

CIS192 Python Programming

Data Visualization

Harry Smith

University of Pennsylvania

April 13, 2016

Outline

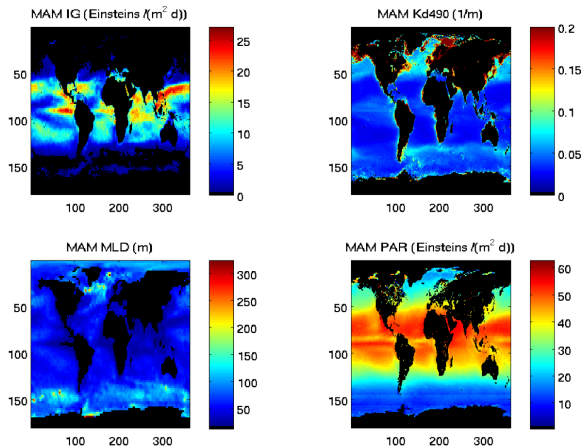
- 1 Introduction and Motivation
- 2 Getting Started
- 3 3D
- 4 Flight Work and Other Interesting Concepts

Motivation

- The first step in using data is understanding it.
- Numbers are complicated and ugly. Colors are pretty.
- Properly visualized data is effective communication on its own.
 - ▶ A scientific paper with well-crafted figures is much more effective than one with dreaded "Data Appendices"

Consider something like this...

Figure: Lookin' good.



...over its original form.

