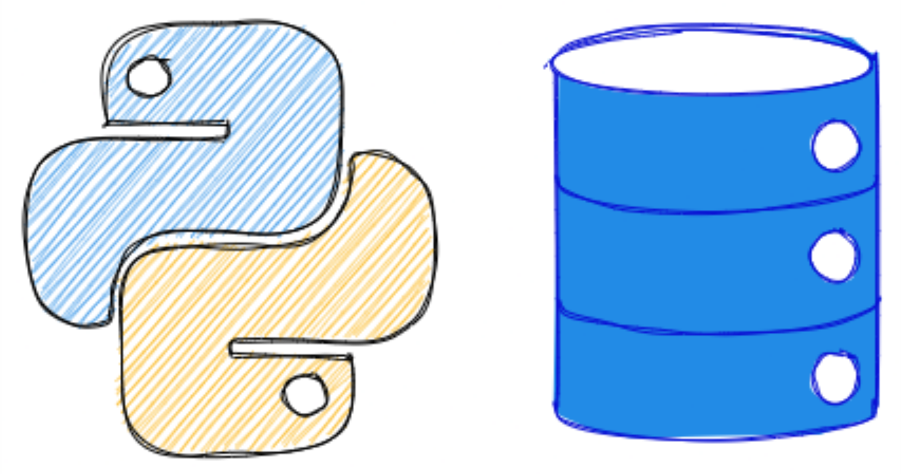
Efficient Python Tricks and Tools for Data Scientists - By Khuyen Tran

Get Data

GitHub View on GitHub   Book View Book

This section covers tools to get some data for your projects.



faker: Create Fake Data in One Line of Code

```
!pip install Faker
```

To quickly create fake data for testing, use faker.

```
>>> from faker import Faker

>>> fake = Faker()

>>> fake.color_name()
```

```
'CornflowerBlue'
```

```
>>> fake.name()
```

```
'Michael Scott'
```

```
>>> fake.address()
```

```
'881 Patricia Crossing\nSouth Jeremy, AR 06087'
```

```
>>> fake.date_of_birth(minimum_age=22)
```

```
datetime.date(1927, 11, 5)
```

```
>>> fake.city()
```

```
'North Donald'
```

```
>>> fake.job()
```

```
'Teacher, secondary school'
```

Link to faker

Link to my full article on faker.

Random User: Generate Random User Data in One Line of Code

Have you ever wanted to create fake user data for testing? Random User Generator is a free API that generates random user data. Below is how to download and use this data in your code.

```
import json
from urllib.request import urlopen
# Show 1 random users
data = urlopen("https://randomuser.me/api?
results=1").read()
users = json.loads(data)["results"]
users
```

```
[{'gender': 'female',
  'name': {'title': 'Miss', 'first': 'Ava',
'last': 'Hansen'},
  'location': {'street': {'number': 3526, 'name':
'George Street'},
   'city': 'Worcester',
   'state': 'Merseyside',
   'country': 'United Kingdom',
   'postcode': 'K7Z 3WB',
   'coordinates': {'latitude': '11.9627',
'longitude': '17.6871'},
   'timezone': {'offset': '+9:00',
    'description': 'Tokyo, Seoul, Osaka, Sapporo,
Yakutsk'}},
  'email': 'ava.hansen@example.com',
  'login': {'username': 'heavywolf743',
   'password': 'cristina'},
  'dob': {'date': '1948-01-21T10:26:00.053Z',
'age': 73},
  'registered': {'date': '2011-11-
19T03:28:46.830Z', 'age': 10},
  'phone': '015242 07811',
  'cell': '0700-326-155',
  'picture': {'large':
'https://randomuser.me/api/portraits/women/60.jpg
',
   'medium':
'https://randomuser.me/api/portraits/med/women/60
.jpg',
   'thumbnail':
'https://randomuser.me/api/portraits/thumb/women/
60.jpg'}}]
```

Link to Random User Generator.

fetch_openml: Get OpenML's Dataset in One Line of Code

OpenML has many interesting datasets. The easiest way to get OpenML's data in Python is to use the `sklearn.datasets.fetch_openml` method.

In one line of code, you get the OpenML's dataset to play with!

```
from sklearn.datasets import fetch_openml

monk = fetch_openml(name="monks-problems-2",
as_frame=True)
print(monk["data"].head(10))
```

```
   attr1 attr2 attr3 attr4 attr5 attr6
0      1     1     1     1     2     2
1      1     1     1     1     4     1
2      1     1     1     2     1     1
3      1     1     1     2     1     2
4      1     1     1     2     2     1
5      1     1     1     2     3     1
6      1     1     1     2     4     1
7      1     1     1     3     2     1
8      1     1     1     3     4     1
9      1     1     2     1     1     1
```

Autoscraper

```
!pip install autoscraper
```

If you want to get the data from some websites, Beautifulsoup makes it easy for you to do so. But can scraping be automated even more? If you are looking for a faster way to scrape some complicated websites such as Stackoverflow, Github in a few lines of codes, try autoscraper.

