11 AUGUST 2016 / PORTFOLIO

The Key to Building a Data Science Portfolio That Will Get You a Job

This is the fourth post in a series of posts on how to build a Data Science Portfolio. You can find links to the other posts in this series at the bottom of the post.

In the past few posts in this series, we've talked about how to build a data science project that tells a story, how to build an end to end machine learning project, and how to setup a data science blog. In this post, we'll take a step back, and focus on your portfolio at a high level. We'll discuss what skills employers want to see a candidate demonstrate, and how to build a portfolio that demonstrates all of those skills effectively. We'll include examples of what each project in your portfolio should look like, and give you suggestions on how to get started.

After reading this post, you should have a good understanding of why you should build a data science portfolio, and how to go about doing it.

What employers look for

When employers hire, they're looking for someone who can add value to their business. Often, this means someone who has skills that can generate revenue and opportunities for the business. As a data scientist, you add value to a business in one of 4 main ways:

- Extracting insights from raw data, and presenting those insights to others.
    - An example would be analyzing ad click rates, and discovering that it's much more cost effective to advertise to people who are 18 to 21 then to people who are 21 to 25 -- this adds business value by allowing the business to optimize its ad spend.
- Building systems that offer direct value to the customer.
    - An example would be a data scientist at Facebook optimizing the news feed to show better results to users -- this generates direct revenue for Facebook because more news feed engagement means more ad engagement.
- Building systems that offer direct value to others in the organization.
    - An example would be building a script that automatically aggregates data from 3 databases and generates a clean dataset for others to analyze -- this adds value by making it faster for others to do their work.
- Sharing your expertise with others in the organization.
    - An example is chatting with a product manager about how to build a feature that requires machine learning algorithms -- this adds value by

preventing unrealistic timelines, or a semi-functional product.

Unsurprisingly, when employers look at candidates to hire, they look at people who can do one or more of the four things above (the exact ones they look at depend on the company and role). In order to demonstrate that you can aid a business in the 4 areas listed above, you need to demonstrate some combination of these skills:

- Ability to communicate
- Ability to collaborate with others
- Technical competence
- Ability to reason about data
- Motivation and ability to take initiative

A well-rounded portfolio should show off your skills in each of the above areas, and be relatively easy for someone to scan -- each portfolio item should be well documented and explained, so a hiring manager is able to quickly evaluate your portfolio.

Why a portfolio

If you have a degree in machine learning or a relevant field from a top-tier school, it's **relatively** easy to get a data science job. Employers trust that you can add value to their business because of the prestige of the institution that issued you the degree, and the fact that it's in a subject that's relevant to their own work. If you don't have a relevant degree from a top-tier school, you have to build that trust yourself.

Think about it this way: employers can have up to *200* applicants for in-demand jobs. Let's say that the hiring manager spends *10* hours total filtering the applications down and deciding who to do a phone chat with. This means that each applicant is only evaluated for *3* minutes on average. The hiring manager starts off with no trust that

you can add value to the business, and you have *3* minutes to build their trust to the point where they decide to do a phone screen.

The great thing about data science is that the work you do on your own building projects often looks exactly like the work you'll do once you're hired. Analyzing credit data as a Data Scientist at Lending Club probably has a lot of similarities to analyzing the anonymous loan data that they release.

Notes offered by Prospectus (https://www.lendingclub.com/info/prospectus.action)												
id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	emp_title	emp_length	home_ownership
1077501	1296599	5000	5000	4975	36 months	10.6500%	162.87	B	B2		10+ years	RENT
1077430	1314167	2500	2500	2500	60 months	15.2700%	59.83	C	C4	Ryder	< 1 year	RENT
1077175	1313524	2400	2400	2400	36 months	15.9600%	84.33	C	C5		10+ years	RENT
1076863	1277178	10000	10000	10000	36 months	13.4900%	339.31	C	C1	AIR RESOURCES BOARD	10+ years	RENT
1075358	1311748	3000	3000	3000	60 months	12.6900%	67.79	B	B5	University Medical Group	1 year	RENT
1075269	1311441	5000	5000	5000	36 months	7.900%	156.46	A	A4	Veolia Transportaton	3 years	RENT
1069639	1304742	7000	7000	7000	60 months	15.9600%	170.08	C	C5	Southern Star Photography	8 years	RENT
1072053	1288686	3000	3000	3000	36 months	18.6400%	109.43	E	E1	MKC Accounting	9 years	RENT
1071795	1306957	5600	5600	5600	60 months	21.2800%	152.39	F	F2		4 years	OWN
1071570	1306721	5375	5375	5350	60 months	12.6900%	121.45	B	B5	Starbucks	< 1 year	RENT
1070078	1305201	6500	6500	6500	60 months	14.6500%	153.45	C	C3	Southwest Rural metro	5 years	OWN
1069908	1305008	12000	12000	12000	36 months	12.6900%	402.54	B	B5	UCLA	10+ years	OWN
1064687	1298717	9000	9000	9000	36 months	13.4900%	305.38	C	C1	Va. Dept of Conservation/Recreation	< 1 year	RENT
1069866	1304956	3000	3000	3000	36 months	9.9100%	96.68	B	B1	Target	3 years	RENT
1069057	1303503	10000	10000	10000	36 months	10.6500%	325.74	B	B2	SFMTA	3 years	RENT
1069759	1304871	1000	1000	1000	36 months	16.2900%	35.31	D	D1	Internal revenue Service	< 1 year	RENT
1065775	1299699	10000	10000	10000	36 months	15.2700%	347.98	C	C4	Chin's Restaurant	4 years	RENT
1069971	1304884	3600	3600	3600	36 months	6.0300%	109.57	A	A1	Duracell	10+ years	MORTGAGE

