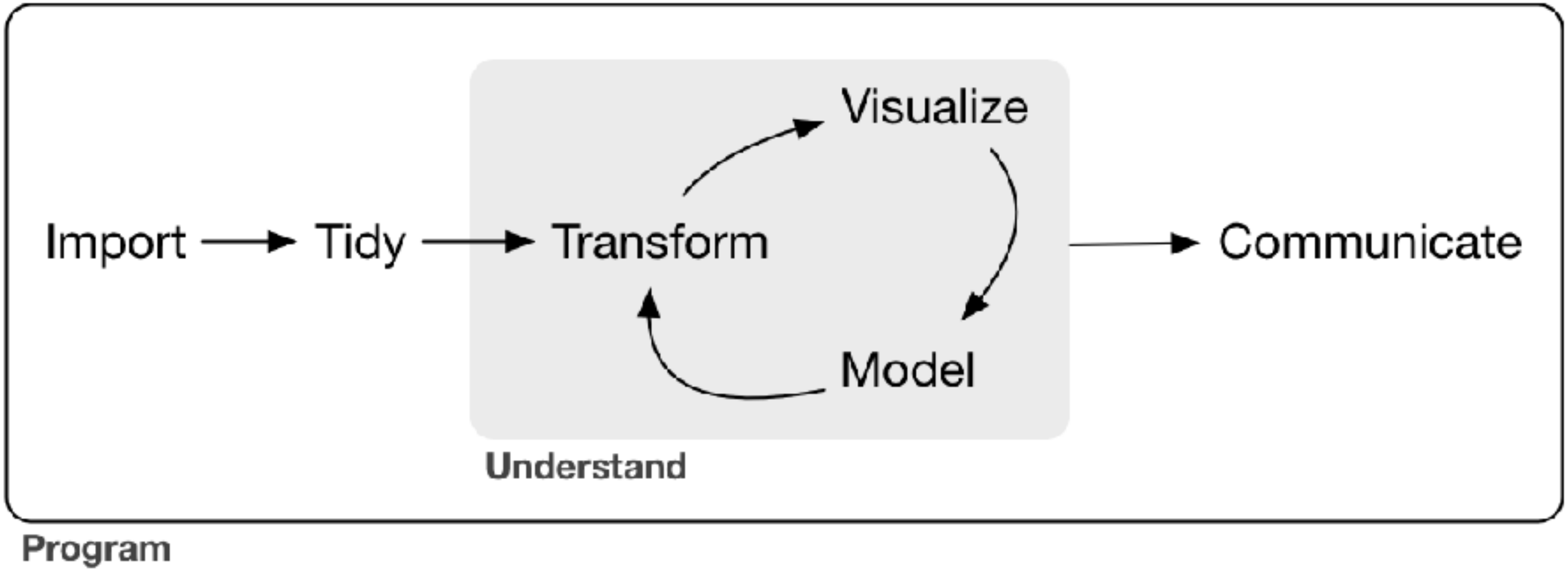
Data Science

using

python™

What is Data Science ?

