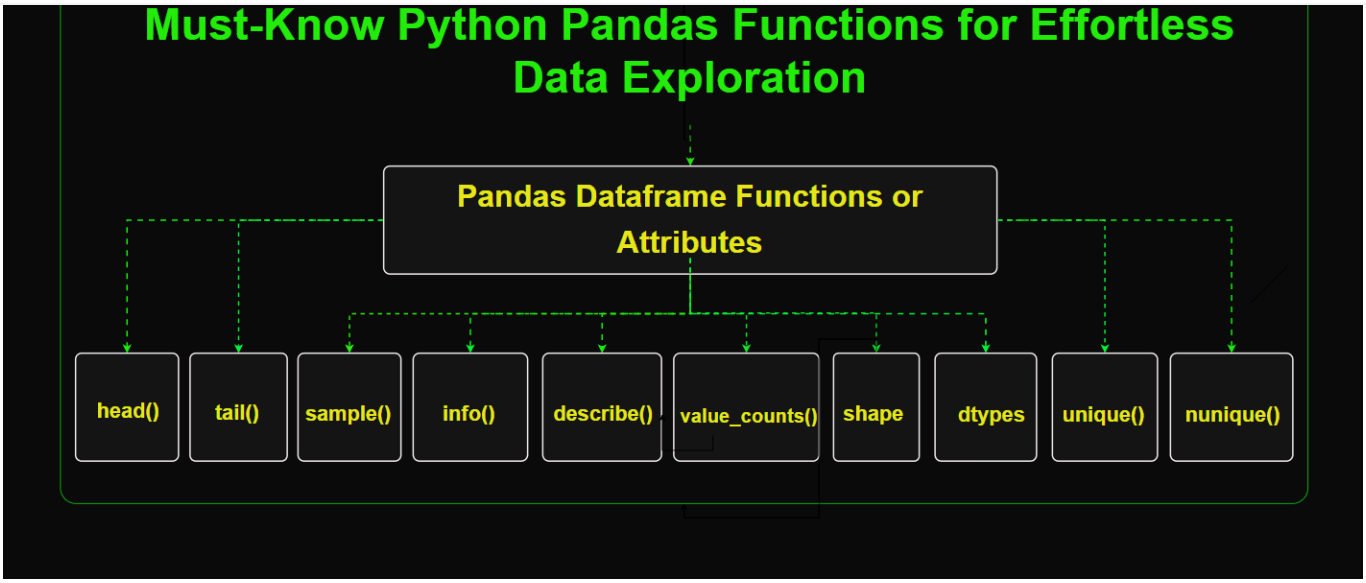
Must-Know Python Pandas Functions for Effortless Data Exploration

Gen. David L. · Follow
5 min read · Nov 27, 2024



The key point of data analysis lies in uncovering the stories behind the data. To achieve this, it's essential to build a solid foundation by thoroughly exploring and understanding the data. In this process, Python's Pandas

library plays a crucial role. Pandas not only offers a powerful set of features but also provides high flexibility, making data exploration both easy and efficient.

In this post we will introduce several must-know Pandas methods for effective data exploration.

