An End to Boring Data

With Visualizations in Python

HEATHER SHAPIRO
TECHNICAL EVANGELIST, MICROSOFT
@microheather

What will we cover?

- Why data visualizations are important
- Case study on NYC Restaurant Ratings
- What libraries in python work best for different types of graphs

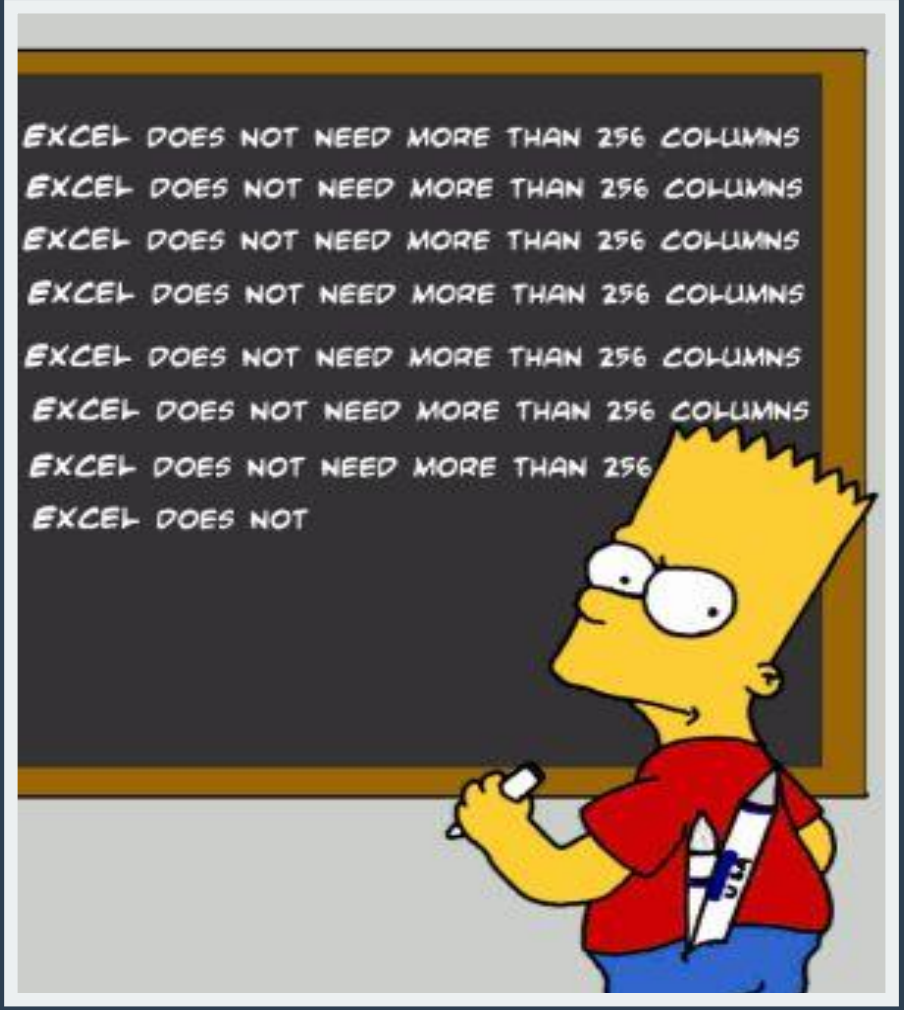
Why we need data viz

HEATHER SHAPIRO | TECHNICAL EVANGELIST, MICROSOFT
@microheather

HARD TO UNDERSTAND

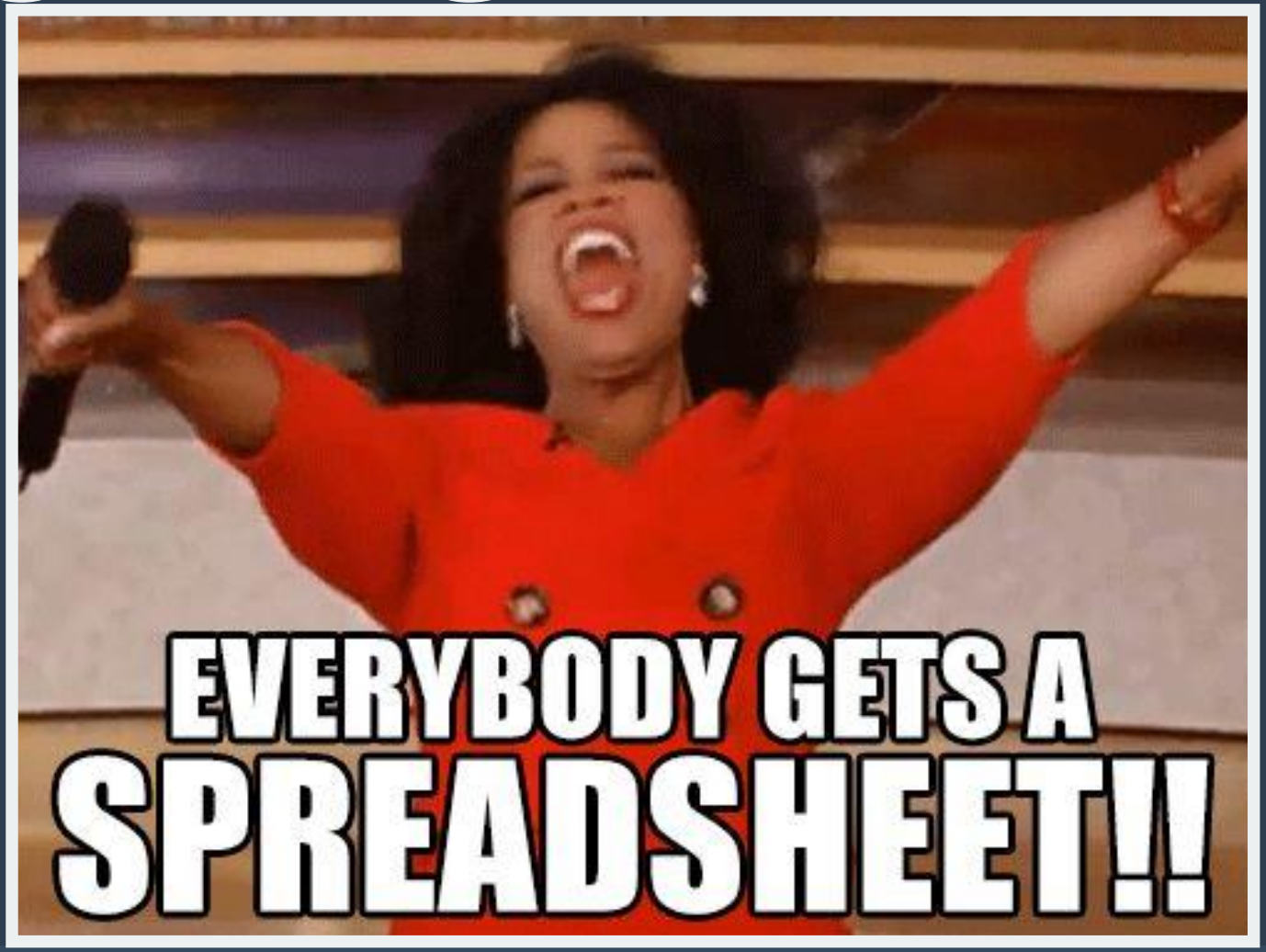
HEATHER SHAPIRO | TECHNICAL EVANGELIST, MICROSOFT
@microheather