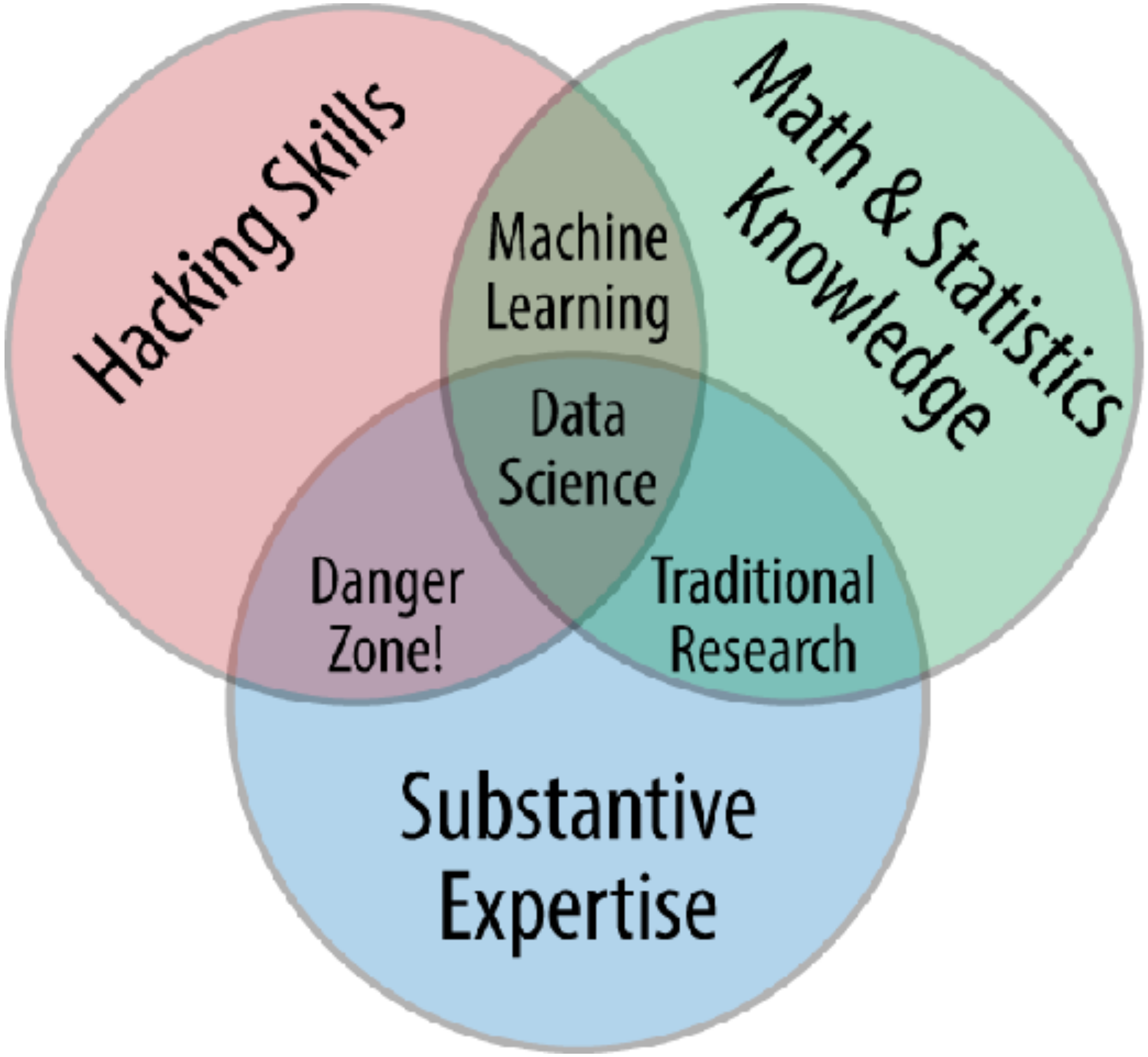
“turning raw data into understanding, insight, and knowledge”.