Create a CSV sample dataset

To better illustrate and demonstrate how to use Pandas functions, let's first create a sample CSV dataset.

```
import random
import csv
from faker import Faker

# Initialize Faker
fake = Faker()

# List of products and their categories
products = [
    {"name": "Laptop", "category": "Electronics", "price": 899.99},
    {"name": "Smartphone", "category": "Electronics", "price": 699.99},
    {"name": "Headphones", "category": "Accessories", "price": 49.99},
    {"name": "Coffee Maker", "category": "Home Appliances", "price": 79.99},
    {"name": "Sneakers", "category": "Fashion", "price": 59.99},
    {"name": "Backpack", "category": "Fashion", "price": 39.99},
    {"name": "Blender", "category": "Home Appliances", "price": 99.99},
    {"name": "Desk Chair", "category": "Furniture", "price": 129.99},
    {"name": "Water Bottle", "category": "Accessories", "price": 19.99},
    {"name": "Notebook", "category": "Stationery", "price": 5.99},
]

# Define a function to generate order data
def generate_order_data(num_rows):
    data = []
    for _ in range(num_rows):
        product = random.choice(products)
        quantity = random.randint(1, 10)
```

```
            total = round(product["price"] * quantity, 2)
            order = {
                "Order ID": fake.uuid4(),
                "Customer Name": fake.name(),
                "Customer Email": fake.email(),
                "Product Name": product["name"],
                "Category": product["category"],
                "Quantity": quantity,
                "Price": product["price"],
                "Total": total,
                "Order Date": fake.date_this_year(),
                "Shipping Address": fake.address(),
            }
            data.append(order)
    return data

# Generate 1000 rows of data
num_rows = 1000
order_data = generate_order_data(num_rows)

# Save the data to a CSV file
output_file = "sample_orders.csv"
with open(output_file, mode="w", newline="", encoding="utf-8") as file:
    writer = csv.DictWriter(file, fieldnames=order_data[0].keys())
    writer.writeheader()
    writer.writerows(order_data)

print(f"Sample dataset with {num_rows} rows has been saved to '{output_file}'.")
```

This code ensures that the sample dataset is saved as a structured CSV file (`sample_orders.csv`) for further data analysis by pandas functions.

Head function head()

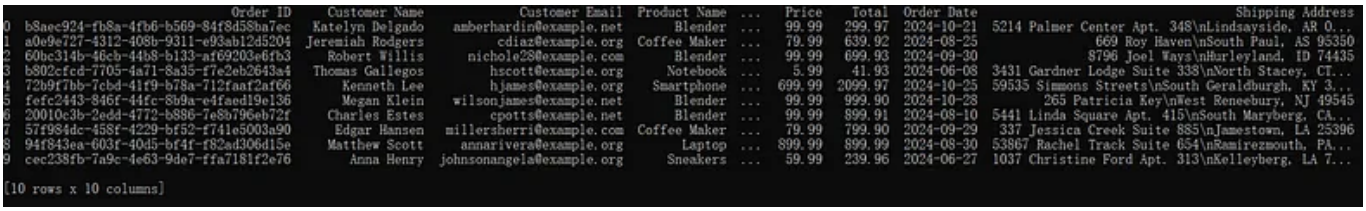
head(): Used to preview the top rows of the sample dataset.

```
import pandas as pd

# Read the sample_orders.csv file into a Pandas DataFrame
```

```
df = pd.read_csv("sample_orders.csv")

# Display the first 10 rows of the dataset
print(df.head(10))
```



```
                                Order ID     Customer Name          Customer Email Product Name  ...   Price    Total Order Date                                     Shipping Address
0  b8aec924-fb8a-4fb6-b569-84f8d58ba7ec    Katelyn Delgado   amberhardin@example.net      Blender  ...   99.99   299.97 2024-10-21  5214 Palmer Center Apt. 348\nLindsayside, AR 0...
1  a0e9e727-4312-408b-9311-e93ab12d5204  Jeremiah Rodgers        cdiaz@example.org  Coffee Maker  ...   79.99   639.92 2024-08-25               669 Roy Haven\nSouth Paul, AS 95350
2  60bc314b-46cb-44b8-b133-af69203e6fb3     Robert Willis    nichole28@example.com      Blender  ...   99.99   699.93 2024-09-30            8796 Joel Ways\nHurleyland, ID 74435
3  b802cfcd-7705-4a71-8a35-f7e2eb2643a4   Thomas Gallegos       hscott@example.org     Notebook  ...    5.99    41.93 2024-06-08  3431 Gardner Lodge Suite 338\nNorth Stacey, CT...
4  72b9f7bb-7cbd-41f9-b78a-712faaf2af66      Kenneth Lee       hjames@example.org   Smartphone  ...  699.99  2099.97 2024-10-25  59535 Simmons Streets\nSouth Geraldburgh, KY 3...
5  fefc2443-846f-44fc-8b9a-e4faed19e136       Megan Klein  wilsonjames@example.net      Blender  ...   99.99   999.90 2024-10-28          265 Patricia Key\nWest Reneebury, NJ 49545
6  20010c3b-2edd-4772-b886-7e8b796eb72f     Charles Estes       cpotts@example.net      Blender  ...   99.99   899.91 2024-08-10  5441 Linda Square Apt. 415\nSouth Maryberg, CA...
7  57f984dc-458f-4229-bf52-f741e5003a90      Edgar Hansen  millersherri@example.com  Coffee Maker  ...   79.99   799.90 2024-09-29   337 Jessica Creek Suite 885\nJamestown, LA 25396
8  94f843ea-603f-40d5-bf4f-f82ad306d15e     Matthew Scott     annarivera@example.org       Laptop  ...  899.99   899.99 2024-08-30  53867 Rachel Track Suite 654\nRamirezmouth, PA...
9  cec238fb-7a9c-4e63-9de7-ffa7181f2e76       Anna Henry  johnsonangela@example.org     Sneakers  ...   59.99   239.96 2024-06-27  1037 Christine Ford Apt. 313\nKelleyberg, LA 7...

[10 rows x 10 columns]
```