TOO MANY NUMBERS



HEATHER SHAPIRO | TECHNICAL EVANGELIST, MICROSOFT
@microheather

BORING MEETINGS

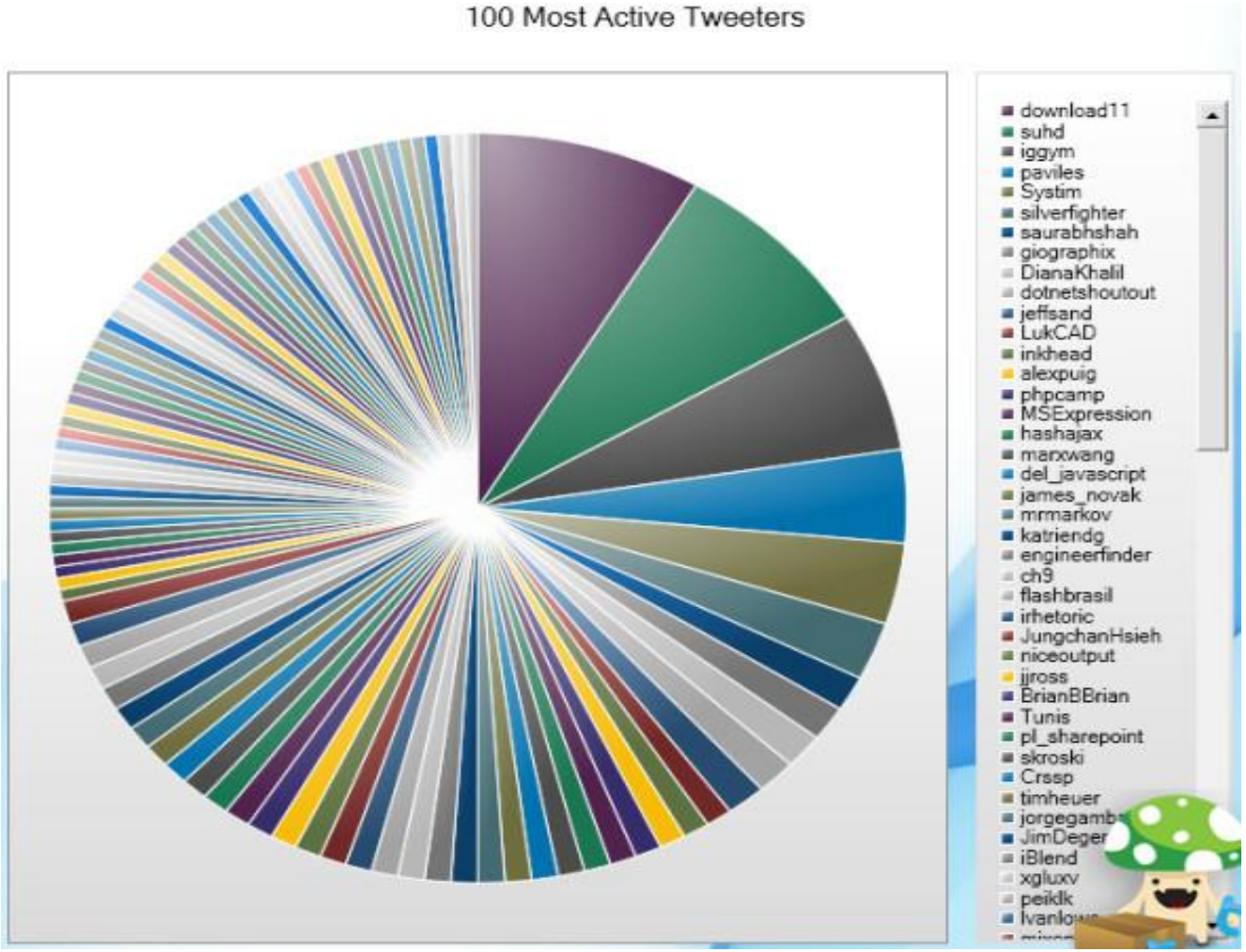


HEATHER SHAPIRO | TECHNICAL EVANGELIST, MICROSOFT