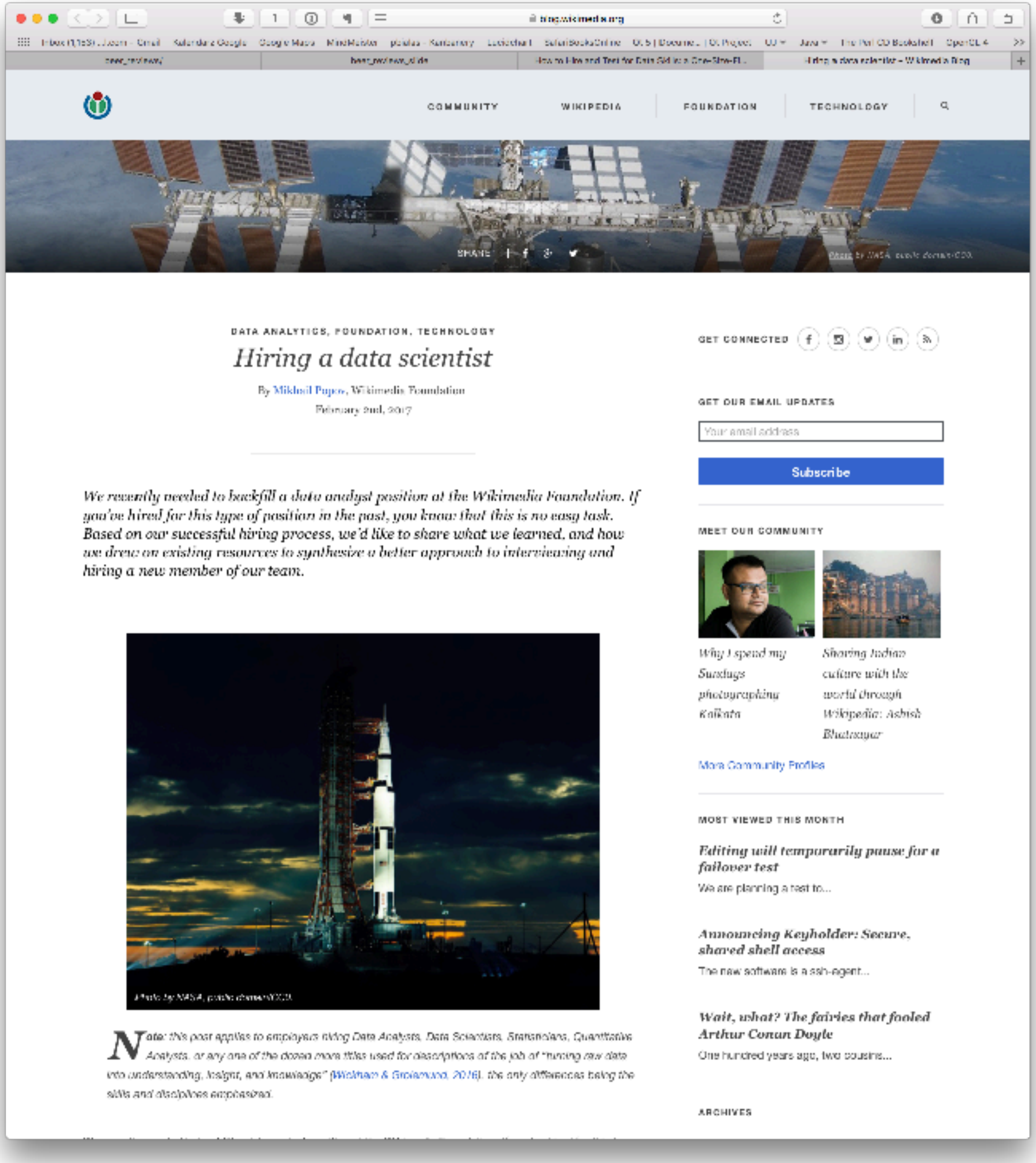
Wickham, H., i Grolemund, G. (2017). *R for Data Science.* O'Reilly Media, Inc.

- the skills of a statistician who knows how to model and summarize datasets (which are growing ever larger);
- the skills of a computer scientist who can design and use algorithms to efficiently store, process, and visualize this data;
- and the domain expertise—what we might think of as ‘classical’ training in a subject—necessary both to formulate the right questions and to put their answers in context.”

VanderPlas, J. (2017). *Python Data Science Handbook.* O'Reilly Media ,Inc.



VanderPlas, J. (2017). *Python Data Science Handbook.* O'Reilly Media ,Inc.



DATA ANALYTICS, FOUNDATION, TECHNOLOGY

Hiring a data scientist

By Mikhail Popov, Wikimedia Foundation
February 2nd, 2017

We recently needed to backfill a data analyst position at the Wikimedia Foundation. If you've hired for this type of position in the past, you know that this is no easy task. Based on our successful hiring process, we'd like to share what we learned, and how we drew on existing resources to synthesize a better approach to interviewing and hiring a new member of our team.