All you need is to give it some texts so it can recognize the rule, and it will take care of the rest for you!

```python
from autoscraper import AutoScraper

url = "https://stackoverflow.com/questions/2081586/web-scraping-with-python"

wanted_list = ["How to check version of python modules?"]

scraper = AutoScraper()
result = scraper.build(url, wanted_list)

for res in result:
    print(res)
```

```
How to execute a program or call a system command?
What are metaclasses in Python?
Does Python have a ternary conditional operator?
Convert bytes to a string
Does Python have a string 'contains' substring method?
How to check version of python modules?
```

Link to autoscraper.

pandas-reader: Extract Data from Various Internet Sources Directly into a Pandas DataFrame

```
!pip install pandas-datareader
```

Have you wanted to extract series data from various Internet sources directly into a pandas DataFrame? That is when pandas_reader comes in handy.

Below is the snippet to extract daily data of AD indicator from 2008 to 2018.

```
import os
from datetime import datetime
import pandas_datareader.data as web

df = web.DataReader(
    "AD",
    "av-daily",
    start=datetime(2008, 1, 1),
    end=datetime(2018, 2, 28),
    api_key=os.gehide-
outputtenv("ALPHAVANTAGE_API_KEY"),
)
```

Link to pandas_reader.

pytrends: Get the Trend of a Keyword on Google Search Over Time

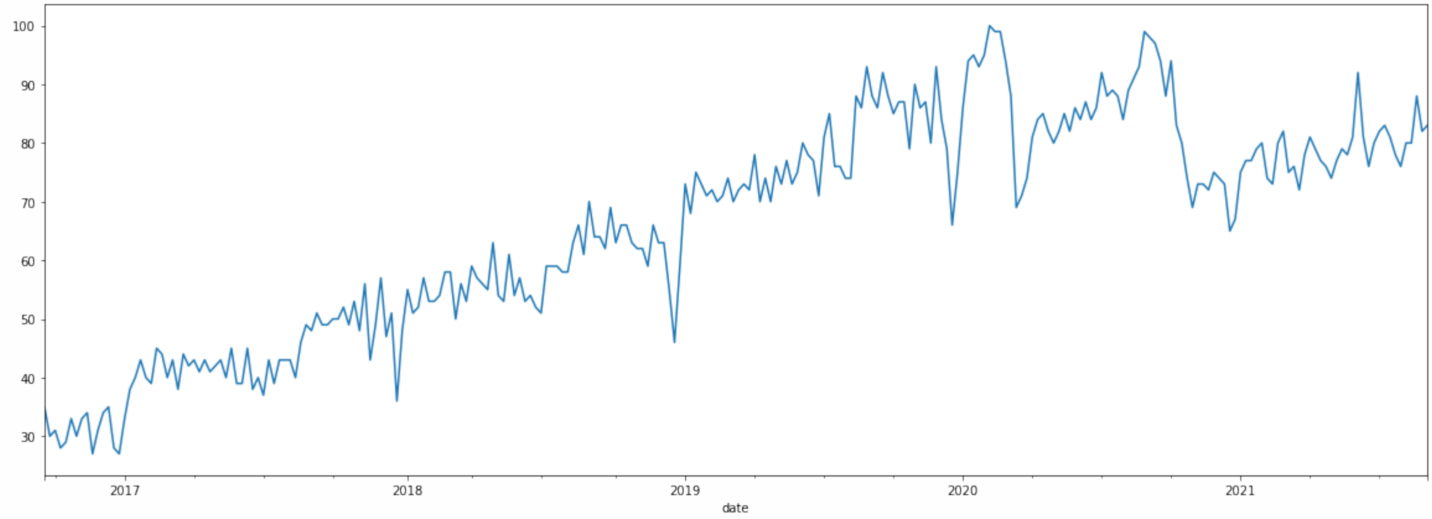
```
!pip install pytrends
```

If you want to get the trend of a keyword on Google Search over time, try pytrends.

In the code below, I use pytrends to get the interest of the keyword "data science" on Google Search from 2016 to 2021.

```python
from pytrends.request import TrendReq
pytrends = TrendReq(hl="en-US", tz=360)
pytrends.build_payload(kw_list=["data science"])

df = pytrends.interest_over_time()
df["data science"].plot(figsize=(20, 7))
```



Link to pytrends

snscrape: Scrape Social Networking Services in Python

If you want to scrape social networking services such as Twitter, Facebook, Reddit, etc, try snscrape.