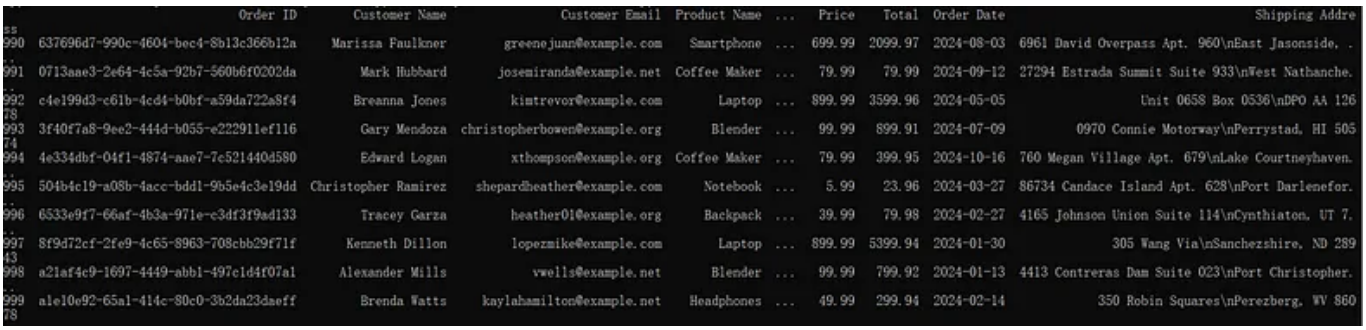
Tail function tail()