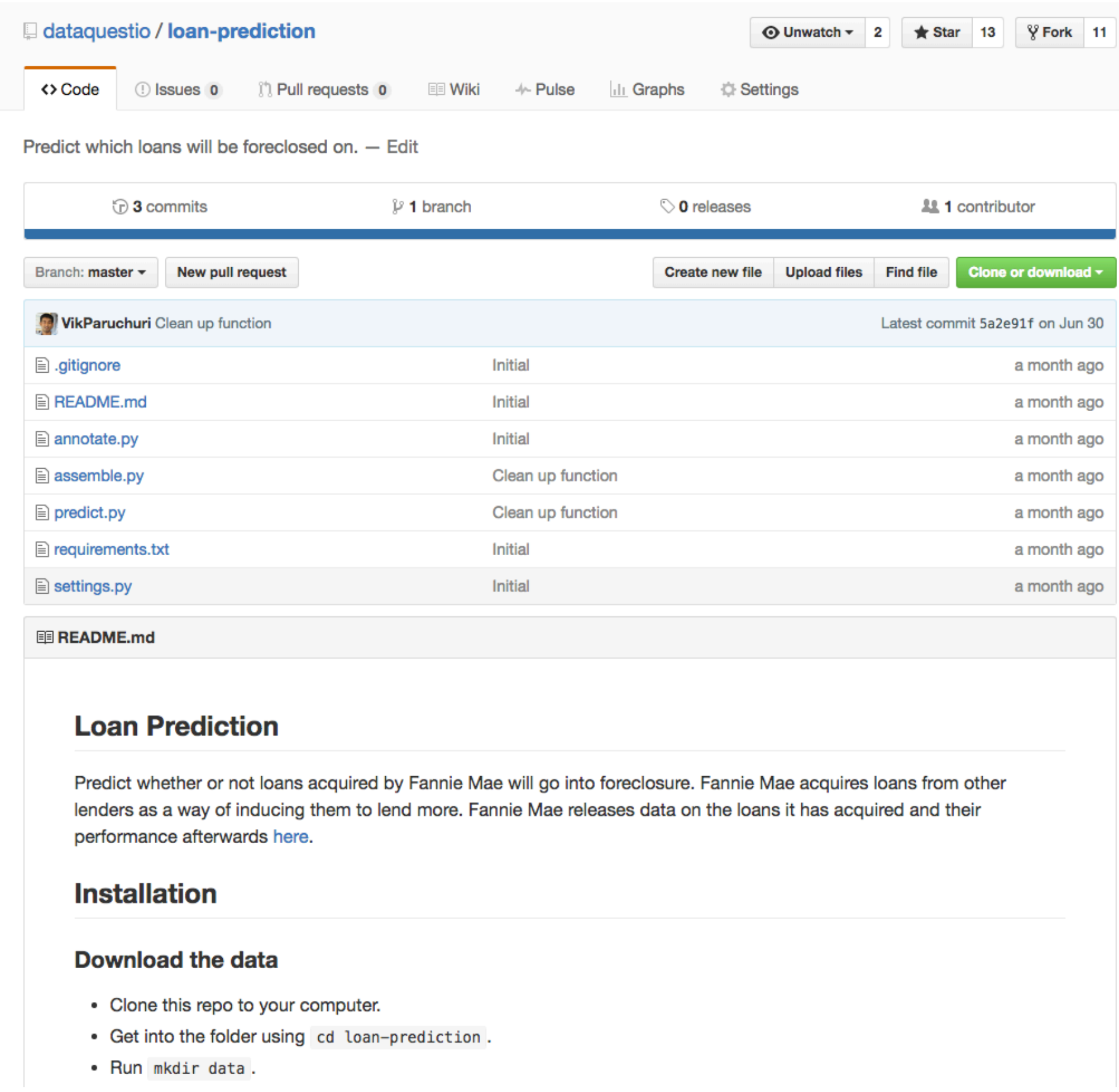
The first few rows of the Lending Club anonymous data.

The number one way to build trust with a hiring manager is to prove you can do the work that they need you to do. With data science, this comes down to building a portfolio of projects. The more "real-world" the projects are, the more the hiring manager will trust that you'll be an asset to the business, and the greater your chances of getting to a phone screen.