@microheather

Visualizations gone wrong

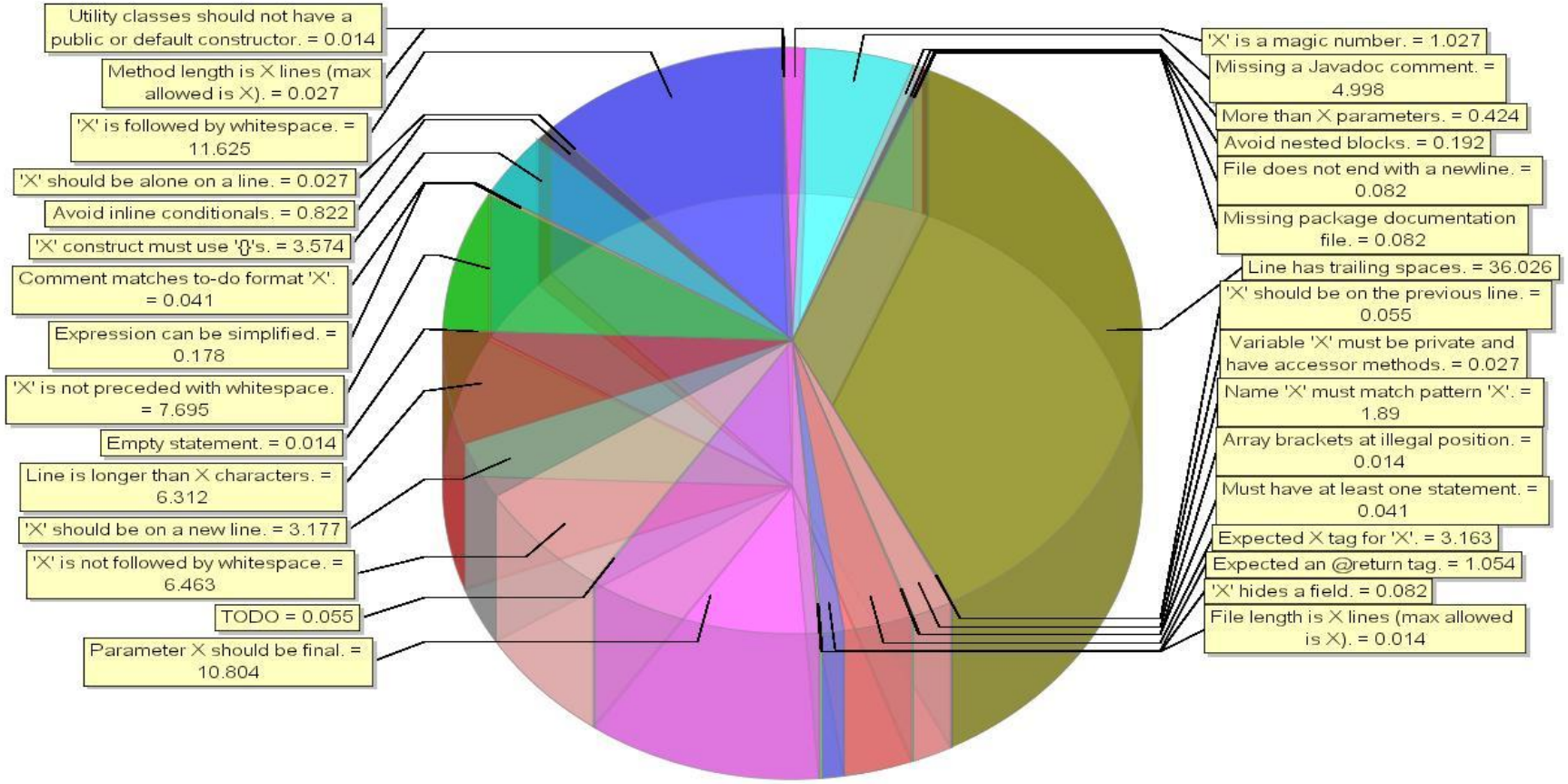
HEATHER SHAPIRO | TECHNICAL EVANGELIST, MICROSOFT
@microheather

