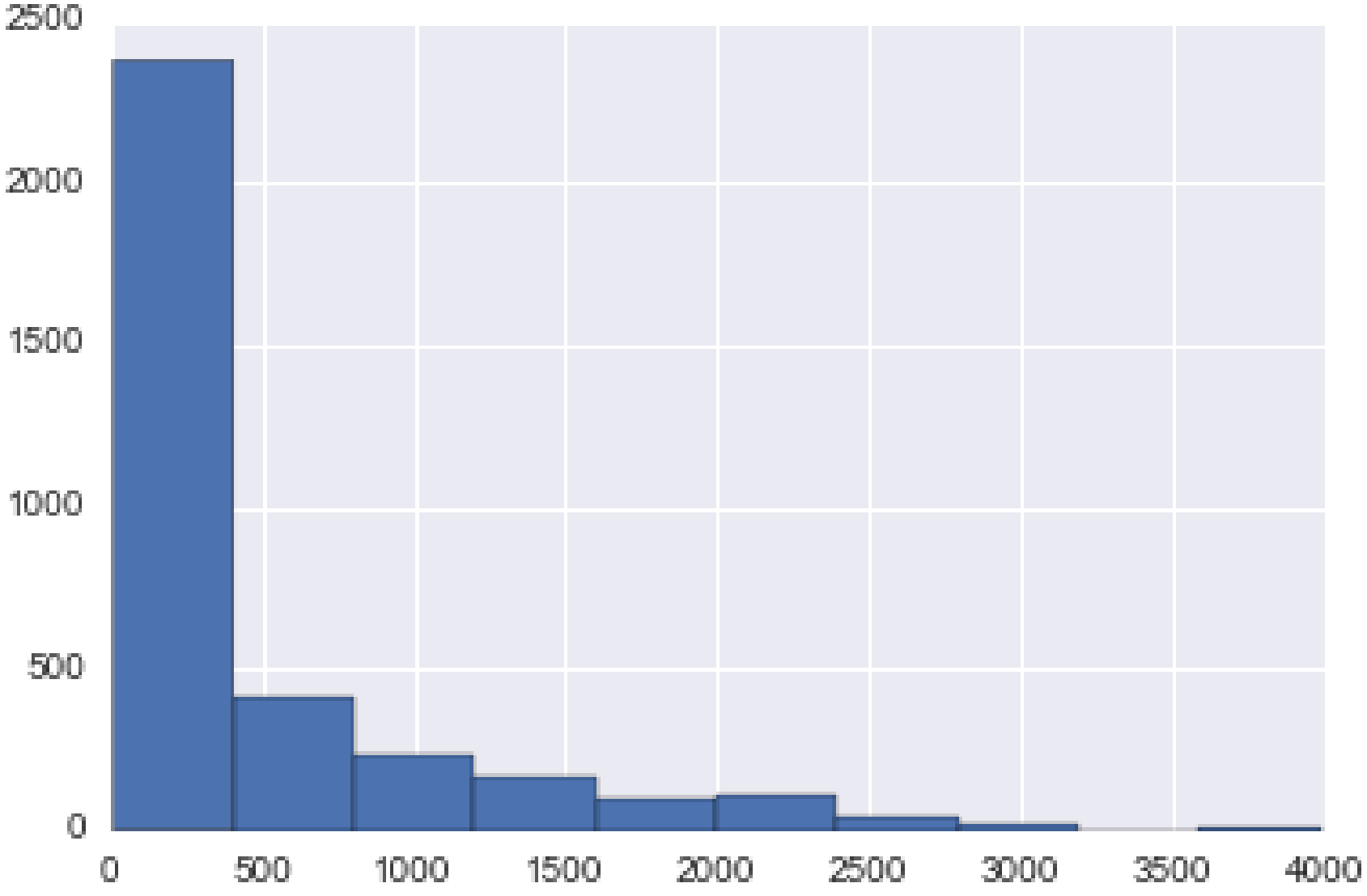
Groupby functions

Method	Effect
.sum(), .mean(), median()	Sum, Mean, Median
.count()	No. of entries
.nunique()	No. of unique entries

Histograms

```
# plot a histogram from a pandas DataFrame

data = df.DP

plt.hist(data, bins=100)
```