What to put in your data science portfolio

Now that we know we need to build a portfolio, we need to figure out what to put into

it. At the minimum, you should have a few projects up on Github or your blog, where the code is visible and well-documented. The easier it is for a hiring manager to find these projects, the easier it is for them to evaluate your skills. Each project should also be as well-documented as possible, with a `README` file both explaining how to set it up, and explaining any quirks about the data.



A well structured project on Github.

We'll walk through a few types of projects that should be in your portfolio. It's suggested to have multiple projects of each type, especially if the type of role you want aligns with one or the other. For example, if you're applying to positions that require a lot of machine learning, building more end to end projects that use machine learning could be useful. On the other hand, if you're applying for analyst positions, data cleaning and storytelling projects are more critical.

Data Cleaning Project

A data cleaning project shows a hiring manager that you can take disparate datasets and make sense of them. This is most of the work a data scientist does, and is a critical skill to demonstrate. This project involves taking messy data, then cleaning it up and doing analysis. A data cleaning project demonstrates that you can reason about data, and can take data from many sources and consolidate it into a single dataset. Data cleaning is a huge part of any data scientist job, and showing that you've done it before will be a leg up.

You'll want to go from raw data to a version that's easy to do analysis with. In order to do this, you'll need to:

- Find a messy dataset
  - Try using data.gov, /r/datasets, or Kaggle Datasets to find something.
  - Avoid picking anything that is already clean -- you want there to be multiple data files, and some nuance to the data.
  - Find any supplemental datasets if you can -- for example, if you downloaded a dataset on flights, are there any datasets you can find via Google that you can combine with it?

    - Try to pick something that interests you personally -- you'll produce a much better final project if you do
- Pick a question to answer using the data
    - Explore the data
    - Identify an interesting angle to explore
- Clean up the data
    - Unify multiple data files if you have them
    - Ensure that exploring the angle you want to is possible with the data
- Do some basic analysis
    - Try to answer the question you picked initially
- Present your results
    - It's recommended to use Jupyter notebook or R Markdown to do the data cleaning and analysis
    - Make sure that your code and logic can be followed, and add as many comments and markdown cells explaining your process as you can
    - Upload your project to Github
    - It's not always possible to include the raw data in your git repository, due to licensing issues, but make sure you at least describe the source data and where it came from

The first part of our earlier post in this series, `Analyzing NYC School Data`, steps you through how to create a complete data cleaning project. You can view it here.


2011 NYC School Survey
Data Dictionary

This data dictionary can be used with the school-level data files from the 2011 NYC School Survey. School-level data is available in one file for all community schools (file name: masterfile11_gened_final) and one file for all District 75 schools (file name: masterfile11_D75_final). These files display one line of information for each school, by DBN, that includes the response rate for each school, the number of surveys submitted, the size of the eligible survey population at each school, question scores, the percentage of responses selected, and the count of responses selected. These fields are detailed below.

Field Name	Field Description
dbn	School identification code (district borough number)
sch_type	School type (Elementary, Middle, High, etc)
location	School name
enrollment	Enrollment size
borough	Borough
principal	Principal name
studentsurvey	Only students in grades 6-12 partipate in the student survey. This field indicates whether or not this school serves any students in grades 6-12.
rr_s	Student Response Rate
rr_t	Teacher Response Rate
rr_p	Parent Response Rate
N_s	Number of student respondents
N_t	Number of teacher respondents
N_p	Number of parent respondents
nr_s	Number of eligible students
nr_t	Number of eligible teachers
nr_p	Number of eligible parents
saf_p_10	Safety and Respect score based on parent responses
com_p_10	Communication score based on parent responses
eng_p_10	Engagement score based on parent responses
aca_p_10	Academic expectations score based on parent responses
saf_t_10	Safety and Respect score based on teacher responses
com_t_10	Communication score based on teacher responses
eng_t_10	Engagement score based on teacher responses
aca_t_10	Academic expectations score based on teacher responses
saf_s_10	Safety and Respect score based on student responses
com_s_10	Communication score based on student responses
eng_s_10	Engagement score based on student responses
aca_s_10	Academic expectations score based on student responses
saf_tot_10	Safety and Respect total score
com_tot_10	Communication total score
eng_tot_10	Engagement total score
aca_tot_10	Academic Expectations total score

A data dictionary of some of the NYC school data.

If you're having trouble finding a good dataset, here are some examples:

- US flight data
- NYC subway turnstile data
- Soccer data

The NYC subway, in all its glory.

If you need some inspiration, here are some examples of good data cleaning projects:

- Analyzing Twitter data
- Cleaning Airbnb data

Data Storytelling Project

A data storytelling project demonstrates your ability to extract insights from data and persuade others. This has a large impact on the business value you can deliver, and is an important piece of your portfolio. This project involves taking a set of data and

telling a compelling narrative with it. For example, you could use data on flights to show that there are significant delays at certain airports, which could be fixed by changing the routing.