TOO MANY VARIABLES



HEATHER SHAPIRO | TECHNICAL EVANGELIST, MICROSOFT
@microheather

...?/


Utility classes should not have a public or default constructor. = 0.014
Method length is X lines (max allowed is X). = 0.027
'X' is followed by whitespace. = 11.625
'X' should be alone on a line. = 0.027
Avoid inline conditionals. = 0.822
'X' construct must use '{}'s. = 3.574
Comment matches to-do format 'X'. = 0.041
Expression can be simplified. = 0.178
'X' is not preceded with whitespace. = 7.695
Empty statement. = 0.014
Line is longer than X characters. = 6.312
'X' should be on a new line. = 3.177
'X' is not followed by whitespace. = 6.463
TODO = 0.055
Parameter X should be final. = 10.804
'X' is a magic number. = 1.027
Missing a Javadoc comment. = 4.998
More than X parameters. = 0.424
Avoid nested blocks. = 0.192
File does not end with a newline. = 0.082
Missing package documentation file. = 0.082
Line has trailing spaces. = 36.026
'X' should be on the previous line. = 0.055
Variable 'X' must be private and have accessor methods. = 0.027
Name 'X' must match pattern 'X'. = 1.89
Array brackets at illegal position. = 0.014
Must have at least one statement. = 0.041
Expected X tag for 'X'. = 3.163
Expected an @return tag. = 1.054
'X' hides a field. = 0.082
File length is X lines (max allowed is X). = 0.014

What data viz provides

- Helps the visual learner
- Makes sense of tremendous amounts of data
- Helps walk through a problem
- Tells a story in seconds

HEATHER SHAPIRO | TECHNICAL EVANGELIST, MICROSOFT
@microheather

CASE STUDY

NYC Restaurant Ratings

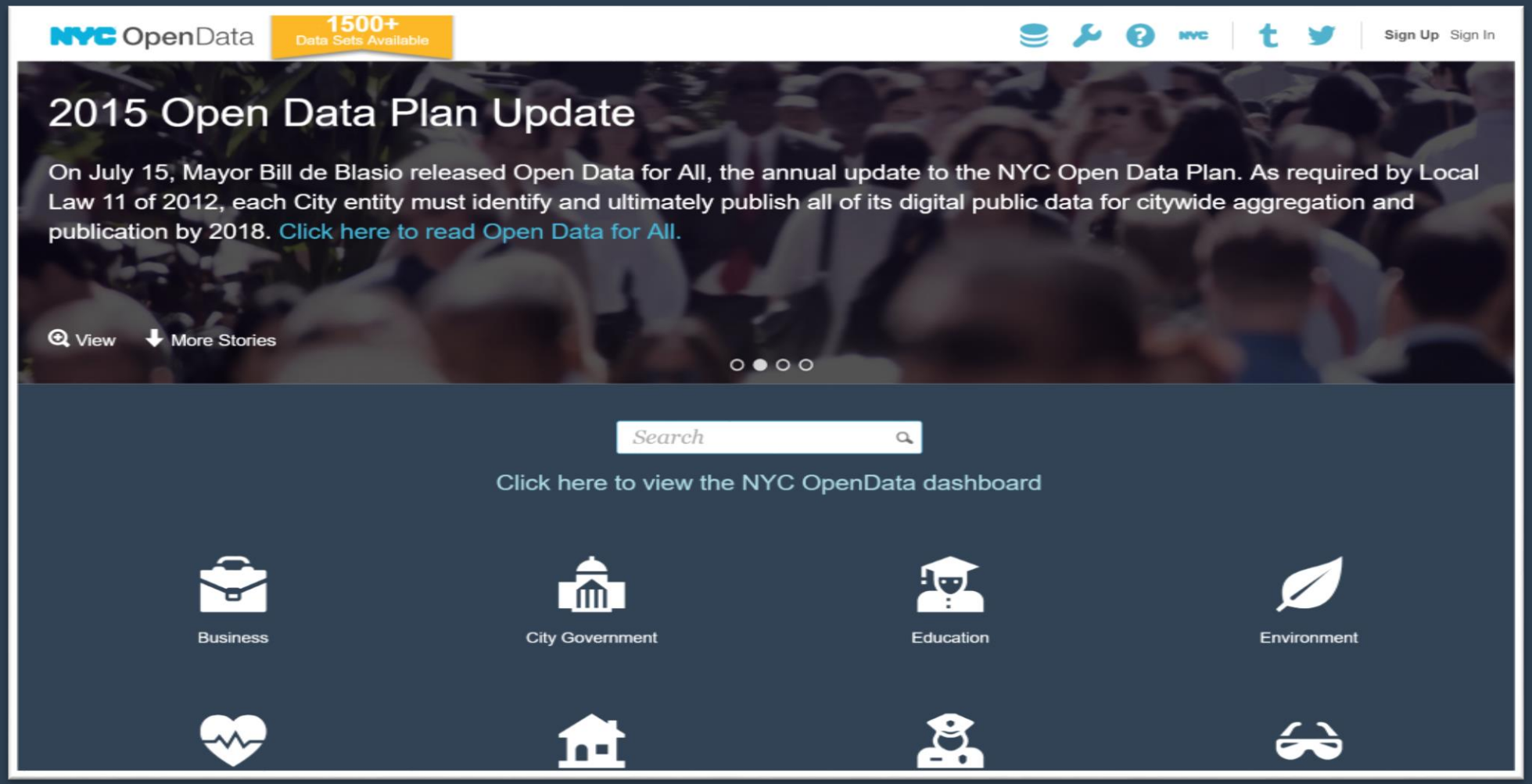
HEATHER SHAPIRO | TECHNICAL EVANGELIST, MICROSOFT
@microheather