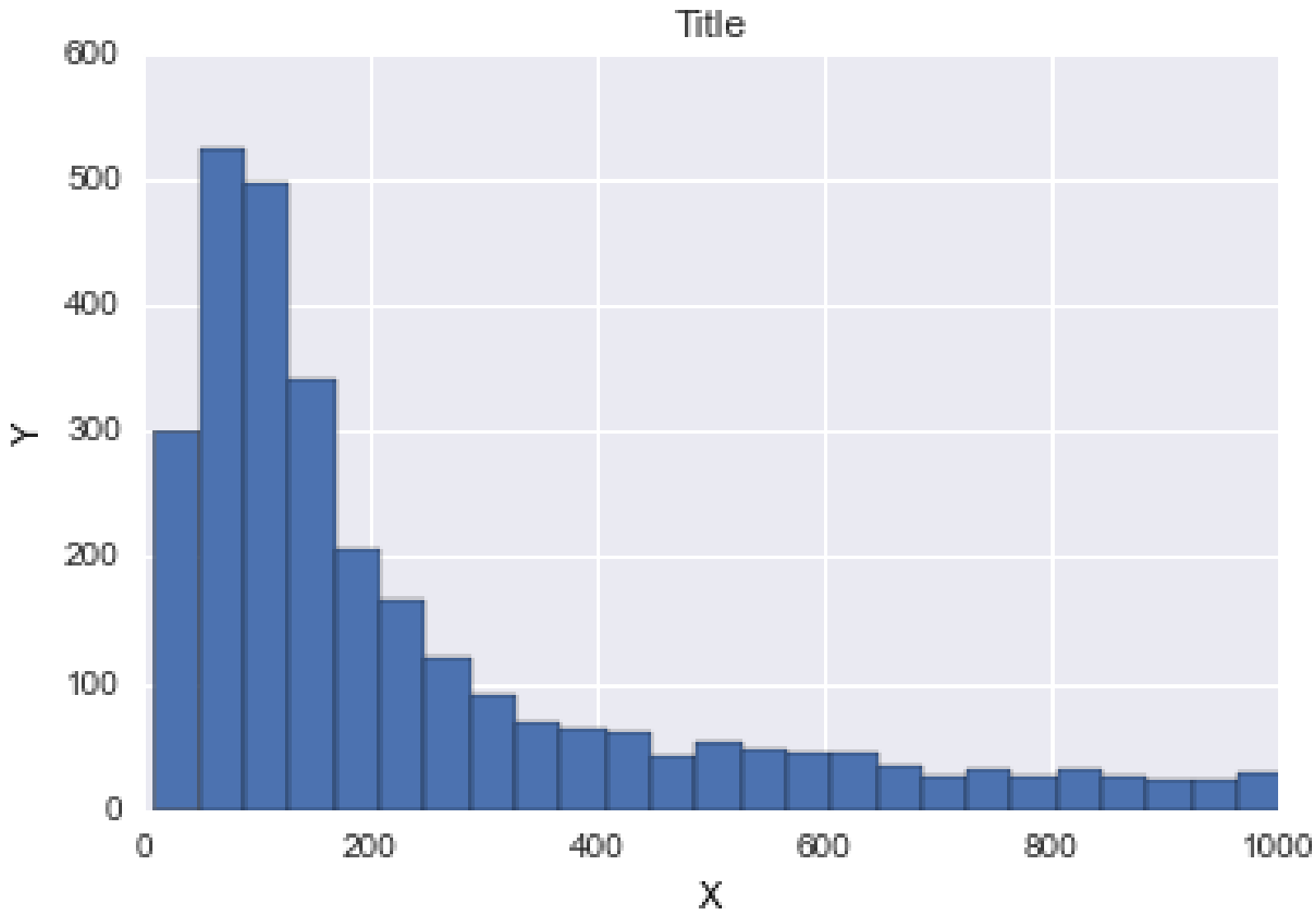
Histograms


2500
2000
1500
1000
500
0
0
500
1000
1500
2000
2500
3000
3500
4000

Histograms

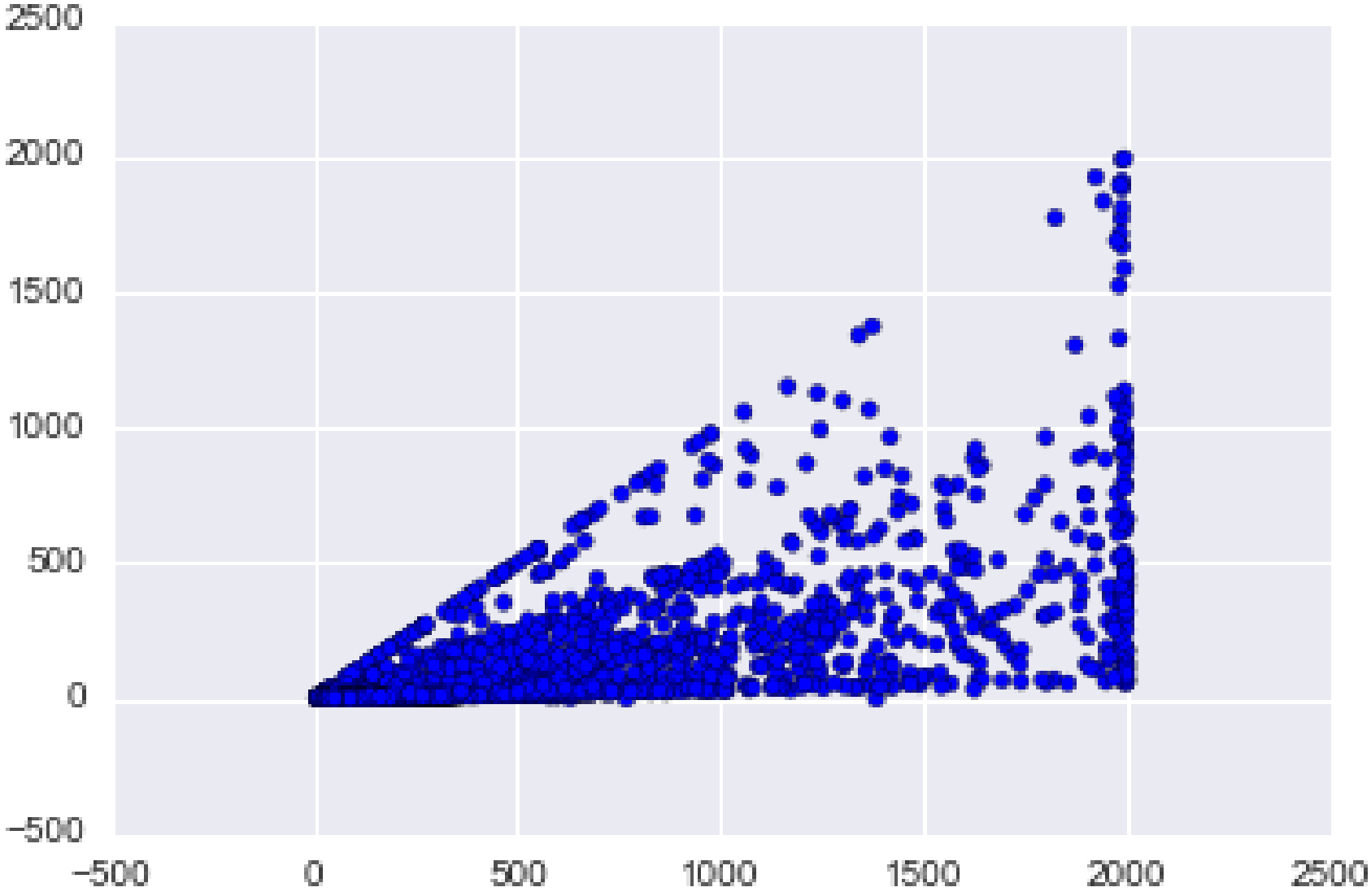
```
# plot a histogram from a pandas DataFrame

data = df.DP

plt.hist(data, bins=100)
plt.xlabel("X")
plt.ylabel("Y")
plt.xlim(0,1000)
plt.title("Title")
```

Histograms



Scatterplots

plt.scatter(df.FDP, df.FAO)



Scatterplots

sns.jointplot("FDP", "FAO", df)