A good storytelling project will make heavy use of visualizations, and will take the reader on a path that lets them see each step of the analysis. Here are the steps you'll need to follow to build a good data storytelling project:

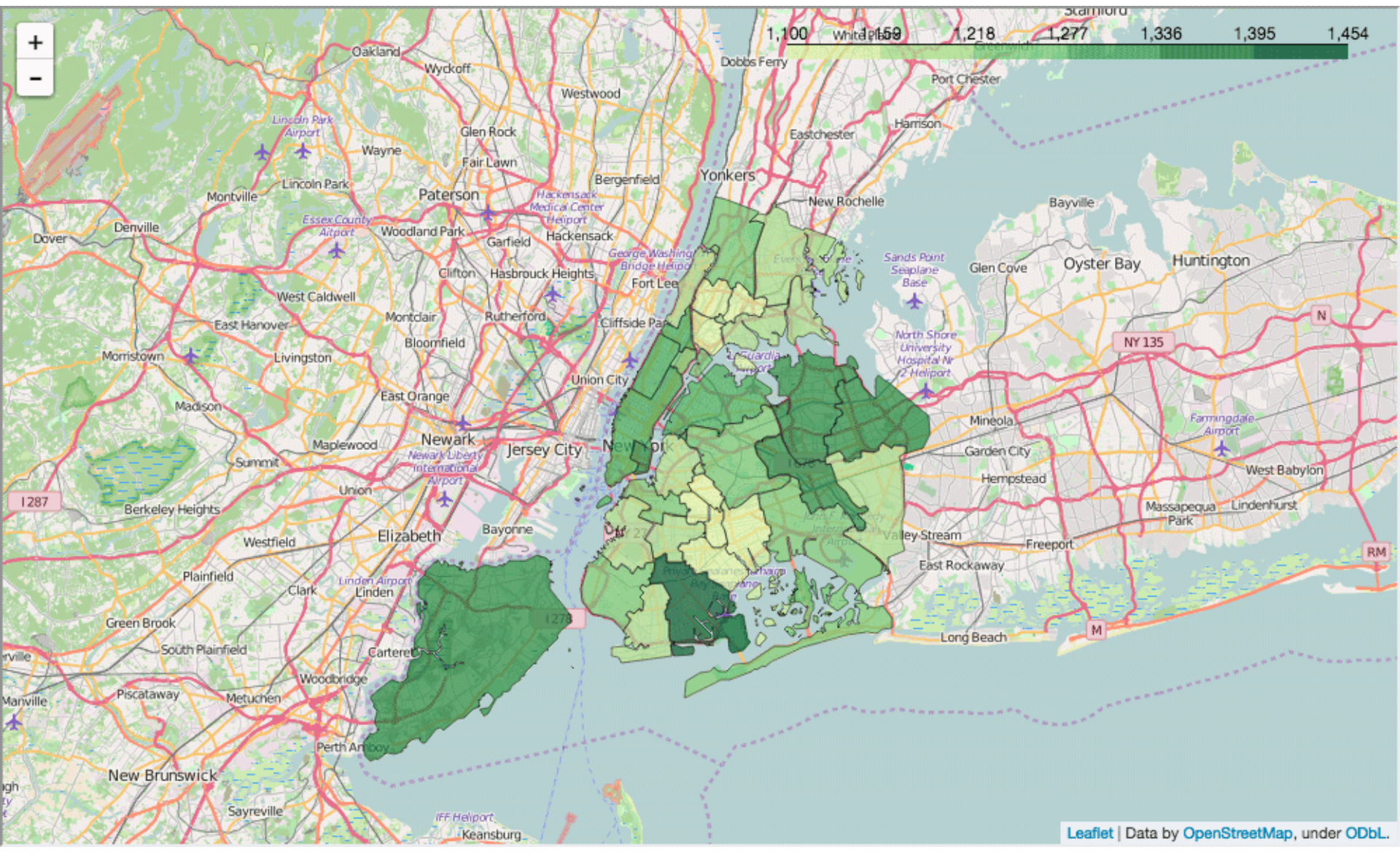
- Find an interesting dataset
  - Try using data.gov, /r/datasets, or Kaggle Datasets to find something.
  - Picking something that is related to current events can be exciting for a reader.
  - Try to pick something that interests you personally -- you'll produce a much better final project if you do
- Explore a few angles in the data
  - Explore the data
  - Identify interesting correlations in the data
  - Create charts and display your findings step-by-step
- Write up a compelling narrative
  - Pick the most interesting angle from your explorations
  - Write up a story around getting from the raw data to the findings you made
  - Create compelling charts that enhance the story
  - Write extensive explanations about what you were thinking at each step, and about what the code is doing
  - Write extensive analysis of the results of each step, and what they tell a

reader

- Teach the reader something as you go through the analysis

- Present your results
  - It's recommended to use Jupyter notebook or R Markdown to do the data analysis
  - Make sure that your code and logic can be followed, and add as many comments and markdown cells explaining your process as you can
  - Upload your project to Github

The second part of our earlier post in this series, `Analyzing NYC School Data`, steps you through how to tell a story with data. You can view it here.