HEATHER SHAPIRO | TECHNICAL EVANGELIST, MICROSOFT
@microheather

HEATHER SHAPIRO | TECHNICAL EVANGELIST, MICROSOFT
@microheather

HEATHER SHAPIRO | TECHNICAL EVANGELIST, MICROSOFT
@microheather

NYC OPEN DATA



HEATHER SHAPIRO | TECHNICAL EVANGELIST, MICROSOFT
@microheather

Dataset (File)	Data Field Name	Data Type	Length	Expected Values	Description
WEBEXTRACT	CAMIS	Varchar	10		This is a unique identifier for the entity (restaurant)
WEBEXTRACT	DBA	varchar	255		This field represents the name (doing business as) of the entity (restaurant)
WEBEXTRACT	BORO	Varchar	1	• 1 = MANHATTAN • 2 = BRONX • 3 = BROOKLYN • 4 = QUEENS • 5 = STATEN ISLAND • Missing	Borough in which the entity (restaurant) is located. NOTE: There may be discrepancies between zip code and listed boro due to differences in an establishment's mailing address and physical location
WEBEXTRACT	BUILDING	Varchar	10		This field represents the building number for the entity (restaurant)
WEBEXTRACT	STREET	Varchar	100		This field represents the street name at which the entity (restaurant) is located.
WEBEXTRACT	ZIPCODE	Varchar	5		Zip code as per the address of the entity (restaurant)
WEBEXTRACT	PHONE	Varchar	20		Phone number
WEBEXTRACT	CUISINE DESCRIPTION	Varchar	200		This field describes the entity (restaurant) cuisine.
WEBEXTRACT	INSPECTION DATE	Datetime	N/A		This field represents the date of inspection. NOTE: Inspection dates of 1/1/1900 mean an establishment has not yet had an inspection
WEBEXTRACT	ACTION	Varchar	150	• Violations were cited in the following area(s). • No violations were recorded at the time of this inspection. • Establishment re-opened by DOHMH • Establishment re-closed by DOHMH • Establishment Closed by DOHMH. Violations were cited in the following area(s) and those requiring immediate action were addressed. • "Missing" = not yet inspected	This field represents the action that is associated with each restaurant inspection.
WEBEXTRACT	VIOLATION CODE	Varchar	3		This field represents each violation associated with a restaurant inspection.
WEBEXTRACT	VIOLATION DESCRIPTION	Varchar	600		This field describes the violation codes
WEBEXTRACT	CRITICAL FLAG	Varchar	1	• Critical • Not Critical • Not Applicable	Critical violations are those most likely to contribute to foodborne illness.
WEBEXTRACT	SCORE	Varchar	3		Total score for a particular inspection; updated based on adjudication results.
WEBEXTRACT	GRADE	Varchar	1	• Not Yet Graded • A = Grade A • B = Grade B • C = Grade C • Z = Grade Pending • P=Grade Pending issued on re-opening following an initial inspection that resulted in a closure	This field represents the grade associated with this inspection. Grades given during a reopening inspection derived from the previous re-inspection.
WEBEXTRACT	GRADE DATE	Datetime	N/A		The date when the grade was issued to the entity (restaurant)
WEBEXTRACT	RECORD DATE	Datetime	N/A		The date when the webextract was run to produce this data set
				•Calorie Posting/ Compliance Inspection	

HEATHER SHAPIRO | TECHNICAL EVANGELIST, MICROSOFT
@microheather

Steps Taken

- Load modules
- Load Restaurant Rating Data
- Understand the data
- Visualize ☺

TOOLS USED

- Pandas
- Matplotlib
- Basemap
- Folium
- Seaborn
- Bokeh
- Plot.ly

GETTING STARTED

HEATHER SHAPIRO | TECHNICAL EVANGELIST, MICROSOFT
@microheather

PANDAS

HEATHER SHAPIRO | TECHNICAL EVANGELIST, MICROSOFT
@microheather

PANDAS

Import pandas as pd

- Python library to provide data analysis features
- Built on *NumPy*, *SciPy*, and *matplotlib*
- Key components
  - Series
  - DataFrames