Photo by NASA, public domain/CC0.

Note: this post applies to employees titled Data Analysts, Data Scientists, Statisticians, Quantitative Analysts, or any one of the dozen more titles used for descriptions of the job of "turning raw data into understanding, insight, and knowledge" (Wickham & Grolemund, 2016), the only differences being the skills and disciplines emphasized.

Why I spend my Sundays photographing Kolkata

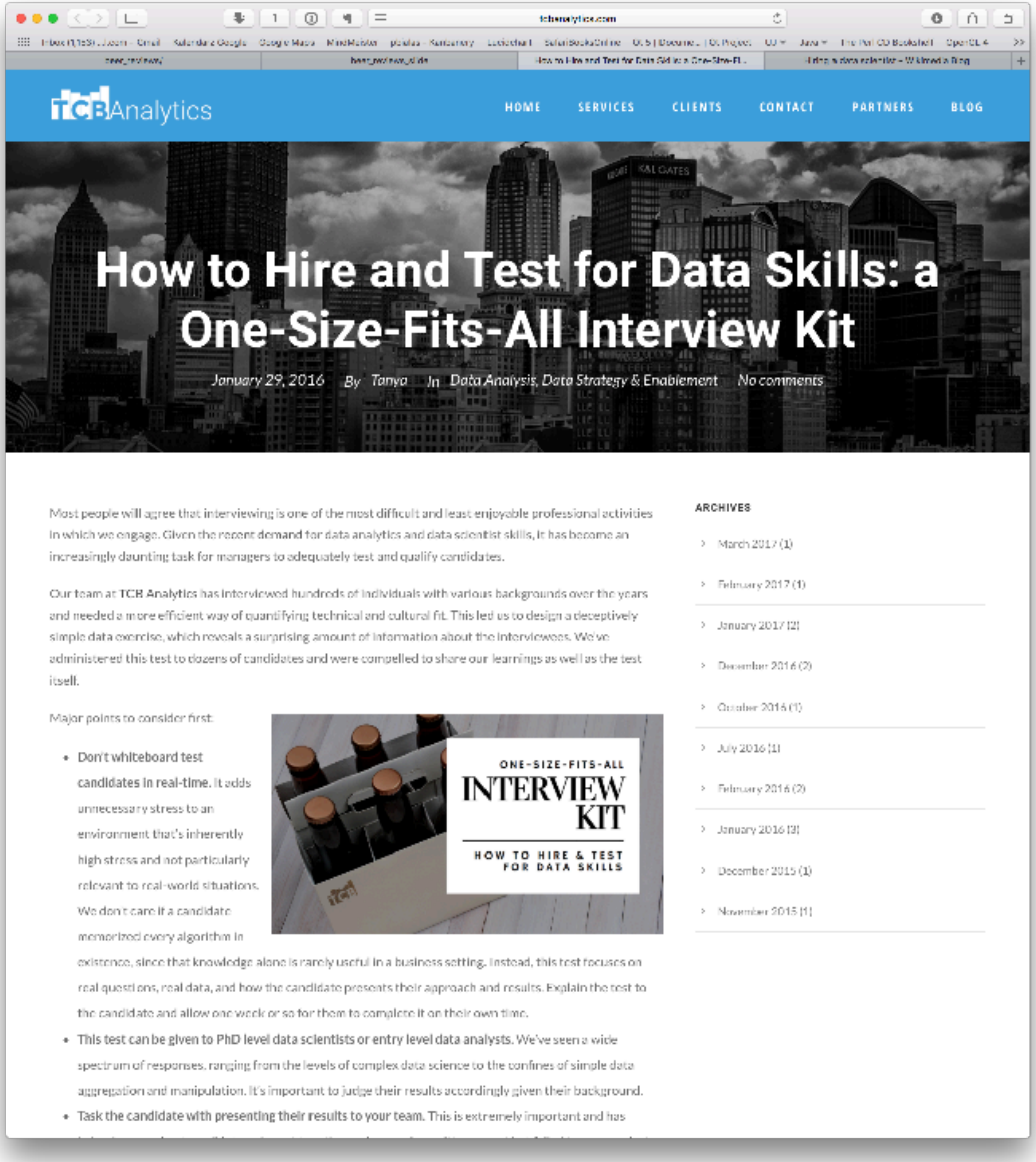
Sharing Indian culture with the world through Wikipedia: Ashish Bhatnagar