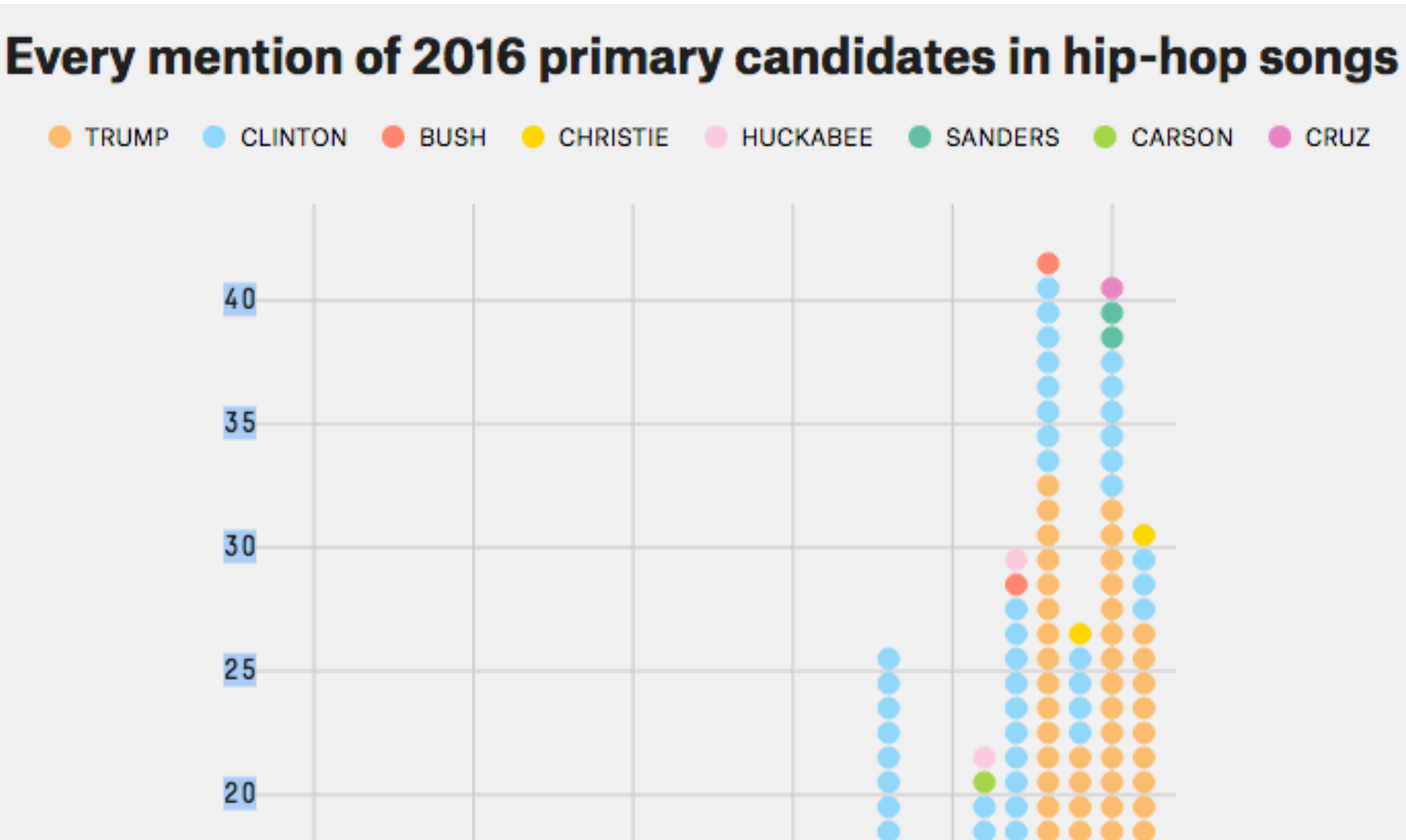
A map of SAT scores by district in NYC.

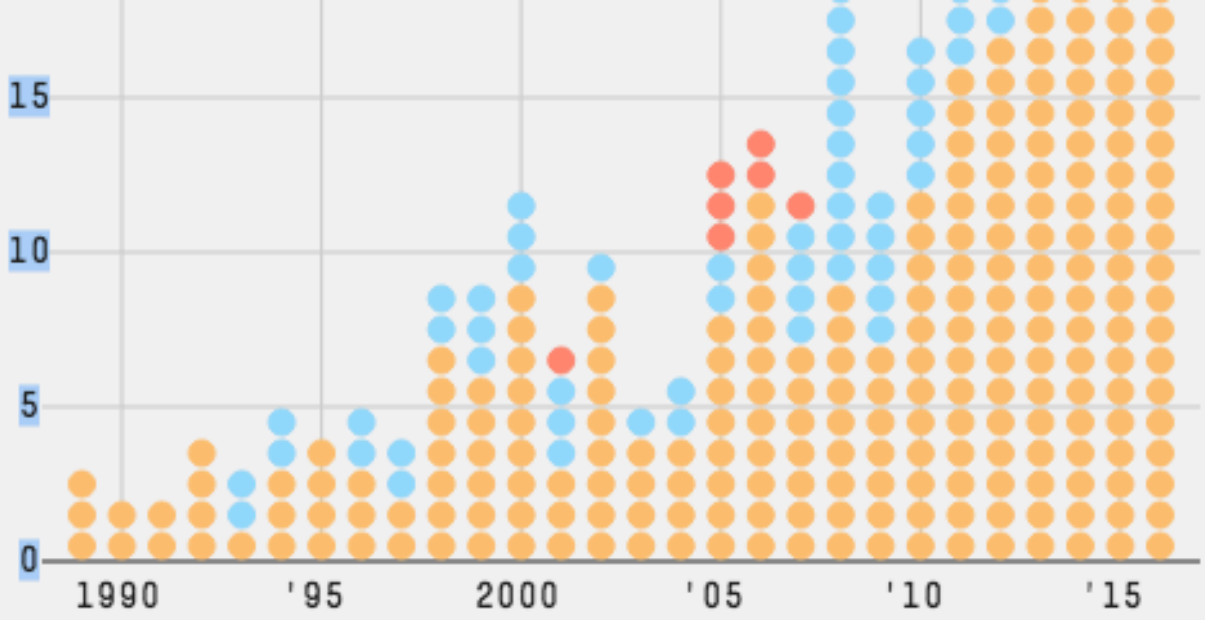
If you're having trouble finding a good dataset, here are some examples:

- Lending club loan data
- FiveThirtyEight's datasets
- Hacker news data

If you need some inspiration, here are some examples of good data storytelling posts:

- Hip-hop and Donald Trump mentions
- Analyzing NYC taxi and Uber data
- Tracking NBA player movements





Lyrics mentioning each primary candidate in the 2016 US elections (from the first project above).

End to End Project

So far, we've covered projects that involve exploratory data cleaning and analysis. This helps a hiring manager who's concerned with how well you can extract insights and present them to others. However, it doesn't show that you're capable of building systems that are customer-facing. Customer-facing systems involve high-performance code that can be run multiple times with different pieces of data to generate different outputs. An example is a system that predicts the stock market -- it will download new market data in every morning, then predict which stocks will do well during the day.

In order to show we can build operational systems, we'll need to build an end to end project. An end to end project takes in and processes data, then generates some output. Often, this is the result of a machine learning algorithm, but it can also be another output, like the total number of rows matching certain criteria.

The key here is to make the system flexible enough to work with new data (like in our stock market data), and high performance. It's also important to make the code easy

to setup and run. Here are the steps you'll need to follow to build a good end to end project:

- Find an interesting topic
  - We won't be working with a single static dataset, so you'll need to find a topic instead
  - The topic should have publicly-accessible data that is updated regularly
  - Some examples:
    - The weather
    - Nba games
    - Flights
    - Electricity pricing
- Import and parse multiple datasets
  - Download as much available data as you're comfortable working with
  - Read in the data
  - Figure out what you want to predict
- Create predictions
  - Calculate any needed features
  - Assemble training and test data
  - Make predictions
- Clean up and document your code
  - Split your code into multiple files
  - Add a README file to explain how to install and run the project

- Add inline documentation
- Make the code easy to run from the command line

- Upload your project to Github

Our earlier post in this series, `Analyzing Fannie Mae loan data`, steps you through how to build an end to end machine learning project. You can view it here.

If you're having trouble finding a good topic, here are some examples:

- Historical S&P 500 data
- Streaming twitter data

S&P 500 (^GSPC) ☆ Add to watchlist
SNP - SNP Delayed Price. Currency in USD

2,185.79 +10.30 (+0.47%)
At close: 4:19 PM EDT

Summary | Conversations | Options | Components | **Historical Data**

Time Period: Aug 11, 2015 - Aug 11, 2016 Show: Historical Prices Frequency: Daily Apply

Currency in USD. Download Data

Date	Open	High	Low	Adj Close*	Volume
Aug 11, 2016	2,177.97	2,188.45	2,177.97	2,185.79	464,238,860
Aug 10, 2016	2,182.81	2,183.41	2,172.00	2,175.49	3,254,950,000
Aug 09, 2016	2,182.24	2,187.66	2,178.61	2,181.74	3,334,300,000
Aug 08, 2016	2,183.76	2,185.44	2,177.85	2,180.89	3,327,550,000
Aug 05, 2016	2,168.79	2,182.87	2,168.79	2,182.87	3,663,070,000
Aug 04, 2016	2,163.51	2,168.19	2,159.07	2,164.25	3,709,200,000

S&P 500 data.

If you need some inspiration, here are some examples of good end to end projects:

- Stock price prediction
- Automatic music generation

Explanatory Post

It's important to be able to understand and explain complex data science concepts, such as machine learning algorithms. This helps a hiring manager understand how good you'd be at communicating complex concepts to other team members and customers. This is a critical piece of a data science portfolio, as it covers a good portion of real-world data science work. This also shows that you understand concepts and how things work at a deep level, not just at a syntax level. This deep understanding is important in being able to justify your choices and walk others through your work.