LIBRARIES FOR STATISTICAL GRAPHS

HEATHER SHAPIRO | TECHNICAL EVANGELIST, MICROSOFT
@microheather

matplotlib

HEATHER SHAPIRO | TECHNICAL EVANGELIST, MICROSOFT
@microheather

MATPLOTLIB

import matplotlib.pyplot as plt

- MATLAB-like plotting framework

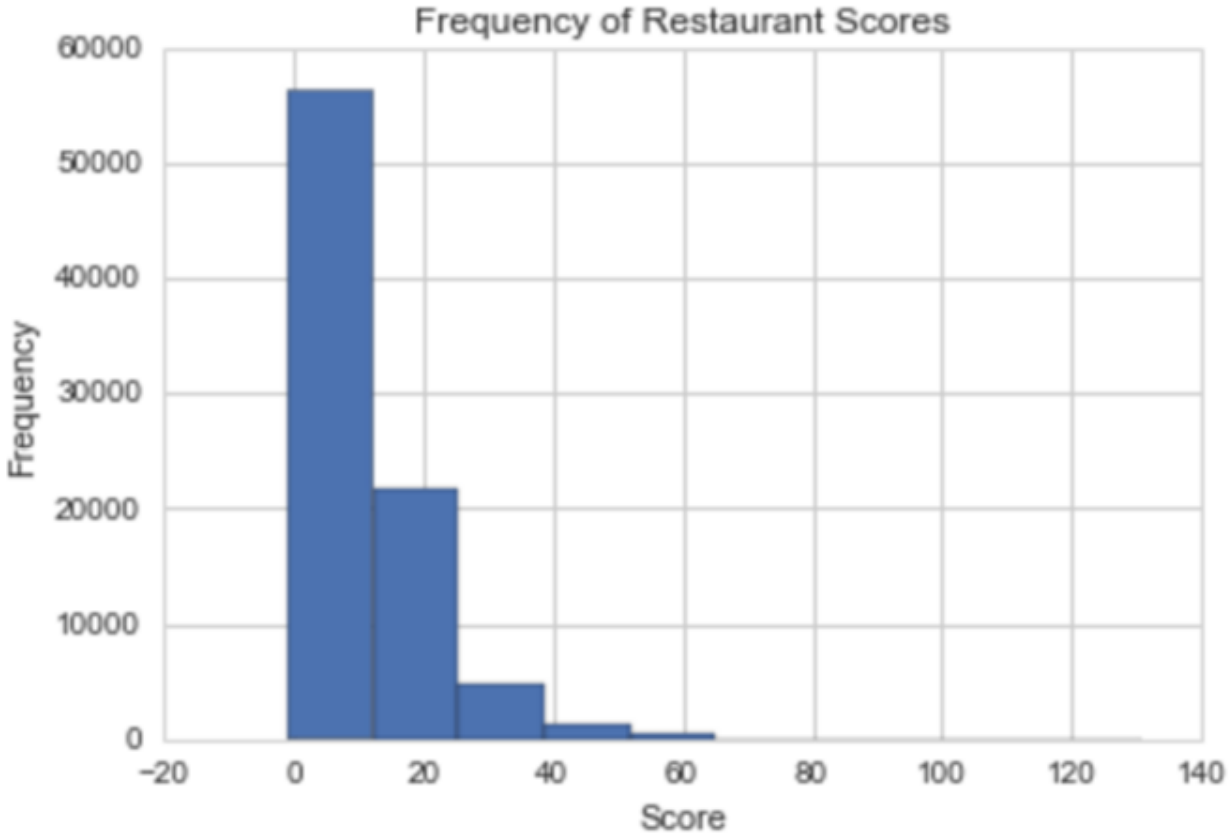
MATPLOTLIB

```
f, ax = plt.subplots() ## creates figure area with axes
# histogram our data with numpy
data = mRests['SCORE']

plt.hist(data)
plt.xlabel('Score')
plt.ylabel('Frequency')
plt.title("Frequency of Restaurant Scores")

plt.show()
```