Figure: YIKES

Stn	Netep	Today	QN1	QN2	QN3	QN4	QN5	QN6	QN7	QN8	QN9	QN0	QN10	QN11	QN12	QN13	QN14	QN15	QN16	QN17	QN18	QN19	QN20
QN10		QN11	QN12	QN13	QN14	QN15	QN16	QN17	QN18	QN19	QN20												
QN21		QN22	QN23	QN24	QN25	QN26	QN27	QN28	QN29	QN30	QN31												
QN32		QN33	QN34	QN35	QN36	QN37	QN38	QN39	QN40	QN41	QN42												
QN43		QN44	QN45	QN46	QN47	QN48	QN49	QN50	QN51	QN52	QN53												
QN54		QN55	QN56	QN57	QN58	QN59	QN60	QN61	QN62	QN63	QN64												
QN65		QN66	QN67	QN68	QN69	QN70	QN71	QN72	QN73	QN74	QN75												
QN67		QN77	QN78	QN79	QN80	QN81	QN82	QN83	QN84	QN85	QN86												
7	QN87	QN88	QN89	QN90	QN91	QN92	QN93	QN94	QN95	QN96	QN97												
	QN98	QN99	QN100	QN101	QN102	QN103	QN104	QN105	QN106	QN107	QN108												
108	QN109	QN110	QN111	QN112	QN113	QN114	QN115	QN116	QN117	QN118	QN119												
M119	QN120	QN121	QN122	QN123	QN124	QN125	QN126	QN127	QN128	QN129	QN130												
QN130	QN131	QN132	QN133	QN134	QN135	QN136	QN137	QN138	QN139	QN140	QN141												
QN141	QN142	QN143	QN144	QN145	QN146	QN147	QN148	QN149	QN150	QN151	QN152												
QN152	QN153	QN154	QN155	QN156	QN157	QN158	QN159	QN160	QN161	QN162	QN163												
QN163	QN164	QN165	QN166	QN167	QN168	QN169	QN170	QN171	QN172	QN173	QN174												
QN174	QN175	QN176	QN177	QN178	QN179	QN180	QN181	QN182	QN183	QN184	QN185												
QN185	QN186	QN187	QN188	QN189	QN190	QN191	QN192	QN193	QN194	QN195	QN196												
QN196	QN197	QN198	QN199	QN200																			
1	25	1 000	2 187E-01	2 181E-01	2 175E-01	2 168E-01	2 162E-01	2 156E-01	2 150E-01	2 143E-01	2 137E-01												
2 131E-01	2 125E-01	2 119E-01	2 112E-01	2 106E-01	2 100E-01	2 094E-01	2 088E-01	2 082E-01	2 076E-01	2 070E-01	2 064E-01												
2 063E-01	2 057E-01	2 051E-01	2 045E-01	2 039E-01	2 033E-01	2 027E-01	2 021E-01	2 015E-01	2 009E-01	2 003E-01	1 997E-01												
1 991E-01	1 985E-01	1 979E-01	1 973E-01	1 967E-01	1 961E-01	1 955E-01	1 949E-01	1 943E-01	1 937E-01	1 931E-01	1 925E-01												
1 929E-01	1 923E-01	1 917E-01	1 911E-01	1 905E-01	1 899E-01	1 893E-01	1 887E-01	1 881E-01	1 875E-01	1 869E-01	1 863E-01												
1 867E-01	1 861E-01	1 855E-01	1 849E-01	1 843E-01	1 837E-01	1 831E-01	1 825E-01	1 819E-01	1 813E-01	1 807E-01	1 801E-01												
1 803E-01	1 797E-01	1 791E-01	1 785E-01	1 779E-01	1 773E-01	1 767E-01	1 761E-01	1 755E-01	1 749E-01	1 743E-01	1 737E-01												
-01	1 739E-01	1 733E-01	1 727E-01	1 721E-01	1 715E-01	1 710E-01	1 704E-01	1 698E-01	1 692E-01	1 686E-01	1 680E-01												
E-01	1 675E-01	1 669E-01	1 664E-01	1 658E-01	1 652E-01	1 646E-01	1 640E-01	1 634E-01	1 628E-01	1 622E-01	1 616E-01												
7E-01	1 612E-01	1 606E-01	1 600E-01	1 594E-01	1 588E-01	1 583E-01	1 577E-01	1 571E-01	1 565E-01	1 559E-01	1 553E-01												
53E-01	1 549E-01	1 542E-01	1 536E-01	1 530E-01	1 524E-01	1 518E-01	1 512E-01	1 507E-01	1 501E-01	1 495E-01	1 489E-01												
489E-01	1 483E-01	1 477E-01	1 471E-01	1 465E-01	1 459E-01	1 454E-01	1 448E-01	1 442E-01	1 436E-01	1 430E-01	1 424E-01												
424E-01	1 418E-01	1 412E-01	1 406E-01	1 400E-01	1 394E-01	1 388E-01	1 382E-01	1 376E-01	1 370E-01	1 364E-01	1 358E-01												
1 358E-01	1 352E-01	1 346E-01	1 339E-01	1 333E-01	1 327E-01	1 321E-01	1 315E-01	1 309E-01	1 303E-01	1 297E-01	1 291E-01												
1 290E-01	1 284E-01	1 278E-01	1 272E-01	1 266E-01	1 259E-01	1 253E-01	1 247E-01	1 241E-01	1 235E-01	1 229E-01	1 223E-01												
1 222E-01	1 215E-01	1 209E-01	1 203E-01	1 197E-01	1 190E-01	1 184E-01	1 178E-01	1 171E-01	1 165E-01	1 159E-01	1 153E-01												
1 152E-01	1 146E-01	1 139E-01	1 133E-01	1 126E-01	1 120E-01	1 113E-01	1 107E-01	1 101E-01	1 094E-01	1 088E-01	1 082E-01												
1 081E-01	1 075E-01	1 068E-01	1 062E-01	1 055E-01	1 049E-01	1 042E-01	1 036E-01	1 029E-01	1 022E-01	1 016E-01	1 010E-01												
1 009E-01	1 003E-01	9 963E-02	9 898E-02	9 832E-02	9 766E-02	9 700E-02	9 634E-02	9 568E-02	9 502E-02	9 436E-02	9 370E-02												
1	49	2 00	2 187E-01	2 179E-01	2 170E-01	2 164E-01	2 158E-01	2 151E-01	2 145E-01	2 139E-01	2 133E-01												
2 127E-01	2 120E-01	2 114E-01	2 108E-01	2 102E-01	2 096E-01	2 090E-01	2 084E-01	2 077E-01	2 071E-01	2 065E-01	2 059E-01												
2 059E-01	2 053E-01	2 047E-01	2 041E-01	2 035E-01	2 029E-01	2 023E-01	2 017E-01	2 011E-01	2 005E-01	1 999E-01	1 993E-01												
1 993E-01	1 987E-01	1 981E-01	1 975E-01	1 969E-01	1 963E-01	1 957E-01	1 951E-01	1 945E-01	1 939E-01	1 933E-01	1 927E-01												
1 927E-01	1 921E-01	1 915E-01	1 909E-01	1 904E-01	1 898E-01	1 892E-01	1 886E-01	1 880E-01	1 874E-01	1 868E-01	1 862E-01												
1 862E-01	1 856E-01	1 851E-01	1 845E-01	1 839E-01	1 833E-01	1 827E-01	1 821E-01	1 816E-01	1 810E-01	1 804E-01	1 798E-01												
01	1 798E-01	1 792E-01	1 786E-01	1 781E-01	1 775E-01	1 769E-01	1 763E-01	1 757E-01	1 752E-01	1 746E-01	1 740E-01												

Outline

- 1 Introduction and Motivation
- 2 Getting Started
- 3 3D
- 4 Flight Work and Other Interesting Concepts

Data Comes in Many Forms

- CSV

- ▶ Use native `csv` library from Python
- ▶ Simple, robust
- ▶ Stands for Comma Separated Values
- ▶ Can also read Tab-Delimited Files

- Excel Spreadsheets

- ▶ Install with: `pip install xlrd`
- ▶ Plays nicely with the Excel models of Books, Sheets, and Cells

- Fixed Width Data Files

- ▶ Use native `struct` library from Python
- ▶ Similar to CSVs but lacking a specific data separator.
- ▶ Implemented in C rather than Python (Cython): very fast!