For example, you can use snsscrape to scrape all tweets from a user or get the latest 100 tweets with the hashtag #python.

```
# Scrape all tweets from @KhuyenTran16
snscrape twitter-user KhuyenTran16

# Save outputs
snscrape twitter-user KhuyenTran16 >>
khuyen_tweets

# Scrape 100 tweets with hashtag python
snscrape --max-results 100 twitter-hashtag python
```

Link to snscrape.

Datacommons: Get Statistics about a Location in One Line of Code

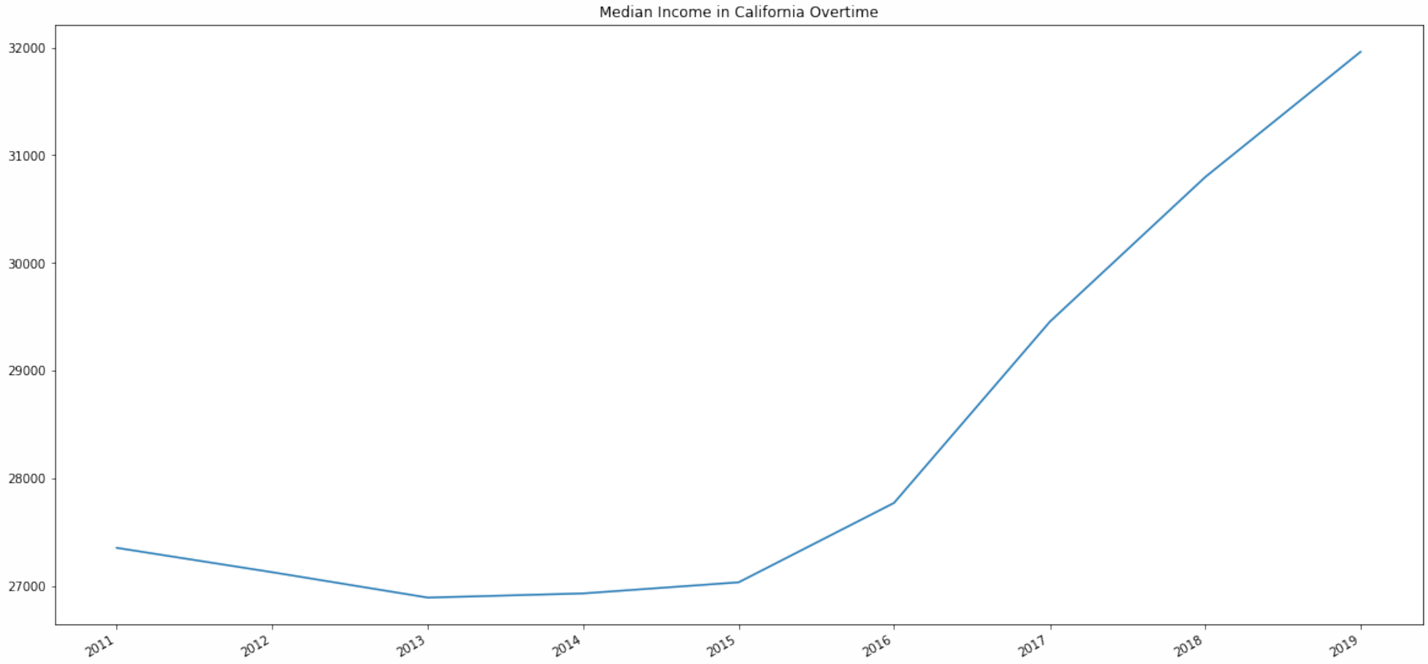
```
!pip install datacommons
```

If you want to get some interesting statistics about a location in one line of code, try Datacommons. Datacommons is a publicly available data from open sources (census.gov, cdc.gov, data.gov, etc.). Below are some statistics extracted from Datacommons.

```
import datacommons_pandas
import plotly.express as px
import pandas as pd
```

Find the Median Income in California Over Time

```
median_income =
datacommons_pandas.build_time_series("geoId/06",
"Median_Income_Person")
median_income.index =
pd.to_datetime(median_income.index)
median_income.plot(
    figsize=(20, 10),
    x="Income",
    y="Year",
    title="Median Income in California Over
Time",
)
```


Median Income in California Overtime
32000
31000
30000
29000
28000
27000
2011
2012
2013
2014
2015
2016
2017
2018
2019

Number of People in the U.S Over Time

```
def process_ts(statistics: str):
    count_person =
datacommons_pandas.build_time_series('country/USA
', statistics)
    count_person.index =
pd.to_datetime(count_person.index)
    count_person.name = statistics
    return count_person
```