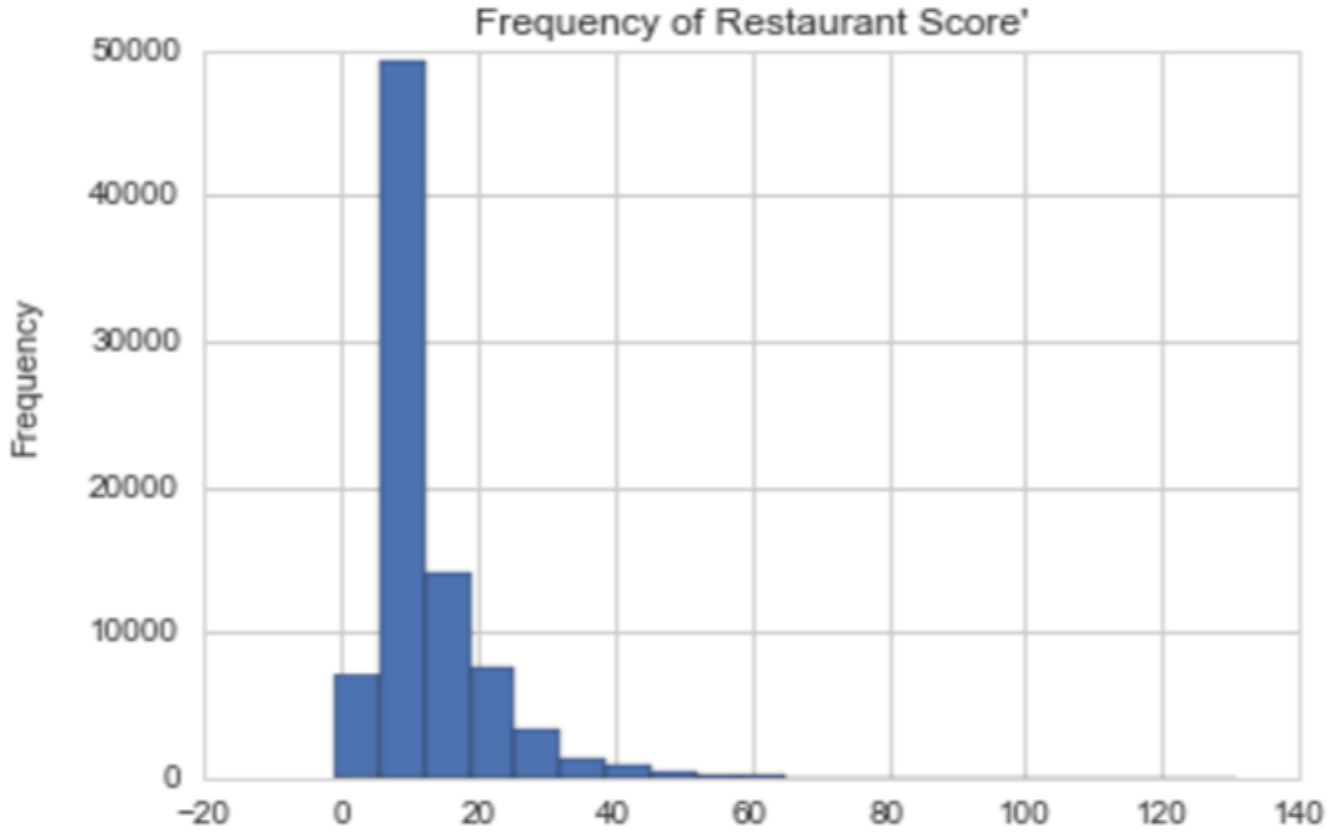
MATPLOTLIB


Frequency of Restaurant Scores
Frequency
60000
50000
40000
30000
20000
10000
0
−20
0
20
40
60
80
100
120
140
Score

PANDAS SHORTHAND

```
mRests["SCORE"].hist(bins=20)
plt.title("Frequency of Restaurant Score"')
```

PANDAS SHORTAHND



HEATHER SHAPIRO | TECHNICAL EVANGELIST, MICROSOFT
@microheather

seaborn

HEATHER SHAPIRO | TECHNICAL EVANGELIST, MICROSOFT
@microheather

SEABORN

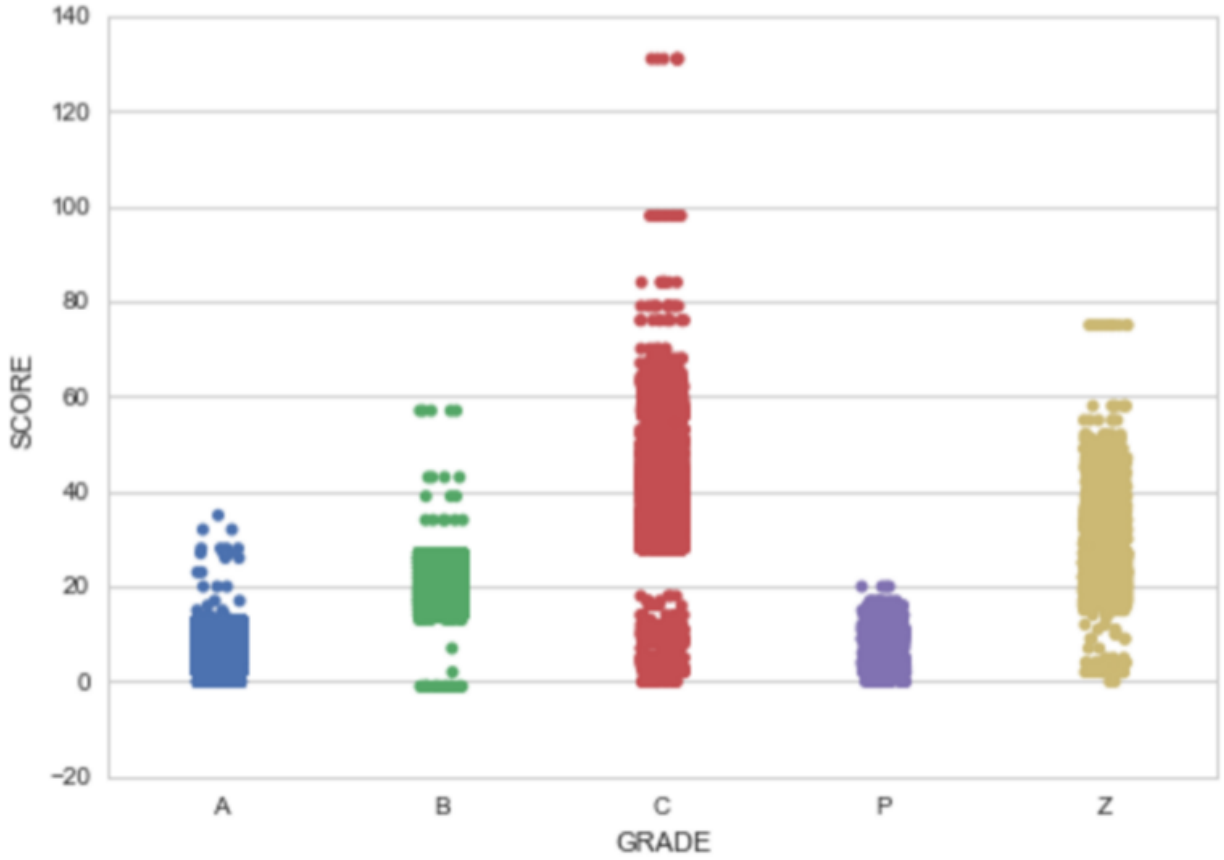
Import seaborn

- Built on top of matplotlib
- Creates more sophisticated graphs
- Look more professional

SEABORN

sns.stripplot(x="GRADE", y = "SCORE", data = mRests, jitter = True)