tail(): Used to preview the bottom rows of the sample dataset.

```
import pandas as pd

# Read the sample_orders.csv file into a Pandas DataFrame
df = pd.read_csv("sample_orders.csv")

# Display the last 10 rows of the dataset
print(df.tail(10))
```



```
                                Order ID     Customer Name          Customer Email Product Name  ...   Price    Total Order Date                                     Shipping Addre
ss
990  637696d7-990c-4604-bec4-8b13c366b12a    Marissa Faulkner     greenejuan@example.com   Smartphone  ...  699.99  2099.97 2024-08-03  6961 David Overpass Apt. 960\nEast Jasonside, .
..
991  0713aae3-2e64-4c5a-92b7-560b6f0202da      Mark Hubbard  josemiranda@example.net  Coffee Maker  ...   79.99    79.99 2024-09-12  27294 Estrada Summit Suite 933\nWest Nathanche.
..
992  c4e199d3-c61b-4cd4-b0bf-a59da722a8f4      Breanna Jones       kimtrevor@example.com       Laptop  ...  899.99  3599.96 2024-05-05                Unit 0658 Box 0536\nDPO AA 126
78
993  3f40f7a8-9ee2-444d-b055-e222911ef116       Gary Mendoza  christopherbowen@example.org      Blender  ...   99.99   899.91 2024-07-09          0970 Connie Motorway\nPerrystad, HI 505
74
994  4e334dbf-04f1-4874-aae7-7c521440d580       Edward Logan       xthompson@example.org  Coffee Maker  ...   79.99   399.95 2024-10-16  760 Megan Village Apt. 679\nLake Courtneyhaven.
..
995  504b4c19-a08b-4acc-bdd1-9b5e4c3e19dd  Christopher Ramirez    shepardheather@example.com     Notebook  ...    5.99    23.96 2024-03-27  86734 Candace Island Apt. 628\nPort Darlenefor.
..
996  6533e9f7-66af-4b3a-971e-c3df3f9ad133       Tracey Garza       heather01@example.org     Backpack  ...   39.99    79.98 2024-02-27  4165 Johnson Union Suite 114\nCynthiaton, UT 7.
..
997  8f9d72cf-2fe9-4c65-8963-708cbb29f71f      Kenneth Dillon       lopezmike@example.com       Laptop  ...  899.99  5399.94 2024-01-30              305 Wang Via\nSanchezshire, ND 289
43
998  a21af4c9-1697-4449-abb1-497c1d4f07a1     Alexander Mills            vwells@example.net      Blender  ...   99.99   799.92 2024-01-13  4413 Contreras Dam Suite 023\nPort Christopher.
..
999  a1e10e92-65a1-414c-80c0-3b2da23daeff       Brenda Watts  kaylahamilton@example.net   Headphones  ...   49.99   299.94 2024-02-14           350 Robin Squares\nPerezberg, WV 860
78
```

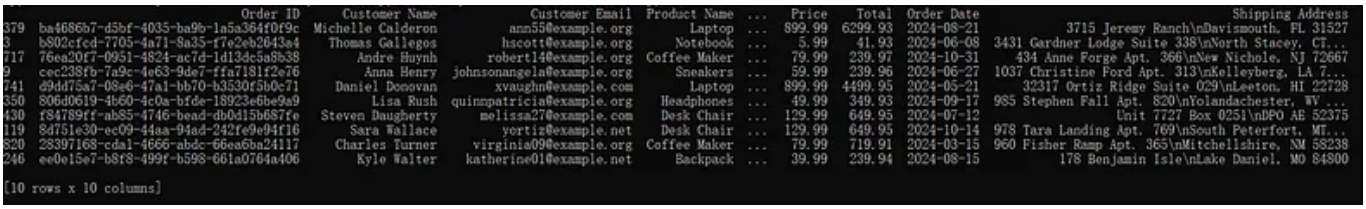
Sample function sample()

sample(): This function is highly valuable when working with large datasets. When we need to extract and analyze a smaller subset from a larger DataFrame, `sample()` helps efficiently retrieve random samples, enabling preliminary data exploration or performance evaluation.

```
import pandas as pd

# Read the sample_orders.csv file into a Pandas DataFrame
df = pd.read_csv("sample_orders.csv")

# Read and display the random 10 rows from the dataset
print(df.sample(10))
```