- JSON

- ▶ Use native `requests` library from Python
- ▶ Get data straight from the web.

Matplotlib and Outputting the Figures

- Simple types of plots to plot
 - 1 `plot()` is a marked scatter plot with the individual data points unenumerated by default.
 - 2 `bar()` is a bar plot.
 - 3 `hist()` is a histogram bar plot.
 - 4 `hbar()` is a *horizontal* bar plot.
 - 5 `boxplot()` is a box and whisker plot.
 - 6 `scatter()` is a scatter plot with line markings turned off by default.

Matplotlib and Formatting the Figures

- Methods of changing the appearance of a plot
 - 1 `subplot(int x)` allows you to choose a section of a figure that you want to plot on. For example, `subplot(311)` means that you have a 3-row 1-column plot and you will plot in the 1st (top) section.
 - 2 `title()` gives the graph a title.
 - 3 `xlim()`, `ylim()` allow for the setting of the ranges of the axes.
 - 4 `xticks()`, `yticks()` allow for the placement of tick marks and labels on the graph's axes.
 - 5 `legend()` generates a legend for your graph. You can specify names for the plotted figures in plotting order or use labels passed in at the time of plotting.
 - 6 `annotate()` allows for the highlighting of a specific value or region.

Unlocking Your Full Matplotlib Potential

- This goes much deeper than the above.
- Visit matplotlib.org to check out all optional parameters for each of the above functions.
 - ▶ color and colormaps
 - ▶ thickness
 - ▶ background coloring
 - ▶ location on plot
 - ▶ formatting modes

- Removing outliers

- ▶ If you know what behavior your data should follow, you can remove outliers to make the picture better.

- Smoothing

- ▶ Sometimes in data presentation, it's better to show the big idea rather than all the minute details.
- ▶ Can use median filters (`matplotlib.signal.medfilt()`) or averaging boxes (`convolve()`).

Be Honest!

Don't misrepresent your data! Use the previous tricks to clarify rather than obfuscate.

Outline

- 1 Introduction and Motivation
- 2 Getting Started
- 3 3D
- 4 Flight Work and Other Interesting Concepts

3D Plotting

- Use `mpl_toolkits.mplot3d`, which features the following classes:
 - 1 `axes3d` is a 3D plotting library that works very similar to typical `matplotlib` 2D plotting
 - 2 `axis3d` is an outdated 3D plotting library that apparently suffers from being buggy and poorly designed. Avoid!
 - 3 `art3d` is a 3D art class which is used to build components of `axes3d`, but has some interesting features of its own right.
 - 4 `proj3d` is the background class for these others.
- When plotting in 3D, you must always be careful to specify your dimensions.

Functions to produce 3D plots

- 1 `Axes3D.plot()` gives a marked scatter
- 2 `Axes3D.scatter()` gives an unmarked scatter plot
- 3 `Axes3D.plot_wireframe()` plots a transparent mesh of a surface.
- 4 `Axes3D.plot_surface()` plots a solid surface
- 5 `Axes3D.plot_trisurf()` plots a solid surface made from a Triangulation object
- 6 `Axes3D.contour()` plots a 3D contour
- 7 Others, like quivers, 2D plots, bar plots, polygon plots.

Outline

- 1 Introduction and Motivation
- 2 Getting Started
- 3 3D
- 4 Flight Work and Other Interesting Concepts

Expanding your visualization vocabulary

- There are many other projects and implementations that you can consider incorporating into your data visualization
- `scikits.audiolab` allows you to analyze sound files and plot their frequencies.
 - ▶ `pip install scikits.audiolab`
- If you're feeling confident with your HTML and JSON vocabulary, you can look into Google Visualization API for plotting to the web
- Basemap is a library that makes coordinate generation easy. It's great if you're looking to plot with respect to space.
- PIL and Images (as you might remember) are excellent libraries for reading in images and using them as data.

Looking on to Airline Delays

- Now we can take a page out of FiveThirtyEight's book.
- We can download some airline delay data from <http://www.transtats.bts.gov/> and play around