SEABORN



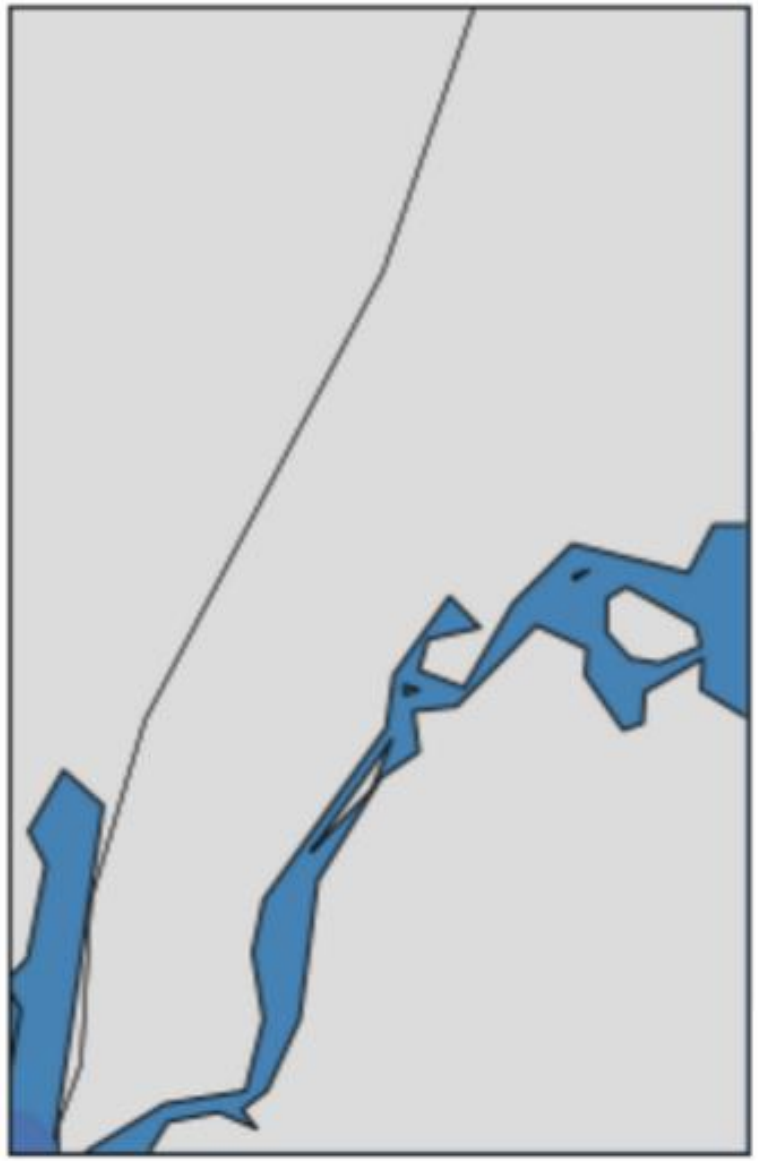
LIBRARIES FOR MAPPING

HEATHER SHAPIRO | TECHNICAL EVANGELIST, MICROSOFT
@microheather

BASEMAP

- Hard to install. There are a lot of prereqs and the documentation isn't there for windows

BASEMAP



HEATHER SHAPIRO | TECHNICAL EVANGELIST, MICROSOFT
@microheather

FOLIUM

Import folium

- Visualize data on a Leaflet map
- Built-in tilesets from:
  - OpenStreetMap, MapQuest Open, MapQuest Open Aerial, Mapbox, and Stamen, and supports custom tilesets with Mapbox or Cloudmade API keys.

LIBRARIES FOR INTERACTIVE PLOTS

HEATHER SHAPIRO | TECHNICAL EVANGELIST, MICROSOFT
@microheather

bokeh

plotly

HEATHER SHAPIRO | TECHNICAL EVANGELIST, MICROSOFT
@microheather

GO PUT THEM IN YOUR BLOGS ☺

HEATHER SHAPIRO | TECHNICAL EVANGELIST, MICROSOFT
@microheather

CLOSING THOUGHTS

- **Pandas** → handy for simple plots but you need to be willing to learn matplotlib to customize.

- **Seaborn** → supports more complex visualization approaches but still requires matplotlib. The color schemes are a nice bonus.

CLOSING THOUGHTS

- **Basemap→** Hard to install. Not very robust and there is not higt granularity for the maps.

- **Folium→** Great documentation for mapping. Wish you could add more interactive widgets.

CLOSING THOUGHTS

- **bokeh** → Overkill for simple scenarios and documentation was not great.

- **Plotly** → most interactive graphs. You can save them offline and create rich web-based visualizations for your blog. Not good with city level data for maps.

What did we cover?

- Introduction to data visualizations in python
- How to walk through a data problem
- Which libraries are useful and for what
- Great way to update that blog

RESOURCES

- My blog ☺ www.Microheather.com
- My github: www.github.com/heatherbshapiro
- NYC Open Data: nycopendata.socrata.com
- Data Sets: data.gov
- Data Science VM in Azure : aka.ms/datasciencevm
- Azure Machine Learning: studio.azureml.net
- Channel9 and MVA

CONTACT ME

- Email: hshapiro@Microsoft.com
- Twitter @microheather

THANKS!

HEATHER SHAPIRO | TECHNICAL EVANGELIST, MICROSOFT
@microheather