	Order ID	Customer Name	Customer Email	Product Name	...	Price	Total	Order Date	Shipping Address
379	ba4686b7-d5bf-4035-ba9b-1a5a364f0f9c	Michelle Calderon	ann55@example.org	Laptop	...	899.99	6299.93	2024-08-21	3715 Jeremy Ranch\nDavismouth, FL 31527
8	b802cfcd-7705-4a71-8a35-f7e2eb2643a4	Thomas Gallegos	hscott@example.org	Notebook	...	5.99	41.93	2024-06-08	3431 Gardner Lodge Suite 338\nNorth Stacey, CI...
717	76ea20f7-0951-4824-ac7d-1d13dc5a8b38	Andre Huynh	robert14@example.org	Coffee Maker	...	79.99	239.97	2024-10-31	434 Anne Forge Apt. 366\nNew Nichole, NJ 72667
9	cec238fb-7a9c-4e63-9de7-ffa7181f2e76	Anna Henry	johnsonangela@example.org	Sneakers	...	59.99	239.96	2024-06-27	1037 Christine Ford Apt. 313\nKelleyberg, LA 7...
741	d9dd75a7-08e6-47a1-bb70-b3530f5b0c71	Daniel Donovan	xvaughn@example.com	Laptop	...	899.99	4499.95	2024-05-21	32317 Ortiz Ridge Suite 029\nLeeton, HI 22728
350	806d0619-4b60-4c0a-bfde-18923e6be9a9	Lisa Rush	quinnpatricia@example.org	Headphones	...	49.99	349.93	2024-09-17	985 Stephen Fall Apt. 820\nYolandachester, WV ...
430	f84789ff-ab85-4746-bead-db0d15b687fe	Steven Daugherty	melissa27@example.com	Desk Chair	...	129.99	649.95	2024-07-12	Unit 7727 Box 0251\nDPO AE 52375
119	8d751e30-ec09-44aa-94ad-242fe9e94f16	Sara Wallace	yortiz@example.net	Desk Chair	...	129.99	649.95	2024-10-14	978 Tara Landing Apt. 769\nSouth Peterfort, MT...
820	28397168-cda1-4666-abdc-66ea6ba24117	Charles Turner	virginia09@example.org	Coffee Maker	...	79.99	719.91	2024-03-15	960 Fisher Ramp Apt. 365\nMitchellshire, NM 58238
246	ee0e15e7-b8f8-499f-b598-661a0764a406	Kyle Walter	katherine01@example.net	Backpack	...	39.99	239.94	2024-08-15	178 Benjamin Isle\nLake Daniel, MO 84800

[10 rows x 10 columns]

Information function info()

info(): This function provides a summary of the dataset, including the number of entries, column names, data types, and memory usage.

```
import pandas as pd

# Read the sample_orders.csv file into a Pandas DataFrame
df = pd.read_csv("sample_orders.csv")

# Display a summary of the dataset
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 10 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Order ID          1000 non-null   object
 1   Customer Name     1000 non-null   object
 2   Customer Email    1000 non-null   object
 3   Product Name      1000 non-null   object
 4   Category          1000 non-null   object
 5   Quantity          1000 non-null   int64
 6   Price             1000 non-null   float64
 7   Total             1000 non-null   float64
 8   Order Date        1000 non-null   object
 9   Shipping Address  1000 non-null   object
dtypes: float64(2), int64(1), object(7)
memory usage: 78.3+ KB
None
```

Describe function describe()

describe(): This function provides basic statistical information about the dataset, such as mean, standard deviation, minimum and maximum values, and quartiles.

```
import pandas as pd

# Read the sample_orders.csv file into a Pandas DataFrame
df = pd.read_csv("sample_orders.csv")

# Display the basic statistical information about the dataset
print(df.describe())
```

```
          Quantity        Price        Total
count  1000.000000  1000.000000  1000.000000
mean      5.437000   201.742000  1071.141630
std       2.863954   297.815676  1863.671317
min       1.000000     5.990000     5.990000
25%       3.000000    39.990000   119.980000
50%       5.000000    59.990000   354.920000
75%       8.000000   129.990000   799.920000
max      10.000000   899.990000  8999.900000
```

Value counts function value_counts()

value_counts(): This method returns the count of all unique values in a column or a pandas Series.

```
import pandas as pd

# Read the sample_orders.csv file into a Pandas DataFrame
df = pd.read_csv("sample_orders.csv")

# Display the count of all unique values in a column,such as "Category"
print(df["Category"].value_counts())
```

Shape attribute

shape: This attribute returns the number of rows and columns in the dataset.

```
import pandas as pd

# Read the sample_orders.csv file into a Pandas DataFrame
df = pd.read_csv("sample_orders.csv")

# Display the number of rows and columns in the dataset
print(df.shape)
```

```
(1000, 10)
```