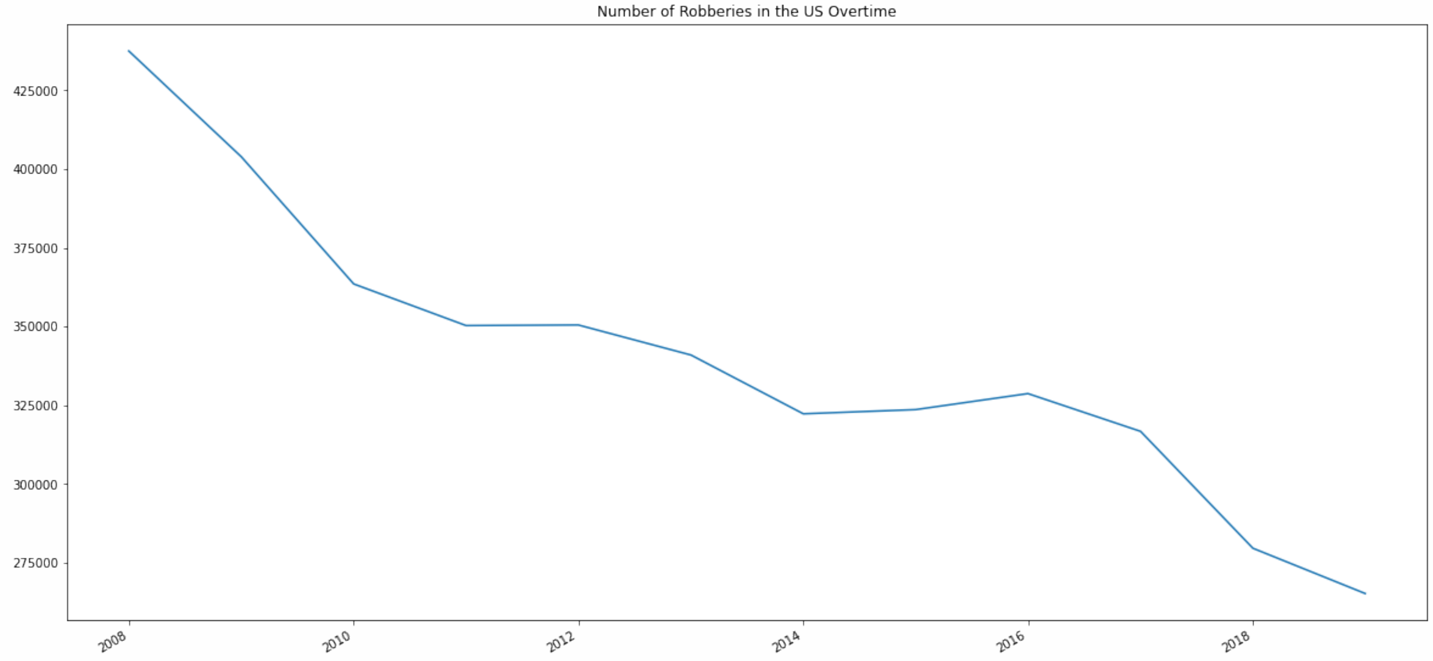
```
count_person_male =
process_ts('Count_Person_Male')
count_person_female =
process_ts('Count_Person_Female')
```

```
count_person = pd.concat([count_person_female,
count_person_male], axis=1)

count_person.plot(
    figsize=(20, 10),
    title="Number of People in the U.S Over
Time",
)
```

Number of Robberies in the US Over Time

```
count_robbery =
datacommons_pandas.build_time_series(
    "country/USA",
"Count_CriminalActivities_Robbery"
)
count_robbery.index =
pd.to_datetime(count_robbery.index)
count_robbery.plot(
    figsize=(20, 10),
    title="Number of Robberies in the US Over
Time",
)
```



Link to Datacommons.

Get Google News Using Python

```
!pip install GoogleNews
```

If you want to get Google news in Python, use GoogleNews. GoogleNews allows you to get search results for a keyword in a specific time interval.

```
from GoogleNews import GoogleNews
googlenews = GoogleNews()
```

```
googlenews.set_time_range('02/01/2022','03/25/2022')
```

```
googlenews.search('funny')
```

```
googlenews.results()
```

```
[{'title': 'Hagan has fastest NHRA Funny Car run
in 4 years',
  'media': 'ESPN',
  'date': 'Feb 26, 2022',
  'datetime': datetime.datetime(2022, 2, 26, 0,
0),
  'desc': '-- Matt Hagan made the quickest Funny
Car run in four years Saturday, \ngiving the new
Tony Stewart Racing NHRA team its first No. 1
qualifier and \nsetting the...',
  'link':
'https://www.espn.com/racing/story/_/id/33381149/
matt-hagan-fastest-nhra-funny-car-pass-4-years',
  'img': 'data:...'},
  ...
  ]
```

Link to GoogleNews.