More Community Profiles

Editing will temporarily pause for a failover test
We are planning a test to...

Announcing Keyholder: Secure, shared shell access
The new software is a ssh-agent...

Wait, what? The fairies that fooled Arthur Conan Doyle
One hundred years ago, two cousins...



How to Hire and Test for Data Skills: a One-Size-Fits-All Interview Kit

January 29, 2016 *By Tanya* *In Data Analysis, Data Strategy & Enablement* *No comments*

Most people will agree that interviewing is one of the most difficult and least enjoyable professional activities in which we engage. Given the recent demand for data analytics and data scientist skills, it has become an increasingly daunting task for managers to adequately test and qualify candidates.

Our team at **TCB Analytics** has interviewed hundreds of individuals with various backgrounds over the years and needed a more efficient way of quantifying technical and cultural fit. This led us to design a deceptively simple data exercise, which reveals a surprising amount of information about the interviewees. We've administered this test to dozens of candidates and were compelled to share our learnings as well as the test itself.

Major points to consider first:

- **Don't whiteboard test candidates in real-time.** It adds unnecessary stress to an environment that's inherently high stress and not particularly relevant to real-world situations. We don't care if a candidate memorized every algorithm in existence, since that knowledge alone is rarely useful in a business setting. Instead, this test focuses on real questions, real data, and how the candidate presents their approach and results. Explain the test to the candidate and allow one week or so for them to complete it on their own time.
- **This test can be given to PhD level data scientists or entry level data analysts.** We've seen a wide spectrum of responses, ranging from the levels of complex data science to the confines of simple data aggregation and manipulation. It's important to judge their results accordingly given their background.
- **Task the candidate with presenting their results to your team.** This is extremely important and has

ARCHIVES

- March 2017 (1)
- February 2017 (1)
- January 2017 (2)
- December 2016 (2)
- October 2016 (1)
- July 2016 (1)
- February 2016 (2)
- January 2016 (3)
- December 2015 (1)
- November 2015 (1)

1. brewery_id
2. brewery_name
3. review_time
4. review_overall
5. review_arome
6. review_appearance
7. review_profilename
8. beer_style
9. review_palate
10. review_taste
11. beer_name
12. beer_abv
13. beer_beerid

```
10325,Vecchio Birraio,1234817823,1.5,2,2.5,stcules,Hefeweizen,1.5,1.5,Sausa Weizen,
5,47986
10325,Vecchio Birraio,1235915097,3,2.5,3,stcules,English Strong Ale,3,3,Red Moon,
6.2,48213
10325,Vecchio Birraio,1235916604,3,2.5,3,stcules,Foreign / Export Stout,3,3,Black Horse
Black Beer,6.5,48215
10325,Vecchio Birraio,1234725145,3,3,3.5,stcules,German Pilsener,2.5,3,Sausa Pils,5,47969
1075,Caldera Brewing Company,1293735206,4,4.5,4,johnmichaelsen,American Double / Imperial
IPA,4,4.5,Cauldron DIPA,7.7,64883
1075,Caldera Brewing Company,1325524659,3,3.5,3.5,oline73,Herbed / Spiced Beer,
3,3.5,Caldera Ginger Beer,4.7,52159
1075,Caldera Brewing Company,1318991115,3.5,3.5,3.5,Reidrover,Herbed / Spiced Beer,
4,4,Caldera Ginger Beer,4.7,52159
1075,Caldera Brewing Company,1306276018,3,2.5,3.5,alpinebryant,Herbed / Spiced Beer,
2,3.5,Caldera Ginger Beer,4.7,52159
1075,Caldera Brewing Company,1290454503,4,3,3.5,LordAdmNelson,Herbed / Spiced Beer,
3.5,4,Caldera Ginger Beer,4.7,52159
```

and 1.5 million more …

- Which brewery produces the strongest beers by ABV%?
- If you had to pick 3 beers to recommend using only this data, which would you pick?
- Which of the factors (aroma, taste, appearance, palette) are most important in determining the overall quality of a beer?
- Lastly, if I typically enjoy a beer due to its aroma and appearance, which beer style should I try?

`Please document your code and explain the reasoning behind your answers.`

Tools

R

python
TM

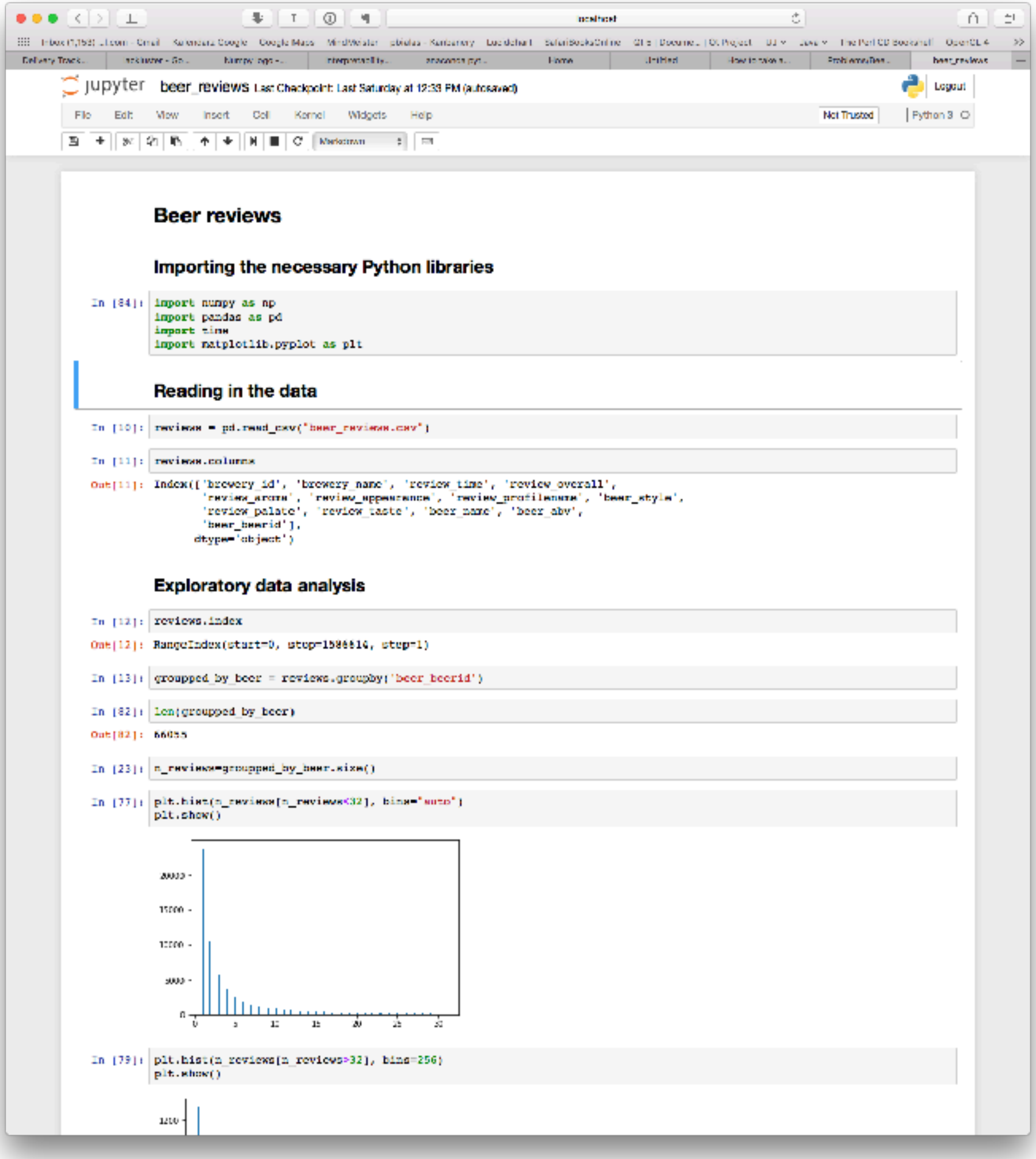
Anaconda/miniconda

pip
conda

Jupyter notebooks

IP[y]:
IPython

jupyter



Beer reviews

Importing the necessary Python libraries

```
In [84]: import numpy as np
         import pandas as pd
         import time
         import matplotlib.pyplot as plt
```

Reading in the data

```
In [10]: reviews = pd.read_csv('beer_reviews.csv')
```

```
In [11]: reviews.columns
```

```
Out[11]: Index(['brewery_id', 'brewery_name', 'review_time', 'review_overall',