Dataframe dtypes attribute

df.dtypes: This attribute returns the data types of all columns.

```
import pandas as pd

# Read the sample_orders.csv file into a Pandas DataFrame
df = pd.read_csv("sample_orders.csv")

# Display the data types of all columns
print(df.dtypes)
```

```
Order ID            object
Customer Name       object
Customer Email      object
Product Name        object
Category            object
Quantity             int64
Price              float64
Total              float64
Order Date          object
Shipping Address    object
dtype: object
```

Unique function unique()

unique(): This method returns all unique values in a column or a pandas Series.

```
import pandas as pd

# Read the sample_orders.csv file into a Pandas DataFrame
df = pd.read_csv("sample_orders.csv")
```

```
# Display all unique values in a column.
print(df["Category"].unique())
```

```
['Home Appliances' 'Stationery' 'Electronics' 'Fashion' 'Accessories'
 'Furniture']
```

Nunique function nunique()

nunique(): This function returns the number of unique values in a DataFrame.

```
import pandas as pd

# Read the sample_orders.csv file into a Pandas DataFrame
df = pd.read_csv("sample_orders.csv")

# Display the count of unique values in the dataset, sorted in descending order
df.nunique().sort_values(ascending=False)
```

```
Order ID           1000
Shipping Address   1000
Customer Email      996
Customer Name       993
Order Date          305
Total               100
Product Name         10
Quantity             10
Price                10
Category              6
dtype: int64
```

In this post we provide a detailed introduction to essential core functions in the Pandas library, which are crucial for data analysis. These functions make

data exploration more straightforward and efficient, helping analysts quickly grasp the structure and characteristics of a dataset.

With these must-know functions, data analysts can gain deeper insights into the data, uncover the stories behind it, and provide data-driven support for decision-making. These Pandas functions are an indispensable part of the data analysis workflow, enhancing both the efficiency and the depth of analysis.

Thanks for your reading.

Python Pandas | Pandas Tutorial | Pandas Data Preprocessing | Python Pandas Dataframe

Written by Gen. David L.

453 Followers · 4 Following

AI practitioner & python coder to record what I learned in python project development

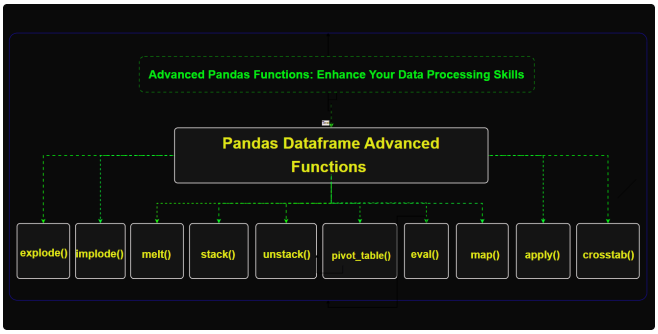
Follow

No responses yet

What are your thoughts?

Respond

More from Gen. David L.



Gen. David L.

Advanced Pandas Features: Enhance Your Data Processing...

The essence of data analysis lies in uncovering the stories behind the data. In thi...

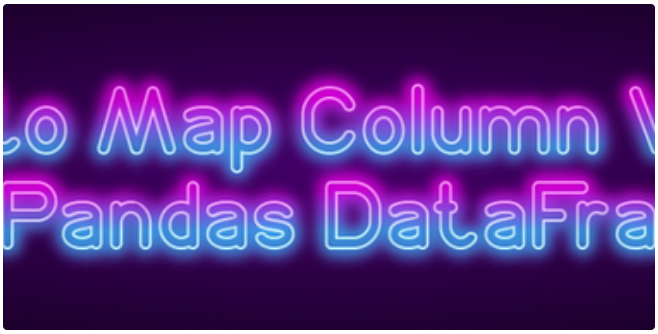
Dec 5, 2024 105

Gen. David L.