In order to build an explanatory post, we'll need to pick a data science topic to explain, then write up a blog post taking someone from the very ground level all the way up to having a working example of the concept. The key here is to use plain, simple, language -- the more academic you get, the harder it is for a hiring manager to tell if you actually understand the concept.

The important steps are to pick a topic you understand well, walk a reader through the concept, then do something interesting with the final concept. Here are the steps you'll need to follow:

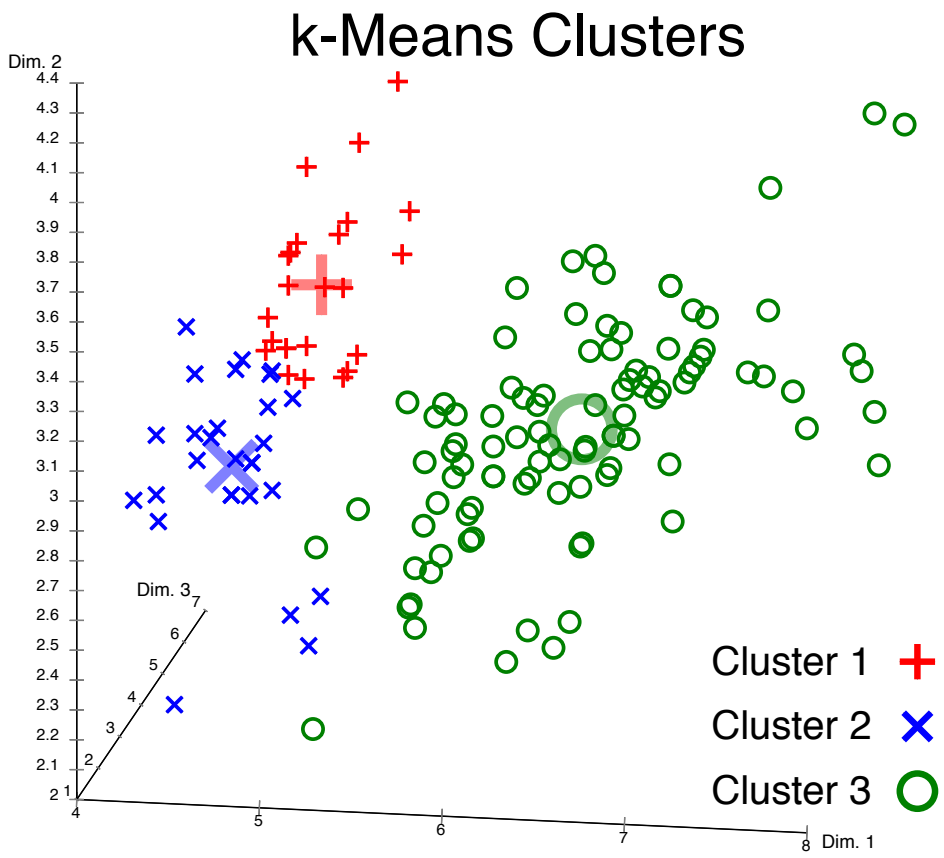
- Find a concept you know well or can learn
    - Machine learning algorithms like k-nearest neigbhors are good

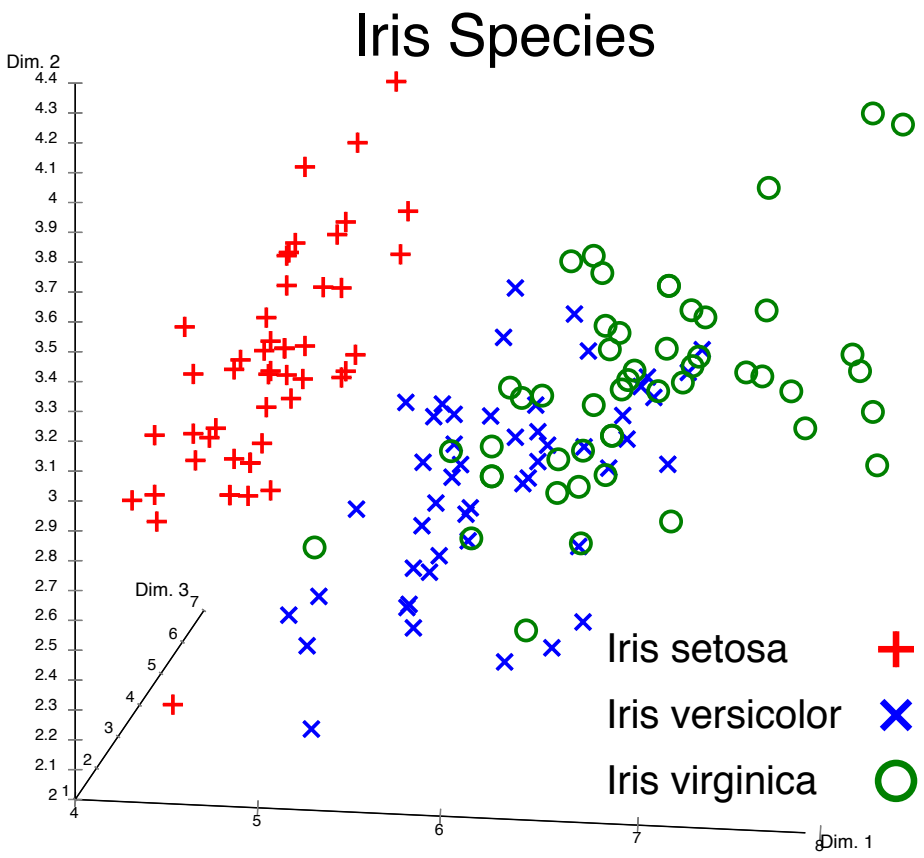
concepts to pick.
    - Statistical concepts are also good to pick.
    - Make sure that the concept has enough nuance to spend some time explaining.
    - Make sure you fully understand the concept, and it's not too complex to explain well.
- Pick a dataset or "scaffold" to help you explain the concept.
    - For instance, if you pick k-nearest neighbors, you could explain k-nearest neighbors by using NBA data (finding similar players).
- Create an outline of your post
    - Assume that the reader has no knowledge of the topic you're explaining
    - Break the concept into small steps
        - For k-nearest neighbors, this might be:
            - Predicting using similarity
            - Measures of similarity
            - Euclidean distance
            - Finding a match using k=1
            - Finding a match with k > 1
- Write up your post
    - Explain everything in clear and straightforward language
    - Make sure to tie everything back to the "scaffold" you picked when possible
    - Try having someone non-technical reading it, and gauge their reaction

- Share your post
  - Preferably post on your own blog
  - If not, upload to Github

If you're having trouble finding a good concept, here are some examples:

- k-means clustering
- Matrix multiplication
- Chi-squared test





Visualizing kmeans clustering.

If you need some inspiration, here are some examples of good explanatory blog posts:

- Linear regression

- Natural language processing
- Naive Bayes
- k-nearest neighbors

Optional portfolio pieces

While the key is to have a set of projects on your blog or Github, it can also be useful to add other components to your project, like Quora answers, talks, and data science competition results. These are often of secondary concern to a hiring manager, but they can be a great way to stand out and prove your skills.

Talk

Talks can be useful ways to help teach others and demonstrate to hiring managers that you know a topic well enough to teach others. This helps a hiring manager understand how good you are at communication and presentation. These skills overlap with some of the portfolio pieces somewhat, but are still useful to demonstrate.

The most common place to give a talk is at a local Meetup. Meetups are all focused around specific topics, like "Python", or "Data visualization using D3".

To give a great talk, here are some good steps:

- Find an interesting project you worked on or concept you know
  - A good place to look is your own portfolio projects and blog posts
  - Whatever you pick should fit with the theme of the meetup
- Break the project down into slides

    - You'll want to break the project down into a series of slides
    - Each slide should have as little text as possible
- Practice your talk a few times
- Give the talk!
- Upload your slides to Github or your blog

If you need some inspiration, here are some examples of good talks:

- Computational statistics
- Scikit-learn vs Spark for ML pipelines
- Analyzing NHL penalties

Data science competition

Data science competitions involve trying to train the most accurate machine learning model on a set of data. These competitions can be a great way to learn. From a hiring manager's perspective, a data science competition can demonstrate technical competence if you do well, initiative if you put in a good amount of effort, and collaboration if you work with others. This overlaps with some of the other portfolio projects, but it can be a nice secondary way to stand out.

Most data science competitions are hosted on Kaggle or DrivenData.

To participate in a data science competition, you just need to sign up for the site and get started! A good competition to get started with is here, and you can find a set of tutorials here.

The leaderboard of a Kaggle competition.

Wrapping up

You should now have a good idea of what skills to demonstrate in your portfolio, and how to go about building your portfolio. Now it's time to start building! If you have a portfolio you'd like to show off, please let us know in the comments!

At Dataquest, our interactive guided projects are designed to help you start building a data science portfolio to demonstrate your skills to employers and get a job in data. If you're interested, you can signup and do our first module for free.

If you liked this, you might like to read the other posts in our 'Build a Data Science Portfolio' series:

- *Storytelling with data*.
- *How to setup up a data science blog*.
- *Building a machine learning project*.
- *17 places to find datasets for data science projects*

- *How to present your data science portfolio on Github*

SUBSCRIBE TO OUR MAILING LIST!

Vik Paruchuri

Read more posts by this author.

Read More

— Dataquest Data Science Blog —

Portfolio

How to present your data science portfolio on GitHub

18 places to find data sets for data science projects

Building a data science portfolio: Machine learning project

See all 5 posts →

LEARN PYTHON

Sep 07, 2016

Working with streaming data: Using the Twitter API to capture tweets

In this post, you'll learn how to work with streaming data for data science in Python, with an example using Twitter's API.

VIK PARUCHURI

DATA SCIENCE PROJECTS

Jul 20, 2016

How I built a Slack bot to help me find an apartment in San Francisco

Learn how to build and deploy a bot to identify the best rental properties using Craigslist, Slack, and Python.

VIK PARUCHURI

Dataquest Data Scie

Latest Posts · Fac