                'review_aroma', 'review_appearance', 'review_profilename', 'beer_style',
                'review_palate', 'review_taste', 'beer_name', 'beer_abv',
                'beer_beerid'],
               dtype='object')
```

Exploratory data analysis

```
In [12]: reviews.index
```

```
Out[12]: RangeIndex(start=0, stop=1586614, step=1)
```

```
In [13]: groupped_by_beer = reviews.groupby('beer_beerid')
```

```
In [82]: len(groupped_by_beer)
```

```
Out[82]: 66055
```

```
In [23]: n_reviews=groupped_by_beer.size()
```

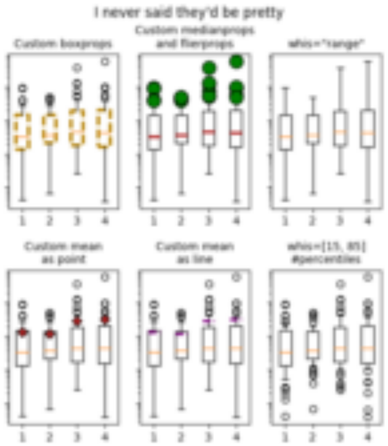
```
In [77]: plt.hist(n_reviews[n_reviews<32], bins='auto')
         plt.show()
```

```
In [79]: plt.hist(n_reviews[n_reviews>32], bins=256)
         plt.show()
```

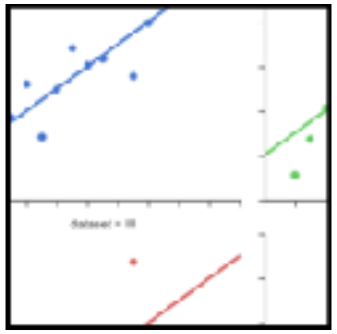
NumPy

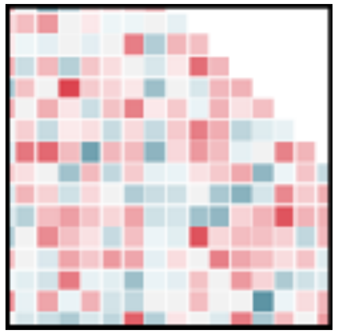
SciPy

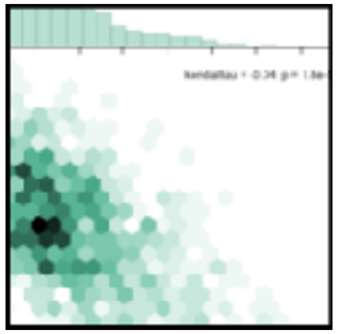
matplotlib/pyplot



Seaborn







pandas