ETL-PIPES: An Efficient Python ETL Data Processing Library

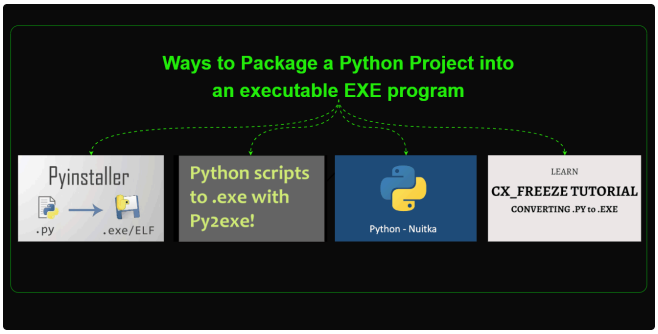
ETL-pipes is a powerful and flexible Python library specifically designed for ETL (Extract...

Nov 30, 2024 45 6



Gen. David L.

How to Map Column Values in a Pandas DataFrame?



Gen. David L.

Four Ways to Package a Python Project into an executable EXE...

Mapping column values refers to replacing specific values in a column with other values,...

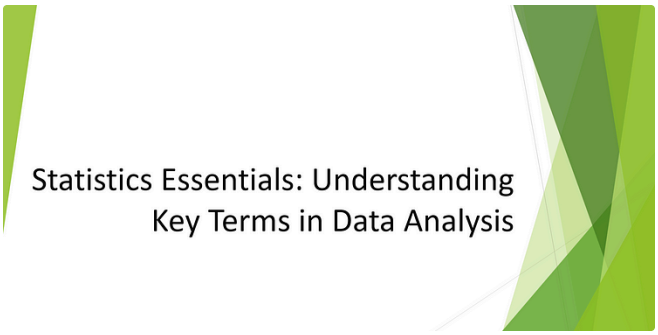
Dec 24, 2024 81

In Python, packaging a project into an executable EXE file is a common task,...

Sep 14, 2024 94

See all from Gen. David L.

Recommended from Medium



Anita Gupta

Statistics Handbook for Data Analysts

Why This Handbook?

Sep 14, 2024 422 11

Harold Finch

Top 25 Python Scripts To Automate Your Daily Tasks

Python is an excellent tool for automating daily tasks, thanks to its simplicity and a wid...

Nov 19, 2024 337 5

Lists

Staff picks

798 stories · 1568 saves

Stories to Help You Level-Up at Work

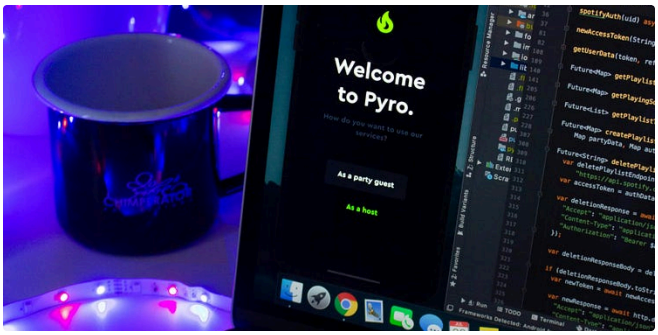
19 stories · 917 saves

Self-Improvement 101

20 stories · 3214 saves

Productivity 101

20 stories · 2716 saves



In The Pythoneers by Kevin Meneses González

5 Ways to Make Money with Python in 2025

The future belongs to those who believe in the beauty of their dreams.” — Eleanor Roosevelt

Dec 8, 2024 282 10

In Python in Plain English by Satyam Sahu

How I Automated Data Cleaning in Python Using Functions and...

Discover the key Python techniques that transformed my data-cleaning workflow fro...

Nov 4, 2024 216 7

Gen. David L.

Lorenzo Uriel

Starting a Python Project

How to Split Columns in a DataFrame Using Python Pandas

Splitting columns is a common data manipulation operation in Pandas. It allows u...

Dec 21, 2024 62

This guide documents the steps I take to set up the foundation of any Python project, wit...

Nov 19, 2024 208 1

